

Αριθμητική Ανάλυση

Τμήμα Ναυπηγών Μηχανικών

Δημήτρης Μητσούδης

Προσεγγίσεις και σφάλματα στον υπολογιστή

Παράσταση αριθμών ως προς οποιαδήποτε βάση

Δεκαδικό σύστημα: Βάση 10, ψηφία: 0, 1, 2, ..., 9.

Π.χ.

$$3.14159 = 3 \cdot 10^0 + 1 \cdot 10^{-1} + 4 \cdot 10^{-2} + 1 \cdot 10^{-3} + 5 \cdot 10^{-4} + 9 \cdot 10^{-5}$$

$$5672 = 5 \cdot 10^3 + 6 \cdot 10^2 + 7 \cdot 10^1 + 2 \cdot 10^0$$

Γενικά, αν δουλεύουμε στο σύνολο αριθμών με βάση β , τότε το σύμβολο

$$\pm(\alpha_N \dots \alpha_0 . \alpha_{-1} \alpha_{-2} \dots)_\beta$$

παριστάνει τον αριθμό με **ακέραιο μέρος**

$$\pm(\alpha_N \beta^N + \alpha_{N-1} \beta^{N-1} + \dots + \alpha_0)$$

και **κλασματικό μέρος**

$$\sum_{k=1}^{\infty} \alpha_{-k} \beta^{-k}$$

Παράδειγμα

$$(100110.11)_2 =$$

Θέση	5	4	3	2	1	0	-1	-2
Τιμή	1	0	0	1	1	0	1	1

$$= 1 \cdot 2^5 + 1 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^{-1} + 1 \cdot 2^{-2} = (38.75)_{10}$$

Αριθμητικά συστήματα στον Η/Υ

Τα συνηθέστερα αριθμητικά συστήματα που χρησιμοποιούνται στον υπολογιστή είναι το δυαδικό, το οκταδικό, και το δεκαεξαδικό.

Σύστημα	Δεκαδικό	Δυαδικό	Οκταδικό	Δεκαεξαδικό
Βάση (β)	10	2	8	16
Ψηφία	0, 1, ..., 9	0, 1	0, 1, ..., 7	0, 1, ..., 9, A, B, C, D, E, F

Μετατροπή αριθμού από ένα σύστημα με βάση β στο δεκαδικό

Απλώς υπολογίζουμε την τιμή της παράστασης:

$$\pm \sum_{k=-\infty}^n a_k \beta^k = \pm a_n \cdot \beta^n + \cdots + a_1 \cdot \beta^1 + a_0 \cdot \beta^0 + a_{-1} \cdot \beta^{-1} + a_{-2} \cdot \beta^{-2} + \cdots$$

Π.χ. Ο αριθμός $(10110.11)_2$ του δυαδικού συστήματος είναι ο αριθμός

$$1 \cdot 2^4 + 0 \cdot 2^3 + 1 \cdot 2^2 + 1 \cdot 2^1 + 0 \cdot 2^0 + 1 \cdot 2^{-1} + 1 \cdot 2^{-2} = (22.75)_{10}.$$

Ο αριθμός $(5137.24)_8$ του οκταδικού συστήματος είναι ο αριθμός

$$5 \cdot 8^3 + 1 \cdot 8^2 + 3 \cdot 8^1 + 7 \cdot 8^0 + 2 \cdot 8^{-1} + 4 \cdot 8^{-2} = 2655 + \frac{5}{16} = (2655.3125)_{10}.$$

Ο αριθμός $(5F9)_{16}$ του δεκαεξαδικού συστήματος είναι ο αριθμός

$$5 \cdot 16^2 + 15 \cdot 16^1 + 9 \cdot 16^0 = (1529)_{10}.$$

Μετατροπή ακεραίου από το δεκαδικό σε ένα σύστημα με βάση β .

Βασίζεται στον αλγόριθμο της διαίρεσης.

Π.χ. θα μετατρέψουμε τον αριθμό $(253)_{10}$ στο δυαδικό σύστημα.

Αναζητούμε δυαδικά ψηφία $a_0, a_1, a_2, \dots$, τ.ω.

$$\begin{aligned}(253)_{10} &= (\dots a_2 a_1 a_0)_2 = \dots + a_2 \cdot 2^2 + a_1 \cdot 2^1 + a_0 \cdot 2^0 \\ &= a_0 + 2(a_1 + 2(a_2 + \dots) \dots). \quad (\text{Σχήμα του Horner})\end{aligned}$$

$$\begin{aligned}(253)_{10} &= (\dots a_2 a_1 a_0)_2 = \dots + a_2 \cdot 2^2 + a_1 \cdot 2^1 + a_0 \cdot 2^0 \\ &= \underbrace{a_0}_{\text{υπόλοιπο}} + 2 \underbrace{(a_1 + 2(a_2 + \dots) \dots)}_{\text{πηλίκιο}}.\end{aligned}$$

Άρα a_0 είναι το υπόλοιπο της διαίρεσης $253 : 2 \Rightarrow a_0 = 1$.

Το πηλίκιο της διαίρεσης είναι ο $(126)_{10}$. Επομένως

$$(126)_{10} = a_1 + 2(a_2 + \dots).$$

Άρα, τώρα, το a_1 είναι το υπόλοιπο της διαίρεσης $126 : 2$, δηλαδή $a_1 = 0$.

