Einführung in die numerische Mathematik

Steffen Suerbier

20. Oktober 2009

Zusammenfassung

Mitschriften basierend auf der Vorlesung Einführung in die numerische Mathematik, gehalten von Prof. Dr. Günter Bärwolff im SS2009 an der TU Berlin

Inhaltsverzeichnis

Kapitel 0

Vorwort

Diese Mitschriften entstanden im Sommersemester 2009 im Laufe der Veranstaltung "Einführung in die numerische Mathematik" von Prof. Dr. Günter Bärwolff. Für Angaben in dieser Mitschrift wird keine Garantie bezüglich Korrektheit und Vollständigkeit übernommen. Verwendete Literatur war:

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

Kapitel 1

Rechnerarithmetik

1. Vorlesung am 15.4.2009

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

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

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 Gleitpunktzahlsystemen wie folgt

Definition 1.1. *Zu gegebener Basis $b \geq 2$ und* ***Mantisse*** *$t \in \mathbb{N}$ 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 Gleitpunktzahlsystem 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 ± 127 darstellen kann. Statt F wird auch M (Maschinenzahlen) als Symbol genutzt, also z.B.

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

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

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 **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 denomalisierten 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 kleinste 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 zwsichen 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 $\text{eps} = \frac{1}{2}b^{-t+1}$ 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 Gleipunktzahlensystemen 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 Gleitpunkzahlensystem $F(b, t, e_{min}, e_{max})$ mit gerader Basis b ist die Funktion* $\mathrm{rd} : \{x \in \mathbb{R} : x_{min} \leq |x| \leq x_{max}\} \to \mathbb{R}$ *durch*

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

für $x = \sigma \cdot (\sum_{k=1}^{t} a_k b^{-k}) \cdot b^e$ erklärt. $\mathrm{rd}(x)$ *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* $\mathrm{rd}(x) \in F$ *und die Minimaleigenschaft*

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

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 $\mathrm{rd}(x)$ 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 $\mathrm{rd}(x) \in F$. □

Es wird nun die Abschätzung

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

gezeigt.

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

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

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

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

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* 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).

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

2. Vorlesung am 20.4.2009

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

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

oder

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

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*

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

wobei im Fall von (1.15) k *gleich* 1 *und im Fall von* (1.16) k *gleich* 2 *ist (*ϵ *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

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

also aus (1.14) □

1.6 Fehlerakkumulation

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

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

mit $\frac{|\Delta x|}{|x|}, \frac{|\Delta y|}{|y|} \leq \epsilon$ Zahlen aus einem Gleitpunktzahlensystem F. $\tilde{\circ}, \circ$ 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

$$(1 + \tau_x) \circ (1 + \tau_y)(1 + \alpha) = 1 + \beta$$

mit einem β mit der Eigenschaft $|\beta| \leq \frac{3\epsilon}{1-3\epsilon}$ gilt (Beweis: 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*

$$(x + \Delta x) \,\tilde{\circ}\, (y + \Delta y) = x \circ y + \eta \tag{1.18}$$

mit $\frac{|\eta|}{|x \circ y|} \leq \frac{3\epsilon}{1-4\epsilon}$.

Die Darstellung (1.18) zeigt, dass die Multiplikation bzw. Division verhältnismäßig gutartig mit einem kleinen relativen Fehler ist. Im Folgenden soll die Fehlerverstärkung bei der Hintereinanderausführung von Addition in einem gegebenen GPZS F betrachtet werden. $\tilde{+}$ und $\tilde{\sum}$ sollen für die Addition bzw. Summation von links nach rechts in F stehen.

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). Dann gilt

$$|\tilde{S_k} - S_k| \leq \underbrace{\left(\sum_{j=1}^{k}(1+\epsilon)^{k-j}(2|x_j| + |S_j|)\right)}_{=:M_k}\epsilon \quad \textit{für} \quad k = 1,..,n \tag{1.19}$$

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

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

Beweis. Vollständige Induktion über k
(1.19) gilt für $k = 1$, denn $|\tilde{S_1} - S_1| = |\Delta x_1| \leq 3|x_1|\epsilon$ wir nehmen jetzt an, dass (1.19) für ein $k \geq 1$ richtig ist. Mit der Notation

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

bezeichnet man mit einer gewissen Zahl $\tau_k \in \mathbb{R}, |\tau_k| \leq \epsilon$

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

und damit

$$|\Delta S_k| \leq (1+\epsilon)|\Delta S_{k-1}| + \epsilon(|S_k| + 2|x_k|) \tag{1.20}$$

Aus (1.20) und der Induktionsannahme folgt die Behauptung (1.19) des Theorems. □

Bemerkung 1.11. Der Faktor $(1+\epsilon)^{n-j}$ in der Abschätzung (1.19) 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 (1.19) 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!

1.7 Stabilität, Vorwärtsanalyse, Rückwärtsanalyse

3. Vorlesung am 22.4.2009

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

$$f : \mathbb{R}^n \to \mathbb{R}^m \tag{1.21}$$

beschrieben werden, wobei eine explizite Formel für f vorliegen soll. Zur Allgemeinheit bedeutet (1.21), dass ausgehend von n Eingangsdaten m Ergebnisse des Problems berechnet werden.
Der besseren Übersichtlichkeit halber betrachten wir später skalarwertige Probleme, d.h. $m = 1$.

Definition 1.12. *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

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

Bemerkung 1.13. $\dot{\leq}$ bedeutet dabei, dass die Relation in erster Näherung gilt. Die verwendeten Normen können z.B. euklidische Normen des $\mathbb{R}^n$ bzw $\mathbb{R}^m$ sein.
κ klein bedeutet grob ein gut konditioniertes Problem
κ gross ein schlecht Konditioniertes

Beispiel 1.14. f differenzierbar

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

Im Folgenden sollen Fehleranalysen durchgeführt werden. Dabei geht es uns um den Fehler, der durch den Übergang vom ursprünglichen Problem f zu einem numerischen Verfahren $\tilde{f}$ entsteht. Statt $f(x)$ wird eine Näherung berechnet.

1.8 Stabilitätskonzepte

1.8.1 Vorwärtsanalyse

Analyse der Vergrößerung von $f(E)$ zu $\tilde{f}(E)$ fragt nach der Stabilität im Sinne der Vorwärtsanalyse. Sie beinhaltet den Einfluss des Eingabefehler ($\tilde{f}$ anstelle von f).

Definition 1.15. *Ein Verfahren heißt stabil, wenn es eine Konstante $\sigma \in \mathbb{R}$ gibt, so dass gilt:*

$$\|f(\tilde{x}) - \tilde{f}(\tilde{x})\| \leq \kappa_{rel} \cdot \sigma \cdot \text{eps}$$

*(*eps *Maschinengenauigkeit). σ quantifiziert die Stabilität im Sinne der Vorwärtsanalyse.*

1.8.2 Rückwärtsanalyse

Geht zurück auf James Hardy Wilkinson. Die Idee besteht darin, ein fehlerbehaftetes Resultat $\tilde{y} = \tilde{f}(\tilde{x})$ durch $\tilde{y} = f(\hat{x}) = f(\tilde{x} + \Delta\tilde{x})$ darzustellen.
Die formale Definition des Rückwärtsfehlers eines Algorithmus $\tilde{f}$ für die fehlerbehafteten (gerundeten) Eingabedaten $\tilde{x}$ mit $\|\tilde{x}\| \neq 0$ lautet:

Definition 1.16 (Rückwärtsfehler)**.**

$$\epsilon_R(\tilde{x}) = \inf \left\{ \frac{\|\Delta\tilde{x}\|}{\|\tilde{x}\|} : \hat{x} = \tilde{x} + \Delta\tilde{x} \in D_f \wedge f(\tilde{x} + \Delta\tilde{x}) = \tilde{f}(\tilde{x}) \right\}$$

Mann nennt einen Algorithmus rückwärtsstabil, wenn der relative Rückwärtsfehler $\epsilon_R(\tilde{x})$ für alle $\tilde{x} \in D_{\tilde{f}}$ kleiner als der unvermeidbar relative Eingabefehler ist, d.h.

$$\epsilon_R(\tilde{x}) \leq \text{eps} \quad \forall \tilde{x} \in D_{\tilde{f}} \tag{1.22}$$

Bemerkung 1.17. Oft interessiert man sich auch nur dafür, ob der relative Rückwärtsfehler überhaupt beschränkt ist. Außerdem schwächt man für manche Anwendungen (1.22) zu

$$\epsilon_R(\tilde{x}) \leq C \cdot \text{eps} \tag{1.23}$$

mit $C > 1$ ab.
Rückwärtsanalyse untersucht das Verhältnis von

$$\tilde{E} = \bigcup_{\tilde{x} \in E} \{\tilde{x} : f(\hat{x}) = \tilde{f}(\tilde{x}) \wedge \|\hat{x} - \tilde{x}\| \text{ minimal}\}$$

zu $E \subset \tilde{E}$

Bemerkung 1.18. Man kann zeigen, dass Rückwärtsstabilität Vorwärtsstabilität impliziert.

Beispiel 1.19 (Addition zweier Zahlen)**.**

$$f(a, b) = a + b$$

Vorwärtsanalyse

$$\begin{aligned} \tilde{f}(\tilde{a}, \tilde{b}) &= \tilde{a} \tilde{+} \tilde{b} = a(1 + \epsilon_1) \tilde{+} b(1 + \epsilon_2) \\ &= (a + a\epsilon_1 + b + b\epsilon_2)(1 + \epsilon_3) \\ &\doteq a + b + \underbrace{a\epsilon_1 + b\epsilon_2 + (a + b)\epsilon_3}_{\tau} \end{aligned}$$

mit $|\epsilon_k| \leq \text{eps}$

$$\Rightarrow \tilde{f}(\tilde{a}, \tilde{b}) = a + b + \tau = f(a, b) + \tau \tag{1.24}$$

$$\Rightarrow \left| \tilde{f}(\tilde{a}, \tilde{b}) - f(a, b) \right| = \tau \leq (|a| + |b| + |a + b|)\text{eps}$$

Rückwärtsanalyse
Ausgehend von (1.24) machen wir den Ansatz

$$\tilde{f}(\tilde{a}, \tilde{b}) = a + b + \tau = \hat{a} + \hat{b} = f(\hat{a}, \hat{b})$$

Gleichung hat unendlich viele Lösungen, aber nur eine, die $\|(\hat{a},\hat{b})-(a,b)\|$ (euklidische Norm) minimal macht, nämlich

$$\hat{a}=a+\frac{\tau}{2},\hat{b}=b+\frac{\tau}{2}$$

$$\epsilon_R(a,b)=\frac{\|(\frac{\tau}{2},\frac{\tau}{2})\|}{\|(a,b)\|}=\frac{\tau}{\sqrt{2}\sqrt{a^2+b^2}}\leq\frac{|a|+|b|+|a+b|}{\sqrt{2}\sqrt{a^2+b^2}}\text{eps}$$

Wie sieht es mit der Kondition der Addition aus?

$$\frac{f(a+\Delta a,b+\Delta b)-f(a,b)}{f(a,b)}=\frac{\Delta a+\Delta b}{a+b}\approx\frac{a}{a+b}\frac{\Delta a}{a}+\frac{b}{a+b}\frac{\Delta b}{b}$$

$$\Rightarrow\left|\frac{\Delta a+\Delta b}{a+b}\right|\leq\left|\frac{a}{a+b}\right|\left|\frac{\Delta a}{a}\right|+\left|\frac{b}{a+b}\right|\left|\frac{\Delta b}{b}\right|\leq\underbrace{\max\left\{\left|\frac{a}{a+b}\right|,\left|\frac{b}{a+b}\right|\right\}}_{\kappa_{\text{rel}}}2\text{eps}$$

$$f(x_1,x_2)=x_1+x_2\Rightarrow\frac{\partial f}{\partial x_1}=1=\frac{\partial f}{\partial x_2}$$

$\Rightarrow$ Addition ist für $\text{sgn}(a)=-\text{sgn}(b)$ und $|a|\approx|b|$ schlecht konditioniert, d.h. der Fehler, der unvermeidbar ist, kann im ungünstigsten Fall

$$\max\left\{\left|\frac{a}{a+b}\right|,\left|\frac{b}{a+b}\right|\right\}2\text{eps}$$

sein.

1.9 Fehlerabschätzungen für lin. Gleichungssysteme

Sei A eine reguläre reelle Matrix vom Typ $(n\times n)$ und $\vec{b}\in\mathbb{R}^n$ ein (Spalten-)Vektor. Zu lösen ist

$$A\vec{x}=\vec{b}\tag{1.25}$$

Beispiel 1.20.

$$A=\begin{bmatrix}3&1.001\\6&1.997\end{bmatrix},\vec{b}=\begin{bmatrix}1.999\\4.003\end{bmatrix}$$

(1.25) liefert dann

$$3x_1+1.001x_2=1.999$$
$$6x_1+1.997x_2=4.003$$

Lösung ist Schnittpunkt von 2 Geraden, und zwar $(x_1, x_2) = (1, -1)$. Ändert man $\vec{b}$ geringfügig zu

$$\tilde{\vec{b}} = \vec{b} + \Delta\vec{b} = \begin{bmatrix} 2.002 \\ 4.000 \end{bmatrix}, \quad \text{d.h.} \quad \Delta\vec{b} = \begin{bmatrix} 0.003 \\ -0.003 \end{bmatrix}$$

dann hat $A\tilde{\vec{x}} = \tilde{\vec{b}}$ die Lösung

$$(\tilde{x_1}, \tilde{x_2}) = (0.4004, 0.8)$$

d.h. eine geringfügige Veränderung der Gerade hat eine recht große Auswirkung auf den Schnittpunkt! Die Geraden sind hier fast parallel.

Benötigen nun Werkzeuge zur mathematischen Beschreibung dieses Phänomens:

1.9.1 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
- $\|\vec{x}\|_\infty = \max_{1 \le k \le n}\{|x_k|\}$ 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. $\|A\vec{x}\|_\sigma \le \|A\|_\sigma \|\vec{x}\|_\sigma$ Verträglichkeit
2. $\|AB\|_\sigma \le \|A\|_\sigma \|B\|_\sigma$ Submultiplikativität

Zurück zum Beispiel 1.20. Es ist

$$\|\Delta\vec{b}\|_\infty = 0.003, \|\vec{b}\|_\infty = 4.003 \Rightarrow \frac{\left\|\Delta\vec{b}\right\|_\infty}{\left\|\vec{b}\right\|_\infty} = 7.5 \cdot 10^{-4}$$

$$\|\Delta\vec{x}\|_\infty = \left\|\vec{x} - \tilde{\vec{x}}\right\|_\infty = 1.8, \|\vec{x}\|_\infty = 1 \Rightarrow \frac{\|\Delta\vec{x}\|_\infty}{\|\vec{x}\|_\infty} = 1.8$$

$$\Rightarrow \frac{\frac{\|\Delta\vec{x}\|_\infty}{\|\vec{x}\|_\infty}}{\frac{\left\|\Delta\vec{b}\right\|_\infty}{\left\|\vec{b}\right\|_\infty}} = 2400$$

Wie hängt die Fehlerverstärkung von A ab?

$$\begin{aligned}
&\left\|\vec{x} - \tilde{\vec{x}}\right\| = \left\|A^{-1}\vec{b} - A^{-1}\tilde{\vec{b}}\right\| = \left\|A^{-1}(\vec{b} - \tilde{\vec{b}})\right\| \\
\Rightarrow \quad &\|\Delta\vec{x}\| = \left\|A^{-1}\Delta\vec{b}\right\| \\
\Rightarrow \quad &\frac{\|\Delta\vec{x}\|}{\|\vec{x}\|} = \frac{\left\|A^{-1}\Delta\vec{b}\right\| \cdot \left\|\vec{b}\right\|}{\|\vec{x}\| \cdot \left\|\vec{b}\right\|} = \frac{\|A\vec{x}\| \cdot \left\|A^{-1}\Delta\vec{b}\right\|}{\|\vec{x}\| \cdot \left\|\vec{b}\right\|} \\
&\qquad \leq \frac{\|A\| \, \|\vec{x}\| \cdot \|A^{-1}\| \left\|\Delta\vec{b}\right\|}{\|\vec{x}\| \left\|\vec{b}\right\|} = \|A\| \left\|A^{-1}\right\| \frac{\left\|\Delta\vec{b}\right\|}{\left\|\vec{b}\right\|}
\end{aligned}$$

Also

$$\frac{\|\Delta\vec{x}\|}{\|\vec{x}\|} \leq \|A\| \left\|A^{-1}\right\| \frac{\left\|\Delta\vec{b}\right\|}{\left\|\vec{b}\right\|} \tag{1.26}$$

Definition 1.21. *Die Zahl*

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

heißt Konditionszahl der Matrix A und ist eine Schranke für die Fehlerverstärkung bei der Lösung von (1.25) *bei gestörter rechter Seite.*

Zurück zu Beispiel 1.20

$$\|A\|_\infty \left\|A^{-1}\right\|_\infty = 4798.2 = \text{cond}(A)$$

1.9.2 Weitere Vektor- und Matrixnormen

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

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

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.

