

Numerische Mathematik für Elektrotechniker

Zusammenfassung

Benjamin Berkels

Karl-Heinz Brakhage, Thomas Jankuhn, Christian Löbbert

Institut für Geometrie und Praktische Mathematik
RWTH Aachen

Wintersemester 2019/2020

Heute in der Vorlesung

Zusammenfassung des Stoffes

Dahmen & Reusken Kap. 1-10

1. Einleitung
2. Fehleranalyse: Kondition, Rundungsfehler, Stabilität
3. Lineare Gleichungssysteme, direkte Lösungsverfahren
4. Lineare Ausgleichsrechnung
5. Nichtlineare Gleichungssysteme, iterative Lösungsverfahren
6. Nichtlineare Ausgleichsrechnung
7. Berechnung von Eigenwerten
8. Interpolation
9. Splinefunktionen
10. Numerische Integration

Klausuraufgaben: Verständnisfragen

Zusammenfassung

Kondition: Berechnung

Wie wird die Kondition analysiert?

- ▶ Wir haben folgende Fälle betrachtet
 - ▶ $f : \mathbb{R} \rightarrow \mathbb{R}$
 - ▶ $f : \mathbb{R}^n \rightarrow \mathbb{R}$
 - ▶ $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$, f linear

Wie sind die elementaren Rechenoperationen konditioniert?

- ▶ Multiplikation und Division sind für alle Eingangsdaten **gut konditioniert**.
- ▶ Addition ist
 - ▶ **gut konditioniert**, wenn beide Zahlen das gleiche Vorzeichen haben;
 - ▶ **sehr schlecht konditioniert**, wenn $x_1 \approx -x_2$.

Kondition: Berechnung

- $f : \mathbb{R}^n \rightarrow \mathbb{R}$ (Eingabe: Vektor, Ausgabe: Skalar)

$$\left| \frac{f(\tilde{x}) - f(x)}{f(x)} \right| \leq \kappa_{\text{rel}}(x) \sum_{j=1}^n \left| \frac{\tilde{x}_j - x_j}{x_j} \right|$$

mit

$$\kappa_{\text{rel}}(x) = \kappa_{\text{rel}}^{\infty}(x) := \max_j \left| \frac{\partial f(x)}{\partial x_j} \cdot \frac{x_j}{f(x)} \right|$$

- $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ (Eingabe: Vektor, Ausgabe: Vektor)

Für $f(x) = Ax$ linear, invertierbar gilt

$$\frac{\|f(\tilde{x}) - f(x)\|}{\|f(x)\|} \leq \underbrace{\|A\| \cdot \|A^{-1}\|}_{\kappa(A)} \frac{\|\tilde{x} - x\|}{\|x\|}$$

Maschinenzahlen

Wie werden Zahlen im Computer dargestellt?

- ▶ Maschinenzahlen $\mathbb{M}(b, m, r, R)$
 $\Rightarrow x_{\text{MIN}}, x_{\text{MAX}}, \text{eps}, |\epsilon| \leq \text{eps}$

Welche Probleme können dabei/deswegen auftreten?

- ▶ Assoziativ- und Distributivgesetz nicht mehr gültig
- ▶ Gefahr der Auslöschung bei $\nabla \in \{+, -\}$

Maschinenzahlen: Darstellung

Seien $b \in \mathbb{N}$, $b > 1$, fest gewählt. Jedes $x \in \mathbb{R}$, $x \neq 0$, lässt sich in der Form

$$x = \pm \left(\sum_{j=1}^{\infty} d_j b^{-j} \right) \cdot b^e$$

darstellen, mit $d_j \in \{0, 1, \dots, b-1\}$, $d_1 \neq 0$,
und e eine ganze Zahl.

Nur endliche Anzahl darstellbar $\Rightarrow$ Maschinenzahlen $\mathbb{M}(b, m, r, R)$

- Betragsmäßig kleinste ($\neq 0$) und größte Zahl:

$$x_{\text{MIN}} = b^{r-1} \quad \text{und} \quad x_{\text{MAX}} = (1 - b^{-m}) \times b^R$$

- **Bildbereich** $\mathbb{D} := [-x_{\text{MAX}}, -x_{\text{MIN}}] \cup [x_{\text{MIN}}, x_{\text{MAX}}]$
- **Unterlauf**, wenn $0 \neq |x| < |x_{\text{MIN}}|$;
- **Überlauf**, wenn $|x| > |x_{\text{MAX}}|$.

Maschinengenauigkeit

- Für den relativen Rundungsfehler erhält man für $x \in \mathbb{D}$

$$\left| \frac{\text{fl}(x) - x}{x} \right| \leq \frac{\frac{b^{-m}}{2} \cdot b^e}{b^{-1} \cdot b^e} = \frac{b^{1-m}}{2}.$$

- Die (relative) **Maschinengenauigkeit**

$$\text{eps} := \frac{b^{1-m}}{2}$$

charakterisiert das Auflösungsvermögen des Rechners, d.h.

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

- Der Rundungsfehler ε erfüllt $|\varepsilon| \leq \text{eps}$ und es gilt

$$\text{fl}(x) = x(1 + \varepsilon).$$

Stabilität vs. Kondition

- ▶ Bei einem **stabilen Lösungsverfahren** bleiben die im Laufe der Rechnung erzeugten Rundungsfehler in der Größenordnung der durch die **Kondition des Problems bedingten unvermeidbaren Fehler**.
- ▶ Kenntnisse über die Kondition eines Problems sind oft für die Interpretation oder Bewertung der Ergebnisse von entscheidender Bedeutung.
 - ▶ “Schlechtes Ergebnis” bedeutet nicht unbedingt gleich “instabiler Algorithmus”, sondern deutet evtl. auf eine schlechte Kondition des Problems hin.
- ▶ In einem Algorithmus sollen (wegen Stabilität) **Auslöschungseffekte vermieden werden**.

Operationen $\cdot$ und $/$ sind für alle Eingangsdaten gut konditioniert. Operationen $+$ und $-$ können schlecht konditioniert sein. Dadurch können Rundungsfehler enorm verstärkt werden: **Auslöschung**.

Lineare Gleichungssysteme: Kondition

Satz 3.9

Sie $x + \Delta x$ die Lösung von $(A + \Delta A)(x + \Delta x) = b + \Delta b$.
Dann gilt falls $\kappa(A) \frac{\|\Delta A\|}{\|A\|} < 1$:

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\kappa(A)}{1 - \kappa(A) \frac{\|\Delta A\|}{\|A\|}} \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|} \right).$$

wobei

$$\kappa_{\|\cdot\|}(A) \equiv \|A^{-1}\| \|A\|$$

die (relative) **Konditionszahl der Matrix A** (bzgl. $\|\cdot\|$) ist.

Der Fehler ist proportional zum Residuum $\tilde{r} = b - A \tilde{x}$

$$\kappa(A)^{-1} \frac{\|\tilde{r}\|}{\|b\|} \leq \frac{\|x - \tilde{x}\|}{\|x\|} \leq \kappa(A) \frac{\|\tilde{r}\|}{\|b\|}$$

Matrixzerlegungen

- ▶ LR-Zerlegung/Gauß-Elimination mit Spaltenpivotisierung
 $PA = LR$
 - ▶ (praktisch) stabile und effiziente Methode
 - ▶ Aufwand $\sim \frac{1}{3} n^3$
- ▶ Cholesky-Zerlegung: $A = LDL^T$
 - ▶ Für symmetrisch positiv definite Matrizen (“symmetrische” LR-Zerlegung)
 - ▶ Berechnung durch elementweise Auswertung von $A = LDL^T$: Cholesky-Verfahren
- ▶ QR-Zerlegung: $A = QR$
 - ▶ Existiert auch für rechteckige Matrizen
 - ▶ Berechnung über Givens-Rotation, Housholder-Transformation (oder Gram-Schmidt)

Gauß-Elimination/ LR-Zerlegung

- ▶ Dreiecksmatrizen ergeben leicht lösbare Systeme:
Aufwand $\approx \frac{1}{2} n^2$ Operationen.
- ▶ $\kappa(A)$ spielt eine zentrale Rolle bei der Kondition des Problems
 $(A, b) \rightarrow x = A^{-1}b$, siehe Satz 3.9.
- ▶ **Residuum** als Maß für die Genauigkeit:
Nur aussagekräftig für $\kappa(A) \approx 1$.
- ▶ Zeilenskalierung vs. Pivotisierung
 - ▶ **Skalierung/Äquilibrierung** verbessert die **Kondition** des Gleichungssystems. (sogenannte Vorkonditionierung)
 - ▶ **Pivotisierung** verbessert die **Stabilität** der Gauß-Elimination/ LR-Zerlegung.

