EINFÜHRUNG IN DIE NUMERIK

Vorlesungsmanuskript, Stand 29.01.2010

Wintersemester 2009/2010

Peter Bastian
Interdisziplinäreszentrum für Wissenschaftliches Rechnen
Universität Heidelberg, Im Neuenheimer Feld 368, D-69120 Heidelberg
Peter.Bastian@iwr.uni-heidelberg.de

INHALTSVERZEICHNIS

26.09.09

Kapitel 1: Grundbegriffe in der Numerik

1.1. Stabilität und Konditionszahl

Betrachte das Auswerten einer Funktionsvorschrift

$$\boxed{\text{Gegeben } x, \text{ berechne } y = F(x).} \qquad (1.1)$$

Wobei $F: X \to Y$ mit X, Y normierte Räume.

(Es geht darum F später näherungsweise im Rechner auszuwerten).

Frage: Wie wirkt sich Änderung in x im Ergebnis $F(x)$ aus?

M.a.W. welche Stetigkeit erfüllt F?

Definition 1.1 (Stabilität) Wir nennen die Auswertung der Funktion $F: X \to Y$

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

d.h. $\forall x \in X \; \exists \delta(x) > 0 \; \exists K_0(x) > 0 : \forall x' \in X : \|x - x'\|_X \leq \delta \Rightarrow \|F(x) - F(x')\|_Y \leq K_0(x) \|x - x'\|_X$

F heißt L-stabil, falls $\Leftrightarrow K_0(x) = K_0$ unabhängig von x gibt. L-stabil ist „stärker" als stabil, d.h. L-stabil $\Rightarrow$ stabil aber nicht umgekehrt.

Die Definition 1.1 b ist in der Praxis nützlicher, da sie eine Quantifizierung der Abhängigkeit $y = F(x)$ erlaubt:

$$\frac{\|\delta y\|_Y}{\|\delta x\|_X} = \frac{\|F(\overbrace{x + \delta x}^{= x'}) - F(x)\|_Y}{\|\delta x\|_X} \leq K_0(x)$$

(Betrag der Ableitung kann nicht unendlich sein).

Definition 1.2 (Konditionszahlen)

Die absolute Kondition einer Abbildung $F: X \to Y$ im Punkt x ist

$$K_{abs}(x) = \sup \left\{ \frac{\|F(x+\delta x) - F(x)\|_Y}{\|\delta x\|} \;\middle|\; \delta x \neq 0,\ x+\delta x \in X \right\}. \quad (1.2)$$

Entsprechend lautet die relative Kondition im Punkt x:

$$K(x) = \sup \left\{ \frac{\|F(x+\delta x) - F(x)\| / \|F(x)\|}{\|\delta x\| / \|x\|} \;\middle|\; \delta x \neq 0,\ x+\delta x \in X \right\} \quad (1.3)$$

(relative Abweichung)

wobei $x \neq 0$ und $F(x) \neq 0$ angenommen sei ☒

$F: X \to Y$ heißt

- gut konditioniert (in x) falls $K(x)$ „klein".
- schlecht " " " " „groß".

Erwartung: Gut konditionierte Vorschriften erlauben eine Berechnung auf dem Computer.

- Bei schlecht konditionierten Vorschriften ist mit „Problemen" zu rechnen. Und zwar unabhängig vom genauen numerischen Verfahren!

Beispiel 1.3 (Kondition der Addition)

Betrachte $y = F(x_1, x_2) = x_1 + x_2$.

Mit $x = (x_1, x_2)^T$ und $\delta x = (\delta x_1, \delta x_2)^T$ lautet die Taylorreihe

$$F(x+\delta x) = F(x) + \nabla F(x) \cdot \delta x + O(\|\delta x\|^2)$$

- Restglied
- Landausche Symbole

$$\Rightarrow \quad |F(x+\delta x) - F(x)| \leq |\nabla F(x) \cdot \delta x| + O(\|\delta x\|^2)$$

$$\xrightarrow{\text{Cauchy-Schwarz}} \leq \|\nabla F(x)\| \, \|\delta x\| + O(\|\delta x\|^2)$$

$$\Leftrightarrow \quad \frac{|F(x+\delta x) - F(x)|}{\|\delta x\|} \leq \|\nabla F(x)\| + O(\|\delta x\|)$$

Nun ist $\nabla F(x) = (1, 1)^T$ und somit

$$K_{abs}(x) \doteq \|(1,1)^T\| = \sqrt{2} \qquad \left(\text{für } \|.\| = \|.\|_2 \text{ euklidsche Norm}\right)$$

↑ „in erster Näherung"

geht auch für andere Normen

Für die relative Kondition gilt

$$K(x) \doteq \|(1,1)^T\| \, \frac{\|x\|}{|x_1 + x_2|}$$

Für $x_1 \approx -x_2$ gilt $|x_1 + x_2| \ll \|x\| = \sqrt{x_1^2 + x_2^2}$.

$\to 0$ $\qquad$ $\to \sqrt{2}\,|x_1|$

Somit ist die Addition schlecht konditioniert für $x_1 \approx -x_2$. ∎

Beispiel 1.4

Wir betrachten die Lösung der quadratischen Gleichung

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

Quadratische Ergänzung liefert

$$\begin{aligned} & x^2 - 2px + p^2 = p^2 - 1 \\ \Leftrightarrow \quad & (x-p)^2 = p^2 - 1 \\ \Leftrightarrow \quad & x_{1,2} = p \pm \sqrt{p^2-1} \end{aligned}$$

Uns interessiert die Vorschrift

$$F(p) = \begin{pmatrix} p+\sqrt{p^2-1} \\ p-\sqrt{p^2-1} \end{pmatrix} \quad \text{mit} \quad p \in [1, \infty)$$

Taylor:

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

$$\| F(p+\delta p) - F(p) \| \leq \left\| \frac{dF}{dp}(p) \right\| |\delta p| + O(|\delta p|^2)$$

$$\Leftrightarrow \quad \frac{\| F(p+\delta p) - F(p) \|}{|\delta p|} \leq \left\| \frac{dF}{dp}(p) \right\| + O(|\delta p|)$$

euklidische Norm:

$$\left\| \frac{dF}{dp} \right\|_2 = \sqrt{\left(1+\frac{p}{\sqrt{p^2-1}}\right)^2 + \left(1-\frac{p}{\sqrt{p^2-1}}\right)^2}$$

$$= \sqrt{1 + \cancel{\frac{p}{\sqrt{p^2-1}}} + \frac{p^2}{p^2-1} + 1 \cancel{- \frac{p}{\sqrt{p^2-1}}} + \frac{p^2}{p^2-1}}$$

$$= \sqrt{2\,\frac{p^2-1+p^2}{p^2-1}} = \sqrt{2}\sqrt{\frac{2p^2-1}{p^2-1}}$$

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

$\|.\|$ euklidische Norm:

$$K_{abs}(p) \doteq \sqrt{2}\sqrt{\frac{2p^2-1}{p^2-1}} \quad \text{schlecht konditioniert für } p=1$$

$$K(p) \doteq \left\| \frac{dF}{dp}(p) \right\| \frac{|p|}{\|F(p)\|} \overset{\|.\| = \|.\|_2}{=} \sqrt{2}\sqrt{\frac{2p^2-1}{p^2-1}} \frac{|p|}{\sqrt{\underset{\to 1}{\left[p+\sqrt{p^2-1}\right]^2} + \underset{\to 1}{\left[p-\sqrt{p^2-1}\right]^2}}}$$

ebenfalls schlecht konditioniert für $p \to 1$.

Übung: Kondition der Lösungsformel von $x^2 - \frac{1+t^2}{t}x + 1$ $t \in [1,\infty]$

$\frac{1+t^2}{t}$ bildet $[1,\infty]$ bijektiv auf sich selbst ab.

1.2 Numerische Auswertung von Funktionsvorschriften

Im Rechner kann die Auswertung $y = F(x)$ nicht exakt realisiert werden.

Stattdessen realisieren wir dort für $k \in \mathbb{N}$

$$y^{(k)} = F^{(k)}(x^{(k)}) \quad \text{mit} \quad F^{(k)} : X^{(k)} \to Y^{(k)}, \tag{1.4}$$

mit normierten Räumen $X^{(k)}, Y^{(k)}$.

Beispiel 1.5

(a)
$$\begin{cases} F^{(k)} : \left(\mathbb{F}(10,k,2)\right)^2 \to \mathbb{F}(10,k,2) \\ F^{(k)}(x_1^{(k)}, x_2^{(k)}) = x_1^{(k)} \oplus x_2^{(k)} \\ F(x_1,x_2) = x_1 + x_2 \end{cases}$$

$k \to \infty$ meint immer mehr Stellen in Dezimaldarstellung.

$\mathbb{F}(10,k,3)$: Zahlen der Form

$$\pm 0,\underbrace{m_1 \dots m_k}_{0 \le m_i < 10} \cdot 10^{\pm e}$$

m_i: Ziffern der Mantisse; e: Exponent mit 2 Stellen

(b) $F^{(k)} : \mathbb{R} \to \mathbb{R}$

$$F^{(k)}(x) = 1 + \sum_{i=1}^{k} \frac{x^k}{k!}$$ abgebr. Reihe für e-Funktion.

$F(x) = \exp(x)$

Hier gilt $X^{(k)} \subseteq X$, $Y^{(k)} \subseteq Y$. Das muss aber nicht so sein.

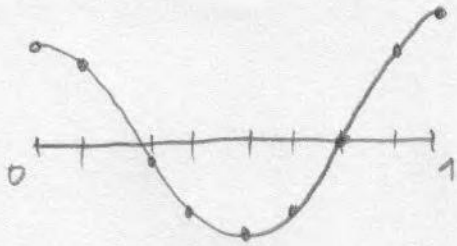
Im folgenden gehen wir davon aus, dass $X^{(k)} \subseteq X$ aber $Y^{(k)} \not\subseteq Y$.

Es gebe aber eine Abbildung $R^{(k)} : Y \to Y^{(k)}$ mit

$R^{(k)}$ linear und lokal L-stabil

$$R^{(k)}(x+x') = R^{(k)}(x) + R^{(k)}(y)$$
$$R^{(k)}(\alpha x) = \alpha R^{(k)}(x)$$

Beispiel 1.6 $Y = C([0,1])$, $Y^{(k)} = \mathbb{R}^{k+1}$



$$(R^{(k)} y)_i = y\left(\frac{i}{k}\right) \qquad 0 \le i \le k.$$

Frage: $x^{(k)} \to x$, geht dann auch $F^{(k)}(x^{(k)}) \to F(x)$?

Mathematisch: Konvergenz der Folge $\{F^{(k)}(x^{(k)})\}$

Definition 1.7 (Konvergenz) Die numerische Auswertung (1.4) heißt konvergent, genau dann wenn für jedes $x \in X$ gilt:

$$\forall \varepsilon > 0 \; \exists k_0(\varepsilon) \in \mathbb{N} \; \exists \delta(\varepsilon, k_0) > 0 :$$

$$\forall k > k_0, \forall x' \in X^{(k)} : \quad \|x - x'\| \le \delta \Rightarrow \underbrace{\|R^{(k)} F(x) - F^{(k)}(x')\|}_{\text{dann ist auch } F(x) - F^{(k)}(x') \text{ klein}} \le \varepsilon$$

Aus Def. 1.7 folgt auch die Konvergenz

$$\|R^{(k)} F(x') - F^{(k)}(x')\| \to 0 \quad \text{für } k \to \infty \text{ und } x' \in X^{(k)}. \tag{1.5}$$

(gleiches Argument!)

Denn sei $x' \in X^{(k)}$, $x \in X$ beliebig

$$\|R^{(k)} F(x') - F^{(k)}(x')\| = \|R^{(k)} F(x') - R^{(k)} F(x) + R^{(k)} F(x) - F^{(k)}(x')\|$$

Dreiecksungl. $\le$ $\underbrace{\|R^{(k)} F(x') - R^{(k)} F(x)\|}_{\text{nur } F!}$ + $\underbrace{\|R^{(k)} F(x) - F^{(k)}(x')\|}_{\le \frac{\varepsilon}{2} \text{ wenn } k > k_0(\frac{\varepsilon}{2}),\ \|x - x'\| \le \delta(\frac{\varepsilon}{2}, k_0)}$

$\le R_0 K_0(x) \|x - x'\|$ wenn F lokal L-stabil

also $\le \frac{\varepsilon}{2}$ falls $\|x - x'\| \le \frac{\varepsilon}{2 R_0 K_0}$

Gilt umgekehrt die Konvergenz für $\|R^{(k)} F(x') - F^{(k)}(x')\|$ für $x' \in X^{(k)}$ und $k \to \infty$, dann

$$\|R^{(k)} F(x) - F^{(k)}(x')\| \le \|R^{(k)} F(x) - R^{(k)} F(x')\| + \|R^{(k)} F(x') - F^{(k)}(x')\|$$

(einschieben)

Bed. in 1.7 ↑ — wieder Stabilität von F — Konvergenz zweiter Art

Somit sind die beiden Bedingungen aus Def. 1.7 und Gl. (1.5) äquivalent.

Bemerkung 1.9

+
26.09.09

Für die Praxis ist diese Aufspaltung praktisch, da

- $\| R^{(k)}F(x) - R^{(k)}F(x') \| \leq \|R^{(k)}\| \, \|F(x) - F(x')\|$
  nur die Stabilität der Auswertung F ist
- $\| R^{(k)}F(x') - F^{(k)}(x') \|$ den Rundungs- oder Abbruchfehler
  darstellt (siehe etwa Beispiel 1.5).

Bisher haben wir die Stabilität der $F^{(k)}$ nicht gefordert. Es zeigt sich, dass diese eine notwendige Voraussetzung für Konvergenz ist.

Wir zeigen: Konvergenz $F^{(k)} \to F \Rightarrow$ Stabilität der $F^{(k)}$

damit gilt dann: ¬ Stabilität $\Rightarrow$ ¬ Konvergenz.

Satz 1.8 Es sei F stabil und $F^{(k)}$ konvergent nach Definition 1.7.
Dann sind die $F^{(k)}$ stabil. (nicht lokal L-stabil!)

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

$$\| F^{(k)}(x) - F^{(k)}(x') \| = \| \underbrace{F^{(k)}(x) - R^{(k)}F(x)} + \underbrace{R^{(k)}F(x) - F^{(k)}(x')} \|$$

$$\leq \| R^{(k)}F(x) - F^{(k)}(x) \| + \| R^{(k)}F(x) - F^{(k)}(x') \|$$

Konvergenz nach Gl. (1.5) — Konvergenz wie in Def. 1.7.

Ist nun $k > k_0(\frac{\varepsilon}{2})$ und $\|x - x'\| \leq \delta(\frac{\varepsilon}{2}, k_0)$ (solche k_0 und δ existieren nach Def. 1.7)
so sind beide Terme jeweils kleiner als $\frac{\varepsilon}{2}$ und $F^{(k)}$, $k > k_0$, ist stabil. ■

1.3 Lösen von Gleichungen

24.09.09

Bisher haben wir nur die Auswertung von Funktionsvorschriften betrachtet.

Oft ist die unbekannte Größe y nur implizit durch eine Gleichung bestimmt:

> Gegeben $x \in X$ („Daten"), finde $y \in Y$ („Lösung") so dass
>
> $$G(y, x) = 0 \tag{1.6}$$

Definition 1.10 (Sachgemäß gestellt).

Das Problem (1.6) heißt sachgemäß gestellt falls gilt:

(a) Zu jedem $x \in X$ gibt es genau eine Lösung $y \in Y$.

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

Die Funktion $F: X \to Y$ mit

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

heißt Lösungsoperator zur Gleichung (1.6).

Sachgemäß gestellt bedeutet also, dass genau ein solches F existiert und F stabil ist.

$\neg$(sachgemäß gestellt) = schlecht gestellt.

Beispiel 1.11 Sei $x \in \mathbb{R}^n$ und $A \in \mathbb{R}^{n \times n}$ invertierbar.

Setze $G(y, x) = Ay - x$. Dann ist $F(x) = A^{-1}x$.

Im Rechner können wir (1.6) nur näherungsweise zu gestörten Daten lösen. Für $k \in \mathbb{N}$ betrachte die Probleme

$$\boxed{\text{Gegeben } x^{(k)} \in X^{(k)}, \text{ finde } y^{(k)} \in Y^{(k)} \text{ sodass } G^{(k)}(y^{(k)}, x^{(k)}) = 0} \quad (1.7)$$

Zu jedem $G^{(k)}$ gebe es einen numerischen Lösungsoperator $F^{(k)} : X^{(k)} \to Y^{(k)}$:

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

Wie oben sei $X^{(k)} \subseteq X$ und $R^{(k)} : Y \to Y^{(k)}$ da $Y^{(k)} \not\subseteq Y$ zulässig ist.

Wie oben gezeigt, ist die Stabilität der $F^{(k)}$ eine notwendige Voraussetzung für die Konvergenz. $R^k y - y^{(k)} \to 0$.

Wir skizzieren nun eine einfache Konvergenztheorie, die in der Praxis häufig bei gewöhnlichen und partiellen DGL Einsatz findet

Zentral ist dabei

Definition 1.12 (Konsistenz)

Das numerische Verfahren (1.7) heißt konsistent zu (1.6) falls gilt

$$G^{(k)}(R^{(k)} y, x + \delta x^{(k)}) \to 0 \text{ für } k \to \infty, \delta x^{(k)} \to 0. \quad \blacksquare$$

Bei vielen praktischen Problemen lässt sich diese Eigenschaft leicht nachweisen. Siehe Beispiel.

Die Funktion $G^{(k)} : Y^{(k)} \times X^{(k)} \to Z^{(k)}$ sei invertierbar bezüglich dem ersten Argument. D.h. es existiere $H^{(k)} : Z^{(k)} \times X^{(k)} \to Y^{(k)}$ so dass

$$\forall x \in X^{(k)}, \forall y \in Y^{(k)}: \quad H^{(k)}\Big(\underbrace{G^{(k)}(y,x)}_{z}, x\Big) = y \qquad (1.8)$$

Definition 1.13

Das numerische Verfahren (1.7) heißt stabil falls $H^{(k)}$ stabil bezüglich dem ersten Argument ist.

Damit zeigt man den zentralen

Satz 1.14 Das numerische Verfahren (1.7) sei konsistent und stabil. Dann ist es auch konvergent.

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

$$\|R^{(k)}y - y^{(k)}\| = \| \underbrace{H^{(k)}\Big(G^{(k)}\big(R^{(k)}y, x+\delta x^{(k)}\big), x+\delta x^{(k)}\Big)}_{= R^{(k)}y \text{ nach (1.8)}} - \underbrace{H^{(k)}\Big(G^{(k)}\big(y^{(k)}, x+\delta x^{(k)}\big), x+\delta x^{(k)}\Big)}_{= y^{(k)}} \|$$

$$\leq K_0 \, \| G^{(k)}\big(R^{(k)}y, x+\delta x^{(k)}\big) - \underbrace{G^{(k)}\big(y^{(k)}, x+\delta x^{(k)}\big)}_{=0 \text{ n. Vor.}} \|$$

(Stabilität von $H^{(k)}$)

$$= K_0 \, \underbrace{\| G^{(k)}\big(R^{(k)}y, x+\delta x^{(k)}\big) \|}_{\to 0 \text{ wg. Konsistenz.}}$$

Bemerkung 1.14

. Der Begriff „Stabilität" tritt in der Numerik häufig auf. Er meint immer einen Zusammenhang der Form $\|f(x) - f(x')\| \leq K(x)\|x - x'\|$. Was f genau ist variiert von Problem zu Problem.

Beispiel 1.15 (Anfangswertproblem) Wir zeigen, dass die abstrakte Definition eine Anwendung haben! 24.09.09

Betrachte die DGL

$$y'(t) = f(t, y(t)) \quad \text{für} \quad t \in I = [0, T],$$
$$y(0) = x \qquad \text{(Anfangswert)}. \tag{1.9}$$

Hier ist $Y = C^1(I)$, $X = \mathbb{R}$, $Z = \mathbb{R}$ und eine mögliche Wahl für G:

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

(Das schreibt man in der Praxis nicht so hin).

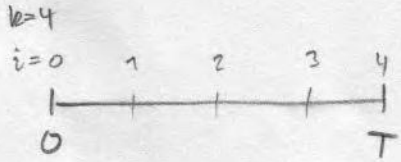
↝ alternative Normen möglich
↝ bietet sich an wegen $y \in C^1(I)$

Lösungsoperator $F : X \to Y$ liefert zu gegebenem Anfangswert x die Lösung.

Wir setzen voraus, dass f in (1.9) derart, dass (1.9) sachgemäß gestellt.

Numerik: „Expliziter Euler". Sei $k \in \mathbb{N}$.

Setze $\Delta t^{(k)} = \frac{T}{k}$ und $t_i^{(k)} = i \cdot \Delta t^{(k)}$, $0 \le i \le k$.



Taylorentwicklung der Lösung von (1.9): Für $i > 0$

$$y(t_i^{(k)}) = y(t_{i-1}^{(k)} + \Delta t^{(k)}) = y(t_{i-1}^{(k)}) + \Delta t^{(k)} y'(t_{i-1}^{(k)}) + \frac{(\Delta t^{(k)})^2}{2} y''(\xi_i^{(k)}), \quad \xi_i^{(k)} \in [t_{i-1}^{(k)}, t_i^{(k)}].$$

(1.9) Einsetzen ↓

$$= y(t_{i-1}^{(k)}) + \Delta t^{(k)} f(t_{i-1}^{(k)}, y(t_{i-1}^{(k)})) + \frac{(\Delta t^{(k)})^2}{2} \underbrace{\left[\partial_t f + (\partial_y f) f\right]}_{\text{kann man ausrechnen}} (\xi_i^{(k)}, y(\xi_i^{(k)}))$$

Umstellen: Für die Lösung $y(t)$ von (1.9) gilt für $0 < i \le k$

$$\frac{1}{\Delta t^{(k)}} \left(y(t_i^{(k)}) - y(t_{i-1}^{(k)}) \right) - f(t_{i-1}^{(k)}, y(t_{i-1}^{(k)})) = O(\Delta t^{(k)}) \tag{1.10}$$

$\to 0$ für $k \to \infty$
„konsistent von erster Ordnung"

Idee: Betrachte (1.10) mit rechter Seite 0 als Definitionsgleichung für $y_i^{(k)} \approx y(t_i^{(k)})$

Formal: $Y^{(k)} = \mathbb{R}^{k+1}$, $X^{(k)} = \mathbb{R}$, $Z^{(k)} = \mathbb{R}^{k+1}$

$G^{(k)} : Y^{(k)} \times X^{(k)} \to Z^{(k)}$ mit den einzelnen Komponenten:

$$G_i^{(k)}\left(Y_0^{(k)}, Y_1^{(k)}, \dots, Y_k^{(k)}, X\right) = \begin{cases} Y_0^{(k)} - x & i = 0 \\ \frac{1}{\Delta t^{(k)}}\left(Y_i^{(k)} - Y_{i-1}^{(k)}\right) - f\left(t_{i-1}^{(k)}, Y_{i-1}^{(k)}\right) & 0 < i \le k \end{cases} \quad (1.11)$$

Der Lösungsoperator $F^{(k)} : X^{(k)} \to Y^{(k)}$ lautet („vorwärts einsetzen")

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

rekursive Definition

Zur Konsistenz: $R^{(k)} : Y \to Y^{(k)}$ also $C^1([0,T]) \to \mathbb{R}^{k+1}$

mit $\left(R^{(k)} y\right)_i = y\left(t_i^{(k)}\right) \quad 0 \le i \le k$

somit ist $G^{(k)}\left(R^{(k)} y, x\right) = O\left(\Delta t^{(k)}\right)$ wegen (1.10).

Damit ist das Verfahren konsistent nach Definition (1.12).

Zur Stabilität: Was ist H aus (1.8)? $H^{(k)}(z^{(k)}, x) = ?$

Für gegebenes $z^{(k)} \in \mathbb{R}^{k+1}$ löse $G^{(k)}(y^{(k)}, x) = z^{(k)}$ nach $y^{(k)}$ auf.

Nach (1.11)

$$\left.\begin{aligned} Y_0^{(k)} - x &= z_0^{(k)} \\ \frac{1}{\Delta t}\left(Y_i^{(k)} - Y_{i-1}^{(k)}\right) - f\left(t_{i-1}^{(k)}, Y_{i-1}^{(k)}\right) &= z_i^{(k)} \end{aligned}\right\} \Rightarrow z_i^{(k)} = H_i^{(k)}\left(z^{(k)}, x\right) = \begin{cases} x + z_0^{(k)} \\ H_{i-1}^{(k)}\left(z^{(k)}, x\right) \\ \quad + \Delta t^{(k)} f\left(t_{i-1}^{(k)}, H_{i-1}^{(k)}\left(z^{(k)}, x\right)\right) \\ \quad + \Delta t^{(k)} z_i^{(k)} \end{cases}$$

Man zeigt: $f(x,y)$ Lipschitzstetig in y, dann ist **H** stabil.

Und damit hat man Konvergenz. Weiteres in Numerik 1.

Kapitel 2 : Fließkommazahlen

Im Computer gibt es verschiedene Typen zur Repräsentation von Zahlen. Etwa in C/C++ :

unsigned int	$\mathbb{N}_0$
int	$\mathbb{Z}$
float	$\mathbb{R}$
double	$\mathbb{R}$
complex<double>	$\mathbb{C}$

int : exakt, aber endlicher Bereich

float, double, .. : Approximation, endlicher Bereich

Was hat das für Folgen?

Beispiel 2.1 (Potenzreihe für e^x). Auf Folie

2.1 Zahlendarstellung

Stellenwertsystem.

$$x = \pm \ldots m_n \beta^n + \ldots + m_1 \beta^1 + m_0 + m_{-1} \beta^{-1} + \ldots + m_{-k} \beta^{-k} + \ldots$$

$\pm$ ist das Vorzeichen (ein Bit genügt).

$\beta \in \mathbb{N}, \beta \geq 2$ heißt Basis.

$m_i \in \{1, 2, \ldots, \beta - 1\}$ heißen Ziffern.

Jede reelle Zahl $x \in \mathbb{R}$ kann mit unendlich vielen Ziffern dargestellt werden.

Geschichte: Siehe [Knuth, Band 2, p. 194]

Babylonier 1750 v. Chr : $\beta = 60$

Basis 10 in Europa ab ca 1585

Blaise Pascal : Jedes $\beta \geq 2$ möglich

Festkommazahlen:

Wähle maximalen und minimalen Exponenten:

$$x = (-1)^s \sum_{i=-k}^{n} m_i \beta^i$$

Problem: Wissenschaftl. Anwendungen brauchen Zahlen sehr unterschiedlicher Größe

Plancksches Wirkungsquantum: $6.6260693 \cdot 10^{-34}\ Js$

Ruhemasse Elektron: $9.11 \cdot 10^{-28}\ g$

Avogadrokonstante: $6.021415 \cdot 10^{23}\ \frac{1}{mol}$

$\Rightarrow \beta = 2 \quad 2^{n+k} \approx 10^{23+34} \rightarrow n+k = \frac{57}{\log_2} \approx \underline{190\ \text{Stellen}}!$

Fließkommazahlen erlauben effiziantere Darstellung unterschiedlich großer Zahlen.

Definition 2.2 (normierte Fließkommazahlen)

Sei $\beta, r, s \in \mathbb{N}$ 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 $x = m(x)\,\beta^{e(x)}$ mit

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

m heißt Mantisse, e Exponent.

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

$m_1 \neq 0$ heißt Normierung. Bedingung b) macht die Fließkommadarstellung eindeutig. ∎

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

$$\beta^{-1} \leq |m(x)| < 1 \quad \text{und damit} \quad \beta^{e(x)-1} \leq |x| < \beta^{e(x)}. \tag{2.1}$$

Beispiel 2.3

a) $\mathbb{F}(10,3,1)$ besteht aus Zahlen der Form:

$$x = \pm \left(m_1 \cdot 0.1 + m_2\, 0.01 + m_3 \cdot 0.001\right) \cdot 10^{\pm e_0}$$

mit $m_1 \neq 0 \vee (m_1 = m_2 = m_3 = 0)$

z.B. $0.999 \cdot 10^{7}$, $0.123 \cdot 10^{-1}$, 0, aber

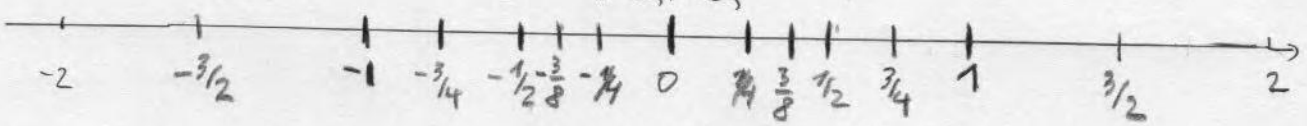
$0,\underbrace{000}\,\underbrace{000}\,\underbrace{000}\,0.14 = 0.140 \cdot 10^{-10} \notin \mathbb{F}(10,3,1)$ da Exponent zu klein.

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

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

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

Lücke wg. Normalisierung.



Mögliche Lösg: Nehme für $|x| < \beta^{-1}$ die nicht normalisierten Zahlen hinzu.

Zahlenbereich:

Größte / kleinste darstellbare Zahl in $\mathbb{F}(\beta, r, s)$

$$X_{+/-} = \pm \underbrace{(\beta-1)\left\{\beta^{-1} + \dots + \beta^{-r}\right\}}_{= 1 - \beta^{-r}} \cdot \beta^{\overbrace{(\beta-1)\{\beta^{s-1} + \dots + \beta^0\}}^{= \beta^s - 1}}$$

größte Ziffer; β^{-r}: Wertigkeit der letzten Stelle

$$= \pm \left(1 - \beta^{-r}\right) \beta^{\beta^s - 1}.$$

Kleinste positive / größte negative Zahl in $\mathbb{F}(\beta, r, s)$:

$$X_{+/-} = \pm \underbrace{\beta^{-1}}_{\text{kleinste Mantisse wg. Normierg}} \cdot \beta^{\overbrace{(-)(\beta-1)\{\beta^{s-1} + \dots + \beta^0\}}^{\beta^s - 1}} = \pm \beta^{\beta^s}.$$

(− : negativ!)

Damit

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

Abstände zwischen Fließkommazahlen

04.10.09

Festkommazahlen: Abstand zwischen Zahlen konstant β^{-k}

Fließkommazahlen: Abstand von $x \in \mathbb{F}$ zum nächsten $x' \in \mathbb{F}$ hängt von x ab.

(Vorzeichen spielt keine Rolle)

$$\beta^{e-1} = 0.1\overset{m_1}{0}\ldots\overset{m_r}{0} \cdot \beta^e$$

nächste Zahl: $0.10\ldots1 \cdot \beta^e$ } Abstand $\beta^{-r}\cdot\beta^e = \underline{\beta^{e-r}}$

$\vdots$

$0.\beta\beta\ldots\beta \cdot \beta^e$ } Abstand β^{e-r} (da Exponent immer noch gleich)

nächste Zahl $+\beta^{e-r}$ ↓

} Abstand β^{e-r} ⇒ Abstand springt!

$\beta^e = 0.10\ldots0 \cdot \beta^{e+1}$

nächste Zahl: $0.10\ldots1 \cdot \beta^{e+1}$ } Abstand $\beta^{-r}\cdot\beta^{e+1} = \underline{\beta^{e+1-r}}$

⇒ Im Intervall $[\beta^{e-1}, \beta^e]$ beträgt der absolute Abstand zwischen zwei Zahlen β^{e-r}.

Sei $x' \in \mathbb{F}$ die nächstliegende Fließkommazahl zu $x \in \mathbb{F}$. Dann gilt

$$|x-x'| = \beta^{e(x)-r} = \frac{|m(x)|\,\beta^{e(x)}}{|m(x)|}\cdot\beta^{-r} = \frac{|x|}{|m(x)|}\,\beta^{-r}$$

($\uparrow$ $x \neq 0$)

Da $\beta^{-1} \leq |m(x)| < 1$ folgt (groß, wenn $|m(x)|$ klein, d.h. β^{-1})

$$|x|\beta^{-1}\beta^{1-r} < |x-x'| \leq |x|\beta^{1-r}$$

Für den relativen Abstand gilt:

$$\beta^{-1}\beta^{1-r} < \frac{|x-x'|}{|x|} \leq \beta^{1-r}$$

Er schwankt um den Faktor β^{-1} je nach Größe von $|x|$. Dieser Faktor heißt **wobble**.

Konsequenz: Kleine β sind besser, β^{-1} dann kleiner ist!

Übung:

β^{1-r}: Die kleinste Zahl, $x \in \mathbb{F}$ so dass $1+x \neq 1$.

Schreibe Programm, dass dieses x bestimmt.

Teste das für float, double.

IEEE 754 / IEC559 Standard

Ziel: Portabilität von Programmen mit Fließkommazahlenarithmetik.

Verabschiedet 1985.

$\beta = 2$ mit vier Genauigkeitsstufen und normierter Darstellung:

	Format			
	single	single-ext	double	double-ext
e_{min}	+127	≥ 1024	1023	≥ 16383
e_{max}	-126	≤ -1021	-1022	≤ -16381
Bits Expon	8	≤ 11	11	15
Bits total	32	≥ 43	64	≥ 79

Betrachte double genauer:

- total 64 bit
- 11 bit für Exponent. Dieser wird vorzeichenlos als Zahl $c \in [1, 2046]$ gespeichert.

  Setze $e = c - 1023 \Rightarrow e \in [\underbrace{-1022}_{2046-1023}, \underbrace{1023}_{1-1023}]$

  Die Werte $c \in \{0, 2047\}$ werden anderweitig genutzt:

  $c = 0 \wedge m = 0$ kodiert die Null
  $c = 2047 \wedge m \neq 0$ kodiert NaN = „not a number" (z.B. Division durch Null)
  $c = 2047 \wedge m = 0$ kodiert ∞ (Überlauf).
- $64 - 11 = 53$ Bit Mantisse, 1 Bit für Vorzeichen, bleiben 52 Nachkommastellen.

  Wg $\beta = 2$ und Normierung ist immer $m_1 = 1$ und diese Ziffer wird nicht gespeichert

  Heißt hidden bit.

  Somit gilt $\boxed{r = 53}$.

2.2 Runden und Rundungsfehler

Um $x \in \mathbb{R}$ im $\mathbb{F}(\beta, r, s)$ zu approximieren brauchen wir

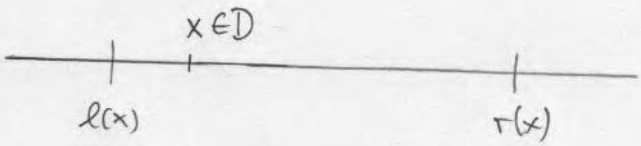
$$rd : \mathbb{D}(\beta, r, s) \longrightarrow \mathbb{F}(\beta, r, s) \qquad (2.2)$$

Achtung: rd setzt voraus, dass x im darstellbaren Bereich liegt!
Im Falle eines Über/Unterlaufes ist r, s zu ändern.

Sinnvollerweise soll für rd gelten:

$$|x - rd(x)| \leq \min_{y \in \mathbb{F}} |x - y| \quad \forall x \in \mathbb{D} \quad \text{(Bestapproximation)}.$$

Mit



$$l(x) = \max\{y \in \mathbb{F} \mid y \leq x\} \qquad r(x) = \min\{y \in \mathbb{F} \mid y \geq x\}$$

gilt dann

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

In letzten Fall ist eine Rundung erforderlich. Dafür gibt es verschiedene Möglichkeiten.

Sei $x = sign(x) \cdot \left(\sum_{i=1}^{\infty} m_i \beta^{-i}\right) \beta^e$ die normierte Darstellung von $x \in \mathbb{D}$.

Natürliche Rundung (das was jeder kennt)

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

(m_{r+1}: nächste Stelle)

(β^{e-r}: Wertigkeit der letzten Stelle)

Gerade Rundung (β sei gerade)

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

Damit ist $m_r\, rd(x)$ immer gerade, wenn gerundet wurde (d.h. $|x-l(x)| = |x-r(x)|$)

- ist $rd(x) = l(x)$ so ist das per Def so
- sonst ist $rd(x) = \underbrace{l(x)}_{m_r \text{ ungerade}} + \underbrace{\beta^{e-r}}_{+1 \text{ in der letzten Stelle}}$. $\Rightarrow m_r$ gerade.

Diese Wahl vermeidet eine mögliche Drift beim Aufrunden.

Definition 2.4 (absoluter und relativer Fehler)

Sei $x' \in \mathbb{R}$ eine Näherung von $x \in \mathbb{R}$. 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:

$$x' = x + \Delta x = x\left(1 + \frac{\Delta x}{x}\right) = x\left(1 + \varepsilon_{x'}\right)$$

(x': abs. Fehler)

Motivation:

Es sei $\Delta x = x' - x = 100\,km$.

Für x = Entfernung Erde-Sonne $\approx 1.5 \cdot 10^8 km$ ist

$$\varepsilon_{x'} = \frac{10^2\,km}{1{,}5 \cdot 10^8 km} \approx 6.6 \cdot 10^{-7}$$

relativ klein. Für x = Entfernung Heidelberg-Paris $\approx 500\,km$ ist

$$\varepsilon_{x'} = \frac{100\,km}{500\,km} = 0.2$$

dagegen relativ groß.

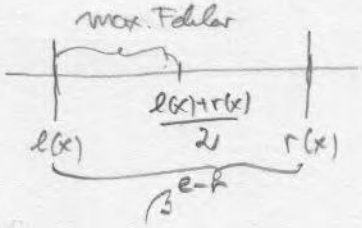
Lemma 2.5 (Rundungsfehler)

Bei der Rundung in $F(\beta, r, s)$ gilt für den absoluten Fehler

$$|x - rd(x)| \leq \frac{1}{2}\beta^{e(x)-r} \qquad (2.2)$$

und für den relativen Fehler $(x \neq 0)$

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



Beweis: Es ist $x = m(x)\beta^{e(x)}$ die normierte Darstellung von x.

Wie oben gezeigt gilt

$$|r(x) - \ell(x)| = \beta^{e(x)-r}$$

zwei aufeinander folg. Zahlen in F, $\ell(x) \in [\beta^{e(x)-1}, \beta^{e(x)}]$

Der maximale Fehler ergibt sich für $x = \frac{\ell(x)+r(x)}{2}$, also

$$|x - rd(x)| \leq \left| \frac{\ell(x)+r(x)}{2} - \ell(x) \right| = \frac{1}{2}|r(x) - \ell(x)| = \frac{1}{2}\beta^{e(x)-r}.$$

(egal)

Für den relativen Fehler $(x \neq 0)$ gilt

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

∎

Die Zahl $eps := \frac{1}{2}\beta^{1-r}$ heißt Maschinengenauigkeit. (Nennt Rannacher auch so).
Wird oft auch mit $\mathbb{u}$ (z.B. in MATLAB) abgekürzt.

Vorsicht: Die Bezeichnungen gehen in der Literatur durcheinander.

[Goldberg] nennt $\frac{1}{2}\beta^{1-r}$ machine epsilon.

[Quarteroni] nennt β^{1-r} machine epsilon und

$\frac{1}{2}\beta^{1-r}$ machine precision.

2.3 Fließkommaarithmetik

05.10.09

Wir benötigen eine Arithmetik auf $\mathbb{F}$:

$$\circledast : \mathbb{F} \times \mathbb{F} \to \mathbb{F} \quad \text{mit } \circledast \in \{\oplus, \ominus, \odot, \oslash\},$$

die den bekannten Operationen $* \in \{+, -, \cdot, /\}$ auf $\mathbb{R}$ entsprechen.

Problem: $x, y \in \mathbb{F} \Rightarrow x \circledast y \notin \mathbb{F}$ in der Regel!

Somit ist das Ergebnis $x \circledast y$ wieder zu runden, d.h. wir definieren

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

Man sagt $\circledast$ ist „exakt gerundet". Dies ist nicht trivial!
(implizite Annahme: $x * y \in \mathbb{D}$!)

Beispiel 2.6 (Guard digit)

Sei $\mathbb{F} = \mathbb{F}(10,3,1)$, $x = 0.215 \cdot 10^8$, $y = 0.125 \cdot 10^{-5}$. Wir betrachten die Subtraktion $x \ominus y = rd(x-y)$.

① $x-y$:

$x = 0.215\,000\,000\,000\,000\,0 \cdot 10^8$

$y = 0.000\,000\,000\,000\,012\,5 \cdot 10^8$ ← Schiebe y auf gleichen Exponenten.

13 Nullen! ... 1 1

$x-y = 0.214\,999\,999\,999\,987\,5 \cdot 10^8$

$y = 0.125 \cdot 10^{-5} = 0.125 \cdot \underbrace{10^{-13}}_{=10^{-5}} \cdot 10^8$

② Runden

$$x \ominus y = rd(0.2149\ldots \cdot 10^8) = 0.215$$

Problem: Schritt 1 erfordert extrem hohe Stellenzahl $O(\beta^s)$!
Geht es einfacher? Z.B. Runde y schon nach dem Schieben.
Dies liefert im obigen Fall das gleiche Ergebnis.

