Technische Universität Hamburg-Harburg **WS 2007/2008**
Institut für Numerische Simulation
Dr. Jens-Peter M. Zemke

Klausur zur Vordiplom-Prüfung

Numerische Verfahren

17. März 2008

Sie haben **90** Minuten Zeit zum Bearbeiten der Klausur.

Bitte kennzeichnen Sie jedes Blatt mit Ihrem Namen und Ihrer Matrikelnummer in DRUCKSCHRIFT.

Tragen Sie bitte zunächst Ihren Namen, Ihren Vornamen und Ihre Matrikelnummer in **Druckschrift** in die folgenden jeweils dafür vorgesehenen Felder ein.

Diese Eintragungen werden auf Datenträger gespeichert.

Name:														
Vorname:														
Matr.-Nr.														

Ich bin darüber belehrt worden, dass meine Ausarbeitung nur dann als Prüfungsleistung bewertet wird, wenn die Nachprüfung durch das Zentrale Prüfungsamt der TUHH meine offizielle Zulassung vor Beginn der Prüfung ergibt.

(Unterschrift)

Lösen Sie die folgenden 12 Aufgaben!

Aufgabe	1	2	3	4	5	6	7	8	9	10	11	12
Punkte												

Σ	

Aufgabe 1: (4 Punkte)
Zur Berechnung der Nullstellen eines reellen quadratischen Polynoms $z(x) = x^2 + px + q$ wird oft die sogenannte „pq-Formel“

$$x_\pm = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q}$$

für *beide* Nullstellen verwendet. Da

$$(x - x_+)(x - x_-) = x^2 - (x_+ + x_-)x + x_+x_- \quad \Rightarrow \quad q = x_+ \cdot x_-$$

gilt, könnte man aber auch x_+ aus der „pq-Formel“ und x_- aus

$$x_- = \frac{q}{x_+}$$

bestimmen. Welche dieser beiden mathematisch äquivalenten Varianten würden Sie im Fall $p = -10^8$ und $q = 1$ in Gleitkommaarithmetik verwenden? Warum?

Lösung zu Aufgabe 1: Die „pq-Formel“ basiert für x_+ auf der Addition zweier positiver Terme, also ist hier keine Auslöschung zu erwarten. Für die Berechnung von x_- werden allerdings zwei fast gleich grosse Terme voneinander abgezogen, und die Berechnung mittels des zweiten Weges ist bei den gegebenen Daten deutlich besser konditioniert. (4 Punkte)

Unter Matlab/Octave liefert die naive Auswertung der „pq-Formel“ approximativ

$$\mathrm{fl}_1(x_+) = 10^8, \quad \mathrm{fl}_1(x_-) \approx 7.45058059692383 \cdot 10^{-9},$$

wohingegen die Auswertung nach der zweiten Idee als Näherung für x_+ logischerweise denselben Wert und für x_- den approximativen Wert

$$\mathrm{fl}_2(x_+) = 10^8, \quad \mathrm{fl}_2(x_-) = 10^{-8}$$

liefert. Die echten Werte unterscheiden sich von denen der zweiten Berechnung nur um

$$\frac{|x_+ - \mathrm{fl}_2(x_+)|}{x_+} \leqslant 1.0000000000000003 \cdot 10^{-16} = 10^{-16} + 3 \cdot 10^{-32},$$
$$\frac{|x_- - \mathrm{fl}_2(x_-)|}{x_-} \leqslant 1.0000000000000001 \cdot 10^{-16} = 10^{-16} + 1 \cdot 10^{-32},$$

wohingegen

$$\frac{|x_- - \mathrm{fl}_1(x_-)|}{x_-} \leqslant 0.2549419403076171 < \frac{1}{4} + \frac{5}{1000}$$

gilt. Der zweite Weg bestimmt also beide Nullstellen bis auf einen relativen Fehler von 10^{-16}, also bis auf (IEEE 754 double precision) Maschinengenauigkeit.

Aufgabe 2: (5+5 Punkte)
Geben Sie die Lagrangsche Form der Polynominterpolation wieder. Berechnen Sie das Interpolationspolynom der Funktion

$$f(x) = \sin(\pi x)$$

zu den Punkten

i	0	1	2	3	4
x_i	-2	-1	0	1	2

.

Lösung zu Aufgabe 2: Die Lagrangsche Interpolation zu den Datenpaaren $\{(x_j, y_j)\}_{j=0}^n$ mit $x_i \neq x_j$ für alle $i \neq j$ ist eine Vorschrift um das eindeutige Polynom p vom Höchstgrad n zu berechnen, dass die Interpolationsbedingungen

$$p(x_j) = y_j, \quad j = 0, 1, \ldots, n$$

erfüllt. Es läßt sich leicht explizit angeben als

$$p(x) = \sum_{j=0}^{n} y_j \ell_j(x),$$

wobei die Lagrangeschen Basispolynome gegeben sind durch die Interpolationspolynome zu den speziellen Aufgaben

$$\ell_j(x_i) = \delta_{ij}, \qquad i = 0, 1, \ldots, n, \quad j = 0, 1, \ldots, n,$$

und explizit angegeben lauten

$$\ell_j(x) = \prod_{\substack{i=0 \\ i \neq j}}^{n} (x - x_i) \Bigg/ \prod_{\substack{i=0 \\ i \neq j}}^{n} (x_j - x_i) \ , \quad j = 0, \ldots, n. \qquad \text{(5 Punkte)}$$

Die Interpolationsaufgabe läßt sich auffassen als die Interpolation der Datenpaare

x_i	-2	-1	0	1	2
y_i	0	0	0	0	0

.

