# Numerik I, aktueller Vorlesungsstand

Günter Bärwolff

27. Januar 2012

# Inhaltsverzeichnis

# Kapitel 0

# Vorwort

Als Literaturempfehlungen seien z.B. die Lehrbücher von

- Robert Plato: Numerische Mathematik kompakt. Grundlagenwissen für Studium und Praxis
- Hans R. Schwarz, Norbert Köckler: Numerische Mathematik
- Günter Bärwollf: Numerik für Ingenieure, Physiker und Informatiker
- Matthias Bollhöfer, Volker Mehrmann: Numerische Mathematik

empfohlen, in denen die Themen der Vorlesung mehr oder wenig ausführlich dargestellt sind.

# Kapitel 1

# Rechnerarithmetik

1. Vorlesung am 17.10.2012

Bei unterschiedlichen "Rechenaufgaben" treten unterschiedliche Fehler auf, und zwar

- Datenfehler aufgrund ungenauer Eingabedaten
- Darstellungsfehler von Zahlen
- Fehler durch ungenaue Rechnungen, z.B. wird man bei der Aufabe $\frac{1}{3} = 0.33333\ldots$ eigentlich nie fertig, d.h. man gibt irgendwann erschöpft auf und macht einen Fehler.

## 1.1 Zahldarstellungen

Aus der Analysis ist bekannt, dass man jede Zahl $x \in \mathbb{R}, x \neq 0$ bei einer gegebenen **Basis** $b \in \mathbb{N}, b \geq 2$ in der Form

$$x = \sigma \sum_{i=-e+1}^{\infty} a_{i+e} b^{-i} = \sigma \left( \sum_{i=1}^{\infty} a_i b^{-i} \right) b^e \tag{1.1}$$

mit $a_1, a_2, \ldots \in \{0, 1, \ldots, b-1\}, e \in \mathbb{Z}, \sigma \in \{+, -\}$ darstellen kann, wobei $a_1 \neq 0$ ist. (Fordert man, dass eine unendliche Teilmenge $\mathbb{N}_1 \subset \mathbb{N}$ gibt mit $a_i \neq b-1$ fur $i \in \mathbb{N}_1$, dann ist die Darstellung (1.1) eindeutig). (1.1) heißt **Gleitpunktdarstellung**. Als Basis $b$ wird oft $b = 10$ (Schule) oder $b = 2$ benutzt. Man spricht vom Dezimal- bzw. Dualsystem.

## 1.2 Allgemeine Gleitpunkt-Zahlensysteme

Da man auf Rechnern nicht beliebig viele Stellen zur Darstellung von Zahlen in der Form (1.1) zur Verfügung hat, z.B. für die Zahlen $\frac{1}{3} = (\sum_{i=1}^{\infty} 3 \cdot 10^{-i}) 10^0$

im Dezimalsystem oder $\frac{2}{3} = (\sum_{i=1}^{\infty} c_i \cdot 2^{-i})2^0$ mit $c_{2k-1} = 0, c_{2k} = 1$ im Dualsystem, arbeitet man mit Gleitpunktzahlensystemen wie folgt

**Definition 1.1.** *Zu gegebener Basis $b \geq 2$ und* ***Mantissenlänge*** *$t \in \mathbb{N}$ sowie für Exponentenschranken $e_{min} < 0 < e_{max}$ ist die Menge $F = F(b, t, e_{min}, e_{max}) \subset \mathbb{R}$ durch*

$$F = \left\{ \sigma \left( \sum_{i=1}^{t} a_i b^{-i} \right) b^e : a_1, \ldots, a_t \in \{0, 1, \ldots, b-1\}, a_1 \neq 0, e \in \mathbb{Z}, \right. \\ \left. e_{min} \leq e \leq e_{max}, \sigma \in \{+, -\} \right\} \cup \{0\} \tag{1.2}$$

*erklärt und wird System von* ***normalisierten*** *Gleitpunktzahlen genannt. Lässt man noch die Kombination $e = e_{min}, a_1 = 0$ zu, dann erhält man mit $\hat{F} \supset F$ das System der* ***denormalisierten*** *Gleitpunktzahlen.*

Statt der Angabe von Exponentenschrankten $e_{\min}, e_{\max} \in \mathbb{Z}$ wird bei einem Gleitpunktzahlensystem auch mit $l$ die Stellenzahl des Exponenten $e$ angegeben, sodass man statt

$$F = F(2, 24, -127, 127) \tag{1.3}$$

auch

$$F = F(2, 24, 7)$$

schreiben kann, da man mit einer 7-stelligen Dualzahl alle Exponenten von 0 bis $\pm 127$ darstellen kann. Statt $F$ wird auch $M$ (Maschinenzahlen) als Symbol genutzt, also z.B.

$$M = F(2, 24, 7) = M(2, 24, 7) \tag{1.4}$$

Die Darstellung (1.3) ist aber oft präziser, da in der Praxis tatsachlich $|e_{\min}| \neq e_{\max}$ ist, was bei (1.4) nicht zu erkennen ist.

## 1.3 Struktur und Eigenschaften des normalisierten Gleitpunktzahlensystems F

2. Vorlesung am 14.4.2011

Es ist offensichtlich, dass die Elemente von $F$ symmetrisch um den Nullpunkt liegen, weshalb hier nur die positiven Elemente betrachtet werden sollen. Konkret betrachten wir $F = F(b, t, e_{\min}, e_{\max})$ und finden mit

$$x_{\min} = (1 \cdot b^{-1} + 0 \cdot b^{-2} + \cdots + 0 \cdot b^{-t}) \cdot b^{e_{\min}} = b^{-1+e_{\min}} \tag{1.5}$$

die **kleinste positive normalisierte Gleitpunktzahl**. Andererseits ergibt sich mit

$$\begin{aligned} x_{\max} &= ((b-1) \cdot b^{-1} + (b-1) \cdot b^{-2} + \cdots + (b-1) \cdot b^{-t}) \cdot b^{e_{\max}} \\ &= (1 - b^{-1} + b^{-1} - b^{-2} + \cdots - b^{-t}) \cdot b^{e_{\max}} \\ &= (1 - b^{-t}) \cdot b^{e_{\max}} \end{aligned} \tag{1.6}$$

**die größte positive normalisierte Gleitpunktzahl**. Für die Mantissen $a$ von Zahlen aus $F$ ergibt sich aus (1.5) und (1.6)

$$b^{-1} \leq a \leq 1 - b^{-t} \tag{1.7}$$

In $\hat{F}$ (Menge der denormalisierten Gleitpunktzahlen) sind kleinere Zahlen als $x_{\min}$ darstellbar und zwar mit

$$\hat{x}_{\min} = (0 \cdot b^{-1} + \cdots + 1 \cdot b^{-t}) \cdot b^{e_{\min}} = b^{-t+e_{\min}} \tag{1.8}$$

**die kleinste positive denormalisierte Gleitpunktzahl**.
Mit der Festlegung einer Mantissenlänge t ist die Anzahl der möglichen Mantissen festgelegt, sodass in jedem Intervall $]b^{e-1}, b^e[$ gleich viele Gleitpunktzahlen liegen, die außerdem äquidistant verteilt sind, und zwar mit dem Abstand

$$\Delta = b^{-t} \cdot b^e = b^{e-t}$$

Der Abstand einer beliebigen reellen Zahl $x \in [b^{e-1}, b^e]$ zum nächstgelegenen Element $z$ aus $F$ ist damit durch $\frac{1}{2}\Delta$ begrenzt, d.h.

$$|z - x| \leq \frac{1}{2} b^{e-t} \tag{1.9}$$

Die Gleichheit wird erreicht, wenn $x$ genau zwischen zwei benachbarten Zahlen aus $F$ liegt, wegen $b^{e-1} \leq x$ folgt aus (1.9)

$$\frac{|z - x|}{|x|} \leq \frac{\frac{1}{2} b^{e-t}}{b^{e-1}} = \frac{1}{2} b^{-t+1} =: \text{eps} \tag{1.10}$$

mit eps $= \frac{1}{2}b^{-t+1}$ der **maximale relative** Abstand der Zahlen $\{x \in \mathbb{R} : x_{\min} \leq |x| \leq x_{\max}\}$ zum nächstgelegenen Element aus $F$.
Mit der Kenntnis von eps lässt sich nun über die Bedingung

$$0.5 \cdot 10^{-n} \leq \text{eps} \leq 5 \cdot 10^{-n} \tag{1.11}$$

eine Zahl $n \in \mathbb{N}$ bestimmen, und man spricht dann beim Gleitpunktzahlensystem $F$ von einer **n-stelligen Dezimalstellenarithmetik**.
Als Beispiele von in der Praxis benutzten Gleitpunktzahlensystemen seien hier IEEE-Standardsystem

- $\hat{F}(2, 24, -125, 128)$ (einfach, real*4)
- $\hat{F}(2, 53, -1021, 1024)$ (doppelt, real*8)

sowie die IBM-Systeme

- $F(16, 6, -64, 63)$ einfach
- $F(16, 14, -64, 63)$ doppelt

genannt.

## 1.4 Rechnen mit Gleitpunktzahlen

Einfache Rechnungen zeigen, dass Gleitpunktzahlensysteme hinsichtlich der Adiition/Subtraktion bzw. Multiplikation/Division nicht abgeschlossen sind, d.h. Addition oder Multiplikation von Zahlen $x, y \in F$ ergibt i.A. keine Zahl aus $F$.

**Beispiel 1.2.** $F(10, 4, -63, 64), x = 0.1502 \cdot 10^2, y = 0.1 \cdot 10^{-4}$

$$x + y = 15.02 + 0.00001 = 15.02001 = 0.1502001 \cdot 10^2$$

Hier reicht die Stellenzahl $t = 4$ nicht aus, um $x + y$ in $F$ exakt darzustellen.

Um in einem Gleitpunktzahlenystem rechnen zu können braucht man letztendlich eine Abbildung aus $\mathbb{R}$ in $F$

**Definition 1.3.** *Zu einem gegebenen Gleitpunktzahlensystem $F(b, t, e_{min}, e_{max})$ mit gerader Basis $b$ ist die Funktion* $\text{rd} : \{x \in \mathbb{R} : x_{min} \leq |x| \leq x_{max}\} \to \mathbb{R}$ *durch*

$$\text{rd}(x) = \begin{cases} \sigma \cdot (\sum_{k=1}^{t} a_k b^{-k}) \cdot b^e & falls \quad a_{t+1} \leq \frac{1}{2}b - 1 \\ \sigma \cdot (\sum_{k=1}^{t} a_k b^{-k} + b^{-t}) \cdot b^e & falls \quad a_{t+1} \geq \frac{1}{2}b \end{cases}$$

*für $x = \sigma \cdot (\sum_{k=1}^{\infty} a_k b^{-k}) \cdot b^e$ erklärt.* $\text{rd}(x)$ *heisst auf* ***t Stellen gerundeter Wert*** *von $x$*

Man kann nun folgende Eigenschaften für das Runden zeigen:

**Theorem 1.4.** *Zu einem gegebenen Gleitpunktzahlensystem $F(b,t,e_{min},e_{max})$ gilt für jede reelle Zahl $x$ mit $|x| \in [x_{min}, x_{max}]$ die Eigenschaft* $\operatorname{rd}(x) \in F$ *und die Minimaleigenschaft*

$$|\operatorname{rd}(x) - x| = \min_{z \in F} |z - x|$$

*Beweis.* Es gilt offensichtlich

$$\sum_{k=1}^{t} a_k b^{-k} \leq \sum_{k=1}^{\infty} a_k b^{-k} \leq \sum_{k=1}^{t} a_k b^{-k} + \sum_{k=t+1}^{\infty} (b-1) \cdot b^{-k} = \sum_{k=1}^{t} a_k b^{-k} + b^{-t}$$

Nach Multiplikation mit $b^e$ erhält man

$$\underbrace{\left(\sum_{k=1}^{t} a_k b^{-k}\right)}_{\geq b^{-1}} \cdot b^e \leq \left(\sum_{k=1}^{\infty} a_k b^{-k}\right) \cdot b^e = |x| \leq \underbrace{\left(\sum_{k=1}^{t} a_k b^{-k} + b^{-t}\right)}_{\leq 1} \cdot b^e$$

d.h. die Schranken von $|x|$ liegen im Intervall $[b^{e-1}, b^e]$ und damit sind die beiden für $\operatorname{rd}(x)$ infrage kommenden Werte

$$\sigma\left(\sum_{k=1}^{t} a_k b^{-k}\right) \cdot b^e \quad \text{und} \quad \sigma\left(\sum_{k=1}^{t} a_k b^{-k} + b^{-t}\right) \cdot b^e$$

die Nachbarn von $x$ aus $F$, also ist $\operatorname{rd}(x) \in F$.
Es wird nun die Abschätzung

$$|\operatorname{rd}(x) - x| \leq \frac{1}{2} b^{-t+e} \tag{1.12}$$

gezeigt.
Für $a_{t+1} \leq \frac{b}{2} - 1$ (abrunden) erhält man

$$\begin{aligned}
|\operatorname{rd}(x) - x| &= \left(\sum_{k=t+1}^{\infty} a_k b^{-k}\right) \cdot b^e = \left(a_{t+1} b^{-(t+1)} + \sum_{k=t+2}^{\infty} a_k b^{-k}\right) \cdot b^e \\
&\leq \left[\left(\frac{b}{2} - 1\right) \cdot b^{-(t+1)} + \sum_{k=t+2}^{\infty} (b-1) \cdot b^{-k}\right] \cdot b^e \\
&= \left[\left(\frac{b}{2} - 1\right) b^{-(t+1)} + b^{-(t+1)}\right] \cdot b^e = \frac{1}{2} b^{-t+e}
\end{aligned}$$

Beim Aufrunden, d.h. $a_{t+1} \geq \frac{b}{2}$, ergibt sich

$$\begin{aligned} |\operatorname{rd}(x) - x| &= \left( b^{-t} - \sum_{k=t+1}^{\infty} a_k b^{-k} \right) \cdot b^e \\ &= \left( b^{-t} - \underbrace{a_{k+1} b^{-(t+1)}}_{\geq \frac{1}{2} b^{-t}} - \underbrace{\sum_{k=t+2}^{\infty} a_k b^{-k}}_{\geq 0} \right) \cdot b^e \leq \frac{1}{2} b^{-t+e} \end{aligned}$$

Da wir früher gezeigt haben, dass $\frac{1}{2}b^{-t+e}$ die Hälfte des Abstandes zweier Nachbarn in $F$ darstellt, folgt aus (1.12)

$$|\operatorname{rd}(x) - x| = \min_{z \in F} |z - x| \tag{1.13}$$

Als Folgerung aus (1.12) erhält man wegen $|x| \geq b^{e_{\min}}$

$$\frac{|\operatorname{rd}(x) - x|}{|x|} \leq \frac{1}{2} b^{-t+1} = \text{eps} \quad \text{(Maschinenepsilon)} \tag{1.14}$$

als Abschätzung für den relativen Rundungsfehler □

**Definition 1.5.** $\text{eps} = \frac{1}{2}b^{-t+1}$ *als Schranke für den relativen Rundungsfehler heißt Maschinengenauigkeit oder roundoff unit u (es werden auch die Bezeichnungen macheps oder* $\text{eps}^*$ *verwendet, es gilt*

$$\text{eps} = \inf\{\delta > 0 : \operatorname{rd}(1 + \delta) > 1\} \, .$$

Neben der Möglichkeit des Rundens mit rd gibt es auch als Alternative das **Abschneiden** (englisch truncate).

2. Vorlesung am 19.10.2012

**Definition 1.6.** *Zu $F = F(b, t, e_{min}, e_{max})$ ist die Funktion*

$$\operatorname{tc} : \{x \in \mathbb{R} : x_{min} \leq |x| \leq x_{max}\} \to F$$

*durch*

$$\operatorname{tc}(x) = \sigma \cdot \left( \sum_{k=1}^{t} a_k b^{-k} \right) \cdot b^e \quad \textit{für} \quad x = \sigma \cdot \left( \sum_{k=1}^{\infty} a_k b^{-k} \right) \cdot b^e$$

*erklärt.*

**Bemerkung 1.7.** Abschneiden ist i.A. ungenauer als Runden und es gilt

$$\frac{|\operatorname{tc}(x) - x|}{|x|} \leq 2 \cdot \text{eps}$$

für $x \in \mathbb{R}$ mit $x_{\min} \leq |x| \leq x_{\max}$.

## 1.5 Ersatzarithmetik

Durch Runden oder abschneiden gelingt es, reelle Zahlen $x$ mit $x_{\min} \leq |x| \leq x_{\max}$ in ein gegebenes Gleitpunktzahlensystem $F(b, t, e_{\min}, e_{\max})$ abzubilden. Deshalb werden die Grundoperationen $\circ \in \{+, -, \cdot, :\}$ oft durch

$$x \,\tilde{\circ}\, y = \mathrm{rd}(x \circ y) \quad \text{für} \quad x, y \in F, x_{\min} \leq |x \circ y| \leq x_{\max} \tag{1.15}$$

oder

$$x \,\tilde{\circ}\, y = \mathrm{tc}(x \circ y) \quad \text{für} \quad x, y \in F, x_{\min} \leq |x \circ y| \leq x_{\max} \tag{1.16}$$

auf Rechner realisiert (bei Division soll $y \neq 0$ sein)

**Theorem 1.8.** *Bezüglich der durch* (1.15) *bzw.* (1.16) *definierten Ersatzoperationen* $\tilde{+}, \tilde{-}, \tilde{\cdot}, \tilde{:}$ *ist* $F$ *abgeschlossen, d.h. im Ergebnis dieser Operationen erhält man Elemente aus* $F$*. Außerdem gilt die Beziehung bzw. Darstellung*

$$x \,\tilde{\circ}\, y = (x \circ y) \cdot (1 + \epsilon) \quad \textit{mit} \quad |\epsilon| \leq k \cdot \mathrm{eps} \tag{1.17}$$

*wobei im Fall von* (1.15) $k$ *gleich* 1 *und im Fall von* (1.16) $k$ *gleich* 2 *ist (*$\epsilon$ *heißt* ***Darstellungsfehler****)*

*Beweis.* Die Abgeschlossenheit von F bezüglich $\tilde{\circ}$ folgt aus Theroem 1.8. Die Darstellung (1.16) ergibt sich im Falle von (1.15) aus

$$\frac{|\,\mathrm{rd}(x \circ y) - (x \circ y)|}{|x \circ y|} \leq \mathrm{eps}$$

also aus (1.14). □

## 1.6 Fehlerakkumulation

Wir betrachten Zahlen $x, y \in \mathbb{R}$. Durch eine eventuelle Rundung erhalten wir mit

$$\mathrm{rd}(x) = x + \Delta x \in F$$
$$\mathrm{rd}(y) = y + \Delta y \in F$$

mit $\frac{|\Delta x|}{|x|}, \frac{|\Delta y|}{|y|} \leq \epsilon$ Zahlen aus einem Gleitpunktzahlensystem $F$ ($\epsilon = eps$ im vorliegenden Fall der Rundung, $\epsilon = 2\,eps$ im Falle des Abschneidens). $\tilde{\circ}, \circ$ sei nun Multiplikation oder Division. Mit (1.15) und (1.17) erhält man

$$\begin{aligned}(x + \Delta x) \,\tilde{\circ}\, (y + \Delta y) &= (x \cdot (1 + \tau_x)) \,\tilde{\circ}\, (y \cdot (1 + \tau_y)), \quad |\tau_x|, |\tau_y| \leq \epsilon \\ &= (x \circ y)((1 + \tau_x) \circ (1 + \tau_y))(1 + \alpha), \quad |\alpha| \leq \epsilon \\ &= (x \circ y)(1 + \beta)\end{aligned}$$

wobei man benutzt, dass

$$(1+\tau_x)\circ(1+\tau_y)(1+\alpha) = (+\tau_x)^{\sigma_1}(1+\tau_y)^{\sigma_2} = 1+\beta\,, \quad \sigma_1,\sigma_2 \in \{-1,+1\}, \tag{1.18}$$

mit einem $\beta$ mit der Eigenschaft $|\beta| \leq \frac{3\epsilon}{1-3\epsilon}$ gilt. Die Beziehung (1.18) ist ein Spezialfall der Beziehung

$$\Pi_{k=1}^{n}(1+\tau_k)^{\sigma_k} = 1+\beta_n\,, \quad |\beta_n| \leq \frac{n\epsilon}{1-n\epsilon}$$

für Zahlen $\tau_k \in \mathbb{R}$ mit $|\tau_k| \leq \epsilon$ und Exponenten $\sigma_k \in \{-1,+1\}$ (für $n\epsilon < 1$), die man mit vollst. Induktion zwar etwas technisch, aber doch recht leicht nachweist (Beweis z.B. bei Plato). Damit ergibt sich für die Multiplikation/-Division das

**Theorem 1.9.** *Zu dem Gleitpunktzahlensystem $F(b,t,e_{min},e_{max})$ seien die Zahlen $x,y \in \mathbb{R}$ und $\Delta x, \Delta y \in \mathbb{R}$ gegeben mit $x+\Delta x \in F, y+\Delta y \in F, \frac{|\Delta x|}{|x|}, \frac{|\Delta y|}{|y|} \leq \epsilon$ mit $\epsilon < \frac{1}{4}$. $\circ$ steht für die Grundoperation $\cdot$ bzw. : und für $x \circ y$ soll $x_{min} \leq |x \circ y| \leq x_{max}$ gelten. Dann gilt die Fehlerdarstellung*

$$(x+\Delta x)\,\tilde{\circ}\,(y+\Delta y) = x \circ y + \eta \tag{1.19}$$

*mit $\frac{|\eta|}{|x\circ y|} \leq \frac{3\epsilon}{1-3\epsilon}$.*

Die Darstellung (1.19) zeigt, dass die Multiplikation bzw. Division verhältnismäßig gutartig mit einem kleinen relativen Fehler ist. Im Folgenden soll noch auf die Fehlerverstärkung bei der Hintereinanderausführung von Addition in einem gegebenen GPZS $F$ hingewiesen werden. Es gilt das

**Theorem 1.10.** *Zu $F(b,t,e_{min},e_{max})$ seien $x_1,\ldots,x_n \in \mathbb{R}$ und $\Delta x_1,\ldots,\Delta x_n \in \mathbb{R}$ Zahlen mit*

$$x_k + \Delta x_k \in F, \frac{|\Delta x_k|}{|x_k|} \leq \epsilon \quad \textit{für} \quad k = 1,..,n$$

*und es bezeichne*

$$\tilde{S}_k := \sum_{j=1}^{k}(x_j + \Delta x_j), \quad S_k := \sum_{j=1}^{k} x_j, \quad k = 1,\ldots,n$$

*die entsprechenden Partialsummen (Summation von links nach rechts), wobei die Summe $\tilde{S}_k$ als Summe im Gleitpunktzahlensystem $F$ zu verstehen ist (die einzelnen Summanden werden also durch $\tilde{+}$ verknüpft). Dann gilt*

$$|\tilde{S}_k - S_k| \leq \underbrace{\left(\sum_{j=1}^{k}(1+\epsilon)^{k-j}(2|x_j| + |S_j|)\right)}_{=:M_k}\epsilon \quad \textit{für} \quad k = 1,..,n \tag{1.20}$$

*falls die Partialsummen (Notation $M_0 = 0$) innerhalb gewisser Schranken liegen:*

$$x_{min} + (M_{k-1} + |x_k|)\epsilon \leq |S_k| \leq x_{max} - (M_{k-1} + |x_k|)\epsilon \quad k = 1, \ldots, n\,. \tag{1.21}$$

*Beweis.* (nach Plato)
Der Beweis erfolgt mit vollst. Induktion. Der Induktionsanfang ($k = 1$) ist wegen $\frac{|\Delta x_1|}{|x_1|} \leq \epsilon$ offensichtlich. Unter der Annahme, dass (1.20) für $k \geq 1$ richtig ist, ergibt sich mit den Verabredungen

$$\Delta S_j = \tilde{S}_j - S_j \quad \text{für } j \geq 1,\ \Delta S_0 = 0$$

die folgende Rechnung für eine Zahl $\tau_k \in \mathbb{R}$ mit $\tau_k \leq \epsilon$

$$\begin{aligned}
\Delta S_k &= \tilde{S}_k - S_k = \tilde{S}_{k-1}\tilde{+}(x_k + \Delta x_k) - S_k \\
&= (S_{k-1} + \Delta S_{k-1})\tilde{+}(x_k + \Delta x_k) - S_k \\
&= (S_{k-1} + x_k + \Delta S_{k-1} + \Delta x_k)(1 + \tau_k) - S_k \\
&= (S_k + \Delta S_{k-1} + \Delta x_k)(1 + \tau_k) - S_k \\
&= (1 + \tau_k)\Delta S_{k-1} + \tau_k S_k + (1 + \tau_k)\Delta x_k
\end{aligned}$$

und damit

$$|\Delta S_k| \leq (1 + \epsilon)|\Delta_{k-1}| + \epsilon(|S_k| + 2|x_k|)\,. \tag{1.22}$$

Aus (1.22) und der Induktionsvoraussetzung folgt die Aussage (1.20). Die Voraussetzung (1.21) sichert, dass die Resultate der Additionen in $F$ im relevanten Bereich $[x_{min}, x_{max}]$ liegen. □

**Bemerkung 1.11.** Der Faktor $(1+\epsilon)^{n-j}$ in der Abschätzung (1.20) ist umso größer, je kleiner $j$ ist. Daher ist es vorteilhaft beim Aufsummieren mit den betragsmäßig kleinen Zahlen zu beginnen. Dies gewährleistet zudem, dass die Partialsummen $S_k$ betragsmäßig nicht unnötig anwachsen. Theorem 1.10 liefert mit (1.20) nur eine Abschätzung für den absoluten Fehler. Der relative Fehler $\frac{|\tilde{S}_n - S_n|}{|S_n|}$ kann jedoch groß ausfallen, falls $|S_n|$ klein gegenüber $\sum_{j=1}^{n-1}(|x_j| + |S_j|) + |x_n|$ ist!

**Definition 1.12.** *(Landausche $\mathcal{O}$-Notation)*
*Sei $h : U \to R^n$ eine Funktion, $U$ offen, $x_0 \in U$, Dann bezeichnet das Landau-Symbol*[1]

$$\mathcal{O}(||h(x)||),\ \ x \to x_0$$

[1] Landau, Edmund Georg Herrmann 1877-1938

*eine (nicht näher spezifizierte) Funktion $\varphi$ mit der Eigenschaft*

$$limsup_{x\to x_0}\frac{||\varphi(x)||}{||h(x)||} < \infty\ .$$

*Das Landau-Symbol*

$$o(||h(x)||),\ \ x\to x_0$$

*beschreibt eine (nicht näher spezifizierte) Funktion $\varphi$ mit der Eigenschaft*

$$lim_{x\to x_0}\frac{||\varphi(x)||}{||h(x)||} = 0\ .$$

*Für Funktionen $g$, $h$ sind die Notationen*

$$\begin{aligned} g &\doteq h \quad x\to x_0,\\ g &\dot{\leq} h \quad x\to x_0, \end{aligned}$$

*eine abkürzende Schreibweise für*

$$\begin{aligned} g(x) &= h(x)+o(||h(x)||),\ \ x\to x_0,\\ g(x) &\dot{\leq} h(x)+o(||h(x)||),\ \ x\to x_0\ (komponentenweise). \end{aligned}$$

**Beispiel 1.13.**

$$\begin{aligned} x\sin x &= \mathcal{O}(x^2),\ x\to 0, \quad x\sin x \doteq x^2\ x\to 0,\\ x\sin x &= o(x),\ x\to 0,\\ 2\cos x &= \mathcal{O}(1),\ x\to 0,\ \ 2\cos x \doteq 2,\ x\to 0,\\ P_n(x)\,e^{-x} &= \mathcal{O}(e^{-\alpha x}),\ x\to\infty \end{aligned}$$

für jedes Polynom $P_n$ vom Grad $n\in\mathbb{N}_0$ und jedes $0<\alpha<1$.

# Kapitel 2

# Stabilität, Vorwärtsanalyse, Rückwärtsanalyse

Nachdem die Fehlerverstärkung bei Grundoperationen in einem GPZS betrachtet wurde, soll nun etwas allgemeiner das Problem der Fehlerfortpflanzung bei Rechenalgorithmen diskutiert werden.
Allgemein beschreibt die Stabilität die Robustheit numerischer Verfahren gegenüber Störungen in den Eingabedaten. Ein gegebenes Problem oder ein Algorithmus soll durch die Funktion

$$f : \mathbb{R}^n \to \mathbb{R}^m \tag{2.1}$$

beschrieben werden, wobei eine explizite Formel für $f$ vorliegen soll. Zur Allgemeinheit bedeutet (2.1), dass ausgehend von $n$ Eingangsdaten $m$ Ergebnisse des Problems berechnet werden.
Der besseren Übersichtlichkeit halber betrachten wir später skalarwertige Probleme, d.h. $m = 1$.

## 2.1 Kondition als Maß für Fehlerverstärkungen

**Definition 2.1.** *Die absolute normweise Kondition des Problems $x \mapsto f(x)$ ist die kleinste Zahl $\kappa_{abs} \geq 0$, sodass*

$$\|f(\tilde{x}) - f(x)\| \dot{\leq} \kappa_{abs} \|\tilde{x} - x\| \quad \tilde{x} \to x$$

*Das Problem heißt schlecht gestellt, falls es keine solche Zahl gibt ($\kappa_{abs} = \infty$). Analog ist die relative normweise Kondition die kleinste Zahl $\kappa_{rel} \geq 0$ mit*

$$\frac{\|f(\tilde{x}) - f(x)\|}{\|f(x)\|} \dot{\leq} \kappa_{rel} \frac{\|\tilde{x} - x\|}{\|x\|} \quad \tilde{x} \to x$$

**Bemerkung 2.2.** $\kappa$ klein bedeutet grob ein gut konditioniertes Problem
$\kappa$ gross ein schlecht Konditioniertes

**Beispiel 2.3.** $f$ differenzierbar

$$\begin{aligned} \|f(\tilde{x}) - f(x)\| &= \|f'(x)(\tilde{x} - x) + o(\|\tilde{x} - x\|)\| \\ &\leq \underbrace{\|f'(x)\|}_{\kappa_{\text{abs}}} \cdot \|\tilde{x} - x\| + o(\|\tilde{x} - x\|) \end{aligned}$$

$$\Rightarrow \frac{\|f(\tilde{(}x) - f(x)\|}{\|f(x)\|} \leq \underbrace{\frac{\|f'(x)\| \cdot \|x\|}{\|f(x)\|}}_{\kappa_{\text{rel}}} \frac{\|\tilde{x} - x\|}{\|x\|}$$

3. Vorlesung am 24.10.2012

## 2.2 Vektor- und Matrixnormen

- $\|\vec{x}\|_1 = \sum_{k=1}^{n} |x_k|$ Summennorm
- $\|\vec{x}\|_2 = \sqrt{\sum_{k=1}^{n} |x_k|^2}$ Euklidische Norm
- $\|\vec{x}\|_\infty = \max_{1\leq k\leq n}\{|x_k|\}$ Maximumsnorm

Durch Vektornormen induzierte Matrixnormen

$$\|A\|_\sigma = \max_{\vec{x}\in\mathbb{R}^n, \vec{x}\neq 0} \frac{\|A\vec{x}\|_\sigma}{\|\vec{x}\|_\sigma} = \max_{\vec{y}\in\mathbb{R}^n, \|\vec{y}\|=1} \|A\vec{y}\|_\sigma$$

mit $\sigma \in \{1, 2, \infty\}$. Es gilt

1. $\|A\vec{x}\|_\sigma \leq \|A\|_\sigma \|\vec{x}\|_\sigma$ Verträglichkeit
2. $\|AB\|_\sigma \leq \|A\|_\sigma \|B\|_\sigma$ Submultiplikativität

**Weitere Vektor- und Matrixnormen**

p-Norm, $p \in \mathbb{N}^+, \vec{x} \in \mathbb{R}^n$

$$\|\vec{x}\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{\frac{1}{p}}$$

induziert $\|A\|_p$
Frobenius-Norm einer $(m \times n)$ Matrix

$$\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2\right)^{\frac{1}{2}}$$

also die euklidsche bzw. 2-Norm von A als Vektor geschrieben.

**Theorem 2.4.** *Für die Berechnung von speziellen induzierten Matrixnormen gilt (A reelle $(m \times n)$ Matrix)*

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}| \quad \textit{Spaltensummennorm} \tag{2.2}$$

$$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| \quad \textit{Zeilensummennorm} \tag{2.3}$$

$$\|A\|_2 = \sqrt{\lambda_{max}} \quad \textit{mit} \quad \lambda_{max} \textit{ größter EW von } A^T A \tag{2.4}$$

*Beweis.*
1) Für den Nachweis von (2.2) erhalten wir für $\vec{x} \in \mathbb{R}^n$

$$\begin{aligned} \|A\vec{x}\|_1 &= \sum_{k=1}^{m} |\sum_{j=1}^{n} a_{kj}x_j| \le \sum_{k=1}^{m}\sum_{j=1}^{n} |a_{kj}|\,|x_j| = \sum_{j=1}^{n}(\sum_{k=1}^{m}|a_{kj}|)|x_j| \\ &\le (\max_{j=1,\ldots,n}\sum_{k=1}^{m}|a_{kj}|)\sum_{j=1}^{n}|x_j| = (\max_{j=1,\ldots,n}\sum_{k=1}^{m}|a_{kj}|)\,\|\vec{x}\|_1 \ , \end{aligned}$$

also $\|A\|_1 \le \max_{j=1,\ldots,n}\sum_{k=1}^{m}|a_{kj}|$. Sei nun $l$ ein beliebiger, aber fester Index und $\vec{e}_l$ der kanonische Einheitsvektor mit einer 1 in der $l$-ten Komponente, dann ist $\|\vec{e}_l\|_1 = 1$ und damit gilt

$$\|A\|_1 \ge \|A\vec{e}_l\|_1 = \sum_{k=1}^{m}|\sum_{j=1}^{n}a_{kj}\delta_{jl}| = \sum_{k=1}^{m}|a_{kl}| \ . \tag{2.5}$$

Da $l$ beliebig gewählt werden kann, folgt die Gleichung (2.5) für alle Spalten der Matrix $A$, also gilt $\|A\|_1 \ge \max_{j=1,\ldots,n}\sum_{k=1}^{m}|a_{kj}|$ und damit (2.2).
2) Für den Nachweis von (2.3) gilt für $\vec{x} \in \mathbb{R}^n$

$$\|A\vec{x}\|_\infty = \max_{k=1,\ldots,m}|\sum_{j=1}^{n}a_{kj}x_j| \le \max_{k=1,\ldots,m}\sum_{j=1}^{n}|a_{kj}|\,|x_j| \le (\max_{k=1,\ldots,m}\sum_{j=1}^{n}|a_{kj}|)\,\|\vec{x}\|_\infty \ ,$$

also $\|A\|_\infty \le \max_{k=1,\ldots,m}\sum_{j=1}^{n}|a_{kj}|$. Nun sei $k$ ein beliebiger, aber fester Index. Für $\vec{x} = (x_j) \in \mathbb{R}^n$ mit

$$x_j = \begin{cases} \frac{|a_{kj}|}{a_{kj}}, & \text{falls } a_{kj} \neq 0, \\ 1, & \text{sonst,} \end{cases} \qquad (j = 1,\ldots,n)$$

gilt $\|\vec{x}\|_\infty = 1$ und damit

$$\|A\|_\infty \ge \frac{\|A\vec{x}\|_\infty}{\|\vec{x}\|_\infty} = \|A\vec{x}\|_\infty \ge |\sum_{j=1}^{n}\underbrace{a_{kj}x_j}_{=|a_{kj}|}| = \sum_{j=1}^{n}|a_{kj}| \ . \tag{2.6}$$

Da $k$ als Zeilenindex frei gewählt wurde, gilt (2.6) für alle Zeilen, also gilt $\|A\|_\infty \ge \max_{k=1,\ldots,m}\sum_{j=1}^{n}|a_{kj}|$ und damit ist (2.3) gezeigt.
Für den Nachweis von (2.4) überlegt man, dass $A^TA$ ähnlich einer Diagonalmatrix mit den EW von $A^TA$ als Diagonalelementen ist. Die EW sind nicht negativ und aus der Definition von $\|A\|_2$ folgt dann schließlich (2.4) □

**Definition 2.5.** *Unter dem Absolutbetrag einer Matrix $A \in \mathbb{C}^{m\times n}$ versteht man die Matrix*

$$|A| = B \quad mit \quad b_{ij} = |a_{ij}|$$

*also die Matrix mit den Absolutbeträgen ihrer Elemente. Gilt für $A, B \in \mathbb{R}^{m\times n}$ dass $a_{ij} \leq b_{ij}$ so schreibt man $A \leq B$*

**Bemerkung 2.6.** Für die "Beträge"von Matrizen gelten die Beziehungen

(i) $|A + B| \leq |A| + |B|$

(ii) $|A \cdot B| \leq |A||B|$

(iii) $A \leq B, C \geq 0, D \geq 0 \Rightarrow CAD \leq CBD$

(iv) $\|A\|_F \leq \||A|\|_p, \quad p \in \mathbb{N}^+$

(v) $\|A\|_p = \||A|\|_p \quad$ für $\quad p \in \{1, \infty, F\}$

(vi) $|A| \leq |B| \Rightarrow \|A\| \leq \|B\| \quad$ für diese Nomen

*Beweis.* Größtenteils trivial □

Die normweise Kondition eines Problems liefert oft eine recht grobe Abschätzung. Im Falle der genügenden Glattheit eines Problems $f : \mathbb{R}^n \to \mathbb{R}$ kann man aus der linearen Approximation

$$f(\tilde{x}) \doteq f(x) + \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(x)(\tilde{x}_j - x_j) \quad \text{für } \tilde{x} \to x$$

die Beziehungen

$$\begin{aligned} |f(\tilde{x}) - f(x)| &\dot{\leq} \sum_{j=1}^{n} |\frac{\partial f}{\partial x_j}(x)|\,|\tilde{x}_j - x_j| \\ &\dot{\leq} (\sum_{j=1}^{n} |\frac{\partial f}{\partial x_j}(x)|) \max_{j=1\ldots n} |\tilde{x}_j - x_j| \quad \text{für } \tilde{x} \to x \end{aligned} \tag{2.7}$$

bzw.

$$\frac{|f(\tilde{x}) - f(x)|}{|f(x)|} \dot{\leq} \sum_{j=1}^{n} \frac{|\frac{\partial f}{\partial x_j}(x)|\,|x_j|}{|f(x)|} \max_{j=1\ldots n} \frac{|\tilde{x}_j - x_j|}{|x_j|} \quad \text{für } \tilde{x} \to x\,. \tag{2.8}$$

Den Verstärkungsfaktor des max. relativen Fehlers

$$\kappa_{rel} = \sum_{j=1}^{n} \frac{|\frac{\partial f}{\partial x_j}(x)|\,|x_j|}{|f(x)|} =: \frac{|f'(x)| \cdot |x|}{|f(x)|} \quad \text{für } \tilde{x} \to x$$

bezeichnet man als relative **komponentenweise** Kondition (die "Beträge" der Matrizen $f'(x)$ bzw. $x$ sind dabei komponentenweise zu verstehen). Allgemein erklärt man die komponentenweise Kondition eines Problems $f$ als die kleinste Zahl $\kappa_{rel} \geq 0$, so dass

$$\frac{|f(\tilde{x}) - f(x)|}{|f(x)|} \dot{\leq} \kappa_{rel} \max_{j=1\ldots n} \frac{|\tilde{x}_j - x_j|}{|x_j|} \quad \text{für } \tilde{x} \to x$$

gilt.

**Beispiel 2.7.** Für die Multiplikation zweier Zahlen $f(x, y) = x\,y$ erhält man $f'(x, y) = [y \;\; x]$ und damit folgt für die relativnee komponentenweise Kondition

$$\kappa_{rel} = \frac{[|y| \;\; |x|]\binom{|x|}{|y|}}{|x\,y|} = \frac{|x\,y| + |x\,y|}{|x\,y|} = 2$$

Im Unterschied dazu findet man für die relative normweise Kondition mit der 1-Norm für $|x| \neq |y|$, wobei o.B.d.A. $|x| > |y|$ angenommen wurde,

$$\kappa_{\text{rel}} = \frac{|x| + |y|}{|y|} \;,$$

d.h. die absolute normweise Kondition kann sehr groß sein.

**Beispiel 2.8.** (Kondition eines linearen Gleichungssystems $Ax = b$, $A$ regulär) Die Lösung $x$ kann man durch die Abbildung

$$f : \mathbb{R}^n \to \mathbb{R}^n, \quad b \mapsto f(b) = A^{-1}b$$

beschreiben. Die Ableitung von $f$ ergibt sich hier im Falle der linearen Abb. zu $f' = A^{-1}$. Für die normweise Kondition erhält man demzufolge

$$\kappa_{abs} = \left\|A^{-1}\right\| \quad \text{und} \quad \kappa_{rel} = \frac{\|b\|}{\|A^{-1}b\|} \left\|A^{-1}\right\| = \frac{\|Ax\|}{\|x\|} \left\|A^{-1}\right\| \leq \|A\| \left\|A^{-1}\right\| \;.$$

Die hierbei erhaltene Schranke für die relative Kondition definiert man als **Konditionszahl**

$$\kappa(A) = \|A\| \left\|A^{-1}\right\|$$

bezüglich einer verträglichen Norm.

## 2.3 Stabilitätskonzepte

Wir wollen 2 Stabilitätskonzepte betrachten. Einmal geht es um die mögliche Auswirkung von Eingabefehlern auf die fehlerhaften Endergebnisse von Algorithmen. Man spricht hier von der sogenannten **Vorwärtsanalyse**, bei der die Kondition und unvermeidbare Fehler von Bedeutung sind.
Bei der sogenannten **Rückwärtsanalyse** interpretiert man ein fehlerhaftes Endergebnis eines Problems $(f, x)$, d.h. $\tilde{y} = \tilde{f}(\tilde{x})$ als exaktes Ergebnis einer gestörten Eingabe $\hat{x}$, d.h. $\tilde{y} = f(\hat{x})$ und falls es mehrere solcher Größen mit $f(\hat{x}) = \tilde{y} = f(\hat{x})$ gibt, wählt man dasjenige mit dem geringsten Abstand zu $\tilde{x}$.

### 2.3.1 Vorwärtsanalyse

Der unvermeidbare Fehler eines Algorithmus lässt sich durch das Produkt der Kondition und des Eingabefehlers, also $\kappa\, eps$ abschätzen. Um Stabilität von Algorithmen bewerten zu können, wird ein **Stabilitätsindikator** $\sigma$ eingeführt, der als Faktor den unvermeidbaren Fehler $\kappa\, eps$ verstärkt.

**Definition 2.9.** *Sei $(f, x)$ ein Problem mit der normweisen relativen Kondition $\kappa_{rel}$ und der Gleitkommarealisierung $\tilde{f}$. Der* **Stabilitätsindikator** *der Vorwärtsanalyse ist die kleinstmögliche Zahl $\sigma \geq 0$, so dass f.a. móglichen Eingabegrößen $\tilde{x}$ gilt*

$$\frac{\left\| \tilde{f}(\tilde{x}) - f(\tilde{x}) \right\|}{\|f(\tilde{x})\|} \dot{\leq} \sigma \kappa_{rel} eps \quad \text{für } eps \to 0 \,. \tag{2.9}$$

*Ein Algorithmus $\tilde{f}$ wird* **stabil** *genannt, wenn $\sigma$ kleiner als die Anzahl der hindereinander ausgeführten Elementaroperationen (Addition, Subtraktion, Multiplikation etc.) ist.*

**Bemerkung 2.10.** Die Elementaroperationen sind stabil, da

$$\sigma\kappa \leq 1$$

gilt, was als Übung nachgewiesen werden sollte.

Wir hatten bereits in einer vergangenen Vorlesung festgestellt, dass es bei Algithmen, die aus mehreren vertauschberen Teilschritten bestehen, mitunter sinnvoll ist mit bestimmten Teilschritten zu beginnen.
Betrachten wir nun ein Problem $(f, x)$, das in 2 Teilprobleme $(g, x)$ und

$(h, g(x)$ (verketteter Algorithmus) aufgeteilt werden kann, d.h. man hat im skalaren Fall

$$f = h \circ g, \quad g : \mathbb{R} \to \mathbb{R},\ h : \mathbb{R} \to \mathbb{R} .$$

Die Stabilitätsindikatoren $\sigma_g$ und $\sigma_h$ seien für die Teilalgorithmen $\tilde{g}$ und $\tilde{h}$ bekannt. Der folgende Satz gibt Auskunft über die Stabilität des verketteten Algorithmus $\tilde{f} = \tilde{h} \circ \tilde{g}$.

**Theorem 2.11.**
*Seien $\kappa_f$, $\kappa_h$, $\kappa_g$ die normweisen relativen Konditionen der Algorithmen $f$, $h$, $g$. Für den Stabilitätsindikator $\sigma_f$ des verketteten Algorithmus $\tilde{f}$ gilt*

$$\sigma_f \kappa_f \leq \sigma_h \kappa_h + \sigma_g \kappa_g \kappa_h .$$

*Beweis.* Es gilt (verwenden $x = \tilde{x}$)

$$\begin{aligned}
\left\| \tilde{f}(x) - f(x) \right\| &= \left\| \tilde{h}(\tilde{g}(x)) - h(g(x)) \right\| \\
&\dot{\leq} \left\| \tilde{h}(\tilde{g}(x)) - h(\tilde{g}(x)) \right\| + \| h(\tilde{g}(x)) - h(g(x)) \| \\
&\dot{\leq} \sigma_h \kappa_h eps\ \| h(\tilde{g}(x)) \| + \kappa_h \frac{\| \tilde{g}(x) - g(x) \|}{\| g(x) \|} \| h(g(x)) \| \\
&\dot{\leq} \sigma_h \kappa_h eps\ \| h(\tilde{g}(x)) \| + \kappa_h \sigma_g \kappa_g eps\ \| h(g(x)) \| \\
&\dot{\leq} (\sigma_h \kappa_h + \sigma_g \kappa_h \kappa_g) eps\, \| f(x) \| \ .
\end{aligned}$$

□

Eine Schlussfolgerung aus diesem Satz ist dann, dass man unbedingt erforderliche Subtraktionen (Auslöschungsproblematik) als Teilprobleme eines verketteten Gesamtproblems möglichst zu Beginn eines Algorithmus ausführt. Sind $f$, $g$, $h$ skalare Funktionen, dann ergibt sich für den Stabilitätsindikator $\sigma_f$ des Algorithmus $\tilde{f} = \tilde{h} \circ \tilde{g}$ direkt aus Satz 2.11 die Beziehung

$$\sigma_f \leq \frac{\sigma_h}{\kappa_g} + \sigma_g ,$$

denn es gilt offensichtlich

$$\kappa_f = \frac{|f'(x)|\,|x|}{|f(x)|} = \frac{|g(x)|\,|h'(g(x))|\,|g'(x)|\,|x|}{|h(g(x))|\,|g(x)|} = \kappa_h \kappa_g .$$

### 2.3.2 Rückwärtsanalyse

**Definition 2.12.** *Der normweise* **Rückwärtsfehler** *des Algorithmus $\tilde{f}$ zur Lösung des Problems $(f, x)$ ist die kleinste Zahl $\eta \geq 0$, für die für alle möglichen Eingaben $\tilde{x}$ ein $\hat{x}$ mit $\tilde{f}(\tilde{x}) = f(\hat{x})$ existiert, so dass*

$$\frac{\|\hat{x} - \tilde{x}\|}{\|\tilde{x}\|} \dot{\leq} \eta \quad \textit{für } \text{eps} \to 0 \ .$$

*Der Algorithmus heißt* **stabil** *bezüglich des relativen Eingabefehlers $\delta$, falls*

$$\eta < \delta$$

*gilt. Für den durch Rundung verursachten Eingabefehler $\delta = eps$ wird durch den Quotienten*

$$\sigma_R := \frac{\eta}{eps}$$

*der* **Stabilitätsindikator der Rückwärtsanalyse** *definiert.*

**Theorem 2.13.** *Für die Stabilitätsindikatoren $\sigma$ und $\sigma_R$ der Vorwärts- bzw. Rückwärtsanalyse gilt*

$$\sigma \leq \sigma_R$$

*(aus der Rückwärtsstabilität folgt die Vorwärtsstabilität).*

*Beweis.* Aus der Definition des Rückwärtsfehlers folgt $f(\hat{x}) = \tilde{f}(\tilde{x})$ und

$$\frac{\|\hat{x} - \tilde{x}\|}{\|\tilde{x}\|} \dot{\leq} \eta = \sigma_R eps \quad \text{für } \text{eps} \to 0 \ .$$

Damit ergibt sich mit

$$\frac{\left\|\tilde{f}(\tilde{x}) - f(\tilde{x})\right\|}{\|f(\tilde{x})\|} = \frac{\|f(\hat{x}) - f(\tilde{x})\|}{\|f(\tilde{x})\|} \dot{\leq} \kappa \frac{\|\hat{x} - \tilde{x}\|}{\|\tilde{x}\|} \dot{\leq} \kappa \sigma_R eps$$

für $eps \to 0$. □

# Kapitel 3

# Lösung linearer Gleichungssysteme

## 3.1 LR-Zerlegung

4. Vorlesung am 26.10.2011

Zu lösen ist $Ax = b$. Dazu soll $A$ als Produkt einer unteren Dreiecksmatrix $L$ und einer oberen Dreiecksmatrix $R$ geschrieben werden, d.h. man hat

$$Ax = b \Leftrightarrow L \underbrace{Rx}_{y} = b$$

und löst zuerst

$$Ly = b$$

und danach

$$Rx = y$$

**Beispiel 3.1.** (praktische Konstruktion einer $LR$-Zerlegung)
Betrachten wir die Matrix

$$A = \begin{pmatrix} 2 & 1 & 3 \\ 4 & 3 & 11 \\ 6 & 5 & 23 \end{pmatrix}$$

Mit dem Gaußschen Algorithmus erhält man nun in den ersten beiden Schritten durch eine geeignete Linearkombination der 1. mit der 2. und 3. Zeile

$$A^{(1)} = \begin{pmatrix} 2 & 1 & 3 \\ 0 & 1 & 5 \\ 0 & 2 & 14 \end{pmatrix}$$

Matrix-technisch gesehen wurde dabei die Matrix $A$ mit der Matrix

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ -4/2 & 1 & 0 \\ -6/2 & 0 & 1 \end{pmatrix}$$

multipliziert, also gilt $A^{(1)} = M_1 A$. Im nächsten Schritt des Gaußschen Algorithmus' erhält man durch eine weitere Linearkombination der 2. mit der 3. Zeile

$$A^{(2)} = \begin{pmatrix} 2 & 1 & 3 \\ 0 & 1 & 5 \\ 0 & 0 & 4 \end{pmatrix}$$

$A^{(2)}$ erhält man dabei aus $A^{(1)}$ durch

$$A^{(2)} = M_2 M_1 A^{(1)}$$

wobei

$$M_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2/1 & 1 \end{pmatrix}$$

gilt. Wir haben also die Gleichung

$$M_2 M_1 A = A^{(2)} := R \tag{3.1}$$

erhalten. Die Matrizen $M_1$ und $M_2$ lassen sich sehr einfach invertieren, denn es gilt

$$M_1^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 4/2 & 1 & 0 \\ 6/2 & 0 & 1 \end{pmatrix} \qquad \text{und} \qquad M_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 2/1 & 1 \end{pmatrix} ,$$

d.h. man muss bloß die Vorzeichen der Nichtdiagonalelemente ändern. Letztendlich erhält man durch die sukzessive Multiplikation der Gleichung (3.1) mit $M_2^{-1}$ und [$M_1^{-1}$ die Gleichung

$$A = M_1^{-1} M_2^{-1} R ,$$

wobei

$$L = M_1^{-1} M_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 4/2 & 1 & 0 \\ 6/2 & 2/1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 2 & 1 \end{pmatrix} ,$$

die gewünschte untere Dreiecksmatrix ist.

### 3.1.1 Realisierung mit dem Gaußschen Eliminationsverfahren

Grundprinzip:
Rangerhaltende Manipulationen der Matrix $[A|b]$ durch Linearkombinationen von Zeilen

$$L_{ij}(\lambda) = \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ & & \lambda & \ddots & \\ 0 & & & & 1 \end{pmatrix}$$

Multiplikation $L_{ij}(\lambda)A$ bewirkt die Addition des $\lambda$-fachen der j-ten Zeile von $A$ zur i-ten Zeile d.h. durch geeignete Wahl von $\lambda$ erzeugt man in

$$\tilde{A} = L_{ij}(\lambda)A$$

an der Position $(i, j)$ z.B. auch eine Null ($A \in \mathbb{R}^{n\times m}$)
$L_{ij}(\lambda)$ hat den Rang $n$ und die Determinante $1 \Rightarrow \operatorname{rg}(\tilde{A}) = \operatorname{rg}(A)$. Durch mehrfache Multiplikation mit

$$L_{jk}, \quad j = k+1, ..., n$$

erhält man bei geeigneter Wahl der $\lambda$ unterhalb von $\tilde{a}_{kk}$ Null-Einträge

**Nun etwas präziser**

$$A = \begin{pmatrix} a_{11} & \cdots & \cdots & a_{1n} \\ & \ddots & & \\ & & a_{kk} & \cdots \\ & & \vdots & \\ & & a_{nk} & \cdots \end{pmatrix} \rightarrow \begin{pmatrix} a_{11} & & \cdots & \\ & \ddots & & \\ & & a_{kk} & \cdots \\ & & 0 & \tilde{a}_{k+1k+1} \\ & & \vdots & \\ & & 0 & \end{pmatrix}$$

Vorraussetzung: $a_{kk} \neq 0$
Setzen

$$t = t^{(k)}(a_k) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ t_{k+1\,k} \\ \vdots \\ t_{nk} \end{pmatrix}$$

mit

$$t_{ik} = \begin{cases} 0 & i = 1, .., k \\ \frac{a_{ik}}{a_{kk}} & i = k+1, ..., n \end{cases}$$

$e_k$ sei der k-te Standardeinheitsvektor

**Definition 3.2.**

$$M_k := \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -t_{k+1k} & \ddots & \\ & & \vdots & & \ddots \\ & & -t_{nk} & & 1 \end{pmatrix} \quad \textit{Frobenius-Matrix, Gaußtrafomatrix}$$

*Man überlegt sich, dass $M_k$ das Produkt der oben diskutierten Matrizen $L_{jk}(-t_j), j = k+1, ..., n$ ist.*

Eigenschaften von $M_k$

$$M_k = E - t^{(k)} e_k^T$$

$$M_k a_k = \begin{bmatrix} a_{k1} \\ \vdots \\ a_{kk} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

d.h. $a_{k1}, ..., a_{kk}$ bleiben bei der Multiplikation mit $M_k$ unverändert

$$\operatorname{rg}(M_k) = n$$
$$\det(M_k) = 1$$
$$M_k^{-1} = E + t^{(k)} e_k^T, \quad da$$
$$M_k^{-1} M_k = E - t^{(k)} \underbrace{e_k^T t^{(k)}}_{=0} e_k^T = E$$

Wenn alles gut geht, d.h. wenn jeweils $\tilde{a}_{kk} \neq 0$ ist, dann erhält man nach der Mutiplikation von $A$ mit den Frobenius-Matrizen $M_1, ..., M_{n-1}$, also

$$M_{n-1} \cdot ... \cdot M_1 A = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ & & \ddots & \vdots \\ 0 & & & r_{nn} \end{bmatrix} =: R$$

eine obere Dreiecksmatrix R. Außerdem hat die Matrix

$$M_{n-1} \cdot ... \cdot M_1$$

die inverse Matrix

$$L = M_1^{-1} \cdot ... \cdot M_{n-1}^{-1} = \begin{bmatrix} 1 & & & & 0 \\ t_{21} & 1 & & & \\ t_{31} & t_{32} & 1 & & \\ \vdots & \vdots & & \ddots & \\ t_{n1} & t_{n2} & & t_{nn-1} & 1 \end{bmatrix}$$

sodass schließlich mit

$$A = LR$$

eine LR-Zerlegung vorliegt.

**Definition 3.3.** *Eine obere oder untere Dreiecksmatrix, deren Diagonalelemente alle gleich eins sind, heißt* ***unipotent****. Die Zerlegung $A = LR$ heißt LR-Zerlegung, wenn $L$ eine unipotente untere Dreiecksmatrix ist.*

**Satz 3.4.** *Eine Matrix $A \in \mathbb{R}^{n \times n}$ besitzt genau dann eine LR-Zerlegung, wenn*

$$\det(A(1:k, 1:k)) \neq 0, \quad k = 1, 2, ..., n-1$$

*Falls die LR-Zerlegung existiert und $A$ regulär ist, dann sind $L$ und $R$ eindeutig bestimmt, und es gilt:*

$$\det A = r_{11} \cdot r_{22} \cdot ... \cdot r_{nn} \ .$$

*Bemerkung: Wir wollen hier die LR-Zerlegung so verstehen, dass der oben beschriebene Algorithmus mit den Frobeniusmatrizen $M_k$, $k = 1, \ldots, n-1$, erfolgreich ist und nicht wegen $a_{kk} = 0$ abbricht.*

*Beweis.* a) $A$ besitze LR-Zerlegung

$$A = \begin{bmatrix} 1 & & & & 0 \\ l_{21} & 1 & & & \\ l_{31} & l_{32} & 1 & & \\ \vdots & \vdots & & \ddots & \\ l_{n1} & l_{n2} & & l_{nn-1} & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ & & \ddots & \vdots \\ 0 & & & r_{nn} \end{bmatrix}$$

Die $r_{jj}$ sind die sogenannten Pivots, durch die in den Schritten $1, ..., n-1$ dividiert werden musste, d.h. $r_{jj} \neq 0, \quad j = 1, ..., n-1$

$$\Rightarrow \qquad \underbrace{\det(L(1:k, 1:k))}_{=1} \cdot \underbrace{\det(R(1:k, 1:k))}_{=\prod_{j=1}^{k} r_{jj}, 1 \leq k \leq n-1} = \det(A(1:k, 1:k)) \neq 0$$

b) Es gelte $\det(A(1:k,1:k)) \neq 0, \quad k = 1,2,...,n-1$ Induktion über k

<u>k=1</u>: nach Voraussetzung ist $a_{11} = \det(A(1:1,1:1)) \neq 0$ d.h. erster Schritt ist möglich

Nun seien $k-1$ Schritte allgemein ausgeführt, und wir zeigen, dass auch Schritt k ausgeführt werden kann

<u>$k-1 \to k$</u> Schritt ist möglich, falls Pivot $a_{kk}^{(k-1)} \neq 0$. Es sei $A^{(k-1)} = M_{k-1} \cdot ... \cdot M_1 A$

$$M_{k-1} \cdot \ldots \cdot M_1 A = \begin{bmatrix} 1 & & & & & & \\ * & \ddots & & & \mathbf{0} & & \\ & * & 1 & & & & \\ & & * & 1 & & 0 & \\ & & & & \ddots & & \\ * & & * & 0 & & 1 & \end{bmatrix} A$$

$$= \begin{bmatrix} a_{11}^{(k-1)} & & & & & & \\ \vdots & \ddots & & & & & \\ 0 & \ldots & a_{k-1k-1}^{(k-1)} & & & & \\ 0 & \ldots & 0 & a_{kk}^{(k-1)} & \cdots & * & \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \\ 0 & \ldots & 0 & * & \cdots & * & \end{bmatrix} = A^{(k-1)}$$

$$\det(A^{(k-1)}(1:k,1:k)) = \prod_{j=1}^{k} a_{jj}^{(k-1)}$$

und

$$\det(A^{(k-1)}(1:k,1:k)) = \underbrace{\det(M^{(k-1)}(1:k,1:k))}_{=1} \cdot \underbrace{\det(A(1:k,1:k))}_{\neq 0, \text{ n.V.}} \neq 0$$

Damit muss $\prod_{j=1}^{k} a_{jj}^{(k-1)} \neq 0$ gelten, weshalb $a_{kk}^{(k-1)} \neq 0$ sein muss. D.h. ein weiterer Schritt ist möglich.

