Numerik I, aktueller Vorlesungsstand

Günter Bärwolff

5.2.2015

Inhaltsverzeichnis

0 Vorwort **1**

1 Rechnerarithmetik **2**
1.1 Zahldarstellungen 2
1.2 Allgemeine Gleitpunkt-Zahlensysteme 2
1.3 Struktur und Eigenschaften des normalisierten Gleitpunktzahlensystems F 4
1.4 Rechnen mit Gleitpunktzahlen 5
1.5 Ersatzarithmetik 8
1.6 Fehlerakkumulation 9

2 Stabilität, Vorwärtsanalyse, Rückwärtsanalyse **13**
2.1 Kondition als Maß für Fehlerverstärkungen 13
2.2 Vektor- und Matrixnormen 14
2.3 Stabilitätskonzepte 18
2.3.1 Vorwärtsanalyse 18
2.3.2 Rückwärtsanalyse 20

3 Lösung linearer Gleichungssysteme **23**
3.1 LR-Zerlegung 23
3.1.1 Realisierung mit dem Gaußschen Eliminationsverfahren 25
3.1.2 LR-Zerlegung mit Spaltenpivotisierung 31
3.2 Cholesky-Zerlegung 34
3.3 Singulärwertzerlegung 37

4 Die iterative Lösung von Gleichungen bzw. Gleichungssystemen **47**
4.1 Die iterative Lösung linearer Gleichungssysteme 48
4.2 Jacobi-Verfahren oder Gesamtschrittverfahren 51
4.3 Gauß-Seidel-Iterationsverfahren oder Einzelschrittverfahren . . 55
4.4 Verallgemeinerung des Gauß-Seidel-Verfahrens 56

4.5 Krylov-Raum-Verfahren zur Lösung linearer Gleichungssysteme 57
4.5.1 Der Ansatz des orthogonalen Residuums (4.20) für symmetrische positiv definite Matrizen 58
4.5.2 Der Ansatz des orthogonalen Residuums (4.20) für gegebene A-konjugierte Basen 59
4.5.3 Das CG-Verfahren für positiv definite, symmetrische Matrizen 60
4.5.4 Konvergenzgeschwindigkeit des CG-Verfahrens 63
4.5.5 CGNR-Verfahren 66
4.5.6 GMRES-Verfahren 66
4.6 Die iterative Lösung nichtlinearer Gleichungssysteme 67
4.6.1 Newton-Verfahren 68
4.6.2 Newton-Verfahren für nichtlineare Gleichungssysteme $F(x) = 0$ 71

5 Orthogonale Matrizen – QR-Zerlegung – Ausgleichsprobleme 76
5.1 Gram-Schmidt-Verfahren zur Orthogonalisierung 78
5.2 Householder-Matrizen/Transformationen 79
5.3 Algorithmus zur Konstruktion der Faktorisierung mittels Householder-Transformationen 80
5.4 Gauß-Newton-Verfahren 82

6 Interpolation 87
6.1 Polynominterpolation 88
6.1.1 Konstruktion des Interpolationspolynoms 89
6.2 Lagrange-Interpolation 90
6.3 Fehler bei der Polynominterpolation 92
6.4 Newton-Interpolation 93
6.4.1 Algorithmische Aspekte der Polynominterpolation - Horner-Schema 97
6.5 Hermite-Interpolation 99
6.6 Spline-Interpolation 101
6.6.1 Interpolierende lineare Splines $s \in S_{\Delta,1}$ 102
6.6.2 Berechnung interpolierender kubischer Splines 103
6.6.3 Gestalt der Gleichungssysteme 104
6.7 Existenz und Eindeutigkeit der betrachteten interpolierenden kubischen Splines 105
6.8 Fehlerabschätzungen für interpolierende kubische Splines . . . 108

7 Numerische Integration **111**
7.1 Interpolatorische Quadraturformeln 111
7.2 Fehler bei der interpolatorischen Quadratur 112
7.3 Numerischen Integration mit Newton-Cotes-Formeln 113
7.4 Summierte abgeschlossene Newton-Cotes-Quadraturformeln . 117
7.5 Gauß-Quadraturen 118
7.5.1 Orthogonale Polynome 120
7.5.2 Konstruktion von Folgen orthogonaler Polynome 121
7.6 Numerische Integration durch Extrapolation 126
7.7 Anwendung des Schemas von Neville-Aitken - Romberg-Verfahren 128

8 Numerische Lösung von Anfangswertaufgaben **131**
8.1 Theorie der Einschrittverfahren 133
8.2 Spezielle Einschrittverfahren 136
8.2.1 Euler-Verfahren 136
8.2.2 Einschrittverfahren der Konsistenzordnung $p = 2$ . . . 136
8.3 Verfahren höherer Ordnung 138
8.4 Einige konkrete Runge-Kutta-Verfahren und deren Butcher-Tabellen 141
8.5 Schrittweitensteuerung bei Einschrittverfahren 145
8.5.1 Schrittweitensteuerung durch Einbettung 145
8.6 Mehrschrittverfahren 147
8.7 Allgemeine lineare Mehrschrittverfahren 150
8.8 Stabilität von Mehrschrittverfahren 154
8.9 Begriff der absoluten Stabilität 156
8.10 BDF-Verfahren 158

Kapitel 0

Vorwort

Als Literaturempfehlungen seien z.B. die Lehrbücher von

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

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

Kapitel 1

Rechnerarithmetik

1. Vorlesung am 20.10.2014

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

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

1.1 Zahldarstellungen

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

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

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

1.2 Allgemeine Gleitpunkt-Zahlensysteme

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

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

Definition 1.1. *Zu gegebener Basis $b \geq 2$ und* ***Mantissenlänge*** *$t \in \mathbb{N}$ sowie für Exponentenschranken $e_{min} < 0 < e_{max}$ ist die Menge*

$$F = F(b, t, e_{min}, e_{max}) \subset \mathbb{R}$$

durch

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

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

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

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

auch

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

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

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

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

1.3 Struktur und Eigenschaften des normalisierten Gleitpunktzahlensystems F

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

- $\hat{F}(2, 23, -126, 127)$ (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 Addition/Subtraktion bzw. Multiplikation/Division nicht abgeschlossen sind, d.h. Addition oder Multiplikation von Zahlen $x, y \in F$ ergibt i.A. keine Zahl aus F.

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

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

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

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

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

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

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

Man kann nun folgende Eigenschaften für das Runden zeigen:

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

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

Beweis. Es gilt offensichtlich

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

Nach Multiplikation mit b^e erhält man

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

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

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

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

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

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

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

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

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

Damit ist (1.12) gezeigt.
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 $|x| \geq b^{e_{\min}}$

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

als Abschätzung für den relativen Rundungsfehler. □

2. Vorlesung am 22.10.2014

Definition 1.5. *$u = \frac{1}{2} b^{-t+1}$ als Schranke für den relativen Rundungsfehler heißt* **Rundungseinheit** *(roundoff unit).*
$\operatorname{eps} = \epsilon_M = b^{-t+1}$ $(= 2\,u)$*, also der relative Abstand zweier Maschinenzahlen aus dem Intervall $[b^{e-1}, b^e]$, heißt* **Maschinenepsilon***.*

Bemerkung 1.6. Kennt man das Gleitpunktzahlsystem, speziell b und t nicht, dann kann man mit

$$\operatorname{eps} = 2 \inf\{\delta > 0 : \operatorname{rd}(1 + \delta) > 1\} = \min_{y \in F \setminus \{1\}} |1 - y|$$

das Maschinenepsilon finden (Abstand von 1 zur nächsten Maschinenzahl).
Überlegen Sie sich als Übung, weshalb dieser Sachverhalt zutrifft.
In Matlab oder Octave kann man eps mit der Befehlsfolge

```
e = 1;
while ( 1+e > 1) e=e/2;
end;
e = 2*e
```

im Handumdrehen bestimmen. Man findet

$$\text{eps} = 2.2204 \cdot 10^{-16} .$$

Da in Matlab standardmäßig mit

$$F(2, 53, -1021, 1024)$$

also mit IEEE-double-Gleitpunktzahlen gearbeitet wird, findet man für eps mit der obigen Formel

$$\text{eps} = 2^{1-53} = 2^{-52} = 2.2204 \cdot 10^{-16}$$

das gleiche Ergebnis wie mit dem kleinen Matlab-Programm.

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

Definition 1.7. *Zu $F = F(b, t, e_{min}, e_{max})$ ist die Funktion*

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

durch

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

erklärt.

Bemerkung 1.8. Abschneiden ist i.A. ungenauer als Runden und es gilt

$$\frac{|\,\text{tc}(x) - x|}{|x|} \leq 2 \cdot u$$

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

1.5 Ersatzarithmetik

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

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

oder

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

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

Satz 1.9. *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*

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

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 Theorem 1.4. Die Darstellung (1.17) ergibt sich im Falle von (1.15) aus

$$\frac{|\operatorname{rd}(x \circ y) - (x \circ y)|}{|x \circ y|} \leq u$$

also aus (1.14). □

1.6 Fehlerakkumulation

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

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

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

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

wobei man benutzt, dass

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

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

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

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

Satz 1.10. *Zu dem Gleitpunktzahlensystem $F(b,t,e_{min},e_{max})$ seien die Zahlen $x,y\in\mathbb{R}$ und $\Delta x,\Delta y\in\mathbb{R}$ gegeben mit $x+\Delta x\in F, y+\Delta y\in F$, $\frac{|\Delta x|}{|x|},\frac{|\Delta y|}{|y|}\leq \epsilon$ mit $\epsilon<\frac{1}{4}$. $\circ$ steht für die Grundoperation $\cdot$ bzw. : und für $x\circ y$ soll $x_{min}\leq|x\circ y|\leq x_{max}$ gelten. Dann gilt die Fehlerdarstellung*

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

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

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

Satz 1.11. *Zu $F(b,t,e_{min},e_{max})$ seien $x_1,\ldots,x_n\in\mathbb{R}$ und $\Delta x_1,\ldots,\Delta x_n\in\mathbb{R}$ Zahlen mit*

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

und es bezeichne

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

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

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

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

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

Beweis. (nach Plato)
Der Beweis erfolgt mit vollst. Induktion. Der Induktionsanfang ($k=1$) ist wegen $\frac{|\Delta x_1|}{|x_1|}\leq\epsilon$ offensichtlich. Unter der Annahme, dass (1.20) für $k-1\geq1$ richtig ist, ergibt sich mit den Verabredungen

$$\Delta S_j=\tilde{S_j}-S_j\quad\text{für } j\geq1,\ \Delta S_0=0$$

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

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

und damit

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

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

Bemerkung 1.12. Der Faktor $(1+\epsilon)^{n-j}$ in der Abschätzung (1.20) ist umso größer, je kleiner j ist. Daher ist es vorteilhaft beim Aufsummieren mit den betragsmäßig kleinen Zahlen zu beginnen. Theorem 1.11 liefert mit (1.20) nur eine Abschätzung für den absoluten Fehler. Der relative Fehler $\frac{|\tilde{S_n} - S_n|}{|S_n|}$ kann jedoch groß ausfallen, falls $|S_n|$ klein gegenüber $\sum_{j=1}^{n-1}(|x_j| + |S_j|) + |x_n|$ ist!

Definition 1.13. *(Landausche $\mathcal{O}$-Notation)*
Sei $h : U \to \mathbb{R}^n$ eine Funktion, U offen, $x_0 \in U$, Dann bezeichnet das Landau-Symbol[1]

$$\varphi(x) = \mathcal{O}(h(x)), \;\; x \to x_0$$

eine (nicht näher spezifizierte) Funktion φ mit der Eigenschaft

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

Das Landau-Symbol

$$\varphi(x) = o(h(x)), \;\; x \to x_0$$

beschreibt eine (nicht näher spezifizierte) Funktion φ mit der Eigenschaft

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

Korollar 1.1. *Es gilt*

$$\varphi(x) = o(h(x)), \; x \to x_0 \quad \Longrightarrow \quad \varphi(x) = \mathcal{O}(h(x)), \; x \to x_0 \; .$$

[1] Landau, Edmund Georg Hermann 1877-1938

Für Funktionen g, h sind die Notationen

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

eine abkürzende Schreibweise für

$$\begin{aligned} g(x) &= h(x) + r(h(x)), \;\; x \to x_0, \\ g(x) &\leq h(x) + r(h(x)), \;\; x \to x_0 \; (\textit{komponentenweise}), \end{aligned}$$

mit jeweils

$$r(h(x)) = o(h(x)), \;\; x \to x_0,$$

also

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

Beispiel 1.14.

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

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

Kapitel 2

Stabilität, Vorwärtsanalyse, Rückwärtsanalyse

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

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

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

2.1 Kondition als Maß für Fehlerverstärkungen

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

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

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

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

Bemerkung 2.2. κ klein bedeutet grob ein gut konditioniertes Problem, κ gross ein schlecht konditioniertes.

Beispiel 2.3. f differenzierbar

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

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

weil

$$err(\tilde{x} - x) = o(\tilde{x} - x) \quad \text{also} \quad \lim_{\tilde{x} \to x} \frac{\|err(\tilde{x} - x)\|}{\|\tilde{x} - x\|} = 0 \, .$$

2.2 Vektor- und Matrixnormen

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

Durch Vektornormen induzierte Matrixnormen

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

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

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

Weitere Vektor- und Matrixnormen

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

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

induziert $\|A\|_p$
Frobeniusnorm 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.

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

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

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

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

3. Vorlesung am 27.10.2014

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Beweis. Größtenteils trivial □

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

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

die Beziehungen

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

bzw.

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

Den Verstärkungsfaktor des max. relativen Fehlers

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

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

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

gilt.

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

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

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

$$\kappa_{\text{rel}} = \frac{\|f'(x, y)\|_1 \left\| \binom{x}{y} \right\|_1}{\|f(x, y)\|_1} = \frac{|x| + |y|}{|y|}\,,$$

wobei hier $\|f'(x, y)\|_1 = \|[y \;\; x]\|_1 = \max\{|y|, |x|\} = |x|$ (Spaltensummennorm) war, d.h. die relative normweise Kondition kann sehr groß sein.
Überprüfen sie als Übung, dass man mit den Normen $\|\;\|_\infty$ das gleiche Resultat für die normweise Kondition erhält.

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

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

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

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

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

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

bezüglich einer verträglichen Norm.

2.3 Stabilitätskonzepte

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

2.3.1 Vorwärtsanalyse

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

$$\frac{\left\|\tilde{f}(x) - f(x)\right\|}{\|f(x)\|}$$

mit dem unvermeidbaren Fehler $\kappa_{rel}\, eps$, der bei gerundeten Eingaben $\tilde{x}$ zu erwarten wäre.

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

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

Ein Algorithmus $\tilde{f}$ wird **vorwärtsstabil** *genannt, wenn σ nicht "übermäßig" groß aber auf jeden Fall beschränkt ist.*

Bemerkung 2.10. Die Elementaroperationen sind vorwärtsstabil.

Wir hatten bereits in einer vergangenen Vorlesung festgestellt, dass es bei Algorithmen, die aus mehreren vertauschbaren Teilschritten bestehen, mitunter sinnvoll ist mit bestimmten Teilschritten zu beginnen.
Betrachten wir nun ein Problem (f, x), das in 2 Teilprobleme (g, x) und $(h, g(x))$ (verketteter Algorithmus) aufgeteilt werden kann, d.h. man hat im skalaren Fall

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

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

Satz 2.11.
Seien κ_f, κ_h, κ_g die normweisen relativen Konditionen der Algorithmen f, h, g. Für den Stabilitätsindikator σ_f des verketteten Algorithmus $\tilde{f}$ gilt

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

Beweis. Es gilt

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

□

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

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

denn es gilt offensichtlich

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

2.3.2 Rückwärtsanalyse

Die Rückwärtsanalyse bedeutet die Interpretation des Ausgabefehlers des Algorithmus als Eingabefehler des Problems:

$$\tilde{f}(\tilde{x}) = f(\hat{x}) \; .$$

Man vergleicht

$$\frac{\|\hat{x} - \tilde{x}\|}{\|\tilde{x}\|}$$

mit eps.

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

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

Der Algorithmus heißt **rückwärtsstabil** *bezüglich des relativen Eingabefehlers δ, falls*

$$\eta \leq c\delta$$

mit nicht "übermäßig" großem, aber in jedem Fall beschränkten c gilt. Für den durch Rundung verursachten Eingabefehler $\delta = \text{eps}$ wird durch den Quotienten

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

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

Satz 2.13. *Für die Stabilitätsindikatoren σ und σ_R der Vorwärts- bzw. Rückwärtsanalyse gilt*

$$\sigma \leq \sigma_R$$

(aus der Rückwärtsstabilität folgt die Vorwärtsstabilität).

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

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

Damit ergibt sich mit

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

für eps $\to 0$, und wegen (2.9) ist der Stabilitätsindikator der Vorwärtsstabilität σ kleiner oder gleich dem Stabilitätsindikator der Rückwärtsstabilität σ_R. $\square$

Beispiel 2.14. (Rückwärtsanalyse)
Wir wollen als Beispiel die Rückwärtsanalyse der Berechnung der Lösung einer quadratischen Gleichung $x^2 - 2px + q = 0$, also

$$f(p,q) = p - \sqrt{p^2 - q}\ ,$$

bzw. die Näherung auf dem Computer

$$\tilde{f}(p,q) = p \tilde{-} \tilde{\sqrt{}}(p \tilde{\times} p \tilde{-} q)$$

durchführen. $\tilde{\sqrt{}} \approx \sqrt{}$ soll die Wurzel ohne großen Fehler ziehen. In der Berechnung stecken 4 fehlerhafte Operationen, die jeweils mit einem relativen Fehler ϵ_i, $|\epsilon_i| \leq eps$ behaftet sind. Man findet nun

$$\begin{aligned}
\tilde{f}(p,q) &= [p - \tilde{\sqrt{}}(p \tilde{\times} p \tilde{-} q)](1+\epsilon_4) \\
&= [p - \sqrt{p \tilde{\times} p \tilde{-} q}(1+\epsilon_3)](1+\epsilon_4) \\
&= [p - \sqrt{(p^2(1+\epsilon_1) - q)(1+\epsilon_2)}(1+\epsilon_3)](1+\epsilon_4)\ . \qquad (2.10)
\end{aligned}$$

Der Algorithmus hat also die folgenden Schritte

$$\begin{aligned}
y_1 &= p^2 \quad \text{relativer Fehler } \epsilon_1 \\
y_2 &= y_1 - q \quad \text{relativer Fehler } \epsilon_2 \\
y_3 &= \sqrt{y_2} \quad \text{relativer Fehler } \epsilon_3 \\
\tilde{f} &= p - y_3 \quad \text{relativer Fehler } \epsilon_4\ .
\end{aligned}$$

Die Rückwärtsanalyse bedeutet nun die Bestimmung von Eingangsfehlern Δp, Δq, so dass gilt

$$\tilde{f}(p,q) = f(p + \Delta p, q + \Delta q),$$

und die Bewertung der fehlerhaften Eingabe.
Dazu wird $\tilde{f}(p,q)$ umgeformt. Ausgehend von (2.10) erhält man

$$\begin{aligned}
\tilde{f} &= p(1+\epsilon_4) - \\
&\quad \sqrt{p^2(1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3)^2(1+\epsilon_4)^2 - q(1+\epsilon_2)(1+\epsilon_3)^2(1+\epsilon_4)^2} \\
&\approx (p+\Delta p) + \sqrt{(p+\Delta p)^2(1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3)^2 - q(1+\epsilon_6)} \\
&\approx (p+\Delta p) - \sqrt{(p+\Delta p)^2(1+\epsilon_5) - q(1+\epsilon_6)} \\
&= (p+\Delta p) - \sqrt{(p+\Delta p)^2 - q\{(1+\epsilon_6) - \epsilon_5 p^2(1+\epsilon_4)^2/q\}} \\
&= (p+\Delta p) - \sqrt{(p+\Delta p)^2 - q(1+\epsilon_7)}
\end{aligned}$$

mit

$$\epsilon_7 = \epsilon_6 - \epsilon_5 \frac{p^2}{q}(1+\epsilon_4)^2 \ , \quad \Delta p = p\epsilon_4 \ , \quad \Delta q = q\epsilon_7 \ .$$

Hierbei haben wir Abschätzungen zur Größenordnung wie folgt vorgenommen:

$$\begin{aligned} (1+\epsilon_3)^2 &= 1 + 2\epsilon_3 + \epsilon_3^2 \sim 1 + 2\epsilon_3 \\ (1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3)^2 &\sim (1+\epsilon_1+\epsilon_2)(1+2\epsilon_3) \\ &\sim 1 + \epsilon_1 + \epsilon_2 + 2\epsilon_3 \\ &=: 1 + \epsilon_5 \ , \end{aligned}$$

wobei $|\epsilon_5| < 4\,eps$ gilt.
Unter Berücksichtigung von $|\epsilon_i| \leq eps$, $i = 1, \ldots, 4$, findet man

$$|\epsilon_5| \leq 4\,eps \ , \quad |\epsilon_6| \leq 5\,eps \ , \quad (1+\epsilon_4)^2\epsilon_5 \dot{\leq} \epsilon_5$$

und damit (unter Nutzung der Näherung $(1+\epsilon_4)^2 \sim 1 + 2\epsilon_4$)

$$|\epsilon_7| < eps\,(5 + 4\frac{p^2}{|q|}) \ .$$

Da die Schranke für ϵ_7 und damit auch für Δq recht gross werden kann (im Fall $p^2 >> |q|$) ist das Berechnungsverfahren nicht in jedem Fall rückwärts stabil.

Kapitel 3

Lösung linearer Gleichungssysteme

3.1 LR-Zerlegung

4. Vorlesung am 29.10.2014

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

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

und löst zuerst

$$Ly = b$$

und danach

$$Rx = y$$

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

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

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

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

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

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

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

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

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

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

wobei

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

gilt. Wir haben also die Gleichung

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

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

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

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

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

wobei

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

die gewünschte untere Dreiecksmatrix ist.

3.1.1 Realisierung mit dem Gaußschen Eliminationsverfahren

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

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

Multiplikation $L_{ij}(\lambda)A$ bewirkt die Addition des 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

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

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

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

mit

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

e_k sei der k-te Standardeinheitsvektor

Definition 3.2.

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

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

Eigenschaften von M_k

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

$$M_k a_k = \begin{bmatrix} a_{1k} \\ \vdots \\ a_{kk} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

d.h. $a_{1k}, ..., a_{kk}$ bleiben bei der Multiplikation mit M_k unverändert

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

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

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

eine obere Dreiecksmatrix R. Außerdem hat die Matrix

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

die inverse Matrix

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

sodass schließlich mit

$$A = LR$$

eine LR-Zerlegung vorliegt.

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

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

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

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

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

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

Beweis. a) A besitze LR-Zerlegung

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

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

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

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

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

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

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

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

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

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

und

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

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

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

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

□

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

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

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

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

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

also die nicht redundanten Elemente von L

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

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

5. Vorlesung am 3.11.2014

Fehleranalyse bei der Konstruktion einer LR-Zerlegung

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

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

mit

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

Beweis. siehe Golub/van Loan, Matrix Computations oder Deuflhard, Numerische Mathematik □

Beim Vorwärtseinsetzen bzw. Rückwärtseinsetzen ("fill in") stellt man durch eine recht einfache Rückwärtsfehleranalyse das folgende Resultat fest:

Korollar 3.1. *Für die auf dem Computer berechneten (fehlerbehafteten) Ergebnisse $\tilde{y}$, $\tilde{x}$ der Gleichungssysteme*

$$L\,y = b\,, \qquad R\,x = y\,,$$

mit einer unipotenten unteren Dreiecksmatrix L und einer oberen Dreiecksmatrix R vom Typ $(n \times n)$ gelten die Beziehungen und Abschätzungen

$$\begin{aligned}(L+F)\tilde{y} &= b\,, \quad |F| \leq n\,\text{eps}|L| + O(\text{eps}^2)\\ (R+G)\tilde{x} &= y\,, \quad |G| \leq n\,\text{eps}|R| + O(\text{eps}^2)\,.\end{aligned}$$

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

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

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

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

Beweis. Rückwärtseinsetzen bzw. die Aussagen des Hilfssatzes 3.1 ergeben

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

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

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

□

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

3.1.2 LR-Zerlegung mit Spaltenpivotisierung

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

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

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

$$P \cdot P = E$$

gilt, d.h. die Matrix gleich ihrer Inversen ist.

Beispiel 3.8. Betrachten wir als Beispiel die Matrix aus dem obigen Beispiel

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

Um in den Frobeniusmatrizen große Einträge zu verhindern, vertauschen wir durch die Multiplikation mit der elementaren Permutationsmatrix

$$P_1 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

und erhalten damit

$$P_1 A = \begin{pmatrix} 6 & 5 & 23 \\ 4 & 3 & 11 \\ 2 & 1 & 3 \end{pmatrix}$$

Mit der Gauß-Transformation

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

erhält man nun

$$M_1 P_1 A = \begin{pmatrix} 6 & 5 & 23 \\ 0 & 3-10/3 & 11-46/3 \\ 0 & 1-5/3 & 3-23/3 \end{pmatrix} = \begin{pmatrix} 6 & 5 & 23 \\ 0 & -1/3 & -13/3 \\ 0 & -2/3 & -14/3 \end{pmatrix}$$

Nun pivotisiert man wieder, indem man die 2. und die 3. Zeile vertauscht, also mit Hilfe der elementaren Permutationsmatrix

$$P_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

das Zwischenergebnis

$$P_2 M_1 P_1 A = \begin{pmatrix} 6 & 5 & 23 \\ 0 & -2/3 & -14/3 \\ 0 & -1/3 & -13/3 \end{pmatrix}$$

erhält. Mit der Gauß-Transformation

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

erhält man schließlich

$$M_2 P_2 M_1 P_1 A = \begin{pmatrix} 6 & 5 & 23 \\ 0 & -2/3 & -14/3 \\ 0 & 0 & -2 \end{pmatrix} =: R\,, \tag{3.2}$$

eine Matrix-Faktorisierung, die man zur Lösung von linearen Gleichungssystemen nutzen kann.

Die Erfahrungen des Beispiels kann man zusammenfassen.

Definition 3.9. *Wir bezeichnen den im Beispiel beschriebenen Algorithmus als Konstruktion einer LR-Zerlegung mit Spaltenpivotisierung (auch Gaußelimination mit partieller Pivotisierung).*

Satz 3.10. *Für die Gaußelimination mit partieller Pivotisierung mit dem Resultat*

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

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

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

mit

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

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

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

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

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

$$\hat{M}_{n-1}\hat{M}_{n-2}\cdots\hat{M}_1PA \quad \text{mit} \quad L = \hat{M}_1^{-1}\cdots\hat{M}_{n-1}^{-1}$$

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.

□

Beispiel 3.11. Zurück zum obigen Beispiel. Mit $P_2P_2 = E$ können wir schreiben

$$M_2P_2M_1P_1A = M_2P_2M_1P_2P_2P_1A = \hat{M}_2\hat{M}_1PA = R\,,$$

wobei

$$\hat{M}_2 = M_2 \quad , \quad \hat{M}_1 = P_2M_1P_2 \quad \text{und}$$

und für die Permutationsmatrix

$$P = P_2P_1 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

gilt. Durch die entsprechende Multiplikation mit den sehr einfach zu bestimmenden Inversen der Frobeniusmatrizen erhält man

$$PA = \hat{M}_2^{-1}\hat{M}_1^{-1}R =: LR = \begin{pmatrix} 1 & 0 & 0 \\ 1/3 & 1 & 0 \\ 2/3 & 1/2 & 1 \end{pmatrix} \begin{pmatrix} 6 & 5 & 23 \\ 0 & -2/3 & -14/3 \\ 0 & 0 & -2 \end{pmatrix}$$

und damit letzendlich die LR-Zerlegung mit partieller Pivotisierung

$$PA = LR\,.$$

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 3.6 führt.

3.2 Cholesky-Zerlegung

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

- numerische Lösung elliptischer und parabolischer Differentialgleichungen
- Spline-Approximation

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

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

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

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

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

(auch Kriterium von Hurwitz)

Satz 3.12. *Sei A symmetrisch und positiv definit. Dann existiert eine untere Dreiecksmatrix $G \in \mathbb{R}^{n\times n}$ mit positiven Diagonalelementen, sodass*

$$A = GG^T$$

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

$$A = LR$$

mit L untere Dreiecksmatrix mit 1-Diagonale und R obere Dreiecksmatrix, d.h. der Algorithmus zur Konstruktion der LR-Zerlegung kann ohne Probleme durchgeführt werden.
Betrachten wir mit M_1 die Frobeniusmatrix zur Erzeugung einer Nullspalte unter dem Element a_{11}, dann erhält man mit

$$M_1 A M_1^T$$

eine symmetrische Matrix der Gestalt

$$\begin{pmatrix} a_{11} & 0 & \ldots & 0 \\ 0 & \# & \ldots & \# \\ \vdots & & & \\ 0 & \# & \ldots & \# \end{pmatrix}$$

und schließlich

$$M_{n-1} M_{n-2} \cdot \cdots \cdot M_1 A M_1^T \cdot \cdots \cdot M_{n-1}^T = D \tag{3.3}$$

mit einer Diagonalmatrix D.
Die Diagonalelemente d_{jj} von D sind positiv, weil einmal

$$\det[M_{n-1} M_{n-2} \cdot \cdots \cdot M_1 A M_1^T \cdot \cdots \cdot M_{n-1}^T](1:k, 1:k) = \det A(1:k, 1:k) = \prod_{j=1}^{k} d_{jj}$$

für alle $k = 1, \ldots, n$ gilt, und mit der Voraussetzung $\det A(1:k, 1:k) > 0$ sukzessiv $d_{jj} > 0$ für $j = 1, \ldots, n$ folgt. Damit können wir

$$D^{1/2} = \operatorname{diag}(\sqrt{d_{11}}, \ldots \sqrt{d_{nn}})$$

bilden und erhalten unter Nutzung von (3.3) durch Multiplikation von links und rechts mit geeigneten Matrizen mit

$$G = M_1^{-1} \cdot \cdots \cdot M_{n-1}^{-1} D^{1/2}$$

die untere Dreiecksmatrix, mit der

$$A = GG^T$$

gilt. □

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 $j > k$

$$a_{jk} = g_{j1}g_{k1} + g_{j2}g_{k2} + \cdots + g_{jk-1}g_{kk-1} + g_{jk}g_{kk}$$

$$\Rightarrow g_{jk} = \frac{1}{g_{kk}} \left(a_{jk} - \sum_{i=1}^{k-1} g_{ji}g_{ki} \right)$$

Pseudocode:

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

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

3.3 Singulärwertzerlegung

6. Vorlesung am 05.11.2014

Motivation

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

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

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

Satz 3.13. *(und Definition)*
Gegeben sei eine Matrix $A \in \mathbb{R}^{m\times n}$. Dann gibt es orthogonale Matrizen $U \in \mathbb{R}^{m\times m}$ und $V \in \mathbb{R}^{n\times n}$ sowie eine Matrix $\Sigma = (s_{ij}) \in \mathbb{R}^{m\times n}$ mit $s_{ij} = 0$ für alle $i \neq j$ und nichtnegativen Diagonalelementen $s_{11} \geq s_{22} \geq \ldots$, für die

$$A = U\Sigma V^T \Longleftrightarrow U^T A V = \Sigma \tag{3.4}$$

gilt. Die Darstellung (3.4) *heißt* **Singulärwertzerlegung** *von A. Die Werte $\sigma_i = s_{ii}$ heißen* **Singulärwerte** *von A.*

Bevor der Satz 3.13 bewiesen wird sei darauf hingewiesen, dass man die Gleichung (3.4) auch in der Form

$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^T \tag{3.5}$$

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

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

$$U^T A V = \begin{pmatrix} \sigma & \mathbf{0} \\ \mathbf{0} & B \end{pmatrix} \tag{3.6}$$

mit einer Zahl σ und einer Matrix $B \in \mathbb{R}^{(m-1)\times(n-1)}$ gilt. Der Beweis der Beziehung (3.4) ergibt sich dann induktiv, indem man B in der Art (3.6) faktorisiert usw.
Sei $\sigma := ||A||_2 = \max_{||x||=1} ||Ax||$. Das Maximum wird angenommen mit Vektoren $u \in \mathbb{R}^m$ und $v \in \mathbb{R}^n$, so dass

$$Av = \sigma u \quad \text{und} \quad ||u||_2 = ||v||_2 = 1 \tag{3.7}$$

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

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

ergänzt und damit sind

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

orthogonale Matrizen. Das Produkt $U^T A V$ hat wegen (3.7) die Form

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

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

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

also

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

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

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

folgt, und damit muss $w = \mathbf{0}$ gelten. Damit ist (3.6) nachgewiesen und mit dem Hinweis auf die Induktion ist der Satz bewiesen. □

Mögliche Konstruktion der Singulärwertzerlegung (konstruktiver Nachweis)

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

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

und aufgrund der Definition von B

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

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

2) Für $j = 1, \ldots, r$ wird

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

gesetzt.

