Inhalt:

0. Gleitkommaarithmetik, etc
1. a) Eliminationsverfahren zur Lsg. (in GL.
1 b) Lineare Ausgleichsrechnung und SVD (Singulärwertzerlegung)
2 a) Interpolation
   b) numerische Quadration
3. Iterationsverfahren
   a) nichtlineare Gleichungssysteme
   b) lineare GL
4. Eigenwertproblem

0. Grundlagen, Zahlendarstellung

Sei $r \in \mathbb{R}$, „p-adische Darstellung"

$$r = \sigma \sum_{k=-\infty}^{n} a_k b^k \qquad b > 1,\ b \in \mathbb{N},\ k \in \mathbb{Z},\ 0 \leq a_k < b$$

$\sigma = \pm 1$ „Vorzeichen" $\qquad a_k < b-1$ für unendlich viele k

$a_k \neq 0 \Leftrightarrow r \neq 0$, b-Basis, Bsp $b = 10$ Dezimal

Gleitkommazahlen (normiert)

$$x = \sigma \cdot \underbrace{f}_{\text{Matisse}} \cdot b^{\overset{\text{Exponent}}{e}} \quad e \in \mathbb{Z} \quad , b^{-1} \leq f < 1$$

Bsp. $x = 3{,}1415 \qquad x = +0{,}31415 \cdot 10^{1} \quad (e=1)$

$f = 0{,}d_1, \ldots, d_m = \sum_{k=1}^{m} d_k b^{-k},\ d_1 \neq 0 \quad (1 \leq d_k < b)_{k=1,\ldots,m}$

m-Stellenzahl: $r \leq e \leq R$

$\mathbb{M}(b, m, (r, R)) = \mathbb{M}(b, m)$

Single Precision: $b=2, m=24$

Double Precision: $b=2,\ m=53,\ r=-1021,\ R=1023$

kleinste Darstellbare Zahl $\quad |x_{min}| = 0{,}1 \cdot b^r = b^{r-1}\ (\neq 0) >$ UNDERFLOW

größte Darstellbare Zahl $\quad |x_{max}| = 0{,}(b-1)\ldots(b-1) \cdot b^R = (1 - b^{-m}) b^R <$ OVERFLOW

Rundungen

$fl: \mathbb{R} \to \mathbb{M}(b,m)$

$x \in \mathbb{R},\ x>0$ „Standardrundung"

$$fl(x) = \begin{cases} \sum_{k=1}^{m} d_k b^{-k}, & \text{falls } d_{m+1} < \frac{b}{2} \\ \sum_{k=1}^{m} d_k b^{-k} + b^{-m}, & \text{falls } d_{m+1} \geq \frac{b}{2} \end{cases}$$

Bsp: $b=10,\ m=3,\ x_1 = 0{,}25367,\ fl(x_1) = 0{,}254$

$x_2 = 0{,}2535,\ fl(x_2) = 0{,}254$

$x_3 = 0{,}2533,\ fl(x_3) = 0{,}253$

absoluter Fehler: $|x - fl(x)|$

relativer Fehler: $\frac{|x - fl(x)|}{|x|} \leq \frac{\frac{b}{2} b^{-(m+1)} \cdot b^e}{b^{-1} \cdot b^e} = \frac{b^{-(m-1)}}{2} = eps$ „Maschinengenauigkeit"

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

Gleitkommaarithmetik

arithm. $(+, -, \cdot, \div)$

Seien $x, y \in \mathbb{M}(b,m)$

$x \circledast y = fl(x \nabla y) \in \mathbb{M}(b,m)$

Es gilt $x \circledast y = (x \nabla y)(1+\varepsilon)$, mit $|\varepsilon| \leq eps$

Bsp.: $\mathbb{M}(10,3)$, $x = 6590 = 0{,}659 \cdot 10^4$

$y = 1 = 0{,}1 \cdot 10^1$, $z = 4 = 0{,}4 \cdot 10^1$

$x + y + z = 6590 + 1 + 4 = 6595 = 0{,}6595 \cdot 10^4$

$fl(x+y+z) = 0{,}660 \cdot 10^4 = 6600$

$x \oplus y = fl(0{,}6591 \cdot 10^4) = 0{,}659 \cdot 10^4$

$(x \oplus y) \oplus z = fl(0{,}6594) \cdot 10^4 = 0{,}659 \cdot 10^4$

$y \oplus z = fl(0{,}1 + 0{,}4) \cdot 10^1 = 0{,}5 \cdot 10^1 = 5$

$x \oplus (y \oplus z) = fl(6590 + 5) = fl(0{,}6595) \cdot 10^4 = 0{,}660 \cdot 10^4$

$\Rightarrow (x \oplus y) \oplus z \neq x \oplus (y \oplus z)$ nicht assoziativ

Grundregel 1: Summiere in Reihenfolge aufsteigender Beträge!

Bsp: $x - y = 0{,}73563 - 0{,}73441 = 0{,}00122 = 0{,}122 \cdot 10^{-2}$ $(\mathbb{M}(10,3))$

$fl(x) \ominus fl(y) = 0{,}736 \ominus 0{,}734 = 0{,}002 = 0{,}2 \cdot 10^{-2}$

Relativer Fehler: $\frac{|0{,}122 \cdot 10^{-2} - 0{,}2 \cdot 10^{-2}|}{|0{,}122 \cdot 10^{-2}|} \approx 0{,}64 = 64\%$! Auslöschung

Grundregel 2: Vermeide Auslöschungseffekte bei Subtraktion bei nahezu gleich großen Zahlen!

1.4. Landau Symbol

Def. $f, g : D \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$

$x_0 \in \overline{D} \subset \overline{\mathbb{R}}$

1) $f(x) = O(g(x))$, falls $\left|\frac{f(x)}{g(x)}\right| \leq C \quad \forall\, x \in U_\delta(x_0)$

2) $f(x) = o(g(x))$, falls $\lim\limits_{x \to x_0} \frac{f(x)}{g(x)} = 0 \quad (x \neq x_0)$

Bsp.: $f(x) = \frac{x^2+1}{x^3+1} \quad (x \neq -1)$,

1) $f(x) = O\left(\frac{1}{x}\right)$, $x \to 0$ (d.h. $x_0 = 0$)

2) $f(x) = o\left(\frac{1}{x^2}\right)$

1.5. Fehlerfortpflanzung

X, Y norm. Räume z.B. $X = \mathbb{R}^n$

$f: X \to Y$ ($f: D \to Y$, $D \subset X$ Gebiet)

$x \mapsto f(x) = y$, $\tilde{x} = x - \Delta x$, $\tilde{y} = y - \Delta y$

$\left.\delta_x = \frac{\|\Delta x\|}{\|x\|} \left(= \frac{|\Delta x|}{|x|}, X, Y = \mathbb{R}\right) \quad \delta_y = \frac{\|\Delta y\|}{\|y\|}\right\}$ relativer Fehler

(abs. Fehler $\|\Delta x\|$

Problem ist gut konditioniert, falls (Problem: $x \to y = f(x)$)

$\mathcal{H}$: $\frac{\delta_y}{\delta_x}$ „klein" ($\delta_x \to 0$) $\tilde{y} = f(\tilde{x}) = y + \Delta y$

Sei $f: \mathbb{R}^n \to \mathbb{R}$, $x \in \mathbb{R}^n$, $x \to f(x) \in \mathbb{R}$, $f \in C^k(\mathbb{R}^n)$, $k = 1, 2, 3$

$\|x\|^2 = \|x\|_2^2 := \sum_{i=1}^n |x_i|^2$, $x \in \mathbb{R}^n$, $\tilde{x} \in U_\delta(x)$, $x = (x_1, \dots, x_n)^T$

$f(\tilde{x}) = f(x) + f'(x)(\tilde{x} - x) + o(\|\tilde{x} - x\|)$

$= f(x) + f'(x)(\tilde{x} - x) + \frac{1}{2}(\tilde{x} - x)^T \underbrace{f''(x)}_{H_f(x)} (\tilde{x} - x) + o(\|\tilde{x} - x\|^2)$ (bzw. $O(\|\tilde{x} - x\|^3)$)

„Taylorentwicklung"

Jakobi-Matrix: $f'(x) = (\nabla f(x))^T = \left(\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x)\right)$

$f'(x): \mathbb{R}^n \to \mathbb{R}$ linear

$f(\tilde{x}) = f(x) + \sum_{j=1}^n \frac{\partial f}{\partial x_j}(\tilde{x}_j - x_j) + o(\|\tilde{x} - x\|)$

lineare Näherung

$$\left|\frac{\Delta y}{y}\right| = \left|\frac{f(\tilde{x})-f(x)}{f(x)}\right| \left(= \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(x) \cdot \frac{(\tilde{x}_j - x)}{f(x)}\right)$$

$$\leq \sum_{j=1}^{n} \underbrace{\left|\frac{\partial f}{\partial x_j}(x) \cdot \frac{x_j}{f(x)}\right|}_{=q_j} \cdot \underbrace{\left|\frac{(\tilde{x}_j - x)}{x_j}\right|}_{\delta x_j} + O(\|\tilde{x}-x\|^2)$$

$q_j, \delta x_j$ „Fehlerverstärkungsfaktoren"

$$\leq \sum_{j=1}^{n} q_j \, \delta x_j \, (+\ldots)$$

<u>Bsp</u>: 1) $y = f(x_1, x_2) := x_1 \cdot x_2$

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = x_2, \quad \frac{\partial f}{\partial x_2}(x_1, x_2) = x_1$$

$$q_1 = \frac{\partial f}{\partial x_1}(x) \frac{x_1}{f(x)} = x_2 \cdot \frac{x_1}{x_1 x_2} = 1 = q_2$$

$$\delta y = \left|\frac{\Delta y}{y}\right| = q_1 \delta x_1 + q_2 \delta x_2 = \delta x_1 + \delta x_2$$

2) $y = f(x_1, x_2) = x_1 + x_2$

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = 1 = \frac{\partial f}{\partial x_2}(x_1, x_2) \Rightarrow q_1 = 1 \cdot \left|\frac{x_1}{x_1 + x_2}\right| = 1 \cdot \left|\frac{x_2}{x_1 + x_2}\right| = q_2$$

$$\delta y = \left|\frac{x_1}{x_1 + x_2}\right| \delta x_1 + \left|\frac{x_2}{x_1 + x_2}\right| \delta x_2.$$ für $x_1 \approx -x_2$ ist $q_1, q_2 \gg 1$

„Auslösung"

3) Gesucht: kleinere Nullstelle y_2 von

$$y^2 - 2x_1 y + x_2 = 0$$

$$y_2 = x_1 - \sqrt{(x_1^2 - x_2)} = f(x_1, x_2)$$

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = 1 - \frac{1}{2} \frac{2x_1}{\sqrt{x_1^2 - x_2}} \Rightarrow q_1 = \left(1 - \frac{x_1}{\sqrt{x_1^2 - x_2}}\right) \frac{|x_1|}{|x_1 - \sqrt{*}|}$$

$$\frac{\partial f}{\partial x_2}(x_1, x_2) = +\frac{1}{2\sqrt{x_1^2 - x_2}} \; * \Rightarrow q_2 = \frac{1}{2\sqrt{*}} \cdot \left|\frac{x_2}{x_1 - \sqrt{\ }}\right| \left(= \frac{1}{2} - \frac{1}{2} q_1\right)$$

Sei $M = (10, 5)$ $(m=5)$

$x_1 = 0{,}60002 \cdot 10^1$, $x_2 = 0{,}1 \cdot 10^{-1}$

$q_1 \approx 1$, $q_2 \approx 1$ „gut konditioniert"

Lösung: $y = 0{,}83336 \cdot 10^{-3}$

Algorithmus 1 zur Berechnung von $y = f(x_1, x_2)$

$a := x_1 + x_1;$

$b := a - x_2;$

$c := \sqrt{b};$

$y := x_1 - c;$ $\left(-x_1 - \sqrt{x_1^2 - x_2}\right)$

$\Rightarrow y = 0{,}9 \cdot 10^{-3}$ δy groß!

Algorithmus 2

$$\left(y = f(x_1, x_2) = \frac{x_2}{x_1 + \sqrt{x_1^2 - x_2}}\right)$$

$a := x_1 + x_1;$

$b := a - x_2;$

$c := \sqrt{b};$

$d := x_1 + c;$

$y = \frac{x_2}{d}$

$\Rightarrow y = 0{,}83333 \cdot 10^{-3}$ δy „klein"

Kondition - Eigenschaft des Problems

Stabilität - Eigenschaft des Lösungsalgorithmus

Hauptsache trennen

Vorwärtsanalyse

x - Daten (ideal)

$\tilde{x}$ - gestörte Daten (real)

Problem gesucht: $y = f(x)$ ideal

berechnet $\tilde{y} = \tilde{f}(\tilde{x})$ real

z.B. absoluter Fehler

$\|\Delta y\| = \|f(x) - \tilde{\varphi}(\tilde{x})\|$

$$\|f(x) - \tilde{\varphi}(\tilde{x})\| = \underbrace{\|f(x) - f(\tilde{x})\|}_{\text{„Datenfehler" (Kondition)}} + \underbrace{\|f(\tilde{x}) - \varphi(\tilde{x})\|}_{\text{„Verfahrenfehler"}} + \underbrace{\|\varphi(\tilde{x}) - \tilde{\varphi}(\tilde{x})\|}_{\text{„akkumulierter Rundungsfehler"}}$$

In der Regel zu pessimistisch. Gibt noch Rückwärtsanalyse

1.6. Komplexität von Algorithmus

Wichtig sind:

1) Aufbau: Abfolge der Operationen sollen eindeutig definiert sein
2) Bestimmtheit: Jede mögliche Situation sollte erfasst sein
3) Endlichkeit: Der Algo sollte nach „endlicher" Zeit terminieren

Komplexität: $n \in \mathbb{N}$ Daten, $T: \mathbb{N} \to \mathbb{N}$, $n \mapsto T(n)$ Anzahl der Operationen,
Maß für den Arbeitsaufwand ($\approx$ Komplexität)

Anforderungen: 1) hohe Genauigkeit (Stabilität ...)
2) geringe Komplexität

2. Direkte Methoden zur Lsg. LGS

Problem: Gegeben. $\begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = b \in \mathbb{R}^n$, $A = (a_{ij})_{i,j=1}^n$

Gesucht: $(x_i)_{i=1}^n = x \in \mathbb{R}^n$ mit $Ax = b$

Voraussetzungen: A regulär, d.h. A nicht singulär

$\Leftrightarrow \det A \neq 0 \Leftrightarrow \exists A^{-1} \in \mathbb{R}^{n \times n}$

2.1.1. Dreiecksmatrix

Def. 2.1.1 $R \in \mathbb{R}^{n \times n}$ ($R = (r_{ij})_{i,j=1}^n$) heißt „obere Δ-Matrix", falls $r_{i,j} = 0$ für alle $i > j$.

$$\begin{pmatrix} r_{11} & r & r \\ 0 & r_{22} & r \\ 0 & 0 & \\ 0 & 0 & \end{pmatrix} = (\nabla)$$

R ist regulär $\Leftrightarrow \det R = \prod_{i=1}^n r_{ii} \neq 0 \Leftrightarrow r_{i,i} \neq 0 \ \forall i = 1, \dots, n$

$$r_{1,1} x_1 + r_{1,2} x_2 \dots + r_{1,n} x_n = b_1$$
$$\dots \quad r_{2,2} x_2 \dots = b_2$$
$$r_{n-1,n-1} x_{n-1} + r_{n-1,n} x_n = b_{n-1}$$
$$r_{n,n} x_n = b_n \Rightarrow x_n = \frac{b_n}{r_{n,n}}$$
$$\Rightarrow x_{n-1} = \frac{1}{r_{n-1,n-1}} (b_{n-1} - r_{n-1,n} x_n$$

$\underset{\text{„vollst. Induktion"}}{\Rightarrow}$ $x_j = \frac{1}{r_{jj}} \left(b_j - \sum_{k=j+1}^{n} r_{j,k} x_k \right)$ $j = n-1, \dots, 1$, $x_n = \frac{b_n}{r_{n,n}}$ „Rückwärts-Substitution"

(j. Zeile $r_{j,j} x_j + r_{j,j+1} x_{j+1} + \dots + r_{j,n} x_n = b_j$)

Komplexität numberofnonzero

$R \in \mathbb{R}^{n\times n}$ $\# R := nnz R = O(n^2)$

$b \in \mathbb{R}^n$ $\# b = nnz b = O(n)$

Für $j = n-1 \,(\searrow 1)\; 1$
in der Schleifen runter

in Zeile j: $(n-1)$ Multiplikationen $(+1$ Division$)$ $(n-j+1$ Add$)$

$$= T(n) = \sum_{j=n-1}^{1} (n-j) + n \qquad \left[\sum_{j=n-1}^{1} (n-j) = \sum_{l=1}^{n-1} l \right.$$

Substitution $l = n-j$
Für $j = n-1 \Rightarrow l = n-n+1 = 1$
$j = 1 = l = n-1$

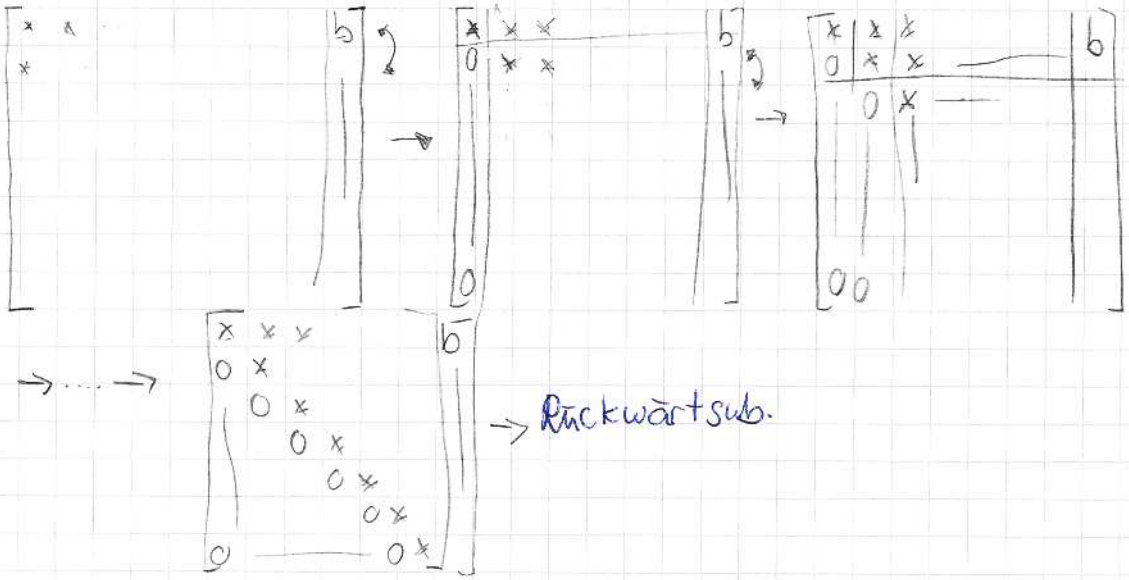
$$= n + \sum_{l=1}^{n-1} l = n + \frac{n(n-1)}{2} = O(n^2)$$

2.1.2 Gauß-Elimination

<u>Ziel:</u> Transformation auf obere Δ-Gestalt mit Elementarumformungen

1) Vertauschen von Zeilen

2) Subtraktion des Vielfachen einer Zeile (k-te Zeile) von einer anderen Zeile (i-te Zeile)



Für $k = 1, \dots, n-1$:

1) Vertausche Zeile 1 mit Zeile i. [Pivotisierung]

$$a_{k,j} \leftarrow \arg\max \{ |a_{i,k}| : i \geq k \}$$

Nehme betragsmäßig höchsten Eintrag in der Spalte und tausche mit darunterliegender Zeile

$j = 1, \dots, n$
(oder $j = k, \dots, n$)

$\rightarrow$ Vertausche

2) Für $i = k+1, \dots, n$
a) $l_{i,k} \leftarrow \dfrac{a_{i,k}}{a_{k,k}}$

Für $j = k+1, \dots, n$ (oder $j = 1, \dots, n$) $\quad a_{i,j} \leftarrow a_{i,j} - l_{i,k}\, a_{k,j}$

Beispel

$$\begin{array}{cc} A & b \end{array}$$

Pivot

$$\left[\begin{array}{ccc|c} 0 & 1 & -3 & 3 \\ 1 & 1 & 3 & -4 \\ 1 & -1 & 3 & 5 \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} 1 & 1 & 3 & -4 \\ 0 & 1 & -3 & 3 \\ 1 & -1 & 3 & 5 \end{array}\right] \xrightarrow[\substack{III-I \\ l_{31}=1}]{} \left[\begin{array}{ccc|c} 1 & 1 & 3 & -4 \\ 0 & 1 & -3 & 3 \\ 0 & -2 & 0 & 9 \end{array}\right]$$

ohne Pivot $\rightarrow$

$$\left[\begin{array}{ccc|c} 1 & 1 & 3 & -4 \\ 0 & 1 & -3 & 3 \\ 0 & -2 & 0 & 9 \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} 1 & 1 & 3 & -4 \\ 0 & 1 & -3 & 3 \\ 0 & 0 & -6 & 15 \end{array}\right]$$

$III-(-2)\cdot II$

$l_{32}=-2$

Δ-Matrix $\Rightarrow$ Backward-Sub

$-6x_3 = 15 \Rightarrow x_3 = -\frac{5}{2}$

$x_2 - 3\cdot(-\frac{5}{2}) = 3$

$\Rightarrow x_2 = -\frac{9}{2}$

$x_1 - \frac{9}{2} + (3)\cdot(-\frac{5}{2}) = -4$

$\Rightarrow x_1 = 8$

$$\rightarrow R = \begin{pmatrix} 1 & 1 & 3 \\ 0 & 1 & -3 \\ 0 & 0 & -6 \end{pmatrix} \cdot L = \begin{pmatrix} 1 & 0 & 0 \\ & 1 & 0 \\ l_{ij} & & 1 \end{pmatrix} \rightsquigarrow L = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & -2 & 0 \end{pmatrix}$$

$$L\cdot R = \begin{pmatrix} 1 & 1 & 3 \\ 0 & 1 & -3 \\ 1 & -1 & 3 \end{pmatrix} \cdot L\cdot R = \underbrace{\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{P} A \quad PLR = A$$

2.1.3 LR-Zerlegung

Erweiterte Koeffizientenmatrix

$(A, b) \in \mathbb{R}^{n \times (n+1)}$

Permutationsmatrizen

$$P = P_{k,r} = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & 0 & & 1 & \\ & & & & 1 & & \\ & & & 1 & & 0 & \\ & & & & & & 1 \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{pmatrix}$$

$\leftarrow$ k-te Zeile

$\leftarrow$ r-te Zeile

r-te Spalte

$\hat{A} = PA$ Multiplikation ~~mit~~ von Links mit der Matrix P vertauscht die Zeilen k mit r.

Proposition 1) $P = P_{k,k}$ $(k \neq r)$ ist regulär

2) $P^2 = I$, d.h. $P^{-1} = P$, 3) $P^T = P$

Satz 2.1.3 (L-R-Zerlegung)

a) Sei $A \in \mathbb{R}^{n\times n}$, $\det A \neq 0$, dann $\exists P = P_{k_1,k_1} \cdots P_{k_d,k_d}$ $(d \leq n)$ Permutation und $L = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ l_{i,j} & & 1 \end{pmatrix}$ und R obere Δ-Matrix mit

(konstruktiv mittel Gauß-Elimination)

$$P A = L \cdot R$$

$$l^{(k)} = (0 \cdots \underset{\substack{\uparrow \\ k\text{-te}}}{0}, l_{k+1,k} \cdots l_{n,k})^T \in \mathbb{R}^n, \quad e_k = (0, \ldots, \underset{\substack{\uparrow \\ k\text{-te Zeile}}}{1}, 0 \ldots 0)^T \in \mathbb{R}^n$$

kanonischer Einheitsvektor

$$G_k := \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ 0 & & 1 & & \\ & & -l_{k+1,k} & \ddots & \\ & & \vdots & & 1 \\ & & -l_{n,k} & & \end{pmatrix} = I - \underset{\substack{\uparrow \\ \text{Spaltenvektor}}}{l^{(k)}} e_k^T$$

$\Rightarrow \det G_k = 1$, d.h. G_k ist regulär und es gilt:

$$G_k^{-1} = I + l^{(k)} e_k^T = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & l_{k+1,k} & \ddots \\ & & \vdots & \\ & & l_{n,k} & & 1 \end{pmatrix}$$

$\langle e_k, l^k \rangle = 0$

$$G_k^{-1} G_k = (I + l^{(k)} e_k^T)(I - l^k e_k^T) = I + \cancel{l^k e_k^T} - \cancel{l^k e_k^T} - \cancel{l^k e_k^T l^k e_k^T} = I$$

Beobachtung $(A, b)^{(k+1)} = G_k \cdot P_{k,r} (A, b)^k \quad r \geq k$
nach k-tem Eliminationsschritt

b) $\tilde{A} = PA = LR$, Diese Zerlegung ist eindeutig!

$\mathcal{L} = \{ L \mid \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ l_{i,j} & & 1 \end{pmatrix} \} \subset \mathbb{R}^{n\times n}, \ l_{i,j} \in \mathbb{R}\}$, $\mathcal{R} \in \{ \begin{pmatrix} r_{11} & & r_{1,j} \\ & \ddots & \\ 0 & & r_{n,n} \end{pmatrix}, r_{ij} \in \mathbb{R} \}$

Beweis: $L_1, L_2 \in \mathcal{L} \Rightarrow L_1 \cdot L_2 \in \mathcal{L}$, $R_1, R_2 \in \mathcal{R} \Rightarrow R_1 \cdot R_2 \in \mathcal{R}$

a) $l^{(k)} := (0, \ldots, 0, \underset{\substack{\uparrow \\ k\text{-te Spalte}}}{l_{k+1,k}}, \ldots, l_{n,k})^T$, $z_k = (0, \ldots, 0, z_{k+1}, \ldots, z_n)^T \in \mathbb{R}^n$

$$G_k = \begin{pmatrix} 1 & & & & 0 \\ 0 & \ddots & & & \\ & & 1 & & \\ & & -l_{k+1,k} & \ddots & \\ & & \vdots & & 1 \\ & & -l_{n,k} & 0 & & 1 \end{pmatrix} = I - \underset{\square}{l^{(k)}} \cdot \underset{\square}{e_k^T} \in \mathcal{L}$$

$\Rightarrow G_k^{-1} = I + l^{(k)} e_k^T$

$P_k = P_{kr}, r > k$ mit k-tem Eliminationsschritt

$G_k \underset{\substack{\uparrow \\ \text{Pivotschritt}}}{P_k} A^{(k)} = A^{(k+1)}$

Es gilt $P_k^T \cdot e_j = e_j$, $k > j$

$$\underbrace{P_k G_j P_k}_{\tilde{G}_j} = P_k (I - z_j e_k^T) P_k = (P_k P_k) - (P_k z_j)(P_k e_j)^T = I - P_k z_j e_j^T$$

nach dem k-ten Schritt

$$\mathcal{R} \ni R := A^n = G_{n-1} P_{n-1} A^{(n-1)} = G_{n-1} P_{n-1} G_{n-2} P_{n-2} \cdots G_1 P_1 A \quad (*)$$

$$\left[(A,b)^n = \text{"} \qquad \text{"} \qquad (A,b) \right]$$

$$(*) = G_{n-1} \underbrace{P_{n-1} G_{n-2} P_{n-1}}_{\text{Vergleich i)}} \underbrace{P_{n-1} P_{n-1}}_{I} P_{n-2}$$

$$= G_{n-1} \tilde{G}_{n-2} P_{n-1} \underbrace{P_{n-2} G_{n-3} P_{n-2}}_{\text{i)}} P_{n-2} P_{n-3}$$

$$= G_{n-1} \tilde{G}_{n-2} \tilde{G}_{n-3} P_{n-1} P_{n-2} P_{n-3} G_{n-4} P_{n-3} P_{n-3} P_{n-4} \cdots$$

$$= G_{n-1} \tilde{G}_{n-2} \tilde{G}_{n-3} \cdots P_{n-1} \cdots P_1 \cdot A \qquad \text{mit } \tilde{G}_j = I - P_{n+1} \cdots P_{j+1} z_j e_j^T = I - z_j e_j^T$$

und $\tilde{G} \in \mathcal{L}$

$$\Rightarrow PA = \tilde{G}^{-1} \cdot R = \underbrace{\tilde{G}_{n-2}^{-1} \cdots \tilde{G}_{n-1}^{-1}}_{\in \mathcal{L}} R = LR$$

b) <u>Eindeutigkeit</u>: Seien $L_1, L_2 \in \mathcal{L}$, $R_1, R_2 \in \mathcal{R}$

mit $PA = L_1 R_1 = L_2 R_2 \quad | L_2^{-1} (\ )$

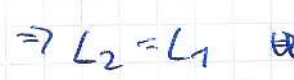
$$L_2^{-1} L_1 R_1 = \underbrace{L_2^{-1} L_2}_{I} R_2$$

$$L_2^{-1} L_1 R_1 = I R_2 \quad | (\ ) R_1^{-1}$$

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

$L \in \mathcal{L} \Rightarrow L^{-1} \in \mathcal{L}$

$$\mathcal{L} \ni L_2^{-1} L_1 = R_2 R_1^{-1} \in \mathcal{R}$$



$$\Rightarrow I = I$$

$$\Rightarrow L_2 = L_1 \text{ und } R_1 = R_2$$

Bem Sei $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \left(= P \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)$. Dann ex. <u>keine</u> LR-Zerlegung (ohne Permutation) (ohne Pivotisierung)

Satz 2.1.4 Falls alle Hauptunterdeterminanten

$0 \neq \det A_k = \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \\ a_{k1} & & a_{kk} \end{pmatrix} \in \mathbb{R}^{k \times k}$ $\forall k = 1, \dots, n$, dann existieren