Nachweis der Eindeutigkeit der Zerlegung.
Existiere $A^{-1}$ und sei $A = L_1 R_1 = L_2 R_2$, wegen $\det(A) \neq 0$ sind auch $R_1, R_2$ regulär $\Rightarrow L_2^{-1} L_1 = R_2 R_1^{-1}$. Die Inverse einer unteren Dreiecksmatrix mit Diagonale 1 ist wieder unipotent und eine untere Dreiecksmatrix, die einer Oberen ist wieder eine Obere. Damit ist

$$L_2^{-1} L_1 = E = R_2 R_1^{-1} \Rightarrow L_1 = L_2, R_1 = R_2 \wedge \det A = \det R$$

□

**Bemerkung.** (1) Man braucht nur den Speicherplatz der Matrix:

Die obere Dreiecksmatrix entsteht durch die sukzessive Multiplikation von $A$ mit Frobenius-Matrizen (Gauß-Transformationen)

$$\begin{bmatrix} r_{11} & \cdots & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ 0 & & \ddots & \vdots \\ & & & r_{nn} \end{bmatrix}$$

An den Positionen $(k+1,k),(k+2,k),...,(n,k)$ wo durch die Gauß-Transformationen (Multiplikation mit $M_k$) Nullen erzeugt werden, können die Elemente $t_{k+1k}, t_{k+2k}, ..., t_{nk}$ sukzessiv für $k = 1, ..., n-2$ eingetragen werden und man erhält

$$\begin{bmatrix} t_{21} & & & \\ t_{31} & t_{32} & & \\ \vdots & & \ddots & \\ t_{n1} & & & t_{nn-1} \end{bmatrix}$$

also die nicht redundanten Elemente von $L$

(2) Berechnung von $L = M_1^{-1} \cdot ... \cdot M_{n-1}^{-1}$ kostet nichts, sondern besteht nur in der Ablage der jeweils bei den Gauß-Transformationen erzeugten $t_{kj}$-Werten $(k > j,\ k = 2, ..., n,\ j = 1, ..., n-1)$

(3) Rechenaufwand ca. $\frac{n^3}{3} \in \mathcal{O}(n^3)$ Multiplikationen (flops, floating point operations).

### Fehleranalyse bei der Konstruktion einer LR-Zerlegung

**Satz 3.5.** *Sei $A \in \mathbb{R}^{n\times n}$ Matrix von Maschinenzahlen. Falls bei der Konstruktion der LR-Zerlegung kein $\tilde{a}_{kk} = 0$ zum Abbruch führt, dann erfüllen die berechneten Faktoren $\tilde{L}, \tilde{R}$ die Gleichung*

$$\tilde{L}\tilde{R} = A + H$$

*mit*

$$|H| \leq 3(n-1)\text{eps}(|A| + |\tilde{L}||\tilde{R}|) + \mathcal{O}(\text{eps}^2)$$

*Beweis.* siehe Bollhöfer/Mehrmann □

**Satz 3.6.** *Sind $\tilde{L}, \tilde{R}$ die Matrizen aus Satz 3.5, so erhält man bei den Algorithmen zum Vorwärts- und Rückwärtseinsetzen*

$$\tilde{L}\tilde{y} = b, \quad \tilde{R}\tilde{x} = \tilde{y}$$

*eine Lösung $\tilde{x}$ von $(A + \Delta)\tilde{x} = b$ mit*

$$|\Delta| \leq n \cdot \text{eps}(3|A| + 5|\tilde{L}||\tilde{R}|) + \mathcal{O}(\text{eps}^2)$$

*Beweis.* Rückwärtseinsetzen ergibt

$$\begin{aligned}
&(\tilde{L} + F)\tilde{y} = b, |F| \leq n \cdot \text{eps}|\tilde{L}| + \mathcal{O}(\text{eps}^2) \\
&(\tilde{R} + G)\tilde{x} = \tilde{y}, |G| \leq n \cdot \text{eps}|\tilde{R}| + \mathcal{O}(\text{eps}^2) \\
\Rightarrow\ &(\tilde{L} + F)(\tilde{R} + G)\tilde{x} = b \\
\Leftrightarrow\ &(\underbrace{\tilde{L}\tilde{R}}_{A+H} + F\tilde{R} + \tilde{L}G + FG)\tilde{x} = b \\
\Leftrightarrow\ &(A + \Delta)\tilde{x} = b
\end{aligned}$$

mit $\Delta = H + F\tilde{R} + \tilde{L}G + FG$. Mit der Abschätzung aus Satz 3.5 für $H$ ergibt sich

$$\begin{aligned}
|\Delta| &\leq |H| + \underbrace{|F|}_{\leq n\text{eps}|\tilde{L}|}|\tilde{R}| + |\tilde{L}|\underbrace{|G|}_{\leq n\text{eps}|\tilde{R}|} + \underbrace{|F||G|}_{\mathcal{O}(\text{eps}^2} \\
&\leq 3(n-1)\text{eps}(|A| + |\tilde{L}||\tilde{R}|) + 2n\text{eps}|\tilde{L}||\tilde{R}| + \mathcal{O}(\text{eps}^2) \\
&\leq n\text{eps}(3|A| + 5|\tilde{L}||\tilde{R}|) + \mathcal{O}(\text{eps}^2)
\end{aligned}$$

□

**Bemerkung.** Problematisch, d.h. recht groß können die Elemente von $|\tilde{L}|$ und $|\tilde{R}|$ werden, wenn bei der Berechnung aller $t_{kj}$ im Rahmen der Gauß-Transformationen große Zahlen entstehen!
Abhilfe: Pivotisierung

### 3.1.2 LR-Zerlegung mit Spaltenpivotisierung

Um zu vermeiden, dass der Algorithmus zur Konstruktion einer LR-Zerlegung aufgrund von $\tilde{a}_{kk} = 0$ abbricht, oder durch betragsmäßig sehr kleine $\tilde{a}_{kk}$ (kleine Pivots) bei der Berechnung der $t_{kj}$ betragsmäßig sehr große Zahlen entstehen, kann man durch Zeilenvertauschungen das betragsmäßig maximale Element in die Diagonalposition bringen.
Zeilenvertauschungen bewirkt man durch Multiplikation mit Permutationsmatrizen $P_k$ (von links).

**Definition 3.7.** *Matrizen $P \in \mathbb{R}^{n\times n}$ die aus der Einheitsmatrix durch Vertauschen von (genau) zwei Zeilen hervorgehen heißen* **elementare Permutationsmatrizen**

Bei den durchgeführten Betrachtungen haben wir benutzt, dass für elementare Permutationsmatrizen

$$P \cdot P = E$$

gilt, d.h. die Matrix gleich ihrer Inversen ist. Die Erfahrungen des Beispiels kann man zusammenfassen.

**Definition 3.8.** *Wir bezeichnen den im Beispiel beschriebenen Algorithmus als Konstruktion einer LR-Zerlegung mit Spaltenpivotisierung (auch Gaußelimination mit partieller Pivotisierung).*

**Satz 3.9.** *Für die Gaußelimination mit partieller Pivotisierung mit dem Resultat*

$$M_{n-1}P_{n-1} \cdot \ldots \cdot M_1 P_1 A = R$$

*gilt $PA = LR$ mit $P = P_{n-1} \cdot \ldots \cdot P_1$. Für $L$ gilt*

$$L = \hat{M}_1^{-1} \cdot \ldots \cdot \hat{M}_{n-1}^{-1}$$

*mit*

$$\begin{aligned} \hat{M}_{n-1} &= M_{n-1} \\ \hat{M}_k &= P_{n-1} \cdot \ldots \cdot P_{k+1} M_k P_{k+1} \cdot \ldots \cdot P_{n-1}, \quad k \leq n-2 \end{aligned}$$

*wobei $\hat{M}_k$ Frobeniusmatrizen sind (deren Inverse trivial zu berechnen ist).*

*Beweis.* Durch die Eigenschaft $PP = E$ von elementaren Permutationsmatrizen überlegt man sich, dass

$$\begin{aligned} & M_{n-1}P_{n-1}M_{n-2}P_{n-2}\cdots M_1P_1A \\ = \; & \underbrace{M_{n-1}}_{\hat{M}_{n-1}} \underbrace{P_{n-1}M_{n-2}P_{n-1}}_{\hat{M}_{n-2}} P_{n-1}P_{n-2}\cdots\cdots \underbrace{M_1P_2\cdots P_{n-1}}_{\hat{M}_1} \underbrace{P_{n-1}\cdots P_2P_1}_{P} A \end{aligned}$$

gilt. Außerdem hat $\hat{M}_k = P_\mu M_k P_\mu$ die gleiche Struktur wie $M_k$, da durch die Multiplikation von $P_\mu$ von links und rechts nur die Reihenfolge der $t_{kl}$ vertauscht wird. Die Multiplikation von

$$\hat{M}_{n-1}\hat{M}_{n-2}\cdots\hat{M}_1 PA \quad \text{mit} \quad L = \hat{M}_1^{-1}\cdots\hat{M}_{n-1}^{-1}$$

ergibt

$$PA = LR$$

Dabei ist $L$ ebenso wie im Fall der LR-Zerlegung ohne Pivotisierung als Produkt von Frobeniusmatrizen eine untere Dreiecksmatrix mit Diagonalelementen gleich eins.

□

**Bemerkung.** Konsequenz dieser LR-Zerlegung mit Spaltenpivotisierung ist, dass $\left|\tilde{L}\right|$ in der Regel wesentlich kleinere Elemente ($\leq 1$) hat, was zu einer Verbesserung der Abschätzung aus Satz 3.6 führt.

5. Vorlesung am 31.10.2011

## 3.2 Cholesky-Zerlegung

Bei vielen Aufgabenstellungen der angewandten Mathematik sind Gleichungssysteme $Ax = b$ mit symmetrischen und positiv definiten Matrizen $A$ zu lösen, z.B.

- numerische Lösung elliptischer und parabolischer Differentialgleichungen
- Spline-Approximation

Voraussetzung: $A \in \mathbb{R}^{n\times n}$ ist positiv definit und symmetrisch, d.h.

$$\forall x \neq 0 : x^T Ax > 0 \quad \text{und} \quad A = A^T$$

Unter diesen Voraussetzungen kann man die Gauß-Elimination (LR-Zerlegung) durch die sogenannte Cholesky-Zerlegung ersetzen und verbessern!

**Satz** (von Sylvester)**.** *Notwendig und hinreichend für positive Definitheit einer symmetrischen Matrix $A \in \mathbb{R}^{n\times n}$ ist die Positivität aller Hauptabschnittsdeterminanten, d.h.*

$$\forall k = 1, \ldots, n : \det A(1:k, 1:k) > 0$$

*(auch Kriterium von Hurwitz)*

**Satz 3.10.** *Sei $A$ symmetrisch und positiv definit. Dann existiert eine untere Dreiecksmatrix $G \in \mathbb{R}^{n\times n}$ mit positiven Diagonalelementen, sodass*

$$A = GG^T$$

*Beweis.* Nach dem Satz von Sylvester gilt $A(1:k, 1:k), k = 1, \ldots, n$ sind positiv definit und $\det A(1:k, 1:k) \neq 0$ sowie $A$ invertierbar (regulär) $\Rightarrow$ nach Satz 3.4 (Existenz eine LR-Zerlegung)

$$A = LR$$

mit $L$ untere Dreiecksmatrix mit 1-Diagonale und $R$ obere Dreiecksmatrix, in diesem Fall gilt

$$\begin{aligned}
&A(1:k, 1:k) = L(1:k, 1:k)R(1:k, 1:k) \\
\Rightarrow\ &0 < \det A(1:k, 1:k) = \underbrace{\det L(1:k, 1:k)}_{=1} \underbrace{\det R(1:k, 1:k)}_{=r_{11}r_{22}\cdots r_{kk}} \\
\Rightarrow\ &0 < \det R(1:k, 1:k) = r_{11}r_{22}\cdots r_{kk} \text{ für alle } k = 1, \ldots, n \\
\Rightarrow\ &\forall j = 1, \ldots, n : r_{jj} > 0
\end{aligned}$$

Nun betrachten wir die Diagonalmatrix

$$D = \operatorname{diag}(r_{11}, \ldots, r_{nn}) =: \operatorname{diag}(d_1, \ldots, d_n), d_k > 0$$

und es gilt

$$R = D\hat{R}$$

mit $\hat{r}_{jj} = 1, j = 1, \ldots, n$. Definiere $D^{\frac{1}{2}} = \operatorname{diag}(d_1^{\frac{1}{2}}, \ldots, d_n^{\frac{1}{2}})$

$$\begin{aligned} &\Rightarrow A = LR = LD\hat{R} = LD^{\frac{1}{2}}D^{\frac{1}{2}}\hat{R} \\ &\Rightarrow D^{-\frac{1}{2}}L^{-1}A = D^{\frac{1}{2}}\hat{R}, \quad D^{-\frac{1}{2}} = \operatorname{diag}(d_1^{-\frac{1}{2}}, \ldots, d_n^{-\frac{1}{2}}) \end{aligned} \tag{3.2}$$

Weiterhin gilt

$$\underbrace{D^{-\frac{1}{2}}L^{-1}A(L^{-1})^T(D^{-\frac{1}{2}})^T}_{\text{symmetrisch}} = \underbrace{D^{\frac{1}{2}}\hat{R}(L^{-1})^T(D^{-\frac{1}{2}})^T}_{\text{obere Dreiecksmatrix mit 1-Diagonale}}$$

$$\begin{aligned} &\Rightarrow D^{\frac{1}{2}}\hat{R}(L^{-1})^T D^{-\frac{1}{2}} = E \\ &\Rightarrow \hat{R}(L^{-1})^T = D^{-\frac{1}{2}}D^{\frac{1}{2}} = E \\ &\Rightarrow \hat{R} = L^T \end{aligned}$$

Einsetzen in (3.2) ergibt

$$\begin{aligned} A &= LD^{\frac{1}{2}}D^{\frac{1}{2}}\hat{R} = LD^{\frac{1}{2}}D^{\frac{1}{2}}L^T \\ &= (LD^{\frac{1}{2}})(LD^{\frac{1}{2}})^T \\ \Rightarrow\ G &= LD^{\frac{1}{2}} \end{aligned}$$

□

**Konstruktion der Choleksy-Zerlegung**

$$GG^T = A \Leftrightarrow \begin{bmatrix} g_{11} & & 0 \\ \vdots & \ddots & \\ g_{n1} & \cdots & g_{nn} \end{bmatrix} \begin{bmatrix} g_{11} & \cdots & g_{n1} \\ & \ddots & \vdots \\ & & g_{nn} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$$

$$\Rightarrow\ a_{kk} = g_{k1}^2 + g_{k2}^2 + \cdots + g_{kk-1}^2 + g_{kk}^2, k = 1, \ldots, n$$

$$\Rightarrow\ k = 1 :\ g_{11}^2 = a_{11}\ \Rightarrow\ g_{11} = \sqrt{a_{11}}$$

$$g_{kk} = \sqrt{a_{kk} - \sum_{j=1}^{k-1} g_{kj}^2}$$

Außerdem für $j > k$

$$a_{kj} = g_{j1}g_{k1} + g_{j2}g_{k2} + \cdots + g_{jk-1}g_{kk-1} + g_{jk}g_{kk}$$
$$\Rightarrow g_{kj} = \frac{1}{g_{kk}} \left( a_{jk} - \sum_{i=1}^{k-1} g_{ji}g_{ki} \right)$$

Pseudocode:

**Algorithmus 1** Berechne Cholesky-Zerlegung von $A \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit

**for** $i = 1$ to $n$ **do**
  $g_{kk} = \left( a_{kk} - \sum_{j=1}^{k-1} g_{kj}^2 \right)^{\frac{1}{2}}$
  **for** $j = k + 1$ to $n$ **do**
    $g_{jk} = \frac{1}{g_{kk}} \left( a_{jk} - \sum_{i=1}^{k-1} g_{ji}g_{ki} \right)$
  **end for**
**end for**

## 3.3 Singulärwertzerlegung

### Motivation

Bekanntlich kann man symmetrische Matrizen kann man vollständig mithilfe ihrer Eigenwerte und Eigenvektoren beschreiben. Mit der Matrix $Q$, in deren Spalten die Eigenvektoren $u_k$ der Matrix $A$ stehen, und der aus den Eigenwerten von $A$ bestehenden Diagonalmatrix $\Lambda$ gilt im Falle einer orthonormalen Eigenvektorbasis

$$AQ = Q\Lambda \quad \text{bzw.} \quad A = Q\Lambda Q^T \ .$$

Diese Darstellung kann unmittelbar zur geometrischen Beschreibung ihrer Wirkung auf Vektoren benutzt werden.
Wir wollen diese Beschreibung auf beliebige Matrizen erweitern. Dies leistet die Singulärwertzerlegung. Singuärwerte können ähnlich gut interpretiert werden wie Eigenwerte symmetrischer Matrizen. Vorteile der Singulärwertzerlegung gegenüber Eigenwerten und Eigenvektoren:
Sie ist nicht auf quadratische Matrizen beschränkt. In der Singulärwertzerlegung einer reellen Matrix treten nur reelle Matrizen auf (kein Rückgriff auf komplexe Zahlen).

**Theorem 3.11.** *(und Definition)*
*Gegeben sei eine Matrix $A \in \mathbb{R}^{m\times n}$. Dann gibt es orthogonale Matrizen $U \in \mathbb{R}^{m\times m}$ und $V \in \mathbb{R}^{n\times n}$ sowie eine Matrix $\Sigma = (s_{ij}) \in \mathbb{R}^{m\times n}$ mit $s_{ij} = 0$ für alle $i \neq j$ und nichtnegativen Diagonalelementen $s_{11} \geq s_{22} \geq \ldots$, für die*

$$A = U\Sigma V^T \Longleftrightarrow U^T AV = \Sigma \tag{3.3}$$

*gilt. Die Darstellung* (3.3) *heißt* **Singulärwertzerlegung** *von $A$. Die Werte $\sigma_i = s_{ii}$ heißen* **Singulärwerte** *von $A$.*

Bevor der Satz 3.11 bewiesen wird sei darauf hingewiesen, dass man die Gleichung (3.3) auch in der Form

$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^T \tag{3.4}$$

aufschreiben kann, wobei $u_j$ der $j$-te Spaltenvektor von $U$ und $v_j$ der $j$-te Spaltenvektor von $V$ ist, sowie $r$ die Zahl der von Null verschieden Singulärwerte ist.
Der folgende Beweis wird nach dem Vorbild von Deuflhard/Hohmann geführt. Eine völlig andere Beweismethode findet man bei Schwarz. Außerdem gebe ich später noch einen konstruktiven Nachweis des Satzes an.

*Beweis.* Es reicht zu zeigen, dass es orthogonale Matrizen $U \in \mathbb{R}^{m\times m}$ und $V \in \mathbb{R}^{n\times n}$ gibt, so dass

$$U^T AV = \begin{pmatrix} \sigma & \mathbf{0} \\ \mathbf{0} & B \end{pmatrix} \tag{3.5}$$

mit einer Zahl $\sigma$ und einer Matrix $B \in \mathbb{R}^{(m-1)\times(n-1)}$ gilt. Der Beweis der Beziehung (3.3) ergibt sich dann induktiv, indem man $B$ in der Art (3.5) faktorisiert usw.
Sei $\sigma := ||A||_2 = \max_{||x||=1} ||Ax||$. Das Maximum wird angenommen mit Vektoren $u \in \mathbb{R}^m$ und $v \in \mathbb{R}^n$, so dass

$$Av = \sigma u \quad \text{und} \quad ||u||_2 = ||v||_2 = 1 \tag{3.6}$$

gilt. Nun werden $v$ und $u$ mit dazu orthonormalen Vektoren $V_2, \ldots, V_n \in \mathbb{R}^n$ und $U_2, \ldots, U_m \in \mathbb{R}^m$ zu Orthonormalbasen

$$\{v = V_1, V_2, \ldots, V_n\} \quad \text{bzw.} \quad \{u = U_1, U_2, \ldots, U_m\}$$

ergänzt und damit sind

$$V = [V_1\ V_2 \ldots V_n] \quad \text{bzw.} \quad U = [U_1\ U_2 \ldots U_m]$$

orthogonale Matrizen. Das Produkt $U^T AV$ hat wegen (3.6) die Form

$$\hat{A} := U^T AV = \begin{pmatrix} \sigma & w^T \\ \mathbf{0} & B \end{pmatrix}$$

mit $w \in \mathbb{R}^{n-1}$. Es ist nun

$$\left\| \hat{A} \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_2^2 \geq (\sigma^2 + \|w\|_2^2)^2 \quad \text{und} \quad \left\| \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_2^2 = \sigma^2 + \|w\|_2^2 \ ,$$

also

$$\left\| \hat{A} \right\|_2^2 \geq \sigma^2 + \|w\|_2^2 \ .$$

Weiterhin ist $\sigma^2 = \|A\|_2^2$ und wegen der Längenerhaltung orthogonaler Abbildungen ist $\|A\|_2 = \left\| \hat{A} \right\|_2$, woraus

$$\sigma^2 = \|A\|_2^2 = \left\| \hat{A} \right\|_2^2 \geq \sigma^2 + \|w\|_2^2 \ ,$$

folgt, und damit muss $w = \mathbf{0}$ gelten. Damit ist (3.5) nachgewiesen und mit dem Hinweis auf die Induktion ist der Satz bewiesen. □

**Mögliche Konstruktion der Singulärwertzerlegung (konstruktiver Nachweis)**

1) Setze $B := A^T A$. $B \in \mathbb{R}^{n \times n}$ ist eine symmetrische Matrix. Wir bestimmen die Eigenwerte $\lambda$ und orthonormale Eigenvektoren $v$ von $B$. Da wegen der Orthogonalität der $v$ einerseits

$$v^T B v = \lambda v^T v$$

und aufgrund der Definition von $B$

$$v^T B v = v^T A^T A v = (Av)^T (Av) \geq 0$$

gilt, sind alle $n$ Eigenwerte $\lambda$ nichtnegativ. Seien o.B.d.A. die EW wie folgt geordnet $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. Da der Rang von $B$ gleich dem Rang von $A$ ist, sind genau die ersten $r$ Eigenwerte $\lambda_1, \ldots, \lambda_r$ positiv. Seien $v_1, \ldots, v_n$ die zu $\lambda_1, \ldots, \lambda_n$ gehörenden Eigenvektoren.

2) Für $j = 1, \ldots, r$ wird

$$u_j := \frac{1}{\sqrt{\lambda_j}} A v_j$$

gesetzt.

3) Man bestimmt $m$-$r$ orthonormale Vektoren $u_{r+1}, \ldots, u_m$, die zu $u_1, \ldots, u_r$ orthogonal sind.

4) Man bildet die Matrizen

$$\begin{aligned} U &:= [u_1\ u_2\ \ldots u_m] \\ V &:= [v_1\ v_2\ \ldots v_n] \end{aligned}$$

aus den orthonormalen (Spalten-) Vektoren $u_1, \ldots, u_m$ und $v_1, \ldots, v_n$. Die Matrix $\Sigma = (s_{ij}) \in \mathbb{R}^{m \times n}$ wird mit

$$s_{ij} = \begin{cases} \sqrt{\lambda_j} & i = j \leq r\ , \\ 0 & \text{sonst}\ , \end{cases}$$

gebildet.

Im Folgenden wird gezeigt, dass $A = U\Sigma V^T$ eine Singulärwertzerlegung von $A$ ist.

i) $V$ ist eine orthogonale Matrix, da $\{v_1, \ldots, v_n\}$ nach Konstruktion eine Orthonormalbasis ist.

ii) $\{u_1, \ldots, u_m\}$ ist eine Orthonormalbasis des $\mathbb{R}^m$, denn für $i, j = 1, \ldots, r$ gilt

$$u_i^T u_j = \frac{1}{\sqrt{\lambda_i \lambda_j}} v_i^T \underbrace{A^T A v_j}_{=\lambda_j v_j} = \frac{\sqrt{\lambda_j}}{\sqrt{\lambda_i}} v_i^T v_j = \begin{cases} \sqrt{\lambda_i/\lambda_i} = 1\ , & i = j, \\ 0 & \text{sonst}, \end{cases}$$

also sind $u_1, \ldots, u_r$ orthonormal, und nach Konstruktion der $u_{r+1}, \ldots, u_m$ ist $\{u_1, \ldots, u_m\}$ damit eine Orthonormalbasis und damit $U$ orthogonal.

iii) $v_{r+1}, \ldots$ sind aus dem Kern von $A$, weil

$$\mathrm{Ker}(A)^T = \mathrm{span}\{v_1, \ldots, v_r\}$$

gilt, denn nimmt man an, dass $Av_j = \mathbf{0}$ für $j = 1, \ldots, r$ gilt, dann folgt aus $Bv_j = A^T A v_j = \mathbf{0} = \lambda_j v_j$, dass $\lambda_j = 0$ sein muss, was ein Widerspruch zu $\lambda_j > 0$ für $j = 1, \ldots, r$ ist.

iv) Es gilt schließlich ausgehend von der Formel (3.4)

$$\begin{aligned}
\sum_{j=1}^{r} \sigma_j u_j v_j^T &= \sum_{j=1}^{r} A v_j v_j^T \qquad \text{nach Def. der } u_j \\
&= \sum_{j=1}^{n} A v_j v_j^T \qquad \text{da } v_{r+1}, \dots \text{ aus dem Kern von } A \text{ sind} \\
&= A \sum_{j=1}^{n} v_j v_j^T = A \underbrace{V V^T}_{=E} \\
&= AE = A \, .
\end{aligned}$$

**Bemerkung 3.12.** Es gilt nun im Weiteren

1. Die Singulärwerte von $A$ und damit $\Sigma$ sind eindeutig bestimmt, $U$ und $V$ sind dies nicht.

2. Es gilt

$$\begin{aligned}
\text{Ker } A &= \text{span } \{v_{r+1}, \dots, v_n\} \\
\text{Im } A &= \text{span } \{u_1, \dots, u_r\} \, .
\end{aligned}$$

3. Die Anzahl der von Null verschiedenen Singulärwerte ist gleich dem Rang $r$ von $A$.

4. Die Bestimmung der Singulärwerte als Quadratwurzeln der Eigenwerte von $A^T A$ kann zu numerischen Ungenauigkeiten führen. Deswegen ist das obige Verfahren zur praktischen Berechnung der Singulärwertzerlegung nicht in allen Fällen geeignet. Andere Verfahren nutzen z. B. Umformungen mittels Householder-Matrizen.

5. Für symmetrische Matrizen $A$ sind die Singulärwerte die Beträge der Eigenwerte. Sind alle Eigenwerte nichtnegativ, so ist die Diagonalisierung (Hauptachsentransformation)

$$A = Q \Lambda Q^T$$

auch eine Singulärwertzerlegung.

Eine wichtige Anwendung der Singulärwertzerlegung ist die Kompression von Bilddaten (Pixel sind dabei die Matrixelemente/Farbintensitäten/Grauwerte einer Matrix $A$). Die Grundlage hierfür liefert der

**Theorem 3.13.** *Es sei $A \in \mathbb{R}^{m\times n}$ eine Matrix vom Rang $r$ mit der Singulärwertzerlegung*

$$A = U\Sigma V^T = \sum_{j=1}^{r} \sigma_j u_j v_j^T \ .$$

*Die Approximationsaufgabe*

$$\min\{\|A - A_k\|_2 \mid A_k \in \mathbb{R}^{m\times n} \text{ und } rg(A_k) \leq k\}$$

*besitzt für $k \leq r$ die Lösung*

$$A_k = \sum_{j=1}^{k} \sigma_j u_j v_j^T \quad \text{mit} \quad \|A - A_k\|_2 = \sigma_{k+1} \ .$$

Der Beweis sollte als Übung durchgeführt werden.

# Kapitel 4

# Die iterative Lösung von Gleichungen bzw. Gleichungssystemen

6. Vorlesung 2.11.11

**Bemerkung 4.1** (Banachscher Fixpunktsatz)**.** Ist $F : A \to A, A \subset \mathbb{R}^n$ abgeschlossen, und gilt

$$\|F(x_1) - F(x_2)\| \leq L \|x_1 - x_2\|$$

mit $L < 1$ für alle $x_1, x_2 \in A$, dann hat $F$ genau einen Fixpunkt $x \in A$ mit

$$F(x) = x$$

und die durch $x_{k+1} = F(x_k)$ definierte Iterationsfolge konvergiert für jeden Anfangspunkt $x_0 \in A$ gegen diesen Fixpunkt. ($\mathbb{R}^n$ ist mit der Metrik $\rho(x, y) = \|x - y\|$ ein Banach-Raum.)

**Bemerkung 4.2.** Aus dem Banachschen Fixpunktsatz ergeben sich die Fehlerabschätzungen

$$\|x_k - \hat{x}\| \leq \frac{L^k}{1 - L} \|x_1 - x_0\| \quad \text{A-priori-Abschätzung} \tag{4.1}$$

$$\|x_k - \hat{x}\| \leq \frac{1}{1 - L} \|x_{k+1} - x_k\| \quad \text{A-posteriori-Abschätzung} \tag{4.2}$$

## 4.1 Die iterative Lösung linearer Gleichungssysteme

Neben der schon beschriebenen direkten Lösung linearer Gleichungssysteme durch den Gaußschen Algorithmus oder durch bestimmte Matrix-Faktorisie-

rungen ist es oft sinnvoll, lineare Gleichungssysteme

$$Ax = b \tag{4.3}$$

mit der regulären Matrix $a$ vom Typ $n \times n$ und $b \in \mathbb{R}^n$ iterativ zu lösen (o.B.d.A. sei $a_{kk} \neq 0, k = 1, \ldots, n$).
Zerlegt man $A$ mit der regulären Matrix $B$ in der Form $A = B + (A - B)$ dann gilt für (4.3).

$$Ax = b \Leftrightarrow Bx = (B - A)x + b \Leftrightarrow x = (E - B^{-1}A)x + B^{-1}b$$

wählt man $B$ als leicht invertierbare Matrix, dann ergibt sich im Fall der Konvergenz der Fixpunktiteration

$$x_k = (E - B^{-1}A)x_{k-1} + B^{-1}b, \quad k = 1, 2, \ldots \tag{4.4}$$

bei Wahl irgendeiner Startnäherung $x_0 \in \mathbb{R}^n$ mit dem Grenzwert $x = \lim_{k\to\infty} x_k$ die Lösung des linearen Gleichungssystems (4.3). Die Lösung ist ein Fixpunkt der Abbildung

$$F : \mathbb{R}^n \to \mathbb{R}^n : x \mapsto (E - B^{-1}A)x + B^{-1}b \tag{4.5}$$

Die Matrix $S = (E - B^{-1}A)$ heißt Iterationsmatrix. Konvergenz liegt dann vor, wenn $\lim_{k\to\infty} \|x - x_k\| = 0$ ist.
Mit $x$ und $\delta x_k = x - x_k$ folgt

$$\delta x_k = (E - B^{-1}A)\delta x_{k-1} = (E - B^{-1}A)^k \delta x_0$$

also gilt für irgendeine Vektornorm und eine dadurch induzierte Matrixnorm

$$\|\delta x_k\| \leq \left\|S^k\right\| \|\delta x_0\| \tag{4.6}$$

Damit konvergiert das Lösungsverfahren, wenn $\lim_{k\to\infty} S^k = 0$ bzw. $\lim_{k\to\infty} \left\|S^k\right\| = 0$ gilt.
Hilfreich zur Konvergenzuntersuchung ist der

**Satz 4.3.** *Sei $S$ eine $(n\times n)$-Matrix. Dann sind folgende Aussagen äquivalent:*

*(a) Der Spektralradius $r(S)$ von $S$ ist kleiner als 1*

*(b) $S^k \to 0$ für $k \to \infty$*

*(c) Es gibt eine Vektornorm, sodass sich für die induzierte Matrixnorm $\|S\| < 1$ ergibt.*

*(d) $S - \lambda E$ ist für alle $\lambda$ mit $|\lambda| \geq 1$ regulär*

*Beweis.* (Auszugsweise) a $\Rightarrow$ b Betrachten die verallgemeinerte Jordansche Normalform $S = T^{-1}JT$ mit einer regulären Matrix $T$ und $J$ mit den Jordan-Blöcken $J_i$

$$J = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_r \end{pmatrix}, \quad J_i = \begin{pmatrix} \lambda_i & \epsilon & & \\ & \lambda_i & \epsilon & \\ & & \ddots & \ddots \\ & & & \lambda_i & \epsilon \end{pmatrix}$$

für die Eigenwerte $\lambda_1, \ldots, \lambda_r$ von $S$, wobei $0 < \epsilon < 1 - |\lambda_i|$ für $i = 1, \ldots, r$ gewählt wurde. Es gilt $\left\|S^k\right\| = \left\|TJ^kT^{-1}\right\|$. Die Potenzen von $J$ enthalten für wachsendes $k$ immer größere Potenzen von $\lambda_i$, sodass wegen $|\lambda_i| < 1$ für alle Eigenwerte $\left\|S^k\right\|$ gegen null geht.
a $\Rightarrow$ c Mit der Zeilensummennorm gilt wegen der Voraussetzung zum Spektralradius von $S$, der gleich dem von $J$ ist, und der Wahl von $\epsilon$

$$\|J\|_\infty = \max_{i=1,\ldots,n} \|J_i\|_\infty < 1$$

Durch

$$\|x\|_T := \|Tx\|_\infty, \quad (x \in \mathbb{R}^n)$$

ist eine Norm auf dem $\mathbb{R}^n$ erklärt. Für die durch $\|\cdot\|_T$ induzierte Matrixnorm gilt $\|S\|_T < 1$, denn es gilt

$$\|Sx\|_T = \|TSx\|_\infty = \|JTx\|_\infty \leq \|J\|_\infty \|Tx\|_\infty = \|J\|_\infty \|x\|_T$$

und damit $\frac{\|Sx\|_T}{\|x\|_T} \leq \|J\|_\infty < 1$ für alle $x \neq 0$.
c $\Rightarrow$ d Annahme: $S - \lambda E$ singulär, d.h. $\exists x \neq 0 : (S - \lambda E)x = 0$, daraus folgt

$$Sx = \lambda x \Leftrightarrow \|Sx\|_T = |\lambda| \|x\|_T \Leftrightarrow \frac{\|Sx\|_T}{\|x\|_T} = |\lambda| \geq 1$$

andererseits ist $1 > \|S\|_T \geq \frac{\|Sx\|_T}{\|x\|_T}$, d.h. es ergibt sich ein Widerspruch und die Annahme war falsch. Damit ist $S - \lambda E$ regulär für $|\lambda| \geq 1$. □

Als Folgerung des Satzes 4.3 erhält man das folgende Konvergenzkriterium

**Satz 4.4.** *Seien $A, B$ reguläre $(n \times n)$-Matrizen. Die Iteration (4.4) konvergiert für alle Startwerte $x_0$ genau dann gegen die eindeutig bestimmte Lösung $x$ von $Ax = b$, wenn der Spektralradius $r = r(S)$ der Iterationsmatrix $S = (E - B^{-1}A)$ kleiner als 1 ist. Ist $S$ diagonalisierbar, dann gilt*

$$\|x_k - x\| \leq Cr^k, \quad C = const \in \mathbb{R} \tag{4.7}$$

Für die weitere Betrachtung konkreter Verfahren stellen wir die quadratische Matrix $A = (a_{ij})$ als Summe der unteren Dreiecksmatrix $L = (l_{ij})$, der Diagonalmatrix $D = (d_{ij})$ und der oberen Dreiecksmatrix $U = (u_{ij})$

$$A = L + D + U \tag{4.8}$$

mit

$$l_{ij} = \begin{cases} a_{ij} & i > j \\ 0 & i \leq j \end{cases}, \quad u_{ij} = \begin{cases} 0 & i \geq j \\ a_{ij} & i < j \end{cases} \quad d_{ij} = \begin{cases} a_{ij} & i = j \\ 0 & i \neq j \end{cases}$$

dar. Bei Iterationsverfahren der Form (4.4) ist für den Aufwand natürlich die einfache Invertierbarkeit von $B$ entscheidend. Das wird bei den nun zu diskutierenden Verfahren auch berücksichtigt.

## 4.2 Jacobi-Verfahren oder Gesamtschrittverfahren

Die Wahl von $B = D$ ergibt die Iterationsmatrix

$$S = E - B^{-1}A = -D^{-1}(L + U) = \begin{pmatrix} 0 & \frac{a_{12}}{a_{11}} & \dots & \frac{a_{1n}}{a_{11}} \\ \frac{a_{21}}{a_{22}} & 0 & & \\ \vdots & & \ddots & \\ \frac{a_{n1}}{a_{nn}} & & & 0 \end{pmatrix} \tag{4.9}$$

Das Verfahren (4.4) mit der durch die Wahl von $B = D$ definierten Iterationsmatrix (4.9) heißt Jacobi-Verfahren oder Gesamtschrittverfahren.
Zur besseren Darstellung von Details der Iterationsverfahren setzen wir den Iterationsindex $k$ nach oben in Klammern, also

$$x^{(k)} = x_k \in \mathbb{R}^n,$$

und die Komponenten von $x^{(k)}$ bezeichnen wir durch $x_j^{(k)}, j = 1, \dots, n$. Damit ergibt sich für die Jacobi-Verfahren koordinatenweise

$$x_j^{(k)} = \frac{1}{a_{jj}} \left( b_j - \sum_{j \neq i = 1}^{n} a_{ji} x_i^{(k-1)} \right), \quad j = 1, \dots, n, k = 1, 2, \dots$$

**Definition 4.5.** *Eine Matrix vom Typ $(n \times n)$ heißt strikt diagonal dominant, wenn gilt*

$$|a_{ii}| > \sum_{j \neq i = 1}^{n} |a_{ij}| \ .$$

Zur Konvergenz des Jacobi-Verfahrens gilt der

**Satz 4.6.** *Sei $A$ eine strikt diagonal dominante $(n \times n)$-Matrix. Dann ist der Spektralradius kleiner als 1 und das Verfahren konvergiert.*

*Beweis.*

$$S = -D^{-1}(L + U)$$

Zeilensummen von S

$$\sum_{j=1}^{n} s_{ij} = \frac{1}{a_{ii}} \sum_{j \neq i}^{n} a_{ij},$$

aufgrund der strikten Diagonaldominanz ist

$$\sum_{i \neq j = 1}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| < 1 \Rightarrow \|S\|_\infty < 1 \Rightarrow r(S) < 1$$

□

Bei der numerischen Lösung von elliptischen Randwertproblemen treten oft Matrizen auf, die nicht strikt diagonal dominant sind, aber die folgenden etwas schwächeren Eigenschaften besitzen.

**Definition 4.7.** *(a) Eine Matrix vom Typ $(n \times n)$ heißt schwach diagonal dominant, wenn gilt*

$$|a_{ii}| \geq \sum_{j \neq i = 1}^{n} |a_{ij}| \ .$$

*(b) Eine $(n \times n)$-Matrix $A = (a_{ij})$ heißt irreduzibel, wenn für alle $i, j \in \{1, 2, \ldots, n\}$ entweder $a_{ij} \neq 0$ oder eine Indexfolge $i_1, \ldots, i_s \in \{1, \ldots, n\}$ existiert, sodass $a_{ii_1} a_{i_1 i_2} \cdots a_{i_s j} \neq 0$ ist. Andernfalls heißt $A$ reduzibel.*

*(c) Eine $(n \times n)$-Matrix $A = (a_{ij})$ heißt irreduzibel diagonal dominant, wenn sie irreduzibel und schwach diagonal dominant ist, sowie wenn es einen Index $l \in \{1, \ldots, n\}$ mit*

$$|a_{ll}| > \sum_{l \neq j = 1}^{n} |a_{lj}|$$

*gibt.*

**Bemerkung.** 1. Man kann entscheiden, ob eine Matrix reduzibel oder irreduzibel ist, indem man für $A = (a_{ij})_{i,j=1,\dots,n}$ einen Graphen mit $n$ Knoten konstruiert, indem eine gerichtete Kante von Knoten $i$ zum Knoten $j$ existiert, wenn $a_{ij} \neq 0$ ist.

Kann man in diesem Graphen ausgehend von einem Knoten alle anderen auf einem gerichteten Weg (Folge von gerichteten Kanten) erreichen, ist $A$ irreduzibel, andernfalls reduzibel.

2. Ein weiteres Kriterium zur Entscheidung ob $A$ vom Typ $n \times n$ irreduzibel oder reduzibel ist, ist Folgendes:

**Lemma 4.8.** *Die $(n \times n)$-Matrix $A$ ist irreduzibel, falls es* keine *Permutationsmatrix $P$ vom Typ $n \times n$ gibt, so dass bei gleichzeitiger Zeilen- und Spaltenpermutation*

$$P^T A P = \begin{pmatrix} F & 0 \\ G & H \end{pmatrix}$$

*gilt, wobei $F$ und $H$ quadratische Matrizen sind und $0$ eine Nullmatrix ist, andernfalls ist $A$ reduzibel.*

**Satz 4.9.** *Für eine irreduzibel diagonal dominante Matrix $A$ ist das Jacobi-Verfahren konvergent.*

*Beweis.* aufwändig (siehe z.B. Schwarz) □

## 4.3 Gauß-Seidel-Iterationsverfahren oder Einzelschrittverfahren

Wählt man ausgehend von der Matrixzerlegung (4.8) $B = L + D$, dann heißt das Iterationsverfahren (4.4) Gauß-Seidel- Verfahren oder Einzelschrittverfahren, d.h. es ergibt sich

$$x^{(k)} = \underbrace{(E - B^{-1}A)}_{S} x^{(k-1)} + B^{-1}b = (L + D)^{-1}(-Ux^{(k-1)} + b), \quad k = 1, 2, \dots \tag{4.10}$$

Die Matrix $B = L + D$ ist eine reguläre untere Dreiecksmatrix und damit leich zu invertieren, was aber keine Arbeit bedeuten wird, wie wir etwas später sehen werden.

**Satz 4.10.** *Das Gauß-Seidel-Verfahren konvergiert für strikt diagonal dominante Matrizen $A$ für beliebige Startiterationen $x^{(0)} \in \mathbb{R}^n$*

*Beweis.* Es ist

$$S = E - B^{-1}A, \quad \lambda v = Sv = (E - B^{-1}A)v = -(L+D)^{-1}Uv, \quad v \neq 0$$

für einen EW $\lambda$ mit dem EV $v$ bzw.

$$\lambda(L+D)v = -Uv; \quad |v_k| = \max_{1 \leq i \leq n} |v_i| > 0$$

Wir betrachten die $k$-te Zeile:

$$\begin{aligned}
& \lambda(a_{kk}v_k + \sum_{k>j=1}^{n} a_{kj}v_j) = -\sum_{k<j=1}^{n} a_{kj}v_j \\
\rightsquigarrow \quad & \lambda\left(1 + \sum_{k>j=1}^{n} \frac{a_{kj}v_j}{a_{kk}v_k}\right) = -\sum_{k<j=1}^{n} \frac{a_{kj}v_j}{a_{kk}v_k}, \quad \left|\frac{v_j}{v_k}\right| \leq 1 \\
\Leftrightarrow \quad & \lambda(1+\alpha) = \beta, \quad |\alpha|, |\beta| < 1 \text{ da } A \text{ strikt diagonal dominant} \\
\Leftrightarrow \quad & \lambda = \frac{\beta}{1+\alpha} \rightsquigarrow |\lambda| = \frac{|\beta|}{|1+\alpha|} \leq \frac{|\beta|}{1-|\alpha|} \quad (*)
\end{aligned}$$

Aus der strengen Diagonaldominanz folgt schließlich

$$|\alpha| + |\beta| = \left|\sum_{k>j=1}^{n} \frac{a_{kj}v_j}{a_{kk}v_k}\right| + \left|\sum_{k<j=1}^{n} \frac{a_{kj}v_j}{a_{kk}v_k}\right| \leq \sum_{k \neq j=1}^{n} \left|\frac{a_{kj}}{a_{kk}}\right| < 1,$$

woraus $|\beta| < 1 - |\alpha|$ bzw. $\frac{|\beta|}{1-|\alpha|} < 1$ und mit $(*)$

$$|\lambda| \leq \frac{|\beta|}{1-|\alpha|} < 1$$

für alle EW $\lambda$ folgt. Damit ist $(r(S) < 1)$ und der Satz bewiesen. □

**Bemerkung 4.11.** Ebenso wie beim Jacobi-Verfahren kann man die Voraussetzung der strikten Diagonaldominanz von $A$ abschwächen.
Das Gauß-Seidel-Verfahren ist für irreduzibel diagonal dominante Matrizen $A$ konvergent.

Wenn man das Gauß-Seidel Verfahren (4.10) in der äquivalenten Form

$$x^{(k)} = D^{-1}(-Lx^{(k)} - Ux^{(k-1)} + b), \quad k = 1, 2, \ldots \tag{4.11}$$

aufschreibt, erkennt man bei der koordinatenweisen Berechnung der neuen Iteration

$$x_j^{(k)} = \frac{1}{a_{jj}}\left(b_j - \sum_{i=1}^{j-1} a_{ji}x_i^{(k)} - \sum_{i=j+1}^{n} a_{ji}x_i^{(k-1)}\right), \quad j = 1, 2, \ldots, n, \ k = 1, 2, \ldots \tag{4.12}$$

zwar, dass auf beiden Seiten der Formeln $x^{(k)}$ vorkommen. Allerdings benötigt man zur Berechnung der $j$-ten Komponenten von $x^{(k)}$ nur die Komponenten $x_1^{(k)}, \ldots, x_{j-1}^{(k)}$ der vorigen Iteration. Diese kennt man aber bereits. Damit kann man die Formel (4.12) für $j = 1, \ldots, n$ sukzessiv zum Update der Koordinaten von $x$ anwenden. Man hat also mit (4.12) eine explizite Berechnungsvorschrift und braucht damit $B = L + D$ nicht wirklich zu invertieren.

## 4.4 Verallgemeinerung des Gauß-Seidel-Verfahrens

Wenn man ausgehend von $x^{(k-1)}$ mit dem Gauß-Seidel-Verfahren eine Näherung

$$\hat{x}^{(k)} = D^{-1}(-Lx^{(k)} - Ux^{(k-1)} + b)$$

bestimmt, und anschließend "relaxiert", d.h. mit $\omega \in ]0, 2[$ die Wichtung

$$x^{(k)} = \omega\hat{x}^{(k)} + (1 - \omega)x^{(k-1)} \tag{4.13}$$

vornimmt, erhält man nach kurzer Rechnung durch

$$x^{(k)} = S_\omega x^{(k-1)} + B^{-1}b, \quad k = 1, 2, \ldots, \ \omega \in ]0, 2[ \tag{4.14}$$

das Gauß-Seidel-Verfahren mit Relaxation, wobei für $S_\omega$ und $B$

$$S = S_\omega = (D + \omega L)^{-1}[(1 - \omega)D - \omega U] = E - \omega(D + \omega L)^{-1}A$$

bzw.

$$B^{-1} = \omega(D + \omega L)^{-1} \Longleftrightarrow B = \frac{1}{\omega}(D + \omega L)$$

gilt. Für $\omega > 1$ spricht man vom sukzessiven Überrelaxtionsverfahren auch SOR-Verfahren genannt. Das SOR-Verfahren konvergiert in allen Fällen, in denen das Gauß-Seidel-Verfahren ($\omega = 1$) konvergiert.
Allerdings kann man in vielen Fällen mit einer Wahl von $\omega > 1$ eine schnellere Konvergenz als mit dem Gauß-Seidel-Verfahren erreichen.

## 4.5 Krylov-Raum-Verfahren zur Lösung linearer Gleichungssysteme

7. Vorlesung 7.11.11

Ziel ist weiterhin die iterative Lösung des linearen Gleichungssystems

$$Ax = b, \quad (n \times n)\text{-Matrix, regulär, } b \in \mathbb{R}^n$$

mit der eindeutigen Lösung $x_* = A^{-1}b$
Hierzu betrachten wir mit

$$\{0\} \subset D_1 \subset \ldots \subset \mathbb{R}^n \tag{4.15}$$

zunächst eine Folge von linearen Unterräumen, die noch präzisiert wird. Im Folgenden werden Ansätze zur Bestimmung von Vektorfolgen $x_k \in D_k, k = 1, \ldots$ betrachtet (mit dem letztendlichen Ziel mit dieser Folge die exakte Lösung $x_*$ zu erreichen).

**Definition 4.12.**

*(a) Für gegebene Ansatzräume* (4.15) *hat der* ***Ansatz des orthogonalen Residuums*** *zur Bestimmung von Vektoren $x_1, x_2, \ldots \in \mathbb{R}^n$ die Form*

$$\left.\begin{array}{c} x_k \in D_k \\ Ax_k - b \in D_k^{\perp} \end{array}\right\} k = 1, 2, \ldots \tag{4.16}$$

*(b) Der* ***Ansatz des minimalen Residuums*** *zur Bestimmung der Vektorfolge hat die Form*

$$\left.\begin{array}{c} x_k \in D_k \\ \|Ax_k - b\|_2 \text{ minimal} \end{array}\right\} k = 1, 2, \ldots \tag{4.17}$$

Bei der Wahl spezieller Ansatzräume (4.15) werden die sogenannten Krylovräume von Bedeutung sein

**Definition 4.13.** *Zu gegebener Matrix $A \in \mathbb{R}^{n \times n}$ und einem Vektor $b \in \mathbb{R}^n$ ist die Folge der Krylovräume durch*

$$K_k(A, b) = \operatorname{span}\{b, Ab, \ldots, A^{k-1}b\} \subset \mathbb{R}^n, \quad k = 0, 1, \ldots$$

*erklärt.*

**Bemerkung.** Im Folgenden werden die in Definition 4.12 angegebenen Ansätze mit den speziellen Räumen $D_k = K_k(A, b)$ betrachtet, wobei wir den Schwerpunkt auf den Ansatz (4.16) legen.

### 4.5.1 Der Ansatz des orthogonalen Residuums (4.16) für symmetrische positiv definite Matrizen

Für positiv definite, symmetrische Matrizen soll nun Existenz und Eindeutigkeit von Vektoren $x_k$ für (4.16) diskutiert werden. Dazu werden die Skalarprodukte und Normen

$$\langle x, y\rangle_2 = x^T y, \quad x, y \in \mathbb{R}^n$$
$$\langle x, y\rangle_A := x^T A y, \quad x, y \in \mathbb{R}^n, \|x\|_A = \langle x, x\rangle_A^{\frac{1}{2}}$$

betrachtet (Nachweis, dass $\langle\cdot,\cdot\rangle_A, \|\cdot\|_A$ Skalarprodukt und Norm im Falle einer positiv definiten, symmetrischen Matrix $A$ sind, ist als Übung zu führen).

**Satz 4.14.** *Zu gegebener symmetrischer positiv definiter Matrix $A \in \mathbb{R}^{n\times n}$ sind für $k = 1, 2, \ldots$ die Vektoren $x_k$ aus dem Ansatz des orthogonalen Residuums* (4.16) *– mit allgemeinen Ansatzräumen $D_k$ gemäß* (4.15) *– eindeutig bestimmt, und es gilt*

$$\|x_k - x_*\|_A = \min_{x\in D_k} \|x - x_*\|_A, \quad k = 1, 2, \ldots \tag{4.18}$$

*Beweis.* Eindeutigkeit: Sei $k$ fest gewählt. Für $x_k, \hat{x}_k$ mit der Eigenschaft (4.16) gilt

$$\langle A(x_k - \hat{x}_k), x_k - \hat{x}_k\rangle_2 = 0 \Rightarrow x_k = \hat{x}_k$$

Existenz: Mit einer beliebigen Basis $d_0, \ldots, d_{m-1}$ von $D_k$ setzt man

$$x_k = \sum_{j=0}^{m-1} \alpha_j d_j \tag{4.19}$$

an und erhält

$$\begin{aligned} x_k \text{ genügt } (4.16) &\Leftrightarrow Ax_k - b \in D_k^{\perp} \\ &\Leftrightarrow \langle Ax_k - b, d_k\rangle_2 = 0 \quad k = 0, \ldots, m-1 \qquad (4.20) \\ &\Leftrightarrow \sum_{j=0}^{m-1} \langle Ad_j, d_k\rangle_2 \alpha_j = \langle b, d_k\rangle_2, \quad k = 0, \ldots, m-1 \qquad (4.21) \end{aligned}$$

(4.21) ist ein lineares Gleichungssystem von $m$ Gleichungen für die Koeffizienten $\alpha_0, \ldots, \alpha_{m-1}$. Da $x_k$ mit (4.16) eindeutig bestimmt ist (wurde schon gezeigt), ist das Gleichungssystem (4.21) eindeutig lösbar, woraus die Existenz von $x_k$ folgt.

Minimalität (4.18) Für $x \in D_k$ findet man

$$\begin{aligned}\|x - x_*\|_A^2 &= \|x_k - x_* + x - x_k\|_A^2 \\ &= \|x_k - x_*\|_A^2 + 2\left\langle \underbrace{A(x_k - x_*)}_{\in D_k^\perp}, \underbrace{x - x_k}_{\in D_k} \right\rangle_2 + \|x - x_k\|_A^2 \quad \geq \|x_k - x_*\|_A^2\end{aligned}$$

□

### 4.5.2 Der Ansatz des orthogonalen Residuums (4.16) für gegebene $A$-konjugierte Basen

Mit dem Beweis von Satz 4.14 ist bereits eine Möglichkeit zur Bestimmung von $x_k$ für (4.16) ausgehend von einer Basis $d_0, \ldots, d_{m-1}$ für $D_k$ mit dem Gleichungssystem (4.21) aufgezeigt worden. Im Folgenden wird ein Spezialfall behandelt, bei dem (4.21) Diagonalgestalt hat.

**Definition 4.15.** *Es sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Gegebene Vektoren $d_0, \ldots, d_{m-1} \in \mathbb{R}^n \setminus \{0\}$ heißen $A$- konjugiert, falls*

$$\langle Ad_i, d_j\rangle_2 = \langle d_i, d_j\rangle_A = 0 \quad i \neq j$$

*gilt.*

**Bemerkung.** Falls eine $A$-konjugierte Basis von $D_k$ gegeben ist, hat (4.21) Diagonalgestalt und damit ist $x_k$ gemäß Ansatz (4.19) sehr einfach berechenbar.

**Satz 4.16.** *Für eine gegebene symmetrische positiv definite Matrix $A \in \mathbb{R}^{n\times n}$ und $A$-konjugierte Vektoren $d_0, \ldots$ gelte*

$$D_k = \operatorname{span}\{d_0, \ldots, d_{k-1}\}, \quad k = 1, 2, \ldots$$

*Dann erhält man für den Ansatz des orthogonalen Residuums (4.16) die folgenden Darstellungen für $k = 1, 2, \ldots$*

$$x_k = \sum_{j=0}^{k-1} \alpha_j d_j \quad mit \quad \alpha_j = -\frac{\langle r_j, d_j\rangle_2}{\langle Ad_j, d_j\rangle_2} \tag{4.22}$$

$$r_j := Ax_j - b, \quad j \geq 1, r_0 = -b \tag{4.23}$$

*Beweis.* Folgt unmittelbar für $k = m$ aus (4.19)-(4.21) □

**Bemerkung.**

(a) Aus (4.22) folgt die Unabhängigkeit der $\alpha_j$ von $k$ und damit gilt

$$x_{k+1} = x_k + \alpha_k d_k, \quad r_{k+1} = r_k + \alpha_k A d_k, \quad k = 0, 1, \ldots ; x_0 = 0 \quad (4.24)$$

(b) Aufgrund der ersten Identität von (4.24) bezeichnet man $d_k$ als Suchrichtung und $\alpha_k$ als Schrittweite

(c) Außerdem wird mit (4.24) klar, dass eine simultane Berechnung der Suchrichtungen und Lösungsapproximationen $x_k$ in der Reihenfolge

$$d_0, x_1, d_1, x_2, \ldots$$

möglich ist. In der Praxis wird im Fall $D_k = K_k(A, b)$ auch so vorgegangen, was im Folgenden behandelt werden soll.

### 4.5.3 Das CG-Verfahren für positiv definite, symmetrische Matrizen

8. Vorlesung 9.11.11

**Definition 4.17.** *Zu gegebener symmetrisch positiv definiter Matrix $A \in \mathbb{R}^{n\times n}$ ist das* ***Verfahren des konjugierten Gradienten*** *gegeben durch den Ansatz* (4.16) *mit der speziellen Wahl*

$$D_k = K_k(A, b), \quad k = 0, 1, \ldots \quad (4.25)$$

*Dieses Verfahren bezeichnet man auch kurz als CG-Verfahren.*

**Bemerkung.** Zur konkreten Bestimmung der Lösungsapproximationen fehlen uns nur noch geeignete Suchrichtungen, am besten $A$-konjugierte Suchrichtungen $d_0, d_1, \ldots$. Das soll nun geschehen.

