Πρόλογος

Οι παρούσες σημειώσεις αποτελούν το μεγαλύτερο μέρος του υλικού που διδάχτηκε στις παραδόσεις του προπτυχιακού μαθήματος της Αριθμητικής Ανάλυσης, το εαρινό εξάμηνο 2007-2008, στο Μαθηματικό τμήμα του Πανεπιστημίου Αιγαίου.

Οι σημειώσεις αυτές γράφτηκαν για την περαιτέρω διευκόλυνση των φοιτητών/τριών στα πλαίσια του «Προγράμματος Αναμόρφωσης Σπουδών» (ΕΠΕΑΕΚ-ΙΙ) και σε καμιά περίπτωση δεν αποτελούν ένα ολοκληρωμένο σύγγραμμα εισαγωγής στην Αριθμητική Ανάλυση.

Πρόκειται για την πρώτη τους έκδοση και επομένως υπάρχουν ελλείψεις, κυρίως σε παραδείγματα και σε αποδείξεις θεωρημάτων, ενώ απαιτούνται αρκετές βελτιώσεις. Σε σχέση με το υλικό που διδάχτηκε στο μάθημα οι περισσότερες ελλείψεις παρουσιάζονται στο 3ο κεφάλαιο, δηλαδή στην επίλυση των συστημάτων γραμμικών εξισώσεων. Επομένως, οι σημειώσεις αυτές πρέπει να χρησιμοποιηθούν μόνο συμπληρωματικά με το επίσημο σύγγραμμα του μαθήματος, δηλαδή την «Εισαγωγή στην Αριθμητική Ανάλυση» των Γ.Δ. Ακρίβη και Β.Α. Δουγαλή, 5η αναθεωρημένη έκδοση, Πανεπιστημιακές Εκδόσεις Κρήτης, 2006.

Η νέα έκδοση των παρόντων σημειώσεων θα ολοκληρωθεί κατά το εαρινό εξάμηνο του ακαδημαϊκού έτους 2008-2009.

Αύγουστος 2008

Κώστας Χουσιάδας

Πίνακας περιεχομένων

Κεφάλαιο 0
Εισαγωγή

Εφαρμοσμένα μαθηματικά: ένας τεράστιος και χαοτικός τομέας των μαθηματικών ο οποίος ασχολείται με τις μαθηματικές τεχνικές που αναπτύσσονται και χρησιμοποιούνται στις άλλες επιστήμες, στις εφαρμογές και την τεχνολογία.

Τι είναι η *Αριθμητική Ανάλυση*: Ίσως ο βασικότερος κλάδος των εφαρμοσμένων μαθηματικών. Η αριθμητική ανάλυση είναι σχεδόν συνώνυμη με τα υπολογιστικά μαθηματικά.

Στόχος: Η προσεγγιστική επίλυση μαθηματικών προβλημάτων που συναντώνται σε όλες τις επιστήμες και την τεχνολογία. Συνήθως έχουμε μαθηματικά μοντέλα τα οποία περιγράφουν διάφορα φαινόμενα ή/και διεργασίες τα οποία εμπλέκουν συνεχείς συναρτήσεις και μεταβλητές. Επειδή η αναλυτική επίλυση είναι σπάνια δυνατή, επιλύουμε το πρόβλημα προσεγγιστικά αφού πρώτα το διακριτοποιήσουμε. Έτσι:
από συνεχείς διαδικασίες $\rightarrow$ σε διακριτές διαδικασίες (τονίζεται ότι ο Η/Υ μπορεί να χειρισθεί μόνο νούμερα)
άπειρες διαδικασίες $\rightarrow$ πεπερασμένες διαδικασίες (οι πρώτες απαιτούν άπειρο χρόνο για να διεκπεραιωθούν)

Το διακριτό πρόβλημα που προκύπτει το ονομάζουμε αριθμητική μέθοδο. Κάθε διακριτό πρόβλημα (ή αριθμητική μέθοδος) για να εφαρμοσθεί (κυρίως στον ηλεκτρονικό υπολογιστή) απαιτεί μια πεπερασμένη, λογική σειρά καλά ορισμένων αριθμητικών πράξεων και λογικών εκφράσεων. Το σύνολο αυτών των βημάτων ονομάζεται αλγόριθμος.

Η αριθμητική ανάλυση χωρίζεται με δύο μέρη:

I. Θεωρητικό μέρος: Κατασκευή αλγορίθμων και μελέτης της ακρίβειάς του και της ευστάθειάς του, δηλαδή ανάλυση των σφαλμάτων τους.
II. Πρακτικό μέρος: Υλοποίηση των αλγορίθμων με τον βέλτιστο τρόπο ή με έναν τρόπο σχεδόν βέλτιστο (σε σχέση με την *ταχύτητα εκτέλεσης* του υπολογιστή και την *απαιτούμενη μνήμη*)

Συνεπώς η διαδικασία επίλυσης ενός μαθηματικού προβλήματος αριθμητικά έχει ως εξής:

Κατασκευάζουμε το μαθηματικό πρόβλημα το οποίο περιγράφεται με συνεχείς συναρτήσεις

↓

(Θεωρία) Κατασκευάζουμε το αντίστοιχο μαθηματικό πρόβλημα το οποίο περιγράφεται με διακριτές συναρτήσεις (αριθμητική μέθοδος) και το οποίο προσεγγίζει το αρχικό πρόβλημα

↓

(Θεωρία) Μελέτη της ακρίβειας και της ευστάθειας

↓

(Πράξη) Κατασκευή αλγορίθμου

↓

(Πράξη) Υλοποίηση αλγόριθμου (με βέλτιστο τρόπο).

<u>*Παράδειγμα*</u>: $I = \int_a^b f(x)dx$ με $f : [a,b] \to \mathbb{R}$

- Ανάπτυξη αριθμητικής μεθόδου-διακριτού σχήματος.

  Η μέθοδος του ορθογωνίου:

  $$I \approx \frac{(b-a)}{n} \sum_{k=0}^{n} f\left(x = a + k\frac{b-a}{n}\right)$$

  Η μέθοδος τραπεζίου:

  $$I \approx \frac{(b-a)}{n}\left[f(x=a) + 2\sum_{k=1}^{n-1} f\left(x = a + k\frac{b-a}{n}\right) + f(x=b)\right]$$

  όπου 'n' ο αριθμός των υποδιαστημάτων (θετικός ακέραιος αριθμός).
- Θεωρητική μελέτη:
  - Πόσο ακριβείς είναι οι παραπάνω μέθοδοι?
  - Είναι ευσταθείς? (η έννοια της ευστάθειας θα εξηγηθεί παρακάτω)
- Πρακτική εφαρμογή:
  - Ποιοι οι αντίστοιχοι αλγόριθμοι?
  - Πως αυτοί οι αλγόριθμοι υλοποιούνται?

Σχόλια πάνω στη θεωρία / πράξη:

I. Θεωρητικά μπορεί μια μέθοδος να είναι ακριβής/ευσταθής, πρακτικά όμως να είναι μη-υλοποιήσιμη.
II. Πρακτικά ένας αλγόριθμος μπορεί να δίνει αποτελέσματα, αλλά χωρίς θεωρητική μελέτη δεν ξέρουμε κατά πόσο μπορούμε να τα εμπιστευτούμε ή όχι.
III. Οι αριθμητικές μέθοδοι για την επίλυση ενός προβλήματος μπορεί να είναι πολλές. Με βάση τα θεωρητικά και πρακτικά χαρακτηριστικά επιλέγουμε κάθε φορά ποια από τις διαθέσιμες θα εφαρμόσουμε.

Μιλήσαμε για προσεγγιστική επίλυση ενός προβλήματος που σημαίνει ότι τα αποτελέσματά μας θα περιέχουν κάποιο σφάλμα σε σχέση με την ακριβή τους τιμή. Για να μετρήσουμε αυτό το σφάλμα, αλλά και άλλους λόγους, χρησιμοποιούμε δύο ποσότητες:

(a) Το απόλυτο σφάλμα: $\boxed{E = \left| x - x_{\pi\rho} \right|}$

(b) Το σχετικό σφάλμα: $\boxed{RE = \dfrac{\left| x - x_{\pi\rho} \right|}{\left| x \right|}, \quad x \neq 0}$

όπου x η πραγματική τιμή του μεγέθους που μας ενδιαφέρει και $x_{\pi\rho}$ η χρησιμοποιούμενη προσεγγιστική του τιμή. Το σχετικό σφάλμα δίνεται συνήθως και ως ποσοστό επί τις εκατό, δηλαδή: $\boxed{RE = 100 \dfrac{\left| x - x_{\pi\rho} \right|}{\left| x \right|} \%, \quad x \neq 0}$

Να σημειωθεί ότι οι παραπάνω ορισμοί μπορούν να συναντηθούν στην βιβλιογραφία χωρίς τις απόλυτες τιμές. Στο σημείο αυτό θέτουμε το παρακάτω ερώτημα: ποια ποσότητα αντιπροσωπεύει καλύτερα την προσέγγισή μας και γιατί? Η μελέτη μερικών παραδειγμάτων θα μας βοηθήσει να απαντήσουμε.

Παράδειγμα I: Έστω $x = 3.1$ και $x_{\pi\rho} = 3.0$. Τότε $E = \left| 3.1 - 3.0 \right| = 0.1$, $RE = \dfrac{\left| 3.1 - 3.0 \right|}{\left| 3.1 \right|} = 0.032 = 3.2\%$.

Παράδειγμα II: Έστω πληθυσμός $x = 10100$ και $x_{\pi\rho} = 10000$. Τότε $E = 100$ και $RE = \dfrac{100}{10100} \simeq 0.01 \simeq 1\%$

Παράδειγμα III: Έστω ποσότητα φαρμάκου που πρέπει να χορηγηθεί σε έναν ασθενή $x = 0.01\,\text{gr}$ και $x_{\pi\rho} = 0.015$ gr η ποσότητα που πραγματικά χορηγείται. Τότε $E = 0.005$ και $RE = \dfrac{0.005}{0.01} = 0.5 = 50\%$.

Από τα παραπάνω είναι φανερό ότι το RE είναι καλύτερος δείκτης ακρίβειας, σε σχέση με το E για την εκτίμηση μίας προσέγγισης.

0.1. Κατηγορίες σφαλμάτων ή λαθών

Διακρίνουμε δύο μεγάλες κατηγορίες σφαλμάτων:

I. Λάθη λόγω μαθηματικού φορμαλισμού
   - μη κατάλληλο σύστημα εξισώσεων
   - ανακρίβειες στις τιμές παραμέτρων του προβλήματος (π.χ. $g = 9.81$ η σταθερά βαρύτητας) ή λάθη στα αρχικά δεδομένα.

II. Λάθη κατά την αριθμητική επίλυση
   - λάθη λόγω προσέγγισης των αριθμών (round-off error)

     π.χ. $\pi = 3.14159\ldots$, $\frac{1}{3} = 0.333\ldots$, δηλαδή όταν αγνοούμε πολλά από τα ψηφία των αριθμών
   - λάθη αποκοπής (truncation error)

     π.χ. $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \underbrace{\cdots}_{\substack{\text{σφαλμα} \\ \text{αποκοπης}}}$, δηλαδή όταν αντικαθιστούμε απειροσειρές με πεπερασμένες σειρές.

Στόχος στο πρώτο κεφάλαιο αυτών των σημειώσεων είναι η μελέτη των λαθών λόγω προσέγγισης των αριθμών, ενώ στα επόμενα κεφάλαια εξετάζονται κυρίως τα λάθη αποκοπής και η επίδρασή τους στα αποτελέσματα των αλγορίθμων.

0.2. Προσέγγιση αριθμών με αποκοπή και στρογγυλοποίηση. Σημαντικά ψηφία.

Έστω ότι θέλουμε να κάνουμε πράξεις με αριθμούς που έχουν είτε άπειρα ψηφία (π.χ. $\sqrt{2}$, π κ.τ.λ.) ή τόσα πολλά που πρακτικά είναι αδύνατο να τις πραγματοποιήσουμε. Τότε χρησιμοποιούμε προσεγγίσεις αυτού του αριθμού σε «k» σημαντικά ψηφία. Λέμε ότι *ένας αριθμός $x_{\pi\rho}$ προσεγγίζει την ακριβή τιμή του αριθμού x με k σωστά σημαντικά ψηφία όταν k είναι ο μεγαλύτερος ακέραιος για τον οποίο ισχύει:*

$$\frac{|x - x_{\pi\rho}|}{|x|} \le 0.5 \times 10^{-k+1}, \quad x \neq 0 \qquad (*)$$

Η προσέγγιση $x_{\pi\rho}$ προκύπτει με δύο διαδικασίες:

I. Αποκοπή: Ξεκινάμε από το πιο αριστερό μη-μηδενικό ψηφίο και μετράμε «k» ψηφία αγνοώντας τα υπόλοιπα.

II. Στρογγυλοποίηση: Παρατηρούμε το «$k+1$» ψηφίο του αριθμού. Αν είναι ≥ 5, τότε αυξάνουμε το «k» τελευταίο ψηφίο κατά 1 και αγνοούμε τα υπόλοιπα.

Παρατήρηση: η προσέγγιση ενός αριθμού γίνεται πιο εύκολα αν φέρουμε τον αριθμό στην κανονική μορφή κινητής υποδιαστολής (βλέπε παρακάτω).

Παράδειγμα: Έστω $\pi = 3.14159265...$ και ότι ζητείται η προσέγγισή του με 5 σημαντικά ψηφία με αποκοπή και στρογγυλοποίηση. Άρα,

- Αποκοπή: $\pi = \underline{3.1415}9265 \Rightarrow \pi_{\alpha\pi.}^{(5)} = 3.1415$

- Στρογγυλοποίηση: $\pi = \underline{3.1415}\,\underset{,6^{o},}{9}\,265 \Rightarrow \pi_{\sigma\tau\rho.}^{(5)} = 3.1416$

-Το απόλυτο και το σχετικό σφάλμα στην αποκοπή είναι $E_{\alpha\pi.} = 0.926536 \times 10^{-4}$ και $RE_{\alpha\pi.} = 0.294914 \times 10^{-4}$, αντίστοιχα. Επομένως, από την (*), προκύπτει $k = 5$.

-Ομοίως, το απόλυτο και το σχετικό σφάλμα στην στρογγυλοποίηση είναι $E_{\sigma\tau\rho.} = 0.734641 \times 10^{-5}$ και $RE_{\sigma\tau\rho.} = 0.233958 \times 10^{-5}$ αντίστοιχα. Επομένως, από την (*), προκύπτει $k = 6$.

Παρατήρηση: γενικά ισχύει ότι το απόλυτο σφάλμα στην στρογγυλοποίηση είναι μικρότερο ή ίσο του απολύτου σφάλματος στην αποκοπή, $E_{\sigma\tau\rho.} \le E_{\alpha\pi.}$ και το ίδιο ισχύει και για το σχετικό σφάλμα, $RE_{\sigma\tau\rho.} \le RE_{\alpha\pi.}$.

Κεφάλαιο 1ο
Αριθμητική του ηλεκτρονικού υπολογιστή

1.1. Αναπαράσταση αριθμών ως προς οποιαδήποτε βάση

Στην καθημερινή μας ζωή χρησιμοποιούμε το 10-δικό σύστημα αναπαράστασης αριθμών. Η βάση είναι το «10» και τα ψηφία τα 0,1,2,...,9. Κάθε αριθμός όμως μπορεί, χωρίς καμία δυσκολία, ως προς οποιαδήποτε βάση. Αυστηρά γράφουμε:

$$\underbrace{\pm}_{\text{προσημο}}\left(\underbrace{a_N a_{N-1} a_{N-2}\ldots a_1 a_0}_{\text{ακεραιο μερος}}.\underbrace{a_{-1}a_{-2}a_3\ldots}_{\text{κλασματικο μερος}}\right)_\beta \quad \text{όπου } 0\le a_i\le \beta-1,\ i=-\infty,\ldots,N$$

Ο κάθε αριθμός έχει δύο μέρη:

- το ακέραιο μέρος, πριν την υποδιαστολή
- το κλασματικό μέρος, μετά την υποδιαστολή

Το κάθε ένα από αυτά μπορεί να γραφεί σε μορφή σειράς

$$\text{Ακεραιο μερος}=\sum_{k=0}^{N} a_k\beta^k = a_0+a_1\beta+a_2\beta^2+\ldots+a_N\beta^N$$

$$\text{Κλασματικο μερος}=\sum_{k=1}^{\infty} a_{-k}\beta^{-k} = a_{-1}\frac{1}{\beta}+a_{-2}\frac{1}{\beta^2}+a_{-3}\frac{1}{\beta^3}+\ldots$$

Επομένως, ο αριθμός μπορεί να γραφεί ως εξής:

$$\pm\left(a_N a_{N-1}\ldots a_1 a_0.a_{-1}a_{-2}a_3\ldots\right)=\pm\sum_{k=-\infty}^{N} a_k\beta^k \quad \text{όπου } 0\le a_k\le\beta-1,\ k=-\infty,\ldots N$$

Συνήθως ισχύει $2\le\beta\le16$ (αν και $\beta>16$ είναι εφικτό, αλλά δεν προσφέρει κάποια πλεονεκτήματα). Στους υπολογιστές $\beta=2,8,16$.

I. Μετατροπή ακεραίου από βάση β σε βάση 10

$$\left(a_n a_{n-1}\ldots a_1 a_0\right)_\beta=\sum_{i=0}^{n} a_i\beta^i$$

- Άμεσος τρόπος
- Σχήμα Horner

II. Μετατροπή κλασματικού x αριθμού $(0 < x < 1)$ από βάση β σε βάση 10

$$x = \left(0.a_{-1}a_{-2}a_{-3}\ldots a_{-n}\right)_\beta = \sum_{k=1}^{n} a_{-k}\beta^{-k}$$

III. Μετατροπή ακεραίου από βάση 10 σε βάση β σύμφωνα με τον αλγόριθμο της διαίρεσης (δείτε και σχήμα Horner)

IV. Μετατροπή κλασματικού x από βάση 10 σε βάση β

Παρατηρήσεις σχετικά με την μετατροπή αριθμών από το ένα σύστημα αριθμών σε ένα άλλο.

1η Παρατήρηση:

Ο ακέραιος παραμένει πάντα ακέραιος σε οποιοδήποτε σύστημα και αν τον εκφράσουμε. Η μετατροπή ενός αριθμού σε βάση $\beta \to$ βάση 10 γίνεται χωρίς καμία δυσκολία.

Παράδειγμα: Ο αριθμός

$$\left(53473\right)_8 = 3 + 7\cdot 8^1 + 4\cdot 8^2 + 3\cdot 8^3 + 5\cdot 8^4 = \left(22331\right)_{10}$$

Μετατροπή στον υπολογιστή

Ο υπολογισμός της ποσότητας $A(\beta) = \sum_{k=0}^{N} a_k \beta^k$, μπορεί να γίνει με δύο τρόπους

a. Άμεσος τρόπος:

$$A(\beta) = a_0 + a_1\beta + a_2\beta^2 + \cdots + a_N\beta^N$$

έχουμε N όρους

- για τον υπολογισμό του ο $N-$οστός όρος απαιτεί N πολλαπλασιασμούς
- ο $N-1$ όρος απαιτεί $N-1$ πολλαπλασιασμούς

$\vdots$

- ο 2ος όρος απαιτεί 1 πολλαπλασιασμό
- ο 1ος όρος απαιτεί 0 πολλαπλασιασμούς

Άρα

$$N+(N-1)+(N-2)+\cdots+1=\sum_{k=1}^{N}k=\frac{N(N+1)}{2}$$

επιπλέον έχουμε $N-1$ προσθέσεις και επομένως το σύνολο των πράξεων είναι

$$\underbrace{\frac{N(N+1)}{2}}_{\substack{\text{πιο σημαντικος ορος}\\ \text{οι πολλαπλασιασμοι}}}+\underbrace{N}_{\substack{\text{προσθεσεις}\\ \text{λιγοτερο}\\ \text{σημαντικος}}}=\frac{N^2}{2}+\frac{N}{2}+N=\frac{N^2}{2}+\frac{3N}{2}$$

b. <u>Σχήμα Horner</u>:

$$A(\beta)=\underbrace{a_0+\beta\left(a_1+\beta\left(a_2+\cdots+\beta\left(a_{N-1}+a_N\beta\right)\right)\cdots\right)}_{\text{συνολο πολλαπλασιασμων: }N}$$

Επομένως ο τρόπος (a) είναι $O\left(N^2\right)$, ενώ ο (b) είναι $O(N)$ και άρα το σχήμα Horner είναι πολύ πιο γρήγορο από ότι ο άμεσος τρόπος.

<u>Επεξήγηση του όρου</u> $O(\bullet)$

<u>Ερώτηση</u>: έχει πραγματικά αξία το γεγονός ότι η μία μέθοδος είναι τάξης $O(N)$ και η άλλη $O\left(N^2\right)$?

<u>Αλγόριθμος</u>:

$p \leftarrow a_N$

για $i=N-1,0,-1$

$p \leftarrow a_i + p\times\beta$ (a)

<u>Κώδικας Fortran</u>:

$p = a(N)$
Do $i = N-1,0,-1$
$p = a(i) + p * \beta$
End do

(b)

(a), (b) $\rightarrow$ 1 flop (floating point operation)

1flop: η πιο συχνή πράξη που χρησιμοποιούμε/συναντούμε στα υπολογιστικά μαθηματικά και στους υπολογιστές. Για τον λόγο αυτό το flop έχει καθιερωθεί ως μονάδα μέτρησης των πράξεων στους αλγορίθμους. Η ταχύτητα ενός επεξεργαστή στους Η/Υ αλλά και των μεγάλων υπερυπολογιστικών συστημάτων μετράται σε αριθμό flops/μονάδα χρόνου.

2η Παρατήρηση:

Η μετατροπή ενός κλασματικού αριθμού, $0 < x < 1$, από σύστημα με βάση το β σε σύστημα με βάση το 10 δεν παρουσιάζει επίσης καμία δυσκολία. Επίσης, ένας κλασματικός παραμένει πάντα κλασματικός σε όποιο σύστημα και αν εκφραστεί, όμως το πλήθος ψηφίων μπορεί από πεπερασμένο να γίνει άπειρο.

Παράδειγμα: $(0.11)_2 = 1 \cdot \frac{1}{2^1} + 1 \cdot \frac{1}{2^2} = (0.5)_{10} + (0.25)_{10} = (0.75)_{10}$

3η Παρατήρηση:

Η μετατροπή ενός ακεραίου με βάση το 10 σε ακέραιο με βάση το β γίνεται σύμφωνα με τον αλγόριθμο της διαίρεσης. Πρώτα εκφράζουμε τον αριθμό στο νέο σύστημα:

$$x = (a_n \ldots a_2 a_1 a_0)_\beta = \sum_{i=0}^{n} a_i \beta^i = a_n \beta^n + a_{n-1} \beta^{n-1} + \ldots + a_1 \beta^1 + a_0$$

Στην συνέχεια τον εκφράζουμε σύμφωνα με το σχήμα Horner:

$$x = \beta \ldots \left(\beta \left(\beta \left(a_n \beta + a_{n-1} \right) + a_{n-2} \right) + a_{n-3} \right) + \ldots + a_0$$

Αν διαιρέσουμε τον παραπάνω αριθμό με «β», το υπόλοιπο είναι το a_0. Αν τον νέο αριθμό που προκύψει τον διαιρέσουμε με «β», το υπόλοιπο είναι το a_1. Συνεχίζουμε μέχρι να φτάσουμε σε ένα αποτέλεσμα διαίρεσης το οποίο να είναι αριθμός μικρότερος του «β», ο οποίος και είναι το ψηφίο a_n.

Παράδειγμα: θέλουμε να εκφράσουμε τον αριθμό 209_{10} στο σύστημα με β=4. Ακολουθούμε τα παρακάτω βήματα.

(1) Διαιρούμε τον αριθμό με το 4. Το ακέραιο μέρος του αποτελέσματος είναι 52 και το υπόλοιπο είναι 1, επομένως $a_0 = 1$.

(2) Διαιρούμε το 52 με το 4 οπότε έχουμε ως αποτέλεσμα το 13 και υπόλοιπο 0, επομένως $a_1 = 0$.

(3) Διαιρούμε το 13 με το 4 οπότε έχουμε ως αποτέλεσμα το 3 και υπόλοιπο 1, επομένως $a_2 = 1$. Επιπλέον, εφόσον το αποτέλεσμα της διαίρεσης είναι 3 < β = 4, άρα αυτό είναι το τελευταίο ψηφίο του αριθμού στο νέο σύστημα, δηλαδή $a_3 = 3$.

Συνολικά έχουμε: $209_{10} = \left(a_3 a_2 a_1 a_0\right)_4 = 3101_4$

<u>4η Παρατήρηση</u>:

Η μετατροπή ενός κλασματικού αριθμού x, $0 < x < 1$, από σύστημα με βάση το 10 σε σύστημα με βάση β, $x_\beta = \left(.a_{-1}a_{-2}...a_{-n}\right)_\beta$, βασίζεται στην εξής διαδικασία: Γράφουμε τον αριθμό σύμφωνα με τον ορισμό του, στο νέο σύστημα:

$$x = \left(.a_{-1}a_{-2}...a_{-n}\right)_\beta = \sum_{i=1}^{n} a_{-i}\beta^{-i} = \frac{a_{-1}}{\beta} + \frac{a_{-2}}{\beta^2} + \frac{a_{-3}}{\beta^3} + ... + \frac{a_{-n}}{\beta^n}$$

Πολλαπλασιάζοντας και τα δύο μέλη με «β», έχουμε:

$\beta x = \left(a_{-1}.a_{-2}...a_{-n}\right)_\beta = a_{-1} + \frac{a_{-2}}{\beta^1} + \frac{a_{-3}}{\beta^2} + ... + \frac{a_{-n}}{\beta^{n-1}}$ όπου είναι προφανές ότι το a_{-1} είναι το ακέραιο μέρος του βx και $\left(.a_{-2}...a_{-n}\right)_\beta$ το κλασματικό μέρος του. Πολλαπλασιάζουμε διαδοχικά με «β» το κλασματικό μέρος του αποτελέσματος και κάθε φορά, προσδιορίζουμε το επόμενο ψηφίο, σύμφωνα με το ορισμό που δίνεται παραπάνω. Έτσι αν έχουμε πολλαπλασιάσει «n» φορές με «β» τότε θα έχουμε:

$\beta^n x = \left(a_{-1}a_{-2}...a_{-n+1}a_{-n}.\right)_\beta = a_{-1}\beta^{n-1} + a_{-2}\beta^{n-2} + ... + a_{-n+1}\beta^1 + a_{-n}$ οπότε τελικά το κλασματικό μέρος του αποτελέσματος είναι μηδέν. Τότε η διαδικασία τερματίζεται.

Παράδειγμα: θέλουμε να εκφράσουμε τον κλασματικό αριθμό 0.90625_{10} στο σύστημα με β=2, δηλαδή στο δυαδικό σύστημα. Ακολουθούμε τα παρακάτω βήματα:

(1) Πολλαπλασιάζουμε τον αριθμό με το 2. Το αποτέλεσμα είναι 1.8125, οπότε το ακέραιο μέρος είναι 1 και το κλασματικό 0.8125, επομένως $a_{-1} = 1$

(2) Πολλαπλασιάζουμε τον κλασματικό αριθμό 0.8125 με το 2. Το αποτέλεσμα είναι 1.625, οπότε το ακέραιο μέρος του αποτελέσματος είναι 1 και το κλασματικό 0.625, επομένως $a_{-2} = 1$.

(3) Πολλαπλασιάζουμε τον κλασματικό αριθμό 0.625 με το 2. Το αποτέλεσμα είναι 1.25, οπότε το ακέραιο μέρος του αποτελέσματος είναι 1 και το κλασματικό είναι 0.25, επομένως $a_{-3} = 1$.

(4) Πολλαπλασιάζουμε τον κλασματικό αριθμό 0.25 με το 2. Το αποτέλεσμα είναι 0.5, οπότε το ακέραιο μέρος του αποτελέσματος είναι 0 και το κλασματικό είναι 0.5, επομένως $a_{-4} = 0$.

(5) Πολλαπλασιάζουμε τον κλασματικό αριθμό 0.5 με το 2. Το αποτέλεσμα είναι 1.0, οπότε το ακέραιο μέρος του αποτελέσματος είναι 1 και το κλασματικό είναι 0, επομένως $a_{-5} = 1$.

Συνολικά έχουμε: $0.90625_{10} = \left(.a_{-1}a_{-2}a_{-3}...a_{-n}\right)_2 = .11101_2$. Εδώ θα πρέπει να σημειωθεί ότι υπάρχει η πιθανότητα η σειρά των δυαδικών ψηφίων a_{-i} να μην τερματίζεται, δηλαδή να μην καταλήγουμε ποτέ σε μηδενικό κλασματικό μέρος. Σε αυτήν την περίπτωση η διαδικασία διακόπτεται όταν υπολογίσουμε όλα τα bits που είναι διαθέσιμα για την συγκεκριμένη μεταβλητή. Τότε ο αντίστοιχος πραγματικός αριθμός θα είναι αποθηκευμένος στον υπολογιστή με κάποιο σφάλμα στρογγύλευσης, κάτι βέβαια που δεν συνέβη με τον αριθμό 0.90625_{10}, στο παραπάνω παράδειγμα.

1.2. Αναπαράσταση των αριθμών στον υπολογιστή

Επειδή η μνήμη του υπολογιστή αποτελείται από εκατομμύρια διακόπτες οι οποίοι μπορούν (ο καθένας από αυτούς) να είναι σε δύο μόνο καταστάσεις, «κλειστός» ή «ανοιχτός», δηλαδή, σε «0» ή «1» κατάσταση και οι οποίοι ονομάζονται bits για αυτό ο κατάλληλος τρόπος αντιπροσώπευσης των αριθμών στον υπολογιστή είναι με βάση το 2, $\beta = 2$. Επίσης ισχύει: 1 byte = 8 bits, 1 word = 1, 2, ή 4 bytes.

Αναπαράσταση ενός ακεραίου με χρήση 1 byte:

Ελάχιστος αριθμός = 0

$\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{0}$

Μέγιστος αριθμός =

$$255 = \boxed{1}\boxed{1}\boxed{1}\boxed{1}\boxed{1}\boxed{1}\boxed{1}\boxed{1} = 1\cdot 2^7 + 1\cdot 2^6 + 1\cdot 2^5 + 1\cdot 2^4 + 1\cdot 2^3 + 1\cdot 2^2 + 1\cdot 2^1 + 1\cdot 2^0$$

Εφόσον σε κάθε θέση έχουμε μόνο δύο δυνατότητες, αν έχω γενικά «n» θέσεις διαθέσιμες τότε προκύπτουν:

$$\underbrace{2\cdot 2\cdot 2\cdot 2\cdot \ldots \cdot 2}_{n} = 2^n \text{ αριθμοί}$$

Επομένως ο ελάχιστος αριθμός είναι ο 0 και μέγιστος ο $2^n - 1$. Άρα με 1 byte μπορούμε να αναπαραστήσουμε συνολικά $2^n = 2^8 = 256$ ακεραίους. Λόγω ότι πρέπει να λάβουμε υπόψη θετικούς και αρνητικούς αριθμούς, $k \in \left[-2^{n-1} - 1, 2^{n-1}\right]$.

Διάφοροι τρόποι έχουν προταθεί για την αναπαράσταση αριθμών στον υπολογιστή. Σήμερα όλοι οι υπολογιστές αναπαριστούν τους πραγματικούς αριθμούς με τους λεγόμενους αριθμούς κινητής υποδιαστολής (floating point numbers), δηλαδή με την μορφή:

$$x = \pm \underbrace{\left(.d_1 d_2 d_3 \ldots\right)}_{mantissa} \times \beta^{e \,\rightarrow\, exponent}$$

όπου

- $\beta -$ η βάση του συστήματος, η οποία είναι πάντα ακέραια (συνήθως $\beta = 2$)
- $e -$ εκθέτης ο οποίος είναι επίσης ακέραιος

$0 \le d_i \le \beta - 1$, $i = 1,2,3,\ldots$

Όταν $d_1 \neq 0$, η μορφή λέγεται <u>κανονικοποιημένη</u> και η αναπαράσταση κάθε αριθμού με τον τρόπο αυτό είναι μοναδική.

<u>Παράδειγμα 1</u>: $-(0.00598)_{10} = -0.598 \times 10^{-2}$

<u>Παράδειγμα 2</u>: $(111.001)_2 = 0.111001 \times 2^3$

<u>Παράδειγμα 3</u>: $(120.005)_{10} = 0.120005 \times 10^3$

Στο σημείο αυτό να τονισθεί ότι η προσέγγιση αριθμών σε «k» ψηφία, είτε με αποκοπή είτε με στρογγυλοποίηση, είναι ιδιαίτερα εύκολη αν ο αριθμός γραφεί στην κανονικοποιημένη μορφή κινητής υποδιαστολής. Απλά πρέπει να κρατήσουμε τα πρώτα «k» ψηφία του κλασματικού μέρους του αριθμού, να αγνοήσουμε τα υπόλοιπα ψηφία του και να αφήσουμε το εκθετικό μέρος του αριθμού ανέπαφο.

<u>Παράδειγμα 1</u>:

Ο αριθμός $\pi = 3.14159265$ σε κανονικοποιημένη μορφή κινητής υποδιαστολής γίνεται ως εξής: $\pi = 0.314159265 \times 10^1$. Η προσέγγιση σε 5 σημαντικά ψηφία με αποκοπή και στρογγυλοποίηση αντίστοιχα είναι $\pi_{αποκ} = 0.31415 \times 10^1$ και $\pi_{στρ} = 0.31416 \times 10^1$.

<u>Παράδειγμα 2</u>:

Ο αριθμός $x = 1329.1689$ σε κανονικοποιημένη μορφή κινητής υποδιαστολής γίνεται $x = 0.13291689 \times 10^4$. Η προσέγγιση σε 5 σημαντικά ψηφία με αποκοπή και

στρογγυλοποίηση αντίστοιχα είναι $x_{αποκ} = 0.13291 \times 10^4 = 1329.1$ και $x_{στρ} = 0.13292 \times 10^4 = 1329.2$. Σε 3 δεκαδικά ψηφία είναι $x_{αποκ} = 0.132 \times 10^4 = 1320$ και $x_{στρ} = 0.133 \times 10^4 = 1330$.

Λόγω του γεγονότος ότι οι πραγματικοί αριθμοί μπορεί να απαιτούν άπειρα ψηφία για να αναπαρασταθούν είμαστε αναγκασμένοι στον Η/Υ να κρατάμε μόνο ένα πεπερασμένο πλήθος ψηφίων, t, δηλαδή μόνο τους ρητούς αριθμούς:

$$x = \pm \underbrace{\left(.d_1 d_2 d_3 ... d_{t-1} d_t\right)}_{mantissa} \times \beta^{e \to exponent}$$

Οι αριθμοί μηχανής ενός υπολογιστή είναι ένα σύνολο ρητών αριθμών, γραμμένων σύμφωνα με την κανονικοποιημένη μορφή κινητής υποδιαστολής. Το σύνολο αυτό χαρακτηρίζεται από 4 παραμέτρους:

a. Την βάση του αριθμητικού συστήματος, β
b. Το πλήθος, t, των ψηφίων του κλάσματος των αριθμών
c. Το κάτω φράγμα του εκθέτη, L (Lower)
d. Το άνω φράγμα του εκθέτη, U (Upper)

Όλοι οι παραπάνω παράμετροι είναι ακέραιοι. Ισχύει ότι $0 \le m < 1$ και $L \le e \le U$. Φυσικά, επειδή το **M** είναι ένα πεπερασμένο σύνολο δεν υπάρχει η δυνατότητα να αναπαρασταθούν ακριβώς όλοι οι πραγματικοί αριθμοί.

Χαρακτηριστικά του συνόλου **M**:

a. Πεπερασμένο πλήθος αριθμών $= 2\left(\beta - 1\right)\beta^{t-1}\left(U - L + 1\right) + 1$

   <u>Σχόλιο</u>: όσο αυξάνουν τα t, U, L τόσο μεγαλώνει το πλήθος των αριθμών που μπορούν να αναπαρασταθούν. <u>Ερώτηση</u>: Γιατί δεν τα αυξάνουμε?

b. Έχει ελάχιστο, κατά απόλυτη τιμή, $m \equiv min\left\{\mathbf{M}\right\}$, όπου $m = 0.\underbrace{100000}_{t-\psi\eta\phi\iota\alpha} \times \beta^L$

c. Έχει μέγιστο, κατά απόλυτη τιμή, $M \equiv max\{\mathbf{M}\}$, όπου $M = \left(1 - \frac{1}{\beta^t}\right) \times \beta^U$

d. τα στοιχεία του συνόλου <u>δεν</u> είναι ισαπέχοντα

e. Κάθε πραγματικός αριθμός x, με $m < x < M$ αναπαριστάται από την μηχανή με τον πιο κοντινό του τον οποίο συμβολίζουμε με $fl(x)$, δηλαδή με κάποιο σφάλμα.

Ισχύει ότι $\left|\frac{fl(x)-x}{x}\right| \le u$, όπου, $u = \begin{cases} \frac{1}{2}\beta^{1-t}, \text{ για στρογγυλοποιηση} \\ \beta^{1-t}, \text{ για αποκοπη} \end{cases}$

Απόδειξη για την περίπτωση της αποκοπής:

Έστω $x = 0.d_1 d_2 d_3 \ldots d_t d_{t+1} \ldots \times \beta^e$ και $fl(x) = 0.d_1 d_2 d_3 \ldots d_t \times \beta^e$, τότε έχουμε

$$\left|\frac{fl(x)-x}{x}\right| = \left|\frac{0.00\ldots 0 d_{t+1} \ldots \times \beta^e}{0.d_1 d_2 d_3 \ldots d_t d_{t+1} \ldots \times \beta^e}\right| = \left|\frac{0.d_{t+1} d_{t+2} \ldots \times \beta^{-t}}{0.d_1 d_2 d_3 \ldots}\right| =$$

$$\left|\frac{0.d_{t+1} d_{t+2} \ldots}{0.d_1 d_2 \ldots d_t d_{t+1} \ldots}\right| \times \beta^{-t} \le \frac{max\{0.d_{t+1} \ldots\}}{min\{\ldots\}} \times \beta^{-t} = \frac{1}{0.1} \times \beta^{-t} = \beta^{-t+1} \Rightarrow$$

$$\Rightarrow \left|\frac{fl(x)-x}{x}\right| \le \beta^{1-t}, \quad \text{για αποκοπη}$$

f. Δεν αποτελεί σώμα.

<u>*Παράδειγμα 1*</u>:

Έστω σύνολο αριθμών κινητής υποδιαστολής $\mathbf{M} = \mathbf{M}(\beta = 10, t = 5, U = 10, L = -10)$ και έστω οι αριθμοί $\alpha, \beta, \gamma \in \mathrm{M}$ με $\alpha = 1.0$, $\beta = 3 \times 10^{-5}$, $\gamma = 3 \times 10^{-5}$. Ζητείται το αποτέλεσμα του αθροίσματος $\alpha + \beta + \gamma$. Έχουμε, $\alpha + \beta + \gamma = (\alpha + \beta) + \gamma = \alpha + (\beta + \gamma)$.

$1^{ος}$ τρόπος, $(\alpha + \beta) + \gamma$:

Αρχικά γράφουμε τους αριθμούς σε μορφή κινητής υποδιαστολής. Δηλαδή $fl(\alpha) = 0.10000 \times 10^{+1}$, $fl(\beta) = 0.30000 \times 10^{-4}$, $fl(\gamma) = 0.30000 \times 10^{-4}$. Έχουμε

$$z = fl(fl(\alpha) + fl(\beta)) = 0.1 \times 10^{+1} = fl\left(0.\underbrace{10000}_{5 \text{ ψηφια}} 3 \times 10^{+1}\right) \Rightarrow z = 0.10000 \times 10^{+1}$$

$$fl\left(z+fl(\gamma)\right)=fl\left(0.10000\times10^{+1}+0.30000\times10^{-4}\right)=0.10000\times10^{+1}$$

2ος τρόπος, $\alpha+(\beta+\gamma)$:

$$fl\left(fl(\alpha)+fl\left(fl(\beta)+fl(\gamma)\right)\right)$$

$$fl(\beta)+fl(\gamma)=0.00003+0.00003=0.00006\Rightarrow fl\left(fl(\beta)+fl(\gamma)\right)=0.00006=0.6\times10^{-4}=z$$

$$fl(\alpha)+z=0.6\times10^{-4}+0.1\times10^{1}=1.0000+0.00006=1.00006\Rightarrow fl\left(fl(\alpha)+z\right)=0.10001\times10^{1}$$

Άρα στον υπολογιστή η πρόσθεση δεν ικανοποιεί την προσεταιριστική ιδιότητα

$(\alpha+\beta)+\gamma\neq\alpha+(\beta+\gamma)$

Επομένως *η σειρά που γίνονται οι πράξεις στον υπολογιστή έχουν σημασία.*

Παράδειγμα 2: Έστω $\beta=10$ και $t=5$. Ο αριθμός $1\in\mathbf{M}(10,5,L,U)$ διότι $1=0.10000\times10^{1}$. Ο αριθμός $10^{-5}\in\mathbf{M}(10,5,L,U)$ διότι $10^{-5}=0.10000\times10^{-4}$. Άθροισμα:

$1+10^{-5}=1.00001=0.\underbrace{100001}_{6\ \psi\eta\phi\iota\alpha}\times10^{1}\notin\mathbf{M}$ διότι έχει 6 σημαντικά ψηφία.

Παράδειγμα 3: Έστω ο αριθμός $0.10000\times\beta^{L}\in\mathbf{M}$ και το τετράγωνο αυτού:

$$\left(0.10000\times\beta^{L}\right)\times\left(0.10000\times\beta^{L}\right)=0.10000^{2}\times\beta^{2L}=0.1\times\beta^{2L-1}$$

Προφανώς $2L-1<L$ και επομένως $0.1\times\beta^{2L-1}\notin\mathbf{M}$.

Γενικά στον ηλεκτρονικό υπολογιστή απλά προσεγγίζουμε τους αριθμούς με άλλους αριθμούς πεπερασμένης ακρίβειας.

Συναντάμε δύο είδη προβλημάτων:

I. Αν $|x|>max\{\mathbf{M}\}\equiv\left(1-\frac{1}{\beta^{t}}\right)\beta^{U}\Rightarrow$ overflow error (λάθος υπερχείλισης)

II. Αν $|x|<min\{\mathbf{M}\}\equiv0.1\times\beta^{L}=\beta^{L-1}\Rightarrow$ underflow error (λάθος υπεκχείλισης)

Πιο σημαντικό είναι το overflow error γιατί έτσι οι υπολογισμοί σταματάνε και η ροή του προγράμματος διακόπτεται. Στο underflow error, αν δηλαδή $|x| < min\{\mathbf{M}\}$ τότε, συνήθως στους περισσότερους υπολογιστές, $x = 0$.

Παράδειγμα 4: Έστω οι αριθμοί $x = 5891.26$, $y = 0.0773414$. Ζητείται το αποτέλεσμα του αθροίσματος σε υπολογιστή με $\beta = 10$, $t = 5$, $U = -L = 10$ με δεδομένο ότι το $fl(\bullet)$ προκύπτει με στρογγυλοποίηση. Άρα έχουμε

Ο x σε μορφή αριθμού κινητής υποδιαστολής

$$0.\underbrace{58912}_{5\ \psi\eta\phi\iota\alpha}6\times10^4 \xrightarrow{\text{προσεγγιση με στρογγυλοποιηση}} \boxed{fl(x) = 0.58913\times10^4}$$

Ο y σε μορφή αριθμού κινητής υποδιαστολής

$$0.\underbrace{77341}_{5\ \psi\eta\phi\iota\alpha}4\times10^{-1} \xrightarrow{\text{προσεγγιση με στρογγυλοποιηση}} \boxed{fl(y) = 0.77341\times10^{-1}}$$

$fl(x) + fl(y) = 0.\underbrace{58913}77341\times10^4 \leftarrow$ πραξη ακριβης

προσέγγιση με στρογγυλοποίηση

$$\boxed{z = fl\left(fl(x) + fl(y)\right) = 0.58914\times10^4}$$

Βρίσκουμε ότι:

$x + y = 5891.3373414$, ακριβές άθροισμα

$fl(x) + fl(y) = 0.5891373414\times10^4$

$fl\left(fl(x) + fl(y)\right) = 0.58914\times10^4$, άθροισμα στον υπολογιστή

$fl(x + y) = 0.58913\times10^4$

Παρατηρούμε ότι όλα τα παραπάνω αθροίσματα είναι διαφορετικά μεταξύ τους!!!

Παράδειγμα 5: Η αυστηρά μαθηματική λύση της εξίσωσης $1 + x = 1$ είναι η $x = 0$. Όμως, έστω $x = 4\times10^{-5} \in \mathrm{M}$ και το $1 \in \mathrm{M}$. Εχουμε $fl(x) = 0.4\times10^{-4}$ και $fl(1.0) = 0.1\times10^1$. Άρα $\boxed{1 + x = 1.00004}$ και $\boxed{fl(1 + x) = fl(1.00004) = 0.1\times10^1}$.

Προφανώς κάθε $x \in \mathbb{R}$ με $0 \le x < \frac{1}{2}\beta^{1-t}$ είναι λύση της εξίσωσης $1 + x = 1$. Η ποσότητα $\frac{1}{2}\beta^{1-t}$ ονομάζεται <u>έψιλον της μηχανής</u> και είναι ο μικρότερος αριθμός ο οποίος αν προστεθεί στην μονάδα δίνει αποτέλεσμα μεγαλύτερο του 1, δηλαδή είναι ο μικρότερος αριθμός για τον οποίο ισχύει $1 + \varepsilon > 1$.

Αλγόριθμος προσεγγιστικού υπολογισμού του ε:

$\varepsilon \leftarrow 1$

εφόσον $1 + \varepsilon > 1$

$\varepsilon \leftarrow \varepsilon / 2$

Αντίστοιχος κώδικας σε fortran-90 για το ε σε μεταβλητές διπλής ακρίβειας.

```
eps = 1.d0
do
 y = 1.d0 + eps
 if( y > 1.d0 ) then
  eps = eps / 2.d0
 elseif( y <= 1.d0 ) then
 exit
endif
enddo
print*, 2.d0*eps
```

Άσκηση: βρείτε όλους τους αριθμούς του συνόλου αριθμών κινητής υποδιαστολής, $\mathbf{M} = \mathbf{M}(\beta = 2, t = 3, U = 1, L = -2)$ *αναλυτικά. Είναι οι αριθμοί ισαπέχοντες? Ποιος είναι ο μέγιστος και ποιος ο ελάχιστος αριθμός αυτού του συνόλου?*

Υπόδειξη: Η μορφή των αριθμών που ανήκουν στο σύνολο αυτό είναι $\pm(0.ddd)_2 \times 2^a, -2 \le a \le 1,\ d = 0,1$. *Υπολογίστε όλους τους θετικούς κλασματικούς αριθμούς (δηλαδή τους* $(0.ddd)_2$) και όλους τους αριθμούς της μορφής 2^a και συνδυάστε τα αποτελέσματα.

1.3. Αριθμητικές πράξεις στον υπολογιστή και επιρροή των σφαλμάτων στρογγύλευσης στους υπολογισμούς

Έστω

$$\square = +, -, \times, \div$$

και ότι ζητάμε το αποτέλεσμα της πράξης

$$x \square y$$

Έχουμε

$$fl\left(fl(x) \square fl(y)\right) = z$$

όπου $z \in \mathrm{M}$, $fl(x) \in \mathrm{M}$, $fl(y) \in \mathrm{M}$.

Παρατήρηση

Έχουμε

$$\left|\frac{fl(x) - x}{x}\right| \le u, \quad u = \begin{cases} \frac{1}{2}\beta^{1-t}, \text{ στρογγυλευση} \\ \beta^{1-t}, \text{ αποκοπη} \end{cases}$$

Η παραπάνω πρόταση είναι ισοδύναμη με την εξής:

$fl(x) = x(1+\varepsilon)$ όπου $\varepsilon = \varepsilon(x)$ με $|\varepsilon| \le u$

Απόδειξη:

$$fl(x) = x(1+\varepsilon) \Rightarrow \frac{fl(x)-x}{x} = \varepsilon \Rightarrow |\varepsilon| = \left|\frac{fl(x)-x}{x}\right| \le u \Rightarrow |\varepsilon| \le u .$$

Πολλαπλασιασμός:

Έστω $x, y \in \mathbb{R}$. Επιπλέον, $\exists\ \varepsilon_1, \varepsilon_2, \varepsilon_3$ με $|\varepsilon_1| \le u$, $|\varepsilon_2| \le u$, και $|\varepsilon_3| \le u$ τέτοια ώστε:

$fl(x) = x(1+\varepsilon_1)$, $fl(y) = y(1+\varepsilon_2)$ ενώ για το γινόμενο στον υπολογιστή θα είναι

$z = fl\big(fl(x)\, fl(y)\big) = \big[x(1+\varepsilon_1)\, y(1+\varepsilon_2)\big](1+\varepsilon_3)$ ή

$$z = xy\left(1+\varepsilon_1+\varepsilon_2+\varepsilon_3+\varepsilon_1\varepsilon_2+\varepsilon_1\varepsilon_3+\varepsilon_2\varepsilon_3+\varepsilon_1\varepsilon_2\varepsilon_3\right) \approx xy\left(1+\varepsilon_1+\varepsilon_2+\varepsilon_3\right)$$

Το σχετικό σφάλμα, σ , θα είναι:

$$\sigma = \left|\frac{z-xy}{xy}\right| \approx \left|\frac{xy\left(1+\varepsilon_1+\varepsilon_2+\varepsilon_3\right)-xy}{xy}\right| = \left|\varepsilon_1+\varepsilon_2+\varepsilon_3\right| \le |\varepsilon_1|+|\varepsilon_2|+|\varepsilon_3|$$

Άρα

$$\left|\frac{z-xy}{xy}\right| \le |\varepsilon_1|+|\varepsilon_2|+|\varepsilon_3|$$

με

$$\left.\begin{aligned} |\varepsilon_1| &\le u \\ |\varepsilon_2| &\le u \\ |\varepsilon_3| &\le u \end{aligned}\right\} \Rightarrow \sum_{i=1}^{3} |\varepsilon_i| \le 3u$$

Συνεπώς

$$\left|\frac{z-xy}{xy}\right| \le 3u \rightarrow \boxed{\sigma \le 3u}$$

Διαίρεση:

$$\frac{x}{y}: \quad z = fl\left(\frac{fl(x)}{fl(x)}\right) = \frac{x(1+\varepsilon_1)}{y(1+\varepsilon_2)}(1+\varepsilon_3) = \frac{x}{y}\frac{(1+\varepsilon_1)(1+\varepsilon_3)}{(1+\varepsilon_2)}$$

Άρα

$$\sigma = \left|\frac{\frac{x}{y}\frac{(1+\varepsilon_1)(1+\varepsilon_3)}{(1+\varepsilon_2)} - \left(\frac{x}{y}\right)}{\left(\frac{x}{y}\right)}\right| = \left|\frac{\frac{(1+\varepsilon_1)(1+\varepsilon_3)}{(1+\varepsilon_2)} - 1}{1}\right| \Rightarrow$$

$$\sigma = \left|\frac{(1+\varepsilon_1)(1+\varepsilon_3)}{(1+\varepsilon_2)} - 1\right| = \left|(1+\varepsilon_1)(1+\varepsilon_3)\left(1-\varepsilon_2+\varepsilon_2^2-\varepsilon_3^3+\ldots\right)-1\right| =$$

$$\left|1+\varepsilon_1+\varepsilon_3-\varepsilon_2+\mathrm{O}\left(\varepsilon_i^2\right)-1\right| \approx \left|\varepsilon_1+\varepsilon_3-\varepsilon_2\right| \le \left|\varepsilon_1\right|+\left|\varepsilon_2\right|+\left|\varepsilon_3\right| \le 3u$$

Επομένως, όπως και προηγουμένως: σφαλμα $\le 3u \;\rightarrow\; \boxed{\sigma \le 3u}$

Υπενθυμίζεται ότι: $\dfrac{1}{1-x} = 1+x+x^2+x^3+x^4+\ldots, \quad |x|<1$

<u>Πρόσθεση και αφαίρεση:</u>

$x+y: \quad z = fl\left(\, fl\left(x\right)+ fl\left(y\right)\right) = \left[x\left(1+\varepsilon_1\right)+y\left(1+\varepsilon_2\right)\right]\left(1+\varepsilon_3\right)$

άρα το σχετικό σφάλμα σ θα δίνεται από την σχέση :

$$\sigma = \left|\frac{z-\left(x+y\right)}{\left(x+y\right)}\right| = \left|\frac{\left[x\left(1+\varepsilon_1\right)+y\left(1+\varepsilon_2\right)\right]\left(1+\varepsilon_3\right)-\left(x+y\right)}{\left(x+y\right)}\right| \Rightarrow$$

$$\sigma \approx \left|\varepsilon_3+\frac{x\varepsilon_1+y\varepsilon_2}{x+y}\right| \le \left|\varepsilon_3\right|+\frac{|x||\varepsilon_1|+|y||\varepsilon_2|}{|x+y|} \le u+\frac{u\left(|x|+|y|\right)}{|x+y|} \quad (**)$$

Επομένως όταν οι αριθμοί x,y είναι ομόσημοι τότε $|x+y|=|x|+|y|$ και άρα η παραπάνω σχέση διαμορφώνεται ως εξής:

$$\sigma \approx \left|\varepsilon_3+\frac{x\varepsilon_1+y\varepsilon_2}{x+y}\right| \le 2u$$

Όταν όμως οι 2 αριθμοί είναι ετερόσημοι και ταυτόχρονα ισχύει $x \approx -y \Rightarrow |x+y| \approx 0$ τότε το άνω φράγμα στην σχέση (**) τείνει στο άπειρο!

<u>Παράδειγμα 1:</u> Έστω $x = 0.45142708$ και $y = -0.45115944$ και έστω σύνολο αριθμών κινητής υποδιαστολής $\mathbf{M} = \mathbf{M}\left(\beta=10, t=5, U=10, L=-10\right)$. Έχουμε ότι:

$$z = fl\left(\, fl\left(x\right)+ fl\left(y\right)\right) = fl\left(0.45143-0.45116\right) = 0.00027 = 0.27000\times 10^{-3}$$

Το σχετικό σφάλμα θα είναι $\sigma = \left|\dfrac{z-\left(x+y\right)}{\left(x+y\right)}\right| \approx 88\times 10^{-4}$ ενώ το αντίστοιχο άνω φράγμα αν οι αριθμοί ήταν ομόσημοι θα ήταν $2u = 2\dfrac{1}{2}\beta^{1-t} = 10^{1-5} = 10^{-4}$ (!!) που δείχνει ότι το σφάλμα στην αφαίρεση μπορεί να είναι πολύ μεγαλύτερο από το σφάλμα στην πρόσθεση (*Άσκηση: βρείτε πόσο ακριβώς είναι το σχετικό σφάλμα αν οι αριθμοί x και y ήταν πράγματι ομόσημοι).*

Παράδειγμα 2: Έστω $x = 451852000$ και $y = -451851000$ και έστω σύνολο αριθμών κινητής υποδιαστολής $\mathbf{M} = \mathbf{M}(\beta = 10, t = 5, U = 10, L = -10)$. Έχουμε ότι $x + y = 1000$. Όμως στο σύνολο που δόθηκε θα είχαμε $z = fl(fl(x) + fl(y)) = fl(0.45185 \times 10^9 - 0.45185 \times 10^9) = 0!!!!!$ Το παράδειγμα δείχνει ότι όταν έχουμε αφαίρεση μεγάλων αριθμών δημιουργείται σημαντικό πρόβλημα.

Παράδειγμα 3: Έστω $x = 7892$ και $y = 7891$ και έστω σύνολο αριθμών κινητής υποδιαστολής $\mathbf{M} = \mathbf{M}(\beta = 10, t = 10, U = 10, L = -10)$. Έχουμε ότι $\sqrt{x} = 0.8883692926 \times 10^2$, $\sqrt{y} = 0.8883130079 \times 10^2$ και $\sqrt{x} - \sqrt{y} = 0.5628470000 \times 10^{-2}$. Τα μηδενικά στο τέλος του αποτελέσματος είναι ένδειξη της απώλειας ακρίβειας. Εναλλακτικά μπορούμε να υπολογίσουμε $\sqrt{x} - \sqrt{y} = \frac{x - y}{\sqrt{x} + \sqrt{y}} = 0.5628468294 \times 10^{-2}$ το οποίο έχει πολύ μεγάλη ακρίβεια.

Παράδειγμα 4: Να υπολογιστεί η συνάρτηση $f(x) = x - sin(x)$ για πολύ μικρές τιμές του x, δηλαδή για $|x| << 1$.

Επειδή $\lim_{x \to 0} \frac{x}{sin(x)} = 1$ θα αντιμετωπίσουμε το πρόβλημα της αφαίρεσης σχεδόν ίσων αριθμών. Έτσι κάνουμε ανάπτυγμα Taylor της συνάρτησης γύρω από το x=0 και έχουμε:

$f(x) = x - \left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + O(x^9) \right) = \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} + O(x^9)$. Άρα για πολύ μικρές τιμές του x μπορούμε να διατηρήσουμε μόνο τον 1ο ή και τον 2ο όρο της σειράς και να μην έχουμε πρόβλημα με τους υπολογισμούς (γιατί?).

Άσκηση 1: *Υπολογίστε την συνάρτηση* $f(x) = \ln\left(\frac{1+x}{1-x}\right)$ *για πολύ μικρές τιμές του* x, *δηλαδή για* $|x| << 1$. *Ομοίως για την συνάρτηση* $f(x) = \frac{\ln(1+x)}{x}$.

<u>Άσκηση 2:</u> Αναδιαμορφώστε την έκφραση $\frac{1}{\sqrt{x}}-\frac{1}{\sqrt{x+1}}$ *για μεγάλες τιμές του* x*. Ποιο πρόβλημα θα παρουσιασθεί στον υπολογισμό της αρχικής έκφρασης? Διορθώνεται με την εναλλακτική έκφραση και γιατί?*

Χρήσιμες σχέσεις (αναπτύγματα Taylor γύρω από το x=0)

$$sin(x)=x-\frac{x^3}{3!}+\frac{x^5}{5!}-\frac{x^7}{7!}+O(x^9)$$

$$cos(x)=1-\frac{x^2}{2!}+\frac{x^4}{4!}-\frac{x^6}{6!}+O(x^8)$$

$$tan(x)=x+\frac{x^3}{3}+\frac{2x^5}{15}+\frac{17x^7}{315}+O(x^9)$$

$$exp(x)=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\frac{x^4}{4!}+O(x^5)$$

$$ln(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+\frac{x^5}{5}+O(x^6)$$

1.4. Σφάλματα στον υπολογισμό αθροισμάτων

Έστω ότι θέλουμε να υπολογίσουμε την σειρά $S_n=1+\sum_{k=1}^{n}\frac{1}{k^2+k}=1+\sum_{k=1}^{n}\frac{1}{k(k+1)}$. Επειδή ισχύει $\frac{1}{k^2+k}=\frac{1}{k}-\frac{1}{k+1}$ έχουμε ότι $S_n=1+\left(1-\frac{1}{2}\right)+\left(\frac{1}{2}-\frac{1}{3}\right)+\left(\frac{1}{3}-\frac{1}{4}\right)+...+\left(\frac{1}{n}-\frac{1}{n+1}\right)$,

$\Rightarrow S_n=2-\frac{1}{n+1}$ οπότε προκύπτει ότι $S_9=1.9,\quad S_{99}=1.99,\quad S_{999}=1.999$ κτλ, και φυσικά $\lim_{n\to\infty}S_n=2$. Ας αγνοήσουμε προς το παρόν ότι γνωρίζουμε τα παραπάνω και ας υποθέσουμε ότι προσπαθούμε να υπολογίσουμε την σειρά απευθείας. Ένας απλός αλγόριθμος για τον υπολογισμό του αθροίσματος είναι ο εξής:

$S_0=1,\quad S_k=S_{k-1}+\frac{1}{k(k+1)}$ για $k=1,2,3,...,N$.

