ΥΠΟΛΟΓΙΣΤΙΚΑ ΜΑΘΗΜΑΤΙΚΑ

ΕΙΣΑΓΩΓΗ

Η ανάπτυξη πολύπλοκων υπολογιστικών συστημάτων, έκανε επιτακτική την ανάγκη οργάνωσης αριθμητικών μεθόδων, για την επίλυση πολύπλοκων προβλημάτων επιστημονικών εφαρμογών. Για την επίλυση ενός πολύπλοκου προβλήματος ακολουθούμε τα παρακάτω βήματα:

ΦΥΣΙΚΟ ΠΡΟΒΛΗΜΑ

↓

ΣΧΕΔΙΑΣΗ ΜΑΘΗΜΑΤΙΚΟΥ ΜΟΝΤΕΛΟΥ

(χαρακτηρίζεται συνήθως από ένα μεγάλο πλήθος εξισώσεων και αγνώστων, που καθιστούν την ακριβή επίλυση του προβλήματος πρακτικά αδύνατη, λόγω του τεράστιου όγκου πράξεων που απαιτούνται.**)**

↓

ΜΕΛΕΤΗ ΜΑΘΗΜΑΤΙΚΟΥ ΜΟΝΤΕΛΟΥ

(Μελέτη δείκτη κατάστασης προβλήματος (βλέπε **§** 1.6))

↓ → **ΥΠΟΛΟΓΙΣΜΟΣ ΣΦΑΛΜΑΤΟΣ ΥΠΟΛ. ΜΟΝΤΕΛΟΥ**

ΥΠΟΛΟΓΙΣΤΙΚΟ ΜΟΝΤΕΛΟ

(χαρακτηρίζεται συνήθως από ένα πεπερασμένο πλήθος πράξεων και βημάτων, που καθιστά εφικτή μία λύση του προβλήματος κατά προσέγγιση).

↓

ΑΛΓΟΡΙΘΜΟΣ ΥΛΟΠΟΙΗΣΗΣ ΥΠΟΛΟΓΙΣΤΙΚΟΥ ΜΟΝΤΕΛΟΥ

(πεπερασμένος αριθμός στοιχειωδών πράξεων, που κάποιος που δε γνωρίζει καθόλου το πρόβλημα να μπορεί να τις εκτελέσει, π.χ. ο Η.Υ.).

↓

ΜΕΛΕΤΗ ΑΛΓΟΡΙΘΜΙΚΩΝ ΣΦΑΛΜΑΤΩΝ

(Καθορισμός ακρίβειας και ανοχής, μελέτη ευστάθειας αλγορίθμου, επίδραση σφαλμάτων στρογγύλευσης και αποκοπής στους υπολογισμούς).

↓

ΠΡΟΣΕΓΓΙΣΤΙΚΗ ΛΥΣΗ

Η εύρεση μιας προσεγγιστικής λύσης ενός μαθηματικού μοντέλου (continuous Mathematics) με χρήση ενός Αριθμητικού μοντέλου μέσω της ανάπτυξης κατάλληλου αλγορίθμου, είναι το αντικείμενο της υπολογιστικής (αριθμητικής) ανάλυσης.

Ανέκαθεν υπήρξε η ανάγκη πρακτικών μαθηματικών υπολογισμών. Ένα από τα παλαιότερα μαθηματικά «κείμενα» είναι η πλάκα των Βαβυλωνίων YBC 7289, που δίνει μία αριθμητική προσέγγιση της $\sqrt{2}$ στο 60-αδικό σύστημα αρίθμησης. Ετσι λοιπόν, η σύγχρονη αριθμητική ανάλυση δεν ψάχνει ακριβείς απαντήσεις, όταν αυτές δεν είναι δυνατόν να επιτευχθούν, αλλά προσεγγιστικές λύσεις με μελέτη των σφαλμάτων. Η Αριθμητική Ανάλυση έχει εφαρμογές στις θετικές και φυσικές επιστήμες, όπως στη μηχανική με την επίλυση διαφορικών εξισώσεων, στη βελτιστοποίηση, στη γραμμική άλγεβρα κλπ.

Από τα προαναφερθέντα, φαίνεται ότι το πεδίο της αριθμητικής Ανάλυσης αναπτύχθηκε πολύ πριν την ανακάλυψη των Η/Υ. Η γραμμική παρεμβολή ήδη χρησιμοποιούνταν 2000 χρόνια πριν. Μεγάλοι μαθηματικοί είχαν αναπτύξει μεθόδους της αριθμητικής Ανάλυσης, όπως φαίνεται και από τα ονόματα σημαντικών αλγορίθμων, όπως της απαλοιφής Gauss, μεθόδου Euler και Newton κλπ. Είναι όμως σαφές ότι η ανακάλυψη των Η/Υ έδωσε τεράστια ώθηση στην Αριθμητική Ανάλυση, διότι επέτρεψε την υλοποίηση πολύπλοκων υπολογισμών.

Από τα παραπάνω, γίνεται σαφές ότι στα υπολογιστικά μαθηματικά ο στόχος είναι διττός:

μας ενδιαφέρει τόσο η ***εύρεση της προσεγγιστικής λύσης*** *ενός πολύπλοκου προβλήματος μέσω της ανάπτυξης κατάλληλου αλγορίθμου, όσο και ο* ***υπολογισμός του σφάλματος*** *ως μέσο εκτίμησης της χρησιμότητας του αριθμητικού μας μοντέλου.*

ΚΕΦΑΛΑΙΟ 1

ΑΡΙΘΜΗΤΙΚΗ ΠΕΠΕΡΑΣΜΕΝΗΣ ΑΚΡΙΒΕΙΑΣ - ΣΦΑΛΜΑΤΑ

§ 1.1 *Αναπαράσταση αριθμών σε οποιαδήποτε βάση*

Ορισμός 1.1.1 Εστω $p= 2, 3, 4,\ldots,\ N = 0,1,\ldots,$ τότε στο p-αδικό σύστημα αρίθμησης, κάθε ακέραιος αριθμός m τέτοιος ώστε $|\, m \,| < p^{N+1}$ εκφράζεται μονοσήμαντα ως ένα πολυώνυμο με βάση τον αριθμό p και συντελεστές $a_i \in \{0,1,\ldots, p-1\}$, $i = 0,1,\ldots$, ως εξής:

$$m = \pm \sum_{i=0}^{N} a_i p^i \ . \tag{1.1}$$

Η σχέση (1.1) καλείται ***p-αδική αναπαράσταση του m*** και οι συντελεστές a_i καλούνται ***ψηφία*** του αριθμού m ως προς τη ***βάση*** p. Στο εξής για συντομία, αντί της σχέσης (1.1) θα γράφουμε:

$$m = \pm\left(a_N a_{N-1}\ldots a_0\right)_p .$$

Ορισμός 1.1.2 Στο p-αδικό σύστημα αρίθμησης, κάθε πραγματικός αριθμός $x \in (0,1)$ εκφράζεται ως εξής:

$$x = \sum_{i=-\infty}^{-1} a_i p^i \ , \tag{1.2}$$

όπου $a_i \in \{0,1,\ldots, p-1\}$, $i = -1,-2,\ldots$ Η σχέση (1.2) καλείται ***p-αδική αναπαράσταση του x*** και οι συντελεστές a_i καλούνται ***ψηφία*** του αριθμού x ως προς τη *βάση* p. Στο εξής αντί της σχέσης (1.2) θα γράφουμε:

$$x = \left(0.\ a_{-1}a_{-2}\ldots\right)_p .$$

Θεώρημα 1.1.1 Κάθε πραγματικός αριθμός y τέτοιος ώστε $|\, y \,| < p^{N+1}$ εκφράζεται στο p-αδικό σύστημα αρίθμησης ως εξής:

$$y \ = \ \pm \sum_{i=-\infty}^{N} a_i p^i = \pm\left(a_N a_{N-1}\ldots a_0 \ .\ a_{-1}a_{-2}\ldots\right)_p . \tag{1.3}$$

Παρατήρηση 1 Επειδή οι Η/Υ χρησιμοποιούν το δυαδικό ή δεκαεξαδικό σύστημα αρίθμησης, ενώ τα εισερχόμενα δεδομένα και τα εξαγόμενα αποτελέσματα παρουσιάζονται στο δεκαδικό σύστημα που εμείς αντιλαμβανόμαστε, η μετατροπή ενός αριθμού από ένα σύστημα αρίθμησης σε άλλο γίνεται εσωτερικά από τον υπολογιστή.