Der folgende Hilfssatz behandelt die Berechnung $A$-konjugierter Suchrichtungen in $K_k(A, b)$ für $k = 0, 1, \ldots$
Ausgehend von den Notationen des Satzes 4.16 wird für den fixierten Index $k$ dabei so vorgegangen, dass – ausgehend von einer bereits konstruierten $A$-konjugierten Basis $d_0, \ldots, d_{k-1}$ f ür $K_k(A, b)$ – eine $A$-konjugierte Basis für $K_{k+1}(A, b)$ gewonnen wird durch eine Gram-Schmidt-Orthogonalisierung der Vektoren $d_0, \ldots, d_{k-1}, -r_k \in \mathbb{R}^n$ bezüglich des Skalarproduktes $\langle \cdot, \cdot \rangle_A$.
Wie sich im Beweis von Lemma 4.18 herausstellt, genügt hier eine Gram-Schmidt-Orthogonalisierung der beiden Vektoren $d_{k-1}, -r_k \in \mathbb{R}^n$.

**Lemma 4.18.** *Zu gegebener symmetrisch positiv definiter Matrix $A \in \mathbb{R}^{n\times n}$ und mit den Notationen des Satzes 4.16 seien die Suchrichtungen speziell wie folgt gewählt:*

$$d_0 = b, \quad d_k = -r_k + \beta_{k-1} d_{k-1}, \quad \beta_{k-1} = \frac{\langle Ar_k, d_{k-1}\rangle_2}{\langle Ad_{k-1}, d_{k-1}\rangle_2}, \quad k = 1, \ldots, k_* - 1 \tag{4.26}$$

*wobei $k_*$ den ersten Index mit $r_{k_*} = 0$ bezeichnet. Mit dieser Wahl sind die Vektoren $d_0, \ldots, d_{k_*-1} \in \mathbb{R}^n$ A-konjugiert und es gilt*

$$\operatorname{span}\{d_0, \ldots, d_{k-1}\} = \operatorname{span}\{b, r_1, \ldots, r_{k-1}\} = K_k(A, b), \quad k = 1, \ldots, k_* \tag{4.27}$$

*Beweis.* Vollständige Induktion über $k = 1, \ldots, k_*$ zum Nachweis der $A$-Konjugiertheit der Vektoren $d_0, \ldots, d_{k-1} \in \mathbb{R}^n$ und der Formeln (4.26) wegen

$$\operatorname{span}\{d_0\} = \operatorname{span}\{b\} = K_1(A, b)$$

ist der Induktionsanfang gemacht.
Im Folgenden sei angenommen, dass (4.26) ein System von $A$- konjugierten Vektoren mit der Eigenschaft (4.27) liefert mit einem fixierten Index $1 \le k \le k_* - 1$
Gemäß dem Ansatz des orthogonalen Residuums (4.16) gilt $r_k \in K_k(A, b)^\perp$ und im Fall $r_k \neq 0$ sind damit die Vektoren $d_0, \ldots, d_{k-1}, -r_k$ linear unabhängig. Eine Gram- Schmidt-Orthogonalisierung dieser Vektoren bzgl. des Skalarproduktes $\langle \cdot, \cdot \rangle$ liefert den Vektor

$$d_k = -r_k + \sum_{j=0}^{k-1} \frac{\langle Ar_k, d_j\rangle_2}{\langle Ad_j, d_j\rangle_2} d_j \overset{(*)}{=} -r_k + \beta_{k-1} d_{k-1} \tag{4.28}$$

wobei $(*)$ aus den Eigenschaften

$$K_{k-1}(A, b) \subset K_k(A, b) \quad \text{sowie} \quad r_k \in K_k(A, b)^\perp$$

folgt, also

$$\langle Ar_k, d_j\rangle_2 = \langle r_k, Ad_j\rangle_2 = 0, \quad j = 0, \ldots, k-2$$

Nach Konstruktion sind die Vektoren $d_0, \ldots, d_{k-1}, d_k$ $A$- konjugiert und es gilt

$$\operatorname{span}\{d_0, \ldots, d_k\} = \operatorname{span}\{b, r_1, \ldots, r_k\}$$

Aufgrund der 2. Formel in (4.24) gilt wegen $Ad_{k-1} \in K_{k+1}(A, b)$ noch

$$\operatorname{span}\{b, r_1, \ldots, r_k\} \subset K_{k+1}(A, b)$$

sodass aus Dimensionsgründen auch hier notwendigerweise Gleichheit vorliegt. □

**Bemerkung.** Mit dem durch Lemma 4.18 beschriebenen Abbruch wird gleichzeitig die Lösung von $Ax = b$ geliefert, es gilt also $x_{k_*} = x_*$. Dabei gilt notwendigerweise

$$k_* \leq n$$

denn aufgrund der linearen Unabhängigkeit der beiden Vektorsysteme in (4.27) erhält man

$$\dim K_k = k$$

für $k = 0, 1, \ldots, k_*$

Im folgenden Lemma werden Darstellungen für die Schrittweiten gezeigt, wie sie auch in numerischen Implementierungen verwendet werden.

**Lemma 4.19.** *In der Situation des Lemma 4.18 gelten die Darstellungen*

$$\lambda_k = \frac{\|r_k\|_2^2}{\langle Ad_k, d_k\rangle_2}, \quad k = 0, 1, \ldots, k_* - 1 \tag{4.29}$$

$$\beta_{k-1} = \frac{\|r_k\|_2^2}{\|r_{k-1}\|_2^2}, \quad k = 1, \ldots, k_* - 1 \quad (r_0 := b) \tag{4.30}$$

*Beweis.* Mit $r_k \in K_k(A, b)^\perp$ sowie der Beziehung (4.28) für die Suchrichtung $d_k$ erhält man $-\langle r_k, d_k\rangle_2 = \|r_k\|_2^2$ und zusammen mit (4.22) ergibt dies (4.29). Diese Darstellung (4.29) für $\alpha_k$ zusammen mit der Identität $r_k = r_{k-1} + \alpha_{k-1} Ad_{k-1}$ aus (4.24) liefert

$$\|r_k\|_2^2 = \underbrace{\langle r_k, r_{k-1}\rangle}_{=0} + \alpha_{k-1} \langle r_k, Ad_{k-1}\rangle_2 = \beta_{k-1} \|r_{k-1}\|_2^2$$

und damit gilt für $\beta_{k-1}$ die Beziehung (4.30) □

### 4.5.4 Konvergenzgeschwindigkeit des CG-Verfahrens

Wir haben bisher festgestellt, dass das CG-Verfahren mit $x_{k_*} = x_*$ nach $k_*$ Schritten die Lösung ergibt. $k_*$ kann aber sehr groß sein und deshalb interessiert auch der Fehler im $k$-ten Schritt ($k = 1, 2, \ldots$). Hilfreich ist

**Lemma 4.20.** *Zu gegebener symmetrisch positiv definiter Matrix $A \in \mathbb{R}^{n\times n}$ sei $(\lambda_j, v_j)_{j=1,\ldots,n}$ ein vollständiges System von Eigenwerten $\lambda_j > 0$ und zugehörigen Eigenvektoren $v_j \in \mathbb{R}^n$, also gilt*

$$Av_j = \lambda_j, \quad v_k^T v_j = \delta_{kj}, \quad k, j = 1, \ldots, n$$

*Mit der Entwicklung*

$$x = \sum_{j=1}^{n} c_j v_j \in \mathbb{R}^n$$

*gelten für jedes Polynom $p$ die folgenden Darstellungen*

$$p(A)x = \sum_{j=1}^{n} c_j p(\lambda_j) v_j \tag{4.31}$$

$$\|p(A)x\|_2 = \left( \sum_{j=1}^{n} c_j^2 p(\lambda_j)^2 \right)^{\frac{1}{2}}, \quad \|p(A)x\|_A = \left( \sum_{j=1}^{n} c_j^2 \lambda_j p(\lambda_j)^2 \right)^{\frac{1}{2}} \tag{4.32}$$

*Speziell gilt also*

$$m^{\frac{1}{2}} \|x\|_2 \leq \|x\|_A \leq M^{\frac{1}{2}} \|x\|_2, \quad x \in \mathbb{R}^n \tag{4.33}$$

*($m := \min_{1 \leq j \leq n} \lambda_j$, $M := \max_{1 \leq j \leq n} \lambda_j$)*

*Beweis.* Mit der angegebenen Entwicklung für $x \in \mathbb{R}^n$ gilt

$$A^\nu x = \sum_{j=1}^{n} c_j \lambda_j^\nu v_j, \quad \nu = 0, 1, \ldots$$

und daraus folgt (4.31). Weiter berechnet man

$$\begin{aligned} \|p(A)x\|_2 &= \left\langle \sum_{k=1}^{n} x_k p(\lambda_k) v_k, \sum_{j=1}^{n} c_j p(\lambda_j) v_j \right\rangle_2^{\frac{1}{2}} \\ &= \left( \sum_{k,j=1}^{n} c_k c_j p(\lambda_k) p(\lambda_j) \underbrace{\langle v_k, v_j \rangle_2}_{=\delta_{kj}} \right)^{\frac{1}{2}} \\ &= \left( \sum_{j=1}^{n} c_j^2 p(\lambda_j)^2 \right)^{\frac{1}{2}} \end{aligned}$$

Und analog erhält man

$$\|p(A)x\|_A = (\sum_{j=1}^{n} c_j^2 \lambda_j p(\lambda_j)^2)^{\frac{1}{2}}$$

□

Es gilt nun noch den Fehler $\|x_k - x_*\|_A$ den man im $k$-ten Schritt des CG-Verfahrens macht, abzuschätzen. Einmal gilt der

**Satz 4.21.** *Zu einer gegebenen symmetrisch positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$ gelten für das CG-Verfahren die folgenden Fehlerabschätzungen:*

$$\|x_k - x_*\|_A \leq \left( \inf_{p\in\Pi_k, p(0)=1} \sup_{\lambda\in\sigma(A)} |p(\lambda)| \right) \|x_*\|_A \tag{4.34}$$

*Beweis.* Für jedes Polynom $p \in \Pi_k$ mit $p(0) = 1$ ist $q(t) := \frac{1-p(t)}{t}$ ein Polynom vom Grad höchstens $k-1$ und damit gilt mit $x := q(A)b$ folgendes:

$$x \in K_k(A, b), \quad x - x_* = -p(A)x_*$$

Mit Lemma 4.20 und der Entwicklung $x_* = \sum_{j=1}^n c_j v_j \in \mathbb{R}^n$ erhält man

$$\begin{aligned}\|x_k - x_*\|_A &\overset{(4.18)}{\leq} \underbrace{\|x - x_*\|_A}_{=\|p(A)x_*\|_A} = \left( \sum_{j=1}^n c_j^2 \lambda_j p(\lambda_j)^2 \right)^{\frac{1}{2}} \\ &\leq \sup_{\lambda\in\sigma(A)} |p(\lambda)| \left( \sum_{j=1}^n c_j^2 \lambda_j \right)^{\frac{1}{2}} = \sup_{\lambda\in\sigma(A)} |p(\lambda)| \, \|x_*\|_A\end{aligned}$$

□

Zur quantitativen Präzisierung der Abschätzung (4.34) des Satzes 4.21 benutzen wir die hier nicht bewiesenen Eigenschaften der Tschebyscheff-Polynome erster Art $T_0, T_1, \ldots$

$$T_k\left(\frac{\kappa+1}{\kappa-1}\right) \geq \frac{1}{2}\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k \quad \text{für} \quad k \in \mathbb{N}, \kappa > 1 \, . \tag{4.35}$$

Außerdem sei daran erinnert, dass die Tschebyscheff-Polynome in der Form $T_k(t) = \cos(k \arccos t)$ darstellbar sind und damit dem Betrage nach durch 1 beschränkt sind.
Es gilt der

**Satz 4.22.** *Zu einer gegebenen symmetrisch positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$ gelten für das CG-Verfahren die Fehlerabschätzungen*

$$\begin{aligned}\|x_k - x_*\|_A &\leq 2\gamma^k \|x_*\|_A \, , \quad k = 0, 1, \ldots \\ \|x_k - x_*\|_2 &\leq 2\sqrt{\kappa_A}\gamma^k \|x_*\|_2 \, , \quad k = 0, 1, \ldots\end{aligned} \tag{4.36}$$

*mit* $\kappa_A = \mathrm{cond}_2(A), \gamma = \frac{\sqrt{\kappa_A}-1}{\sqrt{\kappa_A}+1}$

*Beweis.* Satz 4.21 wird im Fall $\kappa_A > 1$, d.h. $M > m$ angewendet mit dem Polynom

$$p(\lambda) = \frac{T_k[(M+m-2\lambda)/(M-m)]}{T_k[(M+m)/(M-m)]}, \quad \lambda \in \mathbb{R}$$

wobei $m$ und $M$ den kleinsten und größten Eigenwert von $A$ bezeichnen. Offensichtlich ist $p \in \Pi_k$ und $p(0) = 1$, wegen $\sigma(A) \subset [m, M]$ und

$$\max_{m \le \lambda \le M} |p(\lambda)| = \left| T_k\left(\frac{M+m}{M-m}\right)\right|^{-1} = \left| T_k\left(\frac{\kappa_A+1}{\kappa_A-1}\right)\right|^{-1} \overset{(4.35)}{\le} 2\gamma^k$$

folgt aus (4.34) die erste Abschätzung, also (4.36).
Die zweite Abschätzung des Satzes ist eine unmittelbare Konsequenz aus der Ersten unter der Nutzung der Normäquivalenz (4.33). □

### 4.5.5 CGNR-Verfahren

Das CG-Verfahren funktioniert wie besprochen **nur** für symmetrische positiv definite Matrizen. Was kann man tun, wenn die Matrix $A \in \mathbb{R}^{n\times n}$ des Gleichungssystems $Ax = b$ zwar regulär, aber nicht symmetrisch positiv definit ist und man trotzdem die Vorteile des CG-Verfahrens nutzen möchte?
Wir wissen, dass für reguläre Matrizen $A$ das Produkt $M = A^TA$ symmetrisch und positiv definit ist. Da

$$Ax = b \Longleftrightarrow A^TAx =: Mx = \hat{b} := A^Tb$$

kann man nun einfach das Gleichungssystem $Mx = \hat{b}$, das man auch **Normalgleichungssystem** nennt, mit dem CG-Verfahren lösen. Dieses Verfahren nennt man auch **CGNR-Verfahren**, wobei **N** und **R** für Normal bzw. Residuen steht. Die obigen Resultate (Fehlerabschätzungen) lassen sich in gewisser Weise für das CGNR-Verfahren übertragen.

### 4.5.6 GMRES-Verfahren

Lässt man die Voraussetzung der Symmetrie und positiven Definitheit der Matrix $A$ fallen und fordert nur die Regularität, dann ist ein CG-Verfahren zur Lösung von $Ax = b$ nicht möglich. Eine Alternative ist das GMRES-Verfahren

**Definition 4.23.** *Das GMRES-Verfahren ist definiert durch den Ansatz des minimalen Residuums* (4.17) *mit der speziellen Wahl $D_k = K_k(A, b)$, es gilt also*

$$x_k \in K_k(A,b), \quad \|Ax_k - b\|_2 = \min_{x \in K_k(A,b)} \|Ax - b\|_2, \quad k = 0, \ldots, k_*$$

**Bemerkung.** Die Abkürzung "GMRES" hat ihren Ursprung in der Bezeichnung " **g**eneralized **m**inimal **res**idual method"
Detaillierte Konstruktionsmethoden für die Approximationen $x_k$ beim GMRES-Verfahren werden in Plato beschrieben.

## 4.6 Die iterative Lösung nichtlinearer Gleichungssysteme

Ein nichtlineares Gleichungssystem wollen wir durch Nutzung einer Abbildung

$$F : D \to \mathbb{R}^n \,,\; D \subset \mathbb{R}^n \,,$$

und die Suche nach einer Nullstelle, also durch die Gleichung

$$F(x) = \mathbf{0} \Longleftrightarrow x = B(x) := x - F(x) \tag{4.37}$$

beschreiben. Wenn $B$ eine Kontraktion ist, dann kann man nach dem Banachschen Fixpunktsatz mit der der Iterationsfolge

$$x_{(k)} = B(x^{(k-1)})$$

die Lösung annähern.

Im Falle der Differenzierbarkeit von $F$ kann man bekanntlich linear approximieren, d.h. für eine geeignete Näherung der Nullstelle von $F$ die Tangentenabbildung

$$T(x) = F(x^{(0)}) + F'(x^{(0)})(x - x^{(0)}) \approx F(x)$$

ausrechnen, und die Nullstelle von $T$ durch

$$x = x^{(0)} - [F'(x^{(0)})]^{-1} F(x^{(0)})$$

als Näherung einer Nullstelle von $F$ bestimmen. Diesen Prozess kann man nun (sofern er funktioniert und $F'(x_{k-1})$ regulär ist) iterativ fortsetzen und erhält die Iterationsfolge

$$x^{(k)} = x^{(k-1)} - [F'(x^{(k-1)})]^{-1} F(x^{(k-1)} \quad k = 1, 2, \ldots \tag{4.38}$$

die auch **Newtonfolge** genannt wird. Für skalare Gleichungen erhält man die Folge

$$x_k = x_{k-1} - \frac{f(x_{k-1})}{f'(x_{k-1})} \quad k = 1, 2, \ldots \,. \tag{4.39}$$

Die Iterationsvorschriften (4.38) bzw. (4.39) sind bei genauerem Hinsehen nicht anderes als Fixpunktiterationen der Abbildung bzw. Funktion

$$G(x) = x - [F'(x)]^{-1}F(x) \quad \text{bzw.} \quad g(x) = x - \frac{f(x)}{f'(x)} \; .$$

Im Falle der Konvergenz der Folgen (4.38) bzw. (4.39) hat man damit Lösungen (Nullstellen) bestimmt. Diese Verfahren heißen **Newtonverfahren** und mit deren Eigenschaften und Voraussetzungen für die Konvergenz von Newtonfolgen wollen wir uns im Folgenden beschäftigen.

9. Vorlesung 14.11.11

**Satz 4.24.** *Es sei $G \subset \mathbb{R}$ offen und $f : G \to \mathbb{R}$ stetig diff'bar mit einem Fixpunkt $\hat{x} \in G$. Wenn $|f'(\hat{x})| < 1$ gilt, dann existiert ein abgeschlossenes Intervall $D \subset G$ mit $\hat{x} \in D$ und $f(D) \subset D$, auf dem $f$ eine Kontraktion ist.*

*Beweis.* Da $f'$ stetig auf der offenen Menge $G$ ist, existiert eine offene Umgebung $K_{\hat{x},\epsilon} = \{x| \, |x - \hat{x}| < \epsilon\}$ in $G$, auf der die Beträge der Ableitung von $f$ immer noch kleiner als 1 sind. Setzt man $D = [\hat{x} - \frac{\epsilon}{2}, \hat{x} + \frac{\epsilon}{2}]$, so gilt für alle $x_1, x_2 \in D$ aufgrund des Mittelwertsatzes der Differentialrechnung

$$|f(x_1) - f(x_2)| \leq k \, |x_1 - x_2|$$

mit $k = \max_{\xi \in D} |f'(\xi)| < 1$ □

**Bemerkung 4.25.** Ist die Voraussetzung $|f'(\hat{x})| < 1$ des Satzes 4.24 nicht erfüllt, findet man keine Kontraktion. Ist $|f'(\hat{x})| > 1$ dann gilt in der Nähe von $\hat{x}$

$$|f(x) - f(\hat{x})| > |x - \hat{x}|$$

das rechtfertigt die

**Definition 4.26.** *Ein Fixpunkt $\hat{x}$ heißt anziehender Fixpunkt, wenn $|f'(\hat{x})| < 1$ gilt, und $\hat{x}$ heißt abstoßender Fixpunkt, wenn $|f'(\hat{x})| > 1$ ist.*

## 4.7 Das Newton-Verfahren zur Lösung nichtlinearer Gleichungen

Die Grundidee der Lösung der nichtlinearen Gleichung besteht in der Näherung einer existierenden Nullstelle durch die sukzessive Lösung linearer Aufgaben. Man approximiert die diff'bare Funktion $f$ in der Nähe eines geeigneten Startwertes $x_0$ durch die Tangentenfunktion

$$g(x) = f(x_0) + f'(x_0)(x - x_0) \approx f(x)$$

und bestimmt die Nullstelle von $g$, also $x_1$ mit $g(x_1) = 0$, d.h.

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

und allgemein erhält man mit

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, 2, \ldots \tag{4.40}$$

eine Newton-Folge, die im Falle der Konvergenz gegen eine Nullstelle von $f$ geht. Die Folge (4.40) kann man auch anders erhalten. Man definiert die Hilfsfunktion

$$g(x) = x - \frac{f(x)}{f'(x)} \tag{4.41}$$

wobei man $f'(x) \neq 0$ voraussetzt. Ist $g$ eine Kontraktion mit $g(I) \subset I$, dann folgt aus dem Banachschen Fixpunktsatz, dass die Folge $x_{k+1} = g(x_k)$ für einen beliebigen Startwert $x_0 \in I$ (abgeschlossenes Intervall) gegen den in $I$ existierenden Fixpunkt $\hat{x}$ von $g$ konvergiert. Und die Fixpunkt-Folge

$$x_{k+1} = g(x_k) = x_k - \frac{f(x_k)}{f'(x_k)}$$

ist eine Newton-Folge zur Berechnung einer Nullstelle von $f$. Es gilt der folgende

**Satz 4.27.** *Sei $f : I \to \mathbb{R}$ eine auf einem Intervall $I \supset [x_0 - r, x_0 + r], r > 0$, definierte, zweimal stetig diff'bare Funktion mit $f'(x) \neq 0$ für alle $x \in I$. Weiterhin existiere eine reelle Zahl $k$, $0 < k < 1$, mit*

$$\left| \frac{f(x)f''(x)}{f'(x)^2} \right| \leq k \quad \forall x \in I$$

*und*

$$\left| \frac{f(x_0)}{f'(x_0)} \right| \leq (1-k)r$$

*Dann hat $f$ genau eine Nullstelle $\hat{x} \in I$ und die Newton-Folge (4.40) konvergiert quadratisch gegen $\hat{x}$, d.h. es gilt*

$$|x_{k+1} - \hat{x}| \leq C(x_k - \hat{x})^2 \quad \forall k = 0, 1, \ldots$$

*mit einer Konstanten $C$. Außerdem gilt die Fehlerabschätzung*

$$|x_k - \hat{x}| \leq \frac{|f(x_k)|}{M}, \quad \textit{mit} \quad 0 < M < \min_{x \in I} |f'(x)|$$

*Beweis.* Folgt aus dem Banachschen Fixpunktsatz und dem Satz 4.24 über die Existenz einer Kontraktion.

Die quadratische Konvergenz folgt aus

$$\begin{aligned} x_{k+1} - \hat{x} &= x_k - \frac{f(x_k)}{f'(x_k)} - \hat{x} \\ &= x_k - \hat{x} - \frac{f(x_k) - \overbrace{f(\hat{x})}^{=0}}{f'(x_k)} \\ &= \frac{1}{f'(x_k)} \underbrace{[f'(x_k)(x_k - \hat{x}) - f(x_k) + f(\hat{x})]}_{\text{Fehler der Ordnung } \mathcal{O}(|x_k - \hat{x}|^2)} \\ \Rightarrow \quad |x_{k+1} - \hat{x}| &\leq C(x_k - \hat{x})^2 \end{aligned}$$

□

**Bemerkung 4.28.** Die Voraussetzungen des Satzes 4.27 garantieren die Kontraktivität der Hilfsfunktion $g(x) = x - \frac{f(x)}{f'(x)}$ in einer Umgebung von $\hat{x}$. D.h. man ist mit dem Newton-Verfahren immer dann erfolgreich, wenn man nur nah genug an der Nullstelle die Iteration beginnt ($x_0$ nah bei $\hat{x}$)
In diesem Fall ist das Newton-Verfahren auch noch sehr schnell aufgrund der quadratischen Konvergenz.

**Satz 4.29.** *(Nullstelle einer konvexen Funktion)*
*Sei $f : [a, b] \to \mathbb{R}$ zweimal stetig differenzierbar und konvex ($f'(x) \neq 0$ auf $[a, b]$). Die Vorzeichen von $f(a)$ und $f(b)$ seien verschieden. Dann konvergiert die Newton-Folge (4.40) von $f$ für $x_0 = a$, falls $f(a) > 0$ und für $x_0 = b$, falls $f(b) > 0$, gegen die einzige Nullstelle $\bar{x}$ von $f$.*

## 4.8 Newton-Verfahren für nichtlineare Gleichungssysteme $F(x) = 0$

**Satz 4.30.** *$F : D \to \mathbb{R}^n, D \subset \mathbb{R}^n$ sei zweimal stetig partiell diff'bar und besitze eine Nullstelle $\hat{x} \in D$. Weiterhin sei $F'(x)$ für jedes $x \in D$ regulär.*
*Dann folgt:*
*Es gibt eine Umgebung $U$ von $\hat{x}$, sodass die Newton-Folge*

$$x_{k+1} = x_k - [F'(x_k)]^{-1} F(x_k), \quad k = 0, 1, 2, \ldots \tag{4.42}$$

*von einem beliebigen Startpunkt $x_0 \in U$ ausgehend gegen die Nullstelle $\hat{x}$ konvergiert.*
*Die Konvergenz ist quadratisch, d.h. es gibt eine Konstante $C > 0$ mit*

$$\|x_k - \hat{x}\| \leq C \|x_{k-1} - \hat{x}\|^2, \quad k = 1, 2, \ldots$$

*und es gilt die Fehlerabschätzung*

$$\|x_k - \hat{x}\| \leq \|F(x_k)\| \sup_{x \in D} \left\| [F'(x)]^{-1} \right\|$$

*wobei auf der rechten Seite die Matrixnorm* $\|A\| = \sqrt{\sum_{i,j=1}^{n} a_{ij}^2}$ *für eine* $(n \times n)$*-Matrix* $A = (a_{ij})$ *verwendet wurde.*

*Beweis.* Analog zum Beweis von Satz 4.27 im eindimensionalen Fall. □

Es gibt Situationen, da **schießt man über das Ziel hinaus**, d.h. man mach zu große Schritte. In diesen Fällen (also wenn das Newton-Verfahren nicht konvergiert) kann man versuchen, die Schritte zu dämpfen.
Man betrachtet

$$x_{k+1} = x_k - \alpha [F'(x_k)]^{-1} F(x_k), \quad k = 0, 1, \ldots$$

mit $\alpha \in ]0, 1[$, und spricht hier von einem **gedämpften Newton- Verfahren**.
Mit gedämpften Newton-Verfahren erreicht man mitunter Konvergenz der Newton-Folge, wenn das Standard-Newton-Verfahren ($\alpha = 1$) versagt.
Es gilt dann mit $z_{k+1} = x_{k+1} - x_k$ das Gleichungssystem

$$F'(x_k) z_{k+1} = -\alpha F(x_k)$$

zu lösen, und wie üblich erhält man mit

$$x_{k+1} = z_{k+1} + x_k$$

die neue Iterierte.

## 4.9 Sekantenverfahren – Regula falsi

Beim eben dargelegten Newton-Verfahren war die Differenzierbarkeit der Funktion $f$ entscheidend für die Konstruktion des numerischen Lösungsverfahrens. Allerdings ist die Differenzierbarkeit nicht notwendig für die Existenz einer Nullstelle. Wenn für die auf dem Intervall $[a, b]$ stetige Funktion $f$ die Bedingung $f(a)f(b) < 0$ erfüllt ist, dann existiert nach dem Zwischenwertsatz auf jeden Fall mindestens eine Nullstelle $\bar{x} \in ]a, b[$. Diese findet man auf jeden Fall mit dem Bisektionsverfahren (auch Intervallhalbierungsverfahren genannt). Wir setzen $a_0 = a$ und $b_0 = b$. Für den Mittelpunkt $x_1 = \frac{a_0 + b_0}{2}$ (s. Abb. 4.1) gilt auf jeden Fall

$$|x_1 - \bar{x}| \leq \frac{b_0 - a_0}{2} \ .$$

Geht man nun von einer Näherung $x_k$ als Mittelpunkt des Intervalls $[a_{k-1}, b_{k-1}]$ für die Nullstelle $\bar{x}$ aus, dann setzt man im Fall $f(x_k) \neq 0$ (anderenfalls ist man fertig und hat mit $x_k$ eine Nullstelle gefunden)

$$\begin{aligned} a_k &= \begin{cases} a_{k-1} & \text{falls } f(x_k)f(a_{k-1}) < 0 \\ x_k & \text{falls } f(x_k)f(b_{k-1}) < 0 \end{cases}, \\ b_k &= \begin{cases} x_k & \text{falls } f(x_k)f(a_{k-1}) < 0 \\ b_{k-1} & \text{falls } f(x_k)f(b_{k-1}) < 0 \end{cases}, \\ x_{k+1} &= \begin{cases} \frac{a_{k-1}+x_k}{2} & \text{falls } f(x_k)f(a_{k-1}) < 0 \\ \frac{x_k+b_{k-1}}{2} & \text{falls } f(x_k)f(b_{k-1}) < 0 \end{cases} \end{aligned} \tag{4.43}$$

und erhält für den Mittelpunkt $x_{k+1}$ des Intervalls $[a_k, b_k]$ die Abschätzung

$$|x_{k+1} - \bar{x}| \leq \frac{b_0 - a_0}{2^{k+1}} \quad \text{bzw.} \quad |x_{k+1} - \bar{x}| \leq \frac{1}{2}|x_k - \bar{x}| \leq \cdots \leq \frac{1}{2^{k+1}}|x_0 - \bar{x}| \,,$$

d.h., aufgrund von $|x_{k+1} - \bar{x}| \leq \frac{1}{2}|x_k - \bar{x}|^p$ mit $p = 1$ ist die Konvergenzordnung des Verfahrens gleich 1.
Der aufwendig aussehende Algorithmus (4.43) lässt sich recht einfach implementieren. Der Vorteil des Bisektionsverfahrens besteht darin, dass es immer funktioniert, d.h., man findet immer eine Nullstelle. Allerdings findet man mit dem beschriebenen Verfahren nur eine Nullstelle.

**Satz 4.31.** *(Bisektionsverfahren)*
*Sei $f$ eine auf $[a, b]$ stetige Funktion mit $f(a)f(b) < 0$. Mit $x_0 = a$ konvergiert das Bisektionsverfahren (4.43) gegen eine Nullstelle $\bar{x}$ der Funktion $f$. Die Konvergenzordnung ist* 1. *Es gilt*

$$|x_k - \bar{x}| \leq \frac{1}{2^k}|x_0 - \bar{x}| = e^{-\gamma k}|x_0 - \bar{x}|$$

*mit dem Konvergenzexponenten $\gamma = \ln 2$.*

Eine weitere Möglichkeit der Nullstellenbestimmung ohne die Nutzung der Ableitung der Funktion $f$ ist das Sekanten-Verfahren, das auch Regula falsi genannt wird. Man geht von 2 Näherungen $x_k$, $x_{k-1}$ aus dem Intervall $[a, b]$ mit $f(a)f(b) < 0$ aus und führt statt des Newton-Schrittes $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$ den Schritt

$$x_{k+1} = x_k - \frac{f(x_k)}{\frac{f(x_k)-f(x_{k-1})}{x_k - x_{k-1}}} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k) \tag{4.44}$$

aus. Den Schritt (4.44) kann man einmal als genäherten Newton-Schritt mit der Approximation von $f'(x_k)$ durch den Differenzenquotienten $\frac{f(x_k)-f(x_{k-1})}{x_k - x_{k-1}}$

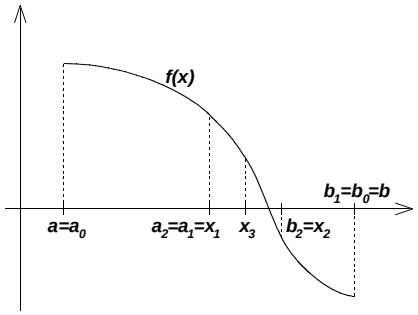



Abbildung 4.1: Bisektions-Verfahren

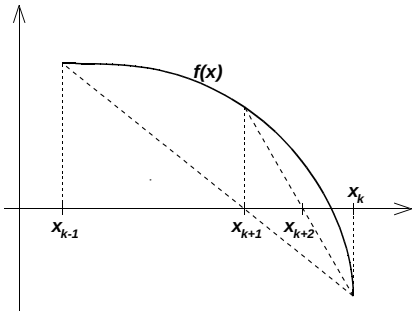



Abbildung 4.2: Sekantenverfahren − Regula falsi

interpretieren. Andererseits bedeutet (4.44) geometrisch die Berechnung des Schnittpunktes der Sekante durch die Punkte $(x_{k-1}, f(x_{k-1}))$ und $(x_k, f(x_k))$ mit der $x$-Achse (s. Abb. 4.2). Beide Interpretationen erklären die Namen "Regula falsi" bzw. "Sekantenverfahren". Das Sekantenverfahren hat den Nachteil, dass die neue Näherung $x_{k+1}$ nicht unbedingt im Intervall $[a, b]$ liegen muss. Das ist nur der Fall, wenn $f(x_k)f(x_{k-1}) < 0$ gilt. Die Modifikation von (4.44) in der Form

$$x_{k+1} = x_k - \frac{f(x_k)}{\frac{f(x_k)-f(x_j)}{x_k-x_j}} = x_k - \frac{x_k - x_j}{f(x_k) - f(x_j)} f(x_k) \tag{4.45}$$

mit $j \leq k-1$ als größtem Index mit $f(x_k)f(x_j) < 0$ ergibt in jedem Fall eine Näherung $x_{k+1} \in [a, b]$, wenn die Startwerte $x_0, x_1 \in [a, b]$ die Eigenschaft $f(x_0)f(x_1) < 0$ haben.
Die Modifikation (4.45) bedeutet gegenüber (4.44) nur den geringen Mehraufwand der Ermittlung des Index $j$. Es gilt der

**Satz 4.32.** *(Sekantenverfahren)*
*Sei $f$ eine auf $[a, b]$ stetige Funktion mit $f(a)f(b) < 0$ und $x_0, x_1 \in [a, b]$ Startwerte mit der Eigenschaft $f(x_0)f(x_1) < 0$. Dann konvergiert das Sekantenverfahren (4.45) gegen eine Nullstelle $\bar{x}$ der Funktion $f$.*

Das Sekantenverfahren (4.44) konvergiert lokal in einer Umgebung um $\bar{x}$ mit einer Konvergenzordnung $p > 1$ schneller als das Bisektionsverfahren. Für eine um $\bar{x}$ zweimal stetig differenzierbare Funktion $f$ mit $f'(\bar{x})$ soll das gezeigt werden. In der Nähe von $\bar{x}$ gilt

$$f(x) \approx f(\bar{x}) + f'(\bar{x})(x-\bar{x}) + \frac{1}{2}f''(\bar{x})(x-\bar{x})^2 = f'(\bar{x})(x-\bar{x}) + \frac{1}{2}f''(\bar{x})(x-\bar{x})^2$$

und mit $\Delta_k = x_k - \bar{x}$ folgt aus (4.44)

$$\Delta_{k+1} \approx \Delta_k - \frac{\Delta_k - \Delta_{k-1}}{f'(\bar{x})(\Delta_k - \Delta_{k-1}) + \frac{1}{2}f''(\bar{x})(\Delta_k^2 - \Delta_{k-1}^2)}(f'(\bar{x})\Delta_k + \frac{1}{2}f''(\bar{x})\Delta_k^2) \ .$$

Mit der realistischen Voraussetzung $|\Delta_k|, |\Delta_{k-1}| \ll 1$ gilt

$$\begin{aligned} \Delta_{k+1} &\approx \Delta_k - \frac{f'(\bar{x})\Delta_k + \frac{1}{2}f''(\bar{x})\Delta_k^2}{f'(\bar{x}) + \frac{1}{2}f''(\bar{x})(\Delta_k + \Delta_{k-1})} \\ &= \frac{\frac{1}{2}f''(\bar{x})\Delta_k\Delta_{k-1}}{f'(\bar{x}) + \frac{1}{2}f''(\bar{x})(\Delta_k + \Delta_{k+1})} \approx \frac{\frac{1}{2}f''(\bar{x})\Delta_k\Delta_{k-1}}{f'(\bar{x})} \ , \end{aligned}$$

d.h.

$$|\Delta_{k+1}| \approx |\frac{f''(\bar{x})}{2f'(\bar{x})}| \, |\Delta_k| \, |\Delta_{k-1}| =: c|\Delta_k| \, |\Delta_{k-1}| \ . \tag{4.46}$$

Mit der Zahl $p > 0$, die der Gleichung $p - 1 = \frac{1}{p}$ genügt, also $p = \frac{1+\sqrt{5}}{2} \approx 1,618$, folgt aus (4.46) für $\tilde{\Delta}_k = c|\Delta_k|$

$$\tilde{\Delta}_{k+1} = \tilde{\Delta}_k\tilde{\Delta}_{k-1} \quad \text{und} \quad \frac{\tilde{\Delta}_{k+1}}{\tilde{\Delta}_k^p} = [\frac{\tilde{\Delta}_{k-1}^p}{\tilde{\Delta}_k}]^{1/p} \ . \tag{4.47}$$

Mit $a_k := \ln \frac{\tilde{\Delta}_{k+1}}{\tilde{\Delta}_k^p}$ erhält man durch Logarithmieren von (4.47) mit $a_k = -a_{k-1}/p$ $(k = 1, 2, \dots)$ eine Folge, die für jeden Startwert $a_0$ wegen $p > 1$ gegen null konvergiert. Daraus folgt aufgrund der Definition von $a_k$ und der Stetigkeit der ln-Funktion

$$\lim_{k\to\infty} a_k = \lim_{k\to\infty} \ln \frac{\tilde{\Delta}_{k+1}}{\tilde{\Delta}_k^p} = 0 = \ln 1 \iff \lim_{k\to\infty} \frac{\tilde{\Delta}_{k+1}}{\tilde{\Delta}_k^p} = 1 \iff \lim_{k\to\infty} \frac{|\Delta_{k+1}|}{|\Delta_k|^p} = c^{p-1}.$$

Schließlich erhalten wir mit

$$|\Delta_{k+1}| \approx c^{p-1}|\Delta_k| \iff |x_{k+1} - \bar{x}| \approx c^{p-1}|x_k - \bar{x}|^p$$

die lokale Konvergenzordnung $p \approx 1,618$ des Sekantenverfahrens.

**Beispiel 4.33.** Zur Berechnung der Nullstelle der Funktion $f(x) = \cos x - x$ erhält man die Nullstelle $\bar{x} \approx 0,73909$ mit einer Genauigkeit von $10^{-10}$ mit 30 Iterationen des Bisektionsverfahrens, 5 Iterationen des Sekantenverfahrens und 4 Iterationen des Newton-Verfahrens bei der Wahl von $x_0 = 2$, $x_1 = 0$. Hinsichtlich der Funktionsauswertungen sind beim Bisektions- und Newton-Verfahren pro Schritt je 2 Funktionswertberechnungen erforderlich, während beim Sekantenverfahren nur eine nötig ist.

# Kapitel 5

# Orthogonale Matrizen – QR-Zerlegung – Ausgleichsprobleme

10. Vorlesung am 16.11.2011

Im Folgenden soll für eine gegebene Matrix $A \in \mathbb{R}^{n\times m}, 1 \leq m \leq n$, eine Faktorisierung der Form

$$A = QS \tag{5.1}$$

bestimmt werden mit einer orthogonalen Matrix $Q$, d.h.

$$Q \in \mathbb{R}^{n\times n}, Q^{-1} = Q^T$$

und einer verallgemeinerten oberen Dreiecksmatrix

$$S = \begin{bmatrix} R \\ - \\ 0 \end{bmatrix} \in \mathbb{R}^{n\times m}, R = \begin{bmatrix} * & & * \\ & \ddots & \\ 0 & & * \end{bmatrix} \in \mathbb{R}^{m\times m} \tag{5.2}$$

Solche Zerlegungen ermöglichen z.B. die stabile Lösung von schlecht konditionierten lösbaren linearen Gleichungssystemen $Ax = b, (m = n)$ oder die stabile Lösung von Ausgleichsproblemen mit $M \in \mathbb{R}^{n\times m}, 1 \leq m \leq n$,

$$\min_{r\in\mathbb{R}^m} \frac{1}{2} \|Mr - b\|_2^2 \ . \tag{5.3}$$

Hier gibt es 2 Möglichkeiten der Lösung.
Wenn man das Funktional

$$F(r) = \frac{1}{2} \|Mr - b\|_2^2 = \frac{1}{2}\langle Mr - b, Mr - b\rangle_2$$

betrachtet, findet man mit der Darstellung

$$F(r+h) = F(r) + \langle M^T Mr - M^T b, h\rangle_2 + \frac{1}{2}\langle M^T Mh, h\rangle_2$$

mit

$$\nabla F(r) = M^T Mr - M^T b \quad \text{und} \quad F''(r) = M^T M$$

Gradient und Hessematrix. Die Auswertung der notwendigen Extremalbedingung

$$\nabla F(r) = \mathbf{0} \Longleftrightarrow M^T Mr = M^T b$$

liefert mit

$$r = [M^T M]^{-1} M^T b$$

die Lösung des Minimumproblems für den Fall, dass die Matrix $M$ den vollen Rang hat, denn dann ist die Hessematrix $M^T M$ symmetrisch und positiv definit.

Eine zweite Möglichkeit der Lösung des Minimuproblems (5.3) ist mit einer $QR$-Zerlegung von $M$ machbar. Dies soll nun im Folgenden besprochen werden. Wir erinnern uns an die Eigenschaften orthogonaler Matrizen:

(i)
$$\|Qx\|_2 = \|x\|_2 = \left\|Q^T x\right\|_2, x \in \mathbb{R}^n \tag{5.4}$$

(ii)
$$\text{cond}(QA) = \text{cond}(A) \tag{5.5}$$

(iii) für $Q_1, Q_2 \in \mathbb{R}^{n\times n}$ orthogonal, gilt $Q_1 Q_2$ ist orthogonal

Faktorisierung $A = QR$ mittels Gram-Schmidt-Orthogonalisierung. Für quadratische reguläre Matrizen $A$ $(m = n)$ hat (5.1), (5.2) die Form

$$A = QR \tag{5.6}$$

mit $Q$ orthogonal und $R$ oberer Dreiecksmatrix vom Typ $(n \times n)$. Schreiben $A, Q, R$ in der Form

$$A = [a_1|a_2|\ldots|a_n], \quad Q = [q_1|\ldots|q_n], \quad R = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{pmatrix}$$

Mit den Spaltenvektoren $a_k, q_k \in \mathbb{R}^n, k = 1, \ldots, n$. (5.6) bedeutet dann

$$a_j = \sum_{i=1}^{j} r_{ij} q_i, \quad j = 1, ..., n, q_1, \ldots, q_n \in \mathbb{R}^n \tag{5.7}$$

## 5.1 Gram-Schmidt-Verfahren zur Orthogonalisierung

(a) Ausgangspunkt: man hat $j-1$ orthonormale Vektoren $q_1, \ldots, q_{j-1} \in \mathbb{R}^n$ mit $\text{span}(a_1, \ldots, a_{j-1}) = \text{span}(q_1, \ldots, q_{j-1}) =: M_{j-1}$

(b) man bestimmt im Schritt $j \geq 1$ das Lot von $a_j$ auf den linearen Unterraum $M_{j-1} \subset \mathbb{R}^n$

$$\hat{q}_j := a_j - \sum_{i=1}^{j-1} \langle a_j, q_i \rangle \, q_i \tag{5.8}$$

Und nach der Normierung

$$q_j = \frac{\hat{q}_j}{\|\hat{q}_j\|}$$

Sind die Vektoren $q_1, \ldots, q_j \in \mathbb{R}^n$ paarweise orhonormal und es gilt

$$\text{span}(a_1, \ldots, a_j) = \text{span}(q_1, \ldots, a_j)$$

Aus der Gleichung (5.8) folgt

$$a_j = \underbrace{\|\hat{q}_j\|_2}_{r_{jj}} q_j + \sum_{i=1}^{j-1} \underbrace{(a_j^T q_i)}_{r_{ij}} q_i = \sum_{i=1}^{j} r_{ij} q_i, \quad j = 1, \ldots, n \tag{5.9}$$

Nach Abschluss der Gram-Schmidt-Orthogonalisierung hat man damit mit (5.9)

$$[a_1|a_2|\ldots|a_n] = [q_1|\ldots|q_n] \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ & & \ddots & \vdots \\ & & & r_{nn} \end{bmatrix}$$

als QR-Zerlegung

**Bemerkung 5.1.** Der beschriebene Algorithmus kann im ungünstigsten Fall Probleme bereiten (nicht gutartig sein), wenn z.B. $\|\hat{q}_j\|$ recht klein wird (Lösung folgt).

## 5.2 Householder-Matrizen/Transformationen

**Definition 5.2.** *Eine Abbildung $H : \mathbb{R}^n \to \mathbb{R}^n, x \mapsto Hx$ mit einer Matrix*

$$H = E - 2ww^T, w \in \mathbb{R}^n, \|w\|_2^2 = w^T w = 1 \tag{5.10}$$

*bezeichnet man als Householder-Transformation und $H$ als Householder-Matrix*

Eigenschaften von $H$:

- $H^T = H$ Symmetrie
- $H^2 = E$ $H$ ist involutorisch
- $H^T H = E$ Orthogonalität

Nachweis als Übung

Wirkung der Householder-Transformation:
Spiegelung von $x \in \mathbb{R}^n$ an der Hyperebene $\{z \in \mathbb{R}^n : z^T w = 0\}$, da die Identität

$$Hx = x - 2(w^T x)w = x - (w^T x)w - (w^T x)w$$

gilt.

**Lemma 5.3.** *Gegeben sei $0 \neq x \in \mathbb{R}^n$ mit $x \notin span\{e_1\}$. Für*

$$w = \frac{x + \sigma e_1}{\|x + \sigma e_1\|_2} \quad mit \quad \sigma = \pm \|x\|_2 \tag{5.11}$$

*gilt*

$$\|w\| = 1, \quad Hx = (E - 2ww^T)x = -\sigma e_1 \tag{5.12}$$

*Beweis.* $\|w\| = 1$ weil $x + \sigma e_1 \neq 0$ und damit (5.11) wohldefiniert ist. Für den Nachweis von (5.12) erhält man

$$\|x + \sigma e_1\|_2^2 = \|x\|_2^2 + 2\sigma e_1^T x + \sigma^2 = \|x\|^2 + 2\sigma e_1^T x + \|x\|^2 = 2(x + \sigma e_1)^T x$$

Und mit (5.11), d.h. $\frac{(x+\sigma e_1)^T}{\|x+\sigma e_1\|_2} = w^T$ folgt:

$$2w^T x = \frac{2(x + \sigma e_1)^T x}{\|x + \sigma e_1\|_2} = \|x + \sigma e_1\|_2$$

die nochmalige Nutzung von (5.11) ergibt

$$\begin{aligned} 2ww^T x &= x + \sigma e_1 \\ \Leftrightarrow x - 2ww^T x &= -\sigma e_1 \end{aligned}$$

was zu zeigen war. □

**Bemerkung.** Um Stellenauslöschungen zu vermeiden, wird in (5.11) $\sigma = \operatorname{sgn}(x_1)\|x\|_2$ gewählt, d.h. z.B. für $x = (-3, 1, 5)^T$ ist $\sigma = -\sqrt{35}$

## 5.3 Algorithmus zur Konstruktion der Faktorisierung mittels Householder-Transformationen

Ausgehend von $A = A^{(1)} \in \mathbb{R}^{n\times m}$ sollen sukzessive Matrizen der Form

$$A^{(j)} = \begin{bmatrix} a_{11}^{(j)} & a_{12}^{(j)} & & \cdots & & a_{1m}^{(j)} \\ & \ddots & & & & \\ & & a_{j-1j-1}^{(j)} & & & a_{j-1m}^{(j)} \\ & & & a_{jj}^{(j)} & \cdots & a_{jm}^{(j)} \\ & & & \vdots & & \vdots \\ & & & a_{nj}^{(j)} & \cdots & a_{nn}^{(j)} \end{bmatrix}, \quad j = 1, \ldots, m \qquad (5.13)$$

berechnet werden, sodass am Ende mit $A^{(m+1)} = S$ die verallgemeinerte obere Dreiecksmatrix vorliegt.
Die Matrizen der Form (5.13) erhält man für $j = 1, \ldots, m-1$ durch Transformationen der Form

$$A^{(j+1)} = \hat{H}_j A^{(j)}, \quad \hat{H}_j = \left[\begin{array}{c|c} E_{j-1} & 0 \\ \hline 0 & H_j \end{array}\right]$$

mit $H_j = E_{n-(j-1)} - 2w_j w_j^T, \|w_j\| = 1$, $E_l$ ist Einheitsmatrix aus $\mathbb{R}^{l\times l}$, und $w_j \in \mathbb{R}^{n-(j-1)}$ ist so zu wählen, dass gilt

$$H_j \underbrace{\begin{pmatrix} a_{jj}^{(j)} \\ \vdots \\ a_{nj}^{(j)} \end{pmatrix}}_{\mathfrak{a}} = \sigma \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, w_j = \frac{\mathfrak{a} + \operatorname{sgn}(\mathfrak{a}_1) \|\mathfrak{a}\|_2 e_1}{\|\mathfrak{a} + \operatorname{sgn}(\mathfrak{a}_1) \|\mathfrak{a}\|_2 e_1\|_2}, \mathfrak{a}, e \in \mathbb{R}^{n-(j-1)}$$

Die Matrizen $\hat{H}_1, \ldots, \hat{H}_{m-1}$ sind aufgrund der Eigenschaften der Matrizen $H_1, \ldots, H_{m-1}$ orthogonal und symmetrisch, sodass man mit

$$S = \hat{H}_{m-1}\hat{H}_{m-2} \cdot \cdots \cdot \hat{H}_1 A, \quad Q = \hat{H}_1\hat{H}_2 \cdot \cdots \cdot \hat{H}_{m-1}$$

die Faktorisierung $A = QS$ konstruiert hat, da $Q$ als Produkt von orthogonalen Matrizen auch eine orthogonale Matrix ist.

Hat man mit Blick auf das Minimumproblem (5.3) nun eine $QR$-Zerlegung von $M$ mit

$$M = QS\,, \qquad S = \begin{bmatrix} R \\ - \\ 0 \end{bmatrix}$$

und orthogonalem $Q$ gegeben, dann ergibt sich mit

$$Q^T b =: \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad b_1 \in \mathbb{R}^m,\ b_2 \in \mathbb{R}^{n-m}$$

durch Nutzung der Eigenschaften von $Q$

$$\begin{aligned} \frac{1}{2}\|Mr - b\|_2^2 &= \frac{1}{2}\|QSr - b\|_2^2 \\ &= \frac{1}{2}\left\|Q^T(QSr - b)\right\|_2^2 \\ &= \frac{1}{2}\left\|Sr - Q^T b\right\|_2^2 \\ &= \frac{1}{2}[\left\|\begin{bmatrix} Rr - b_1 \\ -b_2 \end{bmatrix}\right\|_2^2 \\ &= \frac{1}{2}[\|Rr - b_1\|_2^2 + \|b_2\|_2^2 \\ &\geq \frac{1}{2}\|b_2\|_2^2 \end{aligned}$$

und damit wird das Minimum von $F(r)$ für $r = R^{-1}b_1$ mit

$$\min_{r\in\mathbb{R}^m} \frac{1}{2}\|Mr - b\|_2^2 = \frac{1}{2}\|b_2\|_2^2$$

angenommen, wobei wir hier vorausgesetzt haben, dass $M$ den vollen Rang hat und damit $R$ invertierbar ist.

## 5.4 Gauß-Newton-Verfahren

11. Vorlesung am 21.11.2011

Wir kommen auf die Thematik "Ausgleichsprobleme" zurück. Gegeben sind "Messwerte"

$$x_{i,1}, x_{i,2}, \ldots, x_{i,n} \quad \text{und} \quad y_i \ , \ i = 1, \ldots, k,$$

und gesucht ist ein funktionaler Zusammenhang

$$f(x_1, x_2, \ldots, x_n; a_1, a_2, \ldots, a_p) = y$$

wobei solche Parameter

$$a = (a_1, a_2, \ldots, a_p)$$

gesucht sind, dass das Residuum bzw. die Länge des Residuenvektors $R$

$$R(a) = \begin{pmatrix} f(x_{1,1}, x_{1,2}, \ldots, x_{1,n}; a_1, a_2, \ldots, a_p) - y_1 \\ f(x_{2,1}, x_{2,2}, \ldots, x_{2,n}; a_1, a_2, \ldots, a_p) - y_2 \\ \vdots \\ f(x_{k,1}, x_{k,2}, \ldots, x_{k,n}; a_1, a_2, \ldots, a_p) - y_k \end{pmatrix} =: \begin{pmatrix} r_1(a) \\ r_2(a) \\ \vdots \\ r_k(a) \end{pmatrix}$$

minimal wird. Wir setzen $k \geq p$ voraus. $R$ ist eine Abbildung von $\mathbb{R}^p$ nach $\mathbb{R}^k$.
Wir betrachten den allgemeinen Fall, dass $R$ nichtlinear von $a$ abhängt. Zu lösen ist das Minimum-Problem

$$\min_{a \in \mathbb{R}^p} F(a) \quad \text{für } F(a) = ||R(a)||_2^2 \ .$$

Man geht von einer Näherung $a^{(i)}$ von $a$ aus. Die lineare Approximation von $R$ an der Entwicklungsstelle $a^{(i)}$ ergibt

$$R(a) \approx R(a^{(i)}) + R'(a^{(i)})(a - a^{(i)}) \ ,$$

wobei $R'$ die Ableitung der Abbildung $R$, also die Matrix der partiellen Ableitungen

$$R' = (r'_{ji}) = (\frac{\partial r_j}{\partial a_i}) \ , \quad j = 1, \ldots, k, \ i = 1, \ldots, p \ ,$$

ist. Durch die Lösung $a^*$ des Minimum-Problems

$$\min_{a \in \mathbb{R}^p} ||R(a^{(i)}) + R'(a^{(i)})(a - a^{(i)})||_2^2 \ , \tag{5.14}$$

bestimmt man eine neue Näherung

$$a^{(i+1)} = \alpha a^* + (1 - \alpha) a^{(i)} \ ,$$

wobei man mit $\alpha \in ]0,1]$ die Möglichkeit einer Dämpfung (Relaxation) hat. In vielen Fällen hat man im ungedämpften Fall mit $\alpha = 1$ keine oder nur eine sehr langsame Konvergenz, während man bei geeigneter Wahl von $0 < \alpha < 1$ eine konvergente Folge erhält.
Schreibt man

$$R(a^{(i)}) + R'(a^{(i)})(a - a^{(i)}) \tag{5.15}$$

in der Form

$$Ma - y$$

mit

$$M = R'(a^{(i)}) \ , \quad y = R'(a^{(i)})a^{(i)} - R(a^{(i)})$$

auf, dann kann man die Lösung $a^*$ von

$$\min_{a \in \mathbb{R}^p} ||Ma - y||_2^2$$

entweder mit einer $QR$-Zerlegung von $M$ bestimmen, oder durch die Lösung des Normalgleichungssystems

$$M^T Ma = M^T y \Longleftrightarrow a = [M^T M]^{-1} M^T y$$

erhalten (s. auch vorangegangene Vorlesungen).
Aus Effizienzgründen (jeweiliger Aufbau von $y$) schreibt man (5.15) auch in der Form

$$Ms - \hat{y}$$

mit

$$s = a - a^{(i)} \quad \text{und} \quad \hat{y} = -R(a^{(i)})$$

auf und löst das Minimum-Problem

$$\min_{s \in \mathbb{R}^p} ||Ms - \hat{y}||_2^2$$

und berechnet durch

$$a^* = s + a^{(i)} \qquad a^{(i+1)} = \alpha a^* + (1 - \alpha)a^{(i)}$$

die neue Näherung.
Für den Fall $\alpha = 1$ (keine Dämpfung) bedeutet das Gauß-Newton-Verfahren nichts Anderes als die Fixpunktiteration

$$a^{(i+1)} = a^{(i)} - [R'(a^{(i)})^T R'(a^{(i)})]^{-1} R'(a^{(i)})^T R(a^{(i)}) \ ,$$

und bei Konvergenz gegen $a^*$ hat man (unter der Voraussetzung der Regularität von $[R'^T R']^{-1}$) die Bedingung

$$R'(a^*)^T R(a^*) = \mathbf{0} \tag{5.16}$$

erfüllt. Wenn man den Gradienten von $F(a)$ ausrechnet stellt man fest, dass die Bedingung (5.16) wegen

$$\text{grad}_{a^*} F = 2R'(a^*)^T R(a^*)$$

äquivalent zur notwendigen Extremalbedingung

$$\text{grad}_{a^*} F = \mathbf{0}$$

für das Funktional $F$ ist.
Für die Abbruchbedingung gibt man eine Genauigkeit $\epsilon$ vor und bricht die Iteration dann ab, wenn

$$||a^{(i+1)} - a^{(i)}||_2 < \epsilon$$

erfüllt ist.
Im Unterschied zum Gauß-Newton-Verfahren kann man kritische Punkte als Kandidaten für Extremalstellen des Funktionals $F : \mathbb{R}^p \to \mathbb{R}$

$$F(a) = ||R(a)||_2^2$$

durch die direkte Auswertung der notwendigen Extremalbedingung

$$\text{grad}_a F = \mathbf{0} \tag{5.17}$$

mit dem Newton-Verfahren bestimmen. Diese Methode ist allerdings "teurer" als das Gauß-Newton-Verfahren, da man pro Newton-Iteration jeweils die Jacobi-Matrix von $G : \mathbb{R}^p \to \mathbb{R}^p$

$$G(a) := \text{grad}_a F \ ,$$

d.h. die Hesse-Matrix von $F$ berechnen muss.

