1. Αριθμητική κινητής υποδιαστολής.

9-02-17

Σφάλματα στρογγύλευσης.

Τα αποτελέσματα υπολογισμών με Η/Υ πρέπει να αντιμετωπίζονται με κριτική διάθεση.

Ποιότητα αριθμητικών μεθόδων

- Απαιτούμενη μνήμη
- Απαιτούμενος χρόνος
- Ακρίβεια αποτελεσμάτων

Οι πράξεις γίνονται με πεπερασμένη ακρίβεια και αυτό έχει ως συνέπεια να προκύπτουν σφάλματα στρογγύλευσης. Αλγόριθμοι ευαίσθητοι σε σφάλματα στρογγύλευσης λέγονται ασταθείς και είναι άχρηστοι. Οι αλγόριθμοι που δεν είναι ευαίσθητοι, λέγονται ευσταθείς.

1ο παράδειγμα.

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

Ιδιότητες: $0 < I_{n+1} < I_n \leq \int_0^1 x^n dx, n \in N$ $\quad e^{x-1} \leq 1 \quad \forall x \in [0,1]$

γιατί $x^{n+1} < x^n$ $\quad \forall x [0,1]$

Αναδρομικός τύπος: $I_n = \int_0^1 x^n e^{x-1} d_x = \int_0^1 x^n (e^{x-1})' d_x =$

$$= [x^n e^{x-1}]_{x=0}^{x=1} - \int_0^1 (x^n)' e^{x-1} dv$$

$$\boxed{\int_a^b f(x)\phi(x)\,dx = [f(x)\phi(x)]_a^b - \int_a^b (f(x))'\phi(x)\,dv}$$

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

- $I_1 = 1 - \int_0^1 e^{x-1} dx = 1 - [e^{x-1}]_0^1 = 1 - (1 - e^{-1}) = e^{-1} = 1/e$

10-02-17

1º Παράδειγμα

$$I_n = \int_0^1 x^n e^{x-1}\,dx, \quad n\in \mathbb{N}$$

- $(I_n)_{n\in\mathbb{N}}$ γνησίως φθίνουσα και τείνει στο μηδέν (μηδενική)
- $0 < I_n \leq \frac{1}{n+1}$

$I_1 = 1/e$ $\qquad I_n = 1 - n\cdot I_{n-1}, \ n\geq 2$

1ος Αλγόριθμος

Προσεγγίζουμε το I_1 με έναν αριθμό $\tilde{I}_1$ ($\tilde{I}_1 \neq I_1$) και υποθέτουμε ότι τα $\tilde{I}_n$ υπολογίζονται (ακριβώς) στη συνέχεια από τον τύπο αναδρομικά με το n.

$$\begin{cases} \tilde{I}_n = 1 - n\tilde{I}_{n-1}, & n\geq 2 \\ \tilde{I}_1 & \text{δεδομένα} \end{cases}$$

Ο αλγόριθμος είναι ασταθής.

Αιτιολόγηση της αστάθειας

Έχουμε

(+) $\quad I_n - \tilde{I}_n = -n(I_{n-1} - \tilde{I}_{n-1}), \quad n\geq 2$

Επομένως, επαγωγικά, έχουμε

$$I_n - \tilde{I}_n = (-1)^{n-1} n! (I_1 - \tilde{I}_1) \quad (*) , \ n\in\mathbb{N}$$

Αρχή: $n=1$: $I_1 - \tilde{I}_1 = I_1 - \tilde{I}_1$ ✓

Βήμα: $n \to n+1$:

$$I_{n+1} - \tilde{I}_{n+1} \overset{(+)}{=} -(n+1)\,(I_n - \tilde{I}_n)$$

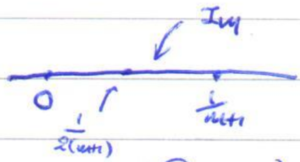
$$\overset{(*)}{=} -(n+1)\,\overbrace{(-1)^{n-1}\, n!\,(I_1 - \tilde{I}_1)}$$

$$= (-1)^n (n+1)!\,(I_1 - \tilde{I}_1) = (-1)^{n+1-1}(n+1)!\,(I_1 - \tilde{I}_1)$$

δηλαδή η σχέση αυτή ισχύει όντως και για n+1. ⊛

<u>Συμπέρασμα</u>: $|I_n - \tilde{I}_n| = n!\,|I_1 - \tilde{I}_1|$ $\left(\begin{array}{l}(-1)^{n+1} \\ |(-1)^{n-1}| = 1\end{array}\right)$

αυξάνει ταχύτατα με το n,
οπότε ο αλγόριθμος είναι <u>ασταθής</u>.



<u>2ος αλγόριθμος</u>.

Προσεγγίζουμε το I_m με έναν αριθμό $\tilde{I}_m$ (για οποιοδήποτε $\tilde{I}_m \in [0, \frac{1}{1+m}]$ ισχύει $|I_m - \tilde{I}_m| \le \frac{1}{m+1}$.

- Βέλτιστη επιλογή θα ήταν το $\tilde{I}_m = \frac{1}{2(m+1)}$.

- Υπολογίζουμε τα $\tilde{I}_{m-1}, \tilde{I}_{m-2}, \dots, \tilde{I}_\ell$ αναδρομικά με τον τύπο.

$$\boxed{\tilde{I}_{n-1} = \frac{1-\tilde{I}_n}{n}, \quad n = m, m-1, \dots, \ell+1.}$$

(θα ξεκινώ από κάτι μεγαλύτερο)

Ο αλγόριθμος είναι <u>ευσταθής</u>

- <u>Αιτιολόγηση της ευστάθειας</u>.

Έχουμε

$$I_{n-1} - \tilde{I}_{n-1} = -\frac{I_n - \tilde{I}_n}{n}, \quad n = m, m-1, \dots, \ell+1.$$

Επαγωγικά έχουμε ότι για οποιοδήποτε $n < m$

$$I_n - \tilde{I}_n = (-1)^{m-n}\frac{1}{(n+1)(n+2)\cdots(m-1)m}(I_m - \tilde{I}_m), \quad \ell \le n < m.$$

Συμπέρασμα,

$$|I_n - \tilde{I}_n| = \frac{1}{(n+1)(n+2)\dots m}|I_m - \tilde{I}_m|$$

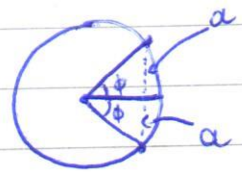
(Το αρχικό σφάλμα διαιρείται)

Οπότε, αλγόριθμος ευσταθής.

$2^{\underline{o}}$ παράδειγμα: $y_n = 2^n \sin\frac{\pi}{2^n}, n \in \mathbb{N}$

- $y_1 = 2$

Γεωμετρική ερμηνεία: Για $n \geq 2$, το $2y_n$ είναι η περίμετρος του κανονικού πολυγώνου με 2^n πλευρές που είναι εγγεγραμμένο στον μοναδιαίο κύκλο.



$\phi = \frac{\pi}{2^n}$, $a = \sin\frac{\pi}{2^n}$

$2\phi = \frac{2\pi}{2^n}$

Συμπέρασμα. Το $2a$ είναι η πλευρά του κανονικού πολυγώνου με 2^n πλευρές που είναι εγγεγραμμένο στον μοναδιαίο κύκλο.

Περίμετρος αυτού του πολυγώνου είναι: $2^n \cdot 2a = 2\cdot 2^n \cdot \sin\frac{\pi}{2^n} = 2y_n$

Η (y_n) είναι γνήσια αύξουσα και συγκλίνει στο π.

$\boxed{14\text{-}02\text{-}17}$

2º παράδειγμα.

$y_n = 2^n \cdot \sin\frac{\pi}{2^n}$, $n\in\mathbb{N}$, $y_1 = 2$.

Ισχυρίζομαι: Η $(y_n)_{n\in\mathbb{N}}$ είναι γνησίως αύξουσα και συγκλίνει στο π.

Απόδειξη.

$\sin(2x) = 2\cdot\sin x\cdot\cos x$ με $x = \frac{\pi}{2^{n+1}}$

α) $y_n = 2^n \sin\frac{\pi}{2^n} = 2^n\cdot 2\sin\frac{\pi}{2^{n+1}}\cos\frac{\pi}{2^{n+1}} =$

$= \underbrace{2^{n+1}\sin\frac{\pi}{2^{n+1}}}_{y_{n+1}}\cos\frac{\pi}{2^{n+1}} = \underbrace{y_{n+1}}_{>0}\underbrace{\cos\frac{\pi}{2^{n+1}}}_{<1} < y_{n+1}$

Η ακολουθία είναι γνησίως αύξουσα

β) Ξέρουμε ότι $\lim\limits_{x\to0}\frac{\sin x}{x} = 1$

Άρα, για οποιοδήποτε $a\in\mathbb{R}$, $a\neq0$ έχουμε ότι $\lim\limits_{x\to0}\frac{\sin(ax)}{ax} = 1$

$\Rightarrow \lim\limits_{x\to0}\frac{\sin(ax)}{ax} = \frac{1}{a}\lim\limits_{x\to0}\frac{\sin(ax)}{x} = 1$

$\Rightarrow \lim\limits_{x\to0}\frac{\sin(ax)}{x} = a.$

Τώρα, $y_n = \frac{\sin\frac{\pi}{2^n}}{\frac{1}{2^n}} \longrightarrow \pi$, $n\to\infty$

$x_n = \frac{1}{2^n}$

$a = \pi.$

Αναδρομικός τύπος.

