Institut für Angewandte und Numerische Mathematik
TU Wien

WS 2002/03

Numerische Mathematik für Informatiker

Prüfung am 28. Januar 2003

Name	Vorname	Kennzahl / Matrikelnummer

Beispiel 1	Beispiel 2	Beispiel 3	Gesamt
a)	a)	a)	
b)	b)	b)	
c)	c)		
	d)		
	e)		
	f)		

Der gesamte Rechengang ist auf den beiliegenden Blättern zu dokumentieren.

Zusätzlich beigefügte Zetteln werden bei der Korrektur nicht berücksichtigt.

1) (10 Punkte)

Der Ausdruck

$$\varphi(x) = \frac{e^x - 1}{x^2}$$

sei für eine gegebene Maschinenzahl $x > 0$ in Gleitpunktarithmetik (mit relativer Genauigkeit *eps*) auszuwerten. Dabei sei angenommen, dass der Exponentialterm e^x mit Hilfe einer vordefinierten Prozedur `exp` "sauber" ausgewertet wird, d.h. mit einem relativen Auswertefehler $\leq eps$.

a) Geben Sie eine möglichst scharfe Schranke für den sich dabei ergebenden relativen Auswertefehler an, d.h. bestimmen Sie ε so, dass[1] $\tilde{\varphi}(x) = \varphi(x)(1+\varepsilon)$, $|\varepsilon| \leq ???\ eps$.

Antwort zu a):

[1] $\tilde{\varphi}$ symbolisiert die Auswertung von φ in Gleitpunktarithmetik.

b) Für welche Werte von x ergibt sich hier ein großer relativer Fehler? Kommentieren Sie Ihr Resultat ausführlich!

Antwort zu b):

c) Betrachten Sie nun die unter b) diagnostizierte numerisch instabile Situation. Modifizieren Sie für diesen Fall den Ausdruck $\varphi(x)$ durch Linearisierung[2] des Terms e^x und schreiben Sie die sich ergebende Approximation für $\varphi(x)$ so an, dass Sie in numerisch stabiler Weise ausgewertet werden kann. (Mit ausführlicher Begründung: worin besteht der entscheidende Unterschied zur direkten Auswertung von $\varphi(x)$?)

Antwort zu c):

[2] mittels nach dem linearen Glied abgebrochener Taylor-Entwicklung von e^x

2) (10 Punkte)

a) Wie löst man mehrere lineare Systeme

$$Ax_1 = b_1, \quad Ax_2 = b_2, \quad \ldots \quad Ax_k = b_k$$

mit möglichst geringem Aufwand?
(Anmerkung: Die Operatoren **factorize** für eine LU-Zerlegung und **solve** für eine Rücksubstitution sollen verwendet werden.)

Antwort zu 3a):

b) Wecher Aufwand ist erforderlich, um ein lineares Gleichungssystem $Tx = b$ mit einer $n \times n$ Tridiagonalmatrix effizient zu lösen?

Anzahl der arithmetischen Operationen $O(n^p)$:

p =

c) Welchen Aufwand kann man reduzieren, wenn man weiß, daß die Matrix A des linearen Systems $Ax = b$ symmetrisch und positiv definit ist?

(i)

(ii)

(iii)

d) Was kann man mit Hilfe der Konditionszahl $\kappa(A) = \|A\| \, \|A^{-1}\|$ abschätzen?

$\ldots\ldots\ldots\ldots \leq \kappa(A) \cdot \ldots\ldots\ldots\ldots$

Verbale Beschreibung:

e) Wann ist ein lineares Gleichungssystem numerisch singulär?

Antwort zu 3e):

f) Kann man durch Einsatz geeigneter Pivot-Strategien numerisch singuläre Systeme mit zufriedenstellender Genauigkeit lösen?

- ○ ja
- ○ nein

Begründung:

3) (10 Punkte)

a) Man bestimme die Koeffizienten a und b in der Funktion

$$y(t) = a \cdot t + b \cdot e^{-t} \tag{1}$$

so, dass sie dem folgenden Datensatz möglichst gut im Sinn der kleinsten quadratischen Ordinatenabweichung angepasst sind:

t_i	0	1	2	3
y_i	−2.10	1.30	7.70	17.90

(Rechnung in Taschenrechnergenauigkeit.)

$a =$

$b =$

b) Machen Sie eine Skizze der soeben errechneten Kurve und vergleichen Sie ihren Verlauf für wachsendes t mit dem Verhalten der Datenpunkte. Geben Sie einen zu (1) alternativen Ansatz, der Ihrer Meinung nach das Verhalten der Daten besser wiedergibt.

Alternativer Ansatz:

(0 Punkte)

Bei Umrechnung von Währungen in Bezug auf den Euro gelten folgende Regeln:[1]

- Um Pfund auf Euro zu konvertieren wird der Pfund-Wert durch den Kurs dividiert.
- Um Euro auf Pfund umzurechnen wird der Euro-Wert mit dem Kurs multipliziert.
- Bei jeder Umrechnung ist optimale Rundung auf den nächsten Cent (1/100 Euro) oder Penny (1/100 Pfund) anzuwenden. (*Round away from zero* im Fall des Mittelpunktes zwischen zwei Werten.)

Den folgenden Rechnungen soll ein Kurs von 1 Euro = 0.654 325 Pfund zugrundegelegt werden.

a) Man konvertiere:

Euro	Pfund
	500
	1000
	7000
1000	

b) Man berechne:

	Pfund	Euro
Item 1	1000	
Item 2	1000	
Item 3	1000	
Item 4	1000	
Item 5	1000	
Item 6	1000	
Item 7	1000	
Summe		

Die Differenz zur Umrechnung von 7000 Pfund in Euro nennt man „vertikalen" Rundungsfehler.

Vertikaler Rundungsfehler:	Euro

[1] Council Regulation (EC) No. 1103/97, Articles 4.5.

Institut für Angewandte und Numerische Mathematik
TU Wien

WS 2001/0.

Numerische Mathematik für Informatiker

Prüfung am 29. Jänner 2002

Name	Vorname	Kennzahl / Matrikelnummer

Beispiel 1	Beispiel 2	Beispiel 3	Gesamt
a)	a)	a)	
b)	b)	b)	
c)	c)	c)	
d)			
e)			

Der gesamte Rechengang ist auf den beiliegenden Blättern zu dokumentieren.

Zusätzlich beigefügte Zetteln werden bei der Korrektur nicht berücksichtigt.

e) Man konvertiere die folgenden Bereichsgrenzen:

Pfund	≤ 1000	1000.01 – 4000	4000.01 – 8000	≥ 8000.01
Euro	$\leq$	–	–	$\geq$

Welches Problem tritt hier auf?