Beispiel:
Gegegen ist eine Wertetabelle

| $k$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $x_k$ | 1 | 2 | 3 | 4 |
| $y_k$ | 2 | 4 | 7 | 3 |

und es gibt die Überlegung nach einer Funktion

$$y = f(x; a, b) = \sin(ax) + b$$

mit solchen Parametern $a, b \in \mathbb{R}$ zu suchen, so dass die Länge des Residuenvektors

$$R(a,b) = \begin{pmatrix} \sin(a) + b - 2 \\ \sin(2a) + b - 4 \\ \sin(3a) + b - 7 \\ \sin(4a) + b - 3 \end{pmatrix}$$

minimal wird, also

$$\min_{(a,b)\in\mathbb{R}^2} ||R(a,b)||_2^2 \, .$$

Ich habe die Aufgabe sowohl mit dem Newtonverfahren zur Bestimmung von Nullstellen des Gradienten des Funktionals

$$F(a,b) = ||R(a,b)||_2^2 \, ,$$

als auch mit dem Gauß-Newton-Verfahren bearbeitet. Beide Verfahren waren erfolgreich (d.h. konvergent), allerdings zeigte sich, dass es mehrere Lösungen gibt. D.h. man findet evtl. mehrere lokale Minima und hat aber das Problem, dass man nicht weiß, wievele es insgesamt gibt.
Bei diesem Beispiel habe ich mit den Startwerten $(a, b) = (1, 5)$ die Extremalstelle $(a, b) = (0.558, 3.197)$ mit beiden Methoden gefunden die Hesse-Matrix von $H = F''(0.558, 3.197)$ ist positiv definit, d.h. es handelt sich um eine lokale Minimalstelle.
Mit dem Startwert $(a, b) = (2, 5)$ findet man mit dem Newton-Verfahren die kritische Stelle $(a, b) = (1.72, 3.91)$, für die die Hessematrix $H = F''(1.72, 3.91)$ einen positiven und einen negativen Eigenwert besitzt. Es handelt sich also um einen Sattelpunkt.
Mit dem Gauß-Newton-Verfahren ergibt sich für den Startwert $(a, b) = (2, 5)$ der Grenzwert $(a, b) = (2.79, 4.11)$ der Iterationsfolge, wobei mit dem Dämpfungsparameter $\alpha = 0.4$ gearbeitet werden musste, da für größere $\alpha$-Werte keine Konvergenz erzielt werden konnte. Um den kritischen Punkt $(a, b) = (2.79, 4.11)$ mit dem Newton-Verfahren zu erhalten, muss man einen näher liegenden Startwert, z.B. $(a, b) = (2.8, 4)$ verwenden. Mit der positiven Definitheit der Hesse-Matrix $H = F''(2.79, 4.11)$ zeigt man, dass es sich bei der Stelle $(a, b) = (2.79, 4.11)$ um eine lokale Minimalstelle handelt.
Im Unterschied zu linearen Ausgleichsproblemen sind bei nichtlinearen Aufgabenstellungen zusätzliche Betrachtungen zur evtl. Mehrdeutigkeit des Problems $\text{grad}_a F = \mathbf{0}$ und zur Bewertung der gefundenen kritischen Stellen (lok.

Maximum/lok. Minimum) durch die Überprüfung hinreichender Extremalbedingungen (Definitheit der Hesse-Matrix) erforderlich.
Zu dem obigen Beispiel ist allerdings auch anzumerken, dass bei den wenigen Werten $(x_k, y_k)$ der Ansatz $y = \sin(ax) + b$ auch etwas gewagt ist. Dass kann und wird sicherlich auch ein Grund für das etwas wilde Extremalverhalten des Funktionals $F(a, b) = ||R(a, b)||_2^2$ sein.
"Fittet" man seine Messwerte mit der Funktion $y = f(x)$ besser, sollte die Bestimmung geeigneter Parameter $a = (a_1, \ldots, a_p)$ auch einfacher werden.

# Kapitel 6

# Interpolation

Oft gibt es die Aufgabe, durch gegebene Punktepaare eine glatte Kurve zu legen, die analytisch leicht zu handhaben ist (Differenzieren, Integrieren), also:
Gegeben: $(x_k, y_k), k = 0, \ldots, N$ gegeben.
Gesucht: Glatte Funktion $P = P(x)$ mit

$$P(x_k) = y_k, \quad k = 0, \ldots, N$$

Mögliche Ansätze für $P$

(i) Polynome

$$P = P(x, a_0, a_1, \ldots, a_n) = a_0 + a_1 x + \cdots + a_n x^n, \quad n = N$$

(ii) Rationale Funktionen

$$P(x) = \frac{a_0 + a_1 x + \cdots + a_n x^n}{a_{n+1} + a_{n+2} x + \cdots + a_{n+m+1} x^m}, \quad n = N$$

(iii) Trigonometrische Polynome, $y_i \in \mathbb{C}$

$$\begin{aligned} P(x) &= a_0 + a_1 e^{ix} + a_2 e^{2ix} + \cdots + a_n e^{nix} \\ &= a_0 + a_1 e^{ix} + a_2 (e^{ix})^2 + \cdots + a_n (e^{ix})^n \end{aligned}$$

(iv) Splines (stückweise Polynome)

**Ziel/Aufgabe der Interpolation**

Bestimmung der Parameter $a_0, \ldots, a_n$, so dass für $P = P(x, a_0, \ldots, a_n)$ aus einer vorzugebenden Funktionenklasse die Beziehungen

$$P(x_k, a_0, \ldots, a_n) = y_k, \quad k = 0, \ldots, n \tag{6.1}$$

zu den vorgegebenen Stützstellen $(x_k, y_k)$ erfüllt sind. (6.1) heißt auch **Interpolationseigenschaft** und die Stützstellen werden auch **Knoten** genannt. (6.1) sind $n+1$ Gleichungen für die $n+1$ Parameter $a_0, \ldots, a_n$.
Sehr einfach: Lineare Splines

# 6.1 Polynominterpolation

**Definition 6.1.** *Unter $\Pi_n$ versteht man die Menge aller reellen Polynome $P : \mathbb{R} \to \mathbb{R}$ von Grad $\leq n$*

Wir wissen:

- im Fall $n = 1$ braucht man 2 Stützpunkte um eine Gerade (Polynom ersten Grades) durchzulegen
- im Fall $n = 2$ braucht man 3 Stützpunkte um eine Parabel (Polynom zweiten Grades) durchzulegen, ...

**Satz 6.2.** *Zu $n+1$ gegebenen Stützstellen $(x_k, y_k), k = 0, \ldots, n$ mit der Eigenschaft $x_i \neq x_j, i \neq j$, gibt es genau ein Polynom $p \in \Pi_n$ mit $P(x_k) = y_k, k = 0, 1, \ldots, n$*

*Beweis.* Ansatz:

$$\begin{aligned} P(x) &= a_0 + a_1 x + \cdots + a_n x^n \\ P(x_k) &= y_k, \quad k = 0, 1, \ldots, n \end{aligned} \tag{6.2}$$

bedeutet

$$\begin{aligned} a_0 + a_1 x_0 + \cdots + a_n x_0^n &= y_0 \\ &\vdots \\ a_0 + a_1 x_n + \cdots + a_n x_n^n &= y_n \end{aligned}$$

$$\Leftrightarrow \underbrace{\begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ & & \vdots & & \\ 1 & x_2 & x_2^2 & \cdots & x_n^n \end{pmatrix}}_{\text{Vandermondesche Matrix } V} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix} \tag{6.3}$$

$V$ ist für paarweise verschiedene $x_k$ regulär, d.h. $a_0, \ldots, a_n$ und damit $P$ sind eindeutig bestimmt.

□

**Definition 6.3.** *Das nach Satz 6.2 eindeutig bestimmte Polynom $P$ mit der Eigenschaft*

$$P(x_k) = y_k, \quad k = 0, 1, \ldots, n$$

*für die vorgegebenen Stützstellen $(x_k, y_k)$ heißt* ***Interpolationspolynom****.*

### 6.1.1 Konstruktion des Interpolationspolynoms

Wir erinnern uns an die Generalvoraussetzung

$$x_i \neq x_j \quad \forall i, j = 0, 1, \ldots, n, i \neq j$$

**Definition 6.4.** *Die Polynome*

$$L_k(x) = \prod_{k \neq i = 0}^{n} \frac{x - x_i}{x_k - x_i} \tag{6.4}$$

*heißen Lagrange-Basispolynome.*

**Definition 6.5.** *Die Polynome*

$$N_k(x) = \prod_{i=0}^{k-1} (x - x_i), \quad k = 1, \ldots, n$$

*mit $N_0(x) = 1$ heißen Newton-Basispolynome.*

**Satz 6.6.** *Die Monombasis*

$$1, x, \ldots, x^n$$

*sowie die Lagrange-Basispolynome*

$$L_k(x), k = 0, \ldots, n$$

*und die Newton-Basispolynome*

$$N_k(x), k = 0, \ldots, n$$

*sind Basen (linear unabhängige erzeugende Funktionensysteme) des Vektorraums der reellen Polynome $\Pi_n$ vom Grad $\leq n$*

*Beweis.* Als Übung empfohlen. □

## 6.2 Lagrange-Interpolation

Zuerst ist anzumerken, dass man das Interpolationspolynom nicht in der Form (6.2) auf der Grundlage der Lösung des Gleichungssystems (6.3) mit der Vandermondeschen Matrix bestimmt, weil das viel zu aufwändig ist.
Besser geht es mit der **Lagrange-Interpolation**.
Für $n = 3$ haben wir zum Beispiel die Basispolynome

$$L_0(x) = \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)}$$
$$L_1(x) = \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}$$
$$L_0(x) = \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)}$$
$$L_0(x) = \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}$$

und erkennen:

$$L_0(x_0) = 1, L_0(x_1) = L_0(x_2) = L_0(x_3) = 0$$

allgemein gilt:

$$L_k(x_j) = \delta_{kj}, \quad k = 0, \ldots, n \tag{6.5}$$

Damit ergibt sich für das Interpolationspolynom:

$$p(x) = \sum_{k=0}^{n} y_k L_k(x) \tag{6.6}$$

da

$$p(x_k) = 0 + 0 + \cdots + y_k L_k(x_k) + \cdots + 0 = y_k$$

gilt. (6.6) heißt **Lagrangsches Interpolationspolynom**.

## 6.3 Newton-Interpolation

Bei der Lagrange-Interpolation haben wir das Interpolationspolynom in der Lagrange-Basis entwickelt. Bei der Newton-Interpolation wird das eindeutig existierende Interpolationspolynom in der Newton-Basis entwickelt.
Ansatz:

$$p(x) = \sum_{k=0}^{n} c_k N_k(x)$$

Durch sukzessives Vorgehen erhalten wir durch Berücksichtigung der Stützstellen $(x_k, y_k), k = 0, \ldots, n$ die Koeffizienten der $N_k(x)$

$$\begin{aligned}
p_n(x_0) &= c_0 && = y_0 \rightsquigarrow c_0 \\
p_n(x_1) &= c_0 + c_1(x_1 - x_0) && = y_1 \rightsquigarrow c_1 \\
p_n(x_2) &= c_0 + c_1(x_1 - x_0) + c_2(x_2 - x_0)(x_2 - x_1) && = y_2 \rightsquigarrow c_2 \\
&\vdots \\
p_n(x_n) &= \sum_{k=0}^{n} c_k N_k(x_n) && = y_n \rightsquigarrow c_n
\end{aligned}$$

**Definition 6.7.**

$$p_n(x) := \sum_{k=0}^{n} c_k N_k(x) \in \Pi_n$$

*heißt Newtonsches Interpolationspolynom.*

**Bemerkung.** $c_n$ ist der Koeffizient von $x^n$ im Interpolationspolynom und $c_k$ ist eindeutig festgelegt durch $x_0, \ldots, x_k, y_0, \ldots y_k$ d.h. durch die ersten $k$ Stützstellen.

**Definition 6.8.** *Wir schreiben $c_k := f[x_0 x_1 \ldots x_k]$ für die Abbildung*

$$\{(x_0, y_0), \ldots, (x_k, y_k)\} \mapsto c_k$$

Betrachtet man Teilmengen der Stützstellen

$$x_{i_0}, \ldots, x_{i_k},$$

dann bezeichnet man das Interpolationspolynom an diesen Stützstellen mit

$$p^*_{i_0 i_1 \ldots i_k}(x)$$

wobei $i_0, \ldots, i_k$ paarweise verschiedene Zahlen aus $\{0, \ldots, k\}$ sind. Nach der Definition eines Interpolationypolynoms muss

$$p^*_{i_0 i_1 \ldots i_k}(x_{i_j}) \equiv y_{i_j}, \quad j = 0, 1, \ldots, k$$

Damit gilt

$$p^*_k(x) \equiv y_k \tag{6.7}$$

für das Polynom 0. Ordnung $p^*_k$ (also $p^*_k(x) \neq p_k(x)$)

**Bemerkung.** $p^*_k$ ist Konstante und $p_k(x)$ Polynom $k$-ter Ordnung, deshalb der Stern

**Lemma 6.9.** *Es gilt für alle* $k \in \{1, 2, \ldots, n\}$

$$p^*_{i_0,\ldots,i_k}(x) = \frac{(x - x_{i_0})p^*_{i_1\ldots i_k}(x) - (x - x_{i_k})p^*_{i_0\ldots i_{k-1}}(x)}{x_{i_k} - x_{i_0}} \tag{6.8}$$

*Beweis.* Induktion Die beiden rechts stehenden Polynome in (6.8) haben einen Grad $\leq k-1$ (damit der gesamte Ausdruck einen Grad $\leq n$).
Anfang ($k = 1$) ist trivial wegen (6.7).
Es ist zu zeigen, dass das rechts in (6.8) stehende Polynom das Interpolationypolynom zu den Stützstellen $x_{i_0}, \ldots, x_{i_k}$ ist (Ausdruck rechts von (6.8) bezeichnen wir mit $q(x)$) $\deg q(x) \leq k$ ist offensichtlich. Weiter ist

$$q(x_{i_0}) = \frac{0 - (x_{i_0} - x_{i_k})p^*_{i_0\ldots i_{k-1}}(x_{i_0})}{x_{i_k} - x_{i_0}} = y_{i_0}$$

und analog

$$q(x_{i_k}) = y_{i_k}$$

Schließlich für die restlichen Stützstellen $1 \leq j \leq k-1$

$$\begin{aligned} q(x_{i_j}) &= \frac{(x_{i_j} - x_{i_0})p^*_{i_1\ldots i_k}(x_{i_j}) - (x_{i_j} - x_{i_k})p^*_{i_0\ldots i_{k-1}}(x_{i_j})}{x_{i_k} - x_{i_0}} \\ &= \frac{(x_{i_j} - x_{i_0})y_{i_j} - (x_{i_j} - x_{i_k})y_{i_j}}{x_{i_k} - x_{i_0}} = y_{i_j} \end{aligned}$$

Damit erfüllt $q$ die Interpolationsbedingung $q(x_{i_j}) = y_{i_j}, j = 0, \ldots, k$ also genau das, was $p^*_{i_0\ldots i_k}(x)$ leistet. Aufgrund der Eindeutigkeit des Interpolationspolynoms gilt also

$$q = p_{i_0\ldots i_k}$$

□

**Satz 6.10.** *Es gilt*

$$f[x_{i_0} \ldots x_{i_k}] = \frac{f[x_{i_1} \ldots x_{i_k}] - f[x_{i_0} \ldots x_{i_{k-1}}]}{x_{i_k} - x_{i_0}}$$

*Beweis.* Nach der Definition von $f[\ldots]$ ist dies gerade der Koeffizient von der höchsten Potenz des Interpolationspolynoms. Betrachten (6.8). Das Polynom links hat in der höchsten Potenz den Term $f[x_{i_0} \ldots x_{i_k}]x^k$, das rechts stehende hat in der höchsten Potenz

$$\frac{x \cdot f[x_{i_1} \ldots x_{i_k}]x^{k-1} - x \cdot f[x_{i_0} \ldots x_{i_{k-1}}]x^{k-1}}{x_{i_k} - x_{i_0}}$$

$\rightsquigarrow$ Behauptung. □

Als Folgerung des Satzes 6.10 findet man das folgende Schema

| | $k=0$ | $k=1$ | $k=2$ |
|---|---|---|---|
| $x_0$ | $y_0 = f[x_0]$ | | |
| $x_1$ | $y_1 = f[x_1]$ | $f[x_0x_1] = \frac{f[x_1]-f[x_0]}{x_1-x_0}$ | |
| $x_2$ | $y_2 = f[x_2]$ | $f[x_1x_2] = \frac{f[x_2]-f[x_1]}{x_2-x_1}$ | $f[x_0x_1x_2] = \frac{f[x_1x_2]-f[x_1x_0]}{x_2-x_0}$ |
| $\vdots$ | | | |

Es wird "Schema der dividierten Differenzen" genannt. Daraus liest man das Newtonsche Interpolationspolynom ab:

$$p_2(x) = f[x_0] + f[x_0x_1](x-x_0) + f[x_0x_1x_2](x-x_0)(x-x_1)$$

12. Vorlesung am 23.11.2011

# 6.4 Algorithmische Aspekte der Polynominterpolation

## 6.4.1 Horner-Schema

Für die Berechnung eines Polynoms in der Form

$$p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$$

Werden $1+2+\cdots+n = \frac{n(n+1)}{2}$ Multiplikationen und $n$ Additionen benötigt. Also $\mathcal{O}(n^2)$ flops

$$p(x) = a_0 + x(a_1 + x(a_2 + \cdots)) = (\cdots(a_nx + a_{n-1})x + a_{n-2})x + \cdots + a_1)x + a_0$$

$\rightsquigarrow n$ Multiplikationen und Additionen, also $2n \in \mathcal{O}(n)$ flops.

Für die Newton Basis ergibt sich

$$p(x) = \sum_{k=0}^{n} c_kN_k(x), \quad c_k \text{ gegeben}$$

$N_k$ rekursiv aufgebaut:

$$N_k(x) = (x-x_0)\cdots(x-x_k)$$
$$\rightsquigarrow N_k(x) = (x-x_{k-1})N_{k-1}(x)$$

$p$ kann in der Form

$$p(x) = c_0 + (x-x_0)(c_1 + (x-x_1)(c_2 + \cdots + c_n(x-x_{n-1}))\cdots)$$

geschrieben werden. Daraus resultiert der Algorithmus:

**Algorithmus 2** Wertet Newton Polynom mittels Horner-Schema aus

$u_{n+1} = 0$
**for** $k = n$ **downto** $0$ **do**
  $u_k = (x - x_k)u_{k+1} + c_k$
**end for**
$p(x) = u_0$

Mit Laufzeit $3n$ flops.

### 6.4.2 Lagrange-Interpolation

Im Unterschied zur Newton-Interpolation ist der Aufwand bei der Lagrange-Interpolation bei Hinzunahme einer Stützstelle recht groß, denn sämtliche Basispolynome ändern sich (Grad wird um 1 erhöht)
Was kann man tun, um hier den Mehraufwand zur Berechnung von $p(x)$ an einer Stelle $x \neq x_j$ klein zu halten?
Man findet:

$$p(x) = \sum_{k=0}^{n} y_k L_k(x) = \sum_{k=0}^{n} y_k \prod_{j=0}^{n} \frac{x - x_k}{x_j - x_k} \tag{6.9}$$

$$= \sum_{k=0}^{n} y_k \frac{1}{x - x_k} \left[ \prod_{k \neq i = 0}^{n} \frac{1}{x_i - x_k} \right] \prod_{j=0}^{n} (x - x_j) \tag{6.10}$$

Die Koeffizienten in den eckigen Klammern

$$\lambda_k = \prod_{k \neq i = 0}^{n} \frac{1}{x_i - x_k} = \frac{1}{\prod (x_i - x_k)}, \quad k = 0, 1, \ldots, n$$

nennt man Stützkoeffizienten. Damit führt man mit

$$\mu_k = \frac{\lambda_k}{x - x_k}, \quad k = 0, 1, \ldots, n$$

Größen ein, die von der Stelle $x$, an der interpoliert werden soll, abhängen. Es ergibt sich

$$p(x) = \left[ \sum_{k=0}^{n} \mu_k y_k \right] \prod_{j=0}^{n} (x - x_j) \tag{6.11}$$

Betrachtet man dies für die speziellen Werte $y_k = 1, k = 0, 1, \ldots, n$, dann ist $p(x) \equiv 1$ das eindeutig bestimmte Interpolationspolynom für die $n + 1$

Stützpunkte $(x_k, 1)$, sodass

$$1 = p(x) = \left[\sum_{k=0}^{n} \mu_k\right] \prod_{j=0}^{n} (x - x_k)$$

$$\Rightarrow \quad \prod_{j=0}^{n} (x - x_k) = \frac{1}{\sum_{k=0}^{n} \mu_k} \tag{6.12}$$

Aus (6.11) und (6.12) folgt mit

$$p(x) = \frac{\sum_{k=0}^{n} \mu_k y_k}{\sum_{k=0}^{n} \mu_k} \tag{6.13}$$

die sogenannte baryzentrische Formel der Lagrange-Interpolation

**Satz 6.11.** *Für die $n+1$ Stützkoeffizienten $\lambda_k^{(n)}$ zu den paarweise verschiedenen Stützstellen $x_0, x_1, \ldots, x_n$ gilt:*

$$\sum_{k=0}^{n} \lambda_k^{(n)} = 0 \tag{6.14}$$

Die Formel (6.13) hat den Vorteil, dass man bei der Hinzunahme einer $(n+2)$-ten Stützstelle $x_{n+1}$ zu $x_0, x_1, \ldots, x_n$ die neuen $\lambda$-Werte $\lambda_k^{(n+1)}$ aus den alten $\lambda_k^{(n)}$ durch die Beziehungen

$$\lambda_k^{(n+1)} = \frac{\lambda_k^{(n)}}{x_k - x_{n+1}}, \quad k = 0, 1, \ldots, n$$

ermitteln kann. Den fehlenden Wert $\lambda_{n+1}^{(n+1)}$ bestimmt man unter Nutzung von (6.14) durch

$$\lambda_{n+1}^{(n+1)} = -\sum_{k=0}^{n} \lambda_k^{(n+1)}$$

Insgesamt braucht man zur Bestimmung der $\mu_k$ $2n$ Multiplikationen und $n$ Additionen und damit zur Polynomwertberechnung mit der baryzentrischen Formel $3n$ Multiplikationen und $3n$ Additionen, wobei der zusätzliche Aufwand bei Hinzunahme einer $(n+2)$-ten Stützstelle mit $n$ Multiplikationen und $n$ Additionen moderat ist.

## 6.5 Verfahren von Neville und Aitken

Es ist vergleichbar mit der Herangehensweise bei der Newton-Interpolation
Aus Lemma 6.9 folgt mit

$$\begin{aligned} y_0 &=: p_0^*(x) \\ &\vdots \\ y_n &=: p_n^*(x) \end{aligned}$$

die Rekursion

$$\begin{aligned} p_{0,1}^*(x) &= \frac{(x-x_0)p_1^*(x) - (x-x_1)p_0^*(x)}{x_1 - x_0} \\ &\vdots \\ p_{n-1,n}^*(x) &= \frac{(x-x_{n-1})p_n^*(x) - (x-x_n)p_{n-1}^*(x)}{x_n - x_{n-1}} \\ &\text{usw.} \\ p_{0,1,2}^*(x) &= \frac{(x-x_0)p_{1,2}^*(x) - (x-x_2)p_{0,1}^*(x)}{x_2 - x_0} \end{aligned}$$

Für den Algorithmus von Neville und Aitken folgt das Schema zur Berechnung von $p$ an der Stelle $x$

$$\begin{array}{c|llll} x_0 & y_0 = p_0^*(x) & & & \\ x_1 & y_1 = p_1^*(x) & p_{0,1}^*(x) & & \\ x_2 & y_2 = p_2^*(x) & p_{1,2}^*(x) & p_{0,1,2}^*(x) & \\ \vdots & \vdots & & \ddots & \\ x_n & y_n = p_n^*(x) & p_{n-1,n}^*(x) & \dots & p_{0,1,\dots,n}^*(x) \end{array}$$

**Beispiel.**

| $x_k$ | 0 | 1 | 3 |
|---|---|---|---|
| $y_k$ | 1 | 3 | 2 |

Polynomwert soll an der Stelle $x = 2$ berechnet werden.

$$\begin{array}{c|lll} 0 & 1 & & \\ 1 & 3 & p_{0,1}(2) = \frac{(2-0)3-(2-1)1}{1-0} = 5 & \\ 3 & 2 & p_{1,2}(2) = \frac{(2-1)2-(2-3)3}{3-1} = \frac{5}{2} & p_{0,1,2}(2) = \frac{(2-0)\frac{5}{2}-(2-3)5}{3-0} = \frac{10}{3} \end{array}$$

## 6.6 Fehlerabschätzung der Polynominterpolation

Handelt es sich bei den Stützpunkten $(x_k, y_k)$ nicht um diskrete Messwerte, sondern um die Wertetabelle einer gegebenen Funktion $f(x)$, dann ist der Fehler $f(x) - p_n(x)$, den man bei der Interpolation macht, von Interesse.
Nimmt man zu den Stützwerten $x_0, \ldots, x_n$ den Wert $x = x_{n+1}$ hinzu, ergibt die Interpolationsbedingung $y = f(x) = p_{n+1}(x)$

$$\underbrace{p_{n+1}(x) = f(x)}_{p_{n+1}(x_{n+1})=y_{n+1}} = p_n(x) + f[x_0, x_1, \ldots, x_n, x] \prod_{k=0}^{n} (x - x_k)$$

bzw.

$$f(x) - p_n(x) = f[x_0, x_1, \ldots, x_n, x](x - x_0)(x - x_1) \cdots (x - x_n) \tag{6.15}$$

Der folgende Satz liefert die Grundlage für die Abschätzung des Interpolationsfehlers (6.15).

**Satz 6.12.** *Sei* $]a, b[ = ]\min_{0\le j\le n} x_j, \max_{0\le j\le n} x_j[$ *und sei* $p_n(x)$ *das Interpolationspolynom zur Wertetabelle* $(x_k, f(x_k))$ *der* $(n+1)$*-mal stetig differenzierbaren Funktion* $f$ *auf* $[a, b]$*, wobei die Stütstellen* $x_k$ *paarweise verschieden sind.*
*Dann gibt es für jedes* $\tilde{x} \in ]a, b[$ *einen Zwischenwert* $\xi = \xi(x_0, \ldots, x_n) \in ]a, b[$ *mit*

$$f(\tilde{x}) - p_n(\tilde{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(\tilde{x} - x_0) \cdots (\tilde{x} - x_n)$$

*Beweis.* Siehe Vorlesung oder Bärwolff □

Aus dem Satz 6.12 folgt direkt für eine $(n+1)$-mal stetig differenzierbare Funktion die Fehlerabschätzung

$$|f(x) - p_n(x)| \le \frac{\max_{\xi\in[a,b]} \left|f^{(n+1)}(\xi)\right|}{(n+1)!} \underbrace{\left|\prod_{k=0}^{n} (x - x_k)\right|}_{=:w(x)} \tag{6.16}$$

Hat man bei den Stützstellen die freie Wahl und soll auf dem Intervall $[a, b]$ interpoliert werden, dann ist die Wahl der Nullstellen des Tschebyscheff-Polynoms $T_{n+1}(x)$ auf $[a, b]$ transformiert, d.h

$$x_k^* = \frac{a+b}{2} + \frac{b-a}{2} \cos\left(\frac{2(n+1-k)-1}{2(n+1)}\pi\right), \quad k = 0, \ldots, n \tag{6.17}$$

von Vorteil, denn für $w^*(x) = \prod_{k=0}^{n}(x - x_k^*)$ gilt:

**Satz 6.13.** *Seien $x_k$ äquidistante und $x_k^*$ gemäß* (6.17) *verteilte Stützstellen des Intervalls $[a, b]$. Dann gilt:*

$$\max_{x\in[a,b]} |w^*(x)| \leq \max_{x\in[a,b]} |w(x)|$$

*und falls $f$ beliebig oft differenzierbar ist, gilt*

$$\lim_{k\to\infty} p_k^*(x) = f(x) \quad auf \quad [a, b]$$

## 6.7 Hermite-Interpolation

Hat man einen Stützpunkt $(x_0, y_0)$ vorgegeben, so ist damit ein Polynom 0-ten Grades festgelegt (Gerade parallel zur $x$-Achse). Hat man an der Stelle noch eine Ableitungsinformation, d.h. $(x_0, y_0')$, dann ist damit eine Gerade durch den Punkt $(x_0, y_0)$ mit dem Anstieg $y_0'$ festgelegt, also ein Polynom 1-ten Grades.

**Satz 6.14.** *Sei $f$ eine $(n+1)$-mal stetig diff'bare Funktion in einem Intervall um den Punkt $x$. Dann gilt*

$$\lim_{x_0\to x\ldots x_n\to x} f[x_0, x_1, \ldots, x_n, x] = \frac{f^{(n+1)}(x)}{(n+1)!}$$

*Beweis.* vollständige Induktion, MWS □

Der Satz 6.14 rechtfertigt

**Definition 6.15.**

$$f[\underbrace{x, x, \ldots, x}_{n+2}] = \frac{f^{(n+1)}(x)}{(n+1)!} \tag{6.18}$$

Auf der Basis dieser Definition enstehen gemischte Differenzen wieder rekursiv, z.B.

$$f[x_0, x_1, x_2] = \frac{f[x_0, x_1] - f[x_1, x_2]}{x_0 - x_1}$$
$$f[x_0, x_0, x_0, x_1] = \frac{f[x_0, x_0, x_0] - f[x_0, x_0, x_1]}{x_0 - x_1}$$

Das Interpolationspolynom ist dann gegeben durch

$$p(x) = \sum_{k=0}^{n} f[x_0 \ldots x_k] \prod_{j=0}^{k-1} (x - x_j)$$

Man überlegt sich, dass zur Bestimmung der Polynomkoeffizienten eines Hermiteschen Interpolationspolynoms (zur Erfüllung von Interpolationsbedingungen bei Berücksichtigung von Ableitungsinformationen) das folgende Schema für die Bedingungen

**Beispiel.**

$$\begin{aligned}(x_0, y_0) &= (0, 1)\\ (x_0, y_0') &= (0, 2)\\ (x_0, y_0'') &= (0, 4)\\ (x_1, y_1) &= (1, 2)\\ (x_1, y_1') &= (1, 3)\end{aligned}$$

die Form

| | $c_0$ | $c_1$ | $c_2$ | $c_3$ | $c_4$ |
|---|---|---|---|---|---|
| 0 | 1 | | | | |
| 0 | 1 | $y_0' = 2$ | | | |
| 0 | 1 | $y_0' = 2$ | $\frac{y_0''}{2} = 2$ | | |
| 1 | 2 | $\frac{1-2}{0-1} = 1$ | $\frac{2-1}{0-1}$ | $\frac{2-(-1)}{0-1} = -3$ | |
| 1 | 2 | $y_1' = 3$ | $\frac{1-3}{0-1} = 2$ | $\frac{-1-2}{0-1} = 3$ | $\frac{-3-3}{0-1} = 6$ |

hat.
Daraus ergibt sich das Hermite-Interpolationspolynom:

$$\begin{aligned}p(x) = f[x_0] + f[x_0, x_0](x - x_0) + f[x_0, x_0, x_0](x - x_0)^2\\ +f[x_0, x_0, x_0, x_1](x - x_0)^3 + f[x_0, x_0, x_0, x_1, x_1](x - x_0)^3(x - x_1)\end{aligned}$$

also für obige Werte

$$p(x) = 1 + 2x + 2x^2 - 3x^3 + 6x^3(x - 1)$$

## 6.8 Spline-Interpolation

13. Vorlesung am 28.11.2011

Problem bei der Polynom-Interpolation:
Eventuell große Oszillationen durch Polynome höheren Grades bei Stützpunktzahlen $\geq 10$
Deshalb:
Statt eines Interpolationspolynoms konstruiert man für $(x_k, y_k), k = 0, 1, \ldots, n$ in jeden Teilintervall einzelne Polynome, die an den Randstellen glatt ineinander übergehen. Betrachten mit

$$\Delta = \{a = x_0 < x_1 < \ldots < x_N = b\}$$

eine fest gewählte Zerlegung von $[a, b]$, wobei die Stützstellen $x_0, \ldots, x_N$ auch als Knoten bezeichnet werden.

**Definition 6.16.** *Eine **Splinefunktion** der Ordnung $l \in \mathbb{N}$ zur Zerlegung $\Delta$ ist eine Funktion $s \in C^{l-1}[a, b]$, die auf jedem Intervall $[x_{k-1}, x_k]$ mit einem Polynom $l$-ten Grades übereinstimmt. Der Raum der Splinefunktionen wird mit $S_{\Delta,l}$ bezeichnet, es gilt also:*

$$S_{\Delta,l} = \{s \in C^{l-1}[a, b] : s|_{[x_{k-1},x_k]} = p_k|_{[x_{k-1},x_k]} \text{ für ein } p_k \in \Pi_l\}$$

*Anstelle Splinefunktionen verwendet man auch einfach **Spline**.*

Splines erster Ordnung nennt man auch lineare, die zweiter Ordnung auch quadratische Splines. Besonders hervorzuheben sind kubische Splines, die in der Praxis besonders häufig verwendet werden.

Da wir vorgegebene Wertetabellen interpolieren wollen, geht es im Folgenden um die Berechnung interpolierender Splinefunktionen, also Splines mit der Eigenschaft

$$s(x_k) = f_k \quad \text{für} \quad k = 0, 1, \ldots, N \tag{6.19}$$

für $(x_k, f_k), k = 0, 1, \ldots, N$

### 6.8.1 Interpolierende lineare Splines $s \in S_{\Delta,1}$

Offensichtlich gilt:

$$s(x) = a_k + b_k(x - x_k), \quad x \in [x_k, x_{k+1}]$$

aus $s_k(x_k) = f_k$ sowie $s_k(x_{k+1}) = f_{k+1}$ folgt

$$a_k = f_k, \quad b_k = \frac{f_{k+1} - f_k}{x_{k+1} - x_k}, \quad k = 0, \ldots, N - 1$$

**Satz 6.17.**

*(a) Zur Zerlegung $\Delta = a = x_0 < \ldots < x_N = b$ und $f_0, \ldots, f_N$ gibt es genau einen Spline $s \in S_{\Delta,1}$ mit der Eigenschaft* (6.19)

*(b) Zu einer Funktion $f \in C^2[a,b]$ sei $s \in S_{\Delta,1}$ der zugehörige interpolierende lineare Spline. Dann gilt*

$$\|s - f\|_\infty \leq \frac{1}{8} \|f''\|_\infty h_{\max}^2$$

*mit $h_{\max} := \max_{k=0,\ldots,N-1}(x_{k+1} - x_k)$*

*Beweis.* (a) nach Konstruktion

(b) Für jedes $k \in 1, \ldots, N$ stimmt $s$ auf $[x_{k-1}, x_k]$ mit demjenigen $p \in \Pi_1$ überein, für das $p(x_{k-1}) = f(x_{k-1})$ und $p(x_k) = f(x_k)$ gilt. Der Fehler bei der Polynominterpolation (Satz 6.12) liefert

$$\begin{aligned} |s(x) - f(x)| &\leq \frac{(x - x_{k-1})(x_k - x)}{2} \max_{\xi \in [x_{k-1}, x_k]} |f''(\xi)| \\ &\leq \frac{1}{8} h_{\max}^2 \|f''\|_\infty, \quad x \in [x_{k-1}, x_k] \end{aligned}$$ □

## 6.8.2 Kubische Splines

Betrachte nun $S_{\Delta,3}$, und verwenden

$$\|u\|_2 := \left( \int_a^b |u(x)|^2 \, \mathrm{d}x \right)^{\frac{1}{2}}$$

**Lemma 6.18.** *Wenn eine Funktion $f \in C^2[a,b]$ und eine kubische Splinefunktion $s \in S_{\Delta,3}$ in den Knoten übereinstimmen, d.h.*

$$s(x_k) = f(x_k) \quad \textit{für} \quad k = 0, \ldots, N$$

*so gilt*

$$\|f'' - s''\|_2^2 = \|f''\|_2^2 - \|s''\|_2^2 - 2([f' - s']s'')(x)\Big|_{x=a}^{x=b} \tag{6.20}$$

*Beweis.*

$$\begin{aligned} \|f'' - s''\|_2^2 &= \int_a^b |f''(x) - s''(x)|^2 \, \mathrm{d}x = \|f''\|_2^2 - 2\int_a^b (f''s'')(x)\mathrm{d}x + \|s''\|_2^2 \\ &= \|f''\|_2^2 - 2\int_a^b ([f'' - s'']s'')(x)\mathrm{d}x - \|s''\|_2^2 \end{aligned}$$

Für den mittleren Term ergibt die partielle Integration

$$\begin{aligned}
&\int_{x_{k-1}}^{x_k} ([f'' - s'']s'')(x)\mathrm{d}x \\
&= ([f' - s']s'')(x)\Big|_{x_{k-1}}^{x_k} - \int_{x_{k-1}}^{x_k} ([f' - s']s''')(x)\mathrm{d}x \\
&= ([f' - s']s'')(x)\Big|_{x_{k-1}}^{x_k} - \underbrace{([f - s]s''')(x)\Big|_{x_{k-1}}^{x_k}}_{=0} + \underbrace{\int_{x_{k-1}}^{x_k} ([f - s]s^{(4)})(x)\mathrm{d}x}_{=0}
\end{aligned}$$

Die Summation über $k = 1, \ldots, N$ ergibt

$$\begin{aligned}
\int_a^b ([f'' - s'']s'')(x)\mathrm{d}x &= \sum_{k=1}^{N} \{([f' - s']s'')(x_k) - ([f' - s']s'')(x_{k-1})\} \\
&= ([f' - s']s'')(b) - ([f' - s']s'')(a)
\end{aligned}$$

□

**Satz 6.19.** *Gegeben sei $f \in C^2[a, b]$ und ein kubischer Spline $s \in S_{\Delta,3}$ mit $s(x_k) = f(x_k), k = 0, \ldots, N$. Dann gilt die Identität*

$$\|f''\|_2^2 - \|s''\|_2^2 = \|f'' - s''\|_2^2 \tag{6.21}$$

*sofern eine der 3 folgenden Bedingungen erfüllt ist*

*(a)* $s''(a) = s''(b) = 0$ *(natürliche RB)*

*(b)* $s'(a) = f'(a),\ s'(b) = f'(b)$ *(vollst. RB)*

*(c)* $f'(a) = f'(b),\ s'(a) = s'(b),\ s''(a) = s''(b)$ *(period. RB)*

*Beweis.* Die Aussage des Satzes ergibt sich durch Berücksichtigung von $(a), (b)$ bzw $(c)$ in der Identität (6.20) □

**Korollar 6.1.** *Zu gegebenen Werten $f_0, \ldots, f_N \in \mathbb{R}$ hat ein interpolierender kubischer Spline $s \in S_{\Delta,3}$ mit $s''(a) = s''(b) = 0$ unter allen hinreichend glatten interpolierenden Funktionen die geringste Krümmung, es gilt also*

$$\|s''\|_2 \leq \|f''\|_2$$

*für jede Funktion $f \in C^2[a, b]$ mit $f(x_k) = f_k$ für $k = 0, \ldots, N$*

*Beweis.* Folgt direkt aus (6.21) □

### 6.8.3 Berechnung interpolierender kubischer Splines

Lokaler Ansatz

$$s(x) = a_k + b_k(x - x_k) + c_k(x - x_k)^2 + d_k(x - x_k)^3, \quad x \in [x_k, x_{k+1}], k = 0, \ldots, N-1 \tag{6.22}$$

für $s : [a, b] \to \mathbb{R}$, wobei $s(x) =: p_k(x)$ auf dem Intervall $[x_k, x_{k+1}]$ verabredet wird.
Aufgabe: Bestimmung von $a_k, \ldots, d_k, k = 0, \ldots, N-1$ so, dass $s$ auf $[a, b]$ zweimal stetig differenzierbar ist und darüberhinaus in den Knoten vorgegebene Werte $f_0, \ldots, f_N \in \mathbb{R}$ interpoliert

$$s(x_k) = f_k, \quad k = 0, \ldots, N$$

Setzen $h_k := x_{k+1} - x_k,\ k = 0, \ldots, N$

**Lemma 6.20.** *Falls $N+1$ reelle Zahlen $s_0'', \ldots, s_N'' \in \mathbb{R}$ den folgenden $N-1$ gekoppelten Gleichungen ($k = 1, \ldots, N-1$)*

$$h_{k-1}\underbrace{s_{k-1}''}_{M_{k-1}} + 2(h_{k-1} + h_k)\underbrace{s_k''}_{M_k} + h_k\underbrace{s_{k+1}''}_{M_{k+1}} = \underbrace{6\frac{f_{k+1} - f_k}{h_k} - 6\frac{f_k - f_{k-1}}{h_{k-1}}}_{g_k} \tag{6.23}$$

*genügen, so liefert der lokale Ansatz* (6.22) *mit den Setzungen*

$$c_k = \frac{M_k}{2}, \quad a_k = f_k, \quad d_k = \frac{M_{k+1} - M_k}{6h_k}, \quad b_k = \frac{f_{k+1} - f_k}{h_k} - \frac{h_k}{6}(M_{k+1} + 2M_k)$$

*für $k = 0, \ldots, N-1$ eine kubische Splinefunktion $s \in S_{\Delta,3}$, die die Interpolationsbedingung $s(x_k) = f_k$ erfüllt.*

*Beweis.* Vorlesung oder Bärwolff bzw. Plato □

**Bemerkung.** Die **Momente** $M_0, \ldots, M_N$ stimmen mit den 2. Ableitungen der Splinefunktion $s$ in den Knoten $x_k$ überein

$$s_k'' = M_k = s''(x_k), \quad k = 0, \ldots, N$$

(6.23) bedeutet: Es liegen $N-1$ Bedingungen für $N+1$ Momente vor, d.h. es gibt 2 Freiheitsgrade. Diese werden durch die folgenden Randbedingungen festgelegt:

- Natürliche RB $s_0'' = s_N'' = 0$
- Vollständige RB $s_0' = f_0',\ s_N' = f_N'$ für geg. $f_0', f_N' \in \mathbb{R}$
- Periodische RB $s_0' = s_N',\ s_0'' = s_N''$

(diese Festlegungen korrelieren mit den Bedinungen $(a), (b), (c)$ des Satzes 6.19)

### 6.8.4 Gestalt der Gleichungssysteme

**Natürliche Randbedingunden**

$$\begin{bmatrix} 2(h_0+h_1) & h_1 & & 0 \\ h_1 & 2(h_1+h_2) & \ddots & \\ & \ddots & & h_{N-2} \\ 0 & & h_{N-2} & 2(h_{N-2}+h_{N-1}) \end{bmatrix} \begin{bmatrix} M_1 \\ \\ \\ M_{N-1} \end{bmatrix} = \begin{bmatrix} g_1 \\ \vdots \\ \\ g_{N-1} \end{bmatrix} \tag{6.24}$$

**Vollständige Randbedingungen**

$$\begin{bmatrix} 2h_0 & h_0 & & & 0 \\ h_0 & 2(h_0+h_1) & \ddots & & \\ & \ddots & 2(h_{N-2}+h_{N-1}) & h_{N-1} & \\ 0 & & h_{N-1} & 2h_{N-1} \end{bmatrix} \begin{bmatrix} M_0 \\ \\ \\ M_N \end{bmatrix} = \begin{bmatrix} g_0 \\ \vdots \\ \\ g_N \end{bmatrix} \tag{6.25}$$

Dieses Gleichungssystem erhält man durch die Beziehungen

$$s'(x_0) = p'_0(x_0) = b_0 = \frac{f_1 - f_0}{h_0} - \frac{h_0}{6}(M_1 + 2M_0) \tag{6.26}$$

$$s'(x_N) = p'_{N-1}(x_N) = \frac{f_N - f_{N-1}}{h_{N-1}} + \frac{h_{N-1}}{6}(M_{N-1} + 2M_N)\,, \tag{6.27}$$

woraus sich mit $s'(x_0) = f'_0$ bzw. $s'(x_N) = f'_N$ die beiden Gleichungen

$$\begin{aligned} 2h_0M_0 + h_0M_1 &= -6s'(x_0) + 6\frac{f_1 - f_0}{h_0} = -6f'_0 + 6\frac{f_1 - f_0}{h_0} =: g_0 \\ h_{N-1}M_{N-1} + 2h_{N-1}M_N &= 6s'(x_N) - 6\frac{f_N - f_{N-1}}{h_{N-1}} = 6f'_N - 6\frac{f_N - f_{N-1}}{h_{N-1}} =: g_N \end{aligned}$$

und damit (6.25) ergeben.

**periodische Randbedingungen**

$$\begin{bmatrix} 2(h_{N-1}+h_0) & h_0 & & h_{N-1} \\ h_0 & 2(h_0+h_1) & \ddots & \\ & \ddots & & h_{N-2} \\ h_{N-1} & & h_{N-2} & 2(h_{N-2}+h_{N-1}) \end{bmatrix} \begin{bmatrix} M_0 \\ \\ \\ M_{N-1} \end{bmatrix} = \begin{bmatrix} g_0 \\ \vdots \\ \\ g_{N-1} \end{bmatrix} \tag{6.28}$$

Die erste Gleichung des Systems ergibt sich unter Nutzung der periodischen Bedingungen $M_0 = M_N$ bzw. $s_0' = s_N'$ und der Beziehungen (6.26), (6.27) zu

$$\frac{f_1 - f_0}{h_0} - \frac{h_0}{6}(M_1 + 2M_0) = \frac{f_N - f_{N-1}}{h_{N-1}} + \frac{h_{N-1}}{6}(M_{N-1} + 2M_0)$$

bzw.

$$2(h_{N-1} + h_0)M_0 + h_0 M_1 + h_{n-1} M_{N-1} = 6\frac{f_1 - f_0}{h_0} - 6\frac{f_N - f_{N-1}}{h_{N-1}} =: g_0 \, .$$

Die letzte Gleichung des Systems (6.28) ergibt sich mit $M_0 = M_N$ unmittelbar.

## 6.9 Existenz und Eindeutigkeit der betrachteten interpolierenden kubischen Splines

10. Vorlesung 20.05.2009

Alle Koeffizientenmatrizen der Gleichungssysteme zur Berechnung der Momente $M_k = s_k''$ haben die Eigenschaft, strikt diagonal dominant zu sein, d.h. es gilt für die Matrix $A = (a_{ij}) \in \mathbb{R}^{N \times N}$

$$\sum_{k \neq j = 1}^{N} |a_{kj}| < |a_{kk}| \, , \quad k = 1, \ldots, N \, . \tag{6.29}$$

**Lemma 6.21.** *Jede strikt diagonal dominante Matrix $A = (a_{kj}) \in \mathbb{R}^{N \times N}$ ist reguär und es gilt*

$$\|x\|_\infty \leq \max_{k=1,\ldots,N} \left\{ (|a_{kk}| - \sum_{k \neq j = 1}^{n} |a_{kj}|)^{-1} \right\} \|Ax\|_\infty \, , \quad x \in \mathbb{R}^n \tag{6.30}$$

*Beweis.* Für $x \in \mathbb{R}^N$ sei der Index $x \in \{1, \ldots, N\}$ so gewählt, dass $|x_k| = \|x\|_\infty$ gilt. Dann findet man

$$\begin{aligned} \|Ax\|_\infty \geq |(Ax)_k| = \left| \sum_{j=1}^{N} a_{kj} x_j \right| &\geq |a_{kk}| \, |x_k| - \sum_{k \neq j = 1}^{N} |a_{kj}| \, |x_j| \\ &\geq |a_{kk}| \, |x_k| - \sum_{k \neq j = 1}^{N} |a_{kj}| \, \|x\|_\infty = \left( |a_{kk}| - \sum_{k \neq j = 1}^{N} |a_{kj}| \right) \|x\|_\infty \\ \Leftrightarrow \quad \|x\|_\infty \leq \left( |a_{kk}| - \sum_{k \neq j = 1}^{N} |a_{kj}| \right)^{-1} &\|Ax\|_\infty \end{aligned}$$

Dies liefert die Gültigkeit von (6.30) woraus die Regularität von $A$ direkt folgt. (Aus $Ax = 0$ folgt $x = 0$ als einzige Lösung) □

**Korollar 6.2.** *Zur Zerlegung $\Delta$ und den Werten $f_0, \ldots, f_N \in \mathbb{R}$ gibt es jeweils genau einen interpolierenden kubischen Spline mit den oben diskutierten Randbedingunen.*

*Beweis.* Die jeweiligen Koeffizientenmatrizen sind strikt diagonal dominant $\rightsquigarrow s_k''$ eindeutig $\rightsquigarrow$ Existenz und Eindeutigkeit der kubischen Splines. $\square$

## 6.10 Fehlerabschätzungen für interpolierende kubische Splines

Zuerst schreiben wir die Gleichungen (6.23) für die Momente durch die jeweilige Division durch $3(h_{k-1} + h_k)$ in der Form

$$\begin{aligned} &\frac{h_{k-1}}{3(h_{k-1}+h_k)} s''_{k-1} + \frac{2}{3} s''_k + \frac{h_k}{3(h_{k-1}+h_k)} s''_{k+1} \\ &= 2\frac{f_{k+1}-f_k}{h_k(h_{k-1}+h_k)} - 2\frac{f_k - f_{k-1}}{h_{k-1}(h_{k-1}+h_k)} =: \hat{g}_k \end{aligned}$$

auf, was für natürliche Randbedingungen auf das Gleichungssystem

$$B := \begin{bmatrix} \frac{2}{3} & \frac{h_1}{3(h_0+h_1)} & & & 0 \\ \frac{h_1}{3(h_1+h_2)} & \frac{2}{3} & \frac{h_2}{3(h_1+h_2)} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{h_{N-3}}{3(h_{N-3}+h_{N-2})} & \frac{2}{3} & \frac{h_{N-2}}{3(h_{N-3}+h_{N-2})} \\ 0 & & & \frac{h_{N-2}}{3(h_{N-2}+h_{N-1})} & \frac{2}{3} \end{bmatrix}$$

$$B \begin{bmatrix} s''_1 \\ \vdots \\ s''_{N-1} \end{bmatrix} = \begin{bmatrix} \hat{g}_1 \\ \vdots \\ \hat{g}_{N-1} \end{bmatrix} \tag{6.31}$$

führt ($h_k = x_{k+1} - x_k$). Die Eigenschaften der Matrix $B$, werden in den Fehlerabschätzungen für interpolierende kubische Splines wesentlich benutzt.

**Lemma 6.22.** *Zu einer gegebenen Funktion $f \in C^4[a,b]$ mit $f''(a) = f''(b) = 0$ bezeichne $s \in S_{\Delta,3}$ den interpolierenden kubischen Spline mit natürlichen Randbedingungen. Dann gilt*

$$\max_{k=1,\ldots,N-1} |s''(x_k) - f''(x_k)| \leq \frac{3}{4} \left\| f^{(4)} \right\|_\infty h_{\max}^2 \tag{6.32}$$

*Beweis.* (Beweisskizze)
Die Aussage des Lemmas wird unter Nutzung einer Beziehung, der Form

$$B \begin{bmatrix} f''(x_1) - s_1''(x_1) \\ \vdots \\ f''(x_{N-1}) - s_{N-1}'' \end{bmatrix} = \begin{bmatrix} \delta_1 - \hat{\delta}_1 \\ \vdots \\ \delta_{N-1} - \hat{\delta}_{N-1} \end{bmatrix} \tag{6.33}$$

nachgewiesen, die man durch Taylorentwicklungen von $f''$ und $f$ erhält, wobei $\delta_j$ und $\hat{\delta}_j$ jeweils von der Ordnung $\mathbf{O}(h_{max}^2)$, $h_{max} = \max_{k=0,\ldots,N-1} x_{k+1} - x_k$, sind.
Für die strikt diagonal dominante Matrix $B$ kann man die Abschätzung

$$\|x\|_\infty \leq (|b_{kk}| - \sum_{k\neq j=1}^{N-1} |b_{kj}|)^{-1} \|Bx\|_\infty$$

nachweisen (Übung), und mit

$$|b_{kk}| - \sum_{k\neq j=1}^{N-1} |b_{kj}| = \frac{2}{3} - \frac{h_k}{3(h_{k+1}+h_k)} - \frac{h_{k+1}}{3(h_{k+1}+h_k)} = \frac{1}{3}\,,\ k = 2,\ldots,N-2,$$

erhält man letztendlich die Beziehung (6.32) (die erste und die letzte Gleichung des Systems (6.33) sind auf Grund der Randbedingungen trivial). □

Das Lemma 6.22 ist die Grundlage für den folgenden Satz zur Fehlerabschätzung der Spline-Interpolation

**Satz 6.23.** *Sei $f \in C^4[a,b]$ und sei $s \in S_{\Delta,3}$ ein interpolierender kubischer Spline. Weiter bezeichne $h_k = x_{k+1} - x_k$ für $k = 0,\ldots,N-1$ und*

$$h_{\max} = \max_{k=0,\ldots,N-1} h_k, \quad h_{\min} = \min_{k=0,\ldots,N-1} h_k$$

*Falls*

$$\max_{k=0,\ldots,N} |s''(x_k) - f''(x_k)| \leq C \left\|f^{(4)}\right\|_\infty h_{\max}^2$$

*erfüllt ist mit einer Konstanten $C > 0$, so gelten mit der Zahl $c := \frac{h_{\max}}{h_{\min}}(C+\frac{1}{4})$ die folgenden Abschätzungen für jedes $x \in [a,b]$*

$$|s(x) - f(x)| \leq c \left\|f^{(4)}\right\|_\infty h_{\max}^4 \tag{6.34}$$

$$|s'(x) - f'(x)| \leq c \left\|f^{(4)}\right\|_\infty h_{\max}^3 \tag{6.35}$$

$$|s''(x) - f''(x)| \leq c \left\|f^{(4)}\right\|_\infty h_{\max}^2 \tag{6.36}$$

$$\left|s^{(3)}(x) - f^{(3)}(x)\right| \leq c \left\|f^{(4)}\right\|_\infty h_{\max} \tag{6.37}$$

*Beweis.* Zuerst wird (6.37) nachgewiesen. $s''$ ist als 2. Ableitung eines Polynoms 3. Grades affin linear auf $[x_k, x_{k+1}]$ für $k = 0, \ldots, N-1$, d.h.

$$s^{(3)}(x) \equiv \frac{s''(x_{k+1}) - s''(x_k)}{h_k} = \text{const}, \quad x_k \le x \le x_{k+1} \tag{6.38}$$

Taylorentwicklung von $f''$ um $x \in [x_k, x_{k+1}]$ liefert

$$f^{(3)}(x) = \frac{f''(x_{k+1}) - f''(x_k)}{h_k} - \frac{(x_{k+1} - x)^2}{2h_k} f^{(4)}(\alpha_k) + \frac{(x - x_k)^2}{2h_k} f^{(4)}(\beta_k) \tag{6.39}$$

für gewisse Zwischenstellen $\alpha_k, \beta_k \in [a, b]$. Subtraktion von (6.38) und (6.39) ergibt

$$\begin{aligned} s^{(3)}(x) - f^{(3)}(x) = &\frac{s''(x_{k+1}) - f''(x_{k+1})}{h_k} - \frac{s''(x_k) - f''(x_k)}{h} \\ &+ \frac{(x_{k+1} - x)^2 f^{(4)}(\alpha_k) - (x - x_k)^2 f^{(4)}(\beta_k)}{2h_k} \end{aligned}$$

$$\begin{aligned} \rightsquigarrow \quad &\left| s^{(3)}(x) - f^{(3)}(x) \right| \\ &\le \left\| f^{(4)} \right\|_\infty \frac{1}{\min\{h_0, \ldots, h_{N-1}\}} (c h_{\max}^2 + c h_{\max}^2 + \frac{h_{\max}^2}{2}) \\ &\le \frac{h_{\max}}{h_{\min}} (2C + \frac{1}{2}) \left\| f^{(4)} \right\|_\infty h_{\max} = 2c \left\| f^{(4)} \right\|_\infty h_{\max} \end{aligned}$$

wobei

$$\begin{aligned} (x_{k+1} - x)^2 + (x - x_k)^2 &= (x_{k+1} - x_k)^2 - 2(x_{k+1} - x)(x - x_k) \\ &\le (x_{k+1} - x_k)^2 \le h_{\max}^2 \quad \forall x \in [x_k, x_{k+1}] \end{aligned}$$

berücksichtigt wurde.
Die restlichen Fehlerabschätzungen (6.36), (6.35), (6.34) erhält man durch sukzessive Integration von (6.37) unter Nutzung des Hauptsatzes der Differential- und Integralrechnung. □

**Bemerkung.** Die wesentliche Voraussetzung des eben bewiesenen Satzes über den Fehler der 2. Ableitungen in den Knoten ist typischerweise erfüllt (siehe auch Hilfssatz 6.22 für den Fall natürlicher Randbedingungen).

## 6.11 Trigonometrische Interpolation

14. Vorlesung am 05.12.2011

Werden periodische Vorgänge "gemessen" oder vermutet man, dass gegebene Stützpunkte zu einer periodischen Funktion gehören, dann bietet sich eine Interpolation durch trigonometrische Funktionen an. O.B.d.A. nehmen wir als periode $T = 2\pi$ an und betrachten das Intervall $[0, 2\pi]$ (sonst Transformation)
Zerlegung:

$$\Delta = \{0 = x_0 < \ldots < x_{n-1} < 2\pi\}$$

mit $x_k = \frac{k}{n}2\pi, k = 0, \ldots, n-1$
Es wird folgender trigonometrischer Ansatz gemacht:

$$\Psi(x) = \begin{cases} \frac{A_0}{2} + \sum_{l=1}^{m} \left(A_l \cos(lx) + B_l \sin(lx)\right), & n = 2m+1 \\ \frac{A_0}{2} + \sum_{l=1}^{m-1} \left(A_l \cos(lx) + B_l \sin(lx)\right) + \frac{A_m}{2}\cos(mx), & n = 2m \end{cases} \tag{6.40}$$

Die Funktion $\Psi(x)$ soll die Interpolationsbedingung

$$\Psi(x_k) = f_k, \quad k = 0, \ldots, n-1 \tag{6.41}$$

mit gegebenen Werten $f_k \in \mathbb{R}$ erfüllen, wobei die Koeffizienten $A_l, B_l$ gesucht sind.
Man kann zwar $A_l, B_l$ aus (6.40) durch Auswertung von (6.41) bestimmen, aber im Komplexen wird es übersichtlicher. Mit

$$\cos\phi = \frac{1}{2}(e^{i\phi} + e^{-i\phi}), \quad \sin\phi = \frac{1}{2i}(e^{i\phi} - e^{-i\phi})$$

folgt nämlich:

$$\cos lx_k = \frac{1}{2}\left(e^{ilx_k} + e^{-ilx_k}\right) = \frac{1}{2}\left(\left(e^{\frac{2\pi il}{n}}\right)^k + \left(e^{\frac{-2\pi il}{n}}\right)^k\right), \quad x_k = \frac{2\pi k}{n}$$

bzw.

$$\sin lx_k = \frac{1}{2i}\left(\left(e^{\frac{2\pi il}{n}}\right)^k - \left(e^{\frac{-2\pi il}{n}}\right)^k\right) \tag{6.42}$$

**Bemerkung.** Wegen der $2\pi$-Periodizität von $e^{i\phi}$ gilt

$$e^{\frac{-2\pi l}{n}i} = e^{\left(\frac{-2\pi l}{n}+2\pi\right)i} = e^{\left(-\frac{2\pi l}{n}+\frac{2\pi n}{n}\right)i} = e^{\frac{(n-l)2\pi}{n}i}$$

Also brauchen keine negativen Potenzen betrachtet zu werden, sondern nur Terme

$$e^{lix_k}, \quad l = 0, \ldots, n-1$$

(6.42) wird in den Ansatz (6.40) eingesetzt, etwas umgeordnet, sodass man mit

$$p(x) = \beta_0 + \beta_1 e^{ix} + \cdots + \beta_{n-1} e^{i(n-1)x} \tag{6.43}$$

ein trigonometrisches Polynom erhält, welches die Interpolationsbedingung erfüllt, d.h.

$$\Psi(x_k) = f_k \Leftrightarrow p(x_k) = f_k, \quad k = 0, \ldots, n-1$$

($\Psi$ und $p$ stimmen nur in den Stützstellen $x_k$ überein, allerdings gilt $\Psi(x) = p(x)$ nicht für beliebige $x$).
Für die Beziehungen zwischen $\beta_k$ und $A_k, B_k$ ergeben sich einfache Formeln, z.B. für $n = 2m + 1$

$$\begin{aligned} \beta_0 = \frac{A_0}{2}, \quad \beta_j = \frac{1}{2}(A_j - iB_j), \quad \beta_{n-j} = \frac{1}{2}(A_j + iB_j), \quad j = 1, \ldots, m \\ A_0 = 2\beta_0, \quad A_l = \beta_l + \beta_{n-l}, \quad B_l = i(\beta_l - \beta_{n-l}), \quad l = 1, \ldots, m \end{aligned}$$

Setzt man $\omega = e^{ix}$, so folgt

$$p(x) = \beta_0 \omega^0 + \beta_1 \omega^1 + \cdots + \beta_{n-1} \omega^{n-1} =: P(\omega) \tag{6.44}$$

Und $P$ ist tatsächlich Polynom in $\omega$.

