

NUMERIK I, AKTUELLER VORLESUNGSSTAND

Günter Bärwolff

30.01.2014

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	8
2	Stabilität, Vorwärtsanalyse, Rückwärtsanalyse	12
2.1	Kondition als Maß für Fehlerverstärkungen	12
2.2	Vektor- und Matrixnormen	13
2.3	Stabilitätskonzepte	17
2.3.1	Vorwärtsanalyse	17
2.3.2	Rückwärtsanalyse	19
3	Lösung linearer Gleichungssysteme	22
3.1	LR-Zerlegung	22
3.1.1	Realisierung mit dem Gaußschen Eliminationsverfahren	24
3.1.2	LR-Zerlegung mit Spaltenpivotisierung	30
3.2	Cholesky-Zerlegung	33
3.3	Singulärwertzerlegung	36
4	Die iterative Lösung von Gleichungen bzw. Gleichungssystemen	46
4.1	Die iterative Lösung linearer Gleichungssysteme	47
4.2	Jacobi-Verfahren oder Gesamtschrittverfahren	50
4.3	Gauß-Seidel-Iterationsverfahren oder Einzelschrittverfahren . .	54
4.4	Verallgemeinerung des Gauß-Seidel-Verfahrens	55

4.5	Krylov-Raum-Verfahren zur Lösung linearer Gleichungssysteme	56
4.5.1	Der Ansatz des orthogonalen Residuums (4.20) für symmetrische positiv definite Matrizen	57
4.5.2	Der Ansatz des orthogonalen Residuums (4.20) für gegebene A -konjugierte Basen	58
4.5.3	Das CG-Verfahren für positiv definite, symmetrische Matrizen	59
4.5.4	Konvergenzgeschwindigkeit des CG-Verfahrens	62
4.5.5	CGNR-Verfahren	65
4.5.6	GMRES-Verfahren	65
4.6	Die iterative Lösung nichtlinearer Gleichungssysteme	66
4.6.1	Newton-Verfahren	68
4.6.2	Sekantenverfahren – Regula falsi	70
4.6.3	Newton-Verfahren für nichtlineare Gleichungssysteme $F(x) = 0$	73
5	Orthogonale Matrizen – QR-Zerlegung – Ausgleichsprobleme	75
5.1	Gram-Schmidt-Verfahren zur Orthogonalisierung	77
5.2	Householder-Matrizen/Transformationen	78
5.3	Algorithmus zur Konstruktion der Faktorisierung mittels Householder-Transformationen	79
5.4	Gauß-Newton-Verfahren	81
6	Interpolation	86
6.1	Polynominterpolation	87
6.1.1	Konstruktion des Interpolationspolynoms	88
6.2	Lagrange-Interpolation	89
6.3	Newton-Interpolation	91
6.3.1	Algorithmische Aspekte der Polynominterpolation - Horner-Schema	94
6.3.2	Verfahren von Neville und Aitken	96
6.4	Fehlerabschätzung der Polynominterpolation	97
6.5	Hermite-Interpolation	98
6.6	Spline-Interpolation	100
6.6.1	Interpolierende lineare Splines $s \in S_{\Delta,1}$	101
6.6.2	Kubische Splines	102
6.6.3	Berechnung interpolierender kubischer Splines	104
6.6.4	Gestalt der Gleichungssysteme	105
6.7	Existenz und Eindeutigkeit der betrachteten interpolierenden kubischen Splines	106

6.8	Fehlerabschätzungen für interpolierende kubische Splines . . .	107
6.9	Trigonometrische Interpolation	109
6.9.1	Beziehungen zwischen den reellen und komplexen Fourierkoeffizienten A_j, B_j, β_j	114
6.9.2	Schnelle Fouriertransformation (FFT)	115
6.9.3	Aufwand der FFT	118
7	Numerische Integration	120
7.1	Interpolatorische Quadraturformeln	120
7.2	Fehler bei der interpolatorischen Quadratur	121
7.3	Numerischen Integration mit Newton-Cotes-Formeln	122
7.4	Summierte abgeschlossene Newton-Cotes-Quadraturformeln	126
7.5	Gauß-Quadraturen	127
7.5.1	Orthogonale Polynome	129
7.5.2	Konstruktion von Folgen orthogonaler Polynome	129
7.6	Numerische Integration durch Extrapolation	135
7.7	Anwendung des Schemas von Neville Aitken - Romberg-Verfahren	137
8	Numerische Lösung von Anfangswertaufgaben	140
8.1	Theorie der Einschrittverfahren	142
8.2	Spezielle Einschrittverfahren	145
8.2.1	Euler-Verfahren	145
8.2.2	Einschrittverfahren der Konsistenzordnung $p = 2$	145
8.3	Verfahren höherer Ordnung	147
8.4	Einige konkrete Runge-Kutta-Verfahren und deren Butcher-Tabellen	150
8.5	Schrittweitensteuerung bei Einschrittverfahren	154
8.5.1	Schrittweitensteuerung durch Einbettung	154
8.5.2	Schrittweitensteuerung durch Extrapolation (nur zur Information, nicht prüfungsrelevant)	156
8.6	Implizite Runge-Kutta-Verfahren	158
8.7	Rundungsfehleranalyse von expliziten Einschrittverfahren	159
8.8	Ein Anwendungsgebiet für Löser von AWP's	160
A	Anhang	162
A.1	Eigenschaften von Matrizen im Ergebnis von FD-Schemen	162
A.2	Schachbrett-, Zebra- und andere Variablennummerierungen	173

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

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 Aufgabe $\frac{1}{3} = 0.33333 \dots$ eigentlich nie fertig, d.h. man gibt irgendwann erschöpft auf und macht einen Fehler.

1. Vor-
lesung
am
20.10.2013

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

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

mit $a_1, a_2, \dots \in \{0, 1, \dots, b-1\}, e \in \mathbb{Z}, \sigma \in \{+, -\}$ darstellen kann, wobei $a_1 \neq 0$ ist. (Fordert man, dass es eine unendliche Teilmenge $\mathbb{N}_1 \subset \mathbb{N}$ gibt mit $a_i \neq b-1$ für $i \in \mathbb{N}_1$, 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

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

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

$$F = F(2, 24, -127, 127) \quad (1.3)$$

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.

$$M = F(2, 24, 7) = M(2, 24, 7) \quad (1.4)$$

Die Darstellung (1.3) ist aber oft präziser, da in der Praxis tatsächlich $|e_{\min}| \neq e_{\max}$ 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

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

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

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

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

$$b^{-1} \leq a \leq 1 - b^{-t} \quad (1.7)$$

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

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

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.

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

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

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

mit $\text{eps} = \frac{1}{2}b^{-t+1}$ der **maximale relative** Abstand der Zahlen $\{x \in \mathbb{R} : x_{\min} \leq |x| \leq x_{\max}\}$ zum nächstgelegenen Element aus F .

Mit der Kenntnis von eps lässt sich nun über die Bedingung

$$0.5 \cdot 10^{-n} \leq \text{eps} \leq 5 \cdot 10^{-n} \quad (1.11)$$

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 Gleitpunktzahlensystem 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 $\text{rd} : \{x \in \mathbb{R} : x_{\min} \leq |x| \leq x_{\max}\} \rightarrow \mathbb{R}$ durch

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

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 $rd(x) \in F$ und die Minimaleigenschaft*

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

Beweis. Es gilt offensichtlich

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

Nach Multiplikation mit b^e erhält man

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

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

Es wird nun die Abschätzung

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

gezeigt.

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

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

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

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

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

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

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

$$\frac{|\text{rd}(x) - x|}{|x|} \leq \frac{1}{2} b^{-t+1} = u \quad (\text{Maschinengenauigkeit}), \quad (1.14)$$

als Abschätzung für den relativen Rundungsfehler. \square

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

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

das sogenannte *Maschinenepsilon* an (Abstand von 1 zur nächsten Maschinenzahl).

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

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

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

durch

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

erklärt.

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

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

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

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

oder

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

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

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

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

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

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

also aus (1.14). □

1.6 Fehlerakkumulation

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

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

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

mit $\frac{|\Delta x|}{|x|}, \frac{|\Delta y|}{|y|} \leq \epsilon$ Zahlen aus einem Gleitpunktzahlensystem F ($\epsilon = u$ im vorliegenden Fall der Rundung, $\epsilon = 2u$ 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

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

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

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

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

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

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

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

und es bezeichne

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

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

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

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

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

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 $k - 1 \geq 1$ richtig ist, ergibt sich mit den Verabredungen

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

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

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

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. \square

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

Definition 1.12. (Landausche \mathcal{O} -Notation)

Sei $h : U \rightarrow \mathbb{R}^n$ eine Funktion, U offen, $x_0 \in U$, Dann bezeichnet das Landau-Symbol¹

$$\varphi(x) = \mathcal{O}(\|h(x)\|), \quad x \rightarrow x_0$$

¹Landau, Edmund Georg Hermann 1877-1938

eine (nicht näher spezifizierte) Funktion φ mit der Eigenschaft

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

Das Landau-Symbol

$$\varphi(x) = o(\|h(x)\|), \quad x \rightarrow x_0$$

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

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

Für Funktionen g, h sind die Notationen

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

eine abkürzende Schreibweise für

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

Beispiel 1.13.

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

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

Kapitel 2

Stabilität, Vorwärtsanalyse, Rückwärtsanalyse

Nachdem die Fehlerverstärkung bei Grundoperationen in einem GPZS betrachtet wurde, soll nun etwas allgemeiner das Problem der Fehlerfortpflanzung bei Rechenalgorithmen diskutiert werden.

Allgemein beschreibt die Stabilität die Robustheit numerischer Verfahren gegenüber Störungen in den Eingabedaten. Ein gegebenes Problem oder ein Algorithmus soll durch die Funktion

$$f : \mathbb{R}^n \rightarrow \mathbb{R}^m \quad (2.1)$$

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. $m = 1$.

3. Vor-
lesung
am
28.10.2013

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

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

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

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

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

Beispiel 2.3. f differenzierbar

$$\begin{aligned} \|f(\tilde{x}) - f(x)\| &= \|f'(x)(\tilde{x} - x) + o(\|\tilde{x} - x\|)\| \\ &\leq \underbrace{\|f'(x)\|}_{\kappa_{\text{abs}}} \cdot \|\tilde{x} - x\| + o(\|\tilde{x} - x\|) \\ \Rightarrow \frac{\|f(\tilde{x}) - f(x)\|}{\|f(x)\|} &\leq \frac{\|f'(x)\| \cdot \|x\|}{\|f(x)\|} \frac{\|\tilde{x} - x\|}{\|x\|} \\ &\quad \underbrace{\hspace{10em}}_{\kappa_{\text{rel}}} \end{aligned}$$

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

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

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

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

Weitere Vektor- und Matrixnormen

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

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

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 euklidische 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)

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

$$\|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}| \quad \text{Zeilensummennorm} \quad (2.3)$$

$$\|A\|_2 = \sqrt{\lambda_{\max}} \quad \text{mit } \lambda_{\max} \text{ größter EW von } A^T A \quad (2.4)$$

Beweis.

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

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

also $\|A\|_1 \leq \max_{j=1, \dots, n} \sum_{k=1}^m |a_{kj}|$. 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

$$\|A\|_1 \geq \|A\vec{e}_l\|_1 = \sum_{k=1}^m \left| \sum_{j=1}^n a_{kj} \delta_{jl} \right| = \sum_{k=1}^m |a_{kl}|. \quad (2.5)$$

Da l beliebig gewählt werden kann, folgt die Gleichung (2.5) für alle Spalten der Matrix A , also gilt $\|A\|_1 \geq \max_{j=1, \dots, n} \sum_{k=1}^m |a_{kj}|$ und damit (2.2).

2) Für den Nachweis von (2.3) gilt für $\vec{x} \in \mathbb{R}^n$

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

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

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

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

$$\|A\|_\infty \geq \frac{\|A\vec{x}\|_\infty}{\|\vec{x}\|_\infty} = \|A\vec{x}\|_\infty \geq \left| \sum_{j=1}^n \underbrace{a_{kj} x_j}_{=|a_{kj}|} \right| = \sum_{j=1}^n |a_{kj}|. \quad (2.6)$$

Da k als Zeilenindex frei gewählt wurde, gilt (2.6) für alle Zeilen, also gilt $\|A\|_\infty \geq \max_{k=1,\dots,m} \sum_{j=1}^n |a_{kj}|$ und damit ist (2.3) gezeigt.

Für den Nachweis von (2.4) überlegt man, dass $A^T A$ ähnlich einer Diagonalmatrix mit den EW von $A^T A$ als Diagonalelementen ist. Die EW sind nicht negativ und aus der Definition von $\|A\|_2$ folgt dann schließlich (2.4) \square

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

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

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

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

Beweis. Größtenteils trivial \square

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

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

die Beziehungen

$$\begin{aligned} |f(\tilde{x}) - f(x)| &\leq \sum_{j=1}^n \left| \frac{\partial f}{\partial x_j}(x) \right| |\tilde{x}_j - x_j| \\ &\leq \left(\sum_{j=1}^n \left| \frac{\partial f}{\partial x_j}(x) \right| \right) \max_{j=1 \dots n} |\tilde{x}_j - x_j| \quad \text{für} \quad \tilde{x} \rightarrow x \end{aligned} \quad (2.7)$$

bzw.

$$\frac{|f(\tilde{x}) - f(x)|}{|f(x)|} \leq \sum_{j=1}^n \frac{\left| \frac{\partial f}{\partial x_j}(x) \right| |x_j|}{|f(x)|} \max_{j=1 \dots n} \frac{|\tilde{x}_j - x_j|}{|x_j|} \quad \text{für} \quad \tilde{x} \rightarrow x. \quad (2.8)$$

Den Verstärkungsfaktor des max. relativen Fehlers

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

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

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

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

$$\zeta_{rel} = \frac{[|y| \ |x|] \begin{pmatrix} |x| \\ |y| \end{pmatrix}}{|x y|} = \frac{|x y| + |x y|}{|x y|} = 2$$

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,

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

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 $\|\cdot\|_\infty$ 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

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

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

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

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

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

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

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

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

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} \leq \sigma \kappa_{rel} eps \quad \text{für } eps \rightarrow 0. \quad (2.9)$$

Ein Algorithmus \tilde{f} wird **vorwärtsstabil** genannt, wenn σ kleiner oder gleich der Anzahl der hintereinander ausgeführten Elementaroperationen (Addition, Subtraktion, Multiplikation etc.) ist.

Bemerkung 2.10. Die Elementaroperationen sind vorwärtsstabil, da

$$\sigma \leq \frac{1}{\kappa_{rel}} \leq 1$$

gilt (s. Vorlesung).

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

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

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 $\kappa_f, \kappa_h, \kappa_g$ 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

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

□

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

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

2.3.2 Rückwärtsanalyse

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

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

4. Vor-
lesung
am
30.10.2013

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

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

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

$$\eta \leq \delta$$

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

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

Damit ergibt sich mit

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

für $\text{eps} \rightarrow 0$, 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

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

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

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

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 $|\epsilon_5| < 4 \text{ eps}$ gilt.

Unter Berücksichtigung von $|\epsilon_i| \leq \text{eps}$, $i = 1, \dots, 4$, findet man

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

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

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

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

Kapitel 3

Lösung linearer Gleichungssysteme

3.1 LR-Zerlegung

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

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

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 $\Rightarrow \text{rg}(\tilde{A}) = \text{rg}(A)$. Durch mehrfache Multiplikation mit

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

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

Nun etwas präziser

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

Vorraussetzung: $a_{kk} \neq 0$

Setzen

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

mit

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

e_k sei der k -te Standardbasisvektor

Definition 3.2.

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

Man überlegt sich, dass M_k das Produkt der oben diskutierten Matrizen $L_{jk}(-t_{jk})$, $j = k + 1, \dots, n$ 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}, \dots, a_{kk} bleiben bei der Multiplikation mit M_k unverändert

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

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

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

eine obere Dreiecksmatrix R . Außerdem hat die Matrix

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

die inverse Matrix

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

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

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

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

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

Bemerkung: Wir wollen hier die LR-Zerlegung so verstehen, dass der oben beschriebene Algorithmus mit den Frobeniusmatrizen M_k , $k = 1, \dots, n-1$, 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 $1, \dots, n-1$ dividiert werden musste, d.h. $r_{jj} \neq 0$, $j = 1, \dots, n-1$

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

□

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 $(k+1, k), (k+2, k), \dots, (n, k)$ wo durch die Gauß-Transformationen (Multiplikation mit M_k) Nullen erzeugt werden, können die Elemente $t_{k+1k}, t_{k+2k}, \dots, t_{nk}$ sukzessiv für $k = 1, \dots, n-2$ 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 $L = M_1^{-1} \cdot \dots \cdot M_{n-1}^{-1}$ kostet nichts, sondern besteht nur in der Ablage der jeweils bei den Gauß-Transformationen erzeugten t_{kj} -Werten ($k > j$, $k = 2, \dots, n$, $j = 1, \dots, n-1$)
- (3) Rechenaufwand ca. $\frac{n^3}{3} \in \mathcal{O}(n^3)$ Multiplikationen (flops, floating point operations).

Fehleranalyse bei der Konstruktion einer LR-Zerlegung

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

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

mit

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

Beweis. siehe Golub/van Loan, Matrix Computations □

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

$$Ly = b, \quad Rx = y,$$

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

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

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

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

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

$$M_{n-1}P_{n-1} \cdot \dots \cdot M_1P_1A = R$$

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

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

mit

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

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

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

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

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

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 letztendlich die LR-Zerlegung mit partieller Pivotisierung

$$PA = LR.$$

Bemerkung. Konsequenz dieser LR-Zerlegung mit Spaltenpivotisierung ist, dass $|\tilde{L}|$ 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.

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

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

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

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

(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 $A(1 : k, 1 : k), k = 1, \dots, n$ 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

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

und schließlich

$$M_{n-1} M_{n-2} \cdots M_1 A M_1^T \cdots M_{n-1}^T = D \quad (3.3)$$

mit einer Diagonalmatrix D .

Die Diagonalelemente d_{jj} von D sind positiv, weil einmal

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

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

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

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

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

die untere Dreiecksmatrix, mit der

$$A = GG^T$$

gilt. □

Konstruktion der Choleksy-Zerlegung

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

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

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

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

Außerdem für $j > k$

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

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

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. Singulärwerte 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 $s_{11} \geq s_{22} \geq \dots$, für die

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

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

6. Vor-
lesung
am
06.11.2013

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

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

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 verschiedenen Singulärwerte ist.

Der folgende Beweis wird nach dem Vorbild von Deuffhard/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

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

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

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

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

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

ergänzt und damit sind

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

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

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

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

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

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

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

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

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

2) Für $j = 1, \dots, r$ wird

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

gesetzt.

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

4) Man bildet die Matrizen

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

aus den orthonormalen (Spalten-) Vektoren u_1, \dots, u_m und v_1, \dots, 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 $\{v_1, \dots, v_n\}$ nach Konstruktion eine Orthonormalbasis ist.
- ii) $\{u_1, \dots, u_m\}$ ist eine Orthonormalbasis des \mathbb{R}^m , denn für $i, j = 1, \dots, r$ 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, \dots, u_r orthonormal, und nach Konstruktion der u_{r+1}, \dots, u_m ist $\{u_1, \dots, u_m\}$ damit eine Orthonormalbasis und damit U orthogonal.

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

$$\text{Ker}(A)^\perp = \text{span}\{v_1, \dots, v_r\}$$

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

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

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

Bemerkung 3.14. 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

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

3. Die Anzahl der von Null verschiedenen Singulärwerte ist gleich dem Rang r von A .
4. Die Bestimmung der Singulärwerte als Quadratwurzeln der Eigenwerte von $A^T A$ kann zu numerischen Ungenauigkeiten führen. Deswegen ist das obige Verfahren zur praktischen Berechnung der Singulärwertzerlegung nicht in allen Fällen geeignet. Andere Verfahren nutzen z. B. Umformungen mittels Householder-Matrizen.
5. Für symmetrische Matrizen A sind die Singulärwerte die Beträge der Eigenwerte. Sind alle Eigenwerte nichtnegativ, so ist die Diagonalisierung (Hauptachsentransformation)

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

auch eine Singulärwertzerlegung.

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

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

wie in der Vorlesung nachgerechnet wurde.

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

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

Die Approximationsaufgabe

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

besitzt für $k < r$ die Lösung

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

A_k ist zugleich Lösung der Approximationsaufgabe

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

Beweis. Für $k = n$ sind die Aussagen trivial, da dann $A = A_k$ gilt. Für die Aufgabe (3.9) soll der Beweis 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

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

gilt. Ohne die Allgemeinheit einzuschränken können wir annehmen, dass die Vektoren x_1, \dots, x_k orthonormal sind. Denn falls sie es nicht sind, können wir sie in eine Orthonormalbasis $\{o_1, \dots, o_k\}$ 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 $\{o_1, \dots, o_k\}$ 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 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:

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

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

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

Da die Singulärwerte in Σ absteigend geordnet sind, ist der Term (1) nicht-negativ. Außerdem ist x_i ein orthonormaler Vektor und die Matrix V ortho-

gonal, so dass der Term (2) ebenfalls nichtnegativ ist. Damit folgt

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

also der Beweis von (3.9).

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, \dots, v_n die rechten Singulärvektoren (Spalten von V) von A , dann hat der Unterraum $\mathcal{V} = \text{span}\{v_1, \dots, v_{k+1}\}$ 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

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

sowie

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

□

Ein weitere Anwendung findet die Singulärwertzerlegung bei der

$$\textbf{Aufgabe} \quad \begin{cases} \text{bestimme } x^* \text{ mit minimaler Euklidischer Länge,} \\ \text{für das } \|Ax^* - b\|_2 = \min_{x \in \mathbb{R}^n} \|Ax - b\|_2 \text{ gilt,} \end{cases} \quad (3.10)$$

für eine gegebene Matrix $A \in \mathbb{R}^{m \times n}$ und $b \in \mathbb{R}^m$. Man kann zeigen, dass die Aufgabe (3.10) 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 $\sigma_1 \geq \dots \geq \sigma_r > \sigma_{r+1} = \dots = \sigma_p = 0$, $p = \min\{m, n\}$. Durch

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

definieren wir mit A^+ die **Pseudoinverse** von A .

Dann ist $A^+ b = x^*$ die Lösung von (3.10).

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

$$\begin{aligned}
\|Ax - b\|_2^2 &= \|U^T A V V^T x - U^T b\|_2^2 = \|\Sigma \tilde{x} - \tilde{b}\|_2^2 \\
&= \sum_{i=1}^r [\sigma_i \tilde{x}_i - \tilde{b}_i]^2 + \sum_{i=r+1}^m \tilde{b}_i^2.
\end{aligned} \quad (3.11)$$

Und aus (3.11) folgt, dass

$\|Ax - b\|_2$ für x^* minimal wird, wenn $\tilde{x}_i = (V^T x^*)_i = \tilde{b}_i/\sigma_i$ für $i = 1, \dots, r$ 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 $i = r + 1, \dots, n$ gilt. Für die Lösung des Problems (3.10) mit minimaler Euklidischer Norm $x^* = A^+b$ erhält man wegen $\tilde{b}_i = (U^T b)_i$, $i = 1, \dots, r$ schließlich

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

□

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.

7. Vor-
lesung
11.11.13

Bemerkung 4.1 (Banachscher Fixpunktsatz). Ist $F : \mathcal{A} \rightarrow \mathcal{A}$, $\mathcal{A} \subset \mathbb{R}^n$ (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 $x_0 \in \mathcal{A}$ 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

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

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

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 $a_{kk} \neq 0, k = 1, \dots, n$).

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

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

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

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

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

$$\|\delta x_k\| \leq \|S^k\| \|\delta x_0\| \quad (4.6)$$

Damit konvergiert das Lösungsverfahren, wenn

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

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) *$S^k \rightarrow 0$ für $k \rightarrow \infty$*
- (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 \Rightarrow b Betrachten die verallgemeinerte Jordansche Normalform $S = T^{-1}JT$ mit einer regulären Matrix T und J mit den Jordan-Blöcken J_i

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

für die Eigenwerte $\lambda_1, \dots, \lambda_r$ von S , wobei $0 < \epsilon < 1 - |\lambda_i|$ für $i = 1, \dots, r$ gewählt wurde. Es gilt $\|S^k\| = \|TJ^kT^{-1}\|$. Die Potenzen von J enthalten für wachsendes k immer größere Potenzen von λ_i , sodass wegen $|\lambda_i| < 1$ für alle Eigenwerte $\|S^k\|$ gegen null geht.

a \Rightarrow c Mit der Zeilensummennorm gilt wegen der Voraussetzung zum Spektralradius von S , der gleich dem von J ist, und der Wahl von ϵ

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

Durch

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

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

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

und damit $\frac{\|Sx\|_T}{\|x\|_T} \leq \|J\|_\infty < 1$ für alle $x \neq 0$.

c \Rightarrow d Annahme: $S - \lambda E$ singular, 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 $|\lambda| \geq 1$. \square

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*

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

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 $U = (u_{ij})$

$$A = L + D + U \quad (4.8)$$

mit

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

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

4.2 Jacobi-Verfahren oder Gesamtschrittverfahren

Die Wahl von $B = D$ ergibt die Iterationsmatrix

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

Das Verfahren (4.4) mit der durch die Wahl von $B = D$ definierten Iterationsmatrix (4.9) heißt Jacobi-Verfahren oder Gesamtschrittverfahren.

