Numerik-Vorlesungen

Teil 2

Anfangs- und Randwertprobleme, lineare Gleichungssysteme

Peter Szyler, Horst Hollatz

Letzte Änderung: 20. Januar 2006

Inhaltsverzeichnis

Kapitel 6

Anfangswertprobleme

6.1. Einführung

Wir betrachten das folgende Problem:
Gegeben seien eine Teilmenge $D \subseteq \mathbb{R}$ und eine Funktion

$$\boldsymbol{f} : D \times \mathbb{R}^d \longrightarrow \mathbb{R}^d.$$

Gesucht ist eine differenzierbare Funktion

$$\boldsymbol{y} : D \longrightarrow \mathbb{R}^d,$$

die die Gleichung

$$\boldsymbol{y}'(x) = \boldsymbol{f}(x, \boldsymbol{y}(x))$$

erfüllt.
Diese vektorielle Differentialgleichung erster Ordnung hat im allgemeinen unendlich viele Lösungen. Oft sucht man nach Lösungen, die gewisse Zusatzbedingungen erfüllen. Sind diese Bedingungen in der Form

$$\boldsymbol{y}(x_0) = \boldsymbol{y}_0$$

gegeben, so spricht man von einem **Anfangswertproblem**[1]. Solche Probleme werden in diesem Kapitels untersucht. Allgemeiner als AWP sind Randwertprobleme[2], die wir im nächsten Kapitel behandeln werden. Bei ihnen sind zusätzliche Bedingungen der Form

$$\boldsymbol{r}(\boldsymbol{y}(a), \boldsymbol{y}(b)) = 0$$

[1] Wir werden statt Anfangswertproblem im weiteren die Abkürzung AWP verwenden.
[2] Für diese werden wir die Abkürzung RWP verwenden.

gegeben.
Bemerkung: Die Beschränkung auf Differentialgleichungen erster Ordnung ist keine Einschränkung. Jede Differentialgleichung m-ter Ordnung lässt sich formal in eine Differentialgleichung erster Ordnung überführen:
Wir definieren

$$\boldsymbol{z}_1(x) = \boldsymbol{y}(x), \quad \boldsymbol{z}_2(x) = \boldsymbol{y}'(x), \quad \boldsymbol{z}_3(x) = \boldsymbol{y}''(x), \ldots, \boldsymbol{z}_m(x) = \boldsymbol{y}^{(m-1)}(x).$$

Dann gilt

$$\boldsymbol{z}_1' = \boldsymbol{z}_2, \quad \boldsymbol{z}_2' = \boldsymbol{z}_3, \quad \boldsymbol{z}_3' = \boldsymbol{z}_4, \ldots, \boldsymbol{z}_m' = \boldsymbol{f}(x, \boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m).$$

Mit

$$\boldsymbol{z}(x) = \begin{pmatrix} \boldsymbol{z}_1(x) \\ \boldsymbol{z}_2(x) \\ \vdots \\ \boldsymbol{z}_m(x) \end{pmatrix} : D \subseteq \mathbb{R} \longrightarrow \mathbb{R}^{m \cdot d}$$

und

$$\boldsymbol{F}(x, \boldsymbol{z}) = \begin{pmatrix} \boldsymbol{z}_2 \\ \boldsymbol{z}_3 \\ \vdots \\ \boldsymbol{z}_m \\ \boldsymbol{f}(x, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_m) \end{pmatrix} : D \times \mathbb{R}^{m \cdot d} \longrightarrow \mathbb{R}^{m \cdot d}$$

erhalten wir die Differentialgleichung erster Ordnung

$$\boldsymbol{z}' = \boldsymbol{F}(x, \boldsymbol{z}).$$

Die numerische Aufgabenstellung unterscheidet sich von der analytischen. Wir suchen nicht nach einem formelmäßigen Ausdruck für die Funktion $\boldsymbol{y}$, die das AWP

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{y}(x_0) = \boldsymbol{y}_0$$

löst, sondern wir werden versuchen, Näherungswerte

$$\boldsymbol{\eta}_i = \boldsymbol{\eta}(x_i)$$

an gewissen Stellen x_i zu bestimmen, die die exakten Werte

$$\boldsymbol{y}_i = \boldsymbol{y}(x_i)$$

möglichst gut approximieren.
Bevor wir zum Konstruieren und Untersuchen solcher Verfahren zum Berechnen der Näherungen $\boldsymbol{\eta}_i$ kommen, ist zu untersuchen, unter welchen Bedingungen ein AWP lösbar ist. Aus der Analysis ist dazu der folgende Satz bekannt.

6.1. Existenz- und Eindeutigkeitssatz: *Die Funktion $\boldsymbol{f}$ sei auf dem Streifen*

$$S = [a,b] \times \mathbb{R}^d \quad (a \text{ und } b \text{ endlich})$$

definiert und stetig. Weiterhin existiere eine Konstante L, so dass für alle $x \in [a,b]$ und alle $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{R}^n$

$$\|\boldsymbol{f}(x,\boldsymbol{y}_1) - \boldsymbol{f}(x,\boldsymbol{y}_2)\| \leqq L\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|$$

gelte. Dann existiert zu jedem $x_0 \in [a,b]$ und jedem $\boldsymbol{y}_0 \in \mathbb{R}^n$ genau eine Funktion $\boldsymbol{y}$ mit

1. $\boldsymbol{y} \in F^1[a,b]$,
2. $\boldsymbol{y}' = \boldsymbol{f}(x,\boldsymbol{y})$ *für alle* $x \in [a,b]$ *und*
3. $\boldsymbol{y}(x_0) = \boldsymbol{y}_0$.

Für eine stetige Funktion $\boldsymbol{f}$, die bezüglich $\boldsymbol{y}$ zusätzlich lipschitzstetig ist, existiert also genau eine Lösung des AWP.
Weiterhin führen wir folgende Bezeichnung ein: Es sei $\boldsymbol{F}^n[a,b]$ die Menge aller Funktionen

$$\boldsymbol{f} : [a,b] \times \mathbb{R}^d \longrightarrow \mathbb{R}^d,$$

für die alle partiellen Ableitungen bis zur Ordnung n auf dem Streifen $S = [a,b] \times \mathbb{R}^d$ existieren, dort stetig und beschränkt sind. Insbesondere erfüllen alle Funktionen $\boldsymbol{f} \in \boldsymbol{F}^1[a,b]$ die Voraussetzungen von Satz 6.1. Aus numerischer Sicht ist die Abhängigkeit der Lösung von den Eingabedaten interessant.

6.2. Beispiel: Wir betrachten das AWP

$$y' = 10\left(y - \frac{x^2}{1+x^2}\right) + \frac{2x}{(1+x^2)^2}, \quad y(0) = y_0 = 0.$$

Die exakte Lösung lautet

$$y(x) = \frac{x^2}{1+x^2}.$$

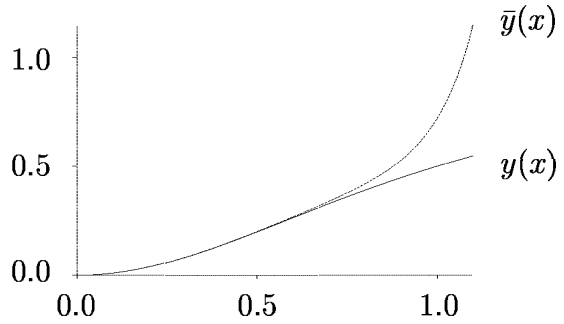
Ändern wir den Anfangswert leicht ab:

$$y(0) = \overline{y_0} = \varepsilon,$$

so erhalten wir die Lösung

$$\overline{y}(x) = \varepsilon e^{10x} + \frac{x^2}{1+x^2}.$$

Für hinreichend großes x wird diese Lösung i. a. beliebig stark von der ersten Lösung abweichen. Das folgende Bild zeigt $y(x)$ und $\overline{y}(x)$ für $\varepsilon = 0.00001$.



♡

Diese starke Abhängigkeit der Lösung von den Anfangswerten ist kein Einzelfall.

6.3. Satz: *Die Funktion $\boldsymbol{f}$ erfülle die Bedingungen aus Satz 6.1. Dann gilt für die Lösung $\boldsymbol{y}(x;\boldsymbol{s})$ des AWP*

$$\boldsymbol{y}' = \boldsymbol{f}(x,\boldsymbol{y}), \quad \boldsymbol{y}(x_0;\boldsymbol{s}) = \boldsymbol{s}$$

für ein beliebiges $x \in [a,b]$ die Abschätzung

$$\|\boldsymbol{y}(x;\boldsymbol{s}_1) - \boldsymbol{y}(x;\boldsymbol{s}_2)\| \leq e^{L|x-x_0|}\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|.$$

Beweis: Es gilt

$$\boldsymbol{y}(x;\boldsymbol{s}) = \boldsymbol{s} + \int_{x_0}^{x} \boldsymbol{f}(t,\boldsymbol{y}(t;\boldsymbol{s}))\,dt.$$

Daraus folgt

$$\boldsymbol{y}(x;\boldsymbol{s}_1) - \boldsymbol{y}(x;\boldsymbol{s}_2) = \boldsymbol{s}_1 - \boldsymbol{s}_2 + \int_{x_0}^{x} \left[\boldsymbol{f}(t,\boldsymbol{y}(t;\boldsymbol{s}_1)) - \boldsymbol{f}(t,\boldsymbol{y}(t;\boldsymbol{s}_2))\right] dt,$$

und weiter

$$\begin{aligned}
\|\boldsymbol{y}(x;\boldsymbol{s}_1) - \boldsymbol{y}(x;\boldsymbol{s}_2)\| &\leq \|\boldsymbol{s}_1 - \boldsymbol{s}_2\| + \left| \int_{x_0}^{x} \|\boldsymbol{f}(t,\boldsymbol{y}(t;\boldsymbol{s}_1)) - \boldsymbol{f}(t,\boldsymbol{y}(t;\boldsymbol{s}_2))\|\,dt \right| \\
&\leq \|\boldsymbol{s}_1 - \boldsymbol{s}_2\| + L\left| \int_{x_0}^{x} \|\boldsymbol{y}(t;\boldsymbol{s}_1) - \boldsymbol{y}(t;\boldsymbol{s}_2)\|\,dt \right|.
\end{aligned}$$

Es sei

$$\mathbf{\Phi}(x) = \int_{x_0}^{x} \|\boldsymbol{y}(t;\boldsymbol{s}_1) - \boldsymbol{y}(t;\boldsymbol{s}_2)\|\, dt$$

und damit

$$\mathbf{\Phi}'(x) = \|\boldsymbol{y}(x;\boldsymbol{s}_1) - \boldsymbol{y}(x;\boldsymbol{s}_2)\|\,.$$

Wir betrachten den Fall $x \geqq x_0$. Für die Funktion

$$\alpha(x) = \mathbf{\Phi}'(x) - L\mathbf{\Phi}(x)$$

gilt dann die Abschätzung

$$\alpha(x) \leqq \|\boldsymbol{s}_1 - \boldsymbol{s}_2\|$$

und $\mathbf{\Phi}$ ist Lösung des AWP

$$\mathbf{\Phi}'(x) = \alpha(x) + L\mathbf{\Phi}(x), \quad \mathbf{\Phi}(x_0) = 0.$$

Die Lösung dieses AWP lässt sich in der Form

$$\mathbf{\Phi}(x) = e^{L(x-x_0)} \int_{x_0}^{x} \alpha(t) e^{-L(t-x_0)}\, dt$$

schreiben. Wegen $\alpha(x) \leqq \|\boldsymbol{s}_1 - \boldsymbol{s}_2\|$ folgt für $x \geqq x_0$ die Abschätzung

$$\begin{aligned} 0 \leqq \mathbf{\Phi}(x) &\leqq e^{L(x-x_0)} \|\boldsymbol{s}_1 - \boldsymbol{s}_2\| \int_{x_0}^{x} e^{-L(t-x_0)}\, dt \\ &= \frac{1}{L} e^{L(x-x_0)} \|\boldsymbol{s}_1 - \boldsymbol{s}_2\| \left(-e^{-L(x-x_0)} + 1\right) \\ &= \frac{1}{L} \|\boldsymbol{s}_1 - \boldsymbol{s}_2\| \left(e^{L(x-x_0)} - 1\right). \end{aligned}$$

Damit gilt

$$\begin{aligned} \|\boldsymbol{y}(x;\boldsymbol{s}_1) - \boldsymbol{y}(x;\boldsymbol{s}_2)\| &= \mathbf{\Phi}'(x) = \alpha(x) + L\mathbf{\Phi}(x) \\ &\leqq \|\boldsymbol{s}_1 - \boldsymbol{s}_2\| + \|\boldsymbol{s}_1 - \boldsymbol{s}_2\| \left(e^{L(x-x_0)} - 1\right) \\ &= \|\boldsymbol{s}_1 - \boldsymbol{s}_2\| e^{L(x-x_0)}. \end{aligned}$$

Im Falle $x < x_0$ folgt aus

$$\|\boldsymbol{y}(x;\boldsymbol{s}_1) - \boldsymbol{y}(x;\boldsymbol{s}_2)\| \leqq \|\boldsymbol{s}_1 - \boldsymbol{s}_2\| - L \int_{x_0}^{x} \|\boldsymbol{y}(t;\boldsymbol{s}_1) - \boldsymbol{y}(t;\boldsymbol{s}_2)\|\, dt$$

und

$$\alpha(x) = \Phi'(x) + L\Phi(x)$$

wieder

$$\alpha(x) \leqq \|\boldsymbol{s}_1 - \boldsymbol{s}_2\|$$

und

$$-\Phi(x) \leqq \frac{1}{L}\|\boldsymbol{s}_1 - \boldsymbol{s}_2\| \left(e^{-L(x-x_0)} - 1\right).$$

Damit gilt dann

$$\|\boldsymbol{y}(x;\boldsymbol{s}_1) - \boldsymbol{y}(x;\boldsymbol{s}_2)\| \leqq \|\boldsymbol{s}_1 - \boldsymbol{s}_2\| \left(e^{-L(x-x_0)} - 1\right) = \|\boldsymbol{s}_1 - \boldsymbol{s}_2\| \left(e^{L|x-x_0|} - 1\right).$$

♣

Dieser Satz sagt aus, dass die Lösung eines AWP stetig von den Anfangswerten abhängt. Trotzdem unterscheiden sich Lösungen beliebig stark voneinander. Wie das Beispiel zeigt, ist die im Satz angegebene Abschätzung scharf.

6.2. Einschrittverfahren

6.2.1. Grundlegende Begriffe und Euler-Verfahren

Wir betrachten zunächst ein AWP im $\mathbb{R}^1$:
Gesucht ist eine Funktion $y \in C^1[a,b]$ mit

$$y' = f(x,y), \quad y(x_0) = y_0.$$

Die Funktion $f : [a,b] \times \mathbb{R} \longrightarrow \mathbb{R}$ erfülle alle Bedingungen aus Satz 6.1. Damit existiert genau eine Lösung des AWP. Als einfachste Methode zur numerischen Behandlung dieses Problems bietet es sich an, die Ableitung $y'(x)$ durch den Differenzenquotienten

$$\frac{y(x+h) - y(x)}{h}$$

mit einer geeigneten Schrittweite h zu ersetzen. Wir erhalten

$$\begin{aligned} \frac{y(x+h)-y(x)}{h} &\approx f(x,y(x)), \\ y(x+h) &\approx y(x)+hf(x,y(x)). \end{aligned}$$

Gehen wir von den exakten Werten $y(x)$ zu Näherungswerten $\eta(x)$ über und ersetzen „$\approx$“ durch „$=$“, so erhalten wir das folgende Verfahren.

6.4. EULER**sches Polygonzugverfahren:**

S0 *Setze* $\boldsymbol{\eta_0 = y_0}$ *und* $k = 0$.

S1 *Berechne*

$$\boldsymbol{\eta_{k+1}} = \boldsymbol{\eta_k} + h\boldsymbol{f}(x_k, \boldsymbol{\eta_k}),$$

$$x_{k+1} = x_k + h.$$

S2 *Setze* $k = k+1$ *und gehe zu Schritt* **S1**.

Das EULERsches Polygonzugverfahren ist ein typisches Einschrittverfahren[3]. Für das Berechnen der neuen Näherung $\boldsymbol{\eta_{k+1}}$ an der Stelle x_{k+1} wird nur **eine** alte Näherung $\boldsymbol{\eta_k}$ an der Stelle x_k benötigt. Ein ESV hat damit folgende Struktur:

6.5. Allgemeines Einschrittverfahren:

S0 *Setze* $\boldsymbol{\eta_0 = y_0}$ *und* $k = 0$.

S1 *Berechne*

$$\boldsymbol{\eta_{k+1}} = \boldsymbol{\eta_k} + h\Phi(x_k, \boldsymbol{\eta_k}; h)$$

$$x_{k+1} = x_k + h.$$

S2 *Setze* $k = k+1$ *und gehe zu Schritt* **S1**.

Bemerkung: Im Schritt **S1** muss nicht jeweils die gleiche Schrittweite verwendet werden. Wir dürfen ihn also zu:

> Wähle Schrittweite h_k und berechne $\boldsymbol{\eta_{k+1}} = \boldsymbol{\eta_k} + h_k\Phi(x_k, \boldsymbol{\eta_k}; h_k)$ und $x_{k+1} = x_k + h_k$

abändern.
Bevor wir weitere Verfahren konstruieren, wollen wir uns mit zwei wichtigen Größen

[3] Wir werden Einschrittverfahren im weiteren Text durch ESV abkürzen.

beschäftigen, durch die wesentliche Eigenschaften von ESV charakterisiert sind.
Es sei $\boldsymbol{z}$ die exakte Lösung des AWP

$$\boldsymbol{z}'(t) = \boldsymbol{f}(t, \boldsymbol{z}(t)), \quad \boldsymbol{z}(x) = \boldsymbol{y}$$

und

$$\Delta(x, \boldsymbol{y}; h) = \begin{cases} \dfrac{\boldsymbol{z}(x+h) - \boldsymbol{z}(x)}{h} = \dfrac{\boldsymbol{z}(x+h) - \boldsymbol{y}}{h} & \text{für} \quad h \neq 0, \\ \boldsymbol{f}(x, \boldsymbol{y}) & \text{für} \quad h = 0. \end{cases}$$

Φ sei die Iterationsfunktion eines allgemeinen ESV. Dann heißt die Größe

$$\boldsymbol{\tau}(x, \boldsymbol{y}; h) = \Delta(x, \boldsymbol{y}; h) - \Phi(x, \boldsymbol{y}; h)$$

lokaler Diskretisierungsfehler an der Stelle $(x, \boldsymbol{y})$ zur Schrittweite h. Führt man ausgehend von einer Stelle $(\overline{x}, \overline{\boldsymbol{y}})$ einen Schritt eines ESV aus, so gibt die Größe $h\boldsymbol{\tau}(\overline{x}, \overline{\boldsymbol{y}}; h)$ an, wie stark die berechnete Näherung

$$\boldsymbol{\eta}(\overline{x} + h) = \overline{\boldsymbol{y}} + h\Phi(\overline{x}, \overline{\boldsymbol{y}};, h)$$

von dem Wert abweicht, der sich bei exakter Lösung des AWP

$$\boldsymbol{z}'(x) = \boldsymbol{f}(x, \boldsymbol{z}(x)), \quad \boldsymbol{z}(\overline{x}) = \overline{\boldsymbol{y}}$$

an der Stelle $\overline{x} + h$ ergeben würde. Oder anders gesagt: Der lokale Diskretisierungsfehler ist ein Maß dafür, wie gut die Gleichung des Näherungsverfahrens im Punkt $(\overline{x}, \overline{\boldsymbol{y}})$ durch die exakte Lösung des entsprechenden AWP erfüllt wird.
Eine erste Forderung an ein Näherungsverfahren wäre die, dass der lokale Diskretisierungsfehler für $h \to 0$ verschwinden soll. Es soll also

$$\lim_{h \to 0} \boldsymbol{\tau}(x, \boldsymbol{y}; h) = \lim_{h \to 0} \Delta(x, \boldsymbol{y}; h) - \lim_{h \to 0} \Phi(x, \boldsymbol{y}; h) = 0$$

und damit

$$\lim_{h \to 0} \Phi(x, \boldsymbol{y}; h) = \boldsymbol{f}(x, \boldsymbol{y})$$

gelten. Ein Verfahren, das diese Bedingung für alle Funktionen $\boldsymbol{f} \in \boldsymbol{F}^1[a, b]$ erfüllt, heißt **konsistent**. Das EULERsche Polygonzugverfahren ist offensichtlich konsistent.
Von praktisch größerem Interesse als der lokale Diskretisierungsfehler ist der globale Diskretisierungsfehler.
Es sei $\boldsymbol{y}$ die exakte Lösung des AWP

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{y}(x_0) = \boldsymbol{y}_0.$$

$\boldsymbol{\eta}(x;h)$ bezeichne die mit einem ESV und der Schrittweite h berechnete Näherung für $\boldsymbol{y}$ an der Stelle x. Dann heißt die Größe

$$\boldsymbol{e}(x;h) = \boldsymbol{\eta}(x;h) - \boldsymbol{y}(x)$$

globaler Diskretisierungsfehler an der Stelle x zur Schrittweite h.
Der globale Diskretisierungsfehler gibt also an, wie stark die berechnete Näherung von der exakten Lösung des AWP abweicht.
Hier wird man natürlich auch fordern, dass der globale Diskretisierungsfehler für $h \to 0$ verschwinden soll, dass also

$$\lim_{h\to 0} \boldsymbol{\eta}(x;h) = \boldsymbol{y}(x)$$

gelte. Ein Verfahren, dass diese Bedingung für alle Funktionen $\boldsymbol{f} \in \boldsymbol{F}^1[a,b]$ und alle $x \in [a,b]$ erfüllt, heißt **konvergent**. Wie wir später sehen werden hängen Konvergenz und Konsistenz von ESV eng zusammen.
Es ist zu erwarten, dass ein ESV um so bessere Näherungen liefert, je schneller der lokale Diskretisierungsfehler mit $h \to 0$ gegen Null konvergiert. Das führt zu folgender Definition. Das durch Φ erzeugte ESV zum Lösen des AWP

$$\boldsymbol{y}' = \boldsymbol{f}(x,\boldsymbol{y}), \quad \boldsymbol{y}(x_0) = \boldsymbol{y}_0$$

hat die Konsistenzordnung p, falls

$$\|\boldsymbol{\tau}(x,\boldsymbol{y};h)\| = O(h^p)$$

für alle $(x,\boldsymbol{y}) \in S$ und alle Funktionen $\boldsymbol{f} \in \boldsymbol{F}^p[a,b]$ gilt.
Die Konsistenzordnung eines Verfahrens lässt sich durch TAYLOR-Entwicklung des lokalen Diskretisierungsfehlers bestimmen. Es gilt

$$\boldsymbol{\tau}(x,\boldsymbol{y};h) = \Delta(x,\boldsymbol{y};h) - \boldsymbol{\Phi}(x,\boldsymbol{y};h) = \frac{\boldsymbol{z}(x+h)-\boldsymbol{y}}{h} - \boldsymbol{\Phi}(x,\boldsymbol{y};h)$$

mit

$$\boldsymbol{z}(x+h) = \boldsymbol{z}(x) + h\boldsymbol{z}'(x) + \frac{h^2}{2}\boldsymbol{z}''(x) + \frac{h^3}{6}\boldsymbol{z}'''(x) + \cdots + \frac{h^p}{p!}\boldsymbol{z}^{(p)}(x+\vartheta h),\ \vartheta \in (0,1).$$

Da $\boldsymbol{z}$ Lösung des AWP

$$\boldsymbol{z}'(t) = \boldsymbol{f}(t,\boldsymbol{z}(t)), \quad \boldsymbol{z}(x) = \boldsymbol{y}$$

ist, gilt

$$\begin{aligned}
\boldsymbol{z}(x) &= \boldsymbol{y}, \\
\boldsymbol{z}'(x) &= \boldsymbol{f}(x,\boldsymbol{z}(x)) = \boldsymbol{f}(x,\boldsymbol{y}) \\
\boldsymbol{z}''(x) &= \boldsymbol{f}_x(x,\boldsymbol{z}(x)) + \boldsymbol{f}_y(x,\boldsymbol{y})\boldsymbol{z}'(x) = \boldsymbol{f}_x(x,\boldsymbol{y}) + \boldsymbol{f}_y(x,\boldsymbol{y})\boldsymbol{f}(x,\boldsymbol{y}) \\
\boldsymbol{z}'''(x) &= \boldsymbol{f}_{xx}(x,\boldsymbol{y}) + 2\boldsymbol{f}_{xy}(x,\boldsymbol{y})\boldsymbol{f}(x,\boldsymbol{y}) + \boldsymbol{f}_{yy}(x,\boldsymbol{y})\left[\boldsymbol{f}(x,\boldsymbol{y})\right]^2 + \\
&\qquad + \boldsymbol{f}_y(x,\boldsymbol{y})\left[\boldsymbol{f}_x(x,\boldsymbol{y}) + \boldsymbol{f}_y(x,\boldsymbol{y})\boldsymbol{f}(x,\boldsymbol{y})\right] \\
\vdots &\quad \vdots
\end{aligned}$$

Damit erhält man

$$\begin{aligned}
\frac{\boldsymbol{z}(x+h)-\boldsymbol{y}}{h} &= \boldsymbol{f} + \frac{h}{2}\left[\boldsymbol{f}_x + \boldsymbol{f}_y\boldsymbol{f}\right] \\
&\quad + \frac{h^2}{6}\left[\boldsymbol{f}_{xx} + 2\boldsymbol{f}_{xy}\boldsymbol{f} + \boldsymbol{f}_{yy}\boldsymbol{f}^2 + \boldsymbol{f}_y\left(\boldsymbol{f}_x + \boldsymbol{f}_y\boldsymbol{f}\right)\right] + \cdots,
\end{aligned}$$

wobei der besseren Lesbarkeit halber die Argumente der Funktion $\boldsymbol{f}$ und ihrer Ableitungen fortgelassen wurden. Hat nun Φ die Entwicklung

$$\Phi(x,\boldsymbol{y};h) = \boldsymbol{f}(x;\boldsymbol{y}) + h\boldsymbol{\varphi}_1(x;\boldsymbol{y}) + h^2\boldsymbol{\varphi}_2(x;\boldsymbol{y}) + h^3\boldsymbol{\varphi}_3(x;\boldsymbol{y}) + \cdots,$$

so ergibt sich

$$\begin{aligned}
\boldsymbol{\tau}(x,\boldsymbol{y};h) &= \frac{h}{2}\left[\boldsymbol{f}_x + \boldsymbol{f}_y\boldsymbol{f} - 2\boldsymbol{\varphi}_1\right] \\
&\quad + \frac{h^2}{6}\left[\boldsymbol{f}_{xx} + 2\boldsymbol{f}_{xy}\boldsymbol{f} + \boldsymbol{f}_{yy}\boldsymbol{f}^2 + \boldsymbol{f}_y\left(\boldsymbol{f}_x + \boldsymbol{f}_y\boldsymbol{f}\right) - 6\boldsymbol{\varphi}_2\right] + \cdots
\end{aligned}$$

Für das EULER-Verfahren gilt $\Phi(x,\boldsymbol{y};h) = \boldsymbol{f}(x;\boldsymbol{y})$ und damit

$$\boldsymbol{\tau}(x,\boldsymbol{y};h) = \frac{h}{2}\left[\boldsymbol{f}_x + \boldsymbol{f}_y\boldsymbol{f}\right] + \cdots,$$

also

$$\|\boldsymbol{\tau}(x,\boldsymbol{y};h)\| = O(h).$$

Das EULER-Verfahren hat die Konsistenzordnung $p = 1$. Um Verfahren höherer Ordnung zu bekommen, könnte man nun die Funktionen $\boldsymbol{\varphi}_1(x;\boldsymbol{y})$, $\boldsymbol{\varphi}_2(x;\boldsymbol{y})$, ..., so wählen, dass die entsprechenden Glieder der Entwicklung des lokalen Diskretisierungsfehlers verschwinden. Man erhält

$$\Phi(x,\boldsymbol{y};h) = \boldsymbol{f}(x;\boldsymbol{y}) + \frac{h}{2}\left[\boldsymbol{f}_x(x;\boldsymbol{y}) + \boldsymbol{f}_y(x;\boldsymbol{y})\boldsymbol{f}(x;\boldsymbol{y})\right]$$

oder

$$\begin{aligned}\Phi(x,\boldsymbol{y};h) &= \boldsymbol{f}(x;\boldsymbol{y})+\frac{h}{2}\left[\boldsymbol{f}_x(x;\boldsymbol{y})+\boldsymbol{f}_y(x;\boldsymbol{y})\boldsymbol{f}(x;\boldsymbol{y})\right]+\\ &\quad+\frac{h^2}{6}\Big[\boldsymbol{f}_{xx}(x;\boldsymbol{y})+2\boldsymbol{f}_{xy}(x;\boldsymbol{y})\boldsymbol{f}+\boldsymbol{f}_{yy}(x;\boldsymbol{y})\left(\boldsymbol{f}(x;\boldsymbol{y})\right)^2+\\ &\qquad+\boldsymbol{f}_y(x;\boldsymbol{y})\left[\boldsymbol{f}_x(x;\boldsymbol{y})+\boldsymbol{f}_y(x;\boldsymbol{y})\boldsymbol{f}(x;\boldsymbol{y})\right]\Big].\end{aligned}$$

Diese Verfahren sind aber praktisch wertlos. In vielen Fällen ist für die Funktion $\boldsymbol{f}$ kein analytischer Ausdruck bekannt, so dass das Berechnen der partiellen Ableitungen nur näherungsweise erfolgen könnte. Ist dagegen ein analytischer Ausdruck für die Funktion $\boldsymbol{f}$ bekannt, so ist die Auswertung der partiellen Ableitungen oft aufwendiger als die Auswertung der Funktion $\boldsymbol{f}$ selbst. Dazu kommt noch ein hoher Programmieraufwand zur Implementierung der Ableitungen von $\boldsymbol{f}$. Im nächsten Abschnitt werden wir eine Möglichkeit zum Konstruieren von ESV höherer Ordnung kennenlernen, die nur Funktionsberechnungen von $\boldsymbol{f}$ benötigen.

6.2.2. Runge-Kutta-Verfahren

Wir machen für die Funktion Φ folgenden Ansatz:

$$\Phi(x,\boldsymbol{y};h)=\sum_{k=0}^{n}\alpha_k\boldsymbol{f}(\xi_k,\boldsymbol{\eta}_k)$$

mit

$$\xi_0=x,\quad \xi_k=x+\vartheta_k h,\quad k=1,\ldots,n$$

und

$$\boldsymbol{\eta}_0=\boldsymbol{y},\quad \boldsymbol{\eta}_k=\boldsymbol{y}+h\sum_{l=0}^{k-1}\beta_{kl}\boldsymbol{f}(\xi_l,\boldsymbol{\eta}_l),\quad k=1,\ldots,n.$$

Die Parameter $\alpha_0,\ldots,\alpha_n$, $\vartheta_1,\ldots,\vartheta_n$ und $\beta_{10},\ldots,\beta_{n,n-1}$ werden so bestimmt, dass in der Entwicklung des lokalen Diskretisierungsfehlers möglichst viele Glieder verschwinden. Betrachten wir die einfachsten Fälle $n=0$ und $n=1$.

$n=0$ Hier gilt $\Phi(x,\boldsymbol{y};h)=\alpha_0\boldsymbol{f}(x,\boldsymbol{y})$, und es folgt sofort $\alpha_0=1$. Man erhält das EULER-Verfahren mit der Konsistenzordnung $p=1$.

$n=1$ Wir machen für $\Phi(x,\boldsymbol{y};h)$ den Ansatz

$$\Phi(x,\boldsymbol{y};h)=\alpha_0\boldsymbol{f}(x,\boldsymbol{y})+\alpha_1\boldsymbol{f}(x+\vartheta_1 h,\boldsymbol{y}+h\beta_{10}\boldsymbol{f}(x,\boldsymbol{y})).$$

Es gilt für $h = 0$

$$\boldsymbol{\Phi}(x, \boldsymbol{y}; 0) = (\alpha_0 + \alpha_1)\boldsymbol{f}(x, \boldsymbol{y}).$$

Differentiation nach h liefert

$$\begin{aligned}\frac{d}{dh}\boldsymbol{\Phi}(x, \boldsymbol{y}; h) &= \alpha_1 \boldsymbol{f}_x(x + \vartheta_1 h, \boldsymbol{y} + h\beta_{10}\boldsymbol{f}(x, \boldsymbol{y}))\vartheta_1 + \\ &\quad + \alpha_1 \boldsymbol{f}_y(x + \vartheta_1 h, \boldsymbol{y} + h\beta_{10}\boldsymbol{f}(x, \boldsymbol{y}))\,\beta_{10}\boldsymbol{f}(x, \boldsymbol{y})\end{aligned}$$

und an der Stelle $h = 0$

$$\left.\frac{d}{dh}\boldsymbol{\Phi}(x, \boldsymbol{y}; h)\right|_{h=0} = \alpha_1\vartheta_1 \boldsymbol{f}_x(x, \boldsymbol{y}) + \alpha_1\beta_{10}\boldsymbol{f}_y(x, \boldsymbol{y})\boldsymbol{f}(x, \boldsymbol{y}).$$

Durch Koeffizientenvergleich mit der TAYLOR-Entwicklung von $\Delta(x, \boldsymbol{y}; h)$ erhält man

$$\begin{aligned} h^0 \;&: & \alpha_0 + \alpha_1 &= 1, \\ h^1 \boldsymbol{f}_x \;&: & \alpha_1\vartheta_1 &= \frac{1}{2}, \\ h^1 \boldsymbol{f}_y \boldsymbol{f} &: & \alpha_1\beta_{10} &= \frac{1}{2}. \end{aligned}$$

als Bestimmungsgleichungen für ein Verfahren der Konsistenzordnung $p = 2$. Aus den beiden letzten Gleichungen folgt $\vartheta_1 = \beta_{10}$. Ansonsten ist ein Parameter frei wählbar. Man erhält verschiedene Verfahren.

- $\alpha_0 = \alpha_1 = 1/2$ und $\beta_{10} = \vartheta_1 = 1$
  Wir erhalten das Verfahren von HEUN(1900):

  $$\boldsymbol{\Phi}(x, \boldsymbol{y}; h) = \frac{1}{2}\left[\boldsymbol{f}(x, \boldsymbol{y}) + \boldsymbol{f}(x + h, \boldsymbol{y} + h\boldsymbol{f}(x, \boldsymbol{y}))\right].$$

  Dieses Verfahren braucht zwei Auswertungen der Funktion $\boldsymbol{f}$ pro Schritt.

- $\alpha_0 = 0$, $\alpha_1 = 1$ und $\beta_{10} = \vartheta_1 = 1/2$
  Wir erhalten das modifizierte EULER-Verfahren (COLLATZ 1960):

  $$\boldsymbol{\Phi}(x, \boldsymbol{y}; h) = \boldsymbol{f}\left(x + \frac{h}{2}, \boldsymbol{y} + \frac{h}{2}\boldsymbol{f}(x, \boldsymbol{y})\right).$$

  Dieses Verfahren erfordert ebenfalls zwei Auswertungen der Funktion $\boldsymbol{f}$ pro Schritt.

Für größere n erhält man entsprechend kompliziertere Bestimmungsgleichungen für die Parameter. Wir geben einige Verfahren nach folgendem Schema an.

$$\begin{array}{c|cccc} \vartheta_1 & \beta_{10} & & & \\ \vartheta_2 & \beta_{20} & \beta_{21} & & \\ \vartheta_3 & \beta_{30} & \beta_{31} & \beta_{32} & \\ \hline & \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 \end{array}$$

$n = 0$; EULER-Verfahren; $p = 1$

$$\begin{array}{c|c} & \\ \hline & 1 \end{array}$$

$n = 1$; HEUN-Verfahren; $p = 2$

$$\begin{array}{c|cc} 1 & 1 & \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

$n = 1$; mod. EULER-Verfahren; $p = 2$

$$\begin{array}{c|cc} \frac{1}{2} & \frac{1}{2} & \\ \hline & 0 & 1 \end{array}$$

$n = 2$; einfache KUTTA-Regel; $p = 3$

$$\begin{array}{c|ccc} \frac{1}{2} & \frac{1}{2} & & \\ 1 & -1 & 2 & \\ \hline & \frac{1}{6} & \frac{4}{6} & \frac{1}{6} \end{array}$$

$n = 3$; RUNGE-KUTTA-Verfahren; $p = 4$

$$\begin{array}{c|cccc} \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array}$$

$n = 3$; $\frac{3}{8}$-Regel; $p = 4$

$$\begin{array}{c|cccc} \frac{1}{3} & \frac{1}{3} & & & \\ \frac{2}{3} & -\frac{1}{3} & 1 & & \\ 1 & 1 & -1 & 1 & \\ \hline & \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{array}$$

$n = 3$; vierstufige ENGLAND-Formel; $p = 4$;

$$\begin{array}{c|cccc} \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & & \\ 1 & 0 & -1 & 2 & \\ \hline & \frac{1}{6} & 0 & \frac{2}{3} & \frac{1}{6} \end{array}$$

$n = 3$; GILL-Modifikation des RUNGE-KUTTA-Verfahrens; $p = 4$

$$\begin{array}{c|cccc} \frac{1}{2} & & \frac{1}{2} & & \\ \frac{1}{2} & \frac{\sqrt{2}-1}{2} & \frac{2-\sqrt{2}}{2} & & \\ 1 & 0 & -\frac{\sqrt{2}}{2} & \frac{2+\sqrt{2}}{2} & \\ \hline & \frac{1}{6} & \frac{2-\sqrt{2}}{6} & \frac{2+\sqrt{2}}{6} & \frac{1}{6} \end{array}$$

$n = 3$; KUNTZMANN-Verfahren; $p = 4$

$$\begin{array}{c|cccc}
\frac{2}{5} & \frac{2}{5} & & & \\
\frac{3}{5} & -\frac{3}{20} & \frac{15}{20} & & \\
1 & \frac{19}{44} & -\frac{15}{44} & \frac{40}{44} & \\
\hline
 & \frac{11}{72} & 0 & \frac{25}{72} & \frac{11}{72}
\end{array}$$

Bemerkungen: (i) Die GILL-Modifikation des RUNGE-KUTTA-Verfahrens besitzt ein günstigeres Rundungsfehlerverhalten als das RUNGE-KUTTA-Verfahren.
(ii) Beim KUNTZMANN-Verfahren fallen in der Entwicklung des lokalen Diskretisierungsfehlers auch noch einige Terme vierter Ordnung weg.

6.2.3. Konvergenz von Einschrittverfahren

Wir betrachten wieder ein AWP

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{y}(x_0) = \boldsymbol{y}_0$$

und ein ESV

$$\left.\begin{array}{rcl}
\boldsymbol{\eta}_0 & = & \boldsymbol{y}_0, \\
\boldsymbol{\eta}_{k+1} & = & \boldsymbol{\eta}_k + h\Phi(x_k, \boldsymbol{\eta}_k; h), \\
x_{k+1} & = & x_0 + (k+1) \cdot h
\end{array}\right\} \quad k = 0, 1, \ldots$$

zur Lösungsberechnung. Wir werden nun untersuchen, wie sich der globale Diskretisierungsfehler

$$\boldsymbol{e}(x; h) = \boldsymbol{\eta}(x; h) - \boldsymbol{y}(x)$$

an einer Stelle x für $h \to 0$ verhält, wobei nur Schrittweiten

$$h \in H_x = \left\{ \frac{x - x_0}{n} \;\middle|\; n = 1, 2, \ldots \right\}$$

zulässig sind. Wir untersuchen

$$\lim_{n\to\infty} \boldsymbol{e}\left(x; \frac{x - x_0}{n}\right) = \lim_{n\to\infty} \boldsymbol{e}(x; h_n).$$

Wir werden zeigen, dass alle ESV der Konsistenzordnung $p > 0$ konvergent sind, und dass darüber hinaus die Konvergenzordnung mit der Konsistenzordnung übereinstimmt. Für den Beweis des Konvergenzsatzes benötigen wir den folgenden Hilfssatz.

6.6. Satz: *Gelten für die Zahlen ξ_i $i = 0, 1, \ldots$ die Ungleichungen*

$$|\xi_{i+1}| \leqq (1+\delta)|\xi_i| + B, \quad i = 0, 1, \ldots$$

mit gewissen Konstanten $\delta > 0$ und $B \geqq 0$, so gilt

$$|\xi_n| \leqq e^{n\delta}|\xi_0| + \frac{e^{n\delta} - 1}{\delta} B, \quad n = 0, 1, \ldots$$

Beweis: Aus der Voraussetzung folgt

$$\begin{aligned}
|\xi_1| &\leqq (1+\delta)|\xi_0| + B \\
|\xi_2| &\leqq (1+\delta)^2|\xi_0| + (1+\delta)B + B \\
&= (1+\delta)^2|\xi_0| + [(1+\delta)+1]\, B \\
|\xi_3| &\leqq (1+\delta)^3|\xi_0| + (1+\delta)\,[(1+\delta)+1]\, B + B \\
&= (1+\delta)^3|\xi_0| + \left[(1+\delta)^2 + (1+\delta) + 1\right] B \\
\vdots &\quad\ \vdots \\
|\xi_n| &\leqq (1+\delta)^n|\xi_0| + \left[(1+\delta)^{n-1} + \cdots + (1+\delta) + 1\right] B \\
&= (1+\delta)^n|\xi_0| + \frac{(1+\delta)^n - 1}{(1+\delta) - 1} B \\
&= (1+\delta)^n|\xi_0| + \frac{(1+\delta)^n - 1}{\delta} B.
\end{aligned}$$

Wegen $0 < 1 + \delta \leqq e^{\delta}$ für $\delta > 0$ folgt daraus

$$|\xi_n| \leqq (1+\delta)^n|\xi_0| + \frac{e^{n\delta} - 1}{\delta} B.$$

❀

Nun beweisen wir den eigentlichen Konvergenzsatz.

6.7. Satz: *Gegeben sei für $x_0 \in [a, b]$ und $\boldsymbol{y}_0 \in \mathbb{R}^d$ das AWP*

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{y}(x_0) = \boldsymbol{y}_0$$

mit der exakten Lösung $\boldsymbol{y}(x)$. Die Iterationsfunktion Φ eines ESV habe folgende Eigenschaften:

1. Φ *ist stetig auf dem Gebiet*

$$G = \left\{ (x, \boldsymbol{y}, h) \mid x \in [a,b], \|\boldsymbol{y} - \boldsymbol{y}(x)\| \leqq \gamma, |h| \leqq h_0 \right\}$$

   mit Konstanten $\gamma > 0$ *und* $h_0 > 0$*.*

2. *Es existiert eine Konstante* $M > 0$*, so dass*

$$\|\Phi(x, \boldsymbol{y}_1; h) - \Phi(x, \boldsymbol{y}_2; h)\| \leqq M \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|$$

   für alle $(x, \boldsymbol{y}_1, h), (x, \boldsymbol{y}_2, h) \in G$ *gilt.*

3. *Es existiert eine Konstante* $N > 0$*, so dass*

$$\|\boldsymbol{\tau}(x, \boldsymbol{y}(x); h)\| = \|\Delta(x, \boldsymbol{y}(x); h) - \Phi(x, \boldsymbol{y}(x); h)\| \leqq N |h|^p$$

   für alle $x \in [a,b]$ *und alle* $|h| \leqq h_0$ *gilt.*

Dann existiert ein $\overline{h}$ *mit* $0 < \overline{h} \leqq h_0$*, so dass*

$$\boldsymbol{e}(x, h_n)\| \leqq N \frac{e^{M|x - x_0|} - 1}{M} |h_n|^p$$

für alle $x \in [a,b]$ *und alle* $h_n = (x - x_0)/n$ *mit* $|h_n| \leqq \overline{h}$ *gilt. Für* $\gamma = \infty$ *ist* $\overline{h} = h_0$*.*

Beweis: Statt der Iterationsfunktion Φ betrachten wir die Funktion

$$\tilde{\Phi} = \begin{cases} \Phi(x, \boldsymbol{y}; h) & (x, \boldsymbol{y}, h) \in G, \\ \Phi\left(x, \boldsymbol{y}(x) + \gamma \frac{\boldsymbol{y} - \boldsymbol{y}(x)}{\|\boldsymbol{y} - \boldsymbol{y}(x)\|}; h\right) & x \in [a,b], \quad |h| \leqq h_0, \quad \|\boldsymbol{y} - \boldsymbol{y}(x)\| > \gamma \end{cases}$$

und das durch sie definierte ESV. $\tilde{\Phi}$ ist offensichtlich stetig auf

$$\tilde{G} = \left\{ (x, \boldsymbol{y}, h) \;\middle|\; x \in [a,b], \boldsymbol{y} \in \mathbb{R}^d, |h| \leqq h_0 \right\}$$

und genügt der LIPSCHITZ-Bedingung

$$\|\tilde{\Phi}(x, \boldsymbol{y}_1; h) - \tilde{\Phi}(x, \boldsymbol{y}_2; h)\| \leqq M \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|$$

für alle $(x, \boldsymbol{y}_1, h), (x, \boldsymbol{y}_2, h) \in \tilde{G}$. Wegen $\tilde{\Phi}(x, \boldsymbol{y}(x); h) = \Phi(x, \boldsymbol{y}(x); h)$ gilt auch die dritte Voraussetzung des Satzes für das durch $\tilde{\Phi}$ definierte ESV:

$$\left.\begin{array}{rcl} \tilde{\boldsymbol{\eta}}_0 & = & \boldsymbol{y}_0, \\ \tilde{\boldsymbol{\eta}}_{k+1} & = & \tilde{\boldsymbol{\eta}}_k + h \tilde{\Phi}(x_k, \tilde{\boldsymbol{\eta}}_k; h), \\ x_{k+1} & = & x_k + h \end{array}\right\} \text{für} \quad k = 0, 1, \ldots$$

Wir wollen den globalen Diskretisierungsfehler $\tilde{\boldsymbol{e}}(x;h)$ dieses Verfahrens abschätzen. Es gilt

$$\begin{aligned}
\tilde{\boldsymbol{e}}_{k+1} &= \tilde{\boldsymbol{\eta}}_{k+1} - \boldsymbol{y}_{k+1} \\
&= \tilde{\boldsymbol{\eta}}_k + h\tilde{\boldsymbol{\Phi}}(x_k, \tilde{\boldsymbol{\eta}}_k; h) - \boldsymbol{y}_{k+1} + \boldsymbol{y}_k - \boldsymbol{y}_k + h\tilde{\boldsymbol{\Phi}}(x_k, \boldsymbol{y}_k; h) - h\tilde{\boldsymbol{\Phi}}(x_k, \boldsymbol{y}_k; h) \\
&= \tilde{\boldsymbol{e}}_k - h\left[\frac{\boldsymbol{y}_{k+1} - \boldsymbol{y}_k}{h} - \tilde{\boldsymbol{\Phi}}(x_k, \boldsymbol{y}_k; h)\right] + h\left[\tilde{\boldsymbol{\Phi}}(x_k, \tilde{\boldsymbol{\eta}}_k; h) - \tilde{\boldsymbol{\Phi}}(x_k, \boldsymbol{y}_k; h)\right] \\
&= \tilde{\boldsymbol{e}}_k - h\boldsymbol{\tau}(x_k, \boldsymbol{y}_k; h) + h\left[\tilde{\boldsymbol{\Phi}}(x_k, \tilde{\boldsymbol{\eta}}_k; h) - \tilde{\boldsymbol{\Phi}}(x_k, \boldsymbol{y}_k; h)\right].
\end{aligned}$$

Damit erhalten wir für die Norm des globalen Diskretisierungsfehlers die Abschätzung

$$\begin{aligned}
\|\tilde{\boldsymbol{e}}_{k+1}\| &\leqq \|\tilde{\boldsymbol{e}}_k\| + |h|\,\|\boldsymbol{\tau}(x, \boldsymbol{y}_k; h)\| + |h|\,\left\|\tilde{\boldsymbol{\Phi}}(x_k, \tilde{\boldsymbol{\eta}}_k; h) - \tilde{\boldsymbol{\Phi}}(x, \boldsymbol{y}_k; h)\right\| \\
&\leqq \|\tilde{\boldsymbol{e}}_k\| + |h|N|h|^p + |h|M\|\tilde{\boldsymbol{\eta}}_k - \boldsymbol{y}_k\| \\
&= (1 + M|h|)\,\|\tilde{\boldsymbol{e}}_k\| + N|h|^{p+1}.
\end{aligned}$$

Mit Satz 6.6 und $\tilde{\boldsymbol{e}}_0 = 0$ folgt daraus

$$\|\tilde{\boldsymbol{e}}_n\| \leqq \frac{e^{n|h|M} - 1}{M|h|} N|h|^{p+1} = \frac{e^{n|h|M} - 1}{M} N|h|^p.$$

Für ein festes $x \in [a, b]$, $x \neq x_0$, ist

$$h = h_n = \frac{x - x_0}{n}$$

und $\tilde{\boldsymbol{e}}_n = \tilde{\boldsymbol{e}}(x, h_n) = \tilde{\boldsymbol{e}}(x_0 + nh_n, h_n)$. Damit folgt

$$\|\tilde{\boldsymbol{e}}(x, h_n)\| \leqq \frac{e^{M|x - x_0|} - 1}{M} N|h|^p \leqq \frac{e^{M|b-a|} - 1}{M} N|h|^p.$$

Diese Abschätzung gilt für beliebige $x \in [a, b]$ und alle $|h_n| \leqq h_0$. Ist nun

$$|h_n| \leqq \bar{h} = \min\left\{\left(\frac{\gamma M}{N\left(e^{M|b-a|} - 1\right)}\right)^{1/p}, h_0\right\} > 0,$$

so gilt für alle h_n mit $|h_n| < \bar{h}$

$$\|\tilde{\boldsymbol{e}}(x, h_n)\| \leqq \frac{e^{M|b-a|} - 1}{M} N \frac{\gamma M}{N\left(e^{M|b-a|} - 1\right)} = \gamma.$$

Damit gilt aber für alle n $\tilde{\boldsymbol{\eta}}_n \in G$ und $\tilde{\boldsymbol{\Phi}}(x, \tilde{\boldsymbol{\eta}}_n; h) = \boldsymbol{\Phi}(x, \tilde{\boldsymbol{\eta}}_n; h)$. Im Falle $|h_n| < \bar{h}$ geht das durch $\tilde{\boldsymbol{\Phi}}$ erzeugte ESV in das durch $\boldsymbol{\Phi}$ erzeugte ESV über. Alle Abschätzungen gelten dann auch für $\boldsymbol{\eta}(x; h_n)$. ❀

Der letzte Satz sagt aus, dass die Konvergenzordnung eines ESV gleich seiner Konsistenzordnung ist, falls die Iterationsfunktion Φ in einer Umgebung der exakten Lösung des AWP lipschitzstetig ist. Für die RUNGE-KUTTA-Verfahren ergibt sich aber die LIPSCHITZ-Stetigkeit von Φ aus der ohnehin geforderten LIPSCHITZ-Stetigkeit von $\boldsymbol{f}$.
Man könnte nun das Ergebnis dieses Satzes dazu nutzen, um die Schrittweite zum Erreichen einer vorgegebenen Genauigkeit abzuschätzen. Dazu müsste man näherungsweise die Konstanten M und N kennen. Falls Φ hinreichend oft partiell differenzierbar ist, läuft dies auf die Abschätzung partieller Ableitungen von Φ bzw. $\boldsymbol{f}$ hinaus.
Für das EULER-Verfahren erhält man

$$M \approx \left\| \frac{\partial \Phi(x, \boldsymbol{y}; h)}{\partial \boldsymbol{y}} \right\| = \left\| \boldsymbol{f}_y(x, \boldsymbol{y}) \right\|, \quad N \approx \frac{1}{2} \left\| \boldsymbol{f}_x(x, \boldsymbol{y}) + \boldsymbol{f}_y(x, \boldsymbol{y}) \boldsymbol{f}(x, \boldsymbol{y}) \right\|.$$

Für das HEUN-Verfahren erhält man

$$M \approx \left\| \frac{\partial \Phi(x, \boldsymbol{y}; h)}{\partial \boldsymbol{y}} \right\| = \left\| \boldsymbol{f}_y(x, \boldsymbol{y}) \right\|$$

und

$$N \approx \frac{1}{12} \left\| \boldsymbol{f}_{xx} + 2\boldsymbol{f}_{xy}\boldsymbol{f} + \boldsymbol{f}_{yy}\boldsymbol{f}^2 - 2\boldsymbol{f}_y(\boldsymbol{f}_x + \boldsymbol{f}_y\boldsymbol{f}) \right\|.$$

Praktisch ist das kaum durchführbar.

6.2.4. Rundungsfehlereinfluss

Wir betrachten das ESV

$$\left.\begin{array}{rcl} \boldsymbol{\eta}_0 & = & \boldsymbol{y}_0, \\ \boldsymbol{\eta}_{k+1} & = & \boldsymbol{\eta}_k + h\Phi(x_k, \boldsymbol{\eta}_k; h), \\ x_{k+1} & = & x_k + h \end{array}\right\}, \quad k = 0, 1, \ldots$$

Auf einem Rechner treten Rundungsfehler auf, so dass das Verfahren in der Form

$$\left.\begin{array}{rcl} \hat{\boldsymbol{\eta}}_0 & = & \boldsymbol{y}_0, \\ \hat{\boldsymbol{\eta}}_{k+1} & = & \hat{\boldsymbol{\eta}}_k + h\Phi(x_k, \hat{\boldsymbol{\eta}}_k; h) + \boldsymbol{\delta}_{k+1}, \\ x_{k+1} & = & x_k + h \end{array}\right\}, \quad k = 0, 1, \ldots$$

realisiert wird; die Größe $\boldsymbol{\delta}_k$ stellt dabei den pro Schritt erzeugten Rundungsfehler dar. Es gilt der folgende Satz.

6.8. Satz: *Es sei zum Lösen eines AWP obiges ESV gegeben. Die Funktion Φ genüge den Voraussetzungen aus Satz 6.7 und der pro Schritt erzeugte Rundungsfehler sei beschränkt:*

$$\|\boldsymbol{\delta}_k\| \leqq \Delta < \infty, \quad k = 1, 2, \ldots$$

Für hinreichend großes γ gilt dann:
Der gesamte Rundungsfehler

$$\boldsymbol{r}(x_k; h) = \hat{\boldsymbol{\eta}}(x_k; h) - \boldsymbol{\eta}(x_k; h)$$

genügt der Abschätzung

$$\|\boldsymbol{r}(x_k; h)\| \leqq \frac{\Delta}{|h|} \frac{e^{M|x_k - x_0|} - 1}{M}.$$

Für den gesamten absoluten Fehler

$$\boldsymbol{v}(x_k; h) = \hat{\boldsymbol{\eta}}(x_k; h) - \boldsymbol{y}(x_k)$$

gilt die Abschätzung

$$\|\boldsymbol{v}(x_k; h)\| \leqq \left(N|h|^p + \frac{\Delta}{|h|} \right) \frac{e^{M|x_k - x_0|} - 1}{M}.$$

Beweis: Wir betrachten wieder, wie im Beweis von Satz 6.7, das durch die Funktion $\tilde{\Phi}$ erzeugte Verfahren

$$\left.\begin{array}{rcl} \tilde{\boldsymbol{\eta}}_0 & = & \boldsymbol{y}_0, \\ \tilde{\boldsymbol{\eta}}_{k+1} & = & \tilde{\boldsymbol{\eta}}_k + h\tilde{\Phi}(x_k, \tilde{\boldsymbol{\eta}}_k; h), \\ x_{k+1} & = & x_k + h \end{array}\right\}, \quad k = 0, 1, \ldots$$

Die Berücksichtigung von Rundungsfehlern führt auf das Verfahren

$$\left.\begin{array}{rcl} \hat{\tilde{\boldsymbol{\eta}}}_0 & = & \boldsymbol{y}_0, \\ \hat{\tilde{\boldsymbol{\eta}}}_{k+1} & = & \hat{\tilde{\boldsymbol{\eta}}}_k + h\tilde{\Phi}(x_k, \hat{\tilde{\boldsymbol{\eta}}}_k; h) + \tilde{\boldsymbol{\delta}}_{k+1}, \\ x_{k+1} & = & x_k + h \end{array}\right\}, \quad k = 0, 1, \ldots$$

mit

$$\|\tilde{\boldsymbol{\delta}}_k\| \leqq \tilde{\Delta} < \infty, \quad k = 1, 2, \ldots$$

Hieraus ergibt sich

$$\begin{aligned}\tilde{\boldsymbol{r}}_{k+1} &= \hat{\tilde{\boldsymbol{\eta}}}_{k+1} - \tilde{\boldsymbol{\eta}}_{k+1} \\ &= \hat{\tilde{\boldsymbol{\eta}}}_k + h\tilde{\boldsymbol{\Phi}}(x_k, \hat{\tilde{\boldsymbol{\eta}}}_k; h) + \tilde{\boldsymbol{\delta}}_{k+1} - \tilde{\boldsymbol{\eta}}_k - h\tilde{\boldsymbol{\Phi}}(x_k, \tilde{\boldsymbol{\eta}}_k; h) \\ &= \tilde{\boldsymbol{r}}_k + h\left[\tilde{\boldsymbol{\Phi}}(x_k, \hat{\tilde{\boldsymbol{\eta}}}_k; h) - \tilde{\boldsymbol{\Phi}}(x_k, \tilde{\boldsymbol{\eta}}_k; h)\right] + \tilde{\boldsymbol{\delta}}_{k+1}\end{aligned}$$

und

$$\begin{aligned}\|\tilde{\boldsymbol{r}}_{k+1}\| &\leqq \|\tilde{\boldsymbol{r}}_k\| + |h| \left\|\tilde{\boldsymbol{\Phi}}(x_k, \hat{\tilde{\boldsymbol{\eta}}}_k; h) - \tilde{\boldsymbol{\Phi}}(x_k, \tilde{\boldsymbol{\eta}}_k; h)\right\| + \|\tilde{\boldsymbol{\delta}}_{k+1}\| \\ &\leqq \|\tilde{\boldsymbol{r}}_k\| + |h| M \left\|\hat{\tilde{\boldsymbol{\eta}}}_k - \tilde{\boldsymbol{\eta}}_k\right\| + \|\tilde{\boldsymbol{\delta}}_{k+1}\| \\ &\leqq (1 + M|h|)\|\tilde{\boldsymbol{r}}_k\| + \tilde{\Delta}.\end{aligned}$$

Mit Satz 6.6 und $\|\tilde{\boldsymbol{r}}_0\| = 0$ folgt die Abschätzung

$$\|\tilde{\boldsymbol{r}}_k\| \leqq \tilde{\Delta}\frac{e^{kM|h|} - 1}{M|h|} = \frac{\tilde{\Delta}}{|h|}\frac{e^{M|x_k - x_0|} - 1}{M}.$$

Für den gesamten absoluten Fehler erhalten wir

$$\begin{aligned}\|\tilde{\boldsymbol{v}}_k\| &\leqq \|\tilde{\boldsymbol{r}}_k\| + \|\tilde{\boldsymbol{e}}_k\| \\ &\leqq \frac{\tilde{\Delta}}{|h|}\frac{e^{M|x_k - x_0|} - 1}{M} + N|h|^p \frac{e^{M|x_k - x_0|} - 1}{M} \\ &= \left(N|h|^p + \frac{\tilde{\Delta}}{|h|}\right)\frac{e^{M|x_k - x_0|} - 1}{M}\end{aligned}$$

Ist nun γ hinreichend groß, so dass Schrittweiten existieren, für die

$$\|\tilde{\boldsymbol{v}}_k\| = \|\tilde{\boldsymbol{v}}(x_k; h)\| \leqq \gamma$$

für alle $x_k \in [a, b]$ gilt, so stimmt $\tilde{\boldsymbol{\Phi}}$ für diese Schrittweiten mit $\boldsymbol{\Phi}$ überein. Alle Abschätzungen gelten dann auch für das ursprüngliche Verfahren mit Δ statt $\tilde{\Delta}$. ❀

Betrachten wir nun noch den pro Schritt erzeugten Rundungsfehler genauer. Wir beschränken uns dabei auf den eindimensionalen Fall $y : D \subset \mathbb{R} \to \mathbb{R}$. Das Berechnen von $\hat{\eta}_{k+1}$ erfolgt in folgenden Schritten:

$$a_k = gl\left(\boldsymbol{\Phi}(x_k, \hat{\eta}_k; h)\right) = \boldsymbol{\Phi}(x_k, \hat{\eta}_k; h)(1 + \alpha_k)$$

$$b_k = gl(ha_k) = ha_k(1 + \beta_k)$$

$$\begin{aligned}
\hat{\eta}_{k+1} &= gl(\hat{\eta}_k + b_k) = (\hat{\eta}_k + b_k)(1+\gamma_k) \\
&= (\hat{\eta}_k + ha_k(1+\beta_k))(1+\gamma_k) \\
&= (\hat{\eta}_k + h\boldsymbol{\Phi}(x_k, \hat{\eta}_k; h)(1+\alpha_k)(1+\beta_k))(1+\gamma_k) \\
&\doteq (\hat{\eta}_k + h\boldsymbol{\Phi}(x_k, \hat{\eta}_k; h)(1+\alpha_k+\beta_k))(1+\gamma_k) \\
&\doteq \hat{\eta}_k + h\boldsymbol{\Phi}(x_k, \hat{\eta}_k; h) + h\boldsymbol{\Phi}(x_k, \hat{\eta}_k; h)(\alpha_k+\beta_k+\gamma_k) + \hat{\eta}_k\gamma_k.
\end{aligned}$$

Damit gilt

$$\delta_{k+1} \doteq h\boldsymbol{\Phi}(x_k, \hat{\eta}_k; h)(\alpha_k+\beta_k+\gamma_k) + \hat{\eta}_k\gamma_k = h\boldsymbol{\Phi}(x_k, \hat{\eta}_k; h)(\alpha_k+\beta_k) + \hat{\eta}_{k+1}\gamma_k.$$

Normalerweise ist die Schrittweite so klein, dass

$$|h\boldsymbol{\Phi}(x_k, \hat{\eta}_k; h)y| \ll |\hat{\eta}_{k+1}|$$

gilt. Dann wird der Rundungsfehler im wesentlichen durch den Term $\hat{\eta}_{k+1}\gamma_k$, also den Fehler bei der Addition bestimmt. Es ist daher nützlich, die Addition mit einer höheren Genauigkeit durchzuführen.

6.2.5. Schrittweitensteuerung

Wie wir gesehen haben, ist die Schätzung einer günstigen Schrittweite zum Erreichen einer vorgegebenen Genauigkeit mittels Satz 6.7 praktisch nicht möglich. In diesem Abschnitt werden wir zwei andere Möglichkeiten zur Schätzung des globalen Diskretisierungsfehlers und damit zur Schrittweitensteuerung kennenlernen. Beide Verfahren beruhen darauf, dass man mit verschiedenen Methoden zwei Näherungen $\boldsymbol{\eta}_I(x)$ und $\boldsymbol{\eta}_{II}(x)$ berechnet und aus der Differenz $\boldsymbol{\eta}_I(x) - \boldsymbol{\eta}_{II}(x)$ auf die Größe des globalen Diskretisierungsfehlers schließt.
Betrachten wir ein ESV p-ter Ordnung. Es gilt

$$\boldsymbol{e}(x; h) = \boldsymbol{\eta}(x; h) - \boldsymbol{y}(x) = \boldsymbol{e}_p(x)h^p + \boldsymbol{O}(h^{p+1})$$

mit $\boldsymbol{e}_p(x_0) = \boldsymbol{o}$. Bei Vernachlässigung des Restgliedes folgt

$$\boldsymbol{\eta}(x; h) \doteq \boldsymbol{y}(x) + \boldsymbol{e}_p(x)h^p.$$

Berechnet man mit der Schrittweite $h/2$ eine weitere Näherung, so gilt

$$\boldsymbol{\eta}\left(x; \frac{h}{2}\right) \doteq \boldsymbol{y}(x) + \boldsymbol{e}_p(x)\left(\frac{h}{2}\right)^p.$$

Nun lässt sich $\boldsymbol{e}_p(x)$ näherungsweise berechnen. Es ergibt sich

$$\boldsymbol{\eta}(x; h) - \boldsymbol{\eta}\left(x; \frac{h}{2}\right) \doteq \boldsymbol{e}_p(x)\left(h^p - \left(\frac{h}{2}\right)^p\right)$$

und

$$\boldsymbol{e}_p(x)\left(\frac{h}{2}\right)^p \doteq \frac{\boldsymbol{\eta}(x;h)-\boldsymbol{\eta}\left(x;\frac{h}{2}\right)}{2^p-1}.$$

Diese Gleichung wird zunächst verwendet, um die berechneten Näherungen zu verbessern. Man erhält

$$\boldsymbol{y}(x) \doteq \overline{\boldsymbol{\eta}}(x) = \boldsymbol{\eta}\left(x;\frac{h}{2}\right) - \frac{\boldsymbol{\eta}(x;h)-\boldsymbol{\eta}\left(x;\frac{h}{2}\right)}{2^p-1} = \frac{2^p\boldsymbol{\eta}\left(x;\frac{h}{2}\right)-\boldsymbol{\eta}(x;h)}{2^p-1}.$$

Dies entspricht dem Extrapolationsprinzip im ROMBERG-Verfahren.
Auch die beiden Näherungen $\boldsymbol{\eta}(x;h)$ und $\boldsymbol{\eta}(x;h/2)$ lassen sich verwenden, um eine bessere Schrittweite zu bestimmen. Wegen $\boldsymbol{e}_p(x_0)=\boldsymbol{o}$ gilt

$$\boldsymbol{e}_p(x) \doteq (x-x_0)\boldsymbol{e}_p'(x_0).$$

Damit folgt weiter

$$\boldsymbol{e}(x_0+h;h) \doteq \boldsymbol{e}_p(x_0+h)h^p \doteq (x_0+h-x_0)\boldsymbol{e}_p'(x_0)h^p = \boldsymbol{e}_p'(x_0)h^{p+1}$$

und

$$\begin{aligned}\boldsymbol{e}\left(x_0+h;\frac{h}{2}\right) &\doteq \boldsymbol{e}_p(x_0+h)\left(\frac{h}{2}\right)^p \\ &\doteq (x_0+h-x_0)\boldsymbol{e}_p'(x_0)\left(\frac{h}{2}\right)^p = \boldsymbol{e}_p'(x_0)h\left(\frac{h}{2}\right)^p.\end{aligned}$$

Aus diesen beiden Gleichungen lässt sich $\boldsymbol{e}_p'(x_0)$ näherungsweise berechnen. Man erhält

$$\boldsymbol{e}_p'(x_0) \doteq \frac{2^p}{2^p-1}\frac{\boldsymbol{\eta}(x_0+h;h)-\boldsymbol{\eta}\left(x_0+h;\frac{h}{2}\right)}{h^{p+1}}.$$

Von einer guten Schrittweite $\overline{h}$ wird man fordern, dass der globale Diskretisierungsfehler unterhalb einer vorgegebenen Genauigkeitsschranke ε liegt:

$$\|\boldsymbol{e}(x_0+\overline{h};\overline{h})\| \doteq \|\boldsymbol{e}_p'(x_0)\|\,|\overline{h}|^{p+1} \mathrel{\dot{\leq}} \varepsilon.$$

Mit dem geschätzten $\boldsymbol{e}_p'(x_0)$ liefert das folgende Schrittweitenschätzung

$$|\overline{h}|^{p+1} \mathrel{\dot{\leq}} \frac{2^p-1}{2^p}\frac{\varepsilon}{\left\|\boldsymbol{\eta}(x_0+h;h)-\boldsymbol{\eta}\left(x_0+h;\frac{h}{2}\right)\right\|}|h|^{p+1}$$

oder

$$|\bar{h}| \stackrel{\cdot}{\leq} |h| \sqrt[p+1]{\frac{2^p-1}{2^p}\frac{\varepsilon}{\left\|\boldsymbol{\eta}(x_0+h;h)-\boldsymbol{\eta}\left(x_0+h;\frac{h}{2}\right)\right\|}}.$$

Damit ergibt sich der folgende Algorithmus:

6.9. Schrittweitensteuerung bei ESV:
Es sei ein ESV mit der Konsistenzordnung p zum Lösen des AWP

$$\boldsymbol{y}' = \boldsymbol{f}(x,\boldsymbol{y}), \quad \boldsymbol{y}(x_0) = \boldsymbol{y}_0$$

gegeben.

S0 *Wähle Grundschrittweite H und eine zu erreichende Genauigkeit ε.*

S1 *Berechne mit dem gegebenen ESV die beiden Näherungen*

$$\boldsymbol{\eta}(x_0+H;H), \quad \boldsymbol{\eta}(x_0+H;H/2).$$

S2 *Berechne*

$$\frac{H}{h} = \sqrt[p+1]{\frac{2^p}{2^p-1}\frac{\left\|\boldsymbol{\eta}(x_0+H;H)-\boldsymbol{\eta}\left(x_0+H;\frac{H}{2}\right)\right\|}{\varepsilon}}.$$

S3 *Ist $H/h \gg 2$, so setze $H = 2h$ und gehe zu Schritt* **S1**.

S4 *Setze*

$$\begin{aligned} x_0 &= x_0 + H, \\ \boldsymbol{y}_0 &= \boldsymbol{\eta}\left(x_0+h;\frac{H}{2}\right), \\ H &= 2h \end{aligned}$$

und gehe zu Schritt **S1**.

Bemerkungen: (i) In der Praxis wird man die Bedingung $H/h \gg 2$ in Schritt **S3** durch $H/h \geq 3..5$ ersetzen.
(ii) ε sollte man nicht kleiner als Keps wählen, wobei K eine obere Schranke für den Betrag der Lösung im betrachteten Bereich ist. Also

$$K \approx \max\{\, \|\boldsymbol{y}(x)\| \mid x \in [x_0, x_0+H] \,\}.$$

Runge-Kutta-Fehlberg-Verfahren

Bei dem oben angegebenen Verfahren **6.9** zur Schrittweitensteuerung benötigt man pro Schritt mindestens drei Auswertungen der Iterationsfunkton Φ. Wir werden sehen, dass man auch mit geringerem Aufwand entsprechende Schätzungen für den globalen Diskretisierungsfehler und damit eine bessere Schrittweite erhält.
Wir betrachten zwei ESV. Diese seien durch die Iterationsfunktionen Φ_1 und Φ_2 gegeben. Die Konsistenzordnung des ersten Verfahrens sei p und die des zweiten Verfahrens sei $p+1$. Für die lokalen Diskretisierungsfehler gilt dann

$$\begin{aligned}\boldsymbol{\tau}_1(x,\boldsymbol{y};h) &= \Delta(x,\boldsymbol{y};h)-\boldsymbol{\Phi}_1(x,\boldsymbol{y};h)=\boldsymbol{C}_1(x)h^p+\boldsymbol{O}(h^{p+1}),\\ \boldsymbol{\tau}_2(x,\boldsymbol{y};h) &= \Delta(x,\boldsymbol{y};h)-\boldsymbol{\Phi}_2(x,\boldsymbol{y};h)=\boldsymbol{C}_2(x)h^{p+1}+\boldsymbol{O}(h^{p+2}).\end{aligned}$$

Ausgehend von einer Näherung $\boldsymbol{\eta}(x)$ berechnen wir mit beiden Verfahren Näherungen an der Stelle $x+H$:

$$\begin{aligned}\boldsymbol{\eta}_1(x+H;H) &= \boldsymbol{\eta}(x)+H\boldsymbol{\Phi}_1(x,\boldsymbol{\eta}(x);H),\\ \boldsymbol{\eta}_2(x+H;H) &= \boldsymbol{\eta}(x)+H\boldsymbol{\Phi}_2(x,\boldsymbol{\eta}(x);H).\end{aligned}$$

Für die Differenz der beiden Näherungen gilt

$$\begin{aligned}\boldsymbol{\eta}_1(x+H;H)-\boldsymbol{\eta}_2(x+H;H) &= H\left(\boldsymbol{\Phi}_1(x,\boldsymbol{\eta}(x);H)-\boldsymbol{\Phi}_2(x,\boldsymbol{\eta}(x);H)\right)\\ &= \boldsymbol{C}_1(x)H^{p+1}+\boldsymbol{O}(H^{p+2}).\end{aligned}$$

In erster Näherung gilt damit

$$\boldsymbol{C}_1(x)\doteq\frac{\boldsymbol{\eta}_1(x+H;H)-\boldsymbol{\eta}_2(x+H;H)}{H^{p+1}}.$$

Von einem erfolgreichen Schritt mit der Schrittweite h werden wir fordern, dass

$$\varepsilon\geqq\|\boldsymbol{\eta}_1(x+h;h)-\boldsymbol{\eta}_2(x+h;h)\|\doteq\|\boldsymbol{C}_1(x)\|\,|h|^{p+1}$$

gilt. Mit der obigen Schätzung für $\boldsymbol{C}_1(x)$ liefert das die Schrittweitenschätzung

$$h\approx H\sqrt[p+1]{\frac{\varepsilon}{\|\boldsymbol{\eta}_1(x+H;H)-\boldsymbol{\eta}_2(x+H;H)\|}}.$$

Es ergibt sich der folgende Algorithmus:

6.10. RUNGE-KUTTA-FEHLBERG**-Verfahren:**
Es seien zwei ESV mit den Konsistenzordnungen p und $p+1$ zum Lösen des AWP

$$\boldsymbol{y}'=\boldsymbol{f}(x,\boldsymbol{y}),\quad \boldsymbol{y}(x_0)=\boldsymbol{y}_0$$

gegeben.

