Numerik 0 – Peter Bastian

Stefan Breunig

13. April 2010

Inhaltsverzeichnis

Lizenz und Copyright . . . 4
LaTeX-Kurs . . . 4
Bierware . . . 4
Verbesserungen . . . 4

1 Grundbegriffe der Numerik **5**
1.1 Stabilität und Kondition . . . 5
1.2 Numerische Auswertung von Funktionsvorschriften . . . 7
1.3 Lösung von Gleichungen . . . 8
Wiederholung: . . . 11
Numerisches Verfahren: . . . 11
„Lösungsoperator“ . . . 11
„Konsistenz“ . . . 11
„Stabilität“ . . . 12

2 Fließkommazahlen **12**
2.1 Zahlendarstellung . . . 12
normierte Fließkommazahlen $\mathbb{F}(\beta, r, s)$ . . . 13
Zahlenbereich: . . . 13
Praxis: IEEE754/IEC559 Standard . . . 14
Double genauer: . . . 14
2.2 Runden und Rundungsfehler . . . 14
natürliche Rundung: . . . 15
Gerade Rundung (β gerade): . . . 15
2.3 Fließkommarithmetik . . . 16
Beispiel: Subtraktion . . . 16
2.4 Fehleranalyse . . . 17
Wichtig: . . . 17
(1) Differentielle Konditionsanalyse . . . 17
(2) Rundungsfehleranalyse . . . 19
2.5 Auslöschung . . . 20

2.6 Die quadratische Gleichung . . . 20
Numerisch stabile Auswertung für den Fall $\frac{p^2}{4} \gg q$ . . . 21

3 Motivation linearer Gleichungssysteme 21
3.1 Strömungen in Rohrleitungsnetzen . . . 21
3.2 Radiositymethode in der Computergraphik . . . 24
Ray-Tracing . . . 24
Radiosity . . . 24
Numerische Lösung mit „Kollokationsmethode“ . . . 25

4 Konditionierung der Lösung linearer Gleichungssysteme 25
4.1 Lösbarkeit . . . 25
4.2 Vektorraum . . . 26
Folge: . . . 28
4.3 Matrixnormen . . . 28
4.4 Eigenwerte und Eigenvektoren . . . 30
4.5 Die Spektralnorm . . . 31
4.6 Positiv definite Matrizen . . . 32
4.7 Störungstheorie . . . 33

5 Eliminationsverfahren zur Lösung linearer Gleichungsysteme 36
5.1 Dreieckssysteme (gestaffelte Gleichungssysteme) . . . 36
5.2 Gauß-Elimination . . . 36
Ziel: . . . 36
Schritt 1: . . . 37
Schritt k: . . . 37
5.3 LR-Zerlegung . . . 39
Wo ist L? . . . 42
Wozu das Ganze? . . . 42
Vorteil: . . . 43
zu den Permuationsmatrizen: . . . 43
Algorithmus zur LR-Zerlegung . . . 43
5.4 Rundungsfehleranalyse der LR-Zerlegung (GEM) . . . 44
Absolutwertnotation . . . 44
fl-Notation (gl-Notation) . . . 44
Rückwärtsanalyse . . . 44
Folgerung . . . 48
5.5 Pivotisierung . . . 48
ABER: . . . 49
Rundungsfehler bei Pivotisierung . . . 49
Totale Pivotisierung . . . 50
Aufwand der Pivotisierung . . . 50
5.6 Spezielle Systeme . . . 50
Symmetrisch positiv definite Matrizen . . . 50

6 Interpolation und Approximation **54**
6.1 Einführung . . . 54
Grundaufgabe der Approximation . . . 54
6.2 Polynominterpolation . . . 55
Lagrange-Interpolation . . . 55
Satz von Rolle: . . . 56
Newton-Polynome . . . 57
Praktische Anwendung . . . 58
Interpolationsfehler . . . 58
Satz von Rolle . . . 59
Diskussion . . . 59
Runges Gegenbeispiel: . . . 59
Konditionierung der Interpolationsaufgabe . . . 60
6.3 Anwendungen der Polynominterpolation . . . 60
Numerische Differentation . . . 60
Bis jetzt: . . . 61
Alternativ: . . . 61
Beispiel: . . . 61
Extrapolatin zum Limes . . . 61
Idee: . . . 62
Warum ist das so gut? . . . 63
Beispiel: . . . 64
6.4 Bernsteinpolynome zur Kurvendarstellung . . . 64
6.5 Splines . . . 67
Idee: . . . 67
Kubische Splines . . . 67
Bedingungen . . . 68
Siehe Folien für Erklärung zu: . . . 71
6.6 Trigonometrische Interpolation . . . 71
Problem: . . . 71
Schnelle Fourier-Transformation . . . 74
Beispiel: N=8 . . . 75
6.7 Approximation von Funktionen . . . 76
Beispiele . . . 76
Approximation mit Orthonormalbasen . . . 79
Fehlerkontrolle . . . 79
Adaptive Approximation mit Haar-Wavelets . . . 81
Orthogonalitätseigenschaft . . . 82
Veranschaulichung . . . 83
effiziente Auswertung . . . 83

7 Numerische Integration **84**
7.1 Newton-Cotes Formeln . . . 84
Berechnung der Gewichte . . . 85

7.2 Summierte Quadratformeln . . . 87
Idee: . . . 88
Beispiele: . . . 89
7.3 Quadraturen höherer Ordnung . . . 90
Romberg-Integration . . . 90
Gauß-Integration . . . 90
7.4 Ausblick . . . 91
Adaptive Quadratur . . . 91
Mehrdimensionale Quadratur . . . 91
Fluch der Dimension . . . 92

8 Iterative Lösung von Gleichungsystemen 92
Intervallschachtelung . . . 92
8.1 Newton Verfahren . . . 93
Bemerkungen zum Newton-Verfahren . . . 96
8.2 Sukzessive Approximation . . . 97
8.3 Iterationsverfahren zur Lösung linearer Gleichungsysteme . . . 99
Beispiele für Iterationsverfahren . . . 100

Lizenz und Copyright

LaTeX-Kurs

An dieser Stelle sei auch auf Arnos LaTeX-Kurs verwiesen, der nächstes Semester wieder angeboten wird, wobei gilt: (Großartigkeit des LaTeX-Kurses $\gg \infty$). Der Beweis wird dem interessierten Hörer überlassen. Sagte ich, dass es ECTS-Punkte gibt? Alte Seite:
`http://www.mpi-hd.mpg.de/blaum/teaching/2009/ss09_latex/index.de.html`

Bierware

Der *Mitschrieb* ist frei und kostenlos, dennoch würde ich mich über ein Bierchen freuen. Das steigert auch ungemein meine Motivation in Numerik 1 weiterzuTeXen ☺. Ansonsten trinkt doch wenigstens ein Bier auf mein Wohl.

Verbesserungen

Dieser *Mitschrieb* ist sicher nicht fehlerfrei und kann nicht als Skriptersatz dienen. Trotzdem freue ich mich über Hinweise, idealerweise als Patch:

stefan@mathphys.fsk.uni-heidelberg.de

1 Grundbegriffe der Numerik

1.1 Stabilität und Kondition

Auswerten von Funktionen:
Gegeben x, berechne $y = F(x)$
$F\colon X \to Y$; X, Y normierte Räume
$x \overset{F(x)}{\to} y$
$x' \overset{F(x')}{\to} y'$
Mathematisch: Stetigkeit!

Definition 1.1 Stabilität
Wir nennen die Auswertung von der Funktion $F\colon X \to Y$

(a) stabil, wenn F stetig ist, d.h.:
$\forall \varepsilon > 0\, \exists\, \delta = \delta_{(\varepsilon)} : \forall x, x' \in X : \|x - x'\|_x \leq \delta \Rightarrow \|F(x) - F(x')\|_y \leq \varepsilon$

(b) lokal L-stabil, wenn F lokal Lipschitz-stetig ist:
$\forall x \in X\, \exists\, \delta(x) > 0\, \exists\, \kappa_0(x) > 0 :$
$\forall x' : \|x - x'\|_x \leq \delta \Rightarrow \|F(x) - F(x')\|_y \leq \kappa_0 \|x - x'\|_x$

(L-stabil $\Rightarrow$ stabil, aber nicht umgekehrt)

Quantifiziere die Abhängigkeit:

$$\frac{\|\delta_y\|}{\|\delta_x\|} = \frac{\|F(\overbrace{x - \delta x}^{x'=}) - F(x)\|}{\|\delta_x\|} \leq \kappa_0(x)$$

Definition 1.2 Konditionszahlen
Die *absolute* Kondition einer Abbildung $F\colon X \to Y$ ist

$$\kappa_{abs}(x) = \sup\left\{ \frac{\|F(x + \delta_x) - F(x)\|}{\|\delta_x\|} \mid \delta_x \neq 0, x + \delta X \in X \right\}$$

entsprechend die *relative* Kondition:

$$\kappa_{rel}(x) = \sup\left\{ \frac{\|F(x + \delta_x) - F(x)\| / \|F(x)\|}{\|\delta_x\| / \|x\|} \mid \delta_x \neq 0, x + \delta x \in X \right\}$$

$(F(x) \neq 0, x \neq 0)$

$F\colon x \to y$

- „gut konditioniert“, wenn $\kappa(x)$ „klein“
- „schlecht konditioniert“ sonst

Erwartung:

- gut konditionierte Vorschriften erlauben eine Berechnung auf dem Computer!

- bei schlecht konditionierten Vorschriften sind Probleme zu erwarten, und zwar **unabhängig** von numerische Verfahren

Beispiel 1.3 (Kondition der Addition)

$$\begin{aligned} y &= F(x_1, x_2) = x_1 + x_2 \\ x &= (x_1, x_2)^T \\ \delta_x &= (\delta_{x_1}, \delta_{x_2})^T \end{aligned}$$

$$\begin{aligned} F(x+\delta_x) &= F(x) + \Delta F(x) \cdot \delta_x + \mathcal{O}(\|\delta_x\|^2) \\ \Rightarrow \quad |F(x+\delta x) - F(x)| &\leq |\Delta F(x) \cdot \delta_x| + \mathcal{O}(\|\delta_x\|^2) \leq \|\Delta F(X)\| \|\delta_x\| + \mathcal{O}(\|\delta_x\|^2) \\ \Leftrightarrow \quad \tfrac{|F(x+\delta_x)-F(x)|}{\|\delta_x\|} &\leq \|\Delta F(x)\| + \mathcal{O}(\|\delta_x\|) \end{aligned}$$

$$\Delta F(x) = (1,1)^T \Rightarrow \|\Delta F(x)\| = \sqrt{2}$$

also: $\kappa_{abs}(x) \doteq \sqrt{2}$
relative Kondition: $\kappa(x) \doteq \sqrt{2}\frac{\|x\|}{|x_1+x_2|}$
$x_1 \approx -x_2, \qquad |x_1 + x_2| \ll \|x\| = \sqrt{x_1^2 + x_2^2}$ Kondition groß!
$\Rightarrow$ Addition ist schlecht konditioniert für $x_1 \approx -x_2$!

Beispiel 1.4 Lösung der Quadratischen Gleichung
Betrachte Lösung von

$$x^2 - 2px + 1 = 0$$

$$\Rightarrow x_{1,2} = p \pm \sqrt{p^2 - 1} \qquad (p \in [1, \infty))$$

$$F\colon \mathbb{R} \to \mathbb{R}^2 \qquad F(p)\begin{pmatrix} p + \sqrt{p^2-1} \\ p - \sqrt{p^2-1} \end{pmatrix}$$

Taylorreihe:

$$F(p + \delta_p) = F(p) + \frac{dF}{dp}(p)\delta_p + \mathcal{O}(|\delta_p|^2)$$

$$\frac{\|F(p-\delta_p) - F(p)\|}{|\delta_p|} \leq \left\|\frac{dF}{dp}(p)\right\| + \mathcal{O}(|\delta_p|)$$

$$\frac{dF}{dp}(p) = \left(1 + \frac{p}{\sqrt{p^2-1}}, 1 - \frac{p}{\sqrt{p^2-1}}\right)$$

$\|\ \ \|$ euklidische Norm

$$\kappa_{abs}(p) \doteq \sqrt{2}\sqrt{\frac{2p^2-1}{p^2-1}}$$

schlecht konditioniert für $p \to 1$

$$\kappa(p) \doteq \sqrt{2}\sqrt{\frac{2p^2-1}{p^2-1}} \cdot \frac{|p|}{\left(p + \sqrt{p^2+1}\right)^2 + \left(p - \sqrt{p^2-1}\right)^2}$$

$\to$ schlecht konditioniert für $p \to 1$!

1.2 Numerische Auswertung von Funktionsvorschriften

$y = F(x)$ kann in der Regel nicht *exakt* realisiert werden. Stattdessen realisieren wir für $k \in \mathbb{N}$:

$$y^{(k)} = F^{(k)}\left(x^{(k)}\right) \text{ mit } F^{(k)} : x^{(k)} \to y^{(k)} \tag{1.1}$$

$x^{(k)}, y^{(k)}$ normierte Räume

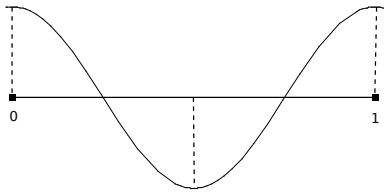
Beispiel 1.5

(a) $F\colon \mathbb{R}^2 \to \mathbb{R}, \qquad F(x_1, x_2) = x_1 + x_2$
$F^{(k)}\colon (\mathbb{F}(10, k, 2))^2 \to \mathbb{F}(10, k, 2), \quad F^{(k)}(x_1^{(k)}, x_2^{(k)} = x_1^{(k)} \oplus x_2^{(k)}$
$0, m_1 m_2 \cdots m_k \cdot 10^{\pm e}$ („Kommazahl“)

(b) $F\colon \mathbb{R} \to \mathbb{R} \qquad F(x) = \exp(x)$
$F^{(k)}\colon \mathbb{R} \to \mathbb{R} \qquad F^{(k)} = 1 + \sum_{i=1}^{k} \frac{x^i}{i!}$

In beiden Beispielen gilt: $X^{(k)} \subset X, Y^{(k)} \subset Y$
Wir setzen voraus: $X^{(k)} \subseteq X$
aber nicht unbedingt: $Y^{(k)} \subseteq Y$!
Es gebe aber eine Abbildung $R^{(k)}\colon Y \to Y^{(k)}$ mit $R^{(k)}$ linear und L-stabil!

Beispiel 1.6
$Y = C^0([0, 1])$
$Y^{(k)} = R^{(k+1)}$
$(R^{(k)}y)_i = y\left(\frac{i}{k}\right) \qquad 0 \le i \le k$



Definition 1.7 (Konvergenz)

Die numerische Auswertung (1.1) heißt konvergent, genau dann wenn für jedes $x \in X$ gilt:
$\forall \varepsilon > 0\, \exists\, k_0(\varepsilon) \in \mathbb{N}\, \exists\, \delta(\varepsilon, k_0) > 0$:
$\quad \forall k > k_0, \forall x' \in X^{(k)} \quad : \quad \|x - x'\| \le \delta \Rightarrow \|R^{(k)}F(x) - F^{(k)}(x')\| \le \varepsilon$

Aus der Definition 1.7 folgt auch die Konvergenz von

$$\|R^{(k)}F(x') - F^{(k)}(x')\| \to 0 \text{ für } k \to \infty,\ x' \in X^{(k)}$$

denn:

$$\begin{aligned}
&\|R^{(k)}F(x') - F^{(k)}(x')\| \\
&\qquad = \|R^{(k)}F(x') - R^{(k)}F(x) + R^{(k)}F(x) - F^{(k)}(x')\| \\
&\qquad \le \underbrace{\|R^{(k)}F(x') - R^{(k)}F(x)\|}_{\substack{R^{(k)}(F(x')-F(x))\to 0,\\ \text{wenn } R^{(k)}, F \text{ stabil}}} + \underbrace{\|R^{(k)}F(x) - F^{(k)}(x')\|}_{\to 0}
\end{aligned}$$

$$\begin{aligned}
&\|R^{(k)}F(x) - F^{(k)}(x')\| \\
&\qquad = \underbrace{\|R^{(k)}F(x) - R^{(k)}F(x')\|} + \underbrace{\|R^{(k)}F(x') - F^{(k)}(x')\|}_{\to 0}
\end{aligned}$$

Wiederholung 1
Funktionsauswertung:
exakt: $F\colon X \to$ Stabilität, Konditionszahl
im Rechner: $F^{(k)}\colon X^{(k)} \to Y^{(k)} \qquad k \in \mathbb{N}$
$X^{(k)} \subset X, \qquad Y^{(k)} \not\subset X$

Definition 1.7
$\forall \varepsilon > 0\, \exists\, k_0(\varepsilon)\, \exists\, \delta(\varepsilon, k_0)$:

(a) $\forall k > k_0, \forall x' \in X^{(k)} : \|x - x'\| \le \delta \Rightarrow \|R^{(k)}F(x) - F^{(k)}(x')\| \le \varepsilon$

(b) $\forall k > k_0, \forall x' \in X^{(k)} : \|R^{(k)}F(x')\| \le \varepsilon$

(a) $\Leftrightarrow$ (b) falls F stabil. (b) $\Rightarrow$ (a)

$$\begin{aligned}\|R^{(k)}F(x) - F^{(k)}(x')\| \\ \le\ \underbrace{\|R^{(k)}F(x) - R^{(k)}F(x')\|}_{\text{„Konditionanalyse“}} + \underbrace{\|R^{(k)}F(x') - F^{(k)}(x')\|}_{\text{„Rundungsfehleranalyse“}}\end{aligned}$$

Bemerkung 1.8

- $\|R^{(k)}F(x) - R^{(k)}F(x')\|$ behandelt die Stabilität von $R^{(k)}F$
- $\|R^{(k)}F(x') - F^{(k)}(x')\|$ behandelt den Rungsungs- oder Abbruchfehler

Was ist mit der Stabilität der $F^{(k)}$?
Wir zeigen: Konvergenz der $F^{(k)} \Rightarrow$ Stabilität der $F^{(k)}$
also damit: „Γ Stabilität $\Rightarrow$ Γ Konvergenz“
d.h. Stabilität der $F^{(k)}$ ist *notwendige* Voraussetzung für Konvergenz!

Satz 1.9
Es sei F stabil und $F^{(k)}$ konvergent (nach 1.7). Dann sind die $F^{(k)}$ stabil.

Beweis .
Sei $x, x' \in X^{(k)}$

$$\begin{aligned}\|F^{(k)}(x) - F^{(k)}(x')\| &= \|F^{(k)}(x) - R^{(k)}F(x) + R^{(k)}F(x) - F^{(k)}(x')\| \\ \|F^{(k)}(x) - F^{(k)}(x')\| &\le \underbrace{\|F^{(k)}(x) - R^{(k)}F(x)\|}_{\to 0\ ((b)\ \text{oben})} + \underbrace{\|R^{(k)}F(x) - F^{(k)}(x')\|}_{\to 0\ ((a)\ \text{oben})}\end{aligned}$$

□

1.3 Lösung von Gleichungen

Oft ist eine unbekannte Größe Y implizit durch eine Gleichung bestimmt:
Gegeben $x \in X$ („Daten“), finde $y \in Y$ („Lösung“) sodass

$$G(y, x) = 0 \tag{1.2}$$

erfüllt ist.

Definition 1.6
Das Problem 1.2 heißt *sachgemäß gestellt* falls:

(a) zu jedem $x \in X$ existiert genau eine Lösung $y \in Y$

(b) Die Lösung y hängt stetig von den Daten x ab.

Beispiel 1.7
$x \in \mathbb{R}^n$; $\quad A \in \mathbb{R}^n$ invertierbar

$$G(y,x) = Ay - x = 0 \qquad (\rightsquigarrow y = A^{-1}x)$$

Die Funktion $F\colon X \to Y$ mit

$$\forall x \in X : G(F(x), x) = 0$$

heißt „Lösungsoperator". Sachgemäß gestellt $\Leftrightarrow \exists! F$ und F stabil!

im Rechner: Ersetze (1.2) durch genähertes System zu genäherten Daten. Formal: Zu $k \in \mathbb{N}$ betrachte

$$\text{Gegeben } x^{(k)} \in X^{(k)}, \text{ finde } y^{(k)} \in Y^{(k)} \text{ sodass } G^{(k)}(y^{(k)}, x^{(k)}) = 0 \tag{1.3}$$

Zu jedem $G^{(k)}$ gibt es den Lösungsoperator $F^{(k)}\colon X^{(k)} \to Y^{(k)}$:

$$\forall x \in X^{(k)} \qquad G^{(k)}(F^{(k)}(x^{(k)}), x^{(k)}) = 0$$

Wie oben gelte $X^{(k)} \subseteq X$, $\quad Y^{(k)} \subseteq Y$ *nicht* notwendigerweise.
Aber es gebe: $R^{(k)}\colon Y \to Y^{(k)} \qquad (R^{(k)}$ stabil, linear).
Ziel: eine einfache „Konvergenztheorie"

Definition 1.8 (Konsistenz)
Das numerische Verfahren (1.3) heißt konsistent zu (1.2), falls

$$G^{(k)}(R^{(k)}y, \underbrace{x + \delta x^{(k)}}_{\in X^{(k)}}) \to 0 \qquad \text{für } k \to \infty \text{ und } \delta x^{(k)} \to 0$$

wobei $G(y,x) = 0$

„Inversion" von $G^{(k)}\colon Y^{(k)} \times X^{(k)} \to Z^{(k)}$
Es existiere Abbildung $H^{(k)}\colon Z^{(k)} \times X^{(k)} \to Y^{(k)}$ sodass:

$$\forall x \in X^{(k)}, \forall y \in Y^{(k)} : \qquad H^{(k)}(\underbrace{G^{(k)}(y,x)}_{z}, x) = y \tag{1.4}$$

Definition 1.9
Das numerische Verfahren (1.3) heißt L-stabil, falls $H^{(k)}$ L-stabil bezüglich dem ersten Argument (z) ist

Satz 1.10
Das numerische Verfahren sei konsistent und L-stabil.
Dann ist es auch konvergent.

Beweis .
Zu $x \in X$ sei $G(y,x) = 0, \quad G^{(k)}(y^{(k)}, x + \delta x^{(k)}) = 0 \qquad \|\delta x^{(k)}\| \to 0$ für $k \to \infty$

$$\begin{aligned}
& \|R^{(k)}y - y^{(k)}\| \\
= \; & \| \underbrace{H^{(k)}\Big(G^{(k)}(R^{(k)}y, x + \delta x^{(k)}), x + \delta x^{(k)}\Big)}_{R^{(k)}y} - \underbrace{H^{(k)}\Big(G^{(k)}(y^{(k)}, x + \delta x^{(k)}), x + \delta x^{(k)}\Big)}_{y^{(k)}} \| \\
\leq \; & \kappa_0 \|G^{(k)}(R^{(k)}y, x + \delta x^{(k)}) - G^{(k)}(y^{(k)}, x + \delta x^{(k)})\| \\
\leq \; & \kappa_0 \|G^{(k)}(R^{(k)}y, X + \delta x^{(k)})\| \\
\to \; & 0 \text{ wegen Konsistenz}
\end{aligned}$$

□

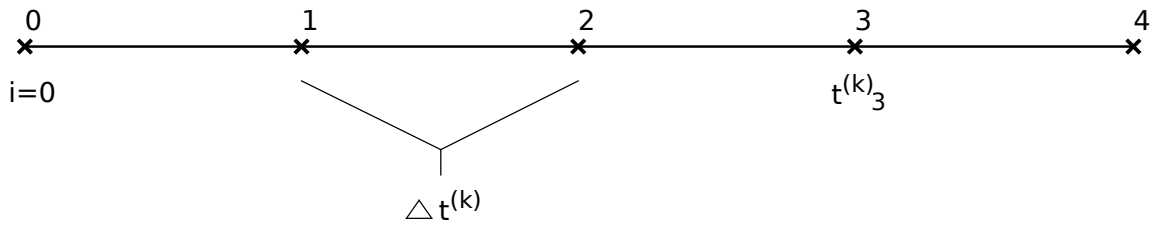
Beispiel 1.11 (Anfangswertproblem)
gewöhnliche Differentialgleichung:

$$\begin{aligned}
y'(t) &= f(t, y(t)) \text{ für } t \in [0, T] \\
y(0) &= x \qquad \text{(Anfangswert)}
\end{aligned}$$

Hier: $X = \mathbb{R}$ (Daten), $Y = C^1(I)$

$$G \colon Y \times X \to Z : \left(G(y,x) = |y(0) - x| + \sup_{t \in I} |y'(t) - f(t, y(t))| \right)$$

Problem sei sachgemäß gestellt.



Numerik: Euler-Verfahren. Sei $k \in \mathbb{N}$. Setze

$$\Delta t^{(k)} = \frac{T}{k}, \qquad T_i^{(k)} i \cdot \Delta t^{(k)} \qquad 0 \leq i \leq k$$

Taylorentwicklung:

$$\begin{aligned}
y(t_i^{(k)}) &= y(t_{i-1}^{(k)} + \Delta t^{(k)}) \\
&= y(t_i^{(k)}) + \Delta t y'(t_{i-1}^{(k)} + \frac{(\Delta t^{(k)})^2}{2} y''(\xi_i^{(k)}), \xi_i^{(k)} \in [t_{i-1}^{(k)}, t_i^{(k)}] \\
&= y(t_{i-1}^{(k)} + \Delta t^{(k)} f(t_{i-1}^{(k)}, y(t_{i-1}^{(k)})) + \frac{(\Delta t^{(k)})^2}{2} [\partial_t f + (\partial_y f) f](\xi_i^{(k)}, y(\xi_i^{(k)}))
\end{aligned}$$

umstellen.

$$\frac{1}{\Delta t^{(k)}} (y(t_i^{(k)}) - y(t_{i-1}^{(k)})) - f(t_{i-1}^{(k)}, y(t_{i-1}^{(k)})) = \mathcal{O}(\Delta t^{(k)}) \tag{1.5}$$

$\rightsquigarrow y_i^{(k)} \approx y(t_i^{(k)})$ werden bestimmt durch Ersetzen der rechten Seite durch $\mathcal{O}$ in (1.5)

Wiederholung:

$$y'(t) = f(t, y(t)) \qquad t \in I = [0, T] \qquad y(0) = x$$

Mit $Y = C^1(I)$, $\qquad Z = \mathbb{R}$

$$G(y, x) = |y(0) - x| + \sup_{t \in I} |y'(t) - f(t, y(t))|$$

„Gitter“ $\qquad k \in \mathbb{N}$

$$\Delta t^{(k)} = \frac{T}{k} \qquad t_i^{(k)} = i \Delta t^{(k)} \qquad 0 \le i \le k$$

Taylorreihe:

$$\frac{1}{\Delta t^{(k)}} \left(y(t_i^{(k)}) - y(t_i^{(k)}) \right) - f\left(t_{i-1}^{(k)}, y(t_{i-1}^{(k)})\right) = \mathcal{O}(\Delta t^{(k)}) \to 0 \text{ für } k \to \infty$$

Numerisches Verfahren:

Definiere Zahlen $y_i^{(k)} \in \mathbb{R} \qquad 0 \le i \le k$ mit $y_0^{(k)} = x$

$$(**) \qquad \frac{1}{\Delta t^{(k)}} \left(y_i^{(k)} - y_{i-1}^{(k)} \right) - f\left(t_{i-1}^{(k)}, y_{i-1}^{(k)}\right) = 0 \qquad 0 < i \le k$$

Mit

$$Y^{(k)} = \mathbb{R}^{k+1} \qquad y^{(k)} = (y_0^{(k)}, \cdots, y_k^{(k)})^T$$
$$X^{(k)} = \mathbb{R} \qquad Z^{(k)} = \mathbb{R}^{k+1}$$

$$G^{(k)}(y^{(k)}, x) = \begin{cases} y_i^{(k)} - x & i = 0 \\ \frac{1}{\Delta t^{(k)}}(y_i^{(k)} - i_{i-1}^{(k)}) - f(t_{i-1}^{(k)}, y_{i-1}^{(k)} & \text{sonst} \end{cases}$$

$$G^{(k)}(y^{(k)}, x) = 0 \qquad \text{Numerische Lösung der AWA}$$

„Lösungsoperator“

$$F^{(k)}: \quad \begin{array}{ll} X^{(k)} & \to Y^{(k)} \\ \mathbb{R} & \to \mathbb{R}^{k+1} \end{array}$$

$$y_i^{(k)} = F_i^{(k)}(x) = \begin{cases} x & i = 0 \\ F_{i-1}^{(k)}(x) + \Delta t^{(k)} f(t_{i-1}^{(k)}, F_{i-1}^{(k)}(x)) & i > 0 \end{cases}$$

„Konsistenz“

$$R^{(k)}: \underbrace{Y}_{(I)} \to \underbrace{Y^{(k)}}_{\mathbb{R}^{k+1}} : (R^{(k)} v)_i = v(t_i^{(k)})$$

$$G^{(k)}(R^{(k)} y, x) \to 0 \text{ für } k \to \infty$$

„Stabilität“
Betrachte $H^{(k)}\colon Z^{(k)} \times X^{(k)} \to Y^{(k)}$ löst $G^{(k)}(y^{(k)}, x) = z^{(k)}$ $F^{(k)}(x) = H^{(k)}(0, x)$

$$Z_i^{(k)} = H_i^{(k)}(Z_i^{(k)} x) = \begin{cases} x + Z_0^{(k)} & i = 0 \\ H_{i-1}^{(k)}(Z_i^{(k)} x) + \Delta t^{(k)} f(t_{i-1}^{(k)}, H_{i-1}^{(k)}(Z_i^{(k)} x)) + \Delta t^{(k)} z_i^{(k)} \end{cases}$$

Zeige: H L-Stabil bezüglich Argument $Z^{(k)}$! $\Rightarrow$ Numerik 1!

2 Fließkommazahlen

Beispiel 2.1
$e^x \to$ Folien

2.1 Zahlendarstellung

Stellenwertsystem: $x \underbrace{\pm}_{\text{Vorzeichen}} \ldots m_2\beta^2 + m_1\beta + m_0 + m_{-1} + m_{-2}\beta^{-2} + \ldots$
β Basis, $\beta \in \mathbb{N}$, $\beta \geq 2$
$m_i \in \{0, 1, 2, \ldots, \beta - 1\}$ heißen *Ziffern*
Geschichte: [Knuth, Band 2, Seite 194]

- Babylonier ($\approx$ -1750), $\beta = 60$
- Basis 10 ab ca. 1580
- Pascal: $\beta \geq 2$ möglich

Festkommazahlen:

$$x = \pm \sum_{i=-k}^{n} m_i \beta^i$$

Problem:

$$\begin{aligned} \text{Plancksches Wirkungsquantum:} \quad & 6.626093 \cdot 10^{-34} Js \\ \text{Avogadrokonstante:} \quad & 6.021415 \cdot 10^{23} \tfrac{1}{\text{mod}} \end{aligned}$$

Definition 2.2 normierte Fließkommazahlen
Sei $\beta, v, s \in \mathbb{N}$ und $\beta \geq 2$. $\mathbb{F}(\beta, r, s) \subset \mathbb{R}$ besteht aus den Zahlen mit folgenden Eigenschaften:

(a) $\forall x \in \mathbb{F}(\beta, r, s)$ gilt $x = m(x) \cdot \beta^{e(x)}$ mit

$$m(x) = \pm \sum_{i=1}^{r} m_i \beta^{i-1} \text{ „Mantisse“} \qquad e(x) = \pm \sum_{j=0}^{s-1} e_j \beta^j \text{ „Exponent“}$$

Problem: doppelte Darstellungen. Beispiel: $\frac{1}{10} = 0.1 \cdot 10^0$ vs $0.01 \cdot 10^1$

(b) $\forall x \in \mathbb{F}(\beta, r, s)$ gilt $x = 0 \vee m_1 \neq 0$

Für $x \in \mathbb{F}(\beta, r, s)$ und $x \neq 0$ gilt wegen b)

$$\frac{1}{\beta} \leq |m(x)| < 1 \qquad \beta^{e(x)-1} \leq |x| \leq \beta^{e(x)} \tag{2.1}$$

Beispiel 2.3

(a) $\mathbb{F}(10, 3, 1)$ besteht aus den Zahlen:
$x = \pm(m_1 \cdot 0.1 + m_2 \cdot 0.01 + m_3 \cdot 0.001) \cdot 10^{\pm e_0}$
mit $m_1 \neq 0 \vee (m_1 = 0 \wedge m_2 = 0 \wedge m_3 = 0)$
z.B. $0.999 \cdot 10^1$ $\quad \cancel{0.024 \cdot 10^3} \rightsquigarrow 0.24 \cdot 10^2$

(b) $\mathbb{F}(2, 2, 1)$ besteht aus Zahlen der Form:

$$x = \pm \underbrace{\left(m_1 \frac{1}{2} + m_2 \frac{1}{4}\right)}_{\{0, \frac{1}{2}, \frac{3}{4}\}} \underbrace{\cdot 2^{\pm e_0}}_{\{2^{-1}, 1, 2^1\}} \qquad m_1 = 1$$

$$\mathbb{F}(2, 2, 1) = \left\{ \underbrace{-\frac{3}{2}, -1}_{\{-\frac{3}{4}, -\frac{1}{2}\} \cdot 2} \underbrace{-\frac{3}{4}, -\frac{1}{2}}_{\{-\frac{3}{4}, -\frac{1}{2}\} \cdot 1} \underbrace{-\frac{3}{8}, -\frac{1}{4}}_{\{-\frac{3}{4}, -\frac{1}{2}\} \cdot \frac{1}{2}}, 0, \frac{1}{4}, \frac{3}{8}, \frac{1}{2}, \frac{3}{4}, 1, \frac{3}{2} \right\}$$

normierte Fließkommazahlen $\mathbb{F}(\beta, r, s)$

(a) $x = \underbrace{\left(\pm \sum_{i=1}^{r} m + i\beta^{-i}\right)}_{m(x)} \cdot \beta^{\underbrace{\sum_{j=0}^{?} e_j \beta^j}_{=e(x)}}$

(b) $x = 0 \vee m_1 \neq 0$

Zahlenbereich:

Größte/kleinste Zahl:

$$X_{+/-} = \pm \underbrace{(\beta - 1)(\beta^{-1} + \ldots + \beta^{-r}}_{1-\beta^{-r}} \beta^{\underbrace{(\beta - 1)(\beta^{s-1} + \ldots + 1)}_{\beta^s - 1}} = \pm(1 - \beta^{-r})\beta^{\beta^s - 1}$$

Betragsmäßig die kleinste Zahlen außer 0:

$$x_{+/-} = \pm \beta^{-1} \beta^{-\underbrace{(\beta-1)(\beta^{s-1}+\ldots+1)}_{\beta^s - 1}} = \pm \beta^{-\beta^s}$$

$\mathbb{F}(\beta, r, s) \subset D(\beta, r, s) = [X_-, x_-] \cup \{0\} \cup [x_+, X_+]$

Praxis: IEEE754/IEC559 Standard

Ziel: Portabilität. Verabschiedet: 1985
$\beta = 2$ mit 4 Genauigkeitsstufen und normierter Darstellung

Format

	single	single-ext	double	double-ext
e_{max}	127	$\geq$ 1024	1023	$\geq$ 16384
e_{min}	-126	$\leq$ -1024	-1022	$\leq$ -16381
Bit Expon.	8	$\geq$ 11	11	$\geq$ 15
Bits total	32	$\geq$ 43	64	$\geq$ 79

Double genauer:

- total 64 bit
- Exponent: 11 bit : $c \in [1, 2046]$ $\quad 2^{11} = 2048 \quad e = c - 1023$
- Die Werte $c \in \{0, 2047\}$ werden anderweitig genutzt.