Zur besseren Darstellung von Details der Iterationsverfahren setzen wir den Iterationsindex k nach oben in Klammern, also

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

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

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

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

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

Zur Konvergenz des Jacobi-Verfahrens gilt der

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

Beweis.

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

Zeilensummen von S

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

aufgrund der strikten Diagonaldominanz ist

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

□

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

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 $i, j \in \{1, 2, \dots, n\}$ entweder $a_{ij} \neq 0$ oder eine Indexfolge $i_1, \dots, i_s \in \{1, \dots, n\}$ existiert, sodass $a_{i_1 i_1} a_{i_1 i_2} \cdots a_{i_s j} \neq 0$ 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 $l \in \{1, \dots, n\}$ mit

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

gibt.

Bemerkung.

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

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

2. 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 $W = \{1, 2, \dots, n\}$ 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 $a_{ii} \neq 0$, $i \in W = \{1, 2, \dots, n\}$, 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.

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

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

$$\sum_{j=1}^n |b_{ij}| \leq 1, \quad i = 1, \dots, n, \quad (4.11)$$

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

$$M = |z_k| = \left| \sum_{j=1}^n b_{kj} z_j \right| \leq \sum_{j=1}^n |b_{kj}| \cdot |z_j|. \quad (4.12)$$

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

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

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, \dots, k_s , so dass

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

ist. Damit muss entweder $b_{kj} \neq 0$ oder $b_{kk_1} b_{k_1 k_2} \cdots b_{k_s j} \neq 0$ 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 $\rho(S_J) \geq 0$. Demnach existiert ein EW μ von S_J mit $|\mu| \geq 1$ und für ihn gelten

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

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 $Q = E - \mu^{-1} S_J$ und man erkennt, dass Q schwach diagonal dominant ist, denn für die Elemente von Q

$$q_{ii} = 1, \quad q_{ij} = -a_{ij}/a_{ii}, \quad (j \neq i),$$

gilt, und wegen der schwachen Diagonaldominanz von A folgt also

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

für alle i , wobei für mindestens einen Index i_0 die strikte Ungleichung gilt. Mit $|\mu^{-1}| \leq 1$ folgt die schwache Diagonaldominanz, 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. \square

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

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

Die Matrix $B = L + D$ ist eine reguläre untere Dreiecksmatrix und damit leicht 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 $x^{(0)} \in \mathbb{R}^n$*

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 bzw.

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

Wir betrachten die k -te Zeile:

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

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. \square

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

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

aufschreibt, erkennt man bei der koordinatenweisen Berechnung der neuen Iteration

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

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 $x_1^{(k)}, \dots, x_{j-1}^{(k)}$ der vorigen Iteration. Diese kennt man aber bereits. Damit kann man die Formel (4.16) für $j = 1, \dots, n$ 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

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

vornimmt, erhält man nach kurzer Rechnung durch

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

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.

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

gilt. Für $\omega > 1$ spricht man vom sukzessiven Überrelaxationsverfahren 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

$$\{0\} \subset D_1 \subset \dots \subset \mathbb{R}^n \quad (4.19)$$

zunächst eine Folge von linearen Unterräumen, die noch präzisiert wird. Im Folgenden werden Ansätze zur Bestimmung von Vektorfolgen $x_k \in D_k, k = 1, \dots$ 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 $x_1, x_2, \dots \in \mathbb{R}^n$ die Form

$$\left. \begin{array}{l} x_k \in D_k \\ Ax_k - b \in D_k^\perp \end{array} \right\} k = 1, 2, \dots \quad (4.20)$$

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

$$\left. \begin{array}{l} x_k \in D_k \\ \|Ax_k - b\|_2 \text{ minimal} \end{array} \right\} k = 1, 2, \dots \quad (4.21)$$

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

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

erklärt.

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 definiten Matrix $A \in \mathbb{R}^{n \times n}$ sind für $k = 1, 2, \dots$ 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

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

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, \dots, d_{m-1} von D_k setzt man

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

an und erhält

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

$$\Leftrightarrow \sum_{j=0}^{m-1} \langle Ad_j, d_k \rangle_2 \alpha_j = \langle b, d_k \rangle_2, \quad k = 0, \dots, m-1 \quad (4.25)$$

(4.25) ist ein lineares Gleichungssystem von m Gleichungen für die Koeffizienten $\alpha_0, \dots, \alpha_{m-1}$. 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 folgt.

Minimalität (4.22) Für $x \in D_k$ findet man

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

□

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

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, \dots, 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 $d_0, \dots, d_{m-1} \in \mathbb{R}^n \setminus \{0\}$ 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

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

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

$$x_k = \sum_{j=0}^{k-1} \alpha_j d_j \quad \text{mit} \quad \alpha_j = -\frac{\langle r_j, d_j \rangle_2}{\langle A d_j, d_j \rangle_2} \quad (4.26)$$

$$r_j := A x_j - b, \quad j \geq 1, r_0 = -b \quad (4.27)$$

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

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

(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 definiten Matrix $A \in \mathbb{R}^{n \times n}$ ist das **Verfahren des konjugierten Gradienten** gegeben durch den Ansatz (4.20) mit der speziellen Wahl

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

Dieses Verfahren bezeichnet man auch kurz als CG-Verfahren.

10.
Vorlesung
20.11.13

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

Der folgende Hilfssatz behandelt die Berechnung A -konjugierter Suchrichtungen in $K_k(A, b)$ für $k = 0, 1, \dots$.

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, \dots, 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 $d_0, \dots, d_{k-1}, -r_k \in \mathbb{R}^n$ 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 definiten Matrix $A \in \mathbb{R}^{n \times n}$ und mit den Notationen des Satzes 4.18 seien die Suchrichtungen speziell wie folgt gewählt:*

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

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

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

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

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

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 $1 \leq k \leq k_* - 1$

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 $d_0, \dots, d_{k-1}, -r_k$ linear unabhängig. Eine Gram-Schmidt-Orthogonalisierung dieser Vektoren bzgl. des Skalarproduktes $\langle \cdot, \cdot \rangle$ liefert den Vektor

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

wobei (*) aus den Eigenschaften

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

folgt, also

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

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

$$\text{span}\{d_0, \dots, d_k\} = \text{span}\{b, r_1, \dots, r_k\}$$

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

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

sodass aus Dimensionsgründen auch hier notwendigerweise Gleichheit vorliegt. \square

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 $k = 0, 1, \dots, k_*$

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*

$$\lambda_k = \frac{\|r_k\|_2^2}{\langle Ad_k, d_k \rangle_2}, \quad k = 0, 1, \dots, k_* - 1 \quad (4.33)$$

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

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) \square

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 definiten 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

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

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

Speziell gilt also

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

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

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

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

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

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

□

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 Fehlerabschätzungen:*

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

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 $x_* = \sum_{j=1}^n c_j v_j \in \mathbb{R}^n$ erhält man

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

□

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, \dots

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

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

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

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

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

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

(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). \square

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

$$x_1 = x_0 + \alpha_0 d_0 = \begin{pmatrix} \frac{3}{2} \\ \frac{3}{2} \\ \frac{3}{2} \end{pmatrix} \quad \text{und} \quad r_1 = r_0 + \alpha_0 Ad_0 = \begin{pmatrix} \frac{1}{2} \\ -1 \\ \frac{1}{2} \end{pmatrix}$$

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

$$x_2 = x_1 + \alpha_1 d_1 = \begin{pmatrix} \frac{3}{2} \\ 2 \\ \frac{3}{2} \end{pmatrix} \text{ und stellt mit } r_2 = r_1 + \alpha_1 Ad_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

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

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

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*

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

Bemerkung. Die Abkürzung “GMRES” hat ihren Ursprung in der Bezeichnung “generalized **m**inimal **r**esidual **m**ethod”

Detaillierte Konstruktionsmethoden für die Approximationen x_k beim GMRES-Verfahren werden in Plato beschrieben.

4.6 Die iterative Lösung nichtlinearer Gleichungssysteme

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

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

Der Fixpunkt soll mit der Fixpunktiteration

$$x^{(k)} = g(x^{(k-1)}), \quad k = 1, 2, \dots \quad (4.41)$$

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

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

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

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

11.
Vorlesung
25.11.13

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 $\lim_{k \rightarrow \infty} \epsilon^{(k)} = 0$ folgt die Behauptung (4.42). \square

In Vorbereitung eines effizienten Iterationsverfahrens zur Lösung nichtlinearer Gleichungen betrachten wir weiterhin den

Satz 4.28. $g : I \rightarrow I = [a, b]$ 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 $\epsilon^{(k)}$

$$\lim_{k \rightarrow \infty} \frac{\epsilon^{(k+1)}}{\epsilon^{(k)^2}} = \frac{1}{2} g''(\hat{x}), \quad (4.43)$$

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

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

Analog zum obigen Beweis ergibt der Limes $k \rightarrow \infty$ die Behauptung. \square

Lemma 4.29. Es sei $G \subset \mathbb{R}$ offen und $f : G \rightarrow \mathbb{R}$ 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 $K_{\hat{x}, \epsilon} = \{x \mid |x - \hat{x}| < \epsilon\}$ 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 $k = \max_{\xi \in D} |f'(\xi)| < 1$ \square

Bemerkung 4.30. Ist die Voraussetzung $|f'(\hat{x})| < 1$ des Hilfssatzes 4.29 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 4.31. 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

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}, \quad k = 0, 1, \dots, \quad (4.44)$$

die bei genauerem Hinsehen eine Fixpunktiteration der Funktion

$$g(x) := x - \frac{f(x)}{f'(x)} \quad \text{mit } g(\hat{x}) = \hat{x}, \quad (4.45)$$

also $x^{(k+1)} = g(x^{(k)})$ ist. Der folgende Satz liefert eine grundlegende Aussage zur Konvergenzordnung der Folge (4.44).

Satz 4.32. 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 $g : I \rightarrow I$ eine kontrahierende Abbildung ist. Für jeden Startwert $x^{(0)} \in I$ ist die Folge (4.44) mindestens quadratisch konvergent.

Beweis. Der Beweis folgt durch Anwendung von Satz 4.28 (s. auch Vorlesung!). \square

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

Satz 4.33. Sei $f : I \rightarrow \mathbb{R}$ 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

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

mit einer Konstanten C . Außerdem gilt die Fehlerabschätzung

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

Beweis. Folgt aus dem Banachschen Fixpunktsatz und dem Satz 4.29 über die Existenz einer Kontraktion.

Die quadratische Konvergenz folgt aus

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

□

Bemerkung 4.34. Die Voraussetzungen des Satzes 4.33 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 4.35. (Nullstelle einer konvexen Funktion)

Sei $f : [a, b] \rightarrow \mathbb{R}$ 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 Sekantenverfahren – Regula falsi

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

$$|x^{(1)} - \hat{x}| \leq \frac{b_0 - a_0}{2}.$$

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

$$\begin{aligned} a_k &= \begin{cases} a_{k-1} & \text{falls } f(x^{(k)})f(a_{k-1}) < 0 \\ x_k & \text{falls } f(x^{(k)})f(b_{k-1}) < 0 \end{cases}, \\ b_k &= \begin{cases} x_k & \text{falls } f(x^{(k)})f(a_{k-1}) < 0 \\ b_{k-1} & \text{falls } f(x^{(k)})f(b_{k-1}) < 0 \end{cases}, \\ x^{(k+1)} &= \begin{cases} \frac{a_{k-1} + x^{(k)}}{2} & \text{falls } f(x^{(k)})f(a_{k-1}) < 0 \\ \frac{x^{(k)} + b_{k-1}}{2} & \text{falls } f(x^{(k)})f(b_{k-1}) < 0 \end{cases} \end{aligned} \quad (4.46)$$

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

$$|x^{(k+1)} - \hat{x}| \leq \frac{b_0 - a_0}{2^{k+1}} \quad \text{bzw.} \quad |x^{(k+1)} - \hat{x}| \leq \frac{1}{2} |x^{(k)} - \hat{x}| \leq \dots \leq \frac{1}{2^{k+1}} |x^{(0)} - \hat{x}|,$$

d.h., aufgrund von $|x^{(k+1)} - \hat{x}| \leq \frac{1}{2} |x^{(k)} - \hat{x}|$ mit $p = 1$ ist die Konvergenzordnung des Verfahrens gleich 1.

Der aufwendig aussehende Algorithmus (4.46) lässt sich recht einfach implementieren. Der Vorteil des Bisektionsverfahrens besteht darin, dass es immer funktioniert, d.h., man findet immer eine Nullstelle. Allerdings findet man mit dem beschriebenen Verfahren nur eine Nullstelle.

Satz 4.36. (Bisektionsverfahren)

Sei f eine auf $[a, b]$ stetige Funktion mit $f(a)f(b) < 0$. Mit $x^{(0)} = a$ konvergiert das Bisektionsverfahren (4.46) gegen eine Nullstelle \bar{x} der Funktion f . Die Konvergenzordnung ist 1. Es gilt

$$|x^{(k)} - \hat{x}| \leq \frac{1}{2^k} |x^{(0)} - \hat{x}| = e^{-\gamma k} |x^{(0)} - \hat{x}|$$

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

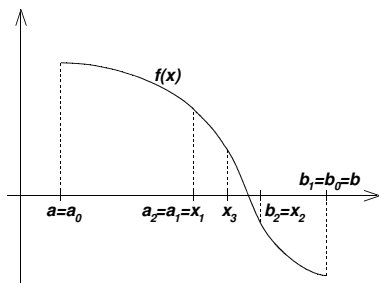


Abbildung 4.1: Bisektions-Verfahren

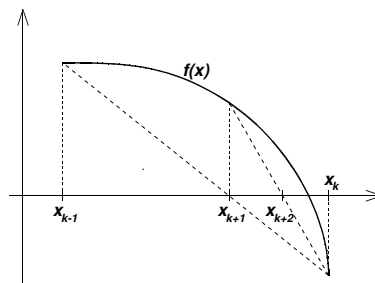


Abbildung 4.2: Sekantenverfahren – Regula falsi

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

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{\frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}} = x^{(k)} - \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})} f(x^{(k)}) \quad (4.47)$$

aus. Den Schritt (4.47) kann man einmal als genäherten Newton-Schritt mit der Approximation von $f'(x^{(k)})$ durch den Differenzenquotienten $\frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}$ interpretieren. Andererseits bedeutet (4.47) geometrisch die Berechnung des Schnittpunktes der Sekante durch die Punkte $(x^{(k-1)}, f(x^{(k-1)}))$ und $(x^{(k)}, f(x^{(k)}))$ mit der x -Achse (s. Abb. 4.2). Beide Interpretationen erklären die Namen "Regula falsi" bzw. "Sekantenverfahren". Das Sekantenverfahren hat den Nachteil, dass die neue Näherung x_{k+1} nicht unbedingt im Intervall $[a, b]$ liegen muss. Das ist nur der Fall, wenn $f(x^{(k)})f(x^{(k-1)}) < 0$ gilt. Die Modifikation von (4.47) in der Form

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{\frac{f(x^{(k)}) - f(x^{(j)})}{x^{(k)} - x^{(j)}}} = x^{(k)} - \frac{x^{(k)} - x^{(j)}}{f(x^{(k)}) - f(x^{(j)})} f(x^{(k)}) \quad (4.48)$$

mit $j \leq k - 1$ als größtem Index mit $f(x^{(k)})f(x^{(j)}) < 0$ ergibt in jedem Fall eine Näherung $x^{(k+1)} \in [a, b]$, wenn die Startwerte $x^{(0)}, x^{(1)} \in [a, b]$ die Eigenschaft $f(x^{(0)})f(x^{(1)}) < 0$ haben.

Die Modifikation (4.48) bedeutet gegenüber (4.47) nur den geringen Mehraufwand der Ermittlung des Index j . Es gilt der

Satz 4.37. (Sekantenverfahren)

Sei f eine auf $[a, b]$ stetige Funktion mit $f(a)f(b) < 0$ und $x^{(0)}, x^{(1)} \in [a, b]$ Startwerte mit der Eigenschaft $f(x^{(0)})f(x^{(1)}) < 0$. Dann konvergiert das Sekantenverfahren (4.48) gegen eine Nullstelle \hat{x} der Funktion f .

Das Sekantenverfahren (4.47) konvergiert lokal in einer Umgebung um \hat{x} mit einer Konvergenzordnung $p > 1$ schneller als das Bisektionsverfahren. Für eine um \hat{x} zweimal stetig differenzierbare Funktion f mit $f'(\hat{x}) \neq 0$ soll das gezeigt werden. In der Nähe von \hat{x} gilt

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

und mit $\Delta_k = x^{(k)} - \hat{x}$ folgt aus (4.47)

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

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

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

d.h.

$$|\Delta_{k+1}| \approx \left| \frac{f''(\hat{x})}{2f'(\hat{x})} \right| |\Delta_k| |\Delta_{k-1}| =: c |\Delta_k| |\Delta_{k-1}|. \quad (4.49)$$

Mit der Zahl $p > 0$, die der Gleichung $p - 1 = \frac{1}{p}$ genügt, also $p = \frac{1+\sqrt{5}}{2} \approx 1,618$, folgt aus (4.49) für $\tilde{\Delta}_k = c|\Delta_k|$

$$\tilde{\Delta}_{k+1} = \tilde{\Delta}_k \tilde{\Delta}_{k-1} \quad \text{und} \quad \frac{\tilde{\Delta}_{k+1}}{\tilde{\Delta}_k^p} = \left[\frac{\tilde{\Delta}_{k-1}}{\tilde{\Delta}_k} \right]^{1/p}. \quad (4.50)$$

Mit $a_k := \ln \frac{\tilde{\Delta}_{k+1}}{\tilde{\Delta}_k^p}$ erhält man durch Logarithmieren von (4.50) mit $a_k = -a_{k-1}/p$ ($k = 1, 2, \dots$) eine Folge, die für jeden Startwert a_0 wegen $p > 1$ gegen null konvergiert. Daraus folgt aufgrund der Definition von a_k und der Stetigkeit der ln-Funktion

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

Schließlich erhalten wir mit

$$|\Delta_{k+1}| \approx c^{p-1} |\Delta_k| \iff |x^{(k+1)} - \hat{x}| \approx c^{p-1} |x^{(k)} - \hat{x}|^p$$

die lokale Konvergenzordnung $p \approx 1,618$ des Sekantenverfahrens (der Wert von $p \approx 1,618$ ist übrigens das Verhältnis der zwei Teile des sogenannten **goldenen Schnitts**).

Beispiel 4.38. Zur Berechnung der Nullstelle der Funktion $f(x) = \cos x - x$ erhält man die Nullstelle $\bar{x} \approx 0,73909$ mit einer Genauigkeit von 10^{-10} mit 30 Iterationen des Bisektionsverfahrens, 5 Iterationen des Sekantenverfahrens und 4 Iterationen des Newton-Verfahrens bei der Wahl von $x_0 = 2, x_1 = 0$. Hinsichtlich der Funktionsauswertungen sind beim Bisektions- und Newton-Verfahren pro Schritt je 2 Funktionswertberechnungen erforderlich, während beim Sekantenverfahren nur eine nötig ist.

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

Satz 4.39. $F : D \rightarrow \mathbb{R}^n, D \subset \mathbb{R}^n$ sei zweimal stetig partiell diff'bar und besitze eine Nullstelle $\hat{x} \in D$. Weiterhin sei $F'(x)$ für jedes $x \in D$ regulär. Dann folgt:

Es gibt eine Umgebung U von \hat{x} , sodass die Newton-Folge

$$x^{(k+1)} = x^{(k)} - [F'(x^{(k)})]^{-1} F(x^{(k)}), \quad k = 0, 1, 2, \dots \quad (4.51)$$

von einem beliebigen Startpunkt $x^{(0)} \in U$ ausgehend gegen die Nullstelle \hat{x} konvergiert.

Die Konvergenz ist quadratisch, d.h. es gibt eine Konstante $C > 0$ mit

$$\|x^{(k)} - \hat{x}\| \leq C \|x^{(k-1)} - \hat{x}\|^2, \quad k = 1, 2, \dots$$

und es gilt die Fehlerabschätzung

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

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

Beweis. Analog zum Beweis von Satz 4.33 im eindimensionalen Fall. □

12.
Vorle-
sung
27.11.13

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

Man betrachtet

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

mit $\alpha \in]0, 1[$, und spricht hier von einem **gedämpften Newton-Verfahren**.

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

Es gilt dann mit $z^{(k+1)} = x^{(k+1)} - x^{(k)}$ das Gleichungssystem

$$F'(x^{(k)}) z^{(k+1)} = -\alpha F(x^{(k)})$$

zu lösen, und wie üblich erhält man mit

$$x^{(k+1)} = z^{(k+1)} + x^{(k)}$$

die neue Iterierte.

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

$$A = QS \quad (5.1)$$

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

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

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 $M \in \mathbb{R}^{n \times m}, 1 \leq m \leq n$,

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

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

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

liefert mit

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

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

Eine zweite Möglichkeit der Lösung des Minimumproblems (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:

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

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

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

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

$$A = QR \quad (5.6)$$

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

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

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

$$a_j = \sum_{i=1}^j r_{ij} q_i, \quad j = 1, \dots, n, \quad q_1, \dots, q_n \in \mathbb{R}^n \quad (5.7)$$

5.1 Gram-Schmidt-Verfahren zur Orthogonalisierung

- (a) Ausgangspunkt: man hat $j - 1$ orthonormale Vektoren $q_1, \dots, q_{j-1} \in \mathbb{R}^n$ mit $\text{span}(a_1, \dots, a_{j-1}) = \text{span}(q_1, \dots, q_{j-1}) =: M_{j-1}$
- (b) man bestimmt im Schritt $j \geq 1$ das Lot von a_j auf den linearen Unterraum $M_{j-1} \subset \mathbb{R}^n$

$$\hat{q}_j := a_j - \sum_{i=1}^{j-1} \langle a_j, q_i \rangle_2 q_i \quad (5.8)$$

Und nach der Normierung

$$q_j = \frac{\hat{q}_j}{\|\hat{q}_j\|_2}$$

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

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

Aus der Gleichung (5.8) folgt

$$a_j = \underbrace{\|\hat{q}_j\|_2}_{r_{jj}} q_j + \sum_{i=1}^{j-1} \underbrace{\langle a_j, q_i \rangle_2}_{r_{ij}} q_i = \sum_{i=1}^j r_{ij} q_i, \quad j = 1, \dots, n \quad (5.9)$$

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

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

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

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, \dots, q_m , da A nur m Spalten besitzt. Damit hat man zwar die obere Dreiecksmatrix 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 $\tilde{Q}_k = [q_1^T q_2^T \dots q_k^T]^T \in \mathbb{R}^{k \times n}$, $k = m, \dots, n-1$, erhält man mit etwas Arbeit die angestrebte orthogonale Matrix $Q = [q_1 \ q_2 \ \dots \ q_m]$ mit der Eigenschaft

$$A = Q \begin{bmatrix} R \\ \mathbf{0} \end{bmatrix}.$$

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 $H : \mathbb{R}^n \rightarrow \mathbb{R}^n$, $x \mapsto Hx$ mit einer Matrix

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

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

Eigenschaften von H :

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

Nachweis als Übung

Wirkung der Householder-Transformation:

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

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

gilt.

13.
Vorle-
sung
am
02.12.2013

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

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

gilt

$$\|w\|_2 = 1, \quad Hx = (E - 2ww^T)x = -\sigma e_1 \quad (5.12)$$

Beweis. $\|w\|_2 = 1$ weil $x + \sigma e_1 \neq 0$ und damit (5.11) wohldefiniert ist. Für den Nachweis von (5.12) erhält man

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

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

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

die nochmalige Nutzung von (5.11) ergibt

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

was zu zeigen war. □

Bemerkung. Um Stellenauslöschungen zu vermeiden, wird in (5.11) $\sigma = \text{sgn}(x_1) \|x\|_2$ 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

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

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

Die Matrizen der Form (5.13) erhält man für $j = 1, \dots, m - 1$ durch Transformationen der Form

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

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

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

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

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

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

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

und orthogonalem Q gegeben, dann ergibt sich mit

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

durch Nutzung der Eigenschaften von Q

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

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

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

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

5.4 Gauß-Newton-Verfahren

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

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

und gesucht ist ein funktionaler Zusammenhang

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

wobei solche Parameter

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

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

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

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

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

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

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

14.
Vorle-
sung
am
04.12.2013

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

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

bestimmt man eine neue Näherung

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

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

Schreibt man

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

in der Form

$$Ma - y$$

mit

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

auf, dann kann man die Lösung a^* von

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

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

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

erhalten (s. auch vorangegangene Vorlesungen).

