1. Lineare Gleichungssysteme – Direkte Löser I

Notation Indexmenge $I \subset \mathbb{N}_0$ (endlich)

Vektoren: $x = (x_i)_{i \in I}$ oder $(x_i)_{1 \le i \le n}$ oder $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$

Matrizen: $A = (a_{ij})_{i \in I^1, j \in I^2}$ oder $(a_{ij})_{\substack{1 \le i \le n \\ 1 \le j \le m}}$

oder $\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ \vdots & & & \vdots \\ a_{n1} & & \cdots & a_{nm} \end{pmatrix}$

Räume von Vektoren, Matrizen: $\mathbb{K} = \mathbb{R}$ oder $\mathbb{C}$

$\mathbb{K}^n$, $\mathbb{K}^I$ bzw. $\mathbb{K}^{n,m}$ oder $\mathbb{K}^{I^1, I^2}$

Kroneckersymbol $\delta_{ij} = \begin{cases} 1 & ; i = j \\ 0 & ; \text{sonst} \end{cases}$

Einheitsmatrix $\mathbb{1} = (\delta_{ij})_{i,j \in I}$

1.1 Betrag in $\mathbb{K}$

$\|x\|$ Vektornorm, $x \in \mathbb{K}^I$

$$\|x\|_2 = \sqrt{\sum_{i \in I} x_i^2} \qquad \|x\|_\infty = \max_{i \in I} |x_i|$$

$|||A|||$ Operatornorm von $A \in \mathbb{K}^{I^1, I^2}$

$$|||A||| = \sup_{0 \neq x \in \mathbb{K}^{I^2}} \frac{\|Ax\|}{\|x\|} \qquad \left(|||A|||_2, |||A|||_\infty \right.$$

zu Norm $\|.\|_2$ bzw $\|.\|_\infty$)

Skalarprodukt auf $\mathbb{K}^I$: $x \cdot y$; $x, y \in \mathbb{K}^I$

quadratische Form $y \cdot Ax$ (andere Schreibweise $y^T A x$)

$$\left(y \cdot A = \left(\sum_{i \in I} y_i a_{ij} \right)_{j \in I} \in \mathbb{K}^{1,I} \right.$$

Zeilenvektor)

d.h. $y \cdot (Ax) = (y \cdot A)\, x$

Skalarprodukt Zeile mal Spalte (Matrizenprodukt)

<u>Landau - Symbole</u>:

$(\,.\,) = O(h^m) \Leftrightarrow |(\,.\,)| \leq C h^m$ für eine Konstante $C > 0$

$(\,.\,) = O(n^\alpha) \Leftrightarrow (\,.\,) \leq C(1 + n^\alpha)$ für $C > 0$

<u>Betrachte</u>: $A \in \mathbb{R}^{n,n}$, $b \in \mathbb{R}^n$ ($\mathbb{C}$ analog)

gesucht $x \in \mathbb{R}^n$ mit $Ax = b$

1.0.1. <u>Satz</u> Falls $\det A \neq 0$, dann existiert genau ein $x \in \mathbb{R}^n$ mit $Ax = b$

<u>Bew</u>: Lineare Algebra

$x = A^{-1} b$ für $A = (a_{ij})_{i,j=1,\dots,n}$

$$\det A = \sum_{\sigma \in S_n} \operatorname{sgn} \sigma \; a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{n\sigma(n)}$$

wobei S_n die Menge aller Permutationen von $\{1,\dots,n\}$

und $\operatorname{sgn} \sigma = (-1)^m$ falls $\sigma = \sigma_1 \circ \sigma_2 \cdots \sigma_m$

mit σ_i Vertauschung zweier Indizes

<u>Cramersche Regel</u> $A^{-1} = \dfrac{1}{\det A} \left((-1)^{i+j} \det A^{ji} \right)$

falls $A^{kl} = (a_{ij})_{i,j=1,\dots,n}$, $i \neq k$, $j \neq l$

(Streichungsmatrix)

$\rightsquigarrow$ x ist berechenbar, aber so <u>nicht</u> effizient

1.1 Gleichungssysteme mit Dreiecksmatrizen

$R \in \mathbb{R}^{n,n}$ $\quad R = (r_{ij})_{i,j=1,\dots,n}$ mit $r_{ij} = 0 \quad \forall_{j<i}$

obere Dreiecksmatrix

$$Rx = b \Leftrightarrow \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

$$\Rightarrow \det R = \prod_{i=1}^{n} r_{ii}$$

1.1.1. Lemma $\quad r_{ii} \neq 0 \ \forall_{i=1,\dots,n} \Leftrightarrow R$ ist invertierbar

Bew (✓)

$\boxed{\text{Rückwärtseinsetzen}}$

$$x_n = \frac{b_n}{r_{nn}}, \quad x_{n-1} = \frac{1}{r_{n-1,n-1}} \left(b_{n-1} - r_{n-1,n} x_n \right)$$

$$\dots \quad x_i = \frac{1}{r_{ii}} \left(b_i - \sum_{j=i+1}^{n} r_{ij} x_j \right)$$

Aufwand: Wir fragen nach Multiplikationen und Divisionen

1.1.2. Bem. für die $i$te Zeile: $(n-i)$ Multiplikationen, 1 Division

insgesamt: $\displaystyle\sum_{i=1}^{n} (n-i) = \sum_{i=0}^{n-1} i = \frac{n(n-1)}{2} = \frac{n^2}{2} + O(n)$

Multiplikationen und n Divisionen

1.1.3. Algorithmus

```
x_n = b_n / r_nn;
for (i = n-1; i ≥ 1; i--) {
    h = b_i;
    for (j = i+1; j ≤ n; j++)
        h = h - r_ij * x_j;
    x_i = h / r_ii;
}
```

Für Gleichungssysteme mit unterer Dreiecksmatrix

$$Lx = b \quad \text{mit} \quad L = \begin{pmatrix} l_{11} & & 0 \\ \vdots & \ddots & \\ l_{n1} & \cdots & l_{nn} \end{pmatrix}$$

verfahren wir analog mittels Vorwärtseinsetzen!

1.2. Die Gauß-Elimination als LR-Zerlegung

Betrachten wir nun eine allgemeine Matrix $A \in \mathbb{R}^{n,n}$ und das Gleichungssystem $Ax = b$ für $b \in \mathbb{R}^n$

Ziel: Transformation in eine obere Dreiecksmatrix durch spaltenweises Eliminieren:

$$\begin{aligned} a_{11} x_1 + a_{12} x_2 + \cdots a_{1n} x_n &= b_1 \\ a_{21} x_2 + a_{22} x_2 + \cdots a_{2n} x_n &= b_2 \\ &\vdots \\ a_{n1} x_1 + a_{n2} x_2 + \cdots a_{nn} x_n &= b_n \end{aligned}$$

$\boxed{\text{1. Schritt}}$ neue ik Zeile = alte $i$te Zeile $- l_{i1}$ alte erste Zeile

mit $l_{i1} = \dfrac{a_{i1}}{a_{11}}$ $\quad i = 2, \ldots n$

Annahme: $a_{11} \neq 0$

$$\rightsquigarrow \underbrace{(a_{i1} - l_{i1} a_{11})}_{=0} x_1 + \underbrace{(a_{i2} - l_{i1} a_{12})}_{a_{i2}^{(2)}} x_2 + \cdots + \underbrace{(a_{1n} - l_{i1} a_{1n})}_{a_{in}^{(2)}} x_n = \underbrace{(b_i - l_{i1} b_1)}_{b_i^{(2)}}$$

nun verbleibt das Gleichungssystem:

$$\begin{aligned} a_{11} x_1 + a_{12} x_2 \cdots + a_{1n} x_n &= b_1 \\ a_{22}^{(2)} x_2 \cdots + a_{2n}^{(2)} x_n &= b_2^{(2)} \\ &\vdots \\ a_{n2}^{(2)} x_2 \cdots + a_{nn}^{(2)} x_n &= b_n^{(2)} \end{aligned}$$

Eliminiere nun $\begin{pmatrix} a_{32}^{(2)} \\ \vdots \\ a_{n2}^{(2)} \end{pmatrix}$, usw.

$$A = A^{(1)} \longrightarrow A^{(2)} \longrightarrow A^{(3)} \longrightarrow \text{---} \longrightarrow A^{(n)} = R$$

obere Dreiecksmatrix

wobei

$$A^{(k)} = \begin{pmatrix} a_{11}^{(1)} & \text{---} & a_{1n}^{(1)} \\ & a_{22}^{(2)} \ \text{---} & a_{2n}^{(2)} \\ & \ddots & \\ & a_{kk}^{(k)} \ \text{---} & a_{kn}^{(k)} \\ 0 & \vdots & \vdots \\ & a_{nk}^{(k)} \ \text{---} & a_{nn}^{(k)} \end{pmatrix}$$

<u>im $k$ten Schritt:</u>

$$l_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \quad ; \quad i = k+1, \text{---} n$$

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - l_{ik}\, a_{kj}^{(k)} \quad ; \quad i,j = k+1, \text{---} n$$

$$b_i^{(k+1)} = b_i^{(k)} - l_{ik}\, b_k^{(k)} \quad ; \quad i = k+1, \text{---} n$$

<u>in Matrix-Notation:</u>

$$L_k = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & -l_{k+1,k} & \ddots \\ & & \vdots & \\ & & -l_{n,k} & & 1 \end{pmatrix}$$

<u>Frobeniusmatrix</u>

jeder Eliminationsschritt ist eine lineare Operation:

$$A^{(k+1)} = L_k A^{(k)} \qquad b^{(k+1)} = L_k b^{(k)}$$

1.2.1. <u>Lemma</u> (i) $L_k^{-1} = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & l_{k+1,k} & \ddots \\ & & \vdots & \\ & & l_{n,k} & & 1 \end{pmatrix}$

(ii) $L := L_1^{-1} L_2^{-1} \text{---} L_{n-1}^{-1} = \begin{pmatrix} 1 & & & 0 \\ l_{2,1} & \ddots & & \\ \vdots & & \ddots & \\ l_{n,1} & \text{---} & l_{n,n-1} & 1 \end{pmatrix}$

<u>Bew</u> (i)

$$\begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & l_{k+1,k} & \ddots & \\ & & \vdots & & \\ & & l_{n,k} & & 1 \end{pmatrix} \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -l_{k+1,k} & \ddots & \\ & & \vdots & & \\ & & -l_{n,k} & & 1 \end{pmatrix} = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}$$

(ii)

$$\begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & l_{k,k-1} & 1 & \\ & & \vdots & & \ddots \\ & & l_{n,k-1} & & & 1 \end{pmatrix} \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & l_{k+1,k} & \ddots & \\ & & \vdots & & \ddots \\ & & l_{n,k} & \cdots & l_{n,n-1} & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & l_{k,k-1} & \ddots & \\ & & \vdots & & \ddots \\ & & l_{n,k-1} & \cdots & l_{n,n-1} & 1 \end{pmatrix}$$

$\rightsquigarrow$ Rest per Induktion $\square$

D.h. wir erhalten falls $a_{kk}^{(k)} \neq 0 \quad \forall_{k=1,\dots,n-1}$:

$$R = A^{(n)} = L_{n-1} \cdots L_2 L_1 A = L_{n-1} \cdots L_2 L_1 b$$

$$\Leftrightarrow \quad \underbrace{L_1^{-1} L_2^{-1} \cdots L_{n-1}^{-1}}_{= L} R = A \quad \Leftrightarrow \quad \boxed{LR = A}$$

<u>Gleichungssystem</u>: $Ax = b \Rightarrow LRx = b$

1.2.2. <u>Algorithmus</u>

1. Bilde LR-Zerlegung von A
2. Löse $Lz = b$ durch Vorwärtseinsetzen
3. Löse $Rx = z$ durch Rückwärtseinsetzen

Speicherung: L und R in der alten Matrix A

$$\begin{pmatrix} r_{11} & \cdots & r_{1n} \\ l_{21} & \ddots & \vdots \\ \vdots & & \\ l_{n1} & \cdots & l_{n,n-1} \; r_{nn} \end{pmatrix}$$

Gaußalgorithmus (ohne Pivotisierung)

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

1.2.3. <u>Bem</u> Der Aufwand der Gaußelimination beträgt $\frac{n^3}{3}$ Multiplikationen

<u>Bew</u> $\boxed{\text{Ü}}$ $$\sum_{k=1}^{n-1} (n-k)^2 = \sum_{k=1}^{n-1} k^2 = \frac{(n-1)\,n\,(2n-1)}{6} = \frac{n^3}{3} + O(n^2)$$

$\square$

1.3. Pivotisierung

1.3.1. Beispiel

(i) $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ $\det A = -1 \Rightarrow$ invertierbar

$a_{11} = 0 \Rightarrow$ LR-Zerlegung nicht durchführbar

<u>Ausweg:</u> Vertauschung der Zeilen von A und der Einträge von b!

(ii) $A = \begin{pmatrix} 10^{-m} & 1 \\ 1 & 1 \end{pmatrix}$ $b = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$

$$A^{(2)} = \begin{pmatrix} 10^{-m} & 1 \\ 0 & 1-10^m \end{pmatrix} \quad b^{(2)} = \begin{pmatrix} 1 \\ 2-10^m \end{pmatrix}$$

<u>Rückwärtseinsetzen:</u> $x_2 = \frac{10^m - 2}{10^m - 1} \approx 1$

$$x_1 = 2 - x_2 = \frac{10^m}{10^m - 1} \approx 1$$

bei weniger als m Nachkommastellen in der Arithmetik:

$$\tilde{A}^{(2)} = \begin{pmatrix} 10^{-m} & 1 \\ 0 & -10^{m} \end{pmatrix}, \quad \tilde{b}^{(2)} = \begin{pmatrix} 1 \\ -10^{m} \end{pmatrix} \Rightarrow \begin{array}{l} \tilde{x}_2 = 1 \\ \tilde{x}_1 = 0 \end{array}$$

Problem: Auslöschung!

Aussage: Vertausche Zeilen

$$\bar{A} = \begin{pmatrix} 1.0 & 1.0 \\ 10^{-m} & 1.0 \end{pmatrix}, \quad \bar{b} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

$$\bar{A}^{(2)} = \begin{pmatrix} 1.0 & 1.0 \\ 0 & 1.0 - 10^{-m} \end{pmatrix} \quad \bar{b}^{(2)} = \begin{pmatrix} 2 \\ 1 - 2\,10^{-m} \end{pmatrix}$$

$$\tilde{\bar{A}}^{(2)} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \tilde{\bar{b}}^{(2)} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \Rightarrow \begin{array}{l} \tilde{\bar{x}}_1 = 1 \\ \tilde{\bar{x}}_2 = 1 \end{array}$$

Generelle Strategie:
- Suche stets das maximale $|a_{ik}^{(k)}|$ für jedes k
- Vertausche dann die $k$te und die $i$te Zeile

1.3.2. Algorithmus (Gauß-Elimination mit Zeilen-Pivotierung)

Eliminationsschritt $A^{(k)} \longrightarrow A^{(k+1)}$:

1. Wähle $p \in \{k, \dots, n\}$, so dass $|a_{pk}^{(k)}| = \max\limits_{j \in \{k, \dots, n\}} |a_{jk}^{(k)}|$
2. definiere $\tilde{A}^{(k)} = (\tilde{a}_{ij}^{(k)})_{ij}$ mit $\tilde{a}_{ij}^{(k)} = \begin{cases} a_{kj}^{(k)} & ; i = p \\ a_{pj}^{(k)} & ; i = k \\ a_{ij}^{(k)} & ; \text{sonst} \end{cases}$
3. eliminiere wie bisher

$$\tilde{A}^{(k)} \longrightarrow A^{(k+1)}$$

1.3.3. <u>Bem</u> • $|l_{ik}| = \dfrac{|\tilde{a}_{ik}^{(k)}|}{|\tilde{a}_{kk}^{(k)}|} \leq 1 \longrightarrow$ wichtig bei der Fehleranalyse später!

• statt der Zeilen kann man auch Spalten vertauschen

• Das Vertauschen von Zeilen lässt sich auch als Multiplikation mit einer Permutationsmatrix schreiben: $e_i = (\delta_{ij})_{j=1,\dots,n}$ i-ter Einheitsvektor

σ Permutation $\in S_n$

$$P_\sigma = \begin{pmatrix} | & & | \\ e_{\sigma(1)} & \cdots & e_{\sigma(n)} \\ | & & | \end{pmatrix} \text{ Permutationsmatrix}$$

mit $P_\sigma e_i = e_{\sigma(i)}$

1.3.4. <u>Lemma</u> Sei $\sigma \in S_n$, dann gilt $P_\sigma \in O(n)$ (orthog. Gruppe), d.h. $P_\sigma^{-1} = P_\sigma^T$

<u>Bew:</u> z.z. $P_\sigma x \cdot y \overset{!}{=} x \cdot P_\sigma^{-1} y \iff$

$$\underbrace{P_\sigma e_i \cdot e_j}_{e_{\sigma(i)} \cdot e_j = \delta_{\sigma(i)j} = \delta_{i\sigma^{-1}(j)} = e_i \cdot \underbrace{P_{\sigma^{-1}}}_{P_\sigma^{-1}} e_j} \overset{!}{=} e_i \cdot P_\sigma^{-1} e_j \quad = \quad \square$$

1.3.5. <u>Satz</u> (Existenz einer LR-Zerlegung mit Zeilenpivotisierung)
Für jede invertierbare Matrix A existiert eine Permutationsmatrix P, so dass PA eine LR-Zerlegung besitzt.
P kann so gewählt werden, dass $\max_{i,j} |L_{ij}| \leq 1$

<u>Bew</u> $\det A \neq 0 \Rightarrow \exists_i\, a_{i1}^{(1)} \neq 0$;
wähle i so dass $|a_{i1}^{(1)}|$ maximal

$$\Rightarrow \quad A^{(2)} = L_1 A^{(1)} = L_1 P_{\sigma_1} A = \left(\begin{array}{c|ccc} a_{11}^{(1)} & * & \text{---} & * \\ \hline 0 & & B^{(2)} & \end{array}\right)$$

Vertauschung zweier Indizes oder $\mathbb{1}$

$$\det A^{(2)} = a_{11}^{(1)} \det B^{(2)}$$
$$= \underbrace{\det L_1}_{=1} \underbrace{\operatorname{sgn} \sigma_1}_{\in \{-1,1\}} \underbrace{\det A}_{\neq 0}$$

$\Rightarrow B^{(2)}$ invertierbar

$\rightsquigarrow$ weiter mit Induktion ...

$$\boxed{R = A^{(n)} = L_{n-1} P_{\sigma_{n-1}} L_{n-2} P_{\sigma_{n-2}} \text{———} L_1 P_{\sigma_1} A} \quad (*)$$

Falls $\pi \in S_n$ nur Zeilen i mit $i \geq k+1$ vertauscht, dann gilt

$$P_\pi L_k P_\pi^{-1} = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -l_{\pi(k+1),k} & \ddots & \\ & & \vdots & & \\ & & -l_{\pi(n),k} & & 1 \end{pmatrix} =: \hat{L}_k$$

$$\uparrow$$

$$\begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & \boxed{\tilde{P}} \end{pmatrix}$$

$\rightarrow$ Einfügen von $\mathbb{1} = P_{\sigma_k}^{-1} P_{\sigma_k}$ liefert:

$$R = L_{n-1} P_{\sigma_{n-1}} L_{n-2} P_{\sigma_{n-1}}^{-1}$$
$$P_{\sigma_{n-1}} P_{\sigma_{n-2}} L_{n-3} P_{\sigma_{n-2}}^{-1} P_{\sigma_{n-1}}^{-1}$$
$$P_{\sigma_{n-1}} P_{\sigma_{n-2}} P_{\sigma_{n-3}} L_{n-4} P_{\sigma_{n-3}}^{-1} P_{\sigma_{n-1}}^{-1} P_{\sigma_{n-1}}^{-1} \text{———}$$
$$(P_{\sigma_{n-1}} \text{———} P_{\sigma_2}) L_1 (P_{\sigma_2}^{-1} \text{———} P_{\sigma_{n-1}}^{-1}) P_{\sigma_{n-1}} P_{\sigma_{n-2}} \text{———} P_{\sigma_2} P_{\sigma_1} A$$
$$= \hat{L}_{n-1} \hat{L}_{n-2} \hat{L}_{n-3} \hat{L}_{n-4} \text{———} \hat{L}_1 P_{\pi_0} A$$

mit $\hat{L}_k = P_{\pi_k} L_k P_{\pi_k}^{-1}$ $\qquad \pi_{n-1} = \mathbb{1}$

$\pi_k = \sigma_{n-1} \cdots \sigma_{k+1}$

für $k \in \{0, \ldots n-2\}$

$$\Rightarrow \boxed{P_{\pi_0} A = LR} \quad \text{mit } P_{\pi_0} = P_{\sigma_{n-1}} \cdots P_{\sigma_1}$$

$$\text{und } L = \begin{pmatrix} 1 & & & \\ \ell_{\pi_1(2),1} & 1 & & 0 \\ \vdots & \ell_{\pi_2(3),2} & \ddots & \\ \ell_{\pi_1(n),1} & \ell_{\pi_2(n),2} & \cdots & \ell_{\pi_{n-1}(n),n-1} \; 1 \end{pmatrix}$$

$$\max_{i,j} \left| \ell_{\pi_i(j),i} \right| \leq 1$$

$\square$

1.3.5. <u>Bem</u> • <u>Im Algorithmus</u>:

Vertauschungen in A auch in L (ganze Zeilen!) ausführen!

Beim Lösen: $LRx = P_{\pi_0} A x = P_{\pi_0} b$

Beim Speichern: neben L, R in der alten Matrix ein Vektor für P_{π_0}

• $\det A = \operatorname{sgn} \pi_0 \left(\prod_{i=1}^{n} r_{ii} \right)$

$\boxed{\text{folgt aus } P_{\pi_0} A = LR}$

1.3.6. <u>Bem</u> Aus der LR-Zerlegung $P_{\pi_0} A = LR$ folgt

$$\det A = \operatorname{sgn} \pi_0 \prod_{i=1}^{n} r_{ii}$$

Die Lösung kann trotz Pivotierung rundungsfehlerbehaftet sein!

<u>Abhilfe</u>: Korrekturgleichung lösen!

1.3.7. <u>Bem</u> (Iterative Verbesserung)

x_0 ungenaue Lösung zu $Ax = b$

$r(y) := b - Ay = A(x-y)$ (Residuum)

$e(x_0) = x - x_0$ Fehler

$A e(x_0) = r(x_0)$ Korrekturgleichung

Löse $Ae = r(x_0)$ $\rightsquigarrow$ approximative Lösung e_0

$x_1 = x_0 + e_0$

$\rightsquigarrow$ <u>Iteration</u>:
- $r_k = b - Ax_k$;
- Löse $Ae_k = r_k$;
- $x_{k+1} = x_k + e_k$;

1.4. Cholesky-Verfahren

A pos. def. $\left(x \cdot Ax \underset{\forall x \neq 0}{>} 0 \right)$, A symmetric

<u>Ziel</u>: $A = LR \underset{R = L^T}{=} LL^T$

1.4.1. <u>Lemma</u> A symm., pos. def. Matrix, dann ist

A invertierbar, $a_{ii} > 0 \ \forall i = 1, \dots, n$, $\max_{i,j} |a_{ij}| = \max_i a_{ii}$,

die $A^{(k)}$ des Gauß-Verfahrens ohne Pivotierung sind ebenfalls symm. und pos. def.

<u>Bem</u>: wegen $a_{ii} > 0$ ist keine Pivotierung notwendig!

<u>Bew</u> (i) $(Ax=0 \Rightarrow Ax\cdot x=0 \Rightarrow x=0)$

$\Rightarrow A$ invertierbar

(ii) $Ae_i\cdot e_i = a_{ij} > 0$

(iii) $A(e_i \pm e_j)\cdot(e_i \pm e_j) = a_{ii} \pm \underbrace{(a_{ij}+a_{ji})}_{2a_{ij}} + a_{jj} > 0$

$\Rightarrow \mp a_{ij} \le \frac{1}{2}(a_{ii}+a_{jj})$

$\le \max_i a_{ii}$

(iv) $A = A^{(1)} = \left(\begin{array}{c|c} a_{11} & -z^T \\ \hline z & B^{(1)} \end{array}\right)$

$\Rightarrow A^{(2)} = L_1 A^{(1)} = \left(\begin{array}{c|c} a_{11} & -z^T- \\ \hline 0 & B^{(2)} \end{array}\right)$ $\quad L_1 = \begin{pmatrix} 1 & & \\ -l_{21} & \ddots & \\ \vdots & & \\ -l_{n1} & & 1 \end{pmatrix}$

$L_1 A^{(1)} L_1^T = \left(\begin{array}{c|c} a_{11} & -0- \\ \hline 0 & B^{(2)} \end{array}\right)$

$\boxed{\text{die } -l_{j1} \text{ treffen auf Nullen!}}$

nun betrachte $\tilde{x} \in \operatorname{span}\{e_2, \dots e_n\}$

$\tilde{x}\cdot(L_1 A^{(1)} L_1^T)\tilde{x} = (L_1^T\tilde{x})\cdot A^{(1)}(L_1^T\tilde{x}) > 0 \quad \forall \tilde{x} \Rightarrow B^{(2)}$ pos. def.

L_1^T invertierbar (Basiswechsel)

$(L_1 A^{(1)} L_1^T)^T = L_1 A^{(1)} L_1^T \Rightarrow B^{(2)}$ symm.

<u>Rest:</u> Induktion

$\square$

Aus dem Beweis sehen wir:

$$L_{n-1} \cdots L_1 A L_1^T \cdots L_{n-1}^T = D = \operatorname{diag}\left(\left(a_{kk}^{(k)}\right)_{k=1,\dots,n}\right)$$

$$\Rightarrow \quad A = \underbrace{L_1^{-1} \cdots L_{n-1}^{-1}}_{=L} D \underbrace{\left(L_1^{-1} \cdots L_{n-1}^{-1}\right)^T}_{=L^T} = L D L^T$$

Da D pos. def. existiert $D^{\frac{1}{2}} = \operatorname{diag}\left(\left(\sqrt{a_{kk}^{(k)}}\right)_{k=1,\dots,n}\right)$

$$\hat{L} = L D^{\frac{1}{2}} \quad \Rightarrow \quad \hat{L}\hat{L}^T = L D^{\frac{1}{2}} (L D^{\frac{1}{2}})^T$$

$$= L D^{\frac{1}{2}} D^{\frac{1}{2}} L^T = L D L^T = A \quad \Rightarrow$$

1.4.2. <u>Satz</u> A symm., pos. def., dann gibt es eine Zerlegung $A = \hat{L}\hat{L}^T$, wobei $\hat{L}$ eine untere Dreiecksmatrix.

zur Berechnung:

$$\begin{pmatrix} l_{11} & & \\ \vdots & \ddots & \\ l_{n1} & \cdots & l_{nn} \end{pmatrix} \begin{pmatrix} l_{11} & \cdots & l_{n1} \\ & \ddots & \vdots \\ & & l_{nn} \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$$

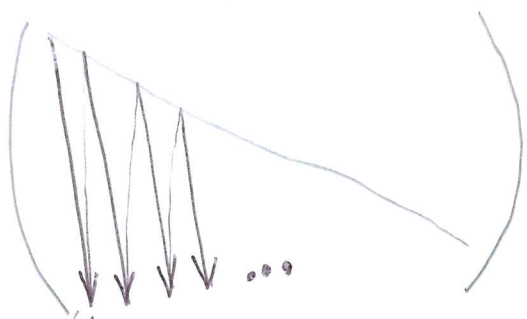
$$\Rightarrow \quad \boxed{a_{kk} = l_{k1}^2 + \cdots l_{k,k-1}^2 + l_{kk}^2 \quad (1)}$$ auflösen!

$$\boxed{(i>k) \quad a_{ik} = l_{i1} l_{k1} + \cdots l_{i,k-1} l_{k,k-1} + l_{ik} l_{kk} \quad (2)}$$ auflösen

$$(1) \Rightarrow \quad l_{kk} = \sqrt{a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2}$$

$$(2) \Rightarrow \quad l_{ik} = \frac{a_{ik} - \sum_{j=1}^{k-1} l_{ij} l_{kj}}{l_{kk}}$$

Schema zur Berechnung der Einträge von $\hat{L}$:



beim Berechnen von l_{ik} sind alle l_{ij}, l_{kj} schon bekannt!

1.4.3. Algorithmus

for $(k=1;\ k \le n;\ k{+}{+})\{$

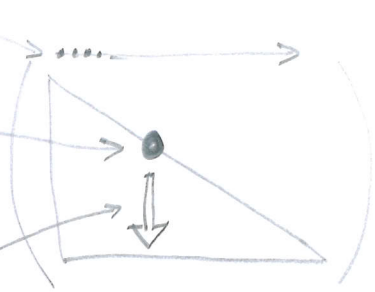
$$l_{kk} = \sqrt{a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2}\ ;$$

if $(k<n)$

for $(i=k+1;\ i \le n;\ i{+}{+})$

$$l_{ik} = \frac{a_{ik} - \sum_{j=1}^{k-1} l_{ij}\, l_{kj}}{l_{kk}}\ ;$$

$\}$



1.4.4. Lemma Der Aufwand der Cholesky Zerlegung ist $\frac{1}{6}n^3 + O(n^2)$ Multiplikationen und n Wurzeln

Bew

n Wurzeln $(\checkmark)$

$$\sum_{k=1}^{n} \sum_{i=k+1}^{n} \sum_{j=1}^{k-1} 1 = \sum_{k=1}^{n} (n-k)(k-1)$$

$$= \sum_{k=1}^{n} (nk - k^2) + O(n^2) = n\frac{n^2}{2} - \sum_{k=1}^{n} k^2 + O(n^2)$$

$$= \frac{n^3}{2} - \frac{n^3}{3} + O(n^2) = \frac{n^3}{6} + O(n^2) \qquad \square$$

2. Fehleranalyse

Fehlerarten:
- Fehler in den Daten (Rundungsfehler, Meßfehler)
- Fehler im Algorithmus (Rundung, Auslöschung)

Fragen: Wie reagiert die Lösung des Problems auf Datenfehler?

$\longrightarrow$ $\boxed{\text{Kondition des Problems}}$

Wie verhält sich die Lösung des Algorithmus bezüglich auftretender Rundungsfehler?

$\longrightarrow$ $\boxed{\text{Stabilität des Algorithmus}}$

2.1. Darstellung von Zahlen im Rechner

z darstellbare Zahl im Rechner:

$$z = a\, d^{e} \qquad d = 2^{m} \quad (d = 2, 8, 16 \text{ fest})$$

$$e \in \mathbb{Z}\,;\ e_{min} \le e \le e_{max} \qquad \underline{\text{Exponent}}$$

$$a = s \sum_{i=1}^{\ell} a_i\, d^{-i} \quad (\ell \text{ fest}) \qquad \underline{\text{Mantisse}}$$

$$a_i \in \{0, \dots, d-1\},\ a_1 \neq 0 \vee a = 0 \qquad \underline{\text{Ziffern}}$$

$$s \in \{-1, 1\} \qquad \underline{\text{Vorzeichen}}$$

Menge der darstellbaren Zahlen:

$$F = \left\{ x \in \mathbb{R} \;\middle|\; \exists\, a = s \sum_{i=1}^{\ell} a_i\, d^{-i},\ \exists\, e \in \{e_{min}, \dots, e_{max}\} \text{ mit } x = a\, d^{e} \right\}$$

relative Genauigkeit: $x \in \mathbb{R}$ mit

$$\underbrace{d^{e_{min}-1}}_{\substack{\text{kleinste positive}\\ \text{darstellbare Zahl}}} \le x \le \underbrace{(1-d^{\ell})\, d^{e_{max}}}_{\substack{\text{größte darstellbare}\\ \text{Zahl}}}$$

wird gerundet auf die nächste darstellbare Zahl $\hat{x} \in F$:

$$\frac{|x - \hat{x}|}{|x|} \le \epsilon := \frac{d^{1-\ell}}{2} \qquad \text{für } x \neq 0$$

Bsp $\underline{d=10,\ \ell=3}$: $x = 1.034 = 0{,}1034 \cdot 10^1 \sim 0{,}103 \cdot 10^1 = 1{,}03 = \hat{x}$

$\epsilon = \frac{0{,}01}{2} = \frac{d^{1-\ell}}{2}$ relative Rechengenauigkeit

2.1.1. Bem Für float in C ist $\epsilon = 2^{-24} = 6 \cdot 10^{-8}$

Für double in C ist $\epsilon = 2^{-53} = 1{,}1 \cdot 10^{-16}$

2.1.2. Bem
- Meßfehler der Daten >> Rundungsfehler
- Zur Realisierung der Grundrechenarten im Rechner

$\circ \in \{+, -, \cdot, /\}$ $\hat{\circ}$ Realisierung, $\hat{x}, \hat{y} \in F$

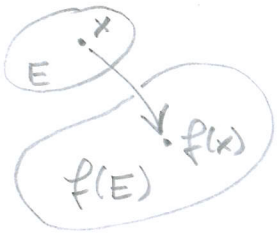
$\hat{x} \hat{\circ} \hat{y} = \hat{x} \circ \hat{y} (1 + \delta(\hat{x}, \hat{y}))$ mit $|\delta(\hat{x}, \hat{y})| \leq \epsilon$

Achtung $\hat{\circ}$ nicht assoziativ

$\left[d=10, \ell=3 : 1{,}0 = (0{,}1 \cdot 10^5 \hat{+} (-0{,}1 \cdot 10^5)) \hat{+} 1{,}0 \right.$
$\left. \neq 0{,}1 \cdot 10^5 \hat{+} ((-0{,}1 \cdot 10^5) \hat{+} 1{,}0) = 0 \right]$

2.2. Kondition von Problemen

2.2.1. Notation



$x \in \mathbb{R}^n$ Eingabedaten

$E \subset \mathbb{R}^n$ Eingabetoleranzbereich

$f: \mathcal{U} \subset \mathbb{R}^n \to \mathbb{R}^m$ Problemlösungsabbildung

$f(E)$ Ausgabebereich

Nur Rundungsfehler $E = \{ \hat{x} \in F \mid \|\hat{x} - x\| \leq \epsilon \|x\| \}$

bei Meßfehlern $E = \{ \hat{x} \in F \mid \|\hat{x} - x\| \leq \delta \}$

Bsp. absolute Genauigkeit

Bsp relative Genauigkeit

$\boxed{\text{Problem } (f, x)}$

Landau Symbol $f(x) = o(g(x))$ für $x \to x_0$

$:\Leftrightarrow \lim_{x \to x_0} \frac{\|f(x)\|}{\|g(x)\|} = 0$

zur Abkürzung: $f(x) \lesssim h(x) \;:\Leftrightarrow\; f(x) \leq h(x) + \mathcal{O}(h(x))$

$f(x) \approx h(x) \;:\Leftrightarrow\; f(x) = h(x) + \mathcal{O}(h(x))$

2.2.2. Def (Kondition)

Die absolute (relative) Kondition K_{abs} (K_{rel}) ist die kleinste Zahl, so dass

$$\| f(\hat{x}) - f(x) \| \lesssim K_{abs} \| \hat{x} - x \|$$

$$\left(\frac{\| f(\hat{x}) - f(x) \|}{\| f(x) \|} \lesssim K_{rel} \frac{\| \hat{x} - x \|}{\| x \|} \right) \quad \text{für } \hat{x} \to x.$$

Statt K_{rel} schreiben wir auch K! (∞ ist auch zugelassen)

2.2.3. Lemma Falls $\mathcal{U}$ offen, f diff.bar in x, dann gilt

$$K_{abs} = |||f'(x)||| \qquad K_{rel} = \frac{\|x\|}{\|f(x)\|} |||f'(x)|||$$

($|||.|||$ Operatornorm auf dem $\mathbb{R}^{n,m}$ zu $\|.\|$ Norm im $\mathbb{R}^n$ bzw $\mathbb{R}^m$)

Betrachten wir die Addition:

$f: \mathbb{R}^2 \to \mathbb{R} \qquad (a,b) \mapsto a+b$

$\|(a,b)\|_1 := |a| + |b| \qquad \|a\| := |a|$

$$|||f'(a,b)||| = \sup_{(\delta a, \delta b) \neq 0} \frac{|(1,1)\cdot(\delta a, \delta b)|}{|\delta a| + |\delta b|} = 1$$

$$K_{abs} = 1 \qquad K_{rel} = \frac{|a| + |b|}{|a+b|}$$

Addition bei gleichem Vorzeichen: $K_{rel} = 1$

Substraktion bei gleichem Vorzeichen:

$b = -(a+\delta) \quad \delta \ll 1 \qquad K_{rel} \gg 1 \qquad \sup_x K(x) = \infty$

(Auslöschung)

2.2.4. <u>Lemma</u> (Kondition lin. Gleichungssystem bzgl. rechten Seite)

$Ax=b$, b Input, x Output, $f: \mathbb{R}^n \to \mathbb{R}^n$ $f(b) = A^{-1}b$

$K_{abs} = |||A^{-1}|||$ $\quad K_{rel} = \frac{\|Ax\| \, |||A^{-1}|||}{\|x\|} \leq |||A||| \, |||A^{-1}|||$

$\|b\|$

<u>Bew</u> (v)

2.2.5. <u>Lemma</u> (Kondition lin. Gleichungssysteme bzgl. der Matrix)

$Ax=b$, A Input, x Output

$K_{abs} \leq |||A^{-1}||| \, \|x\|$

$K_{rel} \leq |||A||| \, |||A^{-1}|||$

<u>Bew</u>

$g: \mathbb{R}^{n,n} \to \mathbb{R}^{n,n}$; $A \mapsto A^{-1}$ ist stetig diffbar auf $GL(n)$

$g'(A)C = \partial_t \, g(A+tC)\big|_{t=0}$

$(A+tC)\,g(A+tC) = \mathbb{1} \Rightarrow \quad 0 = C\,g(A+tC) + (A+tC)\,g'(A+tC)C$

$\underset{t=0}{\Rightarrow} \quad 0 = CA^{-1} + A\,g'(A)C \Rightarrow \boxed{g'(A)C = -A^{-1}CA^{-1}}$

$f(A) = A^{-1}b \to f'(A)C = g'(A)Cb = -A^{-1}CA^{-1}b$

$= -A^{-1}Cx$

$\to K_{abs} = |||f'(A)||| = \sup_{\|C\|=1} \|A^{-1}Cx\| \leq |||A^{-1}||| \, \|x\|$

$K_{rel} \leq \frac{|||A||| \, |||f'(A)|||}{\|x\|} \leq |||A||| \, |||A^{-1}|||$

$\square$

2.2.6 <u>Def</u> (Konditionszahl einer Matrix)

$K(A) := \|A\| \, \|A^{-1}\|$ ($\|.\|$ nicht zwingend die Operatornorm)

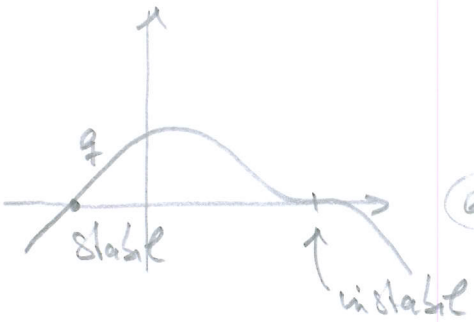
2.2.7. <u>Bsp</u> <u>geometrisch</u> 2×2 Gl. Systeme $\Leftrightarrow$ Schnitt von Geraden

schlecht konditioniert

$$\begin{pmatrix} - a_1 - \\ - a_2 - \end{pmatrix} x = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \qquad a_1 \sim \lambda a_2 \Rightarrow |\det A| \ll 1$$

$$\Rightarrow |||A^{-1}||| \gg 1$$

<u>nichtlineares Problem</u>: Nullstellensuche



$f(x) = a$ $\quad$ a Input, x Output

$a = 0$ $\qquad \|f'(x)\| \ll 1 \Rightarrow |||(f^{-1})'||| \gg 1$

2.2.8. <u>Lemma</u> ($|||.|||_1$, $|||.|||_\infty$ - Normen)

$\|x\|_1 = \sum_{i=1}^{n} |x_i|$, $\|x\|_\infty = \max_{i=1,\dots,n} |x_i|$ für $x \in \mathbb{R}^n$, dann ergeben sich die Operatornormen ($A \in \mathbb{R}^{m,n}$):

$$|||A|||_\infty = \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}|$$

$$|||A|||_1 = \max_{j=1,\dots,n} \sum_{i=1}^{m} |a_{ij}|$$

2.2.9. <u>Lemma</u> (Konditionen verketteter Abbildungen)

$f(x) = (h \circ g)(x)$, dann gilt

$$K_f = K_h \, K_g$$

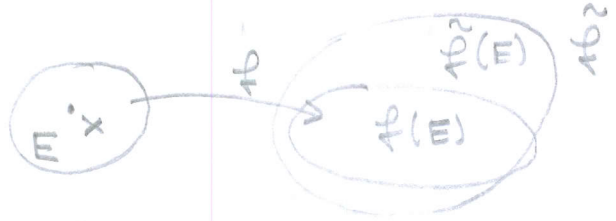
<u>Bew</u>

$$\frac{\|f(\hat{x}) - f(x)\|}{\|f(x)\|} \lesssim K_h \frac{\|g(\hat{x}) - g(x)\|}{\|g(x)\|} \lesssim K_h K_g \frac{\|\hat{x} - x\|}{\|x\|}$$

$\square$

2.3. Stabilität numerischer Algorithmen

Notation



$\tilde{f} \leftrightarrow$ eine gestörte Abb. (algorithmische Realisierung von f)

eigentl eine Menge gestörter Abb.