$$\begin{aligned} c = 0 \quad &\wedge \quad m = 0 \qquad \text{kodiert die Zahl x=0} \\ c = 2047 \quad &\wedge \quad m \neq 0 \qquad \text{Kodiert NaN: „not a number“} \\ c = 2047 \quad &\wedge \quad m = 0 \qquad \text{Kodiert Wert } \infty \text{ (Überlauf)} \end{aligned}$$

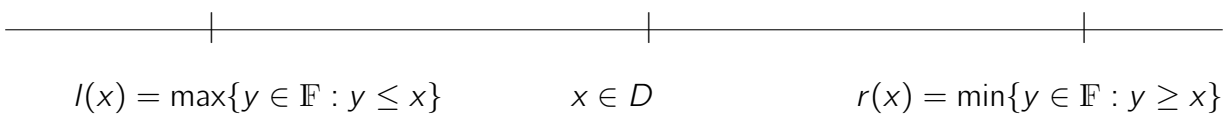
- Mantisse: $64 - 11 = 53$ Bit
  davon: 1 Bit Vorzeichen
  bleiben 52 Bit. Da $\beta = 2$ ist $\underbrace{m_1 \equiv 1}_{\text{(hidden Bit)}}$ damit $r = 53$
- 4 Rundungsarten: nach: $+\infty, -\infty, 0$, „natürliche Rundung“

2.2 Runden und Rundungsfehler

$\mathrm{rd}\colon D(\beta, r, s) \to \mathbb{F}(\beta, r, s)$
Im Fall des Exponenten über- oder Unterlaufs ist β, r, s *anzupassen*!
Sinnvollerweise verlangen wir:

$$|x - \mathrm{rd}(x)| = \min_{y \in \mathbb{F}} |x - y| \qquad \forall x \in D$$

Mit



gilt:

$$\mathrm{rd}(x) = \begin{cases} x & l(x) = r(x) \qquad (x \in F) \\ l(x) & |x - l(x)| < |x - r(x)| \\ r(x) & |x - l(x)| > |x - r(x)| \\ ? & |x - l(x)| = |x - r(x)| \end{cases}$$

Im letzten Fall ist Rundung erforderlich!
Sei $x = \text{sign}(x)\left(\sum_{i=1}^{\infty} m_i\right)\beta^{e(x)}$ normierte Darstellung von $x \in D$

natürliche Rundung:

$$\text{rd}(x) = \begin{cases} l(x) = \text{sign}(x)\left(\sum_{i=1}^{r} m_i\beta^{-i}\right)\beta^{e(x)} & \text{falls } 0 \le m_{r+1} < \frac{\beta}{2} \\ r(x) = l(x) + \beta^{e-r} & \text{falls } \frac{\beta}{2} \le m_{r+1} < \beta \end{cases}$$

Gerade Rundung (β gerade):

$$\text{rd}(x) = \begin{cases} l(x) & (|x - l(x) < |x - r(x)|) \vee (|x - l(x)| = |x - r(x)| \wedge m_r \text{ gerade}) \\ r(x) = l(x) + \beta^{e-r} & \text{sonst} \end{cases}$$

Definition 2.4
Sei $x' \in \mathbb{R}$ eine Näherung für $x \in \mathbb{R}$. Dann heißt

$$\Delta x = x' - x \qquad \text{„absoluter Fehler“}$$

und für $x \neq 0$

$$\varepsilon_{x'} = \frac{\Delta x}{x} \qquad \text{„relativer Fehler“}$$

Umformen liefert: $x' = x + \Delta x \overset{x\neq 0}{=} x\left(1 + \frac{\Delta x}{x}\right) = x\left(1 + \varepsilon_{x'}\right)$
$\Delta x = x' - x = 100km$

$$\begin{array}{rcccl} x & = & \text{Entfernung Erde-Sonne} & \sim & 1,5 \cdot 10^8 km \to \varepsilon_{x'} \approx 6,6 \cdot 10^{-7} \\ x & = & \text{Entfernung HD-Paris} & \sim & 500km \to \varepsilon_{x'} = 0,2 \end{array}$$

Lemma 2.5 Rundungsfehler
Bei der Rundung in $\mathbb{F}(\beta, r, s)$ gilt für den absoluten Fehler

$$|x - \text{rd}(x)| \le \frac{1}{2}\beta^{e(x)-r} \tag{2.2}$$

und für den relativen Fehler:

$$\frac{|x - \text{rd}(x)|}{|x|} \le \frac{1}{2}\beta^{1-r}$$

Beweis .
$x = \underbrace{\cdots}_{\infty \text{ viele Stellen}} m(x) \cdot \beta^{e(x)}$ normierte Darstellung von x:

$$|r(x) - l(x)| = \beta^{e(x)-r} \qquad \begin{array}{rcl} l(x) & = & 0, m_1, m_2, \ldots, m_r\beta^e \\ r(x) & = & 0, m_1, m_2, \ldots, (m_{r+1}) \cdot \beta^e \end{array}$$

(a) max. Fehler für $x = \frac{l(fx)+r(x)}{2}$

$$|x - \text{rd}(x) \le \left|\frac{l(x) + r(x)}{2} - l(x)\right| = \frac{1}{2}\left|r(x) - l(x)\right| = \frac{1}{2}\beta^{e-r}$$

(b)

$$\frac{|x - \mathrm{rd}(x)|}{|x|} \leq \frac{\frac{1}{2}\beta^{e(x)-r}}{|m(x)| \cdot \beta^{e(x)}} = \frac{1}{2} \underbrace{\frac{1}{|m(x)|}}_{|m(x)| \geq \beta^{-1}} \beta^{-r} \leq \frac{1}{2}\beta^{1-r}$$

□

eps $:= \frac{1}{2}\beta^{1-r}$ heißt *Maschinengenauigkeit* (u in MATLAB)

2.3 Fließkommarithmetik

Für $\circledast \in \{\oplus, \ominus, \odot, \oslash\}$ definiere
$\circledast\colon \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ als „Näherung" des entsprechenden $* \in \{+, -, \cdot, /\}$
Problem: $x, y \in F$ dann ist in der Regel $x * y \notin \mathbb{F}$
Wir definieren:

$$x \circledast y = \mathrm{rd}(x * y) \qquad \forall x, y \in \mathbb{F} \tag{2.3}$$

Man sagt dann $\circledast$ ist *exakt gerundet*

Beispiel 2.6
$\mathbb{F}(10, 3, 1) \qquad x = 0,215 \cdot 10^8 \qquad y = 0,125 \cdot 10^{-5}$
Betrachte: $x \ominus y = \mathrm{rd}(x - y)$

1) $x - y$:

$$\begin{array}{rcl} x & = & 0,2150000000000000 \cdot 10^8 \\ y & = & 0,0000000000000125 \cdot 10^8 \\ \hline (x-y) & = & 0,2149999999999875 \cdot 10^8 \end{array}$$

2) $x \ominus y = \mathrm{rd}(x - y) = 0,215 \cdot 10^8$

Wiederholung 2
$\mathbb{F}(\beta, r, s) \subset D(\beta, r, s) \subset \mathbb{R}$
$\mathrm{rd} : D(\beta, r, s) \to \mathbb{F}(\beta, r, s)$
$\frac{|x - \mathrm{rd}(x)|}{|x|} \leq \beta^{1-r} :=$ eps
$\oplus, \ominus, \odot, \oslash : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ exakt gerundet:
$x \circledast y := \mathrm{rd}(\underbrace{x * y}_{\in \mathbb{R}})$

Beispiel: Subtraktion
$m(x)\beta^{e(x)} - m(y)\beta^{e(y)-e(x)+e(x)} = \beta^{e(x)}(m(x) - m(y)\underbrace{\beta^{e(y)-e(x)}}$

Beispiel 2.6 Fortsetzung von 2.6

$$\begin{array}{rcl} x = 0.101 \cdot 10^1 & \longrightarrow & 0.101 \cdot 10^1 \\ y = 0.993 \cdot 10^0 & \xrightarrow{\text{schreiben}} & \underline{0.099 \cdot 10^1} \\ & & 0.002 \cdot 10^1 \end{array}$$

$(\beta = 10, r = 3, s = 1)$
relativer Fehler: $\frac{(x \ominus y)-(x-y)}{(x-y)} = \frac{0.002-0.017}{0.017} \approx 0.176 \approx 35\,\text{eps}$

1	zusätzliche „guard digit“	$\rightarrow$	$\left\lvert\frac{(x \ominus y)(x-y)}{(x-y)}\right\rvert \leq 2\,\text{eps}$
2	zusätzliche „guard digit“s	$\rightarrow$	exakt gerundet

2.4 Fehleranalyse

$\rightsquigarrow$ Fortpflanzung von Rundungsfehlern in Rechnungen
Rechnung $\equiv$ Funktionsauswertung
$F : \mathbb{R}^m \rightarrow \mathbb{R}^n$
$F_1(x_1, \ldots, x_n), \cdots, F_n(x_1, \ldots, x_n)$
$F' = \mathbb{F}^m \rightarrow \mathbb{F}^n$ „numerische Realisierung“ von F
F' wird durch *Algorithmus* realisiert.

- endlich Viele (=Terminierung)
- elementare (=bekannte) Rechenoperationen $(\oplus, \ominus, \odot, \oslash)$

$F'(x) = \varphi_e(\ldots \varphi_2(\varphi_1(x)) \ldots)$

Wichtig:

(i) Zu einem F gibt es in der Regel *viele* Realisierungen im Sinne unterschiedlicher Reihenfolge, da z.B. $(a \oplus b) \oplus c \neq a \oplus (b \oplus c)$

(ii) Jedes φ_i steuert unbekannten Fehler bei

(iii) eigentliche Folge $F^{(k)} : (\mathbb{F}^{(k)})^m \rightarrow (\mathbb{F}^{(k)})^n$

Wie in Kapitel 1 nutze Aufspaltung

$$F(x) - F'(\text{rd}(x)) = \underbrace{F(x) - F(\text{rd}(x))}_{\text{(1) Konditionsanalyse von } F} + \underbrace{F(\text{rd}(x)) - F'(\text{rd}(x))}_{\text{(2) Rundungsfehleranalyse}}$$

(1) Differentielle Konditionsanalyse
F zweimal stetig differenzierbar. Nach Satz von Taylor:

$$F_i(x + \Delta x) = F_i(x) + \sum_{j=1}^{m} \frac{\partial F_i}{\partial x_j}(x)\Delta x_j + R_i^F(x; \Delta x) \qquad i = 1, \cdots, n$$

Für das Restglied: $R_i^F(x; \Delta x) = \mathcal{O}(\|\Delta x\|^2)$

Definition 2.7 (Landausche Symbole)
Man schreibt

$$g(t) = \mathcal{O}(h(t)) \qquad (t \rightarrow \infty)$$

falls es $t_o > 0$ und $c_0 \geq 0$ gibt sodass für alle $t \in (0, t_0]$ die Abschätzung

$$|g(t)| \leq C_0 |h(t)|$$

gilt. Sprechweise: „$g(t)$ geht wie $h(t)$ gegen 0“.
Weiter bedeutet

$$g(t) = o(h(t)) \qquad (t \to 0)$$

dass es $t_0 > 0$ und eine Funktion $c(t); \lim_{t\to 0} c(t) = 0$ gibt sodass

$$|g(t)| \le c(t)|h(t)|$$

„$g(t)$ geht schneller als $h(t)$ gegen 0“ (falls $h(t) \to 0$)

Umformen: $F_i(x+\Delta x) - F_i(x) = \underbrace{\sum_{j=1}^{m} \frac{\partial F_i}{\partial x_j}(x)\Delta x_j}_{\text{„Führende (erste)“ Ordnung}} + \underbrace{\mathcal{O}(\|\Delta x\|^2)}_{\text{höhere Ordnung}}$

Näherung in erster Ordnung: $\doteq$

$$\frac{F_i(x+\Delta) - F_i(x)}{F_i(x)} \doteq \sum_{j=1}^{m} \frac{\partial F_i}{\partial x_j}(x)\frac{\Delta x_j}{F_i(x)} \doteq \sum_{j=1}^{m} \underbrace{\left(\frac{\partial F_i}{\partial x_j}(x)\frac{x_j}{F_i(x)}\right)}_{\substack{\text{Verstärkungs-}\\ \text{faktor } k_{ij}(x)}} \cdot \underbrace{\left(\frac{\Delta x_j}{x_j}\right)}_{\substack{\text{relativer Eingabe-}\\ \text{fehler } \le \text{eps}}}$$

Definition 2.8
Auswertung von $y = F(x)$ heißt „schlecht konditioniert“ in x falls $|k_{ij}(x)| \gg 1$, andernfalls „gut konditioniert“.
$|\overline{k}_{ij}(x)| < 1 \to$ Fehlerdämpfung,
$|\overline{k}_{ij}(x)| > 1 \to$ Fehlerverstärkung

Beispiel 2.9

(a) Addition: $F(x_1, x_2) = x_1 + x_2 \qquad \frac{\partial F}{\partial x_1} = 1 \qquad \frac{\partial F}{\partial x_2} = 1$

$$\frac{F(x_1+\Delta x_1, x_2+\Delta x_2) - F(x_1, x_2)}{F(x_1, x_2)} \doteq \underbrace{1 \cdot \frac{x_1}{x_1+x_2}}_{\overline{k}_{11}} \cdot \underbrace{\frac{\Delta x_1}{x_1}} + \underbrace{1 \cdot \frac{x_2}{x_1+x_2}}_{\overline{k}_{12}} \frac{\Delta x_2}{x_2}$$

schlecht konditioniert für $x_1 \to -x_2$!

(b) $F(x_1, x_2) = x_1^2 - x_2^2 \qquad \frac{\partial F}{\partial x_1} = 2x_1 \qquad \frac{\partial F}{\partial x_2} = -2x_2$

$$\frac{F(x+\Delta x) - F(x)}{F(x)} = \underbrace{2x_1 \cdot \frac{x_1}{x_1^2 - x_2^2}}_{\overline{k}_{11} = 2\frac{x_1^2}{x_1^2 - x_2^2}} \frac{\Delta x + 1}{x_1} - \underbrace{2x + 2\frac{x_2}{x_1^2 - x_2^2}}_{\overline{k}_{12} = -2\frac{x_2^2}{x_1^2 - x_2^2}} \frac{\Delta x_2}{x_2}$$

schlecht konditioniert für $|x_1| \approx |x_2|$

(2) Rundungsfehleranalyse

d.h. $F(x) - F'(x) \qquad x \in \mathbb{F}^m$!

da $\circledast$ exakt gerundet gilt: Es gibt ein ε mit $|\varepsilon| \leq \text{eps}$

$$\frac{(x \circledast y) - x * y}{x * y} = \varepsilon$$

$$\Leftrightarrow x \circledast y = (x * y)(1 + \varepsilon) \qquad \text{für ein } |\varepsilon(x, y)| \leq \text{eps} = \frac{1}{2}\beta^{1-r}$$

Analyse in „erster Näherung“:

Beispiel 2.10

$$\begin{aligned}
F(x_1, x_2) &= x_1^2 - x_2^2 &&= (x_1 - x_2)(x_1 + x_2) \\
F_a(x_1, x_2) &= (\underbrace{x_1 \odot x_1}_{u}) \ominus (\underbrace{x_2 \odot x_2}_{v}) \\
F_b(x_1, x_2) &= &&(\underbrace{x_1 \ominus x_2}_{u}) \odot (\underbrace{x_1 \oplus x_2}_{v})
\end{aligned}$$

(a) $u = x_1 \odot x_1 = (x_1 \cdot x_1)(1 + \varepsilon_1)$
$v = x_2 \odot x_2 = (x_2 \cdot x_2)(1 + \varepsilon_2) \qquad \varepsilon_1 \neq \varepsilon_2$ aber $|\varepsilon_i| \leq \text{eps}$

$$\begin{aligned}
F_a(x_1, x_2) = u \ominus v &= (u - v)(1 + \varepsilon_3) \\
&= \left(x_1^2(1 + \varepsilon_1) - x_2^2(1 + \varepsilon_2)\right)(1 + \varepsilon_3) \\
&= \underbrace{x_1^2 - x_2^2}_{=F(x_1,x_2)} + \underbrace{\varepsilon_1 x_1^2 - \varepsilon_2 x_2^2 + \varepsilon_3 x_1^2 - \varepsilon_3 x_2^2}_{\text{erste Ordnung}} + \underbrace{\varepsilon_1\varepsilon_3 x_1^2 - \varepsilon_2\varepsilon_3 x_2^2}_{\text{zweite Ordnung}}
\end{aligned}$$

relativer Fehler:

$$\frac{F_a(x_1, x_2) - F(x_1, x_2)}{F(x_1, x_2)} \doteq \underbrace{\frac{x_1^2}{x_1^2 - x_2^2}}_{\substack{\text{vergleichbar mit} \\ \text{Verstärkungsfunktion}}} \underbrace{(\varepsilon_1 + \varepsilon_3)} + \underbrace{\frac{x_2^2}{x_2^2 - x_1^2}} \underbrace{(\varepsilon_2 + \varepsilon_2)}$$

(b) $u = x_1 \ominus x_2 = (x_1 - x_2)(1 + \varepsilon_1)$
$v = x_1 \oplus x_2 = (x_1 + x_2)(1 + \varepsilon_2)$

$$\begin{aligned}
F_n(x_1, x_2) = u \odot v &= (u \cdot v)(1 + \varepsilon_3) \\
&= ((x_1 - x_2)(1 + \varepsilon_1) \cdot (x_1 + x_2)(1 + \varepsilon_2))\,(1 + \varepsilon_3) \\
&= (x_1^2 - x_2^2)(1 + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \ldots + \varepsilon_1\varepsilon_2\varepsilon_3)
\end{aligned}$$

$$\frac{F_b(x_1, x_2) - F(x_1, x_2)}{F(x_1, x_2)} \doteq \frac{\cancel{(x_1^2 - x_2^2)}(\varepsilon_1 + \varepsilon_2 + \varepsilon_3)}{\cancel{x_1^2 - x_2^2}} = \varepsilon_1 + \varepsilon_2 + \varepsilon_3$$

Definition 2.11

Wir nennen einen Algorithmus „numerisch stabil“, wenn die im Lauf der Rechnung akkumulierten Rundungsfehler aus (2) den unvermeidbaren Problemfehler aus (1) nicht übersteigen.

2.5 Auslöschung

tritt auf bei

- Addition und $x_1 \approx -x_2$
- Subtraktion und $x_1 \approx x_2$

Bemerkung 2.12
Bei der Auslöschung werden *vor* der entsprechenden Operation gemachte Fehler extrem verstärkt.
Achtung: Sind $x_1, x_2 \in \mathbb{F}$ so gilt natürlich

$$\left| \frac{(x_1 \ominus x_2) - (x_1 - x_2)}{x_1 - x_2} \right| \leq \text{eps}$$

Das Problem tritt erst auf, wenn x_1, x_2 bereits mit Fehlern behaftet sind.

Beispiel 2.13
$\mathbb{F}(10, 4, 1)$

$$\begin{array}{lll} x_1 & = 0,11258762 \quad 10^2 & \overset{\text{rd}}{\to} \quad 0,1126 \cdot 10^2 \\ x_2 & = 0,11244891 \quad 10^2 & \overset{\text{rd}}{\to} \quad 0,1124 \cdot 10^2 \\ \hline & = 0,00013871 \quad \cdot 10^2 & \quad 0,0002 \cdot 10^2 \\ & = 0,13871 \quad \cdot 10^{-1} & \quad = 0,2 \cdot 10^{-1} \end{array}$$

keine gültige Ziffer!
relativer Fehler (von $\text{rd}(x_1) \ominus \text{rd}(x_2)$)

$$\frac{0,1 \cdot 10^{-1} - 0,13871 \cdot 10^{-1}}{0,13871 \cdot 10^{-1}} \approx 0,44 \approx 883 \cdot \underbrace{\frac{1}{2} 10^{-3}}_{=\text{eps}}$$

$\text{eps} = \text{Abstand}(a, x_1) = \text{Abstand}(x_2, b)$

$\frac{2\,\text{eps} - \delta}{\delta}$ kann beliebig groß werden!

Regel 2.14
Setze potentiell gefährliche Operationen $+/-$ möglichst früh in einen Algorithmus ein

2.6 Die quadratische Gleichung

Die Gleichung

$$y^2 - py + q = 0$$

hat für $\frac{p^2}{4} > q \neq 0$ die beiden reellen und verschiedenen Lösungen

$$y_{1,2} = f_{\pm}(p, q) = \frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q}$$

Konditionsanalyse:

$$\frac{f_\pm(p+\Delta p, q+\Delta q) - f_\pm(p,q)}{f_\pm(p,q)}$$
$$\doteq \left(1 \pm \frac{p}{2\sqrt{\frac{p^2}{4}-q}}\right)\left(\frac{p}{p \pm 2\sqrt{\frac{p^2}{4}-q}}\right)\left(\frac{\Delta p}{p}\right) = \mp \frac{q}{\sqrt{\frac{p^2}{4}-q}\left(p \pm 2\sqrt{\frac{p^2}{4}-q}\right)}\left(\frac{\Delta q}{q}\right)$$

$\Rightarrow$ für $\frac{p^2}{4} \gg q$ und $p < 0$ ist $f_-(p,q) = \frac{p}{2} - \sqrt{\frac{p^2}{4} - q}$ gut konditioniert

für $\frac{p^2}{4} \gg q$ und $p > 0$ ist $f_+(p,q) = \frac{p}{2} + \sqrt{\frac{p^2}{4} - q}$ gut konditioniert

für $\frac{p^2}{4} \approx q$ sind f_+ und f_- schlecht konditioniert

Numerisch stabile Auswertung für den Fall $\frac{p^2}{4} \gg q$

- $p < 0 : y_2 = \underbrace{\frac{p}{2} - \underbrace{\sqrt{\frac{p^2}{4} - q}}_{\text{keine Auslöschung}}}_{\text{keine Auslöschung da } p<0}$ berechne $y_1 = \frac{q}{y_2}$

Vieta:
$p = y_1 + y_2$
$q = y_1 \cdot y_2$

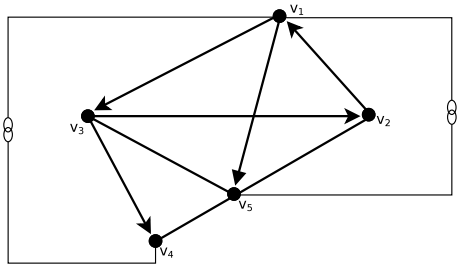
- $p > 0 : y_1 = \frac{p}{2} + \sqrt{\frac{p^2}{4} - q}$ berechne $y_2 = \frac{q}{y_1}$

3 Motivation linearer Gleichungssysteme

3.1 Strömungen in Rohrleitungsnetzen

1) Netzwerk von Röhren = gerichter Graph

- Knotenmenge
  $V = \{v_1, \ldots, v_n\}$ $\quad |V| = n$
- Kantenmenge
  $E = \{e_1, \ldots, e_M\}$
  $E \subseteq V \times V$
  $(v,w) \in E \Rightarrow (w,v) \notin E$
  $(v,w) \in E$ heißt:
  Es gibt Rohr *von v nach w*
- $E = R_R \cup E_p$ „Rohre“ und „Pumpen“
  $E_R = \{e_1, \ldots e_m\}$
  $E_p = \{e_{m+1}, \ldots, e_M\}$ $\quad |E_R| = m$



2) Gesetz von Hagen-Poiseuille

$e = (v, w)$ (Rohr von v nach w)

$$q_e = \frac{\pi r_e^4}{8\nu l_e} \Delta p_e \tag{3.1}$$

(r_e Radius der Röhre, l_e Länge der Röhre, ν dyn. Vikosität)

„q_e“: Volumenstrom in $\left[\frac{m^3}{s}\right] gerichtet$!

„Δ“: Druckdifferenz: $\left[p_a = \frac{N}{m^2}\right]$

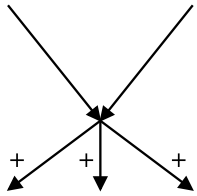
$q_e > 0$: Strom in Richtung des Rohres

$q_e < 0$: Strom entgegen Richtung des Rohres

3) Knotenregel

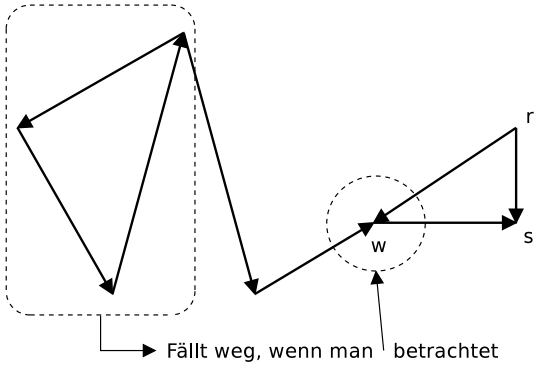
$E_v^+ = \{(u, w) \in E \mid u = v\}$ „ausgehend“
$E_v^- = \{(u, w) \in E \mid w = v\}$ „eingehend“

$$\sum_{e \in E_v^+} q_e - \sum_{e \in E_v^+} q_e = 0 \qquad \forall v \in V \tag{3.2}$$



nur $n - 1$ dieser Beziehungen sind linear unabhängig

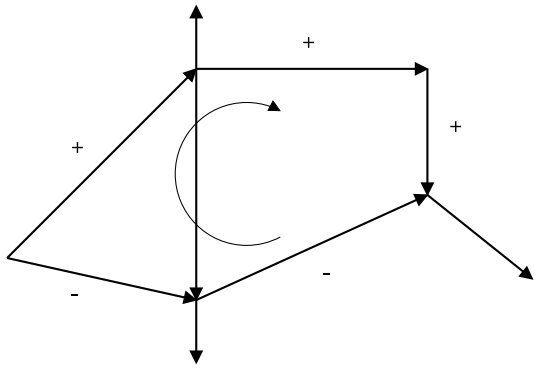
$$\sum_{v \in K\{w\}} \underbrace{\left(\sum_{e \in E_v^-} q_e - \sum_{e \in E_v^+} q_e \right)}_{0} = 0$$



4) Maschenregel:

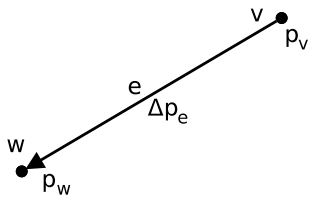
C: sei geschlossener Kantenzug
$C \subseteq E$
$C^+ = \{e \in C \mid e \text{ wie } C \text{ orientiert}\}$
$C^- = C - C^+$
Dann gilt:

$$\sum_{e \in C^+} \Delta p_e - \sum_{e \in C^-} \Delta p_e = 0$$



5) Knotenpotentialvefahren

- Wähle $n-1$ „Knotendrücke“ p_v („Potentiale“) $\rightarrow$ Unbekannte!
- $e = (v, w) \qquad \Delta p_e = p_v - p_w$
- Druck in Referenzknoten p_r ist Null



(a) $\forall e \in E_R$ schreibe Druckdifferenzen:

$$e = (v, w) \in E_R : \Delta p_e = \begin{cases} p_v - p_w & v \neq r \wedge w \neq r \\ p_v & w = r \\ p_w & v = r \end{cases}$$

in Matrixform:

Notation: $[\ ]_{n\times m}$ *bezeichnet eine* $n \times m$ *Matrix*

$$e_k = (v, w)k\,[\Delta p]_m = [B]_{m\times(n-1)}\,[p]_{n-1}$$

$p = (p_{v_1}, \ldots, p^T_{v_{n-1}}) \qquad$ mit $(r = v_n) \qquad \Delta p = (\Delta p_1, \ldots, \Delta p_m)^T$

(b) HP: $q_e = L_e \Delta p_e$

$[q]_m = [L]_{m\times m\ \mathrm{diag}}[\Delta p]_m$

(c) Knotenregel für Knoten $v_1, \ldots, v_{n-1}$:

$$\sum_{e\in E_v^-} q_e - \sum_{e\in E_v^+} = 0 \Leftrightarrow \underbrace{\sum_{e\in E_v^- \cap E_R} q_e - \sum_{e\in E_v^+ \cap E_R}}_{\text{Rohre}} = \underbrace{\sum_{e\in E_v^+ \cap E_p} q_e - \sum_{e\in E_v^- \cap E_p} q_e}_{\text{Pumpen bekannt}}$$

Jetzt!

$$\underbrace{B^T L B p}_{A} = b$$

- $(n-1) \times (n-1)$ Gleichungssystem für die $n-1$ Dreiecke p!
- A ist symmetrisch
- A ist positiv definit $x^T A x 0 \forall x \neq 0$
- A eindeutige Lösung
- $a_{ij} \neq 0$ wenn $(v_i, v_j) \in E \vee (v_j, v_i) \in E$

Wiederholung 3
Knotenpotentialverfahren
n Knoten, m Röhren

(a) Wähle $n-1$ unbekannte Knotendrücke
$p = (p_{v_1}, \cdots, p_{v_{n-1}} \qquad p_{v_n} = 0$
$e = (v, w) \qquad \Delta pe = p_v - p_w$
$\Delta p = Bp \qquad B = [\ ]_{n-1\times m} \qquad$ D.h. die k-te Spalte hat max. eine 1 und -1
Maschenregel erfüllt für alle Maschen

(b) Hagen Poiseuille
$e \in E \qquad q_e = l_e \Delta p_e$
$q = L\Delta p \qquad L = [\ ]_{m\times m} \qquad l_{ii} > 0$

(c) Knotenregel für $v_1, \cdots, v_{n-1}$ $B^T q = b$ (b: Pumpen) $\Rightarrow$

$$\underbrace{B^T L B}_{=A} p = b$$

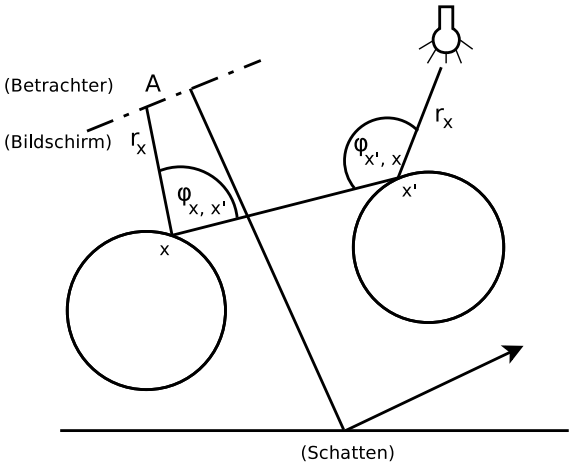
- A symmetrisch $A^T =$
  A ist positiv definit $x^T AX > 0 \forall x \neq 0$
  $x^T Ax = x^T \underbrace{B^T B}_{=A} x = (\underbrace{Bx}_{=:y})^T L (\underbrace{Bx}_{=:y}) = \sum_{i=1}^{n-1} l_{ii} y_i^2 \qquad x \neq 0 \Rightarrow y \neq$
- $0 \Rightarrow x^T AX > 0$
- A ist invertiertbar
- A ist dünn besetzt $|\{(i,j) \mid a_{ij} \neq 0\}| = \mathcal{O}(n)$
- elektrische Netzwerke
  - nur Widerstände
    Druck $p \leftrightarrow$ Spannung U
    Fließrate $q \leftrightarrow$ Strom I
    H.P. $\leftrightarrow$ Ohm'sches Gesetz
  - RLC-Netzwerke, harmonische Anregung $\rightarrow Ax = b$ mit $A \in \mathbb{C}^{n\times n}$

3.2 Radiositymethode in der Computergraphik

„Beleuchtung einer Szene“

Ray-Tracing

Radiosity



besser für „diffuse Belechung“

$S = \{x \in \mathbb{R}^3 \mid x \text{ ist auf Oberfläche des Objekts}\}$

Bestimme $B : S \rightarrow \mathbb{R}$ „Energiedichte“

$$\underbrace{\int_\omega B(s)ds \qquad \omega \leq s}_{\text{„abgestrahlte Engerie“}}$$

Gleichung für B:

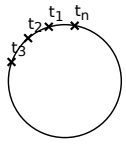
$$B(x) = \underbrace{E(x)}_{\substack{\text{„Eigen-}\\\text{strahlung“}}} + \underbrace{\underbrace{\varphi(x)}_{\substack{\text{Reflexions-}\\\text{faktor}}} \int_S B(x') \frac{\cos\varphi_{x,x'} \cos\varphi_{x',x}}{\|x - x'\|^2} V(x,x') ds}_{\lambda(x,x')} \qquad \varphi_{a,b} = \angle v_a, b - a)$$

„Integralgleichung“ $$V(x,x') = \begin{cases} 1 & x' \text{ von } x \text{ aus sichtbar} \\ 0 & \text{sonst} \end{cases}$$

Integralgleichung für $B(x)$:

$$B(x) - \varphi(x) \int_S B(x')\lambda(x,x')dA = E(x) \qquad \forall x \in S$$

Numerische Lösung mit „Kollokationsmethode“



Zerlege Oberfläche $\tau_k = \{t_1, \cdots, t_n\}$

(a) $t_i \subset S$ offen, $t_i \cap t_j = \qquad \forall i = j$
$\bigcup_{i=1}^n \overline{t_i} = S$ Zu t_i sei x_i der „Mittelpunkt“, „Diskretisierung“

(b) Approximiere $B : S \to \mathbb{R}$ durch eine diskrete Funktion B_h:

$$B_h(x) = \sum_{j=1}^n z_j \varphi_j(x) \qquad \begin{array}{c} z_j \in \mathbb{R} \text{ „Koeffizienten“} \\ \varphi_j : S \to \mathbb{R} \text{ „Basisfunktion“} \end{array} \qquad \varphi_j(x) = \begin{cases} 1 & x \in t_j \\ 0 & \text{sonst} \end{cases}$$

(c) Erfülle Integralgleichung für $x \in X_h = \{x_1, \cdots x_n\}$ ersetze B durch B_h

$$\begin{aligned} & B_h(x_i) - \varphi(x_i) \int_S B_h(x')\lambda(x,x')dA = E(x_i) \qquad \forall x_i \in X_h \\ \Leftrightarrow & \sum_{j=1}^n z_j \varphi)j(x_i) - \varphi(x_i) \int_s \left(\sum_{j=1}^n z_j \varphi_j(x') \right) \lambda(x,x')dA = E(x_i) \\ \Leftrightarrow & \sum_{j=1}^n \underbrace{\left\{ \varphi_j(x_i) - \varphi(x_i) \int_S \varphi)j(x')\lambda(x,x')dA \right\}}_{a_{ij}} = E(x_i) \\ \Leftrightarrow & Ax = n \end{aligned}$$

- Integral in a_{ij} evtentuell numerisch bestimmen ra später
- Man begeht „Diskretisierungsfehler“

$$\|B - B_h\| = \mathcal{O}(H^\alpha) \qquad h: \text{ Größe der } t_i$$

  $\Rightarrow$ „beliebig große“ Gleichungssysteme
- A ist *nicht* dünn besetzt

4 Konditionierung der Lösung linearer Gleichungssysteme

4.1 Lösbarkeit

$$A = (a_{ij})_{i,j=1}^{m,n} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \qquad b = b(b_i)_{i=1}^m = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

Gesucht ist $x = (x_j)_{j=1}^n = (x_1, \cdots, x_n)^T$ sodass

$$\forall 1 \leq i \leq m : \sum_{j=1}^{m} a_{ij} = b_i \tag{4.1}$$

(4.1) $\Leftrightarrow Ax = b$
Die Zahlen a_{ij}, b_i, x_j können aus $\mathbb{C}$ oder $\mathbb{R}$ sein. Das Gleichungssystem (4.1) heißt

- unterbestimmt falls $m < n$
- quadratisch falls $m = n$
- überbestimmt falls $m > n$

Es gibt mindestens eine Lösung falls

$$\text{Rang}(A) = \text{Rang}([A|b]) = \text{Rang}\left(\begin{bmatrix} a_{11} & \dots & a_{1n} & b_1 \\ \vdots & \ddots & \vdots & \vdots \\ a_{m1} & \dots & a_{mn} & b_m \end{bmatrix}\right)$$

Für quadratische Matrizen $A \in \mathbb{K}^{n \times n}$ sind folgende Aussagen äquivalent:

(i) $Ax = b$ ist für jedes b eindeutig lösbar

(ii) $\text{Rang}(A) = n$

(iii) $\det(A) \neq 0$

(iv) A hat keinen Eigenenwert 0

4.2 Vektorraum

Definition 4.1
Eine Abbildung $\|\cdot\| : \mathbb{K}^n \to \mathbb{R}_+$ heißt Norm falls gilt:

(N1) $\|x\| > 0 \qquad \forall x \neq 0$ (Definitheit)

(N2) $\|\alpha x\| = |\alpha| \|x\| \qquad x \in \mathbb{K}^n, \alpha \in \mathbb{K}$ (positive Homogenität)

(N3) $\|x + y\| \leq \|x\| \|y\| \qquad x, y \in \mathbb{K}^n$ (Dreiecksungleichung, Subadditivität)

Beispiel 4.2

$$\begin{array}{lcll} \|x\|_2 & = & (\sum_{i=1}^n |x_i|^2)^{\frac{1}{2}} & \text{euklidische Norm} \\ \|x\|_1 & = & \sum_{i=1}^n |x_i| & l_1 \text{ Norm} \\ \|x\|_\infty & = & \max_{1 \leq i \leq n} |x_i| & \text{Maximumsnorm} \end{array}$$

Konvergenz von Folgen von Vektoren:

$$\forall i = 1, \ldots, n : \lim_{t\to\infty} x_i^{(t)} \Leftrightarrow \lim_{t\to\infty} \|x^{(t)} - x\| = 0$$

(für jede Norm, im Endlichdimensionalen).
Eine wichtige Folgerung aus (N3) ist

$$\|x - y\| \geq |\|x\| - \|y\|| \tag{4.2}$$

$\Rightarrow$ Stetigkeit der Norm als Abbildung von $\mathbb{K}^n \to \mathbb{R}_+$

Beweis von 4.2.

$$\begin{array}{rcl} \|x\| = \|x - y + y\| & \overset{(N3)}{\geq} & \|x - y\| + \|y\| \;\Rightarrow\; \|x - y\| \geq \|x\| - \|y\| \\ \|y\| = \|y - x + x\| & \geq & \|x - y\| + \|x\| \;\Rightarrow\; \|x - y\| \geq \underbrace{\|y\| - \|x\|}_{-(\|x\|-\|y\|)} \end{array}$$

□

Satz 4.3 Äquivalenz aller Normen
Zu $\|\cdot\|$, $\|\cdot\|'$ gibt es Zahlen $m, M > 0$ aus $\mathbb{R}$ sodass gilt:

$$m\|x\|' = \|x\| \leq M\|x\|' \qquad \forall x \in \mathbb{K}^n$$

Beweis .
„$\Rightarrow$“ Es genügt dies für $\|\cdot\|' = \|\cdot\|_\infty$ zu zeigen.
Seien $e^{(k)}, 1 \leq k \leq n$ die Kartesischen Einheitsvektoren. Dann ist

$$x = \sum_{i=1}^n x_i e^{(i)}$$

$$\begin{aligned} \|x\| &= \left\| \sum_{i=1}^n x_i e^{(i)} \right\| \leq \sum_{i=1}^n |x_i| \; \|e^{(i)}\| \\ &\leq \sum_{i=1}^n \underbrace{(\max_{1\leq j\leq n} |x_j|)}_{=\|x\|_\infty} \; \|e^{(i)}\| \leq \|x_\infty\| \sum_{i=1}^n \|e^{(i)}\| = \gamma\|x\|_\infty \end{aligned}$$

„$\Leftarrow$“ Definiere Punktmenge:

$$S = \{x \in \mathbb{K}^n : \|x\|_\infty = 1\}$$

S ist

- beschränkt bezüglich $\|\cdot\|$
- abgeschlossen (Grenzwete von Folgen aus S ist in S)

und damit „kompakt“.

