Αριθμητική Ανάλυση και Περιβάλλοντα Υλοποίησης
Εαρινό 2021
Σετ Ασκήσεων #1

Δούρου Βασιλική Ευαγγελία
Α.Μ.:1072633

28 Μαρτίου 2021

Περιεχόμενα

1.Ερώτηση 1

Το λάπτοπ που χρησιμοποιώ έχει ως επεξεργαστή τον Intel Core i7-8550U.Έπειτα από μία έρευνα, μέσω μίας ιστοσελίδας που σύγκρινε επεξεργαστές της Intel εκ των οποίων ο ένας από τους δύο ήταν ο δικός μου, ανακάλυψα πως χρησιμοποιεί αριθμητική IEEE 754 απλής ακρίβειας (FP32 floating point).

Πηγή: https://gadgetversus.com/processor/intel-core-i5-9400-vs-intel-core-i7-8550u/

2.Ερώτηση 2

(α) NaN
(β) 1

(γ) 0
(δ) 0
(ε) NaN
(στ) Μετά την εκτέλεση των εντολών τα αποτελέσματα ήταν:
a=1, b=NaN, c=1, d=0.

3.Ερώτηση 3

Υπάρχουν αριθμοί κινητής υποδιαστολής για τους οποίους να ισχύει 1+x==x. Αυτό συμβαίνει καθώς το άθροισμα αριθμών που έχουν πολύ μεγάλη διαφορά μεταξύ τους σε μία υπολογιστική μηχανή, δεν υπολογίζεται όπως θα υπέθετε κανείς με ¨θεϊκή' αριθμητική, αλλά του εφαρμόζεται στρογγυλοποίηση. Αυτό παρατηρήθηκε και πειραματικά, καθώς στο περιβάλλον της MATLAB οι ακόλουθες πράξεις έβγαλαν ως αποτέλεσμα 1 ή αλλιώς TRUE:
$1 + 5.8932e + 11 == 5.8932e + 11$
$1 + 5.8932e + 9 == 5.8932e + 9$

4.Ερώτηση 4

Στην αριθμητική κινητής υποδιαστολής διπλής ακρίβειας χρησιμοποιούνται 64 bit για την αναπαράσταση αριθμών σε αντιθέση με την αριθμητική μονής ακρίβειας που χρησιμοποιούνται 32 bit. Θα μπορούσαμε να χρησιμοποιήσουμε τη συνάρτηση eps(x) της MATLAB για να υπολογίσουμε το έψιλον της μηχανής, δηλαδή την απόσταση του x από τον αμέσως επόμενο διαδοχικό του αριθμό κινητής υποδιαστολής, χρησιμοποιώντας ως όρισμα και το 'double' και το 'single'. Έτσι, ανάλογα με το πόσο κοντά είναι δύο διαδοχικοί αριθμοί σε αριθμητική διπλής και απλής ακρίβειας, μπορούμε να αποδείξουμε πως είναι πάνω από δύο φορές πιο ακριβής η πρώτη περίπτωση. Στη MATLAB παίρνουμε τα ακόλουθα αποτελέσματα:
$eps('double') = 2.2204e - 16$
$eps('single') = 1.1921e - 07$

5.Ερώτηση 5

Χρησιμοποιώντας την MATLAB και γράφοντας τα δύο διανύσματα που δίνονται, παρατηρήθηκε πως και τα δύο αναπαριστάθηκαν ακριβώς.

6.Ερώτηση 6

$(fl(x_1)\oplus fl(x_2))\oplus fl(x_3) = (x_1(1+\delta_1)\oplus x_2(1+\delta_1))\oplus x_3(1+\delta_1) = ((x_1+x_2)(1+\delta_1)(1+\delta_2))\oplus x_3(1+\delta_1) = (x_1+x_2)(\delta_1+\delta_2+\delta_1\delta_2)(1+\delta_3)+x_3(1+\delta_1)(1+\delta_3) = = (x_1+x_2)(\delta_1+\delta_2+\delta_1\delta_2+\delta_1\delta_3+\delta_2\delta_3+\delta_1\delta_2\delta_3)+x_3(\delta_1+\delta_3+\delta_1\delta_3).$
Επομένως:
$|(x_1+x_2+x_3)-((fl(x_1)\oplus fl(x_2))\oplus fl(x_3))| \leq |(x_1+x_2)(\delta_1+\delta_2+\delta_1\delta_2+\delta_1\delta_3+\delta_2\delta_3+\delta_1\delta_2\delta_3)+x_3(\delta_1+\delta_3+\delta_1\delta_3)|$
$\leq (|x_1|+|x_2|)(2u+3u^2+u^3)+|x_3|(2u+u^2)$

$\frac{|(x_1+x_2+x_3)-((fl(x_1)\oplus fl(x_2))\oplus fl(x_3))|}{|(x_1+x_2+x_3)|} \leq \frac{|x_1|+|x_2|+|x_3|}{|x_1+x_2+x_3|}(2u+3u^2)$
Αγνοώντας τους όρους $O(u^3)$:

$$\frac{|(x_1+x_2+x_3)-((fl(x_1)\oplus fl(x_2))\oplus fl(x_3))|}{|(x_1+x_2+x_3)|} \leq 2u+3u^2.$$

7.Ερώτηση 7

Όταν γίνταει αφαίρεση παραπλήσιων ομόσημων τιμών και το αποτέλεσμα της αφαίρεσης είναι μικρότερο από τα δεδομένα και μάλιστα τάξης μεγέθους περίπου ίδιας με το θόρυβο που αυτά περιέχουν, τότε μπορεί να έχουμε καταστροφική απαλοιφή. Σε αυτή την περίπτωση πρέπει να αντικαταστήσουμε τη μαθηματική πράξη με μία ισοδύναμη.
Στην δικία μας περίπτωση το $1-\cos x$ για πολύ μικρά $|x|$ είναι ίσο με 0 καθώς το $\cos x$ τείνει στο 1. Η μαθηματική αυτή έκφραση μπορεί να γραφτεί ως $\cos(x-x)-\cos x$ καθώς $\cos 0 = 1$. Από τη στιγμή που το x είναι πολύ μικρό και το -x θα είναι πολύ μικρό άρα είναι της μορφής $\cos(x+\delta)-\cos x$. Άρα, μπορούμε να χρησιμοποιήσουμε τον τύπο $-2\sin(\delta/2)\sin(x+\delta/2)$.
Όπως και φαίνεται και πειραματικά από τη MATLAB για $x = 10^{-9}, \delta = -x = -10^{-9}$:
$1-\cos(10^{-9}) = 0$
$-2\sin((-10^{-9})/2) * sin(10^{-9}+(-10^{-9})/2) = 5.0000e-19$

8.Ερώτηση 8

Ο εκθέτης παίρνει τιμές ανάμεσα στο -3 και στο 4, άρα χρειάζεται 3 bit για την αναπαράσταση του και θα έχει πόλωση 3,άρα ένας κανονικοποιημένος αριθμός μπορεί να έχει ως πολωμένο εκθέτη $0 <$Εκθέτης< 7.
Άν έχουμε προσημασμένους αριθμούς τότε ο ελάχιστος κανονικοποιημένος αριθμός είναι ο $-1.1111 * 2^3$, ενώ αν έχουμε μη προσημασμένους ο ελάχιστος θα είναι ο $1.0000 * 2^{-2}$. Ο μέγιστος θετικός και στις δύο περιπτώσεις θα είναι ο $1.1111 * 2^3$

9.Ερώτηση 9

Μέσω της MATLAB υπολογίστηκε πως το Ιακωβιανό μητρώο της δοθείσας συνάρτησης είναι το
$\begin{pmatrix} 1 & 1 & 2 \end{pmatrix}$.