**Definition 6.24.**

$$\omega := e^{ix}, \quad \omega_k = e^{ix_k} \left( = e^{i\frac{2k\pi}{n}} \right)$$

**Bemerkung.** Wir haben oben $f_k \in \mathbb{R}$ gefordert, darauf kann man auch verzichten und $f_k$ auch aus $\mathbb{C}$ vorgeben.

**Satz 6.25.** *Zu beliebigen Stützstellen $(x_k, f_k), k = 0, \ldots, n-1, f_k \in \mathbb{C}, x_k = k\frac{2\pi}{n}$ gibt es genau ein trigonometrisches Polynom der Form* (6.44) *mit*

$$p(x_k) = f_k = P(\omega_k), \quad k = 0, \ldots, n-1$$

*Dabei gelten die wichtigen Beziehungen*

*(i)* $\omega_k^j = \omega_j^k, \quad \omega_k^{-l} = \overline{\omega_k^l}$

*(ii)* $\sum_{k=0}^{n-1} \omega_k^j \omega_k^{-l} = \begin{cases} n & j = l \\ 0 & j \neq l, 0 \leq l, j \leq n-1 \end{cases}$

*Beweis.* Die Existenz des Polynoms und die Eindeutigkeit folgt analog dem Nachweis der Existenz und Eindeutigkeit der allgemeinen reellen Polynominterpolation (z.B. Lagange-Interpolation)

zu (i) nach Definition

zu (ii) Ist $j = l$

$$\sum_{k=0}^{n-1} \underbrace{\omega_k^j \omega_k^{-l}}_{=1} = \sum_{k=0}^{n-1} 1 = n$$

Weiterhin ist $\omega_p = e^{\frac{2p\pi}{n}i}$ eine der $n$-ten Einheitswurzeln und damit

$$(\omega_p)^n - 1 = 0$$

Ausklammern von $\omega_k - 1$ ergibt

$$(\omega_p - 1)(\omega_p^{n-1} + \omega_p^{n-2} + \cdots + 1) = 0 \tag{6.45}$$

Man findet nun

$$\sum_{k=0}^{n-1} \omega_k^j \omega_k^{-l} = \sum_{k=0}^{n-1} \omega_k^{j-l} \overset{(i)}{=} \sum_{k=0}^{n-1} \omega_{j-l}^k = \sum_{k=0}^{n-1} (\omega_{j-l})^k$$

und da $j \neq l$, ist $\omega_{j-l} \neq 1$, d.h. $\sum(\omega_{j-l})^k$ muss als 2. Faktor der linken Seite von (6.45) $= 0$ sein. □

Aus dem eben bewiesenen Satz ergibt sich die Folgerung

**Korollar.** *Die komplexen Vektoren*

$$\phi_j = \begin{pmatrix} \omega_0^j \\ \vdots \\ \omega_{n-1}^j \end{pmatrix}, \ \phi_l = \begin{pmatrix} \omega_0^l \\ \vdots \\ \omega_{n-1}^l \end{pmatrix} \in \mathbb{C}^n, \quad (\phi_j)_k = \omega_k^j, \quad j \neq l$$

*sind bezüglich des Skalarproduktes*

$$\langle f, g \rangle := \frac{1}{n} \sum_{k=0}^{n-1} f_k \bar{g}_k \tag{6.46}$$

*zueinander orthogonal, d.h.* $\{\phi_0, \ldots, \phi_{n-1}\}$ *ist Orthogonalsystem in* $\mathbb{C}^n$

**Definition 6.26.** *Die Koeffizienten* $\beta_0, \ldots, \beta_{n-1}$ *aus* (6.44)*, d.h. die Koeffizienten von* $P(\omega)$ *heißen* ***Fourierkoeffizienten*** *oder* ***diskrete Fouriertransformierte*** *von* $f_0, \ldots, f_{n-1}$ *falls* $P(\omega_k) = f_k, k = 0, \ldots, n-1$ *gilt.*

**Satz 6.27.** *Für die diskreten Fouriertransformierten $\beta_j$ von $f_j, j = 1, \ldots, n-1$ gilt*

$$\beta_j = \frac{1}{n} \sum_{k=0}^{n-1} f_k e^{-i\frac{jk2\pi}{n}} \tag{6.47}$$

*d.h. sie sind eindeutig bestimmt.*

*Beweis.* Die Interpolationsbedingungen $P(\omega_k) = f_k$ bedeuten

$$\begin{aligned} P(\omega_0) &= \beta_0\omega_0^0 + \cdots + \beta_{n-1}\omega_0^{n-1} = f_0 \\ &\vdots \\ P(\omega_{n-1}) &= \beta_0\omega_{n-1}^0 + \cdots + \beta_{n-1}\omega_{n-1}^{n-1} = f_{n-1} \\ \rightsquigarrow \quad \beta_0\phi_0 + \beta_1\phi_1 + \cdots + \beta_{n-1}\phi_{n-1} &= f := \begin{pmatrix} f_0 \\ \vdots \\ f_{n-1} \end{pmatrix} \end{aligned} \tag{6.48}$$

die skalare Multiplikation mit $\phi_j$ ergibt aufgrund der Orthogonalität

$$\beta_j \langle \phi_j, \phi_j \rangle = \langle f, \phi_j \rangle = \frac{1}{n} \sum_{k=0}^{n-1} f_k \overline{\omega_k^j} = \frac{1}{n} \sum_{k=0}^{n-1} f_k \omega_k^{-j} = \frac{1}{n} \sum_{k=0}^{n-1} f_k e^{-i\frac{2kj\pi}{n}}$$

□

**Bemerkung.** Für die Fourierkoeffizienten oder diskreten Fouriertransformierten $\beta_k$ von $f_k$ wird auch die Notation

$$\mathcal{F}[f_0, \ldots, f_{n-1}] := [\beta_0, \ldots, \beta_{n-1}] \tag{6.49}$$

verwendet.
(6.48) bedeutet das Gleichungssystem

$$\underbrace{\begin{pmatrix} \omega_0^0 & w_0^1 & \ldots & \omega_0^{n-1} \\ \vdots & & & \\ \omega_{n-1}^0 & w_{n-1}^1 & \ldots & \omega_{n-1}^{n-1} \end{pmatrix}}_{=:V=(\omega_k^j)_{j,k=0,\ldots,n-1}} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{n-1} \end{pmatrix} = \begin{pmatrix} f_0 \\ \vdots \\ f_{n-1} \end{pmatrix} \tag{6.50}$$

bzw.

$$\begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{n-1} \end{pmatrix} = \underbrace{\frac{1}{n} \begin{pmatrix} \omega_0^{-0} & w_0^{-1} & \ldots & \omega_0^{-(n-1)} \\ \vdots & & & \\ \omega_{n-1}^{-0} & w_{n-1}^{-1} & \ldots & \omega_{n-1}^{-(n-1)} \end{pmatrix}}_{=\frac{1}{n}\bar{V}=(\frac{1}{n}\omega_k^{-j})_{j,k=0,\ldots,n-1}} \begin{pmatrix} f_0 \\ \vdots \\ f_{n-1} \end{pmatrix} \tag{6.51}$$

**Korollar.** *(i) Es gilt offensichtlich*

$$(\frac{1}{n}\bar{V})^{-1} = V$$

*und jeder Datensatz $f_0, \ldots, f_{n-1} \in \mathbb{C}$ lässt sich aus seiner diskreten Fouriertransformierten*

$$\mathcal{F}[f_0, \ldots, f_{n-1}] = [\beta_0, \ldots, \beta_{n-1}]$$

*durch (siehe (6.48))*

$$f_j = \sum_{k=0}^{n-1} \beta_k e^{i\frac{jk2\pi}{n}}, \quad j = 0, \ldots, n-1$$

*zurückgewinnen. Es wird auch die Notation*

$$\mathcal{F}^{-1}[\beta_0, \ldots, \beta_{n-1}] = [f_0, \ldots, f_{n-1}]$$

*verwendet.*

*(ii) Es gilt*

$$\sum_{k=0}^{n-1} |\beta_k|^2 = \frac{1}{n} \sum_{k=0}^{n-1} |f_k|^2$$

### 6.11.1 Beziehungen zwischen den reellen und komplexen Fourierkoeffizienten $A_j, B_j, \beta_j$

Es galt $\Psi(x_k) = f_k$ und außerdem war $\omega_k = e^{-ix_k}$ definiert. Für ungerades $n = 2m + 1$ folgt

$$\begin{aligned}
\Psi(x_k) &= \frac{A_0}{2} + \sum_{l=1}^{m} \left( A_l \frac{1}{2}(\omega_k^l + \omega_k^{-l}) + B_l \frac{1}{2i}(\omega_k^l - \omega_k^{-l}) \right) \\
&= \frac{A_0}{2} + \sum_{l=1}^{m} \left( A_l \frac{1}{2}(\omega_k^l + \omega_k^{n-l}) + B_l \frac{1}{2i}(\omega_k^l - \omega_k^{n-l}) \right) \\
&= \beta_0 + \beta_1 \omega_k + \cdots + \beta_{n-1} \omega_k^{n-1}
\end{aligned}$$

Daraus folgt

$$A_0 = 2\beta_0 \Leftrightarrow \beta_0 = \frac{A_0}{2}$$

sowie

$$\beta_l = \frac{1}{2}(A_l + \frac{1}{i}B_l) = \frac{1}{2}(A_l - iB_l), \quad l = 1, \ldots, m$$
$$\beta_{n-l} = \frac{1}{2}(A_l - \frac{1}{i}B_l) = \frac{1}{2}(A_l + iB_l), \quad l = 1, \ldots, m$$

$$\rightsquigarrow \quad A_l = \beta_l + \beta_{n-l}, \quad B_l = i(\beta_l - \beta_{n-l}), \quad l = 1, \ldots, m$$

Mit der Formel (6.47) folgt:

$$\begin{aligned} A_l &= \frac{1}{n}\sum_{k=0}^{n-1} f_k \left(e^{-i\frac{kl2\pi}{n}} + e^{-i\frac{k(n-l)2\pi}{n}}\right) \\ &= \frac{2}{n}\sum_{k=0}^{n-1} f_k \frac{1}{2}\left(e^{-i\frac{kl2\pi}{n}} + e^{i\frac{kl2\pi}{n}}\right) = \frac{2}{n}\sum_{k=0}^{n-1} f_k \cos(lx_k) \end{aligned} \tag{6.52}$$

und analog

$$B_l = \frac{2}{n}\sum_{k=0}^{n-1} f_k \sin(lx_k)$$

Die Betrachtungen für gerades $n = 2m$ verlaufen analog.
Zusammengefasst ergibt sich

**Satz 6.28.** *Werden die Koeffizienten gemäß* (6.52) *bestimmt, so erfüllt*

$$\Psi(x) = \begin{cases} \frac{A_0}{2} + \sum_{k=0}^{m}(A_l \cos(kx) + B_k \sin(kx)), & n = 2m+1 \\ \frac{A_0}{2} + \sum_{k=0}^{m-1}(A_l \cos(kx) + B_k \sin(kx)) + \frac{A_m}{2}\cos(mx,) & n = 2m \end{cases}$$

*Die Interpolationsbedingung*

$$\Psi(x_k) = f_k, \quad k = 0, \ldots, n-1$$

*für reelle $f_k$.*

Ziel ist die Reduzierung des Aufwands zur Berechnung der diskreten Fouriertransformierten $\beta_0, \ldots, \beta_{n-1}$ für einen Datensatz $f_0, \ldots, f_{n-1}$ der mit der Auswertung der Berechnungsvorschrift

$$\beta_j = \frac{1}{n}\sum_{k=0}^{n-1} f_k e^{-i\frac{jk2\pi}{n}}, \quad j = 0, \ldots, n-1$$

etwa $\mathcal{O}(n^2)$ komplexe Multiplikationen bedeutet.

### 6.11.2 Schnelle Fouriertransformation (FFT)

**Voraussetzung** $n = 2^p, p \in \mathbb{N}$, d.h. es werden Datensätze mit $n = 2^p$ Daten aus $\mathbb{C}$ betrachtet. Entscheidende Grundlage für die FFT ist der folgende

**Satz 6.29.** *Aus den diskreten Fouriertransformierten der beiden Datensätze*

$$g_0, \ldots, g_{M-1} \quad \textit{und} \quad g_M, \ldots, g_{2M-1}$$

*der Länge $M$ lässt sich die diskrete Fouriertransformierte des Datensatzes*

$$g_0, \ldots, g_{2M-1}$$

*der Länge $2M$ folgendermaßen bestimmen.*

$$\begin{aligned} &\frac{1}{2}\left\{\mathcal{F}_k[g_0, \ldots, g_{M-1}] + e^{-i\frac{k\pi}{M}}\mathcal{F}_k[g_M, \ldots, g_{2M-1}]\right\} \\ &= \mathcal{F}_k[g_0, g_M, g_1, g_{M+1} \ldots, g_{2M-1}] \end{aligned} \tag{6.53}$$

$$\begin{aligned} &\frac{1}{2}\left\{\mathcal{F}_k[g_0, \ldots, g_{M-1}] - e^{-i\frac{k\pi}{M}}\mathcal{F}_k[g_M, \ldots, g_{2M-1}]\right\} \\ &= \mathcal{F}_{M+k}[g_0, g_M, g_1, g_{M+1} \ldots, g_{2M-1}] \end{aligned} \tag{6.54}$$

*Für $k = 0, \ldots, M-1$. Wobei $\mathcal{F}_k$ bzw. $\mathcal{F}_{M+k}$ die $k$-te bzw. $(M+k)$-te Komponente von $\mathcal{F}$ bezeichnen.*

*Beweis.* Für $k = 0, \ldots, M-1$ gilt

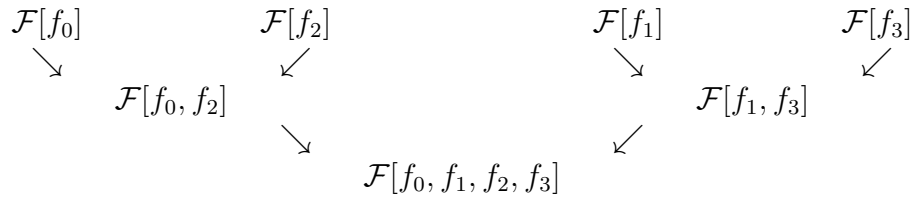
$$\begin{aligned} \mathcal{F}_k[g_0, \ldots, g_{2M-1}] &= \frac{1}{2M}\left(\sum_{j=0}^{M-1} g_j e^{-i\frac{2jk2\pi}{2M}} + \sum_{j=0}^{M-1} g_{M+j} e^{-i\frac{(2j+1)k2\pi}{2M}}\right) \\ &= \frac{1}{2M}\left(\sum_{j=0}^{M-1} g_j e^{-i\frac{jk2\pi}{M}} + e^{-i\frac{k\pi}{M}} \sum_{j=0}^{M-1} g_{M+j} e^{-i\frac{jk2\pi}{M}}\right) \end{aligned}$$

Die Gleichung (6.54) erhält man analog, wobei

$$e^{-i\frac{j(k+M)2\pi}{2M}} = e^{-ij\pi} e^{-i\frac{jk2\pi}{2M}} = (-1)^j e^{-i\frac{jk2\pi}{2M}}$$

berücksichtigt wird. □

Ist $n = 2^p$, dann soll der Satz 6.29 auf einem Datensatz dieser Länge rekursiv angewandt werden. Die Anordnung der Daten wird später erklärt.




Erläuterungen zum Schema

(a) Beim Übergang von Stufe 0 zu Stufe 1 werden 2 diskrete Fouriertransformierte der Länge 2 ausgehend von 4 diskreten Fouriertransformierten der Länge 1 berechnet (Anwendung der Formeln (6.53), (6.54) je 2-mal).

(b) Beim Übergang von Stufe 1 zu Stufe 2 wird 1 diskrete Fouriertransformierte der Länge 4 ausgehend von 2 diskreten FTs der Länge 2 berechnet (zweimalige Anwendung der Formeln (6.53), (6.54))

(c) Schließlich erhält man ausgehend von diesen die gewünschte diskrete FT des Datensatzes $f_0, \ldots, f_3$

(d) Entscheidend für genau dieses Ergebnis war die Anordnung der Daten auf der Stufe 0

(e) Die Anwendung des Satzes 6.29 soll beim Übergang von Stufe 2 zu Stufe 3 erläutert werden:

Setzt man

$$g_0 = f_0, g_1 = f_2, g_2 = f_1, g_3 = f_3$$

dann erhält man ausgehend von

$$\mathcal{F}[g_0, g_2] \quad \text{und} \quad \mathcal{F}[g_1, g_3]$$

mit den Formeln (6.53),(6.54)

$$\mathcal{F}[g_0, g_2, g_1, g_3]$$

also bei Berücksichtigung der Setzungen

$$\mathcal{F}[f_0, f_1, f_2, f_3]$$

**Bemerkung 6.30.** Für Anordnung der Daten auf der Stufe 0 nutzt man das folgende Schema der Bit-Umkehr, die in der folgenden Tabelle für $n = 8 = 2^3$ beschrieben wird:

| $f_k$ Index | Binärwert | Binärwert revers | Index |
|---|---|---|---|
| 0 | 000 | 000 | 0 |
| 1 | 001 | 100 | 4 |
| 2 | 010 | 010 | 2 |
| 3 | 011 | 110 | 6 |
| 4 | 100 | 001 | 1 |
| 5 | 101 | 101 | 5 |
| 6 | 110 | 011 | 3 |
| 7 | 111 | 111 | 7 |

In der letzten Spalte liest man die Indexreihenfolge für die Anordnung der Daten auf der Stufe 0 ab.

### 6.11.3 Aufwand der FFT

Zum Abschluss der Thematik FFT soll nun der Aufwand diskutiert werden.

Bezeichnet man die Stufen der FFT mit $r \in \{0, 1, \ldots, p\}$, also im Fall $8 = n = 2^p = 2^3$ $r \in \{0, 1, 2, 3\}$, dann ergibt sich für den Aufwand der FFT:

Für $r \in \{0, \ldots, p-1\}$ fallen beim Übergang von der $r$-ten zur $(r+1)$-ten Stufe der FFT die folgenden komplexen Multiplikationen an

- Die Berechnung von Zahlen $\omega^2, \ldots, \omega^{2^r-1} \in \mathbb{C}$ ($\omega$ Wert einer komplexen Exponentialfunktion) erfordert $2^r - 2 \leq 2^r$ komplexe Multiplikationen (Faktoren in den Formeln (6.53), (6.54))

- Berechnung der diskreten Fouriertransformierten der Länge $2^{r+1}$ ausgehend von je 2 diskreten Fouriertransformierten der Länge $2^r$, und das insgesamt $2^p - r - 1$-mal ergibt $2^n \cdot 2^{p-r-1} = 2^{p-1}$ komplexe Multiplikationen

- Dazu kommen noch $p - 2 \leq p$ komplexe Multiplikationen zur Berechnung etwa von $\omega_k = \omega_{k+1}^2$

- Für die Ausführung der Übergänge von den Stufen 0 bis p ergibt sich die Gesamtzahl an komplexen Multiplikationen

$$\sum_{r=0}^{p-1}(2^{p-1} + 2^r) + p \leq p2^{p-1} + 2^p + p = \frac{n \log_2 n}{2} + \mathcal{O}(n)$$

Damit gilt der

**Satz 6.31.** *Bei der FFT zur Bestimmung der diskreten Fouriertransformierten eines Datensatzes der Länge $n = 2^p$ fallen nicht mehr als*

$$\frac{n \log_2 n}{2} + \mathcal{O}(n)$$

*komplexe Multiplikationen an.*

**Bemerkung 6.32.** Wir haben für die Fouriertransformation die Formeln

$$\beta_j = \frac{1}{n} \sum_{k=0}^{n-1} f_k e^{-i\frac{jk2\pi}{n}} \tag{6.55}$$

$$f_j = \frac{1}{n} \sum_{k=0}^{n-1} \beta_k e^{i\frac{jk2\pi}{n}}, \quad j = 0, \ldots, n-1$$

für die Hin- resp. Rücktransformation hergeleitet. In vielen Lehrbüchern sind die diskreten Fourierkoeffizienten durch

$$\beta_j = \sum_{k=0}^{n-1} f_k e^{-i\frac{jk2\pi}{n}} \tag{6.56}$$

definiert, also ohne den Faktor $\frac{1}{n}$. Das hat für die Rücktransformation die Konsequenz

$$f_j = \frac{1}{n} \sum_{k=0}^{n-1} \beta_k e^{i\frac{jk2\pi}{n}}, \quad j = 0, \ldots, n-1$$

Eine dritte Möglichkeit ist durch

$$\beta_j = \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} f_k e^{-i\frac{jk2\pi}{n}} \tag{6.57}$$

$$f_j = \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} \beta_k e^{i\frac{jk2\pi}{n}}, \quad j = 0, \ldots, n-1$$

gegeben.

Besonders bei der Nutzung von Numerikprogrammsystemen oder Bibliotheken ist es daher ratsam, die jeweils verwendete Definition der Fouriertransformation und der Rücktransformation zu ermitteln, also (6.55), (6.56) oder (6.57).

# Kapitel 7

# Numerische Integration

15. Vorlesung am 07.12.2011

Ziel ist die Berechnung des bestimmten Integrals

$$\int_a^b f(x)\mathrm{d}x$$

wobei man aus unterschiedlichen Gründen nicht die Berechnung mittels einer Stammfunktion $F(x)$ durch

$$\int_a^b f(x)\mathrm{d}x = F(b) - F(a)$$

nutzen kann oder will. Entweder findet man kein auswertbares $F(x)$ wie im Fall von $f(x) = \frac{e^x}{x}$ oder $f(x) = e^{-x^2}$ oder die Berechnung von $F(b), F(a)$ ist zu mühselig.

## 7.1 Numerischen Integration mit Newton-Cotes-Formeln

- Äquidistante Unterteilung von $[a, b]$

$$x_k = a + kh, \quad k = 0, ..., n, h = \frac{b-a}{n}$$

- Verwendung des Interpolationspolynoms $p_n \in \Pi_n$ für die Stützpunkte $(x_k, f(x_k))$, d.h. es ist

$$p_n(x_k) = f(x_k), \quad k = 0, ..., n$$

- Näherung des Integrals $\int_a^b f(x)\mathrm{d}x$ durch

$$\int_a^b p_n(x)\mathrm{d}x \approx \int_a^b f(x)\mathrm{d}x$$

Mit dem Lagrangschen Interpolationspolynom

$$p_n(x) = \sum_{k=0}^{n} f_k L_k(x), \quad f_k = f(x_k)$$

erhält man

$$\begin{aligned}\int_a^b p_n(x)\mathrm{d}x &= \sum_{k=0}^{n} f_k \int_a^b L_k \mathrm{d}x \\ &= \sum_{k=0}^{n} f_k \int_a^b \prod_{k\neq j=0}^{n} \frac{x-x_j}{x_k-x_j}\mathrm{d}x = (*)\end{aligned}$$

und mit der Substitution $s = \frac{x-a}{h}$, $h\mathrm{d}s = \mathrm{d}x$ folgt

$$(*) = (b-a)\sum_{k=0}^{n} f_k \underbrace{\frac{1}{n}\int_0^n \prod_{k\neq j=0}^{n} \frac{s-j}{k-j}\mathrm{d}s}_{\sigma_k}$$

also

$$\int_a^b p_n(x)\mathrm{d}x = (b-a)\sum_{k=0}^{n} f_k\sigma_k \tag{7.1}$$

mit den Gewichten

$$\sigma_k = \frac{1}{n}\int_0^n \prod_{k\neq j=0}^{n} \frac{s-j}{k-j}\mathrm{d}s, \quad k = 0,\ldots,n \tag{7.2}$$

Für $n = 1$ erhält man

$$\sigma_0 = \int_0^1 \frac{s-1}{0-1}\mathrm{d}s = -\frac{1}{2}(s-1)^2\Big|_0^1 = \frac{1}{2}, \quad \sigma_1 = \frac{1}{2}$$

woraus mit

$$\int_a^b f(x)\mathrm{d}x \approx \int_a^b p_1(x)\mathrm{d}x = \frac{b-a}{2}(f(a)+f(b)) \tag{7.3}$$

die **Trapezregel** folgt.
Für $n = 2$ ergibt sich

$$\begin{aligned}\sigma_0 &= \frac{1}{2}\int_0^2 \frac{s-1}{0-1}\cdot\frac{s-2}{0-2} = \frac{1}{4}\int_0^2 (s^2-3s+2)\mathrm{d}s \\ &= \frac{1}{4}\left[\frac{s^3}{3}-\frac{3s^2}{2}+2s\right] = \frac{1}{4}\left[\frac{8}{3}-6+4\right] = \frac{1}{4}\left[\frac{8-6}{3}\right] = \frac{1}{6}\end{aligned}$$

$$\sigma_2 = \frac{1}{6}, \quad \sigma_1 = \frac{4}{6}$$

woraus mit

$$\int_a^b f(x)\mathrm{d}x \approx \int_a^b p_2(x)\mathrm{d}x = \frac{b-a}{6}\left(f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right) \tag{7.4}$$

Die **Simpson-Regel**, auch **Keplersche Fassregel** genannt, folgt.
Für $n = 3$ findet man auf analoge Weise mit

$$\begin{aligned}\int_a^b f(x)\mathrm{d}x &\approx \int_a^b p_3(x)\mathrm{d}x \\ &= \frac{b-a}{8}\left(f(a)+3f\left(\frac{2a+b}{3}\right)+3f\left(\frac{a+2b}{3}\right)+f(b)\right)\end{aligned} \tag{7.5}$$

die Newtonsche $\frac{3}{8}$-Regel.

**Definition 7.1.** *Die Näherungsformel*

$$Q_n(x) = \int_a^b p_n(x)\mathrm{d}x = (b-a)\sum_{k=0}^{n} f(x_k)\sigma_k \tag{7.6}$$

zu den Stützstellen $x_0, \ldots, x_n$ für das Integral $\int_a^b f(x)\mathrm{d}x$ nennt man **interpolatorische Quadraturformel**.
Gilt für die Stützstellen $x_k = a+kh, h = \frac{b-a}{n}, k = 0, \ldots, n$ spricht man bei der Quadraturformel von einer **abgeschlossenen Newton-Cotes-Quadraturformel**.

**Definition 7.2.** *Mit*

$$E_n[f] = \int_a^b f(x)\mathrm{d}x - Q_n = I - Q_n \tag{7.7}$$

bezeichnet man den Fehler der Quadraturformel $Q_n$. Eine Quadraturformel hat den Genauigkeitsgrad $m \in \mathbb{N}$, wenn sie alle Polynome $p(x)$ bis zum Grad $m$ exakt integriert, d.h. $E_n[p] = 0$ ist, und $m$ die größtmögliche Zahl mit dieser Eigenschaft ist.

Es gilt offensichtlich der folgende

**Satz 7.3.** *Zu den $n+1$ beliebig vorgegebenen paarweise verschiedenen Stützstellen $a \leq x_0 < \cdots < x_n \leq b$ existiert eine eindeutig bestimmte interpolatorische Quadraturformel deren Genauigkeitsgrad mindestens gleich $n$ ist.*

Für die Simpsonregel findet man

$$\begin{aligned} E_2[x^3] &= \int_a^b x^3 \mathrm{d}x - \frac{b-a}{6}\left[a^3 + 4\left(\frac{a+b}{2}\right)^3 + b^3\right] \\ &= \frac{1}{4}(b^4 - a^4) - \frac{b-a}{6}\left[a^3 + \frac{1}{2}(a^3 + 3a^2b + 3ab^2 + b^3) + b^3\right] \\ &= 0 \end{aligned}$$

und

$$E_2[x^4] \neq 0$$

Aufgrund der Additivität und Homogenität des Quadraturfehlers, d.h.

$$E_n[\alpha f + \beta g] = \alpha E_n[f] + \beta E_n[g],$$

ist die Simpsonregel für alle Polynome 3. Grades exakt, allerdings nicht mehr für Polynome 4. Grades. Damit hat sie den Genauigkeitsgrad 3 obwohl ihr nur ein Interpolationspolynom vom Grad 2 zugrunde liegt.
Generell findet man, dass die abgeschlossenen Newton-Cotes Quadraturformeln $Q_n$ für gerades $n$ den Genauigkeitsgrad $n+1$ haben.
Setzt man bei der zu integrierenden Funktion $f$ die $(n+1)$- bzw. $(n+2)$-malige stetige Differenzierbarkeit voraus, dann gilt für Fehler der ersten Newton-Cotes-Quadraturformeln

$$\begin{aligned} E_1[f] &= -\frac{1}{12}h^3 f''(\eta), \quad h = b-a \\ E_2[f] &= -\frac{1}{90}h^5 f^{(4)}(\eta), \quad h = \frac{b-a}{2} \\ E_3[f] &= -\frac{3}{80}h^5 f^{(4)}(\eta), \quad h = \frac{b-a}{3} \\ E_4[f] &= -\frac{8}{945}h^7 f^{(6)}(\eta), \quad h = \frac{b-a}{4} \end{aligned}$$

wobei $\eta \in [a,b]$ jeweils ein geeigneter Zwischenwert ist.

## 7.2 Summierte abgeschlossene Newton-Cotes-Quadraturformeln

Trapezregel $(Q_1)$ und Simpsonregel $(Q_2)$ bedeutet also die Integration von $p_1$ bzw. $p_2$ zur näherungsweisen Berechnung on $I = \int_a^b f(x)\mathrm{d}x$. Bei der Inter-

polation haben wir die Erfahrung gemacht, dass Polynome höheren Grades zu Oszillationen an den Intervallrändern neigen. Man stellt auch fest, dass ab $n = 8$ negative Gewichte $\sigma_k$ auftreten.
Um die Genauigkeit zu erhöhen, verzichtet man auf die Vergrößerung von $n$ und wendet stattdessen z.B. die Trapez- oder Simpson-regel auf $N$ Teilintervallen an.
Zur näherungsweisen Berechnung von $\int_\alpha^\beta f(x)\mathrm{d}x$ unterteilt man das Intervall $[\alpha, \beta]$ durch

$$\alpha = x_{10} < \ldots < x_{1n} = x_{20} < \ldots < x_{N-1n} = x_{N0} < \ldots < x_{Nn} = \beta$$

in $N$ gleichgroße Teilintervalle $[x_{j0}, x_{jn}]$, $j = 1, \ldots, N$ mit jeweils $n + 1$ Stützstellen. Auf den Teilintervallen $[a, b] = [x_{j0}, x_{jn}]$ nähert man das Integral

$$\int_{x_{j0}}^{x_{jn}} f(x)\mathrm{d}x \quad \text{mit} \quad Q_{n,j}$$

zu den Stützstellen $x_{j0}, \ldots, x_{jn}$ an. Die Summation über $j$ ergibt mit

$$S_{n,N} = \sum_{j=1}^{N} Q_{n,j}$$

die sogenannten **summierten abgeschlossenen Newton-Cotes- Formeln**.
Mit $y_{jk} = f(x_{jk})$ erhält man für $n = 1$ die summierte Trapez- Regel ($h = \frac{\beta-\alpha}{N}$)

$$\begin{aligned} S_{1,N} &= h\left[\frac{1}{2}y_{10} + y_{20} + \cdots + y_{N0} + \frac{1}{2}y_{N1}\right] \\ &= h\left[\frac{1}{2}(y_{10} + y_{N1}) + \sum_{k=2}^{N} y_{k0}\right] \end{aligned} \tag{7.8}$$

und für $n = 2$ die aufsummierte Simpson-Regel ($h = \frac{\beta-\alpha}{2N}$)

$$S_{2,N} = \frac{h}{3}\left[(y_{10} + y_{N2}) + 2\sum_{j=1}^{N-1} y_{j2} + 4\sum_{j=1}^{N} y_{j1}\right] \tag{7.9}$$

Für die Quadraturfehler summierter abgeschlossener Newton-Cotes- Formeln gilt der

**Satz 7.4.** *Wenn $f(x)$ in $[\alpha, \beta]$ für gerades $n$ eine stetige $(n+2)$-te Ableitung und für ungerades $n$ eine stetige $(n + 1)$-te Ableitung besitzt, dann existiert ein Zwischenwert $\xi \in ]\alpha, \beta[$, sodass die Beziehungen*

$$E_{S_{n,N}}[f] = Kh^{n+2} f^{(n+2)}(\xi)$$

*für gerades $n$ und*

$$E_{S_{n,N}}[f] = Lh^{n+1} f^{(n+1)}(\xi)$$

*für ungerades $n$ gelten, wobei $K$ und $L$ von $\alpha, \beta$ abhängige Konstanten sind, und $h = \frac{\beta-\alpha}{nN}$ gilt.*

*Beweis.* Plato, Bärwolff □

## 7.3 Gauß-Quadraturen

Bei den Newton-Cotes-Quadraturformeln ist man von einer vorgegebenen Zahl von äquidistanten Stützstellen $x_0, \ldots, x_n$ ausgegangen und hat eine Näherung des Integrals $\int_{x_0}^{x_n} f(x)\mathrm{d}x$ durch das Integral des Interpolationspolynoms $p_n(x)$ für $(x_k, f(x_k)),\ k = 0, \ldots, n$ angenähert. Dabei waren als Freiheitsgrade die Integrationsgewichte $\sigma_k$ zu bestimmen.
Bei den Gauß-Quadraturformeln verzichtet man auf die Vorgabe der Stützstellen und versucht diese so zu bestimmen, dass die Näherung des Integrals besser als bei den Newton-Cotes-Formeln wird.
Bei den Gauß-Quadraturen verwendet man als Bezeichnung für die Stützstellen oft $\lambda_1, \ldots, \lambda_n$, da sie sich letztendlich als Nullstellen eines Polynoms $n$-ten Grades ergeben werden. Wir wollen sie im Folgenden aber weiter mit $x_1, \ldots, x_n$ bezeichnen und beginnen aber im Unterschied zu den Newton-Cotes-Formeln bei $k = 1$ zu zählen.

Ziel ist die Berechnung des Integrals $\int_a^b g(x)\mathrm{d}x$ wobei man die zu integrierende Funktion in der Form $g(x) = f(x)\rho(x)$ mit einer Funktion $\rho(x)$, die mit der evtl. Ausnahme von endlich vielen Punkten auf $[a, b]$ positiv sein soll, vorgibt. $\rho(x)$ heißt **Gewichtsfunktion**. Es ist also das Integral

$$I = \int_a^b f(x)\rho(x)\mathrm{d}x = \int_a^b g(x)\mathrm{d}x$$

numerisch zu berechnen. Im Folgenden geht es darum, Stützstellen $x_k \in [a, b]$ und Integrationsgewichte $\sigma_k$ so zu bestimmen, dass

$$I_n = \sum_{j=1}^{n} \sigma_j f(x_j) \tag{7.10}$$

eine möglichst gute Näherung des Integrals $I$ ergibt. Fordert man, dass die Formel (7.10) für alle Polynome $f(x)$ bis zum Grad $2n - 1$, d.h. für $x^0, x^1, \ldots, x^{2n-1}$ exakt ist und somit $I_n = I$ gilt, dann müssen die Stützstellen $x_1, \ldots, x_n$ und die Gewichte $\sigma_1, \ldots, \sigma_n$ Lösungen des Gleichungssystems

$$\sum_{j=1}^{n} \sigma_j x_j^k = \int_a^b x^k \rho(x)\mathrm{d}x \quad (k = 0, 1, \ldots, 2n-1) \tag{7.11}$$

sein.

Wir werden im Folgenden zeigen, dass das Gleichungssystem (7.11) eindeutig lösbar ist, dass für die Stützstellen $x_k \in ]a, b[$ gilt und dass die Gewichte $\sigma_k$ positiv sind.

16. Vorlesung 12.12.11

Zuerst ein

**Beispiel.** für die Berechnung von $\int_{-1}^{1} f(x)\rho(x)\mathrm{d}x$ mit der Gewichtsfunktion $\rho(x) \equiv 1$ und der Vorgabe von $n = 2$ bedeutet (7.11) mit

$$\int_{-1}^{1} \mathrm{d}x = 2, \quad \int_{-1}^{1} x\mathrm{d}x = 0, \quad \int_{-1}^{1} x^2\mathrm{d}x = \frac{2}{3}, \quad \int_{-1}^{1} x^3 = 0$$

das Gleichungssystem

$$\begin{aligned} \sigma_1 + \sigma_2 &= 2 \\ \sigma_1 x_1 + \sigma_2 x_2 &= 0 \\ \sigma_1 x_1^2 + \sigma_2 x_2^2 &= \frac{2}{3} \\ \sigma_1 x_1^3 + \sigma_2 x_2^3 &= 0 \end{aligned} \tag{7.12}$$

Für (7.12) findet man mit

$$x_1 = -\frac{1}{\sqrt{3}}, \quad x_2 = -\frac{1}{\sqrt{3}}, \quad \sigma_1 = \sigma_2 = 1$$

eine Lösung und damit ist die Quadraturformel

$$I_2 = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$

für alle Polynome $f(x)$ bis zum Grad 3 exakt, d.h. es gilt

$$\int_{-1}^{1} f(x)\mathrm{d}x = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$

Wir sind also *besser* als mit der Trapezregel.

### 7.3.1 Orthogonale Polynome

Die beiden Stützstellen aus dem eben diskutierten Beispiel sind mit $-\frac{1}{\sqrt{3}}$ und $\frac{1}{\sqrt{3}}$ gerade die Nullstellen des Legendre-Polynoms $p_2(x) = x^2 - \frac{1}{3}$ zweiten Grades. Das ist kein Zufall, sondern darin steckt eine Systematik. Deshalb sollen im Foglenden orthogonale Polynome besprochen werden.

Mit einer Gewichtsfunktion $\rho(x)$ statten wir den Vektorraum $P$ aller Polynome über dem Körpert der reellen Zahlen mit dem Skalarprodukt

$$\langle p, q \rangle_\rho := \int_a^b p(x)q(x)\rho(x)\mathrm{d}x \tag{7.13}$$

für $p, q \in P$ aus. Folglich ist durch

$$\|p\|_\rho^2 = \langle p, p \rangle_\rho = \int_a^b p^2(x)\rho(x)\mathrm{d}x \tag{7.14}$$

eine Norm definiert. Der Nachweis, dass (7.13), (7.14) Skalarprodukt bzw. Norm sind, sollte als Übung betrachtet werden.

**Definition 7.5.** *Die Polynome $p, q \in P$ heißen **orthogonal** bezüglich $\langle \cdot, \cdot \rangle_\rho$, wenn*

$$\langle p, q \rangle_\rho = 0$$

*gilt.*
*Ist $V$ ein Unterraum von $P$, dann wird durch*

$$V^\perp = \{f \in P | \langle f, p \rangle_\rho = 0 \quad \forall p \in V\}$$

*das **orthogonale Komplement** von $V$ bezeichnet.*
*Die lineare Hülle der Funktionen $p_1, \ldots, p_n \in P$ wird durch*

$$\mathrm{span}\{p_1, \ldots, p_n\} = \{c_1 p_1 + \cdots + c_n p_n | c_1, \ldots, c_n \in K\}$$

*definiert, wobei $K$ der Zahlkörper ist, über dem der Vektorraum der Polynome $P$ betrachtet wird (und wenn nichts anderes gesagt wird, betrachten wir $K = \mathbb{R}$)*

### 7.3.2 Konstruktion von Folgen orthogonaler Polynome

Wir wissen, dass die Monome $1, x, \ldots, x^n, \ldots$ eine Basis zur Konstruktion von Polynomen bilden. Mit $p_0(x) = 1$ wird durch

$$p_n(x) = x^n - \sum_{j=0}^{n-1} \frac{\langle x^n, p_j \rangle_\rho}{\langle p_j, p_j \rangle_\rho} p_j(x) \tag{7.15}$$

also mit dem Orthogonalisierungsverfahren von Gram-Schmidt eine Folge paarweise orthogonaler Polynome definiert (bezüglich des Skalarproduktes $\langle \cdot, \cdot \rangle_\rho$)

**Beispiel.** Mit $[a,b] = [-1,1]$ und $\rho(x) = 1$ erhält man ausgehend von $p_0(x) = 1$ mit

$$p_1(x) = x, \quad p_2(x) = x^2 - \frac{1}{3}, \quad p_3(x) = x^3 - \frac{3}{5}x, \quad p_4(x) = x^4 - \frac{5}{2}x^2 + \frac{4}{105} \tag{7.16}$$

paarweise orthogonaler Polynome bezüglich des Skalarproduktes

$$\langle p, q \rangle_\rho = \int_{-1}^{1} p(x)q(x)\mathrm{d}x$$

Die eben konstruierten orthogonalen Polynome heißen **Legendre-Polynome**.

**Bemerkung 7.6.** Bezeichnet man durch $P_k = \mathrm{span}\{p_0, \ldots, p_k\}$ en Vektorraum der Polynome bis zum Grad $k$, dann gilt allgemein für die Folge paarweise orthogonaler Polynome $p_0, \ldots, p_n$ mit aufsteigendem Grad

$$p_n \in P_{n-1}^{\perp}$$

**Beispiel.** Mit $[a,b] = [-1,1]$ und der Gewichtsfunktion $\rho(x) = (1-x^2)^{-\frac{1}{2}} = \frac{1}{\sqrt{1-x^2}}$ erhält man mit dem Gram-Schmidt-Verfahren (7.15) ausgehend von $p_0 = 1$ mit

$$p_0(x) = 1, \quad p_1(x) = x, \quad p_2(x) = x^2 - \frac{1}{2}, \quad p_3(x) = x^3 - \frac{3}{4}x \tag{7.17}$$

die orthogonalen **Tschebyscheff-Polynome**.

Sowohl bei den Legendre- als auch bei den Tschebyscheff-Polynomen findet man jeweils einfach reelle Nullstellen, die im Intervall $]a,b[$ liegen. Generell gilt der

**Satz 7.7.** *Die Nullstellen des n-ten Orthogonalpolynoms bezüglich eines Intervalls $[a,b]$ und einer Gewichtsfunktion $\rho$ sind einfach, reell und liegen im Intervall $]a,b[$*

*Beweis.* Es seien $a < \lambda_1 < \cdots < \lambda_j < b$ $(0 \le j \le n)$ die Nullstellen von $p_n$ in $]a,b[$, an denen $p_n$ sein Vorzeichen wechselt (diese Nullstellen haben eine ungerade algebraische Vielfachheit). Es wird nun $j = n$ nachgewiesen.
Für $j \le n-1$ hätte das Polynom

$$q(x) := \Pi_{k=1}^{j}(x - \lambda_k)$$

den Grad $0 \le j \le n-1$, so dass

$$\langle p_n, q \rangle_\rho = 0 \tag{7.18}$$

folgt, weil $p_n$ nach Konstruktion orthogonal zu sämtlichen Polynomen mit Grad kleiner oder gleich $n-1$ ist. Nach dem Fundamentalsatz der linearen Algebra ist $p_n$ als Produkt

$$p_n(x) = v(x)q(x)$$

darstellbar, wobei $v(x)$ auf $[a,b]$ keine Stellen enthält, wo sein Vorzeichen wechselt. Damit wäre aber

$$\langle p_n, q\rangle_\rho = \int_a^b p_n(x)q(x)\rho(x)dx = \int_a^b v(x)q^2(x)\rho(x)dx \neq 0$$

was der Annahme (7.18) (bzw. $j \leq n-1$) widerspricht. Also gilt tatsächlich $j = n$ und damit ist der Satz bewiesen. □

Nun kommen wir zur Definition der Gauß-Quadratur

**Definition 7.8.** *Mit $x_1, \ldots, x_n$ seien die Nullstellen des $n$-ten Orthogonalpolynoms $p_n(x)$ gegeben. Die numerische Integrationsformel*

$$I_n = \sum_{j=1}^{n} \sigma_j f(x_j) \quad \textit{mit} \quad \sigma_j = \langle L_j, 1\rangle_\rho = \int_a^b L_j(x)\rho(x)\mathrm{d}x \tag{7.19}$$

*heißt Gaußsche Quadraturformel der $n$-ten Ordnung oder kurz Gauß-Quadratur zur Gewichtsfunktion $\rho$*

Im Folgenden wird gezeigt, dass die Stützstellen $x_k$ und Gewichte $\sigma_k$ als Lösung des Gleichungssystems (7.11) gerade die Nullstellen des $n$-ten Orthogonalpolynoms $p_n(x)$ bzw. die Gewichte gemäß (7.19) sind und damit die Gleichwertigkeit der Formeln (7.10) und (7.19) nachgewiesen.

**Satz 7.9.** *Mit $x_1, \ldots, x_n$ seien die Nullstellen des $n$-ten Orthogonalpolynoms $p_n(x)$ gegeben.*
*Es existiert eine eindeutig bestimmte Gauß-Quadratur* (7.19)*. Bei der Gauß-Quadratur sind alle Gewichte gemäß* (7.19) *positiv und die Quadratur ist für jedes Polynom vom Grad $m \leq 2n-1$ exakt, d.h. es gilt*

$$\int_a^b p(x)\rho(x)\mathrm{d}x = \langle p, 1\rangle_\rho = \sum_{j=1}^{n} \sigma_j p(x_j), \quad \forall p \in \Pi_{2n-1} \tag{7.20}$$

*Außerdem ist die Quadratur interpolatorisch, d.h. es gilt für das Interpolationspolynom $q_{n-1}$ zu den Stützpunkten $(x_j, f(x_j)), j = 1, \ldots, n$*

$$\int_a^b q_{n-1}(x)\rho(x)\mathrm{d}x = \sum_{j=1}^{n} \sigma_j q_{n-1}(x_j) = \sum_{j=1}^{n} \sigma_j f(x_j)$$

*Beweis.*
Wir betrachten ein Polynom $p \in \Pi_{2n-1}$ mit Grad $m \leq 2n-1$. Durch Polynomdivision findet man für das $n$-te Orthogonalpolynom Polynome $q, r \in P_{n-1}$ mit

$$\frac{p}{p_n} = q + \frac{r}{p_n} \Leftrightarrow p = qp_n + r$$

Mit den Nullstellen $x_1, \ldots, x_n$ von $p_n$ gilt $p(x_j) = r(x_j)$ für $j = 1, \ldots, n$. Das Lagrangsche Interpolationspolynom für $r(x)$ ergibt

$$r(x) = \sum_{j=1}^{n} r(x_j) L_j(x) = \sum_{j=1}^{n} p(x_j) L_j(x)$$

wegen $\langle q, p_n \rangle_\rho = 0$ gilt

$$\begin{aligned} \int_a^b p(x)\rho(x)\mathrm{d}x &= \langle p, 1 \rangle_\rho = \langle r, 1 \rangle_\rho \\ &= \sum_{j=1}^{n} p(x_j) \langle L_j, 1 \rangle_\rho = \sum_{j=1}^{n} \sigma_j p(x_j) \end{aligned}$$

Für $p(x) = L_j^2(x) \in \Pi_{2n-2}$ ergibt die eben nachgewiesene Formel (7.20)

$$0 < \|L_j\|_\rho^2 = \left\langle L_j^2, 1 \right\rangle_\rho = \sum_{k=1}^{n} \sigma_k L_j^2(x_k) = \sigma_j$$

Wegen $L_j^2(x_k) = \delta_{jk}^2$ folgt die Positivität der Gewichte.
Zum Nachweis der Eindeutigkeit der Gauß-Quadratur nimmt man an, dass eine weitere Formel

$$I_n^* = \sum_{j=1}^{n} \sigma_j^* f(x_j^*) \tag{7.21}$$

existiert mit $x_k^* \neq x_j^*$ für $k \neq j$, deren Genauigkeitsgrad gleich $2n-1$ ist. Die Positivität der $\sigma_j^*$ wird analog der Positivität der $\sigma_j$ gezeigt.
Für das Hilfspolynom vom Grad $2n-1$

$$h(x) = L_k^*(x) p_n(x), \quad L_k^*(x) = \prod_{k \neq j=1}^{n} \frac{x - x_j^*}{x_k^* - x_j^*}$$

ergibt (7.21) den exakten Wert des Integrals für $h(x)$, also

$$\begin{aligned} \int_a^b h(x)\rho(x)\mathrm{d}x &= \int_a^b L_k^*(x) p_n(x) \rho(x) \mathrm{d}x \\ &= \sum_{j=1}^{n} \sigma_j^* L_k^*(x_j^*) p_n(x_j^*) = \sigma_k^* p_n(x_k^*) \end{aligned}$$

für alle $k = 1, \ldots, n$. Da das 2. Integral $\int_a^b L_k^*(x) p_n(x) \rho(x) \mathrm{d}x = \langle L_k^*, p_n \rangle_\rho$ wegen der Orthogonalität von $p_n$ zu allen Polynomen bis zum Grad $n-1$ gleich Null ist, folgt $\sigma_k^* p_n(x_k^*) = 0$ für alle $k = 1, \ldots, n$. Wegen der Positivität der Gewichte müssen die $x_k^*$ Nullstellen des $n$-ten Orthogonalpolynoms $p_n(x)$ sein, die eindeutig bestimmt sind. Damit ist die Eindeutigkeit der Gauß-Quadratur bewiesen. □

Auf der Grundlage des Fehlers der Polynominterpolation von $f(x)$ durch ein Polynom $n$-ten Grades kann man den Fehler der Gauß-Quadratur bestimmen, es gilt der

**Satz 7.10.** *Mit den Stützstellen und Gewichten aus Satz 7.9 gilt für auf dem Intervall $[a, b]$ $2n$-mal stetig diffbare Funktionen $f(x)$*

$$\int_a^b f(x)\rho(x)\mathrm{d}x - \sum_{j=1}^{n} \sigma_j f(x_j) = \frac{\|p_n\|_\rho^2}{(2n)!} f^{(2n)}(\xi) \tag{7.22}$$

*mit einem Zwischenwert $\xi \in ]a, b[$.*

Die folgende Tabelle zeigt Intervalle, Gewichtsfunktionen, die zugehörigen Orthogonalpolynome und deren Name ($\alpha, \beta > -1$)

| Intervall | $\rho(x)$ | $p_0, p_1, \ldots$ | Bezeichnung |
|---|---|---|---|
| $[-1, 1]$ | $1$ | $1, x, x^2 - \frac{1}{3}, \ldots$ | Legendre |
| $[-1, 1]$ | $\frac{1}{\sqrt{1-x^2}}$ | $1, x, x^2 - \frac{1}{2}, \ldots$ | Tschebyscheff |
| $[-1, 1]$ | $(1-x)^\alpha (1+x)^\beta$ | $1, \frac{1}{2}[\alpha - \beta + (\alpha + \beta + 2)x]$ | Jacobi |
| $]-\infty, \infty[$ | $e^{-x^2}$ | $1, x, x^2 - \frac{1}{2}, x^3 - \frac{3}{2}x, \ldots$ | Hermite |
| $[0, \infty[$ | $e^{-x} x^\alpha$ | $1, x - \alpha - 1, \ldots$ | Laguerre |

Mit den in der Tabelle angegebenen Polynomen und deren Nullstellen lassen sich Quadraturformeln für endliche Intervalle und unendliche Intervall konstruieren.
Die Tschebyscheffpolynome sind trotz der Gewichtsfunktion gegenüber den Legendrepolynomen attraktiv, weil man die Nullstellen des $n$-ten Tschebyscheffschen Orthogonalpolynoms explizit angeben kann (durch eine Berechnungsformel, s.dazu (6.17) ) ohne die Polynome auszurechnen. Das ist bei den anderen Polynomen aus der Tabelle nicht direkt möglich.

## 7.4 Numerische Integration durch Extrapolation (hier nur zur Information, wird in der Übung behandelt)

Die summierte Trapezregel zur näherungsweisen Berechnung des Integrals $\int_a^b f(x)\,dx$ kann man bei der Verwendung von $N$ Teilintervallen in der Form

$$T(h) := S_{1,N} = h[\frac{1}{2}(f(a)+f(b)) + \sum_{i=1}^{N-1} f(a+i\,h)]$$

mit $h = \frac{b-a}{N}$ aufschreiben.

Die Grundidee der numerischen Integration durch Extrapolation besteht in der Nutzung der Werte der Trapezsumme $T(h)$ für unterschiedliche Schrittweiten $h_0$, $h_1$, um durch Extrapolation auf $h = 0$ zu schließen.

Die entscheidende mathematische Grundlage hierfür ist der

**Satz 7.11.** *Für eine Funktion $f \in C^{2m+2}[a,b]$ besitzt $T(h)$ die Entwicklung*

$$T(h) = \tau_0 + \tau_1 h^2 + \tau_2 (h^2)^2 + \cdots + \tau_m (h^2)^m + R_{m+1}(h) \tag{7.23}$$

*mit*

$$\begin{aligned} \tau_0 &= \int_a^b f(x)\,dx\ , \\ \tau_k &= \frac{B_{2k}}{(2k)!}[f^{(2k-1)}(b) - f^{(2k-1)}(a)] \end{aligned}$$

*($B_{2k}$ sind die Bernoullischen Zahlen, die unabhängig von $h$ sind, und damit sind auch die $\tau_k$ unabhängig von $h$) und dem Restglied $R_{m+1}$*

$$R_{m+1}(h) = \mathcal{O}(h^{2m+2})\ .$$

*Beweis.* Für Interessenten s. Plato □

Es ist offensichtlich, dass nach dem Satz 7.11

$$\int_a^b f(x)\,dx = \tau_0 = \lim_{h \searrow 0} T(h)$$

gilt.

Schreibt man nun die asymptotische Entwicklung (7.23) z.B. für 3 Schrittweiten auf, dann erhält man

$$\begin{aligned} T(h_0) &\approx \tau_0 + \tau_1 h_0^2 + \tau_2 h_0^4 \\ T(h_1) &\approx \tau_0 + \tau_1 h_1^2 + \tau_2 h_1^4 \\ T(h_2) &\approx \tau_0 + \tau_1 h_2^2 + \tau_2 h_2^4 \end{aligned} \tag{7.24}$$

und kann bei Kenntnis der Trapezsummen $T(h_1), T(h_2)$ und $T(h_3)$ daraus $\tau_0$ näherungsweise ermitteln. Bei den Beziehungen (7.24) macht man aufgrund von Satz 7.11 nur Fehler der Ordnung $\mathcal{O}(h^6)$.