Aus Effizienzgründen (jeweiliger Aufbau von y) schreibt man (5.15) auch in der Form

$$Ms - \hat{y}$$

mit

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

auf und löst das Minimum-Problem

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

und berechnet durch

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

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

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

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

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

äquivalent zur notwendigen Extremalbedingung

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

für das Funktional F ist.

Für die Abbruchbedingung gibt man eine Genauigkeit ϵ vor und bricht die Iteration dann ab, wenn

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

erfüllt ist.

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

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

durch die direkte Auswertung der notwendigen Extremalbedingung

$$\text{grad}_a F = \mathbf{0} \quad (5.17)$$

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

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

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

Beispiel:

Gegeben 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, wieviele 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. Das 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 $a = (a_1, \dots, a_p)$ 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: $(x_k, y_k), k = 0, \dots, N$ gegeben.

Gesucht: Glatte Funktion $f = f(x)$ mit

$$f(x_k) = y_k, \quad k = 0, \dots, N$$

Mögliche Ansätze für f

(i) Polynome

$$f = f(x, a_0, a_1, \dots, a_n) = a_0 + a_1x + \dots + a_nx^n, \quad n = N$$

(ii) Rationale Funktionen

$$f(x) = \frac{a_0 + a_1x + \dots + a_nx^n}{a_{n+1} + a_{n+2}x + \dots + a_{n+m+1}x^m}, \quad n = N$$

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

$$\begin{aligned} f(x) &= a_0 + a_1e^{ix} + a_2e^{2ix} + \dots + a_n e^{nix} \\ &= a_0 + a_1e^{ix} + a_2(e^{ix})^2 + \dots + a_n(e^{ix})^n \end{aligned}$$

(iv) Splines (stückweise Polynome)

V ist für paarweise verschiedene x_k regulär, d.h. a_0, \dots, a_n und damit p sind eindeutig bestimmt. □

Definition 6.3. Das nach Satz 6.2 eindeutig bestimmte Polynom p mit der Eigenschaft

$$p(x_k) = y_k, \quad k = 0, 1, \dots, n$$

für die vorgegebenen Stützstellen (x_k, y_k) heißt **Interpolationspolynom**.

6.1.1 Konstruktion des Interpolationspolynoms

Wir erinnern uns an die Generalvoraussetzung

$$x_i \neq x_j \quad \forall i, j = 0, 1, \dots, n, i \neq j$$

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

$$N_k(x) = \prod_{i=0}^{k-1} (x - x_i), \quad k = 1, \dots, n$$

mit $N_0(x) = 1$ heißen *Newton-Basispolynome*.

Satz 6.6. Die Monombasis

$$1, x, \dots, x^n$$

sowie die *Lagrange-Basispolynome*

$$L_k(x), k = 0, \dots, n$$

und die *Newton-Basispolynome*

$$N_k(x), k = 0, \dots, n$$

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

$$\begin{aligned} L_0(x) &= \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} \\ L_1(x) &= \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \\ L_2(x) &= \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} \\ L_3(x) &= \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \end{aligned}$$

und erkennen:

$$L_0(x_0) = 1, L_0(x_1) = L_0(x_2) = L_0(x_3) = 0$$

allgemein gilt:

$$L_k(x_j) = \delta_{kj}, \quad k = 0, \dots, n \quad (6.5)$$

Damit ergibt sich für das Interpolationspolynom:

$$p(x) = \sum_{k=0}^n y_k L_k(x) \quad (6.6)$$

da

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

gilt. (6.6) heißt **Lagrangsches Interpolationspolynom**.

Bemerkung 6.7. Stellt man nun die Lagrange-Interpolationspolynome der Elemente der Monombasis

$$\mathcal{M} = \{1, x, \dots, x^n\}$$

auf, dann ergibt sich

$$\sum_{k=0}^n x_k^j L_k(x) = x^j, \quad j = 0, \dots, n,$$

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

$$\kappa_m = T_{lm}^{-T} \kappa_l. \quad (6.7)$$

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 Newton-Interpolation

Bei der Lagrange-Interpolation haben wir das Interpolationspolynom in der Lagrange-Basis entwickelt. Bei der Newton-Interpolation wird das eindeutig existierende Interpolationspolynom in der Newton-Basis entwickelt.

Ansatz:

$$p(x) = \sum_{k=0}^n c_k N_k(x)$$

Durch sukzessives Vorgehen erhalten wir durch Berücksichtigung der Stützstellen $(x_k, y_k), k = 0, \dots, n$ 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 6.9.

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

heißt *Newtonsches Interpolationspolynom*.

Bemerkung. c_n ist der Koeffizient von x^n im Interpolationspolynom und c_k ist eindeutig festgelegt durch $x_0, \dots, x_k, y_0, \dots, y_k$ d.h. durch die ersten k Stützstellen.

Definition 6.10. Wir schreiben $c_k := f[x_0 x_1 \dots x_k]$ für die Abbildung

$$\{(x_0, y_0), \dots, (x_k, y_k)\} \mapsto c_k$$

Betrachtet man Teilmengen der Stützstellen

$$x_{i_0}, \dots, x_{i_k},$$

dann bezeichnet man das Interpolationspolynom an diesen Stützstellen mit

$$p_{i_0 i_1 \dots i_k}^*(x)$$

wobei i_0, \dots, i_k paarweise verschiedene Zahlen aus $\{0, \dots, k\}$ sind. Nach der Definition eines Interpolationypolynoms muss

$$p_{i_0 i_1 \dots i_k}^*(x_{i_j}) \equiv y_{i_j}, \quad j = 0, 1, \dots, k$$

15.
Vorle-
sung
am
09.12.2013

Damit gilt

$$p_k^*(x) \equiv y_k \quad (6.8)$$

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

$$p_{i_0, \dots, i_k}^*(x) = \frac{(x - x_{i_0})p_{i_1 \dots i_k}^*(x) - (x - x_{i_k})p_{i_0 \dots i_{k-1}}^*(x)}{x_{i_k} - x_{i_0}} \quad (6.9)$$

Beweis. Induktion Die beiden rechts stehenden Polynome in (6.9) haben einen Grad $\leq k - 1$ (damit der gesamte Ausdruck einen Grad $\leq n$).

Anfang ($k = 1$) ist trivial wegen (6.8).

Es ist zu zeigen, dass das rechts in (6.9) stehende Polynom das Interpolationpolynom zu den Stützstellen x_{i_0}, \dots, x_{i_k} ist (Ausdruck rechts von (6.9) bezeichnen wir mit $q(x)$) $\deg q(x) \leq k$ ist offensichtlich. Weiter ist

$$q(x_{i_0}) = \frac{0 - (x_{i_0} - x_{i_k})p_{i_0 \dots i_{k-1}}^*(x_{i_0})}{x_{i_k} - x_{i_0}} = y_{i_0}$$

und analog

$$q(x_{i_k}) = y_{i_k}$$

Schließlich für die restlichen Stützstellen $1 \leq j \leq k - 1$

$$\begin{aligned} q(x_{i_j}) &= \frac{(x_{i_j} - x_{i_0})p_{i_1 \dots i_k}^*(x_{i_j}) - (x_{i_j} - x_{i_k})p_{i_0 \dots i_{k-1}}^*(x_{i_j})}{x_{i_k} - x_{i_0}} \\ &= \frac{(x_{i_j} - x_{i_0})y_{i_j} - (x_{i_j} - x_{i_k})y_{i_j}}{x_{i_k} - x_{i_0}} = y_{i_j} \end{aligned}$$

Damit erfüllt q die Interpolationsbedingung $q(x_{i_j}) = y_{i_j}, j = 0, \dots, k$ also genau das, was $p_{i_0 \dots i_k}^*(x)$ leistet. Aufgrund der Eindeutigkeit des Interpolationpolynoms gilt also

$$q = p_{i_0 \dots i_k}^*$$

□

Satz 6.12. *Es gilt*

$$f[x_{i_0} \dots x_{i_k}] = \frac{f[x_{i_1} \dots x_{i_k}] - f[x_{i_0} \dots x_{i_{k-1}}]}{x_{i_k} - x_{i_0}}$$

Beweis. Nach der Definition von $f[\dots]$ ist dies gerade der Koeffizient von der höchsten Potenz des Interpolationspolynoms. Betrachten (6.9). Das Polynom links hat in der höchsten Potenz den Term $f[x_{i_0} \dots x_{i_k}]x^k$, das rechts stehende hat in der höchsten Potenz

$$\frac{x \cdot f[x_{i_1} \dots x_{i_k}]x^{k-1} - x \cdot f[x_{i_0} \dots x_{i_{k-1}}]x^{k-1}}{x_{i_k} - x_{i_0}}$$

\rightsquigarrow Behauptung. □

Als Folgerung des Satzes 6.12 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 6.13. Stellt man jetzt die Newton-Interpolationspolynome der Elemente der Monombasis

$$\mathcal{M} = \{1, x, \dots, x^n\}$$

auf, dann ergibt sich

$$\sum_{k=0}^n c_k N_k(x) = x^j = \sum_{k=0}^j c_k N_k(x), \quad j = 0, \dots, n,$$

also kann man die Monombasis durch

$$\begin{pmatrix} f_0[x_0] & 0 & 0 & \dots & 0 \\ f_1[x_0] & f_1[x_0x_1] & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_n[x_0] & f_n[x_0x_1] & f_n[x_0x_1x_2] & \dots & f_n[x_0 \dots x_n] \end{pmatrix} \begin{pmatrix} N_0(x) \\ N_1(x) \\ \vdots \\ N_n(x) \end{pmatrix} =: T_{nm} \begin{pmatrix} N_0(x) \\ N_1(x) \\ \vdots \\ N_n(x) \end{pmatrix} = \begin{pmatrix} 1 \\ x^1 \\ \vdots \\ x^n \end{pmatrix}$$

darstellen, wobei die Koeffizienten $f_k[x_0 \dots]$ 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

$$\mathcal{N} = \{N_0(x), N_1(x), \dots, N_n(x)\}$$

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

$$\kappa_m = T_{nm}^{-T} \kappa_n. \quad (6.10)$$

Beispiel 6.14. 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

$$T_{nm}^T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{zu} \quad \kappa_m = T_{nm}^{-T} \kappa_n \iff \kappa_n = T_{nm}^T \kappa_m = \begin{pmatrix} 1 \\ -1 \\ \frac{5}{2} \end{pmatrix},$$

also ergibt sich das Polynom

$$p(x) = 1 - 1(x - 0) + \frac{5}{2}(x - 0)(x - 1) = 1 - x + \frac{5}{2}x(x - 1).$$

6.3.1 Algorithmische Aspekte der Polynominterpolation - Horner-Schema

Für die Berechnung eines Polynoms in der Form

$$p(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$$

Werden $1 + 2 + \dots + n = \frac{n(n+1)}{2}$ Multiplikationen und n Additionen benötigt.
Also $\mathcal{O}(n^2)$ flops

$$p(x) = a_0 + x(a_1 + x(a_2 + \dots)) = (\dots (a_n x + a_{n-1})x + a_{n-2})x + \dots + a_1)x + a_0$$

$\rightsquigarrow n$ 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

$$p(x) = c_0 + (x - x_0)(c_1 + (x - x_1)(c_2 + \dots + c_n(x - x_{n-1}))) \cdots$$

geschrieben werden. Daraus resultiert der Algorithmus:

Algorithmus 2 Wertet Newton Polynom mittels Horner-Schema aus

```
un+1 = 0
for k = n downto 0 do
    uk = (x - xk)uk+1 + ck
end for
p(x) = u0
```

Mit Laufzeit $3n$ flops.

6.3.2 Verfahren von Neville und Aitken

Es ist vergleichbar mit der Herangehensweise bei der Newton-Interpolation
Aus Lemma 6.11 folgt mit

$$\begin{aligned}y_0 &=: p_0^*(x) \\ &\vdots \\ y_n &=: p_n^*(x)\end{aligned}$$

16.
Vorle-
sung
am
11.12.2013

die Rekursion

$$\begin{aligned}p_{0,1}^*(x) &= \frac{(x - x_0)p_1^*(x) - (x - x_1)p_0^*(x)}{x_1 - x_0} \\ &\vdots \\ p_{n-1,n}^*(x) &= \frac{(x - x_{n-1})p_n^*(x) - (x - x_n)p_{n-1}^*(x)}{x_n - x_{n-1}} \\ &\text{usw.} \\ p_{0,1,2}^*(x) &= \frac{(x - x_0)p_{1,2}^*(x) - (x - x_2)p_{0,1}^*(x)}{x_2 - x_0}\end{aligned}$$

Für den Algorithmus von Neville und Aitken folgt das Schema zur Berechnung von p an der Stelle x

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

Beispiel.

x_k	0	1	3
y_k	1	3	2

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

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

6.4 Fehlerabschätzung der Polynominterpolation

Handelt es sich bei den Stützpunkten (x_k, y_k) nicht um diskrete Messwerte, sondern um die Wertetabelle einer gegebenen Funktion $f(x)$, dann ist der Fehler $f(x) - p_n(x)$, den man bei der Interpolation macht, von Interesse.

Nimmt man zu den Stützwerten x_0, \dots, x_n den Wert $x = x_{n+1}$ hinzu, ergibt die Interpolationsbedingung $y = f(x) = p_{n+1}(x)$

$$\underbrace{p_{n+1}(x)}_{p_{n+1}(x_{n+1})=y_{n+1}} = f(x) = p_n(x) + f[x_0x_1 \dots x_n, x] \prod_{k=0}^n (x - x_k)$$

bzw.

$$f(x) - p_n(x) = f[x_0x_1 \dots x_nx](x - x_0)(x - x_1) \cdots (x - x_n) \quad (6.11)$$

Der folgende Satz liefert die Grundlage für die Abschätzung des Interpolationsfehlers (6.11).

Satz 6.15. Sei $]a, b[=]\min_{0 \leq j \leq n} x_j, \max_{0 \leq j \leq n} x_j[$ und sei $p_n(x)$ das Interpolationspolynom zur Wertetabelle $(x_k, f(x_k))$ der $(n+1)$ -mal stetig differenzierbaren Funktion f auf $[a, b]$, wobei die Stützstellen x_k paarweise verschieden sind.

Dann gibt es für jedes $\tilde{x} \in]a, b[$ einen Zwischenwert $\xi = \xi(x_0, \dots, x_n) \in]a, b[$ mit

$$f(\tilde{x}) - p_n(\tilde{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)!} (\tilde{x} - x_0) \cdots (\tilde{x} - x_n)$$

Beweis. Siehe Vorlesung oder Bärwolff, entscheidend genutzt wird das Verhalten der Hilfsfunktion

$$\Phi(x) := f(x) - p_n(x) - c(\tilde{x}) \prod_{k=0}^n (x - x_k) \quad \text{mit} \quad c(\tilde{x}) = \frac{f(\tilde{x}) - p_n(\tilde{x})}{\prod_{k=0}^n (\tilde{x} - x_k)},$$

die \tilde{x} als Nullstelle hat und damit mit den Nullstellen x_k , $k = 0, \dots, n$ insgesamt $n + 2$ Nullstellen hat. Die sukzessive Anwendung des Satzes von Rolle auf die Ableitungen $\Phi^{(j)}(x)$, $j = 0, \dots, n$ beweist den Satz. \square

Aus dem Satz 6.15 folgt direkt für eine $(n + 1)$ -mal stetig differenzierbare Funktion die Fehlerabschätzung

$$|f(x) - p_n(x)| \leq \frac{\max_{\xi \in [a,b]} |f^{(n+1)}(\xi)|}{(n+1)!} \underbrace{\left| \prod_{k=0}^n (x - x_k) \right|}_{=: w(x)} \quad (6.12)$$

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

$$x_k^* = \frac{a+b}{2} + \frac{b-a}{2} \cos\left(\frac{2(n+1-k)-1}{2(n+1)}\pi\right), \quad k = 0, \dots, n \quad (6.13)$$

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 6.16. *Seien x_k äquidistante und x_k^* gemäß (6.13) 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

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

mit p_k^* als dem Interpolationspolynom, das die Interpolationsbedingung

$$p_k^*(x_j^*) = f(x_j^*), \quad j = 0, \dots, k,$$

erfüllt.

6.5 Hermite-Interpolation

Hat man einen Stützpunkt (x_0, y_0) vorgegeben, so ist damit ein Polynom 0-ten Grades festgelegt (Gerade parallel zur x -Achse). Hat man an der Stelle noch eine Ableitungsinformation, d.h. (x_0, y_0') , dann ist damit eine Gerade durch den Punkt (x_0, y_0) mit dem Anstieg y_0' festgelegt, also ein Polynom 1-ten Grades.

Satz 6.17. Sei f eine $(n+1)$ -mal stetig diff'bare Funktion in einem Intervall um den Punkt x . Dann gilt

$$\lim_{x_0 \rightarrow x \dots x_n \rightarrow x} f[x_0 x_1 \dots x_n x] = \frac{f^{(n+1)}(x)}{(n+1)!}$$

Beweis. vollständige Induktion, MWS □

Der Satz 6.17 rechtfertigt

Definition 6.18.

$$f[\underbrace{x, x, \dots, x}_{n+2}] = \frac{f^{(n+1)}(x)}{(n+1)!} \quad (6.14)$$

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	$f[x_0]$	$f[x_0 x_0]$			
x_0	$f[x_0]$	$f[x_0 x_0]$	$f[x_0 x_0 x_0]$		
x_1	$f[x_1]$	$f[x_0 x_1]$	$f[x_0 x_0 x_1]$	$f[x_0 x_0 x_0 x_1]$	
x_1	$f[x_1]$	$f[x_1 x_1]$	$f[x_0 x_1 x_1]$	$f[x_0 x_0 x_1 x_1]$	$f[x_0 x_0 x_0 x_1 x_1]$

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

$$c_0 = f[x_0], \quad c_1 = f[x_0x_0], \quad c_2 = f[x_0x_0x_0], \quad c_3 = f[x_0x_0x_0x_1], \quad c_4 = f[x_0x_0x_0x_1x_1].$$

Führt man nun

$$\tilde{x}_0 = x_0, \quad \tilde{x}_1 = x_0, \quad \tilde{x}_2 = x_0, \quad \tilde{x}_3 = x_1, \quad \tilde{x}_4 = x_1$$

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

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

6.6 Spline-Interpolation

Problem bei der Polynom-Interpolation:

Eventuell große Oszillationen durch Polynome höheren Grades bei Stützpunktzahlen ≥ 10

Deshalb:

17.

Vorle-

sung

am

16.12.2013

Statt eines Interpolationspolynoms konstruiert man für $(x_k, y_k), k = 0, 1, \dots, n$ in jeden Teilintervall einzelne Polynome, die an den Randstellen glatt ineinander übergehen. Betrachten mit

$$\Delta = \{a = x_0 < x_1 < \dots < x_n = b\}$$

eine fest gewählte Zerlegung von $[a, b]$, wobei die Stützstellen x_0, \dots, x_n auch als Knoten bezeichnet werden.

Definition 6.19. 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

$$s(x_k) = f_k \quad \text{für } k = 0, 1, \dots, n \quad (6.15)$$

für $(x_k, f_k), k = 0, 1, \dots, n$

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

$$a_k = f_k, \quad b_k = \frac{f_{k+1} - f_k}{x_{k+1} - x_k}, \quad k = 0, \dots, n-1$$

Satz 6.20.

(a) Zur Zerlegung $\Delta = a = x_0 < \dots < x_n = b$ und f_0, \dots, f_n gibt es genau einen Spline $s \in S_{\Delta, 1}$ mit der Eigenschaft (6.15)

(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 $h_{\max} := \max_{k=0, \dots, n-1} (x_{k+1} - x_k)$

Beweis. (a) nach Konstruktion

(b) Für jedes $k \in 1, \dots, n$ stimmt s auf $[x_{k-1}, x_k]$ mit demjenigen $p \in \Pi_1$ überein, für das $p(x_{k-1}) = f(x_{k-1})$ und $p(x_k) = f(x_k)$ gilt. Der Fehler bei der Polynominterpolation (Satz 6.15) liefert

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

6.6.2 Kubische Splines

Betrachte nun $S_{\Delta,3}$, und verwenden

$$\|u\|_2 := \left(\int_a^b |u(x)|^2 dx \right)^{\frac{1}{2}}$$

Lemma 6.21. Wenn eine Funktion $f \in C^2[a, b]$ und eine kubische Splinefunktion $s \in S_{\Delta,3}$ in den Knoten übereinstimmen, d.h.

$$s(x_k) = f(x_k) \quad \text{für } k = 0, \dots, n$$

so gilt

$$\|f'' - s''\|_2^2 = \|f''\|_2^2 - \|s''\|_2^2 - 2([f' - s']s'')(x) \Big|_{x=a}^{x=b} \quad (6.16)$$

Beweis.

$$\begin{aligned} \|f'' - s''\|_2^2 &= \int_a^b |f''(x) - s''(x)|^2 dx = \|f''\|_2^2 - 2 \int_a^b (f''s'')(x) dx + \|s''\|_2^2 \\ &= \|f''\|_2^2 - 2 \int_a^b ([f'' - s'']s'')(x) dx - \|s''\|_2^2 \end{aligned}$$

Für den mittleren Term ergibt die partielle Integration

$$\begin{aligned}
& \int_{x_{k-1}}^{x_k} ([f'' - s'']s'')(x)dx \\
&= ([f' - s']s'')(x) \Big|_{x_{k-1}}^{x_k} - \int_{x_{k-1}}^{x_k} ([f' - s']s''')(x)dx \\
&= ([f' - s']s'')(x) \Big|_{x_{k-1}}^{x_k} - \underbrace{([f - s]s''')(x) \Big|_{x_{k-1}}^{x_k}}_{=0} + \underbrace{\int_{x_{k-1}}^{x_k} ([f - s]s^{(4)})(x)dx}_{=0}
\end{aligned}$$

Die Summation über $k = 1, \dots, n$ ergibt

$$\begin{aligned}
\int_a^b ([f'' - s'']s'')(x)dx &= \sum_{k=1}^n \{([f' - s']s'')(x_k) - ([f' - s']s'')(x_{k-1})\} \\
&= ([f' - s']s'')(b) - ([f' - s']s'')(a)
\end{aligned}$$

□

Satz 6.22. Gegeben sei $f \in C^2[a, b]$ und ein kubischer Spline $s \in S_{\Delta,3}$ mit $s(x_k) = f(x_k), k = 0, \dots, n$. Dann gilt die Identität

$$\|f''\|_2^2 - \|s''\|_2^2 = \|f'' - s''\|_2^2 \quad (6.17)$$

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 (6.16) □

Korollar 6.1. Zu gegebenen Werten $f_0, \dots, f_n \in \mathbb{R}$ 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 $k = 0, \dots, n$

Beweis. Folgt direkt aus (6.17) □

6.6.3 Berechnung interpolierender kubischer Splines

Lokaler Ansatz

$$s(x) = a_k + b_k(x - x_k) + c_k(x - x_k)^2 + d_k(x - x_k)^3, \quad (6.18)$$

$$x \in [x_k, x_{k+1}], k = 0, \dots, n - 1$$

für $s : [a, b] \rightarrow \mathbb{R}$, wobei $s(x) =: p_k(x)$ auf dem Intervall $[x_k, x_{k+1}]$ verabredet wird.