Vorwärtsanalyse: zu $E, \tilde{f}$ suche $\tilde{R} = \tilde{f}(E)$

Rückwärtsanalyse: zu $E, \tilde{f}$ such $\tilde{E}$, so daß

$$f(\tilde{E}) \supset \tilde{f}(E) = \tilde{R}$$

(Interpretiere Fehler im Algorithmus als Eingabefehler)

2.3.1. Def (Vorwärtsstabilität)

Sei $\tilde{f}$ eine algorithmische Realisierung zum Problem (f, x), dann heißt δ_v Vorwärtsstabilitätsindikator, falls δ_v die kleinste Zahl ist, für die

$$\frac{\| \tilde{f}(\tilde{x}) - f(\tilde{x}) \|}{\| f(\tilde{x}) \|} \lesssim \delta_v K_{rel} \epsilon \quad \text{für } \epsilon \to 0$$

und $\tilde{x} \in E \subset B_\epsilon(x)$

$\tilde{f}$ heißt vorwärtsstabil, falls δ kleiner als die Zahl an elementaren Operationen in f.

2.3.2. Lemma (i) Für obige Realisierung der elementaren Operationen $(+, -, \cdot, /)$ gilt $\delta_v K_{rel} \leq 1$

(ii) $f = h \circ g \quad (g: \mathbb{R}^n \to \mathbb{R}^l, \; h: \mathbb{R}^l \to \mathbb{R}^m)$

dann gilt $\delta_{v,f} K_f \leq \delta_h K_h + \delta_g K_g K_h$

(iii) falls g, h skalar, diffbar $\delta_{v,f} \leq \frac{\delta_{v,h}}{K_g} + \delta_{v,g}$

<u>Bew</u> (i) $$\frac{|v \hat{\circ} w - v \circ w|}{|v \circ w|} \overset{2.1.2.}{=} \frac{|(v \circ w)(1+\delta(v,w)) - v \circ w|}{|v \circ w|}$$

$$= \delta(v,w) \underset{2.1.2.}{\leq} \epsilon$$

(ii) $$\|\tilde{f}(\tilde{x}) - f(\tilde{x})\| = \|\tilde{h}(\tilde{g}(\tilde{x})) - h(g(\tilde{x}))\|$$

$$\leq \|\tilde{h}(\tilde{g}(\tilde{x})) - h(\tilde{g}(\tilde{x}))\| + \|h(\tilde{g}(\tilde{x})) - h(g(\tilde{x}))\|$$

$$\underset{2.2.3.}{\lesssim} \delta_{v,h} K_h \epsilon \|h(\tilde{g}(\tilde{x}))\| + K_h \frac{\|\tilde{g}(\tilde{x}) - g(\tilde{x})\|}{\|g(x)\|} \|h(g(\tilde{x}))\|$$

$$\lesssim \delta_{v,h} K_h \epsilon + \delta_{v,g} K_h K_g \epsilon \|f(\tilde{x})\|$$

(iii) $$K_f = \frac{|x|\,|f'(x)|}{|f(x)|} = \frac{|x|\,|h'(g(x))|}{|h(g(x))|} \frac{|g'(x)|\,|g(x)|}{|g(x)|}$$

$$= K_h K_g$$

$$\overset{(ii)}{\Longrightarrow} \delta_{v,f} \leq \frac{\delta_{v,h} K_h}{K_h K_g} + \delta_{v,g} = \frac{\delta_{v,h}}{K_g} + \delta_{v,g}$$

$\square$

2.3.3. <u>Bem</u> <u>2.3.2. (ii) bei Elementaroperation:</u>

$$\underbrace{\delta_h K_h}_{\sim 1} + \underbrace{\delta_g K_g}_{\sim 1} \underbrace{K_h}_{\gg 1} \quad \text{falls } h = \hat{-}$$

Falls h die zweite Operation $\Rightarrow \delta_f \gg 1$

$\rightsquigarrow$ Subtraktion zuerst ausführen, wenn möglich!

2.3.2. (iii) Falls $K_g \ll 1$ (erste Operation)

$$\Rightarrow \delta_{v,f} \leq \frac{\delta_{v,h}}{\underbrace{K_g}_{\gg 1}} + \delta_{v,g} \Rightarrow \boxed{\delta_{v,f} \gg 1}$$

2.3.4. <u>Lemma</u> (i) $S_n : \mathbb{R}^n \to \mathbb{R}$ $\quad (x_i)_{i=1,\dots,n} \mapsto \sum_{i=1}^n x_i$

ist vorwärtsstabil $\quad \delta_v \le n-1$ (bzgl. $\|\cdot\|_1$)

(ii) $P_n : \mathbb{R}^{2n} \to \mathbb{R}$ $\quad (x,y) \mapsto x \cdot y$ (bzgl. $\|\cdot\|_1$)

ist vorwärtsstabil $\quad \delta_v \le \frac{n}{2}$

<u>Bew:</u> (i)

$$\delta_n K_n \overset{\boxed{2.3.2\,(ii)}}{\le} (\delta_{n-1} + \delta_+ K_+) K_{n-1}$$

bezogen auf S_+

bzgl. $S_n = S_{n-1} \circ S_+$ zwei Terme addieren

$$K_n \overset{2.2.3}{=} \frac{\sum_{i=1}^n |x_i|}{\left|\sum_{i=1}^n x_i\right|} \ge 1$$

$$K_{n-1} = \frac{\sum_{i=3}^n |x_i| + |x_1 + x_2|}{\left|\sum_{i=1}^n x_i\right|} \le K_n$$

$$\Rightarrow \delta_n \le \delta_{n-1} + \underbrace{\delta_+ K_+}_{\substack{2.3.2.(i) \\ \le 1}} \le \delta_{n-1} + 1$$

und $\delta_2 = \delta_+ \le \frac{1}{K_+} \le 1$

$\xrightarrow{\text{Induktion}}$ $\boxed{\delta_v \le n-1}$

(ii) $P_n = S_n \circ P$ $\quad$ mit $P = \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$

$(x,y) \longmapsto (x_i y_i)_{i=1,\dots,n}$

$$\delta_{P_n} K_{P_n} \overset{\boxed{2.3.2.\,(ii)}}{\le} (\delta_n + \underbrace{\delta_P K_P}_{\le 1}) K_n$$

$\le 1 \leftarrow$ da nur elementare Operationen und Norm $\|\cdot\|_1$

$$\le (\delta_n + 1) K_n$$

$$\le n K_n$$

$$\Rightarrow \delta_{P_n} \le n \frac{K_n}{K_{P_n}} = n \frac{\sum_{i=1}^n |x_i y_i| \Big/ \left|\sum_{i=1}^n x_i y_i\right|}{2 \sum_{i=1}^n |x_i y_i| \Big/ \left|\sum_{i=1}^n x_i y_i\right|}$$

$$\Rightarrow \delta_{P_n} \le \frac{n}{2}$$

hierzu: Kondition der Multiplikation $(x,y) \in \mathbb{R}^2 \mapsto x \cdot y \in \mathbb{R}$

$$f(x,y) = xy \quad \Rightarrow \quad |f'(x,y)(\delta x, \delta y)| = |y\,\delta x + x\,\delta y|$$
$$\leq \max\{|x|, |y|\}(|\delta x| + |\delta y|)$$

$$\Rightarrow K_* = \frac{\|(x,y)\|_1 \max\{|x|,|y|\}}{|x \cdot y|}$$
$$\geq \frac{(|x|+|y|)\max\{|x|,|y|\}}{\frac{1}{2}(|x|+|y|)\max\{|x|,|y|\}} = 2$$

daraus folgt:

$$K_{P_n} = \frac{\sum_{i=1}^{n}\left((|x_i|+|y_i|)\max\{|x_i|,|y_i|\}\right)}{\left|\sum_{i=1}^{n} x_i y_i\right|}$$

1. Schritt K_* Abschätzung

$$\geq 2\,\frac{\sum_{i=1}^{n}|x_i y_i|}{\left|\sum_{i=1}^{n} x_i y_i\right|}$$

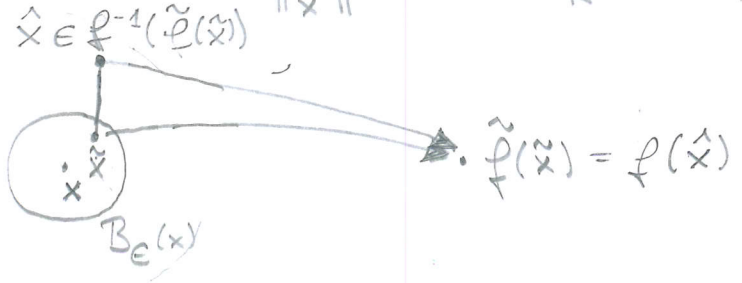
analog 2. Schritt oben

$\square$

2.3.5 <u>Def</u> (Rückwärtsstabilität)

Sei $\tilde{f}$ eine algorithmische Realisierung zum Problem (f, x), dann heißt δ_R Rückwärtsstabilitätsindikator, falls δ_R die kleinste Zahl ist, so dass es für jedes $\tilde{x} \in E \subset B_\epsilon(x)$ ein $\hat{x} \in \mathcal{U}$ gibt mit $f(\hat{x}) = \tilde{f}(\tilde{x})$ und

$$\frac{\|\tilde{x} - \hat{x}\|}{\|\tilde{x}\|} \lesssim \delta_R \epsilon \quad \text{für } \epsilon \to 0.$$



2.3.6. <u>Lemma</u> (i) $\delta_V \leq \delta_R$ (Rückwärtsstabil $\Rightarrow$ Vorwärtsstabil)

(ii) Die elementaren Operationen sind rückwärtsstabil

Bew (i) $\dfrac{\|\tilde f(\tilde x) - f(\tilde x)\|}{\|f(\tilde x)\|} = \dfrac{\|f(\hat x) - f(\tilde x)\|}{\|f(\tilde x)\|}$

$$\lesssim K \frac{\|\hat x - \tilde x\|}{\|\tilde x\|} \lesssim K \delta_R \epsilon$$

$$\Rightarrow K \delta_V \epsilon \underset{\text{Def. 2.3.1.}}{\leq} K \delta_R \epsilon \Rightarrow \delta_V \leq \delta_R$$

(ii) $\tilde x \hat{+} \tilde y = \tilde x \hat{+} \tilde y \,(1 + \delta(\hat x, \hat y))$

$$= \underbrace{\tilde x (1+\delta(\hat x, \hat y))}_{= \hat x} + \underbrace{\tilde y (1+\delta(\hat x, \hat y))}_{= \hat y}$$

$$\Rightarrow \frac{\|(\tilde x, \tilde y) - (\hat x, \hat y)\|}{\|(\tilde x, \tilde y)\|} \leq \delta(\hat x, \hat y) \frac{\|(\tilde x, \tilde y)\|}{\|(\hat x, \hat y)\|} \leq \epsilon$$

$$\Rightarrow \delta_{R,+} \leq 1$$

analog für $-, *, /$ $\square$

2.3.7. Lemma Das Skalarprodukt ist rückwärtsstabil: $\delta_R \leq n$

Bew: $P_n(x,y) = x \cdot y$

$\boxed{\text{Induktion}}$ $\boxed{n=1}$ 2.3.6. (ii)

$\boxed{n-1 \rightsquigarrow n}$ $x^{n-1} = (x_1, \dots x_{n-1})$, $y^{n-1} = (y_1, \dots y_{n-1})$

$$\tilde P_n(x,y) = \Big[x_n \cdot y_n (1+\delta) + \underbrace{\tilde P_{n-1}(x^{n-1}, y^{n-1})}_{= z^{n-1} \cdot y^{n-1}} \Big] (1+\delta')$$

wobei: $\|z^{n-1} - x^{n-1}\|_1 \lesssim (n-1)\epsilon \|x^{n-1}\|_1$ (Ind. Annahme)

nun definiere: $\left.\begin{aligned} \hat x_n &= x_n (1+\delta')(1+\delta) \\ \hat x^{n-1} &= z^{n-1}(1+\delta') \end{aligned}\right\} \Rightarrow \tilde P_n(x,y) = \hat x \cdot y$

Es gilt: $\|\hat x - x\|_1 \leq |\hat x_n - x_n| + \|\hat x^{n-1} - x^{n-1}\|_1$

$$\lesssim \underbrace{2}_{\leq n} \epsilon\, |x_n| + \underbrace{\|\hat{x}^{n-1} - z^{n-1}\|_1}_{\lesssim \epsilon \|z^{n-1}\|_1} + \underbrace{\|z^{n-1} - x\|_1}_{\lesssim (n-1)\epsilon \|x^{n-1}\|_1}$$

$$\lesssim \epsilon \|x^{n-1}\|_1$$

$$\lesssim n \epsilon \left(|x_n| + \|x^{n-1}\|_1\right) = n \epsilon \|x\|_1$$

$\square$

2.4. Anwendung auf lineare Gleichungssysteme

$Ax = b$ $\quad$ $\det A \neq 0$ „im Prinzip" eindt. lösbar!

numerisch: $\quad E = \{ \tilde{A} \mid \|A - \tilde{A}\| \leq \delta \|A\| \}$

A numerisch singulär, falls E eine singuläre Matrix enthält.

genauer: $\dfrac{\|\tilde{x} - x\|}{\|x\|} \leq \kappa(A)\, \delta$ $\quad$ (δ Relative Genauigkeit für A)

2.4.1. Def A mit relativer Genauigkeit δ heißt numerisch singulär, falls

$$\kappa(A)\, \delta \geq 1$$

2.4.2. Lemma Die Implementierung der Vorwärtssubstitution zum Dreiecksgleichungssystem $Lx = b$ berechnet ein $\hat{x}$ zu dem es ein $\hat{L}$ gibt mit

$$\hat{L}\, \hat{x} = b \quad \text{und}$$

$$|||L - \hat{L}|||_\infty \lesssim n \epsilon |||L|||_\infty \quad (\rightsquigarrow \delta_R \leq n)$$

Bew $\boxed{\text{Induktion}}$

$$l_{kk}\, x_k = b_k - l^{k-1} \cdot x^{k-1} \quad \text{mit } l^{k-1} = (l_{k1}, \ldots, l_{k,k-1})$$
$$x^{k-1} = (x_1, \ldots, x_{k-1})$$

fehlerbehaftet:

$$l_{kk}(1+\epsilon')(1+\epsilon)\hat{x}_k = b_k - \tilde{P}_{n-1}(l^{k-1}, \hat{x}^{k-1})$$

relative Fehler der Multipl. $l_{kk} \cdot \hat{x}_k$

relative Fehler der Addition: $b_k - \tilde{P}_{n-1}$

2.3.7 $\Rightarrow$ Es existiert $\hat{l}^{k-1}$ mit

$$\hat{l}^{k-1} \cdot \hat{x}^{k-1} = \hat{p}_{n-1}(l^{k-1}, \hat{x}^{k-1}) \text{ mit}$$

$$\|\hat{l}^{k-1} - l^{k-1}\|_1 \lesssim (k-1)\,\epsilon\, \|l^{k-1}\|_1$$

Nun definiere $\hat{l}_{kk} := l_{kk}(1+\epsilon')(1+\epsilon)$ $\Rightarrow$

$$\hat{l}_{kk}\hat{x}_k = b_k - \hat{l}^{k-1} \cdot \hat{x}^{k-1} \Rightarrow$$

$\hat{L} = (\hat{l}_{ij})_{i,j=1,\dots,n}$ mit $\hat{L}\hat{x} = b$ und

$$|||L - \hat{L}|||_\infty \lesssim n\,\epsilon\, |||L|||_\infty$$

$\max\{k-1, 2\} \le n$ $\forall k \le n$

Zeilensummennorm

$\square$

2.4.3. <u>Lemma</u> A habe eine LR-Zerlegung. Die Gauß-Elimination berechnet numerisch $\hat{L}, \hat{R}$, so dass

$$\hat{A} = \hat{L}\hat{R}$$

für $\hat{A}$ mit $|||\hat{A} - A||| \lesssim 4n\, |||\hat{L}|||_\infty\, |||\hat{R}|||_\infty\, \epsilon$

<u>Bew</u> Notation $|x| = (|x_i|)_{i=1,\dots,n}$ $\quad |A| = (|a_{ij}|)_{i,j=1,\dots,n}$

wir beweisen zunächst:

$$(*) \quad \boxed{\hat{L}\hat{R} = A + \delta A \text{ mit } |\delta A| \lesssim 2n\,\epsilon\,(|A| + |\hat{L}||\hat{R}|)}$$

<u>daraus folgt dann:</u>

$$|\delta A| \lesssim 2n\epsilon\,(\underbrace{|\delta A|}_{\text{von höherer Ordnung}} + |\hat{L}\hat{R}| + |\hat{L}|\,|\hat{R}|) \lesssim 2n\epsilon\,(|\hat{L}||\hat{R}| + |\hat{L}||\hat{R}|)$$

$$= 4n\epsilon\,|\hat{L}||\hat{R}|$$

$$\Rightarrow |||\delta A|||_\infty \lesssim 4n\epsilon\, |||\,|\hat{L}|\,|\hat{R}|\,|||_\infty$$

$$\lesssim 4n\epsilon\, |||\hat{L}|||_\infty\, |||\hat{R}|||_\infty \Rightarrow \text{Beh.}$$

zu ⊛: $\boxed{\text{mit Induktion}}$

$\boxed{n=1}$ $\quad A = 1 \cdot R \quad (R=A) \quad (\checkmark)$

$\boxed{n \rightsquigarrow n+1}$

$$A^{(1)} = \left(\begin{array}{c|c} a_{11} & -w^T- \\ \hline \begin{matrix} | \\ v \\ | \end{matrix} & B^{(1)} \end{array}\right) \qquad A^{(2)} = \left(\begin{array}{c|c} a_{11} & -w^T- \\ \hline 0 & B^{(1)} - \frac{v w^T}{a_{11}} \end{array}\right)$$

numerisch: $$\hat{A}^{(2)} = \left(\begin{array}{c|c} a_{11} & -w^T- \\ \hline 0 & \hat{B}^{(2)} \end{array}\right)$$

Induktionsannahme:

$$\hat{L}^{n-1}\hat{R}^{n-1} = \hat{B}^{(2)} + \widehat{\delta B}^{(2)} \text{ mit}$$
$$|\widehat{\delta B}^{(2)}| \lesssim 2(n-1)\epsilon\left(\hat{B}^{(2)} + |\hat{L}^{n-1}||\hat{R}^{n-1}|\right)$$

genaue Analyse des 1. Schritts:

$e^{(1)} = \frac{v}{a_{11}}$ numerisch: $\hat{e}^{(1)} = \frac{\hat{v}}{a_{11}}$ mit $|\hat{v}-v| \lesssim \epsilon|v|$

$B^{(2)} = B^{(1)} - e^{(1)}w^T$ numerisch: $\hat{B}^{(2)} = \hat{B}^{(1)} - \hat{e}^{(1)} \cdot w^T$

mit $|\hat{B}^{(1)} - B^{(1)}| \leq 2\epsilon |B^{(1)}|$

[eine Addition ⊕
eine Multiplikation]

$$\hat{L}\hat{R} = \left(\begin{array}{c|c} 1 & \\ \hline \hat{e}^{(1)} & \hat{L}^{n-1} \end{array}\right)\left(\begin{array}{c|c} a_{11} & -w^T- \\ \hline 0 & \hat{R}^{n-1} \end{array}\right)$$

$$= \underbrace{\left(\begin{array}{c|c} a_{11} & -w^T- \\ \hline \hat{e}^{(1)} a_{11} & \hat{L}^{n-1}\hat{R}^{n-1} + \hat{e}^{(1)} w^T \end{array}\right)}_{= A + \delta A} \quad \text{mit } |\delta A| \lesssim \left(\begin{array}{c|c} 0 & 0 \\ \hline \epsilon|v| & |\widehat{\delta B}^{(2)}| + 2\epsilon|B^{(1)}| \end{array}\right)$$

hierbei: $\hat{L}^{n-1}\hat{R}^{n-1} + \hat{e}^{(1)} w^T = \hat{B}^{(2)} + \delta\hat{B}^{(2)} + \hat{e}^{(1)} w^T$

$$= \hat{B}^{(1)} + \delta\hat{B}^{(2)}$$

$|\cdot| \leq |B^{(1)}| + 2\epsilon |B^{(1)}|$

weiter: $|\delta\hat{B}^{(2)}| + 2\epsilon |B^{(1)}|$ $\quad \left(|\hat{B}^{(2)}| \leq |\hat{B}^{(1)}| + |\hat{e}^{(1)}||w^T| \right)$

$$\lesssim 2(n-1)\epsilon \Big(\big(|a_{ij}|\big)_{i,j=2,\dots,n} + \big(|\hat{l}_{i1}||\hat{r}_{1j}|\big)_{i,j=2,\dots,n} + \Big(\sum_{k=2}^{n} |\hat{l}_{ik}||\hat{r}_{kj}|\Big)_{i,j=2,\dots,n} \Big)$$

Ind. Annahme

$$+ 2\epsilon \big(|a_{ij}|\big)_{i,j=2,\dots,n}$$

$$\lesssim 2n\epsilon \big(|A| + |\hat{L}||\hat{R}| \big)$$

$\square$

2.4.4. <u>Korollar</u> A habe eine LR-Zerlegung, dann berechnet die Gauß-Elimination für $Ax=b$ numerisch ein $\hat{x}$ mit $\hat{A}\hat{x} = b$ und

$$|||A - \hat{A}|||_\infty \leq 6n\, |||L|||_\infty\, |||R|||_\infty\, \epsilon$$

<u>Bew</u> $\quad (LR = A \rightsquigarrow Lz = b \rightsquigarrow Rx = z)$

<u>numerisch</u>: $\quad (\hat{L} + \delta\hat{L})\hat{z} = b$

$$(\hat{R} + \delta\hat{R})\hat{x} = \hat{z}$$

$$\Rightarrow \quad (\hat{L}\hat{R} + \hat{L}\delta\hat{R} + \delta\hat{L}\hat{R} + \delta\hat{L}\delta\hat{R})\hat{x} = b$$

$\hat{L}\hat{R} = A + \delta A$

$$\Rightarrow \quad \underbrace{(A + \delta A + \hat{L}\delta\hat{R} + \delta\hat{L}\hat{R} + \delta\hat{L}\delta\hat{R})}_{=:\hat{A}}\hat{x} = b$$

$$\Rightarrow |||A - \hat{A}|||_\infty \lesssim \underbrace{6n}_{4n+n+n}\epsilon\, |||L|||_\infty\, |||R|||_\infty$$

$\square$

2.4.5. <u>Satz</u> Die Gauß-Elimination mit Zeilenpivotierung für $Ax=b$ berechnet $\hat{x}$, so dass $\hat{A}\hat{x}=b$ mit

$$\frac{|||A-\hat{A}|||_\infty}{|||A|||_\infty} \leq 6n^3 \varrho_n(A)\,\epsilon$$

wobei $\varrho_n(A) = \dfrac{\max\limits_{k,i,j} |a_{ij}^{(k)}|}{\max\limits_{i,j} |a_{ij}|}$

<u>Bew</u> Seien $\hat{P}, \tilde{L}, \tilde{R}$ und $\hat{x}$ die berechneten Größen bei der Zerlegung $PA = LR$

$\Rightarrow$ $\hat{P}A$ besitzt LR-Zerlegung für ϵ klein und nach 2.4.4. gilt es $\hat{A}$ mit $\hat{A}\hat{x}=b$ und

$$|||A-\hat{A}|||_\infty \lesssim 6n \underbrace{|||\hat{P}^T|||_\infty}_{=1} |||\hat{L}|||_\infty |||\hat{R}|||_\infty \epsilon$$

<u>im Detail:</u>

$\hat{P}A = LR \qquad (\hat{L}+\delta\hat{L})\hat{z} = \hat{P}b \qquad (\hat{R}+\delta\hat{R})\hat{x} = \hat{z}$

$\hat{P}A + \delta A = \hat{L}\hat{R}$

$\Rightarrow \hat{P}^T(\hat{L}+\delta\hat{L})(\hat{R}+\delta\hat{R})\hat{x} = b$

$\Rightarrow \hat{P}^T(\underbrace{\hat{L}\hat{R}}_{\hat{P}A+\delta A} + \delta\hat{L}\hat{R} + \hat{L}\delta\hat{R} + \delta\hat{L}\delta\hat{R})\hat{x} = b$

$\Rightarrow \hat{A} \approx A + \hat{P}^T\delta A + \hat{P}^T\hat{L}\delta\hat{R} + \hat{P}^T\delta\hat{L}\hat{R}$

und $\hat{P}\hat{A}\hat{x} = b$

ferner $|||\hat{A}-A|||_\infty \lesssim \underbrace{|||\hat{P}^T|||_\infty}_{=1} \Big(\underbrace{|||\delta A|||_\infty}_{\leq 4n\epsilon|||\hat{L}|||_\infty|||\hat{R}|||_\infty} + 2n\epsilon|||\hat{L}|||_\infty|||\hat{R}|||_\infty\Big)$

Da Pivotsuche $\Rightarrow$ $|||\hat{L}|||_\infty \le n$

$$|||\hat{R}|||_\infty \le n \max_{i,j} |\hat{r}_{ij}| \le n \max_{i,j,k} |a_{ij}^{(k)}|$$

$$\frac{|||A-\hat{A}|||_\infty}{|||A|||_\infty} \lesssim 6n^3 \epsilon \frac{\max_{i,j,k} |a_{ij}^{(k)}|}{\max_{i,j} |a_{ij}|}$$

$$\boxed{|||A|||_\infty \ge \max_{i,j} |a_{ij}|}$$

$\square$

2.4.6. <u>Bem</u> • $|a_{ij}^{(k+1)}| = |a_{ij}^{(k)} - \underbrace{l_{ik}}_{|\cdot| \le 1} a_{kj}^{(k)}| \le 2 \max_{i,j} |a_{ij}^{(k)}|$

$$\xrightarrow{\text{Iteration}} \quad \rho_n(A) \le 2^{n-1}$$

• Für $A = \begin{pmatrix} 1 & & & 1 \\ -1 & \ddots & 0 & \vdots \\ \vdots & \ddots & \ddots & \\ -1 & \cdots & -1 & 1 \end{pmatrix}$ gilt „=" $\boxed{\text{Ü}}$

• <u>genauere Stabilitätsaussagen</u>:

A tridiagonal : $\rho_n \le 2$

symm., pos. def : $\rho_n \le 1$

3. Lineare Gleichungssysteme, direkte Löser II

Rückblick: $A \rightsquigarrow B^{(1)}A \rightsquigarrow B^{(2)}B^{(1)}A \rightsquigarrow \dots B^{(m)} \dots B^{(1)}A = R$

bisher: $B^{(i)} = L^{(i)} \quad m = n-1$

$\longrightarrow$ Stabilitätsindikator wächst mit der Konditionszahl $K^{(i)}$ der Einzelschritte

bisher: $K^{(i)} = |||L^{(i)}|||_\infty \, |||L^{(i)-1}|||_\infty$

$\underset{\text{mit Pivotierung}}{\leq} 2 \cdot 2 = 4$

falls: $B^{(i)} = Q^{(i)}$ mit $Q^{(i)} \in O(n)$, dann gilt

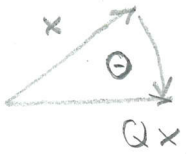
$K^{(i)} = |||Q^{(i)}|||_2 \, |||Q^{(i)T}|||_2 = 1$

$\boxed{\text{bzgl. der 2-Norm}}$

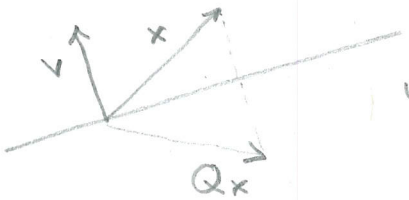
$\rightsquigarrow$ bessere Stabilität!

Frage: Welche orthogonalen Transformationen kommen in Frage?

$\boxed{O(2)}$ Rotationen $\quad Q = \begin{pmatrix} \cos\Theta & \sin\Theta \\ -\sin\Theta & \cos\Theta \end{pmatrix}$



Spiegelung



v normal auf der Spiegelebene

$$x \mapsto x - 2\frac{x \cdot v}{\|x\|^2} v =: Qx \qquad ; \|x\| = \|x\|_2 = \sqrt{\sum_i x_i^2}$$

$$Q = \mathbb{1} - 2\frac{v}{\|v\|}\frac{v^T}{\|v\|}$$

O.B.d.A $\|v\| = 1$:

$$QQ^T = (\mathbb{1} - 2vv^T)(\mathbb{1} - 2vv^T)$$

$$= \mathbb{1} - 4vv^T + 4v\underbrace{v^Tv}_{=1}v^T = \mathbb{1}$$

3.1. QR-Zerlegung mit Givens-Rotationen

3.1.1. Def Matrizen

$$Q_{ij} = \begin{array}{c} \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & c & & & s & \\ & & & & 1 & & & \\ & & & & & \ddots & & \\ & & & s & & & c & \\ & & & & & & & 1 \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{pmatrix} \\ \begin{array}{cccccccccc} & & & i & & & j & & & \end{array} \end{array} \qquad c^2+s^2=1$$

heißen Givensrotationen

geometrisch: Rotation in der durch e_i, e_j aufgespannten Ebene

Ziel: Eliminierung

$\boxed{\text{in 2D}}$ $$\begin{pmatrix} c & s \\ -s & c \end{pmatrix}\begin{pmatrix} x_i \\ x_j \end{pmatrix} = \begin{pmatrix} r \\ 0 \end{pmatrix}$$

$$\Leftrightarrow \quad r = \pm\sqrt{x_i^2 + x_j^2} \qquad c = \frac{x_i}{r} \qquad s = \frac{x_j}{r}$$

(numerisch nicht stabil)

besser: $\boxed{|x_j| > |x_i|}$ $\quad \tau := \frac{x_i}{x_j}$

$(\tau \Leftrightarrow \tan\Theta, \cot\Theta)$ $\quad s := \frac{1}{\sqrt{1+\tau^2}} \qquad c := s\tau$

$\boxed{|x_j| \geq |x_i|}$ $\quad \tau := \frac{x_j}{x_i}$

$c := \frac{1}{\sqrt{1+\tau^2}} \qquad s := c\tau$

$\boxed{A \in \mathbb{R}^{n,n}}$

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \\ a_{n-1,1} & & a_{n-1,n} \\ a_{n,1} & \cdots & a_{n,n} \end{pmatrix} \xrightarrow{Q_{1,n}} \begin{pmatrix} a_{11}^{(2)} & \cdots & a_{1n}^{(2)} \\ a_{21} & \cdots & a_{2n} \\ \vdots & & \vdots \\ a_{n-1} & \cdots & a_{n-1,n} \\ 0 & a_{n,2}^{(2)} \cdots & a_{n,n}^{(2)} \end{pmatrix} \xrightarrow{Q_{1,n-1}} \begin{pmatrix} & * & \\ 0 & & \\ 0 & & \end{pmatrix}$$

$a_{n1} \neq 0$

$a_{11} \neq 0$

$$\xrightarrow{Q_{1,n-2}} \cdots \xrightarrow{Q_{1,2}} \begin{pmatrix} * & & \\ 0 & & \\ \vdots & * & \\ 0 & & \end{pmatrix} \xrightarrow{Q_{2n}} \cdots \xrightarrow{Q_{23}} \begin{pmatrix} * & * & \\ 0 & * & \\ \vdots & \vdots & * \\ 0 & 0 & \end{pmatrix} \cdots \xrightarrow{Q_{n-1,n}} R$$

mit $R = \begin{pmatrix} r_{11} & & * \\ & \ddots & \\ 0 & & r_{nn} \end{pmatrix}$

<u>Bem</u>: Q_{ij} nur ausführen, falls $a_{ij}^{(k)} \neq 0$ im k-ten Schritt

$$R = Q_{n-1,n} \cdots\cdots Q_{1n} A$$

<u>Aufwand</u>: $\frac{(n-1)^2}{2}$ Wurzeln und Divisionen

$$\sum_{i=1}^{n-1} 4(n-i)^2 + O(n^2) = 4\sum_{i=1}^{n-1} i^2 + O(n^2) = \frac{4}{3} n^3 + O(n^2)$$

Multiplikationen

<u>Zsf</u>: tun aber stabil!

<u>Frage</u>: Kann man mit einer Orthogonalen Transformation eine ganze Spalte eliminieren (vgl. Gauß-Elimination)?

3.2. <u>QR-Zerlegung mit Householder-Matrizen</u>

3.2.1. <u>Lemma</u> Für $Q = \mathbb{1} - 2\frac{vv^T}{\|v\|^2}$ (Householder-Matrix) gilt $Q = Q^T$, $Q^{-1} = Q$, $Q^2 = \mathbb{1}$

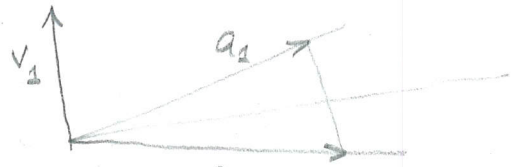
<u>Bew</u> $Q = Q^T$ (✓), $Q^{-1} = Q^T$ (s.o.), $Q^2 = QQ^T = QQ^{-1} = \mathbb{1}$ □

<u>Ziel</u>:

$$Q_1 A = \begin{pmatrix} * & & \\ 0 & & \\ \vdots & & * \\ 0 & & \end{pmatrix} \quad \text{d.h. } Q_1 a_1 = \alpha_1 e_1$$

$$Q_1 A = \begin{pmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{pmatrix}$$

<u>geometrisch</u>:



$$Q_1 a_1 = \alpha_1 e_1 = a_1 - 2\frac{v_1 \cdot a_1}{\|v_1\|^2} v_1$$

da Q_1 orthogonal: $\alpha_1 = \pm \|a_1\|_2$

ferner: $v_1 \in \operatorname{span}\{a_1 - \alpha_1 e_1\}$

<u>wähle:</u> $v_1 = (a_{11} - \alpha_1, a_{21}, \ldots, a_{n1})$

damit es nicht zu Auslöschung kommt:

$$\alpha_1 = -\operatorname{sgn}(a_{11}) \, \|a_1\|_2$$

<u>Berechnung von $Q_1 x$:</u>

$$\|v_1\|^2 = (a_1 - \alpha_1 e_1)\cdot(a_1 - \alpha_1 e_1)$$

$$= \|a_1\|^2 - 2\alpha_1 a_{11} + \alpha_1^2 = -2\alpha_1(a_{11} - \alpha_1)$$

$$\Rightarrow Q_1 x = x + \frac{v_1 \cdot x}{\alpha_1(\alpha_{11} - \alpha_1)} v_1 = x + \frac{v_1 \cdot x}{\alpha_1 v_{11}} v_1$$

<u>damit erhalten wir:</u> $A^{(2)} = Q_1 A^{(1)}$ $\quad A^{(1)} = A$

mit $A^{(2)} = \begin{pmatrix} \alpha_1 & | & & | \\ 0 & & & \\ \vdots & Q_1 a_2 & \cdots & Q_1 a_n \\ 0 & | & & | \end{pmatrix}$

<u>Iteration:</u>

$$A^{(k)} = \begin{pmatrix} \alpha_1 & & * \\ & \ddots & \\ & \alpha_{k-1} & \\ 0 & & \begin{vmatrix} \tilde{a}_k^{(k)} & \cdots & a_n^{(k)} \\ \vdots & & \vdots \end{vmatrix} \end{pmatrix}$$

wähle $Q_k = \begin{pmatrix} \mathbb{1} & \\ & Q_k^{(k)} \end{pmatrix}$ $\quad Q_k^{(k)} \in \mathbb{R}^{n-k+1, n-k+1}$

$e_1^{(k)} = e_k$

$$Q_k = \mathbb{1} - 2\frac{v_k v_k^T}{\|v_k\|^2} \qquad a_k^{(k)} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \tilde{a}_k^{(k)} \\ a_k^{(k)} \end{pmatrix}$$

mit $v_k = a_k^{(k)} - \alpha_k e_1^k$

$$\alpha_k = -\operatorname{sgn}(\tilde{a}_{kk}^{(k)}) \, \|a_k^{(k)}\|_2$$

$k = 1, \ldots, n-1$

$$R = Q_{n-1} A^{(n-1)} = Q_{n-1} \cdots Q_1 A \text{ mit } R = \begin{pmatrix} \alpha_1 & & \\ & \ddots & \\ & & \alpha_{n-1} \\ & & & \alpha_n \end{pmatrix}$$

α_n ergibt sich aus dem Algorithmus

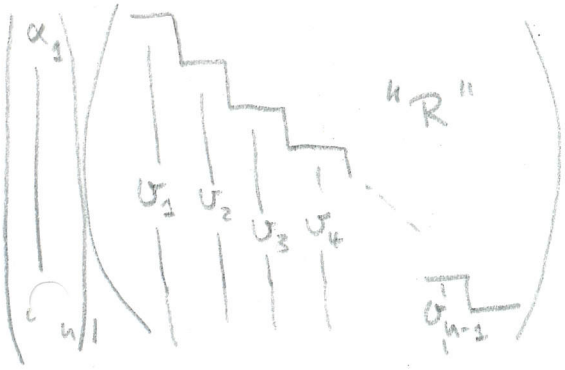
$\Rightarrow \quad A = QR \quad$ mit $\quad Q = Q_1 Q_2 \text{ —— } Q_{n-1}$

$Q_k = Q_k^{-1}$

3.2.2. <u>Bem</u> (Speicherverwaltung)

$R, v_1, \text{—} v_{n-1}$ sind zu speichern

<u>Speichern:</u>
- $(\alpha_i)_{i=1,\text{—}n}$ als Vektor $\quad T_{ii} = \alpha_i$!
- $(T_{ij})_{\substack{i,j=1,\text{—}n \\ j>i}}$ in der Matrix A
- $v_1, \text{—} v_n$ in der Matrix A



3.2.3. <u>Algorithmus</u>

for $(k=1;\ k \le n-1;\ k{+}{+})\ \{$

$\alpha_k = -\operatorname{sgn}(a_{kk}^{(k)})\, \|a_k^k\|_2$;

$v_k = a_k^{(k)} - \alpha_k e_k$;

for $(j = k+1;\ j \le n;\ j{+}{+})$

$$a_j^{(k+1)} = Q_k\, a_j^{(k)} = a_j^{(k)} + \frac{v_k \cdot a_j^{(k)}\ v_k}{\alpha_k\,(a_{kk}^{(k)} - \alpha_k)}$$

$$a_j^{(k)} = \begin{pmatrix} 0 \\ | \\ 0 \\ a_{kj}^{(k)} \\ | \\ a_{nj}^{(k)} \end{pmatrix}$$

3.2.4. <u>Rechenaufwand</u>

$v_k \cdot a_j^{(k)} \mathrel{\hat{=}} (n-k)$ Multipl. $\quad , \left(\frac{v_k \cdot a_j^{(k)}}{(\text{—})}\right) v_k \quad (n-k)$ Multipl.