Ü: Tabellenmacherdilemma. Tabellieren von transzendenten Funktionen.
$\mathbb{F} = \mathbb{F}(10,4,1)$, Erstelle Liste $y = \exp(x)$ $\forall x \in \mathbb{F}$, mit $y = rd(\exp(x))$

Problem: $\exp(1.626) = \underbrace{5.0835}_{5 \text{ Stellen}}$

ich verbiete es nicht.
Wenn 5 eine gültige Ziffer ist,
dann muss man aufrunden

Im allgemeinen ist das gefährlich:

$x = 0.101 \cdot 10^1 \longrightarrow 0.101 \cdot 10^1$

$y = 0.993 \cdot 10^0 \longrightarrow 0.099 \cdot 10^1$ (schieben)

$0.002 \cdot 10^1$ (Schieben und runden)

relativer Fehler im Ergebnis:

$$\frac{(x \ominus y) - (x - y)}{x - y} = \frac{0.02 - 0.017}{0.017} \approx 0.176 \approx 35\,eps$$

mit $eps = \frac{1}{2} 10^{1-3} = 0.005$.

→ Fehler ist 35 mal größer als erwartet.

Eine Stelle mehr im Addierer (also r+1) liefert das exakte Ergebnis!

Mit einer zusätzlichen Stelle erreicht man

$$\frac{(x \ominus y) - (x - y)}{x - y} \leq 2\,eps.$$

Mit zwei Stellen erreicht man exakte Rundung!

Die zusätzlichen Stellen nennt man „guard digits".

Zusätzliche Probleme bei der Arithmetik $\circledast : F \times F \to F$:

- Assoziativ und Distributivgesetz gelten nicht, es kommt also auf die Reihenfolge der Operationen an!
- $\exists\, y \in F$ so dass $x \oplus y = x$
- Allerdings gilt das Kommutativgesetz!
- Es gelten auch folg. einfache Regeln:
  $(-x) \odot y = -(x \odot y)$, $1 \odot x = x$, $x \odot y = 0 \Rightarrow x = 0 \vee y = 0$,
  $x \odot z \leq y \odot z$ falls $x \leq y$ und $z > 0$.

2.4 Fehleranalyse

Fortpflanzung von Rundungsfehlern in Rechnungen.

Wie in Kap. 1 betrachten wir die Funktionsauswertung

- Sei $F: \mathbb{R}^m \to \mathbb{R}^n$, in Komponenten

$$F(x) = \begin{pmatrix} F_1(x_1, \dots, x_m) \\ \\ F_n(x_1, \dots, x_m) \end{pmatrix}$$

- Zur Berechnung von F im Rechner nutze numerische Realisierung $F': \mathbb{F}^m \to \mathbb{F}^n$. F' wird durch einen Algorithmus realisiert, d.h. aus
  - -- endlich vielen (= Terminiery)
  - -- elementaren (= bekannte) Rechenoperationen

  zusammengesetzt: $F'(x) = \varphi_e(\dots \varphi_2(\varphi_1(x))\dots)$.

  Wichtig: i) Zu einem F gibt es i.d.R. viele Realisierungen im Sinne unterschiedlicher Reihenfolgen
  $$a+b+c \approx (a \oplus b) \oplus c \neq a \oplus (b \oplus c) !$$
  ii) Jedes φ_i steuert einen (unbekannten) Fehler bei

  iii) Im Prinzip kann die Rechengenauigkeit beliebig gesteigert werden, d.h. eigentlich Folge $F^{(k)}: (\mathbb{F}^{(k)})^m \to (\mathbb{F}^{(k)})^n$. Das machen wir aber nicht so formal.

Wie in Kap. 1 nutze Aufspaltung:

$$\underbrace{F(x)}_{\text{exakter Ergebnis}} - F'(\underbrace{rd(x)}_{\text{Eingabe runden}}) = \underbrace{F(x) - F(rd(x))}_{=0} + \underbrace{F(rd(x)) - F'(rd(x))}_{=0} \qquad (2.4)$$

$F'(rd(x))$: numerische Auswertung; $rd(x) \in \mathbb{F} \subseteq \mathbb{R}$

① Konditionsanalyse von F

② Rundungsfehleranalyse: Unterschied F, F' bei gleicher Eingabe (aus $\mathbb{F}$)

von hier aus:

- Analyse „in erster Näherung"
- absolute / relative Fehler.
- Normen bilden, das lassen wir aber i.d.R. weg.

① Differentielle Konditionsanalyse

Wir nehmen an, dass $F : \mathbb{R}^m \to \mathbb{R}^n$ zweimal stetig differenzierbar.

Nach dem Satz von Taylor gilt für die F_i:

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

Für das Restglied gilt (unter diesen Voraussetzungen)

$$R_i^F(x;\Delta x) = O(\|\Delta x\|^2)$$

Definition 2.7 (Landausche Symbole)

Man schreibt:

$$g(t) = O(h(t)) \quad (t \to 0)$$

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

$$|g(t)| \leq c\,|h(t)|$$

gilt. Sprechweise: „$g(t)$ geht wie $h(t)$ gegen 0". Man will also quantifizieren „wie schnell" eine Funktion (mindestens) gegen 0 geht.

Weiter bedeutet

$$g(t) = o(h(t)) \quad (t \to 0),$$

dass es $t_0 > 0$ und eine Funktion $c(t)$, $\lim_{t\to 0} c(t) = 0$ gibt, sodass für alle $t \in (0, t_0]$ gilt

$$|g(t)| \leq c(t)\,|h(t)|.$$

Bedeutung: „$g(t)$ geht schneller als $h(t)$ gegen Null" (falls $h(t) \to 0$).

Somit können wir die Taylorformel umformen:

$$F_i(x+\Delta x) - F_i(x) = \underbrace{\sum_{j=1}^{m} \frac{\partial F_i}{\partial x_j}(x)\,\Delta x_j}_{\text{„führende (erste) Ordnung"}} + \underbrace{O(\|\Delta x\|^2)}_{\text{höhere Ordnung}}.$$

Oft lässt man die Terme höherer Ordnung weg und schreibt „$\doteq$" statt „$=$". Sprechweise: „ist in erster Näherung gleich".

Nun gehen wir zu relativen Fehlern über. Sei $F_i(x) \neq 0$ und $x_j \neq 0$

$$\frac{F_i(x+\Delta x) - F_i(x)}{F_i(x)} \doteq \sum_{j=1}^{m} \frac{\partial F_i}{\partial x_j}(x)\,\frac{\Delta x_j}{F_i(x)}$$

$$\doteq \sum_{j=1}^{m} \underbrace{\left(\frac{\partial F_i}{\partial x_j}(x)\,\frac{x_j}{F_i(x)}\right)}_{\text{Verstärkungsfaktor } \bar{k}_{ij}(x)} \left(\frac{\Delta x_j}{x_j}\right) \qquad (2.5)$$

↑ relativer Eingabefehler.

Fasst man die Verstärkungsfaktoren in einer Matrix

$$\left(\bar{K}(x)\right)_{ij} = \bar{k}_{ij}(x)$$

zusammen so kann man (für geeignete Normen) zeigen

$$\|\bar{K}(x)\| \leq K(x)$$

mit $K(x)$ der relativen Konditionszahl aus Kapitel 1.

Ü: Setze für $x = (x_1, \dots, x_m)^T$, $x_i \neq 0$ „$\frac{\Delta x}{x}$" $= \left(\frac{\Delta x_1}{x_1}, \dots, \frac{\Delta x_m}{x_m}\right)^T$

$$\frac{\|\Delta x\|_\infty}{\|x\|_\infty} \leq \left\|\frac{\Delta x}{x}\right\|_\infty \leq \frac{\|\Delta x\|_\infty}{\min_i x_i}$$

$$\Rightarrow \frac{\dfrac{\|F(x+\Delta x) - F(x)\|_\infty}{\|F(x)\|_\infty}}{\dfrac{\|\Delta x\|_\infty}{\|x\|_\infty}} \leq \frac{\dots}{\dfrac{\|\Delta x\|_\infty}{\|x\|_\infty}} = K(x)$$

Definition 2.8 Wir nennen die Auswertung $y = F(x)$ „schlecht konditioniert" im Punkt x, falls $|k_{ij}(x)| >> 1$, andernfalls „gut konditioniert". $|k_{ij}(x)| < 1$ heißt Fehlerdämpfung, $|k_{ij}(x)| > 1$ Fehlerverstärkung.

Warum relative Kondition?

Wegen Lemma 2.5 gilt

$$\left| \frac{x - rd(x)}{x} \right| \leq eps = \frac{1}{2} \beta^{1-r},$$

d.h. es gibt $\varepsilon \in \mathbb{R}$, $|\varepsilon| \leq eps$, sodass

$$\frac{x - rd(x)}{x} = \varepsilon \iff x - rd(x) = \varepsilon x \iff rd(x) = x + \underbrace{\varepsilon x}_{=: \Delta x}$$

d.h. für die relativen Eingabefehler in (2.5) gilt gerade

$$\frac{\Delta x_j}{x_j} = \frac{\varepsilon_j x_j}{x_j} = \varepsilon_j .$$

Beispiel 2.9 **Übung:** Kondition von $F(x_1, x_2) = x \cdot y$, x/y, $F(x) = \sqrt{x}$

a) Addition. $F(x_1, x_2) = x_1 + x_2$ $\frac{\partial F}{\partial x_1} = 1$, $\frac{\partial F}{\partial x_2} = 1$. Nach obiger Formel (2.5):

$$\frac{F(x_1 + \Delta x_1, x_2 + \Delta x_2) - F(x_1, x_2)}{F(x_1, x_2)} = 1 \cdot \frac{x_1}{x_1 + x_2} \underbrace{\frac{\Delta x_1}{x_1}}_{1.1 \leq eps} + 1 \cdot \frac{x_2}{x_1 + x_2} \underbrace{\frac{\Delta x_2}{x_2}}_{1.1 \leq eps}$$

$\Rightarrow$ Für $x_1 \to -x_2$ gehen beide Verstärkungsfaktoren gegen ∞ (sofern $|x_1|, |x_2| > \delta$). Schlecht konditioniert!

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

$$\frac{F(x_1 + \Delta x_1, x_2 + \Delta x_2) - F(x_1, x_2)}{F(x_1, x_2)} = \underbrace{2x_1 \frac{x_1}{x_1^2 - x_2^2}}_{k_1 = 2 \frac{x_1^2}{x_1^2 - x_2^2}} \underbrace{\frac{\Delta x_1}{x_1}}_{1.1 \leq eps} \underbrace{- 2x_2 \frac{x_2}{x_1^2 - x_2^2}}_{k_2 = 2 \frac{x_2^2}{x_2^2 - x_1^2}} \underbrace{\frac{\Delta x_2}{x_2}}_{1.1 \leq eps}$$

Schlecht konditioniert für $|x_1| \approx |x_2|$.

06.10.09

(2) Rundungsfehleranalyse

d.h. Nach (2.4): $F(x) - F'(x)$ mit $x \in \mathbb{F}^m$ Maschinenzahl.

F' „zusammengesetzt" aus Einzeloperationen $\circledast \in \{\oplus, \ominus, \odot, \oslash\}$.

Wegen (2.3) (exakt gerundete Arithmetik) und Lemma 2.5 gilt

$$\frac{(x \circledast y) - (x * y)}{(x * y)} = \varepsilon \qquad \text{mit} \quad |\varepsilon| \leq eps.$$

Vorsicht: ε ist abhängig von x und y, d.h. Für jede Operation verschieden!

Und damit

$$x \circledast y = (x * y)(1+\varepsilon) \quad \text{für ein } |\varepsilon(x,y)| \leq eps.$$

Analyse „in erster Näherung"

Beispiel 2.10 $F(x_1, x_2) = x_1^2 - x_2^2$ mit zwei Realisierungen

$$F_a(x_1, x_2) = (x_1 \odot x_2) \ominus (x_2 \odot x_2)$$
$$F_b(x_1, x_2) = (x_1 \ominus x_2) \odot (x_1 \oplus x_2)$$

a) $u = x_1 \odot x_1 = (x_1 \cdot x_1)(1+\varepsilon_1)$

$v = x_2 \odot x_2 = (x_2 \cdot x_2)(1+\varepsilon_2)$ $\quad \varepsilon_1 \neq \varepsilon_2$ aber $|\varepsilon_i| \leq eps$!

$F_a(x_1, x_2) = u \ominus v = (u - v)(1+\varepsilon_3)$

$$= \left(x_1^2(1+\varepsilon_1) - x_2^2(1+\varepsilon_2)\right)(1+\varepsilon_3)$$
$$= \left(x_1^2 + \varepsilon_1 x_1^2 - x_2^2 - \varepsilon_2 x_2^2\right)(1+\varepsilon_3)$$
$$= \underbrace{x_1^2 - x_2^2}_{= F(x_1,x_2)} + \underbrace{\varepsilon_1 x_1^2 - \varepsilon_2 x_2^2 + \varepsilon_3 x_1^2 - \varepsilon_3 x_2^2}_{\text{erste Ordnung}} + \underbrace{\varepsilon_1\varepsilon_3 x_1^2 - \varepsilon_2\varepsilon_3 x_2^2}_{\text{zweite Ordnung.}}$$

relativer Fehler

$$\frac{F_a(x_1,x_2) - F(x_1,x_2)}{F(x_1,x_2)} \overset{\text{in erster Näherung}}{\doteq} \frac{x_1^2}{x_1^2 - x_2^2}(\varepsilon_1 + \varepsilon_3) + \frac{x_2^2}{x_2^2 - x_1^2}(\varepsilon_2 + \varepsilon_3)$$

Bemerkung: gleiche Verstärkungsfaktoren wie in Bsp. 2.9!

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

$v = x_1 \oplus x_2 = (x_1 + x_2)(1+\varepsilon_2)$

$$F_B(x_1,x_2) = u \odot v = (u \cdot v)(1+\varepsilon_3)$$

$$= \Big((x_1-x_2)(1+\varepsilon_1) \cdot (x_1+x_2)(1+\varepsilon_2) \Big)(1+\varepsilon_3)$$

$$= \underbrace{(x_1-x_2)(x_1+x_2)}\ \underbrace{(1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3)}$$

$$= (x_1^2 - x_2^2)\left(1 + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_1\varepsilon_2 + \ldots + \varepsilon_1\varepsilon_2\varepsilon_3\right)$$

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

$\Rightarrow$ Verstärkungsfaktor 1, besser als vorher.

Definition 2.11

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

a.w. Verstärkungsfaktoren aus Rundungsfehleranalyse $\leq$ denen aus Konditionsanalyse $\Rightarrow$ „numerisch stabil"

Beide Realisierungen a, b aus Bsp. 2.10 sind numerisch stabil.

2.5 Auslöschung

17
06.10.09

Obige Beispiele 2.9, 2.10 enthalten das Phänomen der Auslöschung.

Dies tritt auf bei

- Addition $x_1 + x_2$ mit $x_1 \approx -x_2$,
- Subtraktion $x_1 - x_2$ mit $x_1 \approx x_2$.

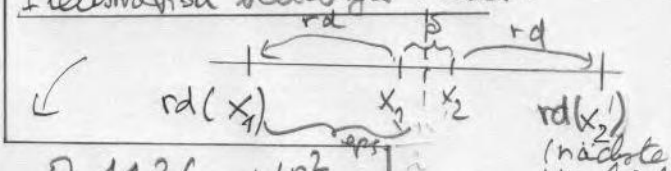
Bemerkung 2.12 Bei der Auslöschung werden vor der entsprechenden Addition bzw. Subtraktion eingeführte Fehler extrem verstärkt. ▨

Sind $x_1, x_2 \in \mathbb{F}$ Maschinenzahlen, so gilt wie oben gezeigt

$$\left| \frac{(x_1 \oplus x_2) - (x_1 - x_2)}{x_1 - x_2} \right| \leq eps.$$

Also kein Problem. Das Problem tritt erst ein, wenn x_1, x_2 selbst schon mit Fehlern behaftet sind.

Illustration beliebiger Fehler:



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

$x_1 = 0.11258762 \cdot 10^2 \rightarrow rd(x_1) = 0.1126 \cdot 10^2$

$x_2 = 0.11244891 \cdot 10^2 \rightarrow rd(x_2) = 0.1124 \cdot 10^2$

$x_1 \ominus x_2 = 0.00013871 \cdot 10^2$

$= 0.13871 \cdot 10^{-1}$

$rd(x_1) - rd(x_2) = 0.0002 \cdot 10^2$

$= \underbrace{0.2 \cdot 10^{-1}}_{\text{keine gültige Ziffer!}}$

Fehler $\frac{2eps - \delta}{\delta} \to \infty$ für $\delta \to 0$

! Überhaupt kein Fehler

relativer Fehler:

$$\frac{0.2 \cdot 10^{-1} - 0.13871 \cdot 10^{-1}}{0.13871 \cdot 10^{-1}} \approx 0.44 \approx 883 \cdot \underbrace{\frac{1}{2} 10^{-3}}_{=eps} \quad !$$ ▨

Hier: Rundung der Eingangsgrößen.

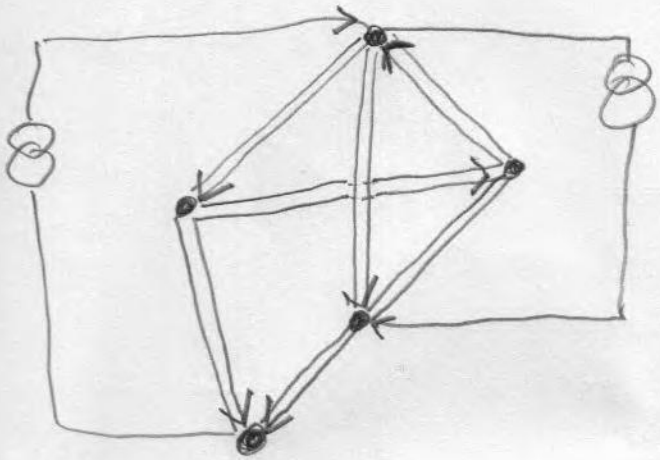
Herkunft der Fehler ist egal, tritt ebenso auf, wenn x_1, x_2 mit Fehlern vorhergehender Rechenschritte behaftet sind.

Regel 2.14 Setze potentiell gefährliche Operationen möglichst früh im Algorithmus ein. (Siehe Beispiel 2.10). ▨

3 Motivation Linearer Gleichungssysteme

07.10.09

3.1 Strömung in Rohrleitungsnetzen



1) Netzwerk von Röhren beschrieben durch gerichteten Graphen.

- Knotenmenge $V = \{v_1, \dots, v_n\}$, $|V| = n$.
- Kantenmenge $E = \{e_1, \dots, e_M\}$, $|E| = M$,

  $E \subseteq V \times V$, mit : $(v,w) \in E \Rightarrow (w,v) \notin E$
- $E = E_R \cup E_P$ „Röhren und Pumpen", $|E_R| = m$,

  $E_R = \{e_1, \dots, e_m\}$, $E_P = \{e_{m+1}, \dots, e_M\}$.

eigentlich hier!

2) Gesetz von Hagen-Poiseuille.

Röhre $e = (v,w)$ (von Knoten v nach Knoten w)

$$(3.1) \quad q_e = \underbrace{\frac{\pi r_e^4}{8 \eta d_e}}_{=: L_e \text{ „Leitfähigkeit"}} \Delta p_e$$

r_e : Radius der Röhre e [m]

d_e : Länge der Röhre e [m]

η : dyn. Viskosität Flüssigk. [Pa s]

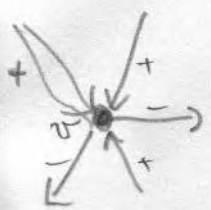
q_e: Volumenstrom $[m^3/s]$

Orientierung: $q_e > 0$ falls Fluss von Knoten v nach w.
$q_e < 0$ Fluss von w nach v

Δp_e: gerichtete Druckdifferenz über Rohr e $[Pa] = [\frac{N}{m^2}]$

$\Delta p_e > 0 \Rightarrow q_e > 0$

3) Knotenregel (1. Kirchhoffsches Gesetz)



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

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

$$\sum_{e \in E_v^+} q_e - \sum_{e \in E_v^-} q_e = 0 \qquad \forall\, v \in V. \qquad (3.2)$$

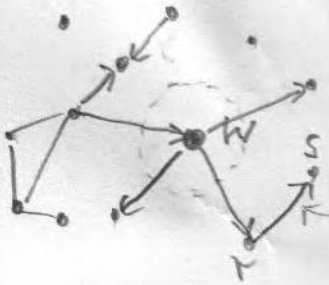
Grund: Massenerhaltung, Knoten speichert keine Flüssigkeit.

Nur $n-1$ der Beziehungen (3.2) sind linear unabhängig:

Wähle $w \in V$.

$$\sum_{v \in V-\{w\}} \Bigg(\underbrace{\sum_{e \in E_v^+} q_e - \sum_{e \in E_v^-} q_e}_{=0!} \Bigg) = \sum_{v \in V-\{w\}} \sum_{e \in E_v^+} q_e - \sum_{v \in V-\{w\}} \sum_{e \in E_v^-} q_e$$

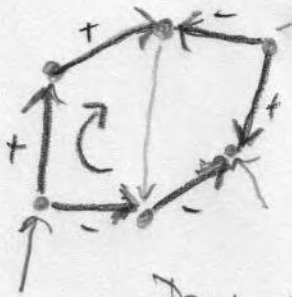
$$= \boxed{\sum_{e \in E_w^-} q_e - \sum_{e \in E_w^+} q_e = 0}$$



$e' = (r,s)$ $r \neq w$, $s \neq w$
kommt genau zweimal vor: $+q_{e'}$ für Knoten s (eingehend)
$-q_{e'}$ für Knoten r (ausgehend)

- Knotenregel für $v \in V-\{w\}$ erfüllt $\Rightarrow$ Knotenregel für w erfüllt.

4) Maschenregel (2. Kirchhoffsches Gesetz)



- Betrachte $C \subseteq E$, C geschlossener (beliebiger) Pfad
- $C^+ = \{ e \in C \mid e \text{ genauso wie } C \text{ orientiert} \}$
  $C^- = \{ e \in C \mid e \text{ entgegen } C \text{ orientiert} \}$

Dann gilt

$$\sum_{e \in C^+} \Delta p_e - \sum_{e \in C^-} \Delta p_e = 0 \qquad \forall \text{ Pfade}$$

Grund:

$\int_v^w \Delta p_e \cdot A \cdot ds$ ist Energie die notwendig ist um Probekörper von v nach w zu bringen

Δp_e: $\frac{N}{m^2}$, A: m^2, ds: m

- Betr. zweifach Knoten: r, s. Energie um K. von r nach s zu bringen unabhängig von Weg.
- Lineare Abhängigkeit!

5) Knotenpotentialverfahren

07.10.09

Folge der Maschenregel:

("Potentiale")

- Man darf $n-1$ Knotendrücke p_v als Unbekannte wählen
- $e=(v,w)$: $\Delta p_e = p_v - p_w$
- Druck in einem (Referenz-) Knoten r kann willkürlich festgelegt werden. z.B. $p_r = 0$. ($p_v, v \in V$ Lösung $\Rightarrow p_v + \text{const}$ auch Lösung)
- Maschenregel ist dann für alle Pfade erfüllt.

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

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

Wähle „Anordnung" $e_k \leftrightarrow k$ $v_i \leftrightarrow i$ Wahl $\boxed{r = v_n}$

$$e_k=(v_i,v_j): k \quad \begin{bmatrix} \Delta p \end{bmatrix} = \begin{bmatrix} B \end{bmatrix} \begin{bmatrix} p \end{bmatrix} \quad \Rightarrow p=(p_{v_1}, \dots, p_{v_{n-1}})^T$$

$$\Delta p = (\Delta p_1, \dots, \Delta p_m)$$

B: $m \times (n-1)$ Matrix, $+1$ in Spalte i, -1 in Spalte j

$|E_R| \quad |V|-1$

b) Leitfähigkeiten

$$e \in E_R: \quad q_e = L_e \Delta p_e$$

als Matrix:

$$\begin{bmatrix} q \end{bmatrix} = \begin{bmatrix} L \end{bmatrix} \begin{bmatrix} \Delta p \end{bmatrix}$$

L: $m \times m$ Diagonalmatrix, $m = |E_R|$

c) Knotenregeln: $n-1$ Stück für Knoten $v_1, \dots, v_{n-1}$ (exklusive Referenzknoten)

$$v \in V - \{v_n\}: \quad \sum_{e \in E_v^+} q_e - \sum_{e \in E_v^-} q_e = 0$$

Pumpenströme auf rechte Seite $\Leftrightarrow$

$$\sum_{e \in E_v^+ \cap E_R} q_e - \sum_{e \in E_v^- \cap E_R} q_e = \sum_{e \in E_v^- \cap E_p} q_e - \sum_{e \in E_v^+ \cap E_p} q_e$$

$$\Leftrightarrow \begin{bmatrix} B^T \end{bmatrix} \begin{bmatrix} q \end{bmatrix} = \begin{bmatrix} b \end{bmatrix} \qquad b=(b_1, \dots, b_{n-1})$$

B^T: $(n-1) \times m$ Matrix; b enthält Pumpenströme

Alles zusammen:

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

$$\Leftrightarrow \quad A p = b$$

- A ist $(n-1)\times(n-1)$ Matrix
- A ist symmetrisch und positiv definit, d.h. $x^T A x > 0 \; \forall x \neq 0$
- damit insbesondere invertierbar
- A ist dünn besetzt:

$$a_{ij} \neq 0 \Leftrightarrow (v_i, v_j) \in E \vee (v_j, v_i) \in E.$$

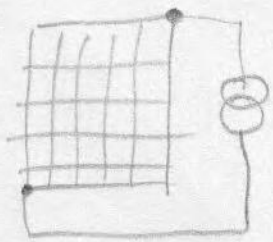
sehr viele Einträge sind Null

$$|\{(i,j) \mid a_{ij} \neq 0\}| = O(n) \quad (\text{statt } n^2).$$

Bew: 1) $B \in \mathbb{R}^{m\times(n-1)}$ hat max. Rang, also $n-1$ da $m \geq n-1$, $n-1$ linear unabh. Spalten!
2) L diagonal mit $l_{ii} > 0$
3) Sei $x \neq 0$, $y := Bx$ damit ist $x^T A x = \sum_i l_{ii} y_i^2 > 0$ da $\dim(\ker(B)) + \operatorname{rang}(B) = n-1$ und $\operatorname{rang}(B) = n-1 \Rightarrow \dim(\ker(B)) = 0$ also $y = 0$ nur für $x = 0$.

ü1 : ein kleines Netzwerk

ü2 : New York City



- völlig analog lassen sich elektrische Netzwerke (aus Widerständen) behandeln

Ohmsches Gesetz: $i_e = \frac{1}{R_e}(u_v - u_w)$

i_e : Strom durch Widerstand R_e
u_v : Knotenpotentiale
R_e : Widerstand

- RLC Netzwerke mit harmonischer Anregung
→ komplexe Ströme und Spannungen

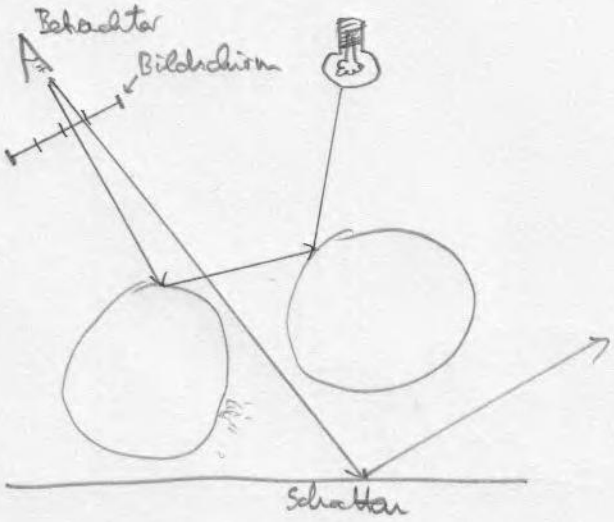
$$Ax = b \quad \text{mit} \quad x, b \in \mathbb{C}^n, \; A \in \mathbb{C}^{n\times n}$$

L: Spulen
C: Kapazitäten.

3.2 Radiosity-Methode in der Computergrafik

Beleuchtung einer „Szene".

Ray-Tracing:



Nachteil: starke Schatten.

Radiosity-Methode:

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

Bestimme $B : S \to \mathbb{R}$ „Energiedichte"

$$\int_\omega B \, dA = \text{von } \omega \subset S \text{ abgestrahlte Energie}$$

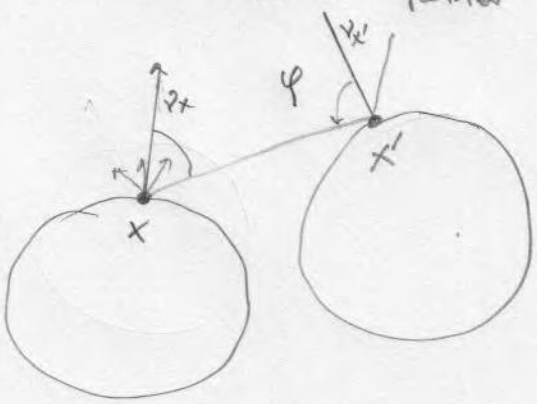
Energie wird von einem Punkt $x \in S$ in alle Richtungen abgestrahlt.

Bestimmungsgleichung für $B(x)$:

reflektiertes Licht

$$B(x) = E(x) + \rho(x) \int_S B(x') \underbrace{\frac{\cos\varphi_{x,x'} \cos\varphi_{x',x}}{\pi \|x-x'\|^2} V(x,x')}_{=: \lambda(x,x') \text{ „Kern"}} dA'$$

- $B(x)$: $x \in S$
- $E(x)$: Eigenstrahlung (Lichtquelle)
- $\rho(x)$: Reflexionsfaktor
- $\int_S$
- $B(x')$: Licht von x'



$$\varphi_{a,b} = \sphericalangle(\nu_a, b-a)$$

$$\cos\varphi_{a,b} = \begin{cases} 1 & \text{Licht trifft auf Richtung } \nu_a \text{ ein} \\ 0 & \text{Licht trifft tangential zur Oberfl. ein} \end{cases}$$

Sichtbarkeit (Visibility)

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

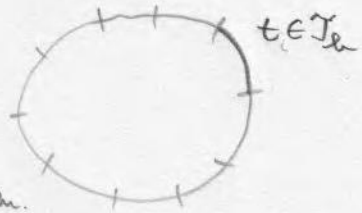
(Eigenschaft der Szene)

Integralgleichung für $B(x) : S \to \mathbb{R}$:

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

Numerische Lösung mit „Kollokationsmethode".

a) Zerlegung der Oberfläche: $\mathcal{T}_h = \{t_1, \dots, t_n\}$



$t_i \subset S$, $t_i \cap t_j = \emptyset \ \forall i \neq j$, $\overset{n}{\underset{i=1}{\cup}} \bar{t}_i = S$

könnte man abschwächen

t_i : offene Gebiete.

Zu $t \in \mathcal{T}_h$ wähle x_t = Mittelpunkt von t.

Dieser Prozess heißt auch Diskretisierung. Üblich bei Differential- und Integralgleichungen.

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

$$B_h(x) = \sum_{j=1}^{n} z_j \varphi_j(x)$$

$z_j \in \mathbb{R}$ Koeffizient

$\varphi_j : S \to \mathbb{R}$ „Basisfunktion"

d.h. $\varphi_1, \dots, \varphi_n$ linear unabhängig.

stückweise konstante Funktionen:

$$\varphi_j(x) = \begin{cases} 1 & \text{falls } x \in t_j \\ 0 & \text{sonst.} \end{cases}$$

c) Erfülle Integralgleichung nur für $x \in X_h = \{x_{t_1}, \dots, x_{t_n}\}$.

$$B_h(x_i) - \varrho(x_i) \int_S B_h(x') \lambda(x_i, x') \, dA' = E(x_i) \qquad i = 1, \dots, n$$

$$\Leftrightarrow \sum_{j=1}^{n} z_j \varphi_j(x_i) - \varrho(x_i) \int_S \sum_{j=1}^{n} z_j \varphi_j(x') \lambda(x_i, x') \, dA' = E(x_i) \quad i = 1, \dots, n$$

$$\Leftrightarrow \sum_{j=1}^{n} z_j \underbrace{\left\{ \varphi_j(x_i) - \varrho(x_i) \int_S \varphi_j(x') \lambda(x_i, x') \, dA' \right\}}_{=: a_{ij}} = \underbrace{E(x_i)}_{b_i} \quad i = 1, \dots, n$$

$$\Leftrightarrow \boxed{Az = b}.$$

- Integral in a_{ij} wird i.d.R. auch numerisch berechnet
  $\Rightarrow$ Methoden später in der Vorlesung.

- Man begeht einen Diskretisierungsfehler

$$\| B - B_h \| = O(h^\alpha) \qquad \text{(Konvergenz)}$$

mit $h = \max_{t \in \mathcal{T}_h} \operatorname{diam}(t)$, α: „Konvergenzordnung".

d.h. je feiner die Unterteilung ($n \to \infty$), desto besser approximiert B_h die gesuchte Funktion B.

$\Rightarrow$ Gleichungssysteme können beliebig groß werden.
(im Gegensatz zum Röhrennetzwerk).

- A ist in diesem Fall nicht dünn besetzt sondern voll besetzt!

4 Kondition der Lösung linearer Gleichungssysteme

4.1 Lösbarkeit

Gegeben Matrix A, Vektor b:

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

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

$$\forall i : \sum_{j=1}^{n} a_{ij} x_j = b_i,$$

bzw.

$$Ax = b. \tag{4.1}$$

Die Zahlen a_{ij}, b_i, x_j können aus $\mathbb{R}$ oder $\mathbb{C}$ sein.

Das Gleichungssystem (4.1) heißt

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

bei maximalem Rang von A:
i. allg. viele Lösungen
genau eine Lösung
$b \in$ Bild(A).

Es gibt mindestens eine Lösung falls

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

Für quadratische Matrizen sind folg. 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 Eigenwert Null.

4.2 Vektornormen

Quantifizieren von Fehlern erfordert Normen.

Im Folgenden sei $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$ der zugrunde liegende Körper.

Definition 4.1 (Vektornorm)

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

(N1)	$\|x\| > 0$	$x \neq 0$	(Definitheit)
(N2)	$\|\alpha \cdot x\| = \|\alpha\| \|x\|$	$x \in \mathbb{K}^n, \alpha \in \mathbb{K}$	(positive Homogenität)
(N3)	$\|x+y\| \leq \|x\| + \|y\|$	$x, y \in \mathbb{K}^n$	(Subadditivität, Dreiecksungl.)

Beispiel 4.2 Häufig verwendete Normen sind:

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}$$ Euklidsche Norm, ℓ_2-Norm

$$\|x\|_\infty = \max_{i=1,\dots,n} |x_i|$$ Maximumnorm, ℓ_∞-Norm

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$ ℓ_1-Norm

Man zeigt über die Konvergenz von Folgen von Vektoren:

$$\forall i = 1,\dots,n: x_i^{(t)} \to x_i \quad (t \to \infty) \iff \|x^{(t)} - x\| \to 0 \quad (t \to \infty)$$

(Konvergenz der Normen äquivalent zu komponentenweiser Konvergenz).
⇒ siehe nächster Satz

Eine wichtige Folgerung aus (N3) ist

$$\|x - y\| \geq \big| \|x\| - \|y\| \big| \quad \forall x, y \in \mathbb{K}^n. \tag{4.2}$$

Beweis:

$$\|x\| = \|x - y + y\| \overset{(N3)}{\leq} \|x-y\| + \|y\| \Rightarrow \|x-y\| \geq \|x\| - \|y\|$$
$$\|y\| = \|y - x + x\| \leq \|y-x\| + \|x\| \Rightarrow \|x-y\| \geq \|y\| - \|x\| = -(\|x\| - \|y\|)$$

Aus (4.2) folgt:

$$\|x^{(t)} - x\| \to 0 \quad (t \to \infty) \Rightarrow \|x^{(t)}\| \to \|x\| \quad (t \to \infty)$$

Satz 4.3 (Äquivalenz aller Normen)

Auf $\mathbb{K}^n$ sind alle Normen im folgenden Sinne äquivalent:
Zu $\|.\|$, $\|.\|'$ gibt es Zahlen $m, M > 0$ aus $\mathbb{R}$ so dass gilt

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

Beweis: Es genügt dies zu zeigen für $\|.\|' = \|.\|_\infty$.
Seien $e_1, \dots, e_n$ die kartesischen Einheitsvektoren. Jedes $x \in \mathbb{K}^n$ hat die Darstellung $x = \sum_{i=1}^n x_i e_i$, also $\|x\| = \|\sum_{i=1}^n x_i e_i\| \le \sum_{i=1}^n |x_i| \|e_i\| \le \left(\max_{1\le i\le n} |x_i|\right) \sum_{i=1}^n \|e_i\| = \|x\|_\infty \underbrace{\sum_i \|e_i\|}_{=\gamma}$
wg. $\|x\| \le \gamma \|x\|_\infty$ ist $\|.\|$ stetig bezügl. komponentenweiser Konvergenz.
Punktmenge $S := \{x \in \mathbb{K}^n : \|x\|_\infty = 1\} \subset \mathbb{K}^n$ ist
- beschränkt (bezgl. $\|.\|$ da $\|x\| \le \gamma$)
- abgeschlossen, d.h. Grenzwerte von Folgen aus S sind in S (da $\mathbb{K}^n$ metrischer Raum).

$\Rightarrow$ „kompakt".

Die Funktion $\|.\| : S \to \mathbb{R}_+$ nimmt auf S ihr Minimum und Maximum an; d.h.:
Es gibt $\underline{x}, \bar{x} \in S$:

$$\underbrace{0 < \|\underline{x}\|}_{\text{da } 0 \notin S,\ \underline{x} \in S \text{ und N1.}} \le \|x\| \le \underbrace{\|\bar{x}\| \le \gamma \underbrace{\|\bar{x}\|_\infty}_{=1 \text{ da } \bar{x} \in S} = \gamma}_{\text{da } \bar{x} \in S} \qquad \forall x \in S$$

Nun sei $y \in \mathbb{K}^n - \{0\}$ beliebig.
Dann ist $y/\|y\|_\infty \in S$ (da Norm 1) und somit

$$\|\underline{x}\| \le \|y/\|y\|_\infty\| \underset{(N2)}{=} \frac{1}{\|y\|_\infty} \|y\| \le \|\bar{x}\|$$

$$\Leftrightarrow \quad \underbrace{\|\underline{x}\|}_{=m} \|y\|_\infty \le \|y\| \le \underbrace{\|\bar{x}\|}_{=M} \|y\|_\infty$$

Vorsicht: m, N hängen i.d.R. von der Dimension n ab.
Aufpassen wenn man etwas für $n \to \infty$ beweist.

Nochmal zur Konvergenz:

$x_i^{(t)} \to x_i$ $(t \to \infty)$ $\Leftrightarrow$ $\|x^{(t)} - x\|_\infty \to 0$ $(t \to \infty)$ überlegt man leicht

Satz 4.3 zeigt dann, dass man auch eine beliebige Norm nehmen darf.

4.3 Matrixnormen

Der $\mathbb{K}^{m\times n}$ stellt auch einen Vektorraum dar und kann mit dem $\mathbb{K}^{mn}$ identifiziert werden.

Damit definiert jede Norm auf $\mathbb{K}^{mn}$ auch eine Norm auf $A\in\mathbb{K}^{m\times n}$.

$A^{(t)}\to A\ (t\to\infty)$ meint dann $a_{ij}^{(t)}\to a\ (t\to\infty)\ \forall i,j\in\{1,\dots,m\}\times\{1,\dots,n\}$

Es zeigt sich, dass folgende Eigenschaften hilfreich sind:

Definition 4.4 Eine Norm $\|\cdot\|$ auf $\mathbb{K}^{n\times n}$ (m≠n würde zwei V-Normen erfordern) heißt verträglich mit einer Vektornorm $\|\cdot\|$ auf $\mathbb{K}^n$ falls gilt

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

Sie heißt Matrixnorm, wenn sie submultiplikativ ist:

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

Beispiel 4.5 Die Frobenius-Norm

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

ist eine Matrixnorm, die verträglich mit der euklidischen Norm ist.

Beweis: (als Übung)

$$\|Ax\|_2^2=\sum_{i=1}^{n}\left(\sum_{j=1}^{n}a_{ij}x_j\right)^2\overset{\text{C.S.}}{\le}\sum_{i=1}^{n}\left(\sum_{j=1}^{n}|a_{ij}|^2\cdot\underbrace{\sum_{j=1}^{n}|x_j|^2}_{=\|x\|_2^2}\right)=\|x\|_2^2\underbrace{\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2}_{=\|A\|_{Fr}^2}.$$

Definition 4.6 (Zugeordnete Matrixnorm)

Es sei $\|\cdot\|$ eine beliebige (Vektor-)Norm auf $\mathbb{K}^n$. Dann heißt

$$\|A\|:=\sup_{x\in\mathbb{K}^n-\{0\}}\frac{\|Ax\|}{\|x\|}\overset{(N2)}{=}\sup_{x\in\mathbb{K}^n,\,\|x\|=1}\|Ax\|$$

die $\|\cdot\|$ zugeordnete (oder natürliche) Matrixnorm, d.h. $\|\cdot\|$ ist verträglich und submultiplikativ.