S0 *Wähle Grundschrittweite H und eine zu erreichende Genauigkeit ε.*

S1 *Berechne mit den gegebenen ESV die Näherungen*

$$\boldsymbol{\eta}_1(x_0+H;H), \quad \boldsymbol{\eta}_2(x_0+H;H).$$

S2 *Berechne*

$$\frac{H}{h} = \sqrt[p+1]{\frac{\|\boldsymbol{\eta}_1(x_0+H;H)-\boldsymbol{\eta}_2(x_0+H;H)\|}{\varepsilon}}.$$

S3 *Ist $H/h \gg 2$, so setze $H = 2h$ und gehe zu Schritt* **S1**.

S4 *Setze*

$$\begin{aligned} x_0 &= x_0+H, \\ \boldsymbol{y}_0 &= \boldsymbol{\eta}_2(x_0+H;H), \\ H &= 2h \end{aligned}$$

und gehe zu Schritt **S1**.

Bemerkungen: (i) Auf den ersten Blick scheint der Aufwand beim Algorithmus **6.10** nicht geringer als beim Algorithmus **6.9** zu sein. Rechnet man beim Algorithmus **6.9** k Auswertungen der Funktion $\boldsymbol{f}$ für einen Iterationsschritt, so benötigt dieser Algorithmus mindestens $3k$ Auswertungen der Funktion $\boldsymbol{f}$ pro Schritt. Rechnet man beim Algorithmus **6.10** k Auswertungen der Funktion $\boldsymbol{f}$ für das erste Verfahren und $k+1$ Auswertungen der Funktion $\boldsymbol{f}$ für das zweite Verfahren, so benötigt dieser Algorithmus mindestens $2k+1$ Auswertungen der Funktion $\boldsymbol{f}$ pro Schritt. Es lassen sich aber ESV der Ordnung $p+1$ konstruieren, bei denen ein ESV der Ordnung p als Zwischenergebnis abfällt. So zum Beispiel

$$\begin{array}{r|cccc} \frac{1}{4} & \frac{1}{4} & & & \\ \frac{27}{40} & -\frac{189}{800} & \frac{729}{800} & & \\ 1 & \frac{214}{891} & \frac{27}{891} & \frac{650}{891} & \\ \hline p=2 & \frac{214}{891} & \frac{27}{891} & \frac{650}{891} & \\ p=3 & \frac{533}{2106} & 0 & \frac{1600}{2106} & -\frac{27}{2106} \end{array}.$$

Wir berechnen

$$\begin{aligned}
\boldsymbol{k}_1 &= \boldsymbol{f}(x,\boldsymbol{y}),\\
\boldsymbol{k}_2 &= \boldsymbol{f}\left(x+\frac{1}{4}H,\boldsymbol{y}+\frac{1}{4}H\boldsymbol{k}_1\right),\\
\boldsymbol{k}_3 &= \boldsymbol{f}\left(x+\frac{27}{40}H,\boldsymbol{y}-\frac{189}{800}H\boldsymbol{k}_1+\frac{79}{800}H\boldsymbol{k}_2\right),\\
\boldsymbol{\eta}_1 &= \boldsymbol{y}+H\left(\frac{214}{891}\boldsymbol{k}_1+\frac{27}{891}\boldsymbol{k}_2+\frac{650}{891}\boldsymbol{k}_3\right),\\
\boldsymbol{k}_4 &= \boldsymbol{f}(x+H\boldsymbol{\eta}_1),\\
\boldsymbol{\eta}_2 &= \boldsymbol{y}+H\left(\frac{533}{2106}\boldsymbol{k}_1+\frac{1600}{2106}\boldsymbol{k}_3+\frac{27}{2106}\boldsymbol{k}_4\right).
\end{aligned}$$

Der Aufwand des gesamten Verfahrens wird durch das Verfahren mit der höheren Konvergenzordnung bestimmt. Beim Vergleich der Algorithmen **6.9** und **6.10** würde man damit ein Verhältnis von $3k$ zu $k+1$ Funktionsauswertungen pro Schritt erhalten. Damit sind die RUNGE-KUTTA-FEHLBERG-Verfahren bedeutend effektiver.
(ii) Man beachte die beiden Bemerkungen zum Algorithmus **6.9**.

Anwendung der Schrittweitensteuerung auf die numerische Integration

Die Aufgabe der numerischen Integration ist auch als AWP formulierbar. Das Berechnen von

$$I(f)=\int_a^b f(x)\,dx$$

ist dem Problem des Berechnens des Wertes $y(b)$ der Lösung des AWP

$$y'(x)=f(x),\quad y(a)=0$$

äquivalent. Ist nun durch

$$Q_n(f)=(b-a)\sum_{i=0}^{n} w_i f(x_i),\quad x_i=a+\vartheta_i(b-a)$$

eine Quadraturformel zum näherungsweisen Berechnen von $I(f)$ gegeben, so entspricht dem das ESV

$$\left.\begin{aligned}
\eta_0 &= y_0,\\
\eta_{k+1} &= \eta_k+h\Phi(x_k,\eta_k;h),\\
x_{k+1} &= x_k+h
\end{aligned}\right\},\quad k=0,1,\ldots$$

mit

$$\mathbf{\Phi}(x_k, \eta_k; h) = \sum_{i=0}^{n} w_i f(x_k + \vartheta_i h).$$

Hat die Quadraturformel den Exaktheitsgrad q, so gilt

$$\int_0^1 x^l \, dx = Q_n(x^l) = \sum_{i=0}^{n} w_i \vartheta_i^l$$

für $l = 0, \ldots, q$. Daraus ergibt sich

$$\sum_{i=0}^{n} w_i \vartheta_i^l = \frac{1}{l+1}, \quad l = 0, \ldots, q.$$

Die ersten q Glieder der TAYLOR-Entwicklung

$$\mathbf{\Phi}(x, y; h) = \varphi_0(x, y) + \varphi_1(x, y)h + \varphi_2(x, y)h^2 + \ldots$$

von $\mathbf{\Phi}$ lauten dann

$$\varphi_l(x, y)h^l = \frac{h^l}{l!} \frac{d^l}{dh^l} \mathbf{\Phi}(x, y; h)|_{h=0} = \frac{h^l}{l!} \sum_{i=0}^{n} w_i f^{(l)}(x) \vartheta_i^l = \frac{h^l}{(l+1)!} f^{(l)}(x).$$

Damit ergibt sich für den lokalen Diskretisierungsfehler

$$\tau(x, y; h) = O(h^{q+1}).$$

Das ESV, das der Quadraturformel $Q_n(f)$ entspricht, hat daher die Konsistenzordnung $p = q + 1$. Damit ist das erste Verfahren zur Schrittweitensteuerung von ESV (Algorithmus **6.9**) unmittelbar auf die numerische Integration anwendbar. Wir erhalten den folgenden Algorithmus.

6.11. Schrittweitensteuerung für Quadraturverfahren:
Es sei durch

$$Q_n(f) = (b - a) \sum_{i=0}^{n} w_i f(a + \vartheta_i (b - a))$$

ein Quadraturverfahren mit dem Exaktheitsgrad q zum näherungsweisen Berechnen von

$$I(f) = \int_a^b f(x) \, dx$$

gegeben.

S0 *Wähle Grundschrittweite $H \leqq b-a$ und eine zu erreichende Genauigkeit ε. Setze $x_0 = a$, $\eta_0 = 0$ und $k = 0$.*

S1 *Falls $x_k = b$ so setze $I(f) = \eta_k$. STOPP*

S2 *Berechne die Näherungen*

$$\overline{\eta}_{k+1} = \eta_k + H \sum_{i=0}^{n} w_i f(a + \vartheta_i H)$$

und

$$\hat{\eta}_{k+1} = \eta_k + \frac{H}{2} \sum_{i=0}^{n} w_i f\left(a + \vartheta_i \frac{H}{2}\right) + \frac{H}{2} \sum_{i=0}^{n} w_i f\left(a + \frac{H}{2} + \vartheta_i \frac{H}{2}\right).$$

S3 *Berechne*

$$\frac{H}{h} = \sqrt[q+2]{\frac{2^{q+1}}{2^{q+1}-1} \frac{\|\overline{\eta}_{k+1} - \hat{\eta}_{k+1}\|}{\varepsilon}}.$$

S4 *Ist $H/h \gg 2$, so setze $H = 2h$ und gehe zu Schritt* **S2**.

S5 *Setze*

$$\begin{aligned} x_{k+1} &= x_k + H, \\ \eta_{k+1} &= \hat{\eta}_{k+1}, \\ H &= \min\{2h, b - x_{k+1}\}, \\ k &= k+1 \end{aligned}$$

und gehe zu Schritt **S1**.

6.2.6. Steife Differentialgleichungen und implizite Verfahren

Viele AWP aus der Praxis bereiten beim Lösen eigenartige Schwierigkeiten. Betrachten wir dazu als Beispiel das AWP

$$y' = \mu y, \quad y(0) = 1$$

mit der exakten Lösung $y(x) = e^{\mu x}$. Ist $\mu < 0$, so strebt die Lösung für $x \to \infty$ gegen Null. Für hinreichend große x ändert sich die Lösung nur noch unwesentlich. Man sollte annehmen, dass man in diesem Bereich mit großen Schrittweiten arbeiten könnte. Verwendet man nun das EULERsche Polygonzugverfahren, erhält man die Näherungen

$$\eta_{k+1} = \eta_k + \mu h \eta_k = (1 + \mu h)\eta_k,$$

folglich

$$\eta_k = (1+\mu h)^k \eta_0 = (1+\mu h)^k.$$

Die η_k konvergieren offensichtlich nur für $|1+\mu h| < 1$ gegen Null. Positive Näherungslösungen erhält man nur für $1+\mu h > 0$. Damit ergibt sich für „vernünftige“ Schrittweiten die Forderung

$$-1 < \mu h < 0,$$

und daraus wegen $\mu < 0$

$$0 < h < -\frac{1}{\mu}.$$

Je größer $|\mu|$, also je schneller die exakte Lösung gegen Null konvergiert, desto kleinere Schrittweiten muss man wählen, um vernünftige Näherungslösungen zu erhalten.
Auch für Verfahren höherer Ordnung ergibt sich die gleiche Situation. Das HEUN-Verfahren liefert für das obige AWP die Näherungslösungen

$$\eta_k = \left(1+\mu h+\frac{\mu^2 h^2}{2}\right)^k.$$

Damit wieder $0 < \eta_k$ und $\eta_k \to 0$ für $k \to \infty$ gilt, muss die Schrittweite h die Ungleichungen

$$0 < 1+\mu h+\frac{\mu^2 h^2}{2} < 1$$

erfüllen. Daraus ergibt sich

$$-1 < (1+\mu h)^2 < 1$$

und weiter

$$|1+\mu h| < 1,$$

also

$$h < -\frac{2}{\mu}.$$

Das ist nicht wesentlich besser als beim EULER-Verfahren. Ein Verfahren, das besser zum Lösen des obigen AWP geeignet ist, erhalten wir folgendermaßen. Wir integrieren die Differentialgleichung von x bis $x+h$ und erhalten

$$\int_x^{x+h} \boldsymbol{y}'(t)\,dt = \int_x^{x+h} \boldsymbol{f}(t,\boldsymbol{y}(t)\,dt,$$

und daraus

$$\boldsymbol{y}(x+h) = \boldsymbol{y}(x) + \int_x^{x+h} \boldsymbol{f}(t,\boldsymbol{y}(t))\,dt.$$

Wendet man die Rechteckregel mit der linken Intervallgrenze zur Approximation des Integrals an, erhält man das bekannte EULER-Verfahren

$$\boldsymbol{\eta}(x+h) = \boldsymbol{\eta}(x) + h\boldsymbol{f}(x,\boldsymbol{\eta}(x)).$$

Wendet man jedoch die Rechteckregel mit der rechten Intervallgrenze an, erhält man

$$\boldsymbol{\eta}(x+h) = \boldsymbol{\eta}(x) + h\boldsymbol{f}(x+h,\boldsymbol{\eta}(x+h)).$$

Hier ergibt sich die neue Näherung $\boldsymbol{\eta}(x+h)$ aus einem im allgemeinen nichtlinearen Gleichungssystem. Das Verfahren wird als **implizites** EULER**-Verfahren** bezeichnet. Für obiges eindimensionales AWP ($\boldsymbol{f}(x,\boldsymbol{y}) = \mu y$) lässt sich das Gleichungssystem nach $\eta(x+h)$ auflösen. Es ergibt sich

$$\eta(x+h) = \frac{1}{1-\mu h}\eta(x).$$

Damit hat die k-te Näherung die Darstellung

$$\eta_k = \eta(x_k) = \frac{1}{(1-\mu h)^k}.$$

Die Forderungen

$$\lim_{k\to\infty} \eta_k = 0, \quad \eta_k > 0$$

liefern dann im Falle $\mu < 0$

$$0 < \frac{1}{1-\mu h} < 1,$$

folglich

$$h > 0.$$

Das Verfahren liefert für beliebige positive Schrittweiten vernünftige Lösungen.
Wir wollen nun Differentialgleichungen mit diesen Schwierigkeiten genauer charakterisieren. Das lineare Differentialgleichungssystem

$$\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y}, \quad \boldsymbol{A} \in \mathbb{R}^{d\times d}$$

heißt **steif**, falls alle Eigenwerte λ_i der Matrix $\boldsymbol{A}$ einen negativen Realteil haben und zusätzlich

$$q = \frac{\max\{\, |\Re(\lambda_i)| \mid i = 1,\ldots,d\,\}}{\min\{\, |\Re(\lambda_i)| \mid i = 1,\ldots,d\,\}} \gg 1$$

gilt. Für $q \approx 10$ heißt das Differentialgleichungssystem **schwach steif**, für $q \geqq 10$ **steif**.
Bemerkung: Der Begriff der Steifheit ist auf beliebige Differentialgleichungen übertragbar. Die Differentialgleichung

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y})$$

wird durch die Einführung der Funktionen $\overline{y}(x) = x$ und

$$\tilde{\boldsymbol{y}}(x) = \begin{pmatrix} \boldsymbol{y}(x) \\ \overline{y}(x) \end{pmatrix}$$

auf die Gestalt

$$\tilde{\boldsymbol{y}}' = \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{y}}) = \begin{pmatrix} \boldsymbol{f}(x, \boldsymbol{y}) \\ 1 \end{pmatrix}$$

gebracht. Durch Linearisierung in der Nähe eines Punktes $\tilde{\boldsymbol{y}}_0$ erhält man dann eine Differentialgleichung der Form $\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y}$. Damit lässt sich für eine beliebige Differentialgleichung lokal der Steifheitsbegriff erklären.
Wendet man nun ein ESV zum Lösen eines AWP der Form

$$\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y}, \quad \boldsymbol{y}(x_0) = \boldsymbol{y}_0,$$

an, lässt sich für die Folge $\{\boldsymbol{\eta}_k\}_{k\in\mathbb{N}}$ der Näherungslösungen meist eine Rekursionsformel

$$\boldsymbol{\eta}_{k+1} = g(h\boldsymbol{A})\boldsymbol{\eta}_k$$

mit einer rationalen Funktion g angeben, die nur vom betrachteten Verfahren abhängt. Man erhält zum Beispiel:

- EULER-Verfahren:

$$\boldsymbol{\eta}_{k+1} = (\boldsymbol{I} + h\boldsymbol{A})\boldsymbol{\eta}_k, \quad g(z) = 1 + z,$$

- implizites EULER-Verfahren:

$$\boldsymbol{\eta}_{k+1} = (\boldsymbol{I} - h\boldsymbol{A})^{-1}\boldsymbol{\eta}_k, \quad g(z) = \frac{1}{1-z},$$

- HEUN-Verfahren:

$$\boldsymbol{\eta}_{k+1} = \left(\boldsymbol{I} + h\boldsymbol{A} + \frac{1}{2}(h\boldsymbol{A})^2\right)\boldsymbol{\eta}_k, \quad g(z) = 1 + z + \frac{z^2}{2},$$

- RUNGE-KUTTA-Verfahren:

$$\boldsymbol{\eta}_{k+1} = \left(\boldsymbol{I} + h\boldsymbol{A} + \frac{1}{2}(h\boldsymbol{A})^2 + \frac{1}{6}(h\boldsymbol{A})^3 + \frac{1}{24}(h\boldsymbol{A})^4\right)\boldsymbol{\eta}_k,$$
$$g(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24}.$$

Genügen die Näherungslösungen eines ESV für das angegebene AWP einer derartigen Rekursionsformel, so gilt

$$\boldsymbol{\eta}_k = [g(h\boldsymbol{A})]^k \boldsymbol{y}_0$$

bzw.

$$\boldsymbol{\eta}(x;h) = [g(h\boldsymbol{A})]^{\frac{x-x_0}{h}} \boldsymbol{y}_0.$$

Falls die Eigenwerte der Matrix $g(h\boldsymbol{A})$ betragsmäßig größer als 1 sind, wachsen die Näherungen $\boldsymbol{\eta}_k$ unbeschränkt. Näherungslösungen, die für große x gegen Null konvergieren, ergeben sich nur für $|\lambda(g(h\boldsymbol{A}))| < 1$. Ist λ Eigenwert von $\boldsymbol{A}$, so ist $g(h\lambda)$ Eigenwert von $g(h\boldsymbol{A})$. Damit spielen gerade die Argumente z von g eine besondere Rolle, für die $|g(z)| < 1$ gilt;. Das lineare AWP

$$\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y}, \quad \boldsymbol{y}(x_0) = \boldsymbol{y}_0$$

werde mit einem ESV gelöst. Für die Näherungslösungen existiere bei konstanter Schrittweite eine Rekursionsbeziehung der Form

$$\left.\begin{array}{rcl} \boldsymbol{\eta}_0 & = & \boldsymbol{y}_0, \\ \boldsymbol{\eta}_{k+1} & = & g(h\boldsymbol{A})\boldsymbol{\eta}_k, \\ x_{k+1} & = & x_k + h \end{array}\right\}, \quad k = 0, 1, \ldots$$

Die Menge

$$M = \{\, z \in \mathbb{C} \mid |g(z)| < 1 \,\}$$

heißt dann **(absolutes) Stabilitätsgebiet** des ESV.
Steife Differentialgleichungen sind gerade dadurch gekennzeichnet, dass alle Eigenwerte der Matrix $\boldsymbol{A}$ einen negativen Realteil haben. Ein ESV wird darum genau dann für jene Schrittweiten $h > 0$ vernünftige Näherungslösungen einer steifen Differentialgleichung liefern, für die $\lambda h \in M$ für alle Eigenwerte λ von $\boldsymbol{A}$ gilt. Ist $h > 0$, so folgt aus

$$\lambda \in \mathbb{C}_- = \{\, z \in \mathbb{C} \mid \Re(z) < 0 \,\}$$

$h\lambda \in \mathbb{C}_-$. Ein ESV ist darum um so besser zur Integration einer steifen Differentialgleichung geeignet, je größer der Durchschnitt zwischen absolutem Stabilitätsgebiet M und der linken komplexen Halbebene $\mathbb{C}_-$ ist. Ein ESV zum Lösen eines AWP heißt **absolut stabil** bzw. **A-stabil**, falls $\mathbb{C}_- \subset M$ gilt. Das implizite EULER-Verfahren ist damit absolut stabil. Weitere absolut stabile Verfahren höherer Ordnung gewinnt man durch ähnliche Ansätze wie bei den RUNGE-KUTTA-Verfahren:

$$\boldsymbol{\Phi}(x, \boldsymbol{y}; h) = \sum_{k=0}^{n} \alpha_k \boldsymbol{K}_k(x, \boldsymbol{y}; h)$$

mit

$$\boldsymbol{K}_k(x, \boldsymbol{y}; h) = \boldsymbol{f}\left(x + \vartheta_k h, \boldsymbol{y} + h \sum_{l=0}^{n} \beta_{kl} \boldsymbol{K}_l(x, \boldsymbol{y}; h) \right), \quad k = 0, \ldots, n.$$

Die Größen $\boldsymbol{K}_k$ ergeben sich hier als Lösungen von i.a. nichtlinearen Gleichungssystemen. Für den Fall $n = 1$ ergibt sich zum Beispiel

$$\begin{aligned} \boldsymbol{\Phi}(x, \boldsymbol{y}; h) &= \alpha_0 \boldsymbol{K}_0 + \alpha_1 \boldsymbol{K}_1, \\ \boldsymbol{K}_0 &= \boldsymbol{f}(x + \vartheta_0 h, \boldsymbol{y} + \beta_{00} \boldsymbol{K}_0 h + \beta_{01} \boldsymbol{K}_1 h), \\ \boldsymbol{K}_1 &= \boldsymbol{f}(x + \vartheta_1 h, \boldsymbol{y} + \beta_{10} \boldsymbol{K}_0 h + \beta_{11} \boldsymbol{K}_1 h). \end{aligned}$$

Die Konstanten $\alpha_0, \ldots, \alpha_n, \vartheta_0, \ldots, \vartheta_n$ und $\beta_{00}, \ldots, \beta_{nn}$ werden wieder so bestimmt, dass man Verfahren möglichst hoher Ordnung erhält. Wir wollen die Koeffizienten einiger Verfahren nach folgendem Schema angeben:

ϑ_0	β_{00}	β_{01}	β_{02}	β_{03}
ϑ_1	β_{10}	β_{11}	β_{12}	β_{13}
ϑ_2	β_{20}	β_{21}	β_{22}	β_{23}
ϑ_3	β_{30}	β_{31}	β_{32}	β_{33}
	α_0	α_1	α_2	α_3

.

Speziell sind die expliziten RUNGE-KUTTA-Formeln in diesem Schema enthalten ($\beta_{ij} = 0$ für $j \geqq i$). Gilt $\beta_{ij} = 0$ für $j > i$, so heißen die Verfahren **diagonalimplizit**.

- GAUSS-LEGENDRE

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array} \qquad p = 2$$

$$\begin{array}{c|cc} \frac{1}{2}-\alpha & \frac{1}{4} & \frac{1}{4}-\alpha \\ \frac{1}{2}+\alpha & \frac{1}{4}+\alpha & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \quad \alpha = \frac{1}{2\sqrt{3}} \quad p = 4$$

- RADAU IA

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \qquad p = 1$$

$$\begin{array}{c|cc} 0 & \frac{1}{4} & -\frac{1}{4} \\ \frac{2}{3} & \frac{1}{4} & \frac{5}{12} \\ \hline & -\frac{1}{4} & \frac{3}{4} \end{array} \qquad p = 3$$

- RADAU IIA

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \qquad p = 1$$

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array} \qquad p = 3$$

- LOBATTO IIIA

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad p = 2$$

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{5}{24} & \frac{1}{3} & -\frac{1}{24} \\ 1 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array} \qquad p = 4$$

Schrittweitensteuerungen sind analog zu expliziten Verfahren möglich. Es existieren auch Verfahren, die ähnlich wie die RUNGE-KUTTA-FEHLBERG-Verfahren funktionieren.

6.3. Mehrschrittverfahren

6.3.1. Prediktor-Korrektor-Verfahren

ESV waren dadurch charakterisiert, dass zum Berechnen einer Näherung $\boldsymbol{\eta}_{k+1}$ nur die Kenntnis der vorigen Näherung $\boldsymbol{\eta}_k$ benötigt wurde. Nun liegt es nahe, Verfahren zu konstruieren, die zum Berechnen der Näherung $\boldsymbol{\eta}_{k+1}$ mehrere vorhergehende Näherungen $\boldsymbol{\eta}_k, \boldsymbol{\eta}_{k-1}, \ldots, \boldsymbol{\eta}_{k-r+1}$ verwenden. Verfahren dieser Art heißen Mehrschrittverfahren.[4] Wir wollen nun einfache MSV konstruieren. Es sei

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{y}(x_0) = \boldsymbol{y}_0$$

das zu lösende AWP. Die Funktion $\boldsymbol{f}$ erfülle die Voraussetzungen aus Satz 6.1. Integrieren wir die Differentialgleichung von $\underline{x}$ bis $\overline{x}$, so erhalten wir

$$\int_{\underline{x}}^{\overline{x}} \boldsymbol{y}'(t)\,dt = \int_{\underline{x}}^{\overline{x}} \boldsymbol{f}(t, \boldsymbol{y}(t))\,dt$$

und damit

$$\boldsymbol{y}(\overline{x}) - \boldsymbol{y}(\underline{x}) = \int_{\underline{x}}^{\overline{x}} \boldsymbol{f}(t, \boldsymbol{y}(t))\,dt.$$

Kennt man einen Näherungswert für $\boldsymbol{y}(\underline{x})$, lässt sich eine Näherung für $\boldsymbol{y}(\overline{x})$ gewinnen, indem man das Integral auf der rechten Seite der obigen Gleichung in geeigneter Weise approximiert. Hier treten zwei Schwierigkeiten auf. Einerseits kennt man die Abhängigkeit des Integranden $\boldsymbol{f}$ von t nicht, da die Funktion $\boldsymbol{y}(t)$ nicht bekannt ist. Andererseits ließe sich das Integral auch bei bekannter Abhängigkeit der Funktion $\boldsymbol{f}$ von t in den meisten Fällen nur näherungsweise berechnen. Betrachten wir das Problem der Approximation des bestimmten Integrals bei bekanntem Integranden. Zum Lösen dieses Problems gehen wir ähnlich wie bei der Herleitung der NEWTON-COTES-Formeln vor. Wir ersetzen den Integranden durch ein geeignetes Interpolationspolynom und integrieren dieses. Zur Konstruktion des Interpolationspolynoms wählen wir äquidistante Stützstellen

$$x_i = x_0 + ih, \quad i = 0, 1, \ldots$$

[4] Wir werden für Mehrschrittverfahren die Abkürzung MSV verwenden.

Weiterhin sei

$$\bar{x} = x_{p+k}, \quad \underline{x} = x_{p-j}.$$

Das gesuchte Interpolationspolynom $\boldsymbol{P}_q(t)$ sei durch

1. Grad$\boldsymbol{P}_q \leq q$ und
2. $\boldsymbol{P}_q(x_i) = \boldsymbol{f}(x_i, \boldsymbol{y}(x_i))$ für $i = p, p-1, \ldots, p-q$

festgelegt. Man beachte, dass $\boldsymbol{P}_q$ im Falle $\boldsymbol{y} : \mathbb{R} \longrightarrow \mathbb{R}^d$ selbst eine Vektorfunktion ist, deren Komponenten Polynome vom Höchstgrad q sind. Mit der LAGRANGEschen Interpolationsformel ist $\boldsymbol{P}_q$ sofort angebbar. Wir erhalten

$$\boldsymbol{P}_q(t) = \sum_{i=0}^{q} \boldsymbol{f}(x_{p-i}, \boldsymbol{y}(x_{p-i})) L_i(t)$$

mit

$$L_i(t) = \prod_{\substack{l=0 \\ l \neq i}}^{q} \frac{t - x_{p-l}}{x_{p-i} - x_{p-l}}, \quad i = 0, 1, \ldots, q.$$

Damit gilt

$$\begin{aligned} \boldsymbol{y}_{p+k} &\approx \boldsymbol{y}_{p-j} + \sum_{i=0}^{q} \boldsymbol{f}(x_{p-i}, \boldsymbol{y}(x_{p-i})) \int_{x_{p-j}}^{x_{p+k}} L_i(t)\, dt \\ &= \boldsymbol{y}_{p-j} + h \sum_{i=0}^{q} \beta_{qi} \boldsymbol{f}(x_{p-i}, \boldsymbol{y}(x_{p-i})) \end{aligned}$$

mit

$$\begin{aligned} \beta_{qi} &= \frac{1}{h} \int_{x_{p-j}}^{x_{p+k}} L_i(t)\, dt \\ &= \frac{1}{h} \int_{x_{p-j}}^{x_{p+k}} \prod_{\substack{l=0 \\ l \neq i}}^{q} \frac{t - x_{p-l}}{x_{p-i} - x_{p-l}}\, dt. \end{aligned}$$

Durch die Substitution $t = x_p + sh = x_0 + (p+s)h$ erhält man daraus

$$\beta_{qi} = \int_{-j}^{k} \prod_{\substack{l=0 \\ l \neq i}}^{q} \frac{s+l}{-j+l}\, ds.$$

Je nach Wahl von k, j und q ergeben sich verschiedene Verfahren der Form

$$\boldsymbol{\eta}_{p+k} = \boldsymbol{\eta}_{p-j} + h \sum_{i=0}^{q} \beta_{qi} \boldsymbol{f}(x_{p-i}, \boldsymbol{\eta}_{p-i}).$$

Wir wollen die wichtigsten vier Verfahrensklassen angeben.

Adams-Bashforth-Verfahren

Wir setzen $k = 1$, $j = 0$ und $q = 0, 1, \ldots$. Es gilt dann

$$\beta_{qi} = \int_0^1 \prod_{\substack{l=0 \\ l \neq i}}^{q} \frac{s+l}{-j+l}\, ds$$

und

$$\boldsymbol{\eta}_{p+1} = \boldsymbol{\eta}_p + h \left[\beta_{q0} \boldsymbol{f}_p + \beta_{q1} \boldsymbol{f}_{p-1} + \cdots + \beta_{qq} \boldsymbol{f}_{p-q}\right]$$

mit

$$\boldsymbol{f}_i = \boldsymbol{f}(x_i, \boldsymbol{\eta}_i), \quad i = 0, 1, \ldots.$$

q	$s\beta_{qi}$					s	Ordnung
0	1					1	1
1	3	-1				2	2
2	23	-16	5			12	3
3	55	-59	37	-9		24	4
4	1901	-2774	2616	-1274	251	720	5

Tabelle 6.1: ADAMS-BASHFORTH-Verfahren

Adams-Moulton-Verfahren

Wir setzen $k = 0$, $j = 1$ und $q = 0, 1, \ldots$. Es gilt dann

$$\beta_{qi} = \int_{-1}^{0} \prod_{\substack{l=0 \\ l \neq i}}^{q} \frac{s+l}{-j+l}\, ds$$

und

$$\boldsymbol{\eta}_p = \boldsymbol{\eta}_{p-1} + h\left[\beta_{q0}\boldsymbol{f}(x_p, \boldsymbol{\eta}_p) + \beta_{q1}\boldsymbol{f}_{p-1} + \cdots + \beta_{qq}\boldsymbol{f}_{p-q}\right].$$

Ersetzt man p durch $p+1$, so erhält man die übliche Form

$$\boldsymbol{\eta}_{p+1} = \boldsymbol{\eta}_p + h\left[\beta_{q0}\boldsymbol{f}(x_{p+1}, \boldsymbol{\eta}_{p+1}) + \beta_{q1}\boldsymbol{f}_p + \cdots + \beta_{qq}\boldsymbol{f}_{p-q+1}\right].$$

q	$s\beta_{qi}$					s	Ordnung
0	1					1	1
1	1	1				2	2
2	5	8	-1			12	3
3	9	19	-5	1		24	4
4	251	646	-264	106	19	720	5

Tabelle 6.2: ADAMS-MOULTON-Verfahren

Nyström-Verfahren

Wir setzen $k = 1$, $j = 1$ und $q = 0, 1, \ldots$. Es gilt dann

$$\beta_{qi} = \int_{-1}^{1} \prod_{\substack{l=0 \\ l\neq i}}^{q} \frac{s+l}{-j+l}\, ds$$

und

$$\boldsymbol{\eta}_{p+1} = \boldsymbol{\eta}_p + h\left[\beta_{q0}\boldsymbol{f}_p + \beta_{q1}\boldsymbol{f}_{p-1} + \cdots + \beta_{qq}\boldsymbol{f}_{p-q}\right].$$

Milne-Verfahren

Wir setzen $k = 0$, $j = 2$ und $q = 0, 1, \ldots$. Es gilt dann

$$\beta_{qi} = \int_{-2}^{0} \prod_{\substack{l=0 \\ l\neq i}}^{q} \frac{s+l}{-j+l}\, ds$$

q	$s\beta_{qi}$					s	Ordnung
0	2					1	2
1	2	0				1	2
2	7	-2	1			3	3
3	8	-5	4	-1		3	4
4	269	-266	294	-146	29	90	5

Tabelle 6.3: NYSTRÖM-Verfahren

und

$$\boldsymbol{\eta}_p = \boldsymbol{\eta}_{p-2} + h\left[\beta_{q0}\boldsymbol{f}(x_p, \boldsymbol{\eta}_p) + \beta_{q1}\boldsymbol{f}_{p-1} + \cdots + \beta_{qq}\boldsymbol{f}_{p-q}\right].$$

Ersetzen wir wieder p durch $p+1$, so erhalten wir die übliche Form

$$\boldsymbol{\eta}_{p+1} = \boldsymbol{\eta}_{p-1} + h\left[\beta_{q0}\boldsymbol{f}(x_{p+1}, \boldsymbol{\eta}_{p+1}) + \beta_{q1}\boldsymbol{f}_{p} + \cdots + \beta_{qq}\boldsymbol{f}_{p-q+1}\right].$$

Bei den Verfahren von ADAMS-BASHFORTH und NYSTRÖM lässt sich aus den be-

q	$s\beta_{qi}$					s	Ordnung
0	2					1	1
1	0	2				1	2
2	1	4	1			3	4
3	1	4	1	0		3	4
4	29	124	24	4	-1	90	5

Tabelle 6.4: MILNE-Verfahren

kannten Näherungen $\boldsymbol{\eta}_p, \ldots, \boldsymbol{\eta}_{p-q}$ leicht die neue Näherung $\boldsymbol{\eta}_{p+1}$ bestimmen. Dazu ist nur eine Auswertung der Funktion $\boldsymbol{f}$ notwendig, nämlich das Berechnen von $\boldsymbol{f}(x_p, \boldsymbol{\eta}_p)$. Bei den Verfahren von ADAMS-MOULTON und MILNE tritt die neue Näherung $\boldsymbol{\eta}_{p+1}$ jedoch auch auf der rechten Seite der Gleichung auf. Damit ergibt sich $\boldsymbol{\eta}_{p+1}$ als Lösung eines i.a. nichtlinearen Gleichungssystems. Die Verfahren von ADAMS-MOULTON und MILNE sind **implizite** Verfahren im Gegensatz zu den **expliziten** Verfahren von ADAMS-BASHFORTH und NYSTRÖM. Zum Lösen der nichtlinearen Gleichungen liegt es nahe, eine Fixpunktiteration folgender Form anzuwenden:

6.12. Fixpunktiteration für Korrektor-Verfahren:

S0 *Wähle Startwert* $\boldsymbol{\eta}_{p+1}^{(0)}$ *und setze* $\nu = 0$. *Berechne für das*

- ADAMS-MOULTON-*Verfahren*

$$\varrho = \boldsymbol{\eta}_p + h\left[\beta_{q1}\boldsymbol{f}_p + \beta_{q2}\boldsymbol{f}_{p-1} + \cdots + \beta_{qq}\boldsymbol{f}_{p-q}\right]$$

- MILNE-*Verfahren*

$$\varrho = \boldsymbol{\eta}_{p-1} + h\left[\beta_{q1}\boldsymbol{f}_p + \beta_{q2}\boldsymbol{f}_{p-1} + \cdots + \beta_{qq}\boldsymbol{f}_{p-q}\right]$$

S1 *Berechne*

$$\boldsymbol{\eta}_{p+1}^{(\nu+1)} = \varrho + h\beta_{q0}\boldsymbol{f}(x_{p+1}, \boldsymbol{\eta}_{p+1}^{(\nu)}).$$

S2 *Setze* $k = k+1$ *und gehe zu Schritt* **S1**.

Die Iteration konvergiert, falls die Funktion $\boldsymbol{f}$ eine LIPSCHITZ-Bedingung

$$\|\boldsymbol{f}(x, \boldsymbol{y}_1) - \boldsymbol{f}(x, \boldsymbol{y}_2)\| \leqq \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|$$

für alle $x \in [a,b]$ und alle $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{R}^d$ erfüllt, und falls die Schrittweite h hinreichend klein gewählt wird. Eine gute Startnäherung $\boldsymbol{\eta}_{p+1}^{(0)}$ beschafft man sich mit einem expliziten MSV. Darum heißen die expliziten MSV auch **Prediktor-Verfahren** und die impliziten MSV **Korrektor-Verfahren**. Die Kombination eines Prediktor-Verfahrens mit einem Korrektor-Verfahren wird **Prediktor-Korrektor-Verfahren** genannt. Für das Berechnen einer neuen Näherung reichen i. a. ein Prediktor-Schritt und ein oder zwei Korrektor-Schritte.

6.3.2. Konvergenz von Mehrschrittverfahren

Die bisher behandelten MSV und ESV sind von der Form

$$\boldsymbol{\eta}_{j+r} + \alpha_{r-1}\boldsymbol{\eta}_{j+r-1} + \cdots + \alpha_0\boldsymbol{\eta}_j = h\boldsymbol{F}(x_j, \boldsymbol{\eta}_{j+r}, \ldots, \boldsymbol{\eta}_j; h; \boldsymbol{f}).$$

Ein derartiges Verfahren wird auch als r**-Schritt-Verfahren** bezeichnet. Bei den ESV ist $r = 1$, $\alpha_0 = -1$ und $\boldsymbol{F} \equiv \Phi$. Bei den im vorigen Abschnitt behandelten MSV hing die Funktion $\boldsymbol{F}$ linear von der Funktion $\boldsymbol{f}$ ab:

$$\begin{aligned}\boldsymbol{F}(x_j, \boldsymbol{\eta}_{j+r}, \ldots, \boldsymbol{\eta}_j; h; \boldsymbol{f}) = {} & \beta_r \boldsymbol{f}(x_{j+r}, \boldsymbol{\eta}_{j+r}) \\ & + \beta_{r-1}\boldsymbol{f}(x_{j+r-1}, \boldsymbol{\eta}_{j+r-1}) + \cdots + \beta_0 \boldsymbol{f}(x_j, \boldsymbol{\eta}_j).\end{aligned}$$

Man spricht von einem linearen r-Schritt-Verfahren. Im Gegensatz dazu hing die Funktion Φ bei den ESV hochgradig nichtlinear von $\boldsymbol{f}$ ab.

Wir wollen nun die Eigenschaften von allgemeinen MSV untersuchen. Dazu werden wir wieder lokale und globale Diskretisierungsfehler betrachten und Konsistenz und Konvergenz der Verfahren untersuchen.
Es seien $\boldsymbol{f} \in F^1[a,b]$, $\boldsymbol{z}(t)$ die exakte Lösung des AWP

$$\boldsymbol{z}'(t) = \boldsymbol{f}(t, \boldsymbol{z}(t)), \quad \boldsymbol{z}(x) = \boldsymbol{y}$$

und $\boldsymbol{F}$ die Verfahrensfunktion eines zugeordneten r-Schritt-Verfahrens. Dann heißt die Größe

$$\boldsymbol{\tau}(x, \boldsymbol{y}; h) = \\ \frac{1}{h}\left[\boldsymbol{z}(x+rh) + \sum_{i=0}^{r-1} \alpha_i \boldsymbol{z}(x+ih) - h\boldsymbol{F}(x_i, \boldsymbol{z}(x+rh), \ldots, \boldsymbol{z}(x+h), \boldsymbol{z}(x); h; \boldsymbol{f})\right]$$

lokaler Diskretisierungsfehler des entsprechenden MSV an der Stelle $(x, \boldsymbol{y})$.
Der lokale Diskretisierungsfehler gibt folglich an, wie gut die exakte Lösung die Gleichung des MSV erfüllt. Wie bei ESV definieren wir nun auch die Konsistenz von Verfahren. Ein MSV heißt **konsistent**, falls für alle $x \in [a,b]$, alle $\boldsymbol{y} \in \mathbb{R}^d$ und alle $\boldsymbol{f} \in F^1[a,b]$

$$\lim_{h \to 0} \boldsymbol{\tau}(x, \boldsymbol{y}; h) = 0$$

gilt. Es ist natürlich wieder wünschenswert, dass der lokale Diskretisierungsfehler möglichst schnell mit h gegen Null konvergiert. Das führt wieder auf den Begriff der Konsistenzordnung. Ein MSV hat die Konsistenzordnung p, falls für alle $x \in [a,b]$, alle $\boldsymbol{y} \in \mathbb{R}^d$ und alle $\boldsymbol{f} \in F^p[a,b]$

$$\|\boldsymbol{\tau}(x, \boldsymbol{y}; h)\| = O(|h|^p)$$

gilt. Die Konsistenzordnung eines Verfahrens lässt sich wieder bestimmen, indem man für den lokalen Diskretisierungsfehler die TAYLOR-Entwicklung bezüglich h verwendet. Die Ordnung des ersten nichtverschwindenden Gliedes gibt dann die Konsistenzordnung an.

6.13. Beispiel: Wir betrachten ein NYSTRÖM-Verfahren mit $q = 2$:

$$\boldsymbol{\eta}_{p+1} - \boldsymbol{\eta}_{p-1} = h\left[\frac{7}{3}\boldsymbol{f}(x_p, \boldsymbol{\eta}_p) - \frac{2}{3}f(x_{p-1}, \boldsymbol{\eta}_{p-1}) + \frac{1}{3}\boldsymbol{f}(x_{p-2}, \boldsymbol{\eta}_{p-2})\right].$$

Setzt man $p = j+2$, so erhält man

$$\boldsymbol{\eta}_{j+3} - \boldsymbol{\eta}_{j+1} = h\left[\frac{7}{3}\boldsymbol{f}(x_{j+2}, \boldsymbol{\eta}_{j+2}) - \frac{2}{3}\boldsymbol{f}(x_{j+1}, \boldsymbol{\eta}_{j+1}) + \frac{1}{3}\boldsymbol{f}(x_j, \boldsymbol{\eta}_j)\right].$$

Damit gilt

$$\begin{aligned}\boldsymbol{\tau}(x,\boldsymbol{z};h) &= \frac{1}{h}[\boldsymbol{z}(x+3h)-\boldsymbol{z}(x+h)]-\frac{7}{3}\boldsymbol{f}(x+2h,\boldsymbol{z}(x+2h))+\\ &\quad+\frac{2}{3}\boldsymbol{f}(x+h,\boldsymbol{z}(x+h))-\frac{1}{3}\boldsymbol{f}(x,\boldsymbol{z}(x))\\ &= \frac{1}{h}[\boldsymbol{z}(x+3h)-\boldsymbol{z}(x+h)]-\frac{7}{3}\boldsymbol{z}'(x+2h)+\frac{2}{3}\boldsymbol{z}'(x+h)-\frac{1}{3}\boldsymbol{z}'(x).\end{aligned}$$

Wir entwickeln die einzelnen Terme nach Potenzen von h:

$$\begin{aligned}\boldsymbol{z}(x+3h) &= \boldsymbol{z}(x)+3h\boldsymbol{z}'(x)+\frac{9h^2}{2}\boldsymbol{z}''(x)+\frac{9h^3}{2}\boldsymbol{z}'''(x)+\frac{27h^4}{8}\boldsymbol{z}^{(4)}(x)+\boldsymbol{O}(h^5),\\ \boldsymbol{z}(x+h) &= \boldsymbol{z}(x)+h\boldsymbol{z}'(x)+\frac{h^2}{2}\boldsymbol{z}''(x)+\frac{h^3}{6}\boldsymbol{z}'''(x)+\frac{h^4}{24}\boldsymbol{z}^{(4)}(x)+\boldsymbol{O}(h^5),\\ \boldsymbol{z}'(x+2h) &= \boldsymbol{z}'(x)+2h\boldsymbol{z}''(x)+2h^2\boldsymbol{z}'''(x)+\frac{4h^3}{3}\boldsymbol{z}^{(4)}(x)+\boldsymbol{O}(h^4),\\ \boldsymbol{z}'(x+h) &= \boldsymbol{z}'(x)+h\boldsymbol{z}''(x)+\frac{h^2}{2}\boldsymbol{z}'''(x)+\frac{h^3}{6}\boldsymbol{z}^{(4)}(x)+\boldsymbol{O}(h^4).\end{aligned}$$

Setzt man diese Entwicklungen in die obige Darstellung von $\boldsymbol{\tau}$ ein, so ergibt sich

$$\begin{aligned}\boldsymbol{\tau}(x,\boldsymbol{z};h) &= \frac{1}{h}\left[\boldsymbol{z}(x)+3h\boldsymbol{z}'(x)+\frac{9h^2}{2}\boldsymbol{z}''(x)+\frac{9h^3}{2}\boldsymbol{z}'''(x)+\frac{27h^4}{8}\boldsymbol{z}^{(4)}(x)-\right.\\ &\quad\left.-\boldsymbol{z}(x)-h\boldsymbol{z}'(x)-\frac{h^2}{2}\boldsymbol{z}''(x)-\frac{h^3}{6}\boldsymbol{z}'''(x)-\frac{h^4}{24}\boldsymbol{z}^{(4)}(x)+\boldsymbol{O}(h^5)\right]-\\ &\quad-\frac{7}{3}\left[\boldsymbol{z}'(x)+2h\boldsymbol{z}''(x)+2h^2\boldsymbol{z}'''(x)+\frac{4h^3}{3}\boldsymbol{z}^{(4)}(x)+\boldsymbol{O}(h^4)\right]+\\ &\quad+\frac{2}{3}\left[\boldsymbol{z}'(x)+h\boldsymbol{z}''(x)+\frac{h^2}{2}\boldsymbol{z}'''(x)+\frac{h^3}{6}\boldsymbol{z}^{(4)}(x)+\boldsymbol{O}(h^4)\right]-\frac{1}{3}\boldsymbol{z}'(x)\\ &= \frac{1}{h}\left[2h\boldsymbol{z}'(x)+4h^2\boldsymbol{z}''(x)+\frac{13h^3}{3}\boldsymbol{z}'''(x)+\frac{10h^4}{3}\boldsymbol{z}^{(4)}(x)+\boldsymbol{O}(h^5)\right]-\\ &\quad-2\boldsymbol{z}'(x)-4h\boldsymbol{z}''(x)-\frac{13h^2}{3}\boldsymbol{z}'''(x)-3h^3\boldsymbol{z}^{(4)}(x)+\boldsymbol{O}(h^4)\\ &= \frac{h^3}{3}\boldsymbol{z}^{(4)}(x)+\boldsymbol{O}(h^4).\end{aligned}$$

Es handelt sich somit um ein Verfahren der Konsistenzordnung 3. ♡

Nun möchte man natürlich wieder Verfahren möglichst hoher Konsistenzordnung konstruieren. Für lineare MSV erscheint dies einfach zu sein. Der Ansatz

$$\boldsymbol{\eta}_{j+r}+\alpha_{r-1}\boldsymbol{\eta}_{j+r-1}+\cdots+\alpha_0\boldsymbol{\eta}_j = h\left[\beta_r\boldsymbol{f}(x_{j+r},\boldsymbol{\eta}_{j+r})+\cdots+\beta_0\boldsymbol{f}(x_j,\boldsymbol{\eta}_j)\right]$$

führt über die entsprechende Entwicklung des lokalen Diskretisierungsfehlers auf ein lineares Gleichungssystem zur Bestimmung der Koeffizienten $\alpha_{r-1},\ldots,\alpha_0$ und $\beta_r,\ldots,\beta_0$. Bei der Konstruktion von ESV über einen entsprechenden Ansatz lieferte jede Lösung des dazugehörigen nichtlinearen Gleichungssystems ein konvergentes ESV. Bei MSV ist dem nicht so. Das zeigt schon das folgende Beispiel.

6.14. Beispiel: Wir wollen ein explizites lineares MSV mit $r=2$ und maximaler Konsistenzordnung konstruieren. Dazu dient der folgende Ansatz:

$$\boldsymbol{\eta}_{j+2}+\alpha_1\boldsymbol{\eta}_{j+1}+\alpha_0\boldsymbol{\eta}_j = h\left[\beta_1\boldsymbol{f}(x_{j+1},\boldsymbol{\eta}_{j+1})+\beta_0\boldsymbol{f}(x_j,\boldsymbol{\eta}_j)\right].$$

Für den lokalen Diskretisierungsfehler erhalten wir

$$\begin{aligned}\boldsymbol{\tau}(x,\boldsymbol{y};h) &= \frac{1}{h}\left[\boldsymbol{z}(x+2h)+\alpha_1\boldsymbol{z}(x+h)+\alpha_0\boldsymbol{z}(x)\right]\\ &\quad -\left[\beta_1\boldsymbol{f}(x+h,\boldsymbol{z}(x+h))+\beta_0\boldsymbol{f}(x,\boldsymbol{z}(x))\right].\end{aligned}$$

Mit $\boldsymbol{z}'(t)=\boldsymbol{f}(t,z(t))$ und $\boldsymbol{z}(x)=\boldsymbol{y}$ ergibt sich daraus

$$\boldsymbol{\tau}(x,\boldsymbol{y};h)=\frac{1}{h}\left[\boldsymbol{z}(x+2h)+\alpha_1\boldsymbol{z}(x+h)+\alpha_0\boldsymbol{z}(x)\right]-\left[\beta_1\boldsymbol{z}'(x+h)+\beta_0\boldsymbol{z}'(x)\right].$$

Wir entwickeln die einzelnen Terme nach Potenzen von h:

$$\begin{aligned}\boldsymbol{z}(x+2h) &= \boldsymbol{z}(x)+2h\boldsymbol{z}'(x)+\frac{4h^2}{2}\boldsymbol{z}''(x)+\frac{8h^3}{6}\boldsymbol{z}'''(x)+\boldsymbol{O}(h^4),\\ \boldsymbol{z}(x+h) &= \boldsymbol{z}(x)+h\boldsymbol{z}'(x)+\frac{h^2}{2}\boldsymbol{z}''(x)+\frac{h^3}{6}\boldsymbol{z}'''(x)+\boldsymbol{O}(h^4),\\ \boldsymbol{z}'(x+h) &= \boldsymbol{z}'(x)+h\boldsymbol{z}''(x)+\frac{h^2}{2}\boldsymbol{z}'''(x)+\boldsymbol{O}(h^3).\end{aligned}$$

Einsetzen dieser Entwicklungen liefert

$$\begin{aligned}\boldsymbol{\tau}(x,\boldsymbol{y};h) &= \frac{1}{h}\Bigg[(1+\alpha_1+\alpha_0)\boldsymbol{z}(x)+(2+\alpha_1)h\boldsymbol{z}'(x)+(4+\alpha_1)\frac{h^2}{2}\boldsymbol{z}''(x)+\\ &\quad +(8+\alpha_1)\frac{h^3}{6}\boldsymbol{z}'''(x)+\boldsymbol{O}(h^4)\Bigg]-\\ &\quad -\left[(\beta_1+\beta_0)\boldsymbol{z}'(x)+\beta_1 h\boldsymbol{z}''(x)+\beta_1\frac{h^2}{2}\boldsymbol{z}'''(x)+\boldsymbol{O}(h^3)\right]\\ &= (1+\alpha_1+\alpha_0)\frac{1}{h}\boldsymbol{z}(x)+(2+\alpha_1-\beta_1-\beta_0)\boldsymbol{z}'(x)\\ &\quad +(4+\alpha_1-2\beta_1)\frac{h}{2}\boldsymbol{z}''(x)+(8+\alpha_1-3\beta_1)\frac{h^2}{6}\boldsymbol{z}'''(x)+\boldsymbol{O}(h^3).\end{aligned}$$

Um ein Verfahren möglichst hoher Konsistenzordnung zu erhalten, müssen die ersten Terme verschwinden. Es ergibt sich das folgende lineare Gleichungssystem.

$$\begin{array}{rcrcrcl} \alpha_1 & + & \alpha_0 & & & = & -1 \\ \alpha_1 & & - & \beta_1 & - \ \beta_0 & = & -2 \\ \alpha_1 & & - & 2\beta_1 & - & = & -4 \\ \alpha_1 & & - & 3\beta_1 & - & = & -8 \end{array}$$

Dieses besitzt die eindeutige Lösung $\alpha_0 = -5$, $\alpha_1 = 4$, $\beta_0 = 2$ und $\beta_1 = 4$. Wir erhalten folglich das Verfahren

$$\boldsymbol{\eta}_{j+2} + 4\boldsymbol{\eta}_{j+1} - 5\boldsymbol{\eta}_j = h\left[4\boldsymbol{f}(x_{j+1}, \boldsymbol{\eta}_{j+1}) + 2\boldsymbol{f}(x_j, \boldsymbol{\eta}_j)\right].$$

mit der Konsistenzordnung 3. Wenden wir dieses Verfahren auf das AWP

$$y' = -y, \quad y(0) = 1$$

mit der exakten Lösung $y(x) = e^{-x}$ an, so lässt sich die Näherungslösung explizit darstellen. Sie erfüllt die homogenen Differenzengleichung

$$\eta_{j+2} + (4+4h)\eta_{j+1} + (-5+2h)\eta_j = 0.$$

Lösungen von homogenen Differenzengleichungen erhält man über den Ansatz

$$\eta_j = \lambda^j.$$

Einsetzen dieses Ansatzes in die Differenzengleichung liefert

$$\lambda^{j+2} + (4+4h)\lambda^{j+1} + (-5+2h)\lambda^j = 0.$$

Diese Gleichung ist für $\lambda = 0$ trivial erfüllt. Wesentliche Lösungen erhält man aus

$$\lambda^2 + (4+4h)\lambda - 5 + 2h = 0.$$

Das sind

$$\lambda_1(h) = -2 - 2h + \sqrt{1 + \frac{2}{3}h + \frac{4}{9}h^2}$$

und

$$\lambda_2(h) = -2 - 2h - \sqrt{1 + \frac{2}{3}h + \frac{4}{9}h^2}.$$

Die allgemeine Lösung der Differenzengleichung lässt sich als Linearkombination dieser speziellen Lösungen darstellen

$$\eta_j = \mu_1 \lambda_1^j + \mu_2 \lambda_2^j.$$

μ_1 und μ_2 werden aus den Anfangsbedingungen ermittelt. Es gelte

$$\begin{aligned} \eta_0 &= y_0 = 1, \\ \eta_1 &= y_1 = e^{-h} \quad \text{(exakter Wert)}. \end{aligned}$$

Daraus folgt

$$\mu_1 = \frac{e^{-h} - \lambda_2}{\lambda_1 - \lambda_2}, \quad \mu_2 = -\frac{e^{-h} - \lambda_1}{\lambda_1 - \lambda_2}.$$

Entwickelt man nun $\lambda_1(h)$, $\lambda_2(h)$, $\mu_1(h)$ und $\mu_2(h)$ nach Potenzen von h und setzt alles in die Darstellung

$$\eta_j = \mu_1(h)\,[\lambda_1(h)]^j + \mu_2(h)\,[\lambda_2(h)]^j$$

ein, so erhält man für festes x und $h_n = x/n$

$$\begin{aligned} \eta_n &= \eta(x; h_n) \\ &= \left[1 + O\left(\frac{x}{n}\right)\right] \left[1 - \frac{x}{n} + O\left(\left(\frac{x}{n}\right)^2\right)\right]^n - \\ &\quad -\frac{1}{216}\left(\frac{x}{n}\right)^4 \left[1 + O\left(\frac{x}{n}\right)\right] \left[-5 - 3\frac{x}{n} + O\left(\left(\frac{x}{n}\right)^2\right)\right]^n. \end{aligned}$$

Für den ersten Term gilt offensichtlich

$$\lim_{n\to\infty} \left[1 + O\left(\frac{x}{n}\right)\right] \left[1 - \frac{x}{n} + O\left(\left(\frac{x}{n}\right)^2\right)\right]^n = e^{-x}.$$

Das ist der Anteil, der die exakte Lösung beschreibt. Für den zweiten Term gilt

$$\begin{aligned} &-\frac{1}{216}\left(\frac{x}{n}\right)^4 \left[1 + O\left(\frac{x}{n}\right)\right] \left[-5 - 3\frac{x}{n} + O\left(\left(\frac{x}{n}\right)^2\right)\right]^n = \\ &\quad -\frac{(-5)^n}{216}\left(\frac{x}{n}\right)^4 \left[1 + O\left(\frac{x}{n}\right)\right] \left[1 + \frac{3}{5}\frac{x}{n} + O\left(\left(\frac{x}{n}\right)^2\right)\right]^n. \end{aligned}$$

Wegen

$$\lim_{n\to\infty}\left[1+\frac{3}{5}\frac{x}{n}+O\left(\left(\frac{x}{n}\right)^2\right)\right]^n = e^{3x/5}$$

verhält sich der zweite Term für großes n wie

$$-\frac{(-5)^n}{216}\left(\frac{x}{n}\right)^4 e^{3x/5}.$$

Dieser Term oszilliert mit wachsendem n immer heftiger.
Genau dieses Verhalten beobachtet man auch bei Anwendung des obigen Verfahrens auf ein beliebiges AWP. Der Grund dafür ist darin zu suchen, dass die allgemeine Lösung der Differenzengleichung den Term

$$\lambda_2^j = [-5+O(h)]^j$$

enthält, der für großes j beliebig groß wird. ♡

Das Beispiel lässt vermuten, dass die Lösungen der Differenzengleichung

$$\boldsymbol{\eta}_{j+r}+\alpha_{r-1}\boldsymbol{\eta}_{j+r-1}+\cdots+\alpha_0\boldsymbol{\eta}_j = 0$$

entscheidenden Einfluss auf das Konvergenzverhalten des MSV haben. Genau dies ist der Fall. Ein MSV heißt **nullstabil**, falls das erste charakteristische Polynom

$$\Psi(\mu) = \mu^r+\alpha_{r-1}\mu^{r-1}+\cdots+\alpha_1\mu+\alpha_0$$

nur Nullstellen λ besitzt, für die $|\lambda|\leqq 1$ gilt, und die im Falle $|\lambda|=1$ einfach sind.
Bemerkung: Für die bisher behandelten MSV ist die Stabilitätsbedingung stets erfüllt. Man erhält für die ADAMS-BASHFORTH- und ADAMS-MOULTON-Verfahren die charakteristischen Polynome

$$\Psi(\mu) = \mu^r-\mu^{r-1} = (\mu-1)\mu^{r-1}$$

und für die NYSTRÖM- und MILNE-Verfahren die charakteristischen Polynome

$$\Psi(\mu) = \mu^r-\mu^{r-2} = (\mu+1)(\mu-1)\mu^{r-2}.$$

Genauere Untersuchungen zeigen, dass es keine r-Schritt-Verfahren der Konsistenzordnung $2r$ gibt, die auch nullstabil sind. Es gilt:

6.15. DAHLQUIST**'s 1. Ordnungsschranke:** *Für die maximale Konsistenzordnung eines nullstabilen r-Schritt-Verfahrens gilt*

$$p = \begin{cases} r+1 & \textit{für } r \textit{ ungerade} \\ r+2 & \textit{für } r \textit{ gerade} \end{cases}.$$

Ein r-Schritt-Verfahren der Konsistenzordnung $r+1$ ist z. B. durch ein Prediktor-Korrektor-Verfahren vom ADAMS-BASHFORTH-MOULTON-Typ gegeben.
Ein r-Schritt-Verfahren der Konsistenzordnung $r+2$ ist das MILNE-Verfahren für $q=2$:

$$\boldsymbol{\eta}_{p+1} = \boldsymbol{\eta}_{p-1} + \left[\frac{1}{3}\boldsymbol{f}(x_{p+1},\boldsymbol{\eta}_{p+1}) + \frac{4}{3}\boldsymbol{f}(x_p,\boldsymbol{\eta}_p) + \frac{1}{3}\boldsymbol{f}(x_{p-1},\boldsymbol{\eta}_{p-1})\right].$$

Löst man mit diesem Verfahren das AWP

$$y' = \kappa y, \quad y(0) = 1$$

mit der exakten Lösung $y(x) = e^{\kappa x}$, so erfüllen die Näherungslösungen $\eta_k = \eta(x_k;h)$ die homogene lineare Differenzengleichung

$$\left(1-\frac{\kappa h}{3}\right)\eta_{k+1} - \frac{4\kappa h}{3}\eta_k - \left(1+\frac{\kappa h}{3}\right)\eta_{k-1} = 0.$$

Die allgemeine Lösung dieser Differenzengleichung lässt sich wieder in der Form

$$\eta_k = \mu_1\lambda_1^k + \mu_2\lambda_2^k$$

angeben, wobei λ_1 und λ_2 Lösungen der quadratischen Gleichung

$$\left(1-\frac{\kappa h}{3}\right)\lambda^2 - \frac{4\kappa h}{3}\lambda - \left(1+\frac{\kappa h}{3}\right) = 0$$

sind. Es ergibt sich

$$\lambda_1 = \frac{1}{1-\frac{\kappa h}{3}}\left(\frac{2\kappa h}{3} + \sqrt{1+\frac{(\kappa h)^2}{3}}\right)$$

und

$$\lambda_2 = \frac{1}{1-\frac{\kappa h}{3}}\left(\frac{2\kappa h}{3} - \sqrt{1+\frac{(\kappa h)^2}{3}}\right).$$

Die Koeffizienten μ_1 und μ_2 sind aus den Anfangsbedingungen zu bestimmen.
Aus $y(0)=0$ folgt $\mu_1+\mu_2=1$. Damit gilt

$$\eta_1 = \mu_1\lambda_1 + (1-\mu_)\lambda_2,$$

folglich

$$\mu_1 = \frac{\eta_1-\lambda_2}{\lambda_1-\lambda_2}.$$

Entwickelt man λ_1 und λ_2 nach Potenzen von h, so gilt in erster Näherung

$$\lambda_1 \doteq \left(1+\frac{\kappa h}{3}\right)\left(\frac{2\kappa h}{3}+1\right) \doteq 1+\kappa h$$

und

$$\lambda_2 \doteq \left(1+\frac{\kappa h}{3}\right)\left(\frac{2\kappa h}{3}-1\right) \doteq -1+\frac{\kappa h}{3}.$$

Für die Näherungslösung an der Stelle x erhalten wir so

$$\eta(x;h) = \mu_1\left(1+\kappa h+O(h^2)\right)^{x/h} + (1-\mu_1)\left(-1+\frac{\kappa h}{3}+O(h^2)\right)^{x/h}.$$

Der erste Term verhält sich wie $e^{\kappa x}$. Er entspricht der exakten Lösung. Der zweite Term dagegen verhält sich wie

$$(-1)^{x/h}e^{-\kappa x/3}.$$

Für $\kappa < 0$, also eine exponentiell abfallende Lösung, liefert dieser Term eine oszillierende exponentiell wachsende Störung. Genau diese Verhalten beobachtet man in der Praxis bei allen r-Schritt-Verfahren der Konsistenzordnung $r+2$, falls diese auf AWP mit exponentiell abklingenden Lösungen angewendet werden. Damit ist die Anwendbarkeit dieser Verfahren stark eingeschränkt.
Neben der Konsistenz von MSV interessiert auch wieder die Konvergenz. Bei ESV folgt aus der Konsistenz die Konvergenz, und Konsistenzordnung und Konvergenzordnung stimmten überein. Bei MSV sind
die Verhältnisse so: Zunächst benötigt man, um ein r-Schritt-Verfahren zu starten, neben dem Anfangswert $\boldsymbol{\eta}_0 = \boldsymbol{y}_0$ weitere $r-1$ Näherungen $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{r-1}$ für die exakten Werte $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{r-1}$. (Diese beschafft man sich zum Beispiel mit einem ESV.) Die Güte dieser Näherungen beeinflusst natürlich die Güte aller weiteren Näherungswerte. Wir nehmen an, dass sich diese Startnäherungen gemäß

$$\begin{aligned}
\boldsymbol{\eta}_0 &= \boldsymbol{y}_0,\\
\boldsymbol{\eta}_1 &= \boldsymbol{y}_1+\boldsymbol{\varepsilon}_1(h),\\
\vdots &\quad \vdots\\
\boldsymbol{\eta}_{r-1} &= \boldsymbol{y}_{r-1}+\boldsymbol{\varepsilon}_{r-1}(h)
\end{aligned}$$

von den exakten Werten unterscheiden. Die berechneten Näherungen $\boldsymbol{\eta}_r, \boldsymbol{\eta}_{r+1}, \ldots$ hängen damit nicht nur von der Schrittweite und dem angewendeten MSV ab, sondern auch von der Güte der Anfangsnäherungen. Folglich ist für die Näherung $\boldsymbol{\eta}$ an der Stelle x zu schreiben

$$\boldsymbol{\eta}(x;\boldsymbol{\varepsilon};h),$$

wobei $\boldsymbol{\varepsilon}$ eine Funktion bezeichnet, für die $\boldsymbol{\varepsilon}_i = \boldsymbol{\varepsilon}(x_i; h)$ für $i = 1, \ldots, r-1$ gilt. Nun definieren wir den globalen Diskretisierungsfehler für MSV. Es sei $\boldsymbol{y}$ die exakte Lösung des AWP

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{y}(x_0) = \boldsymbol{y}_0.$$

$\boldsymbol{\eta}(x; \boldsymbol{\varepsilon}; h)$ bezeichne die mit einem MSV mit der Schrittweite h berechnete Näherung für $\boldsymbol{y}(x)$, wobei die Güte der Startnäherungen durch die Funktion $\boldsymbol{\varepsilon}$ beschrieben wird. Dann heißt die Größe

$$\boldsymbol{e}(x; \boldsymbol{\varepsilon}; h) = \boldsymbol{\eta}(x; \boldsymbol{\varepsilon}; h) - \boldsymbol{y}(x)$$

globaler Diskretisierungsfehler an der Stelle x zur Schrittweite h. Nun lässt sich auch sagen, was unter einem konvergenten MSV zu verstehen ist. Ein MSV zum Lösen von AWP der Form

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{y}(x_0) = \boldsymbol{y}_0$$

heißt **konvergent**, falls

$$\lim_{n \to \infty} \boldsymbol{e}(x; \boldsymbol{\varepsilon}; h_n) = \boldsymbol{o}, \quad h_n = \frac{x - x_0}{n},$$

für alle $x \in [a, b]$, für alle Funktionen $\boldsymbol{f} \in F^1[a, b]$ und für alle Funktionen $\boldsymbol{\varepsilon}(z; h)$ mit

$$\lim_{n \to \infty} \|\boldsymbol{\varepsilon}(z; h_n)\| = 0, \quad z = x_0 + i h_n, \quad i = 1, \ldots, r-1$$

gilt. Der folgende Satz, den wir ohne Beweis angeben, liefert Bedingungen für die Konvergenz von MSV.

6.16. Satz: *Es sei durch*

$$\boldsymbol{\eta}_{j+r} + \alpha_{r-1} \boldsymbol{\eta}_{j+r-1} + \cdots + \alpha_0 \boldsymbol{\eta}_j = h \boldsymbol{F}(x_j, \boldsymbol{\eta}_{j+r}, \ldots, \boldsymbol{\eta}_j; h; \boldsymbol{f})$$

ein konsistentes MSV zum Lösen eines AWP der Form

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{y}(x_0) = \boldsymbol{y}_0$$

gegeben. Die Funktion $\boldsymbol{F}$ erfülle folgende Bedingungen:

1. *$\boldsymbol{F} \equiv \boldsymbol{o}$ für alle $x \in [a, b]$, alle $\boldsymbol{\eta}_i \in \mathbb{R}^d$ und alle $h \in \mathbb{R}$ falls $\boldsymbol{f} \equiv \boldsymbol{o}$.*

2. *Zu jeder Funktion $\boldsymbol{F} \in F^1[a,b]$ gibt es Konstanten $h_0 > 0$ und M, so dass*

$$\|\boldsymbol{F}(x, \boldsymbol{v}_r, \ldots, \boldsymbol{v}_0; h; \boldsymbol{f}) - \boldsymbol{F}(x, \boldsymbol{w}_r, \ldots, \boldsymbol{w}_0; h; \boldsymbol{f})\| \leqq M \sum_{i=0}^{r} \|\boldsymbol{v}_i - \boldsymbol{w}_i\|$$

für alle $x \in [a,b]$, alle $\boldsymbol{v}_i, \boldsymbol{w}_i \in \mathbb{R}$, $i = 0, \ldots, r$ und alle $|h| \leqq h_0$ gilt.

Dann ist das gegebene MSV genau dann konvergent, wenn es nullstabil ist.

Bemerkungen: (i) Die erste Voraussetzung garantiert gemeinsam mit der Nullstabilität, dass das MSV das triviale AWP

$$\boldsymbol{y}' = \boldsymbol{o}, \quad \boldsymbol{y}(x_0) = \boldsymbol{y}_0$$

exakt löst, falls $\boldsymbol{\varepsilon}_1 = \boldsymbol{\varepsilon}_2 = \cdots = \boldsymbol{\varepsilon}_{r-1} = 0$ gilt.
(ii) Die zweite Bedingung bedeutet die LIPSCHITZ-Stetigkeit von $\boldsymbol{F}$ bezüglich der Näherungen $\boldsymbol{\eta}_j, \ldots, \boldsymbol{\eta}_{j+r}$.
(iii) Für lineare MSV ist die erste Bedingung trivialerweise erfüllt. Die zweite Bedingung folgt sofort aus der LIPSCHITZ-Stetigkeit von $\boldsymbol{f}$, die ja schon im Existenz- und Eindeutigkeitssatz gefordert wurde.
Damit folgt die Konvergenz der im Abschnitt 6.3.1. konstruierten Verfahren. Der folgende Satz macht noch eine Aussage über die Konvergenzordnung.

6.17. Satz: *Es sei durch*

$$\boldsymbol{\eta}_{j+r} + \alpha_{r-1}\boldsymbol{\eta}_{j+r-1} + \cdots + \alpha_0\boldsymbol{\eta}_j = h\boldsymbol{F}(x_j, \boldsymbol{\eta}_{j+r}, \ldots, \boldsymbol{\eta}_j; h; \boldsymbol{f})$$

ein nullstabiles MSV der Konsistenzordnung p zum Lösen eines AWP der Form

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{y}(x_0) = \boldsymbol{y}_0$$

gegeben. Die Funktion $\boldsymbol{F}$ erfülle alle Voraussetzungen aus Satz 6.16.
Dann gilt für alle $\boldsymbol{f} \in F^p[a,b]$ und alle $x \in [a,b]$

$$\|\boldsymbol{e}(x; \boldsymbol{\varepsilon}; h_n)\| = O(h_n^p),$$

falls für die Güte der Startnäherungen

$$\|\boldsymbol{\varepsilon}_i(h^p)\| = O(h^p), \quad i = 1, \ldots, r-1$$

gilt.

Um für ein MSV der Konsistenzordnung p auch die Konvergenzordnung p zu erreichen, ist es somit notwendig, die Startnäherungen ebenfalls mit der entsprechenden Güte zu berechnen, folglich zum Beispiel mit einem ESV der Ordnung p. Wenn wir ein Gesamtverfahren betrachten, das aus einem Startverfahren zum Berechnen der Näherungen $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{r-1}$ und einem Prediktor-Korrektor-Verfahren zum Berechnen der weiteren Näherungen besteht, so wird die Ordnung des gesamten Verfahrens durch das Teilverfahren mit der geringsten Ordnung bestimmt.

6.4. Extrapolationsverfahren

Alle bisher betrachteten Verfahren zum Lösen von AWP lieferten Näherungen, die von der verwendeten Schrittweite abhingen. Für die exakte Lösung $\boldsymbol{y}(x)$ an der Stelle x wird eine Näherungslösung $\boldsymbol{\eta}(x;h_n)$ mit $h_n = (x-x_0)/n$ berechnet. Gelingt es nun, für die Abhängigkeit der Näherungslösung $\boldsymbol{\eta}(x;h)$ von der Schrittweite h asymptotische Entwicklungen, analog der Entwicklung der Trapezsumme bei der numerischen Integration, herzuleiten, so lassen sich wie beim ROMBERG-Verfahren durch Extrapolation aus verschiedenen Näherungen, die mit verschiedenen Schrittweiten berechnet wurden, verbesserte Näherungen gewinnen. Leider ist nur in wenigen Spezialfällen die Existenz einer asymptotischen Entwicklung bekannt.