Διαφορετικά μπορούμε να γράψουμε:

$S\leftarrow 1$

Για $k=1,2,3,...,N$ κάνε:

$$S\leftarrow S+\frac{1}{k(k+1)}$$

Αν όμως ο αλγόριθμος αυτός εφαρμοσθεί σε έναν υπολογιστή με $\beta = 10, t = 10$ τότε θα πάρουμε:

$S_9 = 1.900000000$

$S_{99} = 1.990000003$

$S_{999} = 1.999000003$

$S_{9999} = 1.999899972$

Έστω τώρα ότι αλλάζουμε την σειρά με την οποία υπολογίζουμε τους όρους του αθροίσματος. Δηλαδή:

$$T_0 = \frac{1}{n(n+1)}$$

$$T_k = T_{k-1} + \frac{1}{(n-k)(n-k+1)}, \quad k = 1,2,3,...,n-1 \text{ για } k = 1,2,3,...,N.$$

$$T_n = T_{n-1} + 1$$

Ο αντίστοιχος αλγόριθμος είναι:

$$T \leftarrow \frac{1}{n(n+1)}$$

Για $k = 1,2,3,...,n-1$, κάνε

$$T \leftarrow T + \frac{1}{(n-k)(n-k+1)}.$$

$$T \leftarrow T + 1$$

Τότε το πρόβλημα εξαφανίζεται και ο υπολογισμός μέχρι και τον όρο S_{9999} γίνεται με μηδενικό σφάλμα. Παρόλο που ξέρουμε ότι η πρόσθεση δεν έχει την προσεταιριστική ιδιότητα στον υπολογιστή, για πιο λόγο ο δεύτερος αλγόριθμος δίνει καλύτερα αποτελέσματα από τον πρώτο? Ας δούμε το πρόβλημα λίγο πιο γενικά και ας υποθέσουμε ότι δίνονται «n» στο πλήθος αριθμοί των οποίων θέλουμε να βρούμε το άθροισμά τους, $s_n = \sum_{k=1}^{n} a_k$ Για να απλοποιήσουμε την ανάλυση θα θεωρήσουμε ότι όλοι οι όροι του αθροίσματος είναι αριθμοί μηχανής, δηλαδή ότι $fl(a_k) = a_k, \quad k = 1,2,3,...,n$ καθώς επίσης ότι οι αριθμοί είναι διατεταγμένοι σε αύξουσα σειρά. Για λόγους ευκολίας θα συμβολίζουμε $fl(s_k) = \tilde{s}_k$. Ο αλγόριθμος υπολογισμού του αθροίσματος s_n, δίνεται από τον αναδρομικό τύπο, $s_1 = a_1, \quad s_k = s_{k-1} + a_k$ για $k = 2,3,4,...,n$. Επομένως στον υπολογιστή θα έχουμε:

$fl(s_1) = fl(a_1) = a_1, \quad fl(s_k) = fl\left(fl(s_{k-1}) + fl(a_k)\right) \Rightarrow \tilde{s}_k = fl(\tilde{s}_{k-1} + a_k)$ για $k = 2,3,4,...,n$.

Για $k=2$ έχουμε: $\tilde{s}_2 = fl\left(\tilde{s}_1 + a_2\right) = fl\left(a_1 + a_2\right) = \left(a_1 + a_2\right)\left(1+\varepsilon_1\right)$

Για $k=3$ έχουμε:

$$\tilde{s}_3 = fl\left(\tilde{s}_2 + a_3\right) = fl\left(\left(a_1 + a_2\right)\left(1+\varepsilon_1\right) + a_3\right) = \left[\left(a_1 + a_2\right)\left(1+\varepsilon_1\right) + a_3\right]\left(1+\varepsilon_2\right)$$

Για $k=4$ έχουμε: $\tilde{s}_4 = fl\left(\tilde{s}_3 + a_4\right) = fl\left(\left[\left(a_1 + a_2\right)\left(1+\varepsilon_1\right) + a_3\right]\left(1+\varepsilon_2\right) + a_4\right) =$

$$= \left\{\left[\left(a_1 + a_2\right)\left(1+\varepsilon_1\right) + a_3\right]\left(1+\varepsilon_2\right) + a_4\right\}\left(1+\varepsilon_3\right)$$

Για λόγους ευκολίας θα σταματήσουμε στον 4ο όρο χωρίς βλάβη της γενικότητας. Κάνοντας πράξεις στην τελευταία σχέση θα έχουμε:

$$\tilde{s}_4 = \left\{\left[\left(a_1 + a_2\right)\left(1+\varepsilon_1\right) + a_3\right]\left(1+\varepsilon_2\right) + a_4\right\}\left(1+\varepsilon_3\right) = \left\{a_1 + a_2 + a_3 + a_4\right\} + a_1\left(\varepsilon_1 + \varepsilon_2 + \varepsilon_3\right) +$$
$$+a_2\left(\varepsilon_1 + \varepsilon_2 + \varepsilon_3\right) + a_3\left(\varepsilon_2 + \varepsilon_3\right) + a_4\varepsilon_4 + O.Y.T.$$

Όμως $s_4 = a_1 + a_2 + a_3 + a_4$ οπότε έχουμε:

$$\tilde{s}_4 = s_4 + a_1\left(\varepsilon_1 + \varepsilon_2 + \varepsilon_3\right) + a_2\left(\varepsilon_1 + \varepsilon_2 + \varepsilon_3\right) + a_3\left(\varepsilon_2 + \varepsilon_3\right) + a_4\varepsilon_4 + O.Y.T. \Rightarrow$$
$$\left|\tilde{s}_4 - s_4\right| = \left|a_1\left(\varepsilon_1 + \varepsilon_2 + \varepsilon_3\right) + a_2\left(\varepsilon_1 + \varepsilon_2 + \varepsilon_3\right) + a_3\left(\varepsilon_2 + \varepsilon_3\right) + a_4\varepsilon_4 + O.Y.T.\right| \Rightarrow$$
$$\left|\tilde{s}_4 - s_4\right| \le \left|a_1\right|3u + \left|a_2\right|3u + \left|a_3\right|2u + \left|a_4\right|u$$

Η τελευταία σχέση δείχνει ότι το άνω φράγμα για το απόλυτο λάθος του αθροίσματος ελαχιστοποιείται όταν οι αριθμοί είναι σε αύξουσα σειρά διότι σε αυτή την περίπτωση ο μικρότερος αριθμός (κατά απόλυτη τιμή) θα πολλαπλασιάζεται με το μέγιστο συντελεστή σφάλματος (στο συγκεκριμένο παράδειγμα με το $3u$).

<u>Άσκηση</u>: Υλοποιήστε τους αλγορίθμους (*) και (**) χρησιμοποιώντας αρχικά μεταβλητές απλής ακρίβειας (real(4)) και στην συνέχεια μεταβλητές διπλής ακρίβειας (real(8)) και διαπιστώστε την συμπεριφορά τους όσον αφορά τα αποτελέσματα που δίνουν.

1.5. Αριθμητική ευστάθεια αλγορίθμων

Όπως αναφέρθηκε νωρίτερα αλγόριθμο ονομάζουμε την πεπερασμένη σειρά καλά ορισμένων μαθηματικών πράξεων και λογικών εκφράσεων που υλοποιούν ένα διακριτό σχήμα (ή αριθμητική μέθοδο). Λόγω όμως των σφαλμάτων στρογγύλευσης, που πάντα υφίστανται, κάποιοι αλγόριθμοι είναι τέτοιοι ώστε τα σφάλματα αυτά να συσσωρεύονται με τέτοιο τρόπο έτσι ώστε το τελικό αποτέλεσμα να είναι εντελώς ανακριβές. Ένας τέτοιος αλγόριθμος ονομάζεται *αριθμητικά ασταθής αλγόριθμος*. Διαφορετικά πρόκειται για έναν *αριθμητικά ευσταθή αλγόριθμο*. Αν ένας αλγόριθμος είναι ασταθής σε καμία περίπτωση δεν πρέπει να εμπιστευόμαστε τα αποτελέσματά του. Ο τρόπος με τον οποίο

διαπιστώνουμε αν ένας αλγόριθμος είναι ευσταθής είναι διαταράσσοντας κατά μικρές ποσότητες τα δεδομένα του. Αν το αποτέλεσμα αλλάξει κατά πολύ τότε έχουμε σημαντική ένδειξη ότι ο αλγόριθμος είναι ασταθής.
Η ευστάθεια και η ακρίβεια (θεωρητικό σφάλμα προσέγγισης) ενός αλγορίθμου είναι τα κύρια ποιοτικά χαρακτηριστικά του.

Παράδειγμα 1ο: Ο υπολογισμός της ποσότητας e^{-x} για μεγάλα θετικά x, με χρήση της σειράς Taylor. Η συνάρτηση e^{-x} μπορεί να προσεγγιστεί με την σειρά $S_n(x)=1-x+\frac{x^2}{2}-\frac{x^3}{6}+\frac{x^4}{24}+...+\frac{(-1)^{n-1}x^{n-1}}{(n-1)!}$, $n\geq 1$ για την οποία γνωρίζουμε ότι $\lim_{n\to\infty} S_n(x)=e^{-x}$. Έστω επίσης υπολογιστής με $\mathbf{M}=\mathbf{M}(\beta=10,t=5,U=10,L=-10)$ και $x=5.5$. Τότε οι πρώτοι όροι της σειράς είναι οι 1.0000, -5.5000, 15.125, -27.730, 38.129, -41.942, 38.446, -30.208, , 20.768, -12.692, 6.9803, -3.4902, +1.5997 το άθροισμα των οποίων δίνει 0.0026363. Όμως $e^{-5.5}=0.00408677$ και επομένως το προηγούμενο αποτέλεσμα δεν έχει κανένα σημαντικό ψηφίο σωστό! Το πρόβλημα δημιουργείται διότι για όλους τους αριθμούς που είναι μεγαλύτεροι από το 10 το αντίστοιχο απόλυτο σφάλμα προσέγγισης είναι της ίδιας τάξης (δηλαδή είναι του ίδιου μεγέθους) με το αποτέλεσμα!

Παράδειγμα 2ο: Ο αριθμητικός υπολογισμός του ολοκληρώματος $I_n=\int_{x=0}^{x=1} x^n e^{x-1}dx,\ n=1,2,3,...$ για μεγάλες τιμές του n. Ας δούμε πρώτα μερικά χαρακτηριστικά αυτής της σχέσης:

(α) $\left.\begin{array}{l} 0\leq x\leq 1\Rightarrow 0\leq x^n\leq 1 \\ 0\leq x\leq 1\Rightarrow \frac{1}{e}\leq e^{x-1}\leq 1 \end{array}\right\}\Rightarrow 0\leq x^n e^{x-1}\leq 1\Rightarrow 0\leq \int_0^1 x^n e^{x-1}dx\leq \int_0^1 1dx\Rightarrow 0\leq I_n\leq 1$

(β) $\left.\begin{array}{l} 0\leq x\leq 1 \\ n>0 \end{array}\right\}\Rightarrow x^{n+1}<x^n\Rightarrow \int_0^1 x^{n+1}e^{x-1}dx<\int_0^1 x^n e^{x-1}dx\Rightarrow I_{n+1}<I_n$

(γ) $0\leq x\leq 1\Rightarrow \frac{1}{e}\leq e^{x-1}\leq 1\Rightarrow x^n e^{x-1}\leq x^n\Rightarrow \int_0^1 x^n e^{x-1}dx\leq \int_0^1 x^n dx=\frac{1}{n+1}\Rightarrow I_n\leq \frac{1}{n+1}$

από όπου συμπεραίνουμε ότι $\lim_{n\to\infty} I_n\leq \lim_{n\to\infty}\frac{1}{n+1}=0\Rightarrow \lim_{n\to\infty} I_n=0$.

Εύρεση αναδρομικού τύπου: $I_n = \int_{x=0}^{x=1} x^n e^{x-1} dx = \int_{x=0}^{x=1} x^n e^{x-1} d(x-1) = \int_{x=0}^{x=1} x^n d\left(e^{x-1}\right) =$

$= x^n e^{x-1}\Big|_{x=0}^{x=1} - \int_{x=0}^{x=1} e^{x-1} d\left(x^n\right) = 1 - n\int_{x=0}^{x=1} e^{x-1} x^{n-1} dx = 1 - nI_{n-1}$. Για n=1 έχουμε

$I_1 = \int_{x=0}^{x=1} xe^{x-1} dx = xe^{x-1}\Big|_{x=0}^{x=1} - \int_{x=0}^{x=1} e^{x-1} dx = 1 - e^{x-1}\Big|_{x=0}^{x=1} = 1 - \left(1 - \frac{1}{e}\right) = \frac{1}{e}$, οπότε έχουμε τον τύπο:

$$\boxed{I_n = 1 - nI_{n-1}, \quad n > 1, \quad I_1 = \frac{1}{e}} \quad (*)$$

Έστω τώρα υπολογιστής με $\mathbf{M} = \mathbf{M}\left(\beta = 10, t = 6, U = 10, L = -10\right)$. Έχουμε ότι

$I_1 = \frac{1}{e} = 0.367879441171442 \Rightarrow fl\left(I_1\right) = 0.367879$. Εκτελούμε τις πράξεις και βρίσκουμε:

$fl\left(I_2\right) = 0.264242 < 1/3$
$fl\left(I_3\right) = 0.207274 < 1/4$
$fl\left(I_4\right) = 0.170904 < 1/5$
$fl\left(I_5\right) = 0.145480 < 1/6$
$fl\left(I_6\right) = 0.127120 < 1/7$
$fl\left(I_7\right) = 0.110160 < 1/8$
$fl\left(I_8\right) = 0.118720 > 1/9$??
$fl\left(I_9\right) = -0.068480 < 0$??

Παρατηρούμε ότι ο 8ος όρος της ακολουθίας είναι μεγαλύτερος από τον 7ο, ενώ ο 9ος είναι αρνητικός! Τι πήγε στραβά? Μία χοντροειδής ανάλυση σφάλματος θα μας δείξει τι ακριβώς συνέβη. Θα υποθέσουμε ότι όλες οι πράξεις γίνονται χωρίς σφάλμα εκτός από το σφάλμα που εισέρχεται στην τιμή του όρου I_n. Έχουμε δηλαδή

$I_n = 1 - nI_{n-1}, \quad n > 1, \quad I_1 = \frac{1}{e}$ και

$fl\left(I_n\right) = 1 - n\, fl\left(I_{n-1}\right), \quad n > 1, \quad I_1 = 0.367879$

Αφαιρώ τις τελευταίες 2 σχέσεις κατά μέρη και έχουμε:

$\underbrace{fl\left(I_n\right) - I_n}_{\varepsilon_n} = -n\underbrace{\left(fl\left(I_{n-1}\right) - I_{n-1}\right)}_{\varepsilon_{n-1}}, \quad n > 1, \quad \underbrace{fl\left(I_1\right) - I_1}_{\varepsilon_1} = -4.412 \times 10^{-7}$. Είτε:

$\varepsilon_n = -n\varepsilon_{n-1}, \quad n > 1, \quad \varepsilon_1 = -4.412 \times 10^{-7}$ με γενική λύση $\varepsilon_n = (-1)^n n!\varepsilon_1$ που δείχνει ότι όσο μικρό και αν είναι το αρχικό λάθος τελικά $\lim_{n\to\infty}\left|\varepsilon_n\right| = \infty$. Επομένως ο αλγόριθμος αυτός είναι ασταθής διότι τα σφάλματα στρογγύλευσης συσσωρεύονται μέχρι που καθιστούν τα αποτελέσματα εντελώς ανακριβή.

Εναλλακτικός αλγόριθμος με υπολογισμό από μεγάλα προς μικρά n.

$I_n = 1 - nI_{n-1} \Rightarrow I_{n-1} = \dfrac{1 - I_n}{n}$ οπότε ο αναδρομικός τύπος έχει ως εξής:

$$I_k = a < \frac{1}{k+1}, \quad k >> 1$$
$$I_{n-1} = \frac{1 - I_n}{n}, \quad n = k-1, k-2, ..., 2$$
$$I_1 = \frac{1}{e}$$

όπου, προς το παρόν, παραβλέπουμε ότι δεν γνωρίζουμε το α ακριβώς. Ακολουθώντας αντίστοιχη διαδικασία όπως και προηγουμένως θα έχουμε $\varepsilon_2 = (-1)^k \dfrac{\varepsilon_k}{k!}$ που σημαίνει ότι όσο μεγάλο λάθος και αν είχαμε εισάγει στον αρχικό όρο I_k αυτό το σφάλμα θα εκμηδενιζόταν πολύ γρήγορα. Να σημειωθεί ότι αφού $0 < I_k \le \dfrac{1}{k+1}$ άρα το μέγιστο αρχικό σφάλμα στο I_k δεν ξεπερνά την ποσότητα $\dfrac{1}{k+1}$. Ο αλγόριθμος αυτός έχει την ιδιότητα ότι το αρχικό σφάλμα φθίνει, δηλαδή δεν συσσωρεύεται, και έτσι πρόκειται για ευσταθή αλγόριθμο.

<u>*Παράδειγμα 3ο:*</u> Ο υπολογισμός του π μέσου του αναδρομικού τύπου $\Pi_{n+1} = 2^n \sqrt{2\left(1 - \sqrt{1 - \left(\dfrac{\Pi_n}{2^n}\right)^2}\right)}, \Pi_2 = 2^{3/2}, \quad n > 2$, ο οποίος αποτυγχάνει για μεγάλες του n (δείτε την 14η άσκηση της 1ης εργασίας).

1.6. Κατάσταση προβλημάτων

Καλά τοποθετημένα προβλήματα ονομάζονται τα προβλήματα τα οποία (α) έχουν μοναδική λύση και (β) η λύση τους εξαρτάται συνεχώς από τα δεδομένα του προβλήματος.

Αριθμητικά επιλύουμε μόνο προβλήματα που έχουν μοναδική λύση. Όμως λόγω των σφαλμάτων προσέγγισης, κάθε μαθηματικό πρόβλημα λύνεται σε περιβάλλον διαταραχών και επομένως μας ενδιαφέρει κατά πόσο ο αντίστοιχος αλγόριθμος είναι ευαίσθητος σε αλλαγές των δεδομένων του προβλήματος. Έτσι, λέμε ότι ένα πρόβλημα *βρίσκεται σε καλή κατάσταση* όταν μικρές αλλαγές στα δεδομένα του προβλήματος προκαλούν μικρές αλλαγές στην λύση του, δηλαδή όταν το πρόβλημα δεν είναι

ευαίσθητο σε διαταραχές των δεδομένων του. Σε διαφορετική περίπτωση, όταν δηλαδή μικρές αλλαγές στα δεδομένα προκαλούν μεγάλες αλλαγές στην λύση, λέμε ότι το πρόβλημα *βρίσκεται σε κακή κατάσταση*.

Αν ένα πρόβλημα βρίσκεται σε κακή κατάσταση τότε οποιαδήποτε αριθμητική μέθοδο και να ακολουθήσουμε, αυτή θα είναι ασταθής, λόγω της παρουσίας των σφαλμάτων προσέγγισης.

Επομένως, πρακτικά μπορούμε να επιλύσουμε μόνο καλά τοποθετημένα προβλήματα που βρίσκονται σε καλή κατάσταση. Φυσικά ο αντίστοιχος αλγόριθμος μπορεί να είναι ευσταθής είτε ασταθής.

<u>*Παράδειγμα:*</u> Έστω η αλγεβρική εξίσωση $(x-2)^6 = 0 \Rightarrow x = 2$ με πολλαπλότητα 6. Έστω το ελαφρά διαταραγμένο πρόβλημα $(x_\delta - 2)^6 = 10^{-6}$ το οποίο αντιστοιχεί στο αρχικό πρόβλημα αν απλά αλλάξουμε τον σταθερό όρο κατά την ποσότητα 10^{-6}. Οι λύσεις αυτού του προβλήματος είναι $x_\delta = 2 + \frac{1}{10}\exp(i\pi k/3), \quad k = 0,1,2,3,4,5$. Επομένως έχουμε $|x_\delta - x| = \left|2 + \frac{1}{10}\exp(i\pi k/3) - 2\right| = \frac{1}{10}\left|\exp(i\pi k/3)\right| = \frac{1}{10}$ όπου φαίνεται ότι η αλλαγή του ενός και μόνο συντελεστή (δεδομένο) του προβλήματος προκάλεσε αλλαγή στο μέτρο της λύσης κατά 10^{-1}, δηλαδή 10000 φορές μεγαλύτερη από την αντίστοιχη αλλαγή του συντελεστή! Προφανώς το πρόβλημα αυτό δεν βρίσκεται σε καλή κατάσταση.

Κεφάλαιο 2ο
Επίλυση μη-γραμμικών εξισώσεων

2.1. Εισαγωγή

Το πρόβλημα με το οποίο θα ασχοληθούμε στην συνέχεια είναι η αριθμητική επίλυση της μη-γραμμικής αλγεβρικής εξίσωσης $f(x)=0$, όπου θα περιοριστούμε σε πραγματικές συναρτήσεις πραγματικών μεταβλητών. Δηλαδή θα μελετήσουμε το πρόβλημα της προσέγγισης πραγματικών ριζών x^* της συνάρτησης, τέτοιων ώστε $f(x^*)=0$. Όπως είναι γνωστό, υπάρχουν αναλυτικές λύσεις για πολυωνυμικές συναρτήσεις μέχρι και 4ου βαθμού (τύποι Carnado). Στην γενική περίπτωση όμως αναλυτικές λύσεις είτε δεν υπάρχουν είτε είναι εξαιρετικά δύσκολο να βρεθούν.

Συνήθως μία αριθμητική μέθοδος παράγει μία ακολουθία προσεγγίσεων της λύσης x^*, $x_0, x_1, x_2, \ldots$ η οποία ακολουθία υπό κάποιες προϋποθέσεις συγκλίνει, δηλαδή

$$\lim_{n\to\infty} x_n = x^*$$

Συμβολισμός: Έστω $I \subset \mathbb{R}$ και $n \in \mathbb{N}$, τότε

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

$$C^n(I) = \{ f \in C(I) : f \text{ "n" φορες συνεχως παραγωγισιμη στο } I \}$$

Συχνά γράφουμε

$$C[a,b] \text{ αντί για } C([a,b])$$

και

$$C(a,b) \text{ αντί για } C((a,b))$$

ή αντίστοιχα

$$C^n(a,b) \text{ αντί για } C^n((a,b))$$

2.2. Μέθοδος Διχοτόμησης

Είναι η απλούστερη μέθοδος εύρεσης ριζών και βασίζεται στο θεώρημα της ενδιάμεσης τιμής:

Έστω $g \in C[a,b]$ *και* ξ *πραγματικός αριθμός που ανήκει στο διάστημα που ορίζουν τα* $g(a)$ *και* $g(\beta)$. *Τότε, υπάρχει* $x \in [a,b]$, *τέτοιο ώστε* $g(x) = \xi$.

Πριν προχωρήσουμε, ορίζουμε την συνάρτηση προσήμου $\operatorname{sgn}(x)$:

$$\operatorname{sgn}(x) = \begin{cases} +1, & \text{αν } x > 0 \\ -1, & \text{αν } x < 0 \\ 0\ , & \text{αν } x = 0 \end{cases}$$

Η ιδέα της μεθόδου της διχοτόμησης είναι η εξής: έστω $f \in C[a,b]$ και ότι η f έχει ετερόσημες τιμές στα άκρα του $[a,b]$, δηλαδή

$$\operatorname{sgn}(f(a)) \neq \operatorname{sgn}(f(b))$$

τότε, με βάση το θεώρημα της ενδιάμεσης τιμής, το σημείο $\xi = 0$ ανήκει στο διάστημα που ορίζουν τα $f(a)$ και $f(b)$. Άρα υπάρχει $x^* \in [a,b]$ τέτοιο ώστε $f(x^*) = 0$ που σημαίνει ότι το x^* είναι ρίζα της εξίσωσης $f(x) = 0$. Πιο συγκεκριμένα, υπολογίζουμε την ποσότητα

$$x_0 = \frac{a+b}{2}.$$

Τότε για το μέσον του διαστήματος $[a,b]$, x_0 υπάρχουν οι εξής δυνατότητες:

a. $f(x_0) = 0$ ή

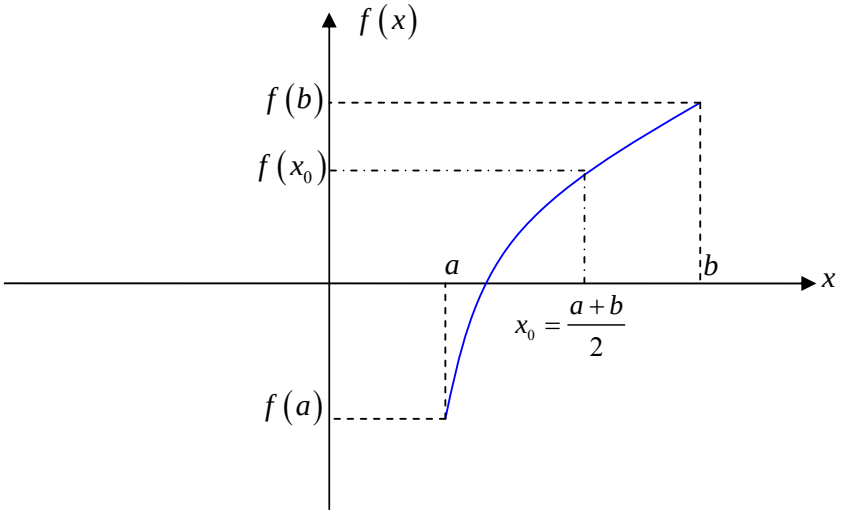
b. $f(x_0) \neq 0 \Rightarrow \begin{cases} f(x_0) < 0 \\ \text{ειτε} \\ f(x_0) > 0 \end{cases}$

Αν ισχύει η περίπτωση (a), τότε προφανώς ρίζα της $f(x) = 0$ είναι το x_0.

Αν ισχύει η περίπτωση (b), τότε η ρίζα βρίσκεται στο $(a, x_0]$ είτε στο $[x_0, b)$. Συγκεκριμένα:

- Αν $\operatorname{sgn}(f(a)) \neq \operatorname{sgn}(f(x_0))$, τότε η ρίζα βρίσκεται στο (a, x_0),
- Αν $\operatorname{sgn} f(b) \neq \operatorname{sgn} f(x_0)$, τότε η ρίζα βρίσκεται στο (x_0, b).

Γεωμετρική ερμηνεία:



$$\left.\begin{array}{l} f(a)<0 \Rightarrow \operatorname{sgn}(f(a))=-1 \\ f(b)>0 \Rightarrow \operatorname{sgn}(f(b))=+1 \end{array}\right\} \Rightarrow \operatorname{sgn}(f(a)) \neq \operatorname{sgn}(f(b))$$

Στο παραπάνω παράδειγμα βλέπουμε ότι $\operatorname{sgn}(f(a)) \neq \operatorname{sgn}(f(x_0))$ και επομένως η ρίζα θα βρίσκεται στο διάστημα (a, x_0) όπως είναι φανερό από το σχήμα. Επαναλαμβάνουμε τη διαδικασία στο νέο διάστημα (a, x_0). Είναι προφανές ότι η ρίζα που ψάχνουμε να βρούμε βρίσκεται στο διάστημα (a, x_0), δηλαδή βρίσκεται εγκλωβισμένη σε ένα διάστημα μικρότερο του αρχικού.

- Μήκος αρχικού διαστήματος:

$$b-a\,(>0)$$

- Μήκος νέου διαστήματος:

$$\frac{a+b}{2}-a=\frac{b-a}{2}>0$$

δηλαδή, ακριβώς το μισό του αρχικού διαστήματος. Επαναλαμβάνουμε την ίδια διαδικασία θεωρώντας ως άκρα του διαστήματος εύρεσης των ριζών τα νέα άκρα όπως προσδιορίστηκαν προηγουμένως.

Η διαδικασία αυτή συγκλίνει με την έννοια ότι σε πεπερασμένο πλήθος βημάτων υπολογίζουμε μία ρίζα της εξίσωσης $f(x)=0$, είτε την εγκλωβίζουμε σε ένα διάστημα, όσο μικρό θέλουμε και όσο βέβαια είναι εφικτό αφού στον υλεκτρονικό υπολογιστή πάντα περιοριζόμαστε από την ακρίβειά του.

Παράδειγμα: Να βρεθεί με την μέθοδο της διχοτόμησης μια ρίζα της $f(x)=x^3+x^2-3x-3$ στο $x\in[1,2]$ κάνοντας τουλάχιστον 4 επαναλήψεις.

Λύση: Είναι:

$$f(1)=1+1-3-3=-4<0$$

και

$$f(2)=8+4-6-3=1>0$$

άρα

$$f(1)\,f(2)<0 \;\Rightarrow\; x^*\in[1,2]$$

Επιπλέον, εξετάζουμε την μονοτονία της συνάρτησης f και βρίσκουμε ότι

$$f'(x)=3x^2+2x-3=3\left(x^2-1\right)+2x>0, \qquad \forall x\in[1,2]$$

άρα η $f(x)$ είναι αύξουσα στο $[1,2]$ και επομένως η ρίζα που θα έχει θα είναι μοναδική.

Υπολογίζουμε:

- ✓ $x_0=\dfrac{1+2}{2}=1.5 \;\rightarrow\; f(1.5)=-1.875<0$, άρα η ρίζα $\in[1.5,2.0]$
- ✓ $x_1=\dfrac{1.5+2}{2}=1.75 \;\rightarrow\; f(1.75)=0.171875>0$, άρα η ρίζα $\in[1.5,1.75]$
- ✓ $x_2=\dfrac{1.5+1.75}{2}=1.625 \;\rightarrow\; f(1.625)=-0.94336<0$, άρα η ρίζα $\in[1.625,1.75]$
- ✓ $x_3=\dfrac{1.625+1.75}{2}=1.6875 \;\rightarrow$ αυτή είναι η ζητούμενη ρίζα.

Επίσης, ισχύει

$$|x_1-x_0|=0.25\,,\quad |x_2-x_1|=0.125\,,\quad |x_3-x_2|=0.0625$$

Πρόταση: Έστω $f\in C[a,b]$ με $\operatorname{sgn}(f(a))\neq\operatorname{sgn}(f(b))$ και $\{x_n\}_{n\in\mathbb{N}}$ η ακολουθία των προσεγγίσεων που προκύπτει από την μέθοδο της διχοτόμησης. Τότε, είτε $x_N=x^*$ για κάποιο N, είτε $\lim\limits_{n\to\infty}x_N=x^*$, όπου $x^*\in(a,b)$ ρίζα της εξίσωσης $f(x)=0$. Για το σφάλμα ισχύει:

$$|x^*-x_n|\le\frac{b-a}{2^n},\ n=1,2,3,\ldots$$

Απόδειξη: Θέτουμε $a_1\equiv a$ και $b_1\equiv b$, οπότε συμβολίζουμε με $I_i\equiv[a_i,b_i]$, άρα και $I_1\equiv[a,b]$ τα διαδοχικά διαστήματα $(i=1,2,\ldots)$ τα οποία δημιουργεί η μέθοδος της

διχοτόμησης. Έστω, x_i το μέσον κάθε διαστήματος I_i, οπότε προφανώς ισχύει $I_{i+1} \subset I_i$. Επειδή σε κάθε I_i υπάρχει ρίζα της εξίσωσης $f(x)=0$, τα διαστήματα αυτά (σε περίπτωση που δεν ισχύει $x_N = x^*$) είναι άπειρα στο πλήθος.

Έχουμε

$$b_n - a_n = \frac{b_{n-1} - a_{n-1}}{2^1} = \frac{\frac{b_{n-2} - a_{n-2}}{2}}{2} = \frac{b_{n-2} - a_{n-2}}{2^2} = \ldots \Rightarrow \boxed{b_n - a_n = \frac{b_{n-k} - a_{n-k}}{2^k}}$$

για $k = n-1$ η παραπάνω ισότητα δίνει:

$$b_n - a_n = \frac{b_1 - a_1}{2^{n-1}} = \frac{b-a}{2^{n-1}} \quad (*)$$

επιπλέον για την αντίστοιχη προσέγγιση της ρίζας ισχύει

$$x_n = \frac{a_n + b_n}{2}$$

Επομένως στην «n» επανάληψη της μεθόδου έχουμε ότι η απόσταση του x_n από το x^* θα είναι κατά απόλυτη τιμή μικρότερη ή ίση του μισού μήκους του τρέχοντος διαστήματος:

$$\left|x_n - x^*\right| \le \frac{b_n - a_n}{2} \overset{(*)}{=} \frac{b-a}{2^n}, n = 0,1,2,...$$

Η παραπάνω σχέση μας λέει ότι το τελικό σφάλμα της ρίζας είναι άνω φραγμένο από την ποσότητα $\frac{b-a}{2^n}$, η οποία τείνει στο μηδέν καθώς $n \to \infty$. Μπορεί να χρησιμοποιηθεί για να υπολογιστεί εκ'των προτέρων ο αριθμός των απαιτούμενων επαναλήψεων της μεθόδου για την εύρεση της ρίζας με συγκεκριμένη ακρίβεια. Πράγματι, αν θέλουμε να υπολογίσουμε την ρίζα μίας εξίσωσης με ακρίβεια μικρότερη ή ίση με ε αυτό σημαίνει ότι:

$$\left|x_n - x^*\right| \le \frac{b-a}{2^n} \le \varepsilon \Rightarrow \frac{b-a}{2^n} \le \varepsilon \Rightarrow b-a \le \varepsilon 2^n \Rightarrow \ln(b-a) \le \ln\varepsilon + n\ln 2 \Rightarrow$$

$$\ln(b-a) - \ln\varepsilon \le n\ln 2 \Rightarrow \ln\left(\frac{b-a}{\varepsilon}\right) \le n\ln 2 \Rightarrow n \ge \frac{\ln\left(\frac{b-a}{\varepsilon}\right)}{\ln 2}$$

Επομένως αν πάρουμε τον κοντινότερο ακέραιο στην ποσότητα $\frac{\ln\left((b-a)/\varepsilon\right)}{\ln 2}$ θα επιτύχουμε την επιθυμητή ακρίβεια.

Αλγόριθμος:

Δίνουμε δεδομένα b, a και πλάτος τελικού διαστήματος ε.

i. Υπολογίζουμε το $n = \text{nint}\left(\dfrac{\ln\left(\dfrac{b-a}{\varepsilon} \right)}{\ln 2} \right)$ και τα $f(a)$ και $f(b)$

ii. Υπολογίζουμε το μέσον του διαστήματος $x = a + \dfrac{b-a}{2}$ και το $f(x)$.

iii. Έλεγχος: εάν $f(x) = 0$ τότε έξοδος. Διαφορετικά, (εάν $f(x) \neq 0$:

Αν $\text{sgn}(f(x)) = \text{sgn}(f(a))$, τότε

$$a \leftarrow x$$

αλλιώς

$$b \leftarrow x$$

iv. Επαναλαμβάνουμε τα βήματα (ii) και (iii) «n» φορές

v. Τυπώνουμε a, b, $|b-a|$, x, $f(x)$

Στο πρώτο βήμα του αλγορίθμου η συνάρτηση nint(x) δίνει τον κοντινότερο ακέραιο στην ποσότητα x. Π.χ. nint(7.3)=7 και nint(1.52)=2.

<u>Σχόλια</u>: Η μέθοδος της διχοτόμησης, με την προϋπόθεση ότι $\text{sgn}(f(a)) \neq \text{sgn}(f(b))$, συγκλίνει εφόσον η $f(x)$ είναι συνεχής στο $[a,b]$, αλλά συγκλίνει αργά.

Η μεγαλύτερη ακρίβεια που μπορούμε να επιτύχουμε στον υπολογιστή (δηλαδή το ελάχιστο ε) είναι φυσικά ίση με την απόσταση δύο αριθμών κινητής υποδιαστολής στο διάστημα που ψάχνουμε να βρούμε την ρίζα.

<u>Παράδειγμα:</u>: Έστω η συνάρτηση $f(x) = x^3 - 2x - 5$ και ότι ζητάμε την ρίζα της εξίσωσης $f(x) = 0$ στο διάστημα (2,3). Παρατηρούμε ότι:

$$\left.\begin{array}{l} f(2) = 8 - 4 - 5 = -1 < 0 \\ f(3) = 27 - 6 - 5 = 16 > 0 \end{array}\right\} \Rightarrow \text{sgn}(f(2)) \neq \text{sgn}(f(3)) \Rightarrow x^* \in (2,3)$$

Επιπλέον, ισχύει:

$$\left.\begin{array}{l} f'(x) = 3x^2 - 2 > 0,\ \forall x \in (2,3) \\ f''(x) = 6x > 0 \end{array}\right\} \Rightarrow \text{η } f(x) \text{ είναι αύξουσα στο } (2,3)$$

συνεπώς, η ρίζα $x^* \in (2,3)$ είναι η μοναδική ρίζα στο διάστημα αυτό. Έστω, ότι θέλουμε να υπολογίσουμε την ρίζα x^* με ακρίβεια 10^{-5}, δηλαδή $|x_n - x^*| \le 10^{-5}$. Εφαρμόζοντας την σχέση (**) για τις ελάχιστες απαιτούμενες επαναλήψεις παίρνουμε

$$n \ge \frac{\ln\left(\frac{b-a}{\varepsilon}\right)}{\ln 2} \approx 17$$

2.3. Επαναληπτικές μέθοδοι

Λόγω του γεγονότος ότι η μέθοδος της διχοτόμησης συγκλίνει πολύ αργά έχουν αναπτυχθεί άλλες μέθοδοι, οι οποίες ονομάζονται *επαναληπτικές*, για τις οποίες όμως δεν γνωρίζουμε εκ' των προτέρων πόσες επαναλήψεις απαιτούνται για να συγκλίνουν με την εκάστοτε επιθυμητή ακρίβεια. Αρχικά, δίνουμε ορισμένους ορισμούς.

Ορισμός σταθερού σημείου:

Έστω φ μία συνάρτηση και x^* ένα σημείο του πεδίου ορισμού της. Αν ισχύει $x^* = \varphi(x^*)$, τότε το x^* λέγεται *σταθερό σημείο* της $\varphi(x)$.

Στις επαναληπτικές μεθόδους έχουμε μια εξίσωση $f(x) = 0$ την οποία γράφουμε σε μία ισοδύναμη μορφή της $x = \varphi(x)$. Ξεκινάμε από μία αρχική τιμή της λύσης x^*, x_0 και παράγουμε μία ακολουθία προσεγγίσεων της λύσης σύμφωνα με την σχέση

$$x_n = \varphi(x_{n-1}),\ n \in \mathbb{N}_0 \quad \text{ή} \quad x_{n+1} = \varphi(x_n),\ n \in \mathbb{N}$$

Αν η $\{x_n\}_{n\in\mathbb{N}}$ συγκλίνει σε ένα σημείο x^* και η $\varphi(x)$ είναι συνεχής στο σημείο αυτό, τότε το x^* είναι σταθερό σημείο της $\varphi(x)$, δηλαδή $x^* = \varphi(x^*)$.

Πράγματι:

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

Άρα $x^* = \varphi(x^*)$ και εφόσον ικανοποιείται η σχέση αυτή, άρα $f(x^*) = 0$ (εφόσον η $f(x) = 0$ είναι ισοδύναμη με την $x = \varphi(x)$).

Πρόταση: Κάθε συνεχής συνάρτηση $\varphi : [a,b] \to [a,b]$ έχει στο διάστημα $[a,b]$ τουλάχιστον ένα σταθερό σημείο.

Απόδειξη:

Λόγω υπόθεσης $\varphi([a,b]) \subset [a,b]$ ισχύει ότι:

$$\{\varphi(a) = a\} \text{ ή } \{\varphi(b) = b\} \text{ ή } \{\varphi(a) > a \text{ και } \varphi(b) < b\}$$

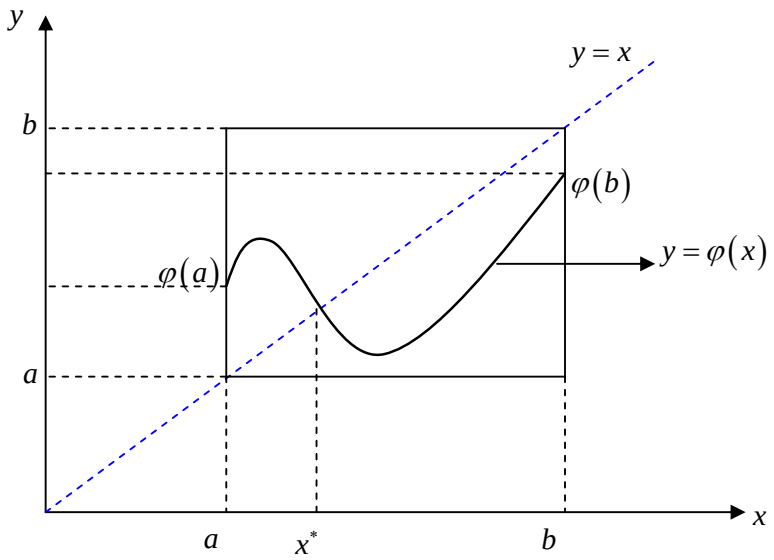
Αν ισχύει μια από τις πρώτες 2 προτάσεις, τότε ήδη υπάρχει ένα σταθερό σημείο στο $[a,b]$. Έστω ότι ισχύει η 3η πρόταση. Ορίζουμε την συνάρτηση $g:[a,b] \to \mathbb{R}$ με $g(x) = x - \varphi(x)$, η οποία φυσικά είναι συνεχής στο $[a,b]$ ως σύνθεση συνεχών συναρτήσεων. Έχουμε:

$$g(a) = a - \varphi(a) < 0$$

$$g(b) = b - \varphi(b) > 0$$

Επομένως, σύμφωνα με το θεώρημα ενδιάμεσης τιμής και $\forall \xi$ που ανήκει στο διάστημα που ορίζουν τα $g(a)$ και $g(b)$, υπάρχει x^*, τέτοιο ώστε $g(x^*) = \xi$. Επιλέγουμε $\xi = 0$, οπότε $g(x^*) = 0 \Rightarrow x^* - \varphi(x^*) = 0 \Rightarrow \boxed{x^* = \varphi(x^*)}$.

Γεωμετρική ερμηνεία



Παρατηρούμε ότι οι συναρτήσεις $y = \varphi(x)$ και $y = x$ έχουν τουλάχιστον ένα κοινό σημείο x^*, το οποίο ονομάζεται σταθερό σημείο της $\varphi(x)$.

Παρατήρηση: Η συνθήκη του θεωρήματος είναι ικανή, αλλά όχι αναγκαία συνθήκη.

Παράδειγμα: Έστω $\varphi:[-1,+1]\to[0,2]$ με $\varphi(x)=2x^2$. Η $\varphi(x)$ είναι συνεχής στο $[-1,+1]$. Επιπλέον, ισχύει $\varphi(0)=0$ και $\varphi\left(\frac{1}{2}\right)=\frac{1}{2}$ όπου $0,\ \frac{1}{2}\ \in[-1,+1]$ είναι σταθερά σημεία της φ, αλλά όμως δεν ισχύει $\varphi([-1,+1])\subset[-1,+1]$.

Ορισμός συνθήκης Lipschitz

Έστω, $I\subset\mathbb{R}$ και $\varphi:I\to\mathbb{R}$, θα λέμε ότι η φ ικανοποιεί την συνθήκη Lipschitz αν υπάρχει $L\geq 0$, τέτοια ώστε $|\varphi(x)-\varphi(y)|\leq L|x-y|\ \ \forall x,y\in I$ ($\Rightarrow$ η φ είναι ομοιόμορφα συνεχής στο I). Η μικρότερη σταθερά για την οποία ισχύει η προηγούμενη ανισότητα ονομάζεται σταθερά Lipschitz. Όταν $L<1$ η φ είναι συστολή στο I.

Παράδειγμα: Έστω $\varphi:[-a,a]\to\mathbb{R}, a>0$ με $\varphi=x^2$. Έχουμε

$$|\varphi(x)-\varphi(y)|=|x^2-y^2|=|(x-y)(x+y)|=|(x-y)||(x+y)|\Rightarrow$$

$$|\varphi(x)-\varphi(y)|=|(x-y)||(x+y)|\leq|(x-y)|(|x|+|y|)\Rightarrow$$

$$|\varphi(x)-\varphi(y)|\leq\max_{x,y\in I}(|x|+|y|)|x-y|=2a|x-y|$$

Άρα, ισχύει $|\varphi(x)-\varphi(y)|\leq L|x-y|$ με $L=2a<\infty\ \ \forall x,y\in I$. Αρα η φ είναι Lipschitz. Αν επιπλέον ισχύει $2a<1\Rightarrow a<\frac{1}{2}$ τότε είναι συστολή.

Σχόλιο: η συνθήκη Lipschitz είναι μια συνθήκη ομαλότητας για συναρτήσεις, η οποία είναι πιο ισχυρή από την απλή συνέχεια και επομένως αν μία συνάρτηση ικανοποιεί την συνθήκη Lipschitz είναι και συνεχής αλλά όχι το αντίθετο. Διαισθητικά σημαίνει ότι η συνάρτηση είναι περιορισμένη στο πόσο γρήγορα αλλάζει. Έτσι, μια γραμμή που συνδέει 2 σημεία της συνάρτησης ποτέ δεν θα έχει μεγαλύτερη κλίση από ένα συγκεκριμένο αριθμό που λέγεται σταθερά Lipschitz.

Παρατήρηση 1η: Κάθε συνάρτηση $\varphi\in C^1[a,b]$ ικανοποιεί την συνθήκη του Lipschitz με $L=\max_{x\in I}|\varphi'(x)|$. Πράγματι, από το θεώρημα μέσης τιμής έχουμε ότι $\forall x,y\in[a,b]$, υπάρχει $\xi\in(a,b)$ τέτοιο ώστε

$$|\varphi(x)-\varphi(y)|=\varphi'(\xi)(x-y)\Rightarrow|\varphi(x)-\varphi(y)|=|\varphi'(\xi)||x-y|\leq\max_{\xi\in I}|\varphi'(\xi)||x-y|=$$

$$\max_{x\in I}\left|\varphi'(x)\right||x-y| \Rightarrow \left|\varphi(x)-\varphi(y)\right| \le L|x-y|$$

όπου παραπάνω έχει χρησιμοποιηθεί το γεγονός ότι εφόσον η $\varphi'(x)$ είναι συνεχής στο $[a,b]$, άρα έχει μέγιστο στο διάστημα αυτό. Έτσι, στο προηγούμενο παράδειγμα εναλλακτικά θα μπορούσαμε να υπολογίζουμε την παράγωγο της συνάρτησης και να βρούμε αν έχει μέγιστο, κατά απόλυτη τιμή, στο διάστημα που μας ενδιαφέρει. Δηλαδή:

$$\varphi'(x) = 2x \Rightarrow \max_{x\in[a,b]}\left|\varphi'(x)\right| = 2a < \infty$$

οπότε και καταλήγουμε στο ίδιο συμπέρασμα όπως και προηγουμένως.

<u>Παρατήρηση 2η</u>: Αν $\varphi \in C^1(a,b)$, τότε δεν ικανοποιεί κατ' ανάγκη την συνθήκη του Lipschitz.

Παράδειγμα:

Έστω $\varphi:(0,1)\to\mathbb{R}$ με $\varphi(x)=\sqrt{x}$ και $\varphi'(x)=\dfrac{1}{2\sqrt{x}}$ στο $(0,1)$.

$$\left|\varphi(x)-\varphi(y)\right| = \left|\sqrt{x}-\sqrt{y}\right| = \frac{|x-y|}{\left|\sqrt{x}+\sqrt{y}\right|} \Rightarrow$$

$$\left|\varphi(x)-\varphi(y)\right| = \frac{1}{\left|\sqrt{x}+\sqrt{y}\right|}|x-y| \le \max_{x,y\in I}\left(\frac{1}{\left|\sqrt{x}\right|+\left|\sqrt{y}\right|}\right)|x-y|$$

Ποιο είναι το $\max\limits_{x,y\in I}\left(\dfrac{1}{\left|\sqrt{x}\right|+\left|\sqrt{y}\right|}\right)$? Παρατηρούμε ότι $\max\limits_{x,y\in I}\left(\dfrac{1}{\left|\sqrt{x}\right|+\left|\sqrt{y}\right|}\right)=\infty$, αφού $\lim\limits_{\substack{x\to 0^+\\ y\to 0^+}}\left(\dfrac{1}{\left|\sqrt{x}\right|+\left|\sqrt{y}\right|}\right)=\infty$ και επομένως δεν υπάρχει μέγιστο. Άρα η $\varphi(x)$ δεν ικανοποιεί τη συνθήκη Lipschitz. Συνεπώς, υπάρχουν συνεχείς συναρτήσεις που δεν ικανοποιούν τη συνθήκη Lipschitz.

<u>*Παράδειγμα 1ο*</u>: Αν $\varphi:\mathbb{R}\to\mathbb{R}$ με $\varphi(x)=x^2$, τότε $\max\limits_{x\in\mathbb{R}}\left|\varphi'(x)\right|=\infty$ $\Rightarrow$ η $\varphi(x)$ δεν είναι Lipschitz.

<u>*Παράδειγμα 2ο*</u>: Έστω $\varphi(x)=|x|$ ή $\varphi(x)=\begin{cases} x, & \text{αν } x>0\\ -x, & \text{αν } x\le 0\end{cases}$ με $\varphi'(x)=\begin{cases} 1, & \text{αν } x>0\\ -1, & \text{αν } x<0\end{cases}$

όμως

$$|\varphi'(x)| = 1 \Rightarrow \max|\varphi'(x)| = 1$$

άρα η φ είναι Lipschitz. Η $\varphi(x)$ δεν είναι παραγωγίσιμη στο x=0 αλλά είναι Lipschitz.

Εναλλακτικά:

$|\varphi(x) - \varphi(y)| = \big||x| - |y|\big| \le |x - y|$, άρα Lipschitz με $L = 1$

2.4. Θεώρημα σταθερού σημείου Banach ή θεώρημα συστολής

Έστω $\varphi : [a,b] \to [a,b]$ μια συστολή με σταθερά $L\,(<1)$, τότε η φ έχει στο $[a,b]$ ένα μοναδικό σημείο, δηλαδή $\exists\, x^* \in [a,b] : \varphi(x^*) = x^*$. Για κάθε $x_0 \in [a,b]$ η $x_n = \varphi(x_{n-1})$ είναι

a) καλά ορισμένη ($\Rightarrow \forall n \in \mathbb{N}, x_n \in [a,b]$),

b) συγκλίνει στο x^*, ($\Rightarrow \lim_{n\to\infty} x_n = x^*$) και

c) σε ότι αφορά τα σφάλματα ισχύουν τα εξής

i. $|x_n - x^*| \le L^n |x_0 - x^*| \le L^n \max(x_0 - a, b - x_0)$

ii. $|x_n - x^*| \le \dfrac{L^n}{1-L} |x_1 - x_0|$

iii. $|x_n - x^*| \le \dfrac{L}{1-L} |x_n - x_{n-1}|$

<u>*Απόδειξη*</u>:

✓ <u>Ύπαρξη λύσης</u>

$\forall x \in [a,b]$ ισχύει $a \le \varphi(x) \le b$

τετριμμένες περιπτώσεις $\to$ $\varphi(a) = a$ ή $\varphi(b) = b$

- Για $x = a$ έχουμε $\varphi(a) \ge a \Rightarrow a - \varphi(a) \le 0$
- Για $x = b$ έχουμε $\varphi(b) \le b \Rightarrow b - \varphi(b) \ge 0$

Εφόσον η φ είναι Lipschitz $\Rightarrow$ η φ είναι συνεχής στο $[a,b]$. Επομένως και η συνάρτηση

$$g(x)=x-\varphi(x)$$

είναι συνεχής στο $[a,b]$. Επιπλέον, ισχύει

$$\left.\begin{array}{l} g(a)=a-\varphi(a)\le 0 \\ g(b)=b-\varphi(b)\ge 0 \end{array}\right\} \Rightarrow \exists x^*\in[a,b]::g(x^*)=0 \Rightarrow \boxed{x^*=\varphi(x^*)}$$

✓ <u>Μοναδικότητα λύσης</u>

Έστω x^* και y^* με $x^*\neq y^*$ σταθερά σημεία της φ, τότε

$$\left|\varphi(x^*)-\varphi(y^*)\right|=\left|x^*-y^*\right|\le L\left|x^*-y^*\right| \Rightarrow L\ge 1 \quad \text{άτοπο}$$

διότι η φ είναι συστολή με σταθερά $L<1 \Rightarrow \boxed{x^*=y^*}$ $\Rightarrow$ x^* μοναδικό

✓ <u>Καλά ορισμένη ακολουθία</u>

Πράγματι, εφόσον $x_0\in[a,b]$, τότε $x_1=\varphi(x_0)\in[a,b]\Rightarrow x_1\in[a,b]$. Όμοια $x_2=\varphi(x_1)\in[a,b]$ Το ίδιο ισχύει και για τα $x_3,x_4,\ldots$, και επαγωγικά έχουμε ότι $x_n\in[a,b]$, $\forall n\in\mathbb{N}$. Άρα η $\{x_n\}_{n\in\mathbb{N}}$ είναι καλά ορισμένη ακολουθία.

✓ <u>Σύγκλιση</u>

Από Lipschitz έχουμε:

$$\left|x_n-x^*\right|=\left|\varphi(x_{n-1})-\varphi(x^*)\right|\le L\left|x_{n-1}-x^*\right|=L\left|\varphi(x_{n-2})-\varphi(x^*)\right|\le L^2\left|x_{n-2}-x^*\right|=\ldots\le L^n\left|x_0-x^*\right|$$

άρα

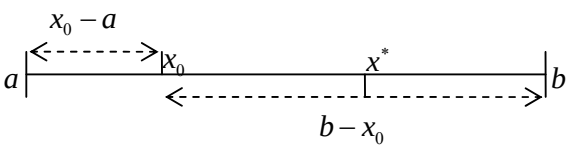
$$\left|x_n-x^*\right|\le L^n\left|x_0-x^*\right|$$

επομένως,

$$\lim_{n\to\infty}\left|x_n-x^*\right|\le\left|x_0-x^*\right|\lim_{n\to\infty}L^n=0\Rightarrow\lim_{n\to\infty}\left|x_n-x^*\right|=0\Rightarrow\lim_{n\to\infty}x_n=x^*$$

Συνεπώς, ισχύει

$$\boxed{\left|x_n-x^*\right|\le L^n\left|x_0-x^*\right|\le L^n\max(x_0-a,b-x_0)}$$



Εκτίμηση του (c.ii), δηλαδή του $\left|x_n-x^*\right|\le\dfrac{L^n}{1-L}\left|x_1-x_0\right|$. Έχουμε

$$|x_2 - x_1| = |\varphi(x_1) - \varphi(x_0)| \le L|x_1 - x_0|$$

$$|x_3 - x_2| = |\varphi(x_2) - \varphi(x_1)| \le L|x_2 - x_1| = L^2|x_1 - x_0|$$

$$\vdots$$

$$|x_{n+1} - x_n| \le L^n |x_1 - x_0|$$

Επίσης, έχουμε $\forall k \in \mathbb{N}$

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

$$|x_{n+k} - x_{n+k-1}| + |x_{n+k-1} - x_{n+k-2}| + \cdots + |x_{n+1} - x_n| \le L^{n+k-1}|x_1 - x_0| + L^{n+k-2}|x_1 - x_0| + \cdots + L^n|x_1 - x_0| =$$

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

Όμως, ισχύει η ταυτότητα

$$\alpha^\nu - \beta^\nu = (\alpha - \beta)\left(\alpha^{\nu-1} + \alpha^{\nu-2}\beta + \cdots + \alpha\beta^{\nu-2} + \beta^{\nu-1}\right)$$

στην οποία αντικαθιστώντας με $\alpha = 1$, $\beta = L$ και $\nu = k$ λαμβάνουμε την σχέση

$$\boxed{1 - L^k = (1 - L)\left(1 + L + \cdots + L^{k-2} + L^{k-1}\right)}$$

άρα

$$|x_{n+k} - x_n| \le L^n \frac{1 - L^k}{1 - L}|x_1 - x_0| \Rightarrow \lim_{k\to\infty}|x_{n+k} - x_n| \le L^n \frac{1}{1 - L}|x_1 - x_0| \Rightarrow$$

$$\boxed{|x^* - x_n| \le L^n \frac{|x_1 - x_0|}{1 - L}}$$

Τέλος, αν θέσουμε $y_0 = x_{n-1}$ και $y_1 = \varphi(y_0) = \varphi(x_{n-1}) = x_n$ θα έχουμε

$$|x^* - y_1| \le \frac{L}{1 - L}|y_1 - y_0| \Rightarrow \boxed{|x^* - x_n| \le \frac{L}{1 - L}|x_n - x_{n-1}|}$$

Να τονισθεί ότι αν στο θεώρημα συστολής η συνθήκη Lipschitz ισχύει με $L = 1$, τότε η ακολουθία $\{x_n\}_{n\in\mathbb{N}}$ με $x_0 \in [a,b]$ δεν συγκλίνει αναγκαστικά. Παράδειγμα η $\varphi : [-1,+1] \to [-1,+1]$ με $x_0 \in [-1,+1]$ και $\varphi(x) = -x$, τότε παίρνουμε την ακολουθία $x_0, -x_0, +x_0, -x_0, \ldots$, η οποία δεν συγκλίνει ποτέ στην ρίζα $x^* = 0$.

<u>Σχόλια για τις εκτιμήσεις σφάλματος</u>:

i. Η εκτίμηση $\left|x_n - x^*\right| \le L^n \left|x_0 - x^*\right|$ δεν είναι πρακτικά χρήσιμη, γιατί το δεξί μέλος ($\left|x_0 - x^*\right|$) περιέχει το άγνωστο σημείο x^*.

ii. Η εκτίμηση $\left|x_n - x^*\right| \le \frac{L}{1-L}\left|x_n - x_{n-1}\right|$ είναι καλύτερη της εκτίμησης $\left|x_n - x^*\right| \le \frac{L^n}{1-L}\left|x_1 - x_0\right|$.

