# 1 Αριθμητική κινητής υποδιαστολής και σφάλματα στρογγύλευσης

Στη συγκεκριμένη ενότητα εξετάζουμε θέματα σχετικά με την αριθμητική πεπερασμένης ακρίβειας που χρησιμοποιούν οι σημερινοί υπολογιστές και τα σφάλματα στρογγύλευσης στους υπολογισμούς τα οποία οφείλονται ακριβώς στην αριθμητική αυτού του τύπου.

## 1.1 Αναπαράσταση αριθμών ως προς οποιαδήποτε βάση

Στην καθημερινή μας ζωή χρησιμοποιούμε αριθμούς στο δεκαδικό σύστημα. Για παράδειγμα, γράφουμε 251 για να δηλώσουμε τον αριθμό

$$2 \cdot 10^2 + 5 \cdot 10^1 + 1 \cdot 10^0$$

Ο αριθμός 10 είναι η λεγόμενη **βάση** του αριθμητικού συστήματος. Κάθε ακέραιος μπορεί να εκφραστεί ως ένα πολυώνυμο με ακέραιους συντελεστές τα **ψηφία** του αριθμού που είναι, φυσικά, οι αριθμοί $0, 1, \ldots, 9$. Στο εξής θα χρησιμοποιούμε το συμβολισμό

$$N = (a_n a_{n-1} \cdots a_0)_{10} = a_n \cdot 10^n + a_{n-1} \cdot 10^{n-1} + \cdots + a_0 \cdot 10^0 \tag{1.1}$$

για να δηλώσουμε οποιονδήποτε θετικό ακέραιο στο αριθμητικό σύστημα με βάση το 10. Δεν υπάρχει κάποιος ιδιαίτερος λόγος να χρησιμοποιούμε τη βάση 10 (εκτος φυσικά από το γεγονός ότι έχουμε 10 δάκτυλα). Κάποιοι πολιτισμοί χρησιμοποίησαν τις βάσεις 12, 20 ή 60 (ποιοί;). Οι σημερινοί υπολογιστές χρησιμοποιούν τη βάση 2, το δυαδικό σύστημα δηλαδή, με ψηφία το 0 και το 1. Κάθε μη αρνητικός ακέραιος θα παριστάνεται στο δυαδικό σύστημα ως

$$N = (a_n a_{n-1} \cdots a_0)_2 = a_n \cdot 2^n + a_{n-1} \cdot 2^{n-1} + \cdots + a_0 \cdot 2^0 \tag{1.2}$$

όπου οι συντελεστές $a_k$ είναι είτε 0 είτε 1. Παρατηρήστε ότι και πάλι ο αριθμός $N$ παριστάνεται ως ένα πολυώνυμο οι συντελεστές του οποίου είναι τα ψηφία 0, 1 του δυαδικού συστήματος.

Η μετατροπή ενός δυαδικού αριθμού στο δεκαδικό σύστημα γίνεται με χρήση του ορισμού (1.2). Για παράδειγμα,

$$(1101)_2 = 1 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 = 13$$

Η μετατροπή ενός κλασματικού αριθμού γραμμένου σε σύστημα με βάση $\beta$, στο δεκαδικό σύστημα γίνεται τελείως ανάλογα. Για παράδειγμα,

$$(0.1101)_2 = 1 \cdot 2^{-1} + 1 \cdot 2^{-2} + 0 \cdot 2^{-3} + 1 \cdot 2^{-4} = \frac{1}{2} + \frac{1}{4} + \frac{1}{16} = \frac{13}{16}$$

