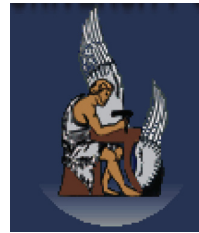


Αριθμητική Ανάλυση

Αργύρης Δελής

Διάλεξη 2η



**Σχολή Μηχανικών Παραγωγής & Διοίκησης
Πολυτεχνείο Κρήτης**

Κεφάλαιο 1ο (συνέχεια)

Αριθμοί Μηχανής (Υπολογιστή)

Η παράσταση ενός αριθμού ($x \neq 0$) στη μορφή

$$x = \pm \overbrace{(0.d_1d_2d_3 \dots d_t)}^{\text{κλάσμα (mantissa)}} \cdot \beta^e = \pm f_\beta(t) \cdot \beta^e, \quad d_1 \neq 0$$

καλείται **κινητής υποδιαστολής (floating point)**.

Αριθμοί Μηχανής (Υπολογιστή)

Η παράσταση ενός αριθμού ($x \neq 0$) στη μορφή

$$x = \pm \overbrace{(0.d_1d_2d_3 \dots d_t)}^{\text{κλάσμα (mantissa)}} \cdot \beta^e = \pm f_\beta(t) \cdot \beta^e, \quad d_1 \neq 0$$

καλείται **κινητής υποδιαστολής (floating point)**.

Ένα σύνολο αριθμών μηχανής $M(\beta, t, L, U)$ χαρακτηρίζεται από:

- τη **βάση** β του αριθμητικού συστήματος,
- την **ακρίβεια** (precision) t = πλήθος των (επιτρεπόμενων) ψηφίων του κλάσματος f_β (mantissa),
- το άνω, U , και κάτω, L , φράγμα του **εκθέτη** (exponent) $e \in \mathbb{Z}$, δηλ. $L \leq e \leq U$ με $L \simeq -U$.

$$M \equiv M(\beta, t, L, U) = \{ \text{όλοι οι αριθμοί: } x_\beta = \pm f_\beta(t) \cdot \beta^e \} \cup \{0\}$$

Πρότυπα της IEEE *	β	t	L	U	β^{1-t}
Απλή Ακρίβεια (Single Precision / Binary 32)	2	24	-125	128	1.2×10^{-7}
Διπλή Ακρίβεια (Double Precision / Binary 64)	2	53	-1021	1024	2.2×10^{-16}
Τετραπλή Ακρίβεια (Quadruple Precision / Binary 128)	2	113	-16381	16384	1.1×10^{-19}

* Σύμφωνα με τον ορισμό των αριθμών μηχανής που υιοθετήσαμε στην προηγούμενη σελίδα.

Ιδιότητες του συνόλου $M(\beta, t, L, U)$

- Για δεδομένα β, t, L, U το σύνολο M είναι πεπερασμένο και έχει κατ' απόλυτη τιμή ένα μέγιστο και ένα ελάχιστο ($\neq 0$) στοιχείο.

Ιδιότητες του συνόλου $M(\beta, t, L, U)$

- Για δεδομένα β, t, L, U το σύνολο M είναι πεπερασμένο και έχει κατ' απόλυτη τιμή ένα μέγιστο και ένα ελάχιστο ($\neq 0$) στοιχείο.

$$\min_{x \in M} \{|x_\beta|\} = (0.10 \cdots 0)_\beta \cdot \beta^L = \frac{1}{\beta} \beta^L = \beta^{L-1}$$

Ιδιότητες του συνόλου $M(\beta, t, L, U)$

- Για δεδομένα β, t, L, U το σύνολο M είναι πεπερασμένο και έχει κατ' απόλυτη τιμή ένα **μέγιστο** και ένα **ελάχιστο** ($\neq 0$) στοιχείο.

$$\min_{x \in M} \{|x_\beta|\} = (0.10 \cdots 0)_\beta \cdot \beta^L = \frac{1}{\beta} \beta^L = \beta^{L-1}$$

$$\begin{aligned} \max_{x \in M} \{|x_\beta|\} &= \left(0.(\beta - 1)(\beta - 1) \cdots (\beta - 1) \right)_\beta \cdot \beta^U \\ &= \left(\sum_{k=1}^t \frac{(\beta - 1)}{\beta^k} \cdot \right) \beta^U \\ &= \left(\sum_{k=0}^{t-1} \frac{1}{\beta^k} - \sum_{k=1}^t \frac{1}{\beta^k} \right) \cdot \beta^U = \left(1 - \frac{1}{\beta^t} \right) \beta^U \end{aligned}$$

Ιδιότητες του συνόλου $M(\beta, t, L, U)$

- Για δεδομένο εκθέτη e η απόσταση, $\Delta^{(e)}x_\beta$, μεταξύ δύο διαδοχικών αριθμών (**με τον ίδιο εκθέτη**) του M είναι σταθερή.