Folge:
Es gibt $\underline{x}, \overline{x} \in S$ sodass

$$\|\underline{x}\| \leq \|x\| \leq \|\overline{x}\| \qquad \forall x \in S$$

Nun sei $y \in \mathbb{K}\backslash\{0\}$, dann ist $\frac{y}{\|y\|_\infty} \in S$ und damit

$$\|\underline{x}\| \leq \left\| \frac{y}{\|y\|_\infty} \right\| = \frac{1}{\|y\|_\infty} \|y\| \leq \|\overline{x}\|$$

also

$$\underbrace{\|\underline{x}\|}_{m} \|y\|_\infty \leq \|y\| \leq \underbrace{\|\overline{x}\|}_{M} \|y\|_\infty$$

□

4.3 Matrixnormen

$\mathbb{K}^{n\times n}$ ist ein Vektorraum und kann mit dem $\mathbb{K}^{nn}$ identifiziert werden.

Definition 4.4
Eine Norm auf $\mathbb{K}^{n\times n}$ heißt *verträglich* mit $\|\cdot\|$ auf $\mathbb{K}^n$ falls gilt:

$$\|Ax\| \leq \|A\| \, \|x\| \qquad x \in \mathbb{K}^n, A \in \mathbb{K}^{n\times n}$$

Sie heißt Matrixnorm, wenn sie submultiplikativ ist:

$$\|AB\| \leq \|A\| \, \|B\| \qquad A, B \in \mathbb{K}^{n\times n}$$

Beispiel 4.5 Die Frobeniusnorm

$$\|A\|_{\mathrm{Fr}} = \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{\frac{1}{2}}$$

ist eine Matrixnorm, die verträglich mit $\|\cdot\|_2$ ist.

Definition 4.6
Es sei $\|\cdot\|$ ein beliebiger Vektorraum auf $\mathbb{K}^n$. Dann heißt

$$\|A\| := \sup_{x\in\mathbb{K}^n\backslash\{0\}} \frac{\|Ax\|}{\|x\|} = \sup_{\substack{\|x\|=1 \\ x\in\mathbb{K}^n}} \|Ax\|$$

zugeordnete Matrixnorm (oder natürliche Matrixnorm).
$\|\cdot\|$ ist verträglich und submultiplikativ:

$$\frac{\|A(\alpha x)\|}{\|\alpha x\|} = \frac{\cancel{|\alpha|} \, \|Ax\|}{\cancel{|\alpha|} \, \|x\|}$$

Hilfssatz 4.7
Die zugeordnete Matrixnorm zu $\|\cdot\|_\infty$ und $\|\cdot\|_1$ sind:

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^n |a_{ij}| \qquad \text{Zeilensummennorm}$$

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^n |a_{ij}| \qquad \text{Spaltensummennorm}$$

Beweis .

(a) Verträglichkeit

$$\|A\|_\infty = \max_{1\le i\le n} \left|\sum_{j=1}^n a_{ij}x_j\right| \le \max_{1\le i\le n} \sum_{j=1}^n |a_{ij}||x_j| \le \|x\|_\infty \underbrace{\max_{1\le i\le n} \sum_{j=1}^n |a_{ij}|}_{\|A\|_\infty} = \|A\|_\infty \|x\|_\infty$$

(b)

$$\sup \|A\|_\infty \overset{(a)}{\le} \sup_{\|x\|_\infty=1} \|A\|_\infty \underbrace{\|x\|_\infty}_{=1} = \|A\|_\infty$$

(c) ZZ ist nun, dass

$$\sup_{\|x\|_\infty=1} \|Ax\|_\infty \ge \|A\|_x$$

Für $A = 0$ damit

$$\sup_{\|x\|_\infty=1} \|Ax\|_\infty = \|A\|_\infty = 0$$

Also $A \ne 0$ und $\|A\|_\infty > 0$. Wähle einen Index $m \in \{1, \cdots, n\}$ mit

$$\sum_{j=1}^n |a_{nj}| = \max_{1\le i\le n} \sum_{j=1}^n |a_{ij}| = \|A\|_\infty$$

und definiere $z \in \mathbb{K}^n$ mit

$$z_k = \begin{cases} \frac{|a_{mk}|}{a_{mk}} & a_m k \ne 0 \\ 0 & \text{sonst} \end{cases}$$

$z_k \in \{-1, 0, +1\} \Rightarrow \|z\|_\infty = 1$

Für $v = Az$ gilt für die n-te Komponente

$$v_m = \sum_{k=1}^n a_{mk} z_k = \sum_{k=1}^n |a_{mk}| = \|A\|_\infty$$

und damit

$$\|A\|_\infty = v_m \le \|v\|_\infty = \|Az\|_\infty \overset{\text{da } \|z\|_\infty=1}{\le} \sup_{\|y\|_\infty=1} \|A\|_\infty$$

□

4.4 Eigenwerte und Eigenvektoren

Sei $A \in \mathbb{K}^{n\times n}$; $\lambda \in \mathbb{K}$ und $e \in \mathbb{K}^n$ sodass

$$Ae = \lambda e$$

dann heißt λ Eigenwert von A und e zugehöriger Eigenvektor. (λ, e) heißt auch „Eigenpaar“. $(A - \lambda I)e = 0$ ist für $e \neq 0$ nur für

$$P(\lambda) = \det(A - \lambda I) = 0$$

erfüllbar. Das charakteristische Polynom $P(\lambda)$ hat genau n Nullstellen in $\mathbb{C}$ (Vielfachheit mitgezählt). Zu jedem Eigenwert λ gibt es mindestens einen Eigenvektor e. Mit den Normregeln folgt für Eigenpaar (λ, e)

$$|\lambda| \overset{\|e\|=1}{=} |\lambda|\ \|e\| = \|\lambda e\| = \|Ae\| \leq \|A\| \underbrace{\|e\|}_{=1} = \|A\|$$

also $|\lambda| \leq \|A\| \forall$ Normen und λ

Definition 4.8
Spezielle Matrizen:

(a) Zu $A \in \mathbb{K}^{m\times n}$ heißt $A^T \in \mathbb{K}^{m\times n}$ transponierte Matrix
$(A^T)_{ij} = (A)_{ji}$

(b) Zu $A \in \mathbb{K}^{m\times n}$ heißt $\overline{A}^T \in \mathbb{K}^{m\times n}$ konjugiert komplexe Matrix
$(A)_{ij} = (\overline{A})_{ij}$

(c) $A \in \mathbb{K}^{n\times n}$ heißt hermitesch falls $A = \overline{A}^T$. D.h.
$(A)_{ij} = (\overline{A})_{ji}$ (manche schreiben $A^H = A^R$)

(d) Für $A \in \mathbb{C}^{n\times n}$ heißt
$A\overline{A}^T = \overline{A}^T A$ normal
$A\overline{A}^T = \overline{A}^T A = I$ unitär $(A^{-1} = \overline{A}^T)$

(e) Für reelle Matrizen $A \in \mathbb{R}^{n\times n}$ heißt
$AA^T = A^T A = I$ orthogonal $(A^{-1} = A^T)$

Definition 4.9 Skalarprodukt
Eine Abbildung $(;) : \mathbb{K}^n \times \mathbb{K}^n \to \mathbb{K}$ heißt Skalarprodukt, falls gilt:

(S1)	$(x, y) = \overline{(x, y)}$	$\forall x, y \in \mathbb{K}^n$	(Symmetrie)
(S2)	$(\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z)$	$\forall x, y, z \in \mathbb{K}^n; \alpha, \beta \in \mathbb{K}$	(Linearität)
(S3)	$(x, x) > 0$	$\forall x \in \mathbb{K}^n \setminus \{0\}$	(Definitheit)

Ein Skalarprodukt definiert eine zugehörige Norm

$$\|x\| = \sqrt{(x, x)} \qquad x \in \mathbb{K}^n$$

Es gilt die Cauchy-Schwarz-Ungleichung:

$$|(x,y)| \leq \|x\| \cdot \|y\|$$

Das euklidische Skalarprodukt lautet:

$$(x,y)_2 = \sum_{i=1}^{n} x_i \overline{y_i} = x^T \overline{y}$$

Dann gilt:

$$A = \overline{A}^T \Leftrightarrow (Ax,y)_2 = (x,Ay)_2 \qquad \forall x,y \in \mathbb{K}^n$$

4.5 Die Spektralnorm

Hilfssatz 4.10
Die der euklidischen Norm zugeordnete Matrixnorm heißt Spektralnorm und es gilt:

(a) Für hermitesche Matrizen A:

$$\|A\|_2 = \max\{|\lambda| : \lambda \text{ ist Eigenwert von } A\}$$

(b) Für jede Matrix $A \in \mathbb{K}^{n\times n}$ gilt:

$$\|A\|_2 = \max\{|\lambda| : \lambda \text{ ist Eigenwert von } \overline{A}^T A\}$$

Wegen

$$\underbrace{\max\{|\lambda| : \lambda \text{ ist Eigenwert von } A\}}_{=\|A\|_2 \text{ für } A \text{ hermitesch}} \leq \|A\|$$

ist damit

$$\|A\|_2 \leq \|A\|$$

Beweis .
A hermitesch: A hat n reelle Eigenwerte und einen vollständigen Satz von orthonormalen Eigenvektoren.

$$\{w^1,\ldots,w^n\} \subset \mathbb{K}^n : Aw^i = \lambda_i w^i \qquad (w^i,w^j) = \delta_{ij} = \begin{cases} 1 & i=j \\ 0 & \text{sonst} \end{cases}$$

$$\mathbb{K}^n \ni x = \sum_{i=1}^{n} \alpha_i w^j \qquad d_i = (x,w^i)_2$$

$$\|A\|_2 = \sup_P x \in \mathbb{K}^n \setminus \{0\} \frac{\|Ax\|_2}{\|x\|_2}$$

$$\|x\|_2^2 = (x,x)_2 = \left(\sum_{i=1}^{n} \alpha_i w^i, \sum_{j=1}^{n} \alpha_j w^j\right)_2 = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i, \overline{\alpha}_j \underbrace{(w^i,w^j)_2}_{\delta_{ij}} = \sum_{j=1}^{n} \lambda_i^2 |\alpha_i|^2$$

$$\|A\|_2^2 = (Ax, Ax)_2 = \left(\sum_{i=1}^n \underbrace{A\alpha_i w^i}_{=\alpha_i \lambda_i w^i}, \sum_{j=1}^n \underbrace{A\alpha_j w^j}_{=\alpha_j \lambda_j w^j} \right) = \sum_{i=1}^n \sum_{j=1}^n \alpha_i, \lambda_i \overline{\alpha}_j \overline{\lambda}_j (w^i, w^j)_2 = \sum_{i=1}^n \lambda_i^2 |\alpha_i|^2$$

$$\|A\|_2 = \sup_{x \in \mathbb{K}^n \setminus \{0\}} \frac{\|Ax\|_2}{\|x\|_2} = \frac{\sum_{i=1}^n |\lambda|^2 |\alpha_i|^2}{\sum_{i=1}^n |\alpha_i|^2} \leq \max_{1 \leq i \leq n} |\lambda_i|^2$$

$$\Rightarrow \|A\|_2 \leq \max_{1 \leq i \leq n} |\lambda_i|$$

Andererseits: $|\lambda| \leq \|A\| \forall$ Eigenwerte und jede Matrixnorm!

$$\Rightarrow \|A\|_2 = \max_{1 \leq i \leq n} |\lambda_i$$

A beliebig:

$$\|A\|_2^2 = (Ax, Ax)_2 = (Ax)^T (\overline{AX}) = x^T A^T (\overline{Ax}) = x^T (\overline{\overline{A}^T Ax}) = (x, \underbrace{\overline{A}^T A}_{\substack{=B \text{ ist} \\ \text{hermitesch}}} x)_2$$

□

4.6 Positiv definite Matrizen

Definition 4.11
Eine Matrix $A \in \mathbb{K}^{n \times n}$ heißt positiv definit in $\mathbb{K}$ wenn:

(a) $(Ax, x)_2 \in \mathbb{R} \forall x \in \mathbb{K}^n$

(b) $(Ax, x)_2 > 0 \forall x \in \mathbb{K}^n \setminus \{0\}$

$\mathbb{K} = \mathbb{R} \rightarrow$ nur b) relevant
$\mathbb{K} = \mathbb{C} \rightarrow$ a) eine zusätzliche Bedingung an die erlaubten Matrizen

Eigenschaft 4.12
Für $A \in \mathbb{C}^{n \times n}$ gilt: $(Ax, x) \in \mathbb{R} \forall x \in \mathbb{C}^n \Leftrightarrow A$ hermitesch

Beweis .
„$\Leftarrow$“ A ist hermitesch: $(Ax, y)_2 = (x, Ay)_2 \forall x, y \in \mathbb{C}^n$

$$\underbrace{(Ax, x)_2}_{z \in \mathbb{C}} \overset{\text{herm.}}{=} (x, Ax) \overset{(S1)}{=} \underbrace{\overline{(Ax, x)}}_{\overline{z}}$$

$$\Rightarrow \operatorname{Im}(z) = 0, \text{ also } z \in \mathbb{R}$$

„$\Rightarrow$“ $(Ax, x)_2 \in \mathbb{R} \forall x \in \mathbb{C}^1$ $\quad$ Z : A hermitesch

1) $(A(x+y), x+y)_2 = (Ax, x)_2 + (Ax, y)_2 + (Ay, x)_2 + (Ay, y)_2$
$\Leftrightarrow (Ax, y)_2 + (Ay, x)_2 = (A(x+y), x+y)_2 - (Ax, x)_2 - (Ay, y)_2 \in \mathbb{R}$ (nach Voraussetzung)

$$\Rightarrow \operatorname{Im}(Ax, y)_2 = -\operatorname{Im}(Ay, x)_2$$

2) Ersetze $1, x$ durch ix:

$$(A(ix), y)_2 + (Ay, ix)_2 = i\underbrace{[(Ax, y)_2 - (Ay, x)_2]}_{\substack{\text{rein imaginär} \\ \text{d.h. Realteil } =0}} \in \mathbb{R} \Rightarrow \mathrm{Re}(Ax, y)_2 = \mathrm{Re}(Ay, x)_2$$

$\mathbb{C} \ni (Ax, y)_2 = \mathrm{Re}(Ax, y)_2 + i\,\mathrm{Im}(Ax, y)_2 = Re(Ay, x)_2 - i\,\mathrm{Im}(Ay, x)_2 = \overline{(Ay, x)}_2 \overset{S1}{=} (x, Ay)_2$
d.h. A hermitesch
d.h. positiv definite Matrizen in $\mathbb{C}$ sind immer hermitesch.
Aber: positiv definite Matrizen in $\mathbb{R}$ sind *nicht notwendigerweise* symmetrisch. □

Lemma 4.13
Eine hermitesche Matrix A ist positiv definit, wenn alle ihre (reellen) Eigenwerte positiv sind. Alle Hauptdiagonalelemente sind positiv.

Beweis .

- A ist hermitesch mit lauter positiven Eigenwerten

$$x = \sum_{i=1}^{n} \alpha_i w^i \qquad w^i \text{ Eigenvektorbasis}$$

$$(Ax, x)_2 = \sum_{i,j=1}^{n} \lambda_i \alpha_i \overline{\alpha}_j (w^i, w^j)_2 \sum_{i=1}^{n} \lambda_i |\alpha_i|^2 > 0$$

da $x \neq 0$ ist mindestens ein $\alpha_i \neq 0$

$$0 < (\underbrace{Aw^i}_{\lambda_i w^i}, w^i)_2 = \lambda_i \underbrace{(w^i, w^i)}_{=1} = \lambda)_i$$

- e^i karthesischer Einheitsvektor $e^i_j = \delta_{ij} \qquad 0 < (Ae^i, e^i)_2 = a_{ii}$

□

Lemma 4.14
Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Dann liegt das betragsmäßig größte Element auf der Hauptdiagonale.

4.7 Störungstheorie

Sei $A \in \mathbb{K}^{n\times n}$ regulär und $x, b \in \mathbb{K}^n$

$$G(x) = Ax - b = 0 \rightsquigarrow x = F(A, b) = A^{-1}b$$

Zunächst: Variere nur b! (Untersuche Kondition)

$$\frac{\|F(A, b+\delta b) - F(A, b)\|}{\|\delta b\|} \cdot \frac{\|b\|}{\|F(A, b)\|} = \frac{\|A^{-1}(b+\delta b) - A^{-1}b\|}{\|\delta b\|}$$

relative Kondition

$$\frac{\|AA^{-1}b\|}{\|A^{-1}b\|} \overset{\substack{\text{verträgliche} \\ \text{Normen}}}{\leq} \frac{\|A^{-1}\| \cdot \cancel{\|\delta b\|}}{\cancel{\|\delta b\|}} \cdot \frac{\|A\| \cdot \cancel{\|A^{-1}b\|}}{\cancel{\|A^{-1}b\|}} = \|A\|^{-1} \cdot \|A\|$$

Definition 4.15
Die Zahl

$$\mathrm{cond}(A) = \|A\| \cdot \|A^{-1}\|$$

heißt *Konditionszahl* einer Matrix (geht für beliebige verträgliche Matrixnormen).

z.B. $\mathrm{cond}_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty$

Hilfssatz 4.16

1) $\forall x \in \mathbb{K}^n$ gilt:

$$\|x\| = \|x + Bx - Bx\| \leq \|x + Bx\| + \|Bx\|$$
$$\Leftrightarrow \|(I+B)x\| \geq \|x\| - \|Bx\| \geq \|x\| - \|B\|\ \|x\| = \underbrace{(1 - \|B\|)}_{>0 \text{ wg. } \|B\|<1} \|x\|$$

damit ist $(I+B)x \neq 0$ für $x \neq 0$ also $J + B$ regulär

2)

$$1 = \|I\| = \|(J+B)(J+B)^{-1}\| = \|(J+B)^{-1} + B(J+B)^{-1}\|$$
$$\geq \|(I+B)^{-1}\| - \|B\| \cdot \|(I+B)^{-1}\| = \underbrace{(1 - \|B\|)}_{0 \text{ nach Vor.}} \|(I+B)^{-1}\|$$

Satz 4.17 Störungssatz
$A \in \mathbb{K}^{n\times n}$ sei regulär und $\|\delta A\| < \frac{1}{\|A^{-1}\|}$. Dann ist $\tilde{A} = A + \delta A$ ebenfalls regulär und es gilt für den relativen Fehler in der Lösung des gestörten Systems $(A + \delta A)(\underbrace{x}_{\substack{\text{Lsg. von}\\ Ax=b}} + \delta x) = (b + \delta b)$:

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\mathrm{cond}(A)}{1 - \mathrm{cond}(A) \cdot \frac{\|\delta A\|}{\|A\|}} \cdot \left\{ \frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right\}$$

Beweis .

1) $A + \delta A = A(I + \underbrace{A^{-1}\delta A}_{„B“})$ und $\|A^{-1}\delta A\| \leq \|A^{-1}\|\ \|\delta A\| < \|A^{-1}\| \cdot \frac{1}{\|A^{-1}\|} = 1$

$\rightsquigarrow A + \delta A$ regulär nach 4.16.

2)

$$(A + \delta A)(x + \delta x) = (b + \delta b)$$
$$\Leftrightarrow \underbrace{Ax}_{b} + \delta A x + (A + \delta A)\delta x = b + \delta b$$
$$\Leftrightarrow (A + \delta A)\delta x = \delta b - \delta A x$$
$$\Leftrightarrow \delta x = (A + \delta A)^{-1}(\delta b - \delta A x)$$

$$
\begin{aligned}
\|\delta x\| &\le \|(A+\delta A)^{-1}\|(\|\delta b\| + \|\delta A\|\cdot\|x\|) \\
&= \|(A(I+A^{-1}\delta A))^{-1}\|(\|\delta b\| + \|\delta A\|\cdot\|x\|) \\
&\le \underbrace{\|(I+A^{-1}\delta A)^{-1}\|}_{4.16}\cdot\|A^{-1}\|(\|\delta b\| + \|\delta A\|\cdot\|x\|) \\
&\le \frac{\|A^{-1}\|}{1-\|A^{-1}\delta A\|}(\|\delta b\| + \|\delta A\|\cdot\|x\|) \\
&\le \frac{\|A^{-1}\|\cdot\|A\|\cdot\|x\|}{1-\|A^{-1}\|\cdot\|\delta A\|\cdot\underbrace{\|A\|\cdot\|A\|^{-1}}_{=1}}\cdot\left(\underbrace{\frac{\|\delta b\|}{\|A\|\cdot\|x\|}}_{\|b\|=\|Ax\|\le\|A\|\cdot\|x\|} + \frac{\|\delta A\|}{\|A\|}\right) \\
&\le \|x\|\cdot\frac{\mathrm{cond}(A)}{1-\mathrm{cond}(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|}\right)
\end{aligned}
$$

□

Beispiel 4.18

Sei $\frac{\|\delta A\|}{\|A\|} \approx \frac{\|\delta A\|}{\|b\|} \approx 10^{-k}$, $\qquad \mathrm{cond}(A) \approx 10^S$
Weiter sei $10^S \cdot 10^{-k} \leqq 1$
Dann gilt:

$$\frac{\|\delta x\|}{\|x\|} \le \underbrace{\frac{10^S}{1-\underbrace{10^S\cdot 10^{-k}}_{\leqq 11}}}_{\text{Nenner}\approx 1} 2\cdot 10^{-k} \approx 10^{S-k}$$

$\Rightarrow$ Man verliert s Stellen an Genauigkeit

Beispiel 4.19

(a) $A\cdot\begin{bmatrix} -1 & 1 \\ 1+\varepsilon & -1 \end{bmatrix}, \qquad A^{-1} = \frac{1}{\varepsilon}\begin{bmatrix} -1 & -1 \\ -(1+\varepsilon) & -1 \end{bmatrix}, \qquad \varepsilon = \det(A)$

Es gilt: $\|A\|_\infty = \max(2, 2+\varepsilon), \qquad \|A^{-1}\|_\infty = \frac{1}{\varepsilon}\max(2, 2+\varepsilon)$

$$\mathrm{cond}_\infty(A) = \frac{\max(2,2+\varepsilon)}{\varepsilon} = \mathcal{O}\left(\frac{1}{\det(A)}\right)$$

Kleine Determinanten = große Kondition?

(b) $B = \begin{bmatrix} 10^{-10} & 0 \\ 0 & 10^{-10} \end{bmatrix} \qquad B^{-1} = \begin{bmatrix} 10^{10} & 0 \\ 0 & 10^{10} \end{bmatrix}$

$\mathrm{cond}_\infty(B) = \|B\|_\infty\cdot\|B^{-1}\|_\infty = 10^{-10}\cdot 10^{10} = 1$

5 Eliminationsverfahren zur Lösung linearer Gleichungsysteme

5.1 Dreieckssysteme (gestaffelte Gleichungssysteme)

Sei $A \in \mathbb{R}^{n\times n}$ von oberer Dreiecksgestalt. Zu lösen ist also:

$$\begin{array}{ccccccccc} a_{11}\cdot x_1 & + & a_{12}\cdot x_2 & + & \cdots & + & a_{1n}\cdot x_n & = & b_1 \\ & + & a_{22}\cdot x_2 & + & \cdots & + & a_{2n}\cdot x_n & = & b_2 \\ & & & & \ddots & & \vdots & & \vdots \\ & & & & & & a_{nn}\cdot x_n & = & b_n \end{array}$$

$$\text{System ist regulär} \Leftrightarrow a_{ii} \neq 0 \quad \forall i = 1, \cdots, n$$

Lösen durch rückwärts einsetzen:

$$x_n = b_n / a_{n,n} i$$

$$x_i = \left(b_i - \sum_{k=i+1}^{n} a_{ik} x_k \right) / a_{ii} i$$

Die Anzahl der Rechenoperationen ist:

$$N_{\substack{\text{Dreiecks-}\\ \text{form}}}(n) = \sum_{i=0}^{n-1} (2i+1) = n^2$$

Unteres Dreieckssystem: analog durch vorwärtseinsetzen

5.2 Gauß-Elimination

$A \in \mathbb{R}^{n\times n}$, regulär

Ziel:

Bringe A durch äquivalente Umformungen auf (obere) Dreiecksgestalt
Benutze

(i) Vertausche zwei Gleichungen

(ii) Addition des Vielfachen einer Gleichung zu einer anderen

$$[A, b] = \begin{bmatrix} a_{11} & \cdots & a_{1n} & b_1 \\ \vdots & \ddots & \vdots & \vdots \\ a_{n1} & \cdots & a_{nn} & b_n \end{bmatrix}$$

sodass nach $n-1$ Schritten $A^{(n-1)}$ obere Dreiecksgestalt hat.

Schritt 1:

$[A^{(0)}, b^{(0)}] \to [A^{(1)}, b^{(1)}]$

(a) $[A^{(0)}, b^{(0)}] \Rightarrow [\tilde{A}^{(0)}, \tilde{b}^{(0)}]$ (nach Zeilenvertauschung)

Bestimme $r \in \{1, \cdots, n\}$ sodass $a_{r1}^{(0)} \neq 0$. Dieses existiert wegen Regulärität von A. Vertausche Zeilen r und 1.

(b) $[\tilde{A}^{(0)}, \tilde{b}^{(0)}] \to [A^{(1)}, b^{(1)}]$

Für $i \in \{2, \cdots, n\}$:

- $q_{i1} = \tilde{a}_{i1}^{(0)} / \tilde{a}_{11}^{(0)}$ $\quad$ $\tilde{a}_{11}^{(0)}$ heißt „*Pivotelement*“
- $\forall j \in \{1, \cdots, n\} : a_{ij}^{(1)} = \tilde{a}_{ij}^{(0)} - q_{i1}\tilde{a}_{1j}^{(0)}$ $\quad$ $b_i^{(1)} = \tilde{b}_i^{(0)} = a_{i1}\tilde{b}_1^{(0)}$

wegen $a_{i1}^{(0)} = \tilde{a}_{i1}^{(0)} - \frac{\tilde{a}_{i1}^{(0)}}{\tilde{a}_{11}^{(0)}} \cdot \tilde{a}_{11}^{(0)} = 0$ für $i \geq 2$ und damit:

$$[A^{(1)}, b^{(1)}] = \left[\begin{array}{c|cccc} \tilde{a}_{11}^{(0)} & \tilde{a}_{12}^{(0)} & \cdots & \tilde{a}_{1n}^{(0)} & \tilde{b}_1^{(0)} \\ \hline 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} & b_n^{(1)} \end{array}\right]$$

$\rightsquigarrow$ rekursiv anwenden!

Schritt k:

$[A^{(k-1)}, b^{(k-1)}] \to [A^{(k)}, b^{(k)}]$ $\qquad 1 \leq \lessgtr n$

$$[A^{(k-1)}, b^{(k-1)}] = \left[\begin{array}{ccc|ccc|c} a_{11}^{(k-1)} & & * & & & & \\ & \ddots & & & * & & * \\ 0 & & a_{k-1\ k-1}^{(k-1)} & & & & \\ \hline & & & a_{kk}^{(k-1)} & \cdots & a_{kn}^{(k-1)} & b_k^{(k-1)} \\ & \underbrace{0}_{k-1} & & \vdots & \ddots & \vdots & \vdots \\ & & & a_{nk}^{(k-1)} & \cdots & a_{nn}^{(k-1)} & b_n^{(k-1)} \end{array}\right]$$

(a) $[A^{(k-1)}, b^{(k-1)}] \to [\tilde{A}^{(k-1)}, \tilde{b}^{(k-1)}]$

Bestimme $r \in \{k, \cdots, n\}$ sodass $a_{r-k}^{(k-1)} = 0$.

Tausche Zeilen r und k.

(b) $[\tilde{A}^{(k-1)}, \tilde{b}^{(k-1)}] \to [A^{(k)}, b^{(k)}]$. Für $i \in \{k+1, \cdots, n\}$

- $q_{ik} = \tilde{a}_{ik}^{(k-1)} / \tilde{a}_{kk}^{(k-1)}$
- $a_{ij}^{(k)} = \tilde{a}_{ij}^{(k-1)} - q_{ik}\tilde{a}_{kj}^{(k-1)}$ $\qquad$ für $j \in \{k, \cdots, n\}$
- $b_i^{(k)} = \tilde{b}_i^{(k-1)} - q_{ik}\tilde{b}_k^{(k-1)}$

Algorithmus 5.1

Merke: Im Skript/Vorlesung: Indizes immer $1, \cdots, n$!

Input: $A \in \mathbb{R}^{n\times n}$, regulär, $b \in \mathbb{R}^n$
Output: System in oberer Dreiecksgestalt $[A^{(n-1)}, b^{(n-1)}]$ gespeichert in A, b!

```
for(k = 1; k < n; k = k+1) {
    Finde r ∈ {k, ···, n} sodass a_rk ≠0; sonst Fehler
    // Schritt a)
    if(r ≠ k) {
        for(j=k; j≤ n; j=j+1) {
            t   = a_kj;
            a_kj = a_rj;
            a_rj = t;
        }
        t   = b_k;
        b_k = br;
        b_r = t;
    }
    // Schritt b)
    for(i=k+1; i ≤ n; i = i+1) {
        q_ik = a_ik / a_kk; // (a_kk „Pivotelement“)
        for(j=k; j ≤ n; j=j+1)
            a_ij = a_ij - a_kj q_ik;
        b_i = b_i - q_ik b_k;
    }
}
```

Lemma 5.2
Der Aufwand der Transformation von A auf obere Dreiecksgestalt beträgt

$$N_{\text{Gauß(n)}} = \frac{2}{3}n^3 + \mathcal{O}(n^2)$$

Beweis.

$$\begin{aligned} N_{\text{Gauß}}(n) &= \sum_{k=1}^{n-1}\left\{\underbrace{n-k}_{\substack{\text{Berechnung}\\ \text{der } q_{ik}}} + (n-k)\left[(n-k)\cdot\underbrace{2}_{\substack{1\text{ Mult}\\ 1\text{ Add}}}\right]\right\} \\ &= 2\sum_{k=1}^{n-1}(n-k)^2 + 3\sum_{k=1}^{n-1} \\ &= \frac{2}{3}n^3 + \mathcal{O}(n^2) \end{aligned}$$

□

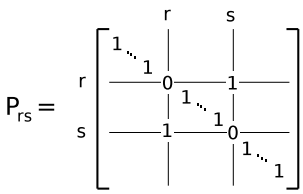
5.3 LR-Zerlegung

„Vorübung“: Kompaktue Formulierung der GEM mittels Matrixmultiplikation. Für Schritt a) (Zeilentauschen) betrachte die sogenannte *Permuationsmatrizen*

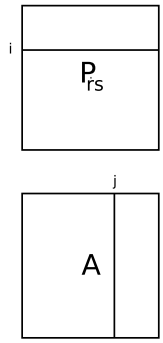
$$P_{rs} \in \mathbb{R}^{n\times n} \qquad 1 \le r \qquad s \le n \qquad r \ne s$$

mit

$$P_{rs} = \begin{cases} 1 & (i=j) \wedge (i \ne r) \wedge (i \ne s) \\ 1 & (i = r \wedge j = s) \vee (i = s \wedge j = r) \\ 0 & \text{sonst} \end{cases}$$



Hilfssatz 5.3
Die Matrix P_{rs} habe die folgenden Eigenschaften:



(i) $\tilde{A} = P_{rs}A$ ist A mit *Zeilen* r und s vertauscht

(ii) $\tilde{A} = AP_{rs}$ ist A mit *Spalten* r und s vertauscht

(iii) $P_{rs}^T = P_{rs}$

(iv) $P_{rs}^{-1} = P_{rs} \Rightarrow P_{rs}^{-1} = P_{rs}^T$ (orthogonal)

Beweis.