10.Ερώτηση 10

Το Ιακωβιανό μητρώο είναι το ακόλουθο:
$\begin{pmatrix} 2x & z & y+1 \\ b+2ax & 0 & 0 \end{pmatrix}$

11.Ερώτηση 11

Η έκφραση $x^2 - y^2$ παρουσιάζει σε πολλές περιπτώσεις καταστροφική απαλοιφή καθώς όταν γίνεται αφαίρεση παραπλήσιων ομόσημων τιμών, το αποτέλεσμα της αφαίρεσης είναι μικρότερο από τα δεδομένα και μάλιστα τάξης μεγέθους περίπου ίδιας με το θόρυβο που αυτά περιέχουν.
Αν όμως τα x, y έχουν πολύ διαφορετικά μεγέθη, τότε η έκφραση $(x-y)(x+y)$ έχει τρία σφάλματα στρογγυλοποίησης, ενώ η $x^2 - y^2$ έχει μόνο δύο, αφού το σφάλμα στρογγυλοποίησης που λαμβάνεται με το τετράγωνο ενός σχετικά μικρού αριθμού (ο μικρότερος των x και y) είναι τόσο αμελητέο που υπάρχουν ουσιαστικά μόνο δύο βήματα σφαλμάτων στρογγυλοποίησης, καθιστώντας τη διαφορά των τετραγώνων πιο ακριβή.

13.Ερώτηση 13

Έστω ότι υπάρχει x_{prog} κοντά στο x τέτοιο ώστε $f_{prog}(x) = f(x_{prog})$. Τότε ο αλγόριθμος χαρακτηρίζεται προς τα πίσω ευσταθής στο x. Αν αυτό συμβαίνει για κάθε x στο πεδίο ορισμού της f ο αλγόριθμος αποκαλείται προς τα πίσω ευσταθής.
Ο υπολογισμός της $(x+y)+z$ είναι πίσω ευσταθής καθώς:
$(x+y)(1+\delta_1)+z = x(1+\delta_1)+y(1+\delta_1)+z = x'+y'+z =$
$(x'+y'+z)(1+\delta_2) = x'(1+\delta_2)+y'(1+\delta_2)+z(1+\delta_2) = x''+y''+z'.$

Και αρθμητική άπειρης ακρίβειας να χρησιμοποιήσουμε, η σειρά με την οποία γίνεται η πρόσθεση έχει σημασία. Οπότε από τη στιγμή που προσθέτουμε πρώτα έναν

πολύ μεγάλο αριθμό με έναν σημαντικά μικρότερο του, η στρογγυλοποίηση που θα γίνει από τη μηχανή μας, θα μας επιστρέψει ουσιαστικά τον ίδιο αριθμό, οπότε στη δεύτερη φάση της πρόσθεσης με τον αντιθετό του, το αποτέλεσμα θα μηδενιστεί.

14.Ερώτηση 14

Σύμφωνα με την πηγή που δόθηκε ο δείκτης κατάστασης της συνάρτησης ax είναι το 1, της $\sin(x)$ είναι το $|x\cot(x)|$, της e^x είναι το $|x|$ και της $\frac{x}{x+a}$ είναι το $(\frac{x}{x+a})'\frac{|x|}{|\frac{x}{x+a}|} = \frac{x'(x+a)-x(x+a)'}{(x+a)^2}|x+a| = \frac{x+a-x}{(x+a)^2}|x+a| = \frac{a}{|x+a|^2}|x+a| = \frac{a}{|x+a|}$.

15.Ερώτηση 15

Σύμφωνα με τη MATLAB αυτός είναι ίσος με 1.

17.Ερώτηση 17

Έστω ότι υπάρχει x_{prog} κοντά στο x τέτοιο ώστε $f_{prog}(x) = f(x_{prog})$. Τότε ο αλγόριθμος χαρακτηρίζεται προς τα πίσω ευσταθής στο x. Αν αυτό συμβαίνει για κάθε x στο πεδίο ορισμού της f ο αλγόριθμος αποκαλείται προς τα πίσω ευσταθής.
Ο υπολογισμός της $x^2 + \sqrt{1+y^2}$ είναι πίσω ευσταθής καθώς:
$x*x(1+\delta_1)+\sqrt{1+y*y(1+\delta_1)} = x'^2+\sqrt{1+y'^2} = x'^2+\sqrt{(1+y'^2)(1+\delta_2)} =$
$x'^2+\sqrt{1(1+\delta_2)+y'^2(1+\delta_2)} = x'^2+\sqrt{1+y''^2} = x'^2+\sqrt{1+y''^2}(1+\delta_3) =$
$x'^2+\sqrt{1+y''^2}' = (x'^2+\sqrt{1+y''^2}')(1+\delta_4) = x'^2(1+\delta_4)+\sqrt{1+y''^2}'(1+\delta_4) =$
$x''^2+\sqrt{1+y''^2}''$.

