Αριθμητική Ανάλυση

Αναστάσιος Τέφας
tefas@aiia.csd.auth.gr
2310-991932

Αριθμοί κινητής υποδιαστολής

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

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

όπου e είναι θετικός ακέραιος που καλείται ***εκθέτης*** και $\pm\,(0.\,b_1 b_2 \ldots)$ 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 $b_1, b_2, \ldots$ είναι όλα ψηφία του p-αδικού συστήματος αρίθμησης.

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

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

όπου

(i) $\tilde{x}_n = \pm\,(0.\, b_1 b_2 \ldots b_n)$ και το n δηλώνει την **ακρίβεια**, δηλαδή το πλήθος των ψηφίων του κλασματικού μέρους του αριθμού,

(ii) ο εκθέτης e παίρνει τις τιμές $e = -c, -c+1, \ldots, c-1, c$ για κάποιο θετικό ακέραιο c,

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

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

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

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. □

Ιδιότητες αριθμών μηχανής

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 $\max_{A_M} = \left(1 - \frac{1}{p^n}\right) p^c$.

Υποχείλιση - Υπερχείλιση

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)***.

Κατανομή αριθμών μηχανής

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

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

Σε κάθε ένα από τα υποδιαστήματα

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

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.

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 $A_M(p,n,c)$.

Σφάλματα

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.

Σφάλματα

Αν λοιπόν υποθέσουμε ότι $x = \pm (0.\, b_1 b_2 \ldots)\, p^e$, $b_1 \neq 0$ 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:

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

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)***.

Σφάλματα

Έστω:

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

(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.

Σφάλματα

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

Σφάλματα

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

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

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

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 $|\rho_y|,\ |\rho_x| << 1$, έχουμε $\rho_{x\,y} \simeq \rho_y + \rho_x$, ή $|\rho_{x\,y}| \leq |\rho_y| + |\rho_x|$. 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***.

Σφάλματα

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^{o} κλπ, άρα:

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

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

έχουμε:

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

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

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.

Ασκήσεις

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

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

Ασκήσεις

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.

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.*

5. *Εστω x, y αριθμοί μηχανής με* $x \approx y$.

(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;*