Antwort:

c) Man berechne (nach obigen Regeln) die folgenden Konversionen Euro → Pfund — Euro:

Euro	Pfund	Euro	Fehler
1.00			
1.01			
1.02			
1.03			
1.04			
1.05			
1.06			
1.07			
1.08			

Die bei der Konversion Euro → Pfund → Euro auftretenden Fehler nennt man „horizontale" Rundungsfehler.

d) Wieso tritt bei der Konversion Pfund → Euro → Pfund *kein* horizontaler Rundungsfehler auf?

Antwort:

b) Lösen Sie die Aufgabe für folgenden Datensatz:

i	x_i	y_i	φ_i
1	0.0	0.0	0.0
2	1.1	1.1	9.9
3	1.1	2.1	14.9
4	2.1	1.1	13.9

Antwort zu b):

a =

b =

c =

c) Kommentieren Sie den Spezialfall, bei dem die (x_i, y_i) auf einer gemeinsamen Geraden liegen.

Antwort zu c):

2) (10 Punkte)

Bei linearen Ausgleichsproblemen, bei denen eine „Ausgleichsebene" $f(x, y) = ax + by + c$ gesucht ist, sind die Parameter $a, b, c \in \mathbb{R}$ so zu bestimmen, dass

$$\sum_{i=1}^{n} [f(x_i, y_i) - \varphi_i]^2 \rightarrow \textit{minimal} \qquad \text{(LS}$$

für einen gegebenen Satz von n Datenpunkten $\{(x_i, y_i; \varphi_i),\ i = 1, 2, \ldots, n\}$ gilt.

a) Wie groß muss die Anzahl n der zur Verfügung stehenden Datenpunkte mindestens sein, damit das Problem (LS) eine eindeutige Lösung (a, b, c) besitzen kann? Ist diese Bedingung auch hinreichend? (Begründung!)

Antwort zu a):

$n \geq$

Diese Bedingung ist ○ hinreichend

○ nicht hinreichend

Begründung:

b) Will man den Wert $f(x)$ für ein x mit $x_i < x < x_{i+1}$ berechnen, so bezieht man die Tabellenwerte $(x_i, f_i = f(x_i))$, $(x_{i+1}, f_{i+1} = f(x_{i+1}))$ ein und wertet die folgende Formel mit $h = x_{i+1} - x_i$ aus:

$$f(x) = \frac{x_{i+1} - x}{h} f_i - \frac{x_i - x}{h} f_{i+1}.$$

Berechnen Sie die Schranke für den absoluten Rechenfehlereffekt, der bei der Auswertung des obigen Ausdruckes in einer Gleitpunktarithmetik ensteht. Nehmen Sie dabei an, dass bei den Subtraktionen $x_{i+1} - x$ und $x_i - x$ keine Rechenfehler enstehen und alle involvierten Größen Maschinenzahlen sind. Die Rechnungen werden in der IEC/IEEE-Arithmetik, im einfachen Format mit optimaler Rundung durchgeführt.

Hinweis: Im Intervall $[0,2]$ ist $f(x)$ positiv und monoton fallend.

Antwort zu b):

3) (10 Punkte)

Für die Funktion

$$f(x) = \frac{1}{\cosh(x)} = \frac{2}{e^x + e^{-x}} \qquad x \in [0,2]$$

wird eine Wertetabelle mit der Eigenschaft benötigt, dass bei einer linearen Interpolation zwischen zwei Tabellenwerten für alle x aus dem gegebenen Intervall maximal ein absoluter Gesamtfehler (Datenfehlereffekt + Verfahrensfehler + Rechenfehler) von 10^{-6} auftritt.

a) Wie groß darf der Abstand Δx der Tabellenwerte höchstens sein, damit der Verfahrensfehler auf dem gesamten Intervall 5×10^{-7} nicht übersteigt? Der Abstand sollte eine Zehnerpotenz sein (oder die Form $5 \times$ Zehnerpotenz haben).

Als Vorüberlegung zeigen Sie, dass folgendes gilt:

$$f''(x) = \frac{2\sinh^2 x - \cosh^2 x}{\cosh^3 x},$$

$$|f''(x)| \leq \frac{3}{\cosh x}.$$

Um letztere Ungleichung zu zeigen, benützen Sie die Abschätzung $\sinh^2 x \leq \cosh^2 x$.

Antwort zu a):

$\Delta x \leq$

c) Kann man garantieren, dass die Schranke für den Gesamtfehler bei der obigen Interpolation gleich 10^{-6} ist, falls für die Datenfehler aller Tabellenwerte $|\tilde{f}_i - f_i| \leq 3 \cdot 10^{-7}$ gilt? Wenn nicht, welche Maßnahme(n) würden Sie ergreifen, um die Toleranzforderung für den absoluten Gesamtfehler von 10^{-6} zu erfüllen?

Antwort zu c):

...it für Angewandte und Numerische Mathematik
...Wien

SS 2001

Numerische Mathematik für Informatiker

Prüfung am 26. Juni 2001

...ame	Vorname	Kennzahl / Matrikelnummer

Beispiel 1	Beispiel 2	Beispiel 3	Gesamt
a)	a)	a)	
b)	b)	b)	
c)		c)	
d)			
e)			

Der gesamte Rechengang ist auf den beiliegenden Blättern zu dokumentieren.

Zusätzlich beigefügte Zetteln werden bei der Korrektur nicht berücksichtigt.

1) (10 Punkte)

Das Gleitpunktzahlensystem der Rechnerfamilie IBM System/390 ist hexadezimal: $\mathbb{F} = \mathbb{F}(16, 6, -64, 63)$ (einfache Genauigkeit, 4 Byte Format). Gleitpunktzahlen $x = \pm m \cdot 16^e \in \mathbb{F}$ sind codiert als

`CCMMMMMM`,

wobei die Hexadezimal-Ziffern[1] `C`, `M` $\in \{$`0`, ..., `9`, `A`, ..., `F`$\}$ wie folgt zu interpretieren sind:

- Im führenden Bit des ersten Bytes `CC` ('Charakteristik') ist das Vorzeichen von x codiert gemäß
$$0 = +, \quad 1 = -.$$
- Die restlichen 7 Bits der Charakteristik `CC` entsprechen einer Binärzahl c zwischen 0 und $127 = 7F_{\text{HEX}}$, die den Exponenten e der Zahl x gemäß
$$e := c - 64 = c - 40_{\text{HEX}}$$
festlegt.
- Die restlichen 3 Bytes bilden die 6-stellige hexadezimale Mantisse m. Wir verwenden hier die Schreibweise mit $m \in [0, 1)$, d.h. m steht für die hexadezimale Festpunktzahl
$$.\texttt{MMMMMM}_{\text{HEX}}.$$

Beispiel: Die Dezimalzahlen $10, -10$ besitzen in diesem Format die Darstellung `41A00000`, `C1A00000`.

Antworten zu folgenden Fragen bitte in Form der hexadezimalen Codierung `CCMMMMMM`:

a) Wie lauten die kleinste und die größte positive normalisierte Zahl $x_{min}, x_{max} \in \mathbb{F}$?