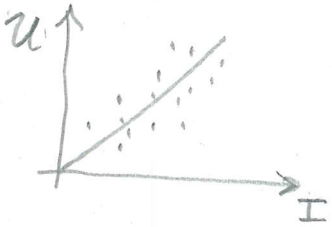
$$\Rightarrow \sum_{k=1}^{n-1} 2(n-k)(n-k) + O(n^2) \text{ (Multiplikation)}$$

$$= 2\sum_{k=1}^{n-1} k^2 + O(n^2) = \frac{2}{3} n^3 + O(n^2)$$

und $(n+1)$ Wurzeln

3.3. Die Methode der kleinsten Quadrate

3.3.1. Bsp (Experimentelle Bestimmung des elektr. Widerstands)



(U_i, I_i) Messungen

Modell $U = RI$ mit Meßfehler:

$$U_i = R I_i + \epsilon_i$$

Ansatz: Gesucht ist R, so dass $\sqrt{\sum_{i=1}^{m} \epsilon_i^2}$ minimal ist

$$E_m(R) = \sum_{i=1}^{m} (U_i - R I_i)^2 = \sum_{i=1}^{m} \epsilon_i^2 \longrightarrow \min$$

notwendige Bedingung: $\partial_R E_m(R) = 2 \sum_{i=1}^{m} (U_i - R I_i) I_i = 0$

$$\Rightarrow \quad R = \frac{\sum_{i=1}^{m} U_i I_i}{\sum_{i=1}^{m} I_i^2}$$

3.3.2. Bem: Allgemeiner betrachten wir

Messungen (a_i, b_i) $i = 1, \ldots m$

mit $a_i = (a_{i1} \ldots a_{im}) \in \mathbb{R}^n$, $b_i \in \mathbb{R}$

und einem Parametervektor $x \in \mathbb{R}^n$

Modell: $b = a \cdot x$

(oben: $b = U$, $a = I$, $x = R$, $n = 1$)

$r = Ax - b$ $\qquad A = (a_{ij})_{\substack{i=1,\ldots m \\ j=1,\ldots n}}$

$\uparrow$

Residuum $\qquad b = (b_i)_{i=1,\ldots m}$

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

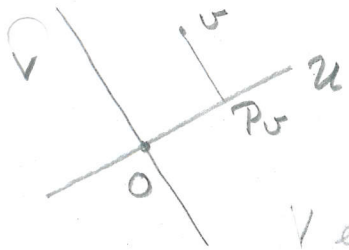
gesucht x, so dass $E(x) = \min_{\tilde{x} \in \mathbb{R}^n} E(\tilde{x})$

Ein kurzer Exkurs:

3.3.3. Satz (Projektionssatz)

V endl. dim Vektorraum mit Skalarprodukt, $\mathcal{U}$ Unterraum, $\mathcal{U}^\perp = \{v \in V \mid v \cdot u = 0 \ \forall u \in \mathcal{U}\}$ orthog. Komplement, dann existiert zu $v \in V$ genau ein $Pv \in \mathcal{U}$, so dass $\|v - Pv\| = \min_{u \in \mathcal{U}} \|v - u\|$; ferner ist P linear und $v - Pv \in \mathcal{U}^\perp$.

Bew (i)



Falls Pv existiert:

$$\|Pv\| \le \|v - Pv\| + \|v\| \le \|v - 0\| + \|v\| \le 2\|v\|$$

$$\Rightarrow \|Pv\| \in \overline{B_{2\|v\|}}(0) \subset V \quad \text{mit } B_R(v) = \{\tilde{v} \in V \mid \|\tilde{v} - v\| \le R\}$$

V endl. dim $\Rightarrow \mathcal{U} \cap \overline{B_{2\|v\|}(0)}$ ist kompakt

$\xRightarrow{\|v - \cdot\| \text{ stetig}}$ Minimum wird angenommen $\Rightarrow Pv$ existiert

(ii) $E(t) = \|v - Pv + tu\|^2$ hat globales Minimum in 0 $\forall u \in \mathcal{U}$

$$\Rightarrow 0 = \frac{d}{dt} E(t)\Big|_{t=0} = 2\left(\frac{d}{dt}(v - Pv + tu)\right) \cdot (v - Pv + tu)\Big|_{t=0}$$

$$= 2\, u \cdot (v - Pv) \qquad \forall u \in \mathcal{U}$$

$$\Rightarrow v - Pv \perp \mathcal{U}$$

(iii) Eindeutigkeit

$\boxed{\text{Annahme}}$ $Pv, P'v$ seien Minima $\Rightarrow$

$$(v - Pv) \cdot u = 0 \ \wedge (v - P'v) \cdot u = 0 \quad \forall u \in \mathcal{U} \Rightarrow$$

$\ominus$, $u = Pv - P'v$

$$\Longrightarrow 0 = (Pv - P'v) \cdot (Pv - P'v) = \|Pv - P'v\|^2$$

$$\Rightarrow Pv = P'v$$

(iv) Linearität Beh: $P(\alpha v + w) = \alpha Pv + Pw$

hinzu: $((\alpha v + w) - (\alpha Pv + Pw)) \cdot u =$

$$\alpha (v - Pv) \cdot u + (w - Pw) \cdot u = 0 \overset{\text{Eindt.}}{\Longrightarrow} \text{Beh} \ \square$$

3.3.4 <u>Satz</u> (Existenz zum Pb. der kleinsten Quadrate)

$A \in \mathbb{R}^{m,n}$, $b \in \mathbb{R}^m$, $m \geq n$, dann ist $x \in \mathbb{R}^n$ Lösung des Problems der kleinsten Quadrate

$$\|Ax - b\|^2 = \min_{\tilde{x} \in \mathbb{R}^n} \|A\tilde{x} - b\|^2$$

genau dann, wenn x die Normalgleichung

$$A^T A x = A^T b \quad \text{löst.}$$

Das Pb ist eindt. lösbar, falls der Rang von A maximal, d.h. Rang $A = n$.

<u>Bew</u> (i) wähle in 3.3.3. $V = \mathbb{R}^m$, $\mathcal{U} = R(A)$ $\overset{3.3.3.}{\Longrightarrow}$

$$\|Ax - b\| = \min \iff (Ax - b) \cdot A\tilde{x} = 0 \quad \forall \tilde{x} \in V$$

$$\iff A^T A x - A^T b = 0$$

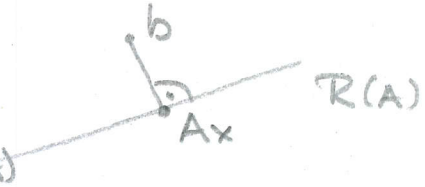
(ii) $\underbrace{A^T A}_{\mathbb{R}^{n,n}}$ invertiert $\iff$ $A^T A$ injektiv

$$A^T A x = 0 \Rightarrow A^T A x \cdot x = 0 \iff Ax \cdot Ax = 0$$

$$\iff \|Ax\|^2 = 0 \Rightarrow Ax = 0 \overset{\text{Rang } A = n}{\Longrightarrow} x = 0$$

$\square$

<u>Bem</u> $A^T A x = A^T b \iff$ geometrisch



$Ax - b \perp R(A)$

3.3.5. <u>Bsp</u> (i) $\boxed{\text{siehe 3.3.1.}}$

$$A = \begin{pmatrix} I_1 \\ | \\ I_m \end{pmatrix}, \quad x = R, \quad b = \begin{pmatrix} u_1 \\ | \\ u_m \end{pmatrix}$$

$$A^T A = \sum_{i=1}^m I_i^2 \qquad A^T b = \sum_{i=1}^m u_i I_i$$

<u>Normalgleichung</u>: $\left(\sum_{i=1}^m I_i^2\right) R = \sum_{i=1}^m u_i I_i$

(ii) Gegeben $(t_i, y_i)_{i=1,\dots,m}$, gesucht

$p(t) = c_0 p_0(t) + \dots c_n p_n(t)$ (z.B. $p_i(t) = t^i$ Polynom)

so dass $\sum_{i=1}^{m} |y_i - p(t_i)|^2 = \min$

<u>wobei</u>: $x = \begin{pmatrix} c_0 \\ \vdots \\ c_n \end{pmatrix}$, $A = \begin{pmatrix} p_0(t_1) & \dots & p_n(t_1) \\ \vdots & & \vdots \\ p_0(t_m) & \dots & p_n(t_m) \end{pmatrix}$, $b = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$

$$A^T A = \left(\sum_{k=1}^{m} p_i(t_k)\, p_j(t_k) \right)_{i,j=0,\dots,m}$$

<u>Bem</u>
- Hier nur lineare Pb ($p(\cdot)$ hängt linear von c_i ab)
- Hier nur minimal in der 2-Norm $\|Ax-b\|_2$

  <u>später</u>: $\|\cdot\|_\infty$

<u>Nun wollen wir die Kondition des Problems analysieren:</u>

<u>Erinnerung</u>: $K_2(A) = |||A|||_2 \, |||A^{-1}|||_2 = \dfrac{\max_{\|x\|=1} \|Ax\|_2}{\min_{\|x\|=1} \|Ax\|_2} \in [0,\infty]$

$|||A|||_2 = \sup_{x \neq 0} \dfrac{\|Ax\|_2}{\|x\|_2}$

allgemein:

λ_{max} = maximale Singulärwert

λ_{min} = minimale Singulärwert

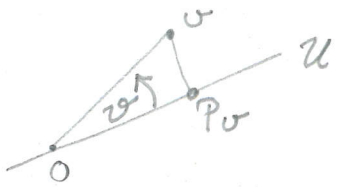
$K_2(A) = \dfrac{\lambda_{max}}{\lambda_{min}}$

$$K_2(A^TA) = \frac{\max_{\|x\|=1} \|A^TAx\|}{\min_{\|x\|=1} \|\underbrace{A^TA}_{\text{symm.}}x\|} = \frac{\lambda_{max}(A^TA)}{\lambda_{min}(A^TA)}$$

$$= \frac{\max_{\|x\|=1} A^TAx \cdot x}{\min_{\|x\|=1} A^TAx \cdot x} = \frac{\max_{\|x\|=1} \|Ax\|^2}{\min_{\|x\|=1} \|Ax\|^2}$$

$$= K(A)^2$$

3.3.6. Lemma $P: \mathbb{R}^m \to \mathcal{U}$ ($\mathcal{U}$ Unterraum) orthogonale Projektion, dann gilt für das Problem (P, v):



$$K_{rel} \leq \frac{1}{\cos\vartheta}$$

wobei ϑ der Winkel zwischen v und $\mathcal{U}$.

Bew $\|Pv\|_2^2 = \|v\|_2^2 - \|v - Pv\|_2^2 \Rightarrow$

$$\frac{\|Pv\|_2^2}{\|v\|_2^2} = 1 - \sin^2\vartheta = \cos^2\vartheta$$

$$K_{rel} = \frac{\|v\|_2}{\|Pv\|_2} \|P'(v)\|_2 = \frac{1}{\cos\vartheta} \underbrace{|||P|||_2}_{=1}$$

□

3.3.7. Satz $A \in \mathbb{R}^{m,n}$, $m \geq n$, $\operatorname{Rang} A = n$, $b \in \mathbb{R}^n$,
x Lösung von $\|Ax - b\|_2 = \min$, $x \neq 0$
$\vartheta = \sphericalangle(b, R(A))$ $\left(\sin\vartheta = \frac{\|Ax-b\|}{\|b\|}\right)$, dann gilt für:

(i) die Kondition K_{rel} bzgl. Störungen von b:

$$K_{rel} \leq \frac{K_2(A)}{\cos\vartheta}$$

(ii) die Kondition K_{rel} bzgl. Störungen von A:

$$K_{rel} \leq K_2(A) + K_2(A)^2 \tan\vartheta$$

Bew: $\boxed{\text{zu (i)}}$ $x = \phi(b) := (A^TA)^{-1}A^Tb \Rightarrow$

$$K_{rel} = \frac{|||\phi'(b)|||_2 \, \|b\|_2}{\|x\|_2} = \frac{|||(A^TA)^{-1}A|||_2 \, \|b\|_2}{\|x\|_2}$$

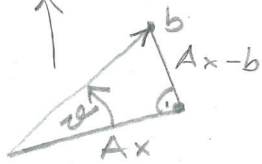
zunächst:

$$K_2(A) = \frac{\max\limits_{\|x\|=1} \|Ax\|}{\min\limits_{\|x\|=1} \|Ax\|} = |||A|||_2 \max_{x \neq 0} \frac{\|x\|}{\|Ax\|}$$

$$= |||A|||_2 \max_{x, y = Ax} \frac{\|x\|}{\|y\|} \underset{A \text{ max Rang}}{=} |||A|||_2 \max_{x, A^Ty = A^TAx} \frac{\|(A^TA)^{-1}A^Ty\|}{\|y\|}$$

$$\underset{A \text{ max. Rang}}{=} |||A|||_2 \max_y \frac{\|(A^TA)^{-1}A^Ty\|}{\|y\|} = |||A|||_2 \; |||(A^TA)^{-1}A^T|||_2$$

$$\Rightarrow K_{rel} = \frac{K_2(A)\;\|b\|_2}{|||A|||_2\;\|x\|_2} \leq K_2(A)\frac{\|b\|_2}{\|Ax\|_2} = K_2(A)\cos\vartheta$$



$\boxed{\text{zu (ii)}}$ $\quad x = \phi(A) = (A^TA)^{-1}A^Tb$

$$\phi'(A)C = \frac{d}{dt}\,\phi(A+tC)\Big|_{t=0}$$

$$(A+tC)^T(A+tC)\,\phi(A+tC) = (A+tC)^Tb$$

$$\overset{\frac{d}{dt}\big|_{t=0}}{\Longrightarrow} \quad (C^TA + A^TC)\underbrace{\phi(A+tC)}_{=x \text{ für } t=0} + A^TA\,\phi'(A)C = C^Tb$$

$$\Rightarrow \quad \phi'(A)C = (A^TA)^{-1}\left(C^T(b-Ax) - A^TCx\right)$$

$$\Rightarrow \quad |||\phi'(A)||| \leq |||(A^TA)^{-1}|||\;\|Ax-b\| + |||(A^TA)^{-1}A^T|||\;\|x\|$$

$$\Rightarrow \quad K_{rel} = \frac{|||\phi'(A)|||\;|||A|||}{\|x\|}$$

$$\leq \underbrace{|||(A^TA)^{-1}|||\;|||A|||^2}_{=K_2(A^TA)=K_2(A)^2}\;\underbrace{\frac{|||Ax-b|||}{|||A|||\;\|x\|}}_{\leq \frac{|||Ax-b|||}{\|b\|}\,\frac{\|b\|}{\|Ax\|} = \frac{\sin\vartheta}{\cos\vartheta}} + \underbrace{|||A|||\;|||(A^TA)^{-1}A|||}_{=K_2(A)}$$

$$\boxed{\begin{aligned} K_2(A^TA) &= \frac{\max\|A^TAx\|}{\min\|A^TAx\|} \\ &= \frac{\lambda_{max}(A^TA)}{\lambda_{min}(A^TA)} \\ &= \lambda_{max}(A^TA)\,\lambda_{max}((A^TA)^{-1}) \\ &= \max_{\|x\|=1}\|Ax\|^2 \max_{\|x=1}\|(A^TA)^{-1}x\| \\ &= |||A|||^2\;|||(A^TA)^{-1}||| \end{aligned}}$$

$$\Rightarrow K_{rel} \leq K_2(A)^2\tan\vartheta + K_2(A)$$

$\square$

zur Lösung der Normalengleichung:

falls Rang $A = n \Rightarrow A^TA$ pos. def., symmetrisch ($\|Ax\|^2 > 0 \ \forall x \neq 0$)

$\rightsquigarrow$ Cholesky-Zerlegung möglich

Rechenaufwand: • Berechnung von A^TA Symmetrie $\frac{n(n+1)}{2} \cdot m$ Multipl.

• Cholesky Zerlegung $\frac{1}{6} n^3 + O(n^2)$ —"—

$\boxed{\text{Für } m \gg n}$ $\frac{1}{2} n^2 m + O(n^3 + nm)$

$\boxed{\text{Kondition für } n = m}$ $K_2(A^TA) = K_2(A)^2 \longrightarrow$ schlechte Kondition

3.3.9. Satz (QR-Verfahren)

Sei $A \in \mathbb{R}^{m,n}$, $m \geq n$, Rang $A = n$, $b \in \mathbb{R}^m$, ferner sei $Q^TA = R$ mit $Q \in O(m)$, $R = \begin{pmatrix} R^1 \\ 0 \end{pmatrix}$ für $R^1 \in \mathbb{R}^{n,n}$ obere Dreiecksmatrix. Dann gilt R^1 ist invertierbar und für $Q^Tb = \begin{pmatrix} b^1 \\ b^2 \end{pmatrix}$ mit $b^1 \in \mathbb{R}^n$, $b^2 \in \mathbb{R}^{m-n}$ erhält man die Lösung zu $\|Ax - b\| = \min$ als $x = R^{-1}b$ und $\|Ax - b\| = \|b^2\|$.

Bew $\|Ax - b\|^2 = \|Q^T(Ax - b)\|^2 = \left\| \begin{pmatrix} R^1x - b^1 \\ b^2 \end{pmatrix} \right\|^2$

$= \|R^1x - b^1\|^2 + \|b^2\|^2$ ⊛

Rang R^1 = Rang $A = n \Rightarrow R^1$ invertierbar

⊛ minimal, wenn $R^1x = b^1 \Leftrightarrow x = (R^1)^{-1}b^1$

ferner ist dann: $\|Ax - b\|^2 = \|b^2\|^2$ □

Um $Q^TA = R$ zu realisieren können wir Householder-Trafo einsetzen:

$$A \rightsquigarrow Q_1 A = \begin{pmatrix} * & \cdots & * \\ 0 & & \\ \vdots & & * \\ 0 & & \end{pmatrix} \rightsquigarrow \dots \; Q_{n-1} Q_{n-2} \cdots Q_1 A = \begin{pmatrix} * & & * \\ 0 & \ddots & \\ \vdots & & 0 \\ 0 & \cdots 0 & * \end{pmatrix}$$

$$\rightsquigarrow Q_n Q_{n-1} \cdots Q_1 A = \begin{pmatrix} * & \cdots & * \\ & \ddots & \vdots \\ 0 & & * \end{pmatrix} = R = \begin{pmatrix} R^1 \\ 0 \end{pmatrix}$$

ein Schritt mehr als bei $n \times n$ Matrizen

Für $Q = Q_1 \cdots Q_n$ gilt $A = QR$ $\rightsquigarrow$

3.3.8. <u>Lemma</u> Sei $A \in \mathbb{R}^{m,n}$, $m \geq n$, $\operatorname{Rang} A = n$, dann gibt es eine QR-Zerlegung $A = QR$ mit $R = \begin{pmatrix} R^1 \\ 0 \end{pmatrix}$, $R^1 \in \mathbb{R}^{n,n}$ obere Dreiecksmatrix und invertierbar.

3.3.9 <u>Rechenaufwand</u>

$$\sum_{k=0}^{n} (n-k) \underbrace{2(m-k)}_{\substack{Q_i \text{ Anwendung} \\ \text{auf Spalten}}} = 2 \sum_{k=0}^{n} k\, m - O(n^2)$$

$$= n^2 m + O(n^2 + nm)$$

<u>Konditionierung</u> ergibt sich aus der Konditionierung von

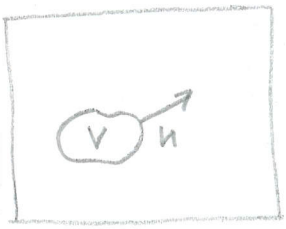
$$R^1 x = b^1 \quad \text{und} \quad \begin{pmatrix} b^1 \\ b^2 \end{pmatrix} = Q^T b$$

$\boxed{\text{für } n = m}$ $\quad K_2(Q^T A) = K_2(A) \leftarrow$ wesentlich besser!

4. Partielle Differentialgleichungen und Finite Elemente

4.1. Typen linear partieller Differentialgleichungen

Bsp Wärmeleitung



$\Omega \subset \mathbb{R}^d$ offen, beschränkt

gesucht ist Temperatur $u : \mathbb{R}_0^+ \times \overline{\Omega} \to \mathbb{R}$

$(t,x) \mapsto u(t,x)$

n äußere Normale zu Randwerten $u(t,x) = u^\partial(t,x) \quad \forall t \in \mathbb{R}^+ \; \forall x \in \partial\Omega$

und Anfangswerten $u(0,x) = u_0(x) \quad \forall x \in \Omega$

Herleitung einer Differentialgleichung:

Sei $V \subset \Omega$ glatt berandetes Gebiet (beliebig), dann gilt:

$$\partial_t \underbrace{\int_V u \, dx}_{\text{„Gesamtwärme" in } V} = \underbrace{-\int_{\partial V} q \cdot n \, da}_{\text{Wärmezufluss von Außen}} + \underbrace{\int_V f \, dx}_{\text{Quellterm (Heizung)}} \qquad \boxed{\text{Erhaltungssatz}}$$

wobei $q : \mathbb{R}_0^+ \times \Omega \to \mathbb{R}^d$ Wärmefluss

$f : \mathbb{R}_0^+ \times \Omega \to \mathbb{R}$ Quellterm

$\boxed{\text{Materialgesetz}}$

$$q(t,x) = -a(x) \nabla u(t,x)$$

Fluss in Richtung der steilsten Abstieg der Wärme (skaliert mit einem Wärmeleitkoeff. $a(\cdot)$)

$\boxed{\text{Satz von Gauß}}$

$$\int_{\partial V} q \cdot n \, da = \int_V \underbrace{\operatorname{div} q}_{\partial_{x_1} q_1 + \dots + \partial_{x_d} q_d} \, dx$$

$$\Rightarrow \int_V \partial_t u(t,x) + \operatorname{div} q(t,x) - f(t,x)\, dx = 0 \quad \forall \text{ Testvolumen } V$$

Glattheit des Integranden $\Rightarrow$

$$\partial_t u(t,x) + \operatorname{div} q(t,x) = f(t,x) \quad \forall t \in \mathbb{R}_0^+,\ x \in \Omega$$

$$\Rightarrow \boxed{\partial_t u(t,x) - \operatorname{div}(a(x)\nabla u(t,x)) = f(t,x) \quad \text{in } \Omega}$$

Wärmeleitglied

$a(x) \equiv 1$: $\boxed{\partial_t u - \Delta u = f}$

<u>Zum Satz von Gauß:</u>

$\boxed{d=1}$ <u>Hauptsatz:</u> $\Omega = (a,b)$

$$\int_a^b q'\, dx = +q(b) - q(a)$$

$+1$ äußere „Normale" bei b
-1 ——— " ——— bei a

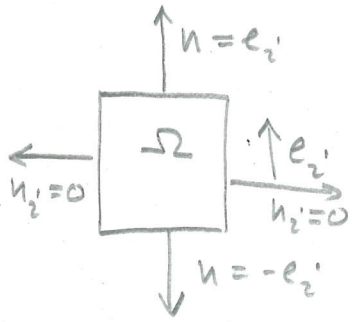
$d > 1$ $\quad \Omega = [a_1, b_1] \times [a_2, b_2] \times \ldots [a_d, b_d]$

$$\int_{a_1}^{b_1} \ldots \int_{a_i}^{b_i} \ldots \int_{a_d}^{b_d} \partial_{x_i} q_i \, dx_1 \ldots dx_d = \int_{a_1}^{b_1} \ldots \int_{a_{i-1}}^{b_{i-1}} \int_{a_{i+1}}^{b_{i+1}} \ldots \int_{a_d}^{b_d} \int_{a_i}^{b_i} \partial_{x_i} q_i \, dx_i \, d\hat{x}_i$$

$$dx_1 \ldots dx_{i-1}\, dx_{i+1} \ldots dx_d$$

$$\underset{\text{Hauptsatz}}{=} \int_{a_1}^{b_1} \ldots \int_{a_{i-1}}^{b_{i-1}} \int_{a_{i+1}}^{b_{i+1}} \ldots \int_{a_d}^{b_d} q_i(b_i) - q_i(a_i)\, d\hat{x}_i$$

$$= \int_{\partial\Omega} q_i(x)\, n_i(x)\, da(x)$$



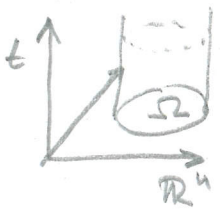
$$\Rightarrow \int_\Omega \operatorname{div} q\, dx = \int_\Omega \partial_{x_1} q_1 + \ldots \partial_{x_d} q_d \, dx_1 \ldots dx_d$$

$$= \int_{\partial\Omega} q_1 n_1 + \ldots q_d n_d \, da(x) = \int_{\partial\Omega} q(x) \cdot n(x)\, da(x)$$

(P) <u>Anfangswertproblem der Wärmeleitung</u>

$f \in C^0(\mathbb{R}_0^+ \times \overline{\Omega})$, $u_0 \in C^0(\overline{\Omega})$, $u^\partial \in C^0(\mathbb{R}_0^+ \times \partial\Omega)$

gesucht $u \in C^0(\mathbb{R}_0^+ \times \overline{\Omega}) \cap C^2(\mathbb{R}^+, \Omega)$, so dass



$$\partial_t u - \Delta u = f \quad \text{in } \mathbb{R}^+ \times \Omega$$
$$u(t,x) = u^\partial(t,x) \quad \text{auf } \mathbb{R}^+ \times \partial\Omega$$
$$u(0,x) = u_0(x) \quad \text{auf } \Omega$$

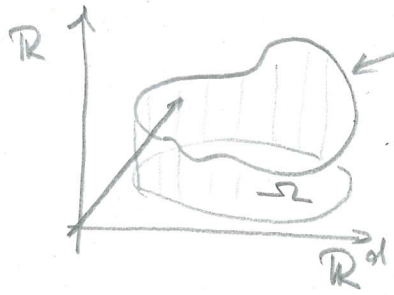
<u>Annahme</u>: für f, g konstant in t konvergiert $u(t,x) \to u^\infty(x)$, dann spricht man von einem stationären Pb (nenne u^∞ wieder u):

(E) <u>Randwertproblem</u>

$\Omega \subset \mathbb{R}^d$, $f \in C^0(\overline{\Omega})$, gesucht $u \in C^0(\overline{\Omega}) \cap C^2(\Omega)$, so dass

$$-\Delta u = f \quad \text{in } \Omega$$
$$u(x) = g(x) \quad \text{auf } \partial\Omega$$

<u>Bsp</u> <u>Membranschwingung</u>



<u>Problemstellung</u>:

gesucht: Auslenkung $u: \mathbb{R}_0^+ \times \overline{\Omega} \to \mathbb{R}$

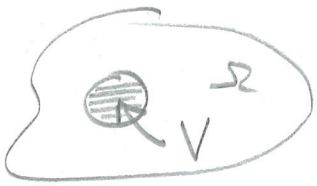
zu Randwerten $u(t,x) = u^\partial(t,x) \quad \forall x \in \partial\Omega \quad \forall t > 0$

Anfangswerten $u(0,x) = u_0(x) \quad \forall x \in \Omega$

$\partial_t u(0,x) = v_0(x) \quad \forall x \in \Omega$

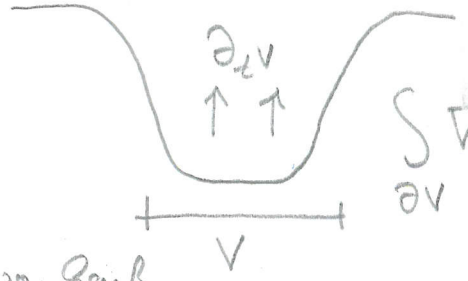
$v(t,x) = \partial_t u(t,x)$ Geschwindigkeit

$\partial_t v(t,x)$ Beschleunigung



$$\partial_t \int_V \underbrace{\rho}_{\text{Dichte} > 0} v(t,x)\, dx = -\int_{\partial V} \underbrace{F}_{\text{Randkräfte}} \cdot n \, da$$

<u>Materialgesetz</u>: $F = -K \nabla u$ <u>Hookesches Gesetz</u>



lokaler Auslenkungsvektor

$$\int_{\partial V} \nabla u \cdot n > 0$$

Satz von Gauß $\Longrightarrow$

$$\int_{\partial V} F \cdot n \, da = \int_V \operatorname{div} F \, dx \Rightarrow 0 = \int_V \partial_t u - \frac{K}{\varrho} \Delta u \, dx$$

$\partial_t^2 u, \Delta u \in C^0(\mathbb{R}^+ \times \Omega) \Longrightarrow$

$$\partial_t^2 u - \frac{K}{\varrho} \Delta u = 0$$

zur Vereinfachung: $\varrho, K = 1$

(H) Anfangsrandwertproblem der Membranschwingung:

$u^\partial \in C^0(\mathbb{R}^+_0 \times \partial\Omega)$, $u_0, v_0 \in C^0(\overline{\Omega})$, gesucht

$u \in C^1(\mathbb{R}^+_0 \times \overline{\Omega}) \cap C^2(\mathbb{R}^+ \times \Omega)$, so dass

$$\begin{aligned} \partial_t^2 u - \Delta u &= 0 && \text{in } \mathbb{R}^+ \times \Omega \\ u(t,x) &= u^\partial(t,x) && \text{in } \mathbb{R}^+ \times \partial\Omega \\ u(0,x) &= u_0(x), \ \partial_t u(0,x) = v_0(x) && \text{auf } \Omega \end{aligned}$$

Annahme: zu u^∂ konstant in t sucht man eine Gleichgewichtslage

(E') Randwertproblem

gesucht $u \in C^0(\Omega) \cap C^2(\overline{\Omega})$ mit

$$\begin{aligned} -\Delta u &= 0 && \text{in } \Omega \\ u &= u^\partial && \text{auf } \partial\Omega \end{aligned}$$

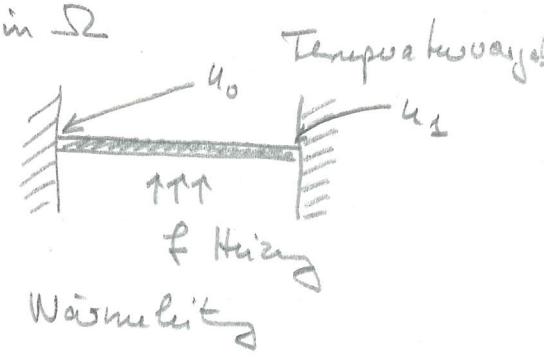
4.1 Das Randwertproblem in 1D

$\Omega = (0,1) \qquad \partial\Omega = \{0,1\}$

Differentialgleichung: $\quad -(a u')' = f \quad$ in Ω

Randwerte: $\quad u(0) = u_0$

$\qquad\qquad\quad u(1) = u_1$



Voraussetzungen:

$a \in L^\infty((0,1))$

$f \in L^2((0,1))$

mit $L^\infty(\Omega) = \{ f : \Omega \to \mathbb{R} \mid f \text{ meßbar}, \|f\|_{L^\infty(\Omega)} < \infty \}$

$$\|f\|_{L^\infty(\Omega)} = \inf_{\substack{N \subset \Omega \\ N \text{ Nullmenge}}} \sup_{x \in \Omega \setminus N} |f(x)|$$

$L^2(\Omega) = \{ f : \Omega \to \mathbb{R} \mid f \text{ meßbar}, \|f\|_{L^2(\Omega)} < \infty \}$

$$\|f\|_{L^2(\Omega)} = \int_\Omega |f(x)|^2 \, dx$$

Schwierigkeit mit klassischen Lösungen $u \in C^2(\Omega)$:

$$a = \begin{cases} 1 & ; x < \frac{1}{2} \quad \leftarrow \text{ niedrige Wärmeleitcoeff.} \\ 2 & ; x \geq \frac{1}{2} \quad \leftarrow \text{ höhere } \text{———} \text{ " } \text{———} \end{cases}$$

unterschiedliche Legierungen

$-(a \underset{\in C^2}{u'})'$ nicht definiert

aber physikalisch relevante Fragestellung!

Ausweg: Ein schwacher Lösungsbegriff!

Wir multiplizieren die Differentialgleichung mit einer (Test-)Funktion $\varphi \in C_0^\infty(\Omega) = \{\varphi \in C^\infty(\overline{\Omega}) \mid \varphi|_{\partial\Omega} = 0\}$ und integrieren:

$$\underbrace{-\int_0^1 (a u')' \varphi \, dx}_{\substack{= \\ \text{partielle Integration}}} = \int_0^1 f \varphi \, dx$$

$$= -\int_0^1 a u' \varphi' \, dx + \underbrace{\left[a u' \varphi\right]_0^1}_{= 0 \text{ wegen } \varphi(0) = \varphi(1) = 0}$$

$$\Rightarrow \quad \underbrace{\int_0^1 a(x)\, u'(x)\, \varphi'(x)\, dx}_{\text{quadratische Form } a(u,\varphi)} = \underbrace{\int_0^1 f(x)\, \varphi(x)\, dx}_{\text{Linearform } \ell(\varphi)} \quad \text{(Variationsformulierung)}$$

<u>Vereinfachung:</u> zunächst $u^\partial = 0$ $(u(0) = u(1) = 0)$

<u>Aufgabe:</u> Gesucht ist $u \in V$, so dass

$$a(u, \varphi) = \ell(\varphi) \quad \forall \varphi \in V$$

<u>Ansatz:</u> $V = C_0^1([0,1]) = C^1([0,1]) \cap \{\varphi(0) = \varphi(1) = 0\}$

1.1. <u>Satz (Riesz)</u>

Sei V ein Hilbertraum (d.h. vollständiger metrischer Raum mit Skalarprodukt $(\cdot,\cdot)_V$ und Norm $\|u\|_V = \sqrt{(u,u)_V}$), $a(\cdot,\cdot)$ symm., beschränkte und koerzive Bilinearform auf V (d.h. $c\,\|u\|_V^2 \le a(u,u) \le C\,\|u\|_V^2$) und ℓ ein beschränktes, lineares Funktional auf V.
(d.h. $\|\ell\|_{V'} = \sup\limits_{\substack{u \in V \\ u \neq 0}} \dfrac{|\ell(u)|}{\|u\|_V} < \infty$, dann existiert genau eine Lösung $u \in V$ des Variationsproblems

$$a(u, \varphi) = \ell(\varphi) \quad \forall \varphi \in V.$$

<u>Anwendung:</u> $a(u,\varphi) = \int_0^1 a u' \varphi' \, dx \qquad \ell(\varphi) = \int_0^1 f \varphi \, dx$

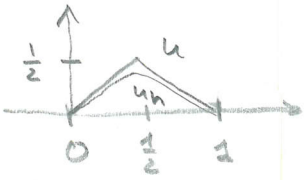
<u>Wahl des Skalarprodukts:</u> $(u,v)_V = \int_0^1 u'v' \, dx$

induzierte Norm $\|u\|_V = \int_0^1 |u'|^2 dx$ zunächst auf $C_0^1(\overline{\Omega})$

$(\|u\|_V = 0 \Rightarrow u' = 0$ fast überall $\Rightarrow u \equiv 0)$

<u>Aber:</u> $C_0^1(\overline{\Omega})$ ist nicht vollständig bzgl. $\|\cdot\|_V$

<u>Bsp</u>



$$u_n(x) = \sqrt{\frac{1}{4} + \frac{1}{n^2}} - \sqrt{\left(x - \frac{1}{2}\right)^2 + \frac{1}{n^2}}$$

$$u_n \xrightarrow[\text{in } \|\cdot\|_V]{n \to \infty} \frac{1}{2} - \left|x - \frac{1}{2}\right| \notin C^1([0,1])$$

<u>Ausweg:</u> Vervollständigung von $C_0^1(\overline{\Omega})$ unter der Norm $\|\cdot\|_V$!

Diesen Raum bezeichnen wir mit $H_0^1(\Omega)$ ← Differenzierbarkeit, ← Randwerte

Per Definition ist $V = H_0^1(\Omega)$ nun Hilbertraum

$\boxed{c \le a(x) \le C}$ $\Rightarrow$ $\int_0^1 a(x) u'(x) u'(x) dx \ge c \|u\|_V^2$

$$\int_0^1 a(x) u'(x) v'(x) dx \le C \int_0^1 u'(x) v'(x) dx$$

$$= C (u,v)_V \le C \|u\|_V \|v\|_V$$

Cauchy-Schwarz Ungleichung

$$|\ell(\varphi)| = \left| \int_0^1 f \varphi \, dx \right| \le \|f\|_{L^2(\Omega)} \underbrace{\|\varphi\|_{L^2(\Omega)}}_{\overset{!}{\le} C \|\varphi\|_V}$$

z.z.