$L \in \mathcal{L}, R \in \mathcal{R}$ mit $A = LR$

Beweis Übung ** (unnötig)

2.1.5 Komplexität des Gauß-Alg.

Satz 2.1.5 Der Eliminationsschritt benötigt für $n = 2$

$\frac{n^3}{3} + \frac{n^2}{2} - \frac{5n}{6}$? Multiplikation / Divisionen D.h. $T(n) = O(n^3)$

Beweis: Im k-ten Schritt benötigt man $(n-k)$ Divisionen

und $(n-k+1)(n-k)$ Multiplikation

$\Rightarrow$ Anz. der (*) (Schritt?) $\sum_{k=1}^{n-1} (n-k)(n-k+2) = \sum_{l=1}^{n-1} l(l+2)$ (*)

$l = n-k$, $k = 1$, $l = n-1$, $k = n-1$, $l = 1$

$(*) = \sum_{l=1}^{n-1} l^2 + 2\sum_{l=1}^{n} l = \frac{n(n-1)(2n-1)}{6} + \frac{2(n-1)n}{2} = \frac{n^3}{3} + \frac{n^2}{2} - \frac{5n}{6} = O(n^3)$

Aufwand für $A \cdot B \rightsquigarrow A, B \in \mathbb{R}^{n \cdot n}$ $\rightarrow O(n^3)$

aber Strassen (1969) – $O(n^{\log_2(7)}) = O(n^{2,7\cdots})$

2.1.5 Vorteile der LR-Zerlegung

1. Berechnung von Determinaten

$PA = LR$, $\det(PA) = \det(LR)$

$\det(PA) = \det(P)\det(A) = \det(L)\det(R) = 1 \cdot \prod_{i=1}^{n} r_{ii}$

$\det(P) = (-1)^{(\#\text{Vertauschungen})} = \sigma \in \{\pm 1\}$

$\Rightarrow \det(A) = \sigma (r_{11} \cdots r_{n,n})$ $\rightarrow$ Aufwand $T(n) = O(n^3)$ (LR Zerlegung)

im Gegensatz zu Entwicklungssatz (Laplace) $T(n) = O(n!)$

2. Mehrere Rechte Seiten $Ax_i = b_1 (\dots b_2 \dots b_p)$

$L\underbrace{Rx}_{=z} = PAx = Pb$ Alg: 1) $c = Pb$ Permutation

2) Löse $Lz = c$; Vorwärtssubstitution $O(n^2)$

3) $Rx = z$ Rückwärtssubstitution $O(n^2)$

3) Berechnung von $A^{-1} = (x_1, \dots, x_n)$

$Ax_i = e_i$, $i = 1, \dots, n$ $(n \cdot O(n^2) = O(n^3))$

4) Nachiteration

Sei $x^* \in \mathbb{R}^n$ eine Näherungslösung von $Ax=b$

Residuum $r = b - Ax^* = A(\underbrace{x - x^*}_{\text{nicht bekannt}}) \neq 0$, setze $x = x^* + \delta \Rightarrow \delta = x - x^*$

mit $\boxed{A\delta = r} = A(x - x^*)$

d.h. Löse $A\delta = r$

2.1.6 Bemerkung zu Pivot-Strategien

- Üblich Spaltenpivotisierung: maximales $|a_{j,k}|$ $j=k,\dots,n$ im k-ten Schritt
- Vollständige oder Restmatrix-Pivotisierung:

  Suche maximales $|a_{j,k}|$ für $j=k,\dots,n$ für $k=k,\dots,n$

  ↳ durchlaufen nicht nur die restliche Spalte, sondern gesamte restliche Matrix

  ↳ Zeilen & Spaltenvertauschung Optimal, aber zu teuer!
- Äquilibierung:

  <u>Def</u>: $A=(a_{i,j})\in\mathbb{R}^{n\times n}$ heißt zeilenweise äquilibriert, wenn
  $$\sum_{j=1}^{n}|a_{i,j}|=1\ (\approx 1)\ \forall i=1,\dots,n$$

  Seien $d_i=\sum_{j=1}^{n}|a_{i,j}|$, $i=1,\dots,n$ ← Summe der Einträge der Zeile i d_i

  und sei $D=\operatorname{diag}(d_i)=\begin{pmatrix} d_1 & & 0\\ & \ddots & \\ 0 & & d_n\end{pmatrix}$

  Aus $Ax=b \Leftrightarrow \bar{A}x=\bar{b}=D^{-1}Ax=D^{-1}b$

  d.h. $\bar{b}=D^{-1}b$, $\bar{A}=D^{-1}A=\operatorname{diag}(d_i^{-1})A$ (leicht zu berechnen)

  $\Rightarrow \bar{A}$ ist äquibriert

  → Strategie: Äquilibriere A und spaltenweise Pivotisierung

  <u>Bsp</u>: $Ax=\begin{pmatrix}0{,}31\cdot 10^{-3} & 1\\ 1 & 1\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}-3\\ -7\end{pmatrix}$

  <u>ohne Pivotisierung</u>: $l_{2,1}=\frac{1}{0{,}31\cdot 10^{3}}$

  erster Eliminationsschritt $\Rightarrow \left[\begin{array}{cc|c}0{,}31\cdot 10^{-3} & 1 & -3\\ 1 & -3225 & 9670\end{array}\right] \Rightarrow x_1^*\approx -6{,}452$ ⇝ relativ schlecht, $x_2^*\approx -2{,}998$ ⇝ relativ gut

  Exakt: $x_1=-4{,}00124$, $x_2=-2{,}998759$ (ohne Runden)

  Ursache: Auslöschung im Rückwärtseinsetzen: $x_2\approx 3$

  und $x_1^*=\underbrace{\frac{1}{0{,}00031}}_{\Rightarrow 1}\underbrace{(-3-x_2^*)}_{\approx 0}$

Mit Spaltenpivotisierung

nach 1. Eliminationsschritt

$$\left[\begin{array}{cc|c} 1 & 1 & -7 \\ 0 & 0{,}9997 & -2{,}998 \end{array}\right] \qquad \left.\begin{array}{l} \bar{x}_1 \approx -4{,}001 \\ \bar{x}_2 \approx -2{,}999 \end{array}\right\} \text{gut}$$

mit Äquilibrierung

Multipliziere 1. Zeile mit 10^4

$\rightsquigarrow$ ähnliches Resultat wie mit Pivotisierung

2.1.7 Spezielle Matrizen - Bandmatrizen

Def $A = (a_{i,j}) \in \mathbb{R}^{n \times n}$ heißt (m,k)-Bandmatrix, falls

$a_{i,j} = 0$ für $i - j > m$ oder $j - i > k$

also $\begin{pmatrix} \ddots & & * \\ & \ddots & \\ 0 & & \ddots \end{pmatrix}$ oder $\begin{pmatrix} \ddots & & 0 \\ & \ddots & \\ * & & \ddots \end{pmatrix}$

Tridiagonalmatrix: $(1,1)$, $m = k = 1$

Hessenbergmatrix: $(1, n-1)$, $m = 1$, $k = n-1$

Tridiagonalmatrix: $(b,a,c) = \begin{pmatrix} a_1 & c_1 & & 0 \\ b_2 & a_2 & \ddots & \\ & \ddots & \ddots & c_{n-1} \\ 0 & & b_n & a_n \end{pmatrix}$

$$\begin{pmatrix} a_1 & c_1 & & 0 \\ b_2 & \ddots & \ddots & \\ & \ddots & \ddots & c_{n-1} \\ 0 & & b_n & a_n \end{pmatrix} = \begin{pmatrix} \alpha_1 & & & 0 \\ b_2 & \ddots & & \\ & \ddots & \ddots & \\ 0 & & b_n & \alpha_n \end{pmatrix} \begin{pmatrix} 1 & \gamma_1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & \gamma_{n-1} \\ 0 & & & 1 \end{pmatrix}$$

Koeffizientenabgleich

$\Rightarrow a_1 = \alpha_1$, $c_1 = \alpha_1 \gamma_1 \Rightarrow \gamma_1 = \frac{c_1}{\alpha_1} = \frac{c_1}{a_1}$

$\Rightarrow a_i = b_i \gamma_{i-1} + \alpha_i \cdot 1 \Rightarrow \alpha_i = a_i - b_i \gamma_{i-1}$

und $c_i = \alpha_i \gamma_i \Rightarrow \gamma_i = \frac{c_i}{\alpha_i}$ Rekursiv

$T = \tilde{L} \cdot \tilde{R}$

T-Tridiagonalmatrix heißt irreduzibel diagonaldominant

falls 1) $0 < |c_i| < |a_1|$;

2) $|a_i| \geq |b_i| + |c_i|$ (bei „strikt" diagonaldominant)

3) $|a_1| > |b_n| > 0$ (hinreichend)

Satz 2.1.9: Sei $A \in \mathbb{R}^{n\times n}$ ($\mathbb{C}^{n\times n}$) irreduzibel diagonaldominant

Dann ist $\det A \neq 0$ und es gelten

1) $|\gamma_i| < 1 \quad \forall i = 1, \dots, n-1$

2) $|a_i| + |b_i| > |\alpha_i| > |a_i| - |b_i| > 0$

Beweis 1) Induktionsanfang: $i = 1$

$$|\gamma_1| = \left|\frac{c_1}{a_1}\right| < 1$$

IS $j-1 \to j$:

$$|\gamma_j| = \left|\frac{c_j}{\alpha_j}\right| = \frac{|c_j|}{|a_j - b_j \gamma_{j-1}|} \leq \frac{|c_j|}{\big|\,|a_j| - \underbrace{|\gamma_{j-1}|}_{<1}|b_j|\,\big|} \leq \frac{|c_j|}{|a_j| - |b_j|} \leq \frac{|c_j|}{|c_j|} = 1$$

2) $|a_j| + |b_j| > |a_j - \gamma_{j-1} b_j| = |\alpha_j| > |a_j| - |b_j| \overset{IV}{\geq} |c_j| > 0$

$\Rightarrow \det T = \det \tilde{L} \cdot \det \tilde{R} = \alpha_1 \alpha_2 \dots \alpha_n 1 > 0$, da $d_i > 0 \ \forall i = 1, \dots, n$ ∎

2.2. Symmetrische Matrizen & Cholesky-Zerlegung

$A \in \mathbb{R}^{n\times n}$ heißt symmetrisch, wenn $A = A^T$

A symmetrisch heißt positiv definit, falls

$\forall x \in \mathbb{R}^n$ mit $x \neq 0$ $\quad x^T A x = \langle x, Ax \rangle > 0$

alternativ: $\exists \gamma > 0 \ \forall x \in \mathbb{R}^n \quad x^T A x \geq \gamma \|x\|^2 = \gamma x^T x$

kurz SPD (symm. pos. definit)

Theorem: A SPD $\Leftrightarrow$ alle Eigenwerte $\lambda_1, \dots, \lambda_n$ von A sind > 0, d.h. $\lambda_i > 0 \ \forall i = 1, \dots, n$

$\Leftrightarrow$ alle Hauptunterdeterminanten von A, also

$\det A_k > 0 \ \forall k = 1, \dots, n$, sind positiv.

Satz 2.2.3 A ist SPD $\Leftrightarrow$ es ex. genau eine linke untere Δ-Matrix $\tilde{L} = (\tilde{l}_{ij})$ ($\tilde{l}_{ij} = 0, j > i$)

mit $A = \tilde{L}\tilde{L}^T$ ($\tilde{L}^T$ ist obere Δ-Matrix)

Beweis „$\Rightarrow$" da $\det A_k > 0$ ex. L, R mit $A = L \cdot R$ und $\det A_k = \det L_k \det R_k = \prod_{i=1}^{k} r_{ii} > 0$

$\Rightarrow r_{k,k} > 0 \ \forall k = 1, \dots, n$ für $\tilde{l}_{i,i} = \sqrt{r_{ii}} \Rightarrow \tilde{r}_{ii} = \frac{r_{ii}}{\sqrt{r_{ii}}}$

$l_{i,j} = \tilde{l}_{i,j}, \ i > j, \quad r_{i,j} = \tilde{r}_{i,j}, \ i < j$

$\Rightarrow A^T = A = \tilde{L} \cdot \tilde{R} = A^T = (\tilde{L}\tilde{R})^T = \tilde{R}^T \tilde{L}^T$

Aus der Eindeutigkeit der LR-Zerlegung $\Rightarrow \tilde{L} = \tilde{R}^T = A = \tilde{L}\tilde{L}^T$

mit der Eigenschaft $\det L > 0$

„$\Leftarrow$" sei $\tilde{L}$ regulär $\Rightarrow \forall x \neq 0 \; x^T A x = x^T \tilde{L} \tilde{L}^T x$

$$= (\tilde{L}^T x)^T (\tilde{L}^T x)$$

$$= \|\tilde{L}^T x\|^2 \geq 0 \; \forall x \neq 0$$

2.2.2. Die Crout-Cholesky-Zerlegung

Gesucht: x mit $Ax=b$

symm. $A=A^T \in \mathbb{R}^n$ | $A^* = A^H = A$ (hermitesch))

A positiv definit (SPD) $\Rightarrow \exists$ $A = L_c L_c^T = (LR)$ (Cholesky-Zerlegung) $(L_{ii}^c)^2 = d_i$ die Diag von dem Cholesky L_c

Sei $A = L \cdot R = L \cdot \underbrace{D \cdot L^T}_{=\tilde{L}^T}$ die Crout-Cholesky-Zerlegung

$l_{ii}=1 \; \forall i=1,..,n$ $\quad D=\mathrm{diag}(d_i)$ Diagonalmatrix

d.h. $(i \geq j)$ $\quad a_{i,j} = \sum_{k=1}^{n} l_{i,k} \tilde{l}_{j,k} = \sum_{k=1}^{n \geq j} l_{i,k} d_k l_{j,k} = \sum_{k=1}^{j-1} l_{i,k} d_k l_{j,k} + l_{i,j} d_j \underset{=1}{l_{j,j}}$

$\tilde{l}_{i,j} = d_j l_{i,j} \Rightarrow l_{i,j} = \frac{\tilde{l}_{i,j}}{d_j}$

$j=1$, $i=1$, $a_{11} = l_{11} d_1 l_{11} = d_1$

$j=1$, $i=2$ $\quad a_{2,1} = l_{2,1} d_1 \Rightarrow \tilde{l}_{2,1} = a_{2,1}$ $l_{2,1} = \frac{a_{2,1}}{d_1}$

$j=1$, $i \geq 2$ $\quad a_{i,1} = l_{i,1} d_1 \Rightarrow l_{i,1} = \frac{a_{i,1}}{d_1}$, $i=2,..,n$

$j=2$:
$i=2$ $\quad a_{2,2} = d_2 + l_{2,1} d_1 l_{2,1}$

$\Rightarrow d_2 = a_{22} - l_{2,1} \underbrace{d_1 l_{2,1}}_{\tilde{l}_{2,1}}$

$j=2$, $i \geq 2$

$(i=j)$ Für $j=1,..,n$: $d_j = a_{j,j} - \sum_{k=1}^{j-1} l_{j,k} \tilde{l}_{j,k} = a_{jj} - \sum_{j=1}^{j-1} l_{j,k}^2 d_k$

Für $i=j+1,..,n$:

$$\left.\begin{aligned} \tilde{l}_{i,j} &= a_{i,j} - \sum_{k=1}^{j-1} l_{i,k} \cdot \tilde{l}_{j,k} \\ l_{i,j} &= \frac{\tilde{l}_{i,j}}{d_j} \end{aligned}\right\} \text{Algo(77)}$$

<u>Beispiel</u>: $A = \begin{pmatrix} 2 & 6 & -2 \\ 6 & 21 & 0 \\ -2 & 0 & 16 \end{pmatrix} \Rightarrow d_1 = a_{11} = 2$

$\Rightarrow j=1, i=2$ $\quad \tilde{l}_{2,1} = a_{2,1} = 6$, $l_{2,1} = \frac{\tilde{l}_{2,1}}{d_1} = \frac{6}{2} = 3$

$j=1, i=3$ $\quad \tilde{l}_{3,1} = a_{3,1} = -2$, $l_{3,1} = \frac{\tilde{l}_{3,1}}{d_1} = -\frac{2}{2} = -1$

$j=2$, $d_2 = a_{2,2} - l_{2,1} \tilde{l}_{2,1} = 21 - 18 = 3 = d_2$

$j=2, i=3$, $\tilde{l}_{3,2} = a_{3,2} - l_{3,1} \tilde{l}_{2,1} = 0 - (-1) \cdot 6 = 6$

$l_{3,2} = \frac{\tilde{l}_{3,2}}{d_2} = \frac{6}{3} = 2$

$j=3$: $d_3 = a_{3,3} - l_{3,1} \tilde{l}_{3,1} - l_{3,2} \tilde{l}_{3,2} = 16 - (-1)(-2) - 2 \cdot 6 = 2$

$$L = \begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ -1 & 2 & 1 \end{pmatrix} \quad D = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

$$A = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ -1 & 2 & 1 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix}}_{D} \underbrace{\begin{pmatrix} 1 & 3 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}}_{L^T} \quad \Rightarrow \underset{i=1,\dots,n}{d_i > 0} \quad A = SPD$$

Die klassische Cholesky-Zerlegung

$A = L_c L_c^T$ ergibt sich aus $A = LDL^T$ mit $r_i = \sqrt{d_i}$ $(d_i > 0)$
$= LD^{1/2}(LD^{1/2})^T$

d.h. $l_{i,j}^{chol} = \underset{\substack{\uparrow \\ \text{berechnet} \\ \text{in Crout-Cholesky}}}{l_{i,j}} \sqrt{d_i}$

Vorteile von Crout-Cholesky gegenüber Cholesky:

1) Cholesky erfordert SPD
Symmetrisch positiv definit

2) Cholesky erfordert (benötigt) Wurzelfunktion

Arbeitsaufwand zur Cholesky Zerlegung:

Anzahl der arithmetischen Operationen $T(n) = \frac{1}{6}n^3 + \frac{1}{2}n^2 - \frac{2}{3}n$

$T(n)|_{Cholesky} \sim \frac{1}{2} T(n)|_{Gauß}$ ohne Pivotisierung

---

Lösen von $Ax = b$ mit $A = LDL^T$ $\quad \underbrace{LDL^T}_{z} x = b$

Löse $Lz = b$, Vorwärts-Substitution $\quad Dy = z$

$y = D^{-1} z = \mathrm{diag}(d_i^{-1}) z$

$x = L^T y$, Rück-Sub

Bemerkung: Es wurde keine Pivotisierung durchgeführt
(Geringere Stabilität als pivot. Gauß)

Orthogonale Projektion (oft =) beste Approximation, QR-Zerlegung Num I, VL

Def (Skalarprodukt) V - reeller Vektorraum,

$\langle \cdot , \cdot \rangle : V \times V \to \mathbb{R}$ heißt Skalarprodukt, falls:

$\langle x, y \rangle = \langle y, x \rangle \quad \forall x, y \in V$

$\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle \quad \forall x, y, z \in V, \forall \alpha, \beta \in \mathbb{R}$

$\langle x, x \rangle \geq 0, \ \langle x, x \rangle = 0 \Leftrightarrow x = 0 \quad \forall x \in V$

Satz (Cauchy-Schwarz-Ungl.) $|\langle x, y \rangle| \leq \|x\| \cdot \|y\|$

Def (Beste Approximation): Sei V normierter Vektorraum über $\mathbb{R}$ und M ein Untervektorraum. y^* heißt Element bester Approximation an $x \in V$, falls $y^* \in M$ und $\forall y \in M$:

$$\underbrace{\|x - y\|}_{\substack{\text{Abstand} \\ x \text{ zu } y}} \geq \underbrace{\|x - y^*\|}_{\substack{\text{Abstand} \\ x \text{ zu } y^*}}$$

Def (orthogonale Projektion) Sei V ein reeller VR mit Skalarprodukt, M ein UVR, y^* heißt orthogonale Projektion von x auf M, falls $y^* \in M$ und $\forall y \in M$: $\langle y, \underbrace{x - y^*}_{\text{Verbindungslinie zw. } x \text{ und seiner orth. Proj.}} \rangle = 0$

Satz Orthogonale Projektion ist eindeutig

Beweis Seien y^* und $\hat{y}$ beide orth. Proj. von x auf M

Es gilt dann $\langle y^* - x, y \rangle = 0 \quad \forall y \in M$

$\langle \hat{y} - x, y \rangle = 0 \quad \forall y \in M$

Daraus folgt: $\langle y^* - \hat{y}, y \rangle = \langle y^* - x, y \rangle - \langle \hat{y} - x, y \rangle = 0 - 0 = 0 \quad \forall y \in M$

Wähle nun $y = y^* - \hat{y} \in M$, dann folgt

$\langle y^* - \hat{y}, y^* - \hat{y} \rangle = 0$ Nach Definitheit des SP folgt $y^* - \hat{y} = 0$, also $y^* - \hat{y} = 0$

Satz Sei $y^* \in M$ ein El. bester Approximation an $x \in V$. Dann ist y^* auch die orthogonale Projektion von x auf M.
(bezgl. einer vom SP induzierten Norm)

Beweis Sei $y^* \in M$ ein beliebiges Element bester Approximation.

Dann liegen $y^* + \alpha v$ auch in M, falls $v \in M$ und die Funktion

$F: \mathbb{R} \to \mathbb{R}: \alpha \mapsto \|x - \underbrace{(y^* + \alpha v)}_{\in M}\|^2$ muss bei $\alpha = 0$ ein Minimum haben

$F(\alpha) = \|(x-y^*) - \alpha v\|^2 = \|x-y^*\| + \alpha^2 \|v\|^2 - 2\alpha \langle x-y^*, v\rangle$

$\Rightarrow$ Bei $\alpha = 0$ muss F' verschwinden

$\Rightarrow 0 = F'(0) = -2\langle x-y^*, v\rangle$

Da v beliebig gewählt folgt die Voraussetzung der orthg. Projektion ∎

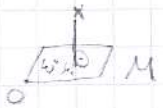
<u>Korollar</u> Das Element bester Approximation ist eindeutig

<u>Beweis</u> Übung

<u>Satz</u> Sei y^* die orthogonale Projektion von $x \in V$ auf M.

Dann ist y^* auch beste Approxi.

<u>Beweis</u> Sei w irgendein Element aus M



$\|x-w\|^2 = \|x-y^* + (y^*-w)\|^2$

$= \langle (x-y^*)+(y^*-w), (x-y^*)+(y^*-w)\rangle$

Linearität, Symm. des SP

$= \langle x-y^*, x-y^*\rangle + \langle y^*-w, y^*-w\rangle + 2\langle x-y^*, y^*-w\rangle$

$= \|x-y^*\|^2 + \underbrace{\|y^*-w\|^2}_{\geq 0} + \underbrace{2\langle x-y^*, \underbrace{y^*-w}_{\in M}\rangle}_{=0,\ \text{weil } (y^*-w)\perp(x-y^*) \text{ bzw. } y^*-w \perp \text{Projekt.}} \geq \|x-y^*\|$

<u>Satz</u> Ist $\dim(M) < \infty$ und $a_1, \dots, a_m$ die lin unab. Spalten von A und $M = \operatorname{span}\{a_1, \dots, a_m\}$. Dann ist $y^* = A\alpha$ genau dann das Element bester Approxi an x wenn

$A(A^TA)^{-1}A^Tx$ ist orth. Proj von x auf $M = \operatorname{span}\{a_1, \dots, a_m\}$ wenn $A = (a_1, \dots, a_m)$

$$A^TA\alpha = A^Tx \iff \|\underbrace{A\alpha}_{=y^*} - x\| \text{ minimal}$$

(Normalengleichung)

<u>Beweis</u> In einer LE wurde gezeigt, dass $y^* = A(A^TA)^{-1}A^Tx$ ∎

<u>Bemerkung</u>: A^TA sym pos definit $\Rightarrow A^TA$ invertierbar

$\Rightarrow \alpha$ eindeutig bestimmt

<u>Bemerkung</u> A^TA zu berechnen numerisch nicht immer sinnvoll!

Beispiel $A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ \varepsilon & & & & \\ & \varepsilon & & & \\ & & \varepsilon & & \\ & & & \varepsilon & \\ & & & & \varepsilon \end{pmatrix}$ hat Rang 5

$A^TA = \begin{pmatrix} 1+\varepsilon^2 & 1 & \cdots & 1 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & 1+\varepsilon^2 \end{pmatrix}$

wenn $\varepsilon < \sqrt{\varepsilon^*}$ ← max rel. Fehler da Rd

folgt $fl(A^TA) = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}$ hat Rang 1

Rd

Was wenn A nicht vollen Rang hat?

Wollen α finden, so dass $\|A\alpha - x\|$ minimal

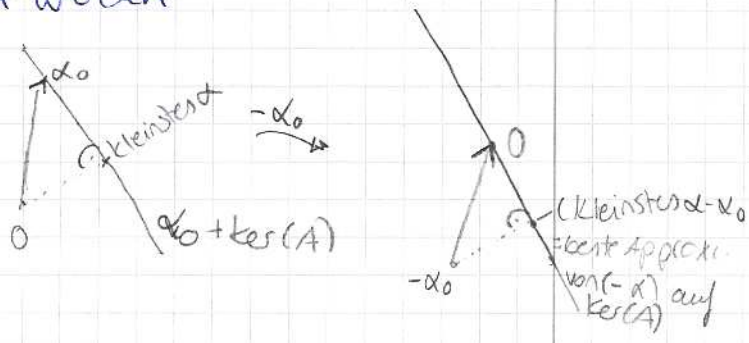
(gesucht ist Element bester Approximation $y^* = A\alpha$ im von den Spalten von A aufgespannten Untervektorraum)

$A\alpha$ eindeutig, aber α nicht eindeutig

Sei α_0 so dass $A\alpha_0 = y^*$ E.b.A $\Rightarrow \alpha_0 + v$ auch gut, wenn $v \in \ker(A)$ $A(\alpha_0 + v) = A\alpha_0 = y^*$

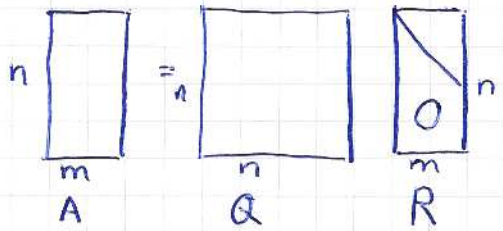
α kann aus affinem UVR $\alpha_0 + \ker(A)$ gewählt werden.

Das kleinste $\alpha \in \alpha_0 + \ker(A)$ kann wieder durch Beste Approximation bestimmt werden



QR-Zerlegung

Def: Die QR-Zerlegung einer $n \times m$ Matrix A ist eine Zerlegung von A in ein Produkt $A = Q \cdot R$, wobei Q eine orthogonale $n \times n$-Matrix und R eine $n \times m$ obere Δ-Matrix ist

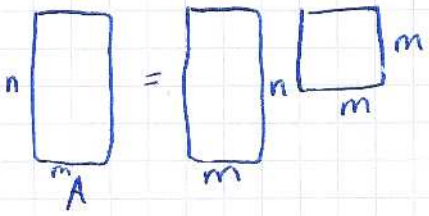


- Eine Matrix Q heißt orthogonal, wenn ihre Spalten jeweils Norm/Länge Eins haben und paarweise orthogonal/senkrecht aufeinander stehen, also wenn $Q^TQ = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$. Es folgt, dass $Q^{-1} = Q^T$ und damit $Q^TQ = Q^{-1}Q = QQ^{-1} = QQ^T = (Q^T)^T Q^T$ und damit ist also auch Q^T orthogonal und deshalb können wir in der Definition einer orth. Matrix das Wort Spalten durch Zeilen ersetzen.

- Q lässt Abstände und Winkel konstant

$$\langle Qx, Qy\rangle = (Qx)^T(Qy) = x^TQ^TQy = x^Ty = \langle x,y\rangle$$

- Wir können die letzten n-m Spalten von Q weglassen und die QR Zerlegung als $A=QR$ mit $Q\in\mathbb{R}^{n\times m}$, $Q^TQ=id$ und $R\in\mathbb{R}^{m\times m}$ obereΔ-Matrix definieren.



- Das <u>Produkt orthogonaler Matrizen ist orthogonal</u>
- Jede Matrix, deren assoziierte lineare Abbildung das Skalarprodukt invariant lässt, ist orthogonal

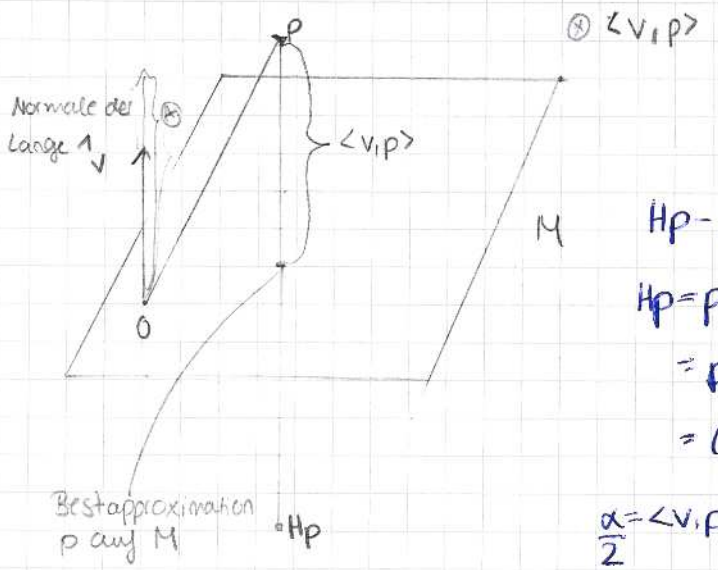
$\Rightarrow$ Drehungen und Spiegelungen sind orthogonal

<u>Spiegelungen</u> (an Hyperebenen)

<u>Def</u> Eine <u>Hyperebene</u> im $\mathbb{R}^n$ ist ein $(n-1)$-dimensionaler möglicherweise affiner Unterraum

<u>Def</u> Eine <u>Spiegelung</u> f an einer Hyperebene M ist definiert durch die Eigenschaft, dass jeder Punkt p auf einen Punkt f(p) abgebildet wird, der <u>denselben Abstand zu M</u> hat, so dass die P und f(P) <u>verbindende Gerade orthogonal auf M</u> steht.

- <u>Householder-Spiegelungstransformation</u> ist eine Spiegelung an einer Hyperebene, die durch 0 geht.
- orthogonales Komplement eines $(n-1)$-dimensionalen Unterraums ist eindimensional und wird von der sogenannten Normalen aufgespannt d.h. wir können Householder-Spiegelung eindeutig durch diese Normale beschreiben.



$Hp - p = \alpha \cdot v$

$Hp = p + \alpha v$

$= p - 2v\langle v,p\rangle$

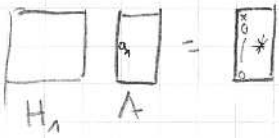
$= (I - 2vv^T)p$

$\frac{\alpha}{2} = \langle v,p\rangle$ bitte überprüfen

<u>Satz</u> Sei v eine Normale der Länge 1 auf M. Dann lässt sich die Spiegelung H an M schreiben als $Hp = (I - 2vv^T)p$

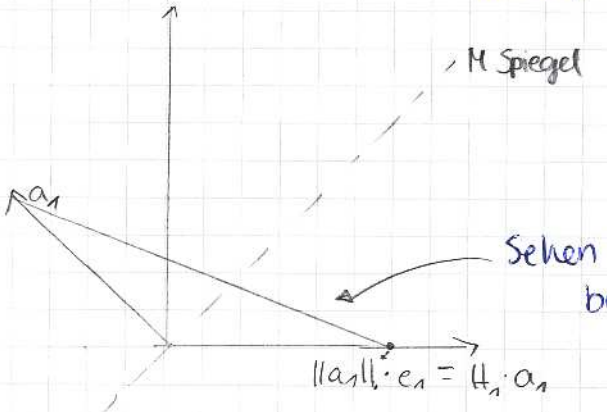
H ist sogar symmetrisch $H = H^T = H^{-1}$

<u>Ziel:</u>



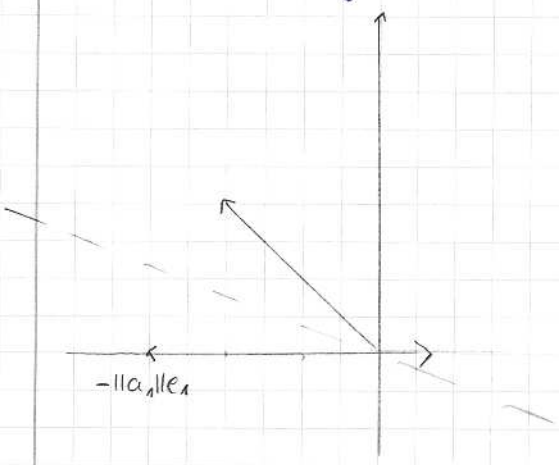
d.h. $H_1 a_1 = \begin{pmatrix} x \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} \pm\|a_1\| \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \pm\|a_1\| \cdot e_1$

weil H_1 ortho. und damit Längen erhaltend



Sehen sofort, dass $a_1 - \|a_1\| e_1$ der benötigte Normalenvektor ist.

Haben zwei Möglichkeiten, zu spiegeln. Die andere ist:



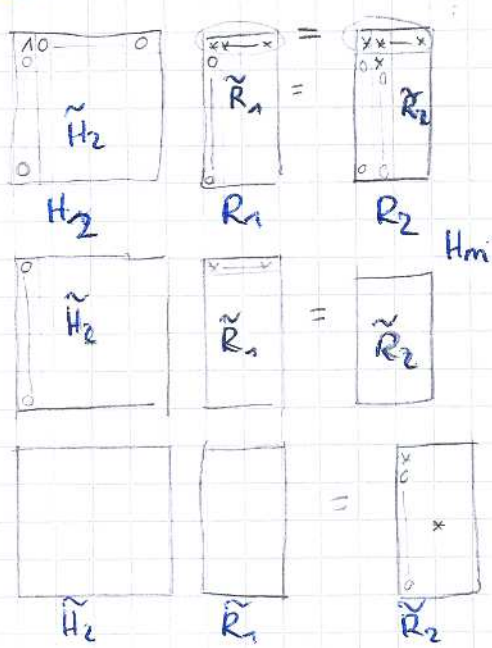
Numerisch stabiler ist aber die, wo $(Hp - p)$ größer ist. Deshalb wählen wir:

$w = a_1 + \operatorname{sgn}(a_{11})\|a_1\| e_1$

Müssen nach normieren und erhalten

$v = \frac{w}{\|w\|}$ als normierte Normale bzw. Einheitsnormale

Next:



Haben die Bestimmung von H_2 also auf die von H_1 zurückgeführt

$H_m \cdot H_2 H_1 \cdot A = R$

$\Rightarrow A = H_1 \cdots H_m R \Rightarrow Q = H_1 \cdot \ldots \cdot H_m$

Auch möglich: Drehungen verwenden: Stichwort „Givens-Rotation"

$$\begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} x & \cdots & x \\ x & \cdots & x \end{pmatrix} = \begin{pmatrix} \uparrow x & \cdots & x \\ 0\, x & \cdots & x \end{pmatrix}$$

23.4 Givens Rotation

$\mathbb{R}^2$: orthogonale Matrizen

$$G = \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \text{ mit } c^2 + s^2 = 1$$

oder $c = \cos\varphi,\ s = \sin\varphi \quad \varphi \in [0, 2\pi]$

Grundaufgabe zu gegebenem $y = \begin{pmatrix} a \\ b \end{pmatrix}$, konstruiere G und $\begin{pmatrix} r \\ 0 \end{pmatrix}$, $r \in \mathbb{R}$ geeignet

$$\begin{pmatrix} r \\ 0 \end{pmatrix} = G \begin{pmatrix} a \\ b \end{pmatrix} \Rightarrow \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} r \\ 0 \end{pmatrix} = \begin{pmatrix} \sqrt{a^2+b^2} \\ 0 \end{pmatrix} \Rightarrow c = \frac{a}{r},\ s = \frac{b}{r}$$

$$\|Gy\|^2 = \|y\|^2 = \left\|\begin{pmatrix} r \\ 0 \end{pmatrix}\right\|^2 = r^2 = \left\|\begin{pmatrix} a \\ b \end{pmatrix}\right\|^2 = a^2 + b^2 \Rightarrow r = \sqrt{a^2+b^2}$$

Probe: $$\begin{pmatrix} \frac{a}{r} & \frac{b}{r} \\ -\frac{b}{r} & \frac{a}{r} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \frac{a^2+b^2}{r} \\ -\frac{ab}{r} + \frac{ab}{r} \end{pmatrix} = \begin{pmatrix} r \\ 0 \end{pmatrix}$$

$$G_{j,k} = \begin{array}{c} \\ j \\ k \end{array} \begin{pmatrix} I & & & & 0 \\ & c & & s & \\ & & I & & \\ & -s & & c & \\ 0 & & & & I \end{pmatrix} \begin{array}{l} \\ \leftarrow j\text{-Zeile} \\ \\ \leftarrow k\text{-Zeile} \\ \\ \end{array} \quad \begin{pmatrix} a_{i1} \\ \vdots \\ a_{ij} \\ a_{ik} \\ \vdots \\ a_{in} \end{pmatrix} = \begin{pmatrix} \vdots \\ \sqrt{a_{ij}^2 + a_{ik}^2} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\rightsquigarrow \underbrace{G_1 \cdots G_n}_{Q} A = R$$

Anwendung z.B.: A Hessenberg-Matrix

Bemerkungen:

1) QR Zerlegung eignet sich besonders für Matrizen, die „nahe zu singulär" sind (schlecht konditioniert (also die Matrizen))

$$\|Ax\| = \|QRx\| = \|Rx\|$$

2) ohne Pivotisierung!

3) dünn besetzte Matrizen $\rightsquigarrow$ Givens Rotation

4) Komplexität: $\approx$ doppelt soviele Operation wie LU-Zerl. (Gauß) für Householder

5) QR-Zerlegung insbesondere Householder Tr. eignen sich zur stabilen Orthogonalisierung

6) ...

2.4. Fehleranalyse beim Lösen lin. Gl.

Sei V Vektorraum z.B. $V=\mathbb{R}^n$

<u>Norm</u> $\|\cdot\|: V \to \mathbb{R}$ mit

1) $\|u\| \geq 0, \forall u \in V$ und $\|u\|=0$ gdw $u=0$

2) $\|\alpha u\| = |\alpha| \, \|u\| \quad \forall u \in V$ und $\alpha \in \mathbb{K}$ (z.B. $\mathbb{R}$)

3) $\|u+v\| \leq \|u\| + \|v\| \quad \forall u,v \in V$

Skalarprodukt $\mathbb{K}=\mathbb{R}$ $\langle \cdot, \cdot \rangle: V \times V \to \mathbb{R}$

mit 1) $\langle u,v \rangle = \langle v,u \rangle$

2) $\langle \alpha u + \beta v, w \rangle = \alpha \langle u,w \rangle + \beta \langle v,w \rangle$

3) $\langle u,u \rangle \geq 0$ f.a. $u \in V$ mit Gleichheit falls $u=0$

$\|u\| := \sqrt{\langle u,u \rangle}$. Falls V vollständig ist. $(V, \langle \cdot, \cdot \rangle)$ ein Hilbertraum

Beispiel $V=\mathbb{R}^n$; $u=(u_1, \dots, u_n)^T \in V$

1) $\langle u,v \rangle = \sum_{i=1}^{n} u_i v_i$

2) Für $1 \leq p < \infty$ def

$\|u\|_p = \left(\sum_{i=1}^{n} |u_i|^p \right)^{1/p}$: p Norm, ℓ_p-Norm

3) $p=2$: $\|u\|_2^2 = \sum_{i=n}^{n} |u_i|^2 = \langle u,u \rangle$ Euklidische Norm

4) $p=\infty$ $\|u\|_\infty = \max_{1 \leq i \leq n} |u_i|$

Proposition: Es gilt $\forall u \in \mathbb{R}^n$

$\|u\|_\infty \leq \|u\|_2 \leq \|u\|_1 \leq \sqrt{n} \, \|u\|_2 \leq n \|u\|_\infty$

Ein Operator $A: V \to W$ (norm. Räume)

- linear: $\forall \alpha, \beta \in \mathbb{R}, u,v \in V$

$A(\alpha u + \beta v) = \alpha Au + \beta Av$

- beschränkt $\forall u \in V: \exists c > 0$

$\|Au\|_W \leq c \|u\|_V$

(A linear: stetig $\Leftrightarrow$ Lipschitz stetig $\Leftrightarrow$ beschränkt

$\|A\| = \sup_{\|v\|_V = 1} \|Av\|_W \; (= \sup_{v \neq 0} \frac{\|Av\|_W}{\|v\|_V}$

Satz 4.9 1) Sei $\mathcal{L}(V;W) = \{A: V \to W : \text{lin + beschränkt}\}$

dann $A \mapsto \|A\|$ eine Norm in $\mathcal{L}(V;W)$

2) Seien $A \in \mathcal{L}(V;W)$, $B \in \mathcal{L}(W;Z) \Rightarrow BA = B \circ A \in \mathcal{L}(V;Z)$

und $\|BA\| \leq \|B\| \|A\|$

Bew 1) $u \in V$ bel. $0 \leq \|Au\|_W \leq c\|u\|$

$\Rightarrow 0 \leq \|A\| \ (\leq c)$; $\|A0\| = \|0\| = 0$

Sei $\|A\| = 0$ d.h. $\sup\limits_{\|u\|_V = 1} \|Au\|_W = 0 \Rightarrow \|Au\|_W = 0 \ \forall u$

$\forall u \in V \quad Au = 0$ d.h. $A = 0$ dh. $\|A\| = 0 \Leftrightarrow A = 0$

$\|\lambda A\| = |\lambda| \, \|A\|$ klar

$\forall A, B \quad \|A + B\| = \sup\limits_{\|v\|_V = 1} \|(A+B)u\|_W$

$\leq \sup\limits_{\|v\|_V = 1} (\|Au\|_W + \|Bu\|_W$

$\leq \sup\limits_{\|u\|_V = 1} \|Au\|_W + \sup\limits_{\|v\|_V = 1} \|Bv\|_W = \|A\| + \|B\|$

2) Aus $\sup\limits_{u \neq 0} \dfrac{\|Au\|_W}{\|u\|_V} = \|A\|$

$\forall u \in V \quad \|Au\|_W \leq \|A\| \, \|u\|_W$

$\Rightarrow \sup\limits_{\|u\|_V = 1} \|BAu\|_Z = \sup\limits_{\|u\|_V = 1} \|B\| \, \|Au\|_W$

$\leq \|B\| \sup\limits_{\|u\|_V = 1} \|Au\|_W = \|B\| \, \|A\|$

Sei $A \in \mathbb{R}^{n \times n}$ $\lambda_1, \dots, \lambda_n \in \mathbb{C}$ Eigenwerte von A

$\rho(A) = \max\limits_{1 \leq i \leq n} |\lambda_i|$ „Spektralradius"

Sei $V = \mathbb{R}^n = W$, $A = (a_{i,j})_{i,j=1}^n \in \mathbb{R}^{n \times n}$

1) Spaltensummennorm

$\|A\|_1 = \max\limits_{1 \leq j \leq n} \sum\limits_{i=1}^n |a_{i,j}|$

2) Zeilensummennorm

$\|A\|_\infty = \max\limits_{1 \leq i \leq n} \sum\limits_{j=1}^n |a_{i,j}|$

3) Spektralnorm

$\|A\|_2 = \sqrt{\rho(A^T A)}$

Satz 1) Sei $V = \mathbb{R}^n$ versehen mit der $\|\cdot\|_1$

dann gilt für $A: V \to V$. $\|A\| = \|A\|_1$

2) Sei $V = \mathbb{R}^n$ versehen mit der $\|\cdot\|_\infty$

dann gilt für $A: V \to V$: $\|A\| = \|A\|_\infty$

3) " $\Rightarrow \|A\| = \|A\|_2$

4) $\|A\|_2 \leq \|A\|_1 \leq \|A\|_\infty$

$(a_{ij}) = A \in \mathbb{R}^{n \times m}$ $A: \mathbb{R}^m \to \mathbb{R}^n$ linear

① $\|A\|_1 = \max\limits_{1 \leq j \leq m} \sum\limits_{i=1}^{n} |a_{ij}|$ Spaltensummennorm

② $\|A\|_2 = \sqrt{\rho(A^T A)}$ Spektralnorm

größter Eigenwert,
sehr schwer zu bestimmen

Bew 1) Sei $\|u\| = \|u\|_1 = \sum\limits_{j=1}^{m} |u_j|$, $u \in \mathbb{R}^m$

[$A: V \to W$, $V = \mathbb{R}^m$, $W = \mathbb{R}^n$

$(V, \|\cdot\|_V) = (\mathbb{R}^m, \|\cdot\|_1)$; $(W, \|\cdot\|_W) = (\mathbb{R}^n, \|\cdot\|_1)$

Operatornorm $\|A\| = \sup\limits_{\|u\|_V = 1} \|Au\|_W$

$\Rightarrow \forall v \in V: \|Av\|_W \leq \|A\| \, \|v\|_V$]

„$\leq$" Sei $u \in V$ mit $\|u\|_1 = 1$

$$\|Au\|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{m} a_{ij} u_j \right| \leq \sum_{i=1}^{n} \sum_{j=1}^{m} |a_{ij}| |u_j| = \|u\|_1$$

$$= \sum_{j=1}^{m} |u_j| \left(\sum_{i=1}^{n} |a_{ij}| \right) \leq \max_{1 \leq j \leq m} \sum_{i=1}^{n} |a_{ij}| \sum_{j=1}^{m} |u_j|$$

$\|u\|_1 = 1$

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

$\Rightarrow \|A\| \leq \|A_1\|$

Operatornorm

Sei j_{max}, sodass $\sum\limits_{i=1}^{n} |a_{i, j_{max}}| = \max\limits_{1 \leq j \leq m} \sum\limits_{i=1}^{n} |a_{ij}|$

$\geq$" Setze $j = j_{max}$ und $u_j = e_j = (0 \dots 0\ \underset{j\text{-te Stelle}}{1}\ 0 \dots 0)^T \Rightarrow \|u_j\|_1 = 1$

$\Rightarrow \|Au_j\| = \|Ae_j\| = \|(a_{i,j})_{i=1}^n\|_1 = \sum_{i=1}^n |a_{i,j}| = \|A\|_1 \leq \|A\|$

2) $p=2$: $B = A^TA = B^T = B$ pos semidefinit,

denn $\langle v, Bv\rangle = v^TBv = v^TA^TAv = (Av)^T(Av) = \langle Av, Av\rangle = \|Av\|^2 \geq 0$

Seien $\lambda_1, \dots, \lambda_m$ Eigenwerte und $u_1, \dots, u_m$, $\|u_j\| = 1$ zugehörige EV.

$Bu_j = \lambda_j u_j$

Sei $u := \sum_{j=1}^m \alpha_j u_j \in \mathbb{R}^m$ mit $1 = \|u\|_2^2 = \sum_{j=1}^m \alpha_j^2$ bel

$\|Au\|_2^2 = \langle Au, Au\rangle = u^TA^TAu$ [u_j bilden ONB] [weil B symm.]

$= u^TB\sum \alpha_j u_j \overset{\text{linear}}{=} u^T \sum_{j=1}^m \lambda_j \alpha_j u_j$

$= \sum_{j,k=1}^m \underbrace{\alpha_k \lambda_j \alpha_j}_{=u^T} \underbrace{u_k^T u_j}_{=1\ \delta_{k,j}} = \sum_{j=1}^m \lambda_j \alpha_j^2$

$\leq \max_{1\leq j\leq m} |\lambda_j| \sum_{j=1}^m \alpha_j^2 = \rho(B) = \rho(A^TA)$ (Spektralradius)

$\Rightarrow \sup_{\|u\|_2=1} \|Au\|_2 \leq \sqrt{\rho(A^TA)}$ d.h. $\|A\| \leq \|A\|_2$

Sei j, sodass $\lambda_j = \max_{1\leq j\leq m} \lambda_j$. Wir nehmen $u := u_j$ EW zu λ_j

$\|u\|_2 = \|u_j\|_2 = 1$

$\|Au\|_2^2 = \|Au_j\|_2^2 = \langle Au_j, Au_j\rangle = u_j^TA^TAu_j$

$= u_j^TBu_j = \lambda_j u_j^Tu_j = \lambda_j = \rho(B) = \rho(A^TA) = \|A\|_2^2$

$\Rightarrow \|A\|_2^2 \leq \|A\|$

3) Frobenius-Norm:

$\|A\|_F^2 = \sum_{i=1}^n \sum_{j=1}^m |a_{ij}|^2$ (ist aber keine Operatornorm). Es gilt $\|A\|_2 \leq \|A_F\|$

Bsp $A = \begin{pmatrix} 1 & -3 \\ 1 & 1 \end{pmatrix} \in \mathbb{R}^{2,2}$

$\Rightarrow$ 1) $\|A\|_1 = \max\{1+1, |-3|+1\} = 4$

2) $A^TA = \begin{pmatrix} 2 & -2 \\ -2 & 10 \end{pmatrix} =: B$, $\lambda_{1,2} = 6 \pm \sqrt{60}$

$\Rightarrow \|A\| = \sqrt{6+\sqrt{60}}$

3) $\|A\|_F^2 = 1 + 1 + (-3)^2 + 1 = 12 \Rightarrow \|A\|_F = \sqrt{12}$

2.4.2 Kondition von Matrizen & Fehleranalyse

Def $A \in \mathbb{R}^{n\times n}$, dann heißt $\quad [\|\cdot\|_2$ Op-Norm in $\ell_2]$

$\mathrm{cond}_2(A) = \|A\|_2 \|A^{-1}\|_2$ die Kondition von A

Sei x Lösung von $Ax=b$ und gestörtes Problem $A(x+\Delta x) = b + \Delta b$ A, b gegeben

$\Rightarrow A\Delta x = \Delta b$ bzw $\Delta x = A^{-1}\Delta b$

wenn A singulär, gibt es Probleme mit cond = „∞"

$$\Rightarrow \frac{\|\Delta x\|}{\|x\|} = \frac{\|A^{-1}\Delta b\|}{\|x\|} \le \frac{\|A^{-1}\|\,\|\Delta b\|}{\|x\|} \frac{\|b\|}{\|b\|}$$

rel. Fehler der Lsg.

$$= \frac{\|A^{-1}\|\,\|b\|}{\|x\|} \frac{\|\Delta b\|}{\|b\|}$$ ← rel. Fehler der Eingabe

$$= \frac{\|A^{-1}\|\,\|Ax\|}{\|x\|} \frac{\|\Delta b\|}{\|b\|} \le \|A^{-1}\|\,\|A\| \frac{\|x\|}{\|x\|} \frac{\|\Delta b\|}{\|b\|} = \mathrm{cond}_2(A) \cdot \frac{\|\Delta b\|}{\|b\|}$$

„Kondition Abschätzung für Fehlerfortpflanzung bei Lösung von linearen Gleichungssystemen mit gestörten Eingabedaten"

Lemma 2.4.3 Sei $A \in \mathbb{R}^{n\times n}$ mit $\|A\| < 1$, dann existiert $(I+A)^{-1}$ und

$$\frac{1}{1+\|A\|} \le \|(I+A)^{-1}\| \le \frac{1}{1-\|A\|}$$

Bew: $(I+A)^{-1} = (I-(-A))^{-1} = \sum_{k=0}^{\infty} (-1)^k A^k$ Neumannsche Reihe

[Exkurs: $(I-(-A))x = b$, $x = -Ax + b = f(x)$ Banach FPS anwenden

$\|f(x) - f(x')\| = \|-Ax + b + Ax' - b\| = \|A(x-x')\| \le \|A\|\,\|(x-x')\|$

$\rightsquigarrow$ Kontraktion ↯ weil $\|A\| < 1$

$\Rightarrow (I+A)^{-1}$

Alternativ:

Sei $u \in \ker(I+A)$ d.h. $u + Au = 0 \Leftrightarrow \|u\| = \|Au\| \overset{\text{für } u \neq 0}{<} \|u\|$ ↯

$\Rightarrow u = 0 \Rightarrow \ker(I+A) = \{0\} \Rightarrow I + A$ regulär

$\|(I+A)x\| = \|x + Ax\| \ge \big|\|x\| - \|Ax\|\big| \ge \|x\| - \|Ax\| \ge \|x\| - \|A\|\,\|x\| = (1-\|A\|)\|x\|$

und $\|(I+A)x\| = \|x + Ax\| \le (1+\|A\|)\|x\|$

$\|y\| = \|(I+A)x\| \ge (1-\|A\|)\|x\|$

$\|y\| \le (1+\|A\|)\|x\|$

$$\Rightarrow \frac{\|y\|}{(1+\|A\|)} \le \|x\| = \|(I+A)^{-1}y\| \le \frac{\|y\|}{(1-\|A\|)} \quad \forall y$$

$$\frac{1}{(1+\|A\|)} \le \sup_{\|y\|=1} \|(I+A)^{-1}y\| \le \frac{1}{1-\|A\|}, \quad \frac{1}{1+\|A\|} \le \|(I+A)^{-1}\| \le \frac{1}{1-\|A\|} \quad \blacksquare$$

$\mathrm{cond}_2(A) = \|A\| \, \|A^{-1}\|$. $\|A\|$ Spektralnorm $\hat{=}$ Op.-Norm in ℓ_2, $\|\|_2$ Num I, VL

Lemma 2.4.13 $A: \mathbb{R}^n \to \mathbb{R}^n$. $\|A\| < 1 \Rightarrow \|(I \pm A)^{-1}\| \leq \dfrac{1}{1 - \|A\|}$

Satz 2.4.14

Sei $A \in \mathbb{R}^{n \times n}$ regulär. Sei $\Delta A \in \mathbb{R}^{n \times n}$ mit $\|\Delta A\| \leq \dfrac{1}{\|A^{-1}\|}$ bzw. $\|\Delta A\| \, \|A^{-1}\| \leq 1$,

$b, \Delta b \in \mathbb{R}^n$ gegeben

$x \in \mathbb{R}^n$ Lsg von $Ax = b$ (d.h. $x = A^{-1}b$) (1)

und $(A + \Delta A)(x + \Delta x) = b + \Delta b$, $x + \Delta x \in \mathbb{R}^n$ Lsg der gestörten Gleichung (2)

Dann gilt $\dfrac{\|\Delta x\|}{\|x\|} \leq \dfrac{\mathrm{cond}_2(A)}{1 - \mathrm{cond}_2(A) \frac{\|\Delta A\|}{\|A\|}} \left(\dfrac{\|\Delta b\|}{\|b\|} + \dfrac{\|\Delta A\|}{\|A\|} \right)$

Beweis (2) $M := (\Delta A + A) \Rightarrow \Delta x = M^{-1}b + M^{-1}\Delta b - x$ [da $M = \underbrace{A}_{\text{reg.}} + \Delta A = \underbrace{A}_{\text{reg}}(I + A^{-1}\Delta A)$

$= M^{-1}b + M^{-1}\Delta b - A^{-1}b$ $\quad \|A^{-1}\Delta A\| \leq \|A^{-1}\| \|\Delta A\| \leq 1$ vorraus

$= M^{-1}\Delta b + M^{-1}(A - M)A^{-1}b$

$= M^{-1}\Delta b + M^{-1}(A - (A + \Delta A)A^{-1}b$ $\Rightarrow M^{-1}(I + A^{-1}\Delta A)^{-1}A^{-1}$ reg (Lemma)

$= M^{-1}(\Delta b - \Delta A A^{-1}b)$

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

$\Leftrightarrow \dfrac{\|\Delta x\|}{\|x\|} \leq M^{-1}\left(\dfrac{\|\Delta b\|}{\underbrace{\|x\|}_{\|A^{-1}b\|}} + \dfrac{\|\Delta A A^{-1}b\|}{\|x\|} \right)$

$\leq \|M^{-1}\| \, \|A\| \left(\dfrac{\|\Delta b\|}{\|A\| \, \underbrace{\|A^{-1}b\|}_{\geq \|AA^{-1}b\|}} + \dfrac{\|\Delta A\| \, \|A^{-1} \cdot \|b\|}{\|A\| \, \|A^{-1}b\|} \right)$

$\leq \|M^{-1}\| \, \|A\| \left(\dfrac{\|\Delta b\|}{\|b\|} + \dfrac{\|\Delta A\|}{\|A\|} \right) \leq \dfrac{1}{\text{''}} \mathrm{cond}_2(A) \left(\text{''} \right) \leq \dfrac{1}{1 - (\text{''})} (\text{''})$

$\|M^{-1}\|$: $M^{-1} = (I + A^{-1}\Delta A)^{-1} A^{-1}$

$\|M^{-1}\| \leq \|(I + A^{-1}\Delta A)^{-1}\| \, \|A^{-1}\|$

$\leq \dfrac{\|A^{-1}\|}{1 - \|A^{-1}\Delta A\|} \leq \dfrac{\|A^{-1}\|}{1 - \|A^{-1}\| \|\Delta A\|} = \dfrac{\|A\|^{-1}}{1 - \mathrm{cond}_2(A) \frac{\|\Delta A\|}{\|A\|}}$

Bsp $A = \begin{pmatrix} 3 & 1{,}001 \\ 6 & 1{,}997 \end{pmatrix}$, $b = \begin{pmatrix} 1{,}999 \\ 4{,}003 \end{pmatrix}$

$\tilde{A} = A + \Delta A = \begin{pmatrix} 3 & 1 \\ 6 & 1{,}997 \end{pmatrix}$,

$\tilde{b} = b + \Delta b = \begin{pmatrix} 2{,}002 \\ 4 \end{pmatrix}$

$\Rightarrow$ mit Satz 2.4.14: $\|A\| \leq \|A\|_\infty = 6 + 1{,}997 = 7{,}997$,

$A^{-1} = \frac{-1}{0{,}015} \begin{pmatrix} 1{,}997 & -1{,}001 \\ -6 & 3 \end{pmatrix}$ $\quad \|A^{-1}\| \leq \|A^{-1}\|_\infty = 600$

$\Rightarrow \mathrm{cond}_2(A) \leq \mathrm{cond}_\infty(A) = 4798{,}2$, $\|\Delta b\|_\infty = 0{,}003$

$\Rightarrow \frac{\|\Delta x\|_\infty}{\|x\|_\infty} \leq 10{,}4898$ Aus (exakter) Rechnung $\Rightarrow \frac{\|\Delta x\|_\infty}{\|x\|_\infty} \approx 2{,}333$

Abschätzung war korrekt, aber sehr pessimistisch

Bemerkung

1) $\|A^{-1}\|$ ist aufwändig zu berechnen.

Aus L-R-Zerlegung (z.B. Cholesky) können $\|L^{-1}\|, \|R^{-1}\|$ leichter abgeschätzt werden.