Übung: Beweis dazu $y\neq0$ beliebig $\frac{\|Ay\|}{\|y\|}\le\sup_{x\neq0}\frac{\|Ax\|}{\|x\|}=\|A\|\Rightarrow\|Ay\|\le\|A\|\,\|y\|$

$$\|AB\|=\sup_{x\neq0}\frac{\|ABx\|}{\|x\|}\le\sup_{x\neq0}\frac{\|A\|\,\|Bx\|}{\|x\|}=\|A\|\sup_{x\neq0}\frac{\|Bx\|}{\|x\|}=\|A\|\,\|B\|.$$

Hilfssatz 4.7

22.10.09

Die zugeordneten Matrixnormen zu $\|\cdot\|_\infty$ und $\|\cdot\|_1$ sind

$$\|A\|_\infty := \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}| \qquad \text{(Zeilensummennorm)},$$

$$\|A\|_1 := \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ij}| \qquad \text{(Spaltensummennorm)}.$$

Beweis: siehe Rannacher, Numerik 0 S. 104. Teile hieraus:

a) $$\|Ax\|_\infty = \max_{i=1\dots n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \le \max_{i=1\dots n} \sum_{j=1}^{n} |a_{ij}||x_j|$$

$$\le \max_{i=1\dots n} \sum_{j=1}^{n} |a_{ij}| \underbrace{\left(\max_{k=1\dots n} |x_k|\right)}_{=\|x\|_\infty}$$

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

Dies zeigt die Verträglichkeit

b) Somit

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

c) zu zeigen ist nun

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

Dies geschieht konstruktiv indem man so einen Vektor angibt → Ra.

4.4 Eigenwerte und Eigenvektoren

Sei $A \in \mathbb{K}^{n\times n}$, $\lambda \in \mathbb{K}$ und $e \in \mathbb{K}^n$ so dass

$$Ae = \lambda e,$$

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

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

erfüllbar.

Das charakteristische Polynom $p(\lambda)$ hat genau n Nullstellen in $\mathbb{C}$ (Vielfachheiten mitgezählt).
Zu jedem Eigenwert λ gibt es mindestens einen Eigenwert.

Mit den Normregeln folgt für ein Eigenpaar (λ, e), $\|e\| = 1$:

$$|\lambda| \underset{\uparrow \|e\|=1}{=} |\lambda|\,\|e\| = \|\lambda e\| = \|Ae\| \leq \|A\| \underbrace{\|e\|}_{=1} = \|A\| \qquad (4.3)$$

Also alle Eigenwerte $\lambda \in \mathbb{C}$ liegen innerhalb eines Kreises um 0 mit Radius $\|A\|$.
Z.B. mit $\|A\|_\infty$ erhält man so eine konkrete Abschätzung für den größten Eigenwert.

Definition 4.8 (Spezielle Matrizen)

Zu $A \in \mathbb{K}^{m\times n}$ heißt $A^T \in \mathbb{K}^{n\times m}$ transponierte Matrix und es ist $(A^T)_{ij} = (A)_{ji}$
Zu $A \in \mathbb{K}^{m\times n}$ ist $\bar{A} \in \mathbb{K}^{m\times n}$ gegeben durch $(\bar{A})_{ij} = \overline{(A)_{ij}}$ ↖ konjugiert komplexe

a) $A \in \mathbb{K}^{n\times n}$ heißt hermitesch falls $A = \bar{A}^T$ d.h. $a_{ij} = \bar{a}_{ji}$
Manche Autoren schreiben auch A^H für $\bar{A}^T$.
Reelle hermitesche Matrizen heißen symmetrisch (dann ist $A = A^T$)

b) Für komplexe Matrizen $A \in \mathbb{C}^{n\times n}$ heißt — also A reell dann ist hermitesch (=) symmetrisch in $\mathbb{C}$ nicht!

$A\bar{A}^T = \bar{A}^T A$ — normal

$A\bar{A}^T = \bar{A}^T A = I$, also $A^{-1} = \bar{A}^T$ — unitär

c) Für reelle Matrizen $A \in \mathbb{R}^{n\times n}$ heißt

$AA^T = A^T A$, also $A^{-1} = A^T$ — orthogonal.

22.10.09

Definition 4.9 (Skalarprodukt)

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

(S1) $(x,y) = \overline{(y,x)}$ $\quad x,y \in \mathbb{K}^n$ (Symmetrie)

(S2) $(\alpha x + \beta y, z) = \alpha (x,z) + \beta (y,z)$ $\quad x,y,z \in \mathbb{K}^n,\ \alpha,\beta \in \mathbb{K}$ (Linearität)

(S3) $(x,x) > 0$ $\quad x \in \mathbb{K}^n - \{0\}$ (Definitheit)

↑ aus (S1) folgt $(x,x) = \overline{(x,x)}$ also $\mathrm{Im}(x,x) = 0$

Ein Skalarprodukt erzeugt immer eine zugehörige Norm

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

Es gilt die Cauchy-Schwarzsche Ungleichung

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

Das Euklidische Skalarprodukt (wird im Folg. oft verwendet) lautet:

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

Damit gilt (andere Schreibweise für hermitesch):

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

4.5 Die Spektralnorm

Die der Euklidischen Norm zugeordnete Matrixnorm $\|A\|_2$ heißt Spektralnorm.

Hilfssatz 4.10 Für die Spektralnorm hermitescher Matrizen gilt

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

Für jede Matrix $A \in \mathbb{K}^{m \times n}$ gilt

$$\|A\|_2 = \max \{ |\lambda|^{1/2} : \lambda \text{ ist Eigenwert von } \overline{A}^T A \}$$

23.10.09

Beweis: a) Sei A hermitesch. Dann hat A n **reelle** Eigenwerte und einen vollständigen Satz von orthonormalen Eigenvektoren:

$$\{w^1, \dots, w^n\} \subset \mathbb{K}^n : \; Aw^i = \lambda_i w^i, \quad (w^i, w^j)_2 = \delta_{ij} \quad i,j = 1, \dots, n$$

↑ euklidisches SP!

Jedes $x \in \mathbb{K}^n$ lässt sich in der Basis $\{w^1, \dots, w^n\}$ darstellen:

$$x = \sum_{i=1}^{n} \alpha_i w^i \quad \text{mit} \quad \alpha_i = (x, w^i)_2 \in \mathbb{K} \text{ (Einelem!)}$$

Und es gilt:

$$\|x\|_2^2 = (x,x)_2 = \Big(\sum_{i=1}^{n} \alpha_i w^i, \sum_{j=1}^{n} \alpha_j w^j\Big)_2$$

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

$$\overline{(a+ib)}\,\overline{(c+id)} = (a-ib)(c+id) = ac - bd - i(ad+bc)$$
$$\overline{(a+ib)(c+id)} = \overline{(ac-bd)+i(ad+bc)} = (ac-bd)-i(ad+bc)$$

$$\|Ax\|_2^2 = (Ax, Ax)_2 = \Big(\sum_i \underbrace{A\alpha_i w^i}_{\alpha_i \lambda_i w^i}, \sum_j \underbrace{A\alpha_j w^j}_{\alpha_j \lambda_j w^j}\Big)_2$$

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

↑ Betrag, da $\alpha \in \mathbb{K}$

Also

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

Andererseits zeigt (4.3) $\|A\| \ge |\lambda|$ für jede Norm und jeden EW, also muss $\|A\|_2 = \max_{1 \le i \le n} |\lambda_i|$ sein.

b) Sei $A \in \mathbb{K}^n$ eine beliebige Matrix. Dann gilt — Rechenregeln für konjug. komplex

$$\|Ax\|_2^2 = (Ax, Ax)_2 = (Ax)^T (\overline{Ax}) = x^T A^T (\overline{Ax}) = x^T \overline{(\overline{A}^T A x)} = (x, \overline{A}^T A x)_2$$

$\overline{A}^T A$ ist hermitesch und habe EW λ_i sowie EV w^i orthonormal wie da

also $$\|Ax\|_2^2 = \Big(\sum_{i=1}^{n} \alpha_i w^i, \sum_{j=1}^{n} \overline{A}^T A \alpha_j w^j\Big)_2 = \sum_{i,j=1}^{n} \overline{\alpha_i} \underbrace{\lambda_j}_{\text{reell}} \overline{\alpha_j} \underbrace{(w^i, w^j)_2}_{=\delta_{ij}} = \sum_{i=1}^{n} \lambda_i |\alpha_i|^2$$

und damit $$\|A\|_2^2 = \sup_{x \in \mathbb{K}^n \setminus \{0\}} \frac{\|Ax\|_2^2}{\|x\|_2^2} = \sup_{x \in \mathbb{K}^n \setminus \{0\}} \frac{\sum_i \lambda_i |\alpha_i|^2}{\sum_i |\alpha_i|^2} \le \max_{1 \le i \le n} |\lambda_i|$$

□

4.6 Positiv definite Matrizen

24.10.09

Wichtige Klasse von Matrizen mit vorteilhaften Eigenschaften.

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

a) $(Ax,x)_2 \in \mathbb{R} \quad \forall x \in \mathbb{K}^n \setminus \{0\} \quad (\mathbb{K}=\mathbb{C})$

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

Ziel: Charakterisierung positiv definiter Matrizen.

Im Fall $\mathbb{K}=\mathbb{C}$ bedeutet die Bedingung a) eine Einschränkung an die Matrix A.

Eigenschaft 4.12 Für $A \in \mathbb{C}^{n\times n}$ gilt A hermitesch genau dann wenn $(Ax,x)_2 \in \mathbb{R} \ \forall x \in \mathbb{C}^n \setminus \{0\}$.

Beweis „$\Rightarrow$" (als ü-Aufgabe!)

$(Ax,x)_2 \in \mathbb{R} \Leftrightarrow \underbrace{(Ax,x)_2}_{a+ib} = \underbrace{\overline{(Ax,x)_2}}_{a-ib}$ geht nur für $b=0$

also $(Ax,x)_2 \overset{\text{hermitesch 5.7}}{=} (x,Ax) \overset{(S1)}{=} \overline{(Ax,x)}$ damit ist „$\Rightarrow$" gezeigt.

„$\Leftarrow$" Aufteilen in Real- und Imaginärteil:

$\mathbb{C}^n \ni x = a+ib \quad a,b \in \mathbb{R}^n \qquad A = V+iW, \ V,W \in \mathbb{R}^{n\times n}$

Und ausrechnen:

$$(Ax,x)_2 = \big((V+iW)(a+ib), a+ib\big)_2 = \big((Va-Wb)+i(Vb+Wa), a+ib\big)_2$$

$$= \underbrace{(Va-Wb,a)_2 + (Vb+Wa,b)}_{\text{Re}} + i\underbrace{\big[(Vb+Wa,a)_2 - (Va-Wb,b)_2\big]}_{\text{Im}\ \overset{!}{\doteq} 0}$$

$(x,\alpha y)_2 = \overline{\alpha}(x,y)_2$! $\qquad i^2=-1$

also $\operatorname{Im}\big((Ax,x)_2\big) = (Vb,a)_2 + (Wa,a)_2 - (Va,b)_2 + (Wb,b)_2 \doteq 0 \qquad (*)$

Fall I) $W=0$ d.h. $A=V$ (reell), aber $x=a+ib$ ist komplex. Dann reduziert sich $(*)$ auf

$(Vb,a)_2 - (Va,b)_2 = 0 \ \forall a,b \in \mathbb{R}^n \Leftrightarrow (Vb,a)_2 = (V^Tb,a)_2 \ \forall a,b \in \mathbb{R}^n$ (alles reell)

damit muss $V=V^T$ sein.

Fall II) Setze $b=0$, dann reduziert sich (*) auf

$$(Wa,a)_2 = 0 \quad \forall a \in \mathbb{C}^n \setminus \{0\}$$

II a: Setze $a = e^i \Rightarrow w_{ii} = 0$. Hier: $e^i_j = \delta_{ij}$ kartesische Einheitsv.

II b: Sei also $w_{ii}=0 \ \forall i$ und setze $a = e^i + e^j$ $(j \neq i)$

dann folgt $w_{ij} + w_{ji} = 0 \quad \forall i,j$

$$\Leftrightarrow w_{ij} = -w_{ji}$$

Damit ist $A = V + iW$ mit $V = V^T$ und $W = -W^T$ also A hermitesch. ∎

Positiv definite Matrizen in $\mathbb{C}^{n\times n}$ sind also immer hermitesch.

Positiv definite Matrizen in $\mathbb{R}^{n\times n}$ sind nicht notwendigerweise symmetrisch.

Charakterisierung über Eigenwerte:

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

Beweis:

- A sei hermitesch mit lauter positiven Eigenwerte.

  $x \in \mathbb{C}^n$ hat Darstellung $x = \sum \alpha_i w^i$ $\quad w^i$: Eigenvektor zum EW λ_i.

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

- A sei hermitesch und positiv definit. A hat also Eigenpaare (λ_i, w^i) mit $\lambda_i \in \mathbb{R}$.
  Zu zeigen ist λ_i positiv

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

- Setze $e^i \in \mathbb{R}^n$, $e^i_j = \delta_{ij}$ kartesische Einheitsvektoren.

$$0 < (Aw_i, w_i)_2 = a_{ii} \in \mathbb{R}.$$ ∎

Die Aussage gilt natürlich auch für reelle symmetrische und positiv definite Matrizen, da auch diese hermitesch sind.

Speziell für reelle Matrizen gilt:

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

Beweis: e^i sei wieder der i-te kartesische Einheitsvektor und a_{ij}, $j \neq i$, das betragsmäßig größte Element. Dann zeigt man den Widerspruch.

$$0 < (A(e^i - \operatorname{sign}(a_{ij})e^j),\ e^i - \operatorname{sign}(a_{ij})e^j)_2$$

$$= (Ae^i, e^i)_2 - 2\operatorname{sign}(a_{ij})\underbrace{(Ae^i, e^j)_2}_{a_{ij}} + \operatorname{sign}(a_{ij})^2 (Ae^j, e^j)$$

$$\overset{A=A^T}{=} a_{ii} - 2|a_{ij}| + a_{jj} \leq 0 \quad \lightning$$

da $|a_{ij}| \geq a_{ii}$, $|a_{ij}| \geq a_{jj}$ ($a_{ii}, a_{jj} > 0$!) ∎

- Manche Autoren (auch Rannacher) verlangen, dass reelle positiv definite Matrizen symmetrisch sind. Wir fordern das extra

Ü $A \in \mathbb{R}^{n\times n}$ A ist positiv definit genau dann wenn $A_S = \frac{1}{2}(A + A^T)$ positiv definit ist.

A pos. definit, dann sind alle Hauptuntermatrizen (definiert) positiv definit

4.7 Störungstheorie

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

$F(A,b) = A^{-1}b$ ist der Lösungsoperator zur Gleichung $G(x) = Ax - b = 0$.

Betrachte die relative Kondition von F, wobei wir zunächst nur Änderungen in b zulassen wollen:

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

$$\leq \frac{\|A^{-1}\|\|\delta b\|}{\|\delta b\|} \frac{\|A\|\|A^{-1}b\|}{\|A^{-1}b\|} = \|A^{-1}\|\,\|A\|.$$

$\leq$: $\|\cdot\|$ sei verträglich und submultiplikativ

Definition 4.15 Die Zahl

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

für irgendeine verträgliche und submultiplikative Matrixnorm heißt Konditionszahl von A.

Nun wollen wir auch Änderungen in A selbst zulassen. Sei $A \in \mathbb{K}^{n\times n}$ regulär. Frage: Wann ist $A+\delta A$ regulär?

Hilfssatz 4.16 $B \in \mathbb{K}^{n\times n}$ habe Norm $\|B\| < 1$. Dann ist $I+B$ regulär und es gilt

$$\|(I+B)^{-1}\| \leq \frac{1}{1-\|B\|}.$$

Beweis: i) Für alle $x \in \mathbb{K}^n$ gilt

$$\|x\| = \|x + Bx - Bx\| \leq \|(I+B)x\| + \|Bx\|$$

$$\Leftrightarrow \|(I+B)x\| \geq \|x\| - \|Bx\| \geq \|x\| - \|B\|\|x\| = (1-\|B\|)\|x\|$$

($\geq$: $\|Bx\| \leq \|B\|\|x\|$)

Damit gilt für $x \neq 0$, dass $(I+B)x \neq 0$ also $I+B$ regulär.

ii)

$$1 = \|I\| = \|(I+B)(I+B)^{-1}\| \overset{\text{erste Klammer auflösen}}{=} \|(I+B)^{-1} + B(I+B)^{-1}\|$$

wie oben: $\geq \|(I+B)^{-1}\| - \|B\|\,\|(I+B)^{-1}\| = \|(I+B)^{-1}\|(1-\|B\|) > 0$ da $\|B\| < 1$

($\|V+W-W\| \leq \|V+W\| + \|W\|$, $\Leftrightarrow \|V+W\| \geq \|V\| - \|W\|$)

Daraus folgt die Behauptung. ▨

Damit gilt der Folgende

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 des gestörten Systems $(A+\delta A)(x+\delta x) = b + \delta b$:

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

Beweis:

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

nach Hilfssatz 4.16 ist $A+\delta A$ regulär.

ii) $(A+\delta A)(x+\delta x) = b + \delta b$

$\Leftrightarrow Ax + \delta A x + (A+\delta A)\delta x = b + \delta b$

($Ax = b$: x Lösung von $Ax = b$)

$\Leftrightarrow (A+\delta A)\delta x = \delta b - \delta A x$

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

$$\|\delta x\| \leq \|(A+\delta A)^{-1}\| \{\|\delta b\| + \|\delta A\| \|x\|\}$$

Normregeln

$$= \|[A(I+A^{-1}\delta A)]^{-1}\| \{\|\delta b\| + \|\delta A\| \|x\|\}$$

$$= \|(I+A^{-1}\delta A)^{-1} A^{-1}\| \{\|\delta b\| + \|\delta A\| \|x\|\}$$

$$\leq \underbrace{\|(I+A^{-1}\delta A)^{-1}\|}_{\text{Hilfssatz 4.16}} \|A^{-1}\| \{\|\delta b\| + \|\delta A\| \|x\|\}$$

$$\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\delta A\|} \{\|\delta b\| + \|\delta A\| \|x\|\}$$

$$\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \|\delta A\|} \{\|\delta b\| + \|\delta A\| \|x\|\}$$

$\|A^{-1}\delta A\| \leq \|A^{-1}\| \|\delta A\| < 1$ nach Vor.!

mehr abziehen macht $1-x$ kleiner, also den Bruch größer!

ausgeklammert

$$= \frac{\|A^{-1}\| \|A\| \|x\|}{1 - \|A^{-1}\| \|\delta A\| \|A\| \|A\|^{-1}} \left\{ \frac{\|\delta b\|}{\|A\| \|x\|} + \frac{\|\delta A\|}{\|A\|} \right\}$$

$= 1$ hinzugefügt

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

$\|b\| = \|Ax\| \leq \|A\| \|x\|$

Nenner verkleinern macht Bruch größer

∎

Beispiel 4.18

Es sei $\frac{\|\delta A\|}{\|A\|} \approx 10^{-k}$ und $\frac{\|\delta b\|}{\|b\|} \approx 10^{-k}$ sowie $\text{cond}(A) \approx 10^{s}$.

Weiter nehmen wir an, dass $10^{s} \cdot 10^{-k} \ll 1$ also etwa $s-k \leq -3$.

Dann gilt

$$\frac{\|\delta x\|}{\|x\|} \leq \underbrace{\frac{10^{s}}{\underbrace{1-\underbrace{10^{s}10^{-k}}_{\ll 1}}_{\text{Nenner} \approx 1}}}_{\approx 1} \cdot 2 \cdot 10^{-k} \approx 10^{s-k} = 10^{-(k-s)}$$

$\Rightarrow$ man verliert s Stellen an Genauigkeit!

Eingabefehler ist in der k-ten Nachkommastelle, Fehler im Ergebnis ist in der $k-s$ ten Stelle. ∎

Man kann zeigen, dass diese Abschätzung im wesentlichen „scharf" ist [Ra]. S. 111.

Beispiel 4.19 (Kondition + Determinante) Evtl als Übung.

(a) Betrachte die 2×2 Matrix

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

Es gilt

$$\|A\|_\infty = \max(2, 2+\varepsilon), \quad \|A^{-1}\|_\infty = \frac{1}{\varepsilon}\max(2, 2+\varepsilon), \quad \text{cond}_\infty(A) = \frac{(2+\varepsilon)^2}{\varepsilon}.$$

Also $\text{cond}_\infty(A) = O\left(\frac{1}{\det(A)}\right)$.

Dies ist aber nicht immer so:

(b)

$$B = \begin{bmatrix} 10^{-10} & 0 \\ 0 & 10^{-10} \end{bmatrix}, \quad B^{-1} = \begin{bmatrix} 10^{10} & 0 \\ 0 & 10^{10} \end{bmatrix}$$

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

$\det(B) = 10^{-20}$! ∎

4.6 Positiv definite Matrizen

24.10.09

Wichtige Klasse von Matrizen mit vorteilhaften Eigenschaften.

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

a) $(Ax,x)_2 \in \mathbb{R} \quad \forall x \in \mathbb{K}^n \setminus \{0\} \quad (\mathbb{K}=\mathbb{C})$

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

Ziel: Charakterisierung positiv definiter Matrizen.

Im Fall $\mathbb{K}=\mathbb{C}$ bedeutet die Bedingung a) eine Einschränkung an die Matrix A.

Eigenschaft 4.12 Für $A \in \mathbb{C}^{n\times n}$ gilt A hermitesch genau dann wenn $(Ax,x)_2 \in \mathbb{R} \ \forall x \in \mathbb{C}^n \setminus \{0\}$.

Beweis „$\Rightarrow$" (als ü-Aufgabe!)

$(Ax,x)_2 \in \mathbb{R} \Leftrightarrow \underbrace{(Ax,x)_2}_{a+ib} = \underbrace{\overline{(Ax,x)_2}}_{a-ib}$ geht nur für $b=0$

also $(Ax,x)_2 \overset{\text{hermitesch 5.7}}{=} (x,Ax) \overset{(S1)}{=} \overline{(Ax,x)}$ damit ist „$\Rightarrow$" gezeigt.

„$\Leftarrow$" Aufteilen in Real- und Imaginärteil:

$\mathbb{C}^n \ni x = a+ib \quad a,b \in \mathbb{R}^n \qquad A = V+iW, \ V,W \in \mathbb{R}^{n\times n}$

Und ausrechnen:

$$(Ax,x)_2 = \big((V+iW)(a+ib), a+ib\big)_2 = \big((Va-Wb)+i(Vb+Wa), a+ib\big)_2$$

$$= \underbrace{(Va-Wb,a)_2 + (Vb+Wa,b)}_{\text{Re}} + i\underbrace{\big[(Vb+Wa,a)_2 - (Va-Wb,b)_2\big]}_{\text{Im}\ \overset{!}{\doteq} 0}$$

$(x,\alpha y)_2 = \overline{\alpha}(x,y)_2$! $\qquad i^2=-1$

also $\operatorname{Im}\big((Ax,x)_2\big) = (Vb,a)_2 + (Wa,a)_2 - (Va,b)_2 + (Wb,b)_2 \doteq 0 \qquad (*)$

Fall I) $W=0$ d.h. $A=V$ (reell), aber $x=a+ib$ ist komplex. Dann reduziert sich $(*)$ auf

$(Vb,a)_2 - (Va,b)_2 = 0 \ \forall a,b \in \mathbb{R}^n \Leftrightarrow (Vb,a)_2 = (V^Tb,a)_2 \ \forall a,b \in \mathbb{R}^n$ (alles reell)

damit muss $V=V^T$ sein.

Fall II) Setze $b=0$, dann reduziert sich (*) auf

$$(Wa,a)_2 = 0 \quad \forall a \in \mathbb{C}^n \setminus \{0\}$$

II a: Setze $a = e^i \Rightarrow w_{ii} = 0$. Hier: $e^i_j = \delta_{ij}$ kartesische Einheitsv.

II b: Sei also $w_{ii}=0 \ \forall i$ und setze $a = e^i + e^j$ $(j \neq i)$

dann folgt $w_{ij} + w_{ji} = 0 \quad \forall i,j$

$$\Leftrightarrow w_{ij} = -w_{ji}$$

Damit ist $A = V + iW$ mit $V = V^T$ und $W = -W^T$ also A hermitesch. ∎

Positiv definite Matrizen in $\mathbb{C}^{n\times n}$ sind also immer hermitesch.

Positiv definite Matrizen in $\mathbb{R}^{n\times n}$ sind nicht notwendigerweise symmetrisch.

Charakterisierung über Eigenwerte:

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

Beweis:

- A sei hermitesch mit lauter positiven Eigenwerte.

  $x \in \mathbb{C}^n$ hat Darstellung $x = \sum \alpha_i w^i$ $\quad w^i$: Eigenvektor zum EW λ_i.

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

- A sei hermitesch und positiv definit. A hat also Eigenpaare (λ_i, w^i) mit $\lambda_i \in \mathbb{R}$.
  Zu zeigen ist λ_i positiv

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

- Setze $e^i \in \mathbb{R}^n$, $e^i_j = \delta_{ij}$ kartesische Einheitsvektoren.

$$0 < (Aw_i, w_i)_2 = a_{ii} \in \mathbb{R}.$$ ∎

Die Aussage gilt natürlich auch für reelle symmetrische und positiv definite Matrizen, da auch diese hermitesch sind.

Speziell für reelle Matrizen gilt:

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

Beweis: e^i sei wieder der i-te kartesische Einheitsvektor und a_{ij}, $j \neq i$, das betragsmäßig größte Element. Dann zeigt man den Widerspruch.

$$0 < (A(e^i - \operatorname{sign}(a_{ij})e^j),\ e^i - \operatorname{sign}(a_{ij})e^j)_2$$

$$= (Ae^i, e^i)_2 - 2\operatorname{sign}(a_{ij})\underbrace{(Ae^i, e^j)_2}_{a_{ij}} + \operatorname{sign}(a_{ij})^2 (Ae^j, e^j)$$

$$\overset{A=A^T}{=} a_{ii} - 2|a_{ij}| + a_{jj} \leq 0 \quad \lightning$$

da $|a_{ij}| \geq a_{ii}$, $|a_{ij}| \geq a_{jj}$ ($a_{ii}, a_{jj} > 0$!) ∎

- Manche Autoren (auch Rannacher) verlangen, dass reelle positiv definite Matrizen symmetrisch sind. Wir fordern das extra

Ü $A \in \mathbb{R}^{n\times n}$ A ist positiv definit genau dann wenn $A_S = \frac{1}{2}(A + A^T)$ positiv definit ist.

A pos. definit, dann sind alle Hauptuntermatrizen (definiert) positiv definit

4.7 Störungstheorie

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

$F(A,b) = A^{-1}b$ ist der Lösungsoperator zur Gleichung $G(x) = Ax - b = 0$.

Betrachte die relative Kondition von F, wobei wir zunächst nur Änderungen in b zulassen wollen:

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

$$\leq \frac{\|A^{-1}\|\|\delta b\|}{\|\delta b\|} \frac{\|A\|\|A^{-1}b\|}{\|A^{-1}b\|} = \|A^{-1}\|\,\|A\|.$$

$\leq$: $\|\cdot\|$ sei verträglich und submultiplikativ

Definition 4.15 Die Zahl

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

für irgendeine verträgliche und submultiplikative Matrixnorm heißt Konditionszahl von A.

Nun wollen wir auch Änderungen in A selbst zulassen. Sei $A \in \mathbb{K}^{n\times n}$ regulär. Frage: Wann ist $A+\delta A$ regulär?

Hilfssatz 4.16 $B \in \mathbb{K}^{n\times n}$ habe Norm $\|B\| < 1$. Dann ist $I+B$ regulär und es gilt

$$\|(I+B)^{-1}\| \leq \frac{1}{1-\|B\|}.$$

Beweis: i) Für alle $x \in \mathbb{K}^n$ gilt

$$\|x\| = \|x + Bx - Bx\| \leq \|(I+B)x\| + \|Bx\|$$

$$\Leftrightarrow \|(I+B)x\| \geq \|x\| - \|Bx\| \geq \|x\| - \|B\|\|x\| = (1-\|B\|)\|x\|$$

($\geq$: $\|Bx\| \leq \|B\|\|x\|$)

Damit gilt für $x \neq 0$, dass $(I+B)x \neq 0$ also $I+B$ regulär.

ii)

$$1 = \|I\| = \|(I+B)(I+B)^{-1}\| \overset{\text{erste Klammer auflösen}}{=} \|(I+B)^{-1} + B(I+B)^{-1}\|$$

wie oben: $\geq \|(I+B)^{-1}\| - \|B\|\,\|(I+B)^{-1}\| = \|(I+B)^{-1}\|(1-\|B\|) > 0$ da $\|B\| < 1$

($\|V+W-W\| \leq \|V+W\| + \|W\|$, $\Leftrightarrow \|V+W\| \geq \|V\| - \|W\|$)

Daraus folgt die Behauptung. ▨

Damit gilt der Folgende

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 des gestörten Systems $(A+\delta A)(x+\delta x) = b + \delta b$:

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

Beweis:

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

nach Hilfssatz 4.16 ist $A+\delta A$ regulär.

ii) $(A+\delta A)(x+\delta x) = b + \delta b$

$\Leftrightarrow Ax + \delta A x + (A+\delta A)\delta x = b + \delta b$

($Ax = b$: x Lösung von $Ax = b$)

$\Leftrightarrow (A+\delta A)\delta x = \delta b - \delta A x$

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

$$\|\delta x\| \leq \|(A+\delta A)^{-1}\| \{\|\delta b\| + \|\delta A\| \|x\|\}$$

Normregeln

$$= \|[A(I+A^{-1}\delta A)]^{-1}\| \{\|\delta b\| + \|\delta A\| \|x\|\}$$

$$= \|(I+A^{-1}\delta A)^{-1} A^{-1}\| \{\|\delta b\| + \|\delta A\| \|x\|\}$$

$$\leq \underbrace{\|(I+A^{-1}\delta A)^{-1}\|}_{\text{Hilfssatz 4.16}} \|A^{-1}\| \{\|\delta b\| + \|\delta A\| \|x\|\}$$

$$\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\delta A\|} \{\|\delta b\| + \|\delta A\| \|x\|\}$$

$$\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \|\delta A\|} \{\|\delta b\| + \|\delta A\| \|x\|\}$$

$\|A^{-1}\delta A\| \leq \|A^{-1}\| \|\delta A\| < 1$ nach Vor.!

mehr abziehen macht $1-x$ kleiner, also den Bruch größer!

ausgeklammert

$$= \frac{\|A^{-1}\| \|A\| \|x\|}{1 - \|A^{-1}\| \|\delta A\| \|A\| \|A\|^{-1}} \left\{ \frac{\|\delta b\|}{\|A\| \|x\|} + \frac{\|\delta A\|}{\|A\|} \right\}$$

$= 1$ hinzugefügt

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

$\|b\| = \|Ax\| \leq \|A\| \|x\|$

Nenner verkleinern macht Bruch größer

∎

Beispiel 4.18

Es sei $\frac{\|\delta A\|}{\|A\|} \approx 10^{-k}$ und $\frac{\|\delta b\|}{\|b\|} \approx 10^{-k}$ sowie $\text{cond}(A) \approx 10^{s}$ $(s, k > 0)$

Weiter nehmen wir an, dass $10^{s} \cdot 10^{-k} \ll 1$ also etwa $s - k \leq -3$.

Dann gilt

$$\frac{\|\delta x\|}{\|x\|} \leq \underbrace{\frac{10^{s}}{1 - \underbrace{10^{s} 10^{-k}}_{\ll 1}}}_{\text{Nenner} \approx 1} \cdot 2 \cdot 10^{-k} \approx 10^{s-k} = 10^{-(k-s)}$$

$\Rightarrow$ man verliert s Stellen an Genauigkeit!

Eingabefehler ist in der k-ten Nachkommastelle, Fehler im Ergebnis ist in der $(k-s)$-ten Stelle. ∎

Man kann zeigen, dass diese Abschätzung im wesentlichen „scharf" ist [Ra]. S. 111.

Beispiel 4.19 (Kondition + Determinante) Evtl als Übung.

(a) Betrachte die 2×2 Matrix

$$A = \begin{bmatrix} -1 & 1 \\ 1-\varepsilon & -1 \end{bmatrix}, \quad A^{-1} = \frac{1}{\varepsilon} \begin{bmatrix} -1 & -1 \\ -(1-\varepsilon) & -1 \end{bmatrix}, \quad \varepsilon = \det(A), \ |\varepsilon| < 1$$

Es gilt

$$\|A\|_\infty = \max(2, 2-\varepsilon), \quad \|A^{-1}\|_\infty = \frac{1}{\varepsilon} \max(2, 2-\varepsilon), \quad \text{cond}_\infty(A) = \frac{(2+\varepsilon)^2}{\varepsilon}.$$

Also $\text{cond}_\infty(A) = O\left(\frac{1}{\det(A)}\right)$.

Dies ist aber nicht immer so:

(b)

$$B = \begin{bmatrix} 10^{-10} & 0 \\ 0 & 10^{-10} \end{bmatrix}, \quad B^{-1} = \begin{bmatrix} 10^{10} & 0 \\ 0 & 10^{10} \end{bmatrix}$$

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

$\det(B) = 10^{-20}$!

(c) $\text{cond}(\varepsilon A) = \text{cond}(A) \quad \forall \varepsilon \neq 0$, aber $\det(\varepsilon A) = \varepsilon^n \det(A)$

Beispiel 4.20 (oder Übung)

$$A = \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix}, \quad A^{-1} = \begin{bmatrix} d_1^{-1} & 0 \\ 0 & d_2^{-1} \end{bmatrix}, \quad d_1, d_2 \neq 0$$

$$\text{cond}_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty = \frac{\max\{|d_1|, |d_2|\}}{\min(|d_1|, |d_2|)}$$

d.h. $\text{cond}\left(\begin{bmatrix} 10^{10} & 0 \\ 0 & 1 \end{bmatrix}\right) = 10^{10}$!

Andererseits sind alle Gleichungen in $Ax = b$ unabhängig und $x_i = b_i / d_i$ ist gut konditioniert.

Problem liegt hier in der Verwendung von Normen, die Abschätzung für einzelne Komponenten ist relativ schlecht.

Man lernt: schlechte Kondition muss nicht unbedingt Probleme beim Lösen des LGS bewirken, es ist jedoch „wahrscheinlich".

für Schritt (2)

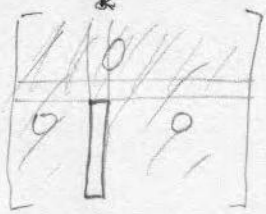
Wir definieren sog. Frobeniusmatrizen welche folgende Struktur haben:

$$G_k \in \mathbb{R}^{n\times n}, \quad 1 \le k < n$$

$$G_k = I + G_k' \quad \text{mit} \quad (G_k')_{ij} = 0 \quad \text{für } (j \neq k) \vee (i \le k),$$

d.h.

$$G_k = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & g_{k+1,k} & \ddots & \\ & & \vdots & & 1 \\ & & g_{n,k} & & \end{pmatrix} \quad (k)$$



Hilfssatz 5.5 Die Frobeniusmatrizen haben folgende Eigenschaften:

i) Für $\tilde{A} = G_k A$ gilt

$$\tilde{a}_{ij} = \begin{cases} a_{ij} & i \le k \\ a_{ij} + g_{ik} a_{kj} & i > k \end{cases}$$

(Dies ist gerade der Eliminationsschritt in der Gauß-Elimination wenn $g_{ik} := -a_{ik}$)

ii) $G_k' G_k' = 0$

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

iv) $G_1 G_2 \cdots G_k = I + \sum_{j=1}^{k} G_k'$, analog: $G_1^{-1} \cdots G_k^{-1} = I - \sum_{j=1}^{k} G_j'$

v) $\forall\, k < \alpha \le \beta$: $P_{\alpha\beta} G_k = (\underbrace{I + P_{\alpha\beta} G_k'}) P_{\alpha\beta}$ — $I + P_{\alpha\beta} G_k'$ ist eine geänderte Frobeniusmatrix!

Das ist genau die Situation in der GEM.

Beweis: Wird alles mit der Nullstruktur von G begründet:

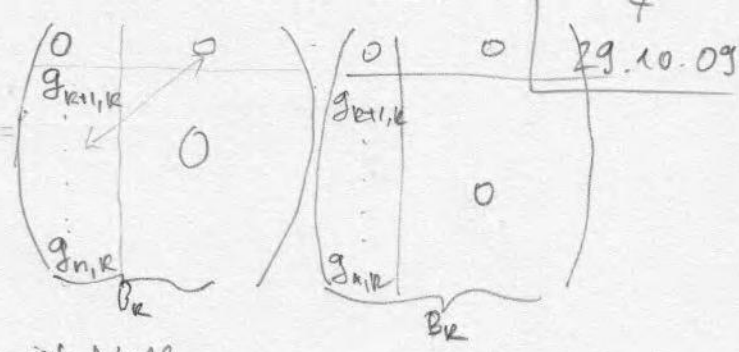
i) $\tilde{a}_{ij} = \sum_{s=1}^{n} (G_k)_{is} a_{sj} = \begin{cases} a_{ij} & \\ g_{ik} a_{kj} + a_{ij} & s = k \end{cases}$

$i \le k$ $(\to (G_k)_{ii} = 1$ else $s = k)$

sonst:
$(G_k)_{ik} = g_{ik}$
$(G_k)_{ii} = 1$

ii) $G_k' = \begin{pmatrix} 0 & 0 \\ 0 & B_k \end{pmatrix}$ ($k-1$ Zeilen, $n-k+1$ Zeilen; $k-1$ Spalten, $n-k+1$ Spalten)

$\begin{pmatrix} 0 & 0 \\ 0 & B_k \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 0 & B_k \end{pmatrix}$



Es genügt $B_k B_k$ zu betrachten, das ist Null

iii) $(\underbrace{I+G_k'}_{G_k})(I-G_k') = I - G_k' + G_k' - \underbrace{G_k' G_k'}_{=0 \text{ wg ii)}} = I$

iv) Per Induktion.

$\underline{k=1}$: $G_1 = I + G_1'$ hat die geforderte Gestalt

$\underline{k-1 \to k}$: Es sei also $G_1 \cdots G_{k-1} = I + \sum_{j=1}^{k-1} G_j'$ bereits gezeigt

$$G_1 \cdots G_k = \left(I + \sum_{j=1}^{k-1} G_j'\right)\left(I + G_k'\right) = I + G_k' + \sum_{j=1}^{k-1} G_j' + \underbrace{\left(\sum_{j=1}^{k-1} G_j'\right) G_k'}_{=0 \text{ ist zu zeigen}}$$

$\left(\sum_{j=1}^{k-1} G_j'\right) G_k' =$ (Zeilen: $k-1$, $n-k+1$; Spalten: $k-1$, $n-k+1$)

$$\begin{pmatrix} 0 & 0 \\ * & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & * \\ \vdots & \vdots \\ 0 & * \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ * & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & * \\ \vdots & \vdots \\ 0 & * \end{pmatrix} = 0$$

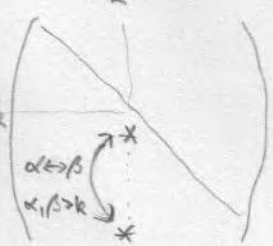
da einer der beiden Faktoren immer Null ist.

G_k, G_k^{-1} unterscheiden sich nur durch Vorzeichen Beweis hängt nur an Nullstrukturen

v) $\boxed{k < \alpha \leq \beta}$: $P_{\alpha\beta} G_k = P_{\alpha\beta}(I + G_k') = P_{\alpha\beta} + P_{\alpha\beta} G_k'$

$$= P_{\alpha\beta} + \underbrace{P_{\alpha\beta} G_k'}_{\text{nur Spalte } k \text{ besetzt}} \underbrace{P_{\alpha\beta}}_{\text{tauscht zwei Spalten } \alpha,\beta > k, \text{ Diese sind Null!}}$$

$$= \left(I + P_{\alpha\beta} G_k'\right) P_{\alpha\beta}$$

$P_{\alpha\beta} G_k' =$ 

Somit ist $\left(I + P_{\alpha\beta} G_k'\right)$ wieder Frob. Matrix

∎

Wir definieren das Produkt von Matrizen:

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

Beachte die Reihenfolge, da die Matrixmultiplikation nicht kommutativ ist

Satz 5.6 (LR-Zerlegung)

Sei $A \in \mathbb{R}^{n \times n}$ regulär, dann gibt es eine Zerlegung

$$PA = LR$$

wobei

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

und $P = \prod_{k=1}^{n-1} P_{k,r_k}$ ein Produkt von Permutationsmatrizen. mit $r_k \geq k$.

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

Beweis a) Wir behandeln zunächst den Fall $P = I$, d.h. ohne Zeilentausch.

Das Gaußsche Eliminationsverfahren lässt sich schreiben als

$$Ax = b$$

Schritt 1: $G_1 A x = G_1 b$ mit Frobeniusmatrix $(G_1)_{i1} = -q_{i1} = -\frac{\tilde{a}_{i1}^{(1)}}{\tilde{a}_{11}^{(1)}}$

Schritt 2: $G_2 G_1 A x = G_2 G_1 b$

nach $n-1$ Schritten $\underbrace{G_{n-1} \cdots G_1 A}\, x = G_{n-1} \cdots G_1 b$ mit $(G_k)_{ik} = -q_{ik}$ aus GEM

Ergebnis der GEM: $G_{n-1} \cdots G_1 A = R$ (rechte obere Dreiecksmatrix).

