Εθνικο Μετσοβιο Πολυτεχνειο
Τμημα Μηχανολογων Μηχανικων

Αριθμητική Ανάλυση για Μηχανικούς

Κ.Χ. Γιαννάκογλου, *Επ. Καθηγητής Ε.Μ.Π.*

Ι. Αναγνωστόπουλος, *Λέκτορας Ε.Μ.Π.*

Γ. Μπεργελές, *Καθηγητής Ε.Μ.Π.*

Αθήνα, 2003
3η Έκδοση

Περιεχόμενα

Κεφάλαιο 3. Επίλυση Συστημάτων

Κεφάλαιο 4. Αριθμητική Παρεμβολή και Προσέγγιση

Κεφάλαιο 5. Αριθμητική Ολοκλήρωση και Παραγώγιση

Κεφάλαιο 6. Αριθμητική Επίλυση Συνήθων Διαφορικών Εξισώσεων

Κεφάλαιο 1

Σφάλματα Αριθμητικών Υπολογισμών

Η αποτελεσματική χρήση των μεθόδων αριθμητικής ανάλυσης προϋποθέτει την κατανόηση της έννοιας του σφάλματος. Η επίτευξη απόλυτης ακρίβειας κατά την επίλυση ενός προβλήματος είναι πρακτικά αδύνατη, ακόμη και αν υπάρχει αναλυτική λύση, επειδή συνήθως υπεισέρχονται διάφορα σφάλματα σε όλη τη διαδικασία επίλυσης, από τη διαμόρφωση του μαθηματικού μοντέλου και την εισαγωγή των δεδομένων μέχρι την ανάπτυξη του αλγορίθμου και την εκτέλεση των αριθμητικών πράξεων. Επιπλέον, οι αριθμητικές μέθοδοι εισάγουν από τη φύση τους σφάλμα, καθώς είναι σχεδιασμένες για να προσεγγίζουν το αποτέλεσμα ενός προβλήματος, όταν δεν είναι δυνατή η αναλυτική του λύση.

Έτσι, το βασικό ερώτημα που πρέπει να απασχολεί έναν Μηχανικό, ο οποίος χρησιμοποιεί μεθόδους αριθμητικής ανάλυσης, δεν είναι εάν ένα αποτέλεσμα είναι ακριβές, αλλά ποιο είναι το μέγεθος του σφάλματος που εμπεριέχεται σε αυτό, καθώς και εάν και με ποιον τρόπο θα μπορούσε αυτό να περιορισθεί. Για να απαντηθεί το ερώτημα πρέπει αρχικά να εντοπισθεί και εκτιμηθεί το μέγεθος κάθε είδους σφάλματος που προκαλείται σε όλα τα στάδια των υπολογισμών, καθώς και η επίδραση που έχει η συσσώρευση και μετάδοση των σφαλμάτων αυτών από στάδιο σε στάδιο, μέχρι τη λήψη του τελικού αποτελέσματος. Στη συνέχεια, πρέπει να διερευνηθεί η δυνατότητα χρήσης των κατάλληλων αριθμητικών μεθόδων και τεχνικών που θα μειώσουν το σφάλμα κάτω από ένα προκαθορισμένο όριο, ώστε το τελικό αποτέλεσμα να είναι αξιόπιστο.

Η σημασία της διαδικασίας ανίχνευσης – εκτίμησης – ελαχιστοποίησης των σφαλμάτων είναι προφανής σε κάθε αριθμητική μέθοδο, καθώς πολλές φορές δεν υπάρχει άλλη δυνατότητα ελέγχου της ορθότητας του αποτελέσματος, ενώ οι συνέπειες ενός εσφαλμένου υπολογισμού σε πρακτικές εφαρμογές μπορεί να κοστίσουν ακριβά σε χρήμα αλλά ακόμη και σε ανθρώπινες ζωές.

Στη παρόν κεφάλαιο, μετά την παρουσίαση μερικών βασικών εννοιών, θα περιγραφούν και θα αναλυθούν οι διάφορες κατηγορίες σφαλμάτων που συναντώνται κατά τη χρήση μεθόδων αριθμητικής ανάλυσης.

1.1. Σφάλματα – Βασικές Έννοιες

Σφάλμα ονομάζεται γενικά η διαφορά μεταξύ της πραγματικής (ή ακριβούς) τιμής ενός μεγέθους ή μιας ποσότητας και μιας προσεγγιστικής τιμής, η οποία προκύπτει από εκτιμήσεις, μετρήσεις ή υπολογισμούς.

Ειδικότερα, *αριθμητικό σφάλμα* είναι το σφάλμα που προκύπτει κατά την εκτέλεση αριθμητικών πράξεων ή την εφαρμογή προσεγγιστικών αριθμητικών μεθόδων. Σε κάθε περίπτωση, το σφάλμα ορίζεται ως *απόλυτο* ή *σχετικό*, με τις ακόλουθες σχέσεις:

$$E_t = x_t - x_a \tag{1.1}$$

$$\varepsilon_t = \frac{E_t}{x_t} = \frac{x_t - x_a}{x_t} = 1 - \frac{x_a}{x_t} \tag{1.2}$$

όπου x_t και x_a είναι η πραγματική και η προσεγγιστική τιμή του μεγέθους αντιστοίχως, ενώ E_t είναι το απόλυτο και ε_t το σχετικό σφάλμα. Το τελευταίο μπορεί να πολλαπλασιασθεί επί 100, ώστε να εκφράζει το % σχετικό σφάλμα.

Όταν η πραγματική τιμή είναι κοντά στη μονάδα, οι δύο παραπάνω εκφράσεις δίνουν ίδιας τάξης μεγέθους αποτέλεσμα. Για μεγάλους αριθμούς η τιμή του σχετικού σφάλματος είναι πιο παραστατική και εποπτική, ενώ για αριθμούς κοντά στο μηδέν είναι προτιμότερη η χρήση του απόλυτου σφάλματος (στην περίπτωση μάλιστα όπου $x_t = 0$, η Εξ. 1.2 δεν ορίζεται).

Εφαρμογή 1.1.

α) Τα βάρη δύο σωμάτων βρέθηκαν μετά από ζύγιση ότι είναι 998 και 22 gr. Εάν τα πραγματικά τους βάρη είναι 1000 και 20 gr αντιστοίχως, να βρεθεί το απόλυτο και το σχετικό σφάλμα της κάθε ζύγισης.

Τα απόλυτα σφάλματα θα είναι:
$E_{t,1} = 1000 - 998 = 2$ gr και $E_{t,2} = 20 - 22 = -2$ gr
ενώ τα αντίστοιχα % σχετικά σφάλματα προκύπτουν από την Εξ. (1.2):

$$\varepsilon_{t,1} = \frac{1000-998}{1000}\,100\% = 0.2\% \quad \text{και} \quad \varepsilon_{t,2} = \frac{20-22}{20}\,100\% = 10\%$$

Επομένως, το απόλυτο σφάλμα των δύο μετρήσεων είναι ίδιο, αλλά προφανώς η ακρίβεια της δεύτερης μέτρησης είναι πολύ μικρότερη, πράγμα που φαίνεται από την τιμή του σχετικού της σφάλματος.

β) Τα βάρη δύο κόκκων σκόνης μετρήθηκαν 0.009 και 0.0009 gr. Εάν τα πραγματικά τους βάρη είναι 10 και 1 mgr αντιστοίχως, να βρεθεί το απόλυτο και το σχετικό σφάλμα της κάθε ζύγισης.

Τα απόλυτα σφάλματα θα είναι:
$E_{t,1} = 0.01 - 0.009 = 0.001$ gr και $E_{t,2} = 0.001 - 0.0009 = 0.0001$ gr
ενώ τα αντίστοιχα % σχετικά σφάλματα:

$$\varepsilon_{t,1} = \frac{0.01-0.009}{0.01}\,100\% = 10\% \quad \text{και} \quad \varepsilon_{t,2} = \frac{0.001-0.0009}{0.001}\,100\% = 10\%$$

Δηλαδή τα σχετικά σφάλματα προκύπτουν ίσα, ενώ η ακρίβεια της δεύτερης ζύγισης είναι σαφώς μεγαλύτερη, πράγμα που φαίνεται στην τιμή του απόλυτου σφάλματος.

Όμως η πραγματική ή ακριβής τιμή του αποτελέσματος δεν είναι συνήθως γνωστή κατά την εφαρμογή μιας αριθμητικής μεθόδου σε πρακτικά προβλήματα. Επομένως, το *πραγματικό* σφάλμα που δίνουν οι Εξ. (1.1) και (1.2) δεν μπορεί να υπολογισθεί. Αντί αυτού, το σφάλμα μπορεί σε ορισμένες περιπτώσεις να εκτιμηθεί κατά προσέγγιση, με διάφορους τρόπους. Η δυνατότητα εκτίμησης του σφάλματος χωρίς να είναι γνωστή η πραγματική τιμή του αποτελέσματος αποτελεί σημαντικό χαρακτηριστικό κάθε αριθμητικής μεθόδου και θα αναφερθεί κατά την παρουσίασή τους στα επόμενα κεφάλαια. Επειδή πολλές μέθοδοι αριθμητικής ανάλυσης είναι επαναληπτικές, δηλαδή παράγουν διαδοχικές

προσεγγίσεις που επιδιώκεται να έχουν όλο και μεγαλύτερη ακρίβεια, ένας συνήθης τρόπος εκτίμησης του σφάλματος είναι η χρήση της σχέσης

$$\varepsilon_a = \left(1 - \frac{x_a^{k-1}}{x_a^k}\right) \cdot 100\% \tag{1.3}$$

η οποία δίνει το *εκτιμώμενο* % σχετικό σφάλμα, χρησιμοποιώντας το προσεγγιστικό αποτέλεσμα δύο διαδοχικών επαναλήψεων, έστω k–1 και k, της μεθόδου. Εφόσον η επαναληπτική διαδικασία συγκλίνει, δηλαδή το εκτιμώμενο σφάλμα βαίνει διαρκώς μειούμενο, τότε όσο βελτιώνεται η προσέγγιση του αποτελέσματος x_a, τόσο το εκτιμώμενο σφάλμα ε_a προσεγγίζει το πραγματικό ε_t. Είναι όμως δυνατόν μια επαναληπτική μέθοδος να συγκλίνει μεν, αλλά όχι στο σωστό αποτέλεσμα. Σε αυτή την περίπτωση, το εκτιμώμενο σφάλμα μπορεί να διαφέρει σημαντικά από το πραγματικό.

Τέλος, αυτό που συνήθως ενδιαφέρει στο τελικό αποτέλεσμα είναι η απόλυτη τιμή του σφάλματος (απόλυτου ή σχετικού), η οποία απαιτείται να είναι μικρότερη μιας προκαθορισμένης, θετικής τιμής, ε_r:

$$|\varepsilon_a| \le \varepsilon_r \tag{1.4}$$

Εφαρμογή 1.2.

Η ρίζα μιας εξίσωσης βρίσκεται με μια επαναληπτική αριθμητική μέθοδο, η οποία τερματίζεται όταν το σχετικό σφάλμα γίνει μικρότερο από 0.1%. Αν το αποτέλεσμα των υπολογισμών είναι $x_a = 21$, να βρεθεί η περιοχή τιμών της ακριβούς λύσης.

Υποθέτοντας ότι το εκτιμώμενο σφάλμα προσεγγίζει το πραγματικό, προκύπτει από τις Εξ. (1.2) και (1.4):

$$|E_t| = |\varepsilon_t| \cdot |x_t| \approx |\varepsilon_a| \cdot |x_a| \le \varepsilon_r \cdot |x_a| = 0.001 \cdot 21 = 0.021$$

άρα $\quad -0.021 \le E_t = (x_t - 21) \le 0.021 \Rightarrow -20.979 \le x_t \le 21.021$

Επομένως, εκτός από την προσεγγιστική τιμή της ρίζας της εξίσωσης, μπορούν να καθορισθούν και τα όρια εντός των οποίων θα βρίσκεται η πραγματική λύση.

1.1.1. Σημαντικά Ψηφία

Κάθε αριθμός μπορεί να παρασταθεί, ακριβώς ή κατά προσέγγιση, με τη μορφή ενός δεκαδικού με πεπερασμένο πλήθος ψηφίων. Στην αρχή του αριθμού μπορεί να υπάρχουν ένα ή περισσότερα μηδενικά ψηφία, που καθορίζουν απλώς τη θέση της υποδιαστολής. Όσα από τα ψηφία του αριθμού που ακολουθούν είναι γνωστά με ακρίβεια, αποτελούν *σημαντικά ψηφία*. Στα σημαντικά ψηφία περιλαμβάνεται επίσης και ένα τελευταίο, εκτιμώμενο ψηφίο.

Παραδείγματα:

- Οι αριθμοί 1.902, 0.1902, 0.001902 έχουν όλοι τέσσερα σημαντικά ψηφία.
- Το πηλίκο 2/3 μπορεί να παρασταθεί με τέσσερα σημαντικά ψηφία ως 0.6666 ή ως 0.6667.
- Οι αριθμοί 0.23, 0.230 και 0.2300 έχουν δύο, τρία και τέσσερα σημαντικά ψηφία αντιστοίχως, επομένως μπορεί να μην είναι ίσοι!

– Οι αριθμός 3545400 μπορεί να έχει από πέντε έως επτά σημαντικά ψηφία, αναλόγως εάν τα δύο τελευταία μηδενικά είναι ακριβή ή είναι αποτέλεσμα στρογγυλοποίησης.

Επειδή, όπως φαίνεται και στα προηγούμενα παραδείγματα, μπορεί να υπάρξει σύγχυση για το πόσα είναι τα σημαντικά ψηφία ενός αριθμού, είναι προτιμητέα η χρήση της επιστημονικής γραφής, δηλαδή της μορφής $a_1.a_2a_3\ldots a_n \cdot 10^m$, όπου $a_1, \ldots a_n$ τα n σημαντικά ψηφία του αριθμού ($a_1 \neq 0$). Έτσι, οι αριθμοί του πρώτου παραδείγματος γράφονται: $1.902\cdot10^0$, $1.902\cdot10^{-1}$ και $1.902\cdot10^{-3}$. Επίσης, ο αριθμός $2.3\cdot10^{-1}$ θα έχει δύο σημαντικά ψηφία, ενώ ο $2.300\cdot10^{-1}$ θα έχει τέσσερα. Τέλος, αν ο αριθμός του τελευταίου παραδείγματος είναι στρογγυλοποιημένος, θα γραφεί $3545.4\cdot10^3$.

Η έννοια των σημαντικών ψηφίων σχετίζεται άμεσα με την ακρίβεια των αριθμητικών πράξεων, καθώς όσο πιο πολλά σημαντικά ψηφία χρησιμοποιούνται, τόσο περισσότερα σωστά ψηφία θα έχει το αποτέλεσμα μιας πράξης. Για παράδειγμα η τιμή π^2 θα προκύψει ~ 9.86 εάν ληφθεί $\pi = 3.14$, ενώ το σωστό αποτέλεσμα είναι ~ 9.87.

Στην περίπτωση ενός αριθμού που προσεγγίζει μια πραγματική τιμή, όπως είναι το αποτέλεσμα μιας αριθμητικής μεθόδου, ονομάζονται *σωστά σημαντικά ψηφία*, όσα αντιστοιχούν στα ψηφία της πραγματική τιμής. Με τη λέξη αντιστοιχούν υπονοείται ότι τα αντίστοιχα ψηφία μπορεί και να μην είναι ίδια, αν και συνήθως είναι. Για παράδειγμα, οι αριθμοί 10.04 και 9.99, οι οποίοι προσεγγίζουν την πραγματική τιμή 10.00, έχουν και οι δύο τρία σωστά σημαντικά ψηφία.

Στις περισσότερες αριθμητικές μεθόδους, εάν το εκτιμώμενο σχετικό σφάλμα ενός αριθμού είναι μικρότερο μιας τιμής ε_r, τότε ο αριθμός θα έχει τουλάχιστον n σωστά σημαντικά ψηφία, σύμφωνα με τη σχέση

$$\varepsilon_r = (0.5\cdot10^{-n}) \qquad \text{ή} \qquad \varepsilon_r = (0.5\cdot10^{2-n})\ \% \tag{1.5}$$

Έτσι, για τους αριθμούς 10.04 και 9.99 του προηγούμενου παραδείγματος, προκύπτει ότι $n \geq 2.1$ και $n \geq 2.7$ αντιστοίχως. Η Εξ. (1.5) μπορεί να χρησιμοποιηθεί και αντίστροφα, για να ορισθεί η τιμή του κριτηρίου ε_r, ώστε το αριθμητικό αποτέλεσμα να έχει τον επιθυμητό αριθμό σωστών σημαντικών ψηφίων.

1.2. Σφάλμα Στρογγυλοποίησης

Κατά την καταχώρηση ενός αριθμού, είτε στο χαρτί είτε στη μνήμη του ηλεκτρονικού υπολογιστή, διατηρείται για λόγους οικονομίας ένας περιορισμένος αριθμός σημαντικών ψηφίων. Έτσι, όλοι αριθμοί που η ακριβής παράστασή τους απαιτεί περισσότερα ή μη πεπερασμένα ψηφία (όπως π.χ. το κλάσμα 1/3 ή ο αριθμός π), δεν μπορούν να αντιπροσωπευθούν και να χρησιμοποιηθούν με ακρίβεια. Επιπλέον, υπάρχουν αριθμοί που, ενώ στο δεκαδικό σύστημα παριστάνονται ακριβώς και με λίγα ψηφία, δεν έχουν πεπερασμένη παράσταση στο δυαδικό σύστημα ενός υπολογιστή (όπως π.χ. ο δεκαδικός αριθμός 0.1, που εκφράζεται στο δυαδικό σύστημα ως 0.000110011001100...). Το σφάλμα που προκαλείται λόγω της παράλειψης σημαντικών ψηφίων ενός αριθμού ονομάζεται *σφάλμα στρογγυλοποίησης*.

Υπάρχουν δύο βασικοί τρόποι στρογγυλοποίησης: στον πρώτο (*απλή στρογγυλοποίηση*) αγνοούνται όλα τα σημαντικά ψηφία που δεν μπορούν να καταχωρηθούν, ενώ στον δεύτερο (*συμμετρική στρογγυλοποίηση*), το τελευταίο ψηφίο που διατηρείται αυξάνεται κατά μία μονάδα εάν το πρώτο ψηφίο που παραλείπεται είναι ίσο ή μεγαλύτερο του 5. Για παράδειγμα, εάν διατηρούνται πέντε σημαντικά ψηφία, ο αριθμός 4.56248 θα καταχωρηθεί ως 4.5624 με απλή και ως 4.5625 με συμμετρική στρογγυλοποίηση.

1.2.1. Καταχώρηση Αριθμών στη Μνήμη Η/Υ

Στα υπολογιστικά συστήματα χρησιμοποιείται για την παράσταση των ακέραιων αριθμών η αριθμητική σταθερής υποδιαστολής, ενώ για την παράσταση των πραγματικών αριθμών η αριθμητική κινητής υποδιαστολής. Σε έναν σύγχρονο 32-bit υπολογιστή διατίθενται συνήθως δύο bytes μνήμης (16 bit) για την καταχώρηση ενός ακέραιου αριθμού και τέσσερα bytes (32 bit) για έναν πραγματικό αριθμό απλής ακρίβειας. Η μνήμη αυτή διπλασιάζεται όταν οι αριθμοί είναι διπλής ακρίβειας.

Η υποδιαστολή στους ακέραιους αριθμούς θεωρείται ότι βρίσκεται πάντοτε μετά το τελευταίο ψηφίο του αριθμού. Ένα από τα 16 bits που διατίθενται για έναν ακέραιο χρησιμοποιείται για το πρόσημό του (0 ή 1 για θετικό ή αρνητικό αριθμό αντιστοίχως). Έτσι τα υπόλοιπα 15 bits μπορούν να παραστήσουν στο δυαδικό σύστημα όλους τους ακέραιους από 0 έως $\pm(2^{15} - 1) = \pm 32767$. Για την ακρίβεια, στη θέση του αριθμού – 0 που δεν χρειάζεται, καταχωρείται ένας ακόμη αρνητικός αριθμός, ο – 32768. Επίσης, οι αρνητικοί ακέραιοι δεν παριστάνονται κανονικά όπως οι θετικοί με απλή αλλαγή του προσήμου: χρησιμοποιείται η μέθοδος συμπληρώματος του 2 (2's complement), κατά την οποία ένας αρνητικός ακέραιος προκύπτει αναστρέφοντας όλα τα δυαδικά ψηφία του αντίστοιχου θετικού αριθμού και προσθέτοντας 1. Με τη μέθοδο αυτή απλοποιείται η πρόσθεση δύο ακεραίων, καθώς η διαδικασία παραμένει ίδια, ανεξαρτήτως του προσήμου των αριθμών, ενώ το πρόσημο του αποτελέσματος προκύπτει αυτόματα. Για ακέραιους αυξημένης ακρίβειας διατίθενται 31 bits συν ένα για το πρόσημο, επομένως μπορούν να παρασταθούν όλοι οι ακέραιοι από –2147483648 έως +2147483647. Διαπιστώνεται δηλαδή ότι σε κάθε περίπτωση υπάρχει άνω και κάτω όριο, πέρα από τα οποία ένας ακέραιος αριθμός δεν μπορεί να καταχωρηθεί στη μνήμη ενός ηλεκτρονικού υπολογιστή. Ένα τέτοιο ενδεχόμενο πρέπει να ελέγχεται με προσοχή, καθώς είναι δυνατόν (ανάλογα και με τη γλώσσα προγραμματισμού) να καταχωρηθεί, αντί για τη σωστή, μια τυχαία τιμή εντός των ορίων, χωρίς διαγνωστικό μήνυμα. Παρόλα αυτά, επειδή με την αριθμητική σταθερής υποδιαστολής οι ακέραιοι αριθμοί καταχωρούνται με ακρίβεια, δεν προκαλείται σφάλμα στρογγυλοποίησης.

Διαφορετικός είναι ο τρόπος με τον οποίο κωδικοποιούνται οι πραγματικοί αριθμοί. Σύμφωνα με την αριθμητική κινητής υποδιαστολής (floating point), κάθε αριθμός εκφράζεται σε εκθετική μορφή, οπότε προκύπτουν δύο ποσότητες: το κλασματικό μέρος ή βάση f (mantissa) και ο εκθέτης m (exponent), δηλαδή

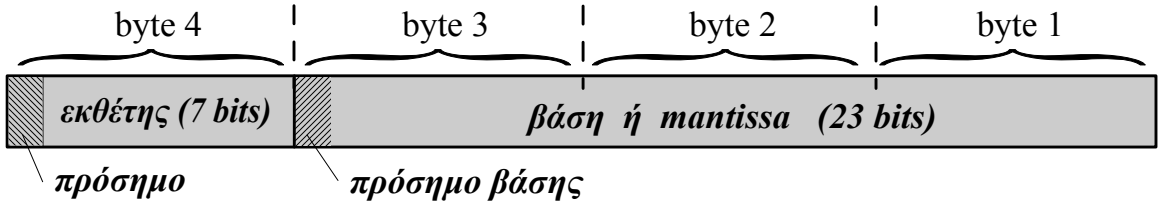
$$x = f \cdot b^m \tag{1.6}$$

όπου b η βάση του συστήματος αρίθμησης (10 ή 2 για δεκαδικό ή δυαδικό αντιστοίχως). Το κλασματικό μέρος είναι *κανονικοποιημένο*, δηλαδή το πρώτο ψηφίο αμέσως μετά την υποδιαστολή είναι διάφορο του μηδενός. Για παράδειγμα, εάν διατίθενται 5 θέσεις (bits) για το κλασματικό μέρος, ο αριθμός 157.45_{10} παριστάνεται ως $0.15745 \cdot 10^3$, ενώ ο αριθμός 10.001_2 ως $0.10001 \cdot 2^2$. Έτσι ισχύει γενικά και η σχέση

$$\frac{1}{b} \le |f| < 1 \tag{1.7}$$

ενώ εξαίρεση αποτελεί το μηδέν, που γράφεται συνήθως ως $0.00...0 \cdot b^{-M}$, όπου M η μέγιστη δυνατή τιμή του εκθέτη.

Από τα 32 bits που διατίθενται για έναν πραγματικό αριθμό απλής ακρίβειας, τα οκτώ bits (1 byte) περιέχουν τον εκθέτη m και τα υπόλοιπα 24 (3 bytes) τη βάση f. Στην πιο απλή περίπτωση, το πρώτο bit του εκθέτη και της βάσης διατίθενται για τον καθορισμό των αντίστοιχων προσήμων, όπως φαίνεται στο Σχήμα 1.1 (αν και τα πρόσημα αυτά μπορούν να κωδικοποιηθούν και με άλλους καλλίτερους τρόπους, εξοικονομώντας ένα πρόσθετο ψηφίο για τον εκθέτη ή/και για το κλασματικό μέρος).



Σχήμα 1.1. Διάταξη μονάδας καταχώρησης πραγματικού αριθμού των 32 bits.

Η μέγιστη δυνατή τιμή που μπορεί να πάρει ο εκθέτης στα επτά διαθέσιμα bits είναι $m = 1111111_2 = (2^7 - 1)_{10} = 127$. Επομένως, ο μέγιστος (κατά απόλυτη τιμή) πραγματικός αριθμός που μπορεί να καταχωρηθεί θα είναι $x = 1 \cdot 2^{127} \approx 2 \cdot 10^{38}$, και έτσι εάν κατά τη διάρκεια αριθμητικών πράξεων προκύψει μεγαλύτερος αριθμός, θα προκληθεί σφάλμα υπερχείλισης (overflow). Αντίστοιχα, ο ελάχιστη δυνατή απόλυτη τιμή ενός αριθμού θα είναι $x = 0.1 \cdot 2^{-127} \approx 0.6 \cdot 10^{-39}$, επομένως ένας αριθμός που πλησιάζει ακόμη περισσότερο στο μηδέν θα προκαλέσει και πάλι σφάλμα (underflow).

Όταν απαιτείται αυξημένη ακρίβεια ή πράξεις με ακόμη μεγαλύτερους αριθμούς, μπορεί να χρησιμοποιηθεί διπλή ακρίβεια (double precision), οπότε διατίθεται διπλάσια μνήμη, 64 bits, για έναν αριθμό. Στην περίπτωση αυτή, τα bits του εκθέτη γίνονται τουλάχιστον 10, οπότε καταχωρούνται αριθμοί μέχρι και $10^{\pm 308}$, ενώ παρέχεται ακρίβεια της τάξης των $15 \div 16$ σημαντικών ψηφίων.

Τέλος, μια άλλη τεχνική κωδικοποίησης δεκαδικών αριθμών είναι ο κώδικας BCD (Binary Coded Decimals), που χρησιμοποιείται συνήθως στους υπολογιστές τσέπης και σε

ορισμένες εμπορικές εφαρμογές. Με τη μέθοδο αυτή, κάθε ψηφίο ενός δεκαδικού αριθμού γράφεται ξεχωριστά στο δυαδικό σύστημα. Έτσι, τα ψηφία 0 έως 9, καθώς και τα πρόσημα συν και πλην μπορούν να κωδικοποιηθούν με μια ομάδα τεσσάρων bits, και κάθε αριθμός γράφεται ως ένα σύνολο τέτοιων ομάδων. Για παράδειγμα, ο αριθμός 159_{10} γράφεται $(0001)(0101)(1001)_{\text{BCD}}$. Επιπλέον, η μνήμη που διατίθεται μεταβάλλεται ανάλογα με τον αριθμό των σημαντικών ψηφίων ενός αριθμού. Οι πράξεις μεταξύ BCD αριθμών εκτελούνται ψηφίο – ψηφίο, όπως γίνεται και στο δεκαδικό σύστημα.

Τα πλεονεκτήματα της μεθόδου είναι ότι δεν χρειάζεται μετατροπή ενός BCD αριθμού στο δυαδικό σύστημα και αντιστρόφως, εξοικονομώντας έτσι τον σχετικό χρόνο, καθώς και ότι δεν υπάρχει πρόβλημα ακρίβειας στην καταχώρηση κάθε αριθμού με λίγα σημαντικά ψηφία (όπως συμβαίνει στο δυαδικό σύστημα, π.χ. για τον αριθμό 0.1). Όμως οι αριθμητικές πράξεις απαιτούν πολύ περισσότερο χρόνο, γι αυτό η μέθοδος χρησιμοποιείται όταν οι υπολογισμοί είναι σχετικά απλοί, αλλά απαιτείται αυξημένη ακρίβεια, όπως για παράδειγμα στην επεξεργασία οικονομικών στοιχείων.

1.2.2. Σφάλμα Στρογγυλοποίησης κατά την Καταχώρηση Αριθμών

Κάθε πραγματικός αριθμός που έχει περισσότερα σημαντικά ψηφία από όσα bits διατίθενται για τη βάση, θα στρογγυλοποιηθεί κατά την καταχώρησή του, με απλή ή συμμετρική στρογγυλοποίηση, όπως έχει ήδη αναφερθεί. Για παράδειγμα, τα 24 bits που διαθέτει ένας 32-bit υπολογιστής αντιστοιχούν σε περίπου επτά σημαντικά ψηφία στο δεκαδικό σύστημα, επομένως ένας αριθμός με περισσότερα σημαντικά ψηφία θα στρογγυλοποιηθεί.

Ένας πραγματικός αριθμός μπορεί να γραφεί γενικά στη μορφή

$$x_t = f \cdot b^m + g \cdot b^{m-k} \tag{1.8}$$

όπου ο όρος f περιέχει μόνο τα πρώτα k σημαντικά ψηφία που μπορούν να καταχωρηθούν (π.χ. $k = 23$), ενώ ο όρος g τα υπόλοιπα σημαντικά ψηφία. Οι δύο όροι είναι κανονικοποιημένοι, επομένως για το f ισχύει η Εξ. (1.7), ενώ για το g θα είναι

$$0 \le |g| < 1 \tag{1.9}$$

Μετά τη στρογγυλοποίηση θα προκύψει ένας προσεγγιστικός αριθμός x_a, που ανάλογα με τη χρησιμοποιούμενη μέθοδο θα είναι:

Απλή στρογγυλοποίηση:

$$x_a = f \cdot b^m \tag{1.10}$$

Συμμετρική στρογγυλοποίηση:

$$x_a = \begin{cases} f \cdot b^m, & |g| < 0.5 \\ f \cdot b^m \pm b^{m-k}, & |g| > 0.5 \end{cases} \tag{1.11}$$

όπου το πρόσημο του όρου b^{m-k} είναι αρνητικό όταν $f < 0$. Επίσης, όταν $|g| = 0.5$, τότε, εάν το ψηφίο k (τελευταίο ψηφίο) του f είναι άρτιο (ή μηδέν), λαμβάνεται η πρώτη εκ των δύο εκφράσεων της (1.11), ενώ όταν είναι περιττό, η δεύτερη (κανόνας του άρτιου ψηφίου).

Το σχετικό σφάλμα στρογγυλοποίησης $(x_t - x_a)/x_t$ εκτιμάται λαμβάνοντας υπόψη ότι (βλ. Εξ. 1.7, 1.8 και 1.9):

$$|x_t| = \left| f \cdot b^m + g \cdot b^{m-k} \right| \geq \left| f \cdot b^m \right| \geq \frac{1}{b} \cdot b^m = b^{m-1}$$

οπότε στην απλή στρογγυλοποίηση θα ισχύει

$$|\varepsilon_t| = \left| \frac{x_t - x_a}{x_t} \right| \leq \left| \frac{g \cdot b^{m-k}}{x_t} \right| \leq \left| \frac{1 \cdot b^{m-k}}{x_t} \right| \leq \frac{b^{m-k}}{b^{m-1}} = b^{1-k} \tag{1.12}$$

Κατά τη συμμετρική στρογγυλοποίηση όμως, το απόλυτο σφάλμα $|x_t - x_a|$ θα είναι ίσο με $|g| \cdot b^{m-k}$ όταν $|g| < 0.5$ και με $(1 - |g|) \cdot b^{m-k}$ όταν $|g| \geq 0.5$ (Εξ. 1.11). Επομένως, σε κάθε περίπτωση το σφάλμα αυτό θα είναι το πολύ ίσο με $0.5 \cdot b^{m-k}$. Έτσι, το σχετικό σφάλμα θα είναι το μισό από όσο κατά την απλή στρογγυλοποίηση.

Συγκεκριμένα, στο δεκαδικό σύστημα θα έχουμε $|\varepsilon_t| \leq 10^{1-k}$ για απλή, και $|\varepsilon_t| \leq 0.5 \cdot 10^{1-k}$ για συμμετρική στρογγυλοποίηση, ενώ στο δυαδικό σύστημα θα είναι $|\varepsilon_t| \leq 2^{1-k}$ και $|\varepsilon_t| \leq 2^{-k}$, αντιστοίχως.

Η απλή στρογγυλοποίηση είναι ευκολότερη διαδικασία και χρησιμοποιείται σε αρκετούς υπολογιστές αντί της πιο χρονοβόρας, συμμετρικής στρογγυλοποίησης, με την προϋπόθεση βέβαια ότι ο αριθμός των σημαντικών ψηφίων είναι αρκετά μεγάλος (π.χ. 24 bits), ώστε το σφάλμα να είναι ούτως ή άλλως πολύ μικρό.

Ένα σημαντικό πρόβλημα που μπορεί να προκαλέσει η στρογγυλοποίηση σε πρακτικούς υπολογισμούς, προέρχεται από το γεγονός ότι οι αριθμοί που καταχωρούνται μέσα σε ένα δεδομένο διάστημα είναι πεπερασμένοι. Για παράδειγμα, σε μια βάση τριών θέσεων, οι αριθμοί που ανήκουν στο διάστημα [0.5, 1] είναι οι εξής: 0.100, 0.101, 0.110, 0.111 και $0.100 \cdot 2^1$, ή αντιστοίχως στο δεκαδικό σύστημα: 0.5, 0.625, 0.75, 0.875 και 1. Επομένως, δύο διαφορετικοί πραγματικοί αριθμοί μπορεί να καταχωρηθούν ως ίσοι, όπως για παράδειγμα οι αριθμοί 0.63 και 0.68, οι οποίοι θα γίνουν 0.625. Επίσης, δύο σχεδόν ίσοι αριθμοί μπορεί να καταχωρηθούν πολύ διαφορετικοί, όπως οι αριθμοί 0.62499 και 0.62501, οι οποίοι στην απλή στρογγυλοποίηση θα γίνουν 0.5 και 0.625 αντιστοίχως.

Αυτό θα πρέπει να λαμβάνεται σοβαρά υπόψη όταν στον υπολογιστικό αλγόριθμο υπάρχουν εντολές ελέγχου που συγκρίνουν δύο πραγματικούς αριθμούς. Έτσι, αντί να ελέγχεται εάν δύο αριθμοί είναι ίσοι, συνιστάται να ελέγχεται εάν η σχετική διαφορά τους είναι μικρότερη από μια προκαθορισμένη τιμή

$$\left| \frac{x_1 - x_2}{x_1} \right| \leq \delta \tag{1.13}$$

όπου η ανοχή δ δεν πρέπει να είναι μικρότερη από το μέγιστο σφάλμα στρογγυλοποίησης, δηλαδή από 2^{1-k}. Το ίδιο ισχύει και στην περίπτωση όπου το δ αποτελεί κριτήριο σύγκλισης μιας επαναληπτικής μεθόδου: αν η τιμή του είναι μικρότερη από το μέγιστο σφάλμα στρογγυλοποίησης, οι υπολογισμοί μπορεί να μην σταματούν, ακόμη και όταν η μέθοδος έχει πρακτικά συγκλίνει.

Επιπλέον, η τήρηση της παραπάνω πρακτικής κατά την ανάπτυξη ενός αλγορίθμου έχει ένα ακόμη σημαντικό πλεονέκτημα: η συμπεριφορά του υπολογιστικού κώδικα δεν εξαρτάται από το εκάστοτε υπολογιστικό σύστημα στο οποίο 'τρέχει'.

Εφαρμογή 1.3.

Να βρεθεί ο αριθμός των θέσεων (bits) που διαθέτει η mantissa ενός Η/Υ, για αριθμούς απλής και διπλής ακρίβειας.

Έστω k οι ζητούμενες θέσεις. Ένας αριθμός που απαιτεί k+1 θέσεις για να καταχωρηθεί με ακρίβεια, θα χάσει το τελευταίο του ψηφίο. Έτσι ο αριθμός
$x = [2^{-1} + 2^{-(k+1)}]$, που στο δυαδικό σύστημα γράφεται: $0.1000\ldots0|1$,
k
θα καταχωρηθεί ως $0.1000....0$, δηλαδή ίσος με 2^{-1}.

Σημειωτέον ότι το ίδιο θα προκύψει και με τους δύο τρόπους στρογγυλοποίησης. Με βάση αυτό το σκεπτικό, η εύρεση του k θα γίνει εκτελώντας διαδοχικές συγκρίσεις των αριθμών $[2^{-1} + 2^{-(n+1)}]$ με τον αριθμό $2^{-1} = 0.5$, αυξάνοντας κάθε φορά το n κατά 1, έως ότου οι δύο αριθμοί να προκύψουν, κατά τον υπολογιστή, ίσοι. Για τον σκοπό αυτόν χρησιμοποιείται ο παρακάτω αλγόριθμος.

Κώδικας 1.1. Εύρεση των bits της mantissa

```
basn = 0.5
DO  n = 1, 100
      reln = basn + 2.-(n+1)
      IF ( reln ≤ basn )  EXIT
END DO
k = n
errel = ( 2.-(k+1) ) / basn
END
```

Εφαρμόζοντας τον Κώδικα 1.1 σε έναν σύγχρονο 32-bit προσωπικό υπολογιστή (σε FORTRAN), λαμβάνεται για απλή ακρίβεια αριθμών, $k = 24$. Επομένως από τα 32 bits, τα 24 διατίθενται για τη βάση, τα επτά για τον εκθέτη και ένα για το πρόσημο του εκθέτη. Επειδή το πρώτο ψηφίο της κανονικοποιημένης βάσης είναι, στο δυαδικό σύστημα, πάντοτε 1, το αντίστοιχο bit χρησιμοποιείται για το πρόσημο της βάσης, χωρίς να μειώνεται η ακρίβειά της. Ο κώδικας υπολογίζει επίσης το μέγιστο σφάλμα για συμμετρική στρογγυλοποίηση *errel*, ίσο περίπου με $6 \cdot 10^{-8}$.

Ο ίδιος κώδικας, αλλά με αριθμούς διπλής ακρίβειας, δίνει $k = 53$ και σφάλμα *errel* $\approx 1 \cdot 10^{-16}$. Επομένως, διατίθενται 53 bits για τη βάση, δέκα για τον εκθέτη και ένα για το πρόσημό του. Στα δέκα bits του εκθέτη μπορούν να καταχωρηθούν 1024 ακέραιοι, επομένως ο μέγιστος δυνατός αριθμός που χωράει στη μνήμη είναι της τάξης του $2^{1024} = 10^{308}$.

Εφαρμογή 1.4.

Να βρεθεί η μέθοδος στρογγυλοποίησης που χρησιμοποιείται σε έναν Η/Υ.

Βρίσκεται πρώτα ο αριθμός των bits που χρησιμοποιούνται για τη βάση, έστω $k = 24$, όπως στην Εφαρμογή 1.3. Στη συνέχεια, με μία πρόσθετη εντολή του αλγορίθμου, υπολογίζεται ο αριθμός $x_t = [\, 2^{-(k+1)} + 2^{-(k+2)} + basn \,]$. Ο αριθμός αυτός, γραμμένος στο δυαδικό σύστημα, θα είναι $0.\underbrace{1000\ldots00}_{k}11$

Εάν στρογγυλοποιηθεί απλά, τότε και τα δύο τελευταία ψηφία του θα παραλειφθούν, οπότε θα καταχωρηθεί ως $x_a = basn$ (= 0.5). Στη συμμετρική στρογγυλοποίηση όμως θα είναι (βλ. και Εξ. 1.8)

$$x_t = f \cdot b^m + g \cdot b^{m-k} = 2^{-1} \cdot 2^0 + (2^{-1} + 2^{-2}) \cdot 2^{0-24} = 0.5 \cdot 2^0 + 0.75 \cdot 2^{-24}$$

Επομένως $|\, g \,| = 0.75 > 0.5$, και σύμφωνα με την εξίσωση ορισμού (1.11) προκύπτει

$$x_a = f \cdot b^m + b^{m-k} = 0.5 \cdot 2^0 + 2^{0-24} = 0.\underbrace{1000\ldots00}_{k}1_2 = 0.50000006_{10}$$

δηλαδή στο 24ο ψηφίο της mantissa θα προστεθεί το 1.

Έτσι, εάν η τιμή του αριθμού x_a που θα εκτυπώσει ο κώδικας είναι 0.50000006, τότε γίνεται συμμετρική στρογγυλοποίηση, ενώ εάν είναι 0.50000000, γίνεται απλή.

1.2.3. Σφάλμα Στρογγυλοποίησης σε Αριθμητικές Πράξεις

Στις περισσότερες μεθόδους αριθμητικής ανάλυσης χρησιμοποιούνται κυρίως οι τέσσερις αριθμητικές πράξεις, οι εκτέλεση των οποίων στην αριθμητική κινητής υποδιαστολής μπορεί να προκαλέσει πρόσθετο σφάλμα, όπως θα αναλυθεί στη συνέχεια.

Για να προστεθούν δύο αριθμοί, πρέπει πρώτα να γραφτούν έτσι ώστε να έχουν τον ίδιο εκθέτη, τον μεγαλύτερο εκ των δύο. Επομένως η βάση του αριθμού με τον μικρότερο εκθέτη παύει να είναι κανονικοποιημένη και κάποια σημαντικά ψηφία του μπορεί να χαθούν. Έστω για παράδειγμα ότι ζητείται η πρόσθεση των αριθμών $0.4658 \cdot 10^1$ και $0.3765 \cdot 10^{-1}$ στο δεκαδικό σύστημα και σε βάση τεσσάρων θέσεων, όπου και είναι καταχωρημένοι με ακρίβεια. Τότε, ο δεύτερος αριθμός θα γραφεί $0.0037 \cdot 10^1$, δηλαδή θα χάσει δύο σημαντικά ψηφία, και το αποτέλεσμα της πρόσθεσης θα είναι $(0.4658 + 0.0037) \cdot 10^1 = 0.4695 \cdot 10^1$, αντί του σωστού $0.469565 \cdot 10^1$.

Γενικά, κατά την πρόσθεση δύο αριθμών που διαφέρουν σημαντικά προκαλείται μια μικρή σχετικά απώλεια ακρίβειας, είναι όμως δυνατό να γίνει σημαντική, όταν εκτελούνται πολλές τέτοιες πράξεις σε έναν υπολογιστικό αλγόριθμο, όπως π.χ. στην Εφαρμογή 1.5.

Η αφαίρεση εκτελείται όπως και η πρόσθεση, μόνο που αλλάζει το πρόσημο του αφαιρετέου. Εδώ όμως προκαλείται ένα πρόσθετο πρόβλημα ακρίβειας, το οποίο μάλιστα επιτείνεται σημαντικά όσο μικρότερη είναι η διαφορά των δύο αριθμών. Έστω για παράδειγμα ότι γίνεται, στο δεκαδικό σύστημα και σε βάση τεσσάρων θέσεων, η αφαίρεση: $0.4658 \cdot 10^1 - 0.4593 \cdot 10^1$. Οι δύο αριθμοί έχουν τον ίδιο εκθέτη, επομένως το αποτέλεσμα θα είναι $0.0066 \cdot 10^1$. Επειδή όμως πρέπει να καταχωρηθεί σε κανονικοποιημένη μορφή, θα γίνει $0.6600 \cdot 10^1$. Εισάγονται δηλαδή δύο μηδενικά ψηφία, τα οποία θα θεωρηθούν σε επόμενες πράξεις ως σημαντικά χωρίς να είναι, πράγμα που μπορεί να προκαλέσει μεγάλο υπολογιστικό σφάλμα. Σε ένα πιο ακραίο παράδειγμα, έστω ότι ζητείται να γίνει η

αφαίρεση 4.65855 – 4.65845 και πάλι σε βάση τεσσάρων θέσεων. Οι δύο αριθμοί πρώτα θα στρογγυλοποιηθούν, έστω συμμετρικά, και θα γραφτούν στη μορφή $0.4659\cdot10^{1}$ – $0.4658\cdot10^{1}$, δίνοντας αποτέλεσμα $0.0001\cdot10^{1} = 0.1000\cdot10^{-2}$, ενώ το σωστό θα ήταν $0.1000\cdot10^{-3}$. Το σφάλμα δηλαδή που γίνεται είναι της τάξης του 900%.

Η απώλεια σημαντικών ψηφίων κατά την αφαίρεση δύο παραπλήσιων αριθμών είναι ένα από τα σημαντικότερα σφάλματα στρογγυλοποίησης και η τυχόν ύπαρξή του σε μια αριθμητική διαδικασία πρέπει να ελέγχεται και να αντιμετωπίζεται με διάφορες τεχνικές, όπως π.χ. στις Εφαρμογές 1.6 και 1.7.

Ένα άλλο αποτέλεσμα της στρογγυλοποίησης είναι ότι δεν ισχύει η προσεταιριστική ιδιότητα κατά την πρόσθεση. Έστω για παράδειγμα ότι ζητείται το άθροισμα των αριθμών $x_1 = 5675$, $x_2 = -5673$ και $x_3 = 3.457$, σε μια βάση τεσσάρων θέσεων. Τότε θα είναι

$$(x_1 + x_2) + x_3 = (0.5675\cdot10^{4} - 0.5673\cdot10^{4}) + 0.3457\cdot10^{1} = 0.0002\cdot10^{4} + 0.3457\cdot10^{1} =$$
$$= 0.2000\cdot10^{1} + 0.3457\cdot10^{1} = 0.5457\cdot10^{1} = 5.457, \quad \text{ενώ}$$
$$x_1 + (x_2 + x_3) = 0.,5675\cdot10^{4} + (-0.5673\cdot10^{4} + 0.0003\cdot10^{4}) = 0.5675\cdot10^{4} - 0.5670\cdot10^{4} =$$
$$= 0.0005\cdot10^{4} = 5.000$$

Επομένως, $(x_1 + x_2) + x_3 \neq x_1 + (x_2 + x_3)$. Γενικότερα, το αποτέλεσμα μιας αριθμητικής παράστασης εξαρτάται από τη σειρά εκτέλεσης των πράξεων, ενώ το τελικό σφάλμα στρογγυλοποίησης μπορεί να περιορισθεί εάν προστεθούν πρώτα οι αριθμοί που δεν διαφέρουν σημαντικά μεταξύ τους (όπως π.χ. στην Εφαρμογή 1.8).

Ο πολλαπλασιασμός δύο αριθμών γίνεται αθροίζοντας τους εκθέτες τους και πολλαπλασιάζοντας τις βάσεις. Επειδή όμως ο πολλαπλασιασμός δύο βάσεων με k ψηφία παράγει γενικά αποτέλεσμα με $2k$ ψηφία, συνήθως το ενδιάμεσο αποτέλεσμα εγγράφεται πρώτα σε έναν καταχωρητή διπλάσιου μήκους, όπου και κανονικοποιείται. Στη συνέχεια στρογγυλοποιείται σε βάση k ψηφίων και αποθηκεύεται στη μνήμη του υπολογιστή. Ανάλογα γίνεται και η διαίρεση, όπου όμως αφαιρούνται οι εκθέτες και διαιρούνται οι βάσεις. Για παράδειγμα, σε μία βάση τριών ψηφίων, το γινόμενο $0.123\cdot10^{1}$ x $0.432\cdot10^{2}$ είναι $0.053136\cdot10^{3}$. Το ενδιάμεσο αυτό αποτέλεσμα κανονικοποιείται σε $0.53136\cdot10^{2}$ και στρογγυλοποιείται, δίνοντας τελικά $0.531\cdot10^{2}$.

Τονίζεται ότι, επειδή σε πολλές μεθόδους αριθμητικής ανάλυσης απαιτείται η εκτέλεση μεγάλου αριθμού διαδοχικών πράξεων, παρόλο που το σφάλμα κάθε πράξης μπορεί να είναι μικρό, η συσσώρευση τέτοιων σφαλμάτων κατά τη διάρκεια των υπολογισμών μπορεί να προκαλέσει σημαντικό σφάλμα στο τελικό αποτέλεσμα. Γενικότερα, το τελικό σφάλμα στρογγυλοποίησης μιας μεθόδου περιορίζεται εάν μειωθεί ο αριθμός των αριθμητικών πράξεων που εκτελούνται (όπως π.χ. στην Εφαρμογή 1.9).

Τέλος, ο απλούστερος τρόπος ελαχιστοποίησης των σφαλμάτων στρογγυλοποίησης είναι η χρήση αριθμών διπλής ακρίβειας (double precision) σε όλες τις πράξεις ή σε εκείνες που εκτιμάται ότι είναι σημαντικό. Όμως αυτή η πρακτική απαιτεί διπλάσια υπολογιστική μνήμη και αυξημένο χρόνο υπολογισμών, έτσι δεν μπορεί να ακολουθείται αβίαστα, ιδίως σε μεγάλους αλγορίθμους, οι οποίοι χρειάζονται πολλή υπολογιστική μνήμη και η ταχύτητα εκτέλεσης αποτελεί σημαντικό χαρακτηριστικό τους. Πάντως, σε αρκετά εμπορικά πακέτα που έχουν ενσωματωμένες μεθόδους αριθμητικής ανάλυσης (π.χ. Excel, Mathcad, Matlab), χρησιμοποιείται αριθμητική διπλής ακρίβειας.

Εφαρμογή 1.5.

Να βρεθεί το αποτέλεσμα της πρόσθεσης N φορές του αριθμού 0.125_{10}, καθώς και του αριθμού 0.1_{10}, σε έναν 32-bit H/Y, με απλή και διπλή ακρίβεια.

Ο αριθμός 0.125 κωδικοποιείται στο δυαδικό σύστημα με ακρίβεια, αφού ισχύει $0.125 = 2^{-3}$, ενώ αντίθετα ο αριθμός 0.1 γράφεται, όπως προαναφέρθηκε, ως 0.000110011001100..., δηλαδή δεν έχει πεπερασμένο αριθμό ψηφίων. Έτσι η καταχώρησή του θα ενέχει πάντα σφάλμα στρογγυλοποίησης, τόσο μεγαλύτερο, όσο λιγότερες θέσεις μνήμης διατίθενται.

Πράγματι, εκτέλεση των υπολογισμών με έναν απλό κώδικα σε FORTRAN, δίνει τα εξής αποτελέσματα:

Πίνακας 1.1. Αποτελέσματα πρόσθεσης των αριθμών 0.1 και 0.125.

	Απλή ακρίβεια αριθμών		Διπλή ακρίβεια
N	$x = 0.1$	$x = 0.125$	$x = 0.1$
10	$0.100000012 \cdot 10^{1}$	$0.1250000 \cdot 10^{1}$	$0.999999999 \cdot 10^{0}$
10^3	$0.999990463 \cdot 10^{2}$	$0.1250000 \cdot 10^{3}$	$0.999999999 \cdot 10^{2}$
10^5	$0.999855664 \cdot 10^{4}$	$0.1250000 \cdot 10^{5}$	$0.100000000 \cdot 10^{5}$
10^7	$0.108793700 \cdot 10^{7}$	$0.1250000 \cdot 10^{7}$	$0.999999998 \cdot 10^{6}$
10^9	$0.209715200 \cdot 10^{8}$	$0.2097152 \cdot 10^{8}$	$0.999999987 \cdot 10^{8}$

Κατά την πρόσθεση του αριθμού 0.1, παρατηρείται ότι το σφάλμα στρογγυλοποίησης για απλή ακρίβεια αριθμών αυξάνει με τον αριθμό των πράξεων, N, φθάνοντας για $N = 10^7$ περίπου στο 8.8 % (σχετικό σφάλμα). Αντίθετα, η πρόσθεση του 0.125 γίνεται με ακρίβεια, όπως αναμενόταν. Και στις δύο περιπτώσεις όμως, το άθροισμα δεν μπορεί να ξεπεράσει την τιμή $0.2907152 \cdot 10^{8}$. Αυτό συμβαίνει επειδή κατά την πρόσθεση του $0.1 = 0.000000001 \cdot 10^{8}$ στον μεγάλο πλέον αυτόν αριθμό, χάνεται κατά την στρογγυλοποίηση το τελευταίο ψηφίο, οπότε οι πράξεις από εδώ και πέρα είναι σαν να μην εκτελούνται καθόλου ή σαν να προστίθεται το μηδέν.

Όταν χρησιμοποιείται διπλή ακρίβεια αριθμών, το αποτέλεσμα είναι ακριβές και για τους δύο αριθμούς, ακόμη και για πολύ μεγάλο αριθμό πράξεων N.

Εφαρμογή 1.6.

Να υπολογισθεί η τιμή της συνάρτησης $f(x) = 1 - \cos x$, για $x = 0.01$ rad, σε μια βάση πέντε ψηφίων και με απλή στρογγυλοποίηση.

Οι αριθμοί 1 και $\cos x$ θα γραφούν αντιστοίχως: $0.10000 \cdot 10^{1}$ και $0.99995 \cdot 10^{0}$. Όμως για να γίνει η αφαίρεση, ο δεύτερος αριθμός θα γραφεί ως $0.09999 \cdot 10^{1}$, δηλαδή με τον μεγαλύτερο εκθέτη, αυτόν του πρώτου αριθμού. Έτσι, το αποτέλεσμα θα είναι $f(0.01) = 0.00001 \cdot 10^{1} = 0.1 \cdot 10^{-3}$.

Επομένως, λαμβάνοντας υπόψη ότι το ακριβές αποτέλεσμα είναι $0.499995 \cdot 10^{-4}$, το σχετικό σφάλμα που γίνεται στη βάση των πέντε ψηφίων είναι 100%.

Η συνάρτηση $f(x)$ μπορεί όμως να γραφεί και με τη μορφή $f(x) = 2\sin^2(x/2)$. Τότε, η τιμή της για $x = 0.01$ θα υπολογισθεί ως εξής:

$$f(0.01) = 0.20000\cdot10^{1} \cdot (0.49999\cdot10^{-2} \cdot 0.49999\cdot10^{-2})$$
$$= 0.20000\cdot10^{1} \cdot 0.24999\cdot10^{-4} = 0.49998\cdot10^{-4}.$$

Δηλαδή, αλλάζοντας τη μαθηματική έκφραση της συνάρτησης, ώστε να μην περιέχει πράξη αφαίρεσης, το σφάλμα ελαχιστοποιείται, ακόμη και για πολύ μικρή τιμή του x.

Εφαρμογή 1.7.

Να υπολογιστούν οι ρίζες της εξίσωσης $x^2 + 1999.999\,x - 2 = 0$.

Οι ρίζες της δευτεροβάθμιας εξίσωσης υπολογίζονται πρώτα από τη γνωστή σχέση

$$x_1,\ x_2 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{1.14}$$

με απλή ακρίβεια αριθμών (32-bit), οπότε προκύπτει $x_1 = 0.9765625\cdot10^{-3}$, $x_2 = -0.2\cdot10^{4}$, ενώ η ακριβής λύση είναι $x_1 = 0.001$, $x_2 = -2000$. Επομένως, η πρώτη ρίζα υπολογίζεται με σχετικό σφάλμα περίπου 2.34%. Αυτό συμβαίνει επειδή στον αριθμητή της παραπάνω σχέσης εμφανίζεται η διαφορά

$$-b + \sqrt{b^2 - 4ac} = -1999.99902 + 2000.00098$$

που δίνει 0.00196 αντί για τη σωστή τιμή 0.002. Το σφάλμα αυτό προκύπτει από τη στρογγυλοποίηση κατά την καταχώρηση των τιμών του συντελεστή b και της διακρίνουσας, και ενισχύεται λόγω της αφαίρεσης δύο παραπλήσιων αριθμών. Στη δεύτερη ρίζα της εξίσωσης οι δύο αριθμοί προστίθενται (έχουν ίδιο πρόσημο), επομένως δεν προκαλείται απώλεια ακρίβειας λόγω αφαίρεσης, και η ρίζα υπολογίζεται χωρίς σφάλμα.

Μία τεχνική για βελτίωση της ακρίβειας είναι η εξής: Ο αριθμητικής και ο παρονομαστής της Εξ. (1.14) πολλαπλασιάζονται επί $b \pm \sqrt{b^2 - 4ac}$, οπότε προκύπτει

$$x_1,\ x_2 = \frac{-2c}{b \pm \sqrt{b^2 - 4ac}} \tag{1.15}$$

Εάν τώρα χρησιμοποιηθεί η νέα αυτή σχέση, θα προκύψουν οι εξής τιμές των ριζών: $x_1 = 0.10000000\cdot10^{-2}$, $x_2 = -0.2048\cdot10^{4}$, δηλαδή η πρώτη ρίζα δεν έχει πλέον σφάλμα, επειδή ο συντελεστής b και η διακρίνουσα προστίθενται, ενώ το αντίθετο ισχύει για τη δεύτερη ρίζα, που έχει σχετικό σφάλμα 2.4%.

Επομένως, η ακρίβεια κατά τον υπολογισμό των ριζών μιας δευτεροβάθμιας εξίσωσης μπορεί γενικά να βελτιωθεί εάν χρησιμοποιούνται κατάλληλα οι δύο παραπάνω εκφράσεις: η πρώτη ρίζα θα υπολογίζεται από την (1.14) ή την (1.15), αναλόγως εάν ο συντελεστής b είναι αρνητικός ή θετικός αντιστοίχως, ενώ η δεύτερη ρίζα το αντίθετο.

Εφαρμογή 1.8.

Να βρεθεί η τιμή της σειράς $f(N)=\sum_{n=1}^{N}\frac{1}{n^2}$ για $N = 10^2, 10^3, 10^4$ και 10^5, αθροίζοντας διαδοχικά τους όρους με την κανονική και με αντίστροφη σειρά (από τον τελευταίο προς τον πρώτο).

Για $N \to \infty$ η σειρά συγκλίνει στην τιμή $f(N) = \pi^2/6 = 1.64493407$. Μετά την εκτέλεση των αριθμητικών υπολογισμών, τα αποτελέσματα με διπλή ακρίβεια αριθμών προκύπτουν ίδια και με τους δύο τρόπους άθροισης, ενώ με απλή ακρίβεια διαφέρουν, όπως φαίνεται στον Πίνακα 1.2.

Πίνακας 1.2. Αποτελέσματα υπολογισμού της σειράς $f(N)$.

	Απλή ακρίβεια αριθμών		Διπλή ακρίβεια
N	$n = 1 \to N$	$n = N \to 1$	$1 \to N \quad \& \quad N \to 1$
10^2	1.63498402	1.63498390	1.63498390
10^3	1.64393485	1.64393449	1.64393457
10^4	1.64472532	1.64483404	1.64483407
10^5	1.64472532	1.64492404	1.64492406

Για απλή ακρίβεια παρατηρείται ότι, ενώ μέχρι περίπου $N = 10^3$ το αποτέλεσμα είναι πρακτικά ακριβές, στη συνέχεια η άθροιση των όρων με την κανονική σειρά δίνει αποτέλεσμα που αποκλίνει από τη σωστή τιμή. Αυτό συμβαίνει επειδή οι όροι της σειράς που προστίθενται μειώνονται συνεχώς, ώσπου η στρογγυλοποίηση του αποτελέσματος ακυρώνει την πρόσθεση, όπως και στην Εφαρμογή 1.5.

Αντίθετα, όταν η άθροιση γίνει με αντίστροφη σειρά, επειδή οι νέοι όροι που προστίθενται αυξάνουν συνεχώς και δεν διαφέρουν πολύ από την τρέχουσα τιμή του αθροίσματος, το σφάλμα στρογγυλοποίησης ελαχιστοποιείται και το αποτέλεσμα είναι ίδιο με εκείνο της διπλής ακρίβειας αριθμών. Έτσι, γίνεται δυνατή η επίτευξη ακριβούς λύσης χωρίς τη χρήση διπλής ακρίβειας, δηλαδή με μικρότερες υπολογιστικές απαιτήσεις.

Εφαρμογή 1.9.

Να υπολογισθεί η τιμή του πολυωνύμου $f(x) = -4.5 + 2x - 4x^2 + x^3$, για $x = 3.84$, υποθέτοντας ότι η βάση (mantissa) έχει τρία ψηφία και ότι γίνεται απλή στρογγυλοποίηση.

Με απλή στρογγυλοποίηση όλων των αριθμών στη βάση των τριών ψηφίων, θα είναι:

$2x = 0.200\cdot10^1 \cdot 0.384\cdot10^1 = 0.768\cdot10^1$.

$4x^2 = 0.400\cdot10^1 \cdot 0.147456\cdot10^2 = 0.400\cdot10^1 \cdot 0.147\cdot10^2 = 0.0588\cdot10^3$

$= 0.588\cdot10^2$

$x^3 = x\cdot(x^2) = 0.384\cdot10^1 \cdot 0.147\cdot10^2 = 0.056448\cdot10^3 = 0.564\cdot10^2$,

Έτσι η τιμή του πολυωνύμου θα προκύψει

$$f(3.84) = (-0.450\cdot10^1 + 0.768\cdot10^1) - 0.588\cdot10^2 + 0.564\cdot10^2 =$$
$$= (0.318\cdot10^1 - 0.588\cdot10^2) + 0.564\cdot10^2 = (0.031\cdot10^2 - 0.588\cdot10^2)$$
$$+ 0.564\cdot10^2 = -0.557\cdot10^2 + 0.564\cdot10^2 = 0.007\cdot10^2 = 0.7$$

Η ακριβής λύση είναι $f_t(3.84) = 0.820704$, επομένως ο αριθμητικός υπολογισμός θα δώσει σχετικό σφάλμα περίπου 14.7 %. Ένας τρόπος μείωσης του σφάλματος αυτού είναι η ελάττωση του αριθμού των πράξεων, που επιτυγχάνεται γράφοντας το πολυώνυμο στη μορφή $f(x) = -4.5 + x\,[2 + x\,(-4 + x)]$. Εκτελώντας τις πράξεις όπως και πριν, θα έχουμε:

$$x\cdot(-4 + x) = 0.384\cdot10^1\cdot(-0.400\cdot10^1 + 0.384\cdot10^1)$$
$$= 0.384\cdot10^1\cdot(-0.160\cdot10^0) = -0.06144\cdot10^1 = -0.614\cdot10^0.$$

Οπότε

$$f(3.84) = -0.450\cdot10^1 + 0.384\cdot10^1\cdot(0.200\cdot10^1 - 0.061\cdot10^1)$$
$$= -0.450\cdot10^1 + 0.384\cdot10^1 \cdot 0.139\cdot10^1 = -0.450\cdot10^1 + 0.053376\cdot10^2$$
$$= -0.450\cdot10^1 + 0.533\cdot10^1 = 0.083\cdot10^1 = 0.83.$$

Το σχετικό σφάλμα επομένως μειώθηκε σημαντικά, σε περίπου 1.1 %.

Γενικά, ένα πολυώνυμο μπορεί να γραφεί και στη μορφή

$$f(x) = a_0 + x(a_1 + x(a_2 + \ldots x(a_{n-1} + x\,a_n)\ldots)) \qquad (1.16)$$

Τότε, για τον υπολογισμό της τιμής του σε δεδομένο x, απαιτούνται μόνο n–1 πολλαπλασιασμοί και n προσθέσεις, ενώ στην κλασική μορφή θα εκτελεστούν $n(n+1)/2$ πολλαπλασιασμοί και n προσθέσεις. Για n=8 π.χ., στην κλασική μορφή θα γίνουν 36 πολλαπλασιασμοί ενώ στη μορφή (1.16) μόνο 7, άρα το σφάλμα στρογγυλοποίησης θα είναι μικρότερο.

1.3. Σφάλμα Αποκοπής

Στις αριθμητικές μεθόδους πολλές φορές δεν χρησιμοποιείται η πλήρης μαθηματική έκφραση μιας παράστασης, συνήθως όταν αυτή περιέχει πολλούς ή άπειρους όρους. Η παράλειψη των μικρότερων όρων παρέχει μια πιο εύχρηστη, αλλά προσεγγιστική μαθηματική έκφραση, το σφάλμα της οποίας ονομάζεται *σφάλμα αποκοπής*.

Για παράδειγμα, εάν για τη σειρά της Εφαρμογής 1.8, η οποία συγκλίνει στην τιμή 1.64493407, χρησιμοποιηθούν μόνο οι πέντε πρώτοι όροι, το αποτέλεσμα θα είναι

$$f(N) = 1 + 1/2^2 + 1/3^2 + 1/4^2 + 1/5^2 = 1.46361111$$

Επομένως προκαλείται σχετικό σφάλμα αποκοπής περίπου 11%. Σε ένα άλλο παράδειγμα, το κλάσμα $1/(1-\delta)$ μπορεί να γραφεί προσεγγιστικά ως $(1+\delta)$, όταν είναι $\delta << 1$:

$$\frac{1}{1-\delta} = (1+\delta) + \left(\frac{\delta^2}{1-\delta}\right) \approx (1+\delta)$$

επειδή ο δεύτερος όρος είναι πολύ μικρότερος του πρώτου. Το απόλυτο σφάλμα αποκοπής θα είναι ίσο με τον όρο που παραλείπεται, π.χ. για $\delta = 0.01$ θα είναι $E_t \cong 1 \cdot 10^{-4}$.

1.3.1. Σειρά Taylor

Κάθε συνάρτηση $f(x)$ που είναι συνεχής στο διάστημα $[a, b]$ και έχει συνεχείς παραγώγους τάξης μέχρι και $n+1$ στο διάστημα αυτό, μπορεί να παρασταθεί με μια *σειρά Taylor*, που έχει τη μορφή

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \ ... + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_n \qquad (1.17)$$

για κάθε x στο $[a, b]$. Το υπόλοιπο, R_n, αντιπροσωπεύει όλους τους (άπειρους) όρους που ακολουθούν και εκφράζεται με τη σχέση

$$R_n = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-a)^{n+1} \qquad (1.18)$$

όπου $\xi \in [a, x]$ είναι ένας συγκεκριμένος αριθμός, που εξαρτάται από το n. Όταν είναι $a = 0$, τότε η σειρά ονομάζεται συχνά και *σειρά Maclaurin*.

Εάν για τον υπολογισμό μιας συνάρτησης χρησιμοποιηθούν μόνο οι όροι μέχρι τάξης n του αναπτύγματός της σε σειρά Taylor, τότε η τιμή της συνάρτησης βρίσκεται κατά προσέγγιση, ενώ το (απόλυτο) σφάλμα αποκοπής θα ισούται με το υπόλοιπο R_n. Στην περίπτωση αυτή, από μία γνωστή τιμή της συνάρτησης και των παραγώγων της σε ένα σημείο x_i, μπορεί να προσεγγισθεί η τιμή της σε ένα άλλο σημείο x_{i+1}. Συνήθως η απόσταση των δύο σημείων ονομάζεται *βήμα* και συμβολίζεται με $h = (x_{i+1} - x_i)$, οπότε από την Εξ. (1.17) προκύπτει

$$f(x_{i+1}) \cong f(x_i) + f'(x_i)h + \frac{f''(x_i)}{2!}h^2 + \frac{f^{(3)}(x_i)}{3!}h^3 + \ ... + \frac{f^{(n)}(x_i)}{n!}h^n \qquad (1.19)$$

ενώ ο όρος του υπολοίπου ή το σφάλμα αποκοπής θα είναι

$$R_n = \frac{f^{(n+1)}(\xi)}{(n+1)!}\, h^{n+1} \tag{1.20}$$

Η έκφραση (1.19) ονομάζεται *προσέγγιση n τάξης* της $f(x_{i+1})$, όταν περιλαμβάνει όρους μέχρι και τάξης n (ο πρώτος όρος είναι μηδενικής τάξης). Αντίστοιχα, το μέγεθος του σφάλματος αποκοπής εκτιμάται σε $R_n = O(h^{n+1})$, δηλαδή της τάξης του h^{n+1}.

Η μεγάλη πρακτική αξία αυτής της μεθόδου έγκειται στο γεγονός ότι, στις περισσότερες περιπτώσεις, αρκούν λίγοι μόνο όροι της σειράς για να προσεγγίσουν την πραγματική τιμή της συνάρτησης με ικανοποιητική ακρίβεια, εφόσον το βήμα h διατηρείται σχετικά μικρό. Επιπλέον, ρυθμίζοντας κατάλληλα την τιμή του βήματος, μπορεί να επιτευχθεί η επιθυμητή κάθε φορά ακρίβεια του αποτελέσματος. Το ανάπτυγμα σε σειρά Taylor χρησιμοποιείται επίσης για την εύρεση προσεγγιστικών εκφράσεων των παραγώγων μιας συνάρτησης (βλ. Κεφ. 5.2).

Εφαρμογή 1.10.

Να υπολογισθεί η τιμή της συνάρτησης $f(x) = \sin x$ στο σημείο $x_{i+1} = \pi/2$, με βάση την τιμή της και τις τιμές των παραγώγων της στη θέση $x_i = \pi/4$ ή στη θέση $x_i = 0$, χρησιμοποιώντας όρους του αναπτύγματος σε σειρά Taylor, μέχρι και έβδομης τάξης.

Η προσέγγιση έβδομης τάξης, με βάση το $x_i = \pi/4$, οπότε $h = \pi/2 - \pi/4 = \pi/4$, είναι

$$f_1\left(\frac{\pi}{2}\right) \cong \sin\left(\frac{\pi}{4}\right) + \cos\left(\frac{\pi}{4}\right)h - \frac{1}{2}\sin\left(\frac{\pi}{4}\right)h^2 - \frac{1}{3!}\cos\left(\frac{\pi}{4}\right)h^3 + \frac{1}{4!}\sin\left(\frac{\pi}{4}\right)h^4 + \frac{1}{5!}\cos\left(\frac{\pi}{4}\right)h^5 - \frac{1}{6!}\sin\left(\frac{\pi}{4}\right)h^6 - \frac{1}{7!}\cos\left(\frac{\pi}{4}\right)h^7$$

ενώ με βάση το $x_i = 0$, οπότε $h = \pi/2$, οι οκτώ πρώτοι όροι θα είναι

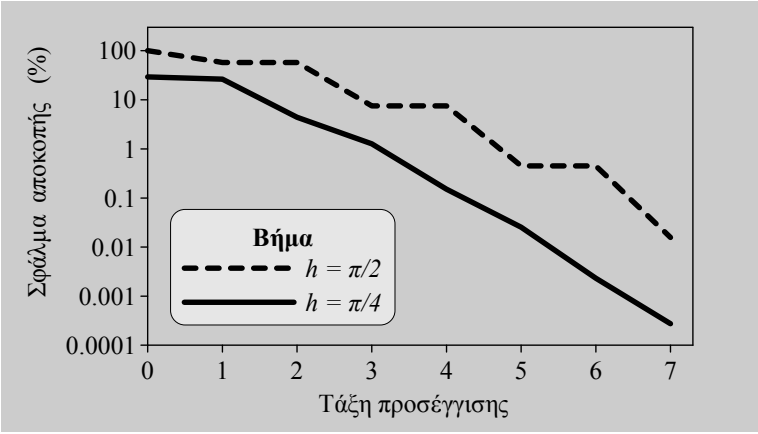
$$f_2\left(\frac{\pi}{2}\right) \cong 0.0 + h - 0.0 - \frac{1}{3!}h^3 + 0.0 + \frac{1}{5!}h^5 - 0.0 - \frac{1}{7!}h^7$$

Στον Πίνακα 1.3 συγκεντρώνονται τα αποτελέσματα των υπολογισμών με τις δύο παραπάνω εκφράσεις, καθώς και τα αντίστοιχα πραγματικά σφάλματα (απόλυτο = σχετικό σφάλμα, επειδή η ακριβής τιμή της συνάρτησης είναι $\sin(\pi/2) = 1$). Οι πράξεις εκτελούνται με αριθμούς διπλής ακρίβειας, ώστε να μην υπεισέρχεται σφάλμα στρογγυλοποίησης.

Πίνακας 1.3. Προσέγγιση της τιμής της συνάρτησης $f(x_{i+1}) = \sin(\pi/2)$

	$x_i = \pi/4$,	$h_1 = \pi/4$	$x_i = 0$,	$h_2 = \pi/2$
τάξη	$f_1(x_{i+1})$	$\varepsilon_{t,1}$ (%)	$f_2(x_{i+1})$	$\varepsilon_{t,2}$ (%)
0	0.70710678	29.29	0.00000000	100.0
1	1.26246710	–26.25	1.57079630	–57.08
2	1.04437764	–4.44	1.57079630	–57.08
3	0.98728194	1.27	0.92483223	7.52
4	0.99849266	$1.51\cdot10^{-1}$	0.92483223	7.52
5	1.00025363	$-2.54\cdot10^{-2}$	1.00452490	$-4.52\cdot10^{-1}$
6	1.00002312	$-2.32\cdot10^{-3}$	1.00452490	$-4.52\cdot10^{-1}$
7	0.99999726	$2.74\cdot10^{-4}$	0.99984310	$1.57\cdot10^{-2}$

Και στις δύο περιπτώσεις, το σφάλμα αποκοπής μειώνεται μετά την προσθήκη κάθε επόμενου όρου της σειράς, ενώ αρκούν οι πρώτοι πέντε – έξι όροι, ώστε η ακρίβεια του αποτελέσματος να είναι ικανοποιητική. Επομένως, ο υπολογισμός περισσότερων όρων δεν προσφέρει πλέον σημαντική βελτίωση του αποτελέσματος, ενώ αντίθετα αυξάνει τις υπολογιστικές απαιτήσεις.



Σχήμα 1.2. Επίδραση του βήματος στο σφάλμα αποκοπής.

Στα αποτελέσματα του Πίνακα 1.3 είναι επίσης φανερό ότι η πρώτη έκφραση με το μικρότερο βήμα ($h = \pi/4$) απαιτεί λιγότερους όρους από τη δεύτερη (που έχει διπλάσιο βήμα), για ίδια ακρίβεια. Το σφάλμα αποκοπής συγκρίνεται και γραφικά στο Σχήμα 1.2: ο ρυθμός μείωσης του σφάλματος είναι πολύ μεγαλύτερος για το μικρότερο βήμα. Αυτό συμβαίνει επειδή το μέγεθος του σφάλματος είναι ανάλογο του h^{n+1}, επομένως ο λόγος των δύο σφαλμάτων θα είναι περίπου

$$\left(\frac{h_1}{h_2}\right)^{n+1} = \left(\frac{\pi/4}{\pi/2}\right)^{n+1} = 0.5^{n+1}$$

δηλαδή θεωρητικά θα υποδιπλασιάζεται μετά από κάθε προσθήκη ενός όρου μεγαλύτερης τάξης. Πράγματι, για $n = 0$, ο λόγος είναι $\varepsilon_{t,1}/\varepsilon_{t,2} \approx 0.3$, ενώ για $n = 6$ είναι $\approx 5\cdot10^{-3}$ (Πίνακας 1.3), δηλαδή περίπου της τάξης του 0.5^1 και 0.5^7, αντιστοίχως.

Τέλος, να σημειωθεί ότι το δεύτερο ανάπτυγμα (σειρά Maclaurin) θα μπορούσε να γραφεί και χωρίς τους μηδενικούς όρους, ως εξής:

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

δηλαδή με τέσσερις μόνο όρους. Παρατηρείται λοιπόν ότι η θέση ενός όρου στο ανάπτυγμα μιας σειράς Taylor μπορεί να μην αντιστοιχεί στην τάξη του, επομένως οι εκφράσεις: "οι n πρώτοι όροι" και "οι όροι έως n τάξης" δεν είναι ταυτόσημες.

1.3.2. Εκτίμηση του Σφάλματος Αποκοπής

Σε πολλές αριθμητικές μεθόδους χρησιμοποιούνται απλές συναρτήσεις, οι οποίες προσεγγίζουν σύνθετες ή μη αναλυτικές εκφράσεις. Το ανάπτυγμα σε σειρά Taylor εφαρμόζεται συχνά για την εκτίμηση του σφάλματος αποκοπής των προσεγγιστικών αυτών λύσεων.

Ο όρος του υπολοίπου της σειράς Taylor (Εξ. 1.20) ισούται με το απόλυτο σφάλμα αποκοπής κατά την προσέγγιση μιας συνάρτησης, μόνο για μια συγκεκριμένη τιμή του ξ στο διάστημα $[x_i, x_{i+1}]$. Για παράδειγμα, η προσέγγιση μηδενικής τάξης της συνάρτησης $\sin x$ στην Εφαρμογή 1.10, έχει απόλυτο σφάλμα $E_t = 0.2929$ (Πίνακας 1.3), οπότε

$$R_0 = f'(\xi)\, h_1 = \cos(\xi)(\pi/4) \cong 0.2929 \;\Rightarrow\; \xi \cong \pi/2.643$$

ενώ η προσέγγιση τρίτης τάξης έχει σφάλμα $E_t = 0.0127$, οπότε

$$R_3 = \frac{f^{(4)}(\xi)}{(3+1)!}\, h^{3+1} = \frac{1}{24}\sin(\xi)\left(\frac{\pi}{4}\right)^4 \cong 0.0127 \;\Rightarrow\; \xi \cong \pi/3.382$$

Η εκάστοτε τιμή του ξ όμως δεν είναι γνωστή, επομένως ο ακριβής υπολογισμός του σφάλματος αποκοπής δεν είναι δυνατός. Παρόλα αυτά, η τάξη μεγέθους του σφάλματος, που εκφράζεται ως $R_n = O(h^{n+1})$, αποτελεί σημαντική πληροφορία όταν πρόκειται να συγκριθεί η ακρίβεια διαφορετικών παραστάσεων της ίδιας συνάρτησης (βλ. Κεφ. 5.2) ή να εκτιμηθεί η επίδραση που έχει στην ακρίβεια μιας παράστασης το μέγεθος του βήματος h (βλ. Εφαρμογή 1.10).

Αναλυτικός υπολογισμός του R_n μπορεί να γίνει όταν οι παράγωγοι μηδενίζονται από κάποια τάξη και πάνω, όπως συμβαίνει σε μια πολυωνυμική συνάρτηση, ή όταν η σειρά συγκλίνει σε κάποια γνωστή τιμή. Σ' αυτές τις περιπτώσεις όμως, η τιμή της συνάρτησης σε κάποια θέση x μπορεί να βρεθεί άμεσα, επομένως δεν χρειάζεται να γίνει προσέγγιση.

Αναφορικά τώρα με τη σχέση που έχει η εκτίμηση της τάξης μεγέθους του σφάλματος με το πραγματικό σφάλμα αποκοπής, μπορούν να γίνουν οι ακόλουθες παρατηρήσεις. Η τάξη μεγέθους του σφάλματος εκφράζει μόνο την οριακή συμπεριφορά του όταν $h \to 0$ και δεν να είναι ανάλογη με το πραγματικό σφάλμα. Έτσι είναι δυνατόν το σφάλμα αποκοπής να μην μηδενίζεται για $h = (x - x_i) \to 0$, όταν οι παράγωγοι ανώτερης τάξης δεν είναι φραγμένες στη θέση x_i. Γενικά, όταν η τιμή των παραγώγων ανώτερης τάξης μιας συνάρτησης μειώνεται ή μένει σταθερή, τότε το μέγεθος του πραγματικού σφάλματος αποκοπής E_t πέφτει κάτω από την εκτιμώμενη τάξη μεγέθους μετά την προσθήκη λίγων μόνο όρων της σειράς Taylor. Αντίθετα, όταν η τιμή των παραγώγων ανώτερης τάξης αυξάνει συνεχώς, τότε το πραγματικό σφάλμα μπορεί να είναι πολύ μεγαλύτερο από το εκτιμώμενο, ιδίως για τους πρώτους όρους της σειράς. Με άλλα λόγια, η εκτιμώμενη τάξη μεγέθους του σφάλματος μπορεί να χρησιμοποιείται με ασφάλεια μόνο για σχετική και όχι για απόλυτη εκτίμηση του μεγέθους του πραγματικού σφάλματος αποκοπής.

Εφαρμογή 1.11.

Να υπολογισθεί με ανάπτυγμα σε σειρά Taylor η τιμή των συναρτήσεων $f_1(x) = e^x$ και $f_2(x) = e^{3x}$ στο σημείο $x_{i+1} = 1.1$, με βάση την τιμή τους στη θέση $x_i = 1$, και να συγκριθεί το εκτιμώμενο με το πραγματικό σφάλμα αποκοπής.

Η προσέγγιση n τάξης των δύο συναρτήσεων, με βήμα $h = 1.1 - 1 = 0.1$, θα είναι:

$$f_1(1.1) \cong e + 0.1e + \frac{0.1^2}{2!}e + \frac{0.1^3}{3!}e + ... + \frac{0.1^n}{n!}e$$

$$f_2(1.1) \cong e^3 + 0.1\cdot 3e^3 + \frac{0.1^2}{2!}3^2 e^3 + \frac{0.1^3}{3!}3^3 e^3 + ... + \frac{0.1^n}{n!}3^n e^3$$

Εκτελώντας τους υπολογισμούς των παραπάνω εκφράσεων για διάφορες τιμές του n και με διπλή ακρίβεια αριθμών, λαμβάνονται τα ακόλουθα αποτελέσματα:

Πίνακας 1.4. Προσέγγιση διαφόρων τάξεων n των συναρτήσεων e^x και e^{3x}

τάξη	$f_1(x_{i+1})$	$E_{t,1}$	$f_2(x_{i+1})$	$E_{t,2}$	E_a
0	2.71828183	0.286	20.0855369	7.027	0.1
1	2.99011001	$0.145\cdot 10^{-1}$	26.1111980	1.001	$0.1\cdot 10^{-1}$
2	3.00370142	$0.46\cdot 10^{-3}$	27.0150472	$0.97\cdot 10^{-1}$	$0.1\cdot 10^{-2}$
3	3.00415447	$0.12\cdot 10^{-4}$	27.1054321	$0.72\cdot 10^{-2}$	$0.1\cdot 10^{-3}$
4	3.00416579	$0.23\cdot 10^{-6}$	27.1122909	$0.43\cdot 10^{-3}$	$0.1\cdot 10^{-4}$
5	3.00416602	$0.38\cdot 10^{-8}$	27.1126177	$0.21\cdot 10^{-4}$	$0.1\cdot 10^{-5}$
.....	.....	.....	.....	.....	.....
8	3.00416602	$0.8\cdot 10^{-14}$	27.1126389	$0.11\cdot x10^{-8}$	$0.1\cdot 10^{-8}$

Η τελευταία στήλη του Πίνακα 1.4 περιέχει την εκτιμώμενη τάξη μεγέθους του σφάλματος αποκοπής, $O(h^{n+1})$, που είναι ίδια και για τις δύο συναρτήσεις (ίδιο h). Όπως παρατηρείται, το πραγματικό σφάλμα αποκοπής της πρώτης συνάρτησης, f_1, γίνεται όσο περίπου το εκτιμώμενο ήδη κατά την προσέγγιση πρώτης τάξης, και στη συνέχεια μειώνεται ταχύτερα. Επομένως, η εκτίμηση του σφάλματος μπορεί να χρησιμοποιηθεί με ασφάλεια, π.χ. για να βρεθεί ο ελάχιστος αριθμός όρων της σειράς, που απαιτούνται για να προσεγγίζεται η συνάρτηση με την επιθυμητή ακρίβεια. Το αποτέλεσμα αυτό οφείλεται στο ότι οι όλες οι παράγωγοι της συνάρτησης είναι σταθερές και ίσες με e^1.

Δεν συμβαίνει όμως το ίδιο για τη δεύτερη συνάρτηση, οι παράγωγοι της οποίας αυξάνουν προοδευτικά: $f_2^{(n)}(1) = 3^n e^3$. Έτσι, στην προσέγγιση πρώτης τάξης το πραγματικό σφάλμα αποκοπής είναι δύο τάξεις μεγέθους μεγαλύτερο από το εκτιμώμενο. Στη συνέχεια αρχίζει να μικραίνει ταχύτερα, αλλά παραμένει μία τάξη μεγέθους μεγαλύτερο ακόμη και μετά την προσθήκη του όρου πέμπτης τάξης, και μόνο μετά και τον όρο όγδοης τάξης συμφωνεί με την εκτίμηση. Παρά την αναντιστοιχία αυτή όμως, οι πρώτοι πέντε – έξι όροι του αναπτύγματος είναι και εδώ αρκετοί, ώστε να ληφθεί προσέγγιση ικανοποιητικής ακρίβειας.

1.4. Μετάδοση Σφάλματος

Σε μια μέθοδο αριθμητικής ανάλυσης, το σφάλμα κάθε ενδιάμεσου αποτελέσματος μεταφέρεται στην επόμενη σχέση που θα χρησιμοποιηθεί, και μπορεί να γίνει μικρότερο ή μεγαλύτερο κατά τους επόμενους υπολογισμούς. Για παράδειγμα, εάν μία ποσότητα x έχει υπολογισθεί με σχετικό σφάλμα 1%, τότε ο όρος x^2 θα έχει σφάλμα 2%, ενώ ο όρος $x^{0.5}$ θα έχει 0.5%. Επομένως είναι σημαντικό να ελέγχεται η επίδραση που θα έχει ένα σφάλμα σε επόμενες πράξεις και να αποφεύγεται μια μεγάλη αύξηση κατά τη μετάδοσή του. Μια τέτοια περίπτωση που έχει ήδη αναφερθεί, είναι η αφαίρεση δύο σχεδόν ίσων αριθμών, κατά την οποία μπορεί να προκληθεί μεγάλο σχετικό σφάλμα.

1.4.1. Συναρτήσεις Μίας Μεταβλητής

Σε μια συνάρτηση μίας μεταβλητής, $f(x)$, έστω x_t και x_a η ακριβής και μια προσεγγιστική τιμή της μεταβλητής της. Τότε η διαφορά μεταξύ της ακριβούς και της προσεγγιστικής τιμής της συνάρτησης θα δώσει το πραγματικό, απόλυτο σφάλμα

$$E_{t,f(x)} = f(x_t) - f(x_a) \tag{1.21}$$

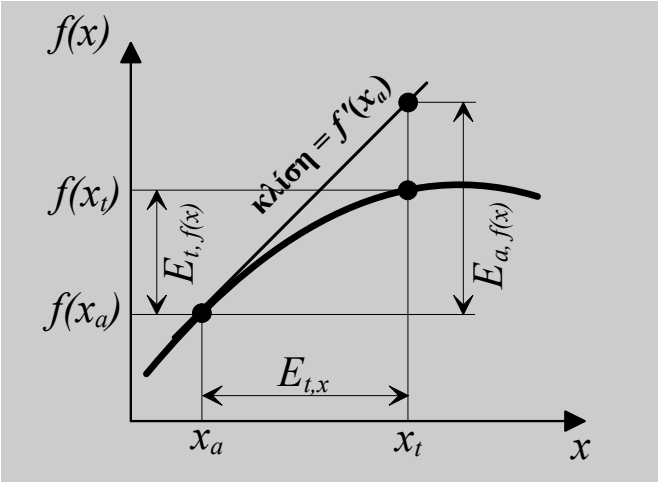
Εάν υποτεθεί ότι η συνάρτηση $f(x)$ είναι συνεχής και παραγωγίσιμη στην περιοχή του x_t, τότε μπορεί αναπτυχθεί σε σειρά Taylor. Η προσέγγιση πρώτης τάξης θα είναι

$$f(x_t) \cong f(x_a) + f'(x_a)(x_t - x_a) \Rightarrow f(x_t) - f(x_a) \cong f'(x_a)(x_t - x_a) \tag{1.22}$$

Από τις δύο προηγούμενες σχέσεις, συνάγεται ότι

$$E_{t,f(x)} \cong f'(x_a) E_{t,x} = E_{a,f(x)} \tag{1.23}$$

όπου $E_{t,x}$ το απόλυτο σφάλμα της μεταβλητής x, και $E_{a,f(x)}$ το εκτιμώμενο, απόλυτο σφάλμα της συνάρτησης $f(x)$.



Σχήμα 1.3. Γραφική εικόνα του εκτιμώμενου σφάλματος συνάρτησης $f(x)$.

Η Εξ. (1.23) δίνει λοιπόν κατά προσέγγιση το σφάλμα που μεταδίδεται από τη μεταβλητή x στην τιμή της συνάρτησης $f(x)$, όταν είναι γνωστή η τιμή της πρώτης

παραγώγου. Όπως φαίνεται και στο γραφικό παράδειγμα του Σχήματος 1.3, η εκτίμηση αυτή θα είναι ακριβής εάν η συνάρτηση είναι γραμμική.

Μια πιο παραστατική εικόνα για τη μετάδοση του σφάλματος δίνει η αντίστοιχη σχέση που προκύπτει για το εκτιμώμενο σχετικό σφάλμα

$$\varepsilon_{a,f(x)} \cong \frac{E_{a,f(x)}}{f(x_a)} \cong \frac{f'(x_a)(x_t - x_a)}{f(x_a)} = \frac{f'(x_a)\, x_a}{f(x_a)} \, \frac{(x_t - x_a)}{x_a} = \frac{f'(x_a)\, x_a}{f(x_a)} \, \varepsilon_{a,x} \qquad (1.24)$$

Επομένως η μετάδοση του σφάλματος εκφράζεται με την παράμετρο $\dfrac{f'(x_a)\, x_a}{f(x_a)}$, η οποία ονομάζεται και *δείκτης κατάστασης* της συνάρτησης στο σημείο x_a. Εάν η απόλυτη τιμή του δείκτη είναι κοντά στη μονάδα ή μικρότερη, τότε το σφάλμα παραμένει περίπου σταθερό ή μειώνεται κατά τη μετάδοση, ενώ εάν είναι μεγαλύτερη, το σφάλμα αυξάνεται. Για παράδειγμα, η συνάρτηση x^b έχει δείκτη κατάστασης ίσο με b, επομένως το σφάλμα του υπολογισμού μιας πολυωνυμικής συνάρτησης θα μεγαλώνει όσο αυξάνει ο βαθμός της.

Για πολύ μεγάλες τιμές του δείκτη, η αύξηση του σφάλματος θα είναι τέτοια που μπορεί να οδηγήσει σε τελείως λανθασμένο τελικό αποτέλεσμα. Τέτοιες περιπτώσεις, που ονομάζονται και 'ασθενείς' καταστάσεις μιας συνάρτησης, πρέπει να λαμβάνονται σοβαρά υπόψη κατά την ανάπτυξη του αλγορίθμου μιας αριθμητικής μεθόδου.

Εφαρμογή 1.12.

Να εκτιμηθεί το βεληνεκές για ένα βλήμα που εκτοξεύεται με ταχύτητα u_0 = 140 m/s σε ομογενές πεδίο βαρύτητας (g = 9.8 m/s^2), όταν η γωνία εκτόξευσης θ ρυθμίζεται στα 0.5 rad, με ακρίβεια ± 0.02 rad.

Το σφάλμα της γωνίας είναι $E_{t,\theta} = \pm 0.02$ rad, δηλαδή $\theta_t = 0.5 \pm 0.02$. Το βεληνεκές δίνεται από τη σχέση $R(\theta) = \dfrac{u_0^2 \cdot \sin 2\theta}{g} = \dfrac{140^2 \cdot \sin 2\theta}{9.8} = 2 \cdot 10^3 \sin 2\theta$

Προκύπτουν επίσης: $R'(\theta) = 4 \cdot 10^3 \cos 2\theta$ και $R(0.5)$ = 1682.94 m. Και από την Εξ. (1.23) λαμβάνεται

$$E_{t,R(\theta)} \cong R'(\theta_a)\, E_{t,\theta} = 4 \cdot 10^3 \cos(2 \cdot 0.5)(\pm 0.02) = \pm 43.224$$

Έτσι, το βεληνεκές εκτιμάται ως $R(\theta_t) \cong 1682.94 \pm 43.224$ m ή ως

$$1639.7 < R(\theta_t) < 1726.2$$

Τα όρια του βεληνεκούς μπορούν να υπολογιστούν και ακριβώς, εάν τοποθετηθούν οι ακραίες τιμές της γωνίας εκτόξευσης στην εξίσωση ορισμού του. Τότε θα ληφθεί $R(0.48)$ = 1638.38 και $R(0.52)$ = 1724.81. Επομένως η προσεγγιστική εκτίμηση με την Εξ. (1.23) είναι πολύ ικανοποιητική.

Εφαρμογή 1.13.

Να εκτιμηθεί ο δείκτης κατάστασης της συνάρτησης $f(x) = \ln x$, για $x = 0.9$, $x = 0.99$ και $x = 10$.

Ο δείκτης κατάστασης θα είναι $\dfrac{f'(x_a)\, x_a}{f(x_a)} = \dfrac{(1/x_a)\, x_a}{\ln x_a} = \dfrac{1}{\ln x_a}$

και για τα δεδομένα x_a: 0.9, 0.99 και 10, παίρνει τις αντίστοιχες (απόλυτες) τιμές: 9.49, 99.5 και 0.43.

Είναι φανερό ότι, όσο η μεταβλητή x_a πλησιάζει την τιμή 1, τόσο πιο ασθενής γίνεται η κατάσταση της συνάρτησης, παρόλο που τόσο η ίδια η συνάρτηση, όσο και η παράγωγός της, είναι πεπερασμένες στο σημείο αυτό. Η συμπεριφορά αυτή εξηγείται με την παρατήρηση ότι, όταν το x πλησιάζει στη μονάδα, η συνάρτηση πλησιάζει στο μηδέν, επομένως η έκφραση του σχετικού σφάλματος παίρνει μεγάλες τιμές. Εάν χρησιμοποιηθεί το απόλυτο σφάλμα, (το οποίο, όπως έχει αναφερθεί, δίνει πιο σωστή εικόνα του σφάλματος για τιμές κοντά στο μηδέν), τότε από την Εξ. (1.23) θα προκύψει ότι το απόλυτο σφάλμα της συνάρτησης είναι περίπου ίσο με το σφάλμα της μεταβλητής x στην περιοχή της μονάδας.

Παρόλα αυτά, σε περίπτωση που η μεταβλητή x παίρνει τιμές κοντά στη μονάδα και η τιμή της συνάρτησης πρόκειται να χρησιμοποιηθεί σε επόμενες αριθμητικές πράξεις, ιδίως πολλαπλασιασμού ή διαίρεσης, τότε πρέπει να εξασφαλισθεί η μέγιστη δυνατή ακρίβεια για το x, αφού το όποιο σφάλμα θα μεταδοθεί πολλαπλάσιο.

Τέλος, για $x = 10$ ο δείκτης κατάστασης είναι μικρότερος της μονάδας, επομένως το σχετικό σφάλμα της συνάρτησης όχι μόνο δεν αυξάνει, αλλά μειώνεται σε λιγότερο από το μισό του σφάλματος της μεταβλητής x.

1.4.2. Συναρτήσεις Δύο ή Περισσότερων Μεταβλητών

Στην περίπτωση συναρτήσεων πολλών ανεξάρτητων μεταβλητών ακολουθείται η ίδια διαδικασία. Μια συνάρτηση $f(x_1, x_2, \ldots x_n)$ μπορεί να αναπτυχθεί σε σειρά Taylor πολλών μεταβλητών και να ληφθεί η προσέγγιση πρώτης τάξης, από την οποία προκύπτει η ακόλουθη γενική σχέση:

$$E_{t,f(x_1,x_2,\ldots x_n)} \cong \frac{\partial f_a}{\partial x_1} E_{t,x_1} + \frac{\partial f_a}{\partial x_2} E_{t,x_2} + \ldots + \frac{\partial f_a}{\partial x_n} E_{t,x_n} \tag{1.25}$$

όπου $f_a = f(x_{1,a}, x_{2,a}, \ldots x_{n,a})$ η τιμή της συνάρτησης για τις προσεγγιστικές τιμές $x_{1,a}, x_{2,a} \ldots x_{n,a}$ των ακριβών τιμών $x_1, x_2 \ldots x_n$.

Η εξίσωση αυτή δίνει κατά προσέγγιση το συνολικό σφάλμα που μεταδίδεται από όλες τις ανεξάρτητες μεταβλητές x_i στην τιμή της συνάρτησης, όταν είναι γνωστές οι τιμές των μερικών παραγώγων πρώτης τάξης και τα σφάλματα των μεταβλητών της. Επίσης, διαιρώντας όλα τα μέλη της με f_a, λαμβάνεται η σχέση για το σχετικό σφάλμα, αντίστοιχη της Εξ. (1.24)

$$\varepsilon_{a,f(x_1,x_2,\ldots x_n)} \cong \frac{\partial f_a}{\partial x_1} \frac{x_{1,a}}{f_a} \varepsilon_{a,x_1} + \frac{\partial f_a}{\partial x_2} \frac{x_{2,a}}{f_a} \varepsilon_{a,x_2} + \ldots + \frac{\partial f_a}{\partial x_n} \frac{x_{n,a}}{f_a} \varepsilon_{a,x_n} \tag{1.26}$$

Οι Εξ. (1.25) και (1.26) αποτελούν ιδιαίτερα χρήσιμα εργαλεία για την εκτίμηση του σφάλματος στην τιμή μιας συνάρτησης, είτε πρόκειται για αριθμητική παράσταση ενός αλγορίθμου (οπότε ενδιαφέρει το σφάλμα που μεταδίδεται), είτε για μαθηματική σχέση που περιγράφει έναν φυσικό μηχανισμό (οπότε ενδιαφέρει η επίδραση τυχόν αβεβαιότητας ή διακύμανσης της τιμής των δεδομένων). Όταν το πρόσημο του πραγματικού σφάλματος των μεταβλητών δεν είναι γνωστό (π.χ. διακύμανση γύρω από μία μέση τιμή), η τιμή της συνάρτησης θα κυμαίνεται επίσης γύρω από μια μέση τιμή. Το μέγιστο εύρος αυτής της διακύμανσης μπορεί να εκτιμηθεί και πάλι από τις παραπάνω εξισώσεις, όπου όμως όλοι οι όροι λαμβάνονται σε απόλυτη τιμή.

Μια απλή αλλά ιδιαίτερη χρήσιμη εφαρμογή των Εξ. (1.25) και (1.26) γίνεται στη συνέχεια, για την εκτίμηση του σφάλματος που μεταδίδεται στο αποτέλεσμα $f(x, y)$ των τεσσάρων αριθμητικών πράξεων μεταξύ δύο μεταβλητών, x και y. Έτσι προκύπτουν οι ακόλουθες εκφράσεις:

Πρόσθεση: $f(x, y) = x + y$

$$E_{t,(x+y)} \cong E_{t,x} + E_{t,y} \qquad \text{και} \qquad \varepsilon_{a,(x+y)} \cong \frac{x}{x+y}\,\varepsilon_{a,x} + \frac{y}{x+y}\,\varepsilon_{a,y} \tag{1.27}$$

Αφαίρεση: $f(x, y) = x - y$

$$E_{t,(x-y)} \cong E_{t,x} - E_{t,y} \qquad \text{και} \qquad \varepsilon_{a,(x-y)} \cong \frac{x}{x-y}\,\varepsilon_{a,x} - \frac{y}{x-y}\,\varepsilon_{a,y} \tag{1.28}$$

Πολλαπλασιασμός: $f(x, y) = x \cdot y$

$$E_{t,(x\cdot y)} \cong y \cdot E_{t,x} + x \cdot E_{t,y} \qquad \text{και} \qquad \varepsilon_{a,(x\cdot y)} \cong \varepsilon_{a,x} + \varepsilon_{a,y} \tag{1.29}$$

Διαίρεση: $f(x, y) = x / y$

$$E_{t,(x/y)} \cong \frac{1}{y}E_{t,x} - \frac{x}{y^2}E_{t,y} \qquad \text{και} \qquad \varepsilon_{a,(x/y)} \cong \varepsilon_{a,x} - \varepsilon_{a,y} \tag{1.30}$$

Στις παραπάνω εξισώσεις έχει παραληφθεί ο δείκτης a των προσεγγιστικών τιμών των μεταβλητών x και y, οι οποίες, επίσης, θεωρούνται ομόσημες στην πρόσθεση και την αφαίρεση.

Όταν δεν είναι γνωστό το πρόσημο του σφάλματος των μεταβλητών x και y, τότε όλοι οι όροι των Εξ. (1.27) έως (1.30) γράφονται με απόλυτη τιμή και αθροίζονται, δίνοντας τη μέγιστη δυνατή απόλυτη τιμή του σφάλματος σε κάθε πράξη. Όπως παρατηρείται, σε όλες τις αριθμητικές πράξεις, εκτός από την αφαίρεση, το μέγιστο σχετικό σφάλμα του αποτελέσματος θα είναι το πολύ ίσο με το άθροισμα των σφαλμάτων των δύο μεταβλητών:

$$\left|\varepsilon_{a,f(x,y)}\right| \le \left|\varepsilon_{a,x}\right| + \left|\varepsilon_{a,y}\right|$$

Έτσι, όταν τα σχετικά σφάλματα $\varepsilon_{a,x}$ και $\varepsilon_{a,y}$ είναι προς την ίδια κατεύθυνση (ομόσημα), το σφάλμα μεταδίδεται αυξανόμενο κατά την πρόσθεση και τον πολλαπλασιαμό, και μειούμενο κατά τη διαίρεση, ενώ το αντίθετο συμβαίνει όταν τα σχετικά σφάλματα είναι ετερόσημα. Τα αριθμητικά σφάλματα συνήθως είναι τυχαία, δηλαδή είτε θετικά είτε αρνητικά, έτσι τα σφάλματα που μεταδίδονται μετά από πολλές αριθμητικές πράξεις δεν αυξάνουν αθροιστικά. Το τελικό σφάλμα, για παράδειγμα, μιας σειράς N πράξεων, που η κάθε μία έχει σφάλμα ε, δεν είναι $\varepsilon \cdot N$, αλλά της τάξης του $\varepsilon \cdot N^{0.5}$.

Η συμπεριφορά του σχετικού σφάλματος κατά την αφαίρεση είναι όμως εντελώς διαφορετική, εξ αιτίας της ύπαρξης του όρου $(x - y)$ στον παρονομαστή της Εξ. (1.28). Έτσι, είναι φανερό ότι το σφάλμα της διαφοράς μπορεί να γίνει πολύ μεγαλύτερο από τα επί

μέρους σφάλματα των μεταβλητών, όταν οι τιμές των τελευταίων είναι σχεδόν ίσες. Για άλλη μια φορά λοιπόν τονίζεται ο κίνδυνος πρόκλησης σοβαρού αριθμητικού σφάλματος κατά την αφαίρεση.

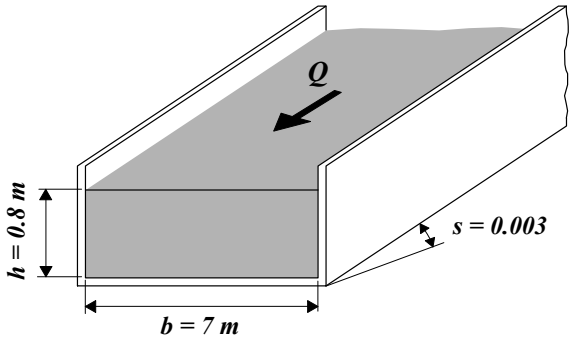
Εφαρμογή 1.14.

Ζητείται να εκτιμηθεί η παροχή Q (m^3/s) ενός ορθογώνιου ανοικτού καναλιού σταθερής ροής, από τη θεωρητική σχέση (τύπος του Manning)

$$Q = \frac{1}{n} \frac{(b\,h)^{5/3}}{(b+2h)^{2/3}} s^{1/2}$$

για τις εξής τιμές: βάθος h = 0.8 m, πλάτος b = 7 m, κλίση s = 0.003 και συντελεστής τραχύτητας n = 0.015. Το σχετικό σφάλμα μέτρησης ή εκτίμησης είναι $\varepsilon_a = \pm 4$ %, για όλα τα μεγέθη.

Σχήμα 1.4.
Ομοιόμορφη ροή σε ορθογώνιο ανοικτό κανάλι



Υπολογίζεται πρώτα από τη θεωρητική σχέση η προσεγγιστική τιμή της παροχής, για τις δεδομένες τιμές των μεταβλητών, και προκύπτει Q_a = 15.362095 m/s. Στη συνέχεια, επειδή πρόκειται για διακύμανση του σχετικού σφάλματος των μεταβλητών, η διακύμανση της παροχής θα ληφθεί από την Εξ. (1.26), με απόλυτες τιμές:

$$\left|\varepsilon_{a,Q}\right| \cong \left|\frac{\partial Q_a}{\partial n}\frac{n_a}{Q_a}\varepsilon_{a,n}\right| + \left|\frac{\partial Q_a}{\partial b}\frac{b_a}{Q_a}\varepsilon_{a,b}\right| + \left|\frac{\partial Q_a}{\partial h}\frac{h_a}{Q_a}\varepsilon_{a,h}\right| + \left|\frac{\partial Q_a}{\partial s}\frac{s_a}{Q_a}\varepsilon_{a,s}\right|$$

Οι μερικές παράγωγοι θα είναι (ο δείκτης a των μεταβλητών παραλείπεται)

$$\frac{\partial Q_a}{\partial n} = -\frac{1}{n^2}\frac{(b\,h)^{5/3}}{(b+2h)^{2/3}} s^{1/2} = -1024.1397$$

$$\frac{\partial Q_a}{\partial b} = \frac{1}{n}\frac{h^{5/3}b^{2/3}(3b+10h)}{3(b+2h)^{5/3}} s^{1/2} = 2.4667816$$

$$\frac{\partial Q_a}{\partial h} = \frac{1}{n}\frac{h^{2/3}b^{5/3}(5b+6h)}{3(b+2h)^{5/3}} s^{1/2} = 29.622644$$

$$\frac{\partial Q_a}{\partial s} = \frac{0.5}{n}\frac{(b\,h)^{5/3}}{(b+2h)^{2/3}}\frac{1}{s^{1/2}} = 2560.3491$$

Αντικαθιστώντας όλες τις τιμές στην εξίσωση του σφάλματος, προκύπτει:

$$\left|\varepsilon_{a,Q}\right| \cong 1.0\cdot\left|\varepsilon_{a,n}\right| + 1.124031\cdot\left|\varepsilon_{a,b}\right| + 1.5426357\cdot\left|\varepsilon_{a,h}\right| + 0.5\cdot\left|\varepsilon_{a,s}\right|$$
$$= 4.1666667\cdot\left|\varepsilon_{a}\right| = 0.16667$$

Επομένως, το σχετικό σφάλμα εκτίμησης της παροχής θα είναι ± 16.67 %, οπότε

$$Q_t = Q_a\cdot(1\pm 0.1667) = 15.36\pm 2.56 \text{ m/s.}$$

Μια εναλλακτική, απλούστερη μεθοδολογία που μπορεί να εφαρμοσθεί εδώ, είναι η ακόλουθη: σε περιπτώσεις όπου η συνάρτηση μπορεί να χωρισθεί σε τμήματα που περιέχουν εξ ολοκλήρου μία ανεξάρτητη μεταβλητή, το σφάλμα της τιμής της μπορεί να προκύψει από τα επί μέρους σφάλματα κάθε τμήματος. Έτσι, ο τύπος του Manning χωρίζεται σε τρία τμήματα:

$$z_1 = \frac{1}{n}, \qquad z_2 = \frac{(bh)^{5/3}}{(b+2h)^{2/3}} \quad \text{και} \quad z_3 = s^{1/2}, \qquad \text{οπότε} \qquad Q = z_1\cdot z_2\cdot z_3$$

Οι όροι z_1 και z_3 αποτελούν συναρτήσεις μιας μεταβλητής, έτσι από την Εξ. (1.24) προκύπτει: $\left|\varepsilon_{a,z_1}\right| \cong \left|\varepsilon_{a,n}\right|$ και $\left|\varepsilon_{a,z_3}\right| \cong 0.5\cdot\left|\varepsilon_{a,s}\right|$

Ο όρος z_2 περιέχει δύο μεταβλητές, όμως με την παρατήρηση ότι μεγιστοποιείται ή ελαχιστοποιείται όταν τα σφάλματα αυτών είναι ομόσημα, και με χρήση των Εξ. (1.27) και (1.30), προκύπτει:

$$\left|\varepsilon_{a,z_2}\right| \cong \frac{5}{3}\cdot\left\{\left|\varepsilon_{a,b}\right| + \left|\varepsilon_{a,h}\right|\right\} - \frac{2}{3}\cdot\left\{\frac{b}{b+2h}\left|\varepsilon_{a,b}\right| + \frac{2h}{b+2h}\left|\varepsilon_{a,h}\right|\right\}$$

Τέλος, χρησιμοποιώντας την Εξ. (1.29), το μέγιστο σφάλμα της παροχής θα είναι

$$\left|\varepsilon_{a,Q}\right| \cong \left|\varepsilon_{a,z_1}\right| + \left|\varepsilon_{a,z_2}\right| + \left|\varepsilon_{a,z_3}\right| = \left|\varepsilon_{a,n}\right| + 1.124031\cdot\left|\varepsilon_{a,b}\right| + 1.5426357\cdot\left|\varepsilon_{a,h}\right| + 0.5\cdot\left|\varepsilon_{a,s}\right|$$

δηλαδή λαμβάνεται ακριβώς το ίδιο αποτέλεσμα, αλλά χωρίς να χρειαστεί να βρεθούν και να υπολογιστούν οι μερικές παράγωγοι της συνάρτησης.

Η πραγματική διακύμανση της τιμής της παροχής μπορεί εδώ επίσης να υπολογισθεί, επειδή η συνάρτηση είναι σχετικά απλή, χρησιμοποιώντας τις ακραίες τιμές των μεταβλητών της. Ο υπολογισμός αυτός δίνει τελικά

$$15.36 - 2.38 \le Q_t \le 15.36 + 2.72 \text{ m/s.}$$

Επομένως, η προσεγγιστική μέθοδος εκτιμά ικανοποιητικά την περιοχή διακύμανσης και σχεδόν ακριβώς το εύρος διακύμανσης της παροχής.

Ένα ακόμη πιο σημαντικό πλεονέκτημα της μεθόδου όμως είναι ότι παρέχει και τη σχετική επίδραση του σφάλματος κάθε μεταβλητής στο συνολικό σφάλμα της συνάρτησης ή αλλιώς, την ευαισθησία της τιμής της συνάρτησης σε κάθε μία από τις ανεξάρτητες μεταβλητές της. Για παράδειγμα, το σφάλμα της κλίσης έχει εδώ τον μικρότερο συντελεστή, 0.5, ενώ το σφάλμα μέτρησης του βάθους του καναλιού έχει τον μεγαλύτερο, 1.543. Έτσι, εάν απαιτείται μεγαλύτερη ακρίβεια εκτίμησης της παροχή του καναλιού, θα πρέπει να βελτιωθεί κυρίως η ακρίβεια μέτρησης του βάθους του και, κατά δεύτερο λόγο, του πλάτους του.

1.5. Έλεγχος Σφαλμάτων Αριθμητικών Μεθόδων

Το σφάλμα στρογγυλοποίησης και το σφάλμα αποκοπής, που αναλύθηκαν σε προηγούμενα κεφάλαια, αποτελούν τις δύο μορφές με τις οποίες εμφανίζονται τα αριθμητικά σφάλματα. Η ορθότητα του αποτελέσματος μιας αριθμητικής μεθόδου δεν εξαρτάται όμως μόνο από το αριθμητικό σφάλμα, αλλά και από μια σειρά άλλων λαθών, που δεν οφείλονται σε υπολογισμούς, αλλά μπορεί να προκύψουν κατά τη διαδικασία ανάπτυξης και εφαρμογής της μεθόδου. Έτσι, στη συνέχεια, μαζί με μερικές πρακτικές οδηγίες και υποδείξεις για τον έλεγχο των αριθμητικών σφαλμάτων, θα γίνει και μια σύντομη περιγραφή και ανάλυση των πρόσθετων αυτών πηγών σφάλματος στο τελικό αριθμητικό αποτέλεσμα.

1.5.1. Αριθμητικά Σφάλματα

Όπως αναφέρθηκε, το συνολικό αριθμητικό σφάλμα, το οποίο τελικά ενδιαφέρει, προκύπτει ως το άθροισμα του σφάλματος στρογγυλοποίησης και αποκοπής. Όμως η προσπάθεια μείωσης του ενός εκ των δύο αυτών σφαλμάτων συνήθως προκαλεί αύξηση του άλλου. Έτσι, η χρησιμοποίηση μικρότερου βήματος ή/και περισσότερων όρων του αναπτύγματος Taylor μειώνει το σφάλμα αποκοπής, αυξάνει όμως το σφάλμα στρογγυλοποίησης, επειδή μεγαλώνει ο αριθμός των αριθμητικών πράξεων, καθώς και η πιθανότητα απώλειας ακρίβειας κατά την πρόσθεση αριθμών που διαφέρουν σημαντικά ή την αφαίρεση δύο παραπλήσιων αριθμών.

Βέβαια, σε έναν σύγχρονο 32-bit υπολογιστή το σφάλμα στρογγυλοποίησης είναι πολύ μικρότερο από το σφάλμα αποκοπής, επομένως η μείωση του τελευταίου αποτελεί την κύρια προτεραιότητα. Υπάρχουν όμως και περιπτώσεις όπου το σφάλμα στρογγυλοποίησης μπορεί να είναι συγκρίσιμο με το σφάλμα αποκοπής ή όπου απαιτείται εξαιρετικά μεγάλη ακρίβεια αποτελέσματος. Τότε είναι επίσης αναγκαία η μείωση και του σφάλματος στρογγυλοποίησης, που μπορεί να επιτευχθεί δραστικά αυξάνοντας τα σημαντικά ψηφία του υπολογιστή, δηλαδή χρησιμοποιώντας αριθμητική διπλής ακρίβειας.

Ακόμη και όταν δεν υπάρχει τέτοια ανάγκη όμως, πρέπει να αντιμετωπίζονται οι πιθανές περιπτώσεις πρόκλησης αυξημένου σφάλματος στρογγυλοποίησης σε κάποια αριθμητική πράξη. Έτσι, κατά την ανάπτυξη του αλγορίθμου πρέπει να γίνεται προσπάθεια αναδιάταξης των αριθμητικών παραστάσεων, ώστε:

- Να αποφεύγεται η αφαίρεση δύο παραπλήσιων αριθμών.
- Να αποφεύγεται η πρόσθεση δύο αριθμών πολύ διαφορετικού μεγέθους.
- Να αθροίζονται πρώτα οι μικρότεροι και μετά οι μεγαλύτεροι όροι μιας παράστασης.
- Να μειώνεται κατά το δυνατόν ο αριθμός των πράξεων.

Τέλος, σε εφαρμογές με μικρές υπολογιστικές απαιτήσεις μνήμης και χρόνου εκτέλεσης, καλό θα είναι να χρησιμοποιούνται πάντοτε αριθμοί διπλής ακρίβειας.

1.5.2. Σφάλματα Μοντελοποίησης

Τα περισσότερα μαθηματικά μοντέλα αποτελούν προσεγγίσεις των αντίστοιχων φυσικοχημικών φαινομένων, είτε επειδή δεν είναι πλήρως κατανοητός ο ακριβής φυσικός μηχανισμός, είτε για να απλοποιηθεί η μαθηματική περιγραφή και η διαδικασία επίλυσης ενός προβλήματος.

Για παράδειγμα, η καταστατική εξίσωση $pv = RT$ ισχύει για τέλεια αέρια, αλλά συνήθως εφαρμόζεται προσεγγιστικά και σε πραγματικά αέρια, αντί για ακριβέστερες αλλά συνθετότερες εκφράσεις, όπως η εξίσωση van der Waals $(p + a/v^2)(v - b) = RT$. Έτσι, η θερμοκρασία ενός πραγματικού αερίου, π.χ. του O_2, υπολογίζεται, για $p = 10^5$ Pa, $v = 0.7$ m^3/kg, $R = 260$, $a \cong 150$ και $b \cong 0.001$, από την πρώτη έκφραση ίση με 269.232 K, ενώ από τη δεύτερη ίση με 269.669 K. Επομένως η χρησιμοποίηση της καταστατικής εξίσωσης τελείων αερίων προκαλεί σφάλμα σχεδόν μισού βαθμού Kelvin, ή περίπου 0.16 %, το οποίο είναι πολύ μεγαλύτερο από το σφάλμα στρογγυλοποίησης. Από την άλλη μεριά, ούτε η εξίσωση van der Waals είναι απόλυτα ακριβής, όμως θεωρητικά έχει πολύ μικρότερο σφάλμα, καθώς συμπεριλαμβάνει και τη μοντελοποίηση λεπτομερέστερων φαινομένων (όπως του όγκου που καταλαμβάνουν τα μόρια του αερίου και των ελκτικών δυνάμεων μεταξύ των μορίων), τα οποία δεν υπάρχουν σε ένα τέλειο αέριο.

Γενικά, όταν οι υπολογισμοί βασίζονται σε ένα ατελές, ανακριβές ή και λανθασμένο ακόμη μαθηματικό μοντέλο, είναι προφανές ότι καμία μέθοδος αριθμητικής ανάλυσης δεν μπορεί να δώσει ακριβές αποτέλεσμα, ενώ το σφάλμα μοντελοποίησης μπορεί πολλές φορές να είναι πολύ μεγαλύτερο από το αριθμητικό σφάλμα, καθιστώντας περιττή κάθε προσπάθεια μείωσης του τελευταίου.

1.5.3. Σφάλματα Δεδομένων

Στις περισσότερες πρακτικές εφαρμογές χρησιμοποιούνται κάποια δεδομένα εισόδου, είτε πρόκειται για τιμές παραμέτρων και φυσικών σταθερών του μαθηματικού μοντέλου, είτε για αποτελέσματα μετρήσεων των αρχικών τιμών διαφόρων μεταβλητών. Σε τελική ανάλυση, και οι δύο αυτές κατηγορίες δεδομένων προέρχονται από μετρήσεις με διάφορες συσκευές και όργανα, επομένως εμπεριέχουν ένα ποσοστό σφάλματος ή αβεβαιότητας.

Για παράδειγμα, στην εξίσωση του βελησνεκούς που υπολογίστηκε στην Εφαρμογή 1.12, η επιτάχυνση της βαρύτητας και η ταχύτητα εκτόξευσης αποτελούν δεδομένα του προβλήματος, τα οποία θεωρήθηκαν ακριβή. Όμως η επιτάχυνση της βαρύτητας κυμαίνεται, από περίπου 9.78 στον ισημερινό έως περίπου 9.83 στους πόλους, επομένως η τιμή της πρέπει να δίνεται ανάλογα με το γεωγραφικό πλάτος εκτόξευσης. Επίσης, η ταχύτητα εκτόξευσης δεν μπορεί παρά να αποτελεί τη μέση τιμή μιας σειράς μετρήσεων με κάποια τυπική απόκλιση, επομένως το εύρος διακύμανσής της θα πρέπει να ληφθεί υπόψη, όπως έγινε ήδη για τη γωνία εκτόξευσης.

Επίσης, μερικές φορές εισάγεται σφάλμα δεδομένων κατά την εκτίμηση της τιμής μιας παραμέτρου από ένα διάγραμμα ή γράφημα, ή κατά τη λήψη μιας τιμής από έναν πίνακα. Για παράδειγμα, η δυναμική συνεκτικότητα (ιξώδες) του νερού δίνεται σε πίνακες ως συνάρτηση της θερμοκρασίας, συνήθως ανά 5 βαθμούς. Έτσι, σε $0\ ^oC$ είναι $\mu = 1.7916 \cdot 10^{-3}$ N s/m^2, ενώ στους $5\ ^oC$ είναι $\mu = 1.5192 \cdot 10^{-3}$ N s/m^2. Εάν τώρα σε ένα πρακτικό πρόβλημα χρειάζεται το ιξώδες του νερού στους $4\ ^oC$, τότε η γραμμική παρεμβολή μεταξύ των δεδομένων του πίνακα θα δώσει την τιμή $\mu = 1.5737 \cdot 10^{-3}$ N s/m^2, που διαφέρει από την ακριβή τιμή $1.568 \cdot 10^{-3}$ N s/m^2, επειδή η μεταβολή του ιξώδους με τη θερμοκρασία δεν είναι γραμμική. Επομένως, από ένα και μόνο δεδομένο εισάγεται ήδη σφάλμα της τάξης του 0.36 %, σε όλους τους μετέπειτα υπολογισμούς.

1.5.4. Σφάλματα στον Αλγόριθμο

Η ανάπτυξη του αλγορίθμου μιας υπολογιστικής μεθόδου σε κάποια γλώσσα προγραμματισμού μπορεί να είναι από απλή έως ιδιαίτερα επίπονη διαδικασία, αναλόγως κυρίως του μεγέθους του κώδικα. Σε κάθε περίπτωση, εύκολα μπορεί να γίνουν σφάλματα κατά το γράψιμο του κώδικα, είτε αυτά είναι συντακτικά είτε λογικά. Τα συντακτικά λάθη, που αφορούν εσφαλμένη σύνταξη ή ορθογραφία εντολών ή μεταβλητών του προγράμματος, είναι εύκολο να διορθωθούν, με τη βοήθεια των διαγνωστικών μηνυμάτων που δίνει ο μεταγλωττιστής (compiler).

Αντίθετα, τα λογικά λάθη δεν μπορούν να διαγνωσθούν, εκτός εάν προκαλέσουν τερματισμό της εκτέλεσης του προγράμματος (π.χ. σφάλμα υπερχείλισης, τετραγωνική ρίζα αρνητικού αριθμού κ.ά.) ή εάν το τελικό αποτέλεσμα είναι εμφανώς λανθασμένο. Έτσι, ένα μικρό σφάλμα σε μια σταθερά ή σε ένα πρόσημο μιας αριθμητικής παράστασης μπορεί να διαφοροποιήσει ελαφρώς την τιμή που θα υπολογισθεί, χωρίς αυτό να γίνει αντιληπτό από τον χρήστη· μια λανθασμένη εντολή ελέγχου μπορεί να παρακάμψει την εκτέλεση ενός τμήματος του προγράμματος· ένας πραγματικός αριθμός μπορεί να χάσει τα δεκαδικά του ψηφία εάν συμβολισθεί με ακέραια μεταβλητή. Λογικό είναι επίσης το ενδεχόμενο σφάλμα εάν δεν γίνει σωστά ο έλεγχος της ισότητας δύο αριθμών ή όταν χρησιμοποιείται λανθασμένο κριτήριο σύγκλισης σε μια επαναληπτική μέθοδο, όπως αναφέρθηκε στην παράγραφο 1.2.2.

Ο κατάλογος των πιθανών λογικών σφαλμάτων είναι τόσο μεγάλος, ώστε ο ασφαλέστερος τρόπος διόρθωσης είναι η προσεκτική ανάγνωση ενός κώδικα εντολή προς εντολή. Αυτό είναι σχετικά εύκολο σε έναν απλό κώδικα, που υπολογίζει για παράδειγμα το άθροισμα των όρων μιας σειράς. Όταν όμως ο αλγόριθμος περιλαμβάνει πολλές εκατοντάδες ή χιλιάδες εντολές, όπως για παράδειγμα ο υπολογισμός των στατικών ενός μεγάλου κτιρίου ή η επίλυση ενός ρευστοδυναμικού προβλήματος, τότε η τήρηση ορισμένων βασικών κανόνων προγραμματισμού είναι απαραίτητη, ώστε να διευκολύνεται όχι μόνο η εύρεση ενός λογικού σφάλματος, αλλά και οποιαδήποτε μελλοντική τροποποίηση του αλγορίθμου. Οι παρακάτω κανόνες είναι σκόπιμο όμως να τηρούνται ακόμη και σε απλά προγράμματα, ώστε να αποκτάται σωστή προγραμματιστική συνήθεια.

- Πριν από το γράψιμο του κώδικα, να καταστρώνεται το λογικό διάγραμμα του αλγορίθμου.
- Ο κώδικας να είναι καλά δομημένος, δηλαδή η σειρά με την οποία γράφονται οι γραμμές του να αντιστοιχεί όσο είναι δυνατό με τη σειρά εκτέλεσης των πράξεων.
- Να επιδιώκεται η χρήση υποπρογραμμάτων ή συναρτήσεων (Functions), ώστε να παραμένει απλή και καθαρή η βασική δομή και λογική του αλγορίθμου.
- Οι εντολές ελέγχου και οι επαναληπτικοί βρόγχοι (IF και DO στη Fortran), να έχουν μία μόνο είσοδο και μία έξοδο.
- Οι εντολές μεταβίβασης ελέγχου (GO TO) να χρησιμοποιούνται όσο το δυνατόν λιγότερο και μόνο υπό συνθήκη.
- Να υπάρχουν επαρκή σχόλια σε όσα σημεία είναι απαραίτητο, ώστε ο κώδικας να γίνεται εύκολα κατανοητός από κάθε χρήστη.
- Τα διάφορα τμήματα του προγράμματος να είναι ευδιάκριτα και ευκρινή, με χρήση π.χ. κενών γραμμών, αυξανόμενων περιθωρίων κ.λ.π.

Πρέπει τέλος να τονιστεί ότι ο χρόνος που μπορεί να χαθεί εκ των υστέρων, για την ανεύρεση ενός ενδεχόμενου λογικού σφάλματος του κώδικα, είναι συνήθως πολλαπλάσιος εκείνου που απαιτείται για την εφαρμογή της παραπάνω πρακτικής κατά το στάδιο του προγραμματισμού.

1.5.5. Αξιολόγηση Αριθμητικών Αποτελεσμάτων

Γενικά, εκτός από ορισμένες απλές περιπτώσεις, η εκτίμηση του σφάλματος που ενυπάρχει σε ένα τελικό αριθμητικό αποτέλεσμα δεν είναι εύκολη διαδικασία, ούτε μπορεί να γενικευτεί, ενώ απαιτεί και εμπειρία και κρίση από την πλευρά του Μηχανικού. Παρόλα αυτά, μπορούν να διατυπωθούν μερικά βασικά σημεία μιας στρατηγικής ελέγχου και αξιολόγησης, που, ανάλογα με τη σπουδαιότητα και την κρισιμότητα των αποτελεσμάτων μιας αριθμητικής μεθόδου, κλιμακώνονται ως εξής:

- Εκτέλεση του προγράμματος με απλή και με διπλή ακρίβεια αριθμών, για προσεγγιστική εκτίμηση του σφάλματος στρογγυλοποίησης.
- Χρήση του αναπτύγματος σε σειρά Taylor, για προσεγγιστική εκτίμηση σφαλμάτων αποκοπής.
- Σύγκριση αποτελεσμάτων με χρήση μαθηματικών εκφράσεων διαφορετικής ακρίβειας (ή όρων) και βήματος.
- Έλεγχος εάν το αποτέλεσμα ικανοποιεί μια συνθήκη, μια εξίσωση ή ένα σύστημα εξισώσεων που επιλύεται.
- Εφαρμογή του κώδικα σε απλές περιπτώσεις, με γνωστή αναλυτική ή αριθμητική λύση.
- Επανάληψη των υπολογισμών με διαφορετικές παραμέτρους, για έλεγχο ευαισθησίας (π.χ. συντελεστές μοντέλων, δεδομένα εισόδου, χρονικό ή χωρικό βήμα, κριτήριο σύγκλισης κλπ.)
- Χρήση διαφορετικών θεωρητικών προσεγγίσεων, μαθηματικών μοντέλων και/ή μεθόδων αριθμητικής ανάλυσης, για την επίλυση του ίδιου προβλήματος.
- Σύγκριση αποτελεσμάτων μεταξύ διαφορετικών, ανεξάρτητων ερευνητικών ομάδων.

Μερικές από τις τεχνικές αυτές θα εφαρμοστούν και για τον έλεγχο της ακρίβειας και της αξιοπιστίας των αριθμητικών αποτελεσμάτων των μεθόδων αριθμητικής ανάλυσης, που παρουσιάζονται στα επόμενα κεφάλαια.

1.6. Ασκήσεις για το Εργαστήριο Η/Υ

α) Να βρεθεί ο αριθμός των θέσεων (bits) που διαθέτει η βάση (mantissa) του Η/Υ που χρησιμοποιείτε, προγραμματίζοντας και εκτελώντας τον αλγόριθμο του Κώδικα 1.1.

β) Να γραφεί κώδικας που να υπολογίζει της ρίζες μιας δευτεροβάθμιας εξίσωσης, σύμφωνα με το συμπέρασμα της Εφαρμογής 1.7.

γ) Να γραφεί κώδικας που να υπολογίζει με όσο το δυνατό μεγαλύτερη ακρίβεια την τιμή e^x χρησιμοποιώντας το ανάπτυγμα

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots$$

και να εφαρμοσθεί για $x = 1,\ x = 10$ και $x = -10$.

δ) Να γραφεί κώδικας που να υπολογίζει τη δυωνυμική σειρά

$$(a+x)^n = a^n + \binom{n}{1} a^{n-1}x + \binom{n}{2} a^{n-2}x^2 + \binom{n}{3} a^{n-3}x^3 + \ldots \qquad \text{με}$$

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} = \frac{n(n-1)(n-2)\ldots(n-k+1)}{k!}$$

για διάφορες τιμές των a, x και n, αλλά πάντα με ακρίβεια πέντε σημαντικών ψηφίων. Θα χρησιμοποιηθεί αριθμητική αυξημένης ακρίβειας, καθώς και οι Εξ. (1.3), (1.4), (1.5).

ε) Να γραφεί κώδικας που να υπολογίζει με ανάπτυγμα σε σειρά Taylor την τιμή της συνάρτησης $f(x) = \ln x$, με βάση τη γνωστή τιμή $f(1) = 0.0$. Ο αλγόριθμος πρέπει να μπορεί να εκτελεί τον υπολογισμό σε n διαδοχικά βήματα, με παράμετρο τον αριθμό των όρων και το x. Να μελετηθεί η ακρίβεια του αποτελέσματος για διάφορες τιμές των x και n.

Κεφάλαιο 2

Επίλυση Μη–Γραμμικών Εξισώσεων

2.1. Γενικά

Η ανάγκη εύρεσης της ρίζας ή των ριζών μιας εξίσωσης προκύπτει συχνά σε προβλήματα του Μηχανικού. Οι εξισώσεις αυτές έχουν τη γενική μορφή

$$f(x) = 0 \tag{2.1}$$

όπου f μια πραγματική συνάρτηση μιας πραγματικής μεταβλητής x και ζητείται η τιμή x_r για την οποία η $f(x_r)$ μηδενίζεται.

Να υπενθυμίσουμε ότι γραμμικές είναι οι εξισώσεις οι οποίες με αναδιάταξη των όρων εμφανίζουν την ανεξάρτητη μεταβλητή υψωμένη μόνο στην πρώτη δύναμη. Αντιθέτως, οι μη–γραμμικές αλγεβρικές εξισώσεις περιέχουν ανεξάρτητες μεταβλητές υψωμένες σε δυνάμεις διάφορες της μονάδας. Μη–γραμμικές θεωρούνται επίσης και οι υπερβατικές εξισώσεις, δηλαδή αυτές που εμπεριέχουν τριγωνομετρικές, εκθετικές κλπ. συναρτήσεις των μεταβλητών.

Ενώ μια γραμμική εξίσωση μπορεί εύκολα να επιλυθεί, μια μη–γραμμική συνάρτηση $f(x)$ συνήθως είναι πεπλεγμένη και δεν μπορεί να αντιστραφεί και να δώσει τη λύση

$$x_r = \varphi(y_r) \tag{2.2}$$

Το ίδιο ισχύει και για συναρτήσεις που δεν έχουν αναλυτική έκφραση, όπως για παράδειγμα η απόκριση ενός οργάνου ή η λύση κάποιας διαφορικής εξίσωσης. Έτσι προκύπτει η ανάγκη μιας αριθμητικής διαδικασίας εύρεσης (με κάποια προσέγγιση) της ρίζας ή των ριζών της Εξ. (2.1).

Ως παράδειγμα δίνεται η εξίσωση Van der Waals

$$\left(P + \frac{\alpha}{V^2}\right)(V - b) = RT \tag{2.3}$$

που είναι η καταστατική εξίσωση των πραγματικών αερίων, όπου α, b σταθερές του αερίου, P η πίεση (N/m^2), V ο ειδικός όγκος (m^3/kg), T η θερμοκρασία (Kelvin, K) και R η σταθερά του αερίου (J/kg/K). Η εξίσωση αυτή είναι γραμμική ως προς την πίεση ή τη θερμοκρασία και μπορεί εύκολα να επιλυθεί ως προς αυτά τα μεγέθη:

$$P = \frac{RT}{V - b} - \frac{\alpha}{V^2} \tag{2.3α}$$

(οπότε βρίσκεται η τιμή της πίεσης για δοσμένες τιμές θερμοκρασίας – όγκου του αερίου),

$$T = \frac{1}{R}\left(P + \frac{\alpha}{V^2}\right)(V - b) \tag{2.3β}$$

(οπότε βρίσκεται η τιμή της θερμοκρασίας για δοσμένες τιμές πίεσης και ειδικού όγκου).

Αντιθέτως, η εξίσωση Van der Waals είναι μη–γραμμική ως προς τον ειδικό όγκο του αερίου (λόγω της παρουσίας του όρου α/V^2) και δεν μπορεί να επιλυθεί δίνοντας μια αντίστοιχη έκφραση της μορφής $V = \varphi(P,T)$.

Από την άλλη μεριά, ενώ η υπερβατική εξίσωση $\alpha + \beta \ln x = 0 \quad (x > 0)$, μπορεί εύκολα να αντιστραφεί ως προς x και να δώσει αμέσως τη ρίζα $x_r = e^{-\alpha/\beta}$, η επίσης απλή εξίσωση $x - e^{-x} = 0$ δεν έχει αναλυτική λύση. Ακόμα, στην ειδική περίπτωση μη–γραμμικής αλγεβρικής εξίσωσης, που αποτελεί ένα πολυώνυμο n βαθμού ($n > 1$)

$$p(x) = \alpha_n x^n + \alpha_{n-1} x^{n-1} + \alpha_{n-2} x^{n-2} + \ldots + \alpha_0 \tag{2.4}$$

όπου α_n σταθεροί αριθμοί (πραγματικοί ή μιγαδικοί), υπάρχουν αναλυτικές εκφράσεις των ριζών μόνο έως $n = 4$. Τέλος, σε ένα άλλο παράδειγμα, η υπερβατικού τύπου εξίσωση

$$x \tan x = C \qquad (C > 0) \tag{2.5}$$

η οποία εκφράζει τη διανομή της θερμοκρασίας σε μονοδιάστατη ράβδο, όχι μόνο δεν επιλύεται αναλυτικά, αλλά έχει άγνωστο εκ των προτέρων πλήθος ριζών.

Γενικά, η δυνατότητα αναλυτικής έκφρασης της λύσης μιας μη–γραμμικής εξίσωσης θα πρέπει να διερευνάται πριν την προσφυγή σε μεθόδους αριθμητικής επίλυσης, καθώς οι τελευταίες είναι σαφώς πιο χρονοβόρες για δεδομένη ακρίβεια. Για παράδειγμα, η εξίσωση

$$\cos x \cdot \sin x - a = 0, \qquad |a| \le 1 \tag{2.6}$$

μπορεί να λυθεί αμέσως ως προς x, εάν χρησιμοποιηθεί η γνωστή ταυτότητα $2\cos x \cdot \sin x = \sin 2x$.

Οι αριθμητικές μέθοδοι επίλυσης των μη–γραμμικών εξισώσεων μπορούν να χωριστούν σε τρεις κατηγορίες, ανάλογα με τον σκοπό χρησιμοποίησής τους, ως εξής:

- Μέθοδοι εντοπισμού της περιοχής τιμών των ριζών μιας εξίσωσης.
- Μέθοδοι προσέγγισης της τιμής μιας ρίζας με προκαθορισμένη ακρίβεια.
- Μέθοδοι υπολογισμού όλων των ριζών (πραγματικών και μιγαδικών) ενός πολυωνύμου.

Στα πρακτικά προβλήματα του Μηχανικού αναζητείται συνήθως μία πραγματική ρίζα της εξίσωσης, που να έχει φυσική έννοια. Για τον σκοπό αυτόν υπάρχουν διάφοροι επαναληπτικοί αλγόριθμοι προσέγγισης, που θα παρουσιαστούν στη συνέχεια.. Η αριθμητική διαδικασία εύρεσης της ρίζας x_r της συνάρτησης $f(x)$ δημιουργεί μια ακολουθία αριθμών $x_1, x_2, x_3, \ldots, x_m$, η οποία προσεγγίζει τη ρίζα x_r με συνεχώς αυξανόμενη ακρίβεια. Θεωρείται ότι η τιμή x_m αποτελεί τη ρίζα της εξίσωσης $f(x)$ όταν ικανοποιηθεί κάποιο κριτήριο ακρίβειας στην προσέγγιση της x_r (π.χ. Εξ. 1.3). Ένας αρχικός εντοπισμός της περιοχής της ρίζας είναι για κάποιες μεθόδους προσέγγισης απαραίτητος, ενώ στις υπόλοιπες μπορεί να εξασφαλίσει ή να επιταχύνει τη σύγκλιση.

2.2. Εντοπισμός Διαστήματος που Εμπεριέχει Ρίζα

Για να εντοπισθεί η (πραγματική) ρίζα ή οι ρίζες μιας εξίσωσης είναι αναγκαίο να γνωρίζουμε τη συμπεριφορά της εξίσωσης σε όλο το πεδίο ορισμού της μονοδιάστατης μεταβλητής $(a \le x \le b)$. Ειδικότερα όμως είναι απαραίτητο να γνωρίζουμε ένα μικρό σχετικά πεδίο τιμών της μεταβλητής x, μέσα στο οποίο αναμένεται η ύπαρξη μιας ρίζας της εξίσωσης $(x_{b1} < x_0 < x_{b2})$. Η οριοθέτηση αυτή δεν είναι εύκολη στις μη–γραμμικές εξισώσεις, επειδή μπορεί να υπάρχουν περισσότερες από μία ή ακόμη και άπειρες πραγματικές ρίζες, όπως στα παραδείγματα $x^2 - 1 = 0$ και $\cos x = 0.5$ αντιστοίχως. Επιπλέον, υπάρχουν εξισώσεις που δεν έχουν πραγματικές ρίζες, όπως π.χ. η $x^2 + 1 = 0$, ή και καθόλου ρίζες, π.χ. $\cos x = 5$.

2.2.1. Μέθοδος Ίσων Διαστημάτων

Μία γενική μέθοδος εντοπισμού των πραγματικών ριζών μιας εξίσωσης βασίζεται στο *θεώρημα της μέσης τιμής*: "Εάν η f είναι συνεχής στο $[a, b]$ και z ένας πραγματικός αριθμός, τέτοιος ώστε $f(a) \le z \le f(b)$ ή $f(a) \ge z \ge f(b)$, τότε υπάρχει $x \in [a, b]$, ώστε $z = f(x)$ ".

Έτσι, εάν υπάρχουν δύο τιμές της $f(x)$ με διαφορετικό πρόσημο $f(a) \cdot f(b) < 0$, τότε υπάρχει μία πραγματική ρίζα της εξίσωσης στο διάστημα (a,b). Ακριβέστερα, αποδεικνύεται ότι μεταξύ δύο ετερόσημων τιμών μιας συνάρτησης υπάρχουν $2n + 1$ πραγματικές ρίζες (δηλαδή τουλάχιστον μία), ενώ μεταξύ δύο ομόσημων τιμών της υπάρχουν $2n$ ρίζες (δηλαδή μπορεί και καμμία).

Την ιδιότητα αυτή μιας συνεχούς συνάρτησης $f(x)$ να αλλάζει πρόσημο για τιμές του x εκατέρωθεν μιας (απλής) ρίζας, την αξιοποιεί η μέθοδος ίσων διαστημάτων, κατά την οποία το πεδίο ορισμού $[a, b]$ του x χωρίζεται σε N ίσα διαστήματα εύρους $\Delta x = (b - a) / N$ και υπολογίζεται η τιμή της $f(x)$ στα σημεία $\alpha + i\,\Delta x$, $i = 0, 1, 2, \ldots, N$. Εάν για κάποιο i προκύψει $f(a + i\Delta x) \cdot f(a + (i+1)\Delta x) < 0$, τότε υπάρχει μία τουλάχιστον πραγματική ρίζα στο διάστημα $(\alpha + i\,\Delta x,\ \alpha + (i+1)\Delta x)$.

Η επιτυχία της μεθόδου εξαρτάται άμεσα από τον αριθμό τον διαστημάτων N. Εάν το εύρος των διαστημάτων είναι μεγάλο, τότε υπάρχει αυξημένη πιθανότητα να μην εντοπισθούν κάποιες ρίζες. Όταν όμως χρησιμοποιείται πολύ μικρό εύρος (πολλά διαστήματα), τότε η μέθοδος γίνεται υπολογιστικά χρονοβόρος και ασύμφορη.

Εφαρμογή 2.1.

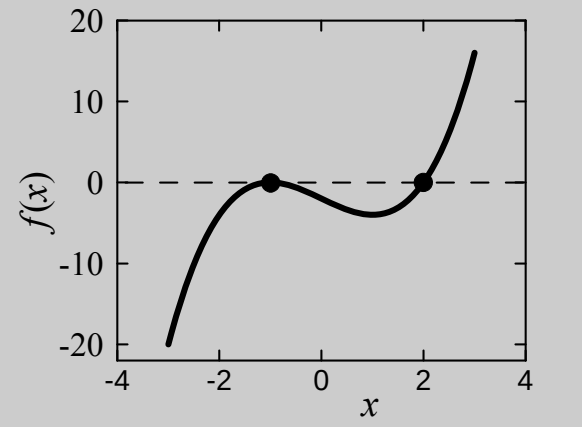
Να εντοπισθούν οι περιοχές που περιέχουν τις πραγματικές ρίζες της πολυωνυμικής εξίσωσης $f(x) = x^3 - 3x - 2$.

Έστω ότι το πεδίο ορισμού του x είναι το $[-3, +3]$ και ισομοιράζεται σε $N = 10$ διαστήματα, εύρους $\Delta x = 0{,}6$. Στη συνέχεια καταστρώνεται ο ακόλουθος πίνακας:

x	-3	-2.4	-1.8	-1.2	-0.6	0	0.6	1.2	1.8	2.4	3
$f(x)$	-20	-9.02	-2.43	-0.13	-0.42	-2	-3.58	-3.87	-1.57	+4.62	+16

Από την επισκόπηση των προσήμων του πίνακα συνάγεται ότι η εξίσωση $f(x)$ έχει μία ρίζα στην περιοχή $1.8 < x < 2.4$. Πράγματι, υπάρχει η ρίζα $x = 2$.

Το διάστημα εντοπισμού μπορεί να γίνει μικρότερο αν αυξηθεί το πλήθος των διαμερίσεων του πεδίου ορισμού του x. Για παράδειγμα, αν $N = 20$, τότε θα προκύψει η περιοχή $1.8 < x < 2.1$.



Σχήμα 2.1.
Γραφική παράσταση της εξίσωσης $f(x) = x^3 - 3x - 2$

Από τη γραφική παράσταση της εξίσωσης στο Σχήμα 2.1 φαίνεται όμως να υπάρχει κι άλλη ρίζα, στην περιοχή [–2, 0]. Πράγματι, η τιμή $x = -1$ αποτελεί επίσης ρίζα της εξίσωσης και μάλιστα διπλή: $f(x) = (x-2)\cdot(x+1)^2$. Αυτός είναι και ο λόγος που η ρίζα δεν εντοπίσθηκε από τη μέθοδο, αφού το πρόσημο της συνάρτησης εκατέρωθεν αυτής είναι το ίδιο (αρνητικό).

Όταν το πεδίο ορισμού της ανεξάρτητης μεταβλητής x έχει μεγάλο εύρος (π.χ. το $[-\infty, +\infty]$), τότε η διαδικασία εντοπισμού των ριζών μιας εξίσωσης μπορεί να επιταχυνθεί σημαντικά με το ακόλουθο τέχνασμα: διερευνάται πρώτα η ύπαρξη ριζών της $f(x) = 0$ στο διάστημα $[-1, +1]$ και μετά η ύπαρξη ριζών της $f(1/x) = 0$ στο ίδιο διάστημα. Με τον δεύτερο αυτόν έλεγχο μπορεί να εντοπισθεί η περιοχή μιας ρίζας $|x_r| > 1$, αφού οι ρίζες της $f(1/x) = 0$ είναι οι αντίστροφες των ριζών της $f(x) = 0$.

2.2.2. Άλλοι Τρόποι Εντοπισμού Ριζών

Η μέθοδος ίσων διαστημάτων έχει το πλεονέκτημα ότι μπορεί να αυτοματοποιηθεί και να προγραμματισθεί σε Η/Υ. Όμως, η χρήση της δεν εξασφαλίζει τον εντοπισμό μιας ρίζας. Ουσιαστική βοήθεια στην κατεύθυνση αυτή μπορεί να προσφέρει η γραφική παράσταση της συνάρτησης $f(x)$, όπως φάνηκε στην προηγούμενη Εφαρμογή 2.1. Η μελέτη του γραφήματος μπορεί επίσης να δώσει πληροφορίες και για την πολλαπλότητα μιας ρίζας, καθώς και για την όλη συμπεριφορά της συνάρτησης στο δεδομένο πεδίο ορισμού.

Κατά την επίλυση πρακτικών προβλημάτων, η περιοχή μιας ζητούμενης ρίζας μπορεί να καθορισθεί ή να οριοθετηθεί λαμβάνοντας υπόψη τη φυσική σημασία που πρέπει να έχει το αποτέλεσμα. Για παράδειγμα, εάν η εξίσωση της Εφαρμογής 2.1 περιγράφει το ύψος στάθμης του υγρού σε μια δεξαμενή, τότε η διερεύνηση αρκεί να γίνει μόνο στο θετικό διάστημα του πεδίου ορισμού του x. Αν επιπλέον αδιαστατοποιηθεί το ύψος της στάθμης με το ύψος της δεξαμενής, τότε η λύση θα αναζητηθεί στο [0, 1], γεγονός που υπογραμμίζει την αξία της αδιαστατοποίησης.

Ένας άλλος τρόπος εκτίμησης της περιοχής τιμών μιας ρίζας είναι η επίλυση μιας απλοποιημένης μορφής της εξίσωσης, που μπορεί να προκύψει π.χ. χρησιμοποιώντας ένα προσεγγιστικό μοντέλο για την περιγραφή ενός φαινομένου. Για παράδειγμα, η καταστατική εξίσωση των τέλειων αερίων

$$PV = RT \tag{2.7}$$

μπορεί να επιλυθεί άμεσα ως προς τον ειδικό όγκο V, δίνοντας έτσι μια προσεγγιστική τιμή της λύσης της εξίσωσης Van der Waals (2.3) για ένα πραγματικό αέριο.

Έχουν επίσης αναπτυχθεί συνθετότερες αλγεβρικές μέθοδοι εντοπισμού, στις οποίες γίνεται χρήση των ριζών της παραγώγου της συνάρτησης, $f'(x)$, που όμως είναι συνήθως εξίσου δύσκολο να επιλυθεί. Τέλος, ειδικά για τις πολυωνυμικές εξισώσεις με πραγματικούς συντελεστές

$$f(x) = \alpha_n x^n + \alpha_{n-1} x^{n-1} + \alpha_{n-2} x^{n-2} + \ldots + \alpha_0$$

προκύπτει ότι όλες οι ρίζες (πραγματικές και μιγαδικές) βρίσκονται στην περιοχή ενός κυκλικού δακτυλίου με εξωτερική και εσωτερική ακτίνα

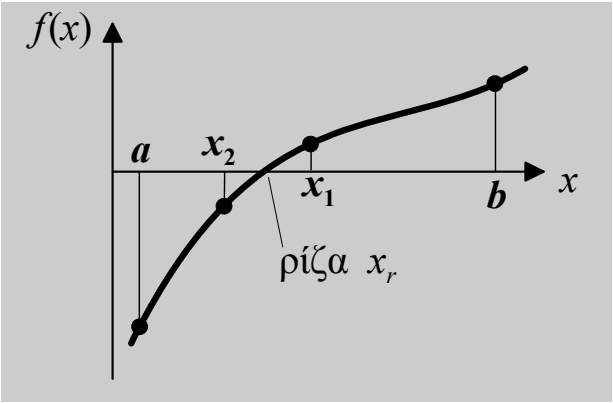
$$\begin{aligned} R_o &= 1 + \max\left(|a_{n-1}|, |a_{n-2}|, \ldots, |a_0|\right) / |a_n| \\ R_i &= |a_n| \Big/ \left[|a_n| + \max\left(|a_n|, |a_{n-1}|, \ldots, |a_1|\right)\right] \end{aligned} \tag{2.8}$$

Έτσι, για την εξίσωση $x^3 - 3x - 2 = 0$ της Εφαρμογής 2.1 προκύπτει: $R_o = 4$ και $R_i = 1/4$, δηλαδή όλες οι πραγματικές ρίζες θα βρίσκονται στο διάστημα $1/4 \le |\, x_r \,| \le 4$, όπως πράγματι ισχύει για τις ρίζες $x = -1$ (διπλή) και $x = 2$.

Έχοντας εντοπίσει το διάστημα ή τα διαστήματα που βρίσκονται οι πραγματικές ρίζες της εξίσωσης, μπορούν στη συνέχεια να χρησιμοποιηθούν διάφορες μέθοδοι για την προσέγγιση της τιμής μιας ρίζας με την επιθυμητή ακρίβεια. Οι πιο συνηθισμένες από αυτές αναλύονται στις επόμενες σελίδες.

2.3. Μέθοδος Διχοτόμησης του Διαστήματος

Η μέθοδος της διχοτόμησης βασίζεται στο θεώρημα της μέσης τιμής, όπως και η τεχνική εντοπισμού των ριζών που παρουσιάστηκε στο Κεφ. 2.2.1. Η μέθοδος απαιτεί τον προκαθορισμό μιας περιοχής τιμών (a, b) της ανεξάρτητης μεταβλητής x, μέσα στην οποία βρίσκεται η ζητούμενη ρίζα x_r της εξίσωσης, οπότε θα ισχύει $f(a)\cdot f(b)<0$ (Σχ. 2.2).



Σχήμα 2.2. Γραφική απεικόνιση της μεθόδου διχοτόμησης του διαστήματος.

Ο αλγόριθμος της μεθόδου ξεκινά με τον έλεγχο του προσήμου της συνάρτησης στο μέσο του αρχικού εύρους τιμών [a, b]: $x_1=(a+b)/2$. Η συνάρτηση $f(x_1)$ θα έχει πρόσημο $[sign\, f(x_1)]$ είτε ίδιο με το $[sign\, f(a)]$, οπότε η ρίζα θα βρίσκεται στο διάστημα $x_1<x_r<b$ είτε ίδιο με το $[sign\, f(b)]$, οπότε η ρίζα θα βρίσκεται στο διάστημα $a<x_r<x_1$ (όπως συμβαίνει στο παράδειγμα του Σχ. 2.2). Έτσι,

Αν $[sign\, f(x_1)]=[sign\, f(a)]$, τότε $x_1<x_r<b$ (2.9α)

Αν $[sign\, f(x_1)]\neq[sign\, f(a)]$, τότε $a<x_r<x_1$ (2.9β)

Η διαδικασία συνεχίζεται στο νέο διάστημα, το οποίο υποδιπλασιάζεται σε κάθε βήμα του αλγορίθμου, έως ότου προσεγγισθεί η ρίζα με την προκαθορισμένη ακρίβεια. Το απόλυτο σφάλμα της μεθόδου θα είναι το πολύ ίσο με το μισό του εύρους του τρέχοντος διαστήματος τιμών. Επομένως ο αλγόριθμος θα τερματισθεί (συγκλίνει) όταν

$$|x_m - x_{m-1}| < E_r \tag{2.10}$$

όπου E_r η επιθυμητή ακρίβεια (απόλυτο σφάλμα) και m ο αριθμός των επαναλήψεων του αλγορίθμου. Ο αλγόριθμος πρέπει επίσης να συγκλίνει και στην απίθανη αλλά όχι αδύνατη περίπτωση που θα συμβεί $f(x_m)=0$. Στη συνέχεια δίνεται ένας απλουστευμένος υπολογιστικός κώδικας της μεθόδου σε γλώσσα Fortran 77, με παράδειγμα την εξίσωση της Εφαρμογής 2.1.

Ο κώδικας αυτός χρειάζεται ακόμη αρκετές βελτιώσεις και προσθήκες, ώστε να γίνει γενικός και να λειτουργεί σωστά σε κάθε περίπτωση. Έτσι πρέπει να προβλεφθεί η περίπτωση εισαγωγής δεδομένων με λάθη, π.χ. όρια αρχικού διαστήματος a και b που να μην δίνουν $f(a)\cdot f(b)<0$, ή η περίπτωση εκτέλεσης άπειρων επαναλήψεων, εάν π.χ. το κριτήριο σύγκλισης e_r είναι μικρότερο από το μέγιστο σφάλμα στρογγυλοποίησης του εκάστοτε υπολογιστή (Κεφ. 1.2.2). Επίσης, ο όρος *term* προκύπτει με πολλαπλασιασμό δύο αριθμών που τείνουν στο μηδέν, επομένως ενδέχεται να προκληθεί σφάλμα 'underflow'.

Ακόμη, η ταχύτητα του αλγορίθμου μπορεί να βελτιωθεί εάν μειωθεί ο αριθμός κλήσεων της συνάρτησης FUNCTION ανά επανάληψη από 2 σε 1 (σημαντικό για πολύπλοκες συναρτήσεις).

Κώδικας 2.1. Μέθοδος Διχοτόμησης του Διαστήματος

```
SUBROUTINE BISECT (a, b, er, xr)
IMPLICIT REAL*8 (a–h, o–z)
xold = a
fold = FUNC (xold)
DO
    xr = (a + b) / 2.
    ea = ABS (xr – xold)
    IF ( ea .LE. er )  RETURN
    term = FUNC (xr) * FUNC (xold)
    IF ( term .LT. 0. )  THEN
        b = xr
    ELSE IF ( term .GT. 0. )  THEN
        a = xr
    ELSE
        ea = 0.
        RETURN
    ENDIF
END DO
RETURN
END
```

```
FUNCTION  FUNC (x)
IMPLICIT REAL*8 (a–h, o–z)
      FUNC (x) = x**3 – 3.*x – 2.
END
```

Η μέθοδος της διχοτόμησης είναι μια ‘κλειστή’ μέθοδος, επειδή οδηγεί πάντοτε στην εύρεση της ρίζας μιας εξίσωσης, υπό τον όρο ότι είναι γνωστή η αρχική περιοχή εντός της οποίας βρίσκεται η ρίζα και ότι δεν πρόκειται για ρίζα πολλαπλότητας $2n$ (η συνάρτηση πρέπει να έχει διαφορετικό πρόσημο εκατέρωθεν της ρίζας). Η σύγκλιση όμως είναι σχετικά αργή (γραμμική σύγκλιση). Επίσης η μέθοδος μπορεί να αποτύχει όταν η συνάρτηση δεν είναι συνεχής.

Στα πλεονεκτήματα της μεθόδου συγκαταλέγεται η δυνατότητα προκαθορισμού των επαναλήψεών της. Συγκεκριμένα, ο αριθμός των επαναλήψεων της διαδικασίας διαμερισμού του διαστήματος εξαρτάται από το αρχικό εύρος τιμών [a, b] και από την επιθυμητή απόλυτη ακρίβεια της λύσης. Σε κάθε επανάληψη το εύρος του διαστήματος στο οποίο βρίσκεται η λύση μειώνεται στο μισό, οπότε για το απόλυτο σφάλμα μετά από m επαναλήψεις θα ισχύει

$$E_t \le \frac{|b-a|}{2^m} \tag{2.11}$$

Επομένως, για να μειωθεί το σφάλμα κάτω από ένα προκαθορισμένο όριο E_r, θα πρέπει

$$\frac{|b-a|}{2^m} = E_r \qquad \text{ή} \qquad m = \lambda n\left(\frac{|b-a|}{E_r}\right) \Big/ \lambda n\, 2 \tag{2.12}$$

Παρ’ όλα αυτά, η χρήση του απόλυτου σφάλματος στο κριτήριο σύγκλισης δεν είναι η πλέον ενδεδειγμένη, αφού η τιμή του εξαρτάται από την τιμή της ζητούμενης ρίζας (βλ. και Κεφ. 1.1). Αν για παράδειγμα η ρίζα είναι πολύ μεγάλη, έστω 10^{20}, δεν έχει νόημα η

επιδίωξη εύρεσής της με απόλυτη ακρίβεια ±0.005. Αντίθετα, αν η ρίζα είναι γύρω στο μηδέν, η εύρεσή της με ακρίβεια ±0.005 θα είναι λανθασμένη.

Ως κριτήριο σύγκλισης δεν μπορεί να ληφθεί ούτε η προσέγγιση της τιμής της συνάρτησης στο μηδέν, αφού υπάρχει περίπτωση να ισχύει $|f(x_m)| \le E_r$, αλλά η τιμή x_m να διαφέρει πολύ από την πραγματική ρίζα x_r. Για παράδειγμα, στην εξίσωση $(x-1)^{10} = 0$ η τιμή $x_m = 0.5$ ικανοποιεί το κριτήριο σύγκλισης για $E_r = 0.001$, αλλά διαφέρει πολύ από την πραγματική ρίζα $x_r = 1$.

Ο ασφαλέστερος έλεγχος της σύγκλισης γίνεται με βάση την τιμή του εκτιμώμενου σχετικού σφάλματος (Κεφ. 1.1)

$$\varepsilon_a = \left| \frac{x_m - x_{m-1}}{x_m} \right| \le \varepsilon_r \qquad (2.13)$$

η χρήση του οποίου δεν επιτρέπει τον προκαθορισμό του αριθμού των επαναλήψεων, εξασφαλίζει όμως την εύρεση της ρίζας με αριθμό σημαντικών ψηφίων όχι μικρότερο του επιθυμητού (βλ. Κεφ. 1.1.1).

Εφαρμογή 2.2.

Να προσεγγισθεί η ρίζα της πολυωνυμικής εξίσωσης: $f(x) = x^3 - 3x - 2$ στην περιοχή $1.8 < x < 2.4$, με ακρίβεια 5 σημαντικών ψηφίων.

Η περιοχή [1.8, 2.4] είναι αυτή που εντοπίστηκε στην Εφαρμογή 2.1. Για επίτευξη ακρίβειας 5 σημαντικών ψηφίων, η Εξ. (1.5) δίνει:

$$\varepsilon_r \le (0.5 \cdot 10^{-n}) = (0.5 \cdot 10^{-5}) \;\Rightarrow\; \varepsilon_r \le 5 \cdot 10^{-6}$$

Ο Πίνακας 2.1 που ακολουθεί περιέχει τα αποτελέσματα της εφαρμογής του τροποποιημένου Κώδικα 2.1 (με χρήση σχετικού αντί απόλυτου σφάλματος). Εκτός από την προσέγγιση της ρίζας x_m, αναγράφονται σε κάθε επανάληψη m τα νέα όρια του διαστήματος, το εκτιμώμενο σφάλμα από την Εξ. (2.13) και το πραγματικό σχετικό σφάλμα: $\varepsilon_t = |1 - x_m / x_r|$ (λαμβάνοντας υπόψη την ακριβή τιμή της ρίζας: $x_r = 2.0$).

Πίνακας 2.1. Αποτελέσματα του κώδικα προσέγγισης ρίζας

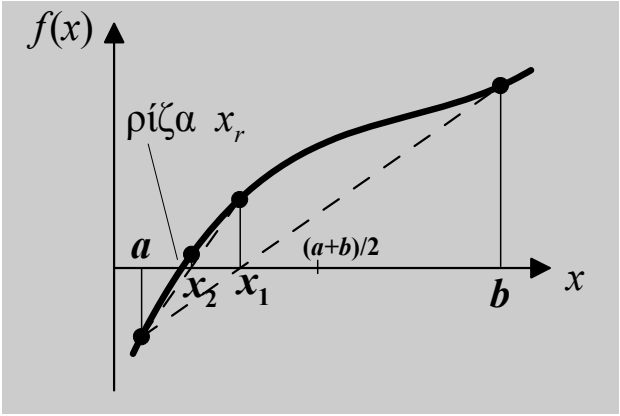
m	***a***	***b***	x_m	ε_a	ε_t
1	1.8000000	2.4000000	2.1000000	1.43E-01	5.00E-02
2	1.8000000	2.1000000	1.9500000	7.69E-02	2.50E-02
3	1.9500000	2.1000000	2.0250000	3.70E-02	1.25E-02
4	1.9500000	2.0250000	1.9875000	1.89E-02	6.25E-03
5	1.9875000	2.0250000	2.0062500	9.35E-03	3.12E-03
6	1.9875000	2.0062500	1.9968750	4.69E-03	1.56E-03
7	1.9968750	2.0062500	2.0015625	2.34E-03	7.81E-04
8	1.9968750	2.0015625	1.9992188	1.17E-03	3.91E-04
9	1.9992188	2.0015625	2.0003906	5.86E-04	1.95E-04
10	1.9992188	2.0003906	1.9998047	2.93E-04	9.77E-05
11	1.9998047	2.0003906	2.0000977	1.46E-04	4.88E-05
12	1.9998047	2.0000977	1.9999512	7.32E-05	2.44E-05
13	1.9999512	2.0000977	2.0000244	3.66E-05	1.22E-05
14	1.9999512	2.0000244	1.9999878	1.83E-05	6.10E-06
15	1.9999878	2.0000244	2.0000061	9.15E-06	3.05E-06
16	1.9999878	2.0000061	1.9999969	4.58E-06	1.53E-06

Παρατηρείται η γραμμική σύγκλιση του αλγορίθμου (το σφάλμα μειώνεται σχεδόν στο μισό σε κάθε επανάληψη), πράγμα που οδηγεί σε αρκετές επαναλήψεις ώστε να επιτευχθεί το κριτήριο σύγκλισης. Το τελευταίο εξασφαλίζει την εύρεση της ρίζας με αριθμό σημαντικών ψηφίων τουλάχιστον ίσο με τον επιθυμητό, αφού το εκτιμώμενο σφάλμα, ως το μέγιστο πιθανό, παραμένει πάντα μικρότερο του πραγματικού.

Η παραπάνω επαναληπτική διαδικασία προσεγγίζει τελικά τη ρίζα της εξίσωσης με την τιμή $x_r \approx x_{16} = 1.9999969$, δηλαδή με ακρίβεια 6 σημαντικών ψηφίων, έναντι των 5 που απαιτήθηκε. Η απόλυτη ακρίβεια της λύσης αυτής είναι: $E_t = 3.1 \cdot 10^{-6}$. Εάν αυτή η τιμή χρησιμοποιηθεί ως απόλυτο κριτήριο σύγκλισης, τότε από την Εξ. (2.12) προκύπτει $m = 17$, δηλαδή ο ίδιος σχεδόν αριθμός επαναλήψεων.

2.3.1. Μέθοδος Εσφαλμένης Θέσης ή Γραμμικής Παρεμβολής

Η μέθοδος αυτή αποτελεί παραλλαγή της μεθόδου διχοτόμησης. Η μόνη διαφορά της έγκειται στον τρόπο υπολογισμού της νέας προσέγγισης της ρίζας σε κάθε επανάληψη του αλγορίθμου: αντί για το μέσο του διαστήματος, γίνεται γραμμική παρεμβολή μεταξύ των άκρων του διαστήματος, δηλαδή σαν να επρόκειτο για γραμμική συνάρτηση (Σχ. 2.3).



Σχήμα 2.3. Γραφική απεικόνιση της μεθόδου εσφαλμένης θέσης.

Επομένως, αντί για $x_1 = (a+b)/2$, λαμβάνεται εδώ

$$x_1 = a - \frac{f(a)\cdot(b-a)}{f(b)-f(a)} \tag{2.14}$$

ενώ ο υπόλοιπος αλγόριθμος παραμένει ακριβώς ίδιος.

Είναι φανερό ότι ο υπολογισμός της Εξ. (2.14) απαιτεί περισσότερες πράξεις, αλλά η μέθοδος πολλές φορές συγκλίνει ταχύτερα από αυτήν της διχοτόμησης (π.χ. όταν η συνάρτηση είναι σχεδόν γραμμική στην περιοχή της ρίζας, όπως στο παράδειγμα του Σχήματος 2.3). Σε ορισμένες όμως μορφές συναρτήσεων μπορεί να γίνει πολύ πιο αργή. Επίσης, το εκτιμώμενο σχετικό σφάλμα από την Εξ. (2.13) δεν είναι πάντα μικρότερο από το πραγματικό, επομένως υπάρχει περίπτωση η μέθοδος να συγκλίνει χωρίς να δώσει την επιθυμητή ακρίβεια. Πάντως, στο παράδειγμα της Εφαρμογής 2.2 η μέθοδος συγκλίνει στην τιμή $x_r \approx 1.9999987$, σε οκτώ μόνο επαναλήψεις αντί για 16 της μεθόδου διχοτόμησης.

2.4. Ανοικτές μέθοδοι προσέγγισης ριζών

Οι 'ανοικτές' επαναληπτικές μέθοδοι υπολογισμού της ρίζας μιας εξίσωσης που θα παρουσιαστούν στη συνέχεια, ονομάζονται έτσι επειδή δεν προαπαιτούν τη γνώση μιας περιοχής που περικλείει τη ρίζα, αλλά μπορούν να εκκινήσουν άμεσα από κάποιο τυχαίο σημείο της συνάρτησης, κατά το δυνατόν βέβαια 'κοντά' στη ρίζα που αναζητείται.

Επιπλέον, εμφανίζουν γενικά πολύ μεγαλύτερη ταχύτητα σύγκλισης από εκείνη των 'κλειστών' μεθόδων όπως της διχοτόμησης, ενώ ορισμένες μπορούν να βρουν και διπλές ή πολλαπλές ρίζες. Όμως έχουν το μειονέκτημα ότι η σύγκλισή τους δεν είναι εξασφαλισμένη, υπάρχει δηλαδή περίπτωση απόκλισης ή ταλάντωσης.

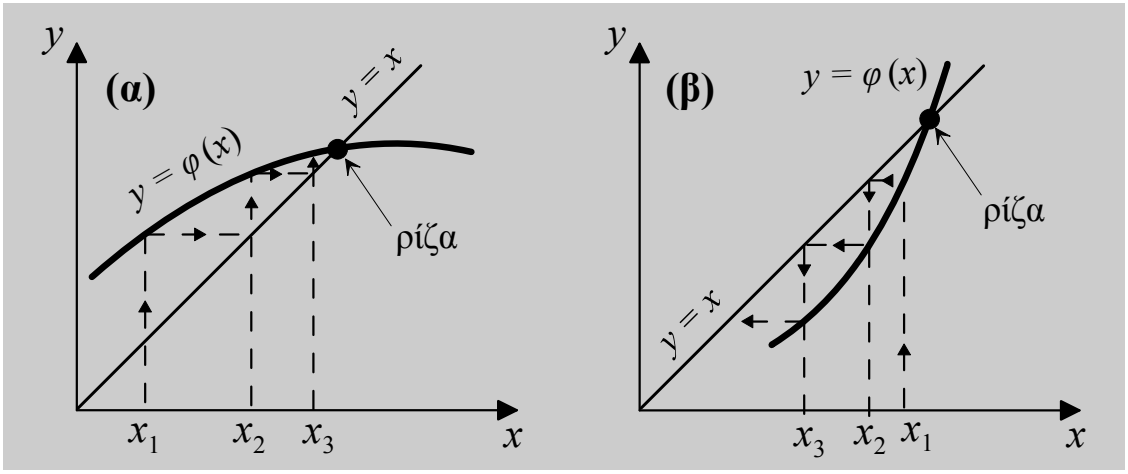
Η αρχή λειτουργίας των ανοικτών μεθόδων συνίσταται στην αναζήτηση ενός νέου σημείου x_m ως συνάρτηση ενός ή περισσότερων προηγούμενων σημείων, έτσι ώστε να προσεγγίζεται προοδευτικά όλο και περισσότερο η ρίζα x_r.

$$x_m = \varphi(x_{m-1},\ x_{m-2},\ ...,\ x_{m-k}) \qquad m = 1,\ 2,\ .... \tag{2.15}$$

Συνήθως χρησιμοποιείται ένα μόνο προηγούμενο σημείο ($k = 1$), οπότε θα είναι

$$x_m = \varphi(x_{m-1}) \qquad m = 1,\ 2,\ .... \tag{2.16}$$

Είναι φανερό ότι η ζητούμενη ρίζα θα βρίσκεται στην τομή των γραμμών $y = x$ και $y = \varphi(x)$ (Σχ. 2.4). Μια τέτοια μέθοδος δεν οδηγεί αναγκαστικά στην εύρεση της ρίζας της εξίσωσης, επειδή η σύγκλισή της εξαρτάται από την ειδική μορφή της συνάρτησης $\varphi(x)$.



Σχήμα 2.4. Γραφική απεικόνιση του αλγορίθμου μιας ανοικτής μεθόδου

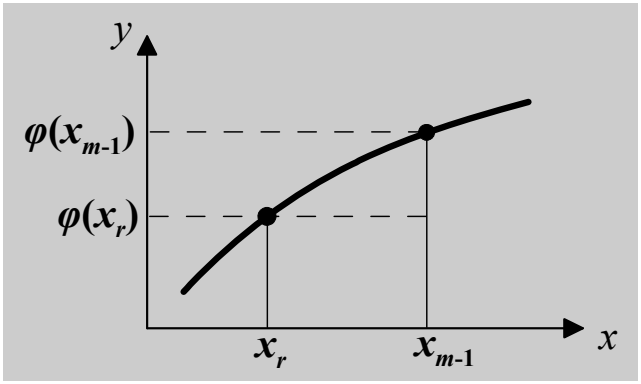
Στο παράδειγμα του Σχήματος 2.4α η μορφή της συνάρτησης $\varphi(x)$ είναι τέτοια ώστε οι διαδοχικές επαναλήψεις να οδηγούν προοδευτικά στη ρίζα της εξίσωσης, ενώ στο Σχήμα 2.4β η μορφή της $\varphi(x)$ προκαλεί απόκλιση της μεθόδου.

Για να συγκλίνει ο παραπάνω αναδρομικός τύπος θα πρέπει το (απόλυτο) σφάλμα, σ, να μειώνεται συνεχώς σε κάθε επανάληψη, δηλαδή

$$\left|\sigma_m\right| < \left|\sigma_{m-1}\right| \quad \acute{\eta} \quad \left|x_m - x_r\right| < \left|x_{m-1} - x_r\right| \tag{2.17}$$

Επειδή είναι $x_r = \varphi(x_r)$, πολύ κοντά στη ρίζα x_r ο όρος $\left|x_m - x_r\right|$ μπορεί να γραφεί εισάγοντας την παράγωγο $\varphi'(x) = d\varphi(x)/dx$, ως εξής (βλ. και Σχήμα 2.5):

$$\left|x_m - x_r\right| = \left|\varphi(x_{m-1}) - \varphi(x_r)\right| \cong \left|\varphi'(x_r)\right| \cdot \left|x_{m-1} - x_r\right| \tag{2.18}$$



Σχήμα 2.5. Επεξήγηση του κριτηρίου σύγκλισης των ανοικτών μεθόδων.

επομένως η συνθήκη σύγκλισης της Εξ. (2.17) γίνεται:

$$\left|\varphi'(x_r)\right| < 1 \tag{2.19}$$

Όταν για μια συνεχή συνάρτηση $\varphi(x)$ ισχύει η σχέση (2.19), τότε υπάρχει πάντοτε ένα διάστημα $[x_r - \delta, x_r + \delta]$, ώστε η Εξ. (2.16) να συγκλίνει στη ρίζα x_r, ξεκινώντας από τυχαίο σημείο x_0 στο διάστημα αυτό. Αυτό σημαίνει ότι η συνθήκη (2.19) είναι αναγκαία αλλά όχι και ικανή, αφού η μέθοδος μπορεί να μην συγκλίνει όταν το αρχικό σημείο ληφθεί εκτός του διαστήματος $[x_r - \delta, x_r + \delta]$, δηλαδή σχετικά 'μακριά' από τη ρίζα.

Η Εξ. (2.18) δείχνει ότι στην περιοχή της ρίζας το σφάλμα μειώνεται γραμμικά, δηλαδή $|\sigma_m| = K \cdot |\sigma_{m-1}|^1$, όπου $K = |\varphi'(x_r)|$, και η σύγκλιση είναι τόσο ταχύτερη, όσο μικρότερη είναι η τιμή $|\varphi'(x_r)|$. Όμως εάν ισχύει $\varphi'(x_r) = 0$, τότε το σφάλμα μειώνεται τετραγωνικά (βλ. Κεφ. 2.4.2). Γενικά, αποδεικνύεται ότι η σύγκλιση θα είναι βαθμού p, όπου p η τάξη της πρώτης μη–μηδενικής παραγώγου της συνάρτησης, δηλ. $\varphi^{(p)}(x_r) \neq 0$.

Η επαναληπτική διαδικασία τερματίζεται όταν δύο διαδοχικές προσεγγίσεις της ρίζας διαφέρουν μεταξύ τους λιγότερο από μια μικρή, προκαθορισμένη τιμή, δηλαδή χρησιμοποιείται το ίδιο κριτήριο τερματισμού όπως και στη μέθοδο διχοτόμησης (Εξ. 2.13).

2.4.1. Μέθοδος των Διαδοχικών Αντικαταστάσεων

Η απλούστερη εφαρμογή της παραπάνω τεχνικής γίνεται στη μέθοδο των διαδοχικών αντικαταστάσεων, κατά την οποία η εξίσωση $f(x) = 0$ γράφεται μετά από αλγεβρικές πράξεις στη μορφή $x = \varphi(x)$, δηλαδή εκφράζεται ως προς την ανεξάρτητη μεταβλητή x. Για παράδειγμα, η εξίσωση $x^3 - 3x - 2 = 0$ μπορεί να γραφεί στη μορφή $x = (3x + 2)^{1/3}$ ή και ως $x = (x^3 - 2)/3$.

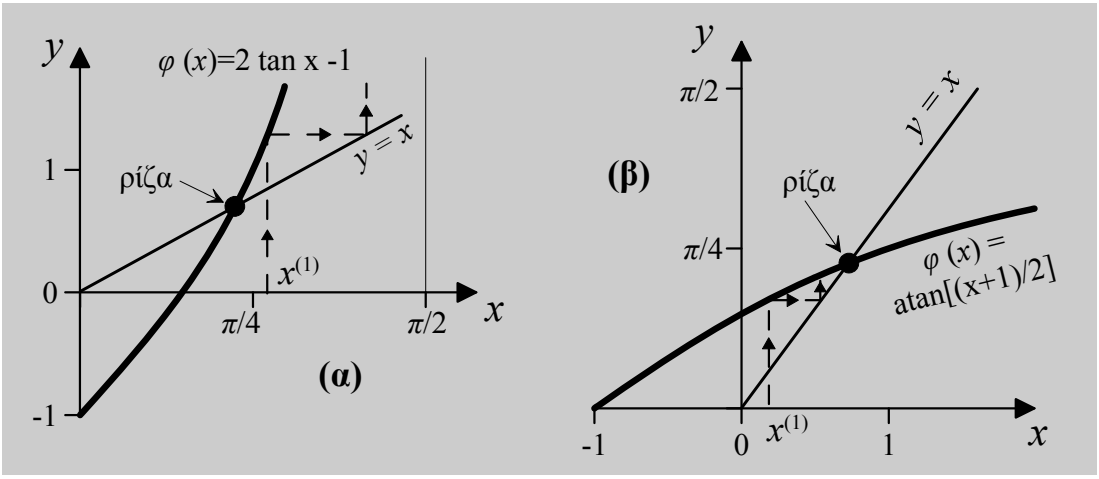
Σύμφωνα με όσα αναφέρθηκαν προηγουμένως, η σύγκλιση της μεθόδου εξαρτάται από τη συγκεκριμένη μορφή που θα χρησιμοποιηθεί. Έστω π.χ. ότι ζητούνται οι ρίζες της εξίσωσης $2\tan x - x - 1 = 0$. Αν γράψουμε

$$x = 2\tan x - 1$$

θα είναι $\varphi(x) = 2\tan x - 1$ και $\varphi'(x) = 2/\cos^2 x$, συνεπώς $\left|\varphi'(x)\right| > 1 \ \forall x$. Άρα η μέθοδος διαδοχικών αντικαταστάσεων όχι μόνο δεν θα συγκλίνει στη ρίζα της εξίσωσης, αλλά θα αποκλίνει συνεχώς. Η εξίσωση όμως μπορεί να γραφεί και ως

$$x = a\tan[(x+1)/2]$$

Τότε $\varphi(x) = a\tan\left[(x+1)/2\right]$ και $\varphi'(x) = \left[1 + \frac{(x+1)^2}{4}\right]^{-1}$, άρα $\left|\varphi'(x)\right| < 1$ πάντοτε. Έτσι η μέθοδος θα συγκλίνει στη ρίζα της εξίσωσης, όποιο κι αν είναι το σημείο εκκίνησης x_0. Οι δύο περιπτώσεις φαίνονται και στο Σχήμα 2.6.



Σχήμα 2.6. Εφαρμογή της μεθόδου διαδοχικών αντικαταστάσεων.

Ένα άλλο χαρακτηριστικό παράδειγμα εφαρμογής της μεθόδου των διαδοχικών αντικαταστάσεων είναι η αριθμητική εύρεση της τετραγωνικής ρίζας θετικού αριθμού A:

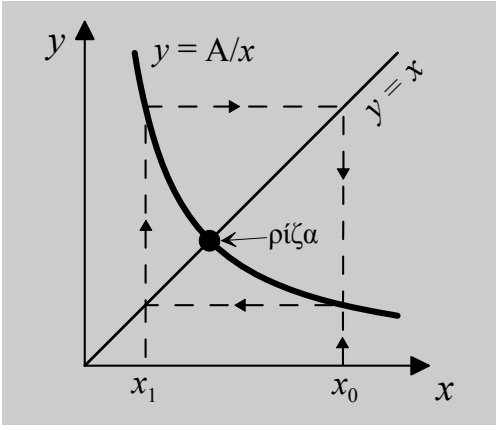
$$x^2 = A$$

Η εξίσωση αυτή μπορεί να γραφεί ως

$$x = A/x,$$

οπότε έχουμε $\varphi(x) = A/x$. Η παράγωγος της συνάρτησης $\varphi(x)$ είναι: $\varphi'(x) = -A/x^2$, οπότε $\left|\varphi'(x_r)\right| = A/x_r{}^2 = A/\left(\sqrt{A}\right)^2 = 1$

Επομένως η μέθοδος δεν θα συγκλίνει ούτε θα αποκλίνει, αλλά θα ταλαντεύεται γύρω από τη ρίζα, ανεξάρτητα από την πρώτη εκτίμηση x_0, ακόμη κι αν $\left|\varphi'(x_0)\right| < 1$. Γραφική παράσταση του αλγορίθμου εύρεσης της ρίζας παρουσιάζεται στο Σχήμα 2.7, όπου



Σχήμα 2.7. Γραφική παράσταση εύρεσης τετραγωνικής ρίζας αριθμού *A*.

διαπιστώνεται η αδυναμία σύγκλισης. Αναλυτικά, για π.χ. $A = 2$ και $x_0 = 2$, είναι $|\varphi'(2)| = 1/2 < 1$, αλλά προκύπτει $x_1 = 1/2, \quad x_2 = 2, \quad x_3 = 1/2, \quad x_4 = 2$ κ.ο.κ.

Έστω τώρα ότι η ίδια εξίσωση γράφεται στην ακόλουθη μορφή:

$$x = A/x \;\Rightarrow\; 2x = A/x + x \;\Rightarrow\; x = (A/x + x)/2 = \varphi(x)$$

Τότε η παράγωγος γίνεται $$|\varphi'(x_r)| = \left| -\frac{A}{2x_r^2} + \frac{1}{2} \right| = \left| -\frac{A}{2\left(\sqrt{A}\right)^2} + \frac{1}{2} \right| = 0$$

δηλαδή ο αλγόριθμος τώρα όχι μόνο θα συγκλίνει, αλλά και με τετραγωνική σύγκλιση (Κεφ. 2.4.2). Πράγματι, για τις τιμές $A = 2$ και $x_0 = 2$ λαμβάνουμε ($\sqrt{2} = 1.4142136$): $x_1 = 1.5 \quad x_2 = 1.4166667 \quad x_3 = 1.4142157$. Το παράδειγμα αυτό υποδεικνύει ότι η μέθοδος μπορεί να επιταχυνθεί σημαντικά, εάν με κάποιο αλγεβρικό τέχνασμα δημιουργηθεί μια συνάρτηση $\varphi(x)$ με πολύ μικρή ή μηδενική κλίση στην περιοχή της ζητούμενης ρίζας.

Εφαρμογή 2.3.

Με τη μέθοδο των διαδοχικών αντικαταστάσεων να προσεγγισθεί μία ρίζα της εξίσωσης: $f(x) = x^3 - 3x - 2$, με ακρίβεια 5 σημαντικών ψηφίων.

Όπως έχει βρεθεί, η εξίσωση έχει μία απλή ρίζα: $x = 2$ και μία διπλή: $x = -1$.
Αν γραφεί στη μορφή: $x = (x^3 - 2)/3 = \varphi(x)$, τότε $|\varphi'(x)| = x^2$, άρα το κριτήριο σύγκλισης δεν ικανοποιείται για καμμία ρίζα.
Η μορφή όμως: $x = (3x + 2)^{1/3}$ δίνει: $|\varphi'(x)| = (3x+2)^{-2/3} < 1$ για $x > -2/3$
Άρα ο αλγόριθμος θα συγκλίνει στην απλή ρίζα $x_r = 2$ για κάθε $x_0 > -2/3$.

Για λόγους σύγκρισης με τη μέθοδο διχοτόμησης, έστω $x_0 = 2.4$. Όπως στην Εφαρμογή 2.2, λαμβάνεται όριο σύγκλισης: $\varepsilon_r = 5 \cdot 10^{-6}$. Στον Πίνακα 2.2 δίνονται τα αποτελέσματα του αλγορίθμου για την προσέγγιση της ρίζας x_m, το εκτιμώμενο σχετικό σφάλμα και την τιμή της συνάρτησης $f(x)$, Η τελευταία πρέπει να ελέγχεται ότι προσεγγίζει το μηδέν (επαλήθευση), ώστε να αποτραπεί τυχόν σφάλμα, π.χ. κατά τη δημιουργία της $\varphi(x)$.

Πίνακας 2.2. Αποτελέσματα κώδικα διαδοχικών αντικαταστάσεων.

m	x_m	ε_a	$f(x_m)$
1	2.0953791	1.454E-01	9.139E-01
2	2.0235660	3.549E-02	2.154E-01
3	2.0058742	8.820E-03	5.308E-02
4	2.0014675	2.202E-03	1.322E-02
5	2.0003668	5.502E-04	3.302E-03
6	2.0000917	1.375E-04	8.253E-04
7	2.0000229	3.439E-05	2.063E-04
8	2.0000057	8.596E-06	5.158E-05
9	2.0000014	2.149E-06	1.289E-05

Η μέθοδος συγκλίνει γραμμικά αλλά ταχύτερα από τη μέθοδο της διχοτόμησης (σε 9 αντί σε 16 επαναλήψεις), επειδή εδώ ισχύει: $|\varphi'(x_r)| = (6+2)^{-2/3} = 0.25$, ενώ στη μέθοδο διχοτόμησης είναι: $|\sigma_m| \approx 0.5 \cdot |\sigma_{m-1}|$.

2.4.2. Μέθοδος Newton – Raphson

Η δημοφιλέστερη ανοικτή μέθοδος είναι η Newton-Raphson (N-R), το βασικό πλεονέκτημα της οποίας είναι ότι επιτυγχάνει τετραγωνική σύγκλιση κοντά στη ρίζα. Ο αναδρομικός τύπος της μεθόδου προκύπτει ως εξής: Έστω ότι μια προσέγγιση στη ρίζα x_r της εξίσωσης $f(x) = 0$ είναι η x_1, οπότε

$$x_r = x_1 + h$$

όπου h μικρός αριθμός και $f(x_1 + h) = 0$. Η σχέση αυτή αναλύεται σε σειρά Taylor:

$$f(x_1 + h) = f(x_1) + h \cdot f'(x_1) + \frac{h^2}{2!} \cdot f''(x_1) + ...$$

και αμελώντας τους όρους δεύτερης τάξης και άνω, προκύπτει

$$h \cong -f(x_1) / f'(x_1)$$

Συνεπώς μπορούμε να δημιουργήσουμε μια καλύτερη προσέγγιση, x_2, της ρίζας

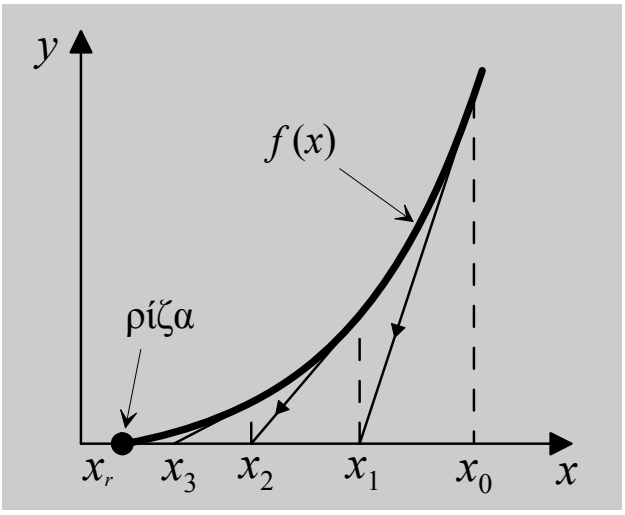
$$x_2 = x_1 - f(x_1) / f'(x_1)$$

Η σχέση αυτή παρέχει τον αναδρομικό τύπο της μεθόδου Newton-Raphson

$$x_m = \varphi(x_{m-1}) = x_{m-1} - \frac{f(x_{m-1})}{f'(x_{m-1})} \qquad m = 1, 2, ... \tag{2.20}$$

Σημειώνεται ότι λόγω της παράλειψης των όρων ανώτερης τάξης στο ανάπτυγμα Taylor, μπορεί να προκληθεί σημαντικό σφάλμα και απόκλιση της μεθόδου εάν η αρχική εκτίμηση x_0 στη ρίζα είναι ανεπιτυχής (δεν είναι αρκετά 'κοντά' στη ρίζα).

Γραφικά η μέθοδος παρίσταται στο Σχήμα 2.8. Η εφαπτομένη (παράγωγος) της συνάρτησης $f(x)$ σε ένα σημείο x_i κοντά στη ρίζα τέμνει τον άξονα $y = 0$ σε σημείο x_{i+1}, το οποίο προσεγγίζει περισσότερο τη ρίζα.



Σχήμα 2.8. Γραφική επεξήγηση της μεθόδου Newton – Raphson.

Η τετραγωνική σύγκλιση της μεθόδου αποδεικνύεται βρίσκοντας την τιμή της παραγώγου της συνάρτησης $\varphi(x)$ (Εξ. 2.20), στη θέση της ρίζας x_r:

$$\varphi'(x_r) = 1 - \frac{[f'(x_r)]^2 - f(x_r)\cdot f''(x_r)}{[f'(x_r)]^2} = \frac{f(x_r)\cdot f''(x_r)}{[f'(x_r)]^2} \tag{2.21}$$

Αλλά $f(x_r) = 0$, επομένως όταν πρόκειται για απλή ρίζα η παραπάνω σχέση δίνει πάντα $\varphi'(x_r) = 0$. Αντίθετα, για πολλαπλή ρίζα ισχύει και $f'(x_r) = 0$ (παρονομαστής της 2.21). Στην περίπτωση αυτή αποδεικνύεται ότι (βλ. Κεφ. 2.6.1)

$$\varphi'(x_r) = 1 - \frac{1}{k} \tag{2.22}$$

δηλαδή η σύγκλιση μεταπίπτει σε γραμμική και μάλιστα με ρυθμό που μειώνεται όσο μεγαλύτερος είναι ο βαθμός πολλαπλότητας k της ρίζας.

Ο ακριβής ρυθμός σύγκλισης της μεθόδου σε απλή ρίζα προκύπτει αναπτύσσοντας σε σειρά Taylor τη συνάρτηση $\varphi(x)$ γύρω από τη ρίζα x_r (έστω $h = x_{m-1} - x_r$):

$$\varphi(x_{m-1}) = \varphi(x_r) + (x_{m-1} - x_r)\cdot\varphi'(x_r) + \frac{(x_{m-1} - x_r)^2}{2!}\cdot\varphi''(x_r) + \ldots\ldots \Rightarrow$$

$$\Rightarrow \quad x_m \cong x_r + \frac{(x_{m-1} - x_r)^2}{2!}\cdot\varphi''(x_r) \;\Rightarrow\; |\sigma_m| \cong \left|\frac{\varphi''(x_r)}{2}\right|\cdot|\sigma_{m-1}|^2 \tag{2.23}$$

Παραγωγίζοντας την Εξ. (2.21), λαμβάνουμε μετά την εκτέλεση αλγεβρικών πράξεων ότι $\varphi''(x_r) = f''(x_r)/f'(x_r)$. Έτσι η (2.23) δίνει τελικά

$$|\sigma_m| \cong \left|\frac{f''(x_r)}{2\cdot f'(x_r)}\right|\cdot|\sigma_{m-1}|^2 \tag{2.24}$$

Επομένως η σύγκλιση είναι βαθμού 2 (τετραγωνική), με ρυθμό που αυξάνει όσο μειώνεται η τιμή του όρου $f''(x_r)/f'(x_r)$. Αν είναι $f''(x_r) = 0$ θα έχουμε κυβική σύγκλιση κ.ο.κ.

Εξετάζοντας και πάλι ως παράδειγμα την εύρεση της τετραγωνικής ρίζας ενός θετικού αριθμού A (βλ. Κεφ. 2.4.1)

$$x^2 = A \quad \text{ή} \quad f(x) = x^2 - A$$

η παράγωγος της $f(x)$ είναι $f'(x) = 2x$, οπότε η μέθοδος Newton-Raphson δίνει τον αναδρομικό τύπο

$$x_m = x_{m-1} - \frac{x_{m-1}^2 - A}{2x_{m-1}} = \frac{1}{2}\cdot\left(x_{m-1} + \frac{A}{x_{m-1}}\right) = \varphi(x_{m-1})$$

Λαμβάνεται δηλαδή η ίδια έκφραση που προέκυψε με αλγεβρικό τέχνασμα και κατά την εφαρμογή της μεθόδου διαδοχικών αντικαταστάσεων στο Κεφ. 2.4.1, ώστε να είναι $\varphi''(x_r) = 0$.

Δίνεται στη συνέχεια μια σχεδόν πλήρης μορφή του κώδικα της μεθόδου Newton-Raphson σε Fortran 77, όπως χρησιμοποιείται στην επόμενη Εφαρμογή 2.4. Δεδομένα εισόδου στην υποπρόγραμμα είναι η αρχική εκτίμηση της ρίζας x_0, το κριτήριο σύγκλισης ε_r και ένα όριο αριθμού επαναλήψεων *nmax*, για την περίπτωση που η μέθοδος δεν συγκλίνει. Εκτός από την προσέγγιση x_r επιστρέφει για επαλήθευση και η τιμή της συνάρτησης f_r. Στον κώδικα χρειάζεται ακόμη να προβλεφθεί η περίπτωση $f(x_0) = 0$, καθώς και το ενδεχόμενο να είναι $x_r = 0$, οπότε το σχετικό σφάλμα ε_a δεν ορίζεται. Επίσης, δεν πρέπει να επιτρέπεται η παράγωγος να πάρει την τιμή μηδέν.

Κώδικας 2.2. Μέθοδος Newton – Raphson

```
SUBROUTINE NEWTR (x0, er, nmax, xr, fr)
IMPLICIT REAL*8 (a–h, o–z)
niter = 0
xold = x0
DO
   niter = niter + 1
   xr = xold – FF (xold) / FD1 (xold)
   fr = FF (xr)
   IF ( fr .EQ. 0. )  THEN
      ea = 0.
      RETURN
   ENDIF
   ea = ABS ((xr – xold) / xr )
   IF ( ea .LE. er )  RETURN
   IF ( niter .EQ. nmax )  RETURN
   xold = xr
END DO
RETURN
END
```

```
FUNCTION  FF (x)
IMPLICIT REAL*8 (a–h, o–z)
      FF (x) = x**3 – 3.*x – 2.
END
```

```
FUNCTION  FD1 (x)
IMPLICIT REAL*8 (a–h, o–z)
      FD1 (x) = 3.*x**2 – 3.
END
```

Εφαρμογή 2.4.

Να βρεθεί με τη μέθοδο Newton-Raphson και με ακρίβεια 5 σημαντικών ψηφίων μία ρίζα της εξίσωσης: $f(x) = x^3 - 3x - 2$.

Ως τιμή εκκίνησης της μεθόδου δίνεται και πάλι η x_0 = 2.4, όπως στις προηγούμενες Εφαρμογές 2.3 & 2.2. Ομοίως το όριο σύγκλισης: $\varepsilon_r = 5 \cdot 10^{-6}$ Με αυτά τα δεδομένα ο Κώδικας 2.2 παράγει τα εξής αποτελέσματα:

Πίνακας 2.3. Αποτελέσματα εφαρμογής της μεθόδου N-R.

m	x_m	ε_a	ε_t	$f(x_m)$
1	2.076190476	1.56E-01	3.81E-02	7.21E-01
2	2.003596011	3.62E-02	1.80E-03	3.24E-02
3	2.000008590	1.79E-03	4.29E-06	7.73E-05
4	2.000000000	4.29E-06	2.46E-11	4.43E-10

Η μέθοδος συγκλίνει τώρα σε 4 μόνο επαναλήψεις και μάλιστα η ρίζα βρίσκεται με ακρίβεια 10 σημαντικών ψηφίων, αφού το εκτιμώμενο σχετικό σφάλμα είναι πολύ μεγαλύτερο του πραγματικού.

Στα αποτελέσματα του Πίνακα 2.3 μπορεί επίσης να παρατηρηθεί αυτό που υπονοεί ο όρος ‘τετραγωνική σύγκλιση’: ο αριθμός των σωστών σημαντικών ψηφίων της προσεγγιστικής λύσης διπλασιάζεται σε κάθε επανάληψη του αλγορίθμου. Πράγματι, εδώ η λύση εμφανίζει κατά σειρά: 1, 3, 5 και 10 σωστά σημαντικά ψηφία.

Μειονέκτημα της μεθόδου Newton-Raphson είναι ότι χρειάζεται την αναλυτική έκφραση της παραγώγου της συνάρτησης, που μπορεί να είναι δύσκολο να βρεθεί σε μια πολύπλοκη συνάρτηση ή να γίνει κάποιο λάθος στην έκφρασή της, ή ακόμη και να μην υπάρχει (μη–παραγωγίσιμη συνάρτηση).

Επίσης, αναφέρθηκε προηγουμένως ότι η μέθοδος N-R είναι ευαίσθητη στην επιλογή της τιμής εκκίνησης και υπάρχει περίπτωση να μην συγκλίνει εάν αυτή δεν βρίσκεται αρκετά 'κοντά' στη ρίζα της εξίσωσης. Ακόμη και τότε όμως, η ύπαρξη της παραγώγου της συνάρτησης στον παρονομαστή του αναδρομικού τύπου δημιουργεί αστάθεια στον αλγόριθμο: μια μικρή τιμή της παραγώγου σε κάποια αρχική επανάληψη μπορεί να προκαλέσει 'μεταπήδηση' σε άλλη περιοχή του πεδίου ορισμού της συνάρτησης, μακριά από τη ρίζα. Επιπλέον, η σύγκλιση γίνεται τετραγωνική μόνο στην περιοχή της ρίζας, ενώ πιο μακριά μπορεί σε ορισμένες συναρτήσεις να έχει πολύ πιο αργό ρυθμό.

Στις περισσότερες όμως πρακτικές περιπτώσεις η μέθοδος N-R λειτουργεί χωρίς προβλήματα και γι' αυτό προτιμάται της ευσταθέστερης αλλά πολύ πιο αργής μεθόδου των διαδοχικών αντικαταστάσεων. Σε δύσκολες περιπτώσεις μπορεί να συνδυάζεται με τη μέθοδο της διχοτόμησης, ώστε να εντοπίζεται πρώτα η περιοχή της ρίζας. Τέλος, όταν δεν μπορεί να βρεθεί η παράγωγος χρησιμοποιείται μια παραλλαγή της, η μέθοδος της τέμνουσας.

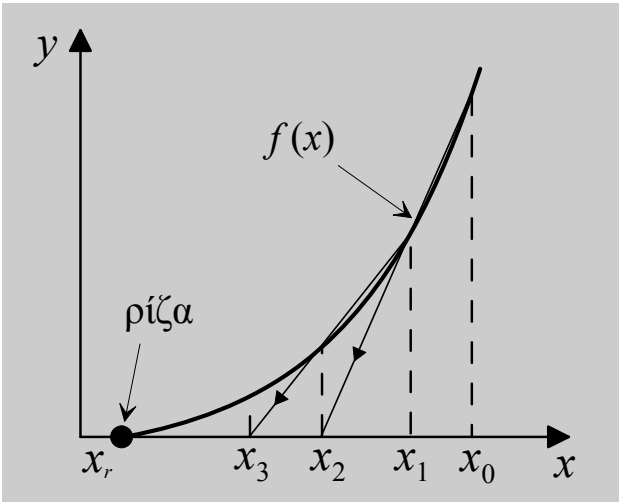
2.4.3. Μέθοδος της Τέμνουσας (Secant)

Στην περίπτωση όπου η παράγωγος της $f(x)$ δεν είναι εύκολο να υπολογισθεί, τότε μπορεί να προσεγγισθεί αριθμητικά με μια έκφραση πεπερασμένων διαφορών (Κεφ. 5.2.1)

$$f'(x_2) = \left[f(x_2) - f(x_1)\right] / (x_2 - x_1)$$

όπου x_1 και x_2 δύο κοντινά σημεία του πεδίου ορισμού της. Έτσι στη μέθοδο N-R μπορούν να χρησιμοποιηθούν δύο διαδοχικές προσεγγίσεις της ρίζας, όπως φαίνεται και στο Σχήμα 2.9, οπότε ο αναδρομικός τύπος (2.20) γίνεται

$$\mathrm{x_m} = \varphi(\mathrm{x_{m-1}}) = \mathrm{x_{m-1}} - \frac{\mathrm{f(x_{m-1})} \cdot (\mathrm{x_{m-1}} - \mathrm{x_{m-2}})}{\mathrm{f(x_{m-1})} - \mathrm{f(x_{m-2})}} \qquad \mathrm{m} = 1,\ 2,\ \ldots. \tag{2.25}$$



Σχήμα 2.9. Γραφική παράσταση της τροποποιημένης μεθόδου N-R ή μεθόδου της τέμνουσας.

Η μέθοδος είναι γνωστή στη βιβλιογραφία ως τροποποιημένη μέθοδος N-R ή ως μέθοδος της τέμνουσας και έχει το μειονέκτημα ότι απαιτεί δύο αρχικές τιμές x_0 και x_1 για την εκκίνησή της. Συνεπώς η μέθοδος μπορεί να εφαρμοσθεί αφού βρεθούν πρώτα δύο διαδοχικές προσεγγίσεις της ρίζας με κάποια άλλη μέθοδο ή και με εκτίμηση. Γίνεται όμως και αυτοκινούμενη, εάν η παράγωγος στην πρώτη επανάληψη υπολογισθεί από τη σχέση

$$f'(x_0) = \left[f(x_0 + \delta x) - f(x_0) \right] / \delta x \tag{2.26}$$

Το διάστημα δx δεν πρέπει να είναι πολύ μικρό, για να μην προκληθεί μεγάλο σφάλμα στρογγυλοποίησης (βλ. Κεφ. 5.2.1).

Η μέθοδος δεν έχει τετραγωνική σύγκλιση, αφού τώρα $\varphi'(x_r) \neq 0$, έτσι η ταχύτητά της είναι μικρότερη της N-R, αλλά μεγαλύτερη της γραμμικής. Για την ακρίβεια, αποδεικνύεται ότι η σύγκλιση είναι 'υπεργραμμική', με βαθμό 1.62:

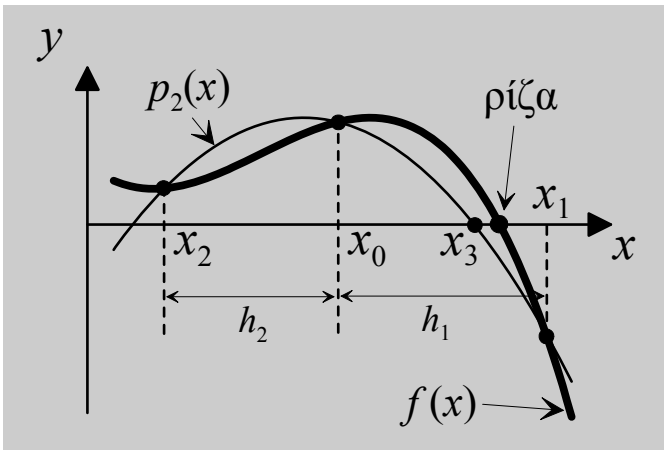
$$|\sigma_m| \cong |g(x_r)| \cdot |\sigma_{m-1}|^{1,62} \tag{2.27}$$

Τέλος, όσον αφορά στην ευστάθεια του αλγορίθμου, η μέθοδος εμφανίζει όλα τα πιθανά προβλήματα της Newton-Raphson, συν ένα πρόσθετο: το σφάλμα στρογγυλοποίησης μπορεί να γίνει σημαντικό ποσοστό της $f(x)$, καθώς η τελευταία τείνει προς το μηδέν ($f(x_r) = 0$). Αυτό ενδέχεται να οδηγήσει σε λάθος τιμή της παραγώγου, με συνέπεια πιθανή απόκλιση ή μεταπήδηση του αλγορίθμου.

Για την εξίσωση της Εφαρμογής 2.4, χρησιμοποιώντας τα ίδια δεδομένα και με σημεία εκκίνησης στα όρια του διαστήματος εντοπισμού της ρίζας $x_0 = 1.8$ και $x_1 = 2.4$, ο αλγόριθμος της μεθόδου συγκλίνει σε 5 επαναλήψεις, δηλαδή μία περισσότερη από τη N-R, και βρίσκει τη λύση με 9 σωστά σημαντικά ψηφία.

2.4.4. Η Μέθοδος του Müller

Σε όλες τις προηγούμενες μεθόδους η συνάρτηση $f(x)$ προσεγγίζεται τοπικά με μια ευθεία γραμμή, η τομή της οποίας με τον άξονα $y = 0$ δίνει την εκάστοτε προσέγγιση της ρίζας. Όμως ένα πολυώνυμο μεγαλύτερου βαθμού αποτελεί καλλίτερη προσέγγιση μιας συνάρτησης και σ' αυτό βασίζεται η μέθοδος του Müller, η οποία χρησιμοποιεί πολυώνυμο δευτέρου βαθμού, που διέρχεται από τρία σημεία της συνάρτησης στις θέσεις x_0, x_1 και x_2, όπως φαίνεται στο Σχήμα 2.10.



Σχήμα 2.10. Γραφική παράσταση της μεθόδου του Müller.

Οι συντελεστές του πολυωνύμου $p_2(x) = \alpha_2 x^2 + \alpha_1 x + \alpha_0$ προκύπτουν από τις εξισώσεις που ισχύουν στα σημεία αυτά $p_2(x_i - x_0) = f(x_i),\ \ i = 0, 1, 2$, ως εξής:

$$\alpha_2 = \frac{h_2 \cdot f(x_1) - (h_1 + h_2) \cdot f(x_0) + h_1 \cdot f(x_2)}{h_1 h_2 \cdot (h_1 + h_2)},$$
$$\alpha_1 = \frac{f(x_1) - f(x_0) - \alpha_2 \cdot h_1^2}{h_1}, \qquad \alpha_0 = f(x_0) \tag{2.28}$$

Η δευτεροβάθμια πολυωνυμική εξίσωση $p_2(x) = 0$ έχει δύο ρίζες, που δίνονται από τη σχέση (βλ. και Εφαρμ. 1.7)

$$x_3 = x_0 - \frac{-2\alpha_0}{\alpha_1 \pm \sqrt{\alpha_1^2 - 4\alpha_2\alpha_0}} \tag{2.29}$$

Από αυτές λαμβάνεται εκείνη που είναι πιο κοντά στο x_0, δηλαδή που έχει τον μεγαλύτερο παρονομαστή (π.χ. λαμβάνεται πρόσημο '+' όταν ο συντελεστής α_1 είναι θετικός).

Η ρίζα x_3 και το σημείο x_0 αποτελούν τα δύο από τα τρία σημεία για την επόμενη επανάληψη της μεθόδου. Το τρίτο είναι είτε το x_1 είτε το x_2, αναλόγως ποιο από τα δύο βρίσκεται στην ίδια μεριά με το x_3 ως προς το x_0. Πριν την εκτέλεση της επόμενης επανάληψης οι δείκτες των σημείων αναδιατάσσονται, ώστε το ενδιάμεσο σημείο να συμβολίζεται πάντα ως x_0.

Παρά το γεγονός ότι απαιτεί τρεις αρχικές τιμές για την εκκίνηση, η μέθοδος Müller πλεονεκτεί της Newton-Raphson στο ότι δεν χρειάζεται την παράγωγο της συνάρτησης, ενώ επίσης είναι λίγο ταχύτερη της μεθόδου της τέμνουσας. Η σύγκλισή της είναι σχεδόν τετραγωνική (σε απλή ρίζα, με θεωρητικό βαθμό $p = 1.84$) και μπορεί να χρησιμοποιηθεί σε κάθε εξίσωση. Ιδιαίτερα αποτελεσματική είναι όμως στις αλγεβρικές εξισώσεις, όπου συγκλίνει σχεδόν για οποιαδήποτε εκλογή αρχικών τιμών. Επίσης, αντίθετα με τις άλλες ανοικτές μεθόδους, μπορεί να συγκλίνει σε μια μιγαδική ρίζα χωρίς να χρειάζεται μιγαδική τιμή εκκίνησης (βλ. Κεφ. 2.6.2).

Η μέθοδος Müller μπορεί επίσης να προγραμματισθεί και ως 'κλειστή' μέθοδος, όπως δηλαδή και η μέθοδος της διχοτόμησης διαστήματος, ώστε να έχει εξασφαλισμένη σύγκλιση. Σε κάθε επανάληψη ο αλγόριθμος επιλέγει τα τρία νέα σημεία έτσι ώστε σε δύο εξ αυτών η συνάρτηση να έχει διαφορετικό πρόσημο. Εξυπακούεται βέβαια ότι αυτό πρέπει να ισχύει και για τα σημεία εκκίνησης της μεθόδου.

2.4.5. Άλλες Ανοικτές Μέθοδοι

Οι προηγούμενες τεχνικές προσέγγισης ριζών είναι οι δημοφιλέστερες μεταξύ πολλών άλλων που έχουν αναπτυχθεί. Γενικά, οι πιο πολύπλοκες μέθοδοι έχουν αυξημένη ταχύτητα σύγκλισης ή μεγαλύτερη ευστάθεια, αλλά απαιτούν περισσότερες πράξεις ανά επανάληψη. Ως παράδειγμα αναφέρεται ο αλγόριθμος Traub, που μοιάζει με τη μέθοδο της τέμνουσας αλλά υπολογίζει την παράγωγο από το ανάπτυγμα Taylor με ακρίβεια δεύτερης τάξης, καθώς και η μέθοδος του Wegstein, που αποτελεί συνδυασμό των μεθόδων της τέμνουσας και των διαδοχικών αντικαταστάσεων, επιτυγχάνοντας μεγαλύτερη ευστάθεια.

Υβριδικά σχήματα που συνδυάζουν δύο ή περισσότερες μεθόδους προσέγγισης χρησιμοποιούνται αρκετές φορές σε πρακτικά προβλήματα, όπου συνήθως ζητούνται πολλαπλές επιλύσεις μιας συνάρτησης γνωστής μορφής. Έτσι, μια σχετικά αργή 'κλειστή' μέθοδος, που δίνει όμως σίγουρα λύση, μπορεί να προηγείται μιας ταχύτερης 'ανοικτής' μεθόδου για να παρέχει καλές αρχικές τιμές ή και να εναλλάσσεται με αυτήν, ώστε να σταθεροποιεί τη διαδικασία σύγκλισης.

2.5. Εύρεση των Ριζών Πολυωνύμων

Οι πολυωνυμικές συναρτήσεις αποτελούν ένα πολύτιμο και συχνά χρησιμοποιούμενο εργαλείο στην αριθμητική ανάλυση, επειδή εμφανίζουν μερικές σημαντικές ιδιότητες που δεν τις έχουν άλλες συναρτήσεις. Μια πολυωνυμική συνάρτηση βαθμού n, της μορφής

$$p_n(x) = \alpha_n x^n + \alpha_{n-1} x^{n-1} + \alpha_{n-2} x^{n-2} + ... + \alpha_0 \tag{2.30}$$

είναι παντού συνεχής και το ίδιο ισχύει για τις παραγώγους και το ολοκλήρωμά της, η αναλυτική έκφραση των οποίων προκύπτει πολύ εύκολα. Επίσης, ο υπολογισμός των τιμών της σε ηλ. υπολογιστή είναι ταχύτατος, αφού περιλαμβάνει μόνο τις τέσσερις πράξεις. Τα πολυώνυμα χρησιμοποιούνται σε πολλές αριθμητικές μεθόδους, κυρίως παρεμβολής και προσέγγισης συναρτήσεων, αλλά και αριθμητικής ολοκλήρωσης, παραγώγισης, κ.ά. Αλλά και στις περισσότερες μεθόδους εύρεσης ρίζας που παρουσιάστηκαν, μια μη–γραμμική εξίσωση προσεγγίζεται με ένα πολυώνυμο $1^{ου}$ βαθμού, του οποίου η ρίζα προκύπτει άμεσα.

Μια άλλη σημαντική εφαρμογή των πολυωνύμων είναι η χρήση τους για τον χαρακτηρισμό δυναμικών συστημάτων (μηχανισμοί, κατασκευές κλπ.), με τη μορφή μιας χαρακτηριστικής εξίσωσης, οι ρίζες της οποίας αντιστοιχούν στις ιδιοτιμές του συστήματος.

Η συστηματική μελέτη των πολυωνύμων έχει οδηγήσει στη διατύπωση πολλών βασικών θεωρημάτων, μερικά μόνο εκ των οποίων, που αφορούν τις ρίζες εξισώσεων, αναφέρονται εδώ (βλ. και θεώρημα οριοθέτησης ριζών, Κεφ. 2.2.2).

- *Θεμελιώδες θεώρημα της άλγεβρας*: Ένα πολυώνυμο n βαθμού έχει ακριβώς n ρίζες (πραγματικές ή μιγαδικές), όπου μια ρίζα πολλαπλότητας k προσμετράται k φορές.
- *Μοναδικότητα*: Από n+1 σημεία περνά ακριβώς μόνο ένα πολυώνυμο n βαθμού.
- *Κανόνας του προσήμου του Descartes*: Οι θετικές πραγματικές ρίζες ενός πολυωνύμου $p_n(x) = 0$ είναι τόσες όσες και οι μεταβολές στο πρόσημο των (πραγματικών) συντελεστών του, α_i, $i = n, n\text{-}1, \ldots, 0$ ή λιγότερες κατά έναν άρτιο αριθμό. Το ίδιο ισχύει για τις αρνητικές ρίζες, όταν λαμβάνονται υπόψη τα πρόσημα του $p_n(-x) = 0$.

Εφαρμογή του τελευταίου θεωρήματος στην πολυωνυμική εξίσωση $p(x) = x^3 - 3x - 2 = 0$ δίνει μία θετική ρίζα (που είναι η x = +2) και 2 ή 0 αρνητικές ρίζες (η διπλή ρίζα x = –1).

Η ευρεία εφαρμογή των πολυωνύμων έχει επίσης συντελέσει στην ανάπτυξη γενικών αριθμητικών μεθόδων για την εύρεση όλων των ριζών τους, πράγμα που δεν συμβαίνει για τις υπερβατικές εξισώσεις. Εξυπακούεται ότι όλες οι αριθμητικές μέθοδοι που αναφέρθηκαν προηγουμένως μπορούν να χρησιμοποιηθούν και για την εύρεση μίας ρίζας ενός πολυωνύμου.

2.5.1. Μέθοδος Newton

Η μέθοδος Newton είναι συνδυασμός της γνωστής μεθόδου Newton-Raphson και της μεθόδου παραγοντοποίησης πολυωνύμων. Η τελευταία, που αναφέρεται και ως σχήμα του Horner, διαιρεί ένα πολυώνυμο βαθμού n με μια ποσότητα $(x - t)$ και παράγει ένα πολυώνυμο βαθμού n–1 και ένα υπόλοιπο (σταθερά):

$$\frac{p_n(x)}{x-t} = g_{n-1}(x) + \frac{r_n}{x-t} \tag{2.31}$$

Η προηγούμενη εξίσωση δίνει

$$p_n(x) = (x-t)\cdot g_{n-1}(x) + r_n \;\Rightarrow\; p_n(t) = r_n \tag{2.32}$$

δηλαδή το υπόλοιπο της διαίρεσης ισούται με την τιμή του πολυωνύμου για $x = t$.

Επίσης, παραγωγίζοντας την προηγούμενη σχέση προκύπτει τελικά

$$p_n'(t) = g_{n-1}(t) \tag{2.33}$$

Δηλαδή η τιμή της παραγώγου του πολυωνύμου $p_n(x)$ στο σημείο $x = t$ ισούται με την τιμή του $g_{n-1}(t)$, η οποία με τη σειρά της θα ισούται με το υπόλοιπο της διαίρεσης του $g_{n-1}(x)$ με την ποσότητα $(x - t)$. Το υπόλοιπο r_n και οι συντελεστές του πολυωνύμου

$$g_{n-1}(x) = b_{n-1}x^{n-1} + b_{n-2}x^{n-2} + ... + b_0 \tag{2.34}$$

υπολογίζονται από τους γνωστούς συντελεστές του $p_n(x)$ (Εξ. 2.30) και με βάση τη σχέση (2.32), με τον ακόλουθο αλγόριθμο:

$$\begin{aligned} b_{n-1} &= a_n \\ b_i &= a_{i+1} + t \cdot b_{i+1}, \qquad \text{i} = n\text{-}2,\ n\text{-}3,\ \ldots.,\ 0. \\ r_n &= a_0 + t \cdot b_0 \end{aligned} \tag{2.35}$$

Είναι αξιοσημείωτο ότι για τον υπολογισμό της τιμής του πολυωνύμου $p_n(t) = r_n$ χρειάζονται με τον αλγόριθμο αυτόν μόνο n πολλαπλασιασμοί, ενώ στην κανονική γραφή της Εξ. (2.30) απαιτούνται $n(n+1)/2$ πολλαπλασιασμοί.

Έτσι, η μέθοδος Newton ξεκινά από μια αρχική εκτίμηση μίας ρίζας του πολυωνύμου, έστω $x_0 = t$ και χρησιμοποιεί επαναληπτικά τον αναδρομικό τύπο (Εξ. 2.20)

$$x_m = x_{m-1} - \frac{p_n(x_{m-1})}{p_n'(x_{m-1})} \qquad m = 1,\ 2,\ ....$$

μόνο που οι τιμές του πολυωνύμου και της παραγώγου του υπολογίζονται εφαρμόζοντας δύο φορές το σχήμα του Horner, μία στο πολυώνυμο $p_n(x)$ και μία στο $g_{n-1}(x)$.

Μετά την εύρεση της πρώτης ρίζας του πολυωνύμου $p_n(x)$, έστω x_{r1}, θα ισχύει

$$p_n(x) = (x - x_{r1}) \cdot g_{n-1}(x) \qquad \left(\ r_n = p_n(x_{r1}) = 0 \ \right)$$

δηλαδή το πολυώνυμο κατώτερου βαθμού $g_{n-1}(x)$ θα έχει τις υπόλοιπες ρίζες πλην αυτής που υπολογίσθηκε. Έτσι, η διαδικασία προσέγγισης ρίζας μπορεί τώρα να επαναληφθεί για το πολυώνυμο αυτό, οπότε δεν υπάρχει περίπτωση να συγκλίνει σε μια ρίζα που ήδη έχει βρεθεί.

Συνεχίζοντας με τον ίδιο τρόπο προσεγγίζονται τελικά όλες οι πραγματικές ρίζες του πολυωνύμου, εφόσον βέβαια δεν προκύψει κάποιο πρόβλημα σύγκλισης, όπως αναφέρθηκε στην περιγραφή της μεθόδου Newton-Raphson (Κεφ. 2.4.2). Τελικά, ενδέχεται να παραμείνει ένα πολυώνυμο κατώτερου βαθμού που θα περιέχει τις μιγαδικές ρίζες. Ο υπολογισμός αυτών των ριζών είναι επίσης εφικτός, με χρήση άλλων μεθόδων.

Ένα πρόβλημα που μπορεί να προκύψει κατά την παραπάνω διαδικασία προέρχεται από τη μετάδοση του σφάλματος στρογγυλοποίησης στις επόμενες πράξεις (βλ. Κεφ. 1.4). Είναι δυνατόν το σφάλμα αυτό να μεγεθύνεται συνεχώς, ώστε από ένα σημείο και μετά τα αποτελέσματα να είναι εντελώς λανθασμένα. Επίσης, επειδή κάθε ρίζα υπολογίζεται με κάποια προσέγγιση, οι συντελεστές του νέου πολυωνύμου κατώτερου βαθμού δεν είναι ακριβείς και η επόμενη ρίζα ενέχει και αυτό το σφάλμα. Υπάρχουν περιπτώσεις πολυωνύμων, κυρίως μεγάλου βαθμού, που μια ελάχιστη μεταβολή της τιμής ενός συντελεστή τους μπορεί να προκαλέσει αλλαγή της τιμής ή ακόμη και του τύπου αρκετών ριζών του. Ένας τρόπος αποφυγής τέτοιων σφαλμάτων είναι κάθε νέα ρίζα που εκτιμάται να προσεγγίζεται με περαιτέρω ακρίβεια, με βάση όμως το αρχικό πολυώνυμο $p_n(x)$, προτού χρησιμοποιηθεί για υποβιβασμό του βαθμού του πολυωνύμου.

Εφαρμογή 2.5.

Να βρεθούν με τη μέθοδο Newton και με ακρίβεια 5 σημαντικών ψηφίων όλες οι ρίζες της πολυωνυμικής εξίσωσης: $p_3(x) = x^3 - 3x - 2$.

Μια αρχική εκτίμηση της ρίζας μπορεί να ληφθεί ως: $x_0 = -\alpha_0 / \alpha_1 = -2/3$. Ο αλγόριθμος (2.35) εφαρμοζόμενος στο πολυώνυμο $p_3(x)$ δίνει:
$b_2 = \alpha_3 = 1$
$b_1 = \alpha_2 + t\, b_2 = t$
$b_0 = \alpha_1 + t\, b_1 = -3 + t^2$ και υπόλοιπο: $p_3(t) = r_3 = a_0 + t\, b_0 = t^3 - 3\, t - 2$
ενώ εφαρμοζόμενος στο πολυώνυμο: $g_2(x) = b_2 x^2 + b_1 x + b_0$ δίνει:
$c_1 = b_2 = 1$
$c_0 = b_1 + t\, c_1 = 2\, t$ και υπόλοιπο: $p_3'(t) = r_2 = b_0 + t\, c_0 = 3\, t^2 - 3$.

Ο αναδρομικός τύπος της Newton-Raphson συγκλίνει μετά από 17 επαναλήψεις στη ρίζα: $x_r = -0.999995504$ (βλ. και επόμενη Εφαρμογή 2.6), επομένως το πολυώνυμο κατώτερου βαθμού που προκύπτει είναι το:

$$g_2(x) = x^2 - 0{,}999995504\, x - 2{,}000009$$

Επανάληψη της ίδιας διαδικασίας στο πολυώνυμο αυτό, με $x_0 = -2$, δίνει σε 5 μόνο επαναλήψεις (αφού πρόκειται τώρα για απλή ρίζα): $x_r = -1.000004497$ και πολυώνυμο κατώτερου βαθμού το: $q_1(x) = x - 2.000008994$. Επομένως οι τρεις ρίζες της πολυωνυμικής εξίσωσης θα είναι οι:

$$x_{r1} = -0.999995504, \quad x_{r2} = -1.000004497 \quad \text{και} \quad x_{r3} = 2.000008994$$

που προσεγγίζουν με την επιθυμητή ακρίβεια τις πραγματικές: $-1, -1$ και $+2$.

Μπορεί όμως να παρατηρηθεί ότι η ακρίβεια της τελευταίας ρίζας είναι λίγο μικρότερη, λόγω μετάδοσης των σφαλμάτων στρογγυλοποίησης. Αν π.χ. είχε χρησιμοποιηθεί η αναλυτική λύση της εξίσωσης $g_2(x) = 0$, θα προέκυπτε η τιμή: $x_{r3} = 1.99999999999$.

Ο αλγόριθμος του Horner (Εξ. 2.35) μπορεί εύκολα να γενικευθεί, ώστε να δίνει το αποτέλεσμα της διαίρεσης ενός πολυωνύμου βαθμού n με ένα άλλο πολυώνυμο βαθμού $m < n$. Έτσι, μία ή περισσότερες ρίζες ενός πολυωνύμου μπορούν να υπολογισθούν με κάποια μέθοδο (όχι απαραίτητα τη Newton-Raphson) και στη συνέχεια να γίνει ο αντίστοιχος υποβιβασμός του βαθμού του αρχικού πολυωνύμου. Εάν π.χ. χρησιμοποιηθεί η μέθοδος Müller, όπου ο διαιρέτης είναι πολυώνυμο $2^{ου}$ βαθμού (Κεφ. 2.4.4), τότε η σύγκλιση της μεθόδου είναι εξασφαλισμένη σχεδόν για κάθε αρχική τιμή.

Η τεχνική υποβιβασμού του βαθμού ενός πολυωνύμου μπορεί επίσης να βρει εφαρμογή για την εύρεση όλων των πραγματικών ριζών και μιας υπερβατικής εξίσωσης. Αφού βρεθεί με κάποια αριθμητική μέθοδο μία ρίζα x_{r1} της εξίσωσης $f(x) = 0$, αναζητείται μία νέα ρίζα x_{r2} στην εξίσωση $f(x)/(x - x_{r1}) = 0$, μετά στην εξίσωση $f(x)/(x - x_{r1})/(x - x_{r2}) = 0$ κ.ο.κ. Προσοχή απαιτεί όμως το γεγονός ότι η νέα εξίσωση δεν είναι συνεχής στα σημεία $x = x_{ri}$.

2.5.2. Άλλες Μέθοδοι για Πολυώνυμα

Πολλές ακόμη τεχνικές έχουν επινοηθεί για τον υπολογισμό όλων των ριζών ενός πολυωνύμου. Η μέθοδος του Bairstow αποτελεί ουσιαστικά επέκταση της μεθόδου Newton, ώστε να υπολογίζονται και οι μιγαδικές ρίζες, χρησιμοποιώντας ως διαιρέτη ένα πολυώνυμο $2^{ου}$ βαθμού: $x^2 - r\,x - s$. Μετά την επαναληπτική εύρεση των τιμών των συντελεστών r και s, η πολυωνυμική αυτή εξίσωση δίνει αναλυτικά δύο ρίζες, είτε πραγματικές είτε συζυγείς μιγαδικές (για αρχικό πολυώνυμο με πραγματικούς συντελεστές).

Η μέθοδος των Graeffe-Lobachevsky υπολογίζει ταυτόχρονα όλες τις ρίζες ενός πολυωνύμου με πραγματικούς συντελεστές, μετατρέποντάς το σε ένα άλλο ίδιου βαθμού αλλά με ρίζες που είναι ίσες με τα τετράγωνα των αντίστοιχων ριζών του αρχικού. Οι νέες ρίζες είναι έτσι πιο απομακρυσμένες μεταξύ τους. Η διαδικασία διασποράς συνεχίζεται αρκετές φορές μέχρι που οι ρίζες να απέχουν τόσο πολύ, ώστε η εκτίμησή τους να γίνεται με ικανοποιητική ακρίβεια, με βάση τα πηλίκα των γειτονικών συντελεστών του τελικού πολυωνύμου. Η μέθοδος δεν απαιτεί αρχική εκτίμηση κάποιας ρίζας, αλλά γίνεται περίπλοκη αν υπάρχουν ρίζες με ίδια απόλυτη τιμή.

Αναφέρονται επίσης και οι μέθοδοι QD (Quotient-Difference), Lehmer και Jenkins-Traub, οι οποίες υπολογίζουν όλες τις ρίζες (πραγματικές και μιγαδικές) ενός πολυωνύμου χωρίς να χρειάζονται κάποια αρχική εκτίμηση, καθώς και η μέθοδος του Laguerre, η οποία απαιτεί αρχική εκτίμηση της ρίζας, αλλά επιτυγχάνει σύγκλιση τρίτου βαθμού (κυβική). Τέλος, υπάρχουν και μέθοδοι εύρεσης των ριζών πολυωνύμων με μιγαδικούς συντελεστές, τα οποία όμως σπάνια συναντώνται σε πρακτικά προβλήματα μηχανικού.

2.6. Ειδικά Θέματα

2.6.1. Εντοπισμός και Προσέγγιση Πολλαπλής Ρίζας

Όπως έχει ήδη αναφερθεί, όταν η ρίζα μιας εξίσωσης είναι διπλή ή πολλαπλή, τότε τόσο ο εντοπισμός της με τη μέθοδο ίσων διαστημάτων, όσο και η προσέγγισή της με τη μέθοδο της διχοτόμησης δεν είναι δυνατή. Στη θέση μιας ρίζας x_r, η οποία είναι βαθμού πολλαπλότητας $k > 1$, η συνάρτηση $f(x)$ εφάπτεται στον άξονα x (βλ. Σχ. 2.1) και ισχύει

$$f^{(p)}(x_r) = 0, \qquad p = 1, 2, ..., k-1 \tag{2.36}$$

Έτσι, ο εντοπισμός μιας τέτοιας ρίζας θα μπορούσε να αναχθεί στον εντοπισμό των ριζών της $f^{(k-1)}(x) = 0$, δηλαδή της παραγώγου k-1 τάξης της συνάρτησης.

Ως παράδειγμα, εξετάζεται η εξίσωση της Εφαρμογής 2.1, $f(x) = x^3 - 3x - 2$, η διπλή ρίζα $x_r = -1$ της οποίας δεν εντοπίστηκε εκεί. Εάν διερευνηθεί με τον ίδιο τρόπο η πρώτη παράγωγος της συνάρτησης $f'(x) = 3x^2 - 3$, θα προκύψουν οι ακόλουθες τιμές:

X	-3	-2.4	-1.8	-1.2	-0.6	0	0.6	1.2	1.8	2.4	3
f ΄(x)	+24	+14.3	+6.72	+1.32	-1.92	-3	-1.92	+1.32	+6.72	+14.3	+24

Επομένως μία διπλή ρίζα είναι πιθανό να υπάρχει στο διάστημα [–1.2, –0.6], αλλά όχι σίγουρο, αφού μπορεί να πρόκειται απλώς για ακρότατο της συνάρτησης $f(x)$, όπως πράγματι συμβαίνει στο διάστημα [+0.6, +1.2] (βλ. Σχ. 2.1).

Η προσέγγιση μιας πολλαπλής ρίζας είναι εφικτή μόνο με μία ανοικτή μέθοδο, ακόμη και τότε όμως η σύγκλιση γίνεται γραμμική. Με βάση την Εξ. (2.36), το ανάπτυγμα Taylor της συνάρτησης σε μια ρίζα x_r πολλαπλότητας k θα είναι

$$f(x) = 0 + (x - x_r)^k \cdot \frac{f^{(k)}(x_r)}{k!} + ... \cong (x - x_r)^k \cdot g(x) \tag{2.37}$$

Υποθέτοντας ότι η συνάρτηση $g(x)$ είναι παραγωγίσιμη έως δύο τάξεις, η (2.37) δίνει

$$f'(x) = k \cdot (x - x_r)^{k-1} \cdot g(x) + (x - x_r)^k \cdot g'(x)$$

οπότε η συνάρτηση της μεθόδου Newton-Raphson γίνεται

$$\varphi(x) = x - \frac{f(x)}{f'(x)} = x - \frac{(x - x_r) \cdot g(x)}{k \cdot g(x) + (x - x_r) \cdot g'(x)} \tag{2.38}$$

Η παράγωγος της συνάρτησης αυτής στη θέση της ρίζας $x = x_r$, παίρνει την απλή μορφή

$$\varphi'(x) = 1 - \frac{1}{k} \tag{2.39}$$

δηλαδή η σύγκλιση θα είναι πρώτου βαθμού, με ρυθμό που μειώνεται όσο αυξάνει η πολλαπλότητα k της ρίζας.

Το γεγονός ότι για $k > 1$ είναι $f'(x_r) = 0$, δηλαδή ο παρονομαστής του αναδρομικού τύπου τείνει στο μηδέν, δεν προκαλεί πρόσθετο πρόβλημα στον αλγόριθμο της μεθόδου N-R, επειδή πάντοτε ο όρος $f(x) / f'(x)$ τείνει στο μηδέν όταν $x \to x_r$.

Το πρόβλημα της μειωμένης ταχύτητας σύγκλισης αντιμετωπίζεται με διάφορες τεχνικές τροποποίησης του βασικού αναδρομικού τύπου, η απλούστερη των οποίων είναι η σχέση των Ralston και Rabinowitz

$$x_m = x_{m-1} - k \cdot \frac{f(x_{m-1})}{f'(x_{m-1})} \tag{2.40}$$

Η μέθοδος αυτή εξασφαλίζει πάντα τετραγωνική σύγκλιση, όμως θα πρέπει να είναι εκ των προτέρων γνωστός ο βαθμός πολλαπλότητας k της ρίζας.

Εφαρμογή 2.6.

Να βρεθεί με τη μέθοδο Newton-Raphson και με ακρίβεια 5 σημαντικών ψηφίων η διπλή ρίζα της εξίσωσης: $f(x) = x^3 - 3x - 2$.

Η ρίζα αυτή εντοπίστηκε προηγουμένως στην περιοχή [−1.2, −0.6]. Έτσι, ως τιμή εκκίνησης λαμβάνεται εδώ η $x_0 = -0.6$. Στη συνέχεια δίνονται τα αποτελέσματα της βασικής μεθόδου N-R (Κώδικας 2.2), όσο και της τροποποιημένης με βάση την Εξ. (2.40), για $k = 2$.

Πίνακας 2.4. Αποτελέσματα εφαρμογής της μεθόδου N-R.

	Newton-Raphson		**N-R τροποποιημένη**	
m	x_m	ε_t	x_m	ε_t
1	-0.8166666667	-1.83E-01	-1.0333333333	3.33E-02
2	-0.9114169215	-8.86E-02	-1.0001821494	1.82E-04
3	-0.9563926793	-4.36E-02	-1.0000000055	5.53E-09
4	-0.9783583384	-2.16E-02		
5	-0.9892186263	-1.08E-02		
6	-0.9946190521	-5.38E-03		
7	-0.9973119455	-2.69E-03		
8	-0.9986565757	-1.34E-03		
9	-0.9993284383	-6.72E-04		
10	-0.9996642568	-3.36E-04		
11	-0.9998321378	-1.68E-04		
12	-0.9999160712	-8.39E-05		
13	-0.9999580362	-4.20E-05		
14	-0.9999790182	-2.10E-05		
15	-0.9999895092	-1.05E-05		
16	-0.9999947546	-5.25E-06		
17	-0.9999973773	-2.62E-06		

Η βασική μέθοδος N-R συγκλίνει στη ρίζα αλλά πολύ αργά, με ταχύτητα ίδια με αυτήν της μεθόδου διχοτόμησης (βλ. Εφαρμ. 2.2), πράγμα αναμενόμενο αφού $\varphi'(x_r) = 1 - 1/2 = 0{,}5$. Η υπεροχή της τροποποιημένης μεθόδου είναι φανερή, καθώς διατηρεί την τετραγωνική σύγκλιση και προσεγγίζει τη ρίζα με ακρίβεια 8 σημαντικών ψηφίων σε μόλις 3 επαναλήψεις.

Σημειώνεται τέλος ότι με τη μέθοδο των διαδοχικών αντικαταστάσεων (Κεφ. 2.4.1) δεν μπορεί να προσεγγισθεί η διπλή ρίζα, επειδή όλες οι συνήθεις αναδιατάξεις της εξίσωσης δίνουν: $\varphi'(x_r) = 1$.

Στη βιβλιογραφία συναντάμε επίσης και άλλες, συνθετότερες εκφράσεις, που δεν έχουν τον περιορισμό της (2.40). Για παράδειγμα, εύκολα αποδεικνύεται ότι η συνάρτηση $g(x) = f(x)/f'(x)$ έχει ρίζες στα ίδια σημεία με την $f(x)$, αλλά μόνο απλές, οι οποίες βρίσκονται εύκολα. Όμως απαιτείται προσοχή όταν η $f(x)$ δεν είναι αλγεβρική συνάρτηση, οπότε η $g(x)$ μπορεί να εμφανίζει ασυνέχειες στα ακρότατα σημεία της συνάρτησης.

Σημειώνεται τέλος ότι μια πολύ δύσκολη περίπτωση υπολογισμού αποτελεί η ύπαρξη διαφορετικών ριζών μιας εξίσωσης που διαφέρουν όμως ελάχιστα μεταξύ τους (σχεδόν πολλαπλή ρίζα). Τότε ο αναδρομικός τύπος (2.40) αποτυγχάνει, ενώ οι άλλες ανοικτές μέθοδοι συγκλίνουν γραμμικά και μάλιστα μόνο όταν το κριτήριο τερματισμού (επιτρεπόμενο σφάλμα) είναι σαφώς μεγαλύτερο από τη σχετική διαφορά των ριζών.

2.6.2. Προσέγγιση Μιγαδικών Ριζών

Η περισσότερες ανοικτές μέθοδοι (Newton-Raphson, Τέμνουσας, Müller κ.ά.), μπορούν να υπολογίσουν άμεσα και τις μιγαδικές ρίζες μιας συνάρτησης, εάν η γλώσσα προγραμματισμού του αλγορίθμου υποστηρίζει αριθμητική μιγαδικών αριθμών (όπως η Fortran). Τότε αρκεί να δοθεί ως αρχική τιμή (ή τιμές) εκκίνησης ένας μιγαδικός αριθμός αρκετά κοντά στη ρίζα (η μέθοδος Müller εκκινεί και από πραγματικό αριθμό).

Εφαρμογή 2.7.

Να βρεθούν με τη μέθοδο Newton-Raphson όλες οι ρίζες της εξίσωσης:

$$f(x) = x^3 + x + 10 .$$

Εντοπισμός: Τα πρόσημα των συντελεστών είναι όλα ίδια, επομένως δεν υπάρχει θετική πραγματική ρίζα (βλ. Κεφ. 2.5). Υπάρχει όμως μία αρνητική ρίζα, επειδή: $f(-x) = -x^3 - x + 10$. Οι άλλες δύο είναι μιγαδικές συζυγείς. Επίσης, από τις Εξ (2.8) προκύπτει ότι όλες οι ρίζες βρίσκονται στον δακτύλιο του μιγαδικού επιπέδου με ακτίνες: $R_o = 10$ και $R_i = 0.5$.

Έτσι, ως αρχικές τιμές λαμβάνονται για μεν την αρνητική ρίζα: $x_0 = -5$, για δε τις μιγαδικές: $x_0 = 3.5 \pm i\ 3.5$ (δηλαδή περίπου στη μέση του δακτυλίου). Ο Κώδικας 2.2 μπορεί εύκολα να τροποποιηθεί ώστε να λειτουργεί για μιγαδικές μεταβλητές (ο τελεστής .LE. δεν δέχεται μιγαδικούς αριθμούς).

Πίνακας 2.5. Αποτελέσματα της μεθόδου N-R για διάφορες αρχικές τιμές.

m	$x_0 = -5$	$x_0 = (3.5,\ 3.5)$
1	-3.421052	(2.299311, 2.500671)
2	-2.494469	(1.503048, 2.003053)
3	-2.086886	(1.068043, 1.928859)
4	-2.003314	(0.996067, 1.997666)
5	-2.000005	(1.000010, 2.000000)
6	-2.000000	(1.000000, 2.000000)

Όπως φαίνεται στον Πίνακα 2.5, η ο κώδικας συγκλίνει τετραγωνικά, τόσο στην περίπτωση της πραγματικής όσο και της μιγαδικής ρίζας της εξίσωσης, δίνοντας τις ακριβείς λύσεις σε 6 επαναλήψεις. Η αρχική τιμή: $x_0 = 3.5 - i\ 3.5$ θα δώσει ομοίως τη συζυγή μιγαδική ρίζα: $x_r = 1.0 - i \cdot 2.0$.

Εάν η γλώσσα προγραμματισμού δεν υποστηρίζει μιγαδικούς αριθμούς, τότε η μιγαδική μεταβλητή x μπορεί να εκφρασθεί ως $x = r + i \cdot c$, όπου r και c πραγματικοί αριθμοί, οπότε η συνάρτηση $f(x)$ γράφεται ισοδύναμα στη μορφή

$$f(x) = f(r + i \cdot c) = u(r,c) + i \cdot v(r,c) \tag{2.41}$$

και η εξίσωση $f(x) = 0$ ισοδυναμεί με το σύστημα μη–γραμμικών εξισώσεων

$$u(r,c) = 0, \qquad v(r,c) = 0 \tag{2.42}$$

το οποίο λύνεται με τις μεθόδους του Κεφαλαίου 3.2 (συνήθως με τη Newton).

Τέλος, σημειώνεται ότι οι μέθοδοι QD, Graeffe-Lobachevsky και Bairstow, που εφαρμόζονται σε πολυώνυμα με πραγματικούς συντελεστές, βρίσκουν και τις μιγαδικές ρίζες χωρίς να χρησιμοποιούν αριθμητική μιγαδικών. Σε πολυώνυμα όμως που έχουν και μιγαδικούς συντελεστές απαιτείται αριθμητική μιγαδικών και μέθοδοι όπως η Newton-Raphson (Κεφ. 2.4).

2.7. Ανακεφαλαίωση

Ο αλγόριθμος υπολογισμού της ρίζας μιας μη–γραμμικής εξίσωσης είναι σχετικά απλός σε όλες σχεδόν τις μεθόδους που παρουσιάστηκαν. Επίσης, όλες οι μέθοδοι μπορούν κατά κανόνα να επιτύχουν την εκάστοτε επιθυμητή προσέγγιση. Επομένως, τα βασικά κριτήρια για την επιλογή μιας συγκεκριμένης μεθόδου είναι αφ' ενός η ασφάλεια και αφ' ετέρου η ταχύτητα εύρεσης της ρίζας. Όπως συμβαίνει συνήθως στις αριθμητικές μεθόδους, τα δύο αυτά χαρακτηριστικά είναι ανταγωνιστικά, δηλαδή για την επίτευξη αυξημένη ασφάλειας απαιτείται περισσότερος υπολογιστικός χρόνος και αντιστρόφως. Στον παρακάτω Πίνακα 2.6 έχουν συγκεντρωθεί τα κυριότερα λειτουργικά στοιχεία των κλειστών και ανοικτών μεθόδων που αναλύθηκαν στα Κεφάλαια 2.3 και 2.4.

Πίνακας 2.6. Σύγκριση χαρακτηριστικών των μεθόδων επίλυσης μη–γραμμικών εξισώσεων.

Χαρακτηριστικά	**Διχοτόμ. Διαστήμ.**	**Εσφαλμ. Θέσης**	**Διαδοχ. Αντικατ.**	**Newton-Raphson**	**Τέμνουσας**	**Müller**
Ταχύτητα (βαθμός p)	1	≥ 1	1	2	1.62	1.84
Σύγκλιση	πάντα	πάντα	υπό συνθήκη	πιθανή απόκλιση	πιθανή απόκλιση	όχι πάντα
Αρχικές τιμές	2 εκατέρωθεν	2 εκατέρωθεν	1 κοντά	1 κοντά	2 κοντά	3 σχετικά κοντά
Πολλαπλή ρίζα	$2k+1$ ναι $2k$ όχι	$2k+1$ ναι $2k$ όχι	όχι	ναι	ναι	ναι
Μιγαδικές ρίζες	όχι	όχι	ναι	ναι	ναι	ναι

Γενικά, εάν η ρίζα μιας εξίσωσης πρόκειται να βρεθεί μία ή λίγες μόνο φορές κατά την επίλυση ενός προβλήματος, τότε είναι προτιμότερη η χρήση μιας κλειστής μεθόδου, με βέβαιη σύγκλιση. Αντίθετα, εάν απαιτούνται πολλαπλές και επαναλαμβανόμενες επιλύσεις, συνήθως ως τμήμα ενός μεγάλου υπολογιστικού κώδικα, τότε ενδιαφέρει ιδιαίτερα η ταχύτητα σύγκλισης, οπότε ενδείκνυνται οι ανοικτές μέθοδοι, σε συνδυασμό με τεχνικές αύξησης της ευστάθειας (υβριδικά σχήματα). Ανάλογα δε με την προς επίλυση εξίσωση, υπάρχει ή μπορεί να αναπτυχθεί η πλέον κατάλληλη μέθοδος.

Σε ειδικές περιπτώσεις που απαιτείται π.χ. η εύρεση πολλαπλής ή μιγαδικής ρίζας, ορισμένες μέθοδοι του Πίνακα 2.6 αποκλείονται, ενώ κάποιες έχουν ειδική συμπεριφορά. Έτσι, για τις μεθόδους Newton-Raphson και Τέμνουσας υπάρχουν σχετικά απλοί τρόποι βελτίωσης της ταχύτητας σύγκλισης σε πολλαπλή ρίζα. Επίσης, η μέθοδος Müller μπορεί να υπολογίσει μια μιγαδική ρίζα ακόμη και από πραγματικές αρχικές τιμές. Βέβαια υπάρχουν κι άλλοι περιορισμοί, όπως το ότι οι κλειστές μέθοδοι χρειάζονται δύο αρχικές τιμές εκατέρωθεν της ρίζας, και ότι η Newton-Raphson θέλει την αναλυτική έκφραση της παραγώγου της συνάρτησης.

Σημειώνεται ότι ο ακριβής ρυθμός σύγκλισης κοντά στη ρίζα δεν εξαρτάται μόνο από τον βαθμό σύγκλισης p της μεθόδου αλλά και από τη μορφή της εξίσωσης τοπικά, η οποία καθορίζει τον συντελεστή της εκθετικής σχέσης μείωσης του σφάλματος. Επίσης, ο συνολικός υπολογιστικός χρόνος που απαιτείται για σύγκλιση σε λύση δεδομένης ακρίβειας εξαρτάται και από τις αρχικές τιμές, καθώς και από τις αριθμητικές πράξεις που εκτελούνται ανά επανάληψη. Έχει βρεθεί για παράδειγμα ότι η μέθοδος της Τέμνουσας γίνεται πιο συμφέρουσα υπολογιστικά από τη μέθοδο Newton-Raphson, αν ο χρόνος

υπολογισμού μιας τιμής της παραγώγου είναι περισσότερος από το 43% του αντίστοιχου χρόνου για την τιμή της ίδιας της συνάρτησης.

Ανάλογες παρατηρήσεις μπορούν να γίνουν και για την επιλογή της κατάλληλης μεθόδου εύρεσης των ριζών μιας πολυωνυμικής εξίσωσης με πραγματικούς συντελεστές. Η μέθοδος Newton που αναλύθηκε στο Κεφ. 2.5.1 υπολογίζει μόνο τις πραγματικές ρίζες, αλλά οι περισσότερες μπορούν να βρουν και τις μιγαδικές ρίζες ενός πολυωνύμου. Σε μερικές μάλιστα, όπως οι Bairstow, Graeffe-Lobachevsky και QD, αυτό γίνεται χωρίς τη χρήση αριθμητικής μιγαδικών αριθμών, επομένως είναι υλοποιήσιμες και σε απλό υπολογιστή τσέπης. Επίσης, οι δύο τελευταίες δεν χρειάζονται αρχική εκτίμηση για εκκίνηση.

Οι πιο εξελιγμένες μέθοδοι, που χρησιμοποιούνται και σε εμπορικά υπολογιστικά πακέτα (όπως οι Lehmer, Jenkins-Traub, Laguerre), συνδυάζουν ασφάλεια και ταχύτητα εύρεσης όλων των ριζών, έχουν όμως αρκετά πολύπλοκο αλγόριθμο. Στη βιβλιογραφία ή στις μαθηματικές βιβλιοθήκες των υπολογιστικών συστημάτων υπάρχουν διαθέσιμα υποπρογράμματα, τα οποία όμως ο χρήστης πρέπει να κατανοήσει πλήρως πριν προχωρήσει στην εφαρμογή τους, ώστε να είναι ενήμερος των δυνατοτήτων και των περιορισμών της μεθόδου που επιλέγει για το πρόβλημά του.

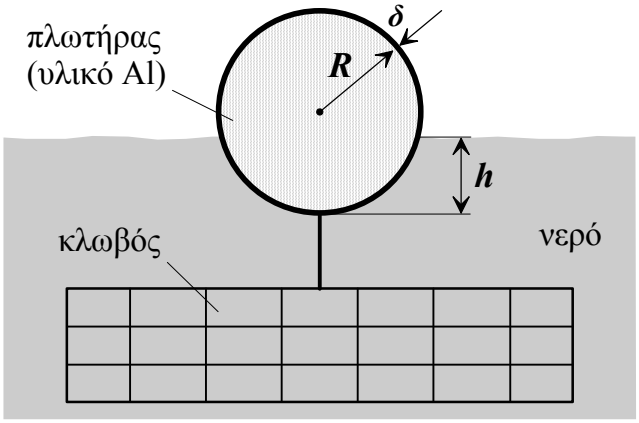
Κλείνοντας, υπογραμμίζεται ότι καμμία από τις μεθόδους επίλυσης μη–γραμμικών εξισώσεων δεν μπορεί να συγκριθεί, ούτε ως προς την ακρίβεια ούτε ως προς την ταχύτητα υπολογισμού, με την αναλυτική λύση μιας εξίσωσης. Επομένως, η επιλογή και χρήση μιας τέτοιας μεθόδου πρέπει να γίνεται μόνο αφού διερευνηθεί και αποκλεισθεί η ύπαρξη αναλυτικής λύσης.

2.8. Πρακτικά παραδείγματα και εφαρμογές

Ακολουθούν ενδεικτικά μερικά προβλήματα μηχανολογικού ενδιαφέροντος και δίνονται υποδείξεις για τη μεθοδολογία επίλυσής τους, με χρήση των αριθμητικών μεθόδων που προηγήθηκαν.

Πρόβλημα 1ο

Α. Περιγραφή:
Από έναν σφαιρικό πλωτήρα ακτίνας R και μάζας m_π, που είναι κενός εσωτερικά και έχει τοίχωμα πάχους δ και πυκνότητας ρ_π, αναρτάται κλωβός εκτροφής ψαριών μάζας m_κ. Να υπολογισθεί το βύθισμα h του πλωτήρα μέσα στο νερό (αγνοείστε την άνωση του κλωβού).



Β. Μαθηματική έκφραση του προβλήματος:
Όπως φαίνεται στο σχήμα, ο όγκος του νερού που εκτοπίζεται για βύθισμα h του πλωτήρα είναι (σφαιρικό τμήμα)

$$V = \frac{\pi h^2}{3}(3R - h)$$

Η ισορροπία των δυνάμεων κατά την κατακόρυφη διεύθυνση δίνει τη σχέση

$$\rho_{H_2O} \cdot g \cdot V = (m_\kappa + m_\pi) \cdot g \quad \text{ή} \quad \rho_{H_2O} \frac{\pi h^2}{3}(3R - h) = (m_\kappa + m_\pi) \tag{2.8.1}$$

Η σχέση (2.8.1) αποτελεί μια μη–γραμμική εξίσωση ως προς h. Για δοσμένα μεγέθη m_π, m_κ και R ζητούνται τα αντίστοιχα βυθίσματα h του πλωτήρα.

Γ. Αδιαστατοποίηση του προβλήματος:
Το βύθισμα h του πλωτήρα μπορεί να αδιαστατοποιηθεί με την ακτίνα R του πλωτήρα. Επίσης η μάζα m_κ του κλωβού να αδιαστατοποιηθεί με τη μάζα m_π του πλωτήρα. Έτσι η εξίσωση (2.8.1) γράφεται

$$\rho_{H_2O} \frac{\pi}{3}(h/R)^2\left(3 - \frac{h}{R}\right) = \frac{m_\pi}{R^3}\left(1 + \frac{m_\kappa}{m_\pi}\right) \tag{2.8.2}$$

Η εξίσωση (2.8.2) μπορεί γενικευθεί περαιτέρω, εισάγοντας το υλικό του πλωτήρα, που θα εκφράζεται με την πυκνότητα ρ_π του υλικού. Αν δ είναι το πάχος του τοιχώματος του πλωτήρα, τότε $m_\pi = 4\pi R^2 \delta \rho_\pi$, οπότε η εξίσωση (2.8.2) γράφεται

$$\left(\frac{h}{R}\right)^2\left(3-\frac{h}{R}\right)=12\frac{\delta}{R}\frac{\rho_\pi}{\rho_{H_2O}}\left(1+\frac{m_\kappa}{m_\pi}\right)$$

Δ. Αριθμητική επίλυση του προβλήματος:
Μετά την προηγούμενη ανάλυση προκύπτει η ανάγκη εύρεσης της ρίζας (ή ριζών) της εξίσωσης

$$f\left(h/R\right)=\left(\frac{h}{R}\right)^2\left(3-\frac{h}{R}\right)-C\cdot\left(1+\frac{m_\kappa}{m_\pi}\right)=0 \tag{2.8.3}$$

όπου η παράμετρος C ισούται με

$$C=12\frac{\delta}{R}\frac{\rho_\pi}{\rho_{H_2O}}$$

Για την εύρεση της ρίζας μπορεί να εφαρμοσθεί η μέθοδος Newton–Raphson. Η παράγωγος της συνάρτησης $f\left(\tilde{h}\right)$, όπου $\tilde{h}=h/R$, γράφεται

$$f'\left(\tilde{h}\right)=2\tilde{h}\left(3-\tilde{h}\right)-\tilde{h}^2=6\tilde{h}-3\tilde{h}^2 \tag{2.8.4}$$

η δε συνάρτηση είναι

$$f\left(\tilde{h}\right)=\tilde{h}^2\left(3-\tilde{h}\right)-C\cdot\left(1+\frac{m_\kappa}{m_\pi}\right) \tag{2.8.5}$$

Αριθμητικό παράδειγμα:
$R=0{,}5\,m,\ \ \delta=1mm,\ \ m_\kappa/m_\pi=10,\ \ \rho_{H_2O}=1000\,kg/m^3,\ \ \rho_\pi=2700\,kg/m^3$ (αλουμίνιο).

Πρόβλημα 2ο

Α. Περιγραφή:
Η εξίσωση κατάστασης Van der Waals, για τα πραγματικά αέρια δίνεται από τη σχέση

$$\left(P+\frac{\alpha}{V^2}\right)\left(V-b\right)=RT \tag{2.8.6}$$

όπου P η πίεση, V ο όγκος, T η απόλυτη θερμοκρασία, a και b σταθερές του αερίου και R η σταθερά των αερίων. Ζητείται να υπολογισθεί ο ειδικός όγκος ενός αερίου για δεδομένες τιμές πίεσης – θερμοκρασίας του.

Β. Μαθηματική έκφραση του προβλήματος:
Η εξίσωση (2.8.6) είναι μη–γραμμική ως προς τον όγκο του αερίου, μπορεί δε να γραφεί ως

$$V=b+\frac{RT}{P+\frac{\alpha}{V^2}} \tag{2.8.7}$$

ή

$$V=\sqrt{\frac{\alpha}{\frac{RT}{V-b}-P}} \tag{2.8.8}$$

(προσοχή στην πιθανότητα εμφάνισης αρνητικού προσήμου στην υπόρριζο ποσότητα).

Οι εξισώσεις (2.8.7) ή (2.8.8) είναι γραμμένες στη μορφή $V = f(V)$, συνεπώς για την εύρεση της ρίζας (ή των ριζών) μπορεί να δοκιμασθεί η μέθοδος των διαδοχικών αντικαταστάσεων. Επίσης η εξίσωση (2.8.6) μπορεί να γραφεί σε πολυωνυμική μορφή

$$PV^3 - (Pb + RT)V^2 + aV - ab = 0$$

Γ. Αδιαστατοποίηση του προβλήματος:
Η εξίσωση Van der Waals μπορεί να γραφεί ως εξής:

$$\frac{PV}{RT}\left(1+\frac{\alpha}{PV^2}\right)\left(1-\frac{b}{V}\right)=1$$

Αν το αέριο ήταν τέλειο θα ίσχυε $\frac{PV}{RT}=1$, άρα η παράμετρος $Z=\left(1+\frac{\alpha}{PV^2}\right)\left(1-\frac{b}{V}\right)$, που καλείται και συντελεστής συμπιεστότητας του αερίου, εκφράζει το βαθμό απόκλισης του πραγματικού αερίου από το τέλειο αέριο. Έτσι η εξίσωση Van der Waals γράφεται

$$PVZ = RT \tag{2.8.9}$$

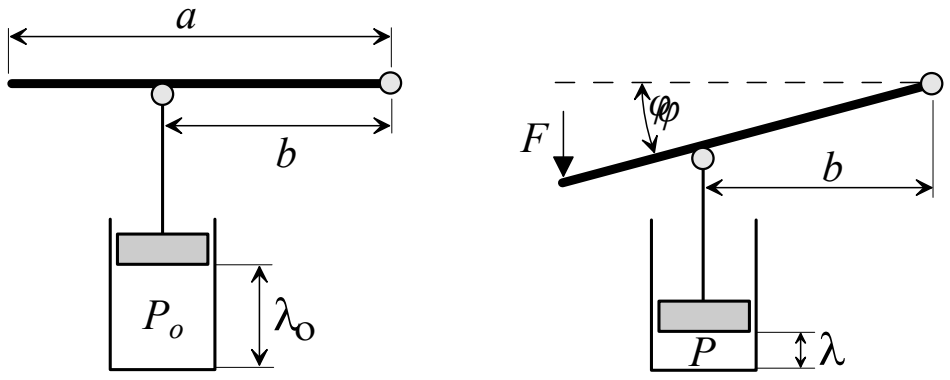
Δ. Αριθμητική επίλυση του προβλήματος:
Για την επίλυση των εξισώσεων (2.8.7) ή (2.8.8) απαιτείται μια αρχική εκτίμηση της ρίζας. Είναι λογικό ότι μια επιτυχής αρχική εκτίμηση θα είναι εκείνη που ικανοποιεί την εξίσωση του τέλειου αερίου

$$V_0 = \frac{RT}{P}$$

Αριθμητικό παράδειγμα: Για τον ατμό οι σταθερές R, a και b ισούνται με R = 461.495 J/kg/K, a = 1703.28 Pa·m^6/kg^2, b = 0.00169099 m^3/kg. Έστω ότι ζητείται ο ειδικός όγκος σε πιέσεις P = 40 bar, 20 bar και 10 bar, και θερμοκρασία T = 573 K.

Πρόβλημα 3^{ο}

Α. Περιγραφή:
Μία πόρτα μάζας m_π ανοίγει με εφαρμογή μιας εξωτερικής δύναμης F στο άκρο της, επαναφέρεται δε με πνευματικό μηχανισμό, όπως στο σχήμα. Να ευρεθεί η γωνία ανοίγματος φ στην οποία ισορροπεί η πόρτα, ως συνάρτηση της δύναμης F. Το έμβολο θεωρείται ότι παραμένει κατακόρυφο και το βάρος του αμελείται.



Β. Μαθηματική έκφραση του προβλήματος:
Με αναφορά στο σχήμα, η ισορροπία ροπών ως προς τον κόμβο στροφής της πόρτας δίνει στις δύο θέσεις της πόρτας ($F = 0$, $\varphi = 0$ και $F - \varphi$) τις ακόλουθες σχέσειςm_π:

$$\frac{\alpha}{2} m_\pi \cdot g = P_0 A\, b$$

$$F \cdot \alpha \cdot \cos\varphi \; + \frac{\alpha}{2} m_\pi \cdot g \cdot \cos\varphi = P\, A b$$

όπου A η διατομή του εμβόλου, και P_0 και P η πίεση του αερίου στο έμβολο στις δύο θέσεις ($\varphi = 0$ και φ) της πόρτας. Η συμπίεση του αερίου στον κύλινδρο γίνεται αδιαβατικά, οπότε η σχέση των πιέσεων συνδέεται με τους αντίστοιχους όγκους του κυλίνδρου,

$$\frac{P}{P_0} = \left(\frac{\lambda_0}{\lambda} \right)^\gamma$$

όπου γ ο εκθέτης της αδιαβατικής μεταβολής ($\gamma = 1.4$ για διατομικό αέριο).
Το μήκος λ συνδέεται με το άνοιγμα της πόρτας με τη σχέση

$$\lambda = \lambda_0 - b \cdot \tan\varphi$$

οπότε προκύπτει η σχέση

$$F \cos\varphi = \frac{m_\pi g}{2} \left[\left(1 - \frac{b}{\lambda_0} \tan\varphi \right)^{-\gamma} - \cos\varphi \right] \tag{2.8.10}$$

η οποία συνδέει την εξωτερική δύναμη F με το άνοιγμα της πόρτας φ.

Γ. Αδιαστατοποίηση του προβλήματος:
Η δύναμη F αδιαστατοποιείται με το βάρος της πόρτας $m_\pi g$, ενώ το μέγεθος b/λ_0 είναι γεωμετρική παράμετρος του προβλήματος:

$$\left(\frac{F}{m_\pi g} + \frac{1}{2} \right) \cos\varphi = \frac{1}{2} \left(1 - \frac{b}{\lambda_0} \tan\varphi \right)^{-\gamma} \tag{2.8.11}$$

Δ. Αριθμητική επίλυση του προβλήματος:
Η εξίσωση (2.8.11) είναι υπερβατικού τύπου και μπορεί να γραφεί ως

$$\varphi = a\cos \left[\left(1 - \frac{b}{\lambda_0} \tan\varphi \right)^{-\gamma} \Big/ \left(\frac{2F}{m_\pi g} + 1 \right) \right] \tag{2.8.12}$$

ή

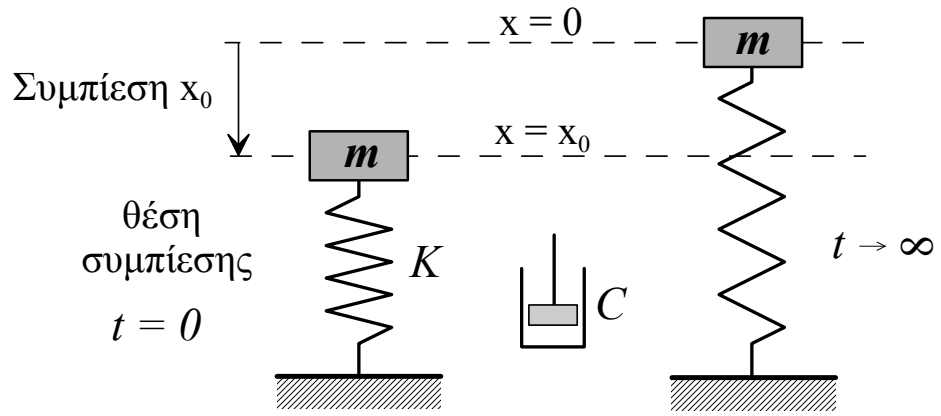
$$\varphi = a\tan \left[\frac{\lambda_0}{b} \left\{ 1 - \left(\frac{2F}{m_\pi g} + 1 \right)^{-1/\gamma} (\cos\varphi)^{-1/\gamma} \right\} \right] \tag{2.8.13}$$

μια πρώτη εκτίμηση της ρίζας στις εξισώσεις μπορεί να είναι η τιμή $\varphi = 0$.

Αριθμητικό παράδειγμα: $F = 500$ N, $m_\pi = 85$ kg, $b/\lambda_0 = 0{,}5$, $g = 9.8$ m/s^2, $\gamma = 1.4$.

Πρόβλημα 4ο

Α. Περιγραφή:
Η ανάρτηση τροχού αυτοκινήτου χαρακτηρίζεται από τη σταθερά k του γραμμικού ελατηρίου και από τον συντελεστή απόσβεσης C. Να μελετηθεί η ταλάντωση του συστήματος για μοναδιαία αρχική συμπίεση του ελατηρίου, χωρίς την επίδραση εξωτερικών δυνάμεων, ώστε να υπολογισθούν οι χρονικές στιγμές όπου το σύστημα έχει απόκλιση ίση με το μισό της αρχικής συμπίεσης x_0.



Β. Μαθηματική έκφραση του προβλήματος:
Με αναφορά στο σχήμα προκύπτει ότι η μετατόπιση της ανάρτησης με το χρόνο δίνεται από τη σχέση

$$\frac{x}{x_0} = 1 - e^{-\zeta \cdot \omega_n \cdot t}\left[\cos(\omega_d t) + \frac{\zeta}{\sqrt{1-\zeta^2}}\sin(\omega_d t)\right] \tag{2.8.14}$$

όπου η ιδιοσυχνότητα ταλάντωσης χωρίς απόσβεση, ω_n, είναι

$$\omega_n = \sqrt{\frac{k}{m}}$$

$$\omega_d = \omega_n\sqrt{1-\zeta^2}, \qquad \zeta = \frac{C}{2\sqrt{km}}$$

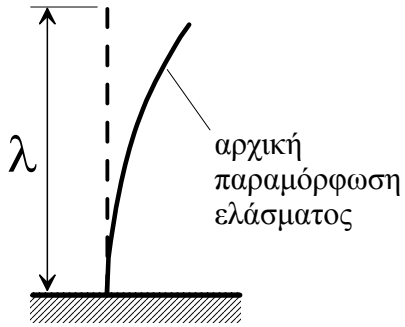
όπου m η μάζα που ταλαντώνεται και ζ η παράμετρος απόσβεσης.

Δ. Αριθμητική επίλυση του προβλήματος:
Η υπερβατική εξίσωση (2.8.14) μπορεί να επιλυθεί με τη μέθοδο της διχοτόμησης του διαστήματος, αφού προηγουμένως με γραφική παράσταση ή με ισοδιαμέριση του πεδίου ορισμού του χρόνου t (π.χ. t = 0 έως 60 sec) βρεθούν οι περιοχές όπου εγκλωβίζονται οι ρίζες της.

Πρόβλημα 5ο

Α. Περιγραφή:
Στις μηχανολογικές κατασκευές είναι συνήθης η περίπτωση εμφάνισης ταλάντωσης. Ιδιαίτερη σημασία για την κατασκευή έχει η γνώση της ιδιοσυχνότητας της κατασκευής, δηλαδή της συχνότητας ταλάντωσης χωρίς την επίδραση εξωτερικής διέγερσης. Η γνώση αυτή επιτρέπει τη λήψη μέτρων αποφυγής φαινομένων συντονισμού. Ζητείται να

υπολογισθεί η ιδιοσυχνότητα ταλάντωσης μιας πακτωμένης κατακόρυφης ράβδου με αρχική απόκλιση του ελεύθερου άκρου της όπως στο σχήμα.



Β. Μαθηματική έκφραση του προβλήματος:
Η μαθηματική τοποθέτηση του προβλήματος οδηγεί στην εύρεση των ιδιοτιμών του πίνακα ακαμψίας του συστήματος, που οδηγεί στην εύρεση των ριζών του πολυωνύμου

$$\alpha_n \lambda^n + \alpha_{n-1} \lambda^{n-1} + \alpha_{n-2} \lambda^{n-2} + ... + \alpha_0 = 0 \qquad (2.8.15)$$

όπου λ η ιδιοπερίοδος της ταλαντούμενης ράβδου-ελάσματος (κυκλική περίοδος). Να σημειωθεί ότι το έλασμα έχει πολλές ιδιοπεριόδους ταλάντωσης. Ποια από όλες θα εμφανισθεί εξαρτάται από τις αρχικές συνθήκες του προβλήματος και ειδικότερα από την αρχική παραμόρφωση του ελάσματος. Για παραμόρφωση της ράβδου όπως φαίνεται στο σχήμα, η ταλάντωση θα γίνει στην πρώτη ιδιοσυχνότητα.

Δ. Αριθμητική επίλυση:
Η πολυωνυμική εξίσωση (2.8.15) μπορεί να επιλυθεί με τη μέθοδο Newton για πολυώνυμα, που παρουσιάσθηκε στο Κεφ. 2.5.1.

Κεφάλαιο 3

Επίλυση Συστημάτων

3.1 Γενικά

Συστήματα εξισώσεων που απαιτούν λύση εμφανίζονται συχνά σε Μηχανολογικές εφαρμογές. Η προσαρμογή καμπυλών σε ομάδα δεδομένων (π.χ. μετρήσεων), η μεθοδολογία πεπερασμένων όγκων για την επίλυση μερικών διαφορικών εξισώσεων, ιδιαίτερα στην περιοχή της Υπολογιστικής Ρευστομηχανικής, η μεθοδολογία πεπερασμένων στοιχείων ή οριακών στοιχείων για επίλυση προβλημάτων Δυναμικής οδηγούν σε συστήματα εξισώσεων που απαιτούν γρήγορη και ακριβή λύση.

Τα συστήματα των εξισώσεων στις περισσότερες των περιπτώσεων είναι γραμμικά ή τα μετατρέπουμε σε γραμμικά, με γραμμικοποίηση του μαθηματικού προβλήματος, ενώ σε λιγότερες περιπτώσεις εμφανίζονται προς επίλυση συστήματα μη γραμμικών εξισώσεων.

Στα γραμμικά συστήματα οι εξισώσεις είναι αλγεβρικές πρώτου βαθμού (οι άγνωστοι εμφανίζονται στην πρώτη δύναμη), ενώ στα μη γραμμικά συστήματα, οι άγνωστοι εμφανίζονται σε δυνάμεις ή και ως ορίσματα υπερβατικών συναρτήσεων, π.χ. ημιτόνου, εκθετικών όρων κλπ.

Τα γραμμικά συστήματα, με τη βοήθεια της γραμμικής άλγεβρας και της θεωρίας των πινάκων, μπορούν εύκολα να λυθούν. Η έμφαση στην αλγοριθμική διαδικασία λύσης εντοπίζεται στην ακρίβεια επίτευξης της λύσης, στην ταχύτητα επίτευξης της λύσης καθώς και στη μνήμη υπολογιστή που απαιτείται.

Αντίθετα για την επίλυση μη γραμμικών συστημάτων, δεν υπάρχει καθιερωμένη αλγοριθμική μεθοδολογία που να εξασφαλίζει τη λύση του συστήματος και η επίτευξη τελικής λύσης είναι περισσότερο θέμα αριθμητικής εμπειρίας του χρήστη και λιγότερο εφαρμογή μιας αυστηρά ορισμένης επιστημονικής διαδικασίας.

3.2 Επίλυση μη γραμμικών συστημάτων – επαναληπτική μέθοδος

Αναφέρθηκε προηγουμένως στο 2ο κεφάλαιο η δυσκολία εύρεσης της λύσης μη γραμμικών εξισώσεων. Η δυσκολία έγκειται στην άγνοιά μας αν υπάρχουν λύσεις και στην άγνοιά μας αν μία συγκεκριμένη μεθοδολογία εξασφαλίζει τη λύση. Στη συνέχεια θα αναφερθούν τρεις μεθοδολογίες που μπορούν να εφαρμοσθούν με πιθανή επιτυχία στην επίλυση μη γραμμικών συστημάτων.

3.2.1 Μέθοδος των διαδοχικών αντικαταστάσεων (Gauss-Seidel)

Η μέθοδος των διαδοχικών αντικαταστάσεων ή Gauss-Seidel αποτελεί επέκταση της αντίστοιχης μεθόδου για την επίλυση εξίσωσης με ένα άγνωστο και βασίζεται στην εύρεση μιας δεύτερης προσέγγισης στη λύση του συστήματος έχοντας γνωστή ή υποθέτοντας μια πρώτη προσέγγιση. Το σύστημα των εξισώσεων μετασχηματίζεται στη μορφή

$$\begin{aligned} x &= X(x, y, z) \\ y &= Y(x, y, z) \\ z &= Z(x, y, z) \end{aligned}$$

και εφαρμόζεται ο αναγωγικός τύπος

$$\begin{aligned} x^{(n+1)} &= X\left[x^{(n)}, y^{(n)}, z^{(n)}\right] \\ y^{(n+1)} &= Y\left[x^{(n+1)}, y^{(n)}, z^{(n)}\right] \\ z^{(n+1)} &= Z\left[x^{(n+1)}, y^{(n+1)}, z^{(n)}\right] \end{aligned} \qquad (3.2.1.1)$$

Η παραπάνω μέθοδος μπορεί να μην οδηγεί στη λύση του συστήματος αλλά να αποκλίνει. Στις περισσότερες όμως περιπτώσεις η μέθοδος συγκλίνει και είναι ευσταθής προς τη λύση της. Για περίπτωση δύο εξισώσεων με δύο αγνώστους

$$\begin{aligned} x &= f(x, y) \\ y &= g(x, y) \end{aligned}$$

οι προσεγγίσεις ριζών της (n) δοκιμής είναι

$$\begin{aligned} x^{(n+1)} &= f\left(x^{(n)}, y^{(n)}\right) \\ y^{(n+1)} &= g\left(x^{(n+1)}, y^{(n)}\right) \end{aligned}$$

όπου στη δεύτερη σχέση χρησιμοποιήθηκε ως τιμή του x η πιο πρόσφατη άρα και η ακριβέστερη τιμή του x που είναι η $x^{(n+1)}$.

Αποδεικνύεται ότι ικανή συνθήκη σύγκλισης της μεθόδου αποτελούν οι σχέσεις

$$\left|\frac{\partial f}{\partial x}\right| + \left|\frac{\partial f}{\partial y}\right| < 1$$
$$\left|\frac{\partial g}{\partial x}\right| + \left|\frac{\partial g}{\partial y}\right| < 1 \qquad (3.2.1.2)$$

Αξίζει να παρατηρηθεί ότι στις εξισώσεις (3.2.1.1) χρησιμοποιούνται οι πλέον πρόσφατες τιμές των αγνώστων. Στην πρώτη εξίσωση χρησιμοποιούνται στο δεύτερο μέρος της εξίσωσης οι τιμές $x^{(n)}, y^{(n)}, z^{(n)}$. Στη δεύτερη όμως εξίσωση χρησιμοποιούνται οι τιμές $x^{(n+1)}, y^{(n)}, z^{(n)}$, ενώ στην τρίτη εξίσωση οι πλέον πρόσφατες τιμές $x^{(n+1)}, y^{(n+1)}, z^{(n)}$. Αλγοριθμικά φυσικά η εξίσωση (3.2.1.1) θα μπορούσε να γραφεί και ως

$$x^{(n+1)} = X\left(x^{(n)}, y^{(n)}, z^{(n)}\right)$$
$$y^{(n+1)} = Y\left(x^{(n)}, y^{(n)}, z^{(n)}\right) \qquad (3.2.1.3)$$
$$z^{(n+1)} = Z\left(x^{(n)}, y^{(n)}, z^{(n)}\right)$$

δηλαδή στο δεύτερο μέρος των εξισώσεων να χρησιμοποιείται πάντα η ίδια τριάδα αριθμών ($x^{(n)}, y^{(n)}, z^{(n)}$). Στην περίπτωση αυτή η επαναληπτική διαδικασία καλείται μέθοδος Jacobi. Η μέθοδος Jacobi έχει μικρότερο ρυθμό σύγκλισης από τη μέθοδο Gauss-Seidel, και απαιτεί περισσότερη υπολογιστική μνήμη .

Η μέθοδος των διαδοχικών αντικαταστάσεων ενδείκνυται για εφαρμογή στην επίλυση μεγάλου πλήθους γραμμικών εξισώσεων , λόγω της αλγεβρικής απλότητάς της, αλλά και της υπολογιστικής της υπεροχής.

3.2.2 Μέθοδος Newton-Raphson

Η μέθοδος Newton-Raphson που παρουσιάστηκε για την εύρεση της ρίζας x_0 της εξίσωσης f(x) μπορεί να επεκταθεί και για την εύρεση των ριζών x_0, y_0 του συστήματος των εξισώσεων

$$f\left(x_0 y_0\right) = 0$$
$$g\left(x_0, y_0\right) = 0$$

Οι εξισώσεις αυτές, με την υπόθεση ότι η εκτίμηση της ρίζας (x_i, y_i) του συστήματος είναι κοντά στη πραγματική τιμή της ρίζας, προσεγγίζονται κατά Taylor αγνοώντας τους όρους δεύτερης τάξης και άνω.

$$\begin{aligned} f(x_0 y_0) &= f(x_i, y_i) + (x_i - x_0) f_x + (y_i - y_0) f_y \\ g(x_0, y_0) &= g(x_i, y_i) + (x_i - x_0) g_x + (y_i - y_0) g_y \end{aligned} \qquad (3.2.2.1)$$

Επίλυση του συστήματος (3.2.21) οδηγεί εύκολα στον αναγωγικό τύπο

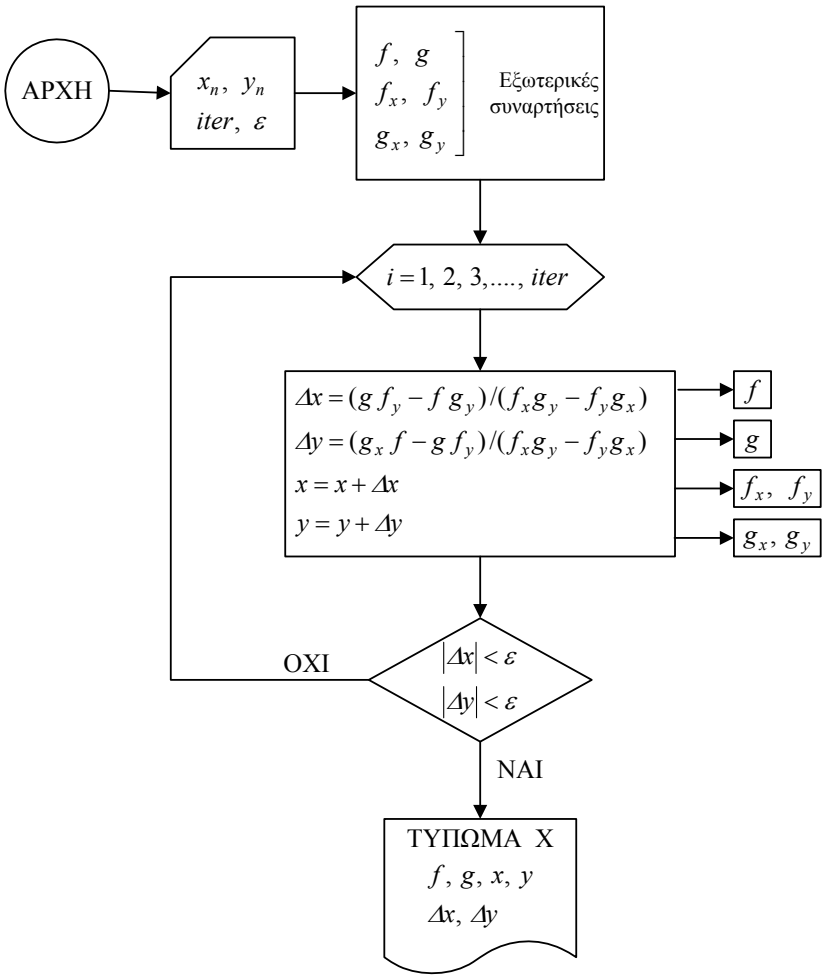
$$\begin{aligned} x^{(n+1)} &= x^{(n)} - \frac{f g_y - g f_y}{J(f,g)} \\ y^{(n+1)} &= y^{(n)} - \frac{g f_x - f g_x}{J(f,g)} \end{aligned} \qquad (3.2.2.2)$$

όπου $(x^{(n)}, y^{(n)})$ η νιοστή προσέγγιση στη ρίζα του συστήματος και f_x, f_y, g_x, g_y είναι οι μερικές παράγωγοι των συναρτήσεων f και g ως προς x και y στη θέση $(x^{(n)}, y^{(n)})$. $J(f,g)$ είναι η Ιακωβιανή των συναρτήσεων f, g υπολογιζόμενη στη θέση $(x^{(n)}, y^{(n)})$,

$$J(f,g) = f_x g_y - g_x f_y$$

Η προηγούμενη αναγωγική σχέση οδηγεί ταχύτατα στη λύση του συστήματος αν οι συναρτήσεις f και g και οι παράγωγοί των είναι συνεχείς συναρτήσεις και η Ιακωβιανή διάφορος του μηδενός. Στο σχήμα 3.2.2.1 δίνεται σε απλό λογικό διάγραμμα η εφαρμογή των προηγούμενων σχέσεων της μεθόδου Newton-Raphson.

Η μέθοδος N-R αποδεικνύεται ότι με επιτυχή πρώτη εκτίμηση των ριζών του συστήματος έχει τετραγωνική σύγκλιση και οδηγεί με ασφάλεια στην εύρεση της ρίζας του συστήματος. Είναι όμως χρονοβόρος γιατί απαιτείται επί πλέον ο υπολογισμός της Ιακωβιανής J για κάθε επανάληψη.



Σχήμα 3.2.2.1: Δομικό διάγραμμα για τη λύση συστήματος δύο εξισώσεων

Τέλος είναι φανερό ότι η μέθοδος N-R μπορεί να επεκταθεί σε σύστημα με N εξισώσεις. Στην περίπτωση αυτή απαιτείται να χρησιμοποιηθεί μια γρήγορη μέθοδος επίλυσης του αντίστοιχου συστήματος (3.2.2.1) όπως απαλοιφής Gauss.

Εφαρμογή: Να βρεθούν τα σημεία τομής του κύκλου

$$x^2 + y^2 = 25$$

και της έλλειψης

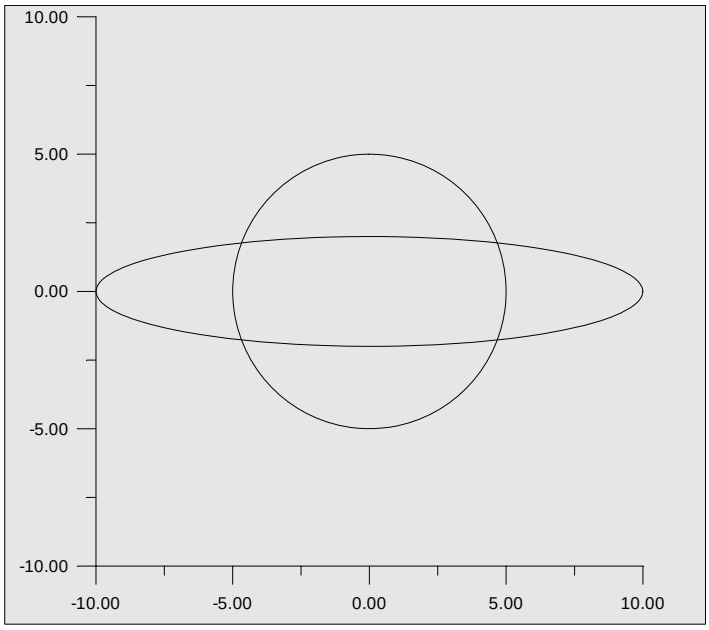
$$\left(\frac{x}{10}\right)^2 + \left(\frac{y}{2}\right)^2 = 1 \qquad \text{Λύση: } x = \pm 4.677, \quad y = \pm 1.678$$

Οι συναρτήσεις $f(x,y)$ και $g(x,y)$ είναι

$$f(x,y) = x^2 + y^2 - 25$$
$$g(x,y) = \left(\frac{x}{10}\right)^2 + \left(\frac{y}{2}\right)^2 - 1$$

Οι μερικές παράγωγοι είναι

$$f_x = 2x, \quad f_y = 2y$$
$$g_x = \frac{x}{50}, \quad g_y = \frac{y}{2}$$



Σχήμα 3.2.2.2 : Τομή κύκλου και έλλειψης

Η Ιακωβιανή του συστήματος είναι

$$J = f_x\, g_y - f_y\, g_x = xy - \frac{xy}{25} = -\frac{24}{25}xy$$

Ο αναγωγικός τύπος είναι τότε

$$x^{(n+1)} = x^{(n)} - \frac{fg_y - gf_y}{J}$$
$$y^{(n+1)} = y^{(n)} - \frac{gf_x - fg_x}{J}$$

Ξεκινώντας με αρχική εκτίμηση $\left(x^{(1)}, y^{(1)}\right) = (5\ ,\ 1)$, έχουμε την ακολουθία δυάδων.

(5 , 1) (4.688 , 2.063) (4.677 , 1.783) (4.677 , 1.768) (4.677 , 1.768)

δηλαδή η μια ρίζα του συστήματος (σημείο τομής κύκλου και έλλειψης) βρέθηκε μέσα σε τέσσερις επαναλήψεις.

3.2.3 Μέθοδος του Στόχου

Στις επαναληπτικές μεθοδολογίες κατατάσσεται και η μέθοδος του στόχου που χρησιμοποιείται σε ορισμένα εμπορικά λογισμικά. Σύμφωνα με τη μέθοδο αυτή η εύρεση της ρίζας του συστήματος βρίσκεται με τον μηδενισμό του «μέτρου» της συνάρτησης. Για παράδειγμα

Η ρίζα του συστήματος

$$f(x, y) = 0$$
$$g(x, y) = 0$$

Ικανοποιεί την εξίσωση

$$f^2 + g^2 = 0$$

Συνεπώς επιδιώκεται ο μηδενισμός (ή η ελαχιστοποίηση) της συνάρτησης $f^2 + g^2$.

Γνωρίζοντας το πεδίο μεταβολής των x και y, μπορούμε να το χωρίσουμε σε N υποδιαστήματα (στις δύο διευθύνσεις) και να υπολογίσουμε τις περιοχές εμφάνισης ελάχιστου (μηδέν) της συνάρτησης $f^2 + g^2$. Τις περιοχές αυτές να τις υποδιαιρέσουμε σε N νέα υποδιαστήματα ακολουθώντας επαναληπτική διαδικασία που θα τερματιστεί μέχρι εντοπισμού των ριζών με την επιθυμητή ακρίβεια.

Η μέθοδος που περιγράφθηκε δεν είναι αυστηρά μαθηματική, είναι όμως ασφαλής και οδηγεί στη λύση (αν αυτή υπάρχει), και βέβαια είναι υπολογιστικά χρονοβόρος.

Είναι φανερό ότι η μέθοδος μπορεί να εφαρμοσθεί και για τον υπολογισμό των μιγαδικών ριζών πολυωνύμου, όπου f είναι το πραγματικό μέρος του πολυωνύμου και g το φανταστικό του μέρος.

Παράδειγμα: Να βρεθεί το σημείο τομής της έλλειψης και του κύκλου με τη μέθοδο του στόχου

$$j := 1..5 \qquad i := 1..5$$

$$x_i := 5 - (i - 1) \qquad y_j := 5 - (j - 1)$$

$$f_{i,j} := (x_i)^2 + (y_j)^2 - 25 \qquad g_{i,j} := \left(\frac{x_i}{10}\right)^2 + \left(\frac{y_j}{2}\right)^2 - 1$$

$$A_{i,j} := (f_{i,j})^2 + (g_{i,j})^2$$

$$A = \begin{pmatrix} 655.25 & 266.563 & 83.25 & 16.063 & 1.25 \\ 285.268 & 58.986 & 1.988 & 25.026 & 64.348 \\ 109.516 & 9.548 & 50.796 & 144.008 & 225.436 \\ 43.984 & 34.242 & 145.664 & 289.002 & 400.504 \\ 28.668 & 73.06 & 226.588 & 400 & 529.548 \end{pmatrix}$$

$$x_1 = 5 \qquad y_5 = 1$$

$$j := 1..5 \qquad i := 1..5$$

$$x_i := 5 - (i - 1)\cdot 0.25 \qquad y_j := 2 - (j - 1)\cdot 0.25$$

$$f_{i,j} := (x_i)^2 + (y_j)^2 - 25 \qquad g_{i,j} := \left(\frac{x_i}{10}\right)^2 + \left(\frac{y_j}{2}\right)^2 - 1$$

$$A_{i,j} := (f_{i,j})^2 + (g_{i,j})^2$$

$$A = \begin{pmatrix} 16.063 & 9.379 & 5.098 & 2.571 & 1.25 \\ 2.492 & 0.391 & 0.08 & 0.913 & 2.341 \\ 0.604 & 2.849 & 6.305 & 10.326 & 14.362 \\ 8.662 & 15.019 & 22.039 & 29.074 & 35.578 \\ 25.026 & 35.259 & 45.64 & 55.518 & 64.348 \end{pmatrix}$$

$$x_2 = 4.75 \quad y_3 = 1.5$$

3.3 Επίλυση γραμμικών συστημάτων

Τα γραμμικά συστήματα έχουν συστηματικά μελετηθεί στη γραμμική άλγεβρα και έχει αποδειχθεί ότι έχουν μοναδική λύση όταν η ορίζουσα των συντελεστών των αγνώστων είναι μη μηδενική. Η μέθοδος που θα αναπτυχθεί στη συνέχεια είναι άμεση μέθοδος επίλυσης γιατί αγνοώντας τα διάφορα σφάλματα του υπολογιστή οδηγεί στην ακριβή λύση με πεπερασμένο πλήθος πράξεων. Έστω το γραμμικό σύστημα n εξισώσεων με n αγνώστους.

$$A \cdot x = B \tag{3.3.1}$$

όπου Α ο τετραγωνικός πίνακας των συντελεστών των αγνώστων **n x n** με στοιχεία $\alpha_{ij} \quad (i = 1,.....n \; και \; j = 1,.....n)$, x ο πίνακας στήλης των αγνώστων με στοιχεία $x_i (i = 1,.....n)$ και Β ο πίνακας στήλης των γνωστών όρων με στοιχεία b_i .

Το σύστημα (3.3.1) έχει μοναδική λύση αν η ορίζουσα του τετραγωνικού πίνακα Α είναι $\det A \neq 0$.

Στην περίπτωση αυτή η γνωστή μέθοδος επίλυσης γραμμικών συστημάτων κατά Cramer οδηγεί στη λύση του συστήματος (3.3.1).

Η μέθοδος κατά Cramer , η οποία είναι γνωστή από τη γραμμική άλγεβρα, απαιτεί τον υπολογισμό $(n+1)$ οριζουσών, κάθε δε ορίζουσα απαιτεί $(n-1)n!$ περίπου πλήθος πράξεων.

Συνεπώς ένας υπολογιστής με ικανότητα εκτέλεσης 10^8 πράξεις στο δευτερόλεπτο θα χρειαζόταν για να επιλύσει ένα σύστημα 20 εξισώσεων κάποιες εκατοντάδες χιλιάδες χρόνια.

Να σημειωθεί ότι στα προβλήματα υπολογιστικής μηχανικής ενδιαφέροντος μηχανικού συνήθως εμφανίζονται συστήματα της τάξης των 100 γραμμικών εξισώσεων .

Έτσι ποτέ σε προβλήματα υπολογιστικής μηχανικής δεν εφαρμόζεται η μέθοδος του Cramer για την επίλυση γραμμικών συστημάτων. Αν η σχέση (3.3.1) πολλαπλασιασθεί με τον αντίστροφο του πίνακα Α,

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

Επειδή $A^{-1} \cdot A = I,$ δηλαδή ο μοναδιαίος πίνακας, προκύπτει η λύση του συστήματος

$$x = A^{-1} \cdot B \tag{3.3.2}$$

Πρέπει όμως να τονισθεί ότι και η αντιστροφή του πίνακα Α είναι χρονοβόρος υπολογιστικά διαδικασία. Αν και υπάρχουν μέθοδοι αντιστροφής πινάκων, η παραπάνω διαδικασία επίλυσης του συστήματος δεν χρησιμοποιείται.

Όλες οι μέθοδοι επίλυσης του συστήματος (3.3.1) βασίζονται στη μέθοδο απαλοιφής κατά Gauss με διάφορες παραλλαγές ή βελτιώσεις της και αυτή η μέθοδος θα παρουσιασθεί στη συνέχεια.

3.3.1 Απαλοιφή κατά Gauss

Η μέθοδος απαλοιφής κατά Gauss για την επίλυση γραμμικών συστημάτων είναι γνωστή από τη γραμμική άλγεβρα. Για να γίνει όμως η αλγοριθμική της παρουσίαση, ακολουθεί ένα αναλυτικό παράδειγμα επίλυσης συστήματος τριών γραμμικών εξισώσεων.

Έστω προς επίλυση το σύστημα

$$\begin{aligned} x_1 + x_2 + x_3 &= 2 \\ 3x_1 + 5x_2 - x_3 &= 14 \\ -x_1 + 3x_2 + 2x_3 &= 3 \end{aligned}$$

Ο τετραγωνικός πίνακας (3 x 3) των συντελεστών των αγνώστων είναι

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 3 & 5 & -1 \\ -1 & 2 & 2 \end{bmatrix}$$

Ο πίνακας στήλης των αγνώστων x είναι

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Ενώ ο πίνακας των γνωστών όρων του δεξιά μέλους είναι

$$B = \begin{bmatrix} 2 \\ 14 \\ 3 \end{bmatrix}$$

Το σύστημα γράφεται και ως

$$\begin{bmatrix} 1 & 1 & 1 \\ 3 & 5 & -1 \\ -1 & 3 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 14 \\ 3 \end{bmatrix}$$

Η απαλοιφή κατά Gauss ακολουθεί τα εξής βήματα

- Πολλαπλασιάζουμε την πρώτη εξίσωση με το 3 (συντελεστή του x_1 της δεύτερης εξίσωσης) και το αποτέλεσμα αφαιρείται από τη δεύτερη εξίσωση. Έτσι προκύπτει το ισοδύναμο σύστημα

$$\begin{aligned} x_1 + x_2 + x_3 &= 2 \\ 2x_2 - 4x_3 &= 8 \\ -x_1 + 3x_2 + 2x_3 &= 3 \end{aligned}$$

Παρατηρούμε ότι με τον τρόπο αυτό απαλείφθηκε ο όρος x_1 από την δεύτερη εξίσωση.

Η διαδικασία επαναλαμβάνεται μεταξύ πρώτης και τρίτης εξίσωσης. Δηλαδή πολλαπλασιάζουμε την πρώτη εξίσωση με –1 και το αποτέλεσμα το αφαιρούμε από την Τρίτη εξίσωση. Έτσι προκύπτει το ισοδύναμο σύστημα

$$\begin{aligned} x_1 + x_2 + x_3 &= 2 \\ 2x_2 - 4x_3 &= 8 \\ 4x_2 + 3x_3 &= 5 \end{aligned}$$

Η διαδικασία επαναλαμβάνεται μεταξύ της δεύτερης και τρίτης εξίσωσης. Δηλαδή πολλαπλασιάζουμε τη δεύτερη εξίσωση με 4/2 και το αποτέλεσμα το αφαιρούμε από την τρίτη εξίσωση, οπότε παίρνουμε το ισοδύναμο σύστημα

$$\begin{aligned} x_1 + x_2 + x_3 &= 2 \\ 2x_2 - 4x_3 &= 8 \\ 11x_3 &= -11 \end{aligned}$$

Το ισοδύναμο σύστημα που δημιουργήθηκε, έχει τριγωνικό πίνακα συντελεστών αγνώστων και γι΄ αυτό η διαδικασία αυτή καλείται και τριγωνοποίηση του συστήματος.

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & -4 \\ 0 & 0 & 11 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 8 \\ -11 \end{bmatrix}$$

Η τελευταία εξίσωση έχει μόνο έναν άγνωστο, τον x_3. Στο στάδιο αυτό (μετά την τριγωνοποίηση) αρχίζει η διαδικασία της πίσω-αντικατάστασης, όπου ξεκινώντας από την τελευταία εξίσωση προχωρούμε προς τα εμπρός, εξίσωση προς εξίσωση και υπολογίζουμε με τη σειρά τους αγνώστους x_3, x_2, x_1.

$$x_3 = -1$$
$$x_2 = (8 + 4x_3)/2 = 2$$
$$x_1 = 2 - x_2 - x_3 = 1$$

δηλαδή η λύση του συστήματος είναι $(x_1, x_2, x_3) = (1,2,-1)$. Η μέθοδος που αναλυτικά παρουσιάσθηκε (τριγωνοποίηση – πίσω αντικατάσταση), μπορεί να γενικευθεί για οποιοδήποτε πλήθος γραμμικών εξισώσεων, όπως φαίνεται στη συνέχεια.

Μέθοδος απαλοιφής κατά Gauss – Γενίκευση

Έστω για επίλυση το σύστημα των m γραμμικών εξισώσεων με m αγνώστους, $x_1, x_2 ....., x_m$.

$$\begin{aligned} &\alpha_{11}x_1 + \alpha_{12}x_2 + ... + a_{1m}x_m = b_1 \\ &\alpha_{21}x_1 + \alpha_{22}x_2 + ... + .\alpha_{2m}x_m = b_2 \\ &..... \\ &\alpha_{m1}x_1 + \alpha_{m2}x_2 + ... + \alpha_{mm}x_m = b_m \end{aligned} \qquad (3.3.1.1)$$

Πολλαπλασιάζουμε την πρώτη των εξισώσεων με α_{21}/α_{11} και αφαιρούμε το αποτέλεσμα από τη δεύτερη των εξισώσεων και το νέο αποτέλεσμα, το οποίο πρέπει να σημειωθεί δεν περιέχει το x_1, το θεωρούμε ως δεύτερη εξίσωση σε αντικατάσταση της δεύτερης εξίσωσης του αρχικού συστήματος. Επαναλαμβάνουμε τη διαδικασία μεταξύ $1^{ης}$ και $3^{ης}$ εξισώσεως, $1^{ης}$ και $4^{ης}$ κλπ. (για την απαλοιφή του x_1). Μετά μεταξύ της $2^{ης}$ και της $3^{ης}$ (καινούργιας εξίσωσης) για την απαλοιφή του x_2, οπότε καταλήγουμε στην παρακάτω ισοδύναμη μορφή ως προς το αρχικό σύστημα έχοντας έτσι επιτύχει την τριγωνοποίηση του πίνακα των συντελεστών των αγνώστων.

$$\begin{aligned} &\alpha_{11}x_1 + \alpha_{12}x_2 + ... + .\alpha_{1m}x_m = b_1 \\ &\qquad \alpha'_{22}x_2 + ... + \alpha'_{2m}x_m = b'_2 \\ &..... \\ &\qquad\qquad\qquad \alpha'_{mm}x_m = b'_m \end{aligned} \qquad (3.3.1.2)$$

κατόπιν ακολουθεί η διαδικασία της πίσω αντικατάστασης.

Η τελευταία εξίσωση μας δίνει την τιμή x_m την οποία αντικαθιστούμε στην προηγούμενη εξίσωση που περιέχει τα x_{m-1} και x_m οπότε υπολογίζουμε την τιμή του x_{m-1} και ούτω καθεξής μέχρι την πρώτη εξίσωση από την οποία θα πάρουμε την τιμή του x_1.

Η αλγοριθμική έκφραση της πίσω αντικατάστασης είναι

$$x_i = \frac{1}{\alpha'_{ii}}\left[b'_i - \sum_{j=i+1}^{m} \alpha'_{ij} x_j \right] \qquad i = m\text{–}1,\ 1$$

Η απαλοιφή κατά Gauss μας οδηγεί πάντα στη λύση του συστήματος αν ο πίνακας των συντελεστών των αγνώστων είναι διάφορος του μηδενός.

Σύμφωνα με τα όσα εκτέθηκαν παραπάνω μπορεί να εκτιμηθεί ότι το πλήθος των πράξεων που απαιτούνται για την επίλυση συστήματος με απαλοιφή κατά Gauss είναι της τάξης του $\frac{n^3}{3}$. Έτσι επανερχόμενοι στο παράδειγμα της παραγράφου 3.3 για την εκτίμηση του χρόνου που απαιτείται για την επίλυση συστήματος 20 γραμμικών εξισώσεων με τη μέθοδο απαλοιφής Gauss προκύπτει χρόνος 0.03 ms (0.03 χιλιοστά του δευτερολέπτου), ενώ για ένα σύστημα 100 γραμμικών εξισώσεων απαιτείται χρόνος μόλις 3.2 ms.

Μέθοδος απαλοιφής κατά Gauss με οδήγηση

Κατά τη διάρκεια της απαλοιφής των αγνώστων από τις εξισώσεις είναι δυνατόν να εμφανιστεί η περίπτωση που ο συντελεστής του αγνώστου της κύριας διαγώνιου (ο συντελεστής a_{ii} του x_i στην **i** εξίσωση) να είναι μηδέν ή σχεδόν μηδέν. Στην περίπτωση αυτή είναι φανερό ότι η διαδικασία δεν μπορεί να συνεχισθεί γιατί στη φάση της πίσω – αντικατάστασης θα γίνει διαίρεση με το μηδέν. Η περίπτωση αυτή και ο τρόπος αντιμετώπισής της γίνεται φανερή με το ακόλουθο παράδειγμα.

$$\begin{aligned} x_1 + x_2 + x_3 &= 2 \\ 3x_1 + 3x_2 - x_3 &= 10 \\ -x_1 + 3x_2 + 2x_3 &= 3 \end{aligned}$$

Η απαλοιφή κατά Gauss οδηγεί στα εξής:

$$\begin{aligned} x_1 + x_2 + x_3 &= 2 \\ 0 \cdot x_2 - 4x_3 &= 4 \\ -x_1 + 3x_2 + 2x_3 &= 3 \end{aligned}$$

όπου διαπιστώνεται ότι ο συντελεστής του x_2 της δεύτερης εξίσωσης είναι μηδέν. Η διαδικασία δεν μπορεί να συνεχιστεί (ακόμα και στην περίπτωση που ο συντελεστής δεν

είναι μηδέν, αλλά ένας πολύ μικρός αριθμός, γιατί η διαίρεση που θα ακολουθούσε στη φάση της πίσω αντικατάστασης θα έδινε αριθμητικά σφάλματα).

Η διαδικασία απαλοιφής όμως μπορεί να συνεχισθεί, μετά τη διαπίστωση της εμφάνισης του μηδενικού συντελεστή με αναδιάταξη των εξισώσεων, όπως

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

Η διαδικασία της απαλοιφής επαναλαμβάνεται μεταξύ της πρώτης και δεύτερης εξίσωσης, οπότε προκύπτει

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

Ο δεύτερος κύκλος της μεθόδου, η πίσω αντικατάσταση δίνει:

$$\begin{aligned} x_3 &= -1 \\ x_2 &= \frac{3 - 3x_3}{3} = 2 \\ x_1 &= 2 - x_3 - x_2 = 1 \end{aligned}$$

δηλαδή η λύση του συστήματος είναι $(x_1, x_2, x_3) = (1,2,-1)$

Είναι φανερό ότι ο έλεγχος της τιμής του κύριου συντελεστή της εξίσωσης που προκύπτει με την απαλοιφή μπορεί να γενικευθεί για σύστημα **n** εξισώσεων και να γίνεται η κατάλληλη αναδιάταξη-οδήγηση των εξισώσεων, οπότε προκύπτει η μέθοδος απαλοιφής Gauss με οδήγηση.

3.3.2 Μέθοδος Gauss – Jordan

Πολλές φορές σε προβλήματα Μηχανικού εμφανίζεται η ανάγκη επίλυσης πολλών γραμμικών συστημάτων της μορφής

$$A \cdot x = B$$

όπου ο πίνακας Α των συντελεστών των αγνώστων παραμένει αμετάβλητος, ενώ μεταβάλλεται ο πίνακας Β των γνωστών όρων (διαφορετικοί πίνακες Β για διαφορετικές περιπτώσεις, π.χ. διαφορετικά φορτία στο δικτύωμα ενός γερανού).

Στη διαδικασία επίλυσης του συστήματος κατά Gauss στους υπολογισμούς υπεισέρχονται οι τιμές του πίνακα Β. Έτσι στο πρόβλημα που τέθηκε (μεταβάλλονται τα εξωτερικά φορτία), στο οποίο οι τιμές του Β μεταβάλλονται, θα χρειαζόταν η εφαρμογή της μεθόδου Gauss τόσες φορές, όσες και το πλήθος των προς επίλυση προβλημάτων.

Στην μέθοδο Gauss-Jordan γίνεται αντιστροφή του πίνακα των συντελεστών των αγνώστων A^{-1}, οπότε η λύση του προβλήματος προκύπτει

$$X = A^{-1} \cdot B \qquad (3.3.2.1)$$

Έτσι η λύση κάθε διαφορετικού προβλήματος βρίσκεται με τον πολλαπλασιασμό των δύο πινάκων A^{-1} και Β, εκ των οποίων ο A^{-1} παραμένει ο ίδιος για κάθε πρόβλημα.

Η μέθοδος Gauss – Jordan απαιτεί περισσότερες πράξεις από τη μέθοδο απαλοιφής Gauss (όπως θα φανεί στην εφαρμογή κάνει δύο φορές απαλοιφή Gauss χωρίς πίσω αντικατάσταση).

Οι πράξεις αυτές είναι της τάξης του $\frac{n^3}{2}$ δηλαδή απαιτούνται 50% περισσότερες πράξεις από τη μέθοδο απαλοιφής Gauss. Τα πλεονεκτήματα όμως της μεθόδου είναι φανερά, όπως η γνώση του αντίστροφου πίνακα και η δυνατότητα λύσης πολλών προβλημάτων, γρήγορα.

Στη συνέχεια παρουσιάζεται σε απλή εφαρμογή επίλυσης του συστήματος τριών γραμμικών εξισώσεων η μέθοδος Gauss-Jordan.

Έστω για επίλυση το σύστημα (n x n, όπου n = 3)

$$\begin{bmatrix} 3 & 3 & 3 \\ 3 & 5 & -1 \\ -1 & 3 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 14 \\ 3 \end{bmatrix}$$

το οποίο έχει λύση $(x_1, x_2, x_3) = (1, 2, -1)$

Από του πίνακες A, x και B δημιουργείται ο επαυξημένος πίνακας 3 x 7, της μορφής

$$\begin{pmatrix} 3 & 3 & 3 & 6 & 1 & 0 & 0 \\ 3 & 5 & -1 & 14 & 0 & 1 & 0 \\ -1 & 3 & 2 & 3 & 0 & 0 & 1 \end{pmatrix}$$

που αποτελείται από τον πίνακα των συντελεστών των αγνώστων A, το διάνυσμα στήλης B και τον μοναδιαίο πίνακα I (n x n).

Βήμα 1^{ο} : Στο πρώτο βήμα οι όροι της πρώτης γραμμής διαιρούνται με τον κύριο συντελεστή της γραμμής (πρώτη γραμμή, πρώτος όρος της γραμμής, δεύτερη γραμμή, δεύτερος όρος της γραμμής, κλπ.) διαδικασία που καλείται κανονικοποίηση (διαιρώντας με 3). Έτσι προκύπτει

$$AB = \begin{pmatrix} 1 & 1 & 1 & 2 & 0.333 & 0 & 0 \\ 3 & 5 & -1 & 14 & 0 & 1 & 0 \\ -1 & 3 & 2 & 3 & 0 & 0 & 1 \end{pmatrix}.$$

Μετά πολλαπλασιάζεται η νέα πρώτη γραμμή με τρία και το αποτέλεσμα αφαιρείται από τη δεύτερη γραμμή, έτσι ώστε να γίνει μηδέν ο πρώτος όρος της δεύτερης σειράς (τριγωνοποίηση).

$$AB = \begin{pmatrix} 1 & 1 & 1 & 2 & 0.333 & 0 & 0 \\ 0 & 2 & -4 & 8 & -1 & 1 & 0 \\ -1 & 3 & 2 & 3 & 0 & 0 & 1 \end{pmatrix}.$$

Η διαδικασία επαναλαμβάνεται μεταξύ πρώτης και τρίτης γραμμής (όπως στην απαλοιφή κατά Gauss), οπότε προκύπτει

$$\begin{bmatrix} 1 & 1 & 1 & 2 & 1/3 & 0 & 0 \\ 0 & 2 & -4 & 8 & -1 & 1 & 0 \\ 0 & 4 & 3 & 5 & 1/3 & 0 & 1 \end{bmatrix}$$

Βήμα 2^{ο} : Στο δεύτερο βήμα επαναλαμβάνεται το 1^{ο} βήμα, αλλά μεταξύ της πρώτης και δεύτερης των εξισώσεων για απαλοιφή του όρου a'_{12}. Έτσι έχουμε

$$\begin{bmatrix} 1 & 1 & 1 & 2 & 1/3 & 0 & 0 \\ 0 & 1 & -2 & 4 & -1/2 & 1/2 & 0 \\ 0 & 4 & 3 & 5 & 1/3 & 0 & 1 \end{bmatrix}$$

(κανονικοποίηση, διαίρεση με το 2)

$$\begin{bmatrix} 1 & 0 & +3 & -2 & 5/6 & -1/2 & 0 \\ 0 & 1 & -2 & 4 & -1/2 & 1/2 & 0 \\ 0 & 4 & 3 & 5 & 1/3 & 0 & 1 \end{bmatrix}$$

(απαλοιφή για την πρώτη σειρά)

Μετά πολλαπλασιάζεται η δεύτερη εξίσωση με το 4 και αφαιρείται από την τρίτη εξίσωση απαλείφοντας τον όρο a_{32}.

$$\begin{bmatrix} 1 & 0 & 3 & -2 & 5/6 & -1/2 & 0 \\ 0 & 1 & -2 & 4 & -1/2 & 1/2 & 0 \\ 0 & 0 & 11 & -11 & -7/3 & -2 & 1 \end{bmatrix}$$

Βήμα 3^{ο} : Κανονικοποίηση τρίτης εξίσωσης(διαίρεση με το 11)

$$\text{AXB} = \begin{pmatrix} 1 & 0 & 3 & -2 & 0.833 & -0.5 & 0 \\ 0 & 1 & -2 & 4 & -0.5 & 0.5 & 0 \\ 0 & 0 & 1 & -1 & 0.212 & -0.182 & 0.091 \end{pmatrix},$$

και μηδενισμός των όρων, α_{23} και α_{13} της δευτέρας και πρώτης εξίσωσης.

$$\text{AXB} = \begin{pmatrix} 1 & 0 & 3 & -2 & 0.833 & -0.5 & 0 \\ 0 & 1 & 0 & 2 & -0.076 & 0.136 & 0.182 \\ 0 & 0 & 1 & -1 & 0.212 & -0.182 & 0.091 \end{pmatrix},$$

ή

$$\text{AXB} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0.197 & 0.045 & -0.273 \\ 0 & 1 & 0 & 2 & -0.076 & 0.136 & 0.182 \\ 0 & 0 & 1 & -1 & 0.212 & -0.182 & 0.091 \end{pmatrix},$$

Παρατηρώντας τον επαυξημένο πίνακα που προέκυψε διαπιστώνεται ότι ο τετραγωνικός πίνακας 3 x 3 (n x n) είναι ο μοναδιαίος πίνακας.

Ο επόμενος πίνακας στήλη (1 x 3) αποτελεί το διάνυσμα x της λύσης του συστήματος. Ενώ ο επόμενος τετραγωνικός πίνακας (3 x 3) αποτελεί τον αντίστροφο πίνακα A^{-1} του πίνακα των συντελεστών των αγνώστων

$$A^{-1} = \begin{pmatrix} 0.197 & 0.045 & -0.273 \\ -0.076 & 0.136 & 0.182 \\ 0.212 & -0.182 & 0.091 \end{pmatrix}.$$

Αποδεικνύεται επίσης ότι η ορίζουσα των συντελεστών των αγνώστων $\det A$, ισούται με το γινόμενο των τριών αριθμών με τους οποίους έγινε στα διάφορα βήματα η κανονικοποίηση των εξισώσεων. Στην παρούσα περίπτωση

$$\det A = 3 \cdot 2 \cdot 11 = 66$$

Αν χρειαστεί στη συνέχεια να επιλυθεί το ίδιο σύστημα αλλά με διαφορετικές τιμές στο δεύτερο μέλος, π.χ.

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

τότε η λύση προκύπτει απλά με τον πολλαπλασιασμό των πινάκων A^{-1} και B.

$$A := \begin{pmatrix} 3 & 3 & 3 \\ 3 & 5 & -1 \\ -1 & 3 & 2 \end{pmatrix} \qquad A^{-1} = \begin{pmatrix} 0.197 & 0.045 & -0.273 \\ -0.076 & 0.136 & 0.182 \\ 0.212 & -0.182 & 0.091 \end{pmatrix} \qquad A \cdot A^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$b := \begin{pmatrix} 6 \\ 5 \\ 2 \end{pmatrix} \qquad xvector := A^{-1} \cdot b \qquad xvector = \begin{pmatrix} 0.864 \\ 0.591 \\ 0.545 \end{pmatrix} \qquad |A| = 66$$

Αλγοριθμικά η διαδικασία κανονικοποίησης εκφράζεται με τη σχέση

$$\alpha_{kj} \leftarrow \frac{\alpha_{kj}}{\alpha_{kk}} \qquad j = n+m, n+m-1, \ldots, k$$
$$k = 1,2,\ldots,n$$

ενώ η διαδικασία απαλοιφής εκφράζεται με τη σχέση

$$\left.\begin{array}{l}\alpha_{ij} \leftarrow \alpha_{ij} - \alpha_{ik}\alpha_{kj} \\ j = n+m, n+m-1, .., k \end{array}\right\} i = 1,2,.....,n \quad (i \neq k)$$
$$k = 1,2,.....n$$

3.3.3 Επίλυση συστήματος γραμμικών εξισώσεων Τριδιαγώνιας μορφής

Πολλές φορές η αριθμητική επίλυση διαφορικών εξισώσεων οδηγεί σε γραμμικό σύστημα αλγεβρικών εξισώσεων του οποίου ο πίνακας των συντελεστών των αγνώστων είναι τριδιαγώνιος, δηλαδή όλα του τα στοιχεία είναι μηδενικά εκτός από τα εκατέρωθεν στοιχεία της κυρίας διαγωνίου.

Ένα τέτοιο σύστημα εξισώσεων μπορεί να γραφεί στη μορφή

$$\begin{Bmatrix} B_1 & C_1 & 0 & & & \\ A_2 & B_2 & C_2 & 0 & & \\ 0 & A_3 & B_3 & C_3 & 0 & \\ & & - & - & - & \\ & & - & - & - & \\ & & 0 & A_{N-1} & B_{N-1} & C_{N-1} \\ & & 0 & 0 & A_N & B_N \end{Bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ - \\ - \\ x_{n-1} \\ x_n \end{Bmatrix} = \begin{Bmatrix} R_1 \\ R_2 \\ R_3 \\ - \\ - \\ R_{N-1} \\ R_N \end{Bmatrix} \qquad (3.3.3.1)$$

Εφαρμόζοντας την απαλοιφή κατά Gauss σταδιακά έχουμε (για n = 3) (κανονικοποίηση και απαλοιφή)

$$\begin{Bmatrix} 1 & C_1/B_1 & 0 \\ A_2 & B_2 & C_2 \\ & A_3 & B_3 \end{Bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} = \begin{Bmatrix} R_1/B_1 \\ R_2 \\ R_3 \end{Bmatrix}$$

ή

$$\begin{Bmatrix} 1 & C_1/B_1 & 0 \\ 0 & (B_2 - A_2C_1/B_1) & C_2 \\ & A_3 & B_3 \end{Bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} = \begin{Bmatrix} R_1/B_1 \\ (R_2 - A_2R_1/B_1) \\ R_3 \end{Bmatrix} \qquad (3.3.3.2)$$

Η δεύτερη εξίσωση (3.3.3.2) αν διαιρεθεί με $B_2 - A_2C_1/B_1$, κατόπιν πολλαπλασιασθεί με A_3 και αφαιρεθεί από την τρίτη εξίσωση (απαλοιφή), δίνει νέα εξίσωση που στη θέση του A_3 έχει το στοιχείο μηδέν.

Τελικά το σύστημα γράφεται στη μορφή

$$\begin{Bmatrix} 1 & C_1' & 0 & & & \\ 0 & 1 & C_2' & 0 & & \\ - & 0 & 1 & C_3' & 0 & \\ & - & - & - & - & \\ & - & - & - & - & \\ & & & 0 & 1 & C'_{n-1} \\ & & & & 0 & 1 \end{Bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ - \\ - \\ x_{n-1} \\ x_n \end{Bmatrix} = \begin{Bmatrix} R'_1 \\ R'_2 \\ R'_3 \\ - \\ - \\ R'_{N-1} \\ R'_N \end{Bmatrix} \qquad (3.3.3.3)$$

Η τελευταία εξίσωση του συστήματος (3.3.3.3) δίνει

$$x_n = R'_n$$

και τότε έχουμε τον αναγωγικό τύπο

$$x_{n-1} = -x_n C'_{n-1} + R'_{n-1} \qquad (3.3.3.4)$$

για την εύρεση των υπολοίπων αγνώστων x_n.

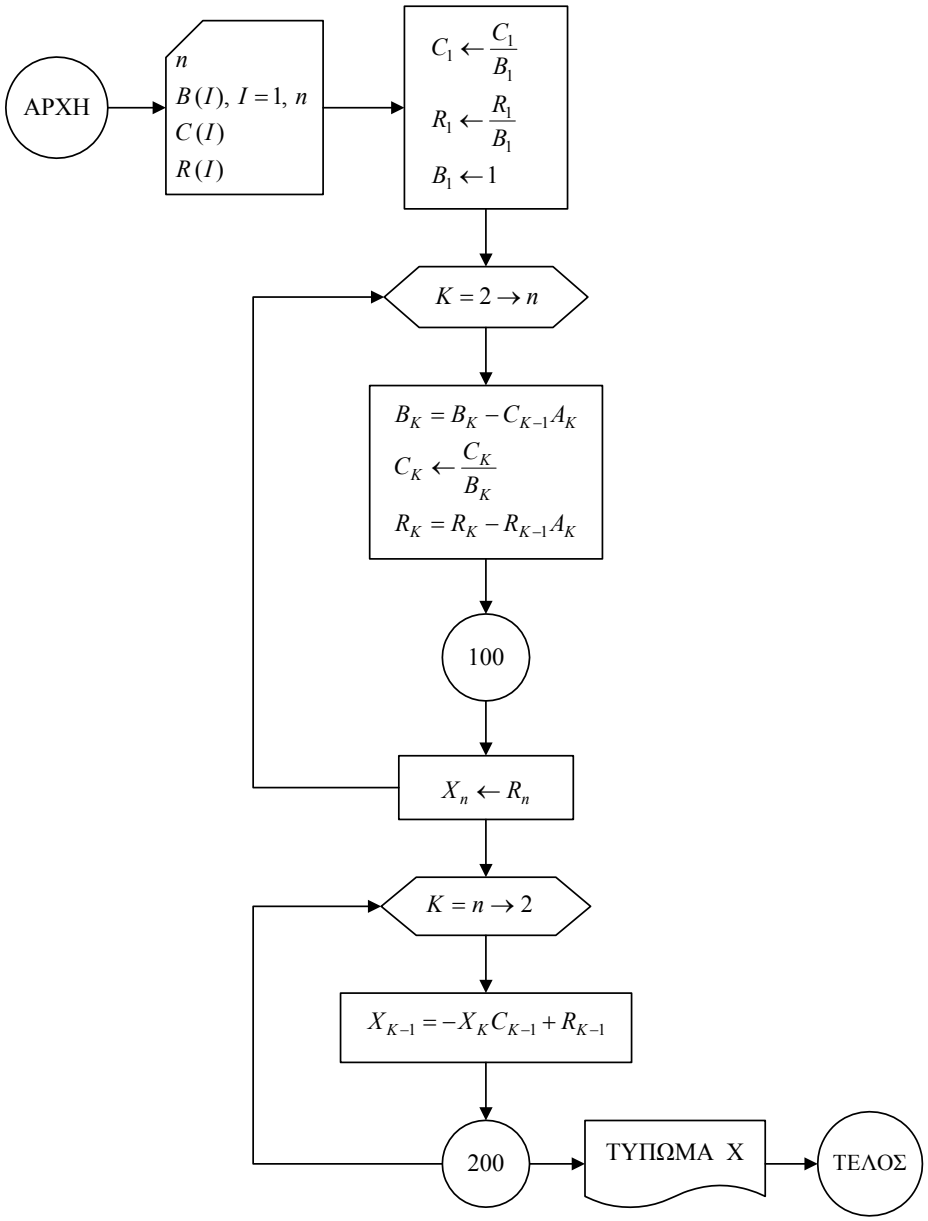
Όπως αναπτύχθηκε, η μέθοδος μπορεί εύκολα να γραφεί σε μορφή αλγορίθμου. Το δομικό διάγραμμα φαίνεται στο σχήμα 3.3.3.1.

Ας εξετάσουμε τώρα τα πλεονεκτήματα της μεθόδου που παρουσιάσθηκε.

Τα μηδενικά στοιχεία του πίνακα παραμένουν μηδενικά στον αλγόριθμο, συνεπώς δεν χρειάζεται να αποθηκευτούν στη μνήμη του υπολογιστή. Έτσι προκύπτει μεγάλη οικονομία στις απαιτήσεις σε μνήμη υπολογιστή. Συγκεκριμένα απαιτείται η αποθήκευση μόνο 5n – 2 στοιχείων σε σύγκριση με $n^2 + 2n$ που θα χρειάζονταν αν ο πίνακας δεν ήταν τριδιαγώνιος.

Οι απαιτούμενες αριθμητικές πράξεις είναι της τάξεως του n ενώ όπως είδαμε στην απαλοιφή κατά Gauss οι πράξεις είναι της τάξεως του n^3. Συνεπώς προκύπτει σημαντικότατη οικονομία και στο χρόνο CPU του υπολογιστή.

Από τα προηγούμενα σημαντικά πλεονεκτήματα της μεθόδου εξηγείται η ευρύτατη χρήση της στις επιλύσεις τριδιαγώνιας μορφής γραμμικών συστημάτων.



Σχήμα 3.3.3.1: Δομικό διάγραμμα επίλυσης συστήματος τριδιαγώνιου χαρακτήρα.

3.3.4 Μέθοδος Ανάλυσης LU

Η μέθοδος ανάλυσης LU είναι η δημοφιλέστερη μέθοδος για την επίλυση γραμμικών συστημάτων,

$$A \cdot x = B \tag{3.3.4.1}$$

Έστω ότι ο πίνακας των συντελεστών των αγνώστων Α μπορεί να αναλυθεί σε γινόμενο δύο τριγωνικών πινάκων ως

$$L \cdot U = A$$

όπου L είναι ένας κάτω τριγωνικός πίνακας (έχει δηλαδή στοιχεία μη μηδενικά μόνο στην κύρια διαγώνιο και κάτω από αυτή, ενώ τα πάνω από την κύρια διαγώνιο είναι μηδενικά) και U είναι ένας πάνω τριγωνικός πίνακας (δηλαδή έχει στοιχεία μη μηδενικά στην κύρια διαγώνιο και πάνω από αυτή, ενώ κάτω από την κύρια διαγώνιο τα στοιχεία του πίνακα είναι μηδενικά).

Για παράδειγμα ο τετραγωνικός πίνακας 3 x 3

$$A = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{bmatrix}$$

Επιδιώκεται να αναλυθεί στον κάτω τριγωνικό L

$$L = \begin{bmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & \lambda_{22} & 0 \\ \lambda_{31} & \lambda_{32} & \lambda_{33} \end{bmatrix}$$

και στον πάνω τριγωνικό U

$$U = \begin{bmatrix} u_{11} & u_{12} & u_{12} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

με στοιχεία λ_{ij} και u_{ij} που πρέπει να υπολογισθούν.

Με την ανάλυση του πίνακα Α σε L και U η εξίσωση (3.3.4.1) γράφεται

$$\left(L \cdot U\right) \cdot x = B$$

ή

$$L \cdot (U \cdot x) = B$$

Ο πίνακας $U \cdot x$ είναι ένα διάνυσμα στήλης Y

$$U \cdot x = Y \tag{3.3.4.2}$$

Έτσι έχουμε

$$L \cdot Y = B \tag{3.3.4.3}$$

Το αρχικό μας πρόβλημα αναλύεται σε δύο επιμέρους προβλήματα

(α) υπολογισμού του ενδιάμεσου διανύσματος Y, εξίσωση (3.3.4.3) και

(β) υπολογισμού της τελικής λύσης X με την επίλυση της εξίσωσης (3.3.4.2).

Η ανάλυση της επίλυσης του αρχικού συστήματος σε δύο επί μέρους προβλήματα αποδεικνύεται ότι είναι υπολογιστικά ταχύτατη, λόγω του τριγωνικού χαρακτήρα των πινάκων L και U.

Η εξίσωση (3.3.4.3) είναι πολύ εύκολο να λυθεί με εμπρός αντικατάσταση και να δώσει το διάνυσμα στήλης Y. Αλγοριθμικά η εμπρός αντικατάσταση γράφεται

$$y_1 = \frac{b_{11}}{\lambda_{11}}$$

$$y_i = \frac{1}{\lambda_{ii}} \left[b_i - \sum_{j=1}^{i-1} \lambda_{ij} \quad y_j \right], \quad i = 2,3,..,N$$

Η εξίσωση (3.3.4.2), η οποία και θα δώσει τη λύση X του αρχικού προβλήματος, επιλύεται επίσης αλλά με πίσω αντικατάσταση.

$$X_N = \frac{y_N}{u_{NN}}$$

$$x_i = \frac{1}{u_{ii}} \left[y_i - \sum_{j=i+1}^{N} u_{ij} \quad x_j \right], \quad i = N-1, N-2,..,1$$

Το πρόβλημα συνεπώς της επίλυσης του αρχικού συστήματος (3.3.4.1) μετατίθεται στην ανάλυση του πίνακα A στους δύο τριγωνικούς πίνακες L και U με τον υπολογισμό των στοιχείων των λ_{ij} και u_{ij} των πινάκων L και U αντίστοιχα.

Από την εξίσωση

$$L \cdot U = A$$

προκύπτει η βασική σχέση υπολογισμού των αγνώστων στοιχείων λ_{ij} και u_{ij}.

$$\lambda_{im} u_{mj} = \alpha_{ij} \quad \text{(άθροιση ως προς δείκτη m)} \tag{3.3.4.4}$$

για $i, j = 1,2,..., N$.

Το πλήθος των αγνώστων στοιχείων για τους δύο πίνακες είναι $N^2 + N$, ενώ το πλήθος των εξισώσεων που υπάρχουν προς επίλυση (3.3.4.4) είναι N^2.

Έτσι υπάρχει η δυνατότητα καθορισμού N αγνώστων στοιχείων και ως τέτοια επιλέγονται τα διαγώνια στοιχεία του πίνακα L τα οποία θεωρούνται ίσα με τη μονάδα

$$(\lambda_{11} = \lambda_{22} = \lambda_{33} = ... = 1).$$

(μέθοδος Doolittle) ή θεωρούνται ίσα με τη μονάδα τα διαγώνια στοιχεία του πίνακα U (μέθοδος Crout).

Η μέθοδος θα εφαρμοσθεί στο ακόλουθο παράδειγμα επίλυσης του συστήματος.

$$\begin{bmatrix} 1 & 1 & 1 \\ 3 & 5 & -1 \\ -1 & 3 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 14 \\ 3 \end{bmatrix}$$

Ζητείται να ευρεθούν οι τριγωνικοί πίνακες L και U

$$L \cdot U = \begin{bmatrix} 1 & 1 & 1 \\ 3 & 5 & -1 \\ -1 & 3 & 2 \end{bmatrix}$$

όπου

$$L = \begin{bmatrix} 1 & 0 & 0 \\ \lambda_{21} & 1 & 0 \\ \lambda_{31} & \lambda_{32} & 1 \end{bmatrix}$$

$$U = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

Ο πολλαπλασιασμός γραμμών (i) του πίνακα L και στηλών (j) του πίνακα U δίνει τα στοιχεία α_{ij}

Για την πρώτη στήλη προκύπτει:

$$\lambda_{11}u_{11} + \lambda_{12}\underset{0}{u_{21}} + \lambda_{13}\underset{0}{u_{31}} = \alpha_{11}$$
$$\lambda_{21}u_{11} + \lambda_{22}\underset{0}{u_{21}} + \lambda_{23}\underset{0}{u_{31}} = \alpha_{21}$$
$$\lambda_{31}u_{11} + \lambda_{32}\underset{0}{u_{21}} + \lambda_{33}\underset{0}{u_{31}} = \alpha_{31}$$

Οι τρεις ανωτέρω σχέσεις δίνουν αμέσως τι τιμές των $u_{11}, \lambda_{21},$ και λ_{31}

με $u_{11} = \alpha_{11}, \quad \lambda_{21} = \alpha_{21} \big/ \alpha_{11}, \quad \lambda_{31} = \alpha_{31} \big/ \alpha_{11}$

Για την δεύτερη στήλη προκύπτει:

$$\lambda_{11}u_{12} + \lambda_{12}u_{22} + \lambda_{13}u_{32} = \alpha_{12}$$
$$\lambda_{21}u_{12} + \lambda_{22}u_{22} + \lambda_{23}u_{32} = \alpha_{22}$$
$$\lambda_{31}u_{12} + \lambda_{32}u_{22} + \lambda_{33}u_{32} = \alpha_{32}$$

Από τις παραπάνω τρεις εξισώσεις, η πρώτη δίνει

$$u_{12} = \alpha_{12}$$

η δεύτερη δίνει

$$u_{22} = a_{22} - \frac{a_{21}}{a_{11}} a_{12}$$

ενώ η τρίτη δίνει

$$\lambda_{32} = \frac{1}{\left(a_{22} - \frac{a_{21}}{a_{11}} a_{12} \right)} \left(a_{32} - \frac{a_{31}}{a_{11}} a_{12} \right)$$

Και για την τρίτη στήλη

$$\lambda_{11}u_{13} + \lambda_{12}u_{23} + \lambda_{13}u_{33} = \alpha_{13}$$
$$\lambda_{21}u_{13} + \lambda_{22}u_{23} + \lambda_{23}u_{33} = \alpha_{23}$$
$$\lambda_{31}u_{13} + \lambda_{32}u_{23} + \lambda_{33}u_{33} = \alpha_{33}$$

Η πρώτη των εξισώσεων δίνει

$$u_{13} = a_{13}$$

Η δεύτερη δίνει

$$u_{23} = a_{23} - \frac{a_{21}}{a_{11}} a_{13}$$

ενώ η τρίτη δίνει

$$u_{33} = a_{33} - \lambda_{32}u_{23} - \lambda_{31}u_{13}$$

Αντικατάσταση των τιμών των όρων a_{ij} της μήτρας Α στις παραπάνω εκφράσεις δίνει τους όρους των πινάκων L και U

$$M := \begin{pmatrix} 1 & 1 & 1 \\ 3 & 5 & -1 \\ -1 & 3 & 2 \end{pmatrix} \qquad M^{-1} = \begin{pmatrix} 0.591 & 0.045 & -0.273 \\ -0.227 & 0.136 & 0.182 \\ 0.636 & -0.182 & 0.091 \end{pmatrix}$$

$$L := \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ -1 & 2 & 1 \end{pmatrix} \qquad U := \begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & -4 \\ 0 & 0 & 11 \end{pmatrix}$$

$$L \cdot U = \begin{pmatrix} 1 & 1 & 1 \\ 3 & 5 & -1 \\ -1 & 3 & 2 \end{pmatrix}$$

Η προηγούμενη εφαρμογή έδειξε ότι με κατάλληλη μεθοδολογία είναι δυνατόν να υπολογισθούν όλοι οι όροι των δύο πινάκων με απλές αλγεβρικές πράξεις, κάτι που μεταφράζεται σε ταχύτητα υπολογισμών.

Πρέπει να σημειωθεί ότι στη διαδικασία ανάλυσης του πίνακα Α δεν υπεισήλθε το δεύτερο μέρος Β. Άρα οι μήτρες L και U μπορούν να υπολογισθούν μια φορά για τον πίνακα Α και να επιλυθούν διάφορα προβλήματα με διαφορετικά δεξιά μέλη των εξισώσεων.

Επειδή τα διαγώνια στοιχεία του πίνακα L (ή U ανάλογα με τη μέθοδο) ισούνται με την μονάδα, εξοικονομείται μνήμη υπολογιστή αν στη θέση του πίνακα Α μετά την αποσύνθεση σε L και U πίνακες αποθηκευθεί το αποτέλεσμα. Αυτό μπορεί να γίνει ως το ακόλουθο παράδειγμα.

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{33} & a_{33} \end{bmatrix} \rightarrow \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ \lambda_{21} & u_{22} & u_{23} \\ \lambda_{31} & \lambda_{32} & u_{33} \end{bmatrix} \qquad (3.3.4.5)$$

όπου υπονοείται ότι τα διαγώνια στοιχεία του πίνακα L ισούνται με τη μονάδα. Έτσι οι πίνακες L και U γράφονται στη συντμημένη μορφή του δεξιού μέλους της (3.3.4.5)

Ο Crout επέλυσε το σύστημα (3.3.4.4) κατά ένα έξυπνο και σύντομο τρόπο με αναδιάταξη των εξισώσεων. Η μέθοδος Crout αλγοριθμικά γράφεται

$$\lambda_{i,1} = a_{i,1} \quad \gamma\iota\alpha \quad i = 1,2,..,n$$

$$u_{1j} = \frac{a_{1j}}{\lambda_{11}} \quad \gamma\iota\alpha \quad j = 2,3,..,n$$

Για $j = 2,3,..,n-1$

$$\lambda_{ij} = a_{ij} - \sum_{k=1}^{j-1} \lambda_{ik} u_{kj} \quad \gamma\iota\alpha \quad i = j, j+1,..,n \quad (i \geq j)$$

$$u_{jk} = \frac{a_{jk} - \sum_{i=1}^{j-1} \lambda_{ji} u_{ik}}{\lambda_{jj}} \quad \gamma\iota\alpha \quad k = j+1, j+2,..,n \quad (k \geq j)$$

και

$$\lambda_{nn} = a_{nn} - \sum_{k=1}^{n-1} \lambda_{nk} u_{kn}$$

3.3.5 Επαναληπτική μέθοδος (Gauss-Seidel)

Σε προβλήματα ερευνητικού επιπέδου τα συστήματα γραμμικών εξισώσεων που εμφανίζονται για επίλυση είναι μεγάλα (σε σύνθετα προβλήματα μπορεί να είναι και 100000 εξισώσεις)

Στην περίπτωση αυτή η μέθοδος απαλοιφής Gauss θα χρειασθεί χρόνο της τάξης των εβδομάδων ή μηνών, δηλαδή υπολογιστικοί χρόνοι τεχνικά μη παραδεκτοί. Το σύστημα όμως των εξισώσεων μπορεί να λυθεί (αν και γραμμικό) με την επαναληπτική μέθοδο Gauss-Seidel (βλ. παρ. 3.2.1).

Για το γραμμικό σύστημα εξισώσεων n x n η επαναληπτική μέθοδος Gauss-Seidel παίρνει την αλγοριθμική μορφή

$$x_{i_0}^{(k+1)} = b_i^1 - \sum_{j=1}^{i-1} \alpha'_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} \alpha'_{ij} x_{jk}^{(k)}$$

όπου

$$b_i' = \frac{b_i}{\alpha_{ii}}, \quad \alpha_{ij}' = \frac{\alpha_{ij}}{\alpha_{ii}}$$

Επειδή συνήθως υπολογιστικά οι νέες τιμές των $x_i^{(k+1)}$ αποθηκεύονται στην ίδια θέση μνήμης με τις τιμές $x_i^{(k)}$, ο ανωτέρω αλγόριθμος απλοποιείται στην

$$x_i = b_i' - \sum_{\substack{j=1 \\ j\neq i}}^{n} \alpha_{ij}' x_j, \qquad i = 1,2,...,n$$

Η προηγούμενη επαναληπτική διαδικασία συγκλίνει υπό προϋποθέσεις, ανάλογα με τη μορφή των προς επίλυση εξισώσεων και τις τιμές των συντελεστών των αγνώστων (για γραμμικά συστήματα απαιτείται υπεροχή του συντελεστή της κυρίας διαγωνίου). Η επαναληπτική διαδικασία θεωρείται ότι συγκλίνει αν οι τιμές των x_i δύο διαδοχικών επαναλήψεων διαφέρουν μιας μικρής τιμής **ε**, του κριτηρίου ακρίβειας

$$\left| x_i^{(k+1)} - x_i^{(k)} \right| < \varepsilon, \qquad \gamma\iota\alpha \qquad i = 1,2,..,n$$

Φυσικά για λόγους υπολογιστικής προστασίας τίθεται μέγιστο πλήθος επαναλήψεων.

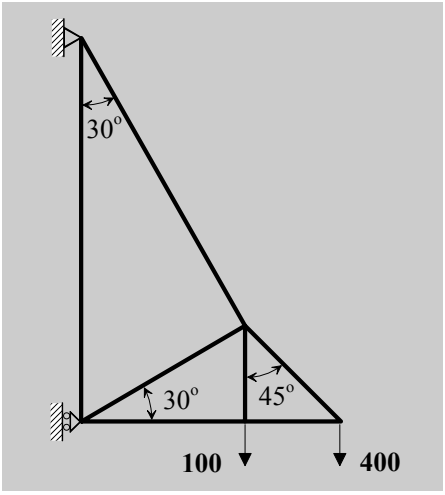
Σύγκριση μεταξύ των μεθόδων Gauss-Seidel και απαλοιφής Gauss δίνει τα εξής συμπεράσματα.

1. Η απαλοιφή Gauss, θεωρητικά δίνει πάντοτε λύση του συστήματος για μη μηδενική ορίζουσα των συντελεστών των αγνώστων, ενώ η Gauss Seidel όχι.
2. Η απαλοιφή κατά Gauss είναι πολύ ευαίσθητη στα σφάλματα από στρογγυλοποίηση, ενώ η μέθοδος Gauss-Seidel όχι.
3. Η μέθοδος Gauss-Seidel είναι πολύ εύκολο να προγραμματισθεί και απαιτεί μικρότερη χωρητικότητα μνήμης υπολογιστή.
4. Αν το σύστημα δεν είναι πλήρες (δηλαδή κάθε εξίσωση δεν περιέχει όλους τους αγνώστους) η εργασία για τη μέθοδο Gauss-Seidel μειώνεται, ενώ για την απαλοιφή Gauss παραμένει η ίδια.
5. Αν το σύστημα των εξισώσεων είναι μη γραμμικό τότε μόνο η μέθοδος Gauss-Seidel μπορεί να εφαρμοσθεί.
6. Ο απαιτούμενος αριθμός πράξεων για τη μέθοδο Gauss-Seidel είναι ανάλογος του m^2 ανά δοκιμή, ενώ για την απαλοιφή Gauss ο συνολικός αριθμός των πράξεων είναι ανάλογος του m^3. Συνεπώς πάνω από ένα αριθμό εξισώσεων *m* η μέθοδος Gauss-Seidel είναι υπολογιστικά οικονομικότερη.

3.4 Υπολογιστικές προσομοιώσεις για το Εργαστήριο Η/Υ

Εφαρμογή 1η

α. Λεκτική περιγραφή: Να υπολογισθούν οι δυνάμεις που καταπονούν τις συνδετήριες δοκούς επίπεδου δικτυώματος απλού ανυψωτικού μηχανήματος καθώς και οι αντιδράσεις στήριξης, όπως στο σχήμα 3.4.1.



Σχήμα 3.4.1: Δικτύωμα απλού ανυψωτικού μηχανήματος.

β. Μαθηματική περιγραφή

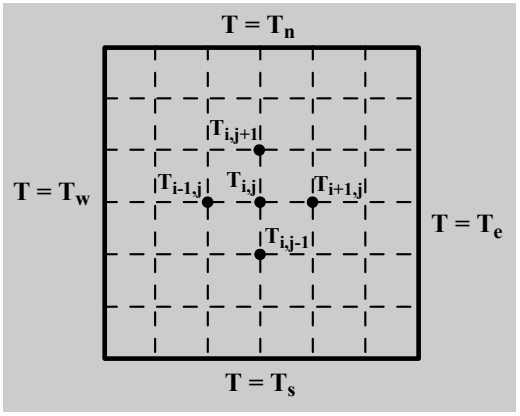
Η μαθηματικοποίηση του προβλήματος βασίζεται στην ισορροπία των δυνάμεων στον οριζόντιο και κατακόρυφο άξονα σε όλους τους κόμβους του δικτυώματος. Έτσι προκύπτει ένα σύστημα γραμμικών εξισώσεων που πρέπει να επιλυθεί με τη μέθοδο Gauss-Jordan ή LU.

Εφαρμογή 2η

α. Λεκτική περιγραφή: Να ευρεθεί η διανομή της θερμοκρασίας σε ράβδο τετραγωνικής διατομής με γνωστές θερμοκρασίες τοιχωμάτων

β. Μαθηματική περιγραφή

Η μέθοδος της χαλάρωσης είναι γνωστή ως μια γρήγορη μέθοδος εύρεσης θερμοκρασιών. Βασίζεται στη σχέση ότι η θερμοκρασία σε κάποιο σημείο μέσα στον χώρο ισούται με το τέταρτο των θερμοκρασιών τεσσάρων γειτονικών σημείων που ισαπέχουν του σημείου, ως στο σχήμα 3.4.2.

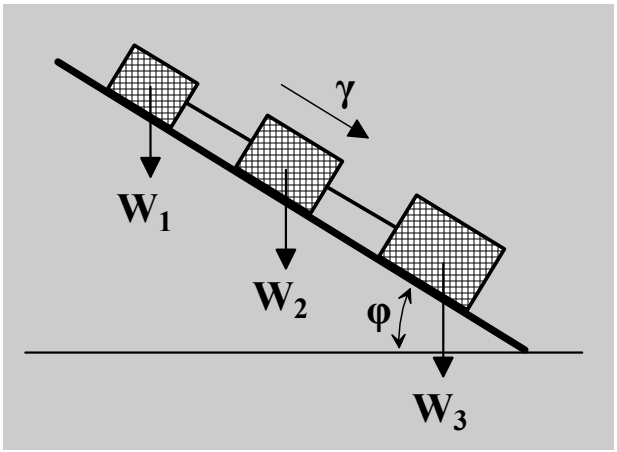


$$T_{i,j} = \frac{T_{i+1,j} + T_{i-1,j} + T_{i,j+1} + T_{i,j-1}}{4}$$

Σχήμα 3.4.2: Πλέγμα υπολογισμού θερμοκρασιών σε τομή ράβδου.

Εφαρμογή 3η

α. Λεκτική περιγραφή: Να ευρεθούν οι τάσεις των καλωδίων που συνδέουν τα δεδομένα βάρη του σχήματος 3.4.3, καθώς και η επιτάχυνση γ με την οποία ολισθαίνουν στην κεκλιμένη επιφάνεια, για δεδομένη κλίση φ και συντελεστές τριβής t_1, t_2 και t_3.



Σχήμα 3.4.3: Ολίσθηση συνδεδεμένων τεμαχίων.

Κεφάλαιο 4

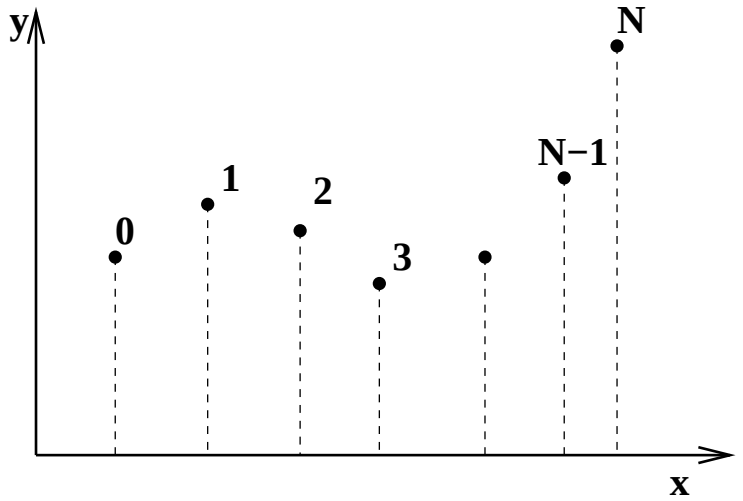
Αριθμητική Παρεμβολή και Προσέγγιση

4.1 Η Ανάγκη για Αριθμητική Παρεμβολή ή Προσέγγιση

4.1.1 Το Πρόβλημα – Μερικοί Βασικοί Ορισμοί

Το κεφάλαιο αυτό διαπραγματεύεται θέματα αριθμητικής παρεμβολής και προσέγγισης καμπυλών στο επίπεδο ή/και στο χώρο. Οι όροι *παρεμβολή* και *προσέγγιση* θα οριστούν αυστηρά στο τέλος της εισαγωγικής ενότητας και, όπως είναι ευνόητο, η παρουσίαση θα εστιασθεί κυρίως σε θέματα καμπυλών στο επίπεδο, αν και αρκετές από τις μεθόδους που θα παρουσιασθούν επεκτείνονται και για τις καμπύλες στο χώρο με αυτονόητες μετατροπές. Οι τεχνικές που θα αναπτυχθούν στις παρακάτω θεματικές ενότητες αλλά κυρίως το αντίστοιχο λογισμικό αποτελούν συχνά πρωτεύοντα ή δευτερεύοντα 'εργαλεία' υποστήριξης του έργου των μηχανικών οποιασδήποτε ειδίκευσης.

Ως τυπική περίπτωση χρήσης τους θα μπορούσαμε να αναφέρουμε ότι, πολύ συχνά, η ανάλυση ενός φυσικού προβλήματος με πειράματα-μετρήσεις ή μοντελοποίηση-υπολογισμούς παράγει μια σειρά τιμών απόκρισης που η κάθε μία τους αντιστοιχεί σε διαφορετική αριθμητική τιμή μιας παραμέτρου εισόδου. Για λόγους απλότητας ας γίνει αρχικά η παραδοχή ότι αναφερόμαστε σε μονοπαραμετρικό πρόβλημα, δηλαδή ότι η *απόκριση* y εξαρτάται από μία και μόνο μία παράμετρο, τη λεγόμενη *παράμετρο εισόδου* x. Αλλά ακόμα και για πολυπαραμετρικά προβλήματα, όπως είναι τα περισσότερα φυσικά προβλήματα, η παρακάτω ανάλυση έχει αξία αν υποτεθεί ότι κατά τη μελέτη μας όλες οι παράμετροι εισόδου πλήν της x διατηρούνται σταθερές. Πρακτικά, το εξαγόμενο κάθε τέτοιας (πειραματικής ή υπολογιστικής) διαδικασίας είναι μια συστοιχία $N+1$ ζευγών της μορφής:



Σχήμα 4.1: $N+1$ δεδομένα σημεία στο επίπεδο (x, y), τα οποία συσχετίζουν τιμές της παραμέτρου εισόδου με τιμές απόκρισης, με διακριτό τρόπο.

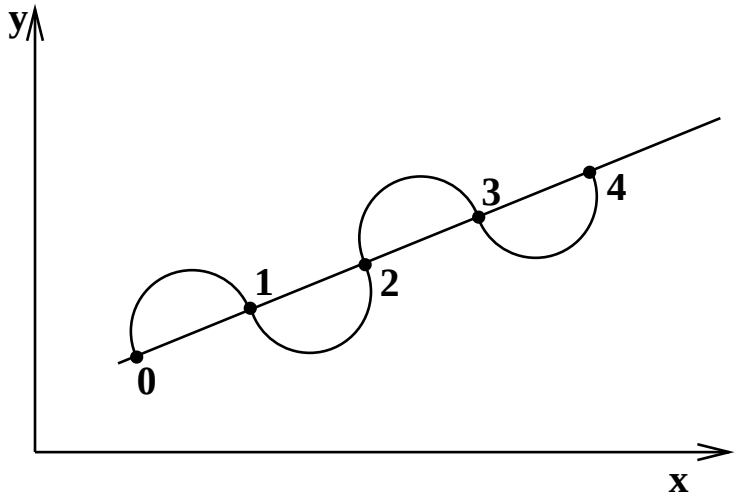
$$\begin{array}{c} (x_0\ ,\ y_0) \\ (x_1\ ,\ y_1) \\ (x_2\ ,\ y_2) \\ \vdots \\ (x_{N-1}\ ,\ y_{N-1}) \\ (x_N\ ,\ y_N) \end{array} \tag{4.1}$$

Χωρίς βλάβη της γενικότητας μπορεί να γίνει η υπόθεση ότι οι $N+1$ τιμές εισόδου είναι διατεταγμένες κατά αύξουσα τιμή εισόδου, $x_0 < x_1 < \cdots < x_{N-1} < x_N$. Μιά γραφική απεικόνιση των αντιστοιχίσεων, όπως αυτή του Σχήματος 4.1, οπτικοποιεί το αποτέλεσμα των πειραμάτων ή των υπολογισμών, πάνω στο οποίο θα στηριχθεί η παρεμβολή ή η προσέγγιση.

Καθώς η απόκριση y είναι μονότιμη συνάρτηση του x, είναι εύλογο να θεωρήσουμε ότι η αντιστοίχιση εισόδων και αποκρίσεων που παριστάνεται στο Σχήμα 4.1 υπονοεί την ύπαρξη μιας συνάρτησης $f(x)$ τέτοιας ώστε για τις αποκρίσεις στα $N+1$ δεδομένα σημεία να ισχύει

$$y_i = f(x_i) \quad , \quad i = 0, \ldots, N \tag{4.2}$$

Η γνώση της αναλυτικής έκφρασης της συνάρτησης $f(x)$,αν βέβαια αυτή υπάρχει, θα ήταν εκτός από επιθυμητή και ο τελικός στόχος κάθε πειράματος ή υπολογισμού. Ας φανταστούμε λ.χ. πόσο απλό και βολικό θα ήταν να μετράμε τη δύναμη που ασκεί ένα σταθερό ηλεκτρικό φορτίο σε ένα άλλο που το τοποθετούμε σε $N+1$ διαφορετικές αποστάσεις από αυτό και, με επεξεργασία των μετρήσεων, να καταλήγουμε στην ακριβή έκφραση του νόμου του Coulomb για το στατικό ηλεκτρισμό. Στην πράξη όμως αυτό δεν είναι δυνατό για πολλούς λόγους. Ένας λόγος είναι ότι το σφάλμα των μετρήσεων ή το σφάλμα στρογγυλοποίησης–αποκοπής στον ηλεκτρονικό υπολογιστή υπεισέρχονται



Σχήμα 4.2: Δύο από τις (πολλές) καμπύλες που είναι δυνατό να δημιουργηθούν χρησιμοποιώντας $N+1$ δεδομένα σημεία στο επίπεδο (x, y).

πάντοτε στις τιμές των αποκρίσεων που καταγράφονται σε πίνακες τιμών της μορφής 4.1. Ένας άλλος λόγος είναι ότι οι διαθέσιμες αντιστοιχίσεις τιμών εισόδου– τιμών απόκρισης, τα διαθέσιμα σημεία δηλαδή στο διάγραμμα, δεν είναι βέβαιο ότι αντιστοιχούν σε μια και μόνο αναλυτική συνάρτηση $f(x)$. Απλό αλλά χαρακτηριστικό είναι το παράδειγμα του Σχήματος 4.2. Τα πέντε διαθέσιμα σημεία $(x_0, y_0), \ldots, (x_4, y_4)$ που 'από σύμπτωση' είναι συνευθειακά και ισαπέχουν διαδοχικά θα μπορούσαν να υποκρύπτουν μια γραμμική συνάρτηση $f(x)$ αλλά και μια ημιτονοειδή συνάρτηση $f(x)$. Και οι δύο σχεδιάζονται στο Σχήμα 4.2. με τα δεδομένα σημεία να βρίσκονται στις τομές τους. Το θέμα επιλογής μιας από τις δύο (ή γιατί όχι και περισσότερων) συναρτήσεων δεν είναι απλό να αναλυθεί σε αυτό το σημείο. Λογική αντίδραση του επιστήμονα που το αντιμετωπίζει είναι να αναζητήσει περισσότερα και ίσως μη-ισαπέχοντα σημεία ή να προβεί σε θεωρητική μοντελοποίηση του σχετικού φυσικού προβλήματος, με όποιες και όσες απλουστευτικές παραδοχές απαιτούνται, ώστε να έχει αντίληψη της μορφής της συνάρτησης–απόκρισης που αναμένεται (γραμμική, παραβολή, εκθετική, ημιτονοειδής, κλπ). Με τον ένα ή τον άλλο τρόπο, αυτό που απαιτείται ώστε να λυθεί το παραπάνω πρόβλημα είναι 'πρόσθετη πληροφορία'.

Η ακριβής αναλυτική συνάρτηση $f(x)$ που αντιστοιχεί στα $N+1$ δεδομένα σημεία του Σχήματος 4.1 έχει λοιπόν, ούτως ή άλλως, δυσκολίες να υπολογισθεί. Αντ' αυτής, αυτό που μπορεί να γίνει (και γίνεται) είναι να προσεγγισθεί η συνάρτηση $f(x)$ με μια άλλη 'κατάλληλη' συνάρτηση $g(x)$. Το ότι η συνάρτηση $g(x)$ προσεγγίζει την $f(x)$ θα συμβολίζεται με

$$f(x) \doteq g(x) \tag{4.3}$$

Η *προσέγγιση* (approximation) είναι βασικό εργαλείο του μηχανικού. Ο βασικότερος λόγος που χρειάζεται η προσέγγιση είναι για να αντιμετωπισθεί το πρόβλημα του υπολογισμού της τιμής της απόκρισης σε μια νέα τιμή εισόδου, διαφορετική από τις διαθέσιμες $N+1$ τιμές. Η χρήση των $N+1$ ζευγών τιμών του Πίνακα 4.1 για τη δημιουργία μιας

προσέγγισης της $f(x)$, της $g(x)$, επιτρέπει στη συνέχεια να εκτιμηθεί άμεσα η τιμή $g(x)$, δηλαδή να προσεγγισθεί η $f(x)$ για κάθε $x \neq x_i$, $i = 0, \ldots, N$.

Η εύρεση της αναλυτικής συνάρτησης $g(x)$ που προσεγγίζει την $f(x)$ εξαρτάται από πολλούς παράγοντες. Συνοψίζουμε τους κυριότερους :

1. Τη γνώση χαρακτηριστικών της συνάρτησης $f(x)$. Για παράδειγμα, αν ένα θεωρητικό μοντέλο υποδεικνύει ότι η συνάρτηση $f(x)$ συμπεριφέρεται ως μια κυβική ή μια εκθετική συνάρτηση του x, είναι λογικό να χρησιμοποιηθούν αντίστοιχα πολυώνυμα τρίτου βαθμού ή εκθετικές συναρτήσεις για τη δημιουργία του $g(x)$.

2. Το πλήθος αλλά και τον τρόπο κτήσης και την ακρίβεια των διαθεσίμων $N + 1$ ζευγών $(x_i, f(x_i))$, $i = 0, \ldots, N$. Για παράδειγμα, είναι λανθασμένο να διαχειριζόμαστε την προσεγγιστική συνάρτηση με ακρίβεια 4 δεκαδικών ψηφίων όταν οι διαθέσιμες τιμές $y_i = f(x_i)$, $i = 0, \ldots, N$ με βάση τις οποίες δημιουργήθηκε έχουν μετρηθεί με πειραματικές διαδικασίες που επιτρέπουν ακρίβεια μόνο μέχρι το δεύτερο δεκαδικό ψηφίο!

3. Το λόγο δημιουργίας και τον τρόπο χρήσης της προσεγγιστικής συνάρτησης $g(x)$, στοιχεία που προφανώς καθορίζουν και τις απαιτήσεις λειότητας και συνέχειας της συνάρτησης αυτής και των παραγώγων της.

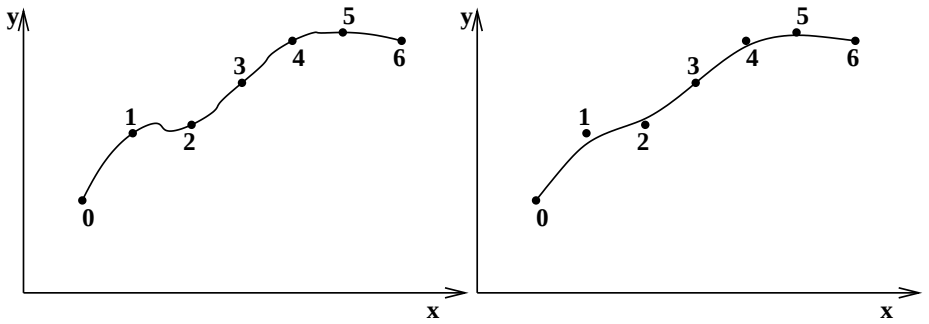
4. Τις απαιτήσεις ακριβείας ως προς την προσέγγιση της $f(x)$.

Σχετικά με το τελευταίο σημείο, θα παρατηρήσουμε ότι όταν δε γνωρίζουμε την αναλυτική έκφραση της $f(x)$ δεν υπάρχει προφανώς τρόπος να υπολογίσουμε την ακριβή τιμή του σφάλματος που προκαλεί η προσέγγιση - αντικατάσταση της $f(x)$ με τη $g(x)$, για κάθε τιμή της παραμέτρου εισόδου x. Όμως, είναι εύκολο να βρεθεί μια εκτίμηση του σφάλματος (σε επίπεδο 'τάξης μεγέθουσ') κάνοντας απλές υποθέσεις, όπως λ.χ. ότι η $f(x)$ είναι λεία ή ότι οι υψηλότερες τάξης παράγωγοί της μπορούν να αμεληθούν κλπ.

Γενικά, η διαχείριση $N + 1$ δεδομένων σημείων (x_i, y_i) , $i = 0, \ldots, N$ με διαφορετικούς τρόπους θα δώσει διαφορετικές συναρτήσεις $g(x)$ που προσεγγίζουν την $f(x)$. Θα διακρίνουμε δύο γενικές κλάσεις μεθόδων που καταλήγουν στον υπολογισμό μιας συνάρτησης $g(x)$, $f(x) \doteq g(x)$, ανάλογα με με το ποιά από τις παρακάτω ιδιότητες ικανοποιούν:

1. Να ισχύει ότι $g(x_i) = y_i$, $i = 0, \ldots, N$, δηλαδή σχεδιάζοντας τη συνάρτηση $g(x)$, αυτή να διέρχεται από τα $N + 1$ σημεία (βλ. Σχήμα 4.3, αριστερά). Στην περίπτωση αυτή αναφερόμαστε σε *συναρτήσεις παρεμβολής* (interpolation functions).

2. Η συνάρτηση $g(x)$ να μην διέρχεται υποχρεωτικά από τα γνωστά σημεία αλλά η μορφή της να αναπαριστά την 'τάση' της $f(x)$ με τη μικρότερη δυνατή απόκλιση από τα δεδομένα σημεία. Πρόκειται για τυπική διαχείριση σημείων που προέκυψαν από πειραματικές μετρήσεις με σφάλματα, και στην περίπτωση αυτή αναφερόμαστε σε *συναρτήσεις προσέγγισης* (approximating functions) ή *βέλτιστης προσαρμογής* (best fit).



Σχήμα 4.3: Παράδειγμα παρεμβολής (αριστερά) και προσέγγισης (δεξιά) των ίδιων $N+1$ σημείων.

Οι δύο αυστηροί και διακριτοί ορισμοί που μόλις δόθηκαν καλύφθηκαν στα όσα προηγήθηκαν αυτών με το γενικό όρο *προσέγγιση*. Στη συνέχεια όμως του κεφαλαίου αυτού οι όροι *παρεμβολή* και *προσέγγιση* θα χρησιμοποιούνται διακριτά και με βάση τους παραπάνω ορισμούς.

Ολοκληρώνοντας την εισαγωγική ενότητα του τρέχοντος κεφαλαίου, θα αναφερθούν ακόμα δύο διαφορετικά παραδείγματα όπου βρίσκει εφαρμογή η προσέγγιση και η παρεμβολή.

- Ο υπολογισμός συναρτήσεων όπως των $ln(x)$, $sin(x)$ κλπ. που ως γνωστό δεν μπορεί να γίνει με ευθείες αριθμητικές πράξεις, βασίζεται στην εύρεση και χρήση προσεγγιστικών συναρτήσεων όπως λ.χ. δυναμοσειρές, όπου κατά τον υπολογισμό της τιμής απόκρισης αποκόπτονται οι λιγότερο σημαντικοί όροι.

- Μηχανικοί σχεδιάζουν μορφές (διδιάστατες ή τριδιάστατες) χρησιμοποιώντας ένα μικρό αριθμό σημείων κερδίζοντας σε ευελιξία και αυξάνοντας τον έλεγχο που έχουν στο αποτέλεσμα τους. Το περίγραμμα διαφόρων συνιστωσών ενός αυτοκινήτου ή ενός αεροσκάφους (κυρίως τα τμήματα από τη μορφή των οποίων εξαρτάται η καλή αεροδυναμική συμπεριφορά ενός τέτοιου οχήματος) σχεδιάζεται συχνά με πολυώνυμα Bezier, τα οποία θα παρουσιασθούν σε επόμενη ενότητα του κεφαλαίου αυτού.

4.1.2 Η Καμπύλη στο Επίπεδο – Τρόποι Περιγραφής

Όπως προαναφέρθηκε, στο κεφάλαιο αυτό θα ασχοληθούμε ιδιαίτερα με τις καμπύλες στο επίπεδο. Οι έννοιες των διακριτών σημείων (x_i, y_i) και των συναρτήσεων $f(x)$ και $g(x)$, έτσι όπως θα χρησιμοποιηθούν στο κεφάλαιο αυτό, δόθηκαν στις σχέσεις 4.2 και 4.3. και στις σχετικές παρατηρήσεις. Η ενότητα αυτή έχει στόχο να παρουσιάσει μια πρώτη, γενική, αλλά με ιδιαίτερη πρακτική σημασία, κατηγοριοποίηση των τρόπων περιγραφής διδιάστατων καμπυλών. Για το σκοπό αυτό, στην ενότητα αυτή θα παρακάμψουμε τους ορισμούς των $f(x)$ και $g(x)$ και θα αναφερόμαστε στη συσχέτιση εισόδου x και απόκρισης y.

Στο επίπεδο (x, y), μια καμπύλη θα περιγράφεται είτε με τις Καρτεσιανές της εξισώσεις

$$F(x,y) \;=\; 0 \tag{4.4}$$

είτε *παραμετρικά*, με τη βοήθεια της παραμέτρου u, ως

$$x = X(u) \qquad , \qquad y = Y(u) \tag{4.5}$$

Είναι σημαντικό να γίνουν εξ αρχής κατανοητές οι δυνατότητες αλλά και οι περιορισμοί που προσφέρουν οι εκφράσεις 4.4 και 4.5. Στη δεύτερη έκφραση, η οποία χρησιμοποιεί την *παραμετροποίηση* (parameterization) της καμπύλης, οι πληροφορίες που εμπεριέχονται είναι σαφώς περισσότερες, αφού η 4.5 περιγράφει μεταξύ άλλων και το ρυθμό εξέλιξης της μορφής της καμπύλης καθώς μεταβάλλεται το u. Η παράμετρος u ορίζεται σε ένα κλειστό διάστημα του $\Re$, συνήθως (αλλά όχι υποχρεωτικά) στο $[0,1]$ και γι' αυτό η 4.5 παριστάνει μια απεικόνιση $\Re \rightarrow \Re^2$. Όμως, το ότι η έκφραση 4.5 είναι περισσότερο πλούσια σε πληροφορία από την 4.4 δεν σημαίνει ότι, σε όλες τις περιπτώσεις, είναι και η περισσότερο βολική για χρήση. Πέραν των ίδιων των μεθοδολογιών παρεμβολής και προσέγγισης, ο μηχανικός οφείλει να αποκτήσει και την 'αντίληψη' για το ποιά από τις δύο εκφράσεις είναι η βολικότερη σε κάθε πρόβλημά του.

Για την κατανόηση των διαφορών των εκφράσεων 4.4 και 4.5 θα τις χρησιμοποιήσουμε στο παράδειγμα της έλλειψης. Μια έλλειψη στο επίπεδο (x,y) περιγράφεται είτε ως

$$F(x,y) \;=\; \frac{x^2}{a^2} \;+\; \frac{y^2}{b^2} \;=\; 1 \tag{4.6}$$

είτε, παραμετρικά, ως

$$x \;=\; a\,cosu \qquad , \qquad y \;=\; b\,sinu \qquad\qquad u \in [0, 2\pi) \tag{4.7}$$

Σημειώνουμε ότι ο δεύτερος τρόπος περιγραφής υπονοεί άμεσα και τον τρόπο διαγραφής της έλλειψης, με φορά δηλαδή που είναι αντίθετη των δεικτών του ρολογιού. Από την άλλη πλευρά, η παραμετρική έκφραση 4.7 δεν είναι μοναδική. Μπορούμε λ.χ. εναλλακτικά να προτείνουμε την έκφραση

$$x \;=\; a\,sinu \qquad , \qquad y \;=\; b\,cosu \qquad\qquad u \in [0, 2\pi) \tag{4.8}$$

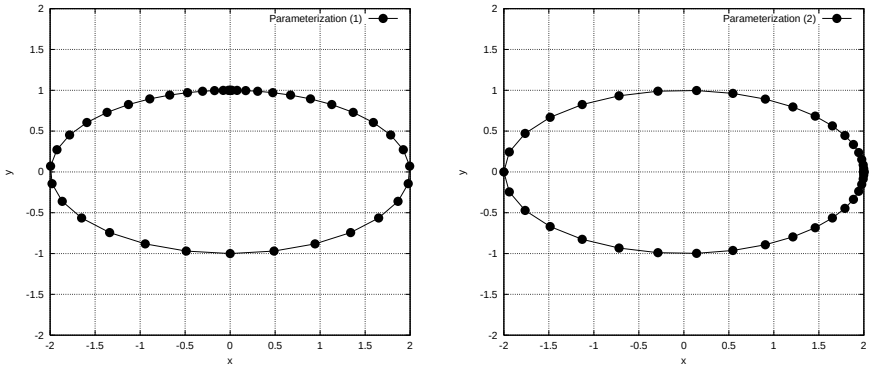
Οι 4.7 και 4.8 παριστούν την ίδια έλλειψη αλλά με διαφορετική αφετηρία (διαφορετικό σημείο του επιπέδου (x,y) αντιστοιχεί στην τιμή $u = 0$) και διαφορετική φορά διαγραφής της καμπύλης (καθώς αυξάνεται το u).

Διευκρινίζεται ότι στο κεφάλαιο αυτό, το σύμβολο y' θα παριστάνει την παράγωγο

$$y' = \frac{dy}{dx} \tag{4.9}$$

ενώ με $\dot{x}$ ή $\dot{y}$ θα παριστάνονται οι παράγωγοι

$$\dot{x} = \frac{dx}{du} \qquad , \qquad \dot{y} = \frac{dy}{du} \tag{4.10}$$



Σχήμα 4.4: Παράδειγμα διαφορετικής κατανομής σημείων σε μια έλλειψη, με βάση τις παραμετροποιήσεις 4.7 και 4.8. Έχουν τοποθετηθεί 41 σημεία, που και στις δύο περιπτώσεις αντιστοιχούν στην ίδια κατανομή (συνημιτονοειδή) της παραμέτρου u.

Κατά την παραμετρική περιγραφή μιάς καμπύλης ορίζουμε ως *ταχύτητα παραμετροποίησης* (parameterization speed) την ποσότητα

$$V(u) \;=\; \left|\dot{\overrightarrow{r}}\right| \;=\; \left|\frac{d\overrightarrow{r}}{du}\right| \tag{4.11}$$

Μια παραμετροποίηση $\overrightarrow{r}(u) = (x(u), y(u))$ ονομάζεται *ομαλή* (regular) όταν η ταχύτητα $V(u)$ δεν μηδενίζεται για οποιαδήποτε τιμή του u στο πεδίο ορισμού του. Η φυσική σημασία της ταχύτητας $V(u)$ είναι απλή, ενώ ο έλεγχος αυτής της ποσότητας είναι σημαντικός γιατί έτσι έχουμε εποπτεία ως προς την γένεση και θέση σημείων σε καμπύλες που περιγράφονται από σχέσεις της μορφής 4.7 ή 4.8, για διακριτές τιμές της παραμέτρου u. Για παράδειγμα, οι παραμετροποιήσεις 4.7 και 4.8 για την έλλειψη οδηγούν σε διαφορετικές εκφράσεις για την ταχύτητα παραμετροποίησης, αντίστοιχα

$$V(u) \;=\; \sqrt{a^2\ sin^2u \;+\; b^2\ cos^2u} \tag{4.12}$$

και

$$V(u) \;=\; \sqrt{a^2\ cos^2u \;+\; b^2\ sin^2u} \tag{4.13}$$

Για το λόγο αυτό, όπως φαίνεται και στο Σχήμα 4.4, η γένεση ενός δεδομένου πλήθους σημείων σε μια έλλειψη, σε μια δεδομένη κατανομή τιμών της παραμέτρου u, θα δημιουργήσει διαφορετικές αλληλουχίες σημείων, που θα αντιστοιχούν σε διαφορετικές ταχύτητες παραμετροποίησης.

Τέλος, αν $\overrightarrow{r}(u) = (x(u), y(u))$, $u \in [a, b]$ είναι μια επίπεδη καμπύλη, το μήκος τόξου από την αφετηρία $u = a$ μέχρι την τυχαία τιμή $u \leq b$ θα δίνεται από τη σχέση

$$s(u) \;=\; \int_a^u ds \;=\; \int_a^u \left(dx^2 + dy^2\right)^{1/2} \;=\; \int_a^u \left(\dot{x}^2 + \dot{y}^2\right)^{1/2} du \;=\; \int_a^u v(u)du \tag{4.14}$$

4.2 Αριθμητική Παρεμβολή

4.2.1 Πολυώνυμα Παρεμβολής

Η χρήση πολυωνύμων διαφόρων βαθμών ως εργαλείο παρεμβολής παρουσιάζει ένα πλήθος πλεονεκτημάτων αλλά και ένα κύριο μειονέκτημα. Τα βασικότερα πλεονεκτήματα είναι:

- τα πολυώνυμα έχουν απλές μαθηματικές εκφράσεις,
- ο υπολογισμός των αγνώστων–συντελεστών των πολυωνύμων ακολουθεί εύκολες και προφανείς διαδικασίες,
- η χρήση ενός πολυωνύμου του οποίου έχουν ήδη βρεθεί οι τιμές των συντελεστών του, ώστε να υπολογιστεί η τιμή απόκρισης y που αντιστοιχεί σε μια οποιαδήποτε τιμή εισόδου x, έχει πρακτικά μηδενικό υπολογιστικό κόστος,
- τα πολυώνυμα είναι συνεχώς διαφορίσιμες συναρτήσεις (πρακτικά, όσες φορές χρειαστεί),
- τα πολυώνυμα έχουν 'ελεγχόμενο' σφάλμα αριθμητικής παρεμβολής.

Από την άλλη πλευρά, το κύριο μειονέκτημά τους είναι ότι ανάλογα με τα κομβικά σημεία (x_i, y_i) , $i = 0, \ldots, N$ με τα οποία δημιουργείται το πολυώνυμο και το βαθμό του πολυωνύμου (βλέπε παρακάτω), εύκολα μπορούν να εμφανιστούν ταλαντωτικές συμπεριφορές που τις περισσότερες φορές δεν συμβαδίζουν με τη 'φυσική' του αντίστοιχου προβλήματος.

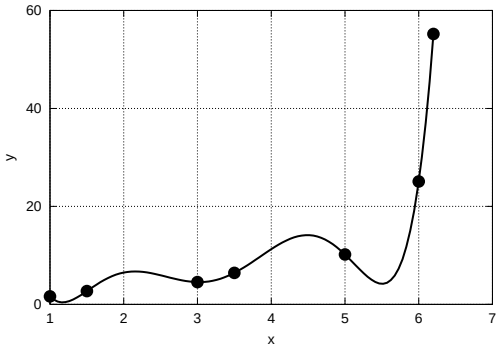
Ένα απλό παράδειγμα παρεμβολής με τη βοήθεια ενός πολυώνυμου έκτου βαθμού μπορεί να βοηθήσει την κατανόηση του μειονεκτήματος που αναφέραμε. Έστω λ.χ. τα παρακάτω 7 σημεία

$$(1.00, 1.64)\ ,\ (1.50, 2.71)\ ,\ (3.00, 4.55)\ ,\ (3.50, 6.43)\ ,$$
$$(5.00, 10.18)\ ,\ (6.00, 25.10)\ ,\ (6.20, 55.20)$$

τα οποία είναι διαθέσιμα ως αποτέλεσμα υπολογισμών ή πειραμάτων και εκφράζουν, με διακριτό τρόπο, την απόκριση μιας συνάρτησης $y = f(x)$ για επτά επιλεγμένες τιμές του x. Ο πιό απλός και προφανής τρόπος είναι να δημιουργηθεί ένα πολυώνυμο παρεμβολής της μορφής

$$g(x) \ = \ a_0 + \ a_1 x + \ a_2 x^2 + \ a_3 x^3 + \ a_4 x^4 + \ a_5 x^5 + \ a_6 x^6$$

που να διέρχεται από τα επτά αυτά σημεία, απαίτηση που καταλήγει στη διατύπωση και επίλυση ενός γραμμικού συστήματος 7×7, για τον υπολογισμό των τιμών των συντελεστών a_i του πολυωνύμου. Το πολυώνυμο που προκύπτει με τον τρόπο αυτό είναι το



Σχήμα 4.5: Γραφική παράσταση του αποτελέσματος μιας πολυωνυμικής παρεμβολής για επτά δεδομένα σημεία, με έμφαση στην ταλαντωτική συμπεριφορά που εμφανίζεται.

$$g(x) \;=\; 182.3503 - 489.5377x + 502.8798x^2 - 251.464x^3 + 65.4774x^4 - 8.4985x^5 + 0.4333x^6$$

του οποίου η γραφική παράσταση δίνεται στο Σχήμα 4.5. Είναι εμφανής η ταλαντωτική συμπεριφορά που παρουσιάζει η γραφική παράσταση του πολυώνυμου παρεμβολής. Πρατηρώντας το Σχήμα 4.5, το εύλογο ερώτημα που τίθεται από τον επιστήμονα που καλείται να το ερμηνεύσει είναι αν η ταλαντωτική συμπεριφορά του πολυωνύμου παρεμβολής αντιστοιχεί (και με τι ακρίβεια) σε μια πραγματική ταλαντωτική συμπεριφορά της απόκρισης του 'φυσικού' προβλήματος ή αν οφείλεται σε εγγενείς μαθηματικές ιδιότητες της συνάρτησης παρεμβολής (στο ότι επιλέχθηκε πολυώνυμο, στο βαθμό αυτού του πολυωνύμου, κλπ).

4.2.2 Ο Βαθμός του Πολυωνύμου

Επιλέγοντας το πολυώνυμο ως εργαλείο παρεμβολής στο πρόβλημά μας, επόμενη βασική επιλογή αποτελεί ο καθορισμός του βαθμού του πολυωνύμου.

Αποδεικνύεται ότι υπάρχει ένα μοναδικό πολυώνυμο, βαθμού N (ή μικρότερου) που παρεμβάλλει τα $N+1$ διακριτά σημεία του Πίνακα 4.1. Η μορφή του πολυωνύμου θα είναι η

$$g(x) \;=\; a_0 + \; a_1 x + \; a_2 x^2 + \; \ldots a_{N-1} x^{N-1} + \; a_N x^N \tag{4.15}$$

και αυτό θα ικανοποιεί $N+1$ εξισώσεις της μορφής

$$y_i \;=\; g(x_i) \qquad , \qquad i = 0, \ldots, N \tag{4.16}$$

οι οποίες οδηγούν και στον υπολογισμό των συντελεστών του. Σε μητρωική γραφή, η εύρεση των συντελεστών a_i απαιτεί την επίλυση του γραμμικού συστήματος

$$C\overrightarrow{a} \;=\; \overrightarrow{y} \tag{4.17}$$

όπου C είναι το τετραγωνικό μητρώο

$$C = \begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^N \\ 1 & x_1 & x_1^2 & \dots & x_1^N \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^N \end{bmatrix} \tag{4.18}$$

και

$$\begin{aligned} \overrightarrow{a} &= (a_0\ ,\ a_1\ ,\ a_2\ ,\ \dots\ ,\ a_N)^T \\ \overrightarrow{y} &= (y_0\ ,\ y_1\ ,\ y_2\ ,\ \dots\ ,\ y_N)^T \end{aligned}$$

Το μητρώο C είναι πάντοτε αναστρέψιμο αφού η ορίζουσά του $detC \neq 0$ (γνωστή και ως ορίζουσα του Vandermonde) όταν $x_i \neq x_j\ ,\ i \neq j$ και συνεπώς το σύστημα 4.17 είναι επιλύσιμο, δηλαδή

$$\overrightarrow{a} = C^{-1}\overrightarrow{y} \tag{4.19}$$

Οι απαιτήσεις σε υπολογιστικά 'εργαλεία' για την εύρεση των τιμών των συντελεστών των πολυωνύμων περιορίζονται σε ένα υποπρόγραμμα (γρήγορης) αντιστροφής ενός μητρώου. Αξίζει, βέβαια, να αναφερθεί η (μη–γραμμική) αύξηση του υπολογιστικού κόστους αν η παρεμβολή γίνεται σε μεγάλο πλήθος δεδομένων σημείων (μεγάλες τιμές του N).

Δεν πρόκειται να ασχοληθούμε, στο σημείο αυτό, με την απόδειξη της μοναδικότητας αυτού του πολυωνύμου. Μια τέτοια απόδειξη συναντάται σε οποιοδήποτε βιβλίο Αριθμητικής Ανάλυσης. Βασίζεται δε στο ότι αν $g(x)$ είναι το πολυώνυμο παρεμβολής και $q(x)$ ένα άλλο πιθανό τέτοιο πολυώνυμο βαθμού $\leq N$ με $y_i = q(x_i)\ ,\ i = 0, \dots, N$, τότε το $g(x)$ θα είναι ή το μηδενικό πολυώνυμο ή ένα πολυώνυμο βαθμού $\geq N + 1$.

4.2.3 Πολυώνυμα Παρεμβολής κατά Lagrange

Η κατά Lagrange πολυωνυμική παρεμβολή πραγματοποιείται κατασκευάζοντας $N + 1$ πολυώνυμα βάσης, τα οποία θα συμβολίζονται με $L_j\ ,\ j = 0, \dots, N$ και θα ικανοποιούν τις βασικές σχέσεις

$$L_j(x_i) = \delta_i^j \qquad ,\ i, j \in [0, N] \tag{4.20}$$

όπου δ_i^j είναι το σύμβολο του Kronecker ($\delta_i^j = 1$ όταν $i = j$, αλλιώς μηδενική τιμή). Ως τέτοια πολυώνυμα επιλέγονται τα

$$L_j = \frac{(x - x_0)(x - x_1)\dots(x - x_{j-1})(x - x_{j+1})\dots(x - x_{N-1})(x - x_N)}{(x_j - x_0)(x_j - x_1)\dots(x_j - x_{j-1})(x_j - x_{j+1})\dots(x_j - x_{N-1})(x_j - x_N)}\ , \quad j = 0, \dots, N \tag{4.21}$$

έτσι ώστε καθένα από αυτά να είναι πολυώνυμο βαθμού N και να ικανοποιεί προφανώς τις σχέσεις 4.20. Το τελικό πολυώνυμο παρεμβολής κατά Lagrange δημιουργείται από το γραμμικό συνδυασμό των πολυωνύμων βάσης L_j , ως εξής

$$g(x) \;=\; \sum_{j=0}^{N} y_j L_j(x) \tag{4.22}$$

Η έκφραση 4.22 αποτελεί τη μαθηματική έκφραση μιας συνάρτησης που παρεμβάλλει (δηλαδή διέρχεται από) τα $N+1$ δεδομένα σημεία, αφού

$$g(x_i) \;=\; \sum_{j=0}^{N} y_j L_j(x_i) \;=\; \sum_{j=0}^{N} y_j \delta_i^j \;=\; y_i \quad , \quad i = 0, \ldots, N \tag{4.23}$$

Πολλές φορές, είναι χρήσιμη η εναλλακτική γραφή των πολυωνύμων L_j στη μορφή

$$L_j \;=\; \frac{\omega(x)}{(x - x_j)\omega'(x_j)} \quad , \quad j = 0, \ldots, N \tag{4.24}$$

όπου

$$\omega(x) = \prod_{k=0}^{N} (x - x_k) \tag{4.25}$$

Εφαρμογή

Έστω τα τρία συνευθειακά σημεία $(0,0)$, $(1,1)$ και $(2,2)$ για τα οποία θέλουμε να υπολογίσουμε το κατά Lagrange πολυώνυμο παρεμβολής. Εφαρμόζοντας τη σχέση 4.21 προκύπτουν διαδοχικά τα

$$\begin{aligned}
L_0(x) &= \frac{(x-1)(x-2)}{(0-1)(0-2)} = \frac{1}{2}(x-1)(x-2) \\
L_1(x) &= \frac{(x-0)(x-2)}{(1-0)(1-2)} = -x(x-2) \\
L_2(x) &= \frac{(x-0)(x-1)}{(2-0)(2-1)} = \frac{1}{2}x(x-1)
\end{aligned}$$

οπότε, το κατά Lagrange πολυώνυμο παρεμβολής, 4.22 είναι το

$$\begin{aligned}
g(x) &= y_0 L_0(x) + y_1 L_1(x) + y_2 L_2(x) \\
&= 0 \cdot \left(\frac{1}{2}(x-1)(x-2) \right) + 1 \cdot (-x(x-2)) + 2 \cdot \left(\frac{1}{2} x(x-1) \right) \\
&= -x^2 + 2x + x^2 - x \\
&= x
\end{aligned}$$

Όπως δηλαδή αναμένονταν, τα τρία συνευθειακά σημεία οδηγούν σε πολυώνυμο παρεμβολής που παριστά την εξίσωση της ευθείας που διέρχεται από αυτά.

4.2.4 Πολυώνυμα Παρεμβολής κατά Hermite

Τα πολυώνυμα παρεμβολής κατά Hermite χρησιμοποιούνται στις περιπτώσεις εκείνες που, εκτός από τις συντεταγμένες των δεδομένων $N+1$ κομβικών σημείων (x_i, y_i) , $i = 0, \ldots, N$, θεωρούνται δεδομένες και επιβάλλονται και οι κλίσεις $y_i^{'} = (dy/dx)_i$ στα ίδια σημεία. Συνεπώς, για κάθε τιμή x_i γνωρίζουμε δύο ποσότητες, το y_i και το $y_i^{'}$. Το ζητούμενο πολυώνυμο θα ικανοποιεί και τα $N+1$ δεδομένα σημεία, αλλά και τις κλίσεις στα σημεία αυτά. Είναι συνεπώς αναμενόμενο ο βαθμός του πολυωνύμου να είναι ίσος με $2(N+1)$.

Αρχικά ορίζονται τα Hermite πολυώνυμα βάσης $H_j(x)$ και $\overline{H}_j(x)$, με τις ιδιότητες

$$\begin{aligned} H_j(x_i) &= \delta_i^j \quad , \quad H_j^{'}(x_i) = 0 \\ \overline{H}_j(x_i) &= 0 \quad , \quad \overline{H}_j^{'}(x_i) = \delta_i^j \quad , \; j = 0, \ldots, N \end{aligned} \tag{4.26}$$

με προφανή ιδέα κάθε πολυώνυμο $H_j(x)$ να συνεισφέρει στην ικανοποίηση της γνωστής τιμής της απόκρισης για καθένα από τα δεδομένα σημεία, χωρίς να εμπλέκεται στην ικανοποίηση της τιμής της κλίσης της, ενώ αντίθετος ακριβώς είναι ο ρόλος των πολυωνύμων $\overline{H}_j(x)$. Με βάση αυτούς στους στόχους, τα πολυώνυμα βάσης της κατά Hermite πολυωνυμικής παρεμβολής ορίζονται ως

$$\begin{aligned} H_j(x) &= \left(1 - 2L_j^{'}(x_j)(x - x_j)\right) \left(L_j(x)\right)^2 \\ \overline{H}_j(x) &= (x - x_j) \left(L_j(x)\right)^2 \quad , \; j = 0, \ldots, N \end{aligned} \tag{4.27}$$

όπου $L_j(x)$ είναι τα γνωστά πολυώνυμα βάσης της κατά Lagrange παρεμβολής.

Η εξίσωση της καμπύλης παρεμβολής κατά Hermite στο Καρτεσιανό επίπεδο είναι και πάλι ένας γραμμικός συνδυασμός των πολυωνύμων βάσης, δηλαδή

$$g(x) = \sum_{j=0}^{N} y_j H_j(x) + \sum_{j=0}^{N} y_j^{'} \overline{H}_j(x) \tag{4.28}$$

Εφαρμογή

Ας επαναλάβουμε την εφαρμογή που ακολούθησε την παρουσίαση της κατά Lagrange πολυωνυμικής παρεμβολής, χρησιμοποιώντας τώρα πολυώνυμα παρεμβολής κατά Hermite. Εκτός από τα τρία σημεία $(0,0)$, $(1,1)$ και $(2,2)$ δίνονται και οι κλίσεις $y^{'}(x)$ σε αυτά, που ας έχουν την κοινή τιμή 1. Υπολογίζουμε αρχικά τις παραγώγους των πολυωνύμων βάσης κατά Lagrange

$$\begin{aligned} L_0^{'}(x) &= \frac{d}{dx}\left(\frac{1}{2}(x-1)(x-2)\right) = x - \frac{3}{2} \\ L_1^{'}(x) &= \frac{d}{dx}\left(-x(x-2)\right) = -2x + 2 \\ L_2^{'}(x) &= \frac{d}{dx}\left(\frac{1}{2}x(x-1)\right) = x - \frac{1}{2} \end{aligned}$$

οπότε είναι

$$\begin{aligned} L_0^{'}(x_0) &= L_0^{'}(0) = -\frac{3}{2} \\ L_1^{'}(x_1) &= L_1^{'}(1) = 0 \\ L_2^{'}(x_2) &= L_2^{'}(2) = \frac{3}{2} \end{aligned}$$

Στη συνέχεια υπολογίζουμε τα πολυώνυμα βάσης κατά Hermite. Είναι

$$\begin{aligned} H_0(x) &= \left[1 - 2\left(-\frac{3}{2}\right)(x-0)\right]\frac{1}{4}(x-1)^2(x-2)^2 = \frac{1}{4}(1+3x)(x-1)^2(x-2)^2 \\ H_1(x) &= [1 - 2\,(0)\,(x-1)]\,x^2(x-2)^2 = x^2(x-2)^2 \\ H_2(x) &= \left[1 - 2\left(\frac{3}{2}\right)(x-2)\right]\frac{1}{4}x^2(x-1)^2 = \frac{1}{4}(7-3x)x^2(x-1)^2 \\ \overline{H}_0(x) &= \frac{1}{4}x(x-1)^2(x-2)^2 \\ \overline{H}_1(x) &= x^2(x-1)(x-2)^2 \\ \overline{H}_2(x) &= \frac{1}{4}x^2(x-1)^2(x-2) \end{aligned}$$

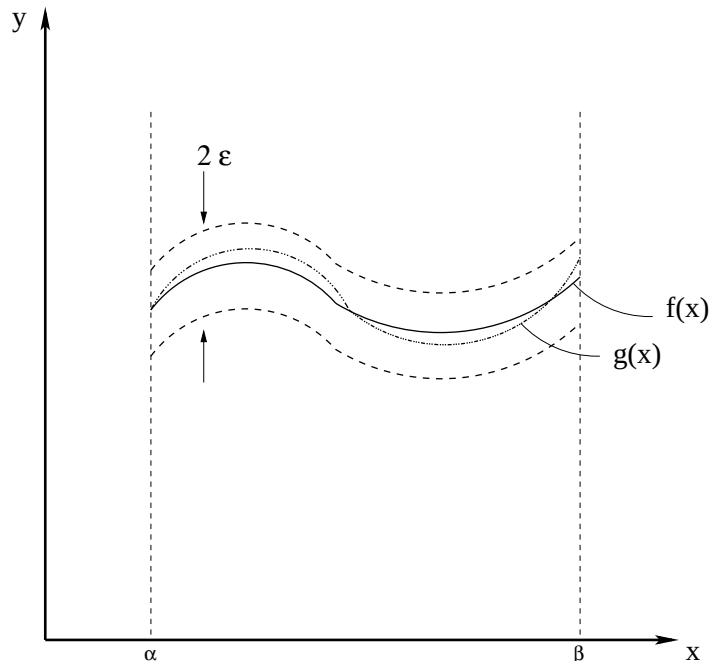
Σύμφωνα με τη σχέση 4.28, το κατά Hermite πολυώνυμο παρεμβολής θα είναι το

$$\begin{aligned} g(x) = \quad & 0 \cdot (1+3x)(x-1)^2(x-2)^2 + 1 \cdot x^2(x-2)^2 + 2 \cdot \frac{1}{4}(7-3x)x^2(x-1)^2 \\ & + 1 \cdot \frac{1}{4}x(x-1)^2(x-2)^2 + 1 \cdot x^2(x-1)(x-2)^2 + 1 \cdot \frac{1}{4}x^2(x-1)^2(x-2) \end{aligned}$$

Η εκτέλεση των πράξεων θα δώσει το αναμενόμενο αποτέλεσμα

$$g(x) = x$$

Το ότι είτε με τα κατά Lagrange ή με τα κατά Hermite πολυώνυμα παρεμβολής, το αποτέλεσμα στην παραπάνω εφαρμογή είναι το ίδιο (όπως άλλωστε θα ήταν αν χρησιμοποιούνταν ένα απλό δευτεροβάθμιο πολυώνυμο παρεμβολής) εξηγείται από το θεώρημα της μοναδικότητας ενός τέτοιου πολυωνύμου που προαναφέραμε. Ας παρατηρήσουμε ακόμη ότι με τη χρήση πολυωνύμων βάσης (Lagrange, Hermite, κλπ) αποφεύγεται μεν η αντιστροφή του μητρώου των συντελεστών (κατά την επίλυση του συστήματος 4.17) χωρίς όμως αυτό να σημαίνει ότι το υπολογιστικό κόστος για τα πολυώνυμα βάσης είναι πάντα αμελητέο.



Σχήμα 4.6: Γεωμετρική ερμηνεία του Θεωρήματος Προσέγγισης του Weierstrass.

4.2.5 Βασικά Θεωρήματα για τα Πολυώνυμα Παρεμβολής

Στην ενότητα αυτή παρατίθενται, χωρίς απόδειξη, δύο βασικά θεωρήματα που διέπουν την αριθμητική παρεμβολή με χρήση πολυωνύμων:

Θεώρημα 4.1 (Θεώρημα Προσέγγισης του Weierstrass:) *Αν f συνεχής συνάρτηση στο διάστημα $[\alpha, \beta]$, τότε για κάθε θετική ποσότητα $\epsilon > 0$ υπάρχει ένα πολυώνυμο $g(x)$ τέτοιο ώστε*

$$\max_{x \in [\alpha,\beta]} |f(x) - g(x)| < \epsilon \tag{4.29}$$

Σχόλια: Η πρακτική αξία του θεωρήματος 4.1 δίνεται εποπτικά στο Σχήμα 4.6, όπου με συνεχή γραμμή σχεδιάζεται η μορφή μιας συνάρτησης $f(x)$ στο διάστημα $[\alpha, \beta]$. Εκατέρωθεν της καμπύλης που αντιστοιχεί στη συνάρτηση ορίζεται μια ζώνη που καλύπτει εύρος από $f(x) - \epsilon$ ως $f(x) + \epsilon$ και παρουσιάζεται με δύο εστιγμένες γραμμές. Το θεώρημα 4.1 εξασφαλίζει ότι υπάρχει ένα πολυώνυμο $g(x)$, βαθμού μη-προσδιοριζόμενου από το θεώρημα αυτό, που προσεγγίζει την $f(x)$ στο $[\alpha, \beta]$ και κείται μέσα στο εύρος που καθορίζουν οι δύο εστιγμένες γραμμές.

Θεώρημα 4.2 *Έστω f συνεχής συνάρτηση στο διάστημα $[\alpha, \beta]$ και $N+1$ διακριτά σημεία ορισμένα στο ίδιο κλειστό διάστημα. Ας συμβολίσουμε με $g(x)$ το (μοναδικό) πολυώνυμο παρεμβολής για το οποίο ισχύει ότι $y_i = f(x_i) = g(x_i)$, $i = 0, \ldots, N$. Τότε, για κάθε $x \in [\alpha, \beta]$ υπάρχει ένα $\xi \in (\alpha, \beta)$ τέτοιο ώστε*

$$E(x) \;=\; f(x) - g(x) = \frac{f^{(N+1)}(\xi)}{(N+1)!}(x-x_0)(x-x_1)\ldots(x-x_N) \tag{4.30}$$

Σχόλια: Η πρακτική αξία του θεωρήματος 4.2 είναι δύσκολο να αναλυθεί, δεδομένου ότι η αναζήτηση του πολυωνύμου παρεμβολής $g(x)$ γίνεται χωρίς προφανώς να γνωρίζουμε την αναλυτική έκφραση της $f(x)$, άρα ο υπολογισμός της $(N+1)$-ιοστής παραγώγου της δεν είναι εφικτός. Ας τονίσουμε, παρόλα αυτά, ότι η μορφή του σφάλματος στην εξίσωση 4.30 εξασφαλίζει το μηδενισμό του στα $N+1$ δεδομένα σημεία, δηλαδή

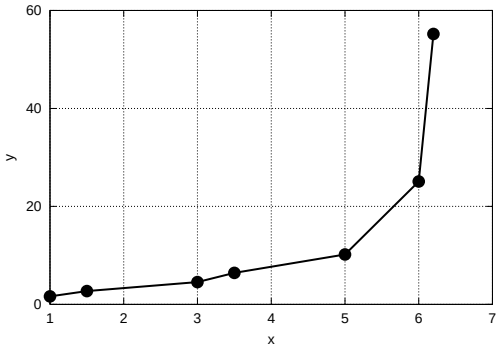
$$E(x) \;=\; 0 \qquad , \quad j = 0, \ldots, N \tag{4.31}$$

Επίσης, αν τα $N+1$ δεδομένα σημεία αντιστοιχούν σε ένα πολυώνυμο βαθμού N (δύο δεδομένα σημεία που καθορίζουν ένα πολυώνυμο πρώτου βαθμού, δηλαδή μια ευθεία, τρία δεδομένα σημεία που καθορίζουν ένα πολυώνυμο δεύτερου βαθμού, κλπ), τότε η $(N+1)$-ιοστή παράγωγος θα είναι πάντοτε μηδενική, άρα $E(x) = 0$. Το αποτέλεσμα αυτό είναι αναμενόμενο αφού το πολυώνυμο $g(x)$ που θα προκύψει θα ταυτίζεται με την πολυωνυμική συνάρτηση $f(x)$ που δημιουργήθηκε από τα $N+1$ σημεία.

4.3 Αριθμητική Παρεμβολή με Τμηματικά Συνεχή Πολυώνυμα

Μέχρι τώρα, η πολυωνυμική παρεμβολή στηρίχθηκε στη δημιουργία ενιαίου πολυωνύμου που ικανοποιούσε τη δεδομένη θέση των $N+1$ σημείων στο επίπεδο (x, y) και, ενδεχομένως, δεδομένες κλίσεις σε αυτά. Εναλλακτικά, είναι δυνατό η παρεμβολή να στηριχθεί και πάλι σε πολυώνυμα, χωρίς όμως το ίδιο το πολυώνυμο να 'ικανοποιεί' και τα $N+1$ δεδομένα σημεία ή τις κλίσεις σε αυτά, αλλά μόνο ένα μικρό υποσύνολό τους. Στην περίπτωση αυτή, η αριθμητική παρεμβολή χρησιμοποιεί τμηματικά συνεχή πολυώνυμα και, γραφικά, η τελική καμπύλη παρεμβολής θα αποτελεί τη σύνθεση διαδοχικών καμπυλών. Καθεμιά από τις καμπύλες αυτές αντιπροσωπεύει χαμηλού βαθμού (άρα εύχρηστα και οικονομικά σε υπολογιστικό κόστος) πολυώνυμα, με συμπληρωματικά πεδία ορισμού. Το πεδίο ορισμού κάθε πολυωνύμου είναι ένα ή περισσότερα διαδοχικά υποδιαστήματα, από την αλληλουχία των N υποδιαστημάτων που ορίζουν τα $N+1$ δεδομένα σημεία. Ο πιο προφανής αντιπρόσωπος της κατηγορίας αυτής είναι η τμηματικά συνεχής γραμμική παρεμβολή, όπου τα $N+1$ κομβικά σημεία ενώνονται με διαδοχικά ευθύγραμμα τμήματα. Αυτή απεικονίζεται στο Σχήμα 4.7. Στην περίπτωση αυτή, το πεδίο ορισμού οποιουδήποτε από τα επιμέρους πολυώνυμα είναι το κλειστό διάστημα μεταξύ των τετμημένων δύο διαδοχικών από τα $N+1$ δεδομένα σημεία. Η συνέχεια τιμής, αλλά όχι παραγώγων, στα κομβικά σημεία είναι το προφανές κόστος που συνεπάγεται η ιδιαίτερη απλότητα αυτής της παρεμβολής.

Σε όσα θα ακολουθήσουν για τα τμηματικά συνεχή πολυώνυμα, είναι χρήσιμο να δίνεται προσοχή στο πλήθος των διαδοχικών κομβικών σημείων που οριοθετούν το



Σχήμα 4.7: Γραφική παράσταση του αποτελέσματος μιας παρεμβολής με τμηματικά συνεχή πολυώνυμα για επτά δεδομένα σημεία.

πεδίο ορισμού του κάθε πολυωνύμου και στις συνθήκες συνέχειας (σίγουρα της τιμής, ενδεχόμενα και παραγώγων μέχρι κάποιου βαθμού) στα σημεία σύνδεσης των επιμέρους πολυωνύμων.

4.3.1 Βασικά Σχήματα

Πέραν της τμηματικά συνεχούς γραμμικής παρεμβολής, με το πολύ περιορισμένο ενδιαφέρον, η κατά τμήματα συνεχής παρεμβολή σε υψηλότερου βαθμού πολυώνυμα πραγματοποιείται συνήθως σε παραμετρική γραφή, δηλαδή με την εισαγωγή της παραμέτρου u, στην οποία αναφερθήκαμε στην αρχή του κεφαλαίου. Έτσι, προκύπτει η ανάγκη να γραφούν τα γνωστά πολυώνυμα παρεμβολής, όπως Lagrange ή Hermite ως συνάρτηση μιας παραμέτρου u που ορίζεται στο διάστημα $[0, 1]$. Σημαντικό είναι να καθοριστεί η αντιστοίχιση μεταξύ τιμών του u και των δεδομένων σημείων. Αν κάθε επιμέρους πολυώνυμο ορίζεται μεταξύ δύο διαδοχικών, έστω των x_i και x_{i+1}, είναι προφανές να αντιστοιχίσουμε την τιμή $u = 0$ στο x_i και την $u = 1$ στο x_{i+1}. Αντίστοιχος επανακαθορισμός των θέσεων $u = 0$ και $u = 1$ γίνεται στο επόμενο διάστημα $[x_{i+1}, x_{i+2}]$, κ.ο.κ. Όπως όμως θα δούμε και σε συγκεκριμένα κατά τμήματα συνεχή πολυώνυμα που θα παρουσιασθούν στη συνέχεια, το πεδίο ορισμού μπορεί να αποτελείται λ.χ. από τρία σημεία, να είναι δηλαδή το $[x_{i-1}, x_{i+1}]$. Σε μια τέτοια περίπτωση, θα αντιστοιχίσουμε την τιμή $u = 0$ στο x_{i-1}, την $u = 1$ στο x_{i+1} και την $u = 1/2$ στο x_i.

Παρακάτω θα ορισθούν ορισμένα βασικά σχήματα παρεμβολής με τμηματικά συνεχή πολυώνυμα. Προηγουμένως, πρέπει να εκφρασθούν τα γνωστά πολυώνυμα βάσης στον παραμετρικό χώρο. Έτσι έχουμε:

(α) *Γραμμικά πολυώνυμα Lagrange*, που ορίζονται ως

$$L_0(u) \;=\; 1 - u \qquad , \qquad L_1(u) \;=\; u \tag{4.32}$$

με τις παραπάνω εκφράσεις να μπορούν εύκολα να αποδειχθούν αφού, παρεμβάλλοντας στα $u = 0$ και $u = 1$, παίρνουμε

$$L_0(u) \;=\; \frac{u-1}{0-1} \quad , \quad L_1(u) \;=\; \frac{u-0}{1-0}$$

(β) *Τετραγωνικά πολυώνυμα Lagrange*, που ορίζονται ως

$$L_0(u) \;=\; 2u^2-3u+1 \quad , \quad L_1(u) \;=\; -4u^2+4u \quad , \quad L_2(u) \;=\; 2u^2-u \tag{4.33}$$

και αποδεικνύονται γράφοντας τις εκφράσεις των πολυωνύμων Lagrange για παρεμβολή στα τρία σημεία $u = 0$, $u = 1/2$ και $u = 1$, δηλαδή

$$\begin{aligned} L_0(u) &= \frac{(u-\frac{1}{2})(u-1)}{(0-\frac{1}{2})(0-1)} \\ L_1(u) &= \frac{(u-0)(u-1)}{(\frac{1}{2}-0)(\frac{1}{2}-1)} \\ L_2(u) &= \frac{(u-0)(u-\frac{1}{2})}{(1-0)(1-\frac{1}{2})} \end{aligned}$$

(γ) *Κυβικά πολυώνυμα Hermite*, που ορίζονται ως

$$\begin{aligned} H_0(u) \;&=\; 2u^3-3u^2+1 \quad , \quad H_1(u) \;=\; -2u^3+3u^2 \\ \overline{H}_0(u) \;&=\; u^3-2u^2+u \quad , \quad \overline{H}_1(u) \;=\; u^3-u^2 \end{aligned} \tag{4.34}$$

Οι προηγούμενες σχέσεις αποδεικνύονται γράφοντας για τα σημεία $u = 0$ και $u = 1$ τις εκφράσεις 4.27, χρησιμοποιώντας βοηθητικά και τα πολυώνυμα βάσης Lagrange της σχέσης 4.32. Έτσι, για παράδειγμα, είναι

$$\begin{aligned} H_0(u) &= \left[1-2L_0'(0)u\right](L_0(u))^2 \;=\; (1+2u)(1-u)^2 \\ H_1(u) &= \left[1-2L_1'(1)(u-1)\right](L_1(u))^2 \;=\; (1-2u+2)u^2 \;=\; (3-2u)u^2 \\ \overline{H}_0(u) &= (u-0)\,(L_0(u))^2 = u(1-u)^2 = u-2u^2+u^3 \\ \overline{H}_1(u) &= (u-1)\,(L_1(u))^2 = (u-1)u^2 = u^3-u^2 \end{aligned}$$

Χρησιμοποιώντας τα πολυώνυμα βάσης των σχέσεων 4.32, 4.33 ή 4.34 ως 'δομικά στοιχεία' σχηματίζεται η τελική καμπύλη παρεμβολής κατά τμήματα. Έτσι έχουμε:

(α) τη γραμμική παρεμβολή μεταξύ δύο διαδοχικών σημείων (x_i, y_i),(x_{i+1}, y_{i+1}), εκφρασμένη με τα γραμμικά πολυώνυμα Lagrange, δηλαδή

$$\begin{aligned} x(u) &= x_i L_0(u) + x_{i+1} L_1(u) \\ y(u) &= y_i L_0(u) + y_{i+1} L_1(u) \end{aligned} \tag{4.35}$$

Με τον τρόπο αυτό όμως, είναι προφανής η ασυνέχεια της πρώτης παραγώγου σε κάθε κομβικό σημείο. Για παράδειγμα, η (σταθερή) παράγωγος στο διάστημα (x_i, x_{i+1}) θα ισούται με

$$\frac{dy}{dx} = \frac{\frac{dy}{du}}{\frac{dx}{du}} = \frac{y_{i+1} - y_i}{x_{i+1} - x_i} \tag{4.36}$$

ενώ για το επόμενο διάστημα (x_{i+1}, x_{i+2}) η τιμή της πρώτης παραγώγου είναι

$$\frac{dy}{dx} = \frac{\frac{dy}{du}}{\frac{dx}{du}} = \frac{y_{i+2} - y_{i+1}}{x_{i+2} - x_{i+1}} \tag{4.37}$$

(β) την τετραγωνική παρεμβολή κατά Lagrange που απαιτεί τρία διαδοχικά κομβικά σημεία, έστω τα (x_{i-1}, y_{i-1}), (x_i, y_i), (x_{i+1}, y_{i+1}), ώστε

$$\begin{aligned} x(u) &= x_{i-1} L_0(u) + x_i L_1(u) + x_{i+1} L_2(u) \\ y(u) &= y_{i-1} L_0(u) + y_i L_1(u) + y_{i+1} L_2(u) \end{aligned} \tag{4.38}$$

με τα πολυώνυμα $L_i(u)$ να ορίζονται από τις σχέσεις 4.33.

(γ) την κυβική παρεμβολή μεταξύ των σημείων (x_i, y_i) και (x_{i+1}, y_{i+1}) κατά Hermite που βασίζεται στις σχέσεις

$$\begin{aligned} x &= x_i + u\,(x_{i+1} - x_i) \\ y &= y_i H_0(u) + y_{i+1} H_1(u) + \dot{y}_i \overline{H}_0(u) + \dot{y}_{i+1} \overline{H}_1(u) \end{aligned} \tag{4.39}$$

Παρατηρούμε την εμπλοκή στη σχέση 4.39 των παραγώγων του y ως προς τη βοηθητική παράμετρο u, των οποίων η γνώση είναι απαραίτητη για την εφαρμογή του σχήματος. Γενικά, η ανάγκη διαθεσιμότητας παραγώγων $\dot{y}$ (ή αντίστοιχα $\dot{x}$) είναι μειονέκτημα για κάθε μέθοδο παρεμβολής. Ειδικά όμως στην τμηματικά συνεχή κυβική παρεμβολή μεταξύ των δύο διαδοχικών σημείων κατά Hermite, η γραμμική συσχέτιση x και u επιτρέπει, στο διάστημα αυτό, να γράψουμε ότι

$\dot{x} = x_{i+1} - x_i$, να χρησιμοποιήσουμε ότι $y^{'} = \dot{y}/\dot{x}$ και να υπολογισθούν εύκολα οι τιμές $\dot{y}$ στους κόμβους από τις γνωστές παραγώγους $y^{'}$. Επιβάλλοντας τις τιμές της κλίσης $\dot{y}$ (αλλά ουσιαστικά $y^{'}$) στα δεδομένα σημεία εξασφαλίζεται τοπικά η συνέχεια της πρώτης παραγώγου, δηλαδή μπορεί να αρθεί το πρόβλημα της ασυνέχειας που δημιουργεί η κατά τμήματα παρεμβολή Lagrange, τουλάχιστο για την πρώτη παράγωγο.

Βασικό στοιχείο για την απόφαση αν λ.χ. θα χρησιμοποιήσουμε την τετραγωνική παρεμβολή κατά Lagrange ή την κυβική παρεμβολή κατά Hermite είναι η διαθεσιμότητα ή όχι των τιμών της κλίσης $y^{'}$ στους κόμβους. Όμως, είτε με την κατά Lagrange είτε με την κατά Hermite παρεμβολή κατά τμήματα, η δεύτερη παράγωγος (δηλαδή η καμπυλότητα της δημιουργούμενης καμπύλης) είναι ασυνεχής στα κομβικά σημεία. Το τελευταίο πρόβλημα αντιμετωπίζεται με τις κυβικές splines, που θα παρουσιάσουμε στη συνέχεια.

4.3.2 Αριθμητική Παρεμβολή με Κυβικές Splines

Οι κυβικές splines χρησιμοποιούνται σήμερα ευρύτατα ως μέθοδος τμηματικά συνεχούς παρεμβολής. Τα δεδομένα κομβικά σημεία (x_i, y_i) αντιμετωπίζονται κατά ζεύγη (διαδοχικών σημείων) όπου μεταξύ τους εφαρμόζεται κυβική παρεμβολή.

Για παράδειγμα, μεταξύ των διαδοχικών κομβικών σημείων (x_i, y_i) και (x_{i+1}, y_{i+1}) τα πολυώνυμα παρεμβολής για τις συντεταγμένες x και y θα δίνονται παραμετρικά από τις εκφράσεις

$$\begin{aligned} g_x(u) &= a_0 + a_1 u + a_2 u^2 + a_3 u^3 \\ g_y(u) &= b_0 + b_1 u + b_2 u^2 + b_3 u^3 \end{aligned} \tag{4.40}$$

με τους οκτώ συντελεστές $a_0, \ldots, b_3$ να έχουν διαφορετικές τιμές στα επιμέρους διαστήματα. Τονίζεται ότι, παρότι θα μπορούσαμε να χρησιμοποιήσουμε διπλούς δείκτες όπως $a_{0,j}$ αντί a_0 κ.ο.κ, το αποφεύγουμε για λόγους απλότητας. Στη συνέχεια, θα ασχοληθούμε με την παρουσίαση του υπολογισμού των συντελεστών a_j , $j = 0, 3$ για το διάστημα (x_i, y_i) ως (x_{i+1}, y_{i+1}), αφού ο υπολογισμός των b_j είναι ακριβώς όμοιος. Για τον υπολογισμό τους επιβάλλουμε αρχικά τις δύο προφανείς απαιτήσεις:

$$\begin{aligned} g_x(0) &= a_0 &&= x_i \\ g_x(1) &= a_0 + a_1 + a_2 + a_3 &&= x_{i+1} \end{aligned} \tag{4.41}$$

Ας συμβολίσουμε με M_i , $i = 0, \ldots, N$ τις τιμές της δεύτερης παραγώγου

$$M_i = (\ \ddot{g}_x(u))_i = \left(\frac{d^2x}{du^2}\right)_i$$

στα $N + 1$ δεδομένα κομβικά σημεία. Με βάση τη μαθηματική έκφραση των συναρτήσεων 4.40 η δεύτερη παράγωγος της είναι

$$\ddot{g}_x(u) = 2a_2 + 6a_3 u$$

ενώ για τα κομβικά σημεία που βρίσκονται στα άκρα κάθε διαστήματος μπορούμε να γράψουμε ότι

$$\begin{aligned} \ddot{g}_x(0) &= M_i = 2a_2 \\ \ddot{g}_x(1) &= M_{i+1} = 2a_2 + 6a_3 \end{aligned}$$

οπότε $a_2 = M_i/2$ και $6a_3 = M_{i+1} - M_i$, από τις οποίες προκύπτει εύκολα η (γραμμική ως προς u) έκφραση για τη δεύτερη παράγωγο, που είναι η

$$\ddot{g}_x(u) = M_i + (M_{i+1} - M_i)u \tag{4.42}$$

Ολοκληρώνοντας στη συνέχεια δύο φορές την εξίσωση 4.42, προκύπτει τελικά ότι

$$g_x(u) = \frac{1}{2}M_i u^2 + \frac{1}{6}(M_{i+1} - M_i)u^3 + A + Bu \tag{4.43}$$

όπου εμφανίζονται δύο σταθερές ολοκλήρωσης A και B. Ο υπολογισμός των A, B γίνεται απαιτώντας να ικανοποιηθούν οι σχέσεις 4.41, από τις οποίες προκύπτει ότι

$$\begin{aligned} A &= x_i \\ B &= x_{i+1} - x_i - \frac{1}{6}M_{i+1} - \frac{1}{3}M_i \end{aligned}$$

και διατυπώνεται έτσι η τελική μορφή της 4.43, ως

$$g_x(u) = x_i + \left[(x_{i+1} - x_i) - \frac{1}{6}M_{i+1} - \frac{1}{3}M_i\right] u + \frac{1}{2}M_i u^2 + \frac{1}{6}(M_{i+1} - M_i)u^3 \tag{4.44}$$

Η σχέση αυτή αντικαθιστά τη γενική γραφή 4.40, με συντελεστές οι οποίοι πλέον διαθέτουν συγκεκριμένες εκφράσεις.

Με βάση τη σχέση 4.44, η υλοποίηση του σχήματος θα ήταν εφικτή αρκεί να ήταν διαθέσιμες οι τιμές των παραγώγων M_i , $i = 0, \ldots, N$. Είναι εύλογο να συμπεράνει κανείς ότι η διαθεσιμότητα τέτοιας πληροφορίας δεν είναι εύκολη και κάθε μέθοδος που την απαιτεί γίνεται εκ προοιμίου ένα πρακτικά μη-χρήσιμο εργαλείο για το μηχανικό. Η ιδέα για να ξεπεραστεί η ανάγκη για την επιπλέον πληροφορία, που εκ πρώτης όψεως απαιτεί η μέθοδος, είναι αυτή να παραχθεί έμμεσα από λογικές συνθήκες που θα επιβληθούν στα δεδομένα κομβικά σημεία. Μια λογική (και, όπως στη συνέχεια θα αποδειχτεί, πολύ βολική) συνθήκη είναι να απαιτήσουμε ότι τα τμηματικά συνεχή πολυώνυμα έχουν συνεχείς πρώτες παραγώγους στα *εσωτερικά* κομβικά σημεία. Με τον

όρο *εσωτερικά* εννοούμε όλα τα σημεία πλην των ακραίων (x_0, y_0) και (x_{N+1}, y_{N+1}). Ας επιλέξουμε λ.χ. αρχικά το διάστημα $(x_i, y_i) \rightarrow (x_{i+1}, y_{i+1})$, στο οποίο η πρώτη παράγωγος του g_x, σύμφωνα με τη σχέση 4.44, είναι

$$\dot{g}_x(u) \quad = \quad x_{i+1} - x_i - \frac{1}{6}M_{i+1} - \frac{1}{3}M_i + M_i u + \frac{1}{2}(M_{i+1} - M_i)u^2 \qquad (4.45)$$

Όμοια, για το προηγούμενο διάστημα $(x_{i-1}, y_{i-1}) \rightarrow (x_i, y_i)$ ισχύει ότι

$$\dot{g}_x(u) \quad = \quad x_i - x_{i-1} - \frac{1}{6}M_i - \frac{1}{3}M_{i-1} + M_{i-1} u + \frac{1}{2}(M_i - M_{i-1})u^2 \qquad (4.46)$$

Η τιμή της 4.45 για $u = 0$ και της 4.46 για $u = 1$ αντιστοιχούν στο ίδιο σημείο, τον κόμβο i, και ως εκ τούτου πρέπει να ταυτίζονται. Εξισώνοντας προκύπτει ότι

$$M_{i-1} + 4M_i + M_{i+1} \;=\; 6\,(x_{i-1} - 2x_i + x_{i+1}) \qquad (4.47)$$

Εξισώσεις της μορφής της 4.47 διατυπώνονται για όλους τους κόμβους $i = 1, \ldots, N-1$, εκτός δηλαδή των δύο ακραίων, όπως προαναφέρθηκε. Έτσι, απομένουν άλλες δύο εξισώσεις ως προς τα M_i, ώστε να διατυπωθεί ένα γραμμικό σύστημα με $N+1$ εξισώσεις για $N+1$ αγνώστους. Οι υπόλοιπες δύο εξισώσεις αναγκαστικά προκύπτουν επιβάλλοντας δύο οριακές συνθήκες στον πρώτο και τον τελευταίο κόμβο. Υπάρχουν πάνω από μια δυνατές περιπτώσεις, οι οποίες δίνονται στη συνέχεια:

- Να γνωρίζουμε ή έστω να υποθέτουμε ότι οι δεύτερες παράγωγοι στα ακραία κομβικά σημεία είναι μηδενικές. Συνεπώς, οι δύο νέες εξισώσεις που συμπληρώνουν το σύστημα είναι οι

  $$M_0 \;=\; 0 \qquad , \qquad M_N \;=\; 0 \qquad (4.48)$$

  και τότε αναφερόμαστε σε *φυσικές splines*.

- Να είναι γνωστές οι κλίσεις, δηλαδή οι τιμές δηλαδή οι τιμές των πρώτων παραγώγων $\dot{g}_x$ και $\dot{g}_y$ στα κομβικά σημεία (x_0, y_0) και (x_{N+1}, y_{N+1}). Η γνωστή τιμή της $\dot{g}_x$ στο (x_0, y_0) , έστω d_0, συσχετίζεται με τις γειτονικές τιμές των x_i και M_i μέσω της σχέσης

  $$2M_0 + M_1 \;=\; 6\,(x_1 - x_0 - d_0) \qquad (4.49)$$

  η οποία αποδεικνύεται με βάση την έκφραση 4.45. Με όμοιο τρόπο, στο διάστημα $(x_{N-1}, y_{N-1}) \rightarrow (x_N, y_N)$ και για $u = 1$ αποδεικνύεται ότι η αντίστοιχη σχέση της 4.49 είναι η

$$-M_{N-1} + 2M_N \;=\; 6\,(-x_N + x_{N-1} + d_N) \tag{4.50}$$

όπου με d_N συμβολίστηκε η γνωστή τιμή της $\dot{g}_x$ στο (x_N, y_N).

- Να γίνεται η υπόθεση ότι υπάρχει σταθερή κλίση δεύτερης παραγώγου στα άκρα, δηλαδή $M_0 \;=\; M_1$ και $M_N \;=\; M_{N-1}$.

Ανεξάρτητα από την επιλογή των οριακών συνθηκών, το αποτέλεσμα είναι ότι διατυπώνεται ένα σύστημα $N+1$ εξισώσεων με $N+1$ αγνώστους, δηλαδή τα M_i , $i = 0, \ldots, N$. Με βάση τα προηγούμενα, το σύστημα αυτό παίρνει τη μορφή

$$\begin{bmatrix} \gamma_{0,0} & \gamma_{0,1} & & & & & & \\ 1 & 4 & 1 & & & & & \\ & 1 & 4 & 1 & & & & \\ & & & \vdots & & & & \\ & & & & 1 & 4 & 1 & \\ & & & & & \gamma_{N,N-1} & \gamma_{N,N} \end{bmatrix} \begin{bmatrix} M_0 \\ M_1 \\ M_2 \\ \vdots \\ M_{N-1} \\ M_N \end{bmatrix} \;=\; 6 \begin{bmatrix} h_0 \\ x_0 - 2x_1 + x_2 \\ x_1 - 2x_2 + x_3 \\ \vdots \\ x_{N-2} - 2x_{N-1} + x_N \\ h_N \end{bmatrix} \tag{4.51}$$

όπου με $\gamma_{i,j}$ και h_i συμβολίστηκαν όροι που προκύπτουν από την επιλογή των δύο οριακών συνθηκών. Η επίλυση του συστήματος 4.51 απαιτεί ένα 'υποπρόγραμμα' που επιλύει γραμμικά συστήματα με το μητρώο των συντελεστών να έχει τριδιαγώνια μορφή. Έχοντας υπολογίσει τους συντελεστές M_i , $i = 0, \ldots, N$, δημιουργείται η συνάρτηση παρεμβολής με χρήση της εξίσωσης 4.44, τμηματικά, σε κάθε διάστημα μεταξύ δύο διαδοχικών από τα δεδομένα κομβικά σημεία. Η ίδια διαδικασία επαναλαμβάνεται και για τη συνάρτηση $g_y(u)$.

Είναι σαφές ότι οι κυβικές splines είναι μια μέθοδος που διορθώνει αρκετά από τα προβλήματα των κατά Lagrange ή Hermite τμηματικά συνεχών πολυωνυμικών παρεμβολών. Όμως, παρόλα αυτά, παρουσιάζει προβλήματα ταλαντώσεων σε κάθε σημείο στο οποίο η δεύτερη παράγωγος δεν είναι συνεχής. Ας σημειωθεί επιπλέον ένα βασικό χαρακτηριστικό των τμηματικά συνεχών κυβικών splines, ότι κάθε μεταβολή στις οριακές συνθήκες (αλλαγή τιμής των $\gamma_{i,j}$) ή η μετακίνηση της θέσης ενός σημείου (x_i, y_i) απαιτεί εκ νέου επίλυση του συστήματος 4.51 και επανακαθορισμό των τιμών των M_i.

Ολοκληρώνοντας την ενότητα αυτή για την παρεμβολή με τμηματικά συνεχείς κυβικές splines, θα αναφερθούμε στους δύο τρόπους που η μέθοδος αυτή θα ήταν χρήσιμη στο μηχανικό. Μια πρώτη χρήση θα ήταν η δημιουργία περισσότερων σημείων σε μια καμπύλη που περιγράφεται αρχικά με $N+1$ σημεία. Για την υλοποίηση αυτού του στόχου, διαμορφώνεται αρχικά και στη συνέχεια επιλύεται το σύστημα 4.51, υπολογίζονται οι συντελεστές M_i και χρησιμοποιούνται οι εκφράσεις 4.44 (και η αντίστοιχη για το y) ώστε να υποδιαιρεθεί κάθε διάστημα $(x_i, y_i) \rightarrow (x_{i+1}, y_{i+1})$ σε υποδιαστήματα ορίζοντας αυθαίρετη διαμέριση του u, στο διάστημα $[0, 1]$. Μια δεύτερη χρήση θα ήταν, με βάση τα $N+1$ δεδομένα σημεία (x_i, y_i), να βρεθεί η τιμή του y για μια εμβόλιμη τιμή του x. Για να λυθεί αυτό το πρόβλημα, απαιτείται αρχικά η εύρεση του

διαστήματος $(x_i, y_i) \rightarrow (x_{i+1}, y_{i+1})$ με $x_i \leq x < x_{i+1}$, ο υπολογισμός (όπως και πριν) των συντελεστών M_i, η επίλυση της εξίσωσης 4.44 για τη δεδομένη τιμή x ώστε να υπολογισθεί το αντίστοιχο u και η χρήση της αντίστοιχης εξίσωσης για να υπολογισθεί η τιμή του y για την τιμή του u που προέκυψε.

Εφαρμογή

Έστω ότι χρησιμοποιείται αριθμητική παρεμβολή με κυβικές splines για τα επόμενα έξι $(N = 5)$ δεδομένα σημεία:

$$\begin{array}{c}(0\ ,\ 0)\\(1\ ,\ 2)\\(3\ ,\ 3)\\(4\ ,\ 3)\\(6\ ,\ 1)\\(8\ ,\ 1)\end{array}$$

Επιπλέον δίνεται ότι στο αρχικό σημείο ισχύει η παραδοχή των φυσικών splines, ενώ στο τελευταίο σημείο υπάρχει σταθερή κλίση δεύτερης παραγώγου. Ζητείται να υπολογισθούν οι συντεταγμένες του σημείου που παρεμβάλλεται στο μέσο (παραμετρικά) του τρίτου διαστήματος.

Λύση:

Για την εύρεση της τιμής του x, διαμορφώνουμε το σύστημα

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 & 0 \\ 0 & 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} M_0 \\ M_1 \\ M_2 \\ M_3 \\ M_4 \\ M_5 \end{bmatrix} = 6 \begin{bmatrix} 0 \\ 1 \\ -1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

Το σύστημα αυτό επιλύεται εύκολα και υπολογίζονται τα M_i, ως

$$M_0 = 0 \ , \quad M_1 = 2.15094 \ , \quad M_2 = -2.60377 \ ,$$
$$M_3 = 2.26415 \ , \quad M_4 = -0.45283 \ , \quad M_5 = -0.45283$$

Η τιμή του x του σημείου που βρίσκεται στο μέσο (παραμετρικά, $u = 1/2$) του τρίτου διαστήματος προκύπτει από τη σχέση 4.44, για $i = 2$, ως

$$x \ = \ x_2 + \left[(x_3 - x_2) - \frac{1}{6}M_3 - \frac{1}{3}M_2\right](\frac{1}{2}) + \frac{1}{2}M_2(\frac{1}{2})^2 + \frac{1}{6}(M_3 - M_2)(\frac{1}{2})^3 \ = \ 3.521$$

Για την εύρεση της τιμής του y, το παραπάνω σύστημα διαφοροποιείται ως προς το δεύτερο μέλος του και γράφεται

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 & 0 \\ 0 & 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} M_0 \\ M_1 \\ M_2 \\ M_3 \\ M_4 \\ M_5 \end{bmatrix} = 6 \begin{bmatrix} 0 \\ -1 \\ -1 \\ -2 \\ 1 \\ 0 \end{bmatrix}$$

με λύσεις

$$M_0 = 0 \ , \ M_1 = -1.40755 \ , \ M_2 = -0.36981 \ ,$$
$$M_3 = -3.11321 \ , \ M_4 = 0.82264 \ , \ M_5 = 0.82264$$

Η αντίστοιχη τιμή του y στο υπόψη σημείο θα είναι

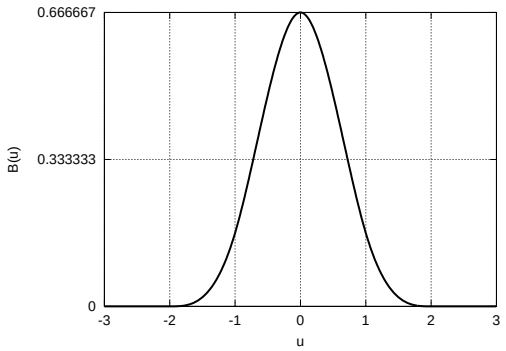
$$y = y_2 + \left[(y_3 - y_2) - \frac{1}{6}M_3 - \frac{1}{3}M_2\right](\frac{1}{2}) + \frac{1}{2}M_2(\frac{1}{2})^2 + \frac{1}{6}(M_3 - M_2)(\frac{1}{2})^3 = 3.218$$

4.3.3 Αριθμητική Παρεμβολή με Κυβικές B–Splines

Η μέθοδος της τμηματικά συνεχούς παρεμβολής μέσω κυβικών splines, που μόλις παρουσιάσθηκε, στηρίχθηκε στη χρήση ενός πολυώνυμου τρίτου βαθμού ως προς την παράμετρο u που εφαρμόσθηκε διαδοχικά στο διάστημα μεταξύ δύο διαδοχικών κομβικών σημείων. Στην ενότητα αυτή θα παρουσιάσουμε ένα τρόπο σχηματισμού πολυωνύμων τρίτου βαθμού ως προς την παράμετρο u τα οποία τώρα θα καταλαμβάνουν εύρος τεσσάρων διαδοχικών διαστημάτων, δηλ. πέντε διαδοχικών κομβικών σημείων, έστω των $(x_{i-2}, y_{i-2}), \ldots, (x_{i+2}, y_{i+2})$. Με βάση αυτά τα πολυώνυμα, θα ορισθεί η μέθοδος παρεμβολής με κυβικές B–splines, που βρίσκει ευρύτατη εφαρμογή.

Επειδή το κάθε πολυώνυμο καταλαμβάνει εύρος τεσσάρων διαστημάτων, το πεδίο ορισμού για την παράμετρο u ορίζεται ως το κλειστό διάστημα $[-2, 2]$. Έτσι, κάθε μία από τις πέντε ακέραιες τιμές του πεδίου ορισμού θα αντιστοιχεί σε καθένα από τα πέντε κομβικά σημεία. Στο διάστημα $[-2, 2]$ ορίζονται διαφορετικές εκφράσεις για το πολυώνυμο αυτό σε κάθε μοναδιαίο εύρος τιμών της u. Αυτές οι εκφράσεις, οι γνωστές και ως κυβικές B-splines είναι οι

$$B(u) = \begin{cases} b_{-2}(u+2) = \frac{(2+u)^3}{6} & -2 \le u \le -1 \\ b_{-1}(u+1) = \frac{4-6u^2-3u^3}{6} & -1 \le u \le 0 \\ b_0(u) = \frac{4-6u^2+3u^3}{6} & 0 \le u \le 1 \\ b_1(u-1) = \frac{(2-u)^3}{6} & 1 \le u \le 2 \\ 0 & \text{αλλού} \end{cases} \tag{4.52}$$



Σχήμα 4.8: Γραφική παράσταση της συνάρτησης κυβικής B-spline

.

Εναλλακτικά, τα τέσσερα παραπάνω πολυώνυμα μπορούν να γραφούν, σε κάθε επιμέρους διάστημα μοναδιαίου εύρους, με χρήση μιας νέας παραμέτρου, της v. Η v ορίζεται πάντα στο $[0,1]$ και συσχετίζεται με τη u με απλούς μετασχηματισμούς της μορφής $v = u+2$ (αν $-2 \leq u \leq -1$), $v = u+1$ (αν $-1 \leq u \leq 0$), κ.ο.κ. Τότε, οι συναρτήσεις b_{-2}, b_{-1}, b_0, b_1 γράφονται και ως

$$\begin{array}{lcl} b_{-2}(v) & = & \frac{v^3}{6} \\ b_{-1}(v) & = & \frac{1+3v+3v^2-3v^3}{6} \\ b_0(v) & = & \frac{4-6v^2+3v^3}{6} \\ b_1(v) & = & \frac{1-3v+3v^2-v^3}{6} \end{array} \tag{4.53}$$

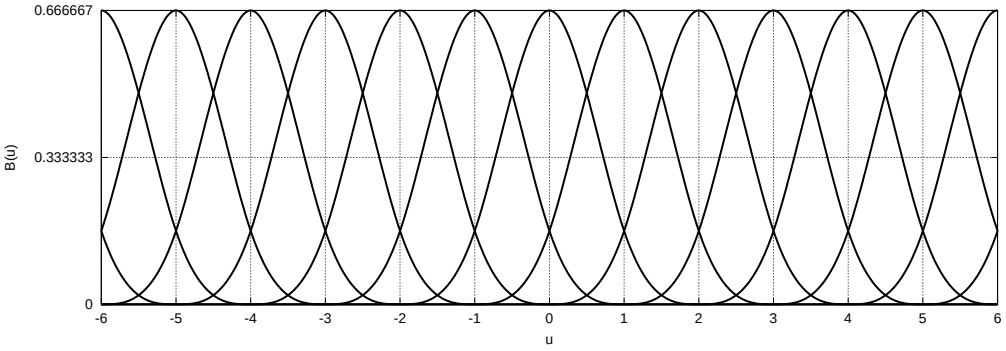
Στο Σχήμα 4.8 δίνεται η γραφική παράσταση της συνάρτησης κυβικής B-spline, δηλαδή της εξίσωσης 4.52. Όπως φαίνεται είναι

$$B(-2) = B(2) = 0 \quad , \quad B(-1) = B(1) = \frac{1}{6} \quad , \quad B(0) = \frac{2}{3} \tag{4.54}$$

Στο σημείο αυτό, είναι χρήσιμο να παραθέσουμε τις εκφράσεις των παραγώγων της κυβικής B-spline, ως προς την παράμετρο u, που θα χρησιμοποιηθεί στη συνέχεια. Είναι

$$\dot{B}(u) \;=\; \begin{cases} \frac{(2+u)^2}{2} & -2 \leq u \leq -1 \\ \frac{-4u-3u^2}{2} & -1 \leq u \leq 0 \\ \frac{-4u+3u^2}{2} & 0 \leq u \leq 1 \\ \frac{-(2-u)^2}{2} & 1 \leq u \leq 2 \\ 0 & \text{αλλού} \end{cases} \tag{4.55}$$

Από τον παραπάνω ορισμό φαίνονται εύκολα οι βασικές ιδιότητες της κυβικής B-spline. Έτσι, αυτή:



Σχήμα 4.9: Γραφική παράσταση της σύνθεσης των συναρτήσεων κυβικής B-spline, στον παραμετρικό χώρο, για μια αλληλουχία αρκετών κόμβων.

- είναι άρτια συνάρτηση, αφού $B(-u) = B(u)$,
- έχει συνεχείς πρώτες και δεύτερες παραγώγους,
- ισχύει ότι

$$B(u-2) + B(u-1) + B(u) + B(u+1) \;=\; 1 \quad , \quad 0 \le u \le 1 \tag{4.56}$$

ή ισοδύναμα

$$b_{-2}(v) + b_{-1}(v) + b_0(v) + b_1(v) \;=\; 1 \quad , \quad 0 \le v \le 1 \tag{4.57}$$

Η σημασία των σχέσεων 4.56 ή 4.57 είναι ουσιώδης. Ερμηνεύοντας λ.χ. τη σχέση 4.56, ας υποτεθεί ότι βρισκόμαστε σε μια θέση στο διάστημα $(x_i, y_i) \rightarrow (x_{i+1}, y_{i+1})$ που αντιστοιχεί σε ορισμένη τιμή του u ($0 \le u < 1$). Από τον ορισμό των κυβικών B-splines, από το σημείο αυτό διέρχονται τέσσερις τέτοιες συναρτήσεις (λογίζονται μόνο όσες έχουν μη–μηδενική τιμή, Σχήμα 4.9), με άθροισμα μονάδα.

Πέραν της παραμέτρου u, η οποία έχει 'τοπικό' χαρακτήρα (με την έννοια του ότι ορίζεται εκ νέου, με αφετηρία τον επόμενο κόμβο, για κάθε μια από τις κυβικές B-splines του Σχήματος 4.9), θα ορισθεί και η παράμετρος μ για την 'ολική' παραμετροποίηση της αλληλουχίας των $N+1$ κομβικών σημείων. Μεταξύ των u και μ θα ισχύει ότι

$$u \;=\; N\mu \tag{4.58}$$

Η παράμετρος μ μεταβάλλεται συνολικά από την τιμή $\mu = 0$ την οποία λαμβάνει στον πρώτο κόμβο (x_0, y_0) ως την τιμή $\mu = 1$ την οποία λαμβάνει στον τελευταίο κόμβο (x_N, y_N). Ενδιάμεσα, και σε κάθε άλλο δεδομένο κομβικό σημείο i, η παράμετρος αυτή λαμβάνει την τιμή $\mu = i/N$.

Η εξίσωση 4.56 ξαναγράφεται ως

$$B(N\mu-2)+B(N\mu-1)+B(N\mu)+B(N\mu+1) \;=\; 1 \quad ,\; 0\le\mu\le\frac{1}{N} \qquad (4.59)$$

Με βάση τους παραπάνω ορισμούς και τις σχέσεις που προέκυψαν, μπορούμε πλέον να παρουσιάσουμε τη χρήση των κυβικών B-splines στην αριθμητική παρεμβολή. Έστω $\overrightarrow{r}(x,y)$ το διάνυσμα θέσης κάθε σημείου της καμπύλης παρεμβολής που αντιστοιχεί σε μια οποιαδήποτε τιμή της παραμέτρου μ , $0\le\mu\le 1$. Τότε η εξίσωση παρεμβολής των $N+1$ δεδομένων σημείων $(x_0,y_0),\ldots,(x_N,y_N)$ μέσω κυβικών B-splines είναι η

$$\overrightarrow{r}(\mu) \;=\; \sum_{i=-1}^{N+1} B(N\mu-i)\overrightarrow{R}_i \qquad (4.60)$$

όπου $\overrightarrow{R}_i \;=\; (X_i,Y_i)$, $i=-1,\ldots,N+1$ είναι ένα σύνολο $N+3$ φανταστικών–βοηθητικών σημείων (ονομάζονται συνήθως *παραμετρικά κομβικά σημεία*, parametric knots) που 'βοηθούν' ώστε η προκύπτουσα καμπύλη να παρεμβάλει τα $N+1$ δεδομένα κομβικά σημεία. Η ύπαρξη δύο επιπλέον σημείων (με δείκτες $i=-1$ και $i=N+1$) εξυπηρετεί την επιβολή οριακών συνθηκών που εμπλέκουν παραγώγους στο πρώτο και το τελευταίο κομβικό σημείο. Η κατασκευή της καμπύλης παρεμβολής με την παραπάνω σχέση απαιτεί αρχικά την εύρεση των συντεταγμένων των $\overrightarrow{R}_i \;=\; (X_i,Y_i)$, $i=-1,\ldots,N+1$. Για τον υπολογισμό αυτό επιβάλλονται οι $N+1$ προφανείς συνθήκες, οι απαιτήσεις δηλαδή του να διέρχεται η καμπύλη από τα $N+1$ δεδομένα σημεία $(x_0,y_0),\ldots,(x_N,y_N)$. Έτσι, με βάση τη σχέση 4.60 και το ότι, όπως προαναφέραμε, ο i–ιοστός κόμβος αντιστοιχεί στην τιμή i/N, προκύπτει ότι

$$\overrightarrow{r}_i \;=\; \overrightarrow{r}(\frac{i}{N}) \;=\; \frac{1}{6}\left(\overrightarrow{R}_{i-1}+4\overrightarrow{R}_i+\overrightarrow{R}_{i+1}\right) \quad , \quad i=0,\ldots,N \qquad (4.61)$$

ενώ χρησιμοποιήθηκαν και οι γνωστές τιμές της συνάρτησης B, δηλαδή ότι $B(u=-1) \;=\; B(u=1) \;=\; \frac{1}{6}$ και $B(u=0) \;=\; \frac{4}{6}$

Έχοντας ήδη διατυπώσει $N+1$ εξισώσεις για τους $N+3$ αγνώστους (κάθε άγνωστος προσμετράται εδώ μια φορά, έστω και αν πρόκειται για τις δύο συντεταγμένες X_i και Y_i στο διδιάστατο χώρο, γίνεται δηλαδή διανυσματική θεώρηση), απομένουν δύο ακόμα εξισώσεις για την επίλυση του συστήματος και τον προσδιορισμό των $\overrightarrow{R}_i$. Οι δύο αυτές εξισώσεις προκύπτουν επιβάλλοντας οριακές συνθήκες για την πρώτη παράγωγο (ως προς μ) στο πρώτο και το τελευταίο κομβικό σημείο. Αν συμβολίσουμε τις τιμές των παραγώγων με $\overrightarrow{d}_0$ και $\overrightarrow{d}_N$, ισχύει ότι

$$\dot{\overrightarrow{r}}(0) \;=\; \overrightarrow{d}_0 \quad , \quad \dot{\overrightarrow{r}}(1) \;=\; \overrightarrow{d}_N$$

Αφού

$$\dot{\overrightarrow{r}}(\mu) \;=\; N \sum_{i=-1}^{N+1} \dot{B}(N\mu - i)\overrightarrow{R}_i \tag{4.62}$$

άρα

$$\begin{aligned}
\dot{\overrightarrow{r}}(0) \;&=\; N\left(\dot{B}(1)\overrightarrow{R}_{-1} + \dot{B}(0)\overrightarrow{R}_0 + \dot{B}(-1)\overrightarrow{R}_1\right) \;=\; d_0 \\
\dot{\overrightarrow{r}}(1) \;&=\; N\left(\dot{B}(1)\overrightarrow{R}_{N-1} + \dot{B}(0)\overrightarrow{R}_N + \dot{B}(-1)\overrightarrow{R}_{N+1}\right) \;=\; d_N
\end{aligned}$$

που, με αντικατάσταση των παραγώγων του B, δίνουν

$$\begin{aligned}
d_0 \;&=\; \frac{N}{2}\left(\overrightarrow{R}_1 - \overrightarrow{R}_{-1}\right) \\
d_N \;&=\; \frac{N}{2}\left(\overrightarrow{R}_{N+1} - \overrightarrow{R}_{N-1}\right)
\end{aligned}$$

Με βάση τα παραπάνω διαμορφώνεται το τελικό σύστημα εξισώσεων

$$\begin{bmatrix}
-N & 0 & N & & & & \\
1 & 4 & 1 & & & & \\
 & 1 & 4 & 1 & & & \\
 & & & \vdots & & & \\
 & & & & 1 & 4 & 1 \\
 & & & & -N & 0 & N
\end{bmatrix}
\begin{bmatrix}
\overrightarrow{R}_{-1} \\ \overrightarrow{R}_0 \\ \overrightarrow{R}_1 \\ \vdots \\ \overrightarrow{R}_N \\ \overrightarrow{R}_{N+1}
\end{bmatrix}
\;=\; 6
\begin{bmatrix}
\frac{1}{3}\overrightarrow{d}_0 \\ \overrightarrow{r}_0 \\ \overrightarrow{r}_1 \\ \vdots \\ \overrightarrow{r}_N \\ \frac{1}{3}\overrightarrow{d}_N
\end{bmatrix} \tag{4.63}$$

Με βάση τα προηγούμενα, η χρήση των κυβικών B–splines για την αριθμητική παρεμβολή με $N+1$ δεδομένα σημεία απαιτεί γνώση ή υπόθεση για τις τιμές των $\overrightarrow{d}_0$ και $\overrightarrow{d}_N$, τη διαμόρφωση και επίλυση του συστήματος 4.63 για τον προσδιορισμό των συντεταγμένων των παραμετρικών κομβικών σημείων $\overrightarrow{R}_i \;=\; (X_i, Y_i)$, $i = -1, \ldots, N+1$ και τέλος τη χρήση της εξίσωσης 4.60 ώστε να παρεμβληθούν νέα σημεία (για τιμές της παραμέτρου μ). Τονίζεται δε, εκ νέου, η διάκριση μεταξύ των $N+1$ δεδομένων κομβικών σημείων $\overrightarrow{r}_i$, $i = 0, N$ και των $N+3$ βοηθητικών παραμετρικών κομβικών σημείων $\overrightarrow{R}_i$, $i = -1, N+1$.

Στην εξίσωση 4.63, παρατηρούμε ότι το μητρώο των συντελεστών των αγνώστων είναι το ίδιο γιά κάθε συνιστώσα του διανύσματος των αγνώστων (δηλαδή για τα X_i, Y_i και ίσως Z_i αν πρόκειται για καμπύλη στον τριδιάστατο χώρο) και συνεπώς απαιτείται η αντιστροφή του μια μόνο φορά, άσχετα με τη διάσταση του προβλήματος. Το μητρώο αυτό μπορεί να λάβει τριδιαγώνια μορφή με απλή διαχείριση της πρώτησ-δεύτερης και τελευταίασ-προτελευταίας γραμμής του.

4.3.4 Σχόλια

Στις προηγούμενες ενότητες γνωρίσαμε μερικές από τις πιο συνηθισμένες μεθόδους αριθμητικής παρεμβολής, με τμηματικά συνεχή πολυώνυμα. Ανακεφαλαιώνοντας, είναι χρήσιμο να αποσαφηνισθούν τα βασικά χαρακτηριστικά των μεθόδων που γνωρίσαμε, οι απαιτήσεις τους σε διαθέσιμη πληροφορία (πέραν των συντεταγμένων των $N+1$ κομβικών σημείων), το υπολογιστικό τους κόστος και οι αδυναμίες τους.

Η τμηματικά συνεχής κατά Lagrange παρεμβολή (με τετραγωνικά πολυώνυμα) παράγει ασυνέχεια στην κλίση της καμπύλης παρεμβολής στα κομβικά σημεία. Το ότι η ασυνέχεια αυτή παύει να υφίσταται αν χρησιμοποιηθεί η τμηματικά συνεχής κατά Hermite παρεμβολή (με κυβικά πολυώνυμα) αντισταθμίζεται με το ότι οφείλουν να είναι γνωστές οι τιμές της πρώτης παραγώγου. Και στις δύο όμως μεθόδους, η δεύτερη παράγωγος στα κομβικά σημεία παραμένει ασυνεχής.

Η αριθμητική παρεμβολή με τμηματικά συνεχείς κυβικές splines ή B–splines εξασφαλίζει τη συνέχεια πρώτης και δεύτερης παραγώγου, χωρίς να απαιτεί επιπλέον πληροφορία για την κλίση στα (εσωτερικά) κομβικά σημεία. Απαιτούνται μόνο υποθέσεις για τις παραγώγους στο πρώτο και το τελευταίο σημείο, κάτι που γενικά πάντα μπορεί να το χειριστεί ο μηχανικός, χωρίς σημαντική επαγόμενη ανακρίβεια στην καμπύλη παρεμβολής.

Οι τμηματικά συνεχείς κυβικές splines ή B–splines επιλέχθηκε να παρουσιασθούν σε παραμετρική γραφή. Και οι δύο συντεταγμένες, x και y, εκφράσθηκαν ως συνάρτηση μιας 'τοπικής' παραμέτρου u. Θα μπορούσε ασφαλώς η παρουσίαση να γίνει σε μη–παραμετρική μορφή. Στην περίπτωση αυτή, οι κυβικές splines λ.χ. θα διέπονταν από ένα πολυώνυμο της μορφής

$$y \;=\; g(x) \quad = \quad c_0 + c_1 x + c_2 x^2 + c_3 x^3 \tag{4.64}$$

σε αντιστοιχία με τα παραμετρικά πολυώνυμα 4.40. Για τα $N+1$ δεδομένα σημεία (N διαστήματα), απαιτούνται να υπολογισθούν $4N$ συντελεστές, χρειάζονται δηλαδή $4N$ εξισώσεις. Σε κάθε κομβικό σημείο πέραν των οριακών, πρέπει να ικανοποιείται η τιμή του y από τα δύο εκατέρωθεν πολυώνυμα (δύο εξισώσεις) και να υπάρχει ταύτιση πρώτης και δεύτερης παραγώγου αυτών. Με τον τρόπο αυτό καταγράφονται $4N-4$ εξισώσεις. Δύο επιπλέον συνθήκες προκύπτουν από τους δύο οριακούς κόμβους, όπου το 'τοπικό' κυβικό πολυώνυμο οφείλει να διέρχεται, αντίστοιχα, από τα σημεία (x_0, y_0) και (x_N, y_N). Έτσι, το σύνολο των διαθέσιμων εξισώσεων ανέρχεται συνολικά σε $4N-2$. Οι δύο απομένοντες βαθμοί ελευθερίας καλύπτονται από υποθέσεις του χρήστη (αν και εφόσον δεν είναι γνωστή επιπλέον πληροφορία στα όρια). Μια δυνατότητα είναι να υποτεθούν μηδενικές δεύτερες παράγωγοι στα ακραία κομβικά σημεία, που έχει ήδη αναφερθεί ως *φυσικές splines*. Άλλες δυνατότητες υπάρχουν, σε συμφωνία με όσα παρουσιάσθηκαν προηγούμενα για τις παραμετρικές κυβικές splines.

4.4 Αριθμητική Προσέγγιση Καμπυλών

Όπως αναφέρθηκε και στην εισαγωγή, η αριθμητική προσέγγιση καμπυλών συνίσταται στην ανάπτυξη και χρήση αλγορίθμων με τους οποίους, έχοντας δεδομένο ένα πλήθος $N+1$ σημείων, να είναι δυνατή η δημιουργία μιας καμπύλης της οποίας η μορφή να προσεγγίζει με βέλτιστο τρόπο τα δεδομένα σημεία, χωρίς αναγκαστικά να διέρχεται από μερικά ή όλα από αυτά. Ως εργαλεία, οι τεχνικές αριθμητικής προσέγγισης καμπυλών στοχεύουν στο να επιτρέψουν τη διαχείριση δεδομένων που περιέχουν ασάφεια ή είναι αποτέλεσμα ανακριβούς λήψης, αλλά και να υποστηρίξουν τη σχεδίαση ή τροποποίηση γεωμετρικών μορφών (λ.χ. του περιγράμματος ενός αυτοκινήτου, μιας πτέρυγας αεροσκάφους, κλπ) με τρόπο που ο μηχανικός, ελέγχοντας τη θέση μόνο λίγων σημείων (*σημεία ελέγχου*) να μπορεί να τροποποιεί ολόκληρο το σχήμα.

4.4.1 Προσέγγιση Καμπυλών μέσω Κυβικών B–Splines

Οι κυβικές B–splines που προηγούμενα χρησιμοποιήθηκαν για την παρεμβολή μιας καμπύλης που περιγράφεται από $N+1$ σημεία, μπορούν, με κατάλληλες αλλαγές, να χρησιμοποιηθούν και για την προσέγγιση μιας καμπύλης που ορίζει μια αλληλουχία τέτοιων σημείων. Για το σκοπό αυτό, η εξίσωση 4.60 επαναδιατυπώνεται χρησιμοποιώντας τα δεδομένα $N+1$ σημεία $\overrightarrow{r}_i$ αντί των παραμετρικών κομβικών σημείων που χρειάστηκαν προηγούμενα (ακριβώς για να 'οδηγήσουν' την καμπύλη να 'περάσει' από τα $N+1$ δεδομένα σημεία).

Ετσι η καμπύλη προσέγγισης γράφεται στη μορφή

$$\overrightarrow{r}(\mu) \;=\; \sum_{i=0}^{N} B(N\mu - i)\,\overrightarrow{r}_i \tag{4.65}$$

όπου ισχύουν πάλι ότι το $\overrightarrow{r}_0$ αντιστοιχεί στο $\mu = 0$ ενώ το $\overrightarrow{r}_N$ στο $\mu = 1$. Η καμπύλη προσέγγισης (σε αντίθεση με την καμπύλη παρεμβολής) γενικά δεν διέρχεται από τα δεδομένα κομβικά σημεία. Συνήθως ο όρος *κομβικά σημεία* στις καμπύλες προσέγγισης αντικαθίσταται από τον όρο *σημεία ελέγχου* (control points), ακριβώς για να τονισθεί αυτή η διαφορά στο ρόλο των σημείων.

Αξίζει να τονισθεί ότι οι δυνατές παραλλαγές χρήσης της μεθόδου αυτής καμπυλών δίνουν έμφαση στο αν τα δύο ακραία σημεία ($\overrightarrow{r}_0$ και $\overrightarrow{r}_N$) της αλληλουχίας ανήκουν ή όχι στην καμπύλη προσέγγισης. Λ.χ., σύμφωνα με τον ορισμό 4.52 των συναρτήσεων $B(u)$ και σύμφωνα με τις τιμές της 4.54 φαίνεται εύκολα ότι

$$\begin{aligned}
\overrightarrow{r}(0) &= B(0)\;\overrightarrow{r}_0 + B(-1)\overrightarrow{r}_1 \;=\; \frac{2}{3}\overrightarrow{r}_0 + \frac{1}{6}\overrightarrow{r}_1 \neq \overrightarrow{r}_0 \\
\overrightarrow{r}(1) &= B(1)\;\overrightarrow{r}_{N-1} + B(0)\overrightarrow{r}_N \;=\; \frac{1}{6}\overrightarrow{r}_{N-1} + \frac{2}{3}\overrightarrow{r}_N \neq \overrightarrow{r}_N
\end{aligned} \tag{4.66}$$

Παρόλα αυτά, είναι εύκολο να ικανοποιηθεί η απαίτηση να περνά η κυβική B-spline τουλάχιστον από το πρώτο και το τελευταίο σημείο ελέγχου, με τη βοήθεια δύο φαν-

ταστικών σημείων ελέγχου (ψευδοσημεία ελέγχου) των $\overrightarrow{r}_{-1}$ και $\overrightarrow{r}_{N+1}$. Η εξίσωση 4.65 παίρνει τότε τη μορφή

$$\overrightarrow{r}(\mu) \;=\; \sum_{i=-1}^{N+1} B(N\mu - i)\overrightarrow{r}_i \tag{4.67}$$

και επιπλέον επιβάλλονται οι σχέσεις

$$\begin{aligned} \overrightarrow{r}_0 &= \overrightarrow{r}(0) = \frac{1}{6}\overrightarrow{r}_{-1} + \frac{2}{3}\overrightarrow{r}_0 + \frac{1}{6}\overrightarrow{r}_1 \\ \overrightarrow{r}_N &= \overrightarrow{r}(1) = \frac{1}{6}\overrightarrow{r}_{N-1} + \frac{2}{3}\overrightarrow{r}_N + \frac{1}{6}\overrightarrow{r}_{N+1} \end{aligned} \tag{4.68}$$

που προσδιορίζουν τα ψευδοσημεία ελέγχου, ως

$$\begin{aligned} \overrightarrow{r}_{-1} &= 2\ \overrightarrow{r}_0 - \overrightarrow{r}_1 \\ \overrightarrow{r}_{N+1} &= 2\ \overrightarrow{r}_N - \overrightarrow{r}_{N-1} \end{aligned} \tag{4.69}$$

Από τα παραπάνω φαίνεται ο ρόλος των ψευδοσημείων ελέγχου: εισάγονται στις εξισώσεις και προσδιορίζονται έτσι ώστε να αναγκάζουν την καμπύλη προσέγγισης να διέρχεται από το πρώτο και το τελευταίο κομβικό σημείο.

Μια ενδιαφέρουσα ιδιότητα των κυβικών B-splines είναι το ότι τα δύο ακραία σημεία ελέγχου (πρώτο–δεύτερο και τελευταίο–προτελευταίο, χωρίς να προσμετρώνται τα ψευδοσημεία ελέγχου) καθορίζουν την κλίση της καμπύλης προσέγγισης στα ακραία σημεία.

Από τη σχέση (η παραγώγιση γίνεται ως προς την παράμετρο μ)

$$\dot{\overrightarrow{r}}(\mu) \;=\; N \sum_{i=-1}^{N+1} \dot{B}(N\mu - i)\overrightarrow{r}_i \tag{4.70}$$

προκύπτει ότι

$$\begin{aligned} \dot{\overrightarrow{r}}(0) &= N \sum_{i=-1}^{N+1} \dot{B}(-i)\overrightarrow{r}_i \;= \\ &= N\left(\dot{B}(1)\overrightarrow{r}_{-1} + \dot{B}(0)\overrightarrow{r}_0 + \dot{B}(-1)\overrightarrow{r}_1\right) \;=\; -\frac{N}{2}\left(\overrightarrow{r}_{-1} - \overrightarrow{r}_1\right) \end{aligned}$$

Αντικαθιστούμε την έκφραση του $\overrightarrow{r}_{-1}$ από τη σχέση 4.69 και τελικά προκύπτει ότι

$$\dot{\overrightarrow{r}}(0) \;=\; N\left(\overrightarrow{r}_1 \;-\; \overrightarrow{r}_0\right) \tag{4.71}$$

Όμοια αποδεικνύεται ότι για $\mu = 1$ ισχύει

$$\dot{\overrightarrow{r}}(1) \;=\; N\,(\overrightarrow{r}_N \;-\; \overrightarrow{r}_{N-1}) \tag{4.72}$$

Σχέσεις όπως η 4.71 (ή η 4.72), έστω και αν εμπλέκουν παραγώγους ως προς την παράμετρο μ είναι εύκολα μετατρέψιμες σε κλίσεις της καμπύλης. Έτσι λ.χ. για την 4.71, έχουμε

$$\frac{dy}{dx}(0) \;=\; \frac{\dot{y}(0)}{\dot{x}(0)} \;=\; \frac{y_1 - y_0}{x_1 - x_0} \tag{4.73}$$

Από την άλλη πλευρά, για κάθε μη-ακέραια τιμή του γινομένου $N\mu$, είναι εμφανές ότι μόνο τέσσερις όροι στη σειρά είναι μη μηδενικοί. Δηλαδή, κάθε διάστημα στην καμπύλη καθορίζεται από το πολύ 4 διαδοχικά σημεία ελέγχου. Πρακτικά, αυτό σημαίνει ότι αν τροποποιηθεί το σημείο $\overrightarrow{r}_k$, οι αλλαγές που θα υποστεί η καμπύλη παρεμβολής θα περιοριστούν στο εύρος από το $k-1$ ως το $k+2$ σημείο.

Η αλληλουχία των σημείων ελέγχου ονομάζεται *πολύγωνο ορισμού της κυβικής B-spline*. Το πολύγωνο ορισμού της B-spline μπορεί να περιλαμβάνει ένα ή περισσότερα σημεία ελέγχου πολλαπλές φορές. Ένα σημείο ελέγχου που δίνεται δύο φορές έλκει την καμπύλη προσέγγισης προς το μέρος του. Τέλος, όταν ένα σημείο ελέγχου δίνεται τρείς φορές, τότε η καμπύλη προσέγγισης περνά αναγκαστικά από αυτό.

4.4.2 Προσέγγιση Καμπυλών με τη Μέθοδο των Ελαχίστων Τετραγώνων

Η προσέγγιση μιας καμπύλης η οποία περιγράφεται διακριτά με $N+1$ δεδομένα σημεία μπορεί πραγματοποιηθεί αναζητώντας το πολυώνυμο εκείνο (βαθμού που συνήθως έχει αποφασισθεί εκ των προτέρων) το οποίο την προσεγγίζει με το *βέλτιστο τρόπο* (best fit). Διατηρώντας χαμηλό το βαθμό του πολυωνύμου (και σαφώς μικρότερο του N, για λόγους που θα φανούν στη συνέχεια) η καμπύλη που παριστά το πολυώνυμο δεν θα διέρχεται από (μερικά ή όλα) από τα δεδομένα σημεία.

Αυτό που προηγουμένως αναφέρθηκε ως *βέλτιστος τρόπος* μπορεί ασφαλώς να ερμηνευθεί και να εκφρασθεί μαθηματικά με πολλούς τρόπους. Η συνήθης ερμηνεία του είναι αυτή που επιβάλλει την ελαχιστοποίηση της *απόκλισης* (deviation) μεταξύ της καμπύλης προσέγγισης και των $N+1$ δεδομένων σημείων. Αλλά και ο όρος *απόκλιση* δέχεται, με τη σειρά του, πολλαπλές ερμηνείες. Εδώ, αν $g(x)$ είναι το πολυώνυμο προσέγγισης, θα ορίσουμε ως απόκλιση για την τιμή εισόδου x_i , $i = 0, \ldots, N$ τη διαφορά

$$e_i \;=\; g(x_i) - y_i \tag{4.74}$$

ενώ η ζητούμενη ελαχιστοποίηση της απόκλισης διατυπώνεται μαθηματικά ως ελαχιστοποίηση του αθροίσματος των τετραγώνων των $N+1$ κομβικών αποκλίσεων. Άρα η απαίτηση είναι

$$\sum_{i=0}^{N} (e_i)^2 \;=\; minimum \tag{4.75}$$

Η τελευταία έκφραση δεν είναι και η μοναδική δυνατότητα. Εναλλακτικά θα μπορούσε να εφαρμοσθεί το λεγόμενο κριτήριο minimax, το οποίο καθορίζει ως βέλτιστη καμπύλη (συνήθως ευθεία) προσέγγισης εκείνη από την οποία ελαχιστοποιείται η μέγιστη απόλυτη τιμή της απόκλισης. Το κριτήριο minimax, παρά το ότι είναι σε χρήση, δεν θα μας απασχολήσει περισσότερο αφού συνήθως παράγει κακής ποιότητας προσεγγίσεις αν υπάρχει ένα σημείο από τα $N+1$ δεδομένα που είναι αρκετά απομακρυσμένο από τα υπόλοιπα.

Στη συνέχεια της ενότητας αυτής θα ασχοληθούμε μόνο με τη μέθοδο πολυωνυμικής προσέγγισης ελαχίστων τετραγώνων. Στη γενική της μορφή, αυτή βασίζεται σε ένα πολυώνυμο βαθμού n ($n < N$ ή και $n << N$)

$$y(x) \;\doteq\; g(x) \;=\; a_0 + \; a_1 x + \; a_2 x^2 + \; \ldots a_{n-1} x^{n-1} + \; a_n x^n \tag{4.76}$$

με την απόκλιση να ορίζεται ως

$$e_i \;=\; g(x_i) - y_i \;=\; a_0 + \; a_1 x_i + \; a_2 x_i^2 + \; \ldots a_{n-1} x_i^{n-1} + \; a_n x_i^n - y_i \tag{4.77}$$

Το άθροισμα των τετραγώνων των αποκλίσεων, η ποσότητα δηλαδή που πρέπει να ελαχιστοποιηθεί σύμφωνα με τη σχέση 4.75, γράφεται

$$E \;=\; \sum_{i=0}^{N} e_i \;=\; \sum_{i=0}^{N} \left(a_0 + \; a_1 x_i + \; a_2 x_i^2 + \; \cdots + a_{n-1} x_i^{n-1} + \; a_n x_i^n - y_i\right)^2 \;=\; minimum \tag{4.78}$$

Η ελαχιστοποίηση της ποσότητας E εξασφαλίζεται για τις τιμές των συντελεστών a_i , $i = 0, \ldots, n$ που μηδενίζουν τις παραγώγους

$$\frac{\partial E}{\partial a_k} \;=\; \sum_{i=0}^{N} 2\left(a_0 + \; a_1 x_i + \; a_2 x_i^2 + \; \cdots + a_{n-1} x_i^{n-1} + \; a_n x_i^n - y_i\right) x^k \;=\; 0 \tag{4.79}$$

$$, \qquad k = 0, \ldots, n$$

και συγχρόνως καθιστούν θετικές τις δεύτερες παραγώγους. Το τελευταίο είναι εύκολο να αποδειχθεί και δεν θα επιμείνουμε περισσότερο σε αυτό. Οι $n+1$ σχέσεις 4.79 οδηγούν στο σύστημα $n+1$ γραμμικών εξισώσεων

$$\begin{bmatrix} (N+1) & \sum_{i=0}^{N} x_i & \ldots & \sum_{i=0}^{N} x_i^n \\ \sum_{i=0}^{N} x_i & \sum_{i=0}^{N} x_i^2 & \ldots & \sum_{i=0}^{N} x_i^{n+1} \\ & & \vdots & \\ \sum_{i=0}^{N} x_i^n & \sum_{i=0}^{N} x_i^{n+1} & \ldots & \sum_{i=0}^{N} x_i^{2n} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} \;=\; \begin{bmatrix} \sum_{i=0}^{N} y_i \\ \sum_{i=0}^{N} y_i x_i \\ \vdots \\ \sum_{i=0}^{N} y_i x_i^n \end{bmatrix} \tag{4.80}$$

το οποίο είναι συμμετρικό και επιλύεται εύκολα. Παρατηρώντας, παρόλα αυτά, τα στοιχεία του συστήματος 4.80 είναι εύκολο να αντιληφθεί κανείς αριθμητικές δυσκολίες (ασθενής κατάσταση μήτρας και τα επαγόμενα σφάλματα στα αποτελέσματα) όταν το πολυώνυμο είναι υψηλού βαθμού. Η αδιαστοποίηση των εξισώσεων και η αναγκαστική χρήση πραγματικών μεταβλητών διπλής ακρίβειας στον υπολογιστή είναι δύο ενέργειες που, σε κάποιες περιπτώσεις, διορθώνουν το πρόβλημα. Πρακτικά, συνιστάται να μην χρησιμοποιούνται τιμές του n μεγαλύτερες του 5 ή του 6.

Ως απλούστερη περίπτωση των παραπάνω, θα παρουσιάσουμε, σε συντομία, και την *γραμμική προσέγγιση ελαχίστων τετραγώνων* (straight–line least squares approximation), όπου αναζητείται η ευθεία με το ελάχιστο άθροισμα τετραγώνων από τα $N+1$ σημεία. Η εξίσωση της ευθείας είναι ($n=1$, βλ. και 4.76)

$$y(x) \doteq g(x) = a_0 + a_1 x \tag{4.81}$$

ενώ, σύμφωνα και με την 4.79, η ελαχιστοποίηση της ποσότητας E εξασφαλίζεται όταν

$$\begin{aligned} \frac{\partial E}{\partial a_0} &= \sum_{i=0}^{N} 2\left(a_0 + a_1 x_i - y_i\right) = 0 \\ \frac{\partial E}{\partial a_1} &= \sum_{i=0}^{N} 2\left(a_0 + a_1 x_i - y_i\right) x_i = 0 \end{aligned} \tag{4.82}$$

που οδηγεί στο σύστημα δύο εξισώσεων με δύο αγνώστους

$$\begin{bmatrix} (N+1) & \sum_{i=0}^{N} x_i \\ \sum_{i=0}^{N} x_i & \sum_{i=0}^{N} x_i^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{N} y_i \\ \sum_{i=0}^{N} y_i x_i \end{bmatrix} \tag{4.83}$$

Το σύστημα 4.83 είναι απλά μια ειδική περίπτωση του 4.80 και το ότι παρουσιάζεται ξεχωριστά οφείλεται μόνο στην ευκολία με την οποία μπορεί κάποιος να επεξεργαστεί τις δεύτερες παράγωγους. Είναι

$$\begin{aligned} \frac{\partial^2 E}{\partial a_0^2} &= \sum_{i=0}^{N} 2 = 2(N+1) \\ \frac{\partial^2 E}{\partial a_1^2} &= 2\sum_{i=0}^{N} x_i^2 \\ \frac{\partial^2 E}{\partial a_0 a_1} &= 2\sum_{i=0}^{N} x_i \end{aligned} \tag{4.84}$$

Οι δύο πρώτες ποσότητες είναι πάντοτε θετικές. Συγχρόνως ισχύει ότι

$$\left(2\sum_{i=0}^{N} x_i\right)^2 - 2(N+1)2\sum_{i=0}^{N} x_i^2 < 0 \tag{4.85}$$

και συνεπώς οι λύσεις a_0 και a_1 του συστήματος 4.83 ελαχιστοποιούν την ποσότητα E.

Εφαρμογή

Έστω ότι μια σειρά μετρήσεων κατέγραψε τιμές της ειδικής θερμοχωρητικότητας του αέρα υπό σταθερή πίεση (C_p, σε $kJ/kg/K$ σε 7 ($N = 6$) διαφορετικές θερμοκρασίες (T, σε K). Οι τιμές παρουσιάζονται στον πίνακα που ακολουθεί:

T	C_p
300	1.0045
400	1.0134
500	1.0296
600	1.0507
700	1.0743
800	1.0984
900	1.1212

Ζητείται να προσεγγισθούν τα διακριτά δεδομένα με τη μέθοδο των ελαχίστων τετραγώνων, χρησιμοποιώντας αφενός γραμμική και αφετέρου τετραγωνική (πολυώνυμο δεύτερου βαθμού) προσέγγιση, δημιουργώντας σχετικό κώδικα γενικής χρήσης σε γλώσσα Fortran 77.

Λύση:

Ας συμβολίσουμε με x τη μεταβλητή εισόδου (T) και y την απόκριση (C_p). Για τη γραμμική προσέγγιση ελαχίστων τετραγώνων εφαρμόζεται το πολυώνυμο της 4.81, ενώ για την τετραγωνική προσέγγιση το

$$y(x) \doteq g(x) = a_0 + a_1 x + a_2 x^2 \tag{4.86}$$

Ο προτεινόμενος κώδικας για το πρόβλημα ακολουθεί, είναι γραμμένος σκόπιμα σε απλή μορφή και όχι κατ΄ ανