Επιστημονικός Υπολογισμός Ι
ΗΥ 343: ΔΙΑΛΕΞΗ 6

Ε. Γαλλόπουλος
Τμήμα Η/Υ & Πληροφορικής
Πανεπιστήμιο Πατρών

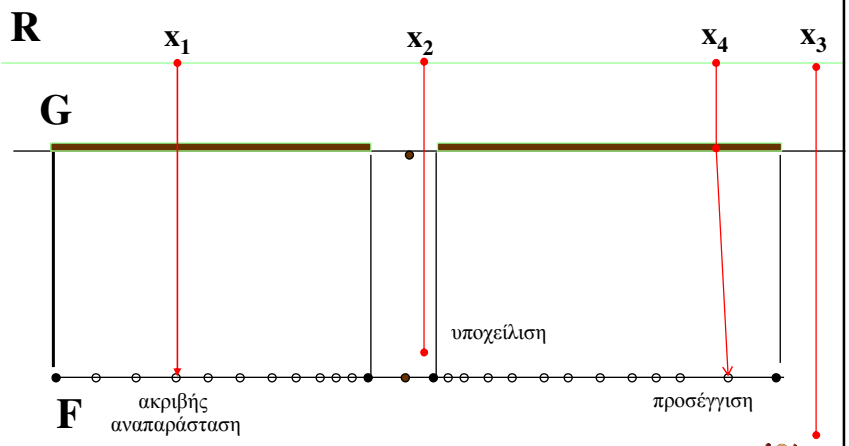
- **Οι α.κ.υ. είναι ένα (μικροσκοπικό) πεπερασμένο υποσύνολο των πραγματικών.**
- **Πρέπει να τους χρησιμοποιήσουμε για να αναπαραστήσουμε όλους τους αριθμούς!**
- **προσεγγίσεις - quantization**

ΤΜΗΥΠ
Πανεπιστήμιο Πατρών

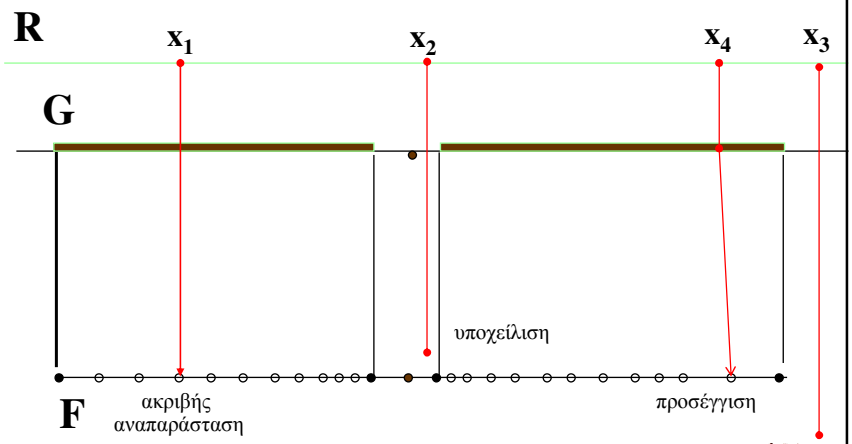
Απεικόνιση fl: $R \rightarrow F$

- συνάρτηση fl: R → F
- Τρείς περιπτώσεις για το y := fl(x)
  - x ∈ F
    ⇒ y = fl(x)
  - x ∉ G
    ⇒ υπερχείλιση ή υποχείλιση
  - x ∈ G και x ∉F
    ⇒ προσέγγιση του x με στοιχείο fl(x) ∈ F

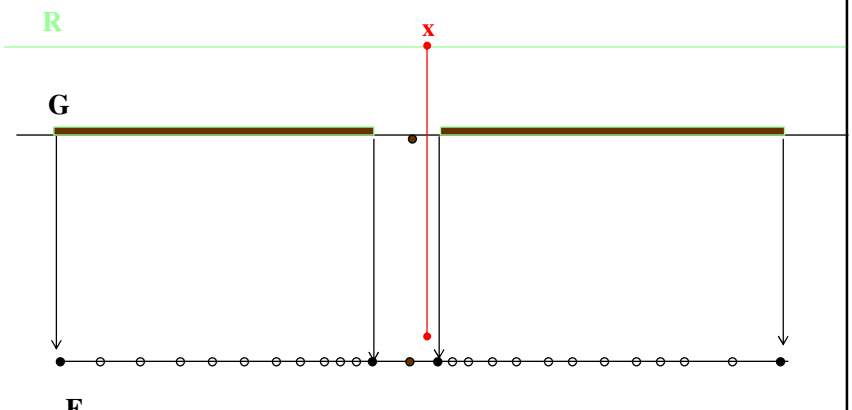
Απεικόνιση πραγματικών σε ακυ



Απεικόνιση πραγματικών σε ακυ



Απεικόνιση πραγματικών σε ακυ

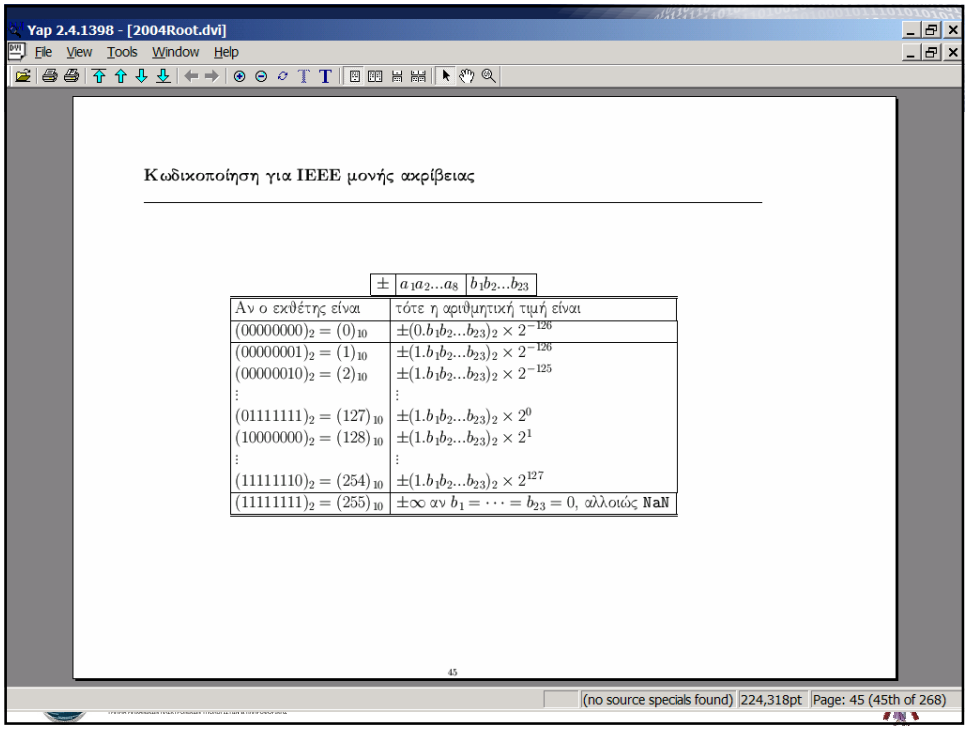


Ειδικοί αριθμοί στο πρότυπο IEEE

- **± Inf (infinity), -0 (μείον μηδέν),**
- **NaN (Not a Number) για αποτέλεσμα άκυρων πράξεων, π.χ. 0/0, $0 \times \infty$, κ.λπ.**

- **Επιτρέπουν να διατηρήσουμε το σύστημα των α.κ.υ. «κλειστό»:**
  - **Στο σύστημα IEEE, κάθε πράξη με α.κ.υ. οδηγεί σε μέλος του συστήματος α.κ.υ.:**
    - **κανονικό ή ειδικό αριθμό του συστήματος α.κ.υ.**

ΤΜΗΥΠ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ & ΠΛΗΡΟΦΟΡΙΚΗΣ
Πανεπιστήμιο Πατρών



Μη κανονικοποιημένοι α.κ.υ.

- Συνήθως οι α.κ.υ. IEEE είναι κανονικοποιημένοι
- Το φάσμα των αναπαραστήσιμων αριθμών της IEEE επεκτείνεται προς τους πολύ μικρούς (κοντά στο 0) με «μη κανονικοιημένους αριθμοί»
  - Πρόκειται για τους α.κ.υ. που έχουν πρώτο ψηφίο 0 (μηδέν), δηλ.
  - $0.* * * \text{---} * \times 2^{e_{min}}$
  - Δηλ.$\{(0.0\text{---}1)_2, (0.0\text{---}10)_2, ..., (0.1\text{----}1)_2\} \times 2^{e_{min}}$
  - Επομένως, ο ελάχιστος (μη κανονικοποιημένος) θετικός αναπαραστήσιμος α.κ.υ. θα είναι ο

$$0.0\text{---}1 \times 2^{e_{min}}$$

  - Σε α.κ.υ. IEEE διπλής ακρίβειας, $\approx 4.9407 \times 10^{-324}$
  - Αν το διαιρέσουμε με y>1 επιστρέφεται 0.