(i) $\tilde{a}_{ij} \sum_{k=1}^{n} (P_{rs})_{ik} a_{kj}$

$$\begin{aligned} i \ne r \wedge i \ne s \quad &: \quad k = i \quad \to \tilde{a}_{ij} = a_{ij} \quad \forall j \\ i = r \quad &: \quad k = s \quad \to \tilde{a}_{rj} = a_{ij} \quad \forall j \\ i = s \quad &: \quad k = r \quad \to \tilde{a}_{sj} = a_{rj} \quad \forall j \end{aligned}$$

(ii) analog

(iii) hinschauen

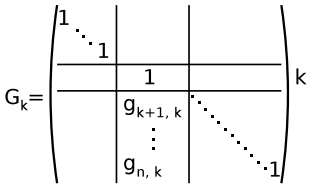
(iv) $P_{rs}P_{rs} = I$ wegen i). D.h. $P_{rs}^{-1} = P_{rs}$

□

Wiederholung 4
LR-Zerlegung: P_{rS} (siehe oben) $\qquad P_{k_i r_k}$ mit $r_k \ge k$

Definition 5.4 Frobeniusmatrizen

$G_k \in \mathbb{R}^{n\times n} \qquad 1 \le k \le n$
$G_k - I + G_k' \qquad \left(G_k'\right)_{ij} = 0$ wenn $j \ne k$ oder $i \le k$



Hilfssatz 5.5
Die Frobeniusmatrizen haben folgende Eigenschaften:

(i) $\tilde{A} = G_k A$ gilt: $\begin{cases} a_{ij} & i \le k \\ a_{ij} + gika_{kj} & i > k \end{cases}$

(ii) $G'_k G'_k = 0$

(iii) $G_k^{-1} = I - G'_k$

(iv) $G_1 G_2 \cdots G_k = I + \sum_{j=1}^{k} G'_j$

analog: $G_1^{-1} \dots G_k^{-1} = I - \sum_{j=1}^{k} G'_k$

(v) $\forall k < \alpha \le \beta \qquad P_{\alpha\beta} G_k = (I + P_{\alpha\beta} G'_k) P_{\alpha\beta}$

Beweis .

(i) $\tilde{a}_{ij} = \sum_{j=1}^{n} (G_k)_{is} a_{sj} = \begin{cases} a_{ij} & i \le k \\ g_{ik} a_{kj} + a_{ij} & i > k \end{cases}$

(ii)

$$G'_k = \begin{array}{c} \left(\begin{array}{c|ccc} 0 & 0 & & \\ \hline & 0 & & \\ 0 & g_{k+1,\,k} & \ddots & \\ & \vdots & & \ddots \\ & g_{n,\,k} & & 0 \end{array}\right) \begin{array}{l} \rangle\ \text{k-1} \\ \\ \\ \\ \\ \end{array} \\ \begin{array}{cc} \underbrace{\quad}_{\text{k-1}} & \underbrace{\qquad\qquad}_{\text{n-k+1}} \end{array} \end{array}$$

$$G'_k G'_k = \left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & B_k \end{array}\right)\left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & B_k \end{array}\right) = \begin{pmatrix} 0 & 0 \\ 0 & \underbrace{B_k B_k}_{=0} \end{pmatrix}$$

$$B_K = \left(\begin{array}{c|c} 0 & 0 \\ \hline * & 0 \end{array}\right)\left(\begin{array}{c|c} 0 & 0 \\ \hline * & 0 \end{array}\right) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

(iii) $(I + G'_k)(I - G'_k) = I - G'_k + G'_k - \underbrace{G'_k G'_k}_{=0} = I$

(iv) Per Induktion

$k = 1 : G_1 = I + G'_1 \checkmark$

$k - 1 \to k$: d.h. $G_1 \cdots G_{k-1} = I + \sum_{j=1}^{k-1} G'_j$

$$\underbrace{G_1 \cdots G_{k-1}}\, G_k = \left(I + \sum_{j=1}^{k-1} G'_j\right)(I + G'_k) = I + \sum_{j=1}^{k-1} G'_j + G'_k + \left(\sum_{j=1}^{k-1} G'_j\right) G'_k$$

$$\begin{array}{c} \text{k-1} \\ \text{n-k+1} \end{array} \left\langle \begin{array}{c} \left(\begin{array}{c|c} \begin{matrix} 0 & & 0 \\ & \ddots & \\ * & & 0 \end{matrix} & 0 \\ \hline * & 0 \end{array}\right) \\ \begin{array}{cc} \underbrace{\quad}_{\text{k-1}} & \underbrace{\qquad}_{\text{n-k+1}} \end{array} \end{array} \right. \left. \begin{array}{c} \left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & \begin{array}{c|c} 0 & \\ \hline * & 0 \end{array} \end{array}\right) \\ \begin{array}{cc} \underbrace{\quad}_{\text{k-1}} & \underbrace{\qquad}_{\text{n-k+1}} \end{array} \end{array} \right\rangle \begin{array}{l} \text{k-1} \\ \text{n-k+1} \end{array} = 0$$

(v)

$$\begin{aligned} P_{\alpha\beta}G_k &= P_{\alpha\beta}(I + G'_k) \\ &= P_{\alpha\beta} + P_{\alpha\beta}G_k \\ &= P_{\alpha\beta} + P_{\alpha\beta}G'_k \underbrace{P_{\alpha\beta}}_{*} \\ &= (\underbrace{I + P_{\alpha\beta}G'_k}_{\substack{\text{Wieder eine} \\ \text{Frobeniusmatrix!}}})P_{\alpha\beta} \end{aligned}$$

$*$: Kein Problem, da $\alpha, \beta > k$ und damit werden nur Null-Spalten getauscht!

□

Wir führen folgende Abkürzung für das Produkt von Matrizen ein:

$$B_b \cdots B_{a+1}B_a = \prod_{i=a}^{b} B_i$$

Satz 5.6 (LR-Zerlegung)
Sei $A \in \mathbb{R}^{n\times n}$ regulär. Dann gibt es eine Zerlung

$$PA = LR$$

wobei

$$L = \begin{pmatrix} 1 & & & 0 \\ l_{21} & \ddots & & \\ \vdots & & \ddots & \\ l_{n1} & \cdots & \cdots & 1 \end{pmatrix} \qquad R = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{pmatrix}$$

$$P = \prod_{k=1}^{n-1} P_{k_1 r_k} \qquad r_k \le k$$

Im Fall $P = I$ ist die Zerlegung eindeutig.

Beweis .

(a) Zunächst sei $P = I$, d.h. kein Zeilentausch notwendig.

$$\begin{array}{rrcl} & Ax & = & b \\ \text{Schritt 1:} & G_1Ax & = & G_1b \\ \text{Schritt 2:} & G_2G_1Ax & = & G_2G_1b \\ \vdots & & & \\ n-1: & \underbrace{G_{n-1}\cdots G_1A}_{R}x & = & G_{n-1}\cdots G_1b \\ \text{d.h.:} & G_{n-1}\cdots G_1A & = & R \\ & A & = & G_1^{-1}\cdots G_{n-1}^{-1}R \\ & & \overset{(iii)}{=} & \underbrace{(I - G'_1)\cdots(I - G'_{n-1})}R \\ & & \overset{(iv)}{=} & \underbrace{\left(I - \sum_{i=1}^{n-1} G'_j\right)}_{=:L}R \end{array}$$

(b) Mit Zeilentausch: GEM („Rüberschaufeln“)

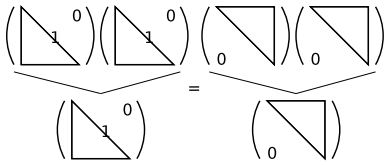
$$
\begin{aligned}
G_{n-1}P_{n-1,r_{n-1}}\cdots G_2P_{2,r_2}G_1P_{1,k}A &= R \qquad \beta \geq \alpha > k \\
G_{n-1}\cdots P_{3,r_3}G_2(I+P_{2,r_2}G_1')P_{2,r_2}P_{1,r_1}A &= R \\
G_{n-1}\cdots(I+P_{3,r_3}G_2')(I+P_{3,r_3}P_{2,r_2}G_1')P_{3,r_3}P_{2,r_2}P_{1,r_1}A &= R \\
&\vdots
\end{aligned}
$$

$$
\prod_{k=1}^{n-1}\left(I+\left(\prod_{\alpha=k+1}^{n-1}P_{\alpha,r_\alpha}\right)G_k'\right)\underbrace{\left(\prod_{k=1}^{n-1}P_{k,r_k}\right)}_{=:P}A = R
$$

$$
\Leftrightarrow \underbrace{\prod_{p=n-1}^{1}\left(I-\left(\prod_{\alpha=k+1}^{n-1}P_{\alpha,r_\alpha}\right)G_k'\right)}_{=:L}R = PA
$$

(c) Eindeutigkeit
Angenommen es gibt zwei Zerlegungen: $A = L_1R_1 = L_2R_2$

$$
I = AA^{-1} = \underbrace{L_1R_1}_{A}\underbrace{R_2^{-1}L_2^{-1}}_{A^{-1}} \Leftrightarrow L_1^{-1}L_2 = R_1R_2^{-1} = I
$$



□

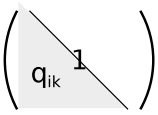
Wo ist L?

Aus dem Beweis:

$$
L = \prod_{p=n-1}^{1}\left(I-\left(\prod_{\alpha=k+1}^{n-1}P_{\alpha,r_\alpha}\right)G_k'\right) = I - \sum_{k=1}^{n-1}\left(\prod_{\alpha=k+1}^{n-1}P_{\alpha,r_\alpha}\right)G_k'
$$

$$
(G_k')_{ik} = -q_{ik} \text{ aus GEM}
$$

- Nichtnullelemente in G_k' nehmen gerade den Platz der im Schritt k zu eliminierenden Einträge von A ein.
- Die Eintrage in G_k' nehmen an allen späteren Zeilentauschen teil



Wozu das Ganze?

Lösen von $Ax = b$ mittels LR-Zerlegung

$$\begin{aligned} Ax &= b \\ \underbrace{PA}\, x &= Pb \\ \Leftrightarrow L\, \underbrace{Rx}_{=:y} &= Pb \\ \Leftrightarrow Ly &= Pb \\ Rx &= y \end{aligned}$$

d.h.

1) Berechne LR-Zerlegung von A: $PA = LR$ (Aufwand: $\mathcal{O}(n^3)$)
2) $b' = Pb$ (Aufwand: $\mathcal{O}(n)$)
3) $Ly = b'$ lösen (Aufwand: $\mathcal{O}(n^2)$)
4) $Rx = y$ lösen (Aufwand: $\mathcal{O}(n^2)$)

Schritt 1-3 $\widehat{=}$ GEM

Vorteil:

Löse $Ax_i = b_i$ für verschiedene b_i $\quad i = 1, \cdots, m$
Aufwand: $\mathcal{O}(n^3) + \mathcal{O}(mn^2)$
$\Rightarrow$ d.h. günstiger bei mehreren rechten Seiten!

zu den Permuationsmatrizen:

$P = P_{n-1,r_{n-1}} \cdots P_{2,r_2} P_{1,r_1}$ mit $r_k \geq k$

- werden nicht gespeichert
- Es werden nur die Indizes $r_1, \cdots, r_{n-1}$ gespeichert
- Anwendung *nicht* als Matrixmultiplikation

Algorithmus zur LR-Zerlegung

Eingabe: $A \in \mathbb{R}^{n\times n}$ (wird überschrieben)
Ausgabe:
$L \in \mathbb{R}^{n\times n}$ in $l_{ij} = a_{ij}$ für $j < i$
$R \in \mathbb{R}^{n\times n}$ in $r_{ij} = a_{ij}$ für $j \geq i$
$P : \{1, \cdots, n\} \to \{1, \cdots, n\}$

```
for(k=1; k ≤ n; k = k+1) {
    Finde r ∈ {k, ···, n} sodass a_rk ≠ 0; sonst Fehler
    for(j=1; j≤n; j=j+1) {
        t  = a_kj;
        a_kj = a_rj;
        a_rj = t;
    }
    p[k] = r; // merke Permuation
    for(i=k+1; i≤n; i=i+1) {
        l_ik=a_ik/a_kk; // setzen von l_ik!
        for(j=k+1; j ≤ n; j=j+1) {
            a_ij=a_ij-l_ik a_kj;
        }
    }
}
```

5.4 Rundungsfehleranalyse der LR-Zerlegung (GEM)

Absolutwertnotation

Sei $A \in \mathbb{R}^{m\times n}$ dann ist

$$B = |A| \Rightarrow b_{ij} = |a_{ij}| \qquad 1 \le i \le m, 1 \le j \le n$$

(analog $|x|$)
Erweitern der Abbildung rd auf $\mathbb{R}^n, \mathbb{R}^{m\times n}$. Dann gilt:

$$\mathrm{rd}(A) = A + A' \text{ mit } \underbrace{|A'| \le |A|\,\mathrm{eps}}_{m\cdot n \text{ Ungleichungen!}}$$

weil

$$\mathrm{rd}(a_{ij}) = a_{ij}(1+\varepsilon_{ij}) = a_{ij} + \underbrace{\varepsilon_{ij}a_{ij}}_{a_{ij}} \qquad |a'_{ij}| \le |a_{ij}|\,\mathrm{eps}$$

fl-Notation (gl-Notation)

Sei E eine Formel (d.h. z.B. $E = ax + by + \cdots$) dann bezeichnet $\mathrm{fl}(E)$ die Auswertung der Formel E in Gleitkommaarithmetrik (Reihenfolge beachten). Man kann auch schreiben:

$$\mathrm{fl}(\sqrt{x}), \mathrm{fl}(\sin(x))$$

Beispiel:

$$\mathrm{fl}(\underbrace{A+B}_{A,B\in\mathbb{F}^{m\times n}}) = (A+B) + H \qquad |H| \le \mathrm{eps}\,|A+B|$$

(kurz für $\mathrm{fl}(a_{ij} + b_{ij} = (a_{ij} + b_{ij}) + \varepsilon_{ij}(a_{ij} + b_{ij}))$

Rückwärtsanalyse

Bisher haben wir Rundungsfehler in der „Vorwärtsanalyse“ betrieben. z.B. für Skalarprodukt:

$$|\,\mathrm{fl}(x^Ty) - x^Ty| \le n\,\mathrm{eps}\,|x^T||y| + \mathcal{O}(\mathrm{eps}^2) \text{ (absolut)}$$

Alternative ist die „Rückwärtsanalyse“: Schreibe Ergebnis der Fließkommarechnung als exaktes Ergebnis eines modifizierten Ausdrucks. z.B. Skalarprodukt

$$\mathrm{fl}(x^Ty) = (x+f)^Ty \qquad |f| \le n\,\mathrm{eps}\,|x| + \mathcal{O}(\mathrm{eps}^2)$$

Beispiel 5.7

Betrachte Lösung von $Ax = b$. Die GEM im Rechner liefert $\hat{x} \in \mathbb{F}^n$.

Vorwärtsanalyse: $\|\hat{x} - x\| \le F(n, \mathrm{eps}, A, b)$

Rückwärtsanalyse: $(A+E)\hat{x} = b \qquad |E| \le F'(n, \mathrm{eps}, A)$

Hilfssatz 5.8

Es gilt für $x, y \in \mathbb{F}^n$

$$\hat{s} = \mathrm{fl}(x^Ty) = (x+f)^Ty \qquad |f| \le n\cdot\mathrm{eps}\,|X| + \mathcal{O}(\mathrm{eps}^2)$$

Beweis .
Per Induktion $n = 1$:

$$\hat{s}_1 = \mathrm{fl}(x_1 y_1) = x_1 y_1 (1 + \delta_1) = (x_1 + \underbrace{x_1 \delta_1}_{=f_1}) y_1$$

also

$$|f_1| \leq |\delta_1||x_1| \leq \mathrm{eps}\,|x_1|$$

$n \geq 2$:

$$\begin{aligned}
\hat{s}_n = \mathrm{fl}(x^T y) = \mathrm{fl}(\tilde{x}^T \tilde{y} + x_n y_n) \qquad x = \begin{pmatrix} \tilde{x} \\ \vdots \\ x_n \end{pmatrix}, y = \begin{pmatrix} \tilde{y} \\ \vdots \\ y_n \end{pmatrix} \\
&= (\mathrm{fl}(\tilde{x}^T \tilde{y}) + \mathrm{fl}(x_n y_n))(1 + \varepsilon_n) \\
&= ((\tilde{x} + \tilde{f})^T \tilde{y} + x_n y_n (1 + \delta_n))(1 + \varepsilon_n) \\
&= (\tilde{x} + \tilde{f}^T \tilde{y}(1 + \varepsilon_n) + x_n y_n (1 + (\delta_n + \varepsilon_n) + \delta_n \varepsilon_n) \\
&= (\tilde{x} + \tilde{f} + \tilde{x}\varepsilon_n \tilde{f})^T \tilde{y} + (x_n + (\delta_n + \varepsilon_n) x_n + \delta_n \varepsilon_n x_n) y_n \\
&= \left[\begin{pmatrix} \tilde{x} \\ \vdots \\ x_n \end{pmatrix} + \underbrace{\begin{pmatrix} \tilde{f} + \varepsilon_n \tilde{x} + \varepsilon_n \tilde{f} \\ \vdots \\ (\delta_n + \varepsilon_n) x_n + \delta_n \varepsilon_n x_n \end{pmatrix}}_{=f} \right] \begin{pmatrix} \tilde{y} \\ \vdots \\ y_n \end{pmatrix}
\end{aligned}$$

mit

$$\begin{aligned}
|\tilde{f} + \varepsilon_n \tilde{x} + \varepsilon_n \tilde{f}| &\leq (n-1)\,\mathrm{eps}\,|\tilde{x}| + \mathrm{eps}\,|\tilde{x}| + \mathcal{O}(\mathrm{eps}^2) \\
|(\delta_n + \varepsilon_n) x_n + \delta_n \varepsilon_n x_n| &\leq 2\,\mathrm{eps}\,|x_n| + \mathcal{O}(\mathrm{eps}^2) \\
\text{wegen } n \geq 2 \text{ gilt: } |f| &\leq n \cdot \mathrm{eps}\,|x| + \mathcal{O}(\mathrm{eps}^2)
\end{aligned}$$

□

Satz 5.10 (Lösen der Dreieckssysteme)
Es seien $\hat{x}, \hat{y}$ die *numerischen* Lösungen des unteren bzw. oberen Dreickssystems $Lx = b$ und $Ry = c$. Dann gilt:

$$\begin{aligned}
(L + F)\hat{x} = b \qquad |F| \leq n\,\mathrm{eps}\,|L| + \mathcal{O}(\mathrm{eps}^2) \\
(R + G)\hat{y} = c \qquad |G| \leq n\,\mathrm{eps}\,|R| + \mathcal{O}(\mathrm{eps}^2)
\end{aligned}$$

Beweis Als Übungsaufgabe.

□

Satz 5.11 (LR-Zerlegung)
Sei $A \in \mathbb{F}^{n \times n}$. Es werden die LR-Zerlegung von A *ohne Zeilentausch* berechnet. Dann gilt für die numerisch berechneten $\hat{L}, \hat{R}$:

$$\hat{L}, \hat{R} = A + H \text{ mit } |H| \leq 3(n-1)\,\mathrm{eps}(|A| + |\hat{L}||\hat{R}|) + \mathcal{O}(\mathrm{eps}^2)$$

Beweis .
$n = 1$: Es ist $\hat{l}_{11} = 1$ und $r_{11} = a_{11}$ und somit $\hat{l}_{11}\hat{r}_{11} = a_{11}$ und $|H| = 0$.
$n \geq 2$: Schreibe

$$A = \begin{bmatrix} \alpha & w^T \\ v & B_{n-1\times n-1} \end{bmatrix}_{n\times n}$$

Dann „macht“ α die LR-Zerlegung:

(a) $\hat{z} = \mathrm{fl}\left(\frac{v}{\alpha}\right)$

(b) $\hat{A}_1 = \mathrm{fl}\left(B - \hat{z}w^T\right)$

(c) Berechne LR-Zerlegung von $\hat{A}_1$

(a) $\hat{z} = \frac{v}{\alpha} + f$ mit $|f| \leq \mathrm{eps}\,\frac{|v|}{|\alpha|}$ (Rundungsfehler in der Division)

(b) Induktionsschritt:

$$\begin{aligned}
\hat{A}_1 &= \mathrm{fl}\left(B - \hat{z}w^T\right) \\
&= B - \mathrm{fl}\left(\hat{z}w^T\right) + G \qquad \text{mit } |G| \leq \mathrm{eps}\left|B - \mathrm{fl}\left(\hat{z}w^T\right)\right| \\
&= B - (\hat{z}w^T + \underbrace{G') + G}_{=F} \quad \text{mit } |G'| \leq \mathrm{eps}\left|\hat{z}w^T\right| \leq \mathrm{eps}\,|\hat{z}||w|^T \\
\hat{A}_1 &= B - \hat{z}w^T + F \text{ mit}
\end{aligned}$$

$$\begin{aligned}
|F| &= |-G' + G| \leq |G'| + |G| \\
&\leq \mathrm{eps}\,|\hat{z}||w|^T + \mathrm{eps}\,|B - \hat{z}w^T - G'| \\
&\leq \mathrm{eps}\,|\hat{z}||w|^T + \mathrm{eps}(|B| + |\hat{z}||w|^T + |G'|) \\
&\leq 2\,\mathrm{eps}(|B| + |\hat{z}||w|^T) + \mathcal{O}(\mathrm{eps}^2)
\end{aligned}$$

(c) $\hat{A}_1$ und LR-Zerlegt und es gilt die Induktionsannahme:

$$\hat{L}_1\hat{R}_1 = \hat{A}_1 + H_1 \qquad |H_1| \leq 3(n-2)\,\mathrm{eps}(|\hat{A}_1| + |\hat{L}_1||\hat{R}_1| + \mathcal{O}(\mathrm{eps}^2)$$

Blockform der LR-Zerlegung

$$\begin{aligned}
\hat{L}\hat{R} &= \begin{bmatrix} 1 & 0 \\ \hat{z} & \hat{L}_1 \end{bmatrix}\begin{bmatrix} \alpha & w^T \\ 0 & \hat{R}_1 \end{bmatrix} = \begin{bmatrix} \alpha & w^T \\ \alpha\hat{z} & \hat{z}w^T + \hat{L}_1\hat{R}_1 \end{bmatrix} \\
&= \begin{bmatrix} \alpha & w^T \\ \alpha\left(\frac{v}{\alpha} + f\right) & \cancel{\hat{z}w^T} + \underbrace{(B - \cancel{\hat{z}w^T} + F)}_{\hat{A}_1} + H_1 \end{bmatrix} = \underbrace{\begin{bmatrix} \alpha & w^T \\ v & B \end{bmatrix}}_{=A} + \underbrace{\begin{bmatrix} 0 & 0 \\ \alpha f & H_1 + F \end{bmatrix}}_{=:H}
\end{aligned}$$

Abschätzung von H:

$$
\begin{aligned}
|\hat{A}_1| &= |B - \hat{z}w^T + F| \le |B| + |\hat{z}||w|^T + \underbrace{|F|} \\
&\le |B| + |\hat{z}||w|^T + \overbrace{2\,\mathrm{eps}(\underbrace{|B| + |\hat{z}||w|^T}) + \mathcal{O}(\mathrm{eps}^2)} \\
&\le (1 + 2\,\mathrm{eps})(|B| + |\hat{z}||w|^T) + \mathcal{O}(\mathrm{eps}^2) \\
|H_1 + F| &\le |H_1| + |F| \\
&\le 3(n-2)\,\mathrm{eps}(|\hat{A}_1| + |\hat{L}_1||\hat{R}_1|) + 2\,\mathrm{eps}(|B| + |\hat{z}||w|^T) + \mathcal{O}(\mathrm{eps}^2) \\
&\le 3(n-2)((1 + 2\,\mathrm{eps})(|B| + |\hat{z}||w|^T) + |\hat{L}_1||\hat{R}_1|) + 2\,\mathrm{eps}(|B| + |\hat{z}||w|^T) + \mathcal{O}(\mathrm{eps}^2) \\
&\le 3(n-1)\,\mathrm{eps}(|B| + |\hat{z}||w|^T + |\hat{L}_1||\hat{R}_1|) + \mathcal{O}(\mathrm{eps}^2)
\end{aligned}
$$

und damit

$$
\begin{aligned}
|H| &= \begin{bmatrix} 0 & 0 \\ \underbrace{|\alpha f|}_{|\alpha||f|} & |H_1 + F| \end{bmatrix} \le \begin{bmatrix} 0 & 0 \\ \mathrm{eps}\,|v| & 3(n-1)\,\mathrm{eps}(|B| + |\hat{z}||w^T| + |\hat{L}_1||\hat{R}_1|) \end{bmatrix} + \mathcal{O}(\mathrm{eps}^2) \\
&\le 3(n-1)\,\mathrm{eps}\begin{bmatrix} 0 & 0 \\ |v| & |B| + |z||w|^T + |\hat{L}_1||\hat{R}_1| \end{bmatrix} + \mathcal{O}(\mathrm{eps}^2) \\
&\le 3(n-1)\,\mathrm{eps}\left(\begin{bmatrix} |\alpha| & |w|^T \\ |v| & |B| \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ |\hat{z}| & |\hat{L}_1| \end{bmatrix}\begin{bmatrix} |\alpha| & |w|^T \\ 0 & |\hat{R}_1| \end{bmatrix}\right) \\
&\le 3(n-1)\,\mathrm{eps}\left(|A| + \begin{bmatrix} |\alpha| & |w|^T \\ |\hat{z}||\alpha| & |\hat{z}||w|^T + |\hat{L}_1||\hat{R}_1| \end{bmatrix}\right)
\end{aligned}
$$

□

Satz 5.12
Seien $\hat{L}$ und $\hat{R}$ die numerisch berechneten Faktoren der LR-Zerlegng aus Satz 5.11. Sei weiter $\hat{y} \in \mathbb{F}^n$ die numerisch berechnete Lösung von $\hat{L}y = b$ und schließlich $\hat{x}$ die numerische Lösung von $\hat{r}x = \hat{y}$. Dann gilt:

$$(A + E)\hat{x} = b \text{ mit } |E| \le n\,\mathrm{eps}(3|A| + 5|\hat{L}||\hat{R}| + \mathcal{O}(\mathrm{eps}^2)$$

Beweis .
Wegen (5.10) gilt:

$$
\begin{aligned}
(\hat{L} + F)\hat{y} &= b & |F| &\le n\,\mathrm{eps}\,|\hat{L}| + \mathcal{O}(\mathrm{eps}^2) \\
(\hat{R} + G)\hat{x} &= \hat{y} & |G| &\le n\,\mathrm{eps}\,|\hat{R}| + \mathcal{O}(\mathrm{eps}^2)
\end{aligned}
$$

Einsetzen

$$(A + E)\hat{x}b \text{ mit } E = H + F\hat{R} + \hat{L}G + \underbrace{FG}$$

und

$$
\begin{aligned}
|E| &\le |H| + |F||\hat{R}| + |\hat{L}||G| + \mathcal{O}(\mathrm{eps}^2) \\
&\le 3(n-1)\,\mathrm{eps}(|A| + |\hat{L}||\hat{R}|) + n\,\mathrm{eps}\,|\hat{L}||\hat{R}| + n\,\mathrm{eps}\,|\hat{L}||\hat{R}| + \mathcal{O}(\mathrm{eps}^2)
\end{aligned}
$$

□

Folgerung

Störungssatz anwenden:

$$\frac{\|\hat{x}-x\|_\infty}{\|x\|_\infty} \underbrace{\leq}_{\frac{\|E\|_\infty}{\|A\|_\infty}\ll 1} \operatorname{cond}_\infty(A)\Bigg\{ \underbrace{3n\,\mathrm{eps}\,\frac{\cancel{\|A\|_\infty}}{\cancel{\|A\|_\infty}}}_{\text{ok, vergleichbar mit Rundungsfehler in Eingabe}} + 5n\,\mathrm{eps}\,\underbrace{\frac{\|\hat{L}\|_\infty\|\hat{R}\|_\infty}{\|\hat{L}\|_i nfty}}_{\text{evtl. Problem}} + \mathcal{O}(\mathrm{eps}^2)\Bigg\}$$

denn:

$$\hat{L}_{ij} = \frac{a_{ij}}{a_{ii}} \text{ sehr groß für } a_{ii} \text{ sehr klein!}$$

Kleine Pivotelemente $\Rightarrow$ großer Rundungsfehler.
Das ist unabhängig von der Kondition von A! z.B. $A = \left[\begin{smallmatrix} \varepsilon & 1 \\ 1 & 0 \end{smallmatrix}\right]$.
$\Rightarrow$ Gauß-Elimination ist in dieser Form *nicht numerisch stabil!*

5.5 Pivotisierung

Idee: Im Schritt k der GEM wähle $r \in \{k, \cdots, n\}$ sodass

$$\left|a_{rk}^{(k)}\right| \geq \left|a_{ik}^{(k)}\right| \forall k \leq i \leq n$$

(mache betragsmäßig größtes Element als Pivotelement) damit dann gilt

$$|\hat{l}_{ij}| \leq 1 \text{ und damit } \|\hat{L}\|_\infty \leq n$$

$\rightarrow$ „maximales Spaltenpivot“, „Spaltenpivotisierung“

Beispiel 5.13

$$\begin{bmatrix} -10^{-5} & 1 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$\curvearrowright$ GEM, exakte Rundung

$$\begin{bmatrix} -10^{-5} & 1 \\ 0 & 1+2\cdot 10^5 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2\cdot 10^5 \end{bmatrix}$$

$x_1 = 0,4999975, x_2 = 0,999995$
Jetzt $\mathbb{F}(10,4,1)$. Multiplikator:

$$\begin{aligned}
q_{21} &= (\overbrace{0,2\cdot 10^1}^{2}) \oslash (\overbrace{-0,1\cdot 10^{-4}}^{-10^{-5}}) = -0,2\cdot 10^6 \\
a_{22}^{(1)} &= \overbrace{0,1\cdot 10^1}^{1} \ominus (\overbrace{-0,2\cdot 10^6}^{a_{21}} \odot \overbrace{(0,1\cdot 10^1)}^{a_{12}^{(0)}}) \\
&= 0,1\cdot 10^1 \oplus 0,2\cdot 10^6 = 0,2\cdot 10^6 \text{ (Rundungsfehler!)} \\
b_2^{(1)} &= 0 \ominus (\overbrace{-0,2\cdot 10^6}^{q_{21}} \odot \overbrace{0,1\cdot 10^1}^{b_1^{(1)}}) = 0,2\cdot 10^6 \\
x_2 &= b_2^{(1)} \oslash a_{22}^{(1)} = 0,2\cdot 10^6 \oslash 0,2\cdot 10^6 = 0,1\cdot 10^1 = \boxed{1} \\
x_1 &= (\underbrace{0,1\cdot 10^1}_{b_1} - \underbrace{0,1\cdot 10^1}_{a_{12}} \cdot \underbrace{0,1\cdot 10^1}_{x_2}) \oslash (\underbrace{-0,1\cdot 10^{-4}}_{a_{11}}) = \boxed{0}
\end{aligned}$$

$\text{cond}_\infty(A) = 3$
Im Sinne einer Rückwärtsanalyse gilt:

$$\begin{bmatrix} -10^{-5} & 1 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

mit einer völlig anderen Lösung!
Spaltenpivotisierung:

$$\begin{bmatrix} 2 & 1 \\ -10^{-5} & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$\Rightarrow$ in $\mathbb{F}(10, 4, 1)$ erhält man:

$$x_2 = 1 \qquad x_1 = -0,5$$

ABER:

Multipliziere ursprüngliches System mit -10^6 in Zeile 1:

$$\begin{bmatrix} 10 & -10^6 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -10^6 \\ 0 \end{bmatrix}$$

- Spaltenpivotisierung tauscht keine Zeilen
- $a_{22}^{(1)} = 1 + 0,2 \cdot 10^6 = 0,2 \cdot 10^6$ in $\mathbb{F}(10, 4, 1)$ selber Fehler wie oben!

Deshalb zusätzlich: Skalierung (Äquilibrierung)

$$Ax = b \to \underbrace{D^{-1}A}_{\tilde{A}}\, x = \underbrace{D^{-1}b}_{\tilde{b}} \text{ mit } d_{ii} = \sum_{j=1}^{n} |a_{ij}| \Rightarrow \|\tilde{A}\|_\infty = 1$$

Alternative: totale Pivotisierung (siehe S 50)

Rundungsfehler bei Pivotisierung

Analoges Resultat für Satz 5.12 mit Spaltenpivotisierung liefert:

$$(A + E)\hat{x} = b \text{ mit } |E| \leq n\,\text{eps}(3|A| + 5P^T|\hat{L}||\hat{R}|) + \mathcal{O}(\text{eps}^2)$$

Nach Konstruktion $\|\hat{L}\|_\infty \leq n$ (Spaltenpivotisierung)
Nun definiert den „Wachstumsfaktor“:

$$\varrho = \max_{i,j,k} \frac{|\hat{a}_{ik}^{(k)}|}{\|A\|_\infty} \text{ (Einflüsse aus } |\hat{R}|)$$

Man kann zeigen, dass:

$$\|E\|_\infty \leq 8n^3 \|A\|_\infty \varrho\, \text{eps}$$

In der Praxis: $\varrho \approx 10$ Worst case: $\tilde{\varrho} 2^n$
Mit Totalpivotisierung gilt:

$$|a_{ij}^{(k)}| \leq k^{\frac{1}{2}} \cdot \sqrt{2 \cdot 3^{\frac{1}{2}} \cdot \ldots \cdot k^{\frac{1}{k-1}}} \cdot \max_{i,j} |a_{ij}|$$

$\Rightarrow$ wächst deutlich weniger stark!

- Totalpivotisierung theoretisch besser
- in der Praxis aber oft nicht notwendig

Totale Pivotisierung

Wähle $r, s \in \{k, \cdots, n\}$ sodass

$$|a_{rs}^{(k)}| \geq |a_{ij}^{(k)}| \forall k \leq i, j \leq n$$

und erreiche $\tilde{a}_{kk}^{(k)} = a_{rs}^{(k)}$ durch Zeilen- *und Spaltentausch*. In Matrixform:

$$\begin{array}{lrcr}
\text{Schritt 1:} & G_1 P_{r_1} A \overbrace{P_{s_1} P_{s_1}}^{=I} x & = & G_1 P_{r_1} b \\
\text{Schritt 2:} & G_2 P_{r_2} G_1 P_{r_1} A P_{s_1} P_{s_2} P) s_2 P_{s_1} x & = & G_2 P_{r_2} G_1 P_{r_1} b \\
\text{Schritt n-1:} & \underbrace{G'_{n-1} \cdots G'_1 \underbrace{P_{r_{n-1}} \cdots P_{r_1}}_{=:P} A \underbrace{P_{s_1} \cdot P_{s_{n-1}}}_{=:Q^T}}_{R} \overbrace{\underbrace{P_{s_{n-1}} \cdots P_{s_1}}_{Z} x}^{Q} & = & G'_{n-1} \cdots G'_1 P_{r_{n-1}} \cdots P_1 b
\end{array}$$

$$\Rightarrow \boxed{PAQ^T = LR}$$

Lösen des LGS gelingt dann mit $\underbrace{PAQ^T}_{L \cdot R} z = Pb$

1) $b' = Pb$ 2) $Ly = b'$ 3) $Rz = y$ 4) $x = Q^T z$

Aufwand der Pivotisierung

- $\frac{1}{2} n^2$ Vergleiche bei Spaltenpivotisierung
- $\frac{1}{3} n^3$ Vergleiche bei Totalpivotisierung

Hinweis: Rechenoperationen in modernen Rechnern eher uninteressant, aber Speicherzugriffe teuer! Daher lieber weniger Operationen durchführen.
$\Rightarrow$ fast doppelte Rechenzeit!