Παράδειγμα 1 Να μετατραπεί ο αριθμός $(14.73)_{10}$ σε σύστημα με βάση το p.

Λύση Θα μετατρέψουμε πρώτα το ακέραιο μέρος του αριθμού και στη συνέχεια το κλασματικό μέρος.

(α) Θέλουμε να υπολογίσουμε τα ψηφία $a_0, a_1, ..., a_N \in \{0, ..., p-1\}$ έτσι ώστε:

$$14 = a_0 + a_1 p + ... + a_N p^N = a_0 + p(a_1 + a_2 p + ... + a_N p^{N-1}) = a_0 + p\ \pi_0(14),$$

όπου $\pi_0(14) = a_1 + a_2 p + ... + a_N p^{N-1}$. Από την παραπάνω σχέση παρατηρούμε ότι

$$\boldsymbol{\pi_0(14)} \textbf{ = πηλίκο της διαίρεσης } \boldsymbol{14/p}$$

και

$$\boldsymbol{a_0} \textbf{ = υπόλοιπο της διαίρεσης } \boldsymbol{14/p.}$$

Εστω $\boldsymbol{\pi_n(14) = a_{n+1} + a_{n+2} p + ... + a_N p^{N-(n+1)}}$, $n = 1, ..., N-1$, συνεχίζοντας με τον ίδιο τρόπο αναδρομικά είναι εύκολο να δει κανείς ότι:

$$\boldsymbol{\pi_n(14)} \textbf{ = πηλίκο της διαίρεσης } \boldsymbol{\pi_{n-1}(14)/p}$$

και

$$\boldsymbol{a_n} \textbf{ = υπόλοιπο της διαίρεσης } \boldsymbol{\pi_{n-1}(14)/p.}$$

(β) Θέλουμε να υπολογίσουμε τα ψηφία $a_{-1}, a_{-2}, ..., \in \{0, ..., p-1\}$ έτσι ώστε:

$$0.73 = a_{-1} p^{-1} + a_{-2} p^{-2} + ...$$

Πολλαπλασιάζουμε και τα δύο μέλη με p και έχουμε:

$$p\ 0.73 = a_{-1} + a_{-2} p^{-1} + a_{-3} p^{-2} + ...,$$

άρα αν $k_{-1}(0.73) = a_{-2}p^{-1} + a_{-3}p^{-2} + \ldots$, τότε:

$$k_{-1}(0.73) = \textbf{κλασματικό μέρος του αριθμού } (p\ 0.73)$$

και

$$a_{-1} = [0.73\ p] = \textbf{ακέραιο μέρος του αριθμού } (p\ 0.73).$$

Εστω $k_n(0.73) = a_{n-1}p^{-1} + a_{n-2}p^{-2} + \ldots$, $n = -2, -3, \ldots$, συνεχίζοντας με τον ίδιο τρόπο αναδρομικά, είναι εύκολο να δει κανείς ότι:

$$k_n(0.73) = \textbf{κλασματικό μέρος του αριθμού } (p\ k_{n+1}(0.73))$$

και

$$a_n = \left[k_{n+1}(0.73)\ p\right] = \textbf{ακέραιο μέρος του αριθμού } (p\ k_{n+1}(0.73)). \quad \square$$

Παρατήρηση 2 Κατά τη μετατροπή ενός αριθμού από ένα σύστημα αρίθμησης σε ένα άλλο, το πλήθος των ψηφίων του κλασματικού μέρους του μπορεί από πεπερασμένο να γίνει άπειρο ή και αντίστροφα.

Παρατήρηση 3 Η μετατροπή ενός αριθμού από ένα p-αδικό σύστημα αρίθμησης στο δεκαδικό σύστημα γίνεται άμεσα με χρήση της σχέσης (1.3).

§ 1.2 *Αριθμοί μηχανής*

Εφ' όσον χρησιμοποιούμε ηλεκτρονικούς υπολογιστές για την επίλυση προβλημάτων αριθμητικής φύσεως, πρέπει να έχουμε έναν τρόπο να **αναπαραστήσουμε αριθμούς σε υπολογιστή**, διότι ενώ οι αριθμοί μπορεί να είναι άπειροι σε πλήθος ή μέγεθος, ο ηλεκτρονικός υπολογιστής έχει πεπερασμένες δυνατότητες μνήμης. Αυτή λοιπόν η αναπαράσταση, που καλείται **αριθμητική πεπερασμένης ακρίβειας** επιφέρει σφάλματα στους υπολογισμούς.

Χωρίς περιορισμό της γενικότητας, υποθέτουμε ότι y είναι ένας πραγματικός αριθμός τέτοιος ώστε $p^N \le |y| < p^{N+1}$, τότε από τη σχέση (1.3) έχουμε:

$$y = \pm\sum_{i=-\infty}^{N} a_i p^i \overset{i=N+1-j}{=} \pm\sum_{j=1}^{\infty} a_{N+1-j} p^{(N+1)-j} = \pm p^{N+1}\sum_{j=1}^{\infty} a_{N+1-j} p^{-j}$$

$$\overset{b_j=a_{N+1-j}}{=} \pm p^{N+1}\sum_{j=1}^{\infty} b_j p^{-j} = \pm p^{N+1}\ \left(0.\ b_1 b_2 ...\right)_p ,$$

Έχουμε λοιπόν:

Ορισμός 1.2.1 Κάθε μη μηδενικός πραγματικός αριθμός x είναι δυνατόν να γραφεί στη λεγόμενη ***κανονική μορφή κινητής υποδιαστολής***:

$$x = \pm\ (0.\ b_1 b_2 ...)\ p^e,\ b_1 \neq 0 ,$$

όπου e είναι θετικός ακέραιος που καλείται ***εκθέτης*** και $\pm\ (0.\ b_1 b_2 ...)$ είναι το μη ακέραιο (δεκαδικό) τμήμα του αριθμού το οποίο καλείται ***βάση*** (mantissa). Πρακτικά, η κανονική μορφή κινητής υποδιαστολής σημαίνει ότι η δεκαδική τελεία μετατοπίζεται, έτσι ώστε όλα τα ψηφία του αριθμού να βρίσκονται στα δεξιά της δεκαδικής τελείας και το πρώτο δεκαδικό ψηφίο b_1 να είναι διάφορο του μηδενός. Τα $b_1, b_2, ...$ είναι όλα ψηφία του p-αδικού συστήματος αρίθμησης.