Συνεχίζουμε με τον τρόπο αυτό (μέχρι το πηλίκιο να γίνει μηδέν) όπως φαίνεται στον ακόλουθο πίνακα:

Μετατροπή ακεραίου από το δεκαδικό σε ένα σύστημα με βάση β .

Αριθμός στο δεκαδικό	Διαίρεση δια β	Πηλίκιο	Υπόλοιπο	Ψηφίο στο σύστημα με βάση β
253	$253 : 2$	126	1	$a_0 = 1$
126	$126 : 2$	63	0	$a_1 = 0$
63	$63 : 2$	31	1	$a_2 = 1$
31	$31 : 2$	15	1	$a_3 = 1$
15	$15 : 2$	7	1	$a_4 = 1$
7	$7 : 2$	3	1	$a_5 = 1$
3	$3 : 2$	1	1	$a_6 = 1$
1	$1 : 2$	0	1	$a_7 = 1$

Επομένως $(253)_{10} = (11111101)_2$.

Μετατροπή κλασματικού από το δεκαδικό σε ένα σύστημα με βάση β .

Για να μετατρέψουμε έναν κλασματικό αριθμό x από το δεκαδικό σε ένα σύστημα με βάση β εργαζόμαστε ως εξής: Έστω

$$x = (0.a_{-1}a_{-2}a_{-3}\dots)_{\beta} = a_{-1} \cdot \beta^{-1} + a_{-2} \cdot \beta^{-2} + a_{-3} \cdot \beta^{-3} \dots$$

Πολλαπλασιάζουμε και τα δύο μέλη επί β , οπότε

$$\beta x = (a_{-1}.a_{-2}a_{-3}\dots)_{\beta}$$

Άρα το a_{-1} είναι το ακέραιο μέρος του βx και το $(0.a_{-2}a_{-3}\dots)_{\beta}$ το κλασματικό μέρος.

Συνεχίζουμε με αυτό τον τρόπο όπως θα δούμε στο επόμενο παράδειγμα.

Μετατροπή κλασματικού από το δεκαδικό σε ένα σύστημα με βάση β .

Θα μετατρέψουμε τον αριθμό $(0.325)_{10}$ στο δυαδικό σύστημα.

Αναζητούμε ψηφία $a_{-1}, a_{-2}, \dots$, τέτοια ώστε $(0.325)_{10} = (0.a_{-1}a_{-2}\dots)_2$.

Αριθμός στο δεκαδικό	Πολ/σμός επί β	Ακέραιο μέρος	Κλασματικό μέρος	Ψηφίο στο σύστημα με βάση β
0.325	$0.325 \cdot 2 = 0.65$	0	0.65	$a_{-1} = 0$
0.65	$0.65 \cdot 2 = 1.3$	1	0.3	$a_{-2} = 1$
0.3	$0.3 \cdot 2 = 0.6$	0	0.6	$a_{-3} = 0$
0.6	$0.6 \cdot 2 = 1.2$	1	0.2	$a_{-4} = 1$
0.2	$0.2 \cdot 2 = 0.4$	0	0.4	$a_{-5} = 0$
0.4	$0.4 \cdot 2 = 0.8$	0	0.8	$a_{-6} = 0$
0.8	$0.8 \cdot 2 = 1.6$	1	0.6	$a_{-7} = 1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Επομένως $(0.325)_{10} = (0.0101001\dots)_2$.

Παρατηρήσεις

- Προφανώς ένας **ακέραιος** σε ένα σύστημα αρίθμησης παραμένει ακέραιος αν τον μετατρέψουμε σε ένα σύστημα με άλλη βάση.
- Ένας **κλασματικός** παραμένει κλασματικός αλλά το πλήθος των ψηφίων του μπορεί από πεπερασμένο να γίνει άπειρο ή αντίστροφα.

Π.χ. αν προσπαθήσουμε να μετατρέψουμε το κλάσμα $1/10$ στο δυαδικό θα διαπιστώσουμε ότι απαιτείται ο υπολογισμός μιας άπειρης σειράς.

Συγκεκριμένα, ισχύει ότι

$$\frac{1}{10} = \sum_{k=1}^{\infty} \left(\frac{1}{2^{4k}} + \frac{1}{2^{4k+1}} \right),$$

επομένως

$$(0.1)_{10} = (0.0001\ 1001\ 1001\ \dots)_2.$$

Απόδειξη ότι: $\sum_{n=1}^{\infty} \left(\frac{1}{2^{4n}} + \frac{1}{2^{4n+1}} \right) = \frac{1}{10}$

Reminder: Γεωμετρική σειρά
 $\sum_{k=0}^{\infty} ar^k = \frac{a}{1-r}, |r| < 1$

$$\sum_{n=1}^{\infty} \frac{1}{2^{4n}} = \sum_{n=1}^{\infty} \frac{1}{16^n} = \sum_{n=0}^{\infty} \frac{1}{16^n} - 1 = \frac{1}{1 - \frac{1}{16}} - 1 = \frac{16}{15} - 1 = \frac{1}{15}$$

Ανάλογα $\sum_{n=1}^{\infty} \frac{1}{2^{4n+1}} = \sum_{n=1}^{\infty} \frac{1}{2 \cdot 2^{4n}} = \frac{1}{2} \sum_{n=1}^{\infty} \frac{1}{2^{4n}} = \frac{1}{2} \cdot \frac{1}{15} = \frac{1}{30}$

$$\therefore \sum_{n=1}^{\infty} \left(\frac{1}{2^{4n}} + \frac{1}{2^{4n+1}} \right) = \frac{1}{15} + \frac{1}{30} = \frac{3}{30} = \frac{1}{10}$$

