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

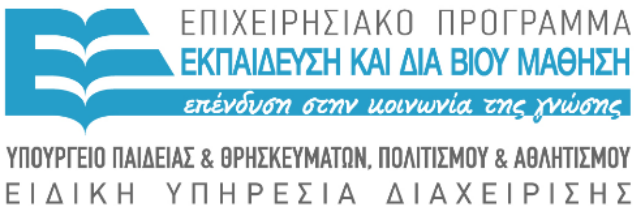
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

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

Υπενθύμιση

Αριθμητική κ.υ. και πρότυπο IEEE-754 (τάξη στο χάος, αλλά ....)

Ειδικοί αριθμοί και αναπαραστάσεις

Υποκανονικοποιημένοι αριθμοί

Χαρακτηριστικοί αριθμοί eps, `realmax`, `realmin`

Είδη στρογγύλευσης προς πλησιέστερο ζυγό, αποκοπή, ...

Αν $\tilde{\odot}$ είναι η υλοποίηση της αριθμητικής πράξης $\odot$,

κλειστό σύστημα: αν $x, y \in F \Rightarrow x \tilde{\odot} y \in F$.

ΑΑΣ (Αρχή ακριβούς (ή ορθής) στρογγύλευσης): Αν $x, y \in F \Rightarrow x \tilde{\odot} y = \mathrm{fl}(x \odot y)$.

Κλασική αναφορά (Gol91)

What Every Computer Scientist Should Know About Floating-Point Arithmetic

Note – This appendix is an edited reprint of the paper *What Every Computer Scientist Should Know About Floating-Point Arithmetic*, by David Goldberg, published in the March, 1991 issue of Computing Surveys. Copyright 1991, Association for Computing Machinery, Inc., reprinted by permission.

Abstract

Floating-point arithmetic is considered an esoteric subject by many people. This is rather surprising because floating-point is ubiquitous in computer systems. Almost every language has a floating-point datatype; computers from PCs to supercomputers have floating-point accelerators; most compilers will be called upon to compile floating-point algorithms from time to time; and virtually every operating system must respond to floating-point exceptions such as overflow. This paper presents a tutorial on those aspects of floating-point that have a direct impact on designers of computer systems. It begins with background on floating-point representation and rounding error, continues with a discussion of the IEEE floating-point standard, and concludes with numerous examples of how computer builders can better support floating-point.

Μελέτη σφαλμάτων: Το «σκηνικό» ((Gau97))

Ας θεωρήσουμε ότι το πρόβλημα αντιστοιχεί στον υπολογισμό της απεικόνισης $f : \mathcal{U} \subset \mathbb{R}^m \to \mathbb{R}^n$ για ορισμένα δεδομένα.

Προσοχή Αν τρέξουμε τον αλγόριθμο διακρίνουμε:

$x \in \mathcal{U}$ τα m στοιχεία στο πεδίο ορισμού της f·

$f(x)$ τα n στοιχεία της τιμής της συνάρτησης στο x, χωρίς λάθη υπολογισμών·

x^* τα στοιχεία εισόδου που χρησιμοποιούνται στην υλοποίηση, οπότε $x^* \in F$. Αν τα μόνα λάθη που υπάρχουν στο x^* είναι αυτά που προέρχονται από τη στρογγύλευση του x, τότε $x^* = \mathrm{fl}(x)$.

$f(x^*)$ η τιμή της $f(x^*)$, υπολογισμένη χωρίς σφάλματα (δηλ. με την αριθμητική του $\mathbb{R}$.

$f_{\mathrm{prog}}(x^*)$ η υλοποίηση του υπολογισμού της $f(x^*)$ με πεπερασμένο πλήθος διακριτών πράξεων α.κ.υ. στο $\mathcal{F}$.

Στόχος

Να εκτιμήσουμε το απόλυτο ή σχετικό σφάλμα

$$\|f_{\text{prog}}(x^*) - f(x)\| \text{ ή αν } f(x) \neq 0 \quad \frac{\|f_{\text{prog}}(x^*) - f(x)\|}{\|f(x)\|}.$$

Αν το σφάλμα είναι μικρό (π.χ. της τάξης του ϵ της μηχανής) ο αλγόριθμος, f, θεωρείται ακριβής.