$\sin^2 x = \frac{1-\cos(2x)}{2}$

$y_{n+1} = 2^{n+1}\cdot\sin\frac{\pi}{2^{n+1}} = 2^{n+1}\sqrt{\frac{1}{2}\left(1-\cos\frac{\pi}{2^n}\right)} = 2^{n+1}\sqrt{\frac{1}{2}\left(1-\sqrt{1-\sin^2\frac{\pi}{2^n}}\right)}$

$\boxed{\sin\frac{\pi}{2^n} = 2^{-n}y_n}$

$= 2^{n+1}\sqrt{\frac{1}{2}\left(1-\sqrt{1-(2^{-n}y_n)^2}\right)}$

1ος αλγόριθμος

$$\begin{cases} y_1 = 2 \\ y_{n+1} = 2^{n+1}\sqrt{\frac{1}{2}\left(1-\sqrt{1-(2^{-n}y_n)^2}\right)}, \quad n=1,2,\dots \end{cases}$$

Ο αλγόριθμος είναι ασταθής

Αιτιολόγηση της αστάθειας: Αν το y_n είναι προσέγγιση του π, τότε για αρκετά μεγάλο n, το $(2^{-n}\cdot y_n)^2$ είναι πολύ κοντά στο 0 (μηδέν). Άρα οδηγούμαστε σε αφαίρεση σχεδόν ίσων αριθμών, και σε αυτό οφείλεται η αστάθεια.

2ος Αλγόριθμος:

$$\sqrt{a}-\sqrt{b} = \frac{(\sqrt{a}-\sqrt{b})(\sqrt{a}+\sqrt{b})}{(\sqrt{a}+\sqrt{b})}$$

$$\sqrt{a}-\sqrt{b} = \frac{a-b}{\sqrt{a}+\sqrt{b}}$$

Έχουμε $1-\sqrt{1-(2^{-n}y_n)^2} = \frac{\cancel{1}-[\cancel{1}-(2^{-n}y_n)^2]}{1+\sqrt{1-(2^{-n}y_n)^2}}$

$$= \frac{(2^{-n}\cdot y_n)^2}{1+\sqrt{1-(2^{-n}y_n)^2}}$$

Επομένως, $y_{n+1} = 2^{n+1}\frac{2^{-n}\cdot y_n}{\sqrt{2}\cdot\sqrt{1+\sqrt{1-(2^{-n}y_n)^2}}} = \frac{\sqrt{2}}{\sqrt{1+\sqrt{1-(2^{-n}y_n)^2}}}\cdot y_n$

$$\begin{cases} y_{n+1} = \sqrt{\frac{2}{1+\sqrt{1-(2^{-n}y_n)^2}}}\cdot y_n, \quad n=1,2,\dots \\ y_1 = 2 \end{cases}$$

Ο αλγόριθμος είναι <u>ευσταθής</u>.

Παράσταση αριθμών ως προς οποιαδήποτε βάση.

Καθημερινή ζωή: δεκαδικό σύστημα

Βάση: 10

Ψηφία: $0, 1, \dots, 9$

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

Γενικά: Έστω $a_N, a_{N-1}, \dots, a_0, a_{-1}, a_{-2}, \dots$ είναι ψηφία του δεκαδικού συστήματος.

Τότε $(a_N a_{N-1} \dots a_0 a_{-1} \dots)_{10} = a_N\cdot 10^N + a_{N-1}\cdot 10^{N-1} + \dots + a_0\cdot 10^0$
$+ a_{-1}\cdot 10^{-1} + a_{-2}\cdot 10^{-2} + \dots$

Ακέραιο μέρος $a_N a_{N-1} \dots a_1 a_0$ είναι η τιμή του πολυωνύμου p,
$p(x) = a_N x^N + a_{N-1} x^{N-1} + \dots + a_1 x + a_0$,
στο σημείο $\boxed{x=10}$

Κλασματικό μέρος: $a_{-1} a_{-2} \dots$ είναι η τιμή της δυναμοσειράς
$\sum_{k=1}^{\infty} a_{-k} x^k$, στο σημείο $x = 1/10$

Μη μοναδικότητα της παράστασης:
$4.130 = 4.1.2\bar{9}$ ← επαναλαμβάνεται άπειρες φορές
$= 4.12999999\dots$

$9 \cdot \sum_{i=1}^{\infty} 10^{-i} = 9 \cdot \frac{\frac{1}{10}}{1-\frac{1}{10}} = 9 \cdot \frac{\frac{1}{10}}{\frac{9}{10}} = 1$

$|w| < 1$
$\sum_{i=0}^{\infty} w^i = \frac{1}{1-w}$

$\sum_{i=n_0}^{\infty} w^i = \frac{w^{n_0}}{1-w}$

$\frac{1}{3} = 0,333333$

Γενικά ισχύει, $\boxed{(b-1)\sum_{i=1}^{\infty} b^{-i} = 1}$

Υπόθεση που μας εξασφαλίζει μοναδικότητα:
$\forall k_0 \in \mathbb{N}\ \exists k > k_0$ τω $a_k \neq 9$

Σύστημα με βάση β, β≥2

Βάση: β

Ψηφία: 0, 1, 2, …, β-1

a_k ψηφία

$$\pm(a_N \dots a_0 \cdot a_{-1} \dots)_\beta = \pm(a_N \beta^N + a_{N-1}\beta^{N-1} + \dots + a_0 \beta^0 + a_{-1}\beta^{-1} + \dots)$$

Παράδειγμα

$$(100110.11)_2 = 1\cdot 2^5 + 0\cdot 2^4 + \dots + 1\times 2^2 + 1\cdot 2^1 + 1\cdot 2^{-1} + 1\cdot 2^{-2} = (38.75)_{10}$$

i) Μετατροπή από ένα σύστημα με βάση β στο δεκαδικό

α) Ακεραίων αριθμών:

Παράδειγμα $(53473)_8 = 5\cdot 8^4 + 3\cdot 8^3 + 4\cdot 8^2 + 7\cdot 8^1 + 3\cdot 8^0$

$$= 3 + 8(7 + 8(4 + 8(\underline{3 + 8(5)})))$$

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

(Σχήμα Horner) $\longrightarrow$ $p(x) = a_N x^N + a_{N-1}x^{N-1} + \dots + a_1 x + a_0$

$$= a_0 + x(a_1 + x(a_2 + x(a_3 + \dots + x(a_{N-1} + x(a_N))$$

$y \leftarrow a_N$, για $i = N-1, N-2, \dots, 0$

$y \leftarrow a_i + x\cdot y$

Τότε $y = p(x)$.

$y \leftarrow a_i + x\cdot y$ (flop)

Το σχήμα του Horner απαιτεί N flop

β) Κλασματικών αριθμών x $\quad 0 < x < 1$

Παράδειγμα: $(.11)_2 = 1\cdot 2^{-1} + 1\cdot 2^{-2} = \frac{1}{2} + \frac{1}{4} = (0,75)_{10}$

ii) Μετατροπή από το δεκαδικό σε ένα σύστημα με βάση 6.

α) Ακεραίων αριθμών

· Βασίζεται στον αλγόριθμο της διαίρεσης.

Παράδειγμα. Μετατροπή του $(369)_{10}$ στο οκταδικό σύστημα

$$(369)_{10} = (\ldots\ldots a_2 a_1 a_0)_8 = a_0 + 8(a_1 + 8(a_2 + \ldots)))$$
$$0 \le a_0 \le 7$$

Άρα το a_0 είναι το υπόλοιπο της διαίρεσης $369 : 8$ και το $a_1 + 8(a_2 + \ldots)$ είναι το πηλίκο της διαίρεσης.

$$\begin{array}{r|l} 369 & 8 \\ \cline{2-2} 49 & 46 \\ 1 & \end{array}$$

Το $\boxed{a_0 = 1}$ και $(46)_{10} = a_1 + 8(a_2 + \ldots)$

Επομένως το a_1 είναι το υπόλοιπο της διαίρεσης $46 : 8$ και το $(a_2 + \ldots)$ είναι το πηλίκο.

$$\begin{array}{r|l} 46 & 8 \\ \cline{2-2} 6 & 5 \end{array}$$

Άρα $a_1 = 6$ και $(5)_{10} = (a_2 + 8(a_3 \ldots)$

Έπεται ότι το $a_2 = 5$ και $a_3 = \cancel{\ldots} = 0$

Συμπέρασμα $(369)_{10} = (561)_8$

Επαλήθευση: $(561)_8 = 5 \cdot 8^2 + 6 \cdot 8 + 1$
$= (369)_{10}$ ✓

16-02-17

α) Μετατροπή από το δεκαδικό σε σύστημα με βάση β.