Αριθμητικές πράξεις

- Ένα σύστημα α.κ.υ. δεν πρέπει να εξασφαλίζει μόνον την *αναπαράσταση* ...
- .... αλλά και τις *αριθμητικές πράξεις* μεταξύ των μελών του.
- Συμβαίνει αριθμητική εξαίρεση όταν το αποτέλεσμα είναι **NaN** ή $\pm\infty$ ή αν

$$fl(x \odot y) \neq x \odot y \text{ ή}$$

- Στην IEEE ορίζεται λεπτομερώς ποιές είναι οι εξαιρέσεις και τι γίνεται όταν προοκύπτουν.

Αριθμητικές εξαιρέσεις (exceptions)

Εξαίρεση	*παράδειγμα*	*αποτέλεσμα*
Invalid op	**0/0, 0 × Inf**	**NaN**
Overflow		**± Inf ή ± N_{max}**
Divide by 0	**Πεπερασμένος αριθμ./0**	**± Inf**
Underflow		**± 0, ± N_{min} ή subnormal**
Inexact	**fl(x ⊙ y) ≠ x ⊙ y**	**στρογγύλευση**

N_{min}, N_{max} είναι οι κανονικοποιημένοι ελάχιστοι και μέγιστοι α.κ.υ.

ΠΡΟΣΟΧΗ: Ο έλεγχος (NaN==NaN) επιστρέφει 0
Το NaN είναι η μόνη «ποσότητα» που έχει αυτή τη συμπεριφορά.

Παραδείγματα

- **Στη MATLAB: προσέξτε τις παρακάτω εντολές**

```
>> realmin
ans =
   2.225073858507201e-308
>> format hex
>> ans
ans =
   0010000000000000
>> realmin/2^51
ans =
   0000000000000002
>> realmin/2^52
ans =
   0000000000000001
>> ans/2
ans =
   0000000000000000
>> format long e
>> realmin/2^52
ans =
   4.940656458412465e-324
```

```
realmax
ans =
  7fefffffffffffff
>> format long e
>> ans
ans =
   1.797693134862316e+308
>> ans*2
ans =
   Inf

Παρόλα αυτά, δείτε το περίεργο
>> realmax+100
ans =
   1.797693134862316e+308
%% γιατί!
```

Ερώτημα: Γιατί στην α.κ.υ. IEEE διπλής ακρίβειας (default της MATLAB) δεν ισχύει ότι

```
z = 100; realmax+z = Inf,
```

Σημειώνουμε ότι το ίδιο ισχύει κι αν προσθέσουμε μεγαλύτερα z (όχι μόνο 100).
Ανάλυση: Προσέξτε ότι σε 16δική αναπαράσταση

```
>> realmax
ans = 7fefffffffffffff
>> realmax+100
ans = 7fefffffffffffff
```

δηλαδή δεν υπάρχει καμμιά αλλαγή. Αυτό οφείλεται στον τρόπο που υλοποιείται η άθροιση α.κ.υ.

Γιατί; Στην αριθμητική κινητής υποδιαστολής, ένα από τα αρχικά βήματα είναι α) η σύγκριση των εκθετών και αύξηση του μικρότερου μέχρι να εξισωθεί με το μεγαλύτερο και β) ισάριθμους υποδιπλασιαμούς της ουράς του.

Ενδέχεται, κατά τη διάρκεια της διαδικασίας, να μηδενιστεί η ουρά του μικρότερου αριθμού και να μην επιδράσει στην άθροιση:

$$x = m_x \times 2^{e_x},\ y = m_y \times 2^{e_y},\ \text{και } x >= y \text{ τότε το } x+y \text{ υπολογίζεται «περίπου» ως εξής}$$

$$x + y = (m_x + (m_y \times 2^{-(e_x - e_y)})) \times 2^{e_x}$$

επομένως αν e_x – e_y >= t (μήκος ουράς) ο όρος $m_y \times 2^{-(e_x-e_y)}$ μηδενίζεται και δεν επιδρά στο αποτέλεσμα.

ΠΡΟΣΟΧΗ: η παραπάνω απορρόφηση συμβαίνει όταν οι αριθμοί που πρέπει να προστεθούν διαφέρουν πάρα πολύ σε μέγεθος εκθέτη π.χ. `1+(1/realmax) =1`.

Ερώτηση: Ποιά διαφορά τάξης μπορεί να δημιουργήσει τέτοιο πρόβλημα στο σύστημα IEEE με διπλή ακρίβεια (t=53);
Η ουρά μηδενίζεται όταν το πρώτο bit (hidden) που αντιστοιχεί στο 1. διολισθίσει δεξιά κατά 53 δυαδικές θέσεις.

Επομένως αρκεί οι εκθέτες να διαφέρουν κατά 2^{53} .

Παράδειγμα (MATLAB):

```
2^52
ans = 4.503599627370496e+015
>> 1+ans
ans = 4.503599627370497e+015
>> 2^53
ans = 9.007199254740992e+015
>> 1+ans
ans = 9.007199254740992e+015
```

Δηλ. $2^{53}+1=2^{53}$.
Προσέξτε επίσης ότι 2^{-52} = `eps`.

Είδη στρογγύλευσης

- Πώς λειτουργεί το **fl(.)** όταν **x** $\in$ **G** αλλά **x** $\notin$ **F;**
- **ΑΡΧΗ: Ο αριθμός που θα χρησιμοποιήσουμε εξαρτάται από τη σχέση του x με τους α.κ.υ. που ορίζουν το ελάχιστο διάστημα που το εγκλείει.**
- **Προς τον πλησιέστερο «ζυγό»:**

$$y = fl(x) \text{ όπου } y = \arg\min_{y^* \in F} |y^* - x|$$

  - **Αν το x ισαπέχει από τα άκρα του διαστήματος, τότε στρογγυλεύουμε προς τον αριθμό που έχει 0 ως τελικό ψηφίο στην ουρά.**
- **Αποκοπή (στρογγύλευση προς 0)**
- **Κατευθυνόμενη (χρήσιμη σε ορισμένες εφαρμογές, π.χ. Αριθμητική διαστημάτων):**
  - **Προς $+\infty$**
  - **Προς $-\infty$**

- **Η απόσταση διαδοχικών κανονικοποιημένων α.κ.υ. για δεδομένο εκθέτη είναι το μήκος του διαστήματος που τα χωρίζει:**

$$d = \beta^{-(t-1)} \times \beta^{e}$$

Επομένως,

$$\max |fl(x)-x| = d/2 = \beta^{-(t-1)} \times \beta^{e} / 2$$

σε δυαδικό σύστημα:

$$\max |fl(x)-x| = 2^{e-t}$$

Σφάλμα στρογγύλευσης

- **Το |fl(x)-x| αποκαλείται**
  ***απόλυτο σφάλμα στρογγύλευσης*.**
- **Το απόλυτο σφάλμα στρογγύλευσης *μεγιστοποιείται* (τοπικά) όταν το «υπό στρογγύλευση» $x \in R$ εξαρτάται από το είδος στρογγύλεθσης που χρησιμοποείται. Έστω ότι $x^- < x < x^+$, όπου $x^-, x^+ \in F$ είναι οι α.κ.υ. που περιβάλλουν το x. Τότε το απόλυτο σφάλμα στρογγύλευσης μεγιστοποιείται όταν το x κείται:**
  - **α) στο μέσο του διαστήματος, αν χρησιμοποιούμε στρογγύλευση προς το πλησιέστερο (το "default" στην α.κ.υ. IEEE).**
  - **β) αμέσως πριν το «δεξιό» άκρο του διαστήματος, αν έχουμε αποκοπή θετικού.**
  - **γ) αμέσως μετά το «αριστερό» άκρο του διαστήματος αν έχουμε στρογγύλευση προς το $+\infty$**
  - **προς τον πλησιέστερο θα είναι ίσο με το μέγεθος του διαστήματος Έτσι υπολογίσαμε το μέγιστο δυνατό σφάλμα στρογγύλευσης.**
- **Το μέγεθος του απολύτου σφάλματος αλλάζει με τον εκθέτη που αντιστοιχεί στην κανονικοποιημένη αναπαράσταση του x.**
- **Συνήθως «αποκλιμακώνουμε» και μελετάμε το *σχετικό σφάλμα στρογγύλευσης:***

$$rel(x) = |fl(x)-x|/|x|$$

Παρατηρήσεις

- **Αξίζει να προσέξετε από τώρα ότι παρόλο που ενδιαφερόμαστε για τα σφάλματα, τις περισσότερες περιπτώσεις δεν μπορούμε να τα υπολογίσουμε ακριβώς!**
- **... αν μπορούσαμε, δεν θα είχαμε αντικείμενο συζήτησης, καθώς θα μπορούσαμε να προσθέσουμε την υπολογισμένη τιμή στο σφάλμα και να λάβουμε την ακριβή τιμή (που γενικά είναι άγνωστη!)**

Επομένως «χαλαρώνουμε» και όταν αναφερόμαστε στην «εύρεση του σφάλματος» συνήθως εννοούμε την εύρεση «καλού άνω φράγματος» για το μέγεθος του σφάλματος

- **... αντί τον ακριβή υπολογισμό του σφάλματος, που γενικά είναι αδύνατος,**
- **... όπως παραπάνω, που διερευνήσαμε ποιο μπορεί να είναι το μέγιστο σφάλμα στρογγύλευσης αν $x \in (x^-, x^+)$ και πότε μεγιστοποιείται.**

Μονάδα στρογγύλευσης

- Επειδή τα σφάλματα είναι συνήθως μικρά, τα μετράμε με βάση τη «μονάδα στρογγύλευσης» που αναδεικνύει τη διακριτότητα της αναπαράστασης.
- Η μονάδα στρογγύλευσης είναι το μέγιστο σχετικό σφάλμα στρογγύλευσης (θεωρούμε ότι ο αριθμός δεν είναι **0**).