Cholesky-Zerlegung

Definition 3.31. (Symmetrisch positiv definite Matrix)

$A \in \mathbb{R}^{n \times n}$ heißt **symmetrisch positiv definit** (s.p.d.), falls

$$A^T = A \quad (\text{Symmetrie})$$

und für alle $x \in \mathbb{R}^n$, $x \neq 0$ gilt

$$x^T A x > 0 \quad (\text{positiv definit}).$$

- ▶ A ist invertierbar, und A^{-1} ist s.p.d.
- ▶ A hat nur strikt positive (insbesondere reelle) Eigenwerte.
- ▶ A hat nur strikt positive Diagonaleinträge und der betragsgrößte Eintrag von A liegt auf der Diagonalen.
- ▶ Bei **Gauß-Elimination ohne Pivotisierung** sind alle **Pivotelemente strikt positiv**.

QR-Zerlegung

Eine Matrix $Q \in \mathbb{R}^{m \times m}$ heißt **orthogonal**, falls

$$Q^T Q = I.$$

Eigenschaften:

- ▶ $Q^{-1} = Q^T$.
- ▶ $\|Q x\|_2 = \|x\|_2$, $\kappa_2(Q) = 1$.
- ▶ $\|Q A\|_2 = \|A\|_2$, $\kappa_2(Q A) = \kappa_2(A)$.

Satz QR-Zerlegung

Zu $A \in \mathbb{R}^{m \times n}$ existiert eine Orthogonale Matrix Q , so dass

$$A = Q R,$$

mit R obere Dreiecksmatrix.

Givens-Rotationen

Grundaufgabe

Gegeben sei $(a, b)^T \in \mathbb{R}^2 \setminus \{0\}$. Finde $c, s, r \in \mathbb{R}$ mit

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} r \\ 0 \end{pmatrix}$$

und

$$c^2 + s^2 = 1.$$

Die Lösung ist: $r = \pm \sqrt{a^2 + b^2}$, $c := \frac{a}{r}$, $s := \frac{b}{r}$

Beachte

Drehung verändert nicht die Euklidische Länge eines Vektors, d.h.

$$\|(r, 0)^T\|_2 = |r| = \sqrt{a^2 + b^2} = \|(a, b)^T\|_2.$$

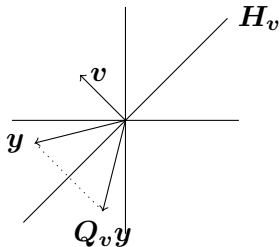
- Die obige (Rotations-)Matrix ist **orthogonal**.

Householder-Transformation

Definition

Für $v = (v_1, \dots, v_n)^T \in \mathbb{R}^n \setminus \{0\}$, ist die Householder-Transformation definiert als

$$Q_v = I - 2 \frac{v v^T}{v^T v}.$$



$$H_v = \{x \in \mathbb{R}^n \mid x^T v = 0\}.$$

Lösung von $Q_v y = \pm \|y\|_2 e^1$

$$v = y + \alpha e^1$$

$$\alpha = \text{sign}(y_1) \|y\|_2$$

$$Q_v y = -\alpha e^1$$

Linearer Ausgleich

- ▶ Aufgabe: $b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$, $m \geq n$,

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

- ▶ Eindeutige Lösung $\Leftrightarrow \text{Rang}(A) = n$
- ▶ Kondition (nur Störung in b):

$$\frac{\|\tilde{x} - x^*\|_2}{\|x^*\|_2} \leq \frac{\kappa_2(A)}{\cos \Theta} \cdot \frac{\|\tilde{b} - b\|_2}{\|b\|_2}$$

$$\text{wobei } \cos \Theta = \frac{\langle Ax^*, b \rangle}{\|Ax^*\| \|b\|} \stackrel{b - Ax^* \perp Ax^*}{=} \frac{\|Ax^*\|}{\|b\|}$$

- ▶ Lösungsverfahren:
 - ▶ über Normalgleichungen $A^T A x = A^T b$
(Cholesky-Verfahren)
 - ▶ über QR -Zerlegung (Householder, Givens)

Linearer Ausgleich: Stabilität

	Normalgleichungen	QR-Zerlegung
Rechenaufwand $(m \gg n)$	$\approx \frac{1}{2} m n^2$	$\approx m n^2$ (Householder)
Stabilität	instabil, wenn $\kappa_2(A) \gg 1$ stabil, wenn $\kappa_2(A)$ moderat	stabil

Singulärwertzerlegung

Satz 4.27

Zu jeder Matrix $A \in \mathbb{R}^{m \times n}$ existieren orthogonale Matrizen $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$ und eine Diagonalmatrix

$$\Sigma := \text{diag}_{mn}(\sigma_1, \dots, \sigma_p) \in \mathbb{R}^{m \times n}, \quad p = \min(m, n),$$

mit

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p \geq 0,$$

so dass

$$A = U \Sigma V^T.$$

Hierbei heißen:

- ▶ Singulärwerte von A : σ_i , $i = 1, \dots, p$
- ▶ Linkssingulärvektoren : Spalten von $U = [u_1 \ u_2 \ \dots \ u_m]$
- ▶ Rechtssingulärvektoren: Spalten von $V = [v_1 \ v_2 \ \dots \ v_n]$

Singulärwertzerlegung: Eigenschaften

Sei $A = U \Sigma V^T$ eine Singulärwertzerlegung von $A \in \mathbb{R}^{m \times n}$ mit

$$\sigma_1 \geq \dots \geq \sigma_r > \sigma_{r+1} = \dots = \sigma_p = 0.$$

- ▶ $\|A\|_2 = \sigma_1 = \sigma_{\max}$
- ▶ $\|A^{-1}\|_2 = \frac{1}{\sigma_n} = \frac{1}{\sigma_{\min}}$, falls $A \in \mathbb{R}^{n \times n}$ regulär
- ▶ $\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{\sigma_1}{\sigma_n}$, falls $A \in \mathbb{R}^{n \times n}$ regulär
- ▶ Die Pseudoinverse $A^+ \in \mathbb{R}^{n \times m}$ ist definiert durch

$$A^+ = V \Sigma^+ U^T \text{ mit } \Sigma^+ = \text{diag}_{n \times m}(\sigma_1^{-1}, \dots, \sigma_r^{-1}, 0, \dots).$$
- ▶ $\kappa_2(A) = \|A\|_2 \|A^+\|_2 = \frac{\sigma_1}{\sigma_r} = \frac{\sigma_{\max}}{\sigma_{\min}}$,
falls $\text{Rang}(A) = r$ mit A^+ die Pseudoinverse.
- ▶ Linearer Ausgleich: $x^* = A^+ b = V \Sigma^+ U^T b$ ist Minimierer
von $\min_{x \in \mathbb{R}^n} \|Ax - b\|_2$ mit $\|x^*\|_2$ minimal.

Nichtlineare Gleichungssysteme

- ▶ Nullstellenproblem $f(x) = 0 \Leftrightarrow$ Fixpunktproblem $\Phi(x) = x$.

Es gibt viele Möglichkeiten für Φ ,

zum Beispiel: für invertierbare Matrix M_x :

$$\Phi(x) = x - M_x f(x).$$

- ▶ Fixpunktiteration:

$$x_{k+1} = \Phi(x_k)$$

- ▶ Banachscher Fixpunktsatz:

$\Phi : E \rightarrow E$, E abgeschlossen, konvex (Selbstabbildung)

Φ Kontraktion auf E

hinreichende Bedingung für Konvergenz der Fixpunktiteration.

- ▶ $n = 1$ (skalares Problem): geometrische Darstellung der Fixpunktiteration.

Nichtlineare Gleichungssysteme: Verfahren

- ▶ **Konvergenzordnung**: Maß für die Konvergenzgeschwindigkeit
- ▶ **Fehlerschätzung** hängt von p (Konvergenzordnung) ab.
 - Skalare Folgen : einfache Formeln $p = 1, p > 1$
 - Vektorfolgen : einfache Formel nur für $p > 1$.
- ▶ Methoden für **skalare** Probleme $f(x) = 0$:
 - Bisektion, Newton-Verfahren, Sekanten-Verfahren.
- ▶ Newton-Verfahren:
 - ▶ $x^{k+1} = x^k + s^k$ mit $f'(x^k) s^k = -f(x^k)$
 - ▶ Konvergenzordnung: $p = 2$, falls $f'(x^*)$ regulär
 - ▶ **lokale** Konvergenz
 - ▶ Stopp-Kriterium: $\|x^* - x^k\| \approx \|x^{k+1} - x^k\|$ "klein"
 - ▶ Vereinfachtes Newton-Verfahren: **feste** Jacobi-Matrix
 - ▶ Wahl des Startvektors: Homotopieverfahren, problemabhängig
 - ▶ über Dämpfung kann man den Konvergenzbereich vergrößern

Das Gauß-Newton-Verfahren

Nichtlineares Ausgleichsproblem

Gegeben $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$, bestimme $x^* \subset \mathbb{R}^n$, so dass

$$x^* \in \arg \min_{x \in \mathbb{R}^n} \|F(x)\|_2.$$

Ansatz:

1. Ersetze $F(x)$ durch lineare Approximation (Taylor-Entwicklung)
2. Schrittweise Annäherung an x^* durch Lösung linearer Probleme in jedem Schritt

Zur Erinnerung:

- ▶ Wir bezeichnen die Lösung am Iterationsschritt k mit

$$x^k = (x_1^k, \dots, x_n^k)^T \in \mathbb{R}^n.$$

- ▶ Taylor-Entwicklung

$$F(x) = F(x^k) + F'(x^k)(x - x^k) + \mathcal{O}(\|x - x^k\|_2^2).$$

Das Gauß-Newton-Verfahren

Algorithmus 6.3. (Gauß-Newton)

Wähle Startwert x^0 .

Für $k = 0, 1, 2, \dots$:

1. Berechne $F(x^k), F'(x^k)$.
2. Finde s^k , so dass

$$s^k \in \arg \min_{s \in \mathbb{R}^n} \|F'(x^k) s + F(x^k)\|_2$$

3. Setze $x^{k+1} = x^k + s^k$.

Beachte

- ▶ Schritt 2 erfordert die Lösung eines linearen Ausgleichsproblems (Normalgleichung, QR-Zerlegung, SVD)
- ▶ Falls $F'(x)$ nicht vollen Rang hat, hat das Ausgleichsproblem keine eindeutige Lösung. → Wähle s^k mit minimaler 2-Norm.

Gauß-Newton als Fixpunktiteration

Gauss-Newton mit $\text{Rang } F'(x) = n$ ist **Fixpunktiteration** mit

$$\Phi(x) = x - [F'(x)^T F'(x)]^{-1} F'(x)^T F(x).$$

Lokale Konvergenz um $x^* = \Phi(x^*)$ hängt von $\rho(K) \|F(x^*)\|_2$, wobei

$$K := A^{-1} \left(\sum_{i=1}^m \frac{F_i(x^*)}{\|F(x^*)\|_2} F_i''(x^*) \right) A^{-1}$$

und $A^2 = F'(x^*)^T F'(x^*)$.

- ▶ Falls x^* **lokales Maximum** oder **Sattelpunkt** von $\phi := \frac{1}{2} \|F\|_2^2$, gilt immer $\rho(K) \|F(x^*)\|_2 \geq 1$, also **abstoßend**.
- ▶ Falls x^* **lokales Minima** mit $\rho(K) \|F(x^*)\|_2 < 1$, dann **lokale Konvergenz**
- ▶ Falls x^* **lokales Minima** mit $\rho(K) \|F(x^*)\|_2 > 1$, dann **abstoßend**

Nichtlinearer Ausgleich: Zusammenfassung

- ▶ Beim Gauß-Newton-Verfahren wird das **nichtlineare Ausgleichsproblem** über eine Folge **linearer Ausgleichsprobleme** gelöst.
- ▶ Lokale Konvergenz im allgemeinen nur **1. Ordnung**, falls $F(x^*) = 0$, sogar von **2. Ordnung**.
- ▶ Es kann lokale **Divergenz** auftreten.
- ▶ Matrix $F'(x^k)$ kann $\text{Rang} < n$ haben.
- ▶ Bei Levenberg-Marquardt:
 - ▶ Parameter μ zur **Vergrößerung des Einzugsbereichs**.
 - ▶ Matrix $\begin{pmatrix} F'(x^k) \\ \mu I \end{pmatrix}$ hat stets **Rang = n**.

Polynominterpolation

Lagrange-Interpolationsaufgabe: Es gibt eindeutiges $P_n \in \Pi_n$, mit

$$P_n(x_j) = f(x_j), \quad j = 0, 1, \dots, n$$

Effiziente Auswertung an einer Stelle mit Neville-Aitken-Schema

Werten Sie $P(f|x_0, x_1, x_2)$ an der Stelle $x = 0.5$ aus.

	$P_{i,0}$	$P_{i,1}$	$P_{i,2}$
$x_0 = 0$	$P_{0,0} = f(x_0) = 1$		
$x_1 = 1$	$P_{1,0} = f(x_1) = 4$	2.5	
$x_2 = 2$	$P_{2,0} = f(x_2) = 2$	5.0	$3\frac{1}{8}$

1. Einsetzen der gegebenen Werte
2. Auswerten der Rekursion

Effiziente Methode zur Auswertung der Potenzform

Sei $p \in \Pi_n$ ein Polynom, das **in der Potenzform vorliegt**, d.h.,

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

mit **bekannten** Koeffizienten $a_0, \dots, a_n$. Es gilt

$$p(x) = a_0 + x \left(a_1 + x \left(a_2 + \cdots + x \left(a_{n-1} + x a_n \right) \right) \right).$$

Algorithmus 8.12. (Horner-Schema)

Setze $b_n = a_n$,

für $k = n - 1, n - 2, \dots, 0$ berechne

$$b_k = a_k + x b_{k+1}$$

Dann ist $p(x) = b_0$.

Der Rechenaufwand wird etwa halbiert.

Lösungsdarstellung

Die eindeutige Lösung $P_n = P(f \mid x_0, \dots, x_n)$ hat **verschiedene Darstellungen** abhängig von der **Wahl der Basis** in Π_n :

1. Lagrange-Basis

$$P_n(x) = \sum_{j=0}^n f(x_j) \ell_{jn}(x), \quad \ell_{jn}(x) = \prod_{\substack{k=0 \\ k \neq j}}^n \frac{x - x_k}{x_j - x_k}.$$

2. monomiale Basis

$$P_n(x) = \sum_{j=0}^n a_j x^j, \quad V_n \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} f(x_0) \\ \vdots \\ f(x_n) \end{pmatrix}.$$

3. Newtonsche Basis

$$P_n(x) = \sum_{j=0}^n [x_0, \dots, x_j] f \omega_j(x), \text{ Knotenpolynome } \omega_j \text{ und}$$

$$[x_0, \dots, x_n] f = \frac{[x_1, \dots, x_n] f - [x_0, \dots, x_{n-1}] f}{x_n - x_0}.$$

Zusammenfassung: Knotenwahl

Definiere

$$\omega_{n+1}(x) := \prod_{j=0}^n (x - x_j) \text{ und } M_{n+1}(f) := \max_{x \in [a,b]} \frac{|f^{(n+1)}(x)|}{(n+1)!}$$

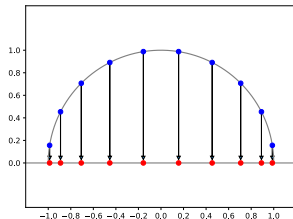
Dann erhält man für die Fehlerschranke

$$|f(x) - P(f|x_0, \dots, x_n)(x)| \leq |\omega_{n+1}(x)| M_{n+1}(f) .$$

Tschebyscheff-Knoten minimieren
den **Fehler der Knotenpolynome**.

Auf $[-1, 1]$ durch gegeben:

$$x_j = \cos\left(\frac{2j+1}{2n+2}\pi\right), \quad j = 0, \dots, n$$



Interpolationsfehler

- ▶ Sei $a := \min\{x_0, \dots, x_n\}$
und $b := \max\{x_0, \dots, x_n\}$ mit n fest.
- ▶ Die Länge des Intervalls $h := b - a$ sei veränderbar.
- ▶ Falls $x \in I := [a, b]$ liegt, erhält man sofort die grobe Abschätzung

$$|\omega_{n+1}(x)| \leq h^{n+1},$$

und somit

$$\|f - P(f|x_0, \dots, x_n)\|_{L^\infty(I)} \leq \frac{h^{n+1}}{(n+1)!} \|f^{n+1}\|_{L^\infty(I)}.$$

- ▶ Der Fehler wird kleiner, wenn die Stützstellen gemäß h zusammenrücken.

Dieser Effekt wird für die numerische Differentiation und Spline-Interpolation genutzt.

Interpolation: Bemerkungen

- ▶ Für die Numerik günstige Darstellung:
 Newtonsche Interpolationsformel
- ▶ Sehr hoher Polynomgrad: oft nicht günstig (Runge-Beispiel).
- ▶ Besser: fester Grad + Wiederholung
 + Verkleinerung des Interpolationsintervalls
- ▶ Bei der Numerischen Differentiation ist bei der Wahl von h ein Kompromiss zwischen Diskretisierungsfehler und Rundungsfehler (Auslöschung) zu finden

Allgemeine Quadraturformel

Satz 10.3 und Lemma 10.4

Für gegebene Knoten $c \leq x_0 < \dots < x_m \leq d$ mit $h = d - c$ gibt es **Gewichte** $c_0, \dots, c_m$ so dass

$$I_m(f) = \int_c^d P(f|x_0, \dots, x_m)(x) dx = h \sum_{j=0}^m c_j f(x_j) .$$

$$c_j = \frac{1}{h} \int_c^d \prod_{\substack{k=0 \\ k \neq j}}^m \frac{x - x_k}{x_j - x_k} dx = \frac{1}{h} \int_c^d \ell_{jm}(x) dx .$$

Die Quadraturformel ist **exakt vom Grade m** , d.h.

für jedes Polynom $Q \in \Pi_m$ gilt $I_m(Q) = \int_c^d Q(x) dx$.