Αριθμητική κινητής υποδιαστολής – Αριθμοί μηχανής

(Κανονική) μορφή κινητής υποδιαστολής:

$$x = \pm(0.d_1d_2 \dots) \cdot \beta^e, \quad \text{όπου } 0 \leq d_i \leq \beta - 1 \text{ και } d_1 \neq 0.$$

Το e είναι κατάλληλος ακέραιος που λέγεται **εκθέτης**, ενώ ο αριθμός $m = (0.d_1d_2 \dots)$ λέγεται **κλασματικό μέρος** (fractional part ή mantissa).

$$\text{Π.χ. } -(0.00598)_{10} = -0.598 \cdot 10^{-2}, \quad (111.001)_2 = .111001 \cdot 2^3$$



Για την παράσταση ενός πραγματικού αριθμού χρειάζονται, γενικά, άπειρα ψηφία, τα οποία φυσικά δεν μπορούν να αποθηκευθούν στην **πεπερασμένη** μνήμη ενός υπολογιστή. Ο υπολογιστής **προσεγγίζει** έναν αριθμό με πεπερασμένη ακρίβεια, συγκεκριμένα με κάποιον από τους λεγόμενους **αριθμούς μηχανής** ($\subset \mathbb{Q}$).

Αριθμοί μηχανής

Κάθε αριθμός μηχανής $x \neq 0$ είναι της μορφής

$$x = \pm.d_1d_2 \dots d_t \cdot \beta^e,$$

όπου

- β είναι η **βάση** του αριθμ. συστήματος
- t είναι η **ακρίβεια**, δηλ. το πλήθος των ψηφίων του κλάσματος των αριθμών
- e είναι ο **εκθέτης**, και είναι ακέραιος με $L \leq e \leq U$. ($L \approx -U$)

Συμβολίζουμε με $M = M(\beta, t, L, U)$ το σύνολο των αριθμών μηχανής.

Μεγαλύτερος αριθμός μηχανής: $x = .d_1d_2 \dots d_t \cdot \beta^U$, με $d_i = \beta - 1$.

Μικρότερος θετικός αριθμός μηχανής: $x = .10 \dots 0 \cdot \beta^L$

Παρατηρήσεις

Δεν θα ασχοληθούμε με τον τρόπο αποθήκευσης των αριθμών μηχανής στον υπολογιστή, δηλαδή με το πώς αποθηκεύεται ο εκθέτης, το κλάσμα, το πρόσημο κ.λ.π.

Σημειώστε ότι πριν από μερικές δεκαετίες η κατάσταση ήταν πολύ μπερδεμένη. Κάθε υπολογιστής είχε το δικό του αριθμητικό σύστημα κινητής υποδιαστολής (δεκαδικό, δυαδικό, ...) και διαφορετική ακρίβεια.

Υπήρχε ακόμα και ένας ρώσικος υπολογιστής με τριαδικό σύστημα! (Setun, 1958)



ANSI/IEEE Standard 754-1985 for Binary Floating-Point Arithmetic

IEEE	β	t	L	U	β^{1-t}
Απλή ακρίβεια	2	24	-125	128	1.2×10^{-7}
Διπλή ακρίβεια	2	53	-1021	1024	2.2×10^{-16}
Επεκτεταμένη διπλή ακρίβεια	2	64	-16381	16384	1.1×10^{-19}

Όλοι οι υπολογιστές που έχουν σχεδιαστεί από το 1985 και μετά χρησιμοποιούν το δυαδικό σύστημα κινητής υποδιαστολής IEEE.

Αυτό δεν σημαίνει ότι όλοι δίνουν ακριβώς τα ίδια αποτελέσματα, αφού υπάρχει κάποια ευελιξία μέσα στο πρότυπο αυτό, αλλά μας δίνει πλέον τη δυνατότητα να μελετήσουμε τη συμπεριφορά ενός συστήματος κινητής υποδιαστολής ανεξάρτητα από το συγκεκριμένο υπολογιστικό σύστημα στο οποίο δουλεύουμε κάθε φορά.

Παράσταση αριθμών στον υπολογιστή

Αν $|x|$ μεγαλύτερο από τον μεγαλύτερο αριθμό μηχανής $\rightarrow$ υπερχείλιση (overflow)

Αν $|x| < 0.1 \cdot \beta^L \rightarrow$ υπεκχείλιση (underflow)

Προσέγγιση πραγματικών αριθμών με αριθμούς μηχανής:

Για να παρασταθεί στον υπολογιστή ένας πραγματικός αριθμός x με $0.1 \cdot \beta^L \leq |x| \leq \beta^U$, προσεγγίζεται από έναν αριθμό μηχανής που βρίσκεται «κοντά» στον x , και τον οποίο συμβολίζουμε με $\text{fl}(x)$.

Η προσέγγιση του x από τον αριθμό μηχανής $\text{fl}(x)$ μπορεί να γίνει:

- α) Με **στρογγύλευση**, όπου επιλέγεται ως $\text{fl}(x)$ ο πλησιέστερος προς τον x αριθμός μηχανής. (Στην πράξη γίνεται στρογγύλευση του t -οστού ψηφίου του x προς τα πάνω ή κάτω.)
- β) Με **αποκοπή**, όπου αποκόπτουμε όλα τα ψηφία του κλάσματος μετά το t -οστό.