2) QR-Zerlegung. Da $\|A\| = \|QR\| \leq \underbrace{\|Q\|}_{=1} \|R\| = \|R\|$

und $\|A^{-1}\| = \|R^{-1}Q^T\| = \|R^{-1}\|$

$\Rightarrow \boxed{\mathrm{cond}_2(A) = \mathrm{cond}_2(R)}$ Stabilität der QR-Zerlegung

3) Die EW λ_{max} und λ_{min} $\Rightarrow \mathrm{cond}_2(R) \approx \frac{|\lambda_{max}|}{|\lambda_{min}|}$ (gute Schätzung)

2.5. Lineare Regression (Ausgleichsrechnung)

- Methode der kleinsten Fehlerquadrate (Gauß) (Least-Squares)

Zu ~~m Stützstellen~~ $x_i \in \mathbb{R}^n$ $i = 1, \dots, m$ existieren Stützwerte $y_i \in \mathbb{R}$ (mit $m > n$)

mit $y_i \approx f(x_i)$; Angenommen f, $y = (y_1, \dots, y_m)^T \in \mathbb{R}^m$ gegeben.

ges $x \in \mathbb{R}^n$ $\Rightarrow y = f(x)$

Bsp $f(x) := Ax$, $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$ $\quad \underset{n}{\square}^{m} \, x = b$

d.h. Gleichungssystem ist überbestimmt

$\Rightarrow$ Gl.-syst. $Ax - b = 0$ i.A. nicht lösbar

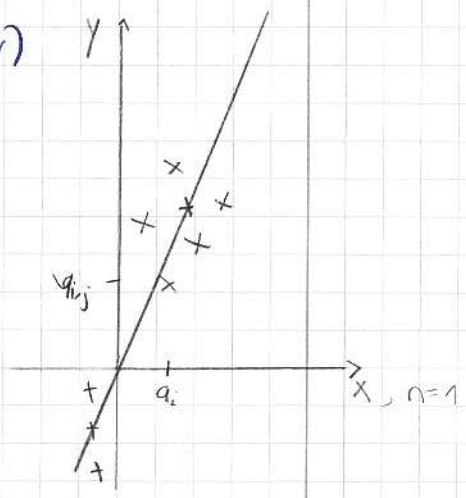
d.h. $0 \neq \underset{r(x) =}{r} := Ax - b$ (Residuum)

Optimierungsaufgabe $\|r\| \to \mathrm{Min}$ d.h. gesucht $x \in \mathbb{R}^n$ mit

$x = \underset{v \in \mathbb{R}^n}{\overset{\text{Argumentstelle von}}{\arg\min}} \|r(v)\|$, $v \in \mathbb{R}^n$ d.h. $\|r(x)\| \leq \|r(v)\| \; \forall v \in \mathbb{R}^n \Leftrightarrow \|r(x)\|^2 \leq \|r(v)\|^2$

$\|r\| := \|r\|_2$

$\Rightarrow x := \underset{v \in \mathbb{R}}{\operatorname{argmin}} \|r(v)\|_2^2 = \langle Av-b, Av-b\rangle = f(v)$



x löst $\nabla f(x) = 0^{\in \mathbb{R}^n}$; $f(x) = \|r(x)\|^2$

notwendig

$\|Ax-b\|_2^2 \to \min$

$A = (a_{ij}),\ y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$

$Ax - b \overset{!}{=} 0$?

$\|Ax-b\| \to \min$

Optimierungsproblem:

$x \to J(x) = \|Ax-b\|^2 = \langle Ax-b, Ax-b\rangle$ in $x \in \mathbb{R}^n$

$J: \mathbb{R}^n \to \mathbb{R}$, $J \in C^1(\mathbb{R}^n)$

Sei $x \in \mathbb{R}^n$, $x = \operatorname{argmin}\{J(x) : x \in \mathbb{R}^n\}$ eine Minimalstelle

$\Rightarrow$ (Ana II) $J'(x): \mathbb{R}^n \to \mathbb{R}$, $J'(x) = \vec{0}$ (notwendige Bedingung)

$J'(x) = \left(\frac{\partial J}{\partial x_1}(x) \dots \frac{\partial J}{\partial x_n}(x)\right)$, $\nabla J(x) = J'(x)^T$

$v \mapsto J(x+v) = J(x) + J'(x)v + o(\|v\|)$

$$J(x+v) = \langle A(x+v)-b, A(x+v)-b\rangle = \langle A(x+v), A(x+v)\rangle - \langle b, A(x+v)\rangle - \langle A(x+v), b\rangle + \langle b, b\rangle$$
$$= \langle A(x+v), A(x+v)\rangle - 2\langle b, A(x+v)\rangle + \langle b, b\rangle$$
$$= \langle A^TAx, x\rangle + 2\langle A^TAx, v\rangle + \|v\|^2 - 2\langle A^Tb, v\rangle - 2\langle A^Tb, x\rangle + \|b\|^2$$
$$= \underbrace{\langle Ax, Ax\rangle - 2\langle b, Ax\rangle + \langle b, b\rangle}_{= \langle Ax-b, Ax-b\rangle = J(x)} + \underbrace{2\langle A^TAx, v\rangle - 2\langle A^Tb, v\rangle}_{2(A^TAx - A^Tb)^T v} + \|v\|^2$$
$$+ 2(\underbrace{A^TAx - A^Tb}_{J'(x)})^T v + o(\|v\|^2)$$

mit $J'(x) = 2(A^TAx - A^Tb)^T$

$x = \underset{v \in \mathbb{R}^n}{\operatorname{argmin}} J(v) \Rightarrow 0 = J'(x) = 2(A^TAx - A^Tb)$, d.h. $\boxed{A^TAx = A^Tb}$

$\rightsquigarrow$ Gaußsche Normalengl.

Satz 2.5.1 (Lineare Regression)

1) Das Optimierungsproblem $\|Ax-b\| \to \min$ besitzt eine Lösung $x \in \mathbb{R}^n$ und $x \in \mathbb{R}^n$ löst die Gaußsche Normalengleichung (GNG)

$A^TAx = A^Tb$

2) Falls Rang A maximal, d.h. Rang $A = n$ $(m \geq n)$ dann ist GNG eindeutig lösbar, d.h. es existiert genau eine Minimalstelle x.

Bew. 1) Sei $\min J(x) = \|Ax-b\|^2$, d.h. x Min-Stelle

$\Rightarrow \frac{1}{2} J'(x)^T = 0 = A^TAx - A^Tb = 0$ (GNG)

„$\Leftarrow$" Sei $\tilde{x} \in \mathbb{R}^n$ mit $A^TA\tilde{x} - A^Tb = 0$ und $x \in \mathbb{R}^n$ bel.

$\Rightarrow 2\langle (x-\tilde{x}), (A^TA\tilde{x} - b)\rangle = 0$

$J(x) = \|Ax-b\|^2 = \|A(x-\tilde{x}) + A\tilde{x} - b\|^2 = \langle \,,\, \rangle$

$= \langle A(x-\tilde{x}), A(x-\tilde{x})\rangle + 2\underbrace{\langle (x-\tilde{x}), A^TA\tilde{x} - A^Tb\rangle}_{\langle A(x-\tilde{x}), A\tilde{x}-b\rangle} + \|A\tilde{x}-b\|^2$

$= \|A(x-\tilde{x})\|^2 + \|A\tilde{x} - b\|^2$ $\quad$ $= 0$, da $\tilde{x}$ Lösung der GNG

$\geq \|A\tilde{x} - b\|^2 = J(\tilde{x}) \Rightarrow \tilde{x} = \operatorname{argmin}_{v \in \mathbb{R}^n} J(v)$

2) $m \geq n$ Rang $A = n \Rightarrow \dim \operatorname{Kern}(A) = 0$

$\forall y \in \mathbb{R}^n: y^TA^TAy = \|Ay\|^2 > 0 \Leftrightarrow Ay = 0 \Leftrightarrow y = 0 \in \mathbb{R}^n$ $\quad$ $\operatorname{Kern}(A) = \{0\}$

$\Rightarrow A^TA$ ist SPD $\Rightarrow A^TA$ ist regulär ∎

Bemerkung:

Sei $A \in \mathbb{R}^{n \times n}$ $\operatorname{cond}_2(A) \gg 1$, dann ist $\operatorname{cond}_2(A^TA) \approx (\operatorname{cond}_2(A))^2 \gg\gg\gg 1$,

d.h. GNG ist oft extrem schlecht konditioniert.

Zur Lösung des Optimierungsproblems bietet sich das QR-Verfahren an.

Dann $J(x) = \|Ax-b\|^2 = \|Q(Rx - Q^Tb)\|^2 = \|Rx - Q^Tb\|$ $\quad$ von $QR = A$

$= \|\tilde{R}x - \tilde{b}_1\|^2 + \|\tilde{b}_2\|^2$ $\quad$ $R = \begin{bmatrix} \tilde{R} \\ 0 \end{bmatrix} = \tilde{R} \in \mathbb{R}^{n \times n}$ obere Δ-M

$\Rightarrow x \in \mathbb{R}^n$ löst $\tilde{R}x = \tilde{b}_1$ (Rücksubstitution) $\quad$ $Q^Tb \in \mathbb{R}^m$

und $\min_{v \in \mathbb{R}^n} J(x) = \|\tilde{b}_2\|^2$ $\quad$ $Q^Tb = \tilde{b} = \begin{bmatrix} \tilde{b}_1 \\ \tilde{b}_2 \end{bmatrix}$, $\tilde{b}_1 \in \mathbb{R}^n$

$(\operatorname{cond}_2(R) = \operatorname{cond}_2(A))$

gut konditioniert, $Q^Tb = \tilde{b} = \begin{pmatrix} \tilde{b}_1 \\ \tilde{b}_2 \end{pmatrix}$, $\tilde{b}_1 \in \mathbb{R}^n$

Beispiel: $A = \begin{pmatrix} 3 & 7 \\ 0 & 12 \\ 4 & 1 \end{pmatrix} \in \mathbb{R}^{3\times 2}$, $b = \begin{pmatrix} 10 \\ 1 \\ 5 \end{pmatrix} \in \mathbb{R}^3$

QR mit Givens-Rot

$$G_{1,3} = \begin{pmatrix} 3/5 & 0 & 4/5 \\ 0 & 1 & 0 \\ -4/5 & 0 & 3/5 \end{pmatrix}, \quad G_{2,3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 12/13 & -5/13 \\ 0 & 5/13 & 12/13 \end{pmatrix}$$

$$G_{2,3} G_{1,3} A = \begin{pmatrix} 5 & 5 \\ 0 & 13 \\ 0 & 0 \end{pmatrix}$$

$$Q^T = G_{2,3} G_{1,3} b = \begin{pmatrix} 10 \\ 37/13 \\ -55/13 \end{pmatrix} = \begin{pmatrix} \tilde{b}_1 \\ \tilde{b}_2 \end{pmatrix} \Rightarrow \tilde{b}_1 = \begin{pmatrix} 10 \\ 37/13 \end{pmatrix}$$

$$\|b_2\| = \left(\frac{55}{13}\right)$$

Löse

$$\tilde{R} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 5 & 5 \\ 0 & 13 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 10 \\ 37/13 \end{pmatrix} \Rightarrow x = \begin{pmatrix} 301/169 \\ 37/169 \end{pmatrix}$$

2.6. Singulärwertzerlegung und Pseudoinverse

Satz 2.6.1. (SVD, Singular value decomposition)

Sei $A \in \mathbb{R}^{m\times n}$ beliebig, dann ex. $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ orthogonal und $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \geq 0$, $r = \min(m,n)$ Singulärwerte (σ_i) mit der Darstellung:

$$A = U \cdot \Sigma V^T \quad \text{bzw.} \quad \Sigma = U^T A V = \operatorname{diag}(\sigma_1, \ldots, \sigma_r) = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \\ & 0 & \\ & \vdots & \\ 0 & & 0 \end{pmatrix} (m > n)$$

oder: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$ $\overset{\nabla}{0}$ $(r \leq \min(m,n))$

$$\Rightarrow \Sigma = \begin{pmatrix} \sigma_1 & & 0 \\ & \sigma_2 & \\ 0 & & 0 \\ \vdots & & \vdots \\ 0 & & 0 \end{pmatrix} \quad m > n$$

~~Eigenwertzerlegung~~ ~~falls $A^T = A$~~: $A = U \Sigma U^T$ d.h. $V = U$
bzw. $U \Sigma U^{-1}$
falls A SPD

Beweisidee: $B = A^T A$ ist symm. und pos. semidef. Die nichtnull Eigenwerte von B sind die Singulärwerte

SVD (2.6.1)

$A=(a_{i,j})\in\mathbb{R}^{m\times n} \Rightarrow \exists\ \sigma_1\geq\ldots\geq\sigma_r>\sigma_{r+1}=0$

$U\in\mathbb{R}^{m\times n}, V\in\mathbb{R}^{n\times n}$ orthogonal

$$A=U\,\mathrm{diag}(\sigma_i)V^T,\ \mathrm{diag}(\sigma_i)=\begin{bmatrix}\sigma_1 & & & & 0\\ & \ddots & & & \\ & & \sigma_r & & \\ & & & 0 & \\ 0 & & & & \ddots\ 0\end{bmatrix}=\Sigma$$

Beweis Sei OBdA $m\leq n$ und $B=A^TA\in\mathbb{R}^{n\times n}$

Symmetr. $u^TBu=u^TA^TAu=(Au)^TAu=\|Au\|^2\geq 0\quad\forall u\in\mathbb{R}^n$

$\Rightarrow Bv_k=\lambda_k v_k\quad \|v_k\|=1 \Rightarrow \lambda_k\geq 0,\ k=1,\ldots,n$

$\ldots\lambda_{k-1}\geq\lambda_k\geq 0$

und $V=(v_1\ldots v_k)\in\mathbb{R}^{n\times n}$ orthogonal

wir definieren $\sigma_i=\sqrt{\lambda_i},\ i=1,\ldots,n\qquad \sigma_1\geq\sigma_2\geq\ldots\geq\sigma_r>\sigma_r=0\ldots=\sigma_n=0$

$[C=AA^T\in\mathbb{R}^{m\times m}]$

Für $j=1,\ldots,r$ Seien $u_j:=\frac{1}{\sigma_j}Av_j\qquad (r\leq\min\{m,n\})$

Für $j=r+1,\ldots,m$ Sei $\{u_j\ j=r+1,\ldots,m\}$ eine ONB von $(\mathrm{span}\{u_i\ i=1,\ldots,r\})^\perp$

$=\{w\in\mathbb{R}^m\ u_i^Tw=0\ \forall i=1,\ldots,r\}$

Zeigen $\langle u_i,u_j\rangle=u_i^Tu_j=\delta_{i,j}\quad (i>r$ oder $j>r)$

Für $i,j=1,\ldots,r \Rightarrow u_i^Tu_j=u_i^TAv_j\frac{1}{\sigma_j}=\frac{1}{\sigma_i}v_i^T\underbrace{A^TA}_{B}v_j\frac{1}{\sigma_j}$

$$=\frac{1}{\sigma_i\sigma_j}\lambda_j\underbrace{v_i^Tv_j}_{=\delta_{i,j}}=\frac{\overbrace{\lambda_j}^{=1}}{\sigma_j^2}\delta_{i,j}$$

$$U^TAV=\mathrm{diag}(\sigma_j)U^TU=(\sigma_j\delta_{i,j})_{\substack{i=1,\ldots,m\\ j=1,\ldots,n}}=\Sigma$$

Korollar 2.6.2: $A=U\Sigma V^T$ aus Satz 2.6.1 $\Rightarrow$ ~~Rang A~~ $=r$ ($\sigma_1\geq\ldots\geq\sigma_r>\sigma_{r+1}$

1) $\mathrm{Rang}A=r$

2) $\mathrm{Im}(A)=\mathrm{Bild}(A)=\mathrm{span}\{u_1,\ldots,u_r\}$

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

$\Leftrightarrow (w\in\mathrm{Kern}A \Leftrightarrow w^Tv_j=0,\ j=1,\ldots,r$

Bemerkung: Rang r - Darstellung von A

$$A=\sum_{k=1}^r\sigma_k u_k\otimes v_k=\sum_{k=1}^r\sigma_k\underbrace{u_k}_{\in\mathbb{R}^m}\underbrace{v_k^T}_{\in\mathbb{R}^n}\ (\text{Komplexität nnz}=\mathcal{O}(\overset{r+nr+mr}{r+nr+mr}))$$

$$=\sum_{k=1}^r u_k\tilde{v}_k^T\quad(\tilde{v}_k=\sigma_k v_k)$$

Alternative Darstellung

$A = \hat{U} R \hat{V}^T$, $\hat{U} \in \mathbb{R}^{m \times m}$, $\hat{V} \in \mathbb{R}^{n \times n}$ orth. $R \in \mathbb{R}^{m \times n}$ regulär

$\hat{I} = GG^{-1}$, $I = FF^{-1}$

$A = \hat{U} R \hat{V}^T = \tilde{U} \tilde{R} \tilde{V}^T$ $\quad \tilde{U} = \hat{U} G \quad \tilde{R} = G^{-1} R F^{-1}$

$\tilde{V} = \hat{V} F$

$\operatorname{span}\{\tilde{u}_j \; j = 1, \dots, m\} = \operatorname{span}\{\hat{u}_j \; j = 1, \dots, r\} = \operatorname{Im}(A)$

$u, \tilde{v} \to \tau(v, \tilde{u}) = \sum_{j=1}^{r} u_j \tilde{v}_j = uv^T$, $\tau: \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r} \to \mathbb{R}^{m \times n}$ bilinear

Korollar 2.6.3

1) Operator Norm $\|A\| = \|A\|_{op} = \sigma_1$

2) Frobenius Norm $\|A\|_F^2 = \sum_{i,j} |a_{ij}|^2 = \sum_{k=1}^{r} \sigma_k^2$

3) Nuklearn (Spur) Norm $\|A\|_1 := \sum_{k=1}^{r} \sigma_k$

4) Sei $A \in \mathbb{R}^{n \times n}$ $(m = n)$ $\operatorname{cond}_2 A = \frac{\sigma_1}{\sigma_n}$

Def. Verallgemeinerte Kondition

$A \in \mathbb{R}^{m \times n}$, $\operatorname{cond}_2 A = \frac{\sigma_1}{\sigma_r}$

Satz 2.2.4 Satz von Schmidt & Mirsky, $A = U \Sigma V^T$

1) $\underset{\operatorname{Rang} B = s}{\arg\min} \|A - B\|_F = \sum_{k=1}^{s} \sigma_k u_k v_k^T$, $\min_{\operatorname{rang} B = s} \|A - B\|_F^2 = \sum_{k=s+1}^{r} \sigma_k^2$

$A - B = \sum_{k=1}^{r} \sigma_k u_k v_k^T - \sum_{k=1}^{s} \sigma_k u_k v_k^T = \sum_{k=s+1}^{r} \sigma_k u_k v_k^T$

Beste Approximation ⇒ SVD und dann kleinsten Singulärwerte weglassen, also nur bis bestimmten s gehen

2) $\underset{\operatorname{Rang} B = s}{\arg\min} \|A - B\|_{op} = \sum_{k=1}^{s} \sigma_k u_k v_k^T$, $\min_{\operatorname{rang} B = s} \|A - B\|_{op} = \sigma_{s+1}$

3) (analog in $\|\cdot\|_{*,1}$)

4) $A, B \in \mathbb{R}^{m \times n}$, mit Singulärwerten $\sigma_{A,i}, \sigma_{B,i} \geq 0$

$\sum_{i=1}^{\min\{m,n\}} |\sigma_{A,i} - \sigma_{B,i}|^2 \leq \|A - B\|_F^2$

Übungsblatt 11 Schwer 1/1 Satz 2 ab morgen Wissenschaft

Beispiel

$$A = \begin{pmatrix} 0{,}48 & -0{,}36 \\ 0{,}006 & 0{,}008 \\ 0{,}64 & 0{,}48 \end{pmatrix} \overset{SVD}{=} \underbrace{\begin{pmatrix} 0{,}6 & 0 & 0{,}8 \\ 0 & 1 & 0 \\ -0{,}8 & 0 & 0{,}6 \end{pmatrix}}_{U} \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 0{,}01 \\ 0 & 0 \end{pmatrix}}_{\Sigma} \underbrace{\begin{pmatrix} 0{,}8 & -0{,}6 \\ 0{,}6 & 0{,}8 \end{pmatrix}}_{V^T}$$

$\sigma_1 = 1$

$\sigma_2 = 0{,}01$

Bemerkung Die Berechnung der SVD mit Eigenwertlösung zur Matrix $B = A^TA$ hat den Nachteil, dass $\text{cond}_2 B \sim (\text{cond}_2 A)^2$ ist schlecht konditioniert

Stabile Alternative: Alg von Golub & Kahan (Matlab)

Definition Sei $A \in \mathbb{R}^{m \times n}$, $m \geq n$

Rang $A = n$

$\Rightarrow \|Ax - b\|^2 \to \min$ Gauß-N.G

$A^TAx = A^Tb$ bzw. $x = (A^TA)^{-1}A^Tb = A^+ b$

$A^+ := (A^TA)^{-1}A^T$, Pseudo-Inverse (Moore-Penrose-Inverse)

Satz 4.6.5: Es gilt: $A = U\Sigma V^T$, $r = n$

$$A^+ = V \underbrace{\Sigma^+}_{\in \mathbb{R}^{n \times m}} U^T, \quad \Sigma^+ = \begin{pmatrix} \sigma_1^{-1} & & 0 \\ & \ddots & \\ 0 & & \sigma_n^{-1} \end{pmatrix}$$

Beweis: $(A^TA)^{-1}A^T = (V\Sigma^T U^T U \Sigma V^T)^{-1}(U\Sigma V^T)^T$

$= (V\Sigma^T U^T U \Sigma V^T)^{-1}(V\Sigma^T U^T)$

$= V(\Sigma^T\Sigma)^{-1}V^T V \Sigma^T U^T$

$= V(\Sigma^T\Sigma)^{-1}\Sigma^T U^T$

$= V\Sigma^+ U^T = A^+$ ∎

Satz (Mirsky-Schmidt) $A \in \mathbb{R}^{m \times n}$, $A = U \Sigma V^T$, $\operatorname{rank} A = r \leq \min\{m,n\}$ Num I, VL

$B = \arg\min \{ \|A-B\|_F : \operatorname{rank} B \leq s \}$, B beste rang s Approximation von A

Dann ist $B = U \Sigma_s V^T = \sum_{k=1}^{s} \sigma_k \underbrace{u_k v_k^T}_{\in \mathbb{R}^{m \times n}}$ $\qquad \Sigma_s = \begin{pmatrix} \sigma_1 & & & & 0 \\ & \ddots & & & \\ & & \sigma_s & & \\ & & & 0 & \\ 0 & & & & \ddots \; 0 \end{pmatrix}$

$\|A-B\|_F^2 = \sum_{k=s+1}^{r} \sigma_k^2$

<u>Beweis</u> $r = s$, $B = A$, $\operatorname{rang} A = r = s$ $(\|A-B\| = 0)$

1) Sei $B = U \Sigma_s V^T \Rightarrow \|A-B\|_F^2 = \|U \Sigma V^T - U \Sigma_s V^T\|_F^2$
$\operatorname{rang} B = s$

$$= \|U(\Sigma - \Sigma_s)V^T\|_F^2 = \|\Sigma - \Sigma_s\|_F^2 = \sum_{k=1}^{r} \sigma_k^2 - \sum_{k=1}^{s} \sigma_k^2$$
$$= \sum_{k=s+1}^{r} \sigma_k^2 \geq 0$$

$$= \sum_{k=1}^{r} \sigma_k^2 - \sum_{k=1}^{s} \sigma_k^2 = \|U \Sigma V^T\|_F^2 - \sum_{k=1}^{s} \sigma_k^2 = \|A\|_F^2 \sum_{k=1}^{s} \sigma_k^2 \quad (*)$$

<u>Lemma 1)</u> Seien $x_k \in \mathbb{R}^m$, $y_k \in \mathbb{R}^n$ bel. $k = 1, \dots, s$. $\|\sum_{k=1}^{s} x_k y_k^T - A\|_F^2$
$$\geq \|A\|_F^2 - \sum_{k=1}^{s} \sigma_k^2$$

$(*) \overset{\text{Lemma 1)}}{\Longrightarrow} \leq \|A-C\|_F^2 \quad \forall C \in \mathbb{R}^{m \times n}, \operatorname{rang} C \leq s$

<u>Beweis Lemma 1):</u>

OBdA seien x_k orthogonal, d.h. $x_k^T x_j = \delta_{kj}$

Andernfalls orthogonalisieren wir $\tilde{x}_k = \sum_{k=1}^{s} S_{k} x_k$; $S \in \mathbb{R}^{s \times s}$ reg. $\Rightarrow x_k = \sum_{k=1}^{s} S^{-1} \tilde{x}_k$

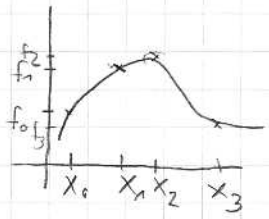
$\sum_{k=1}^{s} x_k y_k^T = \sum_{k,k'=1}^{s} \tilde{x}_k S_{kk'}^{-1} y_{k'}^T$, $\tilde{x}_k$ ON

$= \rightsquigarrow$ Beweis (ganzer) siehe Bärwolff-Skript S. 42 Satz 3.16

3. Interpolation

- kompl. Funktionen zB. $f(x) = \sin(\ln(x^2+1))$
- (x_k, f_k), $k = 0, 1, \dots, n$
- $[0,1\,a]$, $x_k = \frac{k}{10} \cdot 10$, $k = 0, \dots, 10$

  $f_k = f(x_k)$, $(x_k\, f_k)$, $k = 0, \dots, 10$



3.1. Das allgemeine Interpolations-Problem

<u>Def</u>: 3.1. (lin. Funktional)

Sei $\mathcal{F}$ ein lin. Raum (z.B $C[a,b]$, $C^\infty(\Omega)$, $L_2(\Omega)$...).

Eine Abb. $\mu: \mathcal{F} \to \mathbb{R}$ heißt lineares Funktional, falls

$$\mu(\alpha_1 f_1 + \alpha_2 f_2 = \alpha_1 \mu(f_1) + \alpha_2 \mu(f_2),\ \alpha_1, \alpha_2 \in \mathbb{R},\ f_1, f_2 \in \mathcal{F} \quad (3.1)$$

<u>Beispiel</u>
- $\mathcal{F} = C[a,b]$, $x_0 \in [a,b]$, $\mu(f) = f(x_0)$
- $\mathcal{F} = C^1[a,b]$, $x_0 \in [a,b]$, $\mu(f) = f'(x_0)$
- $\mathcal{F} = R[a,b]$ (Raum der riemannintegrierbaren Fkt), $\mu(f) = \int_a^b f(x)\,dx$

<u>Allgemeines Interpolationsproblem</u>

Sei $\mathcal{F}$ ein Funktionenraum und G_n ein $(n+1)$-dim Teilraum von $\mathcal{F}$. Weiter seien lineare Funktionale $\mu_0, \dots, \mu_n$ auf G_n:

$$\begin{cases} \text{zu } f \in \mathcal{F} \text{ finde } g_n \in G_n \text{ mit} \\ \mu_j(g_n) = \mu_j(f),\ j = 0, \dots, n \end{cases} \quad (3.2)$$

<u>Lemma 3.1.1.</u> Existenz und Eindeutigkeit

Die allgemeine Interpolationsaufgabe (3.2) hat genau dann für jedes $f \in \mathcal{F}$ eine eindeutige Lösung g_n, wenn

$\det\left[(\mu_i(\varphi_j))_{i,j=0}^n\right] \neq 0$ für eine (und damit jede) Basis $[\varphi_0, \dots, \varphi_n]$ von G_n

<u>Beweis</u> Sei $g_n \in G_n$, dann existieren eind. bestimmte Koeffizienten $[\alpha_0, \dots, \alpha_n]$ mit $g_n = \sum_{j=0}^n \alpha_j \varphi_j$, also bedeutet (3.2)

$$\mu_i(g_n) \overset{\text{lin.}}{=} \sum_{j=0}^n \alpha_j \mu_i(\varphi_j) = (G\alpha)_i = \mu_i(f) = b_i, \quad G = (\mu_i(\varphi_j))_{i,j=0}^n,$$

$G\alpha = b$ eindeutig lösbares LGS $\qquad \alpha = (\alpha_j)_{j=0}^n,\ b = (b_j)_{j=0}^n$

Bemerkung: $G_n = \mathbb{P}_n \rightarrow$ Polynominterpolation

$G_{2n} = [F(x) = \frac{P(x)}{Q(x)}, P, Q \in \mathbb{P}_n, Q \neq 0] \rightarrow$ rationale Interpolation

3.2 Polynominterpolation

3.2.1. Lagrange-Interpolation

$(x_k, f_k), k = 0, 1, \ldots, n$

gesucht ist ein Polynom $P \in \mathbb{P}_n$ vom Grad $p \leq n$, d.h.

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0, \; a_\mu \in \mathbb{R}, \mu = 0, \ldots, n$$

$$P(x_i) = f_i, \; i = 0, \ldots, n \quad (3.3)$$

Basis $\varphi_j = x^j, j = 0, \ldots, n$ ← Monom-Basis

Für (3.3.) erhalten wir $\mu_i(f) = f(x_i), i = 0, \ldots, n$

$\mu_i(\varphi_j) = x_i^j \rightsquigarrow (\mu_i(\varphi_j))_{i,j=0}^n$ ist die Vandermonde-Matrix

$$\begin{pmatrix} 1 & x_0 & x_0^2 & x_0^3 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & x_1^3 & \ldots & x_1^3 \\ | & & & & & | \\ 1 & x_n & x_n^2 & x_n^3 & \ldots & x_n^3 \end{pmatrix}$$

Definition 3.2.1.

$$L_i^{(n)}(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)}, \; i = 0, \ldots, n \quad (3.4)$$

Lagrange Basispolynome, paarweise verschiedene $x_0, \ldots, x_n$

Bemerkung $L_i^{(n)}(x_k) = \delta_{i,k} = \begin{cases} 1 & \text{für } i = k \\ 0 & \text{für } i \neq k \end{cases}$ (3.5)

$\rightsquigarrow L_i^{(n)}$ sind linear unabhängig

$\rightsquigarrow$ Basis von $\mathbb{P}_n$

Definition: Lagrange-Darstellung

Das Polynom $P(x) = \sum_{i=0}^{n} f_i L_i^{(n)}(x)$ wird Langrange-Darstellung des Interpolationspolynoms zu den Stützstellen $(x_0, f_0), \ldots, (x_n, f_n)$ genannt

$$P(x_k) = \sum_{i=0}^{n} f_i \underbrace{L_i^{(n)}(x_k)}_{\delta_{i,k}} = \sum_{i=0}^{n} f_i \delta_{i,k} = f_k, \; k = 0, \ldots, n$$

Eindeutigkeit: $P, Q \in \mathbb{P}_n, P(x_k) = Q(x_k) = f_k \; (k = 0, \ldots, n)$

$D(x) = P(x) - Q(x) \in \mathbb{P}_n$, D hat $n+1$ paarweise verschiedene NS (Nullstellen), $x_0, \ldots, x_n$ $\rightsquigarrow D \equiv 0$ Nullpolynom

Beispiel

x_k	f_k
0	2
1	4
2	7

$n=2$

$$L_0^{(2)}(x) = \frac{x-1}{0-1} \cdot \frac{x-2}{0-2} = \frac{1}{2}(x-1)(x-2)$$

$$L_1^{(2)}(x) = \frac{x-0}{1-0} \cdot \frac{x-2}{1-2} = -x(x-2)$$

$$L_2^{(2)}(x) = \frac{x-0}{2-0} \cdot \frac{x-1}{2-1} = \frac{1}{2}x(x-1)$$

$$\rightsquigarrow P_2(x) = \sum_{i=0}^{2} f_i L_i^{(2)}(x) = 2 \cdot \frac{1}{2}(x-1)(x-2) + 4(-x(x-2)) + 7 \cdot \frac{1}{2}(x(x-1))$$

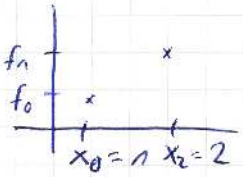
Problem (x_{n+1}, f_{n+1}) soll hinzugefügt werden $\rightsquigarrow$ Basis-Polynome müssen neu berechnet werden

3.2.1. Monombasis

schlecht konditioniert, großer Aufwand

$\{1, x, x^2, \dots, x^n\}$ von $\mathbb{P}_n$.

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \vdots & & & & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_n \end{pmatrix}$$



+ f_0' ist gegeben. $P_2(x_0) = f_0 = 1$

$P_2'(x_0) = f_0' = 0$

$P_2(x_1) = f_1 = 1$

$P_2(x) = a_0 + a_1 x + a_2 x^2$

$P_2'(x) = a + 2a_2 x$

$$\longrightarrow \begin{aligned} a_0 + a_1 + a_2 &= 1 \\ a_1 + 2a_2 &= 0 \\ a_0 + 2a_1 + 4a_2 &= 1 \end{aligned} \longrightarrow \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 1 & 2 & 4 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

3.2.2. Hermite-Interpolation

$a = x_0 \leq x_1 \leq \dots \leq x_n = b$

! Stützstellen können auch mehrfach auftreten

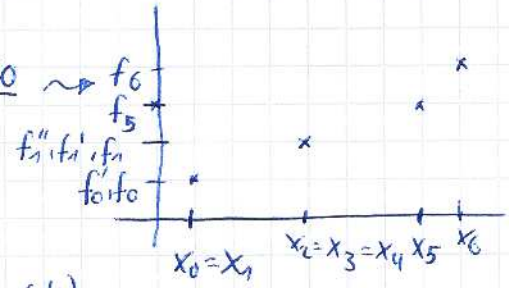
$x_i \rightarrow f(x_i) = f_i, \quad f'(x_i) = f_i', \dots, f^{(k)}(x_i) = f_i^{(k)}$

$\rightsquigarrow x_i$ soll in der Folge der Stützstellen $(k+1)$-mal auftreten

gleich Knoten (Stützstellen) nummerieren wir mit der Vielfachheit

$d_i = \max\{ j \in \mathbb{N} \mid x_i = x_{i-j} \}$

x_i	$x_0 = x_1 < x_2 = x_3 = x_4 < x_5 < x_0$
d_i	0 1 0 1 2 0 0



$\mu_i : C^n[a,b] \to \mathbb{R},\ \mu_i(f) = f^{(d_i)}(x_i),\ i = 0, \dots, n$

Aufgabe der Hermite-Interpolation:

gegeben: $\mu_i\ (i = 0, \dots, n)$, finde $P \in \mathbb{P}_n$ mit $\mu_i(P) = \mu_i(f)$ $i = 0, \dots, n$ (3.8)

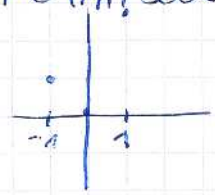
Die Lösung $P(f \mid x_0, \dots, x_n) \in \mathbb{P}_n$ heißt Hermite-Interpolierende.

<u>Satz</u> Zu jeder Funktion $f \in C^n[a,b]$ und jeder monotonen Folge $a = x_0 \le x_1 \le \dots \le x_n = b$ von Stützstellen (Knoten) gibt es genau ein Polynom $P \in \mathbb{P}_n$, sodass $\mu_i(P) = \mu_i(f), i = 0, \dots, n$.

<u>Beispiel</u>: $x_0 = -1,\ f_0 = 1$

$x_1 = 0,\ f_1' = 0$

$x_2 = 1,\ f_2 = 3$



? $P_2(x) = a_0 + a_1 x + a_2 x^2$

$P_2' = a_1 + 2a_2 x$

$P_2(-1) = a_0 - a_1 + a_2 = 1$

$P_2'(0) = a_1 = 0$

$P_2(1) = a_0 + a_1 + a_2 = 3$

$$\rightsquigarrow \begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 3 \end{bmatrix}$$

3.2.3 Auswertung von Polynomen

<u>Ziel</u> $P = P(f \mid x_0, \dots, x_n)$ soll effizient berechnet werden

Schema von Aitken/Neville

Schema der dividierten Differenzen

$x_k,\ k = 0, \dots, n$, paarweise verschieden: Basis von $\mathbb{P}_n \rightsquigarrow$ Monombasis, Lagrange-Basis, Newton-Basis

<u>Definition</u>: Seien $x_0, x_1, \dots, x_{n-1} \in \mathbb{R}$ mit

$\omega_0 := 1,\ \omega_i(x) = \prod_{j=0}^{i-1} (x - x_j),\ i = 1, \dots, n,\ \omega_i \in \mathbb{P}_i$, wir bezeichnen $\{\omega_0, \dots, \omega_n\}$ als Newton-Basis der Polynomraumes $\mathbb{P}_n$

$$\rightsquigarrow P(x) = c_0 \omega_0 + c_1 \omega_1 + \dots + c_n \omega_n$$

$$= c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + \dots + c_n \overbrace{\prod_{j=0}^{n-1} (x - x_j)}^{\omega_n} = \sum_{i=0}^{n} c_i \omega_i(x)$$

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

$\cancel{P(x_0) = c_0,\ P(x_1) = c_0 + c_1(x_1 - x_0)}$

$P(x_0) = c_0 \qquad = f(x_0) = f_0$

$P(x_1) = c_0 + c_1(x - x_0) \qquad = f_1$

$P(x_2) = c_0 + c_1(x_1 - x_0) + c_2(x_2 - x_0)(x_2 - x_1) \qquad = f_2$

$\vdots$

$P(x_n) = c_0 + c_1 \omega_1(x_1) + \dots + c_n \omega_n(x_n) = f_n$

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & (x_1 - x_0) & \cdots & 0 \\ 1 & (x_2 - x_0) & (x_2 - x_0)(x_2 - x_1) & 0 \\ 1 & \cdots & \cdots & \omega_n(x_n) \end{pmatrix} \begin{pmatrix} c_0 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} f_0 \\ \vdots \\ f_n \end{pmatrix}$$

Basen von $\mathbb{P}_n$: *) $\{1, x, x^2, \dots, x^n\}$, $P(x) = a_0 + a_1 x + a_2 x^2 \dots + a_n x^n$

$Va = f$, $a = \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix}$

*) $\left\{ L_i^{(n)}(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j},\ i = 0, \dots, n \right\}$, $P(x) = \sum_{k=0}^{n} b_k L_k^{(n)}(x) \to \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} \begin{pmatrix} b_0 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} f_0 \\ \vdots \\ f_n \end{pmatrix}$

$b_k = f_k$

*) $\{\omega_k(x), k = 0, \dots, n\}$, $\omega_k(x) = \prod_{j=0}^{k-1} (x - x_j)$, $\omega_0(x) := 1$,

$P(x) = \sum_{j=0}^{n} c_j \omega_j(x) \to \begin{pmatrix} \ddots & 0 \\ & \ddots \end{pmatrix} \begin{pmatrix} c_0 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} f_0 \\ \vdots \\ f_n \end{pmatrix}$

$\mathbb{P}_n \to \mathbb{P}_{n+1}$, $\omega_{n+1} = \prod_{j=0}^{n} (x - x_j)$

<u>Lemma</u> (Aitken)

Für das Interpolationspolynom $P = P(f \mid x_0, \dots, x_n)$ gibt die Rekursionsformel $P(f \mid x_0, \dots, x_n)(x) = \underbrace{\frac{(x_0 - x) P(f \mid x_1, \dots, x_n)(x) - (x_n - x) P(f \mid x_0, \dots, x_{n-1})}{x_0 - x_1}}_{\phi(x) \in \mathbb{P}_n}$

insbesondere gilt $P(f \mid x_k) = f_k$

<u>Beweis</u> $\phi(x_i) = \frac{(x_0 - x_i) f_i - (x_n - x_i) f_i}{x_0 - x_n} = f_i$, $i = 1, \dots, n-1$, $\phi(x_0) = \dots = f_0$

$\phi(x_n) = \dots = f_n$

<u>Algorithmus</u> $P(f \mid x_i) = f_i$, $i = 0, \dots, n$

x fest $P_{i,k} := P(f \mid x_{i-k}, \dots, x_i)(x)$, $i \geq k$ $\Big\}$ $\cancel{P_{n,n} = (x_0 - x) P_{n,n-1} - (x_n - x)}$ (*)

(*) $P_{n,n} = \frac{(x_0 - x) P_{n,n-1} - (x_n - x) P_{n-1,n-1}}{x_0 - x_n}$

allgemeiner: $P_{i,k}$, $i \geq k$: $P_{i,k} = \frac{(x_{i-k} - x) P_{i,k-1} - (x_i - x) P_{i-1,k-1}}{x_{i-k} - x_i}$

$= P_{i,k-1} + \frac{x_i - x}{x_{i-k} - x_i} (P_{i,k-1} - P_{i-1,k-1})$

1) $P_{i,0} = f_i \quad i = 0, \dots, n$

2) $P_{i,k} = P_{i,k-1} + \frac{x - x_i}{x_i - x_{i-k}} (P_{i,k-1} - P_{i-1,k-1})$

$$\begin{array}{lll}
f_0 = P_{0,0} & & \\
f_1 = P_{1,0} & P_{1,1} & \\
f_2 = P_{2,0} & & P_{n-1,n-1} \\
\vdots & \dots & \\
f_{n-1} = P_{n-1,0} & & P_{n,n-1} \\
f_n = P_{n,0} & P_{n,1} & & P_{n,n}
\end{array}$$

<u>Newton-Darstellung des Interpolationspolynoms und dividierte Differenzen</u>

$$\underbrace{P(f|x_0, \dots, x_n)}_{\in \mathbb{P}_n}(x) = \underbrace{P(f|x_0, \dots, x_{n-1})}_{\in \mathbb{P}_{n-1}} + \underbrace{Q_n(x)}_{\in \mathbb{P}_n},$$

$$Q_n(x_i) = P(f|x_0, \dots, x_n)(x_i) - P(f|x_0, \dots, x_{n-1})(x_i) = 0$$

$\rightsquigarrow Q_n(x) = a_n (x - x_0)(x - x_1) \cdot \ldots \cdot (x - x_{n-1})$ ← x_n, d.h. vor

$$a_n = \frac{f_n - P(f|x_0, \dots, x_{n-1})(x_n)}{(x_n - x_0) \cdot \ldots \cdot (x_n - x_{n-1})}$$

↳ führender Koeffizient von $P(f|x_0, \dots, x_n)$, d.h. vor x^n

<u>Definition</u> (n-te dividierte Differenz)

Der führende Koeffizient a_n des Interpolationspolynoms $P(f|x_0, \dots, x_n)(x) = a_n x^n + \dots + a_1 x + a_0$ von f zu den Knoten $x_0 \leq x_1 \leq \dots \leq x_n$ heißt n-te dividierte Differenz von f in $x_0, \dots, x_n$ und wird mit

$$[x_0 \dots x_n] f := a_n \quad ([x_0 \dots x_n] = a_n)$$

2. dividierte Differenz: $P(f|x_0 x_1 x_2) = a_2 x^2 + a_1 x + a_0$

<u>Folgerung</u> $P(x) = \sum_{j=0}^{n} c_j \omega_j(x) = \sum_{j=0}^{n} a_n x^j \rightsquigarrow c_n = a_n$

<u>Satz</u> (Newtonsche Interpolationsformel)

$$P(x) = \sum_{i=0}^{n} [x_0 \dots x_i] f \, \omega_i(x) \; (= P(f|x_0 \dots x_n))$$

Gilt $f \in C^{n+1}(\mathbb{R}) \longrightarrow f(x) = P(x) + [x_0 \dots x_n \, x] f \cdot \omega_{n+1}(x)$

Beweis Induktion über $n \in \mathbb{N}$

$n=0 \rightsquigarrow P(x) = \sum_{i=0}^{0} [x_0] f \omega_i = [x_0] f \omega_0 = [x_0] f = f_0$

$n>0 \quad P_{n-1} := P(f|x_0 - x_{n-1}) = \sum_{i=0}^{n-1} [x_0 - x_i] f \omega_i$

$$P_n(x) = [x_0 - x_n] f x^n + a_{n-1} x^{n-1} + \dots + a_0 = [x_0 - x_n] f \omega_n(x) + \underbrace{Q_{n-1}(x)}_{\in \mathbb{P}_{n-1}} = \circledast$$

Es gilt $\omega_n(x_i) = 0, \; i = 0, \dots, n-1 \rightsquigarrow \underbrace{Q_{n-1}}_{} = P_n - [x_0 - x_n] f \omega_n$

erfülle die Interpolationsaufgabe für $x_0, \dots, x_{n-1}$

$$\circledast = [x_0 - x_n] f \omega_n(x) + \sum_{i=0}^{n-1} [x_0 - x_i] f \omega_i(x) = \sum_{i=0}^{n} [x_0 - x_i] f \omega_i(x)$$

Lemma (Eigenschaften der dividierten Differenzen, rekursive Berechnung)

(i) Für $x_i \neq x_k$ gilt $[x_0 - x_n] = \dfrac{[x_0 - \hat{x}_i - x_n] f - [x_0 - \hat{x}_k - x_n] f}{x_k - x_i}$

(ii) Für zusammenfallende Knoten $x_0 = x_1 = \dots = x_n$ gilt $\underbrace{[x_0 - x_0]}_{(n+1)\text{mal}} = \dfrac{f^{(n)}(x_0)}{n!}$

$, [x_0 - \hat{x}_i - x_n] f = [x_0 - x_{i-1} x_{i+1} - x_n] f$

Beweis:

(i) $P(f|x_0 - x_n) = \dfrac{(x_i - x) \overbrace{P(f|x_0 - \hat{x}_i - x_n)}^{\rightarrow [x_0, \dots, \hat{x}_i, \dots, x_n] f x^{n-1} + P_{n-2}} - (x_k - x) P(f|x_0 - \hat{x}_k - x_n)}{x_i - x_k}$

(ii) $P(f|x_0 - x_0)(x) = \sum_{j=0}^{n} \dfrac{(x-x_0)^j}{j!} f^{(j)}(x_0) \rightsquigarrow \dfrac{f^{(j)}(x_0)}{j!} = [x_0 - x_0] f$

Beispiel

$x_0 : [x_0] f = f_0$
$x_1 : [x_1] f = f_1$
$x_2 : [x_2] f = f_2$
$x_3 : [x_3] f = f_3$
$x_4 : [x_4] f = f_4$

$[x_0 x_1] f$, $[x_1 x_2] f$, $[x_2 x_3] f$, $[x_3 x_4] f$

$[x_0 x_1 x_2] f$, $[x_1 x_2 x_3] f$, $[x_2 x_3 x_4] f$

$[x_0 x_1 x_2 x_3] f$, $[x_1 x_2 x_3 x_4] f$

$[x_0 x_1 x_2 x_3 x_4] f$

$$P_4(x) = \underbrace{[x_0] f 1}_{\omega_0(x)} + \underbrace{[x_0 x_1] f (x - x_0)}_{\omega_1(x)} \; \underbrace{[x_0 x_1 x_2] f (x - x_0)(x - x_1)}_{\omega_2(x)}$$

$$+ \underbrace{[x_0 x_1 x_2 x_3] f (x - x_0)(x - x_1)(x - x_2)}_{\omega_3(x)} + \underbrace{[x_0 x_1 x_2 x_3 x_4] f (x - x_0)(x - x_1)(x - x_2)(x - x_3)}_{\omega_4(x)}$$

Beispiel

	x	f
x_0	0	1
x_1	$3/2$	2
x_2	$5/2$	2
x_3	$9/2$	3

0	1			
		$\frac{2-1}{3/2-0} = \frac{2}{3}$		
$3/2$	2		$\frac{0-\frac{2}{3}}{5/2-0} = -\frac{4}{15}$	
		0		$1/45$
$5/2$	2		$= \frac{1}{6}$	
		$-\frac{1}{3}$		
$9/2$	3			

$$\rightarrow P_3(x) = \underbrace{1}_{\omega_0(x)} + \frac{2}{3}\underbrace{(x-0)}_{\omega_1(x)} + \left(-\frac{4}{15}\right)\underbrace{(x-0)(x-3/2)}_{\omega_2(x)} + \frac{1}{45}\underbrace{(x-0)(x-3/2)(x-5/2)}_{\omega_3(x)}$$

Polynom-Interpolation:

$I=[a,b]$, $a \le x_0 \le \dots \le x_n \le b$, x_i: $(n+1)$-Knotenstellen pw. verschieden. Gegeben f_k mit $f_k \approx f(x_k)$, $k=0,\dots,n$

Gesucht: $p_n \in \Pi_n$ ($\deg p_n \le n$) mit $p_n(x_k)=f(x_k)$ $\forall k=0,\dots,n$

Lagrange-Basen $x \mapsto l_k$, $l_k(x_j)=\begin{cases}1 & x_j=x_k\\ 0 & x_j \ne x_k\end{cases}$ $j,k=0,\dots,n$

$$l_k(x)=\prod_{\substack{j=0\\ j\ne k}}^{n} \frac{x-x_j}{x_k-x_j}, \quad p_n(x)=\sum_{k=0}^{n} f_k\, l_k(x)$$

$\Rightarrow p_n(x_j)=\sum_{k=0}^{n} f_k\, l_k(x_j)=f_j\,\delta_{kj}=f_j$, $j=0,\dots,n$ Lagrange-Darstellung

Fehlerabschätzung zur Polynomapproximation

Satz Sei $f \in C^{n+1}(I)$, $x \in I$ bel

Sei $p_n=P(f(x_0,\dots,x_n))$ das Interpolationspolynom zu $f_k=f(x_k)$ $k=0,\dots,n$, dann $\exists\, \xi \in \operatorname{conv}\{x_0,\dots,x_n\} \subseteq I$

$$|f(x)-p_n(x)| \le \frac{f^{n+1}(\xi)}{(n+1)!} \cdot \prod_{k=0}^{n}(x-x_k)$$

Beweis Sei $x' \in I$ bel. und fest und $\alpha \in \mathbb{R}$, sodass

$$F(x)=f(x)-p(x)-\alpha\prod_{k=0}^{n}(x-x_k) \quad F(x')=0$$

d.h. $\alpha=\dfrac{f(x')-p_n(x')}{\prod_{k=0}^{n}(x-x_k)}$

$\Rightarrow F \in C^{n+1}(I)$ und $F(x_k)=\underbrace{f(x_k)-p(x_k)}_{=0}-\alpha\prod_{k=0}^{n}(x_k-x_j)=0$

für $k=0,\dots,n$, $F(x')=0$

Falls $x'<x_k$ für ein k, ist die Aussage trivial $(0=0)$

$\Rightarrow F$ hat mindestens $n+2$ Nullstellen

Anwendung des Lemma von Rolle $\Rightarrow$ es ex. eine Zwischenstelle $\xi \in \operatorname{conv}\{x_0,\dots,x_n\}$ (konvexe Hülle) mit $F^{(n+1)}(\xi)=0=f^{(n+1)}(\xi)+\underbrace{p_n^{(n+1)}(\xi)}_{=0}-\alpha(n+1)!$

$\Rightarrow \alpha=\dfrac{f^{(n+1)}(\xi)}{(n+1)!}$

Grenzen der Polynominterpolation: Bsp: $I=[-5,5]$, $f(x)=\frac{1}{1+x^2} \in C^\infty(\mathbb{R})$

$x_k=-5+\frac{10k}{n}$ $k=0,\dots,n$ (äquidistant)

$p_n(x)$ divergieren in der Nähe des Randes für $n\to\infty$

Folgerung Äquidistante Knotenwahl ist für große n i.A. ungünstig.

Tschebychev-Knoten:

$$x_k = \frac{a+b}{2} + \frac{b-a}{2}\cos\left(\frac{2k+1}{2n+1}\pi\right)$$

ohne Beweis: Seien x_k T.-Knoten und $\tilde{x}_k$ bel. andere Knoten $k=0,\dots,n$

Dann gilt: $\prod_{k=0}^{n} |x-x_k| \le \prod_{k=0}^{n} |x-\tilde{x}_k|$. T-Knoten liefern quasi eine optimale Knoten und es gilt $\|f-p_n\|_\infty \le C_n \cdot \inf_{q_n\in\Pi_n} \|f-q_n\|_\infty$ mit

~~$\inf_{q_n\in\Pi_n} \|f-q_n\|_\infty$, mit~~ $C_n = 1 + \sum_{j=0}^{n} \left\| \prod_{\substack{k=0\\ j\neq k}} \left(\frac{x-x_k}{x_j-x_k}\right) \right\|_\infty$

3.5. Trigonometrische Interpolation

3.5. Diskrete Fourier-Transform.

$f: \mathbb{R} \to \mathbb{C}$ 1-periodisch (d.h. $f(x+k)=f(x)\ \forall x\in\mathbb{R}, k\in\mathbb{Z}$)

$x \mapsto \varphi_k(x) = c_k e^{ikx2\pi}$, $k\in\mathbb{Z}$ (für $\varphi\in\mathbb{R}$ gilt $e^{i\varphi}=\cos\varphi + i\sin\varphi$)

$= c_k (e^{ix2\pi})^k$

Interpolationsaufgabe geg: $x_0,\dots,x_{n-1}$ $(x_n = x_0) \in [0,1)$

ges: $p_n(x) = \sum_{k=0}^{n-1} \alpha_k \varphi_k(x) = \sum_{k=0}^{n} \alpha_k e^{2\pi i k x}$ mit $p_n(x_j) f_j$, $j=0,\dots,n-1$

$f_i = f(x_j)$ d.h. gesucht sind $\alpha_k \in \mathbb{C}$, $k=0,\dots,n-1$

Seien $x_k = \frac{k}{n}$, $\alpha_0,\dots,\alpha_{n-1}$ lösen n Unbekannte, n-Gleichungen

$$\sum_{j=0}^{n-1} \alpha_j e^{2\pi i j \frac{k}{n}} = f_k,\ \forall k=0,\dots,n-1 \quad (*)$$

Sei $\omega_n = e^{2\pi i \frac{1}{n}}$, für $k=0,\dots,n-1$, ω_n^k n-te Einheitswurzel

$\omega_n^0 = 1$, $\omega_n^k = 1$

Lemma: Es gilt für $j=1,\dots,n-1$

$$\sum_{k=0}^{n-1} (\omega_n^j)^k = 0; \quad \text{Für } j=0: \sum_{k=0}^{n-1} (\omega_n^0)^k = \sum_{k=0}^{n-1} 1 = n$$

Beweis $j=0$ siehe oben

$j\neq 0$: Sei $q = \omega_n^j$ $(\Rightarrow |q| = 1)$

$$\sum_{k=0}^{n-1} \omega_n^{jk} = \sum_{k=0}^{n-1} q^k \underset{\text{geom. Reihe}}{=} \frac{1-q^n}{1-q} = \frac{1-(\omega_n^j)^n}{1-\omega_n^j} = \frac{1-e^{2\pi i \frac{jn}{n}}}{1-\omega_n^j} = \frac{1-1}{1-\omega_n^j} = 0$$

Theorem: Die LSg des Interpolations-Problems $(*)$ lautet

$$\alpha_j = \frac{1}{n}\sum_{k=0}^{n-1} f_k e^{-2\pi i \frac{jk}{n}} = (\underline{E}\,\underline{f})(j),\ \underline{E} = \frac{1}{n}\left(e^{-2\pi i \frac{jk}{n}}\right)_{j,k=0,\dots,n-1}\ \underline{f} = (f_k)_{k=0,\dots,n-1}$$

Bemerkung 1) Es gilt $\overline{\underline{E}} = \underline{E}^H = (\overline{E}_{j,k})^T$ (Hermitesch konj: $\underline{F}^H = \underline{E}^{-1} = \left(e^{2\pi i \frac{jk}{n}}\right)_{j,k=0,\dots,n-1}$ d.h. $\sqrt{n}\,F$ ist unitär

2) $p_n(x_k) = \sum_{j=0}^{n} \alpha_j e^{2\pi i \frac{jk}{n}} = \sum_{j=0}^{n-1} \frac{1}{n} \sum_{l=0}^{n-1} f_l e^{2\pi i \frac{j(k-l)}{n}}$

$= \sum_{l=1}^{n-1} f_l \left(\frac{1}{n}\sum_{j=0}^{n-1} e^{2\pi i \frac{j(k-l)}{n}}\right) = \sum_{l=0}^{n-1} f_l \delta_{k,l} \underset{\text{Lemma 5.3.1}}{=} f_k \quad \forall k=0,\dots,n-1$

Technische Universität Berlin
Fakultät II – Mathematik und Naturwissenschaften
Institut für Mathematik
Reinhold Schneider, Benjamin Kutschan

Wintersemester 2016/2017

Numerische Mathematik I

FFT-Lemma

Die Fouriertransformation

$$DFT : (f_0, ..., f_{n-1}) \mapsto (d_0, ..., d_{n-1})$$

ist definiert durch

$$d_j = \frac{1}{n} \sum_{k=0}^{n-1} f_k e^{-2\pi i \frac{k}{n} j}$$

Es fällt auf, dass bei geradem j und geradem n der Term

$$e^{-2\pi i \frac{k}{n} j}$$

sich eine gerade Anzahl an Umdrehungen in der komplexen Ebene um den Nullpunkt dreht. Definiere $n = 2m$ und $j = 2l$. Genauer gesagt gilt dann

$$e^{-2\pi i \frac{k+m}{n} j} = e^{-2\pi i \frac{k+m}{2m} 2l} = e^{-2\pi i \frac{k+m}{m} l} = e^{-2\pi i \frac{k}{m} l} e^{-2\pi i l} = e^{-2\pi i \frac{k}{m} l} = e^{-2\pi i \frac{k}{n} j}.$$

Daraus ergibt sich das folgende Lemma:

Lemma 1. *Sei $n = 2m$ und $DFT(f_0, ..., f_{n-1}) = (d_0, ..., d_{n-1})$. Dann gilt*

$$d_{2l} = \frac{1}{m} \sum_{k=0}^{m-1} \frac{1}{2}(f_k + f_{k+m}) e^{-2\pi i \frac{lk}{m}} = DFT(\frac{1}{2}(f_0 + f_m), ..., \frac{1}{2}(f_{m-1} + f_{2m-1}))$$

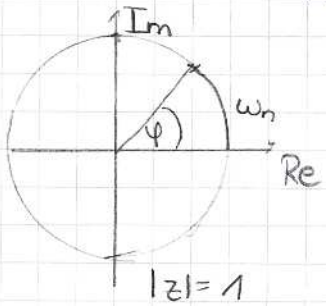
bzw.

$$d_{2l+1} = \frac{1}{m} \sum_{k=0}^{m-1} \frac{1}{2}(f_k - f_{k+m}) e^{-2\pi i \frac{k}{n}} e^{-2\pi i \frac{lk}{m}} = DFT(\frac{1}{2}(f_0 - f_m) e^{-2\pi i \frac{0}{n}}, ..., \frac{1}{2}(f_{m-1} - f_{2m-1}) e^{-2\pi i \frac{m-1}{n}})$$

Beachte, dass die DFTs auf der rechten Seite nur DFTs auf Vektoren von halber Länge sind. Wir können nun also eine DFT der Länge $2m$ berechnen, indem wir zwei DFTs der Länge m berechnen und dann geeignet zusammenfügen.

