Institut für Angewandte und Numerische Mathematik
TU Wien

WS 2001/02

Numerische Mathematik für Informatiker

1. Übungstest am 13. November 2001

Name	Vorname	Kennzahl / Matrikelnummer

Beispiel 1	Beispiel 2	Beispiel 3	Gesamt
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) (17 Punkte)

a) Auf Taschenrechnern kann (modellhaft) angenommen werden, dass das Gleitpunktzahlensystem $\mathbb{F}(10, 10, -98, 100, \mathit{false})$ implementiert ist.

* Welches ist die kleinste positive Gleitpunktzahl in diesem Zahlensystem?

* Welchen Abstand haben die Gleitpunktzahlen in der Umgebung von π?

* Man gebe eine möglichst kleine Schranke für den Abstand von π zur nächstgelegenen Gleitpunktzahl an.

* Warum gibt es in diesem Zahlensystem kein implizites erstes Bit?

b) Zahlen aus $\mathbb{R}$ sollen mittels optimaler Rundung (Variante *round to even*) auf Zahlen aus $\mathbb{F}(10, 4, -9, 9, \mathit{true})$ abgebildet werden.

$x = 33.3333$	$\square\ x =$
$x = 55.5555$	$\square\ x =$
$x = 55.555$	$\square\ x =$
$x = 55.565$	$\square\ x =$
$x = (\sin 710_{\text{rad}})^2$	$\square\ x =$
$x = (\sin 710_{\text{rad}})^3$	$\square\ x =$
$x = (\sin 710_{\text{rad}})^4$	$\square\ x =$

2) (16 Punkte)

Jemand soll im Rahmen einer umfangreicheren numerischen Berechnung die Funktion

$$f(x) = \sqrt{\frac{1 - \cos x}{1 + \cos x}}$$

für mehrere, sehr kleine Agumentwerte $x \approx 0$ auswerten und findet in einer mathematischen Formelsammung unter der Überschrift „Trigonometrische Funktionen doppelter und halber Winkel" u.a. folgende mathematiche Äquivalenzumformung:

$$\tan\left(\frac{x}{2}\right) = \sqrt{\frac{1 - \cos x}{1 + \cos x}} = \frac{\sin x}{1 + \cos x} = \frac{1 - \cos x}{\sin x}. \tag{1}$$

Er überlegt daher, welcher dieser Ausdrücke für seine Berechnungen am geeignetesten ist.

a) Um die Kondition (Empfindlichkeit der f-Werte auf Störungen im Argument x) der Funktionauswertung $f(x)$ für $x \approx 0$ zu studieren, bestimmen Sie die relative Konditionszahl (als Funktion von x) $K_{f(x)\leftarrow x}$ dieser Aufgabe.

Antwort zu a):

relative Konditionszahl $K_{f(x)\leftarrow x} =$

b) Um sich eine Vorstellung über die Größenordung dieser Kondtitionszahl für $x \approx 0$ zu machen, bestimmen Sie den Wert von $K_{f(x)\leftarrow x}$ an der Stelle $x = 0$.

Hinweis: de l'Hospital

Antwort zu b):

Der Wert $K_{f(x)\leftarrow x}(0) =$

c) Es sei nun $f(x)$ für Argumente $x \approx 0$ auszuwerten, wobei sich nach (1) vier alternative Auswertungsvarianten ergeben. Von welchen dieser Varianten ist zu erwarten, dass sie für betragskleine $|x| \ll 1$ numerisch stabil sind, d.h. ein Ergebnis in befriedigender relativer Genauigkeit liefern, und für welche trifft das nicht zu? Geben Sie jeweils eine kurze (aber möglichst präzise) Begründung für Ihre Antwort.

Antwort zu b) (zutreffendes ankreuzen):

Variante	num. stabil	num. instabil	Begründung
$\sqrt{\frac{1-\cos x}{1+\cos x}}$	○	○	
$\tan\left(\frac{x}{2}\right)$	○	○	
$\frac{\sin x}{1+\cos x}$	○	○	
$\frac{1-\cos x}{\sin x}$	○	○	

3) (17 Punkte)

Im Anschluss an Beispiel 2 gebe man Schranken ür den relativen Rundungsfehler in Abhängigkeit von x ($|x| \ll 1$) für folgende Ausdrücke an:

$$\tan\left(\frac{x}{2}\right), \quad \frac{\sin x}{1+\cos x}, \quad \frac{1-\cos x}{\sin x}.$$

Dabei lege man die Gleitkommaarithmetik $\mathbb{F}(10.10.-98.100, \mathit{false})$ mit der Rundungsvorschrift *round to nearest.* zugrunde und nehme an, dass bei der Auswertung der trigonometrischen Funktionen maximal ein elementarer Rundungsfehler auftritt.