4.1.2. <u>Satz</u> (Poincaré) Es gibt eine Konstante C_P, so dass für $u \in H_0^1((0,1))$ gilt

$$\|u\|_{L^2((0,1))} \le C_P \|u\|_V .$$

<u>Bew</u> zunächst sei $u \in C_0^1([0,1])$

$$u(x) = \int_0^x u'(\xi)\, d\xi \Rightarrow |u(x)| \le \int_0^x 1\, u'(x)\, d\xi$$

$$\underset{C.S.}{\le} \underbrace{\sqrt{\int_0^x 1^2\, d\xi}}_{\le 1} \sqrt{\int_0^x |u'(\xi)|^2 d\xi}$$

$$\Rightarrow \int_0^1 |u(x)|^2 \le \int_0^1 1 \int_0^x |u'|^2 d\xi\, dx$$

$$\le \int_0^1 \int_0^1 |u'(\xi)|^2 d\xi\, dx = 1\, \|u\|_V^2$$

$$\Rightarrow \|u\|_{L^2((0,1))} \le \|u\|_V \qquad (C_p \le 1) \qquad \square$$

Anwendung von 4.1.1. liefert eine schwache Lösung!

<u>allgemeiner Fall:</u> $u^\partial|_{\partial\Omega} \neq 0$ $\quad (u^\partial \in C^1([0,1]) : u^\partial(x) = (1-x)u(0) + x\, u(1))$

<u>Ansatz:</u> $w = u - u^\partial$

$$a(u,\varphi) = a(w,\varphi) + a(u^\partial,\varphi)$$

$\rightarrow$ <u>Schwache Formulierung:</u>

gesucht ist $w \in V$ mit $a(w,\varphi) = \ell(\varphi) := \int f\varphi\, dx - a(u^\partial,\varphi)$

$$|\ell(\varphi)| = \left| \int_0^1 f\varphi\, dx + \int_0^1 a(u^\partial)'\varphi'\, dx \right|$$

$$\le C_p \|f\|_{L^2((0,1))} \|\varphi\|_V + C \|u^\partial\|_V \|\varphi\|_V$$

$$= \left(C_p \|f\|_{L^2((0,1))} + C \|u^\partial\|_V\right) \|\varphi\|_V$$

$\Rightarrow$ Existenz von $w \in V$ mit $a(w,\varphi) = \ell(\varphi)$

$\Leftrightarrow$ $u = w + u^\partial$ schwache Lösung.

4.1.3. <u>Bsp</u> $f \equiv 0$ $\quad a = \begin{cases} a_0 \,;\, x \le \frac{1}{2} \\ a_1 \,;\, x > \frac{1}{2} \end{cases}$ $\quad u^D(0) = 0$, $u^D(1) = 1$

$a_0 = 1$, $a_1 = 2$

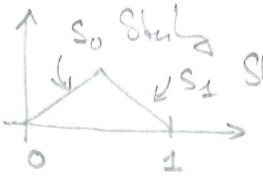
<u>Schwache Formulierung:</u> $\displaystyle\int_0^1 a\, w' \varphi' \, dx = -\int_0^1 a\, \varphi' \, dx$

$w(x) = u(x) - x$

$\varphi \in C_0^\infty((0,\frac{1}{2}))$ oder $C_0^\infty((\frac{1}{2},0)) \xrightarrow[\text{partielle Integration}]{}$

$\displaystyle\int_0^1 a\, w'' \varphi \, dx = 0 \underset{\forall \text{ solchen Testfunktion}}{\Longrightarrow} w'' = 0$ auf $(0,\frac{1}{2}) \cup (\frac{1}{2},1)$

<u>Ansatz:</u> w stückweise linear auf $[0,\frac{1}{2}]$ und $[\frac{1}{2},1]$ mit Steigungen s_0, s_1



Randwerte $\Longrightarrow s_0 + s_1 = 0$

$$\int_0^1 a\, w' \varphi' \, dx \underset{\substack{\text{partielle Integration} \\ \varphi \in C_0^1([0,1])}}{=} 0 + 0 + (a_0 s_0 - a_1 s_1)\varphi(\tfrac{1}{2})$$

$$\| $$

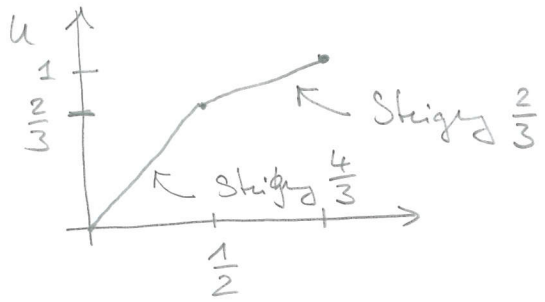
$$-\int_0^1 a\, \varphi' \, dx = 0 + 0 - (a_0 - a_1)\varphi(\tfrac{1}{2})$$

$$\Rightarrow \begin{pmatrix} a_0 s_0 - a_1 s_1 = a_1 - a_0 \\ s_0 + s_1 = 0 \end{pmatrix} \text{ lin. System für } s_0, s_1$$

$$s_1 = -s_0 \Rightarrow (a_0 + a_1) s_0 = a_1 - a_0 \Rightarrow s_0 = \frac{a_0 - a_1}{a_0 + a_1} = \frac{2-1}{1+2} = \frac{1}{3}$$

$$s_1 = \frac{a_1 - a_0}{a_0 + a_1} = -\frac{1}{3}$$

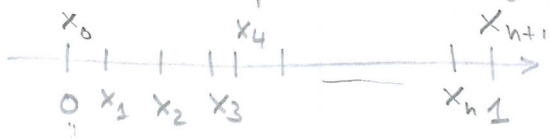
$\Rightarrow$



Nun betrachten wir die numerische Approximation:

$H_0^1((0,1))$ unendlich dimensional

Idee: Einschränkung auf endlich dimensionalen Teilraum von $H_0^1((0,1))$ basierend auf einer Zerlegung des Gebiets:



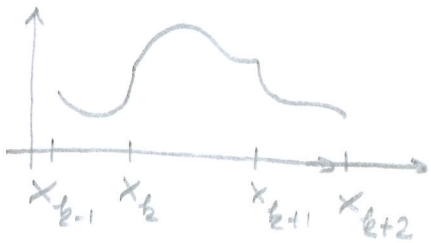
$0 = x_0 < x_1 < x_2 < ___ < x_n < x_{n+1} = 1$

$h = \max\limits_{i=1,\dots,n+1} |x_{i+1} - x_i|$ maximale Gitterweite

Funktionenraum: $V_h = \{ \varphi \in C_0^0((0,1)) \mid \varphi|_{[x_i, x_{i+1}]} \in \mathcal{P}_k \ \forall i = 0,\dots,n \}$

wobei $\mathcal{P}_k$ der Raum der Polynome vom Grad $\leq k$

vgl. Bsp oben $\rightsquigarrow$ $V_h \subset V = H_0^1((0,1))$



V_h nennt man Finite-Elemente-Raum

diskretes Variationsprb:

(E_h) Gesucht $w_h = u_h - u^\partial \in V_h$ mit $\underbrace{a(w_h, \varphi_h)}_{= \int_0^1 a\, w_h' \varphi_h' \, dx} = \underbrace{\ell(\varphi_h)}_{= \int_0^1 f \varphi_h \, dx - \int_0^1 a\, u^\partial \varphi_h' \, dx}$, $\forall \varphi_h \in V_h$

nun kann man nicht unbedingt $a(.,.)$, $\ell(.)$ exakt integrieren $\rightsquigarrow$

$(\hat{E}_h)$ Gesucht $w_h = u_h - u_h^\partial \in V_h$ mit

$a_h(w_h, \varphi_h) = \ell_h(\varphi_h) \quad \forall \varphi_h \in V_h$

wobei $a_h(v_h, w_h) = \int_0^1 a_h(x)\, v_h' w_h' \, dx$

$\ell_h(\varphi_h) = \int_0^1 f_h \varphi_h \, dx - \int_0^1 a_h u_h^\partial \varphi_h' \, dx$

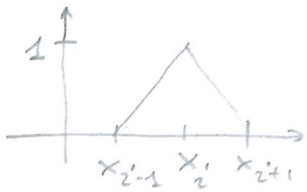
und a_h, f_h, u_h^∂ Approximationen von a, f, u^∂ $(c \le a_h \le C!)$

$u_h^\partial = u^\partial$ in 1D einfach möglich!

<u>Basisdarstellung der Lösung für $k=1$:</u> $w_h(x) = \sum_{i=1}^{n} w_h^i \cdot \varphi_h^i$

mit $\varphi_h^i \in V_h$ eindt. bestimmt durch $\varphi_h^i(x_j) = \delta_{ij}$

(Hütchenbasis) $\rightsquigarrow$ $w_h^i = w_h(x_i)$



Wegen Linearität von $(\hat{E}_h)$ genügt es mit der Basis zu testen, d.h.

$$\underbrace{a_h\Big(\sum_{i=1}^{n} w_h^i \varphi_h^i, \varphi_h^j\Big)}_{} = \underbrace{\ell_h(\varphi_h^j)}_{} \quad \forall j = 1,\dots,n$$

$$\sum_{i=1}^{n} \underbrace{\Big(\int_0^1 a_h(x)(\varphi_h^i)'(\varphi_h^j)'\,dx\Big)}_{a_h^{ji}} w_h^i = \underbrace{\int_0^1 f_h(x)\varphi_h^j\,dx - \int_0^1 a_h(x)(u^\partial)'\varphi_h^{j\prime}\,dx}_{=:b_h^j}$$

$$\Longleftrightarrow \quad \boxed{A_h \overline{w}_h = b_h}$$

$A_h = (a_h^{ji})_{j,i=1,\dots,n}$ $\quad b_h = (b_h^j)_{j=1,\dots,n}$

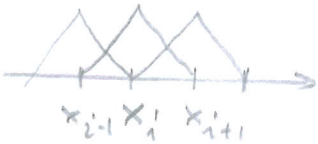
$\overline{w}_h = (w_h^i)_{i=1,\dots,n}$

lineares Gl. system

A_h nennt man <u>Steifigkeitsmatrix</u>

<u>Eigenschaften von A_h:</u>

- tridiagonal

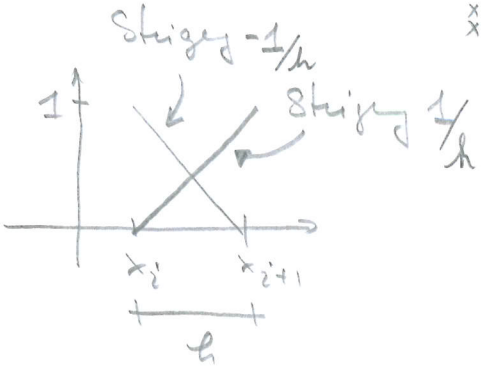


- symmetrisch und positiv definit

$$A_h \overline{v}_h \cdot \overline{v}_h = \sum_{i,j=1,\dots,n} a^{ij} v_h^i v_h^j = \int_0^1 a_h v_h' v_h'\,dx \ge c\,\|v_h'\|^2_{L^2(0,1)} > 0 \quad v_h \ne 0$$

<u>Bsp</u> $x_i = \frac{i}{n+1}$ $\quad h = \frac{1}{n+1}$ $\quad a \equiv a_h \equiv 1$

$$A_h = \frac{1}{h}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}$$



zur Berechnung von A_h, b_h:

- Assemblierung mittels einer Schleife über alle Elemente (Zellen) $[x_{i-1}, x_i]$ $(i = 1, \dots, n+1)$
- Für jedes Element betrachtet man die lokalen Basisfunktionen

$\varphi^k_{h,loc}$ $k = 1,2$

und berechnet alle lokalen Terme:

$$\int_{x_{i-1}}^{x_i} a_h(x)(\varphi^k_{h,loc})'(x)(\varphi^l_{h,loc})'(x)\,dx, \quad \int_{x_{i-1}}^{x_i} f_h(x)\varphi^k_{h,loc}(x) - a_h(x)(u^D)'(x)(\varphi^k_{h,loc})'(x)\,dx$$

- Man setzt die Matrix A_h und die rechte Seite b_h zusammen aus den lokalen Beiträgen (Assemblierung) (Initialisierung mit 0, elementweise Aufaddition basierend auf Zuordnung von lokalen zu globalen Freiheitsgraden)

nun betrachten wir den numerischen Fehler:

$\|u - u_h\|_V$?

kontinuierliches Pb: (E) $a(\overset{\in V}{u}, \varphi) = \ell(\varphi) \quad \forall \varphi \in V$

diskretes Pb: $(\hat{E}_h)$ $a_h(\underset{\in V_h}{u_h}, \varphi_h) = \ell_h(\varphi_h) \quad \forall \varphi_h \in V_h$

mit $V_h \subset V$ (konformer Finite-Elemente-Ansatz)

Voraussetzungen:

$$a(u,v) \le C \|u\|_V \|v\|_V$$
$$a(u,u) \ge c \|u\|_V^2, \quad a_h(u_h, u_h) \ge c \|u_h\|_V^2$$
$$|\ell(v)| \le C \|v\|_V$$

$\|f\|_{L^2}, \|f_h\|_{L^2}, \|a_h\|_{L^\infty}, \|u^D\|_V \le C$

zunächst für: $\boxed{u^\partial = 0,\ a_h = a,\ f_h = f}$

$$\|u-u_h\|_V^2 \le \frac{1}{c}\, a(u-u_h, u-u_h)$$

$$\overset{v_h \in V_h}{=} \frac{1}{c}\Big(a(u-u_h, u-v_h) + \underbrace{a(u-u_h, \overbrace{v_h-u_h}^{\in V_h})}_{= \ell(v_h-u_h) - \ell_{(h)}(v_h-u_h) \overset{(E),(E_h)}{=} 0}\Big)$$

$$\le \frac{C}{c}\|u-u_h\|_V \|u-v_h\|_V$$

$$\Rightarrow \|u-u_h\|_V \le \frac{C}{c}\|u-v_h\|_V \quad \forall v_h \in V_h$$

4.1.4. <u>Lemma</u> (Cea)

Unter obigen Voraussetzungen für (E), (E_h) und einen konformen Finite-Elemente-Ansatz gilt im Fall $u^\partial \equiv 0$, $a_h = a$, $f_h = f$:

$$\underbrace{\|u-u_h\|_V}_{\text{FE-Fehler}} \le \frac{C}{c} \inf_{v_h \in V_h} \underbrace{\|u-v_h\|_V}_{\text{Approximationsfehler}}.$$

$\boxed{\text{nun zum allgemeinen Fall mit } u^\partial = u_h^\partial}$ $\quad (u = w + u^\partial,\ u_h = w_h + u^\partial)$

$(\ell_{(h)}(\varphi) = \int_0^1 f_{(h)}\varphi\, dx)$

$$a(u-u_h, u_h-v_h) = \ell(u_h-v_h) - a(u_h, u_h-v_h)$$

$$= 0 + (\ell-\ell_h)(u_h-v_h) - (a-a_h)(u_h, u_h-v_h)$$

$$= \int_0^1 (f-f_h)(u_h-v_h)\,dx - \int_0^1 (a(x)-a_h(x))\, u_h(x)'(u_h'(x)-v_h'(x))\,dx$$

$$\underset{\text{Cauchy-Schwarz}}{\le} \|f-f_h\|_{L^2(\Omega)}\|u_h-v_h\|_{L^2(\Omega)} + \|a-a_h\|_{L^\infty(\Omega)}\|u_h\|_V\|u_h-v_h\|_V$$

$$\underset{\text{Poincaré}}{\le} \big(C_p\|f-f_h\|_{L^2(\Omega)} + \|a-a_h\|_{L^\infty(\Omega)}\|u_h\|_V\big)\underbrace{\|u_h-v_h\|_V}_{\le \|u-u_h\|_V + \|u-v_h\|_V}$$

$$\|w_h\|_V^2 \le \frac{1}{c}\, a_h(w_h, w_h) = \frac{1}{c}\int f_h w_h - a_h(u^\partial)' w_h'\,dx$$

$$\le \frac{1}{c}\Big(\|f_h\|_{L^2(\Omega)}\|w_h\|_{L^2(\Omega)} + \|a_h\|_{L^\infty(\Omega)}\|u^\partial\|_V\|w_h\|_V\Big)$$

$$\underset{\text{Poincaré}}{\leq} \frac{1}{c}\left(\|f_h\|_{L^2(\Omega)} C_P + \|a_h\|_{L^\infty(\Omega)} \|u^\partial\|_V\right) \|w_h\|_V$$

$$\Rightarrow \quad \|w_h\|_V \leq \frac{1}{c}\left(\|f_h\|_{L^2(\Omega)} C_P + \|a_h\|_{L^\infty(\Omega)} \underbrace{\|u^\partial\|_V}_{\leq C}\right)$$

$$\leq C$$

$$\Rightarrow \quad \|u_h\|_V \leq \|w_h\|_V + \|u^\partial\|_V \leq C^*$$

<u>Zusammenfassend erhalten wir:</u>

$$\|u-u_h\|_V^2 \underset{(s.o.)}{\leq} \frac{C}{c}\|u-u_h\|_V \|u-v_h\|_V$$

$$+ \frac{1}{c}\left(C_P \|f-f_h\|_{L^2(\Omega)} + C^* \|a-a_h\|_{L^\infty(\Omega)}\right) \cdot$$

$$\left(\|u-u_h\|_V + \|u-v_h\|_V\right)$$

$$= \overbrace{\left(\frac{C}{c}\|u-v_h\|_V + \frac{C_P}{c}\|f-f_h\|_{L^2(\Omega)} + \frac{C^*}{c}\|a-a_h\|_{L^\infty(\Omega)}\right)}^{A_1} \underbrace{\|u-u_h\|_V}_{=B_1}$$

$$+ \underbrace{\left(\frac{C_P}{c}\|f-f_h\|_{L^2(\Omega)} + C\|a-a_h\|_{L^\infty(\Omega)}\right)}_{=A_2} \underbrace{\|u-v_h\|_V}_{=B_2}$$

$$\boxed{\begin{aligned} &0 \leq (A_i - B_i)^2 = A_i^2 - 2A_iB_i + B_i^2 \\ &\Rightarrow \quad A_iB_i \leq \frac{A_i^2 + B_i^2}{2} \end{aligned}} \Longrightarrow \quad \left[\text{sowie } (A+B)^2 \leq 2A^2 + 2B^2\right]$$

$$\frac{1}{2}\|u-u_h\|_V^2 \leq C\left(\|u-v_h\|_V^2 + \|f-f_h\|_{L^2(\Omega)}^2 + \|a-a_h\|_{L^\infty(\Omega)}^2\right)$$

Damit folgt:

4.1.5. <u>Satz</u> (<u>abstrakte Fehlerabschätzungen</u>)

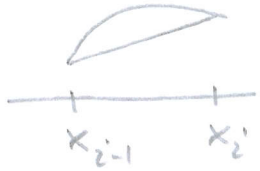
Unter obigen Voraussetzungen für (E), (E_h) gilt

$$\|u-u_h\|_V \leq C\Big(\underbrace{\inf_{v_h \in V_h} \|u-v_h\|_V}_{\text{Approximationsfehler}} + \underbrace{\|f-f_h\|_{L^2(\Omega)}}_{\text{Interpolationsfehler}} + \underbrace{\|a-a_h\|_{L^\infty(\Omega)}}_{\text{Interpolationsfehler}}\Big)$$

$\leq$ Interpolationsfehler

Nun wählen wir spezielle Funktionen v_h und a_h, f_h:

$\boxed{v_h}$ $\quad v_h(x_i) = u(x_i)$ Lagrange-Interpolation



Mittelwertsatz $\Rightarrow \exists\, \xi_i \quad u'(\xi_i) = v_h'(\xi_i)$

$$(u - v_h)'(x) = \int_{\xi}^{x} (u - v_h)''(y)\,dy = \int_{\xi}^{x} 1 \cdot u''(y)\,dy$$

$$\underset{\text{Cauchy Schwarz}}{\leq} \underbrace{\sqrt{\int_{x_{i-1}}^{x_i} 1^2\,dy}}_{\sqrt{x_i - x_{i-1}} \leq \sqrt{h}} \sqrt{\int_{x_{i-1}}^{x_i} u''(y)^2\,dy} \qquad h = \max_{i=1,\dots,n+1} |x_i - x_{i-1}|$$

$$\Rightarrow \|u - v_h\|_V^2 = \sum_{i=1}^{n+1} \int_{x_{i-1}}^{x_i} |(u - v_h)'(x)|^2\,dx$$

$$\leq \sum_{i=1}^{n+1} \int_{x_{i-1}}^{x_i} \left(h \int_{x_{i-1}}^{x_i} |u''(y)|^2\,dy \right) dx \leq \sum_{i=1}^{n+1} h^2 \int_{x_{i-1}}^{x_i} |u''(y)|^2\,dy$$

$$= h^2 \int_0^1 |u''(y)|^2\,dy$$

$$\Rightarrow \boxed{\|u - v_h\|_V \leq h\, \|u''\|_{L^2(\Omega)}}$$

$\boxed{f_h}$ $\quad \boxed{f_h\Big|_{[x_{i-1}, x_i)} = f\left(\frac{x_{i-1} + x_i}{2}\right)}$

Schwerpunktquadratur:

$$\int_{x_{i-1}}^{x_i} f\,\varphi_h\,dx \approx \underbrace{|x_i - x_{i-1}|\,(f\varphi_h)\left(\frac{x_{i-1}+x_i}{2}\right)}_{= \int_{x_{i-1}}^{x_i} f_h(x)\varphi_h(x)\,dx}$$

$$(f - f_h)(x) = \int_{\frac{x_{i-1}+x_i}{2}}^{x} \underbrace{(f - f_h)'}_{= f'}(y)\,dy$$

$$\Rightarrow \int_{x_{i-1}}^{x_i} (f - f_h)^2(x)\,dx \leq \int_{x_{i-1}}^{x_i} \left(\int_{x_{i-1}}^{x_i} 1 \cdot f'\,dy \right)^2 dx$$

$$\underset{\text{Cauchy Schwarz}}{\leq} \int_{x_{i-1}}^{x_i} |x_i - x_{i-1}| \int_{x_{i-1}}^{x_i} |f'|^2\,dy$$

$$\leq h^2 \int_{x_{i-1}}^{x_i} |f'|^2\,dx$$

$$\Rightarrow \|f - f_h\|_{L^2(\Omega)}^2 = \sum_{i=1}^{n+1} \int_{x_{i-1}}^{x_i} (f - f_h)^2(x)\,dx \leq h^2 \int_0^1 |f'|^2\,dx$$

$$\Rightarrow \|f - f_h\|_{L^2(\Omega)} \leq h \, \|f'\|_{L^2(\Omega)}$$

$\boxed{a_h}$ Erneut wählen wir die Schwerpunktintegration:

$$\int_{x_{i-1}}^{x_i} a \, v_h' \, w_h' \, dx \approx \underbrace{|x_i - x_{i-1}| \, (a \, \underset{\uparrow}{v_h'} \, \underset{\uparrow}{w_h'}) \left(\frac{x_{i-1}+x_i}{2}\right)}_{\text{konstant}}$$

$$= \int_{x_{i-1}}^{x_i} a_h \, v_h' \, w_h' \, dx$$

$$\boxed{a_h\big|_{[x_{i-1},x_i)} = a\left(\frac{x_{i-1}+x_i}{2}\right)}$$

$$|a(x) - a_h(x)| = \left| a(x) - a\left(\frac{x_{i-1}+x_i}{2}\right)\right|$$

$$\underset{\uparrow \, [x_{i-1},x_i)}{} \qquad \underset{\text{Mittelwertsatz}}{=} \left| a'(\underset{\uparrow \, [x_{i-1},x_i)}{\xi}) \left(x - \frac{x_{i-1}+x_i}{2}\right)\right|$$

$$\leq \|a'\|_{L^\infty(\Omega)} \, h$$

$$\Rightarrow \boxed{\|a - a_h\|_{L^\infty(\Omega)} \leq h \, \|a'\|_{L^\infty(\Omega)}}$$

wir erhalten:

4.1.6. <u>Satz</u> (konkrete Fehlerabschätzung)

Sei V_h der Raum der stückweise affinen, stetigen Funktionen auf einer Zerlegung von $\Omega = (0,1)$ mit $h = \max_{i=1,\dots,n+1} |x_i - x_{i-1}|$.

Sei $a \in C^1(\overline{\Omega})$, $f, f' \in L^2(\Omega)$,

$f_h\big|_{[x_{i-1},x_i)} = f\left(\frac{x_{i-1}+x_i}{2}\right)$, $a_h\big|_{[x_{i-1},x_i)} = a\left(\frac{x_{i-1}+x_i}{2}\right)$,

dann gilt falls die kontinuierliche Lösung u zweite Ableitungen in $L^2(\Omega)$ besitzt:

$$\|u - u_h\|_V \leq C \, h \left(\|u''\|_{L^2(\Omega)} + \|f'\|_{L^2(\Omega)} + \|a'\|_{L^\infty(\Omega)} \right)$$

(Konvergenz von 1. Ordnung)

<u>Bem</u>: a darf auf Knoten x_i $(i=1,\dots,n)$ springen!

Energieperspektive:

Betrachte $E[u] = \int_0^1 \frac{1}{2} a (u')^2 - f u \, dx$, $E: u^\partial + V \to \mathbb{R}$

notwendige Bedingung für $E[u] \leq E[\tilde{u}] \quad \forall \tilde{u} \in u + V$ ist dann:

$$\underbrace{\frac{d}{d\epsilon} E[u + \epsilon v]\Big|_{\epsilon=0}}_{} = 0 \quad \forall v \in V$$

$$\overset{\text{Glattheit}}{=} \int_0^1 \underbrace{\frac{d}{d\epsilon}\left(\frac{1}{2} a (u' + \epsilon v')^2 - f(u + \epsilon v)\right)}_{= a(u' + \epsilon v') v' - f v}\Big|_{\epsilon=0} dx$$

$$\Rightarrow 0 = \int_0^1 a u' v' - f v \, dx \quad \Leftrightarrow \text{ schwache Formulierung zu } -(a u')' = f \text{ auf } (0,1)$$

diskrete Variante: $E_h : u^\partial + V_h \to \mathbb{R}$; $E_h[u_h] = \int_0^1 \frac{1}{2} a_h (u_h')^2 - f u_h \, dx$

notwendige Bedingung folgt analog für Minimierer von E_h:

$$0 = \int_0^1 a_h u_h' v_h' - f_h v_h \, dx \quad \forall v_h \in V_h$$

in Matrix-Vektor-Notation ($A_h^{ext} = (a_h^{ij})_{i,j=0,\dots,n+1}$):

$$\frac{1}{2} A_h^{ext} \underbrace{(\overline{u}_h^\partial + \overline{w}_h)}_{= \overline{u}_h} \cdot (\overline{u}_h^\partial + \overline{w}_h) - b_h \cdot (\overline{u}_h^\partial + \overline{w}_h) \longrightarrow \min$$

notwendige Bedingung: $A_h \overline{w}_h = b_h \overline{w}_h - (A_h^{ext} \overline{u}_\partial)_{i=1,\dots,n}$

vgl. CG-Verfahren: $Ax = b$ A pos. def., symmetrisch

$$\Leftrightarrow \frac{1}{2} A x \cdot x - b x \leq \frac{1}{2} A \tilde{x} \cdot \tilde{x} - b \tilde{x}$$

Nun betrachten wir $d=2$:

$$u : \overline{\Omega} \to \mathbb{R} \; ; \; u|_{\partial\Omega} = u^\partial \quad (u^\partial : \overline{\Omega} \to \mathbb{R})$$

$$E[u] = \int_\Omega \frac{1}{2} a(x) \|\nabla u(x)\|^2 - f(x)\, u(x)\, dx$$

notwendige Bedingung für $E[u] \leq E[\tilde{u}] \quad \forall \tilde{u} \in u^\partial + V$ ist dann:

$$\frac{d}{d\epsilon} E[u+\epsilon v]\Big|_{\epsilon=0} = 0 \quad \forall v \in V$$

hierbei: V = Vervollständigung von $C_0^1(\overline{\Omega})$ bzgl. der Norm

$$\|u\|_V = \sqrt{\int_\Omega \|\nabla u\|^2 dx}$$

$$(*) \quad 0 = \int_0^1 \frac{d}{d\epsilon} \Big(\frac{1}{2} a(x) \underbrace{\|\nabla u(x) + \epsilon \nabla v(x)\|^2}_{\|\nabla u(x)\|^2 + 2\epsilon \nabla u(x) \cdot \nabla v(x) + \epsilon^2 \|\nabla v(x)\|^2} - f(x)\,(u(x) + \epsilon v(x)) \Big)\Big|_{\epsilon=0} dx$$

$$= \int_0^1 a(x) \nabla u(x) \nabla v(x) - f(x)\, v(x)\, dx$$

Zusammenhang mit einer Differentialgleichung:

Satz von Gauß: $\displaystyle\int_\Omega \operatorname{div} q \, dx = \int_{\partial\Omega} q \cdot n \, da$

nun $q = a(x) \nabla u(x)\, v(x) \Rightarrow$

$$\int \underbrace{\operatorname{div}\big(a(x) \nabla u(x)\, v(x)\big)}_{\substack{= \operatorname{div}(a(x)\nabla u(x))\, v(x) \\ + a(x) \nabla u(x) \nabla v(x)}} dx = \underbrace{\int_{\partial\Omega} a(x) \nabla u(x) \cdot n(x) \underbrace{v(x)}_{=0 \text{ auf } \partial\Omega} da}_{=0}$$

in Koordinaten:

$$\operatorname{div}(a \nabla u\, v) = \sum_{i=1}^{d} \partial_{x_i}(a\, \partial_{x_i} u\, v)$$
$$= \sum_{i=1}^{d} \partial_{x_i}(a\, \partial_{x_i} u) + a\, \partial_{x_i} u\, \partial_{x_i} v$$

Damit folgt aus ⊛:

$$0 = \int_0^1 -\operatorname{div}(a(x)\, \nabla u(x))\, v(x) - f(x)\, v(x)\, dx$$

$$\Rightarrow \quad -\operatorname{div}(a(x)\, \nabla u(x)) = f(x) \quad \forall x \in \Omega \qquad (E)$$

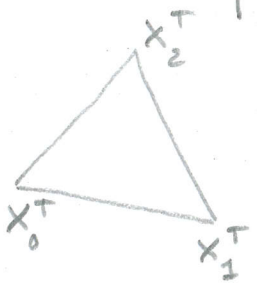
$\forall v \in C_0^1(\overline{\Omega})$

⊛ ist die schwache Formulierung oder Variationsformulierung.

Nun betrachten wir die numerische Approximation:

Für Ω polygonal sei $\mathcal{T}_h$ eine Triangulierung von Ω mit Gitterweite h, d.h.

$$\mathcal{T}_h = \{ T_i \mid i \in 1_m \} \text{ mit}$$



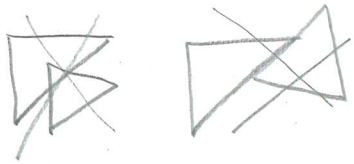
$$T = \{ x = \lambda_0 x_0^T + \lambda_1 x_1^T + \lambda_2 x_2^T \mid \lambda_i \geq 0,\ \sum_{i=0}^{2} \lambda_i = 1 \}$$

und $x_0^T, x_1^T, x_2^T \in \mathbb{R}^2$ mit $\det\begin{pmatrix} | & | \\ x_1 - x_0 & x_2 - x_0 \\ | & | \end{pmatrix} \neq 0$

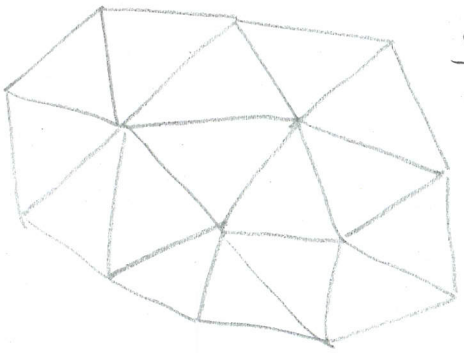
Dreieck mit Ecken x_0^T, x_1^T, x_2^T.

$T_i \cap T_j$ sei entweder leer oder eine gemeinsame Kante oder eine gemeinsame Ecke

$$\overline{\Omega} = \bigcup_{i=1}^{m} T_i$$



$$h = \max_{i=1_m} \max_{k,j=0,1,2} |x_j^{T_i} - x_k^{T_i}|$$



Ω mit Triangulierung $\mathcal{T}_h$

Nun betrachten wir Funktionen v_h, die stückweise affin auf den Dreiecken sind und auf $\overline{\Omega}$ stetig. V_h sei der Raum dieser Funktionen.

Bem wie in 1D gilt $V_h \subset V$ (ohne Beweis hier!)

Insgesamt erhalten wir:

kont. Variationsproblem:

Gesucht u mit $u - u^\partial \in V$ und

$$a(u, \varphi) = \ell(\varphi) \qquad \forall \varphi \in V \qquad (E)$$

diskretes Variationsproblem:

Gesucht u_h mit $u_h - u_h^\partial \in V_h$ und

$$a_h(u_h, \varphi_h) = \ell(\varphi_h) \qquad \forall \varphi_h \in V_h \qquad (\hat{E}_h)$$

hierbei u_h^∂ stückweise affine, stetige Fkt, die auf $\partial\Omega$ u^∂ approximiert und

$$a_{(h)}(u, v) = \int_\Omega a_{(h)}(x) \nabla u(x) \cdot \nabla v(x) - f_{(h)}(x) v(x) \, dx$$

$$\ell_{(h)}(v) = \int_\Omega f_{(h)}(x) v(x) \, dx$$

(man kann man wie oben Approximationen a_h, f_h betrachten)

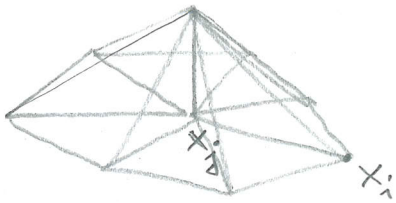
Basisdarstellung der Lösung:

$$w_h = u_h - u_h^\partial$$

$$w_h(x) = \sum_{i=1}^{n} w_h^i \varphi_h^i$$ wobei $\varphi_h^i \in V_h$ eindeutig bestimmt durch $\varphi_h^i(x_j) = \delta_{ij}$ (Hütchenbasis)

i: Knotenindex

n: Anzahl Gitterknoten



$$w_h^i = w_h(x_i)$$

wegen der Linearität von $(\widehat{E}_h)$ genügt es mit der Basis zu testen, d.h.

$$\underbrace{a_h\Big(\sum_{i=1}^{n} w_h^i \varphi_h^i, \varphi_h^j\Big)}_{} = \underbrace{\ell(\varphi_h^j)}_{} \quad \forall_{j=1,\dots,n}$$

$$\sum_{i=1}^{n} \underbrace{\Big(\int_\Omega a_h \nabla\varphi_h^i \cdot \nabla\varphi_h^j \, dx\Big)}_{=:\, a_h^{ji}} w_h^i = \underbrace{\int_\Omega f_h \varphi_h^j \, dx - \int_\Omega a_h \nabla u_h^\partial \cdot \nabla\varphi_h^j \, dx}_{=:\, b_h^j}$$

$$\Leftrightarrow \quad \boxed{A_h \overline{w}_h = b_h}$$

$$A_h = (a_h^{ji})_{j,i=1,\dots,n} \qquad b_h = (b_h^j)_{j=1,\dots,n}$$

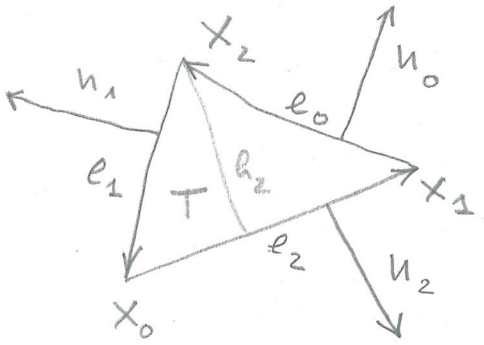
$$\overline{w}_h = (u_h^j)_{j=1,\dots,n}$$

Eigenschaften von A_h:

- dünn besetzt $a_h^{ji} \neq 0$ falls $j=i$ oder x_i, x_j eine Kante verbindet
- symmetrisch, pos. def. (falls $a_h \geq c > 0$)

$$A_h \overline{v}_h \cdot \overline{v}_h = \int_\Omega \underbrace{a_h}_{\geq c > 0} \nabla v_h \cdot \nabla v_h \geq \underbrace{c \int_\Omega |\nabla v_h|^2 \, dx}_{= c \, \|v_h\|_V^2}$$

Berechnung von A_h über lokale Assemblierung:



n_i: Normale auf Kante e_i gegenüber x_i

$$\nabla \varphi_h^i = -\frac{n_i}{h_i}$$

h_i: Höhe von T über e_i

$$a^{ij}_{h,loc} = |T| \frac{n_i \cdot n_j}{h_i h_j} a(S_T)$$

$$= |T| \frac{e_i \cdot e_j \; a(S_T)}{\underbrace{h_i \|e_i\|}_{2|T|} \underbrace{h_j \|e_j\|}_{2|T|}} = \frac{e_i \cdot e_j \; a(S_T)}{4|T|}$$

$$S_T = \frac{x_0 + x_1 + x_2}{3} \quad \text{(Schwerpunkt von } T\text{)}$$

$$a_h|_T \equiv a(S_T)$$

Berechnung von b_h:

- $f_h|_T = f(S_T) \Rightarrow \int_T f_h \varphi_h^i \, dx = \frac{1}{3} |T| f(S_T)$

- $a_h(u_h^\partial, \varphi_h^j)$ wird genauso berechnet wie die Einträge der Steifigkeitsmatrix, nur dass man hier auch Randknoten mit berücksichtigt.

<u>gesamt</u>: (1.) lokale Assemblierung der Matrix A_h^{ext} (bzgl. aller Knoten (s.o.)