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

4) Man bildet die Matrizen

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

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

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

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

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

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

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

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

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

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

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

iv) Es gilt schließlich ausgehend von der Formel (3.5)

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

Bemerkung 3.14. Es gilt nun im Weiteren

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

2. Es gilt

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

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

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

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

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

auch eine Singulärwertzerlegung.

Beispiel 3.15. Mit dem eben diskutierten Verfahren zur Konstruktion einer Singulärwertzerlegung findet man für die Matrix

$$A = \begin{pmatrix} 1 & 0 \\ 2 & 1 \\ 0 & 1 \end{pmatrix}$$

die Faktorisierung

$$A = U\Sigma V^T$$

mit

$$U = \begin{pmatrix} \frac{2}{\sqrt{30}} & \frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{6}} \\ \frac{5}{\sqrt{30}} & 0 & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{30}} & -\frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{6}} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sqrt{6} & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad V = \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}} \end{pmatrix},$$

über die Eigenwerte und Eigenvektoren von

$$B = A^T A = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}$$

wie oben beschrieben.

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

Satz 3.16. *(Schmidt-Mirsky, beste Rang-k-Approximation)*
Es sei $A \in \mathbb{R}^{m\times n}$, $m \geq n$, eine Matrix vom Rang r mit der Singulärwertzerlegung

$$A = U\Sigma V^T = \sum_{j=1}^{r} \sigma_j u_j v_j^T \ . \tag{3.8}$$

Die Approximationsaufgabe

$$\min_{B\in\mathbb{R}^{m\times n}, rg(B)\leq k} \|A - B\|_2$$

besitzt für $k < r$ die Lösung

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

A_k ist zugleich Lösung der Approximationsaufgabe

$$\min_{B\in\mathbb{R}^{m\times n}, rg(B)\leq k} \|A - B\|_F \ , \ mit \quad \|A - A_k\|_F = \sqrt{\sum_{j=k+1}^{r} \sigma_j^2} \ . \tag{3.9}$$

Beweis. (Frobeniusnorm)
Für $k = n$ sind die Aussagen trivial, da dann $A = A_k$ gilt.
Die Aussage (3.9) soll zuerst gezeigt werden. Mit der Frobeniusnorm stellt man fest, dass

$$\begin{aligned} ||A - A_k||_F^2 &= ||U\Sigma V^T - U\Sigma_k V^T||_F^2 = ||U(\Sigma - \Sigma_k)V^T||_F^2 \\ &= \sum_{i=k+1}^{r} \sigma_i^2 = \sum_{i=1}^{r} \sigma_i^2 - \sum_{i=1}^{k} \sigma_i^2 = ||A||_F^2 - \sum_{i=1}^{k} \sigma_i^2 \ , \end{aligned}$$

wobei Σ_k aus Σ durch Nullsetzung aller σ_i mit $i > k$ entsteht. Für den Nachweis, dass A_k die beste Approximation von A bezügl. der Frobeniusnorm ist, genügt es nun z.z., dass

$$||A - \sum_{i=1}^{k} x_i y_i^T||_F^2 \geq ||A||_F^2 - \sum_{i=1}^{k} \sigma_i^2 \tag{3.10}$$

für beliebige Vektoren $x_i \in \mathbb{R}^m$, $y_i \in \mathbb{R}^n$ gilt. Ohne die Allgemeinheit einzuschränken können wir annehmen, dass die Vektoren $x_1, \ldots, x_k$ orthonormal sind. Denn falls sie es nicht sind, können wir sie in eine Orthonormalbasis $\{o_1, \ldots, o_k\}$ entwickeln, und statt

$$\sum_{i=1}^{k} x_i y_i^T \quad \text{mit} \quad \sum_{i=1}^{k} o_i \tilde{y}_i^T$$

arbeiten, wobei die Koordinaten von $\tilde{y}_i$ sich aus den Koordinaten von x_i in der ONB $\{o_1, \ldots, o_k\}$ ergeben. Nun gilt

$$
\begin{aligned}
||A - \sum_{i=1}^{k} x_i y_i^T||_F^2 &= \text{spur}((A - \sum_{i=1}^{k} x_i y_i^T)^T (A - \sum_{i=1}^{k} x_i y_i^T)) \\
= \text{spur}(A^T A + \sum_{i=1}^{k} (y_i - A^T x_i)(y_i - A^T x_i)^T - \sum_{i=1}^{k} A^T x_i x_i^T A) & \\
\geq \text{spur}(A^T A) - \sum_{i=1}^{k} \text{spur}(A^T x_i x_i^T A) & \\
= ||A||_F^2 - \sum_{i=1}^{k} ||A^T x_i||_F^2 \, , &
\end{aligned}
$$

da $\text{spur}((y_i - A^T x_i)(y_i - A^T x_i)^T) \geq 0$ ist. Also genügt es zum Nachweis von (3.10) z.z., dass

$$\sum_{i=1}^{k} ||A^T x_i||_F^2 \leq \sum_{i=1}^{k} \sigma_i^2$$

gilt. Jetzt ersetzen wir A^T mit der Singulärwertzerlegung $U\Sigma V^T$ und teilen diese wie folgt auf:

$$
\begin{aligned}
V_1 &= (v_1 \ldots v_k \, 0 \ldots 0) \\
V_2 &= (0 \ldots 0 \, v_{k+1} \ldots v_m) \, .
\end{aligned}
$$

Σ_1 und Σ_2 werden aus Σ auf analoge Weise gebildet. Damit gilt nun

$$||A^T x_i||_F^2 \;=\; ||U\Sigma V^T x_i||_F^2 = ||\Sigma V^T x_i||_F^2 = ||\Sigma_1 V_1^T x_i||_F^2 + ||\Sigma_2 V_2^T x_i||_F^2 =: (*)$$

Nun addieren wir 2 "nahrhafte Nullen" zu (*) und erhalten

$$
\begin{aligned}
(*) \;&=\; ||\Sigma_1 V_1^T x_i||_F^2 + ||\Sigma_2 V_2^T x_i||_F^2 + \\
&\quad \underbrace{\sigma_k^2 - \sigma_k^2}_{=0} + \sigma_k^2 \underbrace{(||V^T x_i||_F^2 - ||V_1^T x_i||_F^2 - ||V_2^T x_i||_F^2)}_{=0} \\
&=\; \sigma_k^2 + (||\Sigma_1 V_1^T x_i||_F^2 - \sigma_k^2 ||V_1^T x_i||_F^2) - \underbrace{(\sigma_k^2 ||V_2^T x_i||_F^2 - ||\Sigma_2 V_2^T x_i||_F^2)}_{(1)} \\
&\quad - \underbrace{\sigma_k^2 (1 - ||V^T x_i||_F^2)}_{(2)} \, .
\end{aligned}
$$

Da die Singulärwerte in Σ absteigend geordnet sind, ist der Term (1) nichtnegativ. Außerdem ist x_i ein orthonormaler Vektor und die Matrix V orthogonal, so dass der Term (2) ebenfalls nichtnegativ ist. Damit folgt nach

Summation

$$\begin{aligned}\sum_{i=1}^{k}||A^T x_i||_F^2 &\leq k\sigma_k^2 + \sum_{i=1}^{k}(||\Sigma_1 V_1^T x_i||_F^2 - \sigma_k^2||V_1^T x_i||_F^2)\\ &= k\sigma_k^2 + \sum_{i=1}^{k}\sum_{j=1}^{k}(\sigma_j^2-\sigma_k^2)|v_j^T x_i|^2\\ &= k\sigma_k^2 + \sum_{j=1}^{k}(\sigma_j^2-\sigma_k^2)\underbrace{\sum_{i=1}^{k}|v_j^T x_i|^2}_{\leq 1}\\ &\leq \sum_{j=1}^{k}[\sigma_k^2 + (\sigma_j^2-\sigma_k^2)] = \sum_{j=1}^{k}\sigma_j^2\end{aligned}$$

also der Beweis von (3.9). □

7. Vorlesung 10.11.14

Beweis. (Spektralnorm)
Für die entsprechende Aussage hinsichtlich der Spektralnorm sei $B \in \mathbb{R}^{m\times n}$ eine Matrix mit $rg(B) = k < n$. Es gilt dann $\dim(\ker(B)) = n - k$. Sind $v_1, \ldots, v_n$ die rechten Singulärvektoren (Spalten von V) von A, dann hat der Unterraum $\mathcal{V} = \text{span}\{v_1, \ldots, v_{k+1}\}$ die Dimension $k+1$. $\ker(B)$ und $\mathcal{V}$ sind jeweils Unterräume vom $\mathbb{R}^n$ mit

$$\dim(\ker(B)) + \dim(\mathcal{V}) = n - k \ + \ k + 1 = n + 1\ ,$$

so dass $\ker(B) \cap \mathcal{V} \neq \{\mathbf{0}\}$ gilt. Damit finden wir ein $x \in \ker(B) \cap \mathcal{V}$ mit $||x||_2 = 1$, dass man in der Form

$$x = \sum_{j=1}^{k+1}\alpha_j v_j \quad \text{mit} \quad \sum_{j=1}^{k+1}\alpha_j^2 = 1$$

darstellen kann. Es folgt nun

$$(A-B)x = Ax - \underbrace{Bx}_{=0,\, da\, x\in ker(B)} = \sum_{j=1}^{k+1}\alpha_j A v_j = \sum_{j=1}^{k+1}\alpha_j\sigma_j w_j$$

sowie

$$\begin{aligned}
||A-B||_2 &= \max_{||y||_2=1} ||(A-B)y||_2 \geq ||(A-B)x||_2 = ||\sum_{j=1}^{k+1} \alpha_j \sigma_j w_j||_2 \\
&= (\sum_{j=1}^{k+1} |\alpha_j \sigma_j|^2)^{1/2} \quad \text{(weil } w_1, \ldots, w_{k+1} \text{ paarweise orthogonal sind)} \\
&\geq \sigma_{k+1} (\sum_{j=1}^{k+1} |\alpha_j|^2)^{1/2} \quad \text{(weil } \sigma_1 \geq \cdots \geq \sigma_{k+1} \text{ gilt)} \\
&= \sigma_{k+1} = ||A - A_k||_2 \, .
\end{aligned}$$

□

Ein weitere Anwendung findet die Singulärwertzerlegung bei der

$$\textbf{Aufgabe} \quad \left\{ \begin{array}{l} \text{bestimme } x^* \text{ mit minimaler Euklidischer Länge,} \\ \text{für das } ||Ax^* - b||_2 = \min_{x \in R^n} ||Ax - b||_2 \text{ gilt,} \end{array} \right. \tag{3.11}$$

für eine gegebene Matrix $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$. Man kann zeigen, dass die Aufgabe (3.11) eine eindeutige Lösung besitzt, was wir aber als gesichert voraussetzen wollen.
Für die Bestimmung der Lösung können wir nun die Singulärwertzerlegung von A nutzen. Es gilt der

Satz 3.17. *(und Definition)*
Sei $U^T A V = \Sigma$ eine Singulärwertzerlegung von $A \in \mathbb{R}^{m \times n}$ mit Singulärwerten $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0$, $p = \min\{m, n\}$. Durch

$$A^+ = V \Sigma^+ U^T \quad \textit{mit} \quad \Sigma^+ = \textit{diag}(\sigma_1^{-1}, \ldots, \sigma_r^{-1}, 0, \ldots, 0) \in \mathbb{R}^{n \times m}$$

definieren wir mit A^+ die **Pseudoinverse** *von A.*
Dann ist $A^+ b = x^*$ die Lösung von (3.11).

Beweis. Für $b \in \mathbb{R}^m$, $x \in \mathbb{R}^n$ sei $\tilde{b} = U^T b$, $\tilde{x} = V^T x$. Unter Nutzung der Singulärwertzerlegung von A und der Längeninvarianz von orthogonalen Transformationen ergibt sich

$$\begin{aligned}
||Ax - b||_2^2 = ||U^T A V V^T x - U^T b||_2^2 &= ||\Sigma \tilde{x} - \tilde{b}||_2^2 \\
&= \sum_{i=1}^{r} [\sigma_i \tilde{x}_i - \tilde{b}_i]^2 + \sum_{i=r+1}^{m} \tilde{b}_i^2 \, .
\end{aligned} \tag{3.12}$$

Und aus (3.12) folgt, dass

$||Ax - b||_2$ für x^* minimal wird, wenn $\tilde{x}_i = (V^T x^*)_i = \tilde{b}_i/\sigma_i$ für $i = 1, \ldots, r$ gilt. Wegen $||x^*||_2 = ||V^T x^*||_2$ ist $||x^*||_2$ minimal genau dann, wenn $||V^T x^*||_2$ minimal ist, also falls $(V^T x^*)_i = 0$ für $i = r+1, \ldots, n$ gilt. Für die Lösung des Problems (3.11) mit minimaler Euklidischer Norm $x^* = A^+ b$ erhält man wegen $\tilde{b}_i = (U^T b)_i$, $i = 1, \ldots, r$ schließlich

$$V^T x^* = (\frac{\tilde{b}_1}{\sigma_1}, \ldots, \frac{\tilde{b}_r}{\sigma_r}, 0, \ldots, 0)^T = \Sigma^+ U^T b \Longleftrightarrow x^* = V \Sigma^+ U^T b = A^+ b \,.$$

□

Eine weitere Anwendung der Singulärwertzerlegung ist z.B. die Lösung von linearen Gleichungssystemen mit nahezu singulären Koeffizientenmatrizen.

Kapitel 4

Die iterative Lösung von Gleichungen bzw. Gleichungssystemen

Da die iterative Lösung linearer Gleichungssysteme auf ein äquivalentes Fixpunktproblem zurück geführt wird, soll hier an den Banachschen Fixpunktsatz erinnert werden.

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

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

mit $L < 1$ (Kontraktivität) für alle $x_1, x_2 \in \mathcal{A}$, dann hat F genau einen Fixpunkt $\hat{x} \in \mathcal{A}$ mit

$$F(\hat{x}) = \hat{x}$$

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

Bemerkung 4.2. Aus dem Banachschen Fixpunktsatz ergeben sich die Fehlerabschätzungen

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

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

Die Kontraktivität der Fixpunktabbildung werden wir im Folgenden auch als Konvergenz-Bedingung bei den iterativen Lösungsverfahren von linearen Gleichungssystemen wiederfinden.

4.1 Die iterative Lösung linearer Gleichungssysteme

Die iterative Lösung linearer Gleichungssysteme ist immer dann gefragt, wenn die Probleme/Matrizen so groß werden, dass der Computerspeicher nicht mehr ausreicht, um direkte Lösungsverfahren effizient zu implementieren. In den 40er Jahren des vergangenen Jahrhunderts mussten im Zusammenhang mit der Berechnung der Neutronen-Diffusion bei der Entwicklung von Nukleartechnologien sehr große lineare Gleichungssysteme mit sehr dünn besetzten Koeffizientenmatrizen gelöst werden. Da eine direkte Lösung damals kaum möglich war, wurden iterative Verfahren, die wir im folgenden diskutieren wollen, verwendet. Diese benötigten meistens nur einen Speicherplatz der Ordnung $\mathcal{O}(n)$. Obwohl die meisten dieser Verfahren heute durch wesentlich effizientere Verfahren abgelöst wurden, sollen sie dargelegt werden, auch weil viele Begriffe und Matrix-Eigenschaften auch bei heute anstehenden Aufgabenstellungen von Bedeutung sind.
Neben der schon beschriebenen direkten Lösung linearer Gleichungssysteme durch den Gaußschen Algorithmus oder durch bestimmte Matrix-Faktorisierungen ist es wie oben angemerkt in bestimmten Situationen sinnvoll, lineare Gleichungssysteme

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

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

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

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

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

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

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

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

Mit x und $\delta x_k = x - x_k$ folgt

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

also gilt für irgendeine Vektornorm und eine dadurch induzierte Matrixnorm

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

Damit konvergiert das Lösungsverfahren, wenn

$$\lim_{k\to\infty} S^k = 0 \quad \text{bzw.} \quad \lim_{k\to\infty} \left\|S^k\right\| = 0$$

gilt. Hilfreich zur Konvergenzuntersuchung ist der

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

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

(b) $S^k \to 0$ für $k \to \infty$

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

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

Die Punkte (a) und (b) bedeuten gerade die Kontraktivität der Fixpunktabbildung F.

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

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

Durch

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

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

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

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

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

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

Als Folgerung des Satzes 4.3 erhält man das folgende Konvergenzkriterium

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

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

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

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

mit

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

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

4.2 Jacobi-Verfahren oder Gesamtschrittverfahren

Die Wahl von $B = D$ ergibt die Iterationsmatrix

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

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

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

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

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

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

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

Zur Konvergenz des Jacobi-Verfahrens gilt der

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

Beweis.

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

Zeilensummen von S

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

aufgrund der strikten Diagonaldominanz ist

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

□

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

8. Vorlesung 12.11.14

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

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

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

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

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

gibt.

Bemerkung.

1. Man kann entscheiden, ob eine Matrix reduzibel oder irreduzibel ist, indem man für $A = (a_{ij})_{i,j=1,\ldots,n}$ 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. Weitere Kriterien zur Entscheidung ob A vom Typ $n \times n$ irreduzibel oder reduzibel ist, sind in den folgenden Lemmata, die hier nicht bewiesen werden, dargelegt.

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

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

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

Lemma 4.9. *Die $(n\times n)$-Matrix A ist irreduzibel, falls es für zwei beliebige, nichtleere, disjunkte Teilmengen S und T von $W = \{1, 2, \ldots, n\}$ mit $S \cup T = W$ stets Indexwerte $i \in S$ und $j \in T$ existieren, so dass $a_{ij} \neq 0$ ist.*

Da bei der Diskretisierung von elliptischen Randwertproblemen mit FV-, FD- oder FE-Methoden irreduzible diagonal dominante Matrizen entstehen, sollen in den folgenden Sätzen wichtige Eigenschaften dieser Matrizen gezeigt werden.

Satz 4.10. *Eine irreduzibel diagonal dominante Matrix $A \in \mathbb{R}^{n\times n}$ hat nichtverschwindende Diagonalelemente und ist regulär.*

Beweis. (nach Schwarz)
Zum Nachweis $a_{ii} \neq 0,\ i \in W = \{1, 2, \ldots, n\}$, nehmen wir an, dass es einen Index i mit $a_{ii} = 0$ gibt. Wegen der schwachen Diagonaldominanz müsste dann $a_{ij} = 0$ sein für alle $j \neq i$. Mit den Indexmengen $S := \{i\}$ und $T = W \setminus \{i\}$ steht dies im Widerspruch zur Irreduzibilität gemäß Kriterium 4.9, d.h. unsere Annahme war falsch.
Der Nachweis der Regularität von A wird mit der Annahme $\det A = 0$ auch indirekt geführt. Folglich besitzt das homogene lineare Gleichungssystem $Az = \mathbf{0}$ eine nichttriviale Lösung $z \neq \mathbf{0}$. Wegen $a_{ii} \neq 0$ können alle Gleichungen des linearen Systems nach z_i aufgelöst werden, d.h.

$$z_i = -\sum_{j=1, j\neq i}^{n} \frac{a_{ij}}{a_{ii}} z_j = \sum_{j=1}^{n} b_{ij} z_j \ , \quad i = 1, \ldots, n, \tag{4.10}$$

mit $b_{ii} = 0,\ b_{ij} = -a_{ij}/a_{ii},\ (i \neq j)$. Mit der schwachen Diagonaldominanz finden wir

$$\sum_{j=1}^{n} |b_{ij}| \leq 1 \ , \quad i = 1, \ldots, n, \tag{4.11}$$

wobei für mindestens einen Index i_0 in (4.11) die strikte Ungleichung gilt. Wir definieren $M = \max_i |z_i| > 0$ und es sei k ein Index, für den $|z_k| = M$ gilt. Aus (4.10) ergibt sich für die k-te Gleichung

$$M = |z_k| = |\sum_{j=1}^{n} b_{kj} z_j| \leq \sum_{j=1}^{n} |b_{kj}| \cdot |z_j| \ . \tag{4.12}$$

Wegen (4.11) gilt $\sum_{j=1}^{n} |b_{kj}| \cdot M \leq M$ und mit (4.12) ergibt sich

$$\sum_{j=1}^{n} |b_{kj}|(|z_j| - M) \geq 0 \ . \tag{4.13}$$

Da $|z_j| \leq M$ für alle j gilt, kann (4.13) nur dann erfüllt sein, wenn für alle Matrixelemente $b_{kj} \neq 0$ die Gleichheit $|z_j| = M$ gilt.
Aufgrund der Irreduzibilität existiert zu jedem Indexpaar (k, j) mit $k \neq j$ entweder das Matrixelement $a_{kj} \neq 0$ oder eine Indexfolge $k_1, k_2, \ldots, k_s$, so dass

$$a_{kk_1} a_{k_1 k_2} a_{k_2 k_3} \cdot \cdots \cdot a_{k_s j} \neq 0$$

ist. Damit muss entweder $b_{kj} \neq 0$ oder $b_{kk_1} b_{k_1 k_2} \cdot \cdots \cdot b_{k_s j} \neq 0$ sein. Im zweiten Fall kann man die Argumentation auch für den Index k_1 anwenden und wegen $b_{k_1 k_2} \neq 0$ ist $|z_{k_2}| = M$. Die Fortsetzung dieser Schlussweise ergibt schließlich, dass $|z_j| = M$ für jedes beliebige $j \neq k$ gelten muss. Für diejenige Gleichung mit dem Index i_0, für die (4.11) eine strikte Ungleichung ist, folgt wegen $|z_j| = M$ mit

$$M \leq \sum_{j=1}^{n} |b_{i_0 j}| \cdot M < M$$

der angestrebte Widerspruch und damit ist der Satz bewiesen. □

Damit wissen wir zumindest, dass lineare Gleichungssysteme mit irreduzibel diagonal dominanten Matrizen eindeutig lösbar sind. Mit dem folgenden Satz wird iterative Lösbarkeit mit dem Jacobi-Verfahren gezeigt.

Satz 4.11. *Für eine irreduzibel diagonal dominante Matrix A ist das Jacobi-Verfahren konvergent.*

Beweis. (nach Schwarz)
Der Beweis wird indirekt geführt, und zwar mit der Annahme $\rho(S_J) \geq 1$. Demnach existiert ein EW μ von S_J mit $|\mu| \geq 1$ und für ihn gelten

$$\det(S_J - \mu E) = 0 \quad \Longleftrightarrow \quad \det(E - \mu^{-1} S_J) = 0 \, .$$

Aus der Irreduzibilität von A folgt auch die Irreduzibilität von $S_J = D^{-1}(L + U)$, da hier nur die Nichtdiagonalelemente von Interesse sind. Das gleiche gilt für die Matrix $Q = E - \mu^{-1} S_J$. Außerdem erkennt man, dass Q schwach diagonal dominant ist, denn für die Elemente von S_J gilt

$$s_{ii} = 0 \, , \quad s_{ij} = -a_{ij}/a_{ii} \, , \; (j \neq i),$$

gilt, und wegen der schwachen Diagonaldominanz von A folgt also

$$\sum_{j=1, j \neq i}^{n} |s_{ij}| \leq 1$$

für alle i, wobei für mindestens einen Index i_0 die strikte Ungleichung gilt. Mit $|\mu^{-1}| \leq 1$ folgt die schwache Diagonaldominanz von Q, also auch die irreduzible Diagonaldominanz. Damit muss nach Satz 4.10 $\det Q = \det(E - \mu^{-1} S_J) \neq 0$ sein, was unserer Annahme widerspricht, und damit sind alle Eigenwerte von S_J dem Betrage nach kleiner als 1 und damit auch $\rho(S_J) < 1$, was die Konvergenz des Verfahrens impliziert. □

4.3 Gauß-Seidel-Iterationsverfahren oder Einzelschrittverfahren

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

$$x^{(k)} = \underbrace{(E - B^{-1}A)}_{S_{GS}} x^{(k-1)} + B^{-1}b = (L + D)^{-1}(-Ux^{(k-1)} + b), \quad k = 1, 2, \ldots \tag{4.14}$$

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

Satz 4.12. *Das Gauß-Seidel-Verfahren konvergiert für strikt diagonal dominante Matrizen A für beliebige Startiterationen $x^{(0)} \in \mathbb{R}^n$*

9. Vorlesung 17.11.14

Beweis. Es ist

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

für einen EW λ mit dem EV v, also

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

Wir betrachten die k-te Zeile:

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

Aus der strengen Diagonaldominanz folgt schließlich

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

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

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

für alle EW λ folgt. Damit ist $r(S_{GS}) < 1$ und der Satz bewiesen. □

Bemerkung 4.13. Ebenso wie beim Jacobi-Verfahren kann man die Voraussetzung der strikten Diagonaldominanz von A für die Konvergenz abschwächen, allerdings nicht bedingungslos.
Das Gauß-Seidel-Verfahren ist für irreduzibel diagonal dominante Matrizen A ebenso wie das Jacobi-Verfahren konvergent.

Wenn man das Gauß-Seidel Verfahren (4.14) in der äquivalenten Form

$$x^{(k)} = D^{-1}(-Lx^{(k)} - Ux^{(k-1)} + b), \quad k = 1, 2, \ldots \tag{4.15}$$

aufschreibt, erkennt man bei der koordinatenweisen Berechnung der neuen Iteration

$$x_j^{(k)} = \frac{1}{a_{jj}} \left(b_j - \sum_{i=1}^{j-1} a_{ji} x_i^{(k)} - \sum_{i=j+1}^{n} a_{ji} x_i^{(k-1)} \right), \quad j = 1, 2, \ldots, n, \ k = 1, 2, \ldots \tag{4.16}$$

zwar, dass auf beiden Seiten der Formeln $x^{(k)}$ vorkommen. Allerdings benötigt man zur Berechnung der j-ten Komponenten von $x^{(k)}$ nur die Komponenten $x_1^{(k)}, \ldots, x_{j-1}^{(k)}$ der vorigen Iteration. Diese kennt man aber bereits. Damit kann man die Formel (4.16) für $j = 1, \ldots, n$ sukzessiv zum Update der Koordinaten von x anwenden. Man hat also mit (4.16) eine explizite Berechnungsvorschrift und braucht damit $B = L + D$ nicht wirklich zu invertieren.

4.4 Verallgemeinerung des Gauß-Seidel-Verfahrens

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

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

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

$$x^{(k)} = \omega \hat{x}^{(k)} + (1-\omega)x^{(k-1)} \tag{4.17}$$

vornimmt, erhält man nach kurzer Rechnung durch

$$x^{(k)} = S_\omega x^{(k-1)} + B^{-1}b, \quad k = 1,2,\ldots,\ \omega \in]0,2[\tag{4.18}$$

das Gauß-Seidel-Verfahren mit Relaxation, wobei für S_ω und B

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

bzw.

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

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

4.5 Krylov-Raum-Verfahren zur Lösung linearer Gleichungssysteme

Ziel ist weiterhin die iterative Lösung des linearen Gleichungssystems

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

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

$$\{0\} \subset D_1 \subset \ldots \subset \mathbb{R}^n \tag{4.19}$$

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

Definition 4.14.

(a) Für gegebene Ansatzräume (4.19) *hat der* ***Ansatz des orthogonalen Residuums*** *zur Bestimmung von Vektoren* $x_1, x_2, \ldots \in \mathbb{R}^n$ *die Form*

$$\left.\begin{array}{c} x_k \in D_k \\ Ax_k - b \in D_k^\perp \end{array}\right\} k = 1,2,\ldots \tag{4.20}$$

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

$$\left.\begin{array}{c} x_k \in D_k \\ \|Ax_k - b\|_2 \text{ minimal} \end{array}\right\} k = 1, 2, \ldots \tag{4.21}$$

Bei der Wahl spezieller Ansatzräume (4.19) werden die sogenannten Krylovräume von Bedeutung sein

Definition 4.15. *Zu gegebener Matrix $A \in \mathbb{R}^{n\times n}$ und einem Vektor $b \in \mathbb{R}^n$ ist die Folge der Krylovräume durch*

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

erklärt, wobei $K_0(A, b) = \{\mathbf{0}\}$ verabredet wird. Es gilt offensichtlich $K_{k-1}(A, b) \subset K_k(A, b)$, $k = 1, 2, \ldots$.

Bemerkung. Im Folgenden werden die in Definition 4.14 angegebenen Ansätze mit den speziellen Räumen $D_k = K_k(A, b)$ betrachtet, wobei wir den Schwerpunkt auf den Ansatz (4.20) legen.

4.5.1 Der Ansatz des orthogonalen Residuums (4.20) für symmetrische positiv definite Matrizen

Für positiv definite, symmetrische Matrizen soll nun Existenz und Eindeutigkeit von Vektoren x_k für (4.20) 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 4.16. *Zu gegebener symmetrischer positiv definiter Matrix $A \in \mathbb{R}^{n\times n}$ sind für $k = 1, 2, \ldots$ die Vektoren x_k aus dem Ansatz des orthogonalen Residuums (4.20) – mit allgemeinen Ansatzräumen D_k gemäß (4.19) – eindeutig bestimmt, und es gilt*

$$\|x_k - x_*\|_A = \min_{x\in D_k} \|x - x_*\|_A, \quad k = 1, 2, \ldots \tag{4.22}$$

Beweis. Eindeutigkeit: Sei k fest gewählt. Für $x_k, \hat{x}_k$ mit der Eigenschaft (4.20) gilt

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

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

$$x_k = \sum_{j=0}^{m-1} \alpha_j d_j \tag{4.23}$$

an und erhält

$$\begin{aligned} x_k \text{ genügt (4.20)} &\Leftrightarrow Ax_k - b \in D_k^{\perp} \\ &\Leftrightarrow \langle Ax_k - b, d_k \rangle_2 = 0 \quad k = 0, \ldots, m-1 \qquad (4.24) \\ &\Leftrightarrow \sum_{j=0}^{m-1} \langle Ad_j, d_k \rangle_2 \, \alpha_j = \langle b, d_k \rangle_2 , \quad k = 0, \ldots, m-1 \qquad (4.25) \end{aligned}$$

(4.25) ist ein lineares Gleichungssystem von m Gleichungen für die Koeffizienten $\alpha_0, \ldots, \alpha_{m-1}$. Da x_k mit (4.20) eindeutig bestimmt ist (wurde schon gezeigt), ist das Gleichungssystem (4.25) eindeutig lösbar, woraus die Existenz von x_k in der Form (4.23) folgt.
Minimalität (4.22) Für $x \in D_k$ findet man

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

□

4.5.2 Der Ansatz des orthogonalen Residuums (4.20) für gegebene A-konjugierte Basen

10. Vorlesung 19.11.14

Mit dem Beweis von Satz 4.16 ist bereits eine Möglichkeit zur Bestimmung von x_k für (4.20) ausgehend von einer Basis $d_0, \ldots, d_{m-1}$ für D_k mit dem Gleichungssystem (4.25) aufgezeigt worden. Im Folgenden wird ein Spezialfall behandelt, bei dem (4.25) Diagonalgestalt hat.

Definition 4.17. *Es sei $A \in \mathbb{R}^{n \times n}$ symmetrisch und positiv definit. Gegebene Vektoren $d_0, \ldots, d_{m-1} \in \mathbb{R}^n \setminus \{0\}$ heißen A-konjugiert, falls*

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

gilt.

Bemerkung. Falls eine A-konjugierte Basis von D_k gegeben ist, hat (4.25) Diagonalgestalt und damit ist x_k gemäß Ansatz (4.23) sehr einfach berechenbar.

Satz 4.18. *Für eine gegebene symmetrische positiv definite Matrix $A \in \mathbb{R}^{n\times n}$ und A-konjugierte Vektoren $d_0, \dots$ gelte*

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

Dann erhält man für den Ansatz des orthogonalen Residuums (4.20) *die folgenden Darstellungen für $k = 1, 2, \dots$*

$$x_k = \sum_{j=0}^{k-1} \alpha_j d_j \quad mit \quad \alpha_j = -\frac{\langle r_j, d_j\rangle_2}{\langle Ad_j, d_j\rangle_2} \tag{4.26}$$

$$r_j := Ax_j - b, \quad j \geq 1, r_0 = -b \tag{4.27}$$

Beweis. Folgt unmittelbar für $k = m$ aus (4.23)-(4.25) □

Bemerkung.

