

Grundlagen

Matrixmultiplikation:

$$A \in \mathbb{R}^{n \times m}, B \in \mathbb{R}^{m \times l} \quad A \cdot B = C \in \mathbb{R}^{n \times l}$$

mit $c_{ij} = \sum_{k=1}^m a_{ik} b_{kj} \quad i=1 \dots n, j=1 \dots l$

$n \times m \cdot m \times l \rightarrow n \times l$

Spezielle Matrizen:

Invertierbare Matrizen: $A \in \mathbb{R}^{n \times n}$ invertierbar, wenn:

A^{-1} inv wenn 0 kein EW
od
 $\det \neq 0$

es existiert $B \in \mathbb{R}^{n \times n}$ mit $A \cdot B = I$ Einheitsmatrix

$\Rightarrow B$ ist dann die Inverse zu A , $B = A^{-1}$

$$\text{Es gilt: } A^{-1} \cdot A = A \cdot A^{-1} = I$$

Permutationsmatrizen: $P \in \mathbb{R}^{n \times n}$ Permutationsmatrix, falls Spalten (Zeilen) aus

Einheitsvektoren bestehen, wobei jeder e_k genau einmal auftritt

z.B. $P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$

\Rightarrow Für $A \in \mathbb{R}^{m \times n}$ ist AP Matrix mit Spalten von A entspr. P vertauscht

\Rightarrow Für $B \in \mathbb{R}^{n \times m}$ ist PB Matrix mit Zeilen von B entspr. P vertauscht

Es gilt: $(P(s,i))^{-1} = P(s,i)$ \leftarrow tauscht Sp/Zeilen r.i.s

$$(! \text{ i.A. } PP = P^2 = I)$$

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

$$(AB)^T = B^T A^T = BA \quad (A, B \text{ symmetrisch})$$

\Rightarrow Produkt muss nicht symm. sein!

$A \in \mathbb{C}^{n \times n}$ hermitesch, wenn $A = \bar{A}^T$

spd-Matrizen: spd = symmetrisch, positiv definit

$A \in \mathbb{R}^{n \times n}$ positiv definit, wenn $x^T A x > 0 \quad \forall x \in \mathbb{R}^n \setminus \{0\}$

A spd-Matrix, dann gilt: $B A B^T$, $B^T A B$ symm. + pos. semidefinit

Orthogonale Matrizen: $Q \in \mathbb{R}^{n \times n}$ orthogonal, wenn $Q Q^T = I \quad (\Rightarrow Q^{-1} = Q^T)$

\Rightarrow Längenerhaltung bzgl. euklidischer Norm:

$$\|Qx\|_2^2 = (Qx)^T Qx = x^T Q^T Q x = x^T x = \|x\|_2^2$$

Rang einer Matrix: $\text{rang}(A) = \dim(\text{Bild}(A)) \rightarrow$ max. Anz. lin. unabh. Spalten / Zeilen

$A \in \mathbb{R}^{n \times n}$ invertierbar $\Rightarrow \text{rang}(A) = n$

$$\|a+b\| \leq \|a\| + \|b\|$$

$$\|ab\| \leq \|a\| \|b\|$$

Eigenwerte, Eigenvektoren: Gilt $A \cdot v = \lambda \cdot v$ für $A \in \mathbb{R}^{n \times n}$ und $v \in \mathbb{R}^{n \times 1} \neq 0$

v = Eigenvektor von A zu Eigenwert λ

$\rightarrow \text{spec}(A) = \sigma(A) = \text{Menge aller EW von } A$

Berechnung: $c_p(A) = \det(A - \lambda I) \stackrel{!}{=} 0$

- $\rightarrow A, B \in \mathbb{R}^{n \times n}$ ähnlich wenn $S \in \mathbb{R}^{n \times n}$ inv. existiert mit $A = SBS^{-1}$
 \rightsquigarrow haben gleiche EW
- \rightarrow obere/untere Dreiecksmatrizen, Diagonalmatrizen: EW auf Hauptdiagonale
- \rightarrow symm. Matrizen: nur reelle EW, es ex. orth. Matrix $Q \in \mathbb{R}^{n \times n}$ aus EV von A mit $D = QAQ^T$ (D = Diag. Matrix)
 \rightsquigarrow auf Hauptdiag. von D stehen EW von A

Normen, Skalarprodukte: $\| \cdot \| : \mathbb{R}^n \rightarrow \mathbb{R}$ Norm auf \mathbb{R}^n , wenn $\forall x, y \in \mathbb{R}^n \ \lambda \in \mathbb{R}$

(1) pos. Definitheit: $\|x\| \geq 0$ und $\|x\|=0 \Leftrightarrow x=0$

(2) Dreiecksungleit.: $\|x+y\| \leq \|x\| + \|y\|$

(3) Homogenität: $\|\lambda x\| = |\lambda| \|x\|$

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

\rightarrow euklidische Norm $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$

\rightarrow Maximumsnorm $\|x\|_\infty = \max_{i=1,\dots,n} |x_i|$

$\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ (euklidisches) Skalarprodukt, wenn $\forall x, y \in \mathbb{R}^n ; \lambda, \mu \in \mathbb{R}$

(1) Linearität in 2. Komponente: $\langle x, \lambda y + \mu z \rangle = \lambda \langle x, y \rangle + \mu \langle x, z \rangle$

(2) Symmetrie: $\langle x, y \rangle = \langle y, x \rangle$

(3) positive Definitheit: $\langle x, x \rangle \geq 0$ und $\langle x, x \rangle = 0 \Leftrightarrow x=0$

x, y orthogonal bzgl. $\langle \cdot, \cdot \rangle$, falls $\langle x, y \rangle = 0$

Induzierte Norm: $\|x\| = \sqrt{\langle x, x \rangle}$ durch $\langle \cdot, \cdot \rangle$ induzierte Norm

\rightarrow es gilt Cauchy-Schwarz-Ungleichung $|\langle x, y \rangle| \leq \|x\| \|y\| \quad \forall x, y \in \mathbb{R}^n$

euklidisches Skalarprodukt: $\langle x, y \rangle = x^T A y$, wobei $A \in \mathbb{R}^{n \times n}$ Spd-Matrix

Gleitkomma-rechnung $z = a \cdot d^e$; $d = \text{Basis}, e = \text{Exponent}$

$a = \text{Mantisse} = \sqrt[d]{z}$ Mantissenlänge

Jedes $x \neq 0$ mit $d^{\lfloor \log_d x \rfloor} \leq |x| \leq d^{\lfloor \log_d x \rfloor + 1}$ nach Rundung durch Gleitkommazahl darstellbar:

$$x = v(0.a_1a_2\dots a_{\lfloor \log_d x \rfloor} \dots) d^e \quad (a_1 \neq 0, a_i \in \{0, \dots, d-1\})$$

$$\text{rd}(x) = v \cdot a^* d^e \quad \text{wobei } a^* = \begin{cases} a_1a_2\dots a_{\lfloor \log_d x \rfloor} & , 0 \leq a_{\lfloor \log_d x \rfloor} \leq d/2 \\ a_1a_2\dots a_{\lfloor \log_d x \rfloor} + d^{-1} & , a_{\lfloor \log_d x \rfloor} \geq d/2 \end{cases}$$

relativer Fehler = (relative) Maschinengenauigkeit $\text{eps} = \frac{d^{(1-e)}}{2}$

$x \delta y = \text{rd}(x \cdot y) = (x \cdot y)(1+\epsilon) \quad (\epsilon \leq \text{eps}) \quad \text{i.A. nicht assoziativ!}$

$x \hat{+} y = x + y \pm \text{eps} \rightarrow$ nächstgr. Zahl zu x : $x + 2\text{eps}$

lineare Gleichungssysteme

Problem: $A \in \mathbb{R}^{n \times n}$, n sehr groß

Finde Lösungsvektor x mit $Ax = b$

→ besitzt genau dann eine eindeutige Lösung x^* wenn A invertierbar.