Antwort zu a):

relative Rundungsfehlerschranke für $\tan\left(\frac{x}{2}\right)$:

relative Rundungsfehlerschranke für $\frac{\sin x}{1+\cos x}$:

relative Rundungsfehlerschranke für $\frac{1-\cos x}{\sin x}$:

(ii) $L = \begin{pmatrix} & \\ & \end{pmatrix}$ $U = \begin{pmatrix} & \\ & \end{pmatrix}$

(iii) $L = \begin{pmatrix} & \\ & \end{pmatrix}$ $U = \begin{pmatrix} & \\ & \end{pmatrix}$

PRÜFUNGSORDNER - ein Service Deiner Fachschaft Informatik!
LVA: UMMATH UE 1 Preis: 18,-

Institut für Angewandte und Numerische Mathematik
TU Wien

WS 2000/01

Numerische Mathematik für Informatiker

1. Übungstest am 14. November 2000

Name	Vorname	Kennzahl / Matrikelnummer

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

Der gesamte Rechengang ist auf den beiliegenden Blättern zu dokumentieren.

Zusätzlich beigefügte Zetteln werden bei der Korrektur nicht berücksichtigt.

1) (17 Punkte)

In diesem Beispiel wird ein Computer (z. B. ein PC) mit doppelt genauer IEC/IEEE-Arithmetik und optimaler Rundung vorausgesetzt.

a) Welchen Wert für e liefert das folgende Programmstück?
(Anmerkung: Der Vergleichsoperator $\sim=$ liefert TRUE bei Ungleichheit der Operanden, sonst FALSE.)

```
x = 1;  e = 0;  y = 1;  z = x + y;
while x ~= z
   y = y/2;  e = e + 1;  z = x + y;
end
```

Verbale Beschreibung:

Formel (unter Verwendung der Paramater von $\mathbb{F}$):

$e =$

Zahlenwert:

b) Welchen Wert für q liefert das folgende Programmstück?

```
x = 1;  q = 0;
while x > 0
   x = x/2;  q = q + 1;
end
```

Verbale Beschreibung:

Formel: $q =$

Zahlenwert:

c) Welchen Wert für r liefert das folgende Programmstück?
(Anmerkung: `Inf` ist eine symbolische Konstante fuer ∞.)

```
x = 1;  r = 0;
while x ~= Inf
   x = 2*x;  r = r + 1;
end
```

Verbale Beschreibung:

Formel: $r =$

Zahlenwert:

2) (16 Punkte)

a) Um die Kondition (Empfindlichkeit auf Störungen im Argument x) der Funktionauswertung

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

für moderate Werte von x zu studieren, bestimme man die maximale relative Konditionszahl dieser Aufgabe für $0 < |x| \leq \frac{\pi}{4}$.

Antwort zu a):

relative Konditionszahl =

b) Es sei nun $\sin x$ für betragskleine Argumente x auszuwerten. Angenommen, Sie verwenden einen Taschenrechner mit defekter sin-Taste (mit Pokémon-Kaugummi verklebt oder wie auch immer); die anderen trigonometrischen Funktionen cos, tan und cot stehen zur Verfügung.[1] Folgende alternative Auswertungen der Sinus-Funktion bieten sich an:

(i)
$$\sqrt{1 - \cos^2 x}$$

(ii)
$$\cos x \cdot \tan x$$

(iii)
$$\sqrt{\cos^2 x + \frac{1 - \cot^2 x}{1 + \cot^2 x}}$$

(iv)
$$\sqrt{\cos^2 x - \cos 2x}$$

(v)
$$\frac{2 \tan \frac{x}{2}}{1 + \tan^2 \frac{x}{2}}$$

(vi)
$$\frac{2 \cos x}{\cot \frac{x}{2} - \tan \frac{x}{2}}$$

[1] $\sqrt{}$, cos, tan und cot seien korrekt implementiert, d.h. sie liefern einen – im Sinne der relativen Genauigkeit des verwendeten Rechners – genauen Wert (z.B. 10 richtige Stellen auf einem Rechner mit 10-stelliger Arithmetik). Over- oder Underflow-Effekte bleiben außer Betracht.