19.Ερώτηση 19

Για $x = 10^{-1}$ τα αποτελέσματα ήταν τα ακόλουθα:
$exp(10^{-1}) - 1 = 0.1052$
$expm1(10^{-1}) = 0.1052$
$10^{-1}/1! + (10^{-1})^2/2! + (10^{-1})^3/3! + (10^{-1})^4/4! + (10^{-1})^5/5! = 0.1052$
Για $x = 10^{-2}$ τα αποτελέσματα ήταν τα ακόλουθα:
$exp(10^{-2}) - 1 = 0.0101$
$expm1(10^{-2}) = 0.0101$

$10^{-2}/1! + (10^{-2})^2/2! + (10^{-2})^3/3! + (10^{-2})^4/4! + (10^{-2})^5/5! = 0.0101$
Για $x = 10^{-3}$ τα αποτελέσματα ήταν τα ακόλουθα:
$exp(10^{-3}) - 1 = 0.0010$
$expm1(10^{-3}) = 0.0010$
$10^{-3}/1! + (10^{-3})^2/2! + (10^{-3})^3/3! + (10^{-3})^4/4! + (10^{-3})^5/5! = 0.0010$
Για $x = 10^{-7}$ τα αποτελέσματα ήταν τα ακόλουθα:
$exp(10^{-7}) - 1 = 1.0000e - 07$
$expm1(10^{-7}) = 1.0000e - 07$
$10^{-7}/1! + (10^{-7})^2/2! + (10^{-7})^3/3! + (10^{-7})^4/4! + (10^{-7})^5/5! = 1.0000e - 07$
Για $x = 10^{-10}$ τα αποτελέσματα ήταν τα ακόλουθα:
$exp(10^{-10}) - 1 = 1.0000e - 10$
$expm1(10^{-10}) = 1.0000e - 10$
$10^{-10}/1! + (10^{-10})^2/2! + (10^{-10})^3/3! + (10^{-10})^4/4! + (10^{-10})^5/5! = 1.0000e - 10$
Για $x = 10^{-20}$ τα αποτελέσματα ήταν τα ακόλουθα:
$exp(10^{-20}) - 1 = 0$
$expm1(10^{-20}) = 1.0000e - 20$
$10^{-20}/1! + (10^{-20})^2/2! + (10^{-20})^3/3! + (10^{-20})^4/4! + (10^{-20})^5/5! = 1.0000e - 20$
Για $x = 10^{-25}$ τα αποτελέσματα ήταν τα ακόλουθα:
$exp(10^{-25}) - 1 = 0$
$expm1(10^{-25}) = 1.0000e - 25$
$10^{-25}/1! + (10^{-25})^2/2! + (10^{-25})^3/3! + (10^{-25})^4/4! + (10^{-25})^5/5! = 1.0000e - 25$
Για $x = 10^{-30}$ τα αποτελέσματα ήταν τα ακόλουθα:
$exp(10^{-30}) - 1 = 0$
$expm1(10^{-30}) = 1.0000e - 30$
$10^{-30}/1! + (10^{-30})^2/2! + (10^{-30})^3/3! + (10^{-30})^4/4! + (10^{-30})^5/5! = 1.0000e - 30$
Για $x = 10^{-40}$ τα αποτελέσματα ήταν τα ακόλουθα:
$exp(10^{-40}) - 1 = 0$
$expm1(10^{-40}) = 1.0000e - 40$
$10^{-40}/1! + (10^{-40})^2/2! + (10^{-40})^3/3! + (10^{-40})^4/4! + (10^{-40})^5/5! = 1.0000e - 40$

Παρατηρούμε πως ο υπολογισμός της πράξης $e^x - 1$ με τον τύπο $expm1$ της $MATLAB$ είναι πιο ακριβής από ότι με τον τύπο $exp(x) - 1$ για πολύ μικρές τιμές του x.