Η μετατροπή ενός αριθμού από το δεκαδικό σε ένα σύστημα με κάποια άλλη βάση $2 \leq \beta$ γίνεται με τον αλγόριθμο της διαίρεσης. Για να μετατρέψουμε, π.χ. τον αριθμό $(152)_{10}$ στο δυαδικό σύστημα , πρέπει να προσδιορίσουμε ψηφία $d_0, d_1, \ldots$ τέτοια ώστε

$$(152)_{10} = (\ldots d_2 d_1 d_0)_2 = d_0 + 2(d_1 + 2(d_2 + \cdots) \cdots).$$

Από την παραπάνω σχέση αμέσως βλέπουμε ότι το $d_0$ είναι το υπόλοιπο της διαίρεσης $152 : 2$, δηλαδή $d_0 = 0$. Το πηλίκο της διαίρεσης, ο αριθμός $(76)_{10}$, ισούται με $d_1 + 2(d_2 + \cdots)$. Επομένως το $d_1$ είναι το υπόλοιπο της διαίρεσης $76 : 2$, δηλαδή $d_1 = 0$. Συνεχίζοντας με τον ίδιο τρόπο έχουμε τελικά $(152)_{10} = (1001\,1000)_2$.

Η μετατροπή ενός κλασματικού αριθμού $x$ γραμμένου στο δεκαδικό σύστημα σ' ένα σύστημα με βάση $\beta$ γίνεται ως εξής: αν $x = (0.a_{-1}a_{-2}\cdots)_\beta$ τότε πολλαπλασιάζοντας και τα δύο μέλη με $\beta$ έχουμε $\beta x = (a_{-1}.a_{-2}\cdots)_\beta$. Άρα $a_{-1}$ είναι το ακέραιο μέρος του $\beta x$

ενώ το κλασματικό του μέρος είναι το $(0.a_{-2}\cdots)$. Πολλαπλασιάζοντας πάλι με $\beta$ μπορούμε να προσδιορίσουμε το $a_{-2}$ κλπ. Για παράδειγμα, η μετατροπή του αριθμού $x = (0.314)_{10}$ στο δυαδικό σύστημα ακολουθεί τα παρακάτω βήματα (οι αριθμοί $\gamma_i$ που εμφανίζονται παρακάτω δηλώνουν το κλασματικό μέρος του $2x$):

$$\begin{array}{rlllll} 2x = 0.628 & \text{άρα} & a_{-1} = 0 & \text{και} & \gamma_1 = 0.628 \\ 2\gamma_1 = 1.256 & \text{άρα} & a_{-2} = 1 & \text{και} & \gamma_2 = 0.256 \\ 2\gamma_2 = 0.512 & \text{άρα} & a_{-3} = 0 & \text{και} & \gamma_3 = 0.512 \\ 2\gamma_3 = 1.024 & \text{άρα} & a_{-4} = 1 & \text{και} & \gamma_4 = 0.024 \\ \vdots & & & & \end{array}$$

Συνεπώς $(0.314)_{10} = (0.0101\ldots)_2$.

## 1.2 Αριθμοί μηχανής και αριθμητική κινητής υποδιαστολής

Κάθε μη μηδενικός πραγματικός αριθμός $x$, γραμμένος σε ένα αριθμητικό σύστημα με οποιαδήποτε βάση $\beta$, είναι δυνατόν να γραφεί στη λεγόμενη (κανονική) μορφή *κινητής υποδιαστολής*

$$x = \pm(0.d_1d_2\ldots)\cdot\beta^e, \qquad d_1 \neq 0,$$

όπου $d_i$ είναι ψηφία ως προς τη βάση $\beta$ και $e$ κατάλληλος ακέραιος. Εν γένει, χρειάζονται άπειρα ψηφία για την αναπαράσταση ενός πραγματικού αριθμού τα οποία δεν μπορούν να αποθηκευτούν στην πεπερασμένη μνήμη ενός υπολογιστή. Ο υπολογιστής λοιπόν *προσεγγίζει* έναν δεδομένο αριθμό με πεπερασμένη ακρίβεια και συγκεριμένα με κάποιον από τους λεγόμενους *αριθμούς μηχανής*. Οι αριθμοί αυτοί αποτελούν ένα "μικρό" υποσύνολο των ρητών αριθμών με κλάσμα με πεπερασμένο πλήθος ψηφίων και με άνω και κάτω φράγμα του εκθέτη $e$ της βάσης $\beta$. Ένα τέτοιο σύνολο χαρακτηρίζεται από τις παραμέτρους

