ILIAS-Lernkontrollen und -Übungen

Maximilian Göckel

February 7, 2020

Abstract

1 Vorlesung 1

1.1 Lernkontrolle

1. Erklären sie den Aufbau einer normalisierten Gleitkommazahl
   - Gleitkommazahl: $Basis * Mantisse^{Exponent}$
   - Basis: Meistens Zweierpotenz (2,8,16)
   - Exponent: $e_{min} \leq e = e_{min} + \sum_{l=0}^{L_e-1} c_l B^l \leq e_{min} + B^{L_e} - 1 := e_{max}$
   - Mantisse:
     - Entweder 0 oder $m = \sum_{l=1}^{L_m} a_l B1{-}l$
     - $B^{-1} \leq |m| < 1$
   - Normalisierung: Mantisse beginnt immer mit einer 1, also $a_1 \neq 0$
2. Geben Sie die Definition der Maschinengenauigkeit an und erklären Sie, in welcher Weise die Mantissenlänge L_m die relative Genauigkeit der Darstellung reeller Zahlen durch eine normalisierte Gleitpunktzahl bestimmt
   - Maschinengenauigkeit $eps = \frac{B^{(1-L_m)}}{2}$.
   - Ist die Mantisse länger (also L_m größer), so können mehr Nachkommastellen dargestellt werden, da eps kleiner wird.
3. Beschreiben Sie den IEEE-Standard des Double Precision-Formats und erklären Sie, warum $1 + eps$ die kleinste Zahl echt größer 1 in FL ist
   - 64Bit für Gleitpunktdarstellung
     - Basis 2 fest
     - 1Bit Vorzeichen
     - 11Bit Exponent mit $e_{min} = -1022$
     - 52Bit Mantisse

- Darstellbare (pos.) Zahlen von 10^{-308} bis 10^{308}
- Rel. Genauigkeit $eps = 2^{-52}$, circa 16 Nachkommastellen genau
- Jede Zahl x mit $1 < x < 1 + eps$ ist zu klein, es werden Nachkommastellen abgeschnitten
- Bei normalisierten GKZ ist die 1 fest, es können also mehr Bits für die Nachkommastellen der Mantisse genutzt werden und die Genauigkeit steigt auf $\frac{eps}{2}$

4. Nennen Sie mögliche Fehlerquellen der Eingabedaten eines numerischen Verfahrens und innerhalb des Verfahrens selbst beim Lösen eines mathematischen Problems und formulieren Sie die Fragestellungen, welche der Stabilität eines Verfahrens und der Kondition des Problems zugrunde liegen

   - Eingabedaten-Fehler: Rundungsfehler, Messwertfehler
   - Kondition: "Wie wirken sich Störungen der Eingabedaten auf das Resultat aus, unabhängig vom Algorithmus?"
   - Verfahrens-Fehler: Rundungsfehler, Approximationsfehler (wenn das Verfahren von vornherein nicht genau arbeitet)
   - Stabilität: "Wie stark wirken sich Störungen im Algorithmus auf das Ergebnis aus?"

5. Vollziehen Sie die Untersuchung und Deutung der *Kondition der Summe zweier Zahlen* der Vorlesung nach, indem Sie auch das Phänomen der *Auslöschung* erklären

   Kondition von $x+y$: Betrachte $x+\epsilon_x$ und $y+\epsilon_y$ als gestörte Eingabedaten von x und y.
   Die absolute Abweichung von $(x + \epsilon_x) + (y + \epsilon_y)$ zu $x + y$ ist $Abw_{abs} = |((x + \epsilon_x) + (y + \epsilon_y)) - (x + y)|$. Die relative Abweichung ist $Abw_{rel} = \frac{Abw_{abs}}{|(x+y)|} \leq \epsilon \frac{|x|+|y|}{|x+y|}$ fpr ein kleines ϵ. Das Problem ist an sich gut konditioniert, nur für $x \approx (-y)$ kann es zur Auslöschung kommen.