4. Vorlesung am 27.4.2009

Theorem 1.22. *Für die Berechnung von speziellen induzierten Matrixnormen gilt (A reelle $(m \times n)$ Matrix)*

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{m} |a_{ij}| \quad \textit{Spaltensummennorm} \tag{1.27}$$

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \quad \textit{Zeilensummennorm} \tag{1.28}$$

$$\|A\|_2 = \sqrt{\lambda_{max}} \quad \textit{mit} \quad \lambda_{max} \textit{ größter EW von } A^T A \tag{1.29}$$

Beweis. (1.27) und (1.28) als Übung.
Für den Nachweis von (1.29) überlegt man, dass $A^T A$ ähnlich einer Diagonalmatrix mit den EW von $A^T A$ als Diagonalelementen ist. Die EW sind positiv und aus der Definition von $\|A\|_2$ folgt dann schließlich (1.29) □

Definition 1.23. *Unter dem Absolutbetrag einer Matrix $A \in \mathbb{C}^{m\times n}$ versteht man die Matrix*

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

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

Bemerkung 1.24. Für die "Beträge"von Matrizen gelten die Beziehungen

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

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

(iii) $A \le B, C \ge 0, D \ge 0 \Rightarrow CAD \le CBD$

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

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

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

Beweis. Größtenteils trivial □

Wir kehren zurück zur Fehlerfortpflanzung bei der Lösung linearer Gleichungssysteme. Seien A und $\vec{b}$ durch $\Delta A = \epsilon F$ bzw. $\Delta\vec{b} = \epsilon\vec{f}$ gestört, also statt $A\vec{x} = \vec{b}$ wird

$$(A + \epsilon F)\vec{x}(\epsilon) = \vec{b} + \epsilon\vec{f} \tag{1.30}$$

betrachtet. $F, \vec{f}, \epsilon$ sind Störungen und für $\epsilon = 0$ ergibt sich das eigentliche Problem. A sei regulär. Der Satz über implizite Funktionen liefert dann die Existenz und Differenzierbarkeit der Abbildung $\epsilon \mapsto \vec{x}(\epsilon)$ und $\vec{x} = \vec{x}(0)$

$$\dot{\vec{x}}(0) = A^{-1}(\vec{f} - F\vec{x})$$

Für die Taylorreihe folgt

$$\vec{x}(\epsilon) = \vec{x} + \epsilon\dot{\vec{x}}(0) + \mathcal{O}(\epsilon^2)$$

also

$$\vec{x}(\epsilon) - \vec{x} = \epsilon\dot{\vec{x}}(0) + \mathcal{O}(\epsilon^2) = \epsilon A^{-1}(\vec{f} - F\vec{x}) + \mathcal{O}(\epsilon^2)$$

Für den relativen Fehler gilt dann

$$\frac{\|\vec{x}(\epsilon) - \vec{x}\|}{\|\vec{x}\|} \leq \epsilon \frac{\left\|A^{-1}(\vec{f} - F\vec{x})\right\|}{\|\vec{x}\|}$$

Und im Fall einer submultiplikativen (konsistenten) Matrixnorm von A

$$\frac{\|x(\epsilon) - x\|}{\|x\|} \leq \epsilon \left\|A^{-1}\right\| \left\{\frac{\|f\|}{\|x\|} + \|F\|\right\} + \mathcal{O}(\epsilon^2) \tag{1.31}$$

(1.31) sagt noch nichts über die Auswirkungen der relativen Fehler

$$\frac{\|\Delta A\|}{\|A\|} \quad \text{bzw.} \quad \frac{\|\Delta b\|}{\|b\|}$$

aus, was aber von Interesse ist. Mit $b = Ax$ folgt $\|b\| \leq \|A\| \|x\|$, also

$$\frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}$$

und mit $\|F\| = \frac{\|F\|}{\|A\|} \|A\|$ folgt aus (1.31)

$$\begin{aligned}
\frac{\|x(\epsilon) - x\|}{\|x\|} &\leq \|A\| \left\|A^{-1}\right\| \left\{\epsilon\frac{\|F\|}{\|A\|} + \epsilon\frac{\|f\|}{\|b\|}\right\} + \mathcal{O}(\epsilon^2) \\
&= \kappa(A) \left\{\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|}\right\} + \mathcal{O}(\epsilon^2) \\
&\dot{\leq} \kappa(A) \left\{\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|}\right\}
\end{aligned} \tag{1.32}$$

(1.32) gilt für kleine ϵ. Im Folgenden soll eine genauere Analyse vorgenommen werden

Lemma 1.25. *$F \in \mathbb{R}^{n\times n}$ und $\|\cdot\|$ submultiplikativ (konsistent). Für $\|F\| \leq 1$ ist $E - F$ nicht singulär, und es gilt*

$$(E-F)^{-1} = \sum_{k=0}^{\infty} F^k$$

sowie

$$\left\|(E-F)^{-1}\right\| \leq \frac{1}{1-\|F\|}$$

Beweis. In Analogie zur geometrischen Reihe, bzw. Verallgemeinerung derselben. □

Lemma 1.26. *Es gelte $Ax = b$ mit regulärem $A \in \mathbb{R}^{n\times n}$ und $b \in \mathbb{R}^n, b \neq 0$. $F = \Delta A$ und $f = \Delta b$ seien Störungen mit relativem Fehler $\leq \delta$, d.h.*

$$\frac{\|\Delta A\|}{\|A\|} \leq \delta, \quad \frac{\|\Delta b\|}{\|b\|} \leq \delta$$

Falls $\delta\kappa(A) = r < 1$, so ist $A + \Delta A$ regulär und für die Lösung $\tilde{x}$ von $(A + \Delta A)\tilde{x} = b + \Delta b$ gilt

$$\frac{\|\tilde{x}\|}{\|x\|} \leq \frac{1+r}{1-r}$$

Beweis. Es gilt

$$\left\|A^{-1}F\right\| \leq \left\|A^{-1}\right\| \|F\| \leq \delta \underbrace{\left\|A^{-1}\right\| \|A\|}_{\kappa(A)} = r < 1$$

$$\Rightarrow A + F = A(E + A^{-1}F) =: A(E - \tilde{F}) \quad \text{mit} \quad \left\|\tilde{F}\right\| < 1$$

$(E-\tilde{F})^{-1}$ ex. nach Lemma 1.25 also existiert auch $(A+F)^{-1} = (E-\tilde{F})^{-1}A^{-1}$. Nach Definition von $\tilde{x}$ folgt

$$\begin{aligned}
& (A+F)\tilde{x} = b + f = Ax + f \\
\Rightarrow \quad & (E + A^{-1}F)\tilde{x} = x + A^{-1}f \\
\Rightarrow \quad & \|\tilde{x}\| \leq \left\|(E + A^{-1}F)^{-1}\right\| \left\|x + A^{-1}f\right\| \\
& \text{mit} \quad \left\|(E-\tilde{F})^{-1}\right\| \leq \frac{1}{1-\left\|\tilde{F}\right\|} \leq \frac{1}{1-r} \\
\Rightarrow \quad & \|\tilde{x}\| \leq \frac{1}{1-r}(\|x\| + \left\|A^{-1}\right\| \|f\|) \\
& \qquad \leq \frac{1}{1-r}(\|x\| + \delta \|A\| \left\|A^{-1}\right\| \|x\|) = \frac{1+r}{1-r}\|x\|
\end{aligned}$$

□

Zur Abschätzung des relativen Fehler betrachten wir

$$\left.\begin{array}{rcl} A\tilde{x} + F\tilde{x} & = & b + f \\ Ax & = & b \end{array}\right\} \Rightarrow A(\tilde{x} - x) = f - F\tilde{x}$$

$$\Rightarrow \quad \tilde{x} - x = A^{-1}f - A^{-1}F\tilde{x} \tag{1.33}$$

Theorem 1.27. *Unter den Voraussetzungen des Lemmas 1.26 gilt*

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \frac{2r}{1-r}, \quad r = \delta\kappa(A) \tag{1.34}$$

Beweis. Es ergibt sich aus (1.33)

$$\begin{aligned} \|\tilde{x} - x\| &\leq \left\|A^{-1}f\right\| + \left\|A^{-1}F\tilde{x}\right\| \\ &\leq \left\|A^{-1}\right\| \|f\| + \left\|A^{-1}F\right\| \|x\| \\ &\leq \delta\kappa A \|x\| + r\frac{1+r}{1-r} \|x\| \\ &= r\left(1 + \frac{1+r}{1-r}\right)\|x\| = \frac{r \cdot 2}{1-r}\|x\| \end{aligned}$$

□

In den bisherigen Abschätzungen wurde der Fehler F der Matrix A in der Norm $\|F\|$ betrachtet und der Fehler von b ebenso in der Vektornorm $\|f\|$, d.h.

$$\|F\| \leq \delta \|A\| \quad \text{und} \quad \|f\| \leq \delta \|b\|$$

wurden vorausgesetzt. Betrachtet man nur die Komponenten, d.h. fordert man

$$|f_{ij}| \leq \delta |a_{ij}|, \quad i, j = 1, \ldots, n$$

dann lassen sich die Abschätzunden des Lemmas 1.26 und des Theorems 1.27 verbessern. Es gelten:

Lemma 1.28. *Sei $A \in \mathbb{R}^{n\times n}$ regulär und $0 \neq b \in \mathbb{R}^n$ Es gelte $Ax = b$ und $(A + \Delta A)\tilde{x} = b + \Delta b$ und die Abschätzungen*

$$|\Delta A| \leq \delta |A|, \quad |\Delta b| \leq \delta |b|$$

Ist die Bedingung

$$\delta \left\| |A^{-1}||A| \right\|_M = r < 1$$

erfüllt, so ist $A + \Delta A$ regulär und es gilt

$$\frac{\|\tilde{x}\|}{\|x\|} \leq \frac{1+r}{1-r}$$

($\|\cdot\|_M$ ist eine der Matrixnormen $\|\cdot\|_\infty, \|\cdot\|_1, \|\cdot\|_F$ und $\|\cdot\|$ die damit verträgliche Vektornorm).

Theorem 1.29. *Unter den obigen Vorraussetzungen gilt*

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq \frac{2r}{1-r} \tag{1.35}$$

für die obigen Normen.

Bemerkung 1.30. Die Aussagen der Theoreme 1.27 und 1.29 findet man auch oft in der Form

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq \frac{\kappa(A)}{1 - \|A^{-1}\| \, \|\Delta A\|} \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|} \right)$$

Kapitel 2

Lösung linearer Gleichungssysteme

2.1 LR-Zerlegung

Zu lösen ist $Ax = b$. Dazu soll A als Produkt einer unteren Dreiecksmatrix U 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

$$Rx = y$$

2.1.1 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 λ-fachen der j-ten Zeile von A zur i-ten Zeile d.h. durch geeignete Wahl von λ 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 λ unterhalb von $\tilde{a}_{kk}$ Null-Einträge

Nun etwas präziser

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

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 2.1.

$$M_k := \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -t_{k+1k} & \ddots & \\ & & \vdots & & \ddots \\ & & -t_{nk} & & 1 \end{pmatrix}$$ *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

$$\operatorname{rg}(M_k) = n$$

$$\det(M_k) = 1$$

$$M_k^{-1} = E + t^{(k)} e_k^T, \quad da$$

$$M_k^{-1} M_k = E - t^{(k)} \underbrace{e_k^T t^{(k)}}_{=0} e_k = E$$

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.

5. Vorlesung am 29.4.2009

Definition 2.2. *Eine obere oder untere Dreiecksmatri, deren Diagonalelemente alle gleich eins sind, heißt* ***unipotent****. Die Zerlegung $A = LR$ heißt LR-Zerlegung, wenn L eine unipotente untere Dreiecksmatrix ist.*

Satz 2.3. *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:

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

Beweis. a) A besitze LR-Zerlegung

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

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

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

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 $A^{(k-1)} = M_{k-1} \cdot ... \cdot M_1 A$

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

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

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

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. $\Rightarrow$ 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 2.4. *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 2.5. *Sind $\tilde{L}, \tilde{R}$ die Matrizen aus Satz 2.4, 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 2.4 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

2.1.2 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 2.6. *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 2.7. *Wir bezeichnen den im Beispiel beschriebenen Algorithmus als Konstruktion einer LR-Zerlegung mit Spaltenpivotisierung (auch Gaußelimination mit partieller Pivotisierung).*

Satz 2.8. *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

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

gilt. Außerdem hat $\hat{M}_k = P_\mu M_k P_\mu$ die gleiche Struktur wie M_k, da durch die Multiplikation von P_μ 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 (≤ 1) hat, was zu einer Verbesserung der Abschätzung aus Satz 2.5 führt.

2.2 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 2.9. *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 2.3 (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 $D^{\frac{1}{2}} = \operatorname{diag}(d_1^{\frac{1}{2}}, \ldots, d_n^{\frac{1}{2}})$

$$\begin{aligned} &\Rightarrow A = LR = LD\hat{R} = LD^{\frac{1}{2}}D^{\frac{1}{2}}\hat{R} \\ &\Rightarrow D^{-\frac{1}{2}}L^{-1}A = D^{\frac{1}{2}}\hat{R}, \quad D^{-\frac{1}{2}} = \operatorname{diag}(d_1^{-\frac{1}{2}}, \ldots, d_n^{-\frac{1}{2}}) \end{aligned} \tag{2.1}$$

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 (2.1) ergibt

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

□

2.2.1 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

2.3 Orthogonale Matrizen – QR-Zerlegung

6. Vorlesung am 4.5.2009

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

$$A = QS \tag{2.2}$$

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

$$S = \begin{bmatrix} R \\ - \\ 0 \end{bmatrix} \in \mathbb{R}^{n \times m}, R = \begin{bmatrix} * & & * \\ & \ddots & \\ 0 & & * \end{bmatrix} \in \mathbb{R}^{m \times m} \tag{2.3}$$

Solche Zerlegungen ermöglichen z.B. die stabile Lösung von schlecht konditionierten lösbaren linearen Gleichungssystemen $Ax = b, (m = n)$ oder die stabile Lösung von Ausgleichsproblemen

$$\min_{x \in \mathbb{R}^m} \|Ax - b\|_2$$

Wir erinnern uns an die Eigenschaften orthogonaler Matrizen:

(i)

$$\|Qx\|_2 = \|x\|_2 = \left\|Q^T x\right\|_2, x \in \mathbb{R}^n \tag{2.4}$$

(ii)

$$\text{cond}(QA) = \text{cond}(A) \tag{2.5}$$

(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 (2.2), (2.3) die Form

$$A = QR \tag{2.6}$$

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 $a_k, q_k \in \mathbb{R}^n, k = 1, \ldots, n$. (2.6) bedeutet dann

$$a_j = \sum_{i=1}^{j} r_{ij} q_i, \quad j = 1, ..., n, q_1, \ldots, q_n \in \mathbb{R}^n \tag{2.7}$$

2.3.1 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 $M_{j-1} \subset \mathbb{R}^n$

$$\hat{q}_j := a_j - \sum_{i=1}^{j-1} \langle a_j, q_i \rangle \, q_i \tag{2.8}$$

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 (2.8) folgt

$$a_j = \underbrace{\|\hat{q}_j\|_2}_{r_{jj}} q_j + \sum_{i=1}^{j-1} \underbrace{(a_j^T q_i)}_{r_{ij}} q_i = \sum_{i=1}^{j} r_{ij} q_i, \quad j = 1, \ldots, n \tag{2.9}$$

Nach Abschluss der Gram-Schmidt-Orthogonalisierung hat man damit mit (2.9)

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

als QR-Zerlegung

Bemerkung 2.10. Der beschriebene Algorithmus kann im ungünstigsten Fall Probleme bereiten (nicht gutartig sein), wenn z.B. $\|\hat{q}_j\|$ recht klein wird (Lösung folgt).

2.3.2 Householder-Matrizen/Transformationen

Definition 2.11. *Eine Abbildung $H : \mathbb{R}^n \to \mathbb{R}^n, x \mapsto Hx$ mit einer Matrix*

$$H = E - 2ww^T, w \in \mathbb{R}^n, \|w\|_2 = w^T w = 1 \tag{2.10}$$

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

$$w = \frac{x + \sigma e_1}{\|x + \sigma e_1\|_2} \quad mit \quad \sigma = \pm \|x\|_2 \tag{2.11}$$

gilt

$$\|w\| = 1, \quad Hx = (E - 2ww^T)x = -\sigma e_1 \tag{2.12}$$

Beweis. $\|w\| = 1$ weil $x + \sigma e_1 \neq 0$ und damit (2.11) wohldefiniert ist. Für den Nachweis von (2.12) 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 (2.11), 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 (2.12) 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 (2.11) $\sigma = \operatorname{sgn}(x_1)\|x\|_2$ gewählt, d.h. z.B. für $x = (-3, 1, 5)^T$ ist $\sigma = -\sqrt{35}$

2.3.3 Algorithmus zur Konstruktion der Faktorisierung mittels Householder-Transformationen

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

$$A^{(j)} = \begin{bmatrix} a_{11}^{(j)} & a_{12}^{(j)} & & \cdots & & a_{1m}^{(j)} \\ & \ddots & & & & \\ & & a_{j-1j-1}^{(j)} & & & a_{j-1m}^{(j)} \\ & & & a_{jj}^{(j)} & \cdots & a_{jm}^{(j)} \\ & & & \vdots & & \vdots \\ & & & a_{nj}^{(j)} & \cdots & a_{nn}^{(j)} \end{bmatrix}, \quad j = 1, \ldots, m+1 \tag{2.13}$$

berechnet werden, sodass am Ende mit $A^{(m+1)} = S$ die verallgemeinerte obere Dreiecksmatrix vorliegt.

Die Matrizen der Form (2.13) erhält man für $j = 1, \ldots, m$ 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 $\hat{H}_1, \ldots, \hat{H}_m$ sind aufgrund der Eigenschaften der Matrizen $H_1, \ldots, H_m$ orthogonal und symmetrisch, sodass man mit

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

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

2.4 Anwendungen der QR-Zerlegung

7. Vorlesung am 6.5.2009

2.4.1 Lösung eines linearen Gleichungssystems

$$Ax = b, \quad A \in \mathbb{R}^{n \times n} \text{regulär}$$

$\Rightarrow$ mit $A = QR, Q, R \in \mathbb{R}^{n \times n}$ Q orthogonal, R obere Dreiecksmatrix

$$Ax = b \quad \Leftrightarrow \quad Qy = b, y = Q^T b, Rx = y \text{ durch Rückwärtseinsetzen}$$

2.4.2 Ausgleichsprobleme

Gegeben: Wertepaare $(y_k, x_k), k = 1, \ldots, n$ (z.B. Ergebnisse von Messungen)
Gesucht: funktionaler Zusammenhang

$$y_k = f(x_k), \quad k = 1, \ldots, n \tag{2.14}$$

Wobei man f nicht kennt. Man kann mit einem Mehrparameteransatz

$$f = f(x, r_1, r_2, \ldots, r_m), \quad n \gg m$$

und versucht (2.14) im quadratischen Mittel zu lösen

$$\min_{r\in\mathbb{R}^m} \sum_{k=1}^{n} (y_k - f(x_k, r_1, \ldots, r_m))^2$$

Methode der kleinsten Quadrate (Gauß)

Lineares Ausgleichsproblem als Spezialfall

f lineare Funktion von $\vec{r}$. Wir setzen

$$f_k(\vec{r}) := f(x_k, \vec{r}), \quad \vec{r} \in \mathbb{R}^m, \vec{y} = (y_1, \ldots, y_n)^T$$

und damit ergibt sich als Ansatz

$$\begin{bmatrix} f_1(\vec{r}) \\ \vdots \\ f_n(\vec{r}) \end{bmatrix} = M\vec{r}, \quad M \in \mathbb{R}^{n\times m}$$

Beispiel. $(y_k, x_k), k = 1, \ldots, 4$ quadratischer Ansatz

$$y = r_1 + r_2 x + r_3 x^2$$

$$M = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \end{bmatrix}$$

Lineares Ausgleichsproblem

$$\min_{\vec{r}\in\mathbb{R}^m} \|M\vec{r} - \vec{y}\|_2^2 \tag{2.15}$$

$F(\vec{r}) = \|M\vec{r} - \vec{y}\|_2^2 = \langle M\vec{r} - \vec{y}, M\vec{r} - \vec{y}\rangle$ ist differenzierbar und konvex, daraus ergibt sich als notwendige und hinreichende Bedingung für das gesuchte Minimum

$$\nabla F(\vec{r}) = 0$$

$$\begin{aligned} F'(\vec{r})h &= 2\langle M\vec{r} - \vec{y}, Mh\rangle = 2\left\langle M^T(M\vec{r} - \vec{y}), h\right\rangle \\ \Rightarrow \nabla F(\vec{r}) &= 2M^T(M\vec{r} - \vec{y}) = 0 \end{aligned}$$

damit erhält man als Normalengleichungen das lineare System

$$M^T M\vec{r} = M^T\vec{y} \tag{2.16}$$

Im Folgenden werden die Vektorpfeile wieder weggelassen, da dort aus dem Kontext klar werden sollte, ob es sich um einen Vektor oder einen Skalar handelt.

