Σημειώσεις στο μάθημα 'Αριθμητική Ανάλυση Ι'

Διδάσκων: Νοτάρης Σ.

Εθνικό Καποδιστριακό Πανεπιστήμιο Αθηνών

Αριθμητική κινητής υποδιαστολής,
σφάλματα στρογγύλευσης.

Παράσταση αριθμών ως προς οποιαδήποτε βάση.

- Δεκαδικό σύστημα.
  Βάση: 10,
  Ψηφία: 0, 1, 2, 3, $\cdots$, 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:

$$(\alpha_N \alpha_{N-1} \cdots \alpha_0 \alpha_{-1} \alpha_{-2} \cdots)_{10} = \alpha_N \cdot 10^N + \alpha_{N-1} \cdot 10^{N-1} + \cdots + \alpha_0 \cdot 10^0 + \alpha_{-1} + \cdots .$$

  Ακέραιο μέρος: $p(x) = \alpha_N x^N + \alpha_{N-1} x^{N-1} + \cdots + \alpha_0$, $x = 10$,
  Κλασματικό μέρος: $\sum_{\kappa=1}^{\infty} \alpha_\kappa x^\kappa$, $x = \frac{1}{10}$.
  (π.χ. ο αριθμός 4.130 μπορεί να γραφεί σε 'περιοδική μορφή' $4.129 = 4.129999$;;)

- Σύστημα με βάση β

  Ψηφία: 0, 1, 2, $\cdots$, $\beta - 1$.

$$\begin{aligned} \alpha &= 0,999\cdots \\ 10\alpha &= 9,999\cdots \\ 10\alpha &= 9 + 0,999\cdots \ ;;; \\ 10\alpha &= 9 + \alpha \\ \alpha &= 1 \end{aligned}$$

  Παραδείγματα: Στους υπολογιστές χρησιμοποιούνται τα δυαδικό, το οκταδικό και το δεκαεξαδικό σύστημα με ψηφία: 0, 1, 2, $\cdots$, 9 A, B, C, D, E, F,
  π.χ.$(10011011)_2 = 1 \cdot 2^5 + 1 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^{-1} + 1 \cdot 2^{-2}$.

α) Μετατροπή ενός ακεραίου γραμμένου σε σύστημα με βάση β, στο δεκαδικό σύστημα
π.χ. $(53473)_8 = 5 \cdot 8^4 + 3 \cdot 8^3 + 4 \cdot 8^2 + 7 \cdot 8^1 + 3 \cdot 8^0 = (22331)_{10}$
$(53473)_8 = 3 + 8(7 + 8(4 + 8(3 + 8 \cdot 5))) = (22331)_{10}$.
Γενικά για τον υπολογισμό της τιμής του $p(x) = \alpha_N x^N + \alpha_{N-1} x^{N-1} + \cdots + \alpha_0$, με α_i δεδομένους συντελεστές για δεδομένο x προκύπτει από το σχήμα Horner $p(x) = \alpha_0 + x(\alpha_1 + x(\alpha_2 + \cdots + x(\alpha_{N-1} + x\alpha_N)))$.

β) Μετατροπή ενός κλασματικού αριθμού x, $0 < x < 1$ γραμμένου σε σύστημα με βάση β, στο δεκαδικό π.χ. $(.11)_2 = (.75)_{10}$.

γ) Μετατροπή ενός ακεραίου του δεκαδικού συστήματος σ'ενα άλλο σύστημα όπου χρησιμοποιούμε τον αλγόριθμο της διαίρεσης (και το σχήμα Horner).

π.χ. Για να μετατρέψουμε τον αριθμό $(369)_{10}$ στο οκταδικό έχουμε,

$$(369)_{10} = (\cdots \alpha_2\alpha_1\alpha_0)_{10} = \alpha_0 + 8(\alpha_1 + 8(\alpha_2 + \cdots)).$$

- Το α_0 είναι το υπόλοιπο της διαίρεσης $\frac{369}{8}$ δηλαδή 1. Το δε πηλίκο $(46)_{10} = \alpha_1 + 8(\alpha_2 + 8(\cdots))$.
- Το α_1 είναι το υπόλοιπο της διαίρεσης $\frac{46}{8}$ δηλαδή 6. Το δε πηλίκο $(5)_{10} = \alpha_2 + 8(\alpha_3 + 8(\cdots))$.
- Το α_2 είναι το υπόλοιπο της διαίρεσης $\frac{5}{8}$ δηλαδή 5 και $\alpha_3 = \alpha_4 = \cdots = 0$. Άρα $(369)_{10} = (561)_8$.

δ) Μετατροπή ενός κλασματικού αριθμού x, γραμμένου στο δεκαδικό σύστημα, σε ένα άλλο σύστημα με βάση β. Έστω $x = (\cdots \alpha_{-1}\alpha_{-2}\cdots)_\beta = \alpha_{-1}\beta^{-1} + \alpha_{-2}\beta^{-2} + \cdots$.
Πολλαπλασιάζουμε και τα δύο μέλη με β,

$$\beta x = \alpha_{-1}\beta^0 + \alpha_{-2}\beta^{-1} + \cdots$$

όπου το ακέραιο μέρος είναι κάθε φορά το πρώτο ψηφίο του αριθμού.
π.χ. για να μετατρέψουμε τον $(0.372)_{10}$ στο δυαδικό έχουμε $(0.372)_{10} = (\alpha_{-1}\alpha_{-2}\alpha_{-3}\cdots)_2$ όπου

$$\begin{aligned} 2x = 0,744 &\rightarrow \alpha_{-1} = 0 \\ 2 \cdot 0,744 = 1,488 &\rightarrow \alpha_{-2} = 1 \\ 2 \cdot 0,488 = 0,976 &\rightarrow \alpha_{-3} = 0 \\ 2 \cdot 0,976 = 1,952 &\rightarrow \alpha_{-4} = 1 \\ 2 \cdot 0,952 = 1,904 &\rightarrow \alpha_{-5} = 1 \end{aligned}$$

Άρα $(.372)_{10} = (.01011\cdots)_2$.

ε) Ένας ακέραιος σε ένα σύστημα με βάση β_1 παραμένει ακέραιος όταν μετατραπεί σε ένα σύστημα με βάση β_2. Το ίδιο ισχύει με έναν κλασματικό, αλλά το πλήθος των ψηφίων από πεπερασμένο μπορεί να γίνει άπειρο και αντίστροφα για παράδειγμα $\frac{1}{10} = \sum_{n=1}^{\infty}(\frac{1}{2^{4n}} + \frac{1}{2^{4n+1}})$, άρα $(.1)_{10} = (.0001100110011\cdots)_2$.

Αριθμοί μηχανής.

Κανονική μορφή κινητής υποδιαστολής ενός πραγματικού αριθμού x, $x \neq 0$: $x = \pm(d_1 d_2 \cdots)\beta^e$, $d_1 \neq 0$ με d_i πεπερασμένα ή άπειρα ψηφία ως προς βάση β και e κατάλληλος ακέραιος.

Αριθμός μηχανής με πεπερασμένα ψηφία $x = \pm.d_1 d_2 \cdots d_t \beta^e$, $d_1 \neq 0$, με d_i ψηφία ως προς τη βάση β, e ακέραιος με $L \leq e \leq U$ όπου L, U ακέραιοι για τους οποίους $L \cong -U$.
Το σύνολο των αριθμών μηχανής $M = M(\beta, t, L, U)$ της αριθμητικής μονάδας ενός υπολογιστή είναι οι παραπάνω αριθμοί και το 0. Για ένα συγκεκριμένο M με δεδομένα β, t, L, U ισχύουν οι ακόλουθες ιδιότητες:

- Το M είναι πεπερασμένο,
- Το M έχει ένα κατάπόλυτον τιμή μέγιστο στοιχείο, αυτό με $d_i = \beta - 1$, $1 \leq i \leq t$, ??,
- Το M έχει ένα κατάπόλυτον τιμή ελάχιστο στοιχείο, το $.10 \cdots 0\beta^L$,
- Το M δεν είναι σώμα ως προς την πρόσθεση και τον πολλαπλασιασμό. Για παράδειγμα το $1. \cdot \beta^i$ επί τον εαυτόν του δεν βρίσκεται στομ M. Αν $\beta = 10$, $t = 5$ οι αριθμοί $1 = .1 \cdot 10^1$ και $10^{-3} = .1 \cdot 10^{-2}$ ανήκουν στο M αλλά ο $1 + 10^{-5} = 1.00001$ δεν ανήκει στο M.

Είναι ιδιαίτερα σημαντικό το M να είναι όσο το δυνατόν πιο πυκνό.
Παραδείγματα.

Υπολογιστής	β	t	L	U	β^{1-t}
$IEEE$(απλή ακρίβεια)	2	24	-125	128	1.1910^{-7}
$IBM3090$(απλή ακρίβεια)	16	6	-64	63	9.5410^{-7}
$HP33$(χεριού)	10	10	98	100	1.0010^{-9}

Αν προσπαθήσουμε να παραστήσουμε έναν αριθμό μεγαλύτερο του μέγιστου στοιχείου του M παίρνουμε ένα μήνυμα overflow δηλαδή σταματάει το σύστημα. Αντίστοιχα, αν ο αριθμός είναι $0 < x < .1\beta^L$ παίρνουμε το μήνυμα underflow όπου συνήθως ο αριθμός x αντικαθίσταται με το 0. Η προσέγγιση ενός αριθμού x με απόλυτη τιμή μεταξύ $.1\beta^L$ και β^U συμβόλίζεται με $fl(x)$. Αυτό μπορεί να γίνει με δύο τρόπους,

1. Με στρογγύλευση.
   Αν $x = \pm(.d_1 d_2 \cdots d_t d_{t+1} \cdots)\beta^\kappa$, τότε
   $$fl(x) = \begin{cases} \pm(.d_1 d_2 \cdots d_t)\beta^\kappa, & \text{αν } 0 \leq d_{t+1} < \frac{\beta}{2}, \\ \pm(.d_1 d_2 \cdots d_t + \beta^{-t})\beta^\kappa, & \text{αν } \frac{\beta}{2} \leq d_{t+1} < \beta^1. \end{cases}$$
2. Με αποκοπή.
   Αν $x = \pm(.d_1 d_2 \cdots d_t d_{t+1} \cdots)\beta^\kappa$, τότε
   $$fl(x) = \pm(.d_1 d_2 \cdots d_t)\beta^\kappa.$$

Απόλυτο σφάλμα ορίζεται η ποσότητα $|fl(x) - x|$. Το σχετικό σφάλμα της προσέγγισης με στρογγύλευση είναι

$$|\frac{fl(x) - x}{x}| \leq \frac{1}{2}\beta^{1-t}.$$

Απόδειξη.
Αν $fl(x) = x$, τετριμμένο.
Έστω ότι x δεν είναι αριθμός μηχανής και έστω x', x'' οι διαδοχικοί αριθμοί μηχανής
με $x' < x < x''$. Τότε προφανώς,
(σχήμα)

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

οπότε

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

Έστω, χωρίς περιορισμό της γενικότητας, $x > 0$. Αν $x = -.d_1 d_2 \cdots d_t d_{t+1} \cdot \beta^\kappa$ τότε
$x' = .d_1 d_2 \cdots d_t \cdot \beta^\kappa$ και $x'' = (.d_1 d_2 \cdots d_t + \beta^{-t})\beta^\kappa$. Επομένως,

$$|x' - x''| = \beta^{\kappa - t}. \tag{1}$$

Επιπλέον,

$$\begin{aligned} \beta^{-1} \leq .d_1 \leq .d_1 d_2 \cdots d_t d_{t+1} \cdot < 1 \Rightarrow \\ |x| = .d_1 d_2 \cdots d_t d_{t+1} \cdot \beta^\kappa \geq \beta^{\kappa-1} \Rightarrow \\ |x| \geq \beta^{\kappa-1}. \end{aligned} \tag{2}$$

Από (4)–(5) προκύπτει,

$$|\frac{fl(x) - x}{x}| \leq \frac{1}{2}\frac{\beta^{\kappa-t}}{\beta^{\kappa-1}} = \frac{1}{2}\beta^{1-t}.$$

Το σχετικό σφάλμα της προσέγγισης με αποκοπή είναι,

$$|\frac{fl(x) - x}{x}| \leq \beta^{1-t}.$$

Τελικά έχουμε,

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

Αν x, ψ είναι πραγματικοί αριθμοί, μέσα στο εύρος των αριθμών μηχανής, και $\star$ είναι μία απο τις γνωστές πράξεις $+$, $-$, $\cdot$, $/$ τότε το αποτέλεσμα της πράξης $x \star \psi$ είναι ο αριθμός μηχανής

$$z = fl(fl(x) \star fl(\psi)).$$

Η πράξη $fl(x) \star fl(\psi)$ γίνεται με 'άπειρη' ακρίβεια, συνήθως με ακρίβεια $2t$ ψηφίων κλάσματος.
Παράδειγμα:
Σε υπολογιστή με $\beta = 10, t = 5, U = -L = 10$ και fl με στρογγύλευση, θεωρούμε τους $x = 5891,26$ και $0,773414$ και ζητάμε το άθροισμά τους z. Έχουμε $fl(x) = 58913 \cdot 10^4$ και $fl(\psi) = 77341 \cdot 10^{-1}$. Εξισώνοντας, τους εκθέτες παίρνουμε

$$fl(x) + fl(\psi) = 5891377341 \cdot 10^4$$

άρα $z = 58914 \cdot 10^4$. Συγκρίνετε με $x + \psi = 5891,3373414$, $fl(x + \psi) = 58913 \cdot 10^4$ και $fl(x) + fl(\psi)$. Συμπερασματικά ο ορισμός της πράξης δημιουργεί τα εξής παράδοξα,

- Οι συνήθεις ιδιότητες των πράξεων στο $\mathbb{R}$ όπως η προσεταιριστική δεν ισχύουν. Στον υπολογιστή του προηγούμενου παραδείγματος θεωρούμε τους αριθμούς μηχανής $\alpha = 1$, $\beta = \gamma = 3 \cdot 10^{-5}$ τότε $fl(fl(\alpha + \beta) + \gamma) = 1$ ενώ $fl(\alpha + fl(\beta + \gamma)) = 1,0001$.
- Αν κατά την λύση της εξίσωσης $1+x = 1$ στον υπολογιστή του προηγούμενου παραδείγματος το $x = 4 \cdot 10^{-5} (\in M)$, τότε $fl(1 + x) = 1$. Το ίδιο ισχύει για οποιοδήποτε $x \in \mathbb{R}$ με $0 < x < 5 \cdot 10^{-5}$. Αυτό γενικά ισχύει στην στρογγύλευση για οποιδήποτε $x \in \mathbb{R}$ με $0 < x < \frac{1}{2}\beta^{1-t}$, όπου η ποσότητα $\frac{1}{2}\beta^{1-t}$ ονομάζεται 'μηδέν' ή 'έψιλον' της μηχανής.

Επιρροή των σφαλμάτων στρογγύλευσης στους υπολογισμούς.
Αλγόριθμος για το 'μηδέν' ή 'έψιλον' της μηχανής

$$\varepsilon \leftarrow 1$$
$$\text{εφόσον } 1 + \varepsilon > 1$$
$$\varepsilon \leftarrow \tfrac{\varepsilon}{2}$$

Εκτίμηση του σχετικού σφάλματος:

$$\left| \frac{fl(fl(x) \star fl(\psi)) - (x \star \psi)}{x \star \psi} \right|.$$

Παρατηρήσεις,

1. Η εκτίμηση $|\frac{fl(x)-x}{x}| \leq u$ είναι ισοδύναμη με τη σχέση $fl(x) = x(1+\varepsilon)$ για κάποιο $\varepsilon \equiv \varepsilon(x)$, $|\varepsilon| \leq u$?. (Η απόδειξη αφήνεται ως άσκηση).

2. Αν τα ε_i, $1 \le i \le m$ ικανοποιούν τη σχέση $|\varepsilon_i| \le u \le 1$, τότε υπάρχει ε με $|\varepsilon| \le u \le 1$ τέτοιο ώστε
$$\prod_{i=1}^{m}(1+\varepsilon_i) = (1+\varepsilon)^m.$$
(Για την απόδειξη θεωρείστε δεδομένο ότι
$$(1-u)^m \le \prod_{i=1}^{m}(1+\varepsilon_i) \le (1+u)^m$$
και εφαρμόστε το Θεώρημα Ενδιάμεσης Τιμής.)

Έστω x, ψ, $x \star \psi$ μη μηδενικοί αριθμοί στο έυρος των αριθμών μηχανής.
Πολλαπλασιασμός. Έστω ότι $fl(x) = x(1+\varepsilon_1)$, $fl(\psi) = \psi(1+\varepsilon_2)$, με $|\varepsilon_i| \le u$ τότε
$$\begin{aligned} z &= fl(fl(x)\cdot fl(\psi)) \\ &= fl(x(1+\varepsilon_1)\cdot\psi(1+\varepsilon_2)) \\ &= x\psi(1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3) \text{ όπου } \varepsilon_3 \equiv \varepsilon_3(x,\psi),\ \|\varepsilon_3| \le u \\ &= x\psi(1+\varepsilon)^3,\ |\varepsilon| \le u. \end{aligned}$$
Έτσι έχουμε
$$\begin{aligned} |\frac{z - x\cdot\psi}{x\cdot\psi}| = |\frac{x\cdot\psi(1+\varepsilon)^3 - x\psi}{x\psi}| &= |((1+\varepsilon)^3 - 1| \\ &= |\varepsilon^3 + 3\varepsilon^2 + 3\varepsilon| \le |\varepsilon|^3 + 3|\varepsilon|^2 + 3|\varepsilon| \\ &\le u^3 + 3u^2 + 3u \cong 3u \end{aligned}$$
γιατί $u << 1 \Rightarrow u^2 << u$, $u^3 \le u^2 << u$. *Διαίρεση.* Παρόμοια με τον πολλαπλασιασμό.
Πρόσθεση–Αφαίρεση. Έχουμε
$$\begin{aligned} z &= fl(fl(x)+fl(\psi)) = fl(x(1+\varepsilon_1)+\psi(1+\varepsilon_2)) \\ &= x(1+\varepsilon_1)(1+\varepsilon_3) + \psi(1+\varepsilon_2)(1+\varepsilon_3),\ |\varepsilon_3| \le u \\ &= x(1+\varepsilon)^2 + \psi(1+\delta)^2,\quad |\delta|, |\varepsilon| \le u \\ &= x + \psi + 2x\varepsilon + 2\psi\delta + x\varepsilon^2 + \psi\delta^2 \\ &\cong x + \psi + 2(x\varepsilon + \psi\delta) \end{aligned}$$
Έτσι
$$\begin{aligned} \left|\frac{z-(x+\psi)}{x+\psi}\right| \cong 2\left|\frac{x\varepsilon+\psi\delta}{x+\psi}\right| &\le 2\frac{|x||\varepsilon| + |\psi||\delta|}{|x+\psi|} \\ &\le u\frac{|x|+|\psi|}{|x+\psi|}. \end{aligned}$$
Διερεύνηση:

- Αν x, ψ είναι ομόσημοι τότε $|x|+|\psi| = |x+\psi|$ και το σχετικό σφάλμα είναι περίπου $2u$.
- Αν x, ψ είναι ετερόσημοι και $x \cong -\psi$ τότε το φράγμα του σχετικού σφάλματος γίνεται *πολύ μεγάλο*–σφάλμα ακύρωσης (cancelation error).

Παράδειγμα. Στον γωστό υπολογιστή $\beta = 10$, $t = 5$, $U = -L = 10$ και fl με στρογγύλευση, θεωρούμε τους $x = .45142708$ και $\psi = -.45115944$ με ακριβές άθροισμα $x + \psi = .26764 \cdot 10^{-3}$. Έχουμε,

$$\begin{aligned} z &= fl(fl(x) + fl(\psi)) \\ &= fl(.45143 - .45116) \\ &= fl(.00027) = .27000 \cdot 10^{-3} \end{aligned}$$

Για το σφάλμα έχουμε

$$\left| \frac{z - (x+\psi)}{x+\psi} \right| \cong 88 \cdot \underbrace{10^{-4}}_{\rightsquigarrow 2u} = 88 \cdot 2u.$$

Παίρνουμε μεγάλο σφάλμα λόγω της μορφής του x και ψ. Το ίδιο μπορεί να συμβεί κατά την πρόσθεση *μεγάλων σε απόλυτη τιμή* ετερόσημων αριθμών. Αν για παράδειγμα στον ίδιο υπολογιστή $x = 451852000$ και $\psi = -451851000$ οπότε $x + \psi = 1000$ και $z = fl(fl(x) + fl(\psi)) = 0$. Σε κάποιες περιπτώσεις τα προβλήματα αίρονται με τροποποίηση του αλγορίθμου.

Παραδείγματα.

(α) Στον υπολογιστή HP33 ($\beta = 10$, $t = 10$) χειρού θέλουμε να υπολογίσουμε $\sqrt{7892} - \sqrt{7891}$. Έχουμε, $\sqrt{7892} = .8883692926 \cdot 10^2$ και $\sqrt{7891} = .8883130079 \cdot 10^2$ άρα

$$\sqrt{7892} - \sqrt{7891} = .562847 \underbrace{0000}_{\text{απώλεια ακριβείας}} \cdot 10^{-2}.$$

Ισχύει $\boxed{\sqrt{x} - \sqrt{\psi} = \frac{x-\psi}{\sqrt{x}+\sqrt{\psi}}}$

οπότε $\sqrt{7892} - \sqrt{7891} = \dfrac{1}{\sqrt{7892}+\sqrt{7891}} = 5628468294 \cdot 10^{-2}$.

(β) Θέλουμε να υπολογίσουμε με ακρίβεια τις τιμές της συνάρτησης $f(x) = x - \sin x$ για $|x|$ μικρό

$$\lim_{x\to 0} \frac{\sin x}{x} = 1 \rightarrow (\text{οι τιμές του } \sin x \text{ πλησιάζουν πολύ τις τιμές του } x \text{ για } x \to 0).$$

Εφαρμόζουμε Taylor για την

$$\sin x,\ \sin x = x - \frac{x^3}{6} + \varepsilon(x),\ |\varepsilon(x)| \le \frac{|x|^5}{24}.$$

Άρα $f(x) \cong \frac{x^3}{6}$ με σφάλμα μικρότερο του $\frac{|x|^5}{24}$.

Παράδειγμα. Ας υποθέσουμε ότι θέλουμε να υπολογίσουμε το άθροισμα

$$\begin{aligned} S_n &= 1+\sum_{\kappa=1}^{n}\frac{1}{\kappa^2+\kappa}=1+\sum_{\kappa=1}^{n}\frac{1}{\kappa(\kappa+1)} \\ &= 1+\sum_{\kappa=1}^{n}(\frac{1}{\kappa}-\frac{1}{\kappa+1})=1+(1-\frac{1}{2})+(\frac{1}{2}-\frac{1}{3})+\cdots+(\frac{1}{n}-\frac{1}{n+1}) \\ &= 2-\frac{1}{n+1} \end{aligned}$$

1ος Αλγόριθμος

$$\begin{aligned} s_0 &= 1 \\ s_\kappa &= s_{\kappa-1}+\frac{1}{\kappa(\kappa+1)},\ \kappa=1,\,2,\cdots,n \end{aligned}$$

2ος Αλγόριθμος

$$\begin{aligned} T_0 &= \frac{1}{n(n+1)} \\ T_\kappa &= T_{\kappa-1}+\frac{1}{(n-\kappa)(n-\kappa+1)},\ \kappa=1,\,2,\cdots,n-1 \\ T_n &= T_{n-1}+1 \end{aligned}$$

Τώρα στον HP33 ($\beta=10,\ t=10$) παίρνουμε τις ακόλουθες τιμές

n	S_n	$\tilde{S}_n$	$\tilde{T}_n$
9	$1,9$	$1,900000000$	$1,900000000$
99	$1,9$	$1,990000003$	$1,990000000$
999	$1,999$	$1,999000003$	$1,999000000$
9999	$1,9999$	$1,999899972$	$1,999900000$

Έστω N αριθμοί μηχανής $\alpha_i,\ 1\le i\le N$. Για τον υπολογισμό του αθροίσματος $\S_N=\sum_{\kappa=1}^{N}$ χρησιμοποιούμε τον αλγόριθμο

$$\begin{aligned} s_1 &= \alpha_1 \\ s_\kappa &= s_{\kappa-1}+\alpha_\kappa,\ \kappa=2,\,3,\cdots,N \end{aligned}$$

ο οποίος εφαρμοζόμενος στον υπολογιστή παράγει ως μερικά αθροίσματα τους αριθμούς $\tilde{S}_\kappa$, $\kappa=1,\,2,\,\cdots,N$ όπου

$$\begin{aligned} \tilde{s}_1 &= \alpha_1 \\ \tilde{s}_\kappa &= fl(\tilde{s}_{\kappa-1}+\alpha_\kappa),\ \kappa=2,\,3,\,\cdots,N \end{aligned}$$

Μπορεί να αποδειχθεί ότι

$$\left|\frac{\tilde{S}_N - S_N}{S_N}\right| \lessapprox \frac{\gamma_N}{|S_N|}u \tag{3}$$

όπου $\gamma_N = |S_2| + |S_3| + \cdots + |S_N|$. Το πηλίκο $p_N = \dfrac{\gamma_N}{|S_n|}$ λέγεται συντελεστής μετάδοσης του (σχετικού) σφάλματος. Αν $\alpha_\kappa > 0$ για $\kappa = 1, 2, \cdots, N$ τότε $\gamma_N = (N-1)\alpha_1 + (N-1)\alpha_2 + (N-2)\alpha_3 + \cdots + \alpha_N$, δηλαδή για να είναι ο γ_N όσο το δυνατόν μικρότερο, ξεκινάω από τον μικρότερο στον μεγαλύτερο α_i. Σημειώνουμε ότι ο 2ος αλγόριθμος δίνει καλύτερα αποτελέσματα από τον 1ο.