Εμπρός Σφάλμα (forward error).

Το παραπάνω αποκαλείται (απόλυτο ή σχετικό) εμπρός[α΄] σφάλμα.

[α΄] Ο χαρακτηρισμός «εμπρός» θα εξηγηθεί στη συνέχεια των διαλέξεων.

ΠΡΟΣΟΧΗ Γενικά είναι ανέφικτο να υπολογιστεί. Στόχος είναι η εκτίμησή του, π.χ. βρίσκοντας σφικτά φράγματα.

Παρατηρήσεις

- Το $\|\cdot\|$ συμβολίζει κάποια μετρική - π.χ.
- ... απόλυτη τιμή, στους βαθμωτούς,
- ... μία από τις γνωστές νόρμες, για διανύσματα και μητρώα
- ... πίνακας απολύτων τιμών (*απόσταση κατά συνιστώσες*).
- Οι έννοιες του απόλυτου και σχετικού σφάλματος είναι γενικές και ξεπερνούν την α.κ.υ.
- Αν το x^* χρησιμοποιείται ως προσέγγιση του x, τότε

$$\text{ΑΣ: } \|x^* - x\| \quad \text{ΣΣ αν } x \neq 0 \quad \frac{\|x^* - x\|}{\|x\|}$$

- Μερικές φορές θέτουμε ΣΣ το $\frac{\|x^* - x\|}{\|x^*\|}$
- Αν $x = 0$ τότε αυτό το ΣΣ είναι πάντα 1 όσο καλή και να είναι η προσέγγιση x^* !!! (... το άλλο ΣΣ δεν ορίζεται)
- Για προσεγγίσεις πολύ μικρών τιμών χρησιμοποιείται το απόλυτο σφάλμα.
- Χονδρικά, αν $\frac{|x^* - x|}{|x|} \leq 5 \cdot 10^{-d}$ τότε λέμε ότι οι (αριθμοί) x^* και x συμφωνούν σε d δεκαδικά ψηφία.

Από το ϵ_M στο **u** – Μονάδα στρογγύλευσης

- Στη MATLAB για οποιοδήποτε α.κ.υ. `x`, η εντολή `eps(x)` επιστρέφει την απόστασή του από τον διαδοχικό του. Στη βιβλιογραφία αυτό αναφέρεται και ως ulp(x) (units in the last place).
- Αν χρησιμοποιούμε t bits για την αναπαράσταση της ουράς (το 1ο κρυμμένο) και $x = m \times \beta^e$ τότε η απόσταση του x από τον αμέσως επόμενο α.κ.υ. είναι $\text{ulp}(x) = (m + 2^{-(t-1)})2^e - m2^e = 2^{e+1-t}$.
- Η μέγιστη σχετική απόσταση είναι $\frac{2^{e+1-t}}{2^e} = 2^{1-t}$.
- ... προφανώς ίση με ϵ_M.

Θα εκφράζουμε τα σφάλματα ως πολλαπλάσια της μονάδας στρογγυλευσης που θα χρησιμεύσει ως μονάδα μέτρησης των σφαλμάτων.

Μονάδα στρογγύλευσης (unit roundoff)

Η μονάδα στρογγύλευσης είναι το μέγιστο δυνατό σχετικό σφάλμα *για τον επιλεγμένο τρόπο στρογγύλευσης*. Για *στρογγύλευση προς το πλησιέστερο*, $\mathbf{u} = \max_{z \neq 0} \frac{|z - \text{fl}(z)|}{|z|} = \frac{2^{1-t}}{2}$.

IEEE double: $\mathbf{u} = 1.1102\text{e}{-016}$. IEEE single: $\mathbf{u} = 5.9605\text{e}{-008}$.

Αρχή Ακριβούς Στρογγύλευσης → Διάδοση Σφαλμάτων

Αρχή ακριβούς στρογγύλευσης (υπενθύμιση)

Αν $\tilde{\odot}$ είναι η υλοποίηση της αριθμητικής πράξης $\odot$, τότε αν $x, y \in F$ ισχύει ότι $x\tilde{\odot}y = \mathrm{fl}(x \odot y) \in F$.

Το υπολογισμένο αποτέλεσμα είναι ακριβώς ίδιο με το να εκτελούνταν η πράξη με «θεϊκή» αριθμητική και μετά να εφαρμοζόταν στρογγύλευση.