Schnelle FT (Fouriertransformation)

$\omega_n = e^{-2\pi i \frac{1}{n}} = e^{-i\varphi}$ $\quad \varphi = \frac{2\pi}{n}$



$f = (f_k)_{k=0}^{n-1}$, $x_k = 2\pi \frac{k}{n}$, $k = 0, \dots, n-1$

$\underline{d} = (d_k)_{k=0}^{n-1}$

$\sum_{j=0}^{n-1} d_j e^{2\pi i \frac{jk}{n}} = f_k, \ \forall k = 0, \dots, n-1$

$\omega_n^n = (e^{-\frac{2\pi i}{n}})^n = e^{-2\pi i \frac{n}{n}} = 1$

$\sum_{j=0}^{n-1} d_j \omega_n^{-jk} = f_k$ $\quad$ f gegeben, dann

$\Rightarrow d_j = \frac{1}{n} \sum_{k=0}^{n-1} f_k \omega_n^{jk}$, $\underline{d} = \underline{\underline{F}}\,\underline{f}$, $\underline{\underline{E}} = \frac{1}{n} (\omega_n^{jk})_{j,k=0}^{n-1}$

3.5.2

Lemma: Sei $n = 2m$, $\omega_m = e^{-2\pi i \frac{2}{n}} = \omega_n^2$ und $\ell = 0, \dots, m-1$

dann gelten $d_{2\ell} = \frac{1}{m} \sum_{k=0}^{m-1} \frac{1}{2}(f_k + f_{k+m}) \omega_m^{\ell k}$ (1)

$e^{2\pi i} = 1$ $\quad d_{2\ell+1} = \frac{1}{m} \sum_{k=0}^{m-1} \frac{1}{2}(f_k - \omega_n^k f_{k+m}) \omega_m^{\ell k}$ (2)

Beweis 1) $j = 2\ell$

$(\omega_n^2)^{\ell k} = \omega_n^{2\ell(k+m)} = (\omega_n^2)^{\ell \cdot k} \omega_n^{2\ell m}$

$= \omega_m^{\ell k} \underbrace{\omega_m^{\ell m}}_{=1} = \omega_m^{\ell k}$

$\omega_n^2 = \omega_m$, $\omega_m^{\ell m} = (\omega_m^m)^\ell = 1^\ell = 1$

2) $j = 2\ell + 1$ $\quad \omega_n^{(2\ell+1)(k+m)} = \omega_n^{(2\ell k + 2\ell m + k + m)}$

$= \omega_m^{\ell k} \underbrace{\omega_n^{\ell m}}_{=1} \omega_n^k \omega_n^m$

$= \omega_m^{\ell k} \cdot \omega_n^k (-1)$ (Dies lässt sich rekursiv fortsetzen)

$\omega_n^m = e^{-2\pi i \frac{m}{2m}} = e^{-\pi i} = -1$

Lemma Sei $n = 2^p$, $n \in \mathbb{N}$ (2-er Potenz)

Dann gilt für den Aufwand (#arithmetisch. Operationen)

$T(n) = T(2^p) \leq p 2^{p+1} = 2p 2^p = 2n \log_2 n$

Beweis: In jedem Schritt (Lemma 3.5.2) gilt

$T(n) \leq 2n + 2T(\frac{n}{2})$

vollständige Induktion: $p = 1$ $\quad T(2) \leq 2 + 2 = 4 \leq 1 \cdot 2^2 = 4$

$p \Rightarrow p+1$

$n = 2^{p+1}$ $\quad T(2^{p+1}) \leq 2 \cdot 2^{p+1} + 2(p 2^p) = 2(1+p) 2^{p+1}$ ∎

4. Numerische Integration

4.1 Aufgabenstellung

Intervall $[a,b]$, Fktn $f:[a,b]\to\mathbb{R}$ intbar

$I(f)=\int_a^b f(x)dx \approx I_h(f)$ Quadraturformel (*)

approximieren: $I_h(f)=\sum_{k=0}^{n} f(x_k)\omega_k$

$x_k\in[a,b]$ Knotenstellen (pw. verschieden)

Gewichte: $\omega_k\in\mathbb{R}$ ((ω_k, x_k) bestimmen die Quadraturformel

Idee + Bsp: Sei $0=t_0<t_1<t_2<\dots<t_m=b$

$\Rightarrow [a,b]=\underbrace{\bigcup_{j=0}^{m-1}[t_j,t_{j+1}]}_{\text{Zerlegung von } [a,b]}$; und $I(f)=\int_a^b f(x)dx=\sum_{j=0}^{m-1}\int_{t_j}^{t_{j+1}} f(x)dx$

Seien $t_{j+1}-t_j=:h$, äquidistant (Bsp)

1) Rechteckregel: $I_h(f)=\sum_{j=0}^{n-1} hf(t_j)$

2) Trapezregel auf $[t_j,t_{j+1}]$

$\tilde{I}_h(f)=h\,\frac{1}{2}(f(t_j)+f(t_{j+1}))$

zusammengesetzt $I_h(f)=\frac{h}{2}(f(a)+f(b))+h\sum_{j=1}^{m-1} f(t_j)$

3) Mittelpunktregel

$\int_{t_j}^{t_{j+1}} f(x)dx\approx \tilde{I}_h(f)=hf\left(\frac{t_j+t_{j+1}}{2}\right)$

4.2 Newton-Cotes-Formeln

Intervall $[c,d]:=[t_j,t_{j+1}]$ Teilintervall

Quadraturknoten $x_0,\dots,x_n\in[c,d]$ gewählt (Newton-Cotes $\hat{=}$ äquidistant)

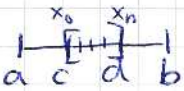
Seien (Stützwerte) $f_k:=f(x_k)$ und $x\mapsto p_n(x)=p_n(f|x_0,\dots,x_n)(x)\in\Pi_n$

das zugehörige Interpolationspolynom

$I(f)=\int_c^d f(x)dx\approx\int_c^d p_n(f|x_0,\dots,x_n)(x)dx=\int_c^d p(x)dx=I_h(f)$

Definition 4.1. Eine Q-Formel (*) heißt exakt von der Ordnung d, falls $\int_c^d p(x)dx=I_h(f)\ \forall p\in\Pi_d$ Polynom vom Grad $\leq d$

Numerische Integration



$$\int_c^d f(t)\,dt = I(f) \quad h := [c,d]$$

$x_0 = c < x_1 < \dots < x_n = d$, Knoten, Fkt-werte $f(x_k) = f_k$

z.B. Newton Cotes Regeln, $x_k = kh + c = \frac{k(d-c)}{n} + c$

„äquidistante Knoten"

$$I_h(f) := \int_c^d p_n(x)\,dx, \quad p_n(x) = P_n(x_0, \dots, x_n | f)(x) \in \Pi_n$$

$$l_k(x) = \prod_{\substack{j=0 \\ j \neq k}}^{n} \frac{x - x_j}{x_k - x_j}, \text{ Lagrange-Polynome}$$

$$\Rightarrow p_n(x) = \sum_{k=0}^{n} f(x_k)\, l_k(x) \Rightarrow I_h(f) = \int_c^d p_n(x)\,dx = \sum_{k=0}^{n} f(x_k) \underbrace{\int_c^d l_k(x)\,dx}_{=: \omega_k}$$

$$= \sum_{k=0}^{n} f(x_k)\,\omega_k$$

Satz 4.2. $I_h(f) = \sum_{k=0}^{n} f(x_k)\,\omega_k$, mit $\omega_k = \int_c^d \left(\prod_{\substack{j=0 \\ j \neq k}}^{n} \frac{x - x_j}{x_k - x_j} \right) dx$

Bemerkung Die Wahl der Knoten bestimmt die Quadraturgewichte.

Satz 4.1 Es gilt $\left| \int_c^d f(x)\,dx - I_h(f) \right| \leq \frac{h}{(n+1)!} \|f^{(n+1)}\|_\infty \max_{x \in [c,d]} \prod_{j=0}^{n} \underbrace{|x - x_j|}_{\leq h}$

für $f \in C^{n+1}([c,d])$ die Fehlerabschätzung

$$\leq \frac{h^{n+2}}{(n+1)!} \|f^{(n+1)}\|_\infty$$

Beweis $\left| \int_c^d (f - p_n)(x)\,dx \right| \leq \int_c^d |f(x) - p_n(x)|\,dx \leq h \|f - p\|_\infty$

$$\underset{\text{Interpolationsabschätzung}}{\leq} h \frac{|f^{(n+1)}(\xi)|}{(n+1)!} \prod_{j=0}^{n} |x - x_j| \leq h \frac{\|f^{(n+1)}\|_\infty}{(n+1)!} \prod_{j=0}^{n} |x - x_j|$$

$$\leq \frac{h^{n+2}}{(n+1)!} \|f^{(n+1)}\|_\infty$$

Bsp zu Newton Cotes, wähle $c = 0$, $d = 1$, ($[0,1]$, $h = 1$)

n	ω_k	Fehlerabsch.
1	$1/2 \quad 1/2$ „Trapezregel"	$1/12\, h^3 f^{(2)}(\xi)$, ξ - Zwischenstelle
2	$1/6 \quad 4/6 \quad 1/6$ „Simpsonregel"	$1/90\, 2^{-5} h^5 f^{(4)}(\xi)$
3	$1/8 \quad 3/8 \quad 3/8 \quad 1/8$ „$\frac{3}{8}$-Regel"	$3/30\, 3^{-5} h^5 f^{(4)}(\xi)$
4	$\frac{7}{90} \quad \frac{32}{90} \quad \frac{12}{90} \quad \frac{32}{90} \quad \frac{7}{90}$ „Milne-Regel"	$\frac{8}{945}\, 4^{-7} h^7 f^{(6)}(\xi)$

Bemerkung: Für N-C-Regeln der Ordnung $n \geq 7$ treten auch neg. Qu.gewichte ($\omega_k < 0$) auf!

$\Rightarrow$ Regeln sind weniger stabil und werden i.A. nicht verwendet

4.3. Romberg-Quadratur und Extrapolation

Zusammengesetzte Trapezregel $h := \frac{b-a}{n}$

$$\int_a^b f(x)dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f\,dx \approx \frac{h}{2}(f(a)+f(b)) + h\sum_{i=1}^{n-1} f(x_i) =: T(h)$$

Satz 4.3.1

$$\left| \int_a^b f(x)dx - T(h) + \sum_{k=1}^{m} \tau_{2k} h^{2k} \right| \leq c_{2m+2} |b-a| h^{m+2}$$

mit $\tau_{2k} = \frac{B_{2k}}{(2k)!}(f^{(2k+1)}(b) - f^{(2k+1)}(a))$

B_{2k} „Bernoulli Zahlen", $b_{2k} = \frac{B_{2k}}{(2k)!}$

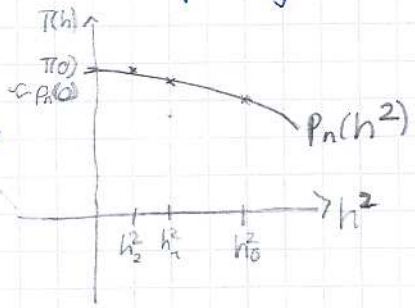
$g(t) = -1 + \frac{1}{2}t + \frac{t}{e^t - 1} = \sum_{k=1}^{\infty} b_{2k} t^{2k}$ „Potenzreihe"

Beweis: „Eulersche Summenformel" („Literatur")

Bemerkung: 1) Falls f periodisch ist (mit $T = b-a$) und analytisch. Dann konvergiert die Trapezregel exponentiell.

2) $c_i := \frac{1}{\sqrt{T}} \int_0^T f(x) e^{i\frac{2\pi}{T}x} dx \approx T(h)$ „Trapezregel $\hat{=}$ diskreter FT

Trick „Extrapolation"



Es gilt $\int_a^b f(x)dx = \lim_{h \to 0} T(h)$;

Sei $h := h_0 > h_1 \dots > h_n$

Bsp: Romberg $h_k = h_0 2^{-k}$ ($h_{k+1} = \frac{h_k}{2}$ ...)

Wir interpolieren $T(h) \approx p_n(h^2)$, $p_n(x) = p_n(h_0, \dots, h_n, T(x))$, $x = h^2$

Definition 4.2 Sei $h \to T(h)$, $h \in \mathbb{R}_+$ eine Funktion, diese besitze eine „asymptotische Entwicklung" für $h \to 0$ in h^p der Ordnung $p \cdot m$, falls Konstanten $\tau_0, \dots, \tau_m \in \mathbb{R}$ ex. mit

$$|T(h) - (\tau_0 + \tau_{1p} h^p + \tau_{2p} h^{2p} + \dots + \tau_{mp} h^{mp})| = O(h^{(m+1)p}), h \to 0$$

d.h. $T(h) = \sum_{k=1}^{m} \tau_{kp} h^{kp} + \tau_0 + O(h^{(m+1)p}), h \to 0$.

Aus Satz 4.3.1 folgt, dass die Trapezregel $T(h)$ eine „asy. Entw." mit $p = 2$ besitzt. ($\tau_0 := \int_a^b f\,dx$)

Korollar (Extrapolation). (p_n - Extrapolationspolynom ($p=2$)

$$\left| \int_a^b f(x)dx - p_n(0) \right| \leq C h^{(m+1)2} \|f^{(m+1)2}\|_\infty$$

4.4 Orthogonale Polynome und Gauß-Quadratur

4.4.1. Orthogonale Polynome

Lemma 4.5 Sei $x \mapsto \omega(x) > 0$, $x \in \Omega$ und f, g messbar, $d\mu = \omega(x)dx$

dann definiert $\int_\Omega f(x)\overline{g(x)}\,\omega(x)\,dx = \langle f, g\rangle_\omega$ ein inneres Produkt

im ~~Prod~~ Hilbertraum $L_2(\Omega, \omega) \overset{ae}{=} \{f : \|f\|^2_{\Omega,\omega} = \langle f, f\rangle_\omega\}$ ae. almost everywhere

Lemma 4.6 Sei $\overline{\Pi}_j = \{p \text{ Polynom } p(x) = x^j + a_{j-1}x^{j-1} \dots a_0,\ a_j \in \mathbb{K}\} \subset \Pi_j$

Dann ex. genau ein $p \in \overline{\Pi}_j$ mit $\langle p_j, q\rangle_\omega = 0 \ \forall q \in \Pi_{j-1}$ $(p_j \perp \Pi_{j-1})$

Beweis: per Induktion

IA: $j = 0$. $p_0 = 1$

IS: Seien $p_0, \dots, p_{j-1} \in \Pi_i, i = 0, \dots, j-1$ ~~Seien~~ gegeben, $\Rightarrow \{(x \mapsto x^j), p_{j-1}, \dots, p_0\}$ sind lin unab.

(Dimension$(\Pi_j) = j+1$); es gilt $\langle p_i, p_k\rangle_\omega = 0 \ i \neq k, i, k < j$

$\overset{x \mapsto}{\Rightarrow} p_j(x) := x^j - \sum_{k=0}^{j-1} p_k(x) \frac{\langle p_k, x^j\rangle_\omega}{\langle p_k, p_k\rangle}$ „Gram Schmidt"

$\Rightarrow p_j \in \Pi_j : \langle p_j, p_k\rangle_\omega = 0, \forall k = 0, \dots, j-1$

$\Rightarrow \langle p_j, q\rangle_\omega = 0, \forall q \in \Pi_{j-1}$ ∎

Bemerkung Beweis ist konstruktiv. $p_j, j \in \mathbb{N}_0$ „Orthogonale Polynome"

bzgl. Gewicht ω

Beispiel 1) $\Omega = [-1, 1]$, $\omega = 1$, Legendre-Polynome

$p_j(x) = \frac{j!}{(2j)!} \frac{d^j}{dx^j}((x^2 - 1)^j), j \in \mathbb{N}_0$

2) Tschebyscheff knoten $\Omega(-1, 1)$

$\omega(x) (\sqrt{1 - x^2})^{-1} = \frac{1}{\sqrt{1-x^2}}$

$p_j(x) = T_j(x) = \cos(j \arccos x)$ $j \in \mathbb{N}_0$

3) Laguerre-Polynome $\Omega = [0, \infty)$, $\omega(t) = e^{-t}$

4) „Hermite-Polynome" $\Omega = \mathbb{R}$, $\omega(t) = e^{-t^2}$ (Stochastik (Statistik))

Lemma 4.8 Sei p_j gemäß oben $x \mapsto p_j(x)$, $y \in [a, b] \in \Omega$, $j \geq 1$

dann ex. genau j einfache Nst $x_i \in [a, b]$, $(a \leq x_1 < \dots < x_j \leq b)$

Beweis Sei $j \geq 1$. Seien $a \leq x_0 < \dots < x_k \leq b$ alle Nst von p_j

Da $\int_a^b p_j(x)\omega(x)dx = \langle p_j, 1\rangle_\omega = \langle p_j, p_0\rangle_\omega = 0 \Rightarrow k \geq 1$ (!!!)

Annahme $1 \leq k < j$

Sei $q(x) = \prod_{i=1}^{k} (x - x_i) \Rightarrow q \in \Pi_k$, $(k<j)$ (*)

$\Rightarrow x \mapsto p_j(x) q(x) = r(x) \underbrace{\prod_{i=0}^{k} (x - x_i)^2}_{\geq 0}$, $r(x) \neq 0 \; \forall x \in [a,b]$

d.h. Vorzeichensign $p_j(x) q(x) = \text{const}$, $\forall x \in [a,b]$

$$0 < \int_a^b \underbrace{p_j(x) q(x) w(x)}_{>0} dx = \langle p_j, q \rangle_w \overset{(*)}{=} 0 \quad \text{z.B.} \; \lightning$$

$\Rightarrow k = j$ ∎

<u>Lemma 4.9.</u> Sei $t_k \in [a,b]$, $k = 1, \dots, j$ $a \leq t_1 < t_2 < \dots < t_j \leq b$

$p_j \in \Pi_j$ (wie oben) $j \in \mathbb{N}_0$ und $A = (a_{i,j})_{i,j=1}^{n} = (p_{i-1}(t_j))_{i,j=1}^{n}$

$\Rightarrow A \in \mathbb{R}^{j \times j}$ ist regulär

<u>Beweis</u> Annahme $\exists t_k$ mit A singulär d.h. $0 \neq u \in \ker(A^T)$

$A^T u = 0$ bzw. $\sum_{k=1}^{j} p_{k-1}(t_i) u_k = 0 \; \forall i = 0, \dots, j-1$

$\Rightarrow q(x) = \sum_{k=1}^{j} p_{k-1}(x) u_k \in \Pi_{j-1}$

und $q(t_i) = \sum_{k=1}^{j} p_{k-1}(t_i) u_k = 0 \; \forall i = 1, \dots, j$

$\Rightarrow q \in \Pi_{j-1}$ hat j verschiedene Nst. $\Rightarrow q \equiv 0 \Rightarrow u_k = 0 \; \forall k$ ∎

4.4.2 Gauß-Quadratur

Sei $a \leq x_1 < \dots < x_j \leq b$ die Nullstellen von p_j, die Knotenpunkte für die Gaußquadratur $\int_a^b f(x) w(x) dx := \int_a^b p_j(x_1, \dots, x_j; f)$

<u>Satz 4.10</u> Sei $v = (v_1, \dots, v_j)^T \in \mathbb{R}^j$ die Lösung von

$A v = e_1 \langle p_0, p_0 \rangle_w$, $e_1 = (1, 0, \dots, 0) \in \mathbb{R}$

$a_{i,j} = p_{k-1}(t_i)$, d.h. $\sum_{k=1}^{j} p_{i-1}(t_k) v_k = \delta_{0,i} \langle p_0, p_0 \rangle_w$, $i = 0, \dots, j-1$ (*)

i) $w_k = v_k > 0$

ii) Sei $p \in \Pi_{2j-1} \Rightarrow \int_a^b p(x) w(x) dx = \sum_{k=1}^{j} p(x_k) v_k$, Gauß-Quad. ist korrekt $\forall p \in \Pi_{2j-1}$ obwohl nur j Punkte gegeben sind ↯

iii) Es existieren keine j Punkte $0 \leq t_1 < \dots < t_j \leq b$ mit und keine w_k, $k = 1, \dots, j$

$\int_a^b p(x) w(x) dx = \sum_{k=1}^{j} p(x_k) w_k \; \forall p \in \Pi_{2j}$ $\Rightarrow$ gilt nicht bei Gauß-Quad.

$\left(\overset{\text{Lemma 4.9.}}{\Rightarrow} \exists! \, v \in \mathbb{R}^j \right)$

Gauß-Quadratur $\langle \cdot,\cdot \rangle_w$, $w(x)>0$

$\langle f,g\rangle_w := \int_I f(x)g(x)w(x)dx$, $p_0,\dots,p_j \in \Pi_j$ orthogonale Polynome

zu p_j, Nullstellen $a \le x_1 < \dots < x_n \le b \Rightarrow$ Quadraturknoten der Gauß-Quadratur:

Seien $A=(a_{i,k})=(p_{i-1}(x_k))_{i,k=1,\dots,n} \in \mathbb{R}^{n\times n}$

$\Rightarrow$ A regulär und $w=(w_1,\dots,w_j)^T \in \mathbb{R}^j$ Lösung von

$Aw = \langle p_0,p_0\rangle_w (1,0,0,\dots,0)^T$

w sind Quadraturgewichte (Behauptung)

j-Quadraturknoten, aber Approximationsordnung ist $2j$

<u>Beweis</u>: ii) Sei $p \in \Pi_{2j-1}$ bel. $\Rightarrow \exists q,r \in \Pi_{j-1}$ mit $r(x)=\sum_{i=0}^{j-1}\alpha_i p_i(x)$

$p(x)=p_j(x)q(x)+r(x)\ \forall x\in\mathbb{R}$ $\qquad \alpha_i = \frac{\langle p_i,r\rangle_w}{\langle p_i,p_i\rangle_w}$

a) $\int_a^b p(x)w(x)dx = \int_a^b q(x)p_j(x)w(x)dx + \int_a^b r(x)w(x)dx$

$= \underbrace{\langle \overset{\in\Pi_{j-1}}{q}, p_j\rangle_w}_{=0} + \sum_{i=0}^{j-1}\alpha_i\langle p_i,p_0\rangle_w = 0 + \alpha_0\langle p_0,p_0\rangle_w = \alpha_0\int_a^b w(x)dx$

b) $\sum_{k=1}^{j} p(x_k)w_k = \underbrace{\sum_{k=1}^{j} q(x_k)\underbrace{p_j(x_k)}_{=0}w_k}_{=0} + \sum_{k=1}^{j} r(x_k)w_k$

$= \sum_{i=0}^{j-1}\sum_{k=1}^{j}\alpha_i \overbrace{p_i(x_k)w_k}^{=Aw} = \sum_{i=0}^{j-1}\alpha_i\,\delta_{i,0}\langle p_0,p_0\rangle_w = \alpha_1\langle p_0,p_0\rangle_w \Rightarrow$ ii)

i) Behauptung $w_i>0\ \forall i=1,\dots,j$.

Sei $x\mapsto q(x) := \prod_{\substack{i=1\\ i\neq k}}^{j}(x-x_i)^2 \ge 0$, $q\in\Pi^{2j-2}$

$0 < \int_a^b q(x)w(x)dx = \sum_{i=1}^{j} q(x_i)w_i = \prod_{\substack{i=1\\ i\neq k}}^{j}(x_k-x_i)^2 w_{i,r} \Rightarrow w_{i,r}>0 \Rightarrow$ ii)

iii) $\exists\, p\in\Pi_{2j}$ mit $\int_a^b p(x)w(x)dx \neq \sum_{k=1}^{j} p(x_k)w_k$

Sei $p(x)=\prod_{i=1}^{j}(x-x_i)^2 \ge 0$, $p\in\Pi_{2j} \Rightarrow 0 < \int_a^b p(x)w(x)dx$

aber $\neq 0 = \sum_{k=0}^{j}\overbrace{p(x_k)}^{=0}w_k$

<u>Satz 4.11</u> p_j, $j\in\mathbb{N}$, $p_j(x_k)=0$ usw gemäß oben

Sei $f\in C^0([a,b])$: $\int_a^b f(x)w(x)dx = \lim_{j\to\infty}\sum_{k=1}^{j} f(x_k)w_k$

<u>Beweis</u> Approximationssatz von Weierstraß: $\forall\varepsilon>0\ \exists j\in\mathbb{N}$ mit $\|f-p_j\|_\infty<\infty$

$|I(f)-I_j(f)| \le |I(f)-I(p_j)| + |I(p_j)-I_j(f)| \le \int_a^b|(f-p)(x)|w(x)dx + \sum_{k=1}^{j}|p_j(x_k)-f(x_k)|w_k$

$\le \|f-p_j\|_\infty\int_a^b w(x)dx + \|f-p_j\|_\infty\sum_{k=1}^{j}w_k = 2\|f-p_j\|_\infty\langle p_0,p_0\rangle_w = 2\langle p_0,p_0\rangle_w\,\varepsilon$

$\xrightarrow{\varepsilon\to 0} 0$ ∎

Satz 4.12 $f \in C^{(2n)}([a,b])$, dann ex. $\xi \in [a,b]$ mit

$$\left| \int_a^b f(x) w(x)\,dx - \sum_{k=1}^{n} f(x_k) w_k \right| \le \frac{|f^{(2n)}(\xi)|}{(2n)!} \langle p_n, p_n \rangle_w$$

Beweis: Hermite-Interpolation:

2n Knoten $a \le \underset{\text{Funktion}}{x_1} = \underset{\text{Ableitung}}{x_1} < \dots < x_n = x_n < b$

$x \in [a,b]$: $f(x) = p(x) + [x, x_1, x_1, \dots, x_n, x_n] f \underbrace{\prod_{i=1}^{j} (x - x_j)^2}_{=(p_n(x))^2}$ $\quad (p_n(x_i) = 0, i = 1, \dots, n$

$$\left| \int_a^b f(x) w(x)\,dx - \sum_{k=1}^{n} f(x_k) w_k \right| \le \left| \int_a^b [x, x_1, x_1, \dots, x_n, x_n] f (p_n(x))^2 w x\,dx \right|$$

$$\le \frac{\|f^{(2n)}\|_\infty}{(2n)!} \int_a^b p_n^2(x) w(x)\,dx = \frac{\|f^{(2n)}\|}{(2n)!} \langle p_n, p_n \rangle_w \quad \blacksquare$$

Bemerkung 1) Man kann zeigen, dass für analytische f die Gauß-Quadratur exponentiell konvergiert, d.h. $\exists \beta > 0$: Fehler $\le c e^{-\beta j}$

Probleme: 1) uneigentliche Integrale bzw. unbeschränkte Integranden, „gradwise Knotenwahl" 2) Problem: nochoziellierende Funktion

5. Numerische Lösung nichtlinearer GLS

5.1. Aufgaben: Geg $D \subset \mathbb{R}^n$, $F: D \to \mathbb{R}^n$

$x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$, $F(x) = (F_1(x_1, \dots, x_n), \dots, F_n(x_1, \dots, x_n))^T \in \mathbb{R}^n$

Gesucht: Nullstellen $x^* \in D$, d.h. $F(x^*) = 0 \in \mathbb{R}^n$ $\begin{pmatrix} F_1(x_1^*, \dots, x_n^*) = 0 \\ \vdots \\ F_n(x_1^*, \dots, x_n^*) = 0 \end{pmatrix}$

Beispiel: $n = 2$, $D = \mathbb{R}^2$

1) $\left.\begin{aligned} F_1(x_1, x_2) &= x_1^2 - x_2 + 0{,}25 \overset{!}{=} 0 \\ F_2(x_1, x_2) &= -x_1 + x_2^2 + 0{,}25 \overset{!}{=} 0 \end{aligned}\right\}$ $x_1 = x_2 = 0{,}5$, $x = \begin{pmatrix} 0{,}5 \\ 0{,}5 \end{pmatrix}$

2) $\left.\begin{aligned} F_1(x_1, x_2) &= x_1 \cdot \sin x_1 - x_2 = 0 \\ F_2(x_1, x_2) &= x_2^2 - x_1 + 1 = 0 \end{aligned}\right\}$ hat abzählbar viele Lsg.

Sei $n = 1$, $f \in C^0([a,b])$ und $f(a) \cdot f(a) < 0 \overset{ZWS}{\Rightarrow} \exists x \in (a,b): f(x) = 0$, $a_0 = 0$, $b_0 = k$

$x_0 = \frac{1}{2}(a,b)$ Für $k = 1, \dots, k_{max}$

Falls $f(a_k) f(x) < 0$, $a_{k+1} = a_k$, $b_{k+1} = x_k$

Andernfalls $b_{k+1} = b_k$, $a_{k+1} = x_k$

5.2 Fixpunktiteration

$f(x^*) = y \Rightarrow F(x^*) = f(x^*) - \overset{\text{gegeben}}{g} = 0$. $F: D \to \mathbb{R}^n$ $\quad (D \subset \mathbb{R}^n)$

Gesucht x^* Nullst. von F

$\phi(x^*) = x^* + F(x^*) = x^*$ „Fixpunktproblem

oder allgemeiner: Sei für $x \in D$

$G(x) \in \mathbb{R}^{n \times n}$ regulär

$F(x^*) = 0$

$\Leftrightarrow \underbrace{x^* - G(x^*)F(x^*)}_{\phi(x^*) :=} = x^*$

(1) Fixpunktgleichung $\phi(x^*) = x^*$

Fixpunktiteration

wähle $x_0 \in D$, $x_1 = \phi(x_0)$, $x_2 = \phi(x_1) \dots x_{n+1} = \phi(x_n)$

Hoffnung: $x_n \to x^*$, $n \to \infty$?

Def. 5.1: Sei $(X, \|\cdot\|)$ ein norm Raum (z.B. $X = \mathbb{R}^n$, $\|\cdot\| = \|\cdot\|_2$)

$D \subset X$, $\phi: D \to X$ heißt

1) Lipschitzstetig, falls $L > 0$ ex. mit $\|\phi(x) - \phi(y)\| \leq L\|x - y\| \; \forall x, y \in D$ ($\Rightarrow$ glm st.)