Newton-Cotes-Formeln

Newton-Cotes-Formeln für äquidistante Stützstellen

$$I_m(f) = h \sum_{j=0}^m c_j f(c + \xi_j h), \quad \text{wobei} \quad \xi_j = \begin{cases} \frac{1}{2}, & j = m = 0 \\ \frac{j}{m}, & j = 0, \dots, m > 0. \end{cases}$$

m		ξ_j	c_j	$I_m(f) - \int_c^d f(x) dx$
1	Trapezregel	0, 1	$\frac{1}{2}, \frac{1}{2}$	$\frac{1}{12} h^3 f^{(2)}(\xi)$
2	Simpson-Regel	0, $\frac{1}{2}$, 1	$\frac{1}{6}, \frac{4}{6}, \frac{1}{6}$	$\frac{1}{90} \left(\frac{1}{2}h\right)^5 f^{(4)}(\xi)$

Summierte Formel: $I_{m,n}(f) = \sum_{k=1}^n h \sum_{j=0}^m c_j f(t_{k-1} + \xi_j h)$

Fehlerschranke (Simpson): $E(h) = I_{m,n}(f) - \int_a^b f(x) dx$

$$E(h) = \frac{h^4}{2880} \sum_{k=1}^n h f^{(4)}(\xi_k) \begin{cases} \leq \frac{h^4}{2880} (b-a) \|f^{(4)}\|_{\infty}, \\ \approx \frac{h^4}{2880} (f^{(3)}(b) - f^{(3)}(a)) . \end{cases}$$

Gauß-Quadratur

Newton-Cotes c_j ändern ab $m \geq 8$ Vorzeichen $\Rightarrow$ **Auslöschung**.

Ziel: c_j positiv und maximaler Exaktheitsgrad $\Rightarrow$ **Gauß-Quadratur**

Satz 10.6

Sei $m \geq 0$. Es **existieren** Stützstellen $x_0, \dots, x_m \in (c, d)$ und **positive** Gewichte $w_0, \dots, w_m$, so dass mit $h = d - c$

$$h \sum_{i=0}^m w_i f(x_i) = \int_c^d f(x) dx + E_f(h)$$

und $E_Q = 0$ für alle $Q \in \Pi_{2m+1}$.

Ferner gilt für passendes $\xi \in [c, d]$

$$|E_f(h)| = \frac{((m+1)!)^4}{((2m+2)!)^3 (2m+3)} h^{2m+3} \left| f^{(2m+2)}(\xi) \right|.$$

Zusammenfassung: Newton-Cotes / Gauß-Quadratur

- ▶ Newton-Cotes: **äquidistante** Stützstellen $x_0, \dots, x_m$
Idee: Integriere $P(f \mid x_0, \dots, x_m)(x)$.
- ▶ Gauß-Quadratur: **spezifisch gewählte** Stützstellen
Idee: Finde $x_0, \dots, x_m, c_0, \dots, c_m$, so dass I_m exakt für alle $Q \in \Pi_{2m+1}$.

Unterschiede

- ▶ Newton-Cotes: einfacher, Exaktheitsgrad m oder $m + 1$
- ▶ Gauß-Quadratur: Gewichte $c_j > 0$, Exaktheitsgrad $2m + 1$

Verständnisfragen VF-1

Es seien $x_{\min}$ bzw. $x_{\max}$ die kleinste bzw. größte (strikt) positive Zahl sowie $\mathbf{eps}$ die relative Maschinengenauigkeit in der Menge $\mathbb{M}(b, m, r, R)$ der Maschinenzahlen gemäß Vorlesung/Buch und $\mathbb{D} := [-x_{\max}, -x_{\min}] \cup [x_{\min}, x_{\max}]$.

Ferner beschreibe $\mathbf{fl} : \mathbb{D} \rightarrow \mathbb{M}(b, m, r, R)$ die Standardrundung, und es sei $\ominus$ (gem. Vorlesung/Buch) der Minusoperator für $\mathbb{M}$, d.h.: $x \ominus y := \mathbf{fl}(\mathbf{fl}(x) - \mathbf{fl}(y))$ wobei wir hier annehmen, dass alle Zwischenergebnisse in $\mathbb{D}$ liegen.

Alle Zahlen sind im Dezimalsystem angegeben.

f Für jedes $x \in \mathbb{D}$ existiert eine Zahl ϵ mit $|\epsilon| \leq \mathbf{eps}$ und $\mathbf{fl}(x) = x + \epsilon$.

w Die Zahl **31** ist in $\mathbb{M}(2, 6, -8, 8)$ exakt darstellbar.

w Es gilt $\frac{|(x \ominus y) - (x - y)|}{|x - y|} \leq \mathbf{eps}$ für alle $x, y \in \mathbb{M}(b, m, r, R)$

Es seien $x_{\min}$ bzw. $x_{\max}$ die kleinste bzw. größte (strikt) positive Zahl sowie eps die relative Maschinengenauigkeit in der Menge $M(h, m, r, R)$ der Maschinenzahlen gemäß Vorlesung/Buch und $D := [-x_{\max}, -x_{\min}] \cup [x_{\min}, x_{\max}]$. Ferner beschreibe $\Pi: \mathbb{D} \rightarrow M(h, m, r, R)$ die Standardrundung, und es sei $\ominus$ (gem. Vorlesung/Buch) der Minusoperator für M , d.h.: $x \ominus y := B(\Pi(x) - \Pi(y))$ wobei wir hier annehmen, dass alle Zwischenergebnisse in $\mathbb{D}$ liegen. Alle Zahlen sind im Dezimalsystem angegeben.

f Für jedes $x \in \mathbb{D}$ existiert eine Zahl ϵ mit $|\epsilon| \leq \text{eps}$ und $\Pi(x) = x + \epsilon$.

w Die Zahl 31 ist in $M(2, 6, -8, 8)$ exakt darstellbar.

w Es gilt $\frac{|\Pi(x) - (x-y)|}{|x-y|} \leq \text{eps}$ für alle $x, y \in M(h, m, r, R)$ mit $x \neq y$.

Frage 1: Kann für (betragsmäßig) sehr große Zahlen ($|x| \gg 1$) nicht gelten:

$\text{eps} \leftrightarrow$ relativer Fehler.

Frage 2: $31 = 11111_2 = 0.11111_2 \cdot 2^5$.

Frage 3: Da x und y Maschinenzahlen sind und nach Voraussetzung alle Zwischenergebnisse in $\mathbb{D}$ liegen, steht hier ($z = x - y$) im Prinzip:

$$\frac{|\text{fl}(z) - z|}{|z|} \leq \text{eps}$$

Verständnisfragen VF-1

Berechnen Sie x_{MAX} für $\mathbb{M}(3, 2, -1, 3)$. 24

f Es gilt $\frac{|(x \ominus y) - (x - y)|}{|x - y|} \leq \text{eps}$ für alle $x, y \in \mathbb{D}$ mit $x \neq y$.

f Bei einem stabilen Algorithmus ist der Ausgabefehler nicht viel größer als der Eingabefehler.

f Die Subtraktion zweier Zahlen mit demselben Vorzeichen ist immer schlecht konditioniert.

w Die Funktion $f(x_1, x_2) = x_2 e^{x_1}$ ist für alle (x_1, x_2) mit $|x_1| \leq 1$ gut konditioniert.

Es sei $f(x) = 1/(1 + x)$ und $\tilde{x}$ ein Näherungswert für $x = 3$, der mit einem relativen Fehler von maximal 2% behaftet ist. Bestimmen Sie in erster Näherung eine (scharfe) Schranke für den relativen Fehler in $f(\tilde{x})$ als Annäherung für $f(x)$. 0.015

Berechnen Sie x_{MAX} für $M(3, 2, -1, 3)$ 24f Es gilt $\frac{|f(x,y) - (x-y)|}{|x-y|} \leq \epsilon$ für alle $x, y \in \mathbb{D}$ mit $x \neq y$.f Bei einem stabilen Algorithmus ist der Ausgabefehler nicht viel größer als der Eingabefehler.f Die Subtraktion zweier Zahlen mit demselben Vorzeichen ist immer schlecht konditioniert.w Die Funktion $f(x_1, x_2) = x_2 e^{x_1}$ ist für alle (x_1, x_2) mit $|x_1| \leq 1$ gut konditioniert.Es sei $f(x) = 1/(1+x)$ und $\tilde{x}$ ein Näherungswert für $x = 3$, der mit einem relativen Fehler von maximal 2% behaftet ist. Bestimmen Sie in erster Näherung eine (scharfe) Schranke für den relativen Fehler in $f(\tilde{x})$ als Annäherung für $f(x)$. 0.015Frage 1: $x_{\text{MAX}} = 0.22_3 \cdot 3^3 = \left(\frac{2}{3} + \frac{2}{9}\right) \cdot 27 = 18 + 6 = 24$.Frage 2: Jetzt sind x und y **keine** Maschinenzahlen. Daher ist bedingt durch die Eingangsroundung die schlechte Kondition der Subtraktion möglich.

Frage 3: Man kann die Kondition nicht umgehen.

Beispiel:

 $f(x) = x^{1000}$, $x = 1 \rightarrow \tilde{x} = 1.0001$. Dieser Eingangsfehler von 0.01% wird (bei exakter Rechnung) zu einem Ausgabefehler von mehr als 10.5%Frage 4: Gegenbeispiel: $x = 101$, $y = 1 : \rightarrow \kappa_{\text{rel}} = 1.01$, also gut konditioniert.Frage 5: Man rechnet leicht nach: $\kappa_{\text{rel}, x_1} = |x_1|$ und $\kappa_{\text{rel}, x_2} = 1$.Frage 6: Hier ist $\kappa_{\text{rel}}(x) = \left| -\frac{1}{(1+x)^2} x \right| / \left| \frac{1}{1+x} \right| = \left| \frac{x}{1+x} \right|$.Also $\kappa_{\text{rel}}(3) = 0.75$ und somit werden aus den 2% in der Eingabe in erster Näherung $1.5\% = 0.015$ in der Ausgabe.

Verständnisfragen VF-2

Es seien $A \in \mathbb{R}^{n \times n}$, beliebig aber regulär, und $b \in \mathbb{R}^n$ und gesucht sei die Lösung $x \in \mathbb{R}^n$ von $Ax = b$.

f Es seien $\tilde{x}$ eine Annäherung der Lösung x und $r := b - A\tilde{x}$ das zugehörige Residuum.

Es gilt $\|r\| \leq \kappa(A) \|\tilde{x} - x\|$, mit $\kappa(A) := \|A\| \|A^{-1}\|$.

f Es existieren stets eine untere Dreiecksmatrix L und eine obere Dreiecksmatrix R , so dass $A = LR$ gilt.

w Falls A orthogonal ist, gilt $A^T A = I$.

w Es seien $B \in \mathbb{R}^{n \times n}$ beliebig aber regulär und $\kappa(\cdot)$ die Konditionszahl bzgl. $\|\cdot\|$. Es gilt $\kappa(AB) \leq \kappa(A) \kappa(B)$.

Es sei $B := DA$ die zeilenäquilibrierte Matrix zu A .

Geben Sie $\|B\|_\infty$ an.

1

f Für die Matrix $A = \begin{pmatrix} 2 & 0 \\ -1 & 3 \end{pmatrix}$ existiert eine Cholesky-Zerlegung

Es seien $A \in \mathbb{R}^{n \times n}$, beliebig aber regulär, und $b \in \mathbb{R}^n$ und gesucht sei die Lösung $x \in \mathbb{R}^n$ von $Ax = b$.

f Es seien $\tilde{x}$ eine Annäherung der Lösung x und $r := b - A\tilde{x}$ das zugehörige Residuum.

Es gilt $\|r\| \leq \kappa(A) \|\tilde{x} - x\|$, mit $\kappa(A) := \|A\| \|A^{-1}\|$.

f Es existieren stets eine untere Dreiecksmatrix L und eine obere Dreiecksmatrix R , so dass $A = LR$ gilt.

w Falls A orthogonal ist, gilt $A^T A = I$.

w Es seien $B \in \mathbb{R}^{n \times n}$ beliebig aber regulär und $\kappa(\cdot)$ die Konditionszahl bzgl. $\|\cdot\|$. Es gilt $\kappa(A B) \leq \kappa(A) \kappa(B)$.

Es sei $B := D A$ die zeilenquiliibrierte Matrix zu A .
Geben Sie $\|B\|_\infty$ an. **1**

f Für die Matrix $A = \begin{pmatrix} 2 & 0 \\ -1 & 3 \end{pmatrix}$ existiert eine Cholesky-Zerlegung

Frage 1: $\|r\| = \|b - A\tilde{x}\| = \|Ax - A\tilde{x}\| = \|A(x - \tilde{x})\| \leq \|A\| \cdot \|x - \tilde{x}\|$
 $\|A^{-1}\|$ kann sehr klein sein. Daher ist diese Aussage falsch.

Frage 2: Gegenbeispiel: $A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$. Die Gauß-Elimination zur Berechnung von LR ist nicht immer ohne Pivotisierung möglich.

Frage 3: Definition von orthogonal.

Frage 4: $\kappa(AB) = \|AB\| \|(AB)^{-1}\|$

$$\left. \begin{aligned} \|AB\| &\leq \|A\| \|B\| \\ \|(AB)^{-1}\| &= \|B^{-1} A^{-1}\| \leq \|B^{-1}\| \|A^{-1}\| \end{aligned} \right\} \text{ Behauptung}$$

Frage 5: Zeilenäquilibrieren $d_{ii} = 1/\sum_j |a_{ij}|$ ("in jeder Zeile")
 also so konzipiert, dass $\|DA\|_\infty = 1 = \|B\|_\infty$

Frage 6: A ist nicht symmetrisch.

Verständnisfragen VF-2

Es seien $A \in \mathbb{R}^{n \times n}$, beliebig aber regulär, $b \in \mathbb{R}^n$ und gesucht sei die Lösung $x \in \mathbb{R}^n$ von $Ax = b$.

f Der Rechenaufwand zur Bestimmung der Lösung x über die Gauß-Elimination mit Spaltenpivotisierung beträgt etwa $\frac{1}{6} n^3$ Operationen (gem. Vorlesung/Buch).

f Pivotisierung verbessert die Kondition der Gauß-Elimination.

w Es sei $PA = LR$ die über den Gauß-Algorithmus mit Spaltenpivotisierung berechnete Faktorisierung. Dann gilt:

$$|\det(A^{-1})| = \frac{1}{|\det(R)|}.$$

Es sei $A = \begin{pmatrix} 3 & 2 & 5 \\ -1 & 0 & 1 \\ 2 & -2 & -2 \end{pmatrix}$. Berechnen Sie $\|A\|_1$.

8

Es seien $A \in \mathbb{R}^{n \times n}$, beliebig aber regulär, $b \in \mathbb{R}^n$ und gesucht sei die Lösung $x \in \mathbb{R}^n$ von $Ax = b$.

f Der Rechenaufwand zur Bestimmung der Lösung x über die Gauß-Elimination mit Spaltenpivotisierung beträgt etwa $\frac{1}{6} n^3$ Operationen (gem. Vorlesung/Buch).

f Pivotisierung verbessert die Kondition der Gauß-Elimination.

w Es sei $PA = LR$ die über den Gauß-Algorithmus mit Spaltenpivotisierung berechnete Faktorisierung. Dann gilt:

$$|\det(A^{-1})| = \frac{1}{|\det(R)|}.$$

Es sei $A = \begin{pmatrix} 3 & 2 & 5 \\ -1 & 0 & 1 \\ 2 & -2 & -2 \end{pmatrix}$. Berechnen Sie $\|A\|_1$. **8**

Frage 1: Gauß/ LR -Zerlegung $\frac{1}{3} n^3$, $\frac{1}{6} n^3$ ist Cholesky/ LDL^T -Zerlegung

Frage 2: Pivotisierung stabilisiert (den Algorithmus)
Zusatz: äquilibrierung liefert eine Matrix mit i. A. besserer ∞ -Kondition.

Frage 3: Da $PA = LR$ gilt $\det P \cdot \det A = \det L \cdot \det R$ also

$$\pm \det A = \det R$$

Da aus $AA^{-1} = I$ zudem $\det A \cdot \det(A^{-1}) = 1$ folgt, gilt

$$\left| \det(A^{-1}) \right| = \frac{1}{|\det A|} = \frac{1}{|\det R|}$$

Frage 4: $\|\cdot\|_1$ Spaltensummennorm $\max\{6, 4, 8\} = 8$

Verständnisfragen VF-3

Es seien A eine symmetrisch positiv definite $n \times n$ -Matrix, $b \in \mathbb{R}^n$ und $A = L D L^T$ die Cholesky-Zerlegung von A .

f Es gilt $\|A\|_2 = \|D\|_2$.

f Das Problem $A x = b$ ist immer gut konditioniert.

w Das Cholesky-Verfahren zur Bestimmung der Cholesky-Zerlegung ist ein stabiles Verfahren.

f Für die stabile Berechnung einer LR -Zerlegung $A = LR$ von A ist Pivotisierung notwendig.