Aufgabe: Bestimmung von $a_k, \dots, d_k, k = 0, \dots, n - 1$ so, dass s auf $[a, b]$ zweimal stetig differenzierbar ist und darüberhinaus in den Knoten vorgegebene Werte $f_0, \dots, f_n \in \mathbb{R}$ interpoliert

$$s(x_k) = f_k, \quad k = 0, \dots, n$$

Setzen $h_k := x_{k+1} - x_k, k = 0, \dots, n$

Lemma 6.23. Falls $n + 1$ reelle Zahlen $s''_0, \dots, s''_n \in \mathbb{R}$ den folgenden $n - 1$ gekoppelten Gleichungen ($k = 1, \dots, n - 1$)

$$h_{k-1} \underbrace{s''_{k-1}}_{m_{k-1}} + 2(h_{k-1} + h_k) \underbrace{s''_k}_{m_k} + h_k \underbrace{s''_{k+1}}_{m_{k+1}} = \underbrace{6 \frac{f_{k+1} - f_k}{h_k} - 6 \frac{f_k - f_{k-1}}{h_{k-1}}}_{c_k} \quad (6.19)$$

genügen, so liefert der lokale Ansatz (6.18) 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 $k = 0, \dots, n - 1$ 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, \dots, m_n stimmen mit den 2. Ableitungen der Splinefunktion s in den Knoten x_k überein

$$s''_k = m_k = s''(x_k), \quad k = 0, \dots, n$$

(6.19) bedeutet: Es liegen $n - 1$ Bedingungen für $n + 1$ Momente vor, d.h. es gibt 2 Freiheitsgrade. Diese werden durch die folgenden Randbedingungen festgelegt:

- Natürliche RB $s''_0 = s''_n = 0$
- Vollständige RB $s'_0 = f'_0, s'_n = f'_n$ für geg. $f'_0, f'_n \in \mathbb{R}$
- Periodische RB $s'_0 = s'_n, s''_0 = s''_n$

(diese Festlegungen korrelieren mit den Bedingungen (a), (b), (c) des Satzes 6.22)

6.6.4 Gestalt der Gleichungssysteme

Natürliche Randbedingungen

$$\begin{bmatrix} 2(h_0 + h_1) & h_1 & & 0 \\ h_1 & 2(h_1 + h_2) & \ddots & \\ & \ddots & & h_{n-2} \\ 0 & & h_{n-2} & 2(h_{n-2} + h_{n-1}) \end{bmatrix} \begin{bmatrix} m_1 \\ \\ \\ m_{n-1} \end{bmatrix} = \begin{bmatrix} c_1 \\ \vdots \\ c_{n-1} \end{bmatrix} \quad (6.20)$$

Vollständige Randbedingungen

$$\begin{bmatrix} 2h_0 & h_0 & & 0 \\ h_0 & 2(h_0 + h_1) & \ddots & \\ & \ddots & 2(h_{n-2} + h_{n-1}) & h_{n-1} \\ 0 & & h_{n-1} & 2h_{n-1} \end{bmatrix} \begin{bmatrix} m_0 \\ \\ \\ m_n \end{bmatrix} = \begin{bmatrix} c_0 \\ \vdots \\ c_n \end{bmatrix} \quad (6.21)$$

Dieses Gleichungssystem erhält man durch die Beziehungen

$$s'(x_0) = p'_0(x_0) = b_0 = \frac{f_1 - f_0}{h_0} - \frac{h_0}{6}(m_1 + 2m_0) \quad (6.22)$$

$$s'(x_n) = p'_{n-1}(x_n) = \frac{f_n - f_{n-1}}{h_{n-1}} + \frac{h_{n-1}}{6}(m_{n-1} + 2m_n), \quad (6.23)$$

woraus sich mit $s'(x_0) = f'_0$ bzw. $s'(x_n) = f'_n$ die beiden Gleichungen

$$2h_0m_0 + h_0m_1 = -6s'(x_0) + 6\frac{f_1 - f_0}{h_0} = -6f'_0 + 6\frac{f_1 - f_0}{h_0} =: c_0$$

$$h_{n-1}m_{n-1} + 2h_{n-1}m_n = 6s'(x_n) - 6\frac{f_n - f_{n-1}}{h_{n-1}} = 6f'_n - 6\frac{f_n - f_{n-1}}{h_{n-1}} =: c_n$$

und damit (6.21) ergeben.

periodische Randbedingungen

$$\begin{bmatrix} 2(h_{n-1} + h_0) & h_0 & & h_{n-1} \\ h_0 & 2(h_0 + h_1) & \ddots & \\ & \ddots & & h_{n-2} \\ h_{n-1} & & h_{n-2} & 2(h_{n-2} + h_{n-1}) \end{bmatrix} \begin{bmatrix} m_0 \\ \\ \\ m_{n-1} \end{bmatrix} = \begin{bmatrix} c_0 \\ \vdots \\ c_{n-1} \end{bmatrix} \quad (6.24)$$

Die erste Gleichung des Systems ergibt sich unter Nutzung der periodischen Bedingungen $m_0 = m_n$ bzw. $s'_0 = s'_n$ und der Beziehungen (6.22), (6.23) 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 (6.24) 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 $A = (a_{ij}) \in \mathbb{R}^{n \times n}$

$$\sum_{k \neq j=1}^n |a_{kj}| < |a_{kk}|, \quad k = 1, \dots, n. \quad (6.25)$$

Lemma 6.24. (s. dazu auch Kapitel 4)

Jede strikt diagonal dominante Matrix $A = (a_{kj}) \in \mathbb{R}^{n \times n}$ ist regulär und es gilt

$$\|x\|_\infty \leq \max_{k=1, \dots, n} \left\{ (|a_{kk}| - \sum_{k \neq j=1}^n |a_{kj}|)^{-1} \right\} \|Ax\|_\infty, \quad x \in \mathbb{R}^n \quad (6.26)$$

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

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

Dies liefert die Gültigkeit von (6.26) woraus die Regularität von A direkt folgt. (Aus $Ax = 0$ folgt $x = 0$ als einzige Lösung) \square

Beweis. (Beweisskizze)

Die Aussage des Lemmas wird unter Nutzung einer Beziehung, der Form

$$B \begin{bmatrix} f''(x_1) - s''_1(x_1) \\ \vdots \\ f''(x_{n-1}) - s''_{n-1} \end{bmatrix} = \begin{bmatrix} \delta_1 - \hat{\delta}_1 \\ \vdots \\ \delta_{n-1} - \hat{\delta}_{n-1} \end{bmatrix} \quad (6.29)$$

nachgewiesen, die man durch Taylorentwicklungen von f'' und f erhält, wobei δ_j und $\hat{\delta}_j$ jeweils von der Ordnung $\mathbf{O}(h_{max}^2)$, $h_{max} = \max_{k=0, \dots, n-1} x_{k+1} - x_k$, 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

$$|b_{kk}| - \sum_{k \neq j=1}^{n-1} |b_{kj}| = \frac{2}{3} - \frac{h_k}{3(h_{k+1} + h_k)} - \frac{h_{k+1}}{3(h_{k+1} + h_k)} = \frac{1}{3}, \quad k = 2, \dots, n-2,$$

erhält man letztendlich die Beziehung (6.28) (die erste und die letzte Gleichung des Systems (6.29) sind auf Grund der Randbedingungen trivial). \square

Das Lemma 6.25 ist die Grundlage für den folgenden Satz zur Fehlerabschätzung der Spline-Interpolation

Satz 6.26. *Sei $f \in C^4[a, b]$ und sei $s \in S_{\Delta,3}$ ein interpolierender kubischer Spline. Weiter bezeichne $h_k = x_{k+1} - x_k$ für $k = 0, \dots, n-1$ und*

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

Falls

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

erfüllt ist mit einer Konstanten $C > 0$, so gelten mit der Zahl $c := \frac{h_{\max}}{h_{\min}} (C + \frac{1}{4})$ die folgenden Abschätzungen für jedes $x \in [a, b]$

$$|s(x) - f(x)| \leq c \|f^{(4)}\|_\infty h_{\max}^4 \quad (6.30)$$

$$|s'(x) - f'(x)| \leq c \|f^{(4)}\|_\infty h_{\max}^3 \quad (6.31)$$

$$|s''(x) - f''(x)| \leq c \|f^{(4)}\|_\infty h_{\max}^2 \quad (6.32)$$

$$|s^{(3)}(x) - f^{(3)}(x)| \leq c \|f^{(4)}\|_\infty h_{\max} \quad (6.33)$$

Beweis. Zuerst wird (6.33) nachgewiesen. s'' ist als 2. Ableitung eines Polynoms 3. Grades affin linear auf $[x_k, x_{k+1}]$ für $k = 0, \dots, n-1$, d.h.

$$s^{(3)}(x) \equiv \frac{s''(x_{k+1}) - s''(x_k)}{h_k} = \text{const}, \quad x_k \leq x \leq x_{k+1} \quad (6.34)$$

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

$$f^{(3)}(x) = \frac{f''(x_{k+1}) - f''(x_k)}{h_k} - \frac{(x_{k+1} - x)^2}{2h_k} f^{(4)}(\alpha_k) + \frac{(x - x_k)^2}{2h_k} f^{(4)}(\beta_k) \quad (6.35)$$

für gewisse Zwischenstellen $\alpha_k, \beta_k \in [a, b]$. Subtraktion von (6.34) und (6.35) ergibt

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

$$\begin{aligned} \rightsquigarrow & |s^{(3)}(x) - f^{(3)}(x)| \\ & \leq \|f^{(4)}\|_\infty \frac{1}{\min\{h_0, \dots, h_{n-1}\}} (ch_{\max}^2 + ch_{\max}^2 + \frac{h_{\max}^2}{2}) \\ & \leq \frac{h_{\max}}{h_{\min}} (2C + \frac{1}{2}) \|f^{(4)}\|_\infty h_{\max} = 2c \|f^{(4)}\|_\infty h_{\max} \end{aligned}$$

wobei

$$\begin{aligned} (x_{k+1} - x)^2 + (x - x_k)^2 &= (x_{k+1} - x_k)^2 - 2(x_{k+1} - x)(x - x_k) \\ &\leq (x_{k+1} - x_k)^2 \leq h_{\max}^2 \quad \forall x \in [x_k, x_{k+1}] \end{aligned}$$

berücksichtigt wurde.

Die restlichen Fehlerabschätzungen (6.32), (6.31), (6.30) erhält man durch sukzessive Integration von (6.33) unter Nutzung des Hauptsatzes der Differential- und Integralrechnung. \square

Bemerkung. Die wesentliche Voraussetzung des eben bewiesenen Satzes über den Fehler der 2. Ableitungen in den Knoten ist typischerweise erfüllt (siehe auch Hilfssatz 6.25 für den Fall natürlicher Randbedingungen).

6.9 Trigonometrische Interpolation

Werden periodische Vorgänge “gemessen” oder vermutet man, dass gegebene Stützpunkte zu einer periodischen Funktion gehören, dann bietet sich eine

Interpolation durch trigonometrische Funktionen an. O.B.d.A. nehmen wir als periode $T = 2\pi$ an und betrachten das Intervall $[0, 2\pi]$ (sonst Transformation)

Zerlegung:

$$\Delta = \{0 = x_0 < \dots < x_{n-1} < 2\pi\}$$

mit $x_k = \frac{k}{n}2\pi, k = 0, \dots, n-1$

Es wird folgender trigonometrischer Ansatz gemacht:

$$\Psi(x) = \begin{cases} \frac{A_0}{2} + \sum_{l=1}^m (A_l \cos(lx) + B_l \sin(lx)), & n = 2m + 1 \\ \frac{A_0}{2} + \sum_{l=1}^{m-1} (A_l \cos(lx) + B_l \sin(lx)) + \frac{A_m}{2} \cos(mx), & n = 2m \end{cases} \quad (6.36)$$

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

$$\Psi(x_k) = f_k, \quad k = 0, \dots, n-1 \quad (6.37)$$

mit gegebenen Werten $f_k \in \mathbb{R}$ erfüllen, wobei die Koeffizienten A_l, B_l gesucht sind.

Man kann zwar A_l, B_l aus (6.36) durch Auswertung von (6.37) bestimmen, aber im Komplexen wird es übersichtlicher. Mit

$$\cos \phi = \frac{1}{2}(e^{i\phi} + e^{-i\phi}), \quad \sin \phi = \frac{1}{2i}(e^{i\phi} - e^{-i\phi})$$

folgt nämlich:

$$\cos lx_k = \frac{1}{2}(e^{ilx_k} + e^{-ilx_k}) = \frac{1}{2} \left(\left(e^{\frac{2\pi il}{n}} \right)^k + \left(e^{-\frac{2\pi il}{n}} \right)^k \right), \quad x_k = \frac{2\pi k}{n}$$

bzw.

$$\sin lx_k = \frac{1}{2i} \left(\left(e^{\frac{2\pi il}{n}} \right)^k - \left(e^{-\frac{2\pi il}{n}} \right)^k \right) \quad (6.38)$$

Bemerkung. Wegen der 2π -Periodizität von $e^{i\phi}$ gilt

$$e^{\frac{-2\pi l}{n}i} = e^{\left(\frac{-2\pi l}{n} + 2\pi\right)i} = e^{\left(-\frac{2\pi l}{n} + \frac{2\pi n}{n}\right)i} = e^{\frac{(n-l)2\pi}{n}i}$$

Also brauchen keine negativen Potenzen betrachtet zu werden, sondern nur Terme

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

(6.38) wird in den Ansatz (6.36) eingesetzt, etwas umgeordnet, sodass man mit

$$p(x) = \beta_0 + \beta_1 e^{ix} + \dots + \beta_{n-1} e^{i(n-1)x} \quad (6.39)$$

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

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

(Ψ und p stimmen nur in den Stützstellen x_k überein, allerdings gilt $\Psi(x) = p(x)$ nicht für beliebige x).

Für die Beziehungen zwischen β_k und A_k, B_k ergeben sich einfache Formeln, z.B. für $n = 2m + 1$

$$\begin{aligned} \beta_0 &= \frac{A_0}{2}, & \beta_j &= \frac{1}{2}(A_j - iB_j), & \beta_{n-j} &= \frac{1}{2}(A_j + iB_j), & j &= 1, \dots, m \\ A_0 &= 2\beta_0, & A_l &= \beta_l + \beta_{n-l}, & B_l &= i(\beta_l - \beta_{n-l}), & l &= 1, \dots, m \end{aligned}$$

Setzt man $\omega = e^{ix}$, so folgt

$$p(x) = \beta_0\omega^0 + \beta_1\omega^1 + \dots + \beta_{n-1}\omega^{n-1} =: P(\omega) \quad (6.40)$$

Und P ist tatsächlich Polynom in ω .

Definition 6.27.

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

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

Satz 6.28. Zu beliebigen Stützstellen $(x_k, f_k), k = 0, \dots, n-1, f_k \in \mathbb{C}, x_k = k\frac{2\pi}{n}$ gibt es genau ein trigonometrisches Polynom der Form (6.40) mit

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

Dabei gelten die wichtigen Beziehungen

$$\begin{aligned} (i) \quad & \omega_k^j = \omega_j^k, \quad \omega_k^{-l} = \overline{\omega_k^l} \\ (ii) \quad & \sum_{k=0}^{n-1} \omega_k^j \omega_k^{-l} = \begin{cases} n & j = l \\ 0 & j \neq l, 0 \leq l, j \leq n-1 \end{cases} \end{aligned}$$

Beweis. Die Existenz des Polynoms und die Eindeutigkeit folgt analog dem Nachweis der Existenz und Eindeutigkeit der allgemeinen reellen Polynominterpolation (z.B. Lagrange-Interpolation)

zu (i) nach Definition

19.
Vorle-
sung
am
06.01.2014

zu (ii) Ist $j = l$

$$\sum_{k=0}^{n-1} \underbrace{\omega_k^j \omega_k^{-l}}_{=1} = \sum_{k=0}^{n-1} 1 = n$$

Weiterhin ist $\omega_p = e^{\frac{2p\pi}{n}i}$ eine der n -ten Einheitswurzeln und damit

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

Ausklammern von $\omega_p - 1$ ergibt

$$(\omega_p - 1)(\omega_p^{n-1} + \omega_p^{n-2} + \dots + 1) = 0 \quad (6.41)$$

Man findet nun

$$\sum_{k=0}^{n-1} \omega_k^j \omega_k^{-l} = \sum_{k=0}^{n-1} \omega_k^{j-l} \stackrel{(i)}{=} \sum_{k=0}^{n-1} \omega_{j-l}^k = \sum_{k=0}^{n-1} (\omega_{j-l})^k$$

und da $j \neq l$, ist $\omega_{j-l} \neq 1$, d.h. $\sum (\omega_{j-l})^k$ muss als 2. Faktor der linken Seite von (6.41) = 0 sein. \square

Aus dem eben bewiesenen Satz ergibt sich die Folgerung

Korollar. Die komplexen Vektoren

$$\phi_j = \begin{pmatrix} \omega_0^j \\ \vdots \\ \omega_{n-1}^j \end{pmatrix}, \quad \phi_l = \begin{pmatrix} \omega_0^l \\ \vdots \\ \omega_{n-1}^l \end{pmatrix} \in \mathbb{C}^n, \quad (\phi_j)_k = \omega_k^j, \quad j \neq l$$

sind bezüglich des Skalarproduktes

$$\langle f, g \rangle := \frac{1}{n} \sum_{k=0}^{n-1} f_k \bar{g}_k \quad (6.42)$$

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

Definition 6.29. Die Koeffizienten $\beta_0, \dots, \beta_{n-1}$ aus (6.40), d.h. die Koeffizienten von $P(\omega)$ heißen **Fourierkoeffizienten** oder **diskrete Fouriertransformierte** von f_0, \dots, f_{n-1} falls $P(\omega_k) = f_k, k = 0, \dots, n-1$ gilt.

Satz 6.30. Für die diskreten Fouriertransformierten β_j von $f_j, j = 0, \dots, n-1$ gilt

$$\beta_j = \frac{1}{n} \sum_{k=0}^{n-1} f_k e^{-i \frac{jk2\pi}{n}} \quad (6.43)$$

d.h. sie sind eindeutig bestimmt.

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

$$\begin{aligned}
 P(\omega_0) &= \beta_0 \omega_0^0 + \cdots + \beta_{n-1} \omega_0^{n-1} = f_0 \\
 &\vdots \\
 P(\omega_{n-1}) &= \beta_0 \omega_{n-1}^0 + \cdots + \beta_{n-1} \omega_{n-1}^{n-1} = f_{n-1} \\
 \rightsquigarrow \quad \beta_0 \phi_0 + \beta_1 \phi_1 + \cdots + \beta_{n-1} \phi_{n-1} &= f := \begin{pmatrix} f_0 \\ \vdots \\ f_{n-1} \end{pmatrix} \quad (6.44)
 \end{aligned}$$

die skalare Multiplikation mit ϕ_j ergibt aufgrund der Orthogonalität

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

□

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

$$\mathcal{F}[f_0, \dots, f_{n-1}] := [\beta_0, \dots, \beta_{n-1}] \quad (6.45)$$

verwendet.

(6.44) bedeutet das Gleichungssystem

$$\underbrace{\begin{pmatrix} \omega_0^0 & \omega_0^1 & \cdots & \omega_0^{n-1} \\ \vdots & \vdots & & \vdots \\ \omega_{n-1}^0 & \omega_{n-1}^1 & \cdots & \omega_{n-1}^{n-1} \end{pmatrix}}_{=: V = (\omega_k^j)_{j,k=0,\dots,n-1}} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{n-1} \end{pmatrix} = \begin{pmatrix} f_0 \\ \vdots \\ f_{n-1} \end{pmatrix} \quad (6.46)$$

bzw.

$$\begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{n-1} \end{pmatrix} = \frac{1}{n} \underbrace{\begin{pmatrix} \omega_0^{-0} & \omega_0^{-1} & \cdots & \omega_0^{-(n-1)} \\ \vdots & \vdots & & \vdots \\ \omega_{n-1}^{-0} & \omega_{n-1}^{-1} & \cdots & \omega_{n-1}^{-(n-1)} \end{pmatrix}}_{=: \frac{1}{n} \bar{V} = (\frac{1}{n} \omega_k^{-j})_{j,k=0,\dots,n-1}} \begin{pmatrix} f_0 \\ \vdots \\ f_{n-1} \end{pmatrix} \quad (6.47)$$

Korollar. (i) Es gilt offensichtlich

$$\left(\frac{1}{n} \bar{V}\right)^{-1} = V$$

und jeder Datensatz $f_0, \dots, f_{n-1} \in \mathbb{C}$ lässt sich aus seiner diskreten Fouriertransformierten

$$\mathcal{F}[f_0, \dots, f_{n-1}] = [\beta_0, \dots, \beta_{n-1}]$$

durch (siehe (6.44))

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

zurückgewinnen. Es wird auch die Notation

$$\mathcal{F}^{-1}[\beta_0, \dots, \beta_{n-1}] = [f_0, \dots, f_{n-1}]$$

verwendet.

(ii) Es gilt

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

6.9.1 Beziehungen zwischen den reellen und komplexen Fourierkoeffizienten A_j, B_j, β_j

Es galt $\Psi(x_k) = f_k$ und außerdem war $\omega_k = e^{-ix_k}$ definiert. Für ungerades $n = 2m + 1$ folgt

$$\begin{aligned} \Psi(x_k) &= \frac{A_0}{2} + \sum_{l=1}^m \left(A_l \frac{1}{2} (\omega_k^l + \omega_k^{-l}) + B_l \frac{1}{2i} (\omega_k^l - \omega_k^{-l}) \right) \\ &= \frac{A_0}{2} + \sum_{l=1}^m \left(A_l \frac{1}{2} (\omega_k^l + \omega_k^{n-l}) + B_l \frac{1}{2i} (\omega_k^l - \omega_k^{n-l}) \right) \\ &= \beta_0 + \beta_1 \omega_k + \dots + \beta_{n-1} \omega_k^{n-1} \end{aligned}$$

Daraus folgt

$$A_0 = 2\beta_0 \Leftrightarrow \beta_0 = \frac{A_0}{2}$$

sowie

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

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

Mit der Formel (6.43) folgt:

$$\begin{aligned} A_l &= \frac{1}{n} \sum_{k=0}^{n-1} f_k \left(e^{-i \frac{kl2\pi}{n}} + e^{-i \frac{k(n-l)2\pi}{n}} \right) \\ &= \frac{2}{n} \sum_{k=0}^{n-1} f_k \frac{1}{2} \left(e^{-i \frac{kl2\pi}{n}} + e^{i \frac{kl2\pi}{n}} \right) = \frac{2}{n} \sum_{k=0}^{n-1} f_k \cos(lx_k) \end{aligned} \quad (6.48)$$

und analog

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

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

Satz 6.31. *Werden die Koeffizienten gemäß (6.48) bestimmt, so erfüllt*

$$\Psi(x) = \begin{cases} \frac{A_0}{2} + \sum_{k=0}^m (A_l \cos(kx) + B_k \sin(kx)), & n = 2m + 1 \\ \frac{A_0}{2} + \sum_{k=0}^{m-1} (A_l \cos(kx) + B_k \sin(kx)) + \frac{A_m}{2} \cos(mx), & n = 2m \end{cases}$$

Die Interpolationsbedingung

$$\Psi(x_k) = f_k, \quad k = 0, \dots, n-1$$

für reelle f_k .

Ziel ist die Reduzierung des Aufwands zur Berechnung der diskreten Fouriertransformierten $\beta_0, \dots, \beta_{n-1}$ für einen Datensatz f_0, \dots, f_{n-1} der mit der Auswertung der Berechnungsvorschrift

$$\beta_j = \frac{1}{n} \sum_{k=0}^{n-1} f_k e^{-i \frac{jk2\pi}{n}}, \quad j = 0, \dots, n-1$$

etwa $\mathcal{O}(n^2)$ komplexe Multiplikationen bedeutet.

6.9.2 Schnelle Fouriertransformation (FFT)

Voraussetzung $n = 2^p, p \in \mathbb{N}$, d.h. es werden Datensätze mit $n = 2^p$ Daten aus \mathbb{C} betrachtet. Entscheidende Grundlage für die FFT ist der folgende

Satz 6.32. Aus den diskreten Fouriertransformierten der beiden Datensätze

$$g_0, \dots, g_{M-1} \quad \text{und} \quad g_M, \dots, g_{2M-1}$$

der Länge M lässt sich die diskrete Fouriertransformierte des Datensatzes

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

der Länge $2M$ folgendermaßen bestimmen.

$$\begin{aligned} & \frac{1}{2} \left\{ \mathcal{F}_k[g_0, \dots, g_{M-1}] + e^{-i\frac{k\pi}{M}} \mathcal{F}_k[g_M, \dots, g_{2M-1}] \right\} \\ &= \mathcal{F}_k[g_0, g_M, g_1, g_{M+1}, \dots, g_{2M-1}] \end{aligned} \quad (6.49)$$

$$\begin{aligned} & \frac{1}{2} \left\{ \mathcal{F}_k[g_0, \dots, g_{M-1}] - e^{-i\frac{k\pi}{M}} \mathcal{F}_k[g_M, \dots, g_{2M-1}] \right\} \\ &= \mathcal{F}_{M+k}[g_0, g_M, g_1, g_{M+1}, \dots, g_{2M-1}] \end{aligned} \quad (6.50)$$

Für $k = 0, \dots, M-1$. Wobei \mathcal{F}_k bzw. \mathcal{F}_{M+k} die k -te bzw. $(M+k)$ -te Komponente von \mathcal{F} bezeichnen.

Beweis. Für $k = 0, \dots, M-1$ gilt

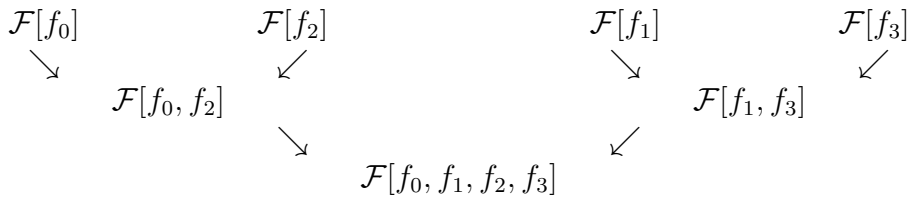
$$\begin{aligned} \mathcal{F}_k[g_0, \dots, g_{2M-1}] &= \frac{1}{2M} \left(\sum_{j=0}^{M-1} g_j e^{-i\frac{2jk2\pi}{2M}} + \sum_{j=0}^{M-1} g_{M+j} e^{-i\frac{(2j+1)k2\pi}{2M}} \right) \\ &= \frac{1}{2M} \left(\sum_{j=0}^{M-1} g_j e^{-i\frac{jk2\pi}{M}} + e^{-i\frac{k\pi}{M}} \sum_{j=0}^{M-1} g_{M+j}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

Die Gleichung (6.50) erhält man analog, wobei

$$e^{-i\frac{j(k+M)2\pi}{2M}} = e^{-ij\pi} e^{-i\frac{jk2\pi}{2M}} = (-1)^j e^{-i\frac{jk2\pi}{2M}}$$

berücksichtigt wird. □

Ist $n = 2^p$, dann soll der Satz 6.32 auf einem Datensatz dieser Länge rekursiv angewandt werden. Die Anordnung der Daten wird später erklärt.