Satz 2.13. *Das lineare Ausgleichsproblem* (2.15) *hat mindestens eine Lösung* r_0. *Für jede andere Lösung* r *gilt* $Mr = Mr_0$. *Das Residuum* $d = y - Mr_0$ *ist eindeutig bestimmt und erfüllt* $M^T d = 0$. *Die Gleichung* (2.16) *ist notwendig und hinreichend dafür, dass* r *Lösung von* (2.15) *ist*

Beweis. Sei $\mathrm{Im}(M)$ Bild von M es gilt

$$\mathbb{R}^n = \mathrm{Im}(M) \oplus \mathrm{Im}(M)^\perp$$

also kann y eindeutig zerlegt werden als

$$y = s + d, \quad s \in \mathrm{Im}(M), d \in \mathrm{Im}(M)^T$$

Nach Definition des Bildes von M gibt es zu s mindestens ein r_0 mit

$$Mr_0 = s$$

außerdem gilt

$$\langle \underbrace{Mr}_{\in \mathrm{Im}(M)}, \underbrace{d}_{\in \mathrm{Im}(M)^\perp} \rangle = 0 \quad \forall r \in \mathbb{R}^m$$

und somit

$$\left\langle r, M^T d \right\rangle = 0 \quad \text{und} \quad M^T d = 0$$

Es folgt

$$M^T y = M^T s + \underbrace{M^T d}_{=0} = M^T M r_0$$

$\Rightarrow r_0$ ist eine Lösung von (2.16).
Sei nun r eine weitere Lösung, d.h.

$$M^T M r = M^T y$$

Dann ist $d = y - Mr$ orthogonal zu $\mathrm{Im}(M)$, denn

$$\langle Mz, d \rangle = \left\langle z, M^T d \right\rangle = \left\langle z, M^T y - M^T M r \right\rangle = 0 \quad \forall z$$

$s = Mr$ gehört zu $\mathrm{Im}(M)$ und

$$y = Mr + (y - Mr) = s + d$$

ist damit wieder eine Zerlegung von y in $s \in \mathrm{Im}(M), d \in \mathrm{Im}(M)^\perp$ und aus der Eindeutigkeit der Zerlegung folgt:

$$Mr = Mr_0$$

□

Bestimmung der Lösung des Minimumproblems

(i) Lösung des Systems

$$M^T M r = M^T y$$

$M^T M$ ist symmetrisch und positiv semidefinit, da

$$\langle r, M^T M r \rangle = \langle Mr, Mr \rangle \geq 0$$

Die Gleichheit mit Null kann nur eintreten, wenn $Mr = 0$. Hat M den vollen Rang, d.h. bei $n \geq m$ also den Rang m, dann kann $r = 0$ nur gelten bei $\langle r, M^T M r \rangle = 0$. Also ist $M^T M$ dann positiv definit. D.h. Cholesky-Zerlegung ist möglich.

- dazu notwendig: Berechnung von $M^T M$ ($\frac{1}{2}m^2 n$ Multiplikationen), bei $n \gg m$ ist der Aufbau von $M^T M$ teurer als das Cholesky Verfahren ($\frac{1}{6}m^3$ Multiplikationen)
- $M^T M$ oft schlecht konditioniert, Fehler in y werden durch die Kondition $\kappa(M^T M)$ verstärkt
- besser ist Methode, die nur M verwendet

(ii) Bestimmung von r mittels QR-Zerlegung

Gute Eigenschaft von orthogonalen Transformationen

$$\kappa_2(Q) = \|Q\|_2 \left\|Q^{-1}\right\|_2 = \|Q\|_2 \left\|Q^T\right\|_2 = 1$$

D.h. sie verstärken nicht den Fehler der Größen, auf die sie angewendet werden.

Satz 2.14. *Sei $M \in \mathbb{R}^{n\times m}, n \geq m$, von vollem Rang und $Q \in \mathbb{R}^{n\times n}$ orthogonal, sodass*

$$Q^T M = \begin{pmatrix} R \\ 0 \end{pmatrix} \quad Q^T y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

mit $y_1 \in \mathbb{R}^m, y_2 \in \mathbb{R}^{n-m}$ und einer invertierbaren oberen Dreiecksmatrix $R \in \mathbb{R}^{m\times m}$ gilt. Dann ist $r = R^{-1}y_1$ die Lösung des linearen Ausgleichsproblems $\min_{r\in\mathbb{R}^m} \|Mr - y\|_2^2$

Beweis. Q ist isometrisch $\Rightarrow \|y - Mr\|_2 = \left\|Q^T(y - Mr)\right\|_2$, daraus folgt:

$$\begin{aligned} \|y - Mr\|_2^2 = \left\|Q^T(y - Mr)\right\|^2 &= \left\| \begin{pmatrix} y_1 - Rr \\ y_2 \end{pmatrix} \right\|^2 \\ &= \|y_1 - Rr\|^2 + \|y_2\|^2 \geq \|y_2\|^2 \quad \forall r \in \mathbb{R}^m \end{aligned}$$

Wählt man $r = R^{-1}y_1$, so hat man die Lösung. □

Bemerkung. Residuum d erfüllt $\|d\|_2 = \|y_2\|_2$.

Kapitel 3

Interpolation

8. Vorlesung am 11.5.

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

$$P(x_k, a_0, \ldots, a_n) = y_k, \quad k = 0, \ldots, n \tag{3.1}$$

zu den vorgegebenen Stützstellen (x_k, y_k) erfüllt sind. (3.1) heißt auch **Interpolationseigenschaft** und die Stützstellen werden auch **Knoten** genannt. (3.1) sind $n+1$ Gleichungen für die $n+1$ Parameter $a_0, \ldots, a_n$.
Sehr einfach: Lineare Splines

3.1 Polynominterpolation

Definition 3.1. *Unter Π_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 3.2. *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:

$$\begin{aligned} P(x) &= a_0 + a_1 x + \cdots + a_n x^n \\ P(x_k) &= y_k, \quad k = 0, 1, \ldots, n \end{aligned} \tag{3.2}$$

bedeutet

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

$$\Leftrightarrow \underbrace{\begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ & & \vdots & & \\ 1 & x_2 & x_2^2 & \cdots & x_n^n \end{pmatrix}}_{\text{Vandermondesche Matrix } V} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix} \tag{3.3}$$

V ist für paarweise verschiedene x_k regulär, d.h. $a_0, \ldots, a_n$ und damit P sind eindeutig bestimmt.

□

Definition 3.3. *Das nach Satz 2.3 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****.*

3.1.1 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 3.4. *Die Polynome*

$$L_k(x) = \prod_{k \neq i = 0}^{n} \frac{x - x_i}{x_k - x_i} \tag{3.4}$$

heißen Lagrange-Basispolynome.

Definition 3.5. *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 3.6. *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 Π_n vom Grad $\leq n$

Beweis. Als Übung empfohlen. □

3.2 Lagrange-Interpolation

Zuerst ist anzumerken, dass man das Interpolationspolynom nicht in der Form (3.2) auf der Grundlage der Lösung des Gleichungssystems (3.3) 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:

$$L_k(x_j) = \delta_{kj}, \quad k = 0, \ldots, n \tag{3.5}$$

Damit ergibt sich für das Interpolationspolynom:

$$p(x) = \sum_{k=0}^{n} y_k L_k(x) \tag{3.6}$$

da

$$p(x_k) = 0 + 0 + \cdots + y_k L_k(x_k) + \cdots + 0 = y_k$$

gilt. (3.6) heißt **Lagrangsches Interpolationspolynom**.

3.3 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 3.7.

$$p_n(x) := \sum_{k=0}^{n} c_k N_k(x) \in \Pi_n$$

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 3.8. *Wir schreiben $C_j := f[x_0 x_1 \ldots x_k]$ 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

$$p^*_{i_0 i_1 \ldots i_k}(x_{ij}) \equiv y_{ij}, \quad j = 0, 1, \ldots, k$$

Damit gilt

$$p^*_k(x) \equiv y_k \tag{3.7}$$

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 3.9. *Es gilt für alle* $k \in \{1, 2, \ldots, n\}$

$$p^*_{i_0,\ldots,i_k}(x) = \frac{(x - x_{i_0})p^*_{i_1\ldots i_k}(x) - (x - x_{i_k})p^*_{i_0\ldots i_{k-1}}(x)}{x_{i_k} - x_{i_0}} \tag{3.8}$$

Beweis. Induktion Die beiden rechts stehenden Polynome in (3.8) haben einen Grad $\leq k-1$ (damit der gesamte Ausdruck einen Grad $\leq n$).
Anfang ($k = 1$) ist trivial wegen (3.7).
Es ist zu zeigen, dass das rechts in (3.8) stehende Polynom das Interpolationypolynom zu den Stützstellen $x_{i_0}, \ldots, x_{i_k}$ ist (Ausdruck rechts von (3.8) 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 3.10. *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 (3.8). 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 3.10 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:

$$p_2(x) = f[x_0] + f[x_0x_1](x-x_0) + f[x_0x_1x_2](x-x_0)(x-x_1)$$

9. Vorlesung am 13.5.09

3.4 Algorithmische Aspekte der Polynominterpolation

3.4.1 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.

3.4.2 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:

$$p(x) = \sum_{k=0}^{n} y_k L_k(x) = \sum_{k=0}^{n} y_k \prod_{j=0}^{n} \frac{x - x_k}{x_j - x_k} \tag{3.9}$$

$$= \sum_{k=0}^{n} y_k \frac{1}{x - x_k} \left[\prod_{k \neq i = 0}^{n} \frac{1}{x_i - x_k} \right] \prod_{j=0}^{n} (x - x_j) \tag{3.10}$$

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

$$p(x) = \left[\sum_{k=0}^{n} \mu_k y_k \right] \prod_{j=0}^{n} (x - x_j) \tag{3.11}$$

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

$$1 = p(x) = \left[\sum_{k=0}^{n} \mu_k\right] \prod_{j=0}^{n} (x - x_k)$$

$$\Rightarrow \quad \prod_{j=0}^{n} (x - x_k) = \frac{1}{\sum_{k=0}^{n} \mu_k} \tag{3.12}$$

Aus (3.11) und (3.12) folgt mit

$$p(x) = \frac{\sum_{k=0}^{n} \mu_k y_k}{\sum_{k=0}^{n} \mu_k} \tag{3.13}$$

die sogenannte baryzentrische Formel der Lagrange-Interpolation

Satz 3.11. *Für die $n+1$ Stützkoeffizienten $\lambda_k^{(n)}$ zu den paarweise verschiedenen Stützstellen $x_0, x_1, \ldots, x_n$ gilt:*

$$\sum_{k=0}^{n} \lambda_k^{(n)} = 0 \tag{3.14}$$

Die Formel (3.13) 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 λ-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 (3.14) durch

$$\lambda_{n+1}^{(n+1)} = -\sum_{k=0}^{n} \lambda_k^{(n+1)}$$

Insgesamt braucht man zur Bestimmung der μ_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.

3.5 Verfahren von Neville und Aitken

Es ist vergleichbar mit der Herangehensweise bei der Newton-Interpolation
Aus Lemma 3.9 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

$$\begin{array}{c|llll} x_0 & y_0 = p_0^*(x) & & & \\ x_1 & y_1 = p_1^*(x) & p_{0,1}^*(x) & & \\ x_2 & y_2 = p_2^*(x) & p_{1,2}^*(x) & p_{0,1,2}^*(x) & \\ \vdots & \vdots & & \ddots & \\ x_n & y_n = p_n^*(x) & p_{n-1,n}^*(x) & \dots & 0 \end{array}$$

Beispiel.

x_k	0	1	3
y_k	1	3	2

Polynomwert soll an der Stelle $x = 2$ berechnet werden.

$$\begin{array}{c|lll} 0 & 1 & & \\ 1 & 3 & p_{0,1}(2) = \frac{(2-0)3-(2-1)1}{1-0} = 5 & \\ 3 & 2 & p_{1,2}(2) = \frac{(2-1)2-(2-3)3}{3-1} = \frac{5}{2} & p_{0,1,2}(2) = \frac{(2-0)\frac{5}{2}-(2-3)5}{3-0} = \frac{10}{3} \end{array}$$

3.6 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 3.12. *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 3.12 rechtfertigt

Definition 3.13.

$$f[\underbrace{x, x, \ldots, x}_{n+2}] = \frac{f^{(n+1)}(x)}{(n+1)!} \tag{3.15}$$

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.

$$(x_0, y_0) = (0, 1)$$
$$(x_0, y_0') = (0, 2)$$
$$(x_0, y_0'') = (0, 4)$$
$$(x_1, y_1) = (1, 2)$$
$$(x_1, y_1') = (1, 3)$$

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:

$$p(x) = f[x_0] + f[x_0, x_0](x - x_0) + f[x_0, x_0, x_0](x - x_0)^2$$
$$+f[x_0, x_0, x_0, x_1](x - x_0)^3 + f[x_0, x_0, x_0, x_1, x_1](x - x_0)^3(x - x_1)$$

also für obige Werte

$$p(x) = 1 + 2x + 2x^2 - 3x^3 + 6x^3(x - 1)$$

3.7 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.

$$f(x) - p_n(x) = f[x_0, x_1, \ldots, x_n, x](x - x_0)(x - x_1) \cdots (x - x_n) \qquad (3.16)$$

Der folgende Satz liefert die Grundlage für die Abschätzung des Interpolationsfehlers (3.16).

Satz 3.14. *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 Bärwolff □

Aus dem Satz 3.14 folgt direkt für eine $(n+1)$-mal stetig differenzierbare Funktion die Fehlerabschätzung

$$|f(x) - p_n(x)| \leq \frac{\max_{\xi\in[a,b]} \left|f^{(n+1)}(\xi)\right|}{(n+1)!} \underbrace{\left|\prod_{k=0}^{n}(x-x_k)\right|}_{=:w(x)} \tag{3.17}$$

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

$$x_k^* = \frac{a+b}{2} + \frac{b-a}{2}\cos\left(\frac{2(n+1-k)-1}{2(n+1)}\pi\right), \quad k = 0,\ldots,n \tag{3.18}$$

von Vorteil, denn für $w^*(x) = \prod_{k=0}^{n}(x-x_k^*)$ gilt:

Satz 3.15. *Seien* x_k *äquidistante und* x_k^* *gemäß* (3.18) *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*

$$\lim_{k\to\infty} p_k^*(x) = f(x) \quad \textit{auf} \quad [a,b]$$

3.8 Spline-Interpolation

10. Vorlesung am 18.5.09

Problem bei der Polynom-Interpolation:
Eventuell große Oszillationen durch Polynome höheren Grades bei Stützpunktzahlen ≥ 10
Deshalb:
Statt eines Interpolationspoly. 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 3.16. *Eine **Splinefunktion** der Ordnung $l \in \mathbb{N}$ zur Zerlegung Δ 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

$$s(x_k) = f_k \quad \text{für} \quad k = 0, 1, \ldots, N \tag{3.19}$$

für $(x_k, f_k), k = 0, 1, \ldots, N$

3.8.1 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 3.17.

(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 (3.19)

(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 3.14) liefert

$$\begin{aligned} |s(x) - f(x)| &\leq \frac{(x - x_{k-1})(x_k - x)}{2} \max_{\xi \in [x_{k-1}, x_k]} |f''(\xi)| \\ &\leq \frac{1}{8} h_{\max}^2 \|f''\|_\infty, \quad x \in [x_{k-1}, x_k] \end{aligned}$$ □

3.8.2 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 3.18. *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

$$\|f'' - s''\|_2^2 = \|f''\|_2^2 - \|s''\|_2^2 - 2([f' - s']s'')(x)\Big|_{x=a}^{x=b} \tag{3.20}$$

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 3.19. *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*

$$\|f''\|_2^2 - \|s''\|_2^2 = \|f'' - s''\|_2^2 \tag{3.21}$$

sofern eine der 3 folgenden Bedingungen erfüllt ist

(a) $s''(a) = s''(b) = 0$

(b) $s'(a) = f'(a),\ s'(b) = f'(a)$

(c) $f'(a) = f'(b),\ s'(a) = s'(b),\ s''(a) = s''(b)$

Beweis. Die Aussage des Satzes ergibt sich durch Berücksichtigung von $(a), (b)$ bzw (c) in der Identität (3.20) □

Korollar 3.1. *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 (3.21) □

3.8.3 Berechnung interpolierender kubischer Splines

Lokaler Ansatz

$$\begin{aligned}s(x) = a_k + b_k(x - x_k) + c_k(x - x_k)^2 + d_k(x - x_k)^3, \\ x \in [x_k, x_{k+1}], k = 0, \ldots, N-1\end{aligned} \tag{3.22}$$

für $s : [a,b] \to \mathbb{R}$
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 3.20. *Falls $N+1$ reelle Zahlen $s_0'', \ldots, s_N'' \in \mathbb{R}$ den folgenden $N-1$ gekoppelten Gleichungen ($k = 1, \ldots, N-1$)*

$$h_{k-1} \underbrace{s_{k-1}''}_{M_{k-1}} + 2(h_{k-1} + h_k) \underbrace{s_k''}_{M_k} + h_k \underbrace{s_{k+1}''}_{M_{k+1}} = \underbrace{6\frac{f_{k+1} - f_k}{h_k} - 6\frac{f_k - f_{k-1}}{h_{k-1}}}_{g_k} \quad (3.23)$$

genügen, so liefert der lokale Ansatz (3.22) *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 Plato, Bärwolff □

Bemerkung. Die **Momente** $M_0, \ldots, M_N$ stimmen mit den 2. Ableitungen der Splinefunktion s in den Knoten x_k überein

$$s_k'' = M_k = s''(x_k), \quad k = 0, \ldots, N$$

(3.23) 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 korrellieren mit den Bedinungen $(a), (b), (c)$ des Satzes 3.19)

3.8.4 Gestalt der Gleichungssysteme

Natürliche Randbedingunden

$$\begin{bmatrix} 2(h_0+h_1) & h_1 & & 0 \\ h_1 & 2(h_1+h_2) & \ddots & \\ & \ddots & & h_{N-2} \\ 0 & & h_{N-2} & 2(h_{N-2}+h_{N-1}) \end{bmatrix} \begin{bmatrix} M_1 \\ \\ \\ M_{N-1} \end{bmatrix} = \begin{bmatrix} g_1 \\ \\ \vdots \\ \\ g_{N-1} \end{bmatrix} \quad (3.24)$$

vollständige Randbedingungen

$$\begin{bmatrix} 2h_0 & h_0 & & & 0 \\ h_0 & 2(h_0+h_1) & \ddots & & \\ & \ddots & 2(h_{N-2}+h_{N-1}) & h_{N-1} & \\ 0 & & h_{N-1} & 2h_{N-1} & \end{bmatrix} \begin{bmatrix} M_0 \\ \\ \\ M_N \end{bmatrix} = \begin{bmatrix} g_0 \\ \vdots \\ \\ g_N \end{bmatrix} \tag{3.25}$$

periodische Randbedingungen

$$\begin{bmatrix} 2(h_{N-1}+h_0) & h_0 & & & h_{N-1} \\ h_0 & 2(h_0+h_1) & \ddots & & \\ & \ddots & & & h_{N-2} \\ h_{N-1} & & & h_{N-2} & 2(h_{N-2}+h_{N-1}) \end{bmatrix} \begin{bmatrix} M_0 \\ \\ \\ M_{N-1} \end{bmatrix} = \begin{bmatrix} g_0 \\ \vdots \\ \\ g_{N-1} \end{bmatrix} \tag{3.26}$$