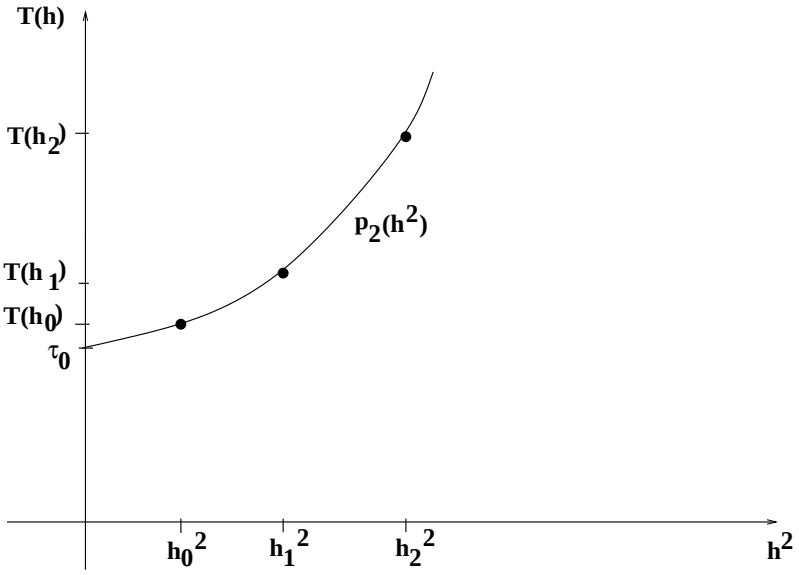



Abbildung 7.1: Polynom $p_2$ und Stützwerte $(h_k^2, T(h_k)), k = 0, 1, 2$

Eine andere Interpretation dieser Extrapolationsidee besteht in der Nutzung der Wertepaare

$$(h_0^2, T(h_0)),\ (h_1^2, T(h_1)),\ (h_2^2, T(h_2))$$

zur Bestimmung des Interpolationspolynoms zweiten Grades in $\tilde{h} = h^2$, also

$$p_2(\tilde{h}) = p_2(h^2) = \tilde{\tau}_0 + \tilde{\tau}_1 h^2 + \tilde{\tau}_2 (h^2)^2$$

mit der Eigenschaft

$$p_2(h_k^2) = T(h_k)\ ,\ k = 0, 1, 2\ .$$

Die Auswertung dieses Polynoms an der Stelle $h^2 = 0$ (Extrapolation, s.auch Abb. 7.1) liefert dann die Näherung

$$\int_a^b f(x)\, dx = \tau_0 \approx p_2(0) = \tilde{\tau}_0\ .$$

## 7.5 Anwendung des Schemas von Neville Aitken - Romberg-Verfahren (hier nur zur Information, wird in der Übung behandelt)

In Anlehnung an die Entwicklung (7.23) sucht man also ein Interpolationspolynoms $p_m$ an den Stützstellen $h_k^2$ mit den Funktionswerten $T(h_k)$ $(k = 0, \ldots, m)$ und möchte dann den Wert des Interpolationspolynoms $p_m$ an der Stelle $\tilde{h} = h^2 = 0$ ausrechnen, dann bietet sich das Schema von Neville-Aitken zur Polynomwertberechnung für das Interpolationspolynom für die Wertepaare $(x_k, f(x_k))$, $k = 0, 1, \ldots, m$, an, also

| $x_i$ | $T_{i,0} = f(x_i)$ | $T_{i,1}$ | $T_{i,2}$ | $\ldots$ | $T_{i,m-1}$ | $T_{i,m}$ |
|---|---|---|---|---|---|---|
| $x_0$ | $T_{0,0} = f(x_0)$ | | | | | |
| $x_1$ | $T_{1,0} = f(x_1)$ | $T_{1,1}$ | | | | |
| $x_2$ | $T_{2,0} = f(x_2)$ | $T_{2,1}$ | $T_{2,2}$ | | | |
| $\vdots$ | | | | | | |
| $x_m$ | $T_{m,0} = f(x_m)$ | $T_{m,1}$ | $T_{m,2}$ | $\ldots$ | $T_{m,m-1}$ | $T_{m,m}$ |

mit $m \geq i \geq k \geq 1$ und

$$\begin{aligned} T_{i,0} &= f(x_i) \\ T_{i,k}(x) &= \frac{(x - x_{i-k})T_{i,k-1}(x) - (x - x_i)T_{i-1,k-1}(x)}{x_i - x_{i-k}}\,, \quad k \geq 1, \end{aligned}$$

woraus

$$\begin{aligned} T_{i,k} &= \frac{(x - x_i)T_{i,k-1} + (x_i - x_{i-k})T_{i,k-1} - (x - x_i)T_{i-1,k-1}}{x_i - x_{i-k}} \\ &= T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{\frac{x - x_{i-k}}{x - x_i} - 1} \end{aligned}$$

folgt (das feste Argument $x$ wurde hier der Übersichtlichkeit halber weg gelassen).
Beim Romberg-Verfahren geht man von der Entwicklung (7.23) von $T(h)$ in $h^2$ aus, und d.h., man muss $x_i = h_i^2$ setzen. Für die Berechnung des Wertes von $p_m$ an der Stelle $\tilde{h} = h^2 = 0$ ergibt das obige Neville-Aitken-Schema

$$\begin{aligned} T_{i,0} &= T(h_i) \\ T_{i,k} &= T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{(\frac{h_{i-k}}{h_i})^2 - 1} \end{aligned}$$

mit $T_{m,m}$ den Näherungswert für $\tau_0 = \int_a^b f(x)\,dx$.
Für $m = 1$ erhält man mit $h_0 = b - a,\ h_1 = (b-a)/2$

$$\begin{aligned} T_{1,1} &= T_{1,0} + \frac{T_{1,0} - T_{0,0}}{(\frac{h_0}{h_1})^2 - 1} = \frac{4}{3}T_{1,0} - \frac{1}{3}T_{0,0} \\ &= \frac{b-a}{3}[f(a) + 2f(\frac{a+b}{2}) + f(b)] - \frac{b-a}{6}[f(a) + f(b)] \\ &= \frac{b-a}{6}[f(a) + 4f(\frac{a+b}{2}) + f(b)]\ , \end{aligned}$$

also die Simpsonregel. Für $h_i = \frac{b-a}{3^i},\ i = 0, 1$ erhält man mit

$$T_{1,1} = \frac{b-a}{8}[f(a) + 3f(\frac{2a+b}{3}) + 3f(\frac{a+3b}{3}) + f(b)]$$

die Newtonsche 3/8-Regel.
Als gängige Folgen $h_i,\ i = 0, \ldots$ werden die **Romberg**-Folge

$$h_0 = b - a, \quad h_1 = \frac{h_0}{2}, \quad h_2 = \frac{h_1}{2}, \quad h_3 = \frac{h_2}{2}, \ldots$$

oder die **Bulirsch**-Folge

$$h_0 = b - a, \quad h_1 = \frac{h_0}{2}, \quad h_2 = \frac{h_0}{3}, \quad h_3 = \frac{h_1}{2}, \quad h_4 = \frac{h_2}{2}, \ldots$$

verwendet. Zur Romberg-Folge ist noch anzumerken, dass man $T(h_{i+1})$ rekursiv aus $T(h_i)$ durch die Formel

$$T(h_{i+1}) = T(\frac{1}{2}h_i) = \frac{1}{2}T(h_i) + h_{i+1}[f(a+h_{i+1}) + f(a+3h_{i+1}) + \cdots + f(b-h_{i+1})]$$

bestimmen kann.

# Kapitel 8

# Numerische Lösung von Anfangswertaufgaben

17. Vorlesung 14.12.11

Anwendungen wie Flugbahnberechnungen, Schwingungsberechnungen oder die Dynamik von Räuber-Beute-Modellen führen auf Anfangswertprobleme für Systeme von gewöhnlichen Differentialgleichungen:

**Definition 8.1.** *Ein **Anfangswertproblem** für ein System von n gewöhnlichen Differentialgleichungen 1. Ordnung ist von der Form*

$$y' = f(t, y), \quad t \in [a, b] \tag{8.1}$$
$$y(a) = y_0 \tag{8.2}$$

*mit einem gegebenen endlichen Intervall $[a, b]$, einem Vektor $y_0 \in \mathbb{R}^n$ und einer Abbildung*

$$f : [a, b] \times \mathbb{R}^n \to \mathbb{R}^n \tag{8.3}$$

*wobei eine differenzierbare Abbildung $y : [a, b] \to \mathbb{R}^n$ mit den Eigenschaften* (8.1) - (8.3) *als **Lösung des Anfangswertproblems** gesucht ist.*

Aussagen zur Existenz und Eindeutigkeit der Lösung liefert

**Satz 8.2.** *Erfüllt $f$ aus* (8.3) *die Bedingung*

$$\|f(t, u) - f(t, v)\| \leq L \|u - v\|, \quad t \in [a, b],\ u, v \in \mathbb{R}^n \tag{8.4}$$

*mit einer Konstanten $L > 0$ in einer beliebigen Vektornorm $\|\cdot\|$ des $\mathbb{R}^n$, dann gelten die Aussagen*

*(a) Das AWP* (8.1),(8.2) *besitzt genau eine stetig diff'bare Lösung $y : [a, b] \to \mathbb{R}^n$ (Picard-Lindelöf)*

*(b) Für differenzierbare Funktionen $y, \hat{y} : [a,b] \to \mathbb{R}^n$ mit*

$$\begin{aligned} y' &= f(t,y), \quad t \in [a,b]; \quad y(a) = y_0 \\ \hat{y}' &= f(t,\hat{y}), \quad t \in [a,b]; \quad \hat{y}(a) = \hat{y}_0 \end{aligned}$$

*gilt die Abschätzung*

$$\|y(t) - \hat{y}(t)\| \leq e^{L(t-a)} \|y_0 - \hat{y}_0\|, \quad t \in [a,b] \tag{8.5}$$

*Beweis.* Vorlesung DGL oder Analysis □

**Bemerkung.**

(1) Mit den Aussagen des Satzes 8.2 hat man die Existenz und Eindeutigkeit der Lösung und die stetige Abhängigkeit der Lösung von den Anfangsdaten unter der Voraussetzung der Lipschitzstetigkeit von $f(t,\cdot)$ vorzuliegen.

(2) Im Folgenden sollen numerische Lösungsverfahren entwickelt werden, wobei wir ohne die Allgemeinheit einzuschränken den Fall $n = 1$ betrachten. Die besprochenen Verfahren gelten allerdings auch im allgemeinen Fall $n > 1$

**Definition 8.3.** *Unter dem Richtungsfeld der Differentialgleichung*

$$y' = f(t,y)$$

*versteht man das Vektorfeld*

$$r(t,y) = \begin{pmatrix} \frac{1}{\sqrt{1+f^2(t,y)}} \\ \frac{f(t,y)}{\sqrt{1+f^2(t,y)}} \end{pmatrix}$$

*d.h. das Vektorfeld der normierten Steigungen*

Betrachtet man um einen beliebigen Punkt $(t_0, y_0)$ der $(t,y)$- Ebene, kann man Lösungskurven $y(t)$ durch diesen Punkt annähren:

**Beispiel.**

$$y' = y^2 + t^2, \quad r(t,y) = \begin{pmatrix} \frac{1}{\sqrt{1+(y^2+t^2)^2}} \\ \frac{y^2+t^2}{\sqrt{1+(y^2+t^2)^2}} \end{pmatrix}$$

(I) $y'(t_0) = y_0^2 + t_0^2$, ($t_0 = a$ entspricht Start in Anfangspunkt $(a, y_0)$)
$t$-Achse wird durch $t_k = t_0 + hk$ äquidistant unterteilt

(II) mit dem Schritt von Punkt

$$(t_0, y_0) \quad \text{zu} \quad (t_0 + h, y_0 + hy'(t_0)) =: (t_1, y_1)$$

bzw. allgemein vom Punkt

$$(t_k, y_k) \quad \text{zu} \quad (t_k + h, y_k + hf(t_k, y_k)) =: (t_{k+1}, y_{k+1})$$

erhält man mit $h = \frac{b-a}{N}$ nach $m$ Schritten mit

$$y_0, y_1, \ldots, y_N$$

unter "günstigen" Umständen eine Approximation der Lösung $y(t)$ an den Stellen

$$a = t_0, t_1, \ldots, t_N = b$$

(III) D.h. man fährt das Richtungsfeld geeignet ab, um eine numerische Lösung $y_k, k = 0, 1, \ldots, N$ zu erhalten

## 8.1 Theorie der Einschrittverfahren

**Definition 8.4.** *Ein Einschrittverfahren zur näherungsweisen Bestimmung einer Lösung des AWP* (8.1),(8.2) *hat die Form*

$$y_{k+1} = y_k + h_k \Phi(t_k, y_k, y_{k+1}, h_k), \quad k = 0, 1, \ldots, N-1 \tag{8.6}$$

*mit einer Verfahrensfunktion*

$$\Phi : [a, b] \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$$

*und einem (noch nicht näher spezifizierten) Gitter bzw. Schrittweiten*

$$\Delta = \{a = t_0 < t_1 < \ldots < t_N \leq b\}, \; h_k := t_{k+1} - t_k, \quad k = 0, 1, \ldots, N-1 \tag{8.7}$$

**Bemerkung.** Hängt die Verfahrensfunktion *nicht* von $y_{k+1}$ ab, ist die Berechnungsvorschrift (8.6) eine explizite Formel zur Berechnung von $y_{k+1}$ und man spricht von einem expliziten Einschrittverfahren.

Zur Klassifizierung und Bewertung von numerischen Lösungsverfahren für AWP benötigen wir im Folgenden einige Begriffe ($y(t)$ bezeichnet hier die exakte Lösung).

**Definition 8.5.** *Unter dem* ***lokalen Diskretisierungsfehler*** *an der Stelle* $t_{k+1}$ *des Verfahrens* (8.6) *versteht man den Wert*

$$d_{k+1} := y(t_{k+1}) - y(t_k) - h_k \Phi(t_k, y(t_k), y(t_{k+1}), h_k) \tag{8.8}$$

**Definition 8.6.** *Unter dem globalen Diskretisierungsfehler* $g_k$ *an der Stelle* $t_k$ *versteht man den Wert*

$$g_k := y(t_k) - y_k$$

**Definition 8.7.** *Ein Einschrittverfahren* (8.6) *besitzt die Fehlerordnung* $p$*, falls für seinen lokalen Diskretisierungsfehler* $d_k$ *die Abschätzungen*

$$|d_k| \leq \text{const.} h_k^{p+1}, \quad k = 1, \ldots, N$$

$$\max_{1 \leq k \leq N} |d_k| \leq D = \text{const.} h_{\max}^{p+1} = \mathcal{O}(h_{\max}^{p+1}) \tag{8.9}$$

*mit* $h_{\max} = \max_{k=0,\ldots,N-1} t_{k+1} - t_k$ *gilt. (Statt Fehlerordnung verwendet man auch den Begriff Konsistenzordnung.) Ist* $p \geq 1$*, dann heißt das Verfahren konsistent.*

Die Bedingungen

$$\begin{aligned} |\Phi(t, u_1, u_2, h) - \Phi(t, v_1, u_2, h)| &\leq L_1 |u_1 - v_1| \\ |\Phi(t, u_1, u_2, h) - \Phi(t, u_1, v_2, h)| &\leq L_2 |u_2 - v_2| \end{aligned} \tag{8.10}$$

für $t \in [a, b], 0 < h \leq b - t, u_j, v_j \in \mathbb{R}$, mit positiven konstanten $L_1, L_2$ sind für die folgenden Konvergenzuntersuchungen von Einschrittverfahren von Bedeutung

**Satz 8.8.** *Ein Einschrittverfahren* (8.6) *zur Lösung des AWP* (8.1), (8.2) *besitze die Konsistenzordnung* $p \geq 1$ *und die Verfahrensfunktion erfülle die Bedinung* (8.10)*. Dann liegt die Konvergenzordnung* $p$ *vor, d.h. es gilt*

$$\max_{k=0,\ldots,N} |y_k - y(t_k)| \leq K h_{\max}^p$$

*Mit einer Konstanten* $K$*, die vom Intervall* $[a, b]$*, Konstanten* $C$ *aus der Abschätzung* (8.9) *und* $L_1, L_1$ *aus* (8.10) *herrührt.*

Bewiesen werden soll der Satz 8.8 für ein explizites Einschrittverfahren (Beweise von allgemeinen Einschrittverfahren in Bärwolff oder Schwarz).
Benötigt wird das

**Lemma 8.9.** *Für Zahlen $L > 0, a_k \geq 0, h_k \geq 0$ und $b \geq 0$ sei*

$$a_{k+1} \leq (1 + h_k L)a_k + h_k b, \quad k = 0, 1, \ldots, N-1$$

*erfüllt. Dann gelten die Abschätzungen*

$$a_k \leq \frac{e^{Lt_k} - 1}{L} b + e^{Lt_k} a_0 \quad \textit{mit} \quad t_k := \sum_{j=0}^{k-1} h_j \quad (k = 0, \ldots, N)$$

*Beweis.* (vollständige Induktion)
Induktionsanfang ist für $k = 0$ offensichtlich gewährleistet. Dr Schritt $k \to k+1$ ergibt sich wie folgt:

$$\begin{aligned} a_{k+1} &\leq (1 + h_k L) \left( \frac{e^{Lt_k} - 1}{L} b + e^{Lt_k} a_0 \right) + h_k b \\ &\leq \left( \frac{e^{L(t_k + h_k)} - 1 - h_k L}{L} + h_k \right) b + e^{L(t_k + h_k)} a_0 \\ &= \frac{e^{Lt_{k+1}} - 1}{L} b + e^{Lt_{k+1}} a_0 \end{aligned}$$

□

*Beweis von Satz 8.8.* Mit den Festlegungen

$$e_k = y_k - y(t_k), \quad k = 0, 1, \ldots, N$$

gilt für $k = 0, 1, \ldots, N-1$

$$\begin{aligned} y(t_{k+1}) &= y(t_k) + h_k \Phi(t_k, y(t_k), h_k) - d_{k+1} \\ y_{k+1} &= y_k + h_k \Phi(t_k, y_k, h_k) \end{aligned}$$

und damit

$$e_{k+1} = e_k + h_k(\Phi(t_k, y_k, h_k) - \Phi(t_k, y(t_k), h_k) + d_{k+1}$$

bzw.

$$\begin{aligned} |e_{k+1}| &\leq |e_k| + h_k \left| \Phi(t_k, y_k, h_k) - \Phi(t_k, y(t_k), h_k) \right| + |d_{k+1}| \\ &\leq (1 + h_k L_1) \left| e_k \right| + h_k C h_{\max}^p \end{aligned}$$

Die Abschätzung des Lemmas 8.9 liefert wegen $e_0 = 0$ die Behauptung des Satzes 8.8 □

## 8.2 Spezielle Einschrittverfahren

### 8.2.1 Euler-Verfahren

Mit der Verfahrensfunktion

$$\Phi(t, y, h_k) = f(t, y)$$

erhält man mit

$$y_{k+1} = y_k + h_k f(t_k, y_k), \quad k = 0, \ldots, N-1 \tag{8.11}$$

das Euler-Verfahren.
Für eine stetig partiell diff'bare Funktion $f : [a, b] \times \mathbb{R} \to \mathbb{R}$ besitzt das Euler-Verfahren die Konsistenzordnung $p = 1$, denn mit der Taylorentwicklung

$$y(t + h) = y(t) + y'(t)h + \frac{h^2}{2} y''(\xi), \quad \xi \in [a, b]$$

erhält man

$$d_{k+1} = y(t_{k+1}) - y(t_k) - h_k f(t_k, y(t_k)) = \frac{h_k^2}{2} y''(\xi)$$

bzw.

$$|d_{k+1}| \leq C h_k^2 \quad \text{mit} \quad C = \frac{1}{2} \max_{\xi \in [a,b]} |y''(\xi)|$$

### 8.2.2 Einschrittverfahren der Konsistenzordnung $p = 2$

Um ein explizites Einschrittverfahren der Konsistenzordnung $p = 2$ zu erhalten, machen wir den Ansatz

$$\Phi(t, y, h) = a_1 f(t, y) + a_2 f(t + b_1 h, y + b_2 h f(t, y)),\ t \in [a, b],\ h \in [0, b - t],\ y \in \mathbb{R} \tag{8.12}$$

mit noch festzulegenden Konstanten $a_j, b_j \in \mathbb{R}$. Es gilt nun der

**Satz 8.10.** *Ein Einschrittverfahren* (8.6) *mit einer Verfahrensfunktion der Form* (8.12) *ist konsistent mit der Ordnung $p = 2$, falls $f : [a, b] \times \mathbb{R} \to \mathbb{R}$ zweimal stetig partiell diff'bar ist und für die Koeffizienten*

$$a_1 + a_2 = 1, \quad a_2 b_1 = \frac{1}{2}, \quad a_2 b_2 = \frac{1}{2} \tag{8.13}$$

*gilt.*

*Beweis.* Taylorentwicklung von $\Phi(t, y(t), \cdot)$ im Punkt $h = 0$ und von der Lösung $y$ in $t$ ergeben

$$\begin{aligned}\Phi(t, y(t), h) &= \Phi(t, y(t), 0) + h\frac{\mathrm{d}\Phi}{\mathrm{d}h}(t, y(t), 0) + \mathcal{O}(h^2) \\ &= (a_1 + a_2) f(t, y(t)) \\ &\quad + h\left(a_2 b_1 \frac{\partial f}{\partial t}(t, y(t)) + a_2 b_2 f(t, y(t)) \frac{\partial f}{\partial y}(t, y(t))\right) + \mathcal{O}(h^2) \\ &= f(t, y(t)) + \frac{h}{2}\frac{\partial f}{\partial t}(t, y(t)) + \frac{h}{2} f(t, y(t)) \frac{\partial f}{\partial y}(t, y(t)) + \mathcal{O}(h^2)\end{aligned}$$

$$\begin{aligned}y(t+h) &= y(t) + hy'(t) + \frac{h^2}{2} y''(t) + \mathcal{O}(h^3) \\ &= y(t) + h\left[f(t, y(t)) + \frac{h}{2} y''(t)\right] + \mathcal{O}(h^3) \\ &= y(t) + h\left[f(t, y(t)) + \frac{h}{2}\left\{\frac{\partial f}{\partial t}(t, y(t)) + f(t, y(t)) \frac{\partial f}{\partial y}(t, y(t))\right\}\right] + \mathcal{O}(h^3) \\ &= y(t) + h\Phi(t, y(t), h) + \mathcal{O}(h^3)\end{aligned}$$

(hier wurde die Differentialgleichung und deren Ableitung benutzt) und damit folgt

$$d_{k+1} = y(t_{k+1}) - y(t_k) - h_k \Phi(t_k, y(t_k), h_k) = \mathcal{O}(h_k^3)$$

also $p = 2$ □

Mit der konkreten Wahl $a_1 = 0, a_2 = 1, b_1 = b_2 = \frac{1}{2}$ erhält man mit

$$y_{k+1} = y_k + h_k f\left(t_k + \frac{h_k}{2}, y_k + \frac{h_k}{2} f(t_k, y_k)\right), \quad k = 0, \ldots, N-1 \tag{8.14}$$

das **modifizierte Euler-Verfahren** (verbesserte Polygonzugmethode) mit der Konsistenzordnung $p = 2$
Mit der Wahl $a_1 = a_2 = \frac{1}{2}, b_1 = b_2 = 1$ erhält man mit

$$y_{k+1} = y_k + \frac{h_k}{2}\left[f(t_k, y_k) + f(t_k + h_k, y_k + h_k f(t_k, y_k))\right], \quad k = 0, \ldots, N-1 \tag{8.15}$$

das **Verfahren von Heun** mit der Konsistenzordnung $p = 2$

## 8.3 Verfahren höherer Ordnung

18. Vorlesung 2.01.12

Die bisher besprochenen Methoden (Euler, Heun) haben wir weitestgehend intuitiv ermittelt. Um systematisch Einschrittverfahren höherer Ordnung zu konstruieren, betrachten wir die zum AWP $y' = f(t, y), y(a) = y_0$ äquivalente Gleichung (nach Integration)

$$y(t) = y_0 + \int_a^t f(s, y(s))\mathrm{d}s \tag{8.16}$$

bzw. für eine Diskretisierung des Intervalls $[a, b]$

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(s, y(s))\mathrm{d}s \tag{8.17}$$

Das letzte Integral aus (8.17) approximieren wir durch eine Quadraturformel

$$\int_{t_k}^{t_{k+1}} f(s, y(s))\mathrm{d}s \tag{8.18}$$

wobei die $s_l$ zu einer Zerlegung von $[t_k, t_{k+1}]$ gehören. (8.17) und (8.18) ergeben

$$y(t_{k+1}) \approx y(t_k) + h_k \sum_{l=1}^{m} \gamma_l f(s_l, y(s_l)) \tag{8.19}$$

wobei wir die Werte $y(s_l)$ nicht kennen. Sie müssen näherungsweise aus $y(t_k)$ bestimmt werden, damit (8.19) als Integrationsverfahren benutzt werden kann.
Wählt man z.B. $m = 2$ und $\gamma_1 = \gamma_2 = \frac{1}{2}$ sowie $s_1 = t_k$ und $s_2 = t_{k+1}$, dann bedeutet (8.19)

$$y(t_{k+1}) \approx y(t_k) + \frac{h_k}{2}\left[f(t_k, y(t_k)) + f(t_{k+1}, y(t_{k+1}))\right]$$

und mit der Approximation

$$y(t_{k+1}) \approx y(t_k) + h_k f(t_k, y(t_k))$$

ergibt sich mit

$$y(t_{k+1}) \approx y(t_k) + \frac{h_k}{2}\left[f(t_k, y(t_k)) + f(t_{k+1}, y(t_k) + h_k f(t_k, y(t_k)))\right]$$

die Grundlage für das Verfahren von Heun.

Im Weiteren wollen wir mit $y_k$ die Verfahrenswerte zur Näherung der exakten Werte $y(t_k)$ bezeichnen und als Näherungen von $f(s_l, y(s_l))$

$$f(s_l, y(s_l)) \approx k_l(t_j, y_j)$$

verwenden. Mit

$$s_l = t_k + \alpha_l h_k, \quad \alpha_l = \sum_{r=1}^{l-1} \beta_{lr}$$

werden die $k_l$ rekursiv definiert:

$$\begin{aligned}
k_1(t_k, y_k) &= f(t_k, y_k) \\
k_2(t_k, y_k) &= f(t_k + \alpha_2 h_k, y_k + h_k \beta_{21} k_1(t_k, y_k)) \\
k_3(t_k, y_k) &= f(t_k + \alpha_3 h_k, y_k + h_k(\beta_{31} k_1 + \beta_{32} k_2)) \\
&\vdots \\
k_m(t_k, y_k) &= f(t_k + \alpha_m h_k, y_k + h_k(\beta_{m1} k_1 + \cdots + \beta_{mm-1} k_{m-1}))
\end{aligned} \tag{8.20}$$

Ausgehend von (8.19) und (8.20) wird durch

$$y_{k+1} = y_k + h_k(\gamma_1 k_1(t_k, y_k) + \cdots + \gamma_m k_m(t_k, y_k)) \tag{8.21}$$

ein explizites numerisches Verfahren zu Lösung des AWP $y' = f(t, y), y(a) = y_0$ definiert.

**Definition 8.11.** *Das Verfahren* (8.21) *heißt* $m$***-stufiges Runge-Kutta-Verfahren*** *mit* $k_l$ *aus* (8.20) *und die* $k_l$ *heißen Stufenwerte.*

**Bemerkung.** Wir haben oben schon festgestellt, dass im Fall $m = 2$ mit $\gamma_1 = \gamma_2 = \frac{1}{2}, \alpha_2 = 1, \beta_{21} = 1$ (8.21) gerade das Heun-Verfahren ergibt, also ein Verfahren mit der Konsistenzordnung $p = 2$. Wir werden nun Bedingungen für die freien Parameter im Verfahren (8.21) formulieren, sodass einmal ein konsistentes Verfahren ($p \geq 1$) entsteht und andererseits eine möglichst große Konsistenzordnung erhalten wird.

Aus der Verwendung der Quadraturformel

$$h_k \sum_{l=1}^{m} \gamma_l f(s_l, y(s_l)) \approx \int_{t_k}^{t_{k+1}} f(s, y(s)) \mathrm{d}s$$

folgt die sinnvolle Forderung

$$1 = \gamma_1 + \gamma_2 + \cdots + \gamma_m \tag{8.22}$$

also haben die $\gamma_l$ die Funktion von Gewichten.
Fordert man vom Verfahren (8.21), dass die Dgl $y' = 1$ ($y$ linear) exakt integriert wird, ergibt sich die Bedingung

$$\alpha_l = \beta_{l1} + \cdots + \beta_{ll-1} \tag{8.23}$$

Es ist nämlich $f(t, y) \equiv 1$ und damit $k_l \equiv 1$ für alle $l$. Ausgangspunkt war

$$k_l(t_k, y_k) \approx f(s_l, y(s_l))$$

und

$$k_l \approx f(t_k + \alpha_l h_k, y(t_k) + h_k(\beta_{l1} k_1 + \cdots + \beta_{ll-1} k_{l-1})) \; .$$

Also steht das $y$-Argument für $y(s_l) = y(t_k + \alpha_l h_k)$. Wir fordern, dass dies bei $f \equiv 1$ exakt ist, also

$$y(s_l) = y(t_k) + h_k(\beta_{l1} + \cdots + \beta_{ll-1}) \tag{8.24}$$

da alle $k_r = 1$ sind. Andererseits ist $y$ als exakte Lösung linear, d.h.

$$y(s_l) = y(t_k) + \alpha_l h_k \tag{8.25}$$

und aus dem Vergleich von (8.24),(8.25) folgt

$$\alpha_l = \beta_{l1} + \cdots + \beta_{ll-1}$$

**Definition 8.12.** *Die Tabelle mit den Koeffizienten $\alpha_l, \beta_{lr}, \gamma_r$ in der Form*

$$\begin{array}{c|ccccc} 0 & & & & & \\ \alpha_2 & \beta_{21} & & & & \\ \alpha_3 & \beta_{31} & \beta_{32} & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ \alpha_m & \beta_{m1} & \beta_{m2} & \cdots & \beta_{mm-1} & \\ \hline & \gamma_1 & \gamma_2 & \cdots & \gamma_{m-1} & \gamma_m \end{array} \tag{8.26}$$

*heißt* ***Butcher-Tabelle*** *und beschreibt das Verfahren* (8.21). *$\alpha_1$ ist hier gleich* 0*, weil explizite Verfahren betrachtet werden.*

**Satz 8.13.** *Ein explizites Runge-Kutta-Verfahren* (8.21)*, dessen Koeffizienten die Bedingungen* (8.22) *und* (8.23) *erfüllen, ist konsistent.*

*Beweis.* Es ist zu zeigen, dass der lokale Diskretisierungsfehler die Ordnung $\mathcal{O}(h_k^{p+1})$ mit $p \geq 1$ hat. Wir setzen $h_k =: h$, da $k$ jetzt fixiert ist.

$$\begin{aligned}
|d_{k+1}| &= |y(t_{k+1}) - y(t_k) - h\Phi(t_k, y(t_k), h)| \\
&= \left| y(t_{k+1}) - y(t_k) - h \sum_{r=1}^{m} \gamma_r k_r(t_k, y(t_k)) \right| \\
&\overset{(8.22)}{=} \left| y(t_{k+1}) - y(t_k) - hf(t_k, y(t_k)) - h \sum_{r=1}^{m} \gamma_r (k_r(t_k, y(t_k)) - f(t_k, y(t_k))) \right| \\
&\leq \underbrace{|y(t_{k+1}) - y(t_k) - hy'(t_k)|}_{\in \mathcal{O}(h^2)} + h \left| \sum_{r=1}^{m} \gamma_r \underbrace{(k_r(t_k, y(t_k)) - f(t_k, y(t_k)))}_{\in \mathcal{O}(h)} \right|
\end{aligned}$$

also

$$|d_{k+1}| \leq Ch^2$$

□

**Bemerkung.** Butcher hat bewiesen, wie groß die maximale Ordnung ist, welche mit einem $m$-stufigen Runge-Kutta-Verfahren erreichbar ist, was in der folgenden Tabelle notiert ist:

| $m$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | für $m \geq 9$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $p$ | 1 | 2 | 3 | 4 | 4 | 5 | 6 | 6 | 7 | $p < m - 2$ |

## 8.4 Einige konkrete Runge-Kutta-Verfahren und deren Butcher-Tabellen

(i) Euler-Verfahren

$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array} \quad m = 1, \gamma_1 = 1$$

$$y_{k+1} = y_k + h_k f(t_k, y_k), \quad p = 1$$

(ii) Modifiziertes Euler-Verfahren

$$\begin{array}{c|cc} 0 & & \\ \frac{1}{2} & \frac{1}{2} & \\ \hline & 0 & 1 \end{array} \quad m = 2, \gamma_1 = 0, \gamma_2 = 1, \alpha_2 = \frac{1}{2}, \beta_{21} = \frac{1}{2}$$

$$
\begin{aligned}
k_1 &= f(t_k, y_k) \\
k_2 &= f(t_k + \frac{1}{2}h_k, y_k + \frac{1}{2}h_k k_1) \\
y_{k+1} &= y_k + h_k k_2, \quad p = 2
\end{aligned}
$$

(iii) Verfahren von Runge von 3. Ordnung

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

$$
m = 3, \gamma_1 = \gamma_2 = 0, \gamma_3 = 1, \alpha_2 = \frac{1}{2}, \alpha_3 = 1, \beta_{21} = \frac{1}{2}, \beta_{31} = 0, \beta_{32} = 1
$$

$$
\begin{aligned}
k_1 &= f(t_k, y_k) \\
k_2 &= f(t_k + \frac{1}{2}h_k, y_k + \frac{1}{2}h_k k_1) \\
k_3 &= f(t_k + h_k, y_k + h_k k_2) \\
y_{k+1} &= y_k + h_k k_3, \quad p = 3
\end{aligned}
$$

(iv) Klassisches Runge-Kutta-Verfahren 4. Ordnung

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

$$
\begin{aligned}
k_1 &= f(t_k, y_k) \\
k_2 &= f(t_k + \frac{1}{2}h_k, y_k + \frac{1}{2}h_k k_1) \\
k_3 &= f(t_k + \frac{1}{2}h_k, y_k + \frac{1}{2}h_k k_2) \\
k_4 &= f(t_k + h_k, y_k + h_k k_3) \\
y_{k+1} &= y_k + h_k \left( \frac{1}{6}k_1 + \frac{1}{3}k_2 + \frac{1}{3}k_3 + \frac{1}{6}k_4 \right), \quad p = 4
\end{aligned}
$$

**Bemerkung.** Die Ordnung eines konkreten Runge-Kutta-Verfahrens kann mit Hilfe von Taylor-Entwicklungen ermittelt werden, wobei man dabei von einer geeigneten Glattheit von $f(t, y)$ ausgeht.

Im Folgenden soll die Ordnung eines 3-stufigen expliziten Runge-Kutta- Verfahrens bestimmt werden.

**Satz 8.14.** *Sei $f$ dreimal stetig partiell diff'bar und gelte für die Parameter*

$$\begin{aligned} \alpha_2 &= \beta_{21} \\ \alpha_3 &= \beta_{31} + \beta_{32} \\ \gamma_1 + \gamma_2 + \gamma_3 &= 1 \end{aligned}$$

*sowie*

$$\begin{aligned} \alpha_2\gamma_2 + \alpha_3\gamma_3 &= \frac{1}{2} \\ \alpha_2\gamma_3\beta_{32} &= \frac{1}{6} \\ \alpha_2^2\gamma_2 + \alpha_3^2\gamma_3 &= \frac{1}{3} \end{aligned}$$

*Dann hat das Runge-Kutta-Verfahren (explizit, 3-stufig) die Fehlerordnung $p = 3$*

*Beweis.* Grundlage für den Beweis ist die Taylor-Approximation

$$\begin{aligned} f(t + \Delta t, y + \Delta y) = f(t, y) + \begin{pmatrix} \frac{\partial f}{\partial t}(t, y) \\ \frac{\partial f}{\partial y}(t, y) \end{pmatrix} \begin{pmatrix} \Delta t \\ \Delta y \end{pmatrix} \\ + \frac{1}{2}(\Delta t, \Delta y) \begin{pmatrix} \frac{\partial^2 f}{\partial t^2}(t, y) & \frac{\partial^2 f}{\partial t \partial y}(t, y) \\ \frac{\partial^2 f}{\partial y \partial t}(t, y) & \frac{\partial^2 f}{\partial y^2}(t, y) \end{pmatrix} \begin{pmatrix} \Delta t \\ \Delta y \end{pmatrix} + \mathcal{O}(\Delta^3) \end{aligned} \tag{8.27}$$

der Funktion $f$, wobei $\frac{\partial^2 f}{\partial t \partial y} = \frac{\partial^2 f}{\partial y \partial t}$ aufgrund der Glattheit von $f$ gilt. Mit

$$\begin{aligned} \bar{k}_1 &= f(t_k, y(t_k)) \\ \bar{k}_2 &= f(t_k + \alpha_2 h, y(t_k) + \beta_{21} h \bar{k}_1) = f(t_k + \alpha_2 h, y(t_k) + \alpha_2 h \bar{k}_1) \\ \bar{k}_3 &= f(t_k + \alpha_3 h, y(t_k) + h(\beta_{31}\bar{k}_1 + \beta_{32}\bar{k}_2)) \end{aligned}$$

gilt es, den lokalen Diskretisierungsfehler

$$d_{k+1} = y(t_{k+1}) - y(t_k) - h(\gamma_1\bar{k}_1 + \gamma_2\bar{k}_2 + \gamma_3\bar{k}_3)$$

abzuschätzen, wobei schon $\alpha_2 = \beta_{21}$ verwendet wurde ($h = h_k$). Mit $\Delta t = \alpha_2 h$ und $\Delta y = \alpha_2 h f(t_k, y(t_k))$ ergibt (8.27) für $\bar{k}_2$

$$\begin{aligned}\bar{k}_2 &= f(t_k + \Delta t, y(t_k) + \Delta y)\\ &= f + \alpha_2 h f_t + \alpha_2 h f f_y + \frac{1}{2}\alpha_2^2 h^2 f_{tt} + \alpha_2^2 h^2 f f_{ty} + \frac{1}{2}\alpha_2^2 h^2 f^2 f_{yy} + \mathcal{O}(h^3)\\ &=: f + \alpha_2 h F + \frac{1}{2}\alpha_2^2 h^2 G + \mathcal{O}(h^3) \qquad (8.28)\end{aligned}$$

$f, f_t, \ldots, f_{yy}$ sind dabei die Funktions- bzw. Ableitunswerte an der Stelle $(t_k, y(t_k))$. Für $\bar{k}_3$ erhält man unter Nutzung von (8.28) und (8.27)

$$\begin{aligned}\bar{k}_3 &= f(t_k + \alpha_3 h, y(t_k) + h(\beta_{31}\bar{k}_1 + \beta_{32}\bar{k}_2)\\ &= f + \alpha_3 h f_t + h(\beta_{31}\bar{k}_1 + \beta_{32}\bar{k}_2) f_y + \frac{1}{2}\alpha_3^2 h^2 f_{tt}\\ &\quad + \alpha_3(\beta_{31}\bar{k}_1 + \beta_{32}\bar{k}_2) h^2 f_{ty} + \frac{1}{2}(\beta_{31}\bar{k}_1 + \beta_{32}\bar{k}_2)^2 h^2 f_{yy} + \mathcal{O}(h^3)\\ &= f + h(\alpha_3 f_t + [\beta_{31} + \beta_{32}] f f_y) + h^2\Big(\alpha_2\beta_{32} F f_y\\ &\quad + \frac{1}{2}\alpha_3^2 f_{tt} + \alpha_3[\beta_{31} + \beta_{32}] f f_{ty} + \frac{1}{2}(\beta_{31} + \beta_{32}) f^2 f_{yy}\Big) + \mathcal{O}(h^3)\\ &= f + \alpha_3 h F + h^2(\alpha_2\beta_{32} F f_y + \frac{1}{2}\alpha_3^2 G) + \mathcal{O}(h^3) \qquad (8.29)\end{aligned}$$

Mit (8.28) und (8.29) folgt für den lokalen Diskretisierungsfehler

$$\begin{aligned}d_{k+1} &= h(1 - \gamma_1 - \gamma_2 - \gamma_3) f + h^2\left(\frac{1}{2} - \alpha_2\gamma_2 - \alpha_3\gamma_3\right) F\\ &+ h^3\left(\left[\frac{1}{6} - \alpha_2\gamma_3\beta_{32}\right] F f_y + \left[\frac{1}{6} - \frac{1}{2}\alpha_2^2\gamma_2 - \frac{1}{2}\alpha_3^2\gamma_3\right] G\right) + \mathcal{O}(h^4)\end{aligned} \qquad (8.30)$$

Aufgrund der Voraussetzungen werden dei Klammerausdrücke gleich Null und es gilt

$$d_{k+1} = \mathcal{O}(h^4)$$

also hat das Verfahren die Fehlerordnung $p = 3$ □

**Korollar.** *Mit Lösungen des Gleichungssystems*

$$\begin{aligned}\gamma_1 + \gamma_2 + \gamma_3 &= 1\\ \alpha_2\gamma_2 + \alpha_3\gamma_3 &= \frac{1}{2}\\ \alpha_2\gamma_3\beta_{32} &= \frac{1}{6}\\ \alpha_2^2\gamma_2 + \alpha_3^2\gamma_3 &= \frac{1}{3}\end{aligned} \qquad (8.31)$$

*hat das dazugehörrige 3-stufige Runge-Kutta-Verfahren die Fehlerordnung* $p = 3$*, wobei* $\alpha_2 = \beta_{21}$ *ist.* (8.31) *hat z.B. mit den Einschränkungen* $\alpha_2 \neq \alpha_3$ *und* $\alpha_2 \neq \frac{2}{3}$ *die Lösungen*

$$\gamma_2 = \frac{3\alpha_3 - 2}{6\alpha_2(\alpha_3 - \alpha_2)}, \qquad \gamma_3 = \frac{2 - 3\alpha_2}{6\alpha_3(\alpha_3 - \alpha_2)} \tag{8.32}$$
$$\gamma_1 = \frac{6\alpha_2\alpha_3 + 2 - 3(\alpha_2 + \alpha_3)}{6\alpha_2\alpha_3}, \qquad \beta_{32} = \frac{\alpha_3(\alpha_3 - \alpha_2)}{\alpha_2(2 - 3\alpha_2)}$$

*für* $\alpha_2, \alpha_3 \in \mathbb{R}$*, also die zweiparametrige Lösungsmenge*

$$\mathcal{M} = \{(\gamma_1, \gamma_2, \gamma_3, \alpha_2, \alpha_3, \beta_{32}) | \gamma_1, \gamma_2, \gamma_3, \beta_{32} \text{ gemäß } (8.32), \\ \alpha_2, \alpha_3 \in \mathbb{R}, \alpha_2 \neq \alpha_3, \alpha_2 \neq \frac{2}{3}\}$$

*Die restlichen Parameter des Verfahrens ergeben sich aus*

$$\beta_{21} = \alpha_2, \quad \beta_{31} = \alpha_3 - \beta_{32}$$

## 8.5 Schrittweitensteuerung bei Einschrittverfahren

19. Vorlesung 4.01.2012

Bei der Konvergenzuntersuchung von Einschrittverfahren werden die lokalen Diskretisierungsfehler in gewissem Sinn summiert und deshalb erscheint eine Beschränkung des Absolutbetrages von $d_k$ durch die Wahl geeigneter Schrittweiten $h_k$ sinnvoll. Man spricht hier von **Schrittweitensteuerung**. Das Prinzip soll am Beispiel des Heun-Verfahrens

$$k_1 = f(t_k, y_k)$$
$$k_2 = f(t_k + h, y_k + hk_1)$$
$$y_{k+1} = y_k + \frac{1}{2}h[k_1 + k_2]$$

erläutert werden. Als lokaler Diskretisierungsfehler ergibt sich

$$d_{k+1}^{(H)} = y(t_{k+1}) - y(t_k) - \frac{1}{2}h[\bar{k}_1 + \bar{k}_2] \tag{8.33}$$

mit $\bar{k}_1 = f(t_k, y(t_k)), \bar{k}_2 = f(t_k + h, y(t_k) + h\bar{k}_1)$
Nun sucht man ein Verfahren höherer Ordnung, also mindestens dritter Ordnung, dessen Steigungen $k_1, k_2$ mit den Steigungen des Heun-Verfahrens übereinstimmen.

Die Forderung der Gleichheit von $k_1$ und $k_2$ bedeutet $\alpha_2 = \beta_{21} = 1$. Die weiteren Parameter ergeben sich aus (8.32) bei der Wahl von $\alpha_3 = \frac{1}{2}$ zu

$$\gamma_3 = \frac{2}{3}, \quad \gamma_2 = \frac{1}{6}, \quad \gamma_1 = \frac{1}{6}, \quad \beta_{32} = \frac{1}{4}, \quad \beta_{31} = \alpha_3 - \beta_{32} = \frac{1}{4}$$

sodass sich das Runge-Kutta-Verfahren 3. Ordnung

$$\begin{aligned} k_1 &= f(t_k, y_k) \\ k_2 &= f(t_k + h, y_k + hk_1) \\ k_3 &= f(t_k + \frac{h}{2}, y_k + \frac{h}{4}(k_1 + k_2)) \\ y_{k+1} &= y_k + \frac{h}{6}[k_1 + k_2 + 4k_3] \end{aligned} \tag{8.34}$$

ergibt. Für den lokalen Diskretisierungsfehler des Verfahrens (8.33) ergibt sich

$$d_{k+1}^{(RK)} = y(t_{k+1}) - y(t_k) - \frac{h}{6}[\bar{k}_1 + \bar{k}_2 + 4\bar{k}_3] \tag{8.35}$$

mit $\bar{k}_3 = f(t_k + \frac{h}{2}, y(t_k) + \frac{h}{4}(\bar{k}_1 + \bar{k}_2))$. Mit (8.33) und (8.35) ergibt sich die Darstellung des lokalen Diskretisierungsfehlers des Heun-Verfahrens

$$d_{k+1}^{(H)} = \frac{h}{6}[\bar{k}_1 + \bar{k}_2 + 4\bar{k}_3] - \frac{h}{2}[\bar{k}_1 + \bar{k}_2] + d_{k+1}^{(RK)}$$

Ersetzt man nun die unbekannten Werte von $\bar{k}_j$ durch die Näherungen $k_j$ und berücksichtigt $d_{k+1}^{(RK)} = \mathcal{O}(h^4)$, so erhält man

$$d_{k+1}^{(H)} = \frac{h}{6}[k_1 + k_2 + 4k_3] - \frac{h}{2}[k_1 + k_2] + \mathcal{O}(h^4) = \frac{h}{3}[2k_3 - k_1 - k_2] + \mathcal{O}(h^4)$$

und damit kann der lokale Diskretisierungsfehler des Heun-Verfahrens mit einer zusätzlichen Steigungsberechnung von $k_3$ durch den Ausdruck $\frac{h}{3}[2k_3 - k_1 - k_2]$ recht gut geschätzt werden.
Aufgrund der Kontrolle des Betrags dieses Ausdrucks kann man eine vorgegebene Schranke $\epsilon_{\text{tol}} > 0$ durch entsprechende Wahl von $h = h_k = t_{k+1} - t_k$

$$h_k < \frac{3\epsilon_{\text{tol}}}{|2k_3 - k_1 - k_2|} \Leftrightarrow \frac{h_k}{3}[2k_3 - k_1 - k_2] < \epsilon_{\text{tol}}$$

unterschreiten. D.h. man kann die aktuelle Schrittweite evtl. vergrößern oder muss sie verkleinern.
Die eben beschriebene Methode der Schrittweitensteuerung bezeichnet man auch als Einbettung des Heun-Verfahrens 2. Ordnung in das Runge-Kutta-Verfahren 3. Ordnung (8.34).

## 8.6 Implizite Runge-Kutta-Verfahren

Explizite Verfahren neigen zur Instabilität und damit besteht die Gefahr der Verstärkung von Rundungsfehlern.
Implizite Verfahren erweisen sich als stabil, speziell, wenn es sich um die Lösung von AWPs mit sogenannten steifen DGL handelt.
Im Unterschied zum Gleichungssystem (8.20) wird beim impliziten Runge-Kutta-Verfahren das Gleichungssystem

$$k_r(t_k, y_k) = f(t_k + \alpha_r h_k, y_k + h_k(\beta_{r1}k_1 + \cdots + \beta_{rm}k_m)), \quad r = 1, \ldots, m \tag{8.36}$$

zur Bestimmung der $k_r$ zugrunde gelegt.
Mit (8.36) wird (8.21) zu einem impliziten Runge-Kutta- Verfahren. Aus (8.36) ergibt sich die Butcher-Tabelle

$$\begin{array}{c|ccc} \alpha_1 & \beta_{11} & \ldots & \beta_{1m} \\ \alpha_2 & \beta_{21} & \ldots & \beta_{2m} \\ \vdots & \vdots & & \vdots \\ \alpha_m & \beta_{m1} & \ldots & \beta_{mm} \\ \hline & \gamma_1 & \ldots & \gamma_m \end{array} \tag{8.37}$$

Die Überlegungen, die bei den expliziten Verfahren die Bedingung (8.23) für die Koeffizienten $\alpha_r, \beta_{rl}$ gerechtfertigt haben, ergeben analog bei den impliziten Runge-Kutta- Verfahren die Bedingung

$$\alpha_r = \beta_{r1} + \beta_{r2} + \cdots + \beta_{rm}, \quad r = 1, \ldots, m \tag{8.38}$$

Zur Lösbarkeit des Gleichungssystems (8.36) gilt der

**Satz 8.15.** *$f$ genüge auf $[a, b] \times \mathbb{R}$ der Lipschitz-Bedingung*

$$|f(t, y_1) - f(t, y_2)| \leq L\,|y_1 - y_2|$$

*und die Schrittweite $h = h_k$ genüge der Bedingung*

$$q = hL \max_{1 \leq j \leq m} \left( \sum_{r=1}^{m} |\beta_{jr}| \right) < 1$$

*Dann hat* (8.36) *zur Bestimmung von $k_1, \ldots, k_m$ genau eine Lösung*

*Beweis.* Aussage folgt aus dem Banachschen Fixpunktsatz (wird am Ende des Semesters behandelt). □

## 8.7 Rundungsfehleranalyse von expliziten Einschrittverfahren

Zur numerischen Lösung eines AWP betrachten wir das Verfahren

$$y_{k+1} = y_k + h_k \Phi(t_k, y_k, h_k), \quad k = 0, 1, 2 \ldots, N-1 \tag{8.39}$$

mit der Verfahrensfunktion $\Phi$. Durch Rundungsfehler arbeitet man statt (8.39) mit einem Verfahren der Form

$$y_{k+1} = y_k + h_k \Phi(t_k, y_k, h_k) + \rho_k, \quad k = 0, 1, \ldots, N-1 \tag{8.40}$$

$$y_0 = y_0 + e_0, \quad |\rho_k| \leq \delta, \quad k = 0, 1, \ldots, N-1, \ |e_0| \leq \epsilon$$

mit gewissen Zahlen $e_0, \rho_k \in \mathbb{R}$
Für die Rundungsfehler infolge des Verfahrens (8.40) gilt der folgende

**Satz 8.16.** *Zur Lösung des AWP $y' = f(t, y), y(a) = y_0$, sei durch* (8.39) *ein Einschrittverfahren mit der Konsistenzordnung $p \geq 1$ gegeben, wobei die Verfahrensfunktion bezüglich der 2. Variablen Lipschitz-stetig mit der Konstanten $L > 0$ ist.*
*Dann gelten für die durch die fehlerbehaftete Verfahrensvorschrift* (8.40) *gewonnenen Approximationen die Abschätzungen*

$$\max_{k=0,\ldots,N} |y_k - y(t_k)| \leq K(h_{\max}^p + \frac{\delta}{h_{\min}}) + e^{L(b-a)}\epsilon \tag{8.41}$$

*mit der Konstanten $K = \frac{\max\{C,1\}}{L}\left[e^{L(b-a)} - 1\right]$. $C$ ist dabei die Konstante aus der Abschätzung $|d_k| \leq Ch_k^{p+1}$ für den lokalen Diskretisierungsfehler.*

## 8.8 Ein Anwendungsgebiet für Löser von AWPs

Eine wichtige Anwendung der numerischen Lösungsverfahren für Anfangswertprobleme ist die Lösung von Zweipunkt-Randwertproblemen mit Schießverfahren.
Schießverfahren zur Lösung von Zweipunkt-Randwertproblemen basieren auf Methoden zur Lösung von Anfangswertproblemen. Beim sogenannten **ersten Randwertproblem**

$$y'' = f(x, y)\ , \quad y(a) = \eta_a\ , \quad y(b) = \eta_b \tag{8.42}$$

nutzt man dabei z.B. die Randbedingung $y(a) = \eta_a$ als Anfangsbedingung und versucht durch eine geeignete Wahl von $\zeta_a = y'(a)$ als Anfangsbedingung für die Ableitung mit einer Lösung des Anfangswertproblems

$$y'' = f(x, y) \ , \quad y(a) = \eta_a \ , \quad y'(a) = \zeta \tag{8.43}$$

die Randbedingung $y(b) = \eta_b$ zu treffen. Für vorgegebenes $\zeta$ sei $y(x, \zeta)$ die Lösung von (8.43). $y(x, \zeta)$ ist dann Lösung des Zweipunkt-Randwertproblems (8.42), wenn $\zeta$ Nullstelle der Funktion

$$g(\zeta) = y(b, \zeta) - \eta_b \tag{8.44}$$

ist. Für eine Funktionswertberechnung von $g$ ist ein Anfangswertproblem (8.42) zu lösen. Eine Möglichkeit zur Bestimmung der Nullstelle von $g$ ist mit dem Bisektionsverfahren gegeben. Allerdings ist es durchaus möglich, dass durch Fehler bei der Lösung des Anfangswertproblems das Vorzeichen von $g$ nicht immer korrekt berechnet werden kann, so dass das Bisektionsverfahren unbrauchbar wird.
Eine andere Möglichkeit zur Bestimmung der Nullstelle von $g$ bietet das Newton-Verfahren. Die Differentiation von $g$ nach $\zeta$ ergibt

$$g'(\zeta) = y_\zeta(b, \zeta) \ , \tag{8.45}$$

wobei $y_\zeta(b, \zeta)$ die partielle Ableitung von $y(x, \zeta)$ nach $\zeta$ ausgewertet an der Stelle $x = b$ ist. Die Differentiation der Gleichung $y''(x, \zeta) = f(x, y(x, \zeta))$ nach $\zeta$ ergibt

$$\frac{\partial}{\partial \zeta}[y''(x, \zeta)] = f_y(x, y(x, \zeta)) y_\zeta(x, \zeta) \ . \tag{8.46}$$

$f_y$ bedeutet dabei die partielle Ableitung von $f(x, y)$ nach $y$. Mit der Voraussetzung der Vertauschbarkeit der Ableitungen nach $\zeta$ und $x$ erhält man aus (8.46) die Differentialgleichung 2. Ordnung

$$y''_\zeta(x, \zeta) = f_y(x, y(x, \zeta)) y_\zeta(x, \zeta) \tag{8.47}$$

für $y_\zeta(x, \zeta)$. Durch Differentiation der Anfangsbedingungen der Aufgabe (8.43) nach $\zeta$ erhält man die Anfangsbedingungen

$$y_\zeta(a, \zeta) = 0 \ , \quad y'_\zeta(a, \zeta) = 1 \ . \tag{8.48}$$

Mit (8.47), (8.48) liegt ein Anfangswertproblem zur Berechnung von $y_\zeta(x, \zeta)$, also auch zur Berechnung der Ableitung von $g$ vor (gemäß (8.45)). Damit kann man durch Lösung der Anfangswertprobleme (8.43) und (8.47), (8.48) Funktionswert und Ableitung von $g(\zeta)$ berechnen und kann somit ein

Newton-Verfahren zur Nullstellenberechnung von $g$ durchführen. Hierzu ist anzumerken, dass man zur Lösung von (8.47), (8.48) die Funktion $y(x, \zeta)$ als Lösung des Anfangswertproblems (8.43) benötigt, um die Funktionswerte von $f_y(x, y(x, \zeta))$ berechnen zu können. Da man die exakte Lösung $y(x, \zeta)$ nicht zur Verfügung hat, verwendet man die Näherungswerte $y_k$ an den Stützstellen $x_k$ des Intervalls $[a, b]$ zur Berechnung von $f_y$ an den Stützstellen $x_k$. Beim Schießverfahren ist es in jedem Fall sinnvoll, ein recht genaues Verfahren zur erforderlichen Lösung der Anfangswertprobleme (8.43) und (8.47), (8.48) zu verwenden, da speziell bei wachsenden Lösungen die Sensibilität der Lösung $y(x, \zeta)$ von $\zeta$ sehr groß sein kann und somit kleine Änderungen von $\zeta$ große Auswirkungen auf $y(b, \zeta)$ haben können. Schießverfahren kann man bei nichtlinearen Problemen anwenden, da bei den benötigten Integrationsverfahren für gewöhnliche Differentialgleichungen die Linearität der Gleichungen nicht notwendig ist.

## 8.9 Mehrschrittverfahren

20. Vorlesung 09.01.2012

Die Klasse der Mehrschrittverfahren zur Lösung von AWP ist dadurch gekennzeichnet, dass man zur Berechnung des Näherungswertes $y_{k+1}$ nicht nur den Wert $y_k$ verwendet, sondern auch weiter zurückliegende Werte, z.B. $y_{k-1}, y_{k-2}$.
Als Ausgangspunkt zur Konstruktion von Mehrschrittverfahren betrachten wir die zum AWP äquivalente Integralbeziehung

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(s, y(s))\mathrm{d}s \tag{8.49}$$

Kennt man die Werte $f_k = f(t_k, y_k), \ldots, f_{k-3} = f(t_{k-1}, y_{k-3})$, dann kann man das Integral auf der rechten Seite von (8.49) i.d.R. besser approximieren als bei den Einschrittverfahren unter ausschließlicher Nutzung des Wertes $f_k$. Für das Interpolationspolynom durch die Stützpunkte $(t_j, f_j)_{j=k-3,\ldots,k}$ ergibt sich

$$p_3(t) = \sum_{j=0}^{3} f_{k-j} L_{k-j}$$

mit den Lagrangschen Basispolynomen

$$L_j(t) = \prod_{j \neq i = k-3}^{k} \frac{t - t_i}{t_j - t_i}, \quad j = k-3, \ldots, k$$

Die Idee der Mehrschrittverfahren besteht nun in der Nutzung von $p_3(t)$ als Approximation von $f(t, y(t))$ im Integral von (8.49), sodass man auf der Grundlage von (8.49) das Mehrschrittverfahren (4-Schritt-Verfahren)

$$\begin{aligned} y_{k+1} &= y_k + \int_{t_k}^{t_{k+1}} \sum_{j=0}^{3} f_{k-j} L_{k-j}(t)\mathrm{d}t \\ &= y_k + \sum_{j=0}^{3} f_{k-j} \int_{t_k}^{t_{k+1}} L_{k-j}(t)\mathrm{d}t \end{aligned}$$

erhält. Im Fall äquidistanter Stützstellen $h = t_{k+1} - t_k$ erhält man für das Integral des 2. Summanden ($j = 1$)

$$I_1 = \int_{t_k}^{t_{k-1}} L_{k-1}(t)\mathrm{d}t = \int_{t_k}^{t_{k+1}} \frac{(t - t_{k-3})(t - t_{k-2})(t - t_k)}{(t_{k-1} - t_{k-3})(t_{k-1} - t_{k-2})(t_{k-1} - t_k)}$$

und nach der Substitution $\xi = \frac{t - t_k}{h}$

$$I_1 = h \int_0^1 \frac{(\xi + 3)(\xi + 2)\xi}{2 \cdot 1 \cdot (-1)}\mathrm{d}\xi = -\frac{h}{2} \int_0^1 (\xi^3 + 5\xi^2 + 6\xi)\mathrm{d}\xi = -\frac{59}{24}h$$

Für die restlichen Integrale erhält man

$$I_0 = \frac{55}{24}h, \quad I_2 = \frac{37}{24}h, \quad I_3 = -\frac{9}{24}h$$

sodass sich mit

$$y_{k+1} = y_k + \frac{h}{24}[55f_k - 59f_{k-1} + 37f_{k-2} - 9f_{k-3}] \tag{8.50}$$

ein explizites Verfahren, bei dem 4 Schritte Verwendung finden, das als **Methode von Adams-Bashforths** (kurz AB-Verfahren) bezeichnet wird.
Durch Taylor-Reihenentwicklung erhält man bei entsprechender Glattheit von $f$ bzw. $y(t)$ den lokalen Diskretisierungsfehler

$$d_{k+1} = \frac{251}{720}h^5 y^{(5)} + \mathcal{O}(h^6) \tag{8.51}$$

d.h. das Verfahren (8.50) ist von 4. Ordnung.

**Definition 8.17.** *Bei Verwendung von $m$ Stützwerten*

$$(t_k, f_k), \ldots, (t_{k+1-m}, f_{k+1-m})$$

*zur Berechnung eines Interpolationspolynoms $p_{m-1}$ zur Approximation von $f$ zwecks näherungsweiser Berechnung des Integrals in* (8.49) *spricht man von einem linearen $m$-Schrittverfahren.*
*Ein $m$-Schrittverfahren hat die Fehlerordngung $p$, falls für seinen lokalen Diskretisierungsfehler $d_k$ die Abschätzung*

$$\max_{m \le k \le N} |d_k| \le K = \mathcal{O}(h^{p+1})$$

*gilt.*

Allgemein kann man für AB-Verfahren ($m$-Schritt)

$$y_{k+1} = y_k + \sum_{j=0}^{m-1} f_{k-j} \int_{t_k}^{t_{k+1}} L_{k-j}(t)\mathrm{d}t$$

bei ausreichender Glattheit der Lösung $y(t)$ zeigen, dass sie die Fehlerordnung $m$ besitzen. Durch Auswertung der entsprechenden Integrale erhält man für $m = 2, 3, 4$ die folgenden 3-,4- und 5- Schritt AB-Verfahren.

$$y_{k+1} = y_k + \frac{h}{12}[23f_k - 16f_{k-1} + 5f_{k-2}] \tag{8.52}$$
$$y_{k+1} = y_k + \frac{h}{24}[55f_k - 59f_{k-1} + 37f_{k-2} - 9f_{k-3}]$$
$$y_{k+1} = y_k + \frac{h}{720}[1901f_k - 2774f_{k-1} + 2616f_{k-2} - 1274f_{k-3} + 251f_{k-4}]$$

Die Formeln der Mehrschrittverfahren "funktionieren" erst ab dem Index $k = m$, d.h. bei einem 3-Schrittverfahren braucht man die Werte $y_0, y_1, y_2$ um $y_3$ mit der Formel (8.52) berechnen zu können.
Die Startwerte $y_1, y_2$ werden meist mit einem Runge-Kutta-Verfahren berechnet.
Es ist offensichtlich möglich die Qualität der Lösungsverfahren für das AWP zu erhöhen, indem man das Integral

$$\int_{t_k}^{t_{k+1}} f(t, y(t))\mathrm{d}t$$

aus der Beziehung (8.49) genauer berechnet. Das kann man durch hinzunahme weiterer Stützpunkte zur Polynominterpolation tun. Nimmt man den (noch unbekannten) Wert $f_{k+1} = f(t_{k+1}, y_{k+1})$ zu den Werten $f_k, \ldots, f_{k-3}$ hinzu, dann erhält man mit

$$p_4(t) = \sum_{j=-1}^{3} f_{k-j} L_{k-j}(t)$$

in Analogie zur Herleitung der AB-Verfahren mit

$$y_{k+1} = y_k + \int_{t_k}^{t_{k+1}} \sum_{j=-1}^{3} f_{k-j} L_{k-j}(t)\mathrm{d}t = y_k + \sum_{j=-1}^{3} f_{k-j} \int_{t_k}^{t_{k+1}} L_{k-j}(t)\mathrm{d}t$$

bzw. nach Auswertung der Integrale

$$y_{k+1} = y_k + \frac{h}{720}[251 f_{k+1} + 646 f_k - 264 f_{k-1} + 106 f_{k-2} - 19 f_{k-3}] \qquad (8.53)$$

Das Verfahren (8.53) ist eine implizite 4-Schritt-Methode und heißt Methode von Adams-Moulton (kurz AM-Verfahren). Das 3-Schritt AM-Verfahren hat die Form

$$y_{k+1} = y_k + \frac{h}{24}[9 f(t_{k+1}, y_{k+1}) + 19 f_k - 5 f_{k-1} + f_{k-2}] \qquad (8.54)$$

Zur Bestimmung der Lösung von (8.54) kann man z.B. eine Fixpunktiteration der Art

$$y_{k+1}^{(j+1)} = y_k + \frac{h}{24}\left[9 f(t_{k+1}, y_{k+1}^{(j)}) + 19 f_k - 5 f_{k-1} + f_{k-2}\right]$$

durchführen (Startwert z.B. $y_{k+1}^{(0)} = y_k$).

Bestimmt man den Startwert $y_{k+1}^{(0)}$ als Resultat eines 3-Schritt AB-Verfahrens und führt nur eine Fixpunktiteration durch, dann erhält man das sogenannte Prädiktor-Korrektor-Verfahren

$$y_{k+1}^{(p)} = y_k + \frac{h}{12}[23f_k - 16f_{k-1} + 5f_{k-2}] \tag{8.55}$$

$$y_{k+1} = y_k + \frac{h}{24}[9f(t_{k+1}, y_{k+1}^{(p)}) + 19f_k - 5f_{k-1} + f_{k-2}] \tag{8.56}$$

Diese Kombination von AB- und AM-Verfahren bezeichnet man als Adams-Bashforth-Moulton-Verfahren (kurz ABM-Verfahren). Das ABM-Verfahren (8.55) hat ebenso wie das AM-Verfahren (8.54) den Diskretisierungsfehler $d_{k+1} \in \mathcal{O}(h^5)$ und damit die Fehlerordnung 4.
Generell kann man zeigen, dass $m$-Schritt-Verfahren von AM- oder ABM-Typ jeweils die Fehlerordnung $p = m + 1$ besitzen.

## 8.10 Allgemeine lineare Mehrschrittverfahren

**Definition 8.18.** *Unter einem linearen $m$-Schrittverfahren ($m > 1$) versteht man eine Vorschrift mit $s = k + 1 - m$*

$$\sum_{j=0}^{m} a_j y_{s+j} = h \sum_{j=0}^{m} b_j f(t_{s+j}, y_{s+j}) \tag{8.57}$$

*wobei $a_m \neq 0$ ist und $a_j, b_j$ geeignet zu wählende reelle Zahlen sind. Die konkrete Wahl dieser Koeffizienten entscheidet über die Ordnung des Verfahrens.*

**Bemerkung.** In den bisher behandelten Verfahren war jeweils $a_m = 1$ und $a_{m-1} = -1$ sowie $a_{m-2} = \cdots = a_0 = 0$. Bei expliziten Verfahren ist $b_m = 0$ und bei impliziten Verfahren ist $b_m \neq 0$

OBdA setzen wir $a_m = 1$. Die anderen freien Parameter $a_j, b_j$ sind so zu wählen, dass die linke und die rechte Seite von (8.57) Approximationen von

$$\alpha[y(t_{k+1}) - y(t_k)] \quad \text{bzw.} \quad \alpha \int_{t_k}^{t_{k+1}} f(t, y(t))\mathrm{d}t$$

sind ($\alpha \neq 0$)
Mit der Einführung der Parameter $a_0, \ldots, a_m$ hat man die Möglichkeit durch die Nutzung der Werte $y_{k+1-m}, \ldots, y_{k+1}$ nicht nur die Approximation von $f$, sondern auch die Approximation von $y'$ mit einer höheren Ordnung durchzuführen.

**Beispiel.** Das 3-Schritt-AB-Verfahren

$$y_{k+1} = y_k + \frac{h}{12}[23f_k - 16f_{k-1} + 5f_{k-2}]$$

ist gleichbedeutend mit

$$\frac{y_{k+1} - y_k}{h} = \frac{1}{12}[23f_k - 16f_{k-1} + 5f_{k-2}]$$

wobei die rechte Seite eine Approximation von $f(t_k, y(t_k))$ der Ordnung $\mathcal{O}(h^3)$ darstellt, während die linke Seite $y'$ an der Stelle $t_k$ nur mit der Ordnung $\mathcal{O}(h)$ approximiert. Das Verfahren ist in der vorliegenden Form 3. Ordnung.
Nutzt man neben $y_{k+1}$ und $y_k$ auch noch $y_{k-1}, y_{k-2}$, dann kann man unter Nutzung der Taylor-Approximationen

$$\begin{aligned}
y(t_{k-2}) &= y(t_k) - 2hy' + 2h^2y'' - \frac{3}{2}h^3y''' + \mathcal{O}(h^4) \\
y(t_{k-1}) &= y(t_k) - hy' + \frac{1}{2}h^2y'' - \frac{1}{6}h^3y''' + \mathcal{O}(h^4) \\
y(t_{k+1}) &= y(t_k) + hy' + \frac{1}{2}h^2y'' - \frac{1}{6}h^3y''' + \mathcal{O}(h^4)
\end{aligned}$$

mit

$$\frac{1}{14h}[5y_{k+1} + 6y_k - 13y_{k-1} + 2y_{k-2}]$$

die linke Seite $y'$ ebenfalls mit der Ordnung $\mathcal{O}(h^3)$ approximieren.
Allerdings ist das daraus resultierende Mehrschrittverfahren

$$y_{k+1} + \frac{6}{5}y_k - \frac{13}{5}y_{k-1} + \frac{2}{5}y_{k-2} = h[\frac{161}{30}f_k - \frac{56}{15}f_{k-1} + \frac{7}{6}f_{k-2}]$$

wie sie später nachrechnen können nur von erster Ordnung.
Die betriebene Mehraufwand ist allerdings in manchen Fällen mit Blick auf einen evtln. Stabilitätszuwachs trotz Ordnungsverlust sinnvoll. In diesem Fall bringt der Mehraufwand nichts, weil man zwar noch ein konsistentes Verfahren hat, das allerdings nicht nullstabil ist (eine Nullstelle des ersten charakteristischen Polynoms ist betragsmäßig größer als Null!).

**Definition 8.19.** *Das lineare Mehrschrittverfahren* (8.57) *hat die Fehlerordnung p, falls in der Entwicklung des lokalen Diskretisierungsfehlers $d_{k+1}$ in eine Potenzreihe von h für eine beliebige Stelle $\tilde{t} \in [t_{k+1-m}, t_{k+1}]$*

$$\begin{aligned}
d_{k+1} &= \sum_{j=0}^{m}[a_j y(t_{s+j}) - hb_j f(t_{s+j}, y(t_{s+j}))] \qquad (8.58) \\
&= c_0 y(\tilde{t}) + c_y hy'(\tilde{t}) + \ldots + c_p h^p y^{(p)}(\tilde{t}) + \ldots
\end{aligned}$$

$c_0 = \ldots = c_p = 0$ *und* $c_{p+1} \neq 0$ *gilt* ($s = k+1-m$). *Ein Mehrschrittverfahren heißt konsistent, wenn es mindestens die Ordnung* $p = 1$ *besitzt.*

Durch eine günstige Wahl von $\tilde{t}$ kann man die Entwicklungskoeffizienten $c_j$ oft in einfacher Form als Linearkombination von $a_j, b_j$ darstellen und erhält mit der Bedingung $c_0 = \ldots = c_p = 0$ Bestimmungsgleichungen für die Koeffizienten des Mehrschrittverfahrens.
Mit der Wahl von $\tilde{t} = t_s$ ergeben sich für $y, y'$ die Taylor-Reihen

$$\begin{aligned} y(t_{s+j}) = y(\tilde{t} + jh) &= \sum_{r=0}^{q} \frac{(jh)^r}{r!} y^{(r)}(\tilde{t}) + R_{q+1} \\ y'(t_{s+j}) = y'(\tilde{t} + jh) &= \sum_{r=0}^{q-1} \frac{(jh)^r}{r!} y^{(r+1)}(\tilde{t}) + R_q \end{aligned} \tag{8.59}$$

Die Substitution der Reihen (8.59) in (8.58) ergibt für die Koeffizienten $c_j$ durch Koeffizientenvergleich

$$\begin{aligned} c_0 &= a_0 + a_1 + \ldots + a_m \\ c_1 &= a_1 + 2a_2 + \ldots + ma_m - (b_0 + b_1 + \ldots + b_m) \\ c_2 &= \frac{1}{2!}(a_1 + 2^2 a_2 + \ldots + m^2 a_m) - \frac{1}{1!}(b_1 + 2b_2 + \ldots + mb_m) \\ &\vdots \\ c_r &= \frac{1}{r!}(a_1 + 2^r a_2 + \ldots + m^r a_m) - \frac{1}{(r-1)!}(b_1 + 2^{r-1} b_2 + \ldots + m^{r-1} b_m) \end{aligned} \tag{8.60}$$

für $r = 2, 3, \ldots, q$

**Beispiel.** Es soll ein explizites 2-Schritt-Verfahren (d.h. $b_2 = 0$)

$$a_0 y_{k-1} + a_1 y_k + a_2 y_{k+1} = h[b_0 f_{k-1} + b_1 f_k] \tag{8.61}$$

der Ordnung 2 bestimmt werden. Mit der Fixierung von $a_2 = 1$ ergibt sich mit

$$\begin{aligned} c_0 &= a_0 + a_1 + 1 = 0 \\ c_1 &= a_1 + 2 - (b_0 + b_1) = 0 \\ c_2 &= \frac{1}{2}(a_1 + 4) - b_1 = 0 \end{aligned}$$

ein Gleichungssystem mit 3 Gleichungen für 4 Unbekannte zur Verfüfgung, d.h. es gibt noch einen Freiheitsgrad.

Wählt man $a_1 = 0$, dann folgt für die restlichen Parameter $a_0 = -1, b_0 = 0$ und $b_1 = 2$, sodass das Verfahren die Form

$$y_{k+1} = y_{k-1} + 2hf_k \tag{8.62}$$

hat.

**Definition 8.20.** *Mit den Koeffizienten $a_j, b_j$ werden durch*

$$\rho(z) = \sum_{j=0}^{m} a_j z^j, \quad \sigma(z) = \sum_{j=0}^{m} b_j z^j \tag{8.63}$$

*das erste und zweite charakteristische Polynom eines m-Schritt-Verfahrens erklärt.*

Aus dem Gleichungssystem (8.60) kann man mit Hilfe der charakteristischen Polynome die folgende notwendige und hinreichende Bedingung für die Konsistenz eines Mehrschrittverfahrens formulieren.

**Satz 8.21.** *Notwendig und hinreichend für die Konsistenz des Mehrschrittverfahrens* (8.57) *ist die Erfüllung der Bedingungen*

$$c_0 = \rho(1) = 0, \quad c_1 = \rho'(1) - \sigma(1) = 0 \tag{8.64}$$

Macht man außer der Wahl von $a_2 = 1$ keine weiteren Einschränkungen an die Koeffizienten des expliziten 2-Schritt-Verfahrens (8.61), dann erreicht man die maximale Ordnung $p = 3$ durch die Lösung des Gleichungssystem (8.60) für $q = 3$, also $c_0 = c_1 = c_2 = c_3 = 0$.
Man findet die eindeutige Lösung

$$a_0 = -5, \quad a_1 = 4, \quad b_0 = 2, \quad b_1 = 4$$

woraus das Verfahren

$$y_{k+1} = 5y_{k-1} - 4y_k + h[4f_k + 2f_{k-1}] \tag{8.65}$$

folgt.

## 8.11 Stabilität von Mehrschrittverfahren

21. Vorlesung 11.01.2012

Obwohl das Verfahren (8.65) die maximale Fehlerordnung $p = 3$ hat, ist es im Vergleich zum Verfahren (8.62) unbrauchbar, weil es nicht stabil ist. Was das konkret bedeutet, soll im Folgenden erklärt und untersucht werden.

Dazu wird die Testdifferentialgleichung (AWP)

$$y' = \lambda y, \quad y(0) = 1,\ \lambda \in \mathbb{R},\ \lambda < 0 \tag{8.66}$$

mit der eindeutig bestimmten Lösung $y(t) = e^{\lambda t}$ betrachtet.
Von einem brauchbaren numerischen Verfahren erwartet man mindestens die Widerspiegelung des qualitativen Lösungsverhaltens.
Mit $f = \lambda y$ folgt aus (8.65)

$$\begin{aligned} & y_{k+1} = 5y_{k-1} - 4y_k + h[4\lambda y_k + 2\lambda y_{k-1}] \\ \Leftrightarrow \quad & (-5 - 2\lambda h)y_{k-1} + (4 - 4\lambda h)y_k + y_{k+1} = 0 \end{aligned} \tag{8.67}$$

Mit dem Lösungsansatz $y_k = z^k, z \neq 0$ ergibt (8.67) nach Division mit $z^{k-1}$

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

bzw. mit dem charakteristischen Polynomen

$$\rho(z) = -5 + 4z + z^2, \quad \sigma = 2 + 4z, \quad \phi(z) = \rho(z) - \lambda h\sigma(z) = 0\ .$$

Als Nullstellen von $\phi$ findet man

$$z_{1,2} = -2 + 2\lambda h \pm \sqrt{(2 - 2\lambda h)^2 + 5 + 2\lambda h}$$

und damit für die Lösung $y_k$ von (8.67)

$$y_k = c_1 z_1^k + c_2 z_2^k \tag{8.68}$$

wobei die Konstanten $c_1, c_2$ aus Anfangsbedingungen der Form $c_1 + c_2 = y_0, z_1 c_1 + z_2 c_2 = y_1$ zu ermitteln sind.
Notwendig für das Abklingen der Lösung $y_k$ in der Formel (8.68) für wachsendes $k$ ist die Bedingung $|z_{1,2}| \leq 1$. Da für $h \to 0$ die Nullstellen von $\phi(z)$ in die Nullstellen von $\rho(z)$ übergehen, dürfen diese dem Betrage nach nicht größer als 1 sein. Im Fall einer doppelten Nullstelle $z$ von $\phi(z)$ eines 2-Schritt-Verfahrens hat die Lösung der entsprechenden DGL statt (8.68) die Form

$$y_k = c_1 z^k + c_2 k z^k$$

sodass für das Abklingen von $y_k$ für wachsendes $k$ die Bedingung $|z| < 1$ erfüllt sein muss. Die durchgeführten Überlegungen rechtfertigen die folgende Definition.

**Definition 8.22.** *Das Mehrschrittverfahren* (8.57) *heißt* ***nullstabil****, falls die Nullstellen $z_j$ des ersten charakteristischen Polynoms $\rho(z)$*

*(a) betragsmäßig nicht größer als 1 sind*

*(b) mehrfache Nullstellen betragsmäßig echt kleiner als 1 sind*

Für das oben konstruierte 2-Schritt-Verfahren mit der maximalen Fehlerordnung $p = 3$ hat das charakteristische Polynom $\rho(z)$ die Nullstellen $z_{1,2} = -2 \pm 3$ und damit ist das Verfahren nicht nullstabil.
Im Unterschied dazu ist das Verfahren (8.62) mit der Ordnung 2 und dem charakteristischen Polynom $\rho(z) = -1 + z^2$ und den Nullstellen $z_{12} = \pm 1$ nullstabil.

**Bemerkung.** Einschrittverfahren sind mit dem charakteristischen Polynom $\rho(z) = -1 + z$ generell nullstabil.
Aufgrund der ersten charakteristischen Polynome der Adams-Bashforth- und Adams-Moulton-Verfahren erkennt man, dass diese auch generell nullstabil sind.

**Bemerkung.** Konsistente und nullstabile Mehrschrittverfahren sind konvergent, falls die Funktion $f(t, y)$ bezüglich $y$ lipschitzstetig ist.
Der Beweis verläuft im Fall expliziter Mehrschrittverfahren analog zum Konvergenzbeweis für konsistente Einschrittverfahren und sollte als Übung erbracht werden.

## 8.12 Begriff der absoluten Stabilität

Bei den Betrachtungen zur Nullstabilität wurde eine Testaufgabe zugrunde gelegt. Um den Begriff der **absoluten Stabilität** zu erläutern, wird die Testaufgabe leicht modifiziert, und zwar zu

$$y' = \lambda y, \quad y(0) = 1, \ \lambda \in \mathbb{R} \text{ oder } \lambda \in \mathbb{C} \tag{8.69}$$

d.h. wir lassen auch Parameter $\lambda$ aus $\mathbb{C}$ zu. Damit sind auch Lösungen der Form $e^{\alpha t}\cos(\beta t)$ möglich. Numerische Lösungsverfahren sollen auch in diesem Fall für $\alpha = \Re(\lambda) < 0$ den dann stattfindenden Abklingprozess korrekt wiedergeben. Für das Eulerverfahren zur Lösung von (8.69) erhält man mit $f(t, y) = \lambda y$

$$y_{k+1} = y_k + h\lambda y_k \ \Leftrightarrow \ y_{k+1} = (1 + h\lambda)y_k =: F(h\lambda)y_k$$

Falls $\lambda > 0$ und reell ist, wird die Lösung, für die

$$y(t_{k+1}) = y(t_k + h) = e^{h\lambda}y(t_k)$$

gilt, in jedem Fall qualitativ richtig wiedergegeben, denn der Faktor $F(h\lambda) = 1 + \lambda h$ besteht gerade aus den beiden ersten Summanden der $e$-Reihe, und es wird ein Fehler der Ordnung 2 gemacht.
Im Fall $\lambda < 0$ wird nur unter der Bedingung $|F(h\lambda)| = |1 + \lambda h| < 1$ das Abklingverhalten der Lösung beschrieben. Des Fall des reellen Parameters $\lambda < 0$ ist deshalb von Interesse.
Beim RK-Verfahren 4. Ordnung

$$\begin{aligned} k_1 &= \lambda y_k \\ k_2 &= \lambda(y_k + \frac{1}{2}hk_1) = (\lambda + \frac{1}{2}h\lambda^2)y_k \qquad (8.70) \\ k_3 &= \lambda(y_k - hk_1 + 2hk_2) = (\lambda + h\lambda^2 + h^2\lambda^3)y_k \\ y_{k+1} &= y_k + \frac{h}{6}[k_1 + 4k_2 + k_3] = (1 + h\lambda + \frac{h^2}{2}\lambda^2 + \frac{h^3}{6}\lambda^3)y_k \qquad (8.71) \end{aligned}$$

also $y_{k+1}$ als Produkt von $y_k$ mit dem Faktor

$$F(h\lambda) = 1 + h\lambda + \frac{h^2}{2}\lambda^2 + \frac{h^3}{6}\lambda^3 \qquad (8.72)$$

Der Faktor (8.72) enthält gerade die ersten 4 Summanden der $e$-Reihe und es wird ein Fehler der Ordnung 4 gemacht, sodass $y(t) = e^{t\lambda}$ durch das Verfahren (8.70) qualitativ beschrieben wird.
Für $\lambda < 0$ reell, muss die Lösung abklingen, was durch die Bedingung $|F(h\lambda)| < 1$ erreicht wird.
Wegen $\lim_{h\lambda\to-\infty} F(h\lambda) = -\infty$ wird das Abklingen nicht für beliebige negative Parameter $\lambda$ gesichert (nur für $\lambda$ mit $|F(h\lambda)| < 1$).
Auch im Fall eines komplexen Parameters $\lambda$ reicht die Bedingung $\alpha = \Re(\lambda) < 0$ nicht aus, um das Abklingen der Lösung der Testaufgabe zu sichern, sondern für $F$ muss $|F(h\lambda)| < 1$ gelten.
Die durchgeführten Überlegungen rechtfertigen die

**Definition 8.23.** *Für ein Einschrittverfahren, das für die Testaufgabe $y' = \lambda y, y(0) = 1, \lambda \in \mathbb{C}$ auf die Vorschrift $y_{k+1} = F(h\lambda)y_k$ führt, nennt man die Menge*

$$B := \{\mu \in \mathbb{C} : |F(\mu)| < 1\}$$

***Gebiet der absoluten Stabilität.***

Das Gebiet der absoluten Stabilität liefert eine Infomration zur Wahl der Schrittweite. Da man aber in den meisten "Ernstfällen" eventuelle Abklingkonstanten nicht kennt, hat man mit der Kenntnis von $B$ keine quantitative Bedingung zur Berechnung der Schrittweite zur Verfügung.

**Beispiel.**

- Euler explizit: $y_{k+1} = y_k + h\lambda y_k$
  $F(\mu) = 1 + \mu$
- Euler implizit: $y_{k+1} = y_k + h\lambda y_{k+1}$
  $F(\mu) = \frac{1}{1-\mu}$
- Runge-Kutta-Verfahren 2. Ordnung

$$k_1 = f(t_k, y_k), \quad k_2 = f(t_k + \frac{h}{2}, y_k + \frac{h}{2}k_1), \quad y_{k+1} = y_k + hk_2$$

$F(\mu) = 1 + \mu + \frac{\mu^2}{2}$

**Bemerkung.** Die Randkurve von $B$ erhält man wegen $\left|e^{i\theta}\right| = 1$ über die Parametrisierung

$$\begin{aligned} & F(\mu) = 1 + \mu + \frac{1}{2}\mu^2 = e^{i\theta} \\ \Leftrightarrow \quad & \mu^2 + 2\mu + 2 - 2e^{i\theta} = 0 \\ \rightsquigarrow \quad & \mu(\theta) = -1 \pm \sqrt{1 - 2 + 2e^{i\theta}}, \quad \theta \in [0, 2\pi] \end{aligned}$$

In der folgenden Tabelle sind die reellen Stabilitätsintervalle, d.h. die Schnittmenge der Gebiete der absoluten Stabilität mit der $\Re(\mu)$-Achse, für explizite $r$-stufige Runge-Kutta-Verfahren angegeben.

| r | |
|---|---|
| 1 | ]-2,0[ |
| 2 | ]-2,0[ |
| 3 | ]-2.51,0[ |
| 4 | ]-2.78,0[ |
| 5 | ]-3.21,0[ |

**Definition 8.24.** *Hat ein Einschrittverfahren als Gebiet der absoluten Stabilität mindestens die gesamte linke Halbebene, also $B \supset \{\mu \in \mathbb{C} : \Re(\mu) < 0\}$, dann nennt man das Verfahren absolut stabil.*

Neben dem impliziten Eulerverfahren (einstufiges RK-Verfahren) sind auch andere implizite RK-Verfahren absolut stabil, z.b. das Verfahren

$$k_1 = f\left(t_k + \frac{h}{2}, y_k + \frac{h}{2}k_1\right), \quad y_{k+1} = y_k + hk_1$$

Mit $f = \lambda y$ erhält man

$$k_1 = \frac{\lambda}{1 - \frac{h\lambda}{2}} y_k$$
$$y_{k+1} = y_k + hk_1 = \frac{1 + \frac{h\lambda}{2}}{1 - \frac{h\lambda}{2}} y_k =: F(h\lambda) y_k$$

und da für negatives $a$ gilt $|1 + a + bi| < |1 - a - bi|$ ist $|F(\mu)| < 1$.

## 8.13 BDF-Verfahren

Lineare Mehrschritt-Verfahren, bei denen bis auf den Koeffizienten $b_m$ alle anderen $b$-Koeffizienten gleich null sind, also Verfahren der Form

$$\sum_{j=0}^{m} a_j y_{k-m+1+j} = h b_m f(t_{k+1}, y_{k+1}) \ , \tag{8.73}$$

werden Rückwärtsdifferentiationsmethoden oder kurz **BDF-Verfahren** (backward differentiation formula) genannt.
Die Koeffizienten $a_j$ für ein $m$-Schritt-BDF-Verfahren bestimmt man, indem man das Interpolationspolynom $P_m$ mit den $m + 1$ Daten

$$(t_{k+1}, y_{k+1}), (t_k, y_k), \ldots, (t_{k+1-m}, y_{k+1-m})$$

bestimmt, und $y'$ an der Stelle $t_{k+1}$ durch $P'_m(t_{k+1})$ approximiert. Damit ergibt sich

$$P_m(t) = \sum_{j=0}^{m} L_j(\frac{t_{k+1} - t}{h}) y_{k+1-j} \quad \text{mit} \quad L_j(t) = \prod_{s=0, s\neq j}^{m} \frac{t - s}{j - s}$$

als den Lagrange'schen Basispolynomen. Für $P'_m(t_{k+1})$ erhält man nun

$$P'_m(t_{k+1}) = \sum_{j=0}^{m} (-\frac{1}{h}) L'_j(0) y_{k+1-j} \ ,$$

so dass sich das $m$-Schritt-BDF-Verfahren mit

$$\sum_{j=0}^{m} L'_j(0) y_{k+1-j} = h f_{k+1}$$

oder in der Standardform

$$y_{k+1} + \sum_{j=1}^{m} L_j'(0)/L_0'(0) y_{k+1-j} = h/(-L_0'(0)) f_{k+1}$$

ergibt. Diese Verfahren werden auch **Gear**-Verfahren genannt und haben die Konsistenzordnung $m$.
Die einfachsten 2- und 3-Schritt-BDF-Verfahren 2. und 3. Ordnung haben die Form

$$y_{k+1} - \frac{4}{3} y_k + \frac{1}{3} y_{k-1} = h \frac{2}{3} f(t_{k+1}, y_{k+1}) \,, \tag{8.74}$$

$$y_{k+1} - \frac{18}{11} y_k + \frac{9}{11} y_{k-1} - \frac{2}{11} y_{k-2} = h \frac{6}{11} f(t_{k+1}, y_{k+1}) \,. \tag{8.75}$$

Das einfachste BDF-Verfahren ist das so genannte Euler-rückwärts-Verfahren

$$y_{k+1} - y_k = h f(t_{k+1}, y_{k+1}) \,. \tag{8.76}$$

Für das Euler-rückwärts-Verfahren findet man für das Testproblem $y' = \lambda y$ schnell mit der Beziehung

$$y_{k+1} = \frac{1}{1 - h\lambda} y_k = F(h\lambda) y_k$$

heraus, dass $|F(h\lambda)| < 1$ für $Re(\lambda) < 0$ ist. D.h., das Euler-rückwärts-Verfahren ist absolut stabil. Das BDF-Verfahren (8.74) hat die charakteristische Gleichung

$$\phi(z) = z^2 - \frac{4}{3} z + \frac{1}{3} - \mu \frac{2}{3} z^2 = 0 \iff \mu(z) = \frac{3z^2 - 4z + 1}{2z^2} \,.$$

Für die Punkte $z = e^{i\theta}$, $\theta \in [0, 2\pi]$ erhält man die in der Abb. 8.1 skizzierte Randkurve $\mu(z(\theta))$ des Gebiets der absoluten Stabilität. Da man z.B. für $\mu = -\frac{1}{2}$ die Lösung $z_{1,2} = \frac{1}{2}$ mit $|z_{1,2}| < 1$ findet, kann man schlussfolgern, dass der Bereich der absoluten Stabilität im Außenbereich der Randkurve liegt. Damit ist das Verfahren (8.74) absolut stabil. Das Verfahren (8.75) ist nicht absolut stabil, weil das Gebiet der absoluten Stabilität nicht die gesamte linke komplexe Halbebene enthält. In der Abb. 8.1 ist der Rand des Gebietes der absoluten Stabilität des Verfahrens skizziert. Das Gebiet liegt wiederum im Außenbereich der Randkurve. In solchen Situationen kann man den Winkel $\alpha$ zwischen der reellen Achse und einer Tangente an die Randkurve durch den Ursprung legen. Bei dem BDF-Verfahren (8.75) ist der Winkel $\alpha = 88^o$, so dass das Verfahren $A(88^o)$-stabil ist. $A(90^o)$-Stabilität bedeutet absolute

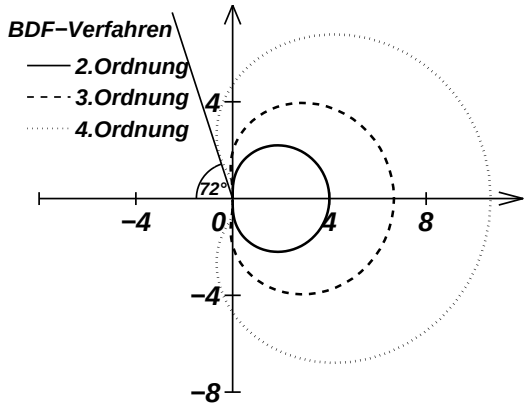



Abbildung 8.1: Gebiete der absoluten Stabilität der BDF-Verfahren (8.74), (8.75) und (8.77)

Stabilität. Liegt der Winkel $\alpha$ nahe bei $90^o$, dann liegt zwar kein absolut stabiles, jedoch ein "sehr" stabiles Verfahren vor. Bei BDF-Verfahren höherer Ordnung wird der Winkel $\alpha$ kleiner, so dass die Stabilität der BDF-Verfahren nachlässt, jedoch zumindest noch $A(\alpha)$-stabil sind. Zur Illustration ist das Gebiet der absoluten Stabilität des 4-Schritt-BDF-Verfahrens

$$y_{k+1} - \frac{48}{25}y_k + \frac{36}{25}y_{k-1} - \frac{16}{25}y_{k-2} + \frac{3}{25}y_{k-3} = h\frac{12}{25}f(t_{k+1}, y_{k+1}) \,, \qquad (8.77)$$

also die Kurve

$$\mu(z(\theta)) = \frac{25z^4 - 48z^3 + 36z^2 - 16z + 3}{12z^4} = \frac{25e^{i4\theta} - 48e^{i3\theta} + 36e^{i2\theta} - 16e^{i\theta} + 3}{12e^{i4\theta}} \,,$$

$\theta \in [0, 2\pi]$, in der Abbildung 8.1 im Vergleich zu den Verfahren (8.74) und (8.75) skizziert. Das Verfahren (8.77) ist $A(72^o)$-stabil.

# Kapitel 9

# Matrix-Eigenwertprobleme

22. Vorlesung am 16.01.2012

In vielen natur- und ingenieurwissenschaftlichen Disziplinen sind Eigenwertwertprobleme zu lösen. Zur Bestimmung von Eigenschwingungen von Bauwerken oder zur Ermittlung von stabilen statischen Konstruktionen sind Eigenwerte zu berechnen. Aber auch bei der Berechnung des Spektralradius bzw. der Norm einer Matrix sind Eigenwerte erforderlich.
Sowohl bei der Lösung von Differentialgleichungssystemen als auch bei Extremwertproblemen sind Eigenwerte von Matrizen Grundlage für die Konstruktion von Lösungen von Differentialgleichungen oder entscheiden über die Eigenschaften von stationären Punkten.
Bei der Berechnung von Eigenwerten und Eigenvektoren werden wir Ergebnisse aus vorangegangenen Semestern, speziell die $QR$-Zerlegung einer Matrix, als wichtiges Hilfsmittel nutzen können.

## 9.1 Problembeschreibung und algebraische Grundlagen

Gegeben ist eine reelle Matrix $A$ vom Typ $n \times n$, zum Beispiel die Koeffizientenmatrix eines linearen Differentialgleichungssystems

$$\begin{array}{rcrrr} x' & = & 2x & +y & -z \\ y' & = & x & +2y & +3z \\ z' & = & -x & +3y & +2z \end{array} \quad \Longleftrightarrow \quad \vec{x}' = A\vec{x},\ A = \begin{pmatrix} 2 & 1 & -1 \\ 1 & 2 & 3 \\ -1 & 3 & 2 \end{pmatrix} . \quad (9.1)$$

Wir werden sehen, dass man mit den Eigenwerten und Eigenvektoren der Matrix $A$ die Lösung des Differentialgleichungssystems (9.1) sehr schnell ermitteln kann.
Das Matrix-Eigenwertproblem ist wie folgt definiert.

**Definition 9.1.** *(Matrix-Eigenwertproblem)*
*Sei $A$ eine Matrix vom Typ $n \times n$. Der Vektor $\vec{x} \neq \vec{0}$ und die Zahl $\lambda$ heißen* ***Eigenvektor*** *bzw.* ***Eigenwert*** *der Matrix $A$, falls*

$$A\vec{x} = \lambda\vec{x} \tag{9.2}$$

*gilt. $\vec{x}$ bezeichnet man als Eigenvektor zum Eigenwert $\lambda$. Die Menge aller Eigenwerte eine Matrix $A$ heißt* ***Spektrum*** *von $A$ und wird durch $\sigma(A)$ bezeichnet. Die Gleichung (9.2) heißt* ***Eigengleichung****.*

Zur Definition 9.1 ist anzumerken, dass auch im Fall einer reellen Matrix $A$ die Eigenwerte und Eigenvektoren durchaus komplex sein können. Wir werden das später bei der Behandlung von Beispielen noch sehen.
Aus der Eigengleichung (9.2) folgt mit der Einheitsmatrix $E$

$$A\vec{x} - \lambda\vec{x} = A\vec{x} - \lambda E\vec{x} = (A - \lambda E)\vec{x} = \vec{0} \tag{9.3}$$

ein homogenes lineares Gleichungssystem, das nur dann eine Lösung $\vec{x} \neq \vec{0}$ hat, wenn die Matrix $A - \lambda E$ singulär ist. Damit gilt zur Bestimmung der Eigenwerte einer Matrix der

**Satz 9.2.** *(Eigenwertkriterium)*
*Für die Eigenwerte $\lambda$ einer Matrix $A$ gilt*

$$\chi_A(\lambda) := \det(A - \lambda E) = 0\ . \tag{9.4}$$

*$\chi_A$ heißt* ***charakteristisches Polynom*** *der Matrix $A$. Die Nullstellen von $\chi_A$ sind die Eigenwerte der Matrix $A$.*
*Die Eigenvektoren zu den Eigenwerten $\lambda$ ergeben sich dann als Lösung des homogenen linearen Gleichungssystems $(A - \lambda E)\vec{x} = \vec{0}$.*

**Beispiel 9.3.** Für Matrix $A$ aus (9.1) erhält man das charakteristische Polynom

$$\begin{aligned}
\det(A - \lambda E) &= \begin{vmatrix} 2-\lambda & 1 & -1 \\ 1 & 2-\lambda & 3 \\ -1 & 3 & 2-\lambda \end{vmatrix} \\
&= (2-\lambda)(2-\lambda)(2-\lambda) - 3 - 3 - 9(2-\lambda) - (2-\lambda) - (2-\lambda) \\
&= -\lambda^3 + 6\lambda^2 - \lambda - 20
\end{aligned}$$

und mit etwas Glück durch Probieren die Nullstelle $\lambda_1 = 5$ sowie nach Polynomdivision die weiteren Nullstellen $\lambda_{2,3} = \frac{1}{2} \pm \frac{\sqrt{17}}{2}$. In der Regel hat man nicht immer solches Glück bei der Eigenwertbestimmung, sondern man muss die Nullstellen numerisch berechnen.

Dabei stellt man bei dem Weg über die Nullstellen des charakteristischen Polynoms sehr schnell fest, dass die Berechnung nicht stabil ist, sondern dass kleine Fehler in den Polynomkoeffizienten mitunter zu gestörten Nullstellen, die sich wesentlich von den exakten unterscheiden, führen können. Im Folgenden werden iterative Methoden zur Bestimmung von Eigenwerten und Eigenvektoren behandelt, ohne das Kriterium 9.2 zu verwenden.
Bevor wir zu den konkreten Berechnungsmethoden von Eigenwerten und Eigenvektoren kommen, fassen wir an dieser Stelle einige wichtige und nützliche Grundlagen der linearen Algebra zum Spektralverhalten von Matrizen zusammen. Eine wichtige Rolle spielen die im Folgenden definierten Begriffe.

**Definition 9.4.** *(ähnliche Matrizen)*
*Die $(n \times n)$-Matrix $\tilde{A}$ ist der Matrix $A$* ***ähnlich****, wenn eine reguläre $(n \times n)$-Matrix $C$ existiert, so dass*

$$\tilde{A} = C^{-1}AC$$

*gilt. Man sagt dann, dass $\tilde{A}$ aus $A$ durch eine reguläre Transformation mit $C$ hervorgegangen ist. Ist die Matrix $C$ eine orthogonale Matrix, dann bezeichnet man $\tilde{A}$ auch als* ***Orthogonaltransformation*** *von $A$ und mit $C^{-1} = C^T$ gilt dann*

$$\tilde{A} = C^T AC \ .$$

*Gibt es eine reguläre Matrix $C$, so dass die Transformation von $A$*

$$D = C^{-1}AC$$

*mit $D$ eine Diagonalmatrix ergibt, dann heißt $A$* ***diagonalisierbar****.*

Für das Spektrum bzw. die Eigenwerte spezieller Matrizen kann man aus der Definition 9.1 folgende Eigenschaften zeigen.

**Satz 9.5.** *(Eigenwerte spezieller Matrizen)*
*Sei $A$ eine $(n \times n)$-Matrix über $\mathbb{C}$. Dann gilt:*

*a) Ist $A$ eine Dreiecksmatrix, dann sind die Diagonalelemente gerade die Eigenwerte.*

*b) Ist $\tilde{A}$ eine reguläre Transformation der Matrix $A$ mit der regulären Matrix $C$, dann haben $\tilde{A}$ und $A$ die gleichen Eigenwerte.*

*c) Sind $\lambda_1, \ldots, \lambda_r$ die Eigenwerte von $A$, so besitzt die Matrix $A_\epsilon = A + \epsilon E$ die Eigenwerte $\mu_j = \lambda_j + \epsilon$ $(j = 1, \ldots, r)$.*

*d) Ist $A$ regulär mit den Eigenwerten $\lambda_1, \ldots, \lambda_r$, dann sind die Eigenwerte verschieden von null und die Inverse $A^{-1}$ hat die Eigenwerte $\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_r}$.*

*e) Die transponierte Matrix $A^T$ hat die gleichen Eigenwerte wie die Matrix $A$.*

Die Aussagen des Satzes 9.5 sind einfach zu zeigen und der Nachweis wird zur Übung empfohlen. Oben wurde schon darauf hingewiesen, dass auch bei Matrizen mit ausschließlich reellen Elementen komplexe Eigenwerte auftreten können. Als Beispiel betrachten wir die Matrix

$$A = \begin{pmatrix} 1 & 5 \\ -1 & 3 \end{pmatrix}$$

und finden als Nullstellen des charakteristischen Polynoms $\chi_A(\lambda) = \lambda^2 - 4\lambda + 8$ die Eigenwerte $\lambda_{1,2} = 2 \pm 2i$. An dieser Stelle sei daran erinnert, dass Polynome mit ausschließlich reellen Koeffizienten, was bei den charakteristischen Polynomen reeller Matrizen der Fall ist, immer eine gerade Zahl $(0, 2, 4, \dots)$ von komplexen Nullstellen haben. Denn wenn überhaupt komplexe Nullstellen auftreten, dann immer als Paar der komplexen Zahl $\lambda$ mit der konjugiert komplexen Zahl $\bar{\lambda}$.
Allerdings gibt es eine große Klasse von reellen Matrizen, die ausschließlich reelle Eigenwerte besitzen. Es gilt der

**Satz 9.6.** *(Eigenschaften symmetrischer reeller Matrizen)*
*Für jede reelle symmetrische $(n \times n)$-Matrix $S$ gilt:*

*a) Alle Eigenwerte von $S$ sind reell.*

*b) Eigenvektoren $\vec{q}_k$, $\vec{q}_j$, die zu verschiedenen Eigenwerten $\lambda_k \neq \lambda_j$ von $S$ gehören, stehen senkrecht aufeinander, d.h., $\vec{q}_k^T \vec{q}_j = \langle \vec{q}_k, \vec{q}_j \rangle = 0$.*

*c) Es gibt $n$ Eigenvektoren $\vec{q}_1, \dots, \vec{q}_n$ von $S$, die eine Orthonormalbasis des $\mathbb{R}^n$ bilden.*

*d) Die Matrix $S$ ist diagonalisierbar.*

*e) Die spezielle symmetrische Matrix $S = A^T A$, wobei $A$ eine beliebige reelle $(n \times n)$-Matrix ist, hat nur nichtnegative Eigenwerte.*

Zum Nachweis von a). Wir bezeichnen mit $\mathbf{x}^*$ den Vektor $\overline{\mathbf{x}}^T$, wobei $\overline{\mathbf{x}}$ der konjugiert komplexe Vektor zu $\mathbf{x}$ ist. Sei nun $\lambda$ ein Eigenwert von $S$ und $\mathbf{x}$ ein zugehöriger Eigenvektor. Damit ist $\mathbf{x}^*\mathbf{x} = |\mathbf{x}|^2 =: r > 0$ reell und es folgt

$$\mathbf{x}^* S\mathbf{x} = \mathbf{x}^* \lambda \mathbf{x} = \lambda \mathbf{x}^* \mathbf{x} = \lambda r \ .$$

Für jede komplexe Zahl $z$, aufgefasst als $(1 \times 1)$-Matrix gilt $z = z^T$. Damit und aus der Symmetrie von $S$ folgt für die komplexe Zahl $\mathbf{x}^* S\mathbf{x}$

$$\mathbf{x}^* S\mathbf{x} = (\mathbf{x}^* S\mathbf{x})^T = \mathbf{x}^T S \mathbf{x}^{*T} = \overline{\mathbf{x}^*} S \overline{\mathbf{x}} = \overline{\mathbf{x}^* S \mathbf{x}} = \overline{\lambda r} = \overline{\lambda} r$$

Es ergibt sich schließlich $\overline{\lambda} r = \lambda r$, d.h., $\lambda$ ist reell.
Wegen der Voraussetzung $\lambda_k \neq \lambda_j$ für die Aussage b) muss einer dieser Eigenwerte von null verschieden sein, z.B. $\lambda_k \neq 0$. Aus $S\vec{q}_k = \lambda_k \vec{q}_k$ folgt

$$\vec{q}_k = \frac{1}{\lambda_k} S\vec{q}_k \quad \text{sowie} \quad \vec{q}_k^T = \frac{1}{\lambda_k} \vec{q}_k^T S^T = \frac{1}{\lambda_k} \vec{q}_k^T S \; .$$

Daraus folgt

$$\vec{q}_k^T \vec{q}_j = \frac{1}{\lambda_k} \vec{q}_k^T S\vec{q}_j = \frac{1}{\lambda_k} \vec{q}_k^T \lambda_j \vec{q}_j = \frac{\lambda_j}{\lambda_k} \vec{q}_k^T \vec{q}_j$$

und aus dieser Gleichung folgt

$$(1 - \frac{\lambda_j}{\lambda_k}) \vec{q}_k^T \vec{q}_j = 0 \Longleftrightarrow \vec{q}_k^T \vec{q}_j = \langle \vec{q}_k, \vec{q}_j \rangle = 0 \; .$$

Zu c) sei nur angemerkt, dass man im Fall eines Eigenwerts $\lambda_k$, der insgesamt $\sigma_k$-mal auftritt (algebraische Vielfachheit gleich $\sigma_k$), als Lösung des homogenen linearen Gleichungssystems $(S - \lambda_k E)\vec{q} = \vec{0}$ immer $\sigma_k$ orthogonale Eigenvektoren $\vec{q}_{k1}, \ldots, \vec{q}_{k\sigma_k}$ finden kann, so dass man auch im Fall mehrfacher Eigenwerte der symmetrischen $(n \times n)$-Matrix $S$ immer $n$ **orthogonale** bzw. nach Normierung **orthonormierte** Eigenvektoren $\vec{q}_1, \ldots, \vec{q}_n$ finden kann.
Die mit den orthonormierten Eigenvektoren gebildete Matrix

$$Q = \begin{pmatrix} | & | & & | \\ \vec{q}_1 & \vec{q}_2 & \ldots & \vec{q}_n \\ | & | & & | \end{pmatrix}$$

ist wegen $\langle \vec{q}_k, \vec{q}_j \rangle = \delta_{kj}$ orthogonal und es gilt für $k = 1, \ldots, n$

$$S\vec{q}_k = \lambda_k \vec{q}_k \;\; (k = 1, \ldots, n) \;\; \Longleftrightarrow \;\; SQ = QD \;\; \Longleftrightarrow \;\; D = Q^T SQ \; ,$$

wobei die Diagonalmatrix $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$ genau die Eigenwerte $\lambda_1, \ldots, \lambda_n$ als Hauptdiagonalelemente hat, also ist $S$ diagonalisierbar.
e) ergibt sich durch die einfache Rechnung mit dem Eigenvektor $\vec{q}$ von $S$ zum Eigenwert $\lambda$

$$\lambda ||\vec{q}||^2 = \langle \lambda\vec{q}, \vec{q} \rangle = \langle S\vec{q}, \vec{q} \rangle = \langle A^T A\vec{q}, \vec{q} \rangle = \langle A\vec{q}, A\vec{q} \rangle = ||A\vec{q}||^2 \geq 0 \; .$$

## 9.2 Abschätzungen und Lokalisierung von Eigenwerten

Zur Lokalisierung der Eigenwerte einer $(n \times n)$-Matrix $A = (a_{ij})$ dient der folgende

**Satz 9.7.** *(Lokalisierung von Eigenwerten in Gerschgorin-Kreisen)*
*Sei $A = (a_{ij})$ eine $(n \times n)$-Matrix mit den* ***Gerschgorin-Kreisen***

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

*a) Dann gilt für das Spektrum $\sigma(A)$ von $A$*

$$\sigma(A) \subset \bigcup_{j=1}^{n} K_j \; ,$$

*d.h., sämtliche Eigenwerte von A liegen in der Vereinigung der Gerschgorin-Kreise.*
*b) Es sei $\{i_1, \ldots, i_k\} \cup \{i_{k+1}, \ldots, i_n\} =: I_1 \cup I_2 = \{1, 2, \ldots, n\}$. Sind die Gerschgorin-Kreise $K_a = \cup_{i \in I_1} K_i$ und $K_b = \cup_{i \in I_2} K_i$ disjunkt, dann liegen in $K_a$ genau $k$ und in $K_b$ genau $n - k$ Eigenwerte von $A$.*

*Beweis.* Zum Nachweis von a) betrachten wir einen zum Eigenwert $\lambda$ gehörenden Eigenvektor $\vec{u}$. $u_j$ sei eine Koordinate von $\vec{u}$ mit

$$|u_j| = ||\vec{u}||_\infty = \max_{k=1,\ldots,n} |u_k| \; .$$

Die $j$-te Gleichung der Eigengleichung $A\vec{u} = \lambda\vec{u}$ ist

$$\sum_{k=1}^{n} a_{jk} u_k = \lambda u_j$$

und es ergibt sich

$$|a_{jj} - \lambda| \, |u_j| = |\sum_{\substack{k=1 \\ k \neq j}}^{n} a_{jk} u_k| \leq ||\vec{u}||_\infty \sum_{\substack{k=1 \\ k \neq j}}^{n} |a_{jk}| = |u_j| \sum_{\substack{k=1 \\ k \neq j}}^{n} |a_{jk}| \; .$$

Daraus folgt $|a_{jj} - \lambda| \leq \sum_{\substack{k=1 \\ k \neq j}}^{n} |a_{jk}|$, d.h., $\lambda$ liegt in $K_j$.
Zum Nachweis von b) betrachten wir mit $D$ die Diagonale von $A$ und $N = A - D$. Sei $A(\epsilon) = D + \epsilon N$ mit den Eigenwerten $\lambda(\epsilon)$. Für $\epsilon = 0$ bestehen die Kreise $K_i(\epsilon)$ aus den durch die Diagonalelemente gegebenen Punkten, die beim stetigen Vergrößern von $\epsilon = 0$ zu $\epsilon = 1$ zu den Gerschgorin-Kreisen $K_i = K_i(1)$ von $A$ anwachsen (die Radien sind proportional zu $\epsilon$ und es gilt $K_i(\epsilon_1) \subset K_i(\epsilon_2)$ für $\epsilon_1 \leq \epsilon_2$). Die Eigenwerte hängen stetig von den Matrixelementen und damit von $\epsilon$ ab und können aufgrund der Aussage a) wegen der Disjunktheit nicht zwischen $K_a$ und $K_b$ wechseln. □

**Beispiel 9.8.** 1) Die Matrix $A = \left(\begin{smallmatrix} 1 & 5 \\ -1 & 3 \end{smallmatrix}\right)$ hat die Gerschgorin-Kreise

$$K_1 = \{z \in \mathbb{C} \,|\, |z-1| \leq 5\} \quad \text{und} \quad K_2 = \{z \in \mathbb{C} \,|\, |z-3| \leq 1\} \,.$$

Die oben berechneten Eigenwerte $\lambda_{1,2} = 2 \pm 2i$ liegen in $K_1 \cup K_2 = K_1$, wie in der Abb. 9.1 zu erkennen ist.

2) Die Matrix $B = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 0,5 & 7 \end{pmatrix}$ hat die Gerschgorin-Kreise

$$K_1 = \{z \in \mathbb{C} \,|\, |z-4| \leq 1\},\ K_2 = \{z \in \mathbb{C} \,|\, |z-2| \leq 2\},\ K_3 = \{z \in \mathbb{C} \,|\, |z-7| \leq 1,5\},$$

die in der Abb. 9.2 dargestellt sind (Eigenwerte $\lambda_1 = 4,26, \lambda_2 = 7,1681, \lambda_3 = 1,5791$).

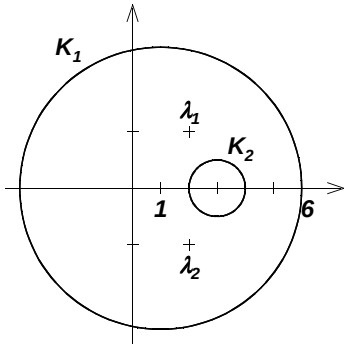



Abbildung 9.1: Gerschgorin-Kreise und Eigenwerte von $A$

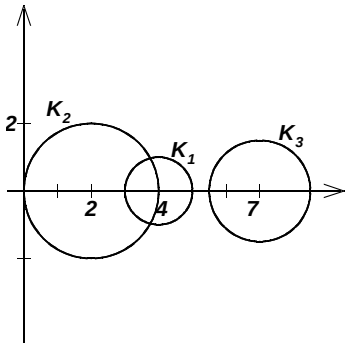



Abbildung 9.2: Gerschgorin-Kreise von $B$

**Definition 9.9.** *Der **Rayleigh-Quotient** von $\vec{x} \neq 0$ bezügl. der Matrix $A$ ist durch*

$$r_A(\vec{x}) = \frac{\langle \vec{x}, A\vec{x} \rangle}{\langle \vec{x}, \vec{x} \rangle}$$

*definiert.*

Der Rayleigh-Quotient ist ein wichtiges Hilfsmittel zur Eigenwertabschätzung. Es gilt der

**Satz 9.10.** *Sei $A$ reell und symmetrisch, $\vec{x} \in \mathbb{R}^n \setminus \{0\}$ beliebig.*
*a) Mit dem kleinsten bzw. größten Eigenwert $\lambda_{min}$ bzw. $\lambda_{max}$ von $A$ gilt*

$$\lambda_{min} \leq r_A(\vec{x}) \leq \lambda_{max} \,.$$

*Die Extremwerte werden für die entsprechenden Eigenvektoren $\vec{x}$ angenommen.*

*b) Eigenwertabschätzung durch den Rayleigh-Quotienten eines Testvektors: es existiert ein Eigenwert $\lambda$ von $A$ mit*

$$|\lambda - r_A(\vec{x})|^2 \leq \underbrace{r_{A^2}(\vec{x}) - [r_A(\vec{x})]^2}_{Auslöschungsgefahr} = \underbrace{\frac{||(A - r_A(\vec{x})E)\vec{x}||_2^2}{\langle \vec{x}, \vec{x}\rangle}}_{numerisch\ stabil} .$$

*Beweis.*
a) Sei $\vec{x}_1, \ldots, \vec{x}_n$ eine Orthonormalbasis von Eigenvektoren ($A\vec{x}_i = \lambda_i \vec{x}_i$). Mit $\vec{x} = \sum_i x_i \vec{x}_i$ folgt

$$r_A(\vec{x}) = \frac{\langle \sum_i x_i\vec{x}_i, \sum_i \lambda_i x_i \vec{x}_i\rangle}{\langle \sum_i x_i\vec{x}_i, \sum_i x_i \vec{x}_i\rangle} = \frac{\sum_i \lambda_i x_i^2}{\sum_i x_i^2} \begin{cases} \leq & \lambda_{max} \\ \geq & \lambda_{min} \end{cases} .$$

b) Sei $\mu$ nicht Eigenwert von $A$, dann gilt

$$1 = \frac{||(A-\mu E)^{-1}(A-\mu E)\vec{x}||_2^2}{||\vec{x}||_2^2} \leq ||(A-\mu E)^{-1}||_2^2 \frac{||(A-\mu E)\vec{x}||_2^2}{||\vec{x}||_2^2} ,$$

und damit

$$\begin{aligned} \frac{||(A-\mu E)\vec{x}||_2^2}{||\vec{x}||_2^2} &\geq \frac{1}{||(A-\mu E)^{-1}||_2^2} = \frac{1}{\rho((A-\mu E)^{-1})^2} \\ &= \frac{1}{\max_i |\lambda_i - \mu|^{-2}} = \min_{i=1,\ldots,n} |\lambda_i - \mu|^2 . \end{aligned}$$

Außerdem folgt für beliebiges $\mu$ auch

$$\frac{\langle (A-\mu E)\vec{x}, (A-\mu E)\vec{x}\rangle}{\langle \vec{x}, \vec{x}\rangle} = \frac{||(A-\mu E)\vec{x}||_2^2}{||\vec{x}||_2^2} \geq \min_{i=1,\ldots,n} |\lambda_i - \mu|^2 .$$

Aufgrund von

$$\frac{\langle (A-\mu E)\vec{x}, (A-\mu E)\vec{x}\rangle}{\langle \vec{x}, \vec{x}\rangle} = \underbrace{\frac{\langle A\vec{x}, A\vec{x}\rangle}{\langle \vec{x}, \vec{x}\rangle}}_{r_{A^2}(\vec{x})} - [\underbrace{\frac{\langle \vec{x}, A\vec{x}\rangle}{\langle \vec{x}, \vec{x}\rangle}}_{r_A(\vec{x})}]^2 + [\mu - \frac{\langle \vec{x}, A\vec{x}\rangle}{\langle \vec{x}, \vec{x}\rangle}]^2$$

wird die Abschätzung optimal für $\mu = r_A(\vec{x})$ und b) gilt. □

**Bemerkung 9.11.** Wenn $\vec{x}$ ein Eigenvektor ist, dann ergibt der Rayleigh-Quotient $r_A(\vec{x})$ den entsprechenden Eigenwert. Rayleigh-Quotienten werden als Hilfsmittel benutzt, um aus einer Approximation eines Eigenvektors eine Approximation eines Eigenwerts abzuleiten.

**Bemerkung 9.12.** Es stellt sich die Frage, wann $\vec{x} \in \mathbb{R}^n$ (bzw. $\mathbb{C}^n$) Approximation eines Eigenvektors ist. Es sei $E$ der Eigenraum zum Eigenwert $\lambda$ und $F$ der von der restlichen Eigen- bzw. Hauptvektoren aufgespannte Raum, so dass $\mathbb{R}^n = E \oplus F$ (bzw. $\mathbb{C}^n = E \oplus F$) gilt. Mit der Zerlegung $\vec{x} = \vec{x}_E + \vec{x}_F$ mit $\vec{x}_E \in E$, $\vec{x}_F \in F$ vereinbart man:

$$\begin{aligned} &\vec{x} \text{ ist approximativer Eigenvektor zum Eigenwert } \lambda \\ &\qquad \Longleftrightarrow \vec{x} \approx \vec{x}_E \Longleftrightarrow ||\vec{x}_F|| << ||\vec{x}_E|| \ . \end{aligned}$$

Für symmetrische Matrizen ist $F$ das orthogonale Komplement von $E$ ($\langle \vec{x}_E, \vec{x}_F \rangle = 0$). Der Winkel $\phi$ zwischen $\vec{x}$ und seiner orthogonalen Projektion $\vec{x}_E$ auf den Eigenraum $E$, definiert durch

$$\cos^2 \phi = \frac{\langle \vec{x}_E, \vec{x}_E \rangle}{\langle \vec{x}, \vec{x} \rangle} \quad \text{bzw.} \quad \sin^2 \phi = \frac{\langle \vec{x}_F, \vec{x}_F \rangle}{\langle \vec{x}, \vec{x} \rangle} \ ,$$

ist ein Maß für den Abstand von $\vec{x}$ zum Eigenraum $E$.

