Αριθμητική Ανάλυση, Εισαγωγή, Βραχάτης
Αριθμητική Ανάλυση, Σοφιανός, Τυχόπουλος } Εύδοξος
Εισαγωγή στην Αριθμητική Ανάλυση, Ακρίβης, Δουγαλής

1. Αριθμητική

Αν $\beta \geq 2$ είναι η "βάση" ενός συστήματος για την παράσταση αριθμών και $\alpha_k = 0, 1, \dots, \beta-1$ τα αντίστοιχα "ψηφία" τότε το σύμβολο $\pm(\alpha_N \cdots \alpha_0 \,.\, \alpha_{-1}\alpha_{-2}\cdots)_\beta$ παριστάνει τον αριθμό $\pm\sum_{k=-\infty}^{N} \alpha_k \beta^k$ στο δεκαδικό σύστημα. Η τελεία ονομάζεται υποδιαστολή.

Αν $\beta = 2, 3, 8, 10, 16, \dots$ λέμε για δυαδικό, τριαδικό, οκταδικό, δεκαδικό, δεκαεξαδικό σύστημα αριθμών. Πεπερασμένη αναπαράσταση (finite representation) αν $\exists k_{min}: \alpha_k = 0,\ \forall k < k_{min}$

π.χ. $\{100110.11\}_2 = 1\cdot 2^5 + 0\cdot 2^4 + 0\cdot 2^3 + 1\cdot 2^2 + 1\cdot 2^1 + 0\cdot 2^0 + 1\cdot 2^{-1} + 1\cdot 2^{-2} = 38.75_{10} = 38.75$

π.χ. $53473_8 = 5\cdot 8^4 + 3\cdot 8^3 + 4\cdot 8^2 + 7\cdot 8^1 + 3\cdot 8^0 = 22331_{10}$

ή ακόμα

$$53473_8 = 3 + 8\cdot(7 + 8\cdot(4 + 8\cdot(3 + 8\cdot 5))) = 3 + 7\cdot 8 + 4\cdot 8^2 + 3\cdot 8^3 + 5\cdot 8^4$$

= κάνουμε τις πράξεις ευκολότερα από "μέσα" προς τα "έξω"

Γενικά για τον ακέραιο $(\alpha_N \dots \alpha_0)_\beta = \alpha_0 + \beta\cdot(\alpha_1 + \beta\cdot(\alpha_2 + \beta(\dots + \beta(\alpha_{N-1} + \beta\alpha_N))\dots)_{10}$ (σχήμα Horner). Η πράξη που αντικαθιστά τη μεταβλητή y με $\alpha_i + \beta\cdot y$ λέγεται flop, επομένως κατά το σχήμα Horner χρειάζονται N flop.

π.χ. $307.17_8 = 3\cdot 8^2 + 0\cdot 8^1 + 7\cdot 8^0 + 1\cdot 8^{-1} + 7\cdot 8^{-2} = 199.234375_{10}$

π.χ. $0.11_2 = 1\cdot 2^{-1} + 1\cdot 2^{-2} = \frac{1}{2} + \frac{1}{4} = 0.75_{10}$

[Η παράσταση ενός αριθμού είναι περιοδική ανν ο αριθμός είναι ρητός, π.χ.

$$4.1299\ldots = 4 + 0.12 + 9\cdot 10^{-3} + 9\cdot 10^{-4} + \dots = 4.12 + 9\cdot\sum_{k=0}^{\infty} 10^{-k} - 9\cdot 10^{-0} - 9\cdot 10^{-1} - 9\cdot 10^{-2} =$$

$$= 4.12 - 0.9 - 0.09 + 9\cdot\frac{1}{1-\frac{1}{10}} - 9 = 3.13 + 1 = 4.13$$]

π.χ. $F2B_{16} = F\cdot 16^2 + 2\cdot 16^1 + B\cdot 16^0 = 15\cdot 16^2 + 2\cdot 16 + 11 = 3883_{10}$

$\{0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F\}$

Μετατροπή ακεραίου από το δεκαδικό σύστημα στο σύστημα με βάση β

π.χ. $x = 369_{10} = (\dots\alpha_2\alpha_1\alpha_0)_8 = \alpha_0 + 8\cdot(\alpha_1 + 8\cdot(\alpha_2 + \dots)\dots)$

$\Rightarrow \alpha_0 = x \bmod 8$, ~~$x:8 = 369:8 = 46+1$~~ $x = 369 \overset{1+}{=} 8\cdot 46 \Rightarrow \alpha_0 = 1$, $y_0 = 46$ πηλίκο

$\Rightarrow y_0 = \alpha_1 + 8\cdot(\alpha_2 + \dots)$, $\alpha_1 = y_0 \bmod 8$, $y_0 = 46 = 6 + 8\cdot 5 \Rightarrow \alpha_1 = 6$, $y_1 = 5$

$\Rightarrow y_1 = \alpha_2 + 8\cdot(\alpha_3 + \dots)$, $\alpha_2 = y_1 \bmod 8$, $y_1 = 5 = 5 + 8\cdot 0 \Rightarrow \alpha_2 = 5$, $y_2 = 0$

$\Rightarrow 369_{10} = 561_8$

Γενικά $\alpha_{k+1} = y_k \bmod \beta$, $k = -1, 0, \dots, N-1$, $y_k = \begin{cases} x, & k=-1 \\ \dfrac{y_{k-1} - \alpha_k}{\beta}, & k = 0, 1, \dots, N \end{cases}$

είναι ο αλγόριθμος

π.χ. $x = 111_{10}$, $k=-1: y_{-1} = x$, $\alpha_0 = y_{-1} \bmod 2 = 111 \bmod 2 = 1$

$k=0: y_0 = \frac{y_{-1} - \alpha_0}{2} = \frac{111-1}{2} = 55$, $\alpha_1 = 55 \bmod 2 = 1$

$k=1: y_1 = \frac{y_0 - \alpha_1}{2} = \frac{55-1}{2} = 27$, $\alpha_2 = 27 \bmod 2 = 1$

$k=2: y_2 = \frac{y_1 - \alpha_2}{2} = \frac{27-1}{2} = 13$, $\alpha_3 = 13 \bmod 2 = 1$

$k=3: y_3 = \frac{y_2 - \alpha_3}{2} = \frac{13-1}{2} = 6$, $\alpha_4 = 6 \bmod 2 = 0$

$k=4: y_4 = \frac{y_3 - \alpha_4}{2} = \frac{6-0}{2} = 3$
$\alpha_5 = 3 \bmod 2 = 1$

$k=5: y_5 = \frac{y_4 - \alpha_5}{2} = \frac{3-1}{2} = 1$
$\alpha_6 = 1 \bmod 2 = 1$

k	y_k		β		y_{k+1}		α_{k+1}
-1	111	:	2	=	55	+	1
0	55	:	2	=	27	+	1
1	27	:	2	=	13	+	1
2	13	:	2	=	6	+	1
3	6	:	2	=	3	+	0
4	3	:	2	=	1	+	1
5	1	:	2	=	0	+	1

$\Longrightarrow 111_{10} = 1101111_2$

Πχ. $3883_{10} \rightarrow (\ldots)_{16}$

k	y_k		β		y_{k+1}		α_{k+1}
-1	3883	:	16	=	242	+	11 = B
0	242	:	16	=	15	+	2
1	15	:	16	=	0	+	15 = F

$\Rightarrow (F2B)_{16}$

Ένας ακέραιος με βάση β_1 παραμένει ακέραιος αν εκφραστεί σε οποιοδήποτε άλλο σύστημα με βάση β_2.

Μετατροπή κλασματικού από το δεκαδικό στο σύστημα με βάση β.

π.χ. $x = 0.372_{10} = (.\alpha_{-1}\alpha_{-2}\alpha_{-3}\dots)_2 = \alpha_{-1}2^{-1} + \alpha_{-2}2^{-2} + \alpha_{-3}2^{-3} + \dots$

$\Rightarrow 2x = 0.744 = \alpha_{-1}2^{0} + \alpha_{-2}2^{-1} + \alpha_{-3}2^{-2} + \dots = (\alpha_{-1}.\alpha_{-2}\alpha_{-3}\dots)_2$

$\Rightarrow \alpha_{-1} = 0, \quad y_{-1} = 0.744 = (.\alpha_{-2}\alpha_{-3}\dots)_2 = \alpha_{-2}2^{-1} + \alpha_{-3}2^{-2} + \dots$

$\Rightarrow 2y_{-1} = 1.488 = \alpha_{-2}2^{0} + \alpha_{-3}2^{-1} + \dots = (\alpha_{-2}.\alpha_{-3}\dots)_2$

$\Rightarrow \alpha_{-2} = 1, \quad y_{-2} = 0.488 = (.\alpha_{-3}\alpha_{-4}\dots)_2 = \alpha_{-3}2^{-1} + \alpha_{-4}2^{-2} + \dots$

$\Rightarrow 2y_{-2} = 0.976 = \alpha_{-3}2^{0} + \alpha_{-4}2^{-1} + \dots = (\alpha_{-3}.\alpha_{-4}\dots)_2$

$\Rightarrow \alpha_{-3} = 0, \quad y_{-3} \overset{0.976}{=} (.\alpha_{-4}\alpha_{-5}\dots)_2 = \alpha_{-4}2^{-1} + \alpha_{-5}2^{-2} + \dots$

$\Rightarrow 2y_{-3} = 1.952 = \alpha_{-4}2^{0} + \alpha_{-5}2^{-1} + \dots = (\alpha_{-4}.\alpha_{-5}\dots)_2$

$\Rightarrow \alpha_{-4} = 1, \quad y_{-4} = 0.952 = (.\alpha_{-5}\alpha_{-6}\dots)_2$

Άρα $(0.372)_{10} = (.0101\dots)_2$

Αλγόριθμος $\alpha_{-k-1} = [\beta y_{-k}], \; k = 0, 1, \dots \qquad y_{-k} = \begin{cases} x, & k=0 \\ \beta y_{-k+1} - \alpha_{-k}, & k = 1, 2, \dots \end{cases}$

π.χ. $k=0: \; y_0 = x = 0.372, \quad \alpha_{-1} = 0$

$k=1: \; y_{-1} = 2y_0 - \alpha_{-1} = 2\cancel{y_0} = 0.744, \quad \alpha_{-2} = 1$

$k=2: \; y_{-2} = 2y_{-1} - \alpha_{-2} = 0.488, \quad \alpha_{-3} = 0$

$k=3: \; y_{-3} = 2y_{-2} - \alpha_{-3} = 0.976, \quad \alpha_{-4} = 1$

ή

k	y_{-k}	β		y_{-k-1}		α_{-k-1}
0	0.372	×2	=	0.744	+	0
1	0.744	×2	=	0.488	+	1
2	0.488	×2	=	0.976	+	0
3	0.976	×2	=	0.952	+	1

π.χ. $x = 0.59375_{10} \rightarrow (\ldots)_2$

k	y_{-k}	β		y_{-k-1}		α_{-k-1}
0	0.59375	$\times 2$	$=$	0.1875	$+$	1
1	0.1875	$\times 2$	$=$	0.375	$+$	0
2	0.375	$\times 2$	$=$	0.75	$+$	0
3	0.75	$\times 2$	$=$	0.5	$+$	1
4	0.5	$\times 2$	$=$	0	$+$	1

$\Rightarrow \{0.10011\}_2$

π.χ. $x = 0.5625_{10} \rightarrow (\ldots)_2$

k	y_{-k}	β		y_{-k-1}		α_{-k-1}
0	0.5625	$\times 2$	$=$	0.125	$+$	1
1	0.125	$\times 2$	$=$	0.25	$+$	0
2	0.25	$\times 2$	$=$	0.5	$+$	0
3	0.5	$\times 2$	$=$	0	$+$	1

$\Rightarrow \{0.1001\}_2$

π.χ. $x = 0.890625_{10} \longrightarrow (\ldots)_8$

k	y_{-k}	β		y_{-k-1}		α_{-k-1}
0	0.890625	$\times 8$	$=$	0.125	$+$	7
1	0.125	$\times 8$	$=$	0	$+$	1

$\Rightarrow \{0.71\}_8$

Αν το πλήθος των ψηφίων του x είναι πεπερασμένο σε κάποιο σύστημα, τότε προφανώς ο x είναι ρητός. Αν λοιπόν σε κάποιο άλλο σύστημα το πλήθος των ψηφίων του είναι άπειρο, αναγκαστικά θα είναι περιοδικός.

π.χ. $0.1_{10} = \frac{1}{10} = \frac{3}{2} \cdot \frac{1}{15} = \frac{3}{2}\left(\frac{16}{15} - 1\right) = \frac{3}{2}\left(\sum_{n=0}^{\infty} \frac{1}{16^n} - 1\right) = \frac{3}{2} \sum_{n=1}^{\infty} \frac{1}{2^{4n}} =$

$$= \sum_{n=1}^{\infty} \frac{1}{2^{4n}} + \sum_{n=1}^{\infty} \frac{1}{2^{4n+1}} = 2^{-4} + 2^{-5} + 2^{-8} + 2^{-9} + 2^{-12} + 2^{-13} + \ldots =$$

$$= (0.000110011001100\ldots)_2$$

Προφανώς, κάθε άρρητος x έχει άπειρα μη περιοδικά ψηφία σε κάθε σύστημα. (γιατί αν σε κάποιο σύστημα ήταν περιοδικά ή πεπερασμένα θα ήταν ρητός)

4a

Αν πάρουμε προσέγγιση (πάρει υπολογισμούς) $x = (0.0001100011)_2 = 2^{-4} + 2^{-5} + 2^{-8} + 2^{-9} = (0.099609375)_{10}$

Αν πάρουμε $x = (0.00011001100011)_2 = 2^{-4} + 2^{-5} + 2^{-8} + 2^{-9} + 2^{-12} + 2^{-13}$

$= (0.0999908447265625)_{10}$

όσο πιο πολύ ακρίβεια, τόσο περισσότερο υπολογιστικό κόστος (μνήμη για αποθήκευση ενδιάμεσων βημάτων, χρόνος...)

ή ακόμα

k	y_{-k}	β		y_{-k-1}		α_{-k-1}
0	0.1	×2	=	0.2	+	0
1	0.2	×2	=	0.4	+	0
2	0.4	×2	=	0.8	+	0
3	0.8	×2	=	0.6	+	1
4	0.6	×2	=	0.2	+	1
5	0.2	×2	=	0.4	+	0
6	0.4	×2	=	0.8	+	0
7	0.8	×2	=	0.6	+	1
8	0.6	×2	=	0.2	+	1

$\Rightarrow (0.000110011\ldots)_2$

περιοδικός

π.χ. $x = 0.3_{10} \longrightarrow (\ldots)_2$

k	y_k	β		y_{-k-1}		α_{-k-1}
0	0.3	×2	=	0.6	+	0
1	0.6	×2	=	0.2	+	1
2	0.2	×2	=	0.4	+	0
3	0.4	×2	=	0.8	+	0
4	0.8	×2	=	0.6	+	1
5	0.6	×2	=	0.2	+	1
6	0.2	×2	=	0.4	+	0
7	0.4	×2	=	0.8	+	0
8	0.8	×2	=	0.6	+	1
9	0.6	×2	=	0.2	+	1

$\Rightarrow (0.0100110011\ldots)_2$

περιοδικός

Σημαντικά ψηφία (significant digits) λέγονται όλα τα ψηφία του αριθμού εκτός των μηδενικών που βρίσκονται στην αρχή του αριθμού.

π.χ. $x=0.0007374$, έχει 4 σημαντικά ψηφία 7374 με πρώτο σημαντικό το 7

π.χ. $x=0.0997$ έχει 3 σημαντικά ψηφία 997 με πρώτο σημαντικό το 9

π.χ. $x=0.099700$ έχει 5 σημαντικά ψηφία 99700 με πρώτο το 9

π.χ. $x=410.7$ έχει 4 σημαντικά 4107 με πρώτο το 4

π.χ. $x=5.70$ έχει 3 σημαντικά 570 με πρώτο το 5

π.χ. $x=0.0079$ έχει 2 σημαντικά 79 με πρώτο το 7.

Κανονικοποιημένη αναπαράσταση (μορφή) κινητής υποδιαστολής (normalized floating point representation) $x=\pm(0.d_1d_2\ldots)\cdot\beta^E$, $d_1=1,2,\ldots,\beta-1$, $d_i=0,1,\ldots,\beta-1\ (i=2,3,\ldots)$, $E\in\mathbb{Z}$.

Προφανώς $\frac{1}{\beta}\le 0.d_1d_2\ldots\le 1$ $\left(0.d_1d_2\ldots = d_1\beta^{-1}+d_2\beta^{-2}+\ldots = \frac{d_1}{\beta}+\frac{d_2}{\beta^2}+\ldots \ge \frac{d_1}{\beta}\ge\frac{1}{\beta}\right.$, $\left.\le \frac{\beta-1}{\beta}+\frac{\beta-1}{\beta^2}+\ldots = \frac{\beta-1}{\beta}\left(1+\frac{1}{\beta}+\ldots\right)=1\right)$

Το E λέγεται εκθέτης (exponent) ή χαρακτηριστικό (characteristic)

Ο $0.d_1d_2\ldots$ λέγεται ουρά ή κλάσμα ή κλασματικό μέρος (fraction, mantissa) του x.