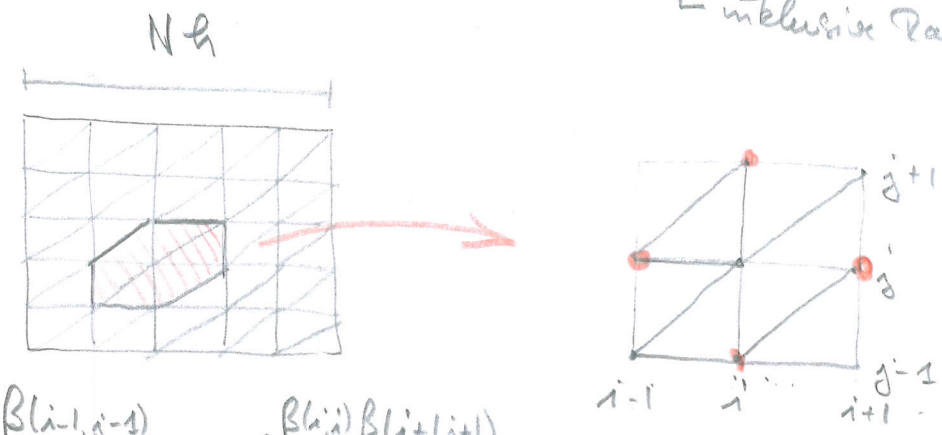
(2.) lokale Assemblierung von $\left(\int_\Omega f_h \varphi_h^j dx\right)_{j=1 \dots m}$

3. $$b_h = \left(\int_\Omega f_h \varphi_h^i\right)_{i=1,\dots,n} - \left(A_h^{ext}\, \overline{u}_h^j\right)_{i=1,\dots,n}$$

↑ inklusive Randknoten

<u>Bsp</u>

$a \equiv 1$



$$a_h^{\beta(i,j),\beta(i-1,j-1)} = a_h^{\beta(i,j),\beta(i+1,j+1)} = 0$$

$$a_h^{\beta(i,j),\beta(i-1,j+1)} = a_h^{\beta(i,j),\beta(i+1,j-1)} = 0$$

$$a_h^{\beta(i,j),\beta(i,j\pm1)} = -\frac{h^2}{4\,\frac{h^2}{2}} - \frac{h^2}{4\,\frac{h^2}{2}} = -1, \quad a_h^{\beta(i,j)\beta(i,j)} = 4$$

$$\| $$

$$a_h^{\beta(i,j),\beta(i\pm1,j)}$$

wobei $\beta(i,j)$ = Index zum Punkt (i,j)

$$\Rightarrow \quad A_h = \begin{pmatrix} \boxed{D} & -\mathbb{1} & & \\ -\mathbb{1} & \boxed{D} & -\mathbb{1} & \\ & -\mathbb{1} & \ddots & \ddots \\ & & \ddots & \ddots \end{pmatrix} \qquad D = \begin{pmatrix} 4 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{pmatrix}$$

vgl. Finite Differenzendiskretisierung (ALMA II)

<u>Bem</u>: • Falls Dreieckswinkel $\leq \frac{\pi}{2}$, dann hat A_h nicht positive Nebendiagonaleinträge und strikt positiven Diagonaleintrag.

5. Lineare Gleichungssysteme – iterative Lösu

(vgl. ALMA II)

5.1. <u>Def</u> (M-Matrix)

$A \in \mathbb{R}^{I,I}$, I Indexmenge, heißt M-Matrix, falls für $A = (a_{ij})_{i,j \in I}$ gilt:

$$a_{ii} > 0 \ \forall_{i \in I} \quad ; \quad a_{ij} \leq 0 \ \forall_{i \neq j}$$

$$A \text{ invertierbar}, \ A^{-1} \geq 0$$

$$(A \geq B :\Leftrightarrow a_{ij} \geq b_{ij} \ \forall_{i,j \in I})$$

5.2. <u>Def</u> (Matrixeigenschaften)

- Für $A \in \mathbb{R}^{I,I}$ ist $i \in I$ mit $j \in I$ verbunden ($j \in I_i$), falls es eine Kette $i = i_1, i_2, \ldots, i_k = j$ gibt mit $a_{i_{\ell-1}, i_\ell} \neq 0$
- A heißt irreduzibel falls i mit j verbunden ist $\forall_{i,j \in I}$
- A heißt (schwach) diagonal-dominant, falls

$$\sum_{j \neq i} |a_{ij}| \underset{(-)}{<} |a_{ii}| \quad \forall_{i \in I}$$

$$\left(\text{und } \forall_{i \in I} \ \exists_{k \in I_i} \text{ mit } \sum_{j \neq k} |a_{kj}| < |a_{kk}|\right)$$

5.3. <u>Lemma</u> (Gerschgorin)

$A = (a_{ij})_{i,j \in I} \in \mathbb{C}^{I,I}$, dann gilt für jeden Eigenwert $\lambda \in \mathbb{C}$ von A:

(i) $$\lambda \in \bigcup_{i \in I} \overline{B_{r_i}}(a_{ii}) \quad \text{mit } r_i = \sum_{j \neq i} |a_{ij}|$$

(ii) $$\lambda \in \bigcup_{i \in I} \left(B_{r_i}(a_{ii}) \cup \left(\bigcap_{j \in I_i} \partial B_{r_j}(a_{jj}) \right) \right)$$

<u>Bew</u> $\boxed{\text{zu (i)}}$ $Ax = \lambda x$ mit $\|x\|_\infty = 1$

$$\Leftrightarrow \quad (a_{ii} - \lambda) x_i = - \sum_{i \neq j} a_{ij} x_j \quad \forall i \in I$$

$$\overset{|x_k| = 1}{\Longrightarrow} \quad |a_{kk} - \lambda| \underbrace{|x_k|}_{=1} \leq \sum_{k \neq j} |a_{kj}| \underbrace{|x_j|}_{\leq 1} \leq r_k \quad \Rightarrow \text{Beh}$$

$\boxed{\text{zu (ii)}}$ $\lambda \notin \bigcup_{i \in I} B_{r_i}(a_{ii})$, $Ax = \lambda x$, $\|x\|_\infty = 1$, $|x_k| = 1$

$$\Rightarrow |a_{kk} - \lambda| \underset{\text{s.o.}}{\leq} r_k \quad \text{wegen Annahme „} = r_k \text{"}$$

$$\Rightarrow |x_j| = 1 \quad \forall j \text{ mit } a_{kj} \neq 0$$

$$\xrightarrow{\text{iterativ}} |x_j| = 1 \quad \forall j \in I_k \quad \Rightarrow |a_{jj} - \lambda| = r_j \quad \forall j \in I_k$$

$\square$

<u>Spalten wie $A \in \mathbb{R}^{I,I}$ nun auf:</u>

$$A = D - B \qquad D = (a_{ij} \delta_{ij})_{i,j \in I} \qquad B = D - A$$

<u>Bem:</u> $A = A_h$ aus Kapitel 4:

$$D, B \geq 0$$

5.4 <u>Satz</u> (M-Matrix)

$A \in \mathbb{R}^{I,I}$, $a_{ii} > 0$, $a_{ij} \underset{i \neq j}{\leq} 0$, dann ist A genau dann eine M-Matrix, wenn

$$\rho(D^{-1} B) < 1 ;$$

insbesondere gilt dies falls A schwach diagonal dominant ist.

<u>Bew</u> (i) zunächst zeigen wir (Neumann Reihe)

$$\rho(A) < 1 \Rightarrow A^k \underset{k \to \infty}{\longrightarrow} 0 \quad \text{und} \quad (\mathbb{1} - A)^{-1} = \sum_{k=0}^{\infty} A^k$$

vollständiger Beweis hierzu: vgl. ALMA II

$\boxed{\text{bis nur für } A \text{ symmetrisch!}}$ $\qquad n = \#I$

$\hookrightarrow$ $\quad A = QDQ^T \qquad Q \in O(n)\,,\ D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix},\ |\lambda_i| < 1 \ \forall i = 1,\dots,n$

$$A^k = QD^kQ^T \xrightarrow[k\to\infty]{} 0 \ (\checkmark)$$

$$\sum_{k=0}^{\infty} A^k = Q \sum_{k=0}^{\infty} D^k Q^T = Q\tilde{D}Q^T \quad \text{mit } \tilde{D} = \begin{pmatrix} (1-\lambda_1)^{-1} & & \\ & \ddots & \\ & & (1-\lambda_n)^{-1} \end{pmatrix}$$

Satz über geom. Reihe

$$= (\mathbb{1} - A)^{-1}$$

(ii) z.z. schwach diagonal dominant $\overset{!}{\Rightarrow}$ $\varrho(D^{-1}B) < 1$:

$$D^{-1}B = (c_{ij})_{i,j\in I} \quad \text{mit} \quad c_{ii} = 0$$

$$c_{ij} = -\frac{a_{ij}}{a_{ii}} \quad i \neq j$$

$$r_j = \sum_{i\neq j} \frac{|a_{ij}|}{|a_{ii}|} \leq 1 \quad \text{„<" für ein } j \in I_i \xrightarrow{\text{Gerschgorin}} \text{Beh}$$

(iii) z.z. $\varrho(D^{-1}B) < 1 \overset{!}{\Rightarrow}$ A M-Matrix:

$$C = D^{-1}B \qquad S = \sum_{k=0}^{\infty} \underbrace{C^k}_{\geq 0} \underset{(i)}{=} (\mathbb{1} - D^{-1}B)^{-1} \Rightarrow S \geq 0$$

$$A^{-1} = (D-B)^{-1} = \underbrace{(\mathbb{1} - D^{-1}B)^{-1}}_{\geq 0} \underbrace{D^{-1}}_{\geq 0} \geq 0$$

(iv) z.z. A M-Matrix $\overset{!}{\Rightarrow}$ $\varrho(D^{-1}B) < 1$:

λ EW zu $D^{-1}B$ zu EV $x \neq 0$ $\Rightarrow$

$$|\lambda| \underbrace{|x|}_{(|x_i|)_{i\in I}} = |\lambda x| = |D^{-1}Bx| \leq \underbrace{D^{-1}}_{\geq 0} \underbrace{B}_{\geq 0} |x|$$

Ferner: $A^{-1}D \geq 0 \Rightarrow -A^{-1}DD^{-1}B|x| \leq -A^{-1}D|\lambda||x|$

$$\Rightarrow |x| = A^{-1}(D-B)|x| = A^{-1}D(\mathbb{1} - D^{-1}B)|x|$$

$$\leq A^{-1}D|x| - A^{-1}D|\lambda||x| = (1 - |\lambda|)\, A^{-1}D|x|$$

also: $|\lambda| \geq 1 \Rightarrow |x| \leq 0 \Rightarrow x = 0$ ↯ $\square$

Nun leiten wir zwei iterative Gleichungslösungen her:

$$Ax = b \qquad A \in \mathbb{R}^{I,I}, \quad b \in \mathbb{R}^I$$

Zerlegung: $A = D - B$ (s.o.) $\Rightarrow$

$$Ax = b \Leftrightarrow Dx = Bx + b \underset{a_{ii} \neq 0}{\Longrightarrow} x = D^{-1}Bx + D^{-1}b$$

5.5. Def (Jacobi-Verfahren / Gesamtschrittverfahren)

x_0 Startwert $\in \mathbb{R}^I$

berechne Folge $(x^k)_{k \in \mathbb{N}}$ mit $x^{k+1} = D^{-1}Bx^k + D^{-1}b$

Bem: $C = D^{-1}B$

$$x^1 = Cx^0 + D^{-1}b$$

$$x^2 = C^2x^0 + (C + \mathbb{1})D^{-1}b$$

$$x^k = \underbrace{C^k x^0}_{\xrightarrow{k\to\infty} 0} + \underbrace{\left(\sum_{j=0}^{k-1} C^j\right)}_{\longrightarrow (\mathbb{1} - D^{-1}B)^{-1}} D^{-1}b$$

$$\Rightarrow \quad x^k \xrightarrow[k\to\infty]{} (\mathbb{1} - D^{-1}B)^{-1}D^{-1}b = (D-B)^{-1}b = A^{-1}b$$

D.h. das Jacobi-Verfahren konvergiert für M-Matrizen

5.6. Algorithmus

```
while (¬ Abbruchkriterium) {
    for (i=1; i≤n; i++)
        y_i = 1/a_ii ( Σ_{j≠i} − a_ij x_j + b_i );
    x = y;
}
```

<u>Bem</u>: Legt man keinen zweiten Vektor y an, sondern verwendet die bereits „verbesserten" Werte x_j mit $j < i$, dann erhält man das Gauß-Seidel-Verfahren.

<u>Genauer</u>:

$$A = D - E - F \qquad D = (a_{ij}\delta_{ij})_{i,j \in I}$$

$$E = (e_{ij})_{i,j \in I} \qquad e_{ij} = \begin{cases} -a_{ij} ; & j < i \\ 0 ; & \text{sonst} \end{cases}$$

$$F = (f_{ij})_{i,j \in I} \qquad f_{ij} = \begin{cases} -a_{ij} ; & j > i \\ 0 ; & \text{sonst} \end{cases}$$

$$Ax = b \Leftrightarrow \quad Dx = Ex + Fx + b$$

<u>Verfahren</u>:

$$x^{k+1} = D^{-1}(Ex^{k+1} + Fx^k + b)$$

oder: $(D - E)x^{k+1} = Fx^k + b$

5.7 <u>Def</u> (Gauß-Seidel-Verfahren, Einzelschrittverfahren)
I angeordnete Indexmenge, $x_0 \in \mathbb{R}^I$ Startwert
berechne Folge $(x^k)_{k \in \mathbb{N}}$ mit $x^{k+1} = (D-E)^{-1}Fx^k + (D-E)^{-1}b$

<u>Bem</u>: • Im Algorithmus (vgl. 5.6.)

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(\sum_{j<i} -a_{ij}\, x_j^{(k+1)} + \sum_{j>i} -a_{ij}\, x_j^{(k)} + b_i\right)$$

fehlen im Algorithmus

• Operation wird direkt auf dem Lösungsvektor ausgeführt!

• Anordnung ist wichtig beim Gauß-Seidel-Verfahren

→ <u>Betrachten wir nun allgemeine Iterationsverfahren</u>:

5.8. <u>Def</u> (<u>Iterationsverfahren zum Lösen lin. Gl. Systeme</u>)

$A \in \mathbb{R}^{I,I}$, $b \in \mathbb{R}^I$, $x \in \mathbb{R}^I$ gesucht mit $Ax = b$; A sei regulär

Ein <u>Iterationsverfahren</u> berechnet zu einem Startwert x^0 eine Folge $(x^k)_{k \geq 0}$ mit $\quad x^{k+1} = \Phi(x^k)$

[x^k <u>Iterierte</u>, x^* mit $x^* = \Phi(x^*)$ <u>Fixpunkt</u>]

Die Iteration <u>konvergiert</u>, falls $\lim_{k \to \infty} x^k$ existiert.

Ein Fixpunkt heißt konsistent zum Gl.S. $Ax = b$, falls

$$x^* \text{ Fixpkt} \Leftrightarrow Ax^* = b$$

5.9. <u>Normalformen linearer Iterationen</u>:

Ein Iteration heißt linear, falls Φ linear in x, b:

$$\Phi(x) = Mx + Nb$$

Eine lineare, konsistente Iteration kann geschrieben werden als:

$$\Phi(x) = x - N(Ax - b) \qquad (*)$$

oder für invertierbares N ($W = N^{-1}$):

$$W(x - \phi(x)) = Ax - b$$

<u>Beweis zu (*)</u>: zu $\bar{x}$ definiere $A\bar{x} = \bar{b} \Rightarrow$ $\bar{x}$ Fixpkt zu $\bar{b}$

$$\bar{x} = M\bar{x} + N\underset{\underset{A\bar{x}}{\parallel}}{\bar{b}} \underset{\forall \bar{x}}{\Rightarrow} \mathbb{1} = M + NA \Rightarrow$$

$$\mathbb{1} - M = NA \quad \text{und damit}$$

$$\phi(x) = Mx + Nb = x - (\mathbb{1} - M)x + Nb = x - NAx + Nb$$

$\square$

<u>Für Jacobi-Verfahren</u>: $\Phi(x) = M^J x + N^J b \quad M^J = D^{-1}B, N^J = D^{-1}$

<u>Für Gauß-Seidel-Verfahren</u>: $\Phi(x) = M^G x + N^G b$ mit

$$M^G = (D - E)^{-1} F, \quad N^G = (D - E)^{-1}$$

wäre $D = \alpha \mathbb{1}$; $\alpha \in \mathbb{R}$ dann ergäbe sich das Jacobi-Verfahren:

$$\Phi(x) = x - \Theta(Ax - b) \quad \text{für } \Theta = \frac{1}{\alpha}$$

5.10. <u>Def</u> (<u>Richardson-Iteration</u>)

$x^0 \in \mathbb{R}^I$ Startwert, $\Theta \in \mathbb{R}^+$, $A \in \mathbb{R}^{I,I}$, dann heißt

$$x^{k+1} = x^k - \Theta(Ax^k - b) \quad \text{Richardson Iteration}$$

(Θ Dämpfungsfaktor) ($M_\Theta^R = (\mathbb{1} - \Theta A)$, $N_\Theta^R = \Theta \mathbb{1}$)

<u>Bem</u>:
- Richardson-Iteration ist ein Gradientenschritt mit fester Schrittweite (vgl. ALHA)
- gedämpftes Jacobi-Verfahren: $x^{k+1} = x^k - \Theta D^{-1}(Ax^k - b)$ mit $\Theta \in (0,1)$
- Relaxations beim Gauß-Seidel-Verfahren:

$$x_i^{k+1} = x_i^k - \omega \frac{1}{a_{ii}} \left(\sum_{j<i} a_{ij} x_j^{k+1} + \sum_{j \geq i} a_{ij} x_j^k - b_i \right)$$

<u>Normalform</u>: $L := D^{-1}E \quad U := D^{-1}(D-F)$

$$(\mathbb{1} - \omega L) x^{k+1} = (\mathbb{1} - \omega U) x^k + \omega D^{-1} b$$

$$\Rightarrow x^{k+1} = M_\omega^{SOR} x^k + N_\omega^{SOR} b \quad \text{mit} \quad M_\omega^{SOR} = (\mathbb{1} - \omega L)^{-1}(\mathbb{1} - \omega U)$$

$$N_\omega^{SOR} = \omega (\mathbb{1} - \omega L)^{-1} D^{-1}$$

$$= \omega (D - \omega E)^{-1}$$

$$W_\omega^{SOR} = \left(\frac{D}{\omega} - E\right)$$

„Successive over relaxation" $\hat{=}$ SOR

($1 \leq \omega < 2$ sinnvoll)

<u>nun betrachten wir die Konvergenz der linearen Iterationsverfahren:</u>

<u>Notation</u> $e^k = x^k - x$ (Fehler) $\quad x$ Lsg. zu $Ax = b$

$d^k = Ax^k - b$ (Residuum)

unter Normalform $\Longrightarrow$

$$e^{k+1} = Mx^k + Nb - x \overset{x \text{ Fixpkt}}{=} Mx^k - Nb - Mx + Nb$$

$$= Me^k = ___ = M^k e^0$$

Form: $Ae^k = Ax^k - Ax = Ax^k - b = d^k$

5.11. Satz Ein lineares, konsistentes Iterationsverfahren konvergiert gdw. $\rho(M) < 1$ ($\rho(M)$ Konvergenzrate)

Bew: „$\Leftarrow$" $e^k = \underbrace{M^k}_{\xrightarrow[k\to\infty]{} 0 \text{ falls } \rho(M)<1 \text{ (vgl. Neumannsche Reihe)}} e^0$

„$\Rightarrow$" Sei $x^0 - x$ EV zu $\lambda = \rho(M) \Rightarrow$

$e^k = M^k e^0 = \lambda^k e^0 \longrightarrow 0 \Leftrightarrow |\lambda| < 1$ □

Bem: A M-Matrix $\Rightarrow \rho(D^{-1}B) < 1 \underset{5.11.}{\Rightarrow}$ Jacobi-Verfahren konv.

5.12 Satz (Konvergenz der Richardson-Iteration)

A pos. def., symm., dann konvergiert die Richardson Iteration gdw.

$$0 < \Theta < \frac{2}{\lambda_{max}(A)} \quad (\lambda_{max}(A) = \rho(A))$$

Die Konvergenzrate ist optimal für $\Theta^* = \dfrac{2}{\lambda_{max}(A) - \lambda_{min}(A)}$

und $\rho(M^R_{\Theta^*}) = \dfrac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}}$

Bew $M^R_\Theta = (\mathbb{1} - \Theta A) \Rightarrow$

$\sigma(M^R_\Theta) \subset [1 - \Theta\lambda_{max}, 1 - \Theta\lambda_{min}] \Rightarrow$

$\rho(M^R_\Theta) = \max\{|1 - \Theta\lambda_{min}|, |1 - \Theta\lambda_{max}|\}$

„$\Leftarrow$" $0 < \Theta < \dfrac{2}{\lambda_{max}} \Rightarrow -1 \leq 1 - 2 < 1 - \Theta\lambda_{max}$

$\leq 1 - \Theta\lambda_{min} < 1 \Rightarrow \rho(M^R_\Theta) < 1$

optimal:

falls $1 - \Theta^*\lambda_{min} = -(1 - \Theta^*\lambda_{max}) \Rightarrow \Theta^*(\lambda_{max} + \lambda_{min}) = 2$

$\Rightarrow \Theta^* = \dfrac{2}{\lambda_{max} + \lambda_{min}} \Rightarrow \rho(M^R_{\Theta^*}) = 1 - \Theta\lambda_{min} = \dfrac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}}$ □

zur Vorbereitung der weiteren Konvergenz-Diskussion:

5.13 Lemma $A \geq 0$, dann ist $\varrho(A)$ Eigenwert von A und es gibt einen nichtnegativen Eigenvektor.

Bew $K = \{ (x, \varrho) \mid x \in \mathbb{R}^I, \|x\|_\infty = 1, x \geq 0, \varrho \geq 0, Ax \geq \varrho x \}$

kompakt (abgeschlossen und beschränkt)

$\Rightarrow \tau = \max \{ \varrho \mid \exists x$ so dass $(x, \varrho) \in K \}$ wird angenommen

$Ax > \tau x$ kann nicht gelten, sonst wäre ϱ nicht maximal

$\boxed{\text{zunächst } A > 0}$ z.z. $\tau > 0$, $\exists y \geq 0$ $\|y\|_\infty = 1$ $Ay = \tau y$

hierzu: $(y, \tau) \in K \Rightarrow z = Ay - \tau y \geq 0$

angenommen $z \neq 0 \Rightarrow 0 < Az = A(Ay - \tau y)$

$= (A - \tau \mathbb{1}) Ay$

$\Rightarrow$ Für $\tilde{y} = \dfrac{Ay}{\|Ay\|_\infty}$ (auswertbar da $A > 0$)

gilt $0 < A\tilde{y} - \tau \tilde{y}$

$\Rightarrow A\tilde{y} > \tau \tilde{y}$ ↯ zu τ maximal

$\Rightarrow z = 0 \Rightarrow A\tilde{y} = \tau \tilde{y}$ mit $\tilde{y} = \dfrac{Ay}{\|Ay\|_\infty}$

damit folgt aber $\tau = \varrho(A)$, denn falls $Ax = \lambda x$, $|\lambda|$ maximal, $\lambda \in \mathbb{C}$

$\Rightarrow A|x| \geq |Ax| = |\lambda| \, |x| \Rightarrow |\lambda| \leq \tau \underset{\tau \text{ Eigenwert}}{\Longrightarrow} |\lambda| = \tau$

$\boxed{\text{nun } A \geq 0}$ Betrachte $A_\epsilon = (a_{ij} + \epsilon)_{i,j \in I}$:

$A_\epsilon > 0 \Rightarrow \exists\, (\lambda_\epsilon, x_\epsilon)$ mit $\lambda_\epsilon = \varrho(A_\epsilon)$; $x_\epsilon \geq 0$;

$$A_\epsilon x_\epsilon = \lambda_\epsilon x_\epsilon$$

$\boxed{\epsilon \to 0}$ Eigenwerte, Eigenvektoren stetig (wählbar) in ϵ $\Rightarrow$

$x_\epsilon \to x$, $\lambda_\epsilon \to \lambda = \varrho(A)$, $Ax = \lambda x$, $|x| \geq 0$ $\square$

5.14. <u>Satz</u> (Konvergenz bei Aufspaltung)

Zu $A = W - R$ definiere die Iteration $W x^{k+1} = R x^k + b$

für W regulär; $W^{-1} \geq 0$; $W^{-1}R \geq 0$ (insbesondere wenn $W \geq A$)

Falls A M-Matrix, dann konvergiert das Iterationsverfahren

und

$$\varrho(W^{-1}R) = \frac{\varrho(A^{-1}R)}{1 + \varrho(A^{-1}R)} < 1 .$$

<u>Bew</u> (i) Falls $W \geq A \Rightarrow R = W - A \geq 0 \underset{W^{-1} \geq 0}{\Longrightarrow} W^{-1}R \geq 0$

(ii) <u>Iterationsmatrix $M = W^{-1}R$</u> :

$$0 \leq M = W^{-1}R = (A^{-1}W)^{-1} A^{-1} R$$

$$= (A^{-1}(A+R))^{-1} A^{-1} R = (\mathbb{1} + C)^{-1} C'$$

mit $C' = A^{-1}R$

5.13 $\Rightarrow$ Es gibt $0 \neq x \geq 0$ EV zum EW $\lambda = \varrho(M) \geq 0$

$\lambda x = Mx = (\mathbb{1} + C)^{-1} C' x \Rightarrow \lambda x + \lambda C' x = C' x$

$\Rightarrow \lambda x = (1 - \lambda) C' x \Rightarrow \lambda \neq 1$ da sonst $x = 0$

$\Rightarrow \dfrac{\lambda}{1-\lambda} x = C' x$

Für später notieren wir:

$$M = (\mathbb{1} + G)^{-1} G \Rightarrow (\mathbb{1} + G) M = G \Rightarrow M = G(\mathbb{1} - M)$$

$$G = M \underbrace{(\mathbb{1} - M)^{-1}}_{\text{invertierbar, da } \lambda \neq 1} = (\mathbb{1} - M)^{-1} M$$

Desweiteren:

$$(\mathbb{1} - \underbrace{W^{-1}R}_{=M}) = W^{-1}(W - R) = W^{-1}A \Rightarrow W^{-1} = (\mathbb{1} - M)A^{-1}$$

$$\mathbb{1} - M^{k+1} \underset{\uparrow}{=} \left(\sum_{i=0}^{k} M^i\right)(\mathbb{1} - M) \Rightarrow (\mathbb{1} - M^{k+1})A^{-1} = \underbrace{\left(\sum_{i=0}^{k} M^i\right) W^{-1}}_{\geq 0}$$

Teleskop-Summe

$$\Rightarrow \quad 0 \leq (\mathbb{1} - \underbrace{M^{k+1}}_{\geq 0}) \underbrace{A^{-1}}_{\geq 0} \leq A^{-1}$$

A M-Matrix

$$\Rightarrow \quad 0 \leq M^{k+1} A^{-1} \leq A^{-1}$$

$\Rightarrow$ M^{k+1} beschränkt $\Rightarrow \rho(M) \leq 1$

A nicht singulär

$\Longrightarrow$ $|\lambda| \leq 1 \xrightarrow{\lambda \neq 1} \lambda < 1 \Rightarrow$ Konvergenz

$\lambda = \rho(M)$ s.o. z.B.: $\lambda \in \mathbb{R}, \lambda \geq 0$

(iii) Aus $G = (\mathbb{1} - M)^{-1} M$ folgt:

$$\lambda' \in \mathbb{C} \text{ EW zu } M \Leftrightarrow \mu' = \frac{\lambda'}{1 - \lambda'} \in \mathbb{C} \text{ EW zu } G$$

$$|\mu'| \underset{\uparrow}{\leq} \frac{|\lambda'|}{1 - |\lambda'|} \underset{\uparrow}{\leq} \frac{|\lambda|}{1 - |\lambda|} =: \mu$$

$\boxed{|1 - \lambda'| \geq 1 - |\lambda'|}$ $\quad \frac{t}{1-t}$ monoton

$|\mu'|$ maximal für $\lambda' = \lambda = \rho(M) \Rightarrow \mu = \rho(G)$

$$\Rightarrow \quad \varrho(G') = \frac{\varrho(M)}{1-\varrho(M)} \quad \Rightarrow \quad \varrho(M) = \frac{\varrho(G')}{1+\varrho(G')}$$

$\square$

5.15. <u>Satz</u> A M-Matrix, dann konvergiert das Gauß-Seidel-Verfahren und $\varrho(M^g) \leq \varrho(M^J) < 1$

<u>Bew</u>: (i) $M^g = (D-E)^{-1}F$ <u>Aufspaltung</u>: $A = W - R$

$W = (D-E)$, $R = F$

<u>Es gilt</u>: $W = D - E \geq D - E - F = A$

W regulär, $W = \underbrace{D}_{\text{strikt positive Diagonale}} - \underbrace{E}_{\leq 0} \Longrightarrow W^{-1} \geq 0$

Dreiecksgl. system:

$$\begin{pmatrix} \ddots & & 0 \\ E & D & \\ & & \ddots \end{pmatrix} (W^{-1})_i = e_i$$

(alternative Argumentation: W M-Matrix)

$\Rightarrow$ Vor. 5.14 erfüllt $\Rightarrow \varrho(M^g) < 1$

(ii) zur Abschätzung:

$W^g = D - E \qquad W^J = D$

$A \leq W^g \leq W^J \underset{A \text{ M-Matrix}}{\Longrightarrow} 0 \leq \underbrace{A^{-1}}_{\geq 0}\underbrace{(W^g - A)}_{=R^g} \leq A^{-1}\underbrace{(W^J - A)}_{=R^J}$

$$\Longrightarrow \varrho(A^{-1}R^g) = \lim_{k\to\infty} |||(A^{-1}R^g)^k|||^{\frac{1}{k}} \leq \lim_{k\to\infty} |||(A^{-1}R^J)^k|||^{\frac{1}{k}} = \varrho(A^{-1}R^J)$$

$$\varrho(A) = \lim_{k\to\infty} |||A^k|||^{\frac{1}{k}}$$

Ü Beweis mit Jordan-Normalform

5.14 $\Longrightarrow$

$$\rho(\underbrace{M^g}_{=W^{-1}R^g}) = \frac{\rho(A^{-1}R^g)}{1+\rho(A^{-1}R^g)} \overset{\frac{t}{1+t}\text{ monoton}}{\leq} \frac{\rho(A^{-1}R^J)}{1+\rho(A^{-1}R^J)} = \rho(\underbrace{M^J}_{=W^{-1}R^J})$$

Im Folgenden betrachten wir speziell positiv definite Matrizen A:

Achtung: $\boxed{\text{Wechsel der Notation}}$

$$A, B \in \mathbb{R}^{I,I} \quad A \underset{(-)}{\leq} B \;:\Leftrightarrow\; Ax\cdot x \underset{(\leq)}{\leq} Bx\cdot x \;\forall x \neq 0$$

5.16 Satz: Sei A pos. def., dann konvergiert das gedämpfte Jacobi-Verfahren falls $0 < \Theta A < 2D$ oder äquivalent falls $\Theta < \frac{2}{\rho(D^{-1}A)}$

Bew zur Erinnerung: $x^{k+1} = x^k - \Theta N^J(Ax^k - b)$

$$N^J = \Theta D^{-1}$$

$$\Leftrightarrow \quad W_\Theta^J(x^k - x^{k+1}) = Ax^k - b, \quad W_\Theta^J = \frac{D}{\Theta}$$

(i) $0 < \Theta A < 2D \Rightarrow 2W = \frac{2}{\Theta}D > A > 0$

Iterationsmatrix $M = M_\Theta^J = (\mathbb{1} - W^{-1}A)$

W pos. def $\Rightarrow W^{\pm\frac{1}{2}}$ existiert und ist pos. def.

$$\tilde{M} := W^{\frac{1}{2}} M W^{-\frac{1}{2}} = \mathbb{1} - W^{-\frac{1}{2}} A W^{-\frac{1}{2}}$$

$$\boxed{2W > A > 0} \Rightarrow 2\underbrace{W^{-\frac{1}{2}} W W^{-\frac{1}{2}}}_{\mathbb{1}} > \underbrace{W^{-\frac{1}{2}} A W^{-\frac{1}{2}}}_{=:\tilde{A}} > 0$$

$\Rightarrow$ Spektrum $\sigma(\tilde{A}) \subset (0,2)$

$\Rightarrow$ Spektrum $\sigma(\tilde{M}) \subset (-1,1)$

$\Rightarrow \rho(\tilde{M}) < 1 \Rightarrow \rho(M) < 1$

$\tilde{M}$ ähnlich zu M (mit Basistrf. $W^{\frac{1}{2}}$)

(ii) zu Äquivalenz:

$$0 < \Theta A < 2D \iff 0 < \frac{\Theta}{2} D^{-\frac{1}{2}} A D^{-\frac{1}{2}} < \mathbb{1}$$

$$\underset{A \text{ pos. def.}}{\iff} \frac{\Theta}{2} \underbrace{\varrho(D^{-\frac{1}{2}} A D^{-\frac{1}{2}})}_{} < 1$$

$$\boxed{\text{Basistrf. } D^{-\frac{1}{2}}} \quad = \varrho(\underbrace{D^{-\frac{1}{2}}(D^{-\frac{1}{2}} A D^{-\frac{1}{2}}) D^{\frac{1}{2}}}_{D^{-1}A})$$

$$\iff \Theta < \frac{2}{\varrho(D^{-1}A)}$$

□

5.17. Lemma

Für Matrizen W der 3ten Normalform ($W = N^{-1}$) gelte $W + W^T > A > 0$, dann ist W regulär und die Iteration konvergiert monoton in der Energie-Norm $\|\cdot\|_A$ mit $\|x\|_A = \sqrt{Ax \cdot x}$. Ferner gilt $\varrho(M) \leq |||M|||_A < 1$ für $M = \mathbb{1} - W^{-1}A$ (Iterationsmatrix)

Bew (i) wäre W singulär $\Rightarrow \exists x \neq 0 \; Wx = 0 \Rightarrow$

$$0 = Wx \cdot x + x \cdot Wx = (W + W^T)x \cdot x > Ax \cdot x > 0 \; ↯$$

(ii) Da $\varrho(M) \leq |||M|||_A$ bleibt z.z. $|||M|||_A < 1$

$\boxed{\text{da Operatornorm}}$

$$\left(\Rightarrow \|e^{k+1}\|_A = \|M e^k\|_A \leq \underbrace{|||M|||_A}_{<1} \|e^k\|_A \right)$$

monotone Konvergenz

hierzu: $|||M|||_A = \sup_{x \neq 0} \sqrt{\frac{AMx \cdot Mx}{Ax \cdot x}} \underset{y = A^{\frac{1}{2}}x}{=} \sup \sqrt{\frac{A^{\frac{1}{2}} M A^{-\frac{1}{2}} y \cdot A^{\frac{1}{2}} M A^{-\frac{1}{2}} y}{y \cdot y}}$

$$\boxed{\|B\|_2 = \sup \frac{\|Bx\|_2}{\|x\|_2}}$$

$$= \|\underbrace{A^{\frac{1}{2}} M A^{-\frac{1}{2}}}_{=: \hat{M}}\|_2 = \|\hat{M}\|_2 \text{ mit}$$

$M = \mathbb{1} - W^{-1}A$

$$\Rightarrow \hat{M} = \mathbb{1} - A^{\frac{1}{2}} W^{-1} A A^{-\frac{1}{2}} = \mathbb{1} - A^{\frac{1}{2}} W^{-1} A^{\frac{1}{2}}$$

$$\Rightarrow \hat{M}^T\hat{M} = \mathbb{1} - A^{\frac{1}{2}}\underbrace{(W^{-T}+W^{-1})}_{=W^{-T}\underbrace{(W+W^T)}_{>A}W^{-1}}A^{\frac{1}{2}} + A^{\frac{1}{2}}W^{-T}AW^{-1}A^{\frac{1}{2}}$$

$$< \mathbb{1}$$

$$\Rightarrow |||\hat{M}|||_2 =: \sup_{x\neq 0}\sqrt{\frac{\hat{M}x\cdot\hat{M}x}{x\cdot x}} = \sup_{x\neq 0}\sqrt{\frac{\hat{M}^T\hat{M}x\cdot x}{x\cdot x}}$$

$$< |||\mathbb{1}|||_2 = 1$$

□

5.18 **Folgerung** Das Gauß-Seidel-Verfahren konvergiert monoton in der Energie-Norm für pos. def. Matrizen A.

Bew $A > 0 \Rightarrow D > 0$

$$W = D - E \underset{E=F^T}{\Longrightarrow} W + W^T = 2D - E - F = D + A > A$$

$\overset{5.17.}{\Rightarrow}$ Beh

□

5.19. **Folgerung** (Ostrowski)

Das SOR-Verfahren konvergiert monoton in der Energienorm falls $0 < \omega < 2$.