3.9 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, eine Eigenschaft, die wie folgt definiert ist

Definition 3.21. *Eine Matrix $A = (a_{ij}) \in \mathbb{R}^{N\times N}$ heißt **strikt diagonal dominant**, falls*

$$\sum_{k\neq j=1}^{N} |a_{kj}| < |a_{kk}|\,, \quad k = 1,\ldots,N \tag{3.27}$$

Lemma 3.22. *Jede strikt diagonal dominante Matrix $A = (a_{kj}) \in \mathbb{R}^{N\times N}$ ist reguär und es gilt*

$$\|x\|_\infty \leq \max_{k=1,\ldots,N} \left\{ (|a_{kk}| - \sum_{k\neq j=1}^{n} |a_{kj}|)^{-1} \right\} \|Ax\|_\infty\,, \quad x \in \mathbb{R}^n \tag{3.28}$$

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

$$\begin{aligned}
\|Ax\|_\infty \geq |(Ax)_k| &= \left|\sum_{j=1}^{N} a_{kj}x_j\right| \geq |a_{kk}|\,|x_k| - \sum_{k\neq j=1}^{N} |a_{kj}|\,|x_j| \\
&\geq |a_{kk}|\,|x_k| - \sum_{k\neq j=1}^{N} |a_{kj}|\,\|x\|_\infty = \left(|a_{kk}| - \sum_{k\neq j=1}^{N} |a_{kj}|\right)\|x\|_\infty \\
\Leftrightarrow \quad \|x\|_\infty &\leq \left(|a_{kk}| - \sum_{k\neq j=1}^{N} |a_{kj}|\right)^{-1} \|Ax\|_\infty
\end{aligned}$$

Die liefert die Gültigkeit von (3.28) woraus die Regularität von A direkt folgt. (Aus $Ax = 0$ folgt $x = 0$ als einzige Lösung) □

Korollar 3.2. *Zur Zerlegung Δ 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. □

3.10 Fehlerabschätzungen für interpolierende kubische Splines

Zuerst schreiben wir die Gleichungen (3.23) 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

$$B := \begin{bmatrix} \frac{2}{3} & \frac{h_1}{3(h_0+h_1)} & & & 0 \\ \frac{h_1}{3(h_1+h_2)} & \frac{2}{3} & \frac{h_2}{3(h_1+h_2)} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{h_{N-3}}{3(h_{N-3}+h_{N-2})} & \frac{2}{3} & \frac{h_{N-2}}{3(h_{N-3}+h_{N-2})} \\ 0 & & & \frac{h_{N-2}}{3(h_{N-2}+h_{N-1})} & \frac{2}{3} \end{bmatrix}$$

$$B \begin{bmatrix} s_1'' \\ \vdots \\ s_{N-1}'' \end{bmatrix} = \begin{bmatrix} \hat{g}_1 \\ \vdots \\ \hat{g}_{N-1} \end{bmatrix} \tag{3.29}$$

führt ($h_k = x_{k+1} - x_k$)

Lemma 3.23. *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*

$$\max_{k=1,\dots,N-1} |s''(x_k) - f''(x_k)| \leq \frac{3}{4} \left\| f^{(4)} \right\|_\infty h_{\max}^2$$

Beweis. Siehe Plato □

Das eben bewiesene Lemma ist die Grundlage für den folgenden Satz zur Fehlerabschätzung der Spline-Interpolation

Satz 3.24. *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 $k = 0, \dots, N-1$ und*

$$h_{\max} = \max_{k=0,\dots,N-1} h_k, \quad h_{\min} = \min_{k=0,\dots,N-1} h_k$$

Falls

$$\max_{k=0,\dots,N} |s''(x_k) - f''(x_k)| \leq C \left\| f^{(4)} \right\|_\infty h_{\max}^2$$

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 $x \in [a,b]$

$$|s(x) - f(x)| \leq c \left\| f^{(4)} \right\|_\infty h_{\max}^4 \tag{3.30}$$

$$|s'(x) - f'(x)| \leq c \left\| f^{(4)} \right\|_\infty h_{\max}^3 \tag{3.31}$$

$$|s''(x) - f''(x)| \leq c \left\| f^{(4)} \right\|_\infty h_{\max}^2 \tag{3.32}$$

$$\left| s^{(3)}(x) - f^{(3)}(x) \right| \leq c \left\| f^{(4)} \right\|_\infty h_{\max} \tag{3.33}$$

Beweis. Zuerst wird (3.33) 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.

$$s^{(3)}(x) \equiv \frac{s''(x_{k+1}) - s''(x_k)}{h_k} = \text{const}, \quad x_k \le x \le x_{k+1} \tag{3.34}$$

Taylorentwicklung von f'' um $x \in [x_k, x_{k+1}]$ liefert

$$f^{(3)}(x) = \frac{f''(x_{k+1}) - f''(x_k)}{h_k} - \frac{(x_{k+1} - x)^2}{2h_k} f^{(4)}(\alpha_k) + \frac{(x - x_k)^2}{2h_k} f^{(4)}(\beta_k) \tag{3.35}$$

für gewisse Zwischenstellen $\alpha_k, \beta_k \in [a, b]$. Subtraktion von (3.34) und (3.35) ergibt

$$\begin{aligned} s^{(3)}(x) - f^{(3)}(x) = &\frac{s''(x_{k+1}) - f''(x_{k+1})}{h_k} - \frac{s''(x_k) - f''(x_k)}{h} \\ &+ \frac{(x_{k+1} - x)^2 f^{(4)}(\alpha_k) - (x - x_k)^2 f^{(4)}(\beta_k)}{2h_k} \end{aligned}$$

$$\begin{aligned} \rightsquigarrow \quad &\left| s^{(3)}(x) - f^{(3)}(x) \right| \\ &\le \left\| f^{(3)} \right\|_\infty \frac{1}{\min\{h_0, \ldots, h_{N-1}\}} (c h_{\max}^2 + c h_{\max}^2 + \frac{h_{\max}^2}{2}) \\ &\le \frac{h_{\max}}{h_{\min}} (2C + \frac{1}{2}) \left\| f^{(4)} \right\|_\infty h_{\max} = 2c \left\| f^{(4)} \right\|_\infty h_{\max} \end{aligned}$$

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 (3.32), (3.31), (3.30) erhält man durch sukzessive Integration von (3.33) 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 3.23 für den Fall natürlicher Randbedingungen).

3.11 Trigonometrische Interpolation

12. Vorlesung am 25.05.09

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:

$$\Psi(x) = \begin{cases} \frac{A_0}{2} + \sum_{l=1}^{m} \left(A_l \cos(lx) + B_l \sin(lx)\right), & n = 2m+1 \\ \frac{A_0}{2} + \sum_{l=1}^{m-1} \left(A_l \cos(lx) + B_l \sin(lx)\right) + \frac{A_m}{2} \cos(mx), & n = 2m \end{cases} \tag{3.36}$$

Die Funktion $\Psi(x)$ soll die Interpolationsbedingung

$$\Psi(x_k) = f_k, \quad k = 0, \ldots, n-1 \tag{3.37}$$

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 (3.36) durch Auswertung von (3.37) 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.

$$\sin lx_k = \frac{1}{2i}\left(\left(e^{\frac{2\pi il}{n}}\right)^k - \left(e^{\frac{-2\pi il}{n}}\right)^k\right) \tag{3.38}$$

Bemerkung. Wegen der 2π-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

$$e^{lix_k}, \quad l = 0, \ldots, n-1$$

(3.38) wird in den Ansatz (3.36) eingesetzt, etwas umgeordnet, sodass man mit

$$p(x) = \beta_0 + \beta_1 e^{ix} + \cdots + \beta_{n-1} e^{i(n-1)x} \tag{3.39}$$

ein trigonometrisches Polynom erhält, welches die Interpolationsbedingung erfüllt, d.h.

$$\Psi(x_k) = f_k \Leftrightarrow p(x_k) = f_k, \quad k = 0, \ldots, n-1$$

wobei $\Psi(x) = p(x)$ nicht gilt.
Für die Beziehungen zwischen β_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

$$p(x) = \beta_0\omega^0 + \beta_1\omega^1 + \cdots + \beta_{n-1}\omega^{n-1} =: P(\omega) \tag{3.40}$$

Und $P(x)$ ist tatsächlich Polynom in ω.

Definition 3.25.

$$\omega := e^{ix}, \quad \omega_k = e^{ix_k} \left(= e^{i\frac{2k\pi}{n}}\right)$$

Bemerkung. Wir haben oben $f_k \in \mathbb{R}$ gefordert, darauf kann man auch verzichten und f_k auch aus $\mathbb{C}$ vorgeben.

Satz 3.26. *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* (3.40) *mit*

$$p(x_k) = P(\omega_k) = f_k, \quad k = 0, \ldots, n-1$$

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 $\omega_k = e^{\frac{2k\pi}{n}i}$ eine der n-ten Einheitswurzeln und damit

$$(\omega_k)^n - 1 = 0$$

Ausklammern von $\omega_k - 1$ ergibt

$$(\omega_k - 1)(\omega_k^{n-1} + \omega_k^{n-2} + \cdots + 1) = 0 \tag{3.41}$$

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 $(3.41) = 0$ 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

$$\langle f, g \rangle := \frac{1}{n} \sum_{k=0}^{n-1} f_k \bar{g}_k \tag{3.42}$$

zueinander orthogonal, d.h. $\{\phi_0, \ldots, \phi_{n-1}\}$ *ist Orthogonalsystem in* $\mathbb{C}^n$

Definition 3.27. *Die Koeffizienten* $\beta_0, \ldots, \beta_{n-1}$ *aus* (3.40)*, 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 3.28. *Für die diskreten Fouriertransformierten* β_j *von* $f_j, j = 1, \ldots, n-1$ *gilt*

$$\beta_j = \frac{1}{n} \sum_{k=0}^{n-1} f_k e^{-i\frac{jk2\pi}{n}} \tag{3.43}$$

d.h. sie sind eindeutig bestimmt.

Beweis. Die Interpolationsbedingungen $P(\omega_k) = f_k$ bedeuten

$$\begin{aligned} P(\omega_0) &= \beta_0\omega_0^0 + \cdots + \beta_{n-1}\omega_0^{n-1} = f_0 \\ &\vdots \\ P(\omega_{n-1}) &= \beta_0\omega_{n-1}^0 + \cdots + \beta_{n-1}\omega_{n-1}^{n-1} = f_{n-1} \end{aligned}$$

$$\rightsquigarrow \quad \beta_0\phi_0 + \beta_1\phi_1 + \cdots + \beta_{n-1}\phi_{n-1} = f := \begin{pmatrix} f_0 \\ \vdots \\ f_{n-1} \end{pmatrix} \tag{3.44}$$

die skalare Multiplikation mit ϕ_j ergibt aufgrund der Orthogonalität

$$\beta_j\langle\phi_j,\phi_j\rangle = \langle f,\phi_j\rangle = \frac{1}{n}\sum_{k=0}^{n-1} f_k\bar{\omega_k^j} = \frac{1}{n}\sum_{k=0}^{n-1} f_k\omega_k^{-j} = \frac{1}{n}\sum_{k=0}^{n-1} f_k e^{-i\frac{2kj\pi}{n}}$$

□

Bemerkung. Für die Fourierkoeffizienten oder diskreten Fouriertransformierten β_k von f_k wird auch die Notation

$$\mathcal{F}[f_0,\ldots,f_{n-1}] := [\beta_0,\ldots,\beta_{n-1}] \tag{3.45}$$

verwendet.
(3.44) bedeutet das Gleichungssystem

$$\underbrace{\begin{pmatrix} \omega_0^0 & w_0^1 & \ldots & \omega_0^{n-1} \\ \vdots & & & \\ \omega_{n-1}^0 & w_{n-1}^1 & \ldots & \omega_{n-1}^{n-1} \end{pmatrix}}_{=:V=(\omega_k^j)_{j,k=0,\ldots,n-1}} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{n-1} \end{pmatrix} = \begin{pmatrix} f_0 \\ \vdots \\ f_{n-1} \end{pmatrix} \tag{3.46}$$

bzw.

$$\begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{n-1} \end{pmatrix} = \underbrace{\frac{1}{n}\begin{pmatrix} \omega_0^{-0} & w_0^{-1} & \ldots & \omega_0^{-(n-1)} \\ \vdots & & & \\ \omega_{n-1}^{-0} & w_{n-1}^{-1} & \ldots & \omega_{n-1}^{-(n-1)} \end{pmatrix}}_{=\frac{1}{n}\bar{V}=(\frac{1}{n}\omega_k^{-j})_{j,k=0,\ldots,n-1}} \begin{pmatrix} f_0 \\ \vdots \\ f_{n-1} \end{pmatrix} \tag{3.47}$$

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 (3.44))

$$f_j = \sum_{k=0}^{n-1} \beta_k e^{i\frac{jk2\pi}{n}}, \quad j = 0, \ldots, n-1$$

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

$$\sum_{k=0}^{n-1} |\beta_k|^2 = \frac{1}{n} \sum_{k=0}^{n-1} |f_k|^2$$

Beziehungen zwischen den reellen und komplexen Fourierkoeffizienten A_j, B_j, β_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

$$\begin{aligned}
\beta_l &= \frac{1}{2}(A_l + \frac{1}{i} B_l) = \frac{1}{2}(A_l - iB_l), \quad l = 1, \ldots, m \\
\beta_{n-l} &= \frac{1}{2}(A_l - \frac{1}{i} B_l) = \frac{1}{2}(A_l + iB_l), \quad l = 1, \ldots, m
\end{aligned}$$

$$\rightsquigarrow \quad A_l = \beta_l + \beta_{n-l}, \quad B_l = i(\beta_l - \beta_{n-l}), \quad l = 1, \ldots, m$$

Mit der Formel (3.43) folgt:

$$\begin{aligned} A_l &= \frac{1}{n} \sum_{k=0}^{n-1} f_k \left(e^{-i\frac{kl2\pi}{n}} + e^{-i\frac{k(n-l)2\pi}{n}} \right) \\ &= \frac{2}{n} \sum_{k=0}^{n-1} f_k \frac{1}{2} \left(e^{-i\frac{kl2\pi}{n}} + e^{i\frac{kl2\pi}{n}} \right) = \frac{2}{n} \sum_{k=0}^{n-1} f_k \cos(lx_k) \end{aligned} \tag{3.48}$$

und analog

$$B_l = \frac{2}{n} \sum_{k=0}^{n-1} f_k \sin(l_{xk})$$

Die Betrachtungen für gerades $n = 2m$ verlaufen analog.
Zusammengefasst ergibt sich

Satz 3.29. *Werden die Koeffizienten gemäß* (3.48) *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.

13. Vorlesung am 27.05.09

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.

3.12 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 3.30. *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 diskreten Fouriertransformierten des Datensatzes

$$g_0, \ldots, g_{2M-1}$$

der Länge $2M$ folgendermaßen bestimmen.

$$\begin{aligned} &\frac{1}{2}\left\{\mathcal{F}_k[g_0, \ldots, g_{M-1}] + e^{-i\frac{k\pi}{M}}\mathcal{F}_k[g_M, \ldots, g_{2M-1}]\right\} \\ &= \mathcal{F}_k[g_0, g_M, g_1, g_{M+1} \ldots, g_{2M-1}] \end{aligned} \tag{3.49}$$

$$\begin{aligned} &\frac{1}{2}\left\{\mathcal{F}_k[g_0, \ldots, g_{M-1}] - e^{-i\frac{k\pi}{M}}\mathcal{F}_k[g_M, \ldots, g_{2M-1}]\right\} \\ &= \mathcal{F}_{M+k}[g_0, g_M, g_1, g_{M+1} \ldots, g_{2M-1}] \end{aligned} \tag{3.50}$$

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

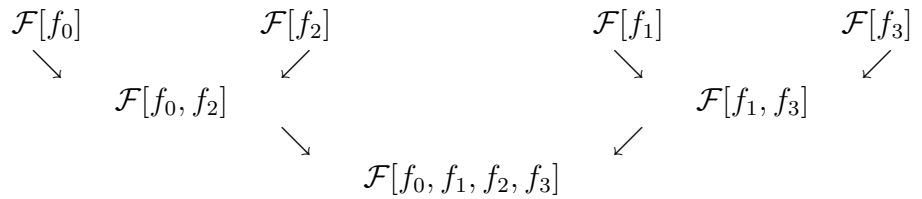
$$\begin{aligned} \mathcal{F}_k[g_0, \ldots, g_{2M-1}] &= \frac{1}{2M}\left(\sum_{j=0}^{M-1} g_j e^{-i\frac{2jk2\pi}{2M}} + \sum_{j=0}^{M-1} g_{M+j} e^{-i\frac{(2j+1)k2\pi}{2M}}\right) \\ &= \frac{1}{2M}\left(\sum_{j=0}^{M-1} g_j e^{-i\frac{jk2\pi}{M}} + e^{-i\frac{k\pi}{M}} \sum_{j=0}^{M-1} g_{M+j} e^{-i\frac{jk2\pi}{M}}\right) \end{aligned}$$

Die Gleichung (3.50) 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 3.30 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 (3.49), (3.50) 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 (3.49), (3.50))

(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 3.30 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 (3.49),(3.50)

$$\mathcal{F}[g_0, g_2, g_1, g_3]$$

also bei Berücksichtigung der Setzungen

$$\mathcal{F}[f_0, f_1, f_2, f_3]$$

Bemerkung 3.31. 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.

3.12.1 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 $\omega^2, \ldots, \omega^{2^r-1} \in \mathbb{C}$ (ω Wert einer komplexen Exponentialfunktion) erfordert $2^r - 2 \leq 2^r$ komplexe Multiplikationen (Faktoren in den Formeln (3.49), (3.50))
- 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 3.32. *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 3.33. Wir haben für die Fouriertransformation die Formeln

$$\beta_j = \frac{1}{n} \sum_{k=0}^{n-1} f_k e^{-i\frac{jk2\pi}{n}} \tag{3.51}$$

$$f_j = \frac{1}{n} \sum_{k=0}^{n-1} \beta_k e^{i\frac{jk2\pi}{n}}, \quad j = 0, \ldots, n-1$$

für die Hin- resp. Rücktransformation hergeleitet. In vielen Lehrbüchern sind die diskreten Fourierkoeffizienten durch

$$\beta_j = \sum_{k=0}^{n-1} f_k e^{-i\frac{jk2\pi}{n}} \tag{3.52}$$

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

$$\beta_j = \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} f_k e^{-i\frac{jk2\pi}{n}} \tag{3.53}$$

$$f_j = \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} \beta_k e^{i\frac{jk2\pi}{n}}, \quad j = 0, \ldots, n-1$$

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 (3.51), (3.52) oder (3.53).