Παράδειγμα

Έστω ένας Η/Υ με παραμέτρους $\beta = 10$, $t = 5$, $U = -L = 10$. Τότε ο πραγματικός αριθμός

$$\pi = 3.141592654 \dots = 0. \underbrace{31415}_{5 \text{ ψηφία}} 92654 \dots \times 10^1$$

θα παρασταθεί στον υπολογιστή ως:

$$\begin{aligned} \text{fl}(\pi) &= 0.31416 \cdot 10^1 && \text{με στρογγύλευση,} \\ \text{fl}(\pi) &= 0.31415 \cdot 10^1 && \text{με αποκοπή.} \end{aligned}$$

...Παράσταση αριθμών στον υπολογιστή

Στη συνέχεια, θα υποθέτουμε ότι αν $x \neq 0$ είναι ένας πραγματικός αριθμός μέσα στο εύρος των αριθμών μηχανής, τότε θα προσεγγίζεται και θα παριστάνεται στον υπολογιστή από τον αριθμό μηχανής $\text{fl}(x)$.

Το **σχετικό σφάλμα** θα είναι

$$\left| \frac{\text{fl}(x) - x}{x} \right| \leq u. \quad (1)$$

Το u λέγεται **μοναδιαίο σφάλμα στρογγύλευσης** και ισούται με

$$u = \begin{cases} \frac{1}{2}\beta^{1-t} & \text{για στρογγύλευση,} \\ \beta^{1-t} & \text{για αποκοπή.} \end{cases} \quad (2)$$

Η εκτίμηση (1) είναι ισοδύναμη με τη σχέση

$$\text{fl}(x) = x(1 + \delta) \quad \text{για κάποιο } \delta = \delta(x), \text{ με } |\delta| \leq u. \quad (3)$$

Μοντέλο πράξεων στον υπολογιστή

Αν $*$ = $+$, $-$, $\times$, $\div$, και x, y είναι πραγματικοί αριθμοί που μπορούν να παρασταθούν κατά προσέγγιση από αριθμούς μηχανής, τότε θα υποθέσουμε ότι η πράξη αυτή στον υπολογιστή θα γίνει ως

$$x \circledast y = \text{fl}(\text{fl}(x) * \text{fl}(y)).$$

Π.χ., έστω Η/Υ με $\beta = 10$, $t = 5$, $U = -L = 10$ και **στρογγύλευση**.

Έστω $x = 5891.26$, $y = 0.0773414$ και θέλω να υπολογίσω το $x \oplus y$

$$(x + y = 5891.3373414)$$

Αρχικά, γράφουμε τα x, y σε κανονική μορφή κινητής υποδιαστολής:

$$x = 0.58912\mathbf{6} \cdot 10^4, \quad y = 0.77341\mathbf{4} \cdot 10^{-1}$$

Άρα

$$\text{fl}(x) = 0.58913 \cdot 10^4, \quad \text{fl}(y) = 0.77341 \cdot 10^{-1} \Rightarrow \text{fl}(x) + \text{fl}(y) = 0.58913\mathbf{77341} \cdot 10^4$$

επομένως

$$x \oplus y = \text{fl}(0.58913\mathbf{77341} \cdot 10^4) = 0.58914 \cdot 10^4 = 5891.4$$

Παράδειγμα 1

Θεωρήστε πάλι τον Η/Υ με $\beta = 10$, $t = 5$, $U = -L = 10$ και **στρογγύλευση**.

Θεωρήστε τους **αριθμούς μηχανής** $a = 1$, $b = 0.3 \cdot 10^{-4}$, $c = 0.4 \cdot 10^{-4}$.

Τότε

$$(a \oplus b) \oplus c = 1 \quad \text{ενώ} \quad a \oplus (b \oplus c) = 1.0001 !!!$$

$\therefore$ στο M δεν ισχύει η προσεταιριστική ιδιότητα.

Παράδειγμα 2

Θεωρήστε πάλι τον Η/Υ με $\beta = 10$, $t = 5$, $U = -L = 10$ και **στρογγύλευση**.

Όταν είμαστε στο $\mathbb{R}$, προφανώς, $1 + x = 1 \Leftrightarrow x = 0$.

Ας πάρουμε $x = 4 \cdot 10^{-5}$. Τότε

$$1 \oplus x = \text{fl}(\text{fl}(1) + \text{fl}(x)) = \text{fl}(1.00004) = \text{fl}(0.100004 \cdot 10^1) = 1$$

Παρατηρήστε ότι $x = 4 \cdot 10^{-5} \gg 10^{-11}$ που είναι σ' αυτόν τον υπολογιστή ο μικρότερος αριθμός μηχανής.

Γενικά, σε έναν υπολογιστή αν προσθέσουμε σε έναν αριθμό της τάξεως του 1 μια ποσότητα που είναι κατ' απόλυτο τιμή μικρότερη του $\frac{1}{2}\beta^{1-t}$ ο αριθμός δεν αλλάζει.

Το όριο αυτό λέγεται **μηδέν της μηχανής** ή **έψιλον της μηχανής** και είναι χρήσιμο να το γνωρίζουμε στους υπολογισμούς μας.

