Νικόλαος Δ. Ατρέας

Αριθμητική Ανάλυση

Α.Π.Θ.

Τμήμα πληροφορικής Α.Π.Θ.

Θεσσαλονίκη 2007

Περιεχόμενα

ΕΙΣΑΓΩΓΗ

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

ΦΥΣΙΚΟ ΠΡΟΒΛΗΜΑ

ΣΧΕΔΙΑΣΗ ΜΑΘΗΜΑΤΙΚΟΥ ΜΟΝΤΕΛΟΥ

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

ΜΕΛΕΤΗ ΜΑΘΗΜΑΤΙΚΟΥ ΜΟΝΤΕΛΟΥ

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

ΥΠΟΛΟΓΙΣΜΟΣ ΣΦΑΛΜΑΤΟΣ ΥΠΟΛ. ΜΟΝΤΕΛΟΥ

ΥΠΟΛΟΓΙΣΤΙΚΟ ΜΟΝΤΕΛΟ

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

ΑΛΓΟΡΙΘΜΟΣ ΥΛΟΠΟΙΗΣΗΣ ΥΠΟΛΟΓΙΣΤΙΚΟΥ ΜΟΝΤΕΛΟΥ

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

ΜΕΛΕΤΗ ΑΛΓΟΡΙΘΜΙΚΩΝ ΣΦΑΛΜΑΤΩΝ

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

↓

ΠΡΟΣΕΓΓΙΣΤΙΚΗ ΛΥΣΗ

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

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

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

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

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

ΚΕΦΑΛΑΙΟ 1

ΑΡΙΘΜΗΤΙΚΗ ΠΕΠΕΡΑΣΜΕΝΗΣ ΑΚΡΙΒΕΙΑΣ - ΣΦΑΛΜΑΤΑ

§ 1.1 *Αναπαράσταση αριθμών σε οποιαδήποτε βάση*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

και

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

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

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

και

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

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

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

Πολλαπλασιάζουμε και τα δύο μέλη με p και έχουμε:

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

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

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

και

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

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

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

και

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

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

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

§ 1.2 *Αριθμοί μηχανής*

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

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

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

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

Έχουμε λοιπόν:

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

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

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

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

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

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

όπου

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

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

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

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

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

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

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

§ *1.3 Ιδιότητες αριθμών μηχανής*

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

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

Απόδειξη Επειδή:

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

πολλαπλασιάζοντας με p^e παίρνουμε το ζητούμενο. □

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

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

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

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

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

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

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

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

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

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

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

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

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

Σε κάθε ένα από τα υποδιαστήματα

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

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

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

Λύση Προφανώς οι αριθμοί μηχανής έχουν τη μορφή $x = \pm\,(0.\, b_1 b_2 b_3)\, 2^e$, όπου $b_i = 0,1,\ b_1 \neq 0.$ Δοθέντος εκθέτη e και λαμβάνοντας υπόψην ότι $b_1 = 1$ έχουμε:

$$(0.\, 1 b_2 b_3)_2 = \{(0.\, 100)_2, (0.\, 101)_2, (0.\, 110)_2, (0.\, 111)_2\} = \left\{\frac{1}{2}, \frac{5}{8}, \frac{3}{4}, \frac{7}{8}\right\}.$$

Αρα το σύνολο των αριθμών μηχανής είναι:

$$x = \pm\,(0.\, b_1 b_2 b_3)_2\ 2^e = \pm\left\{\frac{2^e}{2}, \frac{5\ 2^e}{8}, \frac{3\ 2^e}{4}, \frac{7\ 2^e}{8} :\ e = -2, ..., 2\right\},$$

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

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

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

§ 1.4 Σφάλματα στρογγύλευσης και αποκοπής

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

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

και ***σχετικό σφάλμα*** την ποσότητα

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

§ 1.5 Διαδιδόμενα σφάλματα σε αριθμητικούς υπολογισμούς

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

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

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

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

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

Έστω:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Τέλος παρατηρούμε ότι

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

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

Παράδειγμα 4 *(Μελέτη σφαλμάτων στον υπολογισμό αθροισμάτων)*

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

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

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

έχουμε:

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

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

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

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

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

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

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

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

αγνοώντας ως αμελητέο τον τελευταίο όρο. Αναγωγικά υπολογίζουμε:

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

ή

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

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

άρα:

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

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

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

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

§ 1.6 Ευστάθεια αλγορίθμων

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

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

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

Ο δείκτης κατάστασης ενός προβλήματος ορίζεται ως εξής:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

ΛΥΜΕΝΕΣ ΑΣΚΗΣΕΙΣ

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

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

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

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

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

Για όλους λοιπόν τους αριθμούς μηχανής

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

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

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

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

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

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

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

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

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

(ε) Εστω $x > y$ τότε:

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

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

3. *Θεωρείστε τη δευτεροβάθμια εξίσωση*

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

Δώστε έναν ευσταθή αλγόριθμο για τον υπολογισμό των ριζών της.

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

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

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

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

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

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

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

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

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

και το ακόλουθο:

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

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

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

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

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

Ομοια εργαζόμαστε για την άλλη περίπτωση και παίρνουμε:

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

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

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

Για το σχετικό σφάλμα έχουμε λοιπόν

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

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

ΑΛΥΤΕΣ ΑΣΚΗΣΕΙΣ

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

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

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

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

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

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

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

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

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

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

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

7. Διερευνήστε την κατάσταση επίλυσης του προβλήματος:

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

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

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

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

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

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

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

(γ) Παρατηρήστε ότι

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

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

Εχουμε λοιπόν:

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

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

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

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

ΚΕΦΑΛΑΙΟ 2

ΜΗ ΓΡΑΜΜΙΚΕΣ ΕΞΙΣΩΣΕΙΣ

Η αδυναμία επίλυσης της πλειοψηφίας των μη γραμμικών εξισώσεων με αναλυτικές μεθόδους, ώθησε στην ανάπτυξη αριθμητικών μεθόδων για την προσεγγιστική επίλυσή τους, π.χ. *συν(x) = x*, *ημ(x²)-x-0.2 = 0*, $\sqrt[3]{x^2}+\sqrt[5]{x}-5x=0,\ x>0$, κλπ.

§ 2.1 *Η μέθοδος διχοτόμησης*

Η μέθοδος διχοτόμησης βασίζεται στο ακόλουθο Θεώρημα:

Θεώρημα (Bolzano): Εστω $\alpha,b\in\boldsymbol{R}$, $\alpha < b$ και $f(x):[a,b]\to\boldsymbol{R}$ είναι μία <u>συνεχής</u> συνάρτηση στο κλειστό διάστημα $[\alpha,b]$, με $f(\alpha)\,f(b)<0$. Τότε, υπάρχει μία τουλάχιστον ρίζα της εξίσωσης $f(x)=0$ στο ανοικτό διάστημα (α,b).

Με χρήση του παραπάνω θεωρήματος, δεν γνωρίζουμε αν υπάρχουν περισσότερες της μίας ρίζες, ούτε ποια είναι η τιμή τους.

Η μέθοδος:

Χωρίς περιορισμό της γενικότητας, θεωρούμε μία πραγματική συνάρτηση $f(x)$, συνεχή στο κλειστό διάστημα $[\alpha,b]$, $\alpha < b$, με $f(\alpha)<0$ και $f(b)>0$. Εστω ρ μία ρίζα της εξίσωσης $f(x) = 0$. Κατασκευάζουμε μία ακολουθία προσεγγίσεων $m_n,\ n=1,2,\ldots$ της ρίζας ρ ως εξής:

Βήμα 1ο: Ορίζουμε ως αρχικό διάστημα το $\boldsymbol{I_1 = [\alpha,b]}$ και υπολογίζουμε

$$m_1=\frac{a+b}{2},\ (1^{\eta}\ \text{προσέγγιση της ρίζας})$$

να είναι το μέσον του διαστήματος I_1.

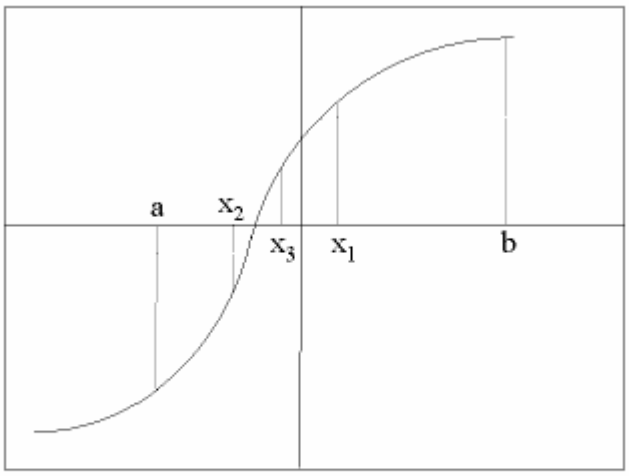
Βήμα 2ο: Υπολογίζουμε **<u>το πρόσημο</u>** της τιμής $f(m_1)$. Αν $f(m_1) = 0$, τότε έχουμε βρει μία ρίζα και σταματούμε, αλλιώς συνεχίζουμε στο επόμενο βήμα.

Βήμα 3ο Ορίζουμε ένα νέο διάστημα

$$I_2 = \begin{cases} (a, m_1), & \text{όταν } f(m_1) f(a) < 0 \\ (m_1, b), & \text{όταν } f(m_1) f(a) \geq 0 \end{cases},$$

εντός του οποίου η συνάρτηση *f(x)* ικανοποιεί τις συνθήκες του θεωρήματος Bolzano και ως εκ τούτου υπάρχει μία τουλάχιστον ρίζα της εξίσωσης *f(x) = 0* στο I_2.

Για να βούμε την 2η προσέγγιση m_2 της ρίζας ρ, επιστρέφουμε στο 1ο βήμα και επαναλαμβάνουμε τα βήματα 1-3 θεωρώντας πλέον ως αρχικό διάστημα το I_2 κλπ. Αποδεικνύεται ότι το $\lim_{n\to\infty} m_n = \rho$.



Σχήμα 2.1: *Μέθοδος της διχοτόμησης*

Είναι φανερό ότι σε κάθε επανάληψη των βημάτων 1-3, το εύρος του αρχικού διαστήματος όπου υπάρχει η ρίζα υποδιπλασιάζεται, διότι το ένα από τα δύο άκρα του νέου διαστήματος, μεταφέρεται ακριβώς στο μέσον του ακριβώς προηγούμενου διαστήματος. Συνεπώς, μετά από *N* επαναλήψεις των βημάτων 1-3, το εύρος (μήκος) *l(N)* του διαστήματος I_N είναι:

$$l(N) = \frac{b-a}{2^{N-1}}.$$

Αν λοιπόν τερματίσουμε τη διαδικασία μετά από *N* επαναλήψεις, δεν θα έχουμε υπολογίσει την ακριβή ρίζα ρ, αλλά μία προσέγγισή της m_N. Επειδή όμως και οι δύο τιμές ρ, m_N θα βρίσκονται εντός του διαστήματος I_N (μάλιστα η m_N είναι το μέσον του I_N), θα ισχύει:

$$|\varepsilon_\rho| \;=\; |\rho - m_N| \;\le\; \frac{l(N)}{2} = \frac{b-a}{2^N},$$

που προφανώς είναι ένα άνω φράγμα για το απόλυτο σφάλμα.

Συνήθως τερματίζουμε τη διαδικασία όταν το εύρος του διαστήματος I_N γίνει μικρότερο από μία θετική παράμετρο ανοχής ε. Θέτουμε λοιπόν

$$|\varepsilon_\rho| < \varepsilon \;\Rightarrow\; \frac{b-a}{2^N} < \varepsilon$$

και λύνοντας την παραπάνω ανίσωση ως προς N προσδιορίζουμε εκ των προτέρων το πλήθος των επαναλήψεων που απαιτούνται, ώστε να έχουμε αποτέλεσμα με την επιθυμητή ακρίβεια ε:

$$N > \frac{\ln(b-a) - \ln\varepsilon}{\ln 2}.$$

Παράδειγμα 1 Εστω *f(x)* <u>συνεχής</u> συνάρτηση στο κλειστό διάστημα *[α,b]*, *α < b* και έστω ότι δίνεται ο ακόλουθος πίνακας:

X	*α*	*b*	*(α+b)/2*	*(α+3b)/4*
Πρόσημο των τιμών *f(x)*	—	+	—	+

Να υπολογίσετε με τη μέθοδο διχοτόμησης μία ρίζα της εξίσωσης *f(x) = 0* για *N = 3* επαναλήψεις και να υπολογίσετε το σφάλμα, εάν *b-α = 0.4*.

Λύση

***<u>1η επανάληψη</u>*:**

1ο βήμα: Ορίζουμε ως **αρχικό διάστημα το $I_1 = (α,b)$**. Υπολογίζουμε το μέσον του διαστήματος I_1

$$m_1 = \frac{a+b}{2}, \text{ (1η προσέγγιση της ρίζας).}$$

Βήμα 2ο: Υπολογίζουμε ***<u>το πρόσημο</u>*** της τιμής $f(m_1)$. Από τον παραπάνω πίνακα τιμών έχουμε ότι $f(m_1) < 0$.

Βήμα 3ο Υπολογίζουμε το γινόμενο του προσήμου της τιμής $f(m_1)$ με το πρόσημο των τιμών της συνάρτησης f στα άκρα του διαστήματος I_1. Επειδή $f(m_1)f(b)<0$, ορίζουμε ένα νέο διάστημα I_2:

$$I_2=\left(\frac{a+b}{2}\ ,\ b\right).$$

***<u>2η επανάληψη</u>*:**

1ο βήμα: Ορίζουμε ως **αρχικό διάστημα το** $I_2=\left(\frac{a+b}{2}\ ,\ b\right)$. Υπολογίζουμε το μέσον του διαστήματος I_2

$$m_2=\frac{\frac{a+b}{2}+b}{2}=\frac{a+3b}{4},\ (2^{\eta}\ \text{προσέγγιση της ρίζας}).$$

Βήμα 2ο: Υπολογίζουμε ***<u>το πρόσημο</u>*** της τιμής $f(m_2)$. Από τον παραπάνω πίνακα τιμών έχουμε ότι $f(m_2)>0$.

Βήμα 3ο Υπολογίζουμε το γινόμενο του προσήμου της τιμής $f(m_2)$ με το πρόσημο των τιμών της συνάρτησης f στα άκρα του διαστήματος I_2. Επειδή $f\left(\frac{a+b}{2}\right)f(m_2)<0$ ορίζουμε ένα νέο διάστημα I_3:

$$I_3=\left(\frac{a+3b}{4}\ ,\ \frac{a+b}{2}\right).$$

***<u>3η επανάληψη</u>*:**

1ο βήμα: Ορίζουμε ως **αρχικό διάστημα το** $I_3=\left(\frac{a+3b}{4}\ ,\ \frac{a+b}{2}\right)$. Υπολογίζουμε το μέσον του διαστήματος I_3

$$m_3=\frac{\frac{a+3b}{4}+\frac{a+b}{2}}{2}=\frac{3a+5b}{8},$$

η οποία τιμή m_3, εφόσον η 3η επανάληψη είναι και η τελευταία, μας δίνει την προσεγγιστική ρίζα της εξίσωσης $f(x) = 0$. Για το σφάλμα έχουμε:

$$|\varepsilon_\rho| \le \frac{b-a}{2^N} = \frac{0.4}{2^3} = \frac{1}{32}.$$

Παράδειγμα 2 Να υπολογίσετε με τη μέθοδο διχοτόμησης τη μοναδική ρίζα της εξίσωσης $2x+2 = e^x$ στο διάστημα $(1,2)$. Η διαδικασία να σταματήσει όταν το εύρος του τελικού διαστήματος γίνει μικρότερο του 0.04.

Λύση Εφόσον

$$|\varepsilon_\rho| \le \frac{b-a}{2^N} \le 0.04 \Rightarrow \frac{1}{2^N} \le 0.04 \Rightarrow 2^N \ge 25 \Rightarrow \ln\left(2^N\right) \ge \ln(25)$$

$$N \ge \frac{\ln 25}{\ln 2} \cong 4.64 \Rightarrow N = 5,$$

άρα χρειαζόμαστε **τουλάχιστον *5* <u>επαναλήψεις</u>**, ώστε το σφάλμα να γίνει μικρότερο του 0.04. Παρατηρούμε ότι $f(x) = 2x+2-e^x$ και $f(1) = 1$, $f(2)<0$, άρα υπάρχει μία τουλάχιστον ρίζα της εξίσωσης $f(x) = 0$ στο ανοικτό διάστημα $(0,1)$.

***<u>1η επανάληψη</u>*:**

1ο βήμα: Ορίζουμε ως **αρχικό διάστημα το $I_1 = (1,2)$**. Υπολογίζουμε το μέσον του διαστήματος I_1

$$m_1 = \frac{1+2}{2} = 1.5.$$

Βήμα 2ο: Υπολογίζουμε ***<u>το πρόσημο</u>*** της τιμής

$$f(1.5) = 2\ 1.5 + 2 - e^{1.5} = 3 + 2 - 4.48169 < 0.$$

Βήμα 3ο Υπολογίζουμε το γινόμενο του προσήμου της τιμής $f(m_1) = f(1.5)$ με το πρόσημο των τιμών της συνάρτησης f στα άκρα του διαστήματος I_1. Επειδή $f(1.5)f(2) < 0$ ορίζουμε ένα νέο διάστημα I_2:

$$I_2 = (1.5\ ,2).$$

***<u>2η επανάληψη</u>*:**

1ο βήμα: Ορίζουμε ως **αρχικό διάστημα το $I_2 = (1.5\ ,\ 2)$**. Υπολογίζουμε το μέσον του διαστήματος I_2

$$m_2 = \frac{1.5+2}{2} = 1.75\,.$$

Βήμα 2ο: Υπολογίζουμε ***<u>το πρόσημο</u>*** της τιμής $f(1.75) = 3.5 + 2 - 5.7546 < 0.$

Βήμα 3ο Υπολογίζουμε το γινόμενο του προσήμου της τιμής $f(m_2) = f(1.75)$ με το πρόσημο των τιμών της συνάρτησης f στα άκρα του διαστήματος I_2. Επειδή $f(1.5)f(1.75) < 0$, ορίζουμε ένα νέο διάστημα I_3:

$$I_3 = \left(1.5\ ,\ 1.75\right).$$

Συνεχίζοντας με τον ίδιο τρόπο, εκτελέστε τις <u>υπόλοιπες 3 επαναλήψεις</u> και διαπιστώστε ότι η προσεγγιστική ρίζα είναι η m_5 = *1.67835.*

§ 2.2 *Επαναληπτικές μέθοδοι. Μέθοδος Newton-Raphson*

Κάθε εξίσωση της μορφής $f(x) = 0$, μπορεί να γραφεί ισοδύναμα στη μορφή $x = \varphi(x)$ με πολλούς τρόπους. Σε τέτοιες παραστάσεις βασίζονται οι λεγόμενες επαναληπτικές μέθοδοι.

Ορισμός 2.2.1 Ένα σημείο x^* του πεδίου ορισμού μιας συνάρτησης φ καλείται σταθερό σημείο της, αν ισχύει $\varphi(x^*) = x^*$.

Στις επαναληπτικές μεθόδους, γράφουμε την εξίσωση $f(x) = 0$ στη μορφή $x = \varphi(x)$ και ξεκινώντας από μία αρχική τιμή x_0, υπολογίζουμε μία ακολουθία προσεγγίσεων ενός σταθερού σημείου της φ από τη σχέση $x_n = \varphi(x_{n-1})$. Αν λοιπόν $x_n \to x^*$ και αν η $\varphi(x)$ είναι συνεχής στο σημείο x^*, τότε το x^* είναι σταθερό σημείο της φ. Πράγματι:

$$x^* = \lim_{n\to\infty} x_n = \lim_{n\to\infty} \varphi(x_{n-1}) = \varphi\left(\lim_{n\to\infty} x_{n-1}\right) = \varphi\left(x^*\right).$$

<u>Υπαρξη και μοναδικότητα σταθερού σημείου</u>

Θεώρημα 2.2.1 Εστω $\varphi:[a,b]\to[a,b]$ συνεχής πραγματική συνάρτηση τέτοια ώστε:

$$\textit{υπάρχει } 0<C<1:\ |\varphi(x)-\varphi(y)|\ \le\ C\ |x-y|\ \ \forall x,y\in[a,b],$$

(μία τέτοια συνάρτηση καλείται ***συστολή***), τότε η συνάρτηση φ έχει ***μοναδικό σταθερό σημείο x****. Επιπλέον, για οποιαδήποτε αρχική τιμή $x_0 \in [\alpha,b]$, η ακολουθία x_n με αναδρομικό τύπο $x_n = \varphi(x_{n-1})$ συγκλίνει προς το x^* . Τέλος, για κάθε φυσικό αριθμό n ισχύει:

$$|x_n - x^*|\ \le\ C\ |x_{n-1} - x^*|. \qquad (2.1)$$

Τάξη σύγκλισης ακολουθίας

Η σχέση (2.1), υπονοεί ότι η ***ακολουθία x_n συγκλίνει (τουλάχιστον) γραμμικά*** στο σταθερό σημείο x^* της $\varphi(x)$. Γενικότερα, θα λέμε ότι η ***σύγκλιση είναι (τουλάχιστον) τάξης p, p > 1***, αν ισχύει

$$|x_n - x^*|\ \le\ C\ |x_{n-1} - x^*|^p,\ \ \text{για κάθε φυσικό αριθμό } n.$$

Προσδιορισμός της τάξης σύγκλισης μιας ακολουθίας

Για τον προσδιορισμό της τάξης σύγκλισης μιας ακολουθίας x_n που είναι γενικά μία δύσκολη υπόθεση, πολύ χρήσιμο είναι το ακόλουθο:

Θεώρημα 2.2.2 Εστω ότι $x_n \neq x^*$ για κάθε φυσικό αριθμό n και έστω ότι ισχύει:

$$\lim_{n\to\infty} \frac{x_{n+1} - x^*}{(x_n - x^*)^p}\ =\ a \neq 0,$$

τότε η τάξη σύγκλισης της ακολουθίας x_n είναι ***ακριβώς*** p.

Ας υποθέσουμε τώρα ότι πλέον των υποθέσεων του θεωρήματος συστολής 2.2.1, έχουμε ότι η συνάρτηση $\varphi(x)$ είναι *συνεχώς παραγωγίσιμη* στο $[\alpha,b]$. Τότε, από το θεώρημα μέσης τιμής του διαφορικού λογισμού, υπάρχει τιμή ξ_n μεταξύ των x_n και x:

$$x_{n+1} - x^* = \varphi(x_n) - \varphi(x^*)\ =\ \varphi'(\xi_n)(x_n - x^*).$$

Εφόσον $x_n \to x^*$, θα ισχύει ότι $\xi_n \to x^*$ και λόγω συνέχειας της $\varphi'(x)$ έχουμε:

$$\lim_{n\to\infty} \frac{x_{n+1} - x^*}{x_n - x^*} = \varphi'(x^*).$$

Αν λοιπόν $0 < |\varphi'(x^*)| < 1$, τότε η τάξη σύγκλισης που παράγει η επαναληπτική μέθοδος $x_n = \varphi(x_{n-1})$ είναι ακριβώς ένα. Για να πάρουμε λοιπόν επαναληπτικές μεθόδους $x_n = \varphi(x_{n-1})$ με τάξη σύγκλισης μεγαλύτερη του 1, θα πρέπει να αναζητήσουμε μεθόδους για τις οποίες ισχύει

$$\varphi'(x^*) = 0.$$

Μία τέτοια είναι και η μέθοδος Newton-Raphson που θα αναπτύξουμε στη συνέχεια.

Παράδειγμα 3 Εστω πραγματική συνάρτηση f, τέτοια ώστε x^* είναι απλή ρίζα της $f(x) = 0$ και η f είναι δύο φορές συνεχώς παραγωγίσιμη σε μία περιοχή του x^*. Να δειχθεί ότι η επαναληπτική μέθοδος

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

συγκλίνει στο x^* και να υπολογισθεί η τάξη σύγκλισης.

Λύση Θεωρούμε την επαναληπτική μέθοδο $x = \varphi(x)$, όπου $\varphi(x) = x - \frac{f(x)}{f'(x)}$. Παραγωγίζουμε και βρίσκουμε:

$$\varphi'(x) = \frac{f(x)f''(x)}{(f'(x))^2} \Rightarrow \varphi'(x^*) = 0,$$

άρα υπάρχει μία περιοχή I του x^* τέτοια ώστε $|\varphi'(x)| \le C < 1$, οπότε

$$|\varphi(x) - \varphi(y)| \le |\varphi'(\xi)|\,|x - y| \le C\,|x - y|,\ x, y \in I,$$

οπότε από το θεώρημα 2.2.1 της συστολής η $\varphi(x)$ έχει μοναδικό σταθερό σημείο x^*, και η ακολουθία (x_n) με αναδρομικό τύπο $x_n = \varphi(x_{n-1}) \to x^*$, $n \to \infty$.

Για να υπολογίσουμε την τάξη σύγκλισης θα χρησιμοποιήσουμε το Θεώρημα 2.2.2. Από τον τύπο του Taylor με κέντρο το σημείο x^* έχουμε:

$$f(x_n) = f(x^*) + f'(x^*)(x_n - x^*) + \frac{f''(\xi_n)}{2}(x_n - x^*)^2,\ \xi_n \in (x^*, x_n) \text{ ή } \xi_n \in (x_n, x^*)$$
$$f'(x_n) = f'(x^*) + f''(\xi_n')(x_n - x^*),\ \xi' \in (x^*, x_n) \text{ ή } \xi_n \in (x_n, x^*)$$

.