Bew $W_\omega^{SOR} = \frac{1}{\omega}D - E$ (regulär ✓)

$$W_\omega^{SOR} + (W_\omega^{SOR})^T = \frac{2}{\omega}D - E - F = A + \left(\frac{2}{\omega} - 1\right)D$$

$$> A$$

$$\Updownarrow$$

$$\frac{2}{\omega} - 1 > 0 \Leftrightarrow \omega > 0 \text{ und } \omega < 2$$

□

Nun wenden wir uns den Gradientenverfahren zu (vgl. ALMA II):

Betrachte $F(x) = \frac{1}{2} \underset{\mathbb{R}^{I,I}}{A} x \cdot x - \underset{\mathbb{R}^{I}}{b} \cdot x$ (A pos. def., symm)

$$F(x^*) = \min_{x \in \mathbb{R}^I} F(x)$$

notwendige Bedingung: $F'(x^*) = 0 \iff F'(x^*)(v) = \frac{d}{dt} F(x^* + \epsilon v)\Big|_{\epsilon=0} = 0$

$$\iff 0 = \frac{d}{d\epsilon}\left(\frac{1}{2} A(x+\epsilon v)\cdot(x+\epsilon v) - b\cdot(x+\epsilon v)\right)\Big|_{\epsilon=0}$$

$$= \frac{1}{2}\left(Av\cdot(x+\epsilon v) + A(x+\epsilon v)\cdot v\right) - b\cdot v\Big|_{\epsilon=0} = \frac{1}{2}(A+A^T)x\cdot v - b\cdot v$$

$$= (Ax-b)\cdot v \quad \forall v \in \mathbb{R}^I \iff Ax = b$$

$\boxed{\text{1D - Minimierung}}$ $f(\lambda) = F(x+\lambda p)$ $\quad f(\lambda^*) = \min_{\lambda \in \mathbb{R}} f(\lambda)$

$$f'(\lambda^*) = 0 = F'(x+\lambda^* p)\cdot p$$

$$= A(x+\lambda^* p)\cdot p - b\cdot p \iff \lambda^* = \frac{(b-Ax)\cdot p}{Ap\cdot p}$$

$\boxed{\text{Abstiegsrichtung}}$ $\boxed{p = -\nabla F(x) = b - Ax}$

Gradientenverfahren:

x^0 Startvektor; $r^0 = b - Ax^0$;

$x^{k+1} = x^k + \lambda^*(x^k, r^k)\, r^k$ mit

$$\lambda^*(x^k, r^k) = \frac{r^k \cdot r^k}{Ar^k \cdot r^k};$$

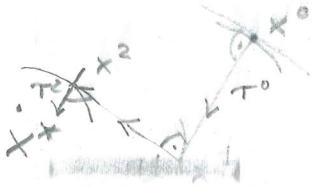
$$r^{k+1} = b - Ax^{k+1} = b - A(x^k + \lambda^*(x^k, r^k)\, r^k)$$

$$= r^k - \lambda^*(x^k, r^k)\, A r^k;$$

5.20 Algorithmus

```
SteepestDescent(x) {
  r = b - Ax;
  while (¬Abbruchkrit.) { h = Ar; λ* = r·r / h·r;
                          x = x + λ* r; r = r - λ* h;
  }
}
```

Problem "zick-zack"



$r^{k+1} \perp r^k$

das $\underbrace{-\nabla F(x^{k+1})}_{r^{k+1}} \perp$ alten Suchrichtung

Bem:
- nichtlineare Iteration, da λ^* nicht konstant ist
- Verfahren konsistent, da $r(x^*) = b - Ax^* = 0$

5.21. **Satz** (Konvergenz) (vgl. ALMA II)

$$\boxed{\begin{aligned} &\|x^k - x^*\|_A = \delta^k \|x^0 - x^*\|_A \qquad \left(\|x\|_A = \sqrt{Ax \cdot x}\right) \\ &\text{mit } \delta = \frac{\lambda_{max}(A) - \lambda_{min}(A)}{\lambda_{max}(A) + \lambda_{min}(A)} \end{aligned}}$$

Bew (vgl. Richardson-Iteration)

$$x^{k+1} = x^k - \Theta^* (Ax^k - b) \qquad \Theta^* \underset{\text{Variabel}}{=} \lambda^*(x^k, r^k)$$

$$e^{k+1} = M_k e^k \qquad M_k = \mathbb{1} - \lambda^*(x^k, r^k) A$$

$$\|e^{k+1}\|_A = \sqrt{Ae^{k+1} \cdot e^{k+1}} = \sqrt{A^{\frac{1}{2}} e^{k+1} \cdot A^{\frac{1}{2}} e^{k+1}} = \|A^{\frac{1}{2}} e^{k+1}\|_2$$

$$= \|A^{\frac{1}{2}} M_k e^k\|_2 = \|M_k A^{\frac{1}{2}} e^k\|_2 \le |||M_k|||_2 \, \|A^{\frac{1}{2}} e^k\|_2$$

$$\Rightarrow \|e^{k+1}\|_A \le |||M_k|||_2 \, \|e^k\|_A$$

und nach der Abschätzung zur Richardson Iteration (M_k symm.)

$$|||M_k|||_2 \le \delta$$

Nun suchen wir nach einer Verbesserung durch Transformation der Variablen:

W pos. def., symm. und W^{-1} einfach berechenbar

$$\boxed{\hat{x} = W^{\frac{1}{2}} x \qquad \hat{A} = W^{-\frac{1}{2}} A W^{-\frac{1}{2}} \qquad \hat{b} = W^{-\frac{1}{2}} b}$$

Es gilt $Ax = b \Leftrightarrow \hat{A}\hat{x} = \hat{b} \Leftrightarrow \hat{F}(\hat{x}) = \frac{1}{2}\hat{A}\hat{x} \cdot \hat{x} - \hat{b} \cdot \hat{x} = \min$

$\boxed{\text{Idee}}$ $\hat{A} \approx \lambda \mathbb{1} \longleftrightarrow$ schnelle Konvergenz

Transformation des Verfahrens:

$$\hat{x}^{k+1} = \hat{x}^k + \hat{\lambda}^* \hat{r}^k \xRightarrow{\cdot W^{-\frac{1}{2}}} x^{k+1} = x^k - \hat{\lambda}^* W^{-\frac{1}{2}} \hat{r}^k$$

$$x^{k+1} = x^k - \hat{\lambda}^* W^{-1} r^k$$

$$\Rightarrow \quad \hat{r}^k = \hat{b}^k - \hat{A}\hat{x}^k = W^{-\frac{1}{2}} r^k$$

$$\hat{r}^{k+1} = \hat{r}^k - \hat{\lambda}^* \hat{A} \hat{r}^k \xRightarrow{\cdot W^{\frac{1}{2}}} r^{k+1} = r^k - \hat{\lambda}^* W^{\frac{1}{2}} W^{-\frac{1}{2}} A W^{-\frac{1}{2}} W^{-\frac{1}{2}} r^k$$

$$= r^k - \hat{\lambda}^* A W^{-1} r^k$$

$$\hat{\lambda}^* = \frac{\hat{r}^k \cdot \hat{r}^k}{\hat{A}\hat{r}^k \cdot \hat{r}^k} = \frac{W^{-\frac{1}{2}} r^k \cdot W^{-\frac{1}{2}} r^k}{W^{-\frac{1}{2}} A W^{-\frac{1}{2}} W^{-\frac{1}{2}} r^k \cdot W^{-\frac{1}{2}} r^k} = \frac{W^{-1} r^k \cdot r^k}{A W^{-1} r^k \cdot W^{-1} r^k}$$

Zusammenfassend erhalten wir das Gradientenverfahren für das transformierte Problem ($\hat{A}\hat{x} = \hat{b}$) in den alten Variablen ($x^k = W^{-\frac{1}{2}} \hat{x}^k$, $r^k = W^{\frac{1}{2}} \hat{r}^k$)!

vorkonditioniertes Gradientenverfahren:

x^0 Startvektor; $r^0 = b - Ax^0$;

$q^k = W^{-1} r^k$; $h^k = A q^k$;

$$\hat{\lambda}^* = \frac{q^k \cdot r^k}{h^k \cdot q^k};$$

$$x^{k+1} = x^k + \hat{\lambda}^* q^k; \quad r^{k+1} = r^k - \hat{\lambda}^* h^k;$$

Bem:

- Im Vergleich zum Standardverfahren ist nur eine Multiplikation mit W^{-1} zusätzlich erforderlich!
- Interpretation: $q^k = W^{-1} r^k$ verbesserte Suchrichtung mit $\hat{\lambda}^* = \lambda^*(x^k, q^k)$

5.22. **Satz** (Konvergenz)

A, W pos. def., symm mit $\gamma W \leq A \leq \Gamma W$ („$\leq$" im Sinn quad. Formen), dann gilt
$$\|x^k - x^*\|_A \leq \left(\frac{\Gamma-\gamma}{\Gamma+\gamma}\right)^k \|x^0 - x^*\|_A$$

Bem

- $W = \mathbb{1}$ (nicht vorhand.) $\rightsquigarrow$ $\gamma = \lambda_{min}(A)$, $\Gamma = \lambda_{max}(A)$
  $\rightsquigarrow$ altes Konvergenzresultat
- $W = A$ $\quad \gamma = \Gamma = 1$ $\rightsquigarrow$ optimale Konvergenz

Bew **trf. Pb** $\quad \hat{F}(\hat{x}) = \frac{1}{2}\hat{A}\hat{x}\cdot\hat{x} - \hat{b}\cdot\hat{x}$

nach 5.21. gilt $\quad \|\hat{x}^k - \hat{x}^*\|_{\hat{A}} \leq \hat{\delta}^k \|\hat{x}^0 - \hat{x}^*\|_{\hat{A}}$

wobei $\|\hat{x}\|_{\hat{A}} = \sqrt{W^{\frac{1}{2}}A\,W^{-\frac{1}{2}}W^{\frac{1}{2}}x \cdot W^{\frac{1}{2}}x} = \sqrt{Ax\cdot x} = \|x\|_A$

ferner
$$\left.\begin{aligned} \hat{x}\cdot\hat{y} &= W^{\frac{1}{2}}x \cdot W^{\frac{1}{2}}y = Wx\cdot y \\ \hat{A}\hat{x}\cdot\hat{y} &= W^{-\frac{1}{2}}AW^{-\frac{1}{2}}W^{\frac{1}{2}}x \cdot W^{\frac{1}{2}}y = Ax\cdot y \end{aligned}\right\} \xrightarrow{\gamma W \leq A \leq \Gamma W}$$

$$\gamma\,\hat{x}\cdot\hat{x} \leq \hat{A}\hat{x}\cdot\hat{y} \leq \Gamma\,\hat{x}\cdot\hat{x}$$

$\Rightarrow \lambda_{max}(\hat{A}) \leq \Gamma$, $\lambda_{min}(\hat{A}) \geq \gamma$

$$\Rightarrow \hat{\delta} = \frac{\lambda_{max}(\hat{A}) - \lambda_{min}(\hat{A})}{\lambda_{max}(\hat{A}) + \lambda_{min}(\hat{A})} \leq \frac{\Gamma-\gamma}{\Gamma+\gamma} \Rightarrow \text{Beh} \quad \square$$

$g(a,b) = \dfrac{a-b}{a+b}$ monoton wachsend in a und monoton fallend in b

nachr.:
$$\partial_a g = \frac{a+b-(a-b)}{(a+b)^2} = \frac{2b}{(a+b)^2} > 0$$
$$\partial_b g = \frac{-(a+b)-(a-b)}{(a+b)^2} = \frac{-2a}{(a+b)^2} < 0$$

Ziel: Finde W (und damit Skalarprodukt $Wx \cdot y$), so dass

(i) W^{-1} einfach berechenbar

(ii) W-Skalarprodukt möglichst äquivalent zum A-Skalarprodukt ($y \sim \Gamma$)

Bem Betrachte lineare Iterationen in der dritten Normalform

$$W(x^k - x^{k+1}) = Ax^k - b \quad (W = A \Rightarrow x^{k+1} = x^*),$$

d.h. auch hier sucht man W mit W^{-1} einfach berechenbar und $W \sim A$.

$\rightsquigarrow$ Verwende W als Vorkond. Matrix

Bsp (i) Jacobi-Verfahren (diagonale Vorkonditionierung)

$$A = D - B, \quad Dx^{k+1} = Bx^k + b \iff$$
$$W(x^k - x^{k+1}) = Ax^k - b \text{ mit } W = D$$

A pos. def., symm $\Rightarrow D$ pos. def.

(ii) symm. Gauß-Seidel-Verfahren

$$A = D - E - F, \quad F = E^T \quad (D-E)x^{k+1} = Fx^k + b \iff$$
$$W(x^k - x^{k+1}) = Ax^k - b \text{ mit } W = D - E$$

Pb: W nicht symm.

Ausweg: Man kann auch die Indexmenge rückw. durchlaufen

$$(D-F)(x^k - x^{k+1}) = Ax^k - b$$

$\rightsquigarrow$ symm Gauß-Seidel-Verfahren:

$$\tilde{x}^{k+1} = (D-E)^{-1}Fx^k + (D-E)^{-1}b \quad \text{(Vorwärts)}$$

$$x^{k+1} = (D-F)^{-1}E\tilde{x}^{k+1} + (D-F)^{-1}b \quad \text{(Rückwärts)}$$

$$= \underbrace{(D-F)^{-1}E(D-E)^{-1}F}_{= M^{sg}} x^k + \underbrace{\left((D-F)^{-1}E(D-E)^{-1} + (D-F)^{-1}\right)}_{= N^{sg}} b$$

$$N^{sg} = (D-F)^{-1}\left(E(D-E)^{-1} + (D-E)(D-E)^{-1}\right)$$
$$= (D-F)^{-1}(E+D-E)(D-E)^{-1} = (D-F)^{-1}D(D-E)^{-1}$$
$$\Rightarrow W^{sg} = (N^{sg})^{-1} = (D-E)D^{-1}(D-F)$$

zu berechnen: $(W^{sg})^{-1} = N^{sg} = (D-F)^{-1}D(D-E)^{-1}$

rückwärts einsetzen vorwärts einsetzen

(Wg) Konsistenz: $Mx^* + Nb = x^* \Rightarrow Nb = (\mathbb{1}-M)x^* \Rightarrow A = N^{-1}(\mathbb{1}-M)$

$$\Rightarrow N^{-1}x^{k+1} = N^{-1}Mx^k + b$$
$$\Rightarrow N^{-1}(x^k - x^{k+1}) = N^{-1}x^k - N^{-1}Mx^k - b = Ax^k - b$$

(iii) Man kann auch das relaxierte Verfahren einsetzen:

SSOR – Vorkonditionierung (symm. – successive- over- relaxation)

$$W^{-1} = (2-\omega)\left(\tfrac{1}{\omega}D - F\right)^{-1}\tfrac{1}{\omega}D\left(\tfrac{1}{\omega}D - E\right)^{-1} \longrightarrow \text{vgl Ü}$$

(iv) unvollständige Eliminationsverfahren

Idee: LR bzw Cholesky-Zerlegung auf dünn besetzten Matrizen und in L (und R) nur Beiträge ablegen, wo bereits Nicht-Nulleinträge stehen.

$\rightsquigarrow$ ILU bzw. IC – Zerlegung

incomplete R Cholesky

zur Vorkonditionierung bietet sich dann das IC-Verfahren an und man wählt für die unvollst. Zerlegung $L^{IC}(L^{IC})^T$:

$$W^{-1} = (L^{IC})^{-T}(L^{IC})^{-1}$$

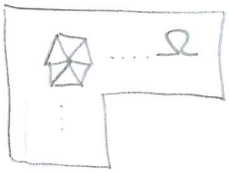
Rückw. –, Vorwärtseinsetzen

Frage: $\boxed{\text{Lohnen sich iterative Löser?}}$

A_h Steifigkeitsmatrix zum Poissonproblem auf einer Triangulierung mit Gitterweite h

Vorkonditionierung: $W = M_h$

Massematrix



$$A_h = \int_\Omega \nabla\varphi_h^i \cdot \nabla\varphi_h^j \, dx \qquad M_h = \int_\Omega \varphi_h^i \varphi_h^j \, dx$$

A_h, M_h pos. def, Frage: $\gamma_h M_h \le A_h \le \Gamma_h M_h$; γ_h, Γ_h?

$$A_h \bar{u}_h \cdot \bar{u}_h = \int_\Omega |\nabla u_h|^2 dx \qquad M_h \bar{u}_h \cdot \bar{u}_h = \int_\Omega |u_h|^2 dx$$

$u_h(x) = \sum_i u_h^i \varphi_h^i(x)$

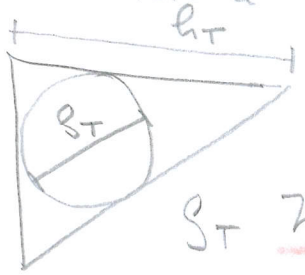
Lemma von Poincaré in $\mathbb{R}^n$: $\|u\|_{L^2(\Omega)} \le C_p \|\nabla u\|_{L^2(\Omega)}$

für $u \in H_0^1(\Omega)$

0 Randwerte

$$\Rightarrow \underbrace{\frac{1}{C_p^2}}_{=\gamma_h} \int_\Omega |u_h|^2 dx \le \int_\Omega |\nabla u_h|^2 dx$$

Eine umgekehrte Abschätzung $\int_\Omega |\nabla u_h|^2 dx \le \Gamma_h \int_\Omega |u_h|^2 dx$ gilt wenn das Gitter gleichmäßig fein ist ($h_{max} \le C h_{min}$) und die Dreiecke regulär sind in dem Sinn, daß

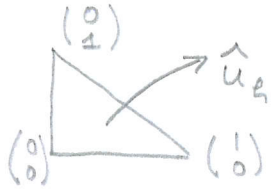


$$h_T \le C \varrho_T$$

ϱ_T Inkreisdurchmesser

nicht regulär

hierzu: (i) Auf dem Einheitsdreieck $\hat{T}$ definieren wir zwei Normen



$$\|\hat{u}_\ell\|_1 := \left(\int_{\hat{T}} |\hat{u}_\ell|^2 + |\nabla \hat{u}_\ell|^2 \, d\hat{x}\right)^{\frac{1}{2}}$$

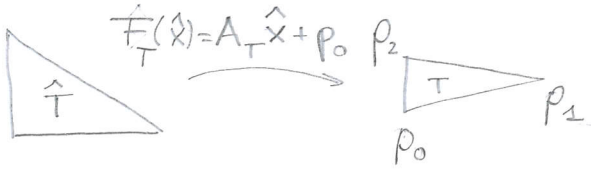
$$\|\hat{u}_\ell\|_0 := \left(\int_{\hat{T}} |\hat{u}_\ell|^2 \, d\hat{x}\right)^{\frac{1}{2}} \leftarrow L^2\text{-Norm}$$

Der Raum der affinen Fkt'n auf $\hat{T}$ ist endlich dimensional, d.h. $\|.\|_0$ und $\|.\|_1$ sind äquivalent:

$$\|\hat{u}_\ell\|_0 \leq \|\hat{u}_\ell\|_1 \leq C \, \|\hat{u}_\ell\|_0$$

insbesondere gilt $\displaystyle\int_{\hat{T}} |\nabla \hat{u}_\ell|^2 \, d\hat{x} \leq C \int_{\hat{T}} |\hat{u}_\ell|^2 \, d\hat{x}$

(ii) Nun transformieren wir elementweise auf das Einheits-Simplex:



$$u_\ell(x) = \hat{u}_\ell(F_T^{-1}(x))$$

$$A_T = \begin{pmatrix} | & | \\ a_1 & a_2 \\ | & | \end{pmatrix} \qquad a_1 = p_1 - p_0, \quad a_2 = p_2 - p_0$$

Dann gilt $\partial_{x_i} u_\ell(x) = \partial_{\hat{x}_j} \hat{u}_\ell(F_T^{-1}(x)) \, (A_T^{-1})_{ji}$

$$\Rightarrow |\nabla_x u_\ell(x)|^2 \leq |||A_T^{-1}|||^2 \, |\nabla_{\hat{x}} \hat{u}_\ell(F_T^{-1}(x))|^2$$

Nun erhalten wir:

$$\int_T |\nabla_x u_\ell(x)|^2 \, dx \leq |\det A_T| \int_{\hat{T}} |\nabla_{\hat{x}} \hat{u}_\ell(\hat{x})|^2 \, d\hat{x} \; |||A_T^{-1}|||^2$$

$$\leq C \, |\det A_T| \int_{\hat{T}} |\hat{u}_\ell(\hat{x})|^2 \, d\hat{x} \; |||A_T^{-1}|||^2$$

$$= C \underbrace{|\det A_T| \, |\det A_T^{-1}|}_{=1} \int_T |u_\ell(x)|^2 \, dx \; \underbrace{|||A_T^{-1}|||^2}_{\overset{!}{\leq} C h_T^{-2}}$$

(*)

zu (*):

$$\frac{\|A_T^{-1}(x-y)\|}{\|x-y\|} = \frac{\|A_T^{-1}x - A_T^{-1}y\|}{\|x-y\|} \le \frac{h_T}{\varrho_T} \le \frac{\boxed{2}}{\frac{1}{C}h_T} = \frac{C}{h_T}$$

Zsf: $\displaystyle\int_T |\nabla u_h|^2 dx \le C h_T^{-2} \int_T |u_h|^2 dx$

und damit $\displaystyle\int_\Omega |\nabla u_h|^2 dx \le \underbrace{C h_{min}^{-2}}_{=\Gamma_h} \int_\Omega |u_h|^2 dx$

Damit erhalten wir als Konvergenzrate für das (mit M_h vorkonditionierte) Gradientenverfahren:

$$\delta_h = \frac{\Gamma_h - \gamma_h}{\Gamma_h + \gamma_h} = \frac{C h_{min}^{-2} - \frac{1}{c_p^2}}{C h_{min}^{-2} + \frac{1}{c_p^2}} = \frac{C C_p^2 - h_{min}^2}{C C_p^2 + h_{min}^2}$$

$$= 1 - O(h_{min}^2)$$ [gilt dimensionsunabh. für $d = 2,3$ ← Tetraedergitter]

$\boxed{\text{Aufwandsabschätzung}}$ $\in$ Fehlertoleranz nach k Schritten

erreicht, falls: $\delta_h^k \|x^0 - x^*\|_A \le \epsilon \Rightarrow$

$$k \ge \frac{|\ln \epsilon| + |\ln \|x^0 - x^n\|_A|}{|\ln \underbrace{\delta_h}_{1 - O(h_{min}^2)}|} = \frac{|\ln \epsilon| + |\ln \|x^0 - x^*\|_A|}{C h_{min}^2} = O(h^2)$$

Gesamtaufwand: $\left.\begin{array}{l} Cd \text{ Einträge} \neq 0 \text{ pro Zeile} \\ h_{min} \ge C h_{max} \text{ (gl.m. feines Gitter)}, \; n = C h^{-d} \end{array}\right\} \Rightarrow$

pro Iteration $O(C h^{-d} d)$

$\Rightarrow$ Der Gesamtaufwand ist

$$C\, d\,(\ln|\epsilon| + \ln\|x^0 - x^*\|_A)\, h^{-d-2}$$

zum Vergleich: Gauß-Elimination oder Cholesky:

Bandbreite $O(h^{-d+1}) \Rightarrow$

Aufwand $O(\underset{\substack{\uparrow \\ \text{Zeilenzahl}}}{h^{-d}}\,(h^{-d+1})^2) = O(h^{-3d+2})$

	$d=1$	$d=2$	$d=3$
Grad. Verf.	$O(h^{-3})$	$O(h^{-4})$	$O(h^{-5})$
Cholesky	$O(h^{-1})$	$O(h^{-4})$	$O(h^{-7})$

$\leftarrow$ [$O(.)$ enthält Abh. von ϵ, $\|x^0 - x^*\|_A$, d

Bem: • Die Vorkond. mit $W = M_h$ ist nicht effektiv (gleich zu $W = I$) (ohne Beweis hier)

• Verbesserung durch das konjugierte Gradientenverfahren

nun: Verfahren der konjugierten Gradienten

Ansatz: Wähle Suchrichtung optimal mit $p^k \in \operatorname{span}\{r^i\}_{i \le k}$ (alte Suchrichtungen und neues Residuum),

so dass

$$\|x^{k+1} - x^*\|_A = \min_{x \in x^0 + \operatorname{span}\{r^i\}_{i \le k}} \|x - x^*\|_A$$

$$r^{k+1} = b - A(x^k + \lambda^k p^k)$$

$$= r^k - \lambda^k A p^k = r^0 + \sum_{i=1}^{k+1} \alpha_i A^i r^0$$

$\boxed{p^0 = r^0,\ p^k \in \operatorname{span}\{r^i\}_{i \le k}}$

$$\Rightarrow r^k \in T_k := \{ r \in \mathbb{R}^I \mid r = r^0 + s,\ s \in S_k \}$$

$$S_k := \operatorname{span}\{A^i r^0\}_{i=1,\dots,k}$$ Krylov-Raum

$$x^k - x^* = A^{-1}(Ax^k - b) = -A^{-1} r^k \Rightarrow$$

$$\|x^k - x^*\|_A = \|r^k\|_{A^{-1}}$$

Optimalitätsbedingung: $\|r^k\|_{A^{-1}} = \min_{r \in T_k} \|r\|_{A^{-1}}$

5.23 Satz A pos. def., symm $\in \mathbb{R}^{I,I}$, $Ax^* = b$, dann gilt

$\|x^{k+1} - x^*\|_A = \min_{x \in x^0 + T_k} \|x - x^*\|_A$, genau dann wenn

$p^k = r^k + \beta^k p^{k-1}$ (bis auf Skalierung)

mit $\beta^k = -\dfrac{Ap^{k-1} \cdot r^k}{Ap^{k-1} \cdot p^{k-1}}$

Ferner gilt: $r^k \perp r^\ell \; \forall \ell \neq k$,

$p^k \perp Ap^\ell \; \forall \ell \neq k$ (A-Orthogonalität),

$p^k = 0 \Rightarrow r^k = 0 \Rightarrow x^k = x^*$,

und $\beta^k = \dfrac{r^k \cdot r^k}{r^{k-1} \cdot r^{k-1}}$ $\quad \lambda^k = \dfrac{r^k \cdot r^k}{Ap^k \cdot p^k}$

Bew (vgl. ALMA II)

5.24 Algorithmus

```
cg (x, A, b) {
  r = b - Ax; p = r, δ^0 = r·r;
  while (δ^0 > ε^2) {          ← ||r^k||_2 ≤ ε Abbruch-
    h = Ap; λ = δ^0/h·p;          kriterium
    x = x + λp; r = r - λh;
    δ^1 = r·r; β = δ^1/δ^0; δ^0 = δ^1;
    p = r + βp;                ← neue Suchrichtung!
  }
```

Exkurs aus des ALMA II zum CG-Verfahren:

Herleitung des Verfahrens basierend auf der Optimalitätsbedingung

Wir zeigen später: $A_\ell \in \mathbb{R}^{\mathring{I}_\ell, \mathring{I}_\ell}$ ist pos. def. und symmetrisch (insbesondere invertierbar!)

Exkurs (Gradientenverfahren)

Wie lösen wir solche Gleichungssystem $A_\ell u_\ell^\circ = f_\ell$?

2.5 Lemma Sei $A \in \mathbb{R}^{I,I}$, $b \in \mathbb{R}^I$, A pos. def., symmetrisch

$F(x) := \frac{1}{2} Ax \cdot x - b \cdot x$, dann ist

x^* Lsg von $Ax = b$ gdw x^* Minimalstelle von F.

Bew „$\Rightarrow$" notwendige Bed. für Minimalstelle $F'(x^*) = 0$

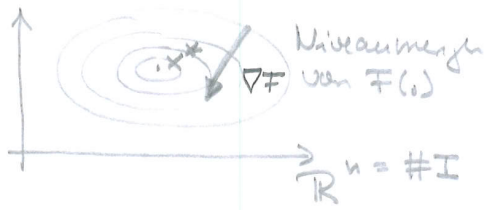
$$F'(x) = Ax - b \quad \left(\partial_k \left(\frac{1}{2} \sum_{i,j} a_{ij} x_i x_j + \sum_i b_i x_i \right) \right.$$

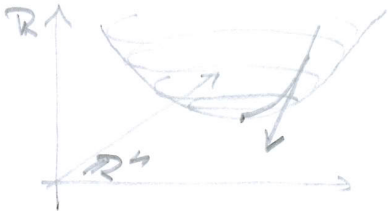
$$= \frac{1}{2} \left(\sum_j a_{kj} x_j + \sum_j a_{jk} x_j - b_k \right)$$

$$\left. = \sum_j a_{kj} x_j - b_k \right)$$

„$\Leftarrow$" $F'(x) = 0$ hat eindt. Lösung x^* und $Ax^* = b$ $\square$

Interpretation





2.6 Lemma (Minimierung in eine Richtung)

$A \in \mathbb{R}^{I,I}$ pos. def., symm., $x \in \mathbb{R}^I$, $p \in \mathbb{R}^I$, $p \neq 0$,

$f(\lambda) = F(x + \lambda p)$, dann ist $\lambda^* \in \mathbb{R}$ mit $f(\lambda^*) = \min_{\lambda \in \mathbb{R}} f(\lambda)$

eindt. bestimmt:

$$\lambda^*(x,p) = \frac{r \cdot p}{Ap \cdot p} \quad \text{mit } r = b - Ap$$

Bew $f'(\lambda^*) = 0 \Rightarrow F'(x + \lambda^* p) \cdot p = 0 \Rightarrow$

$$A(x + \lambda^* p) \cdot p - b \cdot p = 0 \Rightarrow \lambda^* = \frac{(b - Ax) \cdot p}{Ap \cdot p} \quad \square$$

→ p nennt man Suchrichtung

gute Wahl der Suchrichtung: Richtung des (lokal) steilsten Abstiegs $-\nabla F(x)$:

$$-\nabla F(x) = -(Ax-b) = b - Ax = r$$

→ Gradientenverfahren

x^0 Startvektor

$$x^{k+1} = x^k + \lambda^*(x^k, r^k)\, r^k \quad \text{mit } r^k = b - Ax^k$$

$$\text{und } \lambda^*(x^k, r^k) = \frac{r^k \cdot r^k}{Ar^k \cdot r^k}$$

$$r^{k+1} = b - Ax^{k+1} = b - A(x^k + \lambda^*(x^k, r^k)\, r^k)$$

$$= r^k - \lambda^*(x^k, r^k)\, \underbrace{Ar^k}_{\text{bereits berechnet}}$$

2.7. Algorithmus

```
steepestDescent (x, A, b) {
  r = b - Ax; δ = r·r;
  while (e > ε²) {
    h = Ar; λ* = δ / h·r;
    x = x + λ* r; r = r - λ* h;
    δ = r·r;
  }
}
```

2.8. Satz (Konvergenz)

Sei A pos. def., symm., dann konvergiert das Gradientenverfahren und es gilt

$$\|x^k - x^*\|_A \le \delta^k \|x^0 - x^*\|_A$$

für $Ax^* = b$ und $\delta := \dfrac{\lambda_{max}(A) - \lambda_{min}(A)}{\lambda_{max}(A) + \lambda_{min}(A)}$ (< 1)

Bew • Energienorm $\|x\|_A = \sqrt{Ax\cdot x}$

$$\frac{1}{2}\|x-x^*\|_A^2 = \frac{1}{2}A(x-x^*)\cdot(x-x^*)$$

$$\underset{Ax^*=b}{=} \frac{1}{2}Ax\cdot x - b\cdot x + \underbrace{\frac{1}{2}b\cdot x^*}_{=C\text{ konstant}} = F(x) + C$$

$\Rightarrow$ Minimierung von $x \mapsto F(x)$ oder auch $\lambda \mapsto F(x+\lambda p)$ ist äquivalent zur Minimierung des Abstands $\|x^*-x\|_A$ oder $\|x^*-(x+\lambda p)\|_A$. ⊛

• Fortsetzung des Fehlers: $\quad (\lambda^* = \lambda^*(x^k, r^k))$

$$x^{k+1} = x^k - \lambda^*(Ax^k - b)$$

$$\oplus \quad x^* = x^* - \lambda^*(Ax^* - b)$$

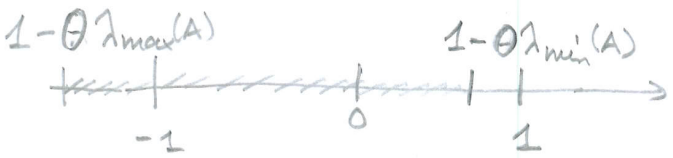
$$e^{k+1} = e^k - \lambda^* A e^k = (\mathbb{1} - \lambda^* A)e^k \Rightarrow$$

$$\|e^{k+1}\|_A = \|(\mathbb{1}-\lambda^* A)e^k\|_A$$

$$\leq \|(\mathbb{1}-\Theta A)e^k\|_A = \|A^{\frac{1}{2}}(\mathbb{1}-\Theta A)e^k\|$$

$$= \|(\mathbb{1}-\Theta A)A^{\frac{1}{2}}e^k\| \leq \underbrace{|||(\mathbb{1}-\Theta A)|||_2}_{\overset{\text{Übung}}{=}|\rho_{\mathbb{1}-\Theta A}|}\|e^k\|_A$$

Optimalität des Schritts



optimaler Fall: $\quad 1-\Theta\lambda_{max}(A) = -(1-\Theta\lambda_{min}(A))$

$$\Leftrightarrow \Theta(\lambda_{max}(A)+\lambda_{min}(A)) = 2$$

$$\Leftrightarrow \Theta = \frac{2}{\lambda_{max}+\lambda_{min}}$$

$$\Rightarrow |||\mathbb{1}-\Theta A|||_2 = 1 - \frac{2\lambda_{min}}{\lambda_{max}+\lambda_{min}} = \frac{\lambda_{max}-\lambda_{min}}{\lambda_{max}+\lambda_{min}}$$

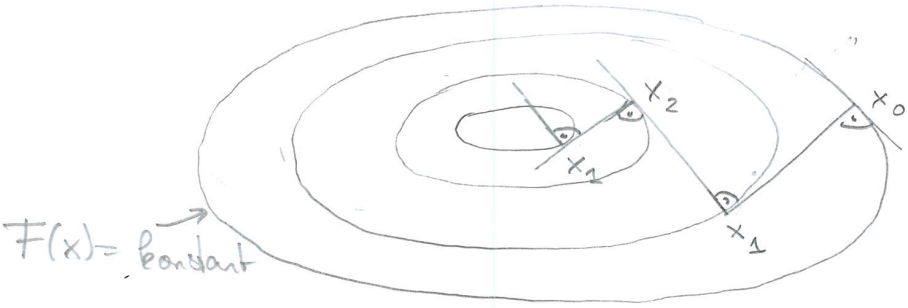
□

2.8 Bem. • Für die Matrix zum diskreten Poissonpb auf $[0,1]^2$ gilt

$$\frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}} = 1 - 2\sin^2\left(\frac{\pi}{2}h\right) = 1 - \frac{1}{2}\pi^2 h^2 + O(h^4)$$

$\rightsquigarrow$ langsame Konvergenz

• geometrische Interpretation: $\lambda_{max} \gg \lambda_{min}$



$r_i = b - Ax_i$

$\perp r_{i-1}$

(da steilster Abstieg!)

D.h. man sucht eine bessere Abstiegsrichtung:

$$x^{k+1} = x^k + \lambda^k p^k \qquad (\text{bisher } p^k = r^k)$$

$$\text{mit } \lambda^k = \frac{r^k \cdot p^k}{Ap^k \cdot p^k} \qquad (p^k \neq 0)$$

Ansatz: p^k sollte von allen bisherigen Richtungen abhängen, d.h. $p^k \in \operatorname{span}\{r^i\}_{i \le k}$, und optimal sein, in dem Sinn, dass

$$\boxed{\|x^{k+1} - x^*\|_A = \min_{x \in x^0 + \operatorname{span}\{r^i\}_{i \le k}} \|x - x^*\|_A}$$

$$r^{k+1} = b - A(x^k + \lambda^k p^k)$$

$$= r^k - \lambda^k A p^k$$

$$\underset{\boxed{p^0 = r^0}}{\overset{\text{Iteration}}{=}} r^0 + \sum_{i=1}^{k+1} \alpha_i A^i r^0$$

$$\Rightarrow \quad r^k \in T_k := \{ r \in \mathbb{R}^I \mid r = r^0 + s,\ s \in S_k \}$$

$$S_k := \operatorname{span}\{ A^i r^0 \}_{i=1,\dots,k}$$

(Krylov-Raum)

$$x^k - x^* = A^{-1}(Ax^k - b) = -A^{-1} r^k \Rightarrow$$

$$\|x^k - x^*\|_A = \sqrt{A A^{-1} r^k \cdot A^{-1} r^k} = \|r^k\|_{A^{-1}}$$

$\Rightarrow$ Optimalitätsbedingung:

$$\boxed{\|r^k\|_{A^{-1}} = \min_{r \in T_k} \|r\|_{A^{-1}}} \iff \boxed{F(x^{k+1}) = \min_{x \in x^0 + T_k} F(x)}$$

2.9. Satz

Voraussetzung wie oben, $\|x^{k+1} - x^*\|_A = \min_{x \in x^0 + T_k} \|x - x^*\|$

gilt gdw. $p^k = r^k + \beta^k p^{k-1}$ (bis auf Skalierung)

mit $\beta^k = -\dfrac{A p^{k-1} \cdot r^k}{A p^{k-1} \cdot p^{k-1}}$.

Ferner gilt: $r^k \perp r^l$ für $l \neq k$

$p^k \perp A p^l$ für $l \neq k$

und $p^k = 0 \Rightarrow r^k = 0 \Rightarrow x^k = x^*$

A orthogonal konjugiert $\rightsquigarrow$ Name der Verfahrens:

Konjugierte Gradienten (CG)

2.10. Lemma

$$\boxed{\beta^k = \frac{r^k \cdot r^k}{r^{k-1} \cdot r^{k-1}} \qquad \lambda^k = \frac{r^k \cdot r^k}{A p^k \cdot p^k}}$$

2.10 Algorithmus

```
cg(x, A, b) {
  r = b - Ax; p = r; δ⁰ = r·r;
  while (δ⁰ > ε²) {          ← ||r^k||_2 ≤ ε  Stop-Kriterium
    h = Ap; λ = δ⁰/h·p;
    x = x + λp;   r = r - λh;
    δ¹ = r·r;     β = δ¹/δ²;
    p = r + βp;
    δ⁰ = δ¹;                 ← Rücksetzen
  }
```

Aufwand: pro Iteration
1. Matrixvektorprodukt
2. Skalarprodukte
3. Operationen ($y = x + \alpha w$, $x, w \in \mathbb{R}^I$, $\alpha \in \mathbb{R}$)

2.11. Bem Das CG-Verfahren berechnet in höchstens $n = \#I$ Schritten die Lösung x^*

(theoretisch interessant, praktisch erwartet man eine gute Approximation von x^* schon viel früher)

hierzu: $r^k \cdot r^i = 0 \quad \forall i < r$

$\Rightarrow$ entweder $r^k = 0 \rightarrow x^k = x^*$

oder $T_{k-1} \subsetneq T_k \Rightarrow k \leq n$ $\square$

Bew von 2.9: Ansatz: $p^k = \sum_{i=0}^{k} \gamma_i^k r^i$ und Optimalität erfüllt

$$\boxed{\frac{d}{dt} F(x^{k+1} + t r^i)\Big|_{t=0} = 0 \text{ (Optimalität)} \Rightarrow \nabla F(x^{k+1}) = r^{k+1} \perp r^i \ \forall_{i \le k}}$$

Falls $r^{k+1} \in \operatorname{span}\{r^i\}_{i \le k} \Leftrightarrow r^{k+1} = 0 \Rightarrow x^{k+1} = x^*$

$\Rightarrow x^{k+1} \neq x^* \Rightarrow r^{k+1} \notin \operatorname{span}\{r^i\}_{i \le k}$

Ferner falls $F(x^k + t p^k) < F(x^k)$ für $t > 0$ klein

$\Rightarrow \underbrace{p^k \cdot r^k}_{\gamma_k} \neq 0$ (sonst keine Abstiegsrichtung) insbesondere $p^k \neq 0$

o.B.d.A. $\boxed{\gamma_k^k = 1}$

$$\Rightarrow p^k = r^k + \gamma_{k-1}^k r^{k-1} + \text{---} \gamma_0^k r^0 \qquad (**)$$

$$\Rightarrow r^k = p^k - \gamma_{k-1}^k r^{k-1} - \text{---} \gamma_0^k r^0$$

Iteration $\rightsquigarrow$ $r^k = p^k - \beta_{k-1}^k p^{k-1} - \text{---} \beta_0^k p^0$

zunächst zeigen wir: $\boxed{A p^k \cdot p^l = 0 \quad \forall_{l<k}}$

hierzu: $r^{k+1} = r^k - \lambda^k A p^k \Rightarrow A p^k = \frac{r^k - r^{k+1}}{\lambda^k}$

$\lambda^k = 0 \rightsquigarrow$ bereits bei x^* und $r^k = 0$

Für $l < k$:

$$A p^k \cdot p^l = \frac{1}{\lambda^k}(r^k - r^{k+1}) \cdot (r^l + \gamma_{l-1}^l r^{l-1} + \text{---} \gamma_0^k r^0)$$

$\boxed{\forall_{k \neq l}\, r^k \cdot r^l = 0}$ $\Downarrow$ $= 0$

weiter: $p^k = r^k + \beta_{k-1}^k p^{k-1} + \text{---} \beta_0^k p^0 \xrightarrow{\cdot A p^i \ (i<k)}$

$$0 = A p^i \cdot r^k + \beta_i^k A p^i \cdot p^i + 0$$

$$\Rightarrow \beta_i^k = -\frac{A p^i \cdot r^k}{A p^i \cdot p^i}$$

$$Ap^i = \frac{r^i - r^{i+1}}{\lambda^i} \Rightarrow Ap^i \cdot r^k = 0 \quad \forall i < k-1 \Rightarrow \beta_i^k = 0 \ \forall i < k-$$

$$\Rightarrow p^k = r^k + \underbrace{\beta_{k-1}^k}_{= -\frac{Ap^{k-1} \cdot r^k}{Ap^{k-1} \cdot p^{k-1}} =: \beta^k} p^{k-1}$$

$\square$

Beweis zu Lemma 2.10:

$$\beta^k = -\frac{Ap^{k-1} \cdot r^k}{Ap^{k-1} \cdot p^{k-1}} = -\frac{\frac{1}{\lambda^{k-1}}(r^k - r^{k-1}) \cdot r^k}{\frac{1}{\lambda^{k-1}}(r^k - r^{k-1}) \cdot (r^{k-1} + \underbrace{s}_{\in \operatorname{span}\{r^i\}_{i<k-1}})}$$

(**)

$r^k \perp r^l \ (k \neq l)$

$$= \frac{r^k \cdot r^k}{r^{k-1} \cdot r^{k-1}}$$

$$\lambda^k = \frac{r^k \cdot p^k}{Ap^k \cdot p^k} = \frac{r^k \cdot (r^k + \underbrace{s}_{\in \operatorname{span}\{r^i\}_{i<k}})}{Ap^k \cdot p^k} = \frac{r^k \cdot r^k}{Ap^k \cdot p^k} \quad \square$$

(**)

2.12 **Satz** (Konvergenz)

Vor wie oben, $K = \frac{\lambda_{max}(A)}{\lambda_{min}(A)}$, dann gilt

$$\|x^k - x^*\|_A \leq 2 \left(\frac{\sqrt{K}-1}{\sqrt{K}+1}\right)^k \|x^0 - x^*\|_A$$

(ohne Beweis hier!)