(a) Aus (4.26) folgt die Unabhängigkeit der α_j von k und damit gilt

$$x_{k+1} = x_k + \alpha_k d_k, \quad r_{k+1} = r_k + \alpha_k A d_k, \quad k = 0, 1, \dots ; x_0 = 0 \tag{4.28}$$

(b) Aufgrund der ersten Identität von (4.28) bezeichnet man d_k als Suchrichtung und α_k als Schrittweite

(c) Außerdem wird mit (4.28) klar, dass eine simultane Berechnung der Suchrichtungen und Lösungsapproximationen x_k in der Reihenfolge

$$d_0, x_1, d_1, x_2, \dots$$

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

4.5.3 Das CG-Verfahren für positiv definite, symmetrische Matrizen

Definition 4.19. *Zu gegebener symmetrisch positiv definiter Matrix $A \in \mathbb{R}^{n\times n}$ ist das **Verfahren des konjugierten Gradienten** gegeben durch den Ansatz* (4.20) *mit der speziellen Wahl*

$$D_k = K_k(A, b), \quad k = 0, 1, \dots \tag{4.29}$$

Dieses Verfahren bezeichnet man auch kurz als CG-Verfahren.

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

Der folgende Hilfssatz behandelt die Berechnung A-konjugierter Suchrichtungen in $K_k(A,b)$ für $k = 0, 1, \ldots$
Ausgehend von den Notationen des Satzes 4.18 wird für den fixierten Index k dabei so vorgegangen, dass – ausgehend von einer bereits konstruierten A-konjugierten Basis $d_0, \ldots, d_{k-1}$ für $K_k(A,b)$ – eine A-konjugierte Basis für $K_{k+1}(A,b)$ gewonnen wird durch eine Gram-Schmidt-Orthogonalisierung der Vektoren $d_0, \ldots, d_{k-1}, -r_k \in \mathbb{R}^n$ bezüglich des Skalarproduktes $\langle \cdot, \cdot \rangle_A$.
Wie sich im Beweis von Lemma 4.20 herausstellt, genügt hier eine Gram-Schmidt-Orthogonalisierung der beiden Vektoren $d_{k-1}, -r_k \in \mathbb{R}^n$.

Lemma 4.20. *Zu gegebener symmetrisch positiv definiter Matrix $A \in \mathbb{R}^{n\times n}$ und mit den Notationen des Satzes 4.18 seien die Suchrichtungen speziell wie folgt gewählt:*

$$d_0 = b, \quad d_k = -r_k + \beta_{k-1} d_{k-1}, \quad \beta_{k-1} = \frac{\langle Ar_k, d_{k-1}\rangle_2}{\langle Ad_{k-1}, d_{k-1}\rangle_2}, \quad k = 1, \ldots, k_* - 1 \tag{4.30}$$

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

$$\operatorname{span}\{d_0, \ldots, d_{k-1}\} = \operatorname{span}\{b, r_1, \ldots, r_{k-1}\} = K_k(A,b), \quad k = 1, \ldots, k_* \tag{4.31}$$

Beweis. Vollständige Induktion über $k = 1, \ldots, k_*$ zum Nachweis der A-Konjugiertheit der Vektoren $d_0, \ldots, d_{k-1} \in \mathbb{R}^n$ und der Formeln (4.30) wegen

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

ist der Induktionsanfang gemacht.
Im Folgenden sei angenommen, dass (4.30) ein System von A-konjugierten Vektoren mit der Eigenschaft (4.31) liefert mit einem fixierten Index $1 \le k \le k_* - 1$
Gemäß dem Ansatz des orthogonalen Residuums (4.20) gilt $r_k \in K_k(A,b)^\perp$ und im Fall $r_k \neq 0$ sind damit die Vektoren $d_0, \ldots, d_{k-1}, -r_k$ linear unabhängig. Eine Gram- Schmidt-Orthogonalisierung dieser Vektoren bzgl. des Skalarproduktes $\langle \cdot, \cdot \rangle$ liefert den Vektor

$$d_k = -r_k + \sum_{j=0}^{k-1} \frac{\langle Ar_k, d_j\rangle_2}{\langle Ad_j, d_j\rangle_2} d_j \overset{(*)}{=} -r_k + \beta_{k-1} d_{k-1} \tag{4.32}$$

wobei $(*)$ aus den Eigenschaften

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

folgt, also

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

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

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

Aufgrund der 2. Formel in (4.28) gilt wegen $Ad_{k-1} \in K_{k+1}(A,b)$ noch

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

sodass aus Dimensionsgründen auch hier notwendigerweise Gleichheit vorliegt. □

Bemerkung. Mit dem durch Lemma 4.20 beschriebenen Abbruch wird gleichzeitig die Lösung von $Ax = b$ geliefert, es gilt also $x_{k_*} = x_*$. Dabei gilt notwendigerweise

$$k_* \leq n$$

denn aufgrund der linearen Unabhängigkeit der beiden Vektorsysteme in (4.31) erhält man

$$\dim K_k = k$$

für $k = 0, 1, \ldots, k_*$

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

Lemma 4.21. *In der Situation des Lemma 4.20 gelten die Darstellungen*

$$\alpha_k = \frac{\|r_k\|_2^2}{\langle Ad_k, d_k \rangle_2}, \quad k = 0, 1, \ldots, k_* - 1 \tag{4.33}$$

$$\beta_{k-1} = \frac{\|r_k\|_2^2}{\|r_{k-1}\|_2^2}, \quad k = 1, \ldots, k_* - 1 \quad (r_0 := b) \tag{4.34}$$

Beweis. Mit $r_k \in K_k(A,b)^\perp$ sowie der Beziehung (4.32) für die Suchrichtung d_k erhält man $-\langle r_k, d_k \rangle_2 = \|r_k\|_2^2$ und zusammen mit (4.26) ergibt dies (4.33). Diese Darstellung (4.33) für α_k zusammen mit der Identität $r_k = r_{k-1} + \alpha_{k-1} Ad_{k-1}$ aus (4.28) liefert

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

und damit gilt für β_{k-1} die Beziehung (4.34) □

4.5.4 Konvergenzgeschwindigkeit des CG-Verfahrens

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

Lemma 4.22. *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 λ_j und zugehörigen Eigenvektoren $v_j \in \mathbb{R}^n$, also gilt*

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

Mit der Entwicklung

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

gelten für jedes Polynom p die folgenden Darstellungen

$$p(A)x = \sum_{j=1}^{n} c_j p(\lambda_j) v_j \tag{4.35}$$

$$\|p(A)x\|_2 = \left(\sum_{j=1}^{n} c_j^2 p(\lambda_j)^2\right)^{\frac{1}{2}}, \quad \|p(A)x\|_A = \left(\sum_{j=1}^{n} c_j^2 \lambda_j p(\lambda_j)^2\right)^{\frac{1}{2}} \tag{4.36}$$

Speziell gilt also

$$m^{\frac{1}{2}} \|x\|_2 \le \|x\|_A \le M^{\frac{1}{2}} \|x\|_2, \quad x \in \mathbb{R}^n \tag{4.37}$$

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

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

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

und daraus folgt (4.35). Weiter berechnet man

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

Und analog erhält man

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

□

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

Satz 4.23. *Zu einer gegebenen symmetrisch positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$ gelten für das CG-Verfahren die folgenden Fehlerabschätzung:*

$$\|x_k - x_*\|_A \leq \left(\inf_{p\in\Pi_k, p(0)=1} \sup_{\lambda\in\sigma(A)} |p(\lambda)| \right) \|x_*\|_A \tag{4.38}$$

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

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

Mit Lemma 4.22 und der Entwicklung $x_* = \sum_{j=1}^{n} c_j v_j \in \mathbb{R}^n$ in der Eigenvektorbasis $\{v_j,\, j = 1, \ldots, n\}$ erhält man

$$\begin{aligned} \|x_k - x_*\|_A &\overset{(4.22)}{\leq} \underbrace{\|x - x_*\|_A}_{=\|p(A)x_*\|_A} = \left(\sum_{j=1}^{n} c_j^2 \lambda_j p(\lambda_j)^2 \right)^{\frac{1}{2}} \\ &\leq \sup_{\lambda\in\sigma(A)} |p(\lambda)| \left(\sum_{j=1}^{n} c_j^2 \lambda_j \right)^{\frac{1}{2}} = \sup_{\lambda\in\sigma(A)} |p(\lambda)|\, \|x_*\|_A \end{aligned}$$

□

Zur quantitativen Präzisierung der Abschätzung (4.38) des Satzes 4.23 benutzen wir die hier nicht bewiesenen Eigenschaften der Tschebyscheff-Polynome erster Art $T_0, T_1, \ldots$

$$T_k\left(\frac{\kappa+1}{\kappa-1}\right) \geq \frac{1}{2}\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k \quad \text{für} \quad k \in \mathbb{N}, \kappa > 1\,. \tag{4.39}$$

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

Satz 4.24. *Zu einer gegebenen symmetrisch positiv definiten Matrix $A \in \mathbb{R}^{n\times n}$ gelten für das CG-Verfahren die Fehlerabschätzungen*

$$\begin{aligned} \|x_k - x_*\|_A &\leq 2\gamma^k \|x_*\|_A, \quad k = 0, 1, \dots \\ \|x_k - x_*\|_2 &\leq 2\sqrt{\kappa_A}\gamma^k \|x_*\|_2, \quad k = 0, 1, \dots \end{aligned} \tag{4.40}$$

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

Beweis. Satz 4.23 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

$$\max_{m\leq\lambda\leq M} |p(\lambda)| = \left|T_k\left(\frac{M+m}{M-m}\right)\right|^{-1} = \left|T_k\left(\frac{\kappa_A+1}{\kappa_A-1}\right)\right|^{-1} \overset{(4.39)}{\leq} 2\gamma^k$$

(weil $\max_{t\in[-1,1]} T_k(t) = 1$) folgt aus (4.38) die erste Abschätzung, also (4.40). Die zweite Abschätzung des Satzes ist eine unmittelbare Konsequenz aus der Ersten unter der Nutzung der Normäquivalenz (4.37). □

Beispiel 4.25. Betrachten wir das Gleichungssystem $Ax = b$ mit

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \quad \text{und} \quad b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Entsprechend der Formeln (4.26), (4.28), (4.30) ergibt sich für die Startlösung $x_0 = b = (1\ 1\ 1)^T$ und die erste Suchrichtung $d_0 = b$

$$r_0 = Ax_0 - b = \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}$$

sowie $\alpha_0 = -\frac{\langle r_0, d_0\rangle_2}{\langle Ad_0, d_0\rangle_2} = \frac{1}{2}$ und damit

$$x_1 = x_0 + \alpha_0 d_0 = \begin{pmatrix} \frac{3}{2} \\ \frac{3}{2} \\ \frac{3}{2} \end{pmatrix} \quad \text{und} \quad r_1 = r_0 + \alpha_0 A d_0 = \begin{pmatrix} \frac{1}{2} \\ -1 \\ \frac{1}{2} \end{pmatrix}$$

sowie $\beta_0 = \frac{\langle Ar_1, d_0\rangle_2}{\langle Ad_0, d_0\rangle_2} = \frac{1}{2}$. Mit der neuen Suchrichtung

$$d_1 = -r_1 + \beta_0 d_0 = \begin{pmatrix} 0 \\ \frac{3}{2} \\ 0 \end{pmatrix}$$

und $\alpha_1 = -\frac{\langle r_1, d_1\rangle_2}{\langle Ad_1, d_1\rangle_2} = \frac{1}{3}$ erhält man

$$x_2 = x_1 + \alpha_0 d_1 = \begin{pmatrix} \frac{3}{2} \\ 2 \\ \frac{3}{2} \end{pmatrix} \text{ und stellt mit } r_2 = r_1 + \alpha_1 A d_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

fest, dass x_2 die Lösung ist.

4.5.5 CGNR-Verfahren

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

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

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

4.5.6 GMRES-Verfahren

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

Definition 4.26. *Das GMRES-Verfahren ist definiert durch den Ansatz des minimalen Residuums* (4.21) *mit der speziellen Wahl $D_k = K_k(A, b)$, es gilt also*

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

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

4.6 Die iterative Lösung nichtlinearer Gleichungssysteme

11. Vorlesung 24.11.14

Die Methode unserer Wahl zur Lösung einer im Allg. nichtlinearen Gleichung $f(x) = 0$ mit einer Abbildung $f : A \to \mathbb{R}$, $A \subset \mathbb{R}$ ist die Lösung der äquivalenten Aufgabe der Bestimmung eines Fixpunkts von

$$g(x) = x - f(x) \, .$$

Der Fixpunkt soll mit der Fixpunktiteration

$$x^{(k)} = g(x^{(k-1)}) \, , \; k = 1, 2, \ldots \tag{4.41}$$

ausgehend von einer Startnäherung $x^{(0)}$ bestimmt werden. Mit

$$\epsilon^{(k)} := x^{(k)} - \hat{x}$$

bezeichnen wir den Fehler nach k Iterationen.
Unter der Voraussetzung der Existenz eines Fixpunktes $\hat{x}$ von g (bzw. einer Nullstelle $\hat{x}$ von f) und der Glattheit von g findet man um $\hat{x}$ ein abgeschlossenes Intervall I mit der Kontraktion $g(I) \subset I$ mit der einer Konstanten $0 < L < 1$ (Beweis: siehe Vorlesung) und nach dem Banachschen Fixpunktsatz konvergiert die Fixpunktiteration (4.41) für eine beliebige Startnäherung gegen $\hat{x}$.
Mit dem folgenden Satz soll die Konvergenzordnung bestimmt werden.

Satz 4.27. *$g : I \to I = [a, b]$ sei auf I Lipschitz-stetig, kontrahierend und einmal stetig differenzierbar. Außerdem sei $g'(x) \neq 0$ auf ganz I. Dann gilt für den Fehler $\epsilon^{(k)}$*

$$\lim_{k \to \infty} \frac{\epsilon^{(k+1)}}{\epsilon^{(k)}} = g'(\hat{x}) \, , \tag{4.42}$$

d.h. die Konvergenz ist mindestens erster Ordnung bzw. linear.

Beweis. Zuerst wird gezeigt, dass für $x^{(0)} \neq \hat{x}$ für alle Iterierten $x^{(k)} \neq \hat{x}$ gilt, d.h., dass die Iteration nicht nach endlich vielen Schritten mit $\epsilon^{(k)} = 0$ abbricht. Wir nehmen das Gegenteil an, d.h. die Existenz eines Index k mit

$$x^{(k-1)} \neq \hat{x} \quad \text{und} \quad x^{(k)} = \hat{x} \, .$$

Wegen $g(x^{(k)}) = x^{(k)} = g(x^{(k-1)})$ folgt aus dem MWS der Differentialrechnung

$$0 = g(x^{(k-1)}) - g(x^{(k)}) = (x^{(k-1)} - x^{(k)}) g'(x^*) \, ,$$

und wegen $(x^{(k-1)} - x^{(k)}) \neq 0$ muss $g'(x^*) = 0$ sein für ein $x^* \in I$, was unserer Voraussetzung widerspricht.
Aus dem MWS folgt weiter

$$\epsilon^{(k+1)} = x^{(k+1)} - \hat{x} = g(x^{(k)}) - g(\hat{x}) = g(\hat{x} + \epsilon^{(k)}) - g(\hat{x}) = \epsilon^{(k)} g'(\hat{x} + \theta_k \epsilon^{(k)})$$

mit $0 < \theta_k < 1$. Da $\epsilon^{(k)} \neq 0$ ist, erhält man

$$\frac{\epsilon^{(k+1)}}{\epsilon^{(k)}} = g'(\hat{x} + \theta_k \epsilon^{(k)})$$

und wegen $\lim_{k\to\infty} \epsilon^{(k)} = 0$ folgt die Behauptung (4.42). □

Lemma 4.28. *Es sei $G \subset \mathbb{R}$ offen und $f : G \to \mathbb{R}$ stetig diff'bar mit einem Fixpunkt $\hat{x} \in G$. Wenn $|f'(\hat{x})| < 1$ gilt, dann existiert ein abgeschlossenes Intervall $D \subset G$ mit $\hat{x} \in D$ und $f(D) \subset D$, auf dem f eine Kontraktion ist.*

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

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

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

Bemerkung 4.29. Ist die Voraussetzung $|f'(\hat{x})| < 1$ des Hilfssatzes 4.28 nicht erfüllt, findet man keine Kontraktion. Ist $|f'(\hat{x})| > 1$ dann gilt in der Nähe von $\hat{x}$

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

das rechtfertigt die

Definition 4.30. *Ein Fixpunkt $\hat{x}$ heißt anziehender Fixpunkt, wenn $|f'(\hat{x})| < 1$ gilt, und $\hat{x}$ heißt abstoßender Fixpunkt, wenn $|f'(\hat{x})| > 1$ ist.*

4.6.1 Newton-Verfahren

Nun betrachten wir zur Bestimmung einer Nullstelle von $f(x)$ die Newton-Iteration

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}\ ,\ k = 0, 1, \ldots, \tag{4.43}$$

die bei genauerem Hinsehen eine Fixpunktiteration der Funktion

$$g(x) := x - \frac{f(x)}{f'(x)} \quad \text{mit } \ g(\hat{x}) = \hat{x}\ , \tag{4.44}$$

also $x^{(k+1)} = g(x^{(k)})$ ist.
Als wichtiges Hilfsmittel zum Konvergenznachwies und zur Konvergenzordnungsbestimmung des Newtonverfahrens betrachten wir das

Lemma 4.31. *$g : I \to I = [a, b]$ sei auf I Lipschitz-stetig, kontrahierend und zweimal stetig differenzierbar. Außerdem gelte $g'(\hat{x}) = 0$ und $g''(x) \neq 0$ auf ganz I. Dann gilt für den Fehler $\epsilon^{(k)}$*

$$\lim_{k\to\infty} \frac{\epsilon^{(k+1)}}{{\epsilon^{(k)}}^2} = \frac{1}{2} g''(\hat{x}) \; , \tag{4.45}$$

d.h. die Konvergenz ist mindestens zweiter Ordnung bzw. quadratisch.

Beweis. Der Beweis erfolgt weitgehend analog zum Beweis von Satz 4.27. Mit einer Taylorentwicklung gilt

$$\epsilon^{(k+1)} = x^{(k+1)} - \hat{x} = g(x^{(k)}) - g(\hat{x}) = g(\hat{x} + \epsilon^{(k)}) - g(\hat{x})$$
$$= g(\hat{x}) + \epsilon^{(k)} g'(\hat{x}) + \frac{1}{2}{\epsilon^{(k)}}^2 g''(\hat{x} + \theta_k \epsilon^{(k)}) - g(\hat{x}) = \frac{1}{2}{\epsilon^{(k)}}^2 g''(\hat{x} + \theta_k \epsilon^{(k)}) \; .$$

Analog zum Beweis des Satzes 4.27 ergibt der Limes $k \to \infty$ die Behauptung. □

Bemerkung 4.32.
Die auf den ersten Blick ungewöhnliche Voraussetzung $g'(\hat{x}) = 0$ im Lemma 4.31 für einen Fixpunkt $\hat{x}$ von g wird gerade von der Funktion

$$g(x) := x - \frac{f(x)}{f'(x)} \; ,$$

die im Newtonverfahren eine zentrale Rolle spielt, erfüllt!

Der folgende Satz liefert eine grundlegende Aussage zur Konvergenzordnung der Folge (4.43), also des Newton-Verfahrens.

Satz 4.33. *Die Funktion $f(x)$ sei dreimal stetig differenzierbar in einem Intervall $I_1 = [a, b]$ mit $a < \hat{s} < b$, und es sei $f'(\hat{x}) \neq 0$, d.h. $\hat{x}$ sei eine einfachen Nullstelle von f. Dann existiert ein Intervall $I = [\hat{x} - \delta, \hat{x} + \delta]$ mit $\delta > 0$, für welches $g : I \to I$ eine kontrahierende Abbildung ist. Für jeden Startwert $x^{(0)} \in I$ ist die Folge (4.43) mindestens quadratisch konvergent.*

Beweis. Der Beweis folgt durch Anwendung von Lemma 4.31 (s. auch Vorlesung!). □

Mit dem folgenden Satz kann man die Glattheitsforderungen an f noch abschwächen.

Satz 4.34. *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 konvergiert quadratisch gegen $\hat{x}$, d.h. es gilt

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

mit einer Konstanten C. Außerdem gilt die Fehlerabschätzung

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

Beweis. Folgt aus dem Banachschen Fixpunktsatz und dem Satz 4.28 über die Existenz einer Kontraktion.
Die quadratische Konvergenz folgt aus

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

□

Bemerkung 4.35. Die Voraussetzungen des Satzes 4.34 garantieren die Kontraktivität der Hilfsfunktion $g(x) = x - \frac{f(x)}{f'(x)}$ in einer Umgebung von $\hat{x}$. D.h. man ist mit dem Newton-Verfahren immer dann erfolgreich, wenn man nur nah genug an der Nullstelle die Iteration beginnt (x_0 nah bei $\hat{x}$)
In diesem Fall ist das Newton-Verfahren auch noch sehr schnell aufgrund der quadratischen Konvergenz.

Satz 4.36. *(Nullstelle einer konvexen Funktion)*
Sei $f : [a, b] \to \mathbb{R}$ zweimal stetig differenzierbar und konvex ($f'(x) \neq 0$ auf $[a, b]$). Die Vorzeichen von $f(a)$ und $f(b)$ seien verschieden. Dann konvergiert die Newton-Folge von f für $x_0 = a$, falls $f(a) > 0$ und für $x_0 = b$, falls $f(b) > 0$, gegen die einzige Nullstelle $\hat{x}$ von f.

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

Die Vorgehensweise bei der Lösung eines nichtlinearen Gleichungssystems $F(x) = \mathbf{0}$ mit einem Newton-Verfahren ist eine Verallgemeinerung des eindimensionalen Problems der Bestimmung einer Nullstelle einer reellen Funktion einer Veränderlichen.
Die Newton-Folge hat im mehrdimensionalen die Form

$$x^{(k+1)} = x^{(k)} - [F'(x^{(k)})]^{-1} F(x^{(k)}), \quad k = 0, 1, 2, \dots \ . \tag{4.46}$$

Bevor wir die lokale Konvergenzaussage beweisen, sollen einige Hilfsmittel bereit gestellt werden.

Lemma 4.37. *Sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar in der offenen Menge $D \subset \mathbb{R}^n$ und Lipschitz-stetiger Ableitung F' in D (mit der Konstante $L > 0$). Sei $z \in D$ fest gewählt und $F'(z)$ regulär. Weiterhin existiere ein $\beta > 0$ so dass $\|F'(z)^{-1}\| \leq \beta$.*
Dann gilt für alle $x \in B(z, \eta) := \{y \in \mathbb{R}^n : \|y - z\| < \eta\}$ mit $0 < \eta < \frac{c}{L\beta}$ mit einem festen $c \in]0, 1[$, dass $F'(x)$ regulär ist, und, dass

$$\left\|F'(x)^{-1}\right\| \leq \frac{\beta}{1 - c}$$

gilt.

Die Beweisskizze dieses Hilfssatzes wurde in der Vorlesung dargelegt.

Lemma 4.38. *Sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar in der offenen konvexen Menge $D \subset \mathbb{R}^n$, $x \in D$ und sei außerdem F' Lipschitz-stetig in einer Umgebung von x aus D (mit der Konstante $L > 0$).*
Dann gilt für alle $x + h \in D$,

$$\|F(x + h) - F(x) - F'(x)\, h\| \leq \frac{L}{2} \|h\|^2 \ .$$

Beweis. Mit der Nutzung des Mittelwertsatzes in Integral-Form ergibt sich

$$\begin{aligned} F(x+h) - F(x) - F'(x) &= \int_0^1 F'(x+th)\, h\, dt \;+\; F'(x)\, h \\ &= \int_0^1 (F'(x+th) - F'(x))\, h\, dt\ . \end{aligned}$$

Die Nutzung der Lipschitz-Stetigkeit von F' ergibt

$$\begin{aligned} \|F(x+h) - F(x) - F'(x)\| &\leq \int_0^1 \|F'(x+th) - F'(x)\|\ dt\ \|h\| \\ &\leq L\,\|h\|^2 \int_0^1 t\, dt \\ &= \frac{L}{2}\,\|h\|^2\ . \end{aligned}$$

□

Die Ergebnisse der Hilfssätze 4.37 und 4.38 erlauben den Beweis des folgenden Satzes.

Satz 4.39. *Sei $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar in der offenen konvexen Menge $D \subset \mathbb{R}^n$. Es existiere ein $\hat{x} \in \mathbb{R}^n$ und $r, \beta > 0$ so dass die offene Kugel $B(\hat{x}, r) \subset D$, $F(\hat{x}) = \mathbf{0}$, $F'(\hat{x})^{-1}$ existiert mit $\|F'(\hat{x})^{-1}\| \leq \beta$, und F' Lipschitz-steig mit der Konstanten L auf $B(\hat{x}, r)$.*
Dann existiert ein $\epsilon > 0$ so dass für alle $x^{(0)} \in B(\hat{x}, r)$ die Newton-Folge $\{x^{(k)}\}$ (4.46) wohl-definiert ist, und gegen $\hat{x}$ konvergiert, und quadratische Konvergenz in Form von

$$\left\|x^{(k+1)} - \hat{x}\right\| \leq \frac{L\beta}{2(1-c}\left\|x^{(k)} - \hat{x}\right\|^2\ ,\ k = 0, 1, \ldots \tag{4.47}$$

für eine festes $c \in]0, \frac{2}{3}[$ gilt.

Beweis. Der Beweis wird induktiv geführt. Für $0 < c \leq \frac{2}{3}$ sei

$$\epsilon = \min\{r, \frac{c}{L\beta}\}\ .$$

Dei Anwendung von Hilfssatz 4.37 mit $z := \hat{x}$ und $x := x^{(0)}$ ergibt die Regularität von $F'(x^{(0)})$ und

$$\left\|F'(x^{(0)})^{-1}\right\| \leq \frac{\beta}{1-c}\ .$$

Damit ist $x^{(1)}$ wohl-definiert und es gilt

$$\begin{aligned} x^{(1)} - \hat{x} &= x^{(0)} - \hat{x} - F'(x^{(0)})^{-1} F(x^{(0)}) \\ &= x^{(0)} - \hat{x} - F'(x^{(0)})^{-1}(F(x^{(0)}) - F(\hat{x})) \\ &= F'(x^{(0)})^{-1}[F(\hat{x}) - F(x^{(0)}) - F'(x^{(0)})(\hat{x} - x^{(0)})] \ . \end{aligned}$$

Die Anwendung von Hilfssatz 4.38 ergibt

$$\begin{aligned} \left\|x^{(1)} - \hat{x}\right\| &\leq \left\|F'(x^{(0)})^{-1}\right\| \left\|F(\hat{x}) - F(x^{(0)}) - F'(x^{(0)})(\hat{x} - x^{(0)})\right\| \\ &\leq \frac{L\beta}{2(1-c)} \left\|x^{(0)} - \hat{x}\right\|^2 \ , \end{aligned}$$

womit der Beweis von (4.47) für $k = 0$ erbracht ist. Da $\left\|x^{(0)} - \hat{x}\right\| \leq \frac{c}{L\beta}$ ist, ergibt sich

$$\left\|x^{(1)} - \hat{x}\right\| \leq \frac{c}{2(1-c)} \left\|x^{(0)} - \hat{x}\right\| \leq \left\|x^{(0)} - \hat{x}\right\| < \epsilon \ .$$

Der Induktionsschritt wird völlig analog gezeigt. □

12. Vorlesung 26.11.14

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

$$x^{(k+1)} = x^{(k)} - \alpha_k [F'(x^{(k)})]^{-1} F(x^{(k)}), \quad k = 0, 1, \ldots \tag{4.48}$$

mit $\alpha_k \in]0, 1[$, und spricht hier von einem **gedämpften Newton-Verfahren**. Man kann es auch in der Form

$$F'(x^{(k)}) z_k = -F(x^{(k)}) \tag{4.49}$$
$$x^{(k+1)} = x^{(k)} + \alpha_k z^{(k)} \tag{4.50}$$

Mit gedämpften Newton-Verfahren erreicht man mitunter Konvergenz der Newton-Folge, wenn das Standard-Newton-Verfahren ($\alpha = 1$) versagt.

Bemerkung 4.40. Beim gedämpften Newtonverfahren hat man allerdings das Problem, dass man im Gegensatz zum klassischen Newtonverfahren keine quadratische Konvergenz erhält. Denn statt von

$$g(x) = x - f'(x)^{-1} f(x)$$

sucht man nach Fixpunkten von

$$\tilde{g}(x) = x - \alpha f'(x)^{-1} f(x)$$

und für $\alpha \neq 1$ gilt für einen Fixpunkt $\hat{x}$ nicht $\tilde{g}'(\hat{x}) = 0$, was ja eine entscheidende Bedingung (s. Bem. 4.32) für den Nachweis der quadratischen Konvergenz war.
Im Prinzip trifft diese Argumentation auch auf den mehrdimensionalen Fall zu, d.h. im Falle eine Dämpfung verliert man eine Konvergenzordnung.

Bemerkung 4.41. Alles was wir bisher zum Newton-Verfahren ausgesagt haben ist lokal (mit Ausnahme der Aussagen für Nullstellen konvexer Funktionen), d.h. die Verfahren sind dann sehr effizient, wenn man Startnäherungen $x^{(0)}$ findet, die nahe genug bei einer existierenden Nullstelle/Lösung $\hat{x}$ liegen. Gute Startiterationen lassen sich mitunter aus dem Kontext der jeweiligen Aufgabenstellung herleiten. Ansonsten muss man **probieren**.

Das oben erwähnte gedämpfte Newton-Verfahren eröffnet auch eine Möglichkeit für globale Konvergenz. Das soll kurz diskutiert werden. Entscheidend ist die jeweilige effiziente Wahl von α_k. Eine gute Wahl von α_k erhält man wie folgt:
Bei gegebenem z_k nach (4.49) wählt man α_k als das erste Element einer Folge $\{\omega^l\}_0^\infty$, $\omega \in]0,1[$ fest gewählt (die Folge ist streng monoton fallend), für das

$$\left\|F(x^{(k)} + \omega^l z^{(k)})\right\| \leq (1 - \nu\omega^l) \left\|F(x^{(k)})\right\| \tag{4.51}$$

gilt, wobei ν ein fest gewählter Parameter aus dem Intervall $]0,1[$ ist. Die Bedingung (4.51) heißt hinreichende Abstiegsbedingung oder auch Armijo-Regel. Eine solche α_k-Wahl impliziert eine streng monoton fallende Folge $\{F(x^{(k)})\}$ und damit geht $F(x^{(k)})$ gegen $\mathbf{0}$. Die Existenz eines Dämpfungsparameters α_k, der die gegenüber (4.51) abgeschwächte Bedingung

$$\left\|F(x^{(k)} + \alpha_k z^{(k)})\right\| < \left\|F(x^{(k)})\right\|$$

sichert, zeigt der folgende

Satz 4.42. *Sei $F : D \to R^n$ eine auf der offenen und konvexen Menge $D \subset \mathbb{R}^n$ stetig differenzierbare Abbildung. Für $x^{(k)} \in D$ mit $F(x^{(k)} \neq \mathbf{0}$ gibt es ein $\mu_k > 0$, so dass*

$$\left\|F(x^{(k)} + \alpha z^{(k)})\right\|^2 < \left\|F(x^{(k)}\right\|^2 \quad \textit{für alle } \alpha \in]0, \mu_k[$$

(als Norm betrachten wir hier die Euklidische Norm).