Πράγματι,

$$\left|x_n - x_{n-1}\right| = \left|\varphi\left(x_{n-1}\right) - \varphi\left(x_{n-2}\right)\right| \le L\left|x_{n-1} - x_{n-2}\right| \Rightarrow$$

$$\left|x_n - x_{n-1}\right| \le L\left|x_{n-1} - x_{n-2}\right| = L\left|\varphi\left(x_{n-2}\right) - \varphi\left(x_{n-3}\right)\right| \le L^2\left|x_{n-2} - x_{n-3}\right| \Rightarrow$$

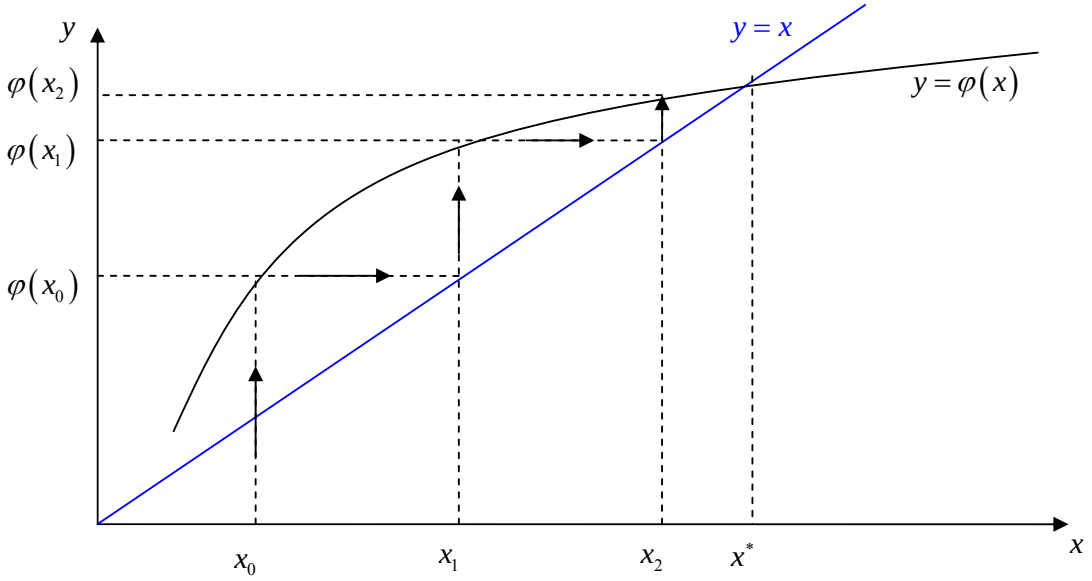
$$\left|x_n - x_{n-1}\right| \le L^{n-1}\left|x_1 - x_0\right|$$

Άρα,

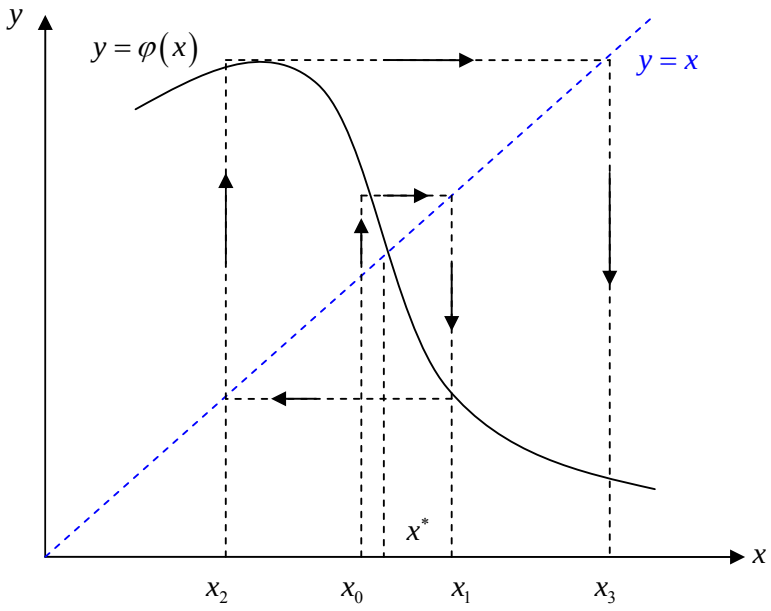
$$\left|x_n - x^*\right| \le \frac{L}{1-L}\left|x_n - x_{n-1}\right| \le \frac{L^n}{1-L}\left|x_1 - x_0\right|$$

iii. Η εκτίμηση $\left|x_n - x^*\right| \le \frac{L}{1-L}\left|x_n - x_{n-1}\right|$ (*) είναι εκτίμηση εκ των υστέρων, ενώ η $\left|x_n - x^*\right| \le \frac{L^n}{1-L}\left|x_1 - x_0\right|$ (**) είναι εκτίμηση εκ των προτέρων. Εκτίμηση εκ των υστέρων σημαίνει ότι για να υπολογίσουμε το άνω φράγμα του σφάλματος, δηλαδή το δεξί μέλος της ανισότητας (*), πρέπει να έχουμε υπολογίσει την ποσότητα που μας ενδιαφέρει (δηλαδή το x_n). Αντίθετα το δεξί μέλος της (**) μπορεί να υπολογιστεί από την αρχή για οποιαδήποτε τιμή του n (με δεδομένα φυσικά τα L, x_0 και αφού υπολογίσουμε το x_1).

Ακολουθούν ποιοτική περιγραφή μίας συγκλίνουσας και μίας αποκλίνουσας ακολουθίας. Τα βέλη δείχνουν την πορεία των υπολογισμών.



Σχήμα 1: Συγκλίνουσα ακολουθία $x_n = \varphi\left(x_{n-1}\right)$



Σχήμα 2: Αποκλίνουσα ακολουθία $x_n = \varphi\left(x_{n-1}\right)$

Έστω ότι θέλουμε να υπολογίσουμε την ρίζα, x^*, της μη-γραμμικής αλγεβρικής εξίσωσης $f\left(x\right)=0$ με την *γενική επαναληπτική μέθοδο* (ή *μέθοδο σταθερού σημείου*):

- ✓ αν δεν μας δίνεται πληροφορία σε πιο διάστημα βρίσκεται η ρίζα ψάχνουμε να βρούμε, κάνοντας δοκιμές, πραγματικούς αριθμούς a και b τέτοιους ώστε $\text{sgn}\left(f\left(a\right)\right) \neq \text{sgn}\left(f\left(b\right)\right) \Rightarrow \exists x^* \in \left(a,b\right)$ για το οποίο $f\left(x^*\right)=0$. Αν το διάστημα

δίνεται διαπιστώνουμε απλά κατά πόσο παραγματικά ισχύει $\operatorname{sgn}\left(f(a)\right) \neq \operatorname{sgn}\left(f(b)\right)$

- ✓ αποδεικνύουμε την μοναδικότητα της ρίζας εξετάζοντας αν η συνάρτηση είναι μονότονη (αύξουσα ή φθίνουσα) στο διάστημα $[a,b]$
- ✓ φέρνουμε την εξίσωση στη μορφή $x=\varphi(x)$ και δείχνουμε ότι:

  (α) $\varphi([a,b]) \subset [a,b]$, δηλαδή ότι $\forall x \in [a,b], \varphi(x) \in [a,b]$ και

  (β) ότι $\forall x, y \in [a,b]$, $|\varphi(x)-\varphi(y)| \leq L|x-y|$ όπου $0 \leq L < 1$, οπότε η φ είναι συστολή.

  Εναλλακτικά, αντί για το (β), διαπιστώνουμε αν $\varphi \in C^1[a,b]$ και υπολογίζουμε το $L \equiv \max_{x\in[a,b]} |\varphi'(x)|$. Αν $L<1$ τότε η φ είναι συστολή.

Έτσι, ικανοποιούνται όλες οι προϋποθέσεις του θεωρήματος της συστολής και επομένως η ακολουθία $\{x_n\}_{n\in\mathbb{N}}$ η οποία ορίζεται από τον αναδρομικό τύπο $x_{n+1}=\varphi(x_n)$ συγκλίνει στο μοναδικό σταθερό σημείο, x^*, της φ στο διάστημα $[a,b]$ για κάθε αρχική τιμή $x_0 \in [a,b]$.

- ✓ Επιλέγουμε ως x_0 το μέσο του διαστήματος, δηλαδή $x_0 = \frac{a+b}{2}$, όπως δηλαδή θα κάναμε με την μέθοδο της διχοτόμησης, και στην συνέχεια υπολογίζουμε τους όρους της ακολουθίας $\{x_n\}_{n\in\mathbb{N}}$ από την την αναδρομική σχέση $x_{n+1}=\varphi(x_n)$. Διακόπτουμε την διαδικασία όταν $|x_{n+1}-x_n| \leq \varepsilon$ όπου ε είναι η επιθυμητή ακρίβεια.

Πράγματι, έχουμε

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

άρα για την ακολουθία $\{\delta_n\}_{n\in\mathbb{N}}$ με $\delta_n \equiv |x_n-x_{n-1}|$ έχουμε ότι $\delta_{n+1} \leq L\delta_n < \delta_n$ επειδή $L<1$, άρα η ακολουθία φθίνει μονότονα. Αν τώρα υπάρχει $N \in \mathbb{N}$ τέτοιο ώστε $|x_{N+1}-x_N| \leq \varepsilon$ ($\delta_{N+1} \leq \varepsilon$), θα έχουμε ότι

$$|x_{n+1}-x^*| \leq \frac{L}{1-L}|x_{n+1}-x_n| \leq \frac{L\varepsilon}{1-L} \Rightarrow |x_{n+1}-x^*| \leq \frac{L\varepsilon}{1-L},\ \forall n \geq N$$

ή

$$\boxed{|\varepsilon_n| \leq \frac{L\varepsilon}{1-L}}$$

έχουμε δηλαδή ένα άνω φράγμα του σφάλματος της μορφής $C\varepsilon$. Έτσι αν σε κάποια επανάληψη επιτύχουμε $|x_n - x_{n-1}| \le \varepsilon$, τότε τότε το μέγιστο απόλυτο λάθος κατά τον προσδιορισμό της ρίζας x^* είναι $\frac{L\varepsilon}{1-L}$. Αν, επιπλέον, ισχύει $\frac{L}{1-L} < 1 \Rightarrow L < 1-L \Rightarrow 2L < 1 \Rightarrow L < \frac{1}{2}$, τότε το τελικό λάθος είναι $< \varepsilon$. Είναι προφανές ότι πρόβλημα υπάρχει μόνο στην περίπτωση που $L \to 1$ οπότε $\frac{L\varepsilon}{1-L} \to \infty$ οπότε και δεν μπορούμε να αποφανθούμε για την σύγκλιση ή όχι της ακολουθίας, που σημαίνει ότι η ακολουθία μπορεί να συγκλίνει αλλά μπορεί και όχι. Να τονισθεί επίσης ότι το ε που θα επιλέξουμε σε καμιά περίπτωση δεν πρέπει να είναι μικρότερο από το αντίστοιχο έψιλον της μηχανής ή, ακριβέστερα, μικρότερο από την απόσταση δύο διαδοχικών αριθμών κινητής υποδιαστολής.

Σε περίπτωση που μας ζητείται να διευρενήσουμε για ποιες αρχικές τιμές του x_0 η ακολουθία $\{x_n\}_{n\in\mathbb{N}}$ συγκλίνει, και υπό την προϋπόθεση ότι $\varphi \in C^1[a,b]$, αρκεί να λύσουμε την ανισότητα $|\varphi'(x)| < 1$. Οπότε, για κάθε x_0 το οποίο ανήκει στην τομή του διαστήματος που θα προκύψει με το διάστημα $[a,b]$ η συνάρτηση φ θα ικανοποιεί τις προϋποθέσεις του θεωρήματος της συστολής και επομένως η αντίστοιχη ακολουθία θα συγκλίνει στην ρίζα της εξίσωσης.

<u>*Παράδειγμα 1*</u>: Να βρεθεί η ρίζα της εξίσωσης $f(x) = 0$ όπου $f(x) = x - 2^{-x}$ στο διάστημα [0,1]. Είναι η ρίζα αυτή μοναδική?
Έχουμε

$$\left.\begin{aligned} f(0) &= 0 - 2^0 = -1 < 0 \\ f(1) &= 1 - 2^{-1} = \frac{1}{2} > 0 \end{aligned}\right\} \Rightarrow \exists x^* \in (0,1) :: f(x^*) = 0$$

Επίσης, $f'(x) = 1 - (-1)\ln(2)2^{-x} = 1 + \frac{\ln(2)}{2^x} > 0 \ \forall x \in [0,1]$, δηλαδή η f είναι γνησίως αύξουσα στο διάστημα αυτό και επομένως η ρίζα αυτή είναι μοναδική. Στην συνέχεια φέρνουμε την εξίσωση στην μορφή $x = \varphi(x)$ όπου $\varphi(x) = 2^{-x}$ η οποία είναι συνεχής στο διάστημα [0,1], έχει παράγωγο $\varphi'(x) = -\ln(2)2^{-x}$ η οποία είναι επίσης συνεχής και

αρνητική στο $[0,1]$. Αρα η φ είναι φθίνουσα και επομένως $\varphi([0,1]) = \left[\varphi(0), \varphi(1)\right] = \left[\frac{1}{2}, 1\right] \subset [0,1]$. Επιπλέον

$$L \equiv \max_{x\in[0,1]} \left|\varphi'(x)\right| = \ln(2) \max_{x\in[0,1]} \left|\frac{1}{2^x}\right| = \ln(2) \approx 0.693 < 1.$$

Έτσι οι προϋποθέσεις του θεωρήματος της συστολής ικανοποιούνται και η ακολουθία $x_{n+1} = 2^{-x_n}$ θα συγκλίνει για κάθε $x_0 \in [0,1]$.

Παράδειγμα 2: Έστω η συνάρτηση $f(x) = x - e^{-x}$. Να βρεθεί η ρίζα της $f(x) = 0$ στο [0.4,0.7].

Λύση: Έχουμε

$$\left.\begin{aligned} f(0.4) &= 0.4 - e^{-0.4} \approx -0.27 < 0 \\ f(0.7) &= 0.7 - e^{-0.7} \approx 0.20 > 0 \end{aligned}\right\} \Rightarrow \exists x^* \in (0.4, 0.7) :: f(x^*) = 0$$

Επίσης, $f'(x) = 1 + e^{-x} > 0$, $\forall x \in (0.4, 0.7) \Rightarrow f(x)$ αύξουσα $\Rightarrow$ το x^* είναι μοναδικό

Ορίζω $x = \varphi(x)$ με $\varphi(x) = e^{-x}$ και $\varphi'(x) = -e^{-x}$. Τότε έχουμε

$$\varphi(0.4) \approx 0.67$$
$$\varphi(0.7) \approx 0.50$$

άρα

$$\varphi([0.4, 0.7]) \approx [0.50, 0.67] \subset [0.4, 0.7]$$

Επιπλέον

$$\varphi'(x) = -e^{-x} \Rightarrow \left|-e^{-x}\right| < 1 \Rightarrow \left|e^{-x}\right| < 1 \Rightarrow -1 < e^{-x} < 1 \Rightarrow 0 < e^{-x} < 1 \Rightarrow -x < 0 \Rightarrow x > 0$$

Άρα $\forall x > 0$, $\left|\varphi'(x)\right| < 1$ και επομένως η φ είναι συστολή. Στο διάστημα $[0.4, 0.7]$ θα έχουμε $L \equiv \max_{x\in[0.4,0.7]} \left|\frac{1}{e^x}\right| = \frac{1}{e^{0.4}} \approx 0.67$. Επόμενως η $x_{n+1} = \varphi(x_n)$ συγκλίνει $\forall x_0 \in [0.4, 0.7]$.

Παράδειγμα 3: Έστω $f(x) = x^2 - 2x - 3$ για την οποία ζητείται $x^* \in [1,4]$ τέτοιο ώστε $f(x^*) = 0$. Να βρεθεί το x^* με την γενική επαναληπτική μέθοδο και $x_0 = 4$ (οι ρίζες της εξίσωσης είναι -1 και 3).

Λύση: Έχουμε

$$\left.\begin{array}{l} f\left(x=1\right)=1-2-3=-4<0 \\ f\left(x=4\right)=16-8-3=5>0 \end{array}\right\} \Rightarrow \exists x^{*}\in\left(1,4\right)::\, f\left(x^{*}\right)=0$$

Επιπλέον

$$f'\left(x\right)=2x-2\Rightarrow f'\left(x\right)=2\left(x-1\right)>0\,,\ \forall x\in\left(1,4\right)$$

άρα η f είναι αύξουσα $\Rightarrow$ έχει μόνο μία ρίζα στο $\left[1,4\right]$ και επομένως η ρίζα είναι μοναδική στο $I\equiv\left[1,4\right]$.

Μπορούμε να δημιουργήσουμε 3 τουλάχιστον μεθόδους σταθερού σημείου

a) $x=\sqrt{2x+3}\,,\ \varphi\left(x\right)=\sqrt{2x+3}\,,$

b) $x=\dfrac{x^{2}-3}{2}\,,\ \varphi\left(x\right)=\dfrac{x^{2}-3}{2}$

c) $x=\dfrac{3}{x-2}\,,\ \varphi\left(x\right)=\dfrac{3}{x-2}\,,$

Για την κάθε περίπτωση αρχικά πρέπει να εξετάσουμε κατά πόσο η συνάρτηση φ είναι συστολή.

✓ Για την 1η περίπτωση

$$\varphi'\left(x\right)=\frac{1}{\sqrt{2x+3}}$$

θα πρέπει

$$\left|\varphi'\left(x\right)\right|<1\Rightarrow -1<\frac{1}{\sqrt{2x+3}}<1\Rightarrow 0<\frac{1}{\sqrt{2x+3}}<1\Rightarrow \sqrt{2x+3}>1\Rightarrow 2x>-2\Rightarrow x>-1$$

Άρα $\forall x\in\left(-1,\infty\right)$ ισχύει $\left|\varphi'\left(x\right)\right|<1$. Επομένως, η φ είναι συστολή στο $\left[1,4\right]\subset\left(-1,\infty\right)$ και έτσι η $x_{n+1}=\varphi\left(x_{n}\right)$ θα συγκλίνει στη ρίζα αρκεί $\varphi\left(\left[1,4\right]\right)\subset\left[1,4\right]$. Πράγματι, $\varphi\left(1\right)\approx 2.23$ και $\varphi\left(4\right)\approx 3.31$ και εφόσον η φ είναι συνεχής και γνησίως αύξουσα στο [1,4] άρα $\varphi\left(\left[1,4\right]\right)\approx\left[2.23,3.31\right]\subset\left[1,4\right]$.

✓ Για την 2η περίπτωση

$\varphi'\left(x\right)=x\Rightarrow\left|\varphi'\left(x\right)\right|<1\Rightarrow\left|x\right|<1\Rightarrow -1<x<1$, άρα για $x_{0}=4$ η μέθοδος θα αποκλίνει.

✓ Για την 3η περίπτωση

$$\varphi'(x) = -\frac{3}{(x-2)^2}$$

η οποία φυσικά δεν ορίζεται για $x = 2$. Από την συνθήκη $|\varphi'(x)| < 1$ παίρνουμε

$$-1 < \frac{3}{(x-2)^2} < 1 \Rightarrow 0 < \frac{3}{(x-2)^2} < 1 \Rightarrow 3 < (x-2)^2 \Rightarrow (x-2)^2 > 3 \Rightarrow x^2 - 4x + 4 > 3 \Rightarrow$$

$$x^2 - 4x + 1 > 0 \Rightarrow x \in \left(-\infty, 2-\sqrt{3}\right) \cup \left(2+\sqrt{3}, +\infty\right) \approx (-\infty, 0.27) \cup (3.73, +\infty)$$

άρα για $x_0 = 4$ δεν μπορούμε να αποφασίσουμε αν θα συγκλίνει ή όχι. Η εφαρμογή αυτής της περίπτωσης στον υπολογιστή δίνει ότι η ακολουθία για $x_0 = 4$ θα συγκλίνει στην ρίζα $x^* = -1$. Επίσης για $x_0 \in (-\infty, 0.27)$ η ακολουθία θα συγκλίνει στην ρίζα -1.

2.5. Σύγκλιση και ταχύτητα σύγκλισης ακολουθιών

- Ορισμός συγκλίνουσας ακολουθίας

Λέμε ότι η ακολουθία $\{x_n\}_{n\in\mathbb{N}}$ συγκλίνει (τουλάχιστον) γραμμικά ή ότι η τάξη σύγκλισης είναι (τουλάχιστον) ένα, αν υπάρχουν $0 \le C < 1$ και $N \in \mathbb{N}$ τέτοια ώστε:

$$|x_{n+1} - x^*| \le C|x_n - x^*|, \quad \forall n \ge N \quad (\text{ή } |\varepsilon_{n+1}| \le C|\varepsilon_n|) \qquad \text{(A)}$$

Η τάξη σύγκλισης είναι $p > 1$ εάν $\exists C > 0$ και $N \in \mathbb{N}$ τέτοια ώστε

$$|x_{n+1} - x^*| \le C|x_n - x^*|^p, \quad \forall n \ge N \quad (\text{ή } |\varepsilon_{n+1}| \le C|\varepsilon_n|^p) \qquad \text{(B)}$$

- Αν $p = 2$: τετραγωνική σύγκλιση
- Αν $p = 3$: κυβική σύγκλιση

Προφανώς, όσο μεγαλύτερο είναι το p, τόσο πιο γρήγορα θα συγκλίνει η $\{x_n\}_{n\in\mathbb{N}}$ στο x^*. Επίσης, αν η τάξη σύγκλισης της ακολουθίας είναι p ($p > 1$), τότε η σύγκλιση θα είναι και οποιασδήποτε άλλης τάξης q με $1 \le q < p$. Να τονισθεί ότι το p μπορεί να είναι πραγματικός αριθμός.

Ο προσδιορισμός της τάξης σύγκλισης μίας ακολουθίας που παράγεται από επαναληπτικές μεθόδους είναι γενικά δύσκολος. Σχετικά εύκολος είναι ο προσδιορισμός όταν η αναδρομική σχέση είναι τέτοια ώστε ο καινούριος όρος της ακολουθίας να εξαρτάται μόνο από τον αμέσως προηγούμενο. Επιπλέον, ενδιαφέρουσα είναι η περίπτωση που $x_n \ne x^*, \forall n\mathbb{N}$ οπότε θα ισχύει $\varepsilon_n = |x_n - x^*| \ne 0, \forall n\mathbb{N}$. Τότε η

σχέση (Β) σημαίνει ότι η ακολουθία $\frac{\varepsilon_{n+1}}{\varepsilon_n^p}$ είναι φραγμένη. Μία ικανή συνθήκη για αυτό είναι η ακολουθία αυτή να συγκλίνει. Αν μάλιστα $\lim_{n\to\infty}\frac{\varepsilon_{n+1}}{\varepsilon_n^p}=\mu\neq 0$ τότε η τάξη σύγκλισης είναι ακριβώς p. Πράγματι, έστω ότι η τάξη σύγκλισης ήταν $p+a\,(a>0)$, άρα σύμφωνα με τον ορισμό θα είχαμε $|\varepsilon_{n+1}|\le C|\varepsilon_n|^{p+a}\Rightarrow\frac{|\varepsilon_{n+1}|}{|\varepsilon_n|^p}\le C|\varepsilon_n|^a$ και παίρνοντας όρια για $n\to\infty$ προκύπτει $\lim_{n\to\infty}\frac{|\varepsilon_{n+1}|}{|\varepsilon_n|^p}\le C\lim_{n\to\infty}|\varepsilon_n|^a\Rightarrow|\mu|\le 0$ το οποίο βέβαια είναι άτοπο και επομένως ή τάξη της ακολουθίας δεν μπορεί να είναι $p+a$ αλλά ακριβώς p. Η σταθερά $\mu\neq 0$ για την οποία η τάξη σύγκλισης είναι ακριβώς p ονομάζεται *«ασυμπτωρική σταθερά του σφάλματος»* της μεθόδου. Συνοπτικά:

Έστω μία συγκλίνουσα ακολουθία $\{x_n\}_{n\in\mathbb{N}}$ για την οποία ισχύει $\lim_{n\to\infty}\frac{\varepsilon_{n+1}}{\varepsilon_n^p}=\mu$. Αν $\mu=0$ τότε η τάξη σύγκλισης της ακολουθίας είναι τουλάχιστον p. Αν $\mu\neq 0$ η τάξη σύγκλισης είναι ακριβώς p. Όταν ισχύει $\mu\neq 0$ η σταθερά μ ονομάζεται *ασυμπτωτική σταθερά της μεθόδου.*

<u>Ταχύτητα σύγκλισης της γενικής επαναληπτικής μεθόδου</u>

Έστω η γενική επαναληπτική μέθοδος $x_{n+1}=\varphi(x_n)$ και ότι η φ ικανοποιεί τις υποθέσεις του θεωρήματος της συστολής, τότε είδαμε ότι

$$|x_{n+1}-x^*|=|\varphi(x_n)-\varphi(x^*)|\le L|x_n-x^*|$$

με $L\equiv\max_{x\in[a,b]}|\varphi'(x)|<1$. Δηλαδή,

$$|x_{n+1}-x^*|\le L|x_n-x^*|,\quad\forall n\ge 0\quad\text{ή}\quad\boxed{|\varepsilon_{n+1}|\le L|\varepsilon_n|}$$

Η σχέση αυτή μας λέει ότι η $\{x_n\}_{n\in\mathbb{N}}$ συγκλίνει, τουλάχιστον, γραμμικά στο σταθερό σημείο της φ, δηλαδή στο x^*. Αν λοιπόν ισχύει $x_n\neq x^*,\forall n$ τότε η τελευταία σχέση συνεπάγεται ότι $\frac{|\varepsilon_{n+1}|}{|\varepsilon_n|}\le L<1$. Ποιο όμως είναι το όριο της $\delta_n\equiv\frac{\varepsilon_{n+1}}{\varepsilon_n},\ n\to\infty$? Ας υποθέσουμε, επιπλέον των υποθέσεων του θεωρήματος συστολής, ότι $\varphi\in C^1[a,b]$, ότι

δηλαδή η φ είναι συνεχώς παραγωγίσιμη στο $[a,b]$. Σύμφωνα με το Θ.Μ.Τ. υπάρχει ξ_n μεταξύ των x_n και x^* τέτοιο ώστε

$$\varphi(x_n)-\varphi(x^*)=\varphi'(\xi_n)(x_n-x^*)\Rightarrow \varphi'(\xi_n)=\frac{\varphi(x_n)-\varphi(x^*)}{x_n-x^*}\Rightarrow$$

$$\lim_{n\to\infty}\varphi'(\xi_n)=\lim_{n\to\infty}\frac{\varphi(x_n)-\varphi(x^*)}{x_n-x^*}\Rightarrow \varphi'\left(\lim_{n\to\infty}\xi_n\right)=\lim_{n\to\infty}\frac{x_{n+1}-x^*}{x_n-x^*}\Rightarrow \varphi'(x^*)=\lim_{n\to\infty}\frac{\varepsilon_{n+1}}{\varepsilon_n}$$

Αν, λοιπόν, ισχύει $\varphi'(x^*)\neq 0$ και επειδή $|\varphi'(x^*)|\le L<1\Rightarrow \boxed{p=1}$, άρα η τάξη σύγκλισης είναι *ακριβώς* γραμμική.

2.6. Ακολουθίες υψηλής τάξης σύγκλισης

Έστω $f(x)=0\Rightarrow x=\varphi(x)$ και έστω η ακολουθία που ορίζεται από την αναδρομική σχέση

$$x_n=\varphi(x_{n-1}),\ n>0 \qquad \text{(i)}$$

Ορίζουμε επίσης την ποσότητα σφάλματος

$$\varepsilon_n=x_n-x^* \text{ και } \varepsilon_{n-1}=x_{n-1}-x^* \qquad \text{(ii)}$$

$$\underset{(ii)}{\overset{(i)}{\Rightarrow}}\varepsilon_n+x^*=\varphi(\varepsilon_{n-1}+x^*)\Rightarrow \varepsilon_n+x^*=\varphi(x^*)+\varepsilon_{n-1}\varphi'(x^*)+\frac{1}{2}\varepsilon_{n-1}^2\varphi''(x^*)+O(\varepsilon_{n-1}^3)$$

ή αν $\varphi\in C^k[a,b]$, τότε

$$\varepsilon_n+x^*=\varphi(x^*)+\varepsilon_{n-1}\varphi'(x^*)+\frac{1}{2}\varepsilon_{n-1}^2\varphi''(x^*)+\cdots+\frac{1}{k!}\varepsilon_{n-1}^k\varphi^{(k)}(\xi_{n-1})$$

όπου ξ_{n-1} μεταξύ των x^* και x_{n-1}. Επειδή $x^*=\varphi(x^*)$, άρα έχουμε την *εξίσωση σφάλματος*

$$\boxed{\varepsilon_n=\varepsilon_{n-1}\varphi'(x^*)+\frac{1}{2}\varepsilon_{n-1}^2\varphi''(x^*)+\cdots+\frac{1}{k!}\varepsilon_{n-1}^k\varphi^{(k)}(\xi_{n-1})} \qquad (*)$$

Η σχέση (*) μας λέει τι θα συμβεί αν $\varphi'(x^*)=0$. Αν επιπλέον $k=2$ δηλαδή αν $\varphi\in C^2[a,b]$, τότε:

$$\varepsilon_n=\frac{1}{2}\varepsilon_{n-1}^2\varphi''(\xi_{n-1})\Rightarrow \frac{\varepsilon_n}{\varepsilon_{n-1}^2}=\frac{1}{2}\varphi''(\xi_{n-1})\Rightarrow \lim_{n\to\infty}\frac{\varepsilon_n}{\varepsilon_{n-1}^2}=\frac{1}{2}\lim_{n\to\infty}\varphi''(\xi_{n-1})\Rightarrow$$

$$\Rightarrow \lim_{n\to\infty}\frac{\varepsilon_n}{\varepsilon_{n-1}^2}=\frac{1}{2}\varphi''\left(\lim_{n\to\infty}\xi_{n-1}\right)=\frac{1}{2}\varphi''(x^*)$$

Η τελευταία σχέση σημαίνει ότι υπάρχει $N \in \mathbb{N}$ τέτοιος ώστε $|\varepsilon_n| \le C|\varepsilon_{n-1}|^2$, $\forall n \ge N$ όπου $C > \frac{1}{2}|\varphi''(x^*)|$ και επομένως η σύγκλιση της μεθόδου θα είναι τετραγωνική.

Απαραίτητη λοιπόν προϋπόθεση για να έχουμε τάξη σύγκλισης της ακολουθίας μεγαλύτερη του ένα είναι να ισχύει $\varphi'(x^*) = 0$.

Να σημειωθεί ότι αν ισχύει

$$\varphi'(x^*) = \varphi''(x^*) = \ldots = \varphi^{(k-1)}(x^*) = 0, \quad \varphi^{(k)}(x^*) \neq 0 \qquad (**)$$

τότε από την (*) θα πάρουμε

$$\varepsilon_n = \frac{1}{k!}\varepsilon_{n-1}^k \varphi^{(k)}(\xi_{n-1}) \Rightarrow \frac{\varepsilon_n}{\varepsilon_{n-1}^k} = \frac{1}{2}\varphi^{(k)}(\xi_{n-1}) \Rightarrow \lim_{n\to\infty}\frac{\varepsilon_n}{\varepsilon_{n-1}^k} = \frac{1}{k!}\lim_{n\to\infty}\varphi^{(k)}(\xi_{n-1}) \Rightarrow$$

$$\Rightarrow \lim_{n\to\infty}\frac{\varepsilon_n}{\varepsilon_{n-1}^k} = \frac{1}{k!}\varphi^{(k)}\left(\lim_{n\to\infty}\xi_{n-1}\right) = \frac{1}{k!}\varphi^{(k)}(x^*)$$

και επομένως η μέθοδος που θα προκύψει σε αυτήν την περίπτωση είναι τάξης k με ασυμπτωτική σταθερά σφάλματος $C = \frac{1}{k!}\varphi^{(k)}(x^*)$. Το να μπορέσουμε όμως να βρούμε κατάλληλη συνάρτηση φ η οποία να ικανοποιεί όλες τις συνθηκες της (**) είναι εξαιρετικά δύσκολο. Για τον λόγο αυτό συνήθως περιοριζόμαστε σε μεθόδους που έχουν τάξη το πολύ 2 ή 3. Θα προσπαθήσουμε στην συνέχεια να κατασκευάσουμε μία μέθοδο με τάξη σύγκλισης μεγαλύτερης της μονάδας ($p > 1$) .

2.7. Επαναληπτική μέθοδος Newton-Raphson

Έστω $f(x) = 0 \Rightarrow \lambda f(x) = 0 \Rightarrow x = \underbrace{x + \lambda f(x)}_{\varphi(x)}$ για $\lambda \neq 0$, οπότε έχουμε $x = \varphi(x)$ με $\varphi(x) = x + \lambda f(x)$. Αν ισχύει $\varphi'(x^*) = 0$ τότε η τάξη σύγκλισης της μεθόδου θα είναι $p > 1$. Πράγματι, έχουμε

$$\left.\begin{array}{l} \varphi'(x) = 1 + \lambda f'(x) \\ \varphi'(x^*) = 0 \end{array}\right\} \Rightarrow \lambda = -\frac{1}{f'(x^*)}$$

όπου για να ορίζεται το λ θα πρέπει να ισχύει $f'(x^*) \neq 0$ και επομένως η ρίζα x^* πρέπει να είναι απλή ρίζα. Έτσι θα έχουμε $\varphi(x) = x - \frac{f(x)}{f'(x^*)}$ και η αντίστοιχη

επαναληπτική μέθοδος που παράγεται είναι η $x_{n+1} = x_n - \frac{f(x_n)}{f'(x^*)}$. Το πρόβλημα με την μέθοδο αυτή όμως είναι ότι το δεξί μέλος περιέχει την ποσότητα που θέλουμε να υπολογίσουμε, δηλαδή το x^*! Επομένως δεν έχει καμία πρακτική αξία. Ευτυχώς μπορούμε εύκολα να την διορθώσουμε. Αν αντί για $f'(x^*)$ θέσουμε $f'(x_n)$ τότε θα έχουμε $\boxed{x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}}$, δηλαδή μία αναδρομική σχέση με αντίστοιχη συνάρτηση επανάληψης $\varphi(x) = x - \frac{f(x)}{f'(x)}$. Κατά πόσο όμως συνεχίζει να ισχύει $\varphi'(x^*) = 0$? Ας εξετάσουμε κατά πόσο η συνθήκη $\varphi'(x^*) = 0$ συνεχίζει να ισχύει.

$$\varphi'(x) = 1 - \frac{f'(x)}{f'(x)} + \frac{f(x) f''(x)}{\left[f'(x)\right]^2} = \frac{f(x) f''(x)}{\left[f'(x)\right]^2} \Rightarrow \varphi'(x^*) = \frac{f(x^*) f''(x^*)}{\left[f'(x^*)\right]^2} = 0$$

διότι $f(x^*) = 0$. Άρα η τροποποιημένη μέθοδος ικανοποιεί την αναγκαία συνθήκη ώστε η τάξη σύγκλισης να είναι μεγαλύτερης της μονάδας. Η μέθοδος που παρήγαμε λέγεται μέθοδος του Νεύτωνα (ή μέθοδος Newton-Raphson).

2.7.1. Γεωμετρική ερμηνεία της Newton-Raphson

Έστω, $(x_n, f(x_n))$ ένα σημείο της $y = f(x)$. Η εφαπτομένη της καμπύλης στο σημείο αυτό έχει εξίσωση:

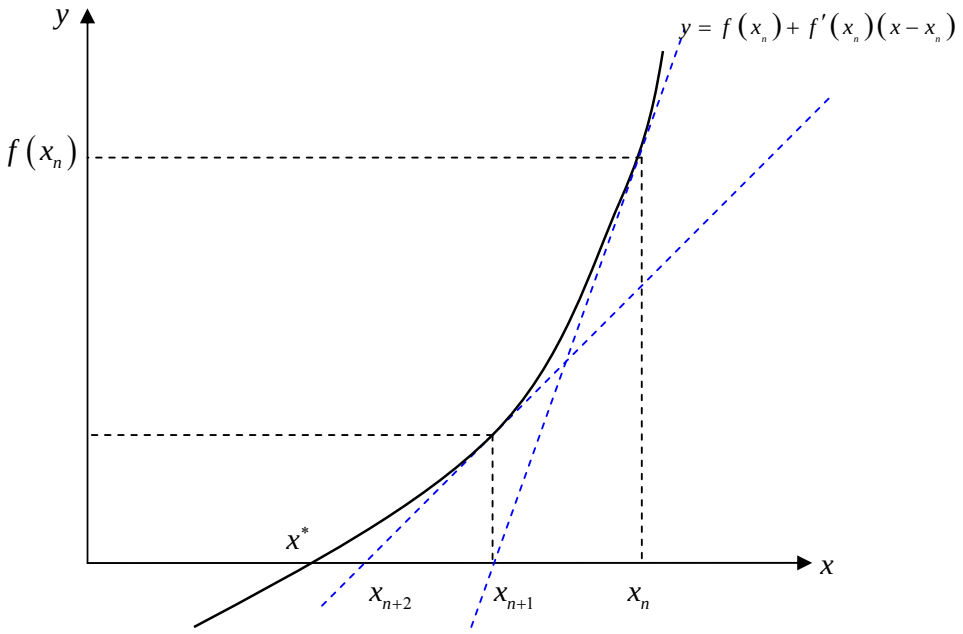
$$y = f(x_n) + f'(x_n)(x - x_n)$$

Φέρνουμε την ευθεία αυτή ώσπου να τμήσουμε τον άξονα x, οπότε στο σημείο τομής x_{n+1} θα ισχύει $y = 0$. Δηλαδή,

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

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

Επαναλαμβάνουμε τη διαδικασία για το νέο σημείο $(x_{n+1}, f(x_{n+1}))$ και παράγουμε το x_{n+2}. Συνεχίζουμε μέχρι να βρούμε x_k τέτοιο ώστε $|x_k - x_{k-1}| < \varepsilon$, όπου ε η ζητούμενη ακρίβεια. Προϋπόθεση φυσικά για την εφαρμογή της μεθόδου είναι ότι $f(x_n) \neq 0 \ \ \forall x_n$.



➢ *Εναλλακτικός τρόπος παραγωγής της Newton-Raphson με σειρά Taylor*

Έχουμε $f(x^*)=0$. Ορίζω $\varepsilon_n = x_n - x^* \Rightarrow x^* = x_n - \varepsilon_n$ και κάνω ανάπτυγμα Taylor γύρω από το σημείο x_n

$$f(x^*)=0 \Rightarrow f(x_n-\varepsilon_n)=0 \Rightarrow f(x_n)-\varepsilon_n f'(x_n)+\frac{1}{2}\varepsilon_n^2 f''(\xi_n)=0 \Rightarrow$$

$$\varepsilon_n f'(x_n) = f(x_n)+\frac{1}{2}\varepsilon_n^2 f''(\xi_n)=0 \Rightarrow \varepsilon_n \equiv x_n - x^* = \frac{f(x_n)}{f'(x_n)}+\frac{1}{2}\varepsilon_n^2\frac{f''(\xi_n)}{f'(x_n)} \Rightarrow$$

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

όπου ξ_n στο διάστημα των x_n και x^*. Εάν $\varepsilon_n \ll 1$, όταν δηλαδή βρισκόμαστε κοντά στην ρίζα, τότε μπορούμε να αμελήσουμε τον όρο $-\frac{1}{2}\varepsilon_n^2\frac{f''(\xi_n)}{f'(x_n)} \ll 1$, οπότε

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

2.7.2 Απόδειξη σύγκλισης της Newton-Raphson (τοπικής σύγκλισης)

Πρόταση: Έστω μια συνεχής συνάρτηση $f:\mathbb{R}\to\mathbb{R}$ και ότι $\exists x^*\in\mathbb{R}$ τέτοιο ώστε $f\left(x^*\right)=0$, με x^* απλή ρίζα, $f'\left(x^*\right)\neq 0$ και η f να είναι δύο φορές συνεχώς παραγωγίσιμη σε ένα μικρό διάστημα κοντά στο x^*. Τότε υπάρχει κλειστό διάστημα I με μέσον το x^* τέτοιο ώστε $\forall x_0\in I$, η ακολουθία $\left\{x_n\right\}_{n\in\mathbb{N}}$ που κατασκευάζεται με την μέθοδο του Νεύτωνα να συγκλίνει τετραγωνικά, δηλαδή $\lim\limits_{n\to\infty}\dfrac{\varepsilon_{n+1}}{\varepsilon_n^2}=\dfrac{f''\left(x^*\right)}{2f'\left(x^*\right)}$.

Απόδειξη: Η συνάρτηση επανάληψης είναι η

$$\varphi(x)=x-\frac{f(x)}{f'(x)}\Rightarrow\varphi'(x)=1-\frac{f'(x)}{f'(x)}+\frac{f(x)f''(x)}{\left[f'(x)\right]^2}=\frac{f(x)f''(x)}{\left[f'(x)\right]^2}$$

Εφόσον η $f\in C^2(I)$, άρα και η $\varphi'(x)=\dfrac{f(x)f''(x)}{\left[f'(x)\right]^2}$ είναι συνεχής, δηλαδή η φ είναι συνεχώς παραγωγίσιμη σε διάστημα του οποίου το μέσον είναι το x^*, τέτοιο ώστε $\max\limits_{x\in I}\left|\varphi'(x)\right|\equiv L<1$. Τότε, $\forall x\in I$ έχουμε

$$\left|\varphi(x)-x^*\right|=\left|\varphi(x)-\varphi\left(x^*\right)\right|\le L\left|x-x^*\right|\le\left|x-x^*\right|$$

και συνεπώς $\varphi(x)\in I$. Εφόσον η φ απεικονίζει το I στον εαυτό του και είναι συστολή, οπότε ικανοποιεί τις προϋποθέσεις του θεωρήματος της συστολής και άρα $\forall x_0\in I$ η ακολουθία $\left\{x_n\right\}_{n\in\mathbb{N}}$ συγκλίνει στο x^*, το οποίο είναι σταθερό σημείο της φ. Επιπλέον, ισχύει

$$f\left(x_n\right)=f\left(x^*+\varepsilon_n\right)=f\left(x^*\right)+\varepsilon_n f'\left(x_n\right)+\frac{1}{2}\varepsilon_n^2 f''\left(\xi_{n1}\right)$$

$$f'\left(x_n\right)=f'\left(x^*+\varepsilon_n\right)=f'\left(x^*\right)+\varepsilon_n f''\left(\xi_{n2}\right)$$

Έτσι από τον ορισμό της αναδρομικής σχέσης και τις παραπάνω δύο σχέσεις έχουμε

$$x_{n+1}=x_n-\frac{f\left(x_n\right)}{f'\left(x_n\right)}\Rightarrow\underbrace{x_{n+1}-x^*}_{\varepsilon_{n+1}}=\underbrace{x_n-x^*}_{\varepsilon_n}-\frac{f\left(x_n\right)}{f'\left(x_n\right)}\Rightarrow\varepsilon_{n+1}=\varepsilon_n-\frac{f\left(x^*\right)+\varepsilon_n f'\left(x^*\right)+\frac{1}{2}\varepsilon_n^2 f''\left(\xi_{n1}\right)}{f'\left(x^*\right)+\varepsilon_n f''\left(\xi_{n2}\right)}$$

$$\Rightarrow\varepsilon_{n+1}=\varepsilon_n^2\frac{f''\left(\xi_{n2}\right)-\frac{1}{2}f''\left(\xi_{n1}\right)}{f'\left(x^*\right)+\varepsilon_n f''\left(\xi_{n2}\right)}\Rightarrow\lim_{n\to\infty}\frac{\varepsilon_{n+1}}{\varepsilon_n^2}=\frac{f''\left(x^*\right)-\frac{1}{2}f''\left(x^*\right)}{f'\left(x^*\right)}=\frac{f''\left(x^*\right)}{2f'\left(x^*\right)}$$

Αν επιπλέον $f''\left(x^*\right)\neq 0$, τότε η σύγκλιση της μεθόδου είναι ακριβώς 2. Στην παραπάνω απόδειξη έχει χρησιμοποιηθεί ότι αφού η μέθοδος συγκλίνει $\lim_{n\to\infty} x_n = x^* \Rightarrow \lim_{n\to\infty} \varepsilon_n = 0$ και εφόσον τα ξ_{n1} και ξ_{n2} βρίσκονται στο διάστημα που ορίζουν τα x_n, x^* άρα $\lim_{n\to\infty} \xi_{n1} = \lim_{n\to\infty} \xi_{n2} = x^*$.

2.7.3. Newton-Raphson για πολλαπλές ρίζες

(α) Περίπτωση όπου η πολλαπλότητα της ρίζας είναι γνωστή.

Εάν η πολλαπλότητα της $f\left(x\right)=0$ είναι $m>1$, τότε η σύγκλιση παύει να είναι τετραγωνική. Ορίζουμε συνάρτηση g τέτοια ώστε $f\left(x\right)=\left(x-x^*\right)^m g\left(x\right)$ με $g\left(x^*\right)\neq 0$. Άρα,

$$f'\left(x\right)=m\left(x-x^*\right)^{m-1} g\left(x\right)+\left(x-x^*\right)^m g'\left(x\right)\Rightarrow f'\left(x^*\right)=0$$

$$\vdots$$

$$f^{(k)}\left(x^*\right)=0,\qquad k=0,1,2,\ldots,m-1$$

και

$$f^{(m)}\left(x^*\right)\neq 0$$

άρα από το ανάπτυγμα Taylor της συνάρτησης f γύρω από το x^* θα έχουμε

$$f\left(x_n\right)=f\left(x^*+\varepsilon_n\right)=f\left(x^*\right)+f'\left(x^*\right)\varepsilon_n+\cdots+f^{(m-1)}\left(x^*\right)\frac{\varepsilon_n^{m-1}}{(m-1)!}+f^{(m)}\left(\xi_{n1}\right)\frac{\varepsilon_n^m}{m!}=f^{(m)}\left(\xi_{n1}\right)\frac{\varepsilon_n^m}{m!}$$

και όμοια από το ανάπτυγμα Taylor της συνάρτησης f' γύρω από το x^*

$$f'\left(x_n\right)=f^{(m)}\left(\xi_{n2}\right)\frac{\varepsilon_n^{m-1}}{(m-1)!}=0$$

όπου τα ξ_{n1} και ξ_{n2} βρίσκονται στο διάστημα που ορίζουν τα x_n και x^*. Τα δύο παραπάνω αναπτύγματα φυσικά προϋποθέτουν ότι η συνάρτηση f είναι m συνεχώς παραγωγίσιμη σε μία περιοχή κοντά στο x^*. Έτσι από τον ορισμό της αναδρομικής σχέσης και τις δύο τελευταίες εκφράσεις (δηλαδή τα αναπτύγματα Taylor) έχουμε:

$$\varepsilon_{n+1}=\varepsilon_n-\frac{f^{(m)}(\xi_{n1})\frac{\varepsilon_n^m}{m!}}{f^{(m)}(\xi_{n2})\frac{\varepsilon_n^{m-1}}{(m-1)!}}=\varepsilon_n\left(1-\frac{f^{(m)}(\xi_{n1})}{mf^{(m)}(\xi_{n2})}\right)\Rightarrow\frac{\varepsilon_{n+1}}{\varepsilon_n}=1-\frac{f^{(m)}(\xi_{n1})}{mf^{(m)}(\xi_{n2})}$$

Εφόσον η μέθοδος συγκλίνει έχουμε $\lim_{n\to\infty}f^{(m)}(\xi_{n1})=f^{(m)}\left(\lim_{n\to\infty}\xi_{n1}\right)=f^{(m)}(x^*)$ και $\lim_{n\to\infty}f^{(m)}(\xi_{n2})=f^{(m)}\left(\lim_{n\to\infty}\xi_{n2}\right)=f^{(m)}(x^*)$ οπότε η παραπάνω σχέση δίνει:

$$\lim_{n\to\infty}\frac{\varepsilon_{n+1}}{\varepsilon_n}=1-\frac{1}{m}\neq 0\quad\text{αν}\quad m\neq 1!$$

Είναι εύκολο να διαπιστώσουμε ότι αν εφαρμόσουμε μία τροποποίηση της μεθόδου ως εξής:

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

τότε η τετραγωνική σύγκλιση επανέρχεται.

(b) Περίπτωση όπου η πολλαπλότητα της ρίζας δεν είναι γνωστή.

Έστω τώρα ότι δεν ξέρουμε την πολλαπλότητα της ρίζας. Ορίζουμε την συνάρτηση $\omega(x)=\frac{f(x)}{f'(x)}$. Επειδή $f(x)=(x-x^*)^m g(x)$, με $g(x^*)\neq 0$ και

$f'(x)=m(x-x^*)^{m-1}g(x)+(x-x^*)^m g'(x)$ άρα

$$\omega(x)=\frac{(x-x^*)^m g(x)}{m(x-x^*)^{m-1}g(x)+(x-x^*)^m g'(x)}=\frac{(x-x^*)g(x)}{mg(x)+(x-x^*)g'(x)}$$

Παρατηρούμε ότι $\omega(x^*)=\frac{0}{mg(x^*)}=0$, δηλαδή το x^* είναι ρίζα της συνάρτησης ω.

Δείχνουμε στην συνέχεια ότι το x^* είναι απλή ρίζα της ω. Αρκεί να ισχύει $\omega'(x^*)\neq 0$. Πράγματι,

$$\omega'(x)=\frac{g(x)+(x-x^*)g'(x)}{mg(x)+(x-x^*)g'(x)}-\frac{(x-x^*)g(x)\left[mg'(x)+g'(x)+(x-x^*)g''(x)\right]}{\left[mg(x)+(x-x^*)g'(x)\right]^2}\Rightarrow$$

$$\omega'(x^*)=\frac{g(x^*)}{mg(x^*)}-\frac{0}{\left[mg(x^*)\right]^2}=\frac{1}{m}\neq 0$$

Αρα η πολλάπλότητα της ρίζας x^* για την συνάρτηση ω είναι 1. Οπότε μπορούμε να εφαρμόσουμε την μέθοδο Newton-Raphson στην συνάρτηση ω για την οποία έχουμε,

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

$$\omega'(x) = \left(\frac{f(x)}{f'(x)} \right)' = \frac{f'(x)}{f'(x)} - \frac{f''(x) f(x)}{\left[f'(x) \right]^2} = 1 - \frac{f''(x) f(x)}{\left[f'(x) \right]^2}$$

άρα

$$\boxed{x_{n+1} = x_n - \frac{\dfrac{f(x_n)}{f'(x_n)}}{1 - \dfrac{f''(x_n) f(x_n)}{\left[f'(x_n) \right]^2}} = x_n - \frac{f(x_n) f'(x_n)}{\left(f'(x_n) \right)^2 - f(x_n) f''(x_n)}}$$

Έτσι, η παραπάνω μέθοδος θα συγκλίνει τετραγωνικά. Το μειονέκτημα της τροποποίησης της οποίας κάναμε είναι ότι αυξήθηκε η πολυπλοκότητα της μεθόδου (παρατηρείστε ότι εμφανίστηκε και η δεύτερη παράγωγος της συνάρτησης της οποίας ζητούμε την ρίζα).

2.8. Μέθοδος εφαπτομένης (ή τέμνουσας)

Η μέθοδος της εφαπτομένης προκύπτει από την μέθοδο Newton-Raphson

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

αν στη θέση της παραγώγου $f'(x_n)$ αντικαταστήσουμε την προσεγγιστική έκφραση

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

Έτσι, παίρνουμε τη μέθοδο

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

Προφανώς, η μέθοδος για να ξεκινήσει απαιτεί δύο σημεία, τα x_0 και x_1, καθώς και τις αντίστοιχες τιμές της συνάρτησης $f(x_0)$ και $f(x_1)$. Συνοπτικά:

- Η τάξη σύγκλισης της μεθόδου είναι $p = \dfrac{1+\sqrt{5}}{2} \approx 1.62 < 2$, άρα η σύγκλιση είναι πιο γρήγορη σε σχέση με την γενική επαναληπτική μέθοδο αλλά πιο αργή σε σχέση μη την μέθοδο Newton-Raphon.

- Χρειάζεται δύο αρχικές τιμές για να ξεκινήσει.
- Δεν απαιτείται γνώση της παραγώγου της συνάρτησης

Γεωμετρική ερμηνεία: Έστω η ευθεία που περνάει από τα σημεία $\left(x_n, f\left(x_n\right)\right)$, $\left(x_{n-1}, f\left(x_{n-1}\right)\right)$, η οποία είναι προσέγγιση της εφαπτομένης της $y = f\left(x\right)$ στο σημείο $\left(x_n, f\left(x_n\right)\right)$. Επεκτείνουμε την ευθεία αυτή ωσπου να τέμνει τον άξονα x. Η τιμή x_{n+1} στην οποία τέμνει τον άξονα είναι μία προσέγγιση της λύσης της εξίσωσης $f\left(x\right)=0$. Η διαδικασία επαναλαμβάνεται μέχρι $\left|x_{n+1}-x_n\right|<\varepsilon$. Η ευθεία η οποία περνά από τα σημεία $\left(x_{n-1}, f\left(x_{n-1}\right)\right)$ και $\left(x_n, f\left(x_n\right)\right)$ είναι η

$$p\left(x\right)=y=f\left(x_n\right)+\frac{f\left(x_n\right)-f\left(x_{n-1}\right)}{\left(x_n-x_{n-1}\right)}\left(x-x_n\right)$$

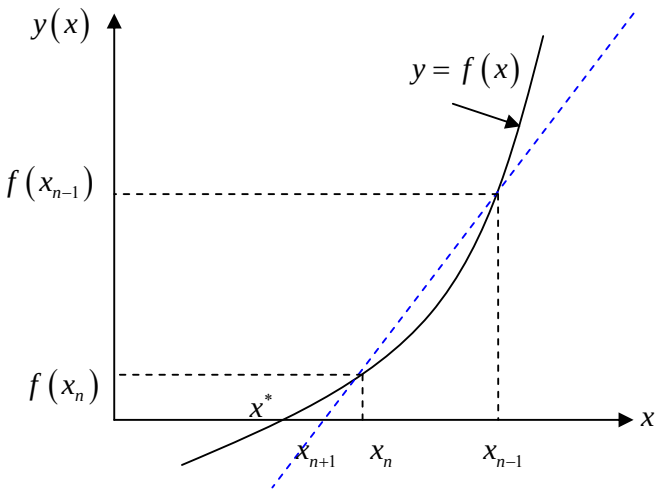
Εύκολα διαπιστώνουμε ότι

- Για $x=x_n$, $y=f\left(x_n\right)$
- Για $x=x_{n-1}$, $y=f\left(x_n\right)+\frac{f\left(x_n\right)-f\left(x_{n-1}\right)}{\left(x_n-x_{n-1}\right)}\left(x_{n-1}-x_n\right)=f\left(x_{n-1}\right)$

Επιπλέον, το σημείο στο οποίο τέμνει τον άξονα η ευθεία αυτή είναι το σημείο $\left(x_{n+1}, p\left(x_{n+1}\right)\right)=\left(x_{n+1},0\right)$. Έτσι, για $y=0$, τότε

$$f\left(x_n\right)+\frac{f\left(x_n\right)-f\left(x_{n-1}\right)}{\left(x_n-x_{n-1}\right)}\left(x_{n+1}-x_n\right)=0\Rightarrow x_{n+1}-x_n=-\frac{f\left(x_n\right)\left(x_n-x_{n-1}\right)}{f\left(x_n\right)-f\left(x_{n-1}\right)}\Rightarrow$$

$$\boxed{x_{n+1}=x_n-\frac{f\left(x_n\right)\left(x_n-x_{n-1}\right)}{f\left(x_n\right)-f\left(x_{n-1}\right)}}$$ "Μέθοδος εφαπτομένης (ή τέμνουσας)"



Θεώρημα Τέμνουσας: Έστω x^* ρίζα της $f(x)=0$ και έστω $(a,b)\subset\mathbb{R}$ με $x^*\in(a,b)$, $f\in C^2(a,b)$, $f'(x^*)\neq 0$ και $f''(x^*)\neq 0$. Τότε υπάρχει διάστημα I και $x^*\in I$, τέτοιο ώστε $\forall x_0, x_1\in I$ με $x_0\neq x_1$ η ακολουθία $\{x_n\}_{n\in\mathbb{N}}$ η οποία ορίζεται από την αναδρομική σχέση

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

και για τις αρχικές τιμές x_0 και x_1 να είναι καλά ορισμένη και να συγκλίνει στο x^*. Η τάξη σύγκλισης είναι $p=\frac{1+\sqrt{5}}{2}\approx 1.62$.

Αν η ρίζα την οποία ζητάμε έχει πολλαπλότητα άρτιου αριθμού, τότε η μέθοδος διχοτόμησης και της εφαπτομένης (ή τέμνουσας) δεν μπορούν να εφαρμοσθούν. Όταν μπορούμε να εφαρμόσουμε την μέθοδο της διχοτόμησης ή την μέθοδο της τέμνουσας αυτές συγκλίνουν πάντα, έστω και αργά. Επίσης, η Newton-Raphson συγκλίνει αρκετά γρήγορα, αλλά όχι πάντα, πρέπει η αρχική προσέγγιση να είναι κοντά στην ρίζα.

Κεφάλαιο 3
Επίλυση συστημάτων γραμμικών εξισώσεων

3.1. Εισαγωγή

Σε αυτό το κεφάλαιο θα μελετηθεί η λύση συστήματος γραμμικών αλγεβρικών εξισώσεων.

Εάν τρεις άγνωστες μεταβλητές x, y και z σχετίζονται μεταξύ τους με εξισώσεις τη μορφής

$$\begin{cases} a_1 x + b_1 y + c_1 z = d_1 \\ a_2 x + b_2 y + c_2 z = d_2 \\ a_3 x + b_3 y + c_3 z = d_3 \end{cases} \qquad (*)$$

όπου a_i, b_i, c_i και d_i $(i = 1, 2, 3)$ σταθερές που δίνονται, τότε οι εξισώσεις (*) αποτελούν ένα σύστημα τριών γραμμικών εξισώσεων. Εφόσον το σύστημα των τριών εξισώσεων περιέχει τρεις αγνώστους, αναμένεται ότι υπάρχει μοναδική λύση, δηλαδή ότι υπάρχουν μοναδικές τιμές των x, y και z οι οποίες ικανοποιούν το σύστημα (*). Θα αποδειχτεί ότι αυτό δεν είναι πάντα αληθές και ο σκοπός στο πρώτο μέρος του κεφαλαίου είναι να επανεξεταστεί κατάλληλα η λύση των γραμμικών εξισώσεων συμεριλαμβάνοντας συνθήκες για την ύπαρξη τέτοιων λύσεων. Για ευκολία οι γραμμικές εξισώσεις αναπαρίστανται υπό μορφή πινάκων. Στην γενική του μορφή ένα γραμμικό σύστημα εξισώσεων έχει την μορφή

$$\left.\begin{matrix} a_{1,1}x_1 + a_{1,2}x_2 + \ldots + a_{1,n}x_n = b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \ldots + a_{2,n}x_n = b_2 \\ \vdots \qquad\qquad\qquad\qquad \vdots \\ a_{n,1}x_1 + a_{n,2}x_2 + \ldots + a_{n,n}x_n = b_n \end{matrix}\right\} \Rightarrow \underbrace{\begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \ldots & a_{1,n} \\ a_{2,1} & a_{2,2} & a_{2,3} & \ldots & a_{2,n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n,1} & a_{n,2} & a_{n,3} & \ldots & a_{n,n} \end{bmatrix}}_{\underline{\underline{A}}} \cdot \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{bmatrix}}_{\underline{x}} = \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_n \end{bmatrix}}_{\underline{b}}$$

ή σε μορφή πινάκων $\underline{\underline{A}} \cdot \underline{x} = \underline{b}$. Σε κάθε περίπτωση, τόσο ο πίνακας $\underline{\underline{A}}$ όσο και το διάνυσμα $\underline{b}$ θεωρούνται ως δεδομένα του προβλήματος. Ανάλογα με την διάσταση του πίνακα $\underline{\underline{A}}$, το πρόβλημα θεωρείται ότι είναι:

(α) μικρού μεγέθους αν $n < 100$

(β) μεσαίου μεγέθους αν $100 < n < 1000$ και

(γ) μεγάλου μεγέθους αν $n > 1000$

Επίσης ο πίνακας $\underline{\underline{A}}$ διακρίνεται σε

(α) *πυκνό* ή *αποθηκεύσιμο* όταν περιέχει μη-μηδενικά σχεδόν παντού σε όλες τις γραμμές και τις στήλες του και

(β) *αραιό* ή *σποραδικό* όταν έχει συγκεκριμένη δομή και τα περισσότερα στοιχεία του είναι μηδενικά.

Οι αριθμητικές μέθοδοι για τους αραιούς πίνακες είναι συνήθως ειδικής μορφής ώστε να εκμεταλλεύονται τα πολλά μηδενικά στοιχεία και την δομή του πίνακα. Αυτό έχει σαν αποτέλεσμα οι αντίστοιχοι αλγόριθμοι να είναι πολύ πιο γρήγοροι σε σχέση με τους αντίστοιχους αλγόριθμους για τους πυκνούς πίνακες.

3.2. Ο αλγόριθμος της πίσω αντικατάστασης

Ο αλγόριθμος της πίσω αντικατάστασης χρησιμοποιείται για την αποδοτική επίλυση γραμμικών συστημάτων της μορφής $\underline{\underline{U}} \cdot \underline{x} = \underline{b}$ όπου $\underline{\underline{U}} \in \mathbb{R}^{n,n}$ είναι δεδομένος άνω τριγωνικός πίνακας, $\underline{b} \in \mathbb{R}^{n}$ είναι γνωστό διάνυσμα και $\underline{x} \in \mathbb{R}^{n}$ είναι το ζητούμενο διάνυσμα. Αναλυτικά το σύστημα των εξισώσεων αυτών έχει την μορφή:

$$
\begin{aligned}
u_{1,1}x_1 + u_{1,2}x_2 + u_{1,3}x_3 + \ldots + u_{1,n}x_n &= b_1 \\
u_{2,2}x_2 + u_{2,3}x_3 + \ldots + u_{2,n}x_n &= b_2 \\
u_{3,3}x_3 + \ldots + u_{3,n}x_n &= b_3 \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\
u_{n-1,n-1}x_{n-1} + u_{n-1,n}x_n &= b_{n-1} \\
u_{n,n}x_n &= b_n
\end{aligned}
$$

Για να έχει το πρόβλημα αυτό μοναδική λύση θα πρέπει ο πίνακας $\underline{\underline{U}}$ να είναι αντιστρέψιμος, δηλαδή η ορίζουσά του να είναι μη-μηδενική, $\det\left(\underline{\underline{U}}\right) \neq 0$. Εφόσον ο πίνακας είναι τριγωνικός η ορίζουσά του είναι ίση με το γινόμενο των διαγωνίων στοιχείων του, δηλαδή $\det\left(\underline{\underline{U}}\right) = \prod_{i=1}^{n} u_{i,i}$ το οποίο συνεπάγεται ότι όλα τα διαγώνια στοιχεία του πρέπει να είναι μη-μηδενικά, δηλαδή $u_{i,i} \neq 0, i = 1,2,3,...,n$.

Η εύρεση της λύσης του συστήματος αυτού βρίσκεται πολύ εύκολα αν ξεκινήσουμε από την τελευταία εξίσωση και λύσουμε ως προς τον μοναδικό άγνωστο της εξίσωσης αυτής, δηλαδή τον x_n , ως εξής:

$$x_n = b_n / u_{n,n}$$

ο οποίος φυσικά είναι καλά ορισμένος διότι όλα τα διαγώνια στοιχεία του πίνακα είναι μη-μηδενικά. Στην συνέχεια παρατηρούμε ότι η 2η από το τέλος εξίσωση έχει μοναδικό άγνωστο τον x_{n-1} τον οποίο μπορούμε να υπολογίσουμε ως εξής:

$$x_{n-1} = \left(b_{n-1} - u_{n-1,n} x_n\right) / u_{n-1,n-1}$$

Όμοια η 3η από το τέλος εξίσωση έχει μοναδικό άγνωστο τον x_{n-2} και τον οποίο μπορούμε να υπολογίσουμε. Αναδρομικά βρίσκουμε ότι η k-οστή από το τέλος εξίσωση έχει ως μοναδικό άγνωστο τον x_{n-k+1} , όπου k=1,2,3,…,n.

Ο αλγόριθμος υπολογισμού προκύπτει αν στην παραπάνω διαδικασία παρατηρήσουμε ότι σε κάθε βήμα της διαδικασίας ο μοναδικός άγνωστος είναι αυτός που βρίσκεται στην πιο αριστερά θέση της τρέχουσας εξίσωσης, ο οποίος μάλιστα είναι και διαγώνιος. Έτσι αν γράψουμε τις εξισώσεις του συστήματος με μορφή αθροίσματος θα έχουμε:

$$\sum_{k=i}^{n} u_{i,k} x_k = b_i, \quad i = 1,2,3,...,n$$

Το παραπάνω άθροισμα μπορεί να εκφραστεί σε 2 μέρη ως εξής:

$$u_{i,i} x_i + \sum_{k=i+1}^{n} u_{i,k} x_k = b_i, \quad i = 1,2,3,...,n$$

και μπορεί να λυθεί ως προς τον άγνωστο x_i , αρκεί να ξεκινήσουμε από το τέλος προς την αρχή, δηλαδή:

$$
\boxed{\begin{aligned} x_n &= b_n / u_{n,n} \\ x_i &= \left(b_i - \sum_{k=i+1}^{n} u_{i,k} x_k \right) \Big/ u_{i,i}\,, \quad i = n-1, n-2, \ldots, 1 \end{aligned}}
$$

3.3 Ο αλγόριθμος της εμπρός αντικατάστασης

Ο αλγόριθμος της εμπρός αντικατάστασης είναι όμοιος με τον προηγούμενο αλγόριθμο. Χρησιμοποιείται για την επίλυση γραμμικών συστημάτων της μορφής $\underline{\underline{L}} \cdot \underline{x} = \underline{b}$ όπου $\underline{\underline{L}} \in \mathbb{R}^{n,n}$ είναι δεδομένος άνω τριγωνικός πίνακας, $\underline{b} \in \mathbb{R}^{n}$ είναι γνωστό διάνυσμα και $\underline{x} \in \mathbb{R}^{n}$ είναι το ζητούμενο διάνυσμα. Αναλυτικά το σύστημα των εξισώσεων αυτών έχει την μορφή:

$$
\begin{aligned}
&\ell_{1,1} x_1 && = b_1 \\
&\ell_{2,1} x_1 + \ell_{2,2} x_2 && = b_2 \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots && \\
&\ell_{n-1,1} x_1 + \ell_{n-1,2} x_2 + \ldots + \ell_{n-1,n-1} x_{n-1} && = b_{n-1} \\
&\ell_{n,1} x_1 + \ell_{n,2} x_2 + \ldots + \ell_{n,n-1} x_{n-1} + \ell_{n,n} x_n && = b_n
\end{aligned}
$$

Για να έχει το πρόβλημα αυτό μοναδική λύση θα πρέπει ο πίνακας $\underline{\underline{L}}$ να είναι αντιστρέψιμος, δηλαδή η ορίζουσά του να είναι μη-μηδενική, $\det\left(\underline{\underline{L}}\right) \neq 0$. Εφόσον ο πίνακας είναι τριγωνικός η ορίζουσά του είναι ίση με το γινόμενο των διαγωνίων στοιχείων του, δηλαδή $\det\left(\underline{\underline{L}}\right) = \prod_{i=1}^{n} \ell_{i,i}$ το οποίο συνεπάγεται ότι όλα τα διαγώνια στοιχεία του πρέπει να είναι μη-μηδενικά, δηλαδή $\ell_{i,i} \neq 0, i = 1,2,3,\ldots,n$.

Η εύρεση της λύσης του συστήματος αυτού βρίσκεται πολύ εύκολα αν ξεκινήσουμε από την πρώτη εξίσωση και λύσουμε ως προς τον μοναδικό άγνωστο της εξίσωσης αυτής, δηλαδή τον x_1, ως εξής:

$$x_1 = b_1 / u_{1,1}$$

ο οποίος φυσικά είναι καλά ορισμένος διότι όλα τα διαγώνια στοιχεία του πίνακα είναι μη-μηδενικά. Στην συνέχεια παρατηρούμε ότι η 2η εξίσωση έχει μοναδικό άγνωστο τον x_2 τον οποίο μπορούμε να υπολογίσουμε ως εξής:

$$x_2 = \left(b_2 - \ell_{2,1} x_1\right) / \ell_{2,2}$$

Όμοια η 3η εξίσωση έχει μοναδικό άγνωστο τον x_3 τον οποίο και μπορούμε να υπολογίσουμε. Αναδρομικά βρίσκουμε ότι η k-οστή εξίσωση έχει ως μοναδικό άγνωστο τον x_k , όπου k=1,2,3,…,n.

Ο αλγόριθμος υπολογισμού προκύπτει αν στην παραπάνω διαδικασία παρατηρήσουμε ότι σε κάθε βήμα της διαδικασίας ο μοναδικός άγνωστος είναι αυτός που βρίσκεται στην πιο δεξιά θέση της τρέχουσας εξίσωσης, ο οποίος μάλιστα είναι και διαγώνιος. Έτσι αν γράψουμε τις εξισώσεις του συστήματος με μορφή αθροίσματος θα έχουμε:

$$\sum_{k=1}^{i} \ell_{i,k} x_k = b_i, \quad i = 1,2,3,...,n$$

Το παραπάνω άθροισμα μπορεί να εκφραστεί σε 2 μέρη ως εξής:

$$\sum_{k=1}^{i-1} \ell_{i,k} x_k + \ell_{i,i} x_i = b_i, \quad i = 1,2,3,...,n$$

και μπορεί να λυθεί ως προς τον άγνωστο x_i , αρκεί να ξεκινήσουμε από την αρχή προς το τέλος, δηλαδή:

$$\boxed{\begin{aligned} x_1 &= b_1 / \ell_{1,1} \\ x_i &= \left(b_i - \sum_{k=i+1}^{n} \ell_{i,k} x_k \right) \Big/ \ell_{i,i}, \quad i = 2,3,...,n \end{aligned}}$$

3.4. Απαλοιφή Gauss

Η απαλοιφή Gauss αποτελεί την κύρια μέθοδο αριθμητικής επίλυσης γραμμικών συστημάτων με πυκνούς πίνακες, δηλαδή για πίνακες που δεν έχουν κάποια συγκεκριμένη δομή. Μπορεί να εφαρμοστεί όταν $\det\left(\underline{\underline{A}}\right) \neq 0$, δηλαδή υπό την προϋπόθεση ότι το γραμμικό σύστημα που θέλουμε να επιλύσουμε έχει μοναδική λύση. Αποτελείται από δύο φάσεις:

- Φάση τριγωνοποίησης. Στην φάση αυτή το αρχικό γραμμικό σύστημα $\underline{\underline{A}} \cdot \underline{x} = \underline{b}$ μετατρέπεται σε ένα ισοδύναμο σύστημα $\underline{\underline{U}} \cdot \underline{x} = \underline{\hat{b}}$ όπου $\underline{\underline{U}}$ είναι ένας άνω τριγωνικός πίνακας και το $\underline{\hat{b}}$ είναι γνωστό διάνυσμα (υπενθυμίζεται ότι δύο συστήματα λέγονται ισοδύναμα όταν έχουν την ίδια λύση).
- Φάση πίσω αντικατάστασης. Στην φάση αυτή, το γραμμικό σύστημα $\underline{\underline{U}} \cdot \underline{x} = \underline{\hat{b}}$ επιλύεται με τον αλγόριθμο της πίσω αντικατάστασης για να υπολογισθεί η λύση $\underline{x}$.

Θα περιγράψουμε εν συντομία την φάση της τριγωνοποίησης (ο αλγόριθμος της πίσω αντικατάστασης αναλύθηκε διεξοδικά στην παράγραφο 3.2). Όπως ήδη αναφέρθηκε, σκοπός αυτής της φάσης είναι να μετατραπεί ο αρχικός πίνακας του συστήματος σε έναν άνω τριγωνικό πίνακα. Η διαδικασία έχεις ως εξής:

(α) Ελέγχουμε αν $a_{1,1} \neq 0$, αλλιώς βρίσκουμε μία εξίσωση του συστήματος, έστω k, της οποίας ο πρώτος συντελεστής είναι διαφορετικός από το μηδέν (δηλαδή $a_{k,1} \neq 0$) και εναλλάσσουμε την πρώτη με την k εξίσωση του συστήματος. Ο νέος πίνακας που προκύπτει είναι ο πίνακας $\underline{\underline{A}}^{(1)}$ του συστήματος με στοιχεία $a_{i,j}^{(1)}, \quad i, j = 1, 2, ..., n$.

$$\begin{bmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & a_{1,3}^{(1)} & ... & a_{1,n}^{(1)} \\ a_{2,1}^{(1)} & a_{2,2}^{(1)} & a_{2,3}^{(1)} & ... & a_{2,n}^{(1)} \\ ..... & & & & \\ a_{2,1}^{(1)} & a_{n,2}^{(2)} & a_{n,3}^{(1)} & ... & a_{n,n}^{(1)} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ ... \\ x_n \end{bmatrix} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \\ ... \\ b_n^{(1)} \end{bmatrix}$$

(β) Ορίζουμε τις ποσότητες $m_{i,1} = \dfrac{a_{i,1}^{(1)}}{a_{1,1}^{(1)}}, \quad i = 2, 3, ..., n$.

(γ) Πολλαπλασιάζουμε την 1η εξίσωση με το $m_{i,1}$ και το αποτέλεσμα το αφαιρώ από την «i» εξίσωση. Το νέο αποτέλεσμα το θέτω στην θέση της «i» εξίσωσης. Την διαδικασία

αυτή την εφαρμόζουμε για $i = 2,3,...,n$. Έτσι το νέο σύστημα που θα προκύψει θα είναι το εξής:

$$\underbrace{\begin{bmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & a_{1,3}^{(1)} & ... & a_{1,n}^{(1)} \\ 0 & a_{2,2}^{(2)} & a_{2,3}^{(2)} & ... & a_{2,n}^{(2)} \\ ..... & & & & \\ 0 & a_{n,2}^{(2)} & a_{n,3}^{(2)} & ... & a_{n,n}^{(2)} \end{bmatrix}}_{\underline{\underline{A}}^{(2)}} \cdot \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ ... \\ x_n \end{bmatrix}}_{\underline{x}} = \underbrace{\begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ ... \\ b_n^{(2)} \end{bmatrix}}_{\underline{b}^{(2)}}$$

Παρατηρούμε δηλαδή ότι με αυτόν τον τρόπο όλα τα στοιχεία της 1ης στήλης κάτω από το στοιχείο $a_{1,1}^{(1)}$ έχουν γίνει μηδέν, οπότε λέμε ότι το πρώτο βήμα της διαδικασίας της τριγωνοποίησης ολοκληρώθηκε. Τα νέα στοιχεία που θα προκύψουν δίνονται από τον τύπο

$$a_{i,j}^{(2)} = a_{i,j}^{(1)} - m_{i,1} a_{1,j}^{(1)}, \quad i, j = 2,3,...,n$$
$$b_i^{(2)} = b_i^{(1)} - m_{i,1} b_1^{(1)}, \quad i = 2,3,...,n$$

Εδώ να τονισθεί ότι η πρώτη εξίσωση του συστήματος δεν άλλαξε.

(δ) Επαναλαμβάνουμε τα βήματα (β) και (γ) για τον υποπίνακα που προκύπτει αν διαγράψουμε την πρώτη στήλη και την πρώτη γραμμή του πίνακα $\underline{\underline{A}}^{(2)}$ που προέκυψε στο προηγούμενο βήμα. Έτσι έχουμε τις ποσότητες $m_{i,2} = \frac{a_{i,2}^{(2)}}{a_{2,2}^{(2)}}, \quad i = 3,4,...,n$. Αν $a_{2,2}^{(2)} = 0$ τότε εφαρμόζουμε την διαδικασία του (α) βήματος, δηλαδή ψάχνουμε όλες τις εξισώσεις από την 3η έως και την τελευταία για να βρούμε ποια έχει $a_{k,2}^{(2)} \neq 0, k = 3,4,...,n$ Στην συνέχεια παίρνουμε τον γραμμικό συνδυασμό των εξισώσεων όπως προηγουμένως οπότε θα προκύψει το εξής σύστημα:

$$\underbrace{\begin{bmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & a_{1,3}^{(1)} & ... & a_{1,n}^{(1)} \\ 0 & a_{2,2}^{(2)} & a_{2,3}^{(2)} & ... & a_{2,n}^{(2)} \\ 0 & 0 & a_{3,3}^{(3)} & ... & a_{3,n}^{(3)} \\ ..... & & & & \\ 0 & 0 & a_{n,3}^{(3)} & ... & a_{n,n}^{(3)} \end{bmatrix}}_{\underline{\underline{A}}^{(3)}} \cdot \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ ... \\ x_n \end{bmatrix}}_{\underline{x}} = \underbrace{\begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(3)} \\ ... \\ b_n^{(3)} \end{bmatrix}}_{\underline{b}^{(3)}}$$

στο οποίο η πρώτη και η δεύτερη γραμμή είναι η ίδια με του προηγούμενου πίνακα, τα στοιχεία της δεύτερης στήλης κάτω από το $a_{2,2}^{(2)}$ έχουν γίνει μηδέν ενώ τα υπόλοιπα θα δίνονται από τον τύπο:

$$a_{i,j}^{(3)} = a_{i,j}^{(2)} - m_{i,2}a_{2,j}^{(2)}, \quad i,j = 3,4,...,n$$
$$b_i^{(3)} = b_i^{(2)} - m_{i,2}b_2^{(2)}, \quad i = 3,4,...,n$$

(ε) Συνεχίζουμε την ίδια διαδικασία και στο τέλος του r-βήματος το σύστημα που προκύπτει είναι $\underline{\underline{A}}^{(r+1)} \cdot \underline{x} = \underline{b}^{(r+1)}$, $1 \le r \le n-1$ όπου τα στοιχεία του $\underline{\underline{A}}^{(r+1)}$ δίνονται από τον τύπο

$$\boxed{\begin{array}{l} \left.\begin{array}{l} a_{i,j}^{(r+1)} = 0, \quad j = 1,2,...,r, \quad i = r+1, r+2,..,n \\ a_{i,j}^{(r+1)} = a_{i,j}^{(r)} - m_{i,r}a_{r,j}^{(r)}, \quad i,j = r+1, r+2,...,n \\ b_i^{(r+1)} = b_i^{(r)} - m_{i,r}b_r^{(r)}, \quad i = r+1, r+2,...,n \end{array}\right\}, \quad r = 1,2,...,n-1 \\ m_{i,r} = a_{i,r}^{(r)} \big/ a_{r,r}^{(r)}, \quad i = r+1, r+2,...,n \end{array}} \quad (*)$$

Ο τύπος (*) αποτελεί ουσιαστικά την απαλοιφή Gauss. Να σημειωθεί ότι τα στοιχεία $a_{r,r}^{(r)}$ ονομάζονται *οδηγοί* και οι ποσότητες $m_{i,r}$ ονομάζονται *πολλαπλασιαστές*.

Στο τέλος του n-2 βήματος θα έχουμε

$$\underbrace{\begin{bmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & a_{1,3}^{(1)} & ... & a_{1,n-1}^{(1)} & a_{1,n}^{(1)} \\ 0 & a_{2,2}^{(2)} & a_{2,3}^{(2)} & ... & a_{2,n-1}^{(2)} & a_{2,n}^{(2)} \\ 0 & 0 & a_{3,3}^{(3)} & ... & a_{3,n-1}^{(3)} & a_{3,n}^{(3)} \\ ..... & & & & & \\ 0 & 0 & 0 & ... & a_{n-1,n-1}^{(n-1)} & a_{n-1,n}^{(n-1)} \\ 0 & 0 & 0 & ... & a_{n,n-1}^{(n-1)} & a_{n,n}^{(n-1)} \end{bmatrix}}_{\underline{\underline{A}}^{(n-1)}} \cdot \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ ... \\ x_{n-1} \\ x_n \end{bmatrix}}_{\underline{x}} = \underbrace{\begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(3)} \\ ... \\ b_{n-1}^{(n-1)} \\ b_n^{(n-1)} \end{bmatrix}}_{\underline{b}^{(n-1)}}$$

και τελικά, εφαρμόζοντας την διαδικασία για τις δύο μόνο τελευταίες εξισώσεις παίρνουμε:

$$\underbrace{\begin{bmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & a_{1,3}^{(1)} & ... & a_{1,n-1}^{(1)} & a_{1,n}^{(1)} \\ 0 & a_{2,2}^{(2)} & a_{2,3}^{(2)} & ... & a_{2,n-1}^{(2)} & a_{2,n}^{(2)} \\ 0 & 0 & a_{3,3}^{(3)} & ... & a_{3,n-1}^{(3)} & a_{3,n}^{(3)} \\ ..... & & & & & \\ 0 & 0 & 0 & ... & a_{n-1,n-1}^{(n-1)} & a_{n-1,n}^{(n-1)} \\ 0 & 0 & 0 & ... & 0 & a_{n,n}^{(n)} \end{bmatrix}}_{\underline{\underline{A}}^{(n)}} \cdot \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ ... \\ x_{n-1} \\ x_n \end{bmatrix}}_{\underline{x}} = \underbrace{\begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(3)} \\ ... \\ b_{n-1}^{(n-1)} \\ b_n^{(n)} \end{bmatrix}}_{\underline{b}^{(n)}}$$

Επομένως στο τέλος της τριγωνοποίησης έχει προκύψει ένα ισοδύναμο σύστημα, $\underline{\underline{A}}^{(n)} \cdot \underline{x} = \underline{b}^{(n)}$, με το αρχικό σύστημα του οποίου όμως ο πίνακας $\underline{\underline{A}}^{(n)}$ είναι άνω τριγωνικός. Έτσι μπορούμε να γράψουμε $\underline{\underline{U}} \cdot \underline{x} = \underline{\hat{b}}$, $\underline{\underline{U}} \equiv \underline{\underline{A}}^{(n)}$, $\underline{b} \equiv \underline{b}^{(n)}$ και να επιλύσουμε το τελευταίο αυτό σύστημα με τον αλγόριθμο της πίσω αντικατάστασης.

Παράδειγμα: (δείτε το παράδειγμα στις σελίδες 73 και 74 του βιβλίου των Ακρίβη και Δουγαλή, Εισαγωγή στην Αριθμητική Ανάλυση, 5η αναθεωρημένη έκδοση).

3.4.1. Απαλοιφή Gauss με οδήγηση

Δυστυχώς, μπορεί να αποδειχτεί ότι ο παραπάνω αλγόριθμος όπως εφαρμόσθηκε είναι ένας ασταθής αλγόριθμος. Η αιτία βρίσκεται στο γεγονός ότι αν τα στοιχεία $a_{r,r}^{(r)}, r = 1,2,...,n-1$, δηλαδή οι *οδηγοί* της απαλοιφής, που απαιτούνται για τον υπολογισμό των *πολλαπλασιαστών* $m_{i,r}$ είναι πολύ μικρά τότε οι πολλαπλασιαστές μπορεί να είναι πολύ μεγαλύτεροι σε σχέση με τα υπόλοιπα στοιχεία που θα πολλαπλασιάσουν, γεγονός που έχει ως αποτέλεσμα την απώλεια ακρίβειας. Όσο η διαδικασία της τριγωνοποίησης προχωρά τα σφάλματα αυτά μεγεθύνονται και τελικά ο αλγόριθμος καθίσταται ασταθής. Όμως, μία μικρή τροποποίησή του μετατρέπει τον αλγόριθμο, στην μεγάλη πλειοψηφία των περιπτώσεων τουλάχιστον, από ασταθή σε ευσταθή. Η τροποποίηση ονομάζεται *οδήγηση* και διακρίνεται σε (α) *μερική οδήγηση* και (β) σε *ολική οδήγηση*. Στην πράξη εφαρμόζεται μόνο η μερική οδήγηση, η οποία είναι ιδιαίτερα αποδοτική και αυξάνει ελάχιστα το υπολογιστικό κόστος του αλγορίθμου.

Όπως περιγράφηκε, πριν την πραγματοποίηση του κάθε βήματος στην τριγωνοποίηση ελέγχουμε κατά πόσο οι οδηγοί ($a_{r,r}^{(r)}, r = 1,2,...,n-1$) είναι μη-μηδενικοί έτσι ώστε οι πολλαπλασιαστές να είναι καλά ορισμένοι. Αν δεν είναι, ψάχνουμε να βρούμε μία εξίσωση, από αυτές που βρίσκονται κάτω από την εξίσωση που καθορίζει τον υπολογισμό των πολλαπλασιαστών, που να έχει μη-μηδενικό στοιχείο στην θέση του οδηγού. Όταν το βρούμε εναλλάσσουμε τις δύο αυτές εξισώσεις. Το στάδιο αυτό λοιπόν το εφαρμόζουμε πάντα με την διαφορά ότι κάθε φορά ψάχνουμε να βρούμε το μέγιστο κατά απόλυτη τιμή των στοιχείων που βρίσκονται στην στήλη κάτω από το οδηγό στοιχείο και στην συνέχεια εναλλάσσουμε τις δύο αυτές εξισώσεις. Εδώ να τονίσουμε ότι η εναλλαγή γραμμών δεν αλλάζει την σειρά των αγνώστων του συστήματος.

Ολική οδήγηση είναι η διαδικασία κατά την οποία ψάχνουμε να βρούμε όχι μόνο το μέγιστο κάτω από το τρέχων οδηγό στοιχείο αλλά το μέγιστο όλων των στοιχείων του υποπίνακα στον οποίο εργαζόμαστε. Στην περίπτωση αυτή όμως απαιτείται να αλλάξει και η σειρά των αγνώστων $x_r, x_{r+1}, ..., x_n$, $r = 1,2,...,n-1$. Στην περίπτωση της ολικής οδήγησης ο αλγόριθμος της απαλοιφής Gauss γίνεται αρκετά πολύπλοκος, το

υπολογιστικό κόστος αυξάνει σε μεγάλο βαθμό ενώ θα πρέπει να σημειωθεί ότι οι περιπτώσεις που η μερική οδήγηση δεν είναι αρκετή για να κάνει τον αλγόριθμο ευσταθή αλλά μπορεί να τον κάνει ευσταθή η ολική οδήγηση είναι ελάχιστες. Για τους λόγους αυτούς η ολική οδήγηση σπάνια χρησιμοποιείται.

Παράδειγμα: (δείτε το παράδειγμα στις σελίδες 79 και 80 του βιβλίου των Ακρίβη και Δουγαλή, Εισαγωγή στην Αριθμητική Ανάλυση, 5η αναθεωρημένη έκδοση).

3.5. Ανάλυση LU

Η ανάλυση LU (Lower-Upper Decomposition) αποτελεί μία μικρή τροποποίηση της απαλοιφής Gauss. Πιο συγκεκριμένα, εκφράζουμε τον πίνακα του γραμμικού μας συστήματος στην μορφή $\underline{\underline{A}} = \underline{\underline{L}} \cdot \underline{\underline{U}}$ όπου $\underline{\underline{U}}$ είναι ένας άνω τριγωνικός πίνακας, ίδιος με αυτόν του τελευταίου βήματος της τριγωνοποίησης κατά την απαλοιφή Gauss και $\underline{\underline{L}}$ είναι ένας κάτω τριγωνικός πίνακας ο οποίος περιέχει μονάδες στην διαγώνιό του και στις στήλες τους πολλαπλασιαστές που προκύπτουν κατά την διαδικασία της τριγωνοποίησης της απαλοιφής Gauss. Ας υποθέσουμε προς στιγμή ότι δεν έχει γίνει καμία εναλλαγή γραμμών κατά την τριγωνοποίηση (δηλαδή σαν να μην έχουμε εφαρμόσει την διαδικασία της μερικής ή ολικής οδήγησης). Σύμφωνα με αυτήν την ανάλυση το γραμμικό σύστημα $\underline{\underline{A}} \cdot \underline{x} = \underline{b}$ γράφεται στην μορφή $\left(\underline{\underline{L}} \cdot \underline{\underline{U}}\right) \cdot \underline{x} = \underline{b}$ είτε $\underline{\underline{L}} \cdot \underbrace{\left(\underline{\underline{U}} \cdot \underline{x}\right)}_{\underline{y}} = \underline{b}$. Στην εξίσωση αυτή παρατηρούμε ότι η ποσότητα μέσα στην παρένθεση είναι ένα διάνυσμα, έστω $\underline{y} = \underline{\underline{U}} \cdot \underline{x}$. Τότε έχουμε $\underline{\underline{L}} \cdot \underline{y} = \underline{b}$, δηλαδή έχουμε ένα γραμμικό σύστημα το οποίο μπορεί να λυθεί πολύ εύκολα με τον αλγόριθμο της προς τα εμπρός αντικατάστασης για να μας δώσει το άγνωστο διάνυσμα $\underline{y}$. Στην συνέχεια επιλύουμε το σύστημα $\underline{y} = \underline{\underline{U}} \cdot \underline{x}$ με τον αλγόριθμο της προς τα πίσω αντικατάστασης, εφόσον αυτό είναι γραμμικό σύστημα με γνωστά τα $\underline{\underline{U}}$ και $\underline{y}$ για να πάρουμε το ζητούμενο διάνυσμα $\underline{x}$.

Στην περίπτωση που εφαρμόζουμε μερική οδήγηση κατά την απαλοιφή Guass η διαδικασία είναι η ίδια που περιγράφηκε παραπάνω μόνο που η απαραίτητη πληροφορία σχετικά με τις εναλλαγές γραμμών αποθηκεύεται σε ένα διάνυσμα και λαμβάνεται υπόψη τόσο στην πίσω όσο και στην εμπρός αντικατάσταση.

Το πλεονέκτημα της ανάλυσης LU βρίσκεται στο γεγονός ότι μπορούμε να λύσουμε το γραμμικό σύστημα για πολλαπλά δεξιά μέλη, δηλαδή $\underline{\underline{A}} \cdot \underline{x} = \underline{b}_j,\ j = 1,2,...,m$ οπότε ουσιαστικά έχουμε να λύσουμε «m» στο πλήθος γραμμικά συστήματα στα οποία ο πίνακας $\underline{\underline{A}}$ είναι ο ίδιος. Με την απαλοιφή Gauss θα έπρεπε κάθε φορά να επαναλάβουμε την διαδικασία της τριγωνοποίησης, η οποία υπενθυμίζεται ότι δεν εξαρτάται από το δεξί μέλος του γραμμικού συστήματος παρά μόνο από τα στοιχεία του πίνακα $\underline{\underline{A}}$. Ένα παράδειγμα στο οποίο απαιτείται πολλαπλή επίλυση γραμμικών συστημάτων με κοινό πίνακα $\underline{\underline{A}}$ είναι όταν πρέπει να υπολογίσουμε τον αντίστροφό του, $\underline{\underline{A}}^{-1}$, το οποίο ισοδυναμεί με την επίλυση «*n*» γραμμικών συστημάτων της μορφής $\underline{\underline{A}} \cdot \underline{x}_j = \underline{e}_j,\quad j = 1,2,...,n$ όπου βέβαια ο $\underline{\underline{A}}$ είναι ένας τετραγωνικός πίνακας nxn με μη-μηδενική ορίζουσα, $\underline{x}_j$ είναι οι στήλες του αντίστροφου πίνακα και $\underline{e}_j$ είναι ένα διάνυσμα διάστασης n το οποίο έχει παντού μηδενικά εκτός από την θέση j στην οποία υπάρχει η μονάδα (τα $\underline{e}_j, j = 1,2,...,n$ αποτελούν μία βάση του χώρου $\mathbb{R}^n$).

3.6. Ανάλυση Cholesky για συμμετρικούς και θετικά ορισμένους πίνακες

Στην παράγραφο αυτή θα μελετήσουμε την επίλυση γραμμικών συστημάτων στα οποία ο πίνακας του συστήματος είναι συμμετρικός και θετικά ορισμένος. Ο ορισμός, ιδιότητες και χαρακτηριστικά των θετικά ορισμένων πινάκων δίνονται στο παράρτημα Π4.

(Η παράγραφος αυτή δεν είναι ολοκληρωμένη. Για τον λόγο αυτό, δείτε την παράγραφο 3.5., σελίδες 93-96, του βιβλίου των Ακρίβη και Δουγαλή, Εισαγωγή στην Αριθμητική Ανάλυση, 5η αναθεωρημένη έκδοση).

3.7. Τριαδιαγώνια συστήματα (σποραδικοί πίνακες)

Στην παράγραφο αυτή θα μελετήσουμε ένα παράδειγμα άμεσης μεθόδου για αραιούς ή σποραδικούς πίνακες, δηλαδή για πίνακες που τα περισσότερα στοιχεία τους είναι μηδέν και έχουν συγκεκριμένη δομή. Πιο συγκεκριμένα θα θεωρήσουμε ότι ο $n \times n$ πίνακας $\underline{\underline{A}}$ του γραμμικού προβλήματος έχει την μορφή:

$$\underline{\underline{A}}=\begin{bmatrix} a_1 & \beta_1 & & & & \\ \gamma_2 & a_2 & \beta_2 & & \underline{\underline{0}} & \\ & \gamma_3 & a_3 & \beta_3 & & \\ & & . & . & . & \\ & \underline{\underline{0}} & & \gamma_{n-1} & a_{n-1} & \beta_{n-1} \\ & & & & \gamma_n & a_n \end{bmatrix}$$

δηλαδή, πρόκειται για έναν πίνακα ο οποίος έχει τα στοιχεία $a_i, i=1,2,...,n$ στην κύρια διαγώνιό του, τα στοιχεία $\beta_i, i=1,2,...,n-1$ ακριβώς δεξιά από την διαγώνιο, τα στοιχεία $\gamma_i, i=2,3,...,n$ ακριβώς αριστερά από την διαγώνιο και όλα τα υπόλοιπα είναι μηδέν (δηλαδή $A_{i,j}=0$ για $j>i+1$ και $j<i-1$). Ο πίνακας αυτός ονομάζεται τριδιαγώνιος και εμφανίζεται πολύ συχνά στις εφαρμογές. Υπό κάποιες συνθήκες, μπορεί να επιλυθεί με έναν ειδικό αλγόριθμο οποίος απαιτεί πολύ λιγότερες πράξεις από τους αλγόριθμους που ήδη μελετήσαμε, για να ολοκληρωθεί. Έστω λοιπόν το γραμμικό σύστημα $\underline{\underline{A}}\cdot\underline{x}=\underline{f}$ όπου $\underline{\underline{A}}\in\mathbb{R}^{n,n}$, $\underline{x},\underline{f}\in\mathbb{R}^n$ και ο πίνακας $\underline{\underline{A}}$ είναι τριδιαγώνιος. Για τα στοιχεία του υποθέτουμε ότι ισχύει $a_i,\beta_i,\gamma_i\neq 0$, $|a_1|>|\beta_1|$, $|a_n|>|\gamma_n|$ και $|a_i|\geq|\beta_i|+|\gamma_i|$, $i=2,3,...,n-1$. Υπό αυτές τις συνθήκες υπάρχει ο αντίστροφος $\underline{\underline{A}}^{-1}$ και μάλιστα ισχύει $\underline{\underline{A}}=\underline{\underline{L}}\cdot\underline{\underline{U}}$ όπου

$$\underline{\underline{L}}=\begin{bmatrix} \delta_1 & 0 & & & & \\ \gamma_2 & \delta_2 & 0 & & \underline{\underline{0}} & \\ & \gamma_3 & \delta_3 & 0 & & \\ & & . & . & . & \\ & \underline{\underline{0}} & & \gamma_{n-1} & \delta_{n-1} & 0 \\ & & & & \gamma_n & \delta_n \end{bmatrix}$$ και

$$\underline{\underline{U}}=\begin{bmatrix} 1 & \varepsilon_1 & & & & \\ 0 & 1 & \varepsilon_2 & & \underline{\underline{0}} & \\ & 0 & 1 & \varepsilon_3 & & \\ & & . & . & . & \\ & \underline{\underline{0}} & & 0 & 1 & \varepsilon_{n-1} \\ & & & & 0 & 1 \end{bmatrix}.$$

Όπως ήδη γνωρίζουμε από την ανάλυση LU αν ισχύει $\underline{\underline{A}}=\underline{\underline{L}}\cdot\underline{\underline{U}}$ όπου ο $\underline{\underline{L}}$ είναι ένας κάτω τριγωνικός πίνακα και ο $\underline{\underline{U}}$ είναι ένας άνω τριγωνικός πίνακας τότε θα έχουμε $\underline{\underline{A}}\cdot\underline{x}=\underline{f}\Rightarrow\underline{\underline{L}}\cdot\underbrace{\underline{\underline{U}}\cdot\underline{x}}_{\underline{y}}=\underline{f}\Rightarrow\underline{\underline{L}}\cdot\underline{y}=\underline{f}$. Το τελευταίο αυτό γραμμικό σύστημα λύνεται πολύ εύκολα με τον αλγόριθμο της προς τα εμπρός αντικατάστασης για να προσδιοριστεί το

άγνωστο διάνυσμα $\underline{y}$ (εφόσον τα $\underline{\underline{L}}, \underline{f}$ είναι γνωστά). Στην συνέχεια επιλύουμε το σύστημα $\underline{\underline{U}} \cdot \underline{x} = \underline{y}$ με τον αλγόριθμο της προς τα πίσω αντικατάστασης και βρίσκουμε το ζητούμενο διάνυσμα $\underline{x}$ (εφόσον τα $\underline{\underline{U}}, \underline{y}$ είναι γνωστά). Οι πίνακες $\underline{\underline{L}}$ και $\underline{\underline{U}}$ γενικά προκύπτουν από την φάση της τριγωνοποίησης. Βλέπουμε όμως ότι στην περίπτωση του τριδιαγώνιου συστήματος η μορφή τους είναι εξαιρετικά απλή. Πιο συγκεκριμένα ο $\underline{\underline{L}}$ είναι ένας κάτω τριγωνικός πίνακας που έχει μη-μηδενικά στοιχεία μόνο στην κύρια διαγώνιό του και ακριβώς αριστερά από αυτήν. Τα στοιχεία μάλιστα αριστερά της διαγωνίου είναι τα ίδια με αυτά του πίνακα $\underline{\underline{A}}$. Ο πίνακας $\underline{\underline{U}}$ είναι ένας άνω τριγωνικός πίνακας με μονάδες στην κύρια διαγώνιό του και μη-μηδενικά στοιχεία ακριβώς δεξιά της κυρίας διαγωνίου. Επομένως τα μόνα στοιχεία των πινάκων αυτήν είναι τα $\varepsilon_i, i = 1,2,...,n-1$ και $\delta_i, i = 1,2,...,n$. Τα στοιχεία αυτά θα προσδιοριστούν από την ισότητα $\underline{\underline{A}} = \underline{\underline{L}} \cdot \underline{\underline{U}}$. Πράγματι αν εκτελέσουμε τον πολλαπλασιασμό των πινάκων $\underline{\underline{L}}, \underline{\underline{U}}$ και εξισώσουμε με τον $\underline{\underline{A}}$ θα πάρουμε:

- Από την 1η γραμμή – 1η στήλη θα έχουμε $\delta_1 = a_1$.
- Από την 1η γραμμή – 2η στήλη θα έχουμε $\delta_1 \varepsilon_1 = \beta_1 \Rightarrow \varepsilon_1 = \beta_1 / \delta_1$.
- Από την 1η γραμμή – k στήλη, με k>2 θα καταλήξουμε σε ταυτότητες της μορφής 0=0 όπως πολύ εύκολα μπορεί να διαπιστωθεί.
- Από την 2η γραμμή – 1η στήλη θα καταλήξουμε σε ταυτότητα.
- Από την 2η γραμμή – 2η στήλη θα έχουμε $\gamma_2 \varepsilon_1 + \delta_2 = a_2 \Rightarrow \delta_2 = a_2 - \gamma_2 \varepsilon_1$
- Από την 2η γραμμή – 3η στήλη θα έχουμε $\varepsilon_2 \delta_2 = \beta_2 \Rightarrow \varepsilon_2 = \beta_2 / \delta_2$
- Από την 2η γραμμή – k στήλη, με k>3 θα καταλήξουμε σε ταυτότητες.
- ….
- Από την k-γραμμή – k-στήλη θα έχουμε $\delta_k = a_k - \gamma_k \varepsilon_{k-1}$
- Από την k-γραμμή – k+1-στήλη θα έχουμε $\varepsilon_k = \beta_k / \delta_k$
- ….
- Από την n-γραμμή – n-στήλη θα έχουμε $\delta_n = a_n - \gamma_n \varepsilon_{n-1}$

Άρα συνολικά ο αλγόριθμος έχει ως εξής:

$$\left\{\begin{array}{l}\delta_1 = a_1, \quad \varepsilon_1 = \beta_1/\delta_1 \\ \text{για } k = 2,3,...,n-1: \\ \quad \delta_k = a_k - \gamma_k \varepsilon_{k-1} \\ \quad \varepsilon_k = \beta_k/\delta_k \\ \text{και} \\ \delta_n = a_n - \gamma_n \varepsilon_{n-1} \\ \end{array}\right\} \quad (1)$$

Ο παραπάνω αλγόριθμος είναι καλά ορισμένος εφόσον $\delta_k \neq 0, \quad k = 1,2,...,n$. Επιπλέον έχουμε ότι:

$$\left.\begin{array}{l}\det\left(\underline{\underline{A}}\right) = \det\left(\underline{\underline{L}} \cdot \underline{\underline{U}}\right) = \det\left(\underline{\underline{L}}\right)\det\left(\underline{\underline{U}}\right) \\ \det\left(\underline{\underline{L}}\right) = \prod_{k=1}^{n} \delta_k \\ \det\left(\underline{\underline{U}}\right) = 1 \end{array}\right\} \Rightarrow \det\left(\underline{\underline{A}}\right) = \prod_{k=1}^{n} \delta_k$$

Υπενθυμίζεται ότι η ορίζουσα ενός τριγωνικού πίνακα είναι ίση με το γινόμενο των στοιχείων της κύριας διαγωνίου του. Όπως επίσης γνωρίζουμε για να είναι αντιστρέψιμος ο $\underline{\underline{A}}$ θα πρέπει η ορίζουσά του να είναι μη-μηδενική, δηλαδή $\det\left(\underline{\underline{A}}\right) \neq 0$ το οποίο φυσικά σημαίνει ότι $\prod_{k=1}^{n} \delta_k \neq 0 \Rightarrow \delta_k \neq 0, k = 1,2,...,n$. Άρα αναγκαία συνθήκη για να είναι ο $\underline{\underline{A}}$ αντιστρέψιμος και ο αλγόριθμος που αναπτύχθηκε καλά ορισμένος είναι όλα τα στοιχεία δ_k να είναι μη-μηδενικά. Παρατηρούμε αρχικά ότι $\delta_1 = a_1 \neq 0$ λόγω των υποθέσεών μας. Επίσης, έχουμε $\varepsilon_1 = \beta_1/\delta_1 = \beta_1/a_1 \Rightarrow |\varepsilon_1| = |\beta_1/a_1| = |\beta_1|/|a_1| < 1$ διότι $|a_1| > |\beta_1|$. Θα αποδείξουμε λοιπόν, με επαγωγικό τρόπο, ότι $\delta_k \neq 0, |\varepsilon_k| < 1, \quad 1 \leq k \leq n$. Η πρόταση ισχύει για $k = 1$. Υποθέτουμε ότι ισχύει για $k-1$, δηλαδή ότι $\delta_{k-1} \neq 0, |\varepsilon_{k-1}| < 1$. Θα δείξουμε ότι ισχύει και για k. Έχουμε $\delta_k = a_k - \gamma_k \varepsilon_{k-1} \Rightarrow |\delta_k| = |a_k - \gamma_k \varepsilon_{k-1}| \overset{(*)}{\geq} |a_k| - |\gamma_k \varepsilon_{k-1}| = |a_k| - |\gamma_k||\varepsilon_{k-1}|$. Άρα:

$$\left.\begin{array}{l}|\delta_k| \geq |a_k| - |\gamma_k||\varepsilon_{k-1}| > |a_k| - |\gamma_k| \quad (\text{διοτι } |\varepsilon_{k-1}| < 1) \\ |a_k| \geq |\beta_k| + |\gamma_k| \ (\text{λογω υποθεσης}) \Rightarrow |a_k| - |\gamma_k| \geq |\beta_k| \end{array}\right\} \Rightarrow |\delta_k| > |\beta_k| > 0 \Rightarrow |\varepsilon_k| \equiv \frac{|\beta_k|}{|\delta_k|} < 1$$

Επομένως δείξαμε ότι $|\delta_k| > 0$ και $|\varepsilon_k| < 1$ γεγονός που ολοκληρώνει την απόδειξη. Άρα ο πίνακας $\underline{\underline{A}}$ είναι αντιστρέψιμος και ο αλγόριθμος (1) είναι καλά ορισμένος. Να σημειωθεί ότι αν είχαμε ακολουθήσει την ανάλυση LU θα χρειαζόμαστε περίπου $\frac{1}{3}n^3$

πράξεις για την φάση της τριγωνοποίησης ενώ ο αλγόριθμος (1) απαιτεί μόνο $2n$ πράξεις!

$$|x| = |(x-y)+y| \le |x-y|+|y| \Rightarrow |x-y| \ge |x|-|y| \qquad (*)$$

3.8. Επαναληπτικές Μέθοδοι Επίλυσης Γραμμικών Συστημάτων

Στόχος εδώ είναι η μελέτη των επαναληπτικών μεθόδων για την αριθμητική επίλυση συστημάτων γραμμικών εξισώσεων, δηλαδή του προβλήματος $\underline{\underline{A}} \cdot \underline{x} = \underline{b}$, όπου $\underline{\underline{A}} \in \mathbb{R}^{n,n}$, $\underline{x}, \underline{b} \in \mathbb{R}^n$. Η φιλοσοφία των επαναληπτικών μεθόδων είναι εντελώς διαφορετική από αυτήν των άμεσων μεθόδων. Αρχικά πρέπει να σημειωθεί ότι οι μέθοδοι αυτοί δεν αλλάζουν τον πίνακα, $\underline{\underline{A}}$, του γραμμικού συστήματος. Το κύριο χαρακτηριστικό τους είναι ότι παράγουν μία ακολουθία προσεγγίσεων της λύσης, $\left\{\underline{x}^{(i)}\right\}_{i=1}^{\infty}$, η οποία ακολουθία, υπό κάποιες προϋποθέσεις που θα δούμε παρακάτω, συγκλίνει στην λύση του συστήματος, $\underline{x}^*$. Πριν περάσουμε στην περιγραφή και μελέτη αυτών των μεθόδων είναι χρήσιμος ο παρακάτω ορισμός.

Ορισμός σύγκλισης ακολουθίας διανυσμάτων: Λέμε ότι η ακολουθία των διανυσμάτων $\left\{\underline{x}^{(i)}\right\}_{i=1}^{\infty}$ όπου $\underline{x}^{(i)} \in \mathbb{R}^n$ συγκλίνει στο διάνυσμα $\underline{x}^* \in \mathbb{R}^n$ και γράφουμε $\lim_{i\to\infty} \underline{x}^{(i)} = \underline{x}^*$ όταν κάθε στοιχείο του διανύσματος, $x_j^{(i)}$, $j = 1,2,3,...,n$, συγκλίνει στο αντίστοιχο στοιχείο του διανύσματος $\underline{x}^*$, δηλαδή στο x_j^*, ή όταν $\lim_{i\to\infty} x_j^{(i)} = x_j^*$.

Για όλες τις διανυσματικές νόρμες ορισμένες στον $\mathbb{R}^n$, το γεγονός ότι μία ακολουθία διανυσμάτων συγκλίνει σημαίνει ότι $\lim_{i\to\infty} \left\| \underline{x}^{(i)} - \underline{x}^* \right\| = 0$, δηλαδή έχουμε:

$$\lim_{i\to\infty} \underline{x}^{(i)} = \underline{x}^* \Leftrightarrow \lim_{i\to\infty} \left\| \underline{x}^{(i)} - \underline{x}^* \right\| = 0$$

Εδώ συνιστάται η προσοχή του αναγνώστη. Αν $\lim_{i\to\infty} \left\| \underline{x}^{(i)} \right\| = \left\| \underline{x}^* \right\|$ τότε δεν συνεπάγεται αναγκαστικά η σύγκλιση της αντίστοιχης ακολουθίας διανυσμάτων. Ας πάρουμε για παράδειγμα την ακολουθία διανυσμάτων $\begin{bmatrix} +1 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ +1 \end{bmatrix}, \begin{bmatrix} +1 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ +1 \end{bmatrix}, ...$ η οποία παράγεται

από την αναδρομική σχέση $\underline{x}^{(i+1)} = -\underline{x}^{(i)}, \quad i = 1,2,...,n, \quad \underline{x}^{(0)} = \begin{bmatrix} +1 \\ -1 \end{bmatrix}$ η οποία προφανώς δεν συγκλίνει σε κάποιο διάνυσμα. Όμως η μέγιστη νόρμα των διανυσμάτων αυτών είναι ίση με 1, όπως ακριβώς και η αντίστοιχη νόρμα του διανύσματος $\underline{x}^* = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

3.8.1. Η μέθοδος Jacobi

Το γραμμικό σύστημα $\underline{\underline{A}} \cdot \underline{x} = \underline{b}$ μπορεί να γραφεί και στην μορφή αθροίσματος, δηλαδή $\sum_{k=1}^{n} a_{i,k} x_k = b_i, \quad i = 1,2,3,...,n$. Αναλύουμε το άθροισμα σε 3 όρους ως εξής:

$$\sum_{k=1}^{i-1} a_{i,k} x_k + a_{i,i} x_i + \sum_{k=i+1}^{n} a_{i,k} x_k = b_i, \quad i = 1,2,3,...,n$$

και επιλύουμε την σχέση αυτή ως προς x_i, δηλαδή:

$$x_i = \frac{1}{a_{i,i}} \left(b_i - \left(\sum_{k=1}^{i-1} a_{i,k} x_k + \sum_{k=i+1}^{n} a_{i,k} x_k \right) \right), \quad i = 1,2,3,...,n$$

Προφανώς για να είναι καλά ορισμένη η παραπάνω σχέση θα πρέπει τα διαγώνια στοιχεία, $a_{i,i}$, του πίνακα $\underline{\underline{A}}$ να είναι μη-μηδενικά. Η τελευταία αυτή σχέση μας οδηγεί στην λεγόμενη *επαναληπτική μέθοδος Jacobi*:

$$\boxed{x_i^{(m+1)} = \frac{1}{a_{i,i}} \left(b_i - \sum_{k=1}^{i-1} a_{i,k} x_k^{(m)} - \sum_{k=i+1}^{n} a_{i,k} x_k^{(m)} \right), \quad i = 1,2,3,...,n \text{ και } m = 0,1,2,...}$$

όπου ο εκθέτης m δηλώνει τον αριθμό επανάληψης της μεθόδου. Για την εκκίνηση της μεθόδου απαιτείται μία αρχική προσέγγιση της λύσης, δηλαδή το διάνυσμα $\underline{x}^{(0)}$. Σε περίπτωση που καμία αρχική προσέγγιση δεν είναι διαθέσιμη μπορούμε να θέσουμε $\underline{x}^{(0)} = \underline{0}$. Κριτήριο τερματισμού είναι συνήθως η συνθήκη $\left\| \underline{x}^{(m+1)} - \underline{x}^{(m)} \right\|_\infty < \varepsilon$, όπου ε η ζητούμενη ακρίβεια, η οποία εκφράζει την μέγιστη διαφορά των στοιχείων μεταξύ δύο διαδοχικών προσεγγίσεων του διανύσματος της λύσης.

3.8.2. Η μέθοδος Gauss-Seidel

Η επαναληπτική μέθοδος Gauss-Seidel είναι μία μικρή παραλλαγή της μεθόδου Jacobi αν και ο ρυθμός σύγκλισης της μεθόδου είναι αρκετά πιο γρήγορος, όπως θα φανεί παρακάτω. Πράγματι κατά την εφαρμογή της μεθόδου Jacobi παρατηρούμε ότι

ο διαδοχικός υπολογισμός των x_i επιτρέπει την χρήση των ήδη υπολογισμένων $x_i^{(m+1)}$, $i=1,2,...,i-1$ στο δεξί μέλος της μεθόδου Jacobi. Έτσι δημιουργείται η παρακάτω επαναληπτική μέθοδος:

$$\boxed{x_i^{(m+1)}=\frac{1}{a_{i,i}}\left(b_i-\sum_{k=1}^{i-1}a_{i,k}x_k^{(m+1)}-\sum_{k=i+1}^{n}a_{i,k}x_k^{(m)}\right),\quad i=1,2,3,...,n\ \text{ και }\ m=0,1,2,...}$$

Όπως και προηγουμένως για την εκκίνηση της μεθόδου απαιτείται μία αρχική προσέγγιση της λύσης, $\underline{x}^{(0)}$. Αν αυτή δεν είναι διαθέσιμη μπορούμε να θέσουμε $\underline{x}^{(0)}=\underline{0}$. Κριτήριο τερματισμού, είναι συνήθως η συνθήκη $\left\|\underline{x}^{(m+1)}-\underline{x}^{(m)}\right\|_\infty<\varepsilon$, όπου ε η ζητούμενη ακρίβεια, όπως ακριβώς και με την μέθοδο Jacobi.

3.8.3. Η μέθοδος διαδοχικής υπερχαλάρωσης (SOR)

Η επαναληπτική μέθοδος επίλυσης γραμμικών συστημάτων διαδοχικής υπερχαλάρωσης SOR (από τα αρχικά των λέξεων successive over relaxation) αποτελεί ουσιαστικά μία βελτίωση της επαναληπτικής μεθόδου Gauss-Seidel. Ο γενικός τύπος της είναι:

$$\boxed{x_i^{(m+1)}=\frac{\omega}{a_{i,i}}\left(b_i-\sum_{k=1}^{i-1}a_{i,k}x_k^{(m+1)}-\sum_{k=i}^{n}a_{i,k}x_k^{(m)}\right)+(1-\omega)x_i^{(m)},\quad i=1,2,...,3}$$

Η παράμετρος ω επιλέγεται με σκοπό την τάχιστη σύγκλιση της μεθόδου. Μπορεί να αποδειχτεί ότι για να μην αποκλίνει η μέθοδος θα πρέπει $0<\omega<2$. Όταν $0<\omega<1$ η μέθοδος ονομάζεται «μέθοδος διαδοχικής υποχαλάρωσης» και όταν $1<\omega<2$ «μέθοδος διαδοχικής υπερχαλάρωσης». Για $\omega=1$ η μέθοδος ανάγεται στην μέθοδο Gauss-Seidel, όπως πολύ εύκολα με απευθείας σύγκριση των αντίστοιχων τύπων.

3.9. Κριτήρια σύγκλισης των επαναληπτικών μεθόδων

Ορισμός: ένας τετραγωνικός πίνακας $\underline{\underline{A}}$ με στοιχεία $a_{i,j}$, $1\le i,j\le n$ λέγεται ότι έχει *αυστηρά κυριαρχική διαγώνιο* όταν $\left|a_{i,i}\right|>\sum_{\substack{j=1\\ j\neq i}}^{n}\left|a_{i,j}\right|$, $1\le i\le n$.

Πρόταση: Κάθε τετραγωνικός πίνακας $\underline{\underline{A}}$ που έχει αυστηρά κυριαρχική διαγώνιο είναι αντιστρέψιμος (το αντίστροφο δεν ισχύει).

Απόδειξη: Έστω ότι ο $\underline{\underline{A}}$ δεν είναι αντιστρέψιμος. Επομένως το γραμμικό, ομογενές σύστημα $\underline{\underline{A}}\cdot\underline{x}=\underline{0}$ έχει μη-μηδενική λύση. Έστω το μέγιστο των απολύτων τιμών των στοιχείων του διανύσματος της λύσης το οποίο φυσικά είναι διάφορο του μηδενός, $m=\max_{1\le j\le n}\left|x_j\right|\neq 0$. Επειδή το σύστημα είναι ομογενές, το διάνυσμα $\underline{y}=\frac{1}{m}\underline{x}$ αποτελεί επίσης λύση του συστήματος και μάλιστα το μέγιστο κατά απόλυτη τιμή των στοιχείων του είναι 1. Ας υποθέσουμε ότι το μέγιστο αυτό βρίσκεται στην θέση k, όπου $k\in\{1,2,...,n\}$. Αν εκφράσουμε αναλυτικά τις εξισώσεις του συστήματος αυτού θα έχουμε $\sum_{j=1}^{n}a_{i,j}y_j=0,\quad 1\le i\le n$, οπότε για $i=k$ έχουμε $\sum_{j=1}^{n}a_{k,j}y_j=0$ και επομένως, αναλύοντας το άθροισμα θα έχουμε $a_{k,k}\underbrace{y_k}_{=1}+\sum_{\substack{j=1\\ j\neq k}}^{n}a_{k,j}y_j=0\Rightarrow$

$$a_{k,k}=-\sum_{\substack{j=1\\ j\neq k}}^{n}a_{k,j}y_j\Rightarrow\left|a_{k,k}\right|=\left|-\sum_{\substack{j=1\\ j\neq i}}^{n}a_{k,j}y_j\right|\Rightarrow\quad\left|a_{k,k}\right|\le\sum_{\substack{j=1\\ j\neq k}}^{n}\left|a_{k,j}y_j\right|=\sum_{\substack{j=1\\ j\neq k}}^{n}\left|a_{k,j}\right|\left|y_j\right|\le\sum_{\substack{j=1\\ j\neq k}}^{n}\left|a_{k,j}\right|$$

όπου η τελευταία ανισότητα ισχύει εφόσον $\left|y_j\right|<1,\ j=1,2,..,k-1,k+1,...,n$. Δηλαδή καταλήξαμε ότι για την k-συνιστώσα του διανύσματος $\underline{\underline{A}}\cdot\underline{y}$ ισχύει $\left|a_{k,k}\right|\le\sum_{\substack{j=1\\ j\neq k}}^{n}\left|a_{k,j}\right|$ το οποίο φυσικά είναι άτοπο λόγω του γεγονότος ότι ο πίνακας έχει αυστηρά κυριαρχική διαγώνιο. Άρα ο πίνακας είναι αντιστρέψιμος.

Η γενική επαναληπτική μέθοδος για την επίλυση γραμμικών συστημάτων

Έστω το γραμμικό σύστημα $\underline{\underline{A}}\cdot\underline{x}=\underline{b}$. Εκφράζουμε τον πίνακα $\underline{\underline{A}}$ ως άθροισμα τριών πινάκων $\underline{\underline{A}}=\underline{\underline{L}}+\underline{\underline{D}}+\underline{\underline{U}}$ όπου:

$\underline{\underline{L}}$ είναι ένας κάτω τριγωνικός πίνακας που περιέχει τα στοιχεία του πίνακα $\underline{\underline{A}}$ κάτω από την κύρια διαγώνιό του, δηλαδή $\underline{\underline{L}} = \begin{bmatrix} 0 & & & \dots & & \\ a_{2,1} & 0 & & \dots & \underline{\underline{0}} & \\ a_{3,1} & a_{3,2} & 0 & \dots & & \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n-1,1} & a_{n-1,2} & a_{n-1,3} & \dots & 0 & \\ a_{n,1} & a_{n,2} & a_{n,3} & \dots & a_{n,n-1} & 0 \end{bmatrix}$.

$\underline{\underline{D}}$ είναι ένας διαγώνιος πίνακας που περιέχει τα στοιχεία της κύριας διαγωνίου του $\underline{\underline{A}}$, δηλαδή $\underline{\underline{D}} = \begin{bmatrix} a_{1,1} & & & \dots & \\ & a_{2,2} & & \dots & \underline{\underline{0}} \\ & & a_{3,3} & \dots & \\ & & & \dots & \\ & \underline{\underline{0}} & & & a_{n-1,n-1} \\ & & & & & a_{n,n} \end{bmatrix}$.

$\underline{\underline{U}}$ είναι ένας άνω τριγωνικός πίνακας που περιέχει τα στοιχεία του πίνακα $\underline{\underline{A}}$ πάνω από την κύρια διαγώνιό του, δηλαδή $\underline{\underline{U}} = \begin{bmatrix} 0 & a_{1,2} & a_{1,3} & \dots & a_{1,n-1} & a_{1,n} \\ & 0 & a_{2,3} & \dots & a_{2,n-1} & a_{2,n} \\ & & 0 & \dots & a_{3,n-1} & a_{3,n} \\ & \underline{\underline{0}} & & \dots & \dots & \dots \\ & & & & 0 & a_{n-1,n} \\ & & & & & 0 \end{bmatrix}$.