Bem:

- $\frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}} = \frac{K-1}{K+1}$
- diskretes Poissonpb auf $[0,1]$: λ_{max}
  $K \doteq C(1 + O(h))\, h^{-2}$

Gradientenverfahren: $\boxed{K = h^{-2}}$

(bis auf Terme höherer Ordnung)

$$\frac{K-1}{K+1} = \frac{h^{-2}-1}{h^{-2}+1} = \frac{1-h^2}{1+h^2}$$

$$= \frac{1+h^2}{1+h^2} - \frac{2h^2}{1+h^2} = 1 - O(h^2)$$

CG-Verfahren:

$$\frac{\sqrt{K}-1}{\sqrt{K}+1} = \frac{h-1}{h+1} = 1 - O(h)$$

---

noch offen ist die Frage ob die dimensionierten Matrizen oben positiv definit sind!

2.13 **Def** $A \in \mathbb{R}^{I,I}$ heißt M-Matrix, falls

$a_{ii} > 0 \quad \forall i \in I, \quad a_{ij} \leq 0 \quad \forall i \neq j;\ i,j \in I$

A nicht singulär, $A^{-1} \geq 0$

$(A \geq B: \quad a_{ij} \geq b_{ij}$ für alle $i,j \in I)$

2.14. **Def** (Matrixeigenschaften)

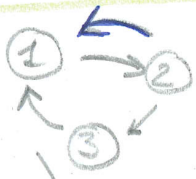
$A \in \mathbb{R}^{I,I}$ gegeben, dann ist $i \in I$ mit $j \in I$ verbunden, wenn es eine Kette $i = i_0, i_1, i_2, \ldots i_k = j$ mit $k \in \mathbb{N}$ gibt, so dass $a_{i_\ell i_{\ell+1}} \neq 0 \quad \forall \ell = 0, \ldots k-1$.

A heißt irreduzibel, falls jeder Index aus I mit jedem Index aus I verbunden ist.

Bsp

- $A = \begin{pmatrix} 1 & \overset{1}{\cancel{0}} & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$ (nicht) irreduzibel

  (zunächst 1 nicht mit 2 und 3 verbunden)



- $\Omega_h = [0,1]^2 \Rightarrow A_h$ zum diskreten Poissonpb irreduzibel

Ende Exkurs aus des ALMA II

5.25 **Bem** [Das CG-Verfahren berechnet in höchstens $n = \mathrm{card}(I)$ Schritten die Lösung x^*]

hierzu: $r^k \cdot r^i = 0 \quad \forall i < k$; d.h. entweder $r^k = 0 \Rightarrow x^k = x^*$ oder $T_k \supsetneq T_{k-1} \ \forall k \Rightarrow k \leq n$ □

- nur theoretisch interessant für große Gl.systeme, da gute Approximation von x^* viel früher erreicht und Rundungsfehler $x^k = x^*$ im Allgemeinen verhindern.

nun: Konvergenzeigenschaften des CG-Verfahrens

Ziel: Finde $M_k \in \mathbb{R}$, so dass $\underbrace{\|x^k - x^*\|_A}_{\|r^k\|_{A^{-1}}} \leq M_k \|x^0 - x^*\|_A$

wir wissen bereits $r^k = r^0 + \sum_{i=1}^{k} \alpha_i^* A^i r^0$

$$= P_k^*(A) r^0$$

mit $P_k^*(t) = 1 + \sum_{i=1}^{k} \alpha_i^* t^i$

Definition: $\mathcal{P}_k^1$ = Menge aller Polynome P mit Grad $\leq k$ und $P(0) = 1$

$P_k^* \in \mathcal{P}_k^1$

$$\Rightarrow \quad T_k = \{ r \mid r = P_k(A) r^0 \, ; \, P_k \in \mathcal{P}_k^1 \}$$

$$\|r^k\|_{A^{-1}} = \|P_k^*(A) r^0\|_{A^{-1}} = \min_{P_k \in \mathcal{P}_k^1} \|P_k(A) r^0\|_{A^{-1}}$$

$$\|P_k(A) r^0\|_{A^{-1}}^2 = A^{-1} P_k(A) r^0 \cdot P_k(A) r^0$$

$$= A^{-1} P_k(A)^2 r^0 \cdot r^0 = \|r^0\|^2_{A^{-1} P_k(A)^2}$$

5.26 Lemma Die Eigenwerte von A seien enthalten in einer Menge S mit $\max_{\lambda \in S} |P_k^*(\lambda)| \le M_k$, dann folgt

$$\|x^k - x^*\|_A \le M_k \|x^0 - x^*\|_A$$

Bew: A pos. def., symm. $\Rightarrow$ $\exists$ ON Basis $\{v_i\}_{i=1,\dots,n}$ aus EV'n zu EW $0 < \lambda_1 \le \lambda_2 \dots \le \lambda_n$

$$\Rightarrow \quad r^0 = \sum_{i=1}^{n} (r^0 \cdot v_i) v_i \qquad r^0 \cdot r^0 = \sum_{i=1}^{n} (r^0 \cdot v_i)^2$$

$v_i \cdot v_j = \delta_{ij}$

$$A^{-1} r^0 = \sum_{i=1}^{n} \lambda_i^{-1} (r^0 \cdot v_i) v_i$$

$$A^{-1} r^0 \cdot r^0 = \sum_{i=1}^{n} \lambda_i^{-1} (r^0 \cdot v_i)^2$$

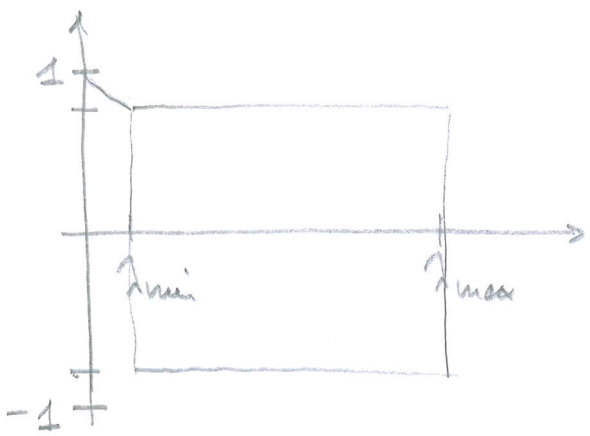
Ferner: $\|r^k\|_{A^{-1}}^2 = A^{-1} P_k^*(A) r^0 \cdot r^0 = \sum_{i=1}^{n} \lambda_i^{-1} P_k^*(\lambda_i)^2 (r^0 \cdot v_i)^2$

$$\Rightarrow \quad \|r^k\|_{A^{-1}}^2 \le M_k^2 \sum_{i=1}^{n} \lambda_i^{-1} (r^0 \cdot v_i)^2 = M_k \|r^0\|_{A^{-1}}^2$$

$\square$

zur Bestimmung von M_k suchen wir Polynom P_k^*, so dass

$$\max_{\lambda \in S} P_k^*(\lambda) = \min_{P_k \in \mathcal{P}_k^1} \max_{\lambda \in S} P_k(\lambda)$$



Bsp $\boxed{k=1}$ (vgl. Richardson-Iteration)

$P_1^*(\lambda) = 1 - \alpha\lambda$, so dass

$$P_1(\lambda_{max}) = -P_1(\lambda_{max}) \Leftrightarrow 1 - \alpha\lambda_{min} = \alpha\lambda_{max} - 1$$

$$\Leftrightarrow \alpha = \frac{2}{\lambda_{min} + \lambda_{max}}$$

$$\Rightarrow \mu^1 = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}}$$

D.h. die Konvergenz des CG-Verfahren ist mindestens so gut, wie die des Gradientenverfahrens!
(nach Def. bereits klar!)

$\boxed{\text{Exkurs:}}$ **Chebyshev Polynome**

Herleitung: betrachte $\cos(k\Theta)$, es gilt

$$\boxed{\cos((k+1)\Theta) = 2\cos\Theta\cos(k\Theta) - \cos((k-1)\Theta)}$$

Bew: $\cos((k+1)\Theta) = \cos(k\Theta + \Theta)$

$$= \cos(k\Theta)\cos\Theta - \underbrace{\sin(k\Theta)}_{\sin((k-1)\Theta+\Theta) = \cos((k-1)\Theta)\sin\Theta + \cos\Theta\cos((k-1)\Theta)}\sin\Theta$$

$$= \cos(k\Theta)\cos\Theta - \cos((k-1)\Theta)(1-\cos^2\Theta) - \sin\Theta\cos\Theta\sin((k-1)\Theta)$$

$$= \cos(k\Theta)\cos\Theta - \cos((k-1)\Theta) + \cos\Theta\underbrace{(\cos\Theta\cos((k-1)\Theta) - \sin\Theta\sin((k-1)\Theta))}_{=\cos(k\Theta)}$$

$$= 2\cos(k\Theta)\cos\Theta - \cos((k-1)\Theta)$$

$\square$

$\Rightarrow$ $\cos(k\Theta) = t_k(\cos\Theta)$ mit t_k Polynom vom Grad k

$\boxed{\cos(0\cdot\Theta),\ \cos\Theta \text{ Polynome über } \cos\Theta \text{ von Grad } 0 \text{ bzw. } 1}$

Mit Variablentf $x = \cos\Theta$ ergibt sich

$$\boxed{\begin{aligned} t_0(x) &= 1 \\ t_1(x) &= x \\ t_{k+1}(x) &= 2x\,t_k(x) - t_{k-1}(x) \end{aligned}} \quad (*)$$

Konstruktion zunächst auf $[-1,1]$, dann $x \in \mathbb{R}$ zugelassen.

5.27 Lemma $\quad t_k(x) = \frac{1}{2}\left[\left(x+\sqrt{x^2-1}\right)^k + \left(x-\sqrt{x^2-1}\right)^k\right]$

Bew $\quad \tilde{t}_k(x) =$ Rechte Seite

$\Rightarrow \quad \tilde{t}_0(x) = 1 \qquad \tilde{t}_1(x) = x$

und induktiv:

$$\tilde{t}_{k+1}(x) = \frac{1}{2}\left(\left(x+\sqrt{x^2-1}\right)^k \left(x+\sqrt{x^2-1}\right) + \left(x-\sqrt{x^2-1}\right)^k \left(x-\sqrt{x^2-1}\right)\right)$$

$$= x\,\tilde{t}_k(x) + \frac{1}{2}\left[\left(x+\sqrt{x^2-1}\right)^k \sqrt{x^2-1} - \left(x-\sqrt{x^2-1}\right)^k \sqrt{x^2-1}\right]$$

$$= x\,\tilde{t}_k(x) + \frac{1}{2}\Big[\left(x^2-1+x\sqrt{x^2-1}\right)\left(x+\sqrt{x^2-1}\right)^{k-1} +$$
$$\left(x^2-1-x\sqrt{x^2-1}\right)\left(x-\sqrt{x^2-1}\right)^{k-1}\Big]$$

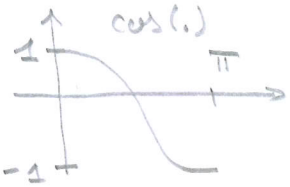
$$= x\,\tilde{t}_k(x) - \tilde{t}_{k-1}(x)$$
$$+\underbrace{\frac{1}{2}\Big[\left(x^2+x\sqrt{x^2-1}\right)\left(x+\sqrt{x^2-1}\right)^{k-1} + \left(x^2-x\sqrt{x^2-1}\right)\left(x-\sqrt{x^2-1}\right)^{k-1}\Big]}_{= x\,\tilde{t}_k}$$

$$= 2x\,\tilde{t}_k(x) - \tilde{t}_{k-1}(x)$$

D.h. es gilt die rekursive Definition ⊛, die $t_k(\cdot)$ def. □

Aus $t_k(x) = \cos(k \cos^{-1}(x))$ für $x \in [-1,1]$ folgt:

$$\boxed{\max_{-1 \le x \le 1} |t_k(x)| = 1}$$

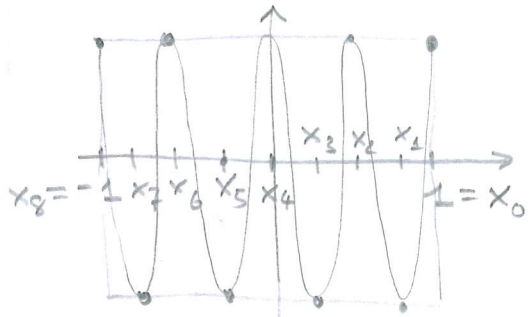


Ferner gilt für $x_i = \cos\left(\frac{i\pi}{k}\right)$ $i = 0, \dots, k$:

$$t_k(x_i) = \cos(i\pi) = (-1)^i$$

$\Rightarrow$ Nullstellen in $[x_i, x_{i+1}]$

$\Rightarrow$ k Nullstellen in $(-1,1)$



$\Rightarrow$ außerhalb $(-1,1)$ keine Nullstellen

5.28 <u>Satz</u> $S = [a,b]$ für $0 < a < b$, dann gilt für

$$P_k^*(x) = \frac{t_k\left(\frac{b+a-2x}{b-a}\right)}{t_k\left(\frac{b+a}{b-a}\right)}$$

$$\max_{a \le x \le b} |P_k^*(x)| = \min_{P_k \in \mathcal{P}_k^1} \max_{a \le x \le b} |P_k(x)|$$

<u>Bew</u> (i) $\hat{x}(x) := \underbrace{-2\frac{x-a}{b-a} + 1}$ ist lineare Abb. von $[a,b] \to [-1,1]$

$$= -\frac{2x+2a+(b-a)}{b-a} = \frac{b+a-2x}{b-a}$$

$$\Rightarrow \underbrace{t_k\left(\frac{b+a-2x}{b-a}\right)}_{} = t_k(\hat{x}(x))$$

Verallgemeinerung der $t_k(-x)$-Konstruktion auf $[a,b]$

$$\Rightarrow \left|t_k\left(\frac{b+a-2x}{b-a}\right)\right| \le 1 \text{ für } x \in [a,b]$$

Ferner oszilliert $t_k(\cdot)$ zwischen 1 und -1 auf den Punkten $a = x_0 < x_1 < x_2 < \text{———} < x_k = b$

mit $x_i = \frac{1}{2}\left(\cos\left(\frac{(k-i)\pi}{k}\right)+1\right)(b-a) + a$

(ii) zunächst sehen wir $P_k^* \in \mathcal{P}_k^1$

Nehmen wir nun an $P_k \in \mathcal{P}_k^1$ existiere mit:

$$\max_{a \le x \le b} |P_k(x)| < \max_{a \le x \le b} |P_k^*(x)|$$

Dann gilt für $e(x) = P_k^*(x) - P_k(x)$

$e(x_i)$ hat das Vorzeichen $(-1)^{(k-i)}$ und $e(x_i) \neq 0$

$\Rightarrow$ in (x_i, x_{i+1}) liegt eine Nullstelle von $e(\cdot)$ für $i = 0, \ldots k-1$

$\Rightarrow$ $e(\cdot)$ hat k Nullstellen in (a,b)

Ferner $e(0) = 1 - 1 = 0$

$\Big\}$ $e(\cdot)$ hat $(k+1)$ Nullstellen

$\overset{e(\cdot) \in P_k}{\Longrightarrow}$ $e \equiv 0$ $\Rightarrow$ $P_k = P_k^*$ ↯ $\square$

$\boxed{\text{Ende Exkurs}}$

$S = [\underbrace{\lambda_{min}(A)}_{=a}, \underbrace{\lambda_{max}(A)}_{=b}]$, P_k^* wie im Exkurs

$$\underbrace{\max_{s \in S} |P_k^*(s)|}_{} \leq |P_k^*(a)|$$

$$\overset{\boxed{\substack{(a<b) \\ s=a}}}{=} \frac{\left|t_k\left(\frac{b+a-2a}{b-a}\right)\right|}{\left|t_k\left(\frac{b+a}{b-a}\right)\right|} = \frac{\overbrace{|t_k(1)|}^{=1}}{\left|t_k\left(\frac{b+a}{b-a}\right)\right|} = t_k\left(\frac{\frac{b}{a}+1}{\frac{b}{a}-1}\right)$$

$$\frac{b}{a} = \frac{\lambda_{max}}{\lambda_{min}} = K(A)$$

$$\frac{K+1}{K-1} \pm \sqrt{\frac{(K+1)^2}{(K-1)^2} - 1} = \frac{K+1}{K-1} \pm \frac{\sqrt{(K+1)^2 - (K-1)^2}}{K-1}$$

$$= \frac{K+1 \pm 2\sqrt{K}}{K-1}$$

$$= \begin{cases} \dfrac{\sqrt{K}^2 + 2\sqrt{K} + 1}{(\sqrt{K}-1)(\sqrt{K}+1)} = \dfrac{\sqrt{K}+1}{\sqrt{K}-1} \\[2ex] \dfrac{\sqrt{K}^2 - 2\sqrt{K} + 1}{(\sqrt{K}-1)(\sqrt{K}+1)} = \dfrac{\sqrt{K}-1}{\sqrt{K}+1} \end{cases}$$

5.27

$$\Rightarrow \quad t_k\left(\frac{K-1}{K+1}\right) = \frac{1}{2}\left[\underbrace{\left(\frac{\sqrt{K}+1}{\sqrt{K}-1}\right)}_{>0}^k + \underbrace{\left(\frac{\sqrt{K}-1}{\sqrt{K}+1}\right)}_{>0}^k\right]$$

$$\overset{\substack{\uparrow \\ K>1}}{\geq} \frac{1}{2}\left(\frac{\sqrt{K}+1}{\sqrt{K}-1}\right)^k$$

$\lambda_{max} > \lambda_{min}$

$$\Rightarrow \quad M_k = \max_{s \in S} |P_k^*(s)| = \frac{1}{t_k\left(\frac{K+1}{K-1}\right)} < 2\left(\frac{\sqrt{K}-1}{\sqrt{K}+1}\right)^k$$

Damit folgt der Konvergenzsatz:

5.29 <u>Satz</u> Für das CG-Verfahren gilt

$$\|x^k - x^*\|_A \leq 2\left(\frac{\sqrt{K}-1}{\sqrt{K}+1}\right)^k \|x^0 - x^*\|_A$$

mit $K = \dfrac{\lambda_{max}}{\lambda_{min}}$

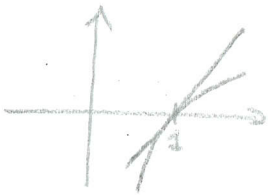
<u>Bew</u>: M_k wie oben; zunächst nur für $\lambda_{max} > \lambda_{min}$

$$\lambda_{max} = \lambda_{min} \rightarrow A = \lambda_{max} \mathbb{1} \Rightarrow x^1 = x^* \quad \square$$

<u>Bem</u> optimal $\|x^k - x^*\|_A \leq t_k\left(\frac{K-1}{K+1}\right)^{-1} \|x^0 - x^*\|_A$

Aufwand: $\left(\frac{\sqrt{K}-1}{\sqrt{K}+1}\right)^k \le \frac{\epsilon}{2\,\|x^0-x^*\|_A}$

$t-1 \ge \ln t$



s.o.

$$\Rightarrow \quad k \ge \frac{\left|\ln\frac{\epsilon}{2}\right| + \left|\ln\|x^0-x^*\|_A\right|}{\left|\ln\left(\frac{\sqrt{K}-1}{\sqrt{K}+1}\right)\right|}$$

$$\ge \frac{\left|\ln\frac{\epsilon}{2}\right| + \left|\ln\|x^0-x^*\|_A\right|}{\left|\frac{\sqrt{K}-1}{\sqrt{K}+1}-1\right|} \longleftarrow = \left|\frac{\sqrt{K}-1-\sqrt{K}-1}{\sqrt{K}+1}\right| = \frac{2}{\sqrt{K}+1}$$

$$\Rightarrow \quad k^{cg} \ge \frac{\left|\ln\frac{\epsilon}{2}\right| + \left|\ln\|x^0-x^*\|_A\right|}{2}\left(\sqrt{K}+1\right)$$

vgl. Gradientenverfahren: $k^{gv} \ge \left(|\ln\epsilon| + |\ln\|x^0-x^*\|_A|\right)\frac{K+1}{2}$

(FE Problem: $K = O(h^{-2})$)

5.30 Satz (Vorkonditionierung)

Für das vorkonditionierte CG-Verfahren erhält man für $\gamma W \le A \le \Gamma W$ die Fehlerabschätzung aus 5.29 mit $K = \frac{\Gamma}{\gamma}$. Für das Verfahren ergibt sich:

$$x^{k+1} = x^k + \lambda^k p^k; \qquad \lambda^k = \frac{r^k\cdot q^k}{Ap^k\cdot p^k}$$

$$r^{k+1} = r^k - \lambda^k A p^k;$$

$$q^{k+1} = W^{-1} r^{k+1};$$

$$\beta^k = \frac{r^{k+1}\cdot q^{k+1}}{r^k\cdot q^k}; \qquad p^{k+1} = q^{k+1} + \beta^k p^k;$$

wobei $r^0 = b - Ax^0$; $q^0 = W^{-1} r^0$; $p^0 = q^0$;

Bew $\boxed{x^k = W^{-\frac{1}{2}}\hat{x}^k \; ; \; r^k = W^{\frac{1}{2}}\hat{r}^k \; ; \; \hat{A} = W^{-\frac{1}{2}}AW^{-\frac{1}{2}}}$

(i) $\hat{p}^{k+1} = \hat{r}^{k+1} + \beta_k \hat{p}^k$ und $\hat{p}^0 = \hat{r}^0 = W^{-\frac{1}{2}} r^0 = W^{-\frac{1}{2}} p^0$

$\underset{\text{Induktion}}{\overset{\cdot W^{-\frac{1}{2}}}{\Longrightarrow}}$ $p^k = W^{-\frac{1}{2}}\hat{p}^k$

(ii) $\hat{x}^{k+1} = \hat{x}^k + \hat{\lambda}^k \hat{p}^k \underset{W^{-\frac{1}{2}}\cdot}{\longrightarrow} x^{k+1} = x^k + \hat{\lambda}^k p^k$

(iii) $\hat{r}^{k+1} = \hat{r}^k - \hat{\lambda}^k \hat{A}\hat{p}^k \underset{W^{\frac{1}{2}}\cdot}{=} r^{k+1} = r^k - \hat{\lambda}^k \cancel{W^{\frac{1}{2}} W^{-\frac{1}{2}}} A \cancel{W^{-\frac{1}{2}} W^{+\frac{1}{2}}} p^k$

$= r^k - \hat{\lambda}^k p^k$

(iv) $\hat{\lambda}^k = \dfrac{\hat{r}^k \cdot \hat{r}^k}{\hat{A}\hat{p}^k \cdot \hat{p}^k} = \dfrac{r^k \cdot W^{-\frac{1}{2}} W^{-\frac{1}{2}} r^k}{W^{-\frac{1}{2}} A W^{-\frac{1}{2}} W^{\frac{1}{2}} p^k \cdot W^{\frac{1}{2}} p^k} = \dfrac{r^k \cdot q^k}{A p^k \cdot p^k}$

(v) $\hat{p}^{k+1} = \hat{r}^{k+1} + \hat{\beta}_k \hat{p}^k \underset{W^{-\frac{1}{2}}\cdot}{\Longrightarrow} p^{k+1} = q^{k+1} + \beta_k p^k$

(vi) $\hat{\beta}_k = \dfrac{\hat{r}^{k+1} \cdot \hat{r}^{k+1}}{\hat{r}^k \cdot \hat{r}^k} = \dfrac{W^{-\frac{1}{2}} W^{-\frac{1}{2}} r^{k+1} \cdot r^{k+1}}{W^{-\frac{1}{2}} W^{-\frac{1}{2}} r^k \cdot r^k} = \dfrac{r^{k+1} \cdot q^{k+1}}{r^k \cdot q^k}$

$\square$

6. Nichtlineare Gleichungen

Bisher: $f(x) = Ax - b \overset{!}{=} 0$ (lin. Gl. system)

und $F(x) = \frac{1}{2} Ax \cdot x - b \cdot x \longrightarrow \min$ (quad. Minimierungsproblem)

nun: $f : \mathbb{R}^n \to \mathbb{R}^n ; x \mapsto f(x)$ nicht linear: gesucht Nullstelle

und $F : \mathbb{R}^n \to \mathbb{R} ; x \mapsto F(x)$ nicht quadratisch: gesucht Minimum

Bsp (hochdimensional): (i) $u_h : \Omega_h \to \mathbb{R}$ stückweise affine Fkt.

$$F[u_h] = \int_{\Omega_h} \sqrt{1 + |\nabla u_h|^2}\, dx$$ Fläche des Graphen von u_h

gesucht u_h mit $u_h|_{\partial\Omega} = u_h^\partial$

$$F[u_h] = \min_{\tilde{u}_h,\ \tilde{u}_h|_{\partial\Omega} = u_h^\partial} F[\tilde{u}_h]$$

(diskrete Minimalfläche)

(ii) **Erinnerung:** Federenergie

$$E[(p_1, \dots p_N)] = \sum_{i,j=1,\dots N} \alpha_{ij} |p_i - p_j|^2$$

$$\alpha_{ij} = \begin{cases} 1 & ; p_i, p_j \text{ durch Kante verbunden} \\ 0 & ; \text{sonst} \end{cases}$$

$\rightsquigarrow$ lineare Gleichungssystem als notwendige Bedingung

nun: $$E[(p_1, \dots p_N)] = \sum_{i,j=1,\dots N} \alpha_{ij} |p_i - p_j|^q$$

$q \geq 1$

$\rightsquigarrow$ für $q \neq 2$ nichtlineares Gleichungssystem

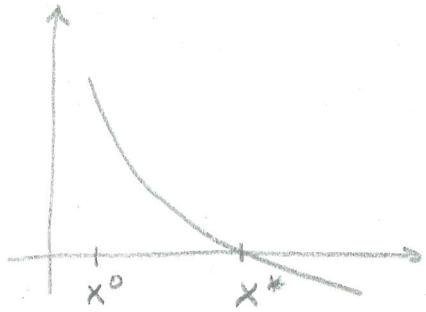
Newton-Verfahren (Motivation (in 1D)):

$f: \mathbb{R}^n \to \mathbb{R}^n$ differenzierbar $(n=1)$

x_0 Näherung für x^* mit $f(x^*)=0$

$p(x) = f(x_0) + f'(x_0)(x - x_0^*)$

affine Approximation



x_1 gesucht mit $p(x_1) = 0 \implies$

$$f'(x_0)(x_1 - x_0^*) = -f(x_0) \implies x_1 = x_0^* - f'(x_0)^{-1} f(x_0)$$

$f'(x_0)$ invertierbar ($\neq 0$ in 1D)

Jacobi-Matrix für $n \geq 1$

$\rightsquigarrow$ **Newton-Verfahren**

x_0 Startwert

$x_{k+1} = x_k - f(x_k)^{-1} f(x_k)$ für $k \geq 1$

$\hookrightarrow$ Fixpunktiteration zur Fixpunktabbildung

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

Exkurs Banach'scher Fixpunktsatz (vgl. Analysis)

6.1 Satz Sei (E, d) ein vollständiger, metrischer Raum und $\phi: E \to E$ kontrahierend $(d(\phi(x), \phi(y)) \leq \Theta\, d(x,y), \Theta < 1)$, dann hat ϕ Fixpunkt $x^* \in E$ $(\phi(x^*) = x^*)$, x^* eindt und die Folge x_k mit $x_0 \in E$ und $x_{k+1} = \phi(x_k)$ konvergiert gegen x^*. Ferner gilt

$$d(x_i, x^*) \leq \frac{\Theta}{1-\Theta} d(x_i, x_{i-1}) \quad \text{(A-Posteriori-Abschätzung)},$$

$$d(x_i, x^*) \leq \frac{\Theta^i}{1-\Theta} d(x_1, x_0) \quad \text{(A-Priori-Abschätzung)}.$$

$d(x,y) = 0 \Leftrightarrow x = y$
$d(x,y) \leq d(x,z) + d(z,y)$
$d(x,y) = d(y,x)$

Bew (i) $d(x_i, x_{i+k}) \leq d(x_i, x_{i+1}) + d(x_{i+1}, x_{i+2}) + \dots d(x_{i+k-1}, x_{i+k})$

$$\leq (1 + \Theta + \dots \Theta^{k-1})\, d(x_i, x_{i+1})$$

$$\underset{\text{geom. Reihe}}{=} \frac{1-\Theta^k}{1-\Theta}\, d(x_i, x_{i+1})$$

$$\leq \frac{1}{1-\Theta}\, \Theta\, d(x_{i-1}, x_i) \longrightarrow \text{1. Absch.}$$

$$\leq \frac{\Theta^i}{1-\Theta}\, d(x_0, x_1) \longrightarrow \text{2. Absch.}$$

(ii) $(i) \Rightarrow (x_i)_{i=0,\dots}$ Cauchy-Folge $\overset{\text{Eval.}}{\Longrightarrow}$ $x_i \xrightarrow[i\to\infty]{} x^* \in E$

$$0 \leq d(x^*, \phi(x^*)) \leq d(x^*, x_i) + d(x_i, \phi(x^*))$$

$$\leq d(x^*, x_i) + \Theta\, d(x_{i-1}, x^*) \xrightarrow[i\to\infty]{} 0 \Rightarrow \boxed{x^* = \phi(x^*)}$$

$$d(y^*, x^*) = d(\phi(y^*), \phi(x^*)) \leq \Theta\, d(y^*, x^*) \Rightarrow d(y^*, x^*) = 0 \Rightarrow y^* = x^*$$

2 Fixpkt'e

$\boxed{\text{eindt. Fixpunkt}}$

$\square$

6.2. **Lemma** X normierter Raum, $E \subset X$ konvex, $F: X \to X$ stetig diff'bar, $F(E) \subset E$, dann ist

F kontrahierend auf E gdw $\sup_{x \in E} |||DF(x)||| \leq \Theta < 1$

Bew $\boxed{\text{"}\Leftarrow\text{"}}$ $\|F(x+\delta) - F(x)\| = \|\int_0^1 \partial_t F(x+t\delta)\, dt\|$

$$= \|\left(\int_0^1 DF(x+t\delta)\, dt\right)\delta\| \leq \sup_{t\in[0,1]} |||DF(x+t\delta)|||\, \|\delta\|$$

$$\leq \Theta\, \|\delta\| \quad (\checkmark)$$

$\boxed{\text{"}\Rightarrow\text{"}}$ $|||DF(y)||| \geq 1$ für ein $y \in E \Rightarrow \exists \delta$

$\|DF(y)\delta\| \geq \|\delta\| \Rightarrow$

$$\Rightarrow \quad \|F(y+\epsilon\delta) - F(y)\| = \left\|\left(\int_0^1 DF(y+\epsilon\delta)\,dt\right)\epsilon\delta\right\|$$

$$\geq \epsilon(1-\tilde{\epsilon})\|\delta\| \quad \text{für } \tilde{\epsilon} \xrightarrow[\epsilon\to 0]{} 0$$

wäre F kontrahierend $\Rightarrow$ $(1-\tilde{\epsilon})\|\epsilon\delta\| \leq \|F(y+\epsilon\delta)-F(y)\| \leq \Theta\|\epsilon\delta\|$

$$\Rightarrow \quad 1-\tilde{\epsilon} \leq \Theta \Rightarrow \Theta = 1 \;\lightning$$

<u>Kontraktion des Newtonverfahren (1D):</u>

$$\phi(x) = x - \frac{f(x)}{f'(x)} \quad \Rightarrow$$

(i) $$|\phi(x)-\phi(y)| = \left|(x-y) - \left(\frac{f(x)}{f'(x)} - \frac{f(y)}{f'(y)}\right)\right|$$

$$= \frac{1}{|f'(x)|}\,|f'(x)(x-y) - f(x) + f(y)| + |f(y)|\underbrace{\left|\frac{1}{f'(x)} - \frac{1}{f'(y)}\right|}_{\leq |f''(\xi)|\frac{1}{|f'(\xi)|^2}|x-y|}$$

$\xi \in (x,y)$

$$\leq \Theta\,|x-y|$$

falls $|f'(x)| > c$, $|x-y| \ll 1$, $|f(y)| \ll 1$

<u>oder speziell:</u> $f \in C^2$, x, y nahe bei Nullstelle x^* und $f'(x^*) \neq 0$

(ii) $\boxed{\text{zu Lemma 6.2.}}$

$$|\phi'(x)| = \left|1 - \frac{f'(x)}{f'(x)} + \frac{f''(x)f(x)}{f'(x)^2}\right| = |f(x)|\frac{|f''(x)|}{|f'(x)|^2} \overset{!}{\leq} \Theta < 1$$

6.3. **Def** (Konvergenz)

X endl. dim. Vektraum, $\|.\|$ Norm auf X, dann konvergiert $(x_k)_{k\in\mathbb{N}}$ von p-ter Ordnung, falls

$$x_k \xrightarrow[k\to\infty]{} x^* \in X \quad \text{und} \quad \|x_{k+1} - x^*\| \le C \|x_k - x^*\|^p$$

($p=1$ dann sei $C<1$, C heißt Konvergenzrate)

und x_k konvergiert superlinear, falls eine Nullfolge $(C_k)_{k\in\mathbb{N}}$ existiert mit $\|x_{k+1} - x^*\| \le C_k \|x_k - x^*\|$.

6.4. **Satz** (superlineare Konvergenz des Newtonverfahrens)

X, Y endl. dim. Vektraume, $\mathcal{U} \subset X$ offen, $F \in C^1(\mathcal{U}, Y)$, $F'(x)$ invertierbar für alle $x \in \mathcal{U}$, $x^* \in \mathcal{U}$ mit $F(x^*) = 0$, dann existiert Umgebung $\mathcal{U}^*$ von x^* mit $\mathcal{U}^* \subset \mathcal{U}$, so dass das Newtonverfahren $x_{k+1} = x_k - DF(x_k)^{-1} F(x_k)$ zu $x_0 \in \mathcal{U}^*$ durchführbar ist und $(x_k)_k$ superlinear gegen x^* konvergiert.

Bew Zu x definiere x_+ als Lösung von $F(x) + F'(x)(x_+ - x) = 0$

$F'(x^*)$ regulär $\Rightarrow$

$$\|x_+ - x^*\| = \|F'(x^*)^{-1} F'(x^*)(x_+ - x^*)\|$$
$$\le |||F'(x^*)^{-1}||| \; \|F'(x^*)(x_+ - x^*)\|$$

$$\Rightarrow \underbrace{|||F'(x^*)^{-1}|||^{-1}}_{=:m} \|x_+ - x^*\| \le |||(F'(x^*) - F'(x))||| \; \|x_+ - x^*\| + \|F'(x)(x_+ - x^*)\|$$

wähle nun $\epsilon < \frac{m}{2}$ $\Rightarrow$ $\exists \delta > 0$, so dass

$$|||F'(x) - F'(x^*)||| < \epsilon \quad \forall x \in B_\delta(x^*) \subset \mathcal{U}$$

und es folgt

$$\frac{m}{2}\|x_+ - x^*\| \le (m-\epsilon)\|x_+ - x^*\|$$

$$\le \|F'(x)(x_+ - x^*)\|$$

$$= \|F'(x)(x_+ - x^*) - \underbrace{(F'(x)(x_+ - x) + F(x))}_{=0} + \underbrace{F(x^*)}_{=0}\|$$

$$= \|F(x^*) - F(x) - F'(x)(x^* - x)\|$$

$$= \|\int_0^1 \partial_t F(x + t(x^* - x))\,dt - F'(x)(x^* - x)\|$$

$$= \|\int_0^1 F'(x + t(x^* - x)) - F'(x)\,dt\,(x^* - x)\|$$

$$\le \underbrace{|||\int_0^1 (F'(x + t(x^* - x)) - F'(x))\,dt|||}_{<\epsilon \quad \text{falls } x \in B_\delta(x^*)} \|x^* - x\|$$

$$\Rightarrow \|x_+ - x^*\| \le \frac{2\epsilon}{m}\|x^* - x\|$$

D.h $\quad x_0 \in B_\delta(x^*) \Longrightarrow x_k \in B_\delta(x^*) \quad \forall k \ge 0$

$$\frac{2\epsilon}{m} < 1 \quad \left(\epsilon < \frac{m}{2}\right)$$

Ferner konvergiert x_k mit Rate $\frac{2\epsilon}{m}$ superlinear $\square$

6.5 <u>Folgerung</u> (<u>quadratische Konvergenz</u>)

Vor. wie in 6.4., F' zusätzlich in x^* Lipschitz-stetig, dann konvergiert das Newtonverfahren (lokal) quadratisch.