α) ακέραιων αριθμών.

Βασίζεται στον αλγόριθμο της διαίρεσης.

β) κλασματικών αριθμών

Έστω $0<x<1$ στο δεκαδικό σύστημα

Μετατροπή του x σε σύστημα με βάση β.

$$x = (.\alpha_{-1}\alpha_{-2}\ldots)_\beta = \alpha_{-1}\cdot\beta^{-1} + \alpha_{-2}\cdot\beta^{-2} + \ldots$$

$$\Rightarrow \beta x = \alpha_{-1} + \alpha_{-2}\cdot\beta^{-1} + \ldots = (\alpha_{-1}.\alpha_{-2}\alpha_{-3}\ldots)_\beta$$

Συμπέρασμα: Το α_{-1} είναι το ακέραιο μέρος του βx

Παράδειγμα: Μετατροπή του $x = (.372)_{10}$ στο δυαδικό σύστημα.

$$(.372)_{10} = (.\alpha_{-1}\alpha_{-2}\alpha_{-3}\ldots)_2$$

Έχουμε: $2x = 0{,}744$ άρα $\alpha_{-1} = 0$, και $y_1 = 0.744$.

$2y_1 = 1.488$, άρα $\alpha_{-2} = 1$, και $y_2 = 0.488$

$2y_2 = 0.976$, άρα $\alpha_{-3} = 0$, και $y_3 = 0.976$.

$2y_3 = 1.952$, άρα $\alpha_{-4} = 1$ και $y_4 = 0.952$

$\vdots$

Επομένως, $(0.372)_{10} = (0.0101\ldots)_2$

Παρατήρηση: Κατά τη μετατροπή από ένα σύστημα σε άλλο, οι ακέραιοι αριθμοί παραμένουν ακέραιοι και οι κλασματικοί παραμένουν κλασματικοί.

Το πλήθος των ψηφίων ενός κλασματικού αριθμού μπορεί από πεπερασμένο να γίνει άπειρο, ή αντίστροφα.

Γνωρίζουμε
$$\sum_{i=n_0}^{\infty} \omega^i = \frac{\omega^{n_0}}{1-\omega}$$

Παράδειγμα:

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

Απόδειξη ισχυρισμού:

$$\sum_{n=1}^{\infty}\left(\frac{1}{2^{4n}} + \frac{1}{2^{4n+1}}\right) = \sum_{n=1}^{\infty}\left(\frac{1}{2^{4n}} + \frac{1}{2\cdot 2^{4n}}\right)$$

$$= \sum_{n=1}^{\infty}\left(1 + \frac{1}{2}\right)\cdot\frac{1}{2^{4n}} = \left(1+\frac{1}{2}\right)\sum_{n=1}^{\infty}\frac{1}{2^{4n}} = \frac{3}{2}\sum_{n=1}^{\infty}\frac{1}{2^{4n}}$$

$$= \frac{3}{2}\sum_{n=1}^{\infty}\left(\frac{1}{2^4}\right)^n = \frac{3}{2}\,\frac{\frac{1}{2^4}}{1-\frac{1}{2^4}} = \frac{3}{2}\cdot\frac{1}{15} = \frac{1}{10}$$

$$\frac{1}{10} = \sum_{n=1}^{\infty}\left(\frac{1}{2^{4n}} + \frac{1}{2^{4n+1}}\right) = \frac{1}{2^4} + \frac{1}{2^5} + \frac{1}{2^8} + \frac{1}{2^9} + \frac{1}{2^{12}} + \frac{1}{2^{13}} + \dots$$

$$\Rightarrow \frac{1}{10} = 2^{-4} + 2^{-5} + 2^{-8} + 2^{-9} + 2^{-12} + 2^{-13} + \dots$$

$$\Rightarrow (0.1)_{10} = (0.0001100110011\dots)_2$$

$$= (0.0\overline{0011})_2$$

Αριθμοί μηχανής

Έχω $x \in \mathbb{R}$, $x \neq 0$

Σε ένα σύστημα με βάση β, ο x μπορεί να γραφεί στην μορφή (*) $x = \pm\, .d_1 d_2 \dots \cdot \beta^e$ με $d_1 \neq 0$ και d_i ψηφία ως προς, τη βάση β και e κατάλληλος εκθέτης.

$153.181 = 0.153181 \cdot 10^3$

$0.00153181 = 0.153181 \cdot 10^{-2}$

Η μορφή (*) λέγεται (κανονική) μορφή κινητής υποδιαστολής.

Το σύνολο των αριθμών μηχανής $M = M(\beta, t, L, U)$
χαρακτηρίζεται από τις παραμέτρους:

- β = βάση του αριθμητικού συστήματος.
- t = ακρίβεια = πλήθος των ψηφίων του κλάσματος του αριθμού.
- L = κάτω φράγμα } του εκθέτη e του β. $(L < e < U)$
- U = άνω φράγμα } L, U ακέραιοι και $L \approx -U$.

Κάθε $x \in M$, $x \neq 0$ είναι της μορφής (+) $x = \pm .d_1 d_2 \ldots d_t \cdot \beta^e$
με $d_1 \neq 0$ και $L \leq e \leq U$.

Το M αποτελείται από το μηδέν και όλους τους αριθμούς αυτής της μορφής (+)

- Το σύνολο M είναι πεπερασμένο.
- Το μέγιστο στοιχείο του M: $x = .d_1 d_2 \ldots d_t \cdot \beta^U$.
  και $d_1 = d_2 = \ldots = d_t = \beta - 1$
- Ελάχιστο θετικό στοιχείο του M:
  $x = .100\ldots0\,\beta^L$.
- Η απόσταση μεταξύ διαδοχικών στοιχείων του M δεν είναι σταθερή.
- Το M δεν είναι κλειστό ως προς τον πολλαπλασιασμό, δηλαδή $x, x^* \in M \not\Rightarrow x \cdot x^* \in M$.

Παράδειγμα:

$$\underbrace{.100\ldots0 \cdot \beta^L}_{\in M} \cdot .100\ldots0 \cdot \beta^L \notin M$$

- Το M δεν είναι κλειστό ως προς την πρόσθεση

Παράδειγμα: $\beta = 10$, $t = 5$

$1, 10^{-5} \in M$

$1 + 10^{-5} = 1.00001 \notin M$.

Μας ενδιαφέρει το M να είναι όσο πιο πυκνό γίνεται και όσο πιο ευρύ γίνεται, δηλαδή να έχει:
- μεγάλο t
- μεγάλο διάστημα $[L, U]$

Προσέγγιση πραγματικών αριθμών με στοιχεία του M (αριθμούς μηχανής).

α) Αν $|x| > .d_1 d_2 \ldots d_t \cdot \beta^{U}$ με $d_i = \beta - 1$, $i = 1, \ldots, t$

Προκύπτει υπερχείλιση (overflow)

Οι υπολογισμοί σταματούν.

β) Αν $0 < |x| < .100 \ldots 0 \cdot \beta^{L}$

Τότε προκύπτει υποχείλιση (underflow)

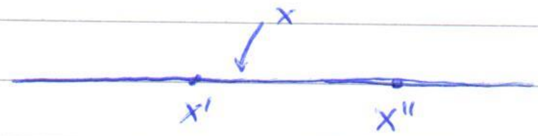
(Συνήθως ο x προσεγγίζεται με το μηδέν και οι υπολογισμοί συνεχίζονται)

γ) Αν $.1 \cdot \beta^{L} \leq |x| \leq$ μέγιστο στοιχείο του M,

Τότε ο x προσεγγίζεται με τον αριθμό $fl(x)$.

Συνήθως ισχύει:

Η απόσταση $|x - fl(x)| \leq |x - y| \quad \forall y \in M$



Ισχυρισμός: Τότε

$$(*) \quad \left| \frac{fl(x) - x}{x} \right| \leq \frac{1}{2} \cdot \beta^{1-t} \quad (\text{για } x \neq 0)$$

σχετικό σφάλμα

α) Αν $fl(x) = x$, τότε η (*) προφανώς ισχύει.