Αντικαθιστούμε τις τιμές $f(x_n), f'(x_n)$ στον αναδρομικό τύπο $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$, και παίρνουμε:

$$x_{n+1} = x_n - \frac{f(x^*) + f'(x^*)(x_n - x^*) + \frac{f''(\xi_n)}{2}(x_n - x^*)^2}{f'(x^*) + f''(\xi_n')(x_n - x^*)},$$

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

$$x_{n+1} - x^* = (x_n - x^*)^2 \frac{f''(\xi_n') - \frac{f''(\xi_n)}{2}}{f'(x^*) + f''(\xi_n')(x_n - x^*)},$$

άρα: $\lim_{n\to\infty} \frac{x_{n+1} - x^*}{(x_n - x^*)^2} = \frac{f''(x^*)}{2f'(x^*)} \neq 0$, άρα η τάξη σύγκλισης είναι 2. □

<u>Η μέθοδος Newton-Raphson</u>

Η μέθοδος Newton-Raphson βασίζεται στο ακόλουθο:

Θεώρημα 2.2.3 Εστω μία πραγματική συνάρτηση *f(x)*:

(α) η *f(x)* είναι δύο φορές παραγωγίσιμη στο διάστημα *[α,b]*, με $f'(x), f''(x) \neq 0$ για κάθε $x \in [\alpha, b]$,

(β) $f(\alpha)\, f(b) < 0$,

τότε υπάρχει <u>μοναδική</u> ρίζα x^* της εξίσωσης $f(x)=0$ στο ανοικτό διάστημα (α,b), η οποία είναι το <u>όριο της αναδρομικής ακολουθίας</u>:

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

όπου το αρχικό σημείο x_0 της αναδρομικής σχέσης εκλέγεται έτσι ώστε

$$f(x_0)\,f''(x_0) > 0\,.$$

Απόδειξη Από τη συνθήκη (β) και την παραγωγισιμότητα της *f(x)* προκύπτει ότι η εξίσωση *f(x) = 0* έχει μία τουλάχιστον ρίζα x^* στο ανοικτό διάστημα *(α,b)*. Εφόσον δε $f'(x) \neq 0$ για κάθε $x \in [\alpha,b]$, προκύπτει ότι η συνάρτηση *f(x)* είναι γνησίως μονότονη στο *[α,b]*, άρα η ρίζα x^* είναι μοναδική. Χωρίς περιορισμό της γενικότητας υποθέτουμε ότι *f(α) < 0*, άρα <u>η f είναι γνησίως αύξουσα</u>. Επιπλέον, έστω x_0: $f(x_0) > 0$, τότε από τη συνθήκη $f(x_0)\,f''(x_0) > 0$ προκύπτει ότι $f''(x_0) > 0$ και εφόσον ισχύει $f''(x) \neq 0$ για κάθε $x \in [\alpha,b]$ θα έχουμε $f''(x) > 0$ για κάθε $x \in [\alpha,b]$, δηλαδή η <u>f είναι κυρτή</u> στο *[α,b]*.

Θεωρούμε τώρα την ακολουθία $x_n = x_{n-1} - \dfrac{f(x_{n-1})}{f'(x_{n-1})},\ n = 1,2,...$

Με τη μέθοδο της επαγωγής θα δείξουμε ότι η (x_n) *είναι κάτω φραγμένη από τη ρίζα* x^*.

Αφού η *f (x)* είναι γνησίως αύξουσα και $f(x_0) > 0 = f(x^*)$, προκύπτει ότι $x_0 > x^*$.Υποθέτουμε ότι $x_n > x^*$, θα δείξουμε ότι $x_{n+1} > x^*$. Από το ανάπτυγμα Taylor της *f(x)* με κέντρο το σημείο x_n έχουμε:

$$f(x) = f(x_n) + f'(x_n)(x - x_n) + \frac{f''(\xi_n)}{2}(x - x_n)^2,\ \xi_n \in (x, x_n),$$

οπότε για $x = x^*$ έχουμε:

$$0 = f(x^*) = f(x_n) + f'(x_n)(x^* - x_n) + \frac{f''(\xi_n)}{2}(x^* - x_n)^2,\ \xi_n \in (x, x_n)$$

άρα $f(x_n) + f'(x_n)(x^* - x_n) < 0$,

(αφού υποθέσαμε ότι η *f* είναι κυρτή) και κάνοντας τις πράξεις παίρνουμε

$$x_n - \frac{f(x_n)}{f'(x_n)} > x^* \Rightarrow x_{n+1} > x^*.$$

Αρα η ακολουθία (x_n) είναι κάτω φραγμένη.

Από την παραπάνω ανισότητα και τη μονοτονία της f ισχύει ότι

$$f(x_{n+1}) > f(x^*) = 0 \text{ για κάθε } n$$

άρα:

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} < x_{n-1} \quad \text{για κάθε } n,$$

δηλαδή η ακολουθία *(x_n) είναι γνησίως φθίνουσα*. Αφού είναι και κάτω φραγμένη συγκλίνει σε έναν αριθμό $\bar{x}$. Τότε:

$$\lim x_n = \lim x_{n-1} - \frac{\lim f(x_{n-1})}{\lim f'(x_{n-1})} \Rightarrow \bar{x} = \bar{x} - \frac{f(\bar{x})}{\lim f'(x_{n-1})} \Rightarrow f(\bar{x}) = 0,$$

άρα το όριο της ακολουθίας x_n είναι η μοναδική ρίζα της εξίσωσης *f(x) = 0*. □

Πόρισμα 2.2.1 Με τις προϋποθέσεις του προηγούμενου θεωρήματος, το σφάλμα κατά την προσέγγιση της μοναδικής ρίζας της εξίσωσης *f(x) = 0* με τη μέθοδο Newton-Raphson από τον όρο x_n δίνεται από τη σχέση

$$|x_n - x^*| \le \frac{M}{2m} |x_n - x_{n-1}|^2,$$

όπου $m = \min_{x\in[a,b]} |f'(x)|,\ M = \max_{x\in[a,b]} |f''(x)|.$

Απόδειξη Εστω x^* η μοναδική ρίζα της εξίσωσης *f(x) = 0,* από το θεώρημα μέσης τιμής του διαφορικού λογισμού ισχύει:

$$f(x^*) - f(x_n) = f'(c_n)(x^* - x_n),\ c_n \in (x, x_n)\ ,$$

άρα $0 - f(x_n) = f'(c_n)(x^* - x_n)$, συνεπώς:

$$|x^* - x_n| = \frac{|f(x_n)|}{|f'(c_n)|} \le \frac{|f(x_n)|}{m}, \qquad (2.2)$$

όπου $m = \min_{x\in[a,b]} f'(x)$. Από το ανάπτυγμα Taylor της *f(x)* με κέντρο το σημείο x_{n-1} έχουμε:

$$f(x) = f(x_{n-1}) + f'(x_{n-1})(x - x_{n-1}) + \frac{f''(\xi_n)}{2}(x - x_{n-1})^2,\ \xi_n \in (x, x_{n-1}),$$

οπότε για $x = x_n$ έχουμε:

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

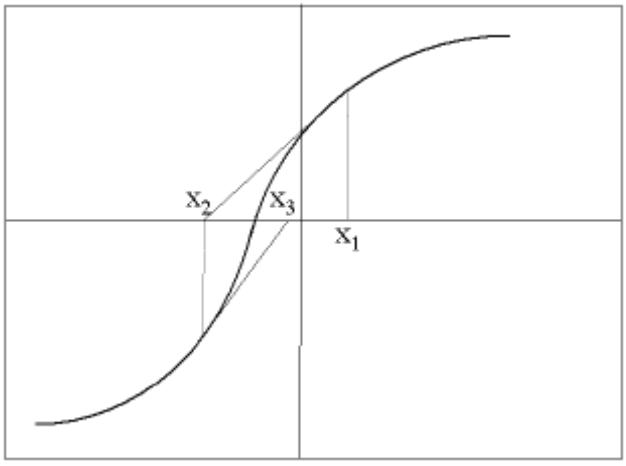
$$= \frac{f''(\xi_n)}{2}(x_n - x_{n-1})^2, \qquad (2.3)$$

διότι $x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \Rightarrow f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1}) = 0$.

Αντικαθιστούμε την (2.3) στην (2.2) και παίρνουμε το ζητούμενο. □

Γεωμετρική ερμηνεία

Η γεωμετρική ερμηνεία της μεθόδου γίνεται σαφής με τη βοήθεια του ακόλουθου σχήματος:



Σχήμα 2: Η μέθοδος Newton

Εστω $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$ η 1^{η} προσέγγιση της ρίζας στην 1^{η} επανάληψη. Η εξίσωση της εφαπτόμενης της συνάρτησης *f(x)* στο σημείο $(x_1, f(x_1))$ είναι:

$$y - f(x_1) = f'(x_1)(x - x_1).$$

Η εφαπτόμενη ευθεία τέμνει τον άξονα των x σε ένα σημείο x_2 το οποίο υπολογίζεται εύκολα αν θέσουμε στην εξίσωση της εφαπτόμενης ευθείας $y = 0$, $x = x_2$, οπότε βρίσκουμε:

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} \ ,$$

η οποία είναι ακριβώς η αναδρομική σχέση του Θεωρήματος 2.2.3 για $n = 2$. Συνεχίζοντας, διαπιστώνουμε η κάθε επόμενη προσέγγιση x_{n+1} είναι το σημείο τομής της γραφικής παράστασης της εφαπτόμενης της $f(x)$ στο σημείο $(x_n, f(x_n))$ με τον άξονα των x.

Παράδειγμα 4 Να προσεγγισθεί με τη μέθοδο Newton – Raphson η μοναδική ρίζα της εξίσωσης $2x - e^{-x} = 0$ στο ανοικτό διάστημα $(0,1)$ για $N = 3$ επαναλήψεις και να υπολογισθεί το σφάλμα της μεθόδου.

Λύση Ορίζουμε $f(x) = 2x - e^{-x}$, και υπολογίζουμε

$$f'(x) = 2 + e^{-x},\ f''(x) = -e^{-x}.$$

Προφανώς $f'(x), f''(x) \neq 0$ για κάθε $x \in (0,1)$, $f(0) = -1$, $f(1) = 1.63212$, άρα ικανοποιούνται οι προϋποθέσεις του Θεωρήματος 2.2.3 και συνεπώς υπάρχει μοναδική ρίζα της εξίσωσης $f(x) = 0$, η οποία είναι το όριο της αναδρομικής ακολουθίας $x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})},\ n = 0,1,...$

Επιλέγουμε ως αρχικό σημείο εκκίνησης το $x_0 = 0$, ή $x_0 = 1$ (ένα από τα δύο άκρα του αρχικού διαστήματος $(0,1)$) βάσει της συνθήκης $f(x_0) f''(x_0) > 0$ του Θεωρήματος 2.2.3. Είναι εύκολο να δει κανείς ότι $f(0) f''(0) > 0$, ενώ $f(1) f''(1) < 0$, άρα

$$x_0 = 0.$$

1η επανάληψη:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 0 - \frac{f(0)}{f'(0)} = 0 - \frac{-1}{2 + e^0} = \frac{1}{3}.$$

2η επανάληψη:

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = \frac{1}{3} - \frac{f\left(\frac{1}{3}\right)}{f'\left(\frac{1}{3}\right)} = 0.351689.$$

3η επανάληψη:

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)} = 0.351689 - \frac{f(0.351689)}{f'(0.351689)} = 0.351734,$$

η οποία είναι και η προσεγγιστική τιμή της ρίζας.

Το σφάλμα υπολογίζεται από το Πόρισμα 2.2.1 για *n = 3*:

$$|x_3 - x^*| \le \frac{M}{2m}|x_3 - x_2|^2 = \frac{M}{2m}|0.351734 - 0.351689|^2,$$

όπου $m = \min_{x\in[0,1]}|f'(x)|$, $M = \max_{x\in[a,b]}|f''(x)|$. Επειδή

$$m = \min_{x\in[0,1]} |f'(x)| = \min_{x\in[0,1]}\left(2 + e^{-x}\right) = 2 + e^{-1} = 2.36788,$$

$$M = \max_{x\in[0,1]}|f''(x)| = \max_{x\in[0,1]} e^{-x} = e^0 = 1,$$

έχουμε: $|x_3 - x^*| \le \frac{1}{2\ 2.36788}|0.351734 - 0.351689|^2 = 4.27598\ 10^{-10}$.

Παρατήρηση: Είναι σαφές από η μέθοδος Newton – Raphson είναι ειδική περίπτωση επαναληπτικής μεθόδου της μορφής $x_n = \varphi(x_{n-1})$, όπου η συνάρτηση επανάληψης είναι της μορφής:

$$\varphi(x) = x - \frac{f(x)}{f'(x)} \quad .$$

Επειδή $\varphi'(x) = -\frac{f(x)f''(x)}{\left(f'(x)\right)^2}$, με την προϋπόθεση ότι $f'(x^*) \neq 0$ (δηλαδή x^* απλή ρίζα της εξίσωσης *f(x) = 0*), προκύπτει ότι $\varphi'(x^*) = 0$, άρα η

τάξη σύγκλισης είναι μεγαλύτερη του 1. στο Θεώρημα 2.2.3 είδαμε ότι η σύγκλιση είναι τετραγωνική.

<u>Γενίκευση της μεθόδου Newton-Raphson σε συστήματα εξισώσεων</u>

Δίδεται το σύστημα εξισώσεων $\begin{cases} f(x,y)=0 \\ g(x,y)=0 \end{cases}$, όπου *f, g* πραγματικές συναρτήσεις δύο μεταβλητών και έστω *(x*,y*)* μία λύση του. Υποθέτοντας ότι σε μία περιοχή του πεδίου ορισμού υπάρχουν οι δεύτερες μερικές παράγωγοι των *f,g* και είναι συνεχείς, έχουμε από το Θεώρημα Taylor για συναρτήσεις πολλών μεταβλητών:

$$0 = f(x^*, y^*) = f(x,y) + \frac{\partial f}{\partial x}(x - x^*) + \frac{\partial f}{\partial y}(y - y^*) + O\left((x-x^*)^2 + (y-y^*)^2\right)$$
$$0 = g(x^*, y^*) = g(x,y) + \frac{\partial g}{\partial x}(x - x^*) + \frac{\partial g}{\partial y}(y - y^*) + O\left((x-x^*)^2 + (y-y^*)^2\right)$$

.

Παραλείποντας τους τετραγωνικούς όρους γράφουμε:

$$\begin{pmatrix} \dfrac{\partial f(x,y)}{\partial x} & \dfrac{\partial f(x,y)}{\partial y} \\ \dfrac{\partial g(x,y)}{\partial x} & \dfrac{\partial g(x,y)}{\partial x} \end{pmatrix} \begin{pmatrix} x - x^* \\ y - y^* \end{pmatrix} \cong -\begin{pmatrix} f(x,y) \\ g(x,y) \end{pmatrix}.$$

Οδηγούμαστε λοιπόν υπό κατάλληλες προϋποθέσεις (η ιακωβιανή ορίζουσα είναι διάφορη του μηδενός και *x, y* αρκετά κοντά στα *x*, y**), σε μία επαναληπτική μέθοδο της μορφής:

$$\begin{pmatrix} \dfrac{\partial f(x_n,y_n)}{\partial x} & \dfrac{\partial f(x_n,y_n)}{\partial y} \\ \dfrac{\partial g(x_n,y_n)}{\partial x} & \dfrac{\partial g(x_n,y_n)}{\partial x} \end{pmatrix} \begin{pmatrix} x_{n+1} - x_n \\ y_{n+1} - y_n \end{pmatrix} \cong -\begin{pmatrix} f(x_n,y_n) \\ g(x_n,y_n) \end{pmatrix}$$

όπου μάλιστα η σύγκλιση είναι τετραγωνική.

§ 2.3 *Μέθοδος τέμνουσας (secant method)*

Η μέθοδος Newton-Raphson απαιτεί γνώση της *f′(x)*. Στην περίπτωση που η παράγωγος δεν είναι γνωστή ή είναι δύσκολο να υπολογισθεί, καταφεύγουμε συνήθως στη μέθοδο της τέμνουσας.

Ας θεωρήσουμε τον αναδρομικό τύπο της μεθόδου Newton-Raphson:

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

και ας θυμηθούμε ότι

$$f'(x) \approx \frac{f(x+h) - f(x)}{(x+h) - x},\ h\ \mu\iota\kappa\rho\dot{o}\,.$$

Για $x = x_n$, $x_{n+1} = x_n + h$, έχουμε:

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

Αν λοιπόν στον αναδρομικό τύπο Newton-Raphson αντικαταστήσουμε αντί της παραγώγου $f'(x_n)$, το δεξιό μέλος της παραπάνω ισότητας, προκύπτει η **μέθοδος της τέμνουσας**

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

η οποία χρειάζεται προφανώς δύο αρχικές συνθήκες x_0, x_1. Η μέθοδος βασίζεται στο ακόλουθο:

Θεώρημα 2.3.1 Εστω μία πραγματική συνάρτηση *f(x)*:

(α) η *f(x)* είναι δύο φορές παραγωγίσιμη στο διάστημα *[α,b]*, με $f'(x), f''(x) \neq 0$ για κάθε $x \in [\alpha,b]$,

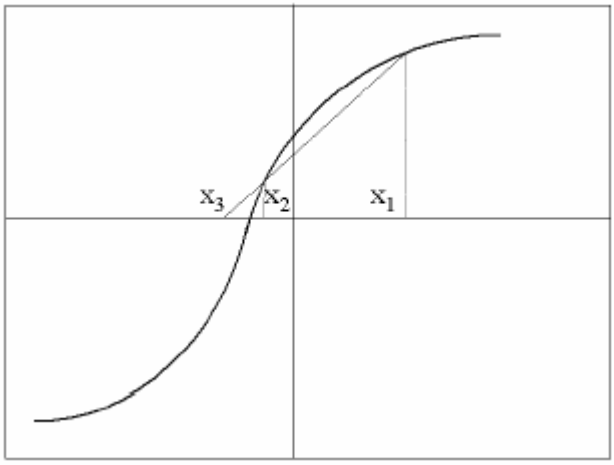
(β) $f(\alpha)\, f(b) < 0$,

τότε υπάρχει μοναδική ρίζα x^* της εξίσωσης $f(x)=0$ στο ανοικτό διάστημα *(α,b)*, η οποία είναι το όριο της αναδρομικής ακολουθίας:

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

όπου ως αρχικά σημεία x_0, x_1 της αναδρομικής σχέσης μπορούν να θεωρηθούν για ευκολία τα άκρα του διαστήματος *(α,b)*, δηλαδή $x_0 = \alpha$, $x_1 = b$. Η τάξη σύγκλισης της μεθόδου είναι $p = \frac{1+\sqrt{5}}{2} \cong 1.62.$

Γραφική επίλυση



Σχήμα 3 Η μέθοδος τέμνουσας

Παράδειγμα 5 Να προσεγγισθεί με τη μέθοδο τέμνουσας μία προσέγγιση της $\sqrt{3}$ στο ανοικτό διάστημα *(1,2)*, για *N = 3* επαναλήψεις.

Λύση Ορίζουμε $f(x) = x^2 - 3$ και εφόσον $f'(x), f''(x) \neq 0$ για κάθε $x \in (1,2)$, $f(1) = -2$, $f(2) = 1$, ικανοποιούνται οι προϋποθέσεις του Θεωρήματος 2.3.1 και συνεπώς υπάρχει μοναδική ρίζα της εξίσωσης $f(x) = 0$, η οποία είναι το όριο της αναδρομικής ακολουθίας

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

Εκλέγουμε $x_0 = 1$, $x_1 = 2$ (τα δύο άκρα του αρχικού διαστήματος *(1,2)*) και υπολογίζουμε

1η επανάληψη:

$$x_2 = x_1 - \frac{f(x_1)\ (x_1 - x_0)}{f(x_1) - f(x_0)} = 2 - \frac{f(2)\ (2-1)}{f(2) - f(1)} = 2 - \frac{1}{3} = 1.\overline{6}.$$

2η επανάληψη:

$$x_3 = x_2 - \frac{f(x_2)\ (x_2 - x_1)}{f(x_2) - f(x_1)} = 1.\overline{6} - \frac{f(1.\overline{6})\ (1.\overline{6} - 2)}{f(1.\overline{6}) - f(2)} = 1.72727\,.$$

3η επανάληψη:

$$x_4 = x_3 - \frac{f(x_3)\ (x_3 - x_2)}{f(x_3) - f(x_2)} = 1.72727 - \frac{f(1.72727)\ (1.72727 - 1.\overline{6})}{f(1.72727) - f(1.\overline{6})} = 1.73214\,,$$

η οποία είναι και η προσεγγιστική τιμή της ρίζας.

ΑΛΥΤΕΣ ΑΣΚΗΣΕΙΣ

1. Δίνονται οι συναρτήσεις: (α) $f(x) = e^{2x} + 2x + 1$, $x \in [-1,0]$,
(β) $g(x) = 2^x - 5x + 2$, $x \in [0,1]$,
(γ) $h(x) = x^3 - x - 1$, $x \in [1,2]$.

A. Αφού δείξετε ότι κάθε μία από τις ανωτέρω συναρτήσεις έχει μοναδική πραγματική ρίζα στα αντίστοιχα διαστήματα, να υπολογίσετε με τη μέθοδο διχοτόμησης μία προσέγγιση της ρίζας για κάθε μία εξ αυτών, ώστε το σφάλμα από την ακριβή τιμή της ρίζας να είναι μικρότερο του *0.06*.

Απάντηση:

(α) N = 6 επαναλήψεις

{$m_1 = -0.5$, $m_2 = -0.75$, $m_3 = -0.625$, $m_4 = -0.6875$, $m_5 = -0.65625$, $m_6 = -0.64063$}

(β) N = 6 επαναλήψεις

{$m_1 = 0.5$, $m_2 = 0.75$, $m_3 = 0.625$, $m_4 = 0.6875$, $m_5 = 0.71875$, $m_6 = 0.73438$}

(γ) N = 6 επαναλήψεις

{$m_1 = 1.5$, $m_2 = 1.25$, $m_3 = 1.375$, $m_4 = 1.3125$, $m_5 = 1.34375$, $m_6 = 1.32813$}

B. Να υπολογίσετε με τη μέθοδο Newton-Raphson μία προσέγγιση της ρίζας για κάθε μία από τις συναρτήσεις (α)-(γ), για $N = 4$ επαναλήψεις και να υπολογίσετε το σφάλμα.

Απάντηση:

(α) $\{ x_0 = 0, x_1 = -0.5, x_2 = -0.634471, x_3 = -0.639227, x_4 = -0.639232\}$

$|e| \leq 1.11699\ 10^{-10}$.

(β) $\{ x_0 = 0, x_1 = 0.69656, x_2 = 0.73212, x_3 = 0.73224, x_4 = 0.73224\}$

$|e| \cong 0$.

(γ) $\{x_0 = 2., x_1 = 1.54545, x_2 = 1.35961, x_3 = 1.3258, x_4 = 1.32472\}$

$|e| \leq 0.00001$.

Γ. Να υπολογίσετε με τη μέθοδο τέμνουσας μία προσέγγιση της ρίζας για κάθε μία από τις συναρτήσεις (α)-(γ), για $N = 4$ επαναλήψεις.

Απάντηση:

(α) $\{x_0 = -1, x_1 = 0, x_2 = -0.69816, x_3 = -0.64981, x_4 = -0.63910, x_5 = -0.63923\}$.

(β) $\{x_0 = 0, x_1 = 1, x_2 = 0.75, x_3 = 0.7317, x_4 = 0.73225, x_5 = 0.73224\}$.

(γ) $\{x_0 = 1, x_1 = 2, x_2 = 1.16667, x_3 = 1.25311, x_4 = 1.33721, x_5 = 1.32385\}$.

Σημείωση: Οι πράξεις να γίνουν με στρογγυλοποίηση στο $5^{ο}$ δεκαδικό ψηφίο.

2. Να δειχθεί ότι η μέθοδος: $x_{n+1} = x_n - \dfrac{f(x_n)}{f(x_n) + f'(x_n)}$ συγκλίνει τετραγωνικά.

Υπόδειξη: Οπως στο παράδειγμα 3 σελ. 31.

3. Δείξτε ότι η ακολουθία $x_{n+1} = \dfrac{1}{2} e^{\left(\frac{x_n}{2}\right)}$, $n = 1,...$ συγκλίνει και το όριό της βρίσκεται στο [0,1].

Υπόδειξη: Δείξτε ότι η $\varphi(x)=\frac{1}{2}e^{x/2}$ είναι συστολή, βλέπε Θεώρημα 2.2.1.

4. Εστω $a_n=\sqrt{2+\sqrt{2+...+2}}$, *n = 1,2,...*Χρησιμοποιώντας μία κατάλληλη επαναληπτική μέθοδο, δείξτε ότι $\lim a_n=2$. Επίσης δείξτε ότι $\lim\frac{a_{n+1}-2}{a_n-2}=\frac{1}{4}$.

5. Δίνεται η επαναληπτική μέθοδος

$$x_{n+1}=x_n+\lambda\left(x_n^2-3\right),\ n=0,1,2...$$

(α) Να βρεθεί διάστημα τιμών της παραμέτρου *λ* ώστε η επαναληπτική μέθοδος να συγκλίνει.
(β) Να βρεθεί η τιμή του *λ* ώστε η σύγκλιση να είναι τετραγωνική.
(γ) Είναι η μέθοδος Newton-Raphson πιο αποτελεσματική μέθοδος από αυτήν που αντιστοιχεί για την τιμή του *λ* που βρέθηκε στο ερώτημα (β);

6. Εστω $f(x) = x\,(x-2)^3$, τότε να επιλέξετε και να εφαρμόσετε την πλέον αποτελεσματική μορφή της μεθόδου Newton-Raphson για τον προσδιορισμό της ρίζας $x = 2$ για $N = 3$ επαναλήψεις. Ποια η τάξη σύγκλισης;

7. Δίνεται το σύστημα εξισώσεων: $\left\{f\left(x,y\right)=0,\ \ g\left(x,y\right)=0\right\}$, όπου $f\left(x,y\right)=1-x-y^2\sigma\upsilon\nu\left(\frac{\pi x}{2}\right)$ και $g\left(x,y\right)=e^{-xy}+5\eta\mu\left(\pi x\right)-2$. Εφαρμόστε την μέθοδο Newton-Raphson για συστήματα εξισώσεων για N=3 επαναλήψεις.

ΚΕΦΑΛΑΙΟ 3

ΑΡΙΘΜΗΤΙΚΗ ΕΠΙΛΥΣΗ ΓΡΑΜΜΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

§ 3.1 *Ο αλγόριθμος Gauss*

Εστω $n = 2,3,\ldots$, με τον όρο γραμμικά συστήματα $n \times n$, εννοούμε συστήματα n εξισώσεων με n αγνώστους της μορφής:

$$\begin{aligned} a_{11}x_1 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \ldots + a_{2n}x_n &= b_2 \\ &\cdots \\ a_{n1}x_1 + \ldots + a_{nn}x_n &= b_n \end{aligned}, \qquad (3.1)$$

όπου x_i, $i = 1,\ldots,n$, είναι οι *άγνωστοι όροι*, $\alpha_{i,j}$, $i,j = 1,\ldots,n$ είναι οι *συντελεστές* των αγνώστων όρων και b_i, $i = 1,\ldots,n$ είναι οι *σταθεροί* όροι. Το σύστημα (3.1) μπορεί να γραφεί με τη μορφή πινάκων ως εξής:

$$A\,x \;=\; B,$$

όπου

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$$

είναι ο ***πίνακας των συντελεστών των αγνώστων***,

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

είναι ο ***πίνακας στήλη των αγνώστων*** και

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

είναι ο ***πίνακας στήλη των σταθερών όρων***.