Von welchen dieser 6 Varianten ist zu erwarten, dass sie für betragskleine $|x| \ll 1$ numerisch stabil sind, d.h. ein Ergebnis in befriedigender relativer Genauigkeit liefern, und für welche trifft das nicht zu? Geben Sie jeweils eine kurze (aber möglichst präzise) Begründung für Ihre Antwort.

Antwort zu b) (zutreffendes ankreuzen):

Variante	num. stabil	num. instabil	Begründung
(i)	○	○	
(ii)	○	○	
(iii)	○	○	
(iv)	○	○	
(v)	○	○	
(vi)	○	○	

3) (17 Punkte)

Der Ausdruck

$$f(x) = \frac{\sin x - \cos x}{\sin x \cdot \cos x} \qquad (1)$$

soll auf einem Computer in Gleitpunktarithmetik ausgewertet werden.

a) Man berechne die absolute Konditionszahl $K_{\text{abs}}(x)$

$K_{\text{abs}}(x) =$

Gibt es für ein $x \in [0, \pi/2]$ Stellen, wo $|K_{\text{abs}}(x)| = \infty$ gilt?

- ○ nein, keine Unendlichkeitsstellen von $K_{\text{abs}}(x)$
- ○ ja, $|K_{\text{abs}}(x)| = \infty$ gilt für

  $x =$

  (Falls Ihre Antwort "ja" lautet, bitte alle Unendlichkeitsstellen von $x \in [0, \pi/2]$ eintragen!)

b) Man berechne eine möglichst scharfe relative Rundungsfehlerschranke $S_{\text{rel}}(x)$ für die Auswertung von $f(x)$ gemäß (1) der Form $S_{\text{rel}}(x) = \mathit{Ausdruck}(x, \mathit{eps})$

$S_{\text{rel}}(x) =$

Gibt es Stellen, wo $S_{\text{rel}}(x) = \infty$ gilt?

- ○ nein
- ○ ja, $S_{\text{rel}}(x) = \infty$ gilt für

  $x =$

  (Falls die Antwort "ja" lautet, bitte alle Unendlichkeitsstellen von $x \in [0, \pi/2]$ eintragen!)

Liegt in allfälligen Unendlichkeitsstellen von $S_{\text{rel}}(x)$ eine numerische Instabilität vor?

- ○ nein
- ○ ja

  Kurze Begründung der Antwort:

Falls Ihrer Ansicht nach in allfälligen Unendlichkeitsstellen von $S_{\text{rel}}(x)$ die Auswertung von $f(x)$ gemäß (1) numerisch instabil ist, ist für diese Stellen dann die mathematisch äquivalente Auswertung gemäß

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

numerisch stabil?

- ○ instabil
- ○ stabil

  Kurze Begründung der Antwort:

Institut für Angewandte und Numerische Mathematik
TU Wien

WS 1999/2000

Numerische Mathematik für Informatiker

1. Übungstest am 16. November 1999

Name	Vorname	Kennzahl / Matrikelnummer

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

Der gesamte Rechengang ist auf den beiliegenden Blättern zu dokumentieren.

Zusätzlich beigefügte Zetteln werden bei der Korrektur nicht berücksichtigt.

1) (16 Punkte)

In diesem Beispiel wird der Gleitpunkt-Zahlenbereich $\mathbb{F} = \mathbb{F}(10, 3, -9, 9, \mathit{true})$ betrachtet.

a) Wie groß ist die Anzahl der denormalisierten Zahlen in $\mathbb{F}$?

b) Wie lautet die kleinste positive Zahl in $\mathbb{F}$?

c) Wie groß ist die relative Maschinengenauigkeit *eps*?

Bei optimaler Rundung:

Bei Abschneiden:

d) Warum kann es bei der Implementierung von $\mathbb{F}$ kein implizites erstes Bit geben? (Begründung!)

e) Welches ist die kleinste Potenz 2^{-p}, $p \in \mathbb{N}$, die in $\mathbb{F}$ exakt darstellbar ist?

$p =$

2^{-p} in $\mathbb{F}$:

f) Welches ist die kleinste Potenz 5^{-p}, $p \in \mathbb{N}$, die in $\mathbb{F}$ exakt darstellbar ist?