- τη βάση $\beta$ του αριθμητικού συστήματος
- την *ακρίβεια* $t$, δηλαδή των πλήθος των ψηφίων του κλάσματος
- το κάτω φράγμα $L$ και το άνω φράγμα $U$ του εκθέτη $e$

Κάθε (κανονικός) αριθμός μηχανής είναι της μορφής

$$\pm 0.d_1d_2\ldots d_t\cdot\beta^e, \tag{1.3}$$

όπου το *κλάσμα* $0.d_1d_2\ldots d_t$ αποτελείται από $t$ ψηφία στη βάση $\beta$ με $d_1 \neq 0$ και ο εκθέτης $e$ είναι ακέραιος με $L \leq e \leq U$. Το *σύνολο των αριθμών μηχανής* $M = M(\beta, t, L, U)$ είναι οι αριθμοί της μορφής (1.3) και το μηδέν. Στους σημερινούς υπολογιστές, σχεδόν πάντα $\beta = 2$. Για παράδειγμα, οι θετικοί αριθμοί μηχανής σ' ένα σύστημα με $\beta = 2$, $t = 3$, και $-L = U = 1$ είναι οι

$$\begin{array}{cccc} (0.100)_2\cdot 2^{-1}, & (0.101)_2\cdot 2^{-1}, & (0.110)_2\cdot 2^{-1}, & (0.111)_2\cdot 2^{-1}, \\ (0.100)_2\cdot 2^{0}, & (0.101)_2\cdot 2^{0}, & (0.110)_2\cdot 2^{0}, & (0.111)_2\cdot 2^{0}, \\ (0.100)_2\cdot 2^{1}, & (0.101)_2\cdot 2^{1}, & (0.110)_2\cdot 2^{1}, & (0.111)_2\cdot 2^{1}, \end{array}$$

με άλλα λόγια οι $1/4, 5/16, 3/8, 7/16, 1/2, 5/8, 3/4, 7/8, 1, 5/4, 3/2, 7/4$ (επιβεβαιώστε το). Για το θέμα της προσέγγισης ενός πραγματικού αριθμού από ένα αριθμό μηχανής ισχύουν τα ακόλουθα:

- αν προσπαθήσουμε να παραστήσουμε έναν αριθμό απολύτως μεγαλύτερο του μέγιστου στοιχείου του $M$ ή όταν ένας τέτοιος αριθμός προκύψει κατά τη διάρκεια των πράξεων, τότε ο υπολογιστής δίνει το μήνυμα της "υπερχείλισης" (overflow) και σταματάει τους υπολογισμούς

- ανάλογα, θα έχουμε "υπεκχείλιση" (underflow) αν προσπαθήσουμε να παραστήσουμε έναν αριθμό απολύτως μικρότερο του ελαχίστου στοιχείου του $M$. Στους περισσότερους υπολογιστές ο συγκεκριμένος αριθμός αντικαθίσταται από το μηδέν και οι πράξεις συνεχίζονται.

- κάθε πραγματικός αριθμός $x$ στο εύρος των αριθμών μηχανής προσεγγίζεται για να παρασταθεί και να χρησιμοποιηθεί σε πράξεις από έναν "κοντινό" αριθμό μηχανής που θα τον συμβολίζουμε με $fl(x)$ και ο οποίος έχει την ιδιότητα $|x - fl(x)| \leq |x - y|$ για κάθε $y \in M$. Σε αυτή την περίπτωση είναι εύκολο να δεί κανείς ότι ισχύει η εκτίμηση
$$\left|\frac{x - fl(x)}{x}\right| \leq \frac{1}{2}\beta^{1-t} \tag{1.4}$$

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

- η επιλογή του $fl(x)$ μπορεί να γίνει και με την λεγόμενη *αποκοπή* κατά την οποία αποκόπτουμε όλα τα ψηφία μετά το $d_t$. Σε αυτή την περίπτωση μπορούμε να δούμε εύκολα ότι ισχύει πάλι η (1.4) αλλά χωρίς τον παράγοντα 1/2 στο δεξί μέλος.

