ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΑΝΟΙΚΤΑ ακαδημαϊκά μαθήματα ΠΠ

Επιστημονικός Υπολογισμός Ι

Ενότητα **4** : Μοντέλο Αριθμητικής και Σφάλματα Υπολογισμού

Ευστράτιος Γαλλόπουλος
Τμήμα Μηχανικών Η/Υ & Πληροφορικής

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ 2007-2013
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο **«Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Πατρών»** έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

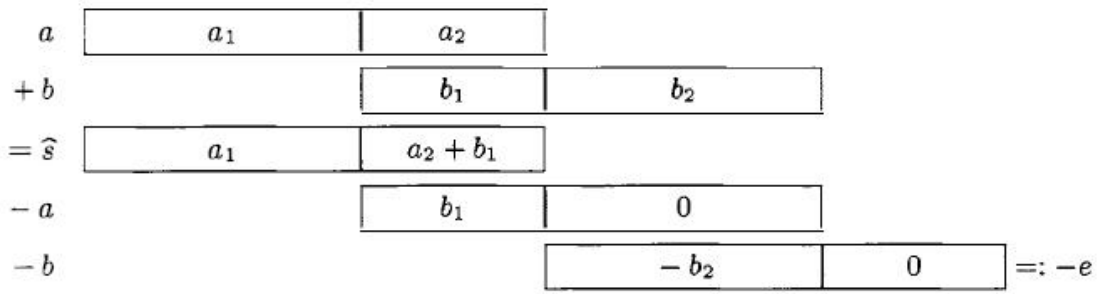
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σκοπός Ενότητας

- Απώλεια πληροφορίας στον επιστημονικό υπολογισμό.
- Αριθμητικό μοντέλο και πρότυπο αριθμητικής κινητής υποδιαστολής IEEE.
- Σφάλματα στρογγύλευσης και διάδοσή τους.
- Σφάλματα στρογγύλευσης και διάδοσή τους.
- Δείκτες κατάστασης προβλήματος και αλγορίθμου.
- Θεωρία και εργαλεία εκτίμησης σφάλματος και ποιότητας υπολογισμών.

Περιεχόμενα

Επανορθωμένη άθροιση I



Σχήμα: Ιδιοφυής ιδέα (Gill'51, Kahan'65) από (Higham'02)

Για κάθε α.κ.υ. $|a| \geq |b|$ μπορούμε να υπολογίσουμε

$$\hat{s} = \mathrm{fl}(a + b), \quad \hat{e} = \mathrm{fl}((a - \hat{s}) + b)$$

χρησιμοποιώντας στρογγύλευση προς το πλησιέστερο. Το υπολογισμένο $\hat{e}$ είναι το ακριβές σφάλμα τής άθροισης! Ειδικότερα,

$$a + b = \hat{s} + \hat{e}$$

Παρόλα αυτά, $\mathrm{fl}(\hat{s} + \hat{e}) = \hat{s}$ γιατί το αποτέλεσμα $\hat{s} = \mathrm{fl}(a + b)$ είναι το καλύτερο δυνατό (λόγω αρχής ακριβούς στρογγύλευσης). Άρα το $\hat{e}$ από μόνο του δεν παρέχει μεγαλύτερη ακρίβεια.

Επανορθωμένη άθροιση II

Ιδέα επανόρθωσης για 3 τιμές: Αν αθροίζονται 3 τιμές, αποθηκεύουμε το ακριβές σφάλμα $\hat{e}_1$ ως την επόμενη πρόσθεση. Τότε αυτό και το ακριβές σφάλμα της επόμενης πρόσθεσης, έστω $\hat{e}_2$, αν προστεθούν σε $e = \hat{e}_1 + \hat{e}_2$ μπορεί η τιμή να είναι τόσο (σχετικά) μεγάλη που να επανορθώνει την τελική τιμή ώστε το υπολογισμένο $(a + b) + c + e$ να είναι ακριβέστερο του υπολογισμένου $(a + b) + c$. Η ιδέα αυτή μπορεί να επεκταθεί με διάφορους τρόπους σε αθροίσματα περισσοτέρων τιμών.
Πλεονεκτήματα επανορθωμένης άθροισης Γενικά πιο ακριβής από την συνηθισμένη άθροιση.

Εμπρός σφάλμα (Knuth, Kahan)

φράσσεται ως

$$|\sum_{j=1}^{n} \xi_j - \mathrm{fl}\left(\sum_{j=1}^{n} \xi_j\right)| \leq (2\mathbf{u} + O(n\mathbf{u}^2)) \sum_{j=1}^{n} |\xi_j|.$$

Πίσω σφάλμα

Ο παράγοντας 1ης τάξης στο φράγμα δεν εξαρτάται από το n:

$$\mathrm{fl}(s) = \sum_{j=1}^{n}(1 + \mu_j)\xi_j, \quad \text{όπου} \quad |\mu_j| \leq 2\mathbf{u} + O(n\mathbf{u}^2)$$

Επανορθωμένη αριθμητική; (εφικτό όπως σκιαγραφήσαμε)

A Floating-Point Technique for Extending the Available Precision

T. J. DEKKER*

Received July 26, 1971

Abstract. A technique is described for expressing multilength floating-point arithmetic in terms of singlelength floating point arithmetic, i.e. the arithmetic for an available (say: single or double precision) floating-point number system. The basic algorithms are exact addition and multiplication of two singlelength floating-point numbers, delivering the result as a doublelength floating-point number. A straightforward application of the technique yields a set of algorithms for doublelength arithmetic which are given as ALGOL 60 procedures.