Erläuterungen zum Schema

- (a) Beim Übergang von Stufe 0 zu Stufe 1 werden 2 diskrete Fouriertransformierte der Länge 2 ausgehend von 4 diskreten Fouriertransformierten der Länge 1 berechnet (Anwendung der Formeln (6.49), (6.50) je 2-mal).
- (b) Beim Übergang von Stufe 1 zu Stufe 2 wird 1 diskrete Fouriertransformierte der Länge 4 ausgehend von 2 diskreten FTs der Länge 2 berechnet (zweimalige Anwendung der Formeln (6.49), (6.50))
- (c) Schließlich erhält man ausgehend von diesen die gewünschte diskrete FT des Datensatzes f_0, \dots, f_3
- (d) Entscheidend für genau dieses Ergebnis war die Anordnung der Daten auf der Stufe 0
- (e) Die Anwendung des Satzes 6.32 soll beim Übergang von Stufe 2 zu Stufe 3 erläutert werden:

Setzt man

$$g_0 = f_0, g_1 = f_2, g_2 = f_1, g_3 = f_3$$

dann erhält man ausgehend von

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

mit den Formeln (6.49),(6.50)

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

also bei Berücksichtigung der Setzungen

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

Bemerkung 6.33. Für Anordnung der Daten auf der Stufe 0 nutzt man das folgende Schema der Bit-Umkehr, die in der folgenden Tabelle für $n = 8 = 2^3$ beschrieben wird:

f_k Index	Binärwert	Binärwert revers	Index
0	000	000	0
1	001	100	4
2	010	010	2
3	011	110	6
4	100	001	1
5	101	101	5
6	110	011	3
7	111	111	7

In der letzten Spalte liest man die Indexreihenfolge für die Anordnung der Daten auf der Stufe 0 ab.

6.9.3 Aufwand der FFT

Zum Abschluss der Thematik FFT soll nun der Aufwand diskutiert werden.

Bezeichnet man die Stufen der FFT mit $r \in \{0, 1, \dots, p\}$, also im Fall $8 = n = 2^3$ $r \in \{0, 1, 2, 3\}$, dann ergibt sich für den Aufwand der FFT:

Für $r \in \{0, \dots, p-1\}$ fallen beim Übergang von der r -ten zur $(r+1)$ -ten Stufe der FFT die folgenden komplexen Multiplikationen an

- Die Berechnung von Zahlen $\omega^2, \dots, \omega^{2^r-1} \in \mathbb{C}$ (ω Wert einer komplexen Exponentialfunktion) erfordert $2^r - 2 \leq 2^r$ komplexe Multiplikationen (Faktoren in den Formeln (6.49), (6.50))
- Berechnung der diskreten Fouriertransformierten der Länge 2^{r+1} ausgehend von je 2 diskreten Fouriertransformierten der Länge 2^r , und das insgesamt $2^p - r - 1$ -mal ergibt $2^n \cdot 2^{p-r-1} = 2^{p-1}$ komplexe Multiplikationen
- Dazu kommen noch $p - 2 \leq p$ komplexe Multiplikationen zur Berechnung etwa von $\omega_k = \omega_{k+1}^2$
- Für die Ausführung der Übergänge von den Stufen 0 bis p ergibt sich die Gesamtzahl an komplexen Multiplikationen

$$\sum_{r=0}^{p-1} (2^{p-1} + 2^r) + p \leq p2^{p-1} + 2^p + p = \frac{n \log_2 n}{2} + \mathcal{O}(n)$$

Damit gilt der

Satz 6.34. *Bei der FFT zur Bestimmung der diskreten Fouriertransformierten eines Datensatzes der Länge $n = 2^p$ fallen nicht mehr als*

$$\frac{n \log_2 n}{2} + \mathcal{O}(n)$$

komplexe Multiplikationen an.

Bemerkung 6.35. Wir haben für die Fouriertransformation die Formeln

$$\beta_j = \frac{1}{n} \sum_{k=0}^{n-1} f_k e^{-i \frac{k2\pi}{n} j} \quad (6.51)$$

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

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

$$\beta_j = \sum_{k=0}^{n-1} f_k e^{-i \frac{jk2\pi}{n}} \quad (6.52)$$

definiert, also ohne den Faktor $\frac{1}{n}$. Das hat für die Rücktransformation die Konsequenz

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

Eine dritte Möglichkeit ist durch

$$\beta_j = \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} f_k e^{-i \frac{jk2\pi}{n}} \quad (6.53)$$

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

gegeben.

Besonders bei der Nutzung von Numerikprogrammsystemen oder Bibliotheken ist es daher ratsam, die jeweils verwendete Definition der Fouriertransformation und der Rücktransformation zu ermitteln, also (6.51), (6.52) oder (6.53).

Kapitel 7

Numerische Integration

Ziel ist die Berechnung des bestimmten Integrals

$$\int_a^b f(x)dx$$

wobei man aus unterschiedlichen Gründen nicht die Berechnung mittels einer Stammfunktion $F(x)$ durch

$$\int_a^b f(x)dx = F(b) - F(a)$$

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.

20.
Vorle-
sung
am
08.01.2014

7.1 Interpolatorische Quadraturformeln

Definition 7.1. *Die Näherungsformel*

$$Q_n(f) = \int_a^b p_n(x)dx = (b-a) \sum_{k=0}^n f(x_k)\sigma_k \quad (7.1)$$

mit dem Interpolationspolynom p_n zu den Stützpunkten $(x_0, f(x_0)), \dots, (x_n, f(x_n))$ für das Integral $\int_a^b f(x)dx$ nennt man **interpolatorische Quadraturformel**.

Definition 7.2. *Mit*

$$E_n[f] = \int_a^b f(x)dx - Q_n = I(f) - Q_n(f) \quad (7.2)$$

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 $a \leq x_0 < \dots < x_n \leq b$ existiert eine eindeutig bestimmte interpolatorische Quadraturformel deren Genauigkeitsgrad mindestens gleich n ist.*

Satz 7.4. *Eine interpolatorische Quadraturformel Q_n besitzt die Gestalt*

$$Q_n(f) = (b-a) \sum_{k=0}^n \sigma_k f(x_k) \quad \text{mit} \quad \sigma_k = \int_0^1 \prod_{\substack{j=0 \\ j \neq k}}^n \frac{t-t_j}{t_k-t_j} dt, \quad t_j = \frac{x_j-a}{b-a}. \quad (7.3)$$

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

$$\frac{1}{b-a} \int_a^b L_k(x) dx = \dots = \sigma_k = \int_0^1 \prod_{\substack{j=0 \\ j \neq k}}^n \frac{t-t_j}{t_k-t_j} dt = \sigma_k.$$

□

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 $f : [a, b] \rightarrow \mathbb{R}$ sei $(r+1)$ -mal stetig diff'bar. Dann gilt die Fehlerabschätzung

$$|I(f) - Q_n(f)| \leq c_r \frac{(b-a)^{r+2}}{(r+1)!} \max_{\xi \in [a,b]} |f^{(r+1)}(\xi)| \quad (7.4)$$

mit

$$c_r = \min_{t_{n+1}, \dots, t_r \in [0,1]} \int_0^1 \prod_{k=0}^r |t - t_k| dt, \quad t_k = \frac{x_k - a}{b - a}, \quad k = 0, 1, \dots, n. \quad (7.5)$$

Wenn mit Werten t_0, \dots, t_n aus (7.5) für eine bestimmte Wahl von $t_{n+1}, \dots, t_r \in [0, 1]$ das Produkt $\prod_{k=0}^r (t - t_k)$ von einem Vorzeichen¹ in $[0, 1]$ ist, so gilt mit einer Zwischenstelle $\xi \in [a, b]$ die Fehlerdarstellung

$$I(f) - Q_n(f) = c'_r \frac{(b-a)^{r+2}}{(r+1)!} f^{(r+1)}(\xi) \quad \text{mit} \quad c'_r = \int_0^1 \prod_{k=0}^r (t - t_k) dt. \quad (7.6)$$

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 $[a, b]$

$$x_k = a + kh, \quad k = 0, \dots, n, \quad h = \frac{b-a}{n}$$

- Verwendung des Interpolationspolynoms $p_n \in \Pi_n$ für die Stützpunkte $(x_k, f(x_k))$, d.h. es ist

$$p_n(x_k) = f(x_k), \quad k = 0, \dots, n$$

- Näherung des Integrals $\int_a^b f(x) dx$ durch

$$\int_a^b p_n(x) dx \approx \int_a^b f(x) dx$$

¹Eine 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

$$\begin{aligned} \int_a^b p_n(x) dx &= \sum_{k=0}^n f_k \int_a^b L_k dx \\ &= \sum_{k=0}^n f_k \int_a^b \prod_{k \neq j=0}^n \frac{x - x_j}{x_k - x_j} dx = (*) \end{aligned}$$

und mit der Substitution $s = \frac{x-a}{h}$, $h ds = dx$ folgt

$$(*) = (b-a) \sum_{k=0}^n f_k \underbrace{\frac{1}{n} \int_0^n \prod_{k \neq j=0}^n \frac{s-j}{k-j} ds}_{\sigma_k}$$

also

$$Q_n(f) = \int_a^b p_n(x) dx = (b-a) \sum_{k=0}^n f_k \sigma_k \quad (7.7)$$

mit den Gewichten

$$\sigma_k = \frac{1}{n} \int_0^n \prod_{k \neq j=0}^n \frac{s-j}{k-j} ds, \quad k = 0, \dots, n \quad (7.8)$$

Für $n = 1$ erhält man

$$\sigma_0 = \int_0^1 \frac{s-1}{0-1} ds = -\frac{1}{2} (s-1)^2 \Big|_0^1 = \frac{1}{2}, \quad \sigma_1 = \frac{1}{2}$$

woraus mit

$$\int_a^b f(x) dx \approx Q_1(f) = \int_a^b p_1(x) dx = \frac{b-a}{2} (f(a) + f(b)) \quad (7.9)$$

die **Trapezregel** folgt.

Für $n = 2$ ergibt sich

$$\begin{aligned} \sigma_0 &= \frac{1}{2} \int_0^2 \frac{s-1}{0-1} \cdot \frac{s-2}{0-2} ds = \frac{1}{4} \int_0^2 (s^2 - 3s + 2) ds \\ &= \frac{1}{4} \left[\frac{s^3}{3} - \frac{3s^2}{2} + 2s \right] = \frac{1}{4} \left[\frac{8}{3} - 6 + 4 \right] = \frac{1}{4} \left[\frac{8-6}{3} \right] = \frac{1}{6} \end{aligned}$$

$$\sigma_2 = \frac{1}{6}, \quad \sigma_1 = \frac{4}{6}$$

woraus mit

$$\int_a^b f(x)dx \approx Q_2(f) = \int_a^b p_2(x)dx = \frac{b-a}{6} \left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right) \quad (7.10)$$

Die **Simpson-Regel**, auch **Keplersche Fassregel** genannt, folgt.

Für $n = 3$ findet man auf analoge Weise mit

$$\begin{aligned} \int_a^b f(x)dx &\approx Q_3(f) = \int_a^b p_3(x)dx \\ &= \frac{b-a}{8} \left(f(a) + 3f\left(\frac{2a+b}{3}\right) + 3f\left(\frac{a+2b}{3}\right) + f(b) \right) \end{aligned} \quad (7.11)$$

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

Gilt für die Stützstellen $x_k = a + kh$, $h = \frac{b-a}{n}$, $k = 0, \dots, n$, 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

$$\begin{aligned} E_2[x^3] &= \int_a^b x^3 dx - \frac{b-a}{6} \left[a^3 + 4\left(\frac{a+b}{2}\right)^3 + b^3 \right] \\ &= \frac{1}{4}(b^4 - a^4) - \frac{b-a}{6} \left[a^3 + \frac{1}{2}(a^3 + 3a^2b + 3ab^2 + b^3) + b^3 \right] \\ &= 0 \end{aligned}$$

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

$$\begin{aligned}
 E_1[f] &= -\frac{1}{12}h^3 f''(\xi), & h &= b - a \\
 E_2[f] &= -\frac{1}{90}h^5 f^{(4)}(\xi), & h &= \frac{b-a}{2} \\
 E_3[f] &= -\frac{3}{80}h^5 f^{(4)}(\xi), & h &= \frac{b-a}{3} \\
 E_4[f] &= -\frac{8}{945}h^7 f^{(6)}(\xi), & h &= \frac{b-a}{4}
 \end{aligned}$$

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

$$E_2[f] = I(f) - Q_2(f) = c'_3 \frac{(b-a)^5}{4!} f^{(4)}(\xi) \quad \text{mit} \quad c'_3 = \int_0^1 \prod_{k=0}^3 (t - t_k) dt, \tag{7.12}$$

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

$$c'_3 = \int_0^1 t(t - \frac{1}{2})(t - 1)(t - \frac{1}{2}) dt = \dots = -\frac{1}{120}.$$

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

$$\sigma_{n-k} = \sigma_k \quad \text{für} \quad k = 0, 1, \dots, n,$$

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.

21.
Vorlesung
am
13.01.2014

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 von $I = \int_a^b f(x)dx$. 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 Simpsonregel auf N Teilintervallen an.

Zur näherungsweisen Berechnung von $\int_\alpha^\beta f(x)dx$ unterteilt man das Intervall $[\alpha, \beta]$ durch

$$\alpha = x_{10} < \dots < x_{1n} = x_{20} < \dots < x_{N-1n} = x_{N0} < \dots < x_{Nn} = \beta$$

in N gleichgroße Teilintervalle $[x_{j0}, x_{jn}]$, $j = 1, \dots, N$ mit jeweils $n + 1$ Stützstellen. Auf den Teilintervallen $[a, b] = [x_{j0}, x_{jn}]$ nähert man das Integral

$$\int_{x_{j0}}^{x_{jn}} f(x)dx \quad \text{mit} \quad Q_{n,j}$$

zu den Stützstellen x_{j0}, \dots, 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 ($h = \frac{\beta-\alpha}{N}$)

$$\begin{aligned} S_{1,N} &= h \left[\frac{1}{2}y_{10} + y_{20} + \dots + y_{N0} + \frac{1}{2}y_{N1} \right] \\ &= h \left[\frac{1}{2}(y_{10} + y_{N1}) + \sum_{k=2}^N y_{k0} \right] \end{aligned} \quad (7.13)$$

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

$$S_{2,N} = \frac{h}{3} \left[(y_{10} + y_{N2}) + 2 \sum_{j=1}^{N-1} y_{j2} + 4 \sum_{j=1}^N y_{j1} \right] \quad (7.14)$$

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 □

7.5 Gauß-Quadraturen

Bei den Newton-Cotes-Quadraturformeln ist man von einer vorgegebenen Zahl von äquidistanten Stützstellen x_0, \dots, x_n ausgegangen und hat eine Näherung des Integrals $\int_{x_0}^{x_n} f(x)dx$ durch das Integral des Interpolationspolynoms $p_n(x)$ für $(x_k, f(x_k))$, $k = 0, \dots, n$ 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 $\lambda_1, \dots, \lambda_n$, da sie sich letztendlich als Nullstellen eines Polynoms n -ten Grades ergeben werden. Wir wollen sie im Folgenden aber weiter mit x_1, \dots, x_n bezeichnen und beginnen aber im Unterschied zu den Newton-Cotes-Formeln bei $k = 1$ zu zählen.

Ziel ist die Berechnung des Integrals $\int_a^b g(x)dx$ 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

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

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 $x^0, x^1, \dots, x^{2n-1}$ exakt ist und somit $I_n = I$ gilt, dann müssen die Stützstellen x_1, \dots, x_n und die Gewichte $\sigma_1, \dots, \sigma_n$ Lösungen des Gleichungssystems

$$\sum_{j=1}^n \sigma_j x_j^k = \int_a^b x^k \rho(x) dx \quad (k = 0, 1, \dots, 2n - 1) \quad (7.16)$$

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 $\int_{-1}^1 f(x) \rho(x) dx$ mit der Gewichtsfunktion $\rho(x) \equiv 1$ und der Vorgabe von $n = 2$ bedeutet (7.16) mit

$$\int_{-1}^1 dx = 2, \quad \int_{-1}^1 x dx = 0, \quad \int_{-1}^1 x^2 dx = \frac{2}{3}, \quad \int_{-1}^1 x^3 dx = 0$$

das Gleichungssystem

$$\begin{aligned} \sigma_1 + \sigma_2 &= 2 \\ \sigma_1 x_1 + \sigma_2 x_2 &= 0 \\ \sigma_1 x_1^2 + \sigma_2 x_2^2 &= \frac{2}{3} \\ \sigma_1 x_1^3 + \sigma_2 x_2^3 &= 0 \end{aligned} \quad (7.17)$$

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

$$\int_{-1}^1 f(x) dx = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$

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 Folgenden 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

$$\langle p, q \rangle_\rho := \int_a^b p(x)q(x)\rho(x)dx \quad (7.18)$$

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

$$\|p\|_\rho^2 = \langle p, p \rangle_\rho = \int_a^b p^2(x)\rho(x)dx \quad (7.19)$$

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

$$V^\perp = \{f \in P \mid \langle f, p \rangle_\rho = 0 \quad \forall p \in V\}$$

das **orthogonale Komplement** von V bezeichnet.

Die lineare Hülle der Funktionen $p_1, \dots, p_n \in P$ wird durch

$$\text{span}\{p_1, \dots, p_n\} = \{c_1p_1 + \dots + c_np_n \mid c_1, \dots, c_n \in K\}$$

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 $1, x, \dots, x^n, \dots$ eine Basis zur Konstruktion von Polynomen bilden. Mit $p_0(x) = 1$ wird durch

$$p_n(x) = x^n - \sum_{j=0}^{n-1} \frac{\langle x^n, p_j \rangle_\rho}{\langle p_j, p_j \rangle_\rho} p_j(x) \quad (7.20)$$

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

$$p_1(x) = x, \quad p_2(x) = x^2 - \frac{1}{3}, \quad p_3(x) = x^3 - \frac{3}{5}x, \quad p_4(x) = x^4 - \frac{5}{2}x^2 + \frac{4}{105} \quad (7.21)$$

paarweise orthogonaler Polynome bezüglich des Skalarproduktes

$$\langle p, q \rangle_\rho = \int_{-1}^1 p(x)q(x)dx$$

Die eben konstruierten orthogonalen Polynome heißen **Legendre-Polynome**.

Bemerkung 7.10. Bezeichnet man durch $P_i_k = \text{span}\{p_0, \dots, p_k\}$ en Vektorraum der Polynome bis zum Grad k , dann gilt allgemein für die Folge paarweise orthogonaler Polynome p_0, \dots, p_n mit aufsteigendem Grad

$$p_n \in P_{n-1}^\perp$$

Beispiel. Mit $[a, b] = [-1, 1]$ und der Gewichtsfunktion $\rho(x) = (1-x^2)^{-\frac{1}{2}} = \frac{1}{\sqrt{1-x^2}}$ erhält man mit dem Gram-Schmidt-Verfahren (7.20) ausgehend von $p_0 = 1$ mit

$$p_0(x) = 1, \quad p_1(x) = x, \quad p_2(x) = x^2 - \frac{1}{2}, \quad p_3(x) = x^3 - \frac{3}{4}x \quad (7.22)$$

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 $a < \lambda_1 < \dots < \lambda_j < b$ ($0 \leq j \leq n$) 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.

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

$$\langle p_n, q \rangle_\rho = 0 \quad (7.23)$$

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 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. \square

Nun kommen wir zur Definition der Gauß-Quadratur

Definition 7.12. *Mit x_1, \dots, x_n seien die Nullstellen des n -ten Orthogonalpolynoms $p_n(x)$ gegeben. Die numerische Integrationsformel*

$$I_n = \sum_{j=1}^n \sigma_j f(x_j) \quad \text{mit} \quad \sigma_j = \langle L_j, 1 \rangle_\rho = \int_a^b L_j(x)\rho(x)dx \quad (7.24)$$

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, \dots, 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

$$\int_a^b p(x)\rho(x)dx = \langle p, 1 \rangle_\rho = \sum_{j=1}^n \sigma_j p(x_j), \quad \forall p \in \Pi_{2n-1} \quad (7.25)$$

22.
Vor-
sung
am
15.01.2014

Außerdem ist die Quadratur interpolatorisch, d.h. es gilt für das Interpolationspolynom q_{n-1} zu den Stützpunkten $(x_j, f(x_j)), j = 1, \dots, n$

$$\int_a^b q_{n-1}(x)\rho(x)dx = \sum_{j=1}^n \sigma_j q_{n-1}(x_j) = \sum_{j=1}^n \sigma_j f(x_j)$$

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

$$\frac{p}{p_n} = q + \frac{r}{p_n} \Leftrightarrow p = qp_n + r \quad (7.26)$$

Mit den Nullstellen x_1, \dots, x_n von p_n gilt $p(x_j) = r(x_j)$ für $j = 1, \dots, n$. 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

$$\begin{aligned} \int_a^b p(x)\rho(x)dx &= \langle p, 1 \rangle_\rho = \langle r, 1 \rangle_\rho \\ &= \sum_{j=1}^n p(x_j) \langle L_j, 1 \rangle_\rho = \sum_{j=1}^n \sigma_j p(x_j) . \end{aligned} \quad (7.27)$$

Für $p(x) = L_j^2(x) \in \Pi_{2n-2}$ ergibt die eben nachgewiesene Formel (7.27)

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

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

$$I_n^* = \sum_{j=1}^n \sigma_j^* f(x_j^*) \quad (7.28)$$

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

$$\begin{aligned} \int_a^b h(x)\rho(x)dx &= \int_a^b L_k^*(x)p_n(x)\rho(x)dx \\ &= \sum_{j=1}^n \sigma_j^* L_k^*(x_j^*)p_n(x_j^*) = \sigma_k^* p_n(x_k^*) \end{aligned}$$

für alle $k = 1, \dots, n$. Da das 2. Integral $\int_a^b L_k^*(x)p_n(x)\rho(x)dx = \langle L_k^*, p_n \rangle_\rho$ 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 $k = 1, \dots, n$. 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. \square

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 $f(x)$*

$$\int_a^b f(x)\rho(x)dx - \sum_{j=1}^n \sigma_j f(x_j) = \frac{\|p_n\|_\rho^2}{(2n)!} f^{(2n)}(\xi) \quad (7.29)$$

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

Die folgende Tabelle zeigt Intervalle, Gewichtsfunktionen, die zugehörigen Orthogonalpolynome und deren Name ($\alpha, \beta > -1$)

Intervall	$\rho(x)$	p_0, p_1, \dots	Bezeichnung
$[-1, 1]$	1	$1, x, x^2 - \frac{1}{3}, \dots$	Legendre
$[-1, 1]$	$\frac{1}{\sqrt{1-x^2}}$	$1, x, x^2 - \frac{1}{2}, \dots$	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}	$1, x, x^2 - \frac{1}{2}, x^3 - \frac{3}{2}x, \dots$	Hermite
$[0, \infty[$	$e^{-x}x^\alpha$	$1, x - \alpha - 1, \dots$	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 (6.13)) 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

$$\langle p, q \rangle = \int_{-\infty}^{\infty} p(x)q(x)e^{-x^2} dx$$

findte man ausgehend von der Monombasis durch Gram-Schmidt-Orthogonalisierung die ersten Orthogonalpolynome

$$\begin{aligned} p_0 &= 1 \\ p_1 &= x - \frac{\langle x, 1 \rangle}{\langle 1, 1 \rangle} 1 = x \\ p_2 &= x^2 - \frac{\langle x^2, 1 \rangle}{\langle 1, 1 \rangle} 1 - \frac{\langle x^2, x \rangle}{\langle x, x \rangle} x = x^2 - \frac{1}{2} \\ &\dots \end{aligned}$$

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

$$I_2(f) = \frac{\sqrt{\pi}}{2} \left[f\left(\frac{1}{\sqrt{2}}\right) + f\left(-\frac{1}{\sqrt{2}}\right) \right]$$

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 näherungsweise Berechnung des Integrals $\int_a^b f(x) dx$ kann man bei der Verwendung von N Teilintervallen in der Form

$$T(h) := S_{1,N} = h \left[\frac{1}{2} (f(a) + f(b)) + \sum_{i=1}^{N-1} f(a + ih) \right]$$

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*

$$T(h) = \tau_0 + \tau_1 h^2 + \tau_2 (h^2)^2 + \dots + \tau_m (h^2)^m + R_{m+1}(h) \quad (7.30)$$

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

$$\begin{aligned} T(h_0) &\approx \tau_0 + \tau_1 h_0^2 + \tau_2 h_0^4 \\ T(h_1) &\approx \tau_0 + \tau_1 h_1^2 + \tau_2 h_1^4 \\ T(h_2) &\approx \tau_0 + \tau_1 h_2^2 + \tau_2 h_2^4 \end{aligned} \quad (7.31)$$

und kann bei Kenntnis der Trapezsummen $T(h_1)$, $T(h_2)$ und $T(h_3)$ daraus τ_0 näherungsweise ermitteln. Bei den Beziehungen (7.31) macht man aufgrund von Satz 7.16 nur Fehler der Ordnung $\mathcal{O}(h^6)$.

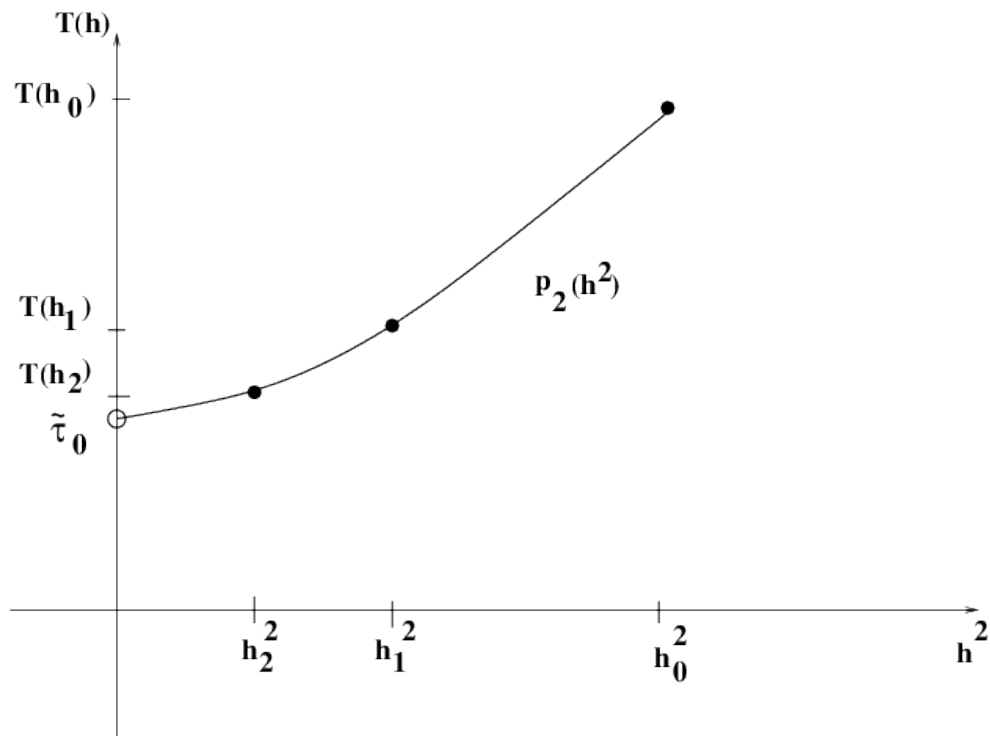


Abbildung 7.1: Polynom p_2 und Stützwerte $(h_k^2, T(h_k))$, $k = 0, 1, 2$

Eine andere Interpretation dieser Extrapolationsidee besteht in der Nutzung der Wertepaare

$$(h_0^2, T(h_0)), (h_1^2, T(h_1)), (h_2^2, T(h_2))$$

zur Bestimmung des Interpolationspolynoms zweiten Grades in $\tilde{h} = h^2$, also

$$p_2(\tilde{h}) = p_2(h^2) = \tilde{\tau}_0 + \tilde{\tau}_1 h^2 + \tilde{\tau}_2 (h^2)^2$$

mit der Eigenschaft

$$p_2(h_k^2) = T(h_k), \quad k = 0, 1, 2.$$

Die Auswertung dieses Polynoms an der Stelle $h^2 = 0$ (Extrapolation, s.auch Abb. 7.1) liefert dann die Näherung

$$\int_a^b f(x) dx = \tau_0 \approx p_2(0) = \tilde{\tau}_0.$$

7.7 Anwendung des Schemas von Neville Aitken - Romberg-Verfahren

In Anlehnung an die Entwicklung (7.30) sucht man also ein Interpolationspolynom p_m an den Stützstellen h_k^2 mit den Funktionswerten $T(h_k)$ ($k = 0, \dots, m$) 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 $(x_k, f(x_k))$, $k = 0, 1, \dots, m$, an, also

x_i	$T_{i,0} = f(x_i)$	$T_{i,1}$	$T_{i,2}$	\dots	$T_{i,m-1}$	$T_{i,m}$
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	$T_{m,0} = f(x_m)$	$T_{m,1}$	$T_{m,2}$	\dots	$T_{m,m-1}$	$T_{m,m}$

mit $m \geq i \geq k \geq 1$ und

$$T_{i,0} = f(x_i)$$

$$T_{i,k}(x) = \frac{(x - x_{i-k})T_{i,k-1}(x) - (x - x_i)T_{i-1,k-1}(x)}{x_i - x_{i-k}}, \quad k \geq 1,$$

woraus

$$T_{i,k} = \frac{(x - x_i)T_{i,k-1} + (x_i - x_{i-k})T_{i,k-1} - (x - x_i)T_{i-1,k-1}}{x_i - x_{i-k}}$$

$$= T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{\frac{x - x_{i-k}}{x - x_i} - 1}$$

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

$$\begin{aligned} T_{i,0} &= T(h_i) \\ T_{i,k} &= T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{\left(\frac{h_{i-k}}{h_i}\right)^2 - 1} \end{aligned}$$

mit $T_{m,m}$ den Näherungswert für $\tau_0 = \int_a^b f(x) dx$.

Für $m = 1$ erhält man mit $h_0 = b - a$, $h_1 = (b - a)/2$

$$\begin{aligned} T_{1,1} &= T_{1,0} + \frac{T_{1,0} - T_{0,0}}{\left(\frac{h_0}{h_1}\right)^2 - 1} = \frac{4}{3}T_{1,0} - \frac{1}{3}T_{0,0} \\ &= \frac{b-a}{3}[f(a) + 2f\left(\frac{a+b}{2}\right) + f(b)] - \frac{b-a}{6}[f(a) + f(b)] \\ &= \frac{b-a}{6}[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)], \end{aligned}$$

also die Simpsonregel. Für $h_i = \frac{b-a}{3^i}$, $i = 0, 1$ erhält man mit

$$T_{1,1} = \frac{b-a}{8}[f(a) + 3f\left(\frac{2a+b}{3}\right) + 3f\left(\frac{a+3b}{3}\right) + f(b)]$$

die Newtonsche 3/8-Regel.

Als gängige Folgen h_i , $i = 0, \dots$ werden die **Romberg-Folge**

$$h_0 = b - a, \quad h_1 = \frac{h_0}{2}, \quad h_2 = \frac{h_1}{2}, \quad h_3 = \frac{h_2}{2}, \dots$$

oder die **Bulirsch-Folge**

$$h_0 = b - a, \quad h_1 = \frac{h_0}{2}, \quad h_2 = \frac{h_0}{3}, \quad h_3 = \frac{h_1}{2}, \quad h_4 = \frac{h_2}{2}, \dots$$

verwendet. Zur Romberg-Folge ist noch anzumerken, dass man $T(h_{i+1})$ rekursiv aus $T(h_i)$ durch die Formel

$$T(h_{i+1}) = T\left(\frac{1}{2}h_i\right) = \frac{1}{2}T(h_i) + h_{i+1}[f(a+h_{i+1}) + f(a+3h_{i+1}) + \dots + f(b-h_{i+1})]$$

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	0.7500000000000000		
1	1/4	0.7083333333333333	0.6944444444444444	
2	1/8	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

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

23.
Vorle-
sung
am
20.01.2014

Definition 8.1. Ein **Anfangswertproblem** für ein System von n gewöhnlichen Differentialgleichungen 1. Ordnung ist von der Form

$$y' = f(t, y), \quad t \in [a, b] \quad (8.1)$$

$$y(a) = y_0 \quad (8.2)$$

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

$$f : [a, b] \times \mathbb{R}^n \rightarrow \mathbb{R}^n \quad (8.3)$$

wobei eine differenzierbare Abbildung $y : [a, b] \rightarrow \mathbb{R}^n$ 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

$$\|f(t, u) - f(t, v)\| \leq L \|u - v\|, \quad t \in [a, b], \quad u, v \in \mathbb{R}^n \quad (8.4)$$

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 $y : [a, b] \rightarrow \mathbb{R}^n$ (Picard-Lindelöf)

(b) Für differenzierbare Funktionen $y, \hat{y} : [a, b] \rightarrow \mathbb{R}^n$ mit

$$\begin{aligned} y' &= f(t, y), & t \in [a, b]; & & y(a) &= y_0 \\ \hat{y}' &= f(t, \hat{y}), & t \in [a, b]; & & \hat{y}(a) &= \hat{y}_0 \end{aligned}$$

gilt die Abschätzung

$$\|y(t) - \hat{y}(t)\| \leq e^{L(t-a)} \|y_0 - \hat{y}_0\|, \quad t \in [a, b] \quad (8.5)$$

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 annähern:

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

$$(t_0, y_0) \quad \text{zu} \quad (t_0 + h, y_0 + hf(t_0, y_0)) =: (t_1, y_1)$$

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

$$y_0, y_1, \dots, y_N$$

unter "günstigen" Umständen eine Approximation der Lösung $y(t)$ an den Stellen

$$a = t_0, t_1, \dots, t_N = b$$

(III) D.h. man fährt das Richtungsfeld geeignet ab, um eine numerische Lösung $y_k, k = 0, 1, \dots, N$ zu erhalten

8.1 Theorie der Einschrittverfahren

Definition 8.4. Ein Einschrittverfahren zur näherungsweise Bestimmung einer Lösung des AWP (8.1),(8.2) hat die Form

$$y_{k+1} = y_k + h_k \Phi(t_k, y_k, y_{k+1}, h_k), \quad k = 0, 1, \dots, N-1 \quad (8.6)$$

mit einer Verfahrensfunktion

$$\Phi : [a, b] \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+ \rightarrow \mathbb{R}$$

und einem (noch nicht näher spezifizierten) Gitter bzw. Schrittweiten

$$\Delta = \{a = t_0 < t_1 < \dots < t_N \leq b\}, \quad h_k := t_{k+1} - t_k, \quad k = 0, 1, \dots, N-1 \quad (8.7)$$

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

$$d_{k+1} := y(t_{k+1}) - y(t_k) - h_k \Phi(t_k, y(t_k), y(t_{k+1}), h_k) \quad (8.8)$$

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

$$\begin{aligned} |d_k| &\leq \text{const} \cdot h_k^{p+1}, \quad k = 1, \dots, N \\ \max_{1 \leq k \leq N} |d_k| &\leq D = \text{const} \cdot h_{\max}^{p+1} = \mathcal{O}(h_{\max}^{p+1}) \end{aligned} \quad (8.9)$$

mit $h_{\max} = \max_{k=0, \dots, N-1} t_{k+1} - t_k$ gilt. (Statt Fehlerordnung verwendet man auch den Begriff *Konsistenzordnung*.) Ist $p \geq 1$, dann heißt das Verfahren *konsistent*.

Die Bedingungen

$$\begin{aligned} |\Phi(t, u_1, u_2, h) - \Phi(t, v_1, u_2, h)| &\leq L_1 |u_1 - v_1| \\ |\Phi(t, u_1, u_2, h) - \Phi(t, u_1, v_2, h)| &\leq L_2 |u_2 - v_2| \end{aligned} \quad (8.10)$$

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 *Bedingung* (8.10). Dann liegt die *Konvergenzordnung* p vor, d.h. es gilt

$$\max_{k=0, \dots, N} |y_k - y(t_k)| \leq K h_{\max}^p$$

Mit einer Konstanten K , die vom Intervall $[a, b]$, Konstanten C aus der Abschätzung (8.9) und L_1, L_2 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

24.
Vor-
lesung
22.01.2014

Lemma 8.9. Für Zahlen $L > 0, a_k \geq 0, h_k \geq 0$ und $b \geq 0$ sei

$$a_{k+1} \leq (1 + h_k L) a_k + h_k b, \quad k = 0, 1, \dots, N-1$$

erfüllt. Dann gelten die Abschätzungen

$$a_k \leq \frac{e^{Lt_k} - 1}{L} b + e^{Lt_k} a_0 \quad \text{mit} \quad t_k := \sum_{j=0}^{k-1} h_j \quad (k = 0, \dots, N)$$

Beweis. (vollständige Induktion)

Induktionsanfang ist für $k = 0$ offensichtlich gewährleistet. Der Schritt $k \rightarrow k+1$ 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

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

gilt für $k = 0, 1, \dots, N-1$

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

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.

$$\begin{aligned} |e_{k+1}| &\leq |e_k| + h_k |\Phi(t_k, y_k, h_k) - \Phi(t_k, y(t_k), h_k)| + |d_{k+1}| \\ &\leq (1 + h_k L_1) |e_k| + h_k C h_{\max}^p \end{aligned}$$

Die Abschätzung des Lemmas 8.9 liefert wegen $e_0 = 0$ die Behauptung des Satzes 8.8 □

8.2 Spezielle Einschrittverfahren

8.2.1 Euler-Verfahren

Mit der Verfahrensfunktion

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

erhält man mit

$$y_{k+1} = y_k + h_k f(t_k, y_k), \quad k = 0, \dots, N-1 \quad (8.11)$$

das Euler-Verfahren.

Für eine stetig partiell diff'bare Funktion $f : [a, b] \times \mathbb{R} \rightarrow \mathbb{R}$ 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.

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

8.2.2 Einschrittverfahren der Konsistenzordnung $p = 2$

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

$$\Phi(t, y, h) = a_1 f(t, y) + a_2 f(t + b_1 h, y + b_2 h f(t, y)), \quad t \in [a, b], \quad h \in [0, b-t], \quad y \in \mathbb{R} \quad (8.12)$$

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 $f : [a, b] \times \mathbb{R} \rightarrow \mathbb{R}$ zweimal stetig partiell diff'bar ist und für die Koeffizienten*

$$a_1 + a_2 = 1, \quad a_2 b_1 = \frac{1}{2}, \quad a_2 b_2 = \frac{1}{2} \quad (8.13)$$

gilt.

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

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

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

(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

$$y_{k+1} = y_k + h_k f \left(t_k + \frac{h_k}{2}, y_k + \frac{h_k}{2} f(t_k, y_k) \right), \quad k = 0, \dots, N-1 \quad (8.14)$$

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

$$y_{k+1} = y_k + \frac{h_k}{2} [f(t_k, y_k) + f(t_k + h_k, y_k + h_k f(t_k, y_k))], \quad k = 0, \dots, N-1 \quad (8.15)$$

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

8.3 Verfahren höherer Ordnung

Die bisher besprochenen Methoden (Euler, Heun) haben wir weitestgehend intuitiv ermittelt. Um systematisch Einschrittverfahren höherer Ordnung zu konstruieren, betrachten wir die zum AWP $y' = f(t, y), y(a) = y_0$ äquivalente Gleichung (nach Integration)

$$y(t) = y_0 + \int_a^t f(s, y(s)) ds \quad (8.16)$$

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

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(s, y(s)) ds \quad (8.17)$$

Das letzte Integral aus (8.17) approximieren wir durch eine Quadraturformel

$$\int_{t_k}^{t_{k+1}} f(s, y(s)) ds \quad (8.18)$$

wobei die s_l zu einer Zerlegung von $[t_k, t_{k+1}]$ gehören. (8.17) und (8.18) ergeben

$$y(t_{k+1}) \approx y(t_k) + h_k \sum_{l=1}^m \gamma_l f(s_l, y(s_l)) \quad (8.19)$$

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)

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

und mit der Approximation

$$y(t_{k+1}) \approx y(t_k) + h_k f(t_k, y(t_k))$$

ergibt sich mit

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

die Grundlage für das Verfahren von Heun.

25.
Vorle-
sung
27.01.2014

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:

$$\begin{aligned} k_1(t_k, y_k) &= f(t_k, y_k) \\ k_2(t_k, y_k) &= f(t_k + \alpha_2 h_k, y_k + h_k \beta_{21} k_1(t_k, y_k)) \\ k_3(t_k, y_k) &= f(t_k + \alpha_3 h_k, y_k + h_k (\beta_{31} k_1 + \beta_{32} k_2)) \\ &\vdots \\ k_m(t_k, y_k) &= f(t_k + \alpha_m h_k, y_k + h_k (\beta_{m1} k_1 + \dots + \beta_{mm-1} k_{m-1})) \end{aligned} \quad (8.20)$$

Ausgehend von (8.19) und (8.20) wird durch

$$y_{k+1} = y_k + h_k (\gamma_1 k_1(t_k, y_k) + \dots + \gamma_m k_m(t_k, y_k)) \quad (8.21)$$

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

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

folgt die sinnvolle Forderung

$$1 = \gamma_1 + \gamma_2 + \dots + \gamma_m \quad (8.22)$$

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

$$\alpha_l = \beta_{l1} + \dots + \beta_{l,l-1} \quad (8.23)$$

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

$$k_l \approx f(t_k + \alpha_l h_k, y(t_k) + h_k(\beta_{l1} k_1 + \dots + \beta_{l,l-1} k_{l-1})) .$$

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

$$y(s_l) = y(t_k) + h_k(\beta_{l1} + \dots + \beta_{l,l-1}) \quad (8.24)$$

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

$$y(s_l) = y(t_k) + \alpha_l h_k \quad (8.25)$$

und aus dem Vergleich von (8.24),(8.25) folgt

$$\alpha_l = \beta_{l1} + \dots + \beta_{l,l-1}$$

Definition 8.12. Die Tabelle mit den Koeffizienten $\alpha_l, \beta_{lr}, \gamma_r$ in der Form

$$\begin{array}{c|cccc}
 0 & & & & \\
 \alpha_2 & \beta_{21} & & & \\
 \alpha_3 & \beta_{31} & \beta_{32} & & \\
 \vdots & \vdots & \vdots & \ddots & \\
 \alpha_m & \beta_{m1} & \beta_{m2} & \dots & \beta_{mm-1} \\
 \hline
 & \gamma_1 & \gamma_2 & \dots & \gamma_{m-1} & \gamma_m
 \end{array} \quad (8.26)$$

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.

$$\begin{aligned}
 |d_{k+1}| &= |y(t_{k+1}) - y(t_k) - h\Phi(t_k, y(t_k), h)| \\
 &= \left| y(t_{k+1}) - y(t_k) - h \sum_{r=1}^m \gamma_r k_r(t_k, y(t_k)) \right| \\
 &\stackrel{(8.22)}{=} \left| y(t_{k+1}) - y(t_k) - hf(t_k, y(t_k)) - h \sum_{r=1}^m \gamma_r (k_r(t_k, y(t_k)) - f(t_k, y(t_k))) \right| \\
 &\leq \underbrace{|y(t_{k+1}) - y(t_k) - hy'(t_k)|}_{\in \mathcal{O}(h^2)} + h \left| \sum_{r=1}^m \gamma_r \underbrace{(k_r(t_k, y(t_k)) - f(t_k, y(t_k)))}_{\in \mathcal{O}(h)} \right|
 \end{aligned}$$

also

$$|d_{k+1}| \leq Ch^2$$

□

Bemerkung. Butcher hat bewiesen, wie groß die maximale Ordnung ist, welche mit einem m -stufigen Runge-Kutta-Verfahren erreichbar ist, was in der folgenden Tabelle notiert ist:

m	1	2	3	4	5	6	7	8	9	für $m \geq 9$
p	1	2	3	4	4	5	6	6	7	$p < m - 2$

8.4 Einige konkrete Runge-Kutta-Verfahren und deren Butcher-Tabellen

(i) Euler-Verfahren

$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array} \quad m = 1, \gamma_1 = 1$$

$$y_{k+1} = y_k + h_k f(t_k, y_k), \quad p = 1$$

(ii) Modifiziertes Euler-Verfahren

$$\begin{array}{c|cc} 0 & & \\ \hline \frac{1}{2} & \frac{1}{2} & \\ \hline & 0 & 1 \end{array} \quad m = 2, \gamma_1 = 0, \gamma_2 = 1, \alpha_2 = \frac{1}{2}, \beta_{21} = \frac{1}{2}$$

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

(iii) Verfahren von Runge von 3. Ordnung

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

$$m = 3, \gamma_1 = \gamma_2 = 0, \gamma_3 = 1, \alpha_2 = \frac{1}{2}, \alpha_3 = 1, \beta_{21} = \frac{1}{2}, \beta_{31} = 0, \beta_{32} = 1$$

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

(iv) Klassisches Runge-Kutta-Verfahren 4. Ordnung

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

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

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

$$\begin{aligned}\bar{k}_2 &= f(t_k + \Delta t, y(t_k) + \Delta y) \\ &= f + \alpha_2 h f_t + \alpha_2 h f f_y + \frac{1}{2} \alpha_2^2 h^2 f_{tt} + \alpha_2^2 h^2 f f_{ty} + \frac{1}{2} \alpha_2^2 h^2 f^2 f_{yy} + \mathcal{O}(h^3) \\ &=: f + \alpha_2 h F + \frac{1}{2} \alpha_2^2 h^2 G + \mathcal{O}(h^3)\end{aligned}\quad (8.28)$$

f, f_t, \dots, 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)

$$\begin{aligned}\bar{k}_3 &= f(t_k + \alpha_3 h, y(t_k) + h(\beta_{31} \bar{k}_1 + \beta_{32} \bar{k}_2)) \\ &= f + \alpha_3 h f_t + h(\beta_{31} \bar{k}_1 + \beta_{32} \bar{k}_2) f_y + \frac{1}{2} \alpha_3^2 h^2 f_{tt} \\ &\quad + \alpha_3 (\beta_{31} \bar{k}_1 + \beta_{32} \bar{k}_2) h^2 f_{ty} + \frac{1}{2} (\beta_{31} \bar{k}_1 + \beta_{32} \bar{k}_2)^2 h^2 f_{yy} + \mathcal{O}(h^3) \\ &= f + h(\alpha_3 f_t + [\beta_{31} + \beta_{32}] f f_y) + h^2 \left(\alpha_2 \beta_{32} F f_y \right. \\ &\quad \left. + \frac{1}{2} \alpha_3^2 f_{tt} + \alpha_3 [\beta_{31} + \beta_{32}] f f_{ty} + \frac{1}{2} (\beta_{31} + \beta_{32}) f^2 f_{yy} \right) + \mathcal{O}(h^3) \\ &= f + \alpha_3 h F + h^2 (\alpha_2 \beta_{32} F f_y + \frac{1}{2} \alpha_3^2 G) + \mathcal{O}(h^3)\end{aligned}\quad (8.29)$$

Mit (8.28) und (8.29) folgt für den lokalen Diskretisierungsfehler

$$\begin{aligned}d_{k+1} &= h(1 - \gamma_1 - \gamma_2 - \gamma_3) f + h^2 \left(\frac{1}{2} - \alpha_2 \gamma_2 - \alpha_3 \gamma_3 \right) F \\ &\quad + h^3 \left(\left[\frac{1}{6} - \alpha_2 \gamma_3 \beta_{32} \right] F f_y + \left[\frac{1}{6} - \frac{1}{2} \alpha_2^2 \gamma_2 - \frac{1}{2} \alpha_3^2 \gamma_3 \right] G \right) + \mathcal{O}(h^4)\end{aligned}\quad (8.30)$$

Aufgrund der Voraussetzungen werden die 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*

$$\begin{aligned}\gamma_1 + \gamma_2 + \gamma_3 &= 1 \\ \alpha_2 \gamma_2 + \alpha_3 \gamma_3 &= \frac{1}{2} \\ \alpha_2 \gamma_3 \beta_{32} &= \frac{1}{6} \\ \alpha_2^2 \gamma_2 + \alpha_3^2 \gamma_3 &= \frac{1}{3}\end{aligned}\quad (8.31)$$

hat das dazugehörige 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

$$\begin{aligned} \gamma_2 &= \frac{3\alpha_3 - 2}{6\alpha_2(\alpha_3 - \alpha_2)}, & \gamma_3 &= \frac{2 - 3\alpha_2}{6\alpha_3(\alpha_3 - \alpha_2)} \\ \gamma_1 &= \frac{6\alpha_2\alpha_3 + 2 - 3(\alpha_2 + \alpha_3)}{6\alpha_2\alpha_3}, & \beta_{32} &= \frac{\alpha_3(\alpha_3 - \alpha_2)}{\alpha_2(2 - 3\alpha_2)} \end{aligned} \quad (8.32)$$

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

$$\mathcal{M} = \{(\gamma_1, \gamma_2, \gamma_3, \alpha_2, \alpha_3, \beta_{32}) \mid \gamma_1, \gamma_2, \gamma_3, \beta_{32} \text{ gemäß (8.32)}, \\ \alpha_2, \alpha_3 \in \mathbb{R}, \alpha_2 \neq \alpha_3, \alpha_2 \neq \frac{2}{3}\}$$

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

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

26.
Vorlesung
29.01.2014

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

erläutert werden. Als lokaler Diskretisierungsfehler ergibt sich

$$d_{k+1}^{(H)} = y(t_{k+1}) - y(t_k) - \frac{1}{2}h[\bar{k}_1 + \bar{k}_2] \quad (8.33)$$

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

$$\begin{aligned} k_1 &= f(t_k, y_k) \\ k_2 &= f(t_k + h, y_k + hk_1) \\ k_3 &= f\left(t_k + \frac{h}{2}, y_k + \frac{h}{4}(k_1 + k_2)\right) \\ y_{k+1} &= y_k + \frac{h}{6}[k_1 + k_2 + 4k_3] \end{aligned} \quad (8.34)$$

ergibt. Für den lokalen Diskretisierungsfehler des Verfahrens (8.33) ergibt sich

$$d_{k+1}^{(RK)} = y(t_{k+1}) - y(t_k) - \frac{h}{6}[\bar{k}_1 + \bar{k}_2 + 4\bar{k}_3] \quad (8.35)$$

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 unbekanntenen 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

0			
1	1		
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	
	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{2}{3}$
	$\frac{1}{6}$	$\frac{1}{6}$	

In den letzten beiden Zeilen stehen erst die Gewichte für das Heun-Verfahren und darunter die Gewichte des Verfahrens, in das eingebettet wird.

8.5.2 Schrittweitensteuerung durch Extrapolation (nur zur Information, nicht prüfungsrelevant)

Zur Lösung des AWP's $y' = f(t, y)$, $y(a) = y_0$ wird für eine Verfahrensfunktion Φ mit der Konsistenzordnung $p \geq 1$ die Vorschrift

$$\left. \begin{aligned} w &= y_k + \frac{h_k}{2} \Phi\left(t_k, y_k, \frac{h_k}{2}\right), \\ y_{k+1} &= w + \frac{h_k}{2} \Phi\left(t_k + \frac{h_k}{2}, w, \frac{h_k}{2}\right), \\ t_{k+1} &:= t_k + h_k, \quad k = 0, 1, \dots \end{aligned} \right\} \quad (8.36)$$

betrachtet. Nun wird eine adaptive Wahl der Schrittweiten h_k diskutiert mit dem Ziel einer effizienten Fehlerkontrolle.

Ausgehend von einer gegebenen Stelle $t_k \in [a, b]$ und einer gegebenen Näherung $y_k \approx y(t_k)$ soll eine Schrittweite $h_k > 0$ bestimmt werden, für die

$$|y_{k+1} - z(t_k + h_k)| \approx \epsilon_{tol} \quad (8.37)$$

erfüllt ist, wobei y_{k+1} aus einem Schritt des Verfahrens (8.36) hervorgeht, $\epsilon_{tol} > 0$ eine vorgegebene Fehlerschranke ist, und $z : [t_k, b] \rightarrow \mathbb{R}$ die Lösung des AWP's

$$z' = f(t, z), \quad t \in [t_k, b]; \quad z(t_k) = y_k, \quad (8.38)$$

ist.

Bemerkung 8.15. Die Forderung (8.37) bedeutet, dass die angestrebte Schrittweitensteuerung auf einer Vorgabe des lokalen Verfahrensfehlers beruht.

Die Lösung des AWP's (8.38) ist nicht bekannt, also insbesondere $z(t_k + h_k)$, und muss erst noch bestimmt werden.

Zur Vereinfachung der Notation wird die Bezeichnung für einen von dem Punkt (t_k, y_k) ausgehenden Verfahrensschritt (8.36) mit der Länge h eingeführt (... 2 Halbschritte mit der Schrittweite $\frac{h}{2}$),

$$y_{2 \times h/2} = w + \frac{h_k}{2} \Phi\left(t_k + \frac{h_k}{2}, w, \frac{h_k}{2}\right) \text{ mit } w = y_k + \frac{h_k}{2} \Phi\left(t_k, y_k, \frac{h_k}{2}\right). \quad (8.39)$$

Zur Bestimmung einer Schrittweite h_k , mit der die Forderung (8.37) annähernd erfüllt wird, geht man von einer nicht zu kleinen Startschrittweite $h^{(0)}$ aus, und für $j = 0, 1, \dots$, führt man den folgenden Algorithmus aus:

- 1) Berechnung von $y_{2 \times h/2}$.
- 2) Ermittlung einer Schätzung für den Fehler $|y_{2 \times h/2} - z(t_k + h)|$ und Abbruch des Iterationsprozesses mit $j_{\epsilon_{tol}} = j$, falls die Schätzung kleiner gleich ϵ_{tol} ausfällt.
- 3) Anderenfalls, falls diese Schätzung größer als ϵ_{tol} ist, wird eine neue Testschrittweite $h^{(s+1)} < h^{(s)}$ bestimmt.

Wie man den unbekanntem Wert $z(t_k + h)$ schätzt und im Falle von 3) die neue Testschrittweite $h^{(s+1)}$ bestimmt, soll im Folgenden beschrieben werden. Der Wert $z(t_k + h_k)$ wird mittels lokaler Extrapolation mittels $z_{h^{(s)}}$ geschätzt, wobei man mit $v_h = y_k + h\Phi(t_k, y_k, h)$, also einem Schritt mit der Schrittweite $h = h^{(s)}$, und $y_{2 \times h/2}$

$$z_h = \underbrace{y_{2 \times h/2} - \frac{v_h - y_{2 \times h/2}}{2^p - 1}}_{z(t_k + h^{(s)}) + O(h^{p+2})} \approx z(t_k + h^{(s)})$$

erhält. Der Fehler $|y_{2 \times h^{(s)}/2} - z(t_k + h^{(s)})|$ berechnet sich dann näherungsweise zu

$$\delta^{(s)} = |y_{2 \times h^{(s)}/2} - z(t_k + h^{(s)})| = \frac{|v_h - y_{2 \times h^{(s)}/2}|}{2^p - 1}. \quad (8.40)$$

Zur Bestimmung der neuen Testschrittweite $h^{(s+1)}$ benutzt man die näherungsweise Darstellung des Fehlers $y_{2 \times h/2} - z(t_k + h)$. Dazu benutzen wir ein Ergebnis der Asymptotik des globalen Verfahrensfehlers, das hier nicht bewiesen wird.

Lemma 8.16. *Mit den Notationen (8.38)-(8.40) gilt*

$$|y_{2 \times h/2} - z(t_k + h)| = \left(\frac{h}{h^{(s)}}\right)^{p+1} \delta^{(s)} + O((h^{(s)})^{p+2}), \quad 0 < h \leq h^{(s)}. \quad (8.41)$$

Gilt also $(h^{(s)})^{p+2} \ll \epsilon_{tol}$, so gewinnt man aus der Darstellung (8.41) unter Vernachlässigung des Restgliedes die neue Testschrittweite

$$h^{(s+1)} = \left(\frac{\epsilon_{tol}}{\delta^{(s)}}\right)^{1/(p+1)} h^{(s)} \quad (8.42)$$

und wiederholt damit den oben beschriebenen Algorithmus mit s um eins erhöht.

8.6 Implizite Runge-Kutta-Verfahren

Explizite Verfahren neigen zur Instabilität und damit besteht die Gefahr der Verstärkung von Rundungsfehlern.

Implizite Verfahren erweisen sich als stabil, speziell, wenn es sich um die Lösung von AWP's mit sogenannten steifen DGL handelt.

Im Unterschied zum Gleichungssystem (8.20) wird beim impliziten Runge-Kutta-Verfahren das Gleichungssystem

$$k_r(t_k, y_k) = f(t_k + \alpha_r h_k, y_k + h_k(\beta_{r1}k_1 + \dots + \beta_{rm}k_m)), \quad r = 1, \dots, m \quad (8.43)$$

zur Bestimmung der k_r zugrunde gelegt.

Mit (8.43) wird (8.21) zu einem impliziten Runge-Kutta-Verfahren. Aus (8.43) ergibt sich die Butcher-Tabelle

$$\begin{array}{c|ccc} \alpha_1 & \beta_{11} & \dots & \beta_{1m} \\ \alpha_2 & \beta_{21} & \dots & \beta_{2m} \\ \vdots & \vdots & & \vdots \\ \alpha_m & \beta_{m1} & \dots & \beta_{mm} \\ \hline & \gamma_1 & \dots & \gamma_m \end{array} \quad (8.44)$$

Die Überlegungen, die bei den expliziten Verfahren die Bedingung (8.23) für die Koeffizienten α_r, β_{rl} gerechtfertigt haben, ergeben analog bei den impliziten Runge-Kutta-Verfahren die Bedingung

$$\alpha_r = \beta_{r1} + \beta_{r2} + \dots + \beta_{rm}, \quad r = 1, \dots, m \quad (8.45)$$

Zur Lösbarkeit des Gleichungssystems (8.43) gilt der

Satz 8.17. f genüge auf $[a, b] \times \mathbb{R}$ der Lipschitz-Bedingung

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

und die Schrittweite $h = h_k$ genüge der Bedingung

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

Dann hat (8.43) zur Bestimmung von k_1, \dots, k_m genau eine Lösung

Beweis. Aussage folgt aus dem Banachschen Fixpunktsatz. \square

8.7 Rundungsfehleranalyse von expliziten Einschrittverfahren

Zur numerischen Lösung eines AWP betrachten wir das Verfahren

$$y_{k+1} = y_k + h_k \Phi(t_k, y_k, h_k), \quad k = 0, 1, 2, \dots, N-1 \quad (8.46)$$

mit der Verfahrensfunktion Φ . Durch Rundungsfehler arbeitet man statt (8.46) mit einem Verfahren der Form

$$y_{k+1} = y_k + h_k \Phi(t_k, y_k, h_k) + \rho_k, \quad k = 0, 1, \dots, N-1 \quad (8.47)$$

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

mit gewissen Zahlen $e_0, \rho_k \in \mathbb{R}$

Für die Rundungsfehler infolge des Verfahrens (8.47) gilt der folgende

Satz 8.18. *Zur Lösung des AWP $y' = f(t, y), y(a) = y_0$, sei durch (8.46) ein Einschrittverfahren mit der Konsistenzordnung $p \geq 1$ gegeben, wobei die Verfahrensfunktion bezüglich der 2. Variablen Lipschitz-stetig mit der Konstanten $L > 0$ ist.*

Dann gelten für die durch die fehlerbehaftete Verfahrensvorschrift (8.47) gewonnenen Approximationen die Abschätzungen

$$\max_{k=0, \dots, N} |y_k - y(t_k)| \leq K \left(h_{\max}^p + \frac{\delta}{h_{\min}} \right) + e^{L(b-a)} \epsilon \quad (8.48)$$

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

8.8 Ein Anwendungsgebiet für Löser von AWP's

Eine wichtige Anwendung der numerischen Lösungsverfahren für Anfangswertprobleme ist die Lösung von Zweipunkt-Randwertproblemen mit Schießverfahren.

Schießverfahren zur Lösung von Zweipunkt-Randwertproblemen basieren auf Methoden zur Lösung von Anfangswertproblemen. Beim sogenannten **ersten Randwertproblem**

$$y'' = f(x, y), \quad y(a) = \eta_a, \quad y(b) = \eta_b \quad (8.49)$$

nutzt man dabei z.B. die Randbedingung $y(a) = \eta_a$ als Anfangsbedingung und versucht durch eine geeignete Wahl von $\zeta_a = y'(a)$ als Anfangsbedingung für die Ableitung mit einer Lösung des Anfangswertproblems

$$y'' = f(x, y), \quad y(a) = \eta_a, \quad y'(a) = \zeta \quad (8.50)$$

die Randbedingung $y(b) = \eta_b$ zu treffen. Für vorgegebenes ζ sei $y(x, \zeta)$ die Lösung von (8.50). $y(x, \zeta)$ ist dann Lösung des Zweipunkt-Randwertproblems (8.49), wenn ζ Nullstelle der Funktion

$$g(\zeta) = y(b, \zeta) - \eta_b \quad (8.51)$$

ist. Für eine Funktionswertberechnung von g ist ein Anfangswertproblem (8.49) zu lösen. Eine Möglichkeit zur Bestimmung der Nullstelle von g ist mit dem Bisektionsverfahren gegeben. Allerdings ist es durchaus möglich, dass durch Fehler bei der Lösung des Anfangswertproblems das Vorzeichen von g nicht immer korrekt berechnet werden kann, so dass das Bisektionsverfahren unbrauchbar wird.

Eine andere Möglichkeit zur Bestimmung der Nullstelle von g bietet das Newton-Verfahren. Die Differentiation von g nach ζ ergibt

$$g'(\zeta) = y_\zeta(b, \zeta), \quad (8.52)$$

wobei $y_\zeta(b, \zeta)$ die partielle Ableitung von $y(x, \zeta)$ nach ζ ausgewertet an der Stelle $x = b$ ist. Die Differentiation der Gleichung $y''(x, \zeta) = f(x, y(x, \zeta))$ nach ζ ergibt

$$\frac{\partial}{\partial \zeta} [y''(x, \zeta)] = f_y(x, y(x, \zeta)) y_\zeta(x, \zeta). \quad (8.53)$$

f_y bedeutet dabei die partielle Ableitung von $f(x, y)$ nach y . Mit der Voraussetzung der Vertauschbarkeit der Ableitungen nach ζ und x erhält man aus (8.53) die Differentialgleichung 2. Ordnung

$$y_\zeta''(x, \zeta) = f_y(x, y(x, \zeta)) y_\zeta(x, \zeta) \quad (8.54)$$

für $y_\zeta(x, \zeta)$. Durch Differentiation der Anfangsbedingungen der Aufgabe (8.50) nach ζ erhält man die Anfangsbedingungen

$$y_\zeta(a, \zeta) = 0, \quad y'_\zeta(a, \zeta) = 1. \quad (8.55)$$

Mit (8.54), (8.55) liegt ein Anfangswertproblem zur Berechnung von $y_\zeta(x, \zeta)$, also auch zur Berechnung der Ableitung von g vor (gemäß (8.52)). Damit kann man durch Lösung der Anfangswertprobleme (8.50) und (8.54), (8.55) Funktionswert und Ableitung von $g(\zeta)$ berechnen und kann somit ein Newton-Verfahren zur Nullstellenberechnung von g durchführen. Hierzu ist anzumerken, dass man zur Lösung von (8.54), (8.55) die Funktion $y(x, \zeta)$ als Lösung des Anfangswertproblems (8.50) benötigt, um die Funktionswerte von $f_y(x, y(x, \zeta))$ berechnen zu können. Da man die exakte Lösung $y(x, \zeta)$ nicht zur Verfügung hat, verwendet man die Näherungswerte y_k an den Stützstellen x_k des Intervalls $[a, b]$ zur Berechnung von f_y an den Stützstellen x_k . Beim Schießverfahren ist es in jedem Fall sinnvoll, ein recht genaues Verfahren zur erforderlichen Lösung der Anfangswertprobleme (8.50) und (8.54), (8.55) zu verwenden, da speziell bei wachsenden Lösungen die Sensibilität der Lösung $y(x, \zeta)$ von ζ sehr groß sein kann und somit kleine Änderungen von ζ große Auswirkungen auf $y(b, \zeta)$ haben können. Schießverfahren kann man bei nicht-linearen Problemen anwenden, da bei den benötigten Integrationsverfahren für gewöhnliche Differentialgleichungen die Linearität der Gleichungen nicht notwendig ist.

Kapitel A

Anhang

A.1 Eigenschaften von Matrizen im Ergebnis von FD-Schemen

Die Diskretisierung eines elliptischen Randwertproblems führt bei einer geeigneten Nummerierung der Unbekannten (Funktionswerte der Gitterfunktionen u_h) auf lineare Gleichungssysteme mit Koeffizientenmatrix im eindimensionalen bzw. zweidimensionalen Fall

$$A = (a_{ij}) = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix}, \quad \mathcal{A} = \begin{pmatrix} A_1 & -T & \mathbf{0} & \dots & \mathbf{0} \\ -T & A_2 & -T & \ddots & \mathbf{0} \\ & \ddots & \ddots & \ddots & \\ \mathbf{0} & \ddots & -T & A_{m-1} & -T \\ \mathbf{0} & \dots & \mathbf{0} & -T & A_m \end{pmatrix}, \quad (\text{A.1})$$

wobei A_j tridiagonale Matrizen und T Diagonalmatrizen sind. Neben der sparsamen Besetztheit sind die Matrizen von FD-Schemen zur numerischen Lösung von elliptischen Randwertproblemen oder parabolischen Rand-Anfangswert-Problemen dadurch gekennzeichnet, dass

$$a_{ij} \leq 0, \quad i \neq j, \quad a_{ii} > 0, \quad \text{und} \quad |a_{ii}| \geq \sum_{j=1, i \neq j}^n |a_{ij}|, \quad i = 1, \dots, n,$$

gilt. Die Matrizen sind also zumindest schwach diagonal dominant. Andere Eigenschaften werden in der folgenden Definition zusammengefasst.

Definition A.1.

Sei $A = (a_{ij}) \in \mathbb{R}^{n \times n}$, dann heißt A

- 1) L_0 -Matrix, wenn $a_{ij} \leq 0$, $i \neq j$ gilt, die Menge aller L_0 -Matrizen aus $\mathbb{R}^{n \times n}$ bezeichnet man auch mit $\mathbb{Z}^{n \times n}$,

2) *L-Matrix*, wenn A eine L_0 -Matrix ist und $a_{ii} > 0$ gilt,

3) *M-Matrix*, wenn $A \in \mathbb{Z}^{n \times n}$, A^{-1} existiert und $A^{-1} \geq 0$ gilt.

Außerdem haben die Matrizen oft die wichtige Eigenschaft, irreduzibel diagonal dominant zu sein,

d.h. die Matrizen sind irreduzibel und es gilt

$$|a_{ii}| \geq \sum_{j=1, j \neq i}^n |a_{ij}| \quad \text{für } i = 1, \dots, n,$$

wobei die Ungleichung für mindestens einen Index i_0 strikt ist.

Bemerkung A.2.

Die Relationen \leq , \geq , $<$ bzw. $>$ für Matrizen oder Vektoren sind so zu verstehen, dass dann jeweils die Relationen für alle Elemente der Matrizen bzw. Komponenten der Vektoren gelten.

Es sollen nun im Folgenden die wichtigsten Aussagen für die in der Def. A.1 erklärten Matrizen zusammen gefasst werden.

Satz A.3. *Die Aussagen*

(a.1) A ist invertierbar und es gilt $A^{-1} \geq 0$,

(a.2) $Ax \leq 0 \implies x \leq 0$,

(a.3) $Ax \leq Ay \implies x \leq y$

sind äquivalent.

Die Eigenschaften (a.2) oder auch (a.3) nennt man auch *Inversmonotonie* von A .

Beweis.

Die Implikation (a.3) \implies (a.2) erhält man mit $y = 0$, und die Implikation (a.2) \implies (a.3) erhält man durch Anwendung von (a.2) auf $z = x - y$.

Nachweis von (a.2) \implies (a.1):

Sei $Ax = 0$, dann ist $A(\pm x) = \pm Ax \leq 0$, woraus $\pm x \leq 0$, also $x = 0$ folgt, was die Injektivität von $l(x) := Ax$ bedeutet. Bijektivität folgt aus der Endlichdimensionalität des \mathbb{R}^n (bzw. \mathbb{C}^n). Damit existiert A^{-1} . (a.2) bedeutet mit $Ax = -y$

$$Ax \leq 0 \implies x \leq 0 \quad \text{bzw.} \quad y \geq 0 \implies A^{-1}y \geq 0.$$

Setzen $y = (\delta_{ik})_{k \in I}$ für festes $k \in I$, damit folgt für alle $l \in W$

$$0 \leq (A^{-1}y)_l = \sum_{i=1}^n (A^{-1})_{li} y_i = (A^{-1})_{lk}.$$

Nachweis von (a.1) \implies (a.2):

Für $y \geq 0$ gilt

$$(A^{-1}y)_i = \sum_{j=1}^n (A^{-1})_{ij} y_j \geq 0 ,$$

also $A^{-1}y \geq 0$. Sei jetzt $Ax \leq 0$, dann ist $y = -Ax \geq 0$ und schließlich

$$A^{-1}y = -x \geq 0 \quad \text{bzw.} \quad x \leq 0 ,$$

und damit ist der Satz vollständig bewiesen. \square

Aus der linearen Algebra sei an das Kriterium von Gerschgorin erinnert:

Satz A.4. (*Gerschgorin-Kriterium*)

Sei $K_r(z) = \{\xi \in \mathbb{C} , |\xi - z| < r\}$. Dann gilt

(a.4) *Alle Eigenwerte von A liegen in den Gerschgorin-Kreisen*

$$\bigcup_{i=1}^n \bar{K}_{r_i}(a_{ii}) \quad \text{mit} \quad r_i = \sum_{j=1, j \neq i}^n |a_{ij}| .$$

(a.5) *Ist A irreduzibel, dann liegen die Eigenwerte sogar in*

$$\left[\bigcup_{i=1}^n K_{r_i}(a_{ii}) \right] \cup \left[\bigcap_{i=1}^n \partial K_{r_i}(a_{ii}) \right] .$$

Beweis. Sei λ ein EW von A mit dem EV x und (o.B.d.A.) $\|x\|_\infty = 1$ und für $j \in W = \{1, 2, \dots, n\}$ gelte $|x_j| = 1$. Aus $|x_j| = 1$ folgt

$$\lambda x_j = (Ax)_j = \sum_{k=1}^n a_{jk} x_k = a_{jj} + \sum_{k=1, k \neq j}^n a_{jk} x_k$$

$$\implies (\lambda - a_{jj}) x_j = \sum_{k=1, k \neq j}^n a_{jk} x_k$$

$$\implies |\lambda - a_{jj}| |x_j| \leq \sum_{k=1, k \neq j}^n |a_{jk}| \underbrace{|x_k|}_{\leq 1} \leq \sum_{k=1, k \neq j}^n |a_{jk}| \quad (\text{A.2})$$

$$\implies |\lambda - a_{jj}| \leq \sum_{k=1, k \neq j}^n |a_{jk}| = r_j \quad (\text{A.3})$$

Damit folgt $\lambda \in \bigcup_{i=1}^n \bar{K}_{r_i}(a_{ii})$, also (a.4).

Zum Nachweis von (a.5) zeigen wir für Eigenwerte λ , die nicht in $\bigcup_{i=1}^n K_{r_i}(a_{ii})$ liegen zuerst, dass im Falle $a_{ji} \neq 0$ gilt:

$$\text{aus } |x_j| = 1 \text{ und } |\lambda - a_{jj}| = r_j \implies |x_i| = 1 \text{ und } |\lambda - a_{ii}| = r_i . \quad (\text{A.4})$$

Nehmen wir an, dass $|x_i| < 1$ ist:

Es gilt nun

$$\begin{aligned}
 r_j &= |\lambda - a_{jj}| \\
 &= |(\lambda - a_{jj})x_j| = \left| \sum_{k=1, k \neq j}^n a_{jk} x_k \right| \\
 &\leq \sum_{k=1, k \neq j, i}^n |a_{jk}| \underbrace{|x_k|}_{\leq 1} + |a_{ji}| \underbrace{|x_i|}_{< 1} \\
 &< \sum_{k=1, k \neq j}^n |a_{jk}| = r_j, \quad ,
 \end{aligned}$$

das ist ein Widerspruch, also war unsere Annahme falsch, und es gilt $|x_i| = 1$. Aufgrund der Irreduzibilität folgt aus (A.4), dass

$$\lambda \in \bigcap_{k=1}^n \partial K_{r_k}(a_{kk})$$

gilt (man fängt in der j -ten Zeile von A (mit $\|x\|_\infty = |x_j|$) an, und kann sich wegen der Irreduzibilität durch die Matrix hangeln), so dass damit letztendlich für alle Indizes $k = 1, \dots, n$ die Gültigkeit von $|\lambda - a_{kk}| = r_k$, also (a.5) folgt. \square

Bemerkung A.5. Eine Folgerung aus dem Satz ist, dass Eigenwerte von irreduziblen Matrizen, die nicht in der Vereinigung der offenen Gerschgorin-Kreise liegen, in der Schnittmenge aller Ränder der Gerschgorin-Kreise liegen.

Nun kann man den Satz zur Konvergenz des Jacobi-Verfahrens für irreduzible diagonal dominante Matrizen unter Nutzung des Satzes A.4 zeigen.