Σφάλματα στον υπολογισμό αθροισμάτων(συνέχεια από το προηγούμενο μάθημα). Έστω ότι η εκτίμηση (6) ισχύει για τα $\tilde{S}_N$ και $\tilde{T}_N$, υποθέτοντας ότι οι όροι έχουν μετατραπεί εκ των προτέρων σε αριθμούς μηχανής.

1ος Αλγόριθμος. Για $N = n+1$ έχουμε $\alpha_1 = 1$, $\alpha_\kappa = \dfrac{1}{(\kappa-1)\kappa}$, $\kappa = 2, 3, \cdots, N$ και $s_N = S_N = 2 - \dfrac{1}{1+n}$ και αφήνεται ως άσκηση το $\gamma_N = 2n - (\frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n+1})$.

(Ορισμός: $f_m \sim g_m$ αν $\lim_{m\to\infty} \dfrac{f_m}{g_m} = 1$, για μεγάλο m.)

Ισχυρίζομαστε ότι $1 + \frac{1}{2} + \cdots + \frac{1}{m} \sim \ln m$ για μεγάλο m. Για την απόδειξη αυτού του ισχυρισμού προσεγγίζουμε το ολοκλήρωμα $\int_1^m \frac{dx}{x} = \ln m$, $m > 1$ με άθροισμα Riemann (πάνω και κάτω) εμβαδών ορθογωνίων μοναδιαίου πλάτους. Άρα,

$$\gamma_N \sim 2n - \ln n \sim 2n (\text{για μεγάλο } n).$$

Τελικά, δεδομένου ότι $S_N \sim 2$, $\Rightarrow p_N = \dfrac{\gamma_N}{|S_N|} \Rightarrow p_N \sim n$.

2ος Αλγόριθμος. Για τον 2ο αλγόριθμο έχουμε:

$$\begin{aligned} s_1 &= T_0 = \frac{1}{n(n+1)}, \\ s_\kappa &= T_{\kappa-1} = \frac{1}{n-\kappa+1} - \frac{1}{n+1},\ \kappa = 2, 3, \cdots, N-1 \\ s_N &= T_n = 2 - \frac{1}{n+1} \end{aligned}$$

Επομένως,

$$\begin{aligned} \gamma_N &= (\frac{1}{n-1} - \frac{1}{n+1}) + (\frac{1}{n-2} - \frac{1}{n+1}) + \cdots + (1 - \frac{1}{n+1}) + (2 - \frac{1}{n+1}) \\ &= 2 + (1 + \frac{1}{2} + \cdots + \frac{1}{n-1}) - \frac{n}{n+1} \end{aligned}$$

Ισχυριζόμαστε ότι $\gamma_N \sim \ln n$. $p_N \sim \dfrac{1}{2}\ln n$ αφού $S_N \sim 2$. Για $n = 10^4$ τότε $\dfrac{1}{2}\ln n \cong 4,6$.

Παράδειγμα. Να προσεγγιστεί το e^{-x} για $x >> 1$ από μερικά αθροίσματα της σειράς Taylor $S_N(x) = 1 - x + \frac{x^2}{2!} - \cdots + (-1)^{N-1}\frac{x^{N-1}}{(N-1)!}$. Για $x = 100$ είναι $e^{-x} \cong 0$ ενώ

$$s_1 = 1,\ s_2 = -99,\ s_3 = 4901,\ s_4 \cong -161766$$

που σημαίνει ότι το $p_N = \frac{\gamma_N}{|S_N|}$ θα είναι πολύ μεγάλο. Επομένως, το σφάλμα θα είναι πολύ μεγάλο. Ο σωστός τρόπος υπολογισμού είναι $e^{-x} = \frac{1}{e^x}$, όπου το e^x προσεγγίζεται με αθροίσματα της σειράς Taylor.

Ευστάθεια Αλγορίθμων

Ο αλγόριθμος μπορεί να είναι αριθμητικά ευσταθής ή αριθμητικά ασταθής.
Παραδείγματα.

(α) Ο αλγόριθμος της προσέγγισης του e^{-x} για $x >> 1$ με μερικά αθροίσματα της σειράς Taylor είναι (αριθμητικά) ασταθής.

(β) Ο αλγόριθμος $e^{-x} = \frac{1}{e^x}$, όπου το e^x προσεγγίζεται με αθροίσματα της σειράς Taylor είναι (αριθμητικά) ευσταθής.

(γ) Ένα ακόμη παράδειγμα. Να υπολογιστεί το ολοκλήρωμα

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

για αρκετά μεγάλο n. Μερικές ιδιότητες των I_n είναι οι ακόλουθες:

1. $I_n > 0,\ n = 1, 2, \cdots$
2. $I_{n+1} < I_n,\ n \geq 1$ (για την απόδειξη:$x^{n+1} < x^n$, $x \in (0, 1)$).
3. $I_n \leq \int_0^1 x^n ds = \frac{1}{n+1}$.
4. $\lim I_n = 0$ για $n \to \infty$.
5. $I_n = \frac{n!}{(-1)^{n+1}}[\frac{1}{e} - \sum_{\kappa=0}^{n} \frac{(-1)^\kappa}{\kappa!}],\ n = 1, 2, \cdots$ (απόδειξη με επαγωγή).

Ο τύπος αυτός για μεγάλο n δίνει έναν ασταθή αλγόριθμο (γιατί·). Χρησιμοποιώντας ολοκλήρωση κατά μέρη έχουμε,

$$\int_0^1 x^n e^{x-1} dx = x^n e^{x-1}\Big|_0^1 - \int_0^1 e^{x-1} n x^{n-1} ds = 1 - n\int_0^1 x^{n-1} e^{x-1} ds$$

$$\Rightarrow I_n = 1 - nI_{n-1},\ n = 2, 3, \cdots, \text{με } I_1 = \frac{1}{e}.$$

Εφαρμόζοντας αυτόν τον αλγόριθμο σε έναν υπολογιστή $\beta = 10$, $t = 6$ παίρνουμε:

n	$\tilde{I}_n$
1	.367879
2	.264242
3	.207274
7	.110160
8	.118720?
9	−.068480?!

το μοναδικό σφάλμα στρογγύλευσης στον υπολογιστή του

$$\tilde{I}_n = 1 - n\tilde{I}_{n-1},\ n = 2, 3, \cdots \text{με } \tilde{I}_1 = I_1 + \varepsilon_1, \text{ όπου } \varepsilon_1 \cong -4,4 \cdot 10^{-7}.$$

Έχουμε στο $n-$ στό βήμα το σφάλμα ε_n,

$$\varepsilon_n = \tilde{I}_n - I_n = 1 - n\tilde{I}_{n-1} - (1 - nI_{n-1})$$

$$= -n(\tilde{I}_{n-1} - I_{n-1}) = -n\varepsilon_{n-1},\ n \geq 2.$$

Επαγωγικά προκύπτει,

$$\varepsilon_n = (-1)^{n-1} n! \varepsilon_1,\ n \geq 2.$$

Παράδειγμα. Για $n = 9$: $\varepsilon_9 = 9!$, $\varepsilon_1 \cong -0,16$. Παρατηρούμε ότι η 'ζημιά' στον προηγούμενο αλγόριθμο προκαλείται από τον δείκτη n, της αναδρομικής σχέσης $I_n = 1 - nI_{n-1}$.

Θεωρούμε τον αναδρομικό τύπο $I_{n-1} = \dfrac{1 - I_n}{n}$. Έτσι αν γνωρίζουμε το I_m, μπορούμε να υπολογίσουμε ένα I_κ, $\kappa < m$, μέσω της ακολουθίας $I_{m-1}, I_{m-2}, I_\kappa$. Επειδή δεν ξέρουμε ακριβώς το I_m, ο αλγόριθμος έχει ως εξής $\boxed{\tilde{I}_m = I_m + \varepsilon_m}$

$$\tilde{I}_{n-1} = \frac{1 - \tilde{I}_n}{n},\ n = m, m-1, \cdots \kappa + 1.$$

Έχουμε,

$$\varepsilon_{n-1} = \tilde{I}_{n-1} - I_{n-1} = \frac{1}{n} - \frac{1}{n}\tilde{I}_n = -\frac{1}{n}(\tilde{I}_n - I_n) = -\frac{1}{n}\varepsilon_n,\ m \geq n \geq \kappa + 1.$$

Άρα,

$$\varepsilon_\kappa = (-1)^{m-\kappa} \frac{1}{\kappa+1} \frac{1}{\kappa+2} \frac{1}{m} \varepsilon_m.$$

Ο παραπάνω είναι ένας ευσταθής αλγόριθμος.
Παράδειγμα. Έστω $m = 20$ και $\tilde{I}_{20} = 0$ (γενικά ισχύει $0 < I_n < \frac{1}{n+1}$) $\Rightarrow |\varepsilon_{20}| < \frac{1}{21}$ τότε στον υπολογιστή με $\beta = 10$, $t = 6$ παίρνουμε μετά από 11 βήματα $\tilde{I}_9 = .916123 \cdot 10^{-1}$.

Ασκήσεις (από βιβλίο Ακριβή-Δουγαλή)

1.2 Βρείτε κατάλληλους τρόπους υπολογισμού των παρακάτω παραστάσεων, έτσι ώστε να μην χάνεται ακρίβεια, όταν οι πράξεις γίνονται με αριθμητική κινητής υποδιαστολής και πεπερασμένη ακρίβεια.

(α) $1 - \cos x$, για μικρό $|x|$ χωρίς Taylor,

(β) $e^{x-\psi}$, για μεγάλα θετικά x, ψ,

(γ) $\log x - \log \psi$ για μεγάλα θετικά x, ψ,

(δ) $\sin \alpha + x - \sin \alpha$ για μικρό $|x|$.

Λύση.

(α) $1 - \cos x = 2\sin^2(\frac{x}{2})$ όταν x κοντά στο 0 τότε $\cos x$ κοντά στο 1, άρα cancelation error,

(β) $e^{x-\psi} = \dfrac{e^x}{e^\psi}$,

(γ) $\log x - \log \psi = \dfrac{\log x}{\log \psi}$,

(δ) $\sin \alpha + x - \sin \alpha = 2\cos(\alpha + \dfrac{x}{2}) \sin \dfrac{x}{2}$, όταν $\alpha + x$ κοντά στο $\alpha \Rightarrow \sin(\alpha + x)$ κοντά στο $\sin \alpha$ άρα cancelation error.

1.3 Θεωρείστε τη δευτεροβάθμια εξίσωση $x^2 - 2\alpha x + \beta = 0$ όπου α, $\beta > 0$ και $\alpha^2 >> \beta$. Δώστε έναν ευσταθή αλγόριθμο για τον υπολογισμό των ριζών της.
Λύση.

$$x_{1,2} = \frac{2\alpha \pm \sqrt{4\alpha^2 - 4\beta}}{2} = \alpha \pm \sqrt{\alpha^2 - \beta}$$

$$\Rightarrow x_1 = \alpha + \sqrt{\alpha^2 - \beta}, \text{ και } x_2 = \alpha - \sqrt{\alpha^2 - \beta} \cong \alpha - \sqrt{\alpha^2} \cong \alpha - \alpha,$$

παρουσιάζεται πρόβλημα γιατί αφαιρούμε ομόσημους αριθμούς $x - \psi$ με $x \cong \psi$. Άρα,

$$x_2 = \frac{(\alpha - \sqrt{\alpha^2 - \beta})(\alpha + \sqrt{\alpha^2 - \beta})}{(\alpha + \sqrt{\alpha^2 - \beta})} = \frac{\alpha^2 - \alpha^2 + \beta}{x_1} = \frac{\beta}{x_1}.$$

1.6 (α) Αποδείξτε ότι η ακολουθία $\psi_\kappa = 2^\kappa \tan\dfrac{\pi}{2^\kappa}$, $\kappa = 2,\, 3, \cdots$ είναι φθίνουσα και συγκλίνει στο π.

(β) Αποδείξτε ότι η ψ_κ παράγεται αναδρομικά από τον αλγόριθμο

$$\begin{aligned} \psi_2 &= 4 \\ \psi_{\kappa+1} &= 2^{2\kappa+1}\frac{\sqrt{1+(2^{-\kappa}\psi_\kappa)^2}-1}{\psi_\kappa},\ \kappa = 2,\, 3, \cdots \end{aligned}$$

(γ) Αν κάνουμε πράξεις με αριθμητική κινητής υποδιαστολής με πεπερασμένη ακρίβεια, παρατηρούμε ότι ο αλγόριθμος είναι ασταθής. Ποιά είναι η αιτία της αστάθειας·

(δ) Βρείτε έναν ευσταθή αλγόριθμο για τον υπολογισμό του ψ_κ.

Λύση.

(α) Έχουμε, $\psi_\kappa = 2^\kappa \tan\dfrac{\pi}{2^\kappa} = 2^\kappa \tan(\dfrac{2\pi}{2^{\kappa+1}})$. Μέσω του τύπου, $\tan 2\varphi = \dfrac{2\tan\varphi}{1-\tan^2\varphi}$, παίρνουμε

$$\psi_\kappa = 2^\kappa \frac{2\tan\dfrac{\pi}{2^{\kappa+1}}}{1-\tan^2\dfrac{\pi}{2^{\kappa+1}}} = \frac{2^{\kappa+1}\tan\dfrac{\pi}{2^{\kappa+1}}}{1-\tan^2\dfrac{\pi}{2^{\kappa+1}}} = \frac{\psi_{\kappa+1}}{1-\tan^2\dfrac{\pi}{2^{\kappa+1}}} \geq \psi_{\kappa+1}.$$

Παρατηρούμε ότι $0 < 1-\tan^2\dfrac{\pi}{2^{\kappa+1}} < 1$ γιατί η ακολουθία $\tan(\dfrac{2\pi}{2^{\kappa+1}})$ είναι φθίνουσα. Άρα η ψ_κ είναι φθίνουσα. Για το όριό της έχουμε,

$$\lim_{\kappa\to\infty}\psi_\kappa = \lim_{\kappa\to\infty} 2^\kappa \tan\frac{\pi}{2^\kappa} = \lim_{\kappa\to\infty}\pi\frac{\tan\dfrac{\pi}{2^\kappa}}{\dfrac{\pi}{2^\kappa}} = \pi\lim_{\kappa\to\infty}\frac{\tan\dfrac{\pi}{2^\kappa}}{\dfrac{\pi}{2^\kappa}} = \pi\cdot 1 = \pi,$$

διότι

$$\lim_{x\to 0}\frac{\tan x}{x} = 1, \text{ αφού } \lim_{x\to 0}\frac{\sin x}{x} = 1.$$

(β) Για τον αναδρομικό τύπο έχουμε,

$$2^{2\kappa+1}\frac{\sqrt{1+(2^{-\kappa}\psi_\kappa)^2}-1}{\psi_\kappa} = 2^{2\kappa+1}\frac{\sqrt{1+\tan^2\dfrac{\pi}{2^\kappa}}-1}{2^\kappa\tan\dfrac{\pi}{2^\kappa}} = 2^{\kappa+1}\frac{\sqrt{\dfrac{1}{\cos^2(\dfrac{\pi}{2^\kappa})}-1}}{\tan\dfrac{\pi}{2^\kappa}}$$

$$= 2^{\kappa+1}\frac{\dfrac{1-\cos\dfrac{\pi}{2^\kappa}}{\cos\dfrac{\pi}{2^\kappa}}}{\dfrac{\sin\dfrac{\pi}{2^\kappa}}{\cos\dfrac{\pi}{2^\kappa}}} = 2^{\kappa+1}\frac{1-\cos\dfrac{\pi}{2^\kappa}}{\sin\dfrac{\pi}{2^\kappa}} = 2^{\kappa+1}\frac{2\sin^2\dfrac{\pi}{2^{\kappa+1}}}{2\sin\dfrac{\pi}{2^{\kappa+1}}\cos\dfrac{\pi}{2^{\kappa+1}}}$$

$$= 2^{\kappa+1}\tan\frac{\pi}{2^{\kappa+1}} = \psi_{\kappa+1}.$$

(γ) Παρατηρούμε ότι,

$$\psi_\kappa \to \pi,\ \kappa\to\infty \Rightarrow\ 2^{-\kappa}\psi_\kappa \to 0,\ \kappa\to\infty \to \sqrt{1+(2^{-\kappa}\psi_\kappa)^2}-1 \cong 1.$$

Άρα $\sqrt{1+(2^{-\kappa}\psi_\kappa)^2}-1$ είναι ασταθής πράξη.

(δ)

$$\psi_{\kappa+1} = 2^{2\kappa+1}\frac{(\sqrt{1+(2^{-\kappa}\psi_\kappa)^2}-1)(\sqrt{1+(2^{-\kappa}\psi_\kappa)^2}+1)}{\psi_\kappa\sqrt{1+(2^{-\kappa}\psi_\kappa)^2}+1} = \cdots = \frac{2_\kappa}{\sqrt{1+(2^{-\kappa}\psi_\kappa)^2}+1},$$

ευσταθής.

1.12 Θέλουμε να υπολογίσουμε για δεδομένη (μεγάλη) σταθερά $\alpha>1$, τους όρους της ακολουθίας

$$\psi_n = \int_0^1 \frac{x^n}{x+\alpha}dx,\ n=0,\,1,\,2,\cdots$$

(α) Αποδείξτε ότι για κάθε $\alpha>0$ η (ψ_n) είναι γνησίως φθίνουσα και ότι $\lim_{n\to\infty}\psi_n=0$,

(β) Αποδείξτε ότι για $n\geq 1$,

$$\psi_n = \sum_{\kappa=0}^{n-1}(-1)^\kappa\binom{n}{\kappa}\alpha^\kappa\frac{(1+\alpha)^{n-\kappa}-\alpha^{n-\kappa}}{n-\kappa}+(-\alpha)^n\log\frac{1+\alpha}{\alpha},$$

είναι ο τρόπος αυτός υπολογισμού ευσταθής,

(γ) Προσδιορίστε αναδρομικό τύπο για τον υπολογισμό του ψ_n συναρτήσει του ψ_{n-1} και υπολογίστε αναλυτικά το ψ_0. Αποδείξτε ότι ο αλγόριθμος που προκύπτει δεν είναι ευσταθής,

(δ) Δώστε έναν ευσταθή αλγόριθμο για τον υπολογισμό π.χ. του ψ_{10}.

Λύση. Υποδείξεις.

(α) Ισχύει $x^{n+1} < x^n$ για $x \in (0,\, 1)$

$$0 < \psi_n = \int_0^1 \frac{x^n}{x+\alpha}dx < \int_0^1 \frac{x^n}{x}dx = \int_0^1 x^{n-1}dx = \frac{1}{n} \Rightarrow \psi_n \to 0,\, n \to 0.$$

(β)

$$\psi_n = \int_o^1 \frac{x^n}{x+\alpha}dx = \int_0^1 \frac{[(x+\alpha)-\alpha]^n}{x+\alpha}dx = \int_0^1 \sum_{\kappa=0}^{n}(-1)^\kappa \binom{n}{\kappa} \frac{(x+\alpha)^{n-\kappa}\alpha^\kappa}{x+\alpha}dx$$

$$= \int_0^1 \sum_{\kappa=0}^{n}(-1)^\kappa \binom{n}{\kappa} (x+\alpha)^{n-\kappa-1}\alpha^\kappa dx = \cdots,$$

όπου βάζουμε ολοκλήρωμα μέσα και παίρνουμε περιπτώσεις για το κ. Για την ευστάθεια έχουμε, αν π.χ. $\alpha = 10$, $n = 10$, $\kappa = 5$ έχουμε $(-1)^5 \binom{10}{5} 10^5 \frac{11^5 - 10^5}{5} \cong -3 \cdot 10^{11}$.

(γ)

$$\psi_n = \int_0^1 \frac{x^n}{x+\alpha}dx = \int_0^1 \frac{x^{n-1}}{x+\alpha}xdx = \int_0^1 \frac{x^{n-1}}{x+\alpha}[(x+\alpha)-\alpha]dx$$

$$= \int_0^1 x^{n-1}dx - \alpha\int_0^1 \frac{x^{n+1}}{x+\alpha}dx = \frac{1}{n} - \alpha\psi_{n-1},$$

και $\psi_0 = \log(\frac{1+\alpha}{\alpha})$ που είναι ασταθής.

(δ) Με το παράδειγμα στο τέλος του προηγούμενου μαθήματος αποδεινύουμε την αστάθεια και βρίσκουμε το ζητούμενο. ;;;

Mέθοδος της διχοτόμησης

Παρατήρηση: Θα ακολουθήσουμε τον εξής συμβολισμό.
Έστω $I \subset \mathbb{R}$, $n \in \mathbb{N}$ τότε

$$C(I) = \{f | f : I \to \mathbb{R},\ f \text{ συνεχής}\}$$

και

$$C^n(I) = \{f | f \in C(I) \text{ και } n \text{ φορές παραγωγίσιμη στο } I\}$$

Γράφουμε $C([a,\, b])$, $C[a,\, b]$, $C(a,\, b)$, $C^n[a,\, b]$, $C^n(a,\, b)$.

Θεώρημα Ενδιάμεσης Τιμής
Έστω $f \in C[a,\, b]$ και $f(a)f(b) < 0$ τότε υπάρχει $x^\star \in (a,\, b)$ τέτοιο ώστε $f(x^\star) = 0$.

Μέθοδος της Διχοτόμησης

Αν $f \in C[a,\, b]$, $f(a)f(b) < 0$ τότε υπάρχει ρίζα της f στο $[a,\, b]$. Έστω $a_1 = a$, $b_1 = b$, $x_1 = \dfrac{a_1 + b_1}{2} = \dfrac{a + b}{2}$

(i) Αν $f(x_1) = 0 \Rightarrow x_1$ ρίζα π.χ. σχήμα

(ii) Αν $f(x_1) \neq 0$ τότε

1. $1^η$ Περίπτωση: $f(a_1)f(x_1) < 0 \Rightarrow \exists$ ρίζα στο $[a_1,\, x_1]$ π.χ. σχήμα
   Θέτουμε $a_2 = a_1$, $b_2 = x_1$.
2. $2^η$ Περίπτωση: $f(a_1)f(x_1) > 0 \Rightarrow f(x_1)f(b_1) < 0\, \exists$ ρίζα στο $[x_1,\, b_1]$ π.χ. σχήμα
   Θέτουμε $a_2 = x_1$, $b_2 = b_1$.

Το $[a_2,\, b_2] = \begin{cases} [a_1, x_1] \\ [x_1, b_1] \end{cases}$ έχει μήκος το μισό του $[a_1,\, b_1]$. Επαναλαμβάνουμε τη διαδικασία στο $[a_2,\, b_2]$.

Πρόταση.
Έστω $f \in C[a,\, b]$ και $f(a)f(b) < 0$ και $(x_n)\, n \in \mathbb{N}$ η ακολουθία την οποία δίνει η μέθοδος της διχοτόμησης. Τότε είτε $x_N = x^\star$ για κάποιο N, είτε $x_n \to x^\star$, $n \to +\infty$, όπου $x^\star \in (a,\, b)$ ρίζα της $f(x) = 0$.
Μάλιστα ισχύει για την εκτίμηση σφάλματος $|x^\star - x_n| \leq \dfrac{b-a}{2^n}$, $n = 1,\, 2, \cdots$ (το απόλυτο σφάλμα

στο $n-$οστό βήμα είναι $\frac{b-a}{2}$).

Απόδειξη.
Έχουμε

$$b_n - a_n = \frac{b_{n-1} - a_{n-1}}{2} = \frac{b_{n-2} - a_{n-2}}{4} = \cdots = \frac{b-a}{2^{n-1}}$$

$$x_n = \frac{a_n + b_n}{2},\ x^\star \in [a_n,\, b_n] \text{ σχήμα}$$

Επομένως,

$$|x^\star - x_n| \leq \frac{b_n - a_n}{2} = \frac{b-a}{2^n}$$

Από θεώρημα ισοσυγκλινουσών (sandwich) προκύπτουν και τα υπόλοιπα συμπεράσματα.
Αριθμός βημάτων έτσι ώστε $|x^\star - x_n| \leq \varepsilon$.
Αρκεί

$$\frac{b-a}{2^n} \leq \varepsilon \Rightarrow 2^n \geq \frac{b-a}{\varepsilon} \Rightarrow n \ln 2 \geq \ln(\frac{b-a}{\varepsilon}) \Rightarrow n \geq \frac{\ln(\frac{b-a}{\varepsilon})}{\ln 2}$$

Παράδειγμα
Να βρεθεί ο απαιτούμενος αριθμός βημάτων όταν $[a,\, b] = [0,\, 1]$ και $\varepsilon = 5 \cdot 10^{-7}$.
Από τον παραπάνω τύπο αρκεί $n \geq \frac{\ln(0,2 \cdot 10^7)}{\ln 2} \cong 20,93 \Rightarrow n = 21$.
Η ακρίβεια της τάξης $5 \cdot 10^{-7}$ έχει σαν συνέπεια πολλά βήματα. Αν το σφάλμα είναι $\varepsilon_n = |x_n - x^\star|$ τότε $\varepsilon_n \cong \frac{\varepsilon_{n-1}}{2}$ (αργή σύγκλιση).