Es ist dann $x^* = A^{-1}b$ nicht praktisch anwendbar

LR-Zerlegung (A invertierbar)

$$Ax = L \cdot R \cdot x = L \cdot y = b \quad \text{mit} \quad L = \text{untere Dr. matrix mit 1en auf Hauptdiag.}$$

Faktoren aus der Zeile 1 bis n

$R = \text{obere Dr. matrix}$

~ Löse $Ly = b$, dann $Rx = y$

In L stehen Eliminierungsfaktoren mit umgekehrtem VZ

Vorwärts-, Rückwärtssubstitution: 1 Operation = 1 Multiplikation + 1 Addition

$$\text{Aufwand: } \sum_{i=1}^{n-1} i = n \frac{n-1}{2} \approx \frac{n^2}{2} \text{ Operationen}$$

Wichtig: $r_{ii} \neq 0$; $i: i \neq 0$ für L, R Dreiecksmatrizen

→ A invertierbar, dann existiert (nicht eindeutige) Permutationsmatrix P derart, dass bzgl. P eindeutige Dreieckszerlegung $P \cdot A = L \cdot R$ möglich
 R ist obere Dreiecksmatrix, L untere Dreiecksmatrix mit 1en auf Hauptdiag.

Algorithmus:

(1) Bestimme P, L, R mit $PA = LR$ durch Gauß-Elimination

(2) Löse $Ly = Pb$ (Vorw. subst.)

(3) Löse $Rx = y$ (Rückw. subst.)

$$\text{~Aufwand: } \sum_{j=1}^{n-1} j^2 = \frac{(n-1)n(2n-1)}{6} \approx \frac{n^3}{3} \text{ Operationen}$$

~ Speicherplatz: 0,1 nicht explizit speichern, Perm. matrix als Vektor $\rightarrow n(n+1)$ Speicherplätze

Spaltenpivotwahl: wähle als Pivotelement im k -ten Schritt das Element $a_{jk}^{(k-1)}$ mit

$$|a_{jk}^{(k-1)}| = \max_{k \leq i \leq n} |a_{ik}^{(k-1)}| \text{ Betragmäßig größtes der "übrigen"}$$

~ bessere Stabilität

$$\text{Bsp.: } A = \begin{pmatrix} 0 & 4 & -1 \\ 2 & 2 & 1 \\ 1 & 1 & 5 \end{pmatrix} \rightsquigarrow \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \begin{bmatrix} 0 & 4 & -1 \\ 2 & 2 & 1 \\ 1 & 1 & 5 \end{bmatrix} \rightsquigarrow \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 5 \\ 2 & 2 & 1 \\ 0 & 4 & -1 \end{bmatrix} \cdot (2)$$

$$\rightsquigarrow \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 5 \\ 2 & 0 & -9 \\ 0 & 4 & -1 \end{bmatrix} \rightsquigarrow \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 5 \\ 0 & 4 & -1 \\ 2 & 0 & -9 \end{bmatrix}$$

$$\Rightarrow L = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix} \quad , \quad R = \begin{pmatrix} 1 & 1 & 5 \\ 0 & 4 & -1 \\ 0 & 0 & -9 \end{pmatrix}, \quad P = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \quad \text{ew als Zeilen}$$

Matrixnorm Jede Norm auf $\mathbb{R}^{n \times n}$ = Matrixnorm ; interessant: $\|Ax\| \leq \|A\| \|x\| \quad \forall x \in \mathbb{R}^n, A \in \mathbb{R}^{n \times n}$

zugehörige Matrixnorm: $\|\cdot\|$ beliebige Norm auf \mathbb{R}^n , dann: zugehörige Matrixnorm auf Raum der quadr. $(n \times n)$ -Matrizen ($A \in \mathbb{R}^{n \times n}$)

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{x \neq 0} \left\| \frac{Ax}{\|x\|} \right\| = \sup_{\substack{\|x\|=1 \\ \text{normiert}}} \|Ax\| = \max_{\|x\|=1} \|Ax\|$$

Es gilt:

- $\|Ax\| \leq \|A\| \|x\|, \forall x \in \mathbb{R}^n, \|A\| \text{ kleinste Zahl mit dieser Eigenschaft}$
- $\|I\| = 1$
- $\|A \cdot B\| \leq \|A\| \cdot \|B\| \quad (\text{Submultiplikativität})$

Sei A quadr. $(n \times n)$ -Matrix:

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

Spektralnorm: $\|A\|_2 = \sqrt{\text{gr. EW von } A^T A}$ symmetrisch \rightarrow nur reelle EW + pos. semidef.

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

Kondition linearer Gleichungssysteme

schätzt Einfluss von Störungen der Eingabegrößen A, b auf Lösung x ab

Störung der rechten Seite: \bar{x} Lösung des gestörten Systems, in dem b durch \bar{b} ersetzt ist
also: $A\bar{x} = \bar{b}$

> absolute Abweichung von \bar{x} zu x gemessen in der Norm $\|\cdot\|$:
 $\|\bar{x} - x\| = \|A^{-1}(b - \bar{b})\| \leq \|A^{-1}\| \|b - \bar{b}\|$

> relative Abweichung von \bar{x} zu x gemessen in der Norm $\|\cdot\|$:

$$\frac{\|\bar{x} - x\|}{\|x\|} \leq \underbrace{\frac{\|b\| \|A^{-1}\|}{\|x\|}}_{\approx \|b\| = \|Ax\| \leq \|A\| \|x\|} \underbrace{\frac{\|b - \bar{b}\|}{\|b\|}}_{\text{relative Störung der rechten Seite}} \quad \approx \frac{\|\bar{x} - x\|}{\|x\|} \leq \underbrace{\|A\| \|A^{-1}\|}_{=\text{cond}(A)} \frac{\|b - \bar{b}\|}{\|b\|}$$

Konditionszahl von A $\text{cond}(A) = \|A\| \|A^{-1}\|$ (Kondition einer Matrix)

\Rightarrow Maß der Sensitivität des rel. Fehlers gegenüber rel. Störung von b
 \rightarrow Sensitivität geringer je kleiner $\text{cond}(A)$

aber: $\text{cond}(A)$ nur obere Schranke dieser Sensitivität

> es gilt: $\lambda = \|A\| = \|A A^{-1}\| \leq \|A\| \|A^{-1}\| = \text{cond}(A)$

> reelles $A = a$: $\text{cond}(A) = 1$

> $\text{cond}(A) = \text{cond}(\alpha A), \alpha \in \mathbb{R} \setminus \{0\}$

$$\text{cond}(A) = \frac{\max_{\|y\|=1} \|Ay\|}{\min_{\|z\|=1} \|Az\|}$$

Rayleigh-Quotienten: A symmetrisch, dann mit $R_A(x) = \frac{x^T A x}{x^T x}$ ($x \neq 0$)

$$\min_{y \neq 0} R_A(y) = \lambda_{\min} \leq R_A(x) \leq \lambda_{\max} = \max_{y \neq 0} R_A(y)$$

$\Rightarrow B^T B$ (B beliebige Matrix) spezielle symm. Matrizen:

- nur EW ≥ 0
- nur $\text{EW} > 0$, falls $\text{rang}(B)$ maximal
- $\text{EW} \propto \lambda^2$, falls B symmetrisch mit EW λ

$$\Rightarrow \text{cond}_2(A) = \frac{\max \{|\lambda| \mid \lambda \text{ EW von } A\}}{\min \{|\lambda| \mid \lambda \text{ EW von } A\}}$$

Störungen von A und b: A invertierbar, $Ax = b$; $\tilde{A}\tilde{x} = \tilde{b}$

$$\text{und } \frac{\|A - \tilde{A}\|}{\|A\|} \leq \epsilon_A, \quad , \quad \frac{\|b - \tilde{b}\|}{\|b\|} \leq \epsilon_b \quad (\text{rel. Abweichungen beschränkt})$$

Dann gilt:

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \frac{\text{cond}(A)}{1 - \epsilon_A \cdot \text{cond}(A)} (\epsilon_A + \epsilon_b) \quad \text{falls } \epsilon_A \cdot \text{cond}(A) < 1$$

$$\text{mit } \epsilon_A, \epsilon_b \leq \epsilon: \frac{\|x - \tilde{x}\|}{\|x\|} \leq 2\epsilon \cdot \text{cond}(A) + O(\epsilon^2)$$

Stabilität der Gauß-Elimination

X mit Algorithmus berechnete Näherungslösung für $Ax = b$

Algorithmus numerisch stabil ...