Έστω $x_\beta = \pm f_\beta \beta^e$ και $\tilde{x}_\beta = \pm \tilde{f}_\beta \beta^e$ τότε

$$\Delta^{(e)}x_\beta = |x_\beta - \tilde{x}_\beta| = |f_\beta - \tilde{f}_\beta| \beta^e = (0.0 \dots 01)_\beta \beta^e = \frac{1}{\beta^t} \beta^e = \beta^{-t} \beta^e$$

Ιδιότητες του συνόλου $M(\beta, t, L, U)$

- Για δεδομένο εκθέτη e η απόσταση, $\Delta^{(e)}x_\beta$, μεταξύ δύο διαδοχικών αριθμών **(με τον ίδιο εκθέτη)** του M είναι σταθερή.

Έστω $x_\beta = \pm f_\beta \beta^e$ και $\tilde{x}_\beta = \pm \tilde{f}_\beta \beta^e$ τότε

$$\Delta^{(e)}x_\beta = |x_\beta - \tilde{x}_\beta| = |f_\beta - \tilde{f}_\beta| \beta^e = (0.0 \dots 01)_\beta \beta^e = \frac{1}{\beta^t} \beta^e = \beta^{-t} \beta^e$$

- Η απόσταση μεταξύ δύο (οποιονδήποτε) διαδοχικών στοιχείων του M δεν είναι σταθερή, λόγω της μεταβολής του εκθέτη e .

Έστω $M(2, 4, -1, 3)$ τότε:

$$\Delta^{(2)}x_2 = (0.0001)2^2 = \frac{1}{2^4}2^2 = \frac{1}{4}, \quad \Delta^{(3)}x_2 = (0.0001)2^3 = \frac{1}{2^4}2^3 = \frac{1}{2}$$

Ιδιότητες του συνόλου $M(\beta, t, L, U)$

- Οι παραστάσιμοι αριθμοί είναι πυκνά κατανεμημένοι κοντά στο μηδέν και αραιά κατανεμημένοι μακριά από το μηδέν.

Ιδιότητες του συνόλου $M(\beta, t, L, U)$

- Οι παραστάσιμοι αριθμοί είναι πυκνά κατανεμημένοι κοντά στο μηδέν και αραιά κατανεμημένοι μακριά από το μηδέν.
- Η διάθεση περισσότερων ψηφίων, t , για την παράσταση της mantissa f , αυξάνει την πυκνότητα των αριθμών μηχανής, ενώ για αύξηση του εκθέτη, e , έχει σαν αποτέλεσμα την αύξηση του διαστήματος παραστασής τους.

Ιδιότητες του συνόλου $M(\beta, t, L, U)$

- Οι παραστάσιμοι αριθμοί είναι πυκνά κατανεμημένοι κοντά στο μηδέν και αραιά κατανεμημένοι μακριά από το μηδέν.
- Η διάθεση περισσότερων ψηφίων, t , για την παράσταση της mantissa f , αυξάνει την πυκνότητα των αριθμών μηχανής, ενώ για αύξηση του εκθέτη, e , έχει σαν αποτέλεσμα την αύξηση του διαστήματος παραστασής τους.
- Το πλήθος των αριθμών με τον ίδιο εκθέτη, e , είναι ανεξάρτητο του e και ίσο με $2(\beta - 1)\beta^{t-1}$.

Το πλήθος **όλων** των στοιχείων του M είναι ίσο με

$$2(U - L + 1)(\beta - 1)\beta^{t-1}.$$

Ιδιότητες του συνόλου $M(\beta, t, L, U)$

- Οι παραστάσιμοι αριθμοί είναι πυκνά κατανεμημένοι κοντά στο μηδέν και αραιά κατανεμημένοι μακριά από το μηδέν.
- Η διάθεση περισσότερων ψηφίων, t , για την παράσταση της mantissa f , αυξάνει την πυκνότητα των αριθμών μηχανής, ενώ για αύξηση του εκθέτη, e , έχει σαν αποτέλεσμα την αύξηση του διαστήματος παραστασής τους.
- Το πλήθος των αριθμών με τον ίδιο εκθέτη, e , είναι ανεξάρτητο του e και ίσο με $2(\beta - 1)\beta^{t-1}$.

Το πλήθος **όλων** των στοιχείων του M είναι ίσο με

$$2(U - L + 1)(\beta - 1)\beta^{t-1}.$$

- Το σύνολο M **δεν είναι κλειστό** ως προς τον πολλαπλασιασμό και την πρόσθεση και επομένως δεν αποτελεί υπόχωρο του $\mathbb{R}$ (για παράδειγμα πολλαπλασιάστε το $0.1 \cdot \beta^L$ με τον εαυτό του).

Ιδιότητες του συνόλου $M(\beta, t, L, U)$

- Εάν θέλουμε να παραστήσουμε έναν αριθμό $x \in \mathbb{R}$ και

$$|x| < \min_{x \in M} \{|x_\beta|\}$$

τότε **δεν υπάρχει** αριθμός μηχανής για να τον παραστήσει και η κατάσταση χαρακτηρίζεται ως **υπεκχείλιση (underflow)**.

Ιδιότητες του συνόλου $M(\beta, t, L, U)$

- Εάν θέλουμε να παραστήσουμε έναν αριθμό $x \in \mathbb{R}$ και

$$|x| < \min_{x \in M} \{|x_\beta|\}$$

τότε **δεν υπάρχει** αριθμός μηχανής για να τον παραστήσει και η κατάσταση χαρακτηρίζεται ως **υπεκχείλιση (underflow)**.

- Εάν θέλουμε να παραστήσουμε έναν αριθμό $x \in \mathbb{R}$ και

$$|x| > \max_{x \in M} \{|x_\beta|\}$$

τότε **δεν υπάρχει** αριθμός μηχανής για να τον παραστήσει και η κατάσταση χαρακτηρίζεται ως **υερχείλιση (overflow)**.

Ιδιότητες του συνόλου $M(\beta, t, L, U)$

- Εάν θέλουμε να παραστήσουμε έναν αριθμό $x \in \mathbb{R}$ και

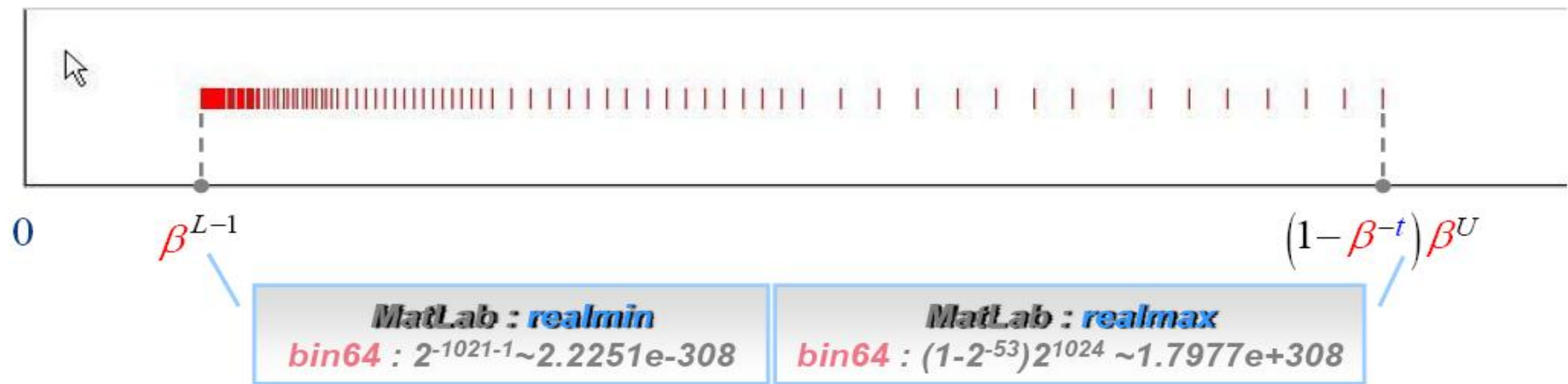
$$|x| < \min_{x \in M} \{|x_\beta|\}$$

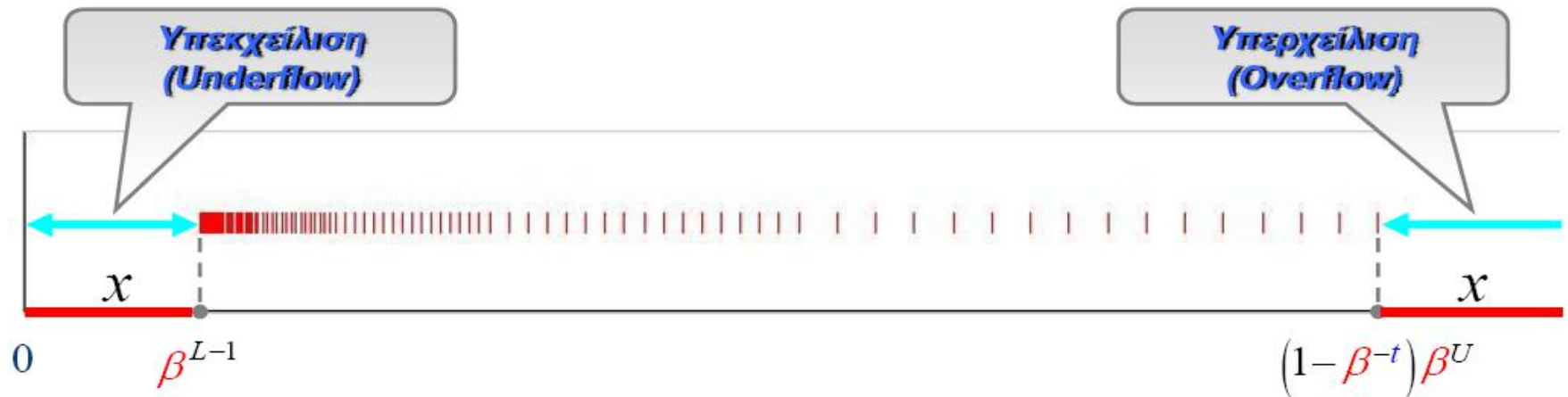
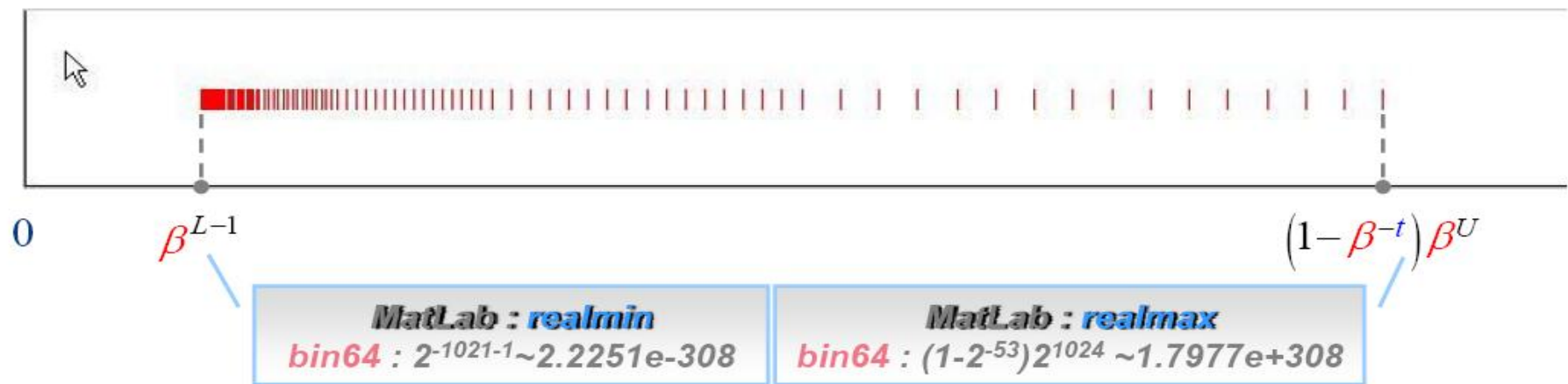
τότε **δεν υπάρχει** αριθμός μηχανής για να τον παραστήσει και η κατάσταση χαρακτηρίζεται ως **υπεκχείλιση (underflow)**.

- Εάν θέλουμε να παραστήσουμε έναν αριθμό $x \in \mathbb{R}$ και

$$|x| > \max_{x \in M} \{|x_\beta|\}$$

τότε **δεν υπάρχει** αριθμός μηχανής για να τον παραστήσει και η κατάσταση χαρακτηρίζεται ως **υερχείλιση (overflow)**.





Παράδειγμα

Αν $M \equiv M(2, 3, -1, 2)$ να βρεθούν και να παρασταθούν οι αριθμοί μηχανής.

Παράδειγμα

Αν $M \equiv M(2, 3, -1, 2)$ να βρεθούν και να παρασταθούν οι αριθμοί μηχανής.

Οι αριθμοί $\in M$ έχουν τη μορφή:

$$x = \pm f \cdot \beta^e = (0.1d_2d_3)_2 \cdot 2^e, \quad 0 \leq d_i \leq 1, i = 2, 3$$

Ο μικρότερος (θετικός) αριθμός του M είναι ο $(0.100)_2 \cdot 2^{-1} = \frac{1}{4}$.

Για δεδομένο e έχουμε τους εξής τέσσερις αριθμούς για την f

$$0.100, \quad 0.101, \quad 0.110, \quad 0.111$$

άρα για $e = -1, 0, 1, 2$ οι θετικοί αριθμοί του M είναι:

. . .Παράδειγμα. . .

$(0.100)_2 \cdot 2^{-1}$ $(\frac{1}{4})$	$(0.101)_2 \cdot 2^{-1}$ $(\frac{5}{16})$	$(0.110)_2 \cdot 2^{-1}$ $(\frac{6}{16})$	$(0.111)_2 \cdot 2^{-1}$ $(\frac{7}{16})$
$(0.100)_2 \cdot 2^0$ $(\frac{1}{2})$	$(0.101)_2 \cdot 2^0$ $(\frac{5}{8})$	$(0.110)_2 \cdot 2^0$ $(\frac{6}{8})$	$(0.111)_2 \cdot 2^0$ $(\frac{7}{8})$
$(0.100)_2 \cdot 2^1$ (1)	$(0.101)_2 \cdot 2^1$ $(\frac{5}{4})$	$(0.110)_2 \cdot 2^1$ $(\frac{6}{4})$	$(0.111)_2 \cdot 2^1$ $(\frac{7}{4})$
$(0.100)_2 \cdot 2^2$ (2)	$(0.101)_2 \cdot 2^2$ $(\frac{5}{2})$	$(0.110)_2 \cdot 2^2$ $(\frac{6}{2})$	$(0.111)_2 \cdot 2^2$ $(\frac{7}{2})$

Επιπλέον, υπάρχει το 0 και το αντίστοιχο σύνολο αρνητικών.

Η γραφική παράσταση των αριθμών είναι η ακόλουθη

. . . Παράδειγμα . . .



Σχήμα: Γραφική παράσταση των αριθμών μηχανής

Παρατηρούμε ότι δεν υπάρχουν αριθμοί στα διαστήματα $(-1/4, 0)$ και $(1/4, 0)$.

Επίσης, οι αριθμοί δεν είναι ομοιόμορφα κατανεμημένοι. Ωστόσο οι αριθμοί με κοινό εκθέτη έχουν ίση απόσταση μεταξύ τους.

Σφάλματα στη παράσταση των αριθμών μηχανής

Έστω το σύνολο $M(\beta, t, L, U)$ και ο πραγματικός αριθμός κινητής υποδιαστολής

$$x = \overbrace{(0.d_1d_2d_3 \dots d_t d_{t+1} \dots)}^q \cdot \beta^k = q \cdot \beta^k.$$

Το μέγιστο πλήθος (επιτρεπόμενων) ψηφείων είναι t , άρα ο αριθμός αυτός δεν μπορεί να παρασταθεί στη μνήμη.

Ερώτημα: **ποιός είναι ο πλησιέστερος αριθμός μηχανής $\in M$ προς τον x ;**

Σφάλματα στη παράσταση των αριθμών μηχανής

Έστω το σύνολο $M(\beta, t, L, U)$ και ο πραγματικός αριθμός κινητής υποδιαστολής

$$x = \overbrace{(0.d_1d_2d_3 \dots d_t d_{t+1} \dots)}^q \cdot \beta^k = q \cdot \beta^k.$$

Το μέγιστο πλήθος (επιτρεπόμενων) ψηφείων είναι t , άρα ο αριθμός αυτός δεν μπορεί να παρασταθεί στη μνήμη.

Ερώτημα: **ποιός είναι ο πλησιέστερος αριθμός μηχανής $\in M$ προς τον x ;**

Ο $x \notin M$ θα πρέπει να **προσεγγιστεί** από έναν αριθμό ο οποίος **συμβολίζεται** $fl(x)$ και $\in M$. Η προσέγγιση αυτή εισάγει κάποιο **σφάλμα**.
Θέλουμε συνεπώς να ισχύει

$$|x - fl(x)| \leq |x - y|, \quad \forall y \in M.$$

Σφάλματα στη παράσταση των αριθμών μηχανής

Έστω το σύνολο $M(\beta, t, L, U)$ και ο πραγματικός αριθμός κινητής υποδιαστολής

$$x = \overbrace{(0.d_1d_2d_3 \dots d_t d_{t+1} \dots)}^q \cdot \beta^k = q \cdot \beta^k.$$

Το μέγιστο πλήθος (επιτρεπόμενων) ψηφείων είναι t , άρα ο αριθμός αυτός δεν μπορεί να παρασταθεί στη μνήμη.

Ερώτημα: **ποιός είναι ο πλησιέστερος αριθμός μηχανής $\in M$ προς τον x ;**

Ο $x \notin M$ θα πρέπει να **προσεγγιστεί** από έναν αριθμό ο οποίος **συμβολίζεται** $fl(x)$ και $\in M$. Η προσέγγιση αυτή εισάγει κάποιο **σφάλμα**.
Θέλουμε συνεπώς να ισχύει

$$|x - fl(x)| \leq |x - y|, \quad \forall y \in M.$$

Αν θεωρήσουμε δύο διαδοχικούς αριθμούς μηχανής x', x'' , μεταξύ των οποίων βρίσκεται ο x , τότε συμβαίνει μια από τις δύο περιπτώσεις:

Προσέγγιση και σφάλμα



Σχήμα: Δύο πιθανές θέσεις του x (x' , x'' αριθμοί μηχανής).

Στην περίπτωση **(α)** ο πλησιέστερος αριθμός μηχανής είναι ο x' και βρίσκεται αν αποκοπούν τα ψηφία $d_{t+1} \dots$ από το δεκαδικό τμήμα του x , δηλ.

$$fl(x) = x' = (0.d_1d_2d_3 \dots d_t) \cdot \beta^k,$$

Στην περίπτωση **(β)** ο πλησιέστερος αριθμός μηχανής είναι ο x'' και

βρίσκεται αν προστεθεί η ποσότητα $(0.\overbrace{00 \dots 01}^t) = \beta^{-t}$ στο t δεκαδικό ψηφείο του x , δηλ.

$$fl(x) = x'' = \left((0.d_1d_2d_3 \dots d_t) + \beta^{-t} \right) \cdot \beta^k,$$

. . .Προσέγγιση και σφάλμα. . .

Ισχύουν τα εξής αν εφαρμόσουμε **στρογγύλευση**, την (β) περίπτωση:

$$|x - x''| = |x - fl(x)| \leq \frac{1}{2}|x' - x''| = \frac{1}{2}\beta^{k-t}, \quad (\text{απόσταση})$$

και

$$\frac{1}{\beta} \leq 0.d_1 \leq q < 1, \quad (\text{κλάσμα}) \Rightarrow |x| = q\beta^k \geq \beta^{k-1}$$

. . .Προσέγγιση και σφάλμα. . .

Ισχύουν τα εξής αν εφαρμόσουμε **στρογγύλευση**, την (β) περίπτωση:

$$|x - x''| = |x - fl(x)| \leq \frac{1}{2}|x' - x''| = \frac{1}{2}\beta^{k-t}, \quad (\text{απόσταση})$$

και

$$\frac{1}{\beta} \leq 0.d_1 \leq q < 1, \quad (\text{κλάσμα}) \Rightarrow |x| = q\beta^k \geq \beta^{k-1}$$

Συνολικά το απόλυτο **σχετικό σφάλμα** στην παράσταση του x από το M

$$\frac{|fl(x) - x|}{|x|} \leq \frac{1}{2} \frac{|x' - x''|}{|x|} \leq \frac{1}{2}\beta^{1-t} \quad (1)$$

Ο τύπος (1) μας δίνει ένα **άνω φράγμα** του σφάλματος.

Τα ίδια ισχύουν και για την περίπτωση (α) του σχήματος.

. . .Προσέγγιση και σφάλμα. . .

Γενικότερα **στρογγύλευση** (rounding) λέγεται η επιλογή του $fl(x)$ ως του πλησιέστερου αριθμού μηχανής προς τον x και στη πράξη γίνεται με στρογγύλευση του t —ψηφίου του x προς τα πάνω η κάτω.

. . .Προσέγγιση και σφάλμα. . .

Γενικότερα **στρογγύλευση** (rounding) λέγεται η επιλογή του $fl(x)$ ως του πλησιέστερου αριθμού μηχανής προς τον x και στη πράξη γίνεται με στρογγύλευση του t —ψηφίου του x προς τα πάνω η κάτω.

Μία άλλη τεχνική είναι αυτή της **αποκοπής** (chopping) όπου ο x προσεγγίζεται πάντα από τα αριστερά, δηλαδή από τον x' (βλ. προηγ. σχήμα). Τότε το ανάλογο σφάλμα είναι

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

. . .Προσέγγιση και σφάλμα. . .

Γενικότερα **στρογγύλευση** (rounding) λέγεται η επιλογή του $fl(x)$ ως του πλησιέστερου αριθμού μηχανής προς τον x και στη πράξη γίνεται με στρογγύλευση του t -ψηφίου του x προς τα πάνω η κάτω.

Μία άλλη τεχνική είναι αυτή της **αποκοπής** (chopping) όπου ο x προσεγγίζεται πάντα από τα αριστερά, δηλαδή από τον x' (βλ. προηγ. σχήμα). Τότε το ανάλογο σφάλμα είναι

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

Παράδειγμα: για $\beta = 10$ και $t = 5$ αν $x = (.a_1a_2a_3a_4a_5a_6 \cdots)10^k$, τότε

$fl(x) = x'' = ((.a_1a_2a_3a_4a_5) + 10^{-5})10^k$ (στρογγύλευση, αν $a_6 \geq 5$)

$fl(x) = x' = (.a_1a_2a_3a_4a_5)10^k$ (αποκοπή, αν $a_6 < 5$)

Σχετικό Σφάλμα Προσέγγισης (ενός αριθμού στον υπολογιστή)

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

η ποσότητα u ονομάζεται **μοναδιαίο σφάλμα στρογγύλευσης**.

Σχετικό Σφάλμα Προσέγγισης (ενός αριθμού στον υπολογιστή)

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

η ποσότητα u ονομάζεται **μοναδιαίο σφάλμα στρογγύλευσης**.

- Αν θέσουμε

$$\frac{fl(x) - x}{x} = \epsilon \implies fl(x) = x(1 + \epsilon), \quad \epsilon = \epsilon(x), \quad |\epsilon| \leq u \ll 1$$

Η «πολύ μικρή» ποσότητα ϵ καλείται μονάδα μηχανής και η παραπάνω σχέση δηλώνει ότι ο $fl(x)$ είναι μία ελαφριά διαταραχή του x .

Σχετικό Σφάλμα Προσέγγισης (ενός αριθμού στον υπολογιστή)

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

η ποσότητα u ονομάζεται **μοναδιαίο σφάλμα στρογγύλευσης**.

- Αν θέσουμε

$$\frac{fl(x) - x}{x} = \epsilon \implies fl(x) = x(1 + \epsilon), \quad \epsilon = \epsilon(x), \quad |\epsilon| \leq u \ll 1$$

Η «πολύ μικρή» ποσότητα ϵ καλείται μονάδα μηχανής και η παραπάνω σχέση δηλώνει ότι ο $fl(x)$ είναι μία ελαφριά διαταραχή του x .

- Υπάρχουν περιπτώσεις που, λόγω **συσσώρευσης** τους π.χ. μετά από ένα μεγάλο πλήθος πράξεων, τα σφάλματα στη παράσταση των αριθμών καταστρέφουν τελείως την ακρίβεια των υπολογισμών.
- Η μελέτη των σφαλμάτων στρογγύλευσης είναι ένα σημαντικό τμήμα της Αριθμητικής Ανάλυσης.



Οι πράξεις στον υπολογιστή

Έστω $x \in \mathbb{R}$ και $fl(x) \in M(\beta, t, L, U)$, συμβολικά στον υπολογιστή (H/Y)

$$x \longmapsto \boxed{H/Y(u, M)} \longmapsto fl(x)$$

Οι πράξεις στον υπολογιστή

Έστω $x \in \mathbb{R}$ και $fl(x) \in M(\beta, t, L, U)$, συμβολικά στον υπολογιστή (H/Y)

$$x \longmapsto \boxed{H/Y(u, M)} \longmapsto fl(x)$$

Συμβολίζουμε με $\star$ την όποια από τις τέσσερις πράξεις, δηλ.

$$\star \equiv \{+, -, \times, \div\}$$

Οι πράξεις στον υπολογιστή

Έστω $x \in \mathbf{R}$ και $fl(x) \in M(\beta, t, L, U)$, συμβολικά στον υπολογιστή (H/Y)

$$x \longmapsto \boxed{\text{H/Y}(u, M)} \longmapsto fl(x)$$

Συμβολίζουμε με $\star$ την όποια από τις τέσσερις πράξεις, δηλ.

$$\star \equiv \{+, -, \times, \div\}$$

Έστω τώρα $x, y \in \mathbf{R}$ και $fl(x), fl(y) \in M(\beta, t, L, U)$

$$x \star y \longmapsto \boxed{\text{H/Y}(u, M)} \longmapsto z \equiv fl\left(fl(x) \star fl(y)\right)$$

Σχετικό σφάλμα (μίας πράξης)

$$\frac{|z - x \star y|}{|x \star y|} = \frac{|fl\left(fl(x) \star fl(y)\right) - x \star y|}{|x \star y|}$$

Ενα Παράδειγμα

Έστω (υποθετικός) υπολογιστής με $M(10, 5, -10, 10)$ και $x, y \in \mathbb{R}$ με

$$x = 5891.26 \longrightarrow fl(x) = 0.58913 \cdot 10^4 \text{ (στρογγύλευση)}$$

$$y = 0.0773414 \longrightarrow fl(y) = 0. \overbrace{77341}^{t=5} 4 \cdot 10^{-1} = 0.77341 \cdot 10^{-1} \text{ (αποκοπή)}$$

Ένα Παράδειγμα

Έστω (υποθετικός) υπολογιστής με $M(10, 5, -10, 10)$ και $x, y \in \mathbb{R}$ με

$$x = 5891.26 \longrightarrow fl(x) = 0.5891\textcolor{red}{3} \cdot 10^4 \text{ (στρογγύλευση)}$$

$$y = 0.0773414 \longrightarrow fl(y) = 0.\overset{t=5}{\overbrace{77341}} 4 \cdot 10^{-1} = 0.77341 \cdot 10^{-1} \text{ (αποκοπή)}$$

$$fl(x) + fl(y) = 0.\overbrace{58913} 77341 \cdot 10^4$$

$$z = fl\left(fl(x) + fl(y)\right) = 0.5891\textcolor{red}{4} \cdot 10^4 \text{ (ο αριθμός στον Η/Υ)}$$

Ένα Παράδειγμα

Έστω (υποθετικός) υπολογιστής με $M(10, 5, -10, 10)$ και $x, y \in \mathbb{R}$ με

$$x = 5891.26 \longrightarrow fl(x) = 0.5891\textcolor{red}{3} \cdot 10^4 \text{ (στρογγύλευση)}$$

$$y = 0.0773414 \longrightarrow fl(y) = 0.\overbrace{77341}^{t=5} 4 \cdot 10^{-1} = 0.77341 \cdot 10^{-1} \text{ (αποκοπή)}$$

$$fl(x) + fl(y) = 0.\overbrace{58913} 77341 \cdot 10^4$$

$$z = fl\left(fl(x) + fl(y)\right) = 0.5891\textcolor{red}{4} \cdot 10^4 \text{ (ο αριθμός στον Η/Υ)}$$

$$\text{Όμως } \textcolor{red}{x} + \textcolor{red}{y} = \textcolor{red}{5891.3373414} (\neq z \neq fl(x + y) \neq fl(x) + fl(y))$$

Ένα Παράδειγμα

Έστω (υποθετικός) υπολογιστής με $M(10, 5, -10, 10)$ και $x, y \in \mathbb{R}$ με

$$x = 5891.26 \longrightarrow fl(x) = 0.5891\textcolor{red}{3} \cdot 10^4 \text{ (στρογγύλευση)}$$

$$y = 0.0773414 \longrightarrow fl(y) = 0.\overbrace{77341}^{t=5} 4 \cdot 10^{-1} = 0.77341 \cdot 10^{-1} \text{ (αποκοπή)}$$

$$fl(x) + fl(y) = 0.\overbrace{58913} 77341 \cdot 10^4$$

$$z = fl\left(fl(x) + fl(y)\right) = 0.5891\textcolor{red}{4} \cdot 10^4 \text{ (ο αριθμός στον Η/Υ)}$$

Όμως $x + y = 5891.3373414 (\neq z \neq fl(x + y) \neq fl(x) + fl(y))$

«Παράδοξα» στους υπολογισμούς

Έστω $a = 1$, $b = 3 \cdot 10^{-5}$, $c = 3 \cdot 10^{-5}$ και $M(10, 5, -10, 10)$

$$(a + b) + c = 1!$$

$$a + (b + c) = 1.0001 \text{ (δηλ. η σειρά πρόσθεσης παίζει ρόλο).}$$

«Παράδοξα» στους υπολογισμούς (συνέχεια)

Έστω ο υπολογιστής με $M(10, 5, -10, 10)$

Αν $x = 4 \cdot 10^{-5} (x \in M)$ τότε

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

«Παράδοξα» στους υπολογισμούς (συνέχεια)

Έστω ο υπολογιστής με $M(10, 5, -10, 10)$

Αν $x = 4 \cdot 10^{-5} (x \in M)$ τότε

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

$$1 + x = 1 \quad !!!$$

Αυτό θα ισχύει $\forall x \in \mathbf{R}$ με $0 < x < 5 \cdot 10^{-5}$.

«Παράδοξα» στους υπολογισμούς (συνέχεια)

Έστω ο υπολογιστής με $M(10, 5, -10, 10)$

Αν $x = 4 \cdot 10^{-5} (x \in M)$ τότε

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

$$1 + x = 1 \quad !!!$$

Αυτό θα ισχύει $\forall x \in \mathbf{R}$ με $0 < x < 5 \cdot 10^{-5}$.

Γενικά $\forall x \in \mathbf{R}$ με $0 < x < u = \frac{1}{2}\beta^{1-t}$ είναι **λύση** της $1 + x = 1$.

Στον υπολογιστή υπάρχει ένα όριο κάτω από το οποίο οι αριθμοί είναι αμελητέοι, πρακτικά «μηδέν», για την πρόσθεση.

«Παράδοξα» στους υπολογισμούς (συνέχεια)

Έστω ο υπολογιστής με $M(10, 5, -10, 10)$

Αν $x = 4 \cdot 10^{-5} (x \in M)$ τότε

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

$$1 + x = 1 \quad !!!$$

Αυτό θα ισχύει $\forall x \in \mathbf{R}$ με $0 < x < 5 \cdot 10^{-5}$.

Γενικά $\forall x \in \mathbf{R}$ με $0 < x < u = \frac{1}{2}\beta^{1-t}$ είναι **λύση** της $1 + x = 1$.

Στον υπολογιστή υπάρχει ένα όριο κάτω από το οποίο οι αριθμοί είναι αμελητέοι, πρακτικά «μηδέν», για την πρόσθεση.

Το όριο αυτό λέγεται **μηδέν ή έψιλον της μηχανής (*eps*)**

Εκτίμηση του μηδέν της μηχανής (αλγόριθμος)

Αρχικά $eps \leftarrow 1$

★ εφ'όσον $1 + eps > 1$ τότε
 $eps \leftarrow \frac{1}{2}eps$
 πήγαινε στο ★

διαφορετικά τέλος τύπωσε το eps

Εκτίμηση του μηδέν της μηχανής (αλγόριθμος)

Αρχικά $eps \leftarrow 1$

★ εφ'όσον $1 + eps > 1$ τότε
 $eps \leftarrow \frac{1}{2}eps$
 πήγαινε στο ★

διαφορετικά τέλος τύπωσε το eps

Για παράδειγμα στη MATLAB και για $\beta = 2, t = 53$ το $eps = 2.2204e - 016$.

Εκτίμηση του μηδέν της μηχανής (αλγόριθμος)

Αρχικά $eps \leftarrow 1$

★ εφ'όσον $1 + eps > 1$ τότε
 $eps \leftarrow \frac{1}{2}eps$
 πήγαινε στο ★

διαφορετικά τέλος τύπωσε το eps

Για παράδειγμα στη MATLAB και για $\beta = 2, t = 53$ το $eps = 2.2204e - 016$.

Άσκηση: Υπάρχουν αριθμοί μηχανής που λειτουργούν σαν το « ∞ της μηχανής» δηλ. ικανοποιούν την εξίσωση $x + 1 = x$;

Άσκηση: Σε κάθε υπολογιστή που έχει βάση το 2 (binary) το eps μπορεί να υπολογιστεί επίσης ως

$$\begin{aligned}a &= \frac{4}{3} \\b &= a - 1 \\c &= b + b + b \\eps &= 1 - c\end{aligned}$$

Άσκηση: Για τον υπολογισμό του $f(x) = y = (x - 1)^7$ στη MATLAB για $x \in [0.99, 1.015]$ έχουμε 2 επιλογές,

$$\gg y = (x - 1)^7$$

$$\gg y = x^7 - 7x^6 + 21x^5 - 35x^4 + 35x^3 - 21x^2 + 7x - 1$$

Σε ποιό τρόπο υπολογισμού αντιστοιχεί το κάθε γράφημα;