Επαναληπτικές Μέθοδοι
$f(x) = 0 \Leftrightarrow f(x) + x = x$. Θέτουμε $\varphi(x) = x \to x$: σταθερό σημείο της φ.
$x_{n+1} = \varphi(x_n)$, $n = 0,\, 1,\, 2,\, \cdots$
Δεδομένο x_0.
Αν η ακολουθία (x_n), $n \in \mathbb{N}$ συγκλίνει στο $x^\star$ και η φ είναι συνεχής στο $x^\star$, έχουμε

$$x^\star = \lim_{n\to\infty} x_n = \lim_{n\to\infty} \varphi(x_{n-1}) = \varphi(\lim_{n\to\infty}) = \varphi(x^\star)$$

Παράδειγμα
$x^3 + 5x - 4 = 0 \Leftrightarrow x = \frac{4 - x^3}{5}$ άρα $\varphi(x) = \frac{4-x^3}{5}$.
Άρα $x_{n+1} = \frac{4 - x_n^3}{5},\ n = 0,\, 1,\, 2,\, \cdots$ (με δεδομένο x_0).
Αν $x_n \to p$, $n \to \infty$ τότε $\lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} \frac{4 - x_n^3}{5} = \frac{4 - (\lim_{n\to\infty} x_n)^3}{5} \Rightarrow p = \frac{4-p^3}{5}$
$\Leftrightarrow p^3 + 5p - 4 = 0 \Rightarrow p$ ρίζα της εξίσωσης.

Πρόταση (Ύπαρξης σταθερού σημείου)
Κάθε συνεχής συνάρτηση $\varphi : [a, b] \to [a, b]$ έχει στο διάστημα $[a, b]$ (τουλάχιστον) ένα σταθερό σημείο.
Απόδειξη.
σχήμα
Αν $\varphi(a) = a$ ή $\varphi(b) = b$ τότε ισχύει.
Αν όχι, τότε $\varphi(a) > a$ και $\varphi(b) < b$. Ορίζουμε $g : [a, b] \to \mathbb{R}$ με $g(x) = \varphi(x) - x, g \in [a, b]$.
$$\left.\begin{array}{l} g(a) = \varphi(a) - a > 0 \\ g(b) = \varphi(b) - b < 0 \end{array}\right\} \Rightarrow g(a)g(b) < 0$$
Από θεώρημα Ενδιάμεσης Τιμής υπάρχει $x^\star \in (a, b) : g(x^\star) = 0 \Leftrightarrow \varphi(x^\star) = x^\star$.
Παράδειγμα
Δείξτε ότι η $\varphi(x) = 3^{-x}$ έχει (τουλάχιστον) ένα σταθερό σημείο στο $[0, 1]$.
Λύση:
Η φ είναι συνεχής, με αρνητική παράγωγο $\varphi'(x) = -3^{-x} \ln 3 < 0$ άρα η φ είναι φθίνουσα.
$\Rightarrow \varphi([0, 1]) = [\varphi(1), \varphi(0)] = [\frac{1}{3}, 1] \subset [0, 1]$.
Άρα η φ έχει (τουλάχιστον) ένα σταθερό σημείο.
Παρατήρηση: Η συνθήκη της πρότασης είναι ικανή αλλά όχι αναγκαία.
π.χ. Η $\varphi : [-1, 1] \to [0, 2]$, $\varphi(x) = 2x^2$ είναι συνεχής και με σταθερά σημεία τα 0, $\frac{1}{2}$ αλλά δεν ισχύει $\varphi([-1, 1]) \subset [-1, 1]$.

Συστολή
Ορισμός. Η συνάρτηση $\varphi : [a, b] \to \mathbb{R}$ ονομάζεται συστολή αν υπάρχει σταθερό L, $0 \leq L \leq 1$, τέτοια ώστε $|\varphi(x) - \varphi(\psi)| \leq L|x - \psi|$, $\forall x, \psi \in [a, b]$.
π.χ. Να δείξετε ότι η $\varphi(x) = \frac{1}{2}x^2$ είναι συστολή
$|\varphi(x) - \varphi(\psi)| = |\frac{1}{2}x^2 - \frac{1}{2}\psi^2| = \frac{1}{2}(x + \psi)|x - \psi| \leq \frac{1}{2}(\frac{1}{2} + \frac{1}{2})|x - \psi| = \frac{1}{2}|x - \psi|$.
Συνεπώς η φ είναι συστολή με σταθερά $L = \frac{1}{2}$.
Παρατήρηση: Αν η $\varphi \in C^1[a, b]$ τότε,
$$|\varphi(x) - \varphi(\psi)| \leq \max_{a \leq \xi \leq b} |\varphi'(\xi)|, \forall x, \psi \in [a, b].$$
Απόδειξη.
Εφαρμόζουμε το Θεώρημα Μέσης Τιμής (άσκηση, για το $max|g(x)|$ ελέγχουμε άκρα και ακρότατο - δηλαδή εκεί που μηδενίζεται η παράγωγος και τα άκρα.)
Παράδειγμα
$\varphi(x) = \frac{x^2 + 5}{6} \Rightarrow \varphi'(x) = \frac{x}{3}, |\varphi'(x)| = |\frac{x}{3}| < 1 \Rightarrow |x| < 3$ άρα η φ είναι συστολή σε κάθε

$[a,\, b] \subseteq (-3,\, 3)$.
(Στο $[-3,\, 3]$ ή σε διάστημα που έχει σημεία εκτός του $(-3,\, 3)$;)

Θεώρημα (Μοναδικού σταθερού σημείου)
Έστω $\varphi : [a,\, b] \to [a,\, b]$ συστολή με σταθερά L, $0 \leq L \leq\leq 1$. Τότε η φ έχει ένα μοναδικό σταθερό σημείο στο $[a,\, b]$.
Απόδειξη.
Η ύπαρξη αποδείχτηκε (από προηγούμενη πρόταση). Έστω ότι υπάρχει $x^\star,\, \psi^\star \in [a,\, b]$ με $x^\star \neq \psi^\star$ τέτοια ώστε $x^\star = \varphi(x^\star)$ και $\psi^\star = \varphi(\psi^\star)$, τότε

$$|x^\star - \psi^\star| = |\varphi(x^\star) - \varphi(\psi^\star)| \leq L|x^\star - \psi^\star| < |x^\star - \psi^\star| \neq 0$$

άτοπο αφού $L < 1$.

Παράδειγμα
Δείξτε ότι η $\varphi(x) = \dfrac{x^2 + 5}{6}$ έχει μοναδικό σταθερό σημείο στο $[-2,\, 2]$. Από το προηγούμενο παράδειγμα η φ είναι συστολή στο $[-2,\, 2] \subset (-3,\, 3)$.

$\varphi([-2,\, 2]) \to [\dfrac{5}{6},\, \dfrac{3}{2}] \subset [-2,\, 2]$ σχήμα

Άρα η φ λεχει μοναδικό σταθερό σημείο στο διάστημα $[-2,\, 2]$. Για να το βρούμε θέτουμε $x = \varphi(x) \Rightarrow 6x^2 = x^2 + 5 \Rightarrow x = 1,\, x = 5$(απορρίπτεται). Άρα στο $[-2,\, 2]$ το σταθερό σημείο είναι το $x = 1$.

Θεώρημα (Σύγκλιση *γενικής* επαναληπτικής μεθόδου)
Έστω $\varphi : [a,\, b] \to [a,\, b]$ συστολή με σταθερά L, $0 \leq L \leq 1$ και $x_0 \in [a,\, b]$. Τότε η ακολουθία $(x_n)_{n\in\mathbb{N}}$ με $x_{n+1} = \varphi(x_n)$ είναι καλώς ορισμένη, δηλαδή $\forall n \in \mathbb{N}_\not{\nu},\, x_n \in [a,\, b]$, συγκλίνει στο μοναδικό σταθερό σημείο $x^\star$ της φ και για το σφάλμα ισχύουν οι εκτιμήσεις:

1. $|x_n - x^\star| \leq \dfrac{L^n}{1 - L}|x_1 - x_0|$
2. $|x_n - x^\star| \leq L^n|x_0 - x^\star|$
3. $|x_n - x^\star| \leq \dfrac{L}{1 - L}|x_n - x_{n-1}|$

Απόδειξη.
Η ύπαρξη και η μοναδικότητα του σταθερού σημείου $x^\star$ έχουν αποδειχθεί.

Επιπλέον, η ακολουθία $(x_n)_{n\in\mathbb{N}}$ είναι καλώς ορισμένη εφόσον $\varphi : [a, b] \to [a, b]$. Θα δείξουμε ότι η (x_n) είναι ακολουθία Cauchy. Έχουμε

$$|x_{n+1} - x_n| = |\varphi(x_n) - \varphi(x_{n-1})| \leq L|x_n - x_{n-1}| = L|\varphi(x_{n-1}) - \varphi(x_{n-2})| \leq L^2|x_{n-1} - x_{n-2}|$$

$$= \cdots \leq \cdots \leq L^n|x_1 - x_0|$$

Επομένως, για $k \in \mathbb{N}$

$$|x_{n+k} - x_n| = |x_{n+k} - x_{n+k-1} + x_{n+k-1} - x_{n+k-2} + \cdots + x_{n+1} - x_n|$$

$$\leq |x_{n+k} - x_{n+k-1}| + |x_{n+k-1} - x_{n+k-2}| + \cdots + |x_{n+1} - x_n|$$

$$\leq (L^{n+k-1} + L^{n+k-2} + \cdots + L^n)|x_1 - x_0| = L^n(1 + L + \cdots + L^{k-1})|x_1 - x_0| =$$

$$L^n\frac{1 - L^k}{1 - L}|x_1 - x_0| \leq \frac{L^n}{1 - L}|x_1 - x_0|.$$

Άρα $|x_{n+k} - x_n| \leq \dfrac{L^n}{1 - L}|x_1 - x_0|$ εφόσον $0 \leq L < 1$, ακολουθία $(x_n)_{n\in\mathbb{N}}$ είναι Cauchy και επομένως συγκλίνουσα με όριο $x^\star \in [a, b]$. Επιπλέον,

$$|x^\star - x_n| = \lim_{n\to\infty} x - n = \lim_{n\to\infty} \varphi(x_{n-1}) \underset{\text{λόγω συνέχειας}}{=} \varphi(\lim_{n\to\infty} x_{n-1}) = \varphi(x^\star)$$

1. Τώρα ορίζοντας τη συνεχή συνάρτηση $g(x) = |x - x_n|$ έχουμε

$$|x^\star - x_n| = g(x^\star) = g(\lim_{k\to\infty} x_{n+k}) = \lim_{k\to\infty} g(x_{n+k}) = \lim_{k\to\infty} |x_{n+k} - x_n| \leq \frac{L^n}{1 - L}|x_1 - x_0|$$

2. Επίσης

$$|x_n - x^\star| = |\varphi(x_{n-1}) - \varphi(x^\star)| \leq L|x_{n-1} - x^\star| = L|\varphi(x_{n-2}) - \varphi(x^\star)| \leq L^2|x_{n-2} - x^\star|$$

$$= \cdots \leq \cdots \leq L^n|x_0 - x^\star|$$

3. Τώρα θέτοντας $\psi_0 = x_{n-1}, \psi_1 = \varphi(x_{n-1}) = x_n$ και εφαρμόζοντας την εκτίμηση 1 με $n = 1$, παίρνουμε
$$|\psi_1 - x^\star| \leq \frac{L}{1 - L}|\psi_1 - \psi_0| \Rightarrow |x_n - x^\star| \leq \frac{L}{1 - L}|x_n - x_{n-1}|$$

Παρατηρήσεις

1. Ισχύει $\dfrac{L^n}{1 - L}|x_1 - x_0| \leq L^n|x_0 - x^\star|\dfrac{1 + L}{1 - L}$ (άσκηση), δηλαδή το φράγμα στην εκτίμηση 1 είναι το πολύ κατά $\dfrac{1 + L}{1 - L}$ μεγαλύτερο από το φράγμα στην εκτίμηση 2.

2. Ισχύει $\frac{L}{1-L}|x_n - x_{n-1}| \leq \frac{L^n}{1-L}|x_1 - x_0|$ (άσκηση), δηλαδή η εκτίμηση 3 είναι καλύτερη απο την 1.

3. Η εκτίμηση 1 είναι a priori ενώ η εκτίμηση 3 είναι a posteriori

Παρατήρηση: Έστω $\varepsilon_n = x_n - x^\star$. Λέμε ότι έχουμε σύγκλιση τάξεως p αν ισχύει $|\varepsilon_{n+1}| \leq C|\varepsilon_n|^p$ όπου $C > 0$ σταθερά. Αν $p = 1$, λέμε ότι έχουμε γραμμική σύγκλιση σε αυτήν την περίπτωση απαιτείται $0 < C < 1$.

Ποια είναι η τάξη σύγκλισης στη *γενική* επαναληπτική μέθοδο· (άσκηση)

σχήμα

$$x_1 = \varphi(x_0)$$
$$x_2 = \varphi(x_1)$$

Παράδειγμα

Έστω $f(x) = x^2 - 6x + 5$ με ρίζες $\rho_1 = 1$ και $\rho_2 = 5$. Λύνοντας ως προς $x = \frac{x^2+5}{6} = \varphi(x)$. Η φ είναι συστολή σε κάθε $[a,\, b] \subset (-3,\, 3)$.

- Άρα $\forall x_0 \in (-3,\, 3)$ η (x_n) συγκλίνει στο $x^\star = \rho_1 = 1$ (από θεώρημα).
- Για $x_0 = \pm 3 \Rightarrow x_1 = \frac{14}{6} = \frac{7}{3} < 3$ άρα $x_1,\, x_2,\, x_3 \in (-3,\, 3)$ οπότε $x_n \to x^\star = \rho_1 = 1$.
- Για $x_0 = 4 \Rightarrow x_1 = \frac{21}{6} > 3$, $x_2 = \frac{69}{24} < 3$ άρα $x_2,\, x_3,\, x_4 \in (-3,\, 3)$ οπότε $x_n \to x^\star = \rho_1 = 1$.
- Ομοίως για $x_0 \in (-5,\, 5) \Rightarrow x_n \to x^\star$, $n \to \infty$.
- Για $x_0 = \pm 5 \Rightarrow x_1 = x_2 = \cdots = 5$, οπότε $x_n \to 5$.
- Τι συμβαίνει για $x > 5$ ή $x < -5$· (άσκηση).

Υπενθύμιση $f(x) = 0 \Leftrightarrow x = \varphi(x)$
$x_{n+1} = \varphi(x_n)$, $x_n = 0,\, 1,\, 2, \cdots$
Δεδομένο x_0
$\varphi : [a,\, b] \to [a,\, b]$ συστολή, $x_0 \in [a,\, b]$ τότε η x_n είναι καλώς ορισμένη και $x_n \to x^\star =$ μοναδικό σταθερό σημείο τη φ.
Για γενικό (τυχαίο) φ:

$$|\varepsilon_{n+1}| = |x_{n+1} - x^\star| = |\varphi(x_n) - \varphi(x^\star)| \leq L|x_n - x^\star| = L|\varepsilon_n|\, 0 \leq L < 1$$

άρα στη γενική επαναληπτική μέθοδο η σύγκλιση είναι γραμμική ($p = 1$).

Η Μέθοδος του Νεύτωνα

Έστω x_n μία προσέγγιση της ρίζας $x^\star$ της $f(x)=0$. Αν η f είναι δύο φορές παραγωγίσιμη σε μία περιοχή U και $x,\, x_n \in U$.

Τύπος του Taylor.

$$f(x) = f(x_n) + (x-x_n)f'(x_n) + \frac{1}{2}(x-x_n)^2 f''(\xi)$$

όπου ξ μεταξύ των $x,\, x_n$. Όπου x βάζω το $x^\star$,

$$0 = f(x^\star) = f(x_n) + (x^\star - x_n)f'(x_n) + \frac{1}{2}(x^\star - x_n)^2 f''(\xi).$$

Υποθέτοντας ότι το x_n είναι "αρκετά κοντά" στο $x^\star$ τότε $f'(x_n) \neq 0$ και

$f(x_n) + (x^\star - x_n)f'(x_n) \cong 0 \Rightarrow x^\star \cong x_n - \dfrac{f(x_n)}{f'(x_n)}$. Δηλαδή αν το x_n είναι μία προσέγγιση του $x^\star$ τότε το $x_n - \dfrac{f(x_n)}{f'(x_n)}$ είναι καλύτερη προσέγγιση.

Θέτουμε $x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_n)}$, $n = 0,\, 1,\, 2, \cdots \rightarrow$ μέθοδος του Νεύτωνα. Παίρνω $\varphi(x) = x - \dfrac{f(x)}{f'(x)}$ άρα ισχύει $x_{n+1} = \varphi(x_n) = x_n - \dfrac{f(x_n)}{f'(x_n)}$.

σχήμα

$$f'(x_n) = \tan\omega = \frac{f(x_n)}{x_n - x_n^\star}$$
$$\Rightarrow x_n^\star = x_n - \frac{f(x_n)}{f'(x_n)} \Rightarrow x_{n+1} = x_n^\star$$

Παραδείγματα:

(α) σχήμα

(β) $x^2 - 6x + 5 = 0$ με ρίζες $\rho_1 = 1$ και $\rho_2 = 5$. σχήμα

- Γενική επαναληπτική μέθοδος

$x_{n+1} = \dfrac{x_n^2 + 5}{6}$

$x_0 = 2$

$x_1 = 1,5$

$x_2 = 1,208\bar{3}$

$x_3 = 1,076678$

$\cdots$

$x_7 = 1,001002$

$x_8 = 1,000334$

άρα $\varepsilon_8 = |x_8 - x^\star| = 3,34 \cdot 10^{-4}$.

Η γενική επαναληπτική μέθοδος δεν μπορεί να συγκλίνει στη ρίζα $\rho_2 = 5$, εκτός αν $x_0 = 5$.

- Μέθοδος του Νεύτωνα

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^2 - 6x_n + 5}{2x_n - 6} = \frac{x_n^2 - 5}{2x_n - 6},\ n = 0,\, 1,\, 2, \cdots$$

$x_0 = 2$
$x_1 = 0,5$
$x_2 = 0,95$
$x_3 = 0,9993902$
$x_4 = 0,999999907$
Άρα $|\varepsilon_3| = 6,1 \cdot 10^{-4}$ και $\varepsilon_4 = 9,3 \cdot 10^{-8}$.
Στη συγκεκριμένη μέθοδο αν πάρω $x_0 > 3$ συγκλίνω στη ρίζα 5. Δεν υπάρχει πρόβλημα με την επιλογή του x_0:
έχουμε σύγκλιση στην πρώτη ρίζα για $x_0 < 3$,
έχουμε σύγκλιση στη δεύτερη ρίζα για $x_0 > 3$. Πρόβλημα υπάρχει στο $x_0 = 3$ γιατί $f'(x) = 2x - 6 = 0 \Rightarrow x = 3$.

Στη μέθοδο του Νεύτωνα είναι σημαντικό η αρχική τιμή x_0 να είναι "καλή" (δηλαδή σχετικά κοντά στη ρίζα. Αυτό μπορεί να ισχύει πλαντα απλά η δεύτερη παράγωγος στον παρανομαστή μπορεί να δημιουργήσει πρόβλημα).
Παράδειγμα:
σχήμα
Πιθανά Κριτήρια Τερματισμού της μεθόδου του Νεύτωνα

(α) $|f(x_k)| < \varepsilon$

(β) $|x_k - x_{k-1}| < \varepsilon$

(γ) $|\dfrac{x_k - x_{k-1}}{x_k}| < \varepsilon$ όπου ε είναι "δεδομένη ανοχή".

Όλα αυτά τα κριτήρια δεν είναι ασφαλή και υπάρχουν περιπτώσεις που η μέθοδος του Νεύτωνα μπορεί να τα ξεγελάσει.
Παραδείγματα:

(α) σχήμα $|f(x_k)| < \varepsilon$ αλλά δεν είμαστε κοντά σε ρίζα

(α) σχήμα $|x_k - x_{k-1}| < \varepsilon$ αλλά δεν είμαστε κοντά σε ρίζα

Γιάυτό:
– Βάζουμε παραπάνω από ένα κριτήρια.
– Βάζουμε έναν μετρητή βημάτων (counter)

$$N \leftarrow N + 1$$

Αν $N > N_{max}$ σταματάμε και ελέγχουμε

Σύγκλιση της μεθόδου του Νεύτωνα.

Θεώρημα (Τυπικά τετραγωνική σύγκλιση της μεθόδου).
Έστω $x^\star$ απλή ρίζα μίας συνάρτησης f, δηλαδή $f(x^\star)=0$, $f'(x^\star)\neq 0$, και έστω ότι η f είναι δύο φορές συνεχώς παραγωγίσιμη σε μια περιοχή του $x^\star$. Τότε υπάρχει κλειστό διάστημα I με μέσον το $x^\star$, τέτοιο ώστε $\forall x_0\in I$ η ακολουθία $(x_n)_{n\in\mathbb{N}}$ που σχηματίζεται με τη μέθοδο του Νεύτωνα $x_{n+1}=x_n-\dfrac{f(x_n)}{f'(x_n)},\ n=0,\,1,\,2,\cdots$ συγκλίνει στο $x^\star$ και μάλιστα ισχύει

$$\lim_{n\to\infty}\frac{\varepsilon_{n+1}}{\varepsilon_n^2}=\lim_{n\to\infty}\frac{x_{n+1}-x^\star}{(x_n-x^\star)^2}=\frac{f''(x^\star)}{2f'(x^\star)}.$$

Απόδειξη.
Έστω $\varphi(x)=x-\dfrac{f(x)}{f'(x)}$. Τότε $\varphi(x^\star)=x^\star$ και

$$\varphi'(x)=\cdots=\frac{f(x)\cdot f''(x)}{[f'(x)]^2}\Rightarrow\varphi'(x^\star)=0.$$

Έχουμε $\varphi'(x^\star)=0$ λόγω του ότι η φ' είναι συνεχής σε μία περιοχή του $x^\star$ υπάρχει ένα κλειστό διάστημα, με μέσον το $x^\star$, τέτοιο ώστε $\max\limits_{x\in I}|\varphi'(x)|=L<1\Rightarrow\varphi$ συστολή στο I με σταθερά L.
Επιπλέον, για $x\in I$

$$|\varphi(x)-x^\star|=|\varphi(x)-\varphi(x^\star)|\leq L|x-x^\star|<|x-x^\star|\Rightarrow\varphi(x)\in I\Rightarrow\varphi(I)\subset I$$

Άρα από το Θεώρημα της Συστολής, $\forall x_0\in I$ η $(x_n)_{n\in\mathbb{N}_0}$ με $x_{n+1}=x_n-\dfrac{f(x_n)}{f'(x_n)}$, $n=0,\,1,\,2,\cdots$ συγκλίνει στο $x^\star$.
Από το ανάπτυγμα του Taylor έχουμε

$$f(x_n)=f(x^\star)+(x_n-x^\star)f'(x^\star)+\frac{(x_n-x^\star)^2}{2}f''(\xi_{n_1})$$

$$f'(x_n)=f'(x^\star)+(x_n-x^\star)f''(\xi_{n_2})$$

όπου ξ_{n_1} και ξ_{n_2} μεταξύ των x_n και $x^\star$. Έτσι $x_{n+1}=x_n-\dfrac{f(x_n)}{f'(x_n)}$

$$\Rightarrow x_{n+1}-x^\star=x_n-x^\star-\frac{(x_n-x^\star)f'(x_n)+\frac{(x_n-x^\star)^2}{2}f''(\xi_{n_1})}{f'(x^\star)+(x_n-x^\star)f''(\xi_{n_2})}$$

$$\Rightarrow\cdots x_{n+1}-x^\star=(x_n-x^\star)^2\frac{f''(\xi_{n_2})-\frac{1}{2}f''(\xi_{n_1})}{f'(x^\star)+(x_n-x^\star)f''(\xi_{n_2})}$$

Άρα

$$\lim_{n\to\infty}\frac{\varepsilon_{n+1}}{\varepsilon_n^2}=\lim_{n\to\infty}\frac{(x_{n+1}-x^\star)}{(x_n-x^\star)^2}=\lim_{n\to\infty}\frac{f''(\xi_{n_2})-\frac{1}{2}f''(\xi_{n_1})}{f'(x^\star)+(x_n-x^\star)f''(\xi_{n_2})}=\frac{f''(x^\star)}{2f'(x^\star)}.$$

Παρατηρήσεις

(α) Η σύγκλιση είναι τουλάχιστον τετραγωνική ($p \geq 2$).
Αν $f''(x^\star) \neq 0$ είναι ακριβώς τετραγωνική.

(β) Στη γενική επαναληπτική μέθοδο έχουμε

$$\varepsilon_{n+1} = x_{n+1} - x^\star = \varphi(x_n) - \varphi(x^\star) = \varphi'(\xi_n)(x_n - x^\star) = \varphi'(\xi_n) \cdot \varepsilon_n$$

όπου ξ_n μεταξύ x_n, $x^\star$. Άρά,

$$\lim_{n\to\infty} \frac{\varepsilon_{n+1}}{\varepsilon_n} = \lim_{n\to\infty} \varphi'(\xi_n) = \varphi'(x^\star).$$

Αν $\varphi'(x^\star) = 0$ τότε μπορώ να πετύχω καλύτερη (μεγαλύτερη) σύγκλιση από την γραμμική.

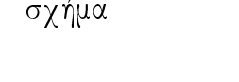
(γ) Το $\lim_{n\to\infty} \frac{\varepsilon_{n+1}}{\varepsilon_n^2} = c^\star = \frac{f''(x^\star)}{2f'(x^\star)}$ σημαίνει ότι για $n >> 1$, $\varepsilon_{n+1} \sim c^\star \varepsilon_n^2$, τάξη σύγκλισης 2.
Βέβαια, $\frac{|\varepsilon_{n+1}|}{|\varepsilon_n^2|} \leq A \Leftrightarrow |\varepsilon_{n+1}| \leq A|\varepsilon_n^2|$, $n = 0, 1, 2, \cdots$
(οι δύο τρόποι ορισμού της τάξης σύγκλισης είναι ισοδύναμοι; (άσκηση)).

(δ) Μειονεκτήματα του θεωρήματος

- Δεν λέει πως να βρεις το I.
- Το I μπορεί να είναι μικρό.

(ε) Υπάρχουν περιπτώσεις που έχω σύγκλιση για x_0 μακρυά από τη ρίζα.
Παράδειγμα $f(x) = x^3 - 2x - 5 = 0$

$(\sqrt{\frac{2}{3}}, -\frac{45 + 4\sqrt{6}}{9})$ τοπικό ελάχιστο
$(-\sqrt{\frac{2}{3}}, -\frac{45 - 4\sqrt{6}}{9})$ τοπικό μέγιστο


σχήμα

$\forall x_0 > \sqrt{\frac{2}{3}}$ έχω πάντα σύγκλιση στη μοναδική ρίζα $x^\star$.

Πρόταση
Έστω ότι η f είναι δύο φορές συνεχώς παραγωγίσιμη και ότι $f'(x) > 0$, $f''(x) > 0$ για $x \geq 0$. Έστω $f(a) < 0$. Τότε η f έχει ακριβώς μία πραγματική ρίζα στο διάστημα $[a, \infty)$. Για οποιοδήποτε $x_0 \geq a$ η ακολουθία $(x_n)_{n\in\mathbb{N}_0}$ που παράγει η μέθοδος του Νεύτωνα συγκλίνει στη p.

Πολλαπλή Ρίζα

Παράδειγμα

σχήμα

$f(x) = x^2 = 0$

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^2}{2x_n} = \frac{x_n}{2},\ n = 0,\ 1,\ 2\cdots$$

Δεδομένο x_0

$$\varepsilon_{n+1} = x_{n+1} - x^\star = x_{n+1} = \frac{x_n}{2} = \frac{1}{2}(x_n - x^\star) = \frac{1}{2}\varepsilon_n \Rightarrow \lim_{n\to\infty} \frac{\varepsilon_{n+1}}{\varepsilon_n} = \frac{1}{2}\,(\text{γραμμική σύγκλιση}).$$

Γενικά αποδεικνύεται ότι:

Αν $x^\star$ είναι ρίζα πολλαπλότητας $m (f(x^\star) = f'(x^\star) = \cdots = f^{m-1}(x^\star) = 0)$ αλλά $f^m(x^\star) \neq 0$, τότε

$$\lim_{n\to\infty} \frac{\varepsilon_{n+1}}{\varepsilon_n} = 1 - \frac{1}{m}.$$

Προκειμένου να έχουμε τετραγωνική σύγκλιση χρησιμοποιούμε την παραλλαγή της μεθόδου του Νεύτωνα

$$x_{n+1} = x_n - m\frac{f(x_n)}{f'(x_n)}.$$

Μέθοδος της Τέμνουσας.

Στη μέθοδο της τέμνουσας

$$f'(x_n) \cong \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

οπότε το σχήμα της μεθόδου του Νεύτωνα είναι

$$x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})},\ n = 1,\ 2, \cdots$$

για δεδομένα x_0, x_1.

σχήμα

Η μέθοδος της τέμνουσας είναι λίγο πιο σταθερή και λίγο πιο αργή από τη μέθοδο του Νεύτωνα. Συγκεκριμένα αποδεικνύεται ότι

$$\lim_{n\to\infty} \frac{\varepsilon_{n+1}}{\varepsilon_n^p} = C \neq 0,\ p = \frac{1+\sqrt{5}}{2} \cong 1,62.$$

Ασκήσεις στην επίλυση ΜΗ ΓΡΑΜΜΙΚΩΝ ΕΞΙΣΩΣΕΩΝ, (βιβλίο Ακριβή-Δουγαλή).

- **άσκηση παρόμοια της 2.1**

  Για την εξίσωση $f(x) = x^3 - 3x - 4 = 0$ να υπολογιστεί η δεύτερη προσέγγιση x_2 της ρίζας $x^\star$ που δίνει η μέθοδος διχοτόμησης. Πόσο το πολύ απέχει η x_2 από τη ρίζα· Πόσα βήματα απαιτούνται για τον υπολογισμό της προσέγγισης που απέχει 10^{-6} το πολύ από την $x^\star$·

  Λύση.

$$f'(x) = 3x^2 - 3 = 3(x^2 - 1) = 0 \Rightarrow x = \pm 1$$

$$f''(x) = 6x = 0 \Rightarrow x = 0$$

$$f(-1) = -2,\ f(0) = -4,\ f(1) = 6,\ f(2) = -2,\ f(3) = 14.$$

x	$-\infty$		-1		0		1		$+\infty$
$f'(x)$		$+$	0	$-$		$-$	0	$+$	
$f''(x)$		$-$		$-$	0	$+$		$+$	
$f(x)$		$\nearrow$	τ.μ.	$\searrow$		$\searrow$	τ.ε.	$\nearrow$	

$[a_1,\, b_1] = [2,\, 3]$
$x_1 = \dfrac{a_1 + b_1}{2} = \dfrac{2+3}{2} = 2,5$
$f(x_1) = 4,125 > 0$ σχήμα
$[a_2,\, b_2] = [2,\, 2,5]$
$x_2 = \dfrac{a_2 + b_2}{2} = \dfrac{2+2,5}{2} = 2,25$
$|x_2 - x^\star| \leq |x_1 - a_1| = \dfrac{1}{2}|2,5 - 2| = \dfrac{1}{4}$ αλλιώς από τον τύπο $|x_n - x^\star| \leq \dfrac{b_1 - a_1}{2^n}$ βρίσκουμε το ίδιο. Για σφάλμα το πολύ 10^{-6} χρησιμοποιώ πάλι τον τύπο

$$|x_n - x^\star| \leq \frac{b-a}{2^n} = \frac{1}{2^n} \leq 10^{-6} \Rightarrow \frac{1}{2^n} \leq 10^{-6} \Rightarrow 2^n \geq 10^6 \Rightarrow n \ln 2 \geq 6 \ln 10$$

$$\Rightarrow n \geq \frac{6 \ln 10}{\ln 2} \cong 19,93 \Rightarrow n = 20.$$

- **άσκηση 2.9**

  Έστω $x_0 \in [0,\, 1]$. Αποδείξτε ότι η ακολουθία $(x_n)_{n \in \mathbb{N}_0}$ με $x_{n+1} = \dfrac{1}{3}(2 + x_n - e^{x_n})$, $n \in \mathbb{N}_0$ συγκλίνει και το όριο της βρίσκεται στο $[0,\, 1]$.

  Λύση.

Θεωρούμε την $\varphi(x) = \frac{1}{3}(2 + x - e^x)$, οπότε $x_{n+1} = \varphi(x_n)$, $n = 0, 1, 2, \cdots$. Αρκεί να δείξουμε ότι

$$\varphi'(x) = \frac{1}{3}(1 - e^x) \leq 0 \text{ για } x \in [0, 1] \Rightarrow \varphi \text{ φθίνουσα} \Rightarrow \varphi([0, 1]) = [\varphi(1), \varphi(0)]$$

$$= [\frac{3-e}{3}, \frac{1}{3}] \subset [0, 1] \Rightarrow |\varphi'(x)| = |\frac{1}{3}(1 - e^x)| = \frac{e^x - 1}{3} \leq \frac{e-1}{3}$$

$$\Rightarrow \text{η } \varphi \text{ συστολή στο } [0, 1] \text{ με σταθερά } L = \frac{e-1}{3} < 1.$$

Συνεπώς, από το Θεώρημα της Συστολής η ακολουθία (x_n) συγκλίνει σε ένα σημείο $x^\star \in [0, 1],\ x^\star = \varphi(x^\star)$.

- **άσκηση παρόμοια της 2.20**

Δείξτε ότι η εξίσωση $f(x) = x^3 - 3x - 4 = 0$ έχει μόνο μία πραγματική ρίζα ρ. Δείξτε ότι $\rho \in [2, 3]$ και ότι αν $x_0 > 1$, η ακολουθία x_n που παράγει η μέθοδος του Νεύτωνα συγκλίνει στη ρ.

Τι μπορεί να συμβεί αν $x_0 \leq 1$·

Λύση.

Για $x > 1$, $f'(x) > 0$ και $f''(x) > 0$. Επίσης,

$$f(x) > 0 \Leftrightarrow x > \rho$$

$$f(x) < 0 \Leftrightarrow x < \rho$$

Μέθοδος του Νεύτωνα

$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ σχήμα

π.χ. για $x_0 = 3 \Rightarrow x_1 \cong 2,417$

Περίπτωση Α: Ας υποθέσουμε ότι $1 < x_0 < \rho$. Τότε $x_1 > \rho$. Από τον τύπο της μεθόδου του Νεύτωνα

$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$ όπου $f(x_0) < 0$ και $f'(x_0) > 0$ άρα $\frac{f(x_0)}{f'(x_0)} < 0 \Rightarrow$

$x_0 - \frac{f(x_0)}{f'(x_0)} > x - 0 \Rightarrow x - 1 > x_0 > 1.$

Από τον τύπο του Taylor

$f(x_1) = \underbrace{f(x_0) + (x_1 - x_0)f'(x_0)}_{0} + \frac{(x_1 - x_0)^2}{2} f''(\xi)$, ξ μεταξύ x_0, x_1

$\Rightarrow f(x_1) = \frac{(x_1 - x_0)^2}{2} f''(\xi)$ όμως $\xi > 1 \Rightarrow f''(\xi) > 0 \Rightarrow f(x_1) > 0 \Rightarrow x_1 > \rho.$

Περίπτωση Β: Ας υποθέσουμε ότι $x_0 > \rho$. Θα δείξουμε ότι $x_n > \rho$, $n = 1, 2, \cdots$ και ότι $x_{n+1} < x_n$, δηλαδή η x_n είναι φθίνουσα. Έστω $x_n > \rho$.
Από τον τύπο του Taylor

$$f(x_{n+1}) = \underbrace{f(x_n) + (x_{n+1} - x_n)f'(x_n)}_{0} + \frac{(x_{n+1} - x_n)^2}{2} f''(\xi_n),\ \xi_n \text{ μεταξύ } x_n,\ x_{n+1}$$

$$\Rightarrow f(x_{n+1}) = \frac{(x_{n+1} - x_n)^2}{2} f''(\xi_n).$$

Από τον τύπο της μεθόδου του Νεύτωνα

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^3 - 3x_n - 4}{3x_n^2 - 3} = \frac{2x_n^3 + 4}{3x_n^2 - 3}$$

Αν

$$x_n > \rho > 1 \Rightarrow x_{n+1} > 0 \Rightarrow \xi_n > 0 \Rightarrow f''(\xi_n) > 0$$

οπότε

$$f(x_{n+1}) > 0 \Rightarrow x_{n+1} > \rho$$

Επίσης, $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ όπου $f(x_n) > 0$ και $f'(x_n) > 0 \Rightarrow x_n - \frac{f(x_n)}{f'(x_n)} < x_n$
$x_{n+1} < x - n \Rightarrow x_n$ φθίνουσα.
Συνεπώς, η x_n είναι συγκλίνουσα και έστω $\lim_{n\to\infty} x_n = \psi \geq \rho$ (θέλω να δείξω ότι $\psi = \rho$).
Έχουμε $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$, η f, f' συνεχείς $\Rightarrow \lim x_{n+1} = \lim x_n - \frac{f(\lim x_n)}{f'(\lim x_n)}$
$\Rightarrow \psi = \psi - \frac{f(\psi)}{f'(\psi)}$ όπου $f'(\psi) > 0$ γιατί $\psi \geq \rho > 1 > 0 \to f(\psi) = 0 \overset{\text{μονότονη}}{\Rightarrow} \psi = \rho$.
Ακόμη, για $x_0 \leq 1$ δεν μπορούμε να εγγυηθούμε τη σύγλιση εφόσον για κάποιο x_n μπορεί $f'(x_n) = 0$.

- **άσκηση 2.19**

  α) Μέθοδος του Νεύτωνα για τον υπολογισμό τετραγωνικής ρίζας.
  Γράψτε τη μέθοδο του Νεύτωνα για την προσέγγιση της θετικής ρίζας της εξίσωσης $f(x) = x^2 - a = 0$ όπου $a > 0$, δηλαδή του αριθμού $\sqrt{a}$, και αποδειξτέ ότι η ακολουθία (x_n), των προσεγγίσεων συγκλίνει στην $\sqrt{a}$ για κάθε $x_0 > 0$.

  Λύση.
  $x_{n+1} = \frac{1}{2}(x_n + \frac{a}{x_n})$
  Ισχύει η πρόταση του βιβλίου· σχήμα

Προτεινόμενες Ασκήσεις:
$1.2 - 1.6, 1.10, 1.12,$
$2.1, 2.8 - 2.10, 2.16, 2.19 - 2.23.$
σχήματα.

Γραμμικά Συστήματα

Γενικά περί γραμμικών συστημάτων
Θεωρούμε το γραμμικό σύστημα:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \vdots + \vdots + \cdots + \vdots &= \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned}$$

ή σε μορφή πινάκων

$$\mathbf{A} = (a_{ij})_{i,\,j=1,\,2,\,\cdots,n} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}, \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}, \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

$\mathbf{Ax} = \mathbf{b}$ όπου οι συντελεστές a_{ij} και τα δεύτερα μέλη b_i είναι δεδομένοι αριθμοί. Το ζητούμενο είναι τα x_i.
Το σύστημα έχει ακριβώς μία λύση αν και μόνον αν ισχύει μία από τις παρακάτω συνθήκες

1. ο πίνακας $\mathbf{A}$ είναι αντιστρέψιμος, δηλαδή υπάρχει ο $\mathbf{A}^{-1}$,
2. $\det \mathbf{A} \neq 0$,
3. το σύστημα $\mathbf{Ax} = \mathbf{0}$ έχει ως μοναδική λύση την τετριμμένη $\mathbf{x} = \mathbf{0}$,
4. οι στήλες ή οι γραμμές του $\mathbf{A}$ είναι γραμμικά ανεξάρτητες

Τρόποι υπολογισμού της λύσης (από γραμμική άλγεβρα):

1. Κανόνας του Cramer:
   Αν $\mathbf{A} = (a^1, a^2, \cdots, a^n)$, a^i, $i = 1, 2, \cdots, n$ οι στήλες του, ορίζω
   $A_i = (a^1, a^2, \cdots, a^{i-1}, b, a^{i+1}, \cdots, a^n)$ τότε

$$x_i = \frac{\det A_i}{\det \mathbf{A}},\ i = i = 1, 2, \cdots, n.$$

   Αν για τον υπολογισμό της $\det \mathbf{A}$ αναπτύξουμε ως προς τη στήλη j

$$\det \mathbf{A} = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det A_{ij},$$

   όπου A_{ij} ο πίνακας που προκύπτει από τον $\mathbf{A}$ δια διαγραφής της i γραμμής και j στήλης και συνεχίζουμε κατάυτόν τον τρόπο μέχρι να καταλήξουμε σε ορίζουσες 2×2. Τότε απαιτούνται περίπου $n!(n-1)$ πολλαπλασιασμοί, π.χ. για $n = 50$ απαιτούνται περίπου $1,49 \cdot 10^{66}$ πολλαπλασιασμοί,

2. Προσδιορισμός του αντιστρόφου $\mathbf{A}^{-1}$, οπότε $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.
   Αυτό γίνεται λύνοντας n γραμμικά συστήματα $\mathbf{Ax} = \mathbf{e}^i$, $i = 1, 2, \cdots, n$ όπου $\{\mathbf{e}^1, \mathbf{e}^2, \cdots, \mathbf{e}^n\}$ η κανονική βάση του $\mathbb{R}^n$. Αν $\mathbf{u}^1, \mathbf{u}^2, \cdots, \mathbf{u}^n$ είναι οι λύσεις αυτών των συστημάτων τότε $\mathbf{A}^{-1} = (\mathbf{u}^1, \mathbf{u}^2, \cdots, \mathbf{u}^n)$.
   Απόδειξη.

$$\mathbf{A} \cdot (\mathbf{u}^1, \mathbf{u}^2, \cdots, \mathbf{u}^n) = (\mathbf{Au}^1, \mathbf{Au}^2, \cdots, \mathbf{Au}^n) = (\mathbf{e}^1, \mathbf{e}^2, \cdots, \mathbf{e}^n) = \mathbf{I}$$

Αριθμητικές Μέθοδοι λύσεως γραμμικών συστημάτων:

(i) Άμεσες μέθοδοι

(ii) Επαναληπτικές μέθοδοι

Τριγωνικά Συστήματα.
Θεωρούμε τον άνω τριγωνικό σύστημα

$$\begin{aligned} u_{11}x_1 + u_{12}x_2 + \cdots + u_{1n}x_n &= b_1 \\ u_{22}x_2 + \cdots + u_{2n}x_n &= b_2 \\ \cdots + \cdots + \cdots + \cdots &= \cdots \\ u_{nn}x_n &= b_n \end{aligned}$$

ή

$$\begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ 0 & 0 & \cdots & \vdots \\ 0 & 0 & 0 & u_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

όπου $u_{ii} \neq 0$, $i = 1, 2, \cdots, n$. Δηλαδή το σύστημα έχει ακριβώς μία λύση η οποία υπολογίζεται με "οπισθοδρόμηση",

$$x_n = \frac{b_n}{u_{nn}}$$

$$x_{n-1} = \frac{b_{n-1} - u_{n-1n}x_n}{u_{n-1n-1}}$$

$$\vdots$$

Γενικά,

$$x_i = \frac{b_i - \sum_{j=i+1}^{n} u_{ij}x_j}{u_{ii}},\ i = n-1, n-2, \cdots, 2, 1.$$

Η λύση αυτή απαιτεί:

- Προσθέσεις: $1 + 2 + \cdots + (n-1) = \dfrac{n(n-1)}{2} = \dfrac{n^2}{2} - \dfrac{n}{2} \cong \dfrac{n^2}{2}$,
- Πολλαπλασιασμοί: $1 + 2 + \cdots + (n-1) = \dfrac{n(n-1)}{2} = \dfrac{n^2}{2} - \dfrac{n}{2} \cong \dfrac{n^2}{2}$,
- Διαιρέσεις: n,
- Θέσεις μνήμης:
  1. για τον πίνακα: $n + (n-1) + \cdots + 1 = \dfrac{n(n+1)}{2} = \dfrac{n^2}{2} + \dfrac{n}{2}$,
  2. για το διάνυσμα $\mathbf{b}$: n,
  3. συνολικά: $\dfrac{n^2}{2} + \dfrac{3n}{2}$

  Το $\mathbf{x}$ αποθηκεύεται στη θέση του $\mathbf{b}$.

Αντίστοιχα λύνονται τα κάτω τριγωνικά συστήματα.

Μέθοδος Απαλοιφής του Gauss
Η μέθοδος περιλαμβάνει δύο στάδια:

- Τριγωνοποίηση: σχήμα

– Οπισθοδρόμηση: Λύση του άνω τριγωνικού συστήματος (όπως πριν).

Έστω ότι το σύστημα $\mathbf{Ax} = \mathbf{b}$ έχει ακριβώς μία λύση.
Τριγωνοποίηση
Θέτουμε $\mathbf{A}^{(1)} = \mathbf{A}$, $\mathbf{b}^{(1)} = \mathbf{b}$ οπότε το σύστημα γράφεται $\mathbf{A}^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$

όπου $\mathbf{A}^{(1)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1}^{(1)} & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{pmatrix}$ και $\mathbf{b}^{(1)} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(1)} \\ \vdots \\ b_n^{(1)} \end{pmatrix}$.

1^o βήμα.
Έστω $a_{11}^{(1)} \neq 0$ διαφορετικά με κατάλληλη αλλαγή γραμών βρίσκουμε πάντα $a_{11}^{(1)} \neq 0$ αφού $|\mathbf{A}| \neq 0$.
Ορίζουμε τους πολλαπλασιαστές

$$u_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}},\ i = 2,\ 3,\ \cdots,\ n$$

Πολλαπλασιάζουμε την πρώτη εξίσωση με u_{i1} και αφαιρούμε από την i-οστή εξίσωση για $i = 2,\ 3,\ \cdots,\ n$. Έτσι παίρνουμε το ισοδύναμο σύστημα

$$\mathbf{A}^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$$

όπου

$$\mathbf{A}^{(2)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & \cdots & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{pmatrix},\ \mathbf{b}^{(1)} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_n^{(2)} \end{pmatrix}$$

με

$$\begin{aligned} a_{ij}^{(2)} &= a_{ij}^{(1)} - u_{i1}a_{ij}^{(1)} \text{ για } i,\ j = 2,\ 3,\ \cdots,\ n \\ b_i^{(2)} &= b_i^{(1)} - u_{i1}b_i^{(1)}. \end{aligned}$$

2^o βήμα.

$$\begin{pmatrix} \star & \star & \cdots & \star \\ 0 & \star & \cdots & \star \\ \vdots & \vdots & \cdots & \vdots \\ 0 & \star & \cdots & \star \end{pmatrix} \rightarrow \begin{pmatrix} \star & \star & \cdots & \star & \\ 0 & \star & \cdots & \star & \\ 0 & 0 & \star & \cdots & \star \\ \vdots & \vdots & \cdots & \vdots & \\ 0 & 0 & \star & \cdots & \star \end{pmatrix}$$

k^o βήμα.
Έχουμε το ισοδύναμο σύστημα $\mathbf{A}^{(k)}\mathbf{x} = \mathbf{b}^{(k)}$, όπου

$$\mathbf{A}^{(k)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & \cdots & \cdots & & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & \cdots & \cdots & & a_{2n}^{(2)} \\ 0 & 0 & & & & & \\ & & \cdots & a_{k-1k-1}^{(k-1)} & & & \\ \vdots & \vdots & \cdots & 0 & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ & & & & & & \\ 0 & 0 & \cdots & 0 & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix}$$

και

$$\mathbf{b}^{(k)} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_{k-1}^{(k-1)} \\ b_k^{(k)} \\ \vdots \\ b_n^{(k)} \end{pmatrix}$$

Έστω $a_{kk}^{(k)} \neq 0$. Ορίζουμε τους πολλαπλασιαστές

$$m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}},\ i = k+1,\ k+2,\ \cdots,\ n.$$

Πολλαπλασιάζουμε την k-οστή εξίσωση με m_{ik} και αφαιρούμε από την i-οστή εξίσωση για $i = k+1,\ k+2,\ \cdots,\ n$. Έτσι παίρνουμε το ισοδύναμο σύστημα $\mathbf{A}^{(k+1)}\mathbf{x} = \mathbf{b}^{(k+1)}$, όπου

$$\mathbf{A}^{(k+1)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & \cdots & \cdots & & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & \cdots & \cdots & & a_{2n}^{(2)} \\ 0 & 0 & & & & & \\ & & \cdots & a_{k-1k-1}^{(k-1)} & & & \\ & & & 0 & a_{kk}^{(k)} & & \\ \vdots & \vdots & \cdots & 0 & a_{k+1k+1}^{(k+1)} & \cdots & a_{k+1n}^{(k+1)} \\ & & & & & & \\ 0 & 0 & \cdots & 0 & a_{nk+1}^{(k+1)} & \cdots & a_{nn}^{(k+1)} \end{pmatrix}$$

και

$$\mathbf{b}^{(k+1)} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_{k-1}^{(k-1)} \\ b_k^{(k)} \\ \vdots \\ b_n^{(k+1)} \end{pmatrix}$$

με

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)} \text{ και } b_i^{(k+1)} = b_i^{(k)} - m_{ik}b_i^{(k)} \text{ όπου } i, j = k+1,\, k+2,\, \cdots,\, n$$

και έχω συνολικά $n-1$ βήματα.

Η μέθοδος απαλοιφής του Gauss (συνέχεια)
$(n-1)^o$ βήμα.
Παίρνουμε το ισοδύναμο σύστημα $\mathbf{A}^{(n)}\mathbf{x} = \mathbf{b}^{(n)}$, όπου

$$\mathbf{A}^{(n)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & \cdots & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & \cdots & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & \cdots & \vdots & & \\ 0 & 0 & \cdots & a_{nn}^{(n)} & & \end{pmatrix},\ \mathbf{b}^{(n)} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_n^{(n)} \end{pmatrix}$$

Οπισθοδρόμηση
Αν στο τελικό σύστημα $\mathbf{A}^{(n)}\mathbf{x} = \mathbf{b}^{(n)}$ θέσουμε

$$\mathbf{A}^{(n)} = \begin{pmatrix} a_{11}^{(n)} & a_{12}^{(n)} & \cdots & \cdots & \cdots & a_{1n}^{(n)} \\ 0 & a_{22}^{(n)} & \cdots & \cdots & \cdots & a_{2n}^{(n)} \\ \vdots & \vdots & \cdots & \vdots & & \\ 0 & 0 & \cdots & a_{nn}^{(n)} & & \end{pmatrix},\ \mathbf{b}^{(n)} = \begin{pmatrix} b_1^{(n)} \\ b_2^{(n)} \\ \vdots \\ b_n^{(n)} \end{pmatrix}$$

έχουμε

$$x_n = \frac{b_n^{(n)}}{a_{nn}^{(n)}}$$
$$x_i = \frac{b_i^{(n)} - \sum_{j=i+1}^{n} a_{ij}^{(n)} x_j}{a_{ii}^{(n)}},\ i = n-1,\, n-2,\, \cdots,\, 1.$$

Παρατήρηση:
Η τριγωνοποίηση είναι δυνατή και στην περίπτωση που ο πίνακας $\mathbf{A}$ είναι μη αντιστρέψιμος (πιθανώς

με κατάλληλες εναλλαγές γραμμών). Βεβαίως στον τελικό πίνακα $\mathbf{A}^{(n)}$ κάποιο/α από τα διαγώνια στοιχεία θα είναι μηδέν.
Παράδειγμα:
Να λυθεί με τη μέθοδο απαλοιφής του Gauss το γραμμικό σύστημα

$$\begin{aligned} x_2 + 2x_3 &= 3 \\ 2x_1 - 2x_2 + x_3 &= 6 \\ 5x_1 + 3x_2 + x_3 &= 4 \end{aligned}$$

Λύση:

$$\begin{aligned} & \begin{aligned} x_2 + 2x_3 &= 3 \\ 2x_1 - 2x_2 + x_3 &= 6 \\ 5x_1 + 3x_2 + x_3 &= 4 \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} x_2 + 2x_3 &= 3 \\ 2x_1 - 2x_2 + x_3 &= 6 \\ 8x_2 - \frac{3}{2}x_3 &= 11 \end{aligned} \\ \Leftrightarrow & \begin{aligned} 2x_1 - 2x_2 + x_3 &= 6 \\ x_2 + 2x_3 &= 3 \\ -\frac{35}{2}x_3 &= 35 \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} x_3 &= 2 \\ x_2 &= 3 - 2x_3 = -1 \\ x_1 &= \frac{6 + 2x_2 - x_3}{2} = 1 \end{aligned} \end{aligned}$$

Τελικά προκύπτει ότι,

$$x_1 = 1,\ x_2 = -1,\ x_3 = 2.$$

Απαιτούμενες Πράξεις

- για την τριγωνοποίηση
  - Υπολογισμός πολλαπλασιαστών
    $1^{ο}$ βήμα θέλω $u_{ii} \leq 2,\ 3,\ \cdots,\ n$. Άρα $n-1$ διαιρέσεις.
  - Υπολογισμός στοιχείων
    $1^{ο}$ βήμα θέλω τα $a_{ij}^{(2)}$ με $i,\ j = 2,\ 3,\ \cdots,\ n$. Άρα $(n-1)^2$ πολλαπλασιασμοί και $(n-1)^2$ προσθέσεις.
- Πίνακας συνολικά: πολλαπλασιασμοί/διαιρέσεις

$$\sum_{i=1}^{n-1}[(n-i)^2 + (n-i)] = \frac{n^3 - n}{3}$$

και προσθέσεις

$$\sum_{i=1}^{n-1}(n-i)^2 = \frac{2n^3 - 3n^2 + n}{6}$$

- Υπολογισμός των στοιχείων b_i.
  1^o βήμα θέλω τα $b_i^{(2)}$ με $i, = 2,\ 3,\ \cdots,\ n$. Άρα $n-1$ πολλαπλασιασμοί και $n-1$ προσθέσεις.

- Δεύτερο μέλος συνολικά: πολλαπλασιασμοί/διαιρέσεις

$$\sum_{i=1}^{n-1}(n-i) = \frac{n(n-1)}{2}$$

  και όμοια προκύπτουν οι προσθέσεις

$$\sum_{i=1}^{n-1}(n-i) = \frac{n(n-1)}{2}.$$

Άρα για την τριγωνοποίηση έχουμε συνολικά: πολλαπλασιασμοί/διαιρέσεις

$$\frac{n^3-n}{3} + \frac{n(n-1)}{2} = \frac{2n^3+3n^2-5n}{6}$$

και προσθέσεις

$$\frac{2n^3+3n^2-5n}{6} + \frac{n(n-1)}{2} = \frac{n^3-n}{3}.$$

Για την οπισθοδρόμηση: πολλαπλασιασμοί/διαιρέσεις

$$\frac{n^2+n}{2}$$

και προσθέσεις

$$\frac{n^2-n}{2}.$$

Συνολικά απαιτούμενες πράξεις για την απαλοιφή Gauss.

− πολλαπλασιασμοί/διαιρέσεις

$$\frac{2n^3+3n^2-5n}{6} + \frac{n^2+n}{2} = \frac{n^3+3n^2-n}{3}$$

− προσθέσεις

$$\frac{n^3-n}{3} + \frac{n^2-n}{2} = \frac{2n^3+3n^2-5n}{6}.$$

Παράδειγμα:
Για να λύσετε ένα γραμμικό σύστημα 20 εξισώσεων με 20 αγνώστους σε έναν υπολογιστή που εκτελεί 10^6 πράξεις το δευτερόλεπτο.

– Με τη μέθοδο απαλοιφής του Gauss (συμπεριλαμβανομένων των προσθέσεων) χρειάζονται $\frac{9180}{3}$ πολλαπλασιασμοί και $\frac{17100}{6}$ προσθέσεις και άρα $5,9 \cdot 10^{-3}$ δευτερόλεπτα.

– Με τον κανόνα του Cramer (υπολογισμός 21 οριζουσών τάξεως, δηλαδή $21(20!) \cdot 19$ πολλαπλασιασμοί και άρα περίπου $3 \cdot 10^5$ αιώνες.

Απαιτούμενη Μνήμη:

– Για την αποθήκευση του πίνακα $\mathbf{A}$: n^2 θέσεις μνήμης,

– Για την αποθήκευση του πίνακα του δεύτερου μέλους n θέσεις μνήμης,

– Για τις πληροφορίες εναλλαγής γραμμών: $O(n)$ θέσεις μνήμης (από 0 έως $n-1$ θέσεις μνήμης),

– Οι πολαπλασιαστές m_{i1} αποθηκεύονται στις θέσεις a_{i1} , όπου $i = 2, 3, \cdots, n$ που αλλιώς θα γέμιζαν με μηδενικά. Γενικά όλοι οι πολλαπλασιαστές αποθηκεύονται στις θέσεις των στοιχείων a_{ij}, $i > j$ του πίνακα, κάτω από τη διαγώνιο,

– Τα καινούρια στοιχεία του πίνακα αποθηκεύονται πάνω στα παλιά. Ομοίως, για τα στοιχεία του δεύτερου μέλους,

– Κατά την οπισθοδρόμηση τα στοιχεία της λύσης $\mathbf{x}$ αποθηκεύονται στις θέσεις του δεύτερου μέλους $\mathbf{b}$.

Συνολικά απαιτούνται $n^2 + n + O(n)$ θέσεις μνήμης.
Παρατηρήσεις.

1. Στην πράξη γίνεται πρώτα η απαλοιφή στον πίνακα $\mathbf{A}$, δηλαδή η τριγωνοποίηση, η οποία απαιτεί $\frac{n^3}{3} + O(n^2)$ πράξεις. Οι δε πολλαπλασιαστές αποθηκεύονται στο κάτω μέρος του πίνακα $\mathbf{A}$. Στη συνέχεια γίνονται οι αλλαγές στο δεύτερο μέλος οι οποίες απαιτούν $\frac{n^2}{2}+O(n)$ και η οπισθοδρόμηση που απαιτεί $\frac{n^2}{2} + O(n)$ πράξεις. Άρα συνολικά $n^2 + O(n)$ πράξεις. Για όλα αυτά απαιτούνται πράξεις. Με τον τρόπο αυτόν μπορούμε να λύσουμε οικονομικά πολλά συστήματα με τον ίδιο πίνακα.

2. Το προηγούμενο βρίσκει εφαρμογή στον υπολογισμό του αντίστροφου ενός πίνακα $\mathbf{A}$, που γίνεται με την επίλυση των γραμμικών συστημάτων.
   $\mathbf{Ax} = \mathbf{e}^i$, $i = 1, 2, \cdots, n$, $\{\mathbf{e}^1, \mathbf{e}^2, \cdots, \mathbf{e}^n\}$ κανονική βάση του $\mathbb{R}^n$.
   Ο συνολικός αριθμός των πράξεων είναι
   $$\frac{n^3}{3} + nn^2 + O(n^2) = \frac{4n^3}{3} + O(n^2).$$
   Επιπλέον απαιτούνται $2n^2 + O(n)$ θέσεις μνήμης.

3. Παράδειγμα.
   Πως θα υπολογίζατε οικονομικά το διάνυσμα $\psi = \mathbf{A}^{-2}\mathbf{b}$ δοθέντων των $\mathbf{A}$, $\mathbf{b}$;

4. Ισχύει ότι $\det \mathbf{A} = (-1)^m \det \mathbf{A}^{(n)} = (-1)^m a_{11}^{(1)} \cdot a_{22}^{(2)} \cdot \cdots \cdot a_{nn}^{(n)}$ όπου m είναι ο αριθμός του πλήθους των εναλλαγών ζευγών γραμμών.

Οδήγηση.
Τα διαγώνια στοιχεία $a_{ii}^{(i)}$, $i = 1, 2, \cdots, n$ του πίνακα $\mathbf{A}^{(n)}$ ονομάζονται οδηγοί.

– Αν κάποιο από αυτά είναι μηδέν, $a_{ii}^{(i)} = 0$, τότε $\nexists \mathbf{A}^{-1}$,

– Λόγω πεπερασμένης ακρίβειας περιμένω προβλήματα αν κάποιο $a_{ii}^{(i)}$ είναι κοντά στο 0.

Παράδειγμα. Να λυθεί με τη μέθοδο του Gauss το γραμμικό σύστημα

$$\begin{aligned} 10^{-4}x_1 + x_2 &= 1 \\ x_1 + x_2 &= 2 \end{aligned}$$

σε έναν υπολογιστή με $\beta = 10$, $t = 3$, $U = -L = 10$ και στρογγύλευση.
Λύση.
Η ακριβής λύση του συστήματος είναι $x_1 = 1,0001$, $x_2 = 0,9999$. Ορίζουσα του πίνακα του συστήματος είναι $10^{-4} - 1$.

$$\begin{aligned} 0,1 \cdot 10^{-3}x_1 + 0,1 \cdot 10^{1}x_2 &= 0,1 \cdot 10^{1} \\ 0,1 \cdot 10^{1}x_1 + 0,1 \cdot 10^{1}x_2 &= 0,2 \cdot 10^{1} \\ m_{21} = \frac{0,1 \cdot 10^{1}}{0,1 \cdot 10^{-3}} = 10^4 = 0,1 \cdot 10^5 \ &\Rightarrow \end{aligned}$$

$$\begin{cases} a_{22}^{(2)} = 0,1 \cdot 10^{1} - 0,1 \cdot 10^{5} \cdot 0,1 \cdot 10^{1} = -0,9999 \cdot 10^4 \overset{\text{στρογγύλευση}}{=} -0,1 \cdot 10^5 \\ b_{2}^{(2)} = 0,2 \cdot 10^{1} - 0,1 \cdot 10^{5} \cdot 0,1 \cdot 10^{1} = -0,9998 \cdot 10^4 \overset{\text{στρογγύλευση}}{=} -0,1 \cdot 10^5 \end{cases}$$

Άρα έχουμε

$$\begin{aligned} 0,1 \cdot 10^{-3}x_1 + 0,1 \cdot 10^{1}x_2 &= 0,1 \cdot 10^{1} \\ -0,1 \cdot 10^{5}x_2 &= -0,1 \cdot 10^{5} \end{aligned}$$

$$\Leftrightarrow x_2 = \frac{-0,1 \cdot 10^5}{-0,1 \cdot 10^5} = 1$$

και

$$x_1 = \frac{0,1 \cdot 10^1 + 0,1 \cdot 10^1 \cdot 1}{0,1 \cdot 10^{-3}} = 0.$$

Στο συγκεκριμένο παράδειγμα το πρόβλημα προκύπτει από το μεγάλο πολλαπλασιαστή και σε σχέση με τα υπόλοιπα στοιχεία.
Εναλλάσοντας τις δύο εξισώσεις

$$0,1 \cdot 10^1 x_1 + 0,1 \cdot 10^1 x_2 = 0,2 \cdot 10^1$$
$$0,1 \cdot 10^{-3} x_1 + 0,1 \cdot 10^1 x_2 = 0,1 \cdot 10^1$$
$$m_{21} = \frac{0,1 \cdot 10^{-3}}{0,1 \cdot 10^1} = 0,1 \cdot 10^{-3}$$

$$\begin{cases} a_{22}^{(2)} = 0,1 \cdot 10^1 - 0,1 \cdot 10^{-3} \cdot 0,1 \cdot 10^1 = 0,9999 \overset{\text{στρογγύλευση}}{=} 0,1 \cdot 10^1 \\ b_2^{(2)} = 0,2 \cdot 10^1 - 0,1 \cdot 10^{-3} \cdot 0,2 \cdot 10^1 = -0,9998 \overset{\text{στρογγύλευση}}{=} 0,1 \cdot 10^1 \end{cases}$$

Έχουμε

$$0,1 \cdot 10^1 x_1 + 0,1 \cdot 10^1 x_2 = 0,2 \cdot 10^1$$
$$+0,1 \cdot 10^1 x_2 = 0,1 \cdot 10^1$$

Οπότε προκύπτουν καλές προσεγγίσεις

$$x_1 = \frac{0,2 \cdot 10^1 - 0,1 \cdot 10^1 \cdot 1}{0,1 \cdot 10^1} \text{ και } x_2 = \frac{0,1 \cdot 10^1}{0,1 \cdot 10^1} = 1.$$

Μέθοδος Απαλοιφής του Gauss με μερική οδήγηση ($k-$οστό βήμα).

- Βρίσκουμε $a_{\rho k}^{(k)}$ τέτοια ώστε $|a_{\rho k}^{(k)}| = \max_{k \le i \le n} |a_{ik}^{(k)}|$.
- Εναλλάσουμε τις γραμμές k και ρ.

Μέθοδος Απαλοιφής του Gauss με ολική οδήγηση ($k-$οστό βήμα).

- Βρίσκουμε $a_{\rho q}^{(k)}$ τέτοια ώστε $|a_{\rho q}^{(k)}| = \max_{k \le i,\, j \le n} |a_{ij}^{(k)}|$.
- Εναλλάσουμε τις γραμμές k και ρ και τις στήλες k και q.

Στη μερική οδήγηση χρειαζόμαστε

$$\sum_{i=1}^{n-1} (n-i) = \frac{n(n-1)}{2} \text{ συγκρίσεις.}$$

Στην ολική οδήγηση χρειαζόμαστε

$$\sum_{i=1}^{n-1} (n-i)^2 = \frac{2n^3 - 3n^2 + n}{6} \text{ συγκρίσεις.}$$

Κατάσταση γραμμικών συστημάτων.

Παράδειγμα. Να λυθεί με τη μέθοδο απαλοιφής του Gauss με ολική οδήγηση το γραμμικό σύστημα

$$0,913x_1 + 0,659x_2 = 0,254$$
$$0,780x_1 + 0,563x_2 = 0,217$$

σε υπολογιστή με $\beta = 10$, $t = 3$, $U = -L = 10$ και αποκοπή. Η ακριβής λύση είναι $x_1 = 1$, $x_2 = -1$ και η ορίζουσα του πίνακα του συστήματος είναι -10^{-6}.
Λύση.
Μετά την τριγωνοποίηση έχω

$$\left.\begin{array}{c} 0,913x_1 + 0,659x_2 = 0,254 \\ 0,001x_2 = 0,001 \end{array}\right\} \Rightarrow \begin{array}{c} x_1 = -0,443 \\ x_2 = 1 \end{array}$$

Αλλάζοντας λίγο τα δεδομένα του δεύτερου μέλους

$$\begin{array}{c} 0,913x_1 + 0,659x_2 = 0,253 \\ 0,780x_1 + 0,001x_2 = 0,218 \end{array} \rightsquigarrow$$

αλλαγή κατά 10^{-3}. Η ακριβής λύση είναι $x_1 = 1,223$ και $x_2 = -1,694$.
Το γραμμικό σύστημα

$$0,913 \cdot 10^6 x_1 + 0,659 \cdot 10^6 x_2 = 0,254 \cdot 10^6$$
$$0,780x_1 + 0,563x_2 = 0,217$$

έχει ορίζουσα του πίνακα -1. Στον ίδιο υπολογιστή η λύση που παίρνουμε είναι $x_1 = -0,443$ και $x_2 = 1$. Άρα το θέμα δεν είναι η τιμή της ορίζουσας.

Νόρμες Διανυσμάτων.

Ορισμός. Έστω X ένας $K-$γραμμικός χώρος με $K = \mathbb{R}$ ή $K = \mathbb{C}$. Μια απεικόνιση $\|\cdot\| : X \to \mathbb{R}$, $x \to \|x\|$ λέγεται νόρμα(norm), αν ισχύουν

$N1$ $x \in X$ $\|x\| = 0 \Leftrightarrow x = 0$

$N2$ $\forall \lambda \in K$ και $x \in X$ $\|\lambda x\| = |\lambda| \cdot \|x\|$

$N3$ $\forall x, \psi \in X$ $\|x + \psi\| \leq \|x\| + \|\psi\|$

Παρατηρήσεις.

1. Ισχύει $\|x\| \geq 0$
   Πράγματι,
   $$0 = \|x - x\| = \|x + (-1)x\| \leq \|x\| + |-1|\|x\| = 2\|x\|$$

2. Ισχύει η τριγωνική ανισότητα προς τα κάτω, δηλαδή
   $$\forall x,\, \psi \in X \big| \|x\| - \|\psi\| \big| \leq \|x - \|\psi\| \Leftrightarrow -\|x - \psi\| \leq \|x\| - \|\psi\| \leq \|x - \psi\|$$
   Πράγματι,
   $$\|x\| = \|x - \psi + \psi\| \leq \|x - \psi\| + \|\psi\| \Rightarrow \|x\| - \|\psi\| \leq \|x - \psi\|$$
   Ομοίως,
   $$-\|x - \psi\| \leq \|x\| - \|\psi\|.$$

Παραδείγματα.

1. $(\mathbb{R}^n,\, \|\cdot\|_1)$ με $\|x\|_1 = \sum_{i=1}^n |x_i|$, $x = (x_1,\, x_2,\, \cdots,\, x_n)^T$. Πράγματι ισχύουν οι ιδιότητες $N1$ και $N2$ οι οποίες αφήνονται ως άσκηση. Για την ιδιότητα $N3$ έχουμε,
   $$x,\, \psi \in \mathbb{R}^n \ \|x + \psi\|_1 = \sum_{i=1}^n |x_i + \psi_i| \leq \sum_{i=1}^n |x_i| + \sum_{i=1}^n |\psi_i| = \|x\|_1 + \|\psi\|_1.$$

2. $(\mathbb{R}^n,\, \|\cdot\|_2)$ με $\|x\|_2 = (\sum_{i=1}^n |x_i|^2)^{\frac{1}{2}}$. Πράγματι ισχύουν οι ιδιότητες $N1$ και $N2$ οι οποίες αφήνονται ως άσκηση. Για την ιδιότητα $N3$ έχουμε,
   Η $\|\cdot\|_2$ παράγεται από το σύνηθες εσωτερικό γινόμενο
   $$(\cdot,\, \cdot) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R},\, (x,\, \psi)_2 = \sum_{i=1}^n (x_i,\, \psi_i).$$
   Πράγματι, $\|x\|_2 = \sqrt{(x,\, x)_2}$. Επιπλέον, ισχύει $\|(x,\, \psi)\|_2 \leq \|x\|_2 \|\psi\|_2$. Πράγματι για $\lambda \in \mathbb{R}$ έχουμε
   $$o \leq (x + \lambda\psi,\, x + \lambda\psi)_2 = \cdots = \|x\|_2^2 + 2\lambda(x,\, \psi)_2 + \lambda^2\|\psi\|_2^2.$$
   Αν $\psi \neq 0$ (αν $\psi = 0$, η ανισότητα ισχύει) η διακρίνουσα του τριωνύμου πρέπει
   $$4(x,\, \psi)_2^2 - 4\|\psi\|_2^2\|x\|_2^2 \leq 0 \Rightarrow |(x,\, \psi)_2| \leq \|\psi\|_2\|x\|_2.$$
   Άρα,
   $$x,\, \psi \in \mathbb{R}^n \ \|x + \psi\|^2 = (x + \psi,\, x + \psi)_2 = \|x\|_2^2 + 2(x,\, \psi)_2 + \|\psi\|_2^2$$
   $$= \|x\|_2^2 + 2\|x\|_2\|\psi\|_2 + \|\psi\|_2^2 = (\|x\|_2 + \|\psi\|_2)^2$$
   $$\Rightarrow \|x + \psi\|_2 \leq \|x\|_2 + \|\psi\|_2.$$

3. $(\mathbb{R}^n, \|\cdot\|_\infty)$ με $\|x\|_\infty = \max_{1\le i\le n}|x_i|$. Πράγματι ισχύουν οι ιδιότητες $N1$ και $N2$ οι οποίες αφήνονται ως άσκηση. Για την ιδιότητα $N3$ έχουμε,

$$x,\, \psi \in \mathbb{R}^n,\ \|x+\psi\|_\infty = \max_{1\le i\le n}|x_i+\psi_i| \le (\text{γιατί;})\max(|x_i|+|\psi_i|) \le (\text{γιατί;})\max|x_i|+\max|\psi_i|$$
$$= \|x\|_\infty + \|\psi\|_\infty$$

Νόρμες Διανυσμάτων.

Παρατήρηση: ο $\mathbb{R}^n$ είναι γραμμικός (ή διανυσματικός) χώρος συνεπώς η απόσταση που ορίζει μεταξύ των διανυσμάτων είναι μετρική. Στον $(\mathbb{R}^n, \|\cdot\|)$ ορίζω $p : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ τότε η $p :$ είναι μετρική (αφήνεται ως άσκηση) $((x,\, \psi) \to \|x-\psi\|)$.
Παίζει ρόλο ποια νόρμα παίρνω για τον ορισμό της μετρικής. Οι τιμές για νόρμα: $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_\infty$ είναι διαφορετικές, αλλά συγκρίνονται; Η απάντηση είναι ναι αλλά με ισοδυναμία νορμών.
Ορισμός. Οι $\|\cdot\|_a$, $\|\cdot\|_b$ είναι "ισοδύναμες ή συγκρίσιμες νόρμες" στον $\mathbb{R}^n$ αν υπάρχουν σταθερές c_1, c_2 που εξαρτώνται από τα a, b έτσι ώστε

$$\forall x \in \mathbb{R}^n :\ c_1\|x\|_a \le \|x\|_b \le c_2\|x\|_a.$$

Πρόταση. Όλες οι νόρμες στον $\mathbb{R}^n$ είναι ισοδύναμες.
(Απόδειξη: Δείχνουμε ότι $\forall p : \|\cdot\|_p \sim \|\cdot\|_\infty$).
Παράδειγμα.
Αν $a = 1$, $b = \infty \Rightarrow \|\cdot\|_a \sim \|\cdot\|_\infty$.
Πράγματι,

$$\frac{1}{n}\|x\|_1 \le \|x\|_\infty \le \|x\|_1\ \forall x \in \mathbb{R}^n \tag{4}$$

ή ισοδύναμα

$$\|x\|_\infty \le \|x\|_1 \le n\|x\|_\infty\ \forall x \in \mathbb{R}^n \tag{5}$$

Απόδειξη της πρότασης. θα αποδείξουμε την σχέση (4).

$$\|x\|_1 = \sum_{k=1}^{n}|x_k| \le n\cdot\max|x_k| = n\|x\|_\infty$$

$$\|x\|_\infty = \max_{1\le k\le n}|x_k| \le \sum_{k=1}^{n}|x_k| = \|x\|_1.$$

Όμοια αποδεικνύεται και η σχέση (5).
Πως προέκυψε η ονομασία "συγκρίσιμες" ;
Αν

$$x = (-1,\, 2,\, -3) \Rightarrow \|x\| = \|x-0\| = \rho(x,\, 0)\,\text{η απόσταση του } x \text{ από την αρχή των αξόνων}$$

$\|x\|_1 = 6$
$\|x\|_2 = \sqrt{1+4+9} = \sqrt{14}$
$\|x\|_\infty = 3$.
Διαφορετικές τιμές μεν, αλλά συγκρίσιμες. Οι τιμές είναι "ίδιας τάξης", δεν είναι π.χ. η μία 5 και η άλλη $3 \cdot 10^5$ ("συγκρίσιμες"). Επομένως, αν για μια ακολουθία $x^{(k)}$, $k \geq 1$ στην $\mathbb{R}^n$ ισχύει $\lim_{k\to\infty} \|x^{(k)}\|_a = 0 \Rightarrow$ ισχύει $\lim_{k\to\infty} \|x^{(k)}\|_b = 0$, $\forall b$. Άρα επιλέγω την ευκολότερα επιλύσιμη νόρμα στο κάθε πρόβλημα.