5.6 Spezielle Systeme

Symmetrisch positiv definite Matrizen

Satz 5.14
Eine s.p.d. Matrix $A \in \mathbb{R}^{n \times n}$ ist stets *ohne* Pivotisierung *stabil* LR-zerlegbar. Für die Diagonalelemente der im Eliminationsprozess auftretenden Matrizen gilt:

$$a_{ii}^{(k)} \geq \lambda_{\min}(A) > 0 \qquad k \leq i \leq n$$

Beweis. Betrifft einen Schritt der LR-Zerlegung:

(a) $A = \begin{bmatrix} \alpha & v^T \\ v & B_{n-1 \times n-1} \end{bmatrix}$

Elimination von v:

$$\begin{bmatrix} 1 & 0 \\ \frac{-v}{\alpha} & I \end{bmatrix} \begin{bmatrix} \alpha & v^T \\ v & B \end{bmatrix} = \begin{bmatrix} \alpha & v^t \\ 0^B - \frac{1}{\alpha} v v^T & \end{bmatrix}$$

$$= \begin{bmatrix} \alpha & 0 \\ 0 & B - \frac{1}{\alpha} v v^T \end{bmatrix} \begin{bmatrix} 1 & \frac{v^T}{\alpha} \\ 0 & I \end{bmatrix} \to \begin{bmatrix} 1 & \frac{v^T}{\alpha} \\ 0 & I \end{bmatrix}^{-1} = \begin{bmatrix} 1 & \frac{-v^T}{\alpha} \\ 0 & I \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} 1 & 0 \\ \frac{-v}{\alpha} & I \end{bmatrix}}_{x^T} \underbrace{\begin{bmatrix} \alpha & v^T \\ v & B \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} 1 & \frac{-v}{\alpha} \\ 0 & I \end{bmatrix}}_{\substack{X\text{hat} \\ \text{vollen Rang}}} = \underbrace{\begin{bmatrix} \alpha & 0 \\ 0 B - \frac{1}{\alpha} v v^T & \end{bmatrix}}_{=:\tilde{A}}$$

X hat vollen Rang, A s.p.d. $\Rightarrow X^T A X = \tilde{A}$ s.p.d.

dmait ist $B - \frac{1}{\alpha} v v^T$ als Hauptuntermatrix einer s.p.d. Matrix auch s.p.d.

(b) Es gilt: (Rayleight-Quotient) $(Ax, x)_2 \geq \lambda_{\min}(A)(x, x)_2$ und damit für $x = e^{(i)}$:

$$a_{ii} = (Ae^{(i)}, e^{(i)})_2 \geq \lambda_{\min}(A)$$

Für die Diagonalelemente von $B - \frac{1}{\alpha} v v^T$ gilt mit $\tilde{e}^{(i)} = \begin{pmatrix} 0 \\ e^{(i)} \end{pmatrix}$

$$\begin{aligned} (B - \frac{1}{\alpha} v v^T)_{ii} &= (\tilde{A}\tilde{e}^{(i)}, \tilde{e}^{(i)})_2 = (X^T A X \tilde{e}^{(i)}, \tilde{e}^{(i)})_2 = (AX\tilde{e}^{(i)}, X\tilde{e}^{(i)})_2 \\ &\geq \lambda_{\min}(A)(X\tilde{e}^{(i)}, X\tilde{e}^{(i)})_2 \geq \lambda_{\min}(A)(1 + \underbrace{\left(\frac{V_i}{\alpha}\right)^2}_{\geq 0}) \geq \lambda_{\min}(A) \end{aligned}$$

$$\begin{bmatrix} 1 & \frac{-v^T}{\alpha} \\ 0 & I \end{bmatrix} \begin{bmatrix} 0 \\ e^{(i)} \end{bmatrix} = \begin{bmatrix} \frac{-v_i}{\alpha} \\ e^{(i)} \end{bmatrix}$$

□

5.15 Cholesky-Zerlegung

Beobachtung: Pivotelemente bei s.p.d. Matrizen sind stets positiv.
Setze $D = \operatorname{diag}(R)$

$$A = LR = LD \underbrace{D^{-1}T}_{U} = LDU \overset{\text{s.p.d.}}{=} LDL^T$$

wegen $A = A^T$ folgt $U = L^T$
Da $d_{ii} > 0$ ist die Matrix „$D^{\frac{1}{2}}$“ mit $\left(D^{\frac{1}{2}}\right)_{ii} = \sqrt{d_{ii}}$ wohldefiniert.

$$A = \underbrace{LD^{\frac{1}{2}}}_{\tilde{L}} \underbrace{D^{\frac{1}{2}}L^T}_{\tilde{L}^T} = \tilde{L}\tilde{L}^T$$

Dies ist die „Cholesky-Zerlegung“ von A.
Diese ist in $\frac{n^3}{3} + \mathcal{O}(n^2)$ Operationen berechenbar.

5.16 Nichtreguläre Systeme

Es sei $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ sowie $\operatorname{Rang}(A)$ sei beliebig. „$Ax = b$“ hat keine, ∞-viele oder genau eine Lösung.
Einige Grundbegriffe aus LA:

- $A : \mathbb{R}^n \to \mathbb{R}^m$
- $\mathrm{Im}(A) = \{y \in \mathbb{R}^m : y = Ax \text{ für ein } x \in \mathbb{R}^n\}$
- $\ker(A) = \{x \in \mathbb{R}^n : Ax = 0\}$
- $\mathrm{Rang}(A) = \dim(\mathrm{Im}(A)) = \mathrm{Rang}(A^T) = \dim(\mathrm{Im}(A^T))$
- orthogonales Komplement:
  $\mathrm{Im}(A)^\perp = \{y \in \mathbb{R}^m : (y, y')_2 = 0 \;\; \forall\, y' \in \mathrm{Im}(A)\}$
  Es gilt:

$$\begin{aligned} (y, y')_2 = 0 \qquad & y' \in \mathrm{Im}(A) \\ \Leftrightarrow (y, Ax)_2 = 0 \qquad & \forall x \in \mathbb{R}^n \\ \Leftrightarrow (A^T y)^T x = 0 \qquad & \forall x \in \mathbb{R}^n \text{ (geht für alle } x \text{ nur, wenn } A^T y = 0) \\ \Leftrightarrow y \in \ker(A^T) & \end{aligned}$$

Damit: $\boxed{\mathrm{Im}(A)^\perp = \ker(A^T)}$

Satz 5.17 (Least Squares Lösung)

(a) Es existiert ein $\overline{x} \in \mathbb{R}^n$ sodass

$$\|A\overline{x} - b\|_2 = \min_{x \in \mathbb{R}^n} \|Ax - b\|_2$$

(b) Diese Bedingung ist äquivalent dazu, dass $\overline{x}$ eine Lösung von

$$A^T A\overline{x} = A^T b$$

der sogenannten „Normalengleichung".

(c) Falls der $\mathrm{Rang}(A) = n$ (dann ist $m \geq n$) ist $\overline{x}$ eindeutig bestimmt, sonst hat jede weitere Lösung die Form $\overline{x} + y$ mit $y \in \ker(A)$

Beweis.
(b) $\Rightarrow$ (a)
$\overline{x}$ sei Lösung der Normalengleichung. Für beliebiges $x \in \mathbb{R}^n$ gilt dann:

$$\begin{aligned} \|Ax - b\|_2^2 &= \|A(x - \overline{x} + \overline{x}) - b\|_2^2 \\ &= (A\overline{x} - b + A(x - \overline{x}), A\overline{x} - b + A(x - \overline{x}))_2 \\ &= (A\overline{x} - b, A\overline{x} - b)_2 + (A(x - \overline{x}), A(x - \overline{x})) \quad + 2(\underbrace{A\overline{x} - b}_{\substack{\in \mathrm{Im}(A)^\perp \\ = \ker(A^T) \text{ da} \\ A^T(A\overline{x} - b) = A^T A\overline{x} - A^T b \overset{\text{Vorr.}}{=} 0}}, \underbrace{A(x - \overline{x})}_{\in \mathrm{Im}(A)})_2 \\ &= \|A\overline{x} - b\|_2^2 + \|\underbrace{A(x - \overline{x})}_{\geq 0}\|_2^2 \\ &\geq \|A\overline{x} - b\|_2^2 \end{aligned}$$

(a) $\Rightarrow$ (b)

$$F(x) = \|Ax - b\|_2^2$$

Notwendige Bedingung für ein Minimum von F:

$$\begin{aligned}
\frac{\partial F}{\partial \frac{x}{\alpha}}(\overline{x}) &= 0 \qquad \forall 1 \le k \le n \Leftrightarrow \nabla F(\overline{x}) = 0 \\
\frac{\partial F}{\partial x_k}(\overline{x}) &= \frac{\partial}{\partial x_k}(Ax-b, Ax-b)_2\Big|_{x=\overline{x}} \\
&= \frac{\partial}{\partial x_k}\left(\sum_{i=1}^{m}\Big[\underbrace{\Big(\sum_{j=1}^{n} a_{ij}x_j\Big)-b}_{(Ax-b)_i}\Big]^2\right)\Bigg|_{x=\overline{x}} \\
&= 2\sum_{i=1}^{m}(A^T)_{k_i}\Big[\underbrace{\Big(\sum_{j=1}^{n} a_{ij}x_j\Big)-b}_{(A\overline{x}-b)_i}\Big] \\
&= 2(A^T(A\overline{x}-b))_k \\
\nabla F(\overline{x}) &= 2A^T(A\overline{x}-b) \overset{!}{=} 0 \\
&\Leftrightarrow \boxed{A^TA\overline{x} = A^Tb}
\end{aligned}$$

c) Lösbarkeit er Normalengleichung

$$\mathbb{R}^m = \mathrm{Im}(A) \oplus \mathrm{Im}(A)^\perp$$

d.h. jedes $b \in \mathbb{R}^m$ besitzt *eindeutige* Zerlegung

$$b = s + r \text{ mit } s \in \mathrm{Im}(A) \qquad r \in \mathrm{Im}(A)^\perp = \ker(A^T)$$

Zu $s \in \mathrm{Im}(A)$ gibt es $\overline{x} \in \mathbb{R}^n$ mit $A\overline{x} = s$. Dafür gilt dann

$$A^TA\overline{x} = A^Ts = A^Ts + \underbrace{A^Tr}_{=0} = A^Tb$$

also löst $\overline{x}$ auch $A^TAx = A^Tb$

$$\mathrm{Rang}(A) = n \text{ (und damit } m \ge n)$$

Betrachte

$$A^T\underbrace{Ax}_{y} = 0 \Leftrightarrow A^Ty = 0 \wedge y = Ax$$

Damit ist $y \in \mathrm{Im}(A) \wedge y \in \ker(A^T) = \mathrm{Im}(A)^\perp$. Das geht nur für $y = 0$.

$\mathrm{Rang}(A) < n$:
Sei x_1 eine weitere Lösung der Normalengleichung (also $A^TAx_1 = A^Tb$), dann gilt:

$$b = \underbrace{Ax_1}_{\in \mathrm{Im}(A)} + (\underbrace{b - Ax_1}_{\substack{\in \ker(A^T),\text{ da } A^T(b-Ax_1) \\ = A^tb - A^TAx_1 = 0}})$$

Zerlegung von b in $\mathrm{Im}(A) \oplus \mathrm{Im}(A)^\perp$ ist eindeutig, also $Ax_1 = A\overline{x}$ und dann $(Ax_1 - \overline{x}) = 0$ also $x_1 - \overline{x}$ im $\ker(A)$ □

6 Interpolation und Approximation

6.1 Einführung

Warum: Darstellung von Funktionen im Rechner
Anwendungen:

- Rekonstruktion eines funktionalen Zusammenhangs aus „gemessenen“ Funktionswerten
- Teuer auszuwertende Funktionen „effizienter“ auswerten
- Darstellung von Fonts (Zeichensätze), Körpern im Rechner. Voraussetzung für Simulationen und Visualisierung
- Lösen von Differential- und Integralgleichungen.
- Datenkompression

Wir beschränken uns auf Funktionen in einer Variablen, z.B.

$$f \in C^r([a,b])$$

Mögliche Teilmengen von Funktionen.

(a) $p(x) = a_0 + a_1 x + \cdots + a_n x^n$ (Polynome)

(b) $r(x) = \frac{a_0 + a_1 x + \cdots + a_n x^n}{b_0 + b_1 x + \cdots + b_n x^n}$ (rationale Funktionen)

(c) $t(x) = \frac{1}{2} a_0 + \sum_{k=1}^{n} (a_k \cos(kx) + b_k \sin(kx)$ (trigonometrische Funktionen)

(d) $e(x) = \sum_{k=1}^{n} a_k \exp(b_k x)$ (Exponentialsumme)

Grundaufgabe der Approximation

Gegeben: eine Menge von Funktionen P (siehe oben) sowie eine Funktion f (z.B. $f \in C^r([a,b])$).
Finde $g \in P$ sodass der Fehler $f - g$ in geeigneter WEise minimal ist.
Beispiele:

(a) $\sqrt{\int_a^b (f-g)^2 dx} \to \min$

(b) $\max_{a \le x \le b} |f(x) - g(x)| \to \min$

(c) $\max_{i=0,\cdots,n} |f(x_i) - g(x_i)| \to \min$ für $a \le x_i \le b \quad i = 0, \cdots, n$

Man spricht von *Interpolation*, falls g durch

$$g(x_i) = y_i := f(x_i) \qquad i = 0, \cdots, n$$

festgelegt wird.

6.2 Polynominterpolation

$$P_n := \{p(x) = \sum_{i=0}^{n} a_i x^i \mid a_i \in' R\}$$

Menge der Polynome über $\mathbb{R}$ vom Grad kleiner gleich $n \in \mathbb{N}$.
P_n ist ein $n+1$-dimensionaler Vektorraum. Die Monome $1, x, x^2, \ldots, x^n$ bilden eine Basis von P_n zu gegebenen $n+1$ paarweise verschiedenen Stützstellen $x_0, x_1, \cdots, x_n$ ist die Interpolationsaufgabe

$$p \in P_n : \quad p(x_i) = y_i \qquad i = 0, \cdots, n$$

$$\underbrace{\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix}}_{\substack{V[x_0,\cdots,x_n] \\ \text{„Vandermonde Matrix“}}} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$$

- $x_i = x_j \Rightarrow$ Zeile $i \Rightarrow=$ Zeile $j \Rightarrow$ Matrix singulär!
- Für paarweise verschiedene x_i ist $V[x_0, \cdots, x_n]$ regulär! (siehe unten)
- V ist schlecht konditioniert für wachsendes n, $\text{cond}_\infty(V) \approx 10^n$

Lagrange-Interpolation

Definition 6.1
Zu den paarweise verschiedenen Stützstellen $x_0, \cdots, x_n$ definiere die „Lagrange-Polynome“

$$L_i^{(n)}(x) = \prod_{\substack{j=0 \\ j\neq i}} \frac{x - x_j}{x_i - x_j} \qquad i = 0, \cdots, n$$

Diese Polynome haben folgende Eigenschaften:

(a) $L_i^{(n)}$ hat Grad n, klar da: $\prod_{\substack{j=0 \\ i\neq j}}^{n} (x - x_j) = x^n + \cdots$

(b) Es gilt:

$$L_i^{(n)}(x_k) = \delta_{ik} = \begin{cases} 1 & i = k \\ 0 & \text{sonst} \end{cases}$$

$$i = k : L_i^{(n)}(x_i) \prod_{\substack{j=0 \\ i\neq j}}^{n} = \underbrace{\frac{x_i - x_j}{x_i - x_j}}_{=1} = 1$$

$$i \neq k : \text{Im Produkt} \prod_{\substack{j=0 \\ i\neq j}}^{n} (x_k - x_j) \text{ kommt } j = k \neq i \text{ vor, d.h. } (x_k - x_k) = 0$$

(c) Die $L_i^{(n)}$ bilden eine Basis von P_n. $L_i^{(n)}(x)$ kann nicht als Linearkombination

$$\sum_{\substack{j=0\\ i\neq j}}^{n} \alpha_j h_j^{(n)}$$

gebildet werden, da

$$L_i^{(n)}(x_i) = 1 \text{ und } L_j^{(n)}(x_j) = 0 \quad j \neq i$$

Lösen der Interpolationsaufgabe

$$p(x) = \sum_{i=0}^{n} y_i L_i^{(n)}(x)$$

$$\text{denn } p(x_k) = \sum_{i=0}^{n} y_i \underbrace{L_i^{(n)}(x_k)}_{\delta_{ik}} = y_k \cdot 1$$

$$\sum_{i=0}^{n} \alpha_i \underbrace{L_i^{(n)}(x_k)}_{(A)_{ki}=\delta_{ik}} \stackrel{!}{=} y_k \qquad \Rightarrow A = I!$$

Satz 6.2
Zu gegebenen paarweise verschiedenen Stützstellen $x_0, \cdots, x_n$ gibt es genau ein Polynom vom Grad n mit

$$p(x_i) = y_i \qquad i = 0, \cdots, n \quad y_i \in \mathbb{R}$$

Beweis.

(I) $p = \sum_{i=0}^{n} v_i L_i^{(n)}(x)$ interpoliert die gegebenen Werte. Die L_i sind eine Basis.

(II) Existenz ist klar mittels Lagrange-Polynome.

Angenommen es gibt zwei Polynome p_1, p_2 mit $p_1 \neq p_2$ aber $p_1(x_i) = p_2(x_i) = y_i \quad i = 0, \cdots, n$.

Dann ist $p_1 - p_2 =: p \in P_n$ mit $p(x_i) = 0 \quad i = 0, \cdots, n$ jat also $n+1$ Nullstellen $\Rightarrow p \equiv 0$

Satz von Rolle:

$u(x)$ sei auf $[a, b]$ stetig und (a, b) differenzierbar sowie $u(a) = u(b) = 0$. Dann existiert mindestens ein $x \in (a, b)$ mit $u'(x) = 0$

Ang. $p \not\equiv$ dann ist also $p = \alpha_m x^m + \cdots$ mit $0 \leq m \leq n \quad \alpha_m \neq 0$

$$\begin{array}{rclcll}
p & = & p^{(0)} & = & \alpha_m x^m + \cdots & \text{hat } n+1 \text{ Nullstellen} \\
p' & = & p^{(1)} & = & \alpha_m m x^{m-1} + \cdots & \text{hat } n \text{ Nullstellen (Satz v. Rolle)} \\
p'' & = & p^{(2)} & = & \alpha_m m(m-1) x^{m-2} + \cdots & \text{hat } n-1 \text{ Nullstellen} \\
 & & & & \vdots & \\
 & & p^{(m)} & = & \alpha_m m! & \text{hat } n+1-m \geq 1 \text{ Nullstellen, da } m \leq n
\end{array}$$

$\Rightarrow \alpha_m = 0$ ` zur Annahme!

(III) (Skript) Beweise erst Eindeutigkeit über Satz von Rolle.

Existenz: Koeffizienten von p sind Lösungen des LGS $Va = y$

V muss regulär sein wegen Endeutigkeit (Setze $y = 0$, dann ` zur Eindeutigkeit des I-Polynoms)

□

Newton-Polynome

Nachteil der Lagrange-Polynome:

- Hinzufügen einer neuen Stützstelle ändert alle bisherigen Basispolynome

Besser ist die Newton-Darstellung:

$$N_0(x) = 1 \qquad i = 0, \cdots, n \qquad N_i(x) = \prod_{j=0}^{i-1}(x - x_j)$$

(a) $N_i(x)$ ist Polynom vom Grad i

(b) $N_0, \cdots, N_r$ bilden eine Basis von P_n

(c) $N_i(x_k) = 0$ für alle $k < i$

Gestaffelte Berechnung:

$$p(x_k) = \sum_{i=0}^{n} a_i N_i(x_k) = \sum_{i=0}^{k} a_i N_i(x) \overset{!}{=} y_i \qquad i = 0, \cdots, n$$

$$\begin{array}{lll} k = 0 & a_0 & = & y_0 \\ k > 0 & a_k & = & \left(y_k - \sum_{i=0}^{k-1} a_i N_i(x_k)\right) \Big/ N_k(x + k) \end{array}$$

(= vorwärts einsetzen zur Lösung eines Dreiecksystems)

Wiederholung 5

$P_n = \{p(x) = \sum_{i=0}^{n} a_i x^i \mid a_i \in \mathbb{R}\}$

$p(x_i) = y_i \qquad i = 0, \cdots, n \qquad x_i \neq x_j \qquad i \neq j$

Monome: $1, x, x^2, \cdots, x^n$

Lagrange: $L_i^{(n)}(x) = \prod_{\substack{j=0 \\ j \neq i}} \frac{(x - x_j)}{x_i - x_j} \Rightarrow p(x) = \sum_{i=0}^{n} y_i L_i^{(n)}(x)$

Newton: $N_0(x) = 1 \qquad N_j(x) = \prod_{j=0}^{i-1}(x - x_j) = (x - x_{i-j}) N_{i-1}(x)$

Satz 6.3 (Dividierte Differenzen)

Man definiert rekursiv die sogenannten „dividierten Differenzen".

$$\forall i = 0, \cdots, n : y[x_i] := y_i$$

$$\forall k = 1, \cdots, n - i : y[\underbrace{x_i, \cdots x_{i+k}}_{k+1}] := \frac{y[\overbrace{x_{i+1}, \cdots, x_{i+k}}^{k \text{ Argumente}}] - y[\overbrace{x_i, \cdots, x_{i+k-1}}^{k \text{ Argumente}}]}{x_{i+k} - x_i}$$

Dann gilt:

$$p(x) = \sum_{i=0}^{n} \underbrace{y[x_0, \cdots, x_i]}_{=a_i} N_i(x)$$

Beweis: [Rannacher, Satz 2.2]

Praktische Anwendung

$$\begin{array}{lllllllllll}
 & & \overbrace{}^{a_0} & & \overbrace{}^{a_1} & & \overbrace{}^{a_2} & & \overbrace{}^{a_{n-1}} & & \overbrace{}^{a_n} \\
y_0 & =: & y[x_0] & \nearrow & y[x_0,x_1] & \nearrow & y[x_0,x_1,x_2] & \cdots & y[x_0,\cdots,x_n-1] & \nearrow & y[x_0,\cdots,x_n] \\
y_1 & =: & y[x_1] & \nearrow & y[x_1,x_2] & \nearrow & y[x_1,x_2,x_3] & \cdots & y[x_1,\cdots,x_n] & & \\
y_2 & =: & y[x_2] & \nearrow & y[x_2,x_3] & \nearrow & \vdots & & & & \\
\vdots & & & & & \nearrow & y[x_{n-2}x_{n-1},x_n] & & & & \\
\vdots & & & \nearrow & y[x_{n-1},x_n] & & & & & & \\
y_n & =: & y[x_n] & & & & & & & &
\end{array}$$

$$\Rightarrow \overbrace{y[x_o, \cdots, x_{n+1}]}^{a_{n+1}}$$

Interpolationsfehler

$y_i = f(x_i) \qquad i = 0, \cdots, n$
$p(x)$ sei Interpolationspolynom

$$e(x) = f(x) - p(x) = \begin{cases} 0 & x = x_i \\ ? & x \neq x_i \end{cases} \text{„Interpolationsfehler“}$$

Satz 6.4 Interpolationsfehler
Sei $f(x) \in C^{n+1}$ auf $[a, b]$ und es sei

$$a \leq x_0 < x_1 < \cdots < x_n \leq b$$

Dann gibt es zu jedem $x \in [a, b]$ ein $\xi_x \in \overline{(x_0, \cdots, x_n, x)}$ (=kleinstes Intervall, welches alle Punkte enthält) so dass

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{j=0}^{n} (x - x_j)$$

Beweis. (I) $x \in \{x_0, \cdots, x_n\} \Rightarrow \prod_{j=0}^{n}(x - x_j) = 0$. Wähle ξ_x beliebig

(II) $x \neq \{x_0, \cdots, x_n\}, x \in [a, b]$ Definiere Hilfsfunktion:

$$F_x(t) = f(t) - p(t) - \underbrace{\frac{f(x) - p(x)}{l(x)}}_{\substack{\text{hängt nur} \\ \text{von } x \text{ ab}}} l(t) \text{ mit } l(t) = \prod_{j=0}^{n}(t - x_j)$$

$F_x(t)$ hat die $n + 2$ Nullstellen $\{\underbrace{x_0, \cdots, x_n}_{n+1}, x\}$

$$i = 0, \cdots, n : F_x(x_i) = \underbrace{f(x_i) - p(x_i)}_{=0} - \frac{f(x) - p(x)}{l(x)} \underbrace{l(x_i)}_{=0} = 0$$

$$x : F_x(x) = f(x) - p(x) - \frac{f(x) - p(x)}{\cancel{l(x)}} \cancel{l(x_i)}$$

Satz von Rolle

$F(t)$ hat $n+2$ Nullstellen (mindestens). Differenzieren nach $t \Rightarrow$
$F_x^{(1)}(t)$ hat $n+1$ Nullstellen. $\Rightarrow \cdots$
$F_x^{(n+1)}(t)$ hat 1 Nullstelle

Zusätzlich gilt: Diese Nullstellen sind alle in $\overline{(x_0, \cdots, x_n, x)}$. Nenne eine der Nullstellen von $F_x^{(n+1)}$ nun xi_x. Für diese gilt:

$$F_x^{(n+1)}(\xi_x) = f^{(n+1)}(\xi_x) - \underbrace{p^{(n+1)}(\xi_x)}_{=0} - \frac{f(x) - p(x)}{l(x)} \underbrace{l^{(n+1)}(\xi_x)}_{=(n+1)!}$$

$$= f^{(n+1)}(\xi_x) - \frac{f(x) - p(x)}{l(x)}(n+1)! \stackrel{!}{=} 0$$

□

Diskussion

Äquidistante Stützstellen. Was passiert bei halbieren des Abstandes h? (Abbildung: Zahlenstrahl von x_0 bis x_n wobei h zwei Marker umfasst)
Betrag bilden:

$$|f(x) - p(x)| \leq \frac{|f^{(n+1)}(\xi_x)|}{(n+1)!} \underbrace{\prod_{j=0}^{n} |x - x_j|}$$

(Abbildung: Zahlenstrahl durch x geteilt. Links davon l Punkte, rechts davon r Punkte.)

$$\prod_{j=0}^{n} |x - x_j| \leq 1h \cdot 2h \cdot lh \cdot 1h \cdot 2h \cdots rh] = h^{n+1} l! r!$$

$$|f(x) - p(x)| \leq |f^{(n+1)}(\xi_x)| \underbrace{\frac{l! r!}{(n+1)!}}_{\substack{\leq 1 \\ l+r=n+1}} h^{n+1}$$

$\Rightarrow$ sehr schnelle Konvergenz, *falls* $f^{(n+1)}$ nicht zu schnell wächst. Dies ist in der Praxis selten!

Runges Gegenbeispiel:

$f(x) = \frac{1}{1+x^a}$ in $[-5, 5]$ Rannacher: $|f^{(n+1)}(x)| \approx 2^n n! \, \mathcal{O}(\frac{1}{|x|^{n+2}})$

Bemerkung 6.5 Approximationssatz von Weierstrass
Jede Funktion $f \in C[a,b]$ kann beliebig gut gleichmäßig auf $[a,b]$ mit Polynomen approximiert werden.

- Kein Widerspruch, da der Satz nicht von Interpolation spricht
- nicht äquidistante Wahl der Stützstellen ist auch schon besser

Bemerkung 6.6
Methoden hoher Ordnung erfordern entsprechende Differenzierbarkeit

Konditionierung der Interpolationsaufgabe
$p(x,y)$ $\quad y = (y_0, \cdots, y_n)^T$ Ordinatenwerte

$$\begin{aligned}\frac{p(x;y+\Delta y)-p(x;y)}{p(x;y)} &= \left(\sum_{j=0}^{n}(y_i+\Delta y_i)L_i^{(n)}(x) - \sum_{i=0}^{j} y_i L_i^{(n)}(x)\right)\Big/ p(x;y) \\ &= \sum_{i=0}^{n} \underbrace{\frac{L_i^{(n)}(x)y_i}{p(x;y)}}_{\substack{\text{Verstärkungs-}\\ \text{faktor}}} \cdot \underbrace{\frac{\Delta y_i}{y_i}}_{\substack{\text{relativer}\\ \text{Eingabef.}}}\end{aligned}$$

6.3 Anwendungen der Polynominterpolation

Numerische Differentation
Problem: Berechne Ableitung einer tabelarisch gegebenen Funktion (oder einer als Programm gegebene Funktion). *Idee:* Erstelle Interpolationspolynom und leite dieses ab. *Zunächst:* Ableitungsordnung = Polynomgrad
Lagrange-Polynome:

$$L_i^{(n)} = \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{(x-x_j)}{(x_i-x_j)} = \underbrace{\left(\prod_{\substack{j=0\\ j\neq i}}^{n} \frac{1}{(x_i-x_j)}\right)}_{=:\lambda_i} x^n + \alpha_{n-1}x^{n+1}$$

n mal ableiten liefert

$$\frac{d^n}{dx^n} L_i^{(n)}(x) = \lambda_i n!$$

also

$$\frac{d^n}{dx^n} \underbrace{\left(\sum_{i=0}^{n} y_i L_i^{(n)}(x)\right)}_{p(x)} = n! \sum_{i=0}^{n} y_i \lambda_i$$

Satz 6.7
Sei $f \in C^n[a,b]$ und $a = x_0 < x_1 < \cdots < x_n = b$. Dann gibt es ein $\xi \in (a,b)$ so dass

$$f^{(n)}(\xi) = n! \sum_{i=0}^{n} y_u \lambda_i$$

Beweisskizze: $g(x) : f(x) - g(x)$ (g hat $n+1$ Nullstellen n-mal Anwenden des Satz von Rolle
Konkret: Verwende äquidistante Stützstellen. Damit

$$\lambda_i = \frac{1}{\underbrace{(x_i - x_0)(x_i - x_1)\cdots(x_i - x_{i-1})}_{i \text{ Faktoren positiv}}\underbrace{(x_i - x_{i+1})\cdots(x_i - x_n)}_{i-1 \text{ Faktoren negativ}}}$$
$$= \frac{1}{h^n i!(n-i)!(-1)^{n-i}} = \frac{(-1)^{n-i}}{h^n n!}\binom{n}{i} \text{ (Binominal-Koeffizient)}$$

und damit

$$f^{(n)}(x) \approx \frac{d^n}{dx^n}\left(\sum_{i=0}^{n} y_i L_i^{(n)}\right) = \frac{1}{h^n}\sum_{i=0}^{n}(-1)^{n-i}\binom{n}{i}y_i$$

$$f^{(1)}(x) = \frac{y_1 - y_0}{h}$$
$$f^{(2)}(x) = \frac{y_2 - 2y_1 + y_0}{h^2}$$
$$f^{(3)}(x) = \frac{y_3 - 3y_2 + 3y_1 - y_0}{h^3}$$

Bis jetzt:

n-te Ableitung aus I-Polynom vom Grad n

Alternativ:

m-te Ableitung aus I-Polynom Grad $n > m$ Näherungswert hängt dann von Auswertestelle x ab!

Beispiel:

$m = 1, n = 2$ äquidistant: $x_0, x_1 = x_0 + h; x_2 = x_0 + 2h$

$$\Rightarrow p'(x_1) = \frac{y_2 - y_0}{2h} \text{ „zentraler Differenzenquotient“}$$

Taylorreihenentwicklung:

$$\begin{aligned} f'(x) &= \tfrac{f(x+h)-f(x-h)}{2h} + \mathcal{O}(h^2) && \text{für } f \in C^3 \\ f''(x) &= \tfrac{f(x+h)-2f(x)+f(x-h)}{h^2} + \mathcal{O}(h^2) && \text{für } f \in C^4 \end{aligned}$$

Extrapolatin zum Limes

Eine Größe $a(h)$ sei im Rechner für $h > 0$ berechenbar, *nicht jedoch für* $h = 0$ Man möchte

$$a(0) = \lim_{k \to 0} a(h)$$

ausrechnen.

Beispiel 6.9

(a)

$$a(0) = \lim_{x\to 0} \frac{\cos(x) - 1}{\sin(x)} \quad (= 0, \text{ nach L'Hopital})$$

(b) Numerische Differentation

$$f'(x) = \lim_{h\to 0} \underbrace{\frac{f(x+h) - f(x)}{h}}_{a(h)}$$

geteilt durch ∅ Auslöschung

(c) Numerische Integration:

$$\int_a^b f(x)dx = \lim_{N\to\infty} \underbrace{\sum_{i=1}^{N} \frac{1}{N} f\left(\frac{1}{2}(\frac{i-1}{N} + \frac{i}{N})\right)}_{=:a(h)}$$

$h \to 0 (\hat{=} N \to \infty)$ nicht möglich wegen Aufwand!

(d) Anfangswertproblem:

$$y'(t) = f(t, y(t)) \qquad y(0) - y_0 \qquad y_n = y_{n+1} + hf(t, y_{n-1}) \qquad h = TN$$

$$\boxed{y(T) \approx y_N}$$

$h \to 0$ nicht möglich wegen Aufwand!

Idee:

Zu $h_0 > h_1 > \cdots > h_n > 0$ bestimme Interpolationspolynom vom Grad n sodass

$$p(h_i) = a(h_i) \qquad i = 0, \cdots, n$$

und berechne $a(0) \approx p(0)$

Beispiel 6.10

$$\begin{array}{ll} & a(h) = \frac{\cos(h)-1}{\sin(h)} \\ h_0 = \frac{1}{8} & a(h_0) = -6.258151 \cdot 10^{-2} \\ h_1 = \frac{1}{16} & a(h_1) = -3.126018 \cdot 10^{-2} \\ h_2 = \frac{1}{32} & a(h_2) = -1.562627 \cdot 10^{-2} \end{array}$$

Extrapolation:

$$a(0) \approx_2 (0) = -1.02 \cdot 10^{-5}$$

Warum ist das so gut?

Sei $a(x)$ $n+1$ mal stetig differenzierbar in einer genügend großen Umgebung von $x=0$. Dann gibt es zu jedem $h>0$ (in dieser Umgebung) ein $\xi_h \in [0,h]$ so dass: (Taylorreihe):

$$a(h) = a(0+h) = \underbrace{a(0) + a'(0)h + \frac{a''(0)}{2}h^2 + \cdots + \frac{a^{(n)}}{n!}h^n}_{\substack{\text{Polynom in h vom Grad n} \\ \text{Koeffizienten hängen NICHT von h ab}}} + \frac{a^{(n+1)}(\xi_h)}{(n+1)!}h^{n+1}$$

Idee: Für verschiedene h_i bilde Linearkombinationen:

$$\sum_{i=0}^{n} c_i a(h_i) = \sum_{i=0}^{n} c_i \left(\sum_{j=0}^{n} a_j h_i^j \right) + \underbrace{\sum_{i=0}^{n} c_i \frac{a^{(n+1)}(\xi_{h_i})}{(n+1)!} h_i^{n+1}} = \sum_{j=0}^{n} a_j \left(\underbrace{\sum_{i=0}^{n} c_i h_i^j} \right) + \text{Fehler} = a(0) + \text{Fehler}$$

wähle c_i sodass

$$\sum_{i=0}^{n} c_i h_i^j = \begin{cases} 0 & j=1 \\ 1 & \text{sonst} \end{cases}$$

Bestimmungsgleichung für Koeffizienten:

$$V^T c = e^{(0)} \qquad e^{(0)} = (1, 0, \cdots, 0)^T$$

$$V = V[h_0, \cdots, h_n] \text{ Vandemonde Matrix}$$

Auswertung:

$$A = \sum_{i=0}^{n} c_i \underbrace{a(h_i)}_{=y_i} = (\underbrace{V^{-T} e^{(0)}}_{=c}) \cdot y = (e^{(0)})^T V^{-1} y$$

$$y = (a(h_0), \cdots, a(h_n))$$

Für das I-Polynom aus der Extrapolation gilt

$$o(h_i) = \sum_{j=0}^{n} b_j h_i^j \stackrel{!}{=} a(h_i) \qquad i = 0, \cdots, n$$

daraus folgt

$$\rightarrow Vb = y \qquad \text{für Koeffizienten in der Monombasis}$$

Auswertung an der Stelle 0:

$$p(0) = b_0 = (e^{(0)})^T b = \boxed{(e^{(0)})^T V^{-1} y}$$

Für den Fehler gilt:

$$|A - a(0)| \underbrace{\leq}_{h_i = hr^i \text{ für } r<1} \left\| V^{-T} \right\|_\infty \left| a^{(n+1)}(\xi_{\max}) \right| \frac{h^{n+1}}{(n+1)!} (1 + r^{n+1})$$

Es geht sogar noch bessser!