$p =$

5^{-p} in $\mathbb{F}$:

g) Für $x = 1.01$ berechne man $(x^3-1)/(x-1) = x^2+x+1$ in $\mathbb{F}$ mit Abschneiden nach jedem Rechenschritt.

	Resultat	relativer Fehler
$\frac{x^3-1}{x-1}$ in $\mathbb{F}$		
x^2+x+1 in $\mathbb{F}$		

h) Man berechne $99-70\sqrt{2} = 1/(99+70\sqrt{2})$ in $\mathbb{F}$ mit optimaler ("traditioneller") Rundung nach jedem Rechenschritt.

	Resultat	relativer Fehler
$99-70\sqrt{2}$ in $\mathbb{F}$		
$\frac{1}{99+70\sqrt{2}}$ in $\mathbb{F}$		

2) (17 Punkte)

Betrachtet wird die Auswertung des Ausdruckes

$$f(x) = \frac{\ln(1+x)}{x}$$

für Werte $x \in (0,1)$.

a) Für kleine x-Werte ist dies eine sehr gut konditionierte (d.h. im Sinne der Fortpflanzung einer relativen Störung in x sehr unempfindliche) Problemstellung. Zeigen Sie dies, indem Sie die relative Konditionszahl $K_{f\leftarrow x}$ von $f(x)$ (als Ausdruck in x) und ihren asymptotischen Wert für $x \to 0$ (d.h. $\lim_{x\to 0} K_{f\leftarrow x}$) berechnen.

Antwort zu a):

- Konditionszahl $K_{f\leftarrow x}$ als Funktion von x:

$$K_{f\leftarrow x} =$$

- Asymptotischer Wert für $x \to 0$:

$$\lim_{x\to 0} K_{f\leftarrow x} =$$

b) In Kontrast zu a) ergibt sich bei Auswertung von $f(x)$ in Gleitpunktarithmetik ein sehr ungenaues Resultat. In einer "Taschenrechner-Arithmetik" $\mathbb{F} = \mathbb{F}(10, 10, \ldots)$ (mit *round-to-nearest* und korrekt implementierter Logarithmus-Funktion) etwa erhält man für $x = 0.1234567890 \cdot 10^{-7} \in \mathbb{F}$ den sehr ungenauen Wert 0.9720000032 (der auf 10 Stellen gerundete exakte Wert ist 0.9999999938).

Analysieren Sie die Situation, indem Sie den relativen Fehler des numerischen Ergebnisses für kleine x ($0 < x \ll 1$) und beliebige relative Maschinengenauigkeit *eps* als Ausdruck in x und *eps* darstellen.

(Hinweis: Höhere Potenzen kleiner Größen wie x, *eps* können dabei vernachlässigt werden; beachte auch $\ln(1+\delta) \approx \delta$ für kleine Werte δ.)

Antwort zu b):

c) Welche der drei durchgeführten Operationen (bzw. der dabei begangene Rundungsfehler), (i) die Addition $1+x$, (ii) die Auswertung des Logarithmus, oder (iii) die Division durch x, ist für das ungenaue Ergebnis verantwortlich? Schriftliche Erläuterung erwünscht.

Antwort zu c):

3) (17 Punkte)

Der Ausdruck

$$f(a) = \frac{\cos a - 1}{a^2}$$

sei für eine gegebene Maschinenzahl $a \in (0, \frac{\pi}{2}]$ in Gleitpunktarithmetik zu berechnen. Dabei soll erreicht werden, dass die relative Genauigkeit des Resultates nicht größer als $m \cdot eps$ ist, mit einer "moderaten" Konstanten m. Der Kosinusterm, $\cos a$, wird mit Hilfe einer vordefinierten Prozedur COS "sauber" ausgewertet, 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 den Ausdruck $m(a) > 0$ so, dass[1]

$$\tilde{f}(a) = f(a)(1+\varepsilon), \quad |\varepsilon| \leq m(a) \cdot eps$$

gilt.

Antwort zu a):

b) Kann man $m(a)$ für $a \in (0, \frac{\pi}{2}]$ gleichmäßig mit $m(a) \leq m$ abschätzen? Für welche Werte von a ergibt sich ein großer relativer Fehler? Kommentieren Sie Ihr Resultat ausführlich!