Es seien Q eine orthogonale Matrix und $Q \begin{pmatrix} 5 \\ 0 \\ 12 \end{pmatrix} = \begin{pmatrix} 0 \\ c \\ 0 \end{pmatrix}$.
Geben Sie $|c|$ an. **13**

f Es gilt $A^{-1} = L D^{-1} L^T$.

Es seien A eine symmetrisch positiv definite $n \times n$ -Matrix, $b \in \mathbb{R}^n$ und $A = L D L^T$ die Cholesky-Zerlegung von A .

☐ Es gilt $\|A\|_2 = \|D\|_2$.

☐ Das Problem $Ax = b$ ist immer gut konditioniert.

☐ Das Cholesky-Verfahren zur Bestimmung der Cholesky-Zerlegung ist ein stabiles Verfahren.

☐ Für die stabile Berechnung einer LR -Zerlegung $A = LR$ von A ist Pivottisierung notwendig.

Es seien Q eine orthogonale Matrix und $Q \begin{pmatrix} 5 \\ 0 \\ 12 \end{pmatrix} = \begin{pmatrix} 0 \\ c \\ 0 \end{pmatrix}$.

Geben Sie $|c|$ an.

☐ Es gilt $A^{-1} = L D^{-1} L^T$.

Frage 1: Dazu müsste z. B. L orthogonal sein. (Also $\|L\| = \|I\|$). Gegenbeispiel:

$$L = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad D = I, \quad L D L^T = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \text{ offenbar } \|I\|_1 \neq \left\| \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \right\|_1$$

Frage 2: Es gibt auch schlecht konditionierte "s.p.d." Probleme.
Beispielsweise Hilbert-Matrizen. (hier mit Shift)

$$H = \begin{pmatrix} \frac{1}{1000} & \frac{1}{1001} \\ \frac{1}{1001} & \frac{1}{1002} \end{pmatrix}, \quad \kappa_2(H) = 42.1513$$

Frage 3: Cholesky ist von der Stabilität mit Skalierung und Pivottisierung bei nicht s.p.d. Matrizen vergleichbar. (Pivottisierung weder notwendig noch sinnvoll.)

Frage 4: Da A s.p.d., hier nicht erforderlich.

Frage 5: Anwendung von Q ändert Länge nicht. $\rightarrow |c| = \sqrt{5^2 + 12^2} = \sqrt{169} = 13$

Frage 6: $A = L D L^T \Rightarrow A^{-1} = L^{-T} D^{-1} L^{-1}$

Verständnisfragen VF-3

Es seien A eine symmetrisch positiv definite $n \times n$ -Matrix, $b \in \mathbb{R}^n$ und $A = L D L^T$ die Cholesky-Zerlegung von A .

w Das Produkt zweier Givens-Rotations-Matrizen ist eine orthogonale Matrix.

f Es seien $Q_v \in \mathbb{R}^{m \times m}$ eine Householder-Transformations-Matrix und $x \in \mathbb{R}^m$ beliebig. Es gilt $\|Q_v x\|_\infty = \|x\|_\infty$.

f Die Berechnung einer QR -Zerlegung $B = Q R$ von $B \in \mathbb{R}^{m \times n}$ über Householder-Transformationen ist nur dann stabil, wenn die Matrix B vollen Spaltenrang hat.

Es sei Q_v eine Householder-Transformation. Geben Sie den Wert des größten Eigenwertes der Matrix Q_v an. **1**

Es seien A eine symmetrisch positiv definite $n \times n$ -Matrix, $b \in \mathbb{R}^n$ und $A = LDL^T$ die Cholesky-Zerlegung von A .

W Das Produkt zweier Givens-Rotations-Matrizen ist eine orthogonale Matrix.

F Es seien $Q_n \in \mathbb{R}^{n \times n}$ eine Householder-Transformations-Matrix und $x \in \mathbb{R}^n$ beliebig. Es gilt $\|Q_n x\|_\infty = \|x\|_\infty$.

F Die Berechnung einer QR -Zerlegung $B = QR$ von $B \in \mathbb{R}^{m \times n}$ über Householder-Transformationen ist nur dann stabil, wenn die Matrix B vollen Spaltenrang hat.

Es sei Q_n eine Householder-Transformation. Geben Sie den Wert des größten Eigenwertes der Matrix Q_n an. **1**

Frage 1: Das Produkt orthogonaler Matrizen ist orthogonal.

$$(Q_1 Q_2) (Q_1 Q_2)^T = Q_1 (Q_2 Q_2^T) Q_1^T = Q_1 Q_1^T = I$$

Frage 2: "falsche" Norm. Gilt nur für die 2-Norm.

Frage 3: Orthogonale Matrizen bilden auch für Matrizen mit Rangdefekt ein stabiles Verfahren zur Berechnung der QR -Zerlegung

Frage 4: $Qx = \lambda x$: für $\|\cdot\|_2$ gilt

$$\|x\|_2 = \|Qx\|_2 = \|\lambda x\|_2 = |\lambda| \|x\|_2 \text{ also } |\lambda| = 1$$

Verständnisfragen VF-4

Es seien $A \in \mathbb{R}^{m \times n}$, mit $\text{Rang}(A) = n < m$, und $b \in \mathbb{R}^m$.
 Weiter seien $Q \in \mathbb{R}^{m \times m}$ eine orthogonale Matrix und $R \in \mathbb{R}^{m \times n}$
 eine obere Dreiecksmatrix so, dass $Q A = R = \begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix}$ gilt, mit
 $\tilde{R} \in \mathbb{R}^{n \times n}$. Es sei $x^* = \arg \min_{x \in \mathbb{R}^n} \|A x - b\|_2 \in \mathbb{R}^n$.
 Weiter sei $\Theta \in [0, \frac{\pi}{2})$ der Winkel zwischen $A x^*$ und b .

w Je kleiner der Winkel Θ , desto kleiner ist die Größe $\frac{\|A x^* - b\|_2}{\|b\|_2}$.

f Es gilt $\tilde{R} x^* = Q^T b$.

w Die Matrix $\tilde{R}$ kann man über Givens-Rotationen bestimmen.

f Es gilt $\det \tilde{R} = \det A$.

Es seien $A \in \mathbb{R}^{m \times n}$, mit $\text{Rang}(A) = n < m$, und $b \in \mathbb{R}^m$.
 Weiter seien $Q \in \mathbb{R}^{m \times m}$ eine orthogonale Matrix und $R \in \mathbb{R}^{m \times n}$
 eine obere Dreiecksmatrix so, dass $Q A = R = \begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix}$ gilt, mit
 $\tilde{R} \in \mathbb{R}^{n \times n}$. Es sei $x^* = \arg \min_{x \in \mathbb{R}^n} \|A x - b\|_2 \in \mathbb{R}^n$.
 Weiter sei $\Theta \in [0, \frac{\pi}{2})$ der Winkel zwischen $A x^*$ und b .

☒ W Je kleiner der Winkel Θ , desto kleiner ist die Größe $\frac{\|A x^* - b\|_2}{\|b\|_2}$.

☐ F Es gilt $\tilde{R} x^* = Q^T b$.

☒ W Die Matrix $\tilde{R}$ kann man über Givens-Rotationen bestimmen.

☐ F Es gilt $\det \tilde{R} = \det A$.

Frage 1: $\sin \theta = \frac{\|A x^* - b\|_2}{\|b\|_2}$

Frage 2: "doppelt" falsch.

a) Dimensionen passen nicht $\tilde{R} x^* \in \mathbb{R}^n$, $Q^T b \in \mathbb{R}^m$

b) $Q A = R$: Aber mit $\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = Q b$ plus $\tilde{R} x^* = b_1$

Frage 3: Givens-Rotationen sind orthogonale Transformationen

Frage 4: $\det A$ ist nicht-quadratische Matrizen nicht definiert.

Verständnisfragen VF-4

w Es gilt $\|Ax - b\|_2 = \|Rx - Qb\|_2$ für beliebiges $x \in \mathbb{R}^n$.

Es seien $m = 4$, $n = 3$ und $Qb = (1, 0, 3, -4)^T$.

Bestimmen Sie $\|Ax^* - b\|_2$. **4**

f Durch eine geeignete Wahl des skalaren Parameters im Levenberg-Marquardt-Verfahren zur Lösung eines nichtlinearen Ausgleichsproblems wird die Konvergenzordnung der Methode in der Regel erhöht.

w Die Gauß-Newton-Methode zur Lösung eines nichtlinearen Ausgleichsproblems kann man als Fixpunktiteration darstellen.

f Die Konvergenzordnung der Gauß-Newton-Methode zur Lösung eines nichtlinearen Ausgleichsproblems ist immer maximal 1.