6.18. Satz von GRAGG**:** *Es sei $\boldsymbol{f} \in F^{2N+2}[a,b]$ und $\boldsymbol{y}$ die exakte Lösung des AWP*

$$\boldsymbol{y}' = \boldsymbol{f}(x,\boldsymbol{y}), \quad \boldsymbol{y}(x_0) = \boldsymbol{y}_0.$$

Für $x \in R_h = \{\, x_0 + ih \mid i = 0,1,\dots \}$ sei $\boldsymbol{\eta}(x;h)$ definiert durch

$$\left.\begin{array}{rcl} \boldsymbol{\eta}_0 & = & \boldsymbol{\eta}(x_0;h) = \boldsymbol{y}_0, \\ \boldsymbol{\eta}_1 & = & \boldsymbol{\eta}(x_1;h) = \boldsymbol{y}_0 + h\boldsymbol{f}(x_0,\boldsymbol{y}_0), \\ \boldsymbol{\eta}_{k+1} & = & \boldsymbol{\eta}(x_{k+1};h) = \boldsymbol{\eta}_{k-1} + 2h\boldsymbol{f}(x_k,\boldsymbol{\eta}_k), \\ x_{k+1} & = & x_k + h \end{array}\right\}, \quad k = 1,2,\dots.$$

Dann besitzt $\boldsymbol{\eta}(x;h)$ eine Entwicklung der Form

$$\boldsymbol{\eta}(x;h) = \boldsymbol{y}(x) + \sum_{i=1}^{N} h^{2i}\left[\boldsymbol{u}_i(x) + (-1)^{\frac{x-x_0}{h}}\boldsymbol{v}_i(x)\right] + h^{2N+2}\boldsymbol{E}_{2N+2}(x;h),$$

die für alle $x \in (a,b)$ und alle $h = (x-x_0)/n$, $n = 1,2,\dots$, gilt. Die Funktionen $\boldsymbol{u}_i$ und $\boldsymbol{v}_i$, $i = 1,2,\dots,N$, sind von h unabhängig. Das Restglied $\boldsymbol{E}_{2N+2}$ bleibt bei festem x für alle $h = (x-x_0)/n$ beschränkt.

Die in diesem Satz angegebene Entwicklung berechtigt uns noch nicht, analog zum ROMBERG-Verfahren vorzugehen. Dazu benötigen wir eine Entwicklung der Form

$$\boldsymbol{\eta}(x;h) = \boldsymbol{y}(x) + \sum_{i=1}^{N} \boldsymbol{\tau}_i(x)h^{2i} + h^{2N+2}\boldsymbol{E}_{2N+2}(x;h),$$

mit von h unabhängigen Funktionen $\boldsymbol{\tau}_i(x)$. In der Entwicklung aus Satz 6.18 hängt der Ausdruck

$$\boldsymbol{u}_i(x) + (-1)^{\frac{x-x_0}{h}}\boldsymbol{v}_i(x)$$

über den Term $(-1)^{(x-x_0)/h}$ von der Schrittweite h ab. Diese Abhängigkeit lässt sich beseitigen, indem man die Schrittweite h so wählt, dass der Exponent $(x-x_0)/h$ immer gerade oder immer ungerade ist. Wir wählen die Schrittweiten folglich in der Form

$$h = \frac{x-x_0}{2n} \quad \text{oder} \quad h = \frac{x-x_0}{2n-1}.$$

Das liefert im ersten Falle Entwicklungen der Form

$$\boldsymbol{\eta}(x;h) = \boldsymbol{y}(x) + \sum_{i=1}^{N} h^{2i} \left[\boldsymbol{u}_i(x) + \boldsymbol{v}_i(x)\right] + h^{2N+2} \boldsymbol{E}_{2N+2}(x;h)$$

und im zweiten Falle Entwicklungen der Form

$$\boldsymbol{\eta}(x;h) = \boldsymbol{y}(x) + \sum_{i=1}^{N} h^{2i} \left[\boldsymbol{u}_i(x) - \boldsymbol{v}_i(x)\right] + h^{2N+2} \boldsymbol{E}_{2N+2}(x;h).$$

Dies sind wirkliche asymptotische Entwicklungen in h^2. Eine weitere Verbesserung der Entwicklung erreicht man durch folgenden Trick (GRAGG), der den Oszillationsterm im ersten Glied prinzipiell beseitigt.

6.19. GRAGG**sche Näherung:**
Es ist für die Stelle x eine Näherungslösung des AWP

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{y}(x_0) = \boldsymbol{y}_0$$

zu berechnen.

S0 *Wähle ein n, setze $h = (x-x_0)/n$ und*

$$\boldsymbol{\eta}_0 = \boldsymbol{\eta}(x_0;h) = \boldsymbol{y}_0,$$
$$\boldsymbol{\eta}_1 = \boldsymbol{\eta}(x_1;h) = \boldsymbol{y}_0 + h\boldsymbol{f}(x_0, \boldsymbol{y}_0).$$

S1 *Für $k = 1, 2, \ldots, n-1$ berechne*

$$\boldsymbol{\eta}_{k+1} = \boldsymbol{\eta}_{k-1} + 2h\boldsymbol{f}(x_k, \boldsymbol{\eta}_k),\ x_{k+1} = x_k + h.$$

S2 *Berechne*

$$\tilde{\boldsymbol{\eta}}(x;h) = \frac{1}{2}\left[\boldsymbol{\eta}_n + \boldsymbol{\eta}_{n-1} + h\boldsymbol{f}(x_n, \boldsymbol{\eta}_n)\right].$$

Für die GRAGGsche Näherung lässt sich unter den Voraussetzungen von Satz 6.18 leicht zeigen, dass sie eine Entwicklung der Form

$$\begin{aligned}\tilde{\boldsymbol{\eta}}(x;h) &= \boldsymbol{y}(x)+h^2\left[\boldsymbol{u}_1(x)+\frac{1}{4}\boldsymbol{y}''(x)\right]+\sum_{i=2}^{N}h^{2i}\left[\tilde{\boldsymbol{u}}_i(x)+(-1)^{\frac{x-x_0}{h}}\tilde{\boldsymbol{v}}_i(x)\right]\\ &\quad +h^{2N+2}\tilde{\boldsymbol{E}}_{2N+2}(x;h)\end{aligned}$$

besitzt.
Damit ergibt sich der folgende Algorithmus.

6.20. Extrapolationsverfahren zum Lösen von AWP:

S0 *Wähle Grundschrittweite H und eine zu erreichende Genauigkeit ε. Setze $x = x_0 + H$, $h_0 = H/2$ und $k = 0$.*

S1 *Berechne mit der Schrittweite h_k eine* GRAGG*'sche Näherung*

$$\boldsymbol{T}_{k0} = \tilde{\boldsymbol{\eta}}(x;h_k).$$

S2 *Für $l = 1,\ldots,k$ berechne*

$$\boldsymbol{T}_{k,l} = \boldsymbol{T}_{k,l-1} + \frac{\boldsymbol{T}_{k,l-1}-\boldsymbol{T}_{k-1,l-1}}{\left(\frac{h_{k-l}}{h_k}\right)^2-1}.$$

S3 *Gilt*

$$\|\boldsymbol{T}_{k,k}-\boldsymbol{T}_{k,k-1}\| \leqq \varepsilon,$$

so setze $\boldsymbol{y}_0 = \boldsymbol{\eta}(x) = \boldsymbol{T}_{k,k}$, $x_0 = x$ und gehe zu Schritt **S0**.

S4 *Setze $h_{k+1} = h_k/2$ und $k = k+1$. Gehe zu Schritt* **S1**.

Bemerkungen: (i) Die Abbruchbedingung im Schritt **S3** lässt sich etwas verfeinern. Man bricht den Aufbau des Interpolationsschemas im Schritt **S2** ab, falls sich zwei benachbarte Werte im Rahmen der Genauigkeit ε nicht mehr unterscheiden. Um zufällige Effekte zu vermeiden, sollte man erst dann abbrechen, wenn diese Bedingung mehrmals hintereinander erfüllt war.
(ii) Man sollte in Schritt **S2** nur $5\ldots7$ Spalten des Interpolationsschemas berechnen, da die Polynominterpolation für Polynome höheren Grades ungenaue Resultate liefert.
(iii) Statt der Polynominterpolation darf wieder Rationale Interpolation verwendet werden. Die Rekursionsformel in Schritt **S2** ist dann gemäß

$$\boldsymbol{T}_{k,l} = \boldsymbol{T}_{k,l-1} + \frac{\boldsymbol{T}_{k,l-1}-\boldsymbol{T}_{k-1,l-1}}{\left(\frac{h_{k-l}}{h_k}\right)^2\left[1-\frac{\boldsymbol{T}_{k,l-1}-\boldsymbol{T}_{k-1,l-1}}{\boldsymbol{T}_{k,l-1}-\boldsymbol{T}_{k-1,l-2}}\right]-1}$$

abzuändern.
(iv) Die Grundschrittweite H wird von Schritt zu Schritt geändert. Man sollte sie so wählen, dass das Interpolationsschema nach $5 \dots 7$ Spalten abbricht.
Die Konvergenzaussagen zur ROMBERG-Integration gelten entsprechend auch für die mit dem Algorithmus 6.20 berechneten Werte. Insbesondere gilt also

$$\lim_{k\to\infty} \boldsymbol{T}_{k,l} = \boldsymbol{y}(x).$$

Somit konvergieren die Spalten des Interpolationsschemas gegen die exakte Lösung. Die Konvergenzordnung ist dabei $2k+2$ falls $\boldsymbol{f} \in F^{2k+2}[a,b]$ und $k \leqq N$.

6.5. Aufgaben

1. Man bestimme die Lösung des Anfangswertproblems

   $$\boldsymbol{y}' = \boldsymbol{A}\boldsymbol{y}, \quad \boldsymbol{y}(0) = \boldsymbol{y}_0$$

   mit $\boldsymbol{y}_0 \in \mathbb{R}^d$ und der $d \times d$-Matrix

   $$\boldsymbol{A} = \begin{bmatrix} \lambda & 1 & & & 0 \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 1 \\ 0 & & & & \lambda \end{bmatrix}, \ \lambda \in \mathbb{R}.$$

2. Gegeben sei das Anfangswertproblem

   $$y' = x^2 + 2x^3 \qquad y(0) = 0.$$

   Zur Schrittweite h sollen mit dem EULER-Verfahren Näherungswerte $\eta(x_j; h)$ für $y(x_j)$, $x_j = jh$ berechnet werden. Man gebe $\eta(x_j; h)$ und den globalen Diskretisierungsfehler $e(x_j; h)$ explizit an und zeige, dass $e(x; h)$ bei festem x für $h = \frac{x}{n} \to 0$ gegen Null geht.

3. Es sei ein Einschrittverfahren durch

   $$\Phi(x, y; h) = \sum_{i=0}^{n} \alpha_i f(\xi_i, \eta_i)$$

mit

$$\xi_0 = x\ , \qquad \xi_i = x + \vartheta_i h\ ,\ i = 1, \ldots, n$$

$$\eta_0 = y\ , \qquad \eta_i = y + h \sum_{j=0}^{i-1} \beta_{ij} f(\xi_j, \eta_j)\ ,\ i = 1, \ldots, n$$

gegeben. Welche Gleichungen haben die Parameter α_i, ϑ_i und β_{ij} zu erfüllen, damit man für $n = 2$ ein Verfahren 3.Ordnung erhält? Man gebe eine Lösung dieser Gleichungen an.

4. Man schreibe ein Programm zum Berechnen der Iterationsfunktion $\Phi(x, y; h)$ für beliebiges n und beliebige Parameter; n und die Parameter sollen beim Aufruf des Programms übergeben werden.

5. Das Anfangswertproblem

$$y' = y^{\frac{1}{2}},\ y(0) = 0$$

hat die Lösung $y(x) = \frac{x^2}{4}$. Die Anwendung des EULER-Verfahrens liefert jedoch $\eta(x; h) = 0$ für alle x und $h = \frac{x}{n},\ n = 1, 2, \ldots$ Man begründe dieses Verhalten.

6. Gegeben sei das Anfangswertproblem

$$y' = \lambda y, \quad y(0) = 1.$$

Welche Schrittweitenschätzung liefert die Schrittweitensteuerung

   (a) nach Algorithmus 6.9 mit dem EULER-Verfahren,

   (b) nach Algorithmus 6.9 mit dem impliziten EULER-Verfahren,

   (c) nach der RUNGE-KUTTA-FEHLBERG-Methode mit EULER- und HEUN-Verfahren?

7. Man zeige, dass das modifizierte EULER-Verfahren die exakte Lösung der Differentialgleichung

$$y' = -2ax$$

liefert.

8. Durch

$$F(x,y;h) = f(x,y) + h\frac{g(x+\frac{h}{3}, y+\frac{hf(x,y)}{3})}{2}$$

mit

$$g(x,y) = f_x(x,y) + f_y(x,y)f(x,y)$$

ist ein Einschrittverfahren gegeben. Man zeige, dass es sich um ein Verfahren 3. Ordnung handelt.

9. Bei den ADAMS-MOULTON-Verfahren wird, ausgehend von den Näherungswerten

$$\eta_{p-j}, \ldots, \eta_p \text{ für } y(x_{p-j}), \ldots, y(x_p)$$

ein Näherungswert η_{p+1} für $y(x_{p+1}), x \in [a,b]$, durch folgende Iterationsvorschrift berechnet:

$\eta_{p+1}^{(0)}$ beliebig

für $i = 0, 1, \ldots$

$\eta_{p+1}^{(i+1)} = \Psi(\eta_{p+1}^{(i)}) = \eta_p + h[\beta_{q0} f(x_{p+1}, \eta_{p+1}^{(i)}) + \beta_{q1} f_p + \cdots + \beta_{qq} f_{p+1-q}]$.

Man zeige: Für einmal stetig partiell differenzierbare Funktionen f gibt es ein $h_0 > 0$, so dass für alle $|h| \leq h_0$ die Folge $\eta_{p+1}^{(i)}$ gegen ein $\eta_{p+1} = \Psi(\eta_{p+1})$ konvergiert.

10. Man prüfe, ob das lineare Mehrschrittverfahren

$$\eta_p - \eta_{p-4} = \frac{h}{3}(8f_{p-1} - 4f_{p-2} + 8f_{p-3})$$

konvergent ist.

11. Man bestimme α, β und γ so, dass das lineare Mehrschrittverfahren

$$\eta_{j+4} - \eta_{j+2} + \alpha(\eta_{j+3} - \eta_{j+1}) = h[\beta(f_{j+3} - f_{j+1}) + \gamma f_{j+2}]$$

die Ordnung 3 hat. Ist das so gewonnene Verfahren stabil?

12. Es werde das durch

$$\eta_{j+2} + a_1\eta_{j+1} + a_0\eta_j = h[b_0 f(x_j, \eta_j) + b_1 f(x_{j+1}, \eta_{j+1})]$$

gegebene PREDIKTOR-Verfahren betrachtet.

(a) Man bestimme a_0, b_0 und b_1 in Abhängigkeit von a_1 so, dass ein Verfahren mindestens 2.Ordnung entsteht.

(b) Für welche a_1-Werte ist das so gewonnene Verfahren stabil?

(c) Welche speziellen Verfahren erhält man für $a_1 = 0$ und $a_1 = -1$?

(d) lässt sich a_1 so wählen, dass man stets ein stabiles Verfahren 3.Ordnung erhält?

Kapitel 7

Randwertprobleme

7.1. Einführung

Allgemeiner als AWP sind Randwertprobleme (RWP):

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{r}(\boldsymbol{y}(a), \boldsymbol{y}(b)) = 0$$

mit

$$\begin{aligned} \boldsymbol{y} &: [a,b] \to \mathbb{R}^n, \\ \boldsymbol{f} &: [a,b] \times \mathbb{R}^n \to \mathbb{R}^n, \\ \boldsymbol{r} &: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n. \end{aligned}$$

Oft liegen die Randbedingungen in linearer Form vor:

$$\boldsymbol{r}(\boldsymbol{y}(a), \boldsymbol{y}(b)) = \boldsymbol{A}\boldsymbol{y}(a) + \boldsymbol{B}\boldsymbol{y}(b) - \boldsymbol{c}, \quad \boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times n}, \quad \boldsymbol{c} \in \mathbb{R}^n$$

gilt. Man spricht von separierten Randbedingungen, falls

$$\boldsymbol{A}_1 \boldsymbol{y}(a) = \boldsymbol{c}_1, \quad \boldsymbol{B}_2 \boldsymbol{y}(b) = \boldsymbol{c}_2$$

gilt, also Permutationsmatrizen $\boldsymbol{P}_1$ und $\boldsymbol{P}_2$ existieren, so dass die Zeilen und Spalten der Matrix $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{c})$ so vertauschbar sind, dass

$$\boldsymbol{P}_1 (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{c}) \boldsymbol{P}_2 = \begin{pmatrix} \boldsymbol{A}_1 & \boldsymbol{0} & \boldsymbol{c}_1 \\ \boldsymbol{0} & \boldsymbol{B}_2 & \boldsymbol{c}_2 \end{pmatrix}$$

gilt. AWP sind spezielle RWP mit $\boldsymbol{A} = \boldsymbol{I}$, $\boldsymbol{B} = \boldsymbol{0}$ und $\boldsymbol{c} = \boldsymbol{y}_0$.
Für RWP existiert kein Existenz- und Eindeutigkeitssatz wie für AWP, der unter schwachen Voraussetzungen die Existenz und Eindeutigkeit der Lösung sichert.

7.1. Beispiel: Es sei

$$\boldsymbol{y}' = \begin{pmatrix} y_1' \\ y_2' \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \boldsymbol{y} = \boldsymbol{A}\boldsymbol{y}.$$

Die exakte Lösung ist

$$\boldsymbol{y}(x) = e^{\boldsymbol{A}x}\boldsymbol{c}, \quad \boldsymbol{c} \in \mathbb{R}^2.$$

Mit

$$\boldsymbol{A} = \frac{1}{\sqrt{2}} \begin{pmatrix} -i & i \\ 1 & 1 \end{pmatrix} \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} i & 1 \\ -i & 1 \end{pmatrix}$$

erhält man

$$\begin{aligned} \exp(\boldsymbol{A}x) &= \frac{1}{2} \begin{pmatrix} -i & i \\ 1 & 1 \end{pmatrix} \begin{pmatrix} e^{ix} & 0 \\ 0 & e^{-ix} \end{pmatrix} \begin{pmatrix} i & 1 \\ -i & 1 \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} -ie^{ix} & ie^{-ix} \\ e^{ix} & e^{-ix} \end{pmatrix} \begin{pmatrix} i & 1 \\ -i & 1 \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} e^{ix} + e^{-ix} & -ie^{ix} + ie^{-ix} \\ ie^{ix} - ie^{-ix} & e^{ix} + e^{-ix} \end{pmatrix} \\ &= \begin{pmatrix} \cos x & \sin x \\ -\sin x & \cos x \end{pmatrix}. \end{aligned}$$

Die allgemeine Lösung der Differentialgleichung lautet damit

$$\begin{aligned} \boldsymbol{y} &= \begin{pmatrix} \cos x & \sin x \\ -\sin x & \cos x \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \\ &= \begin{pmatrix} c_1 \cos x + c_2 \sin x \\ -c_1 \sin x + c_2 \cos x \end{pmatrix}. \end{aligned}$$

Betrachten wir nun verschiedene Randbedingungen.

1. $(1,0)\boldsymbol{y}(0) = 0$ und $(1,0)\boldsymbol{y}(\pi/2) = 1$:
   Es folgt

   $$(1,0) \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = c_1 = 0$$

   und

   $$(1,0) \begin{pmatrix} c_2 \\ -c_1 \end{pmatrix} = c_2 = 1.$$

Damit erhalten wir die eindeutige Lösung

$$\boldsymbol{y} = \begin{pmatrix} \sin x \\ \cos x \end{pmatrix}.$$

2. $(1,0)\boldsymbol{y}(0) = 0$ und $(1,0)\boldsymbol{y}(\pi) = 1$:
   Es folgt

$$(1,0)\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = c_1 = 0$$

und

$$(1,0)\begin{pmatrix} -c_1 \\ -c_2 \end{pmatrix} = -c_1 = 1.$$

Das ist ein Widerspruch. Es existiert keine Lösung dieses RWP.

3. $(1,0)\boldsymbol{y}(0) = 0$ und $(1,0)\boldsymbol{y}(\pi) = 0$:
   Es folgt

$$(1,0)\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = c_1 = 0$$

und

$$(1,0)\begin{pmatrix} -c_1 \\ -c_2 \end{pmatrix} = -c_1 = 0.$$

In diesem Falle ist c_2 beliebig wählbar. Es existieren unendlich viele Lösungen

$$\boldsymbol{y} = \begin{pmatrix} c_2 \sin x \\ c_2 \cos x \end{pmatrix}.$$

♡

Unter starken Voraussetzungen lässt sich ein Existenz- und Eindeutigkeitssatz formulieren.

7.2. Satz: *Für das RWP* $\boldsymbol{y}' = \boldsymbol{f}(x,\boldsymbol{y}), \quad \boldsymbol{r}(\boldsymbol{y}(a),\boldsymbol{y}(b)) = \boldsymbol{o}$ *gelte*

1. $\boldsymbol{f}$ *und* $\frac{\partial}{\partial \boldsymbol{y}}\boldsymbol{f}$ *sind stetig auf dem Streifen* $S = [a,b] \times \mathbb{R}^n$.
2. *Es existiert eine Funktion* $K \in C[a,b]$ *mit*

$$\left\| \frac{\partial}{\partial \boldsymbol{y}} \boldsymbol{f}(x,\boldsymbol{y}) \right\| \leqq K(x) \quad \textit{für alle} \quad (x,\boldsymbol{y}) \in S.$$

3. *Die Matrix*

$$\boldsymbol{P}(\boldsymbol{u},\boldsymbol{v}) = \frac{\partial}{\partial \boldsymbol{u}}\boldsymbol{r}(\boldsymbol{u},\boldsymbol{v}) + \frac{\partial}{\partial \boldsymbol{v}}\boldsymbol{r}(\boldsymbol{u},\boldsymbol{v})$$

besitzt für alle $\boldsymbol{u},\boldsymbol{v} \in \mathbb{R}^n$ *eine Darstellung der Form*

$$\boldsymbol{P}(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{P}_0\,(\boldsymbol{I} + \boldsymbol{M}(\boldsymbol{u},\boldsymbol{v}))$$

mit einer konstanten regulären Matrix $\boldsymbol{P}_0$ *und einer Matrix* $\boldsymbol{M}(\boldsymbol{u},\boldsymbol{v})$*, für die*

$$\left\| \boldsymbol{P}_0^{-1} \frac{\partial}{\partial \boldsymbol{v}}\boldsymbol{r}(\boldsymbol{u},\boldsymbol{v}) \right\| \leqq m$$

und

$$\|\boldsymbol{M}(\boldsymbol{u},\boldsymbol{v})\| \leqq \mu$$

für alle $\boldsymbol{u},\boldsymbol{v} \in \mathbb{R}^n$ *mit Konstanten* $m < \infty$ *und* $\mu < 1$ *gilt.*

4. *Es existiert eine Zahl* λ *mit* $0 < \lambda < 1 - \mu$*, so dass*

$$\int_a^b K(t)\,dt \leqq \ln\left(1 + \frac{\lambda}{m}\right).$$

Dann besitzt das obige RWP genau eine Lösung $\boldsymbol{y}(x)$*.*

Wie das Beispiel zeigte, sind die Voraussetzungen dieses Satzes schon für einfache Fälle nicht erfüllt.

7.2. Das einfache Schießverfahren

Wir wollen das RWP

$$\boldsymbol{y}' = \boldsymbol{f}(x,\boldsymbol{y}), \quad \boldsymbol{r}(\boldsymbol{y}(a),\boldsymbol{y}(b)) = \boldsymbol{o}$$

lösen. Dazu wenden wir die folgende Idee an. Wir betrachten das AWP

$$\boldsymbol{y}' = \boldsymbol{f}(x,\boldsymbol{y}), \quad \boldsymbol{y}(a) = \boldsymbol{s}$$

mit dem frei wählbaren Parameter $\boldsymbol{s}$. Es sei $\boldsymbol{y}(x;\boldsymbol{s})$ die exakte Lösung dieses AWP. An der Stelle $x = b$ hat diese Lösung dann den Wert $\boldsymbol{y}(b;\boldsymbol{s})$. Damit die Lösung des AWP für ein gewisses $\boldsymbol{s}$ auch Lösung des RWP ist, muss

$$\boldsymbol{r}(\boldsymbol{y}(a,\boldsymbol{s}),\boldsymbol{y}(b,\boldsymbol{s})) = \boldsymbol{r}(\boldsymbol{s},\boldsymbol{y}(b,\boldsymbol{s})) = \boldsymbol{o}$$

gelten. Wir versuchen, den Parameter $\boldsymbol{s}$ so zu bestimmen, dass die Lösung $\boldsymbol{y}(x;\boldsymbol{s})$ des AWP auch die Randbedingungen erfüllt. Diese Vorgehensweise wird als einfaches Schießverfahren bezeichnet.

7.3. Einfaches Schießverfahren zum Lösen von RWP:
Gegeben sei das RWP

$$\boldsymbol{y}' = \boldsymbol{f}(x,\boldsymbol{y}), \quad \boldsymbol{r}(\boldsymbol{y}(a),\boldsymbol{y}(b)) = \boldsymbol{o}.$$

S0 *Bestimme den Parameter $\boldsymbol{s}$ so, dass die Lösung $\boldsymbol{y}(x;\boldsymbol{s})$ des AWP*

$$\boldsymbol{y}' = \boldsymbol{f}(x,\boldsymbol{y}), \quad \boldsymbol{y}(a) = \boldsymbol{s}$$

die Randbedingungen

$$\boldsymbol{r}(\boldsymbol{y}(a;\boldsymbol{s}),\boldsymbol{y}(b;\boldsymbol{s})) = \boldsymbol{r}(\boldsymbol{s},\boldsymbol{y}(b;\boldsymbol{s})) = \boldsymbol{o}.$$

erfüllt.

Wir haben ein nichtlineares Gleichungssystem

$$\boldsymbol{F}(\boldsymbol{s}) = \boldsymbol{r}(\boldsymbol{s},\boldsymbol{y}(b;\boldsymbol{s})) = \boldsymbol{o}$$

zu lösen. Dazu ist prinzipiell jedes Verfahren anwendbar, das zum Berechnen von Nullstellen für Funktionen geeignet ist. Die Schwierigkeit besteht darin, dass bei jeder Funktionswertberechnung von $\boldsymbol{F}$ ein AWP zu lösen ist. Im eindimensionalen Fall

$$y : [a,b] \to \mathbb{R}$$

lassen sich die Methoden aus Kapitel 5 anwenden, um eine Nullstelle der entsprechenden nichtlinearen Gleichung $F(s) = r(s,y(b;s)) = 0$ zu berechnen. Die einfachste Möglichkeit wäre, das Bisektionsverfahren anzuwenden.

7.4. Bisektionsverfahren zum Lösen eindimensionaler RWP:
Gegeben sei das RWP

$$y' = f(x,y), \quad r(y(a),y(b)) = 0.$$

Für einen gewissen Parameter s sei $y(x;s)$ die Lösung des AWP

$$y' = f(x,y), \quad y(a) = s.$$

Weiterhin sei eine Funktion F gemäß

$$F(s) = r(s,y(b;s))$$

definiert. Gegeben seien Werte s_0 und t_0 mit

$$F(s_0)F(t_0) < 0.$$

S0 *Setze* $k = 0$.

S1 *Setze* $c = (s_k + t_k)/2$ *und berechne die Lösung des AWP*

$$y' = f(x, y), \quad y(a) = c$$

an der Stelle $x = b$.

S2 *Berechne* $d = F(c)$.

S3 *Für*

$$d \cdot F(s_k) \begin{cases} > 0 & s_{k+1} = c, \quad t_{k+1} = t_k, \\ = 0 & s^* = c, \quad \textit{STOPP}, \\ < 0 & s_{k+1} = s_k, \quad t_{k+1} = c. \end{cases}$$

S4 *Setze* $k = k + 1$ *und gehe zu Schritt* **S1**.

Um schnellere Konvergenz zu erhalten, verwendet man das NEWTON-Verfahren:

$$s_{k+1} = s_k - \frac{F(s_k)}{F'(s_k)}.$$

Dazu ist aber die Kenntnis der ersten Ableitung von $F(s) = r(s, y(b; s))$ notwendig. Es gilt

$$F'(s) = \frac{d}{ds} r(s, y(b; s)) = r_u(s, y(b; s)) + r_v(s, y(b; s)) \frac{d}{ds} y(b; s).$$

Die Ableitung $\frac{d}{ds} y(b; s)$ wird aus einem weiteren AWP berechnet. Durch formale Integration erhält man aus dem AWP

$$y' = f(x, s), \quad y(a) = s$$

mit der exakten Lösung $y(x; s)$

$$\int_a^x y'(t; s)\, dt = \int_a^x f(t, y(t; s))\, dt,$$

$$y(x; s) - y(a; s) = \int_a^x f(t, y(t; s))\, dt,$$

$$y(x; s) = s + \int_a^x f(t, y(t; s))\, dt,$$

$$\frac{\partial}{\partial s}y(x;s) = 1 + \int_a^x \frac{\partial}{\partial y}f(t,y(t;s))\frac{\partial}{\partial s}y(t;s)\,dt.$$

Es sei nun für festes s

$$v(x) = \frac{\partial}{\partial s}y(x;s),$$

also

$$v(x) = 1 + \int_a^x f_y(t,y(t;s))v(t)\,dt,$$

woraus $v' = f_y(x,y(x;s))v, \quad v(a) = 1$ folgt. Damit ergibt sich der folgende Algorithmus.

7.5. Newton-Verfahren zum Lösen von eindimensionalen RWP:

S0 *Wähle s_0 und setze $k = 0$.*

S1 *Berechne simultan die Lösungen $y(b;s_k)$ und $v(b;s_k)$ der AWP*

$$y' = f(x,y), \quad y(a) = s_k$$

$$v' = f_y(x,y)v, \quad v(a) = 1$$

an der Stelle $x = b$.

S2 *Berechne $F(s_k) = r(s_k, y(b;s_k))$.*

S3 *Berechne*

$$F'(s_k) = r_u(s_k,y(b;s_k)) + r_v(s_k,y(b;s_k))v(b;s_k).$$

S4 *Berechne*

$$s_{k+1} = s_k - \frac{F(s_k)}{F'(s_k)}.$$

Setze $k = k+1$ und gehe zu Schritt **S1**.

Wegen der partiellen Ableitung von f nach y ist das zusätzliche AWP im allgemeinen wesentlich komplizierter. Damit wird die Anwendung des NEWTON-Verfahrens aufwendig. Um das Lösen dieses komplizierten AWP zu vermeiden, wendet man ein Quasi-NEWTON-Verfahren an. Hier wird $F'(s)$ durch einen Differenzenquotienten

$$F'(s) \approx \frac{F(s+\Delta s) - F(s)}{\Delta s}$$

approximiert. Zum Berechnen dieses Differenzenquotienten ist ein weiteres AWP mit der Anfangsbedingung $y(a) = s + \Delta s$ zu lösen. In diesem Falle erhalten wir folgendes Verfahren.

7.6. Quasi-Newton-Verfahren zum Lösen von eindimensionalen RWP:

S0 *Wähle* s_0 *und setze* $k = 0$.

S1 *Berechne die Lösung* $y(b; s_k)$ *des AWP*

$$y' = f(x, y), \quad y(a) = s_k$$

an der Stelle $x = b$.

S2 *Wähle ein* Δs_k *und berechne die Lösung* $y(b; s_k + \Delta s_k)$ *des AWP*

$$y' = f(x, y), \quad y(a) = s_k + \Delta s_k$$

an der Stelle $x = b$.

S3 *Berechne* $F(s_k) = r(s_k, y(b; s_k))$ *und*

$$F(s_k + \Delta s_k) = r(s_k + \Delta s_k, y(b; s_k + \Delta s_k)).$$

S4 *Berechne*

$$D_{\Delta s_k} F(s_k) = \frac{F(s_k + \Delta s_k) - F(s_k)}{\Delta s_k}.$$

S5 *Berechne*

$$s_{k+1} = s_k - \frac{F(s_k)}{D_{\Delta s_k} F(s_k)}$$

S6 *Setze* $k = k + 1$ *und gehe zu Schritt* **S1**.

Die Konvergenz des Verfahrens hängt dabei stark von der Wahl der Δs_k ab. Wählt man Δs_k zu groß, so ist $D_{\Delta s_k} F(s_k)$ eine schlechte Approximation für $F'(s_k)$ und die Iteration konvergiert wesentlich schlechter als das NEWTON-Verfahren. Wählt man Δs_k zu klein, so gilt $F(s_k + \Delta s_k) \approx F(s_k)$ und beim Berechnen von $D_{\Delta s_k} F(s_k)$ tritt Auslöschung auf, was zu großen Fehlern in der Lösung führt.
Man hat folglich, um mit diesem Verfahren gute Resultate zu erzielen, zum einen in jedem Schritt Δs_k vernünftig wählen, und zum anderen die AWP zum Berechnen von $F(s_k + \Delta s_k)$ und $F(s_k)$ genau lösen. (Der relative Fehler von $F(s_k + \Delta s_k)$ und $F(s_k)$ sollte nur in der Größenordnung von eps liegen.) Diese Genauigkeit lässt sich mit Extrapolationsverfahren erreichen. Als vernünftige Wahl von Δs_k hat sich

$$\Delta s_k = \sqrt{\text{eps}}\, s_k$$

erwiesen. Hier wird erreicht, dass sich bei der Differenzbildung $F(s_k + \Delta s_k) - F(s_k)$ ungefähr die Hälfte der Mantissenstellen auslöscht.
Betrachten wir nun noch den allgemeinen Fall:

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{r}(\boldsymbol{y}(a), \boldsymbol{y}(b)) = \boldsymbol{o}$$

mit

$$\begin{aligned}\boldsymbol{y} &: [a,b] \to \mathbb{R}^n \\ \boldsymbol{f} &: [a,b] \times \mathbb{R}^n \to \mathbb{R}^n \\ \boldsymbol{r} &: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n.\end{aligned}$$

In diesem Falle ist das nichtlineare Gleichungssystem

$$\boldsymbol{F}(\boldsymbol{s}) = \boldsymbol{r}(\boldsymbol{s}, \boldsymbol{y}(b;\boldsymbol{s})) = \boldsymbol{o}$$

zu lösen. Man wendet dazu das allgemeine NEWTON-Verfahren

$$\boldsymbol{s}^{(k+1)} = \boldsymbol{s}^{(k)} - \left[\boldsymbol{F}'(\boldsymbol{s}^{(k)})\right]^{-1} \boldsymbol{F}(\boldsymbol{s}^{(k)})$$

an. Hierbei bezeichnet $\boldsymbol{F}'$ die JACOBI-Matrix von $\boldsymbol{F}$:

$$\boldsymbol{F}'(\boldsymbol{s}) = \boldsymbol{r}_{\boldsymbol{u}}(\boldsymbol{s}, \boldsymbol{y}(b;\boldsymbol{s})) + \boldsymbol{r}_{\boldsymbol{v}}(\boldsymbol{s}, \boldsymbol{y}(b;\boldsymbol{s}))\boldsymbol{y}_{\boldsymbol{s}}(b;\boldsymbol{s})$$

mit

$$\begin{aligned}\boldsymbol{r}_{\boldsymbol{u}}(\boldsymbol{u},\boldsymbol{v}) &= \left(\frac{\partial \boldsymbol{r}_i(\boldsymbol{u},\boldsymbol{v})}{\partial \boldsymbol{u}_j}\right), \\ \boldsymbol{r}_{\boldsymbol{v}}(\boldsymbol{u},\boldsymbol{v}) &= \left(\frac{\partial \boldsymbol{r}_i(\boldsymbol{u},\boldsymbol{v})}{\partial \boldsymbol{v}_j}\right), \\ \boldsymbol{y}_{\boldsymbol{s}}(b;\boldsymbol{s}) &= \left(\frac{\partial \boldsymbol{y}_i(b;\boldsymbol{s})}{\partial \boldsymbol{s}_j}\right) = \boldsymbol{Z}(b;\boldsymbol{s}).\end{aligned}$$

$\boldsymbol{Z}(b;\boldsymbol{s})$ ergibt sich für festes $\boldsymbol{s}$ wieder als Lösung eines zusätzlichen AWP

$$\boldsymbol{Z}'(x;\boldsymbol{s}) = \boldsymbol{f}_{\boldsymbol{y}}(x, \boldsymbol{y}(x;\boldsymbol{s})), \quad \boldsymbol{Z}(a;\boldsymbol{s}) = \boldsymbol{I}.$$

Dies ist ein Differentialgleichungssystem der Dimension n^2! Das Lösen dieses komplizierten Differentialgleichungssystems wird umgangen, indem man $\boldsymbol{F}'$ durch Differenzenquotienten approximiert

$$\boldsymbol{F}'(\boldsymbol{s}) \approx \boldsymbol{D}_{\Delta \boldsymbol{s}}\boldsymbol{F}(\boldsymbol{s}) = (\boldsymbol{D}_{\Delta s_1}\boldsymbol{F}(\boldsymbol{s}), \boldsymbol{D}_{\Delta s_2}\boldsymbol{F}(\boldsymbol{s}), \ldots, \boldsymbol{D}_{\Delta s_n}\boldsymbol{F}(\boldsymbol{s}))$$

mit

$$\boldsymbol{D}_{\Delta s_i}\boldsymbol{F}(\boldsymbol{s}) = \frac{\boldsymbol{F}(s_1, \ldots, s_{i-1}, s_i + \Delta s_i, s_{i+1}, \ldots, s_n) - \boldsymbol{F}(\boldsymbol{s})}{\Delta s_i}, \quad i = 1, \ldots, n.$$

Zum Berechnen von Approximationen der partiellen Ableitungen ist jeweils wieder ein zusätzliches AWP zu lösen. Wir erhalten damit das folgende Verfahren.

7.7. Quasi-Newton-Verfahren zum Lösen von allgemeinen RWP:

S0 *Wähle* $\boldsymbol{s}^{(0)}$ *und setze* $k = 0$.

S1 *Berechne die Lösung* $\boldsymbol{y}(b; \boldsymbol{s}^{(k)})$ *des AWP*

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{y}(a) = \boldsymbol{s}^{(k)}$$

an der Stelle $x = b$.

S2 *Wähle ein*

$$\Delta \boldsymbol{s}^{(k)} = (\Delta s_1^{(k)}, \ldots, \Delta s_n^{(k)})^T$$

und berechne für $i = 1, \ldots, n$ *die Lösungen*

$$\boldsymbol{y}^{(i)} = \boldsymbol{y}(b; \boldsymbol{s}^{(k)} + \Delta s_i^{(k)} \boldsymbol{e}_i)$$

der AWP

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{y}(a) = \boldsymbol{s}^{(k)} + \Delta s_i^{(k)} \boldsymbol{e}_i$$

an der Stelle $x = b$.

S3 *Berechne*

$$\boldsymbol{F}\left(\boldsymbol{s}^{(k)}\right) = \boldsymbol{r}\left(\boldsymbol{s}^{(k)}, \boldsymbol{y}\left(b; \boldsymbol{s}^{(k)}\right)\right)$$

und für $i = 1, \ldots, n$

$$\boldsymbol{F}\left(\boldsymbol{s}^{(k)} + \Delta s_i^{(k)} \boldsymbol{e}_i\right) = \boldsymbol{r}\left(\boldsymbol{s}^{(k)} + \Delta s_i^{(k)} \boldsymbol{e}_i, \boldsymbol{y}^{(i)}\right).$$

S4 *Berechne für* $i = 1, \ldots, n$

$$\boldsymbol{D}_{\Delta s_i^{(k)}} \boldsymbol{F}(\boldsymbol{s}^{(k)}) = \frac{\boldsymbol{F}(\boldsymbol{s}^{(k)} + \Delta s_i^{(k)} \boldsymbol{e}_i) - \boldsymbol{F}(\boldsymbol{s}^{(k)})}{\Delta s_i^{(k)}}.$$

S5 *Setze*

$$\boldsymbol{D}_{\Delta \boldsymbol{s}^{(k)}} \boldsymbol{F}(\boldsymbol{s}^{(k)}) = \left(\boldsymbol{D}_{\Delta s_1^{(k)}} \boldsymbol{F}(\boldsymbol{s}^{(k)}), \ldots, \boldsymbol{D}_{\Delta s_n^{(k)}} \boldsymbol{F}(\boldsymbol{s}^{(k)})\right).$$

S6 *Löse das lineare Gleichungssystem*

$$\boldsymbol{D}_{\Delta \boldsymbol{s}^{(k)}} \boldsymbol{F}(\boldsymbol{s}^{(k)}) \boldsymbol{d}^{(k)} = \boldsymbol{F}(s^{(k)}).$$

S7 *Setze* $\boldsymbol{s}^{(k+1)} = \boldsymbol{s}^{(k)} - \boldsymbol{d}^{(k)}$.

S8 *Setze* $k = k+1$ *und gehe zu Schritt* **S1**.

Pro Schritt sind hier $n+1$ AWP zu lösen. Die Konvergenz des Verfahrens ist lokal und hängt von der Wahl des Startwertes $\boldsymbol{s}^{(0)}$ und von der Wahl der Größen $\Delta s_i^{(k)}$, $i = 1, \ldots, n$, ab. Praktisch wendet man besser ein modifiziertes NEWTON-Verfahren an, bei dem Schritt **S6** durch

S6' Wähle eine Schrittweite λ_k und setze

$$\boldsymbol{s}^{(k+1)} = \boldsymbol{s}^{(k)} - \lambda_k \boldsymbol{d}^{(k)}.$$

ersetzt wird. Durch geschickte Wahl der Schrittweiten λ_k erreicht man eine bessere Konvergenz.
Für lineare RWP

$$\boldsymbol{y}' = \boldsymbol{T}(x)\boldsymbol{y} + \boldsymbol{g}(x), \quad \boldsymbol{A}\boldsymbol{y}(a) + \boldsymbol{B}\boldsymbol{y}(b) = \boldsymbol{c}$$

lässt sich zeigen, dass das obige Verfahren bei beliebigem Anfangswert $\boldsymbol{s}^{(0)}$ in einem Schritt die Lösung liefert.

Probleme beim einfachen Schießverfahren

Zum Lösen des RWP

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{r}(\boldsymbol{y}(a), \boldsymbol{y}(b)) = \boldsymbol{o}$$

wird beim einfachen Schießverfahren ein Wert $\boldsymbol{s}$ so bestimmt, dass die Lösung des AWP

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{y}(a) = \boldsymbol{s}$$

auch die Randbedingungen erfüllt. Eigentlich interessieren uns aber die Werte der Lösung $\boldsymbol{y}(x)$ des RWP im vorgegebenen Intervall $[a, b]$. Prinzipiell sind Werte dieser Lösung berechnenbar, indem das AWP

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{y}(a) = \boldsymbol{s}$$

näherungsweise für verschiedenen $x \in [a, b]$ gelöst wird. Hierbei tritt aber ein Problem auf. Nach Satz 6.3 hängt die Lösung eines AWP gemäß

$$\|\boldsymbol{y}(x; \boldsymbol{s}_1) - \boldsymbol{y}(x; \boldsymbol{s}_2)\| \leqq \|\boldsymbol{s}_1 - \boldsymbol{s}_2\| e^{L|x-a|}$$

von den Anfangswerten ab. Das bedeutet aber, dass auch bei genauer Bestimmung des Anfangswertes $\boldsymbol{s}$, für den die Lösung des entsprechenden AWP auch die Randbedingung erfüllt, die Lösung $\boldsymbol{y}(x;\boldsymbol{s})$ beliebig stark von der exakten Lösung des RWP abweicht (falls nur der Term $e^{L|x-a|}$ hinreichend groß ist). Das einfache Berechnen eines geeigneten Startwertes $\boldsymbol{s}$ genügt daher nicht, um die Lösung eines RWP mit hinreichender Genauigkeit zu berechnen.

7.8. Beispiel: Wir betrachten das RWP

$$\begin{array}{rclcrcl} y_1' &=& y_2 & , & y_1(0) &=& 1, \\ y_2' &=& 110y_1 + y_2 & , & y_1(10) &=& 1 \end{array}$$

und dazu das AWP

$$\begin{array}{rclcrcl} y_1' &=& y_2 & , & y_1(0) &=& 1, \\ y_2' &=& 110y_1 + y_2 & , & y_2(0) &=& s. \end{array}$$

Die allgemeine Lösung dieses AWP lautet

$$\begin{aligned} y_1(x;s) &= \frac{11-s}{21}e^{-10x} + \frac{10+s}{21}e^{11x}, \\ y_2(x;s) &= -10\frac{11-s}{21}e^{-10x} + 11\frac{10+s}{21}e^{11x}. \end{aligned}$$

Beim einfachen Schießverfahren müßte der Parameter s die Gleichung $y_1(10;s) = 1$ erfüllen. Es ergibt sich

$$\frac{11-s}{21}e^{-100} + \frac{10+s}{21}e^{110} = 1$$

und daraus

$$s = -10 + 21\frac{1-e^{-100}}{e^{110}-e^{-100}} \approx -10 + 3.5 \cdot 10^{-47} \approx -10.$$

Beim Lösen dieses Problems auf einem realen Rechner darf man höchstens erwarten, dass s mit relativer Maschinengenauigkeit berechnet wird. Man erhält eine Näherung $\overline{s} = -10(1+\varepsilon)$ mit $|\varepsilon| \leqq$ eps. Schon für einen Fehler $\varepsilon = -10^{-10}$ erhält man aber

$$y_1(10;\overline{s}) \approx 2.8 \cdot 10^{37}.$$

Selbst beim Berechnen von s mit voller Maschinengenauigkeit weichen die damit berechneten Lösungen des AWP stark von der exakten Lösung des RWP ab. ♡

Aus der Abschätzung

$$\|\boldsymbol{y}(x;\boldsymbol{s}_1) - \boldsymbol{y}(x;\boldsymbol{s}_2)\| \leqq \|\boldsymbol{s}_1 - \boldsymbol{s}_2\| e^{L|x-a|}$$

folgt aber auch, dass man den Einfluss der ungenauen Anfangsbedingungen durch Verkleinern des Intervalls $[a,x]$ beliebig verkleinert. Das führt auf die Mehrzielmethode.

7.3. Die Mehrzielmethode

Zum Lösen des RWP

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{r}(\boldsymbol{y}(a), \boldsymbol{y}(b)) = \boldsymbol{o}$$

gehen wir folgendermaßen vor:
Es sei $\Delta_n : \{x_1, x_2, \ldots, x_m\}$ mit $a = x_1 < x_2 < \cdots < x_m = b$ eine Unterteilung des Intervalls $[a, b]$. Wir wählen zu jedem x_i, $i = 1, \ldots, m-1$, einen Anfangswert $\boldsymbol{s}_i$ und lösen die AWP

$$\boldsymbol{y}' = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{y}(x_i) = \boldsymbol{s}_i, \quad i = 1, \ldots, m-1.$$

Es sei $\boldsymbol{y}(x; x_i, \boldsymbol{s}_i)$ die Lösungsfunktion, die wir für das Intervall $[x_i, x_{i+1})$ erhalten. Sind wir nun in der Lage, die Anfangswerte so zu wählen, dass die aus den Funktionen $\boldsymbol{y}(x; x_i, \boldsymbol{s}_i)$ stückweise zusammengesetzte Funktion stetig ist und noch die Randbedingungen erfüllt, so akzeptieren wir diese Funktion als Lösungsfunktion.
Es sei also $\boldsymbol{s} = (\boldsymbol{s}_1^T, \ldots, \boldsymbol{s}_m^T)^T$ und

$$\boldsymbol{y}(x; \boldsymbol{s}) = \begin{cases} \boldsymbol{y}(x; x_i, \boldsymbol{s}_i) & \text{für} \quad x \in [x_i, x_{i+1}), \quad i = 1, \ldots, m-1 \\ \boldsymbol{s}_m & \text{für} \quad x = x_m = b. \end{cases}$$

Die Stetigkeitsforderung an $\boldsymbol{y}(x; \boldsymbol{s})$ und die zu erfüllenden Randbedingungen führen auf ein nichtlineares Gleichungssystem für die Parameter $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_m$:

$$\boldsymbol{F}(\boldsymbol{s}) = \begin{pmatrix} \boldsymbol{F}_1(\boldsymbol{s}_1, \boldsymbol{s}_2) \\ \boldsymbol{F}_2(\boldsymbol{s}_2, \boldsymbol{s}_3) \\ \vdots \\ \boldsymbol{F}_{m-1}(\boldsymbol{s}_{m-1}, \boldsymbol{s}_m) \\ \boldsymbol{F}_m(\boldsymbol{s}_1, \boldsymbol{s}_m) \end{pmatrix} = \begin{pmatrix} \boldsymbol{y}(x_2; x_1, \boldsymbol{s}_1) - \boldsymbol{s}_2 \\ \boldsymbol{y}(x_3; x_2 \boldsymbol{s}_2) - \boldsymbol{s}_3 \\ \vdots \\ \boldsymbol{y}(x_m; x_{m-1}, \boldsymbol{s}_{m-1}) - \boldsymbol{s}_m \\ \boldsymbol{r}(\boldsymbol{s}_1, \boldsymbol{s}_m) \end{pmatrix} = \boldsymbol{o}.$$

Zum Lösen wird wieder ein NEWTON-Verfahren der Form

$$\boldsymbol{s}^{(k+1)} = \boldsymbol{s}^{(k)} - \left[\boldsymbol{F}'(\boldsymbol{s}^{(k)})\right]^{-1} \boldsymbol{F}(\boldsymbol{s}^{(k)})$$

oder ein modifiziertes NEWTON-Verfahren

$$\boldsymbol{s}^{(k+1)} = \boldsymbol{s}^{(k)} - \lambda_k \left[\boldsymbol{F}'(\boldsymbol{s}^{(k)})\right]^{-1} \boldsymbol{F}(\boldsymbol{s}^{(k)})$$

angewendet. Man beachte, dass die $\boldsymbol{s}_i$ Vektoren aus dem $\mathbb{R}^n$ sind. Die Gleichung $\boldsymbol{F}(\boldsymbol{s}) = \boldsymbol{o}$ stellt damit ein nichtlineares Gleichungssystem mit $n \cdot m$ Gleichungen für

$n \cdot m$ Variable dar. Für die JACOBI-Matrix $\boldsymbol{F}'(\boldsymbol{s})$ erhält man

$$\boldsymbol{F}'(\boldsymbol{s}) = \begin{pmatrix} \boldsymbol{G}_1 & -\boldsymbol{I} & \boldsymbol{O} & \cdots & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{G}_2 & -\boldsymbol{I} & \ddots & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{G}_3 & \ddots & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \ddots & \boldsymbol{G}_{m-2} & -\boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \cdots & \boldsymbol{O} & \boldsymbol{G}_{m-1} & -\boldsymbol{I} \\ \boldsymbol{A} & \boldsymbol{O} & \boldsymbol{O} & \cdots & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{B} \end{pmatrix}$$

mit

$$\boldsymbol{G}_i = \left(\frac{\partial \boldsymbol{F}_i(\boldsymbol{s})}{\partial \boldsymbol{s}_i} \right), \quad \boldsymbol{A} = \left(\frac{\partial \boldsymbol{r}(\boldsymbol{s}_1, \boldsymbol{s}_m)}{\partial \boldsymbol{s}_1} \right), \quad \boldsymbol{B} = \left(\frac{\partial \boldsymbol{r}(\boldsymbol{s}_1, \boldsymbol{s}_m)}{\partial \boldsymbol{s}_m} \right).$$

Man ersetzt wieder die Differentialquotienten durch entsprechende Differenzenquotienten. Zu deren Berechnung ist wieder das Lösen von weiteren $(m-1)n$ AWP notwendig.
Die Berechnung von

$$\left[\boldsymbol{F}'(\boldsymbol{s}^{(k)}) \right]^{-1} \boldsymbol{F}(\boldsymbol{s}^{(k)}),$$

das bedeutet das Lösen des linearen Gleichungssystems

$$\boldsymbol{F}'(\boldsymbol{s}^{(k)})\boldsymbol{d} = \boldsymbol{F}(\boldsymbol{s}^{(k)}),$$

lässt sich wegen der speziellen Struktur der JACOBI-Matrix auf die Behandlung von mehreren linearen Gleichungssystemen geringerer Dimension zurückführen.
Die Voraussetzungen an die Durchführbarkeit der Mehrzielmethode sind bedeutend schwächer als die Voraussetzungen an die Durchführbarkeit des einfachen Schießverfahrens. Besonders die Güte der Startwerte spielt bei der Mehrzielmethode eine geringere Rolle als beim einfachen Schießverfahren.

7.4. Differenzenverfahren

Eine weitere Möglichkeit Verfahren zum Lösen von RWP zu gewinnen, besteht darin, die in der Differentialgleichung auftretenden Differentialquotienten durch Differenzenquotienten zu ersetzen. Wir erläutern das Vorgehen am Beispiel des folgenden linearen RWP 2.Ordnung mit linearen separierten Randbedingungen.

$$\begin{aligned} -y'' + p(x)y' + q(x)y &= g(x), \quad x \in [a,b], \\ \alpha_1 y(a) + \alpha_2 y'(a) &= \alpha_0, \\ \beta_1 y(b) + \beta_2 y'(b) &= \beta_0. \end{aligned}$$

Weiterhin nehmen wir an, dass $y \in C^4[a,b]$ gilt.
Nun ersetzen wir in geeigneter Weise y'' und y' durch Differenzenquotienten. Dazu verwenden wir eine äquidistante Unterteilung des Intervalls $[a,b]$:

$$x_i = a + ih, \quad h = \frac{b-a}{N}.$$

Wir ersetzen $y''(x_i)$ durch

$$\frac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1})}{h^2}, \quad i = 1, \ldots, N-1$$

und $y'(x_i)$ durch

$$\frac{y(x_{i+1}) - y(x_{i-1})}{2h}, \quad i = 1, \ldots, N-1.$$

Durch TAYLOR-Entwicklung von y bestätigt man leicht, dass

$$y''(x_i) = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + O(h^2)$$

und

$$y'(x_i) = \frac{y_{i+1} - y_{i-1}}{2h} + O(h^2)$$

gilt. (Dabei ist $y_i = y(x_i)$ für $i = 0, 1, \ldots, N$.) Setzen wir dies in die Differentialgleichung ein, so erhalten wir

$$-\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + p(x_i)\frac{y_{i+1} - y_{i-1}}{2h} + q(x_i)y_i = g(x_i) + O(h^2),\ i = 1, \ldots, N-1.$$

In den Randbedingungen dürfen wir die ersten Ableitungen nicht durch zentrale Differenzenquotienten ersetzen. Wir verwenden hier einfache Differenzenquotienten

$$\begin{aligned} y'(x_0) &= y'(a) = \frac{y_1 - y_0}{h} + O(h), \\ y'(x_N) &= y'(b) = \frac{y_N - y_{N-1}}{h} + O(h). \end{aligned}$$

Damit ergibt sich

$$\begin{aligned} \alpha_1 y_0 + \alpha_2 \frac{y_1 - y_0}{h} &= \alpha_0 + O(h), \\ \beta_1 y_N + \beta_2 \frac{y_N - y_{N-1}}{h} &= \beta_0 + O(h). \end{aligned}$$

Ersetzen wir nun die y_i durch die Näherungen η_i und lassen die Restglieder in den Gleichungen fort, so ergibt sich zur Bestimmung der η_i das folgende lineare Gleichungssystem (mit $p_i = p(x_i)$, $q_i = q(x_i)$, $g_i = g(x_i)$).

$$\begin{aligned}
(-\alpha_2 + h\alpha_1)\eta_0 + \alpha_2\eta_1 &= h\alpha_0, \\
\left(-1 - \frac{h}{2}p_i\right)\eta_{i-1} + \left(2 + h^2 q_i\right)\eta_i + \left(-1 + \frac{h}{2}p_i\right)\eta_{i+1} &= h^2 g_i,\ i = 1, \ldots, N-1, \\
-\beta_2\eta_{N-1} + (\beta_2 + h\beta_1)\eta_N &= h\beta_0.
\end{aligned}$$

In Matrixschreibweise erhalten wir $\boldsymbol{A}(h)\boldsymbol{\eta} = \boldsymbol{b}(h)$ mit

$$\boldsymbol{A}(h) = \begin{pmatrix}
-\alpha_2 + h\alpha_1 & \alpha_2 & 0 & \cdots & 0 & 0 \\
-1 - \frac{h}{2}p_1 & 2 + h^2 q_1 & -1 + \frac{h}{2}p_1 & \ddots & 0 & 0 \\
0 & -1 - \frac{h}{2}p_2 & 2 + h^2 q_2 & \ddots & 0 & 0 \\
\vdots & & \ddots & \ddots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 2 + h^2 q_{N-1} & -1 + \frac{h}{2}p_{N-1} \\
0 & 0 & 0 & \cdots & -\beta_2 & \beta_2 + h\beta_1
\end{pmatrix}$$

und

$$\boldsymbol{\eta} = \begin{pmatrix} \eta_0 \\ \eta_1 \\ \vdots \\ \eta_{N-1} \\ \eta_N \end{pmatrix}, \quad \boldsymbol{b}(h) = \begin{pmatrix} h\alpha_0 \\ h^2 g_1 \\ \vdots \\ h^2 g_{N-1} \\ h\beta_0 \end{pmatrix}.$$

Wir erhalten genau dann eine eindeutige Lösung $\boldsymbol{\eta}$, wenn $\boldsymbol{A}(h)$ regulär ist. Die Regularität von $\boldsymbol{A}(h)$ ist natürlich in jedem Falle gesondert nachzuprüfen. Für bestimmte Parameterkombinationen lassen sich Aussagen über die Regularität von $\boldsymbol{A}(h)$ machen. Es gilt der folgende Satz (ohne Beweis).

7.9. Satz: *Für das RWP*

$$\begin{aligned}
-y'' + p(x)y' + q(x)y &= g(x), \quad x \in [a,b], \\
\alpha_1 y(a) + \alpha_2 y'(a) &= \alpha_0, \\
\beta_1 y(b) + \beta_2 y'(b) &= \beta_0
\end{aligned}$$

gelte

1. $\alpha_1 > 0$, $\alpha_2 \geqq 0$, $\beta_1 > 0$ *und* $\beta_2 \geqq 0$,
2. $q(x) \geqq 0$ *für alle* $x \in [a,b]$.

Dann ist die Matrix $\boldsymbol{A}(h)$ regulär für alle Schrittweiten h mit

$$0 < h < h_0 = \frac{2}{\sup_{x\in[a,b]} p(x)}.$$

Die Matrix $\boldsymbol{A}(h)$ wird für kleines h (großes N) fast singulär sein, so dass das numerische Lösen des Gleichungssystems $\boldsymbol{A}(h)\boldsymbol{\eta} = \boldsymbol{b}(h)$ schwierig ist.
Für die Güte der Näherungen lässt sich im Falle $\alpha_2 = \beta_2 = 0$ und $p(x) \equiv 0$ die folgende Aussage machen.

7.10. Satz: *Das RWP*

$$\begin{aligned} -y'' + q(x)y &= g(x), \quad x \in [a,b], \\ y(a) &= \alpha_0, \\ y(b) &= \beta_0 \end{aligned}$$

mit $q(x) \geqq 0$ für alle $x \in [a,b]$ habe eine Lösung $y \in C^4[a,b]$. Es seien $|y^{(4)}| \leqq M$ für alle $x \in [a,b]$ und $\boldsymbol{\eta}$ die mit dem Differenzenverfahren erzeugte Näherungslösung. Dann gilt

$$|y(x_i) - \eta_i| \leqq \frac{Mh^2}{24}(x_i - a)(b - x_i).$$

Für dieses spezielle RWP erhalten wir so ein Verfahren 2. Ordnung.
Das obige Differenzenverfahren ist auch zur Behandlung von nichtlinearen RWP anwendbar. Dann erhält man jedoch zum Berechnen der Näherungen $\eta_0, \ldots, \eta_N$ nichtlineare Gleichungssysteme, die selbst mit Hilfe von Iterationsverfahren nur näherungsweise lösbar sind. Um mit Differenzenverfahren hohe Genauigkeiten zu erzielen, ist die Schrittweite h klein zu wählen. Das führt aber auf große Gleichungssysteme, die eventuell noch schlecht konditioniert sind. Differenzenverfahren spielen deshalb nur bei geringen Genauigkeitsanforderungen eine Rolle. Bei hohen Genauigkeitsanforderungen sollte man besser die Mehrzielmethode anwenden.

7.5. Aufgaben

1. Man löse das Randwertproblem

   $$y'' - xy' + y = 1 \qquad x \in [a,b] \qquad y(0) = y'(1) = 1$$

   numerisch mit einem Differenzenverfahren. Als Schrittweite wähle man $h = 0,2$. Anschließend bestimme man die exakte Lösung und vergleiche mit den erhaltenen Näherungen.

Kapitel 8

Lineare Gleichungssysteme

8.1. Allgemeine Grundlagen und Störungstheorie

Wir betrachten in diesem Kapitel lineare Gleichungssysteme

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2, \\ \vdots & \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n, \end{aligned}$$

oder in Matrixschreibweise

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

mit

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}, \quad \boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n.$$

Bevor wir uns mit der Existenz und Eindeutigkeit von Lösungen und Genauigkeitsfragen beschäftigen, sind noch einige Begriffe bereitzustellen.

8.1.1. Vektor- und Matrixnormen

Eine Abbildung

$$\| \circ \| : \mathbb{R}^n \longrightarrow \mathbb{R}_+ = \left\{ \alpha \in \mathbb{R} \mid \alpha \geqq 0 \right\}$$

heißt **Vektornorm**, falls sie folgende Bedingungen erfüllt.

1. $\|\boldsymbol{x}\| \geqq 0, \quad \|\boldsymbol{x}\| = 0 \Longleftrightarrow \boldsymbol{x} = \boldsymbol{o}$.
2. Für alle $\alpha \in \mathbb{R}$ und alle $\boldsymbol{x} \in \mathbb{R}^n$ gilt $\|\alpha\boldsymbol{x}\| = |\alpha|\,\|\boldsymbol{x}\|$.
3. Für alle $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ gilt die Dreiecksungleichung $\|\boldsymbol{x}+\boldsymbol{y}\| \leqq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$.

Bemerkung: Die Dreiecksungleichung ist äquivalent zur Ungleichung

$$\|\boldsymbol{x}-\boldsymbol{y}\| \geqq \big|\|\boldsymbol{x}\| - \|\boldsymbol{y}\|\big|\,.$$

Beispiele für Vektornormen

- Euklidische Norm:

$$\|\boldsymbol{x}\|_2 = \sqrt{\boldsymbol{x}^T\boldsymbol{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

- Betragssummennorm:

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|.$$

- Maximumnorm:

$$\|\boldsymbol{x}\|_\infty = \max_{i=1,\ldots,n} |x_i|.$$

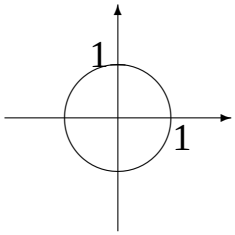
- p-Norm:

$$\|\boldsymbol{x}\|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}.$$

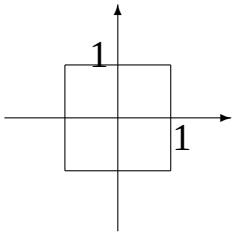
In den folgenden Bildern sind die sogenannten Einheitssphären, die Mengen

$$\{\, \boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\| = 1 \,\},$$

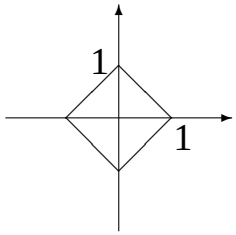
im $\mathbb{R}^2$ dargestellt.



$\|\boldsymbol{x}\|_2 = 1$



$\|\boldsymbol{x}\|_\infty = 1$



$\|\boldsymbol{x}\|_1 = 1$

Zwischen den Vektornormen gibt es gewisse Beziehungen. Die wichtigste kommt in folgendem Satz zum Ausdruck.

8.1. Satz: *Alle Vektornormen im $\mathbb{R}^n$ sind in folgendem Sinne äquivalent:*
Für je zwei Normen $\|\circ\|^{(1)}$ und $\|\circ\|^{(2)}$ gibt es Konstanten $M \geqq m > 0$, so dass für alle $\boldsymbol{x} \in \mathbb{R}^n$

$$m\|\boldsymbol{x}\|^{(2)} \leqq \|\boldsymbol{x}\|^{(1)} \leqq M\|\boldsymbol{x}\|^{(2)}$$

gilt.

So gilt zum Beispiel

$$\begin{array}{rcccl} \|\boldsymbol{x}\|_\infty & \leqq & \|\boldsymbol{x}\|_2 & \leqq & \sqrt{n}\|\boldsymbol{x}\|_\infty, \\ \frac{1}{\sqrt{n}}\|\boldsymbol{x}\|_1 & \leqq & \|\boldsymbol{x}\|_2 & \leqq & \|\boldsymbol{x}\|_1, \\ \|\boldsymbol{x}\|_\infty & \leqq & \|\boldsymbol{x}\|_1 & \leqq & n\|\boldsymbol{x}\|_\infty. \end{array}$$

Auch für Matrizen lassen sich Normen einführen. Eine Abbildung

$$\|\circ\| : \mathbb{R}^{m\times n} \longrightarrow \mathbb{R}_+$$

heißt (m,n)-**Matrixnorm**, falls sie folgende Bedingungen erfüllt:

1. $\|\boldsymbol{A}\| = 0 \Longleftrightarrow \boldsymbol{A} = \boldsymbol{O}$.
2. Für alle $\alpha \in \mathbb{R}$ und alle (m,n)-Matrizen $\boldsymbol{A}$ gilt
$$\|\alpha\boldsymbol{A}\| = |\alpha|\,\|\boldsymbol{A}\|.$$
3. Für alle (m,n)-Matrizen $\boldsymbol{A}, \boldsymbol{B}$ gilt die Dreiecksungleichung
$$\|\boldsymbol{A}+\boldsymbol{B}\| \leqq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|.$$

Bemerkung: Die Dreiecksungleichung lässt sich wieder durch

$$\|\boldsymbol{A}-\boldsymbol{B}\| \geqq \big|\|\boldsymbol{A}\| - \|\boldsymbol{B}\|\big|$$

ersetzen.
Bei Fehlerabschätzungen für lineare Gleichungssysteme treten oft Matrix- und Vektornormen gemeinsam auf. Hier ist darauf zu achten, dass die verwendeten Normen in gewisser Weise verträglich sind. Diese Eigenschaft wird in der folgenden Definition genauer beschrieben.

Eine (m,n)-Matrixnorm $\|\circ\|$ heißt mit den Vektornormen $\|\circ\|^{(m)}$ auf dem $\mathbb{R}^m$ und $\|\circ\|^{(n)}$ auf dem $\mathbb{R}^n$ **verträglich**, falls für alle $\boldsymbol{x} \in \mathbb{R}^n$ und alle (m,n)-Matrizen $\boldsymbol{A}$

$$\|\boldsymbol{A}\boldsymbol{x}\|^{(m)} \leqq \|\boldsymbol{A}\| \cdot \|\boldsymbol{x}\|^{(n)}$$

gilt.
Die (n,n)-Matrixnorm $\|\circ\|$ heißt **submultiplikativ**, falls für alle (n,n)-Matrizen $\boldsymbol{A}, \boldsymbol{B}$

$$\|\boldsymbol{A}\boldsymbol{B}\| \leqq \|\boldsymbol{A}\| \cdot \|\boldsymbol{B}\|$$

gilt. Auch für Matrixnormen gilt ein zu Satz 8.1 analoger Satz.

8.2. Satz: *Alle (m,n)-Matrixnormen sind in folgendem Sinne äquivalent:*
Für je zwei Normen $\|\circ\|^{(1)}$ und $\|\circ\|^{(2)}$ gibt es Konstanten $L \geqq l > 0$, so dass für alle (m,n)-Matrizen $\boldsymbol{A}$

$$l\|\boldsymbol{A}\|^{(2)} \leqq \|\boldsymbol{A}\|^{(1)} \leqq L\|\boldsymbol{A}\|^{(2)}$$

gilt.

Beispiele für Matrixnormen

- Zeilensummennorm:

$$\|\boldsymbol{A}\|_\infty = \max_{i=1,\ldots,m} \sum_{k=1}^{n} |a_{ik}|.$$

- Spaltensummennorm:

$$\|\boldsymbol{A}\|_1 = \max_{j=1,\ldots,n} \sum_{k=1}^{m} |a_{kj}|.$$

- FROBENIUS-Norm:

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}.$$

- Maximumnorm:

$$\|\boldsymbol{A}\|_M = \max_{i=1,\ldots,m} \max_{j=1,\ldots,n} |a_{ij}|.$$

Die Zeilensummennorm, die Spaltensummennorm und die FROBENIUS-Norm sind für quadratische Matrizen submultiplikativ. Die Zeilensummennorm ist mit der Maximumnorm für Vektoren verträglich, die Spaltensummennorm ist mit der Betragsnorm für Vektoren verträglich, und die FROBENIUS-Norm ist mit der euklidischen Vektornorm verträglich.
Unter allen (n,n)-Matrixnormen, die mit einer gegebenen Vektornorm $\|\circ\|$ auf dem $\mathbb{R}^n$ verträglich sind, gibt es genau eine, die dieser Vektornorm besonders gut angepasst ist. Die Matrixnorm

$$\text{lub}(\boldsymbol{A}) = \max_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ \boldsymbol{x}\neq\boldsymbol{o}}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \boldsymbol{a}x \max_{\substack{\boldsymbol{y}\in\mathbb{R}^n \\ \|\boldsymbol{y}\|=1}} \|\boldsymbol{A}\boldsymbol{y}\|$$

heißt **Matrixgrenznorm** [1] oder kurz Grenznorm zur Vektornorm $\|\circ\|$. Die besondere Eigenschaft von Grenznormen kommt in folgendem Satz zum Ausdruck.

8.3. Satz: *Es sei $\|\circ\|$ eine beliebige (n,n)-Matrixnorm, die mit der Vektornorm $\|\circ\|$ auf dem $\mathbb{R}^n$ verträglich ist. Mit der zugehörigen Matrixgrenznorm gilt dann für alle (n,n)-Matrizen $\boldsymbol{A}$*

$$\text{lub}(\boldsymbol{A}) \leqq \|\boldsymbol{A}\|.$$

Der Beweis folgt sofort aus der Definition der Matrixgrenznorm.
Weiterhin ist jede Matrixgrenznorm submultiplikativ und es gilt $\text{lub}(\boldsymbol{I}) = 1$.

Geometrische Interpretation der Matrixgrenznorm

Es sei $\boldsymbol{\varphi}$ die durch die Matrix $\boldsymbol{A}$ vermittelte lineare Abbildung des $\mathbb{R}^n$ in sich. Wir betrachten nun das Bild einer Einheitssphäre:

$$\boldsymbol{\Omega} = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\| = 1 \,\}$$

und

$$\boldsymbol{\varphi}(\boldsymbol{\Omega}) = \{\, \boldsymbol{A}\boldsymbol{x} \mid \boldsymbol{x} \in \boldsymbol{\Omega} \,\}.$$

Dann folgt

$$\text{lub}(\boldsymbol{A}) = \max_{\boldsymbol{y}\in\boldsymbol{\varphi}(\boldsymbol{\Omega})} \{\|\boldsymbol{y}\|\}.$$

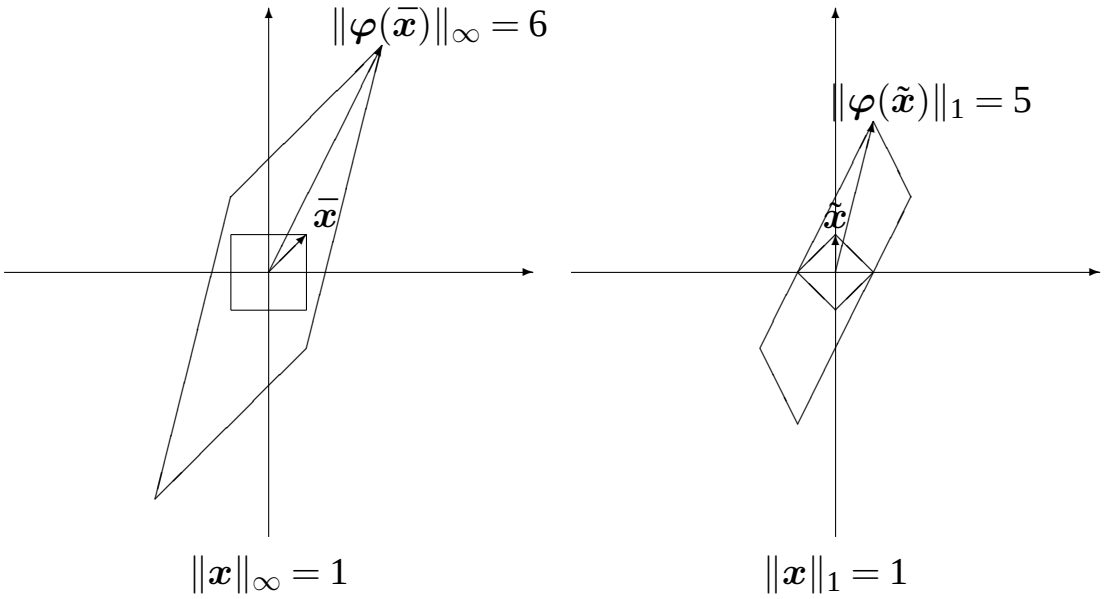
Die Größe $\text{lub}(\boldsymbol{A})$ stellt die größte „Verzerrung“ eines Bildpunktes gegenüber dem Originalpunkt dar.