Επομένως το γραμμικό σύστημα μπορεί να γραφεί στην μορφή $\left(\underline{\underline{L}} + \underline{\underline{D}} + \underline{\underline{U}}\right) \cdot \underline{x} = \underline{b} \Rightarrow$

$$\left.\begin{aligned} \underline{\underline{D}} \cdot \underline{x} + \left(\underline{\underline{L}} + \underline{\underline{U}}\right) \cdot \underline{x} = \underline{b} \\ \left(\underline{\underline{L}} + \underline{\underline{D}}\right) \cdot \underline{x} + \underline{\underline{U}} \cdot \underline{x} = \underline{b} \end{aligned}\right\} \Rightarrow \left.\begin{aligned} \underline{\underline{D}} \cdot \underline{x} = -\left(\underline{\underline{L}} + \underline{\underline{U}}\right) \cdot \underline{x} + \underline{b} \\ \left(\underline{\underline{L}} + \underline{\underline{D}}\right) \cdot \underline{x} = -\underline{\underline{U}} \cdot \underline{x} + \underline{b} \end{aligned}\right\} \Rightarrow \left.\begin{aligned} \underline{x} = \underbrace{-\underline{\underline{D}}^{-1} \cdot \left(\underline{\underline{L}} + \underline{\underline{U}}\right)}_{\underline{\underline{G}}_J} \cdot \underline{x} + \underbrace{\underline{\underline{D}}^{-1} \cdot \underline{b}}_{\underline{f}_J} \\ \underline{x} = \underbrace{-\left(\underline{\underline{L}} + \underline{\underline{D}}\right)^{-1} \cdot \underline{\underline{U}}}_{\underline{\underline{G}}_{GS}} \cdot \underline{x} + \underbrace{\left(\underline{\underline{L}} + \underline{\underline{D}}\right)^{-1} \cdot \underline{b}}_{\underline{f}_{GS}} \end{aligned}\right\} \Rightarrow$$

$$\underline{x} = \underline{\underline{G}}_J \cdot \underline{x} + \underline{f}_J \qquad (1a)$$

$$\underline{x} = \underline{\underline{G}}_{GS} \cdot \underline{x} + \underline{f}_{GS} \qquad (1b)$$

Από τις εξισώσεις αυτές προκύπτουν οι επαναληπτικές σχέσεις

$$\underline{x}^{(m+1)} = \underline{\underline{G}}_J \cdot \underline{x}^{(m)} + \underline{f}_J \qquad (2a)$$

$$\underline{x}^{(m+1)} = \underline{\underline{G}}_{GS} \cdot \underline{x}^{(m)} + \underline{f}_{GS} \qquad (2b)$$

, m=1,2,3...,

Με παρόμοιο τρόπο βρίσκουμε ότι για την μέθοδο SOR:

$\underline{x}^{(m+1)} = \underline{\underline{G}}_{SOR} \cdot \underline{x}^{(m)} + \underline{f}_{SOR}$ (2γ) , m=1,2,3...,

όπου $\underline{\underline{G}}_{GS} = \left(\underline{\underline{D}} + \omega\underline{\underline{L}}\right)^{-1} \cdot \left[\left(1-\omega\right)\underline{\underline{D}} - \omega\underline{\underline{U}}\right]$, $\underline{f}_{SOR} = \left(\underline{\underline{D}} + \omega\underline{\underline{L}}\right)^{-1} \cdot \underline{b}$

Η σχέση (2α) είναι η μέθοδος Jacobi εκφρασμένη σε μορφή πινάκων/διανυσμάτων, η σχέση (2β) είναι η μέθοδος Gauss-Seidel και η σχέση (2γ) είναι η μέθοδος SOR. Οι πίνακες $\underline{\underline{G}}_J$, $\underline{\underline{G}}_{GS}$ και $\underline{\underline{G}}_{SOR}$ ονομάζονται πίνακες επανάληψης των μεθόδων αυτών. Οι μέθοδοι αυτοί είναι καλά ορισμένοι όταν οι πίνακες $\underline{\underline{D}}$, $\underline{\underline{L}} + \underline{\underline{D}}$, $\omega\underline{\underline{L}} + \underline{\underline{D}}$ είναι αντιστρέψιμοι. Επειδή και στις τρεις αυτές περιπτώσεις η ορίζουσα είναι ίση με το γινόμενο των στοιχείων του $\underline{\underline{D}}$, δηλαδή ίση με $\prod_{i=1}^{n} a_{i,i}$, θα πρέπει $a_{i,i} \neq 0 \; \forall i \in \{1,2,...,n\}$, δηλαδή οι μέθοδοι Jacobi, Gauss-Seidel και SOR μπορούν να εφαρμοστούν μόνο εφόσον τα στοιχεία της κύριας διαγώνιου του πίνακα $\underline{\underline{A}}$ είναι μη-μηδενικά.

Και οι τρεις μέθοδοι αποτελούν ειδικές περιπτώσεις της γενικής επαναληπτικής μεθόδου η οποία προκύπτει αν εκφράσουμε τον πίνακα $\underline{\underline{A}}$ ως $\underline{\underline{A}} = \underline{\underline{M}} - \underline{\underline{N}}$ οπότε $\left(\underline{\underline{M}} - \underline{\underline{N}}\right) \cdot \underline{x} = \underline{b} \Rightarrow \underline{\underline{M}} \cdot \underline{x} = \underline{\underline{N}} \cdot \underline{x} + \underline{b}$. Αν ο πίνακας $\underline{\underline{M}}$ είναι αντιστρέψιμος τότε

$\underline{x} = \underbrace{\underline{\underline{M}}^{-1} \cdot \underline{\underline{N}}}_{\underline{\underline{G}}} \cdot \underline{x} + \underbrace{\underline{\underline{M}}^{-1} \cdot \underline{b}}_{\hat{\underline{b}}} \Rightarrow \quad \underline{x} = \underline{\underline{G}} \cdot \underline{x} + \hat{\underline{b}}$ από την οποία κατασκευάζουμε την γενική επαναληπτική μέθοδο $\boxed{\underline{x}^{(m+1)} = \underline{\underline{G}} \cdot \underline{x}^{(m)} + \hat{\underline{b}}, \quad m = 0,1,2,...}$ όπου ο πίνακας $\underline{\underline{G}}$ ονομάζεται πίνακας επανάληψης της μεθόδου. Είναι προφανές ότι αν η ακολουθία $\left\{\underline{x}\right\}_{m=0}^{\infty}$ συγκλίνει σε κάποιο διάνυσμα $\underline{x}^*$ τότε αυτό είναι η λύση του γραμμικού συστήματος $\underline{\underline{A}} \cdot \underline{x} = \underline{b}$. Ας διερευνήσουμε τώρα την σύγκλιση της γενικής επαναληπτικής μεθόδου. Έχουμε:

$$\left.\begin{aligned} \underline{x}^{(m+1)} &= \underline{\underline{G}} \cdot \underline{x}^{(m)} + \hat{\underline{b}} \\ \underline{x} &= \underline{\underline{G}} \cdot \underline{x} + \hat{\underline{b}} \end{aligned}\right\} \Rightarrow \underline{x}^{(m+1)} - \underline{x} = \underline{\underline{G}} \cdot \left(\underline{x}^{(m)} - \underline{x}\right).$$

Η τελευταία σχέση για m=0 δίνει $\underline{x}^{(1)} - \underline{x} = \underline{\underline{G}} \cdot \left(\underline{x}^{(0)} - \underline{x}\right)$. Για m=1 $\underline{x}^{(2)} - \underline{x} = \underline{\underline{G}} \cdot \left(\underline{x}^{(1)} - \underline{x}\right)$ και αντικαθιστώντας το $\underline{x}^{(1)} - \underline{x}$ από την αμέσως προηγούμενη σχέση παίρνουμε $\underline{x}^{(2)} - \underline{x} = \underline{\underline{G}} \cdot \left(\underline{\underline{G}} \cdot \left(\underline{x}^{(0)} - \underline{x}\right)\right) \Rightarrow \underline{x}^{(2)} - \underline{x} = \underline{\underline{G}}^2 \cdot \left(\underline{x}^{(0)} - \underline{x}\right)$. Επαγωγικά μπορούμε να δείξουμε ότι

$$\underline{x}^{(m)} - \underline{x} = \underline{\underline{G}}^m \cdot \left(\underline{x}^{(0)} - \underline{x}\right)$$

Στην συνέχεια θεωρούμε μία οποιαδήποτε διανυσματική νόρμα στον $\mathbb{R}^n$ και την αντίστοιχη, παραγόμενη φυσική νόρμα πινάκων. Παίρνοντας νόρμες στην τελευταία σχέση προκύπτει

$$\left\|\underline{x}^{(m)}-\underline{x}\right\|=\left\|\underline{\underline{G}}^m\cdot\left(\underline{x}^{(0)}-\underline{x}\right)\right\|\le\left\|\underline{\underline{G}}^m\right\|\left\|\left(\underline{x}^{(0)}-\underline{x}\right)\right\|$$

Παίρνοντας το όριο καθώς $m\to\infty$ έχουμε $\lim_{m\to\infty}\left\|\underline{x}^{(m)}-\underline{x}\right\|\le\lim_{m\to\infty}\left(\left\|\underline{\underline{G}}^m\right\|\left\|\left(\underline{x}^{(0)}-\underline{x}\right)\right\|\right)$ και επειδή η ποσότητα $\left\|\left(\underline{x}^{(0)}-\underline{x}\right)\right\|$ είναι ανεξάρτητη από το m έχουμε ότι $\lim_{m\to\infty}\left\|\underline{x}^{(m)}-\underline{x}\right\|\le\left\|\left(\underline{x}^{(0)}-\underline{x}\right)\right\|\lim_{m\to\infty}\left\|\underline{\underline{G}}^m\right\|$. Προφανώς έχουμε ότι αν $\lim_{m\to\infty}\left\|\underline{\underline{G}}^m\right\|=0\Leftrightarrow\lim_{m\to\infty}\underline{\underline{G}}^m=\underline{\underline{0}}$ (από την πρώτη ιδιότητα των νορμών πινάκων) τότε $\lim_{m\to\infty}\left\|\underline{x}^{(m)}-\underline{x}\right\|\le 0$ και επειδή η νόρμα είναι μη-αρνητικός αριθμός θα έχουμε $\lim_{m\to\infty}\left\|\underline{x}^{(m)}-\underline{x}\right\|=0$ και επομένως η μέθοδος συγκλίνει στην λύση του γραμμικού συστήματος $\underline{x}$ για κάθε αρχικό διάνυσμα $\underline{x}^{(0)}$. Επίσης μπορεί να αποδειχτεί και το αντίστροφο οπότε συνολικά έχουμε το παρακάτω θεώρημα:

> Η ακολουθία διανυσμάτων $\left\{\underline{x}^{(m)}\right\}_{m=0}^{\infty}$ που παράγεται από την αναδρομική σχέση $\underline{x}^{(m+1)}=\underline{\underline{G}}\cdot\underline{x}^{(m)}+\hat{\underline{b}}$ συγκλίνει στην λύση του $\underline{\underline{A}}\cdot\underline{x}=\underline{b}$ για κάθε αρχικό διάνυσμα $\underline{x}^{(0)}$ αν και μόνο αν $\lim_{m\to\infty}\underline{\underline{G}}^m=\underline{\underline{0}}$ όπου $\underline{\underline{G}}=\underline{\underline{M}}^{-1}\cdot\underline{\underline{N}}$ ο πίνακας επανάληψης της μεθόδου.

Παραθέτουμε τώρα, χωρίς απόδειξη, τα εξής κριτήρια σύγκλισης της γενικής επαναληπτικής μεθόδου:

(α) $\lim_{m\to\infty}\underline{\underline{G}}^m=\underline{\underline{0}}$

(β) $\left\|\underline{\underline{G}}\right\|<1$

(γ) $\rho\left(\underline{\underline{G}}\right)<1$ όπου $\rho\left(\underline{\underline{G}}\right)$ η φασματική ακτίνα του πίνακα $\underline{\underline{G}}$ (για τον ορισμό της φασματικής ακτίνας δείτε το παράρτημα Π2).

Ειδικά για τις επαναληπτικές μεθόδους Jacobi και Gauss-Seidel οι οποίες αποτελούν ειδική περίπτωση της γενικής επαναληπτικής μεθόδου με πίνακες επανάληψης $\underline{\underline{G}}_J = -\underline{\underline{D}}^{-1} \cdot \left(\underline{\underline{L}} + \underline{\underline{U}}\right)$ και $\underline{\underline{G}}_{GS} = -\left(\underline{\underline{L}} + \underline{\underline{D}}\right)^{-1} \cdot \underline{\underline{U}}$, αντίστοιχα ισχύει και το παρακάτω κριτήριο:

Έστω ότι ο πίνακας $\underline{\underline{A}}$ έχει αυστηρά κυριαρχική διαγώνιο. Τότε για τις μεθόδους Jacobi και Gauss-Seidel ισχύει $\left\|\underline{\underline{G}}_J\right\|_\infty < 1$ και $\left\|\underline{\underline{G}}_{GS}\right\|_\infty < 1$ και επομένως οι μέθοδοι αυτές συγκλίνουν.

Να σημειωθεί ότι όσο μικρότερη είναι η φασματική ακτίνα του πίνακα επανάληψης, τόσο πιο γρήγορα συγκλίνει η αντίστοιχη επαναληπτική μέθοδος. Με βάση το γεγονός αυτό επιλέγεται (όταν φυσικά αυτό είναι δυνατόν) και η παράμετρος ω της μεθόδου διαδοχικής υπερ χαλάρωσης.

Κεφάλαιο 4
Πολυωνυμική παρεμβολή και προσέγγιση συναρτήσεων

4.1. Εισαγωγή

Στο κεφάλαιο αυτό θα αναπτύξουμε μεθόδους με σκοπό
(α) την προσέγγιση γνωστών συναρτήσεων με άλλες απλούστερης μορφής και
(β) την προσέγγιση μίας άγνωστης συνάρτησης με βάση αριθμητικά δεδομένα της συνάρτησης και των παραγώγων της.
Θα περιοριστούμε σε πραγματικές συναρτήσεις μίας πραγματικής μεταβλητής ενώ τόσο για τον πρώτο όσο και για τον δεύτερο σκοπό η μεθοδολογία η οποία ακολουθείται είναι ακριβώς η ίδια και ονομάζεται *παρεμβολή*. Οι νέες, απλούστερες συναρτήσεις οι οποίες χρησιμοποιούνται για την προσέγγιση των αρχικών συναρτήσεων είναι:
(i) πολυώνυμα
(ii) εκθετικές συναρτήσεις
(iii) τριγωνομετρικές συναρτήσεις και
(iv) ρητές συναρτήσεις (δηλαδή πηλίκα πολυωνύμων)
Η πιο εύκολη επιλογή είναι η (i) διότι όλες οι πράξεις μεταξύ πολυώνυμων είναι ιδιαίτερα εύκολες ενώ τόσο η παραγώγιση όσο και η ολοκλήρωσή τους είναι η απλούστερη δυνατή.
Λέμε ότι η νέα συνάρτηση παρεμβάλλεται στην αρχική συνάρτηση ή στην άγνωστη συνάρτηση στα σημεία $\left(x_i, f\left(x_i\right)\right), \quad i=1,2,3,...,n$.

Αρχικά θα αποδείξουμε την ύπαρξη και μοναδικότητα του πολυωνύμου παρεμβολής, στην συνέχεια θα βρούμε μία σχέση για το σφάλμα της παρεμβολής και τέλος θα περιγράψουμε 3 δυνατούς τρόπους για την κατασκευή του.
(1) τον άμεσο τρόπο
(2) την μέθοδο κατά Lagrange και
(3) την μέθοδο κατά Newton.

4.2. Ύπαρξη και μοναδικότητα του πολυωνύμου παρεμβολής

Πρόταση: Έστω $x_0, x_1, x_2, ..., x_n \in \mathbb{R}$ $n+1$ σημεία ανά δύο διαφορετικά μεταξύ τους καθώς και $y_0, y_1, y_2, ..., y_n \in \mathbb{R}$. Τότε υπάρχει ακριβώς ένα πολυώνυμο, p, βαθμού το πολύ n (δηλαδή $p \in P_n$) τέτοιο ώστε $p\left(x_i\right) = y_i, \quad i=0,1,2,3,...,n$.

Απόδειξη: Έστω $p(x)=a_0+a_1x+a_2x^2+...+a_nx^n$ το ζητούμενο πολυώνυμο όπου $a_0,a_1,...,a_n$ είναι οι άγνωστοι συντελεστές του πολυωνύμου οι οποίοι πρέπει να προσδιοριστούν από τις δοσμένες συνθήκες $p(x_i)=y_i,\quad i=0,1,2,3,...,n$. Οι συνθήκες αυτές γράφονται αναλυτικά ως εξής:

$p(x_0)=a_0+a_1x_0+a_2x_0^2+...+a_nx_0^n=y_0$

$p(x_1)=a_0+a_1x_1+a_2x_1^2+...+a_nx_1^n=y_1$

....

$p(x_n)=a_0+a_1x_n+a_2x_n^2+...+a_nx_n^n=y_n$

Τις παραπάνω εξισώσεις μπορούμε να τις γράψουμε σε μορφή πινάκων και διανυσμάτων ως εξής:

$$\underbrace{\begin{bmatrix} 1 & x_0 & x_0^2 & x_0^3 & .... & x_0^n \\ 1 & x_1 & x_1^2 & x_1^3 & .... & x_1^n \\ ..... & & & & & \\ 1 & x_n & x_n^2 & x_n^3 & .... & x_n^n \end{bmatrix}}_{\underline{\underline{V}}} \cdot \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ ... \\ a_n \end{bmatrix}}_{\underline{a}} = \underbrace{\begin{bmatrix} y_0 \\ y_1 \\ ... \\ y_n \end{bmatrix}}_{\underline{y}}$$

είτε $\underline{\underline{V}}\cdot\underline{a}=\underline{y}$, $\underline{\underline{V}}\in\mathbb{R}^{n+1,n+1}$, $\underline{a}\in\mathbb{R}^{n+1}$, $\underline{y}\in\mathbb{R}^{n+1}$ όπου ο πίνακας $\underline{\underline{V}}$ και το διάνυσμα $\underline{y}$ είναι γνωστά εφόσον τα $x_i,y_i,\quad i=0,1,2,3,...,n$ είναι δεδομένα. Ο $\underline{\underline{V}}$ ονομάζεται πίνακας Vandermonde και η ορίζουσα του, η οποία ονομάζεται ορίζουσα Vandermonde, δίνεται από την σχέση $\det\left(\underline{\underline{V}}\right)=\prod_{i>j}\left(x_i-x_j\right)=\prod_{j=0}^{n-1}\prod_{i=j+1}^{n}\left(x_i-x_j\right)$ και είναι μη-μηδενική, $\det\left(\underline{\underline{V}}\right)\neq 0$, εφόσον όλα τα σημεία $x_j,\ j=0,1,..,n$ είναι όλα διαφορετικά μεταξύ τους. Σε αυτή την περίπτωση, όπως γνωρίζουμε, το γραμμικό σύστημα έχει μία και μοναδική λύση, $\underline{a}=\underline{\underline{V}}^{-1}\cdot\underline{y}$.

Στην περίπτωση αυτή θα λέμε ότι το πολυώνυμο $p\in P_n$ παρεμβάλλεται στην συνάρτηση f στα σημεία $x_0,x_1,...,x_n$. Τότε το p ονομάζεται *πολυώνυμο παρεμβολής*.

4.3. Σφάλμα της πολυωνυμικής παρεμβολής

Πρόταση: Έστω $n\in\mathbb{N}$ και $f\in C^{n+1}[a,b]$ καθώς και $x_0,x_1,x_2,...,x_n\in\mathbb{R}$ όλα διαφορετικά μεταξύ τους. Αν $p\in P_n$ το πολυώνυμο που παρεμβάλλεται στην f στα σημεία αυτά

τότε $\forall x\in[a,b],\exists\xi\in(a,b)$ τέτοιο ώστε $\boxed{f(x)-p(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}\prod_{i=0}^{n}(x-x_i)}$ και

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

Απόδειξη: Αν $x\in\{x_0,x_1,...,x_n\}$ τότε το αριστερό μέλος, $f(x)-p(x)$, της σχέσης που θέλουμε να αποδείξουμε είναι μηδέν λόγω του τρόπου κατασκευής του πολυωνύμου παρεμβολής. Μηδέν όμως είναι και η ποσότητα $\prod_{i=0}^{n}(x-x_i)$. Επομένως αν $x\in\{x_0,x_1,...,x_n\}$ η ζητούμενη σχέση ισχύει.

Έστω τώρα ότι $x\in[a,b]$ και $x\notin\{x_0,x_1,...,x_n\}$. Ορίζουμε την βοηθητική συνάρτηση $\Phi(t)=\prod_{i=0}^{n}(t-x_i)$ για την οποία προφανώς ισχύει $\Phi(x_j)=\prod_{i=0}^{n}(x_j-x_i)=0$, $j=0,1,2,...,n$ και $\Phi(x)=\prod_{i=0}^{n}(x-x_i)\neq 0$ αν $x\notin\{x_0,x_1,...,x_n\}$. Επιπλέον ορίζουμε την συνάρτηση $\varphi(t)=f(t)-p(t)-\frac{f(x)-p(x)}{\Phi(x)}\Phi(t)$ όπου $t\in[a,b]$, $x\in[a,b]$, $x\notin\{x_0,x_1,...,x_n\}$. Παρατηρούμε ότι:

Για $t=x_0$, $\varphi(x_0)=\underbrace{f(x_0)-p(x_0)}_{=0}-\frac{f(x)-p(x)}{\Phi(x)}\underbrace{\Phi(x_0)}_{=0}=0$

Για $t=x_1$, $\varphi(x_1)=\underbrace{f(x_1)-p(x_1)}_{=0}-\frac{f(x)-p(x)}{\Phi(x)}\underbrace{\Phi(x_1)}_{=0}=0$

...

Για $t=x_n$, $\varphi(x_n)=\underbrace{f(x_n)-p(x_n)}_{=0}-\frac{f(x)-p(x)}{\Phi(x)}\underbrace{\Phi(x_n)}_{=0}=0$

Για $t=x$, $\varphi(x)=f(x)-p(x)-\frac{f(x)-p(x)}{\Phi(x)}\Phi(x)=0$

Επομένως η συνάρτηση φ έχει ρίζες τα $x_0,x_1,x_2,...,x_n,x$ το πλήθος των οποίων είναι «n+2». Επιπλέον επειδή $f,p,\Phi\in C^{n+1}[a,b]$ άρα και $\varphi\in C^{n+1}[a,b]$. Τότε σύμφωνα με

το θεώρημα του Rolle, εφόσον η φ έχει τουλάχιστον n+2 διαφορετικές ρίζες, η φ' έχει n+1 ρίζες, η φ'' έχει n ρίζες, ... , η $\varphi^{(n+1)}$ έχει 1 ρίζα, έστω ξ, και επομένως

$$\varphi^{(n+1)}(\xi)=0 \tag{1}$$

Από τον ορισμό της φ βρίσκουμε την $\varphi^{(n+1)}$:

$$\varphi^{(n+1)}(t)=f^{(n+1)}(t)-p^{(n+1)}(t)-\frac{f(x)-p(x)}{\Phi(x)}\Phi^{(n+1)}(t) \tag{2}$$

(προσέξτε ότι στον ορισμό της φ η ανεξάρτητη μεταβλητή είναι το t και όχι το x). Εφόσον το p είναι ένα πολυώνυμο το πολύ βαθμού n, άρα η n+1 παράγωγός του είναι μηδέν:

$$p^{(n+1)}(t)=0 \tag{3}$$

Επιπλέον, από τον ορισμό της, η συνάρτηση $\Phi=\Phi(t)$ είναι πολυώνυμο βαθμού n+1 και επομένως:

$$\Phi^{(n+1)}(t)=(n+1)! \tag{4}$$

Από τις (2),(3) και (4) έχουμε:

$$\varphi^{(n+1)}(t)=f^{(n+1)}(t)-\frac{f(x)-p(x)}{\Phi(x)}(n+1)!\overset{(1)}{\Rightarrow}\varphi^{(n+1)}(\xi)=f^{(n+1)}(\xi)-\frac{f(x)-p(x)}{\Phi(x)}(n+1)!=0$$

Λύνοντας την τελευταία ισότητα έχουμε $\frac{f^{(n+1)}(\xi)}{(n+1)!}\Phi(x)=f(x)-p(x)$ και αντικαθιστώντας την $\Phi(x)$ παίρνουμε την ζητούμενη σχέση. Στην συνέχεια, παίρνοντας την μέγιστη νόρμα στην σχέση αυτή και εφαρμόζοντας την τριγωνική ανισότητα καταλήγουμε στην δεύτερη σχέση.

4.4. Κατασκευή του πολυωνύμου παρεμβολής

4.4.1. Άμεση μέθοδος

Η άμεση μέθοδος προκύπτει με απευθείας επίλυση του γραμμικού συστήματος $\underline{\underline{V}}\cdot\underline{a}=\underline{y}$ (για παράδειγμα με απαλοιφή Gauss) η οποία θα δώσει το διάνυσμα $\underline{a}$ το οποίο περιέχει τους άγνωστους συντελεστές του πολυωνύμου.

4.4.2 Μέθοδος Lagrange

Σύμφωνα με την μέθοδο αυτή, το πολυώνυμο παρεμβολής δίνεται από την σχέση:

$$p(x)=\sum_{i=0}^{n} f(x_i)L_i(x)=f(x_0)L_0(x)+f(x_1)L_1(x)+...+f(x_n)L_n(x) \quad (*)$$

όπου $L_i(x)$ είναι τα λεγόμενα πολυώνυμα Lagrange ως προς τα σημεία $\{x_0,x_1,x_2,...,x_n\}$ τα οποία είναι το πολύ βαθμού «n» και έχουν την ιδιότητα

$$L_i(x_j)=\delta_{ij}=\begin{cases}0, & i\neq j\\ 1, & i=j\end{cases} \quad (**)$$

Ας δούμε κατ' αρχήν κατά πόσο το πολυώνυμο παρεμβολής ικανοποιεί τις δοσμένες συνθήκες, $p(x_i)=f(x_i), \quad i=0,1,2,...,n$. Από την σχέση (*) θα έχουμε:

Για $x=x_0$, $p(x_0)=f(x_0)\underbrace{L_0(x_0)}_{=1}+f(x_1)\underbrace{L_1(x_0)}_{=0}+...+f(x_n)\underbrace{L_n(x_0)}_{=0}=f(x_0)$

Για $x=x_1$, $p(x_1)=f(x_0)\underbrace{L_0(x_1)}_{=0}+f(x_1)\underbrace{L_1(x_1)}_{=1}+...+f(x_n)\underbrace{L_n(x_1)}_{=0}=f(x_1)$

....

Για $x=x_n$, $p(x_n)=f(x_0)\underbrace{L_0(x_n)}_{=0}+f(x_1)\underbrace{L_1(x_n)}_{=0}+...+f(x_n)\underbrace{L_n(x_n)}_{=1}=f(x_n)$

Πράγματι λοιπόν το πολυώνυμο στην μορφή (*) ικανοποιεί τις συνθήκες παρεμβολής λόγω των ιδιοτήτων των πολυωνύμων Lagrange (σχέση (**)). Απομένει να κατασκευάσουμε τα πολυώνυμα Lagrange. Λόγω της (**) το L_i πρέπει να περνά από τα σημεία, $\underbrace{(x_0,0),(x_1,0),...,(x_{i-1},0),(x_i,1),(x_{i+1},0),...,(x_n,0)}_{n+1 \text{ σημεια}}$, το πλήθος των οποίων είναι n+1. Άρα θα είναι θα είναι βαθμού n και μάλιστα θα είναι μοναδικό. Τα σημεία αυτά στην πραγματικότητα μας γνωστοποιούν και τις ρίζες του οι οποίες είναι $\underbrace{x_0,x_1,...,x_{i-1},x_{i+1},...,x_n}_{n \text{ ριζες}}$ και επομένως μπορούμε να το εκφράσουμε στην μορφή

$L_i(x)=\alpha\prod_{\substack{j=0\\ i\neq j}}^{n}(x-x_j)$ όπου το α είναι μία σταθερά η οποία θα προσδιοριστεί από την μόνη συνθήκη που δεν έχει ικανοποιηθεί, δηλαδή την $L_i(x_i)=1$. Άρα $L_i(x_i)=\alpha\prod_{\substack{j=0\\ i\neq j}}^{n}(x_i-x_j)=1\Rightarrow\alpha=1\Big/\prod_{\substack{j=0\\ i\neq j}}^{n}(x_i-x_j)$. Έτσι η τελική μορφή των πολυωνύμων Lagrange δίνεται από την σχέση $L_i(x)=\prod_{\substack{j=0\\ i\neq j}}^{n}(x-x_j)\Big/\prod_{\substack{j=0\\ i\neq j}}^{n}(x_i-x_j)=\prod_{j=0}^{n}\frac{x-x_j}{x_i-x_j}$

Επομένως το πολυώνυμο παρεμβολής με την μέθοδο Lagrange δίνεται από την σχέση:

$$p(x)=\sum_{i=0}^{n} f(x_i)L_i(x), \quad L_i(x)=\prod_{\substack{j=0\\ i\neq j}}^{n}\frac{x-x_j}{x_i-x_j}$$

4.4.3 Μέθοδος Newton

Η μέθοδος αυτή είναι άλλη μία προσέγγιση στο πολυώνυμο παρεμβολής και εισήχθηκε από τον Isaac Newton. Έστω τα «n+1» σημεία $(x_i, f_i), i=0,1,2,...,n$ όπου έχει χρησιμοποιηθεί η συντομογραφία $f_i \equiv f(x_i)$, και έστω ότι επιχειρούμε να βρούμε το πολυώνυμο παρεμβολής $p \in P_n$ της άγνωστης συνάρτησης $f(x)$ σε αυτά τα σημεία, στην μορφή:

$$p(x)=a_0+a_1(x-x_0)+a_2(x-x_0)(x-x_1)+...+a_n(x-x_0)(x-x_1)...(x-x_{n-1})$$

Η παραπάνω έκφραση ονομάζεται *πολυώνυμο Newton*. Στο σημείο αυτό να πρέπει να τονισθεί ότι εφόσον το πολυώνυμο παρεμβολής είναι μοναδικό δεν έχει σημασία με ποιον τρόπο θα το εκφράσουμε. Έτσι, είτε εκφραστεί στην μορφή της άμεσης μεθόδου, είτε στην μορφή κατά Lagrange, είτε στην μορφή κατά Newton πρόκειται ακριβώς για το ίδιο πολυώνυμο. Οποιαδήποτε μέθοδος και αν χρησιμοποιηθεί θα πρέπει να δώσει το ίδιο αποτέλεσμα.

Οι συνθήκες κατασκευής του πολυωνύμου $p(x_i)=f_i, \quad i=0,1,2,...,n$, σύμφωνα με την παραπάνω μορφή δίνονται αναλυτικά ως εξής:

$$p(x_0)=f_0=a_0$$
$$p(x_1)=f_1=a_0+a_1(x_1-x_0)$$
$$p(x_2)=f_2=a_0+a_1(x_2-x_0)+a_2(x_2-x_0)(x_2-x_1)$$
...
$$p(x_n)=f_n=a_0+a_1(x_n-x_0)+a_2(x_n-x_0)(x_n-x_1)+...+a_n(x_n-x_0)(x_n-x_1)...(x_n-x_{n-1})$$

Το σύστημα αυτό μπορούμε να το γράψουμε με μορφή πινάκων και διανυσμάτων, δηλαδή ως $\underline{\underline{U}}\cdot\underline{a}=\underline{f}$ όπου $\underline{\underline{U}}\in\mathbb{R}^{n+1,n+1}, \underline{a}\in\mathbb{R}^{n+1}, \underline{f}\in\mathbb{R}^{n+1}$ με

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & .... & 0 \\ 1 & x_1\text{-}x_0 & 0 & 0 & .... & 0 \\ 1 & x_2\text{-}x_0 & (x_2\text{-}x_0)(x_2\text{-}x_1) & 0 & .... & 0 \\ ..... & & & & & \\ 1 & x_n\text{-}x_0 & (x_n\text{-}x_0)(x_n\text{-}x_1) & (x_n\text{-}x_0)(x_n\text{-}x_1)(x_n\text{-}x_2) & .... & U_{n,n} \end{bmatrix}}_{\underline{\underline{U}}} \cdot \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ ... \\ a_n \end{bmatrix}}_{\underline{a}} = \underbrace{\begin{bmatrix} f_0 \\ f_1 \\ ... \\ f_n \end{bmatrix}}_{\underline{f}}$$

όπου $U_{n,n} = (x_n\text{-}x_0)(x_n\text{-}x_1)(x_n\text{-}x_2)\cdot\ldots\cdot(x_n\text{-}x_{n-1})$. Παρατηρούμε ότι πρόκειται για ένα γραμμικό σύστημα, ο πίνακας, $\underline{\underline{U}}$, του οποίου είναι κάτω τριγωνικός και η ορίζουσά του είναι $\det\left(\underline{\underline{U}}\right) = \prod_{i=1}^{n+1} U_{i,i}$. Τα διαγώνια στοιχεία του είναι $U_{1,1} = 1$, $U_{i,i} = \prod_{j=0}^{i-2}\left(x_{i-1} - x_j\right), i \geq 2$ και επομένως, εφόσον τα $x_i,\ i = 0,1,...,n$ είναι διαφορετικά μεταξύ τους, άρα $U_{i,i} \neq 0$ και συνεπώς $\det\left(\underline{\underline{U}}\right) \neq 0$. Άρα το γραμμικό σύστημα $\underline{\underline{U}} \cdot \underline{a} = \underline{f}$ έχει μοναδική λύση (όπως φυσικά ήταν αναμενόμενο). Όπως ήδη γνωρίζουμε, εφόσον ο $\underline{\underline{U}}$ είναι κάτω τριγωνικός το σύστημα αυτό μπορεί να λυθεί πολύ εύκολα με τον αλγόριθμο της προς τα εμπρός αντικατάστασης ο οποίος απαιτεί πράξεις τάξης $O\left(n^2\right)$ (όπου n το πλήθος των εξισώσεων ή αγνώστων).

Μέθοδος Newton με χρήση διαιρεμένων διαφορών

Στην συνέχεια θα δούμε ένα ακόμα τρόπο υπολογισμού του πολυωνύμου παρεμβολής με χρήση των λεγόμενων «διαιρεμένων διαφορών». Το σημείο εκκίνησης είναι το πολυώνυμο Newton που δόθηκε παραπάνω το οποίο βέβαια δημιουργεί ένα γραμμικό σύστημα το οποίο είναι σε τριγωνική μορφή. Ξεκινώντας από την πρώτη εξίσωση και κατεβαίνοντας μπορούμε να προσδιορίσουμε όλους τους άγνωστους συντελεστές $a_i,\ i = 0,1,...,n$. Είναι εύκολο να παρατηρήσουμε ότι ο συντελεστής a_i εξαρτάται από τις τιμές f_j και $x_j,\ j = 0,1,...,i-1$. Χρησιμοποιούμε τον συμβολισμό $a_j = f\left[x_0, x_1, ..., x_j\right]$ για να υποδηλώσουμε αυτήν την εξάρτηση από τους δείκτες και τις αντίστοιχες τιμές της συνάρτησης. Έτσι έχουμε:

$$a_0 = f\left[x_0\right] = f_0$$

$$a_1 = f\left[x_0, x_1\right] = \frac{f_0}{x_0 - x_1} - \frac{f_1}{x_0 - x_1}$$

$$a_2 = f\left[x_0, x_1, x_2\right] = \frac{f\left[x_0, x_1\right] - f\left[x_1, x_2\right]}{x_0 - x_2} = \frac{f_0}{\left(x_0 - x_1\right)\left(x_0 - x_2\right)} + \frac{f_1}{\left(x_1 - x_0\right)\left(x_1 - x_2\right)} + \frac{f_2}{\left(x_0 - x_2\right)\left(x_2 - x_1\right)}$$

κ.τ.λ. για τους υπόλοιπους συντελεστές. Κάνοντας σύγκριση με τους συντελεστές των όρων x^n πολυωνύμου Lagrange καταλήγουμε στην γενική περίπτωση ότι:

$$f\left[x_0, x_1, \ldots, x_n\right] = \frac{f\left[x_0, x_1, \ldots, x_{n-1}\right] - f\left[x_1, x_2, \ldots, x_n\right]}{x_0 - x_n}$$

Οι εκφράσεις αυτές ονομάζονται διαιρεμένες διαφορές (divided differences) τάξεως «n» της συνάρτησης f στα σημεία $x_0, x_1, \ldots, x_n$ και είναι εξαιρετικά εύκολο να υπολογισθούν εάν τοποθετηθούν σε έναν πίνακα όπως φαίνεται παρακάτω:

Πίνακας υπολογισμού διαιρεμένων διαφορών				
x_0	**f[x₀]**			
		f[x₀,x₁]		
x_1	f[x₁]		**f[x₀,x₁,x₂]**	
		f[x₁,x₂]		**f[x₀,x₁,x₂,x₃]**
x_2	f[x₂]		f[x₁,x₂,x₃]	
		f[x₂,x₃]		
x_3	f[x₃]			

Στην περίπτωση αυτή το πολυώνυμο παρεμβολής γράφεται ως εξής:

$$p_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + \ldots +$$
$$+ f[x_0, x_1, \ldots, x_n](x - x_0)(x - x_1)\ldots(x - x_{n-1})$$

Στον πίνακα υπολογισμού των διαιρεμένων διαφορών οι τιμές που μας ενδιαφέρουν είναι αυτές που έχουν σημειωθεί με έντονα γράμματα. Να σημειωθεί ότι οι διαιρεμένες διαφορές είναι ο πιο αποδοτικός τρόπος για να υπολογίσουμε ένα πολυώνυμο $p = p(x)$ ιδιαίτερα στην περίπτωση που θέλουμε να υπολογίσουμε την τιμή του πολυωνύμου για πολλές τιμές του x. Αυτό ισχύει διότι οι διαιρεμένες διαφορές, όπως μπορεί να αποδειχτεί, είναι ανεξάρτητες από το x. Ένα επιπλέον πλεονέκτημα των διαιρεμένων

διαφορών είναι ότι η προσθήκη επιπλέον σημείων στον πίνακα δεν αλλάζει τις ήδη υπολογισμένες τιμές του πίνακα.

Παράδειγμα: Έστω τα σημεία $(1,0),(-1,-3),(2,4)$ και ότι ζητείται να βρεθεί το πολυώνυμο παρεμβολής με χρήση των διαιρεμένων διαφορών. Σχηματίζουμε τον πίνακα υπολογισμού των διαιρεμένων διαφορών όπως φαίνεται παρακάτω:

Πίνακας υπολογισμού διαιρεμένων διαφορών			
+1	0		
		$\frac{0-(-3)}{1-(-1)}=\frac{3}{2}$	
-1	-3		$\frac{3/2-7/3}{1-2}=\frac{5}{6}$
		$\frac{(-3)-4}{(-1)-2}=\frac{7}{3}$	
2	4		

Επομένως έχουμε $p_2(x)=0+\frac{3}{2}(x-1)+\frac{5}{6}(x-1)(x+1)=\frac{1}{6}(5x^2+9x-14)$.

Παράδειγμα: Έστω ότι μας δίνονται τα παρακάτω δεδομένα μιας συνάρτησης $f=f(x)$, για την οποία δεν γνωρίζουμε τον αναλυτικό τύπο. Υπολογίστε το πολυώνυμο παρεμβολής της *f* χρησιμοποιώντας αρχικά τα τέσσερα πρώτα σημεία του πίνακα με (α) την άμεση μέθοδο, (β) την μέθοδο Lagrange (γ) την μέθοδο Newton (είτε με χρήση του αλγορίθμου της προς τα εμπρός αντικατάστασης είτε με χρήση των διαιρεμένων διαφορών). Επαναλάβετε την ίδια διαδικασία χρησιμοποιώντας όλα τα σημεία του πίνακα. Ποια είναι η πρόβλεψη για την $f(x=3.0)$ και στις δύο περιπτώσεις?

x_i	3.2	2.7	1.0	4.8	5.6
$f(x_i)$	22.0	17.8	14.2	38.3	51.7

Λύση: Έστω ότι χρησιμοποιούμε τα 4 πρώτα σημεία $(x_i, f(x_i))$ του παραπάνω πίνακα για να προσδιορίσουμε το πολυώνυμο παρεμβολής τρίτου βαθμού $p_3(x) = a_3x^3 + a_2x^2 + a_1x^1 + a_0$. Ακολουθώντας την άμεση μέθοδο προσδιορισμού των συντελεστών ενός πολυωνύμου, θα έχουμε:

$$a_3(3.2)^3 + a_2(3.2)^2 + a_1(3.2)^1 + a_0 = 22.0$$

$$a_3(2.7)^3 + a_2(2.7)^2 + a_1(2.7)^1 + a_0 = 17.8$$

$$a_3(1.0)^3 + a_2(1.0)^2 + a_1(1.0)^1 + a_0 = 14.2$$

$$a_3(4.8)^3 + a_2(4.8)^2 + a_1(4.8)^1 + a_0 = 38.3$$

Το παραπάνω σύστημα μπορεί να γραφεί σε μορφή πινάκων ως εξής:

$$\begin{bmatrix} 1 & (3.2)^1 & (3.2)^2 & (3.2)^3 \\ 1 & (2.7)^1 & (2.7)^2 & (2.7)^3 \\ 1 & (1.0)^1 & (1.0)^2 & (1.0)^3 \\ 1 & (4.8)^1 & (4.8)^2 & (4.8)^3 \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \end{bmatrix}$$

Επιλύνοντας το σύστημα αυτό λαμβάνουμε $a_3 = -0.52748$, $a_2 = 6.49523$, $a_1 = -16.1177$, $a_0 = 24.3499$ και επομένως $p_3(x) = -0.52748x^3 + 6.49523x^2 - 16.1177x + 24.3499$, οπότε προκύπτει ότι $f(x = 3.0) \approx p_3(x = 3.0) = 20.212$.

Αν είχαμε χρησιμοποιήσει και το τελευταίο σημείο (5.6, 51.7) θα βρίσκαμε το εξής πολυώνυμο, τετάρτου βαθμού:

$$p_4(x) = 0.255838x^4 - 3.52078x^3 + 18.6885x^2 - 36.1836x + 34.96$$

οπότε $f(x = 3.0) \approx p_4(x = 3.0) = 20.2672$.

Ωστόσο, θα πρέπει να σημειωθεί ότι το νέο σύστημα 5x5 που προέκυψε για το πολυώνυμο τετάρτου βαθμού είναι σε ιδιαίτερα κακή κατάσταση, γεγονός που σημαίνει ότι αν αλλαχτούν σε μικρό βαθμό οι τιμές της συνάρτησης θα προκληθούν μεγάλες αλλαγές στους συντελεστές του πολυωνύμου.

Αν αντί για την άμεση μέθοδο χρησιμοποιήσουμε πολυώνυμα Lagrange, δηλαδή την σχέση $p_3(x)=\sum_{i=0}^{3}L_i(x)\,f(x_i)=\sum_{i=0}^{3}\prod_{\substack{j=0\\ i\neq j}}^{3}\frac{x-x_j}{x_i-x_j}f(x_i)$ θα έχουμε:

$$p_3(x)=\frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)}f(x_0)+\frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}f(x_1)+$$
$$+\frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_1)(x_2-x_1)(x_2-x_3)}f(x_2)+\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}f(x_3)$$

Θέτοντας τιμές στην παραπάνω σχέση προκύπτει πολύ εύκολα $f(x=3.0)\approx p_3(x=3.0)=20.212$. Όμοια υπολογίζουμε και το $p_4(x)=\sum_{i=0}^{4}L_i(x)\,f(x_i)=\sum_{i=0}^{4}\prod_{\substack{j=0\\ i\neq j}}^{4}\frac{x-x_j}{x_i-x_j}f(x_i)$. Είναι προφανές ότι η ζητούμενη τιμή της συνάρτησης προκύπτει πολύ πιο εύκολα με την μέθοδο Lagrange παρά με την άμεση μέθοδο, η οποία απαιτεί την επίλυση ενός γραμμικού συστήματος γεγονός που την κάνει μη-αποτελεσματική από πρακτική άποψη.

Στην συνέχεια θα γίνει εφαρμογή της κατασκευής πολυωνύμου με την μέθοδο του Newton, με χρήση των διαιρεμένων διαφορών, όπως φαίνεται στον πίνακα που ακολουθεί:

3.2	**22.0**				
	$\frac{22-17.8}{3.2-2.7}=$ **8.4**				
2.7	17.8		$\frac{8.4-2.11765}{3.2-1}=$ **2.85561**		
	$\frac{17.8-14.2}{2.7-1}=2.11765$			$\frac{2.85561-2.01165}{3.2-4.8}=$ **-0.527475**	
1.0	14.2		$\frac{2.11765-6.34211}{2.7-4.8}=2.01165$		**0.255834**
	$\frac{14.2-38.3}{1-4.8}=6.34211$			$\frac{2.01165-2.26258}{2.7-5.6}=0.0865276$	
4.8	38.3		$\frac{6.34211-16.75}{1.0-5.6}=2.26258$		
	$\frac{38.3-51.7}{4.8-5.6}=16.75$				
5.6	51.7				

Οι αριθμοί με έντονα γράμματα είναι οι ζητούμενοι συντελεστές. Έτσι θα έχουμε

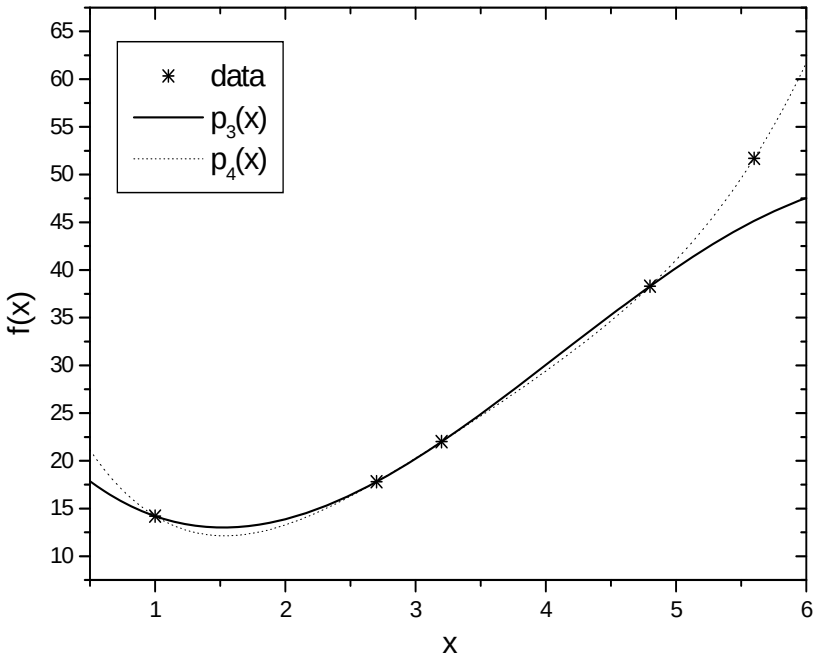
$$p_3(x) = 22 + 8.4(x-3.2) + 2.85561(x-3.2)(x-2.7) - 0.527475(x-3.2)(x-2.7)(x-1.0)$$

και

$$p_4(x) = p_3(x) + 0.255834(x-3.2)(x-2.7)(x-1.0)(x-4.8)$$

Τα δύο αυτά πολυώνυμα για $x = 3.0$ δίνουν $p_3(3.0) \simeq 20.212$ και $p_4(3.0) \simeq 20.269$.

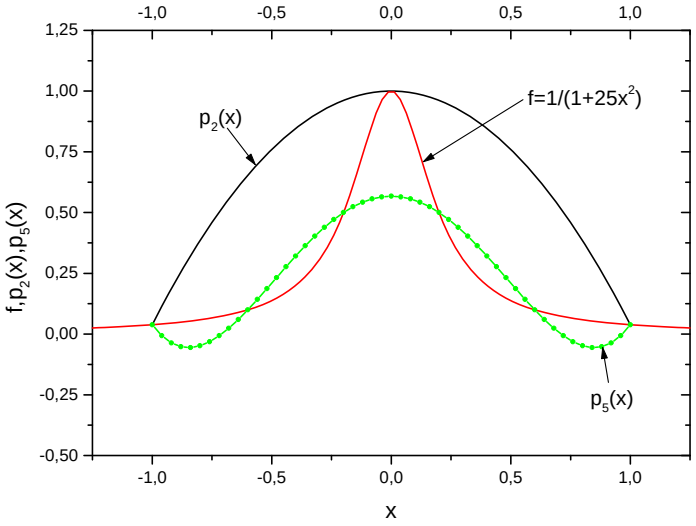
Τόσο τα δεδομένα, όσο και τα πολυώνυμα $p_3(x)$ και $p_4(x)$ απεικονίζονται στο παρακάτω διάγραμμα:

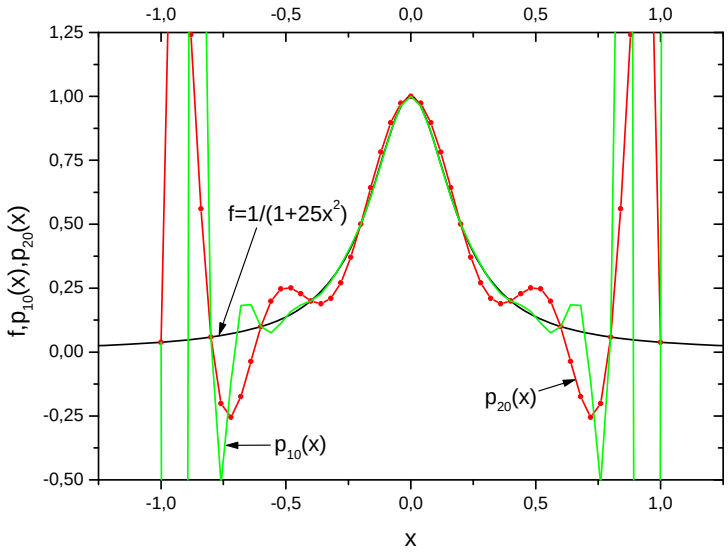


4.5. Οι κίνδυνοι της πολυωνυμικής παρεμβολής και η συνάρτηση του Runge

Εδώ θα πρέπει να σημειωθεί ότι έχει παρατηρηθεί ότι όσο μεγαλώνει ο βαθμός του πολυωνύμου παρεμβολής τότε η παρεμβολή δεν είναι επιτυχής, δηλαδή οι εκτιμήσεις για τις τιμές της συνάρτησης που μας ενδιαφέρουν είναι ιδιαίτερα μη-ακριβείς (συνήθως παρατηρούνται γρήγορες ταλαντώσεις στις τιμές του πολυωνύμου). Ένα κλασσικό παράδειγμα των κινδύνων της πολυωνυμικής παρεμβολής αναφέρεται από τον C. Runge το 1901. Για την ιδιαίτερα απλή συνάρτηση $f(x) = \frac{1}{1+25x^2}$, $x \in [-1, +1]$

η ακολουθία $p_n(x)$ καθώς το «n» μεγαλώνει, αποκλίνει στο διάστημα $0.726 < |x| < 1$. Στα επόμενα δύο διαγράμματα φαίνεται η συνάρτηση καθώς και τα πολυώνυμα $p_2(x), p_5(x)$ (πρώτο διάγραμμα) και τα $p_{10}(x), p_{20}(x)$ (δεύτερο διάγραμμα). Τα πολυώνυμα έχουν κατασκευασθεί στους ομοιόμορφα κατανεμημένους κόμβους $x_j = -1 + \frac{2}{n} j,\ j = 0,1,2,...,n$.





Μπορεί να αποδειχτεί ότι για δεδομένο «n» η ποσότητα $\max_{x\in[a,b]}\left(\prod_{i=0}^{n}\left|(x-x_i)\right|\right)$ από την οποία εξαρτάται το άνω φράγμα του σφάλματος της πολυωνυμικής παρεμβολής ελαχιστοποιείται αν αντί των σημείων $x_j = a + \frac{(b-a)}{n} j,\ j = 0,1,2,...,n$, τα οποία φυσικά είναι ισαπέχοντα, χρησιμοποιηθούν τα λεγόμενα σημεία Chebyshev τα οποία δίνονται από την σχέση $x_j = \cos\left(\frac{\pi}{2}\left(\frac{2j+1}{n+1}\right)\right),\ j = 0,1,2,...,n$ (εδώ απαιτείται προσοχή: τα σημεία Chebyshev δίνονται στο διάστημα [-1,1]. Επομένως αν το αρχικό διάστημα είναι το [a,b] τότε πρέπει πρώτα να γίνει ένας γραμμικός μετασχηματισμός έτσι ώστε η συνάρτηση που μας ενδιαφέρει να ορίζεται στο [-1,1]).

4.6. Παρεμβολή Hermite

Στην περίπτωση που εκτός από δεδομένα για την τιμή της συνάρτησης σε ένα συγκεκριμένο πλήθος σημείων έχουμε και δεδομένα για τις τιμές της παραγώγου της συνάρτησης, τότε χρησιμοποιούμε την λεγόμενη παρεμβολή τύπου Hermite. Θα παρουσιάσουμε στην συνέχεια την παρεμβολή αυτή για την περίπτωση όπου δεδομένα είναι οι τιμές της συνάρτησης και της πρώτης παραγώγου της, δηλαδή τα $(x_i, f(x_i)), (x_i, f'(x_i)), i = 0,1,2,...,n$. Τότε έχουμε 2(n+1) πλήθος δεδομένων οπότε και μπορεί να αποδειχτεί ότι υπάρχει ένα πολυώνυμο το πολύ βαθμού 2n+1 το οποίο να διέρχεται από όλα αυτά τα δεδομένα. Το πολυώνυμο έχει την μορφή

$$p(x) = \sum_{i=0}^{n}\left\{ f(x_i)H_i^{(1)}(x) + f'(x_i)H_i^{(2)}(x)\right\} \quad (1)$$

όπου $H_i^{(1)}(x), H_i^{(2)}(x)$ είναι μοναδικά πολυώνυμα το πολύ βαθμού 2n+1 τα οποία ικανοποιούν τις συνθήκες

$$H_i^{(1)}(x_j) = \delta_{ij} = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases} \quad (2)$$
$$\frac{dH_i^{(1)}}{dx}(x_j) = 0$$

και

$$H_i^{(2)}(x_j) = 0$$
$$\frac{dH_i^{(2)}}{dx}(x_j) = \delta_{ij} = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases} \quad (3)$$

Με βάση τις ιδιότητες αυτές αποδεικνύεται ότι έχουν την μορφή:

$$H_i^{(1)}(x)=\left[1-2(x-x_i)L_i'(x_i)\right]L_i^2(x)$$
$$H_i^{(2)}(x)=(x-x_i)L_i^2(x) \quad (4)$$

όπου $L_i(x), i=0,1,2,...,n$ είναι τα αντίστοιχα πολυώνυμα Lagrange της συνάρτησης f στα σημεία $\{x_0,x_1,...,x_n\}$. Ας επαληθεύσουμε πρώτα ότι πράγματι η μορφή (4) ικανοποιεί τις συνθήκες (2) και (3).

Για $x=x_i$, $H_i^{(1)}(x_i)=\left[1-2(x_i-x_i)L_i'(x_i)\right]L_i^2(x_i)=[1]1^2=1$ αφού $L_i(x_i)=1$ εκ' κατασκευής.

Για $x=x_j$, $H_i^{(1)}(x)=\left[1-2(x_j-x_i)L_i'(x_i)\right]L_i^2(x_j)=\left[1-2(x_j-x_i)L_i'(x_i)\right]0^2=0$ αφού $L_i(x_j)=0$ εκ' κατασκευής. Επιπλέον έχουμε ότι

$$\frac{dH_i^{(1)}(x)}{dx}=-2L_i'(x_i)L_i^2(x)+2\left[1-2(x-x_i)L_i'(x_i)\right]L_i'(x)L_i(x)$$

επομένως για $x=x_j$,

$$\frac{dH_i^{(1)}}{dx}(x_j)=-2L_i'(x_i)L_i^2(x_j)+2\left[1-2(x_j-x_i)L_i'(x_i)\right]L_i'(x_j)L_i(x_j)=0$$ αφού $L_i(x_j)=0$ εκ' κατασκευής.

Ακόμα, $H_i^{(2)}(x_j)=(x_j-x_i)L_i^2(x_j)=0$ αφού $L_i(x_j)=0$ και

$\frac{dH_i^{(2)}(x)}{dx}=L_i^2(x)+(x-x_i)2L_i'(x)L_i(x)$ οπότε

$$\frac{dH_i^{(2)}}{dx}(x_i)=L_i^2(x_i)+(x_i-x_i)2L_i'(x)L_i(x)=1^2=1$$

και $\frac{dH_i^{(2)}}{dx}(x_j)=L_i^2(x_j)+(x_j-x_i)2L_i'(x_j)L_i(x_j)=0$ αφού $L_i(x_j)=0$ εκ' κατασκευής.

Στην συνέχεια αποδεικνύουμε το παρακάτω θεώρημα:

Θεώρημα: Έστω $[a,b]\subset\mathbb{R}$ και $x_i, i=0,1,2,...,n$ σημεία ανά δύο διαφορετικά μεταξύ τους. Τότε υπάρχει μοναδικό πολυώνυμο τύπου Hermite $p\in P_{2n+1}$ που ορίζεται από τις συνθήκες $p(x_i)=f(x_i),\ p'(x_i)=f'(x_i), i=0,1,2,...,n$. Για το πολυώνυμο αυτό και $\forall x\in[a,b],\ \exists\xi\in(a,b)$ τέτοιο ώστε

$$\boxed{f(x)-p(x)=\frac{f^{(2n+2)}(\xi)}{(2n+2)!}\prod_{i=0}^{i=n}(x-x_i)^2} \quad (*) \text{ και } \|f(x)-p(x)\|_\infty\le\frac{\|f^{(2n+2)}\|_\infty}{(2n+2)!}\max_{x\in[a,b]}\prod_{i=0}^{i=n}(x-x_i)^2.$$

Απόδειξη: Αν $x \in \{x_0, x_1, ..., x_n\}$ τότε το αριστερό μέλος της (*) είναι μηδέν λόγω των συνθηκών κατασκευής $p(x_i) = f(x_i), i = 0,1,2,...,n$. Το δεξί μέλος της (*) είναι επίσης μηδέν αφού το γινόμενο $\prod_{i=0}^{i=n} (x - x_i)^2$ μηδενίζεται. Αν $x \in [a,b]$ αλλά $x \notin \{x_0, x_1, ..., x_n\}$ εργαζόμαστε ως εξής. Θεωρούμε, χωρίς βλάβη της γενικότητας ότι τα $x_i, i = 0,1,2,...,n$ είναι διατεταγμένα σε αύξουσα σειρά, δηλαδή $x_0 < x_1 < x_2 < ... < x_n$. Σε αυτήν την περίπτωση μπορούμε να διακρίνουμε 3 περιπτώσεις (α) $x < x_0$, (β) $x \in \left(x_j, x_{j+1}\right), j = 1,2,3,...,n-1$ και (γ) $x > x_n$. Θα μελετήσουμε την (β) περίπτωση, ενώ και οι άλλες δύο αποδεικνύονται με όμοιο τρόπο. Πρώτα θέτουμε την βοηθητική συνάρτηση $\Psi(t) = \prod_{i=0}^{n} \left(t - x_i\right)^2$ για την οποία παρατηρούμε ότι $\Psi(x) = \prod_{i=0}^{n} \left(x - x_i\right)^2 = 0$ αν $x \in \{x_0, x_1, ..., x_n\}$. Επίσης $\Psi'(t) = \frac{d}{dt}\left(\prod_{i=0}^{n} \left(t - x_i\right)^2\right) = \frac{d}{dt}\left(\left(t - x_0\right)^2 \left(t - x_1\right)^2 ....\left(t - x_n\right)^2\right)$ οπότε

$$\Psi'(t) = 2\left(\left(t - x_0\right)\left(t - x_1\right)^2 ....\left(t - x_n\right)^2 + \left(t - x_0\right)^2 \left(t - x_1\right)...\left(t - x_n\right)^2 + ... + \left(t - x_0\right)^2 \left(t - x_1\right)^2 ....\left(t - x_n\right)\right)$$

και επομένως $\Psi'(x) = 0$ αν $x \in \{x_0, x_1, ..., x_n\}$. Παρατηρούμε ότι η πολυωνυμική συνάρτηση Ψ είναι βαθμού 2n+2 και άρα ισχύει $\Psi \in C^{2n+2}[a,b]$. Στην συνέχεια θεωρούμε την συνάρτηση $\varphi(t) = f(t) - p(t) - \frac{f(x) - p(x)}{\Psi(x)}\Psi(t), \quad t \in [a,b]$ και $x \in [a,b], x \notin \{x_0, x_1, ..., x_n\}$, η παράγωγος της οποίας είναι $\varphi'(t) = f'(t) - p'(t) - \frac{f(x) - p(x)}{\Psi(x)}\Psi'(t)$. Εφόσον ισχύει $f, \Psi, p \in C^{2n+2}[a,b]$ άρα και η $\varphi \in C^{2n+2}[a,b]$. Επίσης:

$$\varphi(x_0) = \underbrace{f(x_0) - p(x_0)}_{=0} - \frac{f(x) - p(x)}{\Psi(x)}\Psi(x_0) = 0$$

$$\varphi(x_1) = \underbrace{f(x_1) - p(x_1)}_{=0} - \frac{f(x) - p(x)}{\Psi(x)}\underbrace{\Psi(x_1)}_{=0} = 0$$

....

$$\varphi(x_n) = \underbrace{f(x_n) - p(x_n)}_{=0} - \frac{f(x) - p(x)}{\Psi(x)} \underbrace{\Psi(x_n)}_{=0} = 0$$

και

$$\varphi(x) = f(x) - p(x) - \frac{f(x) - p(x)}{\Psi(x)} \Psi(x) = 0$$

Άρα θα υπάρχουν σημεία $\xi_0, \xi_1, ..., \xi_n$ με

$x_0 < \xi_0 < x_1 < \xi_1 < ... < x_j < \xi_j < x < \xi_{j+1} < x_{j+1} < ... < x_{n-1} < \xi_n < x_n$ τέτοια ώστε

$\varphi'(\xi_i) = 0, \quad i = 0,1,2,..,n$ ("n+1" στο πλήθος ρίζες).

Επιπλέον έχουμε για την παράγωγο της συνάρτησης φ:

$$\varphi'(x_0) = \underbrace{f'(x_0) - p'(x_0)}_{=0} - \frac{f(x) - p(x)}{\Psi(x)} \underbrace{\Psi'(x_0)}_{=0} = 0$$

$$\varphi'(x_1) = \underbrace{f'(x_1) - p'(x_1)}_{=0} - \frac{f(x) - p(x)}{\Psi(x)} \underbrace{\Psi'(x_1)}_{=0} = 0$$

….

$$\varphi'(x_n) = \underbrace{f'(x_n) - p'(x_n)}_{=0} - \frac{f(x) - p(x)}{\Psi(x)} \underbrace{\Psi'(x_n)}_{=0} = 0$$

Άρα ισχύει $\varphi'(x_i) = 0, i = 0,1,2,...,n$ ("n+1" στο πλήθος ρίζες) και επομένως έχουμε συνολικά τουλάχιστον «2n+2» διαφορετικές ρίζες για την συνάρτηση φ' στο διάστημα $[a,b]$. Σύμφωνα με το θεώρημα του Rolle η συνάρτηση φ'' θα έχει τουλάχιστον «2n+1» ρίζες, η φ''' τουλάχιστον «2n» ρίζες, …, και η $\varphi^{(2n+2)}$ τουλάχιστον μία ρίζα $\xi \in (a,b)$, δηλαδή $\varphi^{(2n+2)}(\xi) = 0$. Από τον ορισμό της φ μπορούμε να βρούμε την $\varphi^{(2n+2)}$:

$$\varphi^{(2n+2)}(t) = f^{(2n+2)}(t) - \underbrace{p^{(2n+2)}(t)}_{0} - \frac{f(x) - p(x)}{\Psi(x)} \underbrace{\Psi^{(2n+2)}(t)}_{(2n+2)!} \Rightarrow$$

$$\varphi^{(2n+2)}(t) = f^{(2n+2)}(t) - \frac{f(x) - p(x)}{\Psi(x)} (2n+2)! \overset{t=\xi}{\Rightarrow} 0 = f^{(2n+2)}(\xi) - \frac{f(x) - p(x)}{\Psi(x)} (2n+2)! \Rightarrow$$

$f(x) - p(x) = \frac{\Psi(x)}{(2n+2)!} f^{(2n+2)}(\xi)$ και αντικαθιστώντας την Ψ παίρνουμε την ζητούμενη σχέση. Αν στην συνέχεια πάρουμε την άπειρη νόρμα αυτής της σχέσης και εφαρμόσουμε την τριγωνική ανισότητα καταλήγουμε στην δεύτερη ζητούμενη σχέση.

Παράδειγμα 1ο: Έστω η συνάρτηση $f(x)=1/x$. Χρησιμοποιείστε τις τιμές της συνάρτησης και της παραγώγου της στα σημεία $x_0=1$, $x_1=2$ για να προσδιορίσετε το πολυώνυμο παρεμβολής Hermite. Στην συνέχεια βρείτε το σφάλμα στο x=1.5.

Λύση: Έχουμε $f(x_0)=1/x_0=1$ και $f(x_1)=1/x_1=1/2$. Επιπλέον, $f'(x)=-1/x^2$ και επομένως $f'(x_0)=-1/x_0^2=-1$ και $f'(x_1)=-1/x_1^2=-1/4$. Άρα το πολυώνυμο Hermite θα δίνεται από την σχέση:

$$p(x)=\sum_{i=0}^{n}\left\{f(x_i)H_i^{(1)}(x)+f'(x_i)H_i^{(2)}(x)\right\}\Rightarrow$$

$$p(x)=f(x_0)H_0^{(1)}(x)+f(x_1)H_1^{(1)}(x)+f'(x_0)H_0^{(2)}(x)+f'(x_1)H_1^{(2)}(x)$$

και αντικαθιστώντας τις αριθμητικές τιμές που βρήκαμε παραπάνω παίρνουμε:

$p(x)=H_0^{(1)}(x)+\frac{1}{2}H_1^{(1)}(x)-H_0^{(2)}(x)-\frac{1}{4}H_1^{(2)}(x)$. Απομένει ο προσδιορισμός των πολυωνύμων $H_0^{(1)},H_1^{(1)},H_0^{(2)},H_1^{(2)}$, η βάση των οποίων είναι τα κατάλληλα πολυώνυμα Lagrange της f στα σημεία x_0,x_1. Έχουμε $L_i(x)=\prod_{\substack{j=0\\ i\neq j}}^{n}\frac{x-x_j}{x_i-x_j}$ με n=1 και άρα

$L_0(x)=\prod_{\substack{j=0\\ j\neq 0}}^{1}\frac{x-x_j}{x_0-x_j}=\frac{x-x_1}{x_0-x_1}=\frac{x-1}{1-2}=1-x$ και $L_0'(x)=-1$ ενώ

$L_1(x)=\prod_{\substack{j=0\\ j\neq 1}}^{1}\frac{x-x_j}{x_1-x_j}=\frac{x-x_0}{x_1-x_0}=\frac{x-2}{2-1}=x-2$ και $L_1'(x)=1$. Άρα

$$H_0^{(1)}(x)=\left[1-2(x-x_0)L_0'(x_0)\right]L_0^2(x)=\left[1-2(x-1)(-1)\right](1-x)^2=(2x+3)(1-x)^2$$

$$H_1^{(1)}(x)=\left[1-2(x-x_1)L_1'(x_1)\right]L_1^2(x)=\left[1-2(x-2)(1)\right](x-2)^2=(3-2x)(x-2)^2$$

$$H_0^{(2)}(x)=(x-x_0)L_0^2(x)=(x-1)(2-x)^2$$

$$H_1^{(2)}(x)=(x-x_1)L_1^2(x)=(x-2)(x-1)^2$$

Αντικαθιστώντας τις παραπάνω εκφράσεις στο πολυώνυμο και απλοποιώντας παίρνουμε την τελική έκφραση $p(x)=-\frac{1}{4}x^3+\frac{3}{2}x^2-\frac{13}{4}x+3$. Για x=3/2 παίρνουμε

$p\left(\frac{3}{2}\right)=-\frac{1}{4}\left(\frac{3}{2}\right)^3+\frac{3}{2}\left(\frac{3}{2}\right)^2-\frac{13}{4}\left(\frac{3}{2}\right)+3=$ και άρα το σφάλμα είναι $f\left(\frac{3}{2}\right)-p\left(\frac{3}{2}\right)=$. Αν

φυσικά δεν γνωρίζαμε την ακριβή τιμή της συνάρτησης στο ζητούμενο σημείο θα είχαμε:

$\|f-p\|_\infty \le \frac{\|f^{(4)}\|_\infty}{(2n+2)!}\max_{x\in[a,b]}\left\{\prod_{i=0}^{n}(x-x_i)^2\right\}$ και αντικαθιστώντας τα αριθμητικά δεδομένα μας θα είχαμε $\|f(x)-p(x)\|_\infty \le \frac{\|f^{(4)}\|_\infty}{4!}\max_{x\in[1,2]}\left\{(x-1)^2(x-2)^2\right\}$.

Παράδειγμα 2ο: Έστω η συνάρτηση $f(x)=1/(x^2+2)$. Χρησιμοποιείστε τις τιμές της συνάρτησης και της παραγώγου της στα σημεία $x_0=-1$, $x_1=1$ για να προσδιορίσετε το πολυώνυμο παρεμβολής Hermite. Στην συνέχεια βρείτε το σφάλμα στο x=0.

Λύση: Έχουμε n=1 οπότε η εφαρμογή της σχέσης που δίνει το πολυώνυμο Hermite έχει ως εξής

$$p(x)=f(x_0)H_0^{(1)}(x)+f(x_1)H_1^{(1)}(x)+f'(x_0)H_0^{(2)}(x)+f'(x_1)H_1^{(2)}(x) \quad (*)$$

Εφόσον $f'(x)=-2x/(x^2+2)^2$ έχουμε, $f(x_0)=f(-1)=1/3$, $f(x_1)=f(1)=1/3$, $f'(x_0)=f'(-1)=2/9$, $f'(x_1)=f'(1)=-2/9$. Για τα πολυώνυμα Lagrange έχουμε

$L_0(x)=\prod_{\substack{j=0\\ j\neq 0}}^{1}\frac{x-x_j}{x_0-x_j}=\frac{x-x_1}{x_0-x_1}=\frac{x-1}{(-1)-1}=\frac{1-x}{2}$ και $L_0'(x)=-\frac{1}{2}$. Επίσης,

$L_1(x)=\prod_{\substack{j=0\\ j\neq 1}}^{1}\frac{x-x_j}{x_1-x_j}=\frac{x-x_0}{x_1-x_0}=\frac{x-(-1)}{1-(-1)}=\frac{x+1}{2}$ και $L_1'(x)=\frac{1}{2}$. Άρα

$$H_0^{(1)}(x)=\left[1-2(x-x_0)L_0'(x_0)\right]L_0^2(x)=\left[1-2(x+1)\left(-\frac{1}{2}\right)\right]\left(\frac{1-x}{2}\right)^2=(x+2)\frac{(x-1)^2}{4}$$

$$H_1^{(1)}(x)=\left[1-2(x-x_1)L_1'(x_1)\right]L_1^2(x)=\left[1-2(x-1)\left(\frac{1}{2}\right)\right]\left(\frac{x+1}{2}\right)^2=(3-2x)\frac{(x+1)^2}{4}$$

$$H_0^{(2)}(x)=(x-x_0)L_0^2(x)=(x+1)\frac{(x-1)^2}{4}$$

$$H_1^{(2)}(x)=(x-x_1)L_1^2(x)=(x-1)\frac{(x+1)^2}{4}$$

Αντικαθιστώντας όλα τα παραπάνω στην (*) και απλοποιώντας παίρνουμε $p(x)=-\frac{1}{9}x^2+\frac{4}{9}=\frac{4-x^2}{9}$. Για $x=0$ έχουμε $p(0)=\frac{4}{9}$ και το σφάλμα είναι $f(0)-p(0)=\frac{1}{2}-\frac{4}{9}=\frac{1}{18}$.

4.7. Παρεμβολή με κυβικές splines

Όπως έχει αναφερθεί η διαδικασία της παρεμβολής με υψηλού βαθμού πολυώνυμα αποτυγχάνει (τόσο θεωρητικά όσο και υπολογιστικά) καθώς ο βαθμός του πολυωνύμου αυξάνει. Ωστόσο θα πρέπει να σημειωθεί ότι για μικρού βαθμού πολυωνύμο (μέχρι 3 ή 4 το πολύ) και τοπικά η διαδικασία αυτή έχει αρκετά καλές ιδιότητες. Ας δούμε αρχικά τι θα συμβεί στις εξής περιπτώσεις:

(Α) όταν ένα πολυώνυμο πρώτου βαθμού χρησιμοποιηθεί για να παρεμβάλει μία συνάρτηση $f\in C^2[a,b]$ στα σημεία $x_0=a$, $x_1=b$.

(Β) όταν ένα πολυώνυμο πρώτου βαθμού χρησιμοποιηθεί για να παρεμβάλει μία συνάρτηση $f\in C^3[a,b]$ στα σημεία $x_0=a$, $x_1=\frac{a+b}{2}$, $x_2=b$.

(Γ) όταν ένα πολυώνυμο τρίτου βαθμού χρησιμοποιηθεί για να παρεμβάλει μία συνάρτηση $f\in C^3[a,b]$ στα σημεία $x_0=a$, $x_1=a+\frac{1}{3}(b-a)$, $x_2=a+\frac{2}{3}(b-a)$.

Ας δούμε τις περιπτώσεις αυτές

(Α) Στην περίπτωση αυτή το πολυώνυμο παρεμβολής είναι το εξής:

$$p(x)=f(a)+\frac{f(b)-f(a)}{b-a}(x-a)$$

Για το σφάλμα θα ισχύει:

$$\|f-p\|_\infty\le\frac{\|f''\|_\infty}{2}\max_{x\in[a,b]}\left|\prod_{i=0}^{1}(x-x_i)\right|=\frac{\|f''\|_\infty}{2}\max_{x\in[a,b]}\left|(x-a)(x-b)\right|$$

Για να βρούμε μία καλύτερη έκφραση για το άνω φράγμα της παραπάνω ανισότητας, ορίζουμε την συνάρτηση $g(x)=\left|(x-a)(x-b)\right|=(x-a)(b-x)$, όπου η δεύτερη ισότητα ισχύει εφόσον $x\in[a,b]$, και ζητούμε να βρούμε σε ποιο σημείο εμφανίζει μέγιστο. Για τον σκοπό αυτό πρέπει να βρούμε το σημείο μηδενισμού της πρώτης παραγώγου της, δηλαδή της $g'(x)=-(x-a)+(b-x)=-x+a+b-x=-2x+(a+b)$. Προφανώς έχουμε

$g'(y)=0\Rightarrow -2y+(a+b)=0\Rightarrow y=\dfrac{a+b}{2}$. Για την δεύτερη παράγωγο έχουμε $g''(x)=-2<0$ οπότε στο $x=y$ η συνάρτηση g εμφανίζει μέγιστο, το οποίο είναι $g(y)=(y-a)(b-y)=\left(\dfrac{a+b}{2}-a\right)\left(b-\dfrac{a+b}{2}\right)=\dfrac{(b-a)^2}{4}$. Επομένως έχουμε $\max\limits_{x\in[a,b]}\left|(x-a)(x-b)\right|=\max\limits_{x\in[a,b]}g(x)=\dfrac{(b-a)^2}{4}$ και η ανισότητα διαμορφώνεται ως εξής:

$$\|f-p_1\|_\infty\le\frac{\|f''\|_\infty}{8}(b-a)^2$$

Η τελευταία αυτή σχέση δείχνει πως όσο πιο μικρό είναι το μήκος του διαστήματος [a,b] τόσο μικρότερο είναι το άνω φράγμα του σφάλματος της παρεμβολής.

(Β) Στην περίπτωση αυτή θα έχουμε

$$p_2(x)=f(a)+2\frac{f\left(\frac{a+b}{2}\right)-f(a)}{b-a}(x-a)+2\frac{f(a)-2f\left(\frac{a+b}{2}\right)+f(b)}{(b-a)^2}(x-a)\left(x-\frac{a+b}{2}\right)$$

(ελέγξτε ότι ισχύει $p_2(a)=f(a),\quad p_2\left(\dfrac{a+b}{2}\right)=f\left(\dfrac{a+b}{2}\right),\quad p_2(b)=f(b)$). Για το σφάλμα θα έχουμε:

$$\|f-p_2\|_\infty\le\frac{\|f'''\|_\infty}{6}\max_{x\in[a,b]}\left|\prod_{i=0}^{2}(x-x_i)\right|=\frac{\|f''\|_\infty}{6}\max_{x\in[a,b]}\left|(x-a)\left(x-\frac{a+b}{2}\right)(x-b)\right|$$

Ορίζουμε την συνάρτηση

$$g(x)=\left|(x-a)\left(x-\frac{a+b}{2}\right)(x-b)\right|=\begin{cases}(x-a)\left(x-\dfrac{a+b}{2}\right)(x-b),\ x\in[a,(a+b)/2]\\(x-a)\left(x-\dfrac{a+b}{2}\right)(b-x),\ x\in[(a+b)/2,b]\end{cases}$$

της οποίας ζητούμε το μέγιστο. Θα έχουμε:

$$g'(x)=\begin{cases}\left(\dfrac{a+b}{2}-x\right)(b-x)-(x-a)(b-x)-(x-a)\left(\dfrac{a+b}{2}-x\right),\ x\in[a,(a+b)/2]\\\left(x-\dfrac{a+b}{2}\right)(b-x)+(x-a)(b-x)-(x-a)\left(x-\dfrac{a+b}{2}\right),\ x\in[(a+b)/2,b]\end{cases}$$

είτε απλοποιώντας την παραπάνω έκφραση:

$$g'(x)=\begin{cases}\left(x-\dfrac{a+b}{4}\right)\left(x+3\dfrac{a+b}{4}\right), x\in[a,(a+b)/2]\\ \left(x-\dfrac{a+b}{4}\right)\left(x+3\dfrac{a+b}{4}\right), x\in[(a+b)/2,b]\end{cases}$$

κτλ κτλ κτλ

(Γ) Η περίπτωση αυτή αφήνεται ως άσκηση. Πιο συγκεκριμένα, με όμοιο τρόπο να δείξετε ότι:

$$\|f-p_3\|_\infty \le \frac{(b-a)^4}{1296}\|f^{(4)}\|_\infty$$

Όλα τα παραπάνω δείχνουν ότι αν το μήκος του διαστήματος, $b-a$, είναι αρκετά μικρό και με δεδομένο ότι η συνάρτηση f συμπεριφέρεται ομαλά η πολυωνυμική παρεμβολή δουλεύει καλά. Στην περίπτωση που το διάστημα ενδιαφέροντος [a,b] είναι μεγάλο τότε ωθούμαστε να διερευνήσουμε τι συμβαίνει αν θεωρήσουμε τμηματικά ορισμένες συναρτήσεις. Πιο συγκεκριμένα, τι ακριβώς θα συμβεί να θεωρήσουμε ένα αρκετά λεπτό διαμερισμό του [a,b] και στο κάθε υποδιάστημα ορίσουμε χαμηλού βαθμού πολυώνυμα τάξης 2, 3 ή 4 το πολύ?

Έστω ένα διάστημα $[a,b]\subset\mathbb{R}$ και Δ ένας διαμερισμός του, $\Delta: a\equiv x_0<x_1<x_2<\ldots<x_{n-1}<x_n\equiv b$ καθώς και ένας φυσικός αριθμός $m\in\mathbb{N}$. Τα στοιχεία του γραμμικού χώρου $S_m(\Delta)=\left\{s\in C^{m-1}[a,b]//\,s|_{[x_{i-1},x_i]}\in P_m\right\}$ ονομάζονται splines βαθμού «m» ως προς Δ. Πρόκειται δηλαδή για πολυώνυμα βαθμού το πολύ «m», το καθένα από τα οποία είναι ορισμένα στο υποδιάστημα $[x_{i-1},x_i]$. Να σημειωθεί ότι στις εφαρμογές χρησιμοποιούνται συνήθως οι $S_1(\Delta)$ και $S_3(\Delta)$. Εδώ θα μελετήσουμε μόνο τις $S_3(\Delta)$.

Έστω λοιπόν ο γραμμικός χώρος $S_3(\Delta)=\left\{s\in C^2[a,b]//\,s|_{[x_{i-1},x_i]}\in P_3\right\}$ τα στοιχεία του οποίου είναι 2 φορές συνεχώς παραγωγίσιμα στο κλειστό διάστημα [a,b], είναι πολυώνυμα τρίτου βαθμού το πολύ και γενικά είναι τμηματικά ορισμένα στα υποδιαστήματα $[x_{i-1},x_i], i=1,2,..,n$. Θα ξεκινήσουμε με την κατασκευή των στοιχείων αυτού του χώρου συμβολίζοντας με s_i κάθε τμήμα της συνάρτησης ορισμένης στο $[x_{i-1},x_i]$, δηλαδή

$$s=\begin{cases} s_1(x), & x_0 \le x \le x_1 \\ s_2(x), & x_1 \le x \le x_2 \\ \dots & \\ s_i(x), & x_{i-1} \le x \le x_i \\ \dots & \\ s_n(x), & x_{n-1} \le x \le x_n \end{cases}.$$

Για την κατασκευή της συνάρτησης s θα χρησιμοποιήσουμε όλα τα δεδομένα συν τις ιδιότητες που πρέπει να ικανοποιεί η s. Από τα δεδομένα θα έχουμε:

$s(x_i)=f(x_i), \quad i=0,1,2,\dots,n$ (1)

Εφόσον $s\in C^2[a,b]$ άρα τόσο η συνάρτηση όσο και η πρώτη και δεύτερη παράγωγός της θα πρέπει να είναι συνεχής σε όλα τα εσωτερικά σημεία του διαστήματος [a,b], δηλαδή:

$s_i(x_{i+1})=s_{i+1}(x_i), \quad i=1,2,\dots,n-1$ (2)

$s_i'(x_i)=s_{i+1}'(x_i), \quad i=1,2,\dots,n-1$ (3)

$s_i''(x_i)=s_{i+1}''(x_i), \quad i=1,2,\dots,n-1$ (4)

Συνολικά λοιπόν έχουμε n+1 στο πλήθος συνθήκες από την (1), και 3*(n-1) συνθήκες από τις (2), (3) και (4), δηλαδή σύνολο 4n-2 συνθήκες. Επειδή η κάθε συνάρτηση s_i είναι το πολύ τρίτου βαθμού, μπορούμε να γράψουμε $s_i(x)=a_i+\beta_i x+c_i x^2+d_i x^3, \quad x_{i-1}\le x\le x_i$ άρα θα έχουμε 4 αγνώστους για κάθε κλάδο της συνάρτησης και επομένως 4*n αγνώστους. Απαιτούνται λοιπόν 2 ακόμα συνθήκες έτσι ώστε ο αριθμός των αγνώστων και των συνθηκών να είναι ίσος. Οι δύο αυτές επιπλέον συνθήκες ονομάζονται συνοριακές συνθήκες και συνήθως χρησιμοποιούνται οι εξής επιλογές: (α) $s'(x_0)=f'(a), \quad s'(x_n)=f'(b)$, (β) $s''(x_0)=f''(a), \quad s''(x_n)=f''(b)$, (γ) $s''(x_0)=s''(x_n)=0$. Αν επιλέξουμε την τρίτη περίπτωση τότε προκύπτουν οι λεγόμενες «φυσικές κυβικές splines» και οι οποίες είναι οι ομαλότερες συναρτήσεις οι οποίες να παρεμβάλλονται στα δεδομένα μίας συνάρτησης. Στην συνέχεια δίνουμε το παρακάτω θεώρημα.

Θεώρημα: Έστω $f\in C^1[a,b], \quad n\in\mathbb{N}$ και $\Delta: a\equiv x_0<x_1<x_2<\dots<x_{n-1}<x_n\equiv b$ ένας διαμερισμός του $[a,b]$. Τότε υπάρχει ακριβώς μία συνάρτηση $s\in S_3(\Delta)$ τέτοια ώστε

$\{s(x_i)=f(x_i), s'(a)=f'(a), s'(b)=f'(b), \quad i=0,1,2,\dots,n\}$

Απόδειξη: Εφόσον η s είναι κυβικό πολυώνυμο άρα η δεύτερη παράγωγός του θα είναι ένα γραμμικό πολυώνυμο, έστω $s''(x) = c_0 + c_1 x$. Όμως για $x = x_{i-1}$ θα πρέπει $s''(x_{i-1}) = c_0 + c_1 x_{i-1}$ και για $x = x_i$ θα πρέπει $s''(x_i) = c_0 + c_1 x_i$. Από τις δύο αυτές τελευταίες σχέσεις μπορούμε να απαλείψουμε τις σταθερές c_0, c_1 για να πάρουμε:

$$s''(x) = \frac{1}{h_i}\left(s_i''(x_i)\left(x - x_{i-1}\right) - s_i''(x_{i-1})\left(x - x_i\right)\right), \quad x_{i-1} \le x \le x_i, \quad h_i \equiv x_i - x_{i-1}.$$

Πράγματι, βλέπουμε ότι για $x = x_{i-1}$ έχουμε

$$s''(x_{i-1}) = \frac{1}{h_i}\left(s_i''(x_i)\left(x_{i-1} - x_{i-1}\right) - s_i''(x_{i-1})\underbrace{\left(x_{i-1} - x_i\right)}_{-h_i}\right) = s''(x_{i-1})$$

και για $x = x_i$ έχουμε $s''(x_i) = \frac{1}{h_i}\left(s''(x_i)\underbrace{\left(x_i - x_{i-1}\right)}_{h_i} - s''(x_{i-1})\left(x_i - x_i\right)\right) = s''(x_i)$. Με τον τρόπο αυτό έχουμε ικανοποιήσει την συνέχεια της δεύτερη παραγώγου της ζητούμενης συνάρτησης spline, δηλαδή την συνθήκη (4). Πράγματι έχουμε:

$$s''(x) = \frac{1}{h_i}\left(s''(x_i)\left(x - x_{i-1}\right) - s''(x_{i-1})\left(x - x_i\right)\right) \equiv s_i''(x), \quad x_{i-1} \le x \le x_i, \quad h_i \equiv x_i - x_{i-1} \quad (5)$$

$$s''(x) = \frac{1}{h_{i+1}}\left(s''(x_{i+1})\left(x - x_i\right) - s''(x_i)\left(x - x_{i+1}\right)\right) \equiv s_{i+1}''(x), \quad x_i \le x \le x_{i+1}, \quad h_{i+1} \equiv x_{i+1} - x_i \quad (6)$$

Για $x = x_i$ από την (5) έχουμε $s''(x_i) = s''(x_i) \equiv s_i''(x_i)$

Για $x = x_i$ από την (6) έχουμε $s''(x_i) = s''(x_i) \equiv s_{i+1}''(x_i)$

Επομένως από τις δύο τελευταίες σχέσεις βλέπουμε ότι $s_i''(x_i) = s_{i+1}''(x_i)$.

Την παραγόμενη μορφή μπορούμε να την ολοκληρώσουμε δύο φορές και να πάρουμε:

$$s(x) = \frac{1}{6h_i}\left(s_i''(x_i)\left(x - x_{i-1}\right)^3 - s_i''(x_{i-1})\left(x - x_i\right)^3\right) + \hat{c}_2 x + \hat{c}_3$$

Στο σημείο αυτό μπορούμε να μειώσουμε τις απαιτούμενες πράξεις αν αντί για την προηγούμενη μορφή εκφράσουμε διαφορετικά το γραμμικό κομμάτι της s:

$$s(x) = \frac{1}{6h_i}\left(s_i''(x_i)\left(x - x_{i-1}\right)^3 - s_i''(x_{i-1})\left(x - x_i\right)^3\right) + c_2\left(x - x_{i-1}\right) + c_3\left(x - x_i\right)$$

Στην συνέχεια ικανοποιούμε τα δεδομένα μας, δηλαδή τις συνθήκες (1) και έχουμε:

$$s(x_i)=\frac{1}{6h_i}\left(s''(x_i)\underbrace{\left(x_i-x_{i-1}\right)}_{h_i}^3-s''(x_{i-1})\left(x_i-x_i\right)^3\right)+c_2\underbrace{\left(x_i-x_{i-1}\right)}_{h_i}+c_3\left(x_i-x_i\right)=f\left(x_i\right)\Rightarrow$$

$$\frac{s''(x_i)h_i^3}{6h_i}+c_2h_i=f\left(x_i\right)\Rightarrow c_2=-\frac{s''(x_i)h_i}{6}+\frac{f\left(x_i\right)}{h_i}$$

και

$$s(x_{i-1})=\frac{1}{6h_i}\left(s''(x_i)\left(x_{i-1}-x_{i-1}\right)^3-s''(x_{i-1})\underbrace{\left(x_{i-1}-x_i\right)}_{-h_i}^3\right)+c_2\left(x_{i-1}-x_{i-1}\right)+c_3\underbrace{\left(x_{i-1}-x_i\right)}_{-h_i}\Rightarrow$$

$$\frac{s''(x_{i-1})h_i^3}{6h_i}-c_3h_i=f\left(x_{i-1}\right)\Rightarrow c_3=\frac{s''(x_{i-1})h_i}{6}-\frac{f\left(x_{i-1}\right)}{h_i}$$

επομένως έχουμε

$$s(x)=\frac{1}{6h_i}\left(s''(x_i)\left(x-x_{i-1}\right)^3-s''(x_{i-1})\left(x-x_i\right)^3\right)+\left(\frac{f\left(x_i\right)}{h_i}-\frac{s''(x_i)h_i}{6}\right)\left(x-x_{i-1}\right)+$$
$$-\left(\frac{f\left(x_{i-1}\right)}{h_i}-\frac{s''(x_{i-1})h_i}{6}\right)\left(x-x_i\right),\quad x_{i-1}\le x\le x_i \qquad \text{(A)}$$

Τέλος απομένει να ικανοποιήσουμε την συνέχεια της $s'(x)$ στους εσωτερικούς κόμβους, δηλαδή την συνθήκη (3), $s_i'\left(x_i\right)=s_{i+1}'\left(x_i\right),\quad i=1,2,...,n-1$. Υπολογίζουμε πρώτα την παράγωγο της τελευταίας σχέσης:

$$s'(x)=\frac{1}{2h_i}\left(s''(x_i)\left(x-x_{i-1}\right)^2-s''(x_{i-1})\left(x-x_i\right)^2\right)+\left(\frac{f\left(x_i\right)}{h_i}-\frac{h_is''(x_i)}{6}\right)-\left(\frac{f\left(x_{i-1}\right)}{h_i}-\frac{h_is''(x_{i-1})}{6}\right)$$

η οποία ισχύει για $x_{i-1}\le x\le x_i$ (7)

$$s'(x)=\frac{1}{2h_{i+1}}\left(s''(x_{i+1})\left(x-x_i\right)^2-s''(x_i)\left(x-x_{i+1}\right)^2\right)+\left(\frac{f\left(x_{i+1}\right)}{h_{i+1}}-\frac{h_{i+1}s''(x_{i+1})}{6}\right)-\left(\frac{f\left(x_i\right)}{h_{i+1}}-\frac{h_{i+1}s''(x_i)}{6}\right)$$

η οποία ισχύει για $x_i\le x\le x_{i+1}$ (8)

Για $x=x_i$ από την (7) έχουμε

$$s'(x_i)=\frac{1}{2h_i}\left(s''(x_i)\left(x_i-x_{i-1}\right)^2-s''(x_{i-1})\left(x_i-x_i\right)^2\right)+\left(\frac{f\left(x_i\right)}{h_i}-\frac{h_is''(x_i)}{6}\right)-\left(\frac{f\left(x_{i-1}\right)}{h_i}-\frac{h_is''(x_{i-1})}{6}\right)\Rightarrow$$

$$s'(x_i)=\frac{h_is''(x_i)}{3}+\frac{f\left(x_i\right)-f\left(x_{i-1}\right)}{h_i}+\frac{h_is''(x_{i-1})}{6}\equiv s_i'(x_i),\ x_{i-1}\le x\le x_i \quad (9)$$

Για $x=x_i$ από την (8) έχουμε

$$s'(x_i)=\frac{1}{2h_{i+1}}\left(s''(x_{i+1})(x_i-x_i)^2-s''(x_i)(x_i-x_{i+1})^2\right)+\left(\frac{f(x_{i+1})}{h_{i+1}}-\frac{h_{i+1}s''(x_{i+1})}{6}\right)-\left(\frac{f(x_i)}{h_{i+1}}-\frac{h_{i+1}s''(x_i)}{6}\right)\Rightarrow$$

$$s'(x_i)=-\frac{h_{i+1}s''(x_i)}{3}+\frac{f(x_{i+1})-f(x_i)}{h_{i+1}}-\frac{h_{i+1}s''(x_{i+1})}{6}\equiv s_{i+1}{}'(x_i),\quad x_i\le x\le x_{i+1}\ (10)$$

Από τις δύο τελευταίες σχέσεις και την συνθήκη $s_i''(x_i)=s_{i+1}''(x_i)$ παίρνουμε:

$$\frac{h_is''(x_i)}{3}+\frac{f(x_i)-f(x_{i-1})}{h_i}+\frac{h_is''(x_{i-1})}{6}=-\frac{h_{i+1}s''(x_i)}{3}+\frac{f(x_{i+1})-f(x_i)}{h_{i+1}}-\frac{h_{i+1}s''(x_{i+1})}{6}$$

είτε αναδιαμορφώνοντάς την, παίρνουμε:

$$h_is''(x_{i-1})+2(h_i+h_{i+1})s''(x_i)+h_{i+1}s''(x_{i+1})=6\left(\frac{f(x_{i+1})-f(x_i)}{h_{i+1}}-\frac{f(x_i)-f(x_{i-1})}{h_i}\right)\ (A1)$$

$$h_i=x_i-x_{i-1},\quad h_{i+1}=x_{i+1}-x_i,\quad i=1,2,...,n-1$$

Στην περίπτωση που ο διαμερισμός είναι ομοιόμορφος, αν δηλαδή $h_i=h_{i+1}=h=\frac{b-a}{n}, i=1,2,...,n-1$ τότε η παραπάνω σχέση απλοποιείται ως εξής:

$$s''(x_{i-1})+4s''(x_i)+s''(x_{i+1})=6\frac{f(x_{i+1})-2f(x_i)+f(x_{i-1})}{h^2},\quad h=\frac{b-a}{n}, i=1,2,...,n-1\ (A2)$$

Παρατηρούμε ότι έχουμε n-1 εξισώσεις με n+1 αγνώστους, τους $s''(x_i), i=0,1,2,...,n$. Όπως ήδη έχει αναφερθεί οι δύο εξισώσεις που λείπουν είναι οι συνοριακές συνθήκες οι οποίες στην παρούσα περίπτωση είναι οι $s'(a)=f'(a)$ και $s'(b)=f'(b)$. Η σχέση (7) για $i=1,\ x=x_0=a$ δίνει:

$$s'(a)=-\frac{h_1s''(x_0)}{3}+\frac{f(x_1)-f(x_0)}{h_1}+\frac{h_1s''(x_0)}{6}=f'(a)\Rightarrow$$

$$2s''(x_0)+s''(x_1)=\frac{6}{h_1}\left(\frac{f(x_1)-f(x_0)}{h_1}-f'(a)\right)\ (11)$$

Η σχέση (8) για $i=n-1,\ x=x_n=b$ δίνει

$$s'(b)=\frac{h_ns''(x_n)}{3}+\frac{f(x_n)-f(x_{n-1})}{h_n}+\frac{h_ns''(x_{n-1})}{6}=f'(b)\Rightarrow$$

$$s''(x_{n-1})+2s''(x_n)=\frac{6}{h_n}\left(f'(b)-\frac{f(x_n)-f(x_{n-1})}{h_n}\right)\ (12)$$

Οι σχέσεις (A1) μαζί με τις (11) και (12) είτε στην περίπτωση του ομοιόμορφου διαμερισμού οι σχέσεις (A2) μαζί με τις (11) και (12) αποτελούν ένα γραμμικό σύστημα

n+1 εξισώσεων με n+1 αγνώστους. Συγκεντρωτικά τα δύο αυτά συστήματα έχουν ως εξής:

$$\begin{aligned}
&h_i s''(x_{i-1})+2(h_i+h_{i+1})s''(x_i)+h_{i+1}s''(x_{i+1})=6\left(\frac{f(x_{i+1})-f(x_i)}{h_{i+1}}-\frac{f(x_i)-f(x_{i-1})}{h_i}\right)\\
&h_i=x_i-x_{i-1},\quad h_{i+1}=x_{i+1}-x_i,\quad i=1,2,\ldots,n-1\\
&2s''(x_0)+s''(x_1)=\frac{6}{h_1}\left(\frac{f(x_1)-f(x_0)}{h_1}-f'(a)\right)\\
&s''(x_{n-1})+2s''(x_n)=\frac{6}{h_n}\left(f'(b)-\frac{f(x_n)-f(x_{n-1})}{h_n}\right)
\end{aligned}\qquad (\Sigma 1)$$

ή για ομοιόμορφο διαμερισμό

$$\begin{aligned}
&s''(x_{i-1})+4s''(x_i)+s''(x_{i+1})=6\frac{f(x_{i+1})-2f(x_i)+f(x_{i-1})}{h^2},\quad h=\frac{b-a}{n}, i=1,2,\ldots,n-1\\
&2s''(x_0)+s''(x_1)=\frac{6}{h}\left(\frac{f(x_1)-f(x_0)}{h}-f'(a)\right)\\
&s''(x_{n-1})+2s''(x_n)=\frac{6}{h}\left(f'(b)-\frac{f(x_n)-f(x_{n-1})}{h}\right)
\end{aligned}\qquad (\Sigma 2)$$

Απομένει να αποδείξουμε ότι τα συστήματα (Σ1), (Σ2) έχουν μοναδική λύση. Αυτό μπορούμε να το δούμε αν γράψουμε τα συστήματα αυτά σε μορφή διανυσμάτων και πινάκων ως εξής:

$$\underbrace{\begin{bmatrix}
2 & 1 & 0 & 0 & 0 & \ldots. & 0 & 0 & 0\\
h_1 & 2(h_1+h_2) & h_2 & 0 & 0 & \ldots. & 0 & 0 & 0\\
0 & h_2 & 2(h_2+h_3) & h_3 & 0 & \ldots. & 0 & 0 & 0\\
\ldots\ldots & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & \ldots. & h_{n-1} & 2(h_{n-1}+h_n) & h_n\\
0 & 0 & 0 & 0 & 0 & \ldots. & 0 & 1 & 2
\end{bmatrix}}_{\underline{\underline{J}}}\cdot
\underbrace{\begin{bmatrix} s''(x_0)\\ s''(x_1)\\ s''(x_2)\\ \ldots\\ s''(x_{n-1})\\ s''(x_n)\end{bmatrix}}_{\underline{a}}=
\underbrace{\begin{bmatrix} y_0\\ y_1\\ y_2\\ \ldots\\ y_{n-1}\\ y_n\end{bmatrix}}_{\underline{y}}\qquad (\Sigma 1\beta)$$

$$\underbrace{\begin{bmatrix}
2 & 1 & 0 & 0 & 0 & \ldots. & 0 & 0 & 0\\
1 & 4 & 1 & 0 & 0 & \ldots. & 0 & 0 & 0\\
0 & 1 & 4 & 1 & 0 & \ldots. & 0 & 0 & 0\\
\ldots\ldots & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & \ldots. & 1 & 4 & 1\\
0 & 0 & 0 & 0 & 0 & \ldots. & 0 & 1 & 2
\end{bmatrix}}_{\underline{\underline{\hat{J}}}}\cdot
\underbrace{\begin{bmatrix} s''(x_0)\\ s''(x_1)\\ s''(x_2)\\ \ldots\\ s''(x_{n-1})\\ s''(x_n)\end{bmatrix}}_{\underline{a}}=
\underbrace{\begin{bmatrix} \hat{y}_0\\ \hat{y}_1\\ \hat{y}_2\\ \ldots\\ \hat{y}_{n-1}\\ \hat{y}_n\end{bmatrix}}_{\underline{\hat{y}}}\qquad (\Sigma 1\beta)$$

Παρατηρούμε ότι τόσο το σύστημα (Σ1α), $\underline{\underline{J}}\cdot\underline{a}=\underline{y}$, όσο και το (Σ1β), $\underline{\underline{\hat{J}}}\cdot\underline{a}=\underline{\hat{y}}$, έχουν πίνακες, $\underline{\underline{J}}$ και $\underline{\underline{\hat{J}}}$ αντίστοιχα, ο καθένας από τους οποίους είναι αυστηρά διαγωνίως

υπερτερών. Όπως όμως έχει δειχτεί, κάθε αυστηρά διαγωνίως υπερτερών πίνακας είναι αντιστρέψιμος και επομένως τα (Σ1α) και (Σ1β) έχουν μοναδική λύση, $\underline{a} = \underline{\underline{J}}^{-1} \cdot \underline{y}$ και $\underline{a} = \underline{\underline{\hat{J}}}^{-1} \cdot \underline{\hat{y}}$, αντίστοιχα, γεγονός που ολοκληρώνει την απόδειξη.

Τέλος να τονισθεί ότι λόγω επίσης του γεγονότος ότι οι $\underline{\underline{J}}$ και $\underline{\underline{\hat{J}}}$ είναι τριδιαγώνιοι, τα συστήματα (Σ1α) και (Σ1β) μπορούν να επιλυθούν αποδοτικά με τον αλγόριθμο που αναπτύχθηκε για τα τριδιαγώνια συστήματα στο κεφάλαιο 3 (απαιτούνται προσεγγιστικά πράξεις 4(n+1)).

Προχωράμε τώρα στο επόμενο θεώρημα:

Θεώρημα: Έστω $f \in C^1[a,b]$, $n \in \mathbb{N}$ και $\Delta : a \equiv x_0 < x_1 < x_2 < \ldots < x_{n-1} < x_n \equiv b$ ένας διαμερισμός του $[a,b]$. Τότε υπάρχει ακριβώς μία συνάρτηση $s \in S_3(\Delta)$ τέτοια ώστε

$$\left\{ s(x_i) = f(x_i), s''(a) = f''(a), s''(b) = f''(b), \quad i = 0,1,2,\ldots,n \right\}$$

Απόδειξη: Η απόδειξη είναι ακριβώς ίδια με του προηγούμενου θεωρήματος μόνο που αλλάζουν οι συνοριακές συνθήκες. Έτσι τα συστήματα που προκύπτουν είναι

$$\underbrace{\begin{bmatrix} 2 & 1 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ h_1 & 2(h_1+h_2) & h_2 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & h_2 & 2(h_2+h_3) & h_3 & 0 & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & 0 & \ldots & h_{n-1} & 2(h_{n-1}+h_n) & h_n \\ 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 1 & 2 \end{bmatrix}}_{\underline{\underline{J}}} \cdot \underbrace{\begin{bmatrix} s''(x_0) \\ s''(x_1) \\ s''(x_2) \\ \ldots \\ s''(x_{n-1}) \\ s''(x_n) \end{bmatrix}}_{\underline{a}} = \underbrace{\begin{bmatrix} f''(a) \\ y_1 \\ y_2 \\ \ldots \\ y_{n-1} \\ f''(a) \end{bmatrix}}_{\underline{y}} \quad (\Sigma 2\alpha)$$

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & 0 & \ldots & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 1 \end{bmatrix}}_{\underline{\underline{\hat{J}}}} \cdot \underbrace{\begin{bmatrix} s''(x_0) \\ s''(x_1) \\ s''(x_2) \\ \ldots \\ s''(x_{n-1}) \\ s''(x_n) \end{bmatrix}}_{\underline{a}} = \underbrace{\begin{bmatrix} f''(a) \\ \hat{y}_1 \\ \hat{y}_2 \\ \ldots \\ \hat{y}_{n-1} \\ f''(b) \end{bmatrix}}_{\underline{\hat{y}}} \quad (\Sigma 2\beta)$$

Όμοια αποδεικνύεται και η περίπτωση της «φυσικής κυβικής spline», δηλαδή όταν αναζητούμε $s \in S_3(\Delta)$ η οποία να ικανοποιεί τις συνθήκες

$\{s(x_i) = f(x_i), s''(a) = s''(b) = 0, \quad i = 0,1,2,...,n\}$. Η μόνη διαφορά με τα συστήματα (Σ2α) και (Σ2β) είναι ότι το δεξί μέλος στην πρώτη και τελευταία εξίσωση είναι μηδέν, δηλαδή στην θέση των $f''(a)$ και $f''(b)$ έχουμε το μηδέν.

Να σημειωθεί ότι για την εφαρμογή των συναρτήσεων spline επιλύουμε κατευθείαν τα συστήματα (Σ1α), (Σ1β), (Σ2α) ή (Σ2β), ανάλογα με την περίπτωση, χωρίς να επαναλάβουμε από την αρχή την ανάλυση.

Παράδειγμα: Έστω η συνάρτηση $f = 1/x$. Βρείτε την κυβική συνάρτηση spline η οποία παρεμβάλλεται στην f στα σημεία $x_0 = 1$, $x_1 = 1.5$ και $x_2 = 2$. Ως συνοριακές συνθήκες χρησιμοποιήστε και τις 3 δυνατές επιλογές που αναφέρονται παραπάνω. Σε όλες τις περιπτώσεις βρείτε την πρόβλεψη για την τιμή της συνάρτησης στο $x = 1.25$ και στο $x = 1.75$ καθώς και τα αντίστοιχα απόλυτα σφάλματα.

Λύση: Οι τιμές της συνάρτησης στα σημεία x_0, x_1, x_2 είναι αντίστοιχα $f_0 = 1, f_1 = 2/3, f_2 = 1/2$. Επίσης έχουμε $f'(x) = -1/x^2$ και $f''(x) = 2/x^3$ οπότε για το πρώτο ζεύγος συνοριακών συνθηκών της ζητούμενης κυβική συνάρτηση spline είναι $s'(1) = f'(1) = -1, s'(2) = f'(2) = -1/4$, για το δεύτερο ζεύγος συνοριακών συνθηκών είναι $s''(1) = f''(1) = 2, \quad s''(2) = f''(2) = 1/4$, και για το τρίτο είναι $s''(1) = s''(2) = 0$. Τέλος, ο διαμερισμός είναι ομοιόμορφος με $n = 2$ και $h = (b-a)/n = (2-1)/2 = 1/2$. Έτσι τα συστήματα που πρέπει να επιλυθούν έχουν ως εξής. Για την πρώτη περίπτωση:

$$\begin{bmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} s''(x_0) \\ s''(x_1) \\ s''(x_2) \end{bmatrix} = \begin{bmatrix} \hat{y}_0 \\ \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} \text{ όπου } \hat{y}_0 = \frac{6}{1/2}\left(\frac{2/3-1}{1/2} - (-1) \right) = 4,$$

$$\hat{y}_1 = \frac{6}{h^2}(f_2 - 2f_1 + f_0) = \frac{6}{(1/2)^2}\left(\frac{1}{2} - 2\frac{2}{3} + 1 \right) = 4, \; \hat{y}_2 = \frac{6}{1/2}\left((-1/4) - \frac{1/2 - 2/3}{1/2} \right) = 1.$$

Επιλύουμε το σύστημα και βρίσκουμε $s''(x_0) = \frac{7}{4}, \quad s''(x_1) = \frac{1}{2}, \quad s''(x_2) = \frac{1}{4}$.

Αντικαθιστούμε τις τιμές αυτές στην σχέση (Α) και παίρνουμε την ζητούμενη συνάρτηση:

$$s_a(x) = \frac{1}{24} \times \begin{cases} -10x^3 + 51x^2 - 96x + 79, & 1 \le x \le 3/2 \\ -2x^3 + 15x^2 - 42x + 52, & 3/2 \le x \le 2 \end{cases}$$

Στην δεύτερη περίπτωση θα έχουμε το γραμμικό σύστημα

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 4 & 1 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} s''(x_0) \\ s''(x_1) \\ s''(x_2) \end{bmatrix} = \begin{bmatrix} \hat{y}_0 \\ \hat{y}_1 \\ \hat{y}_2 \end{bmatrix}$$ όπου $\hat{y}_0 = 2$, $\hat{y}_1 = 4$, $\hat{y}_2 = 1/4$, οπότε θα προκύψει $s''(x_0) = 2, \quad s''(x_1) = \frac{7}{16}, \quad s''(x_2) = \frac{1}{4}$ και η ζητούμενη συνάρτηση θα είναι:

$$s_b(x) = \frac{1}{192} \times \begin{cases} -100x^3 + 492x^2 - 883x + 683, & 1 \le x \le 3/2 \\ -12x^3 + 96x^2 - 289x + 386, & 3/2 \le x \le 2 \end{cases}$$

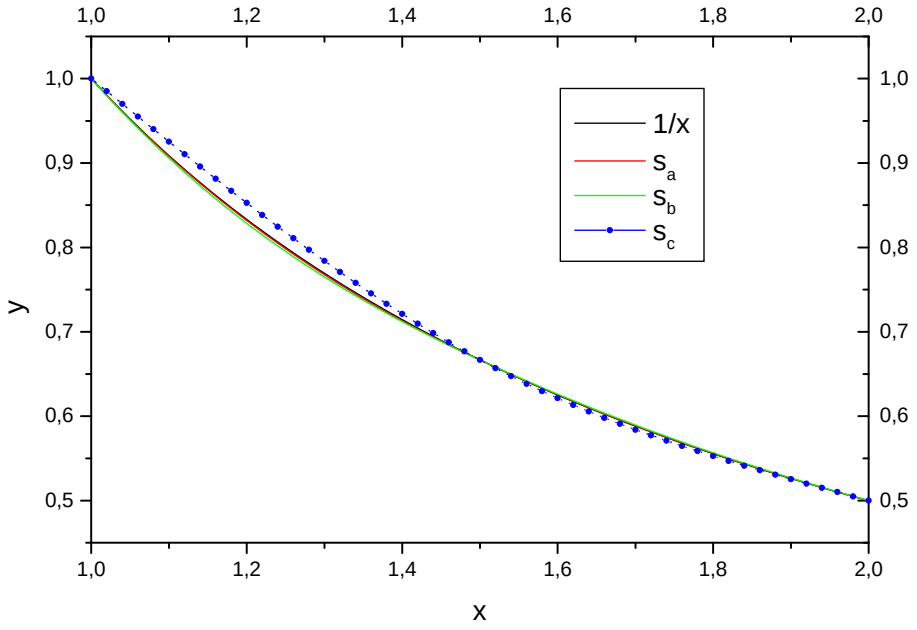
Στην τρίτη περίπτωση θα έχουμε το γραμμικό σύστημα

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 4 & 1 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} s''(x_0) \\ s''(x_1) \\ s''(x_2) \end{bmatrix} = \begin{bmatrix} 0 \\ \hat{y}_1 \\ 0 \end{bmatrix}$$

όπου $\hat{y}_1 = 4$ (όπως και στις προηγούμενες δύο περιπτώσεις) οπότε θα προκύψει $s''(x_0) = 0, \quad s''(x_1) = 1, \quad s''(x_2) = 0$ και η ζητούμενη συνάρτηση θα είναι:

$$s_c(x) = \frac{1}{12} \times \begin{cases} 4x^3 - 12x^2 + 3x + 17, & 1 \le x \le 3/2 \\ -4x^3 + 24x^2 - 51x + 44, & 3/2 \le x \le 2 \end{cases}$$

Αν ζωγραφίσουμε τις 3 αυτές κυβικές splines μαζί με την συνάρτηση f θα πάρουμε το παρακάτω διάγραμμα:



Επιπλέον για $x=1.25$ έχουμε $f=0.8$ ενώ από τις προκύπτουσες συναρτήσεις έχουμε, $s_a \simeq 0.798177$, $s_b \simeq 0.795247$, $s_c \simeq 0.817708$.

Τέλος για $x=1.75$ έχουμε $f=0.571429$ ενώ από τις προκύπτουσες συναρτήσεις έχουμε, $s_a \simeq 0.571615$, $s_b \simeq 0.572591$, $s_c \simeq 0.567708$.

Κεφάλαιο 5ο
Αριθμητική παραγώγιση

5.1. Εισαγωγή

Στο κεφάλαιο αυτό θα μελετήσουμε τρόπους αριθμητικού υπολογισμού των παραγώγων μίας συνάρτησης (γνωστής ή άγνωστης) με βάση αριθμητικά δεδομένα.

Δύο πολύ απλές σχέσεις για τον αριθμητικό υπολογισμό της πρώτης παραγώγου μίας συνάρτησης f σε ένα σημείο x δίνονται μέσω του ορισμού της παραγώγου μίας συνάρτησης:

$$f'(x)=\lim_{h\to 0}\frac{f(x+h)-f(x)}{h},\quad f'(x)=\lim_{h\to 0}\frac{f(x)-f(x-h)}{h}$$

Λόγω όμως του γεγονότος ότι στον υπολογιστή το όριο δεν υφίσταται, είναι δηλαδή μία θεωρητική έννοια, οι παραπάνω σχέσεις δίνουν προσεγγιστικά την τιμή της παραγώγου:

$$f'(x)\approx\frac{f(x+h)-f(x)}{h}\equiv D_{+}^{(1)}f(x),\quad f'(x)\approx\frac{f(x)-f(x-h)}{h}\equiv D_{-}^{(1)}f(x),\quad h<<1$$

Είναι προφανές ότι όσο πιο μικρή η ποσότητα h τόσο πιο ακριβείς γίνονται οι δύο αυτές σχέσεις, οι οποίες αποτελούν τις απλούστερες εκφράσεις πεπερασμένων διαφορών, και οι οποίες αναλύονται παρακάτω. Το σφάλμα των εκφράσεων αυτών βρίσκεται με την χρήση σειρών Taylor. Αν υποθέσουμε ότι η συνάρτηση f είναι δύο φορές συνεχώς παραγωγίσιμη σε μία περιοχή του x, τότε θα έχουμε:

$$f(x+h)=f(x)+hf'(x)+\frac{h^2}{2}f''(\xi_x)\Rightarrow f'(x)=\underbrace{\frac{f(x+h)-f(x)}{h}}_{D_{+}^{(1)}f(x)}-\frac{h}{2}f''(\xi_x),\quad x<\xi_x<x+h$$

$$f(x-h)=f(x)-hf'(x)+\frac{h^2}{2}f''(\xi_x)\Rightarrow f'(x)=\underbrace{\frac{f(x)-f(x-h)}{h}}_{D_{-}^{(1)}f(x)}+\frac{h}{2}f''(\xi_x),\quad x-h<\xi_x<x$$

και επομένως έχουμε, αντίστοιχα:

$$f'(x)-D_{+}^{(1)}f(x)=-\frac{h}{2}f''(\xi_x),\quad x<\xi_x<x+h$$

$$f'(x)-D_{-}^{(1)}f(x)=\frac{h}{2}f''(\xi_x),\quad x-h<\xi_x<x$$

Η ποσότητα στο αριστερό μέλος των παραπάνω σχέσεων είναι το σφάλμα $\varepsilon(x)$ (ακριβής τιμή μείον την προσεγγιστική τιμή) το οποίο μάλιστα, αν η f'' είναι συνεχής στο διάστημα $[x-h, x+h]$, φράσεται ως εξής: $|\varepsilon(x)| = \left|\frac{h}{2} f''(\xi_x)\right| \le \frac{h}{2} M, \quad M \equiv \max_{y\varepsilon[x-h,x+h]} |f''(y)|$. Η σχέση αυτή δείχνει ότι και στις δύο περιπτώσεις το μέγιστο απόλυτο σφάλμα είναι ανάλογο του h το οποίο βέβαια θα τείνει στο μηδέν καθώς $h \to 0$. Η ποσότητα $\frac{h}{2} M$ είναι το θεωρητικό σφάλμα της μεθόδου γνωστό και ως *σφάλμα αποκοπής*.

Θα μελετήσουμε στην συνέχεια δύο μεθοδολογίες παραγωγής προσεγγιστικών τύπων υπολογισμού των παραγώγων μίας συνάρτησης και της εκτίμησης για τα αντίστοιχα σφάλματα.

5.2. Υπολογισμός παραγώγων με χρήση του πολυωνύμου παρεμβολής

Το πολυώνυμο παρεμβολής το οποίο μελετήθηκε στο 4ο κεφάλαιο μπορεί να χρησιμοποιηθεί για να προσεγγιστούν οι παράγωγοι της συνάρτησης στα ζητούμενα σημεία. Αν $p \in P_n$ είναι το πολυώνυμο παρεμβολής της δοσμένης συνάρτησης στα σημεία $\{x_0, x_1, ..., x_n\}$ τότε $p(x) \simeq f(x) \Rightarrow p^{(k)}(x) \simeq f^{(k)}(x), k = 0,1,2,...$. Ας δούμε τι συμβαίνει στην περίπτωση της πρώτης παραγώγου της f. Όπως δείξαμε για το σφάλμα ισχύει $f(x) - p(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{j=0}^{n} (x - x_j), \xi_x \in (a,b)$. Αν παραγωγίσουμε αυτή την σχέση θα έχουμε:

$$f'(x) - p'(x) = \frac{d}{dx}\left(\frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{j=0}^{n} (x - x_j) \right) \Rightarrow$$

$$f'(x) - p'(x) = \frac{1}{(n+1)!} \left\{ \frac{d}{dx}\left(\prod_{i=0}^{n} (x - x_i) \right) f^{(n+1)}(\xi_x) + \prod_{i=0}^{n} (x - x_i) \frac{d}{dx}\left(f^{(n+1)}(\xi_x) \right) \right\}$$

Δυστυχώς η παράγωγος $\frac{d}{dx}\left(f^{(n+1)}(\xi_x) \right)$ δεν μπορεί να εκτιμηθεί αφού δεν ξέρουμε τον τρόπο με τον οποίο μεταβάλλεται το ξ_x ως προς x. Αν όμως επιλέξουμε $x_i \in \{x_0, x_1, ..., x_n\}$ τότε παίρνουμε

$$f'(x_i) - p'(x_i) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\prod_{\substack{j=0\\ j\neq i}}^{n}(x_i - x_j) \quad (*)$$

Το πολυώνυμο Lagrange δίνεται από την σχέση $p(x) = \sum_{j=0}^{n} L_j(x)\, f(x_j)$ και επομένως η παράγωγός του είναι $p'(x) = \frac{d}{dx}\left(\sum_{j=0}^{n} L_j(x)\, f(x_j)\right) \Rightarrow p'(x_j) = \sum_{j=0}^{n} L_j'(x_i)\, f(x_j)$.

Αντικαθιστώντας την έκφραση αυτή στην σχέση (*) θα έχουμε:

$$\boxed{f'(x_i) = \sum_{j=0}^{n} L_j'(x_i)\, f(x_j) + \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\prod_{\substack{j=0\\ i\neq j}}^{n}(x_i - x_j)}$$

Η παραπάνω σχέση ονομάζεται τύπος των «n+1» σημείων για την προσέγγιση της $f'(x_i)$ αφού ουσιαστικά πρόκειται για έναν γραμμικό συνδυασμό των τιμών της συνάρτησης $f(x_j)$, $j = 0,1,2,...,n$. Η σχέση αυτή συνήθως χρησιμοποιείται για $n = 2$.

Θα προχωρήσουμε στην συνέχεια με άλλη, λίγο διαφορετική, μέθοδο η οποία μπορεί εύκολα να χρησιμοποιηθεί για να παραχθούν τύποι για παραγώγους ανώτερης τάξης της συνάρτησης f μαζί με τις αντίστοιχες εκφράσεις για το σφάλμα.

5.3. Τύποι πεπερασμένων διαφορών

Οι τύποι πεπερασμένων διαφορών προκύπτουν εύκολα με την μέθοδο των προσδιοριστέων συντελεστών. Στην συνέχεια θα δοθεί ένα παράδειγμα για την ευκολότερη κατανόηση της μεθόδου. Για περαιτέρω απλούστευση, θα θεωρήσουμε ότι τα σημεία στα οποία γνωρίζουμε την τιμή της συνάρτησης είναι ισαπέχοντα καθώς και ότι η άγνωστη συνάρτηση ικανοποιεί τις συνθήκες ομαλότητας που απαιτούνται από την ανάλυσή μας.

Έστω ότι έχουμε διαθέσιμα τα ζεύγη σημείων $(x-h, f(x-h))$, $(x, f(x))$ και $(x+h, f(x+h))$ και ότι θέλουμε να βρούμε, προσεγγιστικά, την δεύτερη παράγωγο της συνάρτησης f στο σημείο x, την οποία θα ονομάσουμε $D^{(2)} f(x)$. Αρχικά εκφράζουμε την ποσότητα αυτή ως γραμμικό συνδυασμό των δεδομένων μας:

$$D^{(2)} f(x) = af(x-h) + bf(x) + cf(x+h) \quad (1)$$

όπου οι σταθερές a,b,c θα πρέπει να προσδιοριστούν με συνεπή τρόπο. Ορίζουμε το σφάλμα, το οποίο δίνεται ως την διαφορά της πραγματικής τιμής της 2ης παραγώγου της συνάρτησης στο σημείο x από την προσεγγιστική τιμή, δηλαδή,

$$\varepsilon(x) \equiv f^{(2)}(x) - D^{(2)} f(x) \quad (2)$$

Η παραπάνω ποσότητα, λόγω της (1), γίνεται:

$$\varepsilon(x) \equiv f^{(2)}(x) - af(x-h) - bf(x) - cf(x+h) \quad (3)$$

Στην συνέχεια, εκφράζουμε όλες τις ποσότητες που εμφανίζονται στην έκφραση του σφάλματος ως σειρές Taylor γύρω από το σημείο x, έτσι ώστε η τελική σχέση για το σφάλμα να δίνεται μόνο ως έκφραση της συνάρτησης f και των παραγώγων της στο σημείο x. Πράγματι έχουμε ότι:

$$f(x+kh) = f(x) + \frac{kh}{1!} f'(x) + \frac{(kh)^2}{2!} f^{(2)}(x) + \frac{(kh)^3}{3!} f^{(3)}(x) + \frac{(kh)^4}{4!} f^{(4)}(x) + \ldots \quad (4)$$

Για $k = +1, -1$ η παραπάνω σχέση δίνει, αντίστοιχα:

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2} f^{(2)}(x) + \frac{h^3}{6} f^{(3)}(x) + \frac{h^4}{24} f^{(3)}(x) + \ldots \quad (5\alpha)$$

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2} f^{(2)}(x) - \frac{h^3}{6} f^{(3)}(x) + \frac{h^4}{24} f^{(4)}(x) + \ldots \quad (5\beta)$$

Αντικαθιστώντας τις 2 αυτές σχέσεις στην έκφραση του σφάλματος, εξίσωση (3), και συλλέγοντας τους συντελεστές των $f(x), f'(x), f^{(2)}(x), f^{(3)}(x), \ldots$ προκύπτει:

$$\varepsilon(x) \equiv -(a+b+c) f(x) + (ah - ch) f'(x) + \left(1 - a\frac{h^2}{2} - c\frac{h^2}{2}\right) f^{(2)}(x) + \\ + \left(a\frac{h^3}{6} - c\frac{h^3}{6}\right) f^{(3)}(x) - \left(a\frac{h^4}{24} + c\frac{h^4}{24}\right) f^{(4)}(x) + \ldots \quad (6)$$

Η σχέση αυτή θα πρέπει να ισχύει για κάνε τιμή του x. Επιπλέον, είναι φανερό ότι όσοι περισσότεροι όροι της σειράς απαλειφθούν τόσο το καλύτερο (με δεδομένο ότι η συνάρτηση f συμπεριφέρεται ομαλά). Πράγματι, από την σχέση (6) βλέπουμε ότι μπορούμε να μηδενίσουμε τουλάχιστον τους συντελεστές των $f(x), f'(x), f^{(2)}(x)$, δηλαδή:

$$\begin{aligned} &a+b+c=0\\ &ah-ch=0\\ &a\frac{h^2}{2}+c\frac{h^2}{2}=1 \end{aligned}$$ είτε σε μορφή πίνακα: $$\begin{bmatrix} 1 & 1 & 1\\ h & 0 & -h\\ \frac{h^2}{2} & 0 & \frac{h^2}{2} \end{bmatrix}\cdot\begin{bmatrix} a\\ b\\ c \end{bmatrix}=\begin{bmatrix} 0\\ 0\\ 1 \end{bmatrix}$$. Έχουμε δηλαδή ένα γραμμικό, μη-ομογενές γραμμικό σύστημα το οποίο θα έχει μοναδική λύση εφόσον η ορίζουσα του πίνακα στο αριστερό μέλος είναι μη-μηδενική. Πράγματι η ορίζουσα είναι ίση με h^3 και επομένως το πρόβλημα έχει μοναδική λύση για κάθε $h\neq 0$. Η λύση του είναι $a=\frac{1}{h^2}, b=-\frac{2}{h^2}, c=\frac{1}{h^2}$ οπότε η (1) θα δώσει:

$$D^{(2)}f(x)=\frac{f(x-h)-2f(x)+f(x+h)}{h^2}$$

Για το σφάλμα η σχέση (6) θα δώσει:

$$\varepsilon(x)\equiv-(a+c)\frac{h^4}{24}f^{(4)}(x)+\ldots=-\frac{1}{h^2}\frac{h^4}{12}f^{(4)}(x)+\ldots=-\frac{h^2}{12}f^{(4)}(x)+\ldots$$

βλέπουμε δηλαδή ότι λόγω του γεγονότος $a=c$ μηδενίζεται και ο συντελεστής του $f^{(3)}(x)$ (σημείωση: δεν συμβαίνει πάντα αυτό). Το σφάλμα το οποίο προκύπτει είναι το θεωρητικό σφάλμα της μεθόδου και ονομάζεται σφάλμα αποκοπής (για προφανείς λόγους). Θα μπορούσαμε λοιπόν να είχαμε τερματήσει τις σειρές Taylor στον 5ο όρο οπότε θα είχαμε:

$$\varepsilon(x)\equiv-\frac{h^2}{12}f^{(4)}(\xi_x),\quad x-h<\xi_x<x+h$$

Λόγω του γεγονότος ότι το σφάλμα αποκοπής είναι ανάλογο του h^2 η αντίστοιχη σχέση πεπερασμένων διαφορών ονομάζεται 2ης τάξης ακρίβειας. Τώρα, αν η $f^{(4)}(x)$ είναι συνεχής στο $[x-h,x+h]$ τότε εύκολα βρίσκουμε ένα άνω φράγμα του σφάλματος,

$$\left|\varepsilon(x)\right|=\left|\frac{h^2}{12}f^{(4)}(\xi_x)\right|\leq\frac{h^2}{12}M,\quad M\equiv\max_{x-h\leq y\leq x+h}\left|f^{(4)}(y)\right|$$

Είναι προφανές ότι $\frac{h^2}{12}M\underset{h\to 0}{\to}0$. Δυστυχώς όμως, στην πραγματικότητα δεν συμβαίνει αυτό. Αυτό οφείλεται στο ότι οι πράξεις γίνονται πάντα με μία συγκεκριμένη ακρίβεια, δηλαδή εκτός του σφάλματος αποκοπής θα υπάρχει και το σφάλμα στρογγυλοποίησης. Επομένως αντί για την ποσότητα $D^{(2)}f(x)$ θα υπολογιστεί η ποσότητα $\tilde{D}^{(2)}f(x)\equiv fl\left(D^{(2)}f(x)\right)$. Έστω λοιπόν ότι αντί για τις ποσότητες

$f(x-h), f(x), f(x+h)$ έχουμε τις ποσότητες $\tilde{f}_{-1} \equiv fl\left(f(x-h)\right)$, $\tilde{f}_0 \equiv fl\left(f(x)\right)$ και $\tilde{f}_{+1} \equiv fl\left(f(x+h)\right)$ οι οποίες έχουν αντίστοιχα σφάλματα $\sigma_{-1} \equiv f(x-h) - fl\left(f(x-h)\right)$, $\sigma_0 \equiv f(x) - fl\left(f(x)\right), \sigma_1 \equiv f(x+h) - fl\left(f(x+h)\right)$. Άρα θα έχουμε

$$D^{(2)} f(x) = \frac{\left(\tilde{f}_{-1} + \sigma_{-1}\right) - 2\left(\tilde{f}_0 + \sigma_0\right) + \left(\tilde{f}_{+1} + \sigma_{+1}\right)}{h^2} = \tilde{D}^{(2)} f(x) + \frac{\sigma_{-1} - 2\sigma_0 + \sigma_{+1}}{h^2} \Rightarrow$$

$$\tilde{D}^{(2)} f(x) = D^{(2)} f(x) - \frac{\sigma_{-1} - 2\sigma_0 + \sigma_{+1}}{h^2}$$

και το πραγματικό σφάλμα θα είναι:

$$\varepsilon(x) \equiv f^{(2)}(x) - \tilde{D}^{(2)} f(x) = f^{(2)}(x) - \left(D^{(2)} f(x) - \frac{\sigma_{-1} - 2\sigma_0 + \sigma_{+1}}{h^2}\right) \Rightarrow$$

$$\varepsilon(x) = f^{(2)}(x) - D^{(2)} f(x) + \frac{\sigma_{-1} - 2\sigma_0 + \sigma_{+1}}{h^2} \Rightarrow \varepsilon(x) = -\frac{h^2}{12} f^{(4)}(\xi_x) + \frac{\sigma_{-1} - 2\sigma_0 + \sigma_{+1}}{h^2}$$

και επομένως

$$\left|\varepsilon(x)\right| = \left|-\frac{h^2}{12} f^{(4)}(\xi_x) + \frac{\sigma_{-1} - 2\sigma_0 + \sigma_{+1}}{h^2}\right| \le \frac{h^2}{12} M + \left|\frac{\sigma_{-1} - 2\sigma_0 + \sigma_{+1}}{h^2}\right| \le \frac{h^2}{12} M + \frac{\left|\sigma_{-1}\right| + 2\left|\sigma_0\right| + \left|\sigma_{+1}\right|}{h^2} \Rightarrow$$

$$\left|\varepsilon(x)\right| \le \frac{h^2}{12} M + \frac{4\sigma}{h^2}, \quad M \equiv \max_{x-h \le y \le x+h} \left|f^{(4)}(y)\right|, \quad \sigma \equiv \max\left\{\left|\sigma_{-1}\right|, \left|\sigma_0\right|, \left|\sigma_{+1}\right|\right\}$$

Μπορούμε εύκολα να βρούμε που ελαχιστοποιείται το άνω φράγμα του σφάλματος. Θεωρούμε την συνάρτηση $g(h) = \frac{h^2}{12} M + \frac{4\sigma}{h^2}$ για την οποία έχουμε ότι $g'(h) = \frac{h}{6} M - \frac{8\sigma}{h^3}$ και $g''(h) = \frac{1}{6} M + \frac{24\sigma}{h^4} > 0$ για κάθε $h > 0$. Η πρώτη παράγωγος μηδενίζεται στο σημείο $g'(h^*) = \frac{h^*}{6} M - \frac{8\sigma}{h^{*3}} = 0 \Rightarrow h^* = \left(\frac{48\sigma}{M}\right)^{1/4}$ το οποίο βέβαια αποτελεί το σημείο ελαχιστοποίησης της $g = g(h)$ αφού η δεύτερη παράγωγος της συνάρτησης, g'', είναι πάντα θετική.

Θα συνεχίσουμε με ένα ακόμα παράδειγμα. Έστω ότι μας δίνεται ένας συγκεκριμένος τύπος πεπερασμένων διαφορών, $D^{(1)} f(x) = \frac{f(x+h) - f(x-h)}{2h}$, για τον οποίο θέλουμε να υπολογίσουμε το μέγιστο συνολικό σφάλμα, δηλαδή το άνω φράγμα του αθροίσματος του σφάλματος αποκοπής και του σφάλματος στρογγύλευσης.

Λύση: Χρησιμοποιούμε τις σχέσεις (5α) και (5β) και αντικαθιστούμε παίρνουμε

$$D^{(1)}f(x)=\frac{\left\{f(x)+hf'(x)+\frac{h^2}{2}f''(x)+\frac{h^3}{6}f'''(x)+...\right\}-\left\{f(x)-hf'(x)+\frac{h^2}{2}f''(x)-\frac{h^3}{6}f'''(x)+...\right\}}{2h}$$

ή

$$D^{(1)}f(x)=\frac{2hf'(x)+2\frac{h^3}{6}f'''(x)+...}{2h}\Rightarrow D^{(1)}f(x)=f'(x)+\frac{h^2}{6}f'''(x)+...\Rightarrow$$

$$\underbrace{f'(x)-D^{(1)}f(x)}_{\varepsilon(x)}=-\frac{h^2}{6}f'''(x)+...\Rightarrow\varepsilon(x)=-\frac{h^2}{6}f'''(x)+...$$. Έτσι, αν η συνάρτηση είναι 3 φορές συνεχώς παραγωγίσιμη στο διάστημα [x-h,x+h] θα έχουμε $\varepsilon(x)=-\frac{h^2}{6}f'''(\xi_x),\ x-h<\xi_x<x+h$ που μας δείχνει ότι το θεωρητικό σφάλμα (δηλαδή το σφάλμα αποκοπής) είναι δεύτερης τάξης. Λόγω όμως των σφαλμάτων στρογγύλευσης αντί για την ποσότητα $D^{(1)}f(x)$ θα υπολογιστεί η ποσότητα $\tilde{D}^{(1)}f(x)\equiv fl\left(D^{(1)}f(x)\right)$. Έτσι παίρνουμε (όπως προηγουμένως) $\varepsilon(x)=-\frac{h^2}{6}f'''(\xi)+\frac{\sigma_1-\sigma_{-1}}{2h}$ και παίρνοντας απόλυτες τιμές και εφαρμόζοντας την τριγωνική ανισότητα τελικά προκύπτει:

$$\boxed{\left|\varepsilon(x)\right|\le\frac{h^2}{6}M+\frac{\sigma}{h},\quad M\equiv\max_{y\in[x-h,x+h]}\left|f^{(3)}(y)\right|,\quad \sigma\equiv\max\left\{\left|\sigma_{-1}\right|,\left|\sigma_1\right|\right\}}$$

όπου φυσικά ο πρώτος όρος στο δεξί μέλος της ανισότητας αντιστοιχεί στο λάθος αποκοπής και ο δεύτερος στο λάθος στρογγύλευσης. Η ποσότητα στο δεξί μέλος της ανισότητας είναι το μέγιστο απόλυτο συνολικό σφάλμα. Για να βρούμε το σημείο ελαχιστοποίησης του σφάλματος αυτού θεωρούμε την συνάρτηση $g(h)=\frac{h^2}{6}M+\frac{\sigma}{h}$ για την οποία ισχύει $g'(h)=\frac{h}{3}M-\frac{\sigma}{h^2}$ και $g''(h)=\frac{1}{3}M+\frac{2\sigma}{h^3}>0$ για κάθε $h>0$ αφού τα σ και M είναι θετικοί αριθμοί. Το σημείο μηδενισμού της πρώτης παραγώγου είναι $g'(h^*)=\frac{h^*}{3}M-\frac{\sigma}{h^{*2}}=0\Rightarrow h^*=\left(\frac{3\sigma}{M}\right)^{1/3}$ το οποίο φυσικά αποτελεί ελάχιστο της g αφού η g'' είναι πάντα θετική.

Πριν το κλείσιμο αυτού του κεφαλαίου να σημειωθεί ότι οι τύποι των πεπερασμένων διαφορών διακρίνονται σε:

(α) «προς τα εμπρός» αν οι τιμές της συνάρτησης που χρησιμοποιούνται βρίσκονται όλες αριστερά από το σημείο x_i στο οποίο ζητείται να υπολογισθούν οι παράγωγοί της

(β) «προς τα πίσω» αν οι τιμές της συνάρτησης που χρησιμοποιούνται βρίσκονται όλες δεξιά από το σημείο x_i

(γ) «κεντρικούς», αν οι τιμές της συνάρτησης που χρησιμοποιούνται βρίσκονται τόσο αριστερά όσο και δεξιά από το σημείο x_i.

Ακολουθούν οι τύποι πεπερασμένων διαφορών για την πρώτη, δεύτερη και τρίτη παράγωγο που προκύπτουν με την μέθοδο των προσδιοριστέων συντελεστών μαζί με την τάξη ακρίβειας τους. Για το καθένα από τους τύπους αυτούς προσδιορίστε (α) το σφάλμα αποκοπής, (β) το μέγιστο απόλυτο συνολικό σφάλμα (δηλαδή το άνω φράγμα τους αθροίσματος του σφάλματος αποκοπής και του σφάλματος στρογγύλευσης) καθώς και (γ) το βέλτιστο h^* για την ελαχιστοποίηση του σφάλματος αυτού.

5.3.1. Τύποι πεπερασμένες διαφορές προς τα εμπρός.

Πρώτη παράγωγος.

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_i)}{h} + O(h)$$

$$f'(x_i) = \frac{-f(x_{i+2}) + 4f(x_{i+1}) - 3f(x_i)}{2h} + O(h^2)$$

$$f'(x_i) = \frac{-f(x_{i+4}) + \frac{16}{3}f(x_{i+3}) - 12f(x_{i+2}) + 16f(x_{i+1}) - \frac{25}{3}f(x_i)}{4h} + O(h^4)$$

Δεύτερη παράγωγος.

$$f''(x_i) = \frac{f(x_{i+2}) - 2f(x_{i+1}) + f(x_i)}{h^2} + O(h)$$

$$f''(x_i) = \frac{-f(x_{i+3}) + 4f(x_{i+2}) - 5f(x_{i+1}) + 2f(x_i)}{h^2} + O(h^2)$$

$$f''(x_i) = \frac{-10f(x_{i+5}) + 61f(x_{i+4}) - 156f(x_{i+3}) + 214f(x_{i+2}) - 154f(x_{i+1}) + 45f(x_i)}{12h^2} + O(h^4)$$

Τρίτη παράγωγος.

$$f'''(x_i) = \frac{f(x_{i+3}) - 3f(x_{i+2}) + 3f(x_{i+1}) - f(x_i)}{h^3} + O(h)$$

$$f'''(x_i) = \frac{-3f(x_{i+4}) + 14f(x_{i+3}) - 24f(x_{i+2}) + 18f(x_{i+1}) - 5f(x_i)}{2h^3} + O(h^2)$$

$$f'''(x_i) = \frac{-15f(x_{i+6}) + 104f(x_{i+5}) - 307f(x_{i+4}) + 496f(x_{i+3}) - 461f(x_{i+2}) + 232f(x_{i+1}) - 49f(x_i)}{8h^3} + O(h^4)$$

5.3.2. Τύποι πεπερασμένων διαφορών προς τα πίσω.

Πρώτη παράγωγος

$$f'(x_i) = \frac{f(x_i) - f(x_{i-1})}{h} + O(h)$$

$$f'(x_i) = \frac{3f(x_i) - 4f(x_{i-1}) + f(x_{i-2})}{2h} + O(h^2)$$

$$f'(x_i) = \frac{\frac{25}{3}f(x_i) - 16f(x_{i-1}) + 12f(x_{i-2}) - \frac{16}{3}f(x_{i-3}) + f(x_{i-4})}{4h} + O(h^4)$$

Δεύτερη παράγωγος.

$$f''(x_i) = \frac{f(x_i) - 2f(x_{i-1}) + f(x_{i-2})}{h^2} + O(h)$$

$$f''(x_i) = \frac{2f(x_i) - 5f(x_{i-1}) + 4f(x_{i-2}) - f(x_{i-3})}{h^2} + O(h^2)$$

$$f''(x_i) = \frac{45f(x_i) - 154f(x_{i-1}) + 214f(x_{i-2}) - 156f(x_{i-3}) + 61f(x_{i-4}) - 10f(x_{i-5})}{12h^2} + O(h^4)$$

Τρίτη παράγωγος.

$$f'''(x_i) = \frac{f(x_i) - 3f(x_{i-1}) + 3f(x_{i-2}) - f(x_{i-3})}{h^3} + O(h)$$

$$f'''(x_i) = \frac{5f(x_i) - 18f(x_{i-1}) + 24f(x_{i-2}) - 14f(x_{i-3}) + 3f(x_{i-4})}{2h^3} + O(h^2)$$

$$f'''(x_i) = \frac{49f(x_i) - 232f(x_{i-1}) + 461f(x_{i-2}) - 496f(x_{i-3}) + 307f(x_{i-4}) - 104f(x_{i-5}) + 15f(x_{i-6})}{8h^3} + O(h^4)$$

5.3.3. Τύποι κεντρικών πεπερασμένων διαφορών

Πρώτη παράγωγος.

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_{i-1})}{2h} + O(h^2)$$

$$f'(x_i) = \frac{-f(x_{i+2}) + 8f(x_{i+1}) - 8f(x_{i-1}) + f(x_{i-2})}{12h} + O(h^4)$$

Δεύτερη παράγωγος.

$$f''(x_i) = \frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{h^2} + O(h^2)$$

$$f''(x_i) = \frac{-f(x_{i+2}) + 16f(x_{i+1}) - 30f(x_i) + 16f(x_{i-1}) - f(x_{i-2})}{12h^2} + O(h^4)$$

Τρίτη παράγωγος.

$$f'''(x_i) = \frac{f(x_{i+2}) - 2f(x_{i+1}) + 2f(x_{i-1}) - f(x_{i-2})}{2h^3} + O(h^2).$$

$$f'''(x_i) = \frac{-f(x_{i+3}) + 8f(x_{i+2}) - 13f(x_{i+1}) + 13f(x_{i-1}) - 8f(x_{i-2}) + f(x_{i-3})}{8h^3} + O(h^4)$$

Κεφάλαιο 6ο
Αριθμητική ολοκλήρωση

6.1 Εισαγωγή

Σκοπός στο παρόν κεφάλαιο είναι η ανάπτυξη μεθόδων για τον αριθμητικό υπολογισμό ορισμένων ολοκληρωμάτων, δηλαδή ποσοτήτων της μορφής, $I(f)=\int_a^b f(x)dx$ όπου $f:[a,b]\to\mathbb{R}$. Η ανάπτυξη των μεθόδων αυτών ξεκινά με το να θεωρήσουμε ένα διαμερισμό του διαστήματος ενδιαφέροντος [a,b], $\Delta: a\equiv x_1<x_2<\ldots<x_n<x_{n+1}\equiv b$, όχι απαραίτητα ομοιόμορφο, δηλαδή χωρίζουμε το διάστημα [a,b] σε «n» στο πλήθος υποδιαστήματα. Έχουμε:

$$I(f)=\int_a^b f(x)dx=\underbrace{\int_{x_1}^{x_2} f(x)dx}_{I_1(f)}+\underbrace{\int_{x_2}^{x_3} f(x)dx}_{I_2(f)}+\ldots+\underbrace{\int_{x_n}^{x_{n+1}} f(x)dx}_{I_n(f)}=\sum_{i=1}^{n}\int_{x_i}^{x_{i+1}} f(x)dx=\sum_{i=1}^{n} I_i(f)$$

οπότε το πρόβλημα ανάγεται στον υπολογισμό των επιμέρους ολοκληρωμάτων $I_i(f)$. Η σχέση η οποία χρησιμοποιείται για τον υπολογισμό των επιμέρους ολοκληρωμάτων ονομάζεται *κανόνας ολοκλήρωσης*. Η σχέση που προκύπτει για τον υπολογισμό του συνολικού ολοκληρώματος ονομάζεται *σύνθετος κανόνας ολοκλήρωσης*. Στην συνέχεια ακολουθούν διάφοροι κανόνες ολοκλήρωσης, μαζί με τους αντίστοιχους σύνθετους κανόνες, καθώς επίσης και η απαραίτητη ανάλυση σφάλματος.

6.2. Μέθοδος ορθογωνίου

Η μέθοδος αυτή έχει τρεις εκδοχές. Θα περιοριστούμε όμως σε εκείνη που δίνει την μεγαλύτερη ακρίβεια στο τελικό αποτελέσμα, η οποία αναφέρεται συνήθως και ως μέθοδος του μέσου σημείου. Η ιδέα της μεθόδου αυτής είναι η εξής. Για το κάθε υποδιάστημα $[x_i,x_{i+1}]$ ορίζουμε το μέσο του $y_i=x_i+\frac{x_{i+1}-x_i}{2}=\frac{x_i+x_{i+1}}{2}$ και θεωρούμε ότι το ολοκλήρωμα της συνάρτησης, το οποίο βέβαια είναι το εμβαδό του χωρίου κάτω από την καμπύλη της συνάρτησης, είναι προσεγγιστικά ίσο με το εμβαδό του ορθογωνίου που φαίνεται στο σχήμα 1, δηλαδή $I_i(f)\approx(x_{i+1}-x_i)f(y_i)\Rightarrow$ $I_i(f)\approx h_i f(y_i)=R_i(f), i=1,2,\ldots,n$. Επομένως ο σύνθετος κανόνας του ορθογωνίου, $R(f)$, έχει ως εξής:

$$\boxed{R(f)=\sum_{i=1}^{n}R_i(f)=\sum_{i=1}^{n}h_i f(y_i),\quad h_i=x_{i+1}-x_i,\ y_i=(x_{i+1}+x_i)/2}$$

είτε

$$R(f)=\sum_{i=1}^{n}(x_{i+1}-x_i)\,f\left(\frac{x_{i+1}+x_i}{2}\right)$$

και για την περίπτωση του ομοιόμορφου διαμερισμού έχουμε:

$$\boxed{R(f)=h\sum_{i=1}^{n}f\left(\frac{x_{i+1}+x_i}{2}\right),\quad h\equiv(b-a)/n}$$

Ανάλυση σφάλματος για την μέθοδο του ορθογωνίου

Στην συνέχεια, θα υπολογίσουμε το σφάλμα με το οποίο υπολογίζεται το ολοκλήρωμα το οποίο μας ενδιαφέρει. Αρχικά θεωρούμε το ανάπτυγμα Taylor της συνάρτησης f γύρω από το σημείο y_i, υποθέτοντας φυσικά ότι όλες οι παράγωγοι της συνάρτησης υπάρχουν:

$$f(x)=f(y_i)+(x-y_i)\,f'(y_i)+\frac{(x-y_i)^2}{2}f''(y_i)+\frac{(x-y_i)^3}{6}f'''(y_i)+\ldots$$

Το ανάπτυγμα αυτό το ολοκληρώνουμε μεταξύ x_i και x_{i+1} για να πάρουμε:

$$\underbrace{\int_{x_i}^{x_{i+1}}f(x)\,dx}_{I_i(f)}=\underbrace{\int_{x_i}^{x_{i+1}}f(y_i)\,dx}_{R_i(f)}+\int_{x_i}^{x_{i+1}}(x-y_i)\,f'(y_i)\,dx+\int_{xi}^{x_{i+1}}\frac{(x-y_i)^2}{2}f''(y_i)\,dx+\int_{xi}^{x_{i+1}}\frac{(x-y_i)^3}{6}f'''(y_i)\,dx+\ldots$$

είτε

$$\underbrace{I_i(f)-R_i(f)}_{\varepsilon_i(f)}=f'(y_i)\int_{x_i}^{x_{i+1}}(x-y_i)\,dx+\frac{f''(y_i)}{2}\int_{xi}^{x_{i+1}}(x-y_i)^2\,dx+\frac{f'''(y_i)}{6}\int_{xi}^{x_{i+1}}(x-y_i)^3\,dx+\ldots$$

όπου η ποσότητα στο αριστερό μέλος είναι το σφάλμα $\varepsilon_i(f)$, δηλαδή η διαφορά της προσεγγιστικής από την ακριβής τιμή του ολοκληρώματος. Τα ολοκληρώματα στο δεξί μέλος της παραπάνω σχέσης υπολογίζονται εύκολα:

$$\int_{x_i}^{x_{i+1}}(x-y_i)^p\,dx=\frac{1}{p+1}(x-y_i)^{p+1}\Big|_{x_i}^{x_{i+1}}=\frac{1}{p+1}\left((x_{i+1}-y_i)^{p+1}-(x_i-y_i)^{p+1}\right)=\begin{cases}0, & \text{p περιττος}\\ \dfrac{h_i^{p+1}}{2^p(p+1)}, & \text{p αρτιος}\end{cases}$$

και επομένως η σχέση για το σφάλμα διαμορφώνεται ως εξής:

$$\varepsilon_i^R(f)=\frac{f''(y_i)}{24}h_i^3+\frac{f^{(4)}(y_i)}{1920}h_i^5+O\left(h_i^7\right)\quad (*)$$

Αν τώρα αθροίσουμε την παραπάνω σχέση σε όλα τα υποδιαστήματα θα πάρουμε το συνολικό σφάλμα $\varepsilon^R(f)$:

$$\varepsilon^R(f) \equiv \sum_{i=1}^{n} \varepsilon_i^R(f) = \underbrace{\frac{1}{24}\sum_{i=1}^{n} f''(y_i)h_i^3}_{E} + \underbrace{\frac{1}{1920}\sum_{i=1}^{n} f^{(4)}(y_i)h_i^5}_{F} + O\left(h_i^7\right) = E + F + O.Y.T.$$

όπου *O.Y.T.* σημαίνει «όροι υψηλότερης τάξης», δηλαδή λιγότερο σημαντικοί σε μέγεθος όροι. Αν όμως $h_i << 1$ τότε $h_i^3 << h_i^5$ και επομένως η τελευταία σχέση δίνει:

$$\varepsilon^R(f) \approx E \Rightarrow \left|\varepsilon^R\right| \approx |E| = \left|\frac{1}{24}\sum_{i=1}^{n} f''(y_i)h_i^3\right| = \frac{1}{24}\left|\sum_{i=1}^{n} f''(y_i)h_i^3\right| \le \frac{1}{24}\sum_{i=1}^{n}\left(\left|f''(y_i)\right|h_i^3\right) \Rightarrow$$

$$\left|\varepsilon^R\right| \le \frac{M}{24}\sum_{i=1}^{n} h_i^3, \quad M \equiv \max_{a\le x\le b}\left|f''(x)\right|$$

Για ομοιόμορφο διαμερισμό ισχύει $h_i \equiv h = \frac{b-a}{n}$ και επομένως $\sum_{i=1}^{n} h_i^3 = nh^3$ και άρα

$$\left|\varepsilon^R\right| \le \frac{M}{24} n\left(\frac{b-a}{n}\right)^3 = \frac{M}{24} n\frac{(b-a)}{n}\left(\frac{b-a}{n}\right)^2 \Rightarrow \left|\varepsilon^R\right| \le \frac{M}{24}(b-a)h^2 \quad (**)$$

Η σχέση (*) μας δείχνει ότι, τοπικά, η μέθοδος του ορθογωνίου είναι 3ης τάξης ενώ ολικά (δηλαδή σε όλο το διάστημα ενδιαφέροντος) είναι 2ης τάξης. Η σχέση (**) μπορεί να χρησιμοποιηθεί για να γίνει μία εκτίμηση του μήκους h, για ομοιόμορφο διαμερισμό, που απαιτείται για να προσδιορισθεί το ζητούμενο ολοκλήρωμα με μέγιστο σφάλμα, *Σ*. Πράγματι από την (**) έχουμε $\varepsilon_{\max} = \frac{M}{24}(b-a)h^2 \le \Sigma \Rightarrow h \le \sqrt{\frac{24\Sigma}{M(b-a)}} \Rightarrow$

$h_{\max} = \sqrt{\frac{24\Sigma}{M(b-a)}}$. Όμως η σταθερά M είναι γενικά δύσκολο να προσδιορισθεί είτε είναι άγνωστη αν δεν ξέρουμε την f. Στην περίπτωση αυτή θεωρούμε ότι $M = O(1)$ και επομένως παίρνουμε μία εκτίμηση του $h_{\max}$, $h_{\max} \approx \sqrt{\frac{24\Sigma}{b-a}}$. Συνήθης πρακτική είναι να επαναλαμβάνουμε τους υπολογισμούς με $2h_{\max}$ και $h_{\max}/2$ και να συγκρίνουμε τα αποτελέσματα για να διαπιστώσουμε αν πράγματι έχει επιτευχθεί η επιθυμητή ακρίβεια. Αν όχι, συνεχίζουμε με περαιτέρω υποδιπλασιασμό του $h_{\max}$.

6.3. Μέθοδος τραπεζίου

Στην μέθοδο αυτή θεωρούμε ότι το ολοκλήρωμα της συνάρτησης είναι προσεγγιστικά ίσο με το εμβαδόν του τραπεζίου που φαίνεται στο σχήμα 2, δηλαδή

$I_i(f) \approx \frac{x_{i+1}-x_i}{2}\left(f(x_i)+f(x_{i+1})\right) = T_i(f), i=1,2,...,n$. Επομένως ο σύνθετος κανόνας του τραπεζίου, $T(f)$, έχει ως εξής:

$$\boxed{T(f) = \sum_{i=1}^{n} T_i(f) = \sum_{i=1}^{n}\left\{\frac{x_{i+1}-x_i}{2}\left(f(x_i)+f(x_{i+1})\right)\right\}}$$

είτε στην περίπτωση του ομοιόμορφου διαμερισμού:

$$\boxed{T(f) = \frac{h}{2}\left(f(x_1)+2\sum_{i=2}^{n} f(x_i)+f(x_{n+1})\right), \quad h \equiv (b-a)/n}$$

Ανάλυση σφάλματος για την μέθοδο του τραπεζίου

Στην συνέχεια θα υπολογίσουμε το σφάλμα της μεθόδους αυτής, όπως προηγουμένως.

Θεωρούμε το ανάπτυγμα Taylor της συνάρτησης f γύρω από το σημείο $y_i = \frac{x_i + x_{i+1}}{2}$:

$$f(x) = f(y_i)+(x-y_i)f'(y_i)+\frac{(x-y_i)^2}{2}f''(y_i)+\frac{(x-y_i)^3}{6}f'''(y_i)+...$$

Για $x = x_i$ και $x = x_{i+1}$ η παραπάνω σχέση δίνει, αντίστοιχα:

$$f(x_i) = f(y_i)-\frac{h_i}{2}f'(y_i)+\frac{h_i^2}{8}f''(y_i)-\frac{h_i^3}{48}f'''(y_i)+\frac{h_i^4}{384}f^{(4)}(y_i)+...$$

$$f(x_{i+1}) = f(y_i)+\frac{h_i}{2}f'(y_i)+\frac{h_i^2}{8}f''(y_i)+\frac{h_i^3}{48}f'''(y_i)+\frac{h_i^4}{384}f^{(4)}(y_i)+...$$

όπου $h_i \equiv x_{i+1}-x_i$. Προσθέτουμε τις δύο τελευταίες σχέσεις και διαιρούμε δια δύο:

$$\frac{1}{2}\left(f(x_i)+f(x_{i+1})\right) = f(y_i)+\frac{h_i^2}{8}f''(y_i)+\frac{h_i^4}{384}f^{(4)}(y_i)+...$$

Ολοκληρώνουμε την σχέση αυτή στο διάστημα $[x_i, x_{i+1}]$ και παίρνουμε:

$$\underbrace{\frac{(x_{i+1}-x_i)}{2}\left(f(x_i)+f(x_{i+1})\right)}_{T_i(f)} = \underbrace{h_i f(y_i)}_{R_i(f)}+\frac{h_i^3}{8}f''(y_i)+\frac{h_i^5}{384}f^{(4)}(y_i)+...$$

Όμως έχουμε δείξει παραπάνω ότι $I_i(f)-R_i(f) = \frac{f''(y_i)}{24}h_i^3+\frac{f^{(4)}(y_i)}{1920}h_i^5+O\left(h_i^7\right)$

και επομένως οι δύο αυτές τελευταίες σχέσεις δίνουν

$$T_i(f) = I_i(f)-\left\{\frac{f''(y_i)}{24}h_i^3+\frac{f^{(4)}(y_i)}{1920}h_i^5+O\left(h_i^7\right)\right\}+\frac{h_i^3}{8}f''(y_i)+\frac{h_i^5}{384}f^{(4)}(y_i)+...$$

είτε, μετά από στοιχειώδεις πράξεις,

$$\varepsilon_i^T(f) \equiv I_i(f) - T_i(f) = -\frac{f''(y_i)}{12}h_i^3 - \frac{f^{(4)}(y_i)}{480}h_i^5 + O\left(h_i^7\right)...$$

Αθροίζοντας την παραπάνω σχέση σε όλα τα υποδιαστήματα παίρνουμε

$$\varepsilon^T(f) \equiv \sum_{i=1}^{n}\varepsilon_i^T(f) = -2\left(\frac{1}{24}\sum_{i=1}^{n}f''(y_i)h_i^3\right) - 4\left(\frac{1}{1920}\sum_{i=1}^{n}f^{(4)}(y_i)h_i^5\right) + O.Y.T. = -2E - 4F + O.Y.T.$$

Ακολουθώντας την ίδια διαδικασία με την μέθοδο του ορθογωνίου έχουμε:

$$\varepsilon^T(f) \equiv \sum_{i=1}^{n}\varepsilon_i^T(f) \approx -\frac{1}{12}\sum_{i=1}^{n}\left(f''(y_i)h_i^3\right) \Rightarrow \left|\varepsilon^T(f)\right| \approx \frac{1}{12}\left|\sum_{i=1}^{n}\left(f''(y_i)h_i^3\right)\right| \le \frac{1}{12}\sum_{i=1}^{n}\left|f''(y_i)h_i^3\right| \Rightarrow$$

$$\left|\varepsilon^T(f)\right| \le \frac{M}{12}\sum_{i=1}^{n}h_i^3, \quad M \equiv \max_{x\in[a,b]}\left|f''(x)\right|$$

Για ομοιόμορφο διαμερισμό ισχύει $h_i \equiv h = \frac{b-a}{n}$ και επομένως $\sum_{i=1}^{n}h_i^3 = nh^3$ και άρα

$$\left|\varepsilon^T(f)\right| \le \frac{M}{12}nh^3 = \frac{M}{12}n\frac{b-a}{n}\left(\frac{b-a}{n}\right)^2 = \frac{M}{12}(b-a)h^2 \Rightarrow \boxed{\left|\varepsilon^T\right| \le \frac{M}{12}(b-a)h^2}.$$

Όπως ήταν αναμενόμενο, δείχθηκε ότι το συνολικό σφάλμα της μεθόδου του τραπεζίου είναι τάξης 2, ενώ το τοπικό είναι τάξης 3. Η τελευταία σχέση μπορεί να χρησιμοποιηθεί για να βρεθεί το μέγιστο, θεωρητικό, h για ομοιόμορφο διαμερισμό το οποίο πρέπει να χρησιμοποιηθεί για να βρεθεί το ζητούμενο ολοκλήρωμα με μέγιστο σφάλμα Σ. Έτσι προκύπτει $h_{\max} = \sqrt{\frac{12\Sigma}{M(b-a)}}$ και βέβαια επειδή η σταθερά M ίσως να μην μπορεί να εκτιμηθεί παίρνουμε $h_{\max} \approx \sqrt{\frac{12\Sigma}{b-a}}$.

Δείξαμε λοιπόν ότι $\varepsilon^R(f) \approx E + F$ και $\varepsilon^T(f) \approx -2E - 4F$ (έχοντας αμελήσει τους λιγότερο σημαντικούς όρους) που σημαίνει ότι το σφάλμα του σύνθετου κανόνα του τραπεζίου είναι περίπου δύο φορές μεγαλύτερο από το σφάλμα του σύνθετου κανόνα του ορθογωνίου. Στην περίπτωση που $h_i << 1, F << E$ και επομένως από τις δύο αυτές προσεγγιστικές σχέσεις παίρνουμε $\varepsilon^R(f) \equiv I(f) - R(f) \approx E$ και $\varepsilon^T(f) \equiv I(f) - T(f) \approx -2E$ οπότε αν απαλείψουμε το $I(f)$ παίρνουμε $\boxed{E \approx \frac{1}{3}(T(f) - R(f))}$. Επομένως αν υπολογίσουμε το ζητούμενο ολοκλήρωμα με τις μεθόδους ορθογωνίου και τραπεζίου, μπορούμε να εκτιμήσουμε και την ποσότητα E.

6.4. Μέθοδος Simpson

Από τις προηγούμενες δύο μεθόδους μπορούμε να κατασκευάσουμε μία άλλη μέθοδο υψηλότερης τάξης ακρίβειας. Έχουμε:

$$\left.\begin{aligned}\varepsilon^R(f)\equiv I(f)-R(f)\approx E+F \Rightarrow 2I(f)-2R(f)\approx 2E+2F\\ \varepsilon^T(f)\equiv I(f)-T(f)\approx -2E-4F\end{aligned}\right\}\overset{+}{\Rightarrow} 3I(f)-2R(f)-T(f)\approx -2F$$

και επομένως από την τελευταία σχέση προκύπτει $I(f)\approx\frac{1}{3}\big(2R(f)+T(f)\big)-\frac{2}{3}F$ δηλαδή προκύπτει ένας νέος σύνθετος κανόνας ολοκλήρωσης, ο οποίος ονομάζεται κανόνας του Simpson, $S(f)=\frac{1}{3}\big(2R(f)+T(f)\big)$, με σφάλμα προσεγγιστικά ίσο με $-\frac{2}{3}F$. Αντικαθιστώντας τους κανόνες του ορθογωνίου και τραπεζίου παίρνουμε την εξής σχέση:

$$S(f)=\frac{1}{6}\sum_{i=1}^{n}h_i\left(f(x_i)+4f\left(\frac{x_i+x_{i+1}}{2}\right)+f(x_{i+1})\right)$$

Για το σφάλμα θα έχουμε

$$\varepsilon^S(f)\approx-\frac{2}{3}\frac{1}{1920}\sum_{i=1}^{n}\left(f^{(4)}(y_i)h_i^5\right)\Rightarrow\left|\varepsilon^S(f)\right|\approx\frac{1}{2280}\left|\sum_{i=1}^{n}\left(f^{(4)}(y_i)h_i^5\right)\right|\le\frac{1}{2280}\sum_{i=1}^{n}\left|f^{(4)}(y_i)h_i^5\right|\Rightarrow$$

$$\left|\varepsilon^S(f)\right|\le\frac{M}{2280}\sum_{i=1}^{n}h_i^5,\quad M\equiv\max_{x\in[a,b]}\left|f^{(4)}(x)\right|.$$

Αν τώρα μετρήσουμε και τα σημεία $y_i=\big(x_i+x_{i+1}\big)/2,\ \ i=1,2,...,n,$ και επιπλέον θεωρήσουμε ομοιόμορφο διαμερισμό και n άρτιο η παραπάνω σχέση ανάγεται στην:

$$\boxed{S(f)=\frac{h}{3}\left(f(x_0)+2\sum_{i=2}^{n,2}f(x_i)+4\sum_{i=1}^{n-1,2}f(x_i)+f(x_n)\right),\quad n\quad \text{αρτιος}}$$

(έτσι αποφεύγουμε να ορίζουμε τα μέσα των υποδιαστημάτων) ενώ για το ολικό σφάλμα της μεθόδου μπορεί να αποδειχτεί ότι $\boxed{\left|\varepsilon^S(f)\right|\le\frac{M(b-a)}{180}h^4}$. Επομένως ο κανόνας του Simpson είναι 5ης τάξης ακρίβειας τοπικά και 4ης τάξης ακρίβειας συνολικά. Όπως και στις δύο προηγούμενες μεθόδους, αν επιδιώκουμε το μέγιστο σφάλμα στον υπολογισμό του ολοκληρώματος να είναι Σ τότε από την τελευταία σχέση παίρνουμε $h_{\max}=\left(\frac{180\Sigma}{M(b-a)}\right)^{1/4}$. Αν η σταθερά M δεν μπορεί να υπολογισθεί ή να εκτιμηθεί

θεωρούμε $h_{\max} \approx \left(\frac{180\Sigma}{b-a}\right)^{1/4}$, υπολογίζουμε το ζητούμενο ολοκλήρωμα και επαναλαμβάνουμε την διαδικασία με $2h_{\max}$ και $h_{\max}/2$ και ελέγχουμε κατά πόσο τα αποτελέσματα έχουν συγκλίνει με την επιθυμητή ακρίβεια, διαφορετικά συνεχίζουμε με περαιτέρω υποδιπλασιασμό μέχρι σύγκλισης του αποτελέσματος.

Στο σημείο αυτό να σημειωθεί πως, συνήθως, αν μία μέθοδος ολοκλήρωσης έχει τοπική τάξη ακρίβειας $p+1$ τότε η ολική τάξη ακρίβειάς της σε όλο το διάστημα ενδιαφέροντος είναι p. Τέλος, οι εκτιμήσεις της μορφής $\left|\varepsilon^R\right| \le \frac{M}{24}(b-a)h^2$, $\left|\varepsilon^T\right| \le \frac{M}{12}(b-a)h^2$, $\left|\varepsilon^S\right| \le \frac{M}{180}(b-a)h^4$ δεν είναι χρήσιμες μόνο διότι μας δίνουν το μέγιστο σφάλμα της μεθόδου αλλά και γιατί μας προσδιορίζουν τον *ρυθμό σύγκλισης της μεθόδου*. Πράγματι, βλέπουμε ότι σε όλες τις περιπτώσεις που μελετήθηκαν το μέγιστο συνολικό σφάλμα είναι της μορφής $\varepsilon_{\max,h} = ch^p$ όπου p είναι η ολική τάξη της μεθόδου και c είναι μία σταθερά που εξαρτάται από τα a, b, M αλλά όχι από το h. Αν τώρα υποδιπλασιάσουμε το μήκος των υποδιαστημάτων, δηλαδή θεωρήσουμε διαμερισμό με $h/2$ τότε το μέγιστο σφάλμα θα γίνει $\varepsilon_{\max,h/2} = c(h/2)^p$ όπου βέβαια η σταθερά c παραμένει η ίδια. Επομένως ο λόγος των μέγιστων σφαλμάτων θα είναι $\frac{\varepsilon_{\max,h/2}}{\varepsilon_{\max,h}} = \frac{c(h/2)^p}{ch^p} = 2^{-p}$. Για τον κανόνα του ορθογωνίου και του τραπεζίου, οι οποίοι είναι δεύτερης τάξης, $p=2$, θα έχουμε λόγο σφαλμάτων $2^{-2} = 1/4$ που σημαίνει ότι ο υποδιαπλασιασμός του μήκους h υποτετραπλασιάζει το μέγιστο σφάλμα. Αντίστοιχα για τον κανόνα του Simpson, για τον οποίο $p=4$, ο λόγος των σφαλμάτων θα είναι $2^{-4} = 1/16$, δηλαδή ο υποδιπλασιασμός του μήκους h μειώνει το μέγιστο σφάλμα 16 φορές.

Παράδειγμα 1ο: Υπολογίστε αριθμητικά το ολοκλήρωμα $I = \int_0^1 e^{-0.4x} dx$ με τους σύνθετους κανόνες του ορθογωνίου, τραπεζίου και Simpson. Για τους υπολογισμούς θεωρείστε μόνο δύο υποδιαστήματα, δηλαδή $n=2$. Βρείτε πόσο είναι προσεγγιστικά το σφάλμα για τους κανόνες ορθογωνίου και τραπεζίου.

Λύση: Εφόσον $n=2$ άρα $h=\dfrac{b-a}{n}=\dfrac{1-0}{2}=0.5$. Έχουμε:

(α) κανόνας ορθογωνίου:

$$R=h\sum_{i=1}^{2} f\left(\frac{x_{i+1}+x_i}{2}\right)=\frac{1}{2}\left\{f\left(\frac{x_2+x_1}{2}\right)+f\left(\frac{x_3+x_2}{2}\right)\right\}=\frac{1}{2}\left\{f\left(0.25\right)+f\left(0.75\right)\right\}$$

όπου $f\left(0.25\right)=e^{-0.4\times 0.25}\approx 0.904837$ και $f\left(0.75\right)=e^{-0.4\times 0.75}\approx 0.740818$ οπότε προκύπτει

$R=0.822828$

(β) κανόνας τραπεζίου:

$$T=\frac{h}{2}\left(f\left(x_1\right)+2\sum_{i=2}^{n} f\left(x_i\right)+f\left(x_{n+1}\right)\right)=\frac{1/2}{2}\left(f\left(x_1\right)+2f\left(x_2\right)+f\left(x_3\right)\right)$$

όπου $f\left(x_1\right)=e^{-0.4x0.0}=1$, $f\left(x_2\right)=e^{-0.4x0.5}=0.813731$ και $f\left(x_3\right)=e^{-0.4x1.0}=0.67032$ οπότε

$T=0.826945$

(γ) κανόνας Simpson:

$$S=\frac{1}{3}\left(2R+T\right)=0.8242$$

Για την εκτίμηση του σφάλματος έχουμε:

$$E\approx\frac{1}{3}\left(T-R\right)=\frac{1}{3}\left(0.826945-0.822828\right)=0.00137252$$

Επομένως $\varepsilon^R\approx E=0.00137252$ και $\varepsilon^T\approx -2E=-0.00274504$

Έλεγχος:

Η ακριβής τιμή του ολοκληρώματος είναι $I=\int_0^1 e^{-0.4x}dx=-\dfrac{1}{0.4}e^{-0.4x}\Big|_{x=0}^{x=1}=0.8242$

Άρα $\varepsilon^R\approx I-R=0.001372$ και $\varepsilon^T\approx I-T=-0.00274539$ ενώ $\varepsilon^S\approx I-S=0$. Τα αποτελέσματα αυτά δείχνουν ότι ακόμα και με δύο υποδιαστήματα έχουμε υπολογίσει το ζητούμενο ολοκλήρωμα με πολύ καλή ακρίβεια.

Παράδειγμα 2ο: Υπολογίστε εκ' των προτέρων το ελάχιστο πλήθος των υποδιαστημάτων, $n_{\min}^R$, $n_{\min}^T$ και $n_{\min}^S$, που απαιτείται για να επιτευχθεί ακρίβεια 5 δεκαδικών ψηφίων με τις μεθόδους ορθογωνίου, τραπεζίου και Simpson, αντίστοιχα, για το ολοκλήρωμα $I=\int_0^1 e^{-x^2}dx$.

Λύση: Το ολικό σφάλμα των μεθόδων αυτών δίνεται από:

$$\left|\varepsilon^R\right| \le \frac{M}{24}(b-a)\left(h^R\right)^2 < \Sigma \Rightarrow h^R_{\max} = \sqrt{\frac{24\Sigma}{(b-a)M}}, \quad M \equiv \max_{x\in[0,1]}\left|f''(x)\right| \quad (*)$$

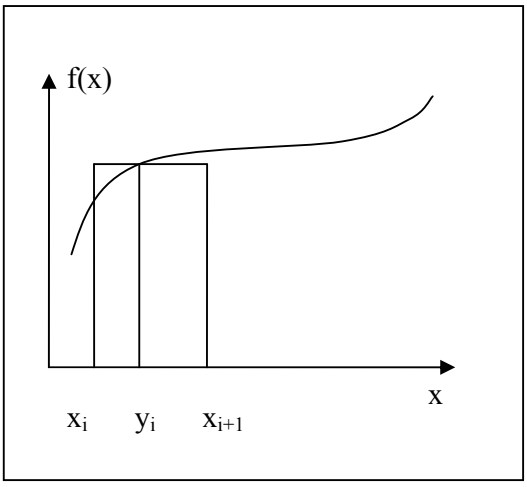
$$\left|\varepsilon^T\right| \le \frac{M}{12}(b-a)\left(h^T\right)^2 < \Sigma \Rightarrow h^T_{\max} = \sqrt{\frac{12\Sigma}{(b-a)M}}, \quad M \equiv \max_{x\in[0,1]}\left|f''(x)\right| \quad (**)$$

$$\left|\varepsilon^S\right| \le \frac{\hat{M}}{180}(b-a)\left(h^S\right)^4 < \Sigma \Rightarrow h^S_{\max} = \sqrt[4]{\frac{180\Sigma}{(b-a)\hat{M}}}, \quad \hat{M} \equiv \max_{x\in[0,1]}\left|f^{(4)}(x)\right| \quad (***)$$

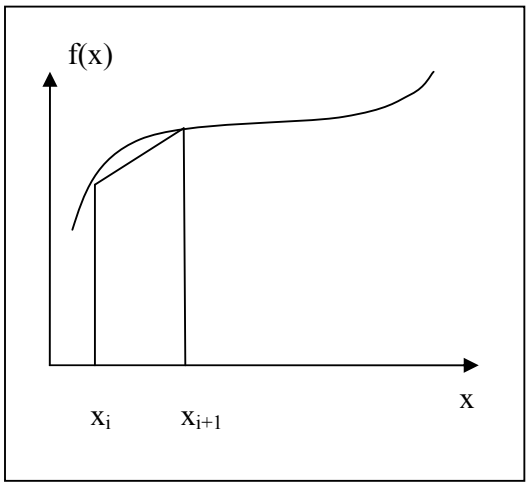
όπου $a=0$, $b=1$ και επειδή ζητούμε ακρίβεια 5 δεκαδικών ψηφίων άρα $\Sigma = 5\times 10^{-6}$. Επιπλέον έχουμε $f(x)=e^{-x^2} \Rightarrow f'(x) = -2xe^{-x^2} \Rightarrow f''(x) = 2\left(2x^2-1\right)e^{-x^2}$ οπότε βρίσκουμε ότι $M = \max_{x\in[0,1]}\left|f''(x)\right| = 2$. Αντικαθιστώντας στην (*) και (**) προκύπτει $h^R_{\max} = \frac{b-a}{n^R_{\min}} \approx 0.007746 \Rightarrow n^R_{\min} \approx 130$ και $h^T_{\max} = \frac{b-a}{n^T_{\min}} \approx 0.005477 \Rightarrow n^T_{\min} \approx 183$, αντίστοιχα.

Επιπλέον μπορούμε να βρούμε ότι $\hat{M} \equiv \max_{x\in[0,1]}\left|f^{(4)}(x)\right| = 12$ οπότε αντικαθιστώντας στην (***) βρίσκουμε $h^S_{\max} = \frac{b-a}{n^S_{\min}} \approx 0.09306 \Rightarrow n^S_{\min} \approx 11$.

Συγκεντρωτικά απαιτούνται 183, 130 και 11 υποδιαστήματα τουλάχιστον για να επιτευχθεί ακρίβεια πέντε δεκαδικών ψηφίων με τους σύνθετους κανόνες τραπεζίου, ορθογωνίου και Simpson, αντίστοιχα.



Σχήμα 1ο



Σχήμα 2ο

Βιβλιογραφία

(1) *Εισαγωγή στην Αριθμητική Ανάλυση,* Γ.Δ. Ακρίβης & Β.Α. Δουγαλής, Πανεπιστημιακές Εκδόσεις Κρήτης, 5η αναθεωρημένη έκδοση, 2006.

(2) *Αριθμητική ανάλυση*, Γ.Σ. Σοφιανός & Ε. Θ. Τυχόπουλος, Εκδόσεις Σταμούλης, 2005.

(3) *Αριθμητικές μέθοδοι και προγράμματα για μαθηματικούς υπολογισμούς*, G.E. Forsythe, M.A. Malcolm & C.B. Moler, μετάφραση από τους Γ.Δ. Ακρίβη & Β.Α. Δουγαλή, Πανεπιστημιακές Εκδόσεις Κρήτης, 1997.

(4) *Numerical methods for scientists and engineers*, R.W.Hamming, 2nd ed., Dover, 1962.

(5) *Theory and applications of numerical analysis*, G.M. Plilips & PJ Taylor, 2nd ed., 1996.

(6) *Introduction to numerical analysis*, F.B. Hildebrand, Dover, 1956.

(7) *A first course in numerical analysis,* A. Ralston & P. Rabinowitz, 2nd ed., Dover, 1965.

Παράρτημα Π.1
Στοιχεία Γραμμικής Άλγεβρας

Πίνακες, διανύσματα

Ένας $m \times n$ πίνακας $\underline{\underline{A}}$, με συνολικό πλήθος στοιχείων $m \times n$, πραγματικούς (ή μιγαδικούς) αριθμούς a_{ij}, με m γραμμές και n στήλες, γράφεται στην μορφή:

$$\underline{\underline{A}} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

Το στοιχείο a_{ij} βρίσκεται στην i-γραμμή και j-στήλη του πίνακα $\underline{\underline{A}}$ ο οποίος είναι τάξεως ή διάστασης $m \times n$. Εάν $m = n$, τότε ο πίνακας $\underline{\underline{A}}$ ονομάζεται ***τετραγωνικός***, τάξεως n. Τα στοιχεία του τετραγωνικού πίνακα a_{ii}, $i = 1, 2, \dots, n$, ονομάζονται *διαγώνια στοιχεία.*. (Πολλές φορές, για ευκολία, θα χρησιμοποιούμε κόμμα μεταξύ των δεικτών στα στοιχεία του πίνακα, για παράδειγμα $a_{i,i}$ ή $a_{i,j}$)

Διάνυσμα στήλη καλείται ο πίνακας που έχει μόνο μία στήλη. Για παράδειγμα,

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

είναι τάξεως $m \times 1$ και καλείται διάνυσμα στήλη, διάστασης m. Συμβολίζουμε με $\mathbb{R}^m$ το σύνολο όλων των m- διάστατων διανυσμάτων στήλης πινάκων με στοιχεία πραγματικούς αριθμούς. ($\mathbb{C}^m$ ορίζεται το αντίστοιχο σύνολο των διανυσμάτων με στοιχεία μιγαδικούς αριθμούς.) $\mathbb{R}^n$ συνήθως αναφέρεται ως ο n-διαστάσεων πραγματικός χώρος.

Διάνυσμα γραμμή είναι ο πίνακας που περιέχει μόνο μία γραμμή. Για παράδειγμα,

$$\underline{x}^T = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}$$

είναι τάξεως $1 \times n$ και καλείται ως διάνυσμα γραμμή διαστάσεως n. Συμβολίζεται με $\underline{x}^T$ ο ανάστροφος του διανύσματος $\underline{x} \in \mathbb{R}^n$.

Δύο πίνακες καλούνται ίσοι εάν έχουν την ίδια τάξη και τα αντίστοιχα στοιχεία τους είναι ίσα, δηλαδή

$\underline{\underline{A}} = \underline{\underline{B}}$ αν $a_{ij} = b_{ij}$ $\forall$ i, j

Δύο πίνακες της ίδιας διάστασης προστίθενται ή αφαιρούνται εάν προστεθούν ή αφαιρεθούν τα αντίστοιχα στοιχεία τους, δηλαδή

$\underline{\underline{C}} = \underline{\underline{A}} \pm \underline{\underline{B}}$ αν $c_{ij} = a_{ij} \pm b_{ij} \ \forall \ i, j$

Επίσης, ισχύει η ***αντιμεταθετική*** καθώς και η ***προσεταιριστική*** ιδιότητα της πρόσθεσης και στους πίνακες, δηλαδή

$\underline{\underline{A}} + \underline{\underline{B}} = \underline{\underline{B}} + \underline{\underline{A}}$

$\underline{\underline{A}} + \left(\underline{\underline{B}} + \underline{\underline{C}} \right) = \left(\underline{\underline{A}} + \underline{\underline{B}} \right) + \underline{\underline{C}}$

Ο μηδενικός πίνακας, τάξεως $m \times n$, συμβολίζεται με $\underline{\underline{0}}$ και κάθε στοιχείο του είναι μηδενικό, έτσι ώστε

$\underline{\underline{A}} + \underline{\underline{0}} = \underline{\underline{A}}$

Αν λ είναι πραγματικός (ή μιγαδικός) αριθμός, τότε ο πίνακας $\lambda \underline{\underline{A}}$ είναι ένας πίνακας με στοιχεία λa_{ij}, δηλαδή κάθε στοιχείο του πίνακα $\underline{\underline{A}}$ είναι πολλαπλασιασμένο με λ. Επίσης, ισχύει ότι:

$-\underline{\underline{A}} = (-1) \underline{\underline{A}}$

$\underline{\underline{A}} + \left(-\underline{\underline{A}} \right) = \underline{\underline{0}}$

$0 \underline{\underline{A}} = \underline{\underline{0}}$

Ο πολλαπλασιασμός των πινάκων είναι λίγο πιο πολύπλοκη πράξη. Πρώτα ορίζουμε το γινόμενο μεταξύ διανυσμάτων. Αν $\underline{x}, \underline{y} \in \mathbb{R}^n$, ορίζεται ως γινόμενο ενός διανύσματος γραμμής και ενός διανύσματος στήλης η πράξη

$$\underline{y}^T \cdot \underline{x} = y_1 x_1 + y_2 x_2 + \ldots + y_n x_n = \sum_{i=1}^{n} y_i x_i$$

Το αποτέλεσμα της παραπάνω πράξης είναι ένας αριθμός και ορίζεται ως το εσωτερικό γινόμενο των $\underline{y}$ και $\underline{x}$.

Αν $\underline{\underline{A}}$ είναι ενας $m \times n$ πίνακας και $\underline{x} \in \mathbb{R}^n$, ορίζεται ότι

$$\underline{\underline{A}} \cdot \underline{x} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11} x_1 + a_{12} x_2 + \ldots + a_{1n} x_n \\ a_{21} x_1 + a_{22} x_2 + \ldots + a_{2n} x_n \\ \vdots \\ a_{m1} x_1 + a_{m2} x_2 + \ldots + a_{mn} x_n \end{bmatrix}$$

Το αποτέλεσμα είναι ένα διάνυσμα $\underline{z} \in \mathbb{R}^n$ του οποίου το $i-$στοιχείο είναι

$$z_i = \sum_{j=1}^{n} a_{ij} x_j = \left[i - \text{γραμμη του } \underline{\underline{A}} \right] \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

Εάν ο $\underline{\underline{A}}$ είναι ένας $m \times n$ πίνακας και ο $\underline{\underline{B}}$ ένας $n \times p$, τότε ορίζεται το γινόμενο $\underline{\underline{C}} = \underline{\underline{A}} \cdot \underline{\underline{B}}$ να είναι ένας πίνακας $m \times p$ με το (i, j) στοιχείο να ορίζεται ως

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} = \left[i - \text{ γραμμη του } \underline{\underline{A}} \right] \cdot \begin{bmatrix} j - \\ \text{στηλη} \\ \text{του } \underline{\underline{B}} \end{bmatrix}$$

Να σημειωθεί ότι ο αριθμός των στοιχείων σε μια γραμμή του πίνακα $\underline{\underline{A}}$ θα πρέπει να είναι ίδιος με τον αριθμό των στοιχείων σε μια στήλη του $\underline{\underline{B}}$ (αριθμός στηλών $\underline{\underline{A}}$ = αριθμός γραμμών $\underline{\underline{B}}$). Οι κανόνες πολλαπλασιασμού για διανύσματα αποτελούν ειδικές περιπτώσεις αυτού του γενικού κανόνα.