**Satz 9.13.** *Für eine symmetrische Matrix $A$ mit den Eigenwerten $\lambda_i$ gilt*

$$(\min_{\lambda_i \neq \lambda} |\lambda - \lambda_i|) \sin^2 \phi \leq |\lambda - r_A(\vec{x})| \leq (\max_{\lambda_i \neq \lambda} |\lambda - \lambda_i|) \sin^2 \phi \ .$$

*Beweis.*

$$\begin{aligned} \lambda - r_A(\vec{x}) &= \lambda - \frac{\langle \vec{x}_E + \vec{x}_F, A(\vec{x}_E + \vec{x}_F) \rangle}{\langle \vec{x}, \vec{x} \rangle} \\ &= \lambda - \frac{\langle \vec{x}_E, A\vec{x}_E \rangle}{\langle \vec{x}, \vec{x} \rangle} - 2\frac{\langle \vec{x}_F, A\vec{x}_E \rangle}{\langle \vec{x}, \vec{x} \rangle} - \frac{\langle \vec{x}_F, A\vec{x}_F \rangle}{\langle \vec{x}, \vec{x} \rangle} \\ &= \lambda - \frac{\langle \vec{x}_E, A\vec{x}_E \rangle}{\langle \vec{x}, \vec{x} \rangle} - \frac{\langle \vec{x}_F, A\vec{x}_F \rangle}{\langle \vec{x}, \vec{x} \rangle} \\ &= \lambda \underbrace{(1 - \frac{\langle \vec{x}_E, \vec{x}_E \rangle}{\langle \vec{x}, \vec{x} \rangle})}_{1-\cos^2 \phi} - \underbrace{\frac{\langle \vec{x}_F, \vec{x}_F \rangle}{\langle \vec{x}, \vec{x} \rangle}}_{\sin^2 \phi} \frac{\langle \vec{x}_F, A\vec{x}_F \rangle}{\langle \vec{x}_F, \vec{x}_F \rangle} \\ &= \sin^2 \phi (\lambda - \frac{\langle \vec{x}_F, A\vec{x}_F \rangle}{\langle \vec{x}_F, \vec{x}_F \rangle}) \ . \end{aligned}$$

Der auf $F$ eingeschränkte Rayleigh-Quotient nimmt in Analogie zu Satz 9.10 a) als Extremwerte einen der von $\lambda$ verschiedenen Eigenwerte von $A$ an, d.h.

$$\min_{\lambda_i \neq \lambda} \lambda_i \leq \frac{\langle \vec{x}_F, A\vec{x}_F \rangle}{\langle \vec{x}_F, \vec{x}_F \rangle} \leq \max_{\lambda_i \neq \lambda} \lambda_i \ ,$$

und damit folgt die Aussage des Satzes. □

**Bemerkung 9.14.** Der Satz 9.13 zeigt aufgrund des Faktors $\sin^2\phi$, dass im Falle von symmetrischen Matrizen bei verhältnismäßig schlechten Eigenvektorapproximationen durch den Rayleigh-Quotienten trotzdem gute Eigenwertapproximationen geliefert werden.

**Satz 9.15.** *Sei $A = T\Lambda T^{-1}$ mit $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$ eine diagonalisierbare Matrix mit den Eigenwerten $\lambda_1, \ldots, \lambda_n$. Für einen beliebigen Eigenwert $\tilde{\lambda}$ einer gestörten Matrix $\tilde{A} = A + \Delta A$ gilt*

$$\min_{i=1,\ldots,n} |\lambda_i - \tilde{\lambda}| \leq cond_p(T)||\Delta A||_p \; .$$

*Beweis.* $\tilde{\lambda}$ sei nicht Eigenwert von $A$ (ansonsten wird es trivial). Es folgt

$$||(A - \tilde{\lambda}E)^{-1}||_p = ||T(\Lambda - \tilde{\lambda}E)^{-1}T^{-1}||_p \leq \mathrm{cond}_p(T)||(\Lambda - \tilde{\lambda}E)^{-1}||_p \; .$$

Da die $p$-Norm einer Diagonalmatrix gleich dem maximalen Betrag der Diagonalelemente ist, gilt

$$||(\Lambda - \tilde{\lambda}E)^{-1}||_p = \max_{i=1,\ldots,n} \frac{1}{|\lambda_i - \tilde{\lambda}|} = \frac{1}{\min_{i=1,\ldots,n} |\lambda_i - \tilde{\lambda}|} \; .$$

Es folgt nun

$$\min_{i=1,\ldots,n} |\lambda_i - \tilde{\lambda}| \leq \frac{\mathrm{cond}_p(T)}{||(A - \tilde{\lambda}E)^{-1}||_p} \; . \tag{9.5}$$

Mit einem Eigenvektor $\vec{y}$ von $\tilde{A}$ zum Eigenwert $\tilde{\lambda}$ ergibt sich

$$\tilde{A}\vec{y} = \tilde{\lambda}\vec{y} \implies (A - \tilde{A})\vec{y} = (A - \tilde{\lambda})\vec{y} \implies (A - \tilde{\lambda}E)^{-1}(A - \tilde{A})\vec{y} = \vec{y}$$

und weiter

$$1 \leq ||(A - \tilde{\lambda}E)^{-1}(A - \tilde{A})||_p \leq ||(A - \tilde{\lambda}E)^{-1}||_p||A - \tilde{A}||_p \; ,$$

also $1/||(A - \tilde{\lambda}E)^{-1}||_p \leq ||\Delta A||_p$. Unter Nutzung von (9.5) folgt die Behauptung. □

Da man symmetrische Matrizen mit orthogonalen Matrizen (bestehend aus den orthogonalen Eigenwerten) diagonalisieren kann, gilt für symmetrische Matrizen $A$ und beliebige Matrizen $\tilde{A} = A + \Delta A$

$$\min_{i=1,\ldots,n} |\lambda_i - \tilde{\lambda}| \leq ||\Delta A||_2 \; ,$$

da man eine Transformationsmatrix $T$ mit $\mathrm{cond}_2(T) = 1$ findet.
Ohne Beweis wird noch ein Vergleichssatz für Eigenwerte symmetrischer Matrizen angegeben.

**Satz 9.16.** *Für symmetrische reelle $(n \times n)$-Matrizen $A$ und $\tilde{A}$ mit den Eigenwerten*

$$\lambda_1 \leq \cdots \leq \lambda_n \text{ von } A \quad \text{bzw.} \quad \tilde{\lambda}_1 \leq \ldots \tilde{\lambda}_n \text{ von } \tilde{A}$$

*gilt*

$$|\lambda_i - \tilde{\lambda}_i| \leq \rho(A - \tilde{A}) \leq ||A - \tilde{A}||$$

*für beliebige Matrixnormen.*

## 9.3 Numerische Methoden zur Eigenwertberechnung

23. Vorlesung am 18.01.2012

Es geht zuerst darum, die Aufgabe der Eigenwertberechnung zu vereinfachen. Dazu werden ausgehend von $A$ einfachere ähnliche Matrizen konstruiert.
Zur Eigenwertberechnung werden dann z.B. Newtonverfahren, Jacobi-Verfahren und die Givensrotation genutzt.

### 9.3.1 Transformation auf Hessenberg- bzw. Tridiagonalform

Das Ziel der nächsten Überlegungen ist die Konstruktion einer Matrix $H$, die der Matrix $A$, von der wir Eigenwerte suchen, ähnlich sind, allerdings eine wesentlich einfachere Gestalt als $A$ haben. Die einfachere Bestimmung der Eigenwerte von $H$ ergibt dann die Lösung des Eigenwertproblems von $A$.

**Definition 9.17.** *Unter einer **Hessenberg**-Matrix versteht man eine Matrix $H = (h_{ij})$, für die $h_{ij} = 0$ für $i > j + 1$ gilt, also eine Matrix der Form*

$$H = \begin{pmatrix} h_{11} & h_{12} & \dots & h_{1\,n-1} & h_{1n} \\ h_{21} & h_{22} & \dots & h_{2\,n-1} & h_{2n} \\ 0 & h_{32} & \dots & h_{3\,n-1} & h_{3n} \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & h_{n\,n-1} & h_{nn} \end{pmatrix},$$

*die unter der Hauptdiagonale nur ein Band besitzt.*

Wir werden nun zeigen, dass man jede Matrix $A$ durch eine orthogonale Ähnlichkeitstransformation auf Hessenberg-Form transformieren kann, d.h., dass es eine orthogonale Matrix $Q$ mit

$$H = Q^T A Q$$

gibt. Betrachten wir dazu mit $\vec{a}_1$ die erste Spalte von $A$. Wir suchen nun eine Householder-Matrix

$$H_1 = E - 2\frac{\vec{u}_1 \vec{u}_1^T}{\langle \vec{u}_1, \vec{u}_1 \rangle},$$

so dass sich mit $\vec{a}_1^{(1)} = H_1 \vec{a}_1 = (a_{11}, *, 0, \dots, 0)^T$ ein Vektor ergibt, der bis auf die ersten beiden Komponenten nur Null-Komponenten besitzt. Analog zum Vorgehen bei der Erzeugung von $QR$-Zerlegungen leistet der Vektor

$$\vec{u}_1 = (0, c + a_{21}, a_{31}, \dots, a_{n1})^T$$

mit $c = \text{sign}(a_{21})\sqrt{a_{21}^2 + \cdots + a_{n1}^2}$ das Geforderte. Es ergibt sich

$$\vec{a}_1^{(1)} = H_1\vec{a}_1 = (a_{11}, -c, 0, \ldots, 0)^T \ .$$

Für die $j$-te Spalte $\vec{a}_j$ von $A$ erzeugt die Householder-Matrix

$$H_j = E - 2\frac{\vec{u}_j\vec{u}_j^T}{\langle \vec{u}_j, \vec{u}_j\rangle} \tag{9.6}$$

mit

$$\vec{u}_j = (0, \ldots, 0, c + a_{j+1\,j}, \ldots, a_{nj})^T \text{ und } c = \text{sign}(a_{j+1\,j})\sqrt{a_{j+1\,j}^2 + \cdots + a_{nj}^2}$$

einen Vektor $\vec{a}_j^{(j)} = H_j\vec{a}_j = (a_{1j}, \ldots, a_{jj}, -c, 0, \ldots, 0)^T$, der bis auf die ersten $j+1$ Komponenten nur Null-Komponenten besitzt. Die Multiplikation einer Matrix $A$ mit der Householder-Matrix $H_j$ (9.6) lässt alle Spalten der Form

$$\vec{s} = (s_1, s_2, \ldots, s_j, 0, \ldots, 0)^T$$

invariant, d.h., es gilt $H_j\vec{s} = \vec{s}$. Damit bleiben durch die Multiplikation von $A$ mit Householder-Matrizen $H_1, \ldots, H_{j-1}$ erzeugte Nullen im unteren Dreieck erhalten, d.h., mit den Householder-Matrizen $H_1, \ldots, H_{n-2}$ erhält man mit

$$G = H_{n-2}H_{n-3}\ldots H_1A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1\,n-1} & a_{1n} \\ g_{21} & g_{22} & \ldots & g_{2\,n-1} & g_{2n} \\ 0 & g_{32} & \ldots & g_{3\,n-1} & g_{3n} \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \ldots & 0 & g_{n\,n-1} & g_{nn} \end{pmatrix}$$

eine Hessenberg-Matrix. Man überprüft durch Nachrechnen, dass die Multiplikation der Matrix $G$ von rechts mit den Householder-Matrizen $H_1, \ldots, H_{n-2}$ die Hessenberg-Form nicht zerstört. Man erkennt nun, dass die Matrix $H_1AH_1$ wieder eine Hessenberg-Matrix ist. Insgesamt erhält man mit

$$H = H_{n-2}H_{n-3}\ldots H_1AH_1H_2\ldots H_{n-2} = \begin{pmatrix} a_{11} & h_{12} & \ldots & h_{1\,n-1} & h_{1n} \\ h_{21} & h_{22} & \ldots & h_{2\,n-1} & h_{2n} \\ 0 & h_{32} & \ldots & h_{3\,n-1} & h_{3n} \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \ldots & 0 & h_{n\,n-1} & h_{nn} \end{pmatrix}$$

die gewünschte Hessenberg-Matrix, die aufgrund der Orthogonalität der Householder-Matrizen $H_i$ eine orthogonale Transformation von $A$ ist. Es gilt

$$H = Q^TAQ \quad \text{mit} \quad Q = H_1H_2\ldots H_{n-2},\ Q^T = H_{n-2}H_{n-3}\ldots H_1 \ .$$

$H$ ist ähnlich zu $A$ und deshalb haben $H$ und $A$ die gleichen Eigenwerte.

**Beispiel 9.18.** Für die Transformation der Matrix

$$A = \begin{pmatrix} 2 & 3 & 4 \\ 3 & 2 & 3 \\ 4 & 1 & 6 \end{pmatrix}$$

ergibt sich mit $\vec{u}_1 = (0, 3+5, 4)^T$ die Householder-Matrix

$$H_1 = E - 2\frac{\vec{u}_1 \vec{u}_1^T}{\langle \vec{u}_1, \vec{u}_1 \rangle} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -\frac{3}{5} & -\frac{4}{5} \\ 0 & -\frac{4}{5} & \frac{3}{5} \end{pmatrix} .$$

Weiter gilt

$$G = H_1 A = \begin{pmatrix} 2 & 3 & 4 \\ -5 & -2 & -\frac{33}{5} \\ 0 & -1 & \frac{6}{5} \end{pmatrix} \quad \text{und} \quad H = H_1 A H_1 = \begin{pmatrix} 2 & -5 & 0 \\ -5 & \frac{162}{25} & -\frac{59}{25} \\ 0 & -\frac{9}{25} & \frac{38}{25} \end{pmatrix} .$$

$H = H_1 A H_1 = H_1^T A H_1$ ist offensichtlich eine Hessenberg-Matrix und eine orthogonale Transformation von $A$.

Fordert man von der zu transformierenden Matrix $A$ die Symmetrie, dann führt der eben dargelegte Algorithmus zur Transformation auf eine symmetrische Hessenberg-Matrix, die folglich eine symmetrische Tridiagonal-Matrix ist.

### 9.3.2 Newton-Verfahren zur Berechnung von Eigenwerten von Hessenberg-Matrizen

Das charakteristische Polynom $\chi(\mu)$ einer Hessenbergmatrix und die zugehörige Ableitung $\chi'(\mu)$ lassen sich jeweils über die Auflösung spezieller gestaffelter linearer Gleichungssysteme berechnen. Dazu betrachten wir den

**Satz 9.19.** *Sei $H = (h_{ij}) \in \mathbb{R}^{n \times n}$ eine Hessenbergmatrix mit $h_{i\,i+1} \neq 0$ für $i = 1 \ldots, n-1$ und charakteristischem Polynom $\chi(\mu) = det\,(H - \mu E)$, $\mu \in \mathbb{R}$. Im Folgenden sei $\mu \in \mathbb{R}$ fest gewählt und kein Eigenwert von $H$, und es bezeichne $\vec{x} = \vec{x}(\mu) = (x_j(\mu)) \in \mathbb{R}^n$ den eindeutig bestimmten Vektor mit*

$$(H - \mu E)\vec{x} = \vec{e}_1 \, , \tag{9.7}$$

*mit $\vec{e}_1 = (1, 0, \ldots, 0)^T \in \mathbb{R}^n$. Dann gelten die folgenden Darstellungen*

$$\chi(\mu) = \frac{(-1)^{n-1} h_{21} h_{32} \cdots h_{n\,n-1}}{x_n(\mu)} \, , \quad \frac{\chi(\mu)}{\chi'(\mu)} = \frac{1}{x_n(\mu)} \frac{d}{d\mu} \left( \frac{1}{x_n(\mu)} \right) . \tag{9.8}$$

*Beweis.* Die Anwendung der Cramerschen Regel auf die Gleichung (9.7) ergibt die erste Aussage in (9.8),

$$
\begin{aligned}
x_n &= \det\left(\begin{bmatrix} h_{11}-\mu & h_{12} & \cdots & h_{1\,n-1} & 1 \\ h_{21} & h_{22}-\mu & & \vdots & 0 \\ & h_{32} & \ddots & \vdots & \vdots \\ & & \ddots & h_{n-1\,n-1}-\mu & \vdots \\ & & & h_{n\,n-1} & 0 \end{bmatrix}\right)/\chi(\mu) \\
&= (-1)^{n-1}\underbrace{\det\left(\begin{bmatrix} h_{21} & h_{22}-\mu & & \vdots \\ & h_{32} & \ddots & \vdots \\ & & \ddots & h_{n-1\,n-1}-\mu \\ & & & h_{n\,n-1} \end{bmatrix}\right)}_{=h_{21}h_{32}\cdots h_{n\,n-1}}/\chi(\mu)\,,
\end{aligned}
$$

wobei die Determinate durch die Entwicklung nach der letzten Spalte berechnet wurde. Damit wurde die erste Aussage von (9.8) gezeigt. Eine anschließende Differentiation ergibt die zweite Aussage. □

**Bemerkung 9.20.** Die Forderung $h_{i\,i+1} \neq 0$ im letzten Satz ist keine wirkliche Einschränkung, da anderenfalls die Hessenbergmatrix in Teilmatrizen zerfällt, die ebenfalls Hessenbergmatrizen sind und dann ebenso behandelt werden können wie die Matrix $H$ im Satz.

**Satz 9.21.** *Mit den Bezeichnungen aus Satz 9.19 erhält man die Werte* $1/x_n(\mu)$ *und* $\frac{d}{d\mu}(\frac{1}{x_n(\mu)})$ *aus den folgenden (durch Umformung und Differentiation von* (9.7) *entstandenen) gestaffelten linearen Gleichungssystemen*

$$
\left.\begin{array}{lllllllllll}
(h_{11}-\mu)v_1 & + & h_{12}v_2 & + & \cdots & + & h_{1\,n-1}v_{n-1} & + & h_{1n} & = & \frac{1}{x_n(\mu)} \\
h_{21}v_1 & + & (h_{22}-\mu)v_2 & + & \cdots & + & h_{2\,n-1}v_{n-1} & + & h_{2n} & = & 0 \\
 & & \ddots & & \ddots & & \vdots & & \vdots & & \\
 & & & & h_{n-1\,n-2}v_{n-2} & - & (h_{n-1\,n-1}-\mu)v_{n-1} & + & h_{n-1\,n} & = & 0 \\
 & & & & & & h_{n\,n-1}v_{n-1} & + & h_{nn}-\mu & = & 0
\end{array}\right\} \quad (9.9)
$$

*beziehungsweise*

$$
\left.\begin{array}{lllllllllll}
(h_{11}-\mu)z_1 & + & h_{12}z_2 & + & \cdots & + & h_{1\,n-1}z_{n-1} & - & v_1 & = & \frac{d}{d\mu}\frac{1}{x_n(\mu)} \\
h_{21}z_1 & + & (h_{22}-\mu)z_2 & + & \cdots & + & h_{2\,n-1}z_{n-1} & - & v_2 & = & 0 \\
 & & \ddots & & \ddots & & \vdots & & \vdots & & \\
 & & & & h_{n-1\,n-2}z_{n-2} & - & (h_{n-1\,n-1}-\mu)z_{n-1} & - & v_{n-1} & = & 0 \\
 & & & & & & h_{n\,n-1}z_{n-1} & - & 1 & = & 0
\end{array}\right\} \quad (9.10)
$$

*die man rekursiv nach den Unbekannten* $v_{n-1},\ldots,v_1$, $1/x_n(\mu)$ *beziehungsweise* $z_{n-1},\ldots,z_1$, $\frac{d}{d\mu}\frac{1}{x_n(\mu)}$ *auflöst.*

*Beweis.* Die Aussage (9.9) erhält man (für $v_j = x_j(\mu)/x_n(\mu)$), indem man die einzelnen Zeilen des Gleichungssystems (9.7) durch $x_n(\mu)$ dividiert. Die Differentiation der Gleichungen in (9.9) nach $\mu$ liefert für $z_j = (\frac{dv_j}{d\mu})(\mu)$ unmittelbar (9.10). □

### 9.3.3 Das Newtonverfahren für tridiagonale Matrizen

Die Transformation einer symmetrischen Matrix auf Hessenbergform ergibt eine tridiagonale Matrix. Deshalb ist es sinnvoll, das Newtonverfahren für tridiagonale Matrizen betrachten, denn $\chi(\mu) = \det(H - \mu E)$ und $\chi'(\mu)$ lassen sich dann auf einfache Weise rekursiv berechnen.

**Lemma 9.22.** *Zu gegebenen Zahlen $\delta_1, \ldots, \delta_n \in \mathbb{R}$ und $\gamma_2, \ldots, \gamma_n \in \mathbb{R}$ gelten für die charakteristischen Polynome*

$$\chi_k(\mu) = det\,(J_k - \mu E)\ , \quad J_k = \begin{bmatrix} \delta_1 & \gamma_2 & & \\ \gamma_2 & \ddots & \ddots & \\ & \ddots & \ddots & \gamma_k \\ & & \gamma_k & \delta_k \end{bmatrix}\ , \quad k = 1, \ldots, n,$$

*die folgenden Rekursionsformeln*

$$\left.\begin{array}{rcl} \chi_1(\mu) & = & \delta_1 - \mu\,, \\ \chi_k(\mu) & = & (\delta_k - \mu)\chi_{k-1}(\mu) - \gamma_k^2 \chi_{k-2}(\mu)\ ,\ k = 2, \ldots, n, \end{array}\right\} \tag{9.11}$$

*mit der Notation $\chi_0(\mu) := 1$. Für die Ableitungen gelten*

$$\begin{array}{rcl} \chi_1'(\mu) & = & -1\ , \\ \chi_k'(\mu) & = & -\chi_{k-1} + (\delta_k - \mu)\chi_{k-1}'(\mu) - \gamma_k^2 \chi_{k-2}'(\mu)\ , k = 2, \ldots, n\ . \end{array}$$

*Beweis.* Die Darstellung für $\chi_1$ ergibt sich unmittelbar, und für $\chi_2$ ist

$$\chi_2(\mu) = \det\left(\begin{bmatrix} \delta_1 - \mu & \gamma_2 \\ \gamma_2 & \delta_2 - \mu \end{bmatrix}\right) = \underbrace{(\delta_1 - \mu)}_{=\chi_1(\mu)}(\delta_2 - \mu) - \gamma_2^2\ ,$$

was die behauptete Darstellung von $\chi_2$ ist. Für $k \geq 3$ erhält man durch

Entwicklung der Determinate nach der letzten Spalte

$$\chi_k(\mu) = \det\left(\begin{bmatrix} \delta_1-\mu & \gamma_2 & & & & \\ \gamma_2 & \ddots & \ddots & & & \\ & \ddots & \delta_{k-2}-\mu & \gamma_{k-1} & & \\ & & \gamma_{k-1} & \delta_{k-1}-\mu & \gamma_k & \\ & & & \gamma_k & \delta_k-\mu \end{bmatrix}\right) \tag{9.12}$$

$$= (\delta_k-\mu)\chi_{k-1}(\mu) - \gamma_k \underbrace{\det\left(\begin{bmatrix} \delta_1-\mu & \gamma_2 & & & & \\ \gamma_2 & \ddots & \ddots & & & \\ & \ddots & \delta_{k-3}-\mu & \gamma_{k-2} & & \\ & & \gamma_{k-2} & \delta_{k-2}-\mu & \gamma_{k-1} & \\ & & & 0 & \gamma_k \end{bmatrix}\right)}_{=\gamma_k\chi_{k-2}(\mu)}, \tag{9.13}$$

womit das Lemma bewiesen wäre. □

Mit den Ergebnissen zur Bestimmung des charakteristischen Polynoms $\chi(\mu)$ und des Quotienten $\frac{\chi(\mu)}{\chi'(\mu)}$ und Informationen zur Lage von Eigenwerten (z.B. nach dem Satz von Gerschgorin), kann man mit dem Newtonverfahren Eigenwerte berechnen.

### 9.3.4 Jacobi-Verfahren zur Eigenwertberechnung

Im Unterschied zum Newtonverfahren geht es beim Jacobi-Verfahren darum, durch die sukzessive Konstruktion von zu $A$ ähnlichen Matrizen $A^{(k)}$ mit Reduktion der Nichtdiagonalelemente die Eigenwerte durch die Diagonaleinträge von $A^{(k)}$ zu approximieren.

#### Approximation der Eigenwerte durch Diagonaleinträge

Um zu verabreden, was unter Konvergenz eines solchen Verfahrens zu verstehen ist, braucht man ein Maß zur Größe des Nichtdiagonalteils einer Matrix.

**Definition 9.23.** *Für eine Matrix $B = (b_{ij}) \in \mathbb{R}^{n\times n}$ ist die Zahl $S(B) \in \mathbb{R}_+$ folgendermaßen erklärt,*

$$S(B) := \sum_{i,j=1, i\neq j}^{n} b_{ij}^2 \,. \tag{9.14}$$

Offensichtlich gilt für $S(B)$ mit der Frobeniusnorm $||\cdot||_F$

$$S(B) := ||B||_F^2 - \sum_{j=1}^{n} b_{jj}^2 = ||B-D||_F^2 \,, \quad \text{mit} \quad D := \operatorname{diag}(b_{11},\ldots,b_{nn}) \,. \tag{9.15}$$

Ist $S(B)$ klein, dann stellen die Diagonalelemente Approximationen für die Eigenwerte dar. Es gilt der

**Satz 9.24.** *Seien $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ die Eigenwerte der symmetrischen Matrix $B = (b_{ij}) \in \mathbb{R}^{n\times n}$, und seien $b_{k_1k_1} \geq b_{k_2k_2} \geq \cdots \geq b_{k_nk_n}$ die der Größe nach geordneten Diagonalelemente von $B$. Dann gilt*

$$|b_{k_jk_j} - \lambda_j| \leq \sqrt{S(B)}\,, \quad j = 1, \ldots, n\,.$$

*Beweis.* Mit $D := \operatorname{diag}(b_{11}, \ldots, b_{nn})$ erhält man

$$\max_{j=1,\ldots,n} |b_{k_jk_j} - \lambda_j| \leq ||B - D||_2 \leq ||B - D||_F = \sqrt{S(B)}\,,$$

wobei erstens ein Störungsresultat aus Satz 9.16 benutzt wurde, und zweitens $||\cdot||_2 \leq ||\cdot||_F$ verwendet wurde. □

### Givensrotation zur Reduktion der Nichtdiagonaleinträge

Das Verfahren von Jacobi zur approximativen Bestimmung der Eigenwerte symmetrischer Matrizen $A \in \mathbb{R}^{n\times n}$ durch die Konstruktion von $A^{(k)}$ mit $S(A) = S(A^{(1)}) > S(A^{(2)}) > \ldots$ beruht auf der sogenannten Givensrotation. Es werden zu $A$ ähnliche Matrizen

$$A^{(k+1)} = S_k^{-1} A^{(k)} S_k\,, \quad k = 1, 2, \ldots \quad \text{mit } A = A^{(1)}$$

konstruiert, wobei die einzelnen Ähnlichkeitstransformationen von der allgemeinen Form

$$\hat{B} := \Omega_{pq}^{-1} B \Omega_{pq}\,, \quad \Omega_{pq} = \begin{bmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & c & & & -s & & & \\ & & & & 1 & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & 1 & & & \\ & & & s & & & & c & & \\ & & & & & & & & 1 & \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{bmatrix} \in \mathbb{R}^{n\times n} \tag{9.16}$$

sind mit einer symmetrischen Matrix $B \in \mathbb{R}^{n\times n}$ und mit speziell zu wählenden Indizes $p \neq q$ und reellen Zahlen

$$c,\, s \in \mathbb{R}\,, \quad c^2 + s^2 = 1\,. \tag{9.17}$$

Die Spalte mit den Zahlen $c$ und $s$ ist die $p$-te Spalte, die Spalte mit den Zahlen $-s$ und $c$ ist die $q$-te Spalte, woraus die entsprechenden Zeilen folgen. Ausgehend von $B = (b_{ij})$ erhält man durch die Transformation

$$\begin{aligned}
\hat{b}_{pp} &= c^2 b_{pp} + 2csb_{pq} + s^2 b_{qq} \ , && (9.18)\\
\hat{b}_{qq} &= s^2 b_{pp} - 2csb_{pq} + c^2 b_{qq} \ , && (9.19)\\
\hat{b}_{pq} &= \hat{b}_{qp} = cs(b_{qq} - b_{pp}) + (c^2 - s^2) b_{pq} \ , && (9.20)\\
\hat{b}_{ij} &= b_{ij} \ , \quad i,j \notin \{p,q\} \ . && (9.21)
\end{aligned}$$

Weiter gilt für die Einträge der $p$-ten und $q$-ten Spalten und Zeilen

$$\hat{b}_{kp} = \hat{b}_{pk} = cb_{kp} + sb_{kq} \ , \quad \hat{b}_{kq} = \hat{b}_{qk} = -sb_{kp} + cb_{kq} \ , \quad \text{für } \ k \notin \{p,q\} \ . \tag{9.22}$$

Bevor der Zusammenhang zwischen $S(\hat{B})$ und $S(B)$ hergestellt wird, soll ein Hilfsresultat hergeleitet werden.

**Lemma 9.25.** *Für jede Matrix $B \in \mathbb{R}^{n\times n}$ und jede orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$ gilt*

$$||Q^{-1}BQ||_F = ||B||_F \ .$$

*Beweis.* Unter der Spur einer Matrix $A$ verstehen wir $\text{spur}\,(A) = \sum_{j=1}^{n} a_{jj}$. Es gelten nun die elementaren Identitäten

$$||A||_F = \text{spur}\,(A^T A) \ , \quad \text{spur}\,(ST) = \text{spur}\,(TS) \quad \text{für alle } \ A,\, S,\, T \in \mathbb{R}^{n\times n} \ ,$$

woraus die Aussage des Lemmas folgt. □

Für den Zusammenhang zwischen $S(\hat{B})$ und $S(B)$ gilt der

**Satz 9.26.** *Für eine symmetrische Matrix $B = (b_{ij}) \in \mathbb{R}^{n\times n}$ gilt mit den Beziehungen aus* (9.16)

$$S(\hat{B}) = S(B) - 2(b_{pq}^2 - \hat{b}_{pq}^2) \ .$$

*Beweis.* Man rechnet

$$S(\hat{B}) = ||\hat{B}||_F^2 - \sum_{j=1}^{n} \hat{b}_{jj}^2 = \underbrace{(||B||_F^2 - \sum_{j=1}^{n} b_{jj}^2)}_{=S(B)} + b_{pp}^2 + b_{qq}^2 - \hat{b}_{pp}^2 - \hat{b}_{qq}^2 \tag{9.23}$$

aus. Die letzten 4 Summanden in (9.23) kann man in der Form

$$\underbrace{\begin{bmatrix} \hat{b}_{pp} & \hat{b}_{pq} \\ \hat{b}_{pq} & \hat{b}_{qq} \end{bmatrix}}_{=:\hat{b}} = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \underbrace{\begin{bmatrix} b_{pp} & b_{pq} \\ b_{pq} & b_{qq} \end{bmatrix}}_{=:b} \begin{bmatrix} c & -s \\ s & c \end{bmatrix}$$

darstellen. Die Matrizen $\hat{b}$ und $b \in \mathbb{R}^{2\times 2}$ sind orthogonal ähnlich zueinander, und damit folgt aus Lemma 9.25

$$\underbrace{\hat{b}_{pp}^2 + \hat{b}_{qq}^2 + 2\hat{b}_{pq}^2}_{=||\hat{b}||_F^2} = \underbrace{b_{pp}^2 + b_{qq}^2 + 2b_{pq}^2}_{=||b||_F^2} , \tag{9.24}$$

und die Identitäten (9.23) und (9.24) ergeben die Behauptung. □

Mit Satz 9.26 wird offensichtlich, dass bei festem Index $(p, q)$ im Fall $\hat{b}_{pq} = 0$ die Zahl $S(\hat{B})$ die größtmögliche Verringerung gegenüber $S(B)$ erfährt.

**Korollar 9.1.** *Wählt man in* (9.16) *die Zahlen* $c$ *und* $s$ *so, dass* $\hat{b}_{pq} = 0$ *erfüllt ist, dann gilt*

$$S(\hat{B}) = S(B) - 2b_{pq}^2 .$$

**Satz 9.27.** *In* (9.16) *erhält man den Eintrag* $\hat{b}_{pq} = \hat{b}_{qp} = 0$ *durch die Wahl der Zahlen* $c$ *und* $s$ *(o.B.d.A. sei* $b_{pq} \neq 0$*)*

$$c = \sqrt{\frac{1+C}{2}} , \quad s = sign\,(b_{pq})\sqrt{\frac{1-C}{2}} \quad mit \quad C = \frac{b_{pp} - b_{qq}}{\sqrt{(b_{pp} - b_{qq})^2 + 4b_{pq}^2}} . \tag{9.25}$$

*Beweis.* Mit den Beziehungen (9.20) folgt

$$\begin{aligned} \hat{b}_{pq} &= \operatorname{sign}(b_{pq})\sqrt{\frac{1-C^2}{4}}(b_{qq} - b_{pp}) + Cb_{pq} \\ &= \frac{\operatorname{sign}(b_{pq})|b_{pq}|(b_{qq} - b_{pp})}{\sqrt{(b_{pp} - b_{qq})^2 + 4b_{pq}^2}} + \frac{b_{pp} - b_{qq}}{\sqrt{(b_{pp} - b_{qq})^2 + 4b_{pq}^2}} b_{pq} = 0 , \end{aligned}$$

wobei der Schritt von der ersten zur zweiten Zeile aus

$$\sqrt{\frac{1-C^2}{4}} = \frac{1}{2}\sqrt{\frac{(b_{pp} - b_{qq})^2 + 4b_{pq}^2 - (b_{pp} - b_{qq})^2}{(b_{pp} - b_{qq})^2 + 4b_{pq}^2}} = \frac{|b_{pq}|}{\sqrt{(b_{pp} - b_{qq})^2 + 4b_{pq}^2}}$$

folgt. □

Das Korollar 9.1 und der folgende Satz liefern einen Hinweis zur jeweiligen Wahl der Indizes $p$ und $q$.

**Satz 9.28.** *Für Indizes $(p,q)$ mit $p \neq q$ sei*

$$|b_{pq}| \geq |b_{ij}| \quad \textit{für } i,j = 1,\ldots,n,\ i \neq j\ , \tag{9.26}$$

*erfüllt. Mit den Bezeichnungen aus* (9.16) *und $c$ und $s$ aus Satz 9.27 gilt die Abschätzung*

$$S(\hat{B}) \leq (1-\eta)S(B)\ , \quad \textit{mit } \eta := \frac{2}{n(n-1)}\ .$$

*Beweis.* Wegen (9.26) gilt die Abschätzung

$$S(B) = \sum_{i,j=1,\ldots,n,i\neq j}^{n} b_{ij}^2 \leq n(n-1)b_{pq}^2\ ,$$

da die Anzahl der Nichtdiagonalelemente gleich $n(n-1)$ ist. Die Aussage des Satzes folgt unter Nutzung des Korollars 9.1. □

**Bemerkung 9.29.** Nach Satz 9.28 gilt für die Messgrößen $S(A^{(k)})$ des Jacobiverfahrens

$$S(A^{(k)}) \leq (1-\eta)^k S(A)\ , \quad \text{für } k = 1,2,\ldots \quad (\eta = \frac{2}{n(n-1)},\ A = A^{(1)})\ .$$

Bei Vorgabe einer Genauigkeit $\epsilon > 0$ für $S(A^{(k)})$ ergibt sich

$$S(A^{(k)}) \leq (1-\eta)^k S(A) < \epsilon \Longleftrightarrow k \geq 2\frac{\log(\sqrt{S(A)}/\epsilon)}{-\log(1-\eta)} \approx n^2 \log((\sqrt{S(A)}/\epsilon)$$

für die durchzuführenden Givensrotationen bei jeweiliger Wahl des betragsgrößten Nichtdiagonalelements zur Ermittlung vom Indexpaar $(p,q)$.

### 9.3.5 Von-Mises-Vektoriteration (zur Information)

24. Vorlesung am 23.01.2012

Bei vielen angewandten Aufgabenstellungen ist der betragsgrößte Eigenwert von besonderer Bedeutung. Bei Schwingungsproblemen ist oft die Grundschwingung von Interesse und für deren Berechnung benötigt man den betragsgrößten Eigenwert. Für den Fall, dass die Matrix $A$ Eigenwerte mit der Eigenschaft

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_N| \tag{9.27}$$

besitzt, kann man ausgehend von einem geeigneten Startvektor $\vec{u}_0$ mit der Iteration

$$\vec{u}_1 = A\vec{u}_0,\ \vec{u}_2 = A\vec{u}_1,\ \ldots, \vec{u}_{k+1} = A\vec{u}_k, \ldots \tag{9.28}$$

den betragsgrößten Eigenwert und den dazugehörigen Eigenvektor berechnen. Betrachten wir als Startvektor

$$\vec{u}_0 = \vec{q}_1 + \vec{q}_2 + \cdots + \vec{q}_N\ ,$$

wobei $\vec{q}_1, \ldots, \vec{q}_N$ die Eigenvektorbasis einer als diagonalisierbar vorausgesetzten Matrix $A$ sind. Mit $A\vec{q}_k = \lambda_k\vec{q}_k$ erhält man mit der Iteration (9.28)

$$\vec{u}_k = A\vec{u}_{k-1} = A^k\vec{u}_0 = \lambda_1^k\vec{q}_1 + \cdots + \lambda_N^k\vec{q}_n \tag{9.29}$$

und bei der Iteration setzt sich die Vektorkomponente mit dem betragsgößten Eigenwert durch, so dass die Iteration in gewisser Weise gegen den Eigenvektor $\vec{q}_1$ strebt. Multipliziert man (9.29) mit einem Testvektor $\vec{z}$, von dem $\langle\vec{z}, \vec{q}_1\rangle \neq 0$ gefordert wird, dann erhält man

$$\langle\vec{u}_k, \vec{z}\rangle \approx \lambda_1\langle\vec{u}_{k-1}, \vec{z}\rangle$$

für genügend große $k$ und es gilt

$$\lambda_1 = \lim_{k\to\infty} \frac{\langle\vec{u}_k, \vec{z}\rangle}{\langle\vec{u}_{k-1}, \vec{z}\rangle}\ ,$$

wobei wir die gesicherte Existenz des Grenzwerts nicht zeigen. Ist $\vec{q}_1$ als Eigenvektor mit einer positiven ersten von null verschiedenen Komponente zum betragsgrößten Eigenwert $\lambda_1$ normiert, dann konvergiert die Folge

$$\vec{v}_k := \zeta_k\frac{\vec{u}_k}{||\vec{u}_k||} \tag{9.30}$$

gegen $\vec{q}_1$, wobei $\zeta_k \in \{+1, -1\}$ so zu wählen ist, dass die erste von null verschiedene Komponente von $\vec{v}_k$ positiv ist. Die durchgeführten Betrachtungen können wir zusammenfassen.

**Satz 9.30.** *(Von-Mises-Vektoriteration)*
*Sei $A$ eine diagonalisierbare $(N\times N)$-Matrix, deren Eigenwerte die Bedingung (9.27) erfüllen. $\vec{q}_j$ seien die Eigenvektoren zu $\lambda_j$. Seien $\vec{u}_k$ und $\vec{v}_k$ durch (9.29) bzw. (9.30) erklärt und gelte $\langle \vec{u}_0, \vec{q}_1\rangle \neq 0$, $\langle \vec{z}, \vec{q}_1\rangle \neq 0$ für die Vektoren $\vec{z}, \vec{u}_0$. Dann konvergiert die Folge $\vec{v}_k$ gegen den Eigenvektor $\vec{q}_1$ und der betragsgrößte Eigenwert $\lambda_1$ ergibt sich als Grenzwert*

$$\lambda_1 = \lim_{k\to\infty} \frac{\langle \vec{u}_k, \vec{z}\rangle}{\langle \vec{u}_{k-1}, \vec{z}\rangle} = \lim_{k\to\infty} \frac{\langle \vec{v}_k, \vec{z}\rangle}{\langle \vec{v}_{k-1}, \vec{z}\rangle} \ . \tag{9.31}$$

*Für die Konvergenzgeschwindigkeit gilt*

$$|\frac{\langle \vec{u}_{k+1}, \vec{z}\rangle}{\langle \vec{u}_k, \vec{z}\rangle} - \lambda_1| \leq K|\frac{\lambda_2}{\lambda_1}|^k \ , \tag{9.32}$$

*wobei die Konstante $K$ von der Wahl von $\vec{z}, \vec{u}_0$ abhängt.*

Zum Satz 9.30 ist anzumerken, dass man auch im Fall

$$\lambda_1 = \cdots = \lambda_r, \quad |\lambda_1| = \cdots = |\lambda_r| > |\lambda_{r+1}| \geq \cdots \geq |\lambda_N|, \ r > 1$$

mit der Von-Mises-Iteration (9.29), (9.30), (9.31) den mehrfachen Eigenwert $\lambda_1$ bestimmen kann. Allerdings konvergiert die Folge (9.30) nur gegen irgendeinen Eigenvektor aus dem Unterraum der Lösungen des linearen Gleichungssystems $(A - \lambda_1 E)\vec{v} = \vec{0}$. Eventuelle weitere Eigenvektoren zum mehrfachen Eigenwert $\lambda_1$ muss man dann auf anderem Weg, z.B. durch die Bestimmung weiterer Lösungen von $(A - \lambda_1 E)\vec{v} = \vec{0}$, berechnen.
Nach der Bestimmung von $\lambda_1$ weiß man, dass für eine symmetrische Matrix $A$ alle Eigenwerte auf jeden Fall im Intervall $[a, b] := [-|\lambda_1|, |\lambda_1|]$ liegen, da sie reell sind. Evtl. kann man das Intervall $[a, b]$ durch die Betrachtung der Gerschgorin-Kreise noch verkleinern.
Mit der folgenden Überlegung kann man unter Umständen Eigenwerte von $A$ schneller bestimmen als mit der Von-Mises-Iteration nach Satz 9.30. Ist $\lambda$ ein Eigenwert von $A$ und $\vec{u}$ ein zu $\lambda$ gehörender Eigenvektor von $A$, dann ist für $\mu \neq \lambda$ wegen

$$A\vec{u} = \lambda\vec{u} \iff (A - \mu E)\vec{u} = (\lambda - \mu)\vec{u} \iff (A - \mu E)^{-1}\vec{u} = \frac{1}{\lambda - \mu}\vec{u}$$

die Zahl $\frac{1}{\lambda-\mu}$ ein Eigenwert von $(A - \mu E)^{-1}$. Wendet man den Satz 9.30 auf das Eigenwertproblem der Matrix $(A - \mu E)^{-1}$ an, dann ergibt sich mit dem folgenden Satz eine effiziente Methode zur Eigenwert- und Eigenvektorbestimmung.

**Satz 9.31.** *(inverse Von-Mises-Vektoriteration)*
*Sei $A$ eine Matrix vom Typ $N \times N$ mit den Eigenwerten $\lambda_1, \ldots, \lambda_N$ und sei $\mu \in \mathbb{C}$ eine komplexe Zahl ungleich allen Eigenwerten von $A$, so dass die Matrix $A$ einen Eigenwert hat, der näher bei $\mu$ als bei allen anderen Eigenwerten liegt, d.h.*

$$0 < |\lambda_1 - \mu| < |\lambda_2 - \mu| \leq \cdots \leq |\lambda_N - \mu|$$

*gilt ($\lambda_1$ ist der Eigenwert, der $\mu$ am nächsten liegt). Mit der Iterationsfolge*

$$\vec{u}_k := (A - \mu E)^{-1} \vec{u}_{k-1} \quad (k = 1, 2, \ldots) \tag{9.33}$$

*gilt*

$$\lim_{k\to\infty} \frac{\langle \vec{u}_k, \vec{z}\rangle}{\langle \vec{u}_{k-1}, \vec{z}\rangle} = \frac{1}{\lambda_1 - \mu} \quad \Longleftrightarrow \quad \lambda_1 = \lim_{k\to\infty} \frac{\langle \vec{u}_{k-1}, \vec{z}\rangle}{\langle \vec{u}_k, \vec{z}\rangle} + \mu \ ,$$

*wobei $\langle \vec{u}_0, \vec{q}_\mu\rangle \neq 0$, $\langle \vec{z}, \vec{q}_\mu\rangle \neq 0$ für den Startvektor $\vec{u}_0$ und den Testvektor $\vec{z}$ mit $\vec{q}_\mu$ als dem zu $\frac{1}{\lambda_1 - \mu}$ gehörenden Eigenvektor der Matrix $(A - \mu E)^{-1}$ gelten muss. Die normalisierten Vektoren $\vec{v}_k = \frac{\vec{u}_k}{||\vec{u}_k||}$ konvergieren gegen den Eigenvektor $\vec{q}_\mu$. Die Iteration (9.33) heißt inverse Von-Mises-Iteration. Für die Konvergenzgeschwindigkeit gilt*

$$\left|\frac{\langle \vec{u}_{k+1}, \vec{z}\rangle}{\langle \vec{u}_k, \vec{z}\rangle} - \frac{1}{\lambda_1 - \mu}\right| \leq K \left|\frac{1/(\lambda_2 - \mu)}{1/(\lambda_1 - \mu)}\right|^k = K \left|\frac{\lambda_1 - \mu}{\lambda_2 - \mu}\right|^k \ .$$

Der Satz 9.31 ist in zweierlei Hinsicht von Bedeutung. Zum einen kann man durch eine günstige Wahl von $\mu$ in der Nähe eines Eigenwertes $\lambda_1$ die Konvergenzgeschwindigkeit der inversen Von-Mises-Iteration groß machen und schnell zu diesem Eigenwert gelangen. Zweitens kann man bei Kenntnis des Intervalls $[\lambda_{\min}, \lambda_{\max}]$ durch die Wahl von $\mu = \frac{\lambda_{\min} + \lambda_{\max}}{2}$ und die Berechnung des Eigenwertes $\lambda_\mu$ von $A$, der $\mu$ am nächsten liegt, mit

$$\mu_1 = \frac{\lambda_{\min} + \lambda_\mu}{2} \ , \quad \mu_2 = \frac{\lambda_\mu + \lambda_{\max}}{2}$$

die Iteration (9.33) für $\mu_1$ und $\mu_2$ durchführen. Die sukzessive Fortsetzung dieses Algorithmus liefert nach evtl. Aussortierung von Punkten, für die (9.33) nicht konvergiert, alle Eigenwerte von $A$. Bei der Wahl der Parameter $\mu$ kann man natürlich auch Informationen zur Lage der Eigenwerte aus dem Satz 9.7 nutzen.
Ein weiterer Weg, sämtliche von null verschiedenen Eigenwerte einer Matrix $A$ durch Von-Mises-Vektoriterations-Methoden zu bestimmen, ist mit Hilfe

der **Deflation** möglich. Kennt man einen Eigenwert $\lambda_1 \neq 0$ der symmetrischen Matrix $A$ und mit $\vec{x}_1$ den dazugehörenden Eigenvektor und bezeichnet die restlichen Eigenwerte von $A$ mit $\lambda_2, \ldots, \lambda_N$, dann hat die Matrix

$$\tilde{A} = (E - \frac{\vec{x}_1 \vec{x}_1^T}{\langle \vec{x}_1, \vec{x}_1 \rangle})A = A - \frac{\lambda_1}{\langle \vec{x}_1, \vec{x}_1 \rangle} \vec{x}_1 \vec{x}_1^T$$

die Eigenwerte $0, \lambda_2, \ldots, \lambda_N$. Außerdem ist jeder Eigenvektor von $A$ auch Eigenvektor von $\tilde{A}$ und umgekehrt. Mit der Deflation transformiert man den Eigenwert $\lambda_1$ auf 0.

**Beispiel 9.32.** Für die Matrix

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$$

findet man die Eigenwerte $\lambda_1 = 2, \lambda_2 = 2 - \sqrt{2}, \lambda_3 = 2 + \sqrt{2}$ mit den Eigenvektoren

$$\vec{x}_1 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \quad \vec{x}_2 = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{\sqrt{2}} \\ \frac{1}{2} \end{pmatrix}, \quad \vec{x}_3 = \begin{pmatrix} -\frac{1}{2} \\ \frac{1}{\sqrt{2}} \\ -\frac{1}{2} \end{pmatrix}.$$

Für $\tilde{A}$ ergibt sich

$$\tilde{A} = A - \frac{\lambda_1}{\langle \vec{x}_1, \vec{x}_1 \rangle} \vec{x}_1 \vec{x}_1^T = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 1 \end{pmatrix}$$

mit den Eigenwerten $0, \lambda_2 = 2 - \sqrt{2}, \lambda_3 = 2 + \sqrt{2}$ und den Eigenvektoren

$$\vec{x}_1 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \quad \vec{x}_2 = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{\sqrt{2}} \\ \frac{1}{2} \end{pmatrix}, \quad \vec{x}_3 = \begin{pmatrix} -\frac{1}{2} \\ \frac{1}{\sqrt{2}} \\ -\frac{1}{2} \end{pmatrix}.$$

Für den allgemeineren Fall der nicht notwendigerweise symmetrischen Matrix $A$ gilt der folgende

**Satz 9.33.** *(Deflation)*
*Sei $\vec{z} \neq \vec{0}$ ein beliebiger Vektor und es sei $\vec{x}_1$ mit $\langle \vec{x}_1, \vec{z} \rangle \neq 0$ ein Eigenvektor der Matrix $A$ zum Eigenwert $\lambda_1$. Dann liefert jeder weitere von $\vec{x}_1$ linear unabhängige Eigenvektor $\vec{x}$ von $A$ zum Eigenwert $\lambda$ mit*

$$\vec{y} = \vec{x} - \frac{\langle \vec{x}, \vec{z} \rangle}{\langle \vec{x}_1, \vec{z} \rangle} \vec{x}_1 \tag{9.34}$$

*einen Eigenvektor der Matrix*

$$\tilde{A} = (E - \frac{\vec{x}_1 \vec{z}^T}{\langle \vec{x}_1, \vec{z} \rangle})A$$

*zum gleichen Eigenwert $\lambda$. Der Eigenvektor $\vec{x}_1$ ist ebenfalls Eigenvektor der Matrix $\tilde{A}$ zum Eigenwert 0. Umgekehrt liefert jeder Eigenvektor $\vec{y}$ von $\tilde{A}$ zum Eigenwert $\lambda$ einen Eigenvektor*

$$\vec{x'} = (A - \lambda_1 E)\vec{y} = (\lambda - \lambda_1)\vec{y} + \frac{\langle A\vec{y}, \vec{z} \rangle}{\langle \vec{x}_1, \vec{z} \rangle}\vec{x}_1 \tag{9.35}$$

*von $A$ zum selben Eigenwert. Alle Eigenvektoren von $\tilde{A}$ zu nichtverschwindenden Eigenwerten stehen senkrecht auf $\vec{z}$.*

$\tilde{A}\vec{y} = \lambda\vec{y}$ und $A\vec{x'} = \lambda\vec{x'}$ rechnet man durch Einsetzen nach. Die Multiplikation von $\vec{z}^T A$ mit (9.34) ergibt

$$\vec{z}^T A\vec{y} = \langle A\vec{y}, \vec{z} \rangle = \langle A\vec{x}, \vec{z} \rangle - \lambda_1 \langle \vec{x}, \vec{z} \rangle \iff \langle A\vec{y}, \vec{z} \rangle = (\lambda - \lambda_1)\langle \vec{x}, \vec{z} \rangle$$

und Einsetzen von $\langle \vec{x}, \vec{z} \rangle = \frac{1}{\lambda - \lambda_1}\langle A\vec{y}, \vec{z} \rangle$ in (9.34) liefert (9.35) mit dem Eigenvektor $\vec{x'} = (\lambda - \lambda_1)\vec{x}$. Die skalare Multiplikation von $\tilde{A}\vec{y}$ mit $\vec{z}$ ergibt unter Nutzung von $\tilde{A}\vec{y} = \lambda\vec{y}$

$$\langle \tilde{A}\vec{y}, \vec{z} \rangle = \langle A\vec{y}, \vec{z} \rangle - \frac{\langle A\vec{y}, \vec{z} \rangle}{\langle \vec{x}_1, \vec{z} \rangle}\langle \vec{x}_1, \vec{z} \rangle = \langle A\vec{y}, \vec{z} \rangle - \langle A\vec{y}, \vec{z} \rangle = \lambda \langle \vec{y}, \vec{z} \rangle \ ,$$

woraus $\langle \vec{y}, \vec{z} \rangle$ für $\lambda \neq 0$ folgt. Damit ist der Satz 9.33 bewiesen.
Mit dem Satz 9.33, d.h., der sukzessiven Deflation, kann man also mit Von-Mises-Iterationen sämtliche Eigenwerte einer Matrix, beginnend mit dem betragsgrößten, und die dazugehörenden Eigenvektoren berechnen.

# Kapitel 10

# Wiederholung/Klausur/Prüfungsthemen

## 10.1 Klausurthemen

- Kondition von Problemen, Vektor- und Matrixnormen,
- Matrixfaktorisierungen, LR-, Cholesky- und QR-Zerlegungen,
- Orthogonale Matrizen, Householder-Transformationen, Gram-Schmidt-Orthogonalisierung,
- Bestapproximation (QR-Zerlegung und Normalgleichungssystem),
- Funktionsinterpolation, Polynominterpolation, Spline-Interpolation (und die basics der trigonometrischen Interpolation),
- Numerische Integration (Newton-Cotes, Gauß, Romberg),
- Iterative Lösung linearer und nichtlinearer Gleichungssysteme (Fixpunktiteration, Newton), CG- und GMRES-Verfahren sind kein Thema,
- Numerische Methoden zur Lösung von Anfangswertproblemen

  qualitative Grundlagen,

  Verfahrensfunktion, lokaler und globaler Diskretisierungsfehler,

  Konsistenz(ordnung),

  Einschrittverfahren, Interpretation von Butchertabellen,

  Ermittlung der max. Ordnung von mehrstuf. Einschrittverfahren,

  Grundlagen der linearen Mehrschrittverfahren,

  Nullstabilität, Gebiete der absoluten Stabilität



25. Vorlesung am 25.01.2012

## 10.2 Beispiel eines Konsistenznachweises

1) Beispielhaft soll hier nochmal ein Konsistenznachweis eines Einschrittverfahrens geführt werden, und zwar soll das Verfahren

$$y_{k+1} = y_k + h\Psi(t_k, y_k, y_{k+1}, h) \quad \text{mit} \quad \Psi(t_k, y_k, y_{k+1}, h) = \frac{1}{2}[f(t_k, y_k) + f(t_k + h, y_{k+1})]$$

untersucht werden (dabei setzen wir die in den Taylorentwicklungen benötigte Glattheit von $f$ und die eindeutige Lösbarkeit der impliziten Verfahrensgleichung voraus). Wir betrachten $\Psi$ als Funktion von $h$

$$\Psi(h) = \frac{1}{2}[f(t, y(t)) + f(t + h, y(t + h))]$$

und damit ist $\Psi(0) = y'(t)$. Für die Ableitungen von $\Psi$ finden wir

$$\Psi'(h) = \frac{1}{2}[f_t(t + h, y(t + h)) + f_y(t + h, y(t + h))y'(t + h)]$$

sowie

$$\Psi'(0) = \frac{1}{2}y''(t) \ ,$$

$$\Psi''(h) = \frac{1}{2}[f_{tt}(...) + f_{ty}(...)y'(t+h) + f_{yt}(...)y'(t+h) + f_{yy}(...)y'(t+h)^2 + f_y(...)y''(t+h)]$$

sowie

$$\Psi''(0) = \frac{1}{2}y'''(t) \ .$$

Für den lokalen Diskretisierungsfehler

$$d(h) = y(t + h) - y(t) - h\Psi(h)$$

finden wir mit den obigen Werten für $\Psi$ und deren Ableitungen an der Stelle 0

$$\begin{aligned}
d(0) &= 0 \ , \\
d'(h) &= y'(t + h) - h\Psi'(h) - \Psi(h) \ , \\
d'(0) &= y'(t) - y'(t) = 0 \ , \\
d''(h) &= y''(t + h) - 2\Psi'(h) - h\Psi(h) \ , \\
d''(0) &= y''(t) - 2\Psi'(0) = y''(t) - y''(t) = 0 \ , \\
d'''(h) &= y'''(t + h) - 3\Psi''(h) - h\Psi'''(h) \ , \\
d'''(0) &= y'''(t) - \frac{3}{2}y'''(t) \neq 0 \ .
\end{aligned}$$

Aus der Entwicklung

$$d(h) = d(0) + hd'(0) + \frac{h^2}{2}d''(0) + \frac{h^3}{6}d'''(\xi) = \frac{h^3}{6}d'''(\xi) = \mathcal{O}(h^3)$$

folgt die Ordnung $p = 2$ für das Verfahren.

2) Für das Verfahren mit der Butchertabelle

| | | |
|---|---|---|
| $a_1$ | $b_1$ | $a_1 - b_1$ |
| $a_2$ | $b_2$ | $a_2 - b_2$ |
| | $\frac{1}{2}$ | $\frac{1}{2}$ |

erkennt man mit den ersten beiden Zeilen der Tabelle, dass das Verfahren mindestens die Ordnung 1 hat. Mit der oben angewendeten Methode kann man für das Verfahren mit

$$\begin{aligned} k_i(h) &= f(t + a_i h, y + h(b_i k_1(h) + (a_i - b_i)k_2(k))) \quad i = 1, 2, \\ \Psi(h) &= \frac{1}{2}(k_1(h) + k_2(h)) \end{aligned}$$

durch die "geschickte" Wahl der Parameter (man muss die Koeffizienten $d^{(k)}(0)$ für $k = 0, 1, 2, ...$ durch die geeignete Parameterwahl zu Null machen) kann man die Ordnung $p = 4$ erreichen. Und zwar schafft man das mit

| | | |
|---|---|---|
| $\frac{1}{2} - \frac{1}{\sqrt{12}}$ | $\frac{1}{4}$ | $\frac{1}{4} - \frac{1}{\sqrt{12}}$ |
| $\frac{1}{2} + \frac{1}{\sqrt{12}}$ | $\frac{1}{4} + \frac{1}{\sqrt{12}}$ | $\frac{1}{4}$ |
| | $\frac{1}{2}$ | $\frac{1}{2}$ |

aber diese Rechnung wäre für die Klausur zu aufwändig.