Nun nutze Hilfssatz 5.5 :

$$A = G_1^{-1} G_2^{-1} \cdots G_{n-1}^{-1} R$$

$$\overset{(iii)}{=} (I - G_1') \cdot (I - G_2') \cdots (I - G_{n-1}') R$$

$$\overset{(iv)}{=} \underbrace{\left(I - \sum_{j=1}^{n-1} G_j'\right)}_{=: L} R$$

L hat die geforderte Gestalt.

b) Nun mit den Zeilenvertauschungen. GEM liefert

$$G_{n-1} P_{n-1,r_{n-1}} \cdots P_{2,r_2} G_1 P_{1,r_1} A = R$$

$\alpha > \beta$ $\quad k$ $\quad k < \alpha \leq \beta$

r_k ist der in Schritt k bestimmte Tauschungsindex $r_k \geq k$.

Wende sukzessive HS 5.5 v) an

$$G_{n-1} P_{n-1,r_{n-1}} \cdots \overset{\curvearrowright}{P_{2,r_2} G_1} P_{1,r_1} A$$

$$= G_{n-1} P_{n-1,r_{n-1}} \cdots \overset{\curvearrowright}{P_{3,r_3} G_2 \left(I + P_{2,r_2} G_1'\right)} P_{2,r_2} P_{1,r_1} A$$

wieder Frobenius m.

$$= G_{n-1} P_{n-1,r_{n-1}} \overset{\curvearrowright}{P_{4,r_4} G_3} \left(I + P_{3,r_3} G_2'\right)\left(I + P_{3,r_3} P_{2,r_2} G_1'\right) P_{3,r_3} P_{2,r_2} P_{1,r_1} A$$

$$\vdots$$

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

„alle späteren" Tauschoperationen

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

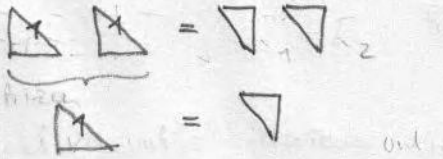
↑ wg Inverse

c) Eindeutigkeit. Es ist $P = I$, also $A = LR$.

Ang. es gebe zwei Zerlegung $A = L_1 R_1 = L_2 R_2$ der geforderten Gestalt.

$A^{-1} = R_1^{-1} L_1^{-1} = R_2^{-1} L_2^{-1}$, also

$$I = AA^{-1} = \underbrace{L_1 R_1}_{A} \underbrace{R_2^{-1} L_2^{-1}}_{A^{-1}} = L_1 R_1 R_2^{-1} L_2^{-1} \iff L_1^{-1} L_2 = R_1 R_2^{-1}$$



- Inverse ist wieder untere (obere Dreiecksm) (mit 1 auf der Hauptdiagonale)
- Produkt ist wieder

geht nur für $L_1^{-1} L_2 = R_1 R_2^{-1} = I$ ■

Wo ist L?

Aus dem Beweis sehen wir, dass

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

(Frobeniusmatrix)

$$= I - \sum_{k=1}^{n+1} \left(\prod_{\alpha=k+1}^{n-1} P_{\alpha, r_\alpha} \right) G'_k \qquad \text{mit } (G'_k)_{ik} = -g_{ik} \text{ aus GEM.}$$

- Nichtnullelemente in G'_k nehmen gerade den Platz der eliminierten Einträge $\tilde{a}_{ik}^{(k)}$ ein → kein zusätzlicher Speicher notwendig (A wird ohnehin überschrieben).
- Die G'_k nehmen an den späteren Zeilenvertauschen ebenso teil wie der Rest $A^{(k+1)}$. ⇒ ganze Zeile tauschen!

Wozu das Ganze?

Lösen von $Ax = b$ mittels LR-Zerlegung:

$$Ax = b$$
$$\iff PAx = Pb$$
$$\iff L \underbrace{Rx}_{=:y} = Pb$$
$$\iff \begin{cases} Ly = Pb & \text{vorwärts einsetzen} \\ Rx = y & \text{rückwärts einsetzen.} \end{cases}$$

also
1. Berechne LR-Zerl. von A und P
2. Für gegebenes b berechne $b' = Pb$
3. Löse $Ly = b'$
4. Löse $Rx = y$

Aufwand

Beobachtung: Transformation auf obere Dreiecksgestalt mittels GEM äquivalent zu LR-Zerlegung berechnen + $Ly = Pb$ lösen.

Somit: Der Aufwand zur Berechnung der LR-Zerlegung ist

$$N_{LR}(n) = \frac{2}{3}n^3 + O(n^2).$$

LR-Zerlegung ist insbesondere interessant, wenn $Ax = b_i$ zu mehreren reellen Seiten b_i zu lösen ist.

Zur Permutation

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

speichert man nicht als Matrix sondern man speichert nur die Zahlen $r_1, \ldots, r_{n-1}$ in einem Vektor (Feld).

Algorithmus zur LR-Zerlegung

Input: $A \in \mathbb{R}^{n \times n}$ (wird überschrieben)

Output: $L \in \mathbb{R}^{n \times n}$ in a_{ij}, $j < i$, $l_{ii} = 1$ implizit

$R \in \mathbb{R}^{n \times n}$ in a_{ij}, $j \geq i$

$p : \{0, \dots, n-1\} \to \{0, \dots, n-1\}$

```
for (k=1; k≤n; k=k+1) {
    Finde r∈{k,...,n} so dass a_rk ≠ 0; sonst Fehler;
    if (r≠k)  // tausche Zeilen
      for (j=1; j≤n; j=j+1) {
        t = a_kj; a_kj = a_rj; a_rj = t;
      }
    p[k] = r;  // merke Permutation

    for (i=k+1; i≤n; i=i+1) {
      a_ik = a_ik / a_kk;
      for (j=k+1; j≤n; j=j+1)
        a_ij = a_ij - a_ik a_kj;
    }
}
```

5.4 Rundungsfehleranalyse der Gauß-Elimination

(LR-Zerlegung)

10.11.09

Absolutwertnotation

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

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

Die Abbildung $rd : \mathbb{R} \to \mathbb{F}$ erweitern wir entsprechend auf $\mathbb{R}^n$, $\mathbb{R}^{m \times n}$.

Dann gilt

$$rd(A) = A + A' \quad \text{mit} \quad \underbrace{|A'| \leq |A| \, eps}_{\text{dies sind } m \cdot n \text{ Ungleichungen}}$$

wdh. aus Analysevar: rd

$$rd(a_{ij}) = a_{ij}(1+\varepsilon_{ij}) = a_{ij} + \underbrace{a_{ij}\varepsilon_{ij}}_{a'_{ij}}$$

$$|a'_{ij}| \leq |a_{ij}| \, eps$$

fl- Notation

→ nächstes Mal früher einführen

Stör: gl

Sei E eine Formel, dann bezeichne $fl(E)$ eine Berechnung der Formel E in Fließkommaarithmetik (dabei wird die Ausführungsreihenfolge meist (links nach rechts, Klammern) offensichtlich sein, sonst ist sie anzugeben).

Somit ist zum Beispiel

$$fl(\underbrace{A+B}_{A,B \in \mathbb{F}^{m \times n}}) = \underbrace{(A+B)}_{\text{exakte Rechg. in } \mathbb{R}} + H \quad \text{mit} \quad |H| \leq eps\,(|A+B|)$$

(Folgt aus

$$fl(a_{ij}+b_{ij}) = a_{ij} \oplus b_{ij}$$

$$= (a_{ij}+b_{ij})(1+\varepsilon_{ij}) \quad |\varepsilon_{ij}| \leq eps$$

$$= (a_{ij}+b_{ij}) + \varepsilon_{ij}(a_{ij}+b_{ij})$$

Man kann auch

$$fl(\sqrt{x})$$

oder

$$fl(\sin(x))$$

schreiben.

10.11.09

Rückwärtsanalyse

Bisher haben wir die „Vorwärtsanalyse" der Rundungsfehler betrieben.
Z.B. gilt für das Skalarprodukt $x^T y$:

$$|fl(x^T y) - x^T y| \leq n\,eps\,|x|^T|y| + O(eps^2) \quad \text{(absolut)}$$

oder

$$\frac{|fl(x^T y) - x^T y|}{|x^T y|} \leq n\,eps\,\frac{|x|^T|y|}{|x^T y|} + O(eps^2) \quad \text{(relativ)}.$$

Eine Alternative ist die sog. „Rückwärtsanalyse".
Dort versucht man das Fließkommaergebnis als exakter Ergebnis eines modifizierten Ausdrucks zu schreiben.

Beispiel 5.7 Betrachte Lösg des LGS $Ax = b$. Die GEM berechnet die numerische Lösung $\hat{x}$.

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

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

Mit dem Störungssatz 4.17 und $\|E\|_\infty = \||E|\|_\infty$ folgt dann für d. Rundungsfehler

$$\frac{\|\hat{x} - x\|_\infty}{\|x\|_\infty} \leq \frac{cond(A)}{1 - cond(A)\frac{\||E|\|_\infty}{\|A\|_\infty}} \underbrace{\frac{\||E|\|_\infty}{\|A\|_\infty}}$$

ist noch entsprechend abzuschätzen, ideal wäre $\||E|\|_\infty \leq n^p\,eps\,\|A\|_\infty$

Rundg.: $rd(A) = A + \delta A$, $|\delta A| \leq |A|\,eps$;
Störungssatz
$\frac{\|\delta A\|_\infty}{\|A\|_\infty} = \frac{\|\,|\delta A|\,\|_\infty}{\|A\|_\infty} \leq \frac{\|\,|A|\,\|_\infty\,eps}{\|A\|_\infty} \leq eps$

Somit ist ein direkter Vergleich mit der Konditionsanalyse möglich.

Rückwärtsanalyse des Skalarproduktes.

Hilfssatz 5.8 Es gilt für $x, y \in \mathbb{F}^n$:

$$\hat{s} = fl(x^T y) = (x+f)^T y, \qquad |f| \leq n\,eps\,|x| + O(eps^2)$$

Beweis: Induktion über n.

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

also $|f_1| = |\delta_1||x_1| \leq eps\,|x_1|$.

$n \geq 2$: $\hat{s}_n = fl(x^T y) = fl\left(fl(\tilde{x}^T \tilde{y}) + x_n y_n \right)$

$x = \begin{pmatrix} \tilde{x} \\ x_n \end{pmatrix}, y = \begin{pmatrix} \tilde{y} \\ y_n \end{pmatrix}$

$$= \left(fl(\tilde{x}^T\tilde{y}) + fl(x_n y_n) \right)(1+\varepsilon_n)$$

$$= \left((\tilde{x}+\tilde{f})^T\tilde{y} + x_n y_n (1+\delta_n) \right)(1+\varepsilon_n)$$

$$= (\tilde{x}+\tilde{f})^T\tilde{y}(1+\varepsilon_n) + x_n y_n \left(1 + (\delta_n+\varepsilon_n) + \delta_n\varepsilon_n\right)$$

$$= (\tilde{x} + \underbrace{\tilde{f} + \varepsilon_n\tilde{x} + \varepsilon_n\tilde{f}}\,)^T\tilde{y} + \left(x_n + (\delta_n+\varepsilon_n)x_n + \delta_n\varepsilon_n x_n\right) y_n$$

$$= \left[\underbrace{\begin{pmatrix} \tilde{x} \\ x_n \end{pmatrix}}_{=x} + \underbrace{\begin{pmatrix} \tilde{f} + \varepsilon_n\tilde{x} + \varepsilon_n\tilde{f} \\ (\delta_n+\varepsilon_n)x_n + \delta_n\varepsilon_n x_n \end{pmatrix}}_{=:f} \right] \underbrace{\begin{pmatrix} \tilde{y} \\ y_n \end{pmatrix}}_{=y}$$

mit $$|\tilde{f} + \varepsilon_n\tilde{x} + \varepsilon_n\tilde{f}| \leq (n-1)\,eps\,|\tilde{x}| + eps\,|\tilde{x}| + O(eps^2)$$
$$\leq n\,eps\,|\tilde{x}| + O(eps^2)$$

und $$|(\delta_n+\varepsilon_n)x_n + \delta_n\varepsilon_n x_n| \leq 2\,eps\,|x_n| + O(eps^2)$$

Wg $n \geq 2$ gilt also

$$|f| \leq n\,eps\,|x| + O(eps^2) \qquad \blacksquare$$

Bemerkung 5.9 $n\,eps$ in 5.8 ist der schlechteste Fall und sehr pessimistisch. Der Rundungsfehler in einer Operation hängt von den Argumenten ab und sie können sich auch wegheben (positiv/negativ!). Besser wäre eine statistische Betrachtung. $\blacksquare$

Satz 5.10 (Lösen von Dreieckssystemen)

Es seien $\hat{x}$ bzw $\hat{y}$ die numerischen Lösungen des unteren bzw. oberen Dreieckssystems $Lx = b$ und $Ry = c$. Dann gilt

$$(L+F)\hat{x} = b \qquad |F| \leq n\,\mathrm{eps}\,|L| + O(\mathrm{eps}^2)$$

$$(R+G)\hat{y} = c \qquad |G| \leq n\,\mathrm{eps}\,|R| + O(\mathrm{eps}^2)$$

Beweis: (evtl als Übung). Induktion über n.

$n=1$ $\quad \ell_{11} x_1 = b_1 \rightarrow \hat{x}_1 = fl(b_1/\ell_{11}) = (b_1/\ell_{11})(1+\varepsilon_1)$

$$\Leftrightarrow \frac{1}{1+\varepsilon_1}\hat{x}_1 = \frac{b_1}{\ell_{11}} \Leftrightarrow \frac{\ell_{11}}{1+\varepsilon_1}\hat{x}_1 = b_1$$

$\frac{1}{1+\varepsilon_1} = 1+$

$\partial_\varepsilon \frac{1}{1+\varepsilon_1} = -\frac{1}{(1+\varepsilon_1)^2}$

$$\Leftarrow \left(1-\varepsilon_1+\tilde{R}(\varepsilon^2)\right)\ell_{11}\hat{x}_1 = b_1$$

$$\Leftrightarrow \left(\ell_{11}\underbrace{-\varepsilon_1\ell_{11} + \ell_{11}\tilde{R}(\varepsilon^2)}_{=f_1}\right)\hat{x}_1 = b_1$$

mit $|f_1| \leq \mathrm{eps}\,|\ell_{11}| + O(\mathrm{eps}^2)$

$n \geq 2$

$$Lx = b \Leftrightarrow \begin{matrix} n-1 \\ 1 \end{matrix}\begin{pmatrix} L_1 & 0 \\ v^T & \alpha \end{pmatrix}\begin{pmatrix} x_1 \\ w \end{pmatrix} = \begin{pmatrix} b_1 \\ \beta \end{pmatrix} \quad \begin{matrix} n-1 \\ 1 \end{matrix}$$

Berechne Lösung $\hat{x}_1$ von $L_1 x_1 = b_1$ und rechne

$$\hat{w} = fl\left((\beta - v^T\hat{x}_1)/\alpha\right)$$

$$= \left(fl(\beta - v^T\hat{x}_1)/\alpha\right)(1+\varepsilon_n) \quad \text{(Fehler in } /\alpha\text{)}$$

$$= \left((\beta - fl(v^T\hat{x}_1))(1+\delta_n)/\alpha\right)(1+\varepsilon_n) = \left((\beta - fl(v^T\hat{x}_1))/\alpha\right)(1+\delta_n)(1+\varepsilon_n) \quad \text{(Skalarprodukt)}$$

HS. 5.8 ↓

$$= \left((\beta - (v+f)^T\hat{x}_1|/\alpha\right)\left(1+(\varepsilon_n+\delta_n)+\varepsilon_n\delta_n\right)$$

Mit $\frac{1}{1+(\varepsilon_n+\delta_n)+\varepsilon_n\delta_n} = 1-(\varepsilon_n+\delta_n)+R(\varepsilon_n^2+\delta_n^2+\varepsilon_n\delta_n)$ gilt

$$\left(1-(\varepsilon_n+\delta_n)+R(..)\right)\hat{w} = \left(\beta-(v+f)^T\hat{x}_1\right)/\alpha$$

$$\Leftrightarrow \quad (v+f)^T\hat{x}_1 + \left(1-(\varepsilon_n+\delta_n)+R(...)\right)\alpha\hat{w} = \beta.$$

Mit der Induktionsvoraussetzung $(L_1+F_1)\hat{x}_1 = b_1$ folgt also für $\hat{x}$:

$$\begin{pmatrix} L_1+F_1 & 0 \\ (v+f)^T & \left(1-(\varepsilon_n+\delta_n)+R(..)\right)\alpha \end{pmatrix}\begin{pmatrix}\hat{x}_1\\ \hat{w}\end{pmatrix} = \begin{pmatrix} b_1 \\ \beta\end{pmatrix}$$

$$\Leftrightarrow \left[\underbrace{\begin{pmatrix} L_1 & 0 \\ v^T & \alpha\end{pmatrix}}_{=L} + \underbrace{\begin{pmatrix} F_1 & 0 \\ f^T & \left(-(\varepsilon_n+\delta_n)+R(...)\right)\alpha\end{pmatrix}}_{=F}\right]\begin{pmatrix}\hat{x}_1\\ \hat{w}\end{pmatrix} = \begin{pmatrix} b_1 \\ \beta\end{pmatrix}$$

Wegen $|F_1| \le (n-1)\,\text{eps}\,|L| + O(\text{eps}^2)$

$|f|^T \le (n-1)\,\text{eps}\,|v|^T + O(\text{eps}^2)$ — hier kommt das $n-1$ her! nicht durch die Reduktion. es geht also nicht besser

$|(-(\varepsilon_n+\delta_n)+R(...))\alpha| \le 2\,\text{eps}\,|\alpha| + O(\text{eps}^2)$

gilt für $n \ge 2$ dann

$$|F| \le n\,\text{eps}\,|L| + O(\text{eps}^2)$$

(möglich wäre auch $\max(n-1, 2)$, ist aber nicht wirklich besser).

Analog für $Ry = c$. ∎

10. 11. 09

Satz 5.11 (Rückwärtsanalyse der LU-Zerlegung)

Sei $A \in \mathbb{F}^{n\times n}$. Es werde die LR-Zerlegung ohne Pivotierung berechnet. Dann gilt für die numerisch berechneten $\hat{L}, \hat{R}$:

$$\hat{L}\hat{R} = A + H \quad \text{mit} \quad |H| \le 3(n-1)\,\text{eps}\left(|A| + |\hat{L}||\hat{R}|\right) + O(\text{eps}^2)$$

Beweis: [Golub/van Loan, THM 3.3.1] Induktion über n.

$\underline{n=1}$: Es ist $\hat{l}_{11} = 1$ und $\hat{r}_{11} = a_{11}$ und damit $H = 0$.

$\underline{n \ge 2}$: Schreibe

$$A = \begin{bmatrix} \overset{1}{\alpha} & \overset{n-1}{w^T} \\ v & B \end{bmatrix}\begin{matrix}1\\ n-1\end{matrix}$$

Dann macht die LR-Zerlegung:

a) $\hat{z} = fl(v/\alpha)$

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

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

Fehler in $\hat{z}$:

$$(\square_1) \quad \hat{z} = \frac{v}{\alpha} + f \quad \text{mit} \quad |f| \le \text{eps}\,\frac{|v|}{|\alpha|}$$

Fehler in $\hat{A}_1$:

$$\hat{A}_1 = fl(B - \hat{z}w^T) \qquad \text{(exakt gerundet)}$$

$$= B - fl(\hat{z}w^T) + G \qquad |G| \le \text{eps}\,|B - fl(\hat{z}w^T)|$$

(Bistritt)

$$= B - (\hat{z}w^T + G') + G \qquad |G'| \le \text{eps}\,|\hat{z}w^T| \le \text{eps}\,|\hat{z}||w|^T$$

(äußeres Produkt! genau eine Operation pro Matrixeintrag.)

$$(\square_2) \quad \hat{A}_1 = B - \hat{z}w^T + F \quad \text{mit} \quad F = -G' + G$$

$$|F| = |-G' + G| \le |G'| + |G|$$

$$\le \underbrace{\text{eps}\,|\hat{z}||w|^T}_{|G'|} + \text{eps}\,|B - \hat{z}w^T - G'|$$

$$\le \text{eps}\,|\hat{z}||w|^T + \text{eps}\left(|B| + |\hat{z}||w|^T\right) + \text{eps}^2\,|\hat{z}||w|^T$$

also $$|F| \le 2\,\text{eps}\left(|B| + |\hat{z}||w|^T\right) + O(\text{eps}^2)$$

Nun wird $\hat{A}_1$ LR-zerlegt und es gilt die Induktionsannahme

10.11.09

$(\square_3)$ $\quad \hat{L}_1 \hat{R}_1 = \hat{A}_1 + H_1$ mit $|H_1| \le 3(n-2)\,\mathrm{eps}\,(|\hat{A}_1| + |\hat{L}_1||\hat{R}_1|) + O(\mathrm{eps}^2)$

also ist die Blockform der LR-Zerlegung von A:

$$\hat{L}\hat{R} = \begin{bmatrix} 1 & 0 \\ \hat{z} & \hat{L}_1 \end{bmatrix} \begin{bmatrix} \alpha & w^T \\ 0 & \hat{R}_1 \end{bmatrix} = \begin{bmatrix} \alpha & w^T \\ \hat{z}\alpha & \hat{z}w^T + \underbrace{\hat{L}_1 \hat{R}_1}_{= \hat{A}_1 + H_1} \end{bmatrix}$$

Darstellung von oben einsetzen $(\square_1), (\square_2), (\square_3)$

$$= \begin{bmatrix} \alpha & w^T \\ \left(\frac{v}{\alpha} + f\right)\alpha & \hat{z}w^T + \underbrace{\left(B - \hat{z}w^T + F\right)}_{= \hat{A}_1} + H_1 \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} \alpha & w^T \\ v & B \end{bmatrix}}_{= A} + \underbrace{\begin{bmatrix} 0 & 0 \\ \alpha f & H_1 + F \end{bmatrix}}_{=: H}$$

Damit hat man schon mal die Form $\hat{L}\hat{U} = A + H$. Bleibt noch H abzuschätzen.

In H steckt $H_1 + F$, in $|H_1|$ steckt $|\hat{A}_1|$ also:

$$|\hat{A}_1| = |B - \hat{z}w^T + F| \le |B| + |\hat{z}||w|^T + |F|$$

Absch. für $|F|$ $\rightarrow \le |B| + |\hat{z}||w|^T + 2\,\mathrm{eps}\,(|B| + |\hat{z}||w|^T) + O(\mathrm{eps}^2)$

$$\le (1 + 2\,\mathrm{eps})(|B| + |\hat{z}||w|^T) + O(\mathrm{eps}^2)$$

$$|H_1 + F| \le |H_1| + |F|$$

Induktionsannahme $\rightarrow \le 3(n-2)\,\mathrm{eps}\,(|\hat{A}_1| + |\hat{L}_1||\hat{R}_1|) + 2\,\mathrm{eps}\,(|B| + |\hat{z}||w|^T) + O(\mathrm{eps}^2)$

Abschätzung für $|\hat{A}_1|$ einsetzen $\rightarrow \le 3(n-2)\,\mathrm{eps}\left[(1 + 2\,\mathrm{eps})(|B| + |\hat{z}||w|^T) + |\hat{L}_1||\hat{R}_1|\right] + 2\,\mathrm{eps}\,(|B| + |\hat{z}||w|^T) + O(\mathrm{eps}^2)$

$$\le 3(n-1)\,\mathrm{eps}\left[|B| + |\hat{z}||w|^T + |\hat{L}_1||\hat{R}_1|\right] + O(\mathrm{eps}^2)$$

Und damit

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

$(\Box_1)$

$$\leq 3(n-1)\,\text{eps} \begin{bmatrix} 0 & 0 \\ |v| & |B| + |\hat{z}||w|^T + |\hat{L}_1||\hat{R}_1| \end{bmatrix} + O(\text{eps}^2)$$

$$\leq 3(n-1)\,\text{eps}\left(\underbrace{\begin{bmatrix} |\alpha| & |w|^T \\ |v| & |B| \end{bmatrix}}_{|A|} + \underbrace{\begin{bmatrix} 1 & 0 \\ |\hat{z}| & |\hat{L}_1| \end{bmatrix}}_{|\hat{L}|} \underbrace{\begin{bmatrix} |\alpha| & |w|^T \\ 0 & |\hat{R}_1| \end{bmatrix}}_{|\hat{R}|} \right)$$

○ Terme sind hinzugefügt.

$0 \leq (1.1)$

$$\begin{bmatrix} |\alpha| & |w|^T \\ |\hat{z}||\alpha| & |\hat{z}||w|^T + |\hat{L}_1||\hat{R}_1| \end{bmatrix}$$

Nun sind noch die Dreieckssysteme aufzulösen.

Satz 5.12

Seien $\hat{L}$ und $\hat{R}$ die numerisch berechnete LR-Zerlegung von $A \in \mathbb{F}^{n\times n}$ aus Satz 5.11. Sei weiter $\hat{y} \in \mathbb{F}^n$ die numerische Lösung von $\hat{L}y = b$ und schließlich $\hat{x} \in \mathbb{F}^n$ die numerische Lösung von $\hat{R}x = \hat{y}$. Dann gilt für $\hat{x}$ die Beziehung

$$(A+E)\hat{x} = b$$

mit

$$|E| \leq n\,\mathrm{eps}\,(3|A| + 5|\hat{L}||\hat{R}|) + O(\mathrm{eps}^2).$$

Beweis:

Wg Satz 5.10 gilt

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

Einsetzen liefert:

$$(\hat{L}+F)\hat{y} = (\hat{L}+F)(\hat{R}+G)\hat{x} = (\underbrace{\hat{L}\hat{R}}_{=A+H} + F\hat{R} + \hat{L}G + FG)\hat{x} = b$$

Wg Satz 5.11 gilt

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

also

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

und

$$|E| \leq |H| + |F||\hat{R}| + |\hat{L}||G| + O(\mathrm{eps}^2)$$

$$\leq \underbrace{3(n-1)\,\mathrm{eps}\,(|A| + |\hat{L}||\hat{R}|)}_{|H|} + \underbrace{n\,\mathrm{eps}\,|\hat{L}||\hat{R}|}_{|F|} + \underbrace{n\,\mathrm{eps}\,|\hat{L}|\,|\hat{R}|}_{|G|} + O(\mathrm{eps}^2)$$

$$\leq 3n\,\mathrm{eps}\,|A| + 5n\,\mathrm{eps}\,|\hat{L}||\hat{R}| + O(\mathrm{eps}^2). \qquad \blacksquare$$

Folgerung

Nach Satz 5.12 ist $\hat{x}$ exakte Lösung des modifizierten Systems

$$(A+E)\hat{x} = b.$$

Mit dem Störungssatz gilt dann in $\|\cdot\|_\infty$-Norm (Beachte: $\|B\|_\infty = \||B|\|_\infty$)

$$\frac{\|\hat{x}-x\|_\infty}{\|x\|_\infty} \leq \mathrm{cond}(A) \cdot \left\{ \underbrace{3n\,\mathrm{eps}\,\frac{\|A\|_\infty}{\|A\|_\infty}}_{=1} + 5n\,\mathrm{eps}\,\underbrace{\frac{\|\hat{L}\|\,\|\hat{R}\|}{\|A\|_\infty}}_{\text{Problem!}} + O(\mathrm{eps}^2) \right\}$$

$3n\,\mathrm{eps}$: vergleichbar mit Rundungsfehler $\|\mathrm{rd}(A)\|/\|A\|$ O.K.

Nehmen an, dass $\|E\|_\infty / \|A\|_\infty \ll 1$ (brauchen ohnehin $\|E\| < \frac{1}{\|A^{-1}\|}$)

$\rightsquigarrow$ Nenner im Vorfaktor ist ≈ 1

- Erster Term vergleichbar mit dem aus der Konditionsanalyse
- Zweiter Term ist möglicherweise problematisch.

$\hat{L}$ enthält Einträge der Form $\frac{\tilde{a}_{ij}}{\tilde{a}_{ii}}$, also $\frac{1}{\text{Pivot element}}$

|Pivotelement| klein $\Rightarrow$ $|\hat{L}|$ groß $\Rightarrow$ großer Rundungsfehler!

- Dies kann trotz guter Kondition von A passieren!

Beispiel $A = \begin{bmatrix} \varepsilon & 1 \\ 1 & 0 \end{bmatrix}$

$\Rightarrow$ Gauß-Elimination (LR-Zerlegung) ist in dieser Form nicht numerisch stabil!

5.5 Pivotisierung

16.11.09

Die Rundungsfehleranalyse in Satz 5.12 führt auf den unvorteilhaften Term $|\hat{L}||\hat{U}|$.

Mit der Wahl von r im Gauß-Algorithmus so dass

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

gilt dann

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

Diese Wahl nennt man „Spaltenpivotisierung".

Beispiel 5.13 Aus [GvL]

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

In exakter Arithmetik führt das Gaußsche Verfahren nach Elimination von a_{21} auf

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

mit der Lösung

$$x_1 = -0.4999975, \qquad x_2 = 0.999995 \quad .$$

Nun führen wir das Verfahren in $\mathbb{F}(10, 4, 1)$ durch. Beim Multiplikator

$$q_{21} = (0.2 \cdot 10^1) \oslash (-0.1 \cdot 10^{-4}) = -0.2 \cdot 10^6$$

ergibt sich kein Rundungsfehler.

Für das neue a_{22} ergibt sich

$$\begin{aligned} a_{22}^{(1)} &= 0.1 \cdot 10^1 \ominus (-0.2 \cdot 10^6) \odot (0.1 \cdot 10^1) \\ &= 0.1 \cdot 10^1 \oplus 0.2 \cdot 10^6 = \boxed{0.2 \cdot 10^6}. \end{aligned}$$

Hier wurde auf vier Stellen gerundet.

Damit ergibt sich (ohne Fehler)

$$b_2^{(1)} = -(-0.2 \cdot 10^6) \odot (0.1 \cdot 10^1) = 0.2 \cdot 10^6$$

und

$$\begin{aligned} x_2 &= b_2^{(1)} \oslash a_{22}^{(1)} = 0.2 \cdot 10^6 \oslash 0.2 \cdot 10^6 = \boxed{1}, \\ x_1 &= (0.1 \cdot 10^1 \ominus 0.1 \cdot 10^1 \odot 1) \oslash (-0.1 \cdot 10^{-4}) = \boxed{0} \quad . \end{aligned}$$

2
16.11.09

Es ist also *keine* Stelle im Ergebnis korrekt obwohl nur an einer *einzigen* Stelle (in der Berechnung von $a_{22}^{(1)}$) ein Rundungsfehler eingeführt wurde.

Darüberhinaus überprüfe man, dass für die Kondition von A gilt:

$$\kappa(A) = 3 \quad .$$

Demnach ist das System gut konditioniert! Der Algorithmus, so wie er ist, ist numerisch nicht stabil.

Das Problem ist offensichtlich der große Multiplikator q_{21} der aus dem sehr kleinen a_{11} resultiert und der dafür sorgt, dass das ursprüngliche a_{22} in $a_{22}^{(1)}$ vollkommen ignoriert wird.

Im Prinzip haben wir in Fließkommaarithmetik das System

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

exakt gelöst, was eine völlig andere Lösung hat als das ursprüngliche (11.1) (Rückwärtsanalyse).

Der große Multiplikator kann ganz einfach vermieden werden indem man eine Zeilenvertauschung durchführt, d. h. wir lösen

$$\begin{bmatrix} 2 & 1 \\ -10^{-5} & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{11.2}$$

Nun erhält man

$$\begin{aligned}
q_{21} &= -0.1 \cdot 10^{-4} \oslash 0.2 \cdot 10^{1} = -0.5 \cdot 10^{-5}, \\
a_{22}^{(1)} &= 0.1 \cdot 10^{1} \ominus (-0.5 \cdot 10^{-5}) \odot 0.1 \cdot 10^{1} = 0.1 \cdot 10^{1} \oplus 0.5 \cdot 10^{-5} = 0.1 \cdot 10^{1}, \\
b_2^{(1)} &= 0.1 \cdot 10^{1} \ominus 0.5 \cdot 10^{-5} \odot 0 = 0.1 \cdot 10^{1}, \\
x_2 &= 0.1 \cdot 10^{1} \oslash 0.1 \cdot 10^{1} = \boxed{1}, \\
x_1 &= (0 \ominus 0.1 \cdot 10^{1} \odot 0.1 \cdot 10^{1}) \oslash 0.2 \cdot 10^{1} = \boxed{-0.5},
\end{aligned}$$

was in $\mathbb{F}(10, 4, 1)$ völlig in Ordnung ist. □

Spaltenpivotisierung ist nicht ausreichend wie folgendes Beispiel zeigt.

Forts. von Beispiel 5.13

(ebenfalls aus [GO96]). Wir betrachten das 2×2 System

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

welches aus (11.1) durch Multiplikation der ersten Zeile mit -10^6 entsteht.

Die Spaltenpivotisierung erfordert keine Vertauschung. Allerdings entsteht für $a_{22}^{(1)} = 1 + 2 \cdot 10^5$ genau dasselbe Problem wie oben! □

16.11.09

Diese Probleme kann man durch eine Skalierung des Gleichungssystems vermindern:

$$Ax = b \rightarrow D^{-1}Ax = D^{-1}b \quad \text{mit } d_{ii} = \sum_{j=1}^{n} |a_{ij}|$$
$$\Leftrightarrow \tilde{A}x = \tilde{b}$$
$$(\text{also } \|\tilde{A}\|_\infty = 1)$$

Rundungsfehleranalyse bei Pivotisierung

Analog zu Satz 5.12 zeigt man, dass für die Lösung $\hat{x}$ bei Spaltenpivotisierung gilt:

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

Nach Konstruktion ist $\|L\|_\infty \leq n$ und mit der Definition

$$\varrho = \frac{\max_{i,j,k} |\hat{a}_{ij}^{(k)}|}{\|A\|_\infty} \quad \text{„Wachstumsfaktor"}$$

Zeigt man [GvL]

$$\|E\|_\infty \leq 8n^3 \|A_\infty\| \varrho\,\text{eps} + O(\text{eps}^2).$$

In der Praxis ist $\varrho \approx 10$

schlechtester Fall bei Spaltenpivotisierung ist $\varrho = 2^{n-1}$.

Mit totaler Pivotisierung erreicht man

$$|a_{ij}^{(k)}| \leq k^{1/2}\left(2 \cdot 3^{1/2} \cdots k^{1/(k-1)}\right)^{1/2} \max |a_{ij}|$$

also deutlich kleineres Wachstum.

Totale Pivotisierung

Wähle $r, s \in \{k, \dots, n\}$ so dass

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

und erreiche durch Zeilen- und Spaltenvertauschung, dass

$$\tilde{a}_{kk}^{(k)} = a_{rs}^{(k)}.$$

In Matrixform:

Schritt 1: $$G_1 P_{r_1} A P_{s_1} \underbrace{P_{s_1} x} = G_1 P_{r_1} b$$

Schritt 2: $$G_2 P_{r_2} G_1 P_{r_1} A P_{s_1} P_{s_2} P_{s_2} P_{s_1} x = G_2 P_{r_2} G_1 P_{r_1} b$$

$$\vdots$$

Schritt n und Umformulierung

$$\underbrace{\underbrace{G_n' \cdots G_1'}_{P} \underbrace{P_{r_n} \cdots P_{r_1}}_{P} A \underbrace{P_{s_1} \cdots P_{s_n}}_{Q^T}}_{=R} \underbrace{\overbrace{P_{s_n} \cdots P_{s_1}}^{Q} x}_{=:z} = G_n' \cdots G_1' P_{r_n} \cdots P_{r_1} b$$

und damit

$$PAQ^T z = Pb$$

$$PAQ^T = LR \qquad z = Qx$$

Lösen des LGS gelingt dann mit

$$b' = Pb$$
$$Ly = b'$$
$$Rz = y$$
$$x = Q^T z$$

Aufwand der Pivotisierung:

- $n^2/2$ Vergleiche bei Spaltenpivotisierung
- $n^3/3$ Vergleiche bei totaler Pivotisierung

Da Speicherzugriffe teurer als eigentliche Rechg $\Rightarrow$ Verdopplung des Zeitbedarfs bei totaler Pivotisierung.

Praktische Erfahrung zeigt keine Vorteile für Rundungsfehler bei totaler Pivotisierung

$\hookrightarrow \Rightarrow$ Spaltenpivotisierung mit Zeilenskalierung ist effizient und numerisch stabil.

16.11.09

5.6 Spezielle Systeme

Symmetrisch positiv definite Matrizen

Satz 5.14 Eine symmetrisch positiv definite Matrix $A \in \mathbb{R}^{n\times n}$ ist stets ohne Pivotisierung LR-zerlegbar. Für die Diagonalelemente der im Eliminationsprozess auftretenden Matrizen gilt

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

Beweis: Betrachte einen Schritt in der LR-Zerlegung

1).

$$A = \begin{bmatrix} \alpha & v^T \\ v & B \end{bmatrix} \begin{matrix} 1 \\ n-1 \end{matrix}$$

$\qquad\qquad 1 \quad n-1$

Elimination der Spalte v liefert

$$\begin{bmatrix} 1 & 0 \\ -v/\alpha & I \end{bmatrix} \underbrace{\begin{bmatrix} \alpha & v^T \\ v & B \end{bmatrix}}_{A} = \begin{bmatrix} \alpha & v^T \\ 0 & B - \frac{1}{\alpha} v v^T \end{bmatrix}$$

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

Mit $\begin{bmatrix} 1 & v^T/\alpha \\ 0 & I \end{bmatrix}^{-1} = \begin{bmatrix} 1 & -v^T/\alpha \\ 0 & I \end{bmatrix}$ folgt:

$$\underbrace{\begin{bmatrix} 1 & 0 \\ -v/\alpha & I \end{bmatrix}}_{X^T} \underbrace{\begin{bmatrix} \alpha & v^T \\ v & B \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} 1 & -v^T/\alpha \\ 0 & I \end{bmatrix}}_{X} = \underbrace{\begin{bmatrix} \alpha & 0 \\ 0 & B - \frac{1}{\alpha} v v^T \end{bmatrix}}_{\tilde{A}}$$

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

$B - \frac{1}{\alpha} v v^T$ ist Hauptuntermatrix von $\tilde{A}$ und somit auch sym. pos. definit.

16-11.09

b) Es gilt (Raleigh-Quotient):

$$(Ax, x)_2 \geq \lambda_{min}(A)\,(x, x)_2$$

und damit für $x = e^{(i)}$ (kartesischer Einheitsvektor)

$$a_{ii} = (Ae^{(i)}, e^{(i)})_2 \geq \lambda_{min}(A)\,\underbrace{(e^{(i)}, e^{(i)})_2}_{=1} = \lambda_{min}(A).$$

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

$$\left(B - \frac{1}{\alpha} v v^T\right)_{ii} = \left(\tilde{A}\tilde{e}^{(i)}, \tilde{e}^{(i)}\right)_2 = \left(X^T A X \tilde{e}^{(i)}, \tilde{e}^{(i)}\right)_2$$

$$= \left(AX\tilde{e}^{(i)}, X\tilde{e}^{(i)}\right)_2$$

$$\geq \lambda_{min}(A)\left(X\tilde{e}^{(i)}, X\tilde{e}^{(i)}\right)_2 = \lambda_{min}(A)\left(1 + \frac{v_i^2}{\alpha^2}\right)$$

$$\underbrace{\begin{bmatrix} 1 & -v^T/\alpha \\ 0 & I \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} 0 \\ e^{(i)} \end{bmatrix}}_{\tilde{e}^{(i)}} = \begin{bmatrix} -\frac{v_i}{\alpha} \\ e^{(i)} \end{bmatrix} \geq \lambda_{min}(A).$$

∎

Dies zeigt, dass die Pivotelemente bei der Gauß-Elimination echt nach unten beschränkt bleiben.

ü) Rundungsfehleranalyse. Kann man $|\hat{L}|\,|\hat{U}|$ abschätzen?
↝ Ramacher Lemma 4.1.

16.11.09

Die Pivotelemente die bei der LR-Zerlegung auftreten sind nach 5.14 stets positiv.

Mit $D = \mathrm{diag}(R)$ gilt

$$A = L D \underbrace{D^{-1} R}_{U} = L D U$$

U ist obere Dreiecksmatrix mit $u_{ii} = 1$. Wegen der Symmetrie gilt $U = L^T$, also

$$A = L D L^T.$$

Da $d_{ii} > 0$ ist die Matrix „$D^{1/2}$"

$$(D^{1/2})_{ii} = \sqrt{d_{ii}}$$

wohldefiniert und es gilt

$$A = \underbrace{L D^{1/2}}_{=:\tilde{L}} \underbrace{D^{1/2} L^T}_{\tilde{L}^T} = \tilde{L}\,\tilde{L}^T.$$

Dies ist die sog. Cholesky-Zerlegung einer symmetrisch positiv definiten Matrix.

Ausnutzung der Symmetrie erlaubt die Berechnung der Cholesky-Zerlegung in $n^3/3 + O(n^2)$ Operationen (Faktor 2 schneller).

Diagonaldominante Matrizen

9
16.11.09

Definition 5.15 Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt diagonaldominant falls

$$\sum_{j=1, j\neq i}^{n} |a_{ij}| \leq |a_{ii}| \qquad i = 1, \dots, n.$$

Satz 5.16 Diagonaldominante, reguläre Matrizen erlauben eine LR-Zerlegung ohne Pivotisierung.