Beweis. Wir betrachten die stetig differenzierbare Fubktion

$$\varphi(\alpha) = \left\|F(x^{(k)} + \alpha z^{(k)})\right\|^2$$

für $\alpha \in]-\epsilon, \epsilon[$, mit $\epsilon > 0$.
Dann gilt

$$\varphi(0) = \left\|F(x^{(k)})\right\|^2 > 0$$

und weiter

$$\varphi'(0) = 2\, F(x^{(k)})^T\, F'(x^{(k)}) z^{(k)} = -2F(x^{(k)})^T F(x^{(k)}) < 0\ .$$

Aus Stetigkeitsgründen gibt es ein Intervall $]0, \mu[$ wo $\varphi(\alpha)$ streng monoton fällt, d.h. wo

$$\varphi(0) = \left\|F(x^{(k)})\right\|^2 > \left\|F(x^{(k)} + \alpha z^{(k)})\right\|^2 = \varphi(\alpha) \quad \text{für alle } \alpha \in]0, \mu[\text{ gilt.}$$

□

Kapitel 5

Orthogonale Matrizen – QR-Zerlegung – Ausgleichsprobleme

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

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

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

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

und einer verallgemeinerten oberen Dreiecksmatrix

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

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

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

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

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

betrachtet, findet man mit der Darstellung

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

mit

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

Gradient und Hessematrix. Die Auswertung der notwendigen Extremalbedingung

$$\nabla F(r) = \mathbf{0} \Longleftrightarrow M^T M r = M^T b\,, \tag{5.4}$$

wobei man das Gleichungssystem (5.4) **Normalgleichungsystem** nennt, liefert mit

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

die Lösung des Minimumproblems für den Fall, dass die Matrix M den vollen Rang hat, denn dann ist die Hessematrix $M^T M$ von F symmetrisch und positiv definit. Bei dieser Methode ist allerdings anzumerken, dass die Kondition der Koeffizientenmatrix im Allg. schlechter ist, als die Kondition der Matrix des linearen Gleichungssystems, das bei der im Folgenden beschriebenen zweiten Lösungsmöglichkeit des Problems (5.3) entsteht.

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

(i)
$$\|Qx\|_2 = \|x\|_2 = \left\|Q^T x\right\|_2, x \in \mathbb{R}^n \tag{5.5}$$

(ii)
$$\mathrm{cond}(QA) = \mathrm{cond}(A) \tag{5.6}$$

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

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

$$A = QR \tag{5.7}$$

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

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

Mit den Spaltenvektoren $a_k, q_k \in \mathbb{R}^n, k = 1, \ldots, n$. (5.7) bedeutet dann

$$a_j = \sum_{i=1}^{j} r_{ij} q_i, \quad j = 1, ..., n, q_1, \ldots, q_n \in \mathbb{R}^n \tag{5.8}$$

5.1 Gram-Schmidt-Verfahren zur Orthogonalisierung

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

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

$$\hat{q}_j := a_j - \sum_{i=1}^{j-1} \langle a_j, q_i \rangle_2 \, q_i \tag{5.9}$$

Und nach der Normierung

$$q_j = \frac{\hat{q}_j}{\|\hat{q}_j\|_2}$$

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

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

Aus der Gleichung (5.9) folgt

$$a_j = \underbrace{\|\hat{q}_j\|_2}_{r_{jj}} q_j + \sum_{i=1}^{j-1} \underbrace{\langle a_j, q_i \rangle_2}_{r_{ij}} q_i = \sum_{i=1}^{j} r_{ij} q_i, \quad j = 1, \ldots, n \tag{5.10}$$

Nach Abschluss der Gram-Schmidt-Orthogonalisierung hat man damit mit (5.10)

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

als QR-Zerlegung

Bemerkung 5.1. Der beschriebene Algorithmus kann im ungünstigsten Fall Probleme bereiten (nicht gutartig sein), wenn z.B. $\|\hat{q}_j\|_2$ recht klein wird (Lösung folgt).

5.2 Householder-Matrizen/Transformationen

13. Vorlesung am 01.12.2014

Für den Fall einer nicht-quadratischen Matrix $A \in \mathbb{R}^{n\times m}$, $n > m$ erhält man mit dem Gram-Schmidt-Verfahren nur m orthogonale Vektoren $q_1, \ldots, q_m$, da A nur m Spalten besitzt. Damit hat man zwar die oberer Dreicksmatrix R bestimmt, Q allerdings nicht. Durch die sukzessive nichttriviale Lösung des linearen Gleichungssystems

$$\tilde{Q}_k \tilde{q}_{k+1} = \mathbf{0}, \quad q_{k+1} = \frac{1}{||\tilde{q}_{k+1}||_2} \tilde{q}_{k+1}$$

mit der Matrix $\tilde{Q}_k = [q_1^T q_2^T \ldots q_k^T]^T \in \mathbb{R}^{k\times n}$, $k = m, \ldots, n-1$, erhält man mit etwas Arbeit die angestrebte orthogonale Matrix $Q = [q_1\ q_2\ \ldots\ q_n]$ mit der Eigenschaft

$$A = Q \left[\frac{R}{\mathbf{0}} \right] .$$

Im folgenden werden wir allerdings eine Konstruktionsmethode für die QR-Zerlegung besprechen, die die Faktoren Q und R auf direktem Wege ergibt.

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

$$H = E - 2ww^T, w \in \mathbb{R}^n, \|w\|_2^2 = w^T w = 1 \tag{5.11}$$

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

Eigenschaften von H:

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

Nachweis als Übung

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

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

gilt.

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

$$w = \frac{x + \sigma e_1}{\|x + \sigma e_1\|_2} \quad mit \quad \sigma = \pm \|x\|_2 \tag{5.12}$$

gilt

$$\|w\|_2 = 1, \quad Hx = (E - 2ww^T)x = -\sigma e_1 \tag{5.13}$$

Beweis. $\|w\|_2 = 1$ weil $x + \sigma e_1 \neq 0$ und damit (5.12) wohldefiniert ist. Für den Nachweis von (5.13) erhält man

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

Und mit (5.12), d.h. $\frac{(x+\sigma e_1)^T}{\|x+\sigma e_1\|_2} = w^T$ folgt:

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

die nochmalige Nutzung von (5.12) 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 (5.12) $\sigma = \operatorname{sgn}(x_1)\|x\|_2$ gewählt, d.h. z.B. für $x = (-3, 1, 5)^T$ ist $\sigma = -\sqrt{35}$

5.3 Algorithmus zur Konstruktion der Faktorisierung mittels Householder-Transformationen

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

$$A^{(j)} = \begin{bmatrix} a_{11}^{(j)} & a_{12}^{(j)} & & \cdots & & a_{1m}^{(j)} \\ & \ddots & & & & \\ & & a_{j-1j-1}^{(j)} & & & a_{j-1m}^{(j)} \\ & & & a_{jj}^{(j)} & \cdots & a_{jm}^{(j)} \\ & & & \vdots & & \vdots \\ & & & a_{nj}^{(j)} & \cdots & a_{nm}^{(j)} \end{bmatrix}, \quad j = 1, \ldots, m \tag{5.14}$$

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

Die Matrizen der Form (5.14) erhält man für $j = 1, \ldots, m-1$ durch Transformationen der Form

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

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

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

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

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

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

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

$$M = QS\,, \qquad S = \left[\begin{array}{c} R \\ - \\ 0 \end{array}\right]$$

und orthogonalem Q gegeben, dann ergibt sich mit

$$Q^T b =: \left[\begin{array}{c} b_1 \\ b_2 \end{array}\right], \quad b_1 \in \mathbb{R}^m,\ b_2 \in \mathbb{R}^{n-m}$$

durch Nutzung der Eigenschaften von Q

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

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

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

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

5.4 Gauß-Newton-Verfahren

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

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

und gesucht ist ein funktionaler Zusammenhang

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

wobei solche Parameter

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

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

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

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

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

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

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

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

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

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

$$\min_{a\in\mathbb{R}^p} ||R(a^{(i)}) + R'(a^{(i)})(a - a^{(i)})||_2^2 \,, \tag{5.15}$$

bestimmt man eine neue Näherung

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

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

$$R(a^{(i)}) + R'(a^{(i)})(a - a^{(i)}) \tag{5.16}$$

in der Form

$$Ma - y$$

mit

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

auf, dann kann man die Lösung a^* von

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

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

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

erhalten (s. auch vorangegangene Vorlesungen).
Aus Effizienzgründen (jeweiliger Aufbau von y) schreibt man (5.16) auch in der Form

$$Ms - \hat{y}$$

mit

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

auf und löst das Minimum-Problem

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

und berechnet durch

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

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

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

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

$$R'(a^*)^T R(a^*) = \mathbf{0} \tag{5.17}$$

erfüllt. Wenn man den Gradienten von $F(a)$ ausrechnet stellt man fest, dass die Bedingung (5.17) wegen

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

äquivalent zur notwendigen Extremalbedingung

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

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

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

erfüllt ist.

14. Vorlesung am 03.12.2014

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

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

durch die direkte Auswertung der notwendigen Extremalbedingung

$$\text{grad}_a F = \mathbf{0} \tag{5.18}$$

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

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

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

Beispiel:
Gegegen ist eine Wertetabelle

k	1	2	3	4
x_k	1	2	3	4
y_k	2	4	7	3

und es gibt die Überlegung nach einer Funktion

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

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

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

minimal wird, also

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

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

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

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

Im Unterschied zu linearen Ausgleichsproblemen sind bei nichtlinearen Aufgabenstellungen zusätzliche Betrachtungen zur evtl. Mehrdeutigkeit des Problems $\text{grad}_a F = \mathbf{0}$ und zur Bewertung der gefundenen kritischen Stellen (lok. Maximum/lok. Minimum) durch die Überprüfung hinreichender Extremalbedingungen (Definitheit der Hesse-Matrix) erforderlich.
Zu dem obigen Beispiel ist allerdings auch anzumerken, dass bei den wenigen Werten (x_k, y_k) der Ansatz $y = \sin(ax) + b$ auch etwas gewagt ist. Dass kann und wird sicherlich auch ein Grund für das etwas wilde Extremalverhalten des Funktionals $F(a, b) = ||R(a, b)||_2^2$ sein.
"Fittet" man seine Messwerte mit der Funktion $y = f(x)$ besser, sollte die Bestimmung geeigneter Parameter $a = (a_1, \ldots, a_p)$ auch einfacher werden.

Kapitel 6

Interpolation

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

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

Mögliche Ansätze für f

(i) Polynome

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

z.B. für schwer handhabbare Funktionen oder Messwert-Interpolation durch die einfach handhabbaren Polynome,

(ii) Rationale Funktionen

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

zur Interpolation von Funktionen mit Polen bzw. Asymptoten,

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

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

zur Interpolation von periodischen Signalen,

(iv) Splines (stückweise Polynome)
zur Konstruktion von (nicht so glatten) Interpolationsfunktionen, um Oszillationen, die bei höhergradigen Interpolationspolynomen auftreten können' zu vermeiden.

Ziel/Aufgabe der Interpolation

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

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

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

6.1 Polynominterpolation

Definition 6.1. *Unter 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 6.2. *Zu $n+1$ gegebenen Stützstellen $(x_k, y_k), k = 0, \ldots, n$ mit der Eigenschaft $x_i \neq x_j, i \neq j$, gibt es genau ein Polynom $p \in \Pi_n$ mit $p(x_k) = y_k, k = 0, 1, \ldots, n$*

Beweis. Ansatz:

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

bedeutet

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

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

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

□

Definition 6.3. *Das nach Satz 6.2 eindeutig bestimmte Polynom p mit der Eigenschaft*

$$p(x_k) = y_k, \quad k = 0, 1, \ldots, n$$

für die vorgegebenen Stützstellen (x_k, y_k) heißt ***Interpolationspolynom****.*

6.1.1 Konstruktion des Interpolationspolynoms

Wir erinnern uns an die Generalvoraussetzung

$$x_i \neq x_j \quad \forall i, j = 0, 1, \ldots, n, i \neq j$$

Definition 6.4. *Die Polynome*

$$L_k(x) = \prod_{k \neq i = 0}^{n} \frac{x - x_i}{x_k - x_i} \tag{6.4}$$

heißen Lagrange-Basispolynome.

Definition 6.5. *Die Polynome*

$$N_k(x) = \prod_{i=0}^{k-1} (x - x_i), \quad k = 1, \ldots, n$$

mit $N_0(x) = 1$ heißen Newton-Basispolynome.

Satz 6.6. *Die Monombasis*

$$1, x, \ldots, x^n$$

sowie die Lagrange-Basispolynome

$$L_k(x), k = 0, \ldots, n$$

und die Newton-Basispolynome

$$N_k(x), k = 0, \ldots, n$$

sind Basen (linear unabhängige erzeugende Funktionensysteme) des Vektorraums der reellen Polynome Π_n vom Grad $\leq n$

Beweis. Als Übung empfohlen, bzw. s. unten. □

6.2 Lagrange-Interpolation

Zuerst ist anzumerken, dass man das Interpolationspolynom nicht in der Form (6.2) auf der Grundlage der Lösung des Gleichungssystems (6.3) mit der Vandermondeschen Matrix bestimmt, weil das viel zu aufwändig ist.
Besser geht es mit der **Lagrange-Interpolation**.
Für $n = 3$ haben wir zum Beispiel die Basispolynome

$$\begin{aligned} L_0(x) &= \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} \\ L_1(x) &= \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \\ L_0(x) &= \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} \\ L_0(x) &= \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \end{aligned}$$

und erkennen:

$$L_0(x_0) = 1, L_0(x_1) = L_0(x_2) = L_0(x_3) = 0$$

allgemein gilt:

$$L_k(x_j) = \delta_{kj}, \quad k = 0, \ldots, n \tag{6.5}$$

Damit ergibt sich für das Interpolationspolynom:

$$p(x) = \sum_{k=0}^{n} y_k L_k(x) \tag{6.6}$$

da

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

gilt. (6.6) heißt **Lagrangsches Interpolationspolynom**.

Bemerkung 6.7. Stellt man nun die Lagrange-Interpolationspolynome der Elemente der Monombasis

$$\mathcal{M} = \{1, x, \ldots, x^n\}$$

auf, dann ergibt sich

$$\sum_{k=0}^{n} x_k^j L_k(x) = x^j \ , \quad j = 0, \ldots, n \ ,$$

also kann man die Monombasis durch

$$\begin{pmatrix} 1 & 1 & \dots & 1 \\ x_0 & x_1 & \dots & x_n \\ \vdots & & & \\ x_0^n & x_1^n & \dots & x_n^n \end{pmatrix} \begin{pmatrix} L_0(x) \\ L_1(x) \\ \vdots \\ L_n(x) \end{pmatrix} =: T_{lm} \begin{pmatrix} L_0(x) \\ L_1(x) \\ \vdots \\ L_n(x) \end{pmatrix} = \begin{pmatrix} 1 \\ x^1 \\ \vdots \\ x^n \end{pmatrix}$$

darstellen. Man erkennt T_{lm} als Transponierte der Vandermondeschen Matrix V, die regulär ist, womit auch der Nachweis, dass es sich bei

$$\mathcal{L} = \{L_0(x), L_1(x), \dots, L_n(x)\}$$

tatsächlich um eine Basis von Π_n handelt, erbracht ist.
Für die Transformation der Koordinaten κ_l eines Polynoms bezüglich der Lagrange-Basis in die Koordinaten des Polynoms bezüglich der Monombasis ergibt sich

$$\kappa_m = T_{lm}^{-T} \kappa_l \,. \tag{6.7}$$

Beispiel 6.8. Betrachten wir die Tabelle

k	0	1	2
x_k	0	1	2
y_k	1	0	4

aus der man die Koordinaten $\kappa_l = (1\ 0\ 4)^T$ bezüglich der Lagrange-Basis abliest. Dann ergibt sich die Matrix

$$T_{lm} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 1 & 4 \end{pmatrix} \,,$$

die die Basistransformation beschreibt, und man errechnet die Koordinaten bezüglich der Monombasis unter Nutzung von

$$T_{lm}^{-T} = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{3}{2} & 2 & -\frac{1}{2} \\ \frac{1}{2} & -1 & \frac{1}{2} \end{pmatrix} \quad \text{zu} \quad \kappa_m = T_{lm}^{-T} \kappa_l = T_{lm}^{-T} \begin{pmatrix} 1 \\ 0 \\ 4 \end{pmatrix} = \begin{pmatrix} 1 \\ -\frac{7}{2} \\ \frac{5}{2} \end{pmatrix}$$

also ergibt sich das Polynom $p(x) = 1 - \frac{7}{2}x + \frac{5}{2}x^2$.

6.3 Fehler bei der Polynominterpolation

Geht es nicht nur um Messwerte, sondern soll eine Funktion f in den Stützstellen $(x_k, f(x_k))$, $k = 0, \ldots, n$, durch ein Polynom interpoliert werden, interessiert der Fehler. Dazu betrachten wir den folgenden

Satz 6.9. *f sei eine reelle, (n+1)-mal stetig diff'bare Funktion auf einem offenen Intervall, das das beschränkte Intervall $[a, b]$ enthält. Für das Interpolationspolynom $p_n(x)$ zu den (n+1) paarweise verschiedenen Stützstellen $x_0, x_1, \ldots, x_n$ mit $a = \min\{x_i\}$, $b = \max\{x_i\}$ und zu den Stützwerten $y_i = f(x_i)$ gilt für jedes $\tilde{x} \in [a, b]$*

$$f(\tilde{x}) - p_n(\tilde{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (\tilde{x} - x_i)\,, \tag{6.8}$$

wobei ξ eine (von $\tilde{x}$ abhängige) Stelle $\in]a, b[$ bedeutet.

Beweis. (s. auch Vorlesung)
Wir betrachten nur den relevanten Fall $\tilde{x} \neq x_i$, $i = 1, \ldots, n$, den anderenfalls ist die Aussage trivial. Zu dem fixierten Wert $\tilde{x}$ wird die Hilfsfunktion

$$F(x) := f(x) - p_n(x) - g(\tilde{x}) \prod_{i=0}^{n} (x - x_i)$$

betrachtet, wobei die Konstante $g(\tilde{x})$ so gewählt ist, dass $F(\tilde{x}) = 0$ gilt, also

$$g(\tilde{x}) = \frac{f(\tilde{x}) - p_n(\tilde{x})}{\prod_{i=0}^{n} (\tilde{x} - x_i)}\,.$$

Man überlegt nun, dass F die (n+2) Nullstellen $x_0, \ldots, x_n, \tilde{x}$ hat, und die sukzessive Anwendung des Satzes von Rolle ergibt die Existenz (mindestens) eines Wertes $\xi \in]a, b[$ mit $F^{(n+1)}(\xi) = 0$. Die Berechnung der (n+1)-ten Ableitung von F ergibt

$$F^{(n+1)}(x) = f^{(n+1)}(x) - g(\tilde{x})(n+1)!\,,$$

woraus die Aussage des Satzes folgt. □

Bemerkung 6.10. Aus (6.8) folgt die Fehler-Abschätzung

$$|f(x) - p_n(x)| \leq \frac{\max_{\xi \in [ab]} |f^{(n+1)}(\xi)|}{(n+1)!} |\prod_{i=0}^{n} (x - x_i)|\,.$$

Damit findet man für die lineare Interpolation ($n = 1$) die (sichere) Abschätzung

$$|f(x) - p_1(x)| \leq \frac{M_2}{2} \frac{(b-a)^2}{4}$$

für den Interpolationsfehler, wobei M_2 eine Schranke für den Betrag der 2-ten Ableitung von f ist.

Hat man bei den Stützstellen die freie Wahl und soll auf dem Intervall $[a, b]$ interpoliert werden, dann ist die Wahl der Nullstellen des Tschebyscheff-Polynoms $T_{n+1}(x)$ auf $[a, b]$ transformiert, d.h

$$x_k^* = \frac{a+b}{2} + \frac{b-a}{2} \cos\left(\frac{2(n+1-k)-1}{2(n+1)}\pi\right), \quad k = 0, \ldots, n \tag{6.9}$$

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

Satz 6.11. *Seien x_k irgendwelche (also auch z.B. äquidistante) und x_k^* gemäß* (6.9) *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

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

mit p_k^* als dem Interpolationspolynom, das die Interpolationsbedingung

$$p_k^*(x_j^*) = f(x_j^*),\ j = 0, \ldots, k,$$

erfüllt.

6.4 Newton-Interpolation

15. Vorlesung am 08.12.2014

Bei der Lagrange-Interpolation haben wir das Interpolationspolynom in der Lagrange-Basis entwickelt. Bei der Newton-Interpolation wird das eindeutig existierende Interpolationspolynom in der Newton-Basis entwickelt.
Ansatz:

$$p(x) = \sum_{k=0}^{n} c_k N_k(x)$$

Durch sukzessives Vorgehen erhalten wir durch Berücksichtigung der Stützstellen $(x_k, y_k), k = 0, \ldots, n$ die Koeffizienten der $N_k(x)$

$$\begin{aligned}
p_n(x_0) &= c_0 && = y_0 \rightsquigarrow c_0 \\
p_n(x_1) &= c_0 + c_1(x_1 - x_0) && = y_1 \rightsquigarrow c_1 \\
p_n(x_2) &= c_0 + c_1(x_1 - x_0) + c_2(x_2 - x_0)(x_2 - x_1) && = y_2 \rightsquigarrow c_2 \\
&\vdots \\
p_n(x_n) &= \sum_{k=0}^{n} c_k N_k(x_n) && = y_n \rightsquigarrow c_n
\end{aligned}$$

Definition 6.12.

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

heißt Newtonsches Interpolationspolynom.

Bemerkung. c_n ist der Koeffizient von x^n im Interpolationspolynom und c_k ist eindeutig festgelegt durch $x_0, \ldots, x_k, y_0, \ldots y_k$ d.h. durch die ersten k Stützstellen.

Definition 6.13. *Wir schreiben $c_k := f[x_0 x_1 \ldots x_k]$ für die Abbildung*

$$\{(x_0, y_0), \ldots, (x_k, y_k)\} \mapsto c_k$$

Betrachtet man Teilmengen der Stützstellen

$$x_{i_0}, \ldots, x_{i_k},$$

dann bezeichnet man das Interpolationspolynom an diesen Stützstellen mit

$$p^*_{i_0 i_1 \ldots i_k}(x)$$

wobei $i_0, \ldots, i_k$ paarweise verschiedene Zahlen aus $\{0, \ldots, k\}$ sind. Nach der Definition eines Interpolationypolynoms muss

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

Damit gilt

$$p^*_k(x) \equiv y_k \tag{6.10}$$

für das Polynom 0. Ordnung p^*_k (also $p^*_k(x) \neq p_k(x)$)

Bemerkung. p^*_k ist Konstante und $p_k(x)$ Polynom k-ter Ordnung, deshalb der Stern

Lemma 6.14. *Es gilt für alle* $k \in \{1, 2, \ldots, n\}$

$$p^*_{i_0,\ldots,i_k}(x) = \frac{(x - x_{i_0})p^*_{i_1\ldots i_k}(x) - (x - x_{i_k})p^*_{i_0\ldots i_{k-1}}(x)}{x_{i_k} - x_{i_0}} \tag{6.11}$$

Beweis. Induktion Die beiden rechts stehenden Polynome in (6.11) haben einen Grad $\leq k-1$ (damit der gesamte Ausdruck einen Grad $\leq k$).
Anfang ($k = 1$) ist trivial wegen (6.10).
Es ist zu zeigen, dass das rechts in (6.11) stehende Polynom das Interpolationypolynom zu den Stützstellen $x_{i_0}, \ldots, x_{i_k}$ ist (Ausdruck rechts von (6.11) bezeichnen wir mit $q(x)$) $\deg q(x) \leq k$ ist offensichtlich. Weiter ist

$$q(x_{i_0}) = \frac{0 - (x_{i_0} - x_{i_k})p^*_{i_0\ldots i_{k-1}}(x_{i_0})}{x_{i_k} - x_{i_0}} = y_{i_0}$$

und analog

$$q(x_{i_k}) = y_{i_k}$$

Schließlich für die restlichen Stützstellen $1 \leq j \leq k-1$

$$\begin{aligned} q(x_{i_j}) &= \frac{(x_{i_j} - x_{i_0})p^*_{i_1\ldots i_k}(x_{i_j}) - (x_{i_j} - x_{i_k})p^*_{i_0\ldots i_{k-1}}(x_{i_j})}{x_{i_k} - x_{i_0}} \\ &= \frac{(x_{i_j} - x_{i_0})y_{i_j} - (x_{i_j} - x_{i_k})y_{i_j}}{x_{i_k} - x_{i_0}} = y_{i_j} \end{aligned}$$

Damit erfüllt q die Interpolationsbedingung $q(x_{i_j}) = y_{i_j}, j = 0, \ldots, k$ also genau das, was $p^*_{i_0\ldots i_k}(x)$ leistet. Aufgrund der Eindeutigkeit des Interpolationspolynoms gilt also

$$q = p^*_{i_0\ldots i_k}$$

□

Satz 6.15. *Es gilt*

$$f[x_{i_0} \ldots x_{i_k}] = \frac{f[x_{i_1} \ldots x_{i_k}] - f[x_{i_0} \ldots x_{i_{k-1}}]}{x_{i_k} - x_{i_0}}$$

Beweis. Nach der Definition von $f[\ldots]$ ist dies gerade der Koeffizient von der höchsten Potenz des Interpolationspolynoms. Betrachten (6.11). 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

$$\frac{x \cdot f[x_{i_1} \ldots x_{i_k}]x^{k-1} - x \cdot f[x_{i_0} \ldots x_{i_{k-1}}]x^{k-1}}{x_{i_k} - x_{i_0}} = \frac{f[x_{i_1} \ldots x_{i_k}] - f[x_{i_0} \ldots x_{i_{k-1}}]}{x_{i_k} - x_{i_0}} x^k$$

$\rightsquigarrow$ Behauptung. □

Als Folgerung des Satzes 6.15 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:

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

Bemerkung 6.16. Stellt man jetzt die Newton-Interpolationspolynome der Elemente der Monombasis

$$\mathcal{M} = \{1, x, \ldots, x^n\}$$

auf, dann ergibt sich

$$\sum_{k=0}^{n} c_k N_k(x) = x^j = \sum_{k=0}^{j} c_k N_k(x)\ , \quad j = 0, \ldots, n\ ,$$

also kann man die Monombasis durch

$$\begin{pmatrix} f_0[x_0] & 0 & 0 & \ldots & 0 \\ f_1[x_0] & f_1[x_0x_1] & 0 & \ldots & 0 \\ \vdots & & & & \\ f_n[x_0] & f_n[x_0x_1] & f_n[x_0x_1x_2] & \ldots & f_n[x_0\ldots x_n] \end{pmatrix} \begin{pmatrix} N_0(x) \\ N_1(x) \\ \vdots \\ N_n(x) \end{pmatrix}$$

$$=: T_{nm} \begin{pmatrix} N_0(x) \\ N_1(x) \\ \vdots \\ N_n(x) \end{pmatrix} = \begin{pmatrix} 1 \\ x^1 \\ \vdots \\ x^n \end{pmatrix}$$

darstellen, wobei die Koeffizienten $f_k[x_0\ldots]$ in der k-ten Zeile der Matrix jeweils rekursiv ausgehend von $f_k[x_j] = x_j^k$ als dividierte Differenzen der Funktionen $f(x) = x^k$ konstruiert werden. Man erkennt T_{nm} als reguläre untere Dreiecksmatrix, womit der Nachweis, dass es sich bei

$$\mathcal{N} = \{N_0(x), N_1(x), \ldots, N_n(x)\}$$

tatsächlich um eine Basis von Π_n handelt, erbracht ist.
Für die Transformation der Koordinaten κ_n eines Polynoms bezüglich der

Newton-Basis in die Koordinaten des Polynoms bezüglich der Monombasis ergibt sich

$$\kappa_m = T_{nm}^{-T}\kappa_n \,. \tag{6.12}$$

Beispiel 6.17. Kehren wir zum obigen Beispiel mit der der Tabelle

k	0	1	2
x_k	0	1	2
y_k	1	0	4

zurück. Die Matrix T_{nm}, die die Basistransformation von Newton-Basis zur Monom-Basis realisiert, hat konkret die Form

$$T_{nm} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \,.$$

Damit errechnet man die Koordinaten bezüglich der Newton-Basis ausgehend von den Koordinaten in der Monombasis $\kappa_m = (1 \;\; -\frac{7}{2} \;\; \frac{5}{2})^T$ unter Nutzung von

$$T_{nm}^T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{zu} \;\; \kappa_m = T_{nm}^{-T}\kappa_n \Longleftrightarrow \kappa_n = T_{nm}^T\kappa_m = \begin{pmatrix} 1 \\ -1 \\ \frac{5}{2} \end{pmatrix} \,,$$

also ergibt sich das Polynom

$$p(x) = 1 - 1(x-0) + \frac{5}{2}(x-0)(x-1) = 1 - x + \frac{5}{2}x(x-1) \,.$$

6.4.1 Algorithmische Aspekte der Polynominterpolation - 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

$$p(x) = \sum_{k=0}^{n} c_k N_k(x), \quad c_k \text{ gegeben}$$

N_k rekursiv aufgebaut:

$$\begin{aligned} N_k(x) &= (x - x_0) \cdots (x - x_k) \\ \rightsquigarrow N_k(x) &= (x - x_{k-1}) N_{k-1}(x) \end{aligned}$$

p kann in der Form

$$p(x) = c_0 + (x - x_0)(c_1 + (x - x_1)(c_2 + \cdots + c_n(x - x_{n-1})) \cdots)$$

geschrieben werden. Daraus resultiert der Algorithmus:

Algorithmus 2 Wertet Newton Polynom mittels Horner-Schema aus

$u_{n+1} = 0$
for $k = n$ **downto** 0 **do**
  $u_k = (x - x_k)u_{k+1} + c_k$
end for
$p(x) = u_0$

Mit Laufzeit $3n$ flops.

6.5 Hermite-Interpolation

16. Vorlesung am 10.12.2014

Hat man einen Stützpunkt (x_0, y_0) vorgegeben, so ist damit ein Polynom 0-ten Grades festgelegt (Gerade parallel zur x-Achse). Hat man an der Stelle noch eine Ableitungsinformation, d.h. (x_0, y_0'), dann ist damit eine Gerade durch den Punkt (x_0, y_0) mit dem Anstieg y_0' festgelegt, also ein Polynom 1-ten Grades.

Satz 6.18. *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. MWS der dividierten Differenzen, Vorlesung □

Der Satz 6.18 rechtfertigt

Definition 6.19.

$$f[\underbrace{x, x, \ldots, x}_{n+2}] = \frac{f^{(n+1)}(x)}{(n+1)!} \tag{6.13}$$

Auf der Basis dieser Definition entstehen gemischte Differenzen wieder rekursiv, z.B.

$$f[x_0 x_1 x_2] = \frac{f[x_0 x_1] - f[x_1 x_2]}{x_0 - x_1}$$
$$f[x_0 x_0 x_0 x_1] = \frac{f[x_0 x_0 x_0] - f[x_0 x_0 x_1]}{x_0 - x_1}$$

Beispiel. 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

$$\begin{aligned}(x_0, y_0) &= (0, 1)\\ (x_0, y_0') &= (0, 2)\\ (x_0, y_0'') &= (0, 4)\\ (x_1, y_1) &= (1, 2)\\ (x_1, y_1') &= (1, 3)\end{aligned}$$

die allgemeine Form

	c_0	c_1	c_2	c_3	c_4
x_0	$f[x_0]$				
x_0	$f[x_0]$	$f[x_0x_0]$			
x_0	$f[x_0]$	$f[x_0x_0]$	$f[x_0x_0x_0]$		
x_1	$f[x_1]$	$f[x_0x_1]$	$f[x_0x_0x_1]$	$f[x_0x_0x_0x_1]$	
x_1	$f[x_1]$	$f[x_1x_1]$	$f[x_0x_1x_1]$	$f[x_0x_0x_1x_1]$	$f[x_0x_0x_0x_1x_1]$

bzw. mit unseren Daten 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, und es ergibt sich

$c_0 = f[x_0],\ c_1 = f[x_0x_0],\ c_2 = f[x_0x_0x_0],\ c_3 = f[x_0x_0x_0x_1],\ c_4 = f[x_0x_0x_0x_1x_1]$.