ΕΠΟΜΕΝΩΣ

$$|x\tilde{\odot}y - x \odot y| \quad = \quad |\mathrm{fl}(x \odot y) - x \odot y| \leq \mathbf{u}|x \odot y|$$

άρα

$$-\mathbf{u}(x \odot y) \leq \mathrm{fl}(x \odot y) - x \odot y \leq \mathbf{u}(x \odot y)$$

$$(1 - \mathbf{u})(x \odot y) \leq x\tilde{\odot}y \leq (1 + \mathbf{u})(x \odot y)$$

Μοντέλο διάδοσης σφαλμάτων στρογγύλευσης

Μοντέλο διάδοσης

Μετά από κάθε αριθμητική πράξη $\odot$ επί δεδομένων α.κ.υ. $x, y \in F$ και δεν υπάρχει υπερ- ή υποχείλιση, ισχύει:

$$\boxed{\mathrm{fl}(x \odot y) = (1+\delta)(x \odot y), \text{ για } \delta \text{ τ.ώ. } |\delta| \leq \mathbf{u}}$$

Σχετικά με το δ:

- φράσσεται σε απόλυτη τιμή! Επομένως $-\mathbf{u} \leq \delta \leq \mathbf{u}$ (με πιο ενδελεχή ανάλυση $|\delta| < \mathbf{u}$.)
- διαφέρει για κάθε πράξη και στοιχεία x, y
- δεν απαιτείται να το γνωρίζουμε ακριβώς - χρησιμοποιούμε το μοντέλο γνωρίζοντας μόνον πώς φράσσεται το δ!
- .... το μοντέλο δεν προβλέπει πότε $\delta = 0$ (αυτή η πληροφορία μπορεί να είναι χρήσιμη, αλλά χάνεται).

Ιδιότητες αριθμητικών πράξεων στο πεδίο των πραγματικών $\mathbb{R}$

Αξιώματα πρόσθεσης

- A0 Το άθροισμα $x + y \in \mathbb{R}$.
- A1 Η πρόσθεση είναι αντιμεταθετική: $x + y = y + x$.
- A2 Η πρόσθεση είναι προσεταιριστική $x + (y + z) = (x + y) + z$.
- A3 Υπάρχει στοιχείο 0 ώστε $x + 0 = x$ για κάθε $x \in \mathbb{R}$.
- A4 Για κάθε $x \in \mathbb{R}$ υπάρχει αντίστροφο στοιχείο $-x \in \mathbb{R}$ ως προς την πρόσθεση, δηλ. $x + (-x) = 0$.

Αξιώματα πολλαπλασιασμού

- Π0 Το γινόμενο $x \times y \in \mathbb{R}$.
- Π1 Αντιμεταθετική ιδιότητα πολλαπλασιασμού: $x \times y = y \times x$.
- Π2 Προσεταιριστική ιδιότητα πολλαπλασιασμού: $x \times (y \times z) = (x \times y) \times z$.
- Π3 Υπάρχει στοιχείο 1 ώστε $x \times 1 = x$ για κάθε $x \in \mathbb{R}$.
- Π4 Για κάθε μη μηδενικό x υπάρχει αντίστροφο ως προς τον πολλαπλασιασμό $\frac{1}{x} \in \mathbb{R}$ ώστε $x \times \left(\frac{1}{x}\right) = 1$.

Ε Επιμεριστική ιδιότητα: $x \times (y + z) = x \times y + x \times z$.

Παράδειγμα: Δεν ισχύει πάντα το A2

$$t_1 = \mathrm{fl}(x + y) \quad s_1 = \mathrm{fl}(y + z)$$
$$t_2 = \mathrm{fl}(t_1 + z) \quad s_2 = \mathrm{fl}(x + s_1)$$

και δεν υπάρχει λόγος να ισχύει πάντα (αν και πολλές φορές ισχύει!)

$$\mathrm{fl}(\mathrm{fl}(x + y) + z) = \mathrm{fl}(x + \mathrm{fl}(y + z)).$$

Παράδειγμα: Δεν ισχύει πάντα το A2

Κώδικας 1: Παράδειγμα

```
1  (1+eps/2)+eps/2 % = 1
2  1+(eps/2+eps/2) % = 1.000000000000000
3  (-10^20+10^20)+1 % = 1
4  -10^20+(10^20+1) % = 0
```

Κώδικας 2: Παράδειγμα

```
1  a = 0.123456700000000;
2  b = 4.711325195312500e+008; c = -b;
3  (a+b)+c % = 0.123456716537476
4  a+(b+c) % = 0.123456700000000
```

Παράδειγμα: Δεν ισχύει πάντα το Π4

Γενικά για $x \in F, \mathrm{fl}(x \cdot \mathrm{fl}(\frac{1}{x})) \neq 1$.

Κώδικας 3: Παράδειγμα

```
1  index = []; for i=1:170
2      if ((1/i)*i ~= 1)
3          index = [index i];
4      end;
5  end;
6  index
7       49     98    103    107    161
```

Προειδοποιήσεις: Προσοχή για τους συγγραφείς μεταφραστών!

Επειδή στο $\mathcal{F}$ δεν ισχύουν όλες οι ιδιότητες πεδίου, δεν μπορούμε να κάνουμε τις ίδιες απλοποιήσεις και μετατροπές των αριθμητικών εκφράσεων που επιτρέπονται στο $\mathbb{R}$.
Π.χ. αν με δύο προσθετές που υπολογίζουν ταυτόχρονα το άθροισμα των α.κ.υ. x_1, x_2 και των x_3, x_4 το άθροισμα $x_1 + x_2 + x_3 + x_4$ μπορεί να υπολογισθεί ταχύτερα από το κλασικό $((x_1 \tilde{+} x_2) \tilde{+} x_3) \tilde{+} x_4$ χησιμοποιώντας $(x_1 \tilde{+} x_2) \tilde{+} (x_3 \tilde{+} x_4)$.

Το σφάλμα και τα αποτελέσματα μπορεί να είναι διαφορετικά!

Προειδοποίηση Οι επεξεργαστές Intel εκτελούν τις πράξεις σε εκτεταμένη διπλή ακρίβεια (80 bits) ... Αυτό ορισμένες φορές οδηγεί σε απρόσμενα αποτελέσματα (double rounding).

Βασικές σχέσεις: Υπενθύμιση

Στρογγύλευση: Από $x \in \mathbb{R}$ στο $\mathrm{fl}(x) \in F$

Αν $x \in G$ τότε μπορούμε να γράψουμε $\mathrm{fl}(x) = x(1+\delta)$ για κάποιο $|\delta| \leq \mathbf{u}$.

Σφάλμα πράξεων

Έστω ότι $x, y \in F$ και $x \odot y \in G$ και ότι $\odot \in \{\pm, \times, , \sqrt{}\}$. Τότε

$$\mathrm{fl}(x \odot y) \quad = \quad (x \odot y)(1+\delta), \text{ για κάποιο } |\delta| < \mathbf{u}$$

και

$$\mathrm{fl}(x \odot y) \quad = \quad \frac{x \odot y}{1+\delta}, \text{ για κάποιο } |\delta| \leq \mathbf{u}$$

ΠΗΓΗ ΠΡΟΒΛΗΜΑΤΩΝ ΣΤΙΣ ±ΣΕΙΣ: Καταστροφική απαλοιφή

Προσοχή Όταν αφαιρούνται δύο αριθμοί που είναι σχεδόν ίσοι και περιέχουν μικρά σφάλματα `αναδύονται σκουπίδια' ακόμα και αν χρησιμοποιούσαμε αριθμητική άπειρης ακρίβειας!
Αν $A = A_1 + \text{θόρυβος}$, $A' = A_2 + \text{θόρυβος}$ και οι τιμές A_1, A_2 είναι σχεδόν ίσες και θέλουμε να υπολογίσουμε το (μικρό) $A_1 - A_2$ αφαιρώντας $A - A'$, τότε:

$$\boxed{(A_1 + \text{θόρυβος}) - (A_2 + \text{θόρυβος}) = (A_1 - A_2) + \text{θόρυβος}}$$

- καταστροφική απαλοιφή όταν $|A_1 - A_2| = O(\text{θόρυβος})$ ή λιγότερο.
- Η μόλυνση από τα σκουπίδια επηρεάζει και έχει καταστροφικά αποτελέσματα αν, για παράδειγμα, τα «ασήμαντα σκουπίδια» πολλαπλασιαστούν με μεγάλους αριθμούς και χρησιμοποιηθούν περαιτέρω.
- Διαβάστε την ιστορία Catastrophic cancellation in the high seas της A. Langville (Lan01).

«Προς τα εμπρός ανάλυση» και εκτίμηση σφάλματος

Παράδειγμα

Θεωρούμε ότι οι μεταβλητές x_j περιέχουν α.κ.υ.

$$\begin{aligned}(x_1 \tilde{+} x_2) \tilde{+} x_3 &= ((x_1 + x_2)(1+\delta_1) + x_3)(1+\delta_2) \\ &= x_1(1+\delta_1)(1+\delta_2) + x_2(1+\delta_1)(1+\delta_2) + x_3(1+\delta_2)\end{aligned}$$

επομένως αν θέσουμε

$$E := (x_1 \tilde{+} x_2) \tilde{+} x_3 - (x_1 + x_2 + x_3)$$

$$E = (x_1 + x_2)(\delta_1 + \delta_2 + \delta_1\delta_2) + x_3\delta_2$$

επομένως μπορούμε να φράξουμε ως εξής:

$$|E| \leq (|x_1| + |x_2|)(2\mathbf{u} + \mathbf{u}^2) + |x_3|\mathbf{u}.$$

Με τον ίδιο τρόπο καταλήγουμε και σε άνω φράγμα για το σφάλμα στον υπολογισμό του $x_1 + (x_2 + x_3)$. Θέτουμε για συντομία
$\hat{E} := x_1 \tilde{+} (x_2 \tilde{+} x_3) - (x_1 + x_2 + x_3)$.
Εντέλει έχουμε τα παρακάτω άνω φράγματα για τους δύο εναλλακτικούς τρόπους υπολογισμού:

$$\begin{aligned} |E| &\leq (|x_1| + |x_2|)(2\mathbf{u} + \mathbf{u}^2) + |x_3|\mathbf{u} \\ |\hat{E}| &\leq |x_1|\mathbf{u} + (2\mathbf{u} + \mathbf{u}^2)(|x_2| + |x_3|). \end{aligned}$$

- Αναδεικνύεται λεπτομερώς η (διαφορετική) συνεισφορά του κάθε όρου στο άνω φράγμα.
- Διαφαίνεται ότι με βάση τα x_j, θα μπορούσαμε να επιλέξουμε σειρά άθροισης που να ελαχιστοποιεί το άνω φράγμα (αλλά όχι κατ' ανάγκη το σφάλμα!).

Ενδιαφέροντα συμπεράσματα

Πώς (με ποια σειρά) αθροίζουμε 3 α.κ.υ; Η παρακάτω συζήτηση θα γίνει αποκλειστικά με βάση τα παραπάνω. Περισσότερα σε επόμενη διάλεξη αφιερωμένη στην άθροιση.

Υπάρχουν 3 μη ισοδύναμοι αριθμητικά τρόποι αθροισης (δεν ισχύει προσεταιριστικότητα):

$$(x_1 + x_2) + x_3, x_1 + (x_2 + x_3), (x_1 + x_3) + x_2$$

Οι υπόλοιποι (9) τρόποι είναι αριθμητικά ισοδύναμοι με έναν απο τους παραπάνω (γιατί ισχύει αντιμεταθετικότητα), π.χ.

$$x_3 \tilde{+} (x_1 \tilde{+} x_2), x_3 \tilde{+} (x_2 \tilde{+} x_1), (x_2 \tilde{+} x_3) \tilde{+} x_1,$$
$$(x_3 \tilde{+} x_2) \tilde{+} x_1, x_2 \tilde{+} (x_1 \tilde{+} x_3), x_2 \tilde{+} (x_3 \tilde{+} x_1),$$
$$x_1 \tilde{+} (x_3 \tilde{+} x_2), (x_3 \tilde{+} x_1) \tilde{+} x_2, (x_2 \tilde{+} x_1) \tilde{+} x_3,$$

Με ποιόν τρόπο προκύπτει το μικρότερο άνω φράγμα; Με βάση τα παραπάνω, αυτός που αφήνει το μεγαλύτερο στοιχείο τελευταίο (εκτός παρένθεσης):

- Αν $|x_3| = \max(|x_1|, |x_2|, |x_3|)$ τότε $\max_{x_j \in F} |E| \leq \max_{x_j \in F} |\hat{E}|$
- Αν $|x_1| = \max(|x_1|, |x_2|, |x_3|)$ τότε $\max_{x_j \in F} |\hat{E}| \leq \max_{x_j \in F} |E|$
- Αν $|x_2| = \max(|x_1|, |x_2|, |x_3|)$ τότε επιλέγουμε $(x_1 \tilde{+} x_3) \tilde{+} x_2$.

- Όμως: Το άνω φράγμα δείχνει την χειρότερη περίπτωση!
- Τα πράγματα μπορεί να είναι πολύ καλυτερα!!
- ... για παράδειγμα αν γνωρίζουμε ότι $x_1 = -x_2$, ή/και το αντίστοιχο δ να είναι 0.
- Αυτά δεν προβλέπονται αν εφαρμόσουμε το μοντέλο χωρίς ειδικές τροποποιήσεις.

Χαλαρώνοντας λίγο!

$$\begin{aligned} |E| &\leq (|x_1| + |x_2|)(2\mathbf{u} + \mathbf{u}^2) + |x_3|\mathbf{u} \\ &\leq (|x_1| + |x_2| + |x_3|)2\mathbf{u} + (|x_1| + |x_2|) \underbrace{\mathbf{u}^2}_{\text{αμελητέο}} \end{aligned}$$

αν αγνοήσουμε όρους με παράγοντα $\mathbf{u}^2$ (δηλ. όρους 2ης τάξης) μπορούμε να φράξουμε αμφότερους όρους $|E|, |\hat{E}|$

$$|E|, |\hat{E}| \leq (|x_1| + |x_2| + |x_3|)2\mathbf{u}.$$

ΠΡΟΣΟΧΗ Από τα παραπάνω και μόνον το σχετικό σφάλμα φράσσεται ως εξής (ομοίως και για το $|\hat{E}|$):

$$\frac{|E|}{|x_1 + x_2 + x_3|} \leq \frac{|x_1| + |x_2| + |x_3|}{|x_1 + x_2 + x_3|} 2\mathbf{u}$$

ΕΡΩΤΗΣΗ: Ποιο είναι το μειονέκτημα αυτού του φράγματος;

Ποιο είναι το μειονέκτημα αυτού του φράγματος;

- Το φράγμα εξαρτάται από τις τιμές των x_1, x_2, x_3 ...
- Ο όρος $\frac{|x_1|+|x_2|+|x_3|}{|x_1+x_2+x_3|}$ μπορεί να γίνει πολύ μεγάλος.

Αν υπολογίσουμε (μη μηδενικό) $x_1 \tilde{\times} x_2 \tilde{\times} x_3$, τότε

$$\begin{aligned}
\frac{|(x_1 \tilde{\times} x_2) \tilde{\times} x_3 - x_1 x_2 x_3|}{|x_1 x_2 x_3|} &= \frac{|(x_1 x_2)(\delta_1) x_3 (1+\delta_2) - x_1 x_2 x_3|}{|x_1 x_2 x_3|} \\
&= \frac{|x_1 x_2 x_3 (\delta_1 + \delta_2 + \delta_1 \delta_2)|}{|x_1 x_2 x_3|} \\
\frac{|(x_1 \tilde{\times} x_2) \tilde{\times} x_3 - x_1 x_2 x_3|}{|x_1 x_2 x_3|} &\leq 2\mathbf{u} + \mathbf{u}^2
\end{aligned}$$

Σημαντική ειδική περίπτωση

Όταν οι όροι x_j είναι ομόσημοι, $|x_1| + |x_2| + |x_3| = |x_1 + x_2 + x_3|$, επομένως

$$\frac{|E|}{|x_1 + x_2 + x_3|} \leq 2\mathbf{u}.$$

Παραδείγματα:

- άθροισμα ομόσημων όρων (π.χ. μη αρνητικών στατιστικών μεγεθών, μετρήσεων, ...)
- υπολογισμός νόρμας πραγματικών διανυσμάτων

Γενικεύσεις

Εύρεση άνω φράγματος του εμπρός σφάλματος για τον υπολογισμό των

- ... $(x_1 + x_2) + x_3$ όταν $x_j \in G$ και όχι κατ' ανάγκη α.κ.υ.
- ... αθροίσματος n αριθμών (για εναλλακτικούς τρόπους άθροισης)
- ... εσωτερικού γινομένου διανυσμάτων ...

Π.χ. στην πρώτη περίπτωση

$$\begin{aligned} \mathrm{fl}((x_1 + x_2) + x_3) &= \mathrm{fl}(\mathrm{fl}(\mathrm{fl}(x_1) + \mathrm{fl}(x_2)) + \mathrm{fl}(x_3)) \\ &= ((x_1(1+\delta_1) + x_2(1+\delta_2))(1+\delta_3) + (1+\delta_4)x_3)(1+\delta_5) \end{aligned}$$

Το παραπάνω μπορεί να ξαναγραφτεί ως:

$$x_1(1+\delta_1)(1+\delta_3)(1+\delta_5) + x_2(1+\delta_2)(1+\delta_3)(1+\delta_5) + x_3(1+\delta_4)(1+\delta_5)$$

Οι εκφράσεις γίνονται πολύπλοκες ακόμα και στην απλή αυτή μελέτη!

Πολύ χρήσιμο εργαλείο

Για την απλοποίηση όρων όπως $p_n = \prod_{i=1}^{n}(1+\delta_i)$ όταν γνωρίζουμε ότι $|\delta_i| \leq u$. Αμέσως βλέπουμε ότι

$$(1-u)^n \leq p_n \leq (1+u)^n.$$

και ότι $p_n = 1 + nu + O(u^2)$.

Ακόμα καλύτερα:

Λήμμα

Αν $|\delta_i| \leq u$ και $\rho_i = \pm 1$ για $i = 1 : n$ και $nu < 1$ τότε υπάρχει κάποια τιμή θ_n τ.ώ.

$$\prod_{i=1}^{n}(1+\delta_i)^{\rho_i} = 1 + \theta_n,$$

όπου

$$|\theta_n| \leq \frac{nu}{1-nu} := \gamma_n.$$

Εφαρμογή

Εύρεση άνω φράγματος του εμπρός σφάλματος για τον υπολογισμό του

- $(x_1 + x_2) + x_3$ όταν $x_j \in G$ και όχι κατ' ανάγκη α.κ.υ.

Με βάση το Λήμμα, υπάρχουν $\theta_3, \zeta_3, \theta_2$ τέτοια ώστε

$$|\theta_3| \leq \gamma_3, |\zeta_3| \leq \gamma_3, |\theta_2| \leq \gamma_2$$

και μπορούμε να γράψουμε

$$\mathrm{fl}((x_1 + x_2) + x_3) \quad = \quad x_1(1 + \theta_3) + x_2(1 + \zeta_3) + x_3(1 + \theta_2)$$

επομένως ένα (χαλαρό) άνω φράγμα για το σφάλμα θα είναι

$$\begin{aligned} |\mathrm{fl}((x_1 + x_2) + x_3) - (x_1 + x_2 + x_3)| &\leq |x_1|\gamma_3 + |x_2|\gamma_3 + |x_3|\gamma_2 \\ &\leq (|x_1| + |x_2| + |x_3|)\gamma_3. \end{aligned}$$

Για ομόσημα x_j, το σχετικό σφάλμα φράσσεται άμεσα από

$$\frac{|\mathrm{fl}((x_1 + x_2) + x_3) - (x_1 + x_2 + x_3)|}{|x_1 + x_2 + x_3|} \leq \frac{(|x_1| + |x_2| + |x_3|)}{|x_1 + x_2 + x_3|}\gamma_3$$

όπου $\gamma_3 = \frac{3\mathbf{u}}{1-3\mathbf{u}}$.

Κριτική

Εκκινώντας από τα στοιχεία εισόδου και παρακολουθώντας το σφάλμα σε κάθε πράξη προσπαθούμε να φράξουμε το μέγιστο απόλυτο ή σχετικό σφάλμα που θα μπορούσε να προκύψει στο τελικό αποτέλεσμα.

Η ιδέα είναι απλή

- ... η εφαρμογή της μπορεί να είναι περίπλοκη
- ... σκληρή άσκηση σε ανισότητες
- ... τεράστιες εκφράσεις, κ.λπ.

Παρατηρήσεις

Έχουν γίνει πολλές προσπάθειες για την αυτοματοποίηση της ανάλυσης των σφαλμάτων που υπεισέρχονται στις υπολογιστικές διαδικασίες με μέτρια ή μεγαλύτερη επιτυχία.

Εμπρός ανάλυση σφάλματος: Θεωρούμε τον αλγόριθμο ως μιά σειρά στοιχειωδών πράξεων. Σε κάθε βήμα, υπολογίζεται μία τιμή α_{k+1} με βάση προηγούμενες τιμές και στοιχεία εισόδου, π.χ. $\alpha_{k+1} = g_k(\alpha_1, ..., \alpha_k)$. Μερικές από τις τιμές μπορεί να είναι δεδομένα εισόδου. Στη συνέχεια, υπολογίζουμε φράγματα για τα σφάλματα στα τελικά αποτελέσματα.

Παρατηρήσεις

Έχουν γίνει πολλές προσπάθειες για την αυτοματοποίηση της ανάλυσης των σφαλμάτων που υπεισέρχονται στις υπολογιστικές διαδικασίες με μέτρια ή μεγαλύτερη επιτυχία.

Εμπρός ανάλυση σφάλματος: Θεωρούμε τον αλγόριθμο ως μιά σειρά στοιχειωδών πράξεων. Σε κάθε βήμα, υπολογίζεται μία τιμή α_{k+1} με βάση προηγούμενες τιμές και στοιχεία εισόδου, π.χ. $\alpha_{k+1} = g_k(\alpha_1, ..., \alpha_k)$. Μερικές από τις τιμές μπορεί να είναι δεδομένα εισόδου. Στη συνέχεια, υπολογίζουμε φράγματα για τα σφάλματα στα τελικά αποτελέσματα.

Ανάλυση διαστημάτων: Θεωρούμε ότι κάθε δεδομένο x εγκλείεται σε κάποιο διάστημα $[x_L, x_U)$ οπότε οι πράξεις επί των δεδομένων εκτελούνται χρησιμοποιώντας «αριθμητική διαστημάτων» (interval arithmetic).

Παρατηρήσεις

Έχουν γίνει πολλές προσπάθειες για την αυτοματοποίηση της ανάλυσης των σφαλμάτων που υπεισέρχονται στις υπολογιστικές διαδικασίες με μέτρια ή μεγαλύτερη επιτυχία.

Εμπρός ανάλυση σφάλματος: Θεωρούμε τον αλγόριθμο ως μιά σειρά στοιχειωδών πράξεων. Σε κάθε βήμα, υπολογίζεται μία τιμή α_{k+1} με βάση προηγούμενες τιμές και στοιχεία εισόδου, π.χ. $\alpha_{k+1} = g_k(\alpha_1, ..., \alpha_k)$. Μερικές από τις τιμές μπορεί να είναι δεδομένα εισόδου. Στη συνέχεια, υπολογίζουμε φράγματα για τα σφάλματα στα τελικά αποτελέσματα.

Ανάλυση διαστημάτων: Θεωρούμε ότι κάθε δεδομένο x εγκλείεται σε κάποιο διάστημα $[x_L, x_U)$ οπότε οι πράξεις επί των δεδομένων εκτελούνται χρησιμοποιώντας «αριθμητική διαστημάτων» (interval arithmetic).

Πίσω ανάλυση σφάλματος: Στη συνέχεια!

Βιβλιογραφία I

W. Gautschi.

Numerical Analysis: An Introduction.

Birkhauser, Boston, 1997.

D. Goldberg.

What every computer scientist should know about floating point arithmetic.

ACM Comput. Surveys, pages 5--48, 1991.

A.N. Langville.

Catastrophic cancellation on the high seas.

The Pi Mu Epsilon Journal, 11(4):205--208, 2001.

http://www4.ncsu.edu/~ipsen/ma798I/langville.pdf.

Ε. Γαλλόπουλος.

Επιστημονικός Υπολογισμός Ι.

Πανεπιστήμιο Πατρών, 2008.

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