Επιρροή των σφαλμάτων στρογγύλευσης στους υπολογισμούς

Γενικά, τα σφάλματα στρογγύλευσης μπορούν να επιφέρουν προβλήματα στους υπολογισμούς μας με έναν υπολογιστή. Παραδείγματα τέτοιων υπολογισμών είναι η αφαίρεση σχεδόν ίσων αριθμών, σφάλματα στον υπολογισμό αθροισμάτων κλπ.

Πριν προχωρήσουμε να δούμε την επιρροή των σφαλμάτων στρογγύλευσης στις βασικές πράξεις (+, −, ·, /) θα αποδείξουμε την ακόλουθη πρόταση:

Πρόταση

Αν δ_i , $1 \leq i \leq m$, είναι αριθμοί τέτοιοι ώστε $|\delta_i| \leq u < 1$, τότε υπάρχει ένας αριθμός δ με $|\delta| \leq u$ τέτοιος ώστε

$$(1 + \delta_1)(1 + \delta_2) \dots (1 + \delta_m) = (1 + \delta)^m.$$

Απόδειξη

Έχουμε ότι: $|\delta_i| \leq u$ για $i=1,2,\dots,m \Rightarrow -u \leq \delta_i \leq u$ για κάθε $i=1,2,\dots,m$
 $\Rightarrow \underbrace{1-u}_{>0} \leq 1+\delta_i \leq 1+u$

Δηλ. $\left. \begin{array}{l} 1-u \leq 1+\delta_1 \leq 1+u \\ 1-u \leq 1+\delta_2 \leq 1+u \\ \vdots \\ 1-u \leq 1+\delta_m \leq 1+u \end{array} \right\} \Rightarrow (1-u)^m \leq \underbrace{(1+\delta_1)(1+\delta_2)\dots(1+\delta_m)}_{=P} \leq (1+u)^m$

Δηλ. $(1-u)^m \leq P \leq (1+u)^m$

Θεωρούμε τη συνάρτηση $f(x) = (1+x)^m$, για $x \in [-u, u]$.

- f συνεχής στο $[-u, u]$
- $f(-u) = (1-u)^m$, $f(u) = (1+u)^m$
- $(1-u)^m \leq P \leq (1+u)^m$

ΘΕΩΡ. ΕΝΔ. ΤΙΜΗΣ $\Rightarrow$ Υπάρχει δε $[\delta]$ στο $[-u, u]$:
 $f(\delta) = P$

Δηλ. $(1+\delta)^m = (1+\delta_1) \dots (1+\delta_m)$

Εκτίμηση σχετικού σφάλματος του πολλαπλασιασμού

Θέλουμε να εκτιμήσουμε την ποσότητα: $\left| \frac{x \odot y - xy}{xy} \right|$. Φυσικά,

$$\begin{aligned} x \odot y &= \text{fl}(\text{fl}(x) \text{fl}(y)) \stackrel{(3)}{=} \text{fl}(x(1 + \delta_1) y(1 + \delta_2)) = (x(1 + \delta_1) y(1 + \delta_2))(1 + \delta_3) \\ &= xy(1 + \delta_1)(1 + \delta_2)(1 + \delta_3) = xy(1 + \delta)^3, \end{aligned}$$

για κάποιο δ με $|\delta| \leq u$. Άρα

$$\begin{aligned} \left| \frac{x \odot y - xy}{xy} \right| &= \left| \frac{xy(1 + \delta)^3 - xy}{xy} \right| = |(1 + \delta)^3 - 1| = |3\delta + 3\delta^2 + \delta^3| \\ &\leq 3|\delta| + 3\delta^2 + |\delta|^3 \leq 3u + 3u^2 + u^3 \leq 3u + \underbrace{3u^2 + u^3}_{\text{μικρό}} \end{aligned}$$

Άρα

$$\left| \frac{x \odot y - xy}{xy} \right| \approx 3u$$

Εκτίμηση σχετικού σφάλματος της διαίρεσης

Ανάλογα μπορούμε να αποδείξουμε ότι

$$\left| \frac{x \oslash y - x/y}{x/y} \right| \approx 3u$$

(Άσκηση)

Εκτίμηση σχετικού σφάλματος της πρόσθεσης-αφαίρεσης

Θέλουμε να εκτιμήσουμε την ποσότητα: $\left| \frac{x \oplus y - (x + y)}{x + y} \right|$. Φυσικά,

$$\begin{aligned} x \oplus y &= \text{fl}(\text{fl}(x) + \text{fl}(y)) = \text{fl}(x(1 + \delta_1) + y(1 + \delta_2)) \\ &= (x(1 + \delta_1) + y(1 + \delta_2))(1 + \delta_3) \\ &= x(1 + \delta_1)(1 + \delta_3) + y(1 + \delta_2)(1 + \delta_3) = x(1 + \delta_1)(1 + \delta_3) + y(1 + \delta_2)(1 + \delta_3) \\ &= x(1 + \delta)^2 + y(1 + \varepsilon)^2, \quad \text{όπου } |\delta|, |\varepsilon| \leq u. \end{aligned}$$

Άρα

$$x \oplus y = (x + y) + 2(\delta x + \varepsilon y) + \delta^2 x + \varepsilon^2 y$$

Άρα

$$x \oplus y = (x + y) + 2(\delta x + \varepsilon y) + \underbrace{\delta^2 x + \varepsilon^2 y}_{\text{μικρό}} \approx (x + y) + 2(\delta x + \varepsilon y).$$