Das Interpolationspolynom p_4 hat den *Höchst*grad vier und ist eindeutig bestimmt. Damit ist das Interpolationspolynom das Nullpolynom. (5 Punkte)

Natürlich konnte man p_4 auch mittels der Lagrangeschen oder der Newtonschen Form der Polynominterpolation berechnen und dabei zu dem selben Ergebnis kommen.

Aufgabe 3: (5+5 Punkte)
Beschreiben Sie die Grundidee der Newtonschen Form der Polynominterpolation und geben Sie diese in Formeln wieder.

Lösung zu Aufgabe 3: Die Grundidee der Newtonschen Form der Polyninterpolation ist es, bei einem bereits berechneten Interpolationspolynom p_n vom Höchstgrad n zu den $n+1$ Datenpaaren $\{(x_i, y_i)\}_{i=0}^n$ mit paarweise verschiedenen Knoten $\{x_i\}_{i=1}^n$ nach Hinzufügen eines weiteren Datums (x_{n+1}, y_{n+1}), wobei selbstverständlich $x_{n+1} \neq x_i$ für alle $i = 0, \ldots, n$, schnell das zugehörige Interpolationspolynom zu den jetzt $n+2$ Datenpaaren berechnen zu können. Dieses geschieht mittels der Addition des bereits berechneten Polynomes p_n und eines polynomialen Termes vom Grad n oder 0, der allen vorherigen Stellen gleich Null ist und dessen Vorfaktor a so gewählt ist, dass

$$p_{n+1}(x_{n+1}) = p_n(x_{n+1}) + a \prod_{i=1}^{n} (x_{n+1} - x_i) = y_{n+1},$$

also

$$a = \frac{y_{n+1} - p_n(x_{n+1})}{\prod_{i=1}^{n}(x_{n+1} - x_i)}$$

gilt. (5 Punkte)

Die Newtonsche Interpolationsformel ist gegeben durch

$$p_n(x) = \sum_{j=0}^{n} [x_0, \ldots, x_j] \prod_{k=0}^{j-1} (x - x_k),$$

wobei die sogenannten dividierten Differenzen $[x_i, \ldots, x_j]$ rekursiv definiert sind durch

$$\begin{aligned} [x_j] &:= y_j, \\ [x_i, \ldots, x_j] &:= \frac{[x_{i+1}, \ldots, x_j] - [x_i, \ldots, x_{k-1}]}{x_j - x_i}, \quad j > i \geqslant 0. \end{aligned}$$

(5 Punkte)

Aufgabe 4: (4+4 Punkte)
Was ist eine Quadraturformel? Berechnen Sie approximativ das Integral

$$\int_0^\pi \cos(x)\, dx = \sin(\pi) - \sin(0) = 0$$

mit einer Quadraturformel.

Lösung zu Aufgabe 4: Eine Quadraturformel ist eine Formel zur numerischen Annäherung eines Integrales, welche nur auf Funktionsauswertungen der Funktion unter dem Integral basiert und allgemein die Gestalt

$$\int_0^1 f(x)\, dx \approx \sum_{j=0}^{n} \omega_j f(x_j)$$

hat. Die Angabe der Auswertungsstellen, der Stützstellen $\{x_j\}_{j=0}^n$ und der sogenannten Gewichte $\{\omega_j\}_{j=0}^n$ bestimmt in eindeutiger Weise eine Quadraturformel. (4 Punkte)

Da in der Aufgabenstellung keine Rede von der Art der zu verwendenden Quadraturformel zur Approximation war, ist viel Freiheit gegeben. Wir verwenden exemplarisch erst die Mittelpunktregel, danach die Trapezregel und zu guter Letzt die optimale Gauß-Formel mit zwei Knoten.

Die Mittelpunkt- oder Rechteckregel ersetzt das Integral durch die Auswertung des Integranden in der Mitte des Intervalles und berechnet den Flächeninhalt des (vorzeichenbehafteten) Rechtecks, welches durch diese (vorzeichenbehaftete) Höhe mal der Länge, also, der Intervallbreite gegeben ist,

$$\begin{aligned}\int_0^\pi \cos(x)\,dx &\approx \cos\left(\frac{\pi}{2}\right)\cdot(\pi-0)\\ &= 0\cdot\pi = 0.\end{aligned}$$ (4 Punkte)

Die Trapezregel verwendet das von den Funktionswerten am Anfang und am Ende des Intervalles aufgespannte Trapez. Der (vorzeichenbehaftete) Flächeninhalt dieses Trapezes berechnet sich aus dem Mittelwert der beiden Funktionswerte mal der Länge des Intervalles, also gemäß

$$\begin{aligned}\int_0^\pi \cos(x)\,dx &\approx \frac{1}{2}\left(\cos(\pi)+\cos(0)\right)\cdot(\pi-0)\\ &= (-1+1)\cdot\pi = 0.\end{aligned}$$ (4 Punkte)