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

Antwort zu b):

c) Betrachten Sie nun die unter b) diagnostizierte numerisch instabile Situation. Modifizieren Sie für diesen Fall den Ausdruck $f(a)$ dadurch, dass Sie $\cos a$ durch die ersten drei Glieder seiner Taylor-Entwicklung[2] um die Stelle $a = 0$ ersetzen. Schreiben Sie die sich ergebende Approximation für $f(a)$ so an, dass diese in numerisch stabiler Weise ausgewertet werden kann. (Mit ausführlicher Begründung; worin besteht der entscheidende Unterschied zur direkten Auswertung von $f(a)$?)

Antwort zu c):

[2] $\cos a = 1 - \frac{a^2}{2!} + \frac{a^4}{4!} - \frac{a^6}{6!} + \ldots$

1) (15 Punkte)

Im Folgenden sei $\alpha \neq 1$ ein reeller Parameter. Durch die Lösung $(x, y) = (x(\alpha), y(\alpha))$ des α-abhängigen Gleichungssystems

$$\begin{aligned} -2x + (1-\alpha)y &= 1 - 2\alpha \\ x - (1-\alpha)y &= -1 + \alpha \end{aligned}$$

sind zwei Funktionen $x(\alpha)$ und $y(\alpha)$ definiert.

a) Bestimmen Sie $x(\alpha)$ und $y(\alpha)$ und dazu die entsprechenden relativen Konditionszahlen[1] $K_{x\leftarrow\alpha}$ und $K_{y\leftarrow\alpha}$ bezüglich kleiner Störungen von α.

Antwort zu a):

$x(\alpha) =$

Relative Konditionszahl $K_{x\leftarrow\alpha}$ als Funktion von α:

$y(\alpha) =$

Relative Konditionszahl $K_{y\leftarrow\alpha}$ als Funktion von α:

b) Es sei $\alpha = 0.99$ gegeben. Für welchen Bereich von Werten $\tilde{\alpha} \approx 0.99$ ist zu erwarten, dass die relative Differenz in der Auswertung von y, d.h. die Größe $\left|\frac{y(\tilde{\alpha})-y(0.99)}{y(0.99)}\right|$ maximal 10^{-3} ist? Man gebe das entsprechende Intervall von $\tilde{\alpha}$-Werten an.
(Antwort auf Grund des unter a) erhaltenen Ergebnisses. Allfällige Rechenfehler bei der Auswertung der Funktion y bleiben hier außer Betracht.)

Antwort zu b): $\tilde{\alpha} \in [0.99-l,\ 0.99+u]$ mit

$l =$ $u =$

[1] Gemeint ist hier die mittels Differentiation zu ermittelnde Konditionszahl im Sinne des Vorlesungsskriptums. Die Tatsache, dass es sich dabei streng genommen um (wenn auch sehr gute) Konditionsschätzungen für den Fall kleiner Störungen handelt, möge im Folgenden ignoriert werden.

Institut für Angewandte und Numerische Mathematik
TU Wien

WS 1998/99

Numerische Mathematik für Informatiker

1. Übungstest am 10. November 1998

Name	Vorname	Kennzahl / Matrikelnummer

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

Punkte maximal: 50

Der gesamte Rechengang ist auf den beiliegenden Blättern zu dokumentieren.

Zusätzlich beigefügte Zetteln werden bei der Korrektur nicht berücksichtigt.

Tragen Sie die Ergebnisse ein und kreuzen Sie die zutreffende Aussage an !

Fehlerschranke für den relativen Rundungsfehler von
$\varphi_1(x) := \frac{\sqrt{x-1}-\sqrt{x}}{2}$:

$|\rho| \leq$

Fehlerschranke für den relativen Rundungsfehler von
$\varphi_2(x) := \frac{-1}{2(\sqrt{x-1}+\sqrt{x})}$:

$|\rho| \leq$

Die Formel $\boxed{\varphi_1(x)}$ $\boxed{\varphi_2(x)}$ ist bei der Auswertung von $\varphi(x)$
für große x-Werte vorzuziehen.
Begründung:

2) (17 Punkte)

Der Ausdruck

$$\varphi(x) = \frac{\sqrt{x-1}-\sqrt{x}}{2} = \frac{-1}{2(\sqrt{x-1}+\sqrt{x})}, \quad x > 1,$$

soll für große Werte von $x \gg 1$ berechnet werden. Dabei ist bekannt, dass diese Auswertung gut konditioniert ist.

a) 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 ρ, mit [2]

$$\square(\varphi(x)) \doteq \varphi(x)(1+\rho)$$

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

Hinweis: Man gehe davon aus, dass die Basis der Arithmetik $b = 2$ ist. Demgemäß muss bei der Division durch 2 kein elementarer Rundungsfehler berücksichtigt werden.