Führt man nun

$$\tilde{x}_0 = x_0,\ \tilde{x}_1 = x_0,\ \tilde{x}_2 = x_0,\ \tilde{x}_3 = x_1,\ \tilde{x}_4 = x_1$$

ein, so kann man das Interpolationspolynom in der Form

$$p(x) = \sum_{k=0}^{n} f[\tilde{x}_0 \dots \tilde{x}_k] \prod_{j=0}^{k-1} (x - \tilde{x}_j)$$

aufschreiben und in unserem Beispiel ergibt sich das Hermite-Interpolationspolynom

$$p(x) = \sum_{k=0}^{4} f[\tilde{x}_0 \dots \tilde{x}_k] \prod_{j=0}^{k-1} (x - \tilde{x}_j)$$

bzw.

$$p(x) = f[x_0] + f[x_0x_0](x - x_0) + f[x_0x_0x_0](x - x_0)^2$$
$$+f[x_0x_0x_0x_1](x - x_0)^3 + f[x_0x_0x_0x_1x_1](x - x_0)^3(x - x_1)$$

also für obige Werte

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

Die eben durchgeführte Konstruktion eines Interpolationspolynoms, bei der auch Ableitungsinformationen vorgegeben waren, basierte auf der Newton-Interpolation. Nun wollen wir für die Aufgabe der Bestimmung eines Interpolationspolynoms $h(x)$, das sowohl die Bedingungen

$$h(x_k) = y_k \ , \tag{6.14}$$

als auch die Bedingungen

$$h'(x_k) = y'_k \tag{6.15}$$

für $k = 0, \ldots, n$ erfüllt (y'_k steht für einen vorgegebenen Wert der Ableitung an der Stelle x, z.B. bei einer vorgegebenen Funktion $y'_k = f'(x_k)$). Man rechnet nun leicht (in der Vorlesung erfolgt, wer nicht da war, sollte es selbst nachrechnen!) nach, dass das Hermite-Polynom

$$h(x) = \sum_{i=0}^{n} L_i^2(x)[(1 - 2(x - x_i)L_i'(x_i))y_i + (x - x_i)y_i']$$

die Bedingungen (6.14) und (6.15) erfüllt.

6.6 Spline-Interpolation

Problem bei der Polynom-Interpolation:
Eventuell große Oszillationen durch Polynome höheren Grades bei Stützpunktzahlen ≥ 10
Deshalb:
Statt eines Interpolationspolynoms konstruiert man für $(x_k, y_k), k = 0, 1, \ldots, n$ in jeden Teilintervall einzelne Polynome, die an den Randstellen glatt ineinander übergehen. Betrachten mit

$$\Delta = \{a = x_0 < x_1 < \ldots < x_n = b\}$$

eine fest gewählte Zerlegung von $[a, b]$, wobei die Stützstellen $x_0, \ldots, x_N$ auch als Knoten bezeichnet werden.

Definition 6.20. *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

$$s(x_k) = f_k \quad \text{für} \quad k = 0, 1, \ldots, n \tag{6.16}$$

für $(x_k, f_k), k = 0, 1, \ldots, n$

6.6.1 Interpolierende lineare Splines $s \in S_{\Delta,1}$

Offensichtlich gilt:

$$s(x) = a_k + b_k(x - x_k), \quad x \in [x_k, x_{k+1}]$$

aus $s_k(x_k) = f_k$ sowie $s_k(x_{k+1}) = f_{k+1}$ folgt

$$a_k = f_k, \quad b_k = \frac{f_{k+1} - f_k}{x_{k+1} - x_k}, \quad k = 0, \ldots, n-1$$

Satz 6.21.

(a) Zur Zerlegung $\Delta = a = x_0 < \ldots < x_n = b$ und $f_0, \ldots, f_n$ gibt es genau einen Spline $s \in S_{\Delta,1}$ mit der Eigenschaft (6.16)

(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 liefert

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

6.6.2 Berechnung interpolierender kubischer Splines

Bevor wir uns mit Eigenschaften von kubischen Splines beschäftigen, wollen wir die Forderung der Glattheit zur Konstruktion nutzen.
Lokaler Ansatz

$$\begin{aligned} s(x) = a_k + b_k(x - x_k) + c_k(x - x_k)^2 + d_k(x - x_k)^3, \\ x \in [x_k, x_{k+1}], k = 0, \ldots, n-1 \end{aligned} \tag{6.17}$$

für $s : [a, b] \to \mathbb{R}$, wobei $s(x) =: p_k(x)$ auf dem Intervall $[x_k, x_{k+1}]$ verabredet wird.
Aufgabe: Bestimmung von $a_k, \ldots, d_k, k = 0, \ldots, n-1$ so, dass s auf $[a, b]$ zweimal stetig differenzierbar ist und darüberhinaus in den Knoten vorgegebene Werte $f_0, \ldots, f_n \in \mathbb{R}$ interpoliert

$$s(x_k) = f_k, \quad k = 0, \ldots, n$$

Setzen $h_k := x_{k+1} - x_k,\ k = 0, \ldots, n$

Lemma 6.22. *Falls $n+1$ reelle Zahlen $s_0'', \ldots, s_n'' \in \mathbb{R}$ den folgenden $n-1$ gekoppelten Gleichungen ($k = 1, \ldots, n-1$)*

$$h_{k-1} \underbrace{s_{k-1}''}_{m_{k-1}} + 2(h_{k-1} + h_k) \underbrace{s_k''}_{m_k} + h_k \underbrace{s_{k+1}''}_{m_{k+1}} = \underbrace{6\frac{f_{k+1} - f_k}{h_k} - 6\frac{f_k - f_{k-1}}{h_{k-1}}}_{c_k} \tag{6.18}$$

genügen, so liefert der lokale Ansatz (6.17) *mit den Setzungen*

$$c_k = \frac{m_k}{2}, \quad a_k = f_k, \quad d_k = \frac{m_{k+1} - m_k}{6h_k}, \quad b_k = \frac{f_{k+1} - f_k}{h_k} - \frac{h_k}{6}(m_{k+1} + 2M_k)$$

für $k = 0, \ldots, n-1$ eine kubische Splinefunktion $s \in S_{\Delta,3}$, die die Interpolationsbedingung $s(x_k) = f_k$ erfüllt.

Beweis. Vorlesung oder Bärwolff bzw. Plato □

Bemerkung. Die **Momente** $m_0, \ldots, m_n$ stimmen mit den 2. Ableitungen der Splinefunktion s in den Knoten x_k überein

$$s_k'' = m_k = s''(x_k), \quad k = 0, \ldots, n$$

(6.18) 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 $s_0' = s_n',\ s_0'' = s_n''$

6.6.3 Gestalt der Gleichungssysteme

Natürliche Randbedingunden

$$\begin{bmatrix} 2(h_0+h_1) & h_1 & & 0 \\ h_1 & 2(h_1+h_2) & \ddots & \\ & \ddots & & h_{n-2} \\ 0 & & h_{n-2} & 2(h_{n-2}+h_{n-1}) \end{bmatrix} \begin{bmatrix} m_1 \\ \\ \\ m_{n-1} \end{bmatrix} = \begin{bmatrix} c_1 \\ \\ \vdots \\ \\ c_{n-1} \end{bmatrix} \tag{6.19}$$

Vollständige Randbedingungen

$$\begin{bmatrix} 2h_0 & h_0 & & & 0 \\ h_0 & 2(h_0+h_1) & \ddots & & \\ & \ddots & 2(h_{n-2}+h_{n-1}) & h_{n-1} & \\ 0 & & h_{n-1} & 2h_{n-1} \end{bmatrix} \begin{bmatrix} m_0 \\ \\ \\ m_n \end{bmatrix} = \begin{bmatrix} c_0 \\ \\ \vdots \\ \\ c_n \end{bmatrix} \tag{6.20}$$

17. Vorlesung 15.12.2014

Dieses Gleichungssystem erhält man durch die Beziehungen

$$s'(x_0) = p_0'(x_0) = b_0 = \frac{f_1 - f_0}{h_0} - \frac{h_0}{6}(m_1 + 2m_0) \tag{6.21}$$

$$s'(x_n) = p_{n-1}'(x_n) = \frac{f_n - f_{n-1}}{h_{n-1}} + \frac{h_{n-1}}{6}(m_{n-1} + 2m_n)\ , \tag{6.22}$$

woraus sich mit $s'(x_0) = f'_0$ bzw. $s'(x_n) = f'_n$ die beiden Gleichungen

$$\begin{aligned} 2h_0m_0 + h_0m_1 &= -6s'(x_0) + 6\frac{f_1 - f_0}{h_0} = -6f'_0 + 6\frac{f_1 - f_0}{h_0} =: c_0 \\ h_{n-1}m_{n-1} + 2h_{n-1}m_n &= 6s'(x_n) - 6\frac{f_n - f_{n-1}}{h_{n-1}} = 6f'_n - 6\frac{f_n - f_{n-1}}{h_{n-1}} =: c_n \end{aligned}$$

und damit (6.20) ergeben.

periodische Randbedingungen

$$\begin{bmatrix} 2(h_{n-1}+h_0) & h_0 & & h_{n-1} \\ h_0 & 2(h_0+h_1) & \ddots & \\ & \ddots & & h_{n-2} \\ h_{n-1} & & h_{n-2} & 2(h_{n-2}+h_{n-1}) \end{bmatrix} \begin{bmatrix} m_0 \\ \\ \\ m_{n-1} \end{bmatrix} = \begin{bmatrix} c_0 \\ \\ \vdots \\ \\ c_{n-1} \end{bmatrix} \tag{6.23}$$

Die erste Gleichung des Systems ergibt sich unter Nutzung der periodischen Bedingungen $m_0 = m_n$ bzw. $s'_0 = s'_n$ und der Beziehungen (6.21), (6.22) zu

$$\frac{f_1 - f_0}{h_0} - \frac{h_0}{6}(m_1 + 2m_0) = \frac{f_n - f_{n-1}}{h_{n-1}} + \frac{h_{n-1}}{6}(m_{n-1} + 2m_0)$$

bzw.

$$2(h_{n-1} + h_0)m_0 + h_0m_1 + h_{n-1}m_{n-1} = 6\frac{f_1 - f_0}{h_0} - 6\frac{f_n - f_{n-1}}{h_{n-1}} =: c_0 \, .$$

Die letzte Gleichung des Systems (6.23) ergibt sich mit $m_0 = m_n$ unmittelbar.

6.7 Existenz und Eindeutigkeit der betrachteten interpolierenden kubischen Splines

Alle Koeffizientenmatrizen der Gleichungssysteme zur Berechnung der Momente $m_k = s''_k$ haben die Eigenschaft, strikt diagonal dominant zu sein, d.h. es gilt für die Matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$

$$\sum_{k\neq j=1}^{n} |a_{kj}| < |a_{kk}| \, , \quad k = 1, \ldots, n \, . \tag{6.24}$$

Lemma 6.23. *(s. dazu auch Kapitel 4)*
Jede strikt diagonal dominante Matrix $A = (a_{kj}) \in \mathbb{R}^{n\times n}$ ist regulär und es gilt

$$\|x\|_\infty \leq \max_{k=1,\dots,n} \left\{ (|a_{kk}| - \sum_{k\neq j=1}^{n} |a_{kj}|)^{-1} \right\} \|Ax\|_\infty , \quad x \in \mathbb{R}^n \tag{6.25}$$

Beweis. Für $x \in \mathbb{R}^n$ sei der Index $x \in \{1,\dots,n\}$ so gewählt, dass $|x_k| = \|x\|_\infty$ gilt. Dann findet man

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

Dies liefert die Gültigkeit von (6.25) woraus die Regularität von A direkt folgt. (Aus $Ax = 0$ folgt $x = 0$ als einzige Lösung) □

Korollar 6.1. *Zur Zerlegung Δ und den Werten $f_0,\dots,f_n \in \mathbb{R}$ 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. □

Betrachte nun $S_{\Delta,3}$, und verwenden

$$\|u\|_2 := \left(\int_a^b |u(x)|^2 \,\mathrm{d}x \right)^{\frac{1}{2}}$$

Lemma 6.24. *Wenn eine Funktion $f \in C^2[a,b]$ und eine kubische Splinefunktion $s \in S_{\Delta,3}$ in den Knoten übereinstimmen, d.h.*

$$s(x_k) = f(x_k) \quad \textit{für} \quad k = 0,\dots,n$$

so gilt

$$\|f'' - s''\|_2^2 = \|f''\|_2^2 - \|s''\|_2^2 - 2([f' - s']s'')(x)\Big|_{x=a}^{x=b} \tag{6.26}$$

Beweis.

$$\|f'' - s''\|_2^2 = \int_a^b |f''(x) - s''(x)|^2 \,\mathrm{d}x = \|f''\|_2^2 - 2\int_a^b (f''s'')(x)\mathrm{d}x + \|s''\|_2^2$$
$$= \|f''\|_2^2 - 2\int_a^b ([f'' - s'']s'')(x)\mathrm{d}x - \|s''\|_2^2$$

Für den mittleren Term ergibt die partielle Integration

$$\int_{x_{k-1}}^{x_k} ([f'' - s'']s'')(x)\mathrm{d}x$$
$$= ([f' - s']s'')(x)\Big|_{x_{k-1}}^{x_k} - \int_{x_{k-1}}^{x_k} ([f' - s']s''')(x)\mathrm{d}x$$
$$= ([f' - s']s'')(x)\Big|_{x_{k-1}}^{x_k} - \underbrace{([f - s]s''')(x)\Big|_{x_{k-1}}^{x_k}}_{=0} + \underbrace{\int_{x_{k-1}}^{x_k} ([f - s]s^{(4)})(x)\mathrm{d}x}_{=0}$$

Die Summation über $k = 1, \ldots, n$ ergibt

$$\int_a^b ([f'' - s'']s'')(x)\mathrm{d}x = \sum_{k=1}^{n} \{([f' - s']s'')(x_k) - ([f' - s']s'')(x_{k-1})\}$$
$$= ([f' - s']s'')(b) - ([f' - s']s'')(a)$$

□

Satz 6.25. *Gegeben sei $f \in C^2[a,b]$ und ein kubischer Spline $s \in S_{\Delta,3}$ mit $s(x_k) = f(x_k), k = 0, \ldots, n$. Dann gilt die Identität*

$$\|f''\|_2^2 - \|s''\|_2^2 = \|f'' - s''\|_2^2 \tag{6.27}$$

sofern eine der 3 folgenden Bedingungen erfüllt ist

(a) $s''(a) = s''(b) = 0$ *(natürliche RB)*

(b) $s'(a) = f'(a),\ s'(b) = f'(b)$ *(vollst. RB)*

(c) $f'(a) = f'(b),\ s'(a) = s'(b),\ s''(a) = s''(b)$ *(period. RB)*

Beweis. Die Aussage des Satzes ergibt sich durch Berücksichtigung von $(a), (b)$ bzw (c) in der Identität (6.26) □

Korollar 6.2. *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 Deformationsenergie (Deformationsenergie eines interpolierenden Stabes g ist $E = \frac{1}{2}\int_a^b g''(x)^2\,dx$), es gilt also*

$$\|s''\|_2 \leq \|f''\|_2$$

für jede Funktion $f \in C^2[a,b]$ mit $f(x_k) = f_k$ für $k = 0, \ldots, n$

Beweis. Folgt direkt aus (6.27) □

6.8 Fehlerabschätzungen für interpolierende kubische Splines

Zuerst schreiben wir die Gleichungen (6.18) für die Momente durch die jeweilige Division durch $3(h_{k-1} + h_k)$ in der Form

$$\begin{aligned}&\frac{h_{k-1}}{3(h_{k-1}+h_k)}s''_{k-1} + \frac{2}{3}s''_k + \frac{h_k}{3(h_{k-1}+h_k)}s''_{k+1}\\ &= 2\frac{f_{k+1}-f_k}{h_k(h_{k-1}+h_k)} - 2\frac{f_k - f_{k-1}}{h_{k-1}(h_{k-1}+h_k)} =: \hat{c}_k\end{aligned}$$

auf, was für natürliche Randbedingungen auf das Gleichungssystem

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

$$B \begin{bmatrix} s''_1 \\ \vdots \\ s''_{n-1} \end{bmatrix} = \begin{bmatrix} \hat{c}_1 \\ \vdots \\ \hat{c}_{n-1} \end{bmatrix} \tag{6.28}$$

führt ($h_k = x_{k+1} - x_k$). Die Eigenschaften der Matrix B, werden in den Fehlerabschätzungen für interpolierende kubische Splines wesentlich benutzt.

Lemma 6.26. *Zu einer gegebenen Funktion $f \in C^4[a,b]$ mit $f''(a) = f''(b) = 0$ bezeichne $s \in S_{\Delta,3}$ den interpolierenden kubischen Spline mit natürlichen Randbedingungen. Dann gilt*

$$\max_{k=1,\ldots,n-1} |s''(x_k) - f''(x_k)| \leq \frac{3}{4} \left\| f^{(4)} \right\|_\infty h_{\max}^2 \tag{6.29}$$

Beweis. (Beweisskizze)
Die Aussage des Lemmas wird unter Nutzung einer Beziehung, der Form

$$B \begin{bmatrix} f''(x_1) - s_1''(x_1) \\ \vdots \\ f''(x_{n-1}) - s_{n-1}'' \end{bmatrix} = \begin{bmatrix} \delta_1 - \hat{\delta}_1 \\ \vdots \\ \delta_{n-1} - \hat{\delta}_{n-1} \end{bmatrix} \tag{6.30}$$

nachgewiesen, die man durch Taylorentwicklungen von f'' und f erhält, wobei δ_j und $\hat{\delta}_j$ jeweils von der Ordnung $\mathbf{O}(h_{max}^2)$, $h_{max} = \max_{k=0,\ldots,n-1} x_{k+1} - x_k$, sind.
Für die strikt diagonal dominante Matrix B kann man die Abschätzung

$$\|x\|_\infty \leq (|b_{kk}| - \sum_{k \neq j=1}^{n-1} |b_{kj}|)^{-1} \|Bx\|_\infty$$

nachweisen (Übung), und mit

$$|b_{kk}| - \sum_{k \neq j=1}^{n-1} |b_{kj}| = \frac{2}{3} - \frac{h_k}{3(h_{k+1} + h_k)} - \frac{h_{k+1}}{3(h_{k+1} + h_k)} = \frac{1}{3} \ , \ k = 2, \ldots, n-2,$$

erhält man letztendlich die Beziehung (6.29) (die erste und die letzte Gleichung des Systems (6.30) sind auf Grund der Randbedingungen trivial). □

Das Lemma 6.26 ist die Grundlage für den folgenden Satz zur Fehlerabschätzung der Spline-Interpolation

Satz 6.27. *Sei $f \in C^4[a,b]$ und sei $s \in S_{\Delta,3}$ ein interpolierender kubischer Spline. Weiter bezeichne $h_k = x_{k+1} - x_k$ für $k = 0, \ldots, n-1$ und*

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

Falls

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

erfüllt ist mit einer Konstanten $C > 0$, so gelten mit der Zahl $c := \frac{h_{\max}}{h_{\min}}(C+\frac{1}{4})$ die folgenden Abschätzungen für jedes $x \in [a,b]$

$$|s(x) - f(x)| \le c \left\|f^{(4)}\right\|_\infty h_{\max}^4 \tag{6.31}$$

$$|s'(x) - f'(x)| \le c \left\|f^{(4)}\right\|_\infty h_{\max}^3 \tag{6.32}$$

$$|s''(x) - f''(x)| \le c \left\|f^{(4)}\right\|_\infty h_{\max}^2 \tag{6.33}$$

$$\left|s^{(3)}(x) - f^{(3)}(x)\right| \le c \left\|f^{(4)}\right\|_\infty h_{\max} \tag{6.34}$$

Beweis. Zuerst wird (6.34) nachgewiesen. s'' ist als 2. Ableitung eines Polynoms 3. Grades affin linear auf $[x_k, x_{k+1}]$ für $k = 0, \ldots, n-1$, d.h.

$$s^{(3)}(x) \equiv \frac{s''(x_{k+1}) - s''(x_k)}{h_k} = \text{const}, \quad x_k \le x \le x_{k+1} \tag{6.35}$$

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

$$f^{(3)}(x) = \frac{f''(x_{k+1}) - f''(x_k)}{h_k} - \frac{(x_{k+1}-x)^2}{2h_k} f^{(4)}(\alpha_k) + \frac{(x-x_k)^2}{2h_k} f^{(4)}(\beta_k) \tag{6.36}$$

für gewisse Zwischenstellen $\alpha_k, \beta_k \in [a,b]$. Subtraktion von (6.35) und (6.36) ergibt

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

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

wobei

$$\begin{aligned} (x_{k+1}-x)^2 + (x-x_k)^2 &= (x_{k+1}-x_k)^2 - 2(x_{k+1}-x)(x-x_k) \\ &\le (x_{k+1}-x_k)^2 \le h_{\max}^2 \quad \forall x \in [x_k, x_{k+1}] \end{aligned}$$

berücksichtigt wurde.
Die restlichen Fehlerabschätzungen (6.33), (6.32), (6.31) erhält man durch sukzessive Integration von (6.34) unter Nutzung des Hauptsatzes der Differential- und Integralrechnung. □

Bemerkung. Die wesentliche Voraussetzung des eben bewiesenen Satzes über den Fehler der 2. Ableitungen in den Knoten ist typischerweise erfüllt (siehe auch Hilfssatz 6.26 für den Fall natürlicher Randbedingungen).

Kapitel 7

Numerische Integration

18. Vorlesung am 17.12.2014

Ziel ist die Berechnung des bestimmten Integrals

$$\int_a^b f(x)\mathrm{d}x$$

wobei man aus unterschiedlichen Gründen nicht die Berechnung mittels einer Stammfunktion $F(x)$ durch

$$\int_a^b f(x)\mathrm{d}x = F(b) - F(a)$$

nutzen kann oder will. Entweder findet man kein auswertbares $F(x)$ wie im Fall von $f(x) = \frac{e^x}{x}$ oder $f(x) = e^{-x^2}$ oder die Berechnung von $F(b), F(a)$ ist zu mühselig.

7.1 Interpolatorische Quadraturformeln

Definition 7.1. *Die Näherungsformel*

$$Q_n(f) = \int_a^b p_n(x)\mathrm{d}x = (b-a)\sum_{k=0}^{n} f(x_k)\sigma_k \tag{7.1}$$

mit dem Interpolationspolynom p_n zu den Stützpunkten $(x_0, f(x_0)), \ldots, (x_n, f(x_n))$ für das Integral $\int_a^b f(x)\mathrm{d}x$ nennt man **interpolatorische Quadraturformel**.

Definition 7.2. *Mit*

$$E_n[f] = \int_a^b f(x)\mathrm{d}x - Q_n = I(f) - Q_n(f) \tag{7.2}$$

bezeichnet man den Fehler der Quadraturformel Q_n. Eine Quadraturformel hat den Genauigkeitsgrad $m \in \mathbb{N}$, wenn sie alle Polynome $q(x)$ bis zum Grad m exakt integriert, d.h. $E_n[q] = 0$ ist, und m die größtmögliche Zahl mit dieser Eigenschaft ist.

Es gilt offensichtlich der folgende

Satz 7.3. *Zu den $n+1$ beliebig vorgegebenen paarweise verschiedenen Stützstellen $a \le x_0 < \cdots < x_n \le b$ existiert eine eindeutig bestimmte interpolatorische Quadraturformel deren Genauigkeitsgrad mindestens gleich n ist.*

Satz 7.4. *Eine interpolatorische Quadraturformel Q_n besitzt die Gestalt*

$$Q_n(f) = (b-a)\sum_{k=0}^{n} \sigma_k f(x_k) \quad \textit{mit } \sigma_k = \int_0^1 \prod_{\substack{j=0 \\ j\neq k}}^{n} \frac{t-t_j}{t_k-t_j}\,dt, \;\; t_j = \frac{x_j-a}{b-a}\,. \tag{7.3}$$

Beweis. Die Beziehungen rechnet man ausgehend von dem Lagrangeschen Interpolationspolynom $p_n(x) = \sum_{k=0}^{n} f(x_k)L_k(x)$ nach und findet mit einer geeigneten Substitution

$$\frac{1}{b-a}\int_a^b L_k(x)\,dx = \cdots = \sigma_k = \int_0^1 \prod_{\substack{j=0 \\ j\neq k}}^{n} \frac{t-t_j}{t_k-t_j}\,dt = \sigma_k\,.$$

□

Bemerkung 7.5. Durch die (exakte) Integration der Funktion $f \equiv 1$ mit einer interpolatorischen Quadraturformel ergibt sich für die Gewichte die charakteristische Eigenschaft

$$\sum_{k=0}^{n} \sigma_k = 1\,.$$

7.2 Fehler bei der interpolatorischen Quadratur

Die Abschätzung der Fehler einer interpolatorischen Quadratur basiert auf dem Fehler, den man bei der Interpolation der Funktion durch das Interpolationspolynom macht. Es gilt der

Satz 7.6. *Die interpolatorische Quadraturformel* $Q_n(f) = (b-a)\sum_{k=0}^{n} \sigma_k f(x_k)$ *besitze mindestens den Genauigkeitsgrad* $r \geq n$*, und die Funktion* $f : [a,b] \to \mathbb{R}$ *sei* $(r+1)$*-mal stetig diff'bar. Dann gilt die Fehlerabschätzung*

$$|I(f) - Q_n(f)| \leq c_r \frac{(b-a)^{r+2}}{(r+1)!} \max_{\xi \in [a,b]} |f^{(r+1)}(\xi)| \tag{7.4}$$

mit

$$c_r = \min_{t_{n+1},\ldots,t_r \in [0,1]} \int_0^1 \prod_{k=0}^{r} |t - t_k| \, dt, \quad t_k = \frac{x_k - a}{b-a}, \; k = 0, 1, \ldots, n \; . \tag{7.5}$$

Wenn mit Werten $t_0, \ldots, t_n$ *aus* (7.5) *für eine bestimmte Wahl von* $t_{n+1}, \ldots, t_r \in [0,1]$ *das Produkt* $\Pi_{k=0}^{r}(t - t_k)$ *von einem Vorzeichen*[1] *in* $[0,1]$ *ist, so gilt mit einer Zwischenstelle* $\xi \in [a,b]$ *die Fehlerdarstellung*

$$I(f) - Q_n(f) = c_r' \frac{(b-a)^{r+2}}{(r+1)!} f^{(r+1)}(\xi) \quad \textit{mit} \;\; c_r' = \int_0^1 \prod_{k=0}^{r} (t - t_k) \, dt \; . \tag{7.6}$$

Beweis. s. Vorlesung oder Plato □

7.3 Numerischen Integration mit Newton-Cotes-Formeln

Als spezielle interpolatorische Quadraturformeln sollen nun die Newton-Cotes-Formeln diskutiert werden.

- 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

$$\int_a^b p_n(x)\mathrm{d}x \approx \int_a^b f(x)\mathrm{d}x$$

[1] Eine reelle Funktion φ heißt **von einem Vorzeichen** auf dem Intervall $[c,d]$, wenn $\varphi(x) \geq 0$ f.a. $x \in [c,d]$ oder $\varphi(x) \leq 0$ f.a. $x \in [c,d]$ gilt.

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

$$Q_n(f) = \int_a^b p_n(x)\mathrm{d}x = (b-a)\sum_{k=0}^{n} f_k\sigma_k \tag{7.7}$$

mit den Gewichten

$$\sigma_k = \frac{1}{n}\int_0^n \prod_{k\neq j=0}^{n} \frac{s-j}{k-j}\mathrm{d}s, \quad k = 0,\ldots,n \tag{7.8}$$

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

$$\int_a^b f(x)\mathrm{d}x \approx Q_1(f) = \int_a^b p_1(x)\mathrm{d}x = \frac{b-a}{2}(f(a)+f(b)) \tag{7.9}$$

die **Trapezregel** folgt.
Für $n = 2$ ergibt sich

$$\begin{aligned}\sigma_0 &= \frac{1}{2}\int_0^2 \frac{s-1}{0-1}\cdot\frac{s-2}{0-2}\mathrm{d}s = \frac{1}{4}\int_0^2 (s^2-3s+2)\mathrm{d}s \\ &= \frac{1}{4}\left[\frac{s^3}{3}-\frac{3s^2}{2}+2s\right] = \frac{1}{4}\left[\frac{8}{3}-6+4\right] = \frac{1}{4}\left[\frac{8-6}{3}\right] = \frac{1}{6}\end{aligned}$$

$$\sigma_2 = \frac{1}{6}, \quad \sigma_1 = \frac{4}{6}$$

woraus mit

$$\int_a^b f(x)\mathrm{d}x \approx Q_2(f) = \int_a^b p_2(x)\mathrm{d}x = \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right) \tag{7.10}$$

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

$$\begin{aligned}\int_a^b f(x)\mathrm{d}x \approx Q_3(f) &= \int_a^b p_3(x)\mathrm{d}x \\ &= \frac{b-a}{8}\left(f(a) + 3f\left(\frac{2a+b}{3}\right) + 3f\left(\frac{a+2b}{3}\right) + f(b)\right)\end{aligned} \tag{7.11}$$

die Newtonsche $\frac{3}{8}$-Regel.
Gilt für die Stützstellen $x_k = a + kh, h = \frac{b-a}{n}, k = 0, \ldots, n$, also $x_0 = a$ und $x_n = b$, spricht man bei der Quadraturformel von einer **abgeschlossenen Newton-Cotes-Quadraturformel**.
Für die Simpsonregel findet man für den Genauigkeitsgrad

$$\begin{aligned}E_2[x^3] &= \int_a^b x^3\mathrm{d}x - \frac{b-a}{6}\left[a^3 + 4\left(\frac{a+b}{2}\right)^3 + b^3\right] \\ &= \frac{1}{4}(b^4 - a^4) - \frac{b-a}{6}\left[a^3 + \frac{1}{2}(a^3 + 3a^2b + 3ab^2 + b^3) + b^3\right] \\ &= 0\end{aligned}$$

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

$$
\begin{aligned}
E_1[f] &= -\frac{1}{12}h^3 f''(\xi), \quad h = b-a \\
E_2[f] &= -\frac{1}{90}h^5 f^{(4)}(\xi), \quad h = \frac{b-a}{2} \\
E_3[f] &= -\frac{3}{80}h^5 f^{(4)}(\xi), \quad h = \frac{b-a}{3} \\
E_4[f] &= -\frac{8}{945}h^7 f^{(6)}(\xi), \quad h = \frac{b-a}{4}
\end{aligned}
$$

wobei $\xi \in [a,b]$ jeweils ein geeigneter Zwischenwert ist. Diese Fehlerdarstellungen ergeben sich nach Satz 7.6. Wir wollen dies für den Fehler $E_2[f]$ der Simpson-Formel zeigen. Nach Satz 7.6 gilt

$$
E_2[f] = I(f) - Q_2(f) = c_3' \frac{(b-a)^5}{4!} f^{(4)}(\xi) \quad \text{mit} \;\; c_3' = \int_0^1 \prod_{k=0}^{3} (t-t_k)\, dt \;, \tag{7.12}
$$

vorausgesetzt, dass $\phi(t) = \prod_{k=0}^{3}(t-t_k)$ für ein geeignetes $t_3 \in [0,1]$ von einem Vorzeichen ist. Bei der Simpsonformel ergeben sich die auf das Intervall $[0,1]$ transformierten Stützstellen zu $t_0 = 0$, $t_1 = 1/2$ und $t_2 = 1$. Damit haben wir

$$
\phi(t) = t(t-\frac{1}{2})(t-1)(t-t_3)
$$

und man erkennt, dass der Faktor $t(t-\frac{1}{2})(t-1)$ im Intervall $[0,\frac{1}{2}]$ größer oder gleich Null ist, und im Intervall $[\frac{1}{2},1]$ kleiner oder gleich Null ist. Damit wird $\phi(t) \leq 0$ für alle $t \in [0,1]$, also von einem Vorzeichen, für $t_3 = \frac{1}{2}$. Damit ergibt sich für c_3'

$$
c_3' = \int_0^1 t(t-\frac{1}{2})(t-1)(t-\frac{1}{2})\, dt = \cdots = -\frac{1}{120} \;.
$$

Für $E_2[f]$ erhält man damit

$$
E_2[f] = I(f) - Q_2(f) = -\frac{1}{120}\frac{(b-a)^5}{4!} f^{(4)}(\xi) = -\frac{(2h)^5}{120 \cdot 24} f^{(4)}(\xi) = -\frac{h^5}{90} f^{(4)}(\xi) \;.
$$

Bemerkung 7.7. Hilfreich für die Berechnung der Gewichte der abgeschlossenen Newton-Cotes-Formeln ist die Symmetrie-Eigenschaft

$$
\sigma_{n-k} = \sigma_k \quad \text{für} \;\; k = 0, 1, \ldots, n \;,
$$

die als Übung (Vorlesung !!!) unter Nutzung der nachzuweisenden Identität

$$
L_{n-k}(x) = L_k(b+a-x) \;, \quad x \in [a,b]
$$

nachgerechnet werden sollte.

7.4 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 Interpolation 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 ($h = \frac{\beta-\alpha}{N}$)

$$\begin{aligned} S_{1,N} &= h\left[\frac{1}{2}y_{10} + y_{20} + \cdots + y_{N0} + \frac{1}{2}y_{N1}\right] \\ &= h\left[\frac{1}{2}(y_{10} + y_{N1}) + \sum_{k=2}^{N} y_{k0}\right] \end{aligned} \tag{7.13}$$

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

$$S_{2,N} = \frac{h}{3}\left[(y_{10} + y_{N2}) + 2\sum_{j=1}^{N-1} y_{j2} + 4\sum_{j=1}^{N} y_{j1}\right] \tag{7.14}$$

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

Satz 7.8. *Wenn $f(x)$ in $[\alpha, \beta]$ für gerades n eine stetige $(n+2)$-te Ableitung und für ungerades n eine stetige $(n+1)$-te Ableitung besitzt, dann existiert ein Zwischenwert $\xi \in]\alpha, \beta[$, sodass die Beziehungen*

$$E_{S_{n,N}}[f] = Kh^{n+2} f^{(n+2)}(\xi)$$

für gerades n und

$$E_{S_{n,N}}[f] = Lh^{n+1} f^{(n+1)}(\xi)$$

für ungerades n gelten, wobei K und L von α, β abhängige Konstanten sind, und $h = \frac{\beta-\alpha}{nN}$ gilt.

Beweis. S. Satz 7.6 bzw. Plato, Bärwolff □

Frohe Weihnachten und ein erfolgreiches Jahr 2015

7.5 Gauß-Quadraturen

19. Vorlesung am 5.01.2015

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 Stützstellen oft $\lambda_1, \ldots, \lambda_n$, da sie sich letztendlich als Nullstellen eines Polynoms n-ten Grades ergeben werden. Wir wollen sie im Folgenden aber weiter mit $x_1, \ldots, x_n$ bezeichnen und beginnen aber im Unterschied zu den Newton-Cotes-Formeln bei $k = 1$ zu zählen.

Ziel ist die Berechnung des Integrals $\int_a^b g(x)\mathrm{d}x$ wobei man die zu integrierende Funktion in der Form $g(x) = f(x)\rho(x)$ mit einer Funktion $\rho(x)$, die mit der

evtl. Ausnahme von endlich vielen Punkten auf $[a, b]$ positiv sein soll, vorgibt. $\rho(x)$ heißt **Gewichtsfunktion**. Es ist also das Integral

$$I = \int_a^b f(x)\rho(x)\mathrm{d}x = \int_a^b g(x)\mathrm{d}x$$

numerisch zu berechnen. Im Folgenden geht es darum, Stützstellen $x_k \in [a, b]$ und Integrationsgewichte σ_k so zu bestimmen, dass

$$I_n = \sum_{j=1}^{n} \sigma_j f(x_j) \tag{7.15}$$

eine möglichst gute Näherung des Integrals I ergibt. Fordert man, dass die Formel (7.15) für alle Polynome $f(x)$ bis zum Grad $2n-1$, d.h. für $x^0, x^1, \ldots, x^{2n-1}$ exakt ist und somit $I_n = I$ gilt, dann müssen die Stützstellen $x_1, \ldots, x_n$ und die Gewichte $\sigma_1, \ldots, \sigma_n$ Lösungen des Gleichungssystems

$$\sum_{j=1}^{n} \sigma_j x_j^k = \int_a^b x^k \rho(x)\mathrm{d}x \quad (k = 0, 1, \ldots, 2n-1) \tag{7.16}$$

sein.
Wir werden im Folgenden zeigen, dass das Gleichungssystem (7.16) 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 (7.16) mit

$$\int_{-1}^{1} \mathrm{d}x = 2, \quad \int_{-1}^{1} x\mathrm{d}x = 0, \quad \int_{-1}^{1} x^2\mathrm{d}x = \frac{2}{3}, \quad \int_{-1}^{1} x^3 = 0$$

das Gleichungssystem

$$\begin{aligned} \sigma_1 + \sigma_2 &= 2 \\ \sigma_1 x_1 + \sigma_2 x_2 &= 0 \\ \sigma_1 x_1^2 + \sigma_2 x_2^2 &= \frac{2}{3} \\ \sigma_1 x_1^3 + \sigma_2 x_2^3 &= 0 \end{aligned} \tag{7.17}$$

Für (7.17) findet man mit

$$x_1 = -\frac{1}{\sqrt{3}}, \quad x_2 = \frac{1}{\sqrt{3}}, \quad \sigma_1 = \sigma_2 = 1$$

eine Lösung und damit ist die Quadraturformel

$$I_2 = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$

für alle Polynome $f(x)$ bis zum Grad 3 exakt, d.h. es gilt

$$\int_{-1}^{1} f(x)\mathrm{d}x = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$

Wir sind also *besser* als mit der Trapezregel.

7.5.1 Orthogonale Polynome

Die beiden Stützstellen aus dem eben diskutierten Beispiel sind mit $-\frac{1}{\sqrt{3}}$ und $\frac{1}{\sqrt{3}}$ gerade die Nullstellen des Legendre-Polynoms $p_2(x) = x^2 - \frac{1}{3}$ zweiten Grades. Das ist kein Zufall, sondern darin steckt eine Systematik. Deshalb sollen im Foglenden orthogonale Polynome besprochen werden.
Mit einer Gewichtsfunktion $\rho(x)$ statten wir den Vektorraum P aller Polynome über dem Körper der reellen Zahlen mit dem Skalarprodukt

$$\langle p, q\rangle_\rho := \int_a^b p(x)q(x)\rho(x)\mathrm{d}x \tag{7.18}$$

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

$$\|p\|_\rho^2 = \langle p, p\rangle_\rho = \int_a^b p^2(x)\rho(x)\mathrm{d}x \tag{7.19}$$

eine Norm definiert. Der Nachweis, dass (7.18), (7.19) Skalarprodukt bzw. Norm sind, sollte als Übung betrachtet werden.

Definition 7.9. *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

$$\mathrm{span}\{p_1, \ldots, p_n\} = \{c_1p_1 + \cdots + c_np_n | c_1, \ldots, c_n \in K\}$$

definiert, wobei K der Zahlkörper ist, über dem der Vektorraum der Polynome P betrachtet wird (und wenn nichts anderes gesagt wird, betrachten wir $K = \mathbb{R}$)

7.5.2 Konstruktion von Folgen orthogonaler Polynome

Wir wissen, dass die Monome $1, x, \ldots, x^n, \ldots$ eine Basis zur Konstruktion von Polynomen bilden. Mit $p_0(x) = 1$ wird durch

$$p_n(x) = x^n - \sum_{j=0}^{n-1} \frac{\langle x^n, p_j \rangle_\rho}{\langle p_j, p_j \rangle_\rho} p_j(x) \tag{7.20}$$

also mit dem Orthogonalisierungsverfahren von Gram-Schmidt eine Folge paarweise orthogonaler Polynome definiert (bezüglich des Skalarproduktes $\langle \cdot, \cdot \rangle_\rho$)

Beispiel. Mit $[a, b] = [-1, 1]$ und $\rho(x) = 1$ erhält man ausgehend von $p_0(x) = 1$ mit

$$p_1(x) = x, \quad p_2(x) = x^2 - \frac{1}{3}, \quad p_3(x) = x^3 - \frac{3}{5}x, \quad p_4(x) = x^4 - \frac{5}{2}x^2 + \frac{4}{105} \tag{7.21}$$

paarweise orthogonaler Polynome bezüglich des Skalarproduktes

$$\langle p, q \rangle_\rho = \int_{-1}^{1} p(x) q(x) \mathrm{d}x$$

Die eben konstruierten orthogonalen Polynome heißen **Legendre-Polynome**.

Bemerkung 7.10. Bezeichnet man durch $\Pi_k = \mathrm{span}\{p_0, \ldots, p_k\}$ den Vektorraum der Polynome bis zum Grad k, dann gilt allgemein für die Folge paarweise orthogonaler Polynome $p_0, \ldots, p_n$ mit aufsteigendem Grad

$$p_n \in \Pi_{n-1}^{\perp}$$

Beispiel. Mit $[a, b] = [-1, 1]$ und der Gewichtsfunktion $\rho(x) = (1 - x^2)^{-\frac{1}{2}} = \frac{1}{\sqrt{1-x^2}}$ erhält man mit dem Gram-Schmidt-Verfahren (7.20) ausgehend von $p_0 = 1$ mit

$$p_0(x) = 1, \quad p_1(x) = x, \quad p_2(x) = x^2 - \frac{1}{2}, \quad p_3(x) = x^3 - \frac{3}{4}x \tag{7.22}$$

die orthogonalen **Tschebyscheff-Polynome**.

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

Satz 7.11. *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. Es seien $a < \lambda_1 < \cdots < \lambda_j < b$ $(0 \leq j \leq n)$ die Nullstellen von p_n in $]a, b[$, an denen p_n sein Vorzeichen wechselt (diese Nullstellen haben eine ungerade algebraische Vielfachheit). Es wird nun $j = n$ nachgewiesen, d.h. damit wird gezeigt, dass n einfache Nullstellen aus $]a, b[$ sind.
Für $j \leq n - 1$ hätte das Polynom

$$q(x) := \prod_{k=1}^{j}(x - \lambda_k)$$

den Grad $0 \leq j \leq n - 1$, so dass

$$\langle p_n, q \rangle_\rho = 0 \tag{7.23}$$

folgt, weil p_n nach Konstruktion orthogonal zu sämtlichen Polynomen mit Grad kleiner oder gleich $n - 1$ ist. Nach dem Fundamentalsatz der Algebra ist p_n als Produkt

$$p_n(x) = v(x)q(x)$$

darstellbar, wobei $v(x)$ auf $[a, b]$ keine Stellen enthält, wo p_n sein Vorzeichen wechselt. Damit wäre aber

$$\langle p_n, q \rangle_\rho = \int_a^b p_n(x)q(x)\rho(x)dx = \int_a^b v(x)q^2(x)\rho(x)dx \neq 0$$

was der Annahme (7.23) (bzw. $j \leq n - 1$) widerspricht. Also gilt tatsächlich $j = n$ und damit ist der Satz bewiesen. □

Nun kommen wir zur Definition der Gauß-Quadratur

Definition 7.12. *Mit $x_1, \ldots, x_n$ seien die Nullstellen des n-ten Orthogonalpolynoms $p_n(x)$ gegeben. Die numerische Integrationsformel*

$$I_n = \sum_{j=1}^{n} \sigma_j f(x_j) \quad \textit{mit} \quad \sigma_j = \langle L_j, 1 \rangle_\rho = \int_a^b L_j(x)\rho(x)\mathrm{d}x \tag{7.24}$$

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 (7.16) gerade die Nullstellen des n-ten Orthogonalpolynoms $p_n(x)$ bzw. die Gewichte gemäß (7.24) sind und damit die Gleichwertigkeit der Formeln (7.15) und (7.24) nachgewiesen.

Satz 7.13. *Mit $x_1, \ldots, x_n$ seien die Nullstellen des n-ten Orthogonalpolynoms $p_n(x)$ gegeben.*
Es existiert eine eindeutig bestimmte Gauß-Quadratur (7.24)*. Bei der Gauß-Quadratur sind alle Gewichte gemäß* (7.24) *positiv und die Quadratur ist für jedes Polynom vom Grad $m \leq 2n-1$ exakt, d.h. es gilt*

$$\int_a^b p(x)\rho(x)\mathrm{d}x = \langle p, 1\rangle_\rho = \sum_{j=1}^{n} \sigma_j p(x_j), \quad \forall p \in \Pi_{2n-1} \tag{7.25}$$

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 $q, r \in \Pi_{n-1}$ mit

$$\frac{p}{p_n} = q + \frac{r}{p_n} \Leftrightarrow p = qp_n + r \tag{7.26}$$

Mit den Nullstellen $x_1, \ldots, x_n$ von p_n gilt $p(x_j) = r(x_j)$ für $j = 1, \ldots, n$. Das Lagrangesche 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 $\langle q, p_n\rangle_\rho = 0$ ergibt die skalare Multiplikation der Darstellung (7.26) von p mit 1

$$\begin{aligned} \int_a^b p(x)\rho(x)\mathrm{d}x &= \langle p, 1\rangle_\rho = \langle r, 1\rangle_\rho \\ &= \sum_{j=1}^{n} p(x_j) \langle L_j, 1\rangle_\rho = \sum_{j=1}^{n} \sigma_j p(x_j)\ . \end{aligned} \tag{7.27}$$

Für $p(x) = L_j^2(x) \in \Pi_{2n-2}$ ergibt die eben nachgewiesene Formel (7.27)

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

Wegen $L_j^2(x_k) = \delta_{jk}^2$ folgt die Positivität der Gewichte.

Zum Nachweis der Eindeutigkeit der Gauß-Quadratur nimmt man an, dass eine weitere Formel

$$I_n^* = \sum_{j=1}^{n} \sigma_j^* f(x_j^*) \tag{7.28}$$

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 (7.28) den exakten Wert des Integrals für $h(x)$, also

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

für alle $k = 1, \ldots, n$. Da das 2. Integral $\int_a^b L_k^*(x) p_n(x)\rho(x)\mathrm{d}x = \langle L_k^*, p_n \rangle_\rho$ wegen der Orthogonalität von p_n zu allen Polynomen bis zum Grad $n-1$ gleich Null ist, folgt $\sigma_k^* p_n(x_k^*) = 0$ für alle $k = 1, \ldots, n$. Wegen der Positivität der Gewichte müssen die x_k^* Nullstellen des n-ten Orthogonalpolynoms $p_n(x)$ sein, die eindeutig bestimmt sind. Damit ist die Eindeutigkeit der Gauß-Quadratur bewiesen. □

Auf der Grundlage des Fehlers der Polynominterpolation von $f(x)$ durch ein Polynom n-ten Grades kann man den Fehler der Gauß-Quadratur bestimmen, es gilt der

Satz 7.14. *Mit den Stützstellen und Gewichten aus Satz 7.13 gilt für auf dem Intervall $[a, b]$ $2n$-mal stetig diffbare Funktionen $f(x)$*

$$\int_a^b f(x)\rho(x)\mathrm{d}x - \sum_{j=1}^{n} \sigma_j f(x_j) = \frac{\|p_n\|_\rho^2}{(2n)!} f^{(2n)}(\xi) \tag{7.29}$$

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

Die folgende Tabelle zeigt Intervalle, Gewichtsfunktionen, die zugehörigen Orthogonalpolynome und deren Name ($\alpha, \beta > -1$)

20. Vorlesung am 12.01.2015

Intervall	$\rho(x)$	$p_0, p_1, \ldots$	Bezeichnung
$[-1,1]$	1	$1, x, x^2 - \frac{1}{3}, \ldots$	Legendre
$[-1,1]$	$\frac{1}{\sqrt{1-x^2}}$	$1, x, x^2 - \frac{1}{2}, \ldots$	Tschebyscheff
$[-1,1]$	$(1-x)^\alpha(1+x)^\beta$	$1, \frac{1}{2}[\alpha - \beta + (\alpha + \beta + 2)x]$	Jacobi
$]-\infty, \infty[$	e^{-x^2}	$1, x, x^2 - \frac{1}{2}, x^3 - \frac{3}{2}x, \ldots$	Hermite
$[0, \infty[$	$e^{-x}x^\alpha$	$1, x - \alpha - 1, \ldots$	Laguerre

Mit den in der Tabelle angegebenen Polynomen und deren Nullstellen lassen sich Quadraturformeln für endliche Intervalle und unendliche Intervall konstruieren.
Die Tschebyscheffpolynome sind trotz der Gewichtsfunktion gegenüber den Legendrepolynomen attraktiv, weil man die Nullstellen des n-ten Tschebyscheffschen Orthogonalpolynoms explizit angeben kann (durch eine Berechnungsformel, s.dazu (6.9)) ohne die Polynome auszurechnen. Das ist bei den anderen Polynomen aus der Tabelle nicht direkt möglich.

Beispiel 7.15. Zur Berechnung von uneigentlichen Integralen sind die orthogonalen Hermite-Polynome von Interesse und mit dem Skalarprodukt

$$\langle p, q\rangle = \int_{-\infty}^{\infty} p(x)q(x)e^{-x^2}\,dx$$

findet man ausgehend von der Monombasis durch Gram-Schmidt-Orthogonalisierung die ersten Orthogonalpolynome

$$\begin{aligned} p_0 &= 1 \\ p_1 &= x - \frac{\langle x, 1\rangle}{\langle 1, 1\rangle}1 = x \\ p_2 &= x^2 - \frac{\langle x^2, 1\rangle}{\langle 1, 1\rangle}1 - \frac{\langle x^2, x\rangle}{\langle x, x\rangle}x = x^2 - \frac{1}{2} \\ &\ldots \end{aligned}$$

wobei wir ausgenutzt haben, dass $x\,e^{-x^2}$ bzw. $x^3\,e^{-x^2}$ ungerade Funktionen sind, und dass man mit partieller Integration

$$\frac{\langle x^2, 1\rangle}{\langle 1, 1\rangle}1 = \frac{1}{2}$$

erhält. Die Nullstellen des 2. Orthogonalpolynoms sind damit $x_1 = -\frac{1}{\sqrt{2}}, x_2 = \frac{1}{\sqrt{2}}$. Mit

$$L_1 = \frac{x - x_2}{x_1 - x_2} = \frac{x - \frac{1}{\sqrt{2}}}{-\frac{2}{\sqrt{2}}}$$

erhält man für die Gewichte

$$\begin{aligned}\sigma_1 &= \langle L_1, 1\rangle = \int_{-\infty}^{\infty} \frac{x-\frac{1}{\sqrt{2}}}{-\frac{2}{\sqrt{2}}} e^{-x^2} dx \\ &= \frac{\sqrt{2}}{2}\frac{1}{\sqrt{2}}\int_{-\infty}^{\infty} e^{-x^2} dx = \frac{\sqrt{\pi}}{2} = \sigma_2 \ .\end{aligned}$$

Damit ist

$$I_2(f) = \frac{\sqrt{\pi}}{2}[f(\frac{1}{\sqrt{2}}) + f(-\frac{1}{\sqrt{2}})]$$

eine Quadratur mit der Genauigkeit $m = 3$ zur Berechnung des Integrals

$$\int_{-\infty}^{\infty} f(x)e^{-x^2}\, dx \ .$$

7.6 Numerische Integration durch Extrapolation

Die summierte Trapezregel zur näherungsweisen Berechnung des Integrals $\int_a^b f(x)\, dx$ kann man bei der Verwendung von N Teilintervallen in der Form

$$T(h) := S_{1,N} = h[\frac{1}{2}(f(a) + f(b)) + \sum_{i=1}^{N-1} f(a + i\, h)]$$

mit $h = \frac{b-a}{N}$ aufschreiben.
Die Grundidee der numerischen Integration durch Extrapolation besteht in der Nutzung der Werte der Trapezsumme $T(h)$ für unterschiedliche Schrittweiten h_0, h_1, um durch Extrapolation auf $h = 0$ zu schließen.
Die entscheidende mathematische Grundlage hierfür ist der

Satz 7.16. *Für eine Funktion $f \in C^{2m+2}[a, b]$ besitzt $T(h)$ die Entwicklung*

$$T(h) = \tau_0 + \tau_1 h^2 + \tau_2 (h^2)^2 + \cdots + \tau_m (h^2)^m + R_{m+1}(h) \tag{7.30}$$

mit

$$\begin{aligned}\tau_0 &= \int_a^b f(x)\, dx \ , \\ \tau_k &= \frac{B_{2k}}{(2k)!}[f^{(2k-1)}(b) - f^{(2k-1)}(a)]\end{aligned}$$

(B_{2k} sind die Bernoullischen Zahlen, die unabhängig von h sind, und damit sind auch die τ_k unabhängig von h) und dem Restglied R_{m+1}

$$R_{m+1}(h) = \mathcal{O}(h^{2m+2}) \ .$$

Beweis. Für Interessenten s. Plato □

Es ist offensichtlich, dass nach dem Satz 7.16

$$\int_a^b f(x)\,dx = \tau_0 = \lim_{h \searrow 0} T(h)$$

gilt.
Schreibt man nun die asymptotische Entwicklung (7.30) z.B. für 3 Schrittweiten auf, dann erhält man

$$\begin{array}{rcl} T(h_0) & \approx & \tau_0 + \tau_1 h_0^2 + \tau_2 h_0^4 \\ T(h_1) & \approx & \tau_0 + \tau_1 h_1^2 + \tau_2 h_1^4 \\ T(h_2) & \approx & \tau_0 + \tau_1 h_2^2 + \tau_2 h_2^4 \end{array} \tag{7.31}$$

und kann bei Kenntnis der Trapezsummen $T(h_0), T(h_1)$ und $T(h_2)$ daraus τ_0 näherungsweise ermitteln. Bei den Beziehungen (7.31) macht man aufgrund von Satz 7.16 nur Fehler der Ordnung $\mathcal{O}(h^6)$.
Eine andere Interpretation dieser Extrapolationsidee besteht in der Nutzung der Wertepaare

$$(h_0^2, T(h_0)),\ (h_1^2, T(h_1)),\ (h_2^2, T(h_2))$$

zur Bestimmung des Interpolationspolynoms zweiten Grades in $\tilde{h} = h^2$, also

$$p_2(\tilde{h}) = p_2(h^2) = \tilde{\tau}_0 + \tilde{\tau}_1 h^2 + \tilde{\tau}_2 (h^2)^2$$

mit der Eigenschaft

$$p_2(h_k^2) = T(h_k)\ ,\ k = 0, 1, 2\ .$$

Die Auswertung dieses Polynoms an der Stelle $h^2 = 0$ (Extrapolation, s.auch Abb. 7.1) liefert dann die Näherung

$$\int_a^b f(x)\,dx = \tau_0 \approx p_2(0) = \tilde{\tau}_0\ .$$

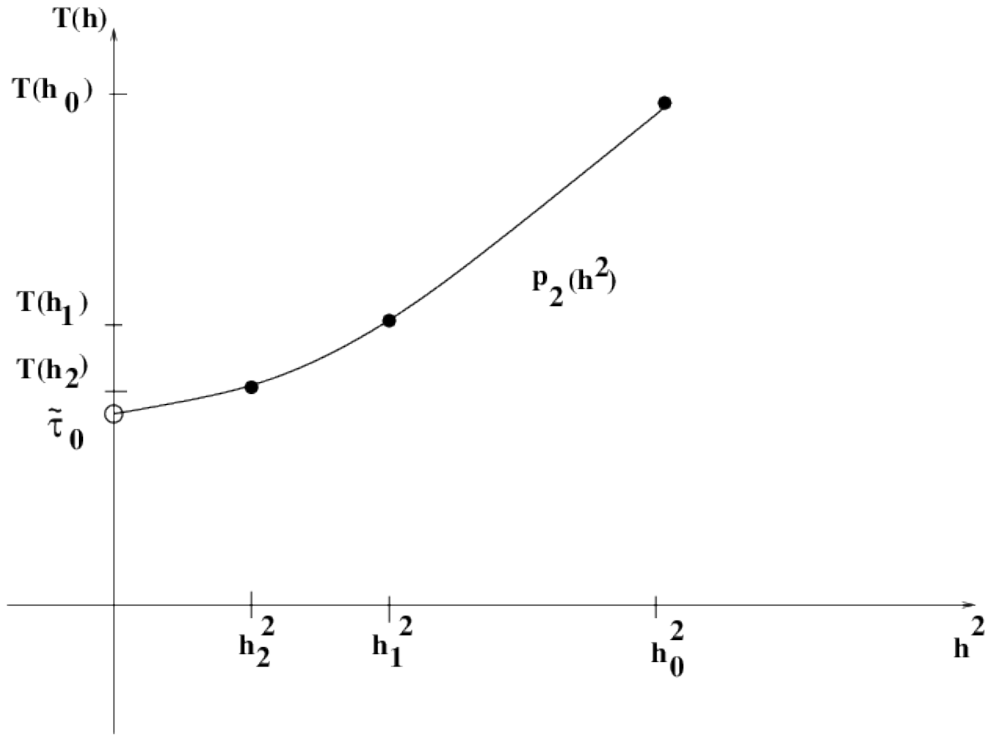

Abbildung 7.1: Polynom p_2 und Stützwerte $(h_k^2, T(h_k)), k = 0, 1, 2$

7.7 Anwendung des Schemas von Neville-Aitken - Romberg-Verfahren

In Anlehnung an die Entwicklung (7.30) sucht man also ein Interpolationspolynoms p_m an den Stützstellen h_k^2 mit den Funktionswerten $T(h_k)$ $(k = 0, \ldots, m)$ und möchte dann den Wert des Interpolationspolynoms p_m an der Stelle $\tilde{h} = h^2 = 0$ ausrechnen, dann bietet sich das Schema von Neville-Aitken zur Polynomwertberechnung für das Interpolationspolynom für die Wertepaare $(x_k, f(x_k))$, $k = 0, 1, \ldots, m$, an, also

x_i	$T_{i,0} = f(x_i)$	$T_{i,1}$	$T_{i,2}$	$\ldots$	$T_{i,m-1}$	$T_{i,m}$
x_0	$T_{0,0} = f(x_0)$					
x_1	$T_{1,0} = f(x_1)$	$T_{1,1}$				
x_2	$T_{2,0} = f(x_2)$	$T_{2,1}$	$T_{2,2}$			
$\vdots$						
x_m	$T_{m,0} = f(x_m)$	$T_{m,1}$	$T_{m,2}$	$\ldots$	$T_{m,m-1}$	$T_{m,m}$

mit $m \geq i \geq k \geq 1$ und

$$\begin{aligned} T_{i,0} &= f(x_i) \\ T_{i,k}(x) &= \frac{(x - x_{i-k})T_{i,k-1}(x) - (x - x_i)T_{i-1,k-1}(x)}{x_i - x_{i-k}}\,, \quad k \geq 1, \end{aligned}$$

woraus

$$\begin{aligned} T_{i,k} &= \frac{(x - x_i)T_{i,k-1} + (x_i - x_{i-k})T_{i,k-1} - (x - x_i)T_{i-1,k-1}}{x_i - x_{i-k}} \\ &= T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{\frac{x - x_{i-k}}{x - x_i} - 1} \end{aligned}$$

folgt (das feste Argument x wurde hier der Übersichtlichkeit halber weg gelassen).
Beim Romberg-Verfahren geht man von der Entwicklung (7.30) als Polynom von $T(h)$ in h^2 aus, und d.h., man muss $x_i = h_i^2$ setzen. Für die Berechnung des Wertes von p_m an der Stelle $\tilde{h} = h^2 = 0$ ergibt das obige Neville-Aitken-Schema

$$\begin{aligned} T_{i,0} &= T(h_i) \\ T_{i,k} &= T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{(\frac{h_{i-k}}{h_i})^2 - 1} \end{aligned}$$

mit $T_{m,m}$ den Näherungswert für $\tau_0 = \int_a^b f(x)\,dx$.
Für $m = 1$ erhält man mit $h_0 = b - a,\ h_1 = (b - a)/2$

$$\begin{aligned} T_{1,1} &= T_{1,0} + \frac{T_{1,0} - T_{0,0}}{(\frac{h_0}{h_1})^2 - 1} = \frac{4}{3}T_{1,0} - \frac{1}{3}T_{0,0} \\ &= \frac{b-a}{3}[f(a) + 2f(\frac{a+b}{2}) + f(b)] - \frac{b-a}{6}[f(a) + f(b)] \\ &= \frac{b-a}{6}[f(a) + 4f(\frac{a+b}{2}) + f(b)]\,, \end{aligned}$$

also die Simpsonregel. Für $h_i = \frac{b-a}{3^i},\ i = 0, 1$ erhält man mit

$$T_{1,1} = \frac{b-a}{8}[f(a) + 3f(\frac{2a+b}{3}) + 3f(\frac{a+2b}{3}) + f(b)]$$

die Newtonsche 3/8-Regel.
Als gängige Folgen $h_i,\ i = 0, \ldots$ werden die **Romberg**-Folge

$$h_0 = b - a, \quad h_1 = \frac{h_0}{2}, \quad h_2 = \frac{h_1}{2}, \quad h_3 = \frac{h_2}{2}, \ldots$$

oder die **Bulirsch**-Folge

$$h_0 = b - a, \quad h_1 = \frac{h_0}{2}, \quad h_2 = \frac{h_0}{3}, \quad h_3 = \frac{h_1}{2}, \quad h_4 = \frac{h_2}{2}, \ldots$$

verwendet. Zur Romberg-Folge ist noch anzumerken, dass man $T(h_{i+1})$ rekursiv aus $T(h_i)$ durch die Formel

$$T(h_{i+1}) = T(\frac{1}{2}h_i) = \frac{1}{2}T(h_i) + h_{i+1}[f(a+h_{i+1}) + f(a+3h_{i+1}) + \cdots + f(b-h_{i+1})]$$

bestimmen kann.

Beispiel 7.17. Wir wollen das Schema der $T_{i,k}$ mal für die Folge $h_0 = b - a$, $h_1 = h_0/2$, $h_2 = h_1/2$ mal aufschreiben, und erhalten

i	h_i	$T_{i,0}$	$T_{1+i,1}$	$T_{2+i,2}$
0	h_0	$T_{0,0}$		
1	h_1	$T_{1,0}$	$T_{1,1}$	
2	h_2	$T_{2,0}$	$T_{2,1}$	$T_{2,2}$

und wenn wir z.B. das Integral $\int_1^2 \frac{1}{x}\,dx = \ln 2$ annähern wollen, erhalten wir das Schema

i	h_i	$T_{i,0}$	$T_{1+i,1}$	$T_{2+i,2}$
0	1	0.750000000000000		
1	1/2	0.708333333333333	0.694444444444444	
2	1/4	0.697023809523809	0.693253968253968	0.693174603174603

und können mit $T_{2,2} = 0.693174603174603$ die Näherung für $\ln 2$ mit einem Fehler der Ordnung $O(h^6)$ ablesen.
Die Ergebnisse habe ich mit einem Octave/Matlab-Programm ausgerechnet (s. dazu auch das Verzeichnis).

Kapitel 8

Numerische Lösung von Anfangswertaufgaben

21. Vorlesung am 14.01.2015

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

Definition 8.1. *Ein **Anfangswertproblem** für ein System von n gewöhnlichen Differentialgleichungen 1. Ordnung ist von der Form*

$$y' = f(t, y), \quad t \in [a, b] \tag{8.1}$$
$$y(a) = y_0 \tag{8.2}$$

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

$$f : [a, b] \times \mathbb{R}^n \to \mathbb{R}^n \tag{8.3}$$

wobei eine differenzierbare Abbildung $y : [a, b] \to \mathbb{R}^n$ mit den Eigenschaften (8.1) - (8.3) *als **Lösung des Anfangswertproblems** gesucht ist.*

Aussagen zur Existenz und Eindeutigkeit der Lösung liefert

Satz 8.2. *Erfüllt f aus* (8.3) *die Bedingung*

$$\|f(t, u) - f(t, v)\| \leq L \|u - v\|, \quad t \in [a, b],\ u, v \in \mathbb{R}^n \tag{8.4}$$

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

(a) Das AWP (8.1),(8.2) *besitzt genau eine stetig diff'bare Lösung $y : [a, b] \to \mathbb{R}^n$ (Picard-Lindelöf)*

(b) Für differenzierbare Funktionen $y, \hat{y} : [a,b] \to \mathbb{R}^n$ mit

$$\begin{aligned} y' &= f(t,y), \quad t \in [a,b]; \quad y(a) = y_0 \\ \hat{y}' &= f(t,\hat{y}), \quad t \in [a,b]; \quad \hat{y}(a) = \hat{y}_0 \end{aligned}$$

gilt die Abschätzung

$$\|y(t) - \hat{y}(t)\| \leq e^{L(t-a)} \|y_0 - \hat{y}_0\|, \quad t \in [a,b] \tag{8.5}$$

Beweis. Vorlesung DGL oder Analysis □

Bemerkung.

(1) Mit den Aussagen des Satzes 8.2 hat man die Existenz und Eindeutigkeit der Lösung und die stetige Abhängigkeit der Lösung von den Anfangsdaten unter der Voraussetzung der Lipschitzstetigkeit von $f(t,\cdot)$ vorzuliegen.