β) Αν $x \notin M$, τότε υπάρχουν διαδοχικοί αριθμοί μηχανής $x', x'' \in M$ τ.ω. $x' < x < x''$. Τότε ισχύει

$$|fl(x) - x| \leq \frac{1}{2} |x' - x''|$$

οπότε

$$\left|\frac{fl(x)-x}{x}\right| \leq \frac{|x-x''|}{2|x|}$$

Έστω ότι $x>0$ (για x αρνητικό ανάλογα αντίστοιχα)

$$x = .d_1 d_2 \dots d_t d_{t+1} \dots \beta^e.$$

Άρα

$$x' = .d_1 d_2 \dots d_t \cdot \beta^e$$

$$x'' = (.d_1 \dots d_t + \beta^{-t}) \cdot \beta^e$$

οπότε $x''-x' = \beta^e \cdot \beta^{-t} = \beta^{e-t}$ (Διαφορά διαδοχικών αριθμών μηχανής δεν είναι σταθερή)

Επομένως,

$$\left.\begin{array}{l} \left|\dfrac{fl(x)-x}{x}\right| \leq \dfrac{\beta^{e-t}}{2x} \\ \text{Όμως ισχύει } x \geq .1 \cdot \beta^e = \beta^{e-1} \end{array}\right\}$$

$$\Rightarrow \left|\frac{fl(x)-x}{x}\right| \leq \frac{1}{2}\frac{\beta^{e-t}}{\beta^{e-1}} = \boxed{\frac{1}{2}\cdot\beta^{1-t}}$$

17-02-17.

Αριθμοί μηχανής

Το σύνολο $M = M(\beta, t, L, U)$ αποτελείται από το μηδέν και τους αριθμούς της εξής μορφής:

$$x = \pm\, .d_1 d_2 \dots d_t \cdot \beta^e \quad \text{με } d_1 \neq 0 \quad \text{και} \quad L \le e \le U$$

Προσέγγιση με αριθμούς μηχανής

Αριθμοί x, στο εύρος των αριθμών μηχανής, προσεγγίζονται με αριθμούς $fl(x) \in M$.

Συνήθως ισχύει

$$① \quad |x - fl(x)| \le |x - y| \quad \forall y \in M$$

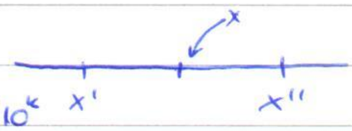
Τότε ισχύει:

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

Στην ① οδηγούμαστε με στρογγύλευση.

Παράδειγμα: $\beta = 10$, $t = 5$

$$x = .a_1 a_2 \dots a_5 \vdots a_6 \dots 10^k$$



- Αν $a_6 \ge 5$, τότε $fl(x) = x'' = (.a_1 a_2 \dots a_5 + 10^{-5}) 10^k$
- Αν $a_6 < 5$, τότε $fl(x) = x' = (.a_1 a_2 \dots a_5) \cdot 10^k$
- Στην περίπτωση που το $a_6 = 5$ και όλα τα $a_i = 0$, $i \ge 7$ τότε μπορούμε να επιλέξουμε είτε τον x' είτε τον x''.

Εναλλακτικός τρόπος προσέγγισης: αποκοπή.

Το $fl(x) = (.a_1 a_2 \dots a_5) 10^k$

Σε αυτήν την περίπτωση αντί για τη ② έχουμε

$$③ \quad \left| \frac{x - fl(x)}{x} \right| \le \beta^{1-t}$$

Συμπέρασμα: $$\left| \frac{x - fl(x)}{x} \right| \le u := \begin{cases} \frac{1}{2} \beta^{1-t} & \text{για στρογγύλευση} \\ \beta^{1-t} & \text{για αποκοπή.} \end{cases}$$

Το u λέγεται μοναδιαίο σφάλμα στρογγύλευσης.

Πράξεις.

$* \in \{+, -, \times, :\}$

Αν $x, y, x*y$ είναι μη μηδενικοί αριθμοί στο εύρος των αριθμών μηχανής, υποθέτουμε ότι αντί για $x*y$ ο υπολογιστής μας δίνει $z = fl(fl(x) * fl(y))$.

Παράδοξα: $\beta=10$, $t=5$, $U=-L=10$
στρογγύλευση.

$a_1 = 1$, $a_2 = 3\cdot 10^{-5}$, $a_3 = 3\cdot 10^{-5}$

$a_1, a_2, a_3 \in M$

Τότε:

$fl(a_1 + a_2) = fl(1.00003) = 1.0000 = 1$

Οπότε

$$fl(fl(a_1+a_2)+a_3) = 1.$$

Αλλά

$$a_2 + a_3 = 6\cdot 10^{-5} \in M$$

οπότε

$$fl(a_1 + (a_2+a_3)) = fl(1.00006)$$
$$= 1.0001.$$

Διαφορετικά αποτελέσματα!

Συμπέρασμα: Έχει σημασία η σειρά με την οποία γίνονται οι πράξεις.

Παρατήρηση: Για κάθε $0<|x|< \ldots \frac{1}{2}\beta^{1-t}$
έχουμε

$$fl(1+fl(x)) = 1.$$

$\frac{1}{2}\beta^{1-t}$ = Έψιλον της μηχανής ή μηδέν της μηχανής

Επιρροή των σφαλμάτων στρογγύλευσης στους υπολογισμούς

$x, y, x * y$ μη μηδενικοί αριθμοί
στο εύρος αριθμών μηχανής.

$z = fl(fl(x) * fl(y))$.

Ερώτημα: Πώς μπορώ να εκτιμήσω το σχετικό σφάλμα;

$$\frac{fl(fl(x) * fl(y)) - x * y}{x * y} \; ;$$

Βοηθητικά αποτελέσματα:

α) $\left|\frac{x - fl(x)}{x}\right| \le u \Rightarrow fl(x) = x \cdot (1+\varepsilon)$
με $\varepsilon = \varepsilon(x)$ τ.ω $|\varepsilon| \le u$.

$fl(x) = x(1+\varepsilon) \Rightarrow \frac{fl(x) - x}{x} = \varepsilon$

$\Rightarrow |\varepsilon| \le u$

β) Αν ε_i, $1 \le i \le m$, τ.ω $|\varepsilon_i| \le u$, $i = 1, \dots, m$.
τότε υπάρχει ε με $|\varepsilon| \le u$ τ.ω.
$$\prod_{i=1}^{m} (1+\varepsilon_i) = (1+\varepsilon)^m$$

Θεωρούμε τη συνάρτηση $\phi(x) = (1+x)^m$, $x \in [-u, u]$
Αυτή είναι συνεχής συνάρτηση. Επίσης,

$$\begin{array}{ccccc} (1-u)^m & \le & \prod_{i=1}^{m}(1+\varepsilon_i) & \le & (1+u)^m \\ \| & & & & \| \\ \phi(-u) & & & & \phi(u) \end{array}$$

Σύμφωνα με το Θεώρημα της ενδιάμεσης τιμής
υπάρχει ένα $\varepsilon \in [-u, u]$ τ.ω $\prod_{i=1}^{m}(1+\varepsilon_i) = \phi(\varepsilon) = (1+\varepsilon)^m$.

α) Πολλαπλασιασμός:

$$z = fl(fl(x)\cdot fl(y))$$
$$= fl(x(1+\varepsilon_1)\cdot y(1+\varepsilon_2))$$
$$= x(1+\varepsilon_1)\cdot y(1+\varepsilon_2)(1+\varepsilon_3)$$
$$= x\cdot y(1+\varepsilon_1)\cdot(1+\varepsilon_2)(1+\varepsilon_3)$$
$$= x\cdot y(1+\varepsilon)^3 \qquad \text{με } |\varepsilon_i| \le u,\ |\varepsilon| \le u.$$

Επομένως,

$$\left|\frac{z - x\cdot y}{xy}\right| = \left|\frac{\cancel{xy} - xy(1+\varepsilon)^3}{\cancel{xy}}\right| = \left|\frac{\cancel{xy}(1+\varepsilon)^3 - \cancel{xy}}{xy}\right|$$
$$= |(1+\varepsilon)^3 - 1| = |\cancel{1} + 3\varepsilon + 3\varepsilon^2 + \varepsilon^3 - \cancel{1}|$$
$$= |3\varepsilon + 3\varepsilon^2 + \varepsilon^3| \le 3\cdot u + \underbrace{3u^2 + u^3}_{O(u^2)}$$

Αφού $u \ll 1$ έχουμε
$u^2 \ll u$.

Λέμε ότι το σχετικό σφάλμα στον πολλαπλασιασμό είναι το πολύ $3u$.