Για την παράσταση ενός πραγματικού αριθμού, χρειάζονται συνήθως άπειρα ψηφία (βλέπε (1.3)), που δεν είναι δυνατόν να αποθηκευτούν στην πεπερασμένη μνήμη ενός Η/Υ. Επομένως, προσεγγίζουμε έναν πραγματικό αριθμό από τους λεγόμενους αριθμούς μηχανής:

Ορισμός 1.2.2 Κάθε αριθμός **κινητής υποδιαστολής** της μορφής:

$$x = \tilde{x}_n\ p^e ,$$

όπου

- ***(i)*** $\tilde{x}_n = \pm\ (0.\ b_1 b_2 ... b_n)$ και το n δηλώνει την **ακρίβεια**, δηλαδή το πλήθος των ψηφίων του κλασματικού μέρους του αριθμού,
- ***(ii)*** ο εκθέτης e παίρνει τις τιμές $e = -c, -c+1, ..., c-1, c$ για κάποιο θετικό ακέραιο c,

καλείται ***αριθμός μηχανής (floating point)***. Το σύνολο

$$A_M\ (p,n,c) = \left\{ \pm\left(0.\ b_1 ... b_n\right)\ p^e\ :\ 0 \le b_i \le p-1,\ b_1 \neq 0,\ |e| \le c \right\}$$

καλείται σύνολο των αριθμών μηχανής ως προς τις παραμέτρους p,n,c.

Παρακάτω δίδεται σχηματικά ο τρόπος αποθήκευσης ενός πραγματικού αριθμού σε λέξη με 32 bits.

1 bit (πρόσημο)	**Bits 2-24** (mantissa)	**bits 25-32** Εκθέτης *e*

Παράσταση στη μνήμη πραγματικού αριθμού σε Η/Υ με λέξη 32 bits στο δυαδικό σύστημα αρίθμησης. Το 1ο bit είναι *0* αν ο αριθμός είναι θετικός και *1* αν είναι αρνητικός. Ακρίβεια *n = 24*, *M = 128*.

§ *1.3 Ιδιότητες αριθμών μηχανής*

Πρόταση 1.3.1 Αν $x = \tilde{x}_n \; p^e$ είναι αριθμός μηχανής, τότε ισχύει:

$$p^{e-1} \;\le\; |x| \;\le\; \left(1-\frac{1}{p^n}\right) p^e .$$

Απόδειξη Επειδή:

$$\frac{1}{p} = \underbrace{(.10....0)_p}_{n-\psi\eta\phi\iota\alpha} \;\le\; \left|\tilde{x}_n = (.b_1...b_n)_p\right| \;\le\; \underbrace{(.aa....a)_p}_{n-\psi\eta\phi\iota\alpha,\ \boldsymbol{a=p-1}} = (p-1)\sum_{i=1}^{n} p^{-i} = 1-\frac{1}{p^n},$$

πολλαπλασιάζοντας με p^e παίρνουμε το ζητούμενο. □

Πόρισμα 1.3.1 Εστω $A_M(p,n,c)$ το σύνολο των αριθμών μηχανής ως προς τις παραμέτρους *p,n,c*, τότε:

(i) το σύνολο $A_M(p,n,c)$, αριθμεί $2\,(2c+1)\,(p-1)\,p^{n-1} + 1$ στοιχεία,

(ii) έχει ελάχιστο στοιχείο $\min_{A_M} = p^{-c-1}$ και μέγιστο στοιχείο $\max_{A_M} = \left(1-\frac{1}{p^n}\right) p^c$.

Απόδειξη Άμεση συνέπεια της Πρότασης 1.3.1 και της ανισότητας $|e| \le c$. □

Ορισμός 1.3.1 Οποιοσδήποτε αριθμός *x* απολύτως μικρότερος του ελαχίστου στοιχείου του συνόλου των αριθμών μηχανής $A_M(p,n,c)$, δηλαδή

$$|x| \;<\; p^{-c-1}$$

δεν μπορεί να αποθηκευθεί στη μνήμη και καλείται ***υποχείλιση (underflow)***. Ομοια, οποιοσδήποτε αριθμός *x* απολύτως μεγαλύτερος του μεγίστου στοιχείου του συνόλου των αριθμών μηχανής $A_M(p,n,c)$, δηλαδή

$$|x| > \left(1-\frac{1}{p^n}\right)p^c$$

επίσης δεν μπορεί να αποθηκευθεί στη μνήμη και καλείται ***υπερχείλιση (overflow)***.

Σημείωση 1 (Κατανομή των αριθμών μηχανής) Από το Πόρισμα 1.3.1 είναι σαφές ότι όλοι οι αριθμοί μηχανής κατανέμονται εντός των διαστημάτων

$$I_1 = \left[p^{-c-1}, \left(1-\frac{1}{p^n}\right)p^c\right], \quad I_2 = \left[-\left(1-\frac{1}{p^n}\right)p^c, -p^{-c-1}\right]. \qquad (1.4)$$

Χωρίς περιορισμό της γενικότητας ας θεωρήσουμε το διάστημα I_1. Για όλες τις τιμές του εκθέτη $e = -c,-c+1,\ldots,c-1,c$, είναι φανερό ότι το διάστημα I_1 διαμερίζεται σε $2c+1$ υποδιαστήματα ξένα μεταξύ τους ανά δύο:

$$\left[p^{-c-1}, \left(1-\frac{1}{p^n}\right)p^c\right] = \left[p^{-c-1}, \left(1-\frac{1}{p^n}\right)p^{-c}\right] \cup \left[p^{-c}, \left(1-\frac{1}{p^n}\right)p^{-c+1}\right] \cup \ldots \cup \left[p^{c-1}, \left(1-\frac{1}{p^n}\right)p^c\right]$$

Σε κάθε ένα από τα υποδιαστήματα

$$\left[p^{e-1}, \left(1-\frac{1}{p^n}\right)p^e\right], \; e = -c-1,\ldots,c.$$

αντιστοιχούν $(p-1)p^{n-1}$ αριθμοί μηχανής ***ισοκατανεμημένοι***. Οσο αυξάνεται ο εκθέτης κατά *1*, *p*-πλασιάζεται το μήκος του επόμενου υποδιαστήματος, συνεπώς ***οι αριθμοί μηχανής δεν είναι όλοι μεταξύ τους ισοκατανεμημένοι***. Πιο συγκεκριμένα, είναι πυκνά κατανεμημένοι πλησίον του μηδενός και αραιά κατανεμημένοι μακριά του μηδενός.

Παράδειγμα 2 Αν $p = 2, n = 3, e = -2,\ldots,2$ να βρεθούν και να παρασταθούν οι αριθμοί μηχανής.

Λύση Προφανώς οι αριθμοί μηχανής έχουν τη μορφή $x = \pm\,(0.\ b_1b_2b_3)\ 2^e$, όπου $b_i = 0,1,\ b_1 \neq 0.$ Δοθέντος εκθέτη e και λαμβάνοντας υπόψην ότι $b_1 = 1$ έχουμε:

$$(0.\ 1b_2b_3)_2 = \{(0.\ 100)_2, (0.\ 101)_2, (0.\ 110)_2, (0.\ 111)_2\} = \left\{\frac{1}{2}, \frac{5}{8}, \frac{3}{4}, \frac{7}{8}\right\}.$$

Αρα το σύνολο των αριθμών μηχανής είναι:

$$x = \pm\,(0.\ b_1b_2b_3)_2\ 2^e = \pm\left\{\frac{2^e}{2}, \frac{5\ 2^e}{8}, \frac{3\ 2^e}{4}, \frac{7\ 2^e}{8} :\ e = -2, ..., 2\right\},$$

$$= \pm\left\{\left\{\frac{1}{8}, \frac{5}{32}, \frac{3}{16}, \frac{7}{32}\right\}, \left\{\frac{1}{4}, \frac{5}{16}, \frac{3}{8}, \frac{7}{16}\right\}, \left\{\frac{1}{2}, \frac{5}{8}, \frac{3}{4}, \frac{7}{8}\right\}, \left\{1, \frac{5}{4}, \frac{3}{2}, \frac{7}{4}\right\}, \left\{2, \frac{5}{2}, 3, \frac{7}{2}\right\}\right\}$$

Επιπλέον υπάρχει και το μηδέν. Παρατηρούμε ότι δεν υπάρχουν αριθμοί στα διαστήματα (0 , 1/8) και (-1/8 , 0). □

Παρατήρηση 4 Το σύνολο των αριθμών μηχανής $A_M(p,n,c)$ δεν έχει τις συνήθεις ιδιότητες των πραγματικών αριθμών π.χ. **δεν είναι κλειστό ως προς την πρόσθεση και τον πολ/σμό**. Για παράδειγμα το γινόμενο των ελαχίστων στοιχείων του συνόλου $A_M(p,n,c)$ **δεν** είναι στοιχείο του $A_M(p,n,c)$.