Satz A.6. Sei A eine L -Matrix und irreduzibel diagonal dominant. Dann gilt mit der Aufspaltung $A = D + L + U = D - R$ und D gleich dem Diagonalanteil von A sowie $R = -(L + U)$ dem negativen Außendiagonalanteil von A

$$\rho(D^{-1}R) < 1.$$

Beweis. Sei $C = D^{-1}R = (c)_{ij}$, dann gilt

$$c_{ij} = -\frac{a_{ij}}{a_{ii}} \quad \text{für } i \neq j, \quad \text{und} \quad c_{ii} = 0.$$

Es gilt nun für

$$r_\alpha = \sum_{\beta=1, \beta \neq \alpha}^n |c_{\alpha\beta}| < 1$$

wegen der irreduziblen Diagonaldominanz für mindestens einen Index α . Für alle $\beta \in W$ gilt $r_\beta \leq 1$. Nach Gerschgorin liegen die EW von C in

$$\underbrace{\left[\bigcup_{i=1}^n K_{r_i}(0) \right]}_{\subset K_1(0)} \cup \left[\bigcap_{i=1}^n \partial K_{r_i}(0) \right].$$

Es bleibt zu zeigen:

$$\bigcap_{i=1}^n \partial K_{r_i}(0) \subset K_1(0).$$

Fall a):

$$r_\beta = r \quad \text{für alle } \beta \in W.$$

Wegen $r_\alpha = r < 1$ folgt

$$\bigcap_{\beta=1}^n \partial K_{r_\beta}(0) = \partial K_r(0) \subset K_1(0),$$

also sind alle EW von C betragsmäßig kleiner als 1, damit gilt $\rho(C) < 1$.

Fall b):

Die r_β sind nicht alle gleich. Dann gilt

$$\bigcap_{\beta=1}^n \partial K_{r_\beta}(0) = \emptyset,$$

d.h. nach Gerschgorin liegen die EW von C in $K_1(0)$, also gilt $|\lambda| < 1$ und damit auch $\rho(C) < 1$. \square

Bemerkung A.7.

Die soeben bewiesene Aussage für irreduzibel diagonal dominante Matrizen gilt auch für strikt diagonal dominante Matrizen (die nicht irreduzibel sein müssen).

Satz A.8.

Sei $A = D - R$ eine L-Matrix ($a_{ii} > 0$, $a_{ij} \leq 0$, $i \neq j$), dann gilt mit D , C und R aus dem obigen Satz

$$A \text{ ist } M\text{-Matrix} \iff \rho(D^{-1}R) < 1.$$

Beweis.

Richtung \implies :

Sei A eine M -Matrix. Sei λ ein EW von $D^{-1}R$ mit dem EV $u \neq 0$. Dann gilt

$$|\lambda||u| = |\lambda u| = |D^{-1}Ru| \leq D^{-1}R|u| ,$$

wegen $A^{-1}D \geq 0$ folgt

$$-A^{-1}DD^{-1}R|u| \leq -A^{-1}D|\lambda||u|$$

und

$$\begin{aligned} |u| &= A^{-1}A|u| = A^{-1}(D - R)|u| \\ &= A^{-1}D(E - D^{-1}R)|u| \\ &\leq A^{-1}D|u| - A^{-1}D|\lambda||u| \\ &= (1 - |\lambda|) \underbrace{A^{-1}D}_{\geq 0} |u| . \end{aligned}$$

Für $|\lambda| \geq 1$ folgt $|u| \leq 0$, d.h. $u = 0$, also muss $|\lambda| < 1$ für alle EW gelten, und damit gilt $\rho(D^{-1}R) < 1$.

Richtung \impliedby :

Sei $\rho(D^{-1}R) < 1$. Damit konvergiert die Neumann-Reihe mit $C = D^{-1}R$ und es gilt

$$S = \sum_{\nu=0}^{\infty} C^{\nu} = (E - C)^{-1} .$$

Wegen $D^{-1} \geq 0$ und $R \geq 0$ folgt

$$C \geq 0 , \quad C^{\nu} \geq 0 , \quad S \geq 0 .$$

Aus

$$E = S(E - C) = SD^{-1}(D - R) = SD^{-1}A$$

folgt

$$A^{-1} = SD^{-1} \implies A^{-1} \geq 0 \implies A \text{ ist } M\text{-Matrix} .$$

□

Aus den Sätzen A.6, A.8 folgt unmittelbar die wichtige Aussage

Satz A.9.

Sei A eine L -Matrix ($a_{ii} > 0$, $a_{ij} \leq 0$, $i \neq j$). Ist A strikt diagonal dominant oder irreduzibel diagonal dominant, dann ist A eine M -Matrix.

Ein weiteres Kriterium für M -Matrizen liefert der folgende

Satz A.10. Sei $A \in \mathbb{Z}^{n \times n}$ ($a_{ij}, i \neq j$) und $D = \text{diag}(a_{11}, \dots, a_{nn})$ die Diagonale von A . Dann sind die Aussagen

1. A regulär und $A^{-1} \geq 0$ und
2. $D > 0, C = E - D^{-1}A = -D^{-1}(L + U) \geq 0, \rho(C) < 1,$

äquivalent, d.h. für jede M -Matrix gilt Aussage 2.

Beweis.

1. \implies 2.

Wir nehmen $a_{ii} \leq 0$ für ein bel. Index i an. Sei a_i die i -te Spalte von A . Da $A^{-1}A = E$ ist, folgt

$$A^{-1}a_i = e_i \text{ (kanonischer Einheitsvektor),}$$

da $a_{ij} \leq 0, i \neq j$ ist, gilt $a_i \leq 0$, also

$$A^{-1}a_i \leq 0,$$

was einen Widerspruch zu $A^{-1}a_i = e_i > 0$ bedeutet.

Mit $a_{ii} > 0$ ist $D > 0$ und regulär. Damit folgt für $\tilde{A} := D^{-1}A$

$$\tilde{A}^{-1} = A^{-1}D \geq 0.$$

$C = E - \tilde{A}$ hat die Diagonaleinträge $c_{ii} = 0$. Für die restlichen Matrixelemente gilt $c_{ij} = 0 - a_{ij}/a_{ii} \geq 0$ und damit $C \geq 0$.

Nach dem Satz von Perron-Frobenius¹ existiert für die Matrix $C \geq 0$ zu $\lambda = \rho(C)$ ein positiver Eigenvektor $x \geq 0$. Damit gilt

$$\begin{aligned} Cx = \lambda x &\iff (E - A^{-1}D)x = A^{-1}D\lambda x \\ &\iff (A^{-1}D - A^{-1}D\lambda)x = x \\ &\iff \tilde{A}^{-1}(1 - \lambda)x = x. \end{aligned}$$

Da sowohl $\tilde{A}^{-1} \geq 0$ als auch $x \geq 0$ gilt, ergibt sich

$$1 - \lambda \geq 0 \implies 0 \leq \rho(C) = \lambda < 1.$$

2. \implies 1.

Da $\rho(C) < 1$ und $C \geq 0$ ist, folgt mit der konvergenten Neumannschen Reihe $(E - C)^{-1} \geq 0$, und damit ergibt sich

$$0 \leq (E - C)^{-1}D^{-1} = (D^{-1}A)^{-1} = A^{-1}DD^{-1} = A^{-1}.$$

□

¹Satz von Perron-Frobenius: $A \in \mathbb{R}^{n \times n}, n > 1$ sei irreduzibel und $A \geq 0$. Dann gilt $\rho(A) > 0$ ist einfacher EW von A , zu $\lambda = \rho(A)$ gehört ein positiver EV $x > 0$. Und es gilt $\rho(B) > \rho(A)$ f.a. $B \not\geq A$

Diesen Satz kann man noch verschärfen, es gilt

Satz A.11. Sei $A \in \mathbb{Z}^{n \times n}$, $D = \text{diag}(a_{11}, \dots, a_{nn})$ und $C = E - D^{-1}A$. Dann sind die Aussagen

1. A regulär und $A^{-1} > 0$ und
2. $D > 0$, $C \geq 0$, $\rho(C) < 1$, C irreduzibel

äquivalent.

Beweis. Es soll nur die Richtung 1. \implies 2. gezeigt werden. Wäre A reduzibel, so ließe sich A in der Form

$$A = \begin{pmatrix} A_{11} & A_{12} \\ \mathbf{0} & A_{22} \end{pmatrix}$$

mit quadratischen Blöcken A_{11} , A_{22} darstellen, und die Inverse hätte die Struktur

$$A^{-1} = \begin{pmatrix} A_{11}^{-1} & -A_{11}^{-1}A_{12}A_{22}^{-1} \\ \mathbf{0} & A_{22}^{-1} \end{pmatrix}$$

wobei der linke untere Null-Block im Widerspruch zu $A^{-1} > 0$ steht, also ist A irreduzibel. Da sich C und A in ihrer Struktur nur auf ihrer Diagonalen unterschieden, ist C auch irreduzibel. \square

Koeffizientenmatrizen A von FD- oder FV-Diskretisierungen sind oftmals M -Matrizen und es gelten meistens die Voraussetzungen des Satzes A.11 und damit auch die Aussage 2, d.h. für die Iterationsmatrix des Jacobi-Verfahrens C gilt dann $C \geq 0$ und $\rho(C) < 1$. Der folgende Satz liefert die Grundlage zum Vergleich der Konvergenz von Verfahren, bei denen die Iterationsmatrizen der Relation $C_1 \geq C_2 \geq 0$ genügen.

Satz A.12. Sei $A \geq 0$, $A \in \mathbb{R}^{n \times n}$. Dann gilt

$$0 < \rho(A) \text{ ist EW von } A, \text{ d.h. } \rho(A) \in \sigma(A), \quad (\text{A.5})$$

$$\text{zu } \lambda = \rho(A) \text{ gehört ein nichtnegativer EV } x \geq 0, \quad (\text{A.6})$$

$$\rho(B) \geq \rho(A) \text{ für alle } B \geq A. \quad (\text{A.7})$$

Beweis. Da der Fall $n = 1$ trivial ist, nehmen wir $n > 1$ an. Wir setzen $A_\epsilon = A + \epsilon E = (a_{ij} + \epsilon)$ für $\epsilon > 0$. A_ϵ ist irreduzibel und nach dem Satz von Perron-Frobenius ist $\lambda_\epsilon = \rho(A_\epsilon)$ ein EW von A_ϵ zum EV $x_\epsilon > 0$, $\|x_\epsilon\|_\infty = 1$. Da die EW als Polynomnullstellen stetig von A_ϵ abhängen, ist

$$\lambda := \lim_{\epsilon \rightarrow 0} \lambda_\epsilon = \lim_{\epsilon \rightarrow 0} \rho(A_\epsilon) = \rho(A)$$

EW von A . Da $\{x \mid \|x\|_\infty = 1\}$ kompakt ist, gibt es eine konvergente Teilfolge $x_{\epsilon_\nu} \rightarrow x$ mit $\|x\|_\infty = 1$ und $x \geq 0$. Aus $A_{\epsilon_\nu} x_{\epsilon_\nu} = \lambda_{\epsilon_\nu} x_{\epsilon_\nu}$ folgt $Ax = \lambda x$, d.h. $x \geq 0$ ist EV.

Nun sei $B_{2\epsilon}$ in Analogie zu A_ϵ definiert. Aus den Ungleichungen $B_{2\epsilon} > A_\epsilon$ und $\rho(B_{2\epsilon}) > \rho(A_\epsilon)$ ergibt sich $\rho(B) \geq \rho(A)$ für $\epsilon \rightarrow 0$. \square

Die am Anfang des Abschnittes angegebene Matrix im Ergebnis der Diskretisierung eines elliptischen Randwertproblems ist eine irreduzibel diagonal dominante L -Matrix, also eine M -Matrix. Mit den bisher diskutierten Eigenschaften dieser Matrizen kann man nun zeigen, dass das Gauß-Seidel-Verfahren schneller als das Jacobi-Verfahren konvergiert. Dazu brauchen wir noch eine Eigenschaft des Verfahrens

$$Bx^{(k+1)} = Rx^{(k)} + b \quad (\text{A.8})$$

zur Lösung von $Ax = b$, wobei A durch

$$A = B - R \quad (\text{A.9})$$

mit einer regulären Matrix B aufgespaltet sein soll.

Definition A.13. Die Matrix $B \in \mathbb{R}^{n \times n}$ aus (A.9) beschreibt eine **reguläre Aufspaltung** von $A \in \mathbb{R}^{n \times n}$, falls

$$B \text{ regulär, } B^{-1} \geq 0, \quad B \geq A. \quad (\text{A.10})$$

Zur Konvergenz eines Iterationsverfahrens (A.8) für eine reguläre Aufspaltung von A betrachten wir

Satz A.14. Sei $A^{-1} > 0$ (es reicht auch aus, wenn A eine M -Matrix ist). B beschreibt eine reguläre Aufspaltung von A ($A = B - R$). Dann konvergiert das Iterationsverfahren (A.8) und es gilt

$$\rho(S) = \rho(B^{-1}R) = \frac{\rho(A^{-1}R)}{1 + \rho(A^{-1}R)} < 1. \quad (\text{A.11})$$

Beweis. Wegen $\rho(A^{-1}R)/(1 + \rho(A^{-1}R)) < 1$ genügt es,

$$\rho(B^{-1}R) = \frac{\rho(C)}{1 + \rho(C)}$$

für $C = A^{-1}R$ zu zeigen. Wegen $S = B^{-1}R \geq 0$ gilt

$$0 \leq S = B^{-1}R = [A^{-1}B]^{-1}A^{-1}R = [A^{-1}(A + R)]^{-1}A^{-1}R = (E + C)^{-1}C.$$

Zu $\lambda = \rho(S) \in \sigma(S)$ gehört wegen $S \geq 0$ nach dem Satz von Perron-Frobenius ein EV $x \not\geq 0$. Aus

$$\lambda x = Sx = (E + C)^{-1}Cx \quad \text{folgt} \quad \lambda x + \lambda Cx = Cx .$$

Damit kann der Wert $\lambda = 1$ nicht auftreten (sonst müsste $x = 0$ sein) und es ist

$$Cx = \frac{\lambda}{1 - \lambda}x . \quad (\text{A.12})$$

Wegen $A^{-1} > 0$ und $R \geq 0$ ist auch $C \geq 0$. Damit folgt aus $x \not\geq 0$ und $Cx \geq 0$ die Ungleichung $\frac{\lambda}{1 - \lambda} \geq 0$, was

$$0 \leq \lambda = \rho(S) < 1$$

impliziert. Aus (A.12) folgt, dass λ' genau dann ein EW von S ist, wenn $\mu' = \frac{\lambda'}{1 - \lambda'}$ ein EW von C ist. Aus $|\lambda'| \leq \lambda = \rho(S)$ schließt man auf

$$|\mu'| = \left| \frac{\lambda'}{1 - \lambda'} \right| \leq \frac{|\lambda'|}{1 - |\lambda'|} \leq \frac{\lambda}{1 - \lambda} =: \mu ,$$

d.h. $|\mu|$ ist maximal für $\lambda' = \lambda = \rho(S) \in \sigma(S)$. Nach dem Satz von Perron-Frobenius ist $\mu = \rho(C) \in \sigma(C)$ der maximale EW von C , also gilt

$$\rho(C) = \frac{\rho(S)}{1 - \rho(S)} \quad \Longleftrightarrow \quad \rho(S) = \frac{\rho(C)}{1 + \rho(C)} .$$

□

Nun kann man folgenden Vergleichssatz formulieren, der den Vergleich von Jacobi- und Gauß-Seidel-Verfahren ermöglicht.

Satz A.15. *Gelte $A^{-1} \geq 0$. Durch B_1 und B_2 seien zwei reguläre Aufspaltungen von A gegeben. Wenn B_1 und B_2 in der Form*

$$A \leq B_1 \leq B_2 \quad (\text{A.13})$$

vergleichbar sind, dann gilt

$$0 \leq \rho(S_1) \leq \rho(S_2) < 1, \quad \text{wobei} \quad S_i := B_i^{-1}R_i, \quad R_i := W_i - A . \quad (\text{A.14})$$

Gilt $A^{-1} > 0$ und für die regulären Aufspaltungen

$$A \not\leq B_1 \not\leq B_2 , \quad (\text{A.15})$$

dann folgt

$$0 < \rho(S_1) < \rho(S_2) < 1 . \quad (\text{A.16})$$

Beweis. Es soll nur die Aussage (A.14) gezeigt werden (zum Beweis dieses Satzes und des Satzes A.14 sei im Übrigen auf die "Bibel" zur iterativen Lösung großer Gleichungssysteme von W. Hackbusch (im Netz verfügbar) verwiesen).

$C_1 = A^{-1}R_1$ und $C_2 = A^{-1}R_2$ erfüllen $0 \leq C_1 \leq C_2$ und damit $0 \leq \rho(C_1) \leq \rho(C_2)$ (s. (A.7)). Aus der Darstellung (A.11) erhält man

$$0 \leq \rho(S_1) = \frac{\rho(C_1)}{1 + \rho(C_1)} \leq \frac{\rho(C_2)}{1 + \rho(C_2)} = \rho(S_2) < 1 .$$

□

Bemerkung A.16. Da für die irreduzible diagonal dominante oder strik diagonal dominante L -Matrizen mit $A = D + L + U$ die Zerlegungen

$$B_1 = D + L \quad \text{und} \quad B_2 = D$$

regulär sind, $A^{-1} > 0$ gilt, und

$$A \not\leq B_1 \not\leq B_2$$

gilt, ist die Konvergenzrate des Gauß-Seidel-Verfahrens größer als die des Jacobi-Verfahrens.

Zum Schluss dieses Abschnittes soll noch das sogenannte M -Kriterium zur Abschätzung der Norm der Inversen einer Matrix A , die man bei Stabilitätsabschätzungen von Diskretisierungen gebrauchen kann, notiert werden.

Satz A.17. (*M -Kriterium*)

Sei A eine M -Matrix und $e > 0$ ein Vektor mit $Ae > 0$. Dann gilt

$$\|A^{-1}y\|_\infty \leq C_e \|y\|_\infty \quad \text{mit} \quad C_e := \frac{\max_i e_i}{\min_i (Ae)_i} ,$$

d.h.

$$\|A^{-1}\|_\infty \leq C_e .$$

Beweis.

Sei $x = A^{-1}y$, also $Ax = y$. Dann ist

$$\pm x_i = \sum_{j=1}^n (A^{-1})_{ij} (\pm y_j) \leq \sum_{j=1}^n (A^{-1})_{ij} \|y\|_\infty .$$

Sei $c = \min_i (Ae)_i$, d.h. $Ae \geq c(1, 1, \dots, 1)^T$. Da A invers monoton ist (aus $Ax \leq 0$ folgt $x \leq 0$), gilt

$$e \geq cA^{-1}(1, 1, \dots, 1)^T , \quad \text{d.h.} \quad e_i \geq c \sum_{j=1}^n (A^{-1})_{ij} .$$

Daraus folgt nun $\pm x_i \leq \frac{e_i}{c} \|y\|_\infty$ und damit

$$\|x\|_\infty \leq \frac{\|e\|_\infty}{c} \|y\|_\infty = C_e \|y\|_\infty ,$$

also die Aussage des Satzes. □

Bemerkung A.18.

Die Matrizen vom Typ (A.1) sind L -Matrizen und irreduzibel diagonal dominant, d.h. sie sind auch M -Matrizen. Damit sind sie invertierbar (regulär) und man kann in der Regel eine Schranke für die Maximumnorm der inversen Matrix A^{-1} angeben, was Stabilität bedeutet.

Allerdings muss man dazu einen Vektor e mit der geforderten Eigenschaft finden. Bei strikt diagonal dominanten L -Matrizen findet man mit

$$e = (1, 1, \dots, 1)^T$$

diesen Vektor sofort.

Beim Beispiel des Poissonschen Randwertproblems auf einem n -dimensionalen Einheitswürfel könnte man die Funktionswerte von

$$v(x) := \frac{1}{2} x_1 (1 - x_1) \quad \text{für } x = (x_1, x_2, \dots, x_n) \in \Omega_h$$

als Komponenten des Vektors $e > 0$ wählen, denn es gilt

$$-L_h v(x) = 1 \quad \text{bzw.} \quad Ae = (1, 1, \dots, 1)^T > 0 .$$

Mit $\|e\|_\infty = \frac{1}{8}$ erhält man dann

$$\|A^{-1}\|_\infty \leq \frac{1}{8} .$$

A.2 Schachbrett-, Zebra- und andere Variablennummerierungen

Die Matrizen (A.1) entstehen bei Diskretisierungen, bei denen man die Unbekannten lexikographisch ordnet. Bei Finite-Differenzen- oder Finite-Volumen-Methoden approximiert man z.B. den Laplace-Operator (2d-Fall) an einem Punkt (x, y) durch Funktionswerte auf einem regelmäßigen Gitter durch den sogenannten 5-Punkte-Stern (s. Abb. A.1)

$$-\Delta u|_{(x,y)} \approx \frac{4u(x, y) - u(x + h, y) - u(x - h, y) - u(x, y + h) - u(x, y - h)}{h^2} , \tag{A.17}$$

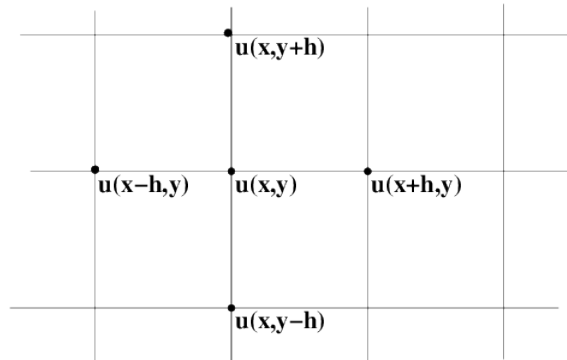


Abbildung A.1: Differenzenstern zur Approximation von Δu

wobei $h > 0$ ein Diskretisierungsparameter ist. Durch Taylorentwicklung zeigt man bei ausreichender Glattheit von u , dass die Approximation (A.17) von der Ordnung $\mathcal{O}(h^2)$ ist. Die Abbildung A.2 zeigt die lexikographische Nummerierung der Unbekannten auf einem Rechteckgitter, die bei der numerischen Lösung eines elliptischen Randwertproblems auf ein lineares Gleichungssystem mit einer Koeffizientenmatrix der Form (A.1) führt. Die Diffe-

17	18	19	20	
13	14	15	16	
9	10	11	12	
5	6	7	8	
1	2	3	4	

Abbildung A.2: lexikographische Nummerierung der Unbekannten

renzengleichung mit der Unbekannten $u_{10} \approx u(2h, 3h)$ ergibt sich z.B. zu

$$\frac{4u_{10} - u_{11} - u_9 - u_{14} - u_6}{h^2} = r_{10} . \quad (\text{A.18})$$

Die Koeffizientenmatrix ergibt sich zu

$$\mathcal{A}_l = \begin{pmatrix} A & -T & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -T & A & -T & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -T & A & -T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -T & A & -T \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -T & A \end{pmatrix} , \quad (\text{A.19})$$

wobei für die Blöcke

$$A = \frac{1}{h^2} \begin{pmatrix} 4 & -1 & 0 & 0 \\ -1 & 4 & -1 & 0 \\ 0 & -1 & 4 & -1 \\ 0 & 0 & -1 & 4 \end{pmatrix} , \quad T = \frac{1}{h^2} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (\text{A.20})$$

gilt. Nummeriert man die Unbekannten schachbrettartig (chequer-board ordering) gemäß Abbildung A.3, dann erhält man als Koeffizientenmatrix

9	19	10	20	
17	7	18	8	
5	15	6	16	
13	3	14	4	
1	11	2	12	

Abbildung A.3: schachbrettartige Nummerierung der Unbekannten

$$\mathcal{A}_c = \begin{pmatrix} D & B \\ B^T & D \end{pmatrix} , \quad (\text{A.21})$$

mit $D = \frac{4}{h^2}E$ und E als Einheitsmatrix aus $\mathbb{R}^{10 \times 10}$, und der Matrix

$$B = -\frac{1}{h^2} \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

Die Gleichung für die Unbekannte $u_{15} \approx u(2h, 3h)$ (entspricht (A.18) bei lexikographischer Nummerierung) ist dann

$$\frac{4u_{15} - u_6 - u_5 - u_7 - u_3}{h^2} = r_{15}.$$

Der Matrixblock B muss nicht notwendig quadratisch sein, das hängt davon ab, ob man in seinem "Schachbrett" die gleiche Anzahl schwarzer und weißer Felder hat (was im Beispiel mit 10 und 10 gegeben ist).

Die unterschiedlichen Nummerierungen ändern im Falle der numerischen Lösung eines elliptischen oder parabolischen Randwert- bzw. Randanfangswertproblems nichts an den wesentlichen Eigenschaften wie Symmetrie, oder Diagonal-Dominanz. Unterschiedliche Nummerierungen erweisen sich aber als sinnvoll mit Blick auf Parallelisierungen der Algorithmen. Außerdem kann man zeigen, dass das Gauß-Seidel-Verfahren im Falle der schachbrettartigen Nummerierung doppelt so schnell wie das Jacobi-Verfahren ist, d.h. für die entsprechenden Iterationsmatrizen $S_{GS_{cb}}$ und S_{jac} gilt

$$\rho(S_{GS_{cb}}) = \rho(S_{jac})^2.$$

Eine sehr ausführliche Diskussion zu dieser Thematik findet bei Hackbusch "Iterative Lösung großer Gleichungssysteme" statt.