- Την ορίζουμε *βάσει του συγκεκριμένου συστήματος α.κ.υ.* και των επιλογών που έχουν γίνει όσον αφορά
  - **1)** στην αναπαράσταση
  - **2)** στο είδος στρογγύλευσης που χρησιμοποιείται.

- Αν π.χ. χρησιμοποιείται η αναπαράσταση **1.**μ_1 **...** μ_{t-1} β^e και *στρογγύλευση προς το πλησιέστερο,*

- **Το u** αποκαλείται «μονάδα στρογγύλευσης» **(unit roundoff)**
  - χρησιμοποιείται εκτενέστατα ως μονάδα μέτρησης σφαλμάτων.

$$\max \mathrm{rel}(x) = \frac{\beta^{e-t+1}}{2\beta^e} = \frac{\beta^{-t+1}}{2} := \mathbf{u}$$

Πανεπιστήμιο Πατρών

Μονάδα στρογγύλευσης

- Επειδή τα σφάλματα είναι συνήθως μικρά, τα μετράμε με βάση τη «μονάδα στρογγύλευσης» που αναδεικνύει τη διακριτότητα της αναπαράστασης.
- Η μονάδα στρογγύλευσης είναι το μέγιστο σχετικό σφάλμα στρογγύλευσης (θεωρούμε ότι ο αριθμός δεν είναι **0**).

- Την ορίζουμε *βάσει του συγκεκριμένου συστήματος α.κ.υ.* και των επιλογών που έχουν γίνει όσον αφορά
  - **1)** στην αναπαράσταση
  - **2)** στο είδος στρογγύλευσης που χρησιμοποιείται.

- Αν π.χ. χρησιμοποιείται η αναπαράσταση **1.**μ_1 **...** μ_{t-1} β^e και *στρογγύλευση προς το πλησιέστερο,*

$$\max \mathrm{rel}(x) = \max_{x \neq 0} \frac{|\mathrm{fl}(x) - x|}{|x|} = \frac{\beta^{e-t+1}}{2\beta^e}$$
$$\mathbf{u} := \frac{\beta^{-t+1}}{2}$$

- **Το u** αποκαλείται «μονάδα στρογγύλευσης» **(unit roundoff)**
  - χρησιμοποιείται εκτενέστατα ως μονάδα μέτρησης σφαλμάτων.

ΠΡΟΣΟΧΗ: Αν αλλάξει ο τρόπος στρογγύλευσης, θα αλλάξει και το u.

Πανεπιστήμιο Πατρών

- **Στην α.κ.υ. IEEE με στρογγύλευση προς πλησιέστερο, β=2, οπότε**
  - **για αριθμητική μονής ακρίβειας**

$$u = 2^{-t} = 2^{-24} \approx 5.96 \times 10^{-8}$$

  - **για αριθμητική διπλής ακρίβειας (default MATLAB)**

$$u = 2^{-t} = 2^{-53} \approx 1.11 \times 10^{-16}$$

Έστω $x \in \mathbb{R}$ και ότι θέλουμε να υπολογίσουμε φράγμα για το σφάλμα που προκύπτει μετά τη στρογγύλευση του x στον πλησιέστερο α.κ.υ. Έστω $[x^-, x^+]$ το μικρότερο διάστημα α.κ.υ. που εγκλείει το x (επομένως $x^-, x^+ \in \mathcal{F}$.)
Το σφάλμα της στρογγύλευσης θα είναι

$$|fl(x) - x| = \min\{|x - x^-|, |x - x^+|\}$$

Το σφάλμα θα είναι μέγιστο όταν το x είναι ακριβώς στη μέση του διαστήματος, οπότε το x ισαπέχει από τα άκρα $[x^-, x^+]$, οπότε:

$$\frac{\beta^{1-t+e}}{2}$$

Αν $x \in [\beta^e, (1+\beta^{-t+1}) \times \beta^e]$ τότε το μέγιστο σχετικό σφάλμα θα είναι

$$\frac{|fl(x) - x|}{|x|} \leq \frac{\beta^{1-t+e}}{2\beta^e} = \frac{\beta^{1-t}}{2}$$

Αν $x \in [(1+\beta^{-t+1})\beta^e, (1+2\beta^{-t+1}) \times \beta^e]$ τότε το μέγιστο σχετικό σφάλμα θα είναι

$$\frac{|fl(x) - x|}{|x|} \leq \frac{\beta^{1-t+e}}{2(1+\beta^{-t+1})\beta^e} < \frac{\beta^{1-t}}{2}$$

και γενικότερα, αν $x \in [(1+\psi)\beta^e, (1+\psi+\beta^{1-t}) \times \beta^e]$, όπου $0 \leq \psi$ οπότε

$$\frac{|fl(x) - x|}{|x|} \leq \frac{\beta^{1-t+e}}{2(1+\psi)\beta^e} \leq \frac{\beta^{1-t}}{2}$$

και προφανώς το ελάχιστο θα επιτευχθεί όταν $\psi = 0$, δηλ. στο υποδιάστημα που βρίσκεται άκρα αριστερά.

1

Έψιλον της μηχανής και ulp

- **eps μηχανής: Είναι η απόσταση του 1 από τον αμέσως επόμενο α.κ.υ., έστω 1^+ δηλ. $\varepsilon_M = 1^+ - 1$**
- **Στο σύστημα α.κ.υ. IEEE $\varepsilon_M = 2^{-t+1} = 2u$**
  - **ε_M διπλής ακρίβειας = $2^{-52} \approx 10^{-16}$**
  - **Το ε_M υπάρχει χρησιμοποιείται σε αλγορίθμους γιατί δείχνει τη διακριτότητα του συστήματος α.κ.υ. Δείτε π.χ. τη συνάρτηση `rank` της MATLAB.**
  - **Το ε_M ορίστηκε βάσει του 1. Η γενίκευσή του σε μεγαλύτερους αριθμούς ονομάζεται ulp (units in the last place)**
  - **Αν $x = m \times 2^E$ τότε $ulp(x) = \varepsilon_M \times 2^E$**

Στη MATLAB 7.5

```
>> eps('single')
ans = 1.1921e-007
>> eps('double')
ans = 2.2204e-016
>> eps
ans = 2.2204e-016
```

Η rank επιστρέφει την τάξη του μητρώου. Επειδή όμως ο ορισμός της τάξης εξαρτάται από το πλήθος των ιδιαζουσών τιμών που είναι μη μηδενικές, τίθεται το θέμα τι σημαίνει όταν λέμε ότι καποια ποσότητα είναι 0; Θα πρέπει να είναι ακριβώς 0; Ή θα δεχτούμε ότι λόγω «θορύβου» στους υπολογισμούς, ό,τι είναι κάτω από κάποιο κατώφλι μπορεί να ληφθεί και αυτό ως 0; Στην πράξη χρησιμοποιείται η δεύτερη ερμηνεία, οπότε χρειάζεται να οριστεί το κατώφλι, κάτι που γίνεται βάσει του eps. Για παράδειγμα, help rank της MATLAB επιστρέφει

```
RANK   Matrix rank.
    RANK(A) provides an estimate of the number of linearly
    independent rows or columns of a matrix A.
    RANK(A,tol) is the number of singular values of A
    that are larger than tol.
    RANK(A) uses the default tol = max(size(A)) * norm(A) * eps.
```

ΤΜΗΥΠ
Πανεπιστήμιο Πατρών

Πράξεις α.κ.υ.

- Η συζήτηση δεν εξαντλείται στην αναπαράσταση των αριθμών!
- Το επόμενο μεγάλο θέμα αφορά στην υλοποίηση των πράξεων στο σύστημα α.κ.υ.
- ... και στα σφάλματα που προκύπτουν από την υλοποίησή τους σε σχέση με τα «θεωρητικά αποτελέσματα» της «αριθμητικής άπειρης ακρίβειας».