§ 1.4 Σφάλματα στρογγύλευσης και αποκοπής

Ορισμός 1.4.1 Αν $\bar{x}$ είναι μία προσέγγιση του x, καλούμε ***σφάλμα*** την ποσότητα

$$e_x = x - \bar{x}$$

και ***σχετικό σφάλμα*** την ποσότητα

$$\rho_x = \frac{x - \bar{x}}{x},\ x \neq 0\,.$$

Οι ποσότητες $|\varepsilon_x|$ και $|\rho_x|$ καλούνται ***απόλυτο σφάλμα*** και ***απόλυτο σχετικό σφάλμα*** αντίστοιχα.

Αν λοιπόν υποθέσουμε ότι $x = \pm\,(0.\,b_1 b_2 ...)\,p^e$, $b_1 \neq 0$ είναι ένας πραγματικός αριθμός κινητής υποδιαστολής εντός των διαστημάτων I_1 ή I_2 (βλέπε (1.4)) και εάν ο μέγιστος αριθμός ψηφίων που μπορούν να αποθηκευτούν στη μνήμη είναι n (βλέπε ορισμό 1.2.2), το ερώτημα που τίθεται είναι:

ποιος ο πλησιέστερος αριθμός μηχανής fl(x) προς τον x;

Υπάρχουν δύο τρόποι υπολογισμού του αριθμού *fl(x):*

(α) στρογγύλευση του νιοστού ψηφίου του *x* προς τα πάνω ή προς τα κάτω (π.χ. ο 13.3456 γίνεται 13.35 με στρογγύλευση στο $2^{ο}$ δεκαδικό ψηφίο ενώ γίνεται 13.3 με στρογγύλευση στο $1^{ο}$ δεκαδικό ψηφίο),

(β) αποκοπή όλων των ψηφίων μετά το νιοστό (π.χ. ο 13.345675 γίνεται 13.3456 με αποκοπή όλων των ψηφίων μετά το $4^{ο}$).

Θεώρημα 1.4.1 Για το σχετικό σφάλμα $\dfrac{|x - fl(x)|}{|x|}$, $x \neq 0$, το οποίο καλείται **μοναδιαίο σφάλμα στρογγύλευσης** ισχύει:

$$\frac{|x - fl(x)|}{|x|} \;\leq\; \begin{cases} \dfrac{1}{2}\,p^{1-n}, & \text{για στρογγύλευση} \\ p^{1-n}, & \text{για αποκοπή} \end{cases}.$$

Απόδειξη (a) Εστω ότι ο x δεν είναι αριθμός μηχανής και ο *fl(x)* υπολογίζεται με στρογγύλευση. Έστω x', x'' οι πλησιέστεροι προς τον x αριθμοί μηχανής έτσι ώστε $x' < x < x''$, τότε

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

Αν λοιπόν $x = \left(0.\,b_1 ... b_n b_{n+1} ...\right)\,p^k$, $-c \leq k \leq c$, τότε $x' = \left(0.\,b_1 ... b_n\right)\,p^k$ και εφόσον $x', x'' \in \left[\,p^{k-1}\,, \left(1 - \dfrac{1}{p^n}\right)\,p^k\,\right]$, όπου οι αριθμοί μηχανής είναι ισοκατανεμημένοι (βλέπε σημείωση 1), έχουμε

$$|x' - x''| = |\tilde{x}'_n - \tilde{x}''_n|\,p^k = p^{k-n}.$$

Εφόσον $x \in \left[p^{k-1} , \left(1 - \frac{1}{p^n}\right) p^k \right]$, δηλαδή $x \geq p^{k-1}$,έχουμε

$$\frac{|x - fl(x)|}{|x|} \leq \frac{|x' - x''|}{2|x|} \leq \frac{p^{k-n}}{2p^{k-1}} = \frac{1}{2} p^{1-n}.$$

(β) Εστω ότι ο x δεν είναι αριθμός μηχανής και ο $fl(x)$ υπολογίζεται με αποκοπή, τότε:

$$\frac{|x - fl(x)|}{|x|} \leq \frac{|x' - x''|}{|x|} \leq \frac{p^{k-n}}{p^{k-1}} = p^{1-n}. \quad \square$$

Σημαντικά ψηφία ενός δεκαδικού αριθμού είναι όλα τα ψηφία του αριθμού που βρίσκονται δεξιά του $1^{ου}$ μη μηδενικού ψηφίου (συμπεριλαμβανομένου και αυτού).

Ορισμός 1.4.2 Το σφάλμα που προκύπτει όταν χρησιμοποιούμε πεπερασμένο πλήθος βημάτων, αντί απείρου πλήθους βημάτων που απαιτείται για την επίτευξη ακριβούς αποτελέσματος καλείται ***σφάλμα αποκοπής (truncation error)***.

Παράδειγμα 3 Οταν η ποσότητα $S_N = \sum_{k=1}^{N} 1/k^2$ χρησιμοποιείται για τον υπολογισμό του αριθμού $\frac{\pi^2}{6} = \sum_{k=1}^{\infty} 1/k^2$, τότε έχουμε ένα σφάλμα αποκοπής όλων των όρων της σειράς μετά τον νιοστό όρο. Τέτοια σφάλματα εμφανίζονται πολύ συχνά σε αριθμητικούς υπολογισμούς ορισμένων ολοκληρωμάτων, σειρών κλπ.

§ 1.5 Διαδιδόμενα σφάλματα σε αριθμητικούς υπολογισμούς

Στο εξής θα δεχθούμε ότι οι πράξεις στον υπολογιστή γίνονται με βάση τον ακόλουθο κανόνα, που αποτελεί αρκετά ρεαλιστικό μοντέλο του πραγματικού μηχανισμού των πράξεων στο Η/Υ:

Αν με $\bullet$ συμβολίσουμε οποιαδήποτε από τις γνωστές πράξεις της αριθμητικής και αν x, y είναι πραγματικοί αριθμοί που μπορούν να

παρασταθούν κατά προσέγγιση από αριθμούς μηχανής, θα θεωρούμε ότι το αποτέλεσμα της πράξης $x \bullet y$ στον υπολογιστή, είναι ο αριθμός

$$x * y = fl\big(fl(x) \bullet fl(y)\big).$$

Υποθέτουμε δηλαδή, ότι πρώτα γίνεται η παράσταση των *x, y* σε αριθμούς μηχανής $fl(x), fl(y)$, έπειτα γίνεται η πράξη $fl(x) \bullet fl(y)$ με άπειρη ακρίβεια (στην πράξη με ακρίβεια *2n* ψηφίων κλάσματος) και το αποτέλεσμα της πράξης αυτής προσεγγίζεται από έναν αριθμό μηχανής.

Έστω:

$$x = fl(x) + \varepsilon_x,\ \ y = fl(y) + \varepsilon_y,\ \ fl(x) \bullet fl(y) = fl\big(fl(x) \bullet fl(y)\big) + \varepsilon_{fl(x) \bullet fl(y)},$$

(βλέπε ορισμό 1.4.1), τότε με προσθαφαίρεση του ιδίου όρου, το σφάλμα

$$\varepsilon = x \bullet y \ - \ x * y = \big(x \bullet y - fl(x) \bullet fl(y)\big) + \big(fl(x) \bullet fl(y) - fl\big(fl(x) \bullet fl(y)\big)\big)$$

$$= \varepsilon_{x \bullet y} + \varepsilon_{fl(x) \bullet fl(y)}. \qquad (1.5)$$

Ο $1^{ος}$ όρος στο δεξιό μέλος της (1.5) είναι το ***διαδιδόμενο σφάλμα*** και ο $2^{ος}$ όρος είναι το ***σφάλμα στρογγύλευσης*** κατά τον υπολογισμό της ποσότητας $fl(x) \bullet fl(y)$. Υποθέτοντας ότι το σφάλμα στρογγύλευσης είναι μικρό (βλέπε πρόταση 1.4.1), θα μελετήσουμε το διαδιδόμενο σφάλμα.

Πρόταση 1.5.1 Η μέγιστη τιμή του απολύτου σφάλματος του αθροίσματος ή της διαφοράς δύο αριθμών, ισούται με το άθροισμα των απολύτων σφαλμάτων των αριθμών αυτών.

Απόδειξη Έστω $x = \overline{x} + \varepsilon_x,\ \ y = \overline{y} + \varepsilon_y,\ x \pm y = \overline{x} \pm \overline{y} + \varepsilon_{x \pm y}$, τότε:

$$\varepsilon_{x \pm y} = \ (x \pm y) - (\overline{x} \pm \overline{y}) = (x - \overline{x}) \pm (y - \overline{y}) = \varepsilon_x \pm \varepsilon_y,$$

άρα $\left|\varepsilon_{x \pm y}\right| \le \ |\varepsilon_x| + |\varepsilon_y|$. □

Πρόταση 1.5.2 $\varepsilon_{x\,y} \simeq x\ \varepsilon_y + y\ \varepsilon_x$ και

$$\varepsilon_{x/y} \simeq \frac{y\ \varepsilon_x - x\ \varepsilon_y}{y^2}.$$

Απόδειξη $\varepsilon_{x\,y} = x\,y - \overline{x}\,\overline{y} = x\,y - (x-\varepsilon_x)\,(y-\varepsilon_y) = x\,\varepsilon_y + y\,\varepsilon_x - \varepsilon_y\,\varepsilon_x$. Λαμβάνοντας υπόψη ότι ο όρος $\varepsilon_y\varepsilon_x$ είναι μικρός, παίρνουμε το ζητούμενο. Όμοια:

$$\varepsilon_{x/y} = \frac{x}{y} - \frac{\overline{x}}{\overline{y}} = \frac{x}{y} - \frac{x-\varepsilon_x}{y-\varepsilon_y} = \frac{y\,\varepsilon_x - x\,\varepsilon_y}{y(y-\varepsilon_y)}.$$

Λαμβάνοντας υπόψη ότι ο όρος ε_y είναι μικρός παίρνουμε το ζητούμενο. □

Από την Πρόταση 1.5.2 συμπεραίνουμε, ότι ***μεγάλες τιμές του x και y ενδέχεται να αυξήσουν το σφάλμα γινομένου, ενώ μικρές τιμές του διαιρέτη y και μεγάλες τιμές του διαιρετέου x ενδέχεται να αυξήσουν το σφάλμα της διαίρεσης. <u>Τέτοιου είδους καταστάσεις θα πρέπει να αποφεύγονται, με αναδιάταξη υπολογισμών</u>***.

Πόρισμα 1.5.1 Η μέγιστη τιμή του απόλυτου σχετικού σφάλματος του γινομένου ή του πηλίκου δύο αριθμών, ισούται κατά προσέγγιση με το άθροισμα των απολύτων σχετικών σφαλμάτων των αριθμών αυτών.

Απόδειξη Από την πρόταση 1.5.2 και τον ορισμό 1.4.1 του σχετικού σφάλματος έχουμε:

$$\rho_{x\,y} = \frac{\varepsilon_{x\,y}}{x\,y} = \frac{x\,\varepsilon_y + y\,\varepsilon_x - \varepsilon_y\,\varepsilon_x}{x\,y} = \rho_y + \rho_x - \rho_x\,\rho_y.$$

Με την προϋπόθεση ότι $|\rho_y|,\ |\rho_x| \ll 1$, έχουμε $\rho_{x\,y} \simeq \rho_y + \rho_x$, ή $|\rho_{x\,y}| \le |\rho_y| + |\rho_x|$. Όμοια για το πηλίκο έχουμε:

$$\rho_{x/y} = \frac{\varepsilon_{x/y}}{x/y} = \frac{\dfrac{y\,\varepsilon_x - x\,\varepsilon_y}{y(y-\varepsilon_y)}}{x/y} = \frac{y\,\varepsilon_x - x\,\varepsilon_y}{x(y-\varepsilon_y)}.$$

Με την προϋπόθεση ότι $|\rho_y| \ll 1$, δηλαδή $|\varepsilon_y| \ll |y|$, έχουμε

$$\rho_{x/y} = \frac{y\,\varepsilon_x - x\,\varepsilon_y}{x(y-\varepsilon_y)} \simeq \frac{y\,\varepsilon_x - x\,\varepsilon_y}{x\,y} = \rho_x - \rho_y,$$

ή $|\rho_{x/y}| \leq |\rho_y| + |\rho_x|$. □

Τέλος παρατηρούμε ότι

$$\rho_{x\pm y} = \frac{\varepsilon_{x\pm y}}{x \pm y} \simeq \frac{\varepsilon_x \pm \varepsilon_y}{x \pm y} = \left(\frac{x}{x \pm y}\right)\rho_x \pm \left(\frac{y}{x \pm y}\right)\rho_y .$$

Η παραπάνω σχέση δηλώνει ότι ***θα πρέπει να αποφεύγεται η πρόσθεση ενός πολύ μεγάλου και ενός πολύ μικρού αριθμού ή η αφαίρεση δύο περίπου ίσων αριθμών***.