Στη συνέχεια θα υποθέτουμε ότι σν $x \neq 0$ είναι πραγματικός αριθμός στο εύρος των αριθμών μηχανής τότε θα παριστάνεται από τον αριθμό μηχανής $fl(x)$ με σχετικό σφάλμα
$$\left|\frac{x - fl(x)}{x}\right| \leq u, \tag{1.5}$$
όπου $u$ είναι το λεγόμενο *μοναδιαίο σφάλμα στρογγύλευσης* το οποίο, σύμφωνα με τα παραπάνω, είναι
$$u = \begin{cases} \frac{1}{2}\beta^{1-t} & \text{για στρογγύλευση,} \\ \beta^{1-t} & \text{για αποκοπή.} \end{cases} \tag{1.6}$$
Καταλαβαίνουμε λοιπόν ότι σφάλματα στρογγύλευσης στην παράσταση των αριθμών οδηγούν σε σφάλματα στις αριθμητικές πράξεις, οι οποίες δεν γίνονται πλέον ακριβώς. Για να μελετήσουμε την επιρροή αυτών των σφαλμάτων στις αριθμητικές πράξεις θα δεχθούμε τον εξής κανόνα (είναι, πραγματικά, ένα ρεαλιστικό μοντέλο του πραγματικού μηχανισμού των πράξεων):

αν $\star$ είναι μια από τις τέσσερεις πράξεις τις αριθμητικής και $x, y$ είναι πραγματικοί αριθμοί που μπορούν να παρασταθούν κατά προσέγγιση από στοιχεία του $M$, τότε το αποτέλεσμα της πράξης $x \star y$ στον υπολογιστή είναι ο αριθμός μηχανής
$$fl(fl(x) \star fl(y)). \tag{1.7}$$

Η παραπάνω σχέση ερμηνεύεται ως εξής: πρώτα γίνεται η παράσταση των αριθμών $x, y$ από τους αριθμούς μηχανής $fl(x), fl(y)$, αντίστοιχα, στη συνέχεια η πράξη $fl(x) \star fl(y)$ γίνεται *ακριβώς* (στην πραγματικότητα με ακρίβεια $2t$ δεκαδικών ψηφίων) και τέλος, το αποτέλεσμα της πράξης αυτής παριστάνεται στο $M$ προσεγγίζοντάς το με την συνάρτηση $fl(\cdot)$.

Ως παράδειγμα, ας θεωρήσουμε ένα υπολογιστικό σύστημα με $\beta = 10$, $t = 5$, $-L = U = 10$ και στο οποίο το $fl(\cdot)$ προκύπτει με στρογγύλευση. Αν $x = 0.589126 \cdot 10^4$ και $y = 0.773414 \cdot 10^{-1}$ τότε ο υπολογισμός του αθροίσματος των $x$ και $y$ με βάση την παραδοχή (1.7) δίνει $fl(x) = 0.58913 \cdot 10^4$, $fl(y) = 0.77341 \cdot 10^{-1}$ και $fl(x) + fl(y) = 0.5891377341 \cdot 10^4$. Άρα $fl(fl(x) + fl(y)) = 0.58914 \cdot 10^4$. Παρατηρήστε ότι $x + y = 0.58913373414 \cdot 10^4$.