- *Το ζήτημα είναι οι «πράξεις α.κ.υ.» που αντιστοιχούν στις συνηθισμένες πράξεις σε αριθμητική άπειρης ακρίβειας.*
  - *Υλοποίηση των πράξεων → θέμα **Computer Arithmetic***
  - *Τα χαρακτηριστικά τους → σφάλματα, ειδικές περιπτώσεις.*

ΤΜΗΥΠ Πανεπιστήμιο Πατρών

Αξιώματα αριθμητικής στο πεδίο R

- **A0:** $x, y \in R \Rightarrow x+y \in R$
- **A1:** $x+y = y+x$
- **A2:** $x+(y+z) = (x+y)+z$
- **A3:** $\exists\ 0 \in R$ τ.ώ. $\forall\ x \in R,\ x+0 = x$
- **A4:** $\forall\ x \in R,\ \exists\ (-x) \in R$ τ.ώ. $x+(-x) = 0$

- **Π0:** $x, y \in R \Rightarrow x \times y \in R$
- **Π1:** $x \times y = y \times x$
- **Π2:** $x \times (y \times z) = (x \times y) \times z$
- **Π3:** $\exists\ 1 \in R$ τ.ώ. $\forall\ x \in R,\ x \times 1 = x$
- **Π4:** $\forall\ x \in R,$ τ.ώ. $x \neq 0, \exists\ x^{-1} \in R$ τ.ώ. $x \times x^{-1} = 1$

- **Ε0:** $x \times (y+z) = x \times y + x \times z$

Αρκετές από τις παραπάνω ιδιότητες δεν ισχύουν στους α.κ.υ.

ΤΜΗΥΠ Πανεπιστήμιο Πατρών

- Οι παρακάτω πράξεις είναι ισοδύναμες και το αποτέλεσμα ίσο με **10**, αλλά σε α.κ.υ. IEEE διπλής ακρίβειας (π.χ. στη **MATLAB**)

$$10^{20} + 20 - 10 - 10^{20} = \quad 0$$

$$10^{20} + 20 - 10^{20} - 10 = \quad -10$$

$$10^{20} - 10 - 10^{20} + 20 = \quad 20$$

- Προσέξτε ότι $10^{20} > 2^{52}$ επομένως τα παραπάνω πρέπει να αναμένονται!
- Ερώτηση: Από τους **24 (=4!)** τρόπους υπολογισμού παραπάνω, ποιοί επιστρέφουν σωστό αποτέλεσμα;

Παράδειγμα παρατυπιών

- Ερώτημα: Αν $x \in$ **F**, ισχύουν τα παρακάτω;
  - $1 \tilde{\times} x = x$
  - $x \neq 0 \Rightarrow x \tilde{/} x = 1$
  - $0.5 \tilde{\times} x = x \tilde{/} 2$
  - $y \in F, x \tilde{-} y = 0 \Rightarrow x = y$

Πριν από το 1980 υπήρχαν Η/Υ για τους οποίους η απάντηση σε ένα ή περισσότερα από τα παραπάνω ερωτήματα ήταν αρνητική.

Στην α.κ.υ. IEEE οι απαντήσεις σε όλα τα παραπάνω είναι ΝΑΙ.

Παράδειγμα

- **Δεν ισχύει πάντα το Α2:**

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

Δεν εξασφαλίζεται η ισότητα

$$t_2 = s_2$$

δηλαδή η ισότητα

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

Αποτυχία του Π4

```
index = [];
for i=1:170
    if ((1/i)*i~= 1)
        index = [index i]
        pause;
    end;
end;
```

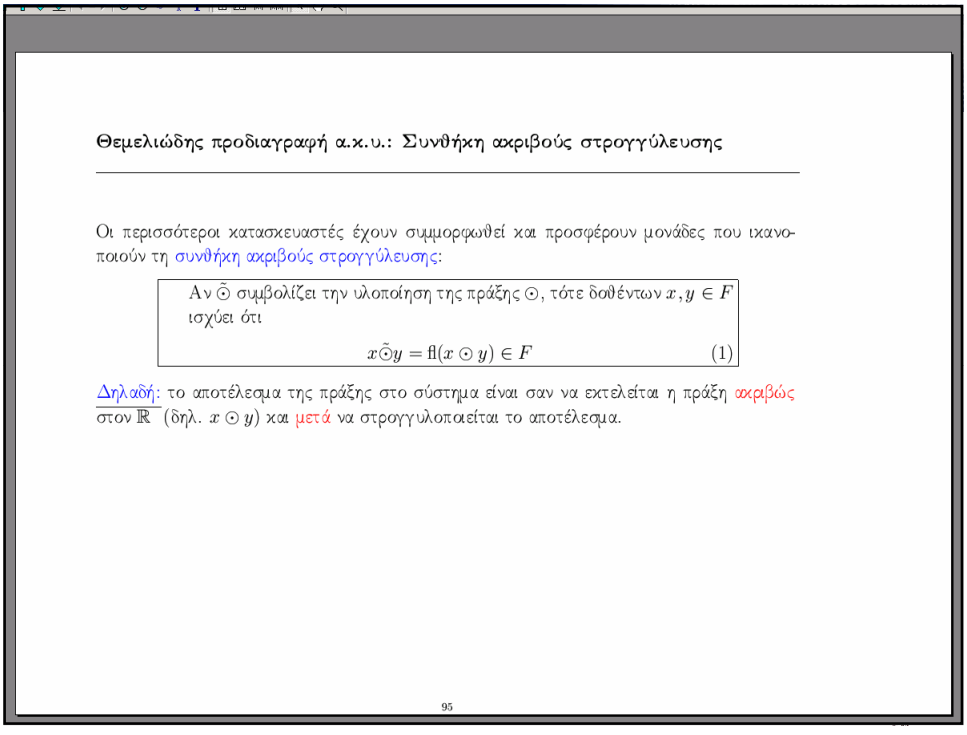
Σε MATLAB @ P4 το αποτέλεσμα ήταν index = [49 98 103 107 161]

Αν τρέξουμε ως 1000, υπάρχουν 82 αριθμοί που αποτυγχάνουν στο παραπάνω τεστ

- **Σε σχέση με το Α3:**
- **Μπορεί να ισχύει**
  x, y \in F\{0} αλλά x y = x
- **Παράδειγμα:**
  |x| ≫ |y| και κατά την εκτέλεση της άθροισης, η εξομοίωση του εκθέτη του y με εκείνον του x οδηγεί σε απώλεια όλων των ψηφίων της ουράς του y.
- **Χαρακτηριστική είναι η περίπτωση που x=1.**
- **Για τους λόγους που αναφέραμε θα υπάρχουν y ≠ 0 τ.ώ. 1 y = 1.**
- **Το μέγιστο αυτό y είναι περίπου ίσο με ε_M**

ΤΜΗΥΠ ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ & ΠΛΗΡΟΦΟΡΙΚΗΣ

Πανεπιστήμιο Πατρών



Θεμελιώδης προδιαγραφή α.κ.υ.: Συνθήκη ακριβούς στρογγύλευσης

Οι περισσότεροι κατασκευαστές έχουν συμμορφωθεί και προσφέρουν μονάδες που ικανοποιούν τη συνθήκη ακριβούς στρογγύλευσης:

> Αν $\tilde{\odot}$ συμβολίζει την υλοποίηση της πράξης $\odot$, τότε δοθέντων $x, y \in F$ ισχύει ότι
>
> $$x \tilde{\odot} y = \mathrm{fl}(x \odot y) \in F \quad (1)$$

Δηλαδή: το αποτέλεσμα της πράξης στο σύστημα είναι σαν να εκτελείται η πράξη ακριβώς στον $\mathbb{R}$ (δηλ. $x \odot y$) και μετά να στρογγυλοποιείται το αποτέλεσμα.

95

Παρατηρήσεις

- **Η διασφάλιση της συνθήκης ακριβούς στρογγύλευσης δεν είναι προφανής:**
  - **Φαίνεται να απαιτούνται μεγάλοι καταχωρητές για την αποθήκευση των ενδιάμεσων αποτελεσμάτων.**
- **Σημαντικό αποτέλεσμα (1980-)**
  - **Για τη σωστή υλοποίηση της συνθήκης αρκούν πολύ λίγα επιπλέον ψηφία/bits:**
    - **Ψηφίο στρογγύλευσης**
    - **Ψηφίο προστασίας**
    - **Κολλώδες (sticky) bit**

Παράδειγμα 7. *Έστω δυαδική αριθμητική με* $t = 3$ *οπότε* $u = 2^{1-3}/2 = 2^{-3}$*:*

	με ψηφίο προστασίας	*χωρίς ψηφίο προστασίας*
$+$	$2^1 \times 0.100$	$2^1 \times 0.100$
$-$	$2^0 \times 0.111$	$2^0 \times 0.111$
	$2^1 \times 0.100$	$2^1 \times 0.100$
$-$	$2^1 \times 0.011\|\mathbf{1}$	$2^1 \times 0.011$
$=$	$2^1 \times 0.000\mathbf{1}$	$2^1 \times 0.001$
$=$	$2^{-2} \times 0.100$	$2^{-1} \times 0.100$

Με ψηφίο προστασίας έχουμε το ακριβές αποτέλεσμα αλλοιώς το σχετικό σφάλμα γίνεται

$$\left|\frac{2^{-2} \times 0.1 - 2^{-1} \times 0.1}{2^{-2} \times 0.100}\right| = 1 = 8u$$

□

97

Το παρακάτω πρόγραμμα υπολογίζει το ελάχιστο y για το οποίο ισχύει το $1+y > 0$ και που είναι της μορφής 1.0×2^p

```
T= 1.0
While (1.0+T > 1.0)
      T=T/2.0;
End
T = T*2.0;
```

Αν τρέξετε το παραπάνω πρόγραμμα στη MATLAB, π.χ. σε P4 (α.κ.υ. IEEE):

T = 2.220446049250313e-016

σε 16κή μορφή (για συνοπτική αναπαράσταση του δυαδικού) είναι

$(3cb0000000000000)_{16}$

Αυτό συμπίπτει με το ε_M που επιστρέφει η MATLAB.

Παρατήρηση: Παρόλα αυτά, ο T *δεν είναι* ο ελάχιστος κανονικοποιημένος θετικός α.κ.υ. για τον οποίο ισχύει ότι $1+T > 1$.
Ο αριθμός αυτός είναι ο

$2^{-53}+2^{-105} = (3ca0000000000001)_{16}$

Παρατηρήσεις

- **Χρήσιμο να υπάρχουν προγράμματα που ανακαλύπτουν και αναδεικνύουν στοιχεία για το σύστημα αριθμητικής**
  - **Παλαιότερα ήταν απολύτως απαραίτητα, π.χ.**
    - **PARANOIA (Kahan)**
    - **MACHAR (Cody)**
    - **Τα προγράμματα αυτά είναι δύσκολο να γραφτούν γιατί λειτουργούν στα άκρα της αριθμητικής (extreme arithmetic!!!)**
    - **Για «φανατικούς» (το συνιστώ θερμά ακόμα και αν δεν το καταλάβετε εντελώς!!!!)**
      - ✓ **Το άρθρο του Kahan, "MATLAB's loss is nobody's gain" (2004)**

Προβληματισμός στα GPUs

GPU Floating-Point Paranoia

Karl E. Hillesland
University of North Carolina at Chapel Hill *

Anselmo Lastra
University of North Carolina at Chapel Hill *

1 Introduction

Up until the late eighties, each computer vendor was left to develop their own conventions for floating-point computation as they saw fit. As a result, programmers needed to familiarize themselves with the peculiarities of each system in order to write effective software and evaluate numerical error. In 1987, a standard was established for floating-point computation to alleviate this problem, and CPU vendors now design to this standard [IEEE 1987].

Today there is an interest in the use of graphics processing units, or GPUs, for non-graphics applications such as scientific computing. GPUs have floating-point representations similar to, and sometimes matching, the IEEE standard. However, we have found that GPUs do not adhere to IEEE standards for floating-point operations, nor do they give the information necessary to establish bounds on error for these operations. Another complication is that this behavior seems to be in a constant state of flux due to the dependent

Operation	R300/arbfp	NV30/fp30
Addition	[-1.000, 0.000]	[-1.000, 0.000]
Subtraction	[-1.000, 1.000]	[-0.750, 0.750]
Multiplication	[-0.989, 0.125]	[-0.782, 0.625]
Division	[-2.869, 0.094]	[-1.199, 1.375]

Table 1: Floating-Point Error in ULPs (Units in Last Place). Note that the R300 has a 16 bit significand, whereas the NV30 has 23 bits. Therefore one ULP on an R300 is equivalent to 2^7 ULPs on an NV30. Division is implemented by a combination of reciprical and multiply on these systems. Cg version 1.2.1. ATI driver 6.14.10.6444. NVIDIA driver 56.72.

Schryer [Schryer 1981]. By testing all combinations of these numbers, we include all the test cases in Paranoia, as well as cases that push the limits of round-off error and cases where the most work must be performed, such as extensive carry propagation. Table 1 gives results for some example systems.

Από MACHAR (Cody)

```
                A = ONE
10              A = A + A
                TEMP = A+ONE
                TEMP1 = TEMP-A
                IF (TEMP1-ONE .EQ. ZERO) GO TO 10
                B = ONE
20              B = B + B
                TEMP = A+B
                ITEMP = INT(TEMP-A)
                IF (ITEMP .EQ. 0) GO TO 20
                IBETA = ITEMP
                BETA = CONV(IBETA)
C-----------------------------------------------------------------------
C  Determine IT, IRND.
C-----------------------------------------------------------------------
                IT = 0
                B = ONE
100             IT = IT + 1
                B = B * BETA
                TEMP = B+ONE
                TEMP1 = TEMP-B
                IF (TEMP1-ONE .EQ. ZERO) GO TO 100
```

Τα κεντρικά θέματα

- Τι αποτελέσματα προκύπτουν μετά από πράξεις α.κ.υ. ;
- Πώς μετράμε το σφάλμα;
- Πώς διαδίδονται τα σφάλματα
  - Αν υπάρχουν στα δεδομένα εισόδου
  - Αυτά που προκύπτουν κατά τη διάρκεια των υπολογισμών;
  - πώς επηρεάζουν τα τελικά αποτελέσματα;
- Πώς μπορούμε να μετρήσουμε την αξιοπιστία
  - των αποτελεσμάτων;
  - του αλγορίθμου και της υλοποίησής του;

Βασικοί κανόνες διάδοσης σφάλματος

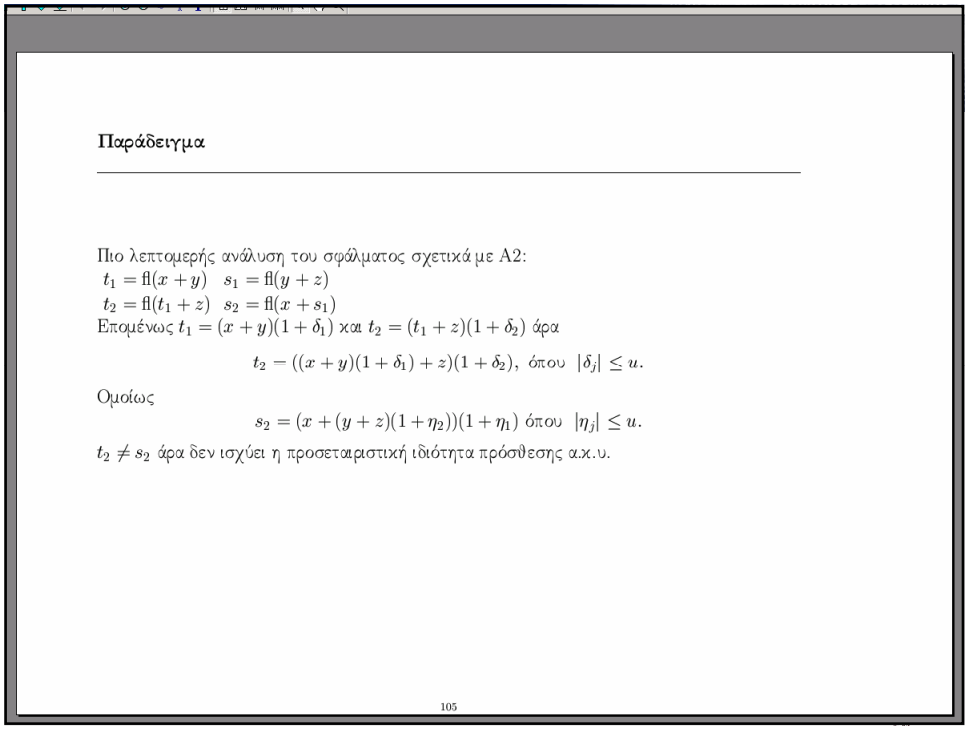
- Έστω **x,y** $\in$ **F** και ότι δεν έχουμε υπερχείλιση ή υποχείλιση στην εκτέλεση του **x** $\odot$ **y**, τότε

$$\frac{|\mathrm{fl}(x \odot y) - (x \odot y)|}{|x \odot y|} \leq u, \quad x \odot y \neq 0, \ \odot \in \{+, -, \times, /\}.$$

- Αν ισχύουν τα παραπάνω τότε ισχύουν επίσης

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

Οι παραπάνω σχέσεις είναι τα βασικά εργαλεία:
Με συστηματική χρήση τους θα μελετήσουμε τη διάδοση και θα εκτιμήσουμε φράγματα για το τελικό σφάλμα υπολογισμών με α.κ.υ.