Beispiel:

$$f''(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} + \text{Fehler}$$

$$f(x+h) = f(x) + hf^{(1)}(x) + \tfrac{h^2}{2}f^{(2)}(x) + \tfrac{h^3}{3!}f^{(3)}(x) + \cdots + \tfrac{h^{2n+2}}{(2n+2)!}f^{(2n+1)}(x) + \cdots + \tfrac{h^{2n+4}}{(2n+4)!}f^{(2n+4)}(\xi_+)$$
$$f(x+h) = f(x) - hf^{(1)}(x) + \tfrac{h^2}{2}f^{(2)}(x) - \tfrac{h^3}{3!}f^{(3)}(x) + \cdots + \tfrac{h^{2n+2}}{(2n+2)!}f^{(2n+1)}(x) + \cdots + \underbrace{\tfrac{h^{2n+4}}{(2n+4)!}f^{(2n+4)}(\xi_-)}_{(\xi\in[x-h,\cdots,x])}$$

$$f(x+h) + f(x-h) = 2f(x) + \cdots + h^2 f^{(2)}(x) + 2\tfrac{h^4}{4!}f^{(4)}(x) + \cdots$$

$$\begin{aligned} a(h) &= \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} \\ &= \underbrace{\underbrace{f^{(2)}(x)}_{„a(0)“} + \frac{2h^2}{4!}f^{(4)}(x) + \cdots + \frac{2h^{2n}}{(2n+2)!}f^{(2n+2)}(x)} + \frac{h^{2n+2}}{(2n+4)!}\left[f^{(2n+4)}(\xi_+ + f^{(2n+4)}(\xi_-)\right] \\ &= P(h^2) + \mathcal{O}(h^{2n+2}) \qquad \text{Praktisch: } P(h_i^2) = a(h_i) \end{aligned}$$

Alle ungeraden Terme in der Fehlerdarstellung fallen „von selbst“ weg $\Rightarrow$ „doppelt“ Konvergenzordnung bei Extrapolation.

6.4 Bernsteinpolynome zur Kurvendarstellung

Nun Approximation statt Interpolation. Speziell geeignet für Kurven

$$u(t) : [a,b] \to \mathbb{R}^d (d = 2,3)$$

Definition 6.11 (Bernstein-Polynome)
Die Polynome

$$\beta_i^{(n)}(t) = \binom{n}{i}(1-t)^{n-i}t^i \qquad i = 0,\cdots,n$$

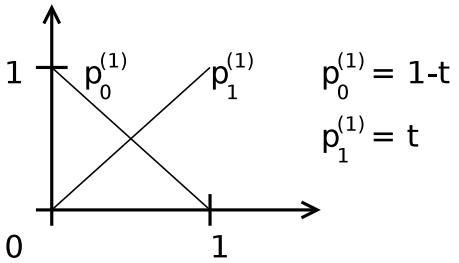
vom Grad n heißen Bernstein-Polynome auf $[0,1]$

Für allgemeine Intervalle benutze

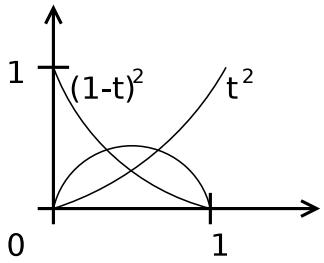
$$\varphi : [a,b] \to [0,1] \qquad \varphi(u) = \frac{u-a}{b-a}$$

$$\begin{aligned} \beta_{i,[a,b]}^{(n)} &= \beta_i^{(n)}(\varphi(u)) \\ &= \binom{n}{i}\left(1 - \frac{n-a}{b-a}\right)^{n-i}\left(\frac{n-a}{b-a}\right)^i \\ &= \binom{n}{i}\left(\frac{1}{b-a}\right)^n (b-n)^{n-i}(n-a)^i \end{aligned}$$

- $n = 1$:



- $n = 2$:



Satz 6.12 (Eigenschaften der Bernstein-Polynome)

(a) $\sum_{i=0}^{n} \beta_i^{(n)}(t) = 1$

$$1 = (1-t+t)^n \overset{\text{Binomische-Formel}}{=} \sum_{i=0}^{n} \underbrace{\binom{n}{i}(1-t)^{n-i}t^i}_{=\beta_i^{(n)}(t)}$$

(b) $t = 0$ ist i-fache Nullstelle von $\beta_i^{(n)}$

(c) $t = 1$ ist $n-i$-fache Nullstelle von $\beta_i^{(n)}$

(d) Symmetrie: $\beta_i^{(n)}(t) = \beta_{n-i}^{(n)}(1-t)$

(e) Positivität:
$0 \le \beta_i^{(n)}(t) \le 1 \qquad \text{für} t \in [0,1]$
$0 < \beta_i^{(n)}(t) < 1 \qquad \text{für} t \in (0,1)$

(f) $\beta_i^{(n)}$ hat in $[0,1]$ genau ein Maximum im Punkt $\frac{i}{n}$

$$\begin{aligned}\frac{d}{dt}\beta_i^{(n)}(t) &= \binom{n}{i}\Big[-(n-i)(1-t)^{n-i-1}t^i + (1-t)^{n-i}it^{i-1}\Big] \\ &= \binom{n}{i}(1-t)^{n-i-1}t^{i-n}\Big(\underbrace{(1-t)i-(n-i)t}_{\substack{i-\cancel{it}-nt+\cancel{it} \\ =i-nt=0\Rightarrow t=i/n}}\Big)\end{aligned}$$

(g) Die $\left\{\beta_i^{(n)}\right\}_{i=0}^{n}$ sind linear unabhängig, bilden also eine Basis von P_n

$$\text{ZZ}: \sum_{i=0}^{n} b_i\beta_i^{(n)}(t) = 0(\forall t) \Rightarrow b_i = 0 \text{ für } i = 0, \cdots, n$$

Betrachte Ableitungen:

$$\frac{d^j}{dt^j}\sum_{i=0}^{n} b_i\beta_i^{(n)}(t) = \sum_{i=0}^{n} b_i\frac{d}{dt^j}\beta_i^{(n)}(t)$$

$j = 0, t = 0$: nur $\beta_0^{(n)} \neq 0 \Rightarrow b_0 = 0$

$j = 1, t = 0$: nur $\frac{d}{dt}\beta_1^{(n)}(0) \neq 0 \Rightarrow b_1 = 0$ (usw.)

(h) Die B-Polynome erlauben folgende rekursive Darstellung über den Grad n:

$$\beta_i^{(n)}(t) = \begin{cases} (1-t)\beta_0^{(n-1)}(t) & i = 0 \\ t\beta_{i-1}^{(n-1)}(t) + (1-t)\beta_i^{(n-1)}(t) & 0 < i < n \\ t \cdot \beta_{n-1}^{(n-1)}(t) & i = n \end{cases}$$

Beweis
Für die Fälle $i = \{0, n\}$. Einsetzen:

$$t\binom{n-1}{i-1}\underbrace{(1-t)^{n-i}t^i} + \binom{n-1}{i}\underbrace{(1-t)^{n-i}t^i} = \left[\binom{n}{i}\binom{n-1}{i-1} + \binom{n-1}{i}\right](1-t)^{n-i}t^i$$

□

(i) Für die erste Ableitung gilt:

$$\frac{d}{dt}\beta_i^{(n)}(t) = \begin{cases} -n\beta_0^{(n-1)}(t) & i = 0 \\ n[\beta_{i-1}^{(n-1)}(t) - \beta_i^{(n-i)}(t)] & 0 < i < n \\ n\beta_{n-1}^{(n-1)}(t) & i = n \end{cases}$$

Definition 6.13 (Bezier-Kurve)
Für gegebene Punke $b_0, \cdots, b_n \in \mathbb{R}^d$ heißt das vektorwertige Polynom

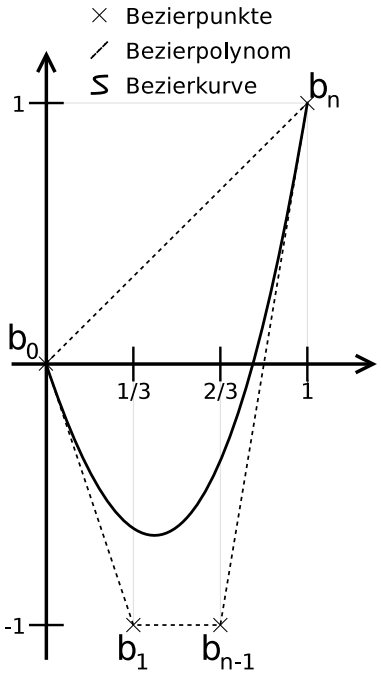
$$B(t) = \sum_{i=0}^{n} b_i \beta_i^{(n)}(t)$$

heißt *Bezier-Kurve*.

Beispiel 6.14

$$b_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, b_1 = \begin{pmatrix} \frac{1}{3} \\ -1 \end{pmatrix}, b_2 = \begin{pmatrix} \frac{2}{3} \\ -1 \end{pmatrix}, b_3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$B(t) = \begin{pmatrix} t \\ t^3 + 3t^2 - 3t \end{pmatrix}$$



konvexe Hülle:

$$b_0(1-t) + b_1 t \qquad \sum c_i b_i \text{ mit } 0 \le c_i \le 1 \qquad \sum c_i = 1$$

- $\Rightarrow$ Bezier-Kurve liegt immer innerhalb des Bezier-Polynoms!
- $B(0) = b_0, B(1) = b_n$
- Die Ableitung (=Tangente an die Kurve!) hat in den Endpunkte die Richtung

$$(b_1 - b_0) \text{ für } t = 0 \text{ bzw. } (b_n - b_{n-1}) \text{ für } t = 1$$

6.5 Splines

Bis jetzt:

- # Stützstellen = Polynomgrad $n+1$
- Großer Polynomgrad = viele Stützstellen $\Rightarrow$ starke Abweichungen zwischen Stützstellen

Idee:

Stückweise Polynome niedrigen Grades

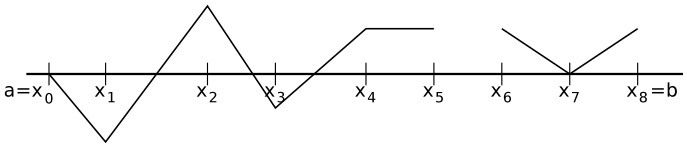
Definition 6.16
Sei $X = (x_0, x_1, \cdots, x_n)$ mit $a = x_0 < x_1 < x_n = b$ eine Zerlegung des Intervalles $[a, b]$ und sei $m \in \mathbb{N}$.

$$S^m(X) = \left\{ s \in C^{m-1}([a,b]), \quad S|_{x_i, x_{i+1}} \in P_M, \quad 0 \le i < n \right\}$$

heißt Spline-Raum vom Grad m über der Zerlegung X

Beispiel 6.17
$S^1(x)$

- $s \in S^1(x)$ ist Polynom vom Grad 1 auf jedem Teilintervall in $[x_1, x_{i-1}]$
- $S^1(x) \subset C^0([a,b])$ also $s \in S^1(x)$ stetig



Stetigkeit $\Rightarrow s \in S^1(x)$ ist eindeutig durch Werte in den Stützstellen beschrieben.

Kubische Splines

In der Praxis ist $S^3(x)$ sehr beliebt
$S^3(X)$ heißt Raum der kubischen Splines. Geschichte: „Straklatte“ zur Konstruktion glatter Kurven im Schniffs- und Flugzeugbau. Dünnes Balsaholz biegt sich unter, Energieminimierung

$$\underbrace{\int_a^b \frac{|y''(t)|^2}{1+|y'(t)|^2} x dy}_{\text{totale Krümmung}} \overset{|y'(t)| \ll 1}{\approx} \int_a^b |y''(t)|^2 dt \to \min$$

Konstruktion $S \in S^3(x)$. Die Funktion setzt sich aus n Polynomen zusammen.

$$S(X) = \begin{cases} P_i(x) & x \in [x_{i-1}, x_i), i \in \{1, \cdots, n\} \\ P_n(x) & \end{cases}$$

Bedingungen

(a) Interpolationsbedingung (Stetigkeit):

$$i = 1, \cdots, n-1: \quad \left.\begin{array}{r} P_i(x_{i-1}) = y_{i-1} \\ P_i(x_i) = y_i \end{array}\right\} \to 2n \text{ Bedingungen}$$

(b) Stetigkeit der ersten und zweiten Ableitung an den inneren Punkten

$$i = 1, \cdots, n-1: \quad \left.\begin{array}{r} P_i'(x_i) = P_{i-1}'(x_i) \\ P_i''(x_i) = P_{i+1}''(x_i) \end{array}\right\} \to 2(n-1) = 2n-2 \text{ Bedingungen}$$

Insgesamt: $4n-2$ Bedingungen

Pro Polynom P_i (vom Grad 3) hat man 4, also insgesamt $4n$ Freiheitsgrade. Die fehlenden zwei Bedingungen erhält man durch *Randbedingungen* an den Stellen x_0 und x_n. Dabei gibt es verschieden Varianten.

(c) Randbedingungen. *Eine* der folgenden Varianten:

(i)	Natürliche Randbedingungen:	$P_1''(x_0) = 0$	$P_n''(x_n) = 0$
(ii)	Hermite-Randbedingungen:	$P_1'(x_0) = f'(x_0)$	$P_n'(x_n) = f'(x_n)$
(iii)	Periodische Randbedingungen:	$P_1'(x_0) = P_n'(x_n)$	$P_1''(x_0) = P_n''(x_n)$

Wir behandeln im folgenden nur die natürlichen Randbedingungen.

Satz 6.18 (Berechnung kubischer Splines)
Wir schreiben Teilpolynome des Splines in der Form

$$P_i(x) = a_0^{(i)} + a_1^{(i)}(x-x_i) + a_2^{(i)}(x-x_i)^2 + a_3^{(i)}(x-x_i)^3 \qquad i = 1, \cdots, n$$

Die a_2 sind dann die Lösung des linearen Gleichungssystem der Dimension $n-1$:

$$h_i a_2^{(i-1)} + 2(h_i + h_{i+1}) a_2^{(i)} + h_{i+1} a_2^{(i+1)} = 3\left(\frac{y_{i+1}-y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i}\right) \tag{6.1}$$

wobei $a_2^{(0)} = a_2^{(n)} = 0$ (natürliche Randbedingung) und $h_i := x_i - x_{i-1}$. Die restlichen Koeffizienten ergeben sich zu:

$$a_0^{(i)} = y_i \tag{6.2}$$

$$a_1^{(i)} = \frac{y_i - y_{i-1}}{h_i} + \frac{h_i}{3}(2a_2^{(i)} + a_2^{(i-1)}) \tag{6.3}$$

$$a_3^{(i)} = \frac{a_2^{(i)} - a_2^{(i-1)}}{3h_i} \tag{6.4}$$

Beweis .

(i) Berechne Ableitungen der Teilpolynome

$$p_i'(x) = a_1^{(i)} + 2a_2^{(i)}(x - x_i) + 3a_3^{(i)}(x - x_i)^2 \tag{6.5}$$

$$p_i''(x) = 2a_2^{(i)} + 6a_3^{(i)}(x - x_i) \tag{6.6}$$

(ii) Interpolationsbedingung nutzen: Einsetzen von x_i, y_i:

$$y_i = p_i(x_i) = a_0^{(i)} \Rightarrow \boxed{a_0^{(i)} = y_i} \tag{6.7}$$

Das ist (6.2).

Einsetzen von x_{i-1}:

$$y_{i-1} = p_i(x_i - 1) = \underbrace{a_0}_{=y_i} + a_1$$

$$\Leftrightarrow y_{i-1} - y_i = -h_i a_1^{(i)} + h_i^2 a_2^{(i)} - h_i^3 a_3^{(i)} \tag{6.8}$$

(iii) Randbedingungen einsetzen. Wir behandeln nur natürliche:

$$0 = P_1''(x_0) = 2a_2^{(1)} - 6a_3^{(1)} h_1 \tag{6.9}$$

$$0 = P_n''(x_n) = 2a_2^{(n)} \Rightarrow \boxed{a_2^{(n)} = 0} \tag{6.10}$$

(iv) Stetigkeit der ersten Ableitung

$$P_i'(x_i) = P_{i+1}'(x_i) \qquad i = 1, \cdots, n-1$$

$$\Leftrightarrow a_1^{(i)} = a_1^{(i+1)} - 2a_2^{(i+1)} h_{i+1} + 3a_3^{(i+1)} h_{i+1}^2 \tag{6.11}$$

(v) Stetigkeit der zweiten Ableitung

$$P_i''(x_i) = P_{i+1}''(x_i) \qquad i = 1, \cdots, n-1$$

$$\Leftrightarrow 2_a 2^{(i)} = 2a_2^{(i+1)} - 6a_3^{(i+1)} h_{i+1} \tag{6.12}$$

(vi) Drücke a_3 durch $a_2^{(i)}$ aus, d.h. lösen (6.12) nach $a_3^{(i+1)}$ auf:

$$a_3^{(i+1)} = \frac{a_2^{(i+1)} - a_2^{(i)}}{3h_{i+1}} \qquad i = 1, \cdots, n-1$$

umnummerieren:

$$\Leftrightarrow \underbrace{\boxed{a_3^{(i)} = \frac{a_2^{(i)} - a_2^{(i+1)}}{3h_i}}}_{\text{Das ist 6.4}} \tag{6.13}$$

$i = 2, \cdots, n$ aus (v) $\qquad$ $i = 1$ aus (6.9) wenn $\boxed{a_2^{(0)} := 0}$ setzt, also $i = 1, \cdots, n$

(vii) $a_1^{(i)}$ durch $a_2^{(i)}$ ausdrücken. Dazu löse dazu (6.8) nach a_1 auf:

$$\begin{aligned} a_1^{(i)} &= \frac{y_i - y_{i-1}}{h_i} + h_i a_2^{(i)} + h_i a_2^{(i)} - h_i^2 a_3^{(i)} \qquad i = 1, \cdots, n \\ &= \frac{y_i - y_{i-1}}{h_i} + h_i a_2^{(i)} - h_i^2 \left(\frac{a_2^{(i)} - a_2^{(i+1)}}{3h_i} \right) \end{aligned}$$

$$\underbrace{\boxed{a_1^{(i)} = \frac{y_i - y_{i-1}}{h_i} + \frac{h_i}{3}\left(2a_2^{(i)} + a_2^{(i-1)}\right)}}_{\text{Das ist 6.3}} \qquad i = 1, \cdots, n$$

.

(viii) Nun setze $a_1^{(i)}$ und $a_3^{(i)}$ in die verbleibende Gleichung (6.11) ein:

$$\begin{aligned} &\frac{y_i - y_{i-1}}{h_i} + \frac{h_i}{3}(2a_2^{(i)} + a_2^{(i-1)}) = \frac{y_{i+1} - y_i}{h_{i+1}} + \frac{h_{i+1}}{3}(2a_2^{(i+1)} + a_2^{(i)}) \\ &= 2a_2^{(i+1)} h_{i+1} + 3h_{i+1}^2 \left(\frac{a_2^{(i+1)} - a_2^{(i)}}{3h_{i+1}} \right) \\ &= \frac{h_i}{3} a_2^{(i+1)} + a_2^{(i)} \left(\frac{2h_i}{3} - \frac{h+1}{3} + h_{i+1} \right) + a_2^{(i+1)} \left(-\frac{2h_{i+1}}{3} + 2h_{i+1} - h_{i+1} \right) \\ &= \frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i} \end{aligned}$$

$$\overset{\text{mal } 3}{\Leftrightarrow} h_i a_2^{(i+1)} + 2(h_i + h_{i+1}) a_2^{(i)} + b_{i+1} a_2^{(i+1)} = 3\left(\frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i} \right)$$

Und das ist (6.1). Beachte, dass in (iii) und (vi) $a_2^{(0)} = a_2^{(n)} = 0$ gesetzt wurde.

□

Satz 6.20 (Fehlerabschätzung)
Sei $f \in C^4([a, b])$. Erfüllt der kubische Spline

$$s''(a) = f''(a) \text{ und } s''(b) = f''(b)$$

so gilt

$$\max_{a \le x \le b} |f(x) - s(x)| \le \frac{1}{2} h^4 \max_{a \le x \le b} |f^{(4)}(x)|$$

für

$$h := \max_{1 \le i \le n} |x_i - x_{i-1}|$$

Beweis siehe Skript

Selbst unter noch (wesentlich) schwächeren Voraussetzungen konvergiert die Spline-Interpolation auch gleichmäßig gegen f.

Siehe Folien für Erklärung zu:

Beweis zu stückweisen Polynomen:

$$\max_{a \le x \le b} |f(x) - S(x)| \le \frac{1}{2} h \max_{a \le x \le b} |f(x)|$$

6.6 Trigonometrische Interpolation

Problem:

Interpolation *periodischer* Funktionen

$$\omega \in \mathbb{R}, \omega > 0 \qquad f(x+\omega) = f(x) \qquad \forall x \in \mathbb{R}$$

Besser: Nehme periodische Funktion als Ausgangspunkt:

$$t_n(x) = \frac{a_0}{2} + \sum_{k=1}^{m} \left\{ a_k \cos\left(\frac{kx2\pi}{\omega}\right) + b_k \sin\left(\frac{kx2\pi}{\omega}\right) \right\}$$

sogenannte „trigonometrische Summe".
$t_n(x)$ ist ω-periodisch und hat $2m+1$ Parameter. Setze:

$$n := 2m$$

Ab jetzt:

$$\boxed{\omega := 2\pi}$$

Außerdem: Äquidistante Stützstellen

$$x_k = \frac{2\pi k}{n+1} \qquad k = 0, \cdots, n$$

Interpolationsaufgabe:

$$t_n(x_k) = y_k := f(x_k) \qquad k = 0, \cdots, n$$

Es zeigt sich: Es ist einfacher dieses Problem in $\mathbb{C}$ zu lösen! Dazu betrachte *komplexe* trigonometrische Summe:

$$t_n^*(x) = \sum_{k=0}^{m} c_k e^{i \underbrace{kx}_{\varphi}} \qquad c_k \in \mathbb{C}$$

mit $i := \sqrt{-1}$ imaginäre Einheit.
Eulersche Formel:

$$e^{i\varphi} = \cos\varphi + i \sin\varphi \qquad \varphi \in \mathbb{R}$$

Hilfssatz 6.22 (Komplexe Einheitswurzeln)
Setze:

$$\boxed{w_k := e^{ix_k}} = e^{i\frac{2\pi k}{n+1}} \qquad \forall k \in \mathbb{Z}$$

und gegebenes $n \in \mathbb{N}$. $w_k \in \mathbb{C}$ heißt „k-te Einheitswurzel".
(Grafik: Einheitskreis im Komplexen mit einigen Einheitswurzeln $w_0, \cdots$ aufgetragen)

(a) $w_k^{n+1} - 1 = 0 \quad \forall\, k \in \mathbb{Z}$
w_k sind Lösungen von $w^{n+1} - 1 = 0$
$w_k^{n+1} = \left(e^{i\frac{2\pi k}{n+1}}\right)^{n+1} = e^{i2\pi k} = \underbrace{\cos(2\pi k)}_{=1} + i\,\underbrace{\sin(2\pi k)}_{=0} = 1$

(b) $w_k^j = w_j^k \quad \forall\, j, k \in \mathbb{Z}$
$w_k^j = \left(e^{i\frac{2\pi k}{n+1}}\right)^j = \left(e^{i\frac{2\pi j}{n+1}}\right)^k = w_j^k$

(c) $w_k^{-j} = w_j^{-k} \quad \forall\, j, k \in \mathbb{Z}$

(d) $w_k^j = w_k^{j \bmod (n+1} = w_{k \bmod (n+1)}^j = w_{k \bmod (n+1)}^{j \bmod (n+1)}$
Sei $j = r(n+1) + s$ mit $0 \le s \le n$:
$w_k^j = e^{i\frac{2\pi k}{n+1}} = e^{i\frac{2\pi k[r(n+1)+s]}{n+1}} = \underbrace{e^{i\frac{2\pi k r\cancel{(n+1)}}{\cancel{n+1}}}}_{1} \cdot e^{i\frac{2\pi k s}{n+1}} = w_k^{j \bmod (n+1)}$

(e) $\sum_{j=0}^n w_k^j = \begin{cases} n+1 & k = j \\ 0 & k \in \mathbb{Z} \setminus \{0\} \end{cases}$
$k = 0: \qquad w + 0^j = e^{i0} = 1 \quad \forall\, j \qquad \sum_{j=0}^n 1 = n+1$
$k \neq 0 \qquad w^{n+1} - 1 = \underbrace{(w-1)}_{\substack{\neq 0 \text{ für} \\ k \neq 0}}\underbrace{(w^n + w^{n-1} + \cdots + w + 1)}_{=0} = 0$

Satz 6.23 (komplexe trigonometrische Interpolation)
Zu gegeben Zahlen $y_0, \cdots, y_n \in \mathbb{C}$ gibt es genau eine Funktion der Gestalt

$$t_n^*(x) = \sum_{k=0}^n c_k e^{ikx}$$

die den Interpolationsbedingungen

$$t_n^*(x_j) = y_j \qquad j = 0, \cdots, n \qquad x_j = \frac{2\pi j}{n+1}$$

genügt. Die Koeffizienten sind gegeben durch

$$c_k = \frac{1}{n+1} \sum_{j=0}^n y_j e^{-ijx_k} \quad \forall\, k = 0, \cdots, n \tag{6.15}$$

Beweis .

$$t_n^*(x) = \sum_{k=0}^n c_k e^{ikx} = \sum_{k=0}^n c_k (\underbrace{e^{ik}}_{=w})^k = \sum_{k=0}^n c_k w^k = p_n(w)$$

Jedem t_n^* entspricht somit eindeutig ein $\boxed{w = e^{ix}}$ Polynom im Komplexen

$$t_n^*(x_j) = p_n(e^{ix_j}) = y_j \qquad j = 0, \cdots, n$$

hat eindeutige Lösung. Berechnung der Koeffizienten:

$$p_n(e^{ix_j}) = \sum_{l=0}^{n} c_l (e^{ix_j})^l = \sum_{l=0}^{n} c_l e^{i\frac{2\pi jl}{n+1}} = \boxed{\sum_{l=0}^{n} c_l w_j^l \overset{!}{=} y_j}$$

Lösen des LGS:

$$\sum_j = 0^n w_k^{-j} \left(\sum_{l=0}^{n} c_l e_j^l \right) = \sum_{l=0}^{n} c_l \underbrace{\left(\sum_{j=0}^{n} w_j^{l-k} \right)}_{\sum_{j=0}^{n} w_{l-k}^j = \begin{cases} n+1 & l-k=0 \\ 0 & l-k \neq 0 \end{cases}} = \sum_{j=0}^{n} w_k^{-j} y_j$$

und damit

$$c_k(n+1) = \sum_{j=0}^{n} w_k^{-j} y_j \Leftrightarrow \boxed{c_k \frac{1}{n+1} \sum_{j=0}^{n} y_j e^{-ijx_k}}$$

□

Satz 6.24 (Diskrete Fourier-Analyse)
Für $n \in \mathbb{N}_0$ gibt es zu geg. reellen Zahlen $y_0, \cdots, y_n$ genau ein trigonometrisches Polynom der Form

$$t_n(x) = \frac{a_0}{2} + \sum_{k=1}^{m} \{a_k \cos(kx) + b_k \sin(kx)\} + \frac{\Theta}{2} a_{m+1} \cos((m+1)x)$$

mit $t_n(x_j) = y_j \qquad j = 0, \cdots, n \qquad x_j = \frac{2\pi j}{n+1}$, sowie

$$\begin{array}{lllllll} \Theta & = & 0, & m & = & \frac{n}{2} & n \text{ gerade} \quad \rightarrow \quad a_0, \cdots, a_m, \quad b_1, \cdots, b_m \\ \Theta & = & 1, & m & = & \frac{n-1}{2} & n \text{ ungerade} \quad \rightarrow \quad a_0, \cdots, a_{m+1}, \quad b_1, \cdots, b_m \end{array}$$

Die Koeffizienten sind bestimmt durch:

$$a_k = \frac{2}{n+1} \sum_{j=0}^{n} y_j \cos(jx_k) \qquad b_k = \frac{2}{n+1} \sum_{j=0}^{n} y_j \sin(jx_k)$$

Beweis .
Berechnung der a_k, b_k aus den c_k erfolgt durch:

$$\begin{array}{rcll} a_0 & = & 2c_0 & \\ a_k & = & c_k + c_{n+1-k} & k = 1, \cdots, m \\ b_k & = & i(c_k - c_{n+1-k}) & k = 1, \cdots, m \\ a_{m+1} & = & 2c_{m+1} & (n = 2m+1 \text{ ungerade}) \end{array}$$

$$
\begin{aligned}
a_k &= c_k + c_{n+1-k} \\
&= \frac{1}{n+1}\sum_{j=0}^{n} y_j(e^{-ijx_k} + e^{-ijx_{m+1-k}} \\
&= \frac{1}{n+1}\sum_{j=0}^{n} y_j(e^{-ijx_k} + e^{ijx_k}) \\
&= \frac{1}{n+1}\sum_{j=0}^{n} y_j(\underbrace{\cos(-jx_k)}_{=\cos(jx_k} + \underbrace{i\sin(-jx_k}_{\cancel{-\sin(ix_k)}} + \cos(jx_k) + i\sin(jx_k)) \\
&= \frac{2}{n+1}\sum_{j=0}^{n} y_j \cos(jx_k)
\end{aligned}
$$

□

Bastian erklärt das Gibbsche Phänomen und zeichnet einen hübschen Graphen. Siehe auf der Wikipedia `http://de.wikipedia.org/wiki/Gibbssches_Phänomen`

Schnelle Fourier-Transformation

Wiederholung 6

$$
\begin{aligned}
c_k &= \frac{1}{N}\sum_{j=0}^{N-1} y_j e^{-i\frac{2\pi jk}{N}} && k = 0, \cdots, N-1 \\
y_j &= \sum_{k=0}^{N-1} c_k e^{i\frac{2\pi}{N}k} \quad (\text{mit } \frac{2\pi}{N} = x_j) && j = 0, \cdots, N-1
\end{aligned}
$$

$N = n+1$ DFT = Diskrete Fouriertransformation
Matrixschreibweise:

$$
\begin{array}{rclrcl}
c &:=& (c_0, \cdots, c_{N-1})^T & y &=& (y_0, \cdots, y_{N-1})^T \\
c &=& \frac{1}{N}Wy & (W)_{k_{ij}} &=& e^{-i\frac{2\pi kj}{N}} = w_k^{-j} \\
y &=& Uc & (U)_{j,k} &=& w_j^k
\end{array}
$$

$$
U\frac{1}{N}W = i \Rightarrow \boxed{w^{-1} = \frac{1}{N}U}
$$

N gerade

$$
\begin{aligned}
c_k^N &= \sum_{j=0}^{N-1} y_j e^{-i\frac{2\pi jk}{N}} \\
&= \underbrace{\sum_{j=0}^{\frac{N}{2}-1} y_{2_j} e^{-i\frac{2\pi 2jk}{N}}}_{\text{gerader Teil}} + \underbrace{\sum_{j=0}^{\frac{N}{2}-1} y_{2_{j+1}} e^{-i\frac{2\pi(2j+1)k}{N}=\frac{2\pi 2jk}{N}+\frac{2\pi k}{N}}}_{\text{ungerader Teil}} \\
&= \underbrace{\sum_{j=0}^{\frac{N}{2}-1} y_{2_j} e^{-i\frac{2\pi jk}{N/2}}}_{=:\tilde{c}_k^g \text{ DFT der Länge } N/2!} + e^{-i\frac{2\pi k}{N}} \underbrace{\sum_{j=0}^{\frac{N}{2}-1} y_{2_{j+1}} e^{-i\frac{2\pi jk}{N/2}}}_{=:\tilde{c}_k^u \text{ DFT der Länge } N/2!}
\end{aligned}
$$

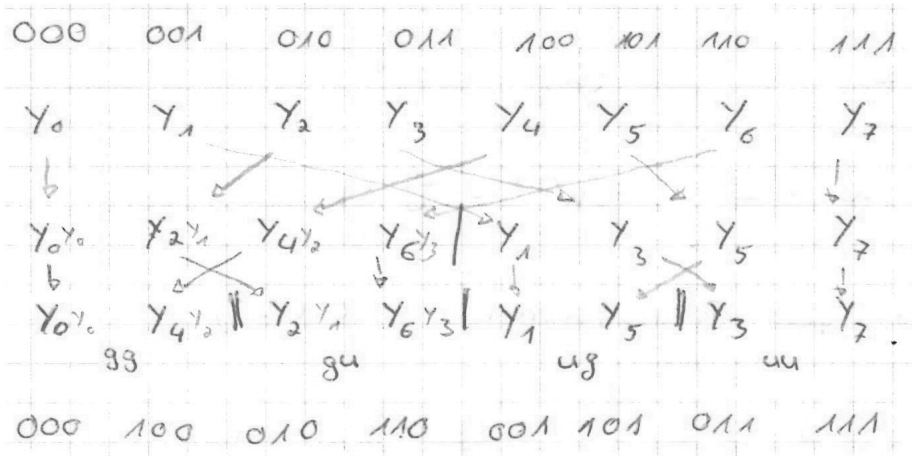
Also:

- $\tilde{c}_k^u, \tilde{c}_k^g$ berechnen sich durch je eine DFT der Länge $\frac{N}{2}$
- Berechung der ursprünglichen Koeffizienten:

$$
\begin{aligned}
\tilde{c}_k &= \tilde{c}_k^g + e^{-i\frac{2\pi k}{N}} \tilde{c}_k^u & 0 \le k < \tfrac{N}{2} \\
\tilde{c}_k &= \tilde{c}_{k-N/2}^g + e^{-i\frac{2\pi k}{N}} \tilde{c}_{k-N/2}^u & \tfrac{N}{2} \le k < N
\end{aligned}
$$

- Prinzip: „Teile und Herrsche“ (Divide and Conquer)
- Rekursive Fortsetzung erfordert $N = 2^d \to$ bis $N = 2$ erreicht wird

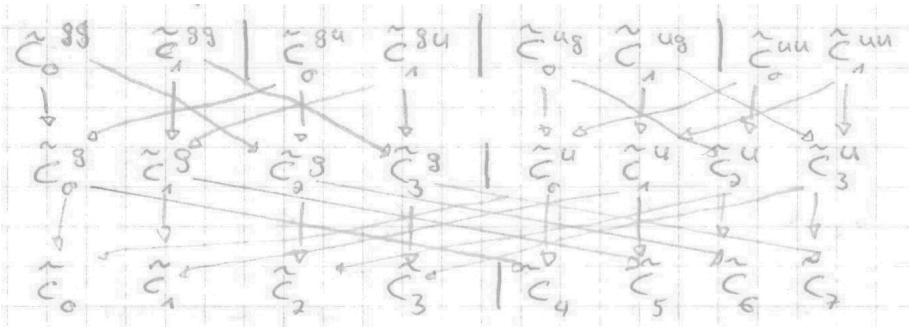
Beispiel: N=8



Permuationen beschrieben durch bit-reversal

$$(b_{d-1}, \cdots, b_0) \to (b_0, b_1, \cdots, b_{d-1})_2$$

Aufwärtsphase:



„perfect shuffle“

Aufwand: $N = 2^d$

$$
\begin{aligned}
A(N) &= 2A(\frac{N}{2}) + c \cdot N \\
&= 2[\overbrace{2 \cdot A\frac{N}{4} + cN2}] + c \cdot N \\
&= 4 \cdot A(\frac{N}{4} + c \cdot N + c \cdot N \text{ (d-mal} \cdots) \\
&= \underbrace{2^d}_{=N} A(1) + \underbrace{cN + \cdots + cN}_{d-1 \text{ Summanden}} \\
&= d \cdot c \cdot N = \mathcal{O}(N \operatorname{ld} N)
\end{aligned}
$$

6.7 Approximation von Funktionen

Wir betrachten Approximationen in sogenannten Prähilberträumen

Definition 6.25
Ein Vektorraum von Funktionen ünber $\mathbb{R}$ und $\mathbb{C}$ mit Skalarprodukt heißt *Prähilbertraum*.