Antwort:

b) Was passiert bei der Gleitpunkt-Division $1. \oslash x_{min}$?

Antwort:

c) Wie lautet die Gleitpunktzahl *eps* (relative Maschinengenauigkeit) bezüglich der Rundung gegen 0 (Rundung durch Abschneiden)?

Antwort:

d) Wie lautet die kleinste positive Maschinenzahl x, für die die Gleitpunkt-Subtraktion $1. \ominus x$ einen Wert < 1 ergibt?

Antwort:

e) Welche Zahl $x \in \mathbb{F}$ liegt der Dezimalzahl -0.1 am nächsten?

Antwort:

[1] mit der naheliegenden 4-Bit-Codierung als Binärzahl: `0` = 0000, `1` = 0001, ..., `F` = 1111

2) (10 Punkte)

a) Gegegeben sei das folgende, sehr speziell strukturierte lineare Gleichungssystem:

$$\begin{pmatrix} a & a & a & a & \dots & a \\ 1 & a & a & a & \dots & a \\ 0 & 1 & a & a & \dots & a \\ 0 & 0 & 1 & a & \dots & a \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \dots & 1 & a \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \tag{1}$$

Man betrachte für diesen einfachen Fall die Gauß-Elimination (ohne Pivotstrategie), die zur Faktorisierung $LUx = b$ führt und gebe die Matrizen L und U an:

$L = \left(\quad \right)$

$U = \left(\quad \right)$

Weiters gebe man x_n als Formelausdruck in a an:

$x_n =$

b) Für den Spezialfall $a = 2$ berechne man U^{-1} und L^{-1}.

$L^{-1} = \left(\quad \right)$

$U^{-1} = \left(\quad \right)$

Neben (1) (im Spezialfall a = 2) betrachte man nun das gestörte System

$$\begin{pmatrix} 2+\epsilon_{11} & 2+\epsilon_{12} & 2+\epsilon_{13} & \cdots & 2+\epsilon_{1n} \\ 1+\epsilon_{21} & 2+\epsilon_{22} & 2+\epsilon_{23} & \cdots & 2+\epsilon_{2n} \\ \epsilon_{31} & 1+\epsilon_{32} & 2+\epsilon_{33} & \cdots & 2+\epsilon_{3n} \\ \epsilon_{41} & \epsilon_{42} & 1+\epsilon_{43} & \cdots & 2+\epsilon_{4n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \epsilon_{ni} & \epsilon_{n2} & \cdots & 1+\epsilon_{n,n-1} & 2+\epsilon_{nn} \end{pmatrix} \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \\ \tilde{x}_4 \\ \vdots \\ \tilde{x}_n \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

Unter der Annahme $|\epsilon_{ij}| \leq 10^{-8}$ schätze man $\frac{\|\tilde{x}-x\|_\infty}{\|\tilde{x}\|_\infty}$ ab.

$$\frac{\|\tilde{x} - x\|_\infty}{\|\tilde{x}\|_\infty} \leq$$

Hinweis: Für das Produkt $M \cdot N$ zweier Matrizen M und N gilt

$$\|M \cdot N\| \leq \|M\| \cdot \|N\|.$$

3) (10 Punkte)

Die CO-Monatsmittelwerte der Luft in Wien sollen durch eine Funktion der Form

$$\mathrm{CO}(t) = a_0 + a_1 \sin\frac{2\pi t}{12} + a_2 \cos\frac{2\pi t}{12}, \qquad t \text{ in Monaten}$$

nach der Methode der kleinsten Quadrate approximiert werden. Folgende Daten (aus dem Jahr 1993) bilden hierfür die Grundlage:

Monat	1	2	3	4	5	6	7	8	9	10	11	12
CO $[mg/m^3]$	1.3	1.4	1.0	0.8	0.7	0.6	0.5	0.6	0.7	1.1	1.2	1.3

a) Man stelle die Normalgleichungen auf.

b) Man bereche a_0, a_1 und a_2.

c) Man skizziere den Funktionsverlauf und trage die Datenpunkte ein. Was läßt sich über die Approximationsqualität sagen?

Institut für Angewandte und Numerische Mathematik
TU Wien

WS 2000/2001

Numerische Mathematik für Informatiker

Prüfung am 30. Januar 2001

Name	Vorname	Kennzahl / Matrikelnummer

Beispiel 1	Beispiel 2	Beispiel 3	Gesamt
a)		a)	
b)		b)	
c)		c)	
d)		d)	

Der gesamte Rechengang ist auf den beiliegenden Blättern zu dokumentieren.

Zusätzlich beigefügte Zetteln werden bei der Korrektur nicht berücksichtigt.

1) (10 Punkte)

Es sei $\mathbb{F} = \mathbb{F}(10, 9, -98, 100)$ (eine typische Taschenrechner-Arithmetik mit 9 signifikanten Dezimalstellen), mit optimaler Rundung. p bezeichne denjenigen Wert der durch Rundung von $\pi = 3.141592653\ldots$ nach $\mathbb{F}$ entsteht.

Im Folgenden bedeutet das Symbol '$*$' Gleitpunkt-Multiplikation in $\mathbb{F}$. Alle Antworten sind genau zu begründen.

a) Wie lautet der größte Wert $K_1 \in \mathbb{N}$ mit der Eigenschaft dass der Wert $k * p \in \mathbb{F}$ für alle $k \leq K_1$ $(k \in \mathbb{N})$ mit dem exakten Wert $k \cdot p$ übereinstimmt?

Antwort zu a):

b) Wie lautet der größte Wert $K_2 \in \mathbb{N}$ so dass

$$k * p \in I(k) := [k \cdot \pi - 10^{-5}, k \cdot \pi + 10^{-5}]$$

für alle $k \leq K_2$ $(k \in \mathbb{N})$ garantiert zutrifft?

Antwort zu b):

c) Bestimmen Sie die maximale relative Konditionszahl für die beiden Funktionen

$$\text{(i) } f(x) = \sin(x) \qquad \text{und} \qquad \text{(ii) } g(x) = \cos(x)$$

für $x \in I(k)$ (das Intervall $I(k)$ ist wie unter b) definiert). Antwort erbeten als Ausdruck in k (für beliebiges $k \in \mathbb{N}_0$).
(Tipp des Tages: Man verwende die Approximation $|\tan\delta| \approx |\delta|$ für kleine Werte $|\delta|$.)

Antwort zu c):

d) Kommentieren Sie Ihr unter c) erhaltenes Ergebnis.
Für beide Funktionen stellt $k = 0$ einen Sonderfall dar. In welchem Sinn jeweils?