Die optimale Gauß-Formel mit zwei Knoten basiert in Nachschlagewerken meist auf dem Intervall $[-1,1]$, ist im Skript aber bereits für das Referenzintervall $[0,1]$ gegeben und hat dort die Knoten

$$x_1 = \frac{1}{2}\left(1-\frac{1}{\sqrt{3}}\right),\quad x_2 = \frac{1}{2}\left(1+\frac{1}{\sqrt{3}}\right)$$

und die Gewichte

$$w_1 = w_2 = \frac{1}{2}.$$

Die lineare Transformation $y = \pi\cdot x$ bildet das Intervall $[0,1]$ auf das Intervall $[0,\pi]$ ab und liefert die (neuen, aber auch mit x_1 und x_2 bezeichneten) Knoten

$$x_1 = \frac{\pi}{2}\left(1-\frac{1}{\sqrt{3}}\right),\quad x_2 = \frac{\pi}{2}\left(1+\frac{1}{\sqrt{3}}\right).$$

Damit ergibt die Gauß-Quadratur mit zwei Knoten unter Berücksichtigung des Additionstheoremes des Kosinus und der Feststellung, dass der Kosinus eine ge-

rade Funktion ist, den Wert

$$\int_0^\pi \cos(x)\,dx \approx \frac{1}{2}\left(\cos\left(\frac{\pi}{2}\left(1-\frac{1}{\sqrt{3}}\right)\right)+\cos\left(\frac{\pi}{2}\left(1+\frac{1}{\sqrt{3}}\right)\right)\right)$$
$$= \cos\left(\frac{\pi}{2}\right)\cos\left(\frac{\pi}{2\sqrt{3}}\right) = 0. \qquad \text{(4 Punkte)}$$

Man sieht also, dass bereits einfache Quadraturformeln gute Ergebnisse liefern können, aber das sollte nicht darüber hinwegtäuschen, dass der Fehler bei den genaueren Formeln (also solchen mit höherer Fehlerordnung und kleinerer Fehlerkonstante) *garantiert* für beliebige genügend oft stetig differenzierbare Funktionen kleiner beschränkt ist.

Aufgabe 5: (5+5 Punkte)
Welche Varianten der LR-Zerlegung kennen Sie? Welche würden Sie im Hinblick auf die numerische Stabilität und den nötigen Arbeitsaufwand verwenden?

Lösung zu Aufgabe 5: Es gibt drei Varianten der LR-Zerlegung, nämlich die LR-Zerlegung ohne Pivotisierung,

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R},$$

wobei $\boldsymbol{L}$ eine linke untere Dreiecksmatrix mit Einsen auf der Diagonalen und $\boldsymbol{R}$ eine rechte obere Dreiecksmatrix ist, die LR-Zerlegung mit partieller Pivotisierung (auch Spaltenpivotsuche genannt),

$$\boldsymbol{P}\boldsymbol{A} = \boldsymbol{L}\boldsymbol{R},$$

wobei $\boldsymbol{L}$ und $\boldsymbol{R}$ wie vorher gewählt sind und $\boldsymbol{P}$ eine Permutationsmatrix ist, welche das größte Element in der jeweiligen aktuell behandelten Spalte in die Diagonale tauscht und die LR-Zerlegung mit vollständiger Pivotisierung,

$$\boldsymbol{P}\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{R},$$

wobei $\boldsymbol{L}$, $\boldsymbol{R}$ und $\boldsymbol{P}$ wie oben sind und $\boldsymbol{Q}$ zusammen mit $\boldsymbol{P}$ dazu verwendet wird um das in der aktuell behandelten Restmatrix größte Element in die aktuell behandelte Position in der Diagonalen zu tauschen. (5 Punkte)

Am billigsten ist die einfache LR-Zerlegung zu implementieren, aber diese ist numerisch nicht stabil, die LR-Zerlegung existiert oft auch gar nicht, bekanntestes Beispiel ist

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Die LR-Zerlegung mit partieller Pivotisierung reicht theoretisch ohne Rundungsfehler aus, eine reguläre Matrix zu zerlegen, ist aber immer noch rundungsfehleranfällig, siehe das Beispiel im Skript. Die LR-Zerlegung mit vollständiger Pivotisierung ist am aufwändigsten, ist aber numerisch sehr stabil. Meist würde man,

sofern keine begründeten Ängste vorliegen, die LR-Zerlegung mit partieller Pivotisierung verwenden und bei unglaubwürdigen Resultaten auf die vollständige Pivotisierung zurückgreifen. (5 Punkte)

Wenn die Ergebnisse dann immer noch nicht die Praxis reflektieren, würde man zu Regularisierung greifen, da dann (vorausgesetzt, es gibt keine groben Modellierungsfehler) meist die Kondition der Matrix zu hoch ist, um sinnvolle numerische Ergebnisse zu erhalten.

Aufgabe 6: (5 Punkte)
Lösen Sie das Gleichungssystem

$$\begin{pmatrix} 110 & 1 \\ 100 & 11 \end{pmatrix} \boldsymbol{x} = \begin{pmatrix} 110 \\ 110 \end{pmatrix}$$

in zweistelliger dezimaler Fließkommazahlenarithmetik, also unter

$$\mathrm{fl}(x \circ y) := \text{Rundung des Ergebnisses von } x \circ y \text{ auf zwei Stellen}$$

für jede Operation $\circ \in \{+, -, \cdot, /\}$, es gilt also z. B.

$$\mathrm{fl}(100 + 10) = 110, \quad \mathrm{fl}(120 + 3) = 120, \quad \mathrm{fl}(120 + 5) = 130,$$

mittels der LR-Zerlegung.