[1] Die Bezeichnung lub ist die Abkürzung von **l**owest **u**pper **b**ound.

8.4. Beispiel: Wir betrachten die Matrix

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 \\ 4 & 2 \end{pmatrix}.$$

Im Bild sind die Einheitssphären und ihre Bilder für die Maximumnorm und die Betragssummennorm dargestellt.



♡

Zu den wichtigsten Vektornormen erhält man die folgenden Grenznormen.

- Betragssummennorm:

$$\|\boldsymbol{A}\|_1 = \mathrm{lub}_1(\boldsymbol{A}) = \max_{\boldsymbol{x}\neq\boldsymbol{o}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_1} = \max_{j=1,\ldots,n} \sum_{k=1}^{n} |a_{kj}|.$$

  Das ist die Spaltensummennorm.

- Maximumnorm:

$$\|\boldsymbol{A}\|_\infty = \mathrm{lub}_\infty(\boldsymbol{A}) = \max_{\boldsymbol{x}\neq\boldsymbol{o}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} = \max_{i=1,\ldots,n} \sum_{k=1}^{n} |a_{ik}|.$$

  Das ist die Zeilensummennorm.

- Euklidische Norm:

$$\|\boldsymbol{A}\|_2 = \mathrm{lub}_2(\boldsymbol{A}) = \max_{\boldsymbol{x}\neq\boldsymbol{o}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = \max_{\boldsymbol{x}\neq\boldsymbol{o}} \frac{\sqrt{(\boldsymbol{A}\boldsymbol{x})^T(\boldsymbol{A}\boldsymbol{x})}}{\sqrt{\boldsymbol{x}^T\boldsymbol{x}}} = \sqrt{\lambda_{max}(\boldsymbol{A}^T\boldsymbol{A})}.$$

Das ist die Spektralnorm. $\lambda_{max}(\boldsymbol{A}^T\boldsymbol{A})$ bezeichnet dabei den größten Eigenwert der Matrix $\boldsymbol{A}^T\boldsymbol{A}$. Die Spektralnorm einer Matrix darf nicht mit dem Spektralradius verwechselt werden. Dieser ist der Betrag des betragsgrößten Eigenwertes von $\boldsymbol{A}$:

$$\varrho(\boldsymbol{A}) = \max_{i=1,\ldots,n} \{|\lambda_i(\boldsymbol{A})|\}.$$

8.1.2. Ordnungen und Beträge

Die in Abschnitt 8.1.1. eingeführten Normen werden dazu verwendet, Vektoren und Matrizen in ihrer Gesamtheit miteinander zu vergleichen. Oft benötigt man aber auch elementweise Abschätzungen. Dazu führen wir eine natürliche Halbordnung für Vektoren aus dem $\mathbb{R}^n$ und für (m,n)-Matrizen ein. Für je zwei Vektoren $\boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^n$ gilt $\boldsymbol{x} \leqq \boldsymbol{y}$ genau dann, wenn $x_i \leqq y_i$ für $i = 1,\ldots,n$ gilt. Für je zwei (m,n)-Matrizen $\boldsymbol{A},\boldsymbol{B}$ gilt $\boldsymbol{A} \leqq \boldsymbol{B}$ genau dann, wenn $a_{ij} \leqq b_{ij}$ für $i = 1,\ldots,m$ und $j = 1,\ldots,n$ gilt. Die so eingeführte Relation ist reflexiv, symmetrisch und transitiv. Aber es gibt Vektoren (Matrizen) für die weder $\boldsymbol{x} \leqq \boldsymbol{y}$ ($\boldsymbol{A} \leqq \boldsymbol{B}$) noch $\boldsymbol{y} \leqq \boldsymbol{x}$ ($\boldsymbol{B} \leqq \boldsymbol{A}$) gilt. Weiterhin führen wir Betragsvektoren und Betragsmatrizen ein. Für $\boldsymbol{x} \in \mathbb{R}^n$ sei

$$|\boldsymbol{x}| = \begin{pmatrix} |x_1| \\ |x_2| \\ \vdots \\ |x_n| \end{pmatrix}.$$

Für eine (m,n)-Matrix $\boldsymbol{A}$ sei

$$|\boldsymbol{A}| = \begin{pmatrix} |a_{11}| & |a_{12}| & \ldots & |a_{1n}| \\ |a_{21}| & |a_{22}| & \ldots & |a_{2n}| \\ \vdots & \vdots & \ddots & \vdots \\ |a_{m1}| & |a_{m2}| & \ldots & |a_{mn}| \end{pmatrix}.$$

Offensichtlich gilt für alle $\boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^n$

- $|\boldsymbol{x}| = 0 \Longleftrightarrow \boldsymbol{x} = \boldsymbol{o}$.
- $|\alpha\boldsymbol{x}| = |\alpha|\,|\boldsymbol{x}|$ für alle $\alpha \in \mathbb{R}$.
- $|\boldsymbol{x}+\boldsymbol{y}| \leqq |\boldsymbol{x}| + |\boldsymbol{y}|$.

Achtung: $|\circ|$ ist trotzdem keine Norm, denn $|\circ| : \mathbb{R}^n \longrightarrow \mathbb{R}^n$.
Für beliebige (m,n)-Matrizen $\boldsymbol{A}$, $\boldsymbol{B}$ und (n,p)-Matrizen $\boldsymbol{C}$ gilt

- $|\boldsymbol{A}| = 0 \Longleftrightarrow \boldsymbol{A} = \boldsymbol{O}$.

- $|\alpha \boldsymbol{A}| = |\alpha|\,|\boldsymbol{A}|$ für alle $\alpha \in \mathbb{R}$.
- $|\boldsymbol{A} + \boldsymbol{B}| \leqq |\boldsymbol{A}| + |\boldsymbol{B}|$.
- $|\boldsymbol{A}\boldsymbol{C}| \leqq |\boldsymbol{A}|\,|\boldsymbol{C}|$.

Insbesondere gilt für alle (m,n)-Matrizen $\boldsymbol{A}$ und alle Vektoren $\boldsymbol{x} \in \mathbb{R}^n$

$$|\boldsymbol{A}\boldsymbol{x}| \leqq |\boldsymbol{A}|\,|\boldsymbol{x}|.$$

Die Ordnungsrelationen und Beträge sind mit den Vektornormen $\|\circ\|_p$ in folgendem Sinne verträglich.

(*) Für alle Vektoren $\boldsymbol{x} \in \mathbb{R}^n$ gilt $\| \, |\boldsymbol{x}| \, \|_p = \|\boldsymbol{x}\|_p$.

(**) Für alle Vektoren $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ gilt

$$|\boldsymbol{x}| \leqq |\boldsymbol{y}| \Longrightarrow \|\boldsymbol{x}\|_p \leqq \|\boldsymbol{y}\|_p.$$

Eine Norm mit der Eigenschaft (*) heißt **absolut**.
Eine Norm mit der Eigenschaft (**) heißt **monoton**. Die Vektornormen $\|\circ\|_p$ sind absolut und monoton. Bei den Matrixnormen sind die Normen $\|\boldsymbol{A}\|_1$, $\|\boldsymbol{A}\|_\infty$ und $\|\boldsymbol{A}\|_F$ absolut und monoton. Die Spektralnorm $\|\boldsymbol{A}\|_2$ ist weder absolut noch monoton. Es gilt jedoch

$$\|\boldsymbol{A}\|_2 \leqq \| \, |\boldsymbol{A}| \, \|_2$$

und

$$|\boldsymbol{A}| \leqq |\boldsymbol{B}| \Longrightarrow \| \, |\boldsymbol{A}| \, \|_2 \leqq \| \, |\boldsymbol{B}| \, \|_2.$$

Die Spektralnorm besitzt dafür (genau wie die Euklidische Vektornorm) eine andere wichtige Eigenschaft:

> Für eine beliebige (m,n)-Matrix $\boldsymbol{A}$ und beliebige orthogonale (m,m)-Matrizen $\boldsymbol{U}$ und (n,n)-Matrizen $\boldsymbol{V}$ gilt
>
> $$\|\boldsymbol{U}\boldsymbol{A}\boldsymbol{V}\|_2 = \|\boldsymbol{A}\|_2.$$

Wir werden im weiteren nur submultiplikative Matrixnormen verwenden. Im wesentlichen werden das die Normen $\|\boldsymbol{A}\|_1$, $\|\boldsymbol{A}\|_2$, $\|\boldsymbol{A}\|_\infty$ und $\|\boldsymbol{A}\|_F$ sein.

8.1.3. Spezielle Transformationsmatrizen

Bei der Entwicklung von Verfahren zum Lösen von linearen Gleichungssystemen werden wir verschiedene Transformationsmatrizen verwenden. Die wichtigsten von ihnen mit ihren speziellen Eigenschaften wollen wir in diesem Abschnitt angeben.

Nichtorthogonale Transformationen

Eine (n,n)-Matrix $\boldsymbol{M}$ det Form

$$\boldsymbol{M} = \boldsymbol{M}_k(\boldsymbol{m}) = \boldsymbol{I} + \boldsymbol{m}\boldsymbol{e}_k^T$$

mit $\boldsymbol{m} \in \mathbb{R}^n$ und $\boldsymbol{e}_k^T\boldsymbol{m} = 0$ heißt **NT-Matrix** [2]. Eine (n,n)-Matrix $\boldsymbol{L}$ der Form

$$\boldsymbol{L} = \boldsymbol{L}_k(\boldsymbol{l}) = \boldsymbol{I} + \boldsymbol{l}\boldsymbol{e}_k^T$$

mit $\boldsymbol{l} \in \mathbb{R}^n$ und $\boldsymbol{e}_i^T\boldsymbol{l} = 0$ für $i = 1,\ldots,k$ heißt **LNT-Matrix** [3]. Eine NT-Matrix hat damit folgendes Aussehen:

$$\boldsymbol{M}_k(\boldsymbol{m}) = \begin{pmatrix} 1 & & & m_1 & & \\ & \ddots & & \vdots & & \boldsymbol{O} \\ & & 1 & m_{k-1} & & \\ & & & 1 & & \\ & & & m_{k+1} & 1 & \\ & \boldsymbol{O} & & \vdots & & \ddots \\ & & & m_n & & & 1 \end{pmatrix}.$$

Eine LNT-Matrix hat dagegen das Aussehen:

$$\boldsymbol{L}_k(\boldsymbol{l}) = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \boldsymbol{O} \\ & & 1 & & & \\ & & & 1 & & \\ & & & l_{k+1} & 1 & \\ & \boldsymbol{O} & & \vdots & & \ddots \\ & & & l_n & & & 1 \end{pmatrix}.$$

Bemerkung: Eine Matrix der Form $\boldsymbol{I} + \boldsymbol{m}\boldsymbol{e}_k^T$ mit beliebigem Vektor $\boldsymbol{m} \in \mathbb{R}^n$ heißt FROBENIUS-**Matrix**.
Für NT-Matrizen und damit auch für LNT-Matrizen gilt der folgende Satz.

[2] Abkürzung für **n**ichtorthogonale **T**ransformationsmatrix
[3] Abkürzung für **l**ower triangular NT-Matrix

8.5. Satz: *Jede NT-Matrix $\boldsymbol{M}_k(\boldsymbol{m})$ ist regulär, und es gilt*

$$\boldsymbol{M}_k(\boldsymbol{m})^{-1} = \boldsymbol{M}_k(-\boldsymbol{m}).$$

Beweis:

$$\begin{aligned}\boldsymbol{M}_k(\boldsymbol{m})\boldsymbol{M}_k(-\boldsymbol{m}) &= (\boldsymbol{I}+\boldsymbol{m}\boldsymbol{e}_k^T)(\boldsymbol{I}-\boldsymbol{m}\boldsymbol{e}_k^T)\\ &= \boldsymbol{I}+\boldsymbol{m}\boldsymbol{e}_k^T-\boldsymbol{m}\boldsymbol{e}_k^T-\boldsymbol{m}\boldsymbol{e}_k^T\boldsymbol{m}\boldsymbol{e}_k^T\\ &= \boldsymbol{I}-(\boldsymbol{e}_k^T\boldsymbol{m})\boldsymbol{m}\boldsymbol{e}_k^T=\boldsymbol{I}.\end{aligned}$$

❀

Die Inverse einer NT-Matrix ist wieder eine NT-Matrix. Speziell ist die Inverse einer LNT-Matrix wieder eine LNT-Matrix.
Eine (n,n)-Matrix $\boldsymbol{L}$ mit $l_{ij}=0$ für $i<j$ heißt **untere Dreiecksmatrix**. Gilt zusätzlich $l_{ii}=1$ für $i=1,\ldots,n$, so heißt die Matrix **untere Einsdreiecksmatrix**.
Diese Matrizen haben daher folgende Struktur:

$$\boldsymbol{L}=\begin{pmatrix} l_{11} & & & & \\ l_{21} & l_{22} & & \boldsymbol{O} & \\ \vdots & \vdots & \ddots & & \\ l_{n-1,1} & l_{n-1,2} & \cdots & l_{n-1,n-1} & \\ l_{n,1} & l_{n,2} & \cdots & l_{n,n-1} & l_{nn}\end{pmatrix}$$

bzw.

$$\boldsymbol{L}=\begin{pmatrix} 1 & & & & \\ l_{21} & 1 & & \boldsymbol{O} & \\ \vdots & \vdots & \ddots & & \\ l_{n-1,1} & l_{n-1,2} & \cdots & 1 & \\ l_{n,1} & l_{n,2} & \cdots & l_{n,n-1} & 1\end{pmatrix}.$$

Eine (n,n)-Matrix $\boldsymbol{U}$ mit $u_{ij}=0$ für $i>j$ heißt **obere Dreiecksmatrix**. Gilt zusätzlich $u_{ii}=1$ für $i=1,\ldots,n$, so heißt die Matrix **obere Einsdreiecksmatrix**.
Eine obere Dreiecksmatrix hat folgende Struktur

$$\boldsymbol{U}=\begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1,n-1} & u_{1n}\\ & u_{22} & \cdots & u_{2,n-1} & u_{2n}\\ & & \ddots & \vdots & \vdots\\ & \boldsymbol{O} & & u_{n-1,n-1} & u_{n-1,n}\\ & & & & u_{nn}\end{pmatrix}.$$

Orthogonale Transformationen

Als orthogonale Transformationsmatrizen werden uns Tauschmatrizen und Permutationsmatrizen begegnen. Eine (n,n)-**Tauschmatrix** $\boldsymbol{T}_{pq}$ ist eine Matrix der Form

$$\begin{aligned}\boldsymbol{T}_{pq} &= (\boldsymbol{e}_1,\ldots,\boldsymbol{e}_{p-1},\boldsymbol{e}_q,\boldsymbol{e}_{p+1},\ldots,\boldsymbol{e}_{q-1},\boldsymbol{e}_p,\boldsymbol{e}_{q+1},\ldots,\boldsymbol{e}_n)\\ &= \begin{pmatrix}\boldsymbol{e}_1^T\\ \vdots\\ \boldsymbol{e}_{p-1}^T\\ \boldsymbol{e}_q^T\\ \boldsymbol{e}_{p+1}^T\\ \vdots\\ \boldsymbol{e}_{q-1}^T\\ \boldsymbol{e}_p^T\\ \boldsymbol{e}_{q+1}^T\\ \vdots\\ \boldsymbol{e}_n^T\end{pmatrix} = \begin{pmatrix}1 & & & & & & & & \\ & \ddots & & & & & & \boldsymbol{O} & \\ & & 1 & & & & & & \\ & & & 0 & & & 1 & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & 1 & & & & 0 & \\ & & & & & & & & 1 \\ & \boldsymbol{O} & & & & & & \ddots & \\ & & & & & & & & 1\end{pmatrix}\begin{matrix}\\ \\ \\ p\text{-te Zeile}\\ \\ \\ \\ q\text{-te Zeile}\\ \\ \\ \\ \end{matrix}.\end{aligned}$$

Multipliziert man eine (n,n)-Matrix $\boldsymbol{A}$ von links mit einer Tauschmatrix $\boldsymbol{T}_{pq}$, so erhält man eine (n,n)-Matrix $\bar{\boldsymbol{A}}$, bei der gegenüber $\boldsymbol{A}$ die p-te und q-te Zeile vertauscht sind. Multiplikation der Matrix $\boldsymbol{A}$ von rechts mit der Tauschmatrix $\boldsymbol{T}_{pq}$ bewirkt einen entsprechenden Spaltentausch. Hat die Matrix $\boldsymbol{A}$ die Zeilen $\boldsymbol{z}_1^T,\ldots,\boldsymbol{z}_n^T$ bzw. die Spalten $\boldsymbol{s}_1,\ldots,\boldsymbol{s}_n$, so ergibt sich

$$\boldsymbol{T}_{pq}\boldsymbol{A} = \begin{pmatrix}\boldsymbol{z}_1^T\\ \vdots\\ \boldsymbol{z}_{p-1}^T\\ \boldsymbol{z}_q^T\\ \boldsymbol{z}_{p+1}^T\\ \vdots\\ \boldsymbol{z}_{q-1}^T\\ \boldsymbol{z}_p^T\\ \boldsymbol{z}_{q+1}^T\\ \vdots\\ \boldsymbol{z}_n^T\end{pmatrix},\quad \boldsymbol{A}\boldsymbol{T}_{pq} = \left(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_{p-1},\boldsymbol{s}_q,\boldsymbol{s}_{p+1},\ldots,\boldsymbol{s}_{q-1},\boldsymbol{s}_p,\boldsymbol{s}_{q+1},\ldots,\boldsymbol{s}_n\right).$$

Eine (n,n)-Matrix $\boldsymbol{P}$ der Form

$$\boldsymbol{P} = (\boldsymbol{e}_{l_1},\boldsymbol{e}_{l_2},\ldots,\boldsymbol{e}_{l_n}),$$

wobei $\{l_1, l_2, \ldots, l_n\}$ eine Permutation der Zahlen $\{1, 2, \ldots, n\}$ ist, nennt man **Permutationsmatrix**. Offenbar ist das Produkt beliebig vieler Tauschmatrizen gleicher Dimension eine Permutationsmatrix. Umgekehrt lässt sich jede Permutationsmatrix als Produkt von Tauschmatrizen darstellen. Tauschmatrizen sind daher spezielle Permutationsmatrizen. Multipliziert man dementsprechend eine (n,n)-Matrix $\boldsymbol{A}$ von links mit einer (n,n)-Permutationsmatrix $\boldsymbol{P}$, so erhält man eine Matrix $\bar{\boldsymbol{A}} = \boldsymbol{PA}$, deren Zeilen gegenüber $\boldsymbol{A}$ entsprechend $\boldsymbol{P}$ permutiert sind. Rechtsmultiplikation einer Matrix $\boldsymbol{A}$ bewirkt in analoger Weise eine Spaltenpermutation.
Es bleibt noch zu zeigen, dass Permutationsmatrizen orthogonal sind. Es gilt:

$$\begin{aligned}\left(\boldsymbol{PP}^T\right)_{ij} &= \boldsymbol{e}_{l_i}{\boldsymbol{e}_{l_j}}^T \\ &= \delta_{l_i l_j} \\ &= \delta_{ij}\end{aligned}$$

wegen $l_i = l_j \Longleftrightarrow i = j$. Damit folgt

$$\boldsymbol{PP}^T = \boldsymbol{I},$$

also die Orthogonalität von $\boldsymbol{P}$.
Bemerkung: Für die Speicherung einer (n,n)-Permutationsmatrix $\boldsymbol{P}$ benötigt man nur ein Feld p der Länge n. Bewirkt die Permutationsmatrix $\boldsymbol{P}$ bei einer Multiplikation $\boldsymbol{Px}$ mit einem Vektor $\boldsymbol{x} \in \mathbb{R}^n$ gerade den Tausch der i-ten Komponente von $\boldsymbol{x}$ auf die Position p_i, so setzt man $p(i) = p_i$.
Weitere orthogonale Transformationen sind die sogenannten GIVENS-Drehungen. Eine (n,n)-Matrix $\boldsymbol{G} = \boldsymbol{G}_{pq}(c,s)$ der Form

$$\boldsymbol{G} = \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & \boldsymbol{O} & \\ & & 1 & & & & & & \\ & & & c & & & s & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & -s & & & c & & \\ & & & & & & & 1 & \\ & \boldsymbol{O} & & & & & & \ddots & \\ & & & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ p\text{-te Zeile} \\ \\ \\ \\ q\text{-te Zeile} \\ \\ \\ \\ \end{matrix} \quad , \quad c^2 + s^2 = 1$$

heißt GIVENS-**Drehung** oder GIVENS-**Matrix**.
Eine GIVENS-Matrix unterscheidet sich von der Einheitsmatrix nur in den Elementen

$(\boldsymbol{G})_{pp}$, $(\boldsymbol{G})_{pq}$, $(\boldsymbol{G})_{qp}$ und $(\boldsymbol{G})_{qq}$. Die Orthogonalität ist offensichtlich. Geometrisch stellt die Transformation

$$\boldsymbol{x} \longrightarrow \boldsymbol{y} = \boldsymbol{G}_{pq}(c,s)\boldsymbol{x}$$

eine Drehung in der (pq)-Ebene um den Winkel $-\varphi$ mit $c = \cos\varphi$ und $s = \sin\varphi$ dar. Von großer Bedeutung sind auch die HOUSEHOLDER-Spiegelungen.
Eine (n,n)-Matrix $\boldsymbol{H} = \boldsymbol{H}(\boldsymbol{u})$ der Form

$$\boldsymbol{H} = \boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^T, \quad \boldsymbol{u} \in \mathbb{R}^n, \quad \boldsymbol{u}^T\boldsymbol{u} = 1$$

heißt HOUSEHOLDER-**Spiegelung** oder HOUSEHOLDER-**Matrix**. Offensichtlich ist $\boldsymbol{H}$ symmetrisch. Auch die Orthogonalität erkennt man leicht. Es gilt

$$\begin{aligned}\boldsymbol{H}\boldsymbol{H}^T &= \boldsymbol{H}\boldsymbol{H} = (\boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^T)(\boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^T) \\ &= \boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^T - 2\boldsymbol{u}\boldsymbol{u}^T + 4\boldsymbol{u}\boldsymbol{u}^T\boldsymbol{u}\boldsymbol{u}^T = \boldsymbol{I} - 4\boldsymbol{u}\boldsymbol{u}^T + 4\boldsymbol{u}\boldsymbol{u}^T = \boldsymbol{I}.\end{aligned}$$

Geometrisch lässt sich die Abbildung

$$\boldsymbol{x} \longrightarrow \boldsymbol{y} = \boldsymbol{H}(\boldsymbol{u})\boldsymbol{x}$$

als Spiegelung des Vektors $\boldsymbol{x}$ an einer Ebene mit dem Normalenvektor $\boldsymbol{u}$ deuten. Verzichtet man auf die Normierung $\boldsymbol{u}^T\boldsymbol{u} = 1$, so stellt sich eine HOUSEHOLDER-Matrix in der Form

$$\boldsymbol{H} = \boldsymbol{I} - \frac{\boldsymbol{v}\boldsymbol{v}^T}{\gamma}, \quad \gamma = \frac{\boldsymbol{v}^T\boldsymbol{v}}{2}$$

dar.

8.1.4. Eigenwerte und Singulärwerte

Bekanntlich lassen sich für quadratische Matrizen sogenannte Eigenwerte und Eigenvektoren definieren. Eine Zahl $\lambda \in \mathbb{C}$ heißt **Eigenwert** der (n,n)-Matrix $\boldsymbol{A}$, falls es einen Vektor $\boldsymbol{x} \in \mathbb{C}^n$, $\boldsymbol{x} \neq \boldsymbol{o}$, gibt, so dass

$$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$$

gilt. Jeder derartige Vektor $\boldsymbol{x}$ heißt **Eigenvektor** zum Eigenwert λ. Eine notwendige und hinreichende Bedingung dafür, dass λ Eigenwert von $\boldsymbol{A}$ ist, ist die Säkulargleichung

$$\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0.$$

Das Polynom

$$\varphi(\mu) = \det(\boldsymbol{A} - \mu \boldsymbol{I})$$

heißt **charakteristisches Polynom** der (n,n)-Matrix $\boldsymbol{A}$. Es gilt $\varphi \in \Pi_n$. Die Eigenwerte einer Matrix sind damit die Nullstellen des charakteristischen Polynoms. Im allgemeinen sind die Eigenwerte und damit auch die Eigenvektoren einer beliebigen reellen Matrix komplex. Es gilt aber der folgende Satz.

8.6. Satz: *Die Eigenwerte einer symmetrischen (n,n)-Matrix $\boldsymbol{A}$ sind sämtlich reell. Zu verschiedenen Eigenwerten gehörende Eigenvektoren sind paarweise orthogonal. Es existiert eine orthogonale (n,n)-Matrix $\boldsymbol{U}$, so dass*

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

gilt. $\lambda_1, \ldots, \lambda_n$ sind dabei die Eigenwerte von $\boldsymbol{A}$, die Spalten von $\boldsymbol{U}$ sind zugehörige normierte Eigenvektoren.

Eine Abschätzung der Eigenwerte einer Matrix liefert der folgende Satz.

8.7. Satz: *Für alle Eigenwerte λ einer (n,n)-Matrix $\boldsymbol{A}$ gilt*

$$|\lambda| \leqq \operatorname{lub}(\boldsymbol{A})$$

bezüglich einer beliebigen Matrixgrenznorm.

Beweis: Ist $\boldsymbol{x} \neq \boldsymbol{o}$ Eigenvektor zum Eigenwert λ, so gilt $\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$. Daraus folgt

$$|\lambda| \, \|\boldsymbol{x}\| = \|\boldsymbol{A}\boldsymbol{x}\| \leqq \operatorname{lub}(\boldsymbol{A}) \|\boldsymbol{x}\|,$$

und nach Division durch $\|\boldsymbol{x}\|$

$$|\lambda| \leqq \operatorname{lub}(\boldsymbol{A}).$$

❀

Bemerkung: Ist $\|\boldsymbol{A}\|$ eine Matrixnorm, die mit irgendeiner Vektornorm verträglich ist, so folgt aus diesem Satz und Satz 8.3 sofort $|\lambda| \leqq \|\boldsymbol{A}\|$ für einen beliebigen Eigenwert λ von $\boldsymbol{A}$.
Beliebige Matrizen lassen sich durch orthogonale Transformationen nicht diagonalisieren. Für sie gilt aber der folgende Satz.

8.8. Singulärwertzerlegung: *Zu jeder (m,n)-Matrix $\boldsymbol{A}$ gibt es eine orthogonale (m,m)-Matrix $\boldsymbol{U}$ und eine orthogonale (n,n)-Matrix $\boldsymbol{V}$, so dass*

$$\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{V} = \boldsymbol{\Sigma} = \text{diag}(\sigma_1,\ldots,\sigma_l)$$

mit $l = \min\{m,n\}$ und $\sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_l \geqq 0$ gilt. Die Zahlen $\sigma_1,\ldots,\sigma_l$ sind eindeutig bestimmt und heißen Singulärwerte der Matrix $\boldsymbol{A}$.

Bemerkung: Die Matrix $\boldsymbol{\Sigma}$ hat folgende Gestalt

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \\ & \boldsymbol{O} & \end{pmatrix}, \quad m > n = l$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1 & & 0 & \\ & \ddots & & \boldsymbol{O} \\ 0 & & \sigma_m & \end{pmatrix}, \quad n > m = l$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{pmatrix}, \quad m = n = l$$

Beweis: Die (n,n)-Matrix $\boldsymbol{A}^T\boldsymbol{A}$ ist symmetrisch und wegen

$$\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{A}\boldsymbol{x})^T(\boldsymbol{A}\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x}\|_2^2 \geqq 0$$

positiv semidefinit. Nach Satz 8.6 existiert damit eine orthogonale (n,n)-Matrix $\boldsymbol{V}$, so dass

$$\boldsymbol{V}^T(\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{V} = \text{diag}(\lambda_1,\ldots,\lambda_n) \geqq \boldsymbol{O}$$

gilt. O.B.d.A. sei

$$\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_r > \lambda_{r+1} = \cdots = \lambda_n = 0$$

angenommen. Weiterhin sei

$$\boldsymbol{V} = (\boldsymbol{v}^{(1)},\ldots,\boldsymbol{v}^{(n)}).$$

Dann gilt

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{v}^{(j)} = \lambda_j\boldsymbol{v}^{(j)}, \quad j = 1,\ldots,n,$$

und damit

$$\lambda_j = \boldsymbol{v}^{(j)^T} \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{v}^{(j)} = \|\boldsymbol{A}\boldsymbol{v}^{(j)}\|_2^2.$$

Es ist also $\boldsymbol{A}\boldsymbol{v}^{(j)} \neq \boldsymbol{o}$ für $j = 1, \ldots, r$ und $\boldsymbol{A}\boldsymbol{v}^{(j)} = \boldsymbol{o}$ für $j = r+1, \ldots, n$. Wir definieren nun

$$\sigma_j = \sqrt{\lambda_j}, \quad j = 1, \ldots, r$$

und

$$\boldsymbol{u}^{(j)} = \frac{\boldsymbol{A}\boldsymbol{v}^{(j)}}{\sigma_j}, \quad j = 1, \ldots, r.$$

Für $i, j = 1, \ldots, r$ gilt dann

$$\boldsymbol{u}^{(i)^T} \boldsymbol{u}^{(j)} = \frac{\boldsymbol{v}^{(i)^T} \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{v}^{(j)}}{\sigma_i \sigma_j} = \frac{\lambda_j}{\sigma_i \sigma_j} \boldsymbol{v}^{(i)^T} \boldsymbol{v}^{(j)} = \delta_{ij}.$$

Die Vektoren $\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(r)}$ sind paarweise orthogonal und auf 1 normiert. Durch Hinzunahme weiterer Vektoren $\boldsymbol{u}^{(r+1)}, \ldots, \boldsymbol{u}^{(m)}$ sind sie zu einer orthonormierten Basis des $\mathbb{R}^m$ ergänzt. Dann ist die Matrix

$$\boldsymbol{U} = \left(\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(m)} \right)$$

orthogonal, und es gilt

$$\left(\boldsymbol{U}^T \boldsymbol{A} \boldsymbol{V} \right)_{ij} = \boldsymbol{u}^{(i)^T} \boldsymbol{A} \boldsymbol{v}^{(j)} = \begin{cases} \sigma_j \boldsymbol{u}^{(i)^T} \boldsymbol{u}^{(j)} & \text{für} \quad j = 1, \ldots, r, \\ \\ 0 & \text{für} \quad j = r+1, \ldots, n. \end{cases}$$

Setzen wir noch $\sigma_{r+1} = \cdots = \sigma_l = 0$, so gilt

$$\boldsymbol{U}^T \boldsymbol{A} \boldsymbol{V} = \boldsymbol{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_l).$$

❀

Die Singulärwerte einer Matrix $\boldsymbol{A}$ sind somit die Wurzeln aus den Eigenwerten der Matrix $\boldsymbol{A}^T \boldsymbol{A}$ (oder $\boldsymbol{A}\boldsymbol{A}^T$). Weiterhin gilt

- $\|\boldsymbol{A}\|_2 = \sigma_1$,
- $\|\boldsymbol{A}\|_F = \sqrt{\sigma_1^2 + \cdots \sigma_r^2}$,

- $\text{rg}(\boldsymbol{A}) = r$.

Für den Zusammenhang zwischen Eigenwerten und Singulärwerten einer quadratischen Matrix gilt

- $\det(\boldsymbol{A}) = \prod_{i=1}^{n} \lambda_i$ und $|\det(\boldsymbol{A})| = \prod_{i=1}^{n} \sigma_i$,
- $\sigma_1 \geqq |\lambda_i| \geqq \sigma_n$ für $i = 1, \ldots, n$,
- $\sigma_i = |\lambda_i|$ für $i = 1, \ldots, n$ falls $\boldsymbol{A}$ symmetrisch ist.

8.1.5. Störungstheorie

Wir betrachten ein lineares Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ mit einer (n,n)-Matrix $\boldsymbol{A}$ und $mb \in \mathbb{R}^n$. Bekanntlich gilt dann der folgende Satz.

8.9. Satz: *Das lineare Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ ist genau dann eindeutig lösbar, wenn die Matrix $\boldsymbol{A}$ regulär ist.*

Dieser Satz ist in der Praxis mit Vorsicht zu genießen, wie das folgende Beispiel zeigt.

8.10. Beispiel: Es sei

$$\boldsymbol{A} = \begin{pmatrix} 1.00 & 0.99 \\ 0.99 & 0.98 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} 1.99 \\ 1.97 \end{pmatrix}.$$

Die exakte Lösung des Gleichungssystems $\boldsymbol{Ax} = \boldsymbol{b}$ ist $\boldsymbol{x} = (1,1)^T$. Die Matrix $\boldsymbol{A}$ ist regulär, denn es gilt $\det(\boldsymbol{A}) = 1 \cdot 0.98 - 0.99^2 = -0.0001 \neq 0$. Behandelt man dieses Gleichungssystems auf einem Rechner mit dem Maschinenzahlbereich $\mathbb{M}(10,2,.,.)$, so ergibt sich

$$\text{gl}\,(\det(\boldsymbol{A})) = \text{gl}\,(0.98 - 0.99^2) = 0.98 - 0.98 = 0.$$

Auf diesem Rechner würde die Matrix als singulär betrachtet werden. ♡

Der Begriff der Regularität der Koeffizientenmatrix reicht daher in der Numerik nicht aus, um die (praktische) Lösbarkeit eines Gleichungssystems zu charakterisieren. Eine (n,n)-Matrix $\boldsymbol{A}$ heißt **numerisch regulär**, falls alle (n,n)-Matrizen $\tilde{\boldsymbol{A}}$ aus einer Umgebung $U(\boldsymbol{A})$ regulär sind. Existiert eine singuläre Matrix $\tilde{\boldsymbol{A}} \in U(\boldsymbol{A})$, so heißt die Matrix $\boldsymbol{A}$ **numerisch singulär**. Die Größe der Umgebung $U(\boldsymbol{A})$ hängt dabei sowohl von der verwendeten Hardware (relative Maschinengenauigkeit), als auch von der verwendeten Software (zur Behandlung der Matrix) ab. Zurück zum Beispiel 8.10.

8.11. Beispiel: Wir betrachten folgende Umgebung der Matrix $\boldsymbol{A}$:

$$U(\boldsymbol{A}) = \left\{ \bar{\boldsymbol{A}} \in \mathbb{R}^{2\times 2} \;\middle|\; |\bar{\boldsymbol{A}} - \boldsymbol{A}| \leqq 0.5 \cdot 10^{-2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \right\}.$$

Das ist gerade die Menge von Matrizen, die bei der Konvertierung in den Maschinenzahlbereich $\mathbb{M}(10, 2, ., .)$ auf die Matrix $\boldsymbol{A}$ abgebildet werden. Wir sehen daher $\boldsymbol{A}$ als Repräsentanten aller Matrizen $\bar{\boldsymbol{A}} \in U(\boldsymbol{A})$ an. Für die Matrix

$$\tilde{\boldsymbol{A}} = \begin{pmatrix} 1.000 & 0.9900 \\ 0.9900 & 0.9801 \end{pmatrix} \in U(\boldsymbol{A})$$

gilt aber $\det(\tilde{\boldsymbol{A}}) = 0$ und $\mathrm{rg}(\tilde{\boldsymbol{A}}) = 1$. Folglich enthält $U(\boldsymbol{A})$ singuläre Matrizen. Die Matrix $\boldsymbol{A}$ ist also in diesem Maschinenzahlbereich als numerisch singulär anzusehen. ♡

Bemerkungen: (i) Hinter der Aussage von Satz 8.9 steckt bei der Prüfung der Regularität von $\boldsymbol{A}$ der Test einer REAL-Zahl auf Gleichheit mit Null. Solche Tests sollte man in Programmen vermeiden. Dafür ist ein Test der Form $|x| \leqq \varepsilon$ mit einer vorgegebenen Genauigkeit ε besser.
(ii) Die Umgebung $U(\boldsymbol{A})$ lässt sich meist in der Form

$$U(\boldsymbol{A}) = \left\{ \bar{\boldsymbol{A}} \in \mathbb{R}^{n\times n} \;\middle|\; \frac{|\bar{a}_{ij} - a_{ij}|}{1 + |a_{ij}|} \leqq K\mathrm{eps};\; i, j = 1, \ldots, n \right\}$$

darstellen. Dabei ist eps die relative Maschinengenauigkeit und K eine Konstante, die von der verwendeten Software abhängt.
Schon das letzte kleine Beispiel zeigt, dass Störungen die Aufgabe mehr oder weniger stark verändern. Den Einfluss dieser Störungen wollen wir nun genauer untersuchen.
Neben dem Gleichungssystem

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

mit der Lösung $\boldsymbol{x}$ betrachten wir ein gestörtes System

$$(\boldsymbol{A} + \boldsymbol{\delta A})(\boldsymbol{x} + \boldsymbol{\delta x}) = \boldsymbol{b} + \boldsymbol{\delta b}.$$

Die Lösung $\boldsymbol{x} + \boldsymbol{\delta x}$ des gestörten Systems weicht um $\boldsymbol{\delta x}$ von der Lösung des ursprünglichen Systems ab. Der Zusammenhang zwischen dieser Störung $\boldsymbol{\delta x}$ und den Eingabefehlern $\boldsymbol{\delta A}$ in der Matrix und $\boldsymbol{\delta b}$ in der rechten Seite soll nun untersucht werden. Wir stellen uns dabei zwei Fragen.

1. Wann besitzt das gestörte System eine eindeutige Lösung?

2. Wie lässt sich $\|\boldsymbol{\delta x}\|$ durch $\|\boldsymbol{\delta A}\|$ und $\|\boldsymbol{\delta b}\|$ abschätzen?
   Wie lässt sich $|\boldsymbol{\delta x}|$ durch $|\boldsymbol{\delta A}|$ und $|\boldsymbol{\delta b}|$ abschätzen?

Die erste Frage ist gerade die Frage nach der numerischen Regularität von $\boldsymbol{A}$. Wir fragen danach, wie groß eine Umgebung der regulären Matrix $\boldsymbol{A}$ maximal ist, so dass sie nur reguläre Matrizen enthält. Für die Einheitsmatrix wird diese Frage mit folgendem Satz beantwortet.

8.12. Satz: *Für die (n,n)-Matrix $\boldsymbol{P}$ gelte $\|\boldsymbol{P}\| < 1$. Dann ist die Matrix $\boldsymbol{I}-\boldsymbol{P}$ regulär, und es gilt die Abschätzung*

$$\|(\boldsymbol{I}-\boldsymbol{P})^{-1}\| \leqq \frac{1}{1-\|\boldsymbol{P}\|}.$$

Beweis: Für jedes $\boldsymbol{x} \in \mathbb{R}^n$ gilt

$$\begin{aligned} \|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{x}\| &= \|\boldsymbol{x}-\boldsymbol{P}\boldsymbol{x}\| \geqq |\,\|\boldsymbol{x}\|-\|\boldsymbol{P}\boldsymbol{x}\|\,| \geqq \|\boldsymbol{x}\|-\|\boldsymbol{P}\boldsymbol{x}\| \\ &\geqq \|\boldsymbol{x}\|-\|\boldsymbol{P}\|\,\|\boldsymbol{x}\| = (1-\|\boldsymbol{P}\|)\|\boldsymbol{x}\| > 0, \quad \boldsymbol{x} \neq \boldsymbol{o}. \end{aligned}$$

Damit ist $\boldsymbol{I}-\boldsymbol{P}$ regulär, und es existiert die Inverse $\boldsymbol{C} = (\boldsymbol{I}-\boldsymbol{P})^{-1}$. Dann gilt

$$\begin{aligned} 1 = \|\boldsymbol{I}\| &= \|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{C}\| = \|\boldsymbol{C}-\boldsymbol{P}\boldsymbol{C}\| \geqq \|\boldsymbol{C}\|-\|\boldsymbol{P}\boldsymbol{C}\| \\ &\geqq \|\boldsymbol{C}\|-\|\boldsymbol{P}\|\,\|\boldsymbol{C}\| = (1-\|\boldsymbol{P}\|)\|\boldsymbol{C}\|. \end{aligned}$$

Wegen $1-\|\boldsymbol{P}\| > 0$ folgt daraus

$$\|\boldsymbol{C}\| = \|(\boldsymbol{I}-\boldsymbol{P})^{-1}\| \leqq \frac{1}{1-\|\boldsymbol{P}\|}.$$

❀

Bemerkung: Die Bedingung $\|\boldsymbol{P}\| < 1$ ist bezüglich einer Grenznorm auch notwendig. Das erkennt man, falls man

$$\boldsymbol{P} = \begin{pmatrix} 1 & & \\ & 0 & \boldsymbol{O} \\ & & \ddots \\ & \boldsymbol{O} & & 0 \end{pmatrix}$$

betrachtet. Hier gilt $\operatorname{rg}(\boldsymbol{I}-\boldsymbol{P}) = n-1$ und $\|\boldsymbol{P}\| = 1$.

8.13. Störungslemma: *Gegeben seien eine reguläre (n,n)-Matrix $\boldsymbol{A}$ und eine Störungsmatrix $\boldsymbol{\delta A}$, die der Bedingung*

$$\kappa = \|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{\delta A}\| < 1$$

genügt. Dann ist auch die Matrix $\boldsymbol{A}+\boldsymbol{\delta A}$ regulär und es gilt

1.

$$\left\|(\boldsymbol{A}+\boldsymbol{\delta A})^{-1}\right\| \leqq \frac{\left\|\boldsymbol{A}^{-1}\right\|}{1-\kappa},$$

2.

$$\left\|(\boldsymbol{A}+\boldsymbol{\delta A})^{-1}-\boldsymbol{A}^{-1}\right\| \leqq \frac{\left\|\boldsymbol{A}^{-1}\right\|^2}{1-\kappa}\|\boldsymbol{\delta A}\|.$$

Beweis: Es sei $\boldsymbol{P} = -\boldsymbol{A}^{-1}\boldsymbol{\delta A}$. Dann gilt

$$\|\boldsymbol{P}\| = \|\boldsymbol{A}^{-1}\boldsymbol{\delta A}\| \leqq \|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{\delta A}\| = \kappa < 1.$$

Nach Satz 8.12 ist dann die Matrix $\boldsymbol{I}-\boldsymbol{P} = \boldsymbol{I}+\boldsymbol{A}^{-1}\boldsymbol{\delta A}$ regulär. Wegen der Regularität von $\boldsymbol{A}$ ist damit aber auch $\boldsymbol{A}+\boldsymbol{\delta A}$ regulär. Für die erste Abschätzung folgt

$$\begin{aligned}\left\|(\boldsymbol{A}+\boldsymbol{\delta A})^{-1}\right\| &= \left\|\left[\boldsymbol{A}\left(\boldsymbol{I}+\boldsymbol{A}^{-1}\boldsymbol{\delta A}\right)\right]^{-1}\right\| = \left\|\left(\boldsymbol{I}+\boldsymbol{A}^{-1}\boldsymbol{\delta A}\right)^{-1}\boldsymbol{A}^{-1}\right\| \\ &\leqq \left\|\left(\boldsymbol{I}+\boldsymbol{A}^{-1}\boldsymbol{\delta A}\right)^{-1}\right\|\left\|\boldsymbol{A}^{-1}\right\| \leqq \frac{\left\|\boldsymbol{A}^{-1}\right\|}{1-\left\|\boldsymbol{A}^{-1}\boldsymbol{\delta A}\right\|} \leqq \frac{\left\|\boldsymbol{A}^{-1}\right\|}{1-\kappa}.\end{aligned}$$

Für die zweite Abschätzung erhalten wir

$$\begin{aligned}(\boldsymbol{A}+\boldsymbol{\delta A})^{-1}-\boldsymbol{A}^{-1} &= (\boldsymbol{A}+\boldsymbol{\delta A})^{-1}\left[\boldsymbol{I}-(\boldsymbol{A}+\boldsymbol{\delta A})\boldsymbol{A}^{-1}\right] \\ &= (\boldsymbol{A}+\boldsymbol{\delta A})^{-1}\left[\boldsymbol{A}-(\boldsymbol{A}+\boldsymbol{\delta A})\right]\boldsymbol{A}^{-1} \\ &= -(\boldsymbol{A}+\boldsymbol{\delta A})^{-1}\boldsymbol{\delta A}\boldsymbol{A}^{-1}.\end{aligned}$$

Damit erhält man

$$\left\|(\boldsymbol{A}+\boldsymbol{\delta A})^{-1}-\boldsymbol{A}^{-1}\right\| \leqq \left\|(\boldsymbol{A}+\boldsymbol{\delta A})^{-1}\right\|\,\|\boldsymbol{\delta A}\|\,\left\|\boldsymbol{A}^{-1}\right\| \leqq \frac{\left\|\boldsymbol{A}^{-1}\right\|^2}{1-\kappa}\|\boldsymbol{\delta A}\|.$$

❀

Bemerkungen: (i) Durch $\kappa = \|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{\delta A}\| < 1$ wird eine Umgebung der regulären Matrix $\boldsymbol{A}$ festgelegt, die nur reguläre Matrizen enthält:

$$U(\boldsymbol{A}) = \left\{ \bar{\boldsymbol{A}} \in \mathbb{R}^{n\times n} \;\middle|\; \|\boldsymbol{A} - \bar{\boldsymbol{A}}\| \leqq \Delta\boldsymbol{A} < \frac{1}{\|\boldsymbol{A}^{-1}\|} \right\}.$$

Die Größe $\|\boldsymbol{A}^{-1}\|$ ist damit ein Maß für die Regularität von $\boldsymbol{A}$.
(ii) Betrachtet man die Inversenbildung als Funktion $\boldsymbol{\Phi} : \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}^{n\times n}$, so erkennt man aus der zweiten Abschätzung in Satz 8.13, dass für ein $\bar{\boldsymbol{A}} \in U(\boldsymbol{A})$

$$\begin{aligned}\left\| \bar{\boldsymbol{A}}^{-1} - \boldsymbol{A}^{-1} \right\| &= \|\boldsymbol{\Phi}(\bar{\boldsymbol{A}}) - \boldsymbol{\Phi}(\boldsymbol{A})\| \\ &\leqq \frac{\|\boldsymbol{A}^{-1}\|^2}{1 - \|\bar{\boldsymbol{A}} - \boldsymbol{A}\|\,\|\boldsymbol{A}^{-1}\|} \|\bar{\boldsymbol{A}} - \boldsymbol{A}\| \\ &\leqq \frac{\|\boldsymbol{A}^{-1}\|^2}{1 - \Delta\boldsymbol{A}\|\boldsymbol{A}^{-1}\|} \|\bar{\boldsymbol{A}} - \boldsymbol{A}\|\end{aligned}$$

und damit

$$\|\boldsymbol{\Phi}(\bar{\boldsymbol{A}}) - \boldsymbol{\Phi}(\boldsymbol{A})\| \leqq L(\boldsymbol{A}, \Delta\boldsymbol{A}) \|\bar{\boldsymbol{A}} - \boldsymbol{A}\|$$

gilt. Die Inversenbildung ist daher für eine reguläre Matrix lokal lipschitzstetig mit der LIPSCHITZ-Konstanten

$$L(\boldsymbol{A}, \Delta\boldsymbol{A}) = \frac{\|\boldsymbol{A}^{-1}\|^2}{1 - \Delta\boldsymbol{A}\|\boldsymbol{A}^{-1}\|} \|\bar{\boldsymbol{A}} - \boldsymbol{A}\|.$$

Nach diesen Vorbereitungen ist der Einfluss von Störungen in einem linearen Gleichungssystem abschätzbar. Es sei durch

$$(\boldsymbol{A} + \boldsymbol{\delta A})(\boldsymbol{x} + \boldsymbol{\delta x}) = \boldsymbol{b} + \boldsymbol{\delta b}$$

ein gestörtes lineares Gleichungssystem gegeben. Die Störung $\boldsymbol{\delta A}$ genüge der Bedingung

$$\kappa = \|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{\delta A}\| < 1.$$

Nach Satz 8.13 ist dann die Matrix $\boldsymbol{A} + \boldsymbol{\delta A}$ regulär, und das gestörte System besitzt eine eindeutige Lösung

$$\boldsymbol{x} + \boldsymbol{\delta x} = (\boldsymbol{A} + \boldsymbol{\delta A})^{-1}(\boldsymbol{b} + \boldsymbol{\delta b}).$$

Dann gilt

$$\begin{aligned}
\boldsymbol{\delta x} &= (\boldsymbol{A}+\boldsymbol{\delta A})^{-1}(\boldsymbol{b}+\boldsymbol{\delta b})-\boldsymbol{A}^{-1}\boldsymbol{b} \\
&= (\boldsymbol{A}+\boldsymbol{\delta A})^{-1}\left[\boldsymbol{b}+\boldsymbol{\delta b}-(\boldsymbol{A}+\boldsymbol{\delta A})\boldsymbol{A}^{-1}\boldsymbol{b}\right] \\
&= (\boldsymbol{A}+\boldsymbol{\delta A})^{-1}\left[\boldsymbol{b}+\boldsymbol{\delta b}-\boldsymbol{b}-\boldsymbol{\delta A}\cdot\boldsymbol{A}^{-1}\boldsymbol{b}\right] \\
&= (\boldsymbol{A}+\boldsymbol{\delta A})^{-1}\left[-\boldsymbol{\delta A}\boldsymbol{x}+\boldsymbol{\delta b}\right] \\
&= \boldsymbol{A}^{-1}\left[-\boldsymbol{\delta A}\boldsymbol{x}+\boldsymbol{\delta b}\right]+\left[(\boldsymbol{A}+\boldsymbol{\delta A})^{-1}-\boldsymbol{A}^{-1}\right]\left[-\boldsymbol{\delta A}\boldsymbol{x}+\boldsymbol{\delta b}\right] \\
&= \boldsymbol{\delta x}'+O(\|\boldsymbol{\delta A}\|(\|\boldsymbol{\delta A}\|+\|\boldsymbol{\delta b}\|)).
\end{aligned}$$

$\boldsymbol{\delta x}' = \boldsymbol{A}^{-1}\left[-\boldsymbol{\delta A}\boldsymbol{x}+\boldsymbol{\delta b}\right]$ stellt dabei den bezüglich $\boldsymbol{\delta A}$ und $\boldsymbol{\delta b}$ linearen Teil des Fehlers $\boldsymbol{\delta x}$ dar. Für $\boldsymbol{\delta x}$ erhält man die Abschätzung

$$\|\boldsymbol{\delta x}\| \leqq \left\|(\boldsymbol{A}+\boldsymbol{\delta A})^{-1}\right\| \, \|-\boldsymbol{\delta A}\boldsymbol{x}+\boldsymbol{\delta b}\| \leqq \frac{\left\|\boldsymbol{A}^{-1}\right\|}{1-\kappa}\left(\|\boldsymbol{\delta A}\|\,\|\boldsymbol{x}\|+\|\boldsymbol{\delta b}\|\right).$$

Für ein inhomogenes Gleichungssystem ($\boldsymbol{x} \neq \boldsymbol{o}$) lässt sich auch der relative Fehler der Lösung abschätzen. Man erhält

$$\begin{aligned}
\frac{\|\boldsymbol{\delta x}\|}{\|\boldsymbol{x}\|} &\leqq \frac{\left\|\boldsymbol{A}^{-1}\right\|}{1-\kappa}\left(\|\boldsymbol{\delta A}\|+\frac{\|\boldsymbol{\delta b}\|}{\|\boldsymbol{x}\|}\right) \\
&= \frac{1}{1-\kappa}\left(\|\boldsymbol{A}\|\left\|\boldsymbol{A}^{-1}\right\|\frac{\|\boldsymbol{\delta A}\|}{\|\boldsymbol{A}\|}+\frac{\left\|\boldsymbol{A}^{-1}\right\|\|\boldsymbol{b}\|}{\|\boldsymbol{x}\|}\cdot\frac{\|\boldsymbol{\delta b}\|}{\|\boldsymbol{b}\|}\right).
\end{aligned}$$

Berücksichtigt man noch

$$\|\boldsymbol{b}\| = \|\boldsymbol{A}\boldsymbol{x}\| \leqq \|\boldsymbol{A}\|\,\|\boldsymbol{x}\|,$$

so ergibt sich

$$\begin{aligned}
\frac{\|\boldsymbol{\delta x}\|}{\|\boldsymbol{x}\|} &\leqq \frac{1}{1-\kappa}\left(\|\boldsymbol{A}\|\left\|\boldsymbol{A}^{-1}\right\|\frac{\|\boldsymbol{\delta A}\|}{\|\boldsymbol{A}\|}+\|\boldsymbol{A}\|\left\|\boldsymbol{A}^{-1}\right\|\frac{\|\boldsymbol{\delta b}\|}{\|\boldsymbol{b}\|}\right) \\
&= \frac{\operatorname{cond}(\boldsymbol{A})}{1-\kappa}\left(\frac{\|\boldsymbol{\delta A}\|}{\|\boldsymbol{A}\|}+\frac{\|\boldsymbol{\delta b}\|}{\|\boldsymbol{b}\|}\right)
\end{aligned}$$

mit $\operatorname{cond}(\boldsymbol{A}) = \|\boldsymbol{A}\|\left\|\boldsymbol{A}^{-1}\right\|$.
Wir fassen die Ergebnisse in einem Satz zusammen.

8.14. Satz: *Gegeben sei das lineare Gleichungssystem $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ mit einer regulären (n,n)-Matrix $\boldsymbol{A}$ und $\boldsymbol{b} \in \mathbb{R}^n$; ferner sei*

$$(\boldsymbol{A}+\boldsymbol{\delta A})(\boldsymbol{x}+\boldsymbol{\delta x}) = \boldsymbol{b}+\boldsymbol{\delta b}$$

ein gestörtes Gleichungssystem, wobei die Störung $\boldsymbol{\delta A}$ der Bedingung

$$\kappa = \left\|\boldsymbol{A}^{-1}\right\| \, \|\boldsymbol{\delta A}\| < 1$$

genüge. Dann ist die Matrix $\boldsymbol{A} + \boldsymbol{\delta A}$ regulär, und das gestörte Gleichungssystem besitzt die eindeutige Lösung

$$\boldsymbol{x} + \boldsymbol{\delta x} = (\boldsymbol{A} + \boldsymbol{\delta A})^{-1}(\boldsymbol{b} + \boldsymbol{\delta b}).$$

Für den Fehler $\boldsymbol{\delta x}$ der Lösung gilt:

1. $$\boldsymbol{\delta x} = \boldsymbol{\delta x}' + O(\|\boldsymbol{\delta A}\|(\|\boldsymbol{\delta A}\| + \|\boldsymbol{\delta b}\|)),$$

   wobei

   $$\boldsymbol{\delta x}' = \boldsymbol{A}^{-1}\left[-\boldsymbol{\delta A}\boldsymbol{x} + \boldsymbol{\delta b}\right]$$

   den bezüglich $\boldsymbol{\delta A}$ und $\boldsymbol{\delta b}$ linearen Teil des Fehlers $\boldsymbol{\delta x}$ darstellt.

2. $$\|\boldsymbol{\delta x}\| \leqq \frac{\left\|\boldsymbol{A}^{-1}\right\|}{1-\kappa}\left(\|\boldsymbol{x}\| \, \|\boldsymbol{\delta A}\| + \|\boldsymbol{\delta b}\|\right).$$

   $\left\|\boldsymbol{A}^{-1}\right\| \, \|\boldsymbol{x}\|$ bzw. $\left\|\boldsymbol{A}^{-1}\right\|$ sind die absoluten partiellen Konditionszahlen bezüglich $\boldsymbol{A}$ und $\boldsymbol{b}$.

3. *Für $\boldsymbol{b} \neq \boldsymbol{o}$ ($\boldsymbol{x} \neq \boldsymbol{o}$) gilt*

   $$\begin{aligned}\frac{\|\boldsymbol{\delta x}\|}{\|\boldsymbol{x}\|} &\leqq \frac{1}{1-\kappa}\left(\|\boldsymbol{A}\| \left\|\boldsymbol{A}^{-1}\right\| \frac{\|\boldsymbol{\delta A}\|}{\|\boldsymbol{A}\|} + \frac{\left\|\boldsymbol{A}^{-1}\right\| \, \|\boldsymbol{b}\|}{\|\boldsymbol{x}\|}\frac{\|\boldsymbol{\delta b}\|}{\|\boldsymbol{b}\|}\right) \\ &\leqq \frac{\operatorname{cond}(\boldsymbol{A})}{1-\kappa}\left(\frac{\|\boldsymbol{\delta A}\|}{\|\boldsymbol{A}\|} + \frac{\|\boldsymbol{\delta b}\|}{\|\boldsymbol{b}\|}\right).\end{aligned}$$

   $\|\boldsymbol{A}\| \left\|\boldsymbol{A}^{-1}\right\|$ und $\left\|\boldsymbol{A}^{-1}\right\| \, \|\boldsymbol{b}\|/\|\boldsymbol{x}\|$ sind die relativen partiellen Konditionszahlen bezüglich $\boldsymbol{A}$ und $\boldsymbol{b}$.
   Die Größe $\operatorname{cond}(\boldsymbol{A}) = \|\boldsymbol{A}\| \left\|\boldsymbol{A}^{-1}\right\|$ *wird schlechthin als* **Kondition** *der Matrix $\boldsymbol{A}$ bezeichnet.*

Bemerkungen: (i) Die Abschätzungen aus Satz 8.14 verkörpern den schlechtesten Fall. Im allgemeinen liefern sie zu pessimistische Schranken. Betrachten wir dazu wieder unser Beispiel. Es ist

$$\boldsymbol{A} = \begin{pmatrix} 1.00 & 0.99 \\ 0.99 & 0.98 \end{pmatrix}, \quad \boldsymbol{A}^{-1} = \begin{pmatrix} -9800 & 9900 \\ 9900 & -10000 \end{pmatrix}.$$

In der Spaltensummennorm gilt $\|\boldsymbol{A}\|_1 = 1.99$ und $\|\boldsymbol{A}^{-1}\|_1 = 19900$. Damit folgt $\text{cond}_1(\boldsymbol{A}) = 39601$. Die Voraussetzung von Satz 8.14 wäre schon scharf. Wir dürften nur Störungen in der Matrix zulassen, für die

$$\|\boldsymbol{\delta A}\|_1 < \frac{1}{\|\boldsymbol{A}^{-1}\|_1} = \frac{1}{19900} \approx 5.025 \cdot 10^{-5}$$

gilt. Schon für die kleine Störung

$$\boldsymbol{\delta A} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} 10^{-4}$$

wäre der Satz nicht anwendbar. Betrachten wir darum zunächst nur Störungen der rechten Seite, also den Fall $\boldsymbol{\delta A} = \boldsymbol{O}$. Für $\boldsymbol{b} \neq \boldsymbol{o}$ gilt dann

$$\frac{\|\boldsymbol{\delta x}\|}{\|\boldsymbol{x}\|} \leqq \text{cond}(\boldsymbol{A}) \frac{\|\boldsymbol{\delta b}\|}{\|\boldsymbol{b}\|}.$$

Für verschiedene rechte Seiten und verschiedene Störungen ergibt sich:

$\boldsymbol{b}$	$\boldsymbol{\delta b}$	$\frac{\|\boldsymbol{\delta b}\|}{\|\boldsymbol{b}\|}$	$\boldsymbol{x}$	$\boldsymbol{\delta x}$	$\frac{\|\boldsymbol{\delta x}\|}{\|\boldsymbol{x}\|}$	$\frac{\|\boldsymbol{\delta x}\| \, \|\boldsymbol{b}\|}{\|\boldsymbol{x}\| \, \|\boldsymbol{\delta b}\|}$
$\begin{pmatrix} 1.00 \\ 0.99 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 10^{-5} \end{pmatrix}$	$\approx 5 \cdot 10^{-6}$	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0.099 \\ 0.100 \end{pmatrix}$	≈ 0.2	39601
$\begin{pmatrix} -1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 10^{-5} \\ 0 \end{pmatrix}$	$\approx 5 \cdot 10^{-6}$	$\begin{pmatrix} 19700 \\ -19900 \end{pmatrix}$	$\begin{pmatrix} -0.0998 \\ 0.0990 \end{pmatrix}$	$\approx 5 \cdot 10^{-6}$	0.995

In der letzten Spalte ist jeweils der wahre Verstärkungsfaktor, mit dem die relativen Fehler in der rechten Seite in die Lösung eingehen, angegeben.
Betrachten wir nun auch Störungen in der Matrix. Es sei

$$\boldsymbol{\delta A} = \begin{pmatrix} 0 & 0 \\ 0 & 2 \cdot 10^{-5} \end{pmatrix}, \quad \|\boldsymbol{\delta A}\|_1 = 2 \cdot 10^{-5}, \quad \frac{\text{cond}_1(\boldsymbol{A})}{1 - \|\boldsymbol{\delta A}\|_1 \|\boldsymbol{A}^{-1}\|_1} \approx 65800.$$

Für verschiedene rechte Seiten erhalten wir folgende Ergebnisse:

$\boldsymbol{b}$	$\boldsymbol{x}$	$\boldsymbol{\delta x}$	$\frac{\|\boldsymbol{\delta x}\|}{\|\boldsymbol{x}\|}$	$\frac{\|\boldsymbol{\delta x}\|\,\|\boldsymbol{b}\|}{\|\boldsymbol{x}\|\,\|\boldsymbol{\delta b}\|}$
$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 100 \\ -100 \end{pmatrix}$	$\begin{pmatrix} 24.75 \\ -25.00 \end{pmatrix}$	≈ 0.24875	≈ 25000
$\begin{pmatrix} -1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 19700 \\ -19900 \end{pmatrix}$	$\begin{pmatrix} 4925.25 \\ -4975.00 \end{pmatrix}$	≈ 0.2500	≈ 25000

Für die Störung

$$\boldsymbol{\delta A} = \begin{pmatrix} -10^{-5} & 10^{-5} \\ -10^{-5} & 10^{-5} \end{pmatrix}, \quad \|\boldsymbol{\delta A}\|_1 = 2 \cdot 10^{-5}, \quad \frac{\mathrm{cond}_1(\boldsymbol{A})}{1 - \|\boldsymbol{\delta A}\|_1 \|\boldsymbol{A}^{-1}\|_1} \approx 65800$$

ergibt sich

$\boldsymbol{b}$	$\boldsymbol{x}$	$\boldsymbol{\delta x}$	$\frac{\|\boldsymbol{\delta x}\|}{\|\boldsymbol{x}\|}$	$\frac{\|\boldsymbol{\delta x}\|\,\|\boldsymbol{b}\|}{\|\boldsymbol{x}\|\,\|\boldsymbol{\delta b}\|}$
$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 100 \\ -100 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	0	0
$\begin{pmatrix} -1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 19700 \\ -19900 \end{pmatrix}$	$\begin{pmatrix} -0.2 \\ 0.2 \end{pmatrix}$	$\approx 10^{-5}$	≈ 0.5

Wir sehen, dass für verschiedene Abschätzungen Fälle existieren, bei denen die jeweiligen maximalen Fehlerverstärkungen voll wirksam werden. Es existieren aber auch Störungen vom gleichen Störungsniveau, die sich fast gar nicht auf die Lösung auswirken. Störungen in der Matrix können sich dabei bedeutend stärker bemerkbar machen, da die Abhängigkeit der Lösung von der Matrix Unstetigkeitsstellen besitzt.

(ii) Verwendet man die euklidische Norm zur Fehlerabschätzung, so lässt sich die dritte Ungleichung in Satz 8.14 etwas verschärfen. Es gilt

$$\frac{\|\boldsymbol{\delta x}\|_2}{\|\boldsymbol{x}\|_2} \leq \frac{1}{1 - \|\boldsymbol{\delta A}\|_2 \|\boldsymbol{A}^{-1}\|_2} \left(\mathrm{cond}_2(\boldsymbol{A}) \frac{\|\boldsymbol{\delta A}\|}{\|\boldsymbol{A}\|} + \|\boldsymbol{A}^{-1}\|_2 \frac{\|\boldsymbol{A}^T \boldsymbol{b}\|_2}{\|\boldsymbol{b}\|_2} \frac{\|\boldsymbol{\delta b}\|_2}{\|\boldsymbol{b}\|_2} \right).$$

Hier wird die Abhängigkeit des Verstärkungsfaktors von der rechten Seite besser berücksichtigt. Für die ersten Beispiele ($\boldsymbol{\delta A} = \boldsymbol{O}$) liefert das die Abschätzungen

$$\begin{aligned} \frac{\|\boldsymbol{\delta x}\|}{\|\boldsymbol{x}\|} &\leqq 36150 \frac{\|\boldsymbol{\delta b}\|}{\|\boldsymbol{b}\|}, \quad \boldsymbol{b} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \\ \frac{\|\boldsymbol{\delta x}\|}{\|\boldsymbol{x}\|} &\leqq 183 \frac{\|\boldsymbol{\delta b}\|}{\|\boldsymbol{b}\|}, \quad \boldsymbol{b} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}. \end{aligned}$$

(iii) In den Abschätzungen tauchen die Normen verschiedener Größen auf. Dabei sind $\|\boldsymbol{b}\|$ und $\|\boldsymbol{A}\|$ meist leicht zu berechnen. Für $\|\boldsymbol{\delta b}\|$ und $\|\boldsymbol{\delta A}\|$ kennt man oft Schranken. Das Berechnen von $\|\boldsymbol{A}^{-1}\|$ bereitet Schwierigkeiten, da die Inverse $\boldsymbol{A}^{-1}$ im allgemeinen natürlich nicht bekannt ist.
Die Aussagen von Satz 8.14 stellen eine Abschätzung des unvermeidbaren Fehlers dar. Mit Hilfe dieses Satzes ist nicht die Güte einer berechneten Lösung bewertbar. Im Sinne der Rückwärtsanalyse müssten wir eine berechnete Lösung $\bar{\boldsymbol{x}}$ des linearen Gleichungssystems $\boldsymbol{Ax} = \boldsymbol{b}$ akzeptieren, falls sie als Lösung eines benachbarten Systems $\bar{\boldsymbol{A}}\bar{\boldsymbol{x}} = \bar{\boldsymbol{b}}$ interpretierbar ist, wobei sich die Matrix $\bar{\boldsymbol{A}}$ und der Vektor $\bar{\boldsymbol{b}}$ von $\boldsymbol{A}$ und $\boldsymbol{b}$ nur in der Größenordnung des Eingabefehlers unterscheiden. Mit dem folgenden Satz lässt sich für eine berechnete Lösung entscheiden, ob sie Lösung eines derartigen benachbarten Systems ist.

8.15. PRAGER-OETTLI**:** *Durch $\boldsymbol{Ax} = \boldsymbol{b}$ sei ein lineares Gleichungssystem gegeben. Mit der Matrix $\Delta\boldsymbol{A} \geqq \boldsymbol{O}$ und dem Vektor $\Delta\boldsymbol{b} \geqq \boldsymbol{o}$ seien folgende Mengen definiert:*

$$\mathcal{A} = \left\{ \hat{\boldsymbol{A}} \in \mathbb{R}^{n\times n} \;\middle|\; |\hat{\boldsymbol{A}} - \boldsymbol{A}| \leqq \Delta\boldsymbol{A} \right\}$$

und

$$\mathcal{B} = \left\{ \hat{\boldsymbol{b}} \in \mathbb{R}^{n} \;\middle|\; |\hat{\boldsymbol{b}} - \boldsymbol{b}| \leqq \Delta\boldsymbol{b} \right\}.$$

Der Vektor $\bar{\boldsymbol{x}}$ ist genau dann exakte Lösung eines Gleichungssystems $\bar{\boldsymbol{A}}\bar{\boldsymbol{x}} = \bar{\boldsymbol{b}}$ mit $\bar{\boldsymbol{A}} \in \mathcal{A}$ und $\bar{\boldsymbol{b}} \in \mathcal{B}$, wenn

$$|\boldsymbol{r}(\bar{\boldsymbol{x}})| \leqq \Delta\boldsymbol{A}|\bar{\boldsymbol{x}}| + \Delta\boldsymbol{b}$$

gilt. Dabei bezeichnet $\boldsymbol{r}(\bar{\boldsymbol{x}}) = \boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{x}}$ das Residuum von $\bar{\boldsymbol{x}}$ bezüglich des ursprünglichen Gleichungssystems.