Der Einfachheit halber nehme man weiters an, dass für die Auswertung der Wurzelfunktionen

$$\square(\sqrt{x-1}) \doteq \sqrt{x-1}\,(1+\rho_1), \quad \square(\sqrt{x}) \doteq \sqrt{x}\,(1+\rho_2)$$

mit $|\rho_1| \leq 2eps$ und $|\rho_2| \leq eps$ gilt.

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

(Bitte die Antworten umseitig eintragen.)

[2] $\square(\varphi(x))$ symbolisiert die jeweilige Auswertung von φ in Gleitpunktarithmetik.

Tragen Sie die Ergebnisse ein und kreuzen Sie die zutreffende Aussage an !

a) $\bar{p} =$

$\overline{eps} =$

b) $\bar{e}_{max} =$

c) $\bar{e}_{min} =$

d) implizites erstes Bit ist realisierbar:

○ JA ○ NEIN

Begründung:

e) $\bar{N} =$ Bit

f) $\varepsilon =$

g) $\bar{\bar{x}}_{min} =$

3) (18 Punkte)

Die einfach genauen IEC/IEEE-Gleitpunktzahlen

$$\mathbb{F}(b, p, e_{min}, e_{max}, denorm) = \mathbb{F}(2, 24, -125, 128, true)$$

sollen durch dezimale Gleitpunktzahlen $\mathbb{F}(10, \bar{p}, \bar{e}_{min}, \bar{e}_{max}, true)$ ersetzt werden, die den IEC/IEEE-Zahlen möglichst "ähnlich" sind.

a) Was ist der kleinste Wert für die Mantissenlänge $\bar{p}$, für den die relative Maschinengenauigkeit $\overline{eps}$ bei optimaler Rundung nicht schlechter als die relative Maschinengenauigkeit *eps* der IEC/IEEE - Zahlen ist?

b) Was ist der kleinste Wert von $\bar{e}_{max}$, für den $x_{max} < \bar{x}_{max}$ gilt?

c) Was ist der größte Wert von $\bar{e}_{min}$, für den $\bar{x}_{min} < x_{min}$ gilt?

d) Kann bei den obigen dezimalen Gleitpunktzahlen ein implizites erstes Bit realisiert werden?

e) Welche Formatbreite $\bar{N}$ (in Bit) benötigt man für die Speicherung dezimaler Gleitpunktzahlen mit den obigen Parametern, wenn jede Dezimalziffer durch 4 Bits codiert wird (BCD - Code)?

f) Was ist die größte dezimale Gleitpunktzahl $\varepsilon > 0$, für die sich in obigem Zahlensystem und bei optimaler Rundung

$$1 + \varepsilon = 1$$

ergibt?

g) Welches ist die kleinste positive denormalisierte Zahl $\bar{\bar{x}}_{min}$ aus obigem Zahlensystem?

(Bitte die Antworten umseitig eintragen.)

Institut für Angewandte und Numerische Mathematik
TU Wien

WS 1997/98

Numerische Mathematik für Informatiker

1. Übungstest am 11. November 1997

Name	Vorname	Kennzahl / Matrikelnummer

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

Der gesamte Rechengang ist auf den beiliegenden Blättern zu dokumentieren.

Zusätzlich beigefügte Zetteln werden bei der Korrektur nicht berücksichtigt.

1) (17 Punkte)

Zugrunde liege eine 6-stellige hexadezimale Gleitpunktarithmetik $\mathbb{F}(16, 6, \ldots)$ mit abschneidender Rundungsvorschrift ('round towards zero'). Man betrachte Matrizen A der Gestalt

$$A = \begin{pmatrix} 0.5 + 2^{-n} & 1.0 + 2^{-n} \\ 8.5 + 2^{-n} & 17.0 + 2^{-n} \end{pmatrix},$$

$n \geq 0$ eine natürliche Zahl, und beantworte folgende Fragen (jeweils mit Begründung):

a) Für welches größte n ist die Matrix A in dem gegebenen Gleitpunktzahlenbereich exakt darstellbar?

b) Für welches kleinste n ergibt die Berechnung (in der gegebenen Gleitpunktarithmetik) der Determinante der nach Rundung von A (gemäß der gegebenen Rundungsvorschrift) entstandenen Matrix den Wert 0?