Lösung zu Aufgabe 6: Die LR-Zerlegung kann hier ohne Pivotisierung erfolgen, das größte Element ist bereits in der Position $(1, 1)$. Da die rechte Seite bereits gegeben ist, transformieren wir die Elemente gleich mit, arbeiten also mit der Matrix $\begin{pmatrix} \boldsymbol{A} & \boldsymbol{b} \end{pmatrix}$,

$$\left(\begin{array}{cc|c} 110 & 1 & 110 \\ 100 & 11 & 110 \end{array}\right).$$

Zur Berechnung des ersten und einzigen interessanten Wertes in L wird 100 durch 110 in Fließkommazahlenarithmetik geteilt,

$$\mathrm{fl}(100/110) = \mathrm{fl}(0.\overline{90}) = 0.91.$$

Nun wird, wiederum in Fließkommazahlenarithmetik, die erste Zeile mit 0.91 malgenommen,

$$\begin{aligned} \mathrm{fl}\left(0.91 \cdot \left(\begin{array}{cc|c} 110 & 1 & 110 \end{array}\right)\right) &= \left(\begin{array}{cc|c} \mathrm{fl}(0.91 \cdot 110) & \mathrm{fl}(0.91 \cdot 1) & \mathrm{fl}(0.91 \cdot 110) \end{array}\right) \\ &= \left(\begin{array}{cc|c} \mathrm{fl}(100.1) & \mathrm{fl}(0.91) & \mathrm{fl}(100.1) \end{array}\right) \\ &= \left(\begin{array}{cc|c} 100 & 0.91 & 100 \end{array}\right). \end{aligned}$$

Jetzt wird, selbstverständlich auch in Fließkommazahlenarithmetik, die so erhaltene Zeile von der zweiten Zeile abgezogen,

$$\begin{aligned} &\mathrm{fl}\left(\left(\begin{array}{cc|c} 100 & 11 & 110 \end{array}\right) - \left(\begin{array}{cc|c} 100 & 0.91 & 100 \end{array}\right)\right) \\ &\qquad = \left(\begin{array}{cc|c} \mathrm{fl}(100 - 100) & \mathrm{fl}(11 - 0.91) & \mathrm{fl}(110 - 100) \end{array}\right) \\ &\qquad\qquad = \left(\begin{array}{cc|c} \mathrm{fl}(0) & \mathrm{fl}(10.01) & \mathrm{fl}(10) \end{array}\right) = \left(\begin{array}{cc|c} 0 & 10 & 10 \end{array}\right). \end{aligned}$$

Die resultierende in Fließkommazahlenarithmetik berechnete LR-Zerlegung hat demnach in Kompaktschreibweise die Form

$$\left(\begin{array}{cc|c} 110 & 1 & 110 \\ \hline 0.91 & 10 & 10 \end{array}\right),$$

womit wir bereits Approximationen für $\boldsymbol{L}$, $\boldsymbol{R}$ aus $\boldsymbol{A} = \boldsymbol{LR}$ und $\boldsymbol{y} = \boldsymbol{L}^{-1}\boldsymbol{b}$ haben, nämlich

$$\tilde{\boldsymbol{L}} = \begin{pmatrix} 1 & 0 \\ 0.91 & 1 \end{pmatrix}, \quad \tilde{\boldsymbol{R}} = \begin{pmatrix} 110 & 1 \\ 0 & 10 \end{pmatrix} \quad \text{und} \quad \tilde{\boldsymbol{y}} = \begin{pmatrix} 110 \\ 10 \end{pmatrix}.$$

Um das Gleichungssystem in Fließkommazahlenarithmetik zu lösen, lösen wir das Gleichungssystem

$$\tilde{\boldsymbol{R}}\boldsymbol{x} = \begin{pmatrix} 110 & 1 \\ 0 & 10 \end{pmatrix} \boldsymbol{x} = \begin{pmatrix} 110 \\ 10 \end{pmatrix} = \tilde{\boldsymbol{y}}$$

mittels Rückwärtsauflösen in Fließkommazahlenarithmetik. Die letzte Komponente der genäherten Lösung ist durch

$$\tilde{x}_2 = \mathrm{fl}(10/10) = \mathrm{fl}(1) = 1$$

gegeben, die erste Komponente erhält man aus

$$\tilde{x}_1 = \mathrm{fl}\left(\mathrm{fl}(110 - 1)/110\right) = \mathrm{fl}\left(\mathrm{fl}(109)/110\right) = \mathrm{fl}\left(110/110\right) = \mathrm{fl}(1) = 1.$$

Die in Fließkommazahlenarithmetik berechntete (approximative) Lösung ist demnach

$$\tilde{\boldsymbol{x}} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \qquad \text{(5 Punkte)}$$

Aufgabe 7: (5+7 Punkte)
Bestimmen Sie die Pseudonormallösungen der beiden Ausgleichsprobleme

$$\left\| \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \\ 1 & 5 \\ 1 & 3 \\ 1 & 1 \end{pmatrix} \boldsymbol{x}^1 - \begin{pmatrix} 1 \\ 2 \\ 3 \\ 3 \\ 2 \\ 1 \end{pmatrix} \right\|_2 = \min, \quad \left\| \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \boldsymbol{x}^2 - \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \right\|_2 = \min.$$

Lösung zu Aufgabe 7: Das erste Ausgleichsproblem wird durch den Vektor

$$\boldsymbol{x}^1 = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

gelöst, wie man leicht sehen kann, da die erste rechte Seite im Spann der Spaltenvektoren der ersten Matrix liegt und die Matrix den vollen Rang Zwei hat. Wenn man diese Lösung nicht sieht, so kann man natürlich mittels Normalgleichungen, QR-Zerlegung oder SVD die Pseudonormallösung bestimmen. (5 Punkte)

Das zweite Ausgleichsproblem hat eine Rang-Eins-Matrix, da

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} = \underbrace{\frac{1}{2}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}}_{=\boldsymbol{U}_1} \underbrace{\sqrt{12}}_{=\boldsymbol{\Sigma}_1} \underbrace{\frac{1}{\sqrt{3}}\begin{pmatrix} 1 & 1 & 1 \end{pmatrix}}_{=\boldsymbol{V}_1^T},$$

wobei die letzte Schreibung $\boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^T$ bereits die (ökonomische Variante der) SVD dieser Rang-Eins-Matrix bezeichnet. Die Pseudonormallösung $\hat{\boldsymbol{x}}^2$ ist charakterisiert durch die Multiplikation der rechten Seite mit der Pseudoinversen,

$$\begin{aligned} \hat{\boldsymbol{x}}^2 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}^{\dagger} \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} &= \boldsymbol{V}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{U}_1^T \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \\ &= \frac{1}{\sqrt{3}}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \frac{1}{\sqrt{12}} \cdot \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \\ &= \frac{1}{12}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} (1+2+3+4) = \frac{5}{6}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}. \end{aligned}$$

(7 Punkte)

Aufgabe 8: (8+4+4 Punkte)
Beschreiben Sie die Potenzmethode. Ist es möglich, dass die Potenzmethode den Eigenwert 2 der Matrix

$$\boldsymbol{G} = \begin{pmatrix} 1 & 1 & -1 \\ 0 & 2 & -2 \\ 0 & 0 & -2 \end{pmatrix}$$

approximiert? Wenn nein, warum nicht? Wenn ja, ist dieses für einen zufällig gewählten Startvektor überhaupt wahrscheinlich?

Lösung zu Aufgabe 8: Die Potenzmethode geht aus von einer quadratischen Matrix $\boldsymbol{A}$ und einem gegebenen (bereits normalisierten) Startvektor $\boldsymbol{u}^0$. Für eine (gewünschte) Anzahl von Indizes $k = 1, 2, \ldots$ wird dann rekursiv (z. B.) berechnet:

$$\boldsymbol{v} = \boldsymbol{A}\boldsymbol{u}^{k-1}, \quad \boldsymbol{u}^k = \boldsymbol{v}/\boldsymbol{\ell}^H\boldsymbol{v},$$

wobei $\boldsymbol{\ell}$ ein fest vorgegebener Vektor ist. Unter geeigneten Vorraussetzungen konvergiert dann die Folge $\{\boldsymbol{u}^k\}_{k=0}^{\infty}$ gegen einen (meist den zum betragsgrößten Eigenwert gehörigen) Eigenvektor und die Skalierungsgröße $\boldsymbol{\ell}^H\boldsymbol{v}$ gegen den zugehörigen Eigenwert. (8 Punkte)

Es ist möglich, dass die Potenzmethode nicht den betragsgrößten Eigenwert und zugehörigen Eigenvektor zurückgibt. Für geeignete Startvektoren kann jedes Eigenpaar zurückgegeben werde, z. B. wenn der Startvektor gleich dem (gesuchten) Eigenvektor ist. (4 Punkte)

Konvergenz gegen den Eigenwert 2 ist der Fall, wenn der Startvektor keinen Anteil am Eigenvektor zum Eigenwert -2 enthält, aber einen nichttrivialen Anteil am Eigenvektor zum Eigenwert 2 hat. Ein zufällig gewählter Startvektor hat mit großer Wahrscheinlichkeit Anteile an jedem Eigenvektor, damit ist die Konvergenz gegen den Eigenwert 2 unter diesen Umständen nahezu (aber nicht völlig) ausgeschlossen. (4 Punkte)

Die Anteile des Startvektors lassen sich in diesem Fall leicht beschreiben. Ein Eigenvektor zum Eigenwert 1 ist gegeben durch

$$\boldsymbol{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},$$

ein Eigenvektor zum Eigenwert 2 durch

$$\boldsymbol{v}_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix},$$

und ein Eigenvektor zum Eigenwert -2 läßt sich herleiten zu

$$\boldsymbol{v}_3 = \begin{pmatrix} 1 \\ 3 \\ 6 \end{pmatrix}.$$

Damit ist eine Matrix von Eigenvektoren und ihre Inverse gegeben als

$$\boldsymbol{V} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 6 \end{pmatrix}, \qquad \boldsymbol{V}^{-1} = \frac{1}{18}\begin{pmatrix} 18 & -18 & 6 \\ 0 & 18 & -9 \\ 0 & 0 & 3 \end{pmatrix}.$$