Νόρμες Πινάκων

Σε ό,τι ακολουθεί ο χώρος είναι ο $\mathbb{R}^{n\times n}$ και τα A, B είναι πίνακες στον $\mathbb{R}^{n\times n}$.
Ισχύουν οι ιδιότητες $N1$, $N2$, $N3$ και η:

$$\|AB\| \leq \|A\| \cdot \|B\|, \ \forall A, B \in \mathbb{R}^{n\times n}.$$

Φυσικές ή παραγόμενες από νόρμες διανυσμάτων.
Ορισμός. Έστω $\|\cdot\|$ νόρμα στον $\mathbb{R}^n$. Η απεικόνιση

$$\|\cdot\| : \mathbb{R}^{n\times n} \to \mathbb{R}^n, \ \|A\| = \max_{x\in\mathbb{R}^n,\, \|x\|\leq 1} \|Ax\|$$

λέγεται φυσική νόρμα πινάκων ή νόρμα πινάκων παραγόμενη απο την στον $\mathbb{R}^n$.

Πρώτον: Δείχνω ότι η απεικόνιση είναι καλώς ορισμένη. Θα δείξουμε ότι το σύνολο

$$\{\|Ax\|, \ x \in \mathbb{R}^n, \ \|x\| \leq 1\}$$

είναι φραγμένο:

$$\|Ax\| \leq c\|Ax\|_\infty = c \max_{1\leq i\leq n} |\sum_{j=1}^n a_{ij}x_j| \leq c \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}| \cdot |x_j|$$

$$\leq c \max_{1\leq j\leq n} |x_j| \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}| = c\|x\|_\infty \cdot \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}| \leq c\cdot c_1 \|x\| \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}| \leq c\cdot c_1 \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}|$$

είναι φραγμένο άρα έχει supremum. Γιατί είναι max ;
(Θεωρούμε την $S : \mathbb{R}^n \to \mathbb{R}$, $S(x) = \|Ax\|$ η οποία είναι συνεχής. Το σύνολο $\|x\| \leq 1$ είναι συμπαγες.)

Επιπλέον, η παραπάνω απεικόνιση πληροί τις ιδιότητες

1. $\|A\| \geq 0 \, \forall A \in \mathbb{R}^{n\times n}$ και $\|A\| = 0 \Leftrightarrow A = 0$,
2. $\|\lambda A\| = |\lambda| \|A\| \, \forall \lambda \in \mathbb{R} \forall A \in \mathbb{R}^{n\times n}$,
3. $\|A + B\| \leq \|A\| + \|B\| \, \forall A, \, B \in \mathbb{R}^{n\times n}$,
4. $\|Ax\| \leq \|A\| \|x\| \, \forall A \in \mathbb{R}^{n\times n}$, $\forall x \in \mathbb{R}^n$,
5. $\|AB\| \leq \|A\| \cdot \|B\| \, \forall A, \, B \in \mathbb{R}^{n\times n}$,
6. $\|I\| = 1$.

Πράγματι,

1. $\|A\| \geq 0 \, \forall A \in \mathbb{R}^{n\times n}$.
   Έστω
   $$\|A\| = 0 \Rightarrow \max_{x\in\mathbb{R}^n, \, \|x\|\leq 1} \|Ax\| = 0 \Rightarrow \|Ax\| = 0 \; x \in \mathbb{R}^n, \; \|x\| \leq 1$$
   $Ax = 0 \, x \in \mathbb{R}^n$, $\|x\| \leq 1$ δεν μπορώ από εδώ να πάω $A = 0$ γιατί δεν ισχύει για κάθε x.
   Ορίζω $\psi = \dfrac{x}{\|x\|} \Rightarrow \|\psi\| = 1$ (αφήνεται ως άσκηση) τότε $A\psi = 0 \Rightarrow \dfrac{Ax}{\|x\|} = 0 \Rightarrow Ax = 0 \Rightarrow A = 0$.
2. $\|\lambda A\| = \max_{x\in\mathbb{R}^n, \, \|x\|\leq 1} \|(\lambda A)x\| = \max \|\lambda(Ax)\| = \max |\lambda| \cdot \|Ax\|$
   $= \|\lambda\| \max \|Ax\| = |\lambda| \cdot \|A\|$.
3. $\|A + B\| = \max \|(A + B)x\| = \max \|Ax + Bx\| \leq \max(\|Ax\| + \|Bx\|)$
   $\leq \max \|Ax\| + \max \|Bx\| = \|A\| + \|B\|$.
4. Αν $x = 0 \Rightarrow \|Ax\| = 0 = \|A\| \cdot \|x\|$.
   Αν $x \neq 0 \Rightarrow$ ορίζω $\psi = \dfrac{x}{\|x\|} \Rightarrow \|\psi\| = 1$ τότε
   $\|A\psi\| = \|A\dfrac{x}{\|x\|}\| \leq \max_{\psi\in\mathbb{R}^n, \, \|\psi\|\leq 1} \|A\psi\| = \|A\|$. Άρα,
   $$\|A\psi\| \leq \|A\| \Leftrightarrow \|\frac{1}{\|x\|} \cdot Ax\| \leq \|A\| \Leftrightarrow \frac{1}{\|x\|}\|Ax\| \leq \|A\| \Leftrightarrow \|Ax\| \leq \|A\| \cdot \|x\|.$$
5. Έστω $x \in \mathbb{R}^n$, $\|x\| \leq 1$, τότε
   $$\|(AB)x\| = \|A(Bx)\| \leq \|A\| \cdot \|Bx\| \leq \|A\| \cdot \|B\| \cdot \|x\| \leq \|A\| \cdot \|B\|$$
   Συνεπώς
   $$\|AB\| = \max_{x\in\mathbb{R}^n, \, \|x\|\leq 1} \|(AB)x\| \leq \|A\| \cdot \|B\| \; \forall A, \, B \in \mathbb{R}^{n\times n}$$

6.

$$I = \max_{x\in\mathbb{R}^n,\, \|x\|\leq 1} \|I\cdot x\| = \max_{x\in\mathbb{R}^n,\, \|x\|=1} \|x\| = 1.$$

Παρατηρήσεις

1.

$$\|A\| = \max_{x\in\mathbb{R}^n,\, \|x\|\leq 1} \|Ax\| = \max_{x\in\mathbb{R}^n,\, \|x\|=1} \|Ax\| = \max_{x\in\mathbb{R}^n,\, \|x\|\neq 0} \frac{\|Ax\|}{\|x\|},$$

2. και στον $\mathbb{R}^{n\times n}$ όλες οι νόρμες είναι ισοδύναμες μεταξύ τους.

- Έστω $(\mathbb{R}^n, \|\cdot\|_\infty)$. Η παραγόμενη από την $\|\cdot\|_\infty$ νόρμα πίνακα στον $\mathbb{R}^{n\times n}$ είναι η λεγόμενη νόρμα του αθροίσματος γραμμών

$$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|.$$

Απόδειξη.
Για $A\neq 0$ (για $A=0 \Rightarrow \|A\|_\infty = 0$) και $\|x\|_\infty \leq 1$, έχω:

$$\|Ax\|_\infty \leq \|x\|_\infty \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| \leq \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|$$

$$\Rightarrow \|A\|_\infty = \max_{x\in\mathbb{R}^n,\, \|x\|_\infty\leq 1} \|Ax\|_\infty \leq \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|.$$

Δείξαμε το $\leq$, τώρα θα δείξουμε το $\geq$.
Αν k τέτοια ώστε $\sum_{j=1}^{n} |a_{kj}| = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|$, τότε ορίζω το διάνυσμα $\psi\in\mathbb{R}^n$:

$$\psi_j = \begin{cases} \dfrac{a_{kj}}{|a_{kj}|}, \text{ για } a_{kj}\neq 0 \\ 0, \text{ αλλιώς} \end{cases} \Rightarrow \|\psi\|_\infty = 1.$$

επίσης

$$\|A\psi\|_\infty = \max \sum a_{ij}\psi_j \geq |\sum_{j=1}^{n} a_{kj}\psi_j| = |\sum_{j=1}^{n} a_{kj}\frac{a_{kj}}{|a_{kj}|}| = |\sum_{j=1}^{n} \frac{a_{kj}^2}{|a_{kj}|}| = \sum \frac{|a_{kj}|^2}{|a_{kj}|}$$

$$= \sum |a_{kj}| = \max_{1\leq i\leq n} \sum |a_{ij}|.$$

Δηλαδή

$$\|A\psi\|_\infty \geq \max \sum |a_{ij}| \Rightarrow \|A\| = \max_{\psi\in\mathbb{R}^n,\, \|\psi\|\leq} \|A\psi\|_\infty \geq \|A\psi\|_\infty \geq \max \sum |a_{ij}|.$$

- Από την $\|\cdot\|_1$ παράγεται η λεγόμενη νόρμα του αθροίσματος στηλών $\|A\|_1 = \max_{1\le j\le n}\sum_{i=1}^{n}|a_{ij}|$.
- Από την $\|\cdot\|_2$ η παραγόμενη νόρμα είναι η λεγόμενη φασματική[1] νόρμα $\|A\|_2 = \max_{1\le i\le n}\sqrt{\lambda_i(AA^T)}$ όπου λ_i ιδιοτιμές του AA^T (ο οποίος είναι συμμτρικός $\Rightarrow \lambda_i \ge 0$).

Δείκτης κατάστασης πίνακα - συστήματος.

$$0,913x_1 + 0,659x_2 = 0,254$$
$$0,780x_1 + 0,563x_2 = 0,217$$

Ερώτημα: Αν στο σύστημα $Ax = b$, $A \in \mathbb{R}^{n\times n}$, A αντιστρέψιμος, $b \in \mathbb{R}^n$, $b \neq 0$ επιφέρουμε μί αλλαγή στο δεύτερο μέλος $\delta b = (\delta b_1, \delta b_2, \cdots, \delta b_n)^\perp$ μπορούμε να εκτιμήσουμε την αλλαγή που επέρχεται στη λύση ; $(\delta x_1, \delta x_2, \cdots, \delta x_n)^\perp$;
Έστω $\|\cdot\|$ μία νόρμα στον $\mathbb{R}^n$ και η αντίστοιχη επαγόμενη (φυσική) νόρμα στον $\mathbb{R}^{n\times n}$.

$$\left.\begin{array}{c} A(x+\delta x) = b + \delta b \\ Ax = b \end{array}\right\} \overset{\text{αφαίρεση κατά μέλη}}{\Longrightarrow} A(\delta x) = \delta b \Rightarrow \delta x = A^{-1}(\delta b)$$

$$\Rightarrow \|\delta x\| = \|A^{-1}\delta b\| \le \|A^{-1}\|\|\delta b\| \tag{6}$$

$$b = Ax \Rightarrow \|b\| = \|Ax\| \le \|A\|\cdot\|x\| \Rightarrow \frac{1}{\|x\|} \le \frac{\|A\|}{\|b\|} \tag{7}$$

Πολλαπλασιάζοντας τις (6)-(7) κατά μέλη έχω

$$\frac{\|\delta x\|}{\|x\|} \le \|A\|\cdot\|A^{-1}\|\cdot\frac{\|\delta b\|}{\|b\|}.$$

Η ποσότητα $\|A\|\cdot\|A^{-1}\|$ συμβολίζεται με $k(A)$ και ονομάζεται *δείκτης κατάστασης* του πίνακα A.
Θεώρημα.
Έστω $\|\cdot\|$ μία νόρμα στον $\mathbb{R}^n$ και η παραγόμενη από αυτήν φυσική νόρμα στον $\mathbb{R}^{n\times n}$. Έστω $A \in \mathbb{R}^{n\times n}$ ένας αντιστρέψιμος πίνακας, $\delta A \in \mathbb{R}^{n\times n}$ και b, $\delta b \in \mathbb{R}^n$, $b \neq 0$. Τότε αν $k(A) = \|A\|\cdot\|A^{-1}\|$ είναι ο δείκτης κατάστασης του A, έχουμε

i Αν $Ax = b$ και $A(x+\delta x) = (b+\delta b)$, τότε

$$\frac{\|\delta x\|}{\|x\|} \le k(A)\cdot\frac{\|\delta b\|}{\|b\|}$$

[1] το $p(A)$, δηλαδή το max των απολύτων τιμών των ιδιοτιμών του A λέγεται η φασματική ακτίνα του A.

ii Αν $Ax = b$, $(A + \delta A)((x + \delta x) = (b + \delta b)$ και $\|A^{-1}\| \cdot \|\delta A\| < 1$, τότε ο $A + \delta A$ είναι αντιστρέψιμος και

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{k(A)}{1 - \|A^{-1}\| \cdot \|\delta A\|} \cdot (\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}).$$

(Η απόδειξη του ερωτήματος υπάρχει στο βιβλίο των Ακρίβη-Δουγαλή).
Παρατηρήσεις.

1. Ο δείκτης κατάστασης εξαρτάται από τη νόρμα που χρησιμοποιούμε, διαφορετικές νόρμες δίνουν διαφορετικές (αλλά συγκρίσιμες) τιμές. Επιπλέον,

$$k(A) = \|A\| \cdot \|A^{-1}\| \geq \|A \cdot A^{-1}\| = \|I\| = 1.$$

   Αν ο δείκτης κατάστασης είναι μικρός, π.χ. $1 \leq k(A) \leq \infty$ λέμε ότι ο πίνακας A έχει καλή κατάσταση. Αντίθετα, αν ο δείκτης κατάστασης είναι μεγάλος, λέμε ότι ο πίνακας A έχει κακή κατάσταση.

2. Το φράγμα στο i είναι καλό. (παράδειγμα όπου επιτυγχάνεται η ισότητα θα δοθεί στις ασκήσεις).

3. Στο γραμμικό σύστημα

$$0,913x_1 + 0,659x_2 = 0,254$$
$$0,780x_1 + 0,563x_2 = 0,217$$

   η αλλαγή στο δεύτερο μέλος κατά 10^{-3} , επέφερε μία αλλαγή στη λύση κατά 10^3. υπεύθυνη γιάυτό είναι η κακή κατάσταση του πίνακα A, που ως προς την $\|\cdot\|_1$, έχει
   $k_1(A) = \|A\|_1 \cdot \|A^{-1}\|_1 \cong 2,7 \cdot 10^6$ και αφήνεται ως άσκηση, με πίνακα $A = \begin{pmatrix} 0,913 & 0,658 \\ 0,780 & 0,583 \end{pmatrix}$.

4. Η ορίζουσα είναι κακός δείκτης για την κατάσταση ενός πίνακα (το μέγεθος της ορίζουσας ουδεμία σχέση έχει με την καλή ή κακή κατάσταση του πίνακα). Ο πίνακας

$$D = \begin{pmatrix} \frac{1}{10} & & \mathbf{0} & \\ & \frac{1}{10} & & \\ & & \vdots & \\ \mathbf{0} & & & \frac{1}{10} \end{pmatrix}$$

έχει $\det D = 10^{-n}$ και είναι αντιστρέψιμος με $D^{-1} = \begin{pmatrix} 10 & & \mathbf{0} \\ & 10 & \\ & & \vdots \\ \mathbf{0} & & 10 \end{pmatrix}$,

και $k_1(A) = \|D\|_1 \|D^{-1}\|_1 = \frac{1}{10} \cdot 10 = 1$.

5. Αν $A = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix}$, $A^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\varepsilon} \end{pmatrix}$, $0 < \varepsilon < 1$ τότε

$$k_1(A) = \|A\|_1 \cdot \|A^{-1}\|_1 = 1 \cdot \frac{1}{\varepsilon} = \frac{1}{\varepsilon}.$$

Καθώς $\varepsilon \to 0$, οπότε ο πίνακας τείνει να γίνει μη αντιστρέψιμος, ο δείκτης $k_1(A) \to \infty$.
Γενικά, για μία νόρμα $\|\cdot\|$ ισχύει ότι

$$\frac{1}{k(A)} \leq \inf\{\frac{\|A-B\|}{\|A\|}, \, B \, \text{μη αντιστρέψιμος}\}$$

.

6. Μια εκτίμηση του δείκτη κατάστασης μπορεί να γίνει ως εξής:

$$Aw = \psi \Leftrightarrow w = A^{-1}\psi \Rightarrow \|w\| = \|A^{-1}\psi\| \leq \|A^{-1}\| \cdot \|\psi\| \Rightarrow \|A^{-1}\| \geq \frac{\|w\|}{\|\psi\|}.$$

Έτσι, επιλέγουμε k διανύσματα ψ_i, $i = 1, 2, \cdots, k$, λύνουμε τα k συστήματα $Aw_i = \psi_i$, $i = 1, 2, \cdots, k$ και κατόπιν παίρνουμε

$$\|A^{-1}\| \cong \max_{1 \leq i \leq k} \frac{\|w_i\|}{\|\psi_i\|} (\max_{1 \leq i \leq k} \frac{\|A^{-1}\psi_i\|}{\|\psi_i\|}).$$

Ακριβής τιμή $\|A^{-1}\| = \sup_{\psi \in \mathbb{R}^n, \, \psi \neq 0} \frac{\|A^{-1}\psi_i\|}{\|\psi_i\|}$.
Συνήθως, $k = 2$ ή 3. Επιπλέον κόστος, είναι $kn^2 + O(n)$ πράξεις.

Επιρροή του Δείκτη Κατάστασης στην Απαλοιφή.

Αν σφάλματα στρογγύλευσης υπεισέρχονται μόνο κατά την παράσταση των στοιχείων του δεύτερου μέλους, ενώ τα στοιχεία του πίνακα παριστάνονται ακριβώς, και όλες οι πράξεις γίνονται ακριβώς, τότε

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq k(A) \frac{1}{2} \beta \beta^{-t} \text{ αφήνεται ως άσκηση.}$$

Στη γενική περίπτωση ο Wilkinson απέδειξε ότι

$$\frac{\|x - \tilde{x}\|}{\|\tilde{x}\|} \leq k(A)p\beta^{-t},$$

όπου $k(A)$ είναι η επιρροή προβλήματος, p η επιρροή του αλγορίθμου και β^{-t} η επιρροή της αριθμητικής.
Το p είναι μικρό για ευσταθή αλγόριθμο και μεγάλο για ασταθή.
Το $p \cong 10$ για απαλοιφή του Gauss με μερική οδήγηση και $p \cong 1$ για απαλοιφή του Gauss με ολική οδήγηση.

Ασκήσεις στην Αριθμητική Γραμμική Άλγεβρα.

3.3 Έστω A, $B \in \mathbb{R}^{n\times n}$ αντιστρέψιμος και $b \in \mathbb{R}^n$. Πως υπολογίζουμε κατά το δυνατόν οικονομικότερα από άποψη πλήθους πράξεων και μνήμης τα διανύσματα $A^{-2}b$, $A^{-1}BA^{-1}b$;
Λύση.

$$A^{-4}b = A^{-1}\cdot A^{-1}\cdot A^{-1}\cdot A^{-1}\cdot b.$$

Συμβολίζω

$$x = A^{-1}b \Leftrightarrow Ax = b$$
$$\psi = A^{-1}x \Leftrightarrow A\psi = x$$
$$w = A^{-1}\psi \Leftrightarrow Aw = \psi$$
$$z = A^{-1}w \Leftrightarrow Az = w$$

Κόστος: $\frac{n^3}{3} + O(n^2)$ πράξεις.
Θέσεις μνήμης: $n^2 + O(n)$ διότι την τριγωνοποίηση την κάνουμε μία φορά στο σύστημα $Ax = b$ και για τις οπισθοδρομήσεις έχουμε μόνο ένα υπόλοιπο της τάξης n^2.

3.35 Έστω $\|\cdot\|$ μία νόρμα στον $\mathbb{R}^n$ και $\|\cdot\|$ η νόρμα στον $\mathbb{R}^{n\times n}$ που παράγεται από αυτή. Αν $A \in \mathbb{R}^{n\times n}$ με $\|A\| < 1$, αποδείξτε ότι ο πίνακας $I_n - A$ είναι αντιστρέψιμος και επιπλέον ότι ισχύει

$$\frac{1}{1+\|A\|} \le \|(I_n - A)^{-1}\| \le \frac{1}{1-\|A\|}.$$

Λύση. Έστω ότι ο $I_n - A$ δεν είναι αντιστρέψιμος. Τότε $\exists x \in \mathbb{R}^n$, $x \neq 0$:
$(I_n - A)x = 0 \Rightarrow x - Ax = 0 \Rightarrow x = Ax$.
Στην τελευταία σχέση βάζουμε νόρμα,

$$\|x\| = \|A\cdot x\| \le \|A\|\cdot\|x\| \Leftrightarrow \|A\| \ge 1 \text{ Άτοπο.}$$

Επιπλέον,

$$
\begin{gathered}
(I_n - A)^{-1}(I_n - A) = I_n \\
\Leftrightarrow (I_n - A)^{-1} - (I_n - A)^{-1}A = I_n \\
\Leftrightarrow (I_n - A)^{-1} = I_n + (I_n - A)^{-1} \cdot A \\
\Rightarrow \|(I_n - A)^{-1}\| = \|I_n + (I_n - A)^{-1} \cdot A\| \leq \\
\leq \|I_n\| + \|(I_n - A)^{-1} \cdot A\| \leq \\
\leq 1 + \|(I_n - A)^{-1}\| \cdot \|A\| \\
\Rightarrow \|(I_n - A)^{-1}\| \leq 1 + \|(I_n - A)^{-1}\| \cdot \|A\| \\
\Rightarrow \|(I_n - A)^{-1}\|(1 - \|A\|) \leq 1 \\
\Rightarrow \|(I_n - A)^{-1}\| \leq \frac{1}{1 - \|A\|}
\end{gathered}
$$

και

$$
\begin{gathered}
(I_n - A)^{-1}\|(I_n - A) = I_n \\
\Rightarrow \|I_n\| = \|(I_n - A)^{-1}(I_n - A)\| \\
\Rightarrow 1 \leq \|(I_n - A)^{-1}\|\|(I_n - A)\| \leq \|(I_n - A)^{-1}\|(\|I_n\| + \|A\|) \\
\Rightarrow 1 \leq \|(I_n - A)^{-1}\|(1 + \|A\|) \\
\Rightarrow \frac{1}{1 + \|A\|} \leq \|(I_n - A)^{-1}\|.
\end{gathered}
$$

Εφαρμογή.

Δείξτε ότι ο πίνακας $B = \begin{pmatrix} 4 & 1 & \cdots & & 0 \\ 1 & 4 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \\ 0 & \cdots\cdots & & 1 & 4 \end{pmatrix}$ τριδιαγώνιος αντιστρέφεται και ότι $k_\infty(B) = \|B\|_\infty \|B^{-1}\|_\infty \leq 3$.

Λύση.

$$
B = \begin{pmatrix} 4 & 1 & \cdots & & 0 \\ 1 & 4 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \\ 0 & \cdots\cdots & & 1 & 4 \end{pmatrix} = 4 \begin{pmatrix} 1 & \frac{1}{4} & 1 & \cdots & & 0 \\ \frac{1}{4} & 1 & \frac{1}{4} & \cdots & & 0 \\ \vdots & \vdots & & \vdots & & \\ 0 & \cdots & \cdots & & \frac{1}{4} & 1 \end{pmatrix}
$$

$$= 4\big[I_n - \begin{pmatrix} 0 & -\frac{1}{4} & 0 & \cdots & & 0 \\ -\frac{1}{4} & 0 & -\frac{1}{4} & \cdots & & 0 \\ \vdots & \vdots & & \vdots & & \\ 0 & \cdots & \cdots & & -\frac{1}{4} & 0 \end{pmatrix}\big].$$

Δηλαδή,

$B = 4(I_n - A)$ όπου $A = \begin{pmatrix} 0 & -\frac{1}{4} & 0 & \cdots & & 0 \\ -\frac{1}{4} & 0 & -\frac{1}{4} & \cdots & & 0 \\ \vdots & \vdots & & \vdots & & \\ 0 & \cdots & \cdots & & -\frac{1}{4} & 0 \end{pmatrix}$ και

$$\|A\|_\infty = |-\frac{1}{4}| + |-\frac{1}{4}| = \frac{1}{2} < 1.$$

Άρα, ο $I_n - A$ αντιστρέφεται, δηλαδή ο $4(I_n - A)$ αντιστρέφεται και άρα ο B αντιστρέφεται. Επιπλέον,

$$\|B^{-1}\|_\infty = \|[4(I_n - A)]^{-1}\| = \|\frac{1}{4}(I_n - A)^{-1}\| \leq \frac{1}{4}\|(I_n - A)^{-1}\| \leq \frac{1}{4}\frac{1}{1 - \|A\|_\infty} = \frac{1}{4}\frac{1}{1 - \frac{1}{2}} = \frac{1}{2}.$$

$$\|B\|_\infty = 1 + 4 + 1 = 6.$$

Άρα,

$$k_\infty(B) = \|B\|_\infty \|B^{-1}\|_\infty \leq 6 \cdot \frac{1}{2} = 3.$$

3.39 Έστω $A, B \in \mathbb{R}^{n\times n}$, A αντιστρέψιμος. Αν $\|\cdot\|$ είναι μια φυσική νόρμα πινάκων και $\|A - B\| < \dfrac{1}{\|A^{-1}\|}$, αποδείξτε ότι ο B είναι αντιστρέψιμος. Οδηγηθείτε στο συμπέρασμα, ότι για οποιονδήποτε μη αντριστέψιμο πίνακα $B \in \mathbb{R}^{n\times n}$ ισχύει

$$\frac{1}{k(A)} \leq \frac{\|A - B\|}{\|A\|}.$$

Λύση.