Beispiele

(a) Raum der stetigen Funktionen $C(a, b)$ mit dem Skalarprodukt

$$(f, g) = \int_a^b f(x)\overline{g(x)}dx$$

FIXME: Graph

(b) Raum der quadratintegierbaren Funktionen

$$L^2(a, b) = \left\{ f \mid \int_a^b |f(x)|^2 dx < \infty \right\} \qquad \text{(Lebesgue-Integral)}$$

$L^2(a, b)$ ist vollständig (jede Cauchyfolge konvergiert) bezüglich der Norm

$$\|f\| = \sqrt{(f, f)}$$

(c) Gegeben:

$\Psi = \{\psi_1, \cdots, \psi_N\}, \psi_i \in C(a,b)$

$S = \operatorname{span}\Psi = \{f : f = \sum_{i=1}^N c_i\psi_i\}$

endlichdimensionaler Hilbertraum (mit SP von oben). Betrachte folgende Aufgabe: $S \subset H$ mit $\dim H < \infty$ zu gegebenem $f \in H$ finde $g \in S$, sodass

$$\|f - g\| \to \min \tag{6.17}$$

wobei $\|f\| = \sqrt{(f,f)}$

Satz 6.26 (Allgemeine Gauß-Approximation)
Die Aufgabe (6.17) hat genau eine Lösung $g \in S$. Diese ist charakterisiert durch:

$$(g,\varphi) = (f,\varphi) \quad \forall \varphi \in S \tag{6.18}$$

Beweis .

(a) Sei $\|f - g\|$ minimal für $g \in S$

$$F_\varphi(t) = \|f - (g + t\varphi)\|^2 \qquad F_\varphi : \mathbb{R} \to \mathbb{R} \text{ für beliebiges } \varphi \in S$$

g Minimum $\Rightarrow \frac{d}{dt}F_\varphi(t)|_{t=0} \stackrel{!}{=} 0$

$$\begin{aligned}
&= \frac{d}{dt}\Big[(f-g,f-g) - 2t(f-g,\varphi) + t^2(\varphi,\varphi)\Big]|_{t=0} \\
&= [-2(f-g,\varphi) + 2t(\varphi,\varphi)\ |_{t=0} \\
&= -2(f-g,\varphi) \stackrel{!}{=} 0
\end{aligned}$$

$$\Leftrightarrow (f-g,\varphi) = 0 \quad \forall \varphi \in S$$

(b) Sei $(f-g,\varphi) = 0 \quad \forall \varphi \in S$ Für beliebiges $g' \in S$ gilt:

$$\begin{aligned}
\|f - g'\|^2 \underbrace{=}_{g'=g+\underbrace{g'-g}_{=:\varphi}} \|f - g - \varphi\|^2 &= (f-g-\varphi, f-g-\varphi) \\
&= (f-g,f-g) - 2t\underbrace{(f-g,\varphi)}_{=0 \quad \forall\varphi} + (\varphi,\varphi) \\
&= \|f-g\|^2 + \|\varphi\|^2 \geq \|f-g\|^2 \text{ also g Minimum}
\end{aligned}$$

Damit

$$\|f - g\| \to \min \Leftrightarrow (f-g,\varphi) = 0 \quad \forall \varphi \in S$$

(c) Eindeutigkeit des Minimums

Angenommen es gäbe zwei g_1 und g_2 und $g_1 \neq g_2$

$$\begin{aligned}\|f-g\|^2 &= \|f-g_2\} + \underbrace{g_2-g_1}_{\varphi}\|^2 \\ &= \|f-g_2\|^2 + \underbrace{\|g_2-g_1\|^2}_{>0 \text{ da } g_1\neq g_2}\|f-g_2\|^{`} \text{ zu } g_1 \text{ min.}\end{aligned}$$

(d) Existens des Minimum:

S endlichdim. Basis $\Psi = \{\psi_1, \cdots, \psi_N\}$

$$g = \sum_{g=1}^{N} \alpha_j \psi_j$$

$$(g,\varphi) = (f,\varphi) \quad \forall \varphi \in S$$

$$\Leftrightarrow \sum_{j=1}^{N} \alpha_j \underbrace{(\psi_j,\psi_i)}_{(A)_{i,j}} = \underbrace{(f,\psi_i)}_{(b)_i} \qquad i = 1, \cdots, N$$

$$\Leftrightarrow \boxed{A\alpha = b}$$

□

H Prä-Hilberträume, $A \subseteq$ endlich dimensional, $g \in S : \|f-g\| \to$ min für $f \in H$

$$\Leftrightarrow (g,\varphi) = (f,\varphi)\varphi \in S$$

$$\Psi - \{\psi_1, \cdots, \psi_N\} \qquad S = \text{span}_\Psi \text{ Basis.}$$

$$\sum_{j=1}^{N} \alpha_j \underbrace{(\psi_j,\psi_i)}_{:=(A)_{ij}} = \underbrace{(f,\psi_i)}_{:=b_i} \qquad i = 1, \cdots, N$$

$$\Leftrightarrow = \boxed{A\alpha = b}$$

A ist symmetrisch und positiv definit

(i) $(A)_{ij} = (\psi_j,\psi_i) = (\psi_i,\psi_j) = (A)_{ji}$

(ii)

$$\begin{aligned}\alpha^T A\alpha &= \sum_{i=1}^{N} \alpha_i \left(\sum_{j=1}^{N} (A)_{ij}\alpha_j\right) = \sum_{i=1}^{N} \alpha_i \left(\sum_{j=1}^{N} (\psi_j,\psi_i)\alpha_j\right) \\ &= \sum_{i=1}^{N} \alpha_i \left(\sum_{j=1}^{N} \alpha_j \psi_j \psi_i\right) = \left(\underbrace{\sum_{j=1}^{N} \alpha_j \psi_j}_{g\in S, g\neq 0}, \underbrace{\sum_{i=1}^{N} \alpha_i \psi_i}_{f\in S, g\neq 0}\right) \\ &= (g,g) = \|g\|^2 > 0 \text{ da } g \neq 0 \Leftrightarrow A \text{ invertiertbar}\end{aligned}$$

$$(g,\varphi) = (f,\varphi) \quad \forall \varphi \in S$$

$$\int_a^b g\varphi dx = \int_a^b f\varphi dx \quad \forall \varphi \in S$$

$$\int_a^b \frac{dg}{dx}\frac{d\varphi}{dx} dx = \int_a^b f\varphi dx \quad \forall \varphi \in S$$

$$\text{„}\Leftrightarrow\text{“} \frac{d^2 g}{dx^2} = f \text{in}(a,b) \text{ „Randwertproblem“}$$

$$g(a) = g(b) = 0$$

Approximation mit Orthonormalbasen

Ψ orthonormal, d.h. $(\psi_i, \psi_j) = \delta_{ij}$. Dann gilt:

$$\sum_{j=1}^N \alpha_j \underbrace{\psi_j, \psi_i)}_{=\delta_{ij}} = \alpha_i = (f, \psi_i) \qquad i = 1, \cdots, N$$

$$g = \sum_{i=1}^N \alpha_i \psi_i = \sum_{i=1}^N (f, \psi_i)\psi_i$$

Beispiel 6.27

Für $N = 2n+1$ ist

$$\Psi_F = \left\{ \frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}}\cos x, \cdots, \frac{1}{\sqrt{\pi}}\cos(mx), \frac{1}{\sqrt{\pi}}\sin x, \cdots, \frac{1}{\sqrt{\pi}}\sin(mx) \right\}$$

eine Orthonormalbasis auf $[-\pi, \pi]$. Dann gilt

$$g(x) = \frac{a_0}{2} + \sum_{k=1}^m \{a_k \cos(kx) + b_k \sin(kx)\}$$

$$a_k = \frac{1}{\pi}\int_{-\pi} \pi f(x)\cos(kx)dx \qquad k = 0, \cdots, m$$

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(kx)dx \qquad k = 1, \cdots, m$$

$m \to \infty$ Fourierreihen

Fehlerkontrolle

Bisher: $S \subset H$, optimales $g \in S$ nerecjmet Fehler $\|f - g\|$ imd akezpetiert.
Verfeinerung der Approximationsaufgabe:

Finde $S \subset C$ mit $\dim S$ möglichst klein, so dass $\|f - g\| \leq TOL$ (vorgegebene Zahl)

Geht geschickt mit Orthonormalbasis!

- Sei $S_N, N \in \mathbb{N}$ sei eine Folge von Approximationsräumen mit $\dim S_N = N$ (Oft gilt $S_N \subset S_{N+1}$)
- Sei g_N die Bestapproximation in S_N
- Ψ_N sei Orthonormalbasis von S_N (oft gilt $\Psi_N \subset \Psi_{N+1}$).

Für den Fehler gilt:

$$
\begin{aligned}
0 \le \|f - g_N\|^2 = (f - g_N, f - g_N) &= (f,f) - 2(f, g_N) - (g_N, g_N) \\
&\overset{\text{ONB}}{=} (f,f) - 2(f, \underbrace{\sum_{i=1}^N \underbrace{(f,\psi_i)}_{=\alpha_i}\psi_i}_{=y_N}) + (\sum_{i=1}^N (f,\psi_i)\psi_i, \sum_{j=1}^N (f,\psi_j)\psi_j) \\
&= (f,f) - 2\sum_{i=1}^N (f,\psi_i)(f,\psi_i) + \sum_{i=1}^N \sum_{j=1}^N (f,\psi_i)(f,\psi_j)\underbrace{(\psi_i,\psi_j)}_{=\delta_{ij}} \\
&= (f,f) - 2\sum_{i=1}^N (f,\psi_i)^2 + \sum_{i-1}^N (f,\psi_i)^2 \\
&= \underbrace{(f,f)}_{\|f\|^2} - \sum_{i=1}^N \underbrace{(f,\psi_i)^2}_{\alpha_i}
\end{aligned}
$$

also

$$
\|f - g\|^2 = \|f\|^2 - \sum_{i=1}^N (f,\psi_i)^2 \text{ „Parsevalsche Formel“}
$$

Fehlerkontrolle:

$$
\begin{aligned}
\|f - g\| = (f,f) - \sum_{i=1}^N (f,\psi_i)^2 &\le TOL(f,f) \\
\Leftrightarrow \boxed{\sum_{i=1}^N (f,\psi_i)^2 \ge (1 - TOL)(f,f)} &\qquad TOL \in (0,1]
\end{aligned}
$$

Bemerkungen:

(a) Datenkompression: H endlichdimensional, dann ist (f,f) exakt bekannt, sonst (f,f) hinreichend genau berechenbar

(b) $\Psi_N \subset \Psi_{N+1}$ (hierarchische Basis) $\rightarrow$ einfaches Hinzufügen von Summanden

(c) Fehler wird nur kleiner:

$$
\begin{aligned}
\|f - g_{N+1}\|^2 = (f,f) - \sum_{i=1}^{N+1} (f,\psi_i)^2 &\overset{\Psi_N \subset \psi_{N+1}}{=} \underbrace{(f,f) - \sum_{i=1}^N (f,\psi_i)^2}_{\|f - g_N\|^2} - (f,\psi_{N+1})^2 \\
&= \|f - g_N\|^2 - \underbrace{(f,\Psi_{N+1})^2}_{\ge 0} \le \|f - g_N\|^2
\end{aligned}
$$

Adaptive Approximation mit Haar-Wavelets

Es sei

$$\mathrm{tr}(f) = \{x : f(x) \neq 0\}$$

der „Träger“ einer Funktion. Oben definierte Fourier-Basis erfüllt

$$\overline{\mathrm{tr}(\psi)} = [-\pi, \pi] \text{ für } \psi \in \Psi_F$$

(Skizze zu Unterschied zwischen Mathematikdaten (total glatt, oder definierte Zacken) und natürlichen Daten (total viele kleine, unbestimmte Zacken, teils mit riesigen Ausschlägen)
Man hätte gerne Funktionen mit folgenden Eigenschaften:

(a) $(\psi_i, \psi_j) = \delta_{ij}$ orthonormal

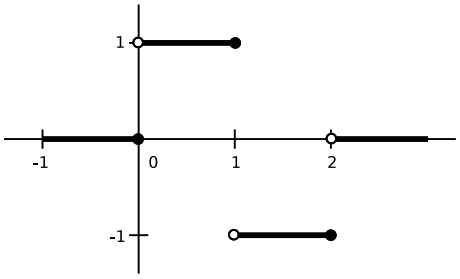
(b) $\Psi_N \subset \Psi_{N+1}$ hierarchisch

(c) $\dim(\mathrm{tr}(\psi_i)) \to 0$ für $i \to \infty$ „lokale Träger“

Definition 6.28 (Haar-Wavelet)

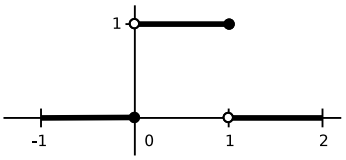
Definiere sogenanntes „mother wavelet“:

$$\Psi(x) = \begin{cases} 1 & 0 < x \leq 1 \\ -1 & 1 < x \leq 2 \\ 0 & \text{sonst} \end{cases}$$



und die Abschneidefunktion

$$\chi(x) = \begin{cases} 1 & 0 < x \leq 1 \\ 0 & \text{sonst} \end{cases}$$



Für $l \in \mathbb{N}_0$ (Stufe) und $0 \leq i < 2^{l-1}$, $i \in \mathbb{N}_0$ (Index) ist das Haar-Wavelet gegeben durch

$$\boxed{\Psi_i^l(x) = \max(\sqrt{2^{l-1}}, 1) \cdot \Psi(2^l x - 2i) \cdot \chi(x)}$$

Die Haar-Waveletbasis der Stufe $l \in \mathbb{N}_0$ ist:

$$\Psi^l = \{\psi_0^0\} \cup \bigcup_{j=1}^{l} \bigcup_{i=0}^{2^{j-1}-1} \{\psi_i^j\}$$

$l = 0$ also $0 \leq i < 2^{0-1} = \frac{1}{2}$

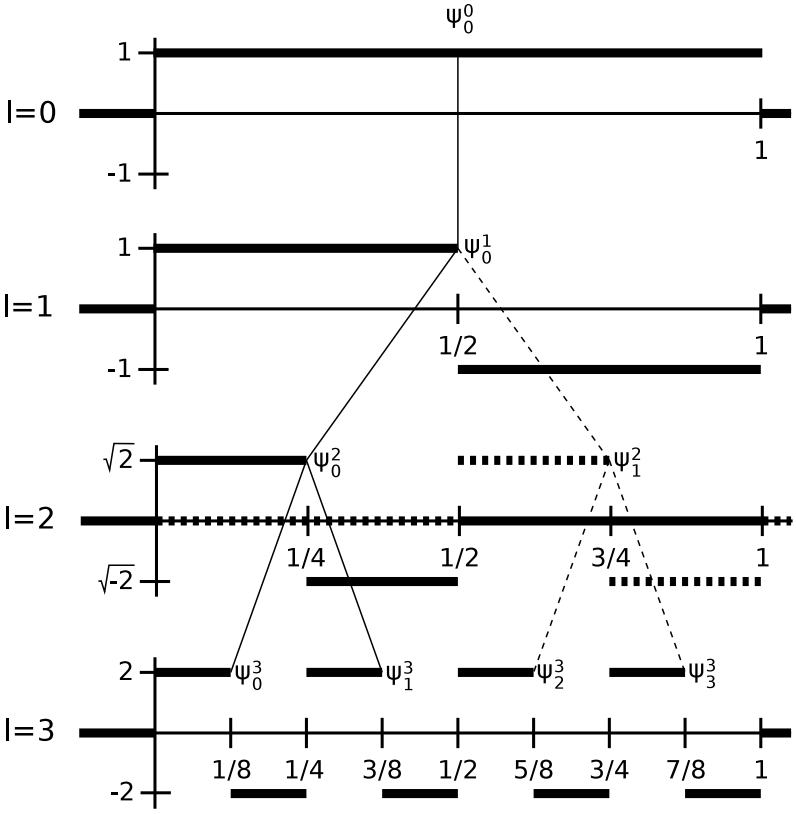
$$\Psi_0^0 = \underbrace{\max(\sqrt{\frac{1}{2}}, 1)}_{=1} \cdot \underbrace{\psi(x)} \cdot \chi(x) = \chi(x)$$

$l = 1$ also $0 \leq i < 2^{1-1} = 1$

$$\Psi_0^1(x) = \underbrace{\max(1,1)}_{=1} \cdot \psi(2x) \cdot \chi(x) = \psi(2x)$$

$l > 0$

$$\Psi_i^l(x) = \begin{cases} \sqrt{2^{l-1}} & \frac{2i}{2^l} < x \leq \frac{2i+1}{2^l} \\ -\sqrt{2^{l-1}} & \frac{2i+1}{2^l} < x \leq \frac{2i+2}{2^l} \\ 0 & \text{sonst} \end{cases}$$



Anmerkung: Träger werden mit jedem Schritt um den Faktor 2 verkleinert (siehe Graphen).
Baumstruktur:

$$\mathrm{tr}(\psi_{2i}^{l+1}) \subset \mathrm{tr}(\psi_i^l)$$
$$\mathrm{tr}(\psi_{2i+1}^{l+1}) \subset \mathrm{tr}(\psi_i^l)$$

Orthogonalitätseigenschaft

Es gilt:

$$(\psi_i^l, \psi_j^k) = \begin{cases} 1 & \text{falls } i = j \text{ und } k = l \\ 0 & \text{sonst} \end{cases}$$

1) $i = j \wedge k = lk = l$

   Für $l = 0 \rightarrow$ offensichtlich erfüllt

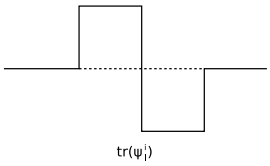
Für $l > 0$

$$(\psi_i^l, \psi_i^l) = \int_{\frac{2i}{2l}}^{\frac{2i+1}{2l}} (\sqrt{2l-1})^2 dx = 2^{l-1}\frac{1}{2^{l-1}} = 1$$

2) $l = k$ aber $i \neq j$ dann gilt $\mathrm{tr}(\psi_i^l) \cap \mathrm{tr}(\psi_j^l) = \emptyset$ also $(\psi_i^l, \psi_j^l) = 0$

3) o.B.d.A. gilt $l > k$, d.h. insbesondere $l > 0$ dann ist ψ_j^k ist konstant auf $\mathrm{tr}(\psi_i^l)$ somit ist

$$(\psi_i^l, \psi_j^l) = \underbrace{c}_{\text{Wert } \psi_j^k} \int_{\frac{2i}{l}}^{\frac{2i+1}{l}} \psi_i^l dx$$



$\psi^l \subset \psi^{l+1}$ erfüllt, da immer nur Basisfunktionen hinzu genommen werden.

Veranschaulichung

Wir definieren die Funktionen

$$\varphi_1^l(x) = \begin{cases} 1 & \frac{i}{2^l} < x \leq \frac{i+1}{2^l} \\ 0 & \text{sonst} \end{cases} \qquad l \in M, 0 \leq i < 2^l$$

(Fixme: Graph der Funktion)

$$\Phi^l = \bigcup_{i=0}^{2^l-1} \left\{\varphi_i^l\right\}$$

$S^l = \operatorname{span} \Phi^l$

S^l ist der Raum der stückweise konstanten Funktionen auf dem Intervall $[0, 1]$ bezüglich der Unterteilung $(\frac{i}{2^l}, \frac{i+1}{2^l}]$. Es gilt

$$\operatorname{span} \psi^l = \operatorname{span} \Psi^l = S^l$$

(Fixme: Graphen zur Veranschaulichung... sieht aus wie Balkendiagramme über die entsprechenden $2^{l-1}, 2^l, \cdots$)

Der Beweisidee folgend kann man auch eine Funktion f bezüglich der Basis Φ^l in eine bezüglich Ψ^l umrechnen. Umgekehrt:

$$f(x) = \sum_l \sum_i c_i \psi_i^l$$

Umrechnung in Ψ^l

effiziente Auswertung

$$f(x) = \sum_{j=0}^{l} \sum_{i=0}^{2^{l-1}-1} c_i^j \psi_i^j$$

ist auszurechnen. Auf $[0, 1]$; für $x \in [0, 1]$ trägt nur eine kleine Menge der Basisfunktion bei, weil Träger lokal begrenzt.

Algorithmus:
$f = c_0^0 \psi_0^0(x)$;
$j = 1, i = 0$
while(true) {
$\quad f = f + c_i^0 \psi_i^j(x)$;
$\quad if(j = l)$ break;
$\quad i = \begin{cases} 2i & x \in \mathrm{tr}(\psi_{2i}^{j+1}) \\ 2i+1 & x \in \mathrm{tr}(\psi_{2i+1}^{j+1}) \end{cases}$
$\quad j = j + 1$;
}

Gegeben: Darstellung von $f \in S$. Gesucht: TEilraum $\tilde{S} \subset S^l$ so dass $\|f - \tilde{f}\| \leq TOL(f, f)$ und $\tilde{S}$ möglichst klein.
Algorithmus:
$I = \{(0,0)\}$; $\qquad I \leq \{(i,l) \mid l \in \mathbb{N}_0, 0 \leq i2^{l-1}\}$
$s = (f, \psi_0^0)$; // Fehler $e = (f, f)$
while($s < (1 - TOL)(f, f)$) {
$\quad$ wähle k und $(k, l+1)$ eines $(i, l) \in I$ mit (f, ψ_k^{l+1}) maximal
$\quad I = I \cup \{(k, l+1)\}$
$\quad s = s + (f, \psi_k^{l+1})$
}

7 Numerische Integration

auch „numerische Quadratur“
Wir beschränken uns auf Integrale einer Raumdimension:

$$I_{(a,b)} = \int_a^b f(x)dx$$

Alle behandelten Verfahren führen auf folgende Form

$$I(f) = \sum_{i=0}^{n} w_i f(x_i) + \text{ Fehler}$$

Hierbei sind

$$w_i \in \mathbb{R} \text{ die Gewichte} \qquad x_i \in \mathbb{R} \text{ Stützstellen}$$

7.1 Newton-Cotes Formeln

interpolatorische Quadraturformeln
Idee: Stelle Interpolationspolynom P auf, bezüglich gewisser Stützstellen berechne Intergral über P exakt
Formel: Stützstellen $(x_i, f(x_i))$, $\quad i = 0, \cdots, n$. Interpolationspolynom in Lagrangedarstellung

$$P_n(x) = \sum_{i=0}^{n} f(x_i) L_i^{(n)}(x)$$

$$L_i^{(n)} = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)}$$

Integral:

$$I(f) \approx I^{(n)}(f) = \int_a^b P_n(x)dx = \sum_{i=0}^{n} f(x_i) \cdot \int_a^b L^{(n)}(x)dx \tag{7.1}$$

Definition 7.1 Ordnung einer Quadratur
Eine Qudraturformel $I^{(n)}(f)$ hat mindestens Ordnung m, wenn sie Polynome vom Grad $m-1$ exakt integriert.

- hier: I-Polynome Grad n, $n+1$ Stützstellen $\Rightarrow$ Ordnung ist mindestens n
- später: geschickte Stütszstellen bis Ordnung $2n+1$ bei $n+1$ Stützstellen
- Für $f \equiv 1$ gilt $p \equiv 1$, damit

$$\int_a^b 1dx = b - a = \sum_{i=0}^{n} w_i$$

Newton-Cotes-Formeln verwenden *äquidistante* Stützstellen.

(a) Abgeschlossene Formeln ($a, b \in$ Stützstellen, $n = 5$)

$$x_j = a + j \cdot H \qquad i = 0, \cdots, n$$

$$H = \frac{b-a}{n}$$

(b) offene Formeln ($a, b \notin$ Stützstellen)

$$x_i = a + (i+1)H \qquad i = 0, \cdots, n$$

$$H = \frac{b-a}{n+2}$$

Berechnung der Gewichte
abgeschlossene Formeln:

$$I^{(n)}(f) \overset{(7.1)}{=} \sum_{i=0}^{n} f(x) \underbrace{\int_0^n L_i^{(n)}(x)dx}_{w_i \text{ (unabhängig von a und b)}}$$

$$= (b-a) \sum_{i=0}^{n} \underbrace{\left(\frac{1}{b-a} \int_a^b L_i^{(n)}(x)dx \right)}_{w_i} \cdot f(x_i)$$

Substituation

$$x = g(s) = a + s \cdot H \qquad s = g^{-1}(x) = \frac{x-a}{H} \qquad g'(s) = H$$

ergibt

$$w_i = \frac{1}{b-a} \in_a^b L_i^{(n)}(x)dx = \frac{1}{b-a}\int_{g^{-1}(a)}^{g^{-1}(b)} L_i^{(n)}(a+sH)\underbrace{g'(s)}_{=H}ds$$
$$= \frac{1}{b-a}\underbrace{\frac{b-a}{n}}_{=H}\int_0^n \prod_{\substack{j=0\\ j\neq i}} \frac{[a+sH-(a+jH]}{[a+iH-(a+jH)]} = \frac{1}{n}\int_0^n \prod_{\substack{j=0\\ j\neq i}} \frac{s-j}{i-j}ds$$

Berechnen der Gewichte

(a) Abgeschlossen:

$$I^{(1)}(f) = \frac{b-a}{2}(f(a)+f(b)) \qquad \text{Trapezregel}$$
$$I^{(2)}(f) = \frac{b-a}{6}(f(a)+4f(\frac{a+b}{2})+f(b)) \qquad \text{Sipsonregel / keplersche Faßregel}$$
$$I^{(3)}(f) = \frac{b-a}{8}(f(a)+3f(a+H)+3f(b-H)+f(b)) \qquad \frac{3}{8}\text{-Regel}$$

(b) offene Formlen

$$I^{(0)}(f) = (b-a)f(\frac{a+n}{2}) \qquad \text{Mittelp.-Regel}$$
$$I^{(1)}(f) = \frac{b-a}{2}(f(a+H)+f(b-H))$$
$$I^{(2)}(f) = \frac{b-a}{3}(2f(a+H)-f(\frac{a+b}{2}+2f(b-H)))$$

Wiederholung 7

$$\sum_{i=1}^{N} f(x_i)w_i \approx \in_a^b f(x)dx$$

Bemerkung 7.2
Ab $n=7$ (abgeschlossene Formeln) und $n=2$ (offene Formeln) ergeben sich auch negative Gewichte w_i.

- Für $f(x) \geq 0$ erwartet man $\int_a^b f(x)dx \geq 0$. Dies gilt aber nicht unbedingt für $I^{(n)}(f)$ falls negative w_i vorkommen.
- Erhöhte Gefahr der Auslöschung
- Kondition ist schlechter:

$$\text{Störung von f: } \tilde{f}(x_i) = f(x_i) + \Delta y + i \qquad |\Delta y_i| \leq \varepsilon$$

$$I^{(n)}(\tilde{f}) = \sum_{i=1}^{n} \tilde{f}(x_i)w_i = \sum_{i=1}^{n}(f(x_i)+\Delta y_i)w_i$$
$$= \underbrace{\sum_{i=1}^{n} f(x_i)w_i}_{I^{(n)}(f)} + \sum_{i=1}^{n}\Delta y_i w_i$$

$$\left|I^{(n)}(\tilde{f}) - I^{(n)}(f)\right| = \sum_{i=1}^{n} \Delta y_i |w_i| \leq \underbrace{\varepsilon \sum_{i=1}^{n} |w_i|}_{(*)}$$

$$(*) = \begin{cases} \text{alle } w_i \text{ positiv} & \sum_{i=1}^{n} |w_i| = \sum_{i=1}^{n} w_i = b-a \\ \text{Gwichte nicht unbedingt positiv} & \sum |w_i| \geq b-a \end{cases}$$

Satz 7.3 (Restglieder)
Für den Fehler in den Quadratformeln gilt:

(i) Trapezregel:

$$I(f) - \frac{b-a}{2}\{f(a)+f(b)\} = -\frac{(b-a)^3}{12} f''(\xi) \qquad f \in C^2[a,b], \xi \in [a,b]$$

(ii) Simpson-Regel

$$I(f) - \frac{b-a}{6}\left\{f(a) + 4f(\frac{a+b}{2}) + f(b)\right\} = \frac{(b-a)^5}{2880} f^{(4)}(\xi) \qquad f \in C^5[a,b], \xi \in [a,b]$$

(iii) Mittelpunktregel

$$I(f) - (b-a) f\left(\frac{a+b}{2}\right) = \frac{(b-a)^3}{24} f''(\xi) \qquad f \in C^2[a,b], \xi \in [a,b]$$

Beweis .

$$\underbrace{\int_a^b f(x) - p_n(x) dx}_{} = \int_a^b \frac{f^{(n+1)}(\eta(x))}{(n+1)!} \prod_{j=0}^{n} (x - x_j) dx$$

$$\text{Trapezregel} \qquad I(f) - I^{(1)}(f) = \frac{1}{2} \int_a^b \underbrace{f''(\eta(x))}_{g(x)} \underbrace{(x-a)(x-b)}_{w(x) \leq 0} dx$$

$$\text{MWS d. Int.rchng.} \qquad = \frac{1}{2} f(\underbrace{\xi}_{\eta(x')} \in_a^b (x-a)(x-b) dx$$

$$= -\frac{(b-a)^3}{12} f''(\xi)$$

- Mittelpunktregel hat halbe Fehler der Trapezregel
- Genrelle Form: $c(b-a)^{m+1} f^{(m)}(\xi)$

□

7.2 Summierte Quadratformeln

Erhöhen des Polynomgrades wenig sinnvoll, da

- negative Gewichte
- Lagrange-Interpolation konvergiert nicht punktweise
- Entsprechende Differenzierbarkeit von f erforderlich

Idee:

Unterteile $[a, b]$ in N Teilintervalle

$$[x_i, x_{i+1}] \qquad x_i = a + ih \qquad i = 0, \cdots, N-1 \qquad h = \frac{b-a}{N}$$

- wende eine der obigen Formeln $I^{(n)}(f)$ in jedem Teilintervall $[x_i, x_{i+1}]$ an
- $\Rightarrow$ „Summierte Quadraturformeln“

Satz 7.4 (Fehler bei summierten Quadraturen)
Für die je Teilintervall verwendete Quadratur gelte die Restglieddarstellung

$$I_{[x_i,x_{i+1}]}(f) - I^{(n)}_{[x_i,x_{i+1}]}(f) = \alpha_n h^{n+2} f^{(n+1)}(\xi_i) \qquad \xi_i \in [x_i, x_{i+1}]$$

Dann gilt für die summierte Formel:

$$I_{[a,b]}(f) - I^{(n)}_h = \alpha_n (b-a) h^{m+1} f^{(m+1)}(\xi) \qquad \xi \in [a, b]$$

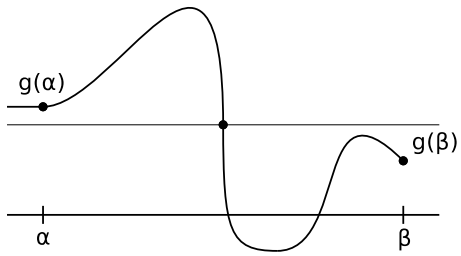
(Trapez: $n = 1, m = 1$; Simpson: $n = 2, m = 3$)

Beweis .

Zwischenwertsatz: $g(x)$ stetig auf $[a, b]$, dann $\exists$ zu jedem

$$u \in [\min(g(x), g(\beta)), \max(g(\alpha), g(\beta))]$$

ein $\eta \in [a, b]$ so dass $g(\eta) = u$.