π.χ. $-0.00598_{10} = -0.598_{10}\times10^{-2}$, $(FBA3.DE2)_{16} = 0.FBA3DE2_{16}\times16^{4}$

$111.001_2 = 0.111001_2\times2^{3}$,

Η μετατροπή μιας μορφής κινητής υποδιαστολής από το ένα σύστημα στο άλλο γίνεται με μετατροπή του κλασματικού μέρους και επιπλέον $\beta^E=10^{E_{10}} \Leftrightarrow E\log_{10}\beta=E_{10}$.

~~Σύνολο αριθμών μηχανής~~ $\cancel{M=\{\pm0.d_1d_2\ldots d_t\cdot\beta^e}$

Σύνολο αριθμών μηχανής $M=\{\pm0.d_1d_2\ldots d_t\cdot\beta^E,\ L\le E\le U\}\cup\{0\}$

t: ακρίβεια ή μήκος

Το ελάχιστο θετικό στοιχείο του M είναι $0.10\ldots0\cdot\beta^L$ ενώ το μέγιστο στοιχείο του M είναι το $(0.\,\beta-1,\beta-1,\ldots,\beta-1)\cdot\beta^U$.

π.χ. $\beta=2$, $t=4$, $L=-1$, $U=2$ $\Rightarrow M=\{0.1000\cdot2^{-1}=\frac{1}{4},\ldots,\ 0.1111\cdot2^{2}=3.75\}\cup\{-\ldots,-\ldots\}\cup\{0\}$

[π.χ. αν $t=3$ και $x=0.000598$ και η μηχανή μπορεί να αποθηκεύσει 3 δεκαδικά ψηφία $\Rightarrow$

Το M δεν είναι σώμα ως προς την πρόσθεση και τον πολ/σμό.

π.χ. το M δεν είναι κλειστό αφού $(0.1\cdot\beta^L)(0.1\cdot\beta^L)\notin M$ $\Rightarrow$

π.χ. για $\beta=10$, $t=5$, $1=0.1\times10^{1}\in M$, $10^{-5}=0.1\times10^{-4}\in M$, αλλά $1+10^{-5}=1.00001\notin M$ (αφού έχει 6 σημαντικά ψηφία)

Υπερχείλιση (overflow) η αδυναμία παράστασης ενός αριθμού x με $|x|>\max M$

Υποχείλιση (underflow) η αδυναμία παράστασης ενός αριθμού x με $0<|x|<|\min M|=0.1\times\beta^L$

6a

⇒ τότε θα καταλήξουμε (αποθηκεύσουμε) $x = 0.000$ δηλ. κανένα σημαντικό ψηφίο. Με την αναδιατύπωση (της πράξης) θα έχουμε $x = 0.598 \times 10^{-3}$ (με σημαντικά ψηφία 5, 9, 8 και εκθέτη -3), άρα η αποθήκευση είναι ακριβέστερη.

⇒ Η αντιμεταθετική, προσεταιριστική, επιμεριστική ιδιότητα δεν ισχύουν στο M.

Αν $x = q\beta^E$, $q = 0.d_1 d_2 \dots d_t d_{t+1} \dots$, τότε οι διαδοχικοί $x', x'' \in M$ με $x' \le x < x''$ είναι $x' = 0.d_1 d_2 \dots d_t \cdot \beta^E$, $x'' = (0.d_1 d_2 \dots d_t + \beta^{-t})\beta^E$ και ισχύει $|x' - x''| = \beta^{E-t}$.

Η προσέγγιση κάποιου x από ένα "κοντινό" του αριθμό μηχανής συμβολίζεται $fl(x)$. Υπάρχουν 2 τρόποι τέτοιας προσέγγισης. Ο ένας είναι η αποκοπή (truncation), δηλαδή αποκόπτονται τα $d_{t+1} \dots$ και άρα $fl(x) = x'$, άρα $|fl(x) - x| \le |x' - x''| = \beta^{E-t}$.
Ο άλλος είναι η στρογγυλοποίηση (round-off) όπου $|fl(x) - x| \le |x - y|$, $\forall y \in M$, άρα π.χ. αν $\beta = 10$ και $d_{t+1} \ge 6$ τότε $fl(x) = x'' = (0.d_1 d_2 \dots d_t + 10^{-t})10^E$, ενώ αν $d_{t+1} \le 4$ τότε $fl(x) = x' = (0.d_1 d_2 \dots d_t) \cdot 10^E$, ενώ αν $d_{t+1} = 5$ με $d_i = 0$, $i \ge t+2$ τότε $fl(x) = x'$ ή x''. Ισχύει $|fl(x) - x| \le \frac{1}{2}|x' - x''| = \frac{1}{2}\beta^{E-t}$ [όσο μικραίνει το t, τόσο χειρότερη η προσέγγιση]
(Το σφάλμα στρογγυλοποίησης σε t δεκαδικά ψηφία είναι $\frac{1}{2}\beta^{-t}$)

π.χ. στρογγυλοποίηση του $x = 2.6457513$ σε 6, 5, 4, 3, 2, 1 δεκαδικά ψηφία
Για $t=6 \rightsquigarrow 2.645751$. Για $t=5 \rightsquigarrow 2.64575$, Για $t=4 \rightsquigarrow 2.6458$,
Για $t=3 \to 2.646$, Για $t=2 \rightsquigarrow 2.65$. Για $t=1 \rightsquigarrow 2.6$
(βέβαια σε μορφή κινητής υποδιαστολής 0.26457513×10^1 ο αριθμός των ψηφίων προσέγγισης διαφέρουν)

π.χ. στρογγυλοποίηση του $x = 3.1415926$ σε 6, 5, 4, 3, 2, 1 δεκαδικά ψηφία
Για $t=6 \rightsquigarrow 3.141593$, Για $t=5 \to 3.14159$, Για $t=4 \rightsquigarrow 3.1416$
Για $t=3 \rightsquigarrow 3.142$, Για $t=2 \rightsquigarrow 3.14$, Για $t=1 \rightsquigarrow 3.1$

Ισχύει $\left|\frac{fl(x) - x}{x}\right| \le \overset{\text{μηδενή έψιλον της μηχανής}}{u} = \begin{cases} \beta^{1-t} & \text{για αποκοπή} \\ \frac{1}{2}\beta^{1-t} & \text{για στρογγυλοποίηση} \end{cases}$

Διότι (αν $x > 0$) αν $\theta = \begin{cases} 1 & \text{για αποκοπή} \\ \frac{1}{2} & \text{για στρογγυλοποίηση} \end{cases}$ τότε $|fl(x) - x| \le \theta |x' - x''| \Rightarrow$

$\left|\frac{fl(x) - x}{x}\right| \le \theta \frac{|x' - x''|}{|x|} = \theta \frac{\beta^{E-t}}{q\beta^E} \le \theta \beta^{-t}\beta = \theta\beta^{1-t} = u$ [διότι π.χ. $q = 0.145 > 0.1 = 10^{-1} = \beta^{-1}$]

Ισχύει $fl(x) = x(1 + \epsilon)$, $|\epsilon| \le u$ (αφού αν $\epsilon = \frac{fl(x) - x}{x} \Rightarrow |\epsilon| = \left|\frac{fl(x) - x}{x}\right| \le u$)

Σφάλμα (error) $\varepsilon = \overset{\text{προσεγγιστική τιμή}}{x^*} - \overset{\text{ακριβής τιμή}}{x}$, $r = -\varepsilon = x - x^*$ διόρθωση (correction)

Απόλυτο σφάλμα $|\varepsilon| = |x^* - x|$

Σχετικό σφάλμα (relative error) $\delta = \frac{\varepsilon}{x} = \frac{x^* - x}{x} \simeq \frac{x^* - x}{x^*}$ διότι $\frac{x^* - x}{x} = \frac{x^* - x}{x^*}\frac{x^*}{x} = \frac{x^* - x}{x^*}\left(1 + \frac{\varepsilon}{x}\right) \simeq \frac{x^* - x}{x^*}$

Απόλυτο σχετικό σφάλμα (absolute relative error) $|\delta| = \left|\frac{\varepsilon}{x}\right| = \left|\frac{x^* - x}{x}\right| \simeq \left|\frac{x^* - x}{x^*}\right|$

Παράδειγμα πολλών Η.Υ.

= δυαδική θέση αποθήκευσης το 0 ή 1

1bit (0 ή 1)	31 bit	1bit	7 bit
⊔	⊔ ⊔ ⊔ . . . . ⊔	⊔	⊔ ⊔ . . . ⊔
για το πρόσημο του x	για το q δηλ., $t=31$	για το πρόσημο του E	για τον εκθέτη E

Συνολικά 40 bits για την αποθήκευση του τυχαίου αριθμού x

= 5 × (8 bits) = 5 bytes

byte = χαρακτήρας

π.χ. αν $x=4$, $x^*=3.96$ είναι $\varepsilon = x^*-x = 3.96-4=-0.04$, $r=-\varepsilon=0.04$,
$|\varepsilon|=|-0.04|=0.04$, $\delta=\frac{\varepsilon}{x}=\frac{-0.04}{4}=-0.01$, $|\delta|=|-0.01|=0.01$
ή 1%.

π.χ. $x=100\mu$, $x^*=101\mu$, $|\varepsilon|=|x^*-x|=1\mu$, $\delta=\frac{\varepsilon}{x}=\frac{1}{100}=0.01$
$y=10000\mu$, $y^*=9999\mu$, $|\tilde\varepsilon|=|y^*-y|=1\mu$, $\tilde\delta=\frac{\tilde\varepsilon}{y}=-\frac{1}{10000}=-0.000$
Άρα το σχετικό σφάλμα είναι πιο κατατοπιστικό.
Αν όμως ήταν $x=0$, τότε $\delta\to\infty$, άρα το ε πιο σημαντικό.

μη ενδεικτικό

π.χ. $x=0.0000001$, $x^*=0.0000005 \leadsto |\varepsilon|=0.0000004$, $\delta=\frac{\varepsilon}{x}=4$
$y=10000001$, $y^*=10000005 \leadsto |\tilde\varepsilon|=4$, $\tilde\delta=\frac{\tilde\varepsilon}{y}=0.0000004$
πιο ενδεικτικό

<u>Σφάλμα αθροίσματος</u> Το απόλυτο σφάλμα του αθροίσματος (διαφορά) δύο αριθμών είναι μικρότερο ή ίσο με το άθροισμα των απολύτων σφαλμάτων των αριθμών αυτών, δηλ.
$|\varepsilon|\le|\varepsilon_1|+|\varepsilon_2|$
Πράγματι, αν $\varepsilon_1=x_1^*-x_1$, $\varepsilon_2=x_2^*-x_2$ τότε
$\varepsilon=(x_1^*+x_2^*)-(x_1+x_2)=(x_1^*-x_1)+(x_2^*-x_2)=\varepsilon_1+\varepsilon_2 \Rightarrow |\varepsilon|=|\varepsilon_1+\varepsilon_2|\le|\varepsilon_1|+|\varepsilon_2|$
Επίσης $\tilde\varepsilon=(x_1^*-x_2^*)-(x_1-x_2)=(x_1^*-x_1)-(x_2^*-x_2)=\varepsilon_1-\varepsilon_2\Rightarrow|\tilde\varepsilon|=|\varepsilon_1-\varepsilon_2|\le|\varepsilon_1|+|\varepsilon_2|$

<u>Σφάλμα γινομένου</u> Το απόλυτο σχετικό σφάλμα του γινομένου (πηλίκου) δύο αριθμών είναι κατά προσέγγιση μικρότερο ή ίσο από το άθροισμα των απόλυτων σχετικών σφαλμάτων των αριθμών αυτών, δηλ. $|\delta|\le|\delta_1|+|\delta_2|$
Πράγματι, αν $\varepsilon_1=x_1^*-x_1$, $\varepsilon_2=x_2^*-x_2$, ~~και~~ $\delta_1=\frac{\varepsilon_1}{x_1}$, $\delta_2=\frac{\varepsilon_2}{x_2}$, τότε
$\varepsilon=x_1^*x_2^*-x_1x_2=(x_1+\varepsilon_1)(x_2+\varepsilon_2)-x_1x_2=\varepsilon_1x_2+\varepsilon_2x_1+\varepsilon_1\varepsilon_2\simeq\varepsilon_1x_2+\varepsilon_2x_1$
$\delta=\frac{\varepsilon}{x_1x_2}=\frac{\varepsilon_1x_2+\varepsilon_2x_1}{x_1x_2}=\frac{\varepsilon_1}{x_1}+\frac{\varepsilon_2}{x_2}=\delta_1+\delta_2\Rightarrow|\delta|=|\delta_1+\delta_2|\le|\delta_1|+|\delta_2|$

Επίσης $\tilde\varepsilon=\frac{x_1^*}{x_2^*}-\frac{x_1}{x_2}=\frac{x_1^*x_2-x_2^*x_1}{x_2^*x_2}=\frac{(x_1+\varepsilon_1)x_2-(x_2+\varepsilon_2)x_1}{x_2^*x_2}=\frac{\varepsilon_1x_2-\varepsilon_2x_1}{x_2^*x_2}$

$\delta=\frac{\tilde\varepsilon}{\frac{x_1}{x_2}}=\tilde\varepsilon\frac{x_2}{x_1}=\frac{\varepsilon_1x_2-\varepsilon_2x_1}{x_2^*x_2}\frac{x_2}{x_1}=\frac{\varepsilon_1x_2-\varepsilon_2x_1}{x_2^*x_1}\simeq\frac{\varepsilon_1x_2-\varepsilon_2x_1}{x_2x_1}=\frac{\varepsilon_1}{x_1}-\frac{\varepsilon_2}{x_2}=\delta_1-\delta_2$

$\Rightarrow|\delta|=|\delta_1-\delta_2|\le|\delta_1|+|\delta_2|$

π.χ. $x = 3.1 x_1 + 2.9 x_2$, όπου x_1, x_2 στρογγυλοποιημένοι σε 2 δεκαδικά ψηφία

$\varepsilon = x^* - x = (3.1 x_1^* + 2.9 x_2^*) - (3.1 x_1 + 2.9 x_2) =$

$= 3.1 (x_1^* - x_1) + 2.9 (x_2^* - x_2) = 3.1\, \varepsilon_1 + 2.9\, \varepsilon_2$

$|\varepsilon_1| \le \frac{1}{2} 10^{-2}$, $|\varepsilon_2| \le \frac{1}{2} 10^{-2}$

$|\varepsilon| = |3.1\, \varepsilon_1 + 2.9\, \varepsilon_2| \le 3.1 |\varepsilon_1| + 2.9 |\varepsilon_2| = 3.1 \times \frac{1}{2} 10^{-2} + 2.9 \times \frac{1}{2} 10^{-2} = 0.003$

π.χ. $a = 0.731$, $b = 9.12 \overset{= 0.912 \times 10^{1}}{}$ στρογγυλοποιημένοι σε 3 σημαντικά ψηφία τότε

$|\varepsilon_1| \le \frac{1}{2} 10^{-3}$, $|\varepsilon_2| \le \frac{1}{2} 10^{-2}$ $\left(\frac{1}{2} \beta^{e-t} = \frac{1}{2} 10^{1-3}\right)$

Το σφάλμα ε του $a+b$ έχει $|\varepsilon| \le |\varepsilon_1| + |\varepsilon_2| \le \frac{1}{2} 10^{-3} + \frac{1}{2} 10^{-2} = 0.0055$

Το απόλυτο σχετικό σφάλμα $|\delta|$ του $\frac{a}{b}$ είναι

$|\delta| \le |\delta_1| + |\delta_2| = \frac{|\varepsilon_1|}{|a|} + \frac{|\varepsilon_2|}{|b|} \le \frac{1}{|a|} \frac{1}{2} 10^{-3} + \frac{1}{|b|} \frac{1}{2} 10^{-2} = \frac{1}{0.731} \frac{10^{-3}}{2} + \frac{1}{9.12} \frac{10^{-2}}{2}$

$= 0.0012321 \le 0.0015$

Πράξεις στον υπολογιστή $\quad x * y \longrightarrow z = fl(fl(x) * fl(y))$

$*$: $+, -, \cdot, :$

π.χ. $\beta = 10$, $t = 5$, $U = -L = 10$, $fl(\cdot)$ με στρογγυλοποίηση.

$x = 5891.26$, $y = 0.0773414$

$fl(x) = 0.58913 \times 10^4$, $\quad fl(y) = 0.77341 \times 10^{-1} = 0.0000077341 \times 10^4$

$fl(x) + fl(y) = 0.5891377341 \times 10^4 = 5891.377341$

$z = fl(fl(x) + fl(y)) = 0.58914 \times 10^4 = 5891.4$

$x + y = 5891.3373414$, $\quad fl(x+y) = 0.58913 \times 10^4 = 5891.3$, $\quad fl(x) + fl(y) = \ldots$

π.χ. $a = 1$, $b = 3 \times 10^{-5}$, $c = 3 \times 10^{-5}$

$a + (b+c) \rightarrow z = fl(fl(a) + fl(fl(b) + fl(c)))$

$fl(a) = 1$, $\quad fl(b) = 0.3 \times 10^{-4} = fl(c) = 0.00003$

$fl(a) + fl(fl(b) + fl(c)) = 1.00006 = 0.100006 \times 10^{+1}$

$z = 0.10001 \times 10 = 1.0001$

$(a+b)+c \rightarrow \tilde{z} = fl(fl(fl(a) + fl(b)) + fl(c))$

$fl(a) + fl(b) = 1.00003 = 0.100003 \times 10^{+1}$, $\quad fl(fl(a) + fl(b)) = 0.10000 \times 10^1 \cancel{= 1.0000}$

$fl(fl(a) + fl(b)) + fl(c) = 0.10000 \times 10^1 + 0.000003 \times 10^1 = 0.100003 \times 10^1$

$\tilde{z} = fl(fl(fl(a) + fl(b)) + fl(c)) = 0.10000 \times 10^1 = 1.0000$

$\Rightarrow z \neq \tilde{z}$

π.χ. $1 + 4 \times 10^{-5} = 1.00004 = 0.100004 \times 10^1$

$fl(1 + 4 \times 10^{-5}) = 0.10000 \times 10^1 = 1.0000 = 1$

όμοια $fl(1+x) = 1$, $\forall\, 0 < x < 5 \times 10^{-5}$ (γενικότερα $\forall\, 0 < x < \frac{1}{2}\beta^{1-t}$ $fl(1+x) = 1$)

άρα αυτά τα x παίζουν το ρόλο του μηδενός.

Ο μικρότερος θετικός αριθμός μηχανής είναι $0.1 \times 10^{-10} = 10^{-11}$, άρα πολύ μικρότερο από το ασφαλές μηδέν της μηχανής.

Καταστροφική ακύρωση σημαντικών ψηφίων

Η αφαίρεση δύο περίπου ίσων προσεγγιστικών αριθμών οδηγεί σε μείωση της ακρίβειας του αποτελέσματος.

π.χ. $e^{-5.5} = \sum_{n=0}^{\infty} \frac{(-5.5)^n}{n!} \simeq \sum_{n=0}^{24} \frac{(-5.5)^n}{n!} = 0.0026363$ με 5 σημαντικά ψηφία, ενώ

$e^{-5.5} = 0.0040868$, δηλαδή κανένα σημαντικό ψηφίο δεν είναι σωστό.

Μπορούμε να χρησιμοποιήσουμε ακρίβεια περισσότερων σημαντικών ψηφίων που κοστίζει σε μνήμη και υπολογιστικό χρόνο.

Εναλλακτικά στο παράδειγμα μπορούμε $e^{-a} = \frac{1}{e^{a}}$ όπου δεν υπάρχει καταστροφική ακύρωση.

π.χ. $\sqrt{708} - \sqrt{707} \overset{\text{4 σημαντικά ψηφία}}{=} 26.61 - 26.59 = 0.02$ [Η πραγματική τιμή $\sqrt{708}-\sqrt{707} = 0.018797791\ldots$ και η απόκλιση της από το 0.02 είναι $0.001202909\ldots$, στη τάξη του χιλιοστού σφάλμα]

Τα σφάλματα στρογγυλοποίησης των αριθμών $\sqrt{708}$, $\sqrt{707}$ είναι $\frac{1}{2}\times 10^{-2}$, άρα το απόλυτο σφάλμα της διαφοράς τους είναι $\frac{1}{2}\times 10^{-2} + \frac{1}{2}\times 10^{-2} = 10^{-2} = 0.01 > \frac{1}{2}\times 10^{-2}$

Αλλιώς $\sqrt{708} - \sqrt{707} \overset{\text{6 σημαντικά}}{\cong} 26.6083 - 26.5895 = 0.0188$ (με σφάλμα από την πραγματική τιμή 0.0000002209 στη τάξη του εκατομμυριοστού)

$\overset{\text{8 σημ.}}{\cong} 0.0187977$ (με σφάλμα στη τάξη του δισεκατομμυριοστού)

Αλλιώς $$\sqrt{708} - \sqrt{707} = \frac{(\sqrt{708}-\sqrt{707})(\sqrt{708}+\sqrt{707})}{\sqrt{708}+\sqrt{707}} = \frac{708-707}{\sqrt{708}+\sqrt{707}} \overset{\text{4 σημ.}}{\cong} \frac{1}{26.61+26.59}$$

$= \frac{1}{53.20} = 0.01880$ (πολύ καλή ακρίβεια με μόλις 4 σημ. ψηφία) στη τάξη του εκατομμυριοστού

π.χ. $x_{1,2} = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$, $b^2 >> 4ac$

$$x_1 = \frac{(-b+\sqrt{b^2-4ac})(b+\sqrt{b^2-4ac})}{2a(b+\sqrt{b^2-4ac})} = \frac{-4ac}{2a(b+\sqrt{b^2-4ac})} = \frac{-2c}{b+\sqrt{b^2-4ac}}$$

π.χ. $\beta=10$, $t=10$, $\sqrt{7892} - \sqrt{7891} = 0.8883692926\times 10^{2} - 0.8883130079\times 10^{2}$

$= 0.5628470000\times 10^{-2}$ σημαντική απώλεια ακρίβειας

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

π.χ. $x - \sin x$, $|x|$ μικρό (αφαίρεση σχεδόν ίσων αριθμών)

Αλλά $\sin x \simeq x - \frac{x^3}{3!} + \varepsilon(x)$, $|\varepsilon(x)| \leq \frac{|x|^5}{120}$

$\Rightarrow x - \sin x \simeq \frac{x^3}{6}$

$$\sin x = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1} = \underset{n=0}{x} - \underset{n=1}{\frac{x^3}{3}} + \underset{n=2}{\frac{x^5}{120}}$$

<u>Σφάλμα πολλαπλασιασμού</u> Αν $z = fl(fl(x)\,fl(y))$ τότε $\left|\dfrac{z-xy}{xy}\right| \lesssim 3u$

Πράγματι, έστω $fl(x) = x(1+\epsilon_1)$, $fl(y) = y(1+\epsilon_2)$, $|\epsilon_i| \le u$

$z = fl(fl(x)\,fl(y)) = xy\,(1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3)$

Ισχύει $\prod_{i=1}^{m}(1+\epsilon_i) = (1+\epsilon)^m$, όπου $|\epsilon_i| \le u < 1$, $|\epsilon| \le u$ διότι αν $\lambda = \prod_{i=1}^{m}(1+\epsilon_i)$ τότε $(1-u)^m = (1-u)\dots(1-u) \le (1+\epsilon_1)\cdots(1+\epsilon_m) \le (1+u)\dots(1+u) = (1+u)^m$ $\Rightarrow (1-u)^m \le \lambda \le (1+u)^m$. Το θεώρημα ενδιάμεσης τιμής για τη συνεχή συνάρτηση $(1+x)^m$ στο $[-u,u]$ δίνει ότι $\exists\, \epsilon \in [-u,u] : (1+\epsilon)^m = \lambda$

Άρα $z = xy(1+\epsilon)^3$, $|\epsilon| \le u$ $\Rightarrow \left|\dfrac{z-xy}{xy}\right| = |1-(1+\epsilon)^3| = |3\epsilon + 3\epsilon^2 + \epsilon^3| \le$

$\le 3|\epsilon| + 3|\epsilon|^2 + |\epsilon|^3 < 3|\epsilon| + 4|\epsilon|^2 \le 3u + 4u^2$

<u>Σφάλμα πηλίκου</u> Αν $z = fl\left(\dfrac{fl(x)}{fl(y)}\right)$ τότε $\left|\dfrac{z - x/y}{x/y}\right| \lesssim 3u$

Πράγματι, $fl(x) = x(1+\epsilon_1)$, $fl(y) = y(1+\epsilon_2)$ τότε

$$z = fl\left(\frac{x(1+\epsilon_1)}{y(1+\epsilon_2)}\right) = \frac{x}{y}\,\frac{1+\epsilon_1}{1+\epsilon_2}(1+\epsilon_3), \quad |\epsilon_i| \le u$$

$$= \frac{x}{y}(1+\epsilon_1)(1+\epsilon_3)\left(1 - \frac{\epsilon_2}{1+\epsilon_2}\right) = \frac{x}{y}(1+\epsilon)^2(1+\delta),$$

$|\epsilon| \le u$, $\delta = -\dfrac{\epsilon_2}{1+\epsilon_2}$

$$\Rightarrow \left|\frac{z - x/y}{x/y}\right| = |1-(1+\epsilon)^2(1+\delta)| = |2\epsilon + \delta + \epsilon^2 + 2\epsilon\delta + \delta\epsilon^2|$$

$\le 2|\epsilon| + |\delta| + |\epsilon|^2 + 2|\epsilon||\delta| + |\delta||\epsilon|^2$

$\le 2u + \dfrac{u}{1-u} + u^2 + \dots \lesssim 3u$

$|\delta| = \dfrac{|\epsilon_2|}{|1+\epsilon_2|} \le \dfrac{|\epsilon_2|}{1-|\epsilon_2|} \le \dfrac{u}{1-u}$

<u>Σφάλμα πρόσθεσης-αφαίρεσης</u> Αν $z = fl(fl(x)+fl(y))$ τότε $\left|\dfrac{z-(x+y)}{x+y}\right| \lesssim 2u\,\dfrac{|x|+|y|}{|x+y|}$

Πράγματι, αν $fl(x) = x(1+\epsilon_1)$, $fl(y) = y(1+\epsilon_2)$, τότε

$z = fl(x(1+\epsilon_1)+y(1+\epsilon_2)) = (x(1+\epsilon_1)+y(1+\epsilon_2))(1+\epsilon_3) = x(1+\epsilon)^2 + y(1+\delta)^2$,

$|\epsilon|, |\delta| \le |u|$

$\Rightarrow z = x+y+2(\epsilon x+\delta y) + \epsilon^2 x + \delta^2 y \simeq (x+y) + 2(\epsilon x + \delta y)$

$$\Rightarrow \left|\frac{z-(x+y)}{x+y}\right| \simeq 2\left|\frac{\epsilon x+\delta y}{x+y}\right| \le 2\,\frac{|\epsilon||x|+|\delta||y|}{|x+y|} \le 2u\,\frac{|x|+|y|}{|x+y|}$$

Αν x, y ομόσημοι, τότε $|x|+|y| = |x+y| \Rightarrow \left|\dfrac{z-(x+y)}{x+y}\right| \le 2u$

Αν x, y ετερόσημοι π.χ. $x=4$, $y=-3$ τότε $\dfrac{|x|+|y|}{|x+y|} = \dfrac{7}{1} = 7$ μεγάλο φράγμα του σχετικού σφάλματος. Αν μάλιστα $y \simeq -x$, τότε το σφάλμα τείνει στο ∞ (καταστροφική επιρροή του σφάλματος στρογγύλευσης αν αφαιρούμε δύο σχεδόν ίσους αριθμούς) π.χ. $x=0.45142708$ $y=-0.45115944$ $x+y=0.26764\times10^{-3}$

$\beta=10$, $t=5$, $U=-L=10$

$z=fl(fl(x)+fl(y))=fl(0.45143-0.45116)=0.00027=0.27000\times10^{-3}$

$\left|\frac{z-(x+y)}{x+y}\right|\simeq 88\times10^{-4}$ δηλ. 88 φορές μεγαλύτερο από το $2u=\beta^{1-t}=10^{-4}$

της αφαίρεσης δύο ομόσημων αριθμών.

Έχουμε βελτίωση βελτιώνοντας την ακρίβεια t, αλλά και πάλι ύστερα από πολλές πράξεις πάλι η κατάσταση δεν διορθώνεται.

Αν ειδικά x,y αριθμοί μηχανής, τότε $z=fl(x+y)=(x+y)(1+\varepsilon)$

$\Rightarrow \left|\frac{z-(x+y)}{x+y}\right|=|\varepsilon|\le|u|$

π.χ. $\beta=10$, $t=5$, $x=0.12346$, $y=-0.12345$,

$z=fl(fl(x)+fl(y))=fl(x+y)=x+y=10^{-5}$