Είναι γνωστό από τη γραμμική άλγεβρα, ότι ικανές και αναγκαίες συνθήκες ώστε το σύστημα (3.1) να έχει μοναδική λύση είναι οι ακόλουθες:

- Ο πίνακας A είναι αντιστρέψιμος.
- Η ορίζουσα του πίνακα A είναι διάφορη του μηδενός.
- Οι γραμμές ή οι στήλες του πίνακα A είναι γραμμικά ανεξάρτητες.
- Το αντίστοιχο ομογενές σύστημα $A\,x = 0$ έχει ως μοναδική λύση την μηδενική.

Αν ένα σύστημα έχει μοναδική λύση, τότε για τον προσδιορισμό αυτής υπενθυμίζουμε τον κανόνα του *Cramer*:

$$x_i = \frac{Det\left(A_i\right)}{Det\left(A\right)},\ i = 1,...,n,$$

όπου A_i είναι ο πίνακας που προκύπτει, αν αντικαταστήσουμε την i-στήλη του πίνακα A με τη στήλη των σταθερών όρων. Ενας άλλος τρόπος είναι:

$$x\ =\ A^{-1}\,B\,,$$

όπου A^{-1} είναι ο αντίστροφος του πίνακα A. Σημειώνουμε όμως ότι και οι δύο παραπάνω τρόποι είναι υπολογιστικά ασύμφοροι. Για την επίλυση με τη μέθοδο Cramer απαιτούνται $(n+1)!+n$ πολλαπλασιασμοί, ενώ για την επίλυση με την εύρεση του αντιστρόφου πίνακα A^{-1}, αναγόμαστε στην επίλυση ενός συστήματος με n^2 αγνώστους, ή ισοδύναμα στην επίλυση n γραμμικών συστημάτων με τον ίδιο πίνακα A και έναν πολλαπλασιασμό επί διάνυσμα. Είναι σαφές ότι για μεγάλα $n = 100, 1000$ κ.λ.π., το υπολογιστικό κόστος γίνεται απαγορευτικό.

Ο συνηθέστερος τρόπος επίλυσης των συστημάτων αυτών, είναι ο αλγόριθμος του Gauss, ο οποίος περιγράφεται από τα ακόλουθα βήματα:

Βήμα 1ο : ορίζουμε τον ***επαυξημένο πίνακα*** των συντελεστών και σταθερών όρων, διάστασης $n\times\left(n+1\right)$:

$$A_{\varepsilon\pi}=\left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{array}\right).$$

Βήμα 2ο: Ξεκινώντας πάντοτε με <u>*οδηγό στοιχείο το 1ο στοιχείο της κυρίας διαγωνίου*</u>, δηλαδή το a_{11}, εκτελούμε μία πράξη μεταξύ της $1^{ης}$ γραμμής και κάθε μίας από τις επόμενες γραμμές, <u>*έτσι ώστε όλα τα στοιχεία κάτω από το οδηγό στοιχείο να μηδενίζονται*</u>. Μετά το πέρας του βήματος αυτού, ο νέος επαυξημένος πίνακας θα έχει τη μορφή:

$$A_{\varepsilon\pi}(1)=\left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} & b_n^{(2)} \end{array}\right).$$

Ο παραπάνω πίνακας, προκύπτει από την πράξη *(i = 2,…,n)*:

(i γραμμή του $A_{\varepsilon\pi}(1)$*)* $= -\dfrac{a_{i1}}{a_{11}}$ *(*$1^{η}$ *γραμμή του* $A_{\varepsilon\pi}$*) + (i γραμμή του* $A_{\varepsilon\pi}$*).*

Βήμα 3ο: Συνεχίζουμε τη διαδικασία, ξεκινώντας τώρα με <u>*οδηγό στοιχείο το 2ο στοιχείο της κυρίας διαγωνίου του πίνακα*</u> $A_{\varepsilon\pi}(1)$, δηλαδή το $a_{22}^{(2)}$, και εκτελούμε μία πράξη μεταξύ της $2^{ης}$ γραμμής και κάθε μίας από τις επόμενες γραμμές, <u>*έτσι ώστε όλα τα στοιχεία κάτω από το οδηγό στοιχείο να μηδενίζονται*</u>. Μετά το πέρας του βήματος αυτού ο νέος επαυξημένος πίνακας θα έχει τη μορφή:

$$A_{\varepsilon\pi}(2)=\left(\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} & b_1 \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3n}^{(3)} & b_3^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & a_{n3}^{(3)} & \cdots & a_{nn}^{(3)} & b_n^{(3)} \end{array}\right).$$

Ο παραπάνω πίνακας, προκύπτει από την πράξη *(i = 3,…,n)*:

(i γραμμή του $A_{\varepsilon\pi}(2)$*)* $= -\dfrac{a_{i2}^{(2)}}{a_{22}^{(2)}}$ *(*$2^{η}$ *γραμμή του* $A_{\varepsilon\pi}(1)$*) + (i γραμμή του* $A_{\varepsilon\pi}(1)$*).*

Συνεχίζοντας με τον ίδιο τρόπο μετά από *n* βήματα καταλήγουμε σε έναν επαυξημένο πίνακα με τριγωνική μορφή:

$$A_{\varepsilon\pi}(n)=\left(\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} & b_1 \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3n}^{(3)} & b_3^{(3)} \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & a_{nn}^{(n)} & b_n^{(n)} \end{array}\right),$$

η οποία επιτρέπει να υπολογισθούν οι λύσεις με τη διαδικασία της προς τα πίσω αντικατάστασης. Δηλαδή, από την εξίσωση:

$$a_{nn}^{(n)}\, x_n \;=\; b_n^{(n)},$$

υπολογίζουμε το x_n κ.λ.π.. Αν συμβεί $a_{nn}^{n} = 0$ και $b_n^{(n)} = 0$, τότε το σύστημα έχει άπειρες λύσεις, ενώ αν $a_{nn}^{n} = 0$ και $b_n^{(n)} \neq 0$, τότε το σύστημα είναι αδύνατο.

<u>Απαιτούμενες πράξεις και μνήμη</u>

Στο $2^{ο}$ βήμα της μεθόδου Gauss για τον υπολογισμό των στοιχείων $a_{ij}^{(2)},\ i,j=2,...,n$ απαιτούνται *(n-1)+(n-1)*^{2} πράξεις. Εύκολα βλέπουμε ότι στα *(n-1)* βήματα της μεθόδου απαιτούνται:

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

πράξεις. Επιπλέον χρειάζονται

$$(n-1)+(n-2)+...+1=\frac{n(n-1)}{2}$$

πράξεις για τον υπολογισμό των $b_i^{(i)},\ i=2,...,n$ και σημειώνουμε ότι κατά τη διαδικασία της προς τα πίσω οπισθοδρόμησης χρειάζονται

$$1+2+...+n=\frac{n(n+1)}{2}$$

πράξεις, δηλαδή συνολικά $\frac{n^3+3n^2-n}{3}$ πράξεις, κόστος δηλαδή πολύ μικρό σε σχέση με το κόστος των $n!(n-1)$ πράξεων που απαιτούνται στον κανόνα του *Cramer*. Οσον αφορά τη μνήμη, αρκεί να αποθηκεύσουμε τα στοιχεία του πίνακα A σε n^2 θέσεις μνήμης και τα στοιχεία του πίνακα b σε n θέσεις μνήμης. Επιπλέον μνήμη δεν χρειάζεται, διότι οι πολλαπλασιαστές $-\frac{a_{i1}}{a_{11}}$ $i=2,...,n$ (βλέπε βήμα 2) θα καταλάβουν τη θέση των στοιχείων $a_{1j}, j=2,...,n$ που θα γίνουν μηδενικά κλπ.

Παράδειγμα 1 Να επιλυθεί με τη μέθοδο Gauss το σύστημα:

$$\begin{aligned} 9x_1+3x_2+4x_3&=7\\ 4x_1+3x_2+4x_3&=8.\\ x_1+x_2+x_3&=3 \end{aligned}$$

Λύση Ορίζουμε τον επαυξημένο πίνακα:

$$A_{\varepsilon\pi}=\left(\begin{array}{ccc|c} 9 & 3 & 4 & 7\\ 4 & 3 & 4 & 8\\ 1 & 1 & 1 & 3 \end{array}\right).$$

1ο βήμα:

$$A_{\varepsilon\pi}(1)=\left(\begin{array}{ccc|c} \boxed{9} & 3 & 4 & 7\\ 0 & 5/3 & 20/9 & 44/9\\ 0 & 2/3 & 5/9 & 20/9 \end{array}\right),$$

όπου η 1η γραμμή του πίνακα $A_{\varepsilon\pi}(1)$ παραμένει αναλλοίωτη,

(2η γραμμή του $A_{\varepsilon\pi}(1)$*)* $= -\frac{4}{9}$ *(1η γραμμή του* $A_{\varepsilon\pi}$*) + (2η γραμμή του* $A_{\varepsilon\pi}$*),*

(3η γραμμή του $A_{\varepsilon\pi}(1)$*)* $= -\frac{1}{9}$ *(1η γραμμή του* $A_{\varepsilon\pi}$*) + (3η γραμμή του* $A_{\varepsilon\pi}$*) .*

2ο βήμα:

$$A_{\varepsilon\pi}(2)=\begin{pmatrix} 9 & 3 & 4 & | & 7 \\ 0 & \boxed{5/3} & 20/9 & | & 44/9 \\ 0 & 0 & -1/3 & | & 12/45 \end{pmatrix},$$

όπου οι 2 πρώτες γραμμές του πίνακα $A_{\varepsilon\pi}(2)$ παραμένουν αναλλοίωτες,

(3η γραμμή του $A_{\varepsilon\pi}(2)$*)* $= -\frac{2}{5}$ *(2η γραμμή του* $A_{\varepsilon\pi}(1)$*)* + *(3η γραμμή του* $A_{\varepsilon\pi}(1)$*)* .

Στη συνέχεια με τη διαδικασία της προς τα πίσω οπισθοδρόμησης, υπολογίζουμε:

$$-1/3\, x_3 = 12/45 \Rightarrow x_3 = -4/5$$

$$5/3\, x_2 + 20/9\, x_3 = 44/9 \Rightarrow x_2 = 4$$

$$9\, x_1 + 3\, x_2 + 4\, x_3 = 7 \Rightarrow x_1 = -1/5\,. \quad \square$$

<u>Αλγόριθμος Gauss με οδήγηση</u>

Στην παραπάνω διαδικασία, μπορεί να προκύψει πρόβλημα: είτε όταν το οδηγό στοιχείο σε κάποιο βήμα είναι το μηδέν, οπότε δεν είναι δυνατόν να διαγραφούν τα στοιχεία που βρίσκονται κάτω από αυτό, είτε όταν το οδηγό στοιχείο σε κάποιο βήμα έχει πολύ μικρή τιμή σε σχέση με τους άλλους αριθμούς του επαυξημένου πίνακα, οπότε δημιουργούνται σφάλματα. Για το λόγο αυτό, πριν ξεκινήσουμε να εφαρμόζουμε τη μέθοδο του Gauss, θα πρέπει να ελέγχουμε αν υπάρχουν οι παραπάνω δυσλειτουργίες. Το πρόβλημα επιλύεται με αντιμετάθεση της γραμμής, της οποίας το οδηγό στοιχείο είναι πολύ μικρό ή μηδέν, με μία άλλη γραμμή που δεν δημιουργεί τέτοια προβλήματα.

Θα ήταν ιδανικό, όλα τα στοιχεία κάτω από το οδηγό στοιχείο, να έχουν τιμές, μικρότερες κατ' απόλυτο τιμή από αυτήν του οδηγού στοιχείου. Για το λόγο αυτό ελέγχουμε την $1^η$ στήλη του επαυξημένου πίνακα, ώστε να βρούμε το μεγαλύτερο κατ' απόλυτο τιμή στοιχείο και αντιμεταθέτουμε την $1^η$ γραμμή με τη γραμμή που περιέχει το συγκεκριμένο στοιχείο. Στη συνέχεια μεταφερόμαστε στη $2^η$ στήλη και ελέγχουμε όλα τα στοιχεία κάτω του οδηγού στοιχείου, ώστε να εντοπίσουμε το μεγαλύτερο κατ' απόλυτο τιμή. Τότε αντιμεταθέτουμε

όλη τη 2^{η} γραμμή, με τη γραμμή που περιέχει το συγκεκριμένο στοιχείο και συνεχίζουμε με τον ίδιο τρόπο για να ελέγξουμε το $3^{ο}$, $4^{ο}$,…, νιοστό οδηγό στοιχείο. Το επιπλέον υπολογιστικό κόστος σε πράξεις είναι της τάξης n^2 και συνεπώς μικρό σε σχέση με το συνολικό κόστος της τριγωνοποίησης.

Παράδειγμα 2 Να επιλυθεί το σύστημα

$$\begin{aligned} 3x_1 + x_2 - 4x_3 + x_4 &= 1 \\ -5x_1 + 2x_2 + x_3 - 2x_4 &= -3 \\ -x_1 + 6x_2 - 3x_3 - 4x_4 &= 2 \\ -2x_1 + x_2 - 4x_3 + 2x_4 &= 0 \end{aligned} .$$

Λύση Επειδή $\max\{|a_{i,1}|: i = 1,...,4\} = 5$, αντιμεταθέτουμε την 1^{η} γραμμή με την 2^{η} και έχουμε:

$$\begin{aligned} -5x_1 + 2x_2 + x_3 - 2x_4 &= -3 \\ 3x_1 + x_2 - 4x_3 + x_4 &= 1 \\ -x_1 + 6x_2 - 3x_3 - 4x_4 &= 2 \\ -2x_1 + x_2 - 4x_3 + 2x_4 &= 0 \end{aligned} .$$

Επειδή $\max\{|a_{i,2}|: i = 2,...,4\} = 6$, αντιμεταθέτουμε την 3^{η} γραμμή με την 2^{η} και έχουμε:

$$\begin{aligned} -5x_1 + 2x_2 + x_3 - 2x_4 &= -3 \\ -x_1 + 6x_2 - 3x_3 - 4x_4 &= 2 \\ 3x_1 + x_2 - 4x_3 + x_4 &= 1 \\ -2x_1 + x_2 - 4x_3 + 2x_4 &= 0 \end{aligned} .$$

Επειδή $\max\{|a_{i,3}|: i = 3,...,4\} = 4$, το σύστημα παραμένει αναλλοίωτο. Συνεχίζουμε με τον ίδιο τρόπο, στη συνέχεια λύνουμε το σύστημα με τη μέθοδο απαλοιφής του Gauss όπως παραπάνω και βρίσκουμε: $x_1 = 21/10$, $x_2 = 152/15$, $x_3 = 187/30$, $x_4 = 19/2$. □

Εφαρμογή 1 (υπολογισμός ορίζουσας) Να υπολογισθεί η ορίζουσα του πίνακα $A=\begin{pmatrix} 9 & 3 & 4 \\ 4 & 3 & 4 \\ 1 & 1 & 1 \end{pmatrix}$.

Λύση Είναι γνωστό ότι η ορίζουσα ενός τριγωνικού πίνακα ισούται με το γινόμενο των στοιχείων της κυρίας διαγωνίου. Μετασχηματίζοντας τον πίνακα A σε άνω τριγωνικό με τη μέθοδο Gauss (βλέπε παράδειγμα 1) προκύπτει ότι

$$Det(A)=Det\begin{pmatrix} \boxed{9} & 3 & 4 \\ 0 & \boxed{5/3} & 20/9 \\ 0 & 0 & \boxed{-1/3} \end{pmatrix}=9\ \frac{5}{3}\left(-\frac{1}{3}\right)=-5. \quad \square$$

<u>Αλγόριθμος Gauss-Jordan</u>

Είναι μία παραλλαγή της μεθόδου του Gauss.

Βήμα 1^o : ορίζουμε τον ***επαυξημένο πίνακα*** των συντελεστών και σταθερών όρων, διάστασης $n\times(n+1)$:

$$A_{\varepsilon\pi}=\left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{array}\right).$$

Βήμα 2^o: Ακολουθούμε τη διαδικασία της μεθόδου Gauss, οπότε μετά το πέρας του βήματος αυτού ο νέος επαυξημένος πίνακας θα έχει τη μορφή:

$$A_{\varepsilon\pi}(1)=\left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} & b_n^{(2)} \end{array}\right).$$

Βήμα 3^o: Συνεχίζουμε τη διαδικασία ξεκινώντας τώρα με <u>οδηγό στοιχείο το 2^o στοιχείο της κυρίας διαγωνίου του πίνακα</u> $A_{\varepsilon\pi}(1)$, δηλαδή το $a_{22}^{(2)}$, και εκτελούμε μία πράξη μεταξύ της $2^{ης}$ γραμμής και κάθε μίας από τις

<u>υπόλοιπες γραμμές, έτσι ώστε όλα τα στοιχεία εκτός από το οδηγό στοιχείο να μηδενίζονται</u>. Μετά το πέρας του βήματος αυτού ο νέος επαυξημένος πίνακας θα έχει τη μορφή:

$$A_{\varepsilon\pi}(2)=\left(\begin{array}{ccccc|c} a_{11} & 0 & a_{13}^{(2)} & \cdots & a_{1n}^{(2)} & b_1^{(2)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3n}^{(3)} & b_3^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & a_{n3}^{(3)} & \cdots & a_{nn}^{(3)} & b_n^{(3)} \end{array}\right).$$

Συνεχίζοντας με τον ίδιο τρόπο μετά από *n* βήματα, καταλήγουμε σε έναν επαυξημένο «***διαγώνιο***» πίνακα της μορφής:

$$A_{\varepsilon\pi}(n)=\left(\begin{array}{ccccc|c} a_{11} & 0 & 0 & \cdots & 0 & b_1^{(n)} \\ 0 & a_{22}^{(2)} & 0 & \cdots & 0 & b_2^{(n)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & 0 & b_3^{(n)} \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & a_{nn}^{(n)} & b_n^{(n)} \end{array}\right),$$

ο οποίος επιτρέπει να υπολογισθούν οι λύσεις απευθείας από τις σχέσεις:

$$a_{ii}^{(i)}\, x_i \;=\; b_i^{(n)}.$$

Παράδειγμα 3 Να επιλυθεί με τη μέθοδο Gauss-Jordan το σύστημα:

$$\begin{aligned} 9x_1+3x_2+4x_3&=7 \\ 4x_1+3x_2+4x_3&=8. \\ x_1+x_2+x_3&=3 \end{aligned}$$

Λύση Ορίζουμε τον επαυξημένο πίνακα:

$$A_{\varepsilon\pi}=\left(\begin{array}{ccc|c} 9 & 3 & 4 & 7 \\ 4 & 3 & 4 & 8 \\ 1 & 1 & 1 & 3 \end{array}\right).$$

1ο βήμα:

$$A_{\varepsilon\pi}(1)=\left(\begin{array}{cccc} \boxed{9} & 3 & 4 & | & 7 \\ 0 & 5/3 & 20/9 & | & 44/9 \\ 0 & 2/3 & 5/9 & | & 20/9 \end{array}\right),$$

(βλέπε παράδειγμα 1).

2ο βήμα:

$$A_{\varepsilon\pi}(2)=\left(\begin{array}{cccc} 9 & 0 & 0 & | & -9/5 \\ 0 & \boxed{5/3} & 20/9 & | & 44/9 \\ 0 & 0 & -1/3 & | & 12/45 \end{array}\right),$$

όπου η 2η γραμμή του πίνακα $A_{\varepsilon\pi}(2)$ παραμένει αναλλοίωτη και

(1η γραμμή του $A_{\varepsilon\pi}(2)$*)* $= -\frac{9}{5}$ *(2η γραμμή του* $A_{\varepsilon\pi}(1)$*)* + *(1η γραμμή του* $A_{\varepsilon\pi}(1)$*)*

(3η γραμμή του $A_{\varepsilon\pi}(2)$*)* $= -\frac{2}{5}$ *(2η γραμμή του* $A_{\varepsilon\pi}(1)$*)* + *(3η γραμμή του* $A_{\varepsilon\pi}(1)$*)* .

3ο βήμα:

$$A_{\varepsilon\pi}(2)=\left(\begin{array}{cccc} 9 & 0 & 0 & | & -9/5 \\ 0 & 5/3 & 0 & | & 60/9 \\ 0 & 0 & -1/3 & | & 12/45 \end{array}\right),$$

όπου η 3η γραμμή του πίνακα $A_{\varepsilon\pi}(2)$ παραμένει αναλλοίωτη και

(2η γραμμή του $A_{\varepsilon\pi}(2)$*)* $= \frac{20}{3}$ *(3η γραμμή του* $A_{\varepsilon\pi}(1)$*)* + *(2η γραμμή του* $A_{\varepsilon\pi}(1)$*)* .

Στη συνέχεια υπολογίζουμε απευθείας:

$$-1/3\ x_3 = 12/45 \Rightarrow x_3 = -4/5$$

$$5/3\ x_2 = 60/9 \Rightarrow\ x_2 = 4$$

$$9\ x_1 = -9/5 \Rightarrow\ x_1 = -1/5\,.\ \square$$

Εφαρμογή 2 (υπολογισμός αντίστροφου πίνακα) Να υπολογισθεί ο αντίστροφος του πίνακα

$$A=\begin{pmatrix}9 & 3 & 4\\ 4 & 3 & 4\\ 1 & 1 & 1\end{pmatrix}.$$

Λύση Θεωρούμε τον επαυξημένο πίνακα

$$A_{\varepsilon\pi}=\left(A \;\;|\;\; I_3\right)=\left(\begin{array}{ccc|ccc}9 & 3 & 4 & 1 & 0 & 0\\ 4 & 3 & 4 & 0 & 1 & 0\\ 1 & 1 & 1 & 0 & 0 & 1\end{array}\right),$$

όπου I_3 είναι ο μοναδιαίος πίνακας 3×3. Με τη μέθοδο Gauss-Jordan μετασχηματίζουμε τον επαυξημένο πίνακα μετά από 3 βήματα στη μορφή

$$A_{\varepsilon\pi}(3)=\left(D_3 \;\;|\;\; B\right)=\left(\begin{array}{ccc|ccc}a_{11}^{(1)} & 0 & 0 & b_{11} & b_{12} & b_{13}\\ 0 & a_{22}^{(2)} & 0 & b_{21} & b_{22} & b_{23}\\ 0 & 0 & a_{33}^{(3)} & b_{31} & b_{32} & b_{33}\end{array}\right),$$

όπου D_3 είναι ένας διαγώνιος πίνακας διάστασης 3×3. Τότε αν $A^{-1}=(c_{ij})_{i,j=1,\ldots 3}$, έχουμε:

$$c_{ij}=\frac{b_{ij}}{a_{ii}^{(i)}}.$$

Πράγματι έχουμε:

$$A_{\varepsilon\pi}(1)=\left(\begin{array}{ccc|ccc}9 & 3 & 4 & 1 & 0 & 0\\ 0 & 5/3 & 20/9 & -4/9 & 1 & 0\\ 0 & 2/3 & 5/9 & -1/9 & 0 & 1\end{array}\right), \qquad (\mathbf{1^o\ βήμα})$$

$$A_{\varepsilon\pi}(2)=\left(\begin{array}{ccc|ccc}9 & 0 & 0 & 9/5 & -9/5 & 0\\ 0 & 5/3 & 20/9 & -4/9 & 1 & 0\\ 0 & 0 & -1/3 & 1/15 & -2/5 & 1\end{array}\right), \qquad (\mathbf{2^o\ βήμα})$$

$$A_{\varepsilon\pi}(3)=\left(\begin{array}{ccc|ccc} 9 & 0 & 0 & 9/5 & -9/5 & 0 \\ 0 & 5/3 & 0 & 0 & -5/3 & 20/3 \\ 0 & 0 & -1/3 & 1/15 & -2/5 & 1 \end{array}\right) \quad (3^{\mathbf{o}}\ \mathbf{βήμα}).$$

Διαιρούμε όλα τα στοιχεία της *i*- γραμμής του πίνακα εκ δεξιών της διακεκκομένης γραμμής με το μη μηδενικό στοιχείο της *i*- γραμμής του διαγώνιου πίνακα και παίρνουμε:

$$A^{-1}=\left(\begin{array}{ccc} 1/5 & -1/5 & 0 \\ 0 & -1 & 4 \\ -1/5 & 6/5 & -3 \end{array}\right). \quad \square$$

§ 3.2 *Δείκτης κατάστασης πίνακα*

Μία κατηγορία συστημάτων που παρουσιάζουν ενδιαφέρον είναι τα λεγόμενα ασταθή ή κακώς ορισμένα. Ένα τέτοιο παράδειγμα είναι το σύστημα

$$x_1 + 2x_2 = 2$$
$$2x_1 + 3.999x_2 = 4.001'$$

το οποίο έχει μοναδική λύση $x_1 = 4$, $x_2 = -1$. Το ελάχιστα διαφορετικό σύστημα

$$x_1 + 2x_2 = 2$$
$$2x_1 + 4.001x_2 = 4.001$$

έχει την τελείως διαφορετική μοναδική λύση $x_1 = 0$, $x_2 = 1$. Τέτοια συστήματα που είναι πολύ ευαίσθητα στις αλλαγές των δεδομένων τους, καλούνται ασταθή ή κακώς ορισμένα και η επίλυσή τους με αριθμητικές μεθόδους θα πρέπει να γίνεται με ιδιαίτερη προσοχή και έμφαση στον περιορισμό των σφαλμάτων στρογγύλευσης. Ένα χρήσιμο εργαλείο διερεύνησης της ευστάθειας ή αστάθειας ενός πίνακα είναι ο δείκτης κατάστασης πίνακα, για τον ορισμό του οποίου απαιτείται προηγουμένως να ορισθεί η έννοιας της νόρμας πινάκων.

Ορισμός 3.2.1 Εστω X ένας διανυσματικός χώρος πάνω από το σύνολο $\boldsymbol{R}$ *ή* $\boldsymbol{C}$ των πραγματικών ή μιγαδικών αριθμών αντίστοιχα. Εστω $K = \boldsymbol{R}$ ή $\boldsymbol{C}$. Μία απεικόνιση:

$$\|.\| \ : X \to \boldsymbol{R}^+, x \to \|x\|$$

καλείται *νόρμα*, αν ισχύουν:

- $\|x\| = 0 \Leftrightarrow x = 0$,
- $\|\lambda\, x\| = |\,\lambda\,|\,\|\,x\|$ για κάθε $\lambda \in K$,
- Για κάθε $x,y \in X$, $\|\,x + y\| \le \|\,x\| + \|\,y\|$.

Από τα παραπάνω, φαίνεται ότι η νόρμα παίζει το ρόλο της απόλυτης τιμής σε διανυσματικούς χώρους.