Kapitel 4

Numerische Integration

14. Vorlesung am 01.06.2009

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.

4.1 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

$$\int_a^b p_n(x)\mathrm{d}x = (b-a)\sum_{k=0}^{n} f_k\sigma_k \tag{4.1}$$

mit den Gewichten

$$\sigma_k = \frac{1}{n}\int_0^n \prod_{k\neq j=0}^{n} \frac{s-j}{k-j}\mathrm{d}s, \quad k = 0,\ldots,n \tag{4.2}$$

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

$$\int_a^b f(x)\mathrm{d}x \approx \int_a^b p_1(x)\mathrm{d}x = \frac{b-a}{2}(f(a)+f(b)) \tag{4.3}$$

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

$$\int_a^b f(x)\mathrm{d}x \approx \int_a^b p_2(x)\mathrm{d}x = \frac{b-a}{6}\left(f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right) \tag{4.4}$$

Die **Simpson-Regel**, auch **Keplersche Fassregel** genannt, folgt.
Für $n = 3$ findet man auf analoge Weise mit

$$\begin{aligned}\int_a^b f(x)\mathrm{d}x &\approx \int_a^b p_3(x)\mathrm{d}x \\ &= \frac{b-a}{8}\left(f(a)+3f\left(\frac{2a+b}{3}\right)+3f\left(\frac{a+2b}{3}\right)+f(b)\right)\end{aligned} \tag{4.5}$$

die Newtonsche $\frac{3}{8}$-Regel.

Definition 4.1. *Die Näherungsformel*

$$Q_n(x) = \int_a^b p_n(x)\mathrm{d}x = (b-a)\sum_{k=0}^{n} f(x_k)\sigma_k \tag{4.6}$$

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 4.2. *Mit*

$$E_n[f] = \int_a^b f(x)\mathrm{d}x - Q_n = I - Q_n \tag{4.7}$$

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 4.3. *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.

4.2 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 σ_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 ($h = \frac{b-a}{n}$)

$$\begin{aligned} S_{1,N} &= h\left[\frac{1}{2}y_{10} + y_{20} + \cdots + y_{N0} + \frac{1}{2}y_{N1}\right] \\ &= h\left[\frac{1}{2}(y_{10} + y_{N1}) + \sum_{k=2}^{N} y_{k0}\right] \end{aligned} \tag{4.8}$$

und für $n = 2$ die aufsummierte Simpson-Regel ($h = \frac{b-a}{2N}$)

$$S_{2,N} = \frac{h}{3}\left[(y_{10} + y_{N1}) + 2\sum_{j=1}^{N-1} y_{j2} + 4\sum_{j=1}^{N} y_{j1}\right] \tag{4.9}$$

Für die Quadraturfehler summierter abgeschlossener Newton-Cotes- Formeln gilt der

Satz 4.4. *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 $\xi \in]a, b[$, 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 α, β abhängige Konstanten sind, und $h = \frac{b-a}{nN}$ gilt.

Beweis. Plato, Bärwolff □

4.3 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 σ_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 St ützstellen 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 σ_k so zu bestimmen, dass

$$I_n = \sum_{j=1}^{n} \sigma_j f(x_j) \tag{4.10}$$

eine möglichst gute Näherung des Integrals I ergibt. Fordert man, dass die Formel (4.10) 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

$$\sum_{j=1}^{n} \sigma_j x_j^k = \int_a^b x^k \rho(x) \mathrm{d}x \quad (k = 0, 1, \ldots, 2n-1) \tag{4.11}$$

sein.
Wir werden im Folgenden zeigen, dass das Gleichungssystem (4.11) eindeutig lösbar ist, dass für die Stützstellen $x_k \in]a, b[$ gilt und dass die Gewichte σ_k positiv sind.

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 (4.11) 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

$$\begin{aligned} \sigma_1 + \sigma_2 &= 2 \\ \sigma_1 x_1 + \sigma_2 x_2 &= 0 \\ \sigma_1 x_1^2 + \sigma_2 x_2^2 &= \frac{2}{3} \\ \sigma_1 x_1^3 + \sigma_2 x_2^3 &= 0 \end{aligned} \tag{4.12}$$

Für (4.12) 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.

4.4 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

$$\langle p, q \rangle_\rho := \int_a^b p(x)q(x)\rho(x)\mathrm{d}x \tag{4.13}$$

für $p, q \in P$ aus. Folglich ist durch

$$\|p\|_\rho^2 = \langle p, p \rangle_\rho = \int_a^b p^2(x)\rho(x)\mathrm{d}x \tag{4.14}$$

eine Norm definiert. Der Nachweis, dass (4.13), (4.14) Skalarprodukt bzw. Norm sind, sollte als Übung betrachtet werden.

Definition 4.5. *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 $K = \mathbb{R}$)

4.4.1 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

$$p_n(x) = x^n - \sum_{j=0}^{n-1} \frac{\langle x^n, p_j \rangle_\rho}{\langle p_j, p_j \rangle_\rho} p_j(x) \tag{4.15}$$

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

$$p_1(x) = x, \quad p_2(x) = x^2 - \frac{1}{3}, \quad p_3(x) = x^3 - \frac{3}{5}x, \quad p_4(x) = x^4 - \frac{5}{2}x^2 + \frac{4}{105} \tag{4.16}$$

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 4.6. 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 (4.15) ausgehend von $p_0 = 1$ mit

$$p_0(x) = 1, \quad p_1(x) = x, \quad p_2(x) = x^2 - \frac{1}{2}, \quad p_3(x) = x^3 - \frac{3}{4}x \tag{4.17}$$

die orthogonalen **Tschebyscheff-Polynome**.

Sowohl bei den Legendre- als auch bei den Tschebyschef-Polynomen findet man jeweils einfach reelle Nullstellen, die im Intervall $]a, b[$ liegen. Generell gilt der

Satz 4.7. *Die Nullstellen des n-ten Orthogonalpolynoms bezüglich eines Intervalls $[a, b]$ und einer Gewichtsfunktion ρ sind einfach, reell und liegen im Intervall $]a, b[$*

Beweis. Plato □

Nun kommen wir zur Definition der Gauß-Quadratur

Definition 4.8. *Mit $x_1, \ldots, x_n$ seien die Nullstellen des n-ten Orthogonalpolynoms $p_n(x)$ gegeben. Die numerische Integrationsformel*

$$I_n = \sum_{j=1}^{n} \sigma_j f(x_j) \quad mit \quad \sigma_j = \langle L_j, 1 \rangle_\rho = \int_a^b L_j(x)\rho(x)\mathrm{d}x \tag{4.18}$$

heißt Gaußsche Quadraturformel der n-ten Ordnung oder kurz Gauß-Quadratur zur Gewichtsfunktion ρ

Im Folgenden wird gezeigt, dass die Stützstellen x_k und Gewichte σ_k als Lösung des Gleichungssystems (4.11) gerade die Nullstellen des n-ten Orthogonalpolynoms $p_n(x)$ bzw. die Gewichte gemäß (4.18) sind und damit die Gleichwertigkeit der Formeln (4.10) und (4.18) nachgewiesen.

Satz 4.9. *Mit $x_1, \ldots, x_n$ seien die Nullstellen des n-ten Orthogonalpolynoms $p_n(x)$ gegeben.*
Es existiert eine eindeutig bestimmte Gauß-Quadratur (4.18)*. Bei der Gauß-Quadratur sind alle Gewichte gemäß* (4.18) *positiv und die Quadratur ist für jedes Polynom vom Grad $m \leq 2n-1$ exakt, d.h. es gilt*

$$\int_a^b p(x)\rho(x)\mathrm{d}x = \langle p, 1 \rangle_\rho = \sum_{j=1}^{n} \sigma_j p(x_j), \quad \forall p \in \Pi_{2n-1} \tag{4.19}$$

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

$$\int_a^b p(x)\rho(x)\mathrm{d}x = \langle p, 1 \rangle_\rho = \langle r, 1 \rangle_\rho$$
$$= \sum_{j=1}^{n} p(x_j) \langle L_j, 1 \rangle_\rho = \sum_{j=1}^{n} \sigma_j p(x_j)$$

Für $p(x) = L_j^2(x) \in \Pi_{2n-2}$ ergibt die eben nachgewiesene Formel (4.19)

$$0 < \|L_j\|_\rho^2 = \left\langle L_j^2, 1 \right\rangle_\rho = \sum_{k=1}^{n} \sigma_k L_j^2(x_k) = \sigma_j$$

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

$$I_n^* = \sum_{j=1}^{n} \sigma_j^* f(x_j^*) \tag{4.20}$$

existiert mit $x_k^* \neq x_j^*$ für $k \neq j$, deren Genauigkeitsgrad gleich $2n-1$ ist. Die Positivität der σ_j^* wird analog der Positivität der σ_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 (4.20) den exakten Wert des Integrals für $h(x)$, also

$$\int_a^b h(x)\rho(x)\mathrm{d}x = \int_a^b L_k^*(x) p_n(x)\rho(x)\mathrm{d}x$$
$$= \sum_{j=1}^{n} \sigma_j^* L_k(x_j^*) p_n(x_j^*) = \sigma_k^* p_n(x_k^*)$$

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 4.10. *Mit den Stützstellen und Gewichten aus Satz 4.9 gilt für auf dem Intervall $[a, b]$ $2n$-mal stetig diffbare Funktionen $f(x)$*

$$\int_a^b f(x)\rho(x)\mathrm{d}x - \sum_{j=1}^{n} \sigma_j f(x_j) = \frac{\|p_n\|_\rho^2}{(2n)!} f^{(2n)}(\xi) \tag{4.21}$$

mit einem Zwischenwert $\xi \in]a, b[$.

16. Vorlesung 08.06.09

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) ohne die Polynome auszurechnen. Das ist bei den anderen Polynomen aus der Tabelle nicht möglich.

Kapitel 5

Iterative Lösung von Gleichungssystemen

Im Folgenden geht es darum, Gleichungen oder Gleichungssysteme zu lösen. Ist G eine Abbildung aus dem $\mathbb{R}^n$ in den $\mathbb{R}^n$, bedeutet

$$G(x) = 0 \tag{5.1}$$

gerade ein Gleichungssystem zur Bestimmung einer Nullstelle $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ der Abbildung G. Definiert man ausgehend von G die Abbildung

$$F(x) := G(x) + x$$

Dann ist die Lösung von (5.1) gleichbedeutend mit der Bestimmung eines Fixpunktes x von F, also

$$F(x) = x \Leftrightarrow G(x) = 0 \tag{5.2}$$

Wenn man keinerlei Vorstellung von der Lösung der Gleichung (5.2) hat, findet man mit der Folge

$$(x_k), \quad x_0 \in D \subset \mathbb{R}^n, \quad x_{k+1} = F(x_k), \quad k = 0, 1, \ldots$$

eine Folge, die, wenn sie konvergiert, im Falle der Stetigkeit der Abbildung gegen einen Fixpunkt von F kovergiert.
Die Grundlagen der iterativen Lösung von Gleichungen (Gleichungssysteme sollen nur nür für den Fall $n = 1$ dargestellt werden.

Definition 5.1. *Sei $f : I \to I$ eine Funktion, die das reelle Intervall I in sich abbildet. Jede Lösung der Gleichung*

$$x = f(x) \tag{5.3}$$

*heißt **Fixpunkt** von f. Die Gleichung* (5.3) *wird Fixpunktgleichung genannt.*

Definition 5.2. *Eine auf einer Teilmenge $D \subset \mathbb{R}$ definierte Funktion $f : D \to \mathbb{R}$ heißt* ***Kontraktion****, wenn eine Konstante $L \in [0, 1[$ existiert, sodass für alle $x_1, x_2 \in D$*

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

mit einer von x_1, x_2 unabhängigen Konstanten $L < 1$, d.h. f ist Kontraktion.

Satz 5.3. *Sei $f : I \to I$ eine reellwertige Funktion, die ein abgeschlossenes Intervall I in sich abbildet und es gelte für alle $x_1, x_2 \in I$ die Ungleichung*

$$|f(x_1) - f(x_2)| \leq L\,|x_1 - x_2| \tag{5.4}$$

mit einer von x_1, x_2 unabhängigen Konstanten $L < 1$, d.h. f ist Kontraktion. Dann hat f genau einen Fixpunkt $\hat{x} \in I$ und die durch die Fixpunktiteration $x_{k+1} = f(x_k)$ definierte Iterationsfolge konvergiert für jeden Anfangspunkt $x_0 \in I$ gegen diesen Fixpunkt

Beweis. Analysis □

Bemerkung 5.4 (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 5.5. Aus dem Banachschen Fixpunktsatz ergeben sich die Fehlerabschätzungen

$$\|x_k - \hat{x}\| \leq \frac{L^k}{1 - L}\,\|x_1 - x_0\| \quad \text{A-priori-Abschätzung} \tag{5.5}$$

$$\|x_k - \hat{x}\| \leq \frac{1}{1 - L}\,\|x_{k+1} - x_k\| \quad \text{A-posteriori-Abschätzung} \tag{5.6}$$

Die Kontraktivität ist essentiell für die Berechnung von Fixpunkten. Unter bestimmten Voraussetzungen kann man die Existenz einer Kontraktion zeigen.

Satz 5.6. *Es sei $G \subset \mathbb{R}$ offen und $f : G \to \mathbb{R}$ stetig diff'bar mit einem Fixpunkt $\hat{x} \in G$. Wenn $|f'(x)| < 1$ 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 5.7. Ist die Voraussetzung $|f'(x)| < 1$ des Satzes 5.6 nicht erfüllt, findet man keine Kontraktion. Ist $|f'(x)| > 1$ dann gilt in der Nähe von $\hat{x}$

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

das rechtfertigt die

Definition 5.8. *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.*

5.1 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

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, 2, \ldots \tag{5.7}$$

eine Newton-Folge, die im Falle der Konvergenz gegen eine Nullstelle von f geht. Die Folge (5.7) kann man auch anders erhalten. Man definiert die Hilfsfunktion

$$g(x) = x - \frac{f(x)}{f'(x)} \tag{5.8}$$

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 5.9. *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 (5.7) 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 5.3 und dem Satz 5.6 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 5.10. Die Voraussetzungen des Satzes 5.9 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.

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

Satz 5.11. *$F : D \to \mathbb{R}^n, D \subset \mathbb{R}^n$ 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

$$x_{k+1} = x_k - [F'(x_k)]^{-1} F(x_k), \quad k = 0, 1, 2, \ldots \tag{5.9}$$

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 5.9 im eindimensionalen Fall. □

Beispiel. Zur Bestimmung der Stützstellen (Nullstellen des n-ten Tschebyscheff-Polynoms) $x_1, \ldots, x_n$ und der Integrationsgewichte $\sigma_1, \ldots, \sigma_n$ ist das Gleichungssystem

$$\sum_{k=1}^{n} \sigma_k x_k^j = \int_{-1}^{1} x^j \mathrm{d}x, \quad j = 0, 1, \ldots, 2n-1 \tag{5.10}$$

zu lösen. Mit $b_j = \int_{-1}^{1} x^j \mathrm{d}x = \frac{1}{j+1}(1-(-1)^{j+1})$ bedeutet, dass die Nullstelle der Abbildung $F : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$

$$F(\sigma_1, \ldots, \sigma_n, x_1, \ldots, x_n) = \begin{pmatrix} \sigma_1 x_1^0 & + & \cdots & + & \sigma_n x_n^0 & - & b_0 \\ & & & \vdots & & & \\ \sigma_1 x_1^{2n-1} & + & \cdots & + & \sigma_n x_n^{2n-1} & - & b_{2n-1} \end{pmatrix}$$

zu bestimmen. Mit der Ableitungsmatrix $F' = (J_{ik})$

$$J_{ik} = \begin{cases} x_k^{i-1}, & 1 \le k \le n, i = 1, \ldots, 2n \\ 0, & i = 1, n+1 \le k \le 2n \\ \sigma_{k-n}(i-2)x_{k-n}^{i-2}, & i = 2, \ldots, 2n, n+1 \le k \le 2n \end{cases}$$

kann man mit dem Newton-Verfahren (5.9) versuchen, das Gleichungssystem (5.10) zu lösen.

5.3 Gedämpftes Newton-Verfahren

Betrachtet man

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

mit $\alpha \in]0, 1[$, spricht man 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.

5.4 Die iterative Lösung linearer Gleichungssysteme

17. Vorlesung 10.06.09

Neben der schon beschriebenen direkten Lösung linearer Gleichungssysteme durch den Gaußschen Algorithmus oder durch bestimmte Matrix-Faktorisierungen ist es oft sinnvoll, lineare Gleichungssysteme

$$Ax = b \tag{5.11}$$

mit der regulären Matrix a vom Typ $n \times n$ und $b \in \mathbb{R}^n$ iterativ zu lösen (oBdA 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 (5.11).

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

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

$$x_k = (E - B^{-1}A)x_{k-1} + B^{-1}b, \quad k = 1, 2, \ldots \tag{5.12}$$

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 (5.11). Die Lösung ist ein Fixpunkt der Abbildung

$$F : \mathbb{R}^n \to \mathbb{R}^n : x \mapsto (E - B^{-1}A)x + B^{-1}b \tag{5.13}$$

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

$$\|\delta x_k\| \leq \left\|S^k\right\| \|\delta x_0\| \tag{5.14}$$

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 5.12. *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* λ *mit* $|\lambda| \geq 1$ *regulär*

Beweis. (Auszugsweise) a ⇒ 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 λ_i, sodass wegen $|\lambda_i| < 1$ für alle Eigenwerte $\left\|S^k\right\|$ gegen null geht.
a ⇒ c Mit der Zeilensummennorm gilt wegen der Voraussetzung zum Spektralradius von S, der gleich dem von J ist, und der Wahl von ϵ

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

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 ⇒ 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 5.12 erhält man das folgende Konvergenzkriterium

Satz 5.13. *Seien A, B reguläre $(n \times n)$-Matrizen. Die Iteration* (5.12) *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*

$$\|x_k - x\| \leq Cr^k, \quad C = const \in \mathbb{R} \tag{5.15}$$

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 $U = (u_{ij})$

$$A = L + D + U \tag{5.16}$$

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 (5.12) ist für den Aufwand natürlich die einfache Invertierbarkeit von B entscheidend. Das wird bei den nun zu diskutierenden Verfahren auch berücksichtigt.