Στον πολλαπλασιασμό πινάκων δεν είναι απαραίτητο να ισχύει η αντιμεταθετική ιδιότητα ακόμα και αν οι $\underline{\underline{A}} \cdot \underline{\underline{B}}$ και $\underline{\underline{B}} \cdot \underline{\underline{A}}$ είναι ορισμένοι. Έτσι, γενικά ισχύει ότι

$$\underline{\underline{A}} \cdot \underline{\underline{B}} \neq \underline{\underline{B}} \cdot \underline{\underline{A}}$$

Για αυτό το λόγο η τάξη μεταξύ δύο πινάκων που πολλαπλασιάζονται είναι σημαντική.

Η προσεταιριστική ιδιότητα ισχύει και στον πολλαπλασιασμό πινάκων. Συνεπώς, υπό τις κατάλληλες συνθήκες για τους πίνακες $\underline{\underline{A}}$, $\underline{\underline{B}}$ και $\underline{\underline{C}}$ ισχύει

$$\underline{\underline{A}} \cdot \left(\underline{\underline{B}} \cdot \underline{\underline{C}} \right) = \left(\underline{\underline{A}} \cdot \underline{\underline{B}} \right) \cdot \underline{\underline{C}}$$

Επιπρόσθετα, ισχύει και η επιμεριστική ιδιότητα υπό τις κατάλληλες συνθήκες. Έτσι,

$$\underline{\underline{A}} \cdot \left(\underline{\underline{B}} + \underline{\underline{C}} \right) = \underline{\underline{A}} \cdot \underline{\underline{B}} + \underline{\underline{A}} \cdot \underline{\underline{C}}$$

και

$$\left(\underline{\underline{A}} + \underline{\underline{B}} \right) \cdot \underline{\underline{C}} = \underline{\underline{A}} \cdot \underline{\underline{C}} + \underline{\underline{B}} \cdot \underline{\underline{C}}$$

Ορίζεται ο $n \times n$ μοναδιαίος πίνακας $\underline{\underline{I}}$ ως

$$\underline{\underline{I}} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Εάν ο πίνακας $\underline{\underline{A}}$ είναι ένας οποιοσδήποτε $m \times n$ πίνακας και ο $\underline{\underline{I}}$ είναι ένας $n \times n$ μοναδιαίος πίνακας, τότε

$\underline{\underline{A}} \cdot \underline{\underline{I}} = \underline{\underline{A}}$

και αν ο $\underline{\underline{I}}$ είναι ένας $m \times m$ μοναδιαίος πίνακας, τότε ισχύει

$\underline{\underline{I}} \cdot \underline{\underline{A}} = \underline{\underline{A}}$

Κάτω από τις κατάλληλες συνθήκες ένας τετραγωνικός πίνακας $\underline{\underline{A}}$ έχει και τον αντίστροφό του ο οποίος συμβολίζεται ως $\underline{\underline{A}}^{-1}$ και ο ικανοποιεί τις συνθήκες

$\underline{\underline{A}} \cdot \underline{\underline{A}}^{-1} = \underline{\underline{I}}$ και $\underline{\underline{A}}^{-1} \cdot \underline{\underline{A}} = \underline{\underline{I}}$

Εάν ένα τέτοιος αντίστροφος $\underline{\underline{A}}^{-1}$ υπάρχει, τότε ο πίνακας $\underline{\underline{A}}$ είναι μη μηδενικός. Το αντίστροφο γινόμενο μεταξύ δύο τετραγωνικών πινάκων $\underline{\underline{A}}$ και $\underline{\underline{B}}$ ικανοποιεί την συνθήκη

$\left(\underline{\underline{A}} \cdot \underline{\underline{B}}\right)^{-1} = \underline{\underline{B}}^{-1} \cdot \underline{\underline{A}}^{-1}$

Κάποιες φορές είναι πιο εύκολο οι πίνακες να γράφονται στην παρακάτω μορφή. Έτσι, ένας $m \times n$ πίνακας $\underline{\underline{A}}$ μπορεί να γραφεί ως

$$\underline{\underline{A}} = \begin{bmatrix} \underline{\underline{A}}_{11} & \underline{\underline{A}}_{12} \\ \underline{\underline{A}}_{21} & \underline{\underline{A}}_{22} \end{bmatrix}$$

όπου ο $\underline{\underline{A}}_{11}$ είναι ένας $r \times p$ πίνακας, ο $\underline{\underline{A}}_{12}$ είναι $r \times (n-p)$, ο $\underline{\underline{A}}_{21}$ είναι $(m-r) \times p$ και ο $\underline{\underline{A}}_{22}$ είναι τάξεως $(m-r) \times (n-p)$. Οι περιορισμοί στους πίνακες είναι απαραίτητοι, έτσι για παράδειγμα , οι πίνακες $\underline{\underline{A}}_{11}$ και $\underline{\underline{A}}_{21}$ έχουν τον ίδιο αριθμό στηλών. Οι πίνακες $\underline{\underline{A}}_{11}$, $\underline{\underline{A}}_{12}$, $\underline{\underline{A}}_{21}$ και $\underline{\underline{A}}_{22}$ καλούνται υποπίνακες του $\underline{\underline{A}}$.

Είναι εύκολο να δει κανείς ότι εαν

$$\underline{\underline{B}} = \begin{bmatrix} \underline{\underline{B}}_{11} & \underline{\underline{B}}_{12} \\ \underline{\underline{B}}_{21} & \underline{\underline{B}}_{22} \end{bmatrix}$$

και ο $\underline{\underline{B}}$ χωρίζεται σε υποπίνακες όπως ακριβώς και ο $\underline{\underline{A}}$, τότε ισχύει

$$\underline{\underline{A}} + \underline{\underline{B}} = \begin{bmatrix} \underline{\underline{A}}_{11} + \underline{\underline{B}}_{11} & \underline{\underline{A}}_{12} + \underline{\underline{B}}_{12} \\ \underline{\underline{A}}_{21} + \underline{\underline{B}}_{21} & \underline{\underline{A}}_{22} + \underline{\underline{B}}_{22} \end{bmatrix}$$

Δεν είναι τόσο προφανές, ωστόσο, εάν ο πίνακας $\underline{\underline{B}}$ και οι υποπίνακές του είναι κατάλληλα ορισμένοι, ισχύει ότι:

$$\underline{\underline{A}}\cdot\underline{\underline{B}}=\begin{bmatrix}\left(\underline{\underline{A}}_{11}\cdot\underline{\underline{B}}_{11}+\underline{\underline{A}}_{12}\cdot\underline{\underline{B}}_{21}\right) & \left(\underline{\underline{A}}_{11}\cdot\underline{\underline{B}}_{12}+\underline{\underline{A}}_{12}\cdot\underline{\underline{B}}_{22}\right)\\ \left(\underline{\underline{A}}_{21}\cdot\underline{\underline{B}}_{11}+\underline{\underline{A}}_{22}\cdot\underline{\underline{B}}_{21}\right) & \left(\underline{\underline{A}}_{21}\cdot\underline{\underline{B}}_{12}+\underline{\underline{A}}_{22}\cdot\underline{\underline{B}}_{22}\right)\end{bmatrix}$$

Δοθέντος οποιουδήποτε $m\times n$ πίνακα, καλείται ο $n\times m$ πίνακας, ο οποίος λαμβάνεται με αντιμετάθεση των γραμμών και στηλών του $\underline{\underline{A}}$, ανάστροφος του $\underline{\underline{A}}$ και συμβολίζεται $\underline{\underline{A}}^T$. Έτσι,

$$\underline{\underline{A}}^T=\begin{bmatrix}a_{11} & a_{21} & \dots & a_{m1}\\ a_{12} & a_{22} & \dots & a_{m2}\\ \vdots & & & \vdots\\ a_{1n} & a_{2n} & \dots & a_{mn}\end{bmatrix}$$

και a_{ij} αποτελούν τα στοιχεία της j γραμμής και i στήλης του $\underline{\underline{A}}^T$.

Για δύο πίνακες $\underline{\underline{A}}$ και $\underline{\underline{B}}$ των οποίων τα γινόμενα είναι ορισμένα, βρίσκεται ότι

$$\left(\underline{\underline{A}}\cdot\underline{\underline{B}}\right)^T=\underline{\underline{B}}^T\cdot\underline{\underline{A}}^T$$

Ένας $n\times n$ πίνακας $\underline{\underline{A}}$ είναι ***συμμετρικός*** εάν $\underline{\underline{A}}^T=\underline{\underline{A}}$, έτσι ώστε $a_{ij}=a_{ji}$ για όλα τα i και j.

Ορίζουσες

Ορισμός: Η ορίζουσα ενός $n\times n$ πίνακα $\underline{\underline{A}}$ είναι ορισμένη αναδρομικά ως ακολούθως

i. Εάν $\underline{\underline{A}}=\left[a_{1,1}\right]$, ο οποίος είναι ένας 1×1 πίνακας, τότε η ορίζουσα του $\underline{\underline{A}}$ γράφεται ως $\det\left(\underline{\underline{A}}\right)$ και είναι $a_{1,1}$.

ii. Εάν ο $\underline{\underline{A}}$ είναι ένας $n\times n$, $n\geq2$, πίνακας, ορίζεται ο ελάσσων πίνακας (minor) A_{ij} του $\underline{\underline{A}}$ να είναι η ορίζουσα του $(n-1)\times(n-1)$ πίνακα ο οποίος λαμβάνεται με απαλοιφή της $i-$γραμμής και $j-$στήλης του $\underline{\underline{A}}$. Έτσι, ορίζεται

$$\det\left(\underline{\underline{A}}\right)=\sum_{i=1}^{n}(-1)^{i+j}a_{i,j}A_{ij},\qquad\forall j,\quad 1\leq j\leq n$$

ή

$$\det\left(\underline{\underline{A}}\right)=\sum_{j=1}^{n}(-1)^{i+j}a_{ij}A_{ij},\qquad\forall i,\quad 1\leq i\leq n$$

Ο πολλαπλασιασμός ενός τετραγωνικού πίνακα $n \times n$, $\underline{\underline{A}}$, με μία μη-μηδενική σταθερά λ έχει την εξής επίπτωση στην ορίζουσά του, $\boxed{\det\left(\lambda \underline{\underline{A}}\right) = \lambda^n \det\left(\underline{\underline{A}}\right)}$.

Έστω δύο τετραγωνικοί πίνακες $\underline{\underline{A}}, \underline{\underline{B}}$. Τότε ισχύει $\boxed{\det\left(\underline{\underline{A}} \cdot \underline{\underline{B}}\right) = \det\left(\underline{\underline{A}}\right) \det\left(\underline{\underline{B}}\right)}$

Η ορίζουσα ενός διαγώνιου πίνακα, $\underline{\underline{L}}$ (κάτω τριγωνικού) ή $\underline{\underline{U}}$ (άνω τριγωνικού), είναι ίση με το γινόμενο των διαγωνίων στοιχείων του, δηλαδή $\det\left(\underline{\underline{L}}\right) = \prod_{i=1}^{n} \ell_{i,i}$ ή $\det\left(\underline{\underline{U}}\right) = \prod_{i=1}^{n} u_{i,i}$, αντίστοιχα. Όμοια, η ορίζουσα ενός διαγώνιου πίνακα $\underline{\underline{D}}$ είναι ίση με το γινόμενο των στοιχείων του $\det\left(\underline{\underline{D}}\right) = \prod_{i=1}^{n} d_{i,i}$.

Ιδιοτιμές και ιδιοδιανύσματα

Έστω τετραγωνικός πίνακας $\underline{\underline{A}} \in \mathbb{R}^{n,n}$ (ή $\underline{\underline{A}} \in \mathbb{C}^{n,n}$). Εάν υπάρχουν $\lambda \in \mathbb{C}$ και $\underline{x} \in \mathbb{C}^n$ με $\underline{x} \neq \underline{0}$ τέτοια ώστε

$\underline{\underline{A}} \cdot \underline{x} = \lambda \underline{x}$ (*)

τότε η σταθερά λ ονομάζεται *ιδιοτιμή* και το διάνυσμα $\underline{x}$ αντίστοιχο *ιδιοδιάνυσμα* του πίνακα $\underline{\underline{A}}$.

Ο προσδιορισμός των ιδιοτιμών/ιδιοδιανυσμάτων ενός πίνακα γίνεται ως εξής. Η σχέση (*) μπορεί ισοδύναμα να γραφεί ως:

$\underline{\underline{A}} \cdot \underline{x} = \lambda \underline{x} = \lambda \underline{\underline{I}} \cdot \underline{x} \Rightarrow \underline{\underline{A}} \cdot \underline{x} - \lambda \underline{\underline{I}} \cdot \underline{x} = \underline{0} \Rightarrow \left(\underline{\underline{A}} - \lambda \underline{\underline{I}}\right) \cdot \underline{x} = \underline{0}$

Η παραπάνω σχέση αποτελεί ένα ομογενές σύστημα «n» γραμμικών εξισώσεων με «n» αγνώστους. Για να έχει μη-μηδενική λύση θα πρέπει η ορίζουσα του πίνακα που πολλαπλασιάζει το άγνωστο διάνυσμα $\underline{x}$ να είναι μηδενική, δηλαδή:

$\det\left(\underline{\underline{A}} - \lambda \underline{\underline{I}}\right) = 0$ (**)

είτε

$$\det\begin{bmatrix} a_{11}-\lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22}-\lambda & \dots & a_{2n} \\ . & & & \\ . & & & \\ . & & & \\ a_{n1} & a_{n2} & \dots & a_{nn}-\lambda \end{bmatrix} = 0$$

Όπως είναι προφανές η παραπάνω εξίσωση είναι μία πολυωνυμική εξίσωση «n» βαθμού και επομένως έχει «n» ρίζες, $\lambda_i, i = 1,2,3,...,n$, οι οποίες είναι οι ιδιοτιμές του πίνακα. Επειδή οι ρίζες ενός πολυωνύμου είναι μοναδικά καθορισμένες έτσι και οι ιδιοτιμές του πίνακα είναι μοναδικές. Το πολυώνυμο που προκύπτει ονομάζεται χαρακτηριστικό πολυώνυμο ενώ η εξίσωση (**) ονομάζεται χαρακτηριστική εξίσωση του πίνακα. Το σύνολο όλων των ιδιοτιμών ενός πίνακα ονομάζεται **φάσμα** του πίνακα $\underline{\underline{A}}$ και συμβολίζεται ως:

$$\sigma\left(\underline{\underline{A}}\right) = \left\{\lambda \in \mathbb{C} \,\|\, \det\left(\underline{\underline{A}} - \lambda \underline{\underline{I}}\right) = 0\right\}$$

Η μέγιστη κατά απόλυτη τιμή ιδιοτιμή του $\underline{\underline{A}}$ ονομάζεται **φασματική ακτίνα** του πίνακα και συμβολίζεται ως:

$$\rho\left(\underline{\underline{A}}\right) = \max_{1\le i\le n}\left\{\left|\lambda_i\right| \,\|\, \lambda_i \in \sigma\left(\underline{\underline{A}}\right)\right\}$$

Η φασματική ακτίνα ενός πίνακα παίζει σημαντικό ρόλο στην σύγκλιση των αριθμητικών μεθόδων επίλυσης συστημάτων γραμμικών εξισώσεων.

Αντίθετα με τις ιδιοτιμές, τα ιδιοδιανύσματα δεν είναι μοναδικά καθορισμένα όπως φαίνεται από την εξίσωση (*). Πράγματι, αν η (*) πολλαπλασιαστεί με μία μη-μηδενική σταθερά μ, $\mu \neq 0$, τότε προκύπτει

$$\underline{\underline{A}} \cdot \left(\mu \underline{x}\right) = \lambda\left(\mu \underline{x}\right)$$

Η σχέση αυτή δείχνει ότι η σταθερά λ είναι ιδιοτιμή του πίνακα $\underline{\underline{A}}$ και το διάνυσμα $\mu\underline{x}$ είναι το αντίστοιχο ιδιοδιάνυσμα. Εάν μάλιστα η σταθερά μ επιλεγεί ως $\mu = 1 \Big/ \sqrt{\sum_{i=1}^{n} x_i^2}$ τότε το μέτρο του ιδιοδιανύσματος είναι ακριβώς μονάδα, είναι δηλαδή κανονικοποιημένο. Τα ιδιοδιανύσματα ενός πίνακα μπορούν πάντα, με την διαδικασία Gram-Schmidt, να είναι ορθογώνια μεταξύ τους και επιπλέον το μέτρο του καθενός να είναι ακριβώς μονάδα. Στην περίπτωση αυτή λέμε ότι τα ιδιοδιανύσματα είναι ορθοκανονικοποιημένα. Έτσι αν $\lambda_i, i = 1,2,3,...,n$ είναι οι ιδιοτιμές ενός τετραγωνικού

πίνακα και $\underline{x}_i, i=1,2,3,...,n$ είναι τα αντίστοιχα ορθοκανονικοποιημένα ιδιοδιανύσματα τότε ισχύει:

$$\underline{x}_i^T \cdot \underline{x}_j = \delta_{ij} = \begin{cases} 1 \text{ αν } i = j \\ 0 \text{ αν } i \neq j \end{cases}$$

Σημειώνουμε, χωρίς απόδειξη, ότι κάθε πραγματικός και συμμετρικός πίνακας έχει πραγματικές ιδιοτιμές.

Απόδειξη: Έστω $\underline{\underline{A}} \in \mathbb{R}^{n,n}$ για τον οποίο ισχύει $\underline{\underline{A}}^T = \underline{\underline{A}}$. Από τον ορισμό των ιδιοτιμών/ιδιοδιανυσμάτων έχουμε $\underline{\underline{A}} \cdot \underline{x} = \lambda \underline{x} \Rightarrow \left(\underline{\underline{A}} \cdot \underline{x}\right)^T = \left(\lambda \underline{x}\right)^T \Rightarrow \underline{\underline{A}}^T \cdot \underline{x} = \lambda \underline{x}$

Θετικά ορισμένοι πίνακες

Έστω τετραγωνικός και συμμετρικός πίνακας $\underline{\underline{A}} \in \mathbb{R}^{n,n}$ ($\underline{\underline{A}}^T = \underline{\underline{A}}$). Αν ισχύει $\underline{x}^T \cdot \underline{\underline{A}} \cdot \underline{x} > 0$ $\forall \underline{x} \in \mathbb{R}^n$ με $\underline{x} \neq \underline{0}$ τότε ο πίνακας $\underline{\underline{A}}$ ονομάζεται ***θετικά ορισμένος***.

Πρόταση 1η: Ο τετραγωνικός, μοναδιαίος πίνακας, $\underline{\underline{I}}$, είναι ένας θετικά ορισμένος.

Απόδειξη: Ο πίνακας $\underline{\underline{I}} = \text{diag}\{1,1,1,...,1\}$ είναι τετραγωνικός, πραγματικός και συμμετρικός πίνακας. Επιπλέον, αν $\underline{x} \in \mathbb{R}^n$ με $\underline{x} \neq \underline{0}$, δηλαδή το διάνυσμα αυτό έχει τουλάχιστον ένα μη-μηδενικό στοιχείο, θα ισχύει ότι $\underline{x}^T \cdot \underline{\underline{I}} \cdot \underline{x} = \underline{x}^T \cdot \underline{x} = x_1^2 + x_2^2 + \ldots + x_n^2 > 0$

Πρόταση 2η: Κάθε θετικά ορισμένος πίνακας $\underline{\underline{A}} \in \mathbb{R}^{n,n}$ είναι αντιστρέψιμος.

Απόδειξη: Εφόσον ο $\underline{\underline{A}}$ είναι θετικά ορισμένος άρα για κάθε $\underline{x} \in \mathbb{R}^n$ με $\underline{x} \neq \underline{0}$ ισχύει $\underline{x}^T \cdot \underline{\underline{A}} \cdot \underline{x} > 0$. Έστω τώρα ότι ο πίνακας $\underline{\underline{A}}$ δεν είναι αντιστρέψιμος. Επομένως το γραμμικό, ομογενές σύστημα αλγεβρικών εξισώσεων $\underline{\underline{A}} \cdot \underline{x} = \underline{0}$ έχει μη τετριμμένη λύση, δηλαδή υπάρχει $\underline{x}^* \in \mathbb{R}^n$ με $\underline{x}^* \neq \underline{0}$ το οποίο ικανοποιεί αυτήν την εξίσωση. Πολλαπλασιάζουμε από αριστερά την εξίσωση με $\underline{x}^{*T}$ και παίρνουμε $\underline{x}^{*T} \cdot \underline{\underline{A}} \cdot \underline{x}^* = 0$, το οποίο φυσικά είναι άτοπο. Άρα η υπόθεσή μας είναι λανθασμένη και επομένως ο πίνακας είναι αντιστρέψιμος.

Πρόταση 3η: Τα διαγώνια στοιχεία ενός θετικά ορισμένου πίνακα $\underline{\underline{A}} \in \mathbb{R}^{n,n}$ είναι θετικοί, μη μηδενικοί αριθμοί.

Απόδειξη: Εφόσον ο $\underline{\underline{A}}$ *είναι θετικά ορισμένος άρα* $\forall \underline{x} \in \mathbb{R}^n$ με $\underline{x} \neq \underline{0}$ ισχύει $\underline{x}^T \cdot \underline{\underline{A}} \cdot \underline{x} > 0$. Επιλέγουμε «n» στο πλήθος γραμμικά ανεξάρτητα διανύσματα τα οποία αποτελούν μία βάση του $\mathbb{R}^n$, δηλαδή τα στοιχεία του συνόλου $\left\{\underline{e}^{(i)}\right\}_{i=1}^{n}$. Επομένως θα έχουμε $\underline{e}^{(i)T} \cdot \underline{\underline{A}} \cdot \underline{e}^{(i)} = a_{ii} > 0$.

Πρόταση 4η: Ένας πραγματικός και συμμετρικός πίνακας $\underline{\underline{A}} \in \mathbb{R}^{n,n}$ είναι θετικά ορισμένος αν και μόνο αν οι ιδιοτιμές του είναι αυστηρά θετικές.

Απόδειξη: Θα αποδείξουμε πρώτα ότι αν ο $\underline{\underline{A}}$ είναι θετικά ορισμένος τότε οι ιδιοτιμές του είναι αυστηρά θετικές. Πράγματι, εφόσον ο $\underline{\underline{A}}$ είναι θετικά ορισμένος άρα ισχύει $\underline{x}^T \cdot \underline{\underline{A}} \cdot \underline{x} > 0 \quad \forall \underline{x} \in \mathbb{R}^n$ με $\underline{x} \neq \underline{0}$ (*), ενώ είναι γνωστό ότι κάθε πραγματικός και συμμετρικός πίνακας έχει πραγματικές ιδιοτιμές. Έτσι έχουμε $\underline{\underline{A}} \cdot \underline{y} = \lambda \underline{y} \Rightarrow \underline{y}^T \cdot \underline{\underline{A}} \cdot \underline{y} = \lambda \underline{y}^T \cdot \underline{y}$ όπου λ ιδιοτιμή και $\underline{y} \neq \underline{0}$ το αντίστοιχο ιδιοδιάνυσμα. Όμως τα ιδιοδιανύσματα μπορεί να είναι ορθοκανονικοποιημένα, επομένως $\underline{y}^T \cdot \underline{y} = 1$ και άρα $\underline{y}^T \cdot \underline{\underline{A}} \cdot \underline{y} = \lambda$. Λόγω όμως της (*) θα πρέπει $\underline{y}^T \cdot \underline{\underline{A}} \cdot \underline{y} = \lambda > 0$.

Θα αποδείξουμε τώρα το αντίστροφο, δηλαδή ότι αν ένας πραγματικός και συμμετρικός πίνακας έχει θετικές, μη-μηδενικές, ιδιοτιμές τότε είναι θετικά ορισμένος. Πράγματι από τον ορισμό ιδιοτιμών/ιδιοδιανυσμάτων έχουμε ότι $\underline{\underline{A}} \cdot \underline{y}_i = \lambda_i \underline{y}_i, \quad \underline{y}_i \neq \underline{0}$ (*) και $\lambda_i > 0$ λόγω της υπόθεσης, όπου $i = 1, 2, ..., n$. Επομένως, αν $\left\{\underline{y}_i\right\}_{i=1}^{n}$ είναι το σύνολο των ιδιοδιανυσμάτων του πίνακα, εξαιτίας του ότι είναι ορθοκανονικοποιημένα αποτελούν μία βάση του $\mathbb{R}^n$. Έτσι κάθε μη-μηδενικό διάνυσμα $\underline{x} \in \mathbb{R}^n$ μπορεί να γραφεί ως γραμμικός συνδυσμός των $\underline{y}_i$, δηλαδή υπάρχουν σταθερές $c_i, i = 1, 2, ..., n$, οι οποίες δεν μπορεί να είναι όλες μηδενικές, τέτοιες ώστε $\underline{x} = c_1 \underline{y}_1 + c_2 \underline{y}_2 + ... + c_n \underline{y}_n = \sum_{i=1}^{n} c_i \underline{y}_i$. Έτσι έχουμε:

$$\underline{\underline{A}} \cdot \underline{x} = \underline{\underline{A}} \cdot \left(c_1 \underline{y}_1 + c_2 \underline{y}_2 + ... + c_n \underline{y}_n\right) = \underline{\underline{A}} \cdot \sum_{i=1}^{n} c_i \underline{y}_i = \sum_{i=1}^{n} c_i \underline{\underline{A}} \cdot \underline{y}_i \overset{(*)}{=} \sum_{i=1}^{n} c_i \lambda_i \underline{y}_i \Rightarrow$$

$$x^T \cdot \underline{\underline{A}} \cdot \underline{x} = \left(\sum_{i=1}^{n} c_i \underline{y}_i \right) \cdot \left(\sum_{j=1}^{n} c_j \lambda_j \underline{y}_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(c_i c_j \lambda_j \underline{y}_i \cdot \underline{y}_j \right).$$

Λόγω όμως ότι τα ιδιοδιανύσματα είναι ορθοκανονικοποιημένα άρα $\underline{y}_i \cdot \underline{y}_j = 0$ αν $i \neq j$ και $\underline{y}_i \cdot \underline{y}_j = 1$ αν $i = j$. Επομένως από το παραπάνω διπλό άθροισμα οι μόνοι μη-μηδενικοί όροι είναι εκείνοι για τους οποίους $i = j$, δηλαδή $x^T \cdot \underline{\underline{A}} \cdot \underline{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(c_i c_j \lambda_j \underline{y}_i \cdot \underline{y}_j \right) = \sum_{i=1}^{n} c_i c_i \lambda_i = \sum_{i=1}^{n} c_i^2 \lambda_i > 0$ αφού υπάρχει τουλάχιστον ένα $c_i \neq 0$ και $\lambda_i > 0, \quad i = 1, 2, ..., n$.

Παράρτημα Π.2
Νόρμες συναρτήσεων, διανυσμάτων, πινάκων

Νόρμες συναρτήσεων

Έστω οι συναρτήσεις $f \in C[a,b]$, δηλαδή οι συνεχείς συναρτήσεις ορισμένες στο κλειστό διάστημα $[a,b]$. Ονομάζουμε «νόρμα της συνάρτησης f» μία απεικόνιση από το σύνολο των συνεχών συναρτήσεων στους θετικούς πραγματικούς αριθμούς $C[a,b] \to \mathbb{R}_+$. Γράφουμε: $\|\cdot\| : C[a,b] \to \mathbb{R}_+$. Η απεικόνιση υπόκειται σε τρία αξιώματα, το τελευταίο από τα οποία είναι γνωστό και ως τριγωνική ανισότητα:

(N1) $\|f\| > 0 \ \ \forall f \in C[a,b]$ εκτός εάν $\|f\| = 0 \Leftrightarrow f = 0$

(N2) $\|\lambda f\| = |\lambda| \|f\| \ \ \forall f \in C[a,b]$ και $\forall \lambda \in \mathbb{R}$

(N3) $\|f + g\| \le \|f\| + \|g\| \ \ \forall f, g \in C[a,b]$

Οι πιο γνωστές νόρμες συναρτήσεων είναι οι λεγόμενες p-νόρμες οι οποίες ορίζονται ως εξής:

$$\|f\|_p := \left(\int_a^b |f(x)|^p \, dx \right)^{1/p}, \quad 1 \le p < \infty$$

Επίσης $\lim_{p \to \infty} \|f\|_p = \|f\|_\infty = \max_{a \le x \le b} |f(x)|$ γνωστή και ως μέγιστη, άπειρη ή νόρμα Chebyshev.

Οι πιο συχνά χρησιμοποιούμενες νόρμες είναι οι:

$$\|f\|_1 := \int_a^b |f(x)| dx$$

$$\|f\|_2 := \sqrt{\int_a^b f^2(x) dx}$$ (γνωστή ως Ευκλίδεια νόρμα)

$$\|f\|_\infty := \max_{a \le x \le b} |f(x)|$$

Νόρμες διανυσμάτων

Οι νόρμες διανυσμάτων είναι απεικονίσεις οι οποίες ορίζονται από έναν K-γραμμικό χώρο (όπου $K = \mathbb{R}$ ή $\mathbb{C}$) στους θετικούς πραγματικούς αριθμούς. Για τις ανάγκες της Αριθμητικής Ανάλυσης θα περιοριστούμε στις νόρμες ορισμένες $\mathbb{R}^n \to \mathbb{R}_+$. Οι νόρμες διανυσμάτων υπόκεινται σε 3 αξιώματα:

(N1) $\|\underline{x}\| > 0 \quad \forall \underline{x} \in \mathbb{R}^n$ εκτός εάν $\|\underline{x}\| = 0 \Leftrightarrow \underline{x} = \underline{0}$

(N2) $\|\lambda \underline{x}\| = |\lambda| \|\underline{x}\| \quad \forall \underline{x} \in \mathbb{R}^n$ και $\forall \lambda \in \mathbb{R}$

(N3) $\|\underline{x} + \underline{y}\| \le \|\underline{x}\| + \|\underline{y}\| \quad \forall \underline{x}, \underline{y} \in \mathbb{R}^n$

Η (N3) όπως και προηγουμένως ονομάζεται τριγωνική ανισότητα. Επιπλέον της (N3) μπορεί να αποδειχτεί και η λεγόμενη *τριγωνική ανισότητα προς τα κάτω*. Έχουμε:

$$\|\underline{x}\| = \|(\underline{x} - \underline{y}) + \underline{y}\| \le \|\underline{x} - \underline{y}\| + \|\underline{y}\| \Rightarrow \|\underline{x} - \underline{y}\| \ge \|\underline{x}\| - \|\underline{y}\| \Rightarrow \|\underline{x}\| - \|\underline{y}\| \le -\|\underline{x} - \underline{y}\|$$

$$\|\underline{y}\| = \|(\underline{y} - \underline{x}) + \underline{x}\| \le \|\underline{x} - \underline{y}\| + \|\underline{x}\| \Rightarrow \|\underline{x} - \underline{y}\| \ge \|\underline{y}\| - \|\underline{x}\| \Rightarrow \|\underline{x}\| - \|\underline{y}\| \ge -\|\underline{x} - \underline{y}\|$$

Από τις δύο τελευταίες σχέσεις έχουμε:

$\|\underline{x}\| - \|\underline{y}\| \le -\|\underline{x} - \underline{y}\|$???

$$\big|\|\underline{x}\| - \|\underline{y}\|\big| \le \|\underline{x} - \underline{y}\| \Rightarrow \|\underline{x} - \underline{y}\| \ge \big|\|\underline{x}\| - \|\underline{y}\|\big|$$

Οι πιο γνωστές νόρμες είναι οι λεγόμενες p-νόρμες οι οποίες ορίζονται ως εξής:

$$\|\underline{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \quad 1 \le p < \infty$$

Επίσης $\lim_{p \to \infty} \|\underline{x}\|_p = \|\underline{x}\|_\infty = \max_{1 \le i \le n} |x_i|$ γνωστή και ως μέγιστη, άπειρη ή νόρμα Chebyshev.

Οι πιο συχνά χρησιμοποιούμενες νόρμες είναι οι:

$$\ell_1 \equiv \|\underline{x}\|_1 := \sum_{i=1}^{n} |x_i|$$

$\ell_2 \equiv \|\underline{x}\|_2 := \sqrt{\sum_{i=1}^{n} x_i^2}$ (γνωστή ως Ευκλίδεια νόρμα)

$$\ell_\infty \equiv \|\underline{x}\|_\infty := \max_{1 \le i \le n} (|x_i|)$$

Στην συνέχεια δίνουμε τους παρακάτω ορισμούς

(α) Έστω δύο διανυσματικές νόρμες ορισμένες στον $\mathbb{R}^n$, $\|\bullet\|$ και $\|\bullet\|'$. Οι νόρμες αυτές λέγονται ισοδύναμες όταν υπάρχουν θετικές πραγματικές σταθερές m και M τέτοιες ώστε $m\|\underline{x}\| \le \|\underline{x}\|' \le M\|\underline{x}\|, \quad \forall \underline{x} \in \mathbb{R}^n$

Από τον ορισμό αυτό εύκολα προκύπτει ότι:

$$\frac{1}{M}\|\underline{x}\|' \le \|\underline{x}\| \le \frac{1}{m}\|\underline{x}\|', \quad \forall \underline{x} \in \mathbb{R}^n$$

(β) Έστω η ακολουθία διανυσμάτων $\left\{\underline{x}^{(i)}\right\}_{i=1}^{\infty} \subset \mathbb{R}^n$ (που συνήθως προκύπτει από τις επαναληπτικές μεθόδους επίλυσης γραμμικών συστημάτων). Λέμε ότι η ακολουθία συγκλίνει ως προς την νόρμα του $\mathbb{R}^n$ $\|\cdot\|$, όταν υπάρχει $\underline{x}^* \in \mathbb{R}^n$ τέτοιο ώστε $\lim_{i\to\infty}\left\|\underline{x}^{(i)} - \underline{x}^*\right\| = 0$. Το διάνυσμα $\underline{x}^*$ ονομάζεται όριο της ακολουθίας ως προς την νόρμα $\|\cdot\|$

Χωρίς απόδειξη σημειώνουμε ότι ***όλες οι διανυσματικές νόρμες ορισμένες στον*** $\mathbb{R}^n$ ***είναι ισοδύναμες μεταξύ τους***. Το γεγονός αυτό έχει επίπτωση στην σύγκλιση ακολουθίας διανυσμάτων $\left\{\underline{x}^{(i)}\right\}_{i=1}^{\infty}$. Έτσι αν μία ακολουθία διανυσμάτων συγκλίνει ως προς μία συγκεκριμένη νόρμα $\|\cdot\|$, δηλαδή υπάρχει το όριο $\underline{x}^*$ της $\left\{\underline{x}^{(i)}\right\}_{i=1}^{\infty}$, λόγω του γεγονότος ότι όλες οι νόρμες είναι ισοδύναμες μεταξύ τους άρα υπάρχουν θετικές σταθερές m και M τέτοιες ώστε $m\left\|\underline{x}^{(i)} - \underline{x}^*\right\| \le \left\|\underline{x}^{(i)} - \underline{x}^*\right\|' \le M\left\|\underline{x}^{(i)} - \underline{x}^*\right\|$. Παίρνοντας όρια έχουμε

$$m\lim_{i\to\infty}\left\|\underline{x}^{(i)} - \underline{x}^*\right\| \le \lim_{i\to\infty}\left\|\underline{x}^{(i)} - \underline{x}^*\right\|' \le M\lim_{i\to\infty}\left\|\underline{x}^{(i)} - \underline{x}^*\right\| \Rightarrow 0 \le \lim_{i\to\infty}\left\|\underline{x}^{(i)} - \underline{x}^*\right\|' \le 0 \Rightarrow \lim_{i\to\infty}\left\|\underline{x}^{(i)} - \underline{x}^*\right\|' = 0$$

και επομένως η ακολουθία συγκλίνει και ως προς οποιαδήποτε άλλη νόρμα του $\mathbb{R}^n$.

Νόρμες πινάκων

Οι νόρμες πινάκων είναι απεικονίσεις που ορίζονται από $\mathbb{R}^{n,n} \to \mathbb{R}_+$ και οι οποίες υπόκεινται σε 4 αξιώματα:

(N1) $\left\|\underline{\underline{A}}\right\| > 0 \quad \forall \underline{\underline{A}} \in \mathbb{R}^{n,n}$ εκτός εάν $\left\|\underline{\underline{A}}\right\| = 0 \Leftrightarrow \underline{\underline{A}} = \underline{\underline{0}}$

(N2) $\left\|\lambda\underline{\underline{A}}\right\| = |\lambda|\left\|\underline{\underline{A}}\right\| \quad \forall \underline{\underline{A}} \in \mathbb{R}^{n,n}$ και $\forall \lambda \in \mathbb{R}$

(N3) $\left\|\underline{\underline{A}} + \underline{\underline{B}}\right\| \le \left\|\underline{\underline{A}}\right\| + \left\|\underline{\underline{B}}\right\| \quad \forall \underline{\underline{A}}, \underline{\underline{B}} \in \mathbb{R}^{n,n}$

(N4) $\left\|\underline{\underline{A}} \cdot \underline{\underline{B}}\right\| \le \left\|\underline{\underline{A}}\right\|\left\|\underline{\underline{B}}\right\| \quad \forall \underline{\underline{A}}, \underline{\underline{B}} \in \mathbb{R}^{n,n}$

Η (N3) είναι γνωστή και ως *τριγωνική ανισότητα για την πρόσθεση* ενώ η (N4) ως *τριγωνική ανισότητα για τον πολλαπλασιασμό*.

Φυσικές νόρμες πινάκων (ή νόρμες τελεστών)

Πρόκειται για ένα υποσύνολο των νορμών πινάκων το οποίο έχει μία επιπλέον ιδιότητα. Για τον ορισμό των φυσικών νορμών είναι απαραίτητο αρχικά να ορίσουμε μία διανυσματική νόρμα $\|\cdot\|$ στον $\mathbb{R}^n$ και στην συνέχεια να ορίσουμε την απεικόνιση

$$\left\|\underline{\underline{A}}\right\| := \sup_{\substack{\underline{x}\in\mathbb{R}^n \\ \underline{x}\neq\underline{0}}} \frac{\left\|\underline{\underline{A}}\cdot\underline{x}\right\|}{\|\underline{x}\|} = \sup_{\|\underline{x}\|=1} \left\|\underline{\underline{A}}\cdot\underline{x}\right\|$$

η οποία ονομάζεται ***φυσική νόρμα πινάκων*** ή ***νόρμα τελεστών***.

Αρχικά θα πρέπει να δείξουμε ότι η ποσότητα $\frac{\left\|\underline{\underline{A}}\cdot\underline{x}\right\|}{\|\underline{x}\|}, \underline{0}\neq\underline{x}\in\mathbb{R}^n$ είναι καλά ορισμένη.

Πράγματι από τον ορισμό των ισοδύναμων νορμών έχουμε ότι υπάρχουν θετικές σταθερές m και M τέτοιες ώστε $m\|\underline{x}\|_\infty \le \|\underline{x}\| \le M\|\underline{x}\|_\infty, \forall \underline{x}\in\mathbb{R}^n$ επομένως θα έχουμε και ότι $m\left\|\underline{\underline{A}}\cdot\underline{x}\right\|_\infty \le \left\|\underline{\underline{A}}\cdot\underline{x}\right\| \le M\left\|\underline{\underline{A}}\cdot\underline{x}\right\|_\infty, \forall \underline{x}\in\mathbb{R}^n$ και $\underline{\underline{A}}\in\mathbb{R}^{n,n}$. Άρα

$$\frac{\left\|\underline{\underline{A}}\cdot\underline{x}\right\|}{\|\underline{x}\|} \le \frac{M\left\|\underline{\underline{A}}\cdot\underline{x}\right\|_\infty}{m\|\underline{x}\|_\infty} = \frac{M}{m}\frac{\max\limits_{1\le i\le n}\left|\sum\limits_{j=1}^n a_{ij}x_j\right|}{\|\underline{x}\|_\infty} \le \frac{M}{m}\frac{\max\limits_{1\le i\le n}\sum\limits_{j=1}^n |a_{ij}x_j|}{\|\underline{x}\|_\infty} \le \frac{M}{m}\frac{\max\limits_{1\le i\le n}\left(\|\underline{x}\|_\infty\sum\limits_{j=1}^n |a_{ij}|\right)}{\|\underline{x}\|_\infty} \Rightarrow$$

$$\frac{\left\|\underline{\underline{A}}\cdot\underline{x}\right\|}{\|\underline{x}\|} \le \frac{M}{m}\max_{1\le i\le n}\left(\sum_{j=1}^n |a_{ij}|\right) := C < \infty$$

Στην συνέχεια πρέπει να δείξουμε ότι μία φυσική νόρμα πινάκων ικανοποιεί τις ιδιότητες (N1)-(N4) των νορμών πινάκων. Έχουμε ότι:

(N1): Γιά κάθε τετραγωνικό πίνακα $\underline{\underline{A}}\in\mathbb{R}^{n,n}$, $\left\|\underline{\underline{A}}\right\| = 0 \Leftrightarrow \sup\limits_{\substack{\underline{x}\in\mathbb{R}^n \\ \underline{x}\neq\underline{0}}} \frac{\left\|\underline{\underline{A}}\cdot\underline{x}\right\|}{\|\underline{x}\|} = 0 \Leftrightarrow \sup\limits_{\substack{\underline{x}\in\mathbb{R}^n \\ \underline{x}\neq\underline{0}}} \left\|\underline{\underline{A}}\cdot\underline{x}\right\| = 0$ Η τελευταία ισότητα μας λεει ότι για κάθε μη-μηδενικό διάνυσμα, $\underline{x}\in\mathbb{R}^n, \underline{x}\neq\underline{0}$, ισχύει $\left\|\underline{\underline{A}}\cdot\underline{x}\right\| = 0$ και επομένως σύμφωνα με το πρώτο αξίωμα της νόρμας διανυσμάτων θα έχουμε $\underline{\underline{A}}\cdot\underline{x} = \underline{0}, \quad \underline{x}\neq\underline{0}$. Επομένως ο πίνακας $\underline{\underline{A}}$ θα είναι ο μηδενικός πίνακας, δηλαδή $\underline{\underline{A}} = \underline{\underline{0}}$

(N2): Γιά κάθε τετραγωνικό πίνακα $\underline{\underline{A}} \in \mathbb{R}^{n,n}$ και κάθε πραγματικό αριθμό $\underline{\underline{A}} \in \mathbb{R}^{n,n}$ $\lambda \in \mathbb{R}$. Όμως από το δεύτερο αξίωμα της νόρμας διανυσμάτων ισχύει $\left\| \lambda \underline{\underline{A}} \cdot \underline{x} \right\| = |\lambda| \left\| \underline{\underline{A}} \cdot \underline{x} \right\|$ οπότε έχουμε $\left\| \lambda \underline{\underline{A}} \right\| = \sup_{\substack{\underline{x} \in \mathbb{R}^n \\ \underline{x} \neq \underline{0}}} \frac{\left\| \lambda \underline{\underline{A}} \cdot \underline{x} \right\|}{\left\| \underline{x} \right\|} = \sup_{\substack{\underline{x} \in \mathbb{R}^n \\ \underline{x} \neq \underline{0}}} \frac{|\lambda| \left\| \underline{\underline{A}} \cdot \underline{x} \right\|}{\left\| \underline{x} \right\|} = |\lambda| \sup_{\substack{\underline{x} \in \mathbb{R}^n \\ \underline{x} \neq \underline{0}}} \frac{\left\| \underline{\underline{A}} \cdot \underline{x} \right\|}{\left\| \underline{x} \right\|} = |\lambda| \left\| \underline{\underline{A}} \right\| \Rightarrow \left\| \lambda \underline{\underline{A}} \right\| = |\lambda| \left\| \underline{\underline{A}} \right\|$

(N3): Γιά κάθε τετραγωνικό πίνακα $\underline{\underline{A}}, \underline{\underline{B}} \in \mathbb{R}^{n,n}$:

$\left\| \underline{\underline{A}} + \underline{\underline{B}} \right\| = \sup_{\substack{\underline{x} \in \mathbb{R}^n \\ \underline{x} \neq \underline{0}}} \frac{\left\| \left(\underline{\underline{A}} + \underline{\underline{B}} \right) \cdot \underline{x} \right\|}{\left\| \underline{x} \right\|} \leq \sup_{\substack{\underline{x} \in \mathbb{R}^n \\ \underline{x} \neq \underline{0}}} \frac{\left\| \underline{\underline{A}} \cdot \underline{x} \right\| + \left\| \underline{\underline{B}} \cdot \underline{x} \right\|}{\left\| \underline{x} \right\|} = \sup_{\substack{\underline{x} \in \mathbb{R}^n \\ \underline{x} \neq \underline{0}}} \frac{\left\| \underline{\underline{A}} \cdot \underline{x} \right\|}{\left\| \underline{x} \right\|} + \sup_{\substack{\underline{x} \in \mathbb{R}^n \\ \underline{x} \neq \underline{0}}} \frac{\left\| \underline{\underline{B}} \cdot \underline{x} \right\|}{\left\| \underline{x} \right\|} = \left\| \underline{\underline{A}} \right\| + \left\| \underline{\underline{B}} \right\|$ οπότε έχουμε $\left\| \underline{\underline{A}} + \underline{\underline{B}} \right\| \leq \left\| \underline{\underline{A}} \right\| + \left\| \underline{\underline{B}} \right\|$

(N4): Γιά κάθε τετραγωνικό πίνακα $\underline{\underline{A}}, \underline{\underline{B}} \in \mathbb{R}^{n,n}$:

$\left\| \underline{\underline{A}} \cdot \underline{\underline{B}} \right\| = \sup_{\substack{\underline{x} \in \mathbb{R}^n \\ \underline{x} \neq \underline{0}}} \frac{\left\| \underline{\underline{A}} \cdot \left(\underline{\underline{B}} \cdot \underline{x} \right) \right\|}{\left\| \underline{x} \right\|} \leq \sup_{\substack{\underline{x} \in \mathbb{R}^n \\ \underline{x} \neq \underline{0}}} \frac{\left\| \underline{\underline{A}} \right\| \left\| \underline{\underline{B}} \cdot \underline{x} \right\|}{\left\| \underline{x} \right\|} = \left\| \underline{\underline{A}} \right\| \sup_{\substack{\underline{x} \in \mathbb{R}^n \\ \underline{x} \neq \underline{0}}} \frac{\left\| \underline{\underline{B}} \cdot \underline{x} \right\|}{\left\| \underline{x} \right\|} = \left\| \underline{\underline{A}} \right\| \left\| \underline{\underline{B}} \right\|$ οπότε έχουμε $\left\| \underline{\underline{A}} \cdot \underline{\underline{B}} \right\| \leq \left\| \underline{\underline{A}} \right\| \left\| \underline{\underline{B}} \right\|$

Τέλος, να σημειωθεί ότι λόγω του ορισμού θα ισχύει ότι $\left\| \underline{\underline{A}} \cdot \underline{x} \right\| \leq \left\| \underline{\underline{A}} \right\| \left\| \underline{x} \right\|$ (για να γίνει κατανοητό αυτό, αρκεί να θεωρήσουμε την απλή περίπτωση, του supremum μίας συνάρτησης, $f : I \to \mathbb{R}, \quad Q = \sup_{x \in I} \left(f(x) \right) \Rightarrow f(x) \leq Q, \forall x \in I$). Συγκεντρωτικά λοιπόν μία φυσική νόρμα πινάκων έχει τις παρακάτω ιδιότητες:

(N1) $\left\| \underline{\underline{A}} \right\| > 0 \quad \forall \underline{\underline{A}} \in \mathbb{R}^{n,n}$ εκτός εάν $\left\| \underline{\underline{A}} \right\| = 0 \Leftrightarrow \underline{\underline{A}} = \underline{\underline{0}}$

(N2) $\left\| \lambda \underline{\underline{A}} \right\| = |\lambda| \left\| \underline{\underline{A}} \right\|, \ \forall \underline{\underline{A}} \in \mathbb{R}^{n,n}$ και $\forall \lambda \in \mathbb{R}$

(N3) $\left\| \underline{\underline{A}} + \underline{\underline{B}} \right\| \leq \left\| \underline{\underline{A}} \right\| + \left\| \underline{\underline{B}} \right\|, \ \forall \underline{\underline{A}}, \underline{\underline{B}} \in \mathbb{R}^{n,n}$

(N4) $\left\| \underline{\underline{A}} \cdot \underline{\underline{B}} \right\| \leq \left\| \underline{\underline{A}} \right\| \left\| \underline{\underline{B}} \right\|, \ \forall \underline{\underline{A}}, \underline{\underline{B}} \in \mathbb{R}^{n,n}$

(N5) $\left\| \underline{\underline{A}} \cdot \underline{x} \right\| \leq \left\| \underline{\underline{A}} \right\| \left\| \underline{x} \right\|, \ \forall \underline{\underline{A}} \in \mathbb{R}^{n,n}$ και $\forall \underline{x} \in \mathbb{R}^{n,n}$

Να σημειωθεί ότι για κάθε φυσική νόρμα πινάκων ισχύει $\left\| \underline{\underline{I}} \right\| = 1$ αφού από τον ορισμό των φυσικών νορμών έχουμε ότι $\left\| \underline{\underline{I}} \right\| = \sup_{\|\underline{x}\|=1} \left\| \underline{\underline{I}} \cdot \underline{x} \right\| = \sup_{\|\underline{x}\|=1} \left\| \underline{x} \right\| = 1$.

Παραδείγματα νορμών πινάκων

Οι πιο γνωστές και συχνά χρησιμοποιούμενες νόρμες πινάκων είναι

(α) Η 1η νόρμα, γνωστή και ως το *μέγιστο του αθροίσματος των στηλών*,

$$\left\|\underline{\underline{A}}\right\|_1 = \max_{1\le j\le n}\left\{\sum_{i=1}^{n}\left|a_{ij}\right|\right\}$$

(β) Η μέγιστη νόρμα ή νόρμα απείρου, γνωστή και ως το *μέγιστο του αθροίσματος των γραμμών*,

$$\left\|\underline{\underline{A}}\right\|_\infty = \max_{1\le i\le n}\left\{\sum_{j=1}^{n}\left|a_{ij}\right|\right\}$$

Στην συνέχεια δείχνουμε την παρακάτω πρόταση. Για κάθε πίνακα $\underline{\underline{A}}\in\mathbb{R}^{n,n}$ ισχύει:

(α) $\left\|\underline{\underline{A}}\right\|_2 = \sqrt{\rho\left(\underline{\underline{A}}^T\cdot\underline{\underline{A}}\right)}$

(β) $\rho\left(\underline{\underline{A}}\right)\le\left\|\underline{\underline{A}}\right\|$

(γ) Αν ο πίνακας $\underline{\underline{A}}$ είναι συμμετρικός, $\left\|\underline{\underline{A}}\right\|_2 = \rho\left(\underline{\underline{A}}\right)$

Απόδειξη:

(α) Ακόμα και όταν ο πίνακας $\underline{\underline{A}}$ δεν είναι συμμετρικός, ο $\underline{\underline{A}}^T\cdot\underline{\underline{A}}$ είναι. Επομένως θα έχει πραγματικές, μη αρνητικές ιδιοτιμές, $\lambda_1,\lambda_2,...,\lambda_n$ και αντίστοιχα ιδιοδιανύσματα $\underline{y}_1,\underline{y}_2,...,\underline{y}_n$ τα οποία αποτελούν μία ορθοκανονική βάση του $\mathbb{R}^n$. Έτσι κάθε $\underline{x}\in\mathbb{R}^n$ μπορεί να γραφεί ως $\underline{x}=\sum_{i=1}^{n}c_i\underline{y}_i$. Τότε έχουμε ότι: $\left\|\underline{x}\right\|_2^2 = \underline{x}^T\cdot\underline{x} = \left(\sum_{i=1}^{n}c_i\underline{y}_i^T\right)\cdot\left(\sum_{i=1}^{n}c_i\underline{y}_i\right) = \sum_{i=1}^{n}c_i^2$. Επιπλέον για το διάνυσμα $\underline{\underline{A}}\cdot\underline{x}$ έχουμε ότι:

$$\left\|\underline{\underline{A}}\cdot\underline{x}\right\|_2^2 = \left(\underline{\underline{A}}\cdot\underline{x}\right)^T\cdot\left(\underline{\underline{A}}\cdot\underline{x}\right) = \underline{x}^T\cdot\left(\underline{\underline{A}}^T\cdot\underline{\underline{A}}\cdot\underline{x}\right) = \left(\sum_{i=1}^{n}c_i\underline{y}_i^T\right)\cdot\left(\sum_{i=1}^{n}\lambda_i c_i\underline{y}_i\right) = \sum_{i=1}^{n}\lambda_i c_i^2$$

Άρα $\dfrac{\left\|\underline{\underline{A}}\cdot\underline{x}\right\|_2^2}{\left\|\underline{x}\right\|_2^2} = \dfrac{\sum_{i=1}^{n}\lambda_i c_i^2}{\sum_{i=1}^{n}c_i^2} \le \dfrac{\max_{1\le i\le n}\left(\lambda_i\right)\sum_{i=1}^{n}c_i^2}{\sum_{i=1}^{n}c_i^2} = \max_{1\le i\le n}\left(\lambda_i\right) = \rho\left(\underline{\underline{A}}^T\cdot\underline{\underline{A}}\right) \Rightarrow \dfrac{\left\|\underline{\underline{A}}\cdot\underline{x}\right\|_2^2}{\left\|\underline{x}\right\|_2^2} \le \rho\left(\underline{\underline{A}}^T\cdot\underline{\underline{A}}\right)$ (*)

Έστω $\max_{1\le i\le n}\left(\lambda_i\right)=\lambda_k$. Τότε

$\left\|\underline{\underline{A}}\cdot\underline{y}_k\right\|_2^2=\underline{y}_k^T\cdot\left(\underline{\underline{A}}^T\cdot\underline{\underline{A}}\cdot\underline{y}_k\right)=\underline{y}_k^T\cdot\left(\lambda_k\underline{y}_k\right)=\lambda_k\underline{y}_k^T\cdot\underline{y}_k=\rho\left(\underline{\underline{A}}^T\cdot\underline{\underline{A}}\right)\left\|\underline{y}_k\right\|_2^2$ και επομένως $\frac{\left\|\underline{\underline{A}}\cdot\underline{y}_k\right\|_2^2}{\left\|\underline{y}_k\right\|_2^2}=\rho\left(\underline{\underline{A}}^T\cdot\underline{\underline{A}}\right)$ (**) το οποίο ουσιαστικά σημαίνει ότι υπάρχει μη-μηδενικό διάνυσμα του $\mathbb{R}^n$ για το οποίο η ισότητα ισχύει στην (*). Όμως από τον ορισμό της φυσικής νόρμας έχουμε:

$$\left\|\underline{\underline{A}}\right\|_2^2=\sup_{\substack{\underline{x}\in\mathbb{R}^n\\ \underline{x}\neq\underline{0}}}\frac{\left\|\underline{\underline{A}}\cdot\underline{x}\right\|_2^2}{\left\|\underline{x}\right\|_2^2}\overset{(*),(**)}{=}\rho\left(\underline{\underline{A}}^T\cdot\underline{\underline{A}}\right)\Rightarrow\left\|\underline{\underline{A}}\right\|_2=\sqrt{\rho\left(\underline{\underline{A}}^T\cdot\underline{\underline{A}}\right)}$$

Για προφανείς λόγους, η φυσική αυτή νόρμα ονομάζεται και ***φασματική νόρμα.***

(β) Έστω $\lambda_i,\underline{x}_i$ η ιδιοτιμή και το αντίστοιχο ιδιοδιάνυσμα του πίνακα $\underline{\underline{A}}$. Τότε από τον ορισμό ιδιοτιμών/ιδιοδιανυσμάτων θα ισχύει: $\underline{\underline{A}}\cdot\underline{x}_i=\lambda_i\underline{x}_i\Rightarrow\left\|\lambda_i\underline{x}_i\right\|=\left\|\underline{\underline{A}}\cdot\underline{x}_i\right\|\Rightarrow\left|\lambda_i\right|\left\|\underline{x}_i\right\|=\left\|\underline{\underline{A}}\cdot\underline{x}_i\right\|\leq\left\|\underline{\underline{A}}\right\|\left\|\underline{x}_i\right\|\Rightarrow\left|\lambda_i\right|\leq\left\|\underline{\underline{A}}\right\|$ το οποίο βέβαια θα ισχύει και για την μέγιστη ιδιοτιμή οπότε $\max_{1\leq i\leq n}\left|\lambda_i\right|\leq\left\|\underline{\underline{A}}\right\|\Rightarrow\rho\left(\underline{\underline{A}}\right)\leq\left\|\underline{\underline{A}}\right\|$

(γ) Αν $\underline{\underline{A}}^T=\underline{\underline{A}}$ τότε από $\underline{\underline{A}}\cdot\underline{x}=\lambda\underline{x}\Rightarrow\underline{\underline{A}}^T\cdot\underline{\underline{A}}\cdot\underline{x}=\lambda\underline{\underline{A}}^T\cdot\underline{x}\Rightarrow\underline{\underline{A}}^2\cdot\underline{x}=\lambda\underline{\underline{A}}\cdot\underline{x}=\lambda^2\underline{x}$. Άρα το λ^2 ιδιοτιμή του $\underline{\underline{A}}^T\cdot\underline{\underline{A}}=\underline{\underline{A}}^2$ και $\underline{x}$ το αντίστοιχο ιδιοδιάνυσμα. Άρα $\rho\left(\underline{\underline{A}}^T\cdot\underline{\underline{A}}\right)=\rho\left(\underline{\underline{A}}^2\right)=\lambda^2=\left[\rho\left(\underline{\underline{A}}\right)\right]^2$. Όμως $\left\|\underline{\underline{A}}\right\|_2=\sqrt{\rho\left(\underline{\underline{A}}^T\cdot\underline{\underline{A}}\right)}=\sqrt{\left[\rho\left(\underline{\underline{A}}\right)\right]^2}=\rho\left(\underline{\underline{A}}\right)$

Επίσης, γνωστή νόρμα είναι και η νόρμα Frobenius $\left\|\underline{\underline{A}}\right\|_F$ ή όπως αλλιώς αναφέρεται, Ευκλίδεια νόρμα, $\left\|\underline{\underline{A}}\right\|_E:\left\|\underline{\underline{A}}\right\|_E=\left\|\underline{\underline{A}}\right\|_F=\sqrt{\sum_{j=1}^{n}\sum_{i=1}^{n}a_{ij}^2}$

Η νόρμα αυτή δεν αποτελεί φυσική νόρμα πινάκων. Αυτό φαίνεται εύκολα απλά αν θεωρήσουμε τον μοναδιαίο πίνακα $\underline{\underline{I}}$ για τον οποίο γνωρίζουμε (δες παράδειγμα παραπάνω) ότι για κάθε φυσική νόρμα πινάκων ισχύει $\left\|\underline{\underline{I}}\right\|=1$. Από τον ορισμό της Frobenius νόρμας όμως προκύπτει $\left\|\underline{\underline{I}}\right\|_E=\left\|\underline{\underline{I}}\right\|_F=\sqrt{\sum_{j=1}^{n}\sum_{i=1}^{n}\delta_{ij}^2}=\sqrt{\sum_{i=1}^{n}\delta_{ii}^2}=\sqrt{n}$ το οποίο φυσικά αποδεικνύει ότι η νόρμα αυτή δεν είναι φυσική νόρμα.