Beweis: ($\Rightarrow$) Wir nehmen an, dass eine Matrix $\bar{\boldsymbol{A}} \in \mathcal{A}$ und ein Vektor $\bar{\boldsymbol{b}} \in \mathcal{B}$ existieren, so dass $\bar{\boldsymbol{A}}\bar{\boldsymbol{x}} = \bar{\boldsymbol{b}}$ gilt. Dann gilt

$$\bar{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{\delta A}, \quad |\boldsymbol{\delta A}| \leqq \Delta\boldsymbol{A}$$

und

$$\bar{\boldsymbol{b}} = \boldsymbol{b} + \boldsymbol{\delta b}, \quad |\boldsymbol{\delta b}| \leqq \Delta \boldsymbol{b}.$$

Damit folgt

$$\begin{aligned} |\boldsymbol{r}(\bar{\boldsymbol{x}})| &= |\boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{x}}| = |\bar{\boldsymbol{b}} - \boldsymbol{\delta b} - (\bar{\boldsymbol{A}} - \boldsymbol{\delta A})\bar{\boldsymbol{x}}| = |\bar{\boldsymbol{b}} - \bar{\boldsymbol{A}}\bar{\boldsymbol{x}} - \boldsymbol{\delta b} + \boldsymbol{\delta A}\bar{\boldsymbol{x}}| \\ &= |-\boldsymbol{\delta b} + \boldsymbol{\delta A}\bar{\boldsymbol{x}}| \leqq |\boldsymbol{\delta b}| + |\boldsymbol{\delta A}|\,|\bar{\boldsymbol{x}}| \leqq \Delta \boldsymbol{b} + \Delta \boldsymbol{A}|\bar{\boldsymbol{x}}|. \end{aligned}$$

($\Leftarrow$) Es gelte

$$|\boldsymbol{r}(\bar{\boldsymbol{x}})| \leqq \Delta \boldsymbol{b} + \Delta \boldsymbol{A}|\bar{\boldsymbol{x}}|.$$

Es sei weiter

$$\boldsymbol{r}(\bar{\boldsymbol{x}}) = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix}, \; \boldsymbol{s} = \Delta \boldsymbol{b} + \Delta \boldsymbol{A}|\bar{\boldsymbol{x}}| = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{pmatrix} \geqq \boldsymbol{o}, \; \Delta \boldsymbol{b} = \begin{pmatrix} \Delta b_1 \\ \Delta b_2 \\ \vdots \\ \Delta b_n \end{pmatrix}. \; \bar{\boldsymbol{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_n \end{pmatrix}$$

und

$$\Delta \boldsymbol{A} = \begin{pmatrix} \Delta a_{11} & \Delta a_{12} & \cdots & \Delta a_{1n} \\ \Delta a_{21} & \Delta a_{22} & \cdots & \Delta a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta a_{n1} & \Delta a_{n2} & \cdots & \Delta a_{nn} \end{pmatrix}.$$

Für die unbekannten Größen $\bar{\boldsymbol{A}}$, $\boldsymbol{\delta A}$, $\bar{\boldsymbol{b}}$, $\boldsymbol{\delta b}$ und $\bar{\boldsymbol{x}}$ machen wir die Ansätze

$$\bar{\boldsymbol{A}} = \begin{pmatrix} \bar{a}_{11} & \bar{a}_{12} & \cdots & \bar{a}_{1n} \\ \bar{a}_{21} & \bar{a}_{22} & \cdots & \bar{a}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{a}_{n1} & \bar{a}_{n2} & \cdots & \bar{a}_{nn} \end{pmatrix}, \quad \boldsymbol{\delta A} = \begin{pmatrix} \delta a_{11} & \delta a_{12} & \cdots & \delta a_{1n} \\ \delta a_{21} & \delta a_{22} & \cdots & \delta a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \delta a_{n1} & \delta a_{n2} & \cdots & \delta a_{nn} \end{pmatrix},$$

$$\bar{\boldsymbol{b}} = \begin{pmatrix} \bar{b}_1 \\ \bar{b}_2 \\ \vdots \\ \bar{b}_n \end{pmatrix}, \quad \boldsymbol{\delta b} = \begin{pmatrix} \delta b_1 \\ \delta b_2 \\ \vdots \\ \delta b_n \end{pmatrix}.$$

Nun definieren wir

$$\delta a_{ij} = \begin{cases} \frac{r_i}{s_i} \Delta a_{ij} \text{sign}(\bar{x}_j) & \text{für} \quad s_i > 0 \\ \\ 0 & \text{für} \quad s_i = 0 \end{cases}$$

und

$$\delta b_i = \begin{cases} -\frac{r_i}{s_i}\Delta b_i & \text{für} \quad s_i > 0 \\ \\ 0 & \text{für} \quad s_i = 0 \end{cases}.$$

Wir haben zu zeigen, dass für die so definierten Größen

$$\bar{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{\delta A}, \quad \bar{\boldsymbol{b}} = \boldsymbol{b} + \boldsymbol{\delta b}$$

einerseits $\bar{\boldsymbol{A}} \in \mathcal{A}$ und $\bar{\boldsymbol{b}} \in \mathcal{B}$ und andererseits $\bar{\boldsymbol{A}}\bar{\boldsymbol{x}} = \bar{\boldsymbol{b}}$ gilt. Die ersten beiden Aussagen folgen sofort aus der Voraussetzung $|\boldsymbol{r}(\bar{\boldsymbol{x}})| \leqq \boldsymbol{s}$. Es folgt $|r_i| \leqq s_i$ für $i = 1, \ldots, n$ und damit

$$|\delta a_{ij}| \leqq \Delta a_{ij}, \quad i, j = 1, \ldots, n$$

und

$$|\delta b_i| \leqq \Delta b_i, \quad i = 1, \ldots, n.$$

Damit folgt aber $|\boldsymbol{\delta A}| \leqq \Delta \boldsymbol{A}$ und $|\boldsymbol{\delta b}| \leqq \Delta \boldsymbol{b}$, daher $\bar{\boldsymbol{A}} \in \mathcal{A}$ und $\bar{\boldsymbol{b}} \in \mathcal{B}$. Weiterhin folgt

$$\boldsymbol{r}(\bar{\boldsymbol{x}}) = \boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{x}} = (\bar{\boldsymbol{b}} - \boldsymbol{\delta b}) - (\bar{\boldsymbol{A}} - \boldsymbol{\delta A})\bar{\boldsymbol{x}} = \bar{\boldsymbol{b}} - \bar{\boldsymbol{A}}\bar{\boldsymbol{x}} - \boldsymbol{\delta b} + \boldsymbol{\delta A}\bar{\boldsymbol{x}}.$$

Für $\boldsymbol{\delta b} - \boldsymbol{\delta A}\bar{\boldsymbol{x}}$ gilt komponentenweise im Falle $s_i > 0$

$$\begin{aligned} (\boldsymbol{\delta b} - \boldsymbol{\delta A}\bar{\boldsymbol{x}})_i &= \delta b_i - \sum_{j=1}^{n} \delta a_{ij}\bar{x}_j = -\frac{r_i}{s_i}\Delta b_i - \sum_{j=1}^{n} \frac{r_i}{s_i}\Delta a_{ij}\text{sign}(\bar{x}_j)\bar{x}_j \\ &= -\frac{r_i}{s_i}\left[\Delta b_i + \sum_{j=1}^{n} \Delta a_{ij}|\bar{x}_j|\right] = -\frac{r_i}{s_i}s_i = -r_i. \end{aligned}$$

Im Falle $s_i = 0$ gilt auch $r_i = 0$, $\delta b_i = 0$ und $\delta a_{ij} = 0$ für $j = 1, \ldots, n$. Damit ist wieder $0 = (\boldsymbol{\delta b} - \boldsymbol{\delta A}\bar{\boldsymbol{x}})_i = -r_i$.
Insgesamt gilt dann $\boldsymbol{\delta b} - \boldsymbol{\delta A}\bar{\boldsymbol{x}} = -\boldsymbol{r}(\bar{\boldsymbol{x}})$ und weiter $\boldsymbol{r}(\bar{\boldsymbol{x}}) = \bar{\boldsymbol{b}} - \bar{\boldsymbol{A}}\bar{\boldsymbol{x}} + \boldsymbol{r}(\bar{\boldsymbol{x}})$, also $\bar{\boldsymbol{A}}\bar{\boldsymbol{x}} = \bar{\boldsymbol{b}}$. ❀

Mit Hilfe des Satzes von PRAGER und OETTLI wird aus der Größe des Residuums auf die Güte der Lösung geschlossen, falls Informationen über die Datenfehler in der Matrix und in der rechten Seite bekannt sind. Das Berechnen des Residuums sollte dabei mit großer Sorgfalt erfolgen (höhere Genauigkeit), da im allgemeinen Auslöschung auftritt. Genügt das Residuum $\boldsymbol{r}(\bar{\boldsymbol{x}})$ der Bedingung

$$|\boldsymbol{r}(\bar{\boldsymbol{x}})| \leqq \Delta \boldsymbol{b} + \Delta \boldsymbol{A}|\bar{\boldsymbol{x}}|,$$

so ist $\bar{\boldsymbol{x}}$ als Lösung zu akzeptieren.
Oft hat man eine spezielle Fehlerkorrelation der Form $\Delta\boldsymbol{A} = \varepsilon|\boldsymbol{A}|$ und $\Delta\boldsymbol{b} = \varepsilon|\boldsymbol{b}|$ vorliegen. Hier sind alle Eingabedaten mit gleicher relativer Genauigkeit gegeben. In diesem Falle ergibt sich

$$|\boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{x}}| \leqq \varepsilon(|\boldsymbol{b}| + |\boldsymbol{A}|\,|\bar{\boldsymbol{x}}|)$$

als Bedingung für die Brauchbarkeit der Lösung $\bar{\boldsymbol{x}}$. Hat man andererseits ein $\bar{\boldsymbol{x}}$ berechnet, so lässt sich über die obige Ungleichung ein kleinstes Fehlerniveau $\underline{\varepsilon}$ berechnen, für das $\bar{\boldsymbol{x}}$ noch als Lösung akzeptierbar ist. Man erhält

$$\underline{\varepsilon} = \max_{i=1,\ldots,n} \frac{|(\boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{x}})_i|}{(|\boldsymbol{b}| + |\boldsymbol{A}|\,|\bar{\boldsymbol{x}}|)_i}.$$

Falls für das wahre Fehlerniveau $\varepsilon < \underline{\varepsilon}$ gilt, ist $\bar{\boldsymbol{x}}$ nicht als Lösung akzeptierbar. Im Falle $\varepsilon \geqq \underline{\varepsilon}$ ist $\bar{\boldsymbol{x}}$ zu akzeptieren.

8.16. Beispiel: Wir betrachten ein Gleichungssystem $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ mit

$$\boldsymbol{A} = \begin{pmatrix} 1.00 & 0.99 \\ 0.99 & 0.98 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} 1.00 \\ 0.99 \end{pmatrix}.$$

Für die Näherungslösung

$$\bar{\boldsymbol{x}} = \begin{pmatrix} 1.099 \\ 0.100 \end{pmatrix}$$

erhält man

$$\boldsymbol{r}(\bar{\boldsymbol{x}}) = \begin{pmatrix} -0.19800 \\ -0.19601 \end{pmatrix}, \quad |\boldsymbol{A}|\,|\bar{\boldsymbol{x}}| + |\boldsymbol{b}| = \begin{pmatrix} 2.19800 \\ 2.17601 \end{pmatrix}.$$

Daraus berechnet man $\underline{\varepsilon} \approx 0.09$. Erst ab einem Fehler der Eingabedaten von 9% oder mehr ist die Näherungslösung akzeptierbar. Für die rechte Seite

$$\boldsymbol{b} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

und die Näherungslösung

$$\bar{\boldsymbol{x}} = \begin{pmatrix} 19600 \\ -20000 \end{pmatrix}$$

erhält man

$$\boldsymbol{r}(\bar{\boldsymbol{x}}) = \begin{pmatrix} 199 \\ 197 \end{pmatrix}, \quad |\boldsymbol{A}|\,|\bar{\boldsymbol{x}}| + |\boldsymbol{b}| = \begin{pmatrix} 39401 \\ 39005 \end{pmatrix}.$$

Hier ergibt sich $\underline{\varepsilon} \approx 0.005$. Schon für relative Fehler der Eingabedaten von 5% oder mehr ist die Lösung zu akzeptieren. ♡

Dieses Beispiel zeigt auch, dass aus einem kleinen Residuum noch nicht auf eine genaue Lösung geschlossen werden darf.

8.2. Direkte Lösungsverfahren

Eine Möglichkeit, ein lineares Gleichungssystem zu lösen, besteht darin, das Lösen auf eine Folge von einfach zu lösenden Gleichungssystemen zurückzuführen. Wir lösen anstelle von

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

Systeme

$$\boldsymbol{A}^{(i)}\boldsymbol{x}^{(i)} = \boldsymbol{b}^{(i)}, \quad i = 1, \ldots, k.$$

Dabei ist die Folge so zu konstruieren, dass einerseits $\boldsymbol{x}^{(k)}$ bei exakter Rechnung die Lösung des ursprünglichen Gleichungssystems ist und andererseits die Systeme mit einem Gesamtaufwand zu lösen sind, der den Aufwand beim Invertieren der Koeffizientenmatrix nicht übersteigt. Als Koeffizientenmatrizen bieten sich dabei Dreiecksmatrizen, Diagonalmatrizen und orthogonale Matrizen an. Ein solches Vorgehen bezeichnen wir als **direktes Lösungsverfahren**. Charakteristisch für ein direktes Lösungsverfahren ist, dass man bei exakter Rechnung nach endlich vielen Rechenschritten die exakte Lösung erhält. Die Anzahl der benötigten Rechenschritte ist dabei a priori abschätzbar.

8.2.1. Die LU-Zerlegung

Schon im alten China nutzte man das heute als GAUSS-Algorithmus bekannte Verfahren zum Lösen linearer Gleichungssysteme. Die Vorgehensweise ist bekannt: Man versucht, durch geeignete Linearkombination von Gleichungen, nach und nach die Variablen zu eliminieren. Bezogen auf die Koeffizientenmatrix $\boldsymbol{A}$ bedeutet das: Man versucht, durch Linearkombination der Zeilen unterhalb der Diagonalen Nullen zu erzeugen. Wir wollen uns einen Transformationsschritt ansehen und gleich überlegen, wie dieser Schritt mit Hilfe geeigneter Transformationsmatrizen darstellbar ist Dabei setzen wir immer eine reguläre Matrix $\boldsymbol{A}$ voraus.
Es sei $\boldsymbol{A}^{(0)} = \boldsymbol{A}$ und $\boldsymbol{b}^{(0)} = \boldsymbol{b}$. Nach k Schritten wurden $\boldsymbol{A}$ und $\boldsymbol{b}$ in $\boldsymbol{A}^{(k)}$ und $\boldsymbol{b}^{(k)}$

transformiert. Die Matrix $\boldsymbol{A}^{(k)}$ besitzt dabei die Struktur

$$\boldsymbol{A}^{(k)} = \begin{pmatrix} a_{11}^{(k)} & \cdots & a_{1k}^{(k)} & a_{1,k+1}^{(k)} & \cdots & a_{1n}^{(k)} \\ 0 & \ddots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \ddots & a_{kk}^{(k)} & a_{k,k+1}^{(k)} & \cdots & a_{kn}^{(k)} \\ 0 & \cdots & 0 & a_{k+1,k+1}^{(k)} & \cdots & a_{k+1,n}^{(k)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{n,k+1}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix} = \left(\begin{array}{c|c} \multicolumn{2}{c}{\boldsymbol{U}^{(k)}} \\ \hline \boldsymbol{O} & \boldsymbol{M}^{(k)} \end{array}\right)$$

mit $\boldsymbol{U}^{(k)} \in \mathbb{R}^{k\times n}$ und $\boldsymbol{M}^{(k)} \in \mathbb{R}^{(n-k)\times(n-k)}$. Nehmen wir nun an, dass $a_{k+1,k+1}^{(k)} \neq 0$ gilt, so sind in der $(k+1)$-ten Spalte unterhalb der Diagonalen Nullen erzeugbar, indem zur i-ten Zeile ($i = k+2,\ldots,n$) von $\boldsymbol{A}^{(k)}$ das $(-a_{i,k+1}^{(k)}/a_{k+1,k+1}^{(k)})$-fache der $(k+1)$-ten Zeile addiert wird.
Bezeichnen wir mit $\boldsymbol{z}_1^{(k)^T},\ldots,\boldsymbol{z}_n^{(k)^T}$ die Zeilen der Matrix $\boldsymbol{A}^{(k)}$, so gilt

$$\boldsymbol{z}_i^{(k+1)^T} = \boldsymbol{z}_i^{(k)^T} = \boldsymbol{e}_i^T \boldsymbol{A}^{(k)}, \quad i = 1,\ldots,k+1$$

und für $i = k+2,\ldots,n$

$$\begin{aligned} \boldsymbol{z}_i^{(k+1)^T} &= \boldsymbol{z}_i^{(k)^T} - \frac{a_{i,k+1}^{(k)}}{a_{k+1,k+1}^{(k)}} \boldsymbol{z}_{k+1}^{(k)^T} \\ &= \boldsymbol{e}_i^T \boldsymbol{A}^{(k)} - \frac{a_{i,k+1}^{(k)}}{a_{k+1,k+1}^{(k)}} \boldsymbol{e}_{k+1}^T \boldsymbol{A}^{(k)} \\ &= \left(\boldsymbol{e}_i^T - \overline{l}_{i,k+1} \boldsymbol{e}_{k+1}^T\right) \boldsymbol{A}^{(k)} \end{aligned}$$

mit

$$\overline{l}_{i,k+1} = \frac{a_{i,k+1}^{(k)}}{a_{k+1,k+1}^{(k)}}.$$

Fasst man diese Gleichungen zusammen, so erhält man

$$\boldsymbol{A}^{(k+1)} = \begin{pmatrix} \boldsymbol{z}_1^{(k+1)^T} \\ \vdots \\ \boldsymbol{z}_n^{(k+1)^T} \end{pmatrix} = \begin{pmatrix} \boldsymbol{e}_1^T \\ \vdots \\ \boldsymbol{e}_{k+1}^T \\ \boldsymbol{e}_{k+2}^T - \bar{l}_{k+2,k+1}\boldsymbol{e}_{k+1}^T \\ \vdots \\ \boldsymbol{e}_n^T - \bar{l}_{n,k+1}\boldsymbol{e}_{k+1}^T \end{pmatrix} \boldsymbol{A}^{(k)} = \boldsymbol{L}_{k+1}(-\bar{\boldsymbol{l}}_{k+1})\boldsymbol{A}^{(k)}.$$

Die Transformationsmatrix $\boldsymbol{L}_{k+1}(-\bar{\boldsymbol{l}}_{k+1})$ ist eine LNT-Matrix. Sie hat die Struktur

$$\boldsymbol{L}_{k+1}(-\bar{\boldsymbol{l}}_{k+1}) = \begin{pmatrix} 1 & & & & & \\ & \ddots & & \boldsymbol{O} & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & -\bar{l}_{k+2,k+1} & 1 & \\ & \boldsymbol{O} & & \vdots & & \ddots \\ & & & -\bar{l}_{n,k+1} & & 1 \end{pmatrix} = \boldsymbol{I} - \bar{\boldsymbol{l}}_{k+1}\boldsymbol{e}_{k+1}^T$$

mit

$$\bar{\boldsymbol{l}}_{k+1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \dfrac{a_{k+2,k+1}^{(k)}}{a_{k+1,k+1}^{(k)}} \\ \vdots \\ \dfrac{a_{n,k+1}^{(k)}}{a_{k+1,k+1}^{(k)}} \end{pmatrix}.$$

Unter der Voraussetzung $a_{k+1,k+1}^{(k)} \neq 0$ lässt sich der $(k+1)$-te Transformationsschritt in der Form

$$\boldsymbol{A}^{(k+1)} = \boldsymbol{L}_{k+1}(-\bar{\boldsymbol{l}}_{k+1})\boldsymbol{A}^{(k)}$$

schreiben. Gilt $a_{k+1,k+1}^{(k)} = 0$, so ist mindestens ein Element $a_{i,k+1}^{(k)}$ mit $i = k+2, \ldots, n$ von Null verschieden. Wäre das nicht der Fall, so wären alle Elemente $a_{i,k+1}^{(k)}$ mit

$i = k+1, \ldots, n$ gleich Null. Damit wäre aber die Matrix $\boldsymbol{A}^{(k)}$ singulär. Da aber alle Transformationsschritte mit regulären LNT-Matrizen durchgeführt wurden, müsste dann auch die Matrix $\boldsymbol{A}$ im Widerspruch zur Voraussetzung singulär sein. Es sei nun $s(k+1) \in \{k+2, \ldots, n\}$ ein Index mit $a^{(k)}_{s(k+1),k+1} \neq 0$. Wir tauschen die Zeilen $k+1$ und $s(k+1)$ in der Matrix $\boldsymbol{A}^{(k)}$ und erhalten eine Matrix[4]

$$\hat{\boldsymbol{A}}^{(k)} = \boldsymbol{T}_{s(k+1),k+1} \boldsymbol{A}^{(k)}.$$

In diesem Falle ergibt sich $\boldsymbol{A}^{(k+1)}$ gemäß

$$\boldsymbol{A}^{(k+1)} = \boldsymbol{L}_{k+1}(-\hat{\boldsymbol{l}}_{k+1}) \hat{\boldsymbol{A}}^{(k)} = \boldsymbol{L}_{k+1}(-\hat{\boldsymbol{l}}_{k+1}) \boldsymbol{T}_{s(k+1),k+1} \boldsymbol{A}^{(k)}.$$

Der Vektor $\hat{\boldsymbol{l}}_{k+1}$ ist durch

$$\hat{\boldsymbol{l}}_{k+1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \frac{\hat{a}^{(k)}_{k+2,k+1}}{\hat{a}^{(k)}_{k+1,k+1}} \\ \vdots \\ \frac{\hat{a}^{(k)}_{n,k+1}}{\hat{a}^{(k)}_{k+1,k+1}} \end{pmatrix}$$

festgelegt, wobei die $\hat{a}^{(k)}_{ij}$ die Elemente der Matrix $\hat{\boldsymbol{A}}^{(k)}$ bezeichnen. Nach $n-1$ Schritten erhält man so eine obere Dreiecksmatrix $\boldsymbol{U} = \boldsymbol{A}^{(n-1)}$. Es gilt

$$\boldsymbol{U} = \boldsymbol{L}_{n-1}(-\hat{\boldsymbol{l}}_{n-1}) \boldsymbol{T}_{s(n-1),n-1} \cdots \boldsymbol{L}_2(-\hat{\boldsymbol{l}}_2) \boldsymbol{T}_{s(2),2} \boldsymbol{L}_1(-\hat{\boldsymbol{l}}_1) \boldsymbol{T}_{s(1),1} \boldsymbol{A}.$$

Im ursprünglichen GAUSS-Algorithmus werden diese Transformationen simultan auf die rechte Seite angewendet. Man erhält

$$\boldsymbol{c} = \boldsymbol{L}_{n-1}(-\hat{\boldsymbol{l}}_{n-1}) \boldsymbol{T}_{s(n-1),n-1} \cdots \boldsymbol{L}_2(-\hat{\boldsymbol{l}}_2) \boldsymbol{T}_{s(2),2} \boldsymbol{L}_1(-\hat{\boldsymbol{l}}_1) \boldsymbol{T}_{s(1),1} \boldsymbol{b}.$$

Damit hat man das Gleichungssystem

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

auf das äquivalente System

$$\boldsymbol{U}\boldsymbol{x} = \boldsymbol{c}$$

[4]Das Element $a_{s(k+1),k+1}$, das nach diesem Tausch auf der Position $(k+1, k+1)$ steht wird als **Pivotelement** bezeichnet.

transformiert. Dies lässt sich wegen der oberen Dreiecksgestalt von $\boldsymbol{U}$ einfach lösen. Diesen Prozess bezeichnet man als **Rücksubstitution**. Einen Algorithmus dazu geben wir am Ende dieses Abschnitts an.
In vielen Anwendungsfällen hat man mehrere Gleichungssysteme mit gleicher Koeffizientenmatrix $\boldsymbol{A}$ aber verschiedenen rechten Seiten $\boldsymbol{b}$ zu lösen. Oft sind die rechten Seiten dieser zu lösenden Gleichungssysteme Funktionen der schon berechneten Lösungen (z.B. bei vielen Iterationsverfahren zur Nullstellenberechnung nichtlinearer Gleichungssysteme). In diesen Fällen ist es günstig, sich die Transformationen $\boldsymbol{L}_i(-\hat{\boldsymbol{l}}_i)\boldsymbol{T}_{s(i),i}$ in geeigneter Weise zu merken. Dazu trennen die eigentlichen Transformationsmatrizen von den Tauschmatrizen. Allgemein gilt für $1 \leqq k < i \leqq j \leqq n$ (siehe Übungsaufgabe 13):

$$\boldsymbol{T}_{ij}\boldsymbol{L}_k(\boldsymbol{l}) = \boldsymbol{L}_k(\boldsymbol{T}_{ij}\boldsymbol{l})\boldsymbol{T}_{ij}.$$

Wir verwenden die Abkürzung

$$\boldsymbol{P}_i = \boldsymbol{T}_{s(n-1),n-1}\boldsymbol{T}_{s(n-2),n-2}\cdots\boldsymbol{T}_{s(i),i}, \quad i = 1,\ldots,n-1,$$

also

$$\boldsymbol{P}_{n-1} = \boldsymbol{T}_{s(n-1),n-1}, \quad \boldsymbol{P}_i = \boldsymbol{P}_{i+1}\boldsymbol{T}_{s(i),i}, \quad i = n-2,\ldots,1.$$

Dann gilt auch für $1 \leqq k < i \leqq n$

$$\boldsymbol{P}_i\boldsymbol{L}_k(\boldsymbol{l}) = \boldsymbol{L}_k(\boldsymbol{P}_i\boldsymbol{l})\boldsymbol{P}_i.$$

Damit folgt

$$\begin{aligned}
&\boldsymbol{L}_{n-1}(-\hat{\boldsymbol{l}}_{n-1})\boldsymbol{T}_{s(n-1),n-1}\boldsymbol{L}_{n-2}(-\hat{\boldsymbol{l}}_{n-2})\boldsymbol{T}_{s(n-2),n-2}\cdots\boldsymbol{L}_1(-\hat{\boldsymbol{l}}_1)\boldsymbol{T}_{s(1),1}\\
&= \boldsymbol{L}_{n-1}(-\hat{\boldsymbol{l}}_{n-1})\boldsymbol{P}_{n-1}\boldsymbol{L}_{n-2}(-\hat{\boldsymbol{l}}_{n-2})\boldsymbol{T}_{s(n-2),n-2}\cdots\boldsymbol{L}_1(-\hat{\boldsymbol{l}}_1)\boldsymbol{T}_{s(1),1}\\
&= \boldsymbol{L}_{n-1}(-\hat{\boldsymbol{l}}_{n-1})\boldsymbol{L}_{n-2}(-\boldsymbol{P}_{n-1}\hat{\boldsymbol{l}}_{n-2})\boldsymbol{P}_{n-2}\cdots\boldsymbol{L}_1(-\hat{\boldsymbol{l}}_1)\boldsymbol{T}_{s(1),1}\\
&\vdots \qquad \vdots\\
&= \boldsymbol{L}_{n-1}(-\hat{\boldsymbol{l}}_{n-1})\boldsymbol{L}_{n-2}(-\boldsymbol{P}_{n-1}\hat{\boldsymbol{l}}_{n-2})\cdots\boldsymbol{L}_2(-\boldsymbol{P}_3\hat{\boldsymbol{l}}_2)\boldsymbol{L}_1(-\boldsymbol{P}_2\hat{\boldsymbol{l}}_1)\boldsymbol{P}_1.
\end{aligned}$$

Insgesamt gilt dann

$$\boldsymbol{U} = \boldsymbol{L}_{n-1}(-\hat{\boldsymbol{l}}_{n-1})\boldsymbol{L}_{n-2}(-\boldsymbol{P}_{n-1}\hat{\boldsymbol{l}}_{n-2})\cdots\boldsymbol{L}_2(-\boldsymbol{P}_3\hat{\boldsymbol{l}}_2)\boldsymbol{L}_1(-\boldsymbol{P}_2\hat{\boldsymbol{l}}_1)\boldsymbol{P}_1\boldsymbol{A}.$$

Setzt man noch

$$\boldsymbol{P} = \boldsymbol{P}_1, \quad \boldsymbol{l}_{n-1} = \hat{\boldsymbol{l}}_{n-1}, \quad \boldsymbol{l}_i = \boldsymbol{P}_{i+1}\hat{\boldsymbol{l}}_i, \quad i = 1,\ldots,n-1,$$

so folgt

$$\boldsymbol{U} = \boldsymbol{L}_{n-1}(-\boldsymbol{l}_{n-1})\boldsymbol{L}_{n-2}(-\boldsymbol{l}_{n-2})\boldsymbol{L}_{n-3}(-\boldsymbol{l}_{n-3})\cdots\boldsymbol{L}_2(-\boldsymbol{l}_2)\boldsymbol{L}_1(-\boldsymbol{l}_1)\boldsymbol{P}\boldsymbol{A}.$$

Multipliziert man diese Gleichung von links mit den Inversen der LNT-Matrizen, so ergibt sich

$$\boldsymbol{P}\boldsymbol{A} = \boldsymbol{L}_1(\boldsymbol{l}_1)\boldsymbol{L}_2(\boldsymbol{l}_2)\cdots\boldsymbol{L}_{n-2}(\boldsymbol{l}_{n-2})\boldsymbol{L}_{n-1}(\boldsymbol{l}_{n-1})\boldsymbol{U}.$$

Nach Übungsaufgabe 14 ist das Produkt der LNT-Matrizen eine untere Einsdreiecksmatrix, und es gilt

$$\boldsymbol{L} = \boldsymbol{L}_1(\boldsymbol{l}_1)\boldsymbol{L}_2(\boldsymbol{l}_2)\cdots\boldsymbol{L}_{n-1}(\boldsymbol{l}_{n-2})\boldsymbol{L}_{n-2}(\boldsymbol{l}_{n-1}) = \boldsymbol{I} + (\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_{n-1}, \boldsymbol{o}).$$

Man erhält somit die Matrix $\boldsymbol{L}$ **ohne** zusätzlichen Rechenaufwand aus den Vektoren $\boldsymbol{l}_1, \ldots, \boldsymbol{l}_{n-1}$. Damit haben wir konstruktiv folgenden Satz bewiesen.

8.17. Satz: *Zu jeder regulären (n,n)-Matrix $\boldsymbol{A}$ existiert eine Permutationsmatrix $\boldsymbol{P}$ derart, dass sich $\boldsymbol{P}\boldsymbol{A}$ eindeutig in das Produkt aus einer unteren Einsdreiecksmatrix $\boldsymbol{L}$ und einer regulären oberen Dreiecksmatrix $\boldsymbol{U}$ zerlegen lässt:*

$$\boldsymbol{P}\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$$
$$= \begin{pmatrix} 1 & & & & \\ l_{21} & 1 & & \boldsymbol{O} & \\ \vdots & \vdots & \ddots & & \\ l_{n-1,1} & l_{n-1,2} & \cdots & 1 & \\ l_{n,1} & l_{n,2} & \cdots & l_{n,n-1} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1,n-1} & u_{1n} \\ & u_{22} & \cdots & u_{2,n-1} & u_{2n} \\ & & \ddots & \vdots & \vdots \\ & \boldsymbol{O} & & u_{n-1,n-1} & u_{n-1,n} \\ & & & & u_{nn} \end{pmatrix}$$

mit $u_{ii} \neq 0$ für $i = 1, \ldots, n$.

Beweis: Es ist noch die Eindeutigkeit der Zerlegung zu zeigen. Nehmen wir dazu an, es gelte

$$\boldsymbol{P}\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U} = \bar{\boldsymbol{L}}\bar{\boldsymbol{U}}$$

mit von $\boldsymbol{L}$ und $\boldsymbol{U}$ verschiedenen Matrizen $\bar{\boldsymbol{L}}$ und $\bar{\boldsymbol{U}}$. Dann folgt

$$\bar{\boldsymbol{L}}^{-1}\boldsymbol{L} = \bar{\boldsymbol{U}}\boldsymbol{U}^{-1} = \boldsymbol{D}.$$

Nach Übungsaufgabe 10 ist $\bar{\boldsymbol{L}}^{-1}\boldsymbol{L}$ untere Einsdreiecksmatrix. Analog ist $\bar{\boldsymbol{U}}\boldsymbol{U}^{-1}$ obere Dreiecksmatrix. Damit ist die Matrix $\boldsymbol{D}$ gleichzeitig untere Einsdreiecksmatrix und obere Dreiecksmatrix, daher gilt $\boldsymbol{D} = \boldsymbol{I}$, daher $\bar{\boldsymbol{L}} = \boldsymbol{L}$ und $\bar{\boldsymbol{U}} = \boldsymbol{U}$. ❀

Hat man einmal von einer Matrix $\boldsymbol{A}$ eine LU-Zerlegung $\boldsymbol{A} = \boldsymbol{P}^T\boldsymbol{L}\boldsymbol{U}$ berechnet, so lässt sich jedes Gleichungssystem $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ lösen, indem man es auf das Lösen von zwei einfacheren Systemen zurückführt. Man hat dabei die beiden Systeme

$$\boldsymbol{L}\boldsymbol{y} = \boldsymbol{P}\boldsymbol{b}, \quad \boldsymbol{U}\boldsymbol{x} = \boldsymbol{y}$$

zu lösen. Dabei wird das Lösen des ersten Systems als Vorwärtssubstitution und das Lösen des zweiten Systems, wie schon erwähnt, als Rückwärtssubstitution bezeichnet.
Der bei der Zerlegung der Matrix $\boldsymbol{A}$ freiwerdende Speicherplatz (alle Elemente unterhalb der hauptdiagonalen) darf genutzt werden, um die Elemente der unteren Einsdreiecksmatrix $\boldsymbol{L}$ zu speichern. Die Zeilenvertauschungen (Permutationsmatrix $\boldsymbol{P}$) merkt man sich auf einem Feld p der Länge n; der Index $p(i)$ gibt dann die wahre Nummer der Zeile an, die aktuell auf der Position i steht.

8.18. LU**-Zerlegung mit Pivotisierung und explizitem Zeilentausch:**
Es ist die reguläre (n,n)-Matrix $\boldsymbol{A}$ in ein Produkt $\boldsymbol{P}\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$ mit einer Permutationsmatrix $\boldsymbol{P}$, einer unteren Einsdreiecksmatrix $\boldsymbol{L}$ und einer oberen Dreiecksmatrix $\boldsymbol{U}$ zu zerlegen. Die Permutationsmatrix $\boldsymbol{P}$ sei dabei auf einem Feld p der Länge n gespeichert.

{Initialisierung}
for $i = 1$ ***to*** n ***do***
  $p(i) = i$
endfor
{LU-Zerlegung}
for $j = 1$ ***to*** $n-1$ ***do***
  {Pivotsuche}
  Wähle einen Index $i^* \in \{\, i \mid a_{ij} \neq 0, \quad i = j, \ldots, n \,\}$.
  {Zeilentausch}
  if $i^* \neq j$ ***then***
    $pp = p(i^*)$; $p(i^*) = p(j)$; $p(j) = pp$
    for $k = 1$ ***to*** n ***do***
      $aa = a_{i^*k}$; $a_{i^*k} = a_{jk}$; $a_{jk} = aa$
    endfor
  endif
  {Transformation der Restmatrix}
  for $i = j+1$ ***to*** n ***do***
    $a_{ij} = a_{ij}/a_{jj}$
    for $k = j+1$ ***to*** n ***do***
      $a_{ik} = a_{ik} - a_{ij} \cdot a_{jk}$
    endfor

endfor
endfor

Aufwand: $\sim n^3/3$ *Additionen/Multiplikationen*

Am Ende des Algorithmus stehen die wesentlichen Elemente der Matrix $\boldsymbol{L}$ auf dem unteren Dreieck der Matrix $\boldsymbol{PA}$. Das obere Dreieck einschließlich der Diagonalen wird von der Matrix $\boldsymbol{U}$ eingenommen:

$$\begin{pmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1,n-2} & u_{1,n-1} & u_{1n} \\ l_{21} & u_{22} & u_{23} & \cdots & u_{2,n-2} & u_{2,n-1} & u_{2n} \\ l_{31} & l_{32} & u_{33} & \cdots & u_{3,n-2} & u_{3,n-1} & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ l_{n-2,1} & l_{n-2,2} & l_{n-2,3} & \cdots & u_{n-2,n-2} & u_{n-2,n-1} & u_{n-2,n} \\ l_{n-1,1} & l_{n-1,2} & l_{n-1,3} & \cdots & l_{n-1,n-2} & u_{n-1,n-1} & u_{n-1,n} \\ l_{n1} & l_{n2} & l_{n3} & \cdots & l_{n,n-2} & l_{n,n-1} & u_{nn} \end{pmatrix}.$$

8.19. Vorwärtssubstitution bei explizitem Zeilentausch:
Es ist das lineare Gleichungssystem $\boldsymbol{Ly} = \boldsymbol{Pb}$ *mit der regulären unteren Einsdreiecksmatrix* $\boldsymbol{L}$*, der Permutationsmatrix* $\boldsymbol{P}$ *und einem beliebigen Vektor* $\boldsymbol{b} \in \mathbb{R}^n$ *zu lösen. Die Matrix* $\boldsymbol{L}$ *ist auf dem unteren Dreieck der Matrix* $\boldsymbol{A}$*, die Permutationsmatrix ist auf einem Feld* p *der Länge* n *gespeichert.*

for $j = 1$ ***to*** n ***do***
  $y_j = b_{p(j)}$
endfor
for $j = 1$ ***to*** n ***do***
  for $i = 1$ ***to*** $j - 1$ ***do***
    $y_i = y_i - a_{ij} \cdot y_j$
  endfor
endfor

Aufwand: $\sim n^2/2$ *Additionen/Multiplikationen*

8.20. Rücksubstitution bei explizitem Zeilentausch:
Es ist das lineare Gleichungssystem $\boldsymbol{Ux} = \boldsymbol{y}$ *mit der regulären oberen Dreiecksmatrix* $\boldsymbol{U}$ *und einem beliebigen Vektor* $\boldsymbol{y} \in \mathbb{R}^n$ *zu lösen. Die Matrix* $\boldsymbol{U}$ *ist auf dem oberen Dreieck der Matrix* $\boldsymbol{A}$ *gespeichert.*

for $j = n$ ***to*** 1 ***step*** -1 ***do***
  $x_j = y_j / a_{jj}$
  for $i = 1$ ***to*** $j - 1$ ***do***
    $y_i = y_i - a_{ij} \cdot x_j$

endfor
endfor

Aufwand: $\sim n^2/2$ *Additionen/Multiplikationen*

Es ist nicht notwendig, den Zeilentausch bei der Zerlegung der Matrix explizit durchzuführen. In der Permutationsmatrix, bzw. im Permutationsvektor ist alle wesentliche Information enthalten. Dieses Vorgehen bezeichnet man als **fiktiven Zeilentausch**. Man erhält folgende Algorithmen.

8.21. LU**-Zerlegung, Pivotisierung und fiktiver Zeilentausch:**
Es ist die reguläre Matrix $\boldsymbol{A}$ *in ein Produkt* $\boldsymbol{PA} = \boldsymbol{LU}$ *mit einer Permutationsmatrix* $\boldsymbol{P}$*, einer unteren Einsdreiecksmatrix* $\boldsymbol{L}$ *und einer oberen Dreiecksmatrix* $\boldsymbol{U}$ *zu zerlegen. Die Permutationsmatrix* $\boldsymbol{P}$ *sei dabei auf einem Feld* p *der Länge* n *gespeichert.*

{Initialisierung}
for $j = 1$ ***to*** n ***do***
  $p(i) = i$
endfor
{LU*-Zerlegung}*
for $j = 1$ ***to*** $n-1$ ***do***
  {Pivotsuche}
  Wähle einen Index

$$i^* \in \left\{ i \;\middle|\; a_{p(i),j} \neq 0, \quad i = j, \ldots, n \right\}.$$

  {Fiktiver Zeilentausch}
  if $i^* \neq j$ ***then***
    $pp = p(i^*)$; $p(i^*) = p(j)$; $p(j) = pp$
  endif
  {Transformation der Restmatrix}
  for $i = j+1$ ***to*** n ***do***
    $a_{p(i),j} = a_{p(i),j} / a_{p(j),j}$
    for $k = j+1$ ***to*** n ***do***
      $a_{p(i),k} = a_{p(i),k} - a_{p(i),j} \cdot a_{p(j),k}$
    endfor
  endfor
endfor

Aufwand: $\sim n^3/3$ *Additionen/Multiplikationen*

Nach diesem Algorithmus liefert das Feld p die Reihenfolge, in der die Zeilen der Matrix verarbeitet wurden: Die Zeile $p(i)$ wurde im i-ten Schritt als Pivotzeile verwendet. Die gleiche Reihenfolge der Zeilen und damit auch der Komponenten der rechten Seite ist natürlich bei der Vorwärtssubstitution zu beachten.

8.22. Vorwärtssubstitution bei fiktivem Zeilentausch:
Es ist das lineare Gleichungssystem $\boldsymbol{Ly} = \boldsymbol{Pb}$ mit der regulären unteren Einsdreiecksmatrix $\boldsymbol{L}$, der Permutationsmatrix $\boldsymbol{P}$ und einem beliebigen Vektor $\boldsymbol{b} \in \mathbb{R}^n$ zu lösen. Die Matrix $\boldsymbol{L}$ ist auf dem unteren Dreieck der Matrix $\boldsymbol{PA}$, die Permutationsmatrix ist auf einem Feld p der Länge n gespeichert.

for $i = 1$ ***to*** n ***do***
  $y_i = b_{p(i)}$
  for $j = 1$ ***to*** $i - 1$ ***do***
    $y_i = y_i - a_{p(i)j} \cdot y_j$
  endfor
endfor

Aufwand: $\sim n^2/2$ Additionen/Multiplikationen

8.23. Rücksubstitution bei fiktivem Zeilentausch:
Es ist das lineare Gleichungssystem $\boldsymbol{Ux} = \boldsymbol{y}$ mit der regulären oberen Dreiecksmatrix $\boldsymbol{U}$ und einem beliebigen Vektor $\boldsymbol{y} \in \mathbb{R}^n$ zu lösen. Die Matrix $\boldsymbol{U}$ ist auf dem oberen Dreieck der Matrix $\boldsymbol{A}$ gespeichert.

for $j = n$ ***to*** 1 ***step*** -1 ***do***
  $x_j = y_j / a_{p(j)j}$
  for $i = 1$ ***to*** $j - 1$ ***do***
    $y_i = y_i - a_{p(i),j} \cdot x_j$
  endfor
endfor

Aufwand: $\sim n^2/2$ Additionen/Multiplikationen

Bemerkung: In den Algorithmen zur LU-Zerlegung müsste in der Praxis noch der Fall berücksichtigt werden, dass kein geeignetes Pivotelement vorhanden ist. Der Schritt **Pivotsuche** ist dazu etwa in folgender Form abzuändern:

> **Pivotsuche:** Bestimme einen Index $i^* \in \{j, \ldots, n\}$ mit $|a_{i^*j}| \geqq \varepsilon$. Gilt für alle $i \in \{j, \ldots, n\}$ $|a_{ij}| < \varepsilon$, so ist die Matrix numerisch singulär. STOPP

8.2.2. Rundungsfehleranalyse der LU-Zerlegung

Führen wir die LU-Zerlegung auf einem realen Rechner durch, so erhalten wir i. a. nicht die exakten Dreiecksfaktoren $\boldsymbol{L}$ und $\boldsymbol{U}$. Durch den Einfluss von Rundungsfehlern ergeben sich nur Näherungen $\boldsymbol{\mathcal{L}}$ und $\boldsymbol{\mathcal{U}}$. Falls die LU-Zerlegung durchführbar war (kein Abbruch wegen numerischer Singularität), ist die berechnete Matrix $\boldsymbol{\mathcal{U}}$ regulär. Die Matrix $\boldsymbol{\mathcal{L}}$ ist als Einsdreiecksmatrix ohnehin regulär. Wir fassen in diesem

Falle $\boldsymbol{\mathcal{L}}$ und $\boldsymbol{\mathcal{U}}$ als exakte Zerlegung einer gestörten Matrix $\boldsymbol{A}+\boldsymbol{\delta A}$ auf. Es gilt dann

$$\boldsymbol{\mathcal{P}}^T\boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{U}} = \boldsymbol{A}+\boldsymbol{\delta A}.$$

Die Matrix $\boldsymbol{\mathcal{P}}$ ist wieder eine Permutationsmatrix. Sie wird sich von der exakten Permutationsmatrix unterscheiden, da der Rundungsfehlereinfluss dazu führen kann, dass die Zeilen der Matrix in anderer Reihenfolge verarbeitet werden. Die Größe der Störung $\boldsymbol{\delta A}$ wollen wir nun abschätzen. Wir führen eine Rückwärtsanalyse der LU-Zerlegung durch. Dazu nehmen wir der Einfachheit halber an, dass während des Algorithmus kein Zeilentausch notwendig ist. Die Zeilen der Matrix seien daher schon in der Reihenfolge geordnet, in der sie verarbeitet werden. Dann gilt $\boldsymbol{\mathcal{P}} = \boldsymbol{P} = \boldsymbol{I}$. Wir erhalten damit

$$\boldsymbol{\delta A} = \boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{U}} - \boldsymbol{A}.$$

Schauen wir uns nun den k-ten Transformationsschritt an. Es gilt

$$\boldsymbol{A}^{(k)} = \boldsymbol{L}_k(-\boldsymbol{l}_k)\boldsymbol{A}^{(k-1)}.$$

In Blockschreibweise erhält man

$$\boldsymbol{A}^{(k)} = \left(\begin{array}{c|c} \boldsymbol{I}^{(k-1)} & \boldsymbol{O} \\ \hline \boldsymbol{O} & \boldsymbol{L}_1(-\bar{\boldsymbol{l}}_k) \end{array}\right)\left(\begin{array}{c|c} \boldsymbol{U}^{(k-1)} & \\ \hline \boldsymbol{O} & \boldsymbol{M}^{(k-1)} \end{array}\right)$$

mit $\boldsymbol{L}_1(-\bar{\boldsymbol{l}}_k) = \boldsymbol{I} - \bar{\boldsymbol{l}}_k\boldsymbol{e}_1^T \in \mathbb{R}^{(n-k+1)\times(n-k+1)}$ und

$$\bar{\boldsymbol{l}}_k = \begin{pmatrix} 0 \\ l_{k+1,k} \\ \vdots \\ l_{nk} \end{pmatrix} \in \mathbb{R}^{n-k+1} \quad \text{falls} \quad \boldsymbol{l}_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ l_{k+1,k} \\ \vdots \\ l_{nk} \end{pmatrix} \in \mathbb{R}^n.$$

$\boldsymbol{I}^{(k-1)}$ bezeichnet die Einheitsmatrix im $\mathbb{R}^{(k-1)\times(k-1)}$. Aus der obigen Blockdarstellung folgt

$$\boldsymbol{A}^{(k)} = \left(\begin{array}{c|c} \boldsymbol{U}^{(k-1)} & \\ \hline \boldsymbol{O} & \bar{\boldsymbol{M}}^{(k-1)} \end{array}\right) = \left(\begin{array}{c|c} \boldsymbol{U}^{(k-1)} & \\ \hline \boldsymbol{O} & \boldsymbol{L}_1(-\bar{\boldsymbol{l}}_k)\boldsymbol{M}^{(k-1)} \end{array}\right).$$

Die Transformation wirkt nur auf die Restmatrix $\boldsymbol{M}^{(k-1)}$. Wir haben daher nur Transformationen der Art

$$\bar{\boldsymbol{M}} = \boldsymbol{L}_1(-\boldsymbol{l})\boldsymbol{M}$$

mit

$$\boldsymbol{M} = \begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1r} \\ m_{21} & m_{22} & \cdots & m_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ m_{r1} & m_{r2} & \cdots & m_{rr} \end{pmatrix} \in \mathbb{R}^{r\times r},$$

$$\boldsymbol{L}_1(-\boldsymbol{l}) = \boldsymbol{I} - \boldsymbol{l}\boldsymbol{e}_1^T \in \mathbb{R}^{r\times r}, \quad \boldsymbol{l} = \begin{pmatrix} 0 \\ \frac{m_{21}}{m_{11}} \\ \vdots \\ \frac{m_{r1}}{m_{11}} \end{pmatrix} \in \mathbb{R}^r$$

zu betrachten. Führen wir die Transformation auf einem realen Rechner aus, so erhalten wir statt der Matrix $\bar{\boldsymbol{M}}$ eine Matrix $\tilde{\boldsymbol{M}}$ mit

$$\tilde{\boldsymbol{M}} = \text{gl}\,(\boldsymbol{L}_1(-\boldsymbol{l})\boldsymbol{M}).$$

Diese Matrix fassen wir nun als exaktes Transformationsergebnis einer gestörten Matrix $\hat{\boldsymbol{M}} = \boldsymbol{M} + \boldsymbol{\delta M}$ auf. Es sei also

$$\tilde{\boldsymbol{M}} = \boldsymbol{L}_1(-\hat{\boldsymbol{l}})\hat{\boldsymbol{M}}$$

mit

$$\hat{\boldsymbol{l}} = \begin{pmatrix} 0 \\ \frac{\hat{m}_{21}}{\hat{m}_{11}} \\ \vdots \\ \frac{\hat{m}_{r1}}{\hat{m}_{11}} \end{pmatrix}.$$

Betrachten wir zunächst nur die zweite Darstellung von $\tilde{\boldsymbol{M}}$. Es gilt

$$\begin{aligned} \tilde{\boldsymbol{M}} &= \boldsymbol{L}_1(-\hat{\boldsymbol{l}})\hat{\boldsymbol{M}} = (\boldsymbol{I} - \hat{\boldsymbol{l}}\boldsymbol{e}_1^T)\hat{\boldsymbol{M}} = \hat{\boldsymbol{M}} - \hat{\boldsymbol{l}}\boldsymbol{e}_1^T\hat{\boldsymbol{M}} \\ &= \begin{pmatrix} \hat{m}_{11} & \hat{m}_{12} & \cdots & \hat{m}_{1r} \\ \hat{m}_{21} & \hat{m}_{22} & \cdots & \hat{m}_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{m}_{r1} & \hat{m}_{r2} & \cdots & \hat{m}_{rr} \end{pmatrix} - \begin{pmatrix} 0 \\ \frac{\hat{m}_{21}}{\hat{m}_{11}} \\ \vdots \\ \frac{\hat{m}_{r1}}{\hat{m}_{11}} \end{pmatrix} (\hat{m}_{11}, \hat{m}_{12}, \ldots, \hat{m}_{1r}), \end{aligned}$$

also

$$\tilde{\boldsymbol{M}} = \begin{pmatrix} \hat{m}_{11} & \hat{m}_{12} & \cdots & \hat{m}_{1r} \\ 0 & \hat{m}_{22} - \frac{\hat{m}_{21}\hat{m}_{12}}{\hat{m}_{11}} & \cdots & \hat{m}_{2r} - \frac{\hat{m}_{21}\hat{m}_{1r}}{\hat{m}_{11}} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \hat{m}_{r2} - \frac{\hat{m}_{r1}\hat{m}_{12}}{\hat{m}_{11}} & \cdots & \hat{m}_{rr} - \frac{\hat{m}_{r1}\hat{m}_{1r}}{\hat{m}_{11}} \end{pmatrix}. \tag{8.1}$$

Andererseits werden die Elemente von $\tilde{\boldsymbol{M}}$ aus den Elementen von $\boldsymbol{M}$ nach den Formeln

$$\begin{array}{rcll} \tilde{m}_{1j} & = & m_{1j} & \text{für } j = 1,\ldots,r \\ \tilde{m}_{i1} & = & 0 & \text{für } i = 2,\ldots,r \\ \tilde{m}_{ij} & = & \text{gl}\left(m_{ij} - \frac{m_{i1}m_{1j}}{m_{11}}\right) & \text{für } i,j = 2,\ldots,r \end{array} \tag{8.2}$$

berechnet. Ein Vergleich von 8.1 und 8.2 liefert:

- Erste Zeile von $\tilde{\boldsymbol{M}}$:

$$\left.\begin{array}{rcl} \hat{m}_{1j} & = & m_{1j} \\ m_{1j} + \delta m_{1j} & = & m_{1j} \\ \delta m_{1j} & = & 0 \end{array}\right\}, \quad j = 1,\ldots,r.$$

- Rest der ersten Spalte von $\tilde{\boldsymbol{M}}$:

  $0 = 0.$

- Restmatrix von $\tilde{\boldsymbol{M}}$:

  $$\hat{m}_{ij} - \frac{\hat{m}_{i1}\hat{m}_{1j}}{\hat{m}_{11}} = \text{gl}\left(m_{ij} - \frac{m_{i1}m_{1j}}{m_{11}}\right), \quad i,j = 2,\ldots,r.$$

  Wegen $\delta m_{1j} = 0$ für $j = 1,\ldots,r$ folgt daraus

  $$m_{ij} + \delta m_{ij} - (m_{i1} + \delta m_{i1})\frac{m_{1j}}{m_{11}} = \left[m_{ij} - \frac{m_{i1}}{m_{11}}(1+\varepsilon_i)m_{1j}(1+\varrho_{ij})\right](1+\vartheta_{ij})$$

  für $i,j = 2,\ldots,r$. Von den Rundungsfehlern nehmen wir die Gültigkeit von

  $$|\varepsilon_i|, |\varrho_{ij}|, |\vartheta_{ij}| \leqq \text{eps}, \quad i,j = 2,\ldots,r$$

  an.

Wir erhalten so $(r-1)^2$ Gleichungen zur Festlegung der restlichen $r(r-1)$ Störungen δm_{ij}, $i=1,\ldots,r$, $j=2,\ldots,r$. Es bleiben daher noch $r-1$ Freiheitsgrade bei der Belastung der einzelnen Element von $\boldsymbol{M}$ mit Störungen. Wir setzen

$$\delta m_{i1} = m_{i1}\varepsilon_i, \quad i=2,\ldots,r.$$

Damit gilt

$$|\delta m_{i1}| \stackrel{\leq}{=} |m_{i1}|\text{eps}, \quad i=2,\ldots,r$$

und

$$\text{gl}\,(l_i) = \text{gl}\left(\frac{m_{i1}}{m_{11}}\right) = \frac{m_{i1}}{m_{11}}(1+\varepsilon_i) = \frac{m_{i1}+\delta m_{i1}}{m_{11}} = \hat{l}_i, \quad i=2,\ldots,r.$$

Es folgt

$$m_{ij}+\delta m_{ij} - \frac{m_{i1}m_{1j}}{m_{11}}(1+\varepsilon_i) = \left[m_{ij} - \frac{m_{i1}}{m_{11}}(1+\varepsilon_i)m_{1j}(1+\varrho_{ij})\right](1+\vartheta_{ij})$$

für $i,j=2,\ldots,r$. Diese Gleichungen lösen wir nach den δm_{ij} auf und erhalten:

$$\delta m_{ij} = m_{ij}\vartheta_{ij} - \frac{m_{i1}m_{1j}}{m_{11}}(1+\varepsilon_i)\left[(1+\varrho_{ij})(1+\vartheta_{ij})-1\right].$$

In erster Näherung folgt daraus

$$\begin{aligned}
\delta m_{ij} &\stackrel{\cdot}{=} m_{ij}\vartheta_{ij} - \frac{m_{i1}m_{1j}}{m_{11}}(1+\varepsilon_i)(\varrho_{ij}+\vartheta_{ij}) && (8.3)\\
&\stackrel{\cdot}{=} m_{ij}\vartheta_{ij} - l_i m_{1j}(\varrho_{ij}+\vartheta_{ij}) && (8.4)\\
|\delta m_{ij}| &\stackrel{\cdot}{\leq} (|m_{ij}|+2|l_i|\,|m_{1j}|)\text{eps}. && (8.5)
\end{aligned}$$

In dieser Abschätzung taucht noch l_i als Element der exakten Transformationsmatrix, also ein Element, das wir eigentlich nicht kennen, auf. Es läst sich aber durch bekannte Elemente ersetzen. Aus der Darstellung

$$\begin{aligned}
\tilde{m}_{ij} &= m_{ij}(1+\vartheta_{ij}) - \frac{m_{i1}m_{1j}}{m_{11}}(1+\varepsilon_i)(1+\varrho_{ij})(1+\vartheta_{ij})\\
&= m_{ij}(1+\vartheta_{ij}) - l_i m_{1j}(1+\varepsilon_i)(1+\varrho_{ij})(1+\vartheta_{ij})
\end{aligned}$$

erhalten wir

$$l_i m_{1j} = \frac{m_{ij}}{(1+\varepsilon_i)(1+\varrho_{ij})} - \frac{\tilde{m}_{ij}}{(1+\varepsilon_i)(1+\varrho_{ij})(1+\vartheta_{ij})}. \qquad (8.6)$$

Setzen wir dies in 8.3 ein, so ergibt sich

$$\begin{aligned}
\delta m_{ij} &\doteq m_{ij}\vartheta_{ij} - m_{ij}\frac{(1+\varepsilon_i)(\varrho_{ij}+\vartheta_{ij})}{(1+\varepsilon_i)(1+\varrho_{ij})} + \tilde{m}_{ij}\frac{(1+\varepsilon_i)(\varrho_{ij}+\vartheta_{ij})}{(1+\varepsilon_i)(1+\varrho_{ij})(1+\vartheta_{ij})} \\
&\doteq m_{ij}\vartheta_{ij} - m_{ij}\frac{\varrho_{ij}+\vartheta_{ij}}{(1+\varrho_{ij})} + \tilde{m}_{ij}\frac{\varrho_{ij}+\vartheta_{ij}}{(1+\varrho_{ij})(1+\vartheta_{ij})} \\
&\doteq m_{ij}\vartheta_{ij} - m_{ij}(\varrho_{ij}+\vartheta_{ij})(1-\varrho_{ij}) + \tilde{m}_{ij}(\varrho_{ij}+\vartheta_{ij})(1-\varrho_{ij})(1-\vartheta_{ij}) \\
&\doteq m_{ij}(\vartheta_{ij} - \varrho_{ij} - \vartheta_{ij}) + \tilde{m}_{ij}(\varrho_{ij}+\vartheta_{ij}) \\
&\doteq -\varrho_{ij}m_{ij} + (\varrho_{ij}+\vartheta_{ij})\tilde{m}_{ij} \\
|\delta m_{ij}| &\dot{\leqq} \text{eps}(|m_{ij}| + 2|\tilde{m}_{ij}|).
\end{aligned}$$

Für die gesamte Matrix $\boldsymbol{\delta M}$ erhalten wir so

$$\begin{aligned}
\delta m_{1j} &= 0 && \text{für } j = 1,\ldots,r, \\
|\delta m_{i1}| &\dot{\leqq} \text{eps}\,|m_{i1}| && \text{für } i = 2,\ldots,r, \\
|\delta m_{ij}| &\dot{\leqq} \text{eps}\,(|m_{ij}| + 2|\tilde{m}_{ij}|) && \text{für } i,j = 2,\ldots,r.
\end{aligned}$$

In Matrixschreibweise lautet dieses Ergebnis für die Transformation

$$\boldsymbol{M}^{(k-1)} \longrightarrow \boldsymbol{M}^{(k)}$$

$$\left|\boldsymbol{\delta M}^{(k-1)}\right| \dot{\leqq} \text{eps}\left[|\boldsymbol{M}^{(k-1)}| + 2\begin{pmatrix} 0 & \boldsymbol{o}^T \\ \boldsymbol{o} & \left|\boldsymbol{M}^{(k)}\right| \end{pmatrix}\right].$$

Bezüglich einer absoluten und monotonen Matrixnorm folgt daraus

$$\|\boldsymbol{\delta M}^{(k-1)}\| \dot{\leqq} \text{eps}\left(\|\boldsymbol{M}^{(k-1)}\| + 2\|\boldsymbol{M}^{(k)}\|\right).$$

Betrachten wir nun wieder einen Transformationsschritt für die gesamte Matrix, so interpretieren wir die aus $\boldsymbol{A}^{(k-1)}$ auf einem realen Rechner berechnete Matrix $\boldsymbol{\mathcal{A}}^{(k)}$ als exaktes Transformationsergebnis einer gestörten Matrix $\hat{\boldsymbol{A}}^{(k-1)} = \boldsymbol{A}^{(k-1)} + \boldsymbol{\delta A}^{(k-1)}$. Für die Störung $\boldsymbol{\delta A}^{(k-1)}$ gilt

$$\boldsymbol{\delta A}^{(k-1)} = \begin{pmatrix} \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\delta M}^{(k-1)} \end{pmatrix}$$

und

$$\|\boldsymbol{\delta A}^{(k-1)}\| = \|\boldsymbol{\delta M}^{(k-1)}\|.$$

Insgesamt gilt dann für die berechnete Matrix $\boldsymbol{\mathcal{A}}^{(k)}$

$$\boldsymbol{\mathcal{A}}^{(k)} = \boldsymbol{\mathcal{L}}^{(k-1)}(\boldsymbol{A}^{(k-1)} + \boldsymbol{\delta A}^{(k-1)})$$

mit der berechneten Transformationsmatrix $\boldsymbol{\mathcal{L}}^{(k-1)} = \boldsymbol{L}_{k-1}(-\bar{\boldsymbol{l}}_{k-1})$. Für den Transformationsvektor $\bar{\boldsymbol{l}}_{k-1}$ gilt

$$\begin{aligned}\bar{\boldsymbol{l}}_{k-1} &= \text{gl}\,(\boldsymbol{l}_{k-1}) \\ &= \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \text{gl}\left(\frac{a^{(k-1)}_{k,k-1}}{a^{(k-1)}_{k-1,k-1}}\right) \\ \vdots \\ \text{gl}\left(\frac{a^{(k-1)}_{n,k-1}}{a^{(k-1)}_{k-1,k-1}}\right) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \text{gl}\left(\frac{m^{(k-1)}_{k,k-1}}{m^{(k-1)}_{k-1,k-1}}\right) \\ \vdots \\ \text{gl}\left(\frac{m^{(k-1)}_{n,k-1}}{m^{(k-1)}_{k-1,k-1}}\right) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \frac{m^{(k-1)}_{k,k-1}+\delta m^{(k-1)}_{n,k-1}}{m^{(k-1)}_{k-1,k-1}} \\ \vdots \\ \frac{m^{(k-1)}_{n,k-1}+\delta m^{(k-1)}_{n,k-1}}{m^{(k-1)}_{k-1,k-1}} \end{pmatrix}.\end{aligned}$$

Die aus $\boldsymbol{A}^{(k-1)}$ berechnete Transformationsmatrix $\boldsymbol{\mathcal{L}}^{(k-1)}$ ist demnach als exakte Transformationsmatrix zur gestörten Matrix $\boldsymbol{A}^{(k-1)} + \boldsymbol{\delta A}^{(k-1)}$ interpretierbar. Für die berechnete Matrix $\boldsymbol{\mathcal{U}}$ erhalten wir dann

$$\begin{aligned}\boldsymbol{\mathcal{U}} = \boldsymbol{A}^{(n-1)} &= \boldsymbol{\mathcal{L}}^{(n-1)}\left(\boldsymbol{A}^{(n-2)} + \boldsymbol{\delta A}^{(n-2)}\right) \\ &= \boldsymbol{\mathcal{L}}^{(n-1)}\left[\boldsymbol{\mathcal{L}}^{(n-2)}\left(\boldsymbol{A}^{(n-3)} + \boldsymbol{\delta A}^{(n-3)}\right) + \boldsymbol{\delta A}^{(n-2)}\right] \\ &= \boldsymbol{\mathcal{L}}^{(n-1)}\boldsymbol{\mathcal{L}}^{(n-2)}\left(\boldsymbol{A}^{(n-3)} + \boldsymbol{\delta A}^{(n-3)} + \left(\boldsymbol{\mathcal{L}}^{(n-2)}\right)^{-1}\boldsymbol{\delta A}^{(n-2)}\right).\end{aligned}$$

Die Matrizen $\left(\boldsymbol{\mathcal{L}}^{(n-2)}\right)^{-1}$ und $\boldsymbol{\delta A}^{(n-2)}$ sind von folgender Struktur

$$\left(\boldsymbol{\mathcal{L}}^{(n-2)}\right)^{-1} = \begin{pmatrix} 1 & & & \\ & \ddots & & \boldsymbol{O} \\ & & 1 & \\ & \boldsymbol{O} & \star & 1 \\ & & \star & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\delta A}^{(n-2)} = \begin{pmatrix} \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \star\ \star \\ & \star\ \star \end{pmatrix}.$$

Man erkennt sofort, dass $\left(\boldsymbol{\mathcal{L}}^{(n-2)}\right)^{-1}\boldsymbol{\delta A}^{(n-2)} = \boldsymbol{\delta A}^{(n-2)}$ gilt. Damit folgt

$$\begin{aligned}
\boldsymbol{\mathcal{U}} &= \boldsymbol{\mathcal{L}}^{(n-1)}\boldsymbol{\mathcal{L}}^{(n-2)}\left(\boldsymbol{A}^{(n-3)} + \boldsymbol{\delta A}^{(n-3)} + \boldsymbol{\delta A}^{(n-2)}\right) \\
&= \boldsymbol{\mathcal{L}}^{(n-1)}\boldsymbol{\mathcal{L}}^{(n-2)}\left[\boldsymbol{\mathcal{L}}^{(n-3)}\left(\boldsymbol{A}^{(n-4)} + \boldsymbol{\delta A}^{(n-4)}\right) + \boldsymbol{\delta A}^{(n-3)} + \boldsymbol{\delta A}^{(n-2)}\right] \\
&= \boldsymbol{\mathcal{L}}^{(n-1)}\boldsymbol{\mathcal{L}}^{(n-2)}\boldsymbol{\mathcal{L}}^{(n-3)} \\
&\quad \left[\boldsymbol{A}^{(n-4)} + \boldsymbol{\delta A}^{(n-4)} + \left(\boldsymbol{\mathcal{L}}^{(n-3)}\right)^{-1}\left(\boldsymbol{\delta A}^{(n-3)} + \boldsymbol{\delta A}^{(n-2)}\right)\right].
\end{aligned}$$

Wegen

$$\left(\boldsymbol{\mathcal{L}}^{(n-3)}\right)^{-1} = \begin{pmatrix} 1 & & & & \\ & \ddots & & \boldsymbol{O} & \\ & & 1 & & \\ & & \star & 1 & \\ & \boldsymbol{O} & \star & 0 & 1 \\ & & \star & 0 & 0 & 1 \end{pmatrix}, \; \boldsymbol{\delta A}^{(n-3)} + \boldsymbol{\delta A}^{(n-2)} = \begin{pmatrix} \boldsymbol{O} & & \boldsymbol{O} \\ & \star & \star & \star \\ \boldsymbol{O} & \star & \star & \star \\ & \star & \star & \star \end{pmatrix}$$

gilt wieder

$$\left(\boldsymbol{\mathcal{L}}^{(n-3)}\right)^{-1}\left(\boldsymbol{\delta A}^{(n-3)} + \boldsymbol{\delta A}^{(n-2)}\right) = \boldsymbol{\delta A}^{(n-3)} + \boldsymbol{\delta A}^{(n-2)}.$$

Setzt man diesen Prozess fort, so erhält man

$$\boldsymbol{\mathcal{U}} = \boldsymbol{\mathcal{L}}^{(n-1)}\boldsymbol{\mathcal{L}}^{(n-2)}\cdots\boldsymbol{\mathcal{L}}^{(1)}\left[\boldsymbol{A} + \boldsymbol{\delta A}^{(0)} + \cdots + \boldsymbol{\delta A}^{(n-3)} + \boldsymbol{\delta A}^{(n-2)}\right].$$

Da sich

$$\boldsymbol{\mathcal{L}} = \left(\boldsymbol{\mathcal{L}}^{(1)}\right)^{-1}\cdots\left(\boldsymbol{\mathcal{L}}^{(n-1)}\right)^{-1}$$

ohne weitere Rundungsfehler aus den $\boldsymbol{\mathcal{L}}^{(i)}$ ergibt, folgt

$$\boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{U}} = \boldsymbol{A} + \boldsymbol{\delta A}, \quad \boldsymbol{\delta A} = \boldsymbol{\delta A}^{(0)} + \boldsymbol{\delta A}^{(1)} + \cdots + \boldsymbol{\delta A}^{(n-2)}.$$

Bezüglich einer absoluten und monotonen Matrixnorm ist dann $\boldsymbol{\delta A}$ abschätzbar. Wir erhalten

$$\begin{aligned}
\|\boldsymbol{\delta A}\| &\leqq \sum_{i=0}^{n-2}\|\boldsymbol{\delta A}^{(i)}\| = \sum_{i=0}^{n-2}\|\boldsymbol{\delta M}^{(i)}\| \leqq \text{eps}\sum_{i=0}^{n-2}\left(\|\boldsymbol{M}^{(i)}\| + 2\|\boldsymbol{M}^{(i+1)}\|\right) \\
&= \text{eps}\left[\|\boldsymbol{M}^{(0)}\| + \sum_{i=1}^{n-2}\|\boldsymbol{M}^{(i)}\| + 2\sum_{i=1}^{n-1}\|\boldsymbol{M}^{(i)}\|\right] \\
&\leqq \text{eps}\left[\|\boldsymbol{A}\| + 3\sum_{i=1}^{n-1}\|\boldsymbol{M}^{(i)}\|\right].
\end{aligned}$$

Damit haben wir den folgenden Satz bewiesen.

8.24. Satz: *Der* GAUSS*sche Algorithmus sei für die (n,n)-Matrix $\boldsymbol{A}$ durchführbar. Die berechneten Dreiecksfaktoren seien $\boldsymbol{\mathcal{L}}$ und $\boldsymbol{\mathcal{U}}$. Dann existiert eine Störung $\boldsymbol{\delta A}$ mit*

$$\boldsymbol{\mathcal{L}\mathcal{U}} = \boldsymbol{A} + \boldsymbol{\delta A}.$$

Die Störung genügt der Abschätzung

$$\|\boldsymbol{\delta A}\| \leqq \operatorname{eps} F(\boldsymbol{A}) \|\boldsymbol{A}\|$$

mit der Fehlerkonstanten

$$F(\boldsymbol{A}) = 1 + 3 \sum_{i=1}^{n-1} \frac{\|\boldsymbol{M}^{(i)}\|}{\|\boldsymbol{A}\|}.$$

Der GAUSSsche Algorithmus ist nach diesem Satz gutartig, falls die Kondition der Matrizen $\boldsymbol{M}^{(i)}$ nicht beliebig groß wird. Ist $\mathcal{A}$ eine Klasse von Matrizen, für die der GAUSSsche Algorithmus durchführbar ist und existiert eine Konstante $F < \infty$, so dass für alle Matrizen $\boldsymbol{A} \in \mathcal{A}$ die Ungleichung $F(\boldsymbol{A}) \leqq F$ gilt, so ist die LU-Zerlegung auf $\mathcal{A}$ numerisch gutartig.
Neben den Rundungsfehlern, die beim Berechnen der LU-Zerlegung auftreten, sind auch jene Rundungsfehler zu beachten, die beim Auflösen der Dreieckssysteme entstehen. Wir werden für die Vorwärts- und die Rücksubstitution wieder eine Rückwärtsanalyse durchführen. Es seien $\hat{\boldsymbol{y}}$ und $\hat{\boldsymbol{x}}$ die berechneten Lösungen der Gleichungssysteme

$$\boldsymbol{\mathcal{L}}\,\boldsymbol{y} = \boldsymbol{b}, \quad \boldsymbol{\mathcal{U}}\boldsymbol{x} = \boldsymbol{y}.$$

Wir nehmen nun an, dass $\hat{\boldsymbol{y}}$ und $\hat{\boldsymbol{x}}$ als exakte Lösungen gestörter Systeme interpretierbar sind. Es gelte also

$$(\boldsymbol{\mathcal{L}} + \boldsymbol{\delta L})\hat{\boldsymbol{y}} = \boldsymbol{b}, \quad (\boldsymbol{\mathcal{U}} + \boldsymbol{\delta U})\hat{\boldsymbol{x}} = \hat{\boldsymbol{y}}.$$

Wir beweisen zunächst den folgenden etwas allgemeineren Satz.

8.25. Satz: *Es sei $\hat{\boldsymbol{x}}$ die berechnete Lösung des Gleichungssystems*

$$\boldsymbol{L}\boldsymbol{x} = \boldsymbol{b}$$

mit der regulären unteren Dreiecksmatrix $\boldsymbol{L}$. Dann existiert eine Störung $\boldsymbol{\delta L}$ mit

$$(\boldsymbol{L} + \boldsymbol{\delta L})\hat{\boldsymbol{x}} = \boldsymbol{b}.$$

Die Störung $\boldsymbol{\delta L}$ ist ebenfalls untere Dreiecksmatrix und genügt der elementweisen Abschätzung

$$|\delta l_{ij}| \leqq j\text{eps}\,|l_{ij}|, \quad i,j = 1,\ldots,n.$$

Beweis: Das Lösen des Gleichungssystems $\boldsymbol{Lx} = \boldsymbol{b}$ erfolgt nach dem Algorithmus

for $i = 1$ **to** n **do**
  $x_i = b_i$
  for $j = 1$ **to** $i-1$ **do**
    $x_i = x_i - l_{ij} \cdot x_j$
  end
  $x_i = x_i / l_{ii}$
end

In Gleitpunktarithmetik gilt dann für das Berechnen einer Komponente $\hat{x}_i$:

$\hat{x}_i = b_i$
for $j = 1$ **to** $i-1$ **do**
  $\hat{x}_i = \text{gl}\,(\hat{x}_i - l_{ij}\hat{x}_j) = (\hat{x}_i - l_{ij}\hat{x}_j(1+\varepsilon_j))(1+\delta_j)$
end
$\hat{x}_i = \text{gl}\,(\hat{x}_i / l_{ii}) = \hat{x}_i/(l_{ii}(1+\varepsilon_i))$

Dabei nehmen wir wieder an, dass $|\varepsilon_j|, |\delta_j| \leqq \text{eps}$ für $j = 1,\ldots,i$ gilt. Damit folgt

$$\begin{aligned}
\hat{x}_i &= \frac{\Big(\cdots\Big((b_i - l_{i1}\hat{x}_1(1+\varepsilon_1))(1+\delta_1) - l_{i2}\hat{x}_2(1+\varepsilon_2)\Big)(1+\delta_2) - \cdots - l_{i,i-1}\hat{x}_{i-1}(1+\varepsilon_{i-1})\Big)(1+\delta_{i-1})}{l_{ii}(1+\varepsilon_i)} \\
&= \frac{b_i(1+\delta_1)(1+\delta_2)\cdots(1+\delta_{i-1}) - l_{i1}\hat{x}_1(1+\varepsilon_1)(1+\delta_1)(1+\delta_2)\cdots(1+\delta_{i-1})}{l_{ii}(1+\varepsilon_i)} + \\
&\quad + \frac{-l_{i2}\hat{x}_2(1+\varepsilon_2)(1+\delta_2)\cdots(1+\delta_{i-1}) - \cdots - l_{i,i-1}\hat{x}_{i-1}(1+\varepsilon_{i-1})(1+\delta_{i-1})}{l_{ii}(1+\varepsilon_i)} \\
&= \frac{b_i - l_{i1}\hat{x}_1(1+\varepsilon_1) - l_{i2}\hat{x}_2\frac{1+\varepsilon_2}{1+\delta_1} - \cdots - l_{i,i-1}\hat{x}_{i-1}\frac{1+\varepsilon_{i-1}}{(1+\delta_1)(1+\delta_2)\cdots(1+\delta_{i-2})}}{l_{ii}\frac{1+\varepsilon_i}{(1+\delta_1)(1+\delta_2)\cdots(1+\delta_{i-1})}}.
\end{aligned}$$

Nun gilt in erster Näherung

$$\frac{1+\varepsilon_j}{(1+\delta_1)(1+\delta_2)\cdots(1+\delta_{j-1})} \doteq 1 + \varepsilon_j - \delta_1 - \delta_2 - \cdots - \delta_{j-1}.$$

Damit folgt

$$\hat{x}_i = \frac{b_i - l_{i1}(1+\beta_{i1})\hat{x}_1 - l_{i2}(1+\beta_{i2})\hat{x}_2 - \cdots - l_{i,i-1}(1+\beta_{i1})\hat{x}_{i-1}}{l_{ii}(1+\beta_{ii})} \tag{8.7}$$

mit

$$\begin{aligned}\beta_{ij} &= \frac{1+\varepsilon_j}{(1+\delta_1)(1+\delta_2)\cdots(1+\delta_{j-1})} \\ &\doteq 1+\varepsilon_j-\delta_1-\delta_2-\cdots-\delta_{j-1}\end{aligned}$$

und

$$|\beta_{ij}| \stackrel{\cdot}{\leqq} j\text{eps}.$$

Interpretieren wir die berechnete Lösung $\hat{\boldsymbol{x}}$ als exakte Lösung eines gestörten Systems, so erhalten wir

$$(\boldsymbol{L}+\boldsymbol{\delta L})\hat{\boldsymbol{x}}=\boldsymbol{b}.$$

Hieraus folgt

$$\hat{x}_i=\frac{b_i-(l_{i1}+\delta l_{i1})\hat{x}_1-(l_{i2}+\delta l_{i2})\hat{x}_2-\cdots-(l_{i,i-1}+\delta l_{i1})\hat{x}_{i-1}}{l_{ii}+\delta l_{ii})}. \tag{8.8}$$

Ein Vergleich von 8.7 und 8.8 liefert

$$\delta l_{ij}=l_{ij}\beta_{ij}$$

und

$$|\delta l_{ij}| \stackrel{\cdot}{\leqq} j\text{eps}\,|l_{ij}|$$

für $i,j=1,\ldots,n$. ❀

Analog gilt für die Lösung von Gleichungssystemen mit einer oberen Dreiecksmatrix der folgende Satz.