Παράδειγμα

Πιο λεπτομερής ανάλυση του σφάλματος σχετικά με A2:
$t_1 = \mathrm{fl}(x+y) \quad s_1 = \mathrm{fl}(y+z)$
$t_2 = \mathrm{fl}(t_1+z) \quad s_2 = \mathrm{fl}(x+s_1)$
Επομένως $t_1 = (x+y)(1+\delta_1)$ και $t_2 = (t_1+z)(1+\delta_2)$ άρα

$$t_2 = ((x+y)(1+\delta_1)+z)(1+\delta_2), \text{ όπου } |\delta_j| \leq u.$$

Ομοίως

$$s_2 = (x+(y+z)(1+\eta_2))(1+\eta_1) \text{ όπου } |\eta_j| \leq u.$$

$t_2 \neq s_2$ άρα δεν ισχύει η προσεταιριστική ιδιότητα πρόσθεσης α.κ.υ.

105

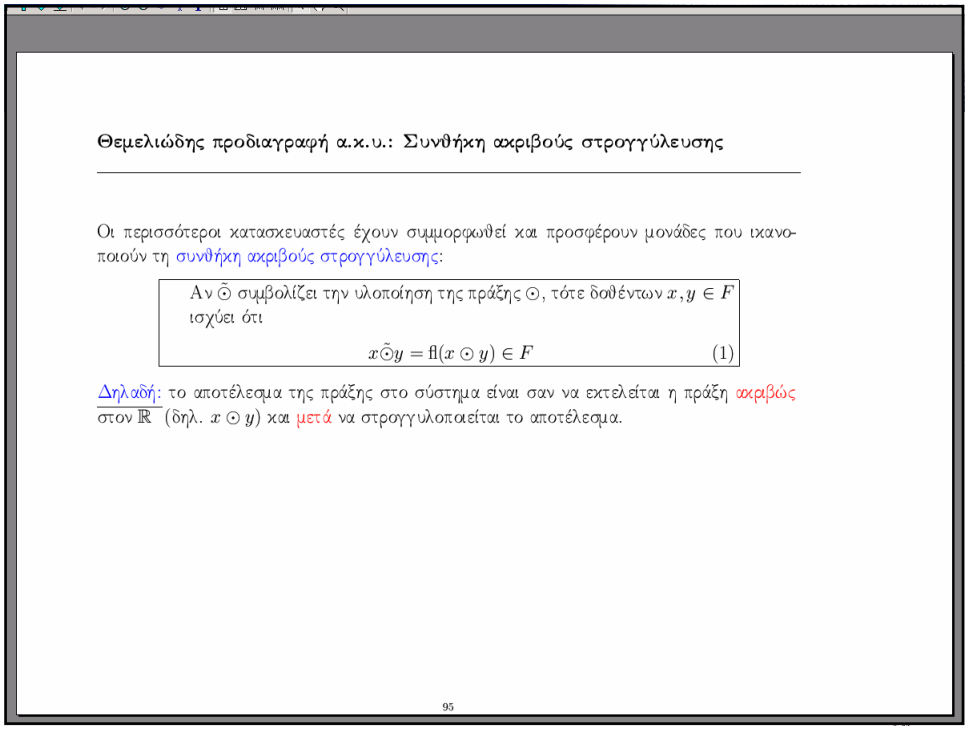
Παρατηρήσεις

- **Αξίζει να προσέξετε από τώρα ότι παρόλο που ενδιαφερόμαστε για τα σφάλματα, τις περισσότερες περιπτώσεις δεν μπορούμε να τα υπολογίσουμε ακριβώς!**
- **… αν μπορούσαμε, δεν θα είχαμε αντικείμενο συζήτησης, καθώς θα μπορούσαμε να προσθέσουμε την υπολογισμένη τιμή στο σφάλμα και να λάβουμε την ακριβή τιμή (που γενικά είναι άγνωστη!)**

Επομένως «χαλαρώνουμε» και όταν αναφερόμαστε στην «εύρεση του σφάλματος» συνήθως εννοούμε την εύρεση «καλού άνω φράγματος» για το μέγεθος του σφάλματος

- **… αντί τον ακριβή υπολογισμό του σφάλματος, που γενικά είναι αδύνατος,**
- **… όπως παραπάνω, που διερευνήσαμε ποιο μπορεί να είναι το μέγιστο σφάλμα στρογγύλευσης αν $x \in (x^-, x^+)$ και πότε μεγιστοποιείται.**

ΤΜΗΥΠ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ
Πανεπιστήμιο Πατρών



Θεμελιώδης προδιαγραφή α.κ.υ.: Συνθήκη ακριβούς στρογγύλευσης

Οι περισσότεροι κατασκευαστές έχουν συμμορφωθεί και προσφέρουν μονάδες που ικανοποιούν τη συνθήκη ακριβούς στρογγύλευσης:

> Αν $\tilde{\odot}$ συμβολίζει την υλοποίηση της πράξης $\odot$, τότε δοθέντων $x, y \in F$ ισχύει ότι
>
> $$x \tilde{\odot} y = \mathrm{fl}(x \odot y) \in F \qquad (1)$$

Δηλαδή: το αποτέλεσμα της πράξης στο σύστημα είναι σαν να εκτελείται η πράξη ακριβώς στον $\mathbb{R}$ (δηλ. $x \odot y$) και μετά να στρογγυλοποιείται το αποτέλεσμα.

95

Βασικοί κανόνες διάδοσης σφάλματος

- **Έστω x,y ∈ F και ότι δεν έχουμε υπερχείλιση ή υποχείλιση στην εκτέλεση του x ⊙ y, τότε**

$$\frac{|\mathrm{fl}(x \odot y) - (x \odot y)|}{|x \odot y|} \leq u, \quad x \odot y \neq 0, \ \odot \in \{+, -, \times, /\}.$$

- **Αν ισχύουν τα παραπάνω τότε ισχύουν επίσης**

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

Οι παραπάνω σχέσεις είναι τα βασικά εργαλεία:
Με συστηματική χρήση τους θα μελετήσουμε τη διάδοση και θα εκτιμήσουμε φράγματα για το τελικό σφάλμα υπολογισμών με α.κ.υ.

ΤΜΗΥΠ Πανεπιστήμιο Πατρών

Παράδειγμα

Πιο λεπτομερής ανάλυση του σφάλματος σχετικά με A2:
$t_1 = \mathrm{fl}(x+y) \quad s_1 = \mathrm{fl}(y+z)$
$t_2 = \mathrm{fl}(t_1+z) \quad s_2 = \mathrm{fl}(x+s_1)$
Επομένως $t_1 = (x+y)(1+\delta_1)$ και $t_2 = (t_1+z)(1+\delta_2)$ άρα

$$t_2 = ((x+y)(1+\delta_1)+z)(1+\delta_2), \text{ όπου } |\delta_j| \leq u.$$

Ομοίως

$$s_2 = (x+(y+z)(1+\eta_2))(1+\eta_1) \text{ όπου } |\eta_j| \leq u.$$

$t_2 \neq s_2$ άρα δεν ισχύει η προσεταιριστική ιδιότητα πρόσθεσης α.κ.υ.

105

Τι μπορούμε να πούμε για το σφάλμα;

- **Από τα παραπάνω μπορούμε να μελετήσουμε το σφάλμα για τις δυο μεθόδους υπολογισμού.**

$$\begin{aligned}
((x+y)(1+\delta_1)+z)(1+\delta_2) &= x(1+\delta_1)(1+\delta_2)+y(1+\delta_1)(1+\delta_2)+z(1+\delta_2) \\
&\Downarrow \\
|((x\tilde{+}y)\tilde{+}z)-((x+y)+z)| &= |t_2-((x+y)+z)| \\
&= |x(\delta_1+\delta_2+\delta_1\delta_2)+y(\delta_1+\delta_2+\delta_1\delta_2)+z\delta_2| \\
&\leq |x|(2\mathbf{u}+\mathbf{u}^2)+|y|(2\mathbf{u}+\mathbf{u}^2)+|z|\mathbf{u} \\
|(x\tilde{+}(y\tilde{+}z))-(x+(y+z))| &\leq |x|\mathbf{u}+|y|(2\mathbf{u}+\mathbf{u}^2)+|z|(2\mathbf{u}+\mathbf{u}^2)
\end{aligned}$$

ΠΡΟΣΕΞΤΕ όπως αναμενόταν, το άνω φραγμα για το σφάλμα εξαρτάται και από τις τιμές των x, y, z και μπορεί να είναι διαφορετικό στους δύο τρόπους. Αν οι τιμές είναι θετικές, φαίνεται να ευνοείται η άθροιση κατά αύξουσα τιμή.

Η ΑΞΙΟΠΙΣΤΗ άθροιση ενός συνόλου αριθμών είναι σημαντικό θέμα (καθότι είναι πυρήνας πολλών υπολογισμών) (δείτε επόμενη διάλεξη).