Έστω ότι ο B δεν είναι αντιστρέψιμος τότε $\exists x \in \mathbb{R}^n$, $x \neq 0 :\ Bx = 0$. Έχουμε,

$(A-B)x = Ax - Bx = Ax \Rightarrow A^{-1}(A-B)x = x \Rightarrow \|x\| = \|A^{-1}(A-B)x\| \leq \|A^{-1}\| \cdot \|A-B\| \cdot \|x\|$,

όπου $\|x\| \neq 0 \Rightarrow 1 \leq \|A^{-1}\| \cdot \|A - B\| \Rightarrow \|A - B\| \geq \dfrac{1}{\|A^{-1}\|}$, Άτοπο. Για B μη αντιστρέψιμο, ισχύει

$$\frac{1}{\|A^{-1}\|} \leq \|A - B\| \underset{A\ \text{αντιστρέψιμος}}{\overset{\|A\|\neq 0}{\Longrightarrow}} \frac{1}{\|A\| \cdot \|A^{-1}\|} \leq \frac{\|A - B\|}{\|A\|} \Rightarrow$$

$$\frac{1}{k(A)} \leq \frac{\|A-B\|}{\|A\|}.$$

3.40

a) Έστω A αντιστρέψιμος άνω ή κάτω τριγωνικός. Αποδείξτε ότι για το δείκτη κατάστασης του A ως προς τη νόρμα $\|\cdot\|_\infty$ ισχύει,

$$k_\infty(A) \geq \frac{\|A\|_\infty}{\min|a_{ii}|},$$

b) Χωρίς να υπολογίσετε τον A^{-1} αποδείξτε ότι για τον $A = \begin{pmatrix} 1,01 & 0,99 \\ 0,99 & 1,01 \end{pmatrix}$ ισχύει ότι $k_\infty(A) \geq 100$.
(Υπόδειξη: χρησιμοποιείστε το τελευταίο αποτέλεσμα της άσκησης 3.39.)

Λύση.

a) Έστω $A = \begin{pmatrix} a_{11} & \cdots\cdots & & \\ 0 & a_{22} & & \cdots \\ 0 & \cdots & & \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix}$, $a_{ii} \neq 0$. Θα δείξουμε ότι $A^{-1} = \begin{pmatrix} \frac{1}{a_{11}} & \cdots\cdots & & \\ 0 & \frac{1}{a_{22}} & & \cdots \\ 0 & \cdots & & \\ 0 & 0 & \cdots & \frac{1}{a_{nn}} \end{pmatrix}$.

Έστω $B = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{pmatrix}$ και $BA = I_n$.

$$\Rightarrow \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

$$\Rightarrow \left.\begin{array}{c} b_{11}a_{11} = 1 \Rightarrow b_{11} = \frac{1}{a_{11}} \\ b_{21}a_{11} = 0 \Rightarrow b_{21} = 0 \\ \cdots\cdots \\ b_{i1}a_{11} = 1 \Rightarrow b_{i1} = 0 \end{array}\right\} \text{ως άσκηση το υπόλοιπο.}$$

Αρκεί να δείξουμε ότι $\|A^{-1}\|_\infty \geq \dfrac{1}{\min|a_{ii}|}$.

Έχουμε

$$\|A^{-1}\|_\infty = \max_{1\leq i\leq n}(\sum_{j=1}^{n}|(A^{-1})_{ij}|) \geq \max|(A^{-1})_{ii}| = \max|\frac{1}{a_{ii}}| = \frac{1}{\min|a_{ii}|}.$$

b) το 3.39 λέει

αν $B \in \mathbb{R}^{n\times n}$ μη αντιστρέψιμος, τότε

$$\frac{1}{k(A)} \leq \frac{\|A-B\|}{\|A\|} \Rightarrow k_\infty(A) \geq \frac{\|A\|_\infty}{\|A-B\|_\infty}.$$

Έστω $B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ τότε $A-B = \begin{pmatrix} 0,01 & -0,01 \\ -0,01 & 0,01 \end{pmatrix} \Rightarrow \|A-B\|_\infty = 0,02$ και $\|A\|_\infty = 2$.

Άρα $k_\infty(A) \geq \dfrac{\|A\|_\infty}{\|A-B\|_\infty} = \dfrac{2}{0,02} = 100$.

3.44 Έστω $0 \neq x \in \mathbb{R}^n$ η ακριβής λύση του συστήματος $Ax = b$, όπου $A \in \mathbb{R}^{n\times n}$ αντιστρέψιμος και $b \in \mathbb{R}^n$. Έστω $\tilde{x} \in \mathbb{R}^n$ μια προσέγγιση της x και έστω $r = A\tilde{x} - b$ το υπόλοιπο της $\tilde{x}$. Αποδείξτε ότι για κάθε νόρμα $\|\cdot\|$ στον $\mathbb{R}^n$ (και αντίστοιχη φυσική νόρμα πινάκων) ισχύει $\dfrac{\|\tilde{x}-x\|}{\|x\|} \leq \underbrace{k(A)\dfrac{\|r\|}{\|b\|}}$ [2], όπου $k(A) = \|A\|\cdot\|A^{-1}\|$. Πώς ερμηνεύετε την ανισότητα αυτή;

Υπόδειξη.

Έχουμε

- $r = A\tilde{x} - b = A\tilde{x} - Ax - A(\tilde{x}-x) \Rightarrow \tilde{x} - x = A^{-1}r \Rightarrow \|\tilde{x}-x\| \leq \|A^{-1}\|\cdot\|r\|$ (1)
- $b = Ax \Rightarrow \|b\| \leq \|A\|\cdot\|x\|$ (2).

Από (1) και (2) προκύπτει το ζητούμενο.

(Σημείωση: $\begin{matrix} x_1 + x_2 = 2 \\ x_1 + 1,01x_2 = 2,01 \end{matrix}$. Ακριβής λύση $\begin{matrix} x_1 = 1 \\ x_2 = 1 \end{matrix}$. Προσεγγιστική λύση $\begin{matrix} \tilde{x}_1 = 10 \\ \tilde{x}_2 = -8 \end{matrix}$

και $r = \begin{pmatrix} 0 \\ -0,99 \end{pmatrix}$).

Προτεινόμενες Ασκήσεις:

$3.32 - 3.35$, $3.38 - 3.41$, $3.44 - 3.48$.

[2] το υπόλοιπο από μόνο του δεν είναι καλός δείκτης για την ακρίβεια της προσέγγισης

Άσκηση. Θεωρούμε το γραμμικό σύστημ $\begin{matrix} 4,1x_1 + 2,8x_2 = 4,1 \\ 9,7x_1 + 6,6x_2 = 9,7 \end{matrix}$ με ακριβή λύση $x_1 = 1$ και $x_2 = 0$. Αλλάζουμε λίγο το δεύτερο μέλος σε $\begin{pmatrix} 4,11 \\ 9,7 \end{pmatrix}$ και η ακριβής λύση γίνεται $x_1 = 0,34$ και $x_2 = 0,97$. Δείξτε ότι ισχύει το i του Θεωρήματος ως ισότητα ως προς την $\|\cdot\|_1$

$$\frac{\|\delta x\|_1}{\|x\|_1} \leq k_1(A)\frac{\|\delta b\|_1}{\|b\|_1}.$$

Παρεμβολή

Πρόβλημα: Η προσέγγιση συναρτήσεων $f(x)$ (με γνωστές ιδιότητες, γνωστές τιμές, κλπ) ή η "ανασυγκρότηση συναρτήσεων" από πίνακες τιμών $\begin{array}{c|ccc} x_i & \cdot & \cdot & \cdot \\ \hline \psi_i & \cdot & \cdot & \cdot \end{array}$ μέσω απλών συναρτήσεων. $p(x)$ πολυώνυμο, κατά τμήματα πολυώνυμο, ρητή συνάρτηση κλπ (που να υπολογίζεται εύκολα) έτσι ώστε $\boxed{p(x_i) = f(x_i) \cdot \eta \ p(x_i) = \psi_i \ \forall i}$.

Πολυώνυμο Παρεμβολής Lagrange.

Έστω $x_0, x_1, \cdots, x_n \in \mathbb{R}$ ανά δύο διάφορα μεταξύ τους σημεία (διακριτά). Μπορούμε να βρούμε ακριβώς ένα πολυώνυμο $L - i \in \mathbb{P}_n$, $i = 0, 1, \cdots, n$ έτσι ώστε

$$L_i(x_j) = \delta_{ij} = \begin{cases} 1, \, i = j \\ 0, \, i \neq j, \, j = 0, 1, \cdots, n. \end{cases}$$

Το L_i μηδενίζεται στα $x_0, x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n$, επομένως

$$L_i(x) = a_i \prod_{\substack{j=0 \\ j\neq i}}^{n} (x - x_j)$$

όπου $a_i \in \mathbb{R}$. Έχουμε

$$1 = L_i(x_i) = a_i \cdot \prod_{j\neq i}(x_i - x_j) \Rightarrow a_i = \frac{1}{\prod_{j\neq i}(x_i - x_j)}.$$

Συνεπώς,

$$L_i(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)}, \ i = 0, 1, \cdots, n,$$

τα οποία καλούμε *(Στοιχειώδη) πολυώνυμα του Lagrange.*
(ύπαρξη) • Τώρα, το πολυώνυμο παρεμβολής $p \in \mathbb{P}_n$ της f στα σημεία $x_0, x_1, \cdots, x_n$ γράφεται

$$p(x) = \sum_{i=0}^{n} f(x_i) L_i(x).$$

Πράγματι,

$$p(x_k) = \sum_{i=0}^{n} f(x_i) L_i(x_k) = \sum_{i=0}^{n} f(x_i)\delta_{ik} = f(x_k),\ k = 0,\ 1,\ \cdots,\ n.$$

(μοναδικότητα) • Έστω ότι υπάρχει και ένα άλλο πολυώνυμο $p_n^\star \in \mathbb{P}_n$ για το οποίο ισχύει $p_n^\star(x_i) = f(x_i),\ i = 0,\ 1,\ \cdots,\ n$. Ορίζουμε το $d_n(x) = p_n(x) - p_n^\star(x)$. Καταρχήν $d_n \in \mathbb{P}_n$ και ικανοποιεί $d_n(x_i) = p_n(x_i) - p_n^\star(x_i) = f(x_i) - f(x_i) = 0,\ i = 0,\ 1,\ \cdots,\ n$. Το d_n έχει (τουλάχιστον) $n+1$ διακριτές ρίζες. Άρα

$$d_n \equiv 0 \Leftrightarrow p_n \equiv p_n^\star.$$

Θεώρημα. (Σφάλμα της Παρεμβολής)
Έστω $n \in \mathbb{N}_0$, $f \in C^{n+1}[a,\ b]$, $x_0,\ x_1,\ \cdots,\ x_n \in [a,\ b]$ ανά δύο διαφορετικά μεταξύ τους σημεία και $p \in \mathbb{P}_n$ το πολυώνυμο το οποίο παρεμβάλλεται στην f στα σημεία $x_0, x_1, \cdots, x_n$. Τότε ισχύουν

$$\forall x \in [a,\ b],\ \exists f = f(x) \in (a,\ b)$$

έτσι ώστε

(1)• $f(x) - p(x) = \dfrac{f^{n+1}(\xi)}{(n+1)!} \cdot \prod_{i=0}^{n}(x - x_i)$ και

(2)• $\|f - p\|_\infty \leq \max_{a \leq x \leq b} |\prod_{i=0}^{n}(x - x_i)| \dfrac{\|f^{n+1}\|_\infty}{(n+1)!}$, όπου $\|g\|_\infty = \max_{a \leq x \leq b} |g(x)|$ με $\xi \in supp(x_0,\ x_1,\ \cdots,\ x_n,\ x)$.

Απόδειξη.

(1) Αν $x \in \{x_0,\ x_1,\ \cdots,\ x_n\}$ τότε προφανώς ισχύει. Έστω $x \in [a,\ b]$, $x \notin \{x_0,\ x_1,\ \cdots,\ x_n\}$. Ορίζουμε την

$$\varphi(t) = f(t) - p(t) - \frac{f(x) - p(x)}{\prod_{i=0}^{n}(x - x_i)} \cdot \prod_{i=0}^{n}(t - x_i)\ t \in [a,\ b].$$

Προφανώς, $\varphi \in C^{n+1}[a,\ b]$. Επιπλέον,

$$\varphi(x_i) = f(x_i) - p(x_i) = 0,\ i = 0,\ 1,\ \cdots,\ n$$

$$\varphi(x) = f(x) - p(x) - \frac{f(x) - p(x)}{\prod(x - x_i)} \prod(x - x_i) = 0.$$

Επομένως, η φ έχει (τουλάχιστον) $n+2$ διακριτές ρίζες. Εφαρμόζοντας επανειλημμένως το θεώρημα του Rolle έχουμε

Η φ' έχει τουλάχιστον $n+1$ διακριτές ρίζες στο $(a,\, b)$.

Η φ'' έχει τουλάχιστον n διακριτές ρίζες στο $(a,\, b)$.

. . . . . .

Η $\varphi^{(n+1)}$ έχει τουλάχιστον 1 ρίζα στο $(a,\, b)$, έστω την $\xi = \xi(x)$.

Τώρα,

$$\varphi^{(n+1)}(t) = f^{(n+1)}(t) - \frac{f(x)-p(x)}{\prod_{i=0}^{n}(x-x_i)} \cdot (n+1)!$$

οπότε

$$0 = \varphi^{(n+1)}(\xi) = f^{(n+1)}(\xi) - \frac{f(x)-p(x)}{\prod_{i=0}^{n}(x-x_i)} \cdot (n+1)! \Rightarrow f(x)-p(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\prod_{i=0}^{n}(x-x_i).$$

(2) Η απόδειξη αφήνεται ως άσκηση.

Παραδείγματα.

(a) Γραμμική Παρεμβολή $(n=1)$. Έστω $x_0 \le x \le x_1$, δηλαδή $[a,\, b] = [x_0,\, x_1]$ και $b = x_1 - x_0$. Τότε,

$$f(x) - p(x) = (x-x-0)(x-x_1)\frac{f''(\xi)}{2},\ x_0 < \xi < x_1$$

και

$$\|f-p\|_\infty \le \max_{x_0 \le x \le x_1} |(x-x-0)(x-x_1)| \frac{\|f''\|_\infty}{2}.$$

Αφήνεται ως άσκηση $\max_{x_0 \le x \le x_1} |(x-x-0)(x-x_1)| = \dfrac{h^2}{4}$. Άρα,

$$\|f-p\| \le \frac{h^2}{8}\|f''\|_\infty.$$

(b) Τετραγωνική Παρεμβολή $(n=2)$ με ισαπέχοντα σημεία (ομοιόμορφος διαμερισμός) x_0, $x_1 = x_0 + h$, $x_2 = x_1 + h = x_0 + 2h$. Για $x_0 \le x \le x_2$,

$$f(x) - p(x) = (x-x-0)(x-x_1)(x-x_2)\frac{f'''(\xi)}{6}$$

και

$$\|f-p\|_\infty \le \frac{h^3}{9\sqrt{3}}\|f'''\|_\infty.$$

(c) Παρεμβολή με $n+1$ ισαπέχοντα σημεία $x_i = x_0 + ih$, $i = 0, 1, \cdots, n$. Όταν το h είναι μικρό και $x_0 \leq x \leq x_n$ το $\xi = \xi(x)$ είναι περιορισμένο σε ένα μικρό διάστημα και το $f^{(n+1)}(\xi)$ δεν μεταβάλλεται πολύ. Έτσι, η συμπεριφορά του σφάλματος εξαρτάται κυρίως από το γινόμενο $\prod_{i=0}^{n}(x - x_i)$.
Η γραφική παράσταση για έχει ως εξής
σχήμα
Στα άκρα μπορεί το κριτήριο παρεμβολής να έχει μεγάλο σφάλμα.

(Υπερνθύμιση)→ x_i $i = 0, 1, \cdots, n$ διακριτά σημεία.
$p(x_i) = f(x_i)$, $i = 0, 1, \cdots, n$
$p \in \mathbb{P}_n$

$$p(x) = \sum_{i=0}^{n} f(x_i) L_i(x),$$

όπου

$$L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}, \ i = 0, 1, \cdots, n$$

και

$$L_i(x_j) = \delta_{ij}, \ i, j = 0, 1, \cdots, n.$$

Μειονέκτημα αυτής της μορφής είναι ότι αν προσθέσω ακόμη ένα σημείο x_{n+1} πρέπει να κάνω όλους τους υπολογισμούς ξανά.

Παράσταση Πολυωνύμων Παρεμβολής σε μορφή Νεύτωνα.
Έστω $x_0, x_1, \cdots, x_n \in \mathbb{R}$ ανά δύο διαφορετικά μεταξύ τους σημεία και $f(x_0), f(x_1), \cdots, f(x_n) \in \mathbb{R}$. Αν γράψουμε το πολυώνυμο παρεμβολής $p \in \mathbb{P}_n$ για το οποίο ισχύει $p(x_i) = f(x_i)$, $i = 0, 1, \cdots, n$ στη μορφή

$$p(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \cdots + a_n(x - x_0) \cdots (x - x_n - 1)$$

τότε μπορούμε να υπολογίσουμε τους συντελεστές αναδρομικά.

$$p(x_0) = f(x_0) \Rightarrow a_0 = f(x_0)$$
$$p(x_1) = f(x_1) \Rightarrow a_0 + a_1(x - x_0) = f(x_1)$$
$$\cdots\cdots \Rightarrow a_1 = \frac{f(x_1) - a_0}{x_1 - x_0} = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

Σύγκλιση του Πολυωνύμου Παρεμβολής.

Έστω ο τριγωνικός πίνακας σημείων παρεμβολής

$$\begin{array}{ccccc} x_0^{(0)} & & & & p_0 \\ x_0^{(1)} & x_1^(1) & & & p_1 \\ x_0^{(2)} & x_1^{(2)} & x_2^{(2)} & & p_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_0^{(n)} & x_1^{(n)} & x_2^{(n)} & \cdots & p_n \end{array}$$

και $p_n(x) = p_n(f, x_0^{(n)}, x_1^{(n)}, \cdots, x_2^{(n)}, x)$ το πολυώνυμο παρεμβολής για τη συνάρτηση f στα σημεία $x_0^{(n)}, x_1^{(n)}, \cdots, x_2^{(n)}$. Λέμε ότι έχουμε σύγκλιση της παρεμβολής κατά Lagrange αν $p_n(x) \to f(x)$ καθώς το $n \to \infty$.
Από την εκτίμηση του σφάλματος (2) έχουμε

$$\|f - p_n\|_\infty \leq \frac{1}{(n+1)!} \max_{a \leq x \leq b} |\prod_{i=0}^{n}(x - x_i)| \cdot \|f^{n+1}\|_\infty.$$

Αν τα σημεία $x_0, x_1, \cdots, x_n$ είναι ισαπέχοντα (ομοιόμορφος διαμερισμός), δηλαδή με $h = \dfrac{b-a}{n}$ έχουμε $x_i = a + ih$, $i = 0, 1, \cdots, n$ τότε μπορεί να αποδειχθεί ότι

$$\max_{a \leq x \leq b} |\prod_{i=0}^{n}(x - x_i)| \leq \frac{n!}{4} h^{n+1}.$$

Επομένως,

$$\|f - p_n\|_\infty \leq \frac{h^{n+1}}{4(n+1)} \|f^{(n+1)}\|_\infty.$$

Όμως ισχύει:

$$[\forall f \in C[a, b] \ \lim_{n \to \infty} \|f - p_n\| = 0]; \quad \text{OXI}$$

Παράδειγμα του Runge:

$f(x) = \dfrac{1}{1+x^2}, \ -5 \leq x \leq 5$ σχήμα

$x_i^{(n)} = -5 + i\dfrac{10}{h}, \ i = 0, 1, \cdots, n$

$$\lim_{n \to \infty} |f(x) - p_n(x)| = \begin{cases} 0, \text{ αν } |x| < 3,633... \\ \infty, \text{ αν } |x| > 3,633... \end{cases}$$

$\lim_{n \to \infty} \|f - p_n\| = \infty$.

Παράδειγμα του Bernstein:
$f(x) = |x|,\ -1 \le x \le 1$ σχήμα
$x_i^{(n)} = -1 + i\dfrac{2}{n},\ i = 0,\, 1,\, \cdots,\, n$
$\lim_{n\to\infty} \|f - p_n\| = \infty\ \forall x \in [-1,\, 1]$ εκτός των $-1,\, 0,\, 1$ σημείων.

Θεώρημα.(Faber 1914)
Για κάθε πίνακα σημείων παρεμβολής $x_{ni} \in [-1,\, 1],\, i = 0,\, 1,\, \cdots,\, n$ υπάρχει συνάρτηση $f \in C[-1,\, 1]$ τέτοια ώστε αν $p_n \in \mathbb{P}_n$ είναι το πολυώνυμο το οποίο παρεμβάλλεται στην f στα σημεία $x_{n0},\, x_{n1},\, \cdots,\, x_{nn}$ τότε ισχύει

$$\lim_{n\to\infty} \sup \|f - p_n\|_\infty = \infty.$$

Πολυώνυμα του Chebyshev πρώτου είδους.

$$T_n(x) = \cos(n \cdot \arccos x)),\ x \in [-1,\, 1]$$

δηλαδή,

$$T_n(\cos\theta) = \cos n\theta,\ 0 \le \theta \le \pi.$$

Επομένως,

$$T_0(x) = 1,\ T_1(x) = x,$$
$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x),\ n = 1,\, 2,\, \cdots$$

π.χ. $T_2(x) = 2x^2 - 1,\ T_3(x) = 4x^3 - 3x.$
Ρίζες του $T_n(x)$

$$x_i = 8\frac{2i-1}{2n} \cdot \pi,\ i = 1,\, 2,\, \cdots,\, n.$$

Έχουμε $T_n(x) = 2^{n-1} \cdot x^n + \cdots\cdots$, οπότε ορίζουμε

$$\widehat{T}_n(x) = \frac{1}{2^{n-1}} T_n(x)$$

το πολυώνυμο Chebyshev πρώτου είδους με συντελεστή μεγιστοβάθμιου όρου τη μονάδα.

Θεώρημα.
Για ένα οποιοδήποτε πολυώνυμο $\hat{p}_n$, βαθμού n, με συντελεστή μεγιστοβάθμιου όρου τη μονάδα, ισχύει

$$\|p_n\|_\infty = \max_{-1\le x\le 1} |\hat{p}_n(x)| \ge \underbrace{\max_{-1\le x\le 1} |\widehat{T}_n(x)| = \frac{1}{2^{n-1}}}_{\text{ως άσκηση}}.$$

. . . . . .

$$\|f - p_n\|_\infty \leq \max_{-1\leq x\leq 1} |\prod_{i=0}^{n}(x - x_i)| \frac{\|f^{(n+1)}\|_\infty}{(n+1)!}$$

$$\max_{-1\leq x\leq 1} |\hat{p}_{n+1}(x)| \geq \max_{-1\leq x\leq 1} |\widehat{T}_{n+1}(x)| = \frac{1}{2^n}.$$

Έτσι αν τα $x_i = \cos\dfrac{(2i+1)}{2(n+1)}\cdot\pi,\ i = 0,\ 1,\ \cdots,\ n$ οι ρίζες του $\widehat{T}_{n+1}$, τότε

$$\max_{-1\leq x\leq 1} |\prod_{i=0}^{n}(x - x_i)| = \frac{1}{2^n}$$

οπότε

$$\forall f \in C^{n+1}[-1,\ 1]\ \|f - p_n\|_\infty \leq \frac{1}{(n+1)!} 2^n \cdot \|f^{(n+1)}\|_\infty.$$

Επιπλέον, για αυτά τα σημεία παρεμβολής ισχύει

$$\forall f \in C^1[-1,\ 1]\ \lim_{n\to\infty} \|f - p_n\|_\infty = 0.$$

<u>υπενθύμιση</u>:

$$\|f - p_n\|_\infty \leq \frac{h^{n+1}}{4(n+1)} \|f^{(n+1)}\|_\infty$$

Για $n = 1:\ h = b - a,\ \|f - p_1\|_\infty \leq \dfrac{h^2}{8}\|f''\|_\infty$ (1) σχήμα
Για $n = 2:\ h = \dfrac{b-a}{2},\ \|f - p_2\|_\infty \leq \dfrac{h^3}{12}\|f'''\|_\infty$ (2) σχήμα
Για $n = 3:\ h = \dfrac{b-a}{3},\ \|f - p_3\|_\infty \leq \dfrac{h^4}{16}\|f^{(4)}\|_\infty$ (3) σχήμα
(1), $\Rightarrow \|f - p_1\|_\infty \leq \dfrac{(b-a)^2}{8}\|f''\|_\infty$
(2), $\Rightarrow \|f - p_2\|_\infty \leq \dfrac{(b-a)^3}{96}\|f'''\|_\infty$
(3), $\Rightarrow \|f - p_3\|_\infty \leq \dfrac{(b-a)^4}{1296}\|f^{(4)}\|_\infty$

Για μικρό διάστημα $[a,\ b]$ και λίγα σημεία (π.χ. $n = 3$) το σφάλμα προσέγγισης της f δεν είναι ιδιαίτερα μεγάλο.