Παραδείγματα: Θεωρούμε το διανυσματικό χώρο $\boldsymbol{R}^n = \{x: x = (x_1,...,x_n)\}$, τότε οι ακόλουθες είναι νόρμες του $\boldsymbol{R}^n$:

1. $\|x\|_\infty = \max\{|\,x_i\,|,\ i = 1,...,N\}$ *(νόρμα μεγίστου).*

2. $\|x\|_1 = \sum_{i=1}^{n} |\,x_i\,|$ *(l_1 νόρμα).*

3. $\|x\|_2 = \left(\sum_{i=1}^{n} |\,x_i\,|^2 \right)^{1/2}$ *(l_2 νόρμα ή ευκλείδια νόρμα).*

Εστω $\boldsymbol{R}^{n,n}$ είναι ο διανυσματικός χώρος των πραγματικών πινάκων διάστασης $n \times n$, τότε μία απεικόνιση $\|.\| : \boldsymbol{R}^{n,n} \to \boldsymbol{R}^+$ που πληροί τα αξιώματα του ορισμού 3.2.1 και επιπλέον $\|A\,B\| \le \|A\|\,\|B\|$ για κάθε $A, B \in \boldsymbol{R}^{n,n}$, καλείται *νόρμα πινάκων*.

Ορισμός 3.2.2 Εστω $\|.\|$ μία νόρμα στο χώρο $\boldsymbol{R}^n$, η απεικόνιση:

$$\|.\| \ : \boldsymbol{R}^{n,n} \to \boldsymbol{R}^+, \|A\| = \boldsymbol{sup}_{x \in \boldsymbol{R}^n} \frac{\|Ax\|}{\|x\|}$$

καλείται *φυσική νόρμα πινάκων*.

Παραδείγματα:

1. Θεωρούμε στο χώρο $\boldsymbol{R}^n$ τη νόρμα μεγίστου $\|x\|_\infty$, τότε η παραγόμενη από την $\|x\|_\infty$ φυσική νόρμα στον $\boldsymbol{R}^{n,n}$ είναι η εξής:

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

2. Θεωρούμε στο χώρο $\boldsymbol{R}^n$ την l_1-νόρμα $\|x\|_1$, τότε η παραγόμενη από την $\|x\|_1$ φυσική νόρμα στον $\boldsymbol{R}^{n,n}$ είναι η εξής:

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{n} |\, a_{ij}\,|.$$

3. Θεωρούμε στο χώρο $\boldsymbol{R}^n$ την l_2-νόρμα $\|x\|_2$ και έστω $\rho(A)$ είναι η *φασματική ακτίνα του πίνακα A*, που ορίζεται ως το μέγιστο των απολύτων τιμών των ιδιοτιμών του πίνακα A, τότε η παραγόμενη από την $\|x\|_2$ φυσική νόρμα στον $\boldsymbol{R}^{n,n}$ είναι η εξής:

$$\|A\|_2 = \left(\rho\left(A^T A\right)\right)^{1/2},$$

όπου A^T είναι ο ανάστροφος του πίνακα A.

Επιστρέφουμε τώρα στο πρόβλημα της κατάστασης των γραμμικών συστημάτων, δηλαδή στη μελέτη της ευαισθησίας των λύσεων του συστήματος

$$A\,x \;=\; b$$

σε διαταραχές των δεδομένων $A \in \boldsymbol{R}^{n,n}$ και $b \in \boldsymbol{R}^n$. Ας αφήσουμε προς στιγμήν τον πίνακα A σταθερό και ας μεταβάλλουμε το διάνυσμα στήλη b, τότε αν $x + \Delta x$ είναι η λύση του διαταραγμένου συστήματος, έχουμε:

$$A\,\left(x+\Delta x\right) \;=\; b+\Delta b \Rightarrow A\,\Delta x \;=\; \Delta b \Rightarrow \Delta x = A^{-1}\,\Delta b,$$

άρα:

$$\|\Delta x\| = \left\|A^{-1}\,\Delta b\right\| \;\le\; \left\|A^{-1}\right\|\,\|\Delta b\|$$

και εφόσον

$$\|b\| \;=\; \|A\,x\| \;\le\; \|A\|\;\|x\| \Rightarrow \frac{1}{\|x\|} \;\le\; \frac{\|A\|}{\|b\|},$$

από το συνδυασμό των παραπάνω ανισοτήτων προκύπτει ότι:

$$\frac{\|\Delta x\|}{\|x\|} \;\le\; \|A\|\;\left\|A^{-1}\right\|\;\frac{\|\Delta b\|}{\|b\|} \Rightarrow \rho_{\Delta x} \le \|A\|\;\left\|A^{-1}\right\|\;\rho_{\Delta b}$$

όπου $\rho_{\Delta x}$, $\rho_{\Delta b}$ τα αντίστοιχα σχετικά σφάλματα. Η ποσότητα

$$\kappa(A) = \|A\|\;\left\|A^{-1}\right\|$$

είναι ένας συντελεστής ευαισθησίας που προσδιορίζει τη μέγιστη δυνατή μεταβολή του σχετικού σφάλματος των αποτελεσμάτων σε σχέση με το σχετικό σφάλμα δεδομένων και καλείται ***δείκτης κατάστασης πίνακα A***, είναι δε πάντα μεγαλύτερος της μονάδας. Αν $\kappa(A) >> 1$, τότε λέμε ότι το πρόβλημα είναι σε κακή κατάσταση. Προφανώς ο δείκτης κατάστασης ορίζεται μόνον για αντιστρέψιμους πίνακες. Αν ο πίνακας A τείνει να γίνει μη αντιστρέψιμος, τότε ο δείκτης κατάστασης αυτού τείνει στο άπειρο. Ο δείκτης κατάστασης $\kappa(A)$ καθορίζει επίσης και το πώς διαταραχές του πίνακα A επηρεάζουν τη λύση.

Θεώρημα 3.2.1

(i) αν $\left(A+\Delta A\right)\left(x+\Delta x\right) = b$ και αν $\left\|A^{-1}\right\|\;\|\Delta A\| < 1$, τότε ο πίνακας $A+\Delta A$ είναι αντιστρέψιμος και ισχύει:

$$\rho_{\Delta x} \le \frac{\kappa(A)}{1-\left\|A^{-1}\right\|\;\|\Delta A\|}\;\rho_{\Delta A}\,.$$

(ii) αν $\left(A+\Delta A\right)\left(x+\Delta x\right) = b+\Delta b$ και αν $\left\|A^{-1}\right\|\;\|\Delta A\| < 1$, τότε ο πίνακας $A+\Delta A$ είναι αντιστρέψιμος και ισχύει:

$$\rho_{\Delta x} \le \frac{\kappa(A)}{1-\left\|A^{-1}\right\|\;\|\Delta A\|}\;\left(\rho_{\Delta A}+\rho_{\Delta b}\right).$$

§ 3.3 *Επαναληπτικές μέθοδοι επίλυσης συστημάτων*

Στην παράγραφο αυτή θα ασχοληθούμε με την αριθμητική επίλυση γραμμικών συστημάτων με τις λεγόμενες επαναληπτικές μεθόδους, όπου ξεκινώντας από μία αυθαίρετη προσέγγιση της λύσης του συστήματος, κατασκευάζουμε μια ακολουθία διαδοχικών προσεγγίσεων της λύσης, η οποία υπό προϋποθέσεις συγκλίνει στην πραγματική λύση.

Μέθοδος Jacobi

Θεωρούμε το σύστημα (3.1) και λύνουμε την *i*-εξίσωση ως προς τον άγνωστο x_i, οπότε:

$$x_i = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j - \sum_{j=i+1}^{n} a_{ij}x_j \right),\ i = 1,...,n\,.$$

Για τυχαία δεδομένη αρχική τιμή $x^{(0)} = \left(x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)}\right)$, η αναδρομική ακολουθία για τον υπολογισμό της λύσης του συστήματος είναι η ακόλουθη:

$$x_i^{(m+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(m)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(m)} \right),\ i,= 1,...,n,\ m = 1,.... \qquad (3.2)$$

Μέθοδος Gauss-Seidel

Θεωρούμε το σύστημα (3.1) και λύνουμε την *i*-εξίσωση ως προς τον άγνωστο x_i, οπότε:

$$x_i = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j - \sum_{j=i+1}^{n} a_{ij}x_j \right),\ i = 1,...,n\,.$$

Για τυχαία δεδομένη αρχική τιμή $x^{(0)} = \left(x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)}\right)$, η αναδρομική ακολουθία για τον υπολογισμό της λύσης του συστήματος είναι η ακόλουθη:

$$x_i^{(m+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(m+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(m)} \right),\ i,= 1,...,n,\ m = 1,.... \qquad (3.3)$$

Η διαφορά από τη μέθοδο Jacobi είναι, ότι για τον υπολογισμό της συνιστώσας $x_i^{(m+1)}$ χρησιμοποιούμε τις ήδη υπολογισθείσες τιμές $x_1^{(m+1)}, x_2^{(m+1)}, \ldots, x_{i-1}^{(m+1)}$ της ίδιας γενιάς.

Για τη ***<u>σύγκλιση των δύο μεθόδων</u>*** ισχύει το ακόλουθο:

Θεώρημα 3.3.1 Εστω ότι ο πίνακας *A* των συντελεστών των αγνώστων ενός γραμμικού συστήματος έχει *κυριαρχική διαγώνιο*, δηλαδή:

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \; i, j = 1, \ldots, n,$$

τότε οι μέθοδοι Jacobi και Gauss-Seidel συγκλίνουν.

Παράδειγμα 4 Να επιλυθεί με τη μέθοδο Jacobi το σύστημα:

$$\begin{aligned} 8x_1 + x_2 + x_3 &= 10 \\ x_1 + 8x_2 + x_3 &= 10, \\ x_1 + x_2 + 8x_3 &= 10 \end{aligned}$$

με ακρίβεια 2 δεκαδικών ψηφίων.

Λύση Παρατηρούμε ότι ο πίνακας των συντελεστών των αγνώστων έχει κυριαρχική διαγώνιο. Πράγματι:

$$8 = |a_{ii}| \geq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| = 2,$$

άρα η μέθοδος Jacobi συγκλίνει για κάθε αρχική τιμή της λύσης. Θεωρούμε αυθαίρετα ότι $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)}) = (0,0,0)$, τότε υπολογίζουμε μία νέα προσέγγιση της λύσης $x^{(1)} = (x_1^{(1)}, x_2^{(1)}, x_3^{(1)})$ από την σχέση (3.2) για *m = 0* :

$$x_1^{(1)} = \frac{10}{8} - \frac{1}{8}x_2^{(0)} - \frac{1}{8}x_3^{(0)} = 1.25$$
$$x_2^{(1)} = \frac{10}{8} - \frac{1}{8}x_1^{(0)} - \frac{1}{8}x_2^{(0)} = 1.25,$$
$$x_3^{(1)} = \frac{10}{8} - \frac{1}{8}x_1^{(0)} - \frac{1}{8}x_2^{(0)} = 1.25$$

άρα: $x^{(1)} = (x_1^{(1)}, x_2^{(1)}, x_3^{(1)}) = (1.25, 1.25, 1.25)$. Συνεχίζουμε για *m = 1* και παίρνουμε:

$$x_1^{(2)} = \frac{10}{8} - \frac{1}{8}x_2^{(1)} - \frac{1}{8}x_3^{(1)} = 0.9375$$
$$x_2^{(2)} = \frac{10}{8} - \frac{1}{8}x_1^{(1)} - \frac{1}{8}x_2^{(1)} = 0.9375,$$
$$x_3^{(2)} = \frac{10}{8} - \frac{1}{8}x_1^{(1)} - \frac{1}{8}x_2^{(1)} = 0.9375$$

άρα: $x^{(2)} = (x_1^{(2)}, x_2^{(2)}, x_3^{(2)}) = (0.9375, 0.9375, 0.9375)$. Συνεχίζουμε για *m = 2* και παίρνουμε

$$x_1^{(3)} = \frac{10}{8} - \frac{1}{8}x_2^{(2)} - \frac{1}{8}x_3^{(2)} = 1.015625$$
$$x_2^{(3)} = \frac{10}{8} - \frac{1}{8}x_1^{(2)} - \frac{1}{8}x_2^{(2)} = 1.015625,$$
$$x_3^{(3)} = \frac{10}{8} - \frac{1}{8}x_1^{(2)} - \frac{1}{8}x_2^{(2)} = 1.015625$$

άρα: $x^{(3)} = (x_1^{(3)}, x_2^{(3)}, x_3^{(3)}) = (1.015625, 1.015625, 1.015625)$. Συνεχίζοντας υπολογίζουμε ότι στην 4^{η} επανάληψη έχουμε ότι

$$x^{(4)} = (x_1^{(4)}, x_2^{(4)}, x_3^{(4)}) = (0.99609375, 0.99609375, 0.99609375),$$

ενώ στην 5^{η} επανάληψη έχουμε ότι

$$x^{(5)} = (x_1^{(5)}, x_2^{(5)}, x_3^{(5)}) = (1.0009766, 1.0009766, 1.0009766).$$

Σταματάμε στην 5^{η} επανάληψη διότι

$$|| x^{(5)} - x^{(4)} ||_\infty = |1.0009766 - 0.99609375| \le 0.005 . \quad \square$$

Παράδειγμα 5 Να επιλυθεί με τη μέθοδο Gauss-Seidel το σύστημα:

$$\begin{aligned} 8x_1 + x_2 + x_3 &= 10 \\ x_1 + 8x_2 + x_3 &= 10 , \\ x_1 + x_2 + 8x_3 &= 10 \end{aligned}$$

με ακρίβεια 2 δεκαδικών ψηφίων μεταξύ διαδοχικών λύσεων.

Λύση Παρατηρούμε ότι ο πίνακας των συντελεστών των αγνώστων έχει κυριαρχική διαγώνιο. Πράγματι:

$$8 \;=\; |a_{ii}| \;\ge\; \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \;=\; 2 ,$$

άρα η μέθοδος Gauss-Seidel συγκλίνει για κάθε αρχική τιμή της λύσης. Θεωρούμε αυθαίρετα ότι $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)}) = (0,0,0)$, τότε υπολογίζουμε μία νέα προσέγγιση της λύσης $x^{(1)} = (x_1^{(1)}, x_2^{(1)}, x_3^{(1)})$ από την σχέση (3.3) για *m = 0* :

$$\begin{aligned} x_1^{(1)} &= \frac{10}{8} - \frac{1}{8} x_2^{(0)} - \frac{1}{8} x_3^{(0)} = 1.25 \\ x_2^{(1)} &= \frac{10}{8} - \frac{1}{8} x_1^{(1)} - \frac{1}{8} x_2^{(0)} = 1.09375 , \\ x_3^{(1)} &= \frac{10}{8} - \frac{1}{8} x_1^{(1)} - \frac{1}{8} x_2^{(1)} = 0.957031 \end{aligned}$$

άρα: $x^{(1)} = (x_1^{(1)}, x_2^{(1)}, x_3^{(1)}) = (1.25, 1.09375, 0.957031)$. Συνεχίζουμε για *m = 1* και παίρνουμε:

$$\begin{aligned} x_1^{(2)} &= \frac{10}{8} - \frac{1}{8} x_2^{(1)} - \frac{1}{8} x_3^{(1)} = 0.993652 \\ x_2^{(2)} &= \frac{10}{8} - \frac{1}{8} x_1^{(2)} - \frac{1}{8} x_2^{(1)} = 1.0061646 , \\ x_3^{(2)} &= \frac{10}{8} - \frac{1}{8} x_1^{(2)} - \frac{1}{8} x_2^{(2)} = 1.0000229 \end{aligned}$$

άρα: $x^{(2)} = (x_1^{(2)}, x_2^{(2)}, x_3^{(2)}) = (0.993652, 1.0061646, 1.0000229)$. Συνεχίζουμε για $m = 2$ και παίρνουμε

$$x_1^{(3)} = \frac{10}{8} - \frac{1}{8}x_2^{(2)} - \frac{1}{8}x_3^{(2)} = 0.9992266$$
$$x_2^{(3)} = \frac{10}{8} - \frac{1}{8}x_1^{(3)} - \frac{1}{8}x_2^{(2)} = 1.0000938\,,$$
$$x_3^{(3)} = \frac{10}{8} - \frac{1}{8}x_1^{(3)} - \frac{1}{8}x_2^{(3)} = 1.0000849$$

άρα: $x^{(3)} = (x_1^{(3)}, x_2^{(3)}, x_3^{(3)}) = (0.9992266, 1.0000938, 1.0000849)$. Συνεχίζοντ ας, υπολογίζουμε ότι στην 4^{η} επανάληψη έχουμε ότι

$$x^{(4)} = (x_1^{(4)}, x_2^{(4)}, x_3^{(4)}) = (0.999978, 0.99999218, 1.0000037),$$

όπου και σταματάμε διότι:

$$\| x^{(4)} - x^{(3)} \|_{\infty} \leq 0.005\,. \quad \square$$

ΑΛΥΤΕΣ ΑΣΚΗΣΕΙΣ

1. Να επιλυθούν με τη μέθοδο Gauss-Jordan τα συστήματα εξισώσεων:

$$(a)\ \begin{matrix} 2x_1 - 3x_2 + x_3 = 1 \\ 3x_1 + x_2 - x_3 = 2 \\ x_1 - x_2 - x_3 = 1 \end{matrix}, \quad (b)\ \begin{matrix} x_1 - 5x_2 + x_3 = 2 \\ 2x_1 + 4x_2 + x_3 = 1 \\ x_1 + x_2 + x_3 = 0 \end{matrix},$$

Επίσης να υπολογισθούν οι ορίζουσες του πίνακα των συντελεστών των αγνώστων των δύο συστημάτων καθώς επίσης και ο αντίστροφος πίνακας.

Απάντ. (α) *$x_1 = 4/7$, $x_2 = -1/14$, $x_3 = -5/14$.*
(b) *$x_1 = 2$, $x_2 = -1/3$, $x_3 = -5/3$.*

Για τον υπολογισμό της ορίζουσας βλέπε εφαρμογή 1

(α) *Det(A) = -14.*
(b) *Det(A) = 6.*

Για τον υπολογισμό του αντίστροφου πίνακα βλέπε εφαρμογή 2.

$$(a)\ A^{-1}=\begin{pmatrix} 1/7 & 2/7 & -1/7 \\ -1/7 & 3/14 & -5/14 \\ 2/7 & 1/14 & -11/14 \end{pmatrix}, \quad (b)\ A^{-1}=\begin{pmatrix} 1/2 & 1 & -3/2 \\ -1/6 & 0 & 1/6 \\ -1/3 & -1 & 7/3 \end{pmatrix}.$$

2. Να επιλυθούν με τη μέθοδο Jacobi τα συστήματα:

$$\begin{aligned} 3x_1 + x_2 - 2_3 &= 1/2 \\ 2x_1 - 4x_2 - 2x_3 &= 1, \\ x_1 + x_2 + 3x_3 &= -2 \end{aligned} \qquad \begin{aligned} -5x_1 + x_2 - 3_3 &= 1 \\ 2x_1 - 4x_2 - x_3 &= 2, \\ x_1 + x_2 + 3x_3 &= -1 \end{aligned}$$

έτσι ώστε το σφάλμα μεταξύ δύο διαδοχικών προσεγγίσεων να είναι μικρότερο του 0.1. Οι πράξεις να γίνουν με στρογγυλοποίηση στο 6 δεκαδικό ψηφίο.

Απάντ: (βλέπε παράδειγμα 4)

(α) Χρειαζόμαστε 5 επαναλήψεις. Τότε

$$x^{(6)} = (\text{-}0.175412,\ \text{-}0.104938,\ \text{-}0.552984)$$

και

$$\| x^{(5)} - x^{(4)} \| \leq 0.0869342.$$

(β) Χρειαζόμαστε 4 επαναλήψεις. Τότε

$$x^{(4)} = (\text{-}0.228333,\ \text{-}0.589722,\ \text{-}0.0772222)$$

και

$$\| x^{(4)} - x^{(3)} \| \leq 0.0647222.$$

3. Να επιλυθούν με τη μέθοδο Gauss-Seidel τα συστήματα:

$$\begin{aligned} 3x_1 + x_2 - 2_3 &= 1/2 \\ 2x_1 - 4x_2 - 2x_3 &= 1, \\ x_1 + x_2 + 3x_3 &= -2 \end{aligned} \qquad \begin{aligned} -5x_1 + x_2 - 3_3 &= 1 \\ 2x_1 - 4x_2 - x_3 &= 2, \\ x_1 + x_2 + 3x_3 &= -1 \end{aligned}$$

έτσι ώστε το σφάλμα μεταξύ δύο διαδοχικών προσεγγίσεων να είναι μικρότερο του 0.1. Οι πράξεις να γίνουν με στρογγυλοποίηση στο 6 δεκαδικό ψηφίο.

Απάντ: (βλέπε παράδειγμα 5)

(α) Χρειαζόμαστε 3 επαναλήψεις. Τότε

$$x^{(3)} = (\text{-}0.212963, \text{-}0.0648148, \text{-}0.574074)$$

και

$$|| x^{(3)} - x^{(2)} || \leq 0.037037.$$

(β) Χρειαζόμαστε 2 επαναλήψεις. Τότε

$$x^{(2)} = (\text{-}0.28, \text{-}0.623333, \text{-}0.0322222)$$

και

$$|| x^{(2)} - x^{(1)} || \leq 0.08.$$

3. Προσδιορίστε τις νόρμες $||.||_\infty$, $||.||_1$, $||.||_2$ των πινάκων:

$$A = \begin{pmatrix} 1 & 0 & 3 \\ -1 & 0 & 1 \\ -2 & -1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

4. Προσδιορίστε το δείκτη κατάστασης του πίνακα:

$$A = \begin{pmatrix} 0.78 & 0.563 \\ 0.913 & 0.659 \end{pmatrix}.$$

5. Εστω $A\,x = b$ η ακριβής λύση ενός συστήματος (A αντιστρέψιμος) και $\tilde{x}$ μία προσεγγιστική λύση, ώστε $r = A\,\tilde{x} - b$ να είναι το υπόλοιπο. Να δείξετε ότι για κάθε νόρμα πινάκων $||.||$ ισχύει:

$$\frac{|| \tilde{x} - x ||}{|| x ||} \leq \kappa(A) \frac{|| r ||}{|| b ||},$$

όπου *κ(A)* είναι ο δείκτης κατάστασης πίνακα Α.

ΚΕΦΑΛΑΙΟ 4

ΑΡΙΘΜΗΤΙΚΕΣ ΜΕΘΟΔΟΙ ΕΥΡΕΣΗΣ ΠΡΑΓΜΑΤΙΚΩΝ ΙΔΙΟΤΙΜΩΝ

§ 4.1 *Γραμμικοί μετασχηματισμοί-Ιδιοτιμές-Ιδιοδιανύσματα*

Εστω $\boldsymbol{R}^n$ είναι ο γνωστός n-διάστατος πραγματικός διανυσματικός χώρος. Μία απεικόνιση $L:\boldsymbol{R}^n \to \boldsymbol{R}^n$ καλείται *γραμμική απεικόνιση* ή *γραμμικός μετασχηματισμός*, αν για κάθε $x, y \in \boldsymbol{R}^n$ και $a, b \in \boldsymbol{R}$ ισχύει

$$L\left(a\,x + b\,y\right) = a\,L(x) + b\,L(y)\,.$$

Είναι επίσης γνωστό ότι σε κάθε γραμμικό μετασχηματισμό αντιστοιχεί ένα μοναδικός πίνακας A έτσι ώστε:

$$L(x) = A\,x\,. \tag{4.1}$$

Εφόσον x, $L(x)$ είναι διανύσματα του $\boldsymbol{R}^n$, μπορούμε να πούμε ότι η δράση του πίνακα Α σε ένα διάνυσμα x (βλέπε (4.1)) έχει ως συνέπεια την κατασκευή ενός νέου διανύσματος $L(x)$, το οποίο εν γένει διαφέρει από το x, τόσο κατά μέτρο όσο και κατά διεύθυνση. Ωστόσο υπάρχουν ορισμένα διανύσματα με την εξής ιδιότητα: *η δράση του γραμμικού μετασχηματισμού (4.1) επιφέρει μεταβολή μόνον του μέτρου τους χωρίς να μεταβάλλεται καθόλου η διεύθυνσή τους*. Τέτοια διανύσματα καλούνται ***ιδιοδιανύσματα***. *Ο λόγος των μέτρων ενός ιδιοδιανύσματος μετά και πριν τη δράση του πίνακα A* σε αυτό, καλείται ***ιδιοτιμή*** του πίνακα A.

Προκύπτει λοιπόν κατά φυσικό τρόπο ο ακόλουθος:

Ορισμός 4.1.1 Εστω A πίνακας διάστασης $n \times n$ και b ένα μη μηδενικό διάνυσμα στήλη. Το b καλείται πραγματικό ιδιοδιάνυσμα του πίνακα A, αν και μόνον αν υπάρχει πραγματικός αριθμός λ τέτοιος ώστε:

$$A\,b = \lambda\,b\,.$$

Ο αριθμός λ καλείται ιδιοτιμή του πίνακα A.

Υπενθυμίζουμε ότι οι ιδιοτιμές ενός τετραγωνικού πίνακα A είναι οι ρίζες της εξίσωσης

$$Det(A-\lambda\ I_n)=0,$$

όπου I_n είναι ο μοναδιαίος πίνακας διάστασης $n\times n$. Στη συνέχεια αντικαθιστούμε κάθε τιμή του λ στην εξίσωση

$$\left(A-\lambda\ I_n\right)x=\boldsymbol{0},$$

όπου $\boldsymbol{0}$ είναι ο μηδενικός πίνακας στήλη και $x=\begin{pmatrix}x_1\\ \vdots\\ x_n\end{pmatrix}$ διάνυσμα στήλη και λύνουμε το αντίστοιχο ομογενές σύστημα. Οι λύσεις που προκύπτουν είναι τα ιδιοδιανύσματα που αντιστοιχούν στην εκάστοτε ιδιοτιμή.

Παράδειγμα 1 Να υπολογισθούν οι ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα

$$A=\begin{pmatrix}1 & 4\\ 4 & 1\end{pmatrix}.$$

Λύση Οι ιδιοτιμές είναι οι ρίζες της εξίσωσης:

$$Det(A-\lambda\ I_n)=0\Leftrightarrow\begin{vmatrix}1-\lambda & 4\\ 4 & 1-\lambda\end{vmatrix}=0\Leftrightarrow\left(1-\lambda\right)^2-4^2=0$$

$$\lambda=-3\ \text{ή}\ \lambda=5.$$

- Για $\lambda=-3$ αντικαθιστούμε στην εξίσωση

$$\left(A-\lambda\ I_n\right)x=\boldsymbol{0}\Rightarrow\left(A+3\ I_n\right)x=\boldsymbol{0}\Rightarrow\begin{pmatrix}4 & 4\\ 4 & 4\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix}$$

$$\begin{matrix}4x_1+4x_2=0\\ 4x_1+4x_2=0\end{matrix}\Rightarrow 4x_1+4x_2=0\Rightarrow x_1=-x_2,$$

άρα:

$$x=\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}-x_2\\ x_2\end{pmatrix}=x_2\begin{pmatrix}-1\\ 1\end{pmatrix},$$

συνεπώς το σύνολο $\left\{ x_2 \begin{pmatrix} -1 \\ 1 \end{pmatrix} : x_2 \in \boldsymbol{R} \right\}$ είναι το σύνολο των ιδιοδιανυσμάτων που αντιστοιχούν στην ιδιοτιμή $\lambda = -3$. Ομοίως:

- Για $\lambda = 5$ αντικαθιστούμε στην εξίσωση

$$\left(A - \lambda\, I_n\right)x = \boldsymbol{0} \Rightarrow \left(A - 5\, I_n\right)x = \boldsymbol{0} \Rightarrow \begin{pmatrix} -4 & 4 \\ 4 & -4 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{matrix} -4x_1 + 4x_2 = 0 \\ -4x_1 + 4x_2 = 0 \end{matrix} \Rightarrow -4x_1 + 4x_2 = 0 \Rightarrow x_1 = x_2\,,$$

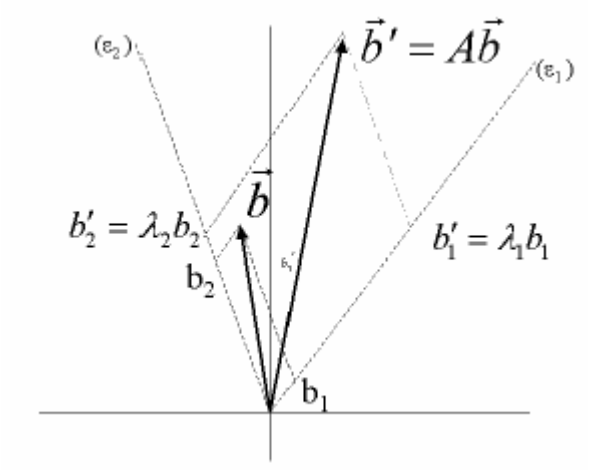
άρα:

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_2 \\ x_2 \end{pmatrix} = x_2 \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

συνεπώς το σύνολο $\left\{ x_2 \begin{pmatrix} 1 \\ 1 \end{pmatrix} : x_2 \in \boldsymbol{R} \right\}$ είναι το σύνολο των ιδιοδιανυσμάτων που αντιστοιχούν στην ιδιοτιμή $\lambda = 5$. □

§ 3.2 *Αριθμητική εύρεση της απολύτως μεγαλύτερης πραγματικής ιδιοτιμής*

Η σημασία της εύρεσης της απολύτως μεγαλύτερης πραγματικής ιδιοτιμής βασίζεται στην ακόλουθη γεωμετρική παρατήρηση:



Σχήμα 4 Μέσω της δράσης του πίνακα Α ένα τυχαίο διάνυσμα έλκεται προς το φορέα των ιδιοδιανυσμάτων που αντιστοιχούν στη μεγαλύτερη κατ' απόλυτο τιμή ιδιοτιμή

Εστω ένα τυχαίο διάνυσμα b (όχι ιδιοδιάνυσμα) και (ε_1) ο φορέας των ιδιοδιανυσμάτων με ιδιοτιμή λ_1 και (ε_2) ο φορέας των ιδιοδιανυσμάτων με ιδιοτιμή λ_2, όπου $|\lambda_1| > |\lambda_2|$. Αν αναλύσουμε το διάνυσμα b σε δύο συνιστώσες b_1 και b_2 επί των φορέων (ε_1) και (ε_2) αντίστοιχα, η δράση του πίνακα A επί των διανυσμάτων b_1 και b_2, δημιουργεί δύο νέα διανύσματα b_1', b_2':

$$b_1' = A\, b_1 = \lambda_1\, b_1, \quad b_2' = A\, b_2 = \lambda_2\, b_2 .$$

Εχουμε λοιπόν:

$$\widehat{\varepsilon\phi(b,b_1)} = \frac{\| b_2 \|}{\| b_1 \|}$$

$$\widehat{\varepsilon\phi(b',b_1)} = \frac{\| b_2' \|}{\| b_1' \|} = \frac{|\lambda_2|}{|\lambda_1|}\ \frac{\| b_2 \|}{\| b_1 \|}$$

και επειδή $|\lambda_1| > |\lambda_2|$ ισχύει

$$\widehat{\varepsilon\phi(b',b_1)} < \widehat{\varepsilon\phi(b,b_1)} \Rightarrow \widehat{(b',b_1)} < \widehat{(b,b_1)},$$

άρα αφού η γωνία μεταξύ των διανυσμάτων b', b_1 είναι μικρότερη της γωνίας μεταξύ των διανυσμάτων b, b_1 *συμπεραίνουμε ότι η δράση ενός πίνακα A με πραγματικές ιδιοτιμές επί τυχαίου διανύσματος με μη μηδενικές προβολές προς τις ιδιοδιευθύνσεις, προκαλεί στροφή του τυχαίου διανύσματος προς την κατεύθυνση του ιδιοδιανύσματος με την απολύτως μεγαλύτερη ιδιοτιμή*.

Είναι σαφές ότι με διαδοχικές δράσεις του πίνακα A επί του διανύσματος b, η διεύθυνση του b τείνει να ταυτισθεί με την διεύθυνση του φορέα (ε_1). Ουσιαστικά λοιπόν η ιδιοδιεύθυνση με την απολύτως μεγαλύτερη ιδιοτιμή έλκει όλα τα διανύσματα του χώρου, με την έννοια ότι διαδοχικές δράσεις του πίνακα A σε σχεδόν οποιοδήποτε διάνυσμα έχουν ως συνέπεια να στραφεί το διάνυσμα ώστε η διεύθυνση του να τείνει να ταυτισθεί με τη διεύθυνση του ιδιοδιανύσματος που αντιστοιχεί στη απόλυτα μεγαλύτερη ιδιοτιμή. Τα μόνα διανύσματα που ξεφεύγουν είναι τα κάθετα στην συγκεκριμένη ιδιοδιεύθυνση, τα οποία όμως με τη σειρά τους έλκονται από τα ιδιοδιανύσματα με τη 2^{η} μεγαλύτερη ιδιοτιμή κλπ.

Αν οι ιδιοτιμές του πίνακα είναι πραγματικές, υπάρχει μία απλή αριθμητική μέθοδος για την εύρεση της απολύτως μεγαλύτερης ιδιοτιμής

του, ενώ ταυτόχρονα βρίσκεται και ένα αντίστοιχο ιδιοδιάνυσμά της. Η μέθοδος, γνωστή ως ***μέθοδος των δυνάμεων*** υλοποιείται με τα ακόλουθα βήματα:

Βήμα 1ο: Εστω $b_0 = \begin{pmatrix} b_{01} \\ \vdots \\ b_{0n} \end{pmatrix}$ ένα τυχαίο διάνυσμα στήλη. Με τη δράση δοθέντος πίνακα A επί του διανύσματος b_0 προκύπτει ένα νέο διάνυσμα b_1:

$$b_1 = \begin{pmatrix} b_{11} \\ \vdots \\ b_{1n} \end{pmatrix} = A\, b_0 .$$

Βήμα 2ο: Διαιρούμε το διάνυσμα b_1 με την πρώτη συνιστώσα b_{11}, εφόσον $b_{11} \neq 0$. Αν $b_{11} = 0$ τότε διαιρούμε την πρώτη κατά σειρά μη μηδενική συνιστώσα μετά την $b_{11} \neq 0$. Εστω $b_{11} \neq 0$, ορίζουμε ένα νέο διάνυσμα:

$$b_1^{(1)} = \frac{1}{b_{11}} \begin{pmatrix} b_{11} \\ \vdots \\ b_{1n} \end{pmatrix} = \begin{pmatrix} 1 \\ b_{11}^{(1)} \\ \vdots \\ b_{1n}^{(1)} \end{pmatrix}.$$

Στη συνέχεια εφαρμόζουμε το βήμα 1 για το διάνυσμα $b_1^{(1)}$, οπότε παίρνουμε ένα νέο διάνυσμα $b_2 = \begin{pmatrix} b_{11}^{(1)} \\ \vdots \\ b_{1n}^{(1)} \end{pmatrix} = A\, b_1^{(1)}$ και με εφαρμογή του βήματος 2 προκύπτει ένα νέο διάνυσμα $b_2^{(2)}$:

$$b_2^{(1)} = \frac{1}{b_{21}} \begin{pmatrix} b_{21} \\ \vdots \\ b_{2n} \end{pmatrix} = \begin{pmatrix} 1 \\ b_{11}^{(1)} \\ \vdots \\ b_{1n}^{(1)} \end{pmatrix}.$$

Αν η ανωτέρω διαδικασία συνεχισθεί N φορές, τότε ο φορέας του διανύσματος $b_N^{(1)}$ τείνει να ταυτισθεί με τον φορέα του ιδιοδιανύσματος

που αντιστοιχεί στη μεγαλύτερη κατ' απόλυτη τιμή ιδιοτιμή, σύμφωνα με όσα αναφέρθηκαν παραπάνω. Θεωρώντας τώρα ότι ισχύει η ακόλουθη ισότητα:

$$b_N = A\, b_{N-1}^{(1)} = \lambda\, b_{N-1}^{(1)} \Rightarrow \begin{pmatrix} b_{1N} \\ \vdots \\ b_{nN} \end{pmatrix} = \lambda \begin{pmatrix} 1 \\ \vdots \\ b_{nN-1}^{(1)} \end{pmatrix} \Rightarrow \lambda = b_{1N},$$

προκύπτει ότι η 1^{η} συνιστώσα πριν την τελευταία κανονικοποίηση μας δίνει την αλγεβρική τιμή της απολύτως μεγαλύτερης ιδιοτιμής.

Σημείωση 1 Με τη μέθοδο της δύναμης μπορούμε επίσης να υπολογίσουμε και την κατ' απόλυτη τιμή μικρότερη ιδιοτιμή, διότι αν λ είναι ιδιοτιμή του πίνακα A, τότε η λ^{-1} είναι ιδιοτιμή του A^{-1}, οπότε εφαρμόζουμε τη μέθοδο της δύναμης για τον πίνακα A^{-1} . Επίσης υπάρχουν και μέθοδοι για την αριθμητική εύρεση και των υπολοίπων ιδιοτιμών, που βασίζονται στην εύρεση ενός πίνακα που περιέχει τις εναπομείνασες ιδιοτιμές και εφαρμογή της μεθόδου της δύναμης για αυτόν τον πίνακα κλπ.

Παράδειγμα 2 Υπολογίστε με τη μέθοδο των δυνάμεων τη μέγιστη κατ΄ απόλυτο τιμή ιδιοτιμή του πίνακα

$$A = \begin{pmatrix} 2 & 1 \\ 3 & 2 \end{pmatrix},$$

με ακρίβεια 2 δεκαδικών ψηφίων.

Λύση Εστω $b_0 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ ένα τυχαίο διάνυσμα στήλη, υπολογίζουμε:

1η επανάληψη:

$$b_1 = A\, b_0 = \begin{pmatrix} 2 & 1 \\ 3 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 4 \\ 7 \end{pmatrix}$$

$$b_1^{(1)} = \frac{1}{b_{11}} b_1 = \frac{1}{4} \begin{pmatrix} \boxed{4} \\ 7 \end{pmatrix} = \begin{pmatrix} 1 \\ 7/4 \end{pmatrix}.$$

2η επανάληψη:

$$b_2 = A\, b_1^{(1)} = \begin{pmatrix} 2 & 1 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 7/4 \end{pmatrix} = \begin{pmatrix} 3.75 \\ 6.5 \end{pmatrix}$$

$$b_2^{(1)} = \frac{1}{b_{21}} b_2 = \frac{1}{3.75} \begin{pmatrix} \boxed{3.75} \\ 6.5 \end{pmatrix} = \begin{pmatrix} 1 \\ 1.73333 \end{pmatrix}.$$

3η επανάληψη:

$$b_3 = A\, b_2^{(1)} = \begin{pmatrix} 2 & 1 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1.73333 \end{pmatrix} = \begin{pmatrix} 3.73333 \\ 6.46666 \end{pmatrix}$$

$$b_3^{(1)} = \frac{1}{b_{31}} b_3 = \frac{1}{3.73333} \begin{pmatrix} \boxed{3.73333} \\ 6.46666 \end{pmatrix} = \begin{pmatrix} 1 \\ 1.73214 \end{pmatrix}.$$

4η επανάληψη:

$$b_4 = A\, b_3^{(1)} = \begin{pmatrix} 2 & 1 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1.73214 \end{pmatrix} = \begin{pmatrix} 3.73214 \\ 6.46428 \end{pmatrix}$$

$$b_4^{(1)} = \frac{1}{b_{41}} b_4 = \frac{1}{3.73214} \begin{pmatrix} \boxed{3.73214} \\ 6.46428 \end{pmatrix} = \begin{pmatrix} 1 \\ 1.73205 \end{pmatrix}.$$

Αφού $\lambda_1 = 4$, $\lambda_2 = 3.75$, $\lambda_3 = 3.73333$, $\lambda_4 = 3.73214$ και $|\lambda_4 - \lambda_3| = 0.00119$ έχουμε ότι $\lambda \cong 3.73333$. Το ιδιοδιάνυσμα που αντιστοιχεί είναι το $\begin{pmatrix} 1 \\ 1.73205 \end{pmatrix}$. □

ΑΛΥΤΕΣ ΑΣΚΗΣΕΙΣ

1. Να υπολογισθούν οι ιδιοτιμές και τα ιδιοδιανύσματα των πινάκων:

$$A = \begin{pmatrix} 25 & 0 & 0 \\ 0 & 34 & -12 \\ 0 & -12 & 41 \end{pmatrix},\ B = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix},\ C = \begin{pmatrix} 2 & 1 \\ -3 & 6 \end{pmatrix}.$$

Απάντ.: Ιδιοτιμές του A: $\lambda = 50, \lambda = 25$ (διπλή).
Ιδιοτιμές του B: $\lambda = 0$ (διπλή).
Ιδιοτιμές του C: $\lambda = 5$ ή $\lambda = 3$.

2. Υπολογίστε με τη μέθοδο της δύναμης την μεγαλύτερη κατ' απόλυτο τιμή ιδιοτιμή και το αντίστοιχο ιδιοδιάνυσμα των πινάκων κάνοντας *N=5* επαναλήψεις. Συγκρίνετε τα αποτελέσματά σας με την πραγματική τιμή της απόλυτα μεγαλύτερης ιδιοτιμής.

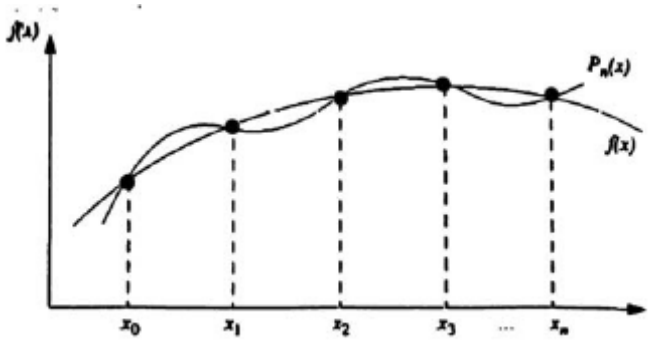
$$A=\begin{pmatrix}1 & 1 & 1\\ 1 & 3 & 3\\ 1 & 3 & 6\end{pmatrix},\ B=\begin{pmatrix}1 & 2\\ 2 & 1\end{pmatrix}.$$

ΚΕΦΑΛΑΙΟ 5

ΠΑΡΕΜΒΟΛΗ

§ 5.1 *Πολυωνυμική παρεμβολή*

Εστω f πραγματική συνάρτηση, της οποίας είναι γνωστές μόνον οι τιμές $f(x_i)$ σε $n+1$ σημεία x_i, $i=0,...,n$ του πεδίου ορισμού της. Το πρόβλημα εύρεσης μιας συνάρτησης φ, (από ένα ορισμένο σύνολο συναρτήσεων Σ), έτσι ώστε η φ να προσδιορίζεται μόνον από τις τιμές $f(x_i)$ και να πληροί τις συνθήκες $\varphi(x_i)=f(x_i)$, $i=0,...,n$ καλείται ***παρεμβολή***. Αν το σύνολο Σ είναι αρκετά «πλούσιο», τότε η τιμή της συνάρτησης $\varphi(x)$ για $x \neq x_i$ μπορεί να θεωρηθεί ότι προσεγγίζει την τιμή $f(x)$.



Σχήμα 5: Πολυωνυμική παρεμβολή

Στο Κεφάλαιο αυτό θα προσεγγίσουμε συναρτήσεις με παρεμβολή με πολυώνυμα, ή με τμηματικά πολυωνυμικές συναρτήσεις. Ο λόγος είναι ότι κατασκευάζονται πολύ εύκολα με πολ/σμούς και προσθαφαιρέσεις, παραγωγίζονται και ολοκληρώνονται πολύ εύκολα και έχουν καλές προσεγγιστικές ιδιότητες. Πράγματι:

Θεώρημα 5.1.1 (Weierstrass) Εστω $f \in C[a,b]$, όπου $C[a,b]$ είναι το σύνολο των συνεχών συναρτήσεων στο κλειστό διάστημα *[α,b]*, τότε για κάθε $\varepsilon > 0$ υπάρχει πολυώνυμο $\pi(x)$ τέτοιο ώστε:

$$\max_{x\in[a,b]} \left| f(x) - \pi(x) \right| < \varepsilon .$$

Θεώρημα 5.1.2 (*Υπαρξης και μοναδικότητας του πολυωνύμου παρεμβολής)* Εστω $n+1$ σημεία του επιπέδου (x_i, y_i), $i=0,...,n$, τότε υπάρχει μοναδικό πολυώνυμο $p(x)$ βαθμού το πολύ n, τέτοιο ώστε:

$$p(x_i) = y_i,\ i = 0,...,n. \qquad (5.1)$$

Απόδειξη: Εστω $p(x) = a_0 + a_1 x + ... + a_n x^n$ πολυώνυμο βαθμού το πολύ n με αγνώστους τους συντελεστές $a_0, a_1, ..., a_n$, τότε από την (5.1) προκύπτει ένα γραμμικό σύστημα $n+1$ εξισώσεων με $n+1$ αγνώστους. Το αντίστοιχο ομογενές σύστημα

$$p(x_i) = 0,\ i = 0,...,n,$$

έχει προφανώς ως μοναδική λύση την τετριμμένη μηδενική λύση, διότι το p ως πολυώνυμο το πολύ n βαθμού έχει το πολύ n ρίζες και όχι $n+1$ που υπονοεί το παραπάνω ομογενές σύστημα. Εφόσον το ομογενές σύστημα έχει μόνον την τετριμμένη λύση, το γραμμικό σύστημα (5.1) έχει μοναδική λύση. □

Εστω f είναι μια πραγματική συνάρτηση και p_n ένα πολυώνυμο βαθμού το πολύ n, τέτοιο ώστε $p_n(x_i) = f(x_i),\ i = 0,...,n$, τότε το μοναδικό πολυώνυμο p_n καλείται πολυώνυμο παρεμβολής της f στα σημεία $x_0, ..., x_n$. Ισχύει δε:

Θεώρημα 5.1.3 (*Σφάλμα προσέγγισης)* Εστω $n = 1,...,$ και $f \in C^{n+1}[a,b]$, όπου $C^{n+1}[a,b]$ είναι το σύνολο των συναρτήσεων που είναι $n+1$ φορές συνεχώς παραγωγίσιμες στο κλειστό διάστημα *[α,b]*. Αν p_n είναι το πολυώνυμο παρεμβολής της f στα σημεία $x_0, ..., x_n \in [a,b]$, τότε:

$$\| f - p_n \|_\infty \ \le\ \frac{\| f^{(n+1)} \|_\infty}{(n+1)!}\ \max_{x\in[a,b]} \prod_{i=0}^{n} |x - x_i|,$$

όπου: $\| f \|_\infty = \max_{x\in[a,b]} |f(x)|$.

Απόδειξη: Εστω $f \in C^{n+1}[a,b]$, $x_0, ..., x_n \in [a,b]$ και $x \neq x_i$, θέτουμε $\Phi(t) = \prod_{i=0}^{n} (t - x_i)$ και ορίζουμε μία βοηθητική συνάρτηση $\varphi_n(t)$ ως εξής:

$$\varphi_n(t) = f(t) - p_n(t) - \frac{f(x) - p_n(x)}{\Phi(x)}\ \Phi(t),\ t \in [a,b].$$

Είναι εύκολο να δει κανείς ότι $\varphi_n(t) \in C^{n+1}[a,b]$ και

$$\varphi_n(x_i)=0,\ i=0,1,...,n,\ \text{και}\ \varphi_n(x)=0,$$

άρα η φ_n έχει στο διάστημα *[α,b]* τουλάχιστον *n+2* διαφορετικές ρίζες. Με χρήση του Θεωρήματος Rolle, προκύπτει ότι η φ_n' έχει στο ανοικτό διάστημα *(α,b)* τουλάχιστον *n+1* διαφορετικές ρίζες, η φ_n'' έχει στο ανοικτό διάστημα *(α,b)* τουλάχιστον *n* διαφορετικές ρίζες κλπ και τέλος η $\varphi_n^{(n+1)}$ έχει στο ανοικτό διάστημα *(α,b)* τουλάχιστον 1 ρίζα *ξ*. Επειδή

$$\varphi_n^{(n+1)}(t)=f^{(n+1)}(t)-\frac{f(x)-p_n(x)}{\Phi(x)}\ (n+1)!,$$

έχουμε:

$$0=\varphi_n^{(n+1)}(\xi)=f^{(n+1)}(\xi)-\frac{f(x)-p_n(x)}{\Phi(x)}\ (n+1)!$$

$$\Rightarrow f(x)-p_n(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}\Phi(x).\quad \square$$

§ 5.2 *Κατασκευή πολυωνύμου παρεμβολής*

(α) Πολυώνυμο Lagrange

Εστω *n+1* σημεία του επιπέδου (x_i,y_i), $i = 0,...,n$, τότε το πολυώνυμο Lagrange που διέρχεται από τα σημεία (x_i,y_i) έχει τη μορφή:

$$p_n(x)=\sum_{i=0}^{n} y_i\ L_i(x), \tag{5.2}$$

όπου τα πολυώνυμα:

$$L_i(x)=\frac{(x-x_0)...(x-x_{i-1})(x-x_{i+1})...(x-x_n)}{(x_i-x_0)...(x_i-x_{i-1})(x_i-x_{i+1})...(x_i-x_n)},\ i=0,...,n$$

καλούνται ***συντελεστές Lagrange***.

Παράδειγμα 5.1 Υπολογίστε το πολυώνυμο παρεμβολής του Lagrange που διέρχεται από τα σημεία: *(-1,-3), (0,-2) (1,-1).*

Λύση Αριθμούμε τα σημεία μας ξεκινώντας πάντοτε από την τιμή $i = 0$ και έχουμε:

	x_0	x_1	x_2
x	-1	0	1
y	-3	-2	-1
	y_0	y_1	y_2

Από τον τύπο (5.2) έχουμε:

$$p_2(x) = \sum_{i=0}^{2} y_i \; L_i(x) = -3L_0(x) - 2L_1(x) - L_2(x) .$$

Υπολογίζουμε:

$$L_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{(x-0)(x-1)}{(-1-0)(-1-1)} = \frac{x^2-x}{2},$$

$$L_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{(x+1)(x-1)}{(0+1)(0-1)} = \frac{x^2-1}{-1},$$

$$L_2(x) = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{(x+1)(x-0)}{(1+1)(1-0)} = \frac{x^2+x}{2}.$$

Αντικαθιστούμε στον τύπο του πολυωνύμου $p_2(x)$ και υπολογίζουμε:

$$p_2(x) = -3\frac{x^2-x}{2} - 2\frac{x^2-1}{-1} - \frac{x^2+x}{2} = x - 2 . \quad \square$$

(β)-Πολυώνυμο Newton

Εστω $n+1$ σημεία του επιπέδου (x_i,y_i), $i = 0,\ldots,n$, τότε το πολυώνυμο του Newton που διέρχεται από τα σημεία (x_i,y_i) έχει τη μορφή:

$$p_n(x) = a_0 + a_1(x-x_0) + a_2(x-x_0)(x-x_1) + \ldots + a_n(x-x_0)\ldots(x-x_{n-1}) , \qquad (5.3)$$

όπου οι συντελεστές $a_0, a_1, \ldots a_n$ υπολογίζονται με *τη μέθοδο των διαιρεμένων ή προσαρτημένων διαφορών* ως εξής:

Κατασκευάζουμε τον ακόλουθο πίνακα:

x	y	Διαιρεμένες διαφορές $1^{ης}$ τάξης	Διαιρεμένες διαφορές $2^{ης}$ τάξης	…	Διαιρεμένες διαφορές *n* τάξης
x_0	y_0	$\Delta_{11} = (y_1-y_0) / (x_1-x_0)$	$\Delta_{21} = (\Delta_{12}-\Delta_{11}) / (x_2-x_0)$	…	$\Delta_{n1} = (\Delta_{n-1,2}-\Delta_{n-1,1}) / (x_n-x_0)$
x_1	y_1	$\Delta_{12} = (y_2-y_1) / (x_2-x_1)$	$\Delta_{22} = (\Delta_{13}-\Delta_{12}) / (x_3-x_1)$	…	
⋮	⋮	⋮	⋮		
x_{n-1}	y_{n-1}	$\Delta_{1,n} = (y_n-y_{n-1}) / (x_n-x_{n-1})$	$\Delta_{2,n-1} = (\Delta_{1,n} - \Delta_{1,n-1}) / (x_n-x_{n-2})$		
x_n	y_n				

ακολουθώντας τους εξής κανόνες:

(1) οι δύο πρώτες στήλες είναι οι στήλες των δεδομένων *x* και *y* όπου τα *x* διατάσσονται από το μικρότερο στο μεγαλύτερο, δηλαδή:

$$x_0 < x_1 < \ldots < x_n .$$

(2) Ορίζουμε τις ***διαιρεμένες διαφορές $1^{ης}$ τάξης*** από τη σχέση:

$$\Delta_{1i} = \frac{y_i - y_{i-1}}{x_i - x_{i-1}},\ i = 1,\ldots n .$$

(3) Ορίζουμε τις ***διαιρεμένες διαφορές i-τάξης i = 1,…,n*** από τη σχέση:

$$\Delta_{ij} = \frac{\Delta_{i-1,j+1} - \Delta_{i-1,j}}{x_{j+i-1} - x_{j-1}},\ j = 1,\ldots n-i+1 .$$

(4) Οι συντελεστές $\alpha_i,\ i = 0,\ldots,n$ του πολυωνύμου του Newton υπολογίζονται ως εξής:

$$a_i = \begin{cases} y_0, & i = 0 \\ \Delta_{i1}, & i = 1,\ldots,n \end{cases} .$$

Παράδειγμα 5.2 Να υπολογισθεί το πολυώνυμο του Newton που παρεμβάλλει μία συνάρτηση *f(x)* στα σημεία:

x	*-1*	*-2*	*0*	*3*	*2*
y	*5*	*3*	*-2*	*0*	*4*

Λύση Κατασκευάζουμε τον ακόλουθο πίνακα διαιρεμένων διαφορών:

x	y	$1^{ης}$ τάξης δ.δ.	$2^{ης}$ τάξης δ.δ	$3^{ης}$ τάξης δ.δ	$4^{ης}$ τάξης δ.δ
-2	**3**	(5-3)/(-1-(-2))=**2**	(-7-2)/(0-(-2))=**-9/2**	(10/3+9/2)/(2-(-2))=**47/24**	(-17/12-47/24)/5=**-81/120**
-1	5	(-2-5)/(0-(-1))=**-7**	(3-(-7))/(2-(-1))=**10/3**	(-7/3-10/3)/(3-(-1))=**-17/12**	

0	-2	(4-(-2))/(2-0)=**3**	(-4-3)/(3-0)=**-7/3**		
2	4	(0-4)/(3-2)=**-4**			
3	0				

Αρα το πολυώνυμο του Newton είναι το εξής:

$$p_4(x) = 3 + 2(x+2) - \frac{9}{2}(x+2)(x+1) + \frac{47}{24}(x+2)(x+1)x - \frac{81}{120}(x+2)(x+1)x(x-2)$$

και προκύπτει από τον τύπο (5.3) *για* $x_0 = -2$, $x_1 = -1$, $x_2 = 0$, $x_3 = 2$, με τους συντελεστές $\alpha_0,\alpha_1,\ldots\ \alpha_4$ να προκύπτουν από τη σχέση (5.3), σε συνδυασμό με τον πίνακα διαιρεμένων διαφορών (βλέπε αριθμούς με κόκκινη απόχρωση) που κατασκευάσαμε:

$$\alpha_0=3,\ \alpha_1=2,\ \alpha_2=-9/2,\ \alpha_3=47/24,\ \alpha_4=-81/120. \quad \square$$

Σημείωση Για πολλές συναρτήσεις *f*, το μέγιστο σφάλμα $\| f - p_n \|_\infty$ κατά την προσέγγιση της *f* με ένα πολυώνυμο παρεμβολής p_n τείνει στο μηδέν. Αυτό όμως δε συμβαίνει πάντα. Από τον Runge δόθηκε το παράδειγμα της συνάρτησης

$$f(x) = \frac{1}{1+25x^2},$$

η οποία είναι απειροδιαφορίσιμη στο κλειστό διάστημα [-1,1] και για την οποία ισχύει $\| f - p_n \|_\infty \to \infty$, $n \to \infty$. Γενικότερα, ο Faber απέδειξε το 1914 ότι για *οποιαδήποτε επιλογή των σημείων παρεμβολής, υπάρχει συνεχής συνάρτηση για την οποία η παρεμβολή αποτυγχάνει*.