   Auslöschung: Bei der Subtraktion von zwei fast gleich großen GKZ (erste paar Nachkommastellen gleich) kann es passieren dass der Unterschied so klein ist, dass er durch die Maschinengenauigkeit beeinflusst wird. Die gleichen Stellen werden ausgelöscht und nur Stellen mit Rundungsfehler bleiben bestehen.
   Beispiel: $a = 2,345678$ und $b = 2,346789$, so ist $(b - a) = 0,001111$. Die niedrigwertigsten Stellen sind sehr anfällig für Rundungsfehler (z.B. aus vorherigen Berechnungen) und die höchsten (korrekten) Stellen löschen sich zu 0 aus.

2 Vorlesung 2

2.1 Lernkontrolle

1. Erklären Sie, wann genau und warum eine Zerlegung $A = LR$ für A existiert und mit welchem Aufwand sich diese berechnen lässt.

   - $A = LR$ existiert $\Leftrightarrow$ Jede Matrix $A_{[1:n,1:n]}$ $(n = 1...N)$ ist regulär
   - Eine LR-Zerlegung kann in $\mathrm{O}(\frac{1}{3}N^3)$ errechnet werden

2. Geben Sie an, für welche spezielle Klasse von Matrizen A auf jeden Fall eine Zerlegung $A = LR$ existiert.

   - Für Matrizen die bereits die Form von L oder R haben (obere o. untere Dreiecksmatrizen)

3. Erklären Sie, wann genau und warum eine Zerlegung $PA = LR$ für A existiert.

   - A regulär $\Rightarrow$ Es existiert eine Permutationsmatrix P sodass für PA die Bedingungen aus 1. gelten
   - P kann während der Berechnung von L, R durch Spaltenpivotwahl berechnet werden

4. Beschreiben Sie, wie sich das lineare Gleichungssystem $Ax = b$ bei Kenntnis einer Zerlegung $PA = LR$ lösen lässt, und geben Sie den Aufwand der einzelnen Schritte an.

   1. Löse $Ly = Pb$ durch Vorwärtssubstitution in $\mathrm{O}(\frac{1}{2}N^2)$
   2. Löse $Rx = y$ durch Rückwärtssubstitution in $\mathrm{O}(\frac{1}{2}N^2)$

   Dies benötigt $2\times\mathrm{O}(\frac{1}{2}N^2) = \mathrm{O}(N^2)$ Operationen

5. Erklären Sie, warum die Kenntnis der Inversen A^{-1} gegenüber einer bekannten Zerlegung $PA = LR$ beim Lösen von $Ax = b$ keinen Vorteil darstellt.

   - Da bei Kenntnis von A^{-1} immer noch $A^{-1}b$ errechnet werden muss. Dies benötigt auch $\mathrm{O}(N^2)$ Schritte

3 Vorlesung 3

3.1 Lernkontrolle

1. Erklären Sie, für welche Matrizen A genau eine Cholesky-Zerlegung existiert und wie bei Kenntnis einer solchen Zerlegung das Gleichungssystem $Ax = b$ gelöst werden kann.

   - A spd-Matrix $\Leftrightarrow$ Cholesky-Zerlegung $A = LL^T$ existiert

- Löse $Ax = b$ wie mit der LR-Zerlegung über Vorwärts- und Rückwärtssubstitition (anstelle von R nehme L^T)

2. Geben Sie an, wie eine QR-Zerlegung einer Matrix $A \in \mathbb{R}^{M \times N}$ definiert ist, indem Sie die Faktoren beschreiben, und erklären Sie, wie sich ein lösbares LGS $Ax = b$ durch eine solche Zerlegung lösen lässt.