Beweis. $A = \begin{bmatrix} \alpha & w^T \\ v & B \end{bmatrix}$, ein Schritt der GEM liefert

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

Da $\alpha \neq 0$ wg. Diagonaldominanz ist dies wohldefiniert.

Aus der Diagonaldominanz von A ergibt sich:

(Zeile 1) $$\sum_{j=1}^{n-1} |w_j| \leq |\alpha| \iff \sum_{j=1}^{n-1} \frac{|w_j|}{|\alpha|} \leq 1$$

(Zeile 2...n) $$|v_i| + \sum_{\substack{j=1 \\ j\neq i}}^{n-1} |b_{ij}| \leq |b_{ii}|$$

Für $B - \frac{1}{\alpha} v w^T$ rechnen wir dann nach:

$$\sum_{\substack{j=1 \\ j\neq i}}^{n-1} \left| b_{ij} - \frac{1}{\alpha} v_i w_j \right| + \left| \frac{v_i w_i}{\alpha} \right| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n-1} |b_{ij}| + |v_i| \sum_{\substack{j=1 \\ j\neq i}}^{n-1} \frac{|w_j|}{|\alpha|} + |v_i| \frac{|w_i|}{|\alpha|}$$

$$= \sum_{\substack{j=1 \\ j\neq i}}^{n-1} |b_{ij}| + |v_i| \underbrace{\sum_{j=1}^{n-1} \frac{|w_i|}{|\alpha|}}_{\leq 1} \leq |v_i| + \sum_{\substack{j=1 \\ j\neq i}}^{n-1} |b_{ij}| \leq |b_{ii}|$$

$$\Longrightarrow \sum_{\substack{j=1 \\ j\neq i}}^{n-1} \left| b_{ij} - \frac{1}{\alpha} v_i w_j \right| \leq |b_{ii}| - \left| \frac{v_i w_i}{\alpha} \right| \leq \left| b_{ii} - \frac{v_i w_i}{\alpha} \right| = \left| \left(B - \frac{1}{\alpha} v w^T \right)_{ii} \right|$$

$|x| = |x - y + y| \leq |x-y| + |y|$
$\iff |x| - |y| \leq |x-y|$ ∎

Somit ist $B - \frac{1}{\alpha} v w^T$ wieder diagonaldominant.

Rangbestimmung

Sei $A \in \mathbb{K}^{n \times n}$. Gilt nach k-Schritten der Gauß-Elimination

$$A^{(k)} = \begin{array}{c} \overbrace{}^{k} \quad \overbrace{}^{n-k} \\ \left[\begin{array}{c|c} U^{(k)} & * \\ \hline 0 & 0 \end{array}\right] \begin{array}{l} \} k \\ \} n-k \end{array} \end{array}$$

mit $u_{ii}^{(k)} \neq 0$ so ist $\text{Rang}(A) = k$. Somit

- Kann die GE zu Ende geführt wurde so gilt $k = n$, mithin $\text{Rang}(A) = n$.
- Rangbestimmung erfordert totales Pivoting, da die erste Spalte der $n-k \times n-k$ Restmatrix 0 sein kann, was aber nicht $\text{Rang}(A) = k$ bedeutet.

Inversenberechnung

Zur Berechnung der Inversen geht man so vor:

- Berechne LR-Zerlegg von A. Aufwand $\frac{2}{3}n^3 + O(n^2)$ bei Spaltenpivotisierung
- Für $i = 1, \ldots, n$ löse $Ax^{(i)} = e^{(i)}$. Aufwand $n \cdot 2n^2$ für Lösen der Dreieckssysteme
- $A^{-1} = [x^{(1)}, \ldots, x^{(n)}]$ besteht spaltenweise aus den $x^{(i)}$
- Gesamtaufwand ist $\frac{8}{3}n^3 + O(n^2)$.

Tridiagonalsysteme

Definition 5.17

$A \in \mathbb{R}^{n \times n}$ heißt Tridiagonalmatrix falls

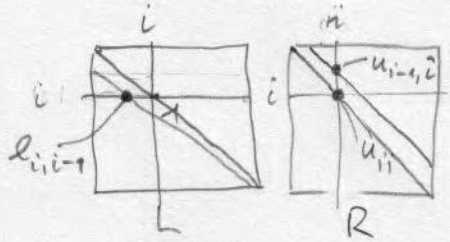
$$a_{ij} = 0 \quad \text{für } |i-j| > 1. \qquad 1 \le i \le n$$

Eine Tridiagonalmatrix ist Spezialfall einer Bandmatrix:

$$a_{ij} = 0 \quad \text{für } j < i - m_\ell \text{ oder } j > i + m_r \qquad 1 \le i \le n$$

Die LR-Zerlegung einer Tridiagonalmatrix sei ohne Pivotisierung durchführbar. Dann gilt:

- L ist Tridiagonalmatrix
- R ist Tridiagonalmatrix

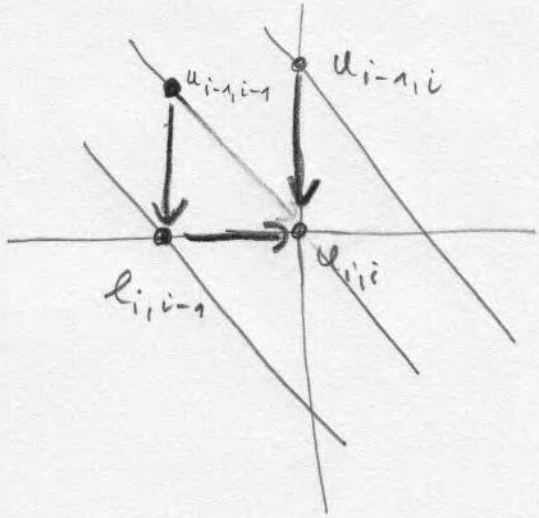


und damit:

Zeile i · Spalte i: $\ell_{i,i-1} \cdot u_{i-1,i} + u_{ii} = a_{ii} \Rightarrow u_{ii} = a_{ii} - \ell_{i,i-1} u_{i-1,i}$

$$\ell_{i,i-1} \cdot u_{i-1,i-1} = a_{i,i-1} \Rightarrow \ell_{i,i-1} = \frac{a_{i,i-1}}{u_{i-1,i-1}}$$

$$1_i \, u_{i-1,i} = a_{i-1,i} \Rightarrow u_{i-1,i} = a_{i-1,i}$$



Nichtreguläre Systeme

Es sei nun $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ wobei $\text{Rang}(A)$ beliebig.
$Ax = b$ hat dann genau eine, unendlich viele oder gar keine Lösung.

Einige Grundbegriffe aus der linearen Algebra:

$$\text{Bild}(A) = \{ y \in \mathbb{R}^m : y = Ax \text{ für ein } x \in \mathbb{R}^n \}$$

$$\text{Kern}(A) = \{ x \in \mathbb{R}^n : Ax = 0 \}$$

$$\text{Rang}(A) = \dim(\text{Bild}(A)) = \text{Rang}(A^T) = \dim(\text{Bild}(A^T))$$

(# l.u. Spalten = # l.u. Zeilen).

Orthogonales Komplement:

$$\text{Bild}(A)^\perp = \{ y \in \mathbb{R}^m : (y, y')_2 = 0 \ \forall y' \in \text{Bild}(A) \}$$

$$\begin{aligned} & (y, y')_2 = 0 \ \forall y' \in \text{Bild}(A) \\ \Leftrightarrow \ & (y, Ax)_2 = 0 \ \forall x \in \mathbb{R}^n \\ \Leftrightarrow \ & (A^T y)^T x = 0 \ \forall x \in \mathbb{R}^n \\ \Leftrightarrow \ & y \in \text{Kern}(A^T) \end{aligned}$$

und damit $\text{Bild}(A)^\perp = \text{Kern}(A^T)$.

Dies zeigt dann

$$\underbrace{\dim(\text{Bild}(A))}_{= \text{Rang}(A)} + \dim(\text{Kern}(A^T)) = m$$

$$\underbrace{\dim(\text{Bild}(A^T))}_{} + \dim(\text{Kern}(A)) = n$$

Den Lösungsbegriff für lineare Gleichungssysteme kann man auf die folgende Weise erweitern.

Satz 5.17 (Least Squares Lösung)

a) Es existiert $\bar{x} \in \mathbb{R}^n$ so dass

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

b) Diese Bedingung ist äquivalent dazu dass $\bar{x}$ Lösung von

$$A^T A\bar{x} = A^T b,$$

der sog. „Normalengleichung".

c) Falls $\text{Rang}(A) = n$ (dann ist zwingend $m \geq n$) ist $\bar{x}$ eindeutig bestimmt, sonst hat jede weitere Lösung die Form $\bar{x} + y$ mit $y \in \text{Kern}(A)$.

Beweis.

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

$$\|Ax - b\|_2^2 = \|A(x - \bar{x} + \bar{x}) - b\|_2^2 = \big(\underbrace{A\bar{x} - b}_{} + \underbrace{A(x-\bar{x})}_{}, \underbrace{A\bar{x} - b}_{} + \underbrace{A(x-\bar{x})}_{}\big)_2$$

$$= (A\bar{x} - b, A\bar{x} - b)_2 + 2\big(\underbrace{A\bar{x} - b}_{\in \text{Kern}(A^T) = \text{Bild}(A)^\perp}, \underbrace{A(x-\bar{x})}_{\in \text{Bild}(A)}\big)_2 + (A(x-\bar{x}), A(x-\bar{x}))_2$$

da $A^T(A\bar{x} - b) = 0$

$$= \|A\bar{x} - b\|_2^2 + \underbrace{\|A(x-\bar{x})\|_2^2}_{\geq 0}$$

$$\geq \|A\bar{x} - b\|_2^2$$

Damit erfüllt $\bar{x}$ die Minimalitätsbedingung.

(a) $\Rightarrow$ (b) Setze $F : \mathbb{R}^n \to \mathbb{R}$, $F(x) = \|Ax - b\|_2^2$. Notwendige Bedingung für ein Minimum ist $\nabla F(\bar{x}) = 0 \Leftrightarrow \frac{\partial F}{\partial x_k}(\bar{x}) = 0 \quad \forall k = 1, \dots, n.$

$$\frac{\partial F}{\partial x_k}(\bar{x}) = \frac{\partial}{\partial x_k}(Ax - b, Ax - b)_2\Big|_{x=\bar{x}} = \frac{\partial}{\partial x_k}\left(\sum_{i=1}^m \left[\left(\sum_{j=1}^n a_{ij} x_j\right) - b_i\right]^2\right)\Bigg|_{x=\bar{x}}$$

$$= \left(\sum_{i=1}^m 2\left[\sum_{j=1}^n a_{ij} x_j - b_i\right] a_{ik}\right)\Bigg|_{x=\bar{x}} = 2\sum_{i=1}^m (A^T)_{ki}\underbrace{\left[\sum_{j=1}^n a_{ij}\bar{x}_j - b_i\right]}_{(A\bar{x} - b)_i}$$

↑ Nachdifferenzieren.

$$= 2\big(A^T(A\bar{x} - b)\big)_k$$

Und damit $\nabla F(\bar{x}) = 2A^T(A\bar{x} - b) \overset{!}{=} 0 \Leftrightarrow A^T A\bar{x} = A^T b,$

die Normalengleichung.

(c) Lösbarkeit der Normalengleichung.

$\mathbb{R}^m = \text{Bild}(A) \oplus \text{Bild}(A)^\perp$,

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

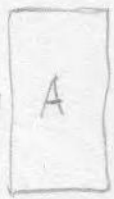
$b = s + r$ mit $s \in \text{Bild}(A)$, $r \in \text{Bild}(A)^\perp = \text{Kern}(A^T)$

Da $s \in \text{Bild}(A)$ gibt es $\bar{x} \in \mathbb{R}^n$ mit $A\bar{x} = s$. Für dieses gilt dann

$$A^T A\bar{x} = A^T s = A^T s + \underbrace{A^T r}_{\substack{=0 \text{ da} \\ r \in \text{Kern}(A^T)}} = A^T b,$$

also löst dieses $\bar{x}$ auch die Normalengleichung.

Sei $\text{Rang}(A) = n$ und damit $m \geq n$ (wg. $\text{Rang}(A) \leq \min(m,n)$).



Betrachte $A^T A x = 0$ also $\text{Kern}(A^T A)$.

$A^T A x = 0 \iff A^T y = 0 \wedge y = Ax$

Nun ist aber $\text{Kern}(A^T) \perp \text{Bild}(A)$ d.h. $\text{Kern}(A^T) \cap \text{Bild}(A) = \{0\}$.

Somit ist $A^T A$ regulär.

(Alternative: $A^T A$ ist symmetrisch und positiv definit wg. max. Rang).

Sei $\text{Rang}(A) < n$. Sei x_1 eine weitere Lösung der Normalengleichung (also $A^T A x_1 = A^T b$). Dann gilt

$$b = \underbrace{Ax_1}_{\in \text{Bild}(A)} + \underbrace{(b - Ax_1)}_{\in \text{Kern}(A^T)}$$

x_1 Lsg der NG.

da $A^T(b - Ax_1) = A^T b - A^T A x_1 \overset{\downarrow}{=} 0$

Da die Zerlegung $\mathbb{R}^m = \text{Bild}(A) \oplus \text{Bild}(A)^\perp$ eindeutig ist muss

$Ax_1 = A\bar{x}$ sein

und damit

$A(\bar{x} - x_1) = 0$, also $\bar{x} - x_1 \in \text{Kern}(A)$. ▮

Bemerkung: $A^T A$ ist symmetrisch und positiv semidefinit.

Lösung prinzipiell mit Cholesky-Zerlegung möglich.

Anwendung: Gaußsche Ausgleichsrechnung.

Gegeben:

(i) n Funktionen $u_1, \dots, u_n : \mathbb{R} \to \mathbb{R}$, sowie

(ii) m Datenpunkte $(x_i, y_i) \in \mathbb{R}^2$, $1 \le i \le m \geqq n$

↖ Bei $=$ ungl. Eindeutig.

Gesucht: n Koeffizienten $c_1, \dots, c_n$ so dass

$$u(x) = \sum_{j=1}^{n} c_j u_j(x)$$

und

(*) $$\sum_{i=1}^{m} (u(x_i) - y_i)^2 \longrightarrow \text{minimal}.$$

Dies läßt sich folgendermaßen formulieren:

$$c = (c_1, \dots, c_n)^T$$

$$y = (y_1, \dots, y_m)^T$$

$$a_{ij} = u_j(x_i)$$

Finde $c \in \mathbb{R}^n$ so dass $\|Ac - y\|_2^2$ minimal.

Denn $$\sum_{i=1}^{m} (u(x_i) - y_i)^2 = \sum_{i=1}^{m} \Big(\underbrace{\sum_{j=1}^{n} c_j u_j(x_i) - y_i}_{(Ac-y)_i} \Big)^2$$

$$= \|Ac - y\|_2^2 .$$

Somit sind nach Satz 5.17 die gesuchten c_i Lösung der Normalengleichung

$$A^T A c = A^T y.$$

Lösung: z.B. mit Cholesky-Zerl.

Ü: Zeige $\text{cond}_2(A^T A) = \text{cond}_2(A)^2$.

6 Interpolation & Approximation

6.1 Einführung

Worum es geht: Darstellung und Auswertung von Funktionen im Rechner.

Anwendungen:

- Rekonstruktion eines funktionalen Zusammenhangs aus „gemessenen" Funktionswerten; Auswertung an Zwischenstellen.
- Teuer auszuwertende Funktionen effizienter auswerten
- Darstellung von Fonts (2D), Körpern (3D) im Rechner. Voraussetzung für Simulation; Szenen in der Computergrafik.
- Lösung von Differential- und Integralgleichungen
- Datenkompression

⇒ Beispiel auf den Folien.

Wir beschränken uns hier weitgehend auf Funktionen in einer Variablen, also etwa

$$f \in C[a,b].$$

$C[a,b]$ ist ein unendlichdimensionaler (Vektor-)raum von Funktionen. Im Rechner betrachten wir Funktionenklassen die durch eine endliche Zahl von Parametern bestimmt sind (müssen keine Teilräume sein!). Z.B.

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

b) $r(x) = \dfrac{a_0 + a_1 x + \ldots + a_n x^n}{b_0 + b_1 x + \ldots + a_m x^m}$ (rationale Funktionen)

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

(trigonometrische Polynome)

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

Grundaufgabe der Approximation:

Gegeben eine Klasse von Funktionen P (siehe oben) sowie eine Funktion f (z.B. in $C[a,b]$).

Finde $g \in P$ so dass der Fehler $f-g$ in geeigneter Weise minimiert wird.

Beispiele:

a) $\left(\int_a^b (f-g)^2 dx \right)^{1/2} \to \min$

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

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

Man spricht von Interpolation falls g durch

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

festgelegt wird.

Dies ist ein Spezialfall der Approximationsaufgabe.

2.2 Polynominterpolation

1.12.09

P_n sei die Menge der Polynome über $\mathbb{R}$ vom Grad kleiner gleich $n \in \mathbb{N}_0$:

$$P_n := \left\{ p(x) = \sum_{i=0}^{n} a_i x^i \;\middle|\; a_i \in \mathbb{R} \right\}$$

P_n ist ein $n+1$-dimensionaler Vektorraum.

Die Monome $1, x, x^2, \dots, x^n$ bilden eine Basis.

Zu gegebenen $n+1$ (paarweise verschiedenen) Stützstellen ist die Interpolationsaufgabe

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

äquivalent zu dem linearen Gleichungssystem

$$\underbrace{\begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ & & & & \\ 1 & x_n & x_n^2 & & x_n^n \end{bmatrix}}_{V[x_0, \dots, x_n]} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$$

- $V[x_0, \dots, x_i]$ heißt Vandermonde-Matrix
- $x_i = x_j \Rightarrow$ Zeile i = Zeile $j \Rightarrow V$ nicht regulär
- aber ist V für paarweise verschiedene x_i regulär? Ja! Siehe unten. (Determinante läßt sich auch direkt ausrechnen).
- V ist schlecht konditioniert für große n $\operatorname{cond}_\infty(V) \approx 2^n$!

Wie macht man es besser und einfacher?

Lagrange-Interpolation

Definition 6.1

Zu den paarweise verschiedenen Stützstellen x_i, $i=0,\dots,n$, definiere die sog. Lagrangepolynome:

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

Diese Polynome haben die Eigenschaften:

a) $L_i^{(n)}$ hat Grad n.

Klarung: $\prod_{\substack{j=0 \\ j\neq i}}^{n} (x-x_j)$ hat n Faktoren.

b) Es gilt

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

$i\neq k$: Im Produkt $\prod_{\substack{j=0 \\ j\neq i}}^{n} (x-x_j)$ kommt $j=k\neq i$ vor und damit $(x_k-x_k)=0$ als ein Faktor.

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

c) Die $L_i^{(n)}$ bilden eine Basis von P_n

$L_k^{(n)}$ ist das einzige Polynom unter den $L_i^{(n)}$ mit $L_i(x_k)=1$, daher ist es unabhängig von den $L_j^{(n)}$, $j\neq k$.

Da P_n $n+1$ dimensional bilden die $L_i^{(n)}$, $i=0,\dots,n$ eine Basis.

Lösung der Interpolationsaufgabe:

Für gegebene (x_i, y_i), $i=0,\dots,n$, erfüllt

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

die Interpolationsaufgabe: $p(x_k) = \sum_{i=0}^{n} y_i L_i^{(n)}(x_k) = \sum_{i=0}^{n} y_i \delta_{ik} = y_k$.

Satz 6.2 (Eindeutigkeit der Polynominterpolation)

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

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

Beweis:

kurz: $p = \sum_{i=0}^{n} y_i L_i^{(n)}$ interpoliert die gegebenen Werte. Da die $L_i^{(n)}$ eine Basis von P_n bilden ist diese Darstellung eindeutig.

lang: $p = \sum_{i=0}^{n} y_i L_i^{(n)}$ zeigt konstruktiv die Existenz des Interpolationspolynoms.

Eindeutigkeit: Ang. es gibt zwei Polynome p_1, p_2, $p_1 \neq p_2$, aber

$p_1(x_i) = p_2(x_i) = y_i \quad i = 0, \dots, n.$

$p := p_1 - p_2 \in P_n$ (Differenzpolynom) erfüllt $p(x_i) = 0$, $i = 0, \dots, n$, hat also $n+1$ Nullstellen, damit muss $p \equiv 0$ gelten. Dies folgt aus

Satz von Rolle: $u(x)$ sei auf $[a, b]$ stetig und in (a, b) differenzierbar sowie $u(a) = u(b) = 0$. Dann existiert mindestens ein $x \in (a, b)$ mit $u'(x) = 0$ ($u(x) = 0$ ist auch möglich).

Angenommen es sei p nicht das Nullpolynom, also

$p = \alpha_m x^m$ mit $0 \leq m \leq n$, $\alpha_m \neq 0$. Dann gilt

$p = p^{(0)} = \alpha_m x^m + \dots$ hat $n+1$ Nullstellen

$p' = p^{(1)} = \alpha_m m x^{m-1} + \dots$ hat n Nullstellen (Rolle!)

$p'' = p^{(2)} = \alpha_m m (m-1) x^{m-2} + \dots$ hat $n-1$ Nullstellen

$\vdots$

$p^{(m)} = \alpha_m m!$ hat $n+1-m \geq 1$ Nullstellen (da $m \leq n$)

Dies geht nur für $\alpha_m = 0$ im Widerspruch zur Annahme. ∎

Newton-Darstellung

Nachteil der Lagrange-Polynome:

- Hinzufügen einer Stützstelle ändert alle bisherigen Basispolynome
- Eignet sich nicht zur „inkrementellen" Erstellung des Interpolationspolynoms.

Besser ist hier die Newton-Darstellung mit den Basis-Polynomen:

$$N_0(x) = 1; \quad i = 1, \dots, n: \quad N_i(x) = \prod_{j=0}^{i-1} (x - x_j)$$

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

b) $N_0, \dots, N_n$ bilden eine Basis von P_n.

c) $N_i(x_k) = 0$ für alle $i > k$, dann $(x - x_k)$ kommt in N_i für $i > k$ vor. ($i-1 \geq k \Leftrightarrow i > k$)

Gestaffelte Berechnung: Bestimme $p(x) = \sum_{i=0}^{n} a_i N_i(x)$ über

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

also $a_0 = y_0$;

$$a_k = \left(y_k - \sum_{i=0}^{k-1} a_i N_i(x_k) \right) \Big/ N_k(x_k);$$

Alternativ lassen sich die Koeffizienten auch berechnen über

Satz 6.3 (Dividierte Differenzen)

Man definiert rekursiv die sog. „Dividierten Differenzen"

$$\forall i = 0, \dots, n \qquad y[x_i] := y_i \quad \text{(die } n+1 \text{ Werte an Stützstellen)},$$

$$\forall k = 1, \dots, n-i \qquad y[\underbrace{x_i, \dots, x_{i+k}}_{k+1 \text{ Stück}}] := \frac{y[x_{i+1}, \dots, x_{i+k}] - y[x_i, \dots, x_{i+k-1}]}{x_{i+k} - x_i}.$$

Dann gilt

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

Beweis: [Ramadan, Satz 2.2]. ∎

Die dividierten Differenzen ordnet man in folgendem Tableau an:

$$
\begin{array}{lccccc}
 & a_0 & a_1 & a_2 & a_3 & \\
 & \| & \| & \| & \| & \\
y_0 = y[x_0] & \to & y[x_0,x_1] & \to & y[x_0,x_1,x_2] & \to \; y[x_0,x_1,x_2,x_3] \\
 & \nearrow & \nearrow & \nearrow & & \\
y_1 = y[x_1] & \to & y[x_1,x_2] & \to & y[x_1,x_2,x_3] & \\
 & \nearrow & \nearrow & & & \\
y_2 = y[x_2] & \to & y[x_2,x_3] & & & \\
 & \nearrow & & & & \\
y_3 = y[x_3] & & & & &
\end{array}
$$

↑ Hinzufügen von y_4 erfordert Berechnung aller Koeffizienten in der Diagonale.

Diese Berechnung der Koeffizienten der Newton-Darstellung ist numerisch stabiler.

Effiziente und numerisch stabile Auswertung des Interpolationspolynoms an <u>einzelnen Stellen</u> (also $p(x)$) gelingt mit der <u>Methode von Neville.</u> Siehe [Ramacher, S. 27].

Interpolationsfehler

8
2.12.09

$y_i = f(x_i)$, $i = 0, \dots, n$ sei die Auswertung einer Funktion f an $n+1$ paarweise verschiedenen Stützstellen

$p(x)$ sei das Polynom vom Grad n welches (x_i, y_i), $i = 0, \dots, n$ interpoliert.

Die Differenz erfüllt

$$e(x) = f(x) - p(x) = \begin{cases} 0 & x = x_i \\ ? & x \neq x_i \quad i = 0, \dots, n. \end{cases}$$

Frage: Wie groß kann die Differenz werden?

Satz 6.4 Sei $f(x)$ $n+1$ mal ~~stetig~~ differenzierbar auf $[a,b]$ und (lässt auch Extrapolation zu)
es sei $a \leq x_0 < x_1 < \dots < x_n \leq b$. Dann gibt es zu jedem $x \in [a,b]$ ein $\xi_x \in \overline{(x_0, \dots, x_n, x)}$ (= kleinstes Intervall welches alle Punkte enthält) sodass:

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

Beweis. Für $x \in \{x_0, \dots, x_n\}$ liefert $\prod_{j=0}^{n} (x - x_j) = 0$ somit ist ξ_x beliebig wählbar.
Sei also $x \in [a,b] \setminus \{x_0, \dots, x_n\}$. Zu diesem x definiere die Funktion

$$F_x(t) = f(t) - p(t) - \underbrace{\frac{f(x) - p(x)}{\ell(x)}}_{\text{eine Zahl abhängig von } x} \ell(t) \quad \text{mit} \quad \ell(t) = \prod_{j=0}^{n} (t - x_j)$$

(t: neue freie Variable)

$F_x(t)$ hat die $n+2$ Nullstellen $\{x_0, \dots, x_n, x\}$, denn

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

$$x: \quad F_x(x) = f(x) - p(x) - \frac{f(x) - p(x)}{\ell(x)} \ell(x) = 0$$

Satz von Rolle (siehe oben):
$F_x(t)$ $n+2$ Nullstellen (mindestens)
nach t ableiten! $F_x^{(1)}(t)$ $n+1$ Nullstellen $-\!\!\shortparallel\!\!-$
$F_x^{(n+1)}(t)$ 1 Nullstelle $(-\!\!\shortparallel\!\!-)$

Zusätzlich gilt: Diese Nullstellen sind in $\overline{(x_0, \dots, x_n, x)}$

Diese Nullstelle von $F_x^{(n+1)}$ sei $\xi_x \in \overline{(x_0,\dots,x_n,x)}$ und es gilt

$$F_x^{(n+1)}(\xi_x) = f^{(n+1)}(\xi_x) - \underbrace{p^{(n+1)}(\xi_x)}_{=0 \text{ da } p \text{ Grad } n} - \frac{f(x)-p(x)}{\ell(x)} \underbrace{\ell^{(n+1)}(\xi_x)}_{\substack{\ell(t)=t^{n+1}+\dots \\ n+1 \text{ mal differenziert:} \\ \ell^{(n+1)}(t)=(n+1)!}}$$

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

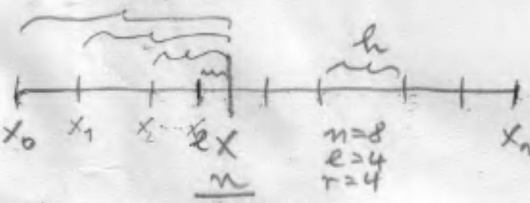
Auflösen nach $f(x)-p(x)$ liefert die Behauptung. ∎

Diskussion des Interpolationsfehlers

Betrag bilden:

$$|f(x)-p(x)| \le \frac{|f^{(n+1)}(\xi_x)|}{n+1!} \prod_{j=0}^{n} |x-x_j| \quad \text{für ein } \xi_x.$$

- Die Stützstellen seien äquidistant gewählt: $|x_i - x_{i+1}| = h$ und $x_0 \le x \le x_n$:



ℓ : # Punkte links von x

r : # Punkte rechts von x

$\Rightarrow \ell + r = n+1$ (Gesamtzahl der Fälle)

und damit $\prod_{j=0}^{n} |x-x_j| \le \underbrace{1h \cdot 2h \cdot \dots \cdot \ell h}_{\text{links von } x} \cdot \underbrace{1h \cdot 2h \cdot \dots \cdot r h}_{\text{rechts von } x}$

$$= \ell!\, r!\, h^{n+1}$$

also $$|f(x)-p(x)| \le |f^{(n+1)}(\xi_x)| \underbrace{\frac{\ell!\, r!}{n+1!}}_{\le 1 \text{ da } \ell+r=n+1} h^{n+1}$$

Für $|f^{(n+1)}|$ beschränkt und $n \to \infty$ geht $|f(x)-p(x)| \to 0$.

Allerdings sind die höheren Ableitungen auch einfacher Funktionen für $n \to \infty$ oft <u>nicht</u> beschränkt sondern wachsen sehr schnell.

TO
2.12.09

- Ramacher: (Runges Gegenbeispiel)

$$f(x) = \frac{1}{1+x^2} \qquad |f^{(n)}(x)| \approx 2^n n! \, O\left(\frac{1}{|x|^{n+2}}\right)$$

$\Rightarrow$ Siehe Folien, insbesondere am Rand des Intervalls.

- Beispiel 4.6. aus dem Numstoch-Skript als Ü-Aufgabe!

Bemerkung 6.5

Approximationssatz von Weierstraß: Jede Funktion $f \in C[a,b]$ kann beliebig gut gleichmäßig auf $[a,b]$ durch Polynome approximiert werden.

Obige Beobachtung ist kein Widerspruch, denn

-- Approximation muss nicht durch Interpolation erfolgen (Der Beweis nutzt sog. Bernstein-Polynome)

-- Mit nichtäquidistanten Stützstellen wird es auch schon sehr viel besser.

Bemerkung 6.6

Allgemein gilt für „Methoden hoher (Polynom-) Ordnung", dass entsprechende Differenzierbarkeit von f vorliegen muss.

Konditionierung

$p(x; y)$ bezeichne das Interpolationspolynom zu den Ordinatenwerten $y = (y_0, \dots, y_n)^T$ und Stützabszissenwerte $(x_0, \dots, x_n)^T$.

$$\frac{p(x; y+\Delta y) - p(x; y)}{p(x; y)} = \left(\sum_{j=0}^{n} (y_i + \Delta y_i) L_i^{(n)}(x) - \sum_{j=0}^{n} y_i L_i^{(n)}(x) \right) \Big/ p(x; y)$$

$$= \sum_{j=0}^{n} \underbrace{\frac{L_i^{(n)}(x) y_i}{p(x; y)}}_{\text{Verstärkungsfaktor}} \cdot \underbrace{\frac{\Delta y_i}{y_i}}_{\text{relativer Eingabefehler}}$$

Für große n kann $L_i^{(n)}(x)$ sehr groß werden, dann ist die Interpolationsaufgabe schlecht konditioniert!

6.3 Anwendungen der Polynominterpolation

Numerische Differentiation

Problem:
- Berechne die Ableitung (der Ordnung k) einer tabellarisch gegebenen Funktion oder einer einer im Rechner als Prozedur gegebenen Funktion.

Idee: Erstelle Interpolationspolynom zu bestimmten Stützstellen, leite dieses ab und werte es aus.

Zunächst = Ableitungsordnung = Polynomgrad.

Lagrange-Polynome sind

$$L_i^{(n)}(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{(x-x_j)}{(x_i-x_j)} = \underbrace{\left(\prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{1}{(x_i-x_j)}\right)}_{=:\lambda_i \in \mathbb{R}} x^n + \alpha_{n-1} x^{n-1} + \ldots + \alpha_0$$

n-maliges differenzieren liefert:

$$\frac{d^n}{dx^n} L_i^{(n)}(x) = \lambda_i \, n! \qquad \text{(konstant, unabhängig von } x\text{)}$$

Damit gilt für die n-te Ableitung eines Interpolationspolynoms vom Grad n:

$$\frac{d^n}{dx^n}\left(\sum_{j=0}^{n} y_i L_i^{(n)}\right)(x) = n! \sum_{j=0}^{n} y_i \lambda_i . \qquad \text{(unabh. von } x\text{)}$$

Eine Aussage über den Fehler liefert:

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

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

Beweisskizze: n-malige Anwendung des Satzes von Rolle auf $g(x) = f(x) - p(x)$.

$g(x) = f(x) - p(x)$ n+1 Nullst.
$g'(x)$ n
$g^{(n)}(x)$ 1 Nullst. → ξ

Um einfachere Formeln zu bekommen verwenden wir nun äquidistante Stützstellen, d.h. $x_i = x_0 + ih$, $0 \le i \le n$.

Damit ist

$$\lambda_i = \frac{1}{\underbrace{(x_i - x_0)}_{ih} \cdots \underbrace{(x_i - x_{i-1})}_{h} \underbrace{(x_i - x_{i+1})}_{-h} \cdots \underbrace{(x_i - x_n)}_{x_0+ih-x_0-nh=-(n-i)h}}$$

(i Stück positiv; $n-i$ Stück negativ)

$$= \frac{1}{h^n (-1)^{n-i} i!(n-i)!} = \frac{(-1)^{n-i}}{h^n n!} \binom{n}{i}$$

← Binomialkoeffizient

und damit

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

$n!$ kürzt sich heraus!

Speziell:

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

Bisher: n-te Ableitung aus Polynom vom Grad n (d.h. $n+1$ Werten)

Es geht auch: m-te Ableitung aus Polynom vom Grad $n > m$.
Wert hängt dann aber von der Auswertestelle ab.

Beispiel: $m=1$, $n=2$. Äquidistante Stützstellen $x_0, x_1 = x_0 + h, x_2 = x_0 + 2h$.

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

Taylorreihenentwicklung (Übung!) zeigt

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} + O(h^2) \quad \text{für } f \in C^3,$$

$$f''(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} + O(h^2) \quad \text{für } f \in C^4.$$

Beispiel 6.8 → Beispiel 4.9 aus Numstoch zur Auslöschung bei numerischer Differentiation

→ Übung!

und zur Motivation der Extrapolation!

Extrapolation zum Limes

13
3.12.09

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

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

mit guter Genauigkeit berechnen.

Beispiel 6.9

a) $$a(0) = \lim_{x\to +0} \frac{\cos(x)-1}{\sin(x)} \qquad (=0)$$

b) Numerische Differentiation

$$f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}$$ (für kleine h tritt Auslösung ein)

c) Numerische Integration

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

wobei $h := \frac{1}{N}$.

$h\to 0$ nicht möglich wg. Aufwand $\to\infty$

d) Numerische Lösung des Anfangswerts

$$y'(t) = f(t, y(t)); \quad y(0) = y_0$$

$y(T) \approx y_N$ und $y_n = y_{n-1} + h f(t, y_{n-1})$; $h = 1/N$;

$h\to 0$ bedeutet $N\to\infty$ und damit Aufwand $\to\infty$. ∎

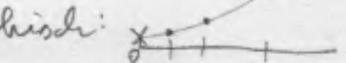
Idee der Extrapolation:

Zu $h_0 > h_1 > \cdots > h_n > 0$ bestimme Interpolationspolynom

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

und berechne

$$a(0) \approx p(0)$$ (Extrapolation, da $0 \notin [h_n, \ldots, h_0]$.

graphisch: 

Beispiel 6.10.

[Rannacher]. Für $a(h) = \dfrac{\cos(h)-1}{\sin(h)}$ erhält man für

$$h_0 = 1/8 \qquad a(h_0) = -6.258\,151 \cdot 10^{-2}$$
$$h_1 = 1/16 \qquad a(h_1) = -3.126\,018 \cdot 10^{-2}$$
$$h_2 = 1/32 \qquad a(h_2) = -1.562\,627 \cdot 10^{-2}$$

} halbiert sich, da h halbiert wird.

und bei Extrapolation mit einem Polynom vom Grad 2:

$$a(0) \approx p_2(0) = -1.02 \cdot 10^{-5}$$

also sehr viel besser! ▨

Übung: Extrapolation bei der numerischen Differentiation!

Warum funktioniert das so gut?

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

$$\underbrace{a(h)}_{\text{Für } h>0 \text{ numerisch berechenbar.}} = a(0+h) = a(0) + h\,a'(0) + \dots + \frac{h^n}{n!}a^{(n)}(0) + \frac{h^{n+1}}{(n+1)!}a^{(n+1)}(\xi_h)$$

gleich dahinschreiben!

$$= \underbrace{a(0)}_{=:a_0} + \underbrace{a'(0)}_{=:a_1} h + \dots + \underbrace{\frac{a^{(n)}(0)}{n!}}_{=:a_n} h^n + \underbrace{\frac{a^{(n+1)}(\xi_h)}{(n+1)!}}_{\text{abhängig von } h!} h^{n+1} \qquad (6.1)$$

Polynom in h vom Grad n

$a^{(k)}(0)$ hängt nicht von h ab!

Idee: Für verschiedene h_i bilde Linearkombination:

$$\sum_{i=0}^{n} c_i\, a(h_i) = \sum_{i=0}^{n} c_i \left(\sum_{j=0}^{m} a_j h_i^j \right) + \sum_{i=0}^{n} \underbrace{c_i \frac{a^{(n+1)}(\xi_{h_i})}{(n+1)!}}_{=:\alpha_i} h_i^{n+1}$$

$$= \sum_{j=0}^{m} a_j \left(\sum_{i=0}^{n} c_i h_i^j \right) + \text{Fehler} = a_0 + \text{Fehler}$$

a_j: unbekannt; $\sum_{i=0}^{n} c_i h_i^j$: Bestimmungsgleichung für die c_i

Fehler $\to$ klein wg. h_i^{n+1}

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

Als Gleichungssystem lautendie Bestimmungsgleichungen

$$V^T c = e^{(0)}$$

mit $e^{(0)} = (1, 0, \dots, 0)^T$

und $V = V[h_0, \dots, h_n]$ der Vandermondematrix.

Die Auswertung mit diesen Koeffizienten ist dann:

$$A := \sum_{i=0}^{n} c_i \underbrace{a(h_i)}_{y_i} = \uparrow_{y=(y_0,\dots,y_n)} \quad \underbrace{\left(V^{-T} e^{(0)}\right)^T}_{=c} y = (e^{(0)})^T V^{-1} y.$$

$$V = \begin{bmatrix} 1 & h_0 & h_0^2 & \cdots & h_0^n \\ & & & & \\ 1 & h_n & & & h_n^n \end{bmatrix}$$

$h_i = h r^i$

$$\to = \begin{bmatrix} 1 & hr & (hr)^2 & & (hr)^n \\ \vdots & & & & \\ 1 & hr^n & (hr^n)^2 & & (hr^n)^n \end{bmatrix}$$

$$= \begin{bmatrix} 1 & r & r^2 & \cdots & r^n \\ & & & & \\ 1 & r^n & r^{2n} & & r^{n^2} \end{bmatrix} \begin{bmatrix} 1 & & & 0 \\ & h & & \\ & & \ddots & \\ 0 & & & h^n \end{bmatrix}$$

Für den Fehler gilt dann

$$|A - \underbrace{a(0)}_{\text{Best.ge. für } c!}| = \left| \sum_{i=0}^{n} c_i \frac{a^{(n+1)}(\xi_i)}{(n+1)!} h_i^{(n+1)} \right|$$

$h_i = h r^i \to$

$$\leq \sum_{i=0}^{n} \overbrace{\|V^{-T} e^{(0)}\|_\infty}^{\max_i c_i} \frac{|a^{(n+1)}(\xi_i)|}{(n+1)!} h^{n+1} r^{i(n+1)}$$

z.B. $r = 1/2$

$$\leq \|V^{-T}\|_\infty \underbrace{\|e^{(0)}\|_\infty}_{=1} |a^{(n+1)}(\xi_{i\max})| \frac{h^{n+1}}{(n+1)!} \underbrace{\sum_{i=0}^{n} r^{i(n+1)}}_{\text{geometrische Reihe}}$$

$$= r^{n+1} + 1$$

$$\leq \|V^{-T}\|_\infty \underbrace{|a^{(n+1)}(\xi_{\max})|}_{\text{beschränkt}} \frac{h^{n+1}}{(n+1)!} \underbrace{(1 + r^{n+1})}_{\leq \frac{5}{4} \text{ für } r \leq \frac{1}{2} \wedge n \geq 1}$$

Nun betrachte das Interpolationspolynom

$$p(h_i) = \sum_{j=0}^{n} b_j h_i^j \overset{!}{=} a(h_i) = y_i$$

und somit $Vb = y$ für dessen Koeffizienten.

Auswerten an der Stelle Null bedeutet

$$p(0) = b_0 = (e^{(0)})^T b = (e^{(0)})^T V^{-1} y = A \text{ von oben!}$$

Oben beschriebene Elimination der Fehlerterme entspricht also genau der Extrapolationsmethode!

Entscheidend ist also die Fehlerdarstellung (6.1).