Antwort zu d):

2) (10 Punkte)

In dem Gleichungssystem $Ax = b$ hat die Matrix $A \in \mathbb{R}^{n\times n}$ die Gestalt

$$A = \begin{pmatrix} 1+\delta_{11} & \delta_{12} & \delta_{13} & \dots & \delta_{1,n-1} & 1+\delta_{1,n} \\ \delta_{21} & 1+\delta_{22} & \delta_{23} & \dots & \delta_{2,n-1} & 1+\delta_{2,n} \\ \delta_{31} & \delta_{32} & 1+\delta_{33} & \dots & \delta_{3,n-1} & 1+\delta_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \delta_{n-1,1} & \delta_{n-1,2} & \delta_{n-1,3} & \dots & 1+\delta_{n-1,n-1} & 1+\delta_{n-1,n} \\ 1+\delta_{n,1} & 1+\delta_{n,2} & 1+\delta_{n,3} & \dots & 1+\delta_{n,n-1} & n+\delta_{n,n} \end{pmatrix}$$

mit $|\delta_{ij}| \le 10^{-8}$. Es sei $\tilde{x}$ die Lösung des vereinfachten Systems $\tilde{A}\tilde{x} = b$ mit

$$\tilde{A} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 1 \\ 0 & 1 & 0 & \dots & 0 & 1 \\ 0 & 0 & 1 & \dots & 0 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 1 \\ 1 & 1 & 1 & \dots & 1 & n \end{pmatrix}.$$

Man schätze $\frac{\|\tilde{x}-x\|}{\|x\|}$ bezüglich der Maximumsnorm $\|\cdot\|_\infty$ möglichst scharf ab!

$\frac{\|\tilde{x}-x\|}{\|x\|} \le$

3) (10 Punkte)

Für die Funktion

$$f(x) = \sinh(x) = \frac{e^x - e^{-x}}{2}, \quad x \in [0, 2]$$

wird eine Wertetabelle benötigt, mit der Eigenschaft, daß bei einer linearen Interpolation zwischen zwei Tabellenwerten für alle x aus dem gegebenen Intervall ungünstigstenfalls ein absoluter Fehler von von 10^{-6} auftritt.

a) Wie groß darf der Abstand der Tabellenwerte höhstens sein, damit der Verfahrensfehler auf dem gesamten Intervall $3 * 10^{-7}$ nicht übersteigt? Der Abstand sollte eine Zehnerpotenz sein oder die Form 5*Zehnertpotenz haben.

Antwort zu a):

b) Wie groß darf die Datenungenauigkeit (in den Funktionswerten) sein, damit der dadurch enstehende Datenfehlereffekt kleiner als 10^{-7} ist?

Antwort zu b):

c) Will man den Wert $f(x)$ für ein x mit $x_i < x < x_{i+1}$ berechnen, so bezieht man die Tabellenwerte $(x_i, f_i = f(x_i))$, $(x_{i+1}, f_{i+1} = f(x_{i+1}))$ ein und wertet die folgende Formel mit $h = x_{i+1} - x_i$ aus:

$$f(x) = \frac{x_{i+1} - x}{h} f_i - \frac{x_i - x}{h} f_{i+1}.$$

Berechnen Sie die Schranke für den absoluten Rechenfehlereffekt, der bei der Auswertung des obigen Ausdruckes in einer Gleitpunktarithmetik ensteht. Nehmen Sie dabei an, daß bei den Subtraktionen $x_{i+1} - x$ und $x_i - x$ keine Rechenfehler enstehen und alle involvierten Größen Maschinenzahlen sind. Die Rechnungen werden in der IEC/IEEE-Arithmetik, im einfachen Format mit optimaler Rundung durchgeführt.
Hinweis: Im Intervall [0,2] ist $f(x)$ positiv und monoton wachsend.

Antwort zu c):

d) Kann man garantieren, daß die Schranke für den Gesamtfehler bei der obigen Interpolation kleiner als 10^{-6} ist? Wenn nicht, welche Maßnahme(n) würden Sie ergreifen, um die Toleranzforderung für den absoluten Gesamtfehler von 10^{-6} zu erfüllen?

Antwort zu d):

Institut für Angewandte und Numerische Mathematik
TU Wien

WS 1998/99

Numerische Mathematik für Informatiker

Prüfung am 1. Dezember 1998

Name	Vorname	Kennzahl / Matrikelnummer

Beispiel 1	Beispiel 2	Beispiel 3	Gesamt
a)	a)	a)	
b)	b)	b)	
c)		c)	
		d)	

Der gesamte Rechengang ist auf den beiliegenden Blättern zu dokumentieren.

Zusätzlich beigefügte Zetteln werden bei der Korrektur nicht berücksichtigt.

1) (10 Punkte)

Die Nullstellen x_1 und x_2 der quadratischen Gleichung $x^2 + bx + c = 0$ seien bekannt. Jemand möchte daraus die Koeffizienten $b = -(x_1 + x_2)$ und $c = x_1 \cdot x_2$ berechnen und anschließend die Größen $d = c/b$ und $e = b/c$. Bei allen Überlegungen dieses Beispiels werde angenommen, daß x_1 und x_2 reelle Zahlen sind, für welche $\frac{1}{10} \leq |x_i| \leq 1$ $(i = 1, 2)$ gilt und daß die Berechnung von b, c, d und e auf einem Taschenrechner mit der Arithmetik $\mathbb{M}(10, 10, +99, -99)$ durchgeführt wird (sodaß sowohl bei der Eingabe von x_1 und x_2 als auch bei der weiteren Rechnung entsprechende Rundungsfehler auftreten). Weiters nehme man an, daß nach jedem Rechenschritt abgeschnitten wird und daß keine Schutzstellen vorhanden sind.

a) Man gebe ein Nullstellenpaar x_1 und x_2 an, sodaß sich bei der Berechnung von b nur vier richtige Stellen ergeben.

b) Nun betrachte man die absoluten Rundungsfehler, die bei der Berechnung von b, c, d und e entstehen. Konkret untersuche man, für welche der Größen b, c, d oder e es Nullstellenpaare x_1 und x_2 gibt, sodaß das absolute Fehlerniveau

(i) in der Größenordnung von 10^{-8} liegt

(ii) in der Größenordnung von 10^{0} liegt

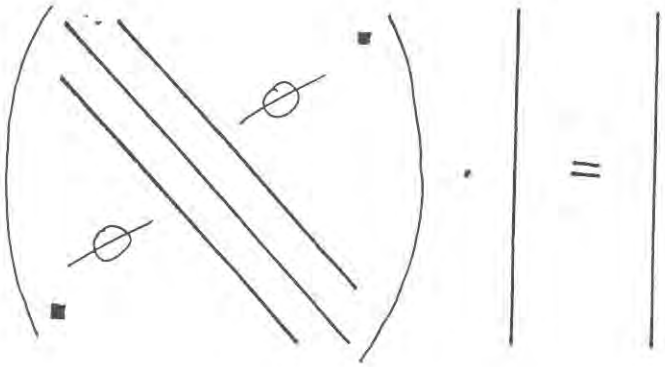
(iii) unendlich groß ist.