   - $A = QR$ mit $Q \in \mathbb{R}^{M \times M}$ orthogonal (also $QQ^T = I_M$) und $R = \begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix}$ mit $\tilde{R} \in \mathbb{R}^{N \times N}$
   - Eine QR-Zerlegung kann mit Householder-Transformationen errechnet werden
   - Löse erst $Qc = b$ ($Q^{-1} = Q^T$, also $c = Q^T b$)
   - $Rx = c$ durch Rückwärtssubstitution

3. Nennen Sie einen Vorteil und einen Nachteil der QR-Zerlegung gegenüber der LR- und Cholesky-Zerlegung.

   - Vorteil: Sehr stabiles Verfahren
   - Nachteil: Mit $\mathrm{O}(\frac{2}{3}N^3)$ am langsamsten

4. Wiederholen Sie Eigenschaften und die Struktur von Householder-Transformationen.

   - Orthogonale Matrix $Q = I_M - 2ww^T$ mit $w \in \mathbb{R}^M, w^T w = 1$ sodass $Qv = \sigma e^1 = (\sigma\ 0 \dots 0)^T$ für ein $v \in \mathbb{R}^M, v \neq 0$

5. Erklären Sie, wie eine Householder-Transformation Q effizient gespeichert werden kann und wie ein Produkt Qy berechnet wird

   - Die Householdervektoren w_i können in A gespeichert werden, für die r_{nn} wird ein N-dim. Vektor benötigt

4 Vorlesung 4

4.1 Lernkontrolle

1. Geben Sie zu einer Norm $||*||$ auf $\mathbb{R}^N$ die zugehörige Matrixnorm an und nennen Sie drei wichtige Beispiele solcher Normen.

   - Allgemeine Matrixnorm $||A|| := sup_{x \neq 0} \frac{||Ax||}{||x||}$ für $A \in \mathbb{R}^{N \times N}$
   - Zeilensummennorm $||A||_\infty$: $max_{m=1\dots N} \sum_{n=1}^{N} |a_{nm}|$
   - Spaltensummennorm $||A||_1$: $max_{m=1\dots N} \sum_{n=1}^{N} |a_{mn}|$
   - Spektralnorm $||A||_2$: $\sqrt{\text{größter EW von } A^T A}$

2. Formulieren Sie die Frage, der sich die Kondition eines Problems widmet.

- "Wie wirken sich Störungen der Eingabegröße auf die Lösung aus, unabh. vom Algorithmus?"

3. Erklären Sie, was genau eine kleine Kondition bzw. eine große Kondition $cond(A)$ über das Problem $Ax = b$ aussagt.

   - Die Kondition von A ist die Sensitivität des rel. Fehlers von A ggü. Störungen von b
   - Eine kleine $cond(A)$ bedeutet geringe Sensitivität

4. Nennen Sie mindestens drei Eigenschaften der Funktion $cond(A)$

   - $cond(A) = ||A||\ ||A^{-1}||$
   - $cond(A) = cond(\lambda A), \quad \lambda \in \mathbb{R}\backslash\{0\}$
   - $cond(A) = \frac{max_{||y||=1}||Ay||}{min_{||x||=1}||Ax||}$

5. Erklären Sie, warum die Kondition orthogonaler Matrizen bezüglich $||A||_2$ gleich 1 ist und wie sich die Kondition von symmetrischen und spd-Matrizen bezüglich dieser Norm berechnen lässt.

   - Da orthogonale Matrizen orthonormal bzgl. des Skalarprodutes sind
   - $cond(A) = \frac{betragl.\ größter\ EW}{betragl.\ kleinster\ EW}$, wenn A symm./spd.

5 Vorlesung 5

5.1 Lernkontrolle

Betrachtet wird das lineare Ausgleichsproblem $||Ax-b|| = min!$ zu $A \in \mathbb{R}^{M\times N}, b \in \mathbb{R}^M$

1. Erklären Sie, in welcher Situation man ein solches lineares Ausgleichsproblem betrachtet.

   - Im Fall $M > N$ ist $Ax = b$ überbestimmt und mglw. nicht lösbar

2. Beschreiben Sie den Zusammenhang des linearen Ausgleichsproblems mit der zugehörigen Normalengleichung und geben Sie diese an.