5.5 Jacobi-Verfahren oder Gesamtschrittverfahren

Die Wahl von $B = D$ ergibt die Iterationsmatrix

$$S = E - B^{-1}A = -D^{-1}(L + U) = \begin{pmatrix} 0 & \frac{a_{12}}{a_{11}} & \cdots & \frac{a_{1n}}{a_{11}} \\ \frac{a_{21}}{a_{22}} & 0 & & \\ \vdots & & \ddots & \\ \frac{a_{n1}}{a_{nn}} & & & 0 \end{pmatrix} \tag{5.17}$$

Das Verfahren (5.12) mit der durch die Wahl von $B = D$ definierten Iterationsmatrix (5.17) 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 $x_j^{(k)}, j = 1, \ldots, n$. Damit ergibt sich für die Jacobi-Verfahren koordinatenweise

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

Zur Konvergenz des Jacobi-Verfahrens gilt der

Satz 5.14. *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

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

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

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

und es gibt einen Index l mit

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

(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.

Bemerkung. 1. Man kann entscheiden, ob eine Matrix reduzibel oder irreduzibel ist, indem man für $A = (a_{ij})_{i,j=1,\ldots,n}$ 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 5.16. *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 5.17. *Für eine irreduzible, schwach diagonal dominante Matrix A ist das Jacobi-Verfahren konvergent.*

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

5.6 Gauß-Seidel-Iterationsverfahren oder Einzelschrittverfahren

Wählt man ausgehend von der Matrixzerlegung (5.16) $B = L+D$, dann heißt das Iterationsverfahren (5.12) Gauß-Seidel- Verfahren oder Einzelschrittverfahren, d.h. es ergibt sich

$$x^{(k)} = \underbrace{(E - B^{-1}A)}_{S} x^{(k-1)} + B^{-1}b = (L+D)^{-1}(-Ux^{(k-1)} + b), \quad k = 1, 2, \ldots \tag{5.18}$$

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 5.18. *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 λ 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:

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

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

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

für alle EW λ folgt $(r(S) < 1)$ □

Bemerkung 5.19. Ebenso wie beim Jacobi-Verfahren kann man die Voraussetzung der strikten Diagonaldominanz von A abschwächen.
Das Gauß-Seidel-Verfahren ist für irreduzible schwach diagonal dominante Matrizen A konvergent.

Wenn man das Gauß-Seidel Verfahren (5.18) in der äquivalenten Form

$$x^{(k)} = D^{-1}(-Lx^{(k)} - Ux^{(k-1)} + b), \quad k = 1, 2, \ldots \tag{5.19}$$

aufschreibt, erkennt man bei der koordinatenweisen Berechnung der neuen Iteration

$$x_j^{(k)} = \frac{1}{a_{jj}}\left(b_j - \sum_{i=1}^{j-1} a_{ji}x_i^{(k)} - \sum_{i=j+1}^{n} a_{ji}x_i^{(k-1)}\right), \quad j = 1, 2, \ldots, n, \ k = 1, 2, \ldots \tag{5.20}$$

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 (5.20) für $j = 1, \ldots, n$ sukzessiv zum Update der Koordinaten von x anwenden. Man hat also mit (5.20) eine explizite Berechnungsvorschrift und braucht damit $B = L + D$ nicht wirklich zu invertieren.

5.7 Verallgemeinerung des Gauß-Seidel-Verfahrens

Wählt man $S = S_\omega = (D + \omega L)^{-1}[(1-\omega)D - \omega U] = E - \omega(D + \omega L)^{-1}A$, bzw. $B = \omega(D + \omega L)^{-1}$, dann hat man mit

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

das Gauß-Seidel-Verfahren mit Relaxation erklärt. Für $\omega > 1$ spricht man von 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.

5.8 Krylov-Raum-Verfahren zur Lösung linearer Gleichungssysteme

18. Vorlesung 12.06.09

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

$$\{0\} \subset D_1 \subset \ldots \subset \mathbb{R}^n \tag{5.21}$$

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 5.20.

(a) Für gegebene Ansatzräume (5.21) *hat der* ***Ansatz des orthogonalen Residuums*** *zur Bestimmung von Vektoren $x_1, x_2, \ldots \in \mathbb{R}^n$ die Form*

$$\left.\begin{array}{c} x_k \in D_k \\ Ax_k - b \in D_k^\perp \end{array}\right\} k = 1, 2, \ldots \tag{5.22}$$

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

$$\left.\begin{array}{c} x_k \in D_k \\ \|Ax_k - b\|_2 \text{ minimal} \end{array}\right\} k = 1, 2, \ldots \tag{5.23}$$

Bei der Wahl spezieller Ansatzräume (5.21) werden die sogenannten Kry-lovräume von Bedeutung sein

Definition 5.21. *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 5.20 angegebenen Ansätze mit den speziellen Räumen $D_k = K_k(A,b)$ betrachtet, wobei wir den Schwerpunkt auf den Ansatz (5.22) legen.

5.8.1 Der Ansatz des orthogonalen Residuums (5.22) für symmetrische positiv definite Matrizen

Für positiv definite, symmetrische Matrizen soll nun Existenz und Eindeutigkeit von Vektoren x_k für (5.22) diskutiert werden. Dazu werden die Skalarprodukte und Normen

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

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 5.22. *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* (5.22) *– mit allgemeinen Ansatzräumen D_k gemäß* (5.21) *– eindeutig bestimmt, und es gilt*

$$\|x_k - x_*\|_A = \min_{x\in D_k} \|x - x_*\|, \quad k = 1, 2, \ldots \tag{5.24}$$

Beweis. Eindeutigkeit: Sei k fest gewählt. Für $x_k, \hat{x}_k$ mit der Eigenschaft (5.22) 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

$$x_k = \sum_{j=0}^{m-1} \alpha_j d_j \tag{5.25}$$

an und erhält

$$\begin{aligned} x_k \text{ genügt (5.22)} &\Leftrightarrow Ax_k - b \in D_k^{\perp} \\ &\Leftrightarrow \langle Ax_k - b, d_k \rangle_2 = 0 \quad k = 0, \ldots, m-1 \qquad (5.26) \\ &\Leftrightarrow \sum_{j=0}^{m-1} \langle Ad_j, d_k \rangle_2 \, \alpha_j = \langle b, d_k \rangle_2 \,, \quad k = 0, \ldots, m-1 \quad (5.27) \end{aligned}$$

(5.27) ist ein lineares Gleichungssystem von m Gleichungen für die Koeffizienten $\alpha_0, \ldots, \alpha_{m-1}$. Da x_k mit (5.22) eindeutig bestimmt ist (wurde schon gezeigt), ist das Gleichungssystem (5.27) eindeutig lösbar, woraus die Existenz von x_k folgt.
Minimalität (5.24) Für $x \in D_k$ findet man

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

□

5.8.2 Der Ansatz des orthogonalen Residuums (5.22) für gegebene A-konjugierte Basen

Mit dem Beweis von Satz 5.22 ist bereits eine Möglichkeit zur Bestimmung von x_k für (5.22) ausgehend von einer Basis $d_0, \ldots, d_{m-1}$ für D_k mit dem Gleichungssystem (5.27) aufgezeigt worden. Im Folgenden wird ein Spezialfall behandelt, bei dem (5.27) Diagonalgestalt hat.

Definition 5.23. *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 (5.27) Diagonalgestalt und damit ist x_k gemäß Ansatz (5.25) sehr einfach berechenbar.

Satz 5.24. *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 (5.22) *die folgenden Darstellungen für* $k = 1, 2, \ldots$

$$x_k = \sum_{j=0}^{k-1} \alpha_j d_j \quad mit \quad \alpha_j = -\frac{\langle r_j, d_j \rangle_2}{\langle A d_j, d_j \rangle_2} \tag{5.28}$$

$$r_j := Ax_j - b, \quad j \geq 1, r_0 = -b \tag{5.29}$$

Beweis. Folgt unmittelbar für $k = m$ aus (5.25)-(5.27) □

Bemerkung.

(a) Aus (5.28) folgt die Unabhängigkeit der α_j von k und damit gilt

$$x_{k+1} = x_k + \alpha_k d_k, \quad r_{k+1} = r_k + \alpha_k A d_k, \quad k = 0, 1, \ldots ; x_0 = 0 \tag{5.30}$$

(b) Aufgrund der ersten Identität von (5.30) bezeichnet man d_k als Suchrichtung und α_k als Schrittweite

(c) Außerderm wird mit (5.30) 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.

5.8.3 Das CG-Verfahren für positiv definite, symmetrische Matrizen

19. Vorlesung 17.06.09

Definition 5.25. *Zu gegebener symmetrisch positiv definiter Matrix* $A \in \mathbb{R}^{n \times n}$ *ist das* ***Verfahren des konjugierten Gradienten*** *gegeben durch den Ansatz* (5.22) *mit der speziellen Wahl*

$$D_k = K_k(A, b), \quad k = 0, 1, \ldots \tag{5.31}$$

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 5.24 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 5.26 herausstellt, genügt hier eine Gram-Schmidt-Orthogonalisierung der beiden Vektoren $d_{k-1}, -r_k \in \mathbb{R}^n$.

Lemma 5.26. *Zu gegebener symmetrisch positiv definiter Matrix $A \in \mathbb{R}^{n\times n}$ und mit dern Notationen des Satzes 5.24 seien die Suchrichtungen speziell wie folgt gewählt:*

$$d_0 = b, \quad d_k = -r_k + \beta_{k-1} d_{k-1}, \quad \beta_{k-1} = \frac{\langle Ar_k, d_{k-1}\rangle_2}{\langle Ad_{k-1}, d_{k-1}\rangle_2}, \quad k = 1, \ldots, k_* - 1 \tag{5.32}$$

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

$$\operatorname{span}\{d_0, \ldots, d_{k-1}\} = \operatorname{span}\{b, r_1, \ldots, r_{k-1}\} = K_k(A,b), \quad k = 1, \ldots, k_* \tag{5.33}$$

Beweis. Vollständige Induktion über $k = 1, \ldots, k_s$ zum Nachweis der A-Konjugiertheit der Vektoren $d_0, \ldots, d_{k-1} \in \mathbb{R}^n$ und der Formeln (5.32) wegen

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

ist der Induktionsanfang gemacht.
Im Folgenden sei angenommen, dass (5.32) ein System von A- konjugierten Vektoren mit der Eigenschaft (5.33) liefert mit einem fixierten Index $1 \le k \le k_* - 1$
Gemäß dem An satz des orthogonalen Residuums (5.22) 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

$$d_k = -r_k + \sum_{j=0}^{k-1} \frac{\langle Ar_k, d_j\rangle_2}{\langle Ad_j, d_j\rangle_2} d_j \overset{(*)}{=} -r_k + \beta_{k-1} d_{k-1} \tag{5.34}$$

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 (5.30) gilt 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 5.26 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 (5.33) 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 5.27. *In der Situation des Lemma 5.26 gelten die Darstellungen*

$$d_k = \frac{\|r_k\|_2^2}{\langle Ad_k, d_k\rangle_2}, \quad k = 0, 1, \ldots, k_* - 1 \tag{5.35}$$

$$\beta_{k-1} = \frac{\|r_k\|_2^2}{\|r_{k-1}\|_2^2}, \quad k = 1, \ldots, k_* - 1 \quad (r_0 := b) \tag{5.36}$$

Beweis. Mit $r_k \in K_k(A, b)^\perp$ sowie der Beziehung (5.34) für die Suchrichtung d_k erhält man $-\langle r_k, d_k\rangle_2 = \|r_k\|_2^2$ und zusammen mit (5.28) ergibt dies (5.35). Diese Darstellung (5.35) für α_k zusammen mit der Identität $r_k = r_{k-1} + \alpha_{k-1} Ad_{k-1}$ aus (5.30) liefert

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

und damit gilt für β_{k-1} die Beziehung (5.36) □

5.8.4 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 ($k = 1, 2, \dots$). Hilfreich ist

Lemma 5.28. *Zu gegebener symmetrisch positiv definiter Matrix $A \in \mathbb{R}^{n\times n}$ sei $(\lambda_j, v_j)_{j=1,\dots,n}$ ein vollständiges System von Eigenwerten $\lambda_j > 0$ und zugehörigen Eigenvektoren $v_j \in \mathbb{R}^n$, also gilt*

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

Mit der Entwicklung

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

gelten für jedes Polynom p die folgenden Darstellungen

$$p(A)x = \sum_{j=1}^{n} c_j p(\lambda_j) v_j \tag{5.37}$$

$$\|p(A)x\|_2 = \left(\sum_{j=1}^{n} c_j^2 p(\lambda_j)^2\right)^{\frac{1}{2}}, \quad \|p(A)x\|_A = \left(\sum_{j=1}^{n} c_j^2 \lambda_j p(\lambda_j)^2\right)^{\frac{1}{2}} \tag{5.38}$$

Speziell gilt also

$$m^{\frac{1}{2}} \|x\|_2 \le \|x\|_A \le M^{\frac{1}{2}} \|x\|_2, \quad x \in \mathbb{R}^n \tag{5.39}$$

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

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

$$Ax = \sum_{j=1}^{n} c_j \lambda_j^\nu v_j, \quad \nu = 0, 1, \dots$$

und daraus folgt (5.37). 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 5.29. *Zu einer gegebenen symmetrisch positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$ gelten für das CG-Verfahren die folgenden Fehlerabschätzungen:*

$$\|x_k - x_*\|_A \leq \left(\inf_{p\in\Pi_k, p(0)=1} \sup_{\lambda\in\sigma(A)} |p(\lambda)| \right) \|x_*\|_A \tag{5.40}$$

Beweis. Für jedes Polynom $p \in \Pi_k$ mit $p(0) = 0$ 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 5.28 und der Entwicklung $x_* = \sum_{j=1}^{n} c_j v_j \in \mathbb{R}^n$ erhält man

$$\begin{aligned} \|x_k - x_*\|_A &\overset{(5.24)}{\leq} \underbrace{\|x - x_*\|_A}_{=\|p(A)x_*\|_A} = \left(\sum_{j=1}^{n} c_j^2 \lambda_j p(\lambda_j)^2 \right)^{\frac{1}{2}} \\ &\leq \sup_{\lambda\in\sigma(A)} |p(\lambda)| \left(\sum_{j=1}^{n} c_j^2 \lambda_j \right)^{\frac{1}{2}} = \sup_{\lambda\in\sigma(A)} |p(\lambda)| \, \|x_*\|_A \end{aligned}$$

□

Zur quantitativen Präzisierung der Abschätzung (5.40) des Satzes 5.29 benutzen wir die hier nicht bewiesenen Eigenschaften der Tschebyscheff-Polynome erster Art $T_0, T_1, \ldots$

$$T_k\left(\frac{\kappa+1}{\kappa-1}\right) \geq \frac{1}{2}\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k \quad \text{für} \quad x \in \mathbb{R}, \kappa > 1 \tag{5.41}$$

Es gilt der

Satz 5.30. *Zu einer gegebenen symmetrisch positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$ gelten für das CG-Verfahren die Fehlerabschätzungen*

$$\begin{aligned} \|x_k - x_*\|_A &\leq 2\gamma^k \|x_*\|_A, \quad k = 0, 1, \dots \\ \|x_k - x_*\|_2 &\leq 2\sqrt{\kappa_A}\gamma^k \|x_*\|_2, \quad k = 0, 1, \dots \end{aligned} \tag{5.42}$$

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

Beweis. Satz 5.29 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

$$\max_{m\leq\lambda\leq M} |p(\lambda)| = \left|T_k\left(\frac{M+m}{M-m}\right)\right|^{-1} = \left|T_k\left(\frac{\kappa_A+1}{\kappa_A-1}\right)\right|^{-1} \overset{(5.41)}{\leq} 2\gamma^k$$

folgt aus (5.40) die erste Abschätzung, also (5.42).
Die zweite Abschätzung des Satzes ist eine unmittelbare Konsequenz aus der Ersten unter der Nutzung der Normäquivalenz (5.39). □

5.8.5 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 5.31. *Das GMRES-Verfahren ist definiert durch den Ansatz des minimalen Residuums* (5.23) *mit der speziellen Wahl $D_k = K_k(A, b)$, es gilt also*

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

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.

Kapitel 6

Numerische Lösung von Anfangswertaufgaben

20. Vorlesung 22.06.09

Anwendungen wie Flugbahnberechnungen, Schwingungsberechnungen oder die Dynamik von Räuber-Beute-Modellen führen auf Anfangswertprobleme für Systeme von gewöhnlichen Differentialgleichungen:

Definition 6.1. *Ein **Anfangswertproblem** für ein System von n gewöhnlichen Differentialgleichungen 1. Ordnung ist von der Form*

$$y' = f(t, y), \quad t \in [a, b] \tag{6.1}$$
$$y(a) = y_0 \tag{6.2}$$

mit einem gegebenen endlichen Intervall $[a, b]$, einem Vektor $y_0 \in \mathbb{R}^n$ und einer Abbildung

$$f : [a, b] \times \mathbb{R}^n \to \mathbb{R}^n \tag{6.3}$$

wobei eine differenzierbare Abbildung $y : [a, b] \to \mathbb{R}^n$ mit den Eigenschaften (6.1) - (6.3) *als **Lösung des Anfangswertproblems** gesucht ist.*

Aussagen zur Existenz und Eindeutigkeit der Lösung liefert

Satz 6.2. *Erfüllt f aus* (6.3) *die Bedingung*

$$\|f(t, u) - f(t, v)\| \leq L \|u - v\|, \quad t \in [a, b],\ u, v \in \mathbb{R}^n \tag{6.4}$$

mit einer Konstanten $L > 0$ in einer beliebigen Vektornorm $\|\cdot\|$ des $\mathbb{R}^n$, dann gelten die Aussagen

(a) Das AWP (6.1),(6.2) *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

$$\|y(t) - \hat{y}\| \leq e^{L(t-a)} \|y_0 - \hat{y}_0\|, \quad t \in [a,b] \tag{6.5}$$

Beweis. Vorlesung DGL oder Analysis □

Bemerkung.