β) Διαίρεση:

$$z = fl\left(\frac{fl(x)}{fl(y)}\right)$$
$$= fl\left(\frac{x(1+\varepsilon_1)}{y(1+\varepsilon_2)}\right) = \frac{x(1+\varepsilon_1)}{y(1+\varepsilon_2)}\cdot(1+\varepsilon_3)$$

$$\frac{1}{1+\varepsilon_2} = 1+\delta$$
$$\Rightarrow \delta = -\frac{\varepsilon_2}{1+\varepsilon_2}$$
$$\Rightarrow |\delta| \le \left|\frac{\varepsilon_2}{1+\varepsilon_2}\right|$$
$$\Rightarrow \frac{|\varepsilon_2|}{1+\varepsilon_2} \le \frac{u}{1-u} = u + O(u^2)$$

Άρα

$$\left|\frac{z - \frac{x}{y}}{\frac{x}{y}}\right| = \left|\frac{\cancel{\frac{x}{y}}(1+\varepsilon_1)(1+\varepsilon_3)(1+\delta) - \cancel{\frac{x}{y}}^{\,1}}{\cancel{x/y}}\right|$$
$$= |\underbrace{(1+\varepsilon_1)(1+\varepsilon_3)}_{(1+\varepsilon)^2}(1+\delta) - 1|$$
$$= |(1+\varepsilon)^2(1+\delta) - 1| = |2\varepsilon + \delta + \varepsilon^2 + 2\varepsilon\delta + \delta\varepsilon^2|$$
$$\le 3u + O(u^2)$$

γ) <u>Πρόσθεση - αφαίρεση</u>:

$$z = fl(fl(x) + fl(y))$$
$$= [x(1+\varepsilon_1) + y(1+\varepsilon_2)](1+\varepsilon_3)$$
$$= x(1+\varepsilon_1)(1+\varepsilon_3) + y(1+\varepsilon_2)(1+\varepsilon_3)$$
$$= x(1+\varepsilon)^2 + y(1+\delta)^2 \quad \text{με } |\varepsilon| \leq u, \ |\delta| \leq u.$$

Άρα $z = x(1+\varepsilon^2) + y(1+\delta)^2$
$$= (x+y) + 2(\varepsilon x + \delta y) + x\varepsilon^2 + y\delta^2$$

$$\Rightarrow \quad z \approx (x+y) + 2(\varepsilon x + \delta y)$$

Επομένως,

παραλείπω όρους του u^2! (πολύ μικροί)

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

1η <u>περίπτωση</u>: Ας υποθέσω ότι x, y ομόσημοι.
Τότε,
$|x+y| = |x| + |y|$, οπότε έχουμε
$$\left|\frac{z-(x+y)}{x+y}\right| \leq 2\cdot u + O(u^2)$$

2η <u>περίπτωση</u>: x, y ετερόσημοι
Στη χειρότερη περίπτωση έχουμε $\varepsilon \approx -\delta$ και $|\varepsilon| \approx u$.
Τότε,
$$\left|\frac{z-(x+y)}{x+y}\right| \approx 2\cdot\frac{|x-y|}{|x+y|}\cdot u$$

$$\left|\frac{\varepsilon x + \delta y}{x+y}\right| \approx \frac{|ux - uy|}{|x+y|} = \frac{|x-y|}{|x+y|}\cdot u$$

Αν τα x και y είναι περίπου αντίθετοι αριθμοί, τότε σχεδόν το κλάσμα $\frac{|x-y|}{|x+y|}$ μπορεί να γίνει πολύ μεγάλο.

<u>Συμπέρασμα</u>, τα σφάλματα στρογγύλευσης στην αφαίρεση σχεδόν ίσων αριθμών μπορούν να έχουν <u>καταστροφική επιρροή</u>.

Άρα, η αφαίρεση σχεδόν ίσων αριθμών πρέπει να αποφεύγεται.

<u>Σημείωση</u>: Η αφαίρεση σχεδόν ίσων αριθμών <u>μηχανής</u> γίνεται χωρίς πρόβλημα.

<u>Παράδειγμα</u>: $b=10$, $t=5$, $u=-L=10$, στρογγύλευση.

$$x = .45142708 \quad , \quad y = -.45115944$$

λείπει το 2

$$x+y = .26764 \cdot 10^{-3} \in M$$

$$z = fl(fl(x)+fl(y)) = fl(0.45147 - .45116)$$

$$= 0.00027 = .27000 \cdot 10^{-3}$$

ένδειξη αφαίρεσης σχεδόν ίσων αριθμών.

$\boxed{2\text{-}03\text{-}17}$

<u>Παράδειγμα</u> $b=10$, $t=5$, $u=-L=10$, στρογγύλευση.

$$x = .45142708 \quad , \quad y = -.45115944$$

$$x+y = .26764 \cdot 10^{-3}$$

$$z = fl(fl(x)+fl(y))$$

$$= fl(.45143 - .45116)$$

$$= \ldots 00027 = .27000 \cdot 10^{-3}$$

Παραδείγματα αποφυγής της <u>αφαίρεσης σχεδόν ίσων αριθμών</u>

1º x, y μεγάλοι θετικοί αριθμοί

$$x \approx y$$

$$\sqrt{x} - \sqrt{y} = \frac{x-y}{\sqrt{x}+\sqrt{y}}$$

$x = 7298$ $\quad$ $\underbrace{\sqrt{x} - \sqrt{y}}$
$y = 7297$ $\quad$ χανεται ακρίβεια

$\frac{x-y}{\sqrt{x}+\sqrt{y}}$ ⤳ δεν χανεται ακρίβεια

2ο $f(x) = x - \sin x$

Θέλαμε να υπολογίσαμε τιμές της f για $|x|$ μικρό.

- Τα x και $\sin x$ είναι ομόσημοι
- $\lim_{x \to 0} \frac{\sin x}{x} = 1$, δηλαδή $\sin x \approx x$ για $|x|$ μικρό.
- Επομένως έχαμε αφαίρεση σχεδόν ίσων αριθμών.

- Ανάπτυγμα Taylor

$$\sin x = x - \frac{x^3}{6} + \varepsilon(x) \leftarrow \text{υπόλοιπο.}$$

$$\text{με } |\varepsilon(x)| \leq \frac{|x|^5}{120}$$

Άρα $f(x) = \cancel{x} - \left(\cancel{x} - \frac{x^3}{6} + \varepsilon(x)\right)$

$$= \frac{x^3}{6} - \varepsilon(x) \approx \frac{x^3}{6}$$

Το $\frac{x^3}{6}$ υπολογίζεται χωρίς πρόβλημα.

Σφάλματα στον υπολογισμό αθροισμάτων.

Θα μελετήσαμε την επιρροή προβλημάτων στρογγύλευσης, λόγω της αριθμητικής κινητής υποδιαστολής με πεπερασμένη ακρίβεια, στον υπολογισμό αθροισμάτων.

<u>Παράδειγμα</u> $S_n = 1 + \sum_{k=1}^{n} \frac{1}{k^2+k}$, $n \in N$

Έχουμε $\frac{1}{k^2+k} = \frac{1}{k} - \frac{1}{k+1}$, οπότε το άθροισμα είναι <u>τηλεσκοπικό</u>,

για $k=1,\ldots$ $S_n = 1 + (1-\cancel{\frac{1}{2}}) + (\cancel{\frac{1}{2}}-\cancel{\frac{1}{3}}) + \ldots + (\cancel{\frac{1}{n}} - \frac{1}{n+1})$

$= 2 - \frac{1}{n+1}$

π.χ $S_{9999} = 1.9999$

$H (S_n)_{n\in N}$ είναι αύξουσα και το $\lim_{n\to\infty} S_n = 2$

<u>1ος Αλγόριθμος</u>: Αθροίζουμε από τον πιο μεγάλο προς τον πιο μικρό όρο (όλοι οι όροι είναι θετικοί)

$S_0 = 1$, $S_k = S_{k-1} + \frac{1}{k(k+1)}$, $k=1,\ldots,n$

Με $b=10$, $t=10$ παίρνουμε $\tilde{S}_{9999} = 1.999\boxed{8999742}$

<u>2ος Αλγόριθμος</u>: Αθροίζουμε από τον μικρότερο προς τον μεγαλύτερο όρο.

$T_0 = \frac{1}{n(n+1)}$

$T_k = T_{k-1} + \frac{1}{(n-k)(n-k+1)}$, $k=1,\ldots,n-1$

$T_n = T_{n-1} + 1$

~~Επομένως~~ Προφανώς $T_n = S_n$

Με $b=10$, $t=10$ παίρνουμε

$\tilde{T}_{9999} = 1.999900000$

Ερώτημα: Γιατί με τον 2º Αλγόριθμο παίρναμε καλύτερο αποτέλεσμα;

Βοηθητικό αποτέλεσμα

Έστω $\lambda, \mu \in \mathbb{R}$ και $\varepsilon_1, \varepsilon_2 \in [-u, u]$

τότε υπάρχει $\varepsilon_3 \in [-u, u]$ τω

$$(*) \quad \lambda \varepsilon_1 + \mu \varepsilon_2 = (|\lambda| + |\mu|) \cdot \varepsilon_3$$

- Αν $\lambda = \mu = 0$, η $(*)$ ισχύει για οποιοδήποτε $\varepsilon_3 \in [-u, u]$
- Διαφορετικά θέτουμε

$$\varepsilon_3 := \frac{\lambda \varepsilon_1 + \mu \varepsilon_2}{|\lambda| + |\mu|}$$

Τότε ισχύει η $(*)$, και

$$|\varepsilon_3| = \frac{|\lambda \varepsilon_1 + \mu \varepsilon_2|}{|\lambda| + |\mu|} = \frac{\overbrace{|\lambda| \cdot |\varepsilon_1| + |\mu| \cdot |\varepsilon_2|}^{\leq u}}{|\lambda| + |\mu|}$$

$$\leq \frac{(|\lambda| + |\mu|) \cdot u}{|\lambda| + |\mu|} = u$$

Πρόβλημα: Έστω $a_1, a_2, \ldots, a_N \in M$

Θέλουμε να υπολογίσουμε το άθροισμα

$$S_N = \sum_{i=1}^{N} a_i$$

Θεωρούμε τον αλγόριθμο

$$s_1 = a_1, \quad S_k = S_{k-1} + a_k, \quad k = 2, \ldots, N$$

Παίρνουμε τις προσεγγίσεις:

$$\tilde{S}_1 = a_1, \quad \tilde{S}_k = fl(\tilde{S}_{k-1} + a_k), \quad k = 2, \ldots, N.$$

Έχουμε:

$$\tilde{S}_2 = fl(\underset{\underset{s_1}{\parallel}}{\tilde{S}_1} + a_2) = (\tilde{S}_1 + a_2)(1 + \overset{|\delta| \leq u}{\delta}) = S_2(1+\delta)$$

$$= S_2 + S_2 \cdot \delta = S_2 + |S_2| \cdot \varepsilon_2 \quad \text{με } |\varepsilon_2| \leq u$$

Παρόμοια: $\tilde{S}_3 = fl(\tilde{S}_2 + a_3) = (\tilde{S}_2 + a_3)(1 + \delta')$

$$= (S_2 + |S_2| \varepsilon_2 + a_3)(1 + \delta') = (S_3 + |S_2| \cdot \varepsilon_2)(1 + \delta')$$

$$= S_3 + (|S_2| \cdot \varepsilon_2 + S_3 \cdot \delta') + |S_2| \cdot \varepsilon_2 \cdot \delta' \rightarrow \text{τάξη } \underline{\underline{u^2}}$$

$$\Rightarrow \tilde{S}_3 \approx S_3 + (|S_2| \cdot \varepsilon_2 + S_3 \cdot \delta')$$

$$\Rightarrow \boxed{\tilde{S}_3 \approx S_3 + (|S_2| + |S_3|)\varepsilon_3}$$

Συνεχίζοντας εντελώς αντίστοιχα παίρνουμε

$$\tilde{S}_n \approx S_N + (|S_2| + |S_3| + \ldots + |S_N|)\,\varepsilon_N$$

με $|\varepsilon_N| \leq u$ και σφάλμα της τάξης $O(u^2)$

Με $\gamma_N := |S_2| + |S_3| + \ldots + |S_N|$

έχουμε

$$\tilde{S}_N \approx S_N + \gamma_N \cdot \varepsilon_N$$

οπότε

$$\left|\frac{\tilde{S}_N - S_N}{S_N}\right| \approx \frac{\gamma_N}{|S_N|} \cdot |\varepsilon_N| \leq \frac{\gamma_N}{|S_N|}\,u$$

Το $\rho_n := \dfrac{\gamma_N}{|S_n|}$ λέγεται συντελεστής μετάδοσης του σχετικού σφάλματος στρογγύλευσης για τον αλγόριθμό μας.

- ρ_N μεγάλο $\to$ ασταθής αλγόριθμος
- ρ_N μικρό $\to$ ευσταθής "

Το ρ_N είναι μεγάλο, αν κάποιο από τα μερικά αθροίσματα S_k έχει απόλυτη τιμή πολύ μεγαλύτερη της απόλυτης τιμής του S_N

Παράδειγμα: Προσέγγιση του e^{-x} για $x \gg 1$

Έχουμε

$$S_N(x) = 1 - x + \frac{x^2}{2} - \frac{x^3}{3!} + \ldots + (-1)^{N-1}\frac{x^{N-1}}{(N-1)!}$$

Τότε $S_N(x) \approx e^{-x}$ για μεγάλο N

Για $x = 100$ έχουμε $e^{-100} \approx 0$, ενώ

$S_1 = 1$, $S_2 = -99$, $S_3 = 4901$, $S_4 \approx -161766$ κλπ.

Ο υπολογισμός του $S_N(100)$ αποτυγχάνει καταγωδώς!

Εναλλακτικά:

$$e^{-x} = \frac{1}{e^x}$$

Το e^x για μεγάλο x προσεγγίζεται χωρίς πρόβλημα! (ανάπτυγμα Taylor)

Ειδική περίπτωση: $a_i > 0$, $i = 1, \ldots, N$

Τότε $y_N = S_2 + S_3 + \ldots + S_N$

$$= (N-1)a_1 + (N-1)a_2 + (N-2)a_3 + (N-3)a_4 + \ldots + 1 \cdot a_N$$

Το y_N παίρνει τη μικρότερη τιμή του αν τα $a_1, a_2 \le a_3 \le a_4 \le \ldots \le a_N$

Αυτή είναι η καλύτερη επιλογή

Χειρότερη επιλογή:

$$a_1, a_2 \ge a_3 \ge a_4 \ge \ldots \ge a_N$$

Τότε το y_N παίρνει τη μέγιστη τιμή του.

Σημείωση: Κάθε άθροισμα με μη ομόσημους όρους γράφεται ως διαφορά αθροισμάτων με θετικούς όρους.

Ευστάθεια αλγορίθμων

- Ένας αλγόριθμος λέγεται ασταθής, αν είναι ευαίσθητος σε σφάλματα στρογγύλευσης δηλαδή αν μικρά σφάλματα που γίνονται κατά την παράσταση των αριθμών και τις πράξεις σε έναν υπολογιστή, είναι δυνατόν να επιφέρουν μεγάλες μεταβολές στο τελικό αποτέλεσμα.
- Ένας αλγόριθμος λέγεται ευσταθής αν τα τελικά αποτελέσματα του δεν επηρεάζονται πολύ από τα σφάλματα στρογγύλευσης

<u>Παραδείγματα:</u> Υπολογισμός του e^{-x} για μεγάλο $x>0$.

$$e^{-x} = 1 - x + \frac{x^2}{2!} + \frac{x^3}{3!} - \dots$$

ασταθής όπως είδαμε.

$$e^{-x} = \frac{1}{e^{x}} = \frac{1}{1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots}$$ ευσταθής.

<u>Άλλα παραδείγματα:</u>

$$I_n = \int_0^1 x^n e^{x-1}\,dx \,,\ n = 1, 2, \dots$$

$$\begin{cases} I_1 = 1/e \\ I_n = 1 - n I_{n-1} \,, & n = 2, 3 \dots \end{cases}$$ ασταθής

Ξεκινώντας από ένα I_m και υπολογίζοντας προς τα πίσω τότε $I_{m-1}, I_{m-2}, \dots$ <u>Ευσταθής</u>

<u>Προσέγγιση του π:</u> $y_n = 2^n \cdot \sin\frac{\pi}{2^n}$, $n = 1, 2, \dots$

$$\begin{cases} y_1 = 2 \\ y_{n+1} = 2^{n+1}\sqrt{\frac{1}{2}\left(1 - \sqrt{1 - (2^{-n} y_n)^2}\right)} \end{cases} \quad n = 1, 2 \dots$$

<u>Ασταθής αλγόριθμος</u>

$$y_1 = 2$$
$$y_{n+1} = \sqrt{\frac{2}{1 + \sqrt{1 - (2^{-n} y_n)^2}}} \quad n = 1, 2, \dots$$

<u>Ευσταθής αλγόριθμος</u>

Κατάσταση προβλημάτων

- Λέμε ότι ένα πρόβλημα έχει καλή κατάσταση, αν μικρές μεταβολές στα δεδομένα του έχουν συνέπεια μικρή μεταβολή της λύσης του.
- Λέμε ότι ένα πρόβλημα έχει κακή κατάσταση, αν μικρές μεταβολές στα δεδομένα του μπορούν να έχουν ως συνέπεια μεγάλη η μεταβολή της λύσης του.

Παράδειγμα: $(x-2)^6 = 0 \iff x^* = 2$.

$$(x-2)^6 = 10^{-6} \Rightarrow x_k = 2 + \frac{1}{10} \cdot e^{\frac{2ik\pi}{6}} \quad \text{για } k=0,..,5.$$

$$\Rightarrow x_k - 2 = \frac{1}{10} \cdot e^{\frac{2ik\pi}{6}} \Rightarrow$$

$$\Rightarrow |x_k - \overset{x^*}{2}| = \frac{1}{10} \cdot \underbrace{|e^{\frac{2ik\pi}{6}}|}_{1}$$

$\boxed{\text{Άσκηση 1.15}}$

$$\Rightarrow |x_k - 2| = 10^{-1} \quad , k=0,..,5$$

Το πρόβλημα έχει κακή κατάσταση.