   - Die Lösung x für die Normalengleichung $A^TAx = A^Tb$ löst auch das Ausgleichsproblem

3. Erklären Sie, warum das lineare Ausgleichsproblem immer lösbar ist und unter welcher Bedingung an A eine eindeutige Lösung vorliegt

   - Die Normalengleichung ist immer lösbar (siehe Skript) $\Rightarrow$ Das lin. Ausgleichsproblem ist immer lösbar
   - Ist $Rang(A)$ maximal, so ist A^TA spd und das lin. Ausgleichsproblem ist eindeutig lösbar

4. Erklären Sie, wie und warum das lineare Ausgleichsproblem mit einer QR-Zerlegung von A gelöst werden kann.

   - Sind Q, R Matrizen aus der QR-Zerlegung mit $R = \begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix}$ mit $\tilde{R} \in \mathbb{R}^{N \times N}$, so ist $x = \tilde{R}^{-1}c$ mit c aus $Q^T b = \begin{pmatrix} c \\ d \end{pmatrix}$ die Lösung des Ausgleichproblems
   - Beweis siehe Skript S. 39 (gut zum Nachrechnen)

5. Erklären Sie, was eine Singulärwertzerlegung von A mit $Rang(A) = R$ ist und wie diese genutzt werden kann, um eine Lösung des linearen Ausgleichsproblems mit minimaler euklidischer Norm zu bestimmen

   - Singulärwertzerlegung von $A = U\Sigma V^T$ mit
     $U \in \mathbb{R}^{M \times M}, V \in \mathbb{R}^{N \times N}$ orthogonal,
     $\Sigma = \begin{pmatrix} \Sigma_R & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{M \times N}$,
     $\Sigma_R = diag(\sigma_1 \dots \sigma_R) \in \mathbb{R}^{R \times R}$
   - bei $M > N$ ist $x = \sum_{r=1}^{R} \frac{(u^r)^T b}{\sigma_r} v^r$ eine Lösung der Normalengleichung mit min. euklidischer Norm (siehe Übungen und ÜBs)

6 Vorlesung 6

6.1 Lernkontrolle

1. Formulieren Sie das Newton-Verfahren zur Berechnung von $\sqrt{a}$ für $a \in (1, 4)$, indem Sie die Funktion f und die Newton-Iteration explizit angeben, und erklären Sie, wie die Berechnung der Wurzel einer normierten Binärzahl $(1 + m)2^e$ darauf zurückgeführt werden kann

   - $f(x) = x^2 - a \Rightarrow x = \sqrt{a}$
   - $2x_k d_k = -x^2 + a \Leftrightarrow 2d_k = \frac{{x_k}^2}{x_k} + \frac{a}{x_k} \Leftrightarrow d_k = \frac{1}{2}(x_k + \frac{a}{x_k})$

2. Geben Sie die allgemeine Newton-Iteration zur Nullstellengleichung $f(x) = 0$ einer Funktion $f : \mathbb{R}^N \to \mathbb{R}^N$ an.

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

3. Nennen Sie zwei wichtige Aspekte im Hinblick auf die Konvergenz des Newton-Verfahrens.

   - Das Newton-Verfahren konvergiert lokal quadratisch
   - Das Newton-Verfahren konvergiert nur bei guter Wahl von x_0, sonst kann es divergieren

4. Erklären Sie die geometrische Deutung des Newton-Verfahrens im Fall $N = 1$

   - Das Newton-Verfahren legt im ersten Schritt eine Tangente an f mit Nullstelle x_0 an
   - In den darauffolgenden Schritten wird die Nullstelle der neuen Tangente der Berührungspunkt der alten Tangente mit f