Es sei $\Theta = 0$. Bestimmen Sie $\|Ax^* - b\|_2$. **0**

w Es gilt $\|Ax - b\|_2 = \|Rx - Qb\|_2$ für beliebiges $x \in \mathbb{R}^n$.

Es seien $m = 4$, $n = 3$ und $Qb = (1, 0, 3, -4)^T$.

Bestimmen Sie $\|Ax^* - b\|_2$.

f Durch eine geeignete Wahl des skalaren Parameters im Levenberg-Marquardt-Verfahren zur Lösung eines nichtlinearen Ausgleichsproblems wird die Konvergenzordnung der Methode in der Regel erhöht.

w Die Gauß-Newton-Methode zur Lösung eines nichtlinearen Ausgleichsproblems kann man als Fixpunktiteration darstellen.

f Die Konvergenzordnung der Gauß-Newton-Methode zur Lösung eines nichtlinearen Ausgleichsproblems ist immer maximal 1.

Es sei $\Theta = 0$. Bestimmen Sie $\|Ax^* - b\|_2$.

Frage 1: $\|Ax - b\|_2 = \|Q(Ax - b)\|_2 = \|Rx - Qb\|_2$

Frage 2: Es gilt: $\|Ax^* - b\|_2 = \|Rx^* - Qb\|_2 = \left\| \begin{pmatrix} \tilde{R}x^* \\ 0 \end{pmatrix} - \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \right\|_2$.
 $\tilde{R}x^* = b_1$, $\|b_2\|_2$ Residuum. Hier $b_2 = (-4)$.

Frage 3: Der Parameter dient zur Dämpfung der Korrektur. Ordnung bleibt. Konvergiert möglicherweise langsamer, dafür aber auch dann, wenn Gauß-Newton nicht konvergiert.

Frage 4: Formal $x^{k+1} = x^k - (F'(x^k)^T F'(x^k))^{-1} F'(x^k)^T x^k$

Frage 5: Kann auch höher sein.
 Falls $\|F(x^*)\| = 0$ mindestens 2 (lokal!)

Frage 6: Mit der erste Frage in diesem Abschnitt folgt $0 = \sin 0 = \frac{\|Ax^* - b\|_2}{\|b\|_2}$, also
 $\|Ax^* - b\|_2 = 0$.

Verständnisfragen VF-5

Es seien $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}^n$ stetig differenzierbar und x^* so, dass $\Phi(x^*) = x^*$ gilt. Für $x_0 \in \mathbb{R}^n$ wird die Fixpunktiteration $x_{k+1} = \Phi(x_k)$, $k = 0, 1, 2, \dots$ definiert.

Weiter sei $\Phi'(x)$ die Ableitung von Φ an der Stelle x .

Für $n = 1$ sei außerdem $\Phi_1(x) := \frac{1}{4}x^2 - 1$.

- f** Es gilt $\|\Phi'(x^*)\| < 1$.
- f** Die Konvergenzordnung der Fixpunktiteration ist maximal 2.
- f** Das Fixpunktproblem $\Phi_1(x) = x$ hat eine eindeutige Lösung $x^* \in \mathbb{R}$.
- w** Für Φ_1 sind auf dem Intervall $[-1, 0]$ alle Voraussetzungen des Banachschen Fixpunktsatzes erfüllt.

Es seien $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}^n$ stetig differenzierbar und x^* so, dass $\Phi(x^*) = x^*$ gilt. Für $x_0 \in \mathbb{R}^n$ wird die Fixpunktiteration $x_{k+1} = \Phi(x_k)$, $k = 0, 1, 2, \dots$ definiert. Weiter sei $\Phi'(x)$ die Ableitung von Φ an der Stelle x . Für $n = 1$ sei außerdem $\Phi_1(x) := \frac{1}{4}x^2 - 1$.

☐ Es gilt $\|\Phi'(x^*)\| < 1$.

☐ Die Konvergenzordnung der Fixpunktiteration ist maximal 2.

☐ Das Fixpunktproblem $\Phi_1(x) = x$ hat eine eindeutige Lösung $x^* \in \mathbb{R}$.

☐ Für Φ_1 sind auf dem Intervall $[-1, 0]$ alle Voraussetzungen des Banachschen Fixpunktsatzes erfüllt.

Frage 1: Es gibt auch "abstoßende" Fixpunkte

Frage 2: $\Phi(x) = x^p$ hat Konvergenzordnung p , denn $x^* = 0$,

$$\Phi'(x^*) = \dots = \Phi^{(p-1)}(x^*) = 0$$

und

$$\Phi^{(p)}(x^*) = p! \neq 0.$$

Frage 3: $\Phi_1(x) = \frac{1}{4}x^2 - 1 = x$ hat genau 2 Lösungen.

$$x^2 - 4x - 4 = 0 \quad \Leftrightarrow \quad x_{1,2} = 2 \pm \sqrt{8}$$

Frage 4: Φ_1 ist auf $[-1, 0]$ monoton fallend, also $\Phi_1([-1, 0]) \subset [-1, -\frac{3}{4}]$

$$\Phi_1'(x) = \frac{1}{2}x \text{ also } |\Phi_1'(x)| \leq \frac{1}{2} \text{ auf } [-1, 0].$$

Verständnisfragen VF-5

Wir betrachten die Fixpunktiteration zur Bestimmung einer Lösung $x^* < 0$ des Fixpunktproblems $\Phi_1(x) = x$, mit einem Startwert x_0 aus einer hinreichend kleinen Umgebung von x^* . Geben Sie die Konvergenzordnung dieser Methode an. 1

f Bei der Sekantenmethode zur Bestimmung einer Nullstelle einer skalaren Funktion f , müssen die Startwerte x_0, x_1 so gewählt werden, dass $f(x_0) f(x_1) < 0$ gilt.

w Es sei $f(x) = x^2 - 3$. Das auf f angewandte Newton-Verfahren konvergiert für jeden Startwert $x_0 > 0$ gegen die Nullstelle $x^* > 0$ dieser Funktion.

f Eine Dämpfungsstrategie beim Newton-Verfahren zur Bestimmung einer Nullstelle kann man nur bei skalarwertigen Funktionen $f : \mathbb{R} \rightarrow \mathbb{R}$ anwenden.

Wir betrachten die Fixpunktiteration zur Bestimmung einer Lösung $x^* < 0$ des Fixpunktproblems $\Phi_1(x) = x$, mit einem Startwert x_0 aus einer hinreichend kleinen Umgebung von x^* . Geben Sie die Konvergenzordnung dieser Methode an. 1

f Bei der Sekantenmethode zur Bestimmung einer Nullstelle einer skalaren Funktion f , müssen die Startwerte x_0, x_1 so gewählt werden, dass $f(x_0)f(x_1) < 0$ gilt.

w Es sei $f(x) = x^2 - 3$. Das auf f angewandte Newton-Verfahren konvergiert für jeden Startwert $x_0 > 0$ gegen die Nullstelle $x^* > 0$ dieser Funktion.

f Eine Dämpfungstrategie beim Newton-Verfahren zur Bestimmung einer Nullstelle kann man nur bei skalarwertigen Funktionen $f: \mathbb{R} \rightarrow \mathbb{R}$ anwenden.

Frage 1: $\Phi'(x)$ hat nur $x = 0$ als Nullstelle $\Rightarrow \Phi'(x^*) \neq 0$
 $\Rightarrow$ Konvergenzordnung maximal 1. Es folgt genau 1

Frage 2: Muss nicht sein. Man verliert den Einschluss ja auch (vorübergehend) während der Iteration

Frage 3: Siehe Beispiel 5.24.

$$\begin{aligned} 0 < x_0 < \sqrt{3} &\rightarrow \sqrt{3} < x_1 \\ x_i = \sqrt{3} &\rightarrow x_{i+1} = \sqrt{3}, \quad i = 0, 1, 2, \dots \\ \sqrt{3} < x_i &\rightarrow \sqrt{3} < x_{i+1} < x_i, \quad i = 0, 1, 2, \dots \end{aligned}$$

Frage 4: Geht mehrdimensional genauso.

Verständnisfragen VF-5

f Es seien $f : \mathbb{R} \rightarrow \mathbb{R}$ zweimal stetig differenzierbar, und $f(x^*) = 0$, $f'(x^*) \neq 0$.

Weiter sei x_0 so gewählt, dass die Newton-Methode

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

mit Startwert x_0 gegen x^* konvergiert.

Dann gilt: $|x^* - x_k| \approx (x_{k+1} - x_k)^2$ für k hinreichend groß.

Es seien $n = 1$ und $\Phi(x) = e^{-\frac{1}{2}x}$.

Wir betrachten das Fixpunktproblem auf dem Intervall $[0, 1]$.