(2) Im Folgenden sollen numerische Lösungsverfahren entwickelt werden, wobei wir ohne die Allgemeinheit einzuschränken den Fall $n = 1$ betrachten. Die besprochenen Verfahren gelten allerdings auch im allgemeinen Fall $n > 1$

Definition 8.3. *Unter dem Richtungsfeld der Differentialgleichung*

$$y' = f(t,y)$$

versteht man das Vektorfeld

$$r(t,y) = \begin{pmatrix} \frac{1}{\sqrt{1+f^2(t,y)}} \\ \frac{f(t,y)}{\sqrt{1+f^2(t,y)}} \end{pmatrix}$$

d.h. das Vektorfeld der normierten Steigungen

Betrachtet man einen beliebigen Punkt (t_0, y_0) der (t,y)- Ebene, kann man Lösungskurven $y(t)$ durch diesen Punkt annähren:

Beispiel.

$$y' = y^2 + t^2, \quad r(t,y) = \begin{pmatrix} \frac{1}{\sqrt{1+(y^2+t^2)^2}} \\ \frac{y^2+t^2}{\sqrt{1+(y^2+t^2)^2}} \end{pmatrix}$$

(I) $y'(t_0) = y_0^2 + t_0^2$, ($t_0 = a$ entspricht Start in Anfangspunkt (a, y_0))
t-Achse wird durch $t_k = t_0 + hk$ äquidistant unterteilt

(II) mit dem Schritt von Punkt

$$(t_0, y_0) \quad \text{zu} \quad (t_0 + h, y_0 + hy'(t_0)) =: (t_1, y_1)$$

bzw. allgemein vom Punkt

$$(t_k, y_k) \quad \text{zu} \quad (t_k + h, y_k + hf(t_k, y_k)) =: (t_{k+1}, y_{k+1})$$

erhält man mit $h = \frac{b-a}{N}$ nach m Schritten mit

$$y_0, y_1, \ldots, y_N$$

unter "günstigen" Umständen eine Approximation der Lösung $y(t)$ an den Stellen

$$a = t_0, t_1, \ldots, t_N = b$$

(III) D.h. man fährt das Richtungsfeld geeignet ab, um eine numerische Lösung $y_k, k = 0, 1, \ldots, N$ zu erhalten

8.1 Theorie der Einschrittverfahren

Definition 8.4. *Ein Einschrittverfahren zur näherungsweisen Bestimmung einer Lösung des AWP* (8.1),(8.2) *hat die Form*

$$y_{k+1} = y_k + h_k \Phi(t_k, y_k, y_{k+1}, h_k), \quad k = 0, 1, \ldots, N-1 \tag{8.6}$$

mit einer Verfahrensfunktion

$$\Phi : [a, b] \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$$

und einem (noch nicht näher spezifizierten) Gitter bzw. Schrittweiten

$$\Delta = \{a = t_0 < t_1 < \ldots < t_N \leq b\}, \; h_k := t_{k+1} - t_k, \quad k = 0, 1, \ldots, N-1 \tag{8.7}$$

Bemerkung. Hängt die Verfahrensfunktion *nicht* von y_{k+1} ab, ist die Berechnungsvorschrift (8.6) eine explizite Formel zur Berechnung von y_{k+1} und man spricht von einem expliziten Einschrittverfahren.

Zur Klassifizierung und Bewertung von numerischen Lösungsverfahren für AWP benötigen wir im Folgenden einige Begriffe ($y(t)$ bezeichnet hier die exakte Lösung).

Definition 8.5. *Unter dem* ***lokalen Diskretisierungsfehler*** *an der Stelle* t_{k+1} *des Verfahrens* (8.6) *versteht man den Wert*

$$d_{k+1} := y(t_{k+1}) - y(t_k) - h_k \Phi(t_k, y(t_k), y(t_{k+1}), h_k) \tag{8.8}$$

Definition 8.6. *Unter dem globalen Diskretisierungsfehler* g_k *an der Stelle* t_k *versteht man den Wert*

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

Definition 8.7. *Ein Einschrittverfahren* (8.6) *besitzt die Fehlerordnung* p*, falls für seinen lokalen Diskretisierungsfehler* d_k *die Abschätzungen*

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

$$\max_{1 \leq k \leq N} |d_k| \leq D = \text{const.} h_{\max}^{p+1} = \mathcal{O}(h_{\max}^{p+1}) \tag{8.9}$$

mit $h_{\max} = \max_{k=0,\ldots,N-1} t_{k+1} - t_k$ *gilt. (Statt Fehlerordnung verwendet man auch den Begriff Konsistenzordnung.) Ist* $p \geq 1$*, dann heißt das Verfahren konsistent.*

Die Bedingungen

$$\begin{aligned} |\Phi(t, u_1, u_2, h) - \Phi(t, v_1, u_2, h)| &\leq L_1 |u_1 - v_1| \\ |\Phi(t, u_1, u_2, h) - \Phi(t, u_1, v_2, h)| &\leq L_2 |u_2 - v_2| \end{aligned} \tag{8.10}$$

für $t \in [a, b], 0 < h \leq b - t, u_j, v_j \in \mathbb{R}$, mit positiven konstanten L_1, L_2 sind für die folgenden Konvergenzuntersuchungen von Einschrittverfahren von Bedeutung

Satz 8.8. *Ein Einschrittverfahren* (8.6) *zur Lösung des AWP* (8.1), (8.2) *besitze die Konsistenzordnung* $p \geq 1$ *und die Verfahrensfunktion erfülle die Bedinung* (8.10)*. Dann liegt die Konvergenzordnung* p *vor, d.h. es gilt*

$$\max_{k=0,\ldots,N} |y_k - y(t_k)| \leq K h_{\max}^p$$

Mit einer Konstanten K*, die vom Intervall* $[a, b]$*, Konstanten* C *aus der Abschätzung* (8.9) *und* L_1, L_1 *aus* (8.10) *herrührt.*

Bewiesen werden soll der Satz 8.8 für ein explizites Einschrittverfahren (Beweis für allgemeines Einschrittverfahren in der Vorlesung).
Benötigt wird das

Lemma 8.9. *Für Zahlen $L > 0, a_k \geq 0, h_k \geq 0$ und $b \geq 0$ sei*

$$a_{k+1} \leq (1 + h_k L)a_k + h_k b, \quad k = 0, 1, \ldots, N-1$$

erfüllt. Dann gelten die Abschätzungen

$$a_k \leq \frac{e^{Lt_k} - 1}{L} b + e^{Lt_k} a_0 \quad \textit{mit} \quad t_k := \sum_{j=0}^{k-1} h_j \quad (k = 0, \ldots, N)$$

Beweis. (vollständige Induktion)
Induktionsanfang ist für $k = 0$ offensichtlich gewährleistet. Dr Schritt $k \to k+1$ ergibt sich wie folgt:

$$\begin{aligned} a_{k+1} &\leq (1 + h_k L) \left(\frac{e^{Lt_k} - 1}{L} b + e^{Lt_k} a_0 \right) + h_k b \\ &\leq \left(\frac{e^{L(t_k + h_k)} - 1 - h_k L}{L} + h_k \right) b + e^{L(t_k + h_k)} a_0 \\ &= \frac{e^{Lt_{k+1}} - 1}{L} b + e^{Lt_{k+1}} a_0 \end{aligned}$$

(es wurde $1 + t \leq e^t$ benutzt). □

Beweis von Satz 8.8. Mit den Festlegungen

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

gilt für $k = 0, 1, \ldots, N-1$

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

und damit

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

bzw.

$$\begin{aligned} |e_{k+1}| &\leq |e_k| + h_k \left| \Phi(t_k, y_k, h_k) - \Phi(t_k, y(t_k), h_k) \right| + |d_{k+1}| \\ &\leq (1 + h_k L_1) |e_k| + h_k C h_{\max}^p \end{aligned}$$

Die Abschätzung des Lemmas 8.9 liefert wegen $e_0 = 0$ die Behauptung des Satzes 8.8 □

8.2 Spezielle Einschrittverfahren

22. Vorlesung 19.01.2015

8.2.1 Euler-Verfahren

Mit der Verfahrensfunktion

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

erhält man mit

$$y_{k+1} = y_k + h_k f(t_k, y_k), \quad k = 0, \ldots, N-1 \tag{8.11}$$

das Euler-Verfahren.
Für eine stetig partiell diff'bare Funktion $f : [a, b] \times \mathbb{R} \to \mathbb{R}$ besitzt das Euler-Verfahren die Konsistenzordnung $p = 1$, denn mit der Taylorentwicklung

$$y(t + h) = y(t) + y'(t)h + \frac{h^2}{2} y''(\xi), \quad \xi \in [a, b]$$

erhält man

$$d_{k+1} = y(t_{k+1}) - y(t_k) - h_k f(t_k, y(t_k)) = \frac{h_k^2}{2} y''(\xi)$$

bzw.

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

8.2.2 Einschrittverfahren der Konsistenzordnung $p = 2$

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

$$\Phi(t, y, h) = a_1 f(t, y) + a_2 f(t + b_1 h, y + b_2 h f(t, y)),\ t \in [a, b],\ h \in [0, b - t],\ y \in \mathbb{R} \tag{8.12}$$

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

Satz 8.10. *Ein Einschrittverfahren* (8.6) *mit einer Verfahrensfunktion der Form* (8.12) *ist konsistent mit der Ordnung $p = 2$, falls $f : [a, b] \times \mathbb{R} \to \mathbb{R}$ zweimal stetig partiell diff'bar ist und für die Koeffizienten*

$$a_1 + a_2 = 1, \quad a_2 b_1 = \frac{1}{2}, \quad a_2 b_2 = \frac{1}{2} \tag{8.13}$$

gilt.

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

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

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

(hier wurde die Differentialgleichung und deren Ableitung benutzt) und damit folgt

$$d_{k+1} = y(t_{k+1}) - y(t_k) - h_k\Phi(t_k, y(t_k), h_k) = \mathcal{O}(h_k^3)$$

also $p = 2$ □

Mit der konkreten Wahl $a_1 = 0, a_2 = 1, b_1 = b_2 = \frac{1}{2}$ erhält man mit

$$y_{k+1} = y_k + h_k f\left(t_k + \frac{h_k}{2}, y_k + \frac{h_k}{2} f(t_k, y_k)\right), \quad k = 0, \ldots, N-1 \tag{8.14}$$

das **modifizierte Euler-Verfahren** (verbesserte Polygonzugmethode) mit der Konsistenzordnung $p = 2$
Mit der Wahl $a_1 = a_2 = \frac{1}{2}, b_1 = b_2 = 1$ erhält man mit

$$y_{k+1} = y_k + \frac{h_k}{2}\left[f(t_k, y_k) + f(t_k + h_k, y_k + h_k f(t_k, y_k))\right], \quad k = 0, \ldots, N-1 \tag{8.15}$$

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

8.3 Verfahren höherer Ordnung

23. Vorlesung 21.01.2015

Die bisher besprochenen Methoden (Euler, Heun) haben wir weitestgehend intuitiv ermittelt. Um systematisch Einschrittverfahren höherer Ordnung zu konstruieren, betrachten wir die zum AWP $y' = f(t, y), y(a) = y_0$ äquivalente Gleichung (nach Integration)

$$y(t) = y_0 + \int_a^t f(s, y(s))\mathrm{d}s \tag{8.16}$$

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

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(s, y(s))\mathrm{d}s \tag{8.17}$$

Das letzte Integral aus (8.17) approximieren wir durch eine Quadraturformel

$$\int_{t_k}^{t_{k+1}} f(s, y(s))\mathrm{d}s \tag{8.18}$$

wobei die s_l zu einer Zerlegung von $[t_k, t_{k+1}]$ gehören. (8.17) und (8.18) ergeben

$$y(t_{k+1}) \approx y(t_k) + h_k \sum_{l=1}^{m} \gamma_l f(s_l, y(s_l)) \tag{8.19}$$

wobei wir die Werte $y(s_l)$ nicht kennen. Sie müssen näherungsweise aus $y(t_k)$ bestimmt werden, damit (8.19) als Integrationsverfahren benutzt werden kann.
Wählt man z.B. $m = 2$ und $\gamma_1 = \gamma_2 = \frac{1}{2}$ sowie $s_1 = t_k$ und $s_2 = t_{k+1}$, dann bedeutet (8.19)

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

und mit der Approximation

$$y(t_{k+1}) \approx y(t_k) + h_k f(t_k, y(t_k))$$

ergibt sich mit

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

die Grundlage für das Verfahren von Heun.

Im Weiteren wollen wir mit y_k die Verfahrenswerte zur Näherung der exakten Werte $y(t_k)$ bezeichnen und als Näherungen von $f(s_l, y(s_l))$

$$f(s_l, y(s_l)) \approx k_l(t_j, y_j)$$

verwenden. Mit

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

werden die k_l rekursiv definiert:

$$\begin{aligned}
k_1(t_k, y_k) &= f(t_k, y_k) \\
k_2(t_k, y_k) &= f(t_k + \alpha_2 h_k, y_k + h_k \beta_{21} k_1(t_k, y_k)) \\
k_3(t_k, y_k) &= f(t_k + \alpha_3 h_k, y_k + h_k(\beta_{31} k_1 + \beta_{32} k_2)) \\
&\vdots \\
k_m(t_k, y_k) &= f(t_k + \alpha_m h_k, y_k + h_k(\beta_{m1} k_1 + \cdots + \beta_{mm-1} k_{m-1}))
\end{aligned} \tag{8.20}$$

Ausgehend von (8.19) und (8.20) wird durch

$$y_{k+1} = y_k + h_k(\gamma_1 k_1(t_k, y_k) + \cdots + \gamma_m k_m(t_k, y_k)) \tag{8.21}$$

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

Definition 8.11. *Das Verfahren* (8.21) *heißt* m***-stufiges Runge-Kutta-Verfahren*** *mit* k_l *aus* (8.20) *und die* k_l *heißen Stufenwerte.*

Bemerkung. Wir haben oben schon festgestellt, dass im Fall $m = 2$ mit $\gamma_1 = \gamma_2 = \frac{1}{2}, \alpha_2 = 1, \beta_{21} = 1$ (8.21) gerade das Heun-Verfahren ergibt, also ein Verfahren mit der Konsistenzordnung $p = 2$. Wir werden nun Bedingungen für die freien Parameter im Verfahren (8.21) formulieren, sodass einmal ein konsistentes Verfahren ($p \geq 1$) entsteht und andererseits eine möglichst große Konsistenzordnung erhalten wird.

Aus der Verwendung der Quadraturformel

$$h_k \sum_{l=1}^{m} \gamma_l f(s_l, y(s_l)) \approx \int_{t_k}^{t_{k+1}} f(s, y(s)) \mathrm{d}s$$

folgt die sinnvolle Forderung

$$1 = \gamma_1 + \gamma_2 + \cdots + \gamma_m \tag{8.22}$$

also haben die γ_l die Funktion von Gewichten.
Fordert man vom Verfahren (8.21), dass die Dgl $y' = 1$ (y linear) exakt integriert wird, ergibt sich die Bedingung

$$\alpha_l = \beta_{l1} + \cdots + \beta_{ll-1} \tag{8.23}$$

Es ist nämlich $f(t, y) \equiv 1$ und damit $k_l \equiv 1$ für alle l. Ausgangspunkt war

$$k_l(t_k, y_k) \approx f(s_l, y(s_l))$$

und

$$k_l \approx f(t_k + \alpha_l h_k, y(t_k) + h_k(\beta_{l1}k_1 + \cdots + \beta_{ll-1}k_{l-1})) \ .$$

Also steht das y-Argument für $y(s_l) = y(t_k + \alpha_l h_k)$. Wir fordern, dass dies bei $f \equiv 1$ exakt ist, also

$$y(s_l) = y(t_k) + h_k(\beta_{l1} + \cdots + \beta_{ll-1}) \tag{8.24}$$

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

$$y(s_l) = y(t_k) + \alpha_l h_k \tag{8.25}$$

und aus dem Vergleich von (8.24),(8.25) folgt

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

Definition 8.12. *Die Tabelle mit den Koeffizienten $\alpha_l, \beta_{lr}, \gamma_r$ in der Form*

$$\begin{array}{c|ccccc} 0 & & & & & \\ \alpha_2 & \beta_{21} & & & & \\ \alpha_3 & \beta_{31} & \beta_{32} & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ \alpha_m & \beta_{m1} & \beta_{m2} & \cdots & \beta_{mm-1} & \\ \hline & \gamma_1 & \gamma_2 & \cdots & \gamma_{m-1} & \gamma_m \end{array} \tag{8.26}$$

heißt ***Butcher-Tabelle*** *und beschreibt das Verfahren* (8.21). *α_1 ist hier gleich* 0*, weil explizite Verfahren betrachtet werden.*

Satz 8.13. *Ein explizites Runge-Kutta-Verfahren* (8.21)*, dessen Koeffizienten die Bedingungen* (8.22) *und* (8.23) *erfüllen, ist konsistent.*

Beweis. Es ist zu zeigen, dass der lokale Diskretisierungsfehler die Ordnung $\mathcal{O}(h_k^{p+1})$ mit $p \geq 1$ hat. Wir setzen $h_k =: h$, da k jetzt fixiert ist.

$$\begin{aligned}
|d_{k+1}| &= |y(t_{k+1}) - y(t_k) - h\Phi(t_k, y(t_k), h)| \\
&= \left| y(t_{k+1}) - y(t_k) - h\sum_{r=1}^{m} \gamma_r k_r(t_k, y(t_k)) \right| \\
&\overset{(8.22)}{=} \left| y(t_{k+1}) - y(t_k) - hf(t_k, y(t_k)) - h\sum_{r=1}^{m} \gamma_r (k_r(t_k, y(t_k)) - f(t_k, y(t_k))) \right| \\
&\leq \underbrace{|y(t_{k+1}) - y(t_k) - hy'(t_k)|}_{\in \mathcal{O}(h^2)} + h \left| \sum_{r=1}^{m} \gamma_r \underbrace{(k_r(t_k, y(t_k)) - f(t_k, y(t_k)))}_{\in \mathcal{O}(h)} \right|
\end{aligned}$$

also

$$|d_{k+1}| \leq Ch^2$$

□

24. Vorlesung 26.01.2015

Bemerkung. Butcher hat bewiesen, wie groß die maximale Ordnung ist, welche mit einem m-stufigen Runge-Kutta-Verfahren erreichbar ist, was in der folgenden Tabelle notiert ist:

m	1	2	3	4	5	6	7	8	9	für $m \geq 9$
p	1	2	3	4	4	5	6	6	7	$p < m - 2$

8.4 Einige konkrete Runge-Kutta-Verfahren und deren Butcher-Tabellen

(i) Euler-Verfahren

$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array} \quad m = 1, \gamma_1 = 1$$

$$y_{k+1} = y_k + h_k f(t_k, y_k), \quad p = 1$$

(ii) Modifiziertes Euler-Verfahren

$$\begin{array}{c|cc} 0 & & \\ \frac{1}{2} & \frac{1}{2} & \\ \hline & 0 & 1 \end{array} \quad m = 2, \gamma_1 = 0, \gamma_2 = 1, \alpha_2 = \frac{1}{2}, \beta_{21} = \frac{1}{2}$$

$$
\begin{aligned}
k_1 &= f(t_k, y_k) \\
k_2 &= f(t_k + \frac{1}{2}h_k, y_k + \frac{1}{2}h_k k_1) \\
y_{k+1} &= y_k + h_k k_2, \quad p = 2
\end{aligned}
$$

(iii) Verfahren von Runge von 3. Ordnung

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

$$
m = 3, \gamma_1 = \gamma_2 = 0, \gamma_3 = 1, \alpha_2 = \frac{1}{2}, \alpha_3 = 1, \beta_{21} = \frac{1}{2}, \beta_{31} = 0, \beta_{32} = 1
$$

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

(iv) Klassisches Runge-Kutta-Verfahren 4. Ordnung

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

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

Bemerkung. Die Ordnung eines konkreten Runge-Kutta-Verfahrens kann mit Hilfe von Taylor-Entwicklungen ermittelt werden, wobei man dabei von einer geeigneten Glattheit von $f(t, y)$ ausgeht.

Im Folgenden soll die Ordnung eines 3-stufigen expliziten Runge-Kutta- Verfahrens bestimmt werden.

Satz 8.14. *Sei f dreimal stetig partiell diff'bar und gelte für die Parameter*

$$\begin{aligned} \alpha_2 &= \beta_{21} \\ \alpha_3 &= \beta_{31} + \beta_{32} \\ \gamma_1 + \gamma_2 + \gamma_3 &= 1 \end{aligned}$$

sowie

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

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

Beweis. Grundlage für den Beweis ist die Taylor-Approximation

$$\begin{aligned} f(t + \Delta t, y + \Delta y) = f(t, y) + \begin{pmatrix} \frac{\partial f}{\partial t}(t, y) \\ \frac{\partial f}{\partial y}(t, y) \end{pmatrix} \cdot \begin{pmatrix} \Delta t \\ \Delta y \end{pmatrix} \\ + \frac{1}{2}(\Delta t, \Delta y) \begin{pmatrix} \frac{\partial^2 f}{\partial t^2}(t, y) & \frac{\partial^2 f}{\partial t \partial y}(t, y) \\ \frac{\partial^2 f}{\partial y \partial t}(t, y) & \frac{\partial^2 f}{\partial y^2}(t, y) \end{pmatrix} \begin{pmatrix} \Delta t \\ \Delta y \end{pmatrix} + \mathcal{O}(\Delta^3) \end{aligned} \tag{8.27}$$

der Funktion f, wobei $\frac{\partial^2 f}{\partial t \partial y} = \frac{\partial^2 f}{\partial y \partial t}$ aufgrund der Glattheit von f gilt. Mit

$$\begin{aligned} \bar{k}_1 &= f(t_k, y(t_k)) \\ \bar{k}_2 &= f(t_k + \alpha_2 h, y(t_k) + \beta_{21} h \bar{k}_1) = f(t_k + \alpha_2 h, y(t_k) + \alpha_2 h \bar{k}_1) \\ \bar{k}_3 &= f(t_k + \alpha_3 h, y(t_k) + h(\beta_{31}\bar{k}_1 + \beta_{32}\bar{k}_2)) \end{aligned}$$

gilt es, den lokalen Diskretisierungsfehler

$$d_{k+1} = y(t_{k+1}) - y(t_k) - h(\gamma_1\bar{k}_1 + \gamma_2\bar{k}_2 + \gamma_3\bar{k}_3)$$

abzuschätzen, wobei schon $\alpha_2 = \beta_{21}$ verwendet wurde ($h = h_k$). Mit $\Delta t = \alpha_2 h$ und $\Delta y = \alpha_2 h f(t_k, y(t_k))$ ergibt (8.27) für $\bar{k}_2$

$$\begin{aligned}\bar{k}_2 &= f(t_k + \Delta t, y(t_k) + \Delta y)\\ &= f + \alpha_2 h f_t + \alpha_2 h f f_y + \frac{1}{2}\alpha_2^2 h^2 f_{tt} + \alpha_2^2 h^2 f f_{ty} + \frac{1}{2}\alpha_2^2 h^2 f^2 f_{yy} + \mathcal{O}(h^3)\\ &=: f + \alpha_2 h F + \frac{1}{2}\alpha_2^2 h^2 G + \mathcal{O}(h^3) \qquad (8.28)\end{aligned}$$

$f, f_t, \ldots, f_{yy}$ sind dabei die Funktions- bzw. Ableitungswerte an der Stelle $(t_k, y(t_k))$. Für $\bar{k}_3$ erhält man unter Nutzung von (8.28) und (8.27)

$$\begin{aligned}\bar{k}_3 &= f(t_k + \alpha_3 h, y(t_k) + h(\beta_{31}\bar{k}_1 + \beta_{32}\bar{k}_2)\\ &= f + \alpha_3 h f_t + h(\beta_{31}\bar{k}_1 + \beta_{32}\bar{k}_2) f_y + \frac{1}{2}\alpha_3^2 h^2 f_{tt}\\ &\quad + \alpha_3(\beta_{31}\bar{k}_1 + \beta_{32}\bar{k}_2) h^2 f_{ty} + \frac{1}{2}(\beta_{31}\bar{k}_1 + \beta_{32}\bar{k}_2)^2 h^2 f_{yy} + \mathcal{O}(h^3)\\ &= f + h(\alpha_3 f_t + [\beta_{31} + \beta_{32}] f f_y) + h^2\Big(\alpha_2\beta_{32} F f_y\\ &\quad + \frac{1}{2}\alpha_3^2 f_{tt} + \alpha_3[\beta_{31} + \beta_{32}] f f_{ty} + \frac{1}{2}(\beta_{31} + \beta_{32}) f^2 f_{yy}\Big) + \mathcal{O}(h^3)\\ &= f + \alpha_3 h F + h^2(\alpha_2\beta_{32} F f_y + \frac{1}{2}\alpha_3^2 G) + \mathcal{O}(h^3) \qquad (8.29)\end{aligned}$$

Mit (8.28) und (8.29) folgt für den lokalen Diskretisierungsfehler

$$\begin{aligned}d_{k+1} &= h(1 - \gamma_1 - \gamma_2 - \gamma_3) f + h^2\left(\frac{1}{2} - \alpha_2\gamma_2 - \alpha_3\gamma_3\right) F\\ &+ h^3\left(\left[\frac{1}{6} - \alpha_2\gamma_3\beta_{32}\right] F f_y + \left[\frac{1}{6} - \frac{1}{2}\alpha_2^2\gamma_2 - \frac{1}{2}\alpha_3^2\gamma_3\right] G\right) + \mathcal{O}(h^4)\end{aligned} \qquad (8.30)$$

Aufgrund der Voraussetzungen werden dei Klammerausdrücke gleich Null und es gilt

$$d_{k+1} = \mathcal{O}(h^4)$$

also hat das Verfahren die Fehlerordnung $p = 3$ □

Korollar. *Mit Lösungen des Gleichungssystems*

$$\begin{aligned}\gamma_1 + \gamma_2 + \gamma_3 &= 1\\ \alpha_2\gamma_2 + \alpha_3\gamma_3 &= \frac{1}{2}\\ \alpha_2\gamma_3\beta_{32} &= \frac{1}{6}\\ \alpha_2^2\gamma_2 + \alpha_3^2\gamma_3 &= \frac{1}{3}\end{aligned} \qquad (8.31)$$

hat das dazugehörrige 3-stufige Runge-Kutta-Verfahren die Fehlerordnung $p = 3$*, wobei* $\alpha_2 = \beta_{21}$ *ist.* (8.31) *hat z.B. mit den Einschränkungen* $\alpha_2 \neq \alpha_3$ *und* $\alpha_2 \neq \frac{2}{3}$ *die Lösungen*

$$\gamma_2 = \frac{3\alpha_3 - 2}{6\alpha_2(\alpha_3 - \alpha_2)}, \qquad \gamma_3 = \frac{2 - 3\alpha_2}{6\alpha_3(\alpha_3 - \alpha_2)} \tag{8.32}$$
$$\gamma_1 = \frac{6\alpha_2\alpha_3 + 2 - 3(\alpha_2 + \alpha_3)}{6\alpha_2\alpha_3}, \qquad \beta_{32} = \frac{\alpha_3(\alpha_3 - \alpha_2)}{\alpha_2(2 - 3\alpha_2)}$$

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

$$\mathcal{M} = \{(\gamma_1, \gamma_2, \gamma_3, \alpha_2, \alpha_3, \beta_{32}) | \gamma_1, \gamma_2, \gamma_3, \beta_{32} \text{ gemäß (8.32)}, \alpha_2, \alpha_3 \in \mathbb{R}, \alpha_2 \neq \alpha_3, \alpha_2 \neq \frac{2}{3}\}$$

Die restlichen Parameter des Verfahrens ergeben sich aus

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

8.5 Schrittweitensteuerung bei Einschrittverfahren

8.5.1 Schrittweitensteuerung durch Einbettung

25. Vorlesung 28.01.2015

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

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

erläutert werden. Als lokaler Diskretisierungsfehler ergibt sich

$$d_{k+1}^{(H)} = y(t_{k+1}) - y(t_k) - \frac{1}{2}h[\bar{k}_1 + \bar{k}_2] \tag{8.33}$$

mit $\bar{k}_1 = f(t_k, y(t_k)), \bar{k}_2 = f(t_k + h, y(t_k) + h\bar{k}_1)$

Nun sucht man ein Verfahren höherer Ordnung, also mindestens dritter Ordnung, dessen Steigungen k_1, k_2 mit den Steigungen des Heun-Verfahrens übereinstimmen.
Die Forderung der Gleichheit von k_1 und k_2 bedeutet $\alpha_2 = \beta_{21} = 1$. Die weiteren Parameter ergeben sich aus (8.32) bei der Wahl von $\alpha_3 = \frac{1}{2}$ zu

$$\gamma_3 = \frac{2}{3}, \quad \gamma_2 = \frac{1}{6}, \quad \gamma_1 = \frac{1}{6}, \quad \beta_{32} = \frac{1}{4}, \quad \beta_{31} = \alpha_3 - \beta_{32} = \frac{1}{4}$$

sodass sich das Runge-Kutta-Verfahren 3. Ordnung

$$\begin{aligned} k_1 &= f(t_k, y_k) \\ k_2 &= f(t_k + h, y_k + hk_1) \\ k_3 &= f(t_k + \frac{h}{2}, y_k + \frac{h}{4}(k_1 + k_2)) \\ y_{k+1} &= y_k + \frac{h}{6}[k_1 + k_2 + 4k_3] \end{aligned} \tag{8.34}$$

ergibt. Für den lokalen Diskretisierungsfehler des Verfahrens (8.33) ergibt sich

$$d_{k+1}^{(RK)} = y(t_{k+1}) - y(t_k) - \frac{h}{6}[\bar{k}_1 + \bar{k}_2 + 4\bar{k}_3] \tag{8.35}$$

mit $\bar{k}_3 = f(t_k + \frac{h}{2}, y(t_k) + \frac{h}{4}(\bar{k}_1 + \bar{k}_2))$. Mit (8.33) und (8.35) ergibt sich die Darstellung des lokalen Diskretisierungsfehlers des Heun-Verfahrens

$$d_{k+1}^{(H)} = \frac{h}{6}[\bar{k}_1 + \bar{k}_2 + 4\bar{k}_3] - \frac{h}{2}[\bar{k}_1 + \bar{k}_2] + d_{k+1}^{(RK)}$$

Ersetzt man nun die unbekannten Werte von $\bar{k}_j$ durch die Näherungen k_j und berücksichtigt $d_{k+1}^{(RK)} = \mathcal{O}(h^4)$, so erhält man

$$d_{k+1}^{(H)} = \frac{h}{6}[k_1 + k_2 + 4k_3] - \frac{h}{2}[k_1 + k_2] + \mathcal{O}(h^4) = \frac{h}{3}[2k_3 - k_1 - k_2] + \mathcal{O}(h^4)$$

und damit kann der lokale Diskretisierungsfehler des Heun-Verfahrens mit einer zusätzlichen Steigungsberechnung von k_3 durch den Ausdruck $\frac{h}{3}[2k_3 - k_1 - k_2]$ recht gut geschätzt werden.
Aufgrund der Kontrolle des Betrags dieses Ausdrucks kann man eine vorgegebene Schranke $\epsilon_{\text{tol}} > 0$ durch entsprechende Wahl von $h = h_k = t_{k+1} - t_k$

$$h_k < \frac{3\epsilon_{\text{tol}}}{|2k_3 - k_1 - k_2|} \Leftrightarrow \frac{h_k}{3}[2k_3 - k_1 - k_2] < \epsilon_{\text{tol}}$$

unterschreiten. D.h. man kann die aktuelle Schrittweite evtl. vergrößern oder muss sie verkleinern.

Die eben beschriebene Methode der Schrittweitensteuerung bezeichnet man auch als Einbettung des Heun-Verfahrens 2. Ordnung in das Runge-Kutta-Verfahren 3. Ordnung (8.34).
Die Einbettung beschreibt man auch mit der **erweiterten** Butcher-Tabelle

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

In den letzten beiden Zeilen stehen erst die Gewichte für das Heun-Verfahren und darunter die Gewichte des Verfahrens, in das eingebettet wird.