... im Sinne der Vorderanalyse, falls $\frac{\|x - \tilde{x}\|}{\|x\|} \leq C \cdot \text{cond}(A) \cdot \text{eps}$ (C nicht zu groß)

... im Sinne der Rückwärtsanalyse, falls das numerische Ergebnis X als exakte Lösung einer gestörten Gleichung $\tilde{A}\tilde{x} = \tilde{b}$ interpretiert werden kann,
mit $\frac{\|A - \tilde{A}\|}{\|A\|} \leq C \cdot \text{eps}, \quad \frac{\|b - \tilde{b}\|}{\|b\|} \leq C \cdot \text{eps}$ (C nicht zu groß)

Anmerkung: $A \leq B \Leftrightarrow a_{ij} \leq b_{ij} \quad \forall i, j; \quad |A| = (a_{ij})_{i,j=1,\dots,n}$

Rückwärtsanalyse der Gauß-Elimination ohne Pivotwahl: $A \in \mathbb{R}^{n \times n}$ bei \mathbb{R}^n (Gleitpunktzahlen)
LR-Zeil. von A: \hat{L}, \hat{R} in Gleitpunkt-Arithmetik berechnet.

Das in Gleitp.-Ar. erhaltene X von $\hat{L}\hat{x} = b$ und $\hat{R}x = \hat{c}$ erfüllt $\tilde{A}\tilde{x} = b$ für Matrix \tilde{A} mit
 $|A - \tilde{A}| \leq 3(n+1)\text{eps}|\hat{L}| |\hat{R}| + O(\text{eps}^2)$

~ Für \hat{R} gilt: $\max_{i,j} |\hat{r}_{ij}| \leq 2^{n-1} \max_{i,j} |a_{ij}| \rightarrow$ schlecht!, aber bei zufällig gewählten A wird $\max_{i,j} |\hat{r}_{ij}| \approx n \cdot \max_{i,j} |a_{ij}|$ beobachtet

... mit Spaltenpivotwahl: Gauß-Elimination mit Spaltenpivotwahl berechnet X für $Ax = b$, sodass $\tilde{A}\tilde{x} = b$ mit

$$\frac{\|A - \tilde{A}\|_\infty}{\|A\|_\infty} \leq 2n^3 \frac{\max_{i,j} |\hat{r}_{ij}|}{\max_{i,j} |a_{ij}|} \text{eps} + O(\text{eps}^2)$$

= $\rho_n(A)$, bestimmender Faktor, aug: $\rho_n(A) \leq 2^{n-1}$

Spd-Matrix: $\rho_n(A) \leq 1$

triagonale Matrizen (Haupt+1 Nebendiag Einträge): $\rho_n(A) \leq 2$

Obere Hessenberg-Matrizen (ob Dreimatrix + untere Nebendiag): $\rho_n(A) \leq n$

Cholesky - Verfahren für spd - Matrizen (A invertierbar)

es existiert Zerlegung $A = LL^T$, wobei L untere Dreiecksmatrix (nicht zwingend ten auf H0).

$$\approx \text{Case } Ax = L \underbrace{L^T x}_{=: y} = b$$

(1) Vorwärtssubst. $Ly = b$

(2) Rückwärtssubst. $L^T x = y$

$A = LL^T$ existiert $\Leftrightarrow A$ ist spd-Matrix (A invertierbar)

! keine Spaltenpivotwahl durchführen!

Aufwand: $\approx \frac{1}{6} n^3$ (\approx halb so groß wie LR)

Berechnung: Ausmultiplizieren der Blockstruktur

$$A = \begin{pmatrix} A_{11} & A_{21} \\ A_{21}^T & A_{22} \end{pmatrix} = \begin{pmatrix} L_{11} & 0 \\ L_{21}^T & L_{22} \end{pmatrix} \begin{pmatrix} L_{11}^T & L_{21} \\ 0 & L_{22} \end{pmatrix} = LL^T$$

wobei A_{11}, L_{11} $(n-1) \times (n-1)$ -Matrizen; $a_{21}, l_{21} \in \mathbb{R}^{n-1}$, $a_{22}, l_{22} \in \mathbb{R}$

CCS - Format

$A \in \mathbb{R}^{N \times M}$ mit $K \leq \max\{N, M\}$ Einträge $\neq 0$

\rightsquigarrow 3 Vektoren $v \in \mathbb{R}^k$ = Einträge der Matrix

$z \in N^k$ = Zeilennummer des Eintrags in v mit gleichem Index.

$\mathbf{p} \in \mathbb{N}^{M+1}$ = Positionen der Spalten von \mathbf{A} in \mathbf{v}

Es gilt für $k \in \{1, \dots, N\}$, $l \in \{1, \dots, M\}$, $m \in \{1, \dots, K\}$:

$$A_{kl} = v_m \Leftrightarrow 2m=k \quad \text{und} \quad p_l \leq m < p_{l+1}$$

→ d.h. p_i gibt Index des ersten ≠ 0-Eintrags in v an, der zur i -ten Spalte von A gehört

$$\text{BGP: } A = \begin{pmatrix} 1 & 0 & 0 & 2 \\ 3 & 4 & 0 & 0 \\ 0 & 6 & 0 & 5 \\ 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 8 \end{pmatrix} \quad v = (1, 3, 4, 6, 7, 2, 5, 8)^T$$

$$z = (1, 2, 2, 3, 4, 1, 2, 4)^T$$

$$p = (1, 3, 6, 6, 9)^T$$

QR-Zerlegung $A \in \mathbb{R}^{m \times n}$ $m \geq n$

konstruiere Zerlegung $A = Q \cdot R$ mit orthog. Matrix $Q \in \mathbb{R}^{m \times m}$ ($QQ^T = I$) und
 $R = \begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix} \in \mathbb{R}^{m \times n}$, $\tilde{R} \in \mathbb{R}^{n \times n}$ obere Dreiecksmatrix

(konstruktion zB mit Householder - Transformation)

Fall $n = m$: löse LGS $Ax = b$

- (1) Bestimme Q, R mittels Househ.-Transf. mit $A = Q \cdot R$
 (2) Löse $Qc = b$ ($Q^{-1} = Q^T$, also $Q^T c = Q^T b$)
 (3) Löse $Rx = c$

→ besonders stabil, aber doppelt so viele Ops wie Gauß-Elimination

Lineare Ausgleichsprobleme $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$

Fall $m > n \rightarrow Ax = b$ überbestimmt, im Alg. keine Lösung

~ Suche nach $x \in \mathbb{R}^n$ mit $\|Ax - b\|_2 = \min$

$x \in \mathbb{R}^n$ Lösung von $\|Ax - b\|_2 = \min \Leftrightarrow x$ löst Normalengleichung $A^T A x = A^T b$

~ Lin. Ausgl. problem eindeutig lösbar $\Leftrightarrow \text{Rang}(A) = n$ (maximal) ~ $A^T A$ spd

LGS $A^T A x = A^T b$ lösbar mit Cholesky für A mit max. Rang

\rightarrow für $A \in \mathbb{R}^{m \times n}$ mit max. Rang $n \leq m$ gilt: $\text{cond}_2(A^T A) = (\text{cond}_2(A))^2$

$\Rightarrow A \in \mathbb{R}^{m \times n}, m \geq n$, A hat vollen Rang, $b \in \mathbb{R}^m$, Q und R aus QR-Zerl. von A , d.h. $Q^T A = R = \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix}$ (\hat{R} invertierbar, $\in \mathbb{R}^{n \times n}$)

Dann ist $x = \hat{R}^{-1} c$ Lösung von $\|Ax - b\|_2 = \min$

mit c def. durch $Q^T b = (d)$

~ $\|r\|_2 = \|Ax - b\|_2 = \|d\|_2$

Algorithmus: (1) Bestimme Q, R mittels Househ. mit $A = QR$

(2) Berechne $Q^T b = (d)$

(3) Löse $\hat{R}x = c$ (Rückw. Subst.)

Nichtlineare Gleichungssysteme

Problem: $f: D \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$, finde $x \in D$ mit $f(x) = 0 \iff$

~ i.A. Näherungslösungen mit Rundungsfehlern + Abbruchfehlern

$$f_1(x_1, \dots, x_n) = 0 \\ \vdots \\ f_n(x_1, \dots, x_n) = 0$$

Fixpunktiteration: Für $F: D \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ finde $x \in D$ mit $F(x) = x$

Element $x^* \in D$ heißt Fixpunkt von F , falls $F(x^*) = x^*$.

~ 1D-Fall: Fixpunkte sind Schnittpunkte von Graph + Winkelhalbierende

~ Nullstellengleichung $f(x) = 0$ äquivalent zur Fixpunktgleichung $F(x) = x$, mit $F(x) = x - A \cdot f(x)$ ($A \in \mathbb{R}^{n \times n}$ invertierbar)

⇒ Forme Nullstellengleichung in Fixpunktgleichung um. Berechne Folge $(x_i)_{i \in \mathbb{N}}$ ausgehend von Startwert x_0 gemäß $x_{k+1} = F(x_k)$
→ Folge konvergiert gegen Fixpunkt x^*

Kontraktion: Abbildung $F: D \rightarrow D \subset \mathbb{R}^n$ ist Kontraktion auf D , falls $0 < \theta < 1$ existiert

$$\text{mit } \|F(x) - F(y)\| \leq \theta \|x - y\| \quad \forall x, y \in D$$

d.h.: Abstand der Bildpunkte von x_k ist kleiner als Abstand von x und y
→ θ konverg. Koeffizient θ über Ableitung von F

Banachscher Fixpunktatz: $F: D \rightarrow D$ Kontraktion auf D , D abgeschlossene Teilmenge des \mathbb{R}^n , mit Kontraktionszahl $0 < \theta < 1$

Dann gilt:

(1) es existiert genau ein Fixpunkt x^* von F

(2) Die Folge $x_{k+1} = F(x_k)$ konvergiert gegen x^* für jeden Startwert $x_0 \in D$

(3) Es gelten die Abschätzungen

$$\|x^* - x_k\| \leq \theta \|x^* - x_{k-1}\| \quad \text{lineare Konvergenz}$$

$$\|x^* - x_k\| \leq \frac{\theta^k}{1-\theta} \|x_0 - x_1\| \quad \text{A-priori-Abschätzung}$$

$$\|x^* - x_k\| \leq \frac{\theta}{1-\theta} \|x_{k-1} - x_k\| \quad \text{A-posteriori-Abschätzung}$$

Newton-Verfahren:

Algorithmus: (1) wähle Startwert x_0

(2) while ($\|\Delta x\| > \text{TOL}$) do

 löse $f'(x_k) \Delta x_k = -f(x_k)$ (LGS, LR-Zerlegung)

 Berechne $x_{k+1} = x_k + \Delta x_k$

dabei ist $f'(x_k)$ die Jacobi-Matrix:

$$f'(x_k) = \begin{pmatrix} \frac{\partial f_1(x_k)}{\partial x_1} & \dots & \frac{\partial f_1(x_k)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n(x_k)}{\partial x_1} & \dots & \frac{\partial f_n(x_k)}{\partial x_n} \end{pmatrix}$$

$$\text{z.B. } f(x) = f(x_1, \dots, x_n) = f(x, y, z) =$$

$$= \begin{pmatrix} x^2+y^2+z \sin(x) \\ z^2+x \sin(y) \end{pmatrix}$$

$$\Rightarrow f'(x, y, z) = \begin{pmatrix} 2x+z \cos(x) & 2y & \sin(x) \\ 0 & z \cos(y) & 2z+\sin(y) \end{pmatrix}$$

- Abbruchkriterium:
- 1) $\|\Delta x_k\| > \text{TOL}$ sinnvoll, da gutes Indiz für Änderung $\rightarrow \|\Delta x_k\| < \text{TOL} \Rightarrow$ kaum noch Änderung
 - 2) $\|f(x_k)\| \leq \text{TOL}$ nicht sinnvoll, da mit $A \cdot f(x) / \|A \cdot f(x)\|$ ($A \neq 0$) Lösung nicht verändert wird, Abbruchkriterium aber beliebig manipulierbar

Geometrische Deutung: Lege Tangente mit Steigung $f'(x_0)$ an Graphen in Punkt $(x_0, f(x_0))$
 \rightarrow Schnittpunkt mit x -Achse ist neue Iterierte x_1
 $\rightarrow \dots$

lokale quadratische Konvergenz: $D \subset \mathbb{R}^n$ offen, $f: D \rightarrow \mathbb{R}^n$ zweimal stetig diff. bar
 Es existiere ein $x^* \in D$ mit $f(x^*) = 0$, Jacobi-Matrix $f'(x^*)$ invertierbar

- (1) es gibt eine Kugel $K = K_p(x^*) = \{x \in \mathbb{R}^n \mid \|x - x^*\|_\infty \leq p\} \subset D$
 so dass x^* einzige Nullstelle von f in K
 $\Rightarrow p$ sehr klein \Rightarrow Startwert zu weit von Nullstelle entfernt: divergierendes NV
- (2) Folgeglieder $x_{k+1} = x_k - f'(x_k)^{-1} f(x_k)$ für alle $k \in \mathbb{N}$ auch in K
- (3) $\lim_{k \rightarrow \infty} x_k = x^*$
- (4) es ex. $C > 0$, $C = \text{const.}$ mit $\|x^* - x_{k+1}\| \leq C \|x^* - x_k\|^2$ ($k \in \mathbb{N}$)
 \rightarrow quadratische Konvergenz

Vereinfachtes Newton-Verfahren

Ersetze teure Ableitung f' durch konstante Matrix $A \approx f'(x_0)$
 \rightarrow nur noch lineare Konvergenz

Algorithmus:

- (1) Wähle Startwert x_0 , berechne QR-Zerlegung von $A \approx f'(x_0)$
- (2) while ($\|\Delta x_k\| > \text{TOL}$) do
 - Löse $A \Delta x_k = -f(x_k)$
 - Berechne $x_{k+1} = x_k + \Delta x_k$

cg-Verfahren für LGS

Löse $Ax = b$ approximativ, mit $A \in \mathbb{R}^{n \times n}$ spd-Matrix und dünn besetzt

Idee: minimiere Funktional $\phi(x) = \frac{1}{2} x^T A x - x^T b$, dann:

x genau dann Lösung von $\phi(x) = \min_{z \in \mathbb{R}^n} \phi(z)$, wenn $Ax = b$

Energienorm: $A \in \mathbb{R}^{n \times n}$ spd-Matrix, Norm auf \mathbb{R}^n (Energienorm):

$$\|x\|_A = \sqrt{x^T A x} ; x \in \mathbb{R}^n$$

entspr. Skalarprodukt:

$$\langle x, y \rangle_A = x^T A y ; x, y \in \mathbb{R}^n$$

→ Für Lösung x^* von $Ax = b$ gilt:

$$\phi(x) - \phi(x^*) = \frac{1}{2} \|x - x^*\|_A^2 \quad (\text{Abweichung des Funktional von seinem Minimum})$$

⇒ HAUPTIDEE: Löse iterativ eindim. Minimierungsprobleme um ϕ zu minimieren

→ Ausgehend von Vektor $x^k \in \mathbb{R}^n$ und Suchrichtung $d^k \in \mathbb{R}^n$ berechne neue Iterierte + minimiere Funktional

→ in jeder Iteration wird Funktional in eine Richtung minimiert

⇒ nach höchstens n Schritten liefert cg-Verfahren exakte Lösung von $Ax = b$

praktisch uninteressant: Rundungsfehler verhindern exakte Lsg.

+ n sehr groß → erwartete brauchbare Näherung für kleinen

Algorithmus:

- (1) wähle beliebiges $x^0 \in \mathbb{R}^n$, def. $d^0 = r^0 = b - Ax^0$ und $k=0$
- (2) while $(\|r^k\| > \text{TOL} \cdot \|b\|)$ berechne iterativ

$$(a) \alpha_k = \frac{\langle r^k, r^k \rangle}{\langle d^k, d^k \rangle} \text{ eukl. Sch}$$

$$(b) x^{k+1} = x^k + \alpha_k d^k$$

$$(c) r^{k+1} = r^k - \alpha_k A d^k$$

$$(d) \beta_k = \frac{\langle r^{k+1}, r^{k+1} \rangle}{\langle r^k, r^k \rangle} \text{ eukl. Sch}$$

$$(e) d^{k+1} = r^{k+1} + \beta_k d^k$$

f) Erhöhe k um 1

Fehlerabschätzungen: q_k beliebiges Polynom mit Grad $\leq k$, $q(0)=1$, dann gilt:

$$\|x^* - x^k\|_A \leq \max_{\lambda \in \text{EW von } A} |q_k(\lambda)| \cdot \|x^* - x^0\|_A$$

Sei x^* Lsg. des LGS, dann: $\|x^* - x^k\|_A \leq 2 \left(\frac{\gamma_{k+1}}{\gamma_k} \right)^k \|x^* - x^0\|_A \quad k=1, 2, \dots$

$$\text{mit } \gamma_k = \text{cond}_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \geq 1$$

$\gamma < 1$ bestimmt Konvergenzgeschw
→ konv. schlecht für große $\text{cond}_2(A)$

→ bessere Konvergenz durch Vorkonditionierung

Interpolation und Approximation

Problem:

f_i : Stützwert
 x_i : Stützstellen

> Interpolation: suche für Stützpunkte $(x_0, f_0), \dots, (x_n, f_n)$ mit $p(x_i) = f_i$ ein Polynom $p(x)$ vom Grad $\leq n$

> Approximation: suche für geg. $f: [a, b] \rightarrow \mathbb{R}$ möglichst einfache auszuwertende Funktion $p: [a, b] \rightarrow \mathbb{R}$, so dass z.B.

- $\int_a^b (f(x) - p(x))^2 dx = \min$
- $\max_{x \in [a, b]} |f(x) - p(x)| = \min$
- zusätzlich zu (a) oder (b): $f(x) = p(x)$ für endlich viele Punkte

Weierstraßscher Approximationssatz: $f: [a, b] \rightarrow \mathbb{R}$ stetig, dann ex. für jedes $\epsilon > 0$ ein $n \in \mathbb{N}$ und ein Polynom $p: [a, b] \rightarrow \mathbb{R}$ mit Grad n , so dass

$$\max_{x \in [a, b]} |f(x) - p(x)| \leq \epsilon$$

Polynominterpolation

no Polynom vom Grad $\leq n$ ist durch $n+1$ Punkte (an paarweise versch. Stellen) eindeutig bestimmt

Lagrangesche Interpolationsformel: zu $n+1$ Stützpunkten (x_i, f_i) ($i=0, \dots, n$) mit paarweise versch. Stützstellen x_i ex. genau ein Interpolationspolynom $p(x)$ vom Grad $\leq n$, welches gegeben ist durch

$$p(x) = \sum_{i=0}^n f_i \cdot l_i(x) \quad \text{wobei } l_i(x) = \prod_{j=0, j \neq i}^n \frac{x - x_j}{x_i - x_j} \quad \begin{matrix} \text{Lagrange-} \\ \text{Polynome} \\ \text{der} \\ \text{Stützstellen} \end{matrix}$$

$$\text{es gilt: } l_i(x_j) = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}$$

Lagrange-Konstante: $\Delta_n = \max_{x \in [a, b]} \sum_{i=0}^n |l_i(x)|$ invariant unter affinen Transformationen, nur abh. von relative Lage der x_i ; Zueinander

Kondition: $p(x), \tilde{p}(x)$ Interpolationspolynome zu (x_i, f_i) bzw. (x_i, \tilde{f}_i) ($i=0, \dots, n$), dann gilt: $\max_{x \in [a, b]} |p(x) - \tilde{p}(x)| \leq \Delta_n \cdot \max_{i=0, \dots, n} |f_i - \tilde{f}_i|$

wobei Δ_n die kleinste Zahl mit dieser Eigenschaft.

Newton'sche Interpolationsformel / Dividierte Differenzen

$p_{i,k}(x)$ IP zu Stützpunkten $(x_i, f_i), \dots, (x_k, f_k)$

Darstellung $p(x)$ mit Grad $\leq n$ zur Newton-Basis $\{w_0(x), \dots, w_n(x)\}$:

$$p(x) = a_0 + a_1 w_1(x) + \dots + a_n w_n(x) \rightarrow \text{auswerten mit Horner-Schema}$$

$$\text{wobei } w_i(x) = \prod_{j=0}^{i-1} (x - x_j)$$

$$\begin{aligned} p(x) &= a_0 + (x - x_0)(a_1 + (x - x_1)(a_2 + \dots + (x - x_{n-2})(a_{n-1} + (x - x_{n-1})a_n))) \end{aligned}$$

Die a_i sind die Koeffizienten bzgl. der Newton-Basis, genannt **dividierte Differenzen von f zu den Stützstellen x_0, \dots, x_n**

$$\text{also: } p_{0,n}(x) = f_{0,0} + f_{0,1} w_1(x) + \dots + f_{0,n} w_n(x)$$

$$\text{Lemma von Aitken: } p_{0,n}(x) = \frac{(x_0 - x) p_{1,n}(x) - (x_n - x) p_{0,n-1}(x)}{(x_0 - x_n)}$$

Newton'sche Interpolationsformel: Zu $n+1$ Stützpunkten (x_i, f_i) ex. mit paarweise versch. x_i , existiert genau ein IP $p(x)$ vom Grad $\leq n$:

$$p(x) = f_{0,0} + f_{0,1}(x-x_0) + \dots + f_{0,n}(x-x_0) \cdots (x-x_{n-1})$$

$$\text{mit } f_{i,i} = f_i \quad f_{i,i+k} = \frac{f_{i,i+k-1} - f_{i+1,i+k}}{x_i - x_{i+k}} \quad i=0, \dots, n \\ 1 \leq k \leq n-i$$

Differenzenschema:

$$\begin{array}{ccccccc} f_0 & = & f_{0,0} & & & & \\ & & \searrow & & & & \\ f_1 & = & f_{1,1} & \rightarrow & f_{0,1} & \searrow & \\ & & \searrow & & & & \\ f_2 & = & f_{2,2} & \rightarrow & f_{1,2} & \rightarrow & f_{0,2} \\ & & \vdots & & \vdots & & \ddots \\ & & f_{n-1} & = & f_{n-1,n-1} & \rightarrow & f_{0,n-1} \\ & & f_n & = & f_{n,n} & \rightarrow & f_{1,n} \rightarrow f_{0,n} \end{array}$$

Interpolationsfehler: untersuche Fehler $f(x) - p(x)$

$f: [a,b] \rightarrow \mathbb{R}$ mind. $(n+1)$ -mal stetig diff. bar., $p(x)$ IP von f in Stellen $x_0, \dots, x_n \in [a,b]$
 $\deg(p(x)) \leq n$.

Dann ex. zu jedem $x \in [a,b]$ eine Zwischenstelle $\xi = \xi(x) \in (a,b)$ mit

$$f(x) - p(x) = w_{n+1}(x) \cdot \underbrace{\frac{f^{(n+1)}(\xi)}{(n+1)!}}_{\text{w. } \xi = \sum_{j=0}^{i-1} (x-x_j)} = f_{0,n+1} \text{ mit } x_{n+1} = \xi + y_i$$

Tschebyscheff-Interpolation

Ziel: Approximation von $f: [a,b] \rightarrow \mathbb{R}$ durch IP mit möglichst guten Stützstellen \rightarrow gute Kondition, optimale Approximation

Sei $[a,b] = [-1,1]$ affine Transformation: $[-1,1] \leftrightarrow [a,b]$

$$x \mapsto \frac{a+b}{2} + \frac{b-a}{2}x = y$$

Tschebyscheff-Polynome: $T_0(x) = 1$

$$T_1(x) = x$$

$$T_{n+1}(x) = 2x T_n(x) - T_{n-1}(x)$$

$$\text{oder } T_n(x) = \cos(n \cdot \arccos(x))$$

$$\Rightarrow \text{(a) Nullstellen von } T_n: \cos\left(\frac{2k+1}{2n}\pi\right) \quad k=0, \dots, n-1$$

$$\text{(b) } T_n\left(\cos\left(\frac{k\pi}{n}\right)\right) = (-1)^k \quad k=0, \dots, n$$

$$\text{(c) } |T_n(x)| \leq 1 \quad (|x| \leq 1)$$

$$\text{(d) Koeffizient von } x^n \text{ ist } 2^{n-1} \text{ für } n \geq 1$$

unter allen $(x_0, \dots, x_n)^\top \in \mathbb{R}^{n+1}$ wird $\max_{x \in [-1, 1]} |w_{n+1}(x)|$ minimal, wenn die x_i Nullstellen von T_{n+1} sind, d.h. wenn $x_k = \cos\left(\frac{2k+1}{2n+2}\pi\right)$ $k=0, \dots, n$.
der minimale Wert ist $\frac{1}{2^n}$.