Geben Sie eine scharfe obere Schranke für die Lipschitzkonstante

$L < 1$ aus dem Banachschen Fixpunktsatz an. **0.5**

Es seien $f: \mathbb{R} \rightarrow \mathbb{R}$ zweimal stetig differenzierbar, und $f(x^*) = 0$, $f'(x^*) \neq 0$.

Weiter sei x_0 so gewählt, dass die Newton-Methode

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

mit Startwert x_0 gegen x^* konvergiert.

Dann gilt: $|x^* - x_k| \approx (x_{k+1} - x_k)^2$ für k hinreichend groß.

Es seien $v_1 = 1$ und $\Phi(x) = e^{-\frac{1}{2}x}$.

Wir betrachten das Fixpunktproblem auf dem Intervall $[0, 1]$.

Geben Sie eine scharfe obere Schranke für die Lipschitzkonstante

$L < 1$ aus dem Banachschen Fixpunktsatz an. 0.5

Frage 1: Richtig wäre

$$|x^* - x_k| \approx |x_{k+1} - x_k|$$

oder

$$|x_{k+1} - x^*| \leq c |x_k - x^*|^2.$$

Frage 2: $[0, 1]$ wird durch $e^{-\frac{1}{2}x}$ auf $[e^{-\frac{1}{2}}, 1]$ abgebildet.

$$\Phi'(x) = -\frac{1}{2}e^{-\frac{1}{2}x} \rightarrow |\Phi'(x)| \leq \frac{1}{2} = 0.5 \text{ auf } [0, 1]$$

Verständnisfragen VF-6

Es sei $P(f | x_0, \dots, x_n)$ das Lagrange-Interpolationspolynom zu den Daten $(x_0, f(x_0)), \dots, (x_n, f(x_n))$ mit $a = x_0 < \dots < x_n = b$. Weiter sei $[x_0, \dots, x_n]f$ die dividierte Differenz der Ordnung n von f .

- f** Es gilt für alle $x \in \mathbb{R}$ $P(f | x_0, \dots, x_n)(x) = (x - x_n)[x_0, \dots, x_n]f + P(f | x_0, \dots, x_{n-1})(x)$.
- w** Es sei Π_n der Raum der reellen Polynome vom Grad maximal n . Die Knotenpolynome $\omega_k(x) := (x - x_0) \cdots (x - x_{k-1})$, $k = 1, \dots, n$, $\omega_0(x) := 1$, bilden eine Basis des Raumes Π_n .
- w** Die Auswertung des Interpolationspolynoms in der monomialen Basis $P(f | x_0, \dots, x_n)(x) = \sum_{k=0}^n a_k x^k$ ist für numerische Zwecke ungünstig, weil das Problem (Auswertung) bezüglich der Koeffizienten a_k oft schlecht konditioniert ist.
- w** Der Fehler $\max_{x \in [a, b]} |P(f | x_0, \dots, x_n) - f(x)|$ hängt von der Wahl der Stützstellen ab.

Es sei $f(x) = 3x^2 + 2$. $[x_0, x_1, x_2, x_3]f$ ist gleich

0

Es sei $P(f | x_0, \dots, x_n)$ das Lagrange-Interpolationspolynom zu den Daten $(x_0, f(x_0)), \dots, (x_n, f(x_n))$ mit $a = x_0 < \dots < x_n = b$. Weiter sei $[x_0, \dots, x_n]f$ die dividierte Differenz der Ordnung n von f .

f Es gilt für alle $x \in \mathbb{R}$: $P(f | x_0, \dots, x_n)(x) = (x - x_n)[x_0, \dots, x_n]f + P(f | x_0, \dots, x_{n-1})(x)$.

w Es sei Π_k der Raum der reellen Polynome vom Grad maximal n . Die Knotenpolynome $\omega_k(x) := (x - x_0) \cdots (x - x_{k-1})$, $k = 1, \dots, n$, $\omega_0(x) := 1$, bilden eine Basis des Raumes Π_n .

w Die Auswertung des Interpolationspolynoms in der monomialen Basis $P(f | x_0, \dots, x_n)(x) = \sum_{k=0}^n a_k x^k$ ist für numerische Zwecke ungünstig, weil das Problem (Auswertung) bezüglich der Koeffizienten a_k oft schlecht konditioniert ist.

w Der Fehler $\max_{x \in [a,b]} |P(f | x_0, \dots, x_n) - f(x)|$ hängt von der Wahl der Stützstellen ab.

Es sei $f(x) = 3x^2 + 2$. $[x_0, x_1, x_2, x_3]f$ ist gleich .

Frage 1: a) Polynomgrad nur $n - 1$

b) Richtig wäre es wenn $(x - x_n)$ durch $\prod_{i=0}^{n-1} (x - x_i)$ ersetzt wird.

Frage 2: Das ist die Newtonsche Basis

Frage 3: Beispiel Kondition bzgl. a_0 .

$$\frac{\partial P(x)}{\partial a_0} = 1 \Rightarrow \kappa_{rel, a_0} = \left| \frac{a_0}{P(x)} \right|$$

a_0 ist groß, $|P(x)|$ klein, daher ist κ_{rel, a_0} groß.

Frage 4: Wir hatten z. B. das Beispiel von Runge und

$$P_n(x) - f(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^n (x - x_i). \text{ Abh. von } x_i.$$

Frage 5: $[x_0, x_1, x_2, x_3]f$ ist der führende Koeffizient von $P(f | x_0, \dots, x_3)$, d.h. der Koeffizient von x^4 , also 0.

Verständnisfragen VF-6

Es sei $f \in C[a, b]$. Das Integral $I(f) = \int_a^b f(x) dx$ soll numerisch approximiert werden durch eine Newton-Cotes-Formel $I_m(f) = (b - a) \sum_{j=0}^m w_j f(x_j)$ mit $a \leq x_0 < \dots < x_m \leq b$. Weiter sei $I_{m,n}(f)$ die aus $I_m(f)$ konstruierte summierte Quadraturformel auf den Teilintervallen $[t_{j-1}, t_j]$, $j = 1, \dots, n$, mit $t_j = a + jh$, $j = 0, 1, \dots, n$, $h = (b - a)/n$.

w Es sei $I_2(f)$ die Simpson-Regel.

Dann gilt $|I_{2,n}(f) - I(f)| \rightarrow 0$ für $n \rightarrow \infty$.

w Es gilt $I_m(p) = I(p)$ für alle Polynome p vom Grad maximal m .

f Es gilt $I_{1,n}(p) = I(p)$ für alle Polynome p vom Grad maximal n .

f Bei den Newton-Cotes-Formeln hängen die Gewichte w_j von dem Intervall $[a, b]$ ab.

Berechnen Sie eine Approximation von $\int_0^2 x^5 dx$ mit Hilfe der summierten Trapezregel $I_{1,2}(f)$.

17

Es sei $f \in C[a, b]$. Das Integral $I(f) = \int_a^b f(x) dx$ soll numerisch approximiert werden durch eine Newton-Cotes-Formel $I_{n,\alpha}(f) = (b-a) \sum_{j=0}^{n-\alpha} w_j f(x_j)$ mit $a \leq x_0 < \dots < x_{n-\alpha} \leq b$. Weiter sei $I_{n,\alpha}(f)$ die aus $I_{n,\alpha}(f)$ konstruierte summierte Quadraturformel auf den Teilintervallen $[x_{j-1}, x_j]$, $j = 1, \dots, n$, mit $x_j = a + jh$, $j = 0, 1, \dots, n$, $h = (b-a)/n$.

w Es sei $I_2(f)$ die Simpson-Regel.
Dann gilt $|I_{2,n}(f) - I(f)| \rightarrow 0$ für $n \rightarrow \infty$.

w Es gilt $I_m(p) = I(p)$ für alle Polynome p vom Grad maximal m .

f Es gilt $I_{n,\alpha}(p) = I(p)$ für alle Polynome p vom Grad maximal n .

f Bei den Newton-Cotes-Formeln hängen die Gewichte w_j von dem Intervall $[a, b]$ ab.
Berechnen Sie eine Approximation von $\int_0^2 x^5 dx$ mit Hilfe der summierten Trapezregel $I_{1,2}(f)$. **17**

Frage 1: Stetigkeit reicht bereits für die Konvergenz der summierten Regeln. Mit $f \in C^4[a, b]$ hätten wir hier sogar $\mathcal{O}(h^4)$, $h = (b-a)/n$

Frage 2: Der Exaktheitsgrad von Newton-Cotes-Formeln ist m oder $m+1$, also ist m immer OK.

Frage 3: Summation ändert **nicht** den Exaktheitsgrad

Frage 4: Weil wir die Darstellung $h \sum_i w_i f(x_i)$ wählen, sind die Gewichte nicht intervallabhängig.

Frage 5: $\frac{1}{2}(0^5 + 2 \cdot 1^5 + 2^5) = \frac{34}{2} = 17$