8.26. Satz: *Es sei $\hat{\boldsymbol{x}}$ die berechnete Lösung des Gleichungssystems*

$$\boldsymbol{U}\boldsymbol{x}=\boldsymbol{b}$$

mit der regulären oberen Dreiecksmatrix $\boldsymbol{U}$. Dann existiert eine Störung $\boldsymbol{\delta U}$ mit

$$(\boldsymbol{U}+\boldsymbol{\delta U})\hat{\boldsymbol{x}}=\boldsymbol{b}.$$

Die Störung $\boldsymbol{\delta U}$ ist ebenfalls obere Dreiecksmatrix und genügt der elementweisen Abschätzung

$$|\delta u_{ij}| \leqq (n-j+1)\text{eps}\,|u_{ij}|, \quad i,j=1,\ldots,n.$$

Mit Hilfe der Sätze 8.24, 8.25 und 8.26 lässt sich der Gesamtrundungsfehler beim Lösen eines linearen Gleichungssystems mit Hilfe einer LU-Zerlegung abschätzen. Die berechnete Lösung $\hat{\boldsymbol{x}}$ ist exakte Lösung des Systems

$$(\boldsymbol{\mathcal{U}}+\boldsymbol{\delta U})\hat{\boldsymbol{x}} = \hat{\boldsymbol{y}}.$$

$\hat{\boldsymbol{y}}$ ist exakte Lösung von

$$(\boldsymbol{\mathcal{L}} + \boldsymbol{\delta L})\hat{\boldsymbol{y}} = \boldsymbol{b}.$$

Insgesamt gilt

$$(\boldsymbol{\mathcal{L}} + \boldsymbol{\delta L})(\boldsymbol{\mathcal{U}}+\boldsymbol{\delta U})\hat{\boldsymbol{x}} = \hat{\boldsymbol{y}}, \tag{8.9}$$

wobei die Störungen $\boldsymbol{\delta L}$ und $\boldsymbol{\delta U}$ den Abschätzungen aus Satz 8.25 bzw. Satz 8.26 genügen. Multipliziert man in Gleichung 8.9 die linke Seite aus, so ergibt sich

$$\begin{aligned}(\boldsymbol{\mathcal{L}\mathcal{U}}+\boldsymbol{\mathcal{L}}\,\boldsymbol{\delta U} + \boldsymbol{\delta L}(\boldsymbol{\mathcal{U}}+\boldsymbol{\delta U}))\hat{\boldsymbol{x}} &= \boldsymbol{b}\\ (\boldsymbol{A}+\boldsymbol{\delta A}+\boldsymbol{\delta A}')\hat{\boldsymbol{x}} &= \boldsymbol{b}.\end{aligned}$$

Die Störung $\boldsymbol{\delta A}$ beinhaltet den Rundungsfehlereinfluss der LU-Zerlegung. Für sie gilt nach Satz 8.24

$$|\boldsymbol{\delta A}| \stackrel{.}{\leq} \text{eps}\, F(\boldsymbol{A})\|\boldsymbol{A}\|.$$

In der Störung $\boldsymbol{\delta A}'$ steckt der Rundungsfehlereinfluss beim Lösen der Dreieckssysteme. Für sie ergibt sich elementweise

$$\begin{aligned}\delta a'_{ij} &= \sum_{k=1}^{\min\{i,j\}} \left[\ell_{ik}\delta u_{kj} + \delta\ell_{ik}(u_{kj}+\delta u_{kj})\right]\\ |\delta a'_{ij}| &= \left|\sum_{k=1}^{\min\{i,j\}} \left[\ell_{ik}\delta u_{kj} + \delta\ell_{ik}(u_{kj}+\delta u_{kj})\right]\right|\\ &\stackrel{.}{\leq} \sum_{k=1}^{\min\{i,j\}} \left[|\ell_{ik}|\,|\delta u_{kj}| + |\delta\ell_{ik}|(|u_{kj}|+|\delta u_{kj}|)\right]\end{aligned}$$

Wendet man die Abschätzungen aus den Sätzen 8.25 und 8.26 an, so folgt

$$\begin{aligned}|\delta a'_{ij}| &\stackrel{.}{\leq} \text{eps} \sum_{k=1}^{\min\{i,j\}} \left[(n-j+1)|\ell_{ik}|\,|u_{kj}| + k|\ell_{ik}|(|u_{kj}|+(n-j+1)\text{eps}\,|u_{kj}|)\right]\\ &\stackrel{.}{\leq} \text{eps} \sum_{k=1}^{\min\{i,j\}} |\ell_{ik}|\,|u_{kj}|\,[n-j+1+k+k(n-j+1)\text{eps}]\\ &\stackrel{.}{\leq} \text{eps} \sum_{k=1}^{\min\{i,j\}} |\ell_{ik}|\,|u_{kj}|(n-j+1+k).\end{aligned}$$

Wegen $k \leqq \min\{i,j\} \leqq j$ gilt $n-j+1+k \leqq n+1$. Damit ergibt sich

$$|\delta a'_{ij}| \stackrel{\cdot}{\leqq} \text{eps}(n+1) \sum_{k=1}^{\min\{i,j\}} |\ell_{ik}|\,|\mathcal{u}_{kj}|.$$

Analog zu Gleichung 8.6 gilt für

$$\ell_{ik}\mathcal{u}_{kj} = \ell_{ik} a_{kj}^{(k-1)}$$

die Abschätzung

$$|\ell_{ik}|\,|\mathcal{u}_{kj}| \stackrel{\cdot}{\leqq} |a_{ij}^{(k)}| + |a_{ij}^{(k-1)}|.$$

Damit erhalten wir

$$\begin{aligned}
|\delta a'_{ij}| &\stackrel{\cdot}{\leqq} \text{eps}(n+1) \left[\sum_{k=1}^{\min\{i,j\}} |a_{ij}^{(k)}| + \sum_{k=1}^{\min\{i,j\}} |a_{ij}^{(k-1)}| \right] \\
&\stackrel{\cdot}{\leqq} \text{eps}(n+1) \left[|a_{ij}^{(0)}| + \sum_{k=1}^{\min\{i,j\}} |a_{ij}^{(k)}| + \sum_{k=1}^{\min\{i,j\}-1} |a_{ij}^{(k)}| \right] \\
&\stackrel{\cdot}{\leqq} \text{eps}(n+1) \left[|a_{ij}^{(0)}| + 2 \sum_{k=1}^{\min\{i,j\}} |a_{ij}^{(k)}| \right].
\end{aligned}$$

In Matrixschreibweise gilt

$$|\boldsymbol{\delta A'}| \stackrel{\cdot}{\leqq} \text{eps}(n+1) \left[|\boldsymbol{A}| + 2 \sum_{k=1}^{n-1} \begin{pmatrix} \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & |\boldsymbol{M}^{(k)}| \end{pmatrix} \right],$$

und bezüglich einer absoluten und monotonen Matrixnorm

$$\|\boldsymbol{\delta A'}\| \stackrel{\cdot}{\leqq} \text{eps}(n+1) \left(\|\boldsymbol{A}\| + 2 \sum_{k=1}^{n-1} \|\boldsymbol{M}^{(k)}\| \right).$$

Damit haben wir folgenden Satz bewiesen.

8.27. Satz: *Der* GAUSS*sche Algorithmus sei für eine (n,n)-Matrix $\boldsymbol{A}$ durchführbar. $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ sei die berechnete Lösung des Gleichungssystems $\boldsymbol{Ax} = \boldsymbol{b}$. Dann existiert eine Störung $\overline{\boldsymbol{\delta A}}$, so dass $\bar{\boldsymbol{x}}$ exakte Lösung des gestörten Systems*

$$(\boldsymbol{A} + \overline{\boldsymbol{\delta A}})\bar{\boldsymbol{x}} = \boldsymbol{b}$$

ist. Für die Störung $\overline{\boldsymbol{\delta A}}$ *gilt*

$$\overline{\boldsymbol{\delta A}} = \boldsymbol{\delta A} + \boldsymbol{\delta A}',$$

wobei $\boldsymbol{\delta A}$ *den Rundungsfehleranteil der* LU*-Zerlegung und* $\boldsymbol{\delta A}'$ *den Rundungsfehleranteil beim Lösen der Dreieckssysteme beschreibt. Es gelten die Abschätzungen*

$$\|\boldsymbol{\delta A}\| \leqq \text{eps}\, F(\boldsymbol{A})\|\boldsymbol{A}\|$$

mit

$$F(\boldsymbol{A}) \doteq 1 + 3\sum_{k=1}^{n-1} \frac{\|\boldsymbol{M}^{(k)}\|}{\|\boldsymbol{A}\|}$$

und

$$\|\boldsymbol{\delta A}'\| \leqq \text{eps}\, F'(\boldsymbol{A})\|\boldsymbol{A}\|$$

mit

$$F'(\boldsymbol{A}) \doteq (n+1)\left(1 + 2\sum_{k=1}^{n-1} \frac{\|\boldsymbol{M}^{(k)}\|}{\|\boldsymbol{A}\|}\right) \stackrel{\cdot}{\leqq} (n+1)F(\boldsymbol{A}).$$

Insgesamt gilt

$$\|\overline{\boldsymbol{\delta A}}\| \leqq \text{eps}\, \bar{F}(\boldsymbol{A})\|\boldsymbol{A}\|$$

mit

$$\bar{F}(\boldsymbol{A}) \doteq (n+2)F(\boldsymbol{A}).$$

Für die Durchführbarkeit des Algorithmus ist

$$\text{eps}\,\text{cond}(\boldsymbol{A})F(\boldsymbol{A}) < 1$$

hinreichend.

Betrachten wir nun noch den erzeugten Rundungsfehler $\boldsymbol{\delta x} = \bar{\boldsymbol{x}} - \boldsymbol{x}$. Für ihn gilt

$$\begin{aligned} \boldsymbol{A\delta x} &= \boldsymbol{A\bar{x}} - \boldsymbol{Ax} = (\boldsymbol{A} + \overline{\boldsymbol{\delta A}} - \overline{\boldsymbol{\delta A}})\bar{\boldsymbol{x}} - \boldsymbol{b} = (\boldsymbol{A} + \overline{\boldsymbol{\delta A}})\bar{\boldsymbol{x}} - \overline{\boldsymbol{\delta A}}\bar{\boldsymbol{x}} - \boldsymbol{b} \\ &= -\overline{\boldsymbol{\delta A}}\bar{\boldsymbol{x}} \quad \boldsymbol{\delta x} = -\boldsymbol{A}^{-1}\overline{\boldsymbol{\delta A}}\bar{\boldsymbol{x}}. \end{aligned}$$

Damit folgt

$$\|\boldsymbol{\delta x}\| \leqq \|\boldsymbol{A}^{-1}\|\,\|\overline{\boldsymbol{\delta A}}\|\,\|\bar{\boldsymbol{x}}\| = \kappa\|\bar{\boldsymbol{x}}\|$$

mit $\kappa = \|\boldsymbol{A}^{-1}\|\,\|\overline{\boldsymbol{\delta A}}\|$. Für $\kappa < 1$ ist $\boldsymbol{A} + \overline{\boldsymbol{\delta A}}$ regulär. Dann ergibt sich

$$\begin{aligned}(\boldsymbol{A}+\overline{\boldsymbol{\delta A}})\boldsymbol{\delta x} &= (\boldsymbol{A}+\overline{\boldsymbol{\delta A}})\overline{\boldsymbol{x}} - (\boldsymbol{A}+\overline{\boldsymbol{\delta A}})\boldsymbol{x} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} - \overline{\boldsymbol{\delta A}}\boldsymbol{x} \\ &= -\overline{\boldsymbol{\delta A}}\boldsymbol{x}, \quad \boldsymbol{\delta x} = -(\boldsymbol{A}+\overline{\boldsymbol{\delta A}})^{-1}\overline{\boldsymbol{\delta A}}\boldsymbol{x}\end{aligned}$$

und weiter

$$\|\boldsymbol{\delta x}\| = \|(\boldsymbol{A}+\overline{\boldsymbol{\delta A}})^{-1}\|\,\|\overline{\boldsymbol{\delta A}}\|\,\|\boldsymbol{x}\| \leqq \frac{\|\boldsymbol{A}^{-1}\|}{1-\kappa}\|\overline{\boldsymbol{\delta A}}\|\,\|\boldsymbol{x}\| = \frac{\kappa}{1-\kappa}\|\boldsymbol{x}\|.$$

Bemerkung: Im allgemeinen sind die Schätzungen für den Rundungsfehlereinfluss beim Lösen der Dreieckssysteme zu pessimistisch. Der erzeugte Rundungsfehler ist oft wesentlich kleiner als die Schranke $\|\boldsymbol{A}^{-1}\|\,\|\overline{\boldsymbol{\delta A}}\|\,\|\overline{\boldsymbol{x}}\|$. Praktische Erfahrungen zeigen, dass die Konstante in Satz 8.27 durch

$$\bar{F} \approx \begin{cases} F, & \text{falls cond}(\boldsymbol{A}) \text{ groß} \\ nF, & \text{falls cond}(\boldsymbol{A}) \text{ klein} \end{cases}$$

ersetzbar ist.

8.2.3. Pivotisierung und Skalierung

Im Schritt **Pivotsuche** der LU-Zerlegung wurde gesichert, dass das Pivotelement a_{i^*j} von Null verschieden ist. Um ein günstiges Fehlerverhalten zu realisieren, ist es nach Satz 8.24 notwendig, die Normen der Matrizen $\boldsymbol{M}^{(i)}$ möglichst klein zu halten. Dies lässt sich durch geschickte Wahl des Pivotelements erreichen. Es gibt im wesentlichen drei Varianten. Dabei müsste der Schritt **Pivotsuche** im Algorithmus zur LU-Zerlegung in folgender Weise abgeändert werden.

- **Spaltenpivotisierung:**
  Bestimme einen Index $i^* \in \{j, \ldots, n\}$, so dass die Ungleichung $|a_{i^*j}| \geqq |a_{ij}|$ für alle $i \in \{j, \ldots, n\}$ gilt.
  Als Pivotelement wird das betragsgrößte Element der Restspalte j genommen.

- **Zeilenpivotisierung:**
  Bestimme einen Index $j^* \in \{j, \ldots, n\}$, so dass die Ungleichung $|a_{jj^*}| \geqq |a_{jk}|$ für alle $k \in \{j, \ldots, n\}$ gilt.
  Als Pivotelement wird das betragsgrößte Element der Restzeile j genommen. Danach erfolgt ein Tausch der Spalten j und j^*, der sich in einem zusätzlichen Permutationsvektor zu merken ist.

- **Totalpivotisierung:**
  Bestimme Indizes $i^*, j^* \in \{j, \ldots, n\}$, so dass die Ungleichung $|a_{i^*j^*}| \geqq |a_{ik}|$ für

alle $i,k \in \{j,\ldots,n\}$ gilt.
Als Pivotelement wird das betragsgrößte Element der Restmatrix genommen. Hier ist ein Zeilen- und ein Spaltentausch vorzunehmen.

Für einen Zerlegungsalgorithmus mit Spaltenpivotisierung wollen wir die Normen der Restmatrizen $\boldsymbol{M}^{(k)}$ abschätzen. Dazu betrachten wir wieder einen Transformationsschritt $\boldsymbol{A}^{(k-1)} \longrightarrow \boldsymbol{A}^{(k)}$. Es gilt

$$a_{ij}^{(k)} = \hat{a}_{ij}^{(k-1)} - \hat{l}_{ik}^{(k-1)} \hat{a}_{kj}^{(k-1)}, \quad i,j = k+1,\ldots,n$$

mit

$$\hat{l}_{ik}^{(k-1)} = \frac{\hat{a}_{ik}^{(k-1)}}{\hat{a}_{kk}^{(k-1)}}.$$

Mit $\hat{a}_{ij}^{(k-1)}$ bezeichnen wir die Elemente der Matrix $\hat{\boldsymbol{A}}^{(k)} = \boldsymbol{T}_{s(k)k}\boldsymbol{A}^{(k)}$. Da bei der Spaltenpivotisierung das Pivotelement $\hat{a}_{kk}^{(k-1)}$ als betragsgrößtes der Restspalte gewählt wird, gilt

$$|\hat{l}_{ik}^{(k-1)}| \leqq 1, \quad i = k+1,\ldots,n.$$

Damit folgt

$$\left|a_{ij}^{(k)}\right| \leqq \left|\hat{a}_{ij}^{(k-1)} - \hat{l}_{ik}^{(k-1)}\hat{a}_{kj}^{(k-1)}\right| \leqq \left|\hat{a}_{ij}^{(k-1)}\right| + \left|\hat{a}_{kj}^{(k-1)}\right| \; i,j = k+1,\ldots,n.$$

Nun lässt sich die ∞-Norm (Zeilensummennorm) von $\boldsymbol{M}^{(k)}$ abschätzen. Wir erhalten:

$$\begin{aligned}
\|\boldsymbol{M}^{(k)}\|_\infty &= \max_{i=k+1,\ldots,n}\left\{\sum_{j=k+1}^{n}\left|a_{ij}^{(k)}\right|\right\} \\
&\leqq \max_{i=k+1,\ldots,n}\left\{\sum_{j=k+1}^{n}\left(\left|\hat{a}_{ij}^{(k-1)}\right| + \left|\hat{a}_{kj}^{(k-1)}\right|\right)\right\} \\
&\leqq \max_{i=k,\ldots,n}\left\{\sum_{j=k}^{n}\left|\hat{a}_{ij}^{(k-1)}\right| + \sum_{j=k}^{n}\left|\hat{a}_{kj}^{(k-1)}\right|\right\} \\
&= \max_{i=k,\ldots,n}\left\{\sum_{j=k}^{n}\left|a_{ij}^{(k-1)}\right| + \sum_{j=k}^{n}\left|a_{s(k)j}^{(k-1)}\right|\right\} \\
&= \|\boldsymbol{M}^{(k-1)}\|_\infty + \sum_{j=k}^{n}\left|a_{s(k)j}^{(k-1)}\right| \leqq 2\|\boldsymbol{M}^{(k-1)}\|_\infty.
\end{aligned}$$

Durch fortlaufende Anwendung der letzten Ungleichung erhält man

$$\|\boldsymbol{M}^{(k)}\|_\infty \leqq 2\|\boldsymbol{M}^{(k-1)}\|_\infty \leqq 4\|\boldsymbol{M}^{(k-2)}\|_\infty \leqq \cdots \leqq 2^k\|\boldsymbol{M}^{(0)}\|_\infty = 2^k\|\boldsymbol{A}\|_\infty.$$

Mit Satz 8.24 folgt dann

$$\|\boldsymbol{\delta A}\|_\infty \leqq \text{eps}F(\boldsymbol{A})\|\boldsymbol{A}\|_\infty$$

mit

$$\begin{aligned} F(\boldsymbol{A}) &= 1+3\sum_{i=1}^{n-1}\frac{\|\boldsymbol{M}^{(i)}\|_\infty}{\|\boldsymbol{A}\|_\infty} \leqq 1+3\sum_{i=1}^{n-1}2^i \\ &= 1+3(2^n-2) = 3\cdot 2^n-5 \leqq 3\cdot 2^n. \end{aligned}$$

Analog erhält man für die Zeilenpivotisierung bezüglich der 1-Norm (Spaltensummennorm) die Abschätzung

$$\|\boldsymbol{\delta A}\|_1 \leqq \text{eps}F(\boldsymbol{A})\|\boldsymbol{A}\|_1$$

mit $F(\boldsymbol{A}) \leqq 3\cdot 2^n$.
Diese Fehlerschranken sind im allgemeinen zu pessimistisch. Für die meisten praktisch auftretenden Gleichungssysteme gilt im Falle der Spalten- oder Zeilenpivotisierung $F(\boldsymbol{A}) \leqq 10n$. Damit reicht die Spaltenpivotisierung für die meisten Fälle zur Stabilisierung des GAUSSschen Algorithmus aus.
Im Falle der der Totalpivotisierung erhält man

$$F(\boldsymbol{A}) \leqq 1.8\cdot n^{2+\ln/4}.$$

Für praktisch auftretende Systeme liegt hier $F(\boldsymbol{A})$ in der Größenordnung von 1. Da aber der Aufwand an Vergleichsoperationen bei der Totalpivotisierung den entsprechenden Aufwand bei der Spalten- oder Zeilenpivotisierung bei weitem übersteigt ($n^3/3 \longleftrightarrow n^2/2$), verwendet man normalerweise nur die Spaltenpivotisierung.

8.28. Beispiel: Wir betrachten das Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ mit

$$\boldsymbol{A} = \begin{pmatrix} 0.005 & 1.000 \\ 1.000 & 1.000 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} 0.5 \\ 1.0 \end{pmatrix}.$$

Die exakte Lösung lautet

$$\boldsymbol{x} = \begin{pmatrix} \frac{500}{995} \\ \frac{495}{995} \end{pmatrix} = \begin{pmatrix} 0.503... \\ 0.497... \end{pmatrix}.$$

Auf einem Rechner mit dem Maschinenzahlbereich $\mathbb{M}(10,2,.,.)$ erhält man ohne Pivotisierung

$$\boldsymbol{A} \Longrightarrow \begin{pmatrix} 1 & 0 \\ 200 & 1 \end{pmatrix} \begin{pmatrix} 0.005 & 1 \\ 0 & -200 \end{pmatrix} = \boldsymbol{\mathcal{L}\mathcal{U}}.$$

Vorwärtselimination und Rücksubstitution liefern

$$\boldsymbol{\mathcal{L}}\,\boldsymbol{y} = \boldsymbol{b} \Longrightarrow \boldsymbol{y} = \begin{pmatrix} 0.5 \\ -99 \end{pmatrix}$$

und

$$\boldsymbol{\mathcal{U}}\hat{\boldsymbol{x}} = \boldsymbol{y} \Longrightarrow \hat{\boldsymbol{x}} = \begin{pmatrix} 0 \\ 0.5 \end{pmatrix}.$$

Mit Spaltenpivotisierung würde das Element a_{21} als Pivotelement gewählt werden. Mit der Permutationsmatrix

$$\boldsymbol{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

ergibt sich

$$\boldsymbol{P}\boldsymbol{A} \Longrightarrow \begin{pmatrix} 1 & 0 \\ 0.005 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \bar{\boldsymbol{\mathcal{L}}}\bar{\boldsymbol{\mathcal{U}}}.$$

Als Lösung erhalten wir:

$$\bar{\boldsymbol{\mathcal{L}}}\bar{\boldsymbol{y}} = \boldsymbol{P}^T\boldsymbol{b} \Longrightarrow \bar{\boldsymbol{y}} = \begin{pmatrix} 1 \\ 0.5 \end{pmatrix}$$

und

$$\bar{\boldsymbol{\mathcal{U}}}\bar{\boldsymbol{x}} = \bar{\boldsymbol{y}} \Longrightarrow \bar{\boldsymbol{x}} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}.$$

Die zweite Lösung ist im Rahmen der Maschinengenauigkeit exakt. Den beiden Zerlegungsvarianten entsprechen folgende Störungen in der Matrix:

$$\boldsymbol{\delta A} = \boldsymbol{A} - \boldsymbol{\mathcal{L}\mathcal{U}} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad \|\boldsymbol{\delta A}\| = 1$$

und

$$\delta\bar{\boldsymbol{A}} = \boldsymbol{A} - \boldsymbol{P}^T\bar{\boldsymbol{\mathcal{L}}}\bar{\boldsymbol{\mathcal{U}}} = \begin{pmatrix} 0 & 0.005 \\ 0 & 0 \end{pmatrix}, \quad \|\delta\bar{\boldsymbol{A}}\| = 0.005.$$

Multipliziert man im ursprünglichen Gleichungssystem die zweite Gleichung mit 200, so erhält man das äquivalente System

$$\tilde{\boldsymbol{A}}\boldsymbol{x} = \tilde{\boldsymbol{b}}$$

mit

$$\tilde{\boldsymbol{A}} = \begin{pmatrix} 1 & 200 \\ 1 & 1 \end{pmatrix}, \quad \tilde{\boldsymbol{b}} = \begin{pmatrix} 100 \\ 1 \end{pmatrix}.$$

Wendet man auf die Matrix $\tilde{\boldsymbol{A}}$ den Algorithmus zur LU-Zerlegung mit Spaltenpivotisierung an, so würde kein Spaltentausch erfolgen. Das Element $\tilde{a}_{11} = 1$ wird als Pivotelement akzeptiert. Im Maschinenzahlbereich $\mathbb{M}(10, 2, ., .)$ erhält man die Dreiecksfaktoren

$$\tilde{\boldsymbol{\mathcal{L}}} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad \tilde{\boldsymbol{\mathcal{U}}} = \begin{pmatrix} 1 & 200 \\ 0 & -200 \end{pmatrix}.$$

Für die Dreieckssysteme werden dann folgende Lösungen berechnet

$$\tilde{\boldsymbol{\mathcal{L}}}\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{b}} \Longrightarrow \tilde{\boldsymbol{y}} = \begin{pmatrix} 100 \\ -99 \end{pmatrix}$$

und

$$\tilde{\boldsymbol{\mathcal{U}}}\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{y}} \Longrightarrow \tilde{\boldsymbol{x}} = \begin{pmatrix} 0 \\ 0.5 \end{pmatrix}.$$

Das ist dieselbe schlechte Lösung wie beim ursprünglichen System ohne Pivotisierung. Als entsprechende Störung der Matrix ergibt sich wieder

$$\delta\tilde{\boldsymbol{A}} = \tilde{\boldsymbol{A}} - \tilde{\boldsymbol{\mathcal{L}}}\tilde{\boldsymbol{\mathcal{U}}} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

♡

Wie das Beispiel zeigte, lässt sich jedes Nichtnullelement der aktuellen Spalte durch Skalierung (Multiplikation der entsprechenden Zeile mit einer hinreichend großen Konstanten) zum betragsgrößten Element machen. Damit ist die einfache Pivotisierung ohne Berücksichtigung der Gesamtmatrix fraglich. Die Spaltenpivotisierung sichert, dass das Verhältnis der Normen $\|\boldsymbol{M}^{(k)}\|_\infty$ zu $\|\boldsymbol{M}^{(k-1)}\|_\infty$ nicht größer als 2 wird. Damit gilt die Abschätzung

$$\|\boldsymbol{M}^{(k)}\|_\infty \leqq 2^k \|\boldsymbol{A}\|_\infty.$$

Die Rückwärtsanalyse der LU-Zerlegung lieferte die Aussage, dass die berechneten Faktoren $\boldsymbol{\mathcal{L}}$ und $\boldsymbol{\mathcal{U}}$ exakte Dreiecksfaktoren der gestörten Matrix $\boldsymbol{A}+\boldsymbol{\delta A}$ sind. Die Störung genügte dabei der Abschätzung

$$\|\boldsymbol{\delta A}\|_\infty \leqq \text{eps} F(\boldsymbol{A})\|\boldsymbol{A}\|_\infty$$

mit

$$F(\boldsymbol{A}) = 1+3\sum_{k=1}^{n-1}\frac{\|\boldsymbol{M}^{(k)}\|_\infty}{\|\boldsymbol{A}\|_\infty}.$$

Für den gesamten erzeugten Rundungsfehler beim Lösen eines Gleichungssystems mittels LU-Zerlegung erhielten wir

$$\|\boldsymbol{\delta x}\|_\infty \leqq \frac{\text{eps}\bar{F}(\boldsymbol{A})\text{cond}_\infty(\boldsymbol{A})}{1-\text{eps}\bar{F}(\boldsymbol{A})\text{cond}_\infty(\boldsymbol{A})}\|\boldsymbol{x}\|_\infty$$

mit $\bar{F}(\boldsymbol{A}) = (n+2)F(\boldsymbol{A})$. Die Fehlerkonstante $F(\boldsymbol{A})$ lässt sich mittels Spaltenpivotisierung klein halten. Durch Skalierung, also Übergang zu einer Matrix $\tilde{\boldsymbol{A}} = \boldsymbol{DA}$, wobei $\boldsymbol{D}$ eine Diagonalmatrix ist, kann aber $\|\tilde{\boldsymbol{A}}\|_\infty$ so vergrößert werden, dass der Effekt der Pivotisierung zunichte gemacht wird. Wir wollen darum versuchen, durch geeignete Skalierung die Norm der Matrix zu verkleinern. Bezüglich der ∞-Norm lässt sich eine optimale Skalierung angeben.

8.29. Satz: *Für die reguläre (n,n)-Matrix $\boldsymbol{A}$ sei die Diagonalmatrix $\boldsymbol{D}$ wie folgt definiert.*

$$\boldsymbol{D} = \text{diag}(d_1,\ldots,d_n)$$

mit

$$d_i = \frac{\|\boldsymbol{A}\|_\infty}{\sum_{j=1}^n |a_{ij}|}, \quad i=1,\ldots,n.$$

Dann gilt für die Matrix $\tilde{\boldsymbol{A}} = \boldsymbol{DA}$

1.
$$\|\tilde{\boldsymbol{A}}\|_\infty = \|\boldsymbol{A}\|_\infty,$$

2.
$$\frac{\|\boldsymbol{A}^{-1}\|_\infty}{\|\boldsymbol{D}\|_\infty} \leqq \|\tilde{\boldsymbol{A}}^{-1}\|_\infty \leqq \|\boldsymbol{A}^{-1}\|_\infty,$$

3.

$$\frac{\text{cond}_\infty(\boldsymbol{A})}{\|\boldsymbol{D}\|_\infty} \leqq \text{cond}_\infty(\tilde{\boldsymbol{A}}) \leqq \text{cond}_\infty(\boldsymbol{A}).$$

Beweis: Siehe Übungsaufgabe 19. ❀

Für die Skalierung aus Satz 8.29 gilt

$$\sum_{j=1}^{n} |\tilde{a}_{ij}| = \|\boldsymbol{A}\|_\infty, \quad i = 1, \ldots, n.$$

Die Betragssummennorm aller Zeilen von $\tilde{\boldsymbol{A}}$ sind also gleich. Eine Matrix mit dieser Eigenschaft bezeichnen wir als **zeilenäquilibriert**. Der nächste Satz zeigt, dass gerade die zeilenäquilibrierten Matrizen die bezüglich der ∞-Norm optimal skalierten sind.

8.30. Satz: *Unter allen durch Zeilenskalierung aus einer (n,n)-Matrix $\boldsymbol{A}$ hervorgegangenen Matrizen $\tilde{\boldsymbol{A}} = \boldsymbol{D}\boldsymbol{A}$ hat jede zeilenäquilibrierte die kleinste Kondition bezüglich der ∞-Norm.*

Beweis: Die Skalierungsmatrix $\boldsymbol{D}$ aus Satz 8.29 erzeugt gerade aus der Matrix $\boldsymbol{A}$ eine zeilenäquilibrierte Matrix $\tilde{\boldsymbol{A}}$. Es sei nun $\bar{\boldsymbol{D}}$ eine weitere Diagonalmatrix, für die ebenfalls $\|\tilde{\boldsymbol{A}}\|_\infty = \|\bar{\boldsymbol{D}}\boldsymbol{A}\|_\infty = \|\boldsymbol{A}\|_\infty$ gilt. (Wegen $\text{cond}(\alpha\boldsymbol{A}) = \|\alpha\boldsymbol{A}\|\,\|(\alpha\boldsymbol{A}]^{-1}\| = |\alpha|\,\|\boldsymbol{A}\|\,\frac{1}{|\alpha|}\|\boldsymbol{A}^{-1}\| = \text{cond}(\boldsymbol{A})$ genügt es, nur Skalierungen mit $\|\bar{\boldsymbol{D}}\boldsymbol{A}\| = \|\boldsymbol{A}\|$ zu betrachten.) Es sei weiterhin $\bar{\boldsymbol{D}} \geqq \boldsymbol{O}$. Dann gilt

$$\|\bar{\boldsymbol{D}}\boldsymbol{A}\|_\infty = \max_{i=1,\ldots,n} \left\{ \bar{d}_i \sum_{j=1}^{n} |a_{ij}| \right\} = \|\boldsymbol{A}\|_\infty.$$

Damit folgt

$$\bar{d}_i \leqq \frac{\|\boldsymbol{A}\|_\infty}{\sum_{j=1}^{n} |a_{ij}|} = d_i, \quad i = 1, \ldots, n,$$

und es gilt

$$\begin{aligned} \|\tilde{\boldsymbol{A}}^{-1}\|_\infty &= \|(\boldsymbol{D}\boldsymbol{A})^{-1}\|_\infty = \|\boldsymbol{A}^{-1}\boldsymbol{D}^{-1}\|_\infty = \|\boldsymbol{A}^{-1}\bar{\boldsymbol{D}}^{-1}\bar{\boldsymbol{D}}\boldsymbol{D}^{-1}\|_\infty \\ &\leqq \|(\bar{\boldsymbol{D}}\boldsymbol{A})^{-1}\|_\infty \|\bar{\boldsymbol{D}}\boldsymbol{D}^{-1}\|_\infty = \|\bar{\boldsymbol{A}}^{-1}\|_\infty \max_{i=1,\ldots,n} \left\{ \frac{\bar{d}_i}{d_i} \right\} \leqq \|\bar{\boldsymbol{A}}^{-1}\|_\infty. \end{aligned}$$

Wegen $\|\tilde{\boldsymbol{A}}\|_\infty = \|\bar{\boldsymbol{A}}\|_\infty$ folgt daraus sofort $\text{cond}(\tilde{\boldsymbol{A}}) \leqq \text{cond}(\bar{\boldsymbol{A}})$. ❀

Eine Zeilenäquilibrierung mit Spaltenpivotisierung führt daher zu einer Verkleinerung der Fehlerschranke

$$\frac{\text{eps}\bar{F}(\boldsymbol{A})\text{cond}_\infty(\boldsymbol{A})}{1-\text{eps}F(\boldsymbol{A})\text{cond}_\infty(\boldsymbol{A})}\|\boldsymbol{x}\|_\infty$$

und damit zu einer Verkleinerung des Rundungsfehlers.

8.31. LU**-Zerlegung mit Spaltenpivotisierung, fiktivem Zeilentausch und fiktiver Skalierung:**
Es ist die reguläre Matrix $\boldsymbol{A}$ *in ein Produkt* $\boldsymbol{PA} = \boldsymbol{LU}$ *mit einer Permutationsmatrix* $\boldsymbol{P}$*, einer unteren Einsdreiecksmatrix* $\boldsymbol{L}$ *und einer oberen Dreiecksmatrix* $\boldsymbol{U}$ *zu zerlegen. Die Permutationsmatrix* $\boldsymbol{P}$ *sei dabei auf einem Feld* p *der Länge* n *gespeichert.*

{Initialisierung}
Wähle eine Genauigkeitsschranke $\varepsilon > 0$.
for $i = 1$ ***to*** n ***do***
  $p(i) = i$; $d_i = 0$
  for $j = 1$ ***to*** n ***do***
    $d_i = d_i + |a_{ij}|$
  endfor
endfor
{LU*-Zerlegung}*
for $j = 1$ ***to*** $n-1$ ***do***
  {Pivotsuche}
  $piv = 0$
  for $k = j$ ***to*** n ***do***
    if $d_j|a_{kj}| > piv$ ***then***
      $piv = d_j|a_{kj}|$
      $i^* = k$
    endif
  endfor
  if $piv \leqq \varepsilon$ ***then***
    STOPP
  endif
  {Fiktiver Zeilentausch}
  if $i^* \neq j$ ***then***
    $pp = p(i^*)$; $p(i^*) = p(j)$; $p(j) = pp$
  endif
  {Transformation der Restmatrix}
  for $i = j+1$ ***to*** n ***do***
    $a_{p(i),j} = a_{p(i),j}/a_{p(j),j}$

for $k = j+1$ ***to*** n ***do***
$a_{p(i),k} = a_{p(i),k} - a_{p(i),j} \cdot a_{p(j),k}$
endfor
endfor
endfor

Aufwand: $\sim n^3/3$ *Additionen/Multiplikationen*

8.32. Beispiel: Wir betrachten wieder das Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ mit

$$\boldsymbol{A} = \begin{pmatrix} 0.005 & 1.000 \\ 1.000 & 1.000 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} 0.5 \\ 1.0 \end{pmatrix}.$$

Für die Matrix $\boldsymbol{A}$ ergibt sich die Skalierungsmatrix

$$\boldsymbol{D} = \begin{pmatrix} \frac{2}{1.005} & 0 \\ 0 & 1 \end{pmatrix}$$

und damit das zeilenäquilibrierte System

$$\tilde{\boldsymbol{A}} = \boldsymbol{DAx} = \begin{pmatrix} \frac{0.01}{1.005} & \frac{2}{1.005} \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{1.005} \\ 1 \end{pmatrix} = \boldsymbol{Db} = \tilde{\boldsymbol{b}}.$$

Spaltenpivotisierung führt für dieses System zur Wahl von $\tilde{a}_{21} = 1$ als Pivotelement genau wie beim Beispiel 8.28 mit Spaltenpivotisierung. ♡

Führt man die Skalierung explizit durch (n^2 Multiplikationen), so treten zusätzliche Rundungsfehler auf. Eigentlich benötigt man die Skalierungsfaktoren aber nur zur Festlegung des jeweiligen Pivotelements. Wir suchen dann bei der Spaltenpivotisierung nicht nach dem betragsgrößten Element der Restspalte j, sondern nach dem Element, für das $d_i|a_{ij}|$ am größten ist. Diese Vorgehensweise nennt man **fiktive Skalierung**. Im obigen Algorithmus ist dies dargestellt.
Meist sind die Eingabedaten (Matrix, rechte Seite) fehlerbehaftet. Der Einfluss dieser Datenfehler übersteigt dabei oft den erzeugten Rundungsfehler. In diesem Falle ist es nicht so wichtig, den Rundungsfehler zu minimieren. Vielmehr ist der Einfluss des Datenfehler genau abzuschätzen. Es sei also die Matrix $\boldsymbol{A}$ mit dem Datenfehler $\boldsymbol{\delta}_0\boldsymbol{A}$ und die rechte Seite $\boldsymbol{b}$ mit dem Fehlervektor $\boldsymbol{\delta}_0\boldsymbol{b}$ belastet. Dann gilt die Abschätzung

$$\|\boldsymbol{\delta x}\| \leqq \frac{\|\boldsymbol{A}^{-1}\|}{1 - \|\boldsymbol{A}^{-1}\| \, \|\boldsymbol{\delta}_0\boldsymbol{A}\|} \left(\|\boldsymbol{\delta}_0\boldsymbol{A}\| \, \|\boldsymbol{x}\| + \|\boldsymbol{\delta}_0\boldsymbol{b}\| \right).$$

8.33. Beispiel: Es sei

$$\boldsymbol{A} = \begin{pmatrix} 1.00 & 0.50 & 0.00 \\ 0.97 & 0.48 & 0.00 \\ 0.75 & 0.00 & -0.75 \end{pmatrix}, \quad \boldsymbol{\delta_0 A} = \boldsymbol{O},$$

$$\boldsymbol{b} = \begin{pmatrix} 3.50 \\ 3.37 \\ -6.75 \end{pmatrix}, \quad \boldsymbol{\delta_0 b} = 10^{-3} \begin{pmatrix} 1 \\ 1 \\ 75 \end{pmatrix}.$$

Es folgt

$$\boldsymbol{A}^{-1} = \begin{pmatrix} -96 & 100 & 0 \\ 194 & -200 & 0 \\ -96 & 100 & -\frac{4}{3} \end{pmatrix}, \quad \boldsymbol{x} = \begin{pmatrix} 1 \\ 5 \\ 10 \end{pmatrix}, \quad \boldsymbol{\delta x} = 10^{-3} \begin{pmatrix} 4 \\ -6 \\ -96 \end{pmatrix}$$

mit $\|\boldsymbol{\delta x}\|_\infty = 0.096$. Die obige Abschätzung liefert aber

$$\|\boldsymbol{\delta x}\|_\infty \leqq \|\boldsymbol{A}^{-1}\|_\infty \|\boldsymbol{\delta_0 b}\|_\infty = 394 \cdot 75 \cdot 10^{-3} = 29.55.$$

Diese Abschätzung ist zu pessimistisch. Multiplizieren wir nun die letzte Gleichung in dem System mit $\frac{1}{75}$, so erhalten wir

$$\bar{\boldsymbol{A}} = \begin{pmatrix} 1.00 & 0.50 & 0.00 \\ 0.97 & 0.48 & 0.00 \\ 0.01 & 0.00 & -0.01 \end{pmatrix}, \quad \boldsymbol{\delta_0} \bar{\boldsymbol{A}} = \boldsymbol{O},$$

$$\bar{\boldsymbol{b}} = \begin{pmatrix} 3.50 \\ 3.37 \\ -0.09 \end{pmatrix}, \quad \boldsymbol{\delta_0} \bar{\boldsymbol{b}} = 10^{-3} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$

$$\bar{\boldsymbol{A}}^{-1} = \begin{pmatrix} -96 & 100 & 0 \\ 194 & -200 & 0 \\ -96 & 100 & -100 \end{pmatrix}, \quad \bar{\boldsymbol{x}} = \boldsymbol{x}, \quad \boldsymbol{\delta} \bar{\boldsymbol{x}} = \boldsymbol{\delta x}.$$

Hier liefert die Fehlerschätzung

$$\|\boldsymbol{\delta} \bar{\boldsymbol{x}}\|_\infty \leqq \|\boldsymbol{\delta} \bar{\boldsymbol{A}}^{-1}\|_\infty \|\boldsymbol{\delta_0} \bar{\boldsymbol{b}}\|_\infty = 394 \cdot 10^{-3} = 0.394.$$

Diese Abschätzung liegt in der Größenordnung des wahren Fehlers. ♡

Das Verhalten, das im letzten Beispiel geschildert wurde, ist auch im allgemeinen gültig. Durch Äquilibrierung des Fehlervektors der rechten Seite erreicht man realistischere Fehlerabschätzungen. Gleiches gilt bezüglich einer Zeilenäquilibrierung des Fehlers der Koeffizientenmatrix. Damit ergeben sich folgende Skalierungsregeln:

R1 Sind die Eingabedaten nicht fehlerbehaftet ($\boldsymbol{\delta}_0\boldsymbol{A} = \boldsymbol{O}$ und $\boldsymbol{\delta}_0\boldsymbol{b} = \boldsymbol{o}$) skaliere so, dass die Matrix zeilenäquilibriert wird:

$$d_i = \frac{\|\boldsymbol{A}\|_\infty}{\sum_{j=1}^n |a_{ij}|}, \quad i = 1, \ldots, n.$$

R2 Ist die rechte Seite fehlerbehaftet mit $|\boldsymbol{\delta}_0\boldsymbol{b}| \leqq \Delta\boldsymbol{b}$ und die Matrix exakt bzw. $\|\boldsymbol{\delta}_0\boldsymbol{A}\| \ll \|\Delta\boldsymbol{b}\|$ skaliere so, dass der Fehlervektor $\Delta\boldsymbol{b}$ äquilibriert wird:

$$d_i = \frac{\max_{j=1,\ldots,n}\{\Delta b_j\}}{\Delta b_i} = \frac{\|\Delta\boldsymbol{b}\|_\infty}{\Delta b_i}, \quad i = 1, \ldots, n.$$

R3 Ist die Matrix fehlerbehaftet mit $|\boldsymbol{\delta}_0\boldsymbol{A}| \leqq \Delta\boldsymbol{A}$ und die rechte Seite exakt bzw. $\|\boldsymbol{\delta}_0\boldsymbol{b}\| \ll \|\Delta\boldsymbol{A}\|$ skaliere so, dass die Fehlermatrix $\Delta\boldsymbol{A}$ zeilenäquilibriert wird:

$$d_i = \frac{\|\boldsymbol{\delta}_0\boldsymbol{A}\|_\infty}{\sum_{j=1}^n |\Delta a_{ij}|}, \quad i = 1, \ldots, n.$$

R4 Sind sowohl die rechte Seite als auch die Matrix fehlerbehaftet mit $|\boldsymbol{\delta}_0\boldsymbol{b}| \leqq \Delta\boldsymbol{b}$ und $|\boldsymbol{\delta}_0\boldsymbol{A}| \leqq \Delta\boldsymbol{A}$ skaliere so, dass der Fehlervektor

$$\Delta\boldsymbol{z} = \Delta\boldsymbol{A}|\boldsymbol{x}| + \Delta\boldsymbol{b}$$

äquilibriert wird:

$$d_i = \frac{\max_{j=1,\ldots,n}\{\Delta z_j\}}{\Delta z_i} = \frac{\|\Delta\boldsymbol{z}\|_\infty}{\Delta z_i}, \quad i = 1, \ldots, n.$$

Bemerkung: Zum Berechnen der Skalierungsfaktoren in Regel **R4** ist die Kenntnis der Lösung $\boldsymbol{x}$ notwendig. Darum wenden wir einen zweistufigen Prozess an.

S1 Skaliere mit **R1** und berechne Lösung $\boldsymbol{x}$.

S2 Berechne $\Delta\boldsymbol{z}$ und $\boldsymbol{D}$ nach **R4** und schätze $\|\boldsymbol{\delta x}\|_\infty$ gemäß

$$\|\boldsymbol{\delta x}\|_\infty \leqq \frac{\|\boldsymbol{A}^{-1}\boldsymbol{D}^{-1}\|_\infty}{1 - \|\boldsymbol{A}^{-1}\boldsymbol{D}^{-1}\|_\infty \|\boldsymbol{D}\Delta\boldsymbol{A}\|_\infty} \|\Delta\boldsymbol{z}\|_\infty.$$

8.2.4. Symmetrische Matrizen

Die gesamte Information einer symmetrischen Matrix ist auf dem oberen (unteren) Dreieck der Matrix einschließlich der Diagonalen enthalten. Man benötigt so gegenüber einer unsymmetrischen Matrix nur etwa den halben Speicherplatz. Beim Lösen von Gleichungssystemen mit symmetrischen Koeffizientenmatrizen lässt sich der Rechenaufwand ebenfalls halbieren. Eine hinreichende Bedingung liefert der folgende Satz.

8.34. Satz: *Der* GAUSS*sche Algorithmus sei für die symmetrische* (n,n)*-Matrix* $\boldsymbol{A}$ *ohne Pivotisierung durchführbar. Dann gilt*

$$\boldsymbol{A} = \boldsymbol{LU} = \boldsymbol{LDL}^T$$

mit $\boldsymbol{D} = \text{diag}(u_{11}, \ldots, u_{nn})$.

Beweis: Es ist zu zeigen, dass $\boldsymbol{U} = \boldsymbol{DL}^T$ gilt. Es sei $\boldsymbol{R} = \boldsymbol{D}^{-1}\boldsymbol{U}$. Dann ist $\boldsymbol{R}$ obere Einsdreiecksmatrix, und es gilt $\boldsymbol{A} = \boldsymbol{LDR}$. Aus der Symmetrie von $\boldsymbol{A}$ folgt

$$\boldsymbol{LDR} = (\boldsymbol{LDR})^T = \boldsymbol{R}^T\boldsymbol{DL}^T.$$

Daraus erhält man

$$\left(\boldsymbol{R}^T\right)^{-1}\boldsymbol{LD} = \boldsymbol{DL}^T\boldsymbol{R}^{-1}.$$

Auf der linken Seite der Gleichung steht eine untere Dreiecksmatrix mit den Diagonalelementen $d_1, \ldots, d_n$. Auf der rechten Seite steht eine obere Dreiecksmatrix mit den Diagonalelementen $d_1, \ldots, d_n$. Das ist aber nur möglich, falls $\boldsymbol{DL}^T\boldsymbol{R}^{-1} = \boldsymbol{D}$ ist. Damit folgt dann $\boldsymbol{L}^T = \boldsymbol{R}$. ❀

Erhält man eine Zerlegung der Form $\boldsymbol{A} = \boldsymbol{LDL}^T$, so war offensichtlich die LU-Zerlegung ohne Pivotisierung durchführbar. Damit ist die Bedingung auch notwendig. Eine Pivotisierung würde im allgemeinen die Symmetrie der Matrix zerstören. Denkbar ist nur eine Diagonalpivotisierung, also ein simultaner Zeilen- und Spaltentausch. Dabei werden nur Diagonalelemente als Pivotelemente ausgewählt.
Wir wollen nun zeigen, dass die LDL^T-Zerlegung einer symmetrischen Matrix etwa nur halb soviel Rechenoperationen erfordert wie die LU-Zerlegung. In Übungsaufgabe 15 wird gezeigt, dass alle Restmatrizen $\boldsymbol{M}^{(k)}$ bei der LU-Zerlegung einer symmetrischen Matrix ebenfalls symmetrisch sind, falls ohne Pivotisierung gearbeitet wird. Wegen $\boldsymbol{U} = \boldsymbol{DL}^T$ dürfen die Elemente l_{ij} auf dem oberen Dreieck von

$\boldsymbol{A}$ gespeichert werden. Mit den Elementen von $\boldsymbol{D}$ werden die entsprechenden Diagonalelemente von $\boldsymbol{A}$ überschrieben. Somit ergibt sich nach k Schritten folgende Situation:

$$\boldsymbol{A}^{(k)} = \begin{pmatrix} d_1 & l_{21} & \cdots & l_{k1} & l_{k+1,1} & \cdots & l_{n1} \\ & d_2 & \ddots & l_{k2} & l_{k+1,2} & & l_{n2} \\ & & \ddots & \ddots & \vdots & \ddots & \vdots \\ & & & d_k & l_{k+1,k} & \cdots & l_{nk} \\ & & & & a^{(k)}_{k+1,k+1} & \cdots & a^{(k)}_{k+1,n} \\ & & & & & \ddots & \vdots \\ & & & & & & a^{(k)}_{nn} \end{pmatrix}.$$

(Dabei ist in der Abbildung nur das obere Dreieck der aktuellen Iterationsmatrix angegeben.) Im nächsten Transformationsschritt wird folgendes berechnet:

$$\begin{aligned} d_{k+1} &= a^{(k)}_{k+1,k+1} \quad (\text{falls} \quad a^{(k)}_{k+1,k+1} \neq 0), \\ l_{i,k+1} &= \frac{a^{(k)}_{k+1,i}}{d_{k+1}}, \quad i = k+2, \ldots, n, \\ a^{(k+1)}_{ij} &= a^{(k)}_{ij} - l_{i,k+1} a^{(k)}_{k+1,j}, \quad i = k+2, \ldots, n, \quad j = i, \ldots, n. \end{aligned}$$

Wir erhalten den folgenden Algorithmus.

8.35. LDL^T**-Zerlegung:**
Es ist die reguläre symmetrische Matrix $\boldsymbol{A}$ in ein Produkt $\boldsymbol{A} = \boldsymbol{LDL}^T$ mit einer unteren Einsdreiecksmatrix $\boldsymbol{L}$ und einer Diagonalmatrix $\boldsymbol{D}$ zu zerlegen. Von der Matrix $\boldsymbol{A}$ ist nur das obere Dreieck einschließlich der Diagonalen gespeichert.

{Initialisierung}
Wähle Genauigkeitsschranke $\varepsilon > 0$.
{(LDL^T-Zerlegung)}
for $k = 1$ ***to*** $n - 1$ ***do***
  if $|a_{kk}| \leqq \varepsilon$ ***then***
    STOPP
  endif
  {Transformation der Restmatrix}
  for $i = k + 1$ ***to*** n ***do***
    $l = a_{ki}/a_{kk}$
    for $j = i$ ***to*** n ***do***
      $a_{ij} = a_{ij} - l \cdot a_{kj}$
    endfor

$a_{ki} = l$
endfor
endfor

Aufwand: $\sim n^3/6$ *Additionen/Multiplikationen*

Ist von einer symmetrischen Matrix eine LDL^T-Zerlegung bekannt, so löst man ein lineares Gleichungssystem in drei Schritten:

$$\boldsymbol{Ly} = \boldsymbol{b}, \quad \boldsymbol{Dz} = \boldsymbol{y}, \quad \boldsymbol{L}^T\boldsymbol{x} = \boldsymbol{z}.$$

Man erhält folgenden Algorithmus.

8.36. Lösen eines linearen Gleichungssystems bei bekannter LDL^T-Zerlegung:

Es ist das lineare Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ *mit der regulären symmetrischen Matrix* $\boldsymbol{A}$ *zu lösen. Von der Matrix* $\boldsymbol{A}$ *sei eine Zerlegung der Form* $\boldsymbol{A} = \boldsymbol{LDL}^T$ *mit einer unteren Einsdreiecksmatrix* $\boldsymbol{L}$ *und einer Diagonalmatrix* $\boldsymbol{D}$ *bekannt. Dabei ist die Matrix* $\boldsymbol{L}^T$ *auf dem oberen Dreieck der Matrix* $\boldsymbol{A}$ *und die Diagonalmatrix* $\boldsymbol{D}$ *auf der Diagonale der Matrix* $\boldsymbol{A}$ *gespeichert.*

{$\boldsymbol{Ly} = \boldsymbol{b}$*}*
*{***b** *wird mit* $\boldsymbol{y}$ *überschrieben.}*
for $k = 1$ **to** $n - 1$ **do**
for $i = k + 1$ **to** n **do**
$b_i = b_i - a_{ki} \cdot b_k$
endfor
endfor
{$\boldsymbol{Dz} = \boldsymbol{y}$*}*
{$\boldsymbol{y}$ *steht auf dem Speicherplatz von* **b** *und wird mit* $\boldsymbol{z}$ *überschrieben.}*
for $i = 1$ **to** n **do**
$b_i = b_i / a_{ii}$
endfor
{$\boldsymbol{L}^T\boldsymbol{x} = \boldsymbol{z}$*}*
{$\boldsymbol{z}$ *steht auf dem Speicherplatz von* **b** *und wird mit* $\boldsymbol{x}$ *überschrieben.}*
for $k = n$ **to** 2 **step** -1 **do**
for $i = 1$ **to** $k - 1$ **do**
$b_i = b_i - a_{ik} \cdot b_k$
endfor
endfor

Aufwand: $\sim n^2$ *Additionen/Multiplikationen*

Es bleibt die Frage zu klären, für welche Matrizen eine LDL^T-Zerlegung möglich ist. In Satz 8.34 hatten wir gesehen, dass das genau dann der Fall ist, wenn der

GAUSSsche Algorithmus ohne Pivotisierung durchführbar ist. Eine Klasse von Matrizen, für die dies zutrifft sind die sogenannten diagonaldominanten Matrizen.
Eine symmetrische (n,n)-Matrix $\boldsymbol{A}$ heißt **streng diagonaldominant**, falls

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$$

für $i = 1,\ldots,n$ gilt. **Bemerkung:** Im unsymmetrischen Falle gilt analog: **Strenge Zeilendiagonal- dominanz** bzw. **Strenge Spaltendiagonaldominanz**, falls

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \quad i = 1,\ldots,n$$

bzw.

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ji}|, \quad i = 1,\ldots,n$$

gilt.

8.37. Satz: *Für streng diagonaldominante symmetrische Matrizen $\boldsymbol{A}$ ist die LDL^T-Zerlegung durchführbar. Die berechneten Faktoren $\boldsymbol{\mathcal{L}}$ und $\boldsymbol{\mathcal{D}}$ sind exakte Faktoren der Zerlegung einer gestörten Matrix*

$$\boldsymbol{A} + \boldsymbol{\delta A} = \boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{D}}\boldsymbol{\mathcal{L}}^T.$$

Die Störung $\boldsymbol{\delta A}$ genügt der Abschätzung

$$\|\boldsymbol{\delta A}\|_\infty \leqq \text{eps}\, F(\boldsymbol{A})\|\boldsymbol{A}\|_\infty$$

mit $F(\boldsymbol{A}) = 1 + 3(n-1) \approx 3n$.

Beweis: In den Übungsaufgaben 15 und 16 wird gezeigt, dass die Symmetrie und die Diagonaldominanz bei einem Transformationsschritt erhalten bleiben. Die Diagonaldominanz wird sogar verstärkt. Damit ist der gesamte Algorithmus durchführbar. Es gilt nach Übungsaufgabe 16

$$\|\boldsymbol{M}^{(n-1)}\|_\infty \leqq \|\boldsymbol{M}^{(n-2)}\|_\infty \leqq \cdots \leqq \|\boldsymbol{M}^{(1)}\|_\infty \leqq \|\boldsymbol{M}^{(0)}\|_\infty = \|\boldsymbol{A}\|_\infty.$$

Mit Satz 8.27 folgt dann sofort die Existenz einer Störung $\boldsymbol{\delta A}$ mit $\boldsymbol{A} + \boldsymbol{\delta A} = \boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{D}}\boldsymbol{\mathcal{L}}^T$ und $\|\boldsymbol{\delta A}\|_\infty \leqq \text{eps}\,(1 + 3(n-1))\|\boldsymbol{A}\|_\infty$. ❀

Bemerkung: Die Abschätzung

$$\|\boldsymbol{\delta A}\|_\infty \leqq \text{eps}\,(1+3(n-1))\|\boldsymbol{A}\|_\infty$$

gilt auch für die LU-Zerlegung einer zeilendiagonaldominanten unsymmetrischen Matrix.
Eine weitere Klasse von Matrizen, für die der GAUSSsche Algorithmus ohne Pivotisierung durchführbar ist, ist die Klasse der positiv definiten Matrizen.
Eine symmetrische Matrix $\boldsymbol{A}$ heißt **positiv definit**, falls für alle Vektoren $\boldsymbol{x} \in \mathbb{R}^n$ mit $\boldsymbol{x} \neq \boldsymbol{o}$ $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > 0$ gilt. Gilt für alle Vektoren $\boldsymbol{x} \in \mathbb{R}^n$ mit $\boldsymbol{x} \neq \boldsymbol{o}$ $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} < 0$, so heißt $\boldsymbol{A}$ **negativ definit**, in den übrigen Fällen **indefinit**. Wir zeigen zunächst die Durchführbarkeit der LDL^T-Zerlegung für positiv definite Matrizen.

8.38. Satz: *Für jede symmetrische, positiv definite Matrix $\boldsymbol{A}$ ist die LDL^T-Zerlegung durchführbar. Man erhält eine Diagonalmatrix*

$$\boldsymbol{D} = \text{diag}(d_1,\ldots,d_n), \quad d_i > 0, \quad i = 1,\ldots,n.$$

Beweis: Wir beweisen den Satz mittels Induktion über die Dimension n der Matrix. Für $n = 1$ ist die Behauptung offensichtlich richtig.
Wir nehmen an, dass jede symmetrische, positiv definite Matrix $\boldsymbol{A} \in \mathbb{R}^{k\times k}$ in der Form $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$ mit $\boldsymbol{D} \geqq \boldsymbol{O}$ zerlegbar ist. Nun betrachten wir eine Matrix $\bar{\boldsymbol{A}} \in \mathbb{R}^{(k+1)\times(k+1)}$. Diese Matrix sei folgendermaßen partitioniert:

$$\bar{\boldsymbol{A}} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{b} \\ \boldsymbol{b}^T & \alpha \end{pmatrix}, \quad \boldsymbol{A} \in \mathbb{R}^{k\times k}.$$

Mit einem beliebigen Vektor $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x} \neq \boldsymbol{o}$ folgt

$$0 < (\boldsymbol{x}^T,0)\bar{\boldsymbol{A}}\begin{pmatrix} \boldsymbol{x} \\ 0 \end{pmatrix} = (\boldsymbol{x}^T,0)\begin{pmatrix} \boldsymbol{A} & \boldsymbol{b} \\ \boldsymbol{b}^T & \alpha \end{pmatrix}\begin{pmatrix} \boldsymbol{x} \\ 0 \end{pmatrix} = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}.$$

Damit ist die Matrix $\boldsymbol{A}$ ebenfalls positiv definit, und es existiert wegen der Voraussetzung eine Zerlegung $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$. Wir erhalten

$$\bar{\boldsymbol{A}} = \begin{pmatrix} \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T & \boldsymbol{b} \\ \boldsymbol{b}^T & \alpha \end{pmatrix}.$$

Nun machen wir für die Faktoren der Zerlegung von $\bar{\boldsymbol{A}}$ den Ansatz

$$\bar{\boldsymbol{L}} = \begin{pmatrix} \boldsymbol{L} & \\ \boldsymbol{c}^T & 1 \end{pmatrix}, \quad \bar{\boldsymbol{D}} = \begin{pmatrix} \boldsymbol{D} & \\ & d \end{pmatrix}.$$

Es folgt

$$\begin{aligned}\bar{\boldsymbol{L}}\bar{\boldsymbol{D}}\bar{\boldsymbol{L}}^T &= \begin{pmatrix} \boldsymbol{L} & \\ \boldsymbol{c}^T & 1 \end{pmatrix}\begin{pmatrix} \boldsymbol{D} & \\ & d \end{pmatrix}\begin{pmatrix} \boldsymbol{L}^T & \boldsymbol{c} \\ & 1 \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T & \boldsymbol{L}\boldsymbol{D}\boldsymbol{c} \\ \boldsymbol{c}^T\boldsymbol{D}\boldsymbol{L}^T & \boldsymbol{c}^T\boldsymbol{D}\boldsymbol{c}+d \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T & \boldsymbol{b} \\ \boldsymbol{b}^T & \alpha \end{pmatrix}.\end{aligned}$$

Daraus erhalten wir

$$\boldsymbol{c} = \boldsymbol{D}^{-1}\boldsymbol{L}^{-1}\boldsymbol{b}$$

und

$$d = \alpha - \boldsymbol{b}^T\left(\boldsymbol{L}^{-1}\right)^T\boldsymbol{D}^{-1}\boldsymbol{L}^{-1}\boldsymbol{b} = \alpha - \boldsymbol{b}^T\boldsymbol{A}^{-1}\boldsymbol{b}.$$

Aus der Regularität von $\boldsymbol{D}$ und $\boldsymbol{L}$ folgt sofort die Existenz von $\boldsymbol{c}$. Es bleibt dann nur noch die Positivität von d zu zeigen. Dazu sei $\boldsymbol{y} \in \mathbb{R}^{k+1}$ die Lösung des Gleichungssystems $\bar{\boldsymbol{L}}^T\boldsymbol{y} = \boldsymbol{e}_{k+1}$. Dann folgt aus der positiven Definitheit von $\bar{\boldsymbol{A}}$

$$0 < \boldsymbol{y}^T\bar{\boldsymbol{A}}\boldsymbol{y} = \boldsymbol{y}^T\bar{\boldsymbol{L}}\bar{\boldsymbol{D}}\bar{\boldsymbol{L}}^T\boldsymbol{y} = \boldsymbol{e}_{k+1}^T\bar{\boldsymbol{D}}\boldsymbol{e}_{k+1} = d.$$

❀

Ohne Beweis seien die Ergebnisse der Rundungsfehleranalyse der LDL^T-Zerlegung einer symmetrischen, positiv definiten Matrix angegeben.

8.39. Satz: *Die LDL^T-Zerlegung ist für die symmetrische, positiv definite Matrix $\boldsymbol{A}$ auf einem realen Rechner mit $d_i > 0$, $i = 1, \ldots, n$, durchführbar, falls*

$$\text{eps}F\|\boldsymbol{A}^{-1}\|_2\|\boldsymbol{A}\|_F \overset{\leq}{=} \frac{1}{2}$$

mit $F = n + 2\ln(n/2) \approx n$ gilt.
Zu den berechneten Faktoren $\boldsymbol{\mathcal{L}}$ und $\boldsymbol{\mathcal{D}}$ existiert eine symmetrische Störung $\boldsymbol{\delta A}$ mit

$$\boldsymbol{A} + \boldsymbol{\delta A} = \boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{D}}\boldsymbol{\mathcal{L}}^T.$$

Die Störung genügt der Abschätzung

$$\|\boldsymbol{\delta A}\|_F \overset{\leq}{=} \text{eps}F\|\boldsymbol{A}\|_F.$$

Die zu $\boldsymbol{b} \in \mathbb{R}^n$ berechnete Lösung $\bar{\boldsymbol{x}}$ ist exakte Lösung des gestörten Problems

$$(\boldsymbol{A} + \overline{\boldsymbol{\delta A}})\boldsymbol{x} = \boldsymbol{b}$$

mit der symmetrischen Störung $\overline{\boldsymbol{\delta A}}$, die der Abschätzung

$$\|\overline{\boldsymbol{\delta A}}\|_p \leqq \text{eps}\bar{F}\|\boldsymbol{A}\|_p$$

mit

$$\bar{F} = \begin{cases} 2n^{3/2} + F \approx 2n^{3/2} & \textit{für} \quad p = F \\ \\ n^{3/2} + n + n^{1/2}F \approx 2n^{3/2} & \textit{für} \quad p = 2 \end{cases}$$

genügt.

Für positiv definite Matrizen lässt sich auch eine Zerlegung der Form $\boldsymbol{A} = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T$ mit einer unteren Dreiecksmatrix $\hat{\boldsymbol{L}}$ angeben. Ein Vergleich mit der LDL^T-Zerlegung zeigt, dass dann $\hat{\boldsymbol{L}} = \boldsymbol{L}\boldsymbol{D}^{1/2}$ mit

$$\boldsymbol{D}^{1/2} = \text{diag}\left(\sqrt{d_1}, \ldots, \sqrt{d_n}\right)$$

gilt. Diese Zerlegung wird als CHOLESKY-**Zerlegung** bezeichnet.

8.40. CHOLESKY-**Zerlegung:**
Es ist die symmetrische, positiv definite Matrix $\boldsymbol{A}$ in ein Produkt $\boldsymbol{A} = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T$ mit einer unteren Dreiecksmatrix $\hat{\boldsymbol{L}}$ zu zerlegen. Von der Matrix $\boldsymbol{A}$ sind nur das untere Dreieck und die Diagonale gespeichert.

{Initialisierung}
Wähle Genauigkeitsschranke $\varepsilon > 0$.
{Zerlegung}
for $k = 1$ *to* n ***do***
  if $a_{kk} \leqq \varepsilon$ ***then***
    STOPP
  endif
  $a_{kk} = \sqrt{a_{kk}}$
  {Transformation der Restmatrix}
  for $i = k+1$ *to* n ***do***
    $a_{ik} = a_{ik}/a_{kk}$
  endfor
  for $i = k+1$ *to* n ***do***
    for $j = i$ *to* n ***do***

$a_{ij} = a_{ij} - l \cdot a_{jk}$
endfor
endfor
endfor

Aufwand: $\sim n^3/6$ *Additionen/Multiplikationen +* n *Wurzeln*

Das Lösen eines Gleichungssystems $\boldsymbol{Ax} = \boldsymbol{b}$ erfolgt dann in zwei Schritten:

$$\hat{\boldsymbol{L}}\boldsymbol{y} = \boldsymbol{b}, \quad \hat{\boldsymbol{L}}^T\boldsymbol{x} = \boldsymbol{y}.$$

Für die Rundungsfehleranalyse der CHOLESKY-Zerlegung gelten alle Aussagen von Satz 8.39 mit $F = n + 1 + \frac{1}{2}\ln n$. Für großes n gilt wieder $F \approx n$ und $\bar{F} \approx 2n^{3/2}$.
Die LDL^T-Zerlegung ist nicht für alle regulären symmetrischen Matrizen durchführbar, auch nicht falls man Diagonalpivotisierung zulässt.

8.41. Beispiel: Für die Matrix

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{pmatrix}$$

existiert keine LDL^T-Zerlegung. ♡

Für derartige Matrizen existieren aber sogenannte Blockfaktorisierungen der Form

$$\boldsymbol{PAP}^T = \boldsymbol{LDL}^T$$

mit

$$\boldsymbol{L} = \begin{pmatrix} \boldsymbol{I}_1 & & & \\ \boldsymbol{G}_{21} & \boldsymbol{I}_2 & & \boldsymbol{O} \\ \vdots & \vdots & \ddots & \\ \boldsymbol{G}_{l1} & \boldsymbol{G}_{l2} & \cdots & \boldsymbol{I}_l \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_1 & & & \\ & \boldsymbol{D}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{D}_l \end{pmatrix}.$$

Die Dimension der Diagonalblöcke $\boldsymbol{I}_i$ bzw. $\boldsymbol{D}_i$ beträgt jeweils 1 oder 2. Für diese Blockzerlegungen existieren auch Pivotstrategien, die die Stabilität der Algorithmen verbessern.

8.2.5. Orthogonalisierungsverfahren

Lösen wir ein lineares Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ mittels einer LU-Zerlegung von $\boldsymbol{A}$, so gilt

$$\boldsymbol{Ly} = \boldsymbol{P}^T\boldsymbol{b}, \quad \boldsymbol{Ux} = \boldsymbol{y}.$$

Für den Einfluss von Fehlern in $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{L}$, $\boldsymbol{U}$ und $\boldsymbol{y}$ gelten die Abschätzungen

$$\frac{\|\boldsymbol{\delta x}\|}{\|\boldsymbol{x}\|} \mathrel{\dot{\leqq}} \operatorname{cond}(\boldsymbol{A}) \left[\frac{\|\boldsymbol{\delta A}\|}{\|\boldsymbol{A}\|} + \frac{\|\boldsymbol{\delta b}\|}{\|\boldsymbol{b}\|} \right]$$

bzw.

$$\begin{aligned} \frac{\|\boldsymbol{\delta y}\|}{\|\boldsymbol{y}\|} &\mathrel{\dot{\leqq}} \operatorname{cond}(\boldsymbol{L}) \left[\frac{\|\boldsymbol{\delta L}\|}{\|\boldsymbol{L}\|} + \frac{\|\boldsymbol{\delta b}\|}{\|\boldsymbol{b}\|} \right], \\ \frac{\|\boldsymbol{\delta x}\|}{\|\boldsymbol{x}\|} &\mathrel{\dot{\leqq}} \operatorname{cond}(\boldsymbol{U}) \left[\frac{\|\boldsymbol{\delta U}\|}{\|\boldsymbol{U}\|} + \frac{\|\boldsymbol{\delta y}\|}{\|\boldsymbol{y}\|} \right]. \end{aligned}$$

Die maximale Fehlerverstärkung beim Lösen der Dreiecksysteme wird daher durch die Größe $\operatorname{cond}(\boldsymbol{L})\operatorname{cond}(\boldsymbol{U})$ beschrieben. Wegen der Submultiplikativität der zugrunde liegenden Matrixnormen gilt aber

$$\operatorname{cond}(\boldsymbol{A}) \leqq \operatorname{cond}(\boldsymbol{L})\operatorname{cond}(\boldsymbol{U}).$$

Dabei wird das Produkt der Konditionszahlen von $\boldsymbol{L}$ und $\boldsymbol{U}$ die Konditionszahl von $\boldsymbol{A}$ bei weitem übertreffen.