Συνεπώς

$$\left| \frac{x \oplus y - (x + y)}{x + y} \right| \approx 2 \left| \frac{\delta x + \varepsilon y}{x + y} \right| \leq 2 \frac{|\delta||x| + |\varepsilon||y|}{|x + y|} \leq \frac{2u(|x| + |y|)}{|x + y|}$$

Εκτίμηση σχετικού σφάλματος της πρόσθεσης

Δηλαδή δείξαμε ότι

$$\left| \frac{x \oplus y - (x + y)}{x + y} \right| \leq 2u \frac{|x| + |y|}{|x + y|}$$

- Αν x, y **ομόσημοι**, τότε $|x| + |y| = |x + y|$ και το σχετικό σφάλμα κατά την πρόσθεσή τους είναι και πάλι μικρό, περίπου ίσο με $2u$.
- Αν $x \approx -y$, τότε $x - y \approx 0$ και άρα το φράγμα του σχετικού σφάλματος γίνεται πολύ μεγάλο.

Άρα στην **αφαίρεση σχεδόν ίσων αριθμών**, περιμένουμε καταστροφική επιρροή του σφάλματος στρογγύλευσης.

Παράδειγμα

Θεωρήστε έναν Η/Υ με παραμέτρους $\beta = 10$, $t = 5$, $U = -L = 10$, που κάνει **στρογγύλευση**.

Έστω $x = 0.39134506$ και $y = -0.39117893$.

Κάνοντας τις πράξεις ακριβώς έχουμε: $x + y = 0.00016613 = 0.16613 \cdot 10^{-3}$.

Στον Η/Υ η πρόσθεση αυτή γίνεται ως εξής:

$$x \oplus y = \text{fl}(\text{fl}(x) + \text{fl}(y)) = \text{fl}(0.39135 - 0.39118) = 0.00017 = 0.17 \cdot 10^{-3}.$$

Άρα το σχετικό σφάλμα είναι:

$$\left| \frac{(x \oplus y) - (x + y)}{(x + y)} \right| \approx 233 \cdot 10^{-4}$$

Άρα σε αυτόν τον υπολογιστή το μοναδιαίο σφάλμα στρογγύλευσης $u = \frac{1}{2} \cdot 10^{-4}$. Δηλαδή εδώ **το σχετικό σφάλμα της πράξης είναι περίπου 466 φορές μεγαλύτερο από το μοναδιαίο σφάλμα στρογγύλευσης**, ενώ είπαμε ότι το σχετικό σφάλμα που προκύπτει όταν προσθέτουμε δύο ομόσημους αριθμούς είναι μόλις διπλάσιο του u !

Παράδειγμα (ακραίο)

Θεωρήστε και πάλι τον υπολογιστή με $\beta = 10$, $t = 5$, $U = -L = 10$ και **στρογγύλευση**.

Αν $x = 451\,852\,000$ και $y = -451\,851\,000$, τότε $x + y = 1000$.

Όμως

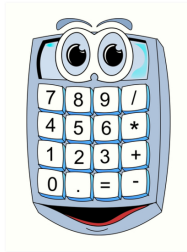
$$\text{fl}(x) = \text{fl}(0.451852 \cdot 10^9) = 0.45185 \cdot 10^9 \text{ και } \text{fl}(y) = \text{fl}(-0.451851 \cdot 10^9) = -0.45185 \cdot 10^9$$

άρα

$$x \oplus y = \text{fl}(\text{fl}(x) + \text{fl}(y)) = 0 !!!$$

Παράδειγμα

Με τον υπολογιστή χειρός θέλουμε να υπολογίσουμε την $\sqrt{7892} - \sqrt{7891}$.



$$\sqrt{7892} = 0.8883692926 \cdot 10^2, \quad \sqrt{7891} = 0.8883130079 \cdot 10^2$$

Αφαιρώντας (με το κομπιουτεράκι) έχουμε $\sqrt{7892} \ominus \sqrt{7891} = 0.5628470000 \cdot 10^{-2}$

Εναλλακτικά:

$$\sqrt{x} - \sqrt{y} = \frac{x - y}{\sqrt{x} + \sqrt{y}} = \frac{1}{\sqrt{7892} + \sqrt{7891}} = 0.5628468294 \cdot 10^{-2}$$

Αριθμητική κινητής υποδιαστολής στο MATLAB

- οι υπολογισμοί γίνονται σύμφωνα με το δυαδικό σύστημα κινητής υποδιαστολής IEEE σε **διπλή ακρίβεια** (double precision).
- Οι μη μηδενικοί αριθμοί κατ' απόλυτο τιμή είναι μεγαλύτεροι από περίπου 10^{-308} και μικρότεροι από περίπου 10^{308} , και το μοναδιαίο σφάλμα στρογγύλευσης είναι $2^{-53} \approx 1.11 \cdot 10^{-16}$
- Η συνάρτηση `eps` επιστρέφει την τιμή του λεγόμενου *μηδέν ή έψιλον της μηχανής*:

```
>> eps
ans =
    2.2204e-16
>> 1+eps
ans =
    1.0000
```


Αριθμητική κινητής υποδιαστολής στο MATLAB

```
>> realmax
ans =
    1.7977e+308
>> 1.3*realmax
ans =
    Inf
>> -2*realmax
ans =
   -Inf
>> realmin
ans =
    2.2251e-308
>> realmin*eps
ans =
    4.9407e-324
>> realmin*eps/2
ans =
    0
```