Und es gelt noch besser! Wir betrachten die Näherung für $f''(x)$:
(n gerade:)

$$f(x+h) = f(x) + h f^{(1)}(x) + \frac{h^2}{2} f''(x) + \frac{h^3}{3!} f^{(3)}(x) + \ldots + \frac{h^{2n+2}}{(2n+2)!} f^{(2n+2)}(x) + \frac{h^{2n+4}}{(2n+4)!} f^{(2n+4)}(\xi_+)$$

$$f(x-h) = f(x) - h f^{(1)}(x) + \frac{h^2}{2} f''(x) - \frac{h^3}{3!} f^{(3)}(x) + \ldots + \frac{h^{2n+2}}{(2n+2)!} f^{(2n+2)}(x) + \frac{h^{2n+4}}{(2n+4)!} f^{(2n+4)}(\xi_-)$$

$$f(x+h) + f(x-h) = 2f(x) + h^2 f''(x) + \ldots + \frac{h^{2n+2}}{2\,(2n+2)!} f^{(2n+2)}(x) + \frac{h^{2n+4}}{(2n+4)!}\left[f^{(2n+4)}(\xi_+) + f^{(2n+4)}(\xi_-)\right]$$

$$a(h) := \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} =$$

$$= f''(x) + \frac{h^2}{2(4!)} f^{(4)}(x) + \ldots + \frac{h^{2n}}{2(2n+2)!} f^{(2n+2)}(x) + \frac{h^{2n+2}}{(2n+4)!}\left[f^{(2n+4)}(\xi_+) + f^{(2n+4)}(\xi_-)\right]$$

$$= p_x(h^2) + O(h^{2n+2})$$

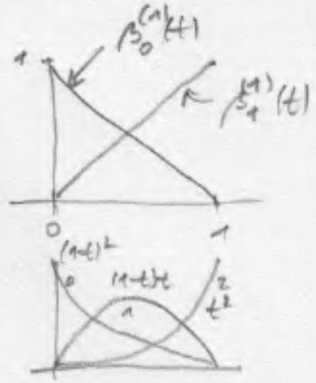
Da alle ungeraden Terme „von selbst" wegfallen kann man mit der gleichen Anzahl von Auswertungen doppelt soviele Fehlerterme eliminieren!

Natürlich muss f dazu entsprechend oft differenzierbar sein.

6.4 Bernsteinpolynome und Kurvendarstellung

7.12.09

Wir gehen nun über von der Interpolation zur Approximation. Speziell für Kurven, d.h. Funktionen $u(t): [a,b] \to \mathbb{R}^d$, $d = 2,3$, haben sich Bernstein-Polynome bewährt,



Definition 6.11 (Bernstein-Polynome)

Die Polynome

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

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

Mittels der Transformation $\varphi: [a,b] \to [0,1]$, $\varphi(u) = \frac{u-a}{b-a}$, definiert man die Bernstein-Polynome auf einem allgemeinen Intervall $[a,b]$:

$$B_{i,[a,b]}^{(n)} = B_i^{(n)}(\varphi(u)) = \binom{n}{i} \left(1 - \frac{u-a}{b-a}\right)^{n-i} \left(\frac{u-a}{b-a}\right)^i$$

$$= \binom{n}{i} \frac{1}{(b-a)^n} (b-u)^{n-i} (u-a)^i.$$

Satz 6.12 (Eigenschaften der Bernstein-Polynome)

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

Beweis: binomischer Lehrsatz: $1 = (\underbrace{1-t}_{} + \underbrace{t}_{})^n = \sum_{i=0}^{n} \underbrace{\binom{n}{i} (1-t)^{n-i} t^i}_{= B_i^{(n)}(t)}$

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

Beweis: Es sei $0 \le j \le i$, dann ist wg Produktregel $\frac{d^j}{dt^j} B_i^{(n)}(t) = \sum_{k=0}^{j} g^{(k)}(t) t^{i-k}$

und damit $\frac{d^j}{dt^j} B_i^{(n)}(0) = 0$ falls $i - j > 0 \Leftrightarrow 0 \le j < i$

d.h. $i=0$: keine Nullstelle

$i=1$: $0 \le j < 1$ eine Nullstelle

$i=2$: $0 \le j < 2$ Nullstelle, Nullstelle der Ableitung

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

$\Rightarrow t=0, t=1$ sind die einzigen Nullstellen

Beweis: analog zu (b).

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

Bew: Einsetzen

(e) Positivität - $0 \leq \beta_i^{(n)} \leq 1$ für $t \in [0,1]$

- $\beta_i^{(n)}(t) > 0$ für $t \in (0,1)$

Beweis: $t \in [0,1] \Rightarrow t \geq 0, 1-t \geq 0$ also $\beta_i^{(n)}(t) \geq 0$

$t \in (0,1) \Rightarrow t > 0, 1-t > 0$ also $\beta_i^{(n)}(t) > 0$

und $\sum_{i=0}^{n} \beta_i^{(n)}(t) = 1 \Leftrightarrow \beta_i^{(n)}(t) = 1 - \sum_{\substack{j=0 \\ j \neq i}}^{n} \underbrace{\beta_j^{(n)}(t)}_{\geq 0} \leq 1.$

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

$$\frac{d}{dt}\beta_i^{(n)}(t) = \binom{n}{i}\left[-(n-i)(1-t)^{n-i-1}t^i + i(1-t)^{n-i}t^{i-1}\right]$$

$$= \binom{n}{i}(1-t)^{n-i-1}t^{i-1}\underbrace{\left((1-t)i - (n-i)t\right)}_{i - it - nt + it}$$

$$= \binom{n}{i}\underbrace{(1-t)^{n-i-1}}_{\substack{n-i-1\text{-fache} \\ \text{Nst bei} \\ t=1}}\underbrace{t^{i-1}}_{\substack{i-1\text{-fache} \\ \text{Nullstelle} \\ \text{bei } t=0}}\underbrace{(i-nt)}_{\substack{\text{Nullstelle} \\ \text{bei } t=i/n}}$$

$n-i-1+i-1+1 = n-1$ Nullstellen. da $\frac{d}{dt}\beta_i^{(n)}$ Polynom vom Grad $n-1$ sind dies alle!

Aus (b) und (c) (und (e)) folgt dass bei i/n ein Maximum vorliegt

(g) Die $\{\beta_i^{(n)}\}_{i=0}^{n}$ sind linear unabhängig und bilden eine Basis von P_n.

Bew: zu zeigen $\sum_{i=0}^{n} b_i \beta_i^{(n)}(t) = 0 \ \forall t \in \mathbb{R} \Rightarrow b_i = 0.$

Betrachte Ableitung $\frac{d^j}{dt^j}\sum_{i=0}^{n} b_i \beta_i^{(n)} = \sum_{i=0}^{n} b_i \frac{d^j}{dt^j}\beta_i^{(n)}(t) = 0 \ \forall t \in \mathbb{R}$ (da $\sum b_i \beta_i$ Nullfunkt.)

Setze $j=0, t=0$: Es ist nur $\beta_0^{(n)}(0) \neq 0$, alle anderen haben dort Nullst. $\Rightarrow b_0 = 0$

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

usw.

(R) Die Bernstein-Polynome erlauben folgende rekursive Darstellung über den Grad n:

$$B_i^{(n)}(t) = \begin{cases} (1-t)\, B_0^{(n-1)}(t) & i=0 \\ t\, B_{i-1}^{(n-1)}(t) + (1-t) B_i^{(n-1)}(t) & 0<i<n \\ t\, B_{n-1}^{(n-1)}(t) & i=n \end{cases}$$

Bew. $i=0, i=n$ sieht man durch einsetzen.

$0<i<n$

$$t \binom{n-1}{i-1} (1-t)^{n-1-(i-1)} t^{i-1} + (1-t) \binom{n-1}{i} (1-t)^{n-1-i} t^i$$

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

Rekursionsformel für Binomialk.

(i) Für die erste Ableitung gilt die Rekursionsformel:

$$\frac{d}{dt} B_i^{(n)}(t) = \begin{cases} -n\, B_0^{(n-1)}(t) & i=0 \\ n \left[B_{i-1}^{(n-1)}(t) - B_i^{(n-1)}(t) \right] & 0<i<n \\ n\, B_{n-1}^{(n-1)}(t) & i=n \end{cases}$$

kein Beweis?

Achtung: Rechts stehen keine Ableitungen sondern Bernstein-Polynome vom Grad $n-1$!

Kurvendarstellung mittels Bernstein-Polynomen beschreibt:

Definition 6.13 (Bezier-Kurven)

Für gegebene Punkte $b_0, \dots, b_n \in \mathbb{R}^d$ heißt das vektorwertige Polynom

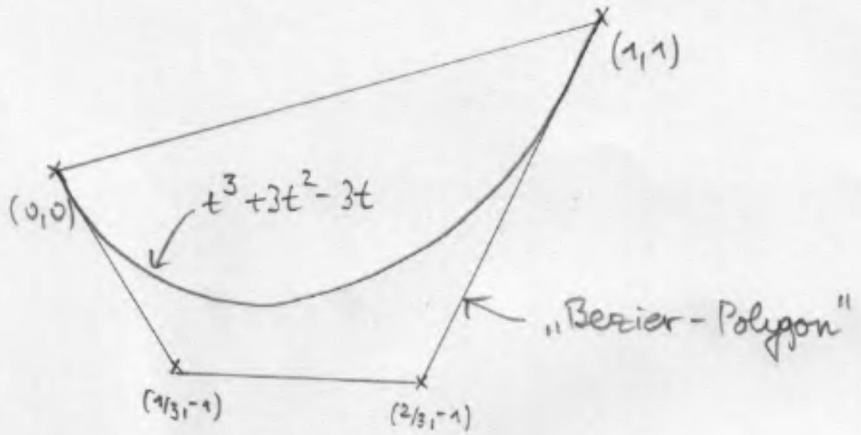
$$B(t) = \sum_{i=0}^{n} b_i B_i^{(n)}(t)$$

Bezier-Kurve.

Beispiel 6.14

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

Die Zugehörige Bezier-Kurve: $B(t) = \sum_{i=0}^{3} b_i \beta_i^{(3)}(t) = \begin{pmatrix} t \\ t^3 + 3t^2 - 3t \end{pmatrix}$



× : Bezier-Punkte $(i/n, b_i)$

Die Verbindung der Punkte $(i/n, b_i)$ nennt man „Bezier-Polygon".

Es gelten folgende Eigenschaften:

- Das Bezier-Polynom liegt in der konvexen Hülle der Bezier-Punkte.

  $B(t) = \sum_{i=0}^{n} b_i \beta_i^{(n)}(t)$ ist wegen $0 \le \beta_i^{(n)}(t) \le 1$ und $\sum_{i=0}^{n} \beta_i^{(n)}(t) = 1$

  eine Konvexkombination.

- Es ist $B(0) = b_0$ und $B(1) = b_1$. (Liegt an den Nullstellen 6.12 b,c)

- Die Ableitung (Tangente an die Kurve) hat den Endpunkten die Richtung $(b_1 - b_0)$ bzw. $(b_n - b_{n-1})$.

Nutze rekursive Darstellung aus Satz 6.12 (i):

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

$$= b_0 \left[-n \underbrace{\beta_0^{(n-1)}(t)}_{=1} \right] + b_1 \left[n \left(\underbrace{\beta_0^{(n-1)}(t)}_{=1} - \beta_1^{(n-1)}(t) \right) \right] \dots$$

$$\dots + b_{n-1} \left[n \left(\beta_{n-2}^{(n-1)}(t) - \underbrace{\beta_{n-1}^{(n-1)}(t)}_{=1} \right) \right] + b_n \left[n \underbrace{\beta_{n-1}^{(n-1)}(t)}_{=1} \right]$$

$$= \begin{cases} t=0 & \text{Fall-} \\ t=1 & \text{unterscheidung} \end{cases}$$

$t=0$: $\beta_0^{(n-1)}(0) = 1$: $-n b_0 + b_1 \cdot n = n(b_1 - b_0)$

$t=1$: $\beta_{n-1}^{(n-1)}(1) = 1$: $b_{n-1}(-n) + b_n n = n(b_n - b_{n-1})$.

Bernstein-Polynome erlauben eine gute Kontrolle der Norte des Polynoms durch dessen Koeffizienten.
Vergleichen wir dies mit anderen Polynombasen:

- Monome, Newton: Koeffizienten erlauben keine einfache Kontrolle der Norte des Polynoms
- Lagrange: Erlaubt exakte Kontrolle an den Stützpunkten, dazwischen aber (sehr) große Abweichungen.
- Bernstein: $\min\{b_0,\dots,b_n\} \le \sum_{i=0}^{n} b_i \beta_i^{(n)}(t) \le \max\{b_0,\dots,b_n\}$

Eine effiziente und numerisch stabile Auswertung einer Bézierkurve erlaubt der

Algorithmus von de Casteljau

Mit Hilfe der Rekursionsformel aus Satz 6.12(2) erhält man

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

($b_i^{(0)}$ ← Eingabekoeffizienten)

$$= b_0^{(0)}(1-t)\beta_0^{(n-1)}(t) + b_1^{(0)}\left[t\beta_0^{(n-1)}(t) + (1-t)\beta_1^{(n-1)}(t)\right]$$
$$+ b_2^{(0)}\left[t\beta_1^{(n-1)}(t) + (1-t)\beta_2^{(n-1)}(t)\right] +$$
$$+ b_{n-1}^{(0)}\left[t\beta_{n-2}^{(n-1)}(t) + (1-t)\beta_{n-1}^{(n-1)}(t)\right] + b_n^{(0)} t\beta_{n-1}^{(n-1)}(t)$$

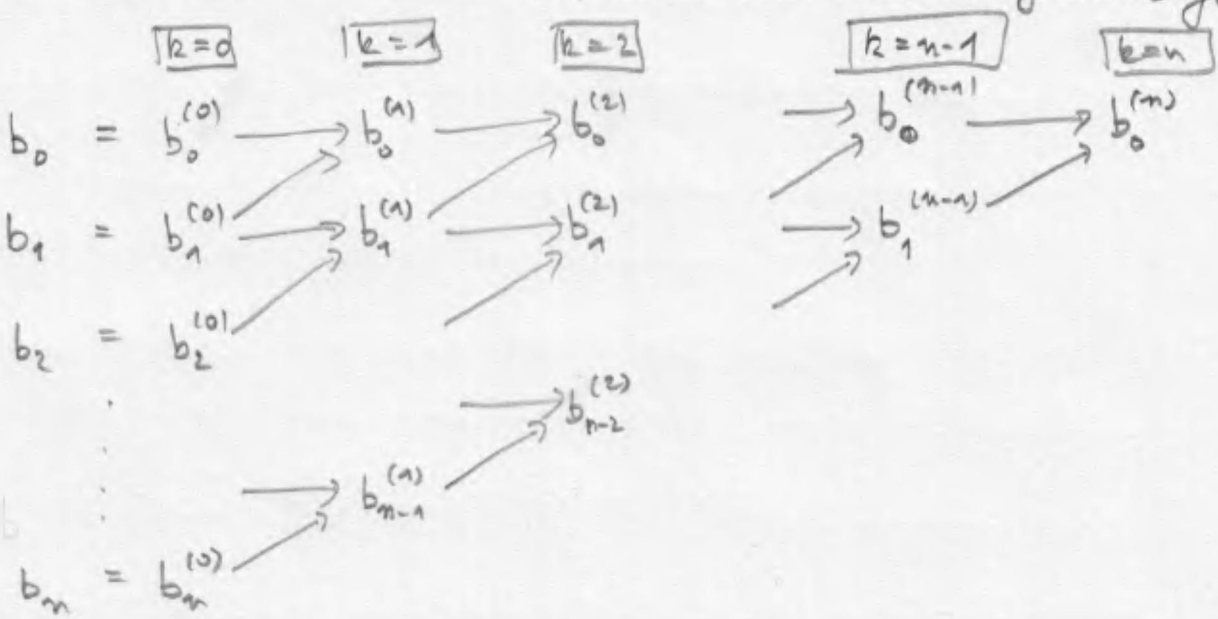
$$= \sum_{i=0}^{n-1} \underbrace{\left[b_i^{(0)}(1-t) + b_{i+1}^{(0)} t\right]}_{=: b_i^{(1)}} \beta_i^{(n-1)}(t)$$

d.h. wir haben nun n neue Koeffizienten $b_i^{(1)}$ für ein Polynom vom Grad n.
Allgemein:

Gegeben $b_i^{(0)} = b_i$ $\quad k=0,\ 0 \le i \le n$

Setze $b_i^{(k)} = b_i^{(k-1)}(1-t) + b_{i+1}^{(k-1)} t$ $\quad 0 < k \le n,\ 0 \le i \le n-k$

Die Auswertung von $B(t)$ erfolgt in einer ähnlichen Struktur wie beim Neville-Verfahren zur Auswertung von Polynomen:

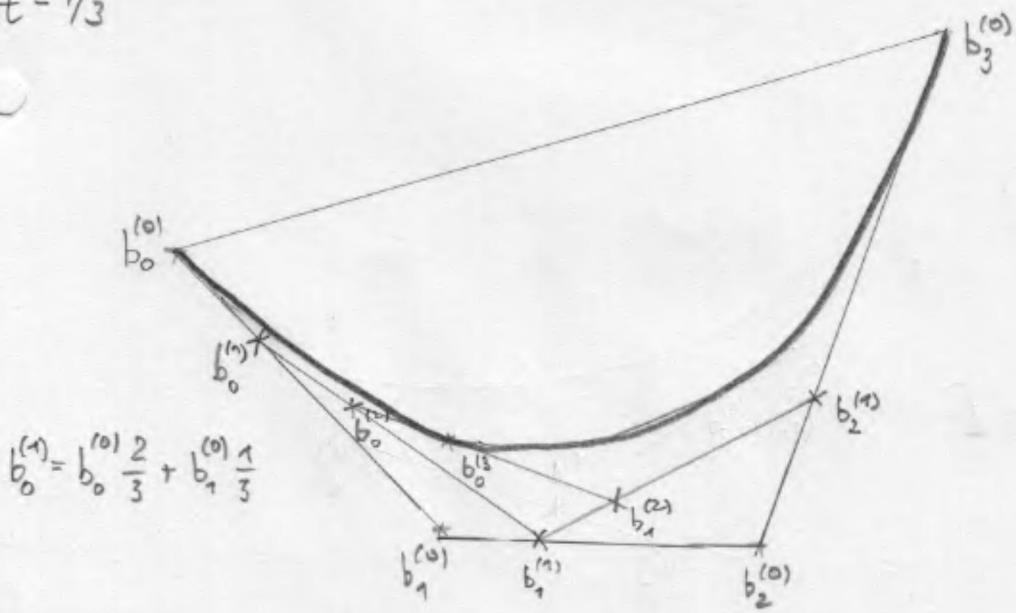


Und damit gilt dann $B(t) = b_0^{(n)} \underbrace{\beta_0^{(0)}(t)}_{=1 \text{ für alle } t} = b_0^{(n)}$

Geometrisch interpretiert man das folgendermaßen (Bernstein-Kurve aus Beispiel 6.14)

Beispiel 6.15 (Forts. von Bsp. 6.14)

$t = 1/3$



6.5 Splines

8.12.09

Wir kehren wieder zurück zur Interpolation.

Bis jetzt

- #Stützstellen = Polynomgrad + 1
- Großer Polynomgrad = viele Stützstellen $\overset{\text{evtl.}}{\Rightarrow}$ starke Abweichg zwischen den Stützstellen.

Idee: Stückweise Polynome niedrigen Grades.

Definition 6.16

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

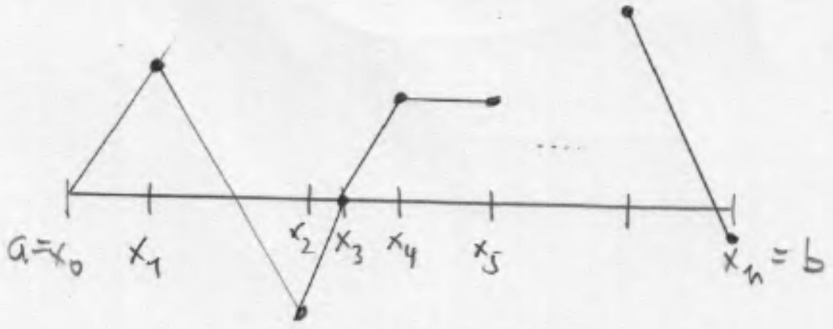
$$S^m(X) = \left\{ s \in C^{m-1}([a,b]) : s|_{[x_i, x_{i+1}]} \in P_m, \ 0 \le i < n \right\}$$

heißt Spline-Raum vom Grad m über der Zerlegung X.

Beispiel 6.17

$S^1(X)$ bedeutet:

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



Stetigkeit $\Rightarrow$ $s \in S^1(X)$ ist eindeutig durch Werte in den Stützstellen bestimmt

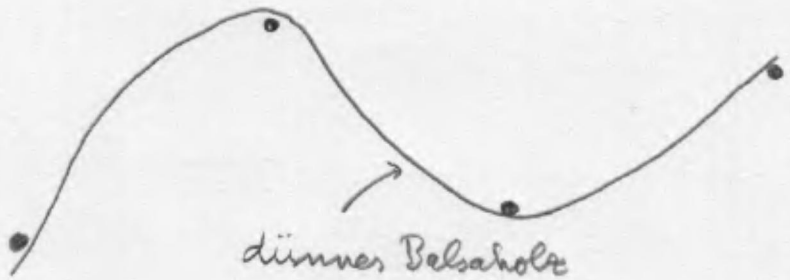
$\dim S^1(X) = |X|$.

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 Schiffs- und Flugzeugbau.



dünner Balsaholz

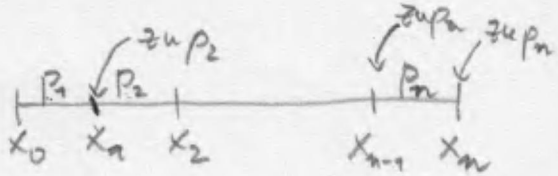
biegt sich unter Energieminimierung

$$\int_a^b \frac{|y''(t)|^2}{1+|y'(t)|^2}\,dt \approx \int_a^b |y''(t)|^2\,dt \to \min.$$

Krümmung. $|y'(t)| \ll 1$

Konstruktion von $s \in S^3(X)$. Die Funktion setzt sich stückweise aus n Polynomen zusammen:

$$s(x) = \begin{cases} p_i(x) & x \in [x_{i-1}, x_i) \quad i \in \{1,\dots,n\} \\ p_n(x_n) & x = x_n \end{cases}$$



Für die Polynome p_i gelten folg. Bed:

(a) Interpolationsbedingung (Stetigkeit)

$$i = 1,\dots,n: \quad \left.\begin{aligned} p_i(x_{i-1}) &= y_{i-1} \\ p_i(x_i) &= y_i \end{aligned}\right\} \text{ 2n Bedingungen}$$

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

$$i = 1,\dots,n-1 \quad \left.\begin{aligned} p_i'(x_i) &= p_{i+1}'(x_i) \\ p_i''(x_i) &= p_{i+1}''(x_i) \end{aligned}\right\} \text{ } 2(n-1) = 2n-2 \text{ Bedingungen}$$

$\Rightarrow$ zusammen $4n-2$ Bedingungen.

Pro Polynom p_i (vom Grad 3) hat man 4, also insgesamt $4n$ Freiheitsgrade.

Die fehlenden 2 Bedingungen erhält man durch Randbedingungen an den Stellen x_0 und x_n. Dabei gibt es verschiedene 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 natürliche RB (c)(i).

Satz 6.18 (Berechnung kubischer Splines)

Wir schreiben die Teilpolynome des Splines in der Form

$$p_i(x) = a_0^{(i)} + a_1^{(i)}(x-x_i) + a_2^{(i)}(x-x_i)^2 + a_3^{(i)}(x-x_i)^3 \quad i=1,\dots,n.$$

Die $a_2^{(i)}$ sind dann die Lösung des linearen Gleichungssystems der Dimension $n-1$:

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

wobei $a_2^{(0)} = a_2^{(n)} = 0$ (natürliche Randbedingung!) und $h_i := x_i - x_{i-1}$.

Die restlichen Koeffizienten ergeben sich zu:

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

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

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

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) Interpolationsbed. nutzen. Einsetzen von x_i:

$$y_i = p_i(x_i) = a_0^{(i)} \quad \Rightarrow \boxed{a_0^{(i)} = y_i} \quad i=1,\dots,n. \tag{6.7}$$

Das ist (6.2).

Einsetzen von x_{i-1}:

$$y_{i-1} = p_i(x_{i-1}) = a_0^{(i)} + a_1^{(i)}(-h_i) + a_2^{(i)} h_i^2 + a_3^{(i)}(-h_i^3) \quad i=1,\dots,n$$

($a_0^{(i)} = y_i$)

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

(iii) Randbedingungen einsetzen. Wir behandeln nur natürliche.

$$0 = p_1''(x_0) = 2a_2^{(1)} - 6a_3^{(1)} h_1 \qquad (6.9)$$

$$0 = p_n''(x_n) = 2a_2^{(n)} \quad \Rightarrow \boxed{a_2^{(n)} = 0} \qquad (6.10)$$

(iv) Stetigkeit der ersten Ableitung

$$p_i'(x_i) = p_{i+1}'(x_i) \qquad i = 1, \dots, n-1$$

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

(v) Stetigkeit der zweiten Ableitung

$$p_i''(x_i) = p_{i+1}''(x_i) \qquad i = 1, \dots, n-1$$

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

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

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

umnummerieren

$$\Leftrightarrow \quad \boxed{a_3^{(i)} = \frac{a_2^{(i)} - a_2^{(i-1)}}{3h_i}} \qquad i = 2, \dots, n \text{ aus (v)} \qquad (6.13)$$

$i = 1$ aus (6.9) wenn man $\boxed{a_2^{(0)} := 0}$ setzt.

also $i = 1, \dots, n$

Das ist (6.4)

(vii) $a_1^{(i)}$ durch $a_2^{(i)}$ ausdrücken. Dazu löse (6.8) nach a_1 auf:

$$a_1^{(i)} = \frac{y_i - y_{i-1}}{h_i} + h_i a_2^{(i)} - h_i^2 a_3^{(i)} \qquad i = 1, \dots, n$$

$$\overset{(6.13)}{\text{eins.}} = \frac{y_i - y_{i-1}}{h_i} + h_i a_2^{(i)} - h_i^2 \left(\frac{a_2^{(i)} - a_2^{(i-1)}}{3h_i} \right)$$

$$\boxed{a_1^{(i)} = \frac{y_i - y_{i-1}}{h_i} + \frac{h_i}{3}\left(2a_2^{(i)} + a_2^{(i-1)}\right)} \qquad i = 1, \dots, n$$

Das ist (6.3)

(viii) Nun setze $a_1^{(i)}$ und $a_3^{(i)}$ in die verbleibende Gleichung (6.11) ein

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

$$-2a_2^{(i+1)}h_{i+1} + 3h_{i+1}^2\left(\frac{a_2^{(i+1)} - a_2^{(i)}}{3h_{i+1}}\right) \qquad i = 1, \dots, n-1$$

$\Leftrightarrow$

$$a_2^{(i-1)}\left(+\frac{h_i}{3}\right) + a_2^{(i)}\left(\frac{2h_i}{3} - \frac{h_{i+1}}{3} + h_{i+1}\right) + a_2^{(i+1)}\left(-\frac{2h_{i+1}}{3} + 2h_{i+1} - h_{i+1}\right)$$

$$= \frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i}$$

mal 3

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

Und das ist (6.1).

Beachte, dass in (iii) und (vi) $a_2^{(0)} = a_2^{(n)} = 0$ gesetzt wurde. ∎

Zur Lösung des Tridiagonalsystem:

- GEM / LR-Zerlegung hat in diesem Fall $O(n)$ Aufwand! → sehr schnell.
- Das Gleichungssystem ist symmetrisch und strikt diagonal dominant

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

⇒ Regularität, stabile LR-Zerlegung ohne Pivotisierung

Beispiel 6.19 Aus Numerisch Beispiel 6.2.

Satz 6.20 (Fehlerabschätzung)

Sei $f \in C^4([a,b])$. Erfüllt der kubische Spline

$$s''(a) = f''(a) \quad \text{und} \quad s''(b) = f''(b)$$ (hatten wir oben nicht)

so gilt

$$\max_{a \le x \le b} |f(x) - s(x)| \le \frac{1}{2} h^4 \max_{a \le x \le b} \cdot |f^{(4)}(x)|$$

für $h := \max_{1 \le i \le n} |x_i - x_{i-1}|$.

Beweis: Warner / Schabacker II (wesentlich), siehe [Ramacher]. ∎

Selbst unter noch schwächeren Voraussetzungen konvergiert die Spline-Interpolierende gleichmäßig gegen f.

Beispiel 6.21

Zur Konvergenzordnung Beispiel 6.5 aus Num Stoch.

6.6 Trigonometrische Interpolation

11.01.10

Problem: Interpolation „periodischer" Funktionen, d.h. es gebe ein $\omega \in \mathbb{R}$, $\omega > 0$ so dass

$$f(x+\omega) = f(x) \qquad \forall x \in \mathbb{R}.$$

Es bietet sich die Interpolation mit „trigonometrischen Summen" an

$$t_n(x) = \frac{a_0}{2} + \sum_{k=1}^{m} \left\{ a_k \cos\left(\frac{k \cdot 2\pi}{\omega}\right) + b_k \sin\left(\frac{k \cdot 2\pi}{\omega}\right) \right\} \qquad (6.14)$$

$\frac{k}{\omega} \in [0,1]$

an, denn jeder Summand ist ω-periodisch: $\cos\left(k(x+\omega)2\pi/\omega\right) =$

$$\cos\left(k \cdot 2\pi/\omega + k\,2\pi\right) = \cos\left(\frac{k \cdot 2\pi}{\omega}\right).$$

(6.14) hat $2m+1$ Parameter, deshalb setze

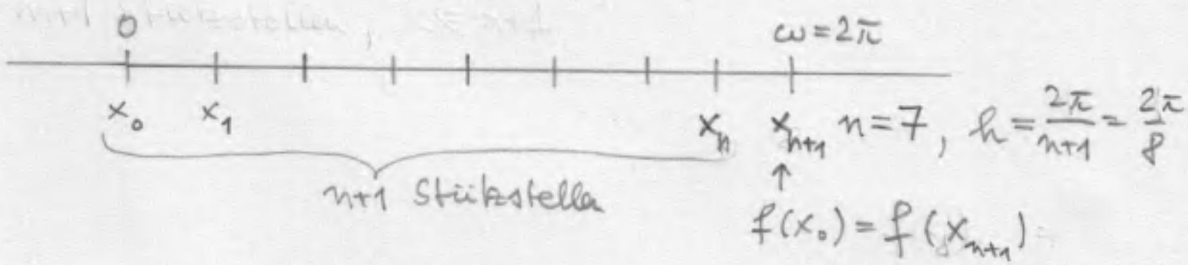
$$n := 2m$$

(bisher: Polynom vom Grad n hatte $n+1$ freie Parameter).

Ab jetzt: Nehme an, dass $\omega = 2\pi$ (sonst Transformation des Arguments).

Ausserdem: Verwende wieder äquidistante Stützstellen

$$x_k = \frac{2\pi k}{n+1} \qquad k = 0, \dots, n.$$



$n = 7$, $h = \frac{2\pi}{n+1} = \frac{2\pi}{8}$

$f(x_0) = f(x_{n+1})$

Es zeigt sich: Die Interpolationsaufgabe $t_n(x_k) = f(x_k)$ $k = 0, \dots, n$ ist zunächst einfacher im Körper $\mathbb{C}$ zu lösen!

D.h. betrachte das komplexe trigonometrische Polynom

$$t_n^*(x) = \sum_{k=0}^{n} c_k e^{ikx}$$

mit $i = \sqrt{-1}$ imaginäre Einheit

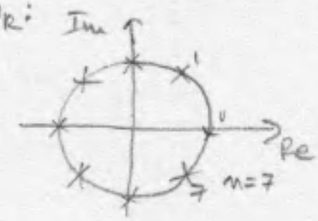
$c_k \in \mathbb{C}$ komplexe Koeffizienten

Eulersche Identität $e^{i\varphi} = \cos\varphi + i \sin\varphi$ für $\varphi \in \mathbb{R}$.

11.01.10

Wir untersuchen zunächst die Eigenschaften der komplexen Exponentialfunktion:

Die w_k:



Hilfssatz 6.22 (komplexe Einheitswurzeln)

Setze $w_k := e^{ix_k} = e^{i\frac{2\pi k}{n+1}}$ für alle $k \in \mathbb{Z}$ und gegebenes $n \in \mathbb{N}$.

$w_k \in \mathbb{C}$ heißt „k-te Einheitswurzel" und hat folgende Eigenschaften

a) $w_k^{n+1} - 1 = 0$ für alle $k \in \mathbb{Z}$.

M.a.W. die w_k sind Lösungen der Gleichung $w^{n+1} - 1 = 0$ in $\mathbb{C}$.

Beweis: $w_k^{n+1} = \left(e^{i\frac{2\pi k}{n+1}}\right)^{n+1} = e^{i 2\pi k} = \underbrace{\cos 2\pi k}_{=1} + i \underbrace{\sin 2\pi k}_{=0} = 1.$

b) $w_k^j = w_j^k \quad \forall\, j,k \in \mathbb{Z}$

Beweis: $w_k^j = \left(e^{i\frac{2\pi k}{n+1}}\right)^j = e^{i\frac{2\pi kj}{n+1}} = \left(e^{i\frac{2\pi j}{n+1}}\right)^k = w_j^k$

c) $w_k^{-j} = w_j^{-k} \quad \forall\, j,k \in \mathbb{Z}$

Beweis genau wie b, ist aber nicht identisch zu b).

d) $w_k^j = w_k^{j \bmod (n+1)} = w_{k \bmod (n+1)}^j = w_{k \bmod (n+1)}^{j \bmod (n+1)} \quad \forall\, j,k \in \mathbb{Z}.$

Zeige nur erste Identität, Rest analog:

Sei $j = r(n+1) + s$ mit $0 \le s < n+1$.

$w_k^j = e^{i\frac{2\pi kj}{n+1}} = e^{i\frac{2\pi[kr(n+1)+ks]}{n+1}} = \underbrace{e^{2\pi kr}}_{=1} \cdot \underbrace{e^{i\frac{2\pi ks}{n+1}}}_{(w_k)^s} = w_k^{j \bmod (n+1)}$

e) $\sum_{j=0}^{n} w_k^j = \begin{cases} n+1 & k \bmod (n+1) = 0 \\ 0 & \text{sonst} \end{cases}$

Sei $\underline{k=0}$: $w_0^j = e^0 = 1 \;\forall j$, also $\sum_{j=0}^{n} 1 = n+1$.

$\underline{k \neq 0}$: w_k ist nach a) Lösung von „Teleskopsumme"

$$w^{n+1} - 1 = (w-1)\left(w^n + w^{n-1} + \dots + 1\right) = 0$$

Für $k \neq 0$ ist $w_k \neq 1$ also $w_k - 1 \neq 0$ somit muss der zweite Faktor, also $\sum_{j=0}^{n} w_k^j = 0$ sein.

Satz 6.23 (Komplexe trigonometrische Interpolation)

Zu gegebenen Zahlen $Y_0, \dots, Y_n \in \mathbb{C}$ gibt es genau eine Funktion der Gestalt

$$t_n^*(x) = \sum_{k=0}^{n} c_k e^{ikx},$$

die den Interpolationsbedingungen

$$t_n^*(x_j) = Y_j \qquad j = 0, \dots, n, \quad x_j = \frac{2\pi j}{n+1}$$

genügt. Die komplexen Koeffizienten sind bestimmt durch

$$c_k = \frac{1}{n+1} \sum_{j=0}^{n} Y_j \underbrace{e^{-ijx_k}}_{= e^{-i\frac{2\pi jk}{n+1}} = w_k^{-j}} \qquad \forall\, k = 0, \dots, n. \tag{6.15}$$

Beweis: Mit der Abkürzung $w = e^{ix}$ gilt

$$t_n^*(x) = \sum_{k=0}^{n} c_k e^{ikx} = \sum_{k=0}^{n} c_k \underbrace{\left(e^{ix}\right)^k}_{=w} = \sum_{k=0}^{n} c_k w^k = p_n(w).$$

Transformation $w = e^{ix}$ eindeutig auf $0 \dots 2\pi$

Jedem t_n^* entspricht also genau ein komplexes Polynom p_n vom Grad n. Übertragung der Interpolationsbedingungen für t_n^* auf p_n ergibt:

$$t_n^*(x_j) = p_n(\underbrace{e^{ix_j}}_{=w_j}) = Y_j \qquad \forall\, j = 0, \dots, n.$$

Da die Polynominterpolation zu paarw. versch. Stützstellen (auch im Komplexen) eindeutig ist, gibt es genau ein solches p_n, also auch t_n^*.

Bleibt die Berechnung der Koeffizienten c_k. Diese ergeben sich durch das lineare Gleichungssystem

$$p_n(e^{ix_j}) = \sum_{\ell=0}^{n} c_\ell \left(e^{ix_j}\right)^\ell = \sum_{\ell=0}^{n} c_\ell e^{i\frac{2\pi \ell j}{n+1}} = \underbrace{\sum_{\ell=0}^{n} c_\ell w_j^\ell \overset{!}{=} Y_j}_{(n+1)\times(n+1)\text{ LGS für die Koeffizienten}} \qquad j = 0, \dots, n.$$

Das Gleichungssystem kann man explizit auflösen:

Für ein $k \in 0, \dots, n$:

$$\sum_{j=0}^{n} w_k^{-j} \underbrace{\left(\sum_{l=0}^{n} c_l w_j^l \right)}_{= y_j} = \sum_{l=0}^{n} c_l \underbrace{\left(\sum_{j=0}^{n} w_j^{l-k} \right)}_{= \sum_{j=0}^{n} w_{l-k}^{j} = \begin{cases} n+1 & l-k=0 \\ 0 & l-k \neq 0 \end{cases}} \overset{!}{=} \sum_{j=0}^{n} w_k^{-j} y_j$$

1) $\sum$ vertauschen
2) $w_k^{-j} = w_j^{-k}$

und damit

$$c_k (n+1) = \sum_{j=0}^{n} y_j w_k^{-j} \iff c_k = \frac{1}{n+1} \sum_{j=0}^{n} y_j e^{-ijx_k}$$

□.

Damit ist auch die reelle Interpolationsaufgabe gelöst, indem man $y_j \in \mathbb{R}$ annimmt.

Zu zeigen ist nun, wie die Koeffizienten a_k, b_k berechnet werden.

Satz 6.24 (Diskrete Fourier-Analyse)

Für $n \in \mathbb{N}_0$ gibt es zu gegebenen reellen Zahlen $y_0, \dots, y_n$ 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 $t_n(x_j) = y_j$, $j = 0, \dots, n$, $x_j = \frac{2\pi j}{n+1}$, sowie

$\theta = 0$, $m = \frac{n}{2}$ $\quad$ n gerade $\rightarrow a_0 \dots a_m, b_1, \dots, b_m$

$\theta = 1$, $m = \frac{n-1}{2}$ $\quad$ n ungerade $\rightarrow a_0, \dots, a_{m+1}, b_1, \dots, b_m$

(n gerade: $2 \cdot (n/2) + 1 = n+1$, n ungerade: $2 \cdot (n-1)/2 + 2 = n+1$).

Die Koeffizienten werden bestimmt durch

$$a_k = \frac{2}{n+1} \sum_{j=0}^{n} y_j \cos(jx_k), \quad b_k = \frac{2}{n+1} \sum_{j=0}^{n} y_j \sin(jx_k).$$

Beweisskizze: (ausführlich siehe Ra-Skript)

Man bestimmt die Koeffizienten c_k des komplexen trigonometrischen Polynoms zu reellen Daten $y_j \in \mathbb{R}$ und setzt dann

$$a_0 = 2c_0$$
$$a_k = c_k + c_{n+1-k} \qquad k = 1, \dots, m$$
$$b_k = i(c_k - c_{n+1-k}) \qquad k = 1, \dots, m$$
$$a_{m+1} = 2c_{m+1} \qquad n = 2m+1 \ (n \text{ ungerade}).$$

Dann rechnet man nach, dass $t_n^*(x_j) = t_n(x_j) = y_j \in \mathbb{R}$.

Wir wollen noch nachrechnen, dass die a_k reell sind:

$$a_k = c_k + c_{n+1-k}$$

$$= \frac{1}{n+1} \sum_{j=0}^{n} y_j \left(e^{-ijx_k} + e^{-ij x_{n+1-k}} \right) \qquad \text{(6.15)}$$

$$e^{-i\frac{2\pi j((n+1)-k)}{n+1}} = e^{i\frac{2\pi jk}{n+1}} = e^{ijx_k}$$

$$= \frac{1}{n+1} \sum_{j=0}^{n} y_j \left(e^{-ijx_k} + e^{ijx_k} \right)$$

$$= \frac{1}{n+1} \sum_{j=0}^{n} y_j \left(\underbrace{\cos(-jx_k)}_{=\cos(jx_k)} + i \underbrace{\sin(-jx_k)}_{=-\sin(jx_k)} + \cos(jx_k) + i\sin(jx_k) \right)$$

$$= \frac{1}{n+1} \sum_{j=0}^{n} y_j \, 2\cos(jx_k).$$

Schnelle Fouriertransformation

Einer der berühmtesten Algorithmen der angewandten Mathematik / Informatik!

Entwickelt von James Cooley und John Tukey 1965.