$\Rightarrow$ Erweiterung auf N Stützstellen

$$\xi_i \in [a, b] \qquad i = 0, \cdots, N-1 \qquad \text{und g stetig}$$

Dann nimmt g jeden Wert zwischen $\min_i g(\xi_i)$ und $\max_i(g(\xi_i))$ an.

$$\min_i g(\xi_i) \le \frac{1}{N}\sum_{i=0}^{N-1} g(\xi_i) \le \max_i g(\xi_i) \Rightarrow \exists \xi \text{ sodass } g(\xi) = \frac{1}{N}\sum_{i=0}^{N-1} g(\xi_i)$$

$$\begin{aligned} I(f) - I^{(n)}_h(f) &= \sum_{i=0}^{N-1} \alpha_n h^{m+2} f^{(m+1)}(\xi_i) \qquad \xi_i \in [x_i, x_{i+1}] \\ &= \alpha_n h^{m+2} \underbrace{\sum_{i=0}^{N-1} f^{(m+1)}(\xi_i)}_{=Nf^{(m+1)}(\xi)} = \alpha_n h^{m+2} N f^{(m+1)}(\xi) \\ &\overset{h=\frac{b-a}{N}}{=} \alpha_n h^{m+2} \frac{b-a}{h} f^{(n+1)}(\xi) = \alpha_n h^{m+1}(b-a) f^{(m+1)}(\xi) \end{aligned}$$

□

Beispiele:

(i) Summierte Trapezregel

$$I_h^{(1)} = \sum_{i=0}^{N-1} \underbrace{\frac{x_{i+1}-x_i}{2}}_{=\frac{h}{2}} \{f(x_i)+f(x_{i+1})\} = h\left\{\frac{f(a)}{2} + \sum_{i=1}^{N-1} f(x_i) + \frac{f(b)}{2}\right\}$$

$$I(f) - I_h^{(1)}(f) = -\frac{b-a}{12}h^2 f''(\xi) \qquad \xi \in [a,b]$$

(ii) Summierte Simpson

$$\begin{aligned} I_h^{(2)}(f) &= \sum_{i=0}^{N-1} \frac{h}{6}\left\{f(x_i) + 4f(\frac{x_i+x_{i+1}}{2})f(x_{i+1})\right\} \\ &= h\left\{\frac{f(a)}{6} + \sum_{i=1}^{N-1}\frac{f(x_i)}{3} + \frac{2}{3}\sum_{i=0}^{N+1} f\left(\frac{x_i+x_{i+1}}{2}\right) + \frac{f(b)}{6}\right\} \\ I(f) - I_h^{(2)}(f) &= -\frac{b-a}{2880}h^4 f^{(4)}(\xi_i) \qquad \xi \in [a,b] \end{aligned}$$

(iii) Summierte Mittelpunktregel

$$I_h^{(0)}(f) = \sum_{i=0}^{N} hf(\frac{x_i+x_{i+1}}{2}) \qquad I(f) - I_h^{(0)} = \frac{b-a}{24}h^2 f''(\xi)$$

Bemerkung 7.5

- $$\underbrace{I_h^{(2)}(f)}_{\text{Simpson}} = \underbrace{\frac{1}{3}I_h^{(1)}(f)}_{\text{Trapez}} + \underbrace{\frac{2}{3}I_h^{(0)}(f)}_{\text{Mittelpkt}}$$

  $\Rightarrow$ Unterschied zwischen

  $$\underbrace{I_h^{(2)}(f)}_{\mathcal{O}(h^4)} \text{ und } \underbrace{I_h^{(1 \text{ oder } 0)}(f)}_{\mathcal{O}(h^2)}$$

  liefert Möglichkeit zur Fehlerkontrolle

- $$I_{h/2}^{(1)} = \frac{1}{2}I_h^{(1)}(f) + \frac{1}{2}I_h^{(0)}(f)$$

7.3 Quadraturen höherer Ordnung

Romberg-Integration

= Extrapolation zum Limes mit Summierter Trapezregel

$$\lim_{h\to 0} I_h^{(1)}(f) = I(f)$$

Euler-MacLaurinsche Summenformel:

$$I(f) - I_h^{(n)}(f) = \sum_{k=1}^{m} h^{2k} \underbrace{\frac{B_{2k}}{(2k)!}(f^{(2k=1)}(b) - f^{(2k-1)}(a))}_{\text{unabhängig von h}} + h^{2m+2}\frac{B_{2m+2}}{(2m+2)!}(b-a)f^{(2m+2)}(\xi)$$

B_i Bernoulli Zahlen

Gauß-Integration

Frage: Lässt sich die Genauigkeit bei freier Wahl der Stützstellen erhöhen?
Idee: Wähe w_i, x_i sodass Polynome möglichst hohen Grades exakt integriert

Satz 7.6
Die maximal erreichbare Ordnung einer Quadraturformel mit $n+1$ Stützstellen ist $2n+2$ (d.h. Polynome vom Grad $2n+1$ werden exakt intgriert)

Beweis .
Angenommen Ordnung $2n+3$, d.h. Polynom vom Grad $2n+2$ wird exakt integriert. Betrachte:

$$q(x) = \prod_{i=0}^{n}(x-x_i)^2$$

- $q(x)$ hat Grad $2n+2$
- $q(x) \geq 0 \quad \forall x$ und $q(x) \not\equiv 0$ also $\int_{-1}^{1} dx > 0$
- andererseits $\sum_{i=0}^{n} \underbrace{q(x_i)}_{=0} w_i = 0$`

□

Satz 7.7 (Gauß-Quadratur)
Es gibt genau eine interpolatiorische Quadratformel zu $n+1$ paarweise verschiedene Stützstellen in $[-1,1]$ mit der Ordnung $2n+2$. Ihre Stützstellen sind die Nullstellen $\lambda_0, \cdots, \lambda_n \in (-1,1)$ des $(n+1)$-ten Legendrepolynoms L_{n+1}:

$$L_0(x) = 1, L_1(x) = x \qquad L_{n+1}(x) = \frac{2n+1}{n+1}L_n(x) - \frac{n}{n+1}L_{n-1}(x)$$

Die Gewichte erhält man mittels

$$w_i = \int_{-1}^{1} \prod_{\substack{j=0\\ j\neq i}}^{n} \left(\frac{x-\lambda_j}{\lambda_i-\lambda_j}\right)^2 dx$$

Beispiele:

$$h = \frac{b-a}{2} \qquad c = \frac{b+a}{2}$$

$$\begin{aligned} n=2: \quad I^{(1)}(f) &= \frac{b-a}{2}\left\{5f(c-\sqrt{\tfrac{1}{3}}-h)+f\left(c+\sqrt{\tfrac{1}{3}}h\right)\right\} && \text{Ordnung: 4} \\ n=2: \quad I^{(2)}(f) &= \frac{b-a}{18}\left\{5f\left(c-\sqrt{\tfrac{3}{5}}-h\right)+8f(c)+5f\left(c+\sqrt{\tfrac{3}{5}}h\right)\right\} && \text{Ordnung: 6} \end{aligned}$$

Entsprechend lassen sich summierte Formeln definieren.

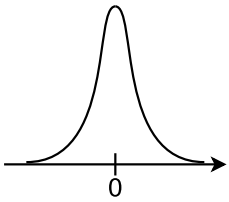
7.4 Ausblick

Adaptive Quadratur

Betrachte

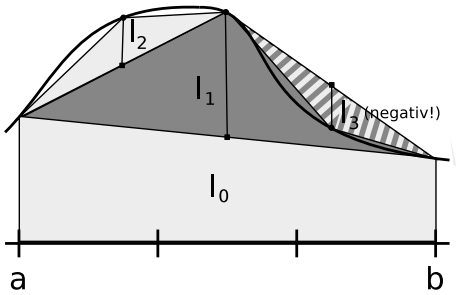
$$f(x) = \frac{1}{10^{-5}+x^2}$$

Summierte Quadratur mit fester Schrittweite ineffizient.



Prinzip von Archimedes:

- $I(f) = I_0 + I_1 + I_2 + \cdots$
- Breche rekursive Unterteilung ab, falls $|I_j|$ klein genug



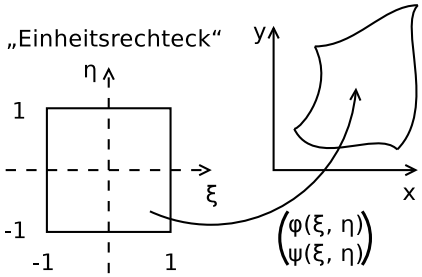
Mehrdimensionale Quadratur

Die Welt ist nicht eindimensional. Für Recheckte ($d=2$), Quader ($d=4$), ... lassen sich obigen Formeln leicht erweitern:

$$\begin{aligned} \int_c^d \int_a^b f(x,y)dxdy &\approx \int_c^d \sum_{i=0}^n f(x_i,y)w_i dy = \sum_{i=0}^n \int_c^d f(x_i,y)dy w_i \\ &\approx \sum_{i=0}^n \sum_{j=0}^n f(x_i,y_j) \underbrace{w_i w_j}_{=w_{ij}} \end{aligned}$$

Allerdings sind nicht alle Gebiete Rechtecke (anders als in 1D!). Koordinatentransformation:

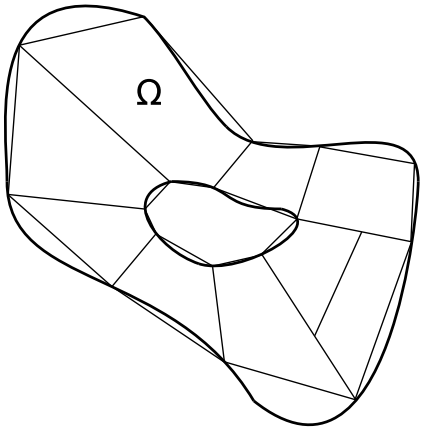
$$\text{Abb.}\begin{pmatrix} \varphi(\xi,\eta) \\ \psi(\xi,\eta) \end{pmatrix} : [-1,1]\times[-1,1] \to \Omega \subset \mathbb{R}^2$$



$$\int_\Omega f(x,y)dxdy = \int_{-1}^1 \int_{-1}^1 f\left(\varphi(\xi,\eta),\psi(\xi,\eta)\right)\left(\frac{\partial(\varphi,\psi)}{\partial(\xi,\eta)}\right)d\xi d\eta$$

$$\left|\frac{\partial(\varphi,\psi)}{\partial(\xi,\eta)}\right| = \det\begin{bmatrix} \frac{\partial\varphi}{\partial\xi}(\xi,\eta) & \frac{\partial\psi}{\partial\xi}(\xi,\eta) \\ \frac{\partial\varphi}{\partial\eta}(\xi,\eta) & \frac{\partial\psi}{\partial\eta}(\xi,\eta) \end{bmatrix} \qquad \text{(Transponierte Jacobimatrix)}$$

Summierte Formeln in mehreren Raumdimensionen. Bei komplexeren, z.B. Gebieten mit Löchern reicht das nicht



- Zerlegung in Teilgebiete die sich auf Rechtecke transformieren lassen. „Gittergenerierung“ nicht trivial und schwierig „automatisch“ zu machen
- Erfordert Beschreibung des Gebietes Ω
- Zusätzlicher Geometriefehler durch nicht exakte Approximation der Geometrie
- Man kann auch Simplizes (= Dreieck, Tetraeder) zur Unterteilung verwenden)

Fluch der Dimension

Ist d sehr groß, so sind die hier behandelten Methoden nicht brauchbar.
Betrachte $\Omega = [0,1]^d$. Zerlegt man $[0,1]$ in zwei Teilintervalle je Richtung so hat man den d-dimensionalen Würfel in 2^d Teilwürfel zerlegt.
$\Rightarrow$ Der Aufwand steigt exponentiell in d an. Dies bezeichnet man als „Fluch der Dimension“. Eine Möglichkeit ist dann die Monte-Carlo Integration

$$I(f) \approx \frac{C}{N}\sum_{i=1}^{N} f(\xi_i) \qquad \text{mit Zufallszahlen } \xi_i \in \Omega$$

8 Iterative Lösung von Gleichungsystemen

In diesem Abschnittt betrachten wir die Lösung von algebraischen Gleichungen

$$f(x) = 0 \text{ mit } f\colon \mathbb{R}^n \to \mathbb{R}^n$$

Dabei beschränken wir uns zunächst auf den Fall $n = 1$ (skalar).

Intervallschachtelung

Idee: Angenommen man kennt ein Teilintervall $[a_0, b_0]$ so dass $f(a_0)f(b_0) < 0$ (unterschiedliche Vorzeichen) und f sei stetig. Dann hat f nach dem Zwischenwertsatz mindestens eine Nullstelle in $[a_0, b_0]$.
Algorithmus: Gegeben: $I_0 = [a_0, b_0]$ mit $f(a_0)f(b_0) < 0$ und $\varepsilon > 0$ t=0;

```
while(b_t - a_t > ε) {
    x_t = (a_t + b_t)/2;
    if(f(x_t) == 0) {
        a_t = x_t − ε; b_t = x_t;
    } else if(f(a_t)f(x_t) < 0) {
        a_{t+1} = a_t; b_{t+1} = x_t;      // Nullstelle in [a_t, x_t]
```

```
    } else {
        a_{t+1} = x_t; b_{t+1} = b_t;      // es ist f(x_t)f(b_t) < 0 da VZ von f(a_t) = VZ von f(x_t)
    }
    t = t + 1;
}
```

Analyse: Es gilt

$$a_t \leq a_{t+1} < b_{b+1} \leq b_t$$

und

$$|b_{t+1} - a_{t+1}| \leq \frac{1}{2}|b_t - a_t| = \left(\frac{1}{2}\right)^{t+1} |b_0 - a_0|$$

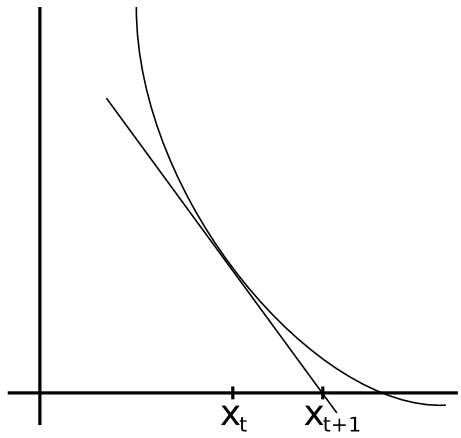
(solange nicht $f(x_t) \equiv 0$).
Bemerkung:

- Konvergenz ist linear mit Rate $\frac{1}{2}$
- Sehr gut geeignet für monotone Funktionen
- nur für reelle Funktionen im $\mathbb{R}^1$ geeignet

8.1 Newton Verfahren

Die Funktion sei (mindestens) einmal stetig differenzierbar.
Idee:
Gegeben: x_t. Da $f \in C'$ gibt es „Tangente".

$$T_t(x) = f'(x_t)(x - x_t) + f(x_t)$$



Nullstelle der Tangente:

$$T_t(x) = 0 \Leftrightarrow x = x_t - \frac{f(x_t)}{f'(x_t)}$$

.
Dies führt zur Iterationsvorschrift

$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}$$

Offensichtlich ist $|f'(x_t)| > 0$ erforderlich, d.h. wir setzen voraus, dass die Nullstelle *einfach* ist.
Das Newton-Verfahren lässt sich auf Systeme $f\colon \mathbb{R}^n \to \mathbb{R}^n$ erweitern:
Es existiere die Taylorentwicklung von f:

$$f_i(x_t + \Delta x) = f_i(x_t) + \sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j}(x_t)\Delta x_i + R_i(x_t, \Delta x) \qquad i = 1, \cdots, n$$

in vektorieller Schreibweise

$$f(x_t, \Delta x) = f(x_t) + I(x_t)\Delta x + R(x_t, \Delta x)$$

$$(J(x_t))_{ij} = \frac{\partial f_i}{\partial x_j}(x_t) \qquad \text{„Jacobimatrix“}$$

Ignorieren des Restgliedes entspricht „Linearisierung von f“

$$f(x_t) = J(x_t)\Delta x \overset{!}{=} 0$$

$$\Leftrightarrow \Delta x = -J^{-1}(x_t)f(x_t)$$

führt zur Iteration

$$x_{t+1} = x_t - J^{-1}(x_t)f(x_t)$$

Jeder schitt erfordert Lösung eines LGS mit der Jacobimatrix!
Nun untersuchen wir die Konvergenz des Newton-Verfahrens. Allerdings nur im $\mathbb{R}^1$.

Satz 8.1 (Newton-Verfahren)
Die Funktion $f \in C^2[a,b]$ habe in (a,b) (Inneres!) eine Nullstelle in Z und es sei

$$m := \min_{a \le x \le b} |f'(x)| > 0 \qquad M := \max_{a \le x \le b} |f''(x)|$$

Es sei $\varrho > 0$ so gewählt, dass

$$q := \frac{M}{2m}\varrho < 1 \qquad K_\varrho(z) := \{x \in \mathbb{R} \mid |x - z| \le \varrho\} \subset [a,b]$$

Dann sind für jeden Startwert $x_0 \in K_\varrho(z)$ die Newton-Iterationen $x_t \in K_\varrho(z)$ definiert und konvergieren gegen die Nullstelle z.
Dabei gilt die a-priori Fehlerabschätzung

$$|x_t - z| \le \frac{2m}{M} q^{(2^t)} \qquad t \in \mathbb{N}$$

(a-priori: nur Abh. von den Voraussetzungen. a-posteriori: auch Abh. von bereits berechneten Iterationen)
und die a-posteriori Fehlerabschätzung:

$$|x_t - z| \le \frac{1}{m}|f(x_t)| \le \frac{M}{2m}|x_t - x_{t-t}|^2 \qquad t \in \mathbb{N}$$

Beweis. Vorbereitungen:

(i) Mittelwert der Differnatialrechnung liefert für alle $x, y \in [a,b], x \neq y$

$$\left|\frac{f(x) - f(y)}{x - y}\right| = |f'(\xi)| \overset{\text{Vor.}}{\ge} m \Leftrightarrow \frac{1}{m}|f(x) - f(y)|$$

- f ist Lipschitz-stetig
- die ullstelle z ist eindeutig, da sonst $0 < |z_1 - z_2| \le \frac{1}{m}|f(z_1) - f(z_2)| = 0$`

(ii) Da $f \subset C^2[a,b]$ gilt folgende Taylordarstellung:

$$f(y) = f(x) + (y-x)f'(x) + \underbrace{\int_x^y (x-t)f''(x)dt}_{=:R(y,x)\text{ Restglied}}$$

Transformation des Integrals mit

$$\varphi(s) = x + s(y-x) \qquad \varphi : [0,1] \to [x,y]$$

liefert für das Restglied

$$\begin{aligned} R(y,x) &= \int_x^y (x-t)f''(t)dt = \int_0^1 (x-\varphi(s))f^1(\varphi(s))ds \\ &= \int_0^1 (\underbrace{x-x}_{=0} - s(y-x))f''(x+s(y-x))(y-x)ds \\ &= -(y-x)^2 \int_0^1 sf''(x+s(y-x))ds \end{aligned}$$

und damit

$$|R(y,x)| \le (y-x)^2 \int_0^1 s\,\underbrace{|f''(x+s(y-x))|}_{\le M \text{ nach Vor.}}\,ds \le \frac{M}{2}|y-x|^2$$

(iii) Nun setze

$$g(x) := x - \frac{f(x)}{f'(x)} \qquad (\text{d.h. } x_{t+1} = g(x_t))$$

Dann gilt:

$$g(x) - z = x - \frac{f(x)}{f'(x)} - z = -\frac{1}{f'(x)}\underbrace{\{f(x) + (z-x)f'(x)\}}_{=-R(z,x)}$$

wh. $\underbrace{f(z)}_{=0} = \underbrace{f(x) + (z-x)f'(x)} + R(z,x)$.

Für $x \in K\varrho(z)$ gilt dann

$$|g(x) - z| = \left|\frac{1}{f'(x)}R(z,x)\right| \le \frac{1}{m}\frac{M}{2}|z-x|^2 = \underbrace{\frac{M}{2m}\underbrace{|x-z|}_{\le\varrho}}_{<1 \text{ n. Wahl v. } \varrho\circledast}\underbrace{|x-z|}_{\le\varrho \text{ da } x\in R_\varrho(z)} \overset{\text{wg. } <1}{<} \varrho$$

Somit folgt aus $x \in K_\varrho(z)$, dass auch $g(x) \in K_\varrho(z)$. g bildet die Menge $K_\varrho(z)$ auf sich selbst ab.

Die Newton-Iteraten sind $x_{t+1} = g(x_t)$. Setze

$$\varrho_t := \frac{M}{2m}|x_t - z| = \frac{M}{2m}|g(x_{t-1}) - z \overset{\circledast}{\le} \frac{M}{2m}|x_{t-1} - z|^2 = \varrho_{t-1}^2$$

Somit gilt nach t Schritten:

$$\varrho_t \leq \varrho_{t-1}^2 \leq \varrho_{t-2}^5 \cdots \leq \varrho_{\underbrace{t-t}_{=0}}^{(2^t)} = \varrho_0^{(2^t)}$$

und damit wegen $|x_t - z| = \frac{2m}{M}\varrho_t$ und $\varrho_0 = \frac{M}{2m}\underbrace{|x_0 - z|}_{\leq \varrho} \leq q < 1$

$$|x_t - z| = \frac{2m}{M}\varrho_t \leq \frac{2m}{M}\varrho_0^{(2^t)} \leq \frac{2m}{M}q^{(2^t)}$$

was zu zeigen war.
A-posteriori Abschätzung folgt aus Taylor-Formel für x_t, x_{t-1}:

$$f(x_t) = \underbrace{f(x_{t-1}) + (x_t - x_{t-1})f'(x_t)}_{=0 \text{ nach Konstruktion!}} + R(x_t, x_{t-1})$$

und

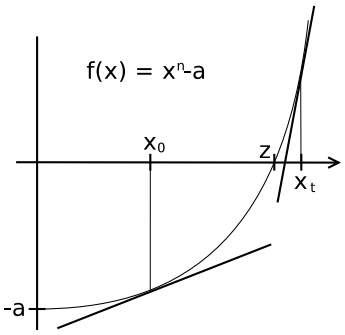
$$|x_t 0 z| \overset{\substack{\text{Lipschitz}\\ \text{vom Anfang}}}{\leq} \frac{1}{m}|f(x_t) - \underbrace{f(z)}_{=0}| = \frac{1}{m}|r(x_t, x_{t-1})| \leq \frac{M}{2m}|x_t - x_{t-1}|^2$$

□

Beispiel 8.2 (Wurzelberechnung mit Newton-Verfahren)
$a > 0, n \geq 1$ löse $x^n = a \Leftrightarrow f(x) = x^n - a = 0, f'(x) - nx^{n-1}$. Also

$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)} = x_t - \frac{x_t^n - a}{nx_t^{n-1}} = \frac{nx_t^n - x_t^n + a}{nx_t^{n-1}} = \frac{n-1(x_t^n) + q}{nx_t^{n-1}} = \frac{1}{n}\left\{(n-1)x_t + \frac{a}{x^{n-1}}\right\}$$

Satz 8.1 behauptet: Iteration konvergiert, falls x_0 nahe genug an z. Hier gilt jedoch: Iteration konvergiert global, d.h. für alle $x_0 > 0$. Aber nicht unbedingt quadratisch von Beginn an.



1) für $x_t > z$ gilt $|x_{t+1} - z| < |x_t - z|$ da $f(x_t) > 0$ und $f'(x_t) > \frac{f(x)}{z - x_0}$.

2) $0 < x_0 < z$ dann ist $x_1 > z$ da $f(x_0) < 0$ und $f'(x_0) < \frac{-f(x_0)}{z - x_0}$.

Man zeigt: für $n = 2$ ist für $|x_0 - \sqrt{a}| \leq 2\sqrt{a}$ die Konvergenz quadratisch.

Bemerkungen zum Newton-Verfahren

- Das Newton-Verfahren konvergiert nur *lokal*, d.h. wenn $|x_0 - z| \leq \varrho \rightarrow$ „Einzugsbereich“. Wobei
  - ϱ i.d.R. unbekannt
  - ϱ möglicherweise sehr klein ist. Oben: $\frac{M}{2m}\varrho < 1 \Rightarrow \varrho < \frac{2m = \min f'}{M = \max f''}$

- Newton Verfahren konvergiert quadratisch

$$|x_t - z| \le c|x_{t-1} - z|^2 \text{ zum Vgl. Intervallsch.: } |x_t - z| \le \frac{1}{2}|x_{t-1} - z|^1$$

- gedämpftes Newton-Verfahren: Verbesserung der Konvergenz *außerhalb* des Einzugsbereichs:

$$x_{t+1} = x_t - \lambda_t \frac{f(x_t)}{f'(x_t)} \qquad \lambda_t \in (0, 1]$$

  Wahl von λ_t „Dämpfungstrategie".
- Mehrfache Nullstellen
  Sei z zunächst zweifache Nullstelle, d.h. $f(z) = f'(z) = 0$ und $f''(z) \neq 0$. Wegen

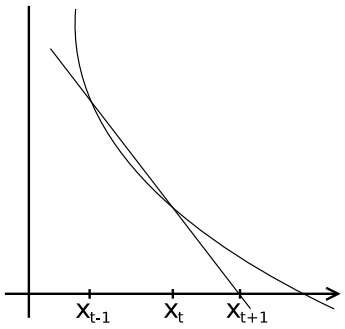
$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)} = x_t - \frac{f(x_t) - f(z)}{f'(x_t) - f'(z)} \overset{\text{Erweitern}}{=} \frac{\frac{f(x_t)-f(z)}{x_t - z}}{\frac{f'(x_t)-f'(z)}{x_t - z}} = x_t - \frac{f'(\xi_t)}{f''(\eta_t)}$$

  und $f''(z) \neq 0$ bleibt die Iteration für $x_t \to z$ (und damit $\eta_t \to z$) wohldefiniert.
  Man zeigt: Für p-facje Nullstelle zeigt

$$x_{t+1} = x_t - p\frac{f(x_t)}{f'(x_t)}$$

  quadratische Konvergenz.
- Sekanten-Methode
  Berechnung der Ableitung unter Umständen teuer. Idee:
  Ersetze Tangente durch eine Sekante:

$$s(x) = f(x_t) + (x - x_t)\frac{f(x_t) - f(x_{t-1})}{x_t - x_{t-1}}$$



  Ansatz: $s(x) \overset{!}{=} 0$ führt auf Iteration

$$x_{t+1} = x_t - f(x_t)\frac{x_t - x_{t-1}}{f(x_t) - f(x_{t-1})}$$

  Konvergenz: lokal mit

$$|x_t - z| \le \frac{2m}{M} q^{\gamma_t} \qquad t \in \mathbb{N} \qquad \substack{\gamma_0=\gamma_1=1 \\ \gamma_{t+1}=\gamma_t+\gamma_{t-1}} \text{ „Fibonaccizahlen"}$$

  Nur eine f-Auswertung pro Iteration notwendig $\gamma_t \tilde{0.723} \cdot (1.618)^t$ Konvergenzordnung 1,618 also zwischen 1 und 2.
  Problem: Auslöschung in $\frac{x_t - x_{t-1}}{f(x_t) - f(x_{t-1})}$.

8.2 Sukzessive Approximation

Mit $g(x) = x - \frac{f(x)}{f'(x)}$ hat das Newton-Verfahren die form $x_{t+1} = g(x_t)$. Da die Nullstelle z von wg. $f(z) = 0$ einen Fixpunkt der Iteration $x_{t+1} = g(x_t)$ ist nennt man das auch Fixpunktiteration.

Hier untersuchen wir nun allgemeine Iterationen dieser Art. Z.B. könnte die Berechnung von $f'(x)$ sehr teuer sein und man wertet f' nur einmal „in der Nähe“ von z aus:

$$x_{t+1} = x - \frac{f(x)}{f'(c)}$$

. Frage: Wann konvergiert so eine Iteration: Insbesondere wollen wir auch $f\colon \mathbb{R}^n \to \mathbb{R}^n$ zulassen. Antwort gibt der sogenannte „Banachsche Fixpunktsatz“.

Satz 8.3 (Sukzessive Approximation)
Sei $G \subset \mathbb{R}^n$ eine nicht leere, abgeschlossene Punktmenge und $gG \to G$ Lipschitz-stetig mit Konstante $q < 1$ d.h.

$$\|g(x) - g(y)\| \le q\|x - y\|$$

Hierbei ist $\|\cdot\|$ eine Vektornorm im $\mathbb{R}^n$ und g nennt man eine „Kontraktion“. Dann existiert genau ein Fixpunkt $z \in G$ von g und für jeden Startpunkt $x^{(0)} \in G$ konvergiert die Folge der Iterierten $x^{(t+1)} = g(x^{(t)})$ gegen z.
Es gelten die a-posteriori und a-priori Fehlerabschätzungen

$$x^{(t)} - z\| \le \frac{q}{1-q}\|x^{(t)} - x^{(t-1)}\| \le \frac{q^t}{1-q}\|x^{(1)} -^{(0)}\|$$

(Wir schreiben den Iterationsindex oben in Klammern, damit bei Vektoren unten Platz für den Komponentenindx bleibt).

Beweis. (i) Da $g\colon G \to G$ ist $x^{(t)} = g(x^{(t-1)}) = g(g(x^{(t-2)})) = \cdots = \underbrace{g(\cdots g}_{t \text{ mal}}(x^{(0)})\cdots)$ wohldefiniert.

(ii) Weiter ist $\|x^{(t+1)} - x^{(t)}\| = \|g((^{(t)} - g(x^{(t-1)}\| \le q\|x^{(t)} - x^{(t-1)}\| \le q^t\|x^{(1)} - x^{(0)}\|$

(iii) Zeige nun, dass die $x^{(t)}$ eine Cauchy-Folge bilden. sei $m \ge 1$ und $\varepsilon > 0$ gegebne. Es ist

$$\begin{aligned}
\|x^{(t+m)} - x^{(t)}\| &= \|\underbrace{x^{(t+m)} - x^{(t+m-1)}} + \underbrace{x^{(t+m-1)} - x^{(t+m-2)}} + \cdots + \underbrace{x^{(t+1)} - x^{(t)}}\| \\
\text{Dreiecksungl.} &\le \|x^{(t+m)} - x^{(t+m-1)}\| + \|x^{(t+m-1)} - x^{(t+m-2)}\| + \cdots + \|x^{(t+1)} - x^{(t)}\| \\
\text{(ii)} &\le q^{t+m-1}\|x^{(1)} - x^{(0)}\| + q^{t+m-2}\|x^{(1)} - x^{(0)}\| + \cdots + q^t\|x^{(1)} - x^{(0)}\| \\
\text{Ausklammern} &= (q^{t+m-1} + q^{t+m-2} + \cdots + q^t)\|x^{(1)} - x^{(0)}\| \\
\text{geom. Reihe} &\le q^t\frac{1-q^m}{1-q}\|x^{(1)} - x^{(0)}\| \le \varepsilon \text{ für } t \ge t(\varepsilon) \text{ hinreichend groß}
\end{aligned}$$

$\mathbb{R}^n$ ist vollständig, jede Cauchy-Folge konvergiert. Also existiert $z = \lim_{t\to\infty} x^{(t)}$ und $z \in G$, da G abgeschlossen. Schließlich ist $z = g(z)$. (Dies zeigt man so:

$$\|z - g(z)\| \overset{\text{t bel.}}{=} \|z - x^{(t)} + x^{(t)} - g(z)\| \le \|z - x^{(t)}\| + q\|\underbrace{x^{(t-1)} - z}_{\to 0 \text{ für } t\to\infty}\| \to 0$$

).

(iv) Fehlerabschätzung

$$\begin{aligned}\|x^{(t+m)} - x^{(t)}\| &\le \|x^{(t+m)} - x^{(t+m-1)}\| + \cdots + \|x^{(t+1)} - x^{(t)}\| \text{ (wie oben)}\\ &\le q^m\|x^{(t)} - x^{(t-1)}\| + \cdots + q\|x^{(t)} - x^{(t-1)}\|\\ &= \underbrace{(q^m + q^{m-1} + \cdots + q)}_{\text{absch. durch. geom. Reihe}}\|x^{(t)} - x^{(t-1)}\| \le \frac{q}{1-q}\|x^{(t)} - x^{(t-1)}\|\end{aligned}$$

Für $m \to \infty$ gilt $x^{(t+m)} \to z$ rechte Seite ist unabhänging von m, also

$$\|z - x^{(t)}\| \le \underbrace{\frac{q}{1-q}}_{\substack{\text{kann man benutzen um Abstand}\\ \text{zur exakten Lösung zu schätzen}}}\|x^{(t)} - x^{(t-1)}\| \le \underbrace{\frac{q}{1-q}q^{t-1}}_{=\frac{q^t}{1-q}}\|x^{(1)} - x^{(0)}\|$$

□

8.3 Iterationsverfahren zur Lösung linearer Gleichungssysteme

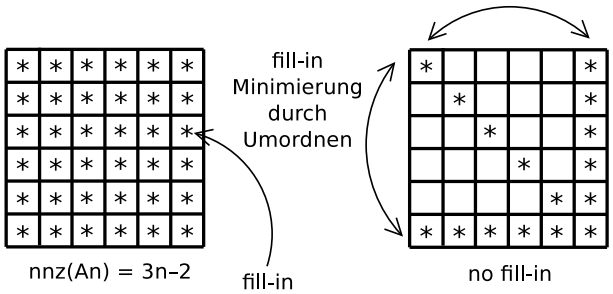
Wir kehren zurück zur Lösung von linearen Gleichungssystemen

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

Definition 8.4
Eine Menge von Matrizen $\{A^{(n)} \mid n \in \mathbb{N}\}$ heißt dünn besetzt, falls

$$\left|\left\{aij^{(n)} \;\middle|\; a_{ij}^{(n)} \neq 0\right\}\right| = nnz(A^{(n)}) = \mathcal{O}(n)$$

Gauß-Elimination ist für dünn besetze Matrizen oft schlecht geeignet aufgrund von fill-in.



Lösen von $Ax = b \Leftrightarrow$ „Nullstellensuche“ $f(x) = b - Ax = 0$
Definiere Iteration

$$x^{(t+1)} = g(x^{(t)}) = x^{(t)} + C^{-1}f(^{(t)}) = x^{(t)} + C^{-1}(b - Ax^{(t)}) = \underbrace{(I - C^{-1})}_{=:B \text{ „Iterationsmatrix“}}x^{(t)} + C^{-1}b$$

Für $x := A^{-1}b$ gilt

$$g(x) = (I - C^{-1}A)\underbrace{A^{-1}b}_{x} + C^{-1}b = A^{-1}b - C^{-1}b + C^{-1}b = A^{-1}b = x$$

also x Fixpunkt von g. Für die Lipschitzkonstante der Funktion g gilt

$$\|g(x) - g(y)\| = \|Bx + C^{-1}b - By - C^{-1}b\| = \|B(x-y)\| \le \|B\|\|x-y\|$$

Falls $\|B\| < 1$ ($\|\cdot\|$ verträgliche Matrixnorm) ist g Kontraktion auf $\mathbb{R}^n$.

Beispiele für Iterationsverfahren

Setze $A = L + D + U$ mit L strikte untere Dreiecksmatrix, D Diagonalmatrix, U obere Dreiecksmatrix.

$$\begin{array}{lll} C = D & \text{also } x^{(t+1)} = x^{(t)} + D^{-1} & (b - Ax^{(t)}) \text{ „Jacobi-Verfahren“} \\ C = L + D & \text{also } x^{(t+1)} = x^{(t)} + (L+D)^{-1} & (b - Ax^{(t)}) \text{ „Gauß-Seidel Verfahren“} \end{array}$$

Iterationsverfahren konvergieren in der Regel nur für bestimmte Klassen von Matrizen. Hier ein Beispiel:

Definition 8.5
Eine Matrix heißt *strikt diagonaldominant* falls

$$\sum_{j \neq i} |a_{ij}| < |a_{ii}| \quad \forall\, i = 1, \cdots, n$$

Beispiel: Splines, Radiosity-Verfahren

Satz 8.6
Das Jacobi-Verfahren konvergiert für strikt diagonaldominante Matrizen.

Beweis. $B = I - D^{-1}A$. Zeige $\|B\|_\infty < 1$ (Zeilensummennorm).

$$B = I - D^{-1}A = I - D^{-1}(L + D + U) = -D^{-1}(L + U)$$

$$\|B\|_\infty = \|D^{-1}(L+D)\|_\infty = \max_{i=1,\cdots,n} \left(\sum_{i \neq j} \left| \frac{a_{ij}}{a_{ii}} \right| \right) = \max_{i=1,\cdots,n} \frac{1}{|a_{ii}|} \underbrace{\sum_{i \neq j} |a_{ij}|}_{<|a_{ii}| \text{ n. Vor.}} < 1$$

□

Es gibt viele weitere solche Aussagen für symmetrisch positiv definite Matrizen, schwach diagonaldominanten Matrizen, M-Matrizen, ...
Aufwand für Iterationsfahren

1) Aufwand für eine Iteration $x^{(t+1)} = x^{(t)} + C^{-1}(b - Ax^{(t)}$ sei $\alpha(n)$. Typischerweise $\alpha(n) = \mathcal{O}(n)$.

2) $\|x^{(t)} - x\| \leq \|B\| \|x^{(t-1)} - x\|$ also
$\|x^{(t)} - x\| \leq \|B\|^t \|x^{(0)} - x\|$ brauche

$$\|B\|^t \leq \varepsilon \Leftrightarrow \underbrace{t \log \underbrace{\|B\|}_{<1}}_{<0} \leq \underbrace{\log \varepsilon}_{<1} \Leftrightarrow t \geq \underbrace{\frac{\log \varepsilon}{\log \|B\|}}_{>0}$$

Gesamtaufwand: $t \cdot \alpha(n) = \dfrac{\log \varepsilon}{log \|B\|} \alpha(n)$
$\|B\|$ problemabhängig, je nach Verfahren auch von n abhängig. Es gibt Verfahren, die relevante Probleme (z.B. Rohrmatrix) in *Gesamtaufwand* $\mathcal{O}(n)$ lösen können!

To Be Continued...