Antwort zu a):

$n =$

Antwort zu b):

$n =$

2) (16 Punkte)

Betrachtet wird der arithmetische Ausdruck

$$\varphi(x) = (1+x)^2 - 1 \qquad \text{für} \quad x \in (0,1).$$

a) Berechnen Sie die relative Konditionszahl $K_{\varphi \leftarrow x}$ von $\varphi(x)$ bezüglich kleiner relativer Störungen von x und geben Sie dafür eine für beliebige $x \in (0,1)$ gültige obere Schranke an.

Konditionszahl $K_{\varphi \leftarrow x}$ als Funktion von x:

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

Schranke für $|K_{\varphi \leftarrow x}|$ $(x \in (0,1))$:

$$|K_{\varphi \leftarrow x}| \leq$$

b) Diskutieren Sie die Frage, ob die Auswertung von $\varphi(x)$ in Gleitpunktarithmetik in der oben gegebenen Gestalt für beliebige $x \in (0,1)$ in numerisch stabiler Weise erfolgt. Begründen Sie Ihre Antwort mittels qualitativer Argumentation (unter Verzicht auf eine formale Rundungsfehleranalyse). Sollte Ihre Diagnose auf numerische Instabilität lauten, so geben Sie eine algebraisch äquivalente Umformulierung von $\varphi(x)$ an, deren Auswertung in Gleitpunktarithmetik für beliebige $x \in (0,1)$ numerisch stabil ist (wiederum mit qualitativer Begründung).

3) (17 Punkte)

Der Ausdruck

$$\varphi(x) = 1 - \frac{1}{1+x} = \frac{x}{1+x}, \quad x \text{ positiv},$$

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

- 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.

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

$$|\varepsilon| \leq$$

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

$$|\varepsilon| \leq$$

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

3) (17 Punkte)

In den Anwendungen der Linearen Algebra sind häufig Matrixausdrücke zu berechnen, in denen Terme

$$X = A^{-1}B \quad \text{bzw.} \quad Y^T = B^T A^{-1} \tag{1}$$

mit einer regulären Matrix $A \in \mathbb{R}^{n \times n}$ und einer beliebigen Matrix $B \in \mathbb{R}^{n \times m}$ vorkommen. Solche Terme lassen sich unter Verwendung einer LU-Faktorisierung von A berechnen.

a) Durch welche Faktorisierungen und Rücksubstitutionen erhält man die Matrizen X und Y?

b) Wieviele Gleitpunktoperationen sind bei der Vorgangsweise nach a) erforderlich und wieviele bei der "naiven" Berechnung von (1), d.h. durch explizite Berechnung von A^{-1} und anschließende Ermittlung von (1)? (Genaue Angabe der Operationszahl ist erforderlich – asymptotische Komplexität reicht nicht aus.)

c) Sollte die Inverse $X = A^{-1}$ ausnahmsweise doch explizit benötigt werden, kann sie als Sonderfall von (1) für $B = I \in \mathbb{R}^{n \times n}$ berechnet werden. Wieviele Gleitpunktoperationen sind hierfür erforderlich?

d) Wieviele Gleitpunktoperationen sind bei der Vorgangsweise nach a) erforderlich, falls A eine Tridiagonalmatrix ist? Welche Argumente sprechen in diesem Spezialfall besonders stark gegen die "naive" Berechnung von (1)?

Inst. f. Angew. u. Numer. Math.
TU Wien

WS 1996/97

Numerische Mathematik für Informatiker

1. Übungstest am 12. Nov. 1996

Name	Vorname	Kennzahl / Matrikelnummer

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

Bitte verwenden Sie nur die beiliegenden Blätter.
Zusätzlich beigefügte Zetteln werden bei der Korrektur nicht berücksichtigt.