(1) Mit den Aussagen des Satzes 6.2 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 6.3. *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

6.1 Theorie der Einschrittverfahren

Definition 6.4. *Ein Einschrittverfahren zur näherungsweisen Bestimmung einer Lösung des AWP* (6.1),(6.2) *hat die Form*

$$y_{k+1} = y_k + h_k \Phi(t_k, y_k, y_{k+1}, h_k), \quad k = 0, 1, \ldots, N-1 \tag{6.6}$$

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

$$\Delta = \{a = t_0 < t_1 < \ldots < t_N \leq b\}, \; h_k := t_{k+1} - t_k, \quad k = 0, 1, \ldots, N-1 \tag{6.7}$$

Bemerkung. Hängt die Verfahrensfunktion *nicht* von y_{k+1} ab, ist die Berechnungsvorschrift (6.6) 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 6.5. *Unter dem* ***lokalen Diskretisierungsfehler*** *an der Stelle* t_{k+1} *des Verfahrens* (6.6) *versteht man den Wert*

$$d_{k+1} := y(t_{k+1}) - y(t_k) - h_k \Phi(t_k, y(t_k), y(t_{k+1}), h_k) \tag{6.8}$$

Definition 6.6. *Unter dem globalen Diskretisierungsfehler* g_k *an der Stelle* t_k *versteht man den Wert*

$$g_k := y(t_k) - y_k$$

Definition 6.7. *Ein Einschrittverfahren* (6.6) *besitzt die Fehlerordnung* p*, falls für seinen lokalen Diskretisierungsfehler* d_k *die Abschätzungen*

$$|d_k| \leq \text{const.} h_k^{p+1}, \quad k = 1, \ldots, N$$

$$\max_{1 \leq k \leq N} |d_k| \leq D = \text{const.} h_{\max}^{p+1} = \mathcal{O}(h_{\max}^{p+1}) \tag{6.9}$$

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

$$\begin{aligned} |\Phi(t, u_1, u_2, h) - \Phi(t, v_1, u_2, h)| &\leq L_1 |u_1 - v_1| \\ |\Phi(t, u_1, u_2, h) - \Phi(t, u_1, v_2, h)| &\leq L_2 |u_2 - v_2| \end{aligned} \tag{6.10}$$

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 6.8. *Ein Einschrittverfahren* (6.6) *zur Lösung des AWP* (6.1), (6.2) *besitze die Konsistenzordnung* $p \geq 1$ *und die Verfahrensfunktion erfülle die Bedinung* (6.10)*. 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* (6.9) *und* L_1, L_1 *aus* (6.10) *herrührt.*

Bewiesen werden soll der Satz 6.8 für ein explizites Einschrittverfahren (Beweise von allgemeinen Einschrittverfahren in Bärwolff oder Schwarz).
Benötigt wird das

Lemma 6.9. *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 6.8. 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 6.9 liefert wegen $e_0 = 0$ die Behauptung des Satzes 6.8 □

6.2 Spezielle Einschrittverfahren

6.2.1 Euler-Verfahren

Mit der Verfahrensfunktion

$$\Phi(t, y, h_k) = f(t, y)$$

erhält man mit

$$y_{k+1} = y_k + h_k f(t_k, y_k), \quad k = 0, \ldots, N-1 \tag{6.11}$$

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.

$$|d_{k+1}| \leq C h_k^2 \quad \text{mit} \quad C = \frac{1}{2} \max_{\xi \in [a,b]} |y''(\xi)|$$

6.2.2 Einschrittverfahren der Konsistenzordnung $p = 2$

Um ein explizites Einschrittverfahren der Konsistenzordnung $p = 2$ zu erhalten, machen wir den Ansatz

$$\Phi(t, y, h) = a_1 f(t, y) + a_2 f(t + b_1 h, y + b_2 h f(t, y)),\ t \in [a, b],\ h \in [0, b - t],\ y \in \mathbb{R} \tag{6.12}$$

mit noch festzulegenden Konstanten $a_j, b_j \in \mathbb{R}$. Es gilt nun der

Satz 6.10. *Ein Einschrittverfahren* (6.6) *mit einer Verfahrensfunktion der Form* (6.12) *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*

$$a_1 + a_2 = 1, \quad a_2 b_1 = \frac{1}{2}, \quad a_2 b_2 = \frac{1}{2} \tag{6.13}$$

gilt.

Beweis. Taylorentwicklung von $\Phi(t, y(t), \cdot)$ im Punkt $h = 0$ und von der Lösung y in t ergeben

$$\begin{aligned}\Phi(t, y(t), h) &= \Phi(t, y(t), 0) + h\frac{\mathrm{d}\Phi}{\mathrm{d}h}(t, y(t), 0) + \mathcal{O}(h^2) \\ &= (a_1 + a_2)f(t, y(t)) + h\left(a_2 b_1 \frac{\partial f}{\partial t}(t, y(t))\right. \\ &\qquad \left. + a_2 b_2 f(t, y(t))\frac{\partial f}{\partial y}(t, y(t))\right) + \mathcal{O}(h^2) \\ &= f(t, y(t)) + \frac{h}{2}\frac{\partial f}{\partial t}(t, y(t)) + \frac{h}{2}f(t, y(t))\frac{\partial f}{\partial y}(t, y(t)) + \mathcal{O}(h^2)\end{aligned}$$

$$\begin{aligned}y(t+h) &= y(t) + hy'(t) + \frac{h^2}{2}y''(t) + \mathcal{O}(h^3) \\ &= y(t) + h\left[f(t, y(t)) + \frac{h}{2}y''(t)\right] + \mathcal{O}(h^3) \\ &= y(t) + h\left[f(t, y(t)) + \frac{h}{2}\left\{\frac{\partial f}{\partial t}(t, y(t))\right.\right. \\ &\qquad \left.\left. + f(t, y(t))\frac{\partial f}{\partial y}(t, y(t))\right\}\right] + \mathcal{O}(h^3) \\ &= y(t) + h\Phi(t, y(t), h) + \mathcal{O}(h^3)\end{aligned}$$

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

$$y_{k+1} = y_k + h_k f\left(t_k + \frac{h_k}{2}, y_k + \frac{h_k}{2}f(t_k, y_k)\right), \quad k = 0, \ldots, N-1 \tag{6.14}$$

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

$$y_{k+1} = y_k + \frac{h_k}{2}\left[f(t_k, y_k) + f(t_k + h_k, y_k + h_k f(t_k, y_k))\right], \quad k = 0, \ldots, N-1 \tag{6.15}$$

das **Verfahren von Heun** mit der Konsistenzordnung $p = 2$

6.3 Verfahren höherer Ordnung

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)

$$y(t) = y_0 + \int_a^t f(s, y(s))\mathrm{d}s \tag{6.16}$$

bzw. für eine Diskretisierung des Intervalls $[a, b]$

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(s, y(s))\mathrm{d}s \tag{6.17}$$

Das letzte Integral aus (6.17) approximieren wir durch eine Quadraturformel

$$\int_{t_k}^{t_{k+1}} f(s, y(s))\mathrm{d}s \tag{6.18}$$

wobei die s_l zu einer Zerlegung von $[t_k, t_{k+1}]$ gehören. (6.17) und (6.18) ergeben

$$y(t_{k+1}) \approx y(t_k) + h_k \sum_{l=1}^{m} \gamma_l f(s_l, y(s_l)) \tag{6.19}$$

wobei wir die Werte $y(s_l)$ nicht kennen. Sie müssen näherungsweise aus $y(t_k)$ bestimmt werden, damit (6.19) 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 (6.19)

$$y(t_{k+1}) \approx y(t_k) + \frac{h_k}{2}\left[f(t_k, y(t_k)) + f(t_{k+1}, y(t_{k+1}))\right]$$

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 $y(t_k)$ beizeichnen 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

$$s_l = t_k + d_l h_k, \quad \alpha_l = \sum_{r=1}^{l-1} \beta_{lr}$$

werden die k_l rekursiv definiert:

$$\begin{aligned}
k_1(t_k, y_k) &= f(t_k, y_k) \\
k_2(t_k, y_k) &= f(t_k + \alpha_2 h_k, y_k + h_k \beta_{21} k_1(t_k, y_k)) \\
k_3(t_k, y_k) &= f(t_k + \alpha_3 h_k, y_k + h_k(\beta_{31} k_1 + \beta_{32} k_2)) \\
&\vdots \\
k_m(t_k, y_k) &= f(t_k + \alpha_m h_k, y_k + h_k(\beta_{m1} k_1 + \cdots + \beta_{mm-1} k_{m-1}))
\end{aligned} \tag{6.20}$$

Ausgehend von (6.19) und (6.20) wird durch

$$y_{k+1} = y_k + h_k(\gamma_1 k_1(t_k, y_k) + \cdots + \gamma_m k_m(t_k, y_k)) \tag{6.21}$$

ein explizites numerisches Verfahren zu Lösung des AWP $y' = f(t, y), y(a) = y_0$ definiert.

Definition 6.11. *Das Verfahren* (6.21) *heißt* m***-stufiges Runge-Kutta-Verfahren*** *mit* k_l *aus* (6.20) *und die* k_l *heißen Stufenwerte.*

Bemerkung. Wir haben oben schon festgestellt, dass im Fall $m = 2$ mit $\gamma_1 = \gamma_2 = \frac{1}{2}, \alpha_2 = 1, \beta_{21} = 1$ (6.21) gerade das Heun-Verfahren ergibt, also ein Verfahren mit der Konsistenzordnung $p = 2$. Wir werden nun Bedingungen für die freien Parameter im Verfahren (6.21) 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

$$1 = \gamma_1 + \gamma_2 + \cdots \gamma_m \tag{6.22}$$

also haben die γ_l die Funktion von Gewichten.
Fordert man vom Verfahren (6.21), dass die Dgl $y' = 1$ (y linear) exakt integriert wird, ergibt sich die Bedingung

$$\alpha_l = \beta_{l1} + \cdots + \beta_{ll-1} \tag{6.23}$$

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

$$y(s_l) = y(t_k) + h_k(\beta_{l1} + \cdots + \beta_{ll-1}) \tag{6.24}$$

da alle $k_r = 1$ sind. Andererseits ist y als exakte Lösung linear, d.h.

$$y(s_l) = y(t_k) + \alpha_l h_k \tag{6.25}$$

und aus dem Vergleich von (6.24),(6.25) folgt

$$\alpha_l = \beta_{l1} + \cdots + \beta_{ll-1}$$

Definition 6.12. *Die Tabelle mit den Koeffizienten $\alpha_l, \beta_{lr}, \gamma_r$ in der Form*

$$\begin{array}{c|ccccc} 0 & & & & & \\ \alpha_2 & \beta_{21} & & & & \\ \alpha_3 & \beta_{31} & \beta_{32} & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ \alpha_m & \beta_{m1} & \beta_{m2} & \dots & \beta_{mm-1} & \\ \hline & \gamma_1 & \gamma_2 & \dots & \gamma_{m-1} & \gamma_m \end{array} \tag{6.26}$$

heißt ***Butcher-Tabelle*** *und beschreibt das Verfahren* (6.21). *α_1 ist hier gleich* 0*, weil explizite Verfahren betrachtet werden.*

Satz 6.13. *Ein explizites Runge-Kutta-Verfahren* (6.21)*, dessen Koeffizienten die Bedingungen* (6.22) *und* (6.23) *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.

$$\begin{aligned}
|d_{k+1}| &= |y(t_{k+1}) - y(t_k) - h\Phi(t_k, y(t_k), h)| \\
&= \left| y(t_{k+1}) - y(t_k) - h\sum_{r=1}^{m} \gamma_r k_r(t_k, y(t_k)) \right| \\
&\overset{(6.22)}{=} \left| y(t_{k+1}) - y(t_k) - hf(t_k, y_k) - h\sum_{r=1}^{m} \gamma_r k_r(t_k, y(t_k)) - f(t_k, y(t_k)) \right| \\
&\leq \underbrace{|y(t_{k+1}) - y(t_k) - hy'(t_k)|}_{\in \mathcal{O}(h^2)} + h \left| \sum_{r=1}^{m} \gamma_r \underbrace{(k_r(t_k, y(t_k)) - f(t_k, y(t_k)))}_{\in \mathcal{O}(h) \ (6.23)} \right|
\end{aligned}$$

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	$p < m - 2$

6.4 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

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

$$
\begin{aligned}
k_1 &= f(t_k, y_k) \\
k_2 &= f(t_k + \frac{1}{2}h_k, y_k + \frac{1}{2}h_k k_1) \\
k_3 &= f(t_k + \frac{1}{2}h_k, y_k + \frac{1}{2}h_k k_2) \\
k_4 &= f(t_k + h_k, y_k + h_k k_3) \\
y_{k+1} &= y_k + h_k \left(\frac{1}{6}k_1 + \frac{1}{4}k_2 + \frac{1}{3}k_3 + \frac{1}{6}k_4 \right), \quad p = 4
\end{aligned}
$$

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 6.14. *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

$$\begin{aligned} \alpha_2\gamma_2 + \alpha_3\gamma_3 &= \frac{1}{2} \\ \alpha_2\gamma_3\beta_{32} &= \frac{1}{6} \\ \alpha_2^2\gamma_2 + \alpha_3^2\gamma_3^6 &= \frac{1}{3} \end{aligned}$$

Dann hat das Runge-Kutta-Verfahren (explizit, 3-stufig) die Fehlerordnung $p = 3$

Beweis. Grundlage für den Beweis ist die Taylor-Approximation

$$\begin{aligned} f(t + \Delta t, y + \Delta y) = f(t, y) + \begin{pmatrix} \frac{\partial f}{\partial t}(t, y) \\ \frac{\partial f}{\partial t}(t, y) \end{pmatrix} \begin{pmatrix} \Delta t \\ \Delta y \end{pmatrix} \\ + \frac{1}{2}(\Delta t, \Delta y) \begin{pmatrix} \frac{\partial^2 f}{\partial t^2}(t, y) & \frac{\partial^2 f}{\partial t \partial y}(t, y) \\ \frac{\partial^2 f}{\partial y \partial t}(t, y) & \frac{\partial^2 f}{\partial y^2}(t, y) \end{pmatrix} \begin{pmatrix} \Delta t \\ \Delta y \end{pmatrix} + \mathcal{O}(\Delta^3) \end{aligned} \tag{6.27}$$

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

$$\begin{aligned} \bar{k}_1 &= f(t_k, y(t_k)) \\ \bar{k}_2 &= f(t_k + \alpha_2 h, y(t_k) + \alpha_2 h \bar{k}_1) \\ \bar{k}_3 &= f(t_k + \alpha_3 h, y(t_k) + h(\beta_{31}\bar{k}_1 + \beta_{32}\bar{k}_2)) \end{aligned}$$

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 (6.27) für $\bar{k}_2$

$$\begin{aligned}\bar{k}_2 &= f(t_k + \Delta t, y(t_k) + \Delta y)\\ &= f + \alpha_2 h f_t + \alpha_2 h f f_y + \frac{1}{2}\alpha_2^2 h^2 f_{tt} + \alpha_2^2 h^2 f f_{ty} + \frac{1}{2}\alpha_2^2 h^2 f^2 f_{yy} + \mathcal{O}(h^3)\\ &=: f + \alpha_2 h F + \frac{1}{2}\alpha_2^2 h^2 G + \mathcal{O}(h^3) \qquad (6.28)\end{aligned}$$

$f, f_t, \ldots, f_{yy}$ sind dabei die Funktions- bzw. Ableitunswerte an der Stelle $(t_k, y(t_k))$. Für $\bar{k}_3$ erhält man unter Nutzung von (6.28) und (6.27)

$$\begin{aligned}\bar{k}_3 &= f(t_k + \alpha_3 h, y(t_k) + h(\beta_{31}\bar{k}_1 + \beta_{32}\bar{k}_2)\\ &= f + \alpha_3 h f_t + h(\beta_{31}\bar{k}_1 + \beta_{32}\bar{k}_2) f_y + \frac{1}{2}\alpha_3^2 h^2 f_{tt}\\ &\quad + \alpha_3(\beta_{31}\bar{k}_1 + \beta_{32}\bar{k}_2) h^2 f_{ty} + \frac{1}{2}(\beta_{31}\bar{k}_1 + \beta_{32}\bar{k}_2)^2 h^2 f_{yy} + \mathcal{O}(h^3)\\ &= f + h(\alpha_3 f_t + [\beta_{31} + \beta_{32}] f f_y) + h^2\Big(\alpha_2\beta_{32} F f_y\\ &\quad + \frac{1}{2}\alpha_3^2 f_{tt} + \alpha_3[\beta_{31} + \beta_{32}] f f_{ty} + \frac{1}{2}(\beta_{31} + \beta_{32}) f^2 f_{yy}\Big) + \mathcal{O}(h^3)\\ &= f + \alpha_3 h F + h^2(\alpha_2\beta_{32} F f_y + \frac{1}{2}\alpha_3^2 G) + \mathcal{O}(h^3) \qquad (6.29)\end{aligned}$$

Mit (6.28) und (6.29) folgt für den lokalen Diskretisierungsfehler

$$\begin{aligned}d_{k+1} &= h(1 - \gamma_1 - \gamma_2 - \gamma_3) f + h^2\left(\frac{1}{2} - \alpha_2\gamma_2 - \alpha_3\gamma_3\right) F\\ &+ h^3\left(\left[\frac{1}{6} - \alpha_2\gamma_3\beta_{32}\right] F f_y + \left[\frac{1}{6} - \frac{1}{2}\alpha_2^2\gamma_2 - \frac{1}{2}\alpha_3^2\gamma_3\right] G\right) + \mathcal{O}(h^4)\end{aligned} \qquad (6.30)$$

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*

$$\begin{aligned}\gamma_1 + \gamma_2 + \gamma_3 &= 1\\ \alpha_2\gamma_2 + \alpha_3\gamma_3 &= \frac{1}{2}\\ \alpha_2\gamma_3\beta_{32} &= \frac{1}{6}\\ \alpha_2^2\gamma_2 + \alpha_3^2\gamma_3 &= \frac{1}{3}\end{aligned} \qquad (6.31)$$

hat das dazugehörrige 3-stufige Runge-Kutta-Verfahren die Fehlerordnung $p = 3$*, wobei* $\alpha_2 = \beta_{21}$ *ist.* (6.31) *hat z.B. mit den Einschränkungen* $\alpha_2 \neq \alpha_3$ *und* $\alpha_2 \neq \frac{2}{3}$ *die Lösungen*

$$\gamma_2 = \frac{3\alpha_3 - 2}{6\alpha_2(\alpha_3 - \alpha_2)}, \qquad \gamma_3 = \frac{2 - 3\alpha_2}{6\alpha_3(\alpha_3 - \alpha_2)} \tag{6.32}$$
$$\gamma_1 = \frac{6\alpha_2\alpha_3 + 2 - 3(\alpha_2 + \alpha_3)}{6\alpha_2\alpha_3}, \qquad \beta_{32} = \frac{\alpha_3(\alpha_3 - \alpha_2)}{\alpha_2(2 - 3\alpha_2)}$$

für $\alpha_2, \alpha_3 \in \mathbb{R}$*, also die zweiparametrige Lösungsmenge*

$$\mathcal{M} = \{(\gamma_1, \gamma_2, \gamma_3, \alpha_2, \alpha_3, \beta_{32}) | \gamma_1, \gamma_2, \gamma_3\beta_{32} \text{ gemäß } (6.32), \\ \alpha_2, \alpha_3 \in \mathbb{R}, \alpha_2 \neq \alpha_3, \alpha_2 \neq \frac{2}{3}\}$$