Die Anteile $\{\alpha_i\}_{i=1}^3$ von $\boldsymbol{u}^0$ an den einzelnen Eigenvektoren erhält man aus

$$\boldsymbol{u}^0 = \sum_{i=1}^{3} \boldsymbol{v}_i\alpha_i = \boldsymbol{V}\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix},$$

also gemäß

$$\begin{pmatrix}\alpha_1\\ \alpha_2\\ \alpha_3\end{pmatrix} = \boldsymbol{V}^{-1}\boldsymbol{u}^0 = \frac{1}{18}\begin{pmatrix}18 & -18 & 6\\ 0 & 18 & -9\\ 0 & 0 & 3\end{pmatrix}\begin{pmatrix}u_{10}\\ u_{20}\\ u_{30}\end{pmatrix}.$$

Um den Eigenwert 2 zu approximieren, muss $\alpha_3 = u_{30}/6$ gleich Null sein und (unter der Vorraussetzung, dass bereits u_{30} gleich Null ist) $\alpha_2 = u_{20}$ ungleich Null sein, mit anderen Worten, jeder Startvektor der Form

$$\boldsymbol{u}^0 = \begin{pmatrix}\star\\ c\\ 0\end{pmatrix}, \quad c \neq 0$$

mit beliebigem Eintrag in der ersten Komponente führt zur Konvergenz der Potenzmethode gegen den Eigenwert 2. Ein zufällig gewählter Startvektor wird mit an Sicherheit grenzender Wahrscheinlichkeit einen Eintrag ungleich Null in der letzten Komponente haben, womit bei einer Skalierung mittels des zweiten Einheitsvektors

$$\boldsymbol{\ell} = \boldsymbol{e}_2 = \begin{pmatrix}0\\ 1\\ 0\end{pmatrix}$$

die Itertierten der Potenzmethode sich im Limes wie die Glieder einer alternierenden Folge, nämlich wie

$$\begin{aligned}\boldsymbol{u}^{2k-1} &\to \frac{\alpha_2\boldsymbol{v}_2 - \alpha_3\boldsymbol{v}_3}{\alpha_2 - 3\alpha_3},\\ \boldsymbol{u}^{2k} &\to \frac{\alpha_2\boldsymbol{v}_2 + \alpha_3\boldsymbol{v}_3}{\alpha_2 + 3\alpha_3},\end{aligned}$$

verhalten, wenn der Startvektor die Gestalt

$$\boldsymbol{u}^0 = \alpha_1\boldsymbol{v}_1 + \alpha_2\boldsymbol{v}_2 + \alpha_3\boldsymbol{v}_3$$

hat. Insbesondere versagt die Skalierung (spätestens bei Konvergenz) bei $\alpha_2 = \pm 3\alpha_3$, was man z. B. an

$$3\boldsymbol{v}_2 - \boldsymbol{v}_2 = \begin{pmatrix}3\\ 3\\ 0\end{pmatrix} - \begin{pmatrix}1\\ 3\\ 6\end{pmatrix} = \begin{pmatrix}2\\ 0\\ -6\end{pmatrix}$$

sofort ablesen kann.

Aufgabe 9: (6 Punkte)
Beschreiben Sie die Berechnung der QR-Zerlegung einer Matrix $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ unter Verwendung von Householdermatrizen (Spiegelungen)

$$\boldsymbol{H}(\boldsymbol{x}) = \boldsymbol{E} - 2\frac{\boldsymbol{x}\boldsymbol{x}^T}{\boldsymbol{x}^T\boldsymbol{x}}, \quad \boldsymbol{x} \in \mathbb{R}^n, \quad \boldsymbol{x} \neq \boldsymbol{o}_n.$$

Der Vektor $\boldsymbol{o}_n \in \mathbb{R}^n$ bezeichne hierbei den Nullvektor, $\boldsymbol{E}$ die $n \times n$-Einheitsmatrix.

Lösung zu Aufgabe 9: Die Idee der QR-Zerlegung unter Verwendung von Householder-Spiegelungen ist es, eine Matrix $\boldsymbol{A}$ spaltenweise in eine obere Dreiecksmatrix zu transformieren. Dazu sucht man einen (Normalen-)Vektor, an dessen zugehöriger Hyperebene der Vektor in der ersten Spalte von $\boldsymbol{A}$ auf ein Vielfaches des ersten Einheitsvektors abgebildet wird. Eine Spiegelung ist eine Kongruenztransfomation, ändert also keine Längen, und das Vielfache des ersten Einheitsvektors ist bis auf das Vorzeichen bereits eindeutig bestimmt zu

$$\pm\|\boldsymbol{a}_1\|_2\boldsymbol{e}_1$$

wobei $\boldsymbol{a}_1$ die erste Spalte von $\boldsymbol{A}$ bezeichne. Die Differenz der Vektoren vorher, also $\boldsymbol{a}_1$, und nachher, also $\pm\|\boldsymbol{a}_1\|_2\boldsymbol{e}_1$, steht senkrecht auf der Hyperebene, welche wir für die Spiegelung nehmen wollen, also bildet

$$\boldsymbol{H}_1 := \boldsymbol{H}(\boldsymbol{z}_\pm) \quad \text{wobei} \quad \boldsymbol{z}_\pm := \boldsymbol{a}_1 \mp \|\boldsymbol{a}_1\|_2\boldsymbol{e}_1$$

die erste Spalte auf $\pm\|\boldsymbol{a}_1\|_2\boldsymbol{e}_1$ ab,

$$\boldsymbol{H}(\boldsymbol{z}_\pm)\boldsymbol{a}_1 = \pm\|\boldsymbol{a}_1\|_2\boldsymbol{e}_1,$$