(Kurze Begründung)

c) Man gebe (mittels '$(1+\varepsilon)$-Technik') eine möglichst scharfe Rundungsfehlerschranke für den absoluten Rundungsfehler bei der Berechnung von e an, die für sämtliche Nullstellenpaare x_1, x_2 mit $\frac{1}{10} \leq |x_i| \leq 1$ gilt.

2) (10 Punkte)

In der graphischen Datenverarbeitung spielen Spline-Funktionen eine wichtige Rolle bei der Darstellung punktförmig gegebener Daten. Bei der Berechnung einer periodischen Spline-Interpolierenden treten lineare Gleichungssysteme der folgenden Bauart auf:



a) Welche Bauart hat die obere Dreiecksmatrix U bei der Dreieckszerlegung? Was folgt daraus für deren Speicherung?

b) Man formuliere einen Algorithmus, der die Lösung des obigen Gleichungssystems – ohne Pivotisierung – unter Berücksichtigung der speziellen Bauart der Matrix liefert.

3) (10 Punkte)

Um den Fehler bei der doppelt genauen Gleitpunktdivision in der Steinzeitversion des Pentium-Prozessors zu umgehen, faßt Fred folgende softwaremäßige Lösung ins Auge:

Um eine vorgegebene doppelt genaue Gleitpunktzahl x zu invertieren, wählt Fred einen geeigneten Näherungswert y_0 für $y := 1/x$ (also z.B. den durch einfach genaue Rechnung erhaltenen Wert) und will ausgehend von y_0 die zu $y = 1/x$ äquivalente Gleichung $f(y) = xy - 1 = 0$ *iterativ* in doppelt genauer Arithmetik lösen. Er meint, daß sich leicht eine Iterationsvorschrift finden lassen müßte, die sehr rasch konvergiert aber keinerlei explizite Divisionen erfordert.

a) Fred denkt an das *Newton - Verfahren* für $f(y) = 0$. Kommentieren Sie diese Idee.

b) Wilma denkt über Freds Vorschlag nach und weiß es besser: Sie meint, es gibt ja auch vereinfachte Varianten des Newton-Verfahrens, bei denen man die in der Iterationsvorschrift auftretende Ableitung nicht exakt bildet sondern dafür einen geeigneten (von y_0 abhängigen) Näherungswert verwendet.

Zeigen Sie, in welcher Weise Wilmas Idee zu einer Iterationsvorschrift führt, die tatsächlich divisionsfrei abläuft und eine sehr gute Konvergenz gegen $y = 1/x$ erwarten läßt (Begründung!).

Schreiben Sie die so entstehende Iteration so an, daß sie (nach einer Initialisierungsphase) in jedem Iterationsschritt nur eine Addition und eine Multiplikation erfordert.

c) Geben Sie für die unter b) eingeführte Iteration eine Abschätzung für die Konvergenzgeschwindigkeit an, d.h. bestimmen Sie ρ so, daß[1]

$$|y_{i+1} - y| \leq \rho\,|y_i - y|$$

gilt. Wodurch ist die Konvergenzgeschwindigkeit, d.h. die Kleinheit von ρ bestimmt?

d) Approximieren Sie den Wert $\varphi(x) := 1/x$ mittels Taylor-Entwicklung von φ bis zum linearen Term (Entwicklung um $x_0 := 1/y_0$). Wie hängt die so entstehende Näherung für $1/x$ mit obiger Iteration zusammen?

[1] Hier bezeichnet $y = 1/x$ das exakte Ergebnis; Rundungsfehlereffekte bei der Iteration bleiben außer Betracht.

Numerische Mathematik f. Informatiker

Prüfung am 6. Mai 1998

(1.) (16) Ausdruck $x^2 + 3x + 2$

2 Auswertungsvarianten

result 1 = $(2 + 3x) + x \cdot x$

result 2 = $2 + (x \cdot (3 + x))$

(Achtung! Beispiel war ungefähr so - aus dem Gedächtnis!)

a.) Für result 2 (Maschinengenauigkeit eps) soll der relative Rundungsfehler berechnet werden. Begründung für den Fehler!

b.) Ist result 1 besser? Ja/Nein? Warum! → nur Begründung keine Rechnung

c.) Gibt es eine bessere Variante → dafür Rundungsfehler berechnen.

(2.) (20) Matrizenbeispiel Nr. 2 von der VO - Prüfung am 27. 1. 97

(3.) (14) Polygonzug - Fehlerberechnung - Beispiel Nr. 3 von —"—

Institut für Angewandte und Numerische Mathematik
TU Wien

SS 1998

Numerische Mathematik für Informatiker

Prüfung am 12. März 1998

Name	Vorname	Kennzahl / Matrikelnummer

Beispiel 1	Beispiel 2	Beispiel 3	Gesamt
a)	a)	a)	
b)	b)	b)	
c)	c)	c)	

Der gesamte Rechengang ist auf den beiliegenden Blättern zu dokumentieren.

Zusätzlich beigefügte Zetteln werden bei der Korrektur nicht berücksichtigt.

1) (10 Punkte)

Der Ausdruck

$$\varphi(x) = \frac{\sqrt{x+1} - \sqrt{x}}{10} = \frac{1}{10(\sqrt{x+1} + \sqrt{x})}, \quad x > 1,$$

soll für große Werte von x berechnet werden.

a) Man gebe die relative Konditionszahl $K_{\varphi(x)\leftarrow x}$ als Funktion von x an.

Man gebe für $K_{\varphi(x)\leftarrow x}$ eine möglichst gute, von x unabhängige Schranke an, d.h. man bestimme ein möglichst kleines $C > 0$, sodaß

$$|K_{\varphi(x)\leftarrow x}| \leq C, \ \forall x > 1$$

gilt. Welche Aussage kann man demgemäß für die Kondition des Problems machen?

b) Man gebe für beide Auswertungsvarianten möglichst gute Fehlerschranken für den relativen Rundungsfehler als Ausdrücke in x und in der Maschinengenauigkeit *eps* an, d.h. man bestimme für beide Fälle ein ε, mit[1]

$$\mathrm{rnd}(\varphi(x)) = \varphi(x)(1+\varepsilon)$$

und schätze es in Abhängigkeit von x und *eps* ab.

Hinweis: Der Einfachheit halber nehme man an, daß für die Auswertung der Wurzelfunktionen

$$\mathrm{rnd}(\sqrt{1+x}) = \sqrt{1+x}\,(1+\varepsilon_1), \quad \mathrm{rnd}(\sqrt{x}) = \sqrt{x}\,(1+\varepsilon_2)$$

mit $|\varepsilon_1| \leq 2eps$ und $|\varepsilon_2| \leq eps$ gilt und vernachlässige den Rundungsfehler bei der Division durch 10.

c) Welche der beiden Auswertungsvarianten ist für große Werte von x vorzuziehen? Ausführliche Begründung!

[1] $\mathrm{rnd}(\varphi(x))$ symbolisiert die jeweilige Auswertung von φ in Gleitpunktarithmetik.