Die restlichen Parameter des Verfahrens ergeben sich aus

$$\beta_{21} = \alpha_2, \quad \beta_{31} = \alpha_3 - \beta_{32}$$

6.5 Implizite Runge-Kutta-Verfahren

Explizite Verfahren neigen zur Instabilität und damit besteht die Gefahr der Verstärkung von Rundungsehlern.
Implizite Verfahren erweisen sich als stabil, speziell, wenn es sich um die Lösung von AWPs mit sogenannten steifen DGL handelt.
Im Unterschied zum Gleichungssystem (6.20) wird beim impliziten Runge-Kutta-Verfahren das Gleichungssystem

$$k_r(t_k, y_k) = f(t_k + \alpha_r h_k, y_k + h_k(\beta_{r1}k_1 + \cdots + \beta_{rm}k_m)), \quad r = 1, \ldots, m \tag{6.33}$$

zur Bestimmung der k_r zugrunde gelegt.
Mit (6.33) wird (6.21) zu einem impliziten Runge-Kutta- Verfahren. Aus (6.33) ergibt sich die Butcher-Tabelle

$$\begin{array}{c|ccc} \alpha_1 & \beta_{11} & \ldots & \beta_{1m} \\ \alpha_2 & \beta_{21} & \ldots & \beta_{2m} \\ \vdots & \vdots & & \vdots \\ \alpha_m & \beta_{m1} & \ldots & \beta_{mm} \\ \hline & \gamma_1 & \ldots & \gamma_m \end{array} \tag{6.34}$$

Die Überlegung, die bei den expliziten Verfahren die Bedingung (6.23) für die Koeffizienten α_r, β_{rl} gerechtfertigt haben, ergeben analog bei den impliziten Runge-Kutta- Verfahren die Bedingung

$$\alpha_r = \beta_{r1} + \beta_{r2} + \cdots + \beta_{rm}, \quad r = 1, \ldots, m \tag{6.35}$$

Zur Lösbarkeit des Gleichungssystems (6.33) gilt der

Satz 6.15. *f genüge auf $[a,b]\times\mathbb{R}$ der Lipschitz-Bedingung*

$$|f(t,y_1)-f(t,y_2)|\le L\,|y_1-y_2|$$

und die Schrittweite $h=h_k$ genüge der Bedingung

$$q=hL\max_{1\le j\le m}\left(\sum_{r=1}^{m}|\beta_{jr}|\right)<1$$

Dann hat (6.33) *zur Bestimmung von $k_1,\dots,k_m$ genau eine Lösung*

Beweis. Aussage folgt aus dem Banachschen Fixpunktsatz □

6.6 Rundungsfehleranalyse von expliziten Einschrittverfahren

Zur numerischen Lösung eines AWP betrachten wir das Verfahren

$$y_{k+1}=y_k+h_k\Phi(t_k,y_k,h_k),\quad k=0,1,2\dots,N-1 \tag{6.36}$$

mit der Verfahrensfunktion Φ. Durch Rundungsfehler arbeitet man statt (6.36) mit einem Verfahren der Form

$$y_{k+1}=y_k+h_k\Phi(t_k,y_k,h_k)+\rho_k,\quad k=0,1,\dots,N-1 \tag{6.37}$$

$$y_0=y_0+e_0,\quad |\rho_k|\le\delta,\quad k=0,1,\dots,N-1$$

mit gewissen Zahlen $e_0,\rho_k\in\mathbb{R}$
Für die Rundungsfehler infolge des Verfahrens (6.37) gilt der folgende

Satz 6.16. *Zur Lösung des AWP $y'=f(t,y), y(a)=y_0$, sei durch* (6.36) *ein Einschrittverfahren mit der Konsistenzordnung $p\ge 1$ gegeben, wobei die Verfahrensfunktion bezüglich der 2. Variablen Lipschitz-stetig mit der Konstanten $L>0$ ist.*
Dann gelten für die durch die fehlerbehaftete Verfahrensvorschrift (6.37) *gewonnenen Approximationen die Abschätzungen*

$$\max_{k=0,\dots,N}|y_k-y(t_k)|\le K(h_{\max}^p+\frac{\delta}{k_{\min}})+e^{L(b-a)}\epsilon \tag{6.38}$$

mit der Konstanten $K=\frac{\max\{C,1\}}{L}\left[e^{L(b-a)}-1\right]$. C ist dabei die Konstante aus der Abschätzung $|d_k|\le Ch_k^{p+1}$ für den lokalen Diskretisierungsfehler.

Beweis. Wir setzen

$$e_k = y_k - y(t_k), \quad k = 0, 1, \ldots, N$$

und damit gilt für $k = 0, 1, \ldots, N$

$$\begin{aligned} y(t_{k+1}) &= y(t_k) + h_k \Phi(t_k, y(t_k), h_k) + d_{k+1} \\ y_{k+1} &= y_k + h_k \Phi(t_k, y_k, h_k) + \rho_k \end{aligned}$$

bzw.

$$e_{k+1} = e_k + h_k[\Phi(t_k, y_k, h_k) - \Phi(t_k, y(t_k), h_k)] + \rho_k - d_{k+1}$$

und folglich

$$\begin{aligned} |e_{k+1}| &\leq |e_k| + h_k \left|\Phi(t_k, y_k, h_k) - \Phi(t_k, y(t_k), h_k)\right| + |\rho_k| + |d_{k+1}| \\ &\leq (1 + h_k L) |e_k| + h_k \left(C h_{\max}^p + \frac{\delta}{h_{\min}} \right) \end{aligned}$$

Aus Lemma 6.9 folgt mit $|e_0| \leq \epsilon$ die Behauptung des Satzes. □

6.7 Schrittweitensteuerung bei Einschrittverfahren

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

$$\begin{aligned} k_1 &= f(t_k, y_k) \\ k_2 &= f(t_k + h, y_k + h k_1) \\ y_{k+1} &= y_k + \frac{1}{2} h [k_1 + k_2] \end{aligned}$$

erläutert werden. Als lokaler Diskretisierungsfehler ergibt sich

$$d_{k+1}^{(H)} = y(t_{k+1}) - y(t_k) - \frac{1}{2} h [\bar{k}_1 + \bar{k}_2] \tag{6.39}$$

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 (6.32) 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

$$\begin{aligned} k_1 &= f(t_k, y_k) \\ k_2 &= f(t_k + h, y_k + hk_1) \\ k_3 &= f(t_k + \frac{h}{2}, y_k + \frac{h}{4}(k_1 + k_2)) \\ y_{k+1} &= y_k + \frac{h}{6}[k_1 + k_2 + 4k_3] \end{aligned} \tag{6.40}$$

ergibt. Für den lokalen Diskretisierungsfehler des Verfahrens (6.39) ergibt sich

$$d_{k+1}^{(RK)} = y(t_{k+1}) - y(t_k) - \frac{h}{6}[\bar{k}_1 + \bar{k}_2 + 4\bar{k}_3] \tag{6.41}$$

mit $\bar{k}_3 = f(t_k + \frac{h}{2}, y(t_k) + \frac{h}{4}(\bar{k}_1 + \bar{k}_2))$. Mit (6.39) und (6.41) 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 (6.40).

6.8 Mehrschrittverfahren

Die Klasse der Mehrschrittverfahren zur Lösung von AWP ist dadruch 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

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(t, y(t))\mathrm{d}t \tag{6.42}$$

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 (6.42) i.d.R. besser approximieren als bei den Einschrittverfahren unter ausschließlicher Nutzung des Wertes f_k. Für das Interpolationspolynom durch die Stützpunkte $(t_j, y_j)_{j=k-3,\ldots,k}$ 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 (6.42), sodass man auf der Grundlage von (6.42) 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 $h = t_{k+1} - t_k$ erh ält man für den 2. Integralsummanden ($j = 1$)

$$I_1 = \int_{t_k}^{t_{k-1}} L_{k-1}(t)\mathrm{d}t = \int_{t_k}^{t_{k+1}} \frac{(t - t_{k-3})(t - t_{k-2})(t - t_k)}{(t_{k-1} - t_{k-3})(t_{k-1} - t_{k-2})(t_{k-1} - t_k)}$$

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

$$y_{k+1} = y_k + \frac{h}{24}[55f_k - 59f_{k-1} + 37f_{k-2} - 9f_{k-3}] \tag{6.43}$$

die **Methode von Adams-Bashforths** (kurz AB-Verfahren) ergibt.
Durch Taylor-Reihenentwicklung erhält man bei entsprechendert Glattheit von f bzw. $y(t)$ den lokalen Diskretisierungsfehler

$$d_{k+1} = \frac{251}{720}h^5 y^{(5)} + \mathcal{O}(h^6) \tag{6.44}$$

d.h. das Verfahren (6.43) ist von 4. Ordnung.

Definition 6.17. *Bei Verwendung von m Stützwerten*

$$(t_k, f_k), \ldots, (t_{k-m+1}, f_{k-m+1})$$

zur Berechnung eines Interpolationspolynoms p_{m-1} zur Approximation von f zwecks näherungsweiser Berechnung des Integrals in (6.42) *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.

$$y_{k+1} = y_k + \frac{h}{12}[23f_k - 16f_{k-1} + 5f_{k-2}] \tag{6.45}$$
$$y_{k+1} = y_k + \frac{h}{24}[55f_k - 59f_{k-1} + 37f_{k-2} - 9f_{k-3}]$$
$$y_{k+1} = y_k + \frac{h}{720}[1901f_k - 2774f_{k-1} + 2616f_{k-2} - 1274f_{k-3} + 251f_{k-4}]$$

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 (6.45) 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 (6.42) 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

$$y_{k+1} = y_k + \frac{h}{720}[251 f_{k+1} + 646 f_k - 264 f_{k-1} + 106 f_{k-2} - 19 f_{k-3}] \qquad (6.46)$$

Das Verfahren (6.46) ist eine implizite 4-Schritt-Methode und heißt Methode von Adams-Moulton (kurz AM-Verfahren). Das 3-Schritt AM-Verfahren hat die Form

$$y_{k+1} = y_k + \frac{h}{24}[9 f(t_{k+1}, y_{k+1}) + 19 f_k - 5 f_{k-1} + f_{k-2}] \qquad (6.47)$$

Zur Bestimmung der Lösung von (6.47) 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

$$y_{k+1}^{(p)} = y_k + \frac{h}{12}[23f_k - 16f_{k-1} + 5f_{k-2}] \tag{6.48}$$

$$y_{k+1} = y_k + \frac{h}{24}[9f(t_{k+1}, y_{k+1}^{(p)}) + 19f_k - 5f_{k-1} + f_{k-2}] \tag{6.49}$$

Diese Kombination von AB- und AM-Verfahren bezeichnet man als Adams-Bashforth-Moulton-Verfahren (kurz ABM-Verfahren). Das ABM-Verfahren (6.48) hat ebenso wie das AM-Verfahren (6.47) 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.

6.9 Allgemeine lineare Mehrschrittverfahren

24. Vorlesung 08.07.2009

Definition 6.18. *Unter einem linearen m-Schrittverfahren ($m > 1$) versteht man eine Vorschrift mit $s = k - m + 1$*

$$\sum_{j=0}^{m} a_j y_{s+j} = h \sum_{j=0}^{m} b_j f(t_{s+j}, y_{s+j}) \tag{6.50}$$

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 (6.50) 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.
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.

Definition 6.19. *Das lineare Mehrschrittverfahren* (6.50) *hat die Fehlerordnung p, falls in der Entwicklung des lokalen Diskretisierungsfehlers d_{k+1} in eine Potenzreihe von h für eine beliebige Stelle $\tilde{t} \in [t_{k+1-m}, t_{k+1}]$*

$$\begin{aligned}
d_{k+1} &= \sum_{j=0}^{m}[a_j y(t_{s+j}) - hb_j f(t_{s+j}, y(t_{s+j}))] \\
&= c_0 y(\tilde{t}) + c_y h y'(\tilde{t}) + \ldots + c_p h^p y^{(p)}(\tilde{t}) + \ldots
\end{aligned} \tag{6.51}$$

$c_0 = \ldots = c_p = 0$ 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

$$y(t_{s+j}) = y(\tilde{t} + jh) = \sum_{r=0}^{q} \frac{(jh)^r}{r!} y^{(r)}(\tilde{t}) + R_{q+1}$$
$$y'(t_{s+j}) = y'(\tilde{t} + jh) = \sum_{r=0}^{q-1} \frac{(jh)^r}{r!} y^{(r+1)}(\tilde{t}) + R_q \tag{6.52}$$

Die Substitution der Reihen (6.52) in (6.51) ergibt für die Koeffizienten c_j durch Koeffizientenvergleich

$$\begin{aligned}
c_0 &= a_0 + a_1 + \ldots + a_m \\
c_1 &= a_1 + 2a_2 + \ldots + ma_m - (b_0 + b_1 + \ldots + b_m) \\
c_2 &= \frac{1}{2!}(a_1 + 2^2 a_2 + \ldots + m^2 a_m) - \frac{1}{1!}(b_1 + 2b_2 + \ldots + mb_m) \\
&\vdots \\
c_r &= \frac{1}{r!}(a_1 + 2^r a_2 + \ldots + m^r a_m) - \frac{1}{(r-1)!}(2b_1 + 2^{r-1} b_2 + \ldots + m^{r-1} b_m)
\end{aligned} \tag{6.53}$$

für $r = 2, 3, \ldots, q$

Beispiel. Es soll ein explizites 2-Schritt-Verfahren

$$a_0 y_{k-1} + a_1 y_k + a_2 y_{k+1} = h[b_0 f_{k-1} + b_1 f_k] \tag{6.54}$$

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

$$y_{k+1} = y_{k-1} + 2hf_k \tag{6.55}$$

hat.

Definition 6.20. *Mit den Koeffizienten a_j, b_j werden durch*

$$\rho(z) = \sum_{j=0}^{m} a_j z^j, \quad \sigma(z) = \sum_{j=0}^{m} b_j z^j \tag{6.56}$$

das erste und zweite charakteristische Polynom eines m-Schritt-Verfahrens erklärt.

Aus dem Gleichungssystem (6.53) kann man mit Hilfe der charakteristischen Polynome die folgende notwendige und hinreichende Bedingung für die Konsistenz eines Mehrschrittverfahrens formulieren.

Satz 6.21. *Notwendig und hinreichend für die Konsistenz des Mehrschrittverfahrens* (6.50) *ist die Erfüllung der Bedingungen*

$$c_0 = \rho(1) = 0, \quad c_1 = \rho'(1) - \sigma(1) = 0 \tag{6.57}$$

Macht man außer der Wahl von $a_2 = 1$ keine weiteren Einschränkungen an die Koeffizienten des expliziten 2-Schritt-Verfahrens (6.54), dann erreicht man die maximale Ordnung $p = 3$ durch die Lösung des Gleichungssystem (6.53) f ür $q = 3$, 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

$$y_{k+1} = 5y_{k-1} - 4y_k + h[4f_k + 2f_{k-1}] \tag{6.58}$$

folgt.
Obwohl das Verfahren (6.58) die maximale Fehlerordnung $p = 3$ hat, ist es im Vergleich zum Verfahren (6.55) unbrauchbar, weil es nicht stabil ist. Was das konkret bedeutet, soll im Folgenden erklärt und untersucht werden.
Dazu wird die Testdifferentialgleichung (AWP)

$$y' = \lambda y, \quad y(0) = 1, \ \lambda \in \mathbb{R}, \ \lambda < 0 \tag{6.59}$$

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 (6.58)

$$\begin{aligned} & y_{k+1} = 5y_{k-1} - 4y_k + h[4\lambda y_k + 2\lambda y_{k-1}] \\ \Leftrightarrow \quad & (-5 - 2\lambda h)y_{k-1} + (4 - 4\lambda h)y_k + y_{k+1} = 0 \end{aligned} \tag{6.60}$$

Mit dem Lösungsansatz $y_k = z^k, z \neq 0$ ergibt (6.60) nach Division mit z^{k-1}

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

bzw. mit dem charakteristischen Polynomen

$$\rho(t) = -5 + 4z + z^2, \quad \sigma(2 + 4z), \quad \phi(z) = \rho(z) - \lambda h \sigma(z) = 0$$

Als Nullstellen von ϕ 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 (6.60)

$$y_k = c_1 z_1^k + c_2 z_2^k \tag{6.61}$$

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 (6.61) 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 (6.61) 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 druchgeführen Überlegungen rechtfertigen die folgende Definition.

Definition 6.22. *Das Mehrschrittverfahren* (6.50) *heißt* ***nullstabil****, falls die Nullstellen z_j des ersten charakteristischen Polynoms $\rho(t)$*

(a) betragsmäßig nicht größer als 1 sind

(b) mehrfache Nulstellen betragsmäßig echt kleiner als 1 sind

Für das oben konstruierte 2-Schritt-Verfahren mit der maximalen Fehlerordnung $p = 3$ hat das charakteristische Polyno $\rho(t)$ die Nullstellen $z_{1,2} = -2 \pm 3$ und damit ist das Verfahren nicht nullstabil.
Im Unterschied dazu ist das Verfahren (6.55) mit der Ordnung 2 und dem charakteristischen Polynom $\rho(t) = -1 + z^2$ 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.

25. Vorlesung 13.7.2009

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.

6.10 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

$$y' = \lambda y, \quad y(0) = 1, \ \lambda \in \mathbb{R} \text{ oder } \lambda \in \mathbb{C} \tag{6.62}$$

d.h. wir lassen auch Parameter λ aus $\mathbb{C}$ zu. Damit sind auch L ösungen 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 (6.62) 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 R-K-Verfahren 3. Ordnung

$$\begin{aligned} k_1 &= \lambda y_k \\ k_2 &= \lambda(y_k + \frac{1}{2}hk_1) = (\lambda + \frac{1}{2}h\lambda^2)y_k \qquad (6.63) \\ k_3 &= \lambda(y_k - hk_1 + 2hk_2) = (\lambda + h\lambda^2 + h^2\lambda^3)y_k \\ y_{k+1} &= y_k + \frac{h}{6}[k_1 + 4k_2 + k_3] = (1 + h\lambda + \frac{h^2}{2}\lambda^2 + \frac{h^3}{6}\lambda^3)y_k \qquad (6.64) \end{aligned}$$

also y_{k+1} als Produkt von y_k mit dem Faktor

$$F(h\lambda) = 1 + h\lambda + \frac{h^2}{2}\lambda^2 + \frac{h^3}{6}\lambda^3 \qquad (6.65)$$

Der Faktor (6.65) enthät 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 (6.63) qualitativ beschrieben wird.
Für $\lambda < 0$ reell, uss 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 λ gesichert (nur für λ mit $|F(h\lambda)| < 1$.
Auch im Fall eines komplexen Parameters λ reicht die Bedingung $\alpha = \Re(\lambda) < 0$ nicht aus, um das Abklingen der Lösungder Testaufgabe zu sicher, sondern für F muss $|F(h\lambda)| < 1$ gelten.
Die durchgeführten Überlegungen rechtfertigen die

Definition 6.23. *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" exentuelle 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 6.24. *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 R-K-Verfahren) sind auch andere implizite R-K-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

$$\begin{aligned} k_1 &= \frac{\lambda}{1 - \frac{h\lambda}{2}} y_k \\ y_{k+1} = y_k + hk_1 &= \frac{1 + \frac{h\lambda}{2}}{1 - \frac{h\lambda}{2}} y_k =: F(h\lambda) y_k \end{aligned}$$

und da für negatives a gilt $|1 + a + bi| < |1 - a - bi|$ ist $|F(\mu)| < 1$