8.6 Mehrschrittverfahren

26. Vorlesung 04.02.2015, und Stoff zum Selbststudium

Die Klasse der Mehrschrittverfahren zur Lösung von AWP ist dadurch gekennzeichnet, dass man zur Berechnung des Näherungswertes y_{k+1} nicht nur den Wert y_k verwendet, sondern auch weiter zurückliegende Werte, z.B. y_{k-1}, y_{k-2}.
Als Ausgangspunkt zur Konstruktion von Mehrschrittverfahren betrachten wir die zum AWP äquivalente Integralbeziehung

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(s, y(s))\mathrm{d}s \tag{8.36}$$

Kennt man die Werte $f_k = f(t_k, y_k), \ldots, f_{k-3} = f(t_{k-1}, y_{k-3})$, dann kann man das Integral auf der rechten Seite von (8.36) i.d.R. besser approximieren als bei den Einschrittverfahren unter ausschließlicher Nutzung des Wertes f_k. Für das Interpolationspolynom durch die Stützpunkte $(t_j, f_j)_{j=k-3,\ldots,k}$ ergibt sich

$$p_3(t) = \sum_{j=0}^{3} f_{k-j} L_{k-j}$$

mit den Lagrangschen Basispolynomen

$$L_j(t) = \prod_{j \neq i = k-3}^{k} \frac{t - t_i}{t_j - t_i}, \quad j = k-3, \ldots, k$$

Die Idee der Mehrschrittverfahren besteht nun in der Nutzung von $p_3(t)$ als Approximation von $f(t, y(t))$ im Integral von (8.36), sodass man auf der

Grundlage von (8.36) das Mehrschrittverfahren (4-Schritt-Verfahren)

$$\begin{aligned} y_{k+1} &= y_k + \int_{t_k}^{t_{k+1}} \sum_{j=0}^{3} f_{k-j} L_{k-j}(t) \mathrm{d}t \\ &= y_k + \sum_{j=0}^{3} f_{k-j} \int_{t_k}^{t_{k+1}} L_{k-j}(t) \mathrm{d}t \end{aligned}$$

erhält. Im Fall äquidistanter Stützstellen $h = t_{k+1} - t_k$ erhält man für das Integral des 2. Summanden ($j = 1$)

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

und nach der Substitution $\xi = \frac{t - t_k}{h}$

$$I_1 = h \int_0^1 \frac{(\xi + 3)(\xi + 2)\xi}{2 \cdot 1 \cdot (-1)} \mathrm{d}\xi = -\frac{h}{2} \int_0^1 (\xi^3 + 5\xi^2 + 6\xi) \mathrm{d}\xi = -\frac{59}{24} h$$

Für die restlichen Integrale erhält man

$$I_0 = \frac{55}{24} h, \quad I_2 = \frac{37}{24} h, \quad I_3 = -\frac{9}{24} h$$

sodass sich mit

$$y_{k+1} = y_k + \frac{h}{24} [55 f_k - 59 f_{k-1} + 37 f_{k-2} - 9 f_{k-3}] \tag{8.37}$$

ein explizites Verfahren, bei dem 4 Schritte Verwendung finden, das als **Methode von Adams-Bashforths** (kurz AB-Verfahren) bezeichnet wird.
Durch Taylor-Reihenentwicklung erhält man bei entsprechender Glattheit von f bzw. $y(t)$ den lokalen Diskretisierungsfehler

$$d_{k+1} = \frac{251}{720} h^5 y^{(5)} + \mathcal{O}(h^6) \tag{8.38}$$

d.h. das Verfahren (8.37) ist von 4. Ordnung.

Definition 8.15. *Bei Verwendung von m Stützwerten*

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

zur Berechnung eines Interpolationspolynoms p_{m-1} zur Approximation von f zwecks näherungsweiser Berechnung des Integrals in (8.36) *spricht man von einem linearen m-Schrittverfahren.*

Ein m-Schrittverfahren hat die Fehlerordngung p, falls für seinen lokalen Diskretisierungsfehler d_k die Abschätzung

$$\max_{m \leq k \leq N} |d_k| \leq K = \mathcal{O}(h^{p+1})$$

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.

$$\begin{aligned} y_{k+1} &= y_k + \frac{h}{12}[23f_k - 16f_{k-1} + 5f_{k-2}] \\ y_{k+1} &= y_k + \frac{h}{24}[55f_k - 59f_{k-1} + 37f_{k-2} - 9f_{k-3}] \\ y_{k+1} &= y_k + \frac{h}{720}[1901f_k - 2774f_{k-1} + 2616f_{k-2} - 1274f_{k-3} + 251f_{k-4}] \end{aligned} \tag{8.39}$$

Die Formeln der Mehrschrittverfahren "funktionieren" erst ab dem Index $k = m$, d.h. bei einem 3-Schrittverfahren braucht man die Werte y_0, y_1, y_2 um y_3 mit der Formel (8.39) berechnen zu können.
Die Startwerte y_1, y_2 werden meist mit einem Runge-Kutta-Verfahren berechnet.
Es ist offensichtlich möglich die Qualität der Lösungsverfahren für das AWP zu erhöhen, indem man das Integral

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

aus der Beziehung (8.36) 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

$$y_{k+1} = y_k + \frac{h}{720}[251 f_{k+1} + 646 f_k - 264 f_{k-1} + 106 f_{k-2} - 19 f_{k-3}] \tag{8.40}$$

Das Verfahren (8.40) ist eine implizite 4-Schritt-Methode und heißt Methode von Adams-Moulton (kurz AM-Verfahren). Das 3-Schritt AM-Verfahren hat die Form

$$y_{k+1} = y_k + \frac{h}{24}[9 f(t_{k+1}, y_{k+1}) + 19 f_k - 5 f_{k-1} + f_{k-2}] \tag{8.41}$$

Zur Bestimmung der Lösung von (8.41) 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

$$y_{k+1}^{(p)} = y_k + \frac{h}{12}[23 f_k - 16 f_{k-1} + 5 f_{k-2}] \tag{8.42}$$

$$y_{k+1} = y_k + \frac{h}{24}[9 f(t_{k+1}, y_{k+1}^{(p)}) + 19 f_k - 5 f_{k-1} + f_{k-2}] \tag{8.43}$$

Diese Kombination von AB- und AM-Verfahren bezeichnet man als Adams-Bashforth-Moulton-Verfahren (kurz ABM-Verfahren). Das ABM-Verfahren (8.42) hat ebenso wie das AM-Verfahren (8.41) den Diskretisierungsfehler $d_{k+1} \in \mathcal{O}(h^5)$ und damit die Fehlerordnung 4.
Generell kann man zeigen, dass m-Schritt-Verfahren von AM- oder ABM-Typ jeweils die Fehlerordnung $p = m + 1$ besitzen.

8.7 Allgemeine lineare Mehrschrittverfahren

Definition 8.16. *Unter einem linearen m-Schrittverfahren ($m > 1$) versteht man eine Vorschrift mit $s = k + 1 - m$*

$$\sum_{j=0}^{m} a_j y_{s+j} = h \sum_{j=0}^{m} b_j f(t_{s+j}, y_{s+j}) \tag{8.44}$$

wobei $a_m \neq 0$ ist und a_j, b_j geeignet zu wählende reelle Zahlen sind. Die konkrete Wahl dieser Koeffizienten entscheidet über die Ordnung des Verfahrens.

Bemerkung. In den bisher behandelten Verfahren war jeweils $a_m = 1$ und $a_{m-1} = -1$ sowie $a_{m-2} = \cdots = a_0 = 0$. Bei expliziten Verfahren ist $b_m = 0$ und bei impliziten Verfahren ist $b_m \neq 0$

OBdA setzen wir $a_m = 1$. Die anderen freien Parameter a_j, b_j sind so zu wählen, dass die linke und die rechte Seite von (8.44) Approximationen von

$$\alpha[y(t_{k+1}) - y(t_k)] \quad \text{bzw.} \quad \alpha \int_{t_k}^{t_{k+1}} f(t, y(t))\mathrm{d}t$$

sind ($\alpha \neq 0$)
Mit der Einführung der Parameter $a_0, \ldots, a_m$ hat man die Möglichkeit durch die Nutzung der Werte $y_{k+1-m}, \ldots, y_{k+1}$ nicht nur die Approximation von f, sondern auch die Approximation von y' mit einer höheren Ordnung durchzuführen.

Beispiel. Das 3-Schritt-AB-Verfahren

$$y_{k+1} = y_k + \frac{h}{12}[23f_k - 16f_{k-1} + 5f_{k-2}]$$

ist gleichbedeutend mit

$$\frac{y_{k+1} - y_k}{h} = \frac{1}{12}[23f_k - 16f_{k-1} + 5f_{k-2}]$$

wobei die rechte Seite eine Approximation von $f(t_k, y(t_k))$ der Ordnung $\mathcal{O}(h^3)$ darstellt, während die linke Seite y' an der Stelle t_k nur mit der Ordnung $\mathcal{O}(h)$ approximiert. Das Verfahren ist in der vorliegenden Form 3. Ordnung.
Nutzt man neben y_{k+1} und y_k auch noch y_{k-1}, y_{k-2}, dann kann man unter Nutzung der Taylor-Approximationen

$$\begin{aligned}
y(t_{k-2}) &= y(t_k) - 2hy' + 2h^2y'' - \frac{3}{2}h^3y''' + \mathcal{O}(h^4) \\
y(t_{k-1}) &= y(t_k) - hy' + \frac{1}{2}h^2y'' - \frac{1}{6}h^3y''' + \mathcal{O}(h^4) \\
y(t_{k+1}) &= y(t_k) + hy' + \frac{1}{2}h^2y'' - \frac{1}{6}h^3y''' + \mathcal{O}(h^4)
\end{aligned}$$

mit

$$\frac{1}{14h}[5y_{k+1} + 6y_k - 13y_{k-1} + 2y_{k-2}]$$

die linke Seite y' ebenfalls mit der Ordnung $\mathcal{O}(h^3)$ approximieren.
Allerdings ist das daraus resultierende Mehrschrittverfahren

$$y_{k+1} + \frac{6}{5}y_k - \frac{13}{5}y_{k-1} + \frac{2}{5}y_{k-2} = h[\frac{161}{30}f_k - \frac{56}{15}f_{k-1} + \frac{7}{6}f_{k-2}]$$

wie sie später nachrechnen können nur von erster Ordnung.
Die betriebene Mehraufwand ist allerdings in manchen Fällen mit Blick auf einen evtln. Stabilitätszuwachs trotz Ordnungsverlust sinnvoll. In diesem Fall bringt der Mehraufwand nichts, weil man zwar noch ein konsistentes Verfahren hat, das allerdings nicht nullstabil ist (eine Nullstelle des ersten charakteristischen Polynoms ist betragsmäßig größer als Null!).

Definition 8.17. *Das lineare Mehrschrittverfahren* (8.44) *hat die Fehlerordnung p, falls in der Entwicklung des lokalen Diskretisierungsfehlers d_{k+1} in eine Potenzreihe von h für eine beliebige Stelle $\tilde{t} \in [t_{k+1-m}, t_{k+1}]$*

$$\begin{aligned} d_{k+1} &= \sum_{j=0}^{m} [a_j y(t_{s+j}) - h b_j f(t_{s+j}, y(t_{s+j}))] \\ &= c_0 y(\tilde{t}) + c_y h y'(\tilde{t}) + \ldots + c_p h^p y^{(p)}(\tilde{t}) + \ldots \end{aligned} \tag{8.45}$$

$c_0 = \ldots = c_p = 0$ und $c_{p+1} \neq 0$ gilt ($s = k+1-m$). Ein Mehrschrittverfahren heißt konsistent, wenn es mindestens die Ordnung $p = 1$ besitzt.

Durch eine günstige Wahl von $\tilde{t}$ kann man die Entwicklungskoeffizienten c_j oft in einfacher Form als Linearkombination von a_j, b_j darstellen und erhält mit der Bedingung $c_0 = \ldots = c_p = 0$ Bestimmungsgleichungen für die Koeffizienten des Mehrschrittverfahrens.
Mit der Wahl von $\tilde{t} = t_s$ ergeben sich für y, y' die Taylor-Reihen

$$\begin{aligned} y(t_{s+j}) = y(\tilde{t} + jh) &= \sum_{r=0}^{q} \frac{(jh)^r}{r!} y^{(r)}(\tilde{t}) + R_{q+1} \\ y'(t_{s+j}) = y'(\tilde{t} + jh) &= \sum_{r=0}^{q-1} \frac{(jh)^r}{r!} y^{(r+1)}(\tilde{t}) + R_q \end{aligned} \tag{8.46}$$

Die Substitution der Reihen (8.46) in (8.45) ergibt für die Koeffizienten c_j

durch Koeffizientenvergleich

$$\begin{aligned}
c_0 &= a_0 + a_1 + \ldots + a_m \\
c_1 &= a_1 + 2a_2 + \ldots + ma_m - (b_0 + b_1 + \ldots + b_m) \\
c_2 &= \frac{1}{2!}(a_1 + 2^2 a_2 + \ldots + m^2 a_m) - \frac{1}{1!}(b_1 + 2b_2 + \ldots + mb_m) \\
&\vdots \\
c_r &= \frac{1}{r!}(a_1 + 2^r a_2 + \ldots + m^r a_m) - \frac{1}{(r-1)!}(b_1 + 2^{r-1} b_2 + \ldots + m^{r-1} b_m)
\end{aligned} \tag{8.47}$$

für $r = 2, 3, \ldots, q$

Beispiel. Es soll ein explizites 2-Schritt-Verfahren (d.h. $b_2 = 0$)

$$a_0 y_{k-1} + a_1 y_k + a_2 y_{k+1} = h[b_0 f_{k-1} + b_1 f_k] \tag{8.48}$$

der Ordnung 2 bestimmt werden. Mit der Fixierung von $a_2 = 1$ ergibt sich mit

$$\begin{aligned}
c_0 &= a_0 + a_1 + 1 = 0 \\
c_1 &= a_1 + 2 - (b_0 + b_1) = 0 \\
c_2 &= \frac{1}{2}(a_1 + 4) - b_1 = 0
\end{aligned}$$

ein Gleichungssystem mit 3 Gleichungen für 4 Unbekannte zur Verfüfgung, d.h. es gibt noch einen Freiheitsgrad.
Wählt man $a_1 = 0$, dann folgt für die restlichen Parameter $a_0 = -1, b_0 = 0$ und $b_1 = 2$, sodass das Verfahren die Form

$$y_{k+1} = y_{k-1} + 2hf_k \tag{8.49}$$

hat.

Definition 8.18. *Mit den Koeffizienten a_j, b_j werden durch*

$$\rho(z) = \sum_{j=0}^{m} a_j z^j, \quad \sigma(z) = \sum_{j=0}^{m} b_j z^j \tag{8.50}$$

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

Aus dem Gleichungssystem (8.47) kann man mit Hilfe der charakteristischen Polynome die folgende notwendige und hinreichende Bedingung für die Konsistenz eines Mehrschrittverfahrens formulieren.

Satz 8.19. *Notwendig und hinreichend für die Konsistenz des Mehrschrittverfahrens* (8.44) *ist die Erfüllung der Bedingungen*

$$c_0 = \rho(1) = 0, \quad c_1 = \rho'(1) - \sigma(1) = 0 \tag{8.51}$$

Macht man außer der Wahl von $a_2 = 1$ keine weiteren Einschränkungen an die Koeffizienten des expliziten 2-Schritt-Verfahrens (8.48), dann erreicht man die maximale Ordnung $p = 3$ durch die Lösung des Gleichungssystem (8.47) für $q = 3$, also $c_0 = c_1 = c_2 = c_3 = 0$.
Man findet die eindeutige Lösung

$$a_0 = -5, \quad a_1 = 4, \quad b_0 = 2, \quad b_1 = 4$$

woraus das Verfahren

$$y_{k+1} = 5y_{k-1} - 4y_k + h[4f_k + 2f_{k-1}] \tag{8.52}$$

folgt.

8.8 Stabilität von Mehrschrittverfahren

Obwohl das Verfahren (8.52) die maximale Fehlerordnung $p = 3$ hat, ist es im Vergleich zum Verfahren (8.49) unbrauchbar, weil es nicht stabil ist. Was das konkret bedeutet, soll im Folgenden erklärt und untersucht werden.
Dazu wird die Testdifferentialgleichung (AWP)

$$y' = \lambda y, \quad y(0) = 1, \ \lambda \in \mathbb{R}, \ \lambda < 0 \tag{8.53}$$

mit der eindeutig bestimmten Lösung $y(t) = e^{\lambda t}$ betrachtet.
Von einem brauchbaren numerischen Verfahren erwartet man mindestens die Widerspiegelung des qualitativen Lösungsverhaltens.
Mit $f = \lambda y$ folgt aus (8.52)

$$\begin{aligned} & y_{k+1} = 5y_{k-1} - 4y_k + h[4\lambda y_k + 2\lambda y_{k-1}] \\ \Leftrightarrow \quad & (-5 - 2\lambda h)y_{k-1} + (4 - 4\lambda h)y_k + y_{k+1} = 0 \end{aligned} \tag{8.54}$$

Mit dem Lösungsansatz $y_k = z^k, z \neq 0$ ergibt (8.54) nach Division mit z^{k-1}

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

bzw. mit dem charakteristischen Polynomen

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

Als Nullstellen von 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 (8.54)

$$y_k = c_1 z_1^k + c_2 z_2^k \tag{8.55}$$

wobei die Konstanten c_1, c_2 aus Anfangsbedingungen der Form $c_1 + c_2 = y_0, z_1 c_1 + z_2 c_2 = y_1$ zu ermitteln sind.
Notwendig für das Abklingen der Lösung y_k in der Formel (8.55) für wachsendes k ist die Bedingung $|z_{1,2}| \leq 1$. Da für $h \to 0$ die Nullstellen von $\phi(z)$ in die Nullstellen von $\rho(z)$ übergehen, dürfen diese dem Betrage nach nicht größer als 1 sein. Im Fall einer doppelten Nullstelle z von $\phi(z)$ eines 2-Schritt-Verfahrens hat die Lösung der entsprechenden DGL statt (8.55) die Form

$$y_k = c_1 z^k + c_2 k z^k$$

sodass für das Abklingen von y_k für wachsendes k die Bedingung $|z| < 1$ erfüllt sein muss. Die durchgeführten Überlegungen rechtfertigen die folgende Definition.

Definition 8.20. *Das Mehrschrittverfahren* (8.44) *heißt* ***nullstabil****, falls die Nullstellen z_j des ersten charakteristischen Polynoms $\rho(z)$*

(a) betragsmäßig nicht größer als 1 sind

(b) mehrfache Nullstellen betragsmäßig echt kleiner als 1 sind

Für das oben konstruierte 2-Schritt-Verfahren mit der maximalen Fehlerordnung $p = 3$ hat das charakteristische Polynom $\rho(z)$ die Nullstellen $z_{1,2} = -2 \pm 3$ und damit ist das Verfahren nicht nullstabil.
Im Unterschied dazu ist das Verfahren (8.49) mit der Ordnung 2 und dem charakteristischen Polynom $\rho(z) = -1 + z^2$ und den Nullstellen $z_{12} = \pm 1$ nullstabil.

Bemerkung. Einschrittverfahren sind mit dem charakteristischen Polynom $\rho(z) = -1 + z$ generell nullstabil.
Aufgrund der ersten charakteristischen Polynome der Adams-Bashforth- und Adams-Moulton-Verfahren erkennt man, dass diese auch generell nullstabil sind.

Bemerkung. Konsistente und nullstabile Mehrschrittverfahren sind konvergent, falls die Funktion $f(t, y)$ bezüglich y lipschitzstetig ist.
Der Beweis verläuft im Fall expliziter Mehrschrittverfahren analog zum Konvergenzbeweis für konsistente Einschrittverfahren und sollte als Übung erbracht werden.

8.9 Begriff der absoluten Stabilität

Bei den Betrachtungen zur Nullstabilität wurde eine Testaufgabe zugrunde gelegt. Um den Begriff der **absoluten Stabilität** zu erläutern, wird die Testaufgabe leicht modifiziert, und zwar zu

$$y' = \lambda y, \quad y(0) = 1, \ \lambda \in \mathbb{R} \text{ oder } \lambda \in \mathbb{C} \tag{8.56}$$

d.h. wir lassen auch Parameter λ aus $\mathbb{C}$ zu. Damit sind auch Lösungen der Form $e^{\alpha t}\cos(\beta t)$ möglich. Numerische Lösungsverfahren sollen auch in diesem Fall für $\alpha = \Re(\lambda) < 0$ den dann stattfindenden Abklingprozess korrekt wiedergeben. Für das Eulerverfahren zur Lösung von (8.56) erhält man mit $f(t, y) = \lambda y$

$$y_{k+1} = y_k + h\lambda y_k \ \Leftrightarrow \ y_{k+1} = (1 + h\lambda)y_k =: F(h\lambda)y_k$$

Falls $\lambda > 0$ und reell ist, wird die Lösung, für die

$$y(t_{k+1}) = y(t_k + h) = e^{h\lambda}y(t_k)$$

gilt, in jedem Fall qualitativ richtig wiedergegeben, denn der Faktor $F(h\lambda) = 1 + \lambda h$ besteht gerade aus den beiden ersten Summanden der e-Reihe, und es wird ein Fehler der Ordnung 2 gemacht.
Im Fall $\lambda < 0$ wird nur unter der Bedingung $|F(h\lambda)| = |1 + \lambda h| < 1$ das Abklingverhalten der Lösung beschrieben. Des Fall des reellen Parameters $\lambda < 0$ ist deshalb von Interesse.
Beim RK-Verfahren 4. Ordnung

$$\begin{aligned} k_1 &= \lambda y_k \\ k_2 &= \lambda(y_k + \frac{1}{2}hk_1) = (\lambda + \frac{1}{2}h\lambda^2)y_k \\ k_3 &= \lambda(y_k - hk_1 + 2hk_2) = (\lambda + h\lambda^2 + h^2\lambda^3)y_k \end{aligned} \tag{8.57}$$

$$y_{k+1} = y_k + \frac{h}{6}[k_1 + 4k_2 + k_3] = (1 + h\lambda + \frac{h^2}{2}\lambda^2 + \frac{h^3}{6}\lambda^3)y_k \tag{8.58}$$

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

$$F(h\lambda) = 1 + h\lambda + \frac{h^2}{2}\lambda^2 + \frac{h^3}{6}\lambda^3 \tag{8.59}$$

Der Faktor (8.59) enthält gerade die ersten 4 Summanden der e-Reihe und es wird ein Fehler der Ordnung 4 gemacht, sodass $y(t) = e^{t\lambda}$ durch das Verfahren (8.57) qualitativ beschrieben wird.

Für $\lambda < 0$ reell, muss die Lösung abklingen, was durch die Bedingung $|F(h\lambda)| < 1$ erreicht wird.
Wegen $\lim_{h\lambda \to -\infty} F(h\lambda) = -\infty$ wird das Abklingen nicht für beliebige negative Parameter 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 Lösung der Testaufgabe zu sichern, sondern für F muss $|F(h\lambda)| < 1$ gelten.
Die durchgeführten Überlegungen rechtfertigen die

Definition 8.21. *Für ein Einschrittverfahren, das für die Testaufgabe $y' = \lambda y, y(0) = 1, \lambda \in \mathbb{C}$ auf die Vorschrift $y_{k+1} = F(h\lambda)y_k$ führt, nennt man die Menge*

$$B := \{\mu \in \mathbb{C} : |F(\mu)| < 1\}$$

Gebiet der absoluten Stabilität.

Das Gebiet der absoluten Stabilität liefert eine Infomration zur Wahl der Schrittweite. Da man aber in den meisten "Ernstfällen" eventuelle Abklingkonstanten nicht kennt, hat man mit der Kenntnis von B keine quantitative Bedingung zur Berechnung der Schrittweite zur Verfügung.

Beispiel.

- Euler explizit: $y_{k+1} = y_k + h\lambda y_k$
  $F(\mu) = 1 + \mu$
- Euler implizit: $y_{k+1} = y_k + h\lambda y_{k+1}$
  $F(\mu) = \frac{1}{1-\mu}$
- Runge-Kutta-Verfahren 2. Ordnung

$$k_1 = f(t_k, y_k), \quad k_2 = f(t_k + \frac{h}{2}, y_k + \frac{h}{2}k_1), \quad y_{k+1} = y_k + hk_2$$

  $F(\mu) = 1 + \mu + \frac{\mu^2}{2}$

Bemerkung. Die Randkurve von B erhält man wegen $\left|e^{i\theta}\right| = 1$ über die Parametrisierung

$$\begin{aligned} & F(\mu) = 1 + \mu + \frac{1}{2}\mu^2 = e^{i\theta} \\ \Leftrightarrow \quad & \mu^2 + 2\mu + 2 - 2e^{i\theta} = 0 \\ \rightsquigarrow \quad & \mu(\theta) = -1 \pm \sqrt{1 - 2 + 2e^{i\theta}}, \quad \theta \in [0, 2\pi] \end{aligned}$$

In der folgenden Tabelle sind die reellen Stabilitätsintervalle, d.h. die Schnittmenge der Gebiete der absoluten Stabilität mit der $\Re(\mu)$-Achse, für explizite r-stufige Runge-Kutta-Verfahren angegeben.

r	
1	]-2,0[
2	]-2,0[
3	]-2.51,0[
4	]-2.78,0[
5	]-3.21,0[

Definition 8.22. *Hat ein Einschrittverfahren als Gebiet der absoluten Stabilität mindestens die gesamte linke Halbebene, also $B \supset \{\mu \in \mathbb{C} : \Re(\mu) < 0\}$, dann nennt man das Verfahren absolut stabil.*

Neben dem impliziten Eulerverfahren (einstufiges RK-Verfahren) sind auch andere implizite RK-Verfahren absolut stabil, z.b. das Verfahren

$$k_1 = f\left(t_k + \frac{h}{2}, y_k + \frac{h}{2}k_1\right), \quad y_{k+1} = y_k + hk_1$$

Mit $f = \lambda y$ erhält man

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

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

8.10 BDF-Verfahren

Lineare Mehrschritt-Verfahren, bei denen bis auf den Koeffizienten b_m alle anderen b-Koeffizienten gleich null sind, also Verfahren der Form

$$\sum_{j=0}^{m} a_j y_{k-m+1+j} = h b_m f(t_{k+1}, y_{k+1}) \,, \tag{8.60}$$

werden Rückwärtsdifferentiationsmethoden oder kurz **BDF-Verfahren** (backward differentiation formula) genannt.

Die Koeffizienten a_j für ein m-Schritt-BDF-Verfahren bestimmt man, indem man das Interpolationspolynom P_m mit den $m+1$ Daten

$$(t_{k+1}, y_{k+1}), (t_k, y_k), \ldots, (t_{k+1-m}, y_{k+1-m})$$

bestimmt, und y' an der Stelle t_{k+1} durch $P'_m(t_{k+1})$ approximiert. Damit ergibt sich

$$P_m(t) = \sum_{j=0}^{m} L_j(\frac{t_{k+1}-t}{h}) y_{k+1-j} \quad \text{mit} \quad L_j(t) = \prod_{s=0, s\neq j}^{m} \frac{t-s}{j-s}$$

als den Lagrange'schen Basispolynomen. Für $P'_m(t_{k+1})$ erhält man nun

$$P'_m(t_{k+1}) = \sum_{j=0}^{m} (-\frac{1}{h}) L'_j(0) y_{k+1-j} \ ,$$

so dass sich das m-Schritt-BDF-Verfahren mit

$$\sum_{j=0}^{m} L'_j(0) y_{k+1-j} = h f_{k+1}$$

oder in der Standardform

$$y_{k+1} + \sum_{j=1}^{m} L'_j(0)/L'_0(0) y_{k+1-j} = h/(-L'_0(0)) f_{k+1}$$

ergibt. Diese Verfahren werden auch **Gear**-Verfahren genannt und haben die Konsistenzordnung m.
Die einfachsten 2- und 3-Schritt-BDF-Verfahren 2. und 3. Ordnung haben die Form

$$y_{k+1} - \frac{4}{3} y_k + \frac{1}{3} y_{k-1} = h \frac{2}{3} f(t_{k+1}, y_{k+1}) \ , \tag{8.61}$$

$$y_{k+1} - \frac{18}{11} y_k + \frac{9}{11} y_{k-1} - \frac{2}{11} y_{k-2} = h \frac{6}{11} f(t_{k+1}, y_{k+1}) \ . \tag{8.62}$$

Das einfachste BDF-Verfahren ist das so genannte Euler-rückwärts-Verfahren

$$y_{k+1} - y_k = h f(t_{k+1}, y_{k+1}) \ . \tag{8.63}$$

Für das Euler-rückwärts-Verfahren findet man für das Testproblem $y' = \lambda y$ schnell mit der Beziehung

$$y_{k+1} = \frac{1}{1 - h\lambda} y_k = F(h\lambda) y_k$$

heraus, dass $|F(h\lambda)| < 1$ für $Re(\lambda) < 0$ ist. D.h., das Euler-rückwärts-Verfahren ist absolut stabil. Das BDF-Verfahren (8.61) hat die charakteristische Gleichung

$$\phi(z) = z^2 - \frac{4}{3}z + \frac{1}{3} - \mu\frac{2}{3}z^2 = 0 \iff \mu(z) = \frac{3z^2 - 4z + 1}{2z^2} .$$

Für die Punkte $z = e^{i\theta}$, $\theta \in [0, 2\pi]$ erhält man die in der Abb. 8.1 skizzierte Randkurve $\mu(z(\theta))$ des Gebiets der absoluten Stabilität. Da man z.B. für $\mu = -\frac{1}{2}$ die Lösung $z_{1,2} = \frac{1}{2}$ mit $|z_{1,2}| < 1$ findet, kann man schlussfolgern, dass der Bereich der absoluten Stabilität im Außenbereich der Randkurve liegt. Damit ist das Verfahren (8.61) absolut stabil. Das Verfahren (8.62) ist nicht

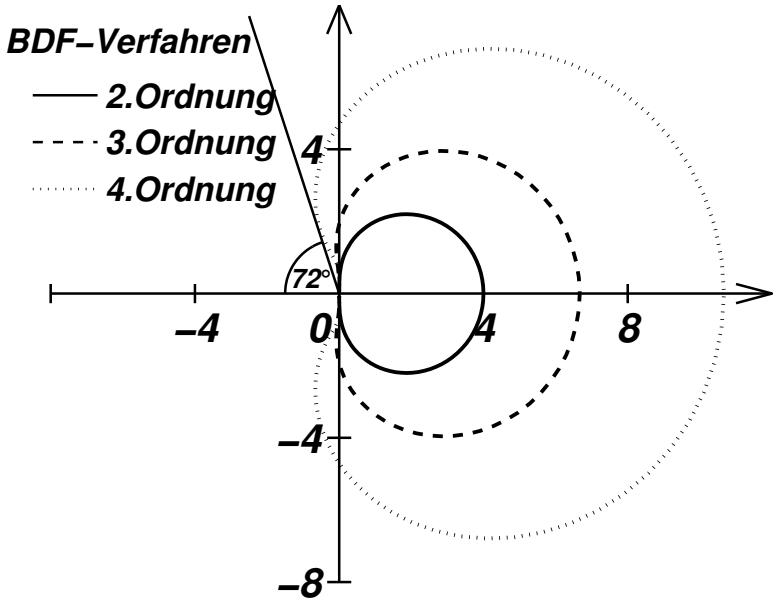

Abbildung 8.1: Gebiete der absoluten Stabilität der BDF-Verfahren (8.61), (8.62) und (8.64)

absolut stabil, weil das Gebiet der absoluten Stabilität nicht die gesamte linke komplexe Halbebene enthält. In der Abb. 8.1 ist der Rand des Gebietes der absoluten Stabilität des Verfahrens skizziert. Das Gebiet liegt wiederum im Außenbereich der Randkurve. In solchen Situationen kann man den Winkel α zwischen der reellen Achse und einer Tangente an die Randkurve durch den Ursprung legen. Bei dem BDF-Verfahren (8.62) ist der Winkel $\alpha = 88^o$, so dass das Verfahren $A(88^o)$-stabil ist. $A(90^o)$-Stabilität bedeutet absolute Stabilität. Liegt der Winkel α nahe bei 90^o, dann liegt zwar kein absolut stabiles, jedoch ein "sehr" stabiles Verfahren vor. Bei BDF-Verfahren höherer

Ordnung wird der Winkel α kleiner, so dass die Stabilität der BDF-Verfahren nachlässt, jedoch zumindest noch $A(\alpha)$-stabil sind. Zur Illustration ist das Gebiet der absoluten Stabilität des 4-Schritt-BDF-Verfahrens

$$y_{k+1} - \frac{48}{25}y_k + \frac{36}{25}y_{k-1} - \frac{16}{25}y_{k-2} + \frac{3}{25}y_{k-3} = h\frac{12}{25}f(t_{k+1}, y_{k+1}) \,, \qquad (8.64)$$

also die Kurve

$$\mu(z(\theta)) = \frac{25z^4 - 48z^3 + 36z^2 - 16z + 3}{12z^4} = \frac{25e^{i4\theta} - 48e^{i3\theta} + 36e^{i2\theta} - 16e^{i\theta} + 3}{12e^{i4\theta}} \,,$$

$\theta \in [0, 2\pi]$, in der Abbildung 8.1 im Vergleich zu den Verfahren (8.61) und (8.62) skizziert. Das Verfahren (8.64) ist $A(72^o)$-stabil.