<u>Bew</u> $\quad |||F'(x + t(x^* - x)) - F'(x)|||$

$$\le |||F'(x + t(x^* - x)) - F'(x^*)||| + |||F'(x^*) - F'(x)|||$$

$$\le 2\underset{\substack{\uparrow \\ \text{Lipschitz-Konstante}}}{L}\|x^* - x\|$$

$$\Rightarrow \|x_+ - x^*\| \le \frac{4L}{m}\|x_+ - x^*\|^2 \qquad \square$$

Das Newtonverfahren als Prototyp eines schnell konvergierenden Verfahrens:

6.6 **Satz** X, Y endl. dim. Vektorräume, $\mathcal{U} \subset X$ offen, $\{x_k\}$ Folge in $\mathcal{U}$, $x_k \longrightarrow x^* \in \mathcal{U}$, $F \in C^1(\mathcal{U}, Y)$, $F(x^*) = 0$, $F'(x^*)$ invertierbar, $x_k \neq x_{k+1}$ für $k \to \infty$, dann sind folgende Aussagen äquivalent:

(i) $x_k \longrightarrow x^*$ superlinear

(ii) $$\frac{\| F(x_k) + F'(x^*)(x_{k+1} - x_k) \|}{\| x_{k+1} - x_k \|} \xrightarrow{k \to \infty} 0$$

(iii) $$\frac{\| F(x_k) + F'(x_k)(x_{k+1} - x_k) \|}{\| x_{k+1} - x_k \|} \xrightarrow{k \to \infty} 0$$

Bew $\boxed{(ii) \Rightarrow (i)}$ $\quad s_k = x_{k+1} - x_k$

Da $x_k \to x^* \Rightarrow \exists_\epsilon \; \forall_{k > k_0(\epsilon)} \; x_k \in B_\epsilon(x^*) \subset \mathcal{U}$

$$\| F(x_{k+1}) \| \leq \| F(x_{k+1}) - F(x_k) - F'(x^*) s_k \| + \| F(x_k) + F'(x^*) s_k \|$$

$$\leq \left\| \int_0^1 F'(x_k + t s_k) - F'(x^*) \, dt \right\| \| s_k \| + \| F(x_k) - F'(x^*) s_k \|$$

$$\Rightarrow \frac{\| F(x_{k+1}) \|}{\| s_k \|} \leq \underbrace{\left\| \int_0^1 F'(x_k + t s_k) - F'(x^*) \, dt \right\|}_{\xrightarrow{} 0 \; F' \text{ stetig}} + \underbrace{\frac{\| F(x_k) - F'(x^*) s_k \|}{\| s_k \|}}_{\xrightarrow[(ii)]{} 0}$$

Da $F'(x^*)$ invertierbar $\exists_{m > 0} \; \forall_{k \text{ groß}}$

$$\| F'(x^*) e_k \| \geq m^- \| e_k \| \text{ mit } m^- = \| F'(x^*)^{-1} \|^{-1}$$

$$e_k = x_k - x^*$$

$$\left(\| F'(x^*) e_k \| \leq m^+ \| e_k \| \text{ mit } m^+ = \| F'(x^*) \| \right)$$

$$\|F(x_{k+1})\| = \|F(x_{k+1}) - F(x^*)\|$$

$$\underset{(\leq)}{\geq} \underbrace{\|F'(x^*)e_{k+1}\|}_{\geq m^- \|e_{k+1}\| \ (\leq)\ m^+\|e_{k+1}\|} \underset{(+)}{-} \underbrace{\|F(x_{k+1}) - F(x^*) - F'(x^*)e_{k+1}\|}_{= o(e_{k+1})}$$

(nur zum Vgl.)

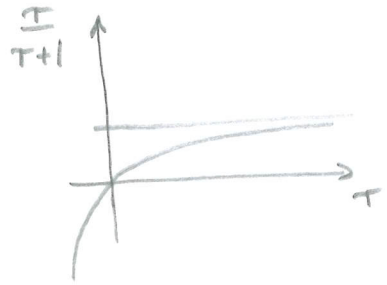
$$\Rightarrow \quad \frac{\|F(x_{k+1})\|}{\|e_{k+1}\|} \begin{cases} \leq m^+ + o(1) \\ \geq (m^- - o(1)) \end{cases}$$

$$\Rightarrow \quad \frac{\|F(x_{k+1})\|}{\underbrace{\|x_{k+1} - x_k\|}_{\to 0}} = \frac{\|F(x_{k+1})\|}{\|x_{k+1} - x^* + x^* - x_k\|} \geq \frac{\|F(x_{k+1})\|}{\|e_{k+1}\| + \|e_k\|} \frac{\|e_k\|}{\|e_{k+1}\|}$$

$$\geq (m^- - o(1)) \frac{\|e_{k+1}\|}{\|e_{k+1}\| + \|e_k\|}$$

$$\Rightarrow \quad \frac{\|e_{k+1}\|}{\underbrace{\|e_{k+1}\| + \|e_k\|}} \xrightarrow{k\to\infty} 0$$

$$\frac{\frac{\|e_{k+1}\|}{\|e_k\|}}{\underbrace{\frac{\|e_{k+1}\|}{\|e_k\|}}_{=:\tau_k} + 1} = \frac{\tau_k}{\tau_k + 1}$$



$$\Rightarrow \quad \tau_k \to 0 \quad \Rightarrow \quad \text{superlinear Konvergenz}$$

$\boxed{(i) \Rightarrow (iii)}$

$s_k = x_{k+1} - x_k = e_{k+1} - e_k$

$$\frac{\|F(x_k) + F'(x_k)s_k\|}{\|s_k\|} \leq$$

$$\frac{\|e_k\|}{\|s_k\|} \left(\frac{\|F(x_k) - F(x^*) - F'(x^*)e_k\|}{\|e_k\|} + \frac{\|F'(x^*)e_k - F'(x_k)e_k\|}{\|e_k\|} + \frac{\|F'(x_k)e_{k+1}\|}{\|e_k\|} \right)$$

$$\leq \underbrace{\frac{\|e_k\|}{\|e_{k+1}-e_k\|}}_{\frac{1}{\left\|\underbrace{\frac{e_{k+1}}{\|e_k\|}}_{\to 0}-\underbrace{\frac{e_k}{\|e_k\|}}_{\|\cdot\|=1}\right\|}\xrightarrow[k\to\infty]{} 1} \underbrace{\left(\mathcal{O}(1) + \overset{F' \text{ stetig}}{\mathcal{O}(1)} + |||F'(x_k)||| \underbrace{\frac{\|e_{k+1}\|}{\|e_k\|}}_{\to 0} \right)}_{\xrightarrow[k\to\infty]{} 0}$$

$$\Rightarrow \quad \frac{\| F(x_k) - F'(x_k) s_k \|}{\| s_k \|} \xrightarrow[k\to\infty]{} 0$$

$\boxed{(iii) \Rightarrow (ii)}$ $$\frac{\|(F'(x_k) - F'(x^*)) s_k\|}{\| s_k \|} \leq |||F'(x_k) - F'(x^*)||| \xrightarrow[k\to\infty]{} 0$$

$\square$

6.7 <u>Def</u> Eine Folge $(x_k)_{k\in\mathbb{N}} \subset \mathcal{U}$ heißt Newton-ähnlich, falls
$$\frac{\| F(x_k) + F'(x_k)(x_{k+1} - x_k) \|}{\| x_{k+1} - x_k \|} \xrightarrow[k\to\infty]{} 0$$
für $x_{k+1} \neq x_k$.

Eine Iteration heißt Newton ähnlich, falls jede Iterationsfolge zu $x_0 \in \mathcal{U}$ Newton-ähnlich ist.

<u>Bem</u>
- Newton-Verfahren ist Newton-ähnlich
- superlineare Konvergenz $\Leftrightarrow$ Newtongleichung
$$F(x_k) + F'(x_k)(x_{k+1} - x_k) = \mathcal{O}(\|x_{k+1} - x_k\|)$$
- Wenn man $F'(x_k)$ durch eine Matrix A_k ersetzt $\left(0 = F(x_k) + A(x_k)(x_{k+1} - x_k) \right)$, so sollte

$$|||A_k - F'(x^*)||| \xrightarrow[k\to\infty]{} 0$$

<u>Nachteil beim Newton-Verfahren</u>: $F'(x_k)$ muß berechnet werden

Ein Ausweg: Approximation mit Differenzenquotienten

$\Rightarrow \quad x_{k+1} \rightsquigarrow \hat{x}_{k+1}$;

interpretiert als gestörte Fixpunktabbildung:

$$x_{k+1} = x_k - F'(x_k)^{-1} F(x_k)$$

$$\hat{x}_{k+1} = \hat{x}_k - A(\hat{x}_k)^{-1} F(\hat{x}_k)$$

verbindet durch

6.8. <u>Satz</u> Vor. wie oben; falls $|||A(x)^{-1} - F'(x)^{-1}||| < \beta$ für β klein, dann $\exists\, \epsilon > 0$, so dass jede Iterationsfolge $(\hat{x}_k)_k$ mindestens linear gegen x^* konvergiert.

<u>Bew</u> $\quad x_+ = x - F'(x)^{-1} F(x)$

$$\forall_{\delta>0}\ \exists_{\epsilon>0} \text{ mit } \|x_+ - x^*\| \le \delta \|x - x^*\| \quad \forall_{x \in B_\epsilon(x^*)}$$

Für $\hat{x}_+ = x - A(x)^{-1} F(x)$ gilt:

$$\frac{\|\hat{x}_+ - x^*\|}{\|x - x^*\|} \le \frac{\|\hat{x}_+ - x_+\| + \|x_+ - x^*\|}{\|x - x^*\|}$$

$$\le \frac{\|(A(x)^{-1} - F'(x)^{-1}) F(x)\| + \|x_+ - x^*\|}{\|x - x^*\|}$$

$$\le \underbrace{\beta \underbrace{\frac{\|F(x) - F(x^*)\|}{\|x - x^*\|}}_{\le L \text{ (Lipschitz-Konstante zu } F)} + \delta}_{\le \Theta << 1 \text{ falls } \beta \text{ klein}} \qquad \square$$

Nun wähle $A_h(x) = \left(\frac{F_i(x+he_j) - F_i(x-he_j)}{2h} \right)_{i,j=1,\dots,n}$

zentrale Differenzenquotienten

$\Rightarrow \quad |||A_h(x) - F'(x)||| \leq C h^2$ falls $F \in C^3$

Ferner: $A_h(x)^{-1} = \left(F'(x) - (F'(x) - A_h(x))\right)^{-1}$

$$= F'(x)^{-1} \Big(\underbrace{\mathbb{1} - (F'(x) - A_h(x)) F'(x)^{-1}}_{\text{Neumann'sche Reihe}} \Big)^{-1}$$

$\Rightarrow$ falls $|||F'(x) - A_h(x)||| \leq |||F'(x)^{-1}|||^{-1}$

dann ist $A_h(x)$ invertierbar und

$$|||A_h(x)^{-1} - F'(x)^{-1}||| \xrightarrow[h \to 0]{} 0$$

6.9 <u>Satz</u> Vor. wie oben, dann gibt es ϵ, h, so dass für $0 < h_k < h$ die Folge

$$x_{k+1}^h = x_k^h - \left(A_{h_k}(x_k^h)\right)^{-1} F(x_k^h)$$

mit $x_0^h \in B_\epsilon(x^*)$ gegen x^* linear konvergiert.

Falls $h_k \xrightarrow[k \to \infty]{} 0$, dann ist die Konvergenz superlinear.

<u>Bew</u> lineare Konvergenz $\Leftarrow$ 6.8.

superlineare Konvergenz $\underset{6.6.}{\Leftrightarrow}$ Newton-ähnlich

$$\frac{\|F(x_k^h) - F'(x_k^h) s_k\|}{\|s_k\|} = \frac{\|\overbrace{F(x_k^h) + A_h(x_k^h) s_k}^{=0} + (F'(x_k^h) - A_h(x_k^h)) s_k\|}{\|s_k\|}$$

$$\leq |||F'(x_k^h) - A_h(x_k^h)||| \leq C h_k^2 \longrightarrow 0 \qquad \square$$

<u>Bem:</u> Die Konvergenz ist sogar quadratisch

Nun betrachten wir approximative Newton-Verfahren, sogenannte Quasi-Newton-Verfahren:

statt: $x_{k+1} = x_k + F'(x_k)^{-1} F(x_k)$

betrachte: $x_{k+1} = x_k + B_k^{-1} F(x_k)$ $\quad x_0$ Startwert

$B_{k+1} = \mathcal{B}(B_k, ___)$ $\quad B_0$ Startwert

wir verlangen: $B_{k+1}(x_{k+1} - x_k) = F(x_{k+1}) - F(x_k)$

(Quasi-Newton-Gleichung)

Rang-1-Bedingung in Approximation von

$$\int_0^1 F'(x_k + t(x_{k+1} - x_k))\,dt\,(x_{k+1} - x_k) = F(x_{k+1}) - F(x_k)$$

nur in Richtung $(x_{k+1} - x_k)$

B_{k+1} Approximation von F' in Richtung $(x_{k+1} - x_k)$

Lösungsmenge: $\boxed{Q(s_k, y_k) := \{ B \in \mathbb{R}^{n,n} \mid B s_k = y_k \} \quad \text{mit } s_k = x_{k+1} - x_k,\ y_k = F(x_{k+1}) - F(x_k)}$

Bem: $Y_{k+1} := \int_0^1 F'(x_k + t s_k)\,dt \in Q(s_k, y_k)$

Ansatz: $B_{k+1} = \mathcal{B}(B_k, s_k, y_k)$

wähle B_{k+1} als beste Approximation von B_k in $Q(s_k, y_k)$ bzgl. einer noch zu wählenden Norm.

wählen wir als Norm: $\|A\|_F = \sqrt{\sum_{i,j=1,\dots,n} a_{ij}^2}$

für $A = (a_{ij})_{i,j=1,\dots,n}$

(Frobenius-Norm)

6.10. <u>Satz</u> $A \in \mathbb{R}^{n,n}$, $s, y \in \mathbb{R}^n$, $s \neq 0$, dann hat das Minimierungsproblem

$$\min_{B \in Q(s,y)} \|B - A\|_F$$

die eindt. Lösung

$$B^* = A + \frac{(y - As)s^T}{s \cdot s}$$

<u>Bew</u> $\|A\|_F^2 = A \cdot A$ mit $A \cdot B = tr(A^T B)$ $\forall A, B \in \mathbb{R}^n$

$\Longrightarrow$ Bestapproximation auf linearen Teilraum ($Q(s,y)$) eindt bestimmt

Projektionssatz

mit $A - B^* \perp_F Q(s,y)$ ($\perp_F$ bzgl. $A \cdot B$)

Für $\tilde{B} = A + \frac{(y - As)s^T}{s \cdot s}$ gilt

$$\|\tilde{B} - A\|_F = \left\|\frac{(y - As)s^T}{s \cdot s}\right\|_F \underset{\substack{y = Bs \\ \forall B \in Q(s,y)}}{=} \left\|(B - A)\frac{ss^T}{s \cdot s}\right\|_F$$

$$\underset{\boxed{(*)\ \|AB\|_F \leq \|A\|_F \|B\|_F}}{\leq} \|B - A\|_F \underbrace{\left\|\frac{ss^T}{s \cdot s}\right\|_F}_{\underset{\boxed{(**)\ \|uv^T\|_F = \|u\|\,\|v\|}}{=} 1} = \|B - A\|_F \quad \forall B \in Q(s,y)$$

$\Longrightarrow \tilde{B} = B^*$

zu (*):

$$\|AB\|_F^2 = \operatorname{tr}((AB)^T AB)$$
$$= \operatorname{tr}(B^T A^T A B)$$
$$= \operatorname{tr}(A^T A B B^T) = (AA^T)\cdot(BB^T)$$
$$\underset{\text{Cauchy Schwarz}}{\leq} \|AA^T\|_F \, \|BB^T\|_F$$
$$= \sqrt{\operatorname{tr}(AA^T AA^T)} \sqrt{\operatorname{tr}(BB^T BB^T)}$$
$$\underset{\text{Cauchy Schwarz}}{\leq} \operatorname{tr}(A^T A) \quad \operatorname{tr}(B^T B)$$
$$= \|A\|_F^2 \, \|B\|_F^2$$

zu (**):

$$\|uv^T\|_F^2 = \operatorname{tr}((uv^T)^T(uv^T))$$
$$= \operatorname{tr}(v \underbrace{u^T u}_{\|u\|^2} v^T) = \|u\|^2 \operatorname{tr}(vv^T)$$
$$= \|u\|^2 \|v\|^2$$

$\square$

$$= \mathbb{1} + A^{-1} u v^T - (1 + v \cdot A^{-1} u) \frac{A^{-1} u v^T}{1 + v \cdot A^{-1} u}$$

$$= \mathbb{1} \qquad \square$$

6.12. <u>Algorithmus</u> (<u>Broyden-Verfahren</u>)

$x_0 \in \mathbb{R}^n$; $B_0 \in \mathbb{R}^{n,n}$ Startwerte;

$$x_{k+1} = x_k - B_k^{-1} F(x_k);$$

$$s_k = x_{k+1} - x_k;$$

$$y_k = F(x_{k+1}) - F(x_k);$$

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k) s_k^T}{s_k \cdot s_k};$$

[<u>Alternativ (inverse Update):</u>

$$B_{k+1}^{-1} = B_k^{-1} - \frac{(B_k^{-1} y_k - s_k) s_k^T B_k^{-1}}{s_k \cdot B_k^{-1} y_k}$$]

<u>zur Alternative:</u> wähle in 6.11. $u = y_k - B_k s_k$, $v = \frac{s_k}{s_k \cdot s_k}$

$$\Rightarrow B_{k+1}^{-1} = B_k^{-1} - \frac{B_k^{-1}(y_k - B_k s_k) s_k^T B_k^{-1} \frac{1}{s_k \cdot s_k}}{1 + \frac{s_k}{s_k \cdot s_k} \cdot B_k^{-1}(y_k - B_k s_k)}$$

$$= B_k^{-1} - \frac{(B_k^{-1} y_k - s_k) s_k^T B_k^{-1}}{s_k \cdot s_k + s_k \cdot B_k^{-1} y_k - s_k \cdot s_k}$$

$$= \text{s.o.} \qquad (\checkmark)$$

<u>Bem</u> • Update iterativ zu berechnen (nur für n klein volle Matrix sp.)

<u>Update B_k</u>: $n \gg 1$ Löse Gl. system iterativ

z.B. bei Minimierungsaufgaben mit CG

Update B_k^{-1}: $n >> 1$ Berechne $B_k^{-1} F(x_k)$ iterativ

- Speichern von Rang 1 - Matrizen:

$\boxed{\text{für } B_k}$ B_0; $(y_k - B_k s_k)$, s_k^T für $k = 1, \ldots$

$\boxed{\text{für } B_k^{-1}}$ B_0^{-1}; $(B_k^{-1} y_k - s_k)$, $B_k^{-1} y_k$, s_k für $k = 1, \ldots$

- nach m Schritten: Neustart

Aufwand: $O(kn)$

4.13 **Lemma** Ein Quasi-Newtonverfahren mit $x_k \to x^*$, $F'(x^*)$ invertierbar, $B_{k+1} - B_k \xrightarrow[k\to\infty]{} 0$ konvergiert superlinear und $F(x^*) = 0$.

Bew $B_k s_k = (B_k - B_{k+1}) s_k + B_{k+1} s_k$

$= (B_k - B_{k+1}) s_k + Y_{k+1} s_k$

$$Y_{k+1} = \underbrace{\int_0^1 F'(x_k + t(x_{k+1} - x_k))\, dt}_{\in Q(s_k, y_k)}$$

$\Rightarrow B_{k+1} s_k = y_k = Y_{k+1} s_k$

$$\Rightarrow \frac{\|(B_k - F'(x_k)) s_k\|}{\|s_k\|} \le \underbrace{|||B_k - B_{k+1}|||}_{\xrightarrow[k\to\infty]{} 0} + \underbrace{|||Y_{k+1} - F'(x_k)|||}_{\xrightarrow[k\to\infty]{} 0}$$

$\Rightarrow$ Newtonähnlich $\Rightarrow$ superlineare Konvergenz

Forw: $F(x_k) + \underbrace{F'(x_k) s_k}_{\xrightarrow[k\to\infty]{} 0} = \underbrace{F(x_k) - B_k s_k}_{=0} + \underbrace{(B_{k+1} - F'(x_k)) s_k}_{= (Y_{k+1} - F'(x_k)) s_k \xrightarrow[k\to\infty]{} 0} + \underbrace{(B_k - B_{k+1}) s_k}_{\xrightarrow[k\to\infty]{} 0}$

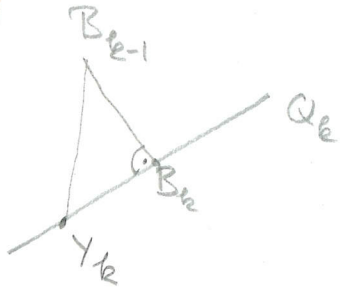
$\Rightarrow F(x_k) \xrightarrow[k\to\infty]{} 0$

$\downarrow$

$F(x^*)$

$\square$

6.14 **Satz** $(Q_k)_{k\in\mathbb{N}_0}$ Folge affiner Unterräume in X (Vektorraum mit Skalarprodukt); es gilt $Y_k \in Q_k$ mit $\sum_{k=0}^{\infty} \|Y_{k+1} - Y_k\| < \infty$, $B_0 \in X$ beliebig, B_k Bestapproximation von B_{k-1} auf Q_k; dann ist $(B_k)_{k\in\mathbb{N}_0}$ beschränkt und $B_{k+1} - B_k \xrightarrow[k\to\infty]{} 0$.



Bew: $B_k - B_{k-1} \perp Q_k \ni B_k - Y_k \Rightarrow$

$$\underbrace{\|B_{k-1} - Y_k\|^2}_{=b} = \|B_k - Y_k\|^2 + \underbrace{\|B_k - B_{k-1}\|^2}_{=a}$$

$$\left[\; 0 \le a \le b \Rightarrow b^2 - a^2 \le b^2 - \frac{2a^2 b}{2b} + \frac{a^4}{4b^2} = \left(b - \frac{a^2}{2b}\right)^2 \;\right]$$

$$\|B_k - Y_k\| = \sqrt{\|B_{k-1} - Y_k\|^2 - \|B_k - B_{k-1}\|^2}$$

$$\le \|B_{k-1} - Y_k\| - \frac{\|B_{k-1} - B_k\|^2}{2\|B_{k-1} - Y_k\|}$$

$$\le \underbrace{\|B_{k-1} - Y_{k-1}\|}_{\text{Induktion}} + \|Y_k - Y_{k-1}\| - \frac{\|B_{k-1} - B_k\|^2}{2\|B_{k-1} - Y_k\|}$$

$$\Rightarrow \|B_k - Y_k\| \le \sum_{i=1}^{k} \|Y_i - Y_{i-1}\| - \sum_{i=1}^{k} \frac{\|B_{i-1} - B_i\|^2}{2\|B_{i-1} - Y_i\|} + \|B_0 - Y_0\|$$

$\Rightarrow (B_k)_k$ beschränkt da Y_k beschränkt

insbesondere $\|B_{i-1} - Y_i\| \le C'$

ferner: $$\sum_{i=1}^{k} \frac{\|B_{i-1} - B_i\|^2}{2\|B_{i-1} - Y_i\|} < \infty$$

$$\Rightarrow \quad C > \frac{1}{2} \sum_{i=1}^{k} \frac{\|B_{i-1} - B_i\|^2}{\underbrace{\sqrt{\|B_{i-1} - B_i\|^2 + \|B_i - Y_i\|^2}}_{\geq \sqrt{\|B_{i-1} - B_i\|^2 + C}}}$$

$$\Rightarrow \quad \|B_{k-1} - B_k\| \xrightarrow[k \to \infty]{} 0$$

$\square$

6.15 Satz Eine Quasi-Newton-Verfahren, für das $x_k \to x^*$, $F'(x^*)$ invertierbar, $F \in C^{1,1}$, $\sum_{k=0}^{\infty} \|x_{k+1} - x_k\| < \infty$, ist superlinear konvergent, $F(x^*) = 0$.

Bew $\|Y_{k+1} - Y_k\|_F = \| \int_0^1 F'(x_k + t s_k) - F'(x_{k-1} + t s_{k-1})\, dt \|_F$

$$\leq L \int_0^1 \|(x_k + t s_k) - (x_{k-1} + t s_k)\|\, dt$$

Lipschitz-Konstante zu F' ($F \in C^{1,1}$)

$= s_{k-1}$

$$\leq L \int_0^1 (1-t)\|s_{k-1}\| + t\|s_k\|\, dt$$

$$\leq \frac{L}{2}(\|s_{k-1}\| + \|s_k\|)$$

$$\overset{\sum_{k=0}^{\infty} \|s_k\| < \infty}{\Longrightarrow} \quad \sum_{k=0}^{\infty} \|Y_{k+1} - Y_k\|_F < \infty \overset{6.14}{\Longrightarrow} (B_k)_k \text{ beschränkt}$$

$X = \mathbb{R}^{n,n}$, $\|\cdot\|_F$ als Norm

$$B_{k+1} - B_k \xrightarrow{k \to \infty} 0$$

Ferner $Y_k \longrightarrow F'(x^*)$ da $x_k \to x^*$

$$\overset{6.13}{\Longrightarrow} \quad x_k \text{ konvergiert superlinear gegen } x^*$$

$\square$

6.16 **Bem:** Wann konvergiert ein Quasi-Newton-Verfahren mit $\sum_{k=0}^{\infty} \|x_{k+1} - x_k\| < \infty$?

$F(x) = \nabla E(x)$ mit E strikt konvex und mit Schrittweitensteuerung:

$$x_{k+1} = x_k - \alpha_k B_k^{-1} F(x_k)$$

α_k so wählbar, dass $\sum_{k=1}^{\infty} \|x_{k+1} - x_k\| < \infty$

$\hookrightarrow$ Schrittweitensteuerung:

$\boxed{\text{Armijo-Regel}}$

$d = -F'(x_k) F(x_k)$ (Suchrichtung)

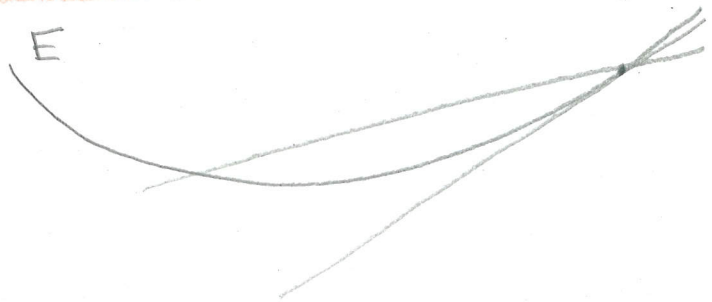
$\lambda = 1$ (voller Schritt)

① if $\|F(x_k + \lambda d)\| > (1 - \alpha\lambda) \|F(x_k)\|$
then $x_{k+1} = x_k + \lambda d$;
else $\{ \lambda = \beta\lambda$; goto ① $\}$

Mindestabnahme

typische Werte: $\beta \in [0.1, 0.5]$
$\alpha = 10^{-4}$

Armijo-Regel für Minimierungsprobleme:



$E(x_k + \lambda d) <$
$E(x_k) + \alpha\lambda \nabla' E(x_k) \cdot d$

Powell-Wolfe-Regel

d Abstiegsrichtung $\quad d = -D^2E(x_k)\nabla E(x_k)$

mit $\quad d \cdot \nabla E(x_k) < 0$

Bedingungen: (i) $E(x_k + \lambda d) < E(x_k) + \alpha \lambda \nabla E(x_k) \cdot d$

(ii) $\nabla E(x_k + \lambda d) \cdot d \geq \beta \nabla E(x_k) \cdot d$

$\boxed{\text{zu (ii)}}$ Optimaler Schritt $\quad E(x_k + \lambda^* d) = \min_{\lambda} E(x_k + \lambda d)$

$$\Longleftrightarrow \quad \nabla E(x_k + \lambda^* d) \cdot d = 0 \geq \beta \underbrace{\nabla E(x_k) \cdot d}_{\text{Abstiegsrichtung}}$$

D.h. $\tilde{\lambda} \mapsto E(x + \tilde{\lambda} d)$ ist weniger steil (nach unten) für $\tilde{\lambda} = \lambda$ wie für $\lambda = 0$.

(i) Schrittweitenbeschränkung

(ii) Bedingung für einen mindestens zu gehenden Schritt

Gauß - Newton - Verfahren

Wir suchen nicht eine Lösung x^* von $F(x^*) = 0$ mit $F: \mathbb{R}^n \to \mathbb{R}^n$, sondern eine Minimalstelle $x^* \in \mathbb{R}^n$ von

$$f: \mathbb{R}^n \to \mathbb{R} \quad \text{mit} \quad f(x) = \frac{1}{2} \sum_{i=1}^{m} |F_i(x)|^2$$

wobei $F(x) = (F_i(x))_{i=1,\dots,m}$ $\quad (F: \mathbb{R}^n \to \mathbb{R}^m)$

Bem: $m > n$ überdeterminiert

$m < n$ unterdeterminiert

$$\nabla f(x) = \sum_{i=1}^{m} F_{i,j}(x) F_i(x) = \underbrace{DF(x)^T}_{\mathbb{R}^{n,m}} \underbrace{F(x)}_{\mathbb{R}^m} \in \mathbb{R}^n$$

notwendige Bedingung für x^*: $0 = \nabla f(x^*) = DF(x^*)^T F(x^*)$

$m < n$, $\operatorname{Rg} DF(x^*) = m \Rightarrow F(x^*) = 0$

Dies gilt nicht im überdeterminierten Fall.

Hesse-Matrix
$$D^2 f(x) = \sum_{i=1}^{m} \left(F_{i,j}(x) F_{i,k}(x) + F_{i,jk}(x) F_i(x) \right)$$
$$= DF(x)^T DF(x) + \sum_{i=1}^{m} D^2 F_i(x) F_i(x)$$

Falls $F(x^*) = 0$, dann ist

$DF(x)^T DF(x)$ eine gute Approximation für $D^2 f(x)$ und $\|x - x^*\|$ klein.

Newton-Verfahren: $x_{k+1} = x_k - D^2 f(x_k) \nabla f(x_k)$

Gauß-Newton-Verfahren: $x_{k+1} = x_k - (DF(x_k)^T DF(x_k))^{-1} DF(x_k)^T F(x_k)$

6.17 Satz Sei x^* lokales Minimum von $\|F(x)\|^2$ für $F: \mathbb{R}^n \to \mathbb{R}^m$, $F \in C^{1,1}$, $DF(x^*)$ habe maximalen Rang, dann gibt es $C > 0$, $\delta > 0$, so dass für den Gauß-Newton-Schritt

$m > n$: Spaltenrang
$m \leq n$: Zeilenrang

$$x_{k+1} = x_k - \left(DF(x_k)^T DF(x_k)\right)^{-1} DF(x_k)\, F(x_k)$$

gilt:

$$\|x_{k+1} - x^*\| \leq C\left(\|x_k - x^*\|^2 + \|F(x^*)\|\,\|x_k - x^*\|\right)$$

falls $\|x_k - x^*\| < \delta$.

Bew δ so klein, dass $\mathrm{Rg}(DF(x)^T DF(x)) = n$

$$x_+ = x - \left(DF(x)^T DF(x)\right)^{-1} DF(x)^T F(x) \quad \Rightarrow$$

$$x_+ - x^* = \left(DF(x)^T DF(x)\right)^{-1} DF(x)^T \underbrace{\left(DF(x)(x - x^*) - F(x)\right)}$$

$$= DF(x)(x - x^*) - F(x) + F(x^*) - F(x^*)$$

$$\|F(x^*) - F(x) - DF(x)(x^* - x)\|$$

$$= \left\|\left(\int_0^1 DF(x + t(x^* - x))\,dt\right)(x^* - x) - DF(x)(x^* - x)\right\|$$

$$\leq L\left(\int_0^1 t\,dt\right)\|x^* - x\|^2 = \frac{L}{2}\|x^* - x\|^2$$

Lipschitz-Konstante von DF

x^* lokales Minimum $\Rightarrow$ $DF(x^*)^T F(x^*) = 0$

$$\Rightarrow \quad -DF(x)^T F(x^*) = \left(DF(x^*) - DF(x)\right)^T F(x^*)$$

$$\Rightarrow \quad \|x_+ - x^*\| \leq \left|\left|\left|\left(DF(x)^T DF(x)\right)^{-1}\right|\right|\right| \cdot$$

$$\underbrace{\left\|DF(x)^T\left(DF(x)(x - x^*) - F(x) + F(x^*) - F(x^*)\right)\right\|}$$

$$\leq |||DF^T(x)|||\,\frac{L}{2}\|x^* - x\|^2 + \left\|\left(DF(x^*) - DF(x)\right)^T F(x^*)\right\|$$

$$\Rightarrow \|x_+ - x^*\| \leq |||(DF(x)^T DF(x))^{-1}||| \cdot$$
$$\left(|||DF(x)^T||| \frac{L}{2} \|x^* - x\|^2 + L \|x^* - x\| \, \|F(x^*)\| \right)$$

$\Rightarrow$ Für $C' = |||(DF(x)^T DF(x))^{-1}||| \max\left[|||DF(x)^T||| \frac{L}{2}, L \right]$

gilt $\|x_+ - x^*\| \leq C' \left(\|x^* - x\|^2 + \|F(x^*)\| \, \|x^* - x\| \right)$

<u>Bem</u> $F(x^*) = 0 \Rightarrow$ quadratische Konvergenz

$C' \|F(x^*)\| < 1 \Rightarrow$ lineare Konvergenz

7. GMRES

GMRES = Generalized Minimum RESidual

Krylov-Raum-Verfahren für nichtsymm. $Ax = b$

$$T_{k-1} = \left\{ x_0 + \sum_{j=0}^{k-1} \alpha_j A^j r_0 \;\middle|\; \alpha_j \in \mathbb{R} \right\}, \quad S_{k-1} = \operatorname{span}\{A^j r_0\}_{j=0,\dots,k-1}$$

k-te Iterierte x_k ist Lösung von $\min\limits_{x \in T_{k-1}} \|b - Ax\|^2$

$$b - Ax_k = b - Ax_0 + \sum_{j=0}^{k-1} \alpha_j A^{j+1} r_0$$

$$= r_0 - \sum_{j=1}^{k} \alpha_{j-1} A^j r_0$$

siehe CG

$$\Rightarrow \quad x_k \in x_0 + T_{k-1} \Rightarrow r = b - Ax = P_1(A) r_0$$

mit $P_1 \in \mathcal{P}_k^1$ (Polynome, Grad $\leq k$, $P_1(0) = 1$)

Für $r_k = b - Ax_k \Rightarrow$

$$\|r_k\|_2 = \min_{P_1 \in \mathcal{P}_k^1} \|P_1(A) r_0\|_2$$

7.1. Satz $A \in \mathbb{C}^{n,n}$, $A = VDV^{-1}$ mit $V^{-1} = V^*$ ($V^* = \overline{V}^T$)

$D \in \mathbb{C}^{n,n}$ diagonal, dann gilt

$$\|r_k\|_2 \leq \underbrace{|||V|||_2 \, |||V^{-1}|||_2}_{K_2(V)} \min_{P_1 \in \mathcal{P}_k^1} \max_{\lambda \in \sigma(A)} |P_1(\lambda)|$$

Bew: $|||P_1(A)|||_2 = |||P_1(VDV^{-1})|||_2 = |||V P_1(D) V^{-1}|||_2$

$$\leq |||V|||_2 \, |||V^{-1}|||_2 \, |||P_1(D)|||_2 = |||V|||_2 \, |||V^{-1}|||_2 \max_{\lambda \in \sigma(A)} |P_1(\lambda)| \quad \square$$

<u>Bem</u> • In Allgemeinen ist A nicht diagonalisierbar.

• Falls A komplex diagonalisierbar ist, so folgt analog zum CG-Verfahren:

GMRES terminiert im m Iterationen, falls A m verschiedene Eigenwerte besitzt (vgl Ü)

zur Implementierung:

$V_k : \mathbb{R}^n \to S_{k-1}$ orthogonale Projektion

berechnet mit Gram-Schmidt-Verfahren:

$$v_1 = \frac{r_0}{\|r_0\|}$$

$$v_{j+1} = \frac{A v_j - \sum_{i=1}^{j} (A v_j \cdot v_i) v_i}{\| A v_j - \sum_{i=1}^{j} (A v_j \cdot v_i) v_i \|} \quad (*)$$

Matrix-Schreibweise: $\underset{\mathbb{R}^{n,n}}{A} \; \underset{\mathbb{R}^{n,k}}{V_k} = \underset{\mathbb{R}^{n,k+1}}{V_{k+1}} \; \underset{\mathbb{R}^{k+1,k}}{H_k}$ $\left(V_k = \begin{pmatrix} | & & | \\ v_1 & \cdots & v_k \\ | & & | \end{pmatrix} \right)$

mit $H_k = (h_{ij}^k)_{\substack{i=1,\dots,k+1 \\ j=1,\dots,k}}$ $\quad h_{ij} = \begin{cases} A v_j \cdot v_i & ; \; i < j \\ \| A v_j - \sum_{j=1}^{i} (A v_j \cdot v_i) v_i \| & ; \; i = j+1 \\ 0 & ; \; i > j+1 \end{cases}$

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

Hessenbergmatrix

$$\boxed{(*) \Leftrightarrow A v_j = \sum_{i=1}^{j+1} V_i h_{ij}}$$

Ziel: Minimierung von $\|r_k\|_2$ mit $r_k = b - A x_k$

und $x_k - x_0 = \underbrace{V_k \underset{\mathbb{R}^k}{y_k}}_{\in S_{k-1}}$

$$\begin{aligned} r_k &= b - A x_k \\ &= b - A x_0 - A(x_k - x_0) \\ &= r_0 - A V_k y_k \\ &= \|r_0\| V_{k+1} \underset{\mathbb{R}^{k+1}}{e_1} - V_{k+1} H_k y_k \\ &= V_{k+1} (\|r_0\| e_1 - H_k y_k) \end{aligned}$$

$$\Rightarrow \quad \|r_k\|_2 = \| V_{k+1} (\|r_0\| e_1 - H_k y_k) \|_2$$

orthogonal

$$= \| \|r_0\| e_1 - H_k y_k \|_2 \longrightarrow \min$$

lösbar mit QR-Verfahren (mittels Givens-Rotationen)

7.2. <u>Algorithmus</u> GMRES$(x, b, A, \epsilon, K, \varrho)$

$r = b - Ax$; $v_1 = \frac{r}{\|r\|_2}$; $\beta = \varrho = \|r\|_2$; $k = 0$;

while $\varrho > \epsilon \|b\|_2$ and $k < K$ {

$k = k+1$;

for $j = 1, \ldots k$ $\quad h_{jk} = A v_k \cdot v_j$

$$v_{k+1} = A v_k - \sum_{j=1}^{k} h_{jk} v_j$$

$$h_{k+1,k} = \| v_{k+1} \|_2$$

$$v_{k+1} = v_{k+1} / h_{k+1,k};$$

minimiere $\| \beta e_1 - H_k \tilde{y}_k \|_2$ über $\mathbb{R}^k$

und erhalte y_k;

$$\varrho = \| \beta e_1 - H_k y_k \|_2$$

}

$$x_k = x_0 + V_k y_k;$$