Τι μπορούμε να πούμε για το σφάλμα;

- **Από τα παραπάνω μπορούμε να μελετήσουμε το σφάλμα για τις δυο μεθόδους υπολογισμού.**

$$
\begin{aligned}
((x+y)(1+\delta_1)+z)(1+\delta_2) &= x(1+\delta_1)(1+\delta_2)+y(1+\delta_1)(1+\delta_2)+z(1+\delta_2) \\
&\Downarrow \\
|((x\tilde{+}y)\tilde{+}z)-((x+y)+z)| &= |t_2-((x+y)+z)| \\
&= |x(\delta_1+\delta_2+\delta_1\delta_2)+y(\delta_1+\delta_2+\delta_1\delta_2)+z\delta_2| \\
&\leq |x|(2\mathbf{u}+\mathbf{u}^2)+|y|(2\mathbf{u}+\mathbf{u}^2)+|z|\mathbf{u} \\
|(x\tilde{+}(y\tilde{+}z))-(x+(y+z))| &\leq |x|\mathbf{u}+|y|(2\mathbf{u}+\mathbf{u}^2)+|z|(2\mathbf{u}+\mathbf{u}^2)
\end{aligned}
$$

ΠΡΟΣΕΞΤΕ όπως αναμενόταν, το άνω φραγμα για το σφάλμα εξαρτάται και από τις τιμές των x, y, z και μπορεί να είναι διαφορετικό στους δύο τρόπους. Αν οι τιμές είναι θετικές, φαίνεται να ευνοείται η άθροιση κατά αύξουσα τιμή.

Η ΑΞΙΟΠΙΣΤΗ άθροιση ενός συνόλου αριθμών είναι σημαντικό θέμα (καθότι είναι πυρήνας πολλών υπολογισμών) (δείτε επόμενη διάλεξη).

Από το απόλυτο στο σχετικό σφάλμα

Με μικρή χαλάρωση του φράγματος, μπορούμε να καταλήξουμε στο ίδιο άνω φράγμα για το μέγιστο σφάλμα και στους δύο τρόπους άθροισης:

$$
\begin{aligned}
|((x\tilde{+}y)\tilde{+}z)-((x+y)+z)| &< (|x|+|y|+|z|)(2\mathbf{u}+\mathbf{u}^2) \\
|(x\tilde{+}(y\tilde{+}z))-(x+(y+z))| &< (|x|+|y|+|z|)(2\mathbf{u}+\mathbf{u}^2)
\end{aligned}
$$

- **Το άνω φράγμα για το απόλυτο σφάλμα εξαρτάται από τα μεγέθη των *x, y, z***

- **Πώς φράσσεται το αντίστοιχο σχετικό σφάλμα;**

$$
\frac{|((x\tilde{+}y)\tilde{+}z)-(x+y+z)|}{|x+y+z|} < \frac{(|x|+|y|+|z|)}{|x+y+z|}(2\mathbf{u}+\mathbf{u}^2)
$$

ΕΡΩΤΗΣΗ: Ποιο είναι το μειονέκτημα αυτού του φράγματος;

Από το απόλυτο στο σχετικό σφάλμα

Με μικρή χαλάρωση του φράγματος, μπορούμε να καταλήξουμε στο ίδιο άνω φράγμα για το μέγιστο σφάλμα και στους δύο τρόπους άθροισης:

$$|((x\tilde{+}y)\tilde{+}z) - ((x+y)+z)| < (|x|+|y|+|z|)(2\mathbf{u}+\mathbf{u}^2)$$
$$|(x\tilde{+}(y\tilde{+}z)) - (x+(y+z))| < (|x|+|y|+|z|)(2\mathbf{u}+\mathbf{u}^2)$$

- **Το άνω φράγμα για το απόλυτο σφάλμα εξαρτάται από τα μεγέθη των x, y, z**

- **Πώς φράσσεται το αντίστοιχο σχετικό σφάλμα;**

$$\frac{|((x\tilde{+}y)\tilde{+}z) - (x+y+z)|}{|x+y+z|} < \frac{(|x|+|y|+|z|)}{|x+y+z|}(2\mathbf{u}+\mathbf{u}^2)$$

ΕΡΩΤΗΣΗ: Ποιο είναι το μειονέκτημα αυτού του φράγματος;

Από το απόλυτο στο σχετικό σφάλμα

1. **Το φράγμα εξαρτάται από τις τιμές των x, y, z**
   - **ουσιαστικά δεν έχουμε καταφέρει να φράξουμε το πίσω σφάλμα**
2. **Το φράγμα μπορεί να γίνει πολύ μεγάλο (αποκλειστικά λόγω του πρώτου όρου), οπότε θα είναι άχρηστο**
   - **δηλ. αν $(|x|+|y|+|z|)/|x+y+z|$ είναι πολύ μεγάλο**
- **Μπορούμε πολύ καλύτερα αν έχουμε περισσότερες πληροφορίες για το πρόβλημα:**
   - **π.χ. αν τα x, y, z είναι ομόσημα**

Άνω φράγμα για το σχετικό σφάλμα

- Όταν x, y, z ομόσημοι τότε

$$\begin{aligned}\frac{|((x\tilde{+}y)\tilde{+}z)-(x+y+z)|}{|x+y+z|} &< \frac{(|x|+|y|+|z|)}{|x+y+z|}(2\mathbf{u}+\mathbf{u}^2)\\ &< 2\mathbf{u}+\mathbf{u}^2\end{aligned}$$

- πολύ καλό άνω φράγμα, μας πληροφορεί ότι το σφάλμα θα είναι μικρό (τάξης u)
- ίδιο φράγμα και για την εναλλακτική άθροιση
  $x+(y+z)$
- είναι όμως απαραίτητο να κάνουμε υποθέσεις για τις τιμές x, y, z

Σε άλλες περιπτώσεις μπορούμε πολύ καλύτερα

$$\begin{aligned}|(x\tilde{\times}y)\tilde{\times}z - x\times y\times z| &= |((x\times y)(1+\delta_1)\times z)(1+\delta_2) - x\times y\times z|\\ &= |xyz(1+\delta_1)(1+\delta_2)-xyz|\\ &= |xyz(\delta_1+\delta_2+\delta_1\delta_2)|\end{aligned}$$

επομένως αν $xyz \neq 0$

$$\begin{aligned}\frac{|(x\tilde{\times}y)\tilde{\times}z - xyz|}{|xyz|} &= |\delta_1+\delta_2+\delta_1\delta_2|\\ &\leq 2\mathbf{u}+\mathbf{u}^2\end{aligned}$$

Σε άλλες περιπτώσεις μπορούμε πολύ καλύτερα

$$\begin{aligned}|(x\tilde{\times}y)\tilde{\times}z - x\times y\times z| &= |((x\times y)(1+\delta_1)\times z)(1+\delta_2) - x\times y\times z| \\ &= |xyz(1+\delta_1)(1+\delta_2) - xyz| \\ &= |xyz(\delta_1+\delta_2+\delta_1\delta_2)|\end{aligned}$$

επομένως αν $xyz \neq 0$

$$\begin{aligned}\frac{|(x\tilde{\times}y)\tilde{\times}z - xyz|}{|xyz|} &= |\delta_1+\delta_2+\delta_1\delta_2| \\ &\leq 2\mathbf{u}+\mathbf{u}^2\end{aligned}$$

Επεκτείνεται ακόμα και αν έχουμε σφάλματα στην είσοδο

Αν τα δεδομένα δεν είναι α.κ.υ. και πρέπει να τα «στρογγυλέψουμε» αρκεί να θέσουμε

$$\tilde{x} = x(1+\delta_3), \tilde{y} = y(1+\delta_4), \tilde{z} = z(1+\delta_5),$$

,

$$\begin{aligned}|(\tilde{x}\tilde{\times}\tilde{y})\tilde{\times}\tilde{z} - x\times y\times z| &= |((x(1+\delta_3)\times y(1+\delta_4)(1+\delta_1)\times z(1+\delta_5)(1+\delta_2) - x\times y\times z| \\ &= |xyz(1+\delta_1)(1+\delta_2)(1+\delta_3)(1+\delta_4)(1+\delta_5) - xyz| \\ &= |xyz(\delta_1+\delta_2+\delta_3+\delta_4+\delta_5+\delta_1\delta_2+\cdots+\delta_1\delta_2\delta_3\delta_4\delta_5)|\end{aligned}$$

Μπορούμε να συμπεράνουμε ότι το σχετικό σφάλμα φράσσεται από

$$\begin{aligned}\frac{|(\tilde{x}\tilde{\times}\tilde{y})\tilde{\times}\tilde{z} - xyz|}{|xyz|} &\leq 5\mathbf{u}+10\mathbf{u}^2+10\mathbf{u}^3+5\mathbf{u}^4+\mathbf{u}^5 \\ &\leq 5\mathbf{u}+O(\mathbf{u}^2)\end{aligned}$$