Απλή ακρίβεια στο MATLAB

```
>> eps('single')
ans =
    1.1921e-07
>> realmax('single')
ans =
    3.4028e+38
>> realmin('single')
ans =
    1.1755e-38
```

Παράδειγμα (Moler)

```
>> format long
>> a = 4/3;
>> b = a - 1;
>> c = 3*b;
>> e = 1 - c
e =
    2.220446049250313e-16
```

Παράδειγμα

Θεωρήστε τους αριθμούς $x = 10\,000\,002$ και $y = 10\,000\,001$. Θέλουμε να υπολογίσουμε με το MATLAB την παράσταση $1/(\sqrt{x} - \sqrt{y})$ χρησιμοποιώντας απλή και διπλή ακρίβεια:

```
>> xx = 10000002;
>> yy = 10000001;
>> x = single(xx);
>> y = single(yy);
>> r1 = 1/(sqrt(x) - sqrt(y))
r1 =
    4096
>> r2 = 1/(sqrt(xx) - sqrt(yy))
r2 =
    6.324555807119124e+03
```

Σχετικό σφάλμα:

```
>> err = abs((r1 - r2)/r2)
err =
    0.3523656
```

Παράδειγμα (συνέχεια)

Μεγάλη απώλεια ακρίβειας λόγω της αφαίρεσης των σχεδόν ίσων αριθμών $\text{sqrt}(x) = 3.1622781\text{e}+03$ και $\text{sqrt}(y) = 3.1622778\text{e}+03$.



$$\begin{aligned}x - y &= (\sqrt{x} - \sqrt{y})(\sqrt{x} + \sqrt{y}) \\ \Leftrightarrow \frac{1}{\sqrt{x} - \sqrt{y}} &= \frac{\sqrt{x} + \sqrt{y}}{x - y} \\ \Rightarrow \frac{1}{\sqrt{10000002} - \sqrt{10000001}} &= \frac{\sqrt{10000002} + \sqrt{10000001}}{1}.\end{aligned}$$

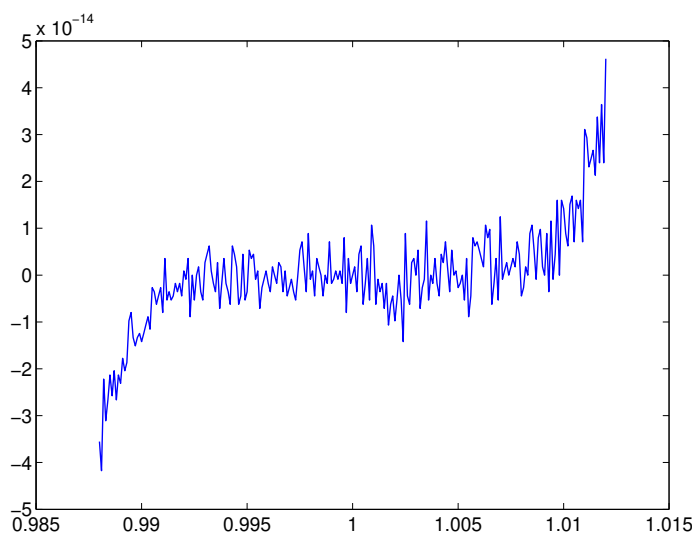
```
>> sqrt(x)+sqrt(y)
ans =
    6.3245557e+03
>> sqrt(xx)+sqrt(yy)
ans =
    6.324555794678388e+03
```

Παράδειγμα (Moler)

Με το MATLAB σχεδιάστε τη γραφική παράσταση του πολυωνύμου 7ου βαθμού

$$p(x) = x^7 - 7x^6 + 21x^5 - 35x^4 + 35x^3 - 21x^2 + 7x - 1, \quad x \in (0.988, 1.012).$$

```
>> x = 0.988 : 0.0001 : 1.012;
>> y = x.^7-7*x.^6+21*x.^5-35*x.^4+35*x.^3-21*x.^2+7*x-1;
>> plot(x,y)
```

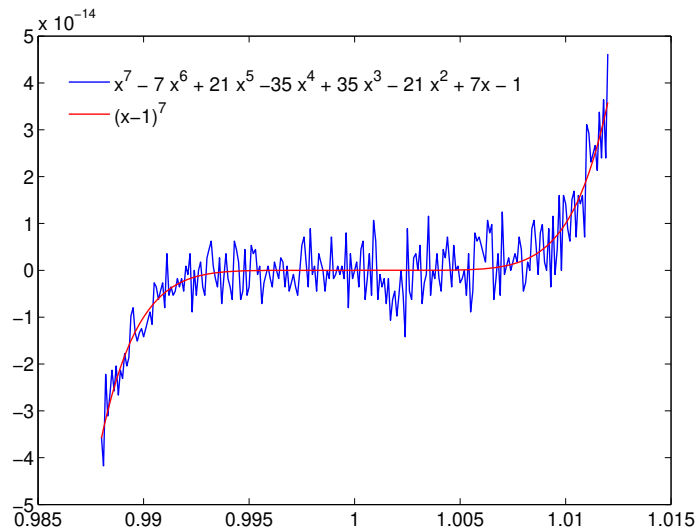


Παράδειγμα (Moler)

Όμως:

$$p(x) = x^7 - 7x^6 + 21x^5 - 35x^4 + 35x^3 - 21x^2 + 7x - 1 = (x-1)^7.$$

```
>> z = (x - 1).^7;  
>> plot(x,y,x,z,'r')
```



Ευστάθεια αλγορίθμων

Θα λέμε ότι ένας αλγόριθμος (σειρά εντολών που υλοποιούν μια αριθμητική μέθοδο) είναι **ασταθής**, αν μικρά σφάλματα που γίνονται κατά την παράσταση των αριθμών και τις πράξεις στον υπολογιστή είναι δυνατόν να επιφέρουν μεγάλες μεταβολές στο τελικό αποτέλεσμα.