1) (17 Punkte)

In Sonderfällen laufen Rechenoperationen in einer Computerarithmetik rundungsfehlerfrei ab. Man überlege, ob in den folgenden Fällen solche Sondersituationen vorliegen.[1] Falls man meint, daß dies nicht der Fall ist, belege man diese Ansicht jeweils mit einem konkreten Beispiel.

a) $x+y$; $x, y \in \mathbb{M}(10, 10, 99, -99)$ bzw. $\mathbb{F}(10, 10, -98, 100, \mathit{false})$; es werde angenommen, daß sowohl bei x als auch bei y nur die drei führenden Stellen $\neq 0$ sind.

$x+y$ ist rundungsfehlerfrei		
ja*	nein*	Beispiel eines rundungsfehlerbehafteten
○	○	Falles:
		$x =$, $y =$

b) $x-y$; $x, y \in \mathbb{M}(10, 6, 9, -9)$ bzw. $\mathbb{F}(10, 6, -8, 10, \mathit{true})$; es werde angenommen, daß die Exponenten von x und y gleich sind.

$x-y$ ist rundungsfehlerfrei		
ja*	nein*	Beispiel eines rundungsfehlerbehafteten
○	○	Falles:
		$x =$, $y =$

c) $x+1$; $x \in \mathbb{M}(10, 3, 9, -9)$ bzw. $\mathbb{F}(10, 3, -8, 10, \mathit{true})$; es werde die Gültigkeit von $10 \leq x \leq 999$ angenommen.

$x+1$ ist rundungsfehlerfrei		
ja*	nein*	Beispiel eines rundungsfehlerbehafteten
○	○	Falles:
		$x =$

d) $\frac{x}{2}$; $x \in \mathbb{M}(10, 10, 99, -99)$ bzw. $\mathbb{F}(10, 10, -98, 100, \mathit{false})$; es werde angenommen: (i) $x > 1$,
(ii) genau die beiden führenden Stellen von x sind ungleich Null.

$\frac{x}{2}$ ist rundungsfehlerfrei		
ja*	nein*	Beispiel eines rundungsfehlerbehafteten
○	○	Falles:
		$x =$

e) $\frac{x}{2}$; $x \in \mathbb{M}(2, 24, 127, -126)$ bzw. $\mathbb{F}(2, 24, -125, 128, \mathit{true})$; es werde angenommen: $x > 0$.

$\frac{x}{2}$ ist rundungsfehlerfrei		
ja*	nein*	Beispiel eines rundungsfehlerbehafteten
○	○	Falles:
		$x =$

[1] Zutreffendes bitte jeweils ankreuzen.

2) (17 Punkte)

In verschiedenen Büchern (Nachschlagewerken, Lehrbüchern etc.) findet man für das lineare Gleichungssystem

$$\begin{aligned} a_{11}\,x_1 + a_{12}\,x_2 &= b_1 \\ a_{21}\,x_1 + a_{22}\,x_2 &= b_2 \end{aligned}$$

die folgende Lösungsformel (Cramersche Regel):

$$\begin{aligned} x_1 &= \frac{a_{22}\,b_1 - a_{12}\,b_2}{a_{11}\,a_{22} - a_{12}\,a_{21}} \\ x_2 &= \frac{a_{11}\,b_2 - a_{21}\,b_1}{a_{11}\,a_{22} - a_{12}\,a_{21}}\,. \end{aligned}$$

Für

$$\begin{aligned} 0.2038\,x_1 + 0.1218\,x_2 &= 0.2014 \\ 0.4071\,x_1 + 0.2436\,x_2 &= 0.4038 \end{aligned}$$

berechne man:

a) die exakte Lösung x_1, x_2;

b) die numerische Lösung $\tilde{x}_1, \tilde{x}_2$ unter Verwendung obiger Formeln und einer vierstelligen ($p = 4$) dezimalen ($b = 10$) Gleitpunktarithmetik mit optimaler Rundung (round to even);

c) die relativen Fehler von $\tilde{x}_1$ und $\tilde{x}_2$. Man gebe eine genaue (verbale) Begründung für die Ursache der Größenordnung dieser Fehler.

d) Eine wievielstellige ($p = ?$) dezimale Gleitpunktarithmetik benötigt man, um das exakte Resultat zu erhalten?

e) Ist die Erhöhung der Rechengenauigkeit (Erhöhung von p) eine sinnvolle Maßnahme, falls die Daten $(a_{11}, \ldots, a_{22}, b_1, b_2)$ des Problems mit Datenfehlern behaftet sind? Man gebe eine genaue (verbale) Begründung der Antwort.