2) Kontraktion, falls $0 < L < 1$

Sei $\phi: D \to \mathbb{R}^n$, $\phi \in C^1(D)$, $x = (x_1, \dots, x_n)^T \in D$

(Frechet/totale) Ableitung $\phi'(x) = \begin{pmatrix} \frac{\partial \phi_1}{\partial x_1} & \cdots & \frac{\partial \phi_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial \phi_n}{\partial x_1} & \cdots & \frac{\partial \phi_n}{\partial x_n} \end{pmatrix}$ $(x_1, \dots, x_n) \in \mathbb{R}^{n \times n}$ Jacobi-Matrix

Satz 5.2 FP Satz von Banach

Sei $(X, \|\cdot\|)$ norm Raum, $D \subset X$ vollständig, $\phi: D \to D$ (Selbstabb) eine Kontraktion

$\Rightarrow$ 1) es ex genau ein FP $x^* \in D$, $x^* = \phi(x^*)$

2) $\forall x_0 \in X$, $x_{n+1} := \phi(x_n)$ konvergiert $x_n \to x^*$, $n \to \infty$

3) a posteriori-Abschätzung: $\|x_n - x^*\| \leq \frac{\overset{\text{Kontraktionskonstante}}{L}}{1-L} \|x_n - x_{n-1}\|$

4) a priori-Abschätzung: $\|x_n - x^*\| \leq \frac{L^n}{1-L} \|x_1 - x_0\|$

5) $\|x^{n+1} - x^*\| \leq L\|x^n - x^*\|$

Korollar 5.3 Sei $\phi: D \to D$, $D \subset \mathbb{R}^n$ abgl., zshg. konvex, $\|\cdot\|$ Norm in $\mathbb{R}^n$ und $\|\phi'(x)\|$: Operatornorm von $\phi'(x) \in \mathbb{R}^{n\times n}$, $\phi \in C^1(D)$

Falls $\max_{x\in D} \|\phi'(x)\| < 1$, dann ist ϕ eine Kontraktion mit $L := \max_{x\in D} \|\phi'(x)\|$; $\Rightarrow$ Aussagen von FD von Banach

Beweis Seien $x, y \in D$ bel. $t \mapsto g(t) = x + t(y-x) = (1-t)x + ty$, $0 \le t \le 1$

$$\|\phi(x) - \phi(y)\| = \left\|\int_0^1 \phi'(x+t(y-x))(x-y)\,dt\right\| \le \int_0^1 \|\phi'(x+t(y-x))\|\,\|x-y\|\,dt$$

$$\le \underbrace{\max_{x\in D} \|\phi'(x)\|}_{=L} \|x-y\| \quad \blacksquare$$

Definition $(X, \|\cdot\|)$ norm Raum, $(x_k)_{k\in\mathbb{N}}$, $x^*, x_k \in X$, $\lim_{k\to\infty} x_k = x^*$.

1) (x_k) heißt konvergent von Ordnung $p > 1$,
falls $c > 0$, $k_0 \in \mathbb{N}$ ex. mit $\|x_{n+1} - x^*\| \le C\|x_n - x^*\|^p$ $\forall n > k_0$ ($p=2$, quadratisch)

2) „lineare Konvergenz" falls $0 < L < 1$ ex mit Konvergenzfehler

$$\|x_{n+1} - x^*\| \le L\|x_n - x^*\|$$

Bsp FP Iteration ist linear konvergent

Newton Verfahren: $G(x) := (F'(x))^{-1}$

$\Rightarrow x_{n+1} := x_n + (F'(x_n))^{-1} F(x_n)$

numerisch: $x_{n+1} = x_n + d_n$, $F'(x_n)\,dx = F(x_n)$ lin. GL. lösen

$F(x) = 0 \overset{!}{=} F(x) = F(x_n) + F'(x_n)(x_{n+1} - x_n) + \cancel{o(\|x - x_n\|)}$ → weglassen da linearisieren

$$\Rightarrow \quad -F(x_n) = F'(x_n)x_{n+1} - F'(x_n)x_n$$

$$F'(x_n)x_{n+1} = F'(x_n)x_n - F(x_n) \quad | \; F'(x_n)^{-1}$$

$$x_{n+1} = x_n - F'(x_n)^{-1}F(x_n)$$

Satz 5.3 (Kantorowich – Newton)

Sei $D \subset \mathbb{R}$ offen und konvex

$F \in C^1(D)$ und $\sup_{x\in D} \|F'(x)\| = \beta$ und $\|F'(x) - F'(y)\| \le L\|x-y\|$, $0 < L$ (Lipschitzkonstante)

und es ex. $x^* \in D$ mit $F(x^*) = 0$;

$U_\omega(x^*) = U_\omega := \{x \in D : \|x - x^*\| < \omega\}$ und $\frac{\beta L \omega}{2} < 1$

$\Rightarrow \forall x_0 \in U_\omega(x^*)$ konvergiert x_n, $(x_{n+1} = x_n - (F'(x_n))^{-1}F(x_n))$

gegen x^* quadratisch (d.h. $p=2$) $\|x_{n+1} - x^*\| \le C\|x_n - x^*\|^2$

$\phi: D \to \mathbb{R}^n,\ D \subset \mathbb{R}^n,\ x = (x_1, \dots, x_n)$

$\phi(x^*) = 0,\ x^* \in D$ gesucht

N.V. $x^{n+1} = x^n + d^n$

mit $\phi'(x_n) d_n = -\phi(x_n)$ $\quad (x^n)_{n\in\mathbb{N}},\ x^n \to x^*,\ n \to \infty$?

1) $\|(\phi'(x))^{-1}\|_{op} \le \beta,\ \forall x \in U_\omega(x^*) \cap D$ konvex

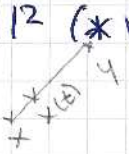
2) $\|\phi'(x) - \phi'(y)\| \le L\|x-y\| \ \forall x, y \in U_\omega(x^*) \cap D$ $\quad$ 3) $\frac{L\beta\omega}{2} < 1$

$\Rightarrow x_0 \in U_\omega \Rightarrow x^n \in U_\omega(x^*),\ n \in \mathbb{N},\ \|x^* - x^{n+1}\| \le C\|x^n - x^*\|^2$

$\to x^n \to x^*,\ n \to \infty$ „quadr."

<u>Beweis</u> wir zeigen $\forall x, y \in U_\omega(x^*)$

$\|\phi(x) - \phi(y) - \phi'(y)(x-y)\| \le \frac{L}{2}\|x-y\|^2$ (* Vermutung)



Sei $t \mapsto \varphi(t) = \phi(\underbrace{y + t(x-y)}_{x(t)}),\ t \in [0,1]$

$\varphi'(t) = \underbrace{\phi'(y + t(x-y))}_{\in \mathbb{R}^{n\times n}} \underbrace{(x-y)}_{\in \mathbb{R}^n}$

$$\|\varphi'(t) - \varphi'(0)\| = \|(\phi'(y + t(x-y)) - \phi'(y))(x-y)\|$$
$$\le \|\phi'(y + t(x-y)) - \phi'(y)\| \, \|x-y\|$$
$$\le L\|y + t(x-y) - y\| \, \|x-y\| \le L_1 t \|x-y\|^2$$

$$\|\phi(x) - \phi(y) - \phi'(y)(x-y)\| = \|\varphi(1) - \varphi(0) - \varphi'(0)\| \le \int \|\varphi'(t) - \varphi'(0)\| \, dt$$
$$\le L\|x-y\|^2 \int_0^1 t \, dt \le \frac{L}{2}\|x-y\|^2 \quad \text{(Beweis von *)} \ \blacksquare$$

Sei $x^n \in U_\omega(x^*)$:

$$x^{n+1} - x^* = x^n - (\phi'(x_n))^{-1}(\phi(x_n) - \underbrace{\phi(x^*)}_{=0}) - x^*$$
$$= (\phi'(x_n))^{-1} \{ \phi'(x_n)(x^n - x^*) - (\phi(x^n) - \phi(x^*)) \}$$

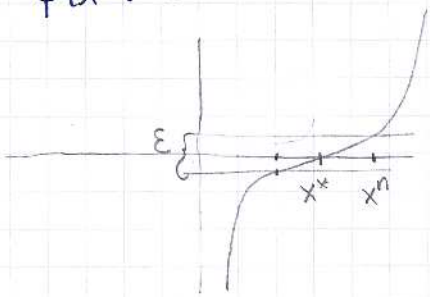
$$\Rightarrow \|x^{n+1} - x^*\| = \|x^n - x^* - (\phi'(x_n))^{-1}(\phi(x^n) - \phi(x^*))\|$$
$$\le \underbrace{\|(\phi'(x_n))^{-1}\|}_{\le \beta} \, \|\phi'(x_n)(x^n - x^*) - (\phi(x^n) - \phi(x^*))\|$$
$$\overset{(*)}{\le} \frac{\beta L}{2} \underbrace{\|x^n - x^*\|}_{< \omega} \, \|x^n - x^*\|$$
$$\le \underbrace{\frac{\beta L \omega}{2}}_{= q < 1} \|x^n - x^*\| = q\|x^n - x^*\|$$

$\Rightarrow x^{n+1} \in U_\omega(x^*),\ x^n \to x^*,\ n \to \infty$ (Banach), mit $p = 2$

Abbruchkriterium $\phi(x^*)=0$

Residuum

$\|\phi(x^n)\| < \varepsilon$



Alternative:

$$\|\underbrace{(\phi'(x^n))^{-1}\phi(x^n)}_{d^n}\| < \varepsilon$$

a posteriori Fehler

Monotoniekriterium (Konvergenztest)

geg. $0<\theta<1$: $\|\phi(x^{n+1})\| \le \theta \|\phi(x^n)\|$ ← die Frage ist ob das gilt

(1) Idee $\|x^{n+1}-x^*\| < \theta' \|x^n - x^*\|$, $\theta' < 1$)

Alternativ: $\|(\phi'(x^{n+1}))^{-1}\phi(x^{n+1})\| \le \theta \|(\phi'(x^n))^{-1}\phi(x^n)\|$ $0<\theta<1$

Gedämpftes Newton Verfahren

$x^{n+1} = x^n + \lambda_k^n d^n$, $(d^n = (\phi'(x^n))^{-1}\phi(x^n))$,

Für $k = 0, \dots, K$ mit $0 < \lambda_k^n \le 1$

$x^{n+1} := x^n + 1/2^k \; d^n = x^n + \lambda_k^n d^n$ $\lambda_0^n = 1, \lambda_1^n = 1/2, \lambda_k^n = 2^{-k}$

↳ Monotonietest → bestanden ⇒ $x^n \to x^{n+1}$ Iteration

→ nicht bestanden ⇒ $\lambda_k^n \to \lambda_{k+1}^n = \frac{\lambda_k^n}{2}$

Vereinfachte N-V

Für $n \gg 1$ ist die Berechnung von $\phi'(x^n)$ bzw. $(\phi'(x^n))^{-1} \in \mathbb{R}^{n\times n}$ sehr teuer.

Vereinfacht NV: $\phi'(x^n) \approx \phi'(x^0)$, konvergiert unter obigen Vor.

d.h. $x^{n+1} := x^n + (\phi(x^0))^{-1}\phi(x^n)$, aber nun linear

Praktisch macht man oft ein Update der Jacobi-Matrix $\phi'(x^n)$ nach k-vereinf-Iteration (z.B. $k=5, 10$ etc)

Quasi-Newton Verfahren

z.B. Sekantenverfahren ($n=1$), „Regulafalsi", $n>1$ „Broyden Verfahren" ($p \approx 1{,}6$) (usw)

Hier $1<p<2$. Oftmals Methoden der Wahl.

Nichtlineare Optimierung (ohne Nebenbed)

Problem Kostenfktn $J: D \to \mathbb{R}$, $D \subseteq \mathbb{R}^n$

gesucht $x^* \in D$, mit $J(x^*) \leq J(x)$ $\forall x \in D$, $x^* = \arg\min\{J(x); x \in D\}$

Sei $J \in C^k(D)$, $k = (1), 2 \ldots$

Notwendige Bedingung $J'(x^*) = 0$ „stationärer Punkt"

Gradienten Verfahren („depest decent")

$$x^{n+1} = x^n + \lambda^n d^n \quad \text{mit } d^n := -\operatorname{grad} J(x^n) = -\nabla J(x^n)$$

Liniensuche für λ^n

(z.B. $\lambda^n_{k+1} = 1/2\, \lambda^n_k$)

Monotonietest: z.B. $J(x^{n+1}) < \theta J(x^n)$, $0 < \theta < 1$

$\Rightarrow$ Konvergenz gegen stationäre Punkte. i.R. lokale Minima.

Alternativ: „Gauß-Newton-Verf"

$x \mapsto \phi(x) = \nabla J(x)$: $D \to \mathbb{R}^n$, zur Lösung $\nabla J(x^*) = \phi(x^*) = 0$

$\phi'(x) = H_J(x) = J''(x)$, ~~Hess~~ Hessematrix von J

6. Iterationsverfahren zur LSg. linearer GL.

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

Ges: $x \in \mathbb{R}^n$ $Ax = b$, $Ax - b = 0$

Eliminationsmeth.: # arith. Op $O(n^3)$

Iterationsref: Matrix-Vektor multiplikation pro Iterationsschritt

$O(n^2)$ falls A vollbesetzt

$O(n)$ falls A dünnbesetzt d.h. # nnz $A = O(n)$

In der Praxis z.B. "FEM" usw. häufig $n \gg 10$
Finite-Elemente-Methode

Fixpunkt(FP)-Iteration: Sei $C (\approx A^{-1}) \in \mathbb{R}^{n \times n}$ „Vorkonditionieren"

$\phi(x) = x - C(Ax - b) = x$ „Fixpunkt-Problem" $\phi: \mathbb{R}^n \to \mathbb{R}^n$, $D = \mathbb{R}^n$, vollst. $\|\cdot\| = \|\cdot\|_p$

FP-Iteration: $x^0 \in \mathbb{R}^n$ (bel. gewählt)

$$x^{n+1} = \phi(x^n) = x^n + C(b - Ax^n) = (I - CA)x^n + Cb$$

$M := (I - CA)$ Iterationsmatrix

ϕ: Kontraktion? $\|\phi(x)-\phi(x')\| = \|(I-CA)x + Cb - \{(I-CA)x' + Cb\}\|$

$= \|(I-CA)x\| \leq \|I-CA\| \, \|x\|$

$\|I-CA\| < 1 \Leftrightarrow \phi$ Kontraktion

<u>Satz 6.1.</u> Sei $\rho := \|I-CA\|_{op} < 1$, dann konvergiert

$\forall x^0 \in \mathbb{R}$, $x^{n+1} := x^n + C(b - Ax^n) \Rightarrow x \in \mathbb{R}^n$ linear

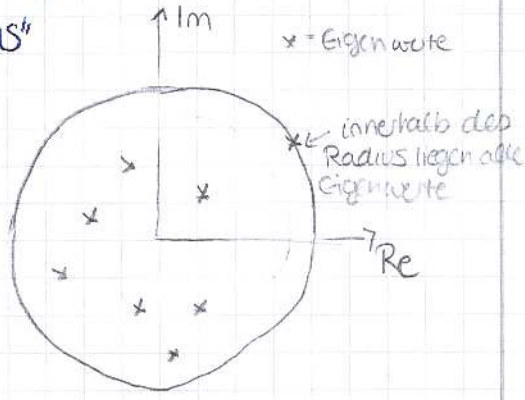
mit 1) $\|x^{n+1} - x\| \leq \rho \|x^n - x\|$, ρ - Fehlerverbesserfaktor

2) $\|x^n - x\| \leq \frac{\rho^n}{1-\rho} \|x^1 - x^0\|$, $\quad (\|x^{n+1} - x\| \leq \rho^{n+1} \|x^0 - x\|)$

<u>Beweis</u> Banach EPS

<u>Def</u> Seien λ_i, $i = 1, \dots, n$ $\lambda_i \in \mathbb{C}$ die Eigenwerte von A

$\rho(A) = \max_{1 \leq i \leq n} |\lambda_i|$ „Spektralradius"



<u>Satz</u> 6.2. Sei $M = (I - CA)$

$x^{n+1} = Mx^n + Cb$, dann konvergiert

$x^n \to x \Leftrightarrow \rho(M) < 1$

<u>Beweis</u> 1) Sei $\rho(M) > 1$ $v = x + x^0 \neq 0$, $x^0 \in \mathbb{R}$ bel. fest, wähle $v = c_{\lambda_{max}}$ Eigenwert

„$\Rightarrow$" $x^{n+1} - x = \cancel{Mx^n - Mv = (I-CA)x^n - (I-CA)}$

$= Mx^n + Cb - (Mx + Cb) = M(x^n - x) = M^{n+1}(x - x^0) = M^{n+1} v$,

$\Rightarrow \|x^n - x\| \leq \|M^n v\| \leq \|\sum_{k=1}^{n} \lambda_k^n \alpha_k c_k\| = \lambda_{max}^n \alpha_{\lambda_{max}} c_{\lambda_{max}}\|$ $\quad v = \alpha_1 c_1 + \dots + \alpha_n c_n = \alpha_{\lambda_{max}} c_{\lambda_{max}}$

$= |\underbrace{\lambda_{max}}_{>1}|^n \underbrace{\|x_0\|}_{\neq 0} \xrightarrow[n \to \infty]{} \infty$

$\Rightarrow x^n$ divergiert

„$\Leftarrow$" $\rho(M) < 1$, $\forall x_0 \in \mathbb{R}^n$, $\Rightarrow v = x^0 - x$

Es gilt $\|x^n - x\| = \|M^n v\|$;

Sei $T \in \mathbb{R}^{n \times n}$ reg. mit $M = T^{-1} J T$ die Jordanzerlegung von M

d.h. $J = \mathrm{diag}(J_j) = J_\ell = \begin{pmatrix} \lambda_\ell & & \\ & \ddots & \\ & & \lambda_\ell \end{pmatrix} = \lambda_\ell I + S_\ell$, $S_\ell = \begin{vmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{vmatrix} \in \mathbb{R}^{m \times m}$

$\Rightarrow S_\ell^m = 0$

$M^n = (T^{-1} J T)^n = T^{-1} J^n T$, es genügt

$J_\ell^n \to 0$, $n \to \infty$ zu zeigen $\forall \ell$ $n > m$, m fest $n \to \infty$

$$J_\ell^n = (\lambda_\ell I + S_\ell)^n = \sum_{k=0}^{n} \binom{n}{k} \lambda_\ell^{n-k} S_\ell^n$$

$$= \sum_{k=0}^{m} \binom{n}{k} \lambda_\ell^{n-k} S_\ell^n \xrightarrow{n\to\infty} 0 \,, \quad \lambda_\ell^{n-k} \xrightarrow[n\to\infty]{} 0 \text{ für } k=0,\dots,m \quad \blacksquare$$

<u>Korollar: 6.2</u> $\forall$ induzierten Operatornormen $\|\cdot\|_{op}$ gilt $\rho(M) \le \|M\|_{op}$

6.2.1 Richardson Iteration

$C = \alpha I, \alpha \in \mathbb{R}$ geeignet

R-It $\Rightarrow x^{n+1} = (I - \alpha A)x^n + \alpha b = x^n - \alpha A x^n + \alpha b$

konvergiert für $\rho(I - \alpha A) < 1$, z.B. für $\|I - \alpha A\| < 1 \;\forall x^0 \in \mathbb{R}^n$

$(A^{-1} \approx \alpha A ??!)$

6.2.2 Gesamtschritt oder Jacobi-Verfahren

wir zerlegen $A = \underbrace{D}_{\text{Diag}} + \underbrace{L}_{\text{untere}} + \underbrace{R}_{\text{obere}}$, $D = \operatorname{diag}(a_{i,i})$, $L = \begin{pmatrix} 0 & & 0 \\ a_{21} & \ddots & \\ & & 0 \end{pmatrix}$, $R = \begin{pmatrix} 0 & & * \\ & \ddots & \\ 0 & & 0 \end{pmatrix}$

Vorauss: Sei $D \in \mathbb{R}^{n\times n}$ regulär, d.h. $a_{i,i} \neq 0 \;\forall i = 1,\dots,n$

Wähle $C := D^{-1} \; (\approx A^{-1})$

$$\Rightarrow x^{n+1} = x^n - D^{-1}(D+L+R)x^n + D^{-1}b = x^n - x^n - D^{-1}(L+R)x^n + D^{-1}b$$

d.h. $x^n = (x_1^n \dots x_n^n)^T \in \mathbb{R}^n$

$$x_k^{n+1} = a_{kk}^{-1}\Big(b_k - \sum_{\substack{j=1 \\ j\neq k}}^{n} a_{k,j}\, x_j^n\Big) \quad k = 1,\dots,n$$

$M = D^{-1}(L+R)$

$\rho(M) < 1$? $\quad \|D^{-1}(L+R)\|_1 = \max_{k\le n} \sum_{j\neq k} \frac{1}{|a_{k,k}|} |a_{k,j}|$

<u>Def</u> $A \in \mathbb{R}^{n\times n}$ heißt <u>strikt diagonaldominant</u>

$$\max_{1\le k\le n} \sum_{\substack{j\neq k \\ j=1}}^{n} |a_{j,k}| \le |a_{k,k}|$$

Sei $A \in \mathbb{R}^{n,n}$ regulär und sei $b \in \mathbb{R}^n$ gegeben $Ax = b$

Gesucht: $x \in \mathbb{R}^n$

$A = D + L + R$, D diagonal, R obere Δ-M, L untere Δ-M

Mit Jakobi (Gesamtschritt)-Verfahren:

$$x_i^{k+1} := (a_{i,i})^{-1}\left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^k - \sum_{j=i+1}^{n} a_{ij} x_j^k\right) \qquad i = 1,\dots,n \qquad C = D^{-1}$$

Mit Gauß-Seidel (Einzelschritt)-Verfahren

$$x_i^{k+1} := (a_{i,i})^{-1}\left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^k - \sum_{j=i+1}^{n} a_{ij} x_j^k\right) \qquad C = (D+L)^{-1} \qquad i = 1,\dots,n$$

Def: $A = (a_{ij})_{i,j=1,\dots,n}$

A strik diagonaldominat, falls gilt

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| < |a_{ii}| \qquad \forall i = 1,\dots,n$$

$$\sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{ij}| < |a_{jj}| \qquad \forall j = 1,\dots,n$$

6.4. A strikt diagonaldominant, dann konvergieren Jakobi und Gauß-Seidel für Wahl des Startvektors (linear)

6.5. A spd, dann konvergiert Gauß-Seidel ($\forall x_k$ linear)

<u>Relaxationsverfahren</u>

Iterationsvorschrift: $x^{k+1} = Gx^k + c$

Verfahren konv $\Leftrightarrow \rho(G) < 1$

Um dies zu verbessern:

$\omega \in \mathbb{R}$ geeignet $\quad x^{k+1} = \omega(Gx^k + c) + (1-\omega)x^k$

ω so gewählt, dass $\rho(\omega G + (1-\omega) I)$ minimiert wird

Angewandt auf Jakobi oder Gauß-Seidel $\Rightarrow$ SOR-Verfahren

6.3. Gradienten und CG-Verfahren

Sei $A \in \mathbb{R}^{n,n}$ spd,

dann definiert $\langle \cdot,\cdot \rangle_A = \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$

ein Skalarprodukt $\langle u,v \rangle_A = \langle u, Av \rangle = \langle Au, v \rangle = u^T A v$

mit der Norm $\|u\|_A = \sqrt{\langle u,u \rangle_A}$

Sei $u \in \mathbb{R}^n$, $u \to J(u) := \frac{1}{2} \langle Au, u \rangle - \langle b, u \rangle$

und $x := \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} J(u)$ erfüllt

$J'(x) = 0 = (Ax)^T - b^T \Leftrightarrow Ax = b$

<u>Ritz-Galerkin-Verfahren</u>

Sei $V_k = \operatorname{span}\{v_j, j = 1, \dots, k\} \subset \mathbb{R}^n$. Wir suchen $x^k \in V_k$

$x^k = \operatorname{argmin} \{ J(u^k) : u^k \in V_k \}$

Gradientenverfahren $(x^k)_{k \in \mathbb{N}}$, $x^k \in \mathbb{R}^n$, Folge mit

$J(x_k) \searrow J(x) (\leq J(u)\ \forall u \in \mathbb{R}^n$, Min) $k \to \inf$

<u>Idee:</u> $x^{k+1} = x^k - \alpha^k \nabla J(x^k)$ ($\alpha^k \in \mathbb{R}_+$ geeignet)

$= x^k + \alpha^k p_k$

$p_k = -\nabla J(x^k) = b - Ax^k =: r^k$

$x^k + \alpha^k p^k =: x + tp$ ($x = x^k$, $p = p^k$ fest)

$J(x+tp) = \tilde{J}(t) = \frac{1}{2} \langle A(x+tp), (x+tp) \rangle - \langle b, (x+tp) \rangle$

$= \frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle + t \langle Ax - b, p \rangle + \frac{1}{2} \langle Ap, p \rangle t^2 \to \min$

Führt zu $\frac{d}{dt} \tilde{J}(t) = 0 = \langle Ax - b, p \rangle + t \langle Ap, p \rangle$

$\Leftrightarrow t = \frac{\langle b - Ax, p \rangle}{\langle Ap, p \rangle}$

$\alpha^n = \frac{\langle r^k, p^k \rangle}{\langle Ap^k, p^k \rangle}$

<u>Gradientenverfahren</u> $p^k = b - Ax^k$

$\alpha^k := \frac{\langle b - Ax^k, b - Ax^k \rangle}{\langle Ap^k, p^k \rangle} = \frac{\langle p^k, p^k \rangle}{\langle Ap^k, p^k \rangle}$

$x^{k+1} = x^k + \alpha^k p^k$

(Vergleich Richardson-Vergleich)

Beobachtung: $x^k \in \operatorname{span}\{b, Ar^0, A^2r^0, \dots, A^{k-1}r^0\} =: \mathcal{K}^k$

$\operatorname{span}\{x^j : j<k\} \oplus p_k$

Frage: existiert eine Linearkombination der Iterationen

d.h. $x_*^k \in \mathcal{K}^k$, sodass das Funktional auf $\mathcal{K}^k$ minimiert?

Ja! $\rightsquigarrow$ CG-Hestens-Stiefel

$Ax = b$, $r = 0 = b - Ax$

$x_0 \in \mathbb{R}^n$ bel (z.B. $x_0 = 0$) $\quad x =: x_0 + x^*$

$A(x_0 + x^*) = b \Leftrightarrow Ax^* = b - Ax_0 =: \tilde{b}$

OBdA: Sei $x_0 := 0$ gewählt $\Rightarrow r_0 = b$

Krylov-Raum $\mathcal{K}_k \subseteq \mathbb{R}^n$: $\mathcal{K}_k := K(k, b) = \operatorname{span}\{r_0, Ar_0, \dots, A^{k-1}r_0\}$

$= \operatorname{span}\{b, Ab, \dots, A^{k-1}b\}$

$= \operatorname{span}\{A^j b \mid j = 0, \dots, k-1\}$

A ist spd: $J: \mathbb{R}^n \to \mathbb{R}$: $J(u) := \frac{1}{2}\langle u, Au\rangle - \langle b, u\rangle = \frac{1}{2} u^T A u - b^T u$

$x = A^{-1}b \Leftrightarrow x = \operatorname{argmin}\{J(u) \mid u \in \mathbb{R}^n\}$

J strikt konvex $\Rightarrow \exists!\, x$

<u>Galerkin-Ritz-Verfahren</u>

Sei $U_k \subset \mathbb{R}^n$, z.B. $U_k := \mathcal{K}_k = \operatorname{span}\{p_1, \dots, p_k\}$

$\exists!\, x^k \in U_k \quad x^k = \operatorname{argmin}\{J(u) \mid u \in U_k\}$ restringiertes Optimierungs-Problem

$U_k (= \mathcal{K}_k)$ ist zulässige Menge (lineare Teilraum, konvex)

$x^k \in U_k$ innerer Pkt $\exists \in C^1$

$\Rightarrow$ Notwendige Bed alle Richtungsableitungen von J um $x_k \in U_k$ sind null, d.h. $0 = J'(x_k)(u) \quad \forall u \in U_k \ (\|u\| = 1)$

$= (\nabla J(x_k))^T u$

$= \langle Ax_k - b, u\rangle \quad \forall u \in U_k$ (Schwache Formulierung)

Ansatz: $U_k \ni x_k = \sum_{j=1}^{k} c_j p_j \Rightarrow \sum \langle p_i, Ap_j\rangle c_j = \langle p_i, b\rangle \ \forall i = 1, \dots, k$

$\Rightarrow \underline{\underline{A}}c = \underline{b} \in \mathbb{R}^k$ <u>Galerkin-Gleichung</u> $\quad \begin{cases} \langle Ax_k, u\rangle \\ = \langle b, u\rangle \end{cases} \forall u$

Lemma Galerkin-Orthogonalität

$$x_k = \arg\{J(u) \mid u \in U_k\} \Leftrightarrow r_k = b - Ax_k \perp U_k$$

$$(0 = \langle r^k, u\rangle = \langle b - Ax_k, u\rangle) \quad \forall u \in U_k$$

C-G-Verfahren

$$U_k := \mathcal{K}_k = \operatorname{span}\{r_0, Ar_0, \dots, A^{k-1} r_0\} = \operatorname{span}\{p_1, \dots, p_k\}$$

mit A-orthogonalem p_k, d.h. $\langle p_j, p_k\rangle_A = \langle p_j, Ap_k\rangle = \delta_{ik} \langle p_k, p_k\rangle_A$

Sei $k < n$, $U_k = \mathcal{K}_k$, $r_k = \underbrace{b - Ax_k}_{\in \mathcal{K}_{k+1},\ \notin \mathcal{K}_k}$ $(x_k = \mathcal{K}_k$

Gram-Schmidt $p_{k+1} = r_k - \sum_{j=1}^{k} \frac{\langle r_k, p_j\rangle_A}{\langle p_j, p_j\rangle_A} p_j$

$$= r_k - \frac{\langle r_k, p_k\rangle_A}{\langle p_k, p_k\rangle_A} p_k - \cancel{\frac{\langle r_k, Ap_{j-1}\rangle}{\langle \dots\rangle}} = 0 \text{ wegen Galerkin-orthogon.}$$

(wollen, dass x_k Galerkin-Lsg sind!!)