§ 5.3 *Παρεμβολή Hermitte*

Εστω *f* μία πραγματική συνάρτηση και ας υποθέσουμε ότι ζητούμε από την παρεμβάλλουσα συνάρτηση *φ(x)* της *f,* να έχει εκτός της ταύτισης με την *f* στα σημεία παρεμβολής $x_0,\ldots,x_n$ και τις ίδιες παραγώγους μέχρι κάποια τάξη με την *f*, όπου η τάξη μπορεί να διαφέρει από σημείο σε σημείο. Τότε μιλούμε για ***παρεμβολή τύπου Hermitte***.

Θεώρημα 5.3.1 (*Υπαρξης και μοναδικότητας του πολυωνύμου παρεμβολής)* Εστω $m_0,\ldots,m_n$ φυσικοί αριθμοί, $N = n + m_0 + \ldots + m_n$, p_N πολυώνυμο βαθμού *N* και $M = max\{m_0,\ldots,m_n\}$. Αν $f \in C^M[a,b]$ και $(x_i, f(x_i))$, $i = 0,\ldots,n$, $n+1$ σημεία του επιπέδου, τότε υπάρχει μοναδικό πολυώνυμο $p_N(x)$ βαθμού το πολύ *N*, τέτοιο ώστε:

$$
\begin{array}{c}
p_N^{(i)}(x_0) = f^{(i)}(x_0),\ i = 0,...,m_0 \\
p_N^{(i)}(x_1) = f^{(i)}(x_1),\ i = 0,...,m_1, \\
\cdots \\
p_N^{(i)}(x_n) = f^{(i)}(x_n),\ i = 0,...,m_n
\end{array}
. \qquad (5.4)
$$

Η πιο συνηθισμένη περίπτωση είναι αυτή κατά την οποία ζητούμε να υπολογίσουμε το πολυώνυμο Hermitte που ικανοποιεί τα δεδομένα:

$$
p_N(x_i) = f(x_i),\ p'_N(x_i) = f'(x_i),\ i = 0,...,n. \qquad (5.5)
$$

Θεώρημα 5.3.2 (*Σφάλμα προσέγγισης)* Εστω *n = 1,...,* $f \in C^{2n+2}[a,b]$, και p_{2n+1} είναι το πολυώνυμο παρεμβολής Hermitte που ικανοποιεί την (5.5) στα σημεία $x_0,...,x_n \in [a,b]$, τότε:

$$
\| f - p_n \|_\infty \ \leq \ \frac{\| f^{(2n+2)} \|_\infty}{(2n+2)!} \ \max_{x\in[a,b]} \prod_{i=0}^{n} |x - x_i|^2 .
$$

Κατασκευή του πολυωνύμου Hermitte

Θεωρούμε τη σχέση (5.4), όπου χωρίς περιορισμό της γενικότητας έχουμε υποθέσει ότι $x_0 < ... < x_n$. Ορίζουμε μία νέα ακολουθία σημείων ως εξής:

$$
\left\{ \underbrace{x_0, x_0, ..., x_0}_{m_0+1\ \phi o \rho \acute{\epsilon} \varsigma}, \underbrace{x_1, x_1, ..., x_1}_{m_1+1\ \phi o \rho \acute{\epsilon} \varsigma}, ..., \underbrace{x_n, x_n, ..., x_n}_{m_n+1\ \phi o \rho \acute{\epsilon} \varsigma} \right\}, \qquad (5.6)
$$

και για την ακολουθία (5.6) κατασκευάζουμε το πολυώνυμο Hermitte να είναι το πολυώνυμο Newton (5.3), όπου οι συντελεστές του υπολογίζονται όπως είδαμε στις σελ. 78-79 με μία μόνον επιπλέον συνθήκη:

(Σ) *όποτε βρίσκουμε στις διαιρεμένες διαφορές i-τάξης την ποσότητα 0/0 (μηδέν διά μηδέν) θα την αντικαθιστούμε με την ποσότητα:*

$$
\Delta_{i,k} \to \frac{f^{(i)}(x_k)}{i!}.
$$

Παράδειγμα 5.3 Να υπολογισθεί το πολυώνυμο του Hermitte που ικανοποιεί τα δεδομένα:

$$f(0)=2, f'(0)=4, f(2)=4, f'(2)=4, f''(2)=4, f(4)=6.$$

Λύση Κατασκευάζουμε την ακολουθία (5.6) ως εξής: προφανώς έχουμε 3 σημεία τα *f(0) = 2, f(2) = 4, f(4) = 6*, τις τετμημένες των οποίων διατάσσουμε κατ' αύξουσα τάξη λόγω του κανόνα (1) της σελ. 78. Ετσι έχουμε $x_0=0, x_1=2, x_2=4$. Στη συνέχεια, κάθε σημείο επαναλαμβάνεται τόσες φορές όσο είναι η μέγιστη τάξη της παραγώγου του, άρα το $x_0=0$ θα επαναληφθεί μία φορά, το $x_1=2$ θα επαναληφθεί 2 φορές, ενώ το $x_0=4$ δεν θα επαναληφθεί. Ετσι κατασκευάζουμε την ακολουθία (5.6):

$$\{0,0,2,2,2,4\}.$$

Το πολυώνυμο του Hermitte κατασκευάζεται όπως το πολυώνυμο Newton με την επιπλέον συνθήκη (Σ) (βλέπε σελ. 80). Κατασκευάζουμε πρώτα τον πίνακα διαιρεμένων διαφορών:

x	y	1ης τάξης δ. δ.	2ης τάξης δ. δ.	3ης τάξης δ. δ.	4ης τάξης δ. δ.
0	**2**	$0/0 \to f'(0)$ = 4	(2-4)/(2-0)=**-1**	(1-(-1))/(2-0) = **1**	(3/2-1)/(2-0)=**1/4**
0	2	(4-0)/(2-0) = 2	(4-2)/(2-0)=1	(4-1)/(2-0) = 3/2	(5/4-3/2)/(4-0)=-1/16
2	4	$0/0 \to f'(2)$ = 4	$0/0 \to f''(2)/2$ = 4	(1-(-3/2))/(4-2)=5/4	
2	4	$0/0 \to f'(2)$ = 4	(1-4)/(4-2) = -3/2		
2	4	(6-4)/(4-2) = 1			
4	6				

5ης τάξης δ. δ.
(-1/16-1/4)/(4-0)=**-5/64**

Αρα το πολυώνυμο του Hermitte είναι το εξής:

$$p_5(x)=2+4(x-0)-1(x-0)(x-0)+1(x-0)(x-0)(x-2)$$

$$+\frac{1}{4}(x-0)(x-0)(x-2)(x-2)-\frac{5}{64}(x-0)(x-0)(x-2)(x-2)(x-2)$$

και προκύπτει από τον τύπο (5.3) για $x_0=0, x_1=0, x_2=2, x_3=2\ x_4=2$ με τους συντελεστές $\alpha_0,\alpha_1,\ldots,\alpha_4,\alpha_5$ να προκύπτουν από τη σχέση (5.3) σε συνδυασμό με τον πίνακα διαιρεμένων διαφορών (βλέπε αριθμούς με κόκκινη απόχρωση) που κατασκευάσαμε:

$$\alpha_0=2, \alpha_1=4, \alpha_2=-1, \alpha_3=1, \alpha_4=1/4, \alpha_5=-5/64. \quad \square$$

§ 5.4 *Splines*

Εστω $a = x_0 < x_1 < ... < x_n = b$ είναι ένας διαμερισμός του κλειστού διαστήματος *[α,b]*, τότε splines ως προς αυτό το διαμερισμό καλούνται γενικά εκείνες οι συναρτήσεις που σε κάθε υποδιάστημα $[x_i,x_{i+1}]$, έχουν μία ορισμένη μορφή, είναι π.χ. πολυώνυμα βαθμού *m*.

Ορισμός 5.4.1 Εστω $a = x_0 < x_1 < ... < x_n = b$ είναι ένας διαμερισμός του κλειστού διαστήματος *[α,b]*. Κάθε συνάρτηση $s \in C^{m-1}[a,b]$ τέτοια ώστε ο περιορισμός $s|_{[x_i,x_{i+1}]}$ $i = 1,...,n$ να είναι πολυώνυμο βαθμού *m* καλείται πολυωνυμική *spline* βαθμού *m.*

Για παράδειγμα οι πολυωνυμικές *splines* βαθμού *1* είναι οι συνεχείς τεθλασμένες γραμμές.

Σημαντικό ρόλο στις εφαρμογές παίζουν οι κυβικές *splines*, δηλαδή οι συναρτήσεις που είναι *2 φορές παραγωγίσιμες στο κλειστό διάστημα [α,b] και είναι κυβικά πολυώνυμα σε κάθε υποδιάστημα ενός οποιουδήποτε διαμερισμού του [α,b]*.

Εστω $a = x_0 < x_1 < ... < x_n = b$ οποιοδήποτε διαμερισμός του *[α,b]*, για τον προσδιορισμό μιας κυβικής *spline* απαιτείται η εύρεση συνολικά *4n* σταθερών, δηλαδή των συντελεστών του αντιστοίχου κυβικού πολυωνύμου

$$s^{(i)}(x) = a_0^{(i)} + a_1^{(i)}x + a_2^{(i)}x^2 + a_3^{(i)}x^3$$

σε κάθε υποδιάστημα $[x_i,x_{i+1}]$, $i=0,...,n-1$. Εχουμε λοιπόν $n+1$ σχέσεις από τις συνθήκες παρεμβολής

$$s^{(i)}(x_i) = f(x_i),\ i = 0,...,n,$$

n-1 σχέσεις από τις συνθήκες συνέχειας στους εσωτερικούς κόμβους

$$s^{(i-1)}(x_i) = s^{(i)}(x_i),\ i = 1,...,n,$$

και *2 (n-1)* σχέσεις από τις συνθήκες παραγωγισιμότητας στους εσωτερικούς κόμβους

$$\left(s^{(i-1)}\right)'(x_i) = \left(s^{(i)}\right)'(x_i),\ i = 1,...,n,$$

$$\left(s^{(i-1)}\right)''(x_i) = \left(s^{(i)}\right)''(x_i),\ i = 1,...,n,$$

συνολικά δηλαδή μπορούμε να προσδιορίσουμε *2 (n-1)* εξισώσεις. Οι υπόλοιπες δύο είναι συνήθως διαφόρων τύπων *συνοριακές εξισώσεις* που αφορούν τους συνοριακούς κόμβους *α* και *b*. Για παράδειγμα αν ισχύει

$$s''(x_0) = s''(x_n) = 0,$$

τότε λέμε ότι έχουμε *φυσικές κυβικές splines*. Ισχύει μάλιστα:

Θεώρημα 5.4.2 Εστω $f \in C^4[a,b]$ και έστω s η κυβική πολυωνυμική spline που παρεμβάλλει την f στα σημεία $a = x_0 < x_1 < ... < x_n = b$, τότε υπάρχουν σταθερές C_m, $m = 0,1,2,3$ τέτοιες ώστε:

$$\| f^{(m)} - s^{(m)} \|_\infty \ \le\ C_m\ h^{4-m}\ \| f^{(4)} \|_\infty ,$$

όπου *h* είναι η μέγιστη τιμή του εύρους μεταξύ των υποδιαστημάτων $[x_i, x_{i+1}]$, $i = 0,...,n-1$.

Για να υπολογίσουμε τις κυβικές *splines* εργαζόμαστε ως εξής:

1. εφόσον η *s* είναι κυβική spline, η *s΄΄* είναι συνεχής συνάρτηση και μάλιστα θα είναι μία τεθλασμένη γραμμή, οπότε κάθε ευθύγραμμο τμήμα αυτής της γραμμής είναι η $\left(s^{(j)}\right)''(x),\ j = 0,...,n-1$, η οποία με χρήση του πολυωνύμου Lagrange που διέρχεται από τα σημεία μπορεί να γραφεί ως εξής:

$$\left(s^{(j)}\right)''(x) = a_j \frac{x - x_{j+1}}{x_j - x_{j+1}} + a_{j+1} \frac{x - x_j}{x_{j+1} - x_j},$$

όπου α_j, α_{j+1} άγνωστοι τους οποίους θέλουμε να προσδιορίσουμε. Με αυτή τη γραφή ισχύει ότι

$$\left(s^{(j)}\right)''\left(x_j\right) = a_j,\ \left(s^{(j)}\right)''\left(x_{j+1}\right) = a_{j+1},\ \left(s^{(j-1)}\right)''\left(x_j\right) = a_j = \left(s^{(j)}\right)''\left(x_j\right).$$

2. Ολοκληρώνοντας 2 φορές και χρησιμοποιώντας τις συνθήκες παρεμβολής προσδιορίζουμε τις άγνωστες σταθερές.

Σημειώνουμε ότι στα περισσότερα μαθηματικά λογισμικά υπάρχουν ειδικές εντολές που υπολογίζουν άμεσα τις κυβικές splines.

Παράδειγμα 4 Εστω $\{x_0 = -1,\ x_1 = 0,\ x_2 = 1\}$ ένας ομοιόμορφος διαμερισμός του διαστήματος $[-1,1]$. Προσδιορίστε τη φυσική κυβική spline που παρεμβάλλεται σε μία συνάρτηση f στα σημεία x_i , $i = 0,1,2$, έτσι ώστε $f(x_0) = 0,\ f(x_1) = 2,\ f(x_2) = 6$.

Λύση Εστω $s(x)$ η φυσική κυβική spline και $s^{(0)}(x), s^{(1)}(x)$ τα κυβικά πολυώνυμα στα υποδιαστήματα $[-1,0]$ και $[0,1]$ αντίστοιχα. Προφανώς κάθε συνάρτηση $\left(s^{(j)}\right)''(x)$ είναι ένα πολυώνυμο $1^{ου}$ βαθμού, δηλαδή ένα ευθύγραμμο τμήμα, οπότε αν $\alpha_{j,k}$, $j,k = 0,1$ είναι άγνωστες σταθερές, έχουμε:

$$\begin{cases} \left(s^{(0)}\right)''(x) = a_{0,0}x + a_{0,1} \\ \left(s^{(1)}\right)''(x) = a_{1,0}x + a_{1,1} \end{cases}.$$

Προφανώς $\left(s^{(0)}\right)''(-1) = 0,\ \left(s^{(1)}\right)''(1) = 0,\ \left(s^{(0)}\right)''(0) = \left(s^{(1)}\right)''(0)$, οπότε βρίσκουμε ότι:

$$a_{0,0} = a_{0,1},\quad a_{1,1} = -a_{1,0},\quad a_{0,1} = a_{1,1},$$

δηλαδή:

$$\begin{cases} \left(s^{(0)}\right)''(x) = a_{0,0}x + a_{0,0} \\ \left(s^{(1)}\right)''(x) = -a_{0,0}x + a_{0,0} \end{cases}$$

άρα η συνάρτηση $s''(x)$ είναι συνεχής (όπως απαιτείται). Ολοκληρώνοντας παίρνουμε:

$$\begin{cases} \left(s^{(0)}\right)'(x) = a_{0,0}\dfrac{x^2}{2} + a_{0,0}x + c_1 \\ \left(s^{(1)}\right)'(x) = -a_{0,0}\dfrac{x^2}{2} + a_{0,0}x + c_2 \end{cases}.$$

Επειδή πρέπει $\left(s^{(0)}\right)'(0)=\left(s^{(1)}\right)'(0)$ παίρνουμε εύκολα ότι $c_1=c_2$ οπότε:

$$\begin{cases}\left(s^{(0)}\right)'(x)=a_{0,0}\dfrac{x^2}{2}+a_{0,0}x+c_1\\ \left(s^{(1)}\right)'(x)=-a_{0,0}\dfrac{x^2}{2}+a_{0,0}x+c_1\end{cases}.$$

άρα η συνάρτηση $s'(x)$ είναι συνεχής (όπως απαιτείται). Ολοκληρώνοντας παίρνουμε:

$$\begin{cases}s^{(0)}(x)=a_{0,0}\dfrac{x^3}{6}+a_{0,0}\dfrac{x^2}{2}+c_1x+d_1\\ s^{(1)}(x)=-a_{0,0}\dfrac{x^3}{6}+a_{0,0}\dfrac{x^2}{2}+c_1x+d_2\end{cases}. \qquad (5.11)$$

Επειδή πρέπει $s^{(0)}(0)=s^{(1)}(0)=2$ παίρνουμε εύκολα ότι $d_1=d_2=2$, οπότε:

$$\begin{cases}s^{(0)}(x)=a_{0,0}\dfrac{x^3}{6}+a_{0,0}\dfrac{x^2}{2}+c_1x+2\\ s^{(1)}(x)=-a_{0,0}\dfrac{x^3}{6}+a_{0,0}\dfrac{x^2}{2}+c_1x+2\end{cases},$$

άρα η συνάρτηση $s(x)$ είναι συνεχής (όπως απαιτείται)..

Τέλος επειδή $s^{(0)}(-1)=0,\ \ s^{(1)}(1)=6$, λύνουμε το σύστημα:

$$\begin{cases}s^{(0)}(-1)=a_{0,0}\dfrac{(-1)^3}{6}+a_{0,0}\dfrac{(-1)^2}{2}+c_1(-1)+2\\ s^{(1)}(1)=-a_{0,0}\dfrac{1^3}{6}+a_{0,0}\dfrac{1^2}{2}+c_1 1+2\end{cases}$$

ως προς $a_{0,0}, c_1$. Εχουμε:

$$\begin{cases} 0 = -a_{0,0}\dfrac{1}{6} + a_{0,0}\dfrac{1}{2} - c_1 + 2 \\ 6 = -a_{0,0}\dfrac{1}{6} + a_{0,0}\dfrac{1}{2} + c_1 + 2 \end{cases},$$

απ' όπου προκύπτει εύκολα ότι:

$$a_{0,0} = 3, \quad c_1 = 3,$$

και τελικά από την (5.11) παίρνουμε:

$$s(x) = \begin{cases} s^{(0)}(x) = \dfrac{1}{2}x^3 + \dfrac{3}{2}x^2 + 3x + 2, & x \in [-1,0) \\ s^{(1)}(x) = -\dfrac{1}{2}x^3 + \dfrac{3}{2}x^2 + 3x + 2, & x \in [0,1] \end{cases}. \qquad \square$$

ΑΛΥΤΕΣ ΑΣΚΗΣΕΙΣ

1. Να προσδιορισθεί το πολυώνυμο Lagrange που διέρχεται από τα δεδομένα: *(i) (-2,-9), (-1,-2), (0,-1), (1,0)*
(ii) (-1,20), (0, 10), (2, -4)
(iii) (0,-3), (5,7).

Απάντ. *(i) $y = x^3$-1, (ii) $y = x^2$-9x+10 (iii) y = 2x-3.*

2. Να προσδιορισθεί το πολυώνυμο Newton που διέρχεται από τα δεδομένα: *(i) (-1,-3), (0,1), (2,15), (1,3)*
(ii) (-1,19), (1, 5), (0, 9).

Απάντ. *(i) $y = 2x^3$- x^2+x+1, (ii) $y = x^2$-9x+10.*

3. Να προσδιορισθεί το πολυώνυμο Hermitte που ικανοποιεί τα δεδομένα: *(i) f(0)=0, f ′(0)=2, f ′ ′(0)=4, f(1)=6, f(2)=8, f(4)=16.*
(ii) f(1)=3, f ′(1)=4, f(2)=6, f ′(2)=4, f ′ ′(2)=2, f ′ ′ ′(2)=6.

Απάντ. *(i) $y = 41/48\ x^5 - 81/16\ x^4 + 149/24\ x^3 + 2x^2$+2x,*
(ii) $y = 3x^5$-26 x^4 + 89 x^3-149x^2+124 x-38.

4. Εστω *{x_0 = -2, x_1 = -1, x_2 = 0, x_3 = 1, x_4 = 2}* ένας ομοιόμορφος διαμερισμός του διαστήματος *[-2,2]*. Προσδιορίστε τη φυσική κυβική

spline που παρεμβάλλεται σε μία συνάρτηση f στα σημεία x_i , $i = 0,1,2,$ έτσι ώστε $f(x_0) = 2, f(x_1) = 4, f(x_2) = 0, f(x_3) = -2, f(x_4) = -6.$

5 Εστω $f(x) = x^3-1$. Να βρεθεί ένα πολυώνυμο παρεμβολής που παρεμβάλλει την f στα σημεία $x_0 = -2, x_1 = -1, x_2 = 1$. Να υπολογίσετε το σφάλμα της παρεμβολής.

6. Εστω $f(x) = e^x$. Θεωρούμε έναν πίνακα τιμών της $f(x)$ στα σημεία $x_i = ih, i = 0,...,N-1$, όπου $N = 1/h$. Να προσδιορίσετε το βήμα h, έτσι ώστε η προσέγγιση της $f(x)$ με ένα πολυώνυμο παρεμβολής $2^{ου}$ βαθμού να δίνει ακρίβεια 3 δεκαδικών ψηφίων.

ΚΕΦΑΛΑΙΟ 6

ΕΛΑΧΙΣΤΑ ΤΕΤΡΑΓΩΝΑ

§ 6.1 *Βέλτιστες προσεγγίσεις σε ευκλείδειους χώρους*

Στο κεφάλαιο αυτό θα ασχοληθούμε με προσεγγίσεις που ελαχιστοποιούν αποστάσεις σε διανυσματικούς χώρους, με νόρμα που προέρχεται από εσωτερικό γινόμενο. Υπενθυμίζουμε ότι

Ορισμός 6.1.1 Εστω X ένας πραγματικός διανυσματικός χώρος. Μία απεικόνιση $(.,.): X \times X \to \boldsymbol{R}$ καλείται ***εσωτερικό γινόμενο*** στο X, αν ισχύουν:

- $(x + y , z) = (x,y) + (x , z)$ για κάθε $x,y,z \in X$
- $(\lambda x , y) = \lambda (x , y)$ για κάθε $x,y \in X$, $\lambda \in \boldsymbol{R}$
- $(x , y) = (y , x)$ για κάθε $x,y \in X$
- $(x , x) > 0$ για κάθε $x,y \in X-\{0\}$.

Ενας πραγματικός διανυσματικός χώρος πεπερασμένης διάστασης, στον οποίο έχει ορισθεί ένα εσωτερικό γινόμενο, καλείται ***ευκλείδειος χώρος***. Αν ισχύει $(x , y) = 0$, θα λέμε ότι τα x,y είναι κάθετα μεταξύ τους ***κάθετα***, ***ή ορθογώνια***. Η ποσότητα $||x|| = \sqrt{(x,x)}$ καλείται ***νόρμα*** του στοιχείου x και η ποσότητα $d(x,y) = || x - y ||$ καλείται ***απόσταση*** μεταξύ των στοιχείων x και y.

Παραδείγματα:

(1) Στο διανυσματικό χώρο $C[a,b]$ των συνεχών συναρτήσεων στο κλειστό διάστημα $[\alpha,b]$, η απεικόνιση:

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

είναι ένα εσωτερικό γινόμενο.

(2) Στο διανυσματικό χώρο $\boldsymbol{R}^n = \{x=(x_1,\ldots,x_n): x_i \in \boldsymbol{R}\}$, η απεικόνιση:

$$(x,y)=\sum_{k=1}^{n}x_i y_i$$

είναι ένα εσωτερικό γινόμενο.