## 1.3 Επιρροή των σφαλμάτων στρογγύλευσης στους υπολογισμούς

Θα μελετήσουμε την επιρροή των σφαλμάτων στρογγύλευσης κατά την πράξη $\star$ εκτιμώντας την απόλυτη τιμή του σχετικού σφάλματος

$$\frac{fl(fl(x) \star fl(y)) - x \star y}{x \star y}$$

χρησιμοποιώντας τις (1.5) και (1.6). Είναι εύκολο να διαπιστώσει κανείς ότι η (1.5) είναι ισοδύναμη με τη σχέση

$$fl(x) = x(1+\epsilon) \qquad \text{για κάποιο } \epsilon = \epsilon(x), \quad |\epsilon| \le u. \tag{1.8}$$

Επίσης, από το θεώρημα ενδιάμεσης τιμής προκύπτει εύκολα ότι αν $\epsilon_i$, $1 \le i \le m$, ικανοποιούν $|\epsilon_i| \le u < 1$, τότε υπάρχει $\epsilon$, $|\epsilon| \le u$, τέτοιο ώστε

$$\prod_{i=1}^{m}(1+\epsilon_i) = (1+\epsilon)^m.$$

Ας δούμε κατ' αρχήν την εκτίμηση του σχετικού σφάλματος του πολλαπλασιασμού (η εκτίμηση του σχετικού σφάλματος της διαίρεσης είναι εντελώς ανάλογη). Βάσει της (1.7) έχουμε $fl(x) = x(1+\epsilon_1)$, $fl(y) = y(1+\epsilon_2)$ με $|\epsilon_i| \le u$. Άρα,

$$z = fl(fl(x) * fl(y)) = xy(1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3) = xy(1+\epsilon)^3$$

για κάποιο $|\epsilon| \le u$. Έτσι,

$$\left|\frac{z-xy}{xy}\right| = \left|1-(1+\epsilon)^3\right| = |3\epsilon + 3\epsilon^2 + \epsilon^3| \le 3u + 4u^2.$$

Επειδή $u^2 \ll u$ λέμε ότι στον πολλαπλασιασμό το σχετικό σφάλμα είναι περίπου τριπλάσιο του μοναδιαίου σφάλματος στρογγύλευσης.

Για την πρόσθεση ή την αφαίρεση έχουμε

$$\begin{aligned} z &= fl(fl(x)+fl(y)) = x(1+\epsilon_1)(1+\epsilon_3) + y(1+\epsilon_2)(1+\epsilon_3) = x(1+\epsilon)^2 + y(1+\delta)^2 \\ &= (x+y) + 2(\epsilon x + \delta y) + \epsilon^2 x + \delta^2 y \end{aligned}$$

όπου $|\epsilon|, |\delta| \le u$. Άρα, με αρκετά καλή προσέγγιση

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

Παρατηρούμε λοιπόν ότι αν $x, y$ είναι ομόσημοι, $|x|+|y| = |x+y|$ και παίρνουμε μικρό σχετικό σφάλμα με φράγμα $2u$. Αν όμως οι $x, y$ είναι ετερόσημοι και $x \approx -y$, το φράγμα του σχετικού σφάλματος είναι πολύ μεγαλύτερο του $2u$. Για παράδειγμα, ας θεωρήσουμε ένα υπολογιστικό σύστημα με $\beta = 10$, $t = 5$, $-L = U = 10$ και στο οποίο το $fl(\cdot)$ προκύπτει με στρογγύλευση. Αν $x = 0.45142708$ και $y = -0.45115944$ τότε $z = fl(0.45143 - 0.45116) = 0.27000 \cdot 10^{-3}$ και

$$\left|\frac{z-(x+y)}{x+y}\right| \approx 88 \cdot 10^{-4},$$

δηλαδή 88 φορές μεγαλύτερο από σχετικό σφάλμα της πρόσθεσης δύο ομόσημων αριθμών! Ο λόγος της μεγάλης αύξησης του σχετικού σφάλματος είναι ότι στο εξαγόμενο $z = 0.27000 \cdot 10^{-3}$ μόνο δύο ψηφία του κλάσματος είναι σωστά και την ίδια στιγμή τα τρία μηδενικά που ακολουθούν δεν οφείλονται σε αφαίρεση ίσων ψηφίων στις θέσεις 3, 4 και 5 του κλάσματος αλλά στην αφαίρεση των πρώτων τριών ψηφίων .451 και την μετατροπή σε αριθμό μηχανής