από δυωνυμικό ανάπτυγμα

κυρίαρχος όρος ανω φράγματος
μπορούμε να είμαστε πολύ ικανοποιημένοι

Επεκτείνεται ακόμα και αν έχουμε σφάλματα στην είσοδο

Αν τα δεδομένα δεν είναι α.κ.υ. και πρέπει να τα «στρογγυλέψουμε» αρκεί να θέσουμε

$$\tilde{x} = x(1+\delta_3), \tilde{y} = y(1+\delta_4), \tilde{z} = z(1+\delta_5),$$

,

$$\begin{aligned} |(\tilde{x}\tilde{\times}\tilde{y})\tilde{\times}\tilde{z} - x \times y \times z| &= |((x(1+\delta_3) \times y(1+\delta_4)(1+\delta_1) \times z(1+\delta_5)(1+\delta_2) - x \times y \times z| \\ &= |xyz(1+\delta_1)(1+\delta_2)(1+\delta_3)(1+\delta_4)(1+\delta_5) - xyz| \\ &= |xyz(\delta_1+\delta_2+\delta_3+\delta_4+\delta_5+\delta_1\delta_2+\cdots+\delta_1\delta_2\delta_3\delta_4\delta_5)| \end{aligned}$$

Μπορούμε να συμπεράνουμε ότι το σχετικό σφάλμα φράσσεται από

$$\begin{aligned} \frac{|(\tilde{x}\tilde{\times}\tilde{y})\tilde{\times}\tilde{z} - xyz|}{|xyz|} &\leq 5\mathrm{u} + 10\mathrm{u}^2 + 10\mathrm{u}^3 + 5\mathrm{u}^4 + \mathrm{u}^5 \\ &\leq 5\mathrm{u} + O(\mathrm{u}^2) \end{aligned}$$

από δυωνυμικό ανάπτυγμα

κυρίαρχος όρος ανω φράγματος
μπορούμε να είμαστε πολύ ικανοποιημένοι

Εμπρός σφάλμα – εμπρός ανάλυση σφάλματος

- **Η διαδικασίες που είδαμε πριν συνοψίζονται ως εξής:**
- **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 –**
  - **αναζητούμε δηλαδή το *ελάχιστο άνω φράγμα*!**
- **Η ιδέα είναι απλή**
  - **.... αλλά η εφαρμογή της μπορεί να είναι περίπλοκη**
  - **.... σκληρή άσκηση σε ανισότητες**
  - **.... τεράστιες εκφράσεις ... κ.λπ.**

Εργαλεία (Η απόδειξη – πραιρετική - στο βιβλίο)

Στις αναλύσεις έχουμε όρους $p_n = \prod_{i=1}^{n}(1+\delta_i)$ όπου γνωρίζουμε ότι $|\delta_i| \le u$. Αμέσως βλέπουμε ότι

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

και ότι $p_n = 1 + nu + \mathrm{O}(u^2)$.
Τι πιο ακριβές μπορούμε να πούμε;
Το ακόλουθο Λήμμα είναι πολύ χρήσιμο εργαλείο:
Λήμμα 1. *Αν $|\delta_i| \le u$ και $\rho_i = \pm 1$ για $i = 1 : n$ και $nu < 1$ τότε*

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

όπου

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

□

Προσέξτε ότι μπορούμε επίσης να γράψουμε:

$$(1+\theta_n)(1+\theta_m) = 1+\theta_{m+n}$$

76

Απόδειξη. Με επαγωγή. Η σχέση ισχύει για $n = 1$. Έστω ότι $\rho_n = 1$, τότε

$$\begin{aligned}
\prod_{i=1}^{n}(1+\delta_i)^{\rho_i} &= (1+\theta_{n-1})(1+\delta_n) \\
1+\theta_n &= 1+\theta_{n-1}+\delta_n+\theta_{n-1}\delta_n \\
|\theta_n| &= |\theta_{n-1}+\delta_n+\theta_{n-1}\delta_n| \le \frac{(n-1)u}{1-(n-1)u} + u + \frac{(n-1)u^2}{1-(n-1)u} \\
&\le \frac{(n-1)u+(n-1)u^2+u-(n-1)u^2}{1-(n-1)u} \le \frac{nu}{1-(n-1)u} \le \gamma_n.
\end{aligned}$$

Αν $\rho_n = -1$

$$\begin{aligned}
\prod_{i=1}^{n}(1+\delta_i)^{\rho_i} &= \frac{1+\theta_{n-1}}{1+\delta_n} \\
\theta_{n-1} &= \theta_n+\delta_n+\theta_n\delta_n \\
\theta_n &= \frac{\theta_{n-1}-\delta_n}{1+\delta_n} \\
|\theta_n| &\le |\frac{\theta_{n-1}+u}{1-u}| \\
|\theta_n| &\le \frac{nu-(n-1)u^2}{1-(n-1)u+(n-1)u^2} \le \gamma_n
\end{aligned}$$

□

77

Μερικές φορές γράφουμε $\prod_{i=1}^{n}(1+\delta_i)^{\rho_i} := < n >$ οπότε

$$< n > \times < k > = < n+k >, \quad < n > / < k > = < n+k >$$

Χρήσιμη είναι και η ανισότητα (υποθέτουμε ότι $nu < 1$):

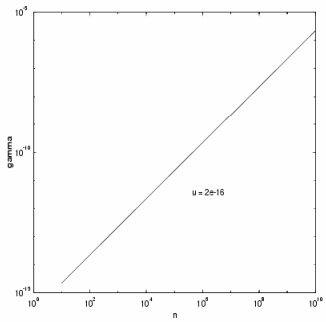
$$\begin{aligned} \gamma_n &= \frac{nu}{1-nu} \\ &\leq nu(1+nu+(nu)^2+\cdots) \\ &= nu+O(u^2) \end{aligned}$$

Επίσης,

$$\prod_{i=1}^{n}(1+\delta_i) \leq (1+u)^n < e^{nu}.$$

78

Μέγεθος του γ_n όταν $u = 2\times 10^{-16}$



79

Χρήση σε προηγούμενο παράδειγμα

Είχαμε δείξει ότι

$$\begin{aligned} |(\tilde{x}\tilde{\times}\tilde{y})\tilde{\times}\tilde{z} - x \times y \times z| &= |((x(1+\delta_3) \times y(1+\delta_4)(1+\delta_1) \times z(1+\delta_5)(1+\delta_2) - x \times y \times z| \\ &= |xyz(1+\delta_1)(1+\delta_2)(1+\delta_3)(1+\delta_4)(1+\delta_5) - xyz| \\ \frac{|(\tilde{x}\tilde{\times}\tilde{y})\tilde{\times}\tilde{z} - x \times y \times z|}{|xyz|} &\leq 5\mathbf{u} + 10\mathbf{u}^2 + 10\mathbf{u}^3 + 5\mathbf{u}^5 + \mathbf{u}^5 \\ &\leq 5\mathbf{u} + O(\mathbf{u}^2) \end{aligned}$$

Με βάση το Λήμμα μπορούμε άμεσα να γράψουμε

$$\begin{aligned} \frac{|(\tilde{x}\tilde{\times}\tilde{y})\tilde{\times}\tilde{z} - xyz|}{|xyz|} &= |(1+\theta_5) - 1| \\ &\leq \gamma_5 = \frac{5\mathbf{u}}{1-5\mathbf{u}} \end{aligned}$$

Χρήση σε προηγούμενο παράδειγμα

Είχαμε δείξει ότι

$$\begin{aligned} |(\tilde{x}\tilde{\times}\tilde{y})\tilde{\times}\tilde{z} - x \times y \times z| &= |((x(1+\delta_3) \times y(1+\delta_4)(1+\delta_1) \times z(1+\delta_5)(1+\delta_2) - x \times y \times z| \\ &= |xyz(1+\delta_1)(1+\delta_2)(1+\delta_3)(1+\delta_4)(1+\delta_5) - xyz| \\ \frac{|(\tilde{x}\tilde{\times}\tilde{y})\tilde{\times}\tilde{z} - x \times y \times z|}{|xyz|} &\leq 5\mathbf{u} + 10\mathbf{u}^2 + 10\mathbf{u}^3 + 5\mathbf{u}^5 + \mathbf{u}^5 \\ &\leq 5\mathbf{u} + O(\mathbf{u}^2) \end{aligned}$$

Με βάση το Λήμμα μπορούμε άμεσα να γράψουμε

$$\begin{aligned} \frac{|(\tilde{x}\tilde{\times}\tilde{y})\tilde{\times}\tilde{z} - xyz|}{|xyz|} &= |(1+\theta_5) - 1| \\ &\leq \gamma_5 = \frac{5\mathbf{u}}{1-5\mathbf{u}} \end{aligned}$$

ΤΜΗΥΠ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ & ΠΛΗΡΟΦΟΡΙΚΗΣ
Πανεπιστήμιο Πατρών