Ορισμός 6.1.2 Εστω X ένας ευκλείδιος διανυσματικός χώρος με νόρμα $||.||$, $x\in X$ και έστω Y ένα υποσύνολο του X. Ένα στοιχείο $y\in Y$ για το οποίο ισχύει:

$$||x\text{-}y|| \le ||x\text{-}z||, \text{ για κάθε } z\in Y,$$

καλείται ***βέλτιστη προσέγγιση*** του X από το Y.

Θεώρημα 6.1.1 Εστω X ένας ευκλείδιος διανυσματικός χώρος με νόρμα $||.||$, $x\in X$ και Y ένας υπόχωρος του X. Ένα στοιχείο $y\in Y$ είναι βέλτιστη προσέγγιση του x από το y, αν και μόνον αν ισχύει:

$$\text{για κάθε } z\in Y,\ \ (x-y,z)=0. \qquad (6.1)$$

Αν λοιπόν ο υπόχωρος Y είναι πεπερασμένης διάστασης n και $\{s_1,...,s_n\}$ είναι μία βάση του, τότε κάθε στοιχείο z του Y μπορεί να γραφεί ως γραμμικός συνδυασμός των στοιχείων της βάσης ως εξής:

$$z=\sum_{k=1}^{n}z_k\ s_k\ ,$$

οπότε η σχέση (6.1) ικανοποιείται αν και μόνον αν:

$$(x-y,s_k)=0,\ i=1,...,n\ ,$$

Συνεπώς, αν $y=\sum_{k=1}^{n}y_k\ s_k$, όπου y_k είναι άγνωστοι συντελεστές, έχουμε:

$$\begin{pmatrix}(s_1,s_1) & (s_1,s_2) & \cdots & (s_1,s_n)\\ (s_2,s_1) & (s_2,s_2) & \cdots & (s_2,s_n)\\ \vdots & \vdots & \ddots & \vdots\\ (s_n,s_1) & (s_n,s_1) & \cdots & (s_n,s_n)\end{pmatrix}\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix}=\begin{pmatrix}(x,s_1)\\ (x,s_2)\\ \vdots\\ (x,s_n)\end{pmatrix}. \qquad (6.2)$$

Το παραπάνω είναι ένα γραμμικό σύστημα ως προς $y_1, \ldots, y_n$, το οποίο λύνεται μονοσήμαντα, αφού το αντίστοιχο ομογενές σύστημα έχει μόνο την τετριμμένη μηδενική λύση. Ο πίνακας των συντελεστών των αγνώστων καλείται ***πίνακας του Gram***. Είναι σαφές από τα παραπάνω, ότι αν η βάση $\{s_1, \ldots, s_n\}$ είναι ορθοκανονική, δηλαδή αν τα στοιχεία της βάσης είναι ανά δύο κάθετα και μοναδιαία, τότε θα ισχύει $(s_i, s_k) = \begin{cases} 1, & i = k \\ 0, & i \neq k \end{cases}$, δηλαδή ο πίνακας του Gram είναι ο μοναδιαίος, άρα παίρνουμε άμεσα ότι:

$$y_k = (x, s_k),$$

συνεπώς η βέλτιστη προσέγγιση y του στοιχείου x είναι η:

$$y = \sum_{k=1}^{n} (x, s_k)\, s_k .$$

Παράδειγμα 6.1 Να προσδιορισθεί το πολυώνυμο 2ου βαθμού, το οποίο είναι η βέλτιστη προσέγγιση της συνάρτησης $f(x) = \eta\mu(\pi x)$ στο διάστημα [-1,1], στο χώρο των πολυωνύμων 2ου βαθμού, ως προς το σύνηθες εσωτερικό γινόμενο του χώρου (βλέπε σελ. 89 παράδειγμα 1). Υπολογίστε το σφάλμα της βέλτιστης προσέγγισης.

Λύση Θεωρούμε τη βάση $\{x^i,\ i=0,1,2\}$ του χώρου των πολυωνύμων 2ου βαθμού, τότε εφόσον για $i,k = 0,1,2$ έχουμε:

$$(s_i, s_k) = \int_{-1}^{1} x^i\ x^k dx = \left[\frac{x^{k+i+1}}{k+i+1} \right]_{-1}^{1} = \frac{1-(-1)^{k+i+1}}{k+i+1},$$

το σύστημα (6.2) γίνεται:

$$\begin{pmatrix} 2 & 0 & 2/3 \\ 0 & 2/3 & 0 \\ 2/3 & 0 & 2/5 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 2/\pi \\ 0 \end{pmatrix},$$

άρα $y_1 = 0,\ y_2 = 3/\pi,\ y_3 = 0$, οπότε:

$$y = \frac{3}{\pi} x.$$

Για το σφάλμα έχουμε:

$$\| f - y \| = \left(\int_{-1}^{1} \left| \eta\mu(\pi x) - \frac{3}{\pi} x \right|^2 dx \right)^{1/2} = 0.626157. \quad \square$$

§ 6.2 *Πολυώνυμα ελαχίστων τετραγώνων*

Δίνονται τα σημεία (x_i, f_i), $i = 1,\ldots,n$ και θέλουμε να προσδιορίσουμε ένα πολυώνυμο βαθμού 1 της μορφής $y = \alpha t + b$, έτσι ώστε το άθροισμα των τετραγώνων:

$$E(a,b) = \sum_{k=1}^{n} \left(f_i - (at_i + b) \right)^2 = \text{ελάχιστο}.$$

Αναγκαία συνθήκη για να ισχύει αυτό είναι:

$$\frac{\partial E(a,b)}{\partial a} = 0, \quad \frac{\partial E(a,b)}{\partial b} = 0,$$

δηλαδή:

$$\begin{cases} n\,a + \left(\sum_{i=1}^{n} t_i \right) b = \left(\sum_{i=1}^{n} f_i \right) \\ \left(\sum_{i=1}^{n} t_i \right) a + \left(\sum_{i=1}^{n} t_i^2 \right) b = \left(\sum_{i=1}^{n} t_i f_i \right) \end{cases}.$$

Η ορίζουσα των συντελεστών του παραπάνω συστήματος είναι πάντα διάφορη του μηδενός, οπότε το σύστημα έχει μοναδική λύση. Το πολυώνυμο που προκύπτει καλείται *πολυώνυμο 1^{ου} βαθμού ελαχίστων τετραγώνων* και χρησιμοποιείται ευρέως στη στατιστική για τη συσχέτιση μεταξύ δύο ποσοτήτων. Ομοίως, αν θέλουμε να προσδιορίσουμε ένα πολυώνυμο $a_0 + a_1 x + \ldots + a_m x^m$ βαθμού m, ώστε το άθροισμα των τετραγώνων να είναι ελάχιστο, καλούμαστε να λύσουμε ένα σύστημα $(m+1) \times (m+1)$ της μορφής:

$$\begin{cases} n\,a_0 + \left(\sum_{i=1}^{n} t_i\right) a_1 + \left(\sum_{i=1}^{n} t_i^2\right) a_2 + \ldots + \left(\sum_{i=1}^{n} t_i^m\right) a_m = \left(\sum_{i=1}^{n} f_i\right) \\ \left(\sum_{i=1}^{n} t_i\right) a_0 + \left(\sum_{i=1}^{n} t_i^2\right) a_1 + \left(\sum_{i=1}^{n} t_i^3\right) a_2 \ldots + \left(\sum_{i=1}^{n} t_i^{m+1}\right) a_m = \left(\sum_{i=1}^{n} t_i f_i\right) \\ \vdots \\ \left(\sum_{i=1}^{n} t_i^m\right) a_0 + \left(\sum_{i=1}^{n} t_i^{m+1}\right) a_1 + \left(\sum_{i=1}^{n} t_i^{m+2}\right) a_2 \ldots + \left(\sum_{i=1}^{n} t_i^{2m}\right) a_m = \left(\sum_{i=1}^{n} t_i^m f_i\right) \end{cases}.$$

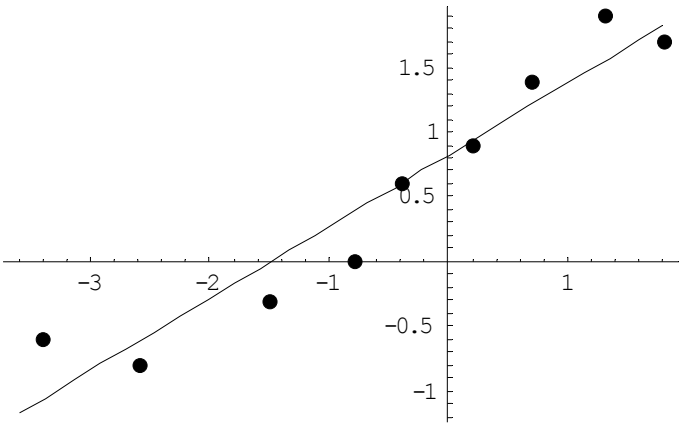
Παράδειγμα 6.2 Υπολογίστε την ευθεία των ελαχίστων τετραγώνων που προσεγγίζει τα σημεία:

t	*-3.4*	*-2.6*	*-1.5*	*-0.8*	*-0.4*	*0.2*	*0.7*	*1.3*	*1.8*
F	*-0.6*	*-0.8*	*-0.3*	*0*	*0.6*	*0.9*	*1.4*	*1.9*	*1.7*

Λύση: Προφανώς έχουμε $n = 9$ σημεία, $\sum_{k=1}^{9} t_k = -4.7$, $\sum_{k=1}^{9} f_k = 4.8$, $\sum_{k=1}^{9} t_k^2 = 26.83$, $\sum_{k=1}^{9} t_k f_k = 11.02$, άρα λύνουμε το σύστημα:

$$\begin{cases} 9\,a - 4.7\,b = 4.8 \\ -4.7\,a + 26.83\,b = 11.02 \end{cases}$$

και βρίσκουμε $\alpha = 0.823129,\ b = 0.554928$. Αρα η ευθεία των ελαχίστων τετραγώνων έχει τη μορφή $y = 0.823129\,t + 0.554928$. □



Σχήμα 6: Η ευθεία των ελαχίστων τετραγώνων.

ΑΛΥΤΕΣ ΑΣΚΗΣΕΙΣ

1. Προσδιορίστε τη βέλτιστη προσέγγιση της συνάρτησης $f(x) = 2\,e^x+1$ από ένα πολυώνυμο $3^{ου}$ βαθμού, ως προς το σύνηθες εσωτερικό γινόμενο του χώρου των συνεχών συναρτήσεων στο διάστημα [-1,1].

2. Προσδιορίστε τη βέλτιστη προσέγγιση της συνάρτησης $f(x) = e^x+συν(\pi x)$ από ένα πολυώνυμο $4^{ου}$ βαθμού, ως προς το σύνηθες εσωτερικό γινόμενο του χώρου των συνεχών συναρτήσεων στο διάστημα [1,4].

3. Υπολογίστε την ευθεία των ελαχίστων τετραγώνων που προσεγγίζει τα σημεία:

t	*-3*	*-2*	*0*	*1*	*4*	*6*
F	*-1.4*	*0*	*1.2*	*5.5*	*7*	*9*

Απάντ: *y = 2.37667 +1.17333 x.*

4. Υπολογίστε την ευθεία των ελαχίστων τετραγώνων που προσεγγίζει τα σημεία:

t	*-4.2*	*-3.6*	*-2.3*	*-1.8*	*-1.1*	*-0.4*	*0.6*	*1.2*
F	*10*	*7*	*13*	*11*	*9*	*12*	*8*	*11*

Απάντ: *y = 10.3441 +0.151099 x.*

ΚΕΦΑΛΑΙΟ 7

ΠΡΟΣΕΓΓΙΣΤΙΚΗ ΟΛΟΚΛΗΡΩΣΗ

Είναι γνωστό ότι για πολλά ορισμένα ολοκληρώματα δεν υπάρχουν αναλυτικές μέθοδοι ακριβούς επίλυσής τους. Ετσι λοιπόν έχουν αναπτυχθεί προσεγγιστικές μέθοδοι υπολογισμού τέτοιων ολοκληρωμάτων. Στο Κεφάλαιο αυτό αναπτύσσουμε τις μεθόδους τραπεζίου, Simpson και Romberg, δίνοντας έμφαση στον τρόπο προσέγγισης τέτοιων προβλημάτων.

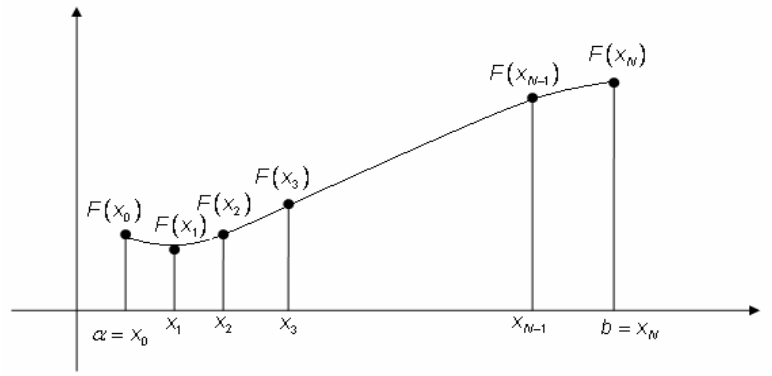
7.1. Η μέθοδος τραπεζίου

Με τη μέθοδο αυτή προσεγγίζουμε την τιμή του ολοκληρώματος μιας συνεχούς συνάρτησης *f* σ' ένα κλειστό και φραγμένο διάστημα $[\alpha,b]$ με χρήση εμβαδών τραπεζίων που προκύπτουν από την προσέγγιση της συνάρτησής μας από μία τεθλασμένη γραμμή. Για ευκολία υποθέτουμε ότι η *f* είναι θετική στο $[\alpha,b]$ όπως φαίνεται στο σχήμα 1. Η διαδικασία που ακολουθούμε είναι η εξής:

- Έστω $\{x_0 = \alpha, x_1, ..., x_N = b\}$ $x_0 < x_1 < .... < x_N$ είναι ένας ομοιόμορφος διαμερισμός του $[\alpha,b]$, δηλαδή χωρίζουμε το $[\alpha,b]$ σε *N* ισομήκη υποδιαστήματα.

- Τότε: $x_i = x_0 + \kappa\dfrac{b-\alpha}{N}, \quad \kappa = 0,...,N.$

- Υπολογίζουμε τις τιμές $f(x_i), \quad i = 0,....,N.$



Σχήμα 1

- Σχηματίζουμε τα διαδοχικά ευθύγραμμα με άκρα τα $f(x_0),...,f(x_N)$ οπότε σχηματίζεται μία τεθλασμένη γραμμή.

- Υπολογίζουμε τα εμβαδά των Ν-τραπεζίων που σχηματίζονται και έχουμε:

$$\int_{\alpha}^{b} f(x)dx \cong E_{\tau\rho\alpha\pi_{\cdot 1}} + \ldots + E_{\tau\rho\alpha\pi_{\cdot N}}$$

$$= \frac{f(x_0)+f(x_1)}{2}(x_1 - x_0) + \frac{f(x_1)+f(x_2)}{2}(x_2 - x_1) + \ldots + \frac{f(x_{N-1})+f(x_N)}{2}(x_N - x_{N-1})$$

$$= \frac{b-\alpha}{2N}(f(x_0)+f(x_1)+f(x_1)+f(x_2)+f(x_2)+f(x_3)\ldots+f(x_{N-1})+f(x_{N-1})+f(x_N)$$

$$= \frac{b-\alpha}{2N}(f(x_0)+f(x_N)+2(f(x_1)+\ldots+f(x_{N-1})))$$

$$= \frac{b-\alpha}{2N}\left(f(x_0)+f(x_N)+2\sum_{k=1}^{N-1}f(x_k)\right) \underset{N\to+\infty}{\to} \int_{\alpha}^{b} f(x)dx. \qquad (1)$$

Σφάλμα:

Είναι γνωστό ότι αν προσεγγίσουμε μία συνεχή συνάρτηση $f(x)$ σε ένα κλειστό διάστημα $[\alpha,b]$ με μία τεθλασμένη γραμμή, δηλαδή με ένα πολυώνυμο $1^{ου}$ βαθμού $p_1(x)$, τότε το σφάλμα είναι:

$$f(x)-p_1(x) = \frac{f''(\xi)}{2}(x-\alpha)(x-b) = -\frac{f''(\xi)}{2}(x-\alpha)(b-x),\ \xi \in (\alpha,b)$$

υπό την προϋπόθεση ότι η f είναι 2 φορές παραγωγίσιμη συνάρτηση.

Επομένως:

$$\int_{\alpha}^{b} f(x)-p_1(x)dx = -\frac{f''(\xi)}{2}\int_{\alpha}^{b}(x-\alpha)(b-x)dx$$

$$= -\frac{f''(\xi)}{2}\frac{(b-\alpha)^3}{6} = -\frac{f''(\xi)}{12}(b-\alpha)^3,$$

άρα εάν e είναι το σφάλμα στην περίπτωση της μεθόδου τραπεζίου, έχουμε:

$$e = \int_{\alpha}^{b} f(x)dx - \left(\frac{b-\alpha}{2N}\left(f(x_0)+f(x_N)+2\sum_{k=1}^{N-1}f(x_k)\right)\right)$$

άρα:

$$e = -\frac{f''(\xi_1)}{12}(x_1 - x_0)^3 - \frac{f''(\xi_2)}{12}(x_2 - x_1)^3 - \ldots - \frac{f''(\xi_N)}{12}(x_N - x_o)^3,\ \xi_i \in (x_i, x_{i+1})$$

$$= -\frac{(b-\alpha)^3}{12N^3}(f''(\xi_1) + \ldots + f''(\xi_N)).$$

Αν λοιπόν $M = \max_{x\in[\alpha,b]}\{|f''(x)|: x \in [\alpha,b]\}$, τότε:

$$|e| \le \frac{(b-\alpha)^3}{12N^3}(M + \ldots + M) = \frac{(b-\alpha)^3}{12N^3}MN = \frac{(b-\alpha)^3}{12N^2}\cdot M. \qquad (2)$$

<u>Παράδειγμα</u> 1 Υπολογίστε την προσεγγιστική τιμή του ολοκληρώματος $\int_0^1 e^{-x^2}dx$, χρησιμοποιώντας $N = 8$ ισομήκεις υποδιαιρέσεις του κλειστού διαστήματος $[0,1]$ με τη μέθοδο τραπεζίου και υπολογίστε το σφάλμα.

Λύση:

Για τον υπολογισμό της προσεγγιστικής τιμής του ολοκληρώματος θα χρησιμοποιήσουμε τον τύπο (1).

- b – α = μήκος διαστήματος ολοκλήρωσης = 1 – 0 = 1.
- N = πλήθος υποδιαστημάτων = πλήθος σημείων – 1 = 8. Επομένως χρειαζόμαστε 9 σημεία.
- Εύρος υποδιαστημάτων = $\frac{b-\alpha}{N} = \frac{1-0}{8} = 0.125$,

άρα:

$$x_0 = 1,\ x_1 = 0.125, x_2 = 0.25,\ x_3 = 0.375, \ldots, x_7 = 0.875,\ x_8 = 1$$

- Εφόσον $f(x) = e^{-x^2}$, υπολογίζουμε τις τιμές $f(x_i),\ i = 0,\ldots,8$ και προκύπτει ο ακόλουθος πίνακας τιμών:

0	0.125	0.25	0.375	0.5	0.625	0.75	0.875	1
1	0.9844	0.9394	0.8688	0.7788	0.6766	0.5697	0.465	0.3678

Χρησιμοποιούμε τον τύπο (1) για τις τιμές του παραπάνω πίνακα και προκύπτει ότι :

$$\int f(x)dx \cong 0.7458 .$$

Για τον υπολογισμό του σφάλματος αρκεί να υπολογίσουμε τη σταθερά M του τύπου 2. Παρατηρούμε ότι $f''(x) = -2e^{-x^2} + 4e^{-x}x^2$ και είναι εύκολο να δει κανείς ότι

$$M = \max\{|f''(x)| : x \in [0,1]\} = |f''(0)| = 2,$$

οπότε:

$$|e| \le \frac{2 \cdot 1^3}{12 \cdot 8^2} = \frac{1}{6 \cdot 64} = \frac{1}{384}.$$

ΑΣΚΗΣΕΙΣ

1. Να γίνει εφαρμογή της μεθόδου τραπεζίου στα δεδομένα:

-1	-0,5	0	0,5	1	1,5	2
-4	2	0	1	4	2	6

2. Υπολογίστε όλες τις συνεχείς συναρτήσεις f για τις οποίες το σφάλμα υπολογισμού του $\int_\alpha^b f(x)dx$ με χρήση της μεθόδου τραπεζίου είναι μηδέν.

7.2. Η μέθοδος Simpson:

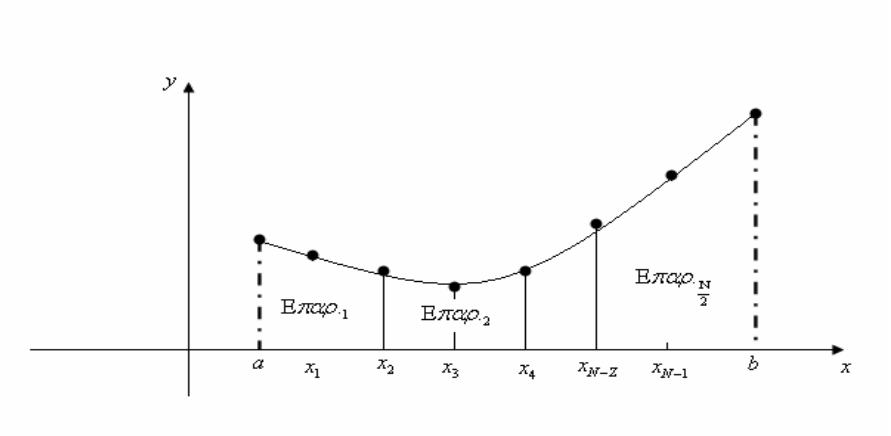
Με τη μέθοδο αυτή προσεγγίζουμε την τιμή του $\int_\alpha^b f(x)dx$ μιας συνεχούς συνάρτησης σε ένα κλειστό και φραγμένο διάστημα $[\alpha,b]$ με χρήση εμβαδών παραβολών, οι οποίες προκύπτουν από την προσέγγιση της συνάρτησής μας σε στοιχειώδη υποδιαστήματα του $[\alpha,b]$ από πολυώνυμα $2^{ου}$ βαθμού, δηλ. από παραβολές. Όπως θα δούμε παρακάτω το σφάλμα (για τον ίδιο αριθμό υποδιαιρέσεων του $[\alpha,b]$) είναι καλύτερο σε σχέση με τη μέθοδο τραπεζίου. Υπάρχουν διάφορες παραλλαγές της μεθόδου αυτής. Για την καλύτερη κατανόηση της μεθόδου αναφέρουμε τη βασική της εκδοχή.

- Έστω $\{x_0 = \alpha, x_1, ..., x_N = b\}$ $x_0 < x_1 < .... < x_N$ είναι ένας ομοιόμορφος διαμερισμός του $[\alpha,b]$, δηλαδή χωρίζουμε το $[\alpha,b]$ σε *N* ισομήκη υποδιαστήματα.

- Τότε: $x_i = x_0 + \kappa\dfrac{b-\alpha}{N}, \qquad \kappa = 0,\ldots,N.$

- Υπολογίζουμε τις τιμές $f(x_i), \quad i = 0,....,N.$

- Σχηματίζουμε τις διαδοχικές παραβολές που διέρχονται από τα σημεία $f(x_i), f(x_{i+1}), f(x_{i+2}),\ i = 0,...,N/2$ οπότε πρέπει **N = ζυγός**.



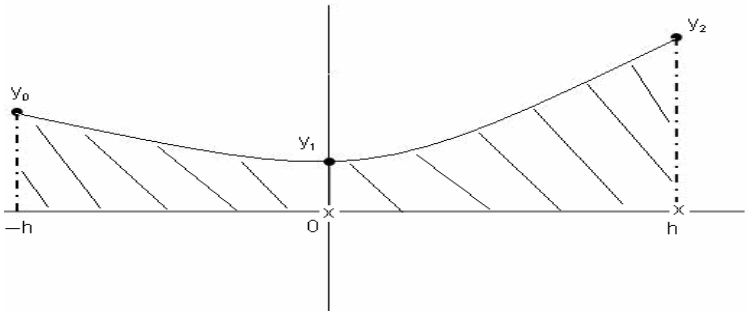
Σχήμα 2

- Υπολογίζουμε τα εμβαδά των N/2-παραβολών που σχηματίζονται και έχουμε:

$$\int_{\alpha}^{b} f(x)dx \cong E_{\text{παραβ.}1} + \ldots + E_{\text{παραβ.}\frac{N}{2}}.$$

Για να υπολογίσουμε τα προαναφερθέντα εμβαδά χρήσιμο είναι το ακόλουθο:

Θεώρημα Εστω παραβολή $y(x) = \alpha x^2 + bx + c$ όπως στο κάτωθι σχήμα:



Σχήμα 3

τότε:

$$E_{\pi\alpha\rho\alpha\beta.} = \int_{-h}^{h} (\alpha x^2 + bx + c)dx = \frac{h}{3}(y_0 + 4y_1 + y_2).$$

Απόδειξη:

$$E_{\pi\alpha\rho\alpha\beta.} = \int_{-h}^{h} (\alpha x^2 + bx + c)dx = \frac{\alpha x^3}{3} + \frac{bx^2}{2} + cx \Big|_{-h}^{h}$$

$$= \frac{\alpha h^3}{3} + \frac{bh^2}{2} + ch - \left(-\frac{\alpha h^3}{3} + \frac{bh^2}{2} - ch\right) = \frac{2\alpha h^3}{3} + 2ch = \frac{h}{3}\left(2\alpha h^2 + 6c\right).$$

Αλλά:

$$\begin{aligned} &\alpha(-h)^2 + b(-h) + c = y_0 && \alpha h^2 - bh + c = y_0 \\ &\alpha 0^2 + b0 + c = y_1 && c = y_1 \\ &\alpha h^2 + bh + c = y_2 && \alpha h^2 + bh + c = y_2 \\ &+ \_\_\_\_\_ \\ & && 2\alpha h^2 + 4c = y_0 + 4y_1 + y_2 \end{aligned}$$

Τελικά:

$$E_{\pi\alpha\rho\alpha\beta} = \frac{h}{3}(y_0 + 4y_1 + y_2). \quad \square$$

Τώρα μπορούμε να υπολογίσουμε:

$$\int_{\alpha}^{b} f(x)dx \cong E_{\pi\alpha\rho\alpha\beta\cdot 1} + \ldots + E_{\pi\alpha\rho\alpha\beta\cdot \frac{N}{2}}$$

$$= \frac{b-\alpha}{3N}\left(f(x_0) + 4f(x_1) + f(x_2)\right) + \frac{b-\alpha}{3N}\left(f(x_2) + 4f(x_3) + f(x_4)\right)$$

$$+ \ldots + \frac{b-\alpha}{3N}\left(f(x_{N-2}) + 4f(x_{N-1}) + f(x_N)\right)$$

$$= \frac{b-\alpha}{3N}(f(x_0) + 4f(x_1) + f(x_2) + f(x_2) + 4f(x_3) + f(x_4) + \ldots$$

$$\ldots + f(x_{N-2}) + 4f(x_{N-1}) + f(x_N))$$

$$\int_{\alpha}^{b} f(x)dx \cong \frac{b-\alpha}{3N}\left(f(x_0)+f(x_N)+2\sum_{i=1}^{\frac{N}{2}-1} f(x_{2i})+4\sum_{i=1}^{\frac{N}{2}} f(x_{2i-1})\right). \qquad (3)$$

Σφάλμα:

Εργαζόμενοι όπως παραπάνω, δηλαδή λαμβάνονται υπόψη ότι ο τύπος Simpson ολοκληρώνει ακριβώς και πολυώνυμα $3^{ου}$ βαθμού έχουμε:

$$f(x)-p_3(x)=\frac{f^{(4)}(\xi)}{24}(x-\alpha)\left(x-\frac{\alpha+b}{2}\right)^2(b-x) \text{ κ.λπ.}$$

οπότε υπολογίζουμε:

$$|e|\le \frac{(b-\alpha)^5}{180N^4}M, \text{ όπου } M=\max\left\{|f^{(4)}(x)|: x\in[\alpha,b]\right\}. \qquad (4)$$

<u>Παράδειγμα</u> 1 Υπολογίστε την προσεγγιστική τιμή του ολοκληρώματος $\int_0^1 e^{-x^2}dx$, χρησιμοποιώντας $N=8$ ισομήκεις υποδιαιρέσεις του κλειστού διαστήματος $[0,1]$ με τη μέθοδο Simpson και υπολογίστε το σφάλμα.