Tragen Sie die Ergebnisse ein und kreuzen Sie die zutreffenden Aussagen an !

$K_{\varphi(x)\leftarrow x} =$

Für $x > 1$ ist $C =$

und deshalb ist die Auswertung [gut] [schlecht] konditioniert.

Fehlerschranke für den relativen Rundungsfehler von $\varphi_1(x) := \frac{\sqrt{x+1}-\sqrt{x}}{10}$:

$|\varepsilon| \leq$

Fehlerschranke für den relativen Rundungsfehler von $\varphi_2(x) := \frac{1}{10(\sqrt{x+1}+\sqrt{x})}$:

$|\varepsilon| \leq$

Die Formel [$\varphi_1(x)$] [$\varphi_2(x)$] ist bei der Auswertung von $\varphi(x)$
für große x-Werte vorzuziehen.

Begründung:

2) (10 Punkte)

Gegeben sei ein 4-dimensionales lineares Gleichungssystem $Ax = b$. Die konkrete Gestalt von A und b tut hier nichts zur Sache; es sei nur bekannt, daß gilt

$$\|A\|_2 = \frac{77}{60}, \qquad \|A^{-1}\|_2 = 63\,420.$$

a) Wie groß ist die relative Konditionszahl von A bezüglich der Norm $\|\cdot\|_2$?

Antwort zu a): 81589

b) Jemand verwendet ein Iterationsverfahren (z.B. Jacobi-Verfahren) und erhält nach einigen Schritten eine Näherungslösung $\tilde{x}$. Diese sieht vernünftig aus, denn ihr Residuum $A\tilde{x} - b$ ist klein, und zwar (auf zwei Stellen genau angegeben:)

$$A\tilde{x} - b = (3.0\text{E}{-4}, 3.0\text{E}{-4}, 5.0\text{E}{-5}, 2.0\text{E}{-4})^T.$$

Dieser Jemand gibt sich mit der Näherungslösung $\tilde{x}$ zufrieden, erkennt aber später, daß sie extrem ungenau ist: In keinem der $\tilde{x}_i$ ist auch nur eine Dezimalstelle richtig.
Erklären Sie die Ursache für diesen Effekt mittels ausführlicher schriftlicher Begründung (ohne Rechnung). Warum ist hier der Fehler noch so groß, obwohl das Residuum schon so klein ist?

Antwort zu b):

$Y = Ax$

$\|A\|$ groß: kleines $\Delta x \Rightarrow$ großes ΔY

$\|A^{-1}\|$ groß: kleines $\Delta Y \Rightarrow$ großes Δx

$Ax = b$ $\qquad (3.0\text{E}{-4}, \dots) = \tilde{b} - b$

$A\tilde{x} = \tilde{b}$

$$\frac{\|\tilde{b}-b\|_2}{\|b\|_2} \leq \|A\|_2 \cdot \frac{\|\tilde{x}-x\|_2}{\|x\|_2}$$

$$\frac{\|\tilde{x}-x\|_2}{\|x\|_2} \leq \underbrace{\|A^{-1}\|_2}_{\text{riesig}} \cdot \underbrace{\frac{\|\tilde{b}-b\|_2}{\|b\|_2}}_{\text{winzig}}$$

} relativ

$\|A\|$ groß: Ur $\Rightarrow$ Bildbereich groß

$\|A^{-1}\|$ groß: Bild $\Rightarrow$ Urbild groß (Glg. lösen!)

$Ax = b \Rightarrow \tilde{x} \Rightarrow A\tilde{x} - b = 0$

Residuum

$\|A\|$ groß $\Rightarrow$ Residuum ist (fast) immer groß

c) Geben Sie – unter Zuhilfenahme der über das Problem vorliegenden Information – an, wie groß dieser Fehler maximal gewesen sein kann; d.h. man schätze die Norm $\|\tilde{x} - x^*\|_2$ nach oben ab[2] (Rechnung inklusive Begründung).

Antwort zu c):

aus Angabe b.) ausrechnen (quadrieren, Summe, Wurzel...)

$\|\tilde{x} - x\| \leq \|A^{-1}\|_2 \cdot \|\tilde{b} - b\|_2$

$\|\tilde{x} - x\| \leq 63420$

Begründung:

[2] x^* bezeichnet die exakte Lösung des Gleichungssystems.

Polynome 1. Grades ($d=1$)

S 148

3) (10 Punkte)

Für die graphische Ausgabe von Funktionen sollen diese durch Polygonzüge (stückweise lineare Funktionen) ersetzt werden.

a) Die Funktion $f(x) = \cos(x)$ soll am Intervall $[0, 1.5]$ durch einen Polygonzug p dargestellt werden, der äquidistante (gleichabständige) Knoten besitzt und

$$|p(x) - f(x)| \leq 10^{-3}, \quad x \in [0,1]$$

erfüllen soll. Wie groß ist der Abstand h der Knoten (Datenpunkte) zu wählen?

b) Wie kann man den Abstand h bestimmen, wenn anstelle des absoluten Fehlers der relative Fehler unter einer vorgegebenen Schranke liegen soll?

c) Für den Fall, daß man von der darzustellenden Funktion f keine Information (Ableitungen etc.) besitzt, die sich zur Fehlerabschätzung eignet, sind Vorschläge zu machen, wie man die algorithmische Bestimmung eines geeigneten Knotenabstandes durchführen könnte.

S 149.-

Antwort zu a): $x_k \leq \xi \leq x_{k+1}$

$$e_1(x) = \frac{\cos(\xi)''}{2!} (x - x_k)\cdot(x - x_{k+1})$$

genaue Erklärung!

$$|e_1(x)| \leq \frac{1}{2} \cdot |-\cos\xi| \cdot |x - x_k| \cdot |x - x_{k+1}|$$

$$\leq \frac{1}{2} \cdot 1 \cdot \frac{h}{2} \cdot \frac{h}{2} = \frac{h^2}{8}$$

$$10^{-3} = 0{,}001 = \frac{h^2}{8} \Rightarrow h = \sqrt{0{,}008} = 0{,}0894$$

Antwort zu b): relative Fehler:

$$|(e_1^r(x))| = \left|\frac{e_1(x)}{\cos x}\right| = \frac{1}{2} \cdot \frac{|-\cos\xi|}{|\cos x|} \cdot |x - x_k| \cdot |x - x_{k+1}|$$

bei kleinem h wegen $x_k \leq x, \xi \leq x_{k+1}$: $\cos x \sim \cos\xi$

$$\frac{1}{2} \cdot |x - x_k| \cdot |x - x_{k+1}| = \frac{h^2}{8}$$

→ ändert sich nichts!

⊙ wie Stützstelle nicht als Pixel!

⊙ ξ irgendein Wert im Intervall (ist vorhanden!)
(wird nicht ausgerechnet, nur abgeschätzt!)

Antwort zu c):

Halbe zw. die Stützstelle!

1.) ein h annehme

2.) an die Halbierungspunkte alle Intervalle messe

3.) wenn an 1 Stelle größer 10^{-3}

$\Rightarrow \frac{h}{2}$

→ solange, bis alle kleiner 10^{-3}

Institut für Angewandte und Numerische Mathematik
TU Wien

WS 1997/98

Numerische Mathematik für Informatiker

Prüfung am 14. Oktober 1997

Name	Vorname	Kennzahl / Matrikelnummer

Beispiel 1	Beispiel 2	Beispiel 3	Gesamt
a) b) c)	a) b)	a) b) c)	

Der gesamte Rechengang ist auf den beiliegenden Blättern zu dokumentieren.

Zusätzlich beigefügte Zetteln werden bei der Korrektur nicht berücksichtigt.

1) (10 Punkte)
Der Ausdruck

$$\varphi(x) = 1 - \frac{1}{1+x} = \frac{x}{1+x}, \quad x \in I := (-\infty, -1.5] \cup [-0.5, +\infty)$$

sei für eine gegebene Maschinenzahl x in Gleitpunktarithmetik, mit einer relativen Maschinengenauigkeit *eps*, zu berechnen.

a) Man gebe die relative Konditionszahl $K_{\varphi(x)\leftarrow x}$ als Funktion von x an. Man gebe für $K_{\varphi(x)\leftarrow x}$ eine möglichst gute, von x unabhängige Schranke an, d.h. man bestimme ein möglichst kleines $C > 0$, sodaß

$$|K_{\varphi(x)\leftarrow x}| \leq C \quad \forall x \in I$$

gilt. Welche Aussage kann man demgemäß für die Kondition des Problems machen?

b) Man gebe für beide Auswertungsvarianten möglichst gute Fehlerschranken für den relativen Rundungsfehler als Ausdrücke in x und in der Maschinengenauigkeit *eps* an, d.h. man bestimme für beide Fälle ein ε, mit[1]

$$\tilde{\varphi}(x) \doteq \varphi(x)(1+\varepsilon)$$

und schätze es in Abhängigkeit von x und *eps* ab.

c) Für welchen Bereich der x-Werte ist eine der beiden Auswertungsvarianten der anderen stark vorzuziehen? Ausführliche Begründung!

[1] $\tilde{\varphi}(x)$ symbolisiert die jeweilige Auswertung von φ in Gleitpunktarithmetik

Tragen Sie die Ergebnisse ein und kreuzen Sie die zutreffenden Aussagen an !

$K_{\varphi(x)\leftarrow x} =$

Für $x \in I$ ist $C =$

und deshalb ist die Auswertung [gut] [schlecht] konditioniert.

Fehlerschranke für den relativen Rundungsfehler von $\varphi_1(x) := 1 - \frac{1}{1+x}$:

$|\varepsilon| \leq$

Fehlerschranke für den relativen Rundungsfehler von $\varphi_2(x) := \frac{x}{1+x}$:

$|\varepsilon| \leq$

Die Formel [$\varphi_1(x)$] [$\varphi_2(x)$] ist bei der Auswertung von $\varphi(x)$

für $x \approx$ vorzuziehen.

Begründung:

2) (10 Punkte)

Gegeben sei die Matrix

$$M = \begin{pmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{pmatrix}$$

a) Berechnen Sie die LU-Zerlegung von M (ohne Pivotsuche):

$L =$

$U =$

b) Verwenden Sie diese Zerlegung, um das Gleichungssystem $M^T x = y$ zu lösen, wobei $y = (-12, -9, -3, 14)^T$. (M^T ... die transponierte Matrix.)

$x =$

Dokumentieren und erläutern Sie Ihre Rechnung

3) (16 Punkte)
Die Datenpunkte

i	1	2	3	4	5	6	7	8
x_i	-2	-1	0	1	2	3	4	5
y_i	0	1	2	3	5	6	7	8

sind mit Hilfe der Akima-Methode zu interpolieren.

a) Man fertige eine Skizze der Datenpunkte an.

b) Man überlege (auf Grund der Definition der Akima-Funktion), wie die Interpolationsfunktion für $x \in [0,3]$ aussieht und trage deren Verlauf in die Skizze ein.

c) Man berechne den Wert der Interpolationsfunktion an der Stelle $x = 1.5$.

Institut für Angewandte und Numerische Mathematik
Technische Universität Wien

SS 1997

Numerische Mathematik für Informatiker

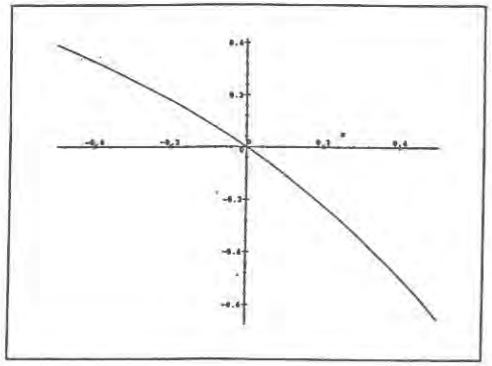
Prüfung am 26. Juni 1997

Name	Vorname	Kennzahl / Matrikelnummer

Beispiel 1	Beispiel 2	Beispiel 3	Gesamt
a) b)	a) b) c) d)	a) b) c)	

Der gesamte Rechengang ist auf den beiliegenden Blättern zu dokumentieren.

Zusätzlich beigefügte Zetteln werden bei der Korrektur nicht berücksichtigt.



1) (16 Punkte)
Die Funktion

$$f(x) := \ln(1 - \sin x)$$

(vgl. obige Skizze) soll in der Nähe der Nullstelle $x = 0$ genau ausgewertet werden. Für $x = 0.1234567890\text{E-}8$ ergibt Auswertung in 10-stelliger dezimaler Gleitpunktarithmetik[1] den Wert

$$x = -0.1199999989\text{E-}8$$

mit nur einer einzigen richtigen Stelle im Vergleich zu dem (auf 10 Stellen) exakten Wert

$$x = -0.1234567891\ldots\text{E-}8.$$

a) Liefern Sie eine ausführliche Erklärung für die Ursache dieser ausgeprägten Ungenauigkeit.

b) Jemand verwendet die Identität $\sin^2 x + \cos^2 x = 1$, um $f(x)$ mittels

$$\ln\left(\frac{\cos^2 x}{1 + \sin x}\right)$$

auszuwerten. Ergibt dies ein genaueres Resultat und wenn ja, warum nicht? (Ausführliche Argumentation!)

[1] Korrekte 10-stellige Implementierung der auftretenden Elementarfunktionen wird vorausgesetzt.

2) (18 Punkte)

Man betrachte die beiden linearen Gleichungssysteme

$$\begin{pmatrix} 1 & & & & \\ 1 & 1 & & \emptyset & \\ 1 & 1 & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ \vdots \\ n \end{pmatrix} \qquad (1)$$

und

$$\begin{pmatrix} 1 & & & & \\ 1 & 1 & & \emptyset & \\ 1 & 1 & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} n \\ n-1 \\ n-2 \\ \vdots \\ 1 \end{pmatrix} \qquad (2)$$

a) Man gebe für den Fall einer allgemeinen rechten Seite, also für das Gleichungssystem

$$\begin{pmatrix} 1 & & & & \\ 1 & 1 & & \emptyset & \\ 1 & 1 & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{pmatrix} \qquad (3)$$

die Lösung an:

$x_1 =$

$x_2 =$

$x_3 =$

$\vdots$

$x_n =$

b) Man gebe die Inverse A^{-1} der Matrix

$$A = \begin{pmatrix} 1 & & & & \\ 1 & 1 & & \emptyset & \\ 1 & 1 & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}$$

an:

$$A^{-1} = \begin{pmatrix} & & \\ & & \\ & & \end{pmatrix}$$

c) Nachfolgend werde als Vektornorm die Maximumsnorm $\|\cdot\|_\infty$ zugrunde gelegt. Man berechne (aufgrund von b)) die relative Konditionszahl $\kappa(A)$:

$\kappa(A) =$

d) Nun betrachte man (1), (2) und (3) mit gestörter rechter Seite b, d.h. im Fall (1) die rechte Seite $\begin{pmatrix}1\\ \vdots \\ n\end{pmatrix} + \begin{pmatrix}\Delta b_1\\ \vdots \\ \Delta b_n\end{pmatrix}$, im Fall (2) die rechte Seite $\begin{pmatrix}n\\ \vdots \\ 1\end{pmatrix} + \begin{pmatrix}\Delta b_1\\ \vdots \\ \Delta b_n\end{pmatrix}$ und im allgemeinen Fall (3) die rechte Seite $\begin{pmatrix}b_1\\ \vdots \\ b_n\end{pmatrix} + \begin{pmatrix}\Delta b_n\\ \vdots \\ \Delta b_n\end{pmatrix}$. Unter der Annahme $\left\|\begin{pmatrix}\Delta b_1\\ \vdots \\ \Delta b_n\end{pmatrix}\right\|_\infty = \|\Delta b\|_\infty \le \Delta$ schätze man aufgrund von der unter Punkt a) gefundenen Lösungsdarstellung möglichst scharf die Größe F in

$$\frac{\|\Delta x\|_\infty}{\|x\|_\infty} \le F \frac{\|\Delta b\|_\infty}{\|b\|_\infty}$$

bes. (1) und (2) ab:

$F =$	für (1)
$F =$	für (2)

und vergleiche jeweils mit $\kappa(A)$:

$F/\kappa(A) =$	für (2)
$F/\kappa(A) =$	für (2)

(10 Punkte)

In einem Programm zur Nullstellenbestimmung skalarer (eindimensionaler) nichtlinearer Gleichungen soll das Konvergenzverhalten der vom Programm erzeugten Folge automatisch untersucht werden.

a) Es soll zwischen konvergentem und nicht-konvergentem Verhalten der Folge von Näherungswerten unterschieden werden. Man mache einen Entwurf für einen Programmabschnitt, der diese Unterscheidung trifft.

b) Im Fall konvergenter Näherungswerte soll zwischen Konvergenzphase und Endphase (wo der Abstand zur Lösung in der Größe der Rechenfehler pro Schritt liegt) unterschieden werden. Man mache einen Entwurf für einen Programmteil, der dies leistet.

c) Für die Konvergenzphase soll die Konvergenzordnung festgestellt werden. Man mache einen Entwurf für einen Teil des Programms, mit dem diese Ermittlung möglich ist.

Anmerkung: Es ist zu beachten, daß die exakte Lösung x^* *nicht* bekannt ist.

Institut für Angewandte und Numerische Mathematik
TU Wien

WS 1996/97

Numerische Mathematik für Informatiker

Prüfung am 28. Januar 1997

Name	Vorname	Kennzahl / Matrikelnummer

Beispiel 1	Beispiel 2	Beispiel 3	Gesamt
a) b) c)	a) b)	a) b)	

1) (16 Punkte)

Zur Auswertung des Ausdrucks

$$a(x) = \frac{1 - \cos x}{x}$$

für x–Werte in der Nähe von Null stehen zwei mathematisch äquivalente Formeln zur Verfügung:

$$f(x) = \frac{1 - \cos x}{x} \quad \text{bzw.} \quad g(x) = \frac{\tan(x/2)\sin x}{x}.$$

a) Man gebe die relative Konditionszahl $K_{a(x)\leftarrow x}$ als Funktion von x an. Man überprüfe die Kondition der Aufgabe für kleine x–Werte, $x \approx 0$. *Hinweis*: Es gilt $\lim_{x\to 0} \frac{x}{\sin x} = 1$.

b) Man gebe für $f(x)$ und $g(x)$ Rundungsfehlerschranken für den relativen Rundungsfehler als Ausdrücke in x und in der Maschinengenauigkeit *eps* an. Man nehme dabei an, daß die relativen Fehler bei der Berechnung von trigonometrischen Funktionen durch *eps* beschränkt sind. *Hinweis*: Ist $y\delta$ klein, so gilt in der ersten Näherung:

$$\tan(y(1+\delta)) = \tan y + \frac{1}{\cos^2 y} y\delta = \tan y \left(1 + \frac{2y}{\sin(2y)}\delta\right).$$

c) Welche der obigen Formeln ist bezüglich der Rundungsfehler bei der Auswertung von $a(x)$ in der Nähe von Null vorzuziehen? Ausführliche Begründung!

Tragen Sie die Ergebnisse ein
und kreuzen Sie die zutreffenden Aussagen an !

$K_{a(x)} =$

Für $x \approx 0$ ist $K_{a(x)\leftarrow x} \approx$

und deshalb ist die Auswertung [gut] [schlecht] konditioniert.

Fehlerschranke für den relativen Rundungsfehler von $f(x)$:

Fehlerschranke für den relativen Rundungsfehler von $g(x)$:

Die Formel $\boxed{f(x)}$ $\boxed{g(x)}$ ist bei der Auswertung von $a(x)$ für $x \approx 0$ vorzuziehen.
Ausführliche Begründung!

2) (20 Punkte)

a) Gegegeben sei das folgende, sehr speziell strukturierte lineare Gleichungssystem:

$$\begin{pmatrix} a & a & a & a & \dots & a \\ 1 & a & a & a & \dots & a \\ 0 & 1 & a & a & \dots & a \\ 0 & 0 & 1 & a & \dots & a \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \dots & 1 & a \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \qquad (1)$$

Man betrachte für diesen einfachen Fall die Gauß-Elimination (ohne Pivotstrategie), die zur Faktorisierung $LUx = b$ führt und gebe die Matrizen L und U an:

$L =$

$U =$

Weiters gebe man x_n als Formelausdruck in a an:

$x_n =$