Seien $x_k = \operatorname{argmin}\{J(u) \mid u \in \mathcal{K}_k\}$

$\Rightarrow p_{k+1} = r_k + \beta_k p_k$ mit $\beta_k := -\frac{\langle r_k, Ap_k\rangle}{\langle p_k, Ap_k\rangle}$

konjugierter Gradient

Wir wählen die Iterationsfolge $x_k \in \mathcal{K}_k$, sodass

$$x_k := \operatorname{argmin}\{J(u) \mid u \in \mathcal{K}_k\}$$

$$\mathcal{K}_{k+1} = \operatorname{span}\{p_1, \dots, p_{k+1}\} = \mathcal{K}_k \oplus \operatorname{span}\{p_{k+1}\}$$

$\Rightarrow x_{k+1} = x_k + \alpha_k p_{k+1}$ (** Behauptung) falls v_k statt p_k dann wie Gradienten-verfahren

$\Rightarrow x_{k+1} = \operatorname{argmin}\{J(u) \mid u \in \mathcal{K}_{k+1}\} \Rightarrow \alpha_k = \frac{\langle p_{k+1}, p_{k+1}\rangle}{\langle Ap_{k+1}, p_{k+1}\rangle}$

Gradienten-update

$$r_k = b - Ax_k = A(x - x_k) = A(x - (x_{k-1} + \alpha_{k-1} p_k))$$

$$= r_{k-1} - \alpha_{k-1} Ap_k$$

Lemma: Es gilt $\langle v_k, v_j\rangle = (v_k)^T v_j = \delta_{kj} \langle r_k^*, r_k\rangle$

Beweis: Galerkin-Orthogonalität. Sei $j < k$ (oBdA)

$\Rightarrow r_j = b - Ax_j \in \mathcal{K}_{j+1}$ u.G.O $\Rightarrow r_k \perp \mathcal{K}_k$

$\Rightarrow j < k$ $r_k \perp \mathcal{K}_{j+1} \Rightarrow \langle r_k, r_j\rangle = 0$ ∎

Allgemeiner Ansatz

$$x_{k+1} = x_k + \alpha_k \left(r_k - \sum_{j=1}^{k} \beta_j p_j\right) \qquad \left[\mathcal{K}_{k+1} = \mathcal{K}_k + \mathrm{span}\{r_k\}\right]$$

$\in \mathcal{K}_{k+1}$ $\quad \langle p_j, p_i \rangle_A = 0, \; i \neq j$

$$(A x_{k+1} - b) \cdot u = \left(A\left(x_k + \alpha_k \left(r_k - \sum_{j=1}^{k} \beta_j p_j\right)\right) - b\right) u$$

Galerkin Gleichung mit $x_{k+1} = \mathrm{argmin}$

$$\left\langle A\left(x_k + \alpha_k \left(r_k - \sum_{j=1}^{k} \beta_j p_j\right)\right) - b, p_i \right\rangle = 0 \quad \forall i = 1, \dots, k+1$$

$$\underbrace{\langle A(x_k - \alpha_k r^k - \alpha_k \beta_k p_k) - b, p_i \rangle}_{=0} - \sum_{j=1}^{k-1} \beta_j \underbrace{\langle A p_j, p_i \rangle}_{= \delta_{ij} \langle p_j, p_j \rangle} = 0 \quad \forall i \leq k-1$$

$$\Rightarrow \beta_j = 0 \quad r_j < k$$

$$\Rightarrow x_{k+1} = x_k + \alpha_k (r_k + \beta_k p_k)$$

wegen $\alpha_k \langle r_k, p_{k+1} \rangle_A = \langle r_k, r_{k+1} - r_k \rangle$

$$= \langle r_k, b - A x_{k+1} - r_k \rangle = \langle r_k, b - A(x_k + \alpha_k p_{k+1}) - (b - A x_k) \rangle$$

$$= \langle r_k, \alpha_k p_{k+1} \rangle_A$$

$$\beta_k = \langle r_k, p_{k+1} \rangle_A = \frac{\langle r_k, r_k \rangle}{\alpha_k} = \frac{\langle r_k, r_k \rangle}{\langle r_{k-1}, r_{k-1} \rangle}$$

$$\alpha_k = \frac{\langle r_{k-1}, r_{k-1} \rangle}{\langle A p_k, p_k \rangle}$$

Im Vergleich zum Gradienten 2fache Matrix-Vektormultiplikation und doppelter Speicher. ABER:

<u>Theorem</u> Es gibt Fehlerabschätzung (wissen wir nicht, schreibweise Fehler wie bei BFS)

$$\| x_k - x \|_A \leq 2 \left(\frac{1 - \sqrt{\varkappa(A)}}{1 + \sqrt{\varkappa(A)}} \right)^k \| x_0 - x \|_A \qquad \left(\text{Iterationszahl } k \sim \sqrt{\varkappa(A)}\right) \gg$$

(Grad. Verfahren:

$$\| x^k - x \|_A \leq C \left(\frac{1 - \varkappa(A)}{1 + \varkappa(A)} \right)^k \| x_0 - x \|_A \qquad \left(\text{Iterationszahl } k \sim \sqrt{\varkappa(A)}\right) !$$

Das CG Verfahren ist das schnellste Iterationsverfahren für Spd-Matrizen

In der Praxis „vorkonditioniertes CG Verfahren"

Nach n Schritten hat man die exakte Lsg.

7. Numerische Behandlung von Eigenwertproblemen

$A \in \mathbb{C}^{n\times n}$; $\langle u,v\rangle := \bar{u}^T v$

$\left[\begin{array}{l} z \in \mathbb{C} \\ z = a+ib,\ Re(z)=a,\ Im(z)=b \\ \bar{z} = a-ib \end{array}\right.$

$(A^* = \bar{A}) = A^H$, adjungierte Matrix zu A: $\langle A^H u, v\rangle = \langle u, Av\rangle$

$A = (a_{ij})_{i,j=1}^n$, $A^H = (\bar{a}_{ji})_{i,j=1}^n$

$(A \in \mathbb{R}^{n\times n} \Rightarrow A^H = A^T)$

Unitäre Matrizen bilden Gruppe

Unitäre Matrizen U: $U^{-1} = U^H$, $(U^H)^H = U$, $U_1 U_2$ ist unitär

Es gilt: $\|Ux\|_2 = \|x\|_2$

(λ, v) ein Eigenpaar, falls i) $v \neq 0$ und ii) $Av = \lambda v$ $|\lambda \in \mathbb{C}, v \in \mathbb{C}^n$

<u>Ähnliche Matrizen</u>: A und B heißen ähnlich, falls eine reguläre Matrix $T \in \mathbb{C}^{n\times n}$ existiert mit $A = T^{-1}BT$ $(A \sim B)$

Falls $A \sim B$, haben A und B die gleichen EW $\lambda_1, \dots, \lambda_n \in \mathbb{C}$

$\Rightarrow$ Falls U unitär existiert mit $A = U^H B U$, dann besitzen A und B die gleichen EW.

$\left[\{\lambda_1, \dots, \lambda_n\} =: \sigma(A)\right.$, Spetrum von A

<u>Satz 7.1</u> (Schur-Normalform)

Zu $A \in \mathbb{C}^{n\times n}$ existiert $Q \in \mathbb{C}^{n\times n}$ <u>unitär</u> und eine rechte obere Δ-Matrix $R \in \mathbb{C}^{n\times n}$ mit $A = Q^H R Q = Q^{-1} R Q$

$A^H = (Q^H R Q)^H = Q^H R^H (Q^H)^H = Q^H R^H Q$

$\Rightarrow$ <u>Korollar</u>: Falls $A^H = A$ (hermitsch) (in reellen symm.) $\Rightarrow R^H = R = diag(\lambda_i)$

<u>Beweis</u>: Es existiert ein Eigenpaar (λ_1, u_1) mit $\|u_1\| = 1$

$\{v_2, \dots, v_n\}$ eine Orthonormalbasis $(\langle v_i, v_j\rangle = \delta_{ij})$ von

$\text{span}\{u_1\}^\perp = \{v \in \mathbb{C}^n, \langle u_1, v\rangle = 0\} \Rightarrow \{u_1, v_2, \dots, v_n\}$ ist ONB von $\mathbb{C}^n$

Sei $P = (u_1, v_2, \dots, v_n) \Rightarrow P^H = P^{-1}$ ist unitär (Bew. Übung)

$$P^H A P = \begin{pmatrix} \lambda_1 & * & \cdots & * \\ 0 & & & \\ \vdots & & & \\ 0 & * & \cdots & * \end{pmatrix} = \left(\begin{array}{c|c} \lambda_1 & R_1 \\ \hline 0 & \\ \vdots & \tilde{A}_{n-1} \\ 0 & \end{array}\right) \exists \left(\begin{array}{c|c|c} \lambda_1 & * \cdots * & \\ \hline 0 & \lambda_2 & R_2 \\ \hline \vdots & 0 & \\ 0 & 0 & A_{n-2} \end{array}\right) = P_2^H P_1^H A P_1 P_2 \qquad P_2 = \begin{pmatrix} I & 0 \\ 0 & \tilde{P}_2 \end{pmatrix}$$

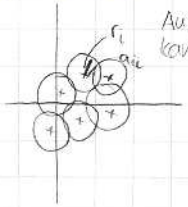
$$\underbrace{P_n^H \cdots P_2^H P_1^H}_{=:U} A P_1 P_2 \cdots P_n = UAU^H = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \Rightarrow A = U^H R U \quad \blacksquare$$

<u>Satz 7.3</u> „Kreise des Gerschgorin"

Seien $r_i = \sum_{\substack{j=1 \\ j\neq i}}^n |a_{ij}|$, $c_j = \sum_{\substack{i=1 \\ i\neq j}}^n |a_{ij}|$

$\Rightarrow \sigma(A) \subset \bigcup_{i=1}^n \{z \in \mathbb{C} \mid |z - a_{i,i}| < r_i\}$

$\subset \bigcup_{j=1}^n \{z \in \mathbb{C} : |z - a_{jj}| \leq c_j\}$



Außerhalb der Kreise kann kein Eigenwert liegen

Beweis Sei $\lambda \in \mathbb{C}$ mit $\overset{|a_{ii}-\lambda|}{|\lambda + a_{ii}|} > r_i \; \forall i = 1,\dots,n$

$B = A - \lambda I \Rightarrow |b_{ii}| = |a_{ii} - \lambda| \;\; \forall i = 1,\dots,n \;\; \forall i \neq j \; b_{ij} = a_{ij}$

$\Rightarrow \sum_{\substack{j=1 \\ j \neq i}}^{n} |b_{ij}| = \sum_{j \neq i} |a_{ij}| = r_i < |b_{ii}| \;\; \forall i = 1,\dots,n$

d.h. B ist diagonaldominant $\Rightarrow B$ ist regulär, d.h. λ ist kein EW von A

Satz 7.4 (Satz von Bauer & Fike)

Sei A diagonalisierbar, d.h. es ex. $T \in \mathbb{C}^{n \times n}$ reg.

mit $\operatorname{diag}(\lambda_i)_{i=1,\dots,n} =: D = T^{-1}AT$ (Eigenwerte von A)

$\delta A \in \mathbb{C}^{n \times n}$, λ ein EW von $B = A + \delta A$ und $1 \leq p \leq \infty$ dann gilt

$$\min_{1 \leq j \leq n} |\lambda_j - \lambda| \leq \|\delta A\|_p \operatorname{cond}_p(T)$$

Beweis $B = D - \lambda I + T^{-1}\delta A T$

$TBT^{-1} = TDT^{-1} + \lambda TT^{-1} + \delta A$ $\qquad \|A\|_p = \|A\|_{op}$ in ℓ_p

$= A + \delta A - \lambda I$ singulär!

$(D - \lambda I)^{-1} B = I + (D - \lambda I)^{-1} T^{-1} \delta A T$ sing

$\Rightarrow \rho((D - \lambda I)^{-1} T^{-1} \delta A T) > 1$

$\Rightarrow 1 < \rho(\,) \leq \|(D - \lambda I)^{-1} T^{-1} \delta A T\|_p$

$\leq \|(D - \lambda I)^{-1}\|_p \|T^{-1}\|_p \|\delta A\|_p \|T\|_p$

$\leq \dfrac{1}{\min_{1 \leq j \leq n} |\lambda_j - \lambda|} \|\delta A\|_p \operatorname{cond}(T)$ ∎

Transformation auf Hessenbergestalt

„Hessenbergmatrix" $\begin{pmatrix} \times & \times & \\ \times & \ddots & \ddots \\ 0 & \times & \times \end{pmatrix} = H$ (symm. Hessenb.m. sind tridiagonal)

Konstruiere Q unitär mit $Q^H A Q = H \Rightarrow \sigma(A) = \sigma(H)$

z.B. durch Givens Rotation aber auch durch Householder

Iterative Näherungsverfahren zur Berechnung des EW Problems symm. Matrizen

$A \in \mathbb{R}^{n \times n}$, $A^T = A$ $\quad (A^H = A)$

$\Rightarrow$ dann ex $Q \in \mathbb{R}^{n \times n}$ orth. (Q unitär)

mit $A = Q^T \operatorname{diag}(\lambda_i) Q$, EW $\lambda_1, \dots, \lambda_n \in \mathbb{R}$

Def 7.1 $\mu: \mathbb{R}^n \setminus \{0\} \to \mathbb{R}$ Rayleigh-Quotient von A

$\mu(x) = \frac{\langle Ax, x\rangle}{\langle x, x\rangle}$. ~~Es gilt das folgende~~

~~Satz 7.6 „Courant Fischer"~~

Es gilt das folgende Extremalprinzip

$\lambda_{max} = \max\{\mu(x) \mid x \in \mathbb{R}^n\}$, $\lambda_{min} = \min\{\mu(x) \mid x \in \mathbb{R}^n\} = \min\{\langle Ax, x\rangle : x \in \mathbb{R}^n, \|x\|_2^2 = 1\}$

nt? ukr

~~Satz 7.6 „Courant Fischer"~~

Eigenwertprobleme symmtr. Matrizen

$A \in \mathbb{R}^{n\times n}$, $A^T = A$ ($A^H = A$)

$A = U^T \operatorname{diag}(\lambda_i) U$, $U = (u_1, \dots, u_n)$

Satz 7.6 „Courant Fischer"

? Norm, wenn nix gesagt/geschrieben wird

Sei $\lambda_1 \le \lambda_2 \le \dots \le \lambda_n \in \omega$ mit $\in V$ $u_1, \dots, u_n$ $\|u_i\| = 1$ $i = 1, \dots, n$

1) $\lambda_i = \max\{\frac{\langle Ax, x\rangle}{\langle x, x\rangle} : 0 \ne x \in \operatorname{span}\{u_1, \dots, u_i\}\} \Rightarrow \lambda_{max} = \max_{x \in \mathbb{R}^n} \frac{\langle Ax, x\rangle}{\langle x, x\rangle}$

$= \max\{\langle Ax, x\rangle : x \in \operatorname{span}\{u_1, \dots, u_i\}, \|x\| = 1\}$

2) $\lambda_i = \min\{\frac{\langle Ax, x\rangle}{\langle x, x\rangle} : 0 \ne x \in \operatorname{span}\{u_i, u_{i+1}, \dots, u_n\}\}$

Beweis $U \in \mathbb{R}^{n\times n}$ orthog. (unitär) $A = U^T \operatorname{diag}(\lambda_i) U$

1) $\lambda_i = \max\{\lambda_k \mid 1 \le k \le i\} = \max\{\langle Au_i, u_i\rangle \mid 1 \le k \le i, \|u_i\| = 1\}$

$\lambda_i = \langle Au_i, u_i\rangle \ge \langle Ax, x\rangle, \forall x \in \operatorname{span}\{u_1, \dots, u_i\}, \|x\| = 1$

$x = \sum_{k=1}^{i} \alpha_k u_k \Rightarrow |\alpha_k| \le 1$

2) analog ∎

Potenzmethode (Poweriteration, von Mises) (Vektoriteration)

$x_0 \in \mathbb{R}^n$ gewählt: $x_{k+1} = A \cdot x_k = (A^{k+1} x_0)$, $y_{k+1} = \frac{x_{k+1}}{\|x_{k+1}\|}$

$\hat{x}_{k+1} = A x_k \Rightarrow x_{k+1} = \frac{\hat{x}_{k+1}}{\|\hat{x}_{k+1}\|}$

nur andere Schreibweise

$(\|y_k\| = 1, \forall k \in \mathbb{N})$

Satz 7.7 λ_i EW von A, $|\lambda_1| \ge |\lambda_2| \ge \dots$ $\lambda_{max,k} = \langle Ay_k, y_k\rangle$

$\langle u_1, x_0\rangle \ne 0$, $x_0 \in \mathbb{R}^n$ bel., $x_{k+1} = Ax_k$, $y_{k+1} = \frac{x_{k+1}}{\|x_{k+1}\|}$

so nicht orth um EV ansonsten erforen

$\lim_{k\to\infty} y_k = \pm u_1$, Eigenvektor zu $\lambda_1 = \langle Au_1, u_1\rangle = \lambda_{max}$

Beweis Sei $x_0 = \sum_{k=1}^{n} \alpha_k u_k$ ($\|x_0\| = 1$), da $0 \ne \langle x_0, u_1\rangle \Rightarrow \alpha_1 \ne 0$

$x_k = Ax_{k-1} = A^k x_0 = A^k (\sum_{j=1}^{n} \alpha_j u_j) = \sum_{j=1}^{n} \alpha_j A^k u_j = \sum_{j=1}^{n} \alpha_j \lambda_j^k u_j$

$= \lambda_1^k (\alpha_1 u_1 + \underbrace{\sum_{j=2}^{n} \left(\frac{\lambda_j}{\lambda_1}\right)^k \alpha_j u_j}_{\to 0 \text{ für } k\to\infty, \text{ da } q = |\frac{\lambda_j}{\lambda_1}| < 1}) \to \frac{x_k}{\|x_k\|} \to u_1$ für $k \to \infty$

zu Beweis: Konvergenzrate hängt von $\left|\frac{\lambda_2}{\lambda_1}\right| < 1$ ab.

<u>Bemerkung</u>: 1) Für $|\lambda_1| \approx |\lambda_2|$ wird das Verfahren langsam

2) Berechnet immer nur den maximalen EW.

$|\lambda_1| > \ldots \geq |\lambda_n|$, λ_i EW von A

Dann gilt: $\left|\frac{1}{\lambda_1}\right| < \ldots \leq \left|\frac{1}{\lambda_n}\right|$ EW von A^{-1} } mit den gleichen EV

<u>Inverse Iteration</u> („Wielandt")

Löse $Ax_{k+1} = x_k$, d.h. $x_{k+1} = A^{-1}x_k$, $u_{k+1} = \frac{x_{k+1}}{\|x_{k+1}\|}$

<u>Beschleunigen durch „Shift + Invert"</u>:

Sei $\mu \notin \sigma(A)$, d.h. $\min_{1 \leq j \leq n} |\mu - \lambda_j| > 0$ und $\lambda = \operatorname{argmin}_{1 \leq j \leq n} |\mu - \lambda_j|$

Sei $\rho = \frac{1}{\lambda - \mu}$, sei $u \in V$ zu $\lambda (\in \sigma(A))$.

$\Rightarrow (A - \mu I)u = (\lambda - \mu)u$, d.h. $(\lambda - \mu) \in \sigma(A - \mu I)$ und der kleinste Eigenwert von $(A - \mu I)$

$\Rightarrow (A - \mu I)^{-1}u = \frac{1}{\lambda - \mu}u = \rho u$, ρ größter? EW von $(A - \mu I)^{-1}$

<u>Potenzmethode</u> (Vektoriteration) mit $(A - \mu I)^{-1}$

d.h. $(A - \mu I)x_{k+1} = x_k$. (Problem: $\operatorname{cond}(A - \mu I) \gg 1$!

Ausweg: Jakobi-Davidson)

bester Algo um einen Eigenwert/Vektor? zu berechnen

<u>Das QR-Verfahren von Francis</u>:

$A_1 := A = Q_1 R_1$ Zerlegung (z.B. nach Householder)

$R_1 = Q_1^T A$

$A_2 := R_1 Q_1 = Q_1^T A Q_1$ $\qquad Q_k \in \mathbb{R}^{n \times n}$

$A_{k+1} = R_k Q_k \qquad \Rightarrow A_k^T = A_k$

$\Rightarrow A_k \to \operatorname{diag}(\lambda_k)$, $Q_k \cdots Q_1 \xrightarrow{k \to \infty} U$

<u>Satz 7.9</u> i) $\lim_{k \to \infty} Q_k = I$

$A^k (a_{ij}^k)_{i,j=1}^n$ ii) $\lim_{k \to \infty} R_k = \lim_{k \to \infty} A_k = \operatorname{diag}(\lambda_i)$

$|\lambda_1| > \ldots >$

iii) $|a_{ij}^{(k)}| = O\left(\left|\frac{\lambda_i}{\lambda_j}\right|\right) \forall j > i$ $\qquad$ Beweis per Induktion

cos?

Lanzcos Verfahren

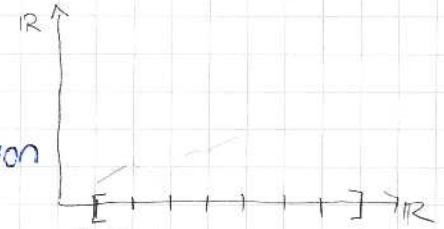
Krylov Raummethode: $\lambda_{max}^{(k)} = \max\left\{ \frac{\langle Ax, x\rangle}{\langle x, x\rangle} \mid 0 \neq x \in \mathcal{K}^k \right\}$

$\mathcal{K}^k = \operatorname{span}\{p_j : j = 1, \dots, k\}$ wie bei CG?

$(\langle Ap_i, p_j\rangle) = \begin{pmatrix} * & & 0 \\ & \ddots & \\ 0 & & * \end{pmatrix} \in \mathbb{R}^{k \times k}$

krr(?)

3.6 Splines



Splines: Stückweise Polynome mit maximaler Glattheit. Trägers von Splines (schön?)

$k \geq 2$

$a = \tau_0 < \quad \tau_\ell < b = \tau_{\ell+1} \quad \tau = \{\tau_0, \dots\}$

$S_{k,\tau} = \{ f \in C^{k-1}([a,b]) \mid f|_{[\tau_j, \tau_{j+1})} \in \Pi_{k-1}, j = 0, \dots, \ell \}$ stückweise stetig für $k = 2$?

Spline fkt der Ordnung k (Grad + 1)

Prop 3.6.1 1) $\dim S_{k,\tau} = k + \ell$

$(f(x))_+ = \begin{cases} f(x), & f(x) \geq 0 \\ 0, & \text{sonst} \end{cases}$

2) $S_{k,\tau} = \operatorname{span}\{ \Pi_{k-1}|_{[a,b]}, (x - \tau_j)_+^{k-1}, j = 1, \dots, \ell - 1 \}$

Beweis 2) lin-unab. $\Pi_k \in S_{k,\tau}$ $(\tau)_+^{k-1} \in C^{k-2}$

3.6.2. B-Splines

$k = 1$, $N_{j,1}(x) = \chi_{[\tau_j, \tau_{j+1})}(x) = \begin{cases} 1, & x \in [\tau_j, \tau_{j+1}] \\ 0, & \text{sonst} \end{cases}$

$k = 2$, $N_{j,2}(x) = \begin{cases} \frac{x - \tau_j}{\tau_{j+1} - \tau_j}, & x \in [\tau_j, \tau_{j+1}] \\ \frac{\tau_{j+2} - x}{\tau_{j+1} - \tau_{j+1}}, & x \in [\tau_{j+1}, \tau_{j+2}] \\ 0, & \text{sonst} \end{cases}$

Def: $N_{j,k}(x) = (\tau_{j+k} - \tau_j) [\tau_{j-1}, \tau_{j+k}] (\cdot - x)_+^{k-1}$

B Splines der Ordnung k und Knoten j

Satz 3.6.3 Es gilt die Rekursionsformel

1) $N_{j,k}(x) = \frac{x - \tau_j}{\tau_{k+j-1} - \tau_j} N_{j,k-1}(x) + \frac{\tau_{j+k} - x}{\tau_{k+j} - \tau_{j+1}} N_{j+1,k-1}(x)$

2) $N_{j,k}'(x) = (k-1) \left\{ \frac{N_{j,k-1}(x)}{\tau_{k+j-1} - \tau_j} - \frac{N_{j+1,k-1}(x)}{\tau_{k+j} - \tau_{j+1}} \right\}$

Bew Skizze $(\cdot - x)_+^{k-1}$ $(\cdot - x)_+^{k-2}$ $(\cdot - x)$

$[\ ](\cdot - x)_+^{k-1} = [\ ]$ " „Leibniz Formel für Dividierte Differenzen"

Satz 3.6.2 i) $N_{j,k} \in \operatorname{span}\{(\tau_i - x)_+^{k-1}, i = j, \dots, j+k\}$

ii) $\operatorname{supp} N_{j,k} = \{x : N_{j,k}(x) \neq 0\} \subset [\tau_j, \tau_{j+k}]$,

iii) $N_{j,k}(x) \geq 0$

Konvention: $\frac{0}{0} := 0$

Bew i) klar

ii 2) $x < \tau_j$, $N_{j,k}(x) \subset \Pi_{k-1}$, $\psi \in U_\delta(x) \Rightarrow [\tau_{j-1}, \tau_{j+k}](\cdot - x)_+^{k-1} = 0$

$x > \tau_{j+k}$, $N_{j,k}(x) = 0 \Rightarrow$ ii)

iii) klar

Lemma 3.6.4 Marsden Identität

$$(x-y)^{k-1} = \sum_{i=1}^{k+\ell} \prod_{j=1}^{k-1} (\tau_{i+j} - y) N_{i,k}(x)$$

Beweis Induktion über k

i) $k = 1$ klar

ii) Sei $k > 1$, $1 \leq i \leq k-1+\ell$ bewiesen

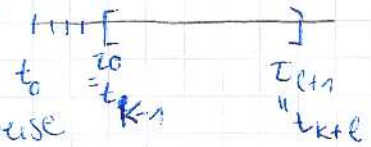
$$\varphi_{i,k} := \prod_{j=1}^{k-1} (\tau_{i+j} - y)$$

$$\sum_{i=1}^{k+\ell} (\varphi_{i,k}(y) N_{i,k}(x)) = \sum_{i=2}^{k+\ell} \left\{ \frac{x - \tau_i}{\tau_{i+k-1} - \tau_i} \varphi_{i,k}(y) + \frac{\tau_{i+k-1} - x}{\tau_{i+k-1} - \tau_i} \varphi_{i-1,k}(y) \right\} N_{i,k-1}(x)$$

$$= \sum_{i=2}^{k+\ell} \prod_{j=1}^{k-2} (\tau_{i+j} - y) \cdot \underbrace{\left\{ \frac{x - \tau_i}{\tau_{i+k-1} - \tau_i} (\tau_{i+k-1} - y) + \frac{\tau_{i+k-1} - x}{\tau_{i+k-1} - \tau_i} (\tau_i - y) \right\}}_{(x-y)}$$

$$= (x-y) \sum_{i=2}^{k+\ell} N_{i,k-1}(x) = (x-y)(x-y)^{k-2} \quad \blacksquare$$

Erweiterte Knotenfolge



t_i nicht notwendigerweise verschieden $\Rightarrow \tau_i \Rightarrow t_i$

$y \mapsto f(y) = (x-y)^{k-1}$, $f^{(\ell)}(0) = (k-1) \cdots (k-\ell)(-1)^\ell x^{k-\ell-1} = \sum_{i=1}^{k+\ell} y_{ik}^{(\ell)}(0) N_{i,k}(x)$

$\Rightarrow x^m \in \operatorname{span}\{N_{i,k} : i = 0, \dots, k+\ell\} \; \forall \; m \leq k-1$

$1 = \sum_{i=1}^{k+\ell} N_{i,k}(x)$ „Zerlegung der Eins"

Satz 3.6.5 1) $S_{k,\tau} = \operatorname{span}\{N_{i,k} : i = 1, \dots, k+\ell\}$

2) $1 = \sum_{i=1}^{k+\ell} N_{i,k}(x)$, $\Pi_{k-1} \subset \operatorname{span}\{N_{i,k}(x), i = 1, \dots, k+\ell\}$

3) $\inf_{v_k \in S_{k,\tau}} \|f - v_k\|_\infty \leq c\, h^k \sup_{x \in [a,b]} |f^{(k)}(x)|$

$$h := \max\{\tau_{i+1} - \tau_i\}$$

<u>De Boer Algorithmus</u> (Carl de Boer)

Sei $x \mapsto S(x) = \sum_{j=1}^{k+l} c_j N_{j,k}(x)$, $x \in [a,b]$

$$S(x) = \sum_{j=2}^{k+l} c_j^{[1]}(x) N_{j,k-1}(x), \quad c_j^{[1]}(x) = \frac{x - t_j}{t_{j+k-1} - t_j} c_j + \frac{t_{j+k-1} - x}{t_{j+k-1} - t_j} c_{j-1}$$

$$S(x) = \sum_{j=1+n}^{k+l} c_j^{[n]}(x) N_{j,k-n}(x)$$

$$\text{mit } c_j^{[n]}(x) = \begin{cases} \frac{x - t_j}{t_{j+k-1} - t_j} c_j^{[n-1]}(x) \\ + \frac{t_{j+k-1} - x}{t_{j+k-n} - t_j} c_{j-1}^{[n-1]}(x) \end{cases}$$

$$= \sum_{j=k}^{l+k} c_j^{[k-1]} N_{j,1}(x) = c_m^{[n-1]}(x) \quad x \in [t_m, t_{m+1}]$$

Analog lassen sich damit auch die Ableitungen von S d.h. $S^{(m)}(x)$ berechnen