Λύση:

Για τον υπολογισμό της προσεγγιστικής τιμής του ολοκληρώματος θα χρησιμοποιήσουμε τον τύπο (3).

- $b-\alpha$ = μήκος διαστήματος ολοκλήρωσης = $1-0=1$.
- N = πλήθος υποδιαστημάτων = πλήθος σημείων – 1 = 8. Επομένως χρειαζόμαστε 9 σημεία.
- Εύρος υποδιαστημάτων = $\frac{b-\alpha}{N}=\frac{1-0}{8}=0.125$,

άρα:

$$x_0=1,\ x_1=0.125, x_2=0.25,\ x_3=0.375,...,x_7=0.875,\ x_8=1$$

- Εφόσον $f(x)=e^{-x^2}$, υπολογίζουμε τις τιμές $f(x_i),\ i=0,...,8$ και προκύπτει ο ακόλουθος πίνακας τιμών:

0	0.125	0.25	0.375	0.5	0.625	0.75	0.875	1
1	0.9844	0.9394	0.8688	0.7788	0.6766	0.5697	0.465	0.3678

Χρησιμοποιούμε τον τύπο (3) για τις τιμές του παραπάνω πίνακα και προκύπτει ότι:

$$f(x)dx \cong 0.7467 .$$

Για τον υπολογισμό του σφάλματος αρκεί να υπολογίσουμε τη σταθερά M του τύπου (4). Με διαδοχικές παραγωγίσεις είναι εύκολο να δει κανείς ότι

$$M = \max\left\{\left|f^{(4)}(x)\right| : x \in [0,1]\right\} = 12,$$

οπότε:

$$|e| \le \frac{12}{180 \bullet 8^4} 1^5 = 0.000016 .$$

ΑΣΚΗΣΕΙΣ

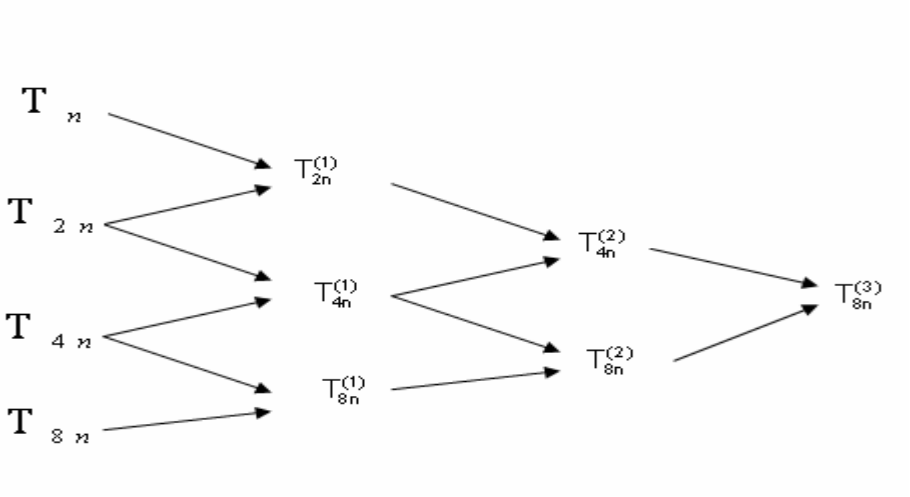
1. Να γίνει εφαρμογή της μεθόδου Simpson στα δεδομένα:

-1	-0.5	0	0.5	1	1.5	2
-4	2	0	1	4	2	6

2. Υπολογίστε όλες τις συνεχείς συναρτήσεις f για τις οποίες το σφάλμα υπολογισμού του $\int_{\alpha}^{b} f(x)dx$ με χρήση της μεθόδου Simpson είναι μηδέν.

7.3. **Ολοκλήρωση Romberg**

Χρησιμοποιεί μία τεχνική διαδοχικών διχοτομήσεων του διαστήματος ολοκλήρωσης με στόχο τη μείωση του σφάλματος αποκοπής. Αν T_n είναι η προσέγγιση του ολοκληρώματος που προκύπτει από τον κανόνα τραπεζίου με n-υποδιαστήματα υπολογίζουμε τις T_{2n}, T_{4n}, T_{8n} κ.λπ. Συνδυάζοντας τις προσεγγίσεις αυτές μπορούμε να πάρουμε ακόμα καλύτερες προσεγγίσεις με τον ακόλουθο τρόπο:



Σχήμα 4

όπου εάν το συνολικό πλήθος των υποδιαστημάτων του [α,b] είναι $N = 2^{\mu}$ n, τότε:

$$T_{2i}^{(j)} = T_{2i}^{(j-1)} + \frac{T_{2i}^{(j-1)} - T_{i}^{(j-1)}}{4^{j} - 1},\ i, j = 1,2,...\mu . \qquad (5)$$

Παράδειγμα: Υπολογίστε με τη μέθοδο Romberg το $\int_0^{0.8} \frac{\eta\mu x}{x} dx$ χρησιμοποιώντας 8 υποδιαιρέσεις του [0,0.8] όπως παρακάτω:

0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
1	0.9983	0.9933	0.9851	0.9735	0.9589	0.9411	0.9203	0.8967

Λύση: Αρχικά υπολογίζουμε το $\int_0^{0.8} \frac{\eta\mu x}{x} dx$ με χρήση του τύπου (1) του τραπεζίου για n =1, 2, 4, 8 ισομήκεις υποδιαιρέσεις του διαστήματος [0,0.8]:

$$n = 1:\ T_1 = \frac{b-\alpha}{2}(f_0 + f_8) = 0.7586 .$$

$$n = 2:\ T_2 = \frac{b-\alpha}{4}(f_0 + 2(f_4 + f_8)) = 0.7687 .$$

$$n = 4:\ T_4 = \frac{b-\alpha}{8}(f_0 + 2(f_2 + f_4 + f_6) + f_8) = 0.7712 .$$

$$n = 8:\ T_8 = \frac{b-\alpha}{16}(f_0 + 2(f_1 + ... + f_7) + f_8) = 0.7718 .$$

Στη συνέχεια χρησιμοποιούμε τον τύπο (5) (βλέπε σχήμα 4) για j = 1:

$$T_2^{(1)} = T_2^{(0)} + \frac{T_2^{(0)} - T_1^{(0)}}{4-1} = 0.7221.$$

$$T_4^{(1)} = T_4^{(0)} + \frac{T_4^{(0)} - T_2^{(0)}}{4-1} = 0.772097\,.$$

$$T_8^{(1)} = T_8^{(0)} + \frac{T_8^{(0)} - T_4^{(0)}}{4-1} = 0.772095\,.$$

έπειτα για j = 2: $$T_4^{(2)} = T_4^{(1)} + \frac{T_4^{(1)} - T_2^{(1)}}{4^2-1} = 0.77209577\,.$$

$$T_8^{(2)} = T_8^{(1)} + \frac{T_8^{(1)} - T_4^{(1)}}{4^2-1} = 0.77209578\,,$$

και τελικά:

$$I \cong T_8^{(2)} + \frac{T_8^{(2)} - T_4^{(2)}}{4^3-1} = 0.77209578\,.$$

ΚΕΦΑΛΑΙΟ 8

ΑΡΙΘΜΗΤΙΚΕΣ ΜΕΘΟΔΟΙ ΕΠΙΛΥΣΗΣ ΣΥΝΗΘΩΝ ΔΙΑΦΟΡΙΚΩΝ ΕΞΙΣΩΣΕΩΝ

Έστω $[\alpha,b]\subset\mathbb{R}$, $f:[\alpha,b]\times\mathbb{R}\to\mathbb{R}$, $y_o\in\mathbb{R}$. Το τυπικό πρόβλημα αρχικών τιμών που θα μας απασχολήσει, είναι το ακόλουθο:

Ζητείται μια συνάρτηση $y:[\alpha,b]\to\mathbb{R}$:

$$\begin{cases} y'(t)=f(t,y(t)) \\ y(\alpha)=y_0 \end{cases},\ \alpha\le t\le b. \qquad (1)$$

Η θεωρία των διαφορικών εξισώσεων μελετά συνθήκες στην f που εξασφαλίζουν την ύπαρξη και μοναδικότητα της λύσης, καθώς επίσης και την εξάρτηση της λύσης από τα δεδομένα, π.χ. εάν $\|\tilde{y}_o-y_o\|$ είναι μικρή, τότε η διαφορά $\|y-\tilde{y}\|$ πόση μικρή είναι. Ειδικά στην περίπτωση που η f είναι πολυώνυμο $1^{ου}$ βαθμού ως προς $y(t)$ δηλαδή:

$$y'(t)=p(t)y(t)+q(t)$$

η δ.ε. είναι γραμμική, με λύση

$$y(t)=e^{\int_\alpha^t p(t)dt}\left\{y_0+\int_\alpha^t q(s)e^{-\int_\alpha^s p(r)dr}ds\right\}.$$

Για γενική f όμως όχι μόνο δεν μπορούμε να δώσουμε λύση σε κλειστή μορφή, αλλά ούτε να εγγυηθούμε την ύπαρξη και μοναδικότητά της. Προβλήματα με μη μονοσήμαντη λύση είναι δύσκολο να προσεγγισθούν αριθμητικά. Στην περίπτωση μοναδικής λύσης τα πράγματα είναι αρκετά απλούστερα.

Θεώρημα 1 Έστω $f\in C[\alpha,b]\times\mathbb{R}$ είναι μία συνάρτηση που ικανοποιεί τη συνθήκη Lipchitz δηλαδή:

$$\exists\ L\ge 0:\forall t\in[\alpha,b]\quad \forall y_1,y_2\in\mathbb{R}\quad |f(t,y_1)-f(t,y_2)|\le L|y_1-y_2|,$$

τότε το πρόβλημα (1) λύνεται μονοσήμαντα
Η συνθήκη αυτή είναι περιοριστική. Αντικαθιστώντας την με μία «τοπική» συνθήκη Lipchitz, τα πράγματα γίνονται καλύτερα. Έτσι:

Θεώρημα 2 Εστω $fC[\alpha,b]\times[y_o-c,y_0+c]$ και

$$\exists\ L\geq 0:\ \forall t\in[\alpha,b]\ \forall y_1,y_2\in[y_o-c,y_o+c]\ \ \left|f(t,y_1)-f(t,y_2)\right|\leq L\left|y_1-y_2\right|,$$

τότε το πρόβλημα αρχικών τιμών λύνεται μονοσήμαντα τουλάχιστον στο διάστημα $[\alpha,b']$, όπου $b'=\min\left(b,\alpha+\frac{c}{A}\right)$, $A=\max\limits_{\substack{\alpha\leq t\leq b\\ y_o-c\leq y\leq y_o+c}}\{|f(t,y)|\}$.

8.1 Μέθοδος Eüler

Υποθέτουμε ότι το πρόβλημα των αρχικών τιμών λύνεται μονοσήμαντα. Θεωρούμε έναν ομοιόμορφο διαμερισμό $\alpha=t_o<t_1<...t_N=b$ του $[\alpha,b]$. Εστω $y_1,...,y_N$ είναι οι προσεγγίσεις των πραγματικών τιμών $y(t_n)$, n=0,...,N-1 της λύσης της δ.ε. (1), τότε οι $y_1,...,y_N$ προσδιορίζονται από τον αναδρομικό τύπο:

$$y_{n+1} = y_n + hf(t_n,y_n)\ \text{n=0,...,N-1}.$$

Αυτό τεκμηριώνεται ως εξής: Αν $h\to 0$, τότε:

$$f'(x)\cong\frac{f(x+h)-f(x)}{h}\Rightarrow f(x+h)\cong f(x)+hf'(x),$$

οπότε εάν $t_o, t_1, ..., t_N$ ομοιόμορφος διαμερισμός, τότε $t_{n+1}=t_n+h$, συνεπώς χρησιμοποιώντας την παραπάνω με ισότητα αντί $\cong$ γράφουμε:

$$y_{n+1}=y_n+hy_n'=y_n+hf\left(t_n,y_n\right),\ \ \text{n=0,...,N-1},$$

όπου $y_n=y\left(t_n\right)$.

Θεώρημα 3 Αν $y_o, ...,y_N$ είναι οι προσεγγίσεις τις οποίες δίνει η μέθοδος Eüler για τον ομοιόμορφο διαμερισμό του $[\alpha,b]$ με βήμα $h=\frac{b-\alpha}{N}$, τότε:

$$\max\left|y_n-y\left(t_n\right)\right|\leq\frac{M}{2L}\left(e^{L(b-\alpha)}-1\right)h$$

όπου $M=\max\limits_{t\in[\alpha,b]}|y''(t)|$ και L η ανισότητα στη συνθήκη Lipchitz.

Απόδειξη: Από το πολυώνυμο Taylor 2^{ου} βαθμού έχουμε:

$$y(t_{n+1}) = y(t_n) + hy'(t_n) + \frac{h^2}{2} y''(\xi_n),\ \xi_n \in (t_n, t_{n+1})$$

άρα:

$$y(t_{n+1}) - y_{n+1} = y(t_n) - y_n + h\big(f(t_n, y(t_n)) - f(t_n, y_n)\big) + \frac{h^2}{2} y''(\xi_n)$$

$$|e_{n+1}| \ \le\ |e_n| + hL|y(x_n) - y_n| + \frac{M}{2}h^2$$

$$|e_{n+1}| \ \le\ |e_n| + hL|e_n| + \frac{M}{2}h^2 = (1+Lh)|e_n| + \frac{M}{2}h^2$$

$$|e_{n+1}| \ \le\ (1+Lh)|e_n| + \frac{Mh^2}{2}$$

Επειδή όμως:

$$d_{n+1} \le (1+\delta)d_n + K \Rightarrow d_n \le d_o e^{n\delta} + K\frac{e^{n\delta} - 1}{\delta}$$

έχουμε:

$$|e_{n+1}| \le e^{nLh}|e_o| \frac{Mh^2}{2} \frac{e^{nhL} - 1}{hL}$$

και επειδή $nh \le b - c$ έχουμε τελικά

$$|e_{n+1}| \le \frac{Mh}{2} \frac{e^{L(b-\alpha)} - 1}{L}. \quad \square$$

Το παραπάνω σφάλμα που υπολογίσαμε είναι το λεγόμενο ολικό σφάλμα που είναι της τάξης 1. Το σφάλμα:

$$\delta_n = \big(y(t_n) + hf\big(t_n, y(t_n)\big)\big) - y\big(t_{n+1}\big)$$

καλείται τοπικό σφάλμα και αν αναπτύξουμε με Taylor την $y(t)$ για $t = t_{n+1}$ παίρνουμε:

$$\delta_n = \big(y(t_n) + hf\big(t_n, y(t_n)\big)\big) - \left(y(t_n) + y'(t_n)h + \frac{y''(\xi_n)}{2}\right)h^2 = -\frac{f''(\xi_n)}{2},$$

δηλαδή το τοπικό σφάλμα είναι τάξης 2 και αυτό είναι ένα γενικό φαινόμενο, διότι τα σφάλματα συσσωρεύονται.

Παράδειγμα 1 Δίνεται η διαφ. εξίσωση $y'(t) = 2\,t\,y(t) + 1,\ \ y(0) = 1.$ Θεωρείστε έναν ομοιόμορφο διαμερισμό του $[0,1]$ με βήμα $h = 0.25$ και υπολογίστε τις προσεγγίσεις της λύσης y της δ.ε. στα σημεία του διαμερισμού με τη μέθοδο του Eüler. Στη συνέχεια χρησιμοποιήστε την παρεμβολή Newton για να προσεγγίσετε τη λύση της δ. ε.

Λύση: Εφαρμόζουμε τον αναδρομικό τύπο της μεθόδου Eüler:

$$y_1 = y_0 + hf\left(t_0, y\left(t_0\right)\right) = 1 + 0.25\bullet(2\bullet 0\bullet 1 + 1) = 1.25$$

$$y_2 = y_1 + hf\left(t_1, y_2\right) = 1.25 + 0.25\bullet(2\bullet 0.25\bullet 1.25 + 1) = 1.65625$$

$$y_3 = y_2 + hf\left(t_2, y_2\right) = 1.65625 + 0.25(2\bullet 0.5\bullet 1.65625 + 1) = 2.984375$$

$$y_4 = y_3 + hf\left(t_3, y_3\right) = 2.984375 + 0.25(2\bullet 0.75\bullet 2.984375 + 1) = 7.091796$$

και έτσι σχηματίζουμε τον ακόλουθο πίνακα τιμών της προσεγγιστικής λύσης της δ.ε. στα σημεία 0, 0.25, 0.5, 0.75, 1:

	t_0	t_1	t_2	t_3	T_4
t	0	0.25	0.5	0.75	1
y	1	1.25	1.65625	2.984375	7.091796
	y_0	y_1	y_2	y_3	y_4

Στη συνέχεια εφαρμόζουμε την μέθοδο παρεμβολής του Newton (βλέπε Κεφ. 5) για τα προαναφερθέντα σημεία και παίρνουμε:

0	1	1	1.25	8.16667	11.64559
0.25	1.25				
0.5	1.65625	1.625	7.375		
0.75	2.984375	5.3125	22.2342	19.81226	
1	7.091796	16.4296			

άρα η προσεγγιστική λύση της δ.ε. είναι η ακόλουθη:

$$y(x) = 1 + (x-0) + 1.25(x-0)(x-0.25) + 8.16667(x-0)(x-0.25)(x\text{-}0.5)$$

$$+11.64559(x-0)(x-0.25)(x-0.5)(x-0.75).$$

Ευστάθεια της μεθόδου Eüler

Έστω f ικανοποιεί τις προϋποθέσεις του Θεωρήματος 2. Συνεπώς για δοσμένες αρχικές τιμές $y_o, z_o \in \mathbb{R}$ τα προβλήματα αρχικών τιμών:

$$\begin{cases} y'(t) = f(t, y(t)) \\ y(\alpha) = y_o \end{cases} \qquad \begin{cases} z'(t) = f(t, z(t)) \\ z(\alpha) = z_o \end{cases}$$

έχουν μονοσήμαντες λύσεις $y, z \in C[\alpha, b]$.

Έστω $\varepsilon(t) = \mathrm{y}(t) - \mathrm{z}(t)$, τότε έχουμε:

$$\varepsilon'(t) = \mathrm{y}'(t) - \mathrm{z}'(t) = f(t, y(t)) - f(t, z(t)),$$

άρα:

$$\frac{1}{2}\frac{d}{dt}\varepsilon^2(t) = \varepsilon(t)\varepsilon'(t) = \varepsilon(t)\big(f(t, y(t)) - f(t, z(t))\big)$$

$$\le |\varepsilon(t)| \; |f(t, y(t)) - f(t, z(t))| \; \le \; |\varepsilon(t)| L |y(t) - z(t)| = L\varepsilon^2(t).$$

Αν λοιπόν $\varphi(t) = \varepsilon^2(t)$, τότε:

$$\varphi' - 2L\varphi \le 0 \Rightarrow e^{-2Lt}\varphi'(t) - 2L\varphi(t)e^{-2Lt} = \frac{d}{dt}\left(e^{-2Lt}\varphi(t)\right) \le 0$$

άρα η $e^{-2Lt}\varphi(t)$ είναι φθίνουσα $\searrow \forall t \in [\alpha, b]$, άρα:

$$e^{-2Lt}\varphi(t) \le e^{-2L\alpha}\varphi(\alpha),$$

οπότε:

$$|e(t)| \le e^{L(t-\alpha)}|e(\alpha)| \Rightarrow \; \max\{|y(t) - z(t)| \le e^{L(b-\alpha)}|y_o - z_o|.$$

Η έννοια αυτή δεν είναι πολύ χρήσιμη στην πράξη. Έστω

$$\begin{cases} y'(t) = \lambda y(t) \\ y(0) = 1 \end{cases}, \; \lambda < 0, \; t \in [0, \infty)$$

είναι γραμμική δ.ε. η οποία έχει μοναδική λύση $y = e^{\lambda t} \xrightarrow[t \to +\infty]{} 0$. Έστω $t_n = nh$, τότε:

$$y_{n+1} = y_n + hy_n\lambda = (1 + h\lambda)y_n = (1 + h\lambda)^n = \left(1 + \frac{\lambda t}{n}\right)^n \xrightarrow[n \to +\infty]{} e^{\lambda t}.$$

Στην πράξη οι υπολογισμοί γίνονται με μικρό αλλά θετικό h. Έτσι:

$$|y_n| \to 0 \qquad \alpha\nu \qquad |1+h\lambda| < 1 = -1 < h\lambda < 1 \Rightarrow -2 < h\lambda < 0$$

Λέμε ότι η μέθοδος είναι απόλυτα ευσταθής για $h > 0$, αν όταν εφαρμοσθεί σ' αυτό το πρόβλημα δίνει προσεγγίσεις που τείνουν στο μηδέν όταν $n \to +\infty$.

Γενικότερα: Έστω το πρόβλημα

$$\begin{cases} y'(t) = \lambda y(t) \\ y(0) = 1 \end{cases}, \ \lambda \in \mathbb{C}, \ \mathrm{Re}(\lambda) < 0 \qquad (2)$$

Μία αριθμητική μέθοδος επίλυσης προβλήματος αρχικών τιμών καλείται **απόλυτα ευσταθής** για κάποιο $h > 0$ αν όταν εφαρμοσθεί στο πρόβλημα (2) δίνει προσεγγίσεις για τις οποίες $|y_n| \to 0$. Η περιοχή του μιγαδικού ημιεπιπέδου $S = \{z \in \mathbb{C} : \mathrm{Re}(z) < 0\}$ ώστε η μέθοδος να είναι απόλυτα ευσταθής αν $h\lambda \in S$ καλείται περιοχή απόλυτης ευστάθειας της μεθόδου.

8.2. Μέθοδος Runge-Kutta

Η μέθοδος Eüler είναι ειδική περίπτωση της κατηγορίας Runge-Kutta. Μία μέθοδος Runge-Kutta με *q* ενδιάμεσα στάδια, υπολογίζει το y_{n+1} από το y_n μέσω ενός τύπου:

$$y_{n+1} = y_n + h\sum_{j=1}^{q} b_j f\left(t_{n,j}, y_{n,j}\right)$$

όπου οι b_j $(j = 1, ..., q)$ είναι δοσμένες σταθερές. Οι q ενδιάμεσες τιμές $y_{n,j}$ και $t_{n,j}$ δίνονται από τις σχέσεις

$$y_{n,i} = y_n + h\sum_{j=1}^{q} \alpha_{ij} f\left(t_{n,j}, y_{n,j}\right) \quad i = 1,...,q$$

όπου $t_{n,j} = t_n + T_j h$, T_j, a_{ij} δεδομένες σταθερές. Έτσι μία μέθοδος Runge-Kutta με q στάδια ορίζεται από $q^2 + 2q$ σταθερές γραμμένες ως εξής:

$$\begin{array}{ccc|c} \alpha_{11} & & \alpha_{1q} & T_1 \\ \vdots & & & \vdots \end{array}$$

α_{q1}		α_{qq}	T_q
b_1	…	b_q	

Η κλασσική μέθοδος Runge-Kutta ορίζεται από τις κάτωθι σταθερές:

0	0	0	0	0
$\frac{1}{2}$	0	0	0	$\frac{1}{2}$
0	$\frac{1}{2}$	0	0	$\frac{1}{2}$
0	0	1	0	1
$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$	

Έχουμε λοιπόν:

$$y_{n,i} = y_n + \sum_{i=1}^{q} \alpha_{ij} f\left(t_{n,j}, y_{n,j}\right) \quad i = 1,.....,q$$

$$\Rightarrow \begin{bmatrix} y_{n,1} \\ y_{n,2} \\ y_{n,3} \\ y_{n,4} \end{bmatrix} = \begin{bmatrix} y_n \\ y_n \\ y_n \\ y_n \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} f(t_n, y_{n,1}) \\ f\left(t_n + \frac{h}{2}, y_{n,2}\right) \\ f\left(t_n + \frac{h}{2}, y_{n,3}\right) \\ f\left(t_n + h, y_{n,4}\right) \end{bmatrix},$$

άρα:

$$y_{n+1} = y_n + \sum_{i=1}^{4} b_i \ f\left(t_{n,i}, y_{n,i}\right)$$

$$= y_n + \frac{1}{6} f\left(t_{n,1}, y_{n,1}\right) + \frac{1}{3} f\left(t_{n,2}, y_{n,2}\right) + \frac{1}{3} f\left(t_{n,3}, y_{n,3}\right) + \frac{1}{6} f\left(t_{n,4}, y_{n,4}\right)$$

$$= y_n + \frac{1}{6} f\left(t_n, y_{n,1}\right) + \frac{1}{3} f\left(t_n + \frac{h}{z}, y_{n,2}\right) + \frac{1}{3} f\left(t_n + \frac{h}{a}, y_{n,3}\right) + \frac{1}{6} f\left(t_n + h, y_{n,4}\right).$$