με μετακίνση της υποδιαστολής και μεταβολή του εκθέτη. Η παρουσία τέτοιων μηδενικών είναι τυπική ένδειξη απώλειας ακρίβειας, δηλαδή απώλειας σημαντικών ψηφίων.

Σε ορισμένες περιπτώσεις είναι δυνατόν να αποφύγουμε την αφαίρεση σχεδόν ίσων αριθμών αν χρησιμοποιήσουμε ένα διαφορετικό αλγόριθμο. Για παράδειγμα είναι προτιμότερο να χρησιμοποιήσουμε την σχέση

$$\frac{1}{\sqrt{x+1}+\sqrt{x}}$$

αντί της $\sqrt{x+1}-\sqrt{x}$, ιδίως όταν $x \gg 1$. Ως δεύτερο παράδειγμα, ας θεωρήσουμε το πρόβλημα του υπολογισμού των ριζών της εξίσωσης $ax^2+bx+c=0$. Γνωρίζουμε ότι οι ρίζες δίνονται από τον τύπο

$$x=\frac{-b\pm\sqrt{b^2-4ac}}{2a}.$$

Αν υποθέσουμε ότι $b>0$, $b^2-4ac>0$ και ενδιαφερόμαστε να υπολογίσουμε την ρίζα με την μικρότερη απόλυτη τιμή, δηλαδή την

$$x=\frac{-b+\sqrt{b^2-4ac}}{2a}. \tag{1.9}$$

Αν $4ac \ll b^2$ τότε ο υπολογισμός του αριθμητή της παραπάνω έκφρασης θα υποφέρει από απώλεια σημαντιικών ψηφίων. Για παράδειγμα, ο υπολογισμός της μικρότερης κατ' απόλυτη τιμή ρίζας της εξίσωσης $x^2+111.11x+1.2121=0$ με ακρίβεια 5 ψηφίων από τον τύπο (1.9) δίνει ως αποτέλεσμα το $-0.01000$ ενώ στην πραγματικότητα $x_1=-0.010910$. Αν όμως υπολογίσουμε την ρίζα από τη σχέση

$$x_1=\frac{-2c}{b+\sqrt{b^2-4ac}} \tag{1.10}$$

τότε παίρνουμε τη σωστή ρίζα με ακρίβεια 5 δεκαδικών ψηφίων.

## 1.4 Ευστάθεια αλγορίθμων

Ένας *αλγόριθμος*, δηλαδή μια σειρά εντολών που υλοποιούν μια αριθμητική μέθοδο, ονομάζεται *ευσταθής* αν τα τελικά αποτελέσματά του δεν επηρεάζονται πολύ από μικρές διαταραχές που τα σφάλματα στρογγύλευσης προκαλούν σε κάθε βήμα του. Διαφορετικά ονομάζεται *ασταθής*. Το αριθμητικό πρόβλημα το οποίο παρουσιάζουμε παρακάτω είναι ένα κλασσικό παράδειγμα όπου ο προφανής αλγόριθμος της λύσης του είναι ασταθής. Οφείλεται στους Forsythe, Malcolm και Moler (1977). Ενδιαφερόμαστε να υπολογίσουμε τους όρους της ακολουθίας