8.42. Beispiel: Für die Matrix

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

ergeben sich die Faktoren

$$\boldsymbol{L} = \begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ -1 & -1 & 1 & \\ -1 & -1 & -1 & 1 \end{pmatrix}, \quad \boldsymbol{U} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ & 1 & 0 & 2 \\ & & 1 & 4 \\ & & & 8 \end{pmatrix}.$$

Weiterhin ist

$$\boldsymbol{A}^{-1} = \frac{1}{8} \begin{pmatrix} 4 & -2 & -1 & -1 \\ 0 & 4 & -2 & -2 \\ 0 & 0 & 4 & -4 \\ 4 & 2 & 1 & 1 \end{pmatrix},$$

$$\boldsymbol{L}^{-1} = \begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ 2 & 1 & 1 & \\ 4 & 2 & 1 & 1 \end{pmatrix}, \quad \boldsymbol{U}^{-1} = \frac{1}{8} \begin{pmatrix} 8 & 0 & 0 & -1 \\ & 8 & 0 & -2 \\ & & 8 & -4 \\ & & & 1 \end{pmatrix}.$$

Damit gilt

$$\begin{aligned} \text{cond}_1(\boldsymbol{A}) &= \|\boldsymbol{A}\|_1\|\boldsymbol{A}^{-1}\|_1 = 4\cdot 1 = 4, \\ \text{cond}_1(\boldsymbol{L}) &= \|\boldsymbol{L}\|_1\|\boldsymbol{L}^{-1}\|_1 = 4\cdot 8 = 32, \\ \text{cond}_1(\boldsymbol{U}) &= \|\boldsymbol{U}\|_1\|\boldsymbol{U}^{-1}\|_1 = 15\cdot 1 = 15. \end{aligned}$$

Hier ist $4 = \text{cond}(\boldsymbol{A}) \ll \text{cond}_1(\boldsymbol{L})\text{cond}_1(\boldsymbol{U}) = 480$. ♡

Eine Forderung an Zerlegungen einer Matrix $\boldsymbol{A}$ der Form $\boldsymbol{A} = \boldsymbol{A}_1\boldsymbol{A}_2\cdots\boldsymbol{A}_k$ wäre damit die, dass das Produkt der Konditionszahlen der Faktoren $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_k$ nicht wesentlich größer sein sollte als die Konditionszahl von $\boldsymbol{A}$. Noch besser wäre es, falls

$$\text{cond}(\boldsymbol{A}) = \text{cond}(\boldsymbol{A}_1)\cdots\text{cond}(\boldsymbol{A}_k)$$

gelten würde.
Bezüglich der Spektralnorm lassen sich solche Zerlegungen sofort angeben. Da die Spektralnorm invariant gegenüber orthogonalen Transformationen ist, gilt für eine Zerlegung der Form $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ mit einer orthogonalen Matrix $\boldsymbol{Q}$:

$$\|\boldsymbol{A}\|_2 = \|\boldsymbol{Q}\boldsymbol{R}\|_2 = \|\boldsymbol{R}\|_2, \quad \|\boldsymbol{A}^{-1}\|_2 = \|\boldsymbol{R}^{-1}\boldsymbol{Q}^T\|_2 = \|\boldsymbol{R}^{-1}\|_2.$$

Damit folgt $\text{cond}_2(\boldsymbol{A}) = \text{cond}_2(\boldsymbol{R})$. Für orthogonale Matrizen gilt $\text{cond}_2(\boldsymbol{Q}) = 1$, so dass sich insgesamt $\text{cond}_2(\boldsymbol{A}) = \text{cond}_2(\boldsymbol{Q})\text{cond}_2(\boldsymbol{R})$ ergibt. Ist nun noch $\boldsymbol{R}$ eine obere Dreiecksmatrix, so verwendet man die Zerlegung $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ zum Lösen von linearen Gleichungssystemen $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. Man erhält $\boldsymbol{x}$ als Lösung von

$$\boldsymbol{R}\boldsymbol{x} = \boldsymbol{Q}^T\boldsymbol{b}.$$

Die Frage nach der Existenz einer solchen Zerlegung lässt sich sofort beantworten. Es gilt

8.43. Satz: *Für jede reguläre Matrix $\boldsymbol{A}$ existieren eine orthogonale Matrix $\boldsymbol{Q}$ und eine obere Dreiecksmatrix $\boldsymbol{R}$, so dass $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ gilt.*

Beweis: Man wende das SCHMIDTsche Orthogonalisierungsverfahren auf die Spalten von $\boldsymbol{A}$ an. ❀

Mit dem SCHMIDTsche Orthogonalisierungsverfahren hätte man sofort einen Algorithmus zum Berechnen einer QR-Zerlegung. Leider ist das SCHMIDTsche Orthogonalisierungsverfahren aber nicht numerisch gutartig, wie das folgende Beispiel zeigt.

8.44. Beispiel: Wir betrachten die Matrix

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

Für ε gelte $\varepsilon >$ eps, aber $\varepsilon^2 <$ eps, so dass gl $(1+\varepsilon^2) = 1$. Das SCHMIDTsche Orthogonalisierungsverfahren liefert dann

$$\boldsymbol{Q} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \varepsilon & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & -1 \\ 0 & \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & 0 \end{pmatrix}, \quad \boldsymbol{R} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & \varepsilon\sqrt{2} & 0 & 0 \\ 0 & 0 & \varepsilon\sqrt{2} & 0 \\ 0 & 0 & 0 & \varepsilon \end{pmatrix}.$$

Damit ist zwar eine exakte Zerlegung gegeben: $\boldsymbol{A} = \boldsymbol{QR}$, aber die Matrix $\boldsymbol{Q}$ ist extrem nichtorthogonal:

$$\boldsymbol{Q}^T\boldsymbol{Q} = \begin{pmatrix} 1+\varepsilon^2 & -\frac{\varepsilon}{\sqrt{2}} & -\frac{\varepsilon}{\sqrt{2}} & -\varepsilon \\ -\frac{\varepsilon}{\sqrt{2}} & 1 & \frac{1}{2} & \frac{1}{\sqrt{2}} \\ -\frac{\varepsilon}{\sqrt{2}} & \frac{1}{2} & 1 & \frac{1}{\sqrt{2}} \\ -\varepsilon & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 1 \end{pmatrix}.$$

Würde man mit dieser Zerlegung ein Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ gemäß $\boldsymbol{Rx} = \boldsymbol{Q}^T\boldsymbol{b}$ lösen, so würde man einen großen Fehler erhalten. Zum Beispiel ergibt sich für die rechte Seite

$$\boldsymbol{b} = \begin{pmatrix} 4 \\ \varepsilon \\ \varepsilon \\ \varepsilon \end{pmatrix} \quad \text{mit exaktem} \quad \boldsymbol{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad \text{die Lösung} \quad \bar{\boldsymbol{x}} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix}.$$

♡

Das SCHMIDTsche Orthogonalisierungsverfahren ist daher in dieser Form für Computerrechnung nicht geeignet. Es existiert aber eine Modifizierung des Verfahrens, die diese numerischen Schwierigkeiten überwindet. Wir wollen jedoch ein anderes Verfahren zum Berechnen einer QR-Zerlegung einer Matrix $\boldsymbol{A}$ behandeln.

Das Householder-Verfahren

Erinnern wir uns zunächst an den GAUSSschen Algorithmus. Dort wurde mit Hilfe von elementaren Transformationsmatrizen (LNT-Matrizen) die Matrix schrittweise auf obere Dreiecksform transformiert. Im k-ten Schritt wurden durch die Transformation mit $\boldsymbol{L}_k(-\boldsymbol{l}_k)$ in der k-ten Spalte unterhalb der Diagonalen Nullen erzeugt. Wir wollen nun orthogonale Transformationsmatrizen konstruieren, mit denen wir ähnliches erreichen. Wir suchen nach einer orthogonalen Matrix $\boldsymbol{P}$, die einen Vektor $\boldsymbol{a} \in \mathbb{R}^k$ auf ein Vielfaches des ersten Einheitsvektors $\boldsymbol{e}_1 \in \mathbb{R}^k$ transformiert:

$$\boldsymbol{P}\boldsymbol{a} = \varrho \boldsymbol{e}_1.$$

Wir werden sehen, dass dies mit HOUSEHOLDER-Spiegelungen möglich ist. Es sei daher $\boldsymbol{P}$ eine HOUSEHOLDER-Matrix

$$\boldsymbol{P} = \boldsymbol{H} = \boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^T, \quad \boldsymbol{u}^T\boldsymbol{u} = 1.$$

Den Vektor $\boldsymbol{u}$ bestimmen wir so, dass

$$\boldsymbol{H}\boldsymbol{a} = (\boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^T)\boldsymbol{a} = \varrho \boldsymbol{e}_1$$

erfüllt ist. Da $\boldsymbol{H}$ orthogonal ist, gilt

$$\|\boldsymbol{a}\|_2 = \|\boldsymbol{H}\boldsymbol{a}\|_2 = \|\varrho \boldsymbol{e}_1\|_2 = |\varrho| \, \|\boldsymbol{e}_1\|_2 = |\varrho|.$$

Damit folgt $\varrho = \pm\|\boldsymbol{a}\|_2$ und

$$\begin{aligned}
(\boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^T)\boldsymbol{a} &= \pm\|\boldsymbol{a}\|_2 \boldsymbol{e}_1, && (8.10)\\
\boldsymbol{a} - 2\boldsymbol{u}(\boldsymbol{u}^T\boldsymbol{a}) &= \pm\|\boldsymbol{a}\|_2 \boldsymbol{e}_1, && (8.11)\\
\boldsymbol{u} &= \frac{\boldsymbol{a} \mp \|\boldsymbol{a}\|_2 \boldsymbol{e}_1}{2(\boldsymbol{u}^T\boldsymbol{a})}. && (8.12)
\end{aligned}$$

Aus Gleichung 8.11 ergibt sich

$$\begin{aligned}
\boldsymbol{a}^T\left[\boldsymbol{a} - 2\boldsymbol{u}(\boldsymbol{u}^T\boldsymbol{a})\right] &= \pm\|\boldsymbol{a}\|_2 \boldsymbol{a}^T\boldsymbol{e}_1,\\
\|\boldsymbol{a}\|_2^2 - 2(\boldsymbol{u}^T\boldsymbol{a})^2 &= \pm\|\boldsymbol{a}\|_2 \boldsymbol{a}^T\boldsymbol{e}_1,\\
(\boldsymbol{u}^T\boldsymbol{a})^2 &= \frac{\mp\|\boldsymbol{a}\|_2 \boldsymbol{a}^T\boldsymbol{e}_1 - \|\boldsymbol{a}\|_2^2}{2} = \frac{\varrho a_1 - \varrho^2}{2},\\
\boldsymbol{u}^T\boldsymbol{a} &= \sqrt{\frac{\varrho a_1 - \varrho^2}{2}}
\end{aligned}$$

wobei a_1 die erste Komponente des Vektors $\boldsymbol{a}$ bezeichnet. Setzt man dies in die Gleichung 8.12 ein, so erhält man

$$\boldsymbol{u} = \frac{\boldsymbol{a} - \varrho \boldsymbol{e}_1}{\sqrt{2\varrho(\varrho - a_1)}}$$

und

$$\begin{aligned} \boldsymbol{H} &= \boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^T = \boldsymbol{I} - 2\frac{(\boldsymbol{a} - \varrho\boldsymbol{e}_1)(\boldsymbol{a} - \varrho\boldsymbol{e}_1)^T}{2\varrho(\varrho - a_1)} \\ &= \boldsymbol{I} - \frac{(\boldsymbol{a} - \varrho\boldsymbol{e}_1)(\boldsymbol{a} - \varrho\boldsymbol{e}_1)^T}{\varrho(\varrho - a_1)} = \boldsymbol{I} - \frac{\boldsymbol{v}\boldsymbol{v}^T}{\gamma} \end{aligned}$$

mit

$$\boldsymbol{v} = \boldsymbol{a} - \varrho\boldsymbol{e}_1 = \begin{pmatrix} a_1 - \varrho \\ a_2 \\ \vdots \\ a_k \end{pmatrix}, \quad \gamma = \varrho(\varrho - a_1).$$

Damit haben wir die gesuchte Transformation gefunden. Für die Wahl von ϱ stehen uns zwei Möglichkeiten zur Verfügung:

$$\varrho = \|\boldsymbol{a}\|_2 \quad \text{oder} \quad \varrho = -\|\boldsymbol{a}\|_2.$$

Wir werden das Vorzeichen von ϱ so wählen, dass beim Berechnen von $\varrho - a_1$ keine Auslöschung auftritt: ϱ und a_1 müssen entgegengesetzte Vorzeichen haben. Wir erhalten den folgenden Algorithmus.

8.45. Festlegen der HOUSEHOLDER**-Matrix zur Transformation eines Vektors auf ein Vielfaches des ersten Einheitsvektors:**
Es sei ein Vektor $\boldsymbol{a} \in \mathbb{R}^k$ gegeben.

S0 *Berechne* $\varrho = \sqrt{\boldsymbol{a}^T\boldsymbol{a}} = \|\boldsymbol{a}\|_2$.

S1 **if** $\varrho = 0$ **then** $\boldsymbol{H} = \boldsymbol{I}$ **STOPP.**

S2 **if** $a_1 > 0$ **then** $\varrho = -\varrho$.
Berechne $\boldsymbol{v} = \boldsymbol{a} - \varrho\boldsymbol{e}_1$ *und* $\gamma = \varrho(\varrho - a_1)$.
Die Matrix $\boldsymbol{H} = \boldsymbol{I} - \boldsymbol{v}\boldsymbol{v}^T/\gamma$ *ist dann orthogonal und es gilt*

$$\boldsymbol{H}\boldsymbol{a} = \varrho\boldsymbol{e}_1.$$

Aufwand: $\sim n$ *Additionen/Multiplikationen, eine Quadratwurzel.*

Das Berechnen der Transformation eines Vektors $\boldsymbol{x}$ erfolgt nach folgendem Algorithmus.

8.46. HOUSEHOLDER**-Transformation:**
Gegeben sei eine (n,n)-HOUSEHOLDER*-Matrix*

$$\boldsymbol{H} = \boldsymbol{I} - \boldsymbol{v}\boldsymbol{v}^T/\gamma.$$

Zu berechnen ist für einen Vektor $\boldsymbol{x} \in \mathbb{R}^n$ *der Vektor*

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}.$$

S0 *Berechne* $\beta = \boldsymbol{v}^T\boldsymbol{x}/\gamma$.
S1 *Berechne* $\boldsymbol{y} = \boldsymbol{x} - \beta\boldsymbol{v}$
Aufwand: $\sim 2n$ *Additionen/Multiplikationen.*

Bemerkung: Die Matrix $\boldsymbol{H}$ braucht explizit nicht berechnet zu werden. Die gesamten Informationen sind im Vektor $\boldsymbol{v}$ und in γ enthalten. Das explizite Berechnen der HOUSEHOLDER-Matrix ist aufwendig ($\sim n^2/2$ Additionen/Multiplikationen). Das Berechnen des Vektors $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}$ würde sogar $\sim n^2$ Operationen kosten, falls man ihn in der Form Matrix mal Vektor berechnet.
Mit Hilfe der letzten beiden Algorithmen entwickeln wir nun einen Algorithmus zum Transformieren einer Matrix auf obere Dreiecksform. Wir nehmen an, dass wir nach $k-1$ Schritten die Matrix auf die Form

$$\boldsymbol{A}^{(k-1)} = \begin{pmatrix} a_{11}^{(k-1)} & \cdots & a_{1,k-1}^{(k-1)} & a_{1k}^{(k-1)} & \cdots & a_{1n}^{(k-1)} \\ 0 & \ddots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \ddots & a_{k-1,k-1}^{(k-1)} & a_{k-1,k}^{(k-1)} & \cdots & a_{k-1,n}^{(k-1)} \\ 0 & \cdots & 0 & a_{kk}^{(k-1)} & \cdots & a_{kn}^{(k-1)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{nk}^{(k-1)} & \cdots & a_{nn}^{(k-1)} \end{pmatrix}$$

$$= \left(\begin{array}{c|c} \boldsymbol{R}^{(k-1)} & \\ \hline \boldsymbol{O} & \boldsymbol{M}^{(k-1)} \end{array}\right)$$

mit $\boldsymbol{R}^{(k-1)} \in \mathbb{R}^{(k-1)\times n}$, $\boldsymbol{M}^{(k-1)} \in \mathbb{R}^{(n-k+1)\times(n-k+1)}$ transformiert haben. Im nächsten Schritt wählen wir eine HOUSEHOLDER-Matrix $\boldsymbol{H}^{(k)}$ so, dass die Teilmatrix $\boldsymbol{R}^{(k-1)}$ unverändert bleibt und die erste Spalte von $\boldsymbol{M}^{(k-1)}$ auf ein Vielfaches des entsprechenden Einheitsvektors abgebildet wird. Damit hat $\boldsymbol{H}^{(k)}$ folgende Struktur

$$\boldsymbol{H}^{(k)} = \begin{pmatrix} \boldsymbol{I}^{(k-1)} & \boldsymbol{O} \\ \boldsymbol{O} & \tilde{\boldsymbol{H}}^{(k)} \end{pmatrix}$$

mit

$$\tilde{\boldsymbol{H}}^{(k)}\boldsymbol{M}^{(k-1)} = \begin{pmatrix} \star & \star & \cdots & \star \\ 0 & \star & \cdots & \star \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \star & \cdots & \star \end{pmatrix} = \left(\begin{array}{c|c} \star & \star \ \cdots \ \star \\ \hline 0 & \\ \vdots & \boldsymbol{M}^{(k)} \\ 0 & \end{array}\right).$$

Damit gilt

$$\begin{aligned} \boldsymbol{H}^{(k)}\boldsymbol{A}^{(k-1)} &= \begin{pmatrix} \boldsymbol{I}^{(k-1)} & \boldsymbol{O} \\ \boldsymbol{O} & \tilde{\boldsymbol{H}}^{(k)} \end{pmatrix} \left(\begin{array}{c|c} \boldsymbol{R}^{(k-1)} & \\ \hline \boldsymbol{O} & \boldsymbol{M}^{(k-1)} \end{array}\right) \\ &= \left(\begin{array}{c|c} \boldsymbol{R}^{(k-1)} & \\ \hline \boldsymbol{O} & \tilde{\boldsymbol{H}}^{(k)}\boldsymbol{M}^{(k-1)} \end{array}\right) \\ &= \left(\begin{array}{c|c} \boldsymbol{R}^{(k-1)} & \\ \hline \boldsymbol{O} & \begin{array}{c} \star \ \star \ \cdots \ \star \\ 0 \\ \vdots \quad \boldsymbol{M}^{(k)} \\ 0 \end{array} \end{array}\right) = \left(\begin{array}{c|c} \boldsymbol{R}^{(k)} & \\ \hline \boldsymbol{O} & \boldsymbol{M}^{(k)} \end{array}\right) \end{aligned}$$

mit $\boldsymbol{R}^{(k)} \in \mathbb{R}^{k\times n}$ und $\boldsymbol{M}^{(k)} \in \mathbb{R}^{(n-k)\times(n-k)}$. Nach $n-1$ Schritten ist schließlich $\boldsymbol{M}^{(n-1)} \in \mathbb{R}^{1\times 1}$. Dann ist $\boldsymbol{A}^{(n-1)}$ obere Dreiecksmatrix, und es gilt

$$\boldsymbol{R} = \boldsymbol{A}^{(n-1)} = \boldsymbol{H}^{(n-1)}\boldsymbol{H}^{(n-2)}\cdots\boldsymbol{H}^{(2)}\boldsymbol{H}^{(1)}\boldsymbol{A}$$

bzw.

$$\boldsymbol{A} = \boldsymbol{H}^{(1)}\boldsymbol{H}^{(2)}\cdots\boldsymbol{H}^{(n-2)}\boldsymbol{H}^{(n-1)}\boldsymbol{R} = \boldsymbol{Q}\boldsymbol{R}$$

mit der orthogonalen Matrix

$$\boldsymbol{Q} = \boldsymbol{H}^{(1)}\boldsymbol{H}^{(2)}\cdots\boldsymbol{H}^{(n-2)}\boldsymbol{H}^{(n-1)}.$$

Die Matrix $\boldsymbol{Q}$ braucht nicht explizit berechnet zu werden. Es genügt, die Vektoren $\boldsymbol{v}^{(k)}$ und die γ_k zu speichern. So sind die HOUSEHOLDER-Transformationen

$$\boldsymbol{H}^{(k)} = \boldsymbol{I} - \boldsymbol{v}^{(k)}\boldsymbol{v}^{(k)^T}/\gamma_k$$

und damit die Matrix $\boldsymbol{Q}$ jederzeit kostengünstig rekonstruierbar. Die Vektoren $\boldsymbol{v}^{(k)}$, $k = 1,\ldots,n-1$, dürfen ähnlich wie bei der LU-Zerlegung auf dem freiwerdenden

Speicherplatz unterhalb der Diagonalen abgelegt werden. Man benötigt in der k-ten Spalte $n-k+1$ Speicherplätze für $\boldsymbol{v}^{(k)}$. Das ist ein Speicherplatz mehr als unterhalb der Diagonalen frei wird. Darum benutzt man üblicherweise die Positionen k bis n in der k-ten Spalte zum Speichern von $\boldsymbol{v}^{(k)}$. Die $\varrho_k = r_{kk}$ und die γ_k werden in zwei zusätzlichen Feldern der Länge n gespeichert. Mit diesen Konventionen geben wir nun den Algorithmus an.

8.47. QR**-Zerlegung einer Matrix nach** HOUSEHOLDER**:**
Es ist die reguläre (n,n)-Matrix $\boldsymbol{A}$ in ein Produkt $\boldsymbol{A} = \boldsymbol{QR}$ mit einer orthogonalen Matrix $\boldsymbol{Q}$ und einer oberen Dreiecksmatrix $\boldsymbol{R}$ zu zerlegen.

{Initialisierung}
Wähle eine Genauigkeitsschranke $\varepsilon > 0$.
{QR-Zerlegung}
for $k = 1$ *to* $n-1$ ***do***
  {Berechnen der Transformationsmatrix $\boldsymbol{H}^{(k)}$)}
  $\varrho_k = \sqrt{a_{kk}^2 + a_{k+1,k}^2 + \cdots + a_{nk}^2}$
  if $\varrho_k \leqq \varepsilon$ ***then***
    STOPP
  endif
  if $a_{kk} > 0$ ***then***
    $\varrho_k = -\varrho_k$
  endif
  $a_{kk} = a_{kk} - \varrho_k$
  $\gamma_k = -\varrho_k \cdot a_{kk}$
  {Transformation der Restmatrix}
  for $j = k+1$ *to* n ***do***
    $\beta = 0$
    for $i = k$ *to* n ***do***
      $\beta = \beta + a_{ik} \cdot a_{ij}$
    endfor
    $\beta = \beta / \gamma_k$
    for $i = k$ *to* n ***do***
      $a_{ij} = a_{ij} - \beta \cdot a_{ik}$
    endfor
  endfor
endfor
$\varrho_n = a_{nn}$

Aufwand: $\sim 2n^3/3$ *Additionen/Multiplikationen,* n *Quadratwurzeln.*

Bei gegebener QR-Zerlegung einer Matrix $\bar{\boldsymbol{A}}$ lässt sich dann jedes lineare System $\boldsymbol{Ax} = \boldsymbol{b}$ in zwei Schritten lösen:

$$\boldsymbol{c} = \boldsymbol{Q}^T\boldsymbol{b}, \quad \boldsymbol{Rx} = \boldsymbol{c}.$$

Wurde die Matrix $\boldsymbol{Q}$ als Produkt von HOUSEHOLDER-Matrizen abgespeichert, so stellt sich $\boldsymbol{Q}^T$ ebenfalls als Produkt von HOUSEHOLDER-Matrizen dar. Aus

$$\boldsymbol{Q} = \boldsymbol{H}^{(1)}\boldsymbol{H}^{(2)}\ldots\boldsymbol{H}^{(n-2)}\boldsymbol{H}^{(n-1)}$$

folgt wegen der Symmetrie der HOUSEHOLDER-Matrizen

$$\boldsymbol{Q}^T = \boldsymbol{H}^{(n-1)}\boldsymbol{H}^{(n-2)}\ldots\boldsymbol{H}^{(2)}\boldsymbol{H}^{(1)}.$$

Damit lässt sich $\boldsymbol{c} = \boldsymbol{Q}^T\boldsymbol{b}$ nach folgendem Algorithmus berechnen.

8.48. Lösen eines Gleichungssystems $\boldsymbol{Ax} = \boldsymbol{b}$ bei gegebener QR-Zerlegung der Matrix $\boldsymbol{A}$:
Es ist das Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ zu lösen.
Für die Matrix $\boldsymbol{A}$ wurde mit obigem Algorithmus eine QR-Zerlegung berechnet.

{Berechnen von $\boldsymbol{c} = \boldsymbol{Q}^T\boldsymbol{b}$}
for $k = 1$ ***to*** $n-1$ ***do***
  $\beta = 0$
  for $i = k$ ***to*** n ***do***
    $\beta = \beta + b_i \cdot a_{ik}$
  endfor
  $\beta = \beta/\gamma_k$
  for $i = k$ ***to*** n ***do***
    $b_i = b_i - \beta \cdot a_{ik}$
  endfor
endfor
{Lösen von $\boldsymbol{Rx} = \boldsymbol{c}$}
for $k = n$ ***to*** 1 ***step*** -1 ***do***
  for $i = k+1$ ***to*** n ***do***
    $b_k = b_k - b_i \cdot a_{ki}$
  endfor
  $b_k = b_k/\varrho_k$
endfor

Aufwand: $\sim n^2$ Additionen/Multiplikationen.

Ohne Beweis sei ein Satz über das Rundungsfehlerverhalten des HOUSEHOLDER-Verfahrens angegeben.

8.49. Satz: *Die* HOUSEHOLDER*-Orthogonalisierung ist für eine reguläre Matrix* $\boldsymbol{A}$ *mit* $\varrho_k = r_{kk} \neq 0$, $k = 1, \ldots, n$, *durchführbar, falls*

$$\kappa = \text{eps} F \text{cond}(\boldsymbol{A}) < 1$$

mit

$$F = 3.14 \left(1 + \frac{3}{n}\right) n^{5/2} \approx 4n^{5/2}$$

gilt. Es existiert eine orthogonale Matrix $\hat{\boldsymbol{Q}}$, *so dass für den berechneten Dreiecksfaktor* $\boldsymbol{\mathcal{R}}$ *gilt*

$$\hat{\boldsymbol{Q}}\boldsymbol{\mathcal{R}} = \boldsymbol{A} + \boldsymbol{\delta A}$$

mit

$$\|\boldsymbol{\delta A}\| \leqq \text{eps} F \|\boldsymbol{A}\|.$$

Für die durch die Vektoren $\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(n-1)}$ *repräsentierte Matrix* $\boldsymbol{\mathcal{Q}}$ *gilt*

$$\tilde{\boldsymbol{c}} = \text{gl}\left(\boldsymbol{\mathcal{Q}}^T \boldsymbol{b}\right) = \hat{\boldsymbol{Q}}^T (\boldsymbol{b} + \boldsymbol{\delta b})$$

mit

$$\|\boldsymbol{\delta b}\| \leqq \text{eps} \frac{F}{\sqrt{n}} \|\boldsymbol{b}\|.$$

Bemerkung: Das zu lösende Dreieckssystem $\boldsymbol{\mathcal{R}}\boldsymbol{x} = \tilde{\boldsymbol{c}}$ entsteht damit durch eine exakte orthogonale Transformation mit der Matrix $\hat{\boldsymbol{Q}}$ aus dem gestörten Gleichungssystem

$$(\boldsymbol{A} + \boldsymbol{\delta A})\boldsymbol{x} = \boldsymbol{b} + \boldsymbol{\delta b}.$$

Das Lösen eines linearen Gleichungssystems mit dem HOUSEHOLDER-Verfahren ist demnach ein numerisch gutartiger Prozess.

8.50. Beispiel: Wir betrachten wieder die Matrix

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

mit $\varepsilon >$ eps und $\varepsilon^2 <$ eps.
Das HOUSEHOLDER-Verfahren liefert im Maschinenzahlbereich $\mathbb{M}(10,2,.,.)$ die Faktoren

$$\boldsymbol{\mathcal{Q}} = \begin{pmatrix} -1 & \frac{\varepsilon}{\sqrt{2}} & \frac{\varepsilon}{\sqrt{6}} & -\frac{\varepsilon}{\sqrt{3}} \\ -\varepsilon & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ 0 & 0 & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{pmatrix}, \quad \boldsymbol{\mathcal{R}} = \begin{pmatrix} -1 & -1 & -1 & -1 \\ 0 & \varepsilon\sqrt{2} & \frac{\varepsilon}{\sqrt{2}} & \frac{\varepsilon}{\sqrt{2}} \\ 0 & 0 & \frac{3\varepsilon}{\sqrt{6}} & \frac{\varepsilon}{\sqrt{6}} \\ 0 & 0 & 0 & -\frac{\varepsilon}{\sqrt{3}} \end{pmatrix}.$$

Damit gilt

$$\boldsymbol{\delta A} = \boldsymbol{\mathcal{Q}}\boldsymbol{\mathcal{R}} - \boldsymbol{A} = \begin{pmatrix} 0 & \varepsilon^2 & \varepsilon^2 & \varepsilon^2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

und

$$\boldsymbol{\mathcal{Q}}^T\boldsymbol{\mathcal{Q}} - \boldsymbol{I} = \begin{pmatrix} \varepsilon^2 & 0 & 0 & 0 \\ 0 & -\frac{\varepsilon^2}{2} & \frac{\sqrt{3}\varepsilon^2}{6} & -\frac{\sqrt{6}\varepsilon^2}{6} \\ 0 & \frac{\sqrt{3}\varepsilon^2}{6} & \frac{\varepsilon^2}{6} & -\frac{\sqrt{2}\varepsilon^2}{3} \\ 0 & -\frac{\sqrt{6}\varepsilon^2}{6} & -\frac{\sqrt{2}\varepsilon^2}{3} & \frac{\varepsilon^2}{3} \end{pmatrix}.$$

Das ist eine im Rahmen der Maschinengenauigkeit exakte Zerlegung mit einer im Rahmen der Maschinengenauigkeit exakten orthogonalen Matrix. ♡

8.3. Iterative Verfahren

8.3.1. Iterationsverfahren und ihre Konvergenz

Neben den direkten Lösungsverfahren für lineare Gleichungssysteme, die nach einer wohlbestimmten Maximalzahl von Rechenoperationen eine Lösung liefern, existieren iterative Verfahren, die zu einem linearen Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ und einem Startvektor $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ eine Folge $\{\boldsymbol{x}^{(i)}\}_{i\in\mathbb{N}} \subset \mathbb{R}^n$ liefern, die gegen die Lösung des Gleichungssystems konvergiert. Charakteristisch für die direkten Verfahren war, dass die Koeffizientenmatrix des Systems in ein Produkt einfacherer Matrizen zerlegt wurde. Damit dürfen wir die direkten Verfahren auch als Zerlegungsmethoden oder Faktorisierungsverfahren bezeichnen. Charakteristisch für die iterativen Verfahren

wird sein, dass man die Matrix $\boldsymbol{A}$ in unveränderter Form im Algorithmus verwendet. Wie wir später sehen werden, benötigen wir eigentlich nicht die Matrix $\boldsymbol{A}$ selbst, sondern nur einen Algorithmus, der zu gegebenem $\boldsymbol{x} \in \mathbb{R}^n$ den Vektor $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ berechnet. Damit sind diese Verfahren besonders für große Systeme mit schwach besetzten Matrizen attraktiv.
Damit ein Iterationsverfahren sinnvoll und gegenüber einem direkten Verfahren konkurrenzfähig ist, sollte es folgende Bedingungen erfüllen:

A1 Der Aufwand für den Übergang $\boldsymbol{x}^{(i)} \longrightarrow \boldsymbol{x}^{(i+1)}$ sollte möglichst gering sein. Die Anzahl der benötigten Rechenoperationen für einen Iterationsschritt sollte in der Größenordnung n^2 oder kleiner liegen. Der Aufwand für einen Iterationsschritt sollte den Aufwand für das Berechnen von Matrix $\times$ Vektor nicht wesentlich überschreiten.

A2 Das Verfahren sollte für beliebige Startvektoren $\boldsymbol{x}^{(0)}$ konvergieren. Die Konvergenzgeschwindigkeit sollte möglichst groß sein.

Wir betrachten folgenden allgemeinen Ansatz: Es sei $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ ein lineares Gleichungssystem mit der regulären Matrix $\boldsymbol{A}$; der Vektor $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$ ist die exakte Lösung dieses Systems. Weiterhin sei $\boldsymbol{B}$ eine beliebige reguläre Matrix. Dann gilt

$$\begin{aligned}
(\boldsymbol{A} - \boldsymbol{B} + \boldsymbol{B})\boldsymbol{x} &= \boldsymbol{b}, \\
\boldsymbol{B}\boldsymbol{x} &= (\boldsymbol{B} - \boldsymbol{A})\boldsymbol{x} + \boldsymbol{b}, \\
\boldsymbol{x} &= \boldsymbol{B}^{-1}(\boldsymbol{B} - \boldsymbol{A})\boldsymbol{x} + \boldsymbol{B}^{-1}\boldsymbol{b}, \\
\boldsymbol{x} &= (\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A})\boldsymbol{x} + \boldsymbol{B}^{-1}\boldsymbol{b}.
\end{aligned}$$

Das lineares Gleichungssystem $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ ist zur Fixpunktaufgabe

$$\boldsymbol{x} = \Phi(\boldsymbol{x}) = (\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A})\boldsymbol{x} + \boldsymbol{B}^{-1}\boldsymbol{b}$$

äquivalent. Zum Lösen der Fixpunktaufgabe liegt es nahe, ein Iterationsverfahren anzuwenden:

Wähle $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$,

Für $k = 0, 1, \ldots$ berechne $\boldsymbol{x}^{(k+1)}$ durch Lösen von

$$\boldsymbol{B}\boldsymbol{x}^{(k+1)} = (\boldsymbol{B} - \boldsymbol{A})\boldsymbol{x}^{(k)} + \boldsymbol{b}.$$

Je nach Wahl von $\boldsymbol{B}$ erhält man verschiedene Iterationsverfahren. Um der Anforderung **A1** zu genügen, ist $\boldsymbol{B}$ möglichst einfach zu wählen. Es bieten sich wieder Diagonalmatrizen, Dreiecksmatrizen oder orthogonale Matrizen an. Die Forderung **A2** ist offensichtlich erfüllt, falls Φ auf dem gesamten $\mathbb{R}^n$ ein kontrahierender Operator ist. Das ist eine Forderung an die Matrix $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A}$. Es gilt der

8.51. Satz:

1. *Das Iterationsverfahren*

   $$\boldsymbol{x}^{(k+1)} = (\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A})\boldsymbol{x}^{(k)} + \boldsymbol{B}^{-1}\boldsymbol{b}$$

   ist konvergiert dann und nur dann, wenn der Spektralradius der Matrix $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A}$ *kleiner als* 1 *ist:*

   $$\varrho(\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A}) < 1.$$

2. *Hinreichend für die Konvergenz ist die Bedingung*

   $$\|(\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A}\| < 1$$

   bezüglich einer beliebigen Matrixnorm, die mit einer Vektornorm verträglich ist.

Beweis:

1. Wir nehmen zunächst an, dass die Iteration konvergiert. Dann gilt

   $$\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}$$

   Mit $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. Für den Fehlervektor $\boldsymbol{f}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}$ folgt

   $$\begin{aligned}
   \boldsymbol{f}^{(k+1)} &= \boldsymbol{x}^{(k+1)} - \boldsymbol{x} \\
   &= (\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A})\boldsymbol{x}^{(k)} + \boldsymbol{B}^{-1}\boldsymbol{b} - \boldsymbol{x} \\
   &= (\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A})(\boldsymbol{x}^{(k)} - \boldsymbol{x}) + (\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A})\boldsymbol{x} + \boldsymbol{B}^{-1}\boldsymbol{b} - \boldsymbol{x} \\
   &= (\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A})\boldsymbol{f}^{(k)} + \boldsymbol{x} - \boldsymbol{B}^{-1}\boldsymbol{b} + \boldsymbol{B}^{-1}\boldsymbol{b} - \boldsymbol{x} \\
   &= (\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A})\boldsymbol{f}^{(k)}.
   \end{aligned}$$

   Aus der Konvergenz folgt weiterhin

   $$\lim_{k\to\infty} \boldsymbol{f}^{(k)} = \boldsymbol{o}.$$

Wählt man den Startvektor so, dass $\boldsymbol{f}^{(0)}$ ein Eigenvektor der Matrix $\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A}$ zum Eigenwert λ ist, so gilt

$$\boldsymbol{f}^{(k)} = (\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A})^k \boldsymbol{f}^{(0)} = \lambda^k \boldsymbol{f}^{(0)}.$$

Wegen $\boldsymbol{f}^{(0)} \neq \boldsymbol{o}$ muss dann aber

$$\lim_{k\to\infty} \lambda^k = 0$$

und damit $|\lambda| < 1$ gelten. Da λ ein beliebiger Eigenwert von $\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A}$ war, folgt $\varrho(\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A}) < 1$.
Nun nehmen wir an, dass $\varrho(\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A}) < 1$ gilt. Es folgt

$$\lim_{k\to\infty} \varrho\left((\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A})^k\right) = \lim_{k\to\infty} \left(\varrho(\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A})\right)^k = 0,$$

also

$$\lim_{k\to\infty} (\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A})^k = \boldsymbol{O}$$

und

$$\lim_{k\to\infty} \boldsymbol{f}^{(k)} = (\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A})^k \boldsymbol{f}^{(0)} = \boldsymbol{o}.$$

2. Nach Satz 8.7 folgt aus $\|\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A}\| < 1$ sofort

$$\varrho(\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A}) \leqq \text{lub}(\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A}) \leqq \|\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A}\| < 1$$

und mit der ersten Aussage des Satzes die Konvergenz des Verfahrens.

❀

Die Konvergenzgeschwindigkeit des Verfahrens hängt von $\varrho(\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A})$ ab. Je kleiner der Spektralradius, desto schneller die Konvergenz.

8.3.2. Das Jacobi- und das Gauß-Seidel-Verfahren

Es sei $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ eine Matrix mit $a_{ii} \neq 0$ für $i = 1, \ldots, n$. Wir wählen

$$\boldsymbol{B} = \text{diag}(a_{11}, a_{22}, \ldots, a_{nn}).$$

Damit ergibt sich aus dem allgemeinen Ansatz für Iterationsverfahren das JACOBI-**Verfahren**:

8.52. JACOBI**-Verfahren:**
Es ist das Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ zu lösen. Für die Matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ gelte

$$a_{ii} \neq 0, \quad i = 1, \ldots, n.$$

S0 *Wähle einen Startvektor $\boldsymbol{x}^{(0)} \in \mathbb{R}^{n\times n}$ und setze $i = 0$.*

S1 *Berechne*
for $k = 1$ **to** n **do**

$$x_k^{(i+1)} = \frac{1}{a_{kk}} \left(b_k - \sum_{\substack{j=1 \\ j\neq k}}^{n} a_{kj} x_j^{(i)} \right)$$

end

S2 *Setze $i = i + 1$ und gehe zu Schritt* **S1**.

Das JACOBI-Verfahren wird auch als **Iteration in Gesamtschritten** bezeichnet. Als Iterationsmatrix $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A}$ erhält man

$$\boldsymbol{M} = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & \cdots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \cdots & -\frac{a_{2n}}{a_{22}} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & \cdots & 0 \end{pmatrix}.$$

Nach Satz 8.51 konvergiert das Gesamtschrittverfahren genau dann, wenn $\varrho(\boldsymbol{M}) < 1$ gilt. Diese Bedingung läßt sich im allgemeinen schwer nachprüfen. Hinreichende Konvergenzbedingungen erhält man jedoch leicht.

8.53. Satz:

1. **Starkes Zeilensummenkriterium**
   Das Gesamtschrittverfahren konvergiert für alle Matrizen $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ mit

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \quad i = 1, \ldots, n.$$

2. **Starkes Spaltensummenkriterium**
   Das Gesamtschrittverfahren konvergiert für alle Matrizen $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ mit

$$|a_{jj}| > \sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{ij}| \quad \textit{für} \quad j = 1, \ldots, n.$$

Beweis:

1. Aus

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \quad i = 1, \ldots, n$$

folgt

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1 \quad i = 1, \ldots, n$$

und

$$1 > \max_{1 \leq i \leq n} \left\{ \sum_{\substack{j=1 \\ j\neq i}}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| \right\} = \|\boldsymbol{M}\|_{\infty}.$$

Damit ist das hinreichende Konvergenzkriterium aus Satz 8.51 erfüllt.

2. Ist $\boldsymbol{A}$ streng spaltendiagonaldominant, so ist $\boldsymbol{A}^T$ streng zeilendiagonaldominant. Damit konvergiert das JACOBI-Verfahren für die Matrix $\boldsymbol{A}^T$. Dann ist nach Satz 8.51 der Spektralradius der Iterationsmatrix $\bar{\boldsymbol{M}} = \boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A}^T$ kleiner als 1. Da die Matrix

$$\boldsymbol{B}\bar{\boldsymbol{M}}\boldsymbol{B}^{-1} = \boldsymbol{I} - \boldsymbol{A}^T\boldsymbol{B}^{-1} = \left(\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A}\right)^T$$

dieselben Eigenwerte wie $\bar{\boldsymbol{M}}$ hat, folgt

$$\varrho(\boldsymbol{M}) = \varrho(\boldsymbol{M}^T) = \varrho(\boldsymbol{B}\bar{\boldsymbol{M}}\boldsymbol{B}^{-1}) = \varrho(\bar{\boldsymbol{M}}) < 1.$$

Damit konvergiert das Gesamtschrittverfahren auch für die Matrix $\boldsymbol{A}$.

❀

Benutzt man in der Iterationsvorschrift des JACOBI-Verfahrens

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij} x_j^{(k)} \right)$$

beim Berechnen der i-ten Komponente des neuen Iterationsvektors statt der Komponenten

$$x_1^{(k)}, \ldots, x_{i-1}^{(k)}$$

die schon berechneten neuen Komponenten

$$x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)},$$

so ergibt sich das GAUSS-SEIDEL-**Verfahren**.

8.54. GAUSS-SEIDEL-**Verfahren:**
Es ist das Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ zu lösen. Für die Matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ gelte

$$a_{ii} \neq 0, \quad i = 1, \ldots, n.$$

S0 *Wähle einen Startvektor $\boldsymbol{x}^{(0)} \in \mathbb{R}^{n\times n}$ und setze $i = 0$.*

S1 *Berechne*
for $k = 1$ **to** n **do**

$$x_k^{(i+1)} = \frac{1}{a_{kk}} \left(b_k - \sum_{j=1}^{k-1} a_{kj} x_j^{(i+1)} - \sum_{j=k+1}^{n} a_{kj} x_j^{(i)} \right)$$

end

S2 *Setze $i = i + 1$ und gehe zu Schritt* **S1**.

Dieses Verfahren wird auch als **Einzelschrittverfahren** bezeichnet. Dem Verfahren entspricht die Wahl

$$\boldsymbol{B} = \begin{pmatrix} a_{11} & & & \\ a_{21} & a_{22} & & \boldsymbol{O} \\ \vdots & \vdots & \ddots & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}.$$

Als hinreichende Bedingungen für die Konvergenz des Einzelschrittverfahrens ergeben sich ebenfalls das starke Zeilen- oder Spaltensummenkriterium aus Satz 8.53.

8.55. Satz:

1. **Starkes Zeilensummenkriterium**
   Das Einzelschrittverfahren konvergiert für alle Matrizen $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ mit

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \quad i = 1, \ldots, n.$$

2. **Starkes Spaltensummenkriterium**
Das Einzelschrittverfahren konvergiert für alle Matrizen $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ mit

$$|a_{jj}| > \sum_{\substack{i=1\\ i\neq j}}^{n} |a_{ij}| \quad j = 1,\ldots,n.$$

Beweis: Wir beweisen nur die erste Aussage. Es sei

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} = -\boldsymbol{E} + \boldsymbol{D} - \boldsymbol{F}$$

mit

$$\boldsymbol{E} = -\begin{pmatrix} 0 & \cdots & \cdots & 0 \\ a_{21} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{pmatrix}, \qquad \boldsymbol{D} = \text{diag}(a_{11}, a_{22}, \ldots, a_{nn})$$

und

$$\boldsymbol{F} = -\begin{pmatrix} 0 & a_{12} & \cdots & a_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & a_{n-1,n} \\ 0 & \cdots & \cdots & 0 \end{pmatrix}.$$

Aus der Zeilendiagonaldominanz der Matrix $\boldsymbol{A}$ folgt $a_{ii} \neq 0$ für $i = 1,\ldots,n$. Damit ist $\boldsymbol{D}$ regulär. Wir definieren weiterhin die Matrizen $\boldsymbol{L} = \boldsymbol{D}^{-1}\boldsymbol{E}$ und $\boldsymbol{U} = \boldsymbol{D}^{-1}\boldsymbol{F}$. Als Iterationsmatrizen für das JACOBI- bzw. GAUSS-SEIDEL-Verfahren erhalten wir

$$\boldsymbol{M}_J = \boldsymbol{I} - \boldsymbol{D}^{-1}(-\boldsymbol{E} + \boldsymbol{D} - \boldsymbol{F}) = \boldsymbol{L} + \boldsymbol{U}$$

und

$$\boldsymbol{M}_{GS} = \boldsymbol{I} - (-\boldsymbol{E} + \boldsymbol{D})^{-1}(-\boldsymbol{E} + \boldsymbol{D} - \boldsymbol{F}) = (-\boldsymbol{D}\boldsymbol{L} + \boldsymbol{D})^{-1}\boldsymbol{D}\boldsymbol{U} = (\boldsymbol{I} - \boldsymbol{L})^{-1}\boldsymbol{U}.$$

Für eine streng zeilendiagonaldominante Matrix $\boldsymbol{A}$ gilt dann $\|\boldsymbol{M}_J\|_\infty < 1$. Wir zeigen, dass

$$\|\boldsymbol{M}_{GS}\|_\infty \leqq \|\boldsymbol{M}_J\|_\infty < 1$$

gilt. Aus $\|\boldsymbol{M}_J\|_\infty < 1$ folgt

$$|\boldsymbol{M}_J|\boldsymbol{e} \leqq \|\boldsymbol{M}_J\|_\infty \boldsymbol{e} < \boldsymbol{e} \quad \text{mit} \quad \boldsymbol{e} = (1,1,\ldots,1)^T \in \mathbb{R}^n.$$

Weiterhin gilt $|\boldsymbol{M}_J| = |\boldsymbol{L}| + |\boldsymbol{U}|$ wegen der speziellen Struktur der Matrizen $\boldsymbol{L}$ und $\boldsymbol{U}$. Daraus folgt

$$|\boldsymbol{U}|\boldsymbol{e} = (|\boldsymbol{M}_J| - |\boldsymbol{L}|)\boldsymbol{e} \leqq (\|\boldsymbol{M}_J\|_\infty \boldsymbol{I} - |\boldsymbol{L}|)\boldsymbol{e}. \tag{8.13}$$

$\boldsymbol{L}$ und $|\boldsymbol{L}|$ sind untere Dreiecksmatrizen mit verschwindender Diagonale. Man sieht leicht ein, dass für diese Matrizen

$$\boldsymbol{L}^n = |\boldsymbol{L}|^n = \boldsymbol{O}$$

gilt. Damit existieren die Inversen von $\boldsymbol{I} - \boldsymbol{L}$ und $\boldsymbol{I} - |\boldsymbol{L}|$, und es gilt

$$\boldsymbol{O} \leqq \left|(\boldsymbol{I} - \boldsymbol{L})^{-1}\right| = \left|\boldsymbol{I} + \boldsymbol{L} + \cdots + \boldsymbol{L}^{n-1}\right| \leqq \boldsymbol{I} + |\boldsymbol{L}| + \cdots + |\boldsymbol{L}|^{n-1} = (\boldsymbol{I} - |\boldsymbol{L}|)^{-1}.$$

Mit dieser Ungleichung folgt aus 8.13

$$\begin{aligned} |\boldsymbol{M}_{GS}|\boldsymbol{e} &\leqq \left|(\boldsymbol{I} - \boldsymbol{L})^{-1}\right| |\boldsymbol{U}|\boldsymbol{e} \\ &\leqq (\boldsymbol{I} - |\boldsymbol{L}|)^{-1} (\|\boldsymbol{M}_J\|_\infty \boldsymbol{I} - |\boldsymbol{L}|)\boldsymbol{e} \\ &= (\boldsymbol{I} - |\boldsymbol{L}|)^{-1} (\boldsymbol{I} - |\boldsymbol{L}| + (\|\boldsymbol{M}_J\|_\infty - 1)\boldsymbol{I})\,\boldsymbol{e} \\ &= \left(\boldsymbol{I} + (\|\boldsymbol{M}_J\|_\infty - 1)\,(\boldsymbol{I} - |\boldsymbol{L}|)^{-1}\right)\boldsymbol{e}. \end{aligned}$$

Wegen $(\boldsymbol{I} - |\boldsymbol{L}|)^{-1} \geqq \boldsymbol{I}$ und $\|\boldsymbol{M}_J\|_\infty < 1$ ergibt sich

$$|\boldsymbol{M}_{GS}|\boldsymbol{e} \leqq [\boldsymbol{I} + (\|\boldsymbol{M}_J\|_\infty - 1)\boldsymbol{I}]\,\boldsymbol{e} = \|\boldsymbol{M}_J\|_\infty \boldsymbol{e}$$

und damit

$$\|\boldsymbol{M}_{GS}\|_\infty \leqq \|\boldsymbol{M}_J\|_\infty.$$

❀

Welches von beiden Verfahren für eine beliebige Matrix konvergiert und welches schneller konvergiert, läßt sich nicht ohne weiteres angeben.

8.56. Beispiel: Für die Matrix

$$\boldsymbol{A} = \begin{pmatrix} 1 & -2 & 2 \\ -1 & 1 & -1 \\ -2 & -2 & 1 \end{pmatrix}$$

ergeben sich für das JACOBI- bzw. GAUSS-SEIDEL-Verfahren die Iterationsmatrizen

$$\boldsymbol{M}_J = \begin{pmatrix} 0 & 2 & -2 \\ 1 & 0 & 1 \\ 2 & 2 & 0 \end{pmatrix} \quad \text{und} \quad \boldsymbol{M}_{GS} = \begin{pmatrix} 0 & 2 & -2 \\ 0 & 2 & -1 \\ 0 & 8 & -6 \end{pmatrix}$$

mit $\varrho(\boldsymbol{M}_J) = 0$ und $\varrho(\boldsymbol{M}_{GS}) = 2(1+\sqrt{2})$. Hier konvergiert das JACOBI-Verfahren, das GAUSS-SEIDEL-Verfahren dagegen nicht. Für die Matrix

$$\boldsymbol{A} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ -1 & 1 & -1 \\ -\frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix}$$

ergeben sich

$$\boldsymbol{M}_J = \begin{pmatrix} 0 & -\frac{1}{2} & -\frac{1}{2} \\ 1 & 0 & 1 \\ \frac{1}{2} & -\frac{1}{2} & 0 \end{pmatrix} \quad \text{und} \quad \boldsymbol{M}_{GS} = \begin{pmatrix} 0 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & -\frac{1}{2} & \frac{1}{2} \\ 0 & 0 & -\frac{1}{2} \end{pmatrix}$$

mit

$$\varrho(\boldsymbol{M}_J) = \sqrt{5}/2, \quad \varrho(\boldsymbol{M}_{GS}) = 1/2.$$

Hier konvergiert das GAUSS-SEIDEL-Verfahren; das JACOBI-Verfahren jedoch konvergiert nicht. ♡

Das starke Zeilen- bzw. Spaltensummenkriterium läßt sich abschwächen. Dazu führen wir den Begriff der unzerlegbaren Matrizen ein. Eine Matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ heißt **unzerlegbar (irreduzibel)**, falls keine Permutationsmatrix $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ existiert, so dass $\boldsymbol{P}^T\boldsymbol{A}\boldsymbol{P}$ die Gestalt

$$\boldsymbol{P}^T\boldsymbol{A}\boldsymbol{P} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{O} & \boldsymbol{A}_{22} \end{pmatrix}$$

mit $\boldsymbol{A}_{11} \in \mathbb{R}^{k\times k}$, $\boldsymbol{A}_{22} \in \mathbb{R}^{(n-k)\times(n-k)}$ und $1 \leqq k \leqq n-1$ besitzt. Die Unzerlegbarkeit einer Matrix läßt sich mit Hilfe der Graphentheorie überprüfen. Wir ordnen dazu einer Matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ in folgender Weise einen gerichteten Graphen $G(\boldsymbol{A})$ zu. $G(\boldsymbol{A})$ besitzt die Knoten $P_1, \ldots, P_n$. Vom Knoten P_i zum Knoten P_j existiert genau dann eine gerichtete Kante e_{ij}, wenn $a_{ij} \neq 0$ gilt. Die Matrix $\boldsymbol{A}$ ist genau dann unzerlegbar, wenn der Graph $G(\boldsymbol{A})$ zusammenhängend ist. Dabei heißt ein

Graph zusammenhängend, falls für jedes Paar (i,j) mit $i \neq j$ ein gerichteter Weg vom Knoten P_i zum Knoten P_j existiert.
Nun können wir ein schwächeres Konvergenzkriterium für das JACOBI-Verfahren formulieren.

8.57. Satz: *Für eine unzerlegbare Matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ gilt:*

1. **Schwaches Zeilensummenkriterium**
   Das JACOBI*-Verfahren konvergiert, falls*

$$|a_{ii}| \geqq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \quad \textit{für} \quad i = 1,\ldots,n$$

   und für mindestens ein i^*

$$|a_{i^*i^*}| > \sum_{\substack{j=1 \\ j\neq i^*}}^{n} |a_{i^*j}|$$

   gilt.

2. **Schwaches Spaltensummenkriterium**
   Das JACOBI*-Verfahren konvergiert, falls*

$$|a_{jj}| \geqq \sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{ij}| \quad \textit{für} \quad j = 1,\ldots,n$$

   und für mindestens ein j^*

$$|a_{j^*j^*}| > \sum_{\substack{i=1 \\ i\neq j^*}}^{n} |a_{ij^*}|$$

   gilt.

Beweis: Wir beweisen nur das Schwache Zeilensummenkriterium. Die Gültigkeit des Schwachen Spaltensummenkriteriums folgt dann analog zum Beweis des Starken Spaltensummenkriteriums. Aus dem Schwachen Zeilensummenkriterium folgt für das JACOBI-Verfahren

$$\|\boldsymbol{M}_J\|_\infty = \|\boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{A}\|_\infty \leqq 1.$$

Als hinreichende Bedingung für die Konvergenz benötigen wir aber $\|\boldsymbol{M}_J\|_\infty < 1$. Mit $\boldsymbol{e} = (1,1,\ldots,1)^T$ folgt

$$|\boldsymbol{M}|\boldsymbol{e} \leqq \boldsymbol{e} \quad \text{und} \quad |\boldsymbol{M}|\boldsymbol{e} \neq \boldsymbol{e}.$$

Wir zeigen, dass $|\boldsymbol{M}|^n\boldsymbol{e} < \boldsymbol{e}$ gilt. Daraus folgt dann

$$1 > \||\boldsymbol{M}|^n\|_\infty \geqq (\||\boldsymbol{M}|\|_\infty)^n = (\|\boldsymbol{M}\|)^n \geqq (\varrho(\boldsymbol{M}))^n$$

und endlich $\varrho(\boldsymbol{M}) < 1$. Um die Beziehung $|\boldsymbol{M}|^n\boldsymbol{e} < \boldsymbol{e}$ zu zeigen, betrachten wir die Ungleichungskette

$$\boldsymbol{e} \begin{array}{c} \geqq \\ \neq \end{array} |\boldsymbol{M}|\boldsymbol{e} \geqq |\boldsymbol{M}|^2\boldsymbol{e} \geqq \cdots \geqq |\boldsymbol{M}|^i\boldsymbol{e} \geqq |\boldsymbol{M}|^{i+1}\boldsymbol{e} \geqq \cdots$$

Daraus folgt

$$\boldsymbol{o} \begin{array}{c} \leqq \\ \neq \end{array} \boldsymbol{e} - |\boldsymbol{M}|\boldsymbol{e} \leqq \boldsymbol{e} - |\boldsymbol{M}|^2\boldsymbol{e} \leqq \cdots \leqq \boldsymbol{e} - |\boldsymbol{M}|^i\boldsymbol{e} \leqq \boldsymbol{e} - |\boldsymbol{M}|^{i+1}\boldsymbol{e} \leqq \cdots$$

Von den Vektoren $\boldsymbol{t}^{(i)} = \boldsymbol{e} - |\boldsymbol{M}|^i\boldsymbol{e}$ sind damit gewisse Komponenten gleich Null und die restlichen Komponenten größer als Null. Dann existiert eine Permutationsmatrix $\boldsymbol{P}$, so dass

$$\boldsymbol{P}\boldsymbol{t}^{(i)} = \boldsymbol{P}\left(\boldsymbol{e} - |\boldsymbol{M}|^{i+1}\boldsymbol{e}\right) = \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{o} \end{pmatrix} \quad \text{mit} \quad \boldsymbol{a} \in \mathbb{R}^p \quad \text{und} \quad \boldsymbol{a} > \boldsymbol{o}.$$

Wegen $0 \neq \boldsymbol{t}^{(i)}$ gilt $p < n$. Für den Vektor $\boldsymbol{t}^{(i+1)} = \boldsymbol{e} - |\boldsymbol{M}|^i\boldsymbol{e}$ gilt $\boldsymbol{t}^{(i+1)} \geqq \boldsymbol{t}^{(i)}$, daher

$$\boldsymbol{P}\boldsymbol{t}^{(i+1)} \geqq \boldsymbol{P}\boldsymbol{t}^{(i)} = \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{o} \end{pmatrix}.$$

Es gilt somit

$$\boldsymbol{P}\boldsymbol{t}^{(i+1)} = \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{o} \end{pmatrix} \quad \text{mit} \quad \boldsymbol{b} \in \mathbb{R}^q, \quad \boldsymbol{b} > \boldsymbol{o}$$

und $q \geqq p$. ❀

8.3.3. Nachiteration

Gegeben sei eine näherungsweise LU-Zerlegung der Matrix $\boldsymbol{A}$

$$\boldsymbol{P}^T\boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{U}} = \boldsymbol{A} + \boldsymbol{\delta A}.$$

Wir wählen im allgemeinen Iterationsansatz

$$\boldsymbol{B} = \boldsymbol{P}^T \boldsymbol{\mathcal{L}} \boldsymbol{\mathcal{U}} = \boldsymbol{A} + \boldsymbol{\delta A}$$

und erhalten

$$\begin{aligned} \boldsymbol{M} &= \boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{B}^{-1}(\boldsymbol{B} - \boldsymbol{\delta A}) \\ &= \boldsymbol{B}^{-1}\boldsymbol{\delta A} = (\boldsymbol{A} + \boldsymbol{\delta A})^{-1}\boldsymbol{\delta A}. \end{aligned}$$

Es gilt

$$\|\boldsymbol{M}\| = \|(\boldsymbol{A} + \boldsymbol{\delta A})^{-1}\boldsymbol{\delta A}\| \leqq \|(\boldsymbol{A} + \boldsymbol{\delta A})^{-1}\| \, \|\boldsymbol{\delta A}\|$$

und im Falle $\kappa = \|\boldsymbol{\delta A}\| \, \|\boldsymbol{A}^{-1}\| < 1$

$$\|\boldsymbol{M}\| \leqq \frac{\kappa}{1-\kappa}.$$

Für die LU-Zerlegung mit Spaltenpivotisierung gilt

$$\kappa = \text{eps} F(\boldsymbol{A}) \text{cond}(\boldsymbol{A})$$

mit $F(\boldsymbol{A}) \approx 3 \cdot 2^n$. Die hinreichende Bedingung $\|\boldsymbol{M}\| < 1$ ist erfüllt, falls $\kappa < 1/2$ gilt. Wir erhalten folgenden Algorithmus.

8.58. Nachiteration für die LU-Zerlegung:
Zu lösen sei das lineare Gleichungssystem $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ mit $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ und $\boldsymbol{b} \in \mathbb{R}^n$. Durch $\boldsymbol{\mathcal{L}}$ und $\boldsymbol{\mathcal{U}}$ seien die berechneten Dreiecksfaktoren einer LU-Zerlegung der Matrix $\boldsymbol{A}$ gegeben. Es gelte

$$\boldsymbol{A} + \boldsymbol{\delta A} = \boldsymbol{P}^T \boldsymbol{\mathcal{L}} \boldsymbol{\mathcal{U}}.$$

S0 *Wähle einen Startvektor $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ und setze $i = 0$.*

S1 *Berechne das Residuum*

$$\boldsymbol{r}^{(i)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(i)}.$$

S2 *Löse das Gleichungssystem $\boldsymbol{A}\boldsymbol{\delta x}^{(i)} = \boldsymbol{r}^{(i)}$ näherungsweise gemäß*

$$\boldsymbol{\mathcal{L}}\boldsymbol{y} = \boldsymbol{P}\boldsymbol{r}^{(i)} \quad \textit{und} \quad \boldsymbol{\mathcal{U}}\boldsymbol{\delta x}^{(i)} = \boldsymbol{y}.$$

S3 *Setze $\boldsymbol{x}^{(i+1)} = \boldsymbol{x}^{(i)} + \boldsymbol{\delta x}^{(i)}$, $i = i + 1$ und gehe zu Schritt* **S1**.

Bemerkungen: (i) Wenn das Verfahren konvergiert, werden die $\boldsymbol{x}^{(i)}$ gute Näherungen für die Lösung des Gleichungssystems $\boldsymbol{Ax} = \boldsymbol{b}$ sein. Dann tritt aber beim Berechnen des Residuums in Schritt **S1** Auslöschung auf. Damit das ganze Verfahren noch sinnvoll bleiben soll, muss man an dieser Stelle besonders sorgfältig vorgehen. Am besten berechnet man das Residuum in einer höheren Genauigkeit.
(ii) Üblicherweise wählt man als Startvektor $\boldsymbol{x}^{(0)} = \boldsymbol{o}$. Damit entspricht der erste Schritt des Verfahrens der normalen Lösung des Gleichungssystems mittels LU-Zerlegung.
(iii) Die Konvergenz des Verfahrens ist im allgemeinen gut, falls $\boldsymbol{A}$ nicht zu schlecht konditioniert ist. Es werden nur wenige Schritte ausgeführt.
Es gilt $\boldsymbol{Ax}^{(i)} = \boldsymbol{b} - \boldsymbol{r}^{(i)}$. Wir können $\boldsymbol{r}^{(i)}$ als eine Störung $\boldsymbol{\delta b}$ interpretieren. Nach Satz 8.14 gilt dann ($\boldsymbol{\delta A} = \boldsymbol{O}$)

$$\frac{1}{\operatorname{cond}(\boldsymbol{A})} \frac{\|\boldsymbol{r}^{(i)}\|}{\|\boldsymbol{b}\|} \leqq \frac{\|\boldsymbol{\delta x}^{(i)}\|}{\|\boldsymbol{x}^{(i)}\|} \leqq \operatorname{cond}(\boldsymbol{A}) \frac{\|\boldsymbol{r}^{(i)}\|}{\|\boldsymbol{b}\|}.$$

Diese Ungleichungen werden benutzt, um bei der Nachiteration die Konditionszahl von $\boldsymbol{A}$ abzuschätzen. Man erhält

$$\operatorname{cond}(\boldsymbol{A}) \geqq \frac{\|\boldsymbol{\delta x}^{(i)}\|}{\|\boldsymbol{x}^{(i)}\|} \Bigg/ \frac{\|\boldsymbol{r}^{(i)}\|}{\|\boldsymbol{b}\|}$$

beziehungsweise

$$\operatorname{cond}(\boldsymbol{A}) \geqq \frac{\|\boldsymbol{r}^{(i)}\|}{\|\boldsymbol{b}\|} \Bigg/ \frac{\|\boldsymbol{\delta x}^{(i)}\|}{\|\boldsymbol{x}^{(i)}\|}.$$

Analoge Nachiterationsverfahren sind natürlich auch für andere direkte Verfahren möglich. Für die LDL^T-Zerlegung erhält man folgenden Algorithmus.

8.59. Nachiteration für die LDL^T-Zerlegung:
Zu lösen ist das lineare Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ mit der symmetrischen Matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ und $\boldsymbol{b} \in \mathbb{R}^n$. Durch $\boldsymbol{\mathcal{L}}$ und $\boldsymbol{\mathcal{D}}$ seien die berechneten Faktoren einer LDL^T-Zerlegung der Matrix $\boldsymbol{A}$ gegeben. Es gelte

$$\boldsymbol{A} + \boldsymbol{\delta A} = \boldsymbol{\mathcal{L}}\boldsymbol{\mathcal{D}}\boldsymbol{\mathcal{L}}^T.$$

S0 *Wähle einen Startvektor $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ und setze $i = 0$.*
Berechne das Residuum

$$\boldsymbol{r}^{(i)} = \boldsymbol{b} - \boldsymbol{Ax}^{(i)}.$$

S1 *Löse das Gleichungssystem* $\boldsymbol{A}\boldsymbol{\delta x}^{(i)} = \boldsymbol{r}^{(i)}$ *näherungsweise gemäß*

$$\boldsymbol{\mathcal{L}}\,\boldsymbol{y} = \boldsymbol{r}^{(i)} \quad \textit{und} \quad \boldsymbol{\mathcal{L}}^T \boldsymbol{\delta x}^{(i)} = \boldsymbol{\mathcal{D}}^{-1}\boldsymbol{y}.$$

S2 *Setze* $\boldsymbol{x}^{(i+1)} = \boldsymbol{x}^{(i)} + \boldsymbol{\delta x}^{(i)}$, $i = i+1$ *und gehe zu Schritt* **S1**.

Für ein Orthogonalisierungsverfahren ergibt sich das folgende Nachiterationsverfahren.

8.60. Nachiteration für die QR-Zerlegung:
Zu lösen ist das lineare Gleichungssystem $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ *mit* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *und* $\boldsymbol{b} \in \mathbb{R}^n$. *Durch* $\boldsymbol{\mathcal{Q}}$ *und* $\boldsymbol{\mathcal{R}}$ *seien die berechneten Faktoren einer* QR*-Zerlegung der Matrix* $\boldsymbol{A}$ *gegeben. Es gelte*

$$\boldsymbol{A} + \boldsymbol{\delta A} = \boldsymbol{\mathcal{Q}}\boldsymbol{\mathcal{R}}.$$

S0 *Wähle einen Startvektor* $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ *und setze* $i = 0$.

S1 *Berechne das Residuum*

$$\boldsymbol{r}^{(i)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(i)}.$$

S2 *Löse das Gleichungssystem* $\boldsymbol{A}\boldsymbol{\delta x}^{(i)} = \boldsymbol{r}^{(i)}$ *näherungsweise gemäß*

$$\boldsymbol{y} = \boldsymbol{Q}^T\boldsymbol{r}^{(i)} \quad \textit{und} \quad \boldsymbol{\mathcal{R}}\boldsymbol{\delta x}^{(i)} = \boldsymbol{y}.$$

S3 *Setze* $\boldsymbol{x}^{(i+1)} = \boldsymbol{x}^{(i)} + \boldsymbol{\delta x}^{(i)}$, $i = i+1$ *und gehe zu Schritt* **S1**.

Analoge Konditionszahlschätzungen sind möglich. Das Residuum sollte wieder in höherer Genauigkeit berechnet werden.

8.3.4. Das cg-Verfahren von Hestenes und Stiefel

Das jetzt zu behandelnde Verfahren gehört nicht zu den Iterationsverfahren, wie sie sich aus dem allgemeinen Ansatz in Abschnitt8.3.1. ergeben. Bei genauer Betrachtung müßte man es sogar zu den direkten Verfahren zählen, da man bei exakter Rechnung nach einer a priori bekannten Maximalzahl von Rechenoperationen die exakte Lösung des Gleichungssystems erhält. Da aber bei diesem Verfahren auch eine Folge von Vektoren konstruiert wird, die den exakten Lösungsvektor immer genauer approximieren, behandelt wir das Verfahren an dieser Stelle.
Wir betrachten zunächst ein lineares Gleichungssystem $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ mit einer symmetrischen, positiv definiten Matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Diesem Gleichungssystem ordnen wir ein Optimierungsproblem zu. Es sei

$$F(\boldsymbol{z}) = \frac{1}{2}\left(\boldsymbol{A}\boldsymbol{z} - \boldsymbol{b}\right)^T \boldsymbol{A}^{-1}\left(\boldsymbol{A}\boldsymbol{z} - \boldsymbol{b}\right) = \frac{1}{2}\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z} - \boldsymbol{b}^T\boldsymbol{z} + \frac{1}{2}\boldsymbol{b}^T\boldsymbol{A}^{-1}\boldsymbol{b}.$$

Wegen der positiven Definitheit von $\boldsymbol{A}$ (und damit auch von $\boldsymbol{A}^{-1}$) gilt $F(\boldsymbol{z}) \geqq 0$ für alle $\boldsymbol{z} \in \mathbb{R}^n$ und $F(\boldsymbol{z}) = 0$ genau dann, wenn $\boldsymbol{Az} - \boldsymbol{b} = \boldsymbol{o}$. Das Minimierungsproblem

$$\boldsymbol{x} = \arg\min_{\boldsymbol{z} \in \mathbb{R}^n} F(\boldsymbol{z}) \tag{8.14}$$

ist dem Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ äquivalent. Minimierungsprobleme dieser Art löst man üblicherweise, indem man sie auf viele eindimensionale Minimierungsprobleme zurückführt.

S0 Wähle Startvektor $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ und setze $i = 0$.

S1 Wähle eine Richtung $\boldsymbol{h}^{(i)} \in \mathbb{R}^n$ und löse das eindimensionale Minimierungsproblem

$$\alpha_i = \arg\min_{\alpha \in \mathbb{R}} F(\boldsymbol{x}^{(i)} + \alpha \boldsymbol{h}^{(i)}).$$

S2 Setze $\boldsymbol{x}^{(i+1)} = \boldsymbol{x}^{(i)} + \alpha_i \boldsymbol{h}^{(i)}$.

S3 Setze $i = i + 1$ und gehe zu Schritt **S1**.