Zurück zur komplexen trigonometrischen Interpolation:

$$c_k = \frac{1}{N} \sum_{j=0}^{N-1} y_j e^{-i \frac{2\pi jk}{N}} \qquad k = 0, \dots, N-1 \quad \text{(Hintransformation)} \tag{6.15a}$$

$$y_j = \sum_{k=0}^{N-1} c_k e^{i \frac{2\pi kj}{N}} \qquad j = 0, \dots, N-1 \quad \text{(Rücktransformation)} \tag{6.15b}$$

(x_j)

- Hier wurde $N = n+1$ gesetzt.
- Die Berechnung der y_j aus den c_k ist schlicht die Definition des komplexen trig. Polynoms
- 6.15 bezeichnet man als diskrete Fouriertransformation.

Aufwand: Setze

$$c := (c_0, \dots, c_{N-1})^T, \quad y := (y_0, \dots, y_{N-1})$$

Dann ist (6.15a) äquivalent zu

$$c = \frac{1}{N} W y$$

mit $(W)_{kj} = e^{-i \frac{2\pi jk}{N}} = w_k^{-j}$ (Einheitswurzel)

Die Rücktransformation (6.15b) lautet

$$y = Uc \quad \text{mit} \quad (U)_{j,k} = w_j^k.$$

Wegen $U \frac{1}{N} W = I$ ist $W^{-1} = \frac{1}{N} U$, man hat also explizite Darst. der Inversen.

Da W, U voll besetzt sind beträgt der Aufwand für Hin- und Rücktransformation jeweils $O(N^2)$.

12.01.10

Betrachte (6.15a) ohne den Vorfaktor $\frac{1}{N}$, nenne das $\tilde{c}_k$.
Es sei $\underline{N \text{ gerade}}$, dann gilt:

$$\tilde{c}_k = \sum_{j=0}^{N-1} y_j e^{-i\frac{2\pi jk}{N}}$$

$$= \sum_{j=0}^{N/2-1} y_{2j} e^{-i\frac{2\pi 2jk}{N}} + \sum_{j=0}^{N/2-1} y_{2j+1} e^{-i\frac{2\pi(2j+1)k}{N}}$$

Aufspalten der Summe in gerade und ungar. Indizes →

gerader Teil | ungerader Teil

$$= \sum_{j=0}^{N/2-1} y_{2j} e^{-i\frac{2\pi jk}{N/2}} + e^{-i\frac{2\pi k}{N}} \sum_{j=0}^{N/2-1} y_{2j+1} e^{-i\frac{2\pi jk}{N/2}}$$

→ bringe N/2 im Nenner

Ausgeklammert!

$=: \tilde{c}_k^g$ $\qquad$ $=: \tilde{c}_k^u$

Wegen der $\underline{N/2}$-Periodizität der Einheitswurzeln $e^{-i\frac{2\pi k}{N/2}}$ gilt

$$\tilde{c}^g_{k+N/2} = \tilde{c}^g_k \quad \text{und} \quad \tilde{c}^u_{k+N/2} = \tilde{c}^u_k \qquad \forall\, k = 0, \dots, N/2-1.$$

Damit gilt:

- $\tilde{c}^g_k$, $\tilde{c}^u_k$, $k = 0, \dots, N/2-1$ berechnen sich jeweils durch eine DFT der Länge N/2
- Daraus berechnet man dann die ursprünglich gesuchten Koeffizienten

$$\tilde{c}_k = \tilde{c}^g_k + e^{-i\frac{2\pi k}{N}} \tilde{c}^u_k \qquad 0 \le k < N/2 \qquad (6.16a)$$

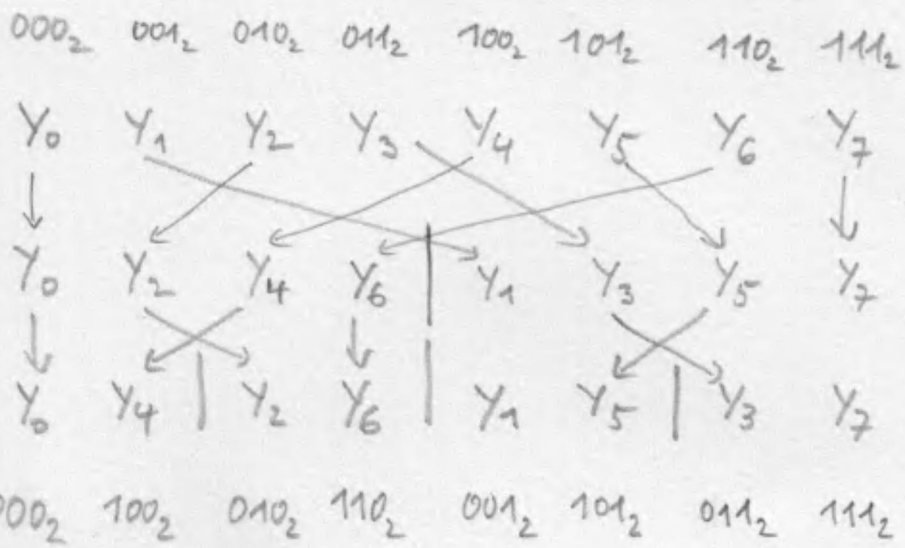
$$\tilde{c}_k = \tilde{c}_{k-N/2} + e^{-i\frac{2\pi k}{N}} \tilde{c}^u_{k-N/2} \qquad N/2 \le k < N \qquad (6.16b)$$

- Falls N/2 wieder gerade kann man das Prinzip rekursiv fortsetzen.
- Ist $N = 2^d$ eine Zweierpotenz erhält man schließlich eine DFT der Länge 2, die einfach durchführbar ist.

Beispiel $N = 8$.

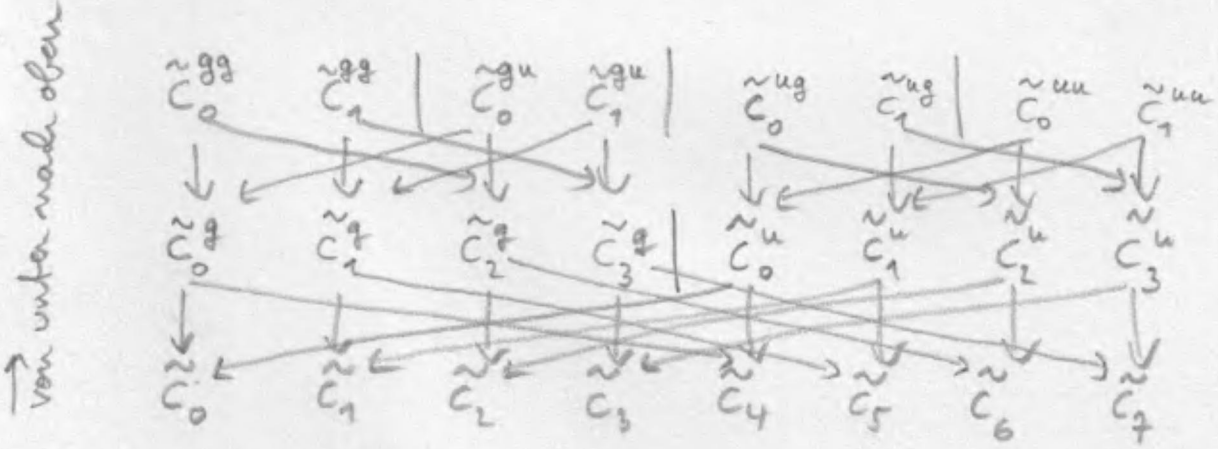
„Abstiegsphase" der Rekursion: Umsortieren der Eingabedaten.

Am besten schreibt man die Indizes zur Basis 2



Permutation $(b_{d-1} \dots b_0)_2 \rightarrow (b_0 \dots b_{d-1})_2$ heißt „bit reversal"

„Aufstiegsphase": Rekombination der Koeffizienten nach (6.16)



Dieses Verknüpfungsmuster nennt man „perfect shuffle".

Aufwand der FFT: $N = 2^d$ Zweierpotenz
$\Rightarrow d = ld\, N$.

$$A(N) = \underbrace{2A\left(\frac{N}{2}\right)}_{\text{2 Transf. der Länge } N/2 \text{ Ber.}} + \underbrace{cN}_{\text{Aufwand für (6.16)}}.$$

$A(N)$: # Gleitkommaoperationen für DFT der Länge N

$$= 2\left[2A\left(\frac{N}{4}\right) + c\frac{N}{2}\right] + cN$$

$$= 4A\left(\frac{N}{4}\right) + cN + cN$$

$$= 4\left(2A\left(\frac{N}{8}\right) + c\frac{N}{4}\right) + cN + cN$$

$$= 8A\left(\frac{N}{8}\right) + cN + cN + cN$$

d mal

$$= 2^d A(1) + \underbrace{cN + \ldots + cN}_{d-1 \text{ mal}} = cdN$$

($2^d = N$, $A(1) = c$)

$$= O(N ld N)$$

Deutlich schneller für große N.

Praktisches zur DFT:

- Spektralanalyse
- Beispiele auf dem Computer.

6.7 Approximation von Funktionen

12.01.10

Wir betrachten Approximation von Funktionen in Prähilberträumen:

Definition 6.25 Ein Vektorraum von Funktionen über $\mathbb{R}$ oder $\mathbb{C}$ mit Skalarprodukt heißt Prähilbertraum.

Beispiele:

a) Raum der stetigen Funktionen $C(a,b)$ mit dem Skalarprodukt

$$(f,g) = \int_a^b f(x)\,\overline{g(x)}\,dx.$$

b) Raum der quadratintegrierbaren Funktionen

$$L^2(a,b) = \left\{ f : \int_a^b |f(t)|^2\,dx < \infty \right\}$$

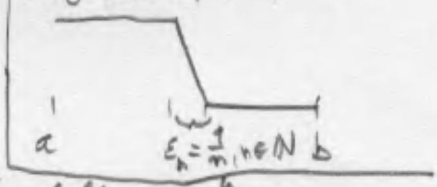
Dabei ist $\int_a^b \cdot\, dx$ das „Lebesgue-Integral" (→ siehe Funktionalanalysis).

Im Gegensatz zu $C(a,b)$ ist $L^2(a,b)$ vollständig bezüglich der Norm (jede Cauchy-Folge konvergiert.)

$$\|f\| = \sqrt{(f,f)}$$

und damit ein Hilbertraum.

Gegenbeispiel:



$L^2(a,b)$ ist das Analogon des $\mathbb{R}^n$ für Funktionenräume.

c) Gegeben $\Psi = \{\psi_1, \dots, \psi_N\}$, $\psi_i \in C(a,b)$ (z.B.)

$S = \operatorname{span}\Psi = \left\{ f : f = \sum_{i=1}^{N} c_i \psi_i \right\}$ ist endlichdimensionaler Prähilbertraum.

Im Folgenden sei H Prähilbertraum und $S \subseteq H$ ein endlichdimensionaler Teilraum.

Wir betrachten folgende Aufgabe:

Zu $f \in H$ finde $g \in S$ so dass

$$\| f - g \| \to \min \tag{6.17}$$

wobei $\|f\| = (f,f)^{1/2}$ die durch das Skalarprodukt induzierte Norm ist.

Satz 6.26 (Allgemeine Gauß-Approximation)

Die Aufgabe (6.17) hat genau eine Lösung $g \in S$.
Dies ist charakterisiert durch

$$(g, \varphi) = (f, \varphi) \quad \forall \varphi \in S. \tag{6.18}$$

Beweis.

a) Sei $\|f-g\|$ minimal für $g \in S$.

$F_\varphi(t) := \| f - (g + t\varphi) \|^2$, $t \in \mathbb{R}$

ist für beliebiges $\varphi \in S$ eine stetig differenzierbare Funktion mit Minimum bei $t=0$.

Also

$$\frac{d}{dt} F_\varphi(t)\Big|_{t=0} = \frac{d}{dt} (f-g-t\varphi, f-g-t\varphi)\Big|_{t=0}$$

$$= \frac{d}{dt} \left[(f-g, f-g) - 2t(f-g, \varphi) + t^2(\varphi, \varphi) \right]\Big|_{t=0}$$

$$= \left[2(f-g, \varphi) + 2t(\varphi, \varphi) \right]\Big|_{t=0}$$

$$= (f-g, \varphi) \stackrel{!}{=} 0 \quad \text{(für alle } \varphi \in S\text{)}$$

Dies ist (6.18).

Geometrisch: „Fehler" $f-g$ steht senkrecht auf allen $\varphi \in S$.

b) Sei nun $(f-g, \varphi) = 0 \ \forall \varphi \in S$. Für beliebiges $g' \in S$ gilt dann

$$\|f-g'\|^2 = \|f-g-\varphi\|^2 = (f-g-\varphi, f-g-\varphi)$$

($g' = g + \underbrace{g'-g}_{=:\varphi} = g+\varphi$)

$$= (f-g, f-g) - 2t \underbrace{(f-g, \varphi)}_{=0 \text{ n. Vor.}} + (\varphi, \varphi)$$

$$= \|f-g\|^2 + \underbrace{\|\varphi\|^2}_{\geq 0} \geq \|f-g\|^2 \quad \text{also } g \text{ Minimum.}$$

Damit ist die Aquivalenz gezeigt:

$$\|f-g\| \to \min \iff (g,\varphi) = (f,\varphi) \quad \forall \varphi \in S.$$

c) Eindeutigkeit des Minimums.

Angenomma es gibt zwei Minima g_1, g_2 mit $g_1 \neq g_2$. Dann ist

$$\|f-g_1\|^2 = \|f-g_2+\underbrace{g_1-g_2}_{=:\varphi}\|^2$$

$$= \|f-g_2\|^2 + \underbrace{\|g_1-g_2\|^2}_{>0 \text{ da } g_1 \neq g_2} > \|f-g_2\|^2 \quad \lightning \text{ zu } g_1 \text{ Minimum.}$$

d) Existenz des Minimums. Dies zeigen wir konstruktiv und erhalten damit auch einen Algorithmus zur Berechnung von g.

Da S endlich-dimensional gibt es eine Basis $\Psi = \{\psi_1, \dots, \psi_N\}$, $N = \dim S$. Das gesuchte Element g hat die Darstellung

$$g = \sum_{j=1}^{N} \alpha_j \psi_j .$$

Einsetzen der Basisdarstellung in (6.18) liefert

$$(g,\varphi) = (f,\varphi) \quad \forall \varphi \in S$$

$$\iff \left(\sum_{j=1}^{N} \alpha_j \psi_j, \psi_i\right) = (f,\psi_i) \qquad i = 1,\dots,N$$

$$\iff \sum_{j=1}^{N} \alpha_j (\psi_j,\psi_i) = (f,\psi_i) \qquad i = 1,\dots,N$$

$$\iff A\alpha = b$$

mit $(A)_{i,j} = (\psi_j,\psi_i)$ Massenmatrix oder Gramsche Matrix

$(b)_i = (f,\psi_i)$

A ist symmetrisch und positiv definit, dann

i) $(A)_{ij} = (\psi_j, \psi_i) \underset{\text{Sym. des SP.}}{=} (\psi_i, \psi_j) = (A)_{ji}$

ii) $$\alpha^T A \alpha = \sum_{i=1}^N \alpha_i \left(\sum_{j=1}^N (A)_{ij} \alpha_j \right) = \sum_{i=1}^N \alpha_i \left(\sum_{j=1}^N (\psi_j, \psi_i) \alpha_j \right)$$

$$= \sum_{i=1}^N \alpha_i \left(\sum_{j=1}^N \alpha_j \psi_j, \psi_i \right) = \Big(\underbrace{\sum_{j=1}^N \alpha_j \psi_j}_{=: g}, \underbrace{\sum_{i=1}^N \alpha_i \psi_i}_{=: g} \Big)$$

$= (g, g) > 0$ da $g \neq 0$ wenn $\alpha \neq 0$.

(Wir nehmen hier $\mathbb{K} = \mathbb{R}$ an, sonst hermitesch).

Da A s.p.d. hat das LGS $A\alpha = b$ genau eine Lösung für jedes b.

Nach a)-c) ist $\sum \alpha_j \psi_j$ die eindeutig bestimmte Bestapproximation. ∎

Approximation mit Orthonormalbasen

Besonders einfach wird die Lösung der Approximationsaufgabe wenn Ψ eine Orthonormalbasis ist, d.h. $(\psi_i, \psi_j) = \delta_{ij}$.

Dann gilt

$$\sum_{j=1}^N \alpha_j (\psi_j, \psi_i) = \alpha_i \overset{!}{=} (f, \psi_i) \qquad i = 1, \dots, N$$

und somit

$$g = \sum_{j=1}^N \alpha_j \psi_j = \sum_{j=1}^N (f, \psi_j) \psi_j$$

Beispiel 6.27 (Fourierreihe)

Für $N = 2m+1$, $m \in \mathbb{N}$ ist

$$\Psi_F = \left\{ \frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}} \cos x, \dots, \frac{1}{\sqrt{\pi}} \cos(mx), \frac{1}{\sqrt{\pi}} \sin x, \dots, \frac{1}{\sqrt{\pi}} \sin(mx) \right\}$$

eine Orthonormalbasis von Funktionen auf dem Intervall $[-\pi, \pi]$. (nachrechnen!).

Damit gilt dann

$$g(x) = \frac{a_0}{2} + \sum_{k=1}^{m} \{ a_k \cos(kx) + b_k \sin(kx) \}$$

und

$$a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(kx)\, dx, \qquad k = 0, \dots, m,$$

$$b_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(kx)\, dx, \qquad k = 1, \dots, m.$$

Für unendlich viele Glieder ($m = \infty$) nennt man die Reihe Fourier-Reihe. Diese konvergiert gegen ein Element aus $L^2(-\pi, \pi)$.

Ü: DFT lässt sich als näherungsweise Auswertung der Integrale mittels Trapezregel verstehen.

Fehlerkontrolle

Bisher: $S \subset H$ fest gewählt; Berechne das „optimale" g.
Der Fehler $\|f - g\|$ wird akzeptiert.

Bsp.: Fourierreihe für $m \to \infty$ sollte $\|f - g\|$ immer kleiner werden.

Verfeinerung der Approximationsaufgabe:

Finde $S \subset H$ mit $\dim S$ möglichst klein, so dass

$$\|f - g\| \leq TOL \quad (= \text{vorgegebene Zahl}).$$

Spezialfall „Kompression": H ist selbst schon endlich-dimensional!

Mittels Orthonormalbasen lässt sich der Fehler recht einfach messen.

Sei $S_N \subset H$, $N \in \mathbb{N}$, eine Folge von Approximationsräumen mit $\dim S_N = N$. Oft gilt (sogar) $S_N \subset S_{N+1}$.

Sei nun g_N die Bestapproximation von f in S_N.

Weiter sei Ψ_N eine Orthonormalbasis von S_N (oft gilt $\Psi_N \subset \Psi_{N+1}$).

Dann gilt für den Fehler:

$$0 \le \|f - g_N\|^2 = (f - g_N, f - g_N)$$

$$= (f,f) - 2(f, g_N) + (g_N, g_N)$$

Basis einsetzen ↘

$$= (f,f) - 2\Big(f, \sum_{i=1}^{N} \underbrace{(f,\psi_i)}_{\alpha_i}\psi_i\Big) + \Big(\sum_{i=1}^{N}(f,\psi_i)\psi_i, \sum_{j=1}^{N}(f,\psi_j)\psi_j\Big)$$

$$= (f,f) - 2\sum_{i=1}^{N}(f,\psi_i)(f,\psi_i) + \sum_{i=1}^{N}\sum_{j=1}^{N}\underbrace{(f,\psi_i)}_{\alpha_i}\underbrace{(f,\psi_j)}_{\alpha_j}\underbrace{(\psi_i,\psi_j)}_{=\delta_{ij}}$$

$$= (f,f) - 2\sum_{i=1}^{N}(f,\psi_i)^2 + \sum_{i=1}^{N}(f,\psi_i)^2$$

$$= (f,f) - \sum_{i=1}^{N}(f,\psi_i)^2 .$$

Wir wollen

$$\|f - g_N\|^2 = (f,f) - \sum_{i=1}^{N}(f,\psi_i)^2 \le \underbrace{TOL \cdot (f,f)}_{\text{relative Toleranz}}$$

$$\Longleftrightarrow \quad \sum_{i=1}^{N}(f,\psi_i)^2 \ge (1 - TOL)(f,f) \qquad (6.18)$$

$TOL \in [0,1)$

$TOL = 1 \Rightarrow N = 1$

$TOL = 0 \Rightarrow N = \infty$ und $g = f$.

(Kompression: $H = S_M$, dann ist $N = M$ und $\|f - g_N\| = 0$ möglich)

Bemerkung:

a) (f,f) sei berechenbar (zumindest mit ausreichender Genauigkeit. bei Kompression hat man das.

b) Falls $\Psi_N \subset \Psi_{N+1}$ sind einfach soviele Basisfunktionen hinzuzufügen bis (6.18) erreicht ist.

c) Für die Fourierreihe gilt $\Psi_N \subset \Psi_{N+1}$

d) Aus der obigen Fehlerdarstellung folgt bei $\Psi_N \subset \Psi_{N+1}$ unmittelbar

$$\|f - g_{N+1}\|^2 = (f,f) - \sum_{i=1}^{N+1}(f,\psi_i)^2 = \underbrace{(f,f) - \sum_{i=1}^{N}(f,\psi_i)^2}_{\|f - g_N\|^2} - \underbrace{(f,\psi_{N+1})^2}_{\ge 0}$$

$$\le \|f - g_N\|^2 .$$

Der Fehler nimmt also höchstens ab!

Adaptive Approximation mit Haar-Wavelets

Es bezeichne

$$tr(f) = \{x : f(x) \neq 0\}$$

den Träger einer Funktion.

In Oben gewählte ONB aus Sinus u. Cosinusfunktion haben alle ψ_i im wesentl. globalen Träger (d.h. $\overline{tr(f)} = [a,b]$)

Oft variiert die zu approximierende Funktion aber nur lokal sehr stark. In diesem Fall möchte man ein Funktionensystem mit folgenden Eigenschaften:

a) $(\psi_i, \psi_j) = \delta_{ij}$ Orthonormalität

b) $\Psi_N \subset \Psi_{N+1}$ Geschachtelt heit → „inkrementelles Hinzufügen"

c) $diam(tr(\psi_i)) \to 0$ für $i \to \infty$

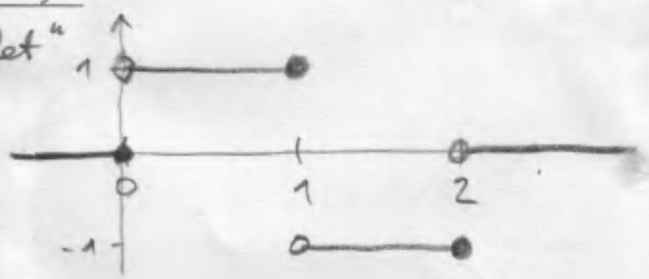
Diese Eigenschaften besitzen sog. Waveletfunktionen.

Wir untersuchen das sog. Haar-Wavelet.

Definition 6.28 (Haar-Wavelet)

Wir definieren das „mother wavelet"

$$\psi(x) = \begin{cases} 1 & 0 < x \leq 1 \\ -1 & 1 \leq x \leq 2 \\ 0 & \text{sonst} \end{cases}$$



und die Abschneidefunktion auf $(0,1]$:

$$\chi(x) = \begin{cases} 1 & 0 < x \leq 1 \\ 0 & \text{sonst} \end{cases}$$

Für $\ell \in \mathbb{N}_0$ (Stufe) und $0 \leq i < 2^{\ell-1}$ (Index) ist das Haar-Wavelet gegeben durch

$$\psi_i^\ell(x) = \max\left(\sqrt{2^{\ell-1}}, 1\right) \cdot \psi(2^\ell x - 2i) \cdot \chi(x).$$

Die Haar-Waveletbasis der Stufe ℓ ist

$$\Psi^\ell = \{\psi_0^0\} \cup \bigcup_{j=1}^{\ell} \bigcup_{i=0}^{2^{j-1}} \{\psi_i^j\}.$$

Wir fassen einige Eigenschaften der Haar-Wavelets zusammen: 16.01.10

a) Für $\ell > 0$ gilt

$$\psi_i^\ell(x) = \begin{cases} 0 & x \le \frac{2i}{2^\ell} \vee x > \frac{2i+2}{2^\ell} \\ \sqrt{2^{\ell-1}} & \frac{2i}{2^\ell} < x \le \frac{2i+1}{2^\ell} \\ -\sqrt{2^{\ell-1}} & \frac{2i+1}{2^\ell} < x \le \frac{2i+2}{2^\ell} \end{cases}$$

Betrachte Argument der mother wavelet

$2^\ell x - 2i \le 0 \iff x \le \frac{2i}{2^\ell} \to \psi_i^\ell(x) = 0$ für $x \le \frac{2i}{2^\ell}$

$2^\ell x - 2i \le 1 \iff x \le \frac{2i+1}{2^\ell} \to \psi_i^\ell(x) = 1$ für $\frac{2i}{2^\ell} < x \le \frac{2i+1}{2^\ell}$

$2^\ell x - 2i \le 2 \iff x \le \frac{2i+2}{2^\ell} \to \psi_i^\ell(x) = -1$ für $\frac{2i+1}{2^\ell} < x \le \frac{2i+2}{2^\ell}$

und $\psi_i^\ell(x) = 0$ für $x > \frac{2i+2}{2^\ell}$

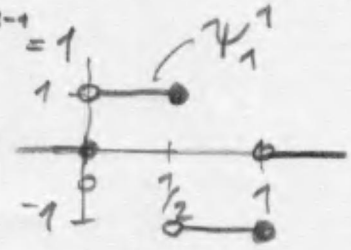
Der letzte Index ist $2^{\ell-1} - 1 \in \mathbb{N}_0$ (für $\ell > 0$!) und damit

$$\frac{2i+2}{2^\ell} \le \frac{2(2^{\ell-1}-1)+2}{2^\ell} = \frac{2^\ell - 2 + 2}{2^\ell} = 1.$$

Für $\ell > 0$ ist also $\psi(2^\ell x - 2i) = 0$ für $x \le 0$ und $x > 1$ und somit $\chi(x)$ überflüssig

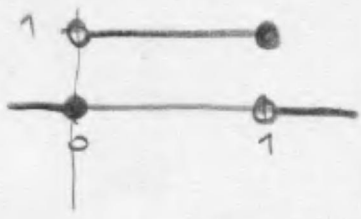
b) Speziell für $\ell = 1$ (in a) enthalten) gilt $0 \le i < 2^{1-1} = 1$

$$\psi_0^1(x) = \underbrace{\max\left(\sqrt{2^{1-1}}, 1\right)}_{=1} \underbrace{\psi(2x)}_{=0 \text{ für } x \le 0,\ x > 1} \cdot \chi(x) = \psi(2x)$$



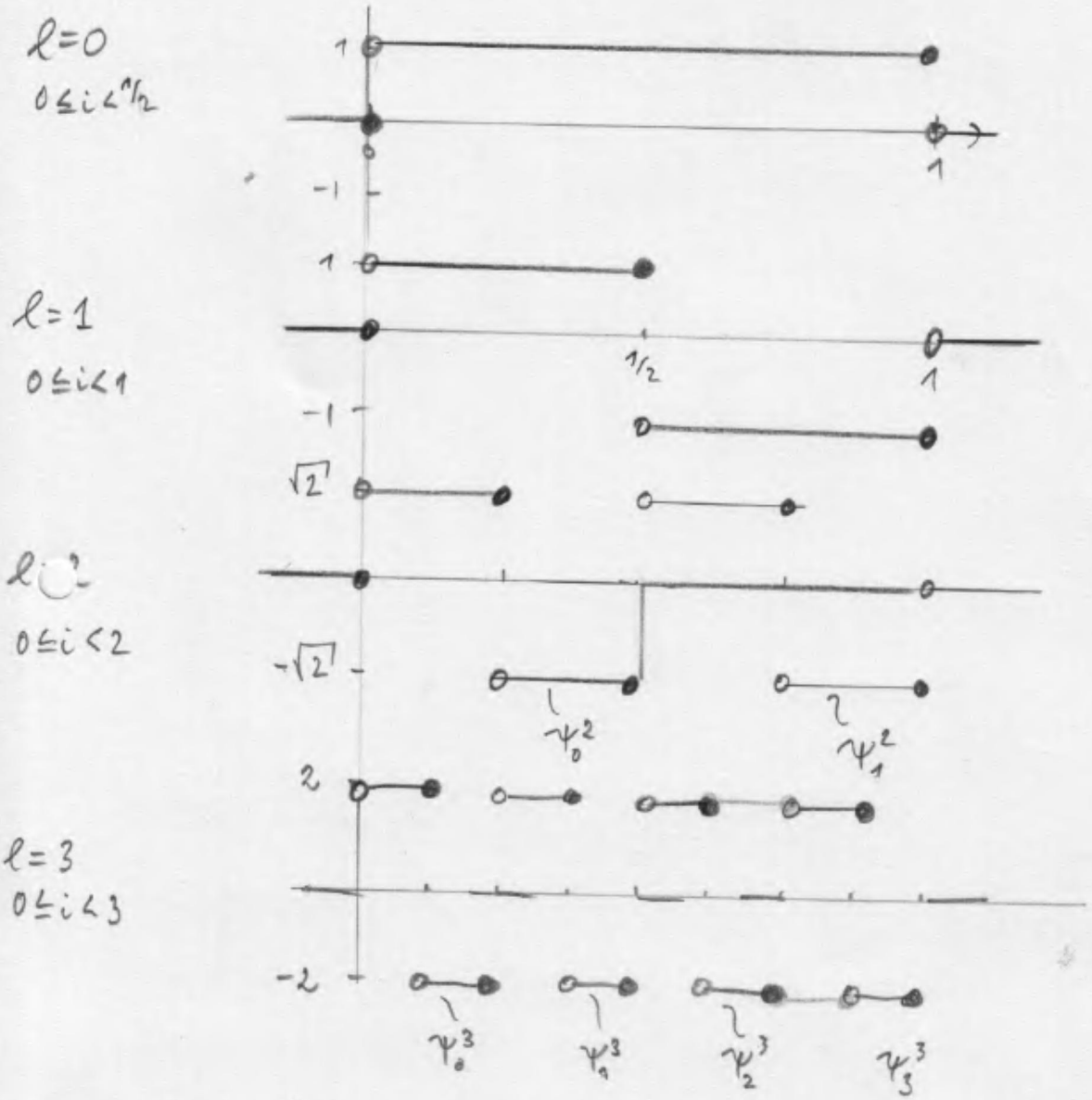
c) Für $\ell = 0$ gilt $0 \le i < 2^{-1} = 1/2 \to$ Ein Index $i = 0$

$$\psi_0^0(x) = \underbrace{\max\left(\sqrt{1/2}, 1\right)}_{=1} \cdot \underbrace{\psi(x)}_{=1 \text{ für } 0 < x \le 1} \cdot \chi(x) = \chi(x)$$



d) Graphische Veranschaulichung der Haar-Wavelets

16.01.10



$$\Rightarrow tr(\psi_i^l) = tr(\psi_{2i}^{l+1}) \cup tr(\psi_{2i+1}^{l+1})$$

Die Wavelets bilden eine Baumstruktur bezüglich der Inklusion der Träger.

ψ_j^{l+1} ist Kind von ψ_i^l genau dann wenn $tr(\psi_j^{l+1}) \subseteq tr(\psi_i^l)$

17.01.10

e) Orthogonalitätseigenschaften

Es gilt $(\psi_i^\ell, \psi_j^k) = \begin{cases} 1 & i=j \wedge \ell=k \\ 0 & \text{sonst.} \end{cases}$

1) $i=j \wedge \ell=k$

Für $\ell=0$ stimmt es offensichtlich

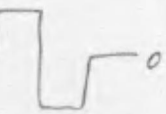
Für $\ell>0$ gilt $(\psi_i^\ell, \psi_i^\ell) = \int_{\frac{2i}{2^\ell}}^{\frac{2i+2}{2^\ell}} (\sqrt{2^{\ell-1}})^2 \, dx = 2^{\ell-1} \cdot \frac{1}{2^{\ell-1}} = 1$

2) Sei $\ell=k$ aber $i=j$

dann gilt $tr(\psi_i^\ell) \cap tr(\psi_j^k) = \emptyset$ also $(\psi_i^\ell, \psi_j^k) = 0$

3) O.B.d.A. sei $\ell > k$ und somit insb. $\ell > 0$

Dann ist ψ_j^k konstant auf $tr(\psi_i^\ell)$ und somit

$(\psi_i^\ell, \psi_j^k) = c \int_{\frac{2i}{2^\ell}}^{\frac{2i+2}{2^\ell}} \psi_i^\ell(x) \, dx = 0$ da ψ_i^ℓ 

(c = Wert von ψ_j^k)

f) Offensichtlich erfüllt die Konstruktion

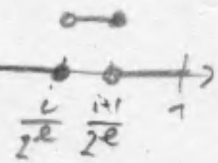
$$\Psi^\ell \subset \Psi^{\ell+1}$$

da immer nur Basisfunktionen hinzugenommen werden.

Damit sind alle geforderten Eigenschaften erfüllt.

g) Um die durch Ψ^ℓ darstellbaren Funktionen noch anschaulicher zu machen definieren wir die Funktionen

$$\varphi_i^\ell(x) = \begin{cases} 1 & \frac{i}{2^\ell} < x \leq \frac{i+1}{2^\ell} \\ 0 & \text{sonst} \end{cases}$$

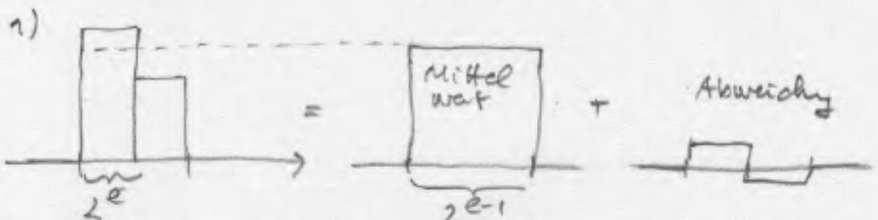
$\ell \in \mathbb{N}_0, \; 0 \leq i < 2^\ell$ 

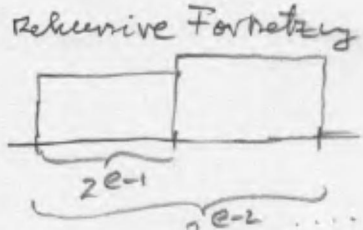
$$\Phi^\ell = \bigcup_{i=0}^{2^\ell-1} \{\varphi_i^\ell\}$$

$S^\ell = \operatorname{span} \Phi^\ell$ ist der Raum der stückweise konstanten Funktionen auf dem Intervall $(0,1]$ bezüglich der Unterteilung $(\frac{i}{2^\ell}, \frac{i+1}{2^\ell}]$.

Es gilt $\operatorname{span} \Psi^\ell = \operatorname{span} \Phi^\ell = S^\ell$.

Beweisidee:

1) 

2) Rekursive Fortsetzung 

Die Beweisidee zeigt auch wie man die Darstellung einer Funktion f bezüglich der Basis Φ^ℓ in eine Bezüglich der Basis Ψ^ℓ umrechnet.

Umgekehrt bedeutet $f(x) = \sum_\ell \sum_i c_i \psi_i^\ell$ die Umrechnung in die Basis Φ^ℓ (diesen Punkt kann man noch ausführlicher machen)

h) Effiziente Auswertung.

Hat man $f(x) = \sum_{j=0}^{\ell} \sum_{i=0}^{2^{j-1}-1} c_i^j \psi_i^j$ zu berechnen,

so spielen für ein $x \in (0,1]$ wegen der lokalen Träger nur wenige x eine Rolle.

Algorithmus:

$f = c_0^0 \cdot \psi_0^0(x)$;

$j = 1, i = 0$;

while (true) {

$f = f + c_i^j \psi_i^j(x)$;

if ($j == \ell$) break;

$i := \begin{cases} 2i & x \in \mathrm{tr}(\psi_{2i}^{j+1}) \\ 2i+1 & x \in \mathrm{tr}(\psi_{2i+1}^{j+1}) \end{cases}$

$j = j+1$;

}

Datenkompression mit Wavelets

Anwendung von (6,18). Geg: Darst von $f \in S^\ell$ in Basis Ψ^ℓ.
Gesucht: Teilraum $\tilde{S} \subset S^\ell$ so dass $\|f - \tilde{f}\| \leq TOL \cdot (f,f)$ und $\tilde{S}$ möglichst klein.

Algorithmus: $I = \{(0,0)\}$; $\quad I \subseteq \{(i,\ell) \mid \ell \in \mathbb{N}_0, 0 \leq i < 2^{\ell-1}\}$

Charakteristiat eine beliebige Menge von Wavelet Basen.

$\xi = (f, \psi_0^0)^2$;

while ($\varepsilon < (1-TOL)(f,f)$) {

Wähle Kind $(k, \ell+1)$ eines $(i,\ell) \in I$ mit $(f, \psi_k^{\ell+1})$ maximal.

$I = I \cup \{(k, \ell+1)\}$;

$\xi = \xi + (f, \psi_k^{\ell+1})^2$;

}

Ausblick

- Eine leichte Änderung von (6.18):

Finde $u \in S$ so dass

$$\int_\Omega \nabla u \cdot \nabla \varphi \, dx = \int_a^b f \varphi \, dx \qquad \forall \varphi \in S$$

führt auf die (numerische) Lösung der partiellen Differentialgleichung

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f \qquad \text{in } \Omega,$$

den Physikern als „Laplace-Gleichung" bekannt.

Mehr davon in Numerik 2 ....

- Die Welt ist nicht eindimensional.

Einen kurzen Ausflug ins Mehrdimensionale bei der Integration.

7 Numerische Integration

18.01.10

auch: „numerische Quadratur".
Wir behandeln die numerische Berechnung bestimmter Integrale in einer Raumdimension:

$$I_{(a,b)} = \int_a^b f(x)\,dx.$$

Alle von uns behandelten Verfahren führen auf folgende Form

$$I(f) = \sum_{i=0}^{n} w_i f(x_i) + \text{Fehler}.$$

Hierbei sind

$w_i \in \mathbb{R}$ die Gewichte und

$x_i \in \mathbb{R}$ die Stützstellen.

7.1 Newton-Cotes Formeln

sind sog. interpolatorische Quadraturformeln.
Idee: Stelle Interpolationspolynom p zu gewissen Stützstellen auf und berechne das Integral über p exakt.

Formel: Stützstellen $(x_i, f(x_i))$ $i = 0, \dots, n$

Interpolationspolynom in Lagrange-Darstellung

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

also

$$I(f) \approx I^{(n)}(f) = \int_a^b p_n(x)\,dx = \sum_{i=0}^{n} f(x_i) \int_a^b L_i^{(n)}(x)\,dx \qquad (7.1)$$

Bevor wir explizite Formeln für die w_i angeben noch eine

Definition 7.1 (Ordnung einer Quadraturformel)

18.01.10

Eine Quadraturformel $I^{(n)}(f)$ hat mindestens die Ordnung m, wenn sie Polynome vom Grad $m-1$ exakt integriert. ▨

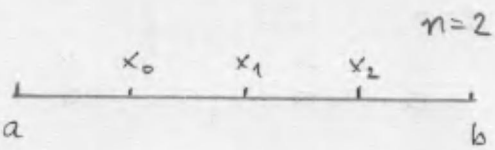
- Hier: $n+1$ Stützstellen $\Rightarrow$ Polynom vom Grad n exakt $\Rightarrow$ Ordnung mind. m.
- Später: Bei geschickter Wahl der Stützstellen max. Ordnung $2n+1$ bei $n+1$ Stützst.
- Für $f \equiv 1$ gilt $p \equiv 1$ und damit $\int_a^b 1\,dx = b-a = \sum_{i=0}^{n} w_i$.