Αντίθετα, αν τα τελικά αποτελέσματά του δεν επηρεάζονται πολύ από τις μικρές διαταραχές που τα σφάλματα στρογγύλευσης προκαλούν σε κάθε βήμα του, τότε λέμε ότι ο αλγόριθμος είναι **ευσταθής**.

Παράδειγμα

Θέλουμε να υπολογίσουμε το ολοκλήρωμα

$$I_n = \int_0^1 x^n e^{x-1} dx$$

Παρατηρήστε ότι $I_n \leq \int_0^1 x^n dx = \frac{1}{n+1} \rightarrow 0$.

Επίσης, αν ολοκληρώσουμε κατά μέρη έχουμε

$$\begin{aligned} I_n &= \int_0^1 x^n e^{x-1} dx = 1 - n \int_0^1 x^{n-1} e^{x-1} dx = 1 - nI_{n-1} \\ \Rightarrow \quad I_n &= 1 - nI_{n-1} \end{aligned}$$

Αλγόριθμος (αναδρομικός τύπος):

$$I_1 = \frac{1}{e}, \quad I_n = 1 - nI_{n-1}, \quad n = 2, 3, \dots$$

n	$\tilde{I}_n$
1	0.367879
2	0.264241
$\vdots$	$\vdots$
12	0.071773
13	0.066947
14	0.062736
15	0.058961
16	0.056621
17	0.037447
18	; 0.325949
19	; -5.193040
20	;! 104.860796

Ασταθής αλγόριθμος!

Ας δοκιμάσουμε το εξής

$$I_n = 1 - nI_{n-1} \Rightarrow I_{n-1} = \frac{1 - I_n}{n}$$

Αλγόριθμος (αναδρομικός τύπος):

$$I_{30} = 0, \quad I_{n-1} = \frac{1 - I_n}{n}, \quad n = 30, 29, \dots$$

n	$\tilde{I}_n$
30	0.
29	0.033333
$\vdots$	$\vdots$
19	0.047723
18	0.050120
17	0.052771
16	0.055719
15	0.059018
14	0.062732
13	0.066948
12	0.071773

Ευσταθής αλγόριθμος!

Ασκήσεις

- 1 Πώς θα υπολογίζατε με ακρίβεια τις τιμές της συνάρτησης $f(x) = \sqrt{x+4} - 2$, για πολύ μικρές τιμές του x ; (Υποθέστε ότι οι πράξεις γίνονται με αριθμητική κινητής υποδιαστολής και πεπερασμένη ακρίβεια.)
- 2 Θεωρήστε έναν υπολογιστή με σύνολο αριθμών μηχανής με παραμέτρους $\beta = 10$, $t = 3$, $-L = U = 10$ και αποκοπή. Υπολογίστε την παράσταση

$$A = (345 + 0.7) - (151.499 + 193.701)$$

κάνοντας πράξεις στον παραπάνω υπολογιστή. Πόσο είναι το σχετικό σφάλμα;

- 3 Θεωρήστε έναν Η/Υ με σύνολο αριθμών μηχανής με παραμέτρους $\beta = 10$, $t = 16$, $-L = U = 100$ και στρογγύλευση. Αν x, y, z είναι θετικοί αριθμοί μηχανής, εκτιμήστε το σχετικό σφάλμα στον υπολογισμό του

$$x \odot (y \oplus z).$$

Ασκήσεις

- 4 Εκτιμήστε το σχετικό σφάλμα για τον υπολογισμό της παράστασης

$$\alpha \cdot (\beta - \gamma), \quad \alpha \neq 0, \beta \neq \gamma$$

- (i) αν α, β, γ είναι **αριθμοί μηχανής**,
(ii) αν $\alpha, \beta, \gamma \in \mathbb{R}$ (μέσα στο εύρος των αριθμών μηχανής).

- 5 Έστω x, y οποιοιδήποτε θετικοί **αριθμοί μηχανής**. Αν $x \simeq y$, πώς θα υπολογίζατε την ποσότητα $x^2 - y^2$;

α) Ως $(x \odot x) \ominus (y \odot y)$, ή β) ως $(x \ominus y) \odot (x \oplus y)$;

Αιτιολογήστε την απάντησή σας εκτιμώντας τα σχετικά σφάλματα και για τους δύο τρόπους υπολογισμού.

- 6 Θεωρήστε τη δευτεροβάθμια εξίσωση

$$x^2 - 2ax + b = 0,$$

όπου $a, b > 0$ και $a^2 \gg b$ (το a^2 είναι πολύ μεγαλύτερο του b). Προτείνετε έναν ευσταθή αλγόριθμο για τον υπολογισμό των ριζών της.