$$I_n=\int_0^1 x^n e^{x-1}\,dx, \qquad n=1,2,\ldots, \tag{1.11}$$

για αρκετά μεγάλο $n$. Παρατηρούμε κατ' αρχήν ότι

$$0<I_n\le\int_0^1 x^n\,dx=\frac{1}{n+1}, \qquad n=1,2,\ldots$$

και επειδή $x^{n+1}<x^n$ στο $(0,1)$ ισχύει ότι $I_{n+1}<I_n$. Επομένως η ακολουθία $(I_n)_{n\in\mathbb{N}}$ είναι γνησίως φθίνουσα και τείνει στο μηδέν καθώς $n\to\infty$. Χρησιμοποιώντας ολοκλήρωση κατά μέρη στην (1.11) έχουμε

$$\int_0^1 x^n e^{x-1}\,dx=x^n e^{x-1}\big|_{x=0}^1-n\int_0^1 x^{n-1}e^{x-1}\,dx,$$

δηλαδή,

$$I_1 = \frac{1}{e}, \qquad I_n = 1 - nI_{n-1}, \quad n = 2, 3, \ldots. \tag{1.12}$$

Η σχέση (1.12) αποτελεί έναν απλό αλγόριθμο για τον υπολογισμό των $I_n$ που, επιπλέον, απαιτεί πολύ λιγες πράξεις. Υπολογίζοντας με αριθμητική κινητής υποδιαστολής με $\beta = 10$ και $t = 6$ έχουμε $\tilde{I}_1 = 0.367879$, $\tilde{I}_2 = 0.264242, \ldots, \tilde{I}_9 = -0.068489$, με την τελευταία τιμή να είναι προφανώς λάθος. Στον παραπάνω υπολογισμό, το μόνο σφάλμα στρογγύλευσης είναι η προσέγγιση του $1/e$ από την ποσότητα $\tilde{I}_1$ με σφάλμα $\epsilon_1 \approx -4.4 \cdot 10^{-4}$. Συνεπώς οι τιμές $\tilde{I}_n$ δίνονται από τις σχέσεις

$$\tilde{I}_1 = I_1 + \epsilon_1, \qquad \tilde{I}_n = 1 - n\tilde{I}_{n-1}, \quad n = 2, 3, \ldots. \tag{1.13}$$

Για τη διαφορά $\epsilon_n = \tilde{I}_n - I_n$, έχουμε, αφαιρώντας κατά μέλη τις (1.12) και (1.13), $\epsilon_n = -n\epsilon_{n-1}, \quad n \geq 2$, δηλαδή $\epsilon_n = (-1)^{n-1} n! \epsilon_1$. Αυτό δείχνει ότι ο αλγόριθμος είναι ασταθής γιατί το αρχικό σφάλμα πολλαπλασιάζεται με τον παράγοντα $n!$ που αυξάνει πολύ γρήγορα.

Έχοντας διαπιστώσει ότι υπεύθυνος για την αστάθεια του αλγορίθμου είναι ο πολλαπλασιαμός $nI_{n-1}$ , γράφουμε την αναδρομική σχέση στην (1.12) στη μορφή $I_{n-1} = (1 - I_n)/n$ και "αντιστρέφουμε" τη σειρά των υπολογισμών. Αν γνωρίζουμε κάποια τιμή $I_m$ και ενδιαφερόμαστε για τον υπολογισμό του $I_k$ για $k < m$, υπολογίζουμε τους όρους $I_{m-1}, I_{m-2}, \ldots, I_k$ βάσει της αναδρομικής σχέσης που μόλις δώσαμε. Η ανάλυση της ευστάθειας του συγκεκριμένου αλγορίθμου γίνεται ως εξής: έστω $\tilde{I}_m$ μια προσέγγιση της ακριβούς τιμής $I_m$ με σφάλμα $\epsilon_m$ και υπολογίζουμε τους όρους $\tilde{I}_{m-1}, \tilde{I}_{m-2}, \ldots, I_k$ από την αναδρομική σχέση

$$\tilde{I}_m = I_m + \epsilon_m, \qquad \tilde{I}_{n-1} = (1 - n\tilde{I}_n, \quad n = m, \ldots, k+1. \tag{1.14}$$