das Vorzeichen sucht man jetzt gemeinhin so aus, dass in der ersten Komponente von $\boldsymbol{z}_\pm$ keine Auslöschung stattfindet. Jetzt hat man, da Householder-Matrizen orthogonal und symmetrisch sind, also die Inverse gleich der Matrix ist, $\boldsymbol{A}$ bereits zerlegt in

$$\boldsymbol{A} = \boldsymbol{H}_1\boldsymbol{A}_1,$$

wobei $\boldsymbol{A}_1$ in der ersten Spalte ein Vielfaches des ersten Einheitsvektors stehen hat. Nun spiegelt man in der Matrix, die man erhält, wenn man von $\boldsymbol{A}_1$ die erste Spalte und Zeile entfernt, die neue erste Spalte in den (jetzt um Eins kürzeren) Einheitsvektor mit einer Matrix $\boldsymbol{H}_2 \in \mathbb{R}^{(n-1)\times(n-1)}$ analog zu dem eben beschriebenen Vorgehen. Damit gilt

$$\boldsymbol{A}_1 = \begin{pmatrix} 1 & \boldsymbol{o}_n^T \\ \boldsymbol{o}_n & \boldsymbol{H}_2 \end{pmatrix} \boldsymbol{A}_2,$$

zusammen also bereits

$$\boldsymbol{A} = \boldsymbol{H}_1 \begin{pmatrix} 1 & \boldsymbol{o}_n^T \\ \boldsymbol{o}_n & \boldsymbol{H}_2 \end{pmatrix} \boldsymbol{A}_2.$$

Nun wird dieses Schema iterativ durchgeführt, bis die Matrix, die nach Multiplikation mit einer Householder-Matrix und wegnehmen der ersten Zeile und Spalte keine Elemente mehr hat. Am Ende gilt (hier am Beispiel des für ein Ausgleichsproblem typischen Falles $m \leqslant n$)

$$\boldsymbol{A} = \boldsymbol{H}_1 \begin{pmatrix} 1 & \boldsymbol{o}_n^T \\ \boldsymbol{o}_n & \boldsymbol{H}_2 \end{pmatrix} \cdots \begin{pmatrix} \boldsymbol{E}_{m-1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{H}_m \end{pmatrix} \boldsymbol{R},$$

wobei $\boldsymbol{R}$ nach Konstruktion eine obere Dreiecksmatrix ist. Das Produkt der erweiterten Householder-Matrizen ist ein Produkt von orthogonalen Matrizen, also wiederum orthogonal. Damit ist die QR-Zerlegung von $\boldsymbol{A}$ gegeben als

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}, \quad \text{wobei} \quad \boldsymbol{Q} := \boldsymbol{H}_1 \begin{pmatrix} 1 & \boldsymbol{o}_n^T \\ \boldsymbol{o}_n & \boldsymbol{H}_2 \end{pmatrix} \cdots \begin{pmatrix} \boldsymbol{E}_{m-1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{H}_m \end{pmatrix}. \qquad \text{(6 Punkte)}$$

Selbstverständlich multipliziert man $\boldsymbol{Q}$ nie aus, sondern verwendet die Darstellung als Produkt von Householder-Matrizen, wobei auf die Reihenfolge zu achten ist, da Matrixmultiplikation bekanntlich nicht kommutativ ist. Die Householder-Matrizen werden auch nur in Form einer Skalierung und eines Vektors gespeichert und angewendet, was sowohl Speicherplatz als auch Arbeitsaufwand einspart.

Aufgabe 10: (6+3 Punkte)
Der zentrale Differenzenquotient

$$D_2 f(x;h) := \frac{f(x+h) - f(x-h)}{2h}$$

zur Approximation der ersten Ableitung an der Stelle x hat die Fehlerordnung 2. Beschreiben Sie in eigenen Worten, welches Verhalten Sie auf einem Rechner mit einer Maschinengenauigkeit von ca. 10^{-16} in der Genauigkeitsfunktion

$$G(h) := |f'(x) - D_2 f(x;h)|$$

für zweimal stetig differenzierbare Funktionen f, z. B. $f(x) = \cos(x)$, an einer Auswertungsstelle, z. B. $x = 1$ erwarten. In welcher Größenordnung wählen Sie h, damit die Formel trotz Rundungsfehlern „optimale“ Ergebnisse liefert?

Lösung zu Aufgabe 10: Die Genauigkeit bestimmt sich aus zwei Einflüssen. Wenn h klein gewählt wird, so beschreibt die Fehlerordnung, in welcher Potenz von h der Differenzenquotient gegen $f'(x)$ strebt, durch die Auslöschung ist aber immer ein Anteil gegeben, der für kleine h sich invers linear (siehe Skript), also wie

$$\frac{C_1}{h} 10^{-16}$$

verhält. Insgesamt kann man den Verlauf der Genauigkeit durch eine Funktion der Art

$$G(h) \approx V(h) := \frac{C_1}{h} 10^{-16} + C_2 h^2 + O(10^{-16}), \quad C_1, C_2 > 0$$

beschreiben. (6 Punkte)

Die Wahl des optimalen h orientiert sich an der Approximation V, man berechnet das optimale h als Minimum der gegebenen Approximation, also unter Verwendung der Ableitung

$$V'(h) = -\frac{C_1}{h^2} 10^{-16} + 2C_2 h = 0 \quad \Rightarrow \quad -C_1 10^{-16} + 2C_2 h^3 = 0$$

zu

$$h_{\text{optimal}} = \sqrt[3]{\frac{C_1}{2C_2}} \cdot \sqrt[3]{10^{-16}} = O(10^{-16/3}),$$

ist also bis auf eine Konstante gleich der dritten Wurzel aus der Maschinengenauigkeit.