Παρεμβολή με Splines.
Παρεμβολή με τμηματικά γραμμικές συναρτήσεις.
Έστω $\Delta : a = x_0 < x_1 < \cdots < x_n = b$ μία διαμέριση του διαστήματος $[a, b]$ θέλουμε να κατασκευάσουμε ένα τμηματικά γραμμικά πολυώνυμο $s(x)$.

σχήμα

$s(x_i) = f(x_i),\ i = 0, 1, \cdots, n.$
Έχουμε,

$$s(x) = f(x_i) + \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}(x - x_i),\ x \in [x_i, x_{i+1}]$$

Θεώρημα.
Έστω $f \in C^2[a, b]$, $\Delta : a = x_0 < x_1 < \cdots < x_n = b$ μία διαμέριση του $[a, b]$ και s ένα τμηματικά γραμμικό πολυώνυμο που παρεμβάλλεται στην f στα σημεία $x_0, x_1, \cdots, x_n$. Αν $h = \max_{0\le i\le n-1}(x_{i+1} - x_i)$ όπου h λεπτότητα διαμέρισης ή πλάτος, τότε

$$\|f - s\|_\infty \le \frac{h^2}{8}\|f''\|_\infty.$$

Απόδειξη.
Έχουμε,

$$\|f - s\|_\infty = \max_{a\le x\le b}|f(x) - s(x)| = |f(\bar{x}) - s(\bar{x})| \underset{\bar{x}\in[x_k, x_{k+1}]}{=} \max_{x_k\le x\le x_{k+1}}|f(x) - s(x)|$$

$$\le \frac{(x_{k+1} - x_k)^2}{8}\max_{x_k\le x\le x_{k+1}}\|f''(x)\| \le \frac{h^2}{8}\max_{a\le x\le b}|f''(x)| = \frac{h^2}{8}\|f''\|_\infty$$

Παράδειγμα
Να βρεθεί ο αριθμός των σημείων ομοιόμορφης διαμέρισης ώστε το σφάλμα στην παρεμβολή της συνάρτησης $f(x) = e^x$, $x \in [0, 1]$ με μία γραμμική spline, να είναι $\le 10^{-6}$.
Λύση.
Έχουμε,

$$|e^x - s(x)| \le \max_{0\le x\le 1}|e^x - s(x)| \le \frac{h^2}{8}\max_{0\le x\le 1}|e^x| = \frac{h^2}{8}e$$

όπου $h = \dfrac{b-a}{n} = \dfrac{1}{n}$.
Πρέπει

$$\frac{h^2}{8}e \le 10^{-6} \Rightarrow \frac{e}{8n^2} \le 10^{-6} \Rightarrow \cdots \Rightarrow n \ge 1000\sqrt{\frac{e}{8}} \Rightarrow n = 583\,\text{υποδιαστήματα, (δηλαδή 584 σημεία).}$$

Μειονεκτήματα γραμμικής splines.

- χρειάζονται πολλά σημεία για απλή ακρίβεια (10^{-6})
- χαμηλή τάξη σφάλματος (h^2)
- η πρώτη παράγωγος δεν προσεγγίζεται συνεχώς

Παρεμβολή με τμηματικά κυβικές συναρτήσεις.
Έστω $\Delta : a = x_0 < x_1 < \cdots < x_n = b$ μία διαμέριση του διαστήματος $[a, b]$ θέλουμε να κατασκευάσουμε ένα τμηματικά κυβικό πολυώνυμο $s(x)$ έτσι ώστε

1. Το $s(x)$ να είναι (γενικά) κυβικό πολυώνυμο σε κάθε υποδιάστημα $[x_i, x_{i+1}]$, $i = 0, 1, \cdots, n$,
2. $s(x) \in C^2[a, b]$
3. $s(x_i) = f(x_i)$, $i = 0, 1, \cdots, n$

σχήμα

Αν $s(x)|_{x\in[x_i, x_{i+1}]} = s_i(x)$, $i = 0, 1, \cdots, n$ τότε

$$s_i(x) = a_i + b_i x + c_i x^2 + d_i x^3.$$

Επομένως, υπάρχουν $4n$ σταθερές (παράμετροι) που πρέπει να προσδιοριστούν.

Τώρα $\left.\begin{array}{c} s_i(x_i) = f(x_i),\ i = 0, 1, \cdots, n-1 \\ s_i(x_{i+1}) = f(x_{i+1}) \end{array}\right\}$ $2n$ συνθήκες.
(με τις παραπάνω συνθήκες έχω καλύψει τη συνέχεια της s).

$$s_i'(x_{i+1}) = s_{i+1}'(x_{i+1})\ i = 0, 1, \cdots, n-2 \to n-1 \text{ συνθήκες}$$
$$s_i''(x_{i+1}) = s_{i+1}''(x_{i+1})\ i = 0, 1, \cdots, n-2 \to n-1 \text{ συνθήκες}$$

Σύνολο $4n - 2$ συνθήκες.
Άρα χρειαζόμαστε δύο επιπλέον συνθήκες που μπορεί να είναι

$$s'(a) = f'(a),\ s'(b) = f'(b)$$

ή

$$s''(a) = f''(a),\ s''(b) = f''(b)$$

ή

$$s''(a) = 0,\ s''(b) = 0.$$

Βασικά βήματα της κατασκευής.
Αν γνωρίζουμε τα $s_i''(x_i) = z_i$, $x \in [x_i, x_{i+1}]$, $i = 0, 1, \cdots, n-1$ τότε μπορούμε να υπολογίσουμε την s.

σχήμα

$$h_i = x_{i+1} - x_i$$
$$2i,\ i = 0, 1, \cdots, n$$

Πράγματι,

$$s_i(x) = \frac{z_{i+1}}{6h_i}(x - x_i)^3 + \frac{z_i}{6h_i}(x_{i+1} - x)^3 + (\frac{f(x_{i+1})}{h_i} - \frac{z_{i+1}h_i}{6})(x - x_i) + (\frac{f(x_i)}{h_i} - \frac{z_i h_i}{6})(x_{i+1} - x)$$

με $x_i \le x \le x_{i+1},\ i = 0, 1, \cdots, n-1$.
Χρησιμοποιώντας τη συνέχεια της s', δηλαδή,

$$s'_{i-1}(x_i) = s'_i(x_i),\ i = 0, 1, \cdots, n-1$$

παίρνουμε ένα τριδιαγώνιο γραμμικό σύστημα στο οποίο έχουμε n αγνώστους z_i και $n-1$ εξισώσεις.

$$h_{i-1}z_{i-1} + 2(h_{i-1} + h_i)z_i + h_i z_{i+1} = \frac{6[f(x_{i+1}) - f(x_i)]}{h_i} - \frac{6[f(x_i) - f(x_{i-1})]}{h_{i-1}},\ 1 \le i \le n-1.$$

Με τις δύο επιπλέον συνθήκες που πήραμε στην αρχή το σύστημα έχει ακριβώς μία λύση.
Θεώρημα.
Έστω $\Delta : a = x_0 < x_1 < \cdots < x_n = b$ μία διαμέριση του διαστήματος
$[a, b]$, $b = \max_{x_0 \le i \le x_{n-1}}(x_{i+1} - x_i)$,

$$M = \frac{h}{\min(x_{i+1} - x_i)},\ f \in C^4[a, b]$$

και s η κυβική spline για την οποία ισχύει

$$s(x_i) = f(x_i),\ i = 0, 1, \cdots, n,\ f'(x_0) = s'(x_0),\ f'(x_n) = s'(x_n)$$

τότε υπάρχουν σταθερές C_m, $m = 0, 1, 2, 3$ ανεξάρτητες των f και h, τέτοιες ώστε

$$\|f^{(m)} - s^{(m)}\|_\infty \le C_m h^{4-m}\|f^{(4)}\|_\infty,\ m = 0, 1, 2, 3.$$

Παρατήρηση: Η προηγούμενη εκτίμηση ισχύει με τις ακόλουθες τιμές σταθερών,

$$c_0 = \frac{5}{384},\ c_1 = \frac{1}{24},\ c_2 = \frac{3}{8},\ c_3 = \max\{2,\ \frac{1}{2}(M + \frac{1}{M})\},$$

από τις οποίες βέλτιστες είναι οι δύο πρώτες.

Ασκήσεις στην Παρεμβολή (βίβλιο Ακρίβη-Δουγαλή).

4.4 Έστω $p \in \mathbb{P}_3$ με $p(x_i) = \log x_i$, $x_i = i+1$, $i = 0, 1, 2, 3$. Αποδείξτε ότι η συνάρτηση ε, $\varepsilon(x) = \ln(x) - p(x)$ (συνάρτηση σφάλματος) έχει στο διάστημα $[1, 4]$ ακριβώς τέσσερις ρίζες.

Λύση.

Έστω $f(x) = \ln(x)$, $x \in [1, 4]$, το p είναι το πολυώνυμο παρεμβολής της f στα σημεία x_i, $i = 0, 1, 2, 3$ το οποίο υπάρχει και είναι μοναδικό.

Έχουμε,

$$\varepsilon(x) = \ln(x) - p(x) = f(x) - p(x) = (x - x_0)(x - x_1)(x - x_2)(x - x_3)\frac{f^{(4)}(\xi)}{4!},$$

$$f'(x) = \frac{1}{x},\ f''(x) = -\frac{2}{x^2},\ f'''(x) = \frac{2}{x^3},\ f^{(4)}(x) = -\frac{6}{x^4}.$$

Συνεπώς,

$$\varepsilon(x) = (x - x_0)(x - x_1)(x - x_2)(x - x_3)\big(-\frac{6}{\xi^4 \cdot 4!}\big),\ 1 < \xi < 4,\ x \in [1, 4].$$

Άρα έχει 4 ρίζες $x_0, x_1, x_2, x_3 \in [1, 4]$.

4.5 Έστω $p \in \mathbb{P}_3$ τέτοιο ώστε $p(i) = e^i$, $i = 1, 2, 3, 4$. Αποδείξτε ότι $\forall x \in (2, 3)$ $e^x > p(x)$.

Λύση.

Το p είναι το πολυώνυμο παρεμβολής της συνάρτησης $f(x) = e^x$ στα σημεία 1, 2, 3, 4 το οποίο υπάρχει και είναι μοναδικό. Επίσης $f \in C^4[1, 4]$. Οπότε

$$e^x - p(x) = (x-1)(x-2)(x-3)(x-4)f^{(4)}(\xi) \cdot \frac{1}{4!} = (x-1)(x-2)(x-3)(x-4)\frac{e^\xi}{24}.$$

Για

$$x \in (2, 3) :\ x - 1 > 0,\ x - 2 > 0,\ x - 3 < 0,\ x - 4 < 0,\ e^\xi > 0.$$

Άρα

$$e^x - p(x) > 0 \Leftrightarrow e^x > p(x) \forall x \in (2, 3).$$

Θέμα 4^0 (Ιούνιος 1997).

(a) Υπολογίστε το πολυώνυμο παρεμβολής $p \in \mathbb{P}_2$ που παρεμβάλλεται στις τιμές της $f(x) = \ln x$ στα σημεία 2, 3 και 4 και δείξτε ότι αν $\varepsilon(x) = f(x) - p(x)$, τότε $-\frac{1}{64} \le \varepsilon(3,5) \le -\frac{1}{512}$ (χωρίς ακριβή υπολογισμό του σφάλματος).

Λύση.

Πολυώνυμο παρεμβολής σε μορφή Νεύτωνα

$$p(x) = a_0 + a_1(x-2) + a_2(x-2)(x-3)$$

$$p(2) = \ln 2 \Leftrightarrow a_0 = \ln 2$$

$$p(3) = \ln 3 \Leftrightarrow a_0 + a_1 = \ln 3 \Rightarrow a_1 = \ln 3 - \ln 2$$

$$p(4) = \ln 4 \Leftrightarrow a_0 + 2a_1 + 2a_2 = \ln 4 \Rightarrow a_2 = \frac{\ln 4 - 2\ln 3 + \ln 2}{2}.$$

Συνεπώς,

$$p(x) = \ln 2 + (\ln 3 - \ln 2)(x-2) + \frac{\ln 4 - 2\ln 3 + \ln 2}{2}(x-2)(x-3) \Rightarrow$$

$$p(x) = \frac{\ln 4 - 2\ln 3 + \ln 2}{2}x^2 - \frac{5\ln 4 - 12\ln 3 + 7\ln 2}{2}x + 3ln4 - 8\ln 3 + 6\ln 2.$$

Έχουμε, $f \in C^3[2,\, 4]$

$$\varepsilon(x) = f(x) - p(x) = (x-2)(x-3)(x-4) \cdot \frac{f'''(\xi)}{3!},\ 2 < \xi < 4$$

$$\Rightarrow \varepsilon(x) = (x-2)(x-3)(x-4) \cdot \frac{2}{\xi^3}\frac{1}{3!}.$$

Άρα

$$\varepsilon(3,5) = (3,5-2)(3,5-3)(3,5-4) \cdot \frac{2}{\xi^3}\frac{1}{3!} = -\frac{3}{2}\frac{1}{2} \cdot \frac{1}{2}\frac{2}{\xi^3}\frac{1}{3!} = -\frac{1}{8} \cdot \frac{2}{\xi^3}.$$

Η $\dfrac{2}{\xi^3}$ είναι φθίνουσα. Άρα,

$$\frac{1}{2^3} \geq \frac{2}{\xi^3} \geq \frac{1}{4^3} \Rightarrow$$

$$\Rightarrow -\frac{1}{8}\frac{1}{2^3} \leq -\frac{1}{8}\frac{2}{\xi^3} \leq -\frac{1}{8}\frac{1}{4^3}$$

$$\Rightarrow -\frac{1}{64} \leq \varepsilon(3,5) \leq -\frac{1}{512}.$$

(*b*) Θεωρήστε την $f:\, [0,\, 4] \to \mathbb{R}:\, f(x) = \begin{cases} x-1,\, 1 \leq x \\ 3-x,\, 2 \leq x \leq 3 \\ 0,\, \text{διαφορετικά} \end{cases}$.

Έστω s η κυβική spline παρεμβολής της f στα σημεία $x_i = i$, $0 \leq i \leq 4$, με συνοριακές συνθήκες δευτέρων παραγώγων στα άκρα 0 και 4. Μία από τις παρακάτω είναι η σωστή γραφική παράσταση της $s(x)$. Προσδιορίστε την δικαιολογώντας πλήρως της απόρριψη των υπολοίπων.

(*i*) $s(x) = f(x)$
σχήμα

(*ii*) $s \in C^2[0,\, 4]$
σχήμα

(iii) $s \in C^2[0,\, 4]$	(iv) $s(x) = (x-1)(3-x)$
σχήμα	σχήμα

Λύση.
Η (i) απορρίπτεται διότι δεν είναι $C^2[0,\, 4]$.
Η (ii) απορρίπτεται για τον ίδιο λόγο.
Για την (iii):
Στο διάστημα $[1,\, 2]$: $s(x) = ax^3 + bx^2 + cx + d$.
Έχουμε

$$\left.\begin{array}{c} \lim_{x\to 1^-} s(x) = \lim_{x\to 1^+} s(x) \Rightarrow 0 = a+b+c+d \\ \lim_{x\to 1^-} s'(x) = \lim_{x\to 1^+} s'(x) \Rightarrow 0 = 3a+2b+c \\ \lim_{x\to 1^-} s''(x) = \lim_{x\to 1^+} s''(x) \Rightarrow 0 = 6a+2b \end{array}\right\} \Rightarrow \begin{array}{c} a = a \\ b = -3a \\ c = 3a \\ d = -a \end{array}$$

Συνεπώς, $s(x) = a(x-1)^3$. Άρα, $s(x) = a(x-1)^3$, $x \in [1,\, 2]$. Αντίστοιχα,

$$s(x) = b(x-3)^3,\ x \in [2,\, 3].$$

Όμως πρέπει

$$\left.\begin{array}{c} \lim_{x\to 2^-} s(x) = \lim_{x\to 2^+} s(x) = 1 \Rightarrow a = -b = 1 \\ \lim_{x\to 2^-} s'(x) = \lim_{x\to 2^+} s'(x) \Rightarrow 3a = 3b \Rightarrow a = b \end{array}\right\} \text{Άτοπο.}$$

Άρα η (iii) απορρίπτεται.

4.21

$$f(x) = \begin{cases} 0, & 0 \le x \le 1 \\ (x-1)^4, & 1 < x \le 2 \end{cases}$$

(a) προσεγγίστε με κατά τμήματα πολυώνυμο

$$p(x) = \begin{cases} 0, & 0 \le x \le 1 \\ a + b(x-1) + c(x-1)^2 + d(x-1)^3, & 1 < x \le 2 \end{cases}$$

έτσι ώστε

$$p \in C^1[0,\, 2],\ p(0) = f(0),\ p'(0) = f'(0),\ p(1) = f(1),\ p(2) = f(2),\ p'(2) = f'(2).$$

(b) Συμπίπτει η p με την κυβική spline;

Προτεινόμενες ασκήσεις:
$4.3 - 4.6$, 4.15, 4.18, 4.21, 4.25.

Αριθμητική Ολοκλήρωση

Έστω $[a, b] \subset \mathbb{R}$ και $f : [a, b] \to \mathbb{R}$ φραγμένη και (κατάρχήν) ολοκληρώσιμη. Θέλουμε να προσεγγίσουμε το $\int_a^b f(x)dx$.
Αν F είναι μια παράγουσα της f, τότε

$$\int_a^b f(x)dx = F(b) - F(a).$$

Όμως:

- Μία παράγουσα F σπάνια μπορεί να υπολογιστεί αναλυτικά.

- Πολλές φορές η παράγουσα F μπορεί να είναι ιδιαίτερα περίπλοκη.

Παράδειγμα, μία παράγουσα F της $f(x) = \dfrac{1}{1+x^4}$ είναι η

$$F(x) = \frac{1}{4\sqrt{2}} \ln \frac{x^2 + x\sqrt{2} + 1}{x^2 - x\sqrt{2} + 1} - \frac{1}{2\sqrt{2}} \cdot \big(\arctan \frac{x}{\sqrt{2} - x} + \arctan \frac{x}{\sqrt{2} + x} \big).$$

Η προσέγγιση του ολοκληρώματος γίνεται με ένα άθροισμα της μορφής

$$Q_{n+1}(f) = \sum_{i=0}^{n} w_i f(x_i)$$

όπου $x_i \in [a, b]$ →κόμβοι και w_i →βάρη.

Ο τύπος του Τραπεζίου.

Ο τύπος αυτός προκύπτει από την εφαρμογή γραμμικής παρεμβολής της f στα σημεία $x_0 = a$, $x_1 = b$. Έτσι $h = b - a$ και

$$f(x) = \frac{x-b}{a-b} f(a) + \frac{x-a}{b-a} f(b) + (x-a)(x-b)\frac{f''(\xi_x)}{2}$$

όπου $f \in C^2[a, b]$ (τύπος του σφάλματος από το πολυώνυμο Lagrange και $\xi = \xi(x) \in (a, b)$.

σχήμα!!!

Οπότε

$$\int_a^b f(x)dx = \big(\int_a^b \frac{x-b}{a-b} dx \big) f(a) + \big(\int_a^b \frac{x-a}{b-a} dx \big) f(b) + \int_a^b (x-a)(x-b)\frac{f''(\xi_x)}{2} dx$$

$$\Rightarrow \int_a^b f(x)dx = -\frac{a-b}{2}f(a) + \frac{b-a}{2}f(b) - \frac{1}{2}\int_a^b (x-a)(x-b)f''(\xi_x)dx.$$

Δεδομένου ότι $(x-a)(b-x) \geq 0$, $x \in [a,\, b]$ τι Θεώρημα Μέσης Τιμής για ολοκληρώματα[3] δίνει

$$\int_a^b f(x)dx = \frac{b-a}{2}[f(a)-f(b)] - \frac{f''(\zeta)}{2}\int_a^b (x-a)(b-x)dx = \frac{b-a}{2}[f(a)-f(b)] - \frac{f''(\zeta)}{2}\frac{(b-a)^3}{6}.$$

Άρα τελικά έχουμε,

$$\int_a^b f(x)dx = \frac{b-a}{2}[f(a)-f(b)] - f''(\zeta)\frac{(b-a)^3}{12},\ \zeta \in (a,\, b).$$

Ο παραπάνω τύπος είναι ο τύπος του τραπεζίου,
$\int_a^b f(x)dx$ →ολοκλήρωμα
$\frac{b-a}{2}[f(a)-f(b)]$ →προσέγγιση
$f''(\zeta)\frac{(b-a)^3}{12}$ →σφάλμα (μικρό για διαστήματα μικρού μήκους).

Αν τώρα θεωρήσουμε έναν ομοιόμορφο διαμερισμό του $[a,\, b]$ με βήμα $h = \frac{b-a}{n}$, $x_i = a + ih$, $i = 0,\, 1,\, \cdots,\, n$ και εφαρμόσουμε σε κάθε ένα από τα διαστήματα $[x_i,\, x_{i+1}]$ τον τύπο του τραπεζίου, έχουμε

$$\int_a^b f(x)dx = \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} f(x)dx =$$

$$= \sum_{i=0}^{n-1}\frac{x_{i+1}-x_i}{2}[f(x_i)+f(x_{i+1})] - \frac{(x_{i+1}-x_i)^3}{12}f''(\xi_i),\ \xi_i \in (x_i,\, x_{i+1})$$

$$= \sum_{i=0}^{n-1}\frac{h}{2}[f(x_i)+f(x_{i+1})] - \frac{h^3}{12}f''(\xi_i)$$

$$= \frac{h}{2}\sum_{i=0}^{n-1}[f(x_i)+f(x_{i+1})] - \frac{h^3}{12}\sum_{i=0}^{n-1}f''(\xi_i)$$

$$= \frac{h}{2}[f(x_0)+2f(x_1)+\cdots+2f(x_{n-1})+f(x_n)] - \frac{h^3}{12}\sum_{i=0}^{n-1}f''(\xi_i)$$

$$= h[\frac{1}{2}f(x_0)+f(x_1)+\cdots+f(x_{n-1})+\frac{1}{2}f(x_n)] - \frac{h^3}{12}\sum_{i=0}^{n-1}f''(\xi_i)$$

[3]ΘΜΤ για ολοκληρώματα: $g(x) \geq 0$ ή $g(x) \leq 0$, $f \in C[a,\, b]$ τότε υπάρχει η τέτοιο ώστε

$$\int_a^b g(x)f(x)dx = f(\eta)\int_a^b g(x)dx$$

όπου $\eta \in (a,\, b)$.

$$= h[\frac{1}{2}f(x_0) + f(x_1) + \cdots + f(x_{n-1}) + \frac{1}{2}f(x_n)] - \frac{(b-a)}{12}h^2\frac{1}{\eta}\sum_{i=0}^{n-1} f''(\xi_i).$$

Τώρα δεδομένου ότι $f \in C^2[a,\, b]$

$$\min_{a\le x\le b} f''(x) \le f''(\xi_i) \le \max_{a\le x\le b} f''(x)$$

$$n \cdot \min f''(x) \le \sum_{i=0}^{n-1} f''(\xi_i) \le n \cdot \max f''(x)$$

$$\min f''(x) \le \frac{1}{n}\sum_{i=0}^{n-1} f''(\xi_i) \le \max f''(x).$$

Από Θεώρημα Ενδιάμεσης Τιμής

$$\exists \xi \in (a,\, b) \text{ τέτοιο ώστε } \frac{1}{n}\sum_{i=0}^{n-1} f''(\xi_i) = f''(\xi).$$

Συνεπώς

$$\int_a^b f(x)dx = h[\frac{1}{2}f(x_0) + f(x_1) + \cdots + f(x_{n-1}) + \frac{1}{2}f(x_n)] - \frac{(b-a)}{12}h^2 f''(\xi),\ \xi \in (a,\, b),$$

ο οποίος είναι ο *Σύνθετος τύπος του τραπεζίου*.

$$Q_{n+1}^T(f) = h[\frac{1}{2}f(x_0) + f(x_1) + \cdots + f(x_{n-1}) + \frac{1}{2}f(x_n)]$$

$$R_{n+1}^T(f) = -\frac{b-a}{12}h^2 f''(\xi),$$

από όπου έπεται αμέσως

$$|R_{n+1}^T(f)| \le \frac{b-a}{12}h^2 f''(\xi) \Rightarrow \lim_{n\to\infty} R_{n+1}^T(f) = 0,\ f \in C^2[a,\, b].$$

Αν η f είναι γραμμική συνάρτηση, τότε ο τύπος του τραπεζίου δίνει την ακριβή τιμή του ολοκληρώματος. Λέμε ότι ο τύπος του τραπεζίου έχει βαθμό ακριβείας 1.
Γενικά, ο *μέγιστος* βαθμός των πολυωνύμων που ένας τύπος αριθμητικής ολοκλήρωσης ολοκληρώνει ακριβώς (με σφάλμα 0) ονομάζεται βαθμός ακριβείας του τύπου αυτού.