$(X-2)^6 \iff \boxed{x-2=0} \rightsquigarrow$ καλή κατάσταση

$\boxed{3-03-17}$

Άσκηση 1.15

$i = \sqrt{-1}$

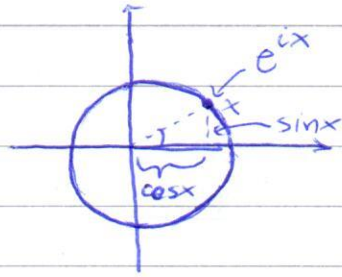
Οι ν-οστές ρίζες της μονάδας.

Έστω $\nu \in \mathbb{N}$. Οι λύσεις της εξίσωσης $z^\nu = 1$ είναι

$$z_k = e^{i \cdot \frac{2\pi k}{\nu}}, \quad k = 0, \dots, \nu-1$$

Για $x \in \mathbb{R}$ ισχύει

$$e^{ix} = \cos x + i \sin x \quad (\text{τύπος του Euler})$$



$$e^{ix} = 1 \iff x = \lambda \cdot 2\pi \quad \text{με } \lambda \in \mathbb{Z}$$

$x, y \in \mathbb{R}$

$$e^{ix} = e^{iy} \iff e^{i(x-y)} = 1 \Rightarrow$$
$$\Rightarrow x - y = 2\pi \cdot \lambda \quad \text{με } \lambda \in \mathbb{Z}$$

Έχουμε

$$(z_k)^\nu = \left(e^{i \cdot \frac{2\pi k}{\nu}}\right)^\nu$$

$$= e^{i(2\pi k)} = 1$$

Συμπέρασμα,

Τα $z_0, \dots, z_{\nu-1}$ είναι ν-οστές ρίζες της μονάδας.

Οι αριθμοί $\frac{2\pi \cdot 0}{\nu}, \frac{2\pi \cdot 1}{\nu}, \dots, \frac{2\pi(\nu-1)}{\nu}$ ανήκουν στο $[0, 2\pi]$ και είναι διαφορετικοί μεταξύ τους.

Άρα τα $z_0, \dots, z_{\nu-1}$ είναι διαφορετικά μεταξύ τους.
Επειδή το $z^\nu - 1$ είναι πολυώνυμο βαθμού ν, δεν έχει άλλες ρίζες πέραν των $z_0, \dots, z_{\nu-1}$

$\underline{\nu=1}$ $z_0=1$

$\underline{\nu=2}$ $z_0=1,\ z_1=-1$

Για $\nu \geq 3$ έχουμε $z_0=1$ και τα $z_0, z_1, \ldots, z_{\nu-1}$ είναι οι κορυφές του κανονικού ν-γώνου που είναι εγγεγραμμένο στον μοναδιαίο κύκλο.

<u>Εφαρμογή</u>

$$(x-2)^6 = 10^{-6} \Leftrightarrow$$

$$(x-2)^6 \cdot 10^6 = 1 \Leftrightarrow \quad [\underbrace{10\cdot(x-2)}_{z}]^{\overset{\nu}{6}} = 1 \Leftrightarrow$$

$$\Leftrightarrow 10(x_k-2) = e^{i\cdot\frac{2\pi k}{6}} \quad , k=0,\ldots,\overset{\nu-1}{5}$$

$$\Leftrightarrow x_k = 2+\frac{1}{10}\cdot e^{i\cdot\frac{2\pi k}{6}} \quad , k=0,\ldots,5$$

$$\boxed{|x_k-2| = 1/10}$$

<u>Άσκηση 1.2.</u>

α) $1-\cos x$, $|x|$ μικρή (χωρίς ανάπτυγμα Taylor)

$$1-\cos x = 2\sin^2\frac{x}{2}$$

$$\cos(2\alpha) = \cos^2\alpha - \sin^2\alpha \quad (\cos^2\alpha = 1-\sin^2\alpha)$$
$$= 1-2\sin^2\alpha$$
$$\Leftrightarrow 1-\cos(2\alpha) = \sin^2\alpha$$

β) $e^{x-y} = \dfrac{e^x}{e^y}$

γ) $\log x - \log y = \log\dfrac{x}{y}$

δ) $\sin(\alpha+x) - \sin\alpha =$, $|x|$ μικρή
$= 2\sin\frac{x}{2}\cdot\cos\left(\alpha+\frac{x}{2}\right)$

$\sin x - \sin y =$
$= 2\sin\frac{x-y}{2}\cdot\cos\frac{x+y}{2}$

Άσκηση 1.3

$$x^2 - 2ax + b = 0 \qquad a, b > 0, \quad a^2 \gg b$$

ρίζες $\underbrace{x_1 = a + \sqrt{a^2 - b}}_{\text{υπολογίζεται χωρίς πρόβλημα}}$, $x_2 = a - \sqrt{a^2-b}$ $\approx a$

αφαίρεση σχεδόν ίσων αριθμών!

$$x_1 \cdot x_2 = b \Rightarrow \boxed{x_2 = \frac{b}{x_1}} \Rightarrow x_2 = \frac{b}{a + \sqrt{a^2 - b}}$$

$p(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_0$

$x_1, \ldots, x_n$ ρίζες

Τότε $p(x) = a_n (x - x_1) \ldots (x - x_n)$

Άσκηση 1.4

$$x^3 + 3px + 2q = 0 \qquad p, q \in \mathbb{R}, \quad p^3 + q^2 > 0$$

α) Υπάρχει ακριβώς μια πραγματική ρίζα ρ, και

$\rho = u - v$ με $\begin{cases} u = \left(\sqrt{p^3 + q^2} - q\right)^{1/3} \\ v = \left(\sqrt{p^3+q^2} + q\right)^{1/3} \end{cases}$ (Τύπος του Cardano)

$\boxed{10-03-14}$

Άσκηση 1.4 (συνέχεια)

α) <u>Απόδειξη</u> $\qquad f(x) = x^3 + 3px + 2q$ (συνεχής)

<u>Ύπαρξη ρίζας.</u>

$$\lim_{x \to \infty} f(x) = \infty \qquad f(x) = x^3 \underbrace{\left(1 + \frac{3p}{x^2} + \frac{q}{x^3}\right)}_{1 \text{ για } x \to \infty}$$

$$\lim_{x \to -\infty} f(x) = -\infty$$

Η f παίρνει τόσο θετικές όσο και αρνητικές τιμές. Σύμφωνα με το Θεώρημα της ενδιάμεσης τιμής παίρνει και την τιμή 0, δηλαδή έχει (τουλάχιστον μία) πραγματική ρίζα.

Μοναδικότητα ρίζας.

$$f'(x) = 3x^2 + 3p = 3(x^2+p)$$

• $p \geq 0$: Τότε $f'(x) > 0 \quad \forall x \neq 0$

Επομένως η f είναι γνήσια αύξουσα.

Άρα η f έχει το πολύ μια ρίζα.

• $p < 0$:

$$f'(x) = 0 \iff x^2 + p = 0 \iff x^2 = -p \iff \begin{cases} x_1 = -\sqrt{-p} \\ x_2 = \sqrt{-p} \end{cases}$$

		$-\sqrt{-p}$		$\sqrt{-p}$	
f'	+	0	−	0	+
f	↗		↘		↗

Έχουμε ότι

$$f(-\sqrt{-p}) = p\sqrt{-p} - 3p\sqrt{-p} + 2q = 2(q - p\sqrt{-p})$$
$$f(\sqrt{-p}) = \ldots = 2(q + p\sqrt{-p})$$

οπότε

$$f(-\sqrt{-p}) \cdot f(\sqrt{-p}) = 4(q^2 + p^3) > 0$$

<u>1η περίπτωση</u>: $f(-\sqrt{-p}) > 0$

Τότε η f έχει ακριβώς μια ρίζα στο $(-\infty, -\sqrt{-p})$, και στα διαστήματα $[-\sqrt{-p}, \sqrt{-p}]$ και $(\sqrt{-p}, \infty)$ δεν έχει ρίζα (παίρνει θετικές τιμές). Άρα η f έχει ακριβώς μια ρίζα.

<u>2η περίπτωση</u>: $f(-\sqrt{-p}) < 0$

Τότε η f δεν έχει καμία ρίζα στα διαστήματα $(-\infty, -\sqrt{-p})$ και $[-\sqrt{-p}, \sqrt{-p}]$ (παίρνει αρνητικές τιμές) και έχει ακριβώς μια ρίζα στο $[\sqrt{-p}, \infty)$.

Τώρα,

$$f(\rho) = f(u-v) = (u-v)^3 + 3p(u-v) + 2q$$

$$= u^3 - 3u^2v + 3uv^2 - v^3 + 3p(u-v) + 2q$$

$$= u^3 - v^3 - 3uv(u-v) + 3p(u-v) + 2q$$

$$= -\cancel{2q} - 3vu(u-v) + 3p(u-v) + \cancel{2q}$$

$$= 3\underbrace{(p-uv)}_{0}(u-v) = 0$$

Όμως,

$$uv = \Big[\underbrace{\left(\sqrt{p^3+q^2} - q\right)\left(\sqrt{p^3+q^2} + q\right)}_{p^3+\cancel{q^2}-\cancel{q^2}}\Big]^{1/3} = (p^3)^{1/3} = p$$

β) $p^3 >> q^2 \rightsquigarrow$ ο τύπος του Cardano έχει πρόβλημα ευστάθειας.

Τότε $u \approx \sqrt{p}$

$v \approx \sqrt{p}$

(Υπολογίζοντας το ρ από τον τύπο $\rho = u - v$, αφαιρούμε σχεδόν ίσους αριθμούς).

γ) $u^3 - v^3 = (u-v)(u^2+uv+v^2)$

$$\Rightarrow \rho = u - v = \frac{\overbrace{u^3-v^3}^{=-2q}}{u^2+\underbrace{uv}_{p}+v^2}$$

$$\Rightarrow \rho = -\frac{2q}{u^2+\underbrace{p}_{>0}+v^2}$$

Δεν υπάρχει πρόβλημα ευστάθειας σε αυτόν τον υπολογισμό.

Άσκηση 1.7.

$$y_n = n \cdot \sin\frac{\pi}{n}$$
$$Y_n = n \cdot \tan\frac{\pi}{n}$$

Με ανάπτυγμα Taylor.

$$\left.\begin{aligned} y_n &= \pi - \frac{\pi^3}{6n^2} + O\left(\frac{1}{n^4}\right) \Rightarrow 2y_n = 2\pi - \frac{\pi^3}{3n^2} + O\left(\frac{1}{n^4}\right) \\ Y_n &= \pi + \frac{\pi^3}{3n^2} + O\left(\frac{1}{n^4}\right) \Rightarrow Y_n = \pi + \frac{\pi^3}{3n^2} + O\left(\frac{1}{n^4}\right) \end{aligned}\right\} \Rightarrow$$

$$\Rightarrow 2y_n + Y_n = 3\pi + O\left(\frac{1}{n^4}\right) \Rightarrow \underbrace{\frac{2}{3}y_n + \frac{Y_n}{3}}_{= z_n} = \pi + O\left(\frac{1}{n^4}\right)$$

Προέκταση κατά Richardson
(Richardson extrapolation)

Άσκηση 1.12.

$$y_n = \int_0^1 \frac{x^n}{x+\alpha} \cdot dx \quad , \quad n = 0, 1, 2, \ldots$$

α) Για $\alpha > 0$ η $(y_n)_{n \in \mathbb{N}}$ είναι γνησίως φθίνουσα και μηδενική.

$$y_{n+1} = \int_0^1 \frac{x^{n+1}}{x+\alpha} \cdot dx = \int_0^1 \frac{x^n \cdot \overset{\leq 1}{x}}{x+\alpha} \cdot dx$$

$$< \int_0^1 \frac{x^n}{x+\alpha} dx = y_n$$

$$\phi(x) \geq f(x) \Rightarrow \int_a^b f(x)\,dx \leq \int_a^b \phi(x)\,dx$$

$$0 \leq y_n = \int_0^1 \frac{x^n}{\underbrace{x+\alpha}_{\geq \alpha}} dx \leq \int_0^1 \frac{x^n}{\alpha} dx = \frac{1}{\alpha}\int_0^1 x^n dx$$

$$\Rightarrow 0 \leq y_n \leq \frac{1}{(n+1)\alpha}$$

$$\downarrow$$

$$0 \text{ για } n \to \infty$$

$$\Rightarrow y_n \to 0 \text{ για } n \to \infty$$

β) $y_n = \sum_{k=0}^{n-1} (-1)^k \binom{n}{k} a^k \frac{(1+a)^{n-k} - a^{n-k}}{n-k} + (-a)^n \log\frac{1+a}{a}$

$a >> 1$

Πώς προκύπτει

$y_n = \int_0^1 \frac{x^n}{x+a}\,dx \underset{z=x+a}{=} \int_a^{1+a} \frac{(z-a)^n}{z}\,dz$

Ευστάθεια;

$\rho_n = \frac{|S_2| + |S_3| + \dots + |S_n|}{|S_n|}$

- Η $|y_n|$ είναι μικρή.
- Για μεγάλο a, κάποιοι από τους όρους $\binom{n}{k} a^k \frac{(1+a)^{n-k} - a^{n-k}}{n-k}$ είναι μεγάλοι.

Άρα αφού το τελικό αποτέλεσμα είναι μικρό, κάποιο από τα ενδιάμεσα αθροίσματα έχει μεγάλη απόλυτη τιμή. Επομένως, ο αλγόριθμος είναι ασταθής.

γ) $y_{n-1} \to y_n$

y_0; ευστάθεια;

$$y_0 = \int_0^1 \frac{x^0}{x+a}\,dx = \int_0^1 \frac{1}{x+a}\,dx = \Big[\log(x+a)\Big]_{x=0}^{x=1}$$

$$= \log(1+a) - \log a = \log\left(\frac{1+a}{a}\right)$$

$$y_n = \int_0^1 \frac{x^n}{a+x}\,dx = \int_0^1 \frac{x^{n-1}(x+a) - a x^{n-1}}{x+a}\,dx$$

$$= \underbrace{\int_0^1 \frac{x^{n-1}\cancel{(x+a)}}{\cancel{x+a}}\,dx}_{=\frac{1}{n}} - \int_0^1 \frac{a x^{n-1}}{x+a}\,dx$$

$$= \frac{1}{n} - a \underbrace{\int_0^1 \frac{x^{n-1}}{x+a}\,dx}_{y_{n-1}} = \frac{1}{n} - a \cdot y_{n-1}$$

$$\begin{cases} y_n = \frac{1}{n} - a \cdot y_{n-1} \\ y_0 = \log\frac{a+1}{a} \end{cases}$$

$$\begin{cases} \tilde{y}_n = \frac{1}{n} - a\tilde{y}_{n-1}, \quad n=1,2,\dots \\ \tilde{y}_0 \end{cases}$$

Τότε, $y_n - \tilde{y}_n = -a(y_{n-1} - \tilde{y}_{n-1})$

$$\underset{\text{επαγωγικά}}{\Rightarrow} \quad y_n - \tilde{y}_n = (-a)^n (y_0 - \tilde{y}_0)$$

$$\Rightarrow |y_n - \tilde{y}_n| = a^n |y_0 - \tilde{y}_0|$$

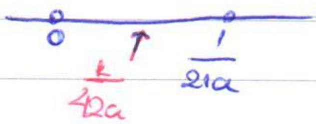
Το a^n αυξάνει ταχύτατα.

Άρα, ο αλγόριθμος είναι ασταθής.

δ) Ευσταθής αλγόριθμος για τον υπολογισμό του y_{10}.

$0 \le y_n \le \frac{1}{(n+1)a}$

$$y_n = \frac{1}{n} - a y_{n-1} \Rightarrow y_{n-1} = \frac{1}{a}\cdot\left(\frac{1}{n} - y_n\right)$$



$$\begin{cases} \tilde{y}_{20} = 0 \dots \\ \tilde{y}_{n-1} = \frac{1}{a}\left(\frac{1}{n} - \tilde{y}_n\right), \quad n = 20, 19, \dots, 11 \end{cases}$$

Άσκηση 1.13

$$\left.\begin{aligned} x + y &= 1 \\ x + (1-a)y &= 0 \end{aligned}\right\} \quad a \in \mathbb{R} \qquad \text{κατάσταση;}$$

Για $a=0$ το σύστημα δεν έχει λύση.

Έστω $a \neq 0$. Θεωρούμε το σύστημα

$$\begin{aligned} \tilde{x} + \tilde{y} &= 1 + \varepsilon_1 \\ \tilde{x} + (1-a)\tilde{y} &= 0 + \varepsilon_2 \end{aligned}$$

Θέτουμε $u = \tilde{x} - x$ και $v = \tilde{y} - y$

και αφαιρούμε τις αντίστοιχες εξισώσεις και παίρνουμε:

$$\left.\begin{aligned} u + v &= \varepsilon_1 \\ u + (1-a)v &= \varepsilon_2 \end{aligned}\right\}$$

$$\Rightarrow \left.\begin{aligned} u &= \varepsilon_1 - \frac{\varepsilon_1 - \varepsilon_2}{a} \\ v &= \frac{\varepsilon_1 - \varepsilon_2}{a} \end{aligned}\right\}$$

Για $|a|$ μεγάλη, έχουμε $|u|, |v|$ είναι μικρά, οπότε η κατάσταση είναι καλή. Αντίθετα, για $|a|$ μικρή, τα $|u|, |v|$ μπορεί να είναι μεγάλα, η κατάσταση είναι κακή.