Παράδειγμα 4 *(Μελέτη σφαλμάτων στον υπολογισμό αθροισμάτων)*

Εστω ότι θέλουμε να υπολογίσουμε το άθροισμα $S = \sum_{k=1}^{N} x_k$, όπου x_k είναι αριθμοί κινητής υποδιαστολής που έχουν ήδη αποθηκευτεί στη μνήμη. Προσθέτουμε λοιπόν τους 2 πρώτους, στο αποτέλεσμα προσθέτουμε τον $3^{ο}$ κλπ, άρα:

$S_2 = fl(x_1 + x_2)$ και επειδή από τον ορισμό 1.4.1 προκύπτει ο τύπος:

$$\bar{x} = x(1 - \rho_x),$$

έχουμε:

$$S_2 = fl(x_1 + x_2) = (x_1 + x_2)(1 - \rho_{x_1+x_2}) = (x_1 + x_2) - (x_1 + x_2)\rho_{x_1+x_2} . \qquad (1.6)$$

όπου $|\rho_{x_1+x_2}| \leq \frac{1}{2} p^{1-n}$ (βλέπε Θεώρημα 1.4.1). Συνεχίζοντας έχουμε:

$$S_3 = fl(S_2 + x_3) = (S_2 + x_3)(1 - \rho_{S_2+x_3}),$$

όπου $|\rho_{S_2+x_3}| \leq \frac{1}{2} p^{1-n}$ και αντικαθιστώντας την τιμή της S_2 από τη σχέση (1.6), παίρνουμε:

$$S_3 = ((x_1 + x_2) - (x_1 + x_2)\rho_{x_1+x_2} + x_3)(1 - \rho_{S_2+x_3})$$

$$= (x_1 + x_2 + x_3) - (x_1 + x_2)\rho_{x_1+x_2} - (x_1 + x_2 + x_3)\rho_{S_2+x_3}$$

$$+(x_1+x_2)\rho_{x_1+x_2}\rho_{S_2+x_3}$$

$$\cong (x_1+x_2+x_3)-(x_1+x_2)\rho_{x_1+x_2}-(x_1+x_2+x_3)\rho_{S_2+x_3},$$

αγνοώντας ως αμελητέο τον τελευταίο όρο. Αναγωγικά υπολογίζουμε:

$$S_N \cong \sum_{k=1}^{N} x_k - (x_1+x_2)\rho_{x_1+x_2} - (x_1+x_2+x_3)\rho_{S_2+x_3} - \ldots - (x_1+\ldots+x_N)\rho_{S_{N-1}+x_N},$$

ή

$$S_N - S \cong -(x_1+x_2)\left(\rho_{x_1+x_2}+\ldots+\rho_{S_{N-1}+x_N}\right) - x_3\left(\rho_{S_2+x_3}+\ldots+\rho_{S_{N-1}+x_N}\right)$$

$$-x_4\left(\rho_{S_3+x_4}+\ldots+\rho_{S_{N-1}+x_N}\right)-\ldots-x_N\rho_{S_{N-1}+x_N},$$

άρα:

$$|S_N - S| \le (|x_1|+|x_2|)\left(|\rho_{x_1+x_2}|+\ldots+|\rho_{S_{N-1}+x_N}|\right)+|x_3|\left(|\rho_{S_2+x_3}|+\ldots+|\rho_{S_{N-1}+x_N}|\right)$$

$$+|x_4|\left(|\rho_{S_3+x_4}|+\ldots+|\rho_{S_{N-1}+x_N}|\right)+\ldots+|x_N||\rho_{S_{N-1}+x_N}|.$$

Παρατηρούμε ότι για να ελαχιστοποιηθεί το απόλυτο σφάλμα $|S_N - S|$, πρέπει εκ των προτέρων οι όροι του αθροίσματος να διαταχθούν έτσι ώστε

$$|x_1| \le |x_1| \le \ldots \le |x_N|.$$

§ 1.6 Ευστάθεια αλγορίθμων

Ενας αλγόριθμος που είναι ευαίσθητος σε σφάλματα στρογγύλευσης, δηλαδή όταν μικρά σφάλματα επιφέρουν μεγάλες αλλαγές στα τελικά αποτελέσματα, καλείται ***ασταθής***, διαφορετικά καλείται ***ευσταθής***.

Θα λέμε επίσης ότι ένα πρόβλημα είναι σε ***καλή κατάσταση***, όταν μικρές μεταβολές των δεδομένων προκαλούν μικρές μεταβολές στα

αποτελέσματα, διαφορετικά θα λέμε ότι το πρόβλημα είναι σε ***κακή κατάσταση***.

Ο δείκτης κατάστασης ενός προβλήματος ορίζεται ως εξής:

$$K_p = \frac{\textit{απόλυτο σχετικό σφάλμα αποτελεσμάτων}}{\textit{απόλυτο σχετικό σφάλμα δεδομένων}}.$$

Παράδειγμα 5 Η λύση της πολυωνυμικής εξίσωσης $(x-2)^6 = 0$ είναι προφανώς η $x = 2$ πολλαπλότητας 6. Μεταβάλλοντας όμως ελάχιστα το σταθερό συντελεστή του πολυωνύμου, π.χ. αντικαθιστώντας το 0 με το 10^{-6} παίρνουμε την εξίσωση $(x-2)^6 = 10^{-6}$ η οποία έχει τις (μιγαδικές) ρίζες

$$x_k = 2 + \frac{1}{10} e^{2\pi i k/6},\ k = 0,...,5.$$

Δηλαδή μία μικρή διαταραχή στα δεδομένα του προβλήματος, επιφέρει μία αρκετά μεγάλη διαταραχή στη λύση, αφού $|x_k - 2| = \frac{1}{10}$. Το πρόβλημά μας είναι δηλαδή σε κακή κατάσταση. Είναι προφανές ότι όταν ένα πρόβλημα είναι σε κακή κατάσταση, τότε κάθε μέθοδος για την επίλυσή του είναι ασταθής, λόγω της παρουσίας σφαλμάτων στρογγύλευσης.

Παράδειγμα 6 Εστω $y_k = 2^k \varepsilon\phi\left(\frac{\pi}{2^k}\right),\ k = 2,\ldots.$

(α) ΝΔΟ η ακολουθία y_k παράγεται από την αναδρομική σχέση:

$$y_{k+1} = \begin{cases} 4, & k = 2 \\ 2^{2k+1} \dfrac{\sqrt{1 + \left(2^{-k} y_k\right)^2} - 1}{y_k} & k > 2 \end{cases}.$$

(β) Αν κάνουμε τις πράξεις με αριθμητική κινητής υποδιαστολής, παρατηρούμε ότι ο αλγόριθμος είναι ασταθής. Εξηγήστε την αιτία της αστάθειας και βρείτε έναν ευσταθή αλγόριθμο για τον υπολογισμό των τιμών της ακολουθίας y_k.

Λύση: **(α)** Επειδή $\lim_{x\to 0} \frac{\varepsilon\phi(a\,x)}{x} = a$, έχουμε ότι $\lim_{k\to\infty} y_k = \pi$.

$$y_{k+1}=2^{k+1}\varepsilon\phi\left(\frac{\pi}{2^{k+1}}\right)=2^{k+1}\frac{\eta\mu\left(\frac{\pi}{2^{k+1}}\right)}{\sigma\upsilon\nu\left(\frac{\pi}{2^{k+1}}\right)}=2^{k+1}\frac{2\eta\mu^2\left(\frac{\pi}{2^{k+1}}\right)}{2\eta\mu\left(\frac{\pi}{2^{k+1}}\right)\sigma\upsilon\nu\left(\frac{\pi}{2^{k+1}}\right)}$$

$$=2^{k+1}\frac{1-\sigma\upsilon\nu\left(\frac{\pi}{2^k}\right)}{\eta\mu\left(\frac{\pi}{2^k}\right)}=2^{k+1}\frac{\frac{1-\sigma\upsilon\nu\left(\frac{\pi}{2^k}\right)}{\sigma\upsilon\nu\left(\frac{\pi}{2^k}\right)}}{\varepsilon\phi\left(\frac{\pi}{2^k}\right)}=2^{2k+1}\frac{\frac{1}{\sigma\upsilon\nu\left(\frac{\pi}{2^k}\right)}-1}{y_k}$$

$$=2^{2k+1}\frac{\sqrt{\frac{1}{\sigma\upsilon\nu^2\left(\frac{\pi}{2^k}\right)}}-1}{y_k}=2^{2k+1}\frac{\sqrt{1+\varepsilon\phi^2\left(\frac{\pi}{2^k}\right)}-1}{y_k}=2^{2k+1}\frac{\sqrt{1+\left(y_k2^{-k}\right)^2}-1}{y_k}.$$

(β) Προφανώς επειδή $y_k2^{-k}\to 0,\ k\to\infty$ στον αριθμητή της αναδρομικής σχέσης έχουμε αφαίρεση σχεδόν ίσων αριθμών, που καλό είναι να αποφεύγεται σε αριθμητική πεπερασμένης ακρίβειας. Για να αποφύγουμε λοιπόν ενδεχόμενη καταστροφή σημαντικών ψηφίων, πολ/ζουμε και διαιρούμε με τη συζυγή παράσταση:

$$y_{k+1}=2^{2k+1}\frac{\sqrt{1+\left(2^{-k}y_k\right)^2}-1}{y_k}=2^{2k+1}\frac{\left(\sqrt{1+\left(2^{-k}y_k\right)^2}-1\right)\left(\sqrt{1+\left(2^{-k}y_k\right)^2}+1\right)}{y_k\left(\sqrt{1+\left(2^{-k}y_k\right)^2}+1\right)}$$

$$y_{k+1}=2^{2k+1}\frac{\left(2^{-k}y_k\right)^2}{y_k\left(\sqrt{1+\left(2^{-k}y_k\right)^2}+1\right)}=\frac{2y_k}{\sqrt{1+\left(2^{-k}y_k\right)^2}+1}.\qquad\square$$

ΛΥΜΕΝΕΣ ΑΣΚΗΣΕΙΣ

1. *Υπάρχουν αριθμοί μηχανής που ικανοποιούν την εξίσωση x+1 = x; Προσδιορίστε μία καλή προσέγγιση του μεγαλύτερου αριθμού x τέτοιου ώστε sin(x) = x.*

Λύση Εστω $x = (0.b_1...b_n)_p \; p^e$ αριθμός μηχανής, τότε για να μην είναι ο *x + 1* αριθμός μηχανής θα πρέπει $|x - x'| > 2$ όπου x' ο πλησιέστερος του *x* αριθμός μηχανής. Εφόσον

$$x' = (0.b_1...b_n + p^{-n})_p \; p^e = x + p^{e-n}$$

θα πρέπει να ισχύει $|x - x'| > 2$, δηλαδή

$$p^{e-n} > 2 \Rightarrow p^{e-n} > p^{\log_p 2} \Rightarrow e > n + \log_p 2,$$

Για όλους λοιπόν τους αριθμούς μηχανής

$$x = (0.b_1...b_n)_p \; p^e \;\; : \;\; e \geq n + log_p 2 \geq n+1$$

έχουμε ότι ο *x + 1* δεν είναι αριθμός μηχανής και *fl(x+1) = x*.

(β) Επειδή η μοναδική ρίζα της εξίσωσης *sin(x) = x* είναι η τιμή *x = 0,* ο πλησιέστερος αριθμός μηχανής είναι ο p^{-c-1} (βλέπε σημείωση 1).

2. *Βρείτε κατάλληλους τρόπους ώστε να μην χάνεται ακρίβεια όταν οι πράξεις γίνονται με αριθμητική κινητής υποδιαστολής πεπερασμένης ακρίβειας.*

(α) 1-cos(x) για μικρό |x|
(β) e^{x-y}*, x, y θετικοί*
(γ) log(x) – log(y) για μεγάλα θετικά x,y
(δ) sin(α+x) – sin(α) για μικρό |x|
(ε) τοξεφ(x) – τοξεφ(y) για μεγάλα θετικά x,y

Λύση *(α) 1-cos(x) =* $2 \sin^2(x/2)$.

(β) $e^{x-y} = \dfrac{e^x}{e^y} = \dfrac{\sum_{k=0}^{\infty} \frac{x^k}{k!}}{\sum_{k=0}^{\infty} \frac{y^k}{k!}} = \dfrac{\sum_{k=0}^{N} \frac{x^k}{k!} + \varepsilon_x}{\sum_{k=0}^{N} \frac{y^k}{k!} + \varepsilon_y} \cong \dfrac{\sum_{k=0}^{N} \frac{x^k}{k!}}{\sum_{k=0}^{N} \frac{y^k}{k!}}$ (βλέπε επίσης Πρόταση 1.5.2, πόρισμα 1.5.1 για το σφάλμα και σχετικό σφάλμα πηλίκου).

(γ) $\log(x) - \log(y) = \log(x/y)$ (ιδιότητα λογαρίθμου).

(δ) $\sin(\alpha+x) - \sin(\alpha) = 2\cos(a + x/2)\sin(x/2)$ (τύπος τριγωνομετρίας)

(ε) Εστω $x > y$ τότε:

$$\tau o \xi \varepsilon \varphi(x) - \tau o \xi \varepsilon \varphi(y) = \int_y^x \frac{1}{1+t^2} dt = \sum_{k=0}^{N} \frac{1}{1+\left(y + k\frac{x-y}{N}\right)^2} \frac{1}{N} + \varepsilon_N,$$

όπου $|\varepsilon_N| \leq \dfrac{1}{N}$. □

3. *Θεωρείστε τη δευτεροβάθμια εξίσωση*

$$x^2 - 2ax + b = 0,\ a,b > 0,\ a^2 >> b.$$

Δώστε έναν ευσταθή αλγόριθμο για τον υπολογισμό των ριζών της.

Λύση Προφανώς $\rho_{1,2} = a \pm \sqrt{a^2 - b}$. Επειδή $a^2 >> b$, $\sqrt{a^2 - b} \cong a$, οπότε έχουμε αστάθεια που προκαλείται από την αφαίρεση δύο σχεδόν ίσων αριθμών σε μία από τις δύο ρίζες. Υπολογίζουμε λοιπόν τη ρίζα $\rho_1 = a + \sqrt{a^2 - b}$ για την οποία δεν παρατηρείται καμία αστάθεια και στη συνέχεια για τον υπολογισμό της ρίζας $\rho_2 = a - \sqrt{a^2 - b}$ χρησιμοποιούμε τον τύπο του γινομένου ριζών $\rho_1 \rho_2 = b \Rightarrow \rho_2 = \dfrac{b}{\rho_1} = \dfrac{b}{a + \sqrt{a^2 - b}}$. □

4. *Σε αριθμητικό σύστημα με βάση p = 10, n = 4 και c = 4 να βρεθούν: (α) η μονάδα μηχανής ε (β) το μοναδιαίο σφάλμα στρογγύλευσης (γ) πότε συμβαίνει υποχείλιση και υπερχείλιση.*

Λύση (α) Η μονάδα μηχανής είναι μία ποσότητα ε, που αν προσθεθεί στον αριθμό *1* τον αφήνει αναλλοίωτο. Δηλαδή όλοι οι αριθμοί εντός του

διαστήματος *(-ε,ε)* παίζουν το ρόλο του «μηδενός» του Η/Υ. Υπολογίζεται από τη σχέση $|\varepsilon| \le \frac{1}{2} p^{1-n} = \frac{1}{2}\, 10^{1-4} = 0.0005$ (βλέπε Θεώρημα 1.4.1).
(β)-(γ) Αμεσες συνέπειας της θεωρίας. □

5. *Εστω x, y αριθμοί μηχανής με x≈y.*

(α) Εκτιμήστε το σχετικό σφάλμα κατά την αφαίρεση x-y.
(β) Πως θα υπολογίζατε το x^2-y^2: ως (x-y) (x+y) ή ως x x- y y;

Λύση (α) $\rho = \left|\frac{fl(x-y)-(x-y)}{x-y}\right| \le \begin{cases} 1, & \text{αν έχουμε υποχείλιση} \\ \frac{1}{2}\, p^{1-n}, & \text{αν δεν έχουμε υποχείλιση (βλέπε Θεώρημα 1.4.1)} \end{cases}$

(β) Στην 1η περίπτωση, πρώτα θα γίνει η πράξη *x-y* με άπειρη ακρίβεια, μετά θα μετατραπεί η ποσότητα *x-y* σε αριθμό μηχανής, στη συνέχεια θα γίνει ο πολ\σμός των αριθμών μηχανής *fl(x-y) fl(x-y)* με άπειρη ακρίβεια και τέλος το αποτέλεσμα θα μετατραπεί σε αριθμό μηχανής. Λαμβάνοντας υπόψη τον ορισμό του σχετικού σφάλματος:

$$\rho_x = \frac{x-\overline{x}}{x} \Rightarrow \overline{x} = x(1-\rho_x),$$

και το ακόλουθο:

Θεώρημα Α: Αν $|\rho_i| \le c < 1$, *i = 1,...,m* τότε υπάρχει $|\rho'| \le c < 1$:

$$|\rho'| \le c < 1:\ \prod_{i=1}^{m}(1+\rho_i) = (1+\rho')^m,$$

έχουμε: $fl\big(fl(x-y)\, fl(x+y)\big) = fl(x-y)\, fl(x+y)(1-\rho_1)$

$$= (x-y)(1-\rho_2)(x+y)(1-\rho_3)(1-\rho_1) = (x^2-y^2)(1-\rho)^3,$$

(βλέπε Θεώρημα Α)

$$= (x^2-y^2)(1-3\rho+3\rho^2-\rho^3).$$

Αρα: $|\rho_{(x-y)(x+y)}| \simeq 3|\rho|$, θεωρώντας ότι $|\rho^2|, |\rho^3| << 1$.

Ομοια εργαζόμαστε για την άλλη περίπτωση και παίρνουμε:

$$fl\left(fl\left(x\,x\right)\text{-}fl\left(y\,y\right)\right)=\left(fl\left(x\,x\right)\text{-}fl\left(y\,y\right)\right)\left(1-\rho_1\right)$$

$$=\left(x^2\left(1-\rho_2\right)-y^2\left(1-\rho_3\right)\right)\left(1-\rho_1\right)$$

$$=x^2\left(1-\rho_1'\right)^2-y^2\left(1-\rho_2'\right)^2,$$

Για το σχετικό σφάλμα έχουμε λοιπόν

$$\left|\frac{fl\left(fl\left(x\,x\right)\text{-}fl\left(y\,y\right)\right)-\left(x^2-y^2\right)}{\left(x^2-y^2\right)}\right|=\left|\frac{x^2\rho_1'(\rho_1'+2)-y^2\rho_2'(\rho_2'+2)}{\left(x^2-y^2\right)}\right|,$$

το οποίο λόγω παρανομαστή ενδέχεται να οδηγήσει σε μεγάλα σφάλματα. Αρα προτιμούμε την $1^{η}$ μέθοδο. □

ΑΛΥΤΕΣ ΑΣΚΗΣΕΙΣ

1. Να μετατραπεί ο αριθμός $(17.015625)_{10}$ στο δυαδικό σύστημα αρίθμησης.

2. Να γίνει (α) αποκοπή και (β) στρογγυλοποίηση, με *6* σημαντικά ψηφία στους ακόλουθους αριθμούς: *0.674595821, 0.003742534, 0.123455578*. Στην κάθε περίπτωση υπολογίστε το απόλυτο και το σχετικό σφάλμα.

3. Να γίνουν οι παρακάτω πράξεις: (α) ακριβώς (β) με αποκοπή διατηρώντας *3* σημαντικά ψηφία (γ) με στρογγυλοποίηση διατηρώντας *3* σημαντικά ψηφία σωστά.

(i) 13.*2 + 0.0841* (ii) *0.0314 * 129* (iii) (*132+0.713) - (112+22)*.

4. Έστω $q = 1000$. Να βρεθεί μεταξύ ποιών τιμών βρίσκεται η προσέγγιση q^*, όταν το φράγμα του σχετικού σφάλματος με στρογγυλοποίηση είναι $0.5x10^{-4}$.

5. Αν $p = 2,\ n = 4,\ e = -3,\ldots,3$ να βρεθούν και να παρασταθούν οι αριθμοί μηχανής.

6. Θεωρείστε το σύνολο των αριθμών μηχανής του Παραδείγματος 2 (σελ. 8). Επιλέξτε 3 οποιουσδήποτε αριθμούς μηχανής *x, y, z* και εξετάστε αν ισχύουν οι ισότητες:

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

$$(x\, y)\, z = x\, (y\, z)$$

$$x\, (y + z) = x\, y + x\, z.$$

Υπόδειξη: Θεωρείστε ότι οι πράξεις γίνονται στον Η/Υ (βλέπε σελ. 12 και λυμένη άσκηση 5).

7. Διερευνήστε την κατάσταση επίλυσης του προβλήματος:

$$\begin{cases} x + y = 1 \\ x + (1 - a)y = 0 \end{cases}.$$

8. Θεωρούμε την ακολουθία: $y_n = \int_0^1 \frac{x^n}{x + a} dx,\ n = 0,1,\ldots,\ \alpha >> 1.$

(α) Δείξτε ότι η ακολουθία y_n είναι γνησίως φθίνουσα και τείνει στο μηδέν.
(β) Υπολογίστε τον τύπο της ακολουθίας y_n
(γ) Προσδιορίστε αναδρομικό τύπο για τον προσδιορισμό της ακολουθίας συναρτήσει της y_{n-1} και δείξτε ότι ο αλγόριθμος που προκύπτει είναι ασταθής.
(δ) Δώστε ένα ευσταθή αλγόριθμο για τον υπολογισμό της y_n.

Υπόδειξη: (όπως παράδειγμα 1.1 σελ. 18 βιβλίου).
(α) Χρησιμοποιήστε το Θεώρημα Μέσης Τιμής του Ολοκληρωτικού Λογισμού, τότε:

$$y_n = \int_0^1 \frac{x^n}{x + a} dx = \xi^n \int_0^1 \frac{1}{x + a} dx,\ \xi \in (0,1).$$

(β) Κάντε τη διαίρεση x^n διά $(x+\alpha)$ και υπολογίστε το ολοκλήρωμα.

$$x^n = (x + a)\left(x^{n-1} - a\, x^{n-2} + \ldots + (-1)^{n-1} a^{n-1}\right) + (-1)^n a^n.$$

(γ) Παρατηρήστε ότι

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

$$= \int_0^1 x^{n-1} dx - a\int_0^1 \frac{x^{n-1}}{x+a}\ dx = \frac{1}{n} - a\ y_{n-1}.$$

Εχουμε λοιπόν:

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

$$\varepsilon_n = -a\ \varepsilon_{n-1} = a^2 \varepsilon_{n-2} = \ldots \left(-1\right)^n a^n\ \varepsilon_0,$$

και εφόσον α >>1 το σφάλμα αυξάνεται εκθετικά, άρα ο αλγόριθμος είναι ασταθής.

(δ) $y_n = \frac{1}{n} - a\ y_{n-1} \Rightarrow y_{n-1} = \frac{n^{-1} - y_n}{a}$ (ευσταθής αλγόριθμος, γιατί;).