$$w_{n+1}(x) = \frac{T_{n+1}}{2^n}$$

Lebesgue-Konstanten zu Tschebyscheff-Stützstellen:

$$\Delta_n \leq 3, \quad n \leq 20$$

$$\Delta_n \leq 4, \quad n \leq 100$$

$$\Delta_n \propto \frac{2}{\pi} \log n, \quad n \rightarrow \infty$$

Tschebyscheff-Interpolationsformel: $n+1$ Stützpunkte (x_i, f_i) ($i=0, \dots, n$), Stützstellen = Nullstellen von

$$\text{IP } p(x) = p_{0,n}(x) \text{ (eindeutig, Grad } \leq n): \quad p(x) = \frac{1}{2} c_0 + c_1 T_1(x) + \dots + c_n T_n(x)$$

$$\text{mit } c_k = \frac{2}{n+1} \sum_{i=0}^n f_i \cos\left(k \frac{2i+1}{2n+2}\pi\right) \quad k=0, \dots, n$$

$$\Rightarrow x_i = \cos\left(\frac{2i+1}{2n+2}\pi\right)$$

Clenshaw-Algorithmus: $p(x) = \frac{1}{2} c_0 + c_1 T_1(x) + \dots + c_n T_n(x)$,

$$d_{n+2} = d_{n+1} > 0 \quad \text{und} \quad d_k = c_k + 2x \cdot d_{k+1} - d_{k+2} \quad (k=n, n-1, \dots, 0)$$

$$\text{dann: } p(x) = \frac{1}{2} (d_0 - d_2)$$

Stabilität: sei $0 = \tilde{d}_{n+2} = \tilde{d}_{n+1}$; $\tilde{d}_k = c_k + 2x \cdot \tilde{d}_{k+1} - \tilde{d}_{k+2} + \varepsilon_k$ ($k=n, n-1, \dots, 0$)
(ε_k z.B. Rundungsfehler in Schritt k)

$$\text{dann: } \underbrace{\left(\frac{1}{2} (\tilde{d}_0 - \tilde{d}_2) - p(x) \right)}_{= \tilde{p}(x)} = |\tilde{p}(x) - p(x)| \leq \sum_{i=0}^n |\varepsilon_i|$$

Linearer Aufwand

Kubische Spline-Interpolation

Problem: suche glatte Funktion $s: [a, b] \rightarrow \mathbb{R}$, die durch vorgegebene Punkte $a = x_0 < x_1 < \dots < x_n = b$ geht

glatt: s zweimal stetig diffbar und $\int_a^b |s''(x)|^2 dx$ minimal

Min-Eigenschaft kubischer Splines $a = x_0 < x_1 < \dots < x_n = b$, $f, s: [a, b] \rightarrow \mathbb{R}$ 2mal stetig diff. bar mit $s(x_i) = f(x_i)$, $i=0, \dots, n$

+ es gelte eine der Bedingungen

$$(i) \quad s'(a) = f'(a) \quad \text{und} \quad s'(b) = f'(b) \quad \text{eingespannter natürlicher}$$

$$(ii) \quad s''(a) = 0 \quad \text{und} \quad s''(b) = 0$$

$$(iii) \quad s^{(k)}(a) = s^{(k)}(b) \quad \text{für } k=1, 2 \quad \text{und} \quad f^{(k)}(a) = f^{(k)}(b) \quad \text{periodischer für } k=0, 1$$

} interpolierender kubischer Spline

Falls s auf jedem Teilintervall $[x_{i-1}, x_i]$ kubisches Polynom, gilt:

$$\int_a^b |s''(x)|^2 dx \leq \int_a^b |f''(x)|^2 dx$$

Def. kubischer interpol. Spline: $s: [x_{i-1}, x_i] \in P_3, i=1, \dots, n$ $s \in C^2, s(x_i) = f_i$ ($i=0, \dots, n$)

Konstruktion: $a = x_0 < x_1 < \dots < x_n = b$, (x_i, y_i) gegeben

fordere: $s_i := s|_{[x_{i-1}, x_i]}$ Polynom dritten Grades mit

$$s_i(x_{i-1}) = y_{i-1}$$

$$s_i(x_i) = y_i$$

sorgt für $= 0$ an Rändern

$$\rightsquigarrow s_i(x) = \underbrace{y_{i-1} + (x - x_{i-1}) y_{i-1,i}}_{\text{interpolierender Anteil}} + \underbrace{(x - x_{i-1})(x - x_i)}_{\text{glättender Anteil}} [\alpha_i (x - x_{i-1}) + \beta_i (x - x_i)]$$

$$y_{i-1,i} = \frac{y_{i-1} - y_i}{(x_{i-1} - x_i)}$$

$\rightsquigarrow n$ Teilintervalle $\rightarrow n$ kubische Funktionen mit 2 Unbekannten $\rightarrow 2n$ Unbekannte

Bedingungen: $s_i(x_i) = y_i$

$$s_i(x_{i+1}) = y_{i+1} \quad i = 0, \dots, n-1 \quad \left. \begin{array}{l} \\ \end{array} \right\} 2n \text{ Bedingungen}$$

$$s_i'(x_{i+1}) = s_{i+1}'(x_{i+1})$$

$$s_i''(x_{i+1}) = s_{i+1}''(x_{i+1})$$

+ 2 Bed. eingesp. / nat. / per.

$$\rightsquigarrow \text{def. } y_{i-1} = s_i''(x_{i-1})$$

$$y_i = s_i''(x_i)$$

$$d_i = y_{i+1} - y_{i-1,i}$$

$$d_0 = y_{0,1} - v_0 = y_{0,1} - s_1'(x_0)$$

$$d_n = v_n - y_{n-1,n} = s_n'(x_n) - y_{n-1,n}$$

h_i = Gitterweiten

$$\Rightarrow \frac{1}{6} \begin{pmatrix} 2h_1 & h_1 & & & & \\ h_1 & 2(h_1+h_2) & h_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & h_{n-1} & 2(h_{n-1}+h_n) & h_n & \\ & & & h_n & z_{h_n} & \\ & & & & & \end{pmatrix} \begin{pmatrix} f_0 \\ y_1 \\ \vdots \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix} = \begin{pmatrix} d_0 \\ d_1 \\ \vdots \\ \vdots \\ d_{n-1} \\ d_n \end{pmatrix}$$

\rightarrow eindeutig lösbar + gut konditioniert

Kondition:

$s(x)$ eingesp. interpol. Spline zu (x_i, y_i) , $\tilde{s}(x)$ zu gestörten Daten (x_i, \tilde{y}_i)

$$\text{es gilt: } s(x) - \tilde{s}(x) = \sum_{i=0}^n (y_i - \tilde{y}_i) l_i(x)$$

mit $l_i(x_j) = \begin{cases} 1 & : i=j \\ 0 & : i \neq j \end{cases}$ eingesp. Lagrange-Spline, $l_i'(a) = 0$, $l_i'(b) = 0$

$$\Rightarrow |s(x) - \tilde{s}(x)| \leq \sum_{i=0}^n |y_i - \tilde{y}_i| l_i(x) \leq \Delta_n \max_{i=0, \dots, n} |y_i - \tilde{y}_i|$$

$$\text{mit } \Delta_n = \max_{x \in [a, b]} \sum_{i=0}^n \|l_i(x)\|$$

\rightarrow bei aquidistenter Unterteilung gilt $\Delta_n \leq 2 \quad \forall n \in \mathbb{N}$

Fehlerabschätzung: $f \in C^4[a,b]$, s eingesp. interp. Spline

dann: $|f(x) - s(x)| \leq \frac{5}{384} h^4 \max_{\xi \in [a,b]} |f^{(4)}(\xi)|$

mit $h = \max_{1, \dots, n} h$:

entscheidender Faktor

durch Problem gegeben

Bézier-Technik

- Speichern, berechnen, zeichnen von Polynomen effizient machen

Polynom in \mathbb{R}^d : Polynom vom Grad n in \mathbb{R}^d bzw. polynomiale Kurve in \mathbb{R}^d ist Abbildung $p: \mathbb{R} \rightarrow \mathbb{R}^d$ mit

$$p(x) = \sum_{k=0}^n a_k x^k, \quad a_k \in \mathbb{R}^d \text{ für } k=0, \dots, n; \quad a_n \neq 0$$

→ Polynom ist in jeder Komponente Polynom vom Grad $\leq n$

→ Menge \mathbb{P}_n^d ist Menge der Polynome vom Grad $\leq n$ in \mathbb{R}^d , $\dim(\mathbb{P}_n^d) = d(n+1)$