Newton-Cotes-Formeln nutzen äquidistante Stützstellen:

Variante a) Abgeschlossene Formeln: $a, b \in$ Stützstellen

$$x_i = a + iH \quad i = 0, \dots, n, \quad H = \frac{b-a}{n}.$$

Variante b) Offene Formeln: $a, b \notin$ Stützstellen



$$x_i = a + (i+1)H \quad i = 0, \dots, n, \quad H = \frac{b-a}{n+2}.$$

Berechnung der Gewichte (abgeschlossene Form):

$$I^{(n)}(f) \overset{(7.1)}{=} \sum_{i=0}^{n} f(x_i) \int_a^b L_i^{(n)}(x)\,dx = (b-a) \sum_{i=0}^{n} \underbrace{\left(\frac{1}{b-a} \int_a^b L_i^{(n)}(x)\,dx \right)}_{=: w_i \text{ unabhängig von } a,b!} f(x_i)$$

Mittels Substitution $x = g(s) = a + sH \Rightarrow s = g^{-1}(x) = \frac{x-a}{H}$, $g'(x) = H$ ergibt sich:

$$w_i = \frac{1}{b-a} \int_a^b L_i^{(n)}(x)\,dx = \frac{1}{b-a} \int_{g^{-1}(a)}^{g^{-1}(b)} L_i^{(n)}(a+sH) \underbrace{g'(s)}_{=H}\,ds$$

$$= \frac{1}{b-a} \frac{b-a}{n} \int_0^n \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{[a+sH-(a+jH)]}{[a+iH-(a+jH)]}\,ds$$

$$= \frac{1}{n} \int_0^n \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{s-j}{i-j}\,ds.$$

19.01.10

Berechnet man diese Gewichte so erhält man:

a) Abgeschlossene Formeln $n = 1, 2, 3$, $H = \frac{b-a}{n}$

$$I^{(1)}(f) = \frac{b-a}{2}\{f(a)+f(b)\}$$ Trapez, Sehnen-Trapezregel

$$I^{(2)}(f) = \frac{b-a}{6}\left\{f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right\}$$ Simpsonregel / Keplersche Fassregel

$$I^{(3)}(f) = \frac{b-a}{8}\{f(a)+3f(a+H)+3(b-H)+f(b)\}$$ $\frac{3}{8}$-Regel

b) Offene Formeln $n = 0, 1, 2$, $H = \frac{b-a}{n+2}$

$$I^{(0)}(f) = (b-a)f\left(\frac{a+b}{2}\right)$$ Mittelpunkt, Tangenten-Trapez, Rechteckregel

$$I^{(1)}(f) = \frac{b-a}{2}\{f(a+H)+f(b-H)\}$$

$$I^{(2)}(f) = \frac{b-a}{3}\left\{2f(a+H)-f\left(\frac{a+b}{2}\right)+2f(b-H)\right\}$$

Bemerkung 7.2

Ab $n=7$ für abgeschlossene und $n=2$ für offene Formeln treten negative Gewichte w_i auf. Dies ist ungünstig weil

- Für $f(x) \geq 0 \; \forall x$ garantieren positive w_i dass $I^{(n)}(f) \geq 0$, sonst nicht.
- Erhöhte Gefahr der Auslöschung
- Konditionierung wird schlechter:

Sei $\tilde{f}(x_i) = f(x_i) + \Delta y_i$ mit $|\Delta y_i| \leq \varepsilon$ so gilt:

$$I^{(n)}(\tilde{f}) = \sum_{i=0}^{n} w_i(f(x_i)+\Delta y_i) = I^{(n)}(f) + \sum_{i=0}^{n} w_i \Delta y_i$$

also

$$|I^{(n)}(\tilde{f}) - I^{(n)}(f)| = \left|\sum_{i=0}^{n} w_i \Delta y_i\right| \leq \varepsilon \sum_{i=0}^{n} |w_i|$$

Sind alle w_i positiv so gilt $\sum_{i=0}^{n} |w_i| = \sum_{i=0}^{n} w_i = b-a$.

Sonst kann Kondition schlechter werden. ∎

Satz 7.3 (Restglieder)

19.01.10

Den begangenen Fehler kann man folgendermaßen abschätzen:

(i) Trapezregel

$$I(f) - \frac{b-a}{2}\{f(a)+f(b)\} = -\frac{(b-a)^3}{12} f''(\xi), \quad f \in C^2[a,b].$$

(ii) Simpson-Regel

$$I(f) - \frac{b-a}{6}\left\{f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right\} = -\frac{(b-a)^5}{2880} f^{(4)}(\xi), \quad f \in C^4[a,b].$$

(iii) Mittelpunktregel:

$$I(f) - (b-a) f\left(\frac{a+b}{2}\right) = \frac{(b-a)^3}{24} f''(\xi), \quad f \in C^2[a,b].$$

Wobei $\xi \in [a,b]$.

Beweis:

$$\int_a^b f(x) - p_n(x)\,dx \overset{\text{I-Fehler}}{=} \int_a^b \frac{f^{(n+1)}(\eta(x))}{(n+1)!} \prod_{j=0}^{n} (x-x_j)$$

Speziell für die Trapezregel gilt ($n=1$)

$$I(f) - I^{(1)}(f) = \frac{1}{2}\int_a^b f''(\eta(x)) \underbrace{(x-a)(x-b)}_{=:g(x)\leq 0}\,dx$$

Mittelwertsatz der Integralrechg. (Er fordert $g(x) \geq 0$ oder $g(x) \leq 0$) →

$$= \frac{1}{2} f''(\xi) \int_a^b (x-a)(x-b)\,dx = -\frac{(b-a)^3}{12} f''(\xi), \quad \xi \in [a,b].$$

Die beiden anderen Fälle sind etwas schwieriger wg. Vorzeichenwechsel von g, siehe Rannacher. ∎

- Mittelpunktregel hat den halben Fehler der Trapezregel bei nur einer Auswertung von f (genau).
- Restglieder haben immer die typische Form $c\,(b-a)^{m+1} f^{(m)}(\xi)$ (Kopplung!)

7.2 Summierte Quadraturformeln

19.01.10

Erhöhen des Polynomgrades ist wenig sinnvoll, da

- negative Gewichte auftreten
- Lagrange-Interpol. nicht punktweise konvergiert
- Entsprechende Differenzierbarkeit von f gegeben sein muss (↑ bei allen gekappt)

Idee: - Unterteile $[a,b]$ in N Teilintervalle

$$[x_i, x_{i+1}] \quad x_i = a + ih, \; i = 0, \dots, N-1, \; h = \frac{b-a}{N}$$

- Wende eine der obigen Formeln in jedem Teilintervall an
- Das Ergebnis nennt man „summierte Quadraturformel"

Satz 7.4 (Restglied für summierte Quadraturen)

Für die je Teilintervall verwendete Quadraturformel gelte

$$I_{[x_i,x_{i+1}]}(f) - I^{(n)}_{[x_i,x_{i+1}]}(f) = \underbrace{\alpha_n}_{\text{unabh. von } [x_i,x_{i+1}]} h^{m+2} f^{(m+1)}(\xi_i), \quad \xi_i \in [x_i, x_{i+1}].$$

Dann gilt für die summierte Quadraturformel

$$I(f) - \underbrace{I^{(n)}_h}_{\text{summierte Formel}}(f) = \alpha_n (b-a) h^{m+1} f^{(m+1)}(\xi), \quad \xi \in [a,b].$$

Trapezregel: $n = m = 1$, also $O(h^2)$. Die Ordnung ist (mind.) 2 (dies motiviert die Def. der Ordnung).

Simpson: $n = 2$, $m = 3$, also $O(h^4)$. Die Ordnung ist mindestens 3 (Pol. v. Grad 2)

Beweis: Zwischenwertsatz aus der Analysis:

Sei $g(x)$ stetig auf $[\alpha, \beta]$ dann $\exists$ zu jedem $u \in [\min(g(\alpha), g(\beta)), \max(g(\alpha), g(\beta))]$ mind. ein $\eta \in [\alpha, \beta]$ mit $g(\eta) = u$.

Aus dem Zwischenwertsatz folgt: Sei $\xi_i \in [a,b]$, $i = 0, \dots, N-1$ und g stetig auf $[a,b]$. Dann nimmt g jeden Wert zwischen $\min_i g(\xi_i)$ und $\max_i g(\xi_i)$ an.

Wegen $\min_i g(\xi_i) \le \frac{1}{N} \sum_{i=0}^{N-1} g(\xi_i) \le \max_i g(\xi_i)$ existiert $\xi \in [a,b]$ mit $\sum_{i=0}^{N-1} g(\xi_i) = N g(\xi)$.

Also: $I(f) - I^{(n)}_h(f) \overset{\text{Vor.}}{=} \sum_{i=0}^{N-1} \alpha_n h^{m+2} f^{(m+1)}(\xi_i) = \alpha_n h^{m+2} \sum_{i=0}^{N-1} f^{(m+1)}(\xi_i) = \alpha_n h^{m+2} N f^{(m+1)}(\xi)$

$$\overset{h = \frac{b-a}{N}}{=} \alpha_n h^{m+2} \frac{b-a}{h} f^{(m+1)}(\xi) = \alpha_n (b-a) h^{m+1} f^{(m+1)}(\xi). \qquad \blacksquare$$

19.01.10

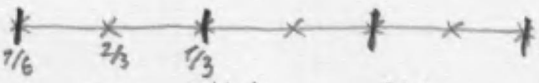
Beispiele für summierte Formeln:

(i) Summierte Trapezregel

$$I_h^{(1)}(f) = \sum_{i=0}^{N-1} \underbrace{\frac{x_{i+1}-x_i}{2}}_{=h} \left\{ f(x_i) + f(x_{i+1}) \right\} = h \left\{ \frac{f(a)}{2} + \sum_{i=1}^{N-1} f(x_i) + \frac{f(b)}{2} \right\}$$

$$I(f) - I_h^{(1)}(f) = -\frac{b-a}{12} h^2 f''(\xi), \quad \xi \in [a,b].$$

(ii) Summierte Simpson Regel



$$I_h^{(2)}(f) = \sum_{i=0}^{N-1} \underbrace{\frac{x_{i+1}-x_i}{6}}_{=h} \left\{ f(x_i) + 4 f\left(\frac{x_i+x_{i+1}}{2}\right) + f(x_{i+1}) \right\} = h \left\{ \frac{f(a)}{6} + \sum_{i=1}^{N-1} \frac{f(x_i)}{3} + \frac{2}{3} \sum_{i=0}^{N-1} f\left(\frac{x_i+x_{i+1}}{2}\right) + \frac{f(b)}{6} \right\}$$

$$I(f) - I_h^{(2)}(f) = -\frac{b-a}{2880} h^4 f^{(4)}(\xi), \quad \xi \in [a,b].$$

(iii) Summierte Mittelpunktregel

$$I_h^{(0)}(f) = \sum_{i=0}^{N-1} (x_{i+1}-x_i) f\left(\frac{x_i+x_{i+1}}{2}\right) = h \sum_{i=0}^{N-1} f\left(\frac{x_i+x_{i+1}}{2}\right)$$

$$I(f) - I_h^{(0)}(f) = \frac{b-a}{24} h^2 f''(\xi), \quad \xi \in [a,b].$$

Übung: Man zeige folgende Formeln:

$$\underbrace{I_h^{(2)}(f)}_{\text{Simpson}} = \frac{1}{3} \underbrace{I_h^{(1)}(f)}_{\text{Trapez}} + \frac{2}{3} \underbrace{I_h^{(0)}(f)}_{\text{Mittelpunktregel}}$$

$O(h^4)$ — je $O(h^2)$

$$I_{h/2}^{(1)} = \frac{1}{2} I_h^{(1)} + \frac{1}{2} I_h^{(0)}(f)$$

↑ halbe Gitterweite

Diese Formeln können zur Fehlerkontrolle verwendet werden. ↖ $|I(f) - I_h(f)| \leq TOL$.

Beispiel 7.5 (Zur Quadratur)

Siehe Rechner.

7.3 Quadraturen höherer Ordnung

19.01.10

Wie gesehen ist mit Newton-Cotes maximal $n=7$ ($n=2$, offen) sinnvoll.

Wie erreicht man höhere Ordnung?

Romberg-Integration (= Extrapolation zum Limes mit Trapezregel)

Mit den summierten Formeln ist man an $\lim_{h\to 0} I_h^{(n)}(f)$ interessiert. Extrapolation zum Limes ist anwendbar.

Herzstück ist die Euler-Maclaurinsche Summenformel:

$$I(f) - \underbrace{I_h^{(1)}(f)}_{\text{Trapezsumme!}} = \sum_{k=1}^{m} h^{2k} \frac{B_{2k}}{(2k)!} \left(f^{(2k-1)}(b) - f^{(2k-1)}(a)\right)$$

$$+ h^{2m+2} \frac{B_{2m+2}}{(2m+2)!} (b-a) f^{(2m+2)}(\xi), \quad \xi \in [a,b].$$

Die Trapezsumme hat also eine Fehlerentwicklung in gerade Potenzen. Damit ist die Extrapolation besonders effizient.

B_i: Bernoulli Zahlen.

Gauß-Integration

Frage: Lässt sich die Genauigkeit von Quadraturformeln durch nichtäquidistante Stützstellen verbessern?

Idee: Wähle w_i, x_i so dass Polynome von möglichst hohem Grad exakt integriert werden (Ordnungsmaximierung).

Satz 7.6 Die maximale Ordnung einer Quadraturformel mit $n+1$ Stützstellen ist $2n+2$ (d.h. Polynome vom Grad $2n+1$ werden exakt integriert).

Beweis: Ang. man könnte Polynome vom Grad $2n+2$ bei $n+1$ Stützst. exakt integrieren (Ordn. $2n+3$). Betrachte

$$q(x) = \prod_{i=0}^{n} (x-x_i)^2$$

- $q(x)$ hat Grad $2n+2$
- $q(x) \geq 0$ $\forall x$ und $q(x) \not\equiv 0$ also $\int_{-1}^{1} q(x)\,dx > 0$
- Andererseits: $q(x_i) = 0$ und damit $\sum_{i=0}^{n} q(x_i) w_i = 0$, d.h. q wird nicht exakt integriert ↯ ∎.

Die maximal mögliche Ordnung wird mit der Gauß-Quadratur erreicht:

Satz 7.7 (Gauß-Quadratur)

Es gibt genau eine interpolatorische Quadraturformel zu $n+1$ paarweise verschiedenen Stützstellen in $[-1,1]$ mit der Ordnung $2n+2$.
Ihre Stützstellen sind die Nullstellen $\lambda_0, \dots, \lambda_n \in (-1,1)$ des $(n+1)$-ten Legendrepolynoms L_{n+1}

$$L_0(x) = 1, \quad L_1(x) = x, \quad L_{n+1}(x) = \frac{2n+1}{n+1} L_n(x) - \frac{n}{n+1} L_{n-1}(x).$$

Die Gewichte erhält man mittels

$$w_i = \int_{-1}^{1} \prod_{\substack{j=0 \\ j \neq i}}^{n} \left(\frac{x - \lambda_j}{\lambda_i - \lambda_j} \right)^2 dx > 0 \qquad i = 0, \dots, n.$$

Bem: Die Legendrepolynome bilden Orthogonalsystem $\int_{-1}^{1} L_n(x) L_m(x) = 0$ $n \neq m$.

Beispiele: $h = \frac{b-a}{2}$, $c = \frac{b+a}{2}$

$n=1$: $I^{(1)}(f) = \frac{b-a}{2} \left\{ f(c - \sqrt{1/3}\, h) + f(c + \sqrt{1/3}\, h) \right\}$ Ordnung 4

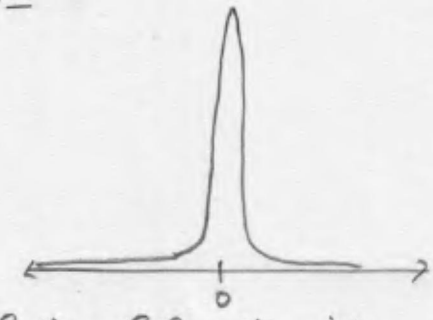
$n=2$: $I^{(2)}(f) = \frac{b-a}{18} \left\{ 5 f(c - \sqrt{3/5}\, h) + 8 f(c) + 5 f(c + \sqrt{3/5}\, h) \right\}$ Ord. 6.

Entsprechend lassen sich summierte Formeln definieren.

7.4 Ausblick

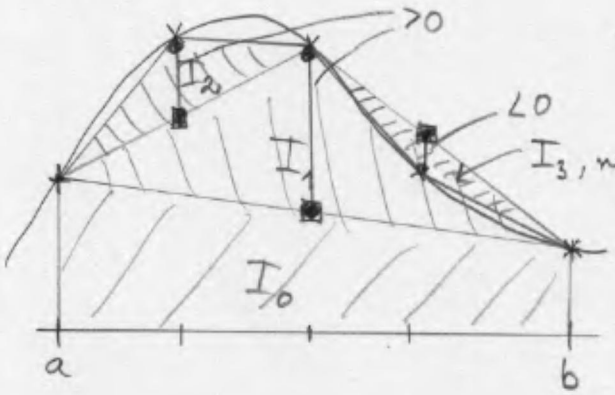
Adaptive Quadratur

Betr. $f(x) = \frac{1}{10^{-5}+x^2}$



Summierte Quadraturen mit fester Schrittweite ineffizient.

Prinzip von Archimedes:



1) $I(f) = I_0 + I_1 + I_2 + \dots\dots$

2) Breche rekursive Unterteilung ab falls $|I_j|$ klein genug.

Mehrdimensionale Quadraturen

Die Welt ist nicht eindimensional.

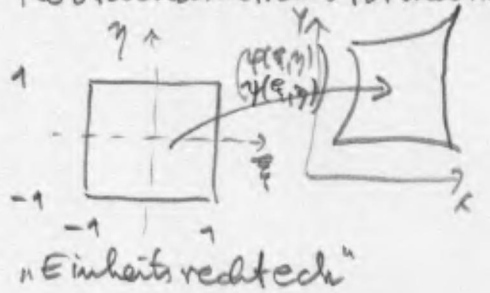
Für Rechtecke ($d=2$), Quader ($d=3$), ... lassen sich obige Formeln leicht erweitern:

$$\int_c^d \int_a^b f(x,y)\,dx\,dy \approx \int_c^d \sum_{i=0}^{n} f(x_i,y)\,w_i\,dy = \sum_{i=0}^{n} \int_c^d f(x_i,y)\,dy\; w_i$$

$$\approx \sum_{i=0}^{n}\sum_{j=0}^{n} f(x_i,y_j)\,\underbrace{w_i\,w_j}_{=w_{ij}}$$

Allerdings sind nicht alle Gebiete Rechtecke (anders als in 1D!)

Koordinatentransformation:



$$\text{Abb} \begin{pmatrix}\varphi(\xi,\eta)\\ \psi(\xi,\eta)\end{pmatrix} : [-1,1]\times[-1,1] \to \Omega \subset \mathbb{R}^2$$

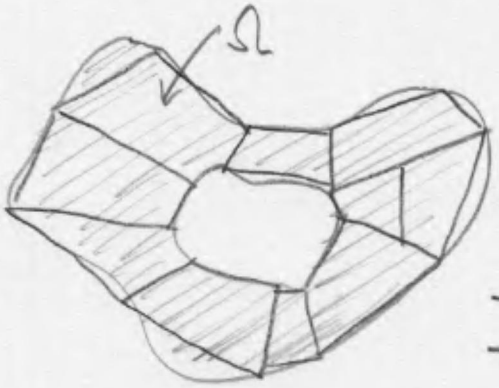
$$\int_\Omega f(x,y)\,dx\,dy = \int_{-1}^{1}\int_{-1}^{1} f(\varphi(\xi,\eta),\psi(\xi,\eta)) \left|\frac{\partial(\varphi,\psi)}{\partial(\xi,\eta)}\right| d\xi\,d\eta$$

$$\left|\frac{\partial(\varphi,\psi)}{\partial(\xi,\eta)}\right| = \det \begin{bmatrix} \frac{\partial\varphi}{\partial\xi}(\xi,\eta) & \frac{\partial\psi}{\partial\xi}(\xi,\eta) \\ \frac{\partial\varphi}{\partial\eta}(\xi,\eta) & \frac{\partial\psi}{\partial\eta}(\xi,\eta) \end{bmatrix}$$

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 $\underline{2^d}$ 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

$$I(f) \approx \frac{C}{N} \sum_{i=1}^{N} f(\xi_i) \quad \text{mit Zufallszahlen } \xi_i \in \Omega.$$

8 Iterative Lösung von Gleichungssystemen

25.01.10

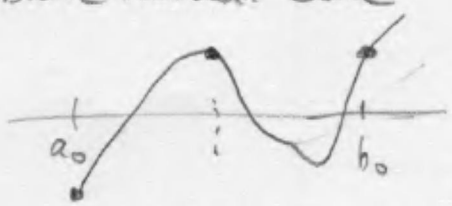
In diesem Abschnitt betrachten wir die Lösung von algebraischen Gleichungen

$$f(x) = 0 \qquad \text{mit} \qquad f: \mathbb{R}^n \to \mathbb{R}^n.$$

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 mind. eine Nullstelle in $[a_0, b_0]$.



Algorithmus:

Geg. $I_0 = [a_0, b_0]$ 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:

25.01.10

Es gilt $a_t \leq a_{t+1} < b_{t+1} \leq b_t$

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 $f(x_t) \equiv 0$).

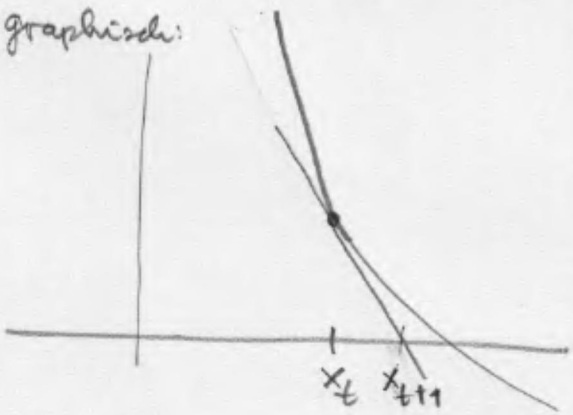
Bemerkung:

- 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: graphisch:



geg. x_t. Da $f \in C^1$ gibt es a 'Tangente'

$$T_t(x) = f'(x_t)(x - x_t) + f(x_t)$$

Nullstelle der Tangente:

$$T_t(x) = 0 \iff x = x_t - \frac{f(x_t)}{f'(x_t)}$$

Das 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 $f: \mathbb{R}^n \to \mathbb{R}^n$ erweitern:

Es existiere die Taylorentwicklung von f:

$$f_i(x_t+\Delta x) = f_i(x_t) + \sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j}(x_t)\,\Delta x_i + R_i(x_t, \Delta x) \quad i = 1, \dots, n$$

- in vektorieller Schreibweise

$$f(x_t+\Delta x) = f(x_t) + J(x_t)\,\Delta x + R(x_t, \Delta x)$$

$$\left(J(x_t)\right)_{ij} = \frac{\partial f_i}{\partial x_j}(x_t) \quad \text{„Jacobimatrix"}$$

Ignorieren des Restgliedes entspricht „Linearisierung von f".

$$f(x_t) + J(x_t)\,\Delta x \stackrel{!}{=} 0$$

$$\Leftrightarrow \quad \Delta x = -J^{-1}(x_t)\, f(x_t)$$

führt zur Iteration

$$x_{t+1} = x_t - J^{-1}(x_t)\, f(x_t)$$

Jeder Schritt erfordert Lösung eines LGS mit der Jacobimatrix!

Nun untersuchen wir die Konvergenz des Newton-Verfahrens. Allerdings nur im $\mathbb{R}^1$.

25.01.10

Satz 8.1 (Newton-Verfahren)

Die Fkt $f \in C^2[a,b]$ habe in (a,b) (Innen!) eine Nullstelle 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

$$q := \frac{M}{2m} \varrho < 1 \;, \quad K_\varrho(z) := \{ x \in \mathbb{R} : |x - z| \le \varrho \} \subset [a,b].$$

Dann sind für jeden Startwert $x_0 \in K_\varrho(z)$ die Newton-Iteraten $x_t \in K_\varrho(z)$ definiert und konvergieren gegen die Nullstelle z.

Dabei gilt die a-priori Fehlerabschätzung

$$|x_t - z| \le \frac{2m}{M} q^{(2^t)}, \quad t \in \mathbb{N}$$

a-priori: nur Abh. von den Voraussetzge

und die a-posteriori Fehlerabschätzung

$$|x_t - z| \le \frac{1}{m} |f(x_t)| \le \frac{M}{2m} |x_t - x_{t-1}|^2, \quad t \in \mathbb{N}.$$

a-posteriori: auch Abh. von bereits berechnete Iteraten.

Beweis: Vorbereitung

i) Mittelwert der Differentialrechnung liefert für alle $x, y \in [a,b]$, $x \neq y$:

$$\left| \frac{f(x) - f(y)}{x - y} \right| = |f'(\xi)| \underset{\text{Vor.}}{\ge} m \quad \Leftrightarrow |x - y| \le \frac{1}{m} |f(x) - f(y)|$$

- f ist Lipschitz-stetig
- die Nullstelle z ist eindeutig, da sonst $0 < |z_1 - z_2| \le \frac{1}{m} |f(z_1) - f(z_2)| = 0$ ↯

ii) Da $f \in C^2[a,b]$ gilt Folg. Taylordarstellg:

$$f(y) = f(x) + (y - x) f'(x) + \underbrace{\int_x^y (x - t) f''(t)\, dt}_{=: R(y;x)}$$

Restglied

Transformation des Integrals mit

$$\varphi(s) = x + s(y-x) \qquad \varphi: [0,1] \to [x,y]$$

liefert für das Restglied

$$R(y;x) = \int_x^y (x-t) f''(t)\,dt = \int_0^1 (x-\varphi(s)) f''(\varphi(s)) \varphi'(s)\,ds$$

$$= \int_0^1 (\underbrace{x-x}_{=0} - s(y-x)) f''(x+s(y-x)) (y-x)\,ds$$

$$= -(y-x)^2 \int_0^1 s f''(x+s(y-x))\,ds.$$

Und damit

$$|R(y;x)| \le (y-x)^2 \int_0^1 s \underbrace{|f''(x+s(y-x))|}_{\le M \text{ n.Vor}} ds \le \frac{M}{2} |y-x|^2.$$

iii) Nun setze $g(x) := x - \frac{f(x)}{f'(x)}$ (d.h. $x_{t+1} = g(x_t)$)

Dann gilt:

$$g(x) - z = x - \frac{f(x)}{f'(x)} - z = -\frac{1}{f'(x)} \underbrace{\{ f(x) + (z-x) f'(x) \}}_{= -R(z;x)}$$

wg: $\underbrace{f(z)}_{=0} = f(x) + (z-x) f'(x) + R(z;x)$

Für $x \in K_\rho(z)$ gilt dann

$$|g(x)-z| = \left| \frac{1}{f'(x)} R(z;x) \right| \le \frac{1}{m} \frac{M}{2} |z-x|^2 = \underbrace{\frac{M}{2m}}_{<1 \text{ n. Wahl von } \rho} \underbrace{|x-z|}_{\le \rho} \underbrace{|x-z|}_{\le \rho \text{ da } x \in K_\rho(z)} < \rho$$

m: min. macht es größer; wg: < 1

Somit folgt aus $x \in K_\rho(z)$, dass auch $g(x) \in K_\rho(z)$.
g bildet die Menge $K_\rho(z)$ auf sich selbst ab. (X)

Die Newton-Iteriarten sind $x_{t+1} = g(x_t)$. 25.01.10

(*) von oben

Setze $\varrho_t := \frac{M}{2m} |x_t - z|$. Dann gilt mit der Absch. von oben

$$\varrho_t = \frac{M}{2m} |x_t - z| = \frac{M}{2m} |g(x_{t-1}) - z| \overset{(*)}{\leq} \frac{M}{2m} \frac{M}{2m} |x_{t-1} - z|^2 = \varrho_{t-1}^2$$

Somit gilt nach t Schritten:

$$\varrho_t \leq \varrho_{t-1}^2 \leq \varrho_{t-2}^4 \cdots \leq \varrho_{\underbrace{t-t}_{=0}}^{(2^t)} = \varrho_0^{(2^t)}$$

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

$$|x_t - z| \leq \frac{1}{m} |f(x_t) - \underbrace{f(z)}_{=0}| = \frac{1}{m} |R(x_t; x_{t-1})| \leq \frac{M}{2m} |x_t - x_{t-1}|^2$$

↑ Lipschitz von Anfang

∎

Beispiel 8.2 (Wurzelberechnung mit Newton-Verfahren)

25.01.10

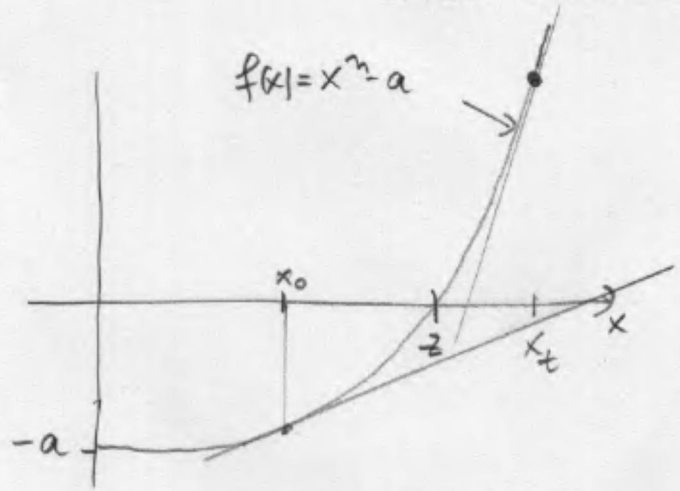
$a>0$, $n \geq 1$, löse $x^n = a \iff f(x) = x^n - a = 0$, $f'(x) = n x^{n-1}$

Also

$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)} = x_t - \frac{x_t^n - a}{n x_t^{n-1}} = \frac{n x_t^n - x_t^n + a}{n x_t^{n-1}} = \frac{(n-1)x_t^n + a}{n x_t^{n-1}}$$

$$= \frac{1}{n}\left\{ (n-1)x_t + \frac{a}{x^{n-1}} \right\}$$

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 $f'(x_t) > \frac{f(x)}{x_t - z}$

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 <u>lokal</u>, d.h. wenn $|x_0 - z| \leq \rho$. → „Einzugsbereich". Wobei
  -- ρ i.d.R. unbekannt
  -- ρ möglicherweise sehr klein ist. Oben: $\frac{M}{2m}\rho < 1$
  $\Rightarrow \rho < \frac{2m}{M}$ ($m \leq \min f'$, $M \geq \max f''$)
- Newton-Verfahren konvergiert quadratisch.
  $|x_t - z| \leq c\,|x_{t-1} - z|^2$. zum Vgl Intervallsch: $|x_t - z| \leq \frac{1}{2}|x_{t-1} - z|$
- gedämpftes Newton-Verfahren: Verbesserung der Konvergenz <u>ausserhalb</u> des Einzugsbereichs:

$$x_{t+1} = x_t - \lambda_t \frac{f(x_t)}{f'(x_t)}, \quad \lambda_t \in (0,1]$$

Wahl von λ_t „Dämpfungsstrategie".

- Mehrfache Nullstellen

Sei z zunächst zweifache Nullstelle, d.h. $f(z) = f'(z) = 0$ und $f''(z) \neq 0$.

Wegen

$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)} = x_t - \frac{f(x_t) - \overbrace{f(z)}^{=0}}{f'(x_t) - \underbrace{f'(z)}_{=0}} \overset{\text{Erweitern}}{\downarrow} = x_t - \frac{\frac{f(x_t)-f(z)}{x_t - z}}{\frac{f'(x_t)-f'(z)}{x_t - z}} = x_t - \frac{f'(\xi_t)}{f''(\eta_t)}$$

und $f''(z) \neq 0$ bleibt die Iteration für $x_t \to z$ (und damit $\eta_t \to z$) wohldefiniert.

Man zeigt: Für p-fache Nullstelle zeigt

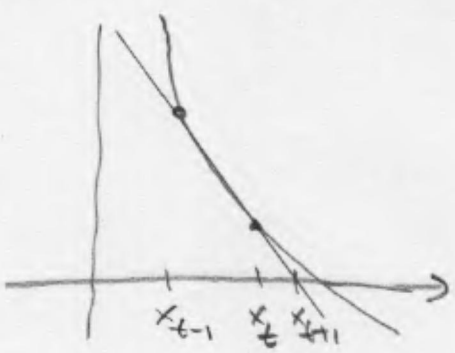
$$x_{t+1} = x_t - p \frac{f(x_t)}{f'(x_t)}$$

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

$$s(x) \overset{!}{=} 0 \quad \Leftrightarrow$$

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

$$|x_t - z| \leq \frac{2m}{M} q^{\gamma_t}, \quad t \in \mathbb{N}$$

$\gamma_0 = \gamma_1 = 1$

$\gamma_{t+1} = \gamma_t + \gamma_{t-1}$ „Fibonacci zahlen"

Nur eine f-Auswertung pro Iteration notwendig

$\gamma_t \sim 0.723 \cdot (1{,}618)^t$: Konvergenzordnung 1,618 also zw. 1 u. 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. Insbesonder wollen wir auch $f: \mathbb{R}^n \to \mathbb{R}^n$ zulassen.

Antwort gibt der sog. „Banachsche Fixpunktsatz".

Satz 8.3 (Sukzessive Approximation)

Sei $G \subseteq \mathbb{R}^n$ eine nichtleere, abgeschlossene Punktmenge und $g: G \to G$ Lipschitz-stetig mit Konstante $q < 1$, d.h.

$$\| g(x) - g(y) \| \leq q \, \| x - y \|.$$

Hierbei ist $\|.\|$ 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 aposteriori und a priori Fehlerabschätzungen

$$\| x^{(t)} - z \| \leq \frac{q}{1-q} \| x^{(t)} - x^{(t-1)} \| \leq \frac{q^t}{1-q} \| x^{(1)} - x^{(0)} \|.$$

(Wir schreiben den Iterationsindex oben in Klammer damit bei Vektoren unten Platz für den Komponentenindex bleibt)

Beweis:

i) Da $g: G \to G$ ist $x^{(t)} = g(x^{(t-1)}) = g(g(x^{(t-2)})) = \ldots = \underbrace{g(\ldots g}_{t\text{ mal}}(x^{(0)})\ldots)$ wohldefiniert.

ii) Weiter ist

$$\|x^{(t+1)} - x^{(t)}\| = \|g(x^{(t)}) - g(x^{(t-1)})\| \leq q\,\|x^{(t)} - x^{(t-1)}\|$$

$$\vdots$$

$$\leq q^t\,\|x^{(1)} - x^{(0)}\|$$

iii) Zeige nun, dass die $x^{(t)}$ eine Cauchy-Folge bilden.
Sei $m \geq 1$ und $\varepsilon > 0$ gegeben. Es ist

$$\|x^{(t+m)} - x^{(t)}\| \leq \|\underbrace{x^{(t+m)} - x^{(t+m-1)}}_{} + \underbrace{x^{(t+m-1)} - x^{(t+m-2)}}_{} + \ldots + \underbrace{x^{(t+1)} - x^{(t)}}_{}\|$$

Dreiecksungl. $\to \leq \|x^{(t+m)} - x^{(t+m-1)}\| + \|x^{(t+m-1)} - x^{(t+m-2)}\| + \ldots + \|x^{(t+1)} - x^{(t)}\|$

(ii) $\to \leq q^{t+m-1}\|x^{(1)} - x^{(0)}\| + q^{t+m-2}\|x^{(1)} - x^{(0)}\| + \ldots + q^t\|x^{(1)} - x^{(0)}\|$

Ausklammern $\to = \left(q^{t+m-1} + q^{t+m-2} + \ldots + q^t\right)\|x^{(1)} - x^{(0)}\|$

geom. Reihe $\to \leq q^t\,\frac{1-q^m}{1-q}\,\|x^{(1)} - x^{(0)}\| \leq \varepsilon$ für $t \geq t(\varepsilon)$ hinreichend groß.

$\mathbb{R}^n$ 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:

$$\|z - g(z)\| = \|z - x^{(t)} + x^{(t)} - g(z)\|$$

$$\overset{t\text{ bel.}}{\leq} \underbrace{\|z - x^{(t)}\|}_{\to 0 \text{ für } t\to\infty} + q\,\underbrace{\|x^{(t-1)} - z\|}_{\to 0 \text{ für } t\to\infty} \to 0$$

iv) Fehlerabschätzung

$$\|x^{(t+m)} - x^{(t)}\| \leq \underbrace{\|x^{(t+m)} - x^{(t+m-1)}\| + \ldots + \|x^{(t+1)} - x^{(t)}\|}_{\downarrow m\text{ mal}} \quad \text{(wie oben)}$$

$$\leq q^m\,\|x^{(t)} - x^{(t-1)}\| + \ldots + q\,\|x^{(t)} - x^{(t-1)}\|$$

$$= \underbrace{(q^m + q^{m-1} + \ldots + q)}_{\text{abschätzen durch geom. Reihe}}\,\|x^{(t)} - x^{(t-1)}\| \leq \frac{q}{1-q}\,\|x^{(t)} - x^{(t-1)}\|$$

Für $m \to \infty$ gilt $x^{(t+m)} \to z$, rechte Seite ist unabh. von m, also

$$\|z - x^{(t)}\| \leq \underbrace{\frac{q}{1-q}}_{\substack{\text{kann man benutzen} \\ \text{um Abstand zur exakten} \\ \text{Lösung zu schätzen}}}\,\|x^{(t)} - x^{(t-1)}\| \leq \underbrace{\frac{q}{1-q}\,q^{t-1}}_{= \frac{qt}{1-q}}\,\|x^{(1)} - x^{(0)}\|.$$

∎

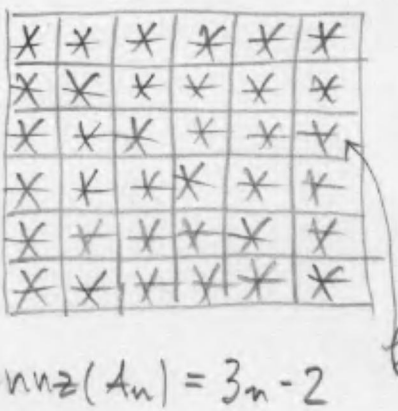
8.3 Iterationsverfahren zur Lösung linearer Gleichungssysteme

Wir kehren zurück zur Lösung von linearen Gleichungssystemen

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

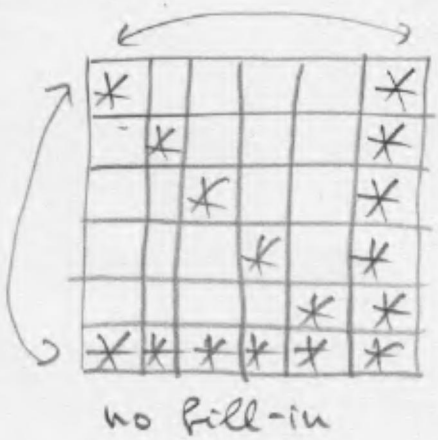
Definition 8.4 Eine Menge von Matrizen $\{A^{(n)} \mid n \in \mathbb{N}\}$ heißt dünn besetzt falls $\left|\{ a_{ij}^{(n)} \mid a_{ij}^{(n)} \neq 0 \}\right| = \operatorname{nnz}(A^{(n)}) = O(n)$. ▨

Gauß-Elimination ist für dünn besetzte Matrizen oft schlecht geeignet aufgrund von fill-in.



$\operatorname{nnz}(A_n) = 3n - 2$

fill-in Minimierung durch Umordnen



no fill-in

Lösen von $Ax = b$ $\Leftrightarrow$ „Nullstellensuche" $f(x) = b - Ax = 0$.

Definiere Iteration

$$\begin{aligned} x^{(t+1)} = g(x^{(t)}) &= x^{(t)} + C^{-1} f(x^{(t)}) \\ &= x^{(t)} + C^{-1}(b - Ax^{(t)}) \\ &= \underbrace{(I - C^{-1}A)}_{=: B \text{ „Iterationsmatrix"}} x^{(t)} + C^{-1}b \end{aligned}$$

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

$$\|g(x)-g(y)\| = \|Bx + C^{-1}b - By - C^{-1}b\| = \|B(x-y)\|$$
$$\leq \|B\|\,\|x-y\|$$

Falls $\|B\| < 1$ ($\|.\|$ verträgliche Matrixnorm) ist g Kontraktion auf $\mathbb{R}^n$.

Beispiele für Iterationsverfahren.

Setze $A = L + D + U$

L strikte untere Dreiecksmatrix
D Diagonalmatrix
U obere Dreiecksmatrix

$C = D$, also $x^{(t+1)} = x^{(t)} + D^{-1}(b - Ax^{(t)})$ „Jacobi-Verfahren"

$C = L+D$, also $x^{(t+1)} = x^{(t)} + (L+D)^{-1}(b - Ax^{(t)})$ „Gauß-Seidel Verf."

Iterationsverfahren konvergieren in der Regel nur für bestimmte Klassen von Matrizen. Hier ein Beispiel.

Definition 8.5 Eine Matrix heißt **strikt diagonaldominant** falls

$$\sum_{j \neq i} |a_{ij}| < |a_{ii}| \quad \forall i = 1, \dots, n.$$

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).

$$B = I - D^{-1}A = I - D^{-1}(L+D+U) = -D^{-1}(L+U)$$

$$\|B\|_\infty = \|D^{-1}(L+U)\|_\infty = \max_{i=1,\dots,n} \left(\sum_{j \neq i} \left| \frac{a_{ij}}{a_{ii}} \right| \right) = \max_{i=1,\dots,n} \frac{1}{|a_{ii}|} \underbrace{\sum_{j \neq i} |a_{ij}|}_{< |a_{ii}| \text{ n. Vor.}} < 1.$$

∎

Es gibt viele weitere solche Aussagen für sym. pos. def. Matrizen, schwach diagonaldom. Matrizen, M-Matrizen, ...

Aufwand für Iterationsverfahren

1) Aufwand für eine Iteration $x^{(t+1)} = x^{(t)} + C^{-1}(b - Ax^{(t)})$ sei $\alpha(n)$. Typischerweise $\alpha(n) = O(n)$.

2) $\| x^{(t)} - x \| \leq \| B \| \, \| x^{(t-1)} - x \|$

also $\| x^{(t)} - x \| \leq \| B \|^t \, \| x^{(0)} - x \|$

brauche $\| B \|^t \leq \varepsilon \iff t \underbrace{\log \underbrace{\| B \|}_{<1}}_{<0} \leq \underbrace{\log \varepsilon}_{<1} \iff t \geq \underbrace{\frac{\log \varepsilon}{\log \| B \|}}_{>0}$

Gesamtaufwand: $t \cdot \alpha(n) = \frac{\log \varepsilon}{\log \| B \|} \alpha(n)$

$\| B \|$ problemabhängig, je nach Verfahren auch von n abhängig.

Es gibt Verfahren, die relevante Probleme (z.B. Rdvmatrix) im Gesamtaufwand $O(n)$ lösen können!