Für das gegebene Funktional $F(\boldsymbol{z})$ und beliebiges $\boldsymbol{x}$ und eine beliebige Richtung $\boldsymbol{h}$ läßt sich die Lösung des eindimensionalen Minimierungsproblems in Schritt **S1** leicht angeben. Es gilt der folgende

8.61. Satz: *Für die symmetrische, positiv definite Matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ und einen Vektor $\boldsymbol{b} \in \mathbb{R}^n$ sei das Funktional $F : \mathbb{R}^n \to \mathbb{R}_+$ folgendermaßen definiert:*

$$F(\boldsymbol{z}) = \frac{1}{2} (\boldsymbol{Az} - \boldsymbol{b})^T \boldsymbol{A}^{-1} (\boldsymbol{Az} - \boldsymbol{b}).$$

Weiterhin seien ein Vektor $\boldsymbol{x} \in \mathbb{R}^n$ und ein Vektor $\boldsymbol{h} \in \mathbb{R}^n$ mit $\boldsymbol{h} \neq \boldsymbol{o}$ gegeben. Dann hat das Minimierungsproblem

$$\alpha^* = \arg\min_{\alpha \in \mathbb{R}} F(\boldsymbol{x} + \alpha \boldsymbol{h})$$

die eindeutige Lösung

$$\alpha^* = \frac{\boldsymbol{h}^T \boldsymbol{r}}{\boldsymbol{h}^T \boldsymbol{A} \boldsymbol{h}} \quad \textit{mit} \quad \boldsymbol{r} = \boldsymbol{b} - \boldsymbol{Ax}.$$

Es gilt

$$\boldsymbol{h}^T [\boldsymbol{A}(\boldsymbol{x} + \alpha^* \boldsymbol{h}) - \boldsymbol{b}] = 0.$$

Beweis: Aus

$$F(\boldsymbol{z}) = \frac{1}{2}\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z} - \boldsymbol{b}^T\boldsymbol{z} + \frac{1}{2}\boldsymbol{b}^T\boldsymbol{A}^{-1}\boldsymbol{b}$$

folgt $F'(\boldsymbol{z}) = \boldsymbol{A}\boldsymbol{z} - \boldsymbol{b}$ und

$$\frac{d}{d\alpha}F(\boldsymbol{x}+\alpha\boldsymbol{h}) = \boldsymbol{h}^T\left[\boldsymbol{A}(\boldsymbol{x}+\alpha\boldsymbol{h}) - \boldsymbol{b}\right].$$

Eine notwendige Bedingung für ein Extremum ist

$$\frac{d}{d\alpha}F(\boldsymbol{x}+\alpha\boldsymbol{h})\bigg|_{\alpha^*} = 0.$$

Daraus ergibt sich

$$\boldsymbol{h}^T\left[\boldsymbol{A}(\boldsymbol{x}+\alpha^*\boldsymbol{h}) - \boldsymbol{b}\right] = 0$$

und

$$\alpha^* = \frac{\boldsymbol{h}^T(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x})}{\boldsymbol{h}^T\boldsymbol{A}\boldsymbol{h}} = \frac{\boldsymbol{h}^T\boldsymbol{r}}{\boldsymbol{h}^T\boldsymbol{A}\boldsymbol{h}}.$$

Es ist noch zu zeigen, dass α^* ein Minimum liefert. Es gilt

$$\begin{aligned}
F(\boldsymbol{x}+\alpha^*\boldsymbol{h}) &= \frac{1}{2}(\boldsymbol{A}\boldsymbol{x}+\alpha^*\boldsymbol{A}\boldsymbol{h}-\boldsymbol{b})^T\boldsymbol{A}^{-1}(\boldsymbol{A}\boldsymbol{x}+\alpha^*\boldsymbol{A}\boldsymbol{h}-\boldsymbol{b}) \\
&= \frac{1}{2}(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}+\alpha^*\boldsymbol{A}\boldsymbol{h})^T\boldsymbol{A}^{-1}(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}+\alpha^*\boldsymbol{A}\boldsymbol{h}) \\
&= F(\boldsymbol{x}) + \frac{1}{2}\alpha^*\boldsymbol{h}^T(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}) + \frac{1}{2}\alpha^*(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})^T\boldsymbol{h} + \frac{1}{2}\alpha^{*2}\boldsymbol{h}^T\boldsymbol{A}\boldsymbol{h} \\
&= F(\boldsymbol{x}) - \alpha^*\boldsymbol{h}^T\boldsymbol{r} + \frac{1}{2}\alpha^{*2}\boldsymbol{h}^T\boldsymbol{A}\boldsymbol{h} \\
&= F(\boldsymbol{x}) - \frac{(\boldsymbol{h}^T\boldsymbol{r})^2}{\boldsymbol{h}^T\boldsymbol{A}\boldsymbol{h}} + \frac{1}{2}\frac{(\boldsymbol{h}^T\boldsymbol{r})^2}{(\boldsymbol{h}^T\boldsymbol{A}\boldsymbol{h})^2}\boldsymbol{h}^T\boldsymbol{A}\boldsymbol{h} \\
&= F(\boldsymbol{x}) - \frac{1}{2}\frac{(\boldsymbol{h}^T\boldsymbol{r})^2}{\boldsymbol{h}^T\boldsymbol{A}\boldsymbol{h}} \\
&\leqq F(\boldsymbol{x}).
\end{aligned}$$

❀

Wir werden sehen, dass die Richtungen $\boldsymbol{h}^{(i)}$ so wählbar sind, dass nach höchstens n Schritten die Lösung des Minimierungsproblems 8.14 berechnet wird. Dazu definieren wir den Vektorraum

$$H_{i+1} = \text{span}(\boldsymbol{h}^{(0)}, \boldsymbol{h}^{(1)}, \ldots, \boldsymbol{h}^{(i)}).$$

Das ist der Vektorraum, der von den Richtungen $\boldsymbol{h}^{(0)}, \boldsymbol{h}^{(1)}, \ldots, \boldsymbol{h}^{(i)}$ aufgespannt wird. Nach dem Algorithmus ergibt sich $\boldsymbol{x}^{(i+1)}$ aus $\boldsymbol{x}^{(0)}$ nach

$$\boldsymbol{x}^{(i+1)} = \boldsymbol{x}^{(0)} + \alpha_0 \boldsymbol{h}^{(0)} + \alpha_1 \boldsymbol{h}^{(1)} + \cdots + \alpha_i \boldsymbol{h}^{(i)}.$$

Es gilt also $\boldsymbol{x}^{(i+1)} - \boldsymbol{x}^{(0)} \in H_{i+1}$.
Wir wollen die Richtungen $\boldsymbol{h}^{(i)}$ und damit die Räume H_{i+1} so wählen, dass einerseits die Minimierung des Funktionals $F(\boldsymbol{z})$ auf dem gesamten $\mathbb{R}^n$ auf eine schrittweise Minimierung des Funktionals auf den Teilräumen H_{i+1} zurückgeführt ist, und dass andererseits die Minimierung auf den Teilräumen möglichst einfach wird.
Wir nehmen an, wir hätten einen Teilraum H_i konstruiert und ein $\boldsymbol{x}^{(i)} \in \boldsymbol{x}^{(0)} + H_i$ bestimmt, das $F(\boldsymbol{z})$ auf $\boldsymbol{x}^{(0)} + H_i$ minimiert. $\boldsymbol{h}^{(i)}$ sei eine weitere Richtung, durch die H_{i+1} festgelegt ist. Jeder Vektor $\boldsymbol{y} \in \boldsymbol{x}^{(0)} + H_{i+1}$ ist dann in der Form

$$\boldsymbol{y} = \boldsymbol{x} + \alpha \boldsymbol{h}^{(i)}$$

mit $\boldsymbol{x} \in \boldsymbol{x}^{(0)} + H_i$ und $\alpha \in \mathbb{R}$ zerlegbar. Wenden wir das Funktional F auf $\boldsymbol{y}$ an, so erhalten wir

$$\begin{aligned}
F(\boldsymbol{y}) &= \frac{1}{2}\left[\boldsymbol{A}\left(\boldsymbol{x} + \alpha \boldsymbol{h}^{(i)}\right) - \boldsymbol{b}\right]^T \boldsymbol{A}^{-1}\left[\boldsymbol{A}\left(\boldsymbol{x} + \alpha \boldsymbol{h}^{(i)}\right) - \boldsymbol{b}\right] \\
&= F(\boldsymbol{x}) + \frac{1}{2}\alpha {\boldsymbol{h}^{(i)}}^T (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) + \frac{1}{2}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^T \boldsymbol{h}^{(i)} + \frac{1}{2}\alpha^2 {\boldsymbol{h}^{(i)}}^T \boldsymbol{A}\boldsymbol{h}^{(i)} \\
&= F(\boldsymbol{x}) + \alpha {\boldsymbol{h}^{(i)}}^T (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) + \frac{1}{2}\alpha^2 {\boldsymbol{h}^{(i)}}^T \boldsymbol{A}\boldsymbol{h}^{(i)} \\
&= F(\boldsymbol{x}) + \alpha {\boldsymbol{h}^{(i)}}^T \boldsymbol{A}\left(\boldsymbol{x} - \boldsymbol{x}^{(0)}\right) + \alpha {\boldsymbol{h}^{(i)}}^T \boldsymbol{A}\boldsymbol{x}^{(0)} - \alpha {\boldsymbol{h}^{(i)}}^T \boldsymbol{A}\boldsymbol{b} + \frac{1}{2}\alpha^2 {\boldsymbol{h}^{(i)}}^T \boldsymbol{A}\boldsymbol{h}^{(i)}.
\end{aligned}$$

Der zweite Summand in der letzten Gleichung hängt sowohl von $\boldsymbol{x}$ als auch von $\boldsymbol{h}^{(i)}$ ab. Können wir die Richtung $\boldsymbol{h}^{(i)}$ so wählen, dass dieser Term für alle $\boldsymbol{x} \in \boldsymbol{x}^{(0)} + H_i$ verschwindet, so zerfällt das Minimierungsproblem von F auf $\boldsymbol{x}^{(0)} + H_{i+1}$ in zwei getrennte Minimierungsprobleme

$$\min_{\boldsymbol{x} \in \boldsymbol{x}^{(0)} + H_i} F(\boldsymbol{x})$$

und

$$\min_{\alpha \in \mathbb{R}} \left\{ \frac{1}{2}\alpha^2 {\boldsymbol{h}^{(i)}}^T \boldsymbol{A}\boldsymbol{h}^{(i)} - \alpha {\boldsymbol{h}^{(i)}}^T \boldsymbol{A}\boldsymbol{b} + \alpha {\boldsymbol{h}^{(i)}}^T \boldsymbol{A}\boldsymbol{x}^{(0)} \right\}.$$

Die Forderung

$$\boldsymbol{h}^{(i)^T}\boldsymbol{A}\left(\boldsymbol{x}-\boldsymbol{x}^{(0)}\right)=0$$

für alle $\boldsymbol{x}\in\boldsymbol{x}^{(0)}+H_i$ ist genau dann erfüllt, wenn

$$\boldsymbol{h}^{(i)^T}\boldsymbol{A}\boldsymbol{h}^{(j)}=0 \quad \text{für} \quad j=0,1,\ldots,i-1$$

gilt. Zwei Vektoren $\boldsymbol{x}\in\mathbb{R}^n$ und $\boldsymbol{y}\in\mathbb{R}^n$, für die $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{y}=0$ bezüglich einer symmetrischen, positiv definiten Matrix $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ gilt, heißen **$\boldsymbol{A}$-konjugiert** oder **$\boldsymbol{A}$-orthogonal**. Es gilt der folgende

8.62. Satz: *Sind die Vektoren $\boldsymbol{h}^{(0)},\ldots,\boldsymbol{h}^{(i)}$ bezüglich der positiv definiten Matrix $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ paarweise konjugiert, so sind sie linear unabhängig.*

Beweis: Es sei $\boldsymbol{h}=\alpha_0\boldsymbol{h}^{(0)}+\cdots\alpha_i\boldsymbol{h}^{(i)}=\boldsymbol{o}$. Wir müssen zeigen, dass dann $\alpha_0=\alpha_1=\cdots=\alpha_i=0$ folgt. Es gilt

$$0=\boldsymbol{h}^T\boldsymbol{A}\boldsymbol{h}=\left(\sum_{j=1}^{i}\alpha_j\boldsymbol{h}^{(j)}\right)^T\boldsymbol{A}\left(\sum_{j=1}^{i}\alpha_j\boldsymbol{h}^{(j)}\right)=\sum_{j=1}^{i}\sum_{k=1}^{i}\alpha_j\alpha_k\boldsymbol{h}^{(j)^T}\boldsymbol{A}\boldsymbol{h}^{(k)}.$$

Wegen der $\boldsymbol{A}$-Konjugiertheit der Vektoren entfallen alle Summanden mit $j\neq k$. Wir erhalten

$$0=\sum_{j=1}^{i}\alpha_j^2\boldsymbol{h}^{(j)^T}\boldsymbol{A}\boldsymbol{h}^{(j)}.$$

Aus der positiven Definitheit von $\boldsymbol{A}$ folgt $\boldsymbol{h}^{(j)^T}\boldsymbol{A}\boldsymbol{h}^{(j)}>0$ und daraus sofort

$$\alpha_0=\alpha_1=\cdots=\alpha_i=0.$$

❀

Betrachten wir nun wieder die Minimierung des Funktionals auf $\boldsymbol{x}^{(0)}+H_{i+1}$. Wurde die Richtung $h^{(i)}$ so gewählt, dass $\boldsymbol{h}^{(j)^T}\boldsymbol{A}\boldsymbol{h}(i)=0$ für $j=0,\ldots,i-1$ gilt, so folgt

$$\begin{aligned}
\min_{\boldsymbol{y}\in\boldsymbol{x}^{(0)}+H_{i+1}}F(\boldsymbol{y}) &= \min_{\boldsymbol{x}\in\boldsymbol{x}^{(0)}+H_i,\alpha\in\mathbb{R}}F(\boldsymbol{x}+\alpha\boldsymbol{h}^{(i)})\\
&= \min_{\alpha\in\mathbb{R}}\min_{\boldsymbol{x}\in\boldsymbol{x}^{(0)}+H_i}F(\boldsymbol{x}+\alpha\boldsymbol{h}^{(i)})\\
&= \min_{\alpha\in\mathbb{R}}F(\boldsymbol{x}^{(i)}+\alpha\boldsymbol{h}^{(i)})\\
&= F(\boldsymbol{x}^{(i)}+\alpha_i\boldsymbol{h}^{(i)})\\
&= F(\boldsymbol{x}^{(i)})-\frac{1}{2}\frac{\left(\boldsymbol{r}^{(i)^T}\boldsymbol{h}^{(i)}\right)^2}{\boldsymbol{h}^{(i)^T}\boldsymbol{A}\boldsymbol{h}^{(i)}}
\end{aligned}$$

mit

$$\alpha_i = \frac{\boldsymbol{r}^{(i)^T}\boldsymbol{h}^{(i)}}{\boldsymbol{h}^{(i)^T}\boldsymbol{A}\boldsymbol{h}^{(i)}}.$$

Damit gilt

8.63. Satz: *Es sei $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ eine symmetrische, positiv definite Matrix.*

$$\boldsymbol{h}^{(0)}, \boldsymbol{h}^{(1)}, \ldots, \boldsymbol{h}^{(n-1)}, \quad \boldsymbol{h}^{(i)} \neq \boldsymbol{o}, \quad i = 0, 1, \ldots, n-1$$

seien n paarweise $\boldsymbol{A}$-konjugierte Richtungen.
Für einen beliebigen Startvektor $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ seien die Vektoren $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}$ rekursiv durch

$$\boldsymbol{x}^{(i+1)} = \boldsymbol{x}^{(i)} + \alpha_i \boldsymbol{h}^{(i)}$$

mit

$$\alpha_i = \frac{\boldsymbol{r}^{(i)^T}\boldsymbol{h}^{(i)}}{\boldsymbol{h}^{(i)^T}\boldsymbol{A}\boldsymbol{h}^{(i)}} = \frac{\left(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(i)}\right)^T \boldsymbol{h}^{(i)}}{\boldsymbol{h}^{(i)^T}\boldsymbol{A}\boldsymbol{h}^{(i)}}$$

definiert. Dann gilt

1. $$\boldsymbol{x}^{(i)} = \arg \min_{\boldsymbol{x} \in \boldsymbol{x}^{(0)} + H_i} F(\boldsymbol{x}) \quad \textit{für} \quad i = 1, \ldots, n.$$

2. $$\boldsymbol{A}\boldsymbol{x}^{(n)} = \boldsymbol{b}.$$

3. $$\boldsymbol{h}^{(j)^T}\boldsymbol{r}^{(k)} = 0 \quad \textit{für alle} \quad 0 \leqq j < k \leqq n.$$

Beweis:

1. Diese Aussage folgt sofort aus der Konstruktion der Vektoren $\boldsymbol{x}^{(i)}$.
2. Mit Satz 8.62 folgt $H_n = \mathbb{R}^n$. Dann gilt

   $$\boldsymbol{x}^{(n)} = \arg\min \boldsymbol{x} \in \mathbb{R}^n F(\boldsymbol{x})$$

   und $\boldsymbol{A}\boldsymbol{x}^{(n)} = \boldsymbol{b}$.

3. Wir beweisen die Aussage mittels Induktion über k.
   - Für $k = 1$ gilt
   $$\begin{aligned}
   \boldsymbol{h}^{(0)^T}\boldsymbol{r}^{(1)} &= \boldsymbol{h}^{(0)^T}\left(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}^{(1)}\right)\\
   &= \boldsymbol{h}^{(0)^T}\left(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}^{(0)}-\alpha_0\boldsymbol{A}\boldsymbol{h}^{(0)}\right)\\
   &= \boldsymbol{h}^{(0)^T}\boldsymbol{r}^{(0)}-\alpha_0\boldsymbol{h}^{(0)^T}\boldsymbol{A}\boldsymbol{h}^{(0)}\\
   &= \boldsymbol{h}^{(0)^T}\boldsymbol{r}^{(0)}-\frac{\boldsymbol{h}^{(0)^T}\boldsymbol{r}^{(0)}}{\boldsymbol{h}^{(0)^T}\boldsymbol{A}\boldsymbol{h}^{(0)}}\boldsymbol{h}^{(0)^T}\boldsymbol{A}\boldsymbol{h}^{(0)}\\
   &= 0.
   \end{aligned}$$
   - Es sei für festes k $\boldsymbol{h}^{(j)^T}\boldsymbol{r}^{(k)} = 0$ für $j = 0,\ldots,k-1$. Wir zeigen, dass dann $\boldsymbol{h}^{(j)^T}\boldsymbol{r}^{(k+1)} = 0$ für $j = 0,\ldots,k$ gilt.
   Wegen
   $$\boldsymbol{r}^{(k+1)} = \boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}^{(k+1)} = \boldsymbol{b}-\boldsymbol{A}\left(\boldsymbol{x}^{(k)}+\alpha_k\boldsymbol{h}^{(k)}\right) = \boldsymbol{r}^{(k)}-\alpha_k\boldsymbol{A}\boldsymbol{h}^{(k)}$$
   folgt
   $$\boldsymbol{h}^{(j)^T}\boldsymbol{r}^{(k+1)} = \boldsymbol{h}^{(j)^T}\boldsymbol{r}^{(k)}-\alpha_k\boldsymbol{h}^{(j)^T}\boldsymbol{A}\boldsymbol{h}^{(k)}.$$
   Betrachten wir den Fall $j \leqq k-1$. Wegen der Induktionsvoraussetzung gilt hier $\boldsymbol{h}^{(j)^T}\boldsymbol{r}^{(k)} = 0$ und wegen der paarweisen $\boldsymbol{A}$-Konjugiertheit der Richtungen $\boldsymbol{h}^{(j)^T}\boldsymbol{A}\boldsymbol{h}^{(k)} = 0$. Damit ist die Aussage bis $j = k-1$ bewiesen.
   Für $j = k$ ergibt sich
   $$\begin{aligned}
   \boldsymbol{h}^{(k)^T}\boldsymbol{r}^{(k+1)} &= \boldsymbol{h}^{(k)^T}\boldsymbol{r}^{(k)}-\alpha_k\boldsymbol{h}^{(k)^T}\boldsymbol{A}\boldsymbol{h}^{(k)}\\
   &= \boldsymbol{h}^{(k)^T}\boldsymbol{r}^{(k)}-\frac{\boldsymbol{h}^{(k)^T}\boldsymbol{r}^{(k)}}{\boldsymbol{h}^{(k)^T}\boldsymbol{A}\boldsymbol{h}^{(k)}}\boldsymbol{h}^{(k)^T}\boldsymbol{A}\boldsymbol{h}^{(k)}\\
   &= 0.
   \end{aligned}$$
   ❀

Ist man nun in der Lage, effizient $\boldsymbol{A}$-konjugierte Richtungen $\boldsymbol{h}^{(0)},\boldsymbol{h}^{(1)},\ldots,\boldsymbol{h}^{(n-1)}$ zu konstruieren, so erhält man mit dem Verfahren aus Satz 8.63 nach höchstens n Schritten die Lösung des Gleichungssystems $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. Satz 8.63 zeigt auch, dass das Residuum $\boldsymbol{r}^{(i)} = \boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}^{(i)}$ orthogonal zu allen $\boldsymbol{h} \in H_i$ ist. Es gilt daher entweder $\boldsymbol{r}^{(i)} = \boldsymbol{o}$ oder $\boldsymbol{r}^{(i)} \notin H_i$. Im ersten Falle ist $\boldsymbol{x}^{(i)}$ Lösung des Gleichungssystems. Im zweiten Falle bietet es sich an, für die neue Richtung $\boldsymbol{h}^{(i)}$ den Ansatz

$$\boldsymbol{h}^{(i)} = \boldsymbol{r}^{(i)}+\sum_{j=0}^{i-1}\beta_{ij}\boldsymbol{h}^{(j)} \tag{8.15}$$

zu verwenden. Dann gilt für die Residuen

$$\boldsymbol{r}^{(i)} = \boldsymbol{h}^{(i)} - \sum_{j=0}^{i-1} \beta_{ij} \boldsymbol{h}^{(j)} \in H_{i+1} \quad \text{für} \quad i = 0, 1, \ldots$$

Nach der dritten Aussage aus Satz 8.63 sind dann alle Residuen paarweise orthogonal. Außerdem folgt aus

$$\boldsymbol{x}^{(i+1)} = \boldsymbol{x}^{(i)} + \alpha_i \boldsymbol{h}^{(i)}$$

für die Residuen die Rekursionsformel

$$\boldsymbol{r}^{(i+1)} = \boldsymbol{r}^{(i)} - \alpha_i \boldsymbol{A}\boldsymbol{h}^{(i)}. \tag{8.16}$$

Dabei läßt sich für

$$\alpha_i = \frac{{\boldsymbol{r}^{(i)}}^T \boldsymbol{h}^{(i)}}{{\boldsymbol{h}^{(i)}}^T \boldsymbol{A}\boldsymbol{h}^{(i)}}$$

noch eine andere Darstellung finden. Multipliziert man Gleichung 8.15 von links mit ${\boldsymbol{r}^{(i)}}^T$, so erhält man

$${\boldsymbol{r}^{(i)}}^T \boldsymbol{h}^{(i)} = {\boldsymbol{r}^{(i)}}^T \boldsymbol{r}^{(i)}$$

und damit

$$\alpha_i = \frac{{\boldsymbol{r}^{(i)}}^T \boldsymbol{r}^{(i)}}{{\boldsymbol{h}^{(i)}}^T \boldsymbol{A}\boldsymbol{h}^{(i)}}.$$

Es sind zwei Fälle möglich:

1. $\alpha_i = 0$
   Es folgt $\boldsymbol{r}^{(i)} = \boldsymbol{o}$. Dann ist $\boldsymbol{x}^{(i)}$ Lösung des Gleichungssystems $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$.

2. $\alpha_i \neq 0$
   Es folgt aus Gleichung 8.16

$$-\boldsymbol{A}\boldsymbol{h}^{(i)} = \frac{\boldsymbol{r}^{(i+1)} - \boldsymbol{r}^{(i)}}{\alpha_i}.$$

Nun können wir darangehen, die β_{ij} so zu bestimmen, dass $\boldsymbol{h}^{(i)}$ zu allen $\boldsymbol{h}^{(k)}$, $k = 0,\ldots,i-1$, $\boldsymbol{A}$-konjugiert ist. Wir erhalten

$$\begin{aligned} 0 &= \boldsymbol{h}^{(k)^T}\boldsymbol{A}\boldsymbol{h}^{(i)} = \boldsymbol{h}^{(k)^T}\boldsymbol{A}\boldsymbol{r}^{(i)} + \sum_{j=0}^{i-1}\beta_{ij}\boldsymbol{h}^{(k)^T}\boldsymbol{A}\boldsymbol{h}^{(j)} \\ &= \boldsymbol{h}^{(k)^T}\boldsymbol{A}\boldsymbol{r}^{(i)} + \beta_{ik}\boldsymbol{h}^{(k)^T}\boldsymbol{A}\boldsymbol{h}^{(k)}. \end{aligned}$$

Daraus folgt

$$\beta_{ik} = -\frac{\boldsymbol{h}^{(k)^T}\boldsymbol{A}\boldsymbol{r}^{(i)}}{\boldsymbol{h}^{(k)^T}\boldsymbol{A}\boldsymbol{h}^{(k)}} = -\frac{\left(\boldsymbol{A}\boldsymbol{h}^{(k)}\right)^T\boldsymbol{r}^{(i)}}{\boldsymbol{h}^{(k)^T}\boldsymbol{A}\boldsymbol{h}^{(k)}}.$$

Im Falle $\boldsymbol{r}^{(k)} \neq \boldsymbol{o}$ ($\alpha_k \neq 0$) ergibt sich

$$\beta_{ik} = \frac{\left(\boldsymbol{r}^{(k+1)} - \boldsymbol{r}^{(k)}\right)^T}{\alpha_k}\frac{\boldsymbol{r}^{(i)}}{\boldsymbol{h}^{(k)^T}\boldsymbol{A}\boldsymbol{h}^{(k)}} \quad \text{für} \quad k = 0,\ldots,i-1.$$

Für $k \leqq i-2$ folgt dann wegen der paarweisen Orthogonalität der Residuen $\beta_{ik} = 0$. Für $k = i-1$ erhalten wir

$$\begin{aligned} \beta_{i,i-1} &= \frac{1}{\alpha_{i-1}}\frac{\boldsymbol{r}^{(i)^T}\boldsymbol{r}^{(i)}}{\boldsymbol{h}^{(i-1)^T}\boldsymbol{A}\boldsymbol{h}^{(i-1)}} = \frac{\boldsymbol{h}^{(i-1)^T}\boldsymbol{A}\boldsymbol{h}^{(i-1)}}{\boldsymbol{r}^{(i-1)^T}\boldsymbol{r}^{(i-1)}}\frac{\boldsymbol{r}^{(i)^T}\boldsymbol{r}^{(i)}}{\boldsymbol{h}^{(i-1)^T}\boldsymbol{A}\boldsymbol{h}^{(i-1)}} \\ &= \frac{\boldsymbol{r}^{(i)^T}\boldsymbol{r}^{(i)}}{\boldsymbol{r}^{(i-1)^T}\boldsymbol{r}^{(i-1)}}. \end{aligned}$$

Damit haben wir das cg-Verfahren von HESTENES und STIEFEL hergeleitet. [5]

8.64. cg-Verfahren von HESTENES **und** STIEFEL**:**
Zu lösen ist das lineare Gleichungssystem $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ mit einer symmetrischen, positiv definiten Matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$.

S0 *Wähle einen Startvektor $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$.*

S1 *Berechne $\boldsymbol{r}^{(0)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(0)}$.*
Setze $\boldsymbol{h}^{(0)} = \boldsymbol{r}^{(0)}$ und $i = 0$.

S2 *Falls $\boldsymbol{h}^{(i)} = \boldsymbol{o}$ STOPP. $\boldsymbol{x}^{(i)}$ ist die gesuchte Lösung.*

[5] Der Name leitet sich aus dem Englischen ab. cg steht für **c**onjugate **g**radients.

S3 *Berechne*

$$\begin{aligned} \alpha_i &= \frac{\boldsymbol{r}^{(i)^T}\boldsymbol{r}^{(i)}}{\boldsymbol{h}^{(i)^T}\boldsymbol{A}\boldsymbol{h}^{(i)}}, \\ \boldsymbol{x}^{(i+1)} &= \boldsymbol{x}^{(i)} + \alpha_i \boldsymbol{h}^{(i)}, \\ \boldsymbol{r}^{(i+1)} &= \boldsymbol{r}^{(i)} - \alpha_i \boldsymbol{A}\boldsymbol{h}^{(i)}, \\ \beta_i &= \frac{\boldsymbol{r}^{(i+1)^T}\boldsymbol{r}^{(i+1)}}{\boldsymbol{r}^{(i)^T}\boldsymbol{r}^{(i)}}, \\ \boldsymbol{h}^{(i+1)} &= \boldsymbol{r}^{(i+1)} + \beta_i \boldsymbol{h}^{(i)}. \end{aligned}$$

S4 *Setze* $i = i + 1$ *und gehe zu Schritt* **S2**.

Bemerkungen: (i) Im Algorithmus wird die Matrix (wie bei den anderen Iterationsverfahren) nicht verändert. Man benötigt wieder nur ein Unterprogramm, das zu einem gegebenen Vektor $\boldsymbol{x}$ den Vektor $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ berechnet.
(ii) Nach Satz 8.63 bricht das Verfahren bei exakter Rechnung nach höchstens n Schritten mit der Lösung des linearen Gleichungssystems ab.
(iii) Pro Schritt sind im Algorithmus ein Produkt Matrix$\times$Vektor, zwei Skalarprodukte und drei Operationen Vektor+Skalar$\times$Vektor zu berechnen. Für eine vollbesetzte Matrix beträgt der Aufwand maximal rund n^2 Additionen/Multiplikationen. Er wird durch die Operation Matrix$\times$Vektor bestimmt. Für einen schwach besetzte Matrix liegt der Aufwand für die Operation Matrix$\times$Vektor nur in der Größenordnung von n. Damit beträgt der Aufwand pro Schritt des Algorithmus nur Kn Additionen/Multiplikationen, wobei K wesentlich kleiner als n ist.
(iv) Der Gesamtaufwand beträgt für eine vollbesetzte Matrix rund n^3 Operationen, ist also größer als bei den bisher behandelten direkten Verfahren. Für schwach besetzte Matrizen erhalten wir jedoch einen Gesamtaufwand in der Größenordnung von n^2. Hier ist das cg-Verfahren effektiv.
Satz 8.63 zeigt, dass für das Verhalten des cg-Verfahrens gerade die paarweise $\boldsymbol{A}$-Konjugiertheit der Richtungen $\boldsymbol{h}^{(i)}$ und damit die paarweise Orthogonalität der Residuen $\boldsymbol{r}^{(i)}$ wesentlich ist. In der Praxis werden diese Beziehungen durch den Einfluß von Rundungsfehlern oft verletzt sein. Man wird daher erwarten, dass man mehr als n Iterationen benötigt, um eine brauchbare Lösung zu berechnen. Praktische Erfahrungen zeigen aber, dass man gerade für großes n oft mit weniger Iterationsschritten auskommt. So wurde zum Beispiel 1971 von REID mit dem cg-Verfahren ein Gleichungssystem mit 4080 Gleichungen in 40 Schritten gelöst. Eine Begründung für dieses Verhalten liefert der folgende

8.65. Satz: *Es sei* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *eine symmetrische, positiv definite Matrix mit genau* k *paarweise verschiedenen Eigenwerten* $\lambda_1, \ldots, \lambda_k$. *Dann liefert das cg-Verfahren bei*

beliebigem Startvektor $\boldsymbol{x}^{(0)}$ nach höchstens k Schritten die Lösung des Gleichungssystems $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$.

Beweis: Zu den k Eigenwerten $\lambda_1, \ldots, \lambda_k$ gehören k Eigenräume $\boldsymbol{\mathcal{U}}_1, \ldots, \boldsymbol{\mathcal{U}}_k$. Da die Matrix $\boldsymbol{A}$ symmetrisch ist, gilt

$$\dim(\boldsymbol{\mathcal{U}}_1) + \cdots + \dim(\boldsymbol{\mathcal{U}}_k) = n$$

und

$$\boldsymbol{\mathcal{U}}_i \perp \boldsymbol{\mathcal{U}}_j \quad \text{für} \quad i \neq j.$$

Damit lässt sich das Residuum $\boldsymbol{r}^{(0)}$ in folgender Weise zerlegen:

$$\boldsymbol{r}^{(0)} = p_1^{(0)}\boldsymbol{u}_1 + p_2^{(0)}\boldsymbol{u}_2 + \cdots + p_k^{(0)}\boldsymbol{u}_k$$

mit $\boldsymbol{u}_i \in \boldsymbol{\mathcal{U}}_i$ für $i = 1, \ldots, k$. In Matrixschreibweise ergibt sich

$$\boldsymbol{r}^{(0)} = \boldsymbol{U}\boldsymbol{p}^{(0)}$$

mit der spaltenorthogonalen Matrix $\boldsymbol{U} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k) \in \mathbb{R}^{n \times k}$ und $\boldsymbol{p}^{(0)} \in \mathbb{R}^k$. Für die Matrix $\boldsymbol{U}$ gilt

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{U} &= \boldsymbol{A}(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k) \\
&= (\boldsymbol{A}\boldsymbol{u}_1, \boldsymbol{A}\boldsymbol{u}_2, \ldots, \boldsymbol{A}\boldsymbol{u}_k) \\
&= (\lambda_1\boldsymbol{u}_1, \lambda_2\boldsymbol{u}_2, \ldots, \lambda_k\boldsymbol{u}_k) \\
&= \boldsymbol{U}\boldsymbol{\Lambda}
\end{aligned}$$

mit $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_k) \in \mathbb{R}^{k \times k}$.
Mittels vollständiger Induktion zeigen wir nun, dass alle Vektoren $\boldsymbol{r}^{(i)}$ und $\boldsymbol{h}^{(i)}$, die vom cg-Verfahren erzeugt werden, Darstellungen der Form

$$\boldsymbol{r}^{(i)} = \boldsymbol{U}\boldsymbol{p}^{(i)} \quad \text{und} \quad \boldsymbol{h}^{(i)} = \boldsymbol{U}\boldsymbol{g}^{(i)}$$

mit $\boldsymbol{p}^{(i)} \in \mathbb{R}^k$ und $\boldsymbol{g}^{(i)} \in \mathbb{R}^k$ besitzen.

- Induktionsanfang: Für $i = 0$ gilt $\boldsymbol{h}^{(0)} = \boldsymbol{r}^{(0)} = \boldsymbol{U}\boldsymbol{p}^{(0)}$.
- Induktionsschritt: Es sei

  $$\boldsymbol{r}^{(i)} = \boldsymbol{U}\boldsymbol{p}^{(i)} \quad \text{und} \quad \boldsymbol{h}^{(i)} = \boldsymbol{U}\boldsymbol{g}^{(i)}$$

mit $\boldsymbol{p}^{(i)} \in \mathbb{R}^k$ und $\boldsymbol{g}^{(i)} \in \mathbb{R}^k$. Ein Schritt des cg-Verfahrens liefert

$$\begin{aligned} \boldsymbol{r}^{(i+1)} &= \boldsymbol{r}^{(i)} + \alpha_i \boldsymbol{A}\boldsymbol{h}^{(i)} = \boldsymbol{U}\boldsymbol{p}^{(i)} + \alpha_i \boldsymbol{A}\boldsymbol{U}\boldsymbol{g}^{(i)} \\ &= \boldsymbol{U}\boldsymbol{p}^{(i)} + \alpha_i \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{g}^{(i)} = \boldsymbol{U}\left(\boldsymbol{p}^{(i)} + \alpha_i \boldsymbol{\Lambda}\boldsymbol{g}^{(i)}\right) = \boldsymbol{U}\boldsymbol{p}^{(i+1)} \end{aligned}$$

mit $\boldsymbol{p}^{(i+1)} = \boldsymbol{p}^{(i)} + \alpha_i \boldsymbol{\Lambda}\boldsymbol{g}^{(i)} \in \mathbb{R}^k$ und

$$\boldsymbol{h}^{(i+1)} = \boldsymbol{r}^{(i+1)} + \beta_i \boldsymbol{h}^{(i)} = \boldsymbol{U}\boldsymbol{p}^{(i)} + \beta_i \boldsymbol{U}\boldsymbol{g}^{(i)} = \boldsymbol{U}\left(\boldsymbol{p}^{(i+1)} + \beta_i \boldsymbol{g}^{(i)}\right) = \boldsymbol{U}\boldsymbol{g}^{(i+1)}$$

mit $\boldsymbol{g}^{(i+1)} = \boldsymbol{p}^{(i+1)} + \beta_i \boldsymbol{g}^{(i)} \in \mathbb{R}^k$.
Nun sind die $\boldsymbol{r}^{(i)}$ paarweise orthogonal, es gilt daher

$$\boldsymbol{r}^{(i)^T} \boldsymbol{r}^{(j)} = 0 \quad \text{für} \quad i \neq j.$$

Daraus folgt

$$0 = \boldsymbol{r}^{(i)^T} \boldsymbol{r}^{(j)} = \left(\boldsymbol{U}\boldsymbol{p}^{(i)}\right)^T \left(\boldsymbol{U}\boldsymbol{p}^{(j)}\right) = \boldsymbol{p}^{(i)^T} \boldsymbol{U}^T \boldsymbol{U}\boldsymbol{p}^{(j)} = \boldsymbol{p}^{(i)^T} \boldsymbol{p}^{(j)}$$

für $i \neq j$. Die Vektoren $\boldsymbol{p}^{(i)}$ sind daher ebenfalls paarweise orthogonal. Wegen $\boldsymbol{p}^{(i)} \in \mathbb{R}^k$ gibt es nur k vom Nullvektor verschiedene Vektoren $\boldsymbol{p}^{(i)}$. Damit ist spätestens $\boldsymbol{p}^{(k)} = \boldsymbol{o}$ und damit $\boldsymbol{r}^{(k)} = \boldsymbol{o}$.

❀

Das modifizierte cg-Verfahren

In der ursprünglichen Form sind mit dem cg-Verfahren nur Gleichungssysteme lösbar, deren Koeffizientenmatrix symmetrisch und positiv definit ist. Das Verfahren läßt sich aber so modifizieren, dass es auf beliebige Gleichungssysteme anwendbar ist. Dazu betrachtet man das zur Lösung des linearen Gleichungssystems $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ äquivalente Minimierungsproblem

$$x = \arg\min_{\boldsymbol{z} \in \mathbb{R}^n} F(\boldsymbol{z}),$$

wobei das Funktional F hier die Darstellung

$$F(\boldsymbol{z}) = \frac{1}{2} (\boldsymbol{A}\boldsymbol{z} - \boldsymbol{b})^T \left(\boldsymbol{A}\boldsymbol{A}^T\right)^{-1} (\boldsymbol{A}\boldsymbol{z} - \boldsymbol{b})$$

besitzt. Die Matrix $\boldsymbol{A}\boldsymbol{A}^T$ ist genau dann positiv definit wenn $\boldsymbol{A}$ regulär ist. Damit gilt

$$F(\boldsymbol{z}) = 0 \Longleftrightarrow \boldsymbol{A}\boldsymbol{z} - \boldsymbol{b} = \boldsymbol{o}.$$

Eine analoge Vorgehensweise wie beim ursprünglichen cg-Verfahren liefert den folgenden Algorithmus.

8.66. Modifiziertes cg-Verfahren:
Lösen des lineare Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ mit einer regulären Matrix.

S0 *Wähle einen Startvektor $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$.*

S0 *Berechne $\boldsymbol{r}^{(0)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(0)}$ und $\boldsymbol{h}^{(0)} = \boldsymbol{A}^T\boldsymbol{r}^{(0)}$*
Setze $i = 0$.

S0 *Falls $\boldsymbol{h}^{(i)} = \boldsymbol{o}$ STOPP. $\boldsymbol{x}^{(i)}$ ist die gesuchte Lösung.*

S0 *Berechne*

$$\begin{aligned}
\alpha_i &= \frac{\boldsymbol{r}^{(i)^T}\boldsymbol{r}^{(i)}}{\boldsymbol{h}^{(i)^T}\boldsymbol{h}^{(i)}}, \\
\boldsymbol{x}^{(i+1)} &= \boldsymbol{x}^{(i)} + \alpha_i \boldsymbol{h}^{(i)}, \\
\boldsymbol{r}^{(i+1)} &= \boldsymbol{r}^{(i)} - \alpha_i \boldsymbol{A}\boldsymbol{h}^{(i)}, \\
\beta_i &= \frac{\boldsymbol{r}^{(i+1)^T}\boldsymbol{r}^{(i+1)}}{\boldsymbol{r}^{(i)^T}\boldsymbol{r}^{(i)}}, \\
\boldsymbol{h}^{(i+1)} &= \boldsymbol{A}^T\boldsymbol{r}^{(i+1)} + \beta_i \boldsymbol{h}^{(i)}.
\end{aligned}$$

S0 *Setze $i = i + 1$ und gehe zu Schritt* **S2.**

Beim modifizierten cg-verfahren sind die Richtungen und Residuen paarweise orthogonal:

$$\boldsymbol{h}^{(i)^T}\boldsymbol{h}^{(j)} = 0 \quad \text{und} \quad \boldsymbol{r}^{(i)^T}\boldsymbol{r}^{(j)} = 0 \quad \text{für} \quad i \neq j.$$

Außerdem gilt

$$\boldsymbol{h}^{(i)^T}\boldsymbol{A}^{-1}\boldsymbol{r}^{(i)} = \boldsymbol{r}^{(i)^T}\boldsymbol{r}^{(i)}.$$

Diese Aussagen folgen wie die entsprechenden Aussagen beim ursprünglichen cg-Verfahren aus der Herleitung der Rekursionsformeln. Aus der paarweisen Orthogonalität der Richtungen und Residuen folgt wieder, dass das Verfahren nach höchstens n Schritten die Lösung des linearen Gleichungssystems liefert. Analog zum Satz 8.65 ist die Anzahl der benötigten Iterationen genauer abschätzbar.

8.67. Satz: *Gegeben sei ein lineares Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ mit einer regulären (n,n)-Matrix $\boldsymbol{A}$, die genau k paarweise verschiedene Singulärwerte $\sigma_1, \ldots, \sigma_k$ haben möge.*
Dann liefert das modifizierte cg-Verfahren zu jedem Startvektor $\boldsymbol{x}^{(0)}$ nach höchstens k Schritten die Lösung des Gleichungssystems.

Beweis: Es sei $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ die Singulärwertzerlegung der Matrix $\boldsymbol{A}$ mit

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1 & & & & & & \\ & \ddots & & & \boldsymbol{O} & & \\ & & \sigma_1 & & & & \\ & & & \ddots & & & \\ & & & & \sigma_k & & \\ & \boldsymbol{O} & & & & \ddots & \\ & & & & & & \sigma_k \end{pmatrix}.$$

Gemäß der Vielfachheit der Singulärwerte können die Matrizen $\boldsymbol{U}$ und $\boldsymbol{V}$ partitioniert werden:

$$\boldsymbol{U} = (\boldsymbol{U}_1, \ldots, \boldsymbol{U}_k) \quad \text{und} \quad \boldsymbol{V} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_k).$$

Das zum Startvektor $\boldsymbol{x}^{(0)}$ gehörende Residuum $\boldsymbol{r}^{(0)}$ ist in der Form

$$\boldsymbol{r}^{(0)} = p_1^{(0)}\boldsymbol{u}_1 + \cdots + p_k^{(0)}\boldsymbol{u}_k$$

mit

$$\boldsymbol{u}_i \in \boldsymbol{\mathcal{U}}_i = \text{span}(\boldsymbol{U}_i), \quad i = 1, \ldots, k$$

zerlegbar. Mit $\boldsymbol{p}^{(0)} = \left(p_1^{(0)}, \ldots, p_k^{(0)}\right)^T \in \mathbb{R}^k$ und $\bar{\boldsymbol{U}} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k) \in \mathbb{R}^{n\times k}$ gilt $\boldsymbol{r}^{(0)} = \bar{\boldsymbol{U}}\boldsymbol{p}^{(0)}$. Weiterhin gilt $\bar{\boldsymbol{U}}^T\bar{\boldsymbol{U}} = \boldsymbol{I}$ und für $i = 1, \ldots, k$

$$\boldsymbol{A}^T\boldsymbol{u}_i = \sigma_i\boldsymbol{v}_i$$

mit $\boldsymbol{v}_i \in \boldsymbol{\mathcal{V}}_i = \text{span}\,(\boldsymbol{V}_i)$. Mit den Bezeichnungen $\bar{\boldsymbol{V}} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k) \in \mathbb{R}^{n\times k}$ und $\bar{\boldsymbol{\Sigma}} = \text{diag}(\sigma_1, \ldots, \sigma_k) \in \mathbb{R}^{k\times k}$ folgt

$$\boldsymbol{A}^T\bar{\boldsymbol{U}} = \bar{\boldsymbol{V}}\bar{\boldsymbol{\Sigma}} \quad \text{und} \quad \boldsymbol{A}\bar{\boldsymbol{V}} = \bar{\boldsymbol{U}}\bar{\boldsymbol{\Sigma}}.$$

Wir zeigen nun mittels Induktion, dass alle Richtungen und Residuen Darstellungen der Form

$$\boldsymbol{r}^{(i)} = \bar{\boldsymbol{U}}\boldsymbol{p}^{(i)} \quad \text{und} \quad \boldsymbol{h}^{(i)} = \bar{\boldsymbol{V}}\boldsymbol{g}^{(i)} \quad \text{mit} \quad \boldsymbol{p}^{(i)} \in \mathbb{R}^k \quad \text{und} \quad \boldsymbol{g}^{(i)} \in \mathbb{R}^k$$

für $i = 1, \ldots, k$ besitzen.

- Induktionsanfang: Wir haben schon gezeigt, dass $\boldsymbol{r}^{(0)} = \bar{\boldsymbol{U}}\boldsymbol{p}^{(0)}$ mit $\boldsymbol{p}^{(0)} \in \mathbb{R}^k$ gilt. Es folgt

  $$\boldsymbol{h}^{(0)} = \boldsymbol{A}^T\boldsymbol{r}^{(0)} = \boldsymbol{A}^T\bar{\boldsymbol{U}}\boldsymbol{p}^{(0)} = \bar{\boldsymbol{V}}\bar{\boldsymbol{\Sigma}}\boldsymbol{p}^{(0)} = \bar{\boldsymbol{V}}\boldsymbol{g}^{(0)}$$

  mit $\boldsymbol{g}^{(0)} = \bar{\boldsymbol{\Sigma}}\boldsymbol{p}^{(0)} \in \mathbb{R}^k$.

- Induktionsschritt: Wir setzen voraus, dass die Beziehungen

$$\boldsymbol{r}^{(i)} = \bar{\boldsymbol{U}}\boldsymbol{p}^{(i)} \quad \text{und} \quad \boldsymbol{h}^{(i)} = \bar{\boldsymbol{V}}\boldsymbol{g}^{(i)} \quad \text{mit} \quad \boldsymbol{p}^{(i)} \in \mathbb{R}^k \quad \text{und} \quad \boldsymbol{g}^{(i)} \in \mathbb{R}^k$$

gelten. Dann folgt

$$\begin{aligned} \boldsymbol{r}^{(i+1)} &= \boldsymbol{r}^{(i)} - \alpha_i \boldsymbol{A}\boldsymbol{h}^{(i)} = \bar{\boldsymbol{U}}\boldsymbol{p}^{(i)} - \alpha_i \boldsymbol{A}\bar{\boldsymbol{V}}\boldsymbol{g}^{(i)} \\ &= \bar{\boldsymbol{U}}\boldsymbol{p}^{(i)} - \alpha_i \bar{\boldsymbol{U}}\bar{\boldsymbol{\Sigma}}\boldsymbol{g}^{(i)} = \bar{\boldsymbol{U}}\left(\boldsymbol{p}^{(i)} - \alpha_i \bar{\boldsymbol{\Sigma}}\boldsymbol{g}^{(i)}\right) = \bar{\boldsymbol{U}}\boldsymbol{p}^{(i+1)} \end{aligned}$$

mit $\boldsymbol{p}^{(i+1)} = \boldsymbol{p}^{(i)} - \alpha_i \bar{\boldsymbol{\Sigma}}\boldsymbol{g}^{(i)} \in \mathbb{R}^k$ und

$$\begin{aligned} \boldsymbol{h}^{(i+1)} &= \boldsymbol{A}^T\boldsymbol{r}^{(i+1)} + \beta_i \boldsymbol{h}^{(i)} = \boldsymbol{A}^T\bar{\boldsymbol{U}}\boldsymbol{p}^{(i+1)} + \beta_i \bar{\boldsymbol{V}}\boldsymbol{g}^{(i)} \\ &= \bar{\boldsymbol{V}}\bar{\boldsymbol{\Sigma}}\boldsymbol{p}^{(i+1)} + \beta_i \bar{\boldsymbol{V}}\boldsymbol{g}^{(i)} = \bar{\boldsymbol{V}}\left(\bar{\boldsymbol{\Sigma}}\boldsymbol{p}^{(i+1)} + \beta_i \boldsymbol{g}^{(i)}\right) = \bar{\boldsymbol{V}}\boldsymbol{g}^{(i+1)} \end{aligned}$$

mit $\boldsymbol{g}^{(i+1)} = \bar{\boldsymbol{\Sigma}}\boldsymbol{p}^{(i+1)} + \beta_i \boldsymbol{g}^{(i)} \in \mathbb{R}^k$.

Aus der paarweisen Orthogonalität der Residuen folgt wieder die paarweise Orthogonalität der $\boldsymbol{p}^{(i)}$. Wegen $\boldsymbol{p}^{(i)} \in \mathbb{R}^k$ gibt es jedoch höchstens k vom Nullvektor verschiedene $\boldsymbol{p}^{(i)}$ die paarweise orthogonal sind. Damit ist spätestens $\boldsymbol{p}^{(k)} = \boldsymbol{o}$, $\boldsymbol{r}^{(k)} = \boldsymbol{o}$ und $\boldsymbol{A}\boldsymbol{x}^{(k)} = \boldsymbol{o}$. ❀

Vorkonditionierung der cg-Verfahren

Wir hatten schon bemerkt, dass für die Konvergenzaussagen des cg-Verfahrens die $\boldsymbol{A}$-Konjugiertheit der Richtungen wesentlich ist. Bei praktischen Rechnungen wird diese durch den Einfluß von Rundungsfehlern gestört. Andererseits wissen wir aus Satz 8.14, dass der Einfluß von Störungen um so geringer ist, je kleiner die Kondition der Matrix ist. Man kann nun versuchen, das Gleichungssystem $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ in ein solches System zu transformieren, bei dem die Koeffizientenmatrix besser konditioniert ist. Diese Vorgehen nennt man **Vorkonditionierung**. Es sei dazu $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{A} + \boldsymbol{\delta A}$ eine näherungsweise CHOLESKY-Zerlegung der Matrix $\boldsymbol{A}$. Wir wenden das cg-Verfahren auf das zu $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ äquivalente System

$$\left(\boldsymbol{L}^{-1}\boldsymbol{A}\left(\boldsymbol{L}^{-1}\right)^T\right)\boldsymbol{L}^T\boldsymbol{x} = \boldsymbol{L}^{-1}\boldsymbol{b}$$

an. Die Matrix $\bar{\boldsymbol{A}} = \boldsymbol{L}^{-1}\boldsymbol{A}\left(\boldsymbol{L}^{-1}\right)^T = \boldsymbol{I} - \boldsymbol{L}^{-1}\boldsymbol{\delta A}\left(\boldsymbol{L}^{-1}\right)^T$ wird sich um so weniger von der Einheitsmatrix unterscheiden, je genauer die Zerlegung ist. Unter der Voraussetzung

$$\left\|\boldsymbol{L}^{-1}\boldsymbol{\delta A}\left(\boldsymbol{L}^{-1}\right)^T\right\| < 1$$

ergibt sich die Abschätzung

$$\text{cond}(\bar{\boldsymbol{A}}) \leq \frac{1+\left\|\boldsymbol{L}^{-1}\boldsymbol{\delta A}\left(\boldsymbol{L}^{-1}\right)^T\right\|}{1-\left\|\boldsymbol{L}^{-1}\boldsymbol{\delta A}\left(\boldsymbol{L}^{-1}\right)^T\right\|} = 1+2\frac{\left\|\boldsymbol{L}^{-1}\boldsymbol{\delta A}\left(\boldsymbol{L}^{-1}\right)^T\right\|}{1-\left\|\boldsymbol{L}^{-1}\boldsymbol{\delta A}\left(\boldsymbol{L}^{-1}\right)^T\right\|}.$$

Für $\|\boldsymbol{\delta A}\| \ll 1$ gilt dann $\text{cond}(\bar{\boldsymbol{A}}) \approx 1$. Es ist zu erwarten, dass das Rundungsfehlerverhalten und damit auch das Konvergenzverhalten des transformierten Systems günstiger sind als beim ursprünglichen Problem. Nach einigen Umformungen erhalten wir den folgenden Algorithmus.

8.68. Vorkonditioniertes cg-Verfahren:
Zu lösen ist das lineare Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ mit einer symmetrischen, positiv definiten Matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, von der eine näherungsweise CHOLESKY*-Zerlegung $\boldsymbol{LL}^T = \boldsymbol{A} + \boldsymbol{\delta A}$ bekannt ist.*

S0 *Wähle einen Startvektor $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$.*

S1 *Berechne $\boldsymbol{r}^{(0)} = \boldsymbol{b} - \boldsymbol{Ax}^{(0)}$ und $\boldsymbol{h}^{(0)} = \left(\boldsymbol{LL}^T\right)^{-1}\boldsymbol{r}^{(0)}$. Setze $i = 0$.*

S2 *Falls $\boldsymbol{h}^{(i)} = \boldsymbol{o}$ STOPP. $\boldsymbol{x}^{(i)}$ ist die gesuchte Lösung.*

S3 *Berechne*

$$\begin{aligned}
\alpha_i &= \frac{\boldsymbol{r}^{(i)^T}\left(\boldsymbol{LL}^T\right)^{-1}\boldsymbol{r}^{(i)}}{\boldsymbol{h}^{(i)^T}\boldsymbol{Ah}^{(i)}},\\
\boldsymbol{x}^{(i+1)} &= \boldsymbol{x}^{(i)} + \alpha_i\boldsymbol{h}^{(i)},\\
\boldsymbol{r}^{(i+1)} &= \boldsymbol{r}^{(i)} - \alpha_i\boldsymbol{Ah}^{(i)},\\
\beta_i &= \frac{\boldsymbol{r}^{(i+1)^T}\left(\boldsymbol{LL}^T\right)^{-1}\boldsymbol{r}^{(i+1)}}{\boldsymbol{r}^{(i)^T}\left(\boldsymbol{LL}^T\right)^{-1}\boldsymbol{r}^{(i)}},\\
\boldsymbol{h}^{(i+1)} &= \left(\boldsymbol{LL}^T\right)^{-1}\boldsymbol{r}^{(i+1)} + \beta_i\boldsymbol{h}^{(i)}.
\end{aligned}$$

S4 *Setze $i = i+1$ und gehe zu Schritt* **S2**.

Auch das modifizierte cg-Verfahren für beliebige reguläre Matrizen läßt sich vorkonditionieren. Dazu benutzt man eine näherungsweise LU-Zerlegung der Matrix.

8.69. Vorkonditioniertes, modifiziertes cg-Verfahren:
Zu lösen ist das lineare Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ mit einer regulären Matrix, von der eine näherungsweise LU-Zerlegung $\boldsymbol{P}^T\boldsymbol{LU} = \boldsymbol{A} + \boldsymbol{\delta A}$ bekannt ist.

S0 *Wähle einen Startvektor $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$.*

S1 *Berechne* $\boldsymbol{r}^{(0)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(0)}$ *und* $\boldsymbol{h}^{(0)} = \left(\boldsymbol{U}^T\boldsymbol{U}\right)^{-1}\boldsymbol{A}^T\boldsymbol{P}^T\left(\boldsymbol{L}\boldsymbol{L}^T\right)^{-1}\boldsymbol{P}\boldsymbol{r}^{(0)}$. *Setze* $i = 0$.

S2 *Falls* $\boldsymbol{h}^{(i)} = \boldsymbol{o}$ *STOPP.* $\boldsymbol{x}^{(i)}$ *ist die gesuchte Lösung.*

S3 *Berechne*

$$\begin{aligned}
\alpha_i &= \frac{\left(\boldsymbol{P}\boldsymbol{r}^{(i)}\right)^T\left(\boldsymbol{L}\boldsymbol{L}^T\right)^{-1}\left(\boldsymbol{P}\boldsymbol{r}^{(i)}\right)}{\boldsymbol{h}^{(i)^T}\left(\boldsymbol{U}^T\boldsymbol{U}\right)\boldsymbol{h}^{(i)}},\\
\boldsymbol{x}^{(i+1)} &= \boldsymbol{x}^{(i)} + \alpha_i\boldsymbol{h}^{(i)},\\
\boldsymbol{r}^{(i+1)} &= \boldsymbol{r}^{(i)} - \alpha_i\boldsymbol{A}\boldsymbol{h}^{(i)},\\
\beta_i &= \frac{\left(\boldsymbol{P}\boldsymbol{r}^{(i+1)}\right)^T\left(\boldsymbol{L}\boldsymbol{L}^T\right)^{-1}\left(\boldsymbol{P}\boldsymbol{r}^{(i+1)}\right)}{\left(\boldsymbol{P}\boldsymbol{r}^{(i)}\right)^T\left(\boldsymbol{L}\boldsymbol{L}^T\right)^{-1}\left(\boldsymbol{P}\boldsymbol{r}^{(i)}\right)},\\
\boldsymbol{h}^{(i+1)} &= \left(\boldsymbol{U}^T\boldsymbol{U}\right)^{-1}\boldsymbol{A}^T\boldsymbol{P}^T\left(\boldsymbol{L}\boldsymbol{L}^T\right)^{-1}\boldsymbol{P}\boldsymbol{r}^{(i+1)} + \beta_i\boldsymbol{h}^{(i)}.
\end{aligned}$$

S4 *Setze* $i = i + 1$ *und gehe zu Schritt* **S2**.

8.4. Aufgaben

1. Man zeige:

   (a)

   $$\mathrm{lub}_1(\boldsymbol{A}) = \max_{1 \leqq j \leqq n}\left\{\sum_{k=1}^{n}|a_{kj}|\right\}$$

   (b)

   $$\mathrm{lub}_\infty(\boldsymbol{A}) = \max_{1 \leqq i \leqq n}\left\{\sum_{k=1}^{n}|a_{ik}|\right\}.$$

2. Man beweise: Jede Grenznorm ist submultiplikativ.

3. Am Beispiel der Matrizen

   $$\boldsymbol{A} = \begin{pmatrix} 0.6 & 0.6 \\ 0.6 & 0.6 \end{pmatrix}, \quad \boldsymbol{B} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

   ist zu zeigen, dass $\|\boldsymbol{A}\|_2$ weder eine absolute noch eine monotone Norm ist.

4. Man zeige: Für reguläre Matrizen $\boldsymbol{A}$ gilt:

$$\frac{1}{\text{lub}(\boldsymbol{A}^{-1})} = \min_{\boldsymbol{y}\neq\boldsymbol{o}} \frac{\|\boldsymbol{A}\boldsymbol{y}\|}{\|\boldsymbol{y}\|}.$$

5. Man zeige für eine reguläre Matrix $\boldsymbol{A}$ und Vektoren $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$:

   (a) Ist $\boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u} \neq -1$, so gilt

$$(\boldsymbol{A}+\boldsymbol{u}\boldsymbol{v}^T)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{A}^{-1}}{1+\boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u}}.$$

   (b) Ist $\boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u} = -1$, so ist $(\boldsymbol{A}+\boldsymbol{u}\boldsymbol{v}^T)$ singulär.
   (Hinweis: Finde einen Vektor $\boldsymbol{z} \neq \boldsymbol{o}$ mit $(\boldsymbol{A}+\boldsymbol{u}\boldsymbol{v}^T)\boldsymbol{z} = \boldsymbol{o}$!)

6. Es sei $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)$ eine reguläre Matrix mit den Spalten $\boldsymbol{a}_i$, $i = 1, \ldots, n$.

   (a) $\hat{\boldsymbol{A}} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}, \boldsymbol{b}, \boldsymbol{a}_{i+1}, \ldots, \boldsymbol{a}_n)$ mit $\boldsymbol{b} \in \mathbb{R}^n$ sei die Matrix, in der gegenüber $\boldsymbol{A}$ die i-te Spalte $\boldsymbol{a}_i$ durch $\boldsymbol{b}$ ersetzt wurde.
   Man untersuche mit Hilfe der Formel aus Aufgabe 5 unter welchen Bedingungen $\hat{\boldsymbol{A}}^{-1}$ existiert und zeige, dass dann $\hat{\boldsymbol{A}}^{-1} = \boldsymbol{F}\boldsymbol{A}^{-1}$ gilt, wobei $\boldsymbol{F}$ eine FROBENIUS-Matrix ist.

   (b) Es sei $\boldsymbol{A} = (a_{ik})_{n,n}$ regulär. $\boldsymbol{A}_\alpha$ entstehe aus $\boldsymbol{A}$ dadurch, dass ein einziges Element a_{ik} zu $a_{ik} + \alpha$ abgeändert wird. Für welche α existiert $\boldsymbol{A}_\alpha^{-1}$?

7. Es sei $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_n)$ mit $\sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_n > 0$ die Singulärwertzerlegung der (n,n)-Matrix $\boldsymbol{A}$.

   (a) Wie lautet $\text{cond}_2(\boldsymbol{A})$ ausgedrückt in den σ_i?

   (b) Man gebe mit Hilfe von $\boldsymbol{U}$ diejenigen Vektoren $\boldsymbol{b}$ bzw. $\boldsymbol{\delta b}$ an, die in den Abschätzungen

   i.
$$\|\boldsymbol{\delta x}\|_2 \leqq \|\boldsymbol{A}^{-1}\|_2 \|\boldsymbol{\delta b}\|_2,$$

   ii.
$$\frac{\|\boldsymbol{\delta x}\|_2}{\|\boldsymbol{x}\|_2} \leqq \|\boldsymbol{A}\|_2\|\boldsymbol{A}^{-1}\|_2 \frac{\|\boldsymbol{\delta b}\|_2}{\|\boldsymbol{b}\|_2} = \text{cond}_2(\boldsymbol{A})\frac{\|\boldsymbol{\delta b}\|_2}{\|\boldsymbol{b}\|_2} \quad \text{und}$$

   iii.
$$\|\boldsymbol{b}\|_2 \leqq \|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2$$

Gleichheit ergeben.

(c) Gibt es ein $\boldsymbol{b}$, so dass für alle $\boldsymbol{\delta b}$ in 7(b)ii. gilt:

$$\frac{\|\boldsymbol{\delta x}\|_2}{\|\boldsymbol{x}\|_2} \leqq \frac{\|\boldsymbol{\delta b}\|_2}{\|\boldsymbol{b}\|_2}?$$

Man bestimme solche Vektoren $\boldsymbol{b}$ mit Hilfe von $\boldsymbol{U}$.
(Hinweis: Man betrachte Vektoren $\boldsymbol{b}$ mit $\|\boldsymbol{A}^{-1}\|_2\|\boldsymbol{b}\|_2 = \|\boldsymbol{x}\|_2$.)

8. Man zeige, dass die FROBENIUS-Norm einer Matrix mit der euklidischen Vektornorm verträglich ist.

9. Es sei

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T, \quad \boldsymbol{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_l, 0, \ldots, 0)$$

und

$$\sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_r > 0, \quad \sigma_{r+1} = \cdots = \sigma_l = 0, \quad l = min(m,n)$$

die Singulärwertzerlegung der (n,n)-Matrix $\boldsymbol{A}$. Man zeige:

(a) $\text{rg}(\boldsymbol{A}) = r$

(b) $\|\boldsymbol{A}\|_2 = \sigma_1$

(c) $\|\boldsymbol{A}\|_F = \sqrt{\sigma_1^2 + \cdots \sigma_r^2}$

10. Man zeige:

(a) Es seien $\boldsymbol{L}$ und $\bar{\boldsymbol{L}}$ untere Einsdreiecksmatrizen. Dann ist das Produkt $\boldsymbol{L}\bar{\boldsymbol{L}}$ wieder untere Einsdreiecksmatrix.

(b) Mit $\boldsymbol{L}$ ist auch $\boldsymbol{L}^{-1}$ untere Einsdreiecksmatrix.

11. Man zeige, dass für beliebige Vektoren $\boldsymbol{x} \in \mathbb{R}^n$

(a)

$$\|\boldsymbol{x}\|_\infty \leqq \|\boldsymbol{x}\|_2 \leqq \sqrt{n}\|\boldsymbol{x}\|_\infty \quad \text{und}$$

(b)

$$\frac{1}{\sqrt{n}}\|\boldsymbol{x}\|_1 \leqq \|\boldsymbol{x}\|_2 \leqq \|\boldsymbol{x}\|_1.$$

gilt, und dass die auftretenden Konstanten nicht verbesserbar sind.

12. Wie groß ist der Aufwand an arithmetischen Operationen zum Lösen des Gleichungssystem

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},$$

falls von der (n,n)-Matrix $\boldsymbol{A}$ eine Zerlegung der Form $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$ bekannt ist?

13. Man zeige, dass für $1 \leqq k < i \leqq s \leqq n$

$$\boldsymbol{T}_{is}\boldsymbol{L}_k(\boldsymbol{l}) = \boldsymbol{L}_k(\boldsymbol{T}_{is}\boldsymbol{l})\boldsymbol{T}_{is}$$

gilt.

14. Man zeige:

$$\boldsymbol{L}_1(\boldsymbol{l}_1)\boldsymbol{L}_2(\boldsymbol{l}_2)\cdots\boldsymbol{L}_{n-1}(\boldsymbol{l}_{n-1}) = \boldsymbol{I} + (\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_{n-1}, 0).$$

15. Man zeige:
Wenn $\boldsymbol{A}$ symmetrisch ist und s(1) = 1 wählbar ist, so ist auch $\boldsymbol{M}^{(1)}$ symmetrisch.

16. Man zeige, dass für zeilendiagonaldominante Matrizen:

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \quad i = 1, \ldots, n$$

bei Durchführung des GAUSSschen Algorithmus ohne Pivotisierung gilt:

(a) Alle Matrizen $\boldsymbol{M}^{(k)}$ sind zeilendiagonaldominant,

(b) $\|\boldsymbol{M}^{(k)}\|_\infty \leqq \|\boldsymbol{M}^{(k-1)}\|_\infty \leqq \|\boldsymbol{A}\|_\infty$ für $k = 1, 2, \ldots, n-1$.

(Hinweis: siehe KIELBASINSKI/SCHWETLICK Numerische lineare Algebra, S. 166.)

17. Man überlege sich, dass für die Matrix

$$\boldsymbol{A} = \begin{pmatrix} 1 & & & & & a \\ -1 & 1 & & & & a \\ -1 & -1 & 1 & & & a \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ -1 & -1 & -1 & \ldots & 1 & a \\ -1 & -1 & -1 & \ldots & -1 & a \end{pmatrix}$$

im Falle der Spaltenpivotisierung

$$\|\boldsymbol{A}\|_\infty = |a| + n - 1, \quad \|\boldsymbol{M}^{(k)}\|_\infty = 2^k|a| + n - k - 1$$

gilt, so dass der Quotient $\frac{\|\boldsymbol{M}^{(k)}\|_\infty}{\|\boldsymbol{A}\|_\infty}$ der Schranke 2^k für $|a| \longrightarrow \infty$ beliebig nahe kommt.

18. Von der symmetrischen (n,n)-Matrix $\boldsymbol{A}$ sei das obere Dreieck

    (a) zeilenweise,

    (b) spaltenweise

    eindimensional gespeichert. Man schreibe ein Programm zum Berechnen von $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ bei vorgegebenem $\boldsymbol{x} \in \mathbb{R}^n$.

19. Es sei $\boldsymbol{A}$ eine reguläre Matrix mit den Zeilen $\boldsymbol{a}_i^T$, $i = 1, \ldots, n$:

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{a}_1^T \\ \boldsymbol{a}_2^T \\ \vdots \\ \boldsymbol{a}_n^T \end{pmatrix}.$$

    Weiterhin sei $\boldsymbol{D}$ eine Diagonalmatrix mit

$$\boldsymbol{D} = \text{diag}\left(\frac{\|\boldsymbol{A}\|_\infty}{\|\boldsymbol{a}_1^T\|_1}, \ldots, \frac{\|\boldsymbol{A}\|_\infty}{\|\boldsymbol{a}_n^T\|_1}\right).$$

    Man zeige, dass dann für die Matrix $\bar{\boldsymbol{A}} = \boldsymbol{D}\boldsymbol{A}$ folgendes gilt:

    (a) $\|\bar{\boldsymbol{A}}\|_\infty = \|\boldsymbol{A}\|_\infty$,

    (b) $\frac{\|\boldsymbol{A}^{-1}\|_\infty}{\max_i d_i} \leqq \|\bar{\boldsymbol{A}}^{-1}\|_\infty \leqq \|\boldsymbol{A}^{-1}\|_\infty$ und

    (c) $\frac{\text{cond}_\infty(\boldsymbol{A})}{\max_i d_i} \leqq \text{cond}_\infty(\bar{\boldsymbol{A}}) \leqq \text{cond}_\infty(\boldsymbol{A})$.

20. Es sei $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ die Singulärwertzerlegung der Matrix $\boldsymbol{A}$. Mit $\boldsymbol{c} = \boldsymbol{U}^T\boldsymbol{b}$ kann die Lösung von $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ in der Form $\boldsymbol{x} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{c}$ geschrieben werden. Man zeige, dass unter der Voraussetzung

$$c_n^2 \leqq \frac{0.01(c_1^2 + \cdots + c_n^2)}{n}$$

in der 2-Norm die Abschätzung

$$K_b(\boldsymbol{A},\boldsymbol{b}) = \frac{\|\boldsymbol{A}^{-1}\|_2\|\boldsymbol{b}\|_2}{\|\mathbf{x}\|_2} \leq 10\cdot\sqrt{n}$$

unabhängig von $\text{cond}_2(\boldsymbol{A})$ gilt.

21. Für das Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ mit

$$\boldsymbol{A} = \begin{pmatrix} 1.00 & 1.00 \\ 1.00 & 0.99 \end{pmatrix} \quad \boldsymbol{b} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

ist

$$\boldsymbol{A}^{-1} = \begin{pmatrix} -99 & 100 \\ 100 & -100 \end{pmatrix} \quad \boldsymbol{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

und $\text{cond}_\infty(\boldsymbol{A}) = 400$. Für die Störungen

(a)

$$\boldsymbol{\delta A} = 10^{-3}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \quad \boldsymbol{\delta b} = 10^{-3}\begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

(b)

$$\boldsymbol{\delta A} = 10^{-3}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad \boldsymbol{\delta b} = 10^{-3}\begin{pmatrix} -1 \\ -1 \end{pmatrix}$$

ist $\boldsymbol{\delta x}$ sowie $\|\boldsymbol{\delta x}\|_\infty$ zu berechnen und mit den Schranken gemäß der Vorlesung zu vergleichen.

22. Gegeben sei die positiv definite symmetrische Matrix $\bar{\boldsymbol{A}}$ mit der Partitionierung

$$\bar{\boldsymbol{A}} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^T & \boldsymbol{C} \end{pmatrix}, \quad \boldsymbol{A} \in \mathbb{R}^{m\times m}.$$

Weiterhin sei $\bar{\boldsymbol{A}} = \boldsymbol{L}\boldsymbol{L}^T$ mit

$$\boldsymbol{L} = \begin{pmatrix} \boldsymbol{L}_{11} & \boldsymbol{O} \\ \boldsymbol{L}_{21} & \boldsymbol{L}_{22} \end{pmatrix}, \quad \boldsymbol{L}_{11} \in \mathbb{R}^{m\times m}.$$

Man zeige:

(a) Die Matrix $\boldsymbol{M} = \boldsymbol{C} - \boldsymbol{B}^T \boldsymbol{A}^{-1} \boldsymbol{B}$ ist positiv definit.

(b) Es gilt $\boldsymbol{L}_{22} \boldsymbol{L}_{22}^T = \boldsymbol{M}$.

(c) Es gilt die Abschätzung

$$l_{ii}^2 \geqq \min_{\boldsymbol{x} \neq \boldsymbol{o}} \left(\frac{\boldsymbol{x}^T \bar{\boldsymbol{A}} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} \right) = \frac{1}{\text{lub}_2(\boldsymbol{L}^{-1})^2}, \quad i = 1, \ldots, n.$$

(d) Es gilt die Abschätzung

$$\text{lub}_2(\boldsymbol{L})^2 = \max_{\boldsymbol{x} \neq \boldsymbol{o}} \left(\frac{\boldsymbol{x}^T \bar{\boldsymbol{A}} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} \right) \geqq l_{ii}^2, \quad i = 1, \ldots, n.$$

(e) Es gilt die Abschätzung

$$\text{cond}_2(\boldsymbol{L}) \geqq \max_{1 \leqq i,k \leqq n} \left| \frac{l_{ii}}{l_{kk}} \right|.$$

23. Gegeben seien $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ mit $\boldsymbol{a} \neq \boldsymbol{b}$ und $\|\boldsymbol{a}\|_2 = \|\boldsymbol{b}\|_2$.
Man konstruiere eine HOUSEHOLDER-Spiegelung $\boldsymbol{H}$, für die $\boldsymbol{H}\boldsymbol{a} = \boldsymbol{b}$ gilt.

24. Man löse das Gleichungssystem $\mathbf{A}\mathbf{x} = \boldsymbol{b}$ mit

$$\boldsymbol{A} = \begin{pmatrix} \frac{1}{3} & -1 & \frac{5}{6} \\ \frac{2}{3} & 0 & \frac{1}{6} \\ \frac{2}{3} & \frac{1}{5} & \frac{1}{6} \end{pmatrix} \quad \boldsymbol{b} = \begin{pmatrix} \frac{1}{6} \\ \frac{5}{6} \\ \frac{31}{30} \end{pmatrix}$$

durch das HOUSEHOLDER-Verfahren!

25. Man zeige für eine (n,n)-Matrix $\boldsymbol{P}$, $\quad \|\boldsymbol{P}\| < 1$:

$$(\boldsymbol{I} - \boldsymbol{P})^{-1} = \boldsymbol{I} + \sum_{i=1}^{\infty} \boldsymbol{P}^i$$

Hinweis: Man zeige für die Partialsummen

$$\boldsymbol{S}_n = \boldsymbol{I} + \sum_{i=1}^{n} \boldsymbol{P}^i \quad \text{gilt} \quad (\boldsymbol{I} - \boldsymbol{P})^{-1} - \boldsymbol{S}_n = (\boldsymbol{I} - \boldsymbol{P})^{-1} \boldsymbol{P}^{n+1}.$$

26. Wann konvergiert das JACOBI-Verfahren bzw. das GAUSS-SEIDEL-Verfahren für die Matrix

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{S} \\ \boldsymbol{S}^T & \boldsymbol{I} \end{pmatrix}, \quad \boldsymbol{I}, \boldsymbol{S} \in \mathbb{R}^{n \times n}?$$

27. Man zeige: Ist die (n,n)-Matrix $\boldsymbol{A}$ unzerlegbar, so ist die Matrix

$$\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A}$$

für das JACOBI-Verfahren ebenfalls unzerlegbar.
Hinweis: $\boldsymbol{B}$ ist Diagonalmatrix.

28. Gegeben sei

$$\boldsymbol{A} = \begin{pmatrix} 2 & 0 & -1 & -1 \\ 0 & 2 & -1 & -1 \\ -1 & -1 & 2 & 0 \\ -1 & -1 & 0 & 2 \end{pmatrix}.$$

Man zeige:

(a) $\boldsymbol{A}$ ist unzerlegbar.

(b) Das Gesamtschrittverfahren konvergiert nicht.

Index