→ Graph Γ_p eines $p \in \mathbb{P}_n^d$, $\Gamma_p: \mathbb{R} \rightarrow \mathbb{R}^{d+1}$, $x \mapsto (x, p(x))$

auffassbar als Polynom in \mathbb{R}^{d+1} : $\Gamma_p(x) = (a_0^0) + (a_1^1)x + (a_2^2)x^2 + \dots + (a_n^n)x^n$

Z.B.: $p: \mathbb{R} \rightarrow \mathbb{R}$, $p(x) = 1 - 7x + 14x^2 - 9x^3$

→ Graph in \mathbb{R}^2 : $\Gamma_p(x) = (1) + (-7)x + (14)x^2 - (9)x^3$

Bernstein-Polynome, Bernstein-Darstellung

Bernstein-Polynome: Das i -te Bernstein-Polynom vom Grad n bezüglich $[0,1]$ ist

$$B_i^n(x) = \binom{n}{i} (1-x)^{n-i} x^i \quad i=0, \dots, n \quad \binom{n}{i} = \frac{n!}{i!(n-i)!}$$

Eigenschaften:

(i) $B_i^n(x) \geq 0$ auf $[0,1]$ ($i=0, \dots, n$)

(ii) Symmetrie: $B_i^n(x) = B_{n-i}^n(1-x)$ ($i=0, \dots, n$)

(iii) B_i^n bilden Partition der Eins: $1 = \sum_{i=0}^n B_i^n(x)$

(iv) B_i^n hat in $[0,1]$ genau ein Maximum, bei $x = \frac{i}{n}$

(v) B_i^n Basis von \mathbb{P}_n^d

→ für $i=1, \dots, n-1$, $x \in \mathbb{R}$ gilt:

$$B_i^n(x) = x B_{i-1}^{n-1}(x) + (1-x) B_i^{n-1}(x)$$

→ Polynom $p \in \mathbb{P}_n^d$ in Bernstein-Darstellung: $p(x) = \sum_{i=0}^n b_i B_i^n(x)$

$b_i \in \mathbb{R}^d$ Kontrollpunkte von p , der durch sie verlaufende Streckenzug heißt Bézier-Polygon

Die Kontrollpunkte von $p(x) = x$ sind $0, \frac{1}{n}, \frac{2}{n}, \dots, 1$

⇒ Bernstein-Darstellung von $p(x) = x$: $\Gamma_p(x) = \sum_{i=0}^n \binom{i/n}{b_i} B_i^n(x)$

Weitere Eigenschaft der $B_i^n(x)$:

$$B_i^0(0) = 0 \quad B_i^n(0) = 1 \quad i=1, \dots, n$$

$$B_j^n(1) = 0 \quad B_n^n(1) = 1 \quad j = 0, \dots, n-1$$

$$\Rightarrow p(0) = b_0$$

$$p(1) = b_n \quad (\text{auf } [0,1])$$

• Polygonzug im Anfangs- und Endpunkt tangential zur Kurve im \mathbb{R}^d

Ableitung der Bernstein-Polynome:

$$\frac{d}{dx} B_i^n(x) = \begin{cases} -n B_0^{n-1}(x) & , i=0 \\ n(B_{i-1}^{n-1}(x) - B_i^{n-1}(x)) & , i=1, \dots, n-1 \\ n B_{n-1}^{n-1}(x) & , i=n \end{cases}$$

$$\Rightarrow p'(0) = n(b_1 - b_0)$$

$$p'(1) = n(b_n - b_{n-1})$$

~ Das Bild $p([0,1])$ von $p \in \mathbb{P}_n^d$, $p(x) = \sum_{i=0}^n b_i B_i^n$ liegt in der konvexen Hülle der Kontrollpunkte b_i : $p(x) \in \text{conv}\{b_0, \dots, b_n\}, x \in [0,1]$

$$\text{conv}\{x_1, \dots, x_n\} = \{x \mid x = \sum_{i=1}^n \lambda_i x_i, \lambda_i \geq 0 \text{ und } \sum_{i=1}^n \lambda_i = 1\}$$

Algorithmus von de Casteljau

zerlege Polynom in Polynome niedrigeren Grades:

$$\text{für } p \in \mathbb{P}_n^d, p(x) = \sum_{i=0}^n b_i B_i^n(x) : \beta_i^k(x) = \sum_{j=0}^k b_{i+j} B_j^k(x) \quad ; k=0, \dots, n; i=0, \dots, n-k$$

β_i^k sind Teilpolynome, $\beta_i^k \in \mathbb{P}_k^d$, $k = \text{Grad}$

$$\rightarrow \beta_0^0(x) = p(x)$$

$$\beta_i^0 = b_i$$

$$\text{rekursive Berechnung: } \beta_i^k(x) = (1-x)\beta_i^{k-1}(x) + x\beta_{i+1}^{k-1}(x) \quad ; k=0, \dots, n; i=0, \dots, n-k$$

→ für $x \in [0,1]$ stabil, da Störungen gedämpft werden

$$b_0 = \beta_0^0 \quad \rightarrow$$

$$b_1 = \beta_1^0 \quad \rightarrow \quad \beta_0^1 \quad \rightarrow$$

$$b_2 = \beta_2^0 \quad \rightarrow \quad \beta_1^1 \rightarrow \beta_0^2$$

:

:

$$b_{n-1} = \beta_{n-1}^0 \quad \rightarrow \quad \beta_{n-2}^1 \rightarrow \dots \rightarrow \beta_0^{n-1}$$

$$b_n = \beta_n^0 \quad \rightarrow \quad \beta_{n-1}^1 \rightarrow \dots \rightarrow \beta_1^{n-1} \rightarrow \beta_0^n$$

Bernstein-Polynome bzgl. beliebiger Intervalle

- i-tes Bernstein-Polynom vom Grad n bzgl. $[a, b]$:

$$B_i^n(x; a, b) = \frac{1}{(b-a)^n} \binom{n}{i} (b-x)^{n-i} (x-a)^i \quad i = 0, \dots, n$$

rekursiv: $B_i^k(x; a, b) = \frac{x-a}{b-a} B_{i-1}^{k-1}(x; a, b) + \frac{b-x}{b-a} B_{i+1}^{k-1}(x; a, b) \quad i = 1, \dots, k-1$

Teilpolynome auf $[a, b]$: , $k=0, \dots, n$, $i = 0, \dots, n-k$

$$\beta_i^k(x; a, b) = \frac{b-x}{b-a} \beta_i^{k-1}(x; a, b) + \frac{x-a}{b-a} \beta_{i+1}^{k-1}(x; a, b)$$

→ bessere Approximation: Unterteilung in Teilintervalle

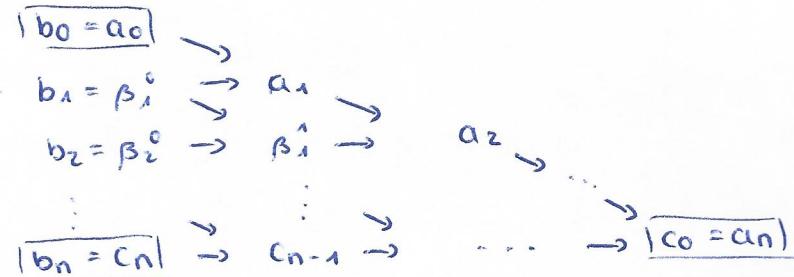
$$p(x) = \sum_{i=0}^n b_i B_i^n(x; a, b) \text{ zu } [a, b]$$

dann: $p(x) = \sum_{i=0}^n a_i B_i^n(x; a, c) = \sum_{i=0}^n c_i B_i^n(x; c, b)$

bzgl. $[a, c]$ und $[c, b]$ ($c \in (a, b)$)

mit $a_k = \beta_0^k(c; a, b)$ und $c_k = \beta_k^{n-k}(c; a, b) \quad k=0, \dots, n$

Schema:



→ Bézier-Polygone konvergieren bei wiederholter Unterteilung gegen die Kurve

Nummerische Integration

Problem: berechne für $f: [a, b] \rightarrow \mathbb{R}$: $I(f) = \int_a^b f(x) dx$

Rechteckregel: approx. $I(f)$ durch Rechteck: $I(f) \approx (b-a)f(a)$

Mittelpunktsregel: werte Fkt. im Mittelpunkt aus: $I(f) \approx (b-a) f\left(\frac{a+b}{2}\right)$

Trapezregel: lineare Fkt. durch Punkte $(a, f(a)), (b, f(b))$: $I(f) \approx (b-a) \frac{f(a)+f(b)}{2}$