Θέτοντας $\epsilon_n = \tilde{I}_n - I_n$, βρίσκουμε

$$\epsilon_{n-1} = -\frac{1}{n}\epsilon_n, \qquad m \geq n \geq k+1,$$

δηλαδή

$$\epsilon_k = (-1)^{m-k} \frac{1}{k+1} \cdot \frac{1}{k+2} \cdot \cdots \cdot \frac{1}{m} \epsilon_m \,.$$

Από την τελευταία σχέση αντιλαμβανόμαστε ότι ο αλγόριθμος που προτείνεται είναι ευσταθής μια και το αρχικό σφάλμα $\epsilon_m$ όχι μόνο δεν αυξάνει αλλά καταστέλλεται. Αν π.χ., πάρουμε $\tilde{I}_m = 0$ υπολογίζουμε (με $beta = 10, t = 6$) $\tilde{I}_9 = 0.916123 \cdot 1{-}^{-1}$ με πολύ καλή ακρίβεια.

## Ασκήσεις

**Άσκηση 1.1** *Μετατρέψτε τους αριθμούς* $(1010)_2, (100101)_2$, $(10000001)_2$ *και* $(0.1100011)_2$ *στο δεκαδικό σύστημα.*

**Άσκηση 1.2** *Μετατρέψτε τους αριθμούς* $(82)_{10}, (109)_{10}, (3433)_{10}$ *και* $(0.614)_{10}$ *στο δυαδικό σύστημα.*

**Άσκηση 1.3** *Εκτελέστε τις πράξεις* $a \star b$ *όπου* $a = 0.4523 \cdot 10^4$ *και* $b = 0.2583 \cdot 10^1$ *σε ένα υπολογιστικό σύστημα με* $\beta = 10$, $t = 5$ *και* $fl(\cdot)$ *να προκύπτει με στρογγυλοποίηση. Εδώ* $\star$ *δηλώνει κάθε μια από τις τέσσερεις πράξεις της αριθμητικής. Σε κάθε περίπτωση υπολογίστε το σχετικό σφάλμα.*

**Άσκηση 1.4** *Εξηγείστε πως προκύπτει η σχέση* (1.10) *και βρείτε την μικρότερη, κατ' απόλυτη τιμή, ρίζα της εξίσωσης* $x^2+0.4002x+0.8\cdot 10^{-4}=0$ *σε ένα υπολογιστικό σύστημα με* $\beta=10$, $t=4$ *και* $fl(\cdot)$ *να προκύπτει με στρογγυλοποίηση. Συγκρίνετε με την ακριβή τιμή και της ρίζας και προτείνετε ένα τρόπο για τον υπολογισμό της δεύτερης ρίζας.*

**Άσκηση 1.5** *Προτείνετε τρόπους για τον ακριβή υπολογισμό των παρακάτω ποσοτήτων:* $(x-\sin x)/\tan x, (\alpha+x)^n-\alpha^n$ *και* $\sin(\alpha+x)-\sin\alpha$ *δεδομένου του ότι* $|x|\ll 1$.

**Άσκηση 1.6** *Θέλουμε να υπολογίσουμε, για δεδομένη σταθερά* $\alpha\gg 1$, *τους όρους της ακολουθίας*

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

*Αποδείξτε ότι η ακολουθία* $\{y_n\}$ *είναι γνησίως φθίνουσα και ότι* $\lim_{n\to\infty} y_n=0$. *Στη συνέχεια, προσδιορίστε αναδρομικό τύπο για τον υπολογισμό του* $y_n$ *συναρτήσει του* $y_{n-1}$ *και προσδιορίστε αναλυτικά το* $y_0$. *Είναι ο αλγόριθμος που προκύπτει ευσταθής; Προτείνετε έναν ευσταθή αλγόριθμο για τον υπολογισμό του* $y_n$.