Aufgabe 11: (5+4+4 Punkte)
Berechnen Sie die Pseudoinversen der Matrizen

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} 0 & 1 \\ 1 & 2 \\ 2 & 0 \end{pmatrix}, \quad \boldsymbol{C} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

Lösung zu Aufgabe 11: Die Pseudoinverse ist die eindeutige lineare Abbildung, die einer gegebenen rechten Seite die Pseudonormallösung eines Ausgleichsproblemes zuordnet. Da im Falle einer Matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ mit $m \geqslant n$ mit vollem (Spalten-)Rang n die Pseudonormallösung die eindeutige Lösung der Normalgleichungen ist, gilt im Falle eines vollen Ranges

$$\boldsymbol{A}^\dagger = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T.$$

Die Matrix $\boldsymbol{A}$ hat nicht vollen Rang, sondern ist eine Rang-Eins-Matrix, welche die folgende (speicherplatzsparende Variante der) SVD hat,

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix} = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \cdot \sqrt{42} \cdot \frac{1}{\sqrt{14}} \begin{pmatrix} 1 & 2 & 3 \end{pmatrix} = \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^T.$$

Damit ist die Pseudoinverse gegeben als

$$\boldsymbol{A}^\dagger = \boldsymbol{V}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{U}_1^T = \frac{1}{42} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} = \frac{1}{42} \begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 3 & 3 & 3 \end{pmatrix} = \frac{1}{42}\boldsymbol{A}^T. \quad \text{(5 Punkte)}$$

Die Matrizen $\boldsymbol{B}$ und $\boldsymbol{C}$ haben vollen Spaltenrang, also können wir die obige Formel zur Berechnung der Pseudoinversen heranziehen, es gilt

$$\begin{aligned} \boldsymbol{B}^\dagger = (\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T &= \begin{pmatrix} 5 & 2 \\ 2 & 5 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 1 & 2 \\ 1 & 2 & 0 \end{pmatrix} \\ &= \frac{1}{21} \begin{pmatrix} 5 & -2 \\ -2 & 5 \end{pmatrix} \begin{pmatrix} 0 & 1 & 2 \\ 1 & 2 & 0 \end{pmatrix} \\ &= \frac{1}{21} \begin{pmatrix} -2 & 1 & 10 \\ 5 & 8 & -4 \end{pmatrix} \end{aligned} \quad \text{(4 Punkte)}$$

und

$$\boldsymbol{C}^\dagger = (\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T = \frac{1}{4}\begin{pmatrix}1 & 1 & 1 & 1\end{pmatrix}. \qquad \text{(4 Punkte)}$$

Aufgabe 12: (4+4 Punkte)
Berechnen Sie die Singulärwertzerlegungen (SVD) der folgenden Matrizen:

$$\boldsymbol{Q} = \begin{pmatrix}1 & 2 & 2\\ 2 & -2 & 1\\ 2 & 1 & -2\end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix}1\\2\\2\end{pmatrix}.$$

Lösung zu Aufgabe 12: Die erste Matrix ist eine Vielfache einer orthogonalen Matrix, alle Zeilen und Spalten sind paarweise orthogonal und die Norm ist jeweils 3. Damit ist z. B.

$$\boldsymbol{Q} = \left(\frac{1}{3}\boldsymbol{Q}\right)\cdot(3\boldsymbol{E})\cdot(\boldsymbol{E}) = \boldsymbol{U}_Q\boldsymbol{\Sigma}_Q\boldsymbol{V}_Q^T,$$

wobei $\boldsymbol{E}$ die 3×3-Einheitsmatrix bezeichnet, eine (nicht eindeutig bestimmte) SVD, z. B. wäre auch

$$\boldsymbol{Q} = (\boldsymbol{E})\cdot(3\boldsymbol{E})\cdot\left(\frac{1}{3}\boldsymbol{Q}\right) = \tilde{\boldsymbol{U}}_Q\tilde{\boldsymbol{\Sigma}}_Q\tilde{\boldsymbol{V}}_Q^T$$

eine SVD. (4 Punkte)

Da $\boldsymbol{y}$ die erste Spalte von $\boldsymbol{Q}$ ist, gilt

$$\boldsymbol{y} = \left(\frac{1}{3}\boldsymbol{Q}\right)\cdot\begin{pmatrix}3\\0\\0\end{pmatrix}\cdot 1 = \boldsymbol{U}_y\boldsymbol{\Sigma}_y\boldsymbol{V}_y^T$$

und dieses ist eine (vollständige) SVD. (4 Punkte)

Man hätte auch ohne Kenntnis der SVD von $\boldsymbol{Q}$ die ökonomische Variante

$$\boldsymbol{y} = \left(\frac{1}{3}\boldsymbol{y}\right)\cdot 3\cdot 1 = \boldsymbol{U}_y^{(0)}\boldsymbol{\Sigma}_y^{(0)}(\boldsymbol{V}_y^{(0)})^T$$

der SVD von $\boldsymbol{y}$ angeben können.