Simpsonregel: Parabel durch $(a, f(a)), \left(\frac{a+b}{2}, f\left(\frac{a+b}{2}\right)\right), (b, f(b))$.

$$I(f) \approx \frac{b-a}{6} \left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right) \quad (\text{Fläche unter der Parabel})$$

Eigenschaften des bestimmten Integrals:

$$\rightarrow c \in [a, b]: \int_a^b f(x) dx = \int_a^c f(x) dx + \int_c^b f(x) dx$$

$$\bullet \lambda, \mu \in \mathbb{R}, f, g \text{ stetig}: I(\lambda f + \mu g) = \lambda I(f) + \mu I(g)$$

$$\rightarrow \text{Monotonie: } f \geq g \text{ auf } [a, b] \text{ dann: } \int_a^b f(x) dx \geq \int_a^b g(x) dx \\ \Rightarrow \left| \int_a^b f(x) dx \right| \leq \int_a^b |f(x)| dx$$

Kondition:

$$\text{Norm: } \|f\|_1 = \int_a^b |f(x)| dx = |I(f)|$$

$$\bar{f}, f: [a, b] \rightarrow \mathbb{R} \text{ stetig, dann: } \frac{|I(f) - I(\bar{f})|}{|I(f)|} \leq \text{cond}_1 \frac{\|f - \bar{f}\|_1}{\|f\|_1} \quad \text{mit } \text{cond}_1 = \frac{|I(f)|}{\|f\|_1}$$

\rightarrow schlecht konditioniert, wenn Integral über Betrag der Fkt. im Verhältnis zu Betrag des Integrals sehr groß, z.B. bei oszillierenden Integranden

Quadraturformeln:

$$\int_a^b f(x) dx \approx (b-a) \underbrace{\sum_{i=1}^s b_i f(a + c_i(b-a))}_{\text{gewichtetes Mittel der Funktionswerte an den Stützstellen}} , \text{ schreibe } (b_i, c_i)_{i=1, \dots, s}$$

b_i = Gewichte

c_i = Knoten, typischerweise in $[0, 1]$

Rechteckregel: $s=1 \quad b_1=1$

$$c_1=0$$

Mittelpunktsregel: $s=1 \quad b_1=1$

$$c_1=\frac{1}{2}$$

Trapezregel: $s=2 \quad b_1=b_2=\frac{1}{2}$

$$c_1=0, c_2=1$$

Simpsonregel: $s=3 \quad b_1=b_3=\frac{1}{6} \quad b_2=\frac{4}{6}$

$$c_1=0, c_2=\frac{1}{2}, c_3=1$$

Ordnung einer Quadraturformel

Eine Quadraturformel $(b_i, c_i)_{i=1,\dots,s}$ besitzt die **Ordnung** p , falls sie exakte Lösungen für alle Polynome vom Grad $\leq p-1$ liefert, wobei p maximal ist.

\Rightarrow besitzt Ordnung $p \Leftrightarrow \frac{1}{q} = \sum_{i=1}^s b_i c_i^{q-1} \quad \forall q = 1, \dots, p$, aber nicht mehr für $q=p+1$

Rechteckregel: $p=1$ ($s=1$)

Mittelpunktregel: $p=2$ ($s=1$)

Trapezregel: $p=2$ ($s=2$)

Simpsonregel: $p=4$ ($s=3$)

Knoten $c_1 < \dots < c_s$ vorgegeben. Verlange von $(b_i, c_i)_{i=1,\dots,s}$ mind. Ordnung s , so

sind Gewichte eindeutig bestimmt:

$$b = C^{-1} \gamma \quad \text{wobei} \quad C = \begin{pmatrix} c_1^0 & c_1^1 & \dots & c_1^s \\ c_2^0 & c_2^1 & \dots & c_2^s \\ \vdots & \vdots & \ddots & \vdots \\ c_s^0 & c_s^1 & \dots & c_s^s \end{pmatrix}$$

$$\text{bzw. } b_i = \int_0^1 l_i(x) dx \quad l_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{s-1} \frac{x - c_j}{c_i - c_j}$$

Symmetrische Quadraturformeln

$(b_i, c_i)_{i=1,\dots,s}$ **symmetrisch**, falls

$$c_i = 1 - c_{s+1-i} \quad , \quad b_i = b_{s+1-i}$$

d.h. Knoten symmetrisch zum Punkt $\frac{1}{2}$ verteilt, Gewichtsvektor von oben nach unten / unten nach oben identisch
Die Ordnung einer symm. Qudr.formel ist gerade.

\rightarrow symm. Knoten ergeben symm. Gewichte:

symm. Knoten $c_1 < \dots < c_s$ vorgegeben ($c_i = 1 - c_{s+1-i}$), $(b_i, c_i)_{i=1,\dots,s}$ mind. Ordnung s \Rightarrow Gewichte auch symm. ($b_i = b_{s+1-i}$)

Quadraturformeln mit erhöhter Ordnung

$(b_i, c_i)_{i=1,\dots,s}$ mit Ordnung $p \geq s$. Ordnung = stm genau dann wenn

$\int_0^1 H(x) g(x) dx = 0 \quad \text{für alle } g(x) \text{ vom Grad } \leq m-1, \text{ aber nicht mehr}$
 $\text{für ein Polynom vom Grad } m$

$$H(x) = (x - c_1)(x - c_2) \dots (x - c_s)$$

\Rightarrow die maximale Ordnung einer QF ist $2s$

\Rightarrow es ex. eine eindeutige QF der Ordnung $2s$, mit $c_i = \frac{1}{2}(1 + y_i)$ ($i=1, \dots, s$)
wobei y_i die Nullstellen des Legendre-Polynoms vom Grad s ; **Gauß-Quadraturformeln** (symmetrisch)

Quadraturfehler:

Fehler: $\int_a^b f(x) dx - (b-a) \sum_{i=1}^s b_i f(a+c_i(b-a)) = (b-a) \underbrace{\left[\int_0^1 g(\tau) d\tau - \sum_{i=1}^s b_i g(c_i) \right]}_{= R(g)} = R(g)$

$$g(\tau) = f(a + \tau(b-a))$$

$(b_i, c_i)_{i=1, \dots, s}$ habe Ordnung p , $g: [0, 1] \rightarrow \mathbb{R}$ q -mal stetig diff. bar, dann:

$$R(g) = \int_0^1 k_q(t) g^{(q)}(t) dt \quad \text{falls } 2 \leq q \leq p$$

mit $k_q(t) = \frac{(1-t)^q}{q!} - \sum_{i=1}^s b_i \frac{(c_i-t)^{q-1}}{(q-1)!}$

$$x_+ = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases} \quad (\mathbb{R} \rightarrow \mathbb{R}^{>0})$$

• $(x-t)_+^n = \begin{cases} (x-t)^n, & x > t \\ 0, & x \leq t \end{cases} \quad (\mathbb{R} \rightarrow \mathbb{R}^{>0}, n \geq 1)$

$$|R(g)| \leq \max_{t \in [0, 1]} |g^{(q)}(t)| \int_0^1 |k_q(t)| dt$$

Es gilt: $\int_0^1 k_p(t) dt = \frac{1}{p!} \left[\frac{1}{p+1} - \sum_{i=1}^s b_i c_i^p \right]$

hat $k_p(t)$ konstantes Vorzeichen:

$$\int_0^1 |k_p(t)| dt = \frac{1}{p!} \left| \frac{1}{p+1} - \sum_{i=1}^s b_i c_i^p \right|$$

⇒ insgesamt: $\int_a^b f(x) dx - (b-a) \sum_{i=1}^s b_i f(a+c_i(b-a)) = (b-a)^{p+1} \int_0^1 f^{(p)}(a+t(b-a)) k_p(t) dt$

→ rechter Kontrollierbar für $\underbrace{b-a}_{h_j} \ll 1 \rightarrow$ Teilintervalle $[x_j, x_{j+1}] \quad j=0, \dots, N-1$

→ Fehler auf Gesamtintervall:

$$\left| \int_a^b f(x) dx - \sum_{j=0}^{N-1} h_j \sum_{i=1}^s b_i f(x_j + c_i h_j) \right| \leq (b-a) \max_{j=0, \dots, N-1} \text{err}_j$$

→ $\max_j \text{err}_j = O((\max_j h_j)^p)$

↑ max. Fehler auf
Teilintervallen