5. Erklären Sie, was mit dem Vereinfachten Newton-Verfahren gemeint ist und was die Vereinfachung für Konsequenzen hat.

   - Das berechnen von $f'(x_k)$ in $d_k f'(x_k) = -f(x_k)$ ist sehr kompliziert, da im mehrdim. Fall eine Jacobimatrix berechnetw erden muss
   - Beim vereinfachten Newton-Verfahren wird anstelle von $f'(x_k)$ immer die Matrix $A \approx f'(x_0)$ genommen, also $d_k A = -f(x_k)$
   - Die Konvergenz des Verfahrens wird dann linear anstelle von quadratisch
   - Für die Matrix A kann im ersten Schritt eine LR-Zerlegung errechnet werden, welche dann immer weiter genutzt werden kann

7 Vorlesung 7

7.1 Lernkontrolle

1. Formulieren Sie die Fragestellung der Polynominterpolation.

   - Gegeben sind $N+1$ paarw. verschiedene Stützstellen $(x_0, f(x_0)), \dots (x_N, f(x_N))$
   - Wir suchen ein Polynom p mit $deg(p) \leq N$ mit $p(x_i) = f(x_i)$ für alle $i \in (0, N)$

2. Begründen Sie, warum das Problem der Polynominterpolation eine eindeutige Lösung besitzt

   - Ein Polynom p mit $deg(p) \leq N$ ist durch $N + 1$ Stützstellen exakt definiert
   - Bsp.: Eine Gerade aus $\mathbb{P}^1$ ist durch 2 Punkte exakt definiert

3. Vollziehen Sie nach, welche Größe als Maß für die Kondition des Problems gilt und warum Interpolationspolynome von hohem Grad zumindest bei äquidistanten Stützstellen mit Vorsicht zu genießen sind.

   - Das Maß für die Kondition der Interpolation ist die Lebesque-Konstante: $\Lambda_N := max_{x \in [a,b]} \sum_{n=0}^{N} |L_n(x)|$
   - Die Lebesque-Konstante misst die Auswirkung von Störungen der Stützstellen $(x_0, \tilde{f}(x_0)), \dots (x_N, \tilde{f}(x_N))$ und dem gestörten Interpolationspolynom $\tilde{p}$

- $|p(x) - \tilde{p}(x)| = |\sum (f_n - \tilde{f}_n) L_n(x)|$
  $\leq \sum |(f_n - \tilde{f}_n)||L_n(x)|$
  $\leq max_{n=0\ldots N}|(f_n - \tilde{f}_n)| \sum |L_n(x)|$
- Bei Polynomen mit hohem Grad und äquidistanten Stützstellen wächst Λ_N sehr stark an

4. Geben Sie die Newton-Darstellung des Interpolationspolynoms an und vergleichen Sie diese mit der Lagrange-Darstellung.

   - Bei der Newton-Darstellung wird versucht, das Interpolationspolynom schrittweise aus Polynomen mit niedrigerem Grad aufzubauen
   - Newton-Darstellung: $p_{0,N}(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \ldots + a_N(x - x_0) \cdot \ldots \cdot (x - x_{N-1})$
   - Lagrange-Darstellung: $p(x) = \sum_{n=0}^{N} f_n L_n(x)$ mit $L_n(x) = \prod_{m=0, m\neq n}^{N} \frac{x - x_m}{x_n - x_m}$

5. Formulieren Sie das Lemma von Aitken und erklären Sie, wie daraus die Formel für die dividierten Differenzen hergeleitet werden kann.

   - Das Lemma besagt, dass die Interpolationspolynome rekursiv aufgabeut werden können
   - Die Leitkoeffizieniten der Rekursion des Lemmas sind $f_{n,k} = \frac{f_{n,k-1} - f_{n+1,k}}{x_n - x_k}$, welche auch die a_i beim Polynom in Newton-Darstellung sind