Let x and y be singlelength floating-point numbers and let

$$z = fl(x+y);$$

i.e. z is the result of a singlelength floating-point addition of x and y. Let zz be the correction term exactly satisfying

$$z+zz=x+y.$$

It will be shown that, under various conditions, zz can be obtained by the formula

$$zz=fl((x-z)+y).$$

Μερικές υλοποιήσεις

Κώδικας 1: Επανόρθωση στην άθροιση 2 αριθμών

```
1 function [s,e] = kahan_cs(a,b);
2 if (abs(a)< abs(b))
3    temp =a; a=b; b=temp;
4 end
5 s = a+b;
6 e = (a-s)+b;
```

Κώδικας 2: Επανόρθωση χωρίς ελέγχους στην άθροιση 2 αριθμών

```
1 function [s,e] = nocomp2sum(x,y);
2 % following Muller et al.
3 s = a+b;
4 a_dot = s-b; b_dot = s-a_dot;
5 da = a-a_dot; db = b-b_dot;
6 e = da + db;
```

Άθροιση $n \geq 3$ στοιχείων (ιδέα Pichat)

Κώδικας 3: Επανόρθωση στην άθροιση

```
1 function [s] = casc_sum(x);
2 s = x(1);
3 e = 0;
4 for i=2:length(x);
5     [s,c(i)] = kahan_cs(s,x(i));
6     e = e + c(i);
7 end
8 s = s+e;
```

Παρατήρηση: Εναλλακτικά μπορούμε να εφαρμόσουμε αναδρομικά τον ίδιο αλγόριθμο για την άθροιση των στοιχείων του `c`.

Άθροιση $n \geq 3$ στοιχείων (ιδέα Pichat)

Κώδικας 4: Επανόρθωση στην άθροιση

```
1 function [s] = casc_sum(x);
2 s = x(1);
3 e = 0;
4 for i=2:length(x);
5     [s,c(i)] = kahan_cs(s,x(i));
6     e = e + c(i);
7 end
8 s = s+e;
```

Παρατήρηση: Εναλλακτικά μπορούμε να εφαρμόσουμε αναδρομικά τον ίδιο αλγόριθμο για την άθροιση των στοιχείων του `c`.

ΠΑΡΑΔΕΙΓΜΑ: `x = 1/n*ones(n,1)`

n	sum(x)	casc_sum(x)	abs(1-sum(x))
10	9.999999999999999e-01	1	1.1102e-016
100	1.000000000000001e+00	1	6.6613e-016
10^4	9.999999999980838e-01	1	1.9162e-012

Εντολή Fused Multiply-Add (FMA) I

- Ορισμένοι επεξεργαστές περιέχουν εντολή FMA (Fused Multiply and Add).
- Πλεονέκτημα στο χρόνο εκτέλεσης: Εκτελεί την πράξη $z + x * y$ στον ίδιο περίπου χρόνο που απαιτεί η $x * y$ ή η $x + y$.
- Πλεονέκτημα στην ακρίβεια: συνεπάγεται μόνον ένα αριθμητικό σφάλμα στρογγύλευσης.

$$\mathrm{fl}(z + x * y) = (z + x * y)(1 + \delta), \quad |\delta| \leq \mathbf{u}.$$

αντί

$$\mathrm{fl}(z + \mathrm{fl}(xy)) = (z + xy(1 + \delta_1))(1 + \delta_2)$$

- Περιέχεται στο πρότυπο IEEE-754-2008.
- FMA3: (a,b,c) <- a + b*c· FMA4: (a,b,c,d) <- a + b*c
- Δείτε (M^+10), Wikipedia
- Αρχικά (1990) στο IBM RS-6000. Σήμερα στα Intel Haswell, AMD Piledriver, Bulldozer

Βιβλιογραφία I

J.-M. Muller et al.

Handbook of Floating-Point Arithmetic.

Birkhäuser Boston, 2010.

Ε. Γαλλόπουλος.

Επιστημονικός Υπολογισμός Ι.

Πανεπιστήμιο Πατρών, 2008.

Χρήση Έργου Τρίτων Ι

1. http://link.springer.com/article/10.1007%2FBF01397083 (βλ. σελ 7)

Σημείωμα Αναφοράς

``Επιστημονικός Υπολογισμός Ι'', Έκδοση: 1.0, Πάτρα 2013-2014.

Διαθέσιμο από τη δικτυακή διεύθυνση: https://eclass.upatras.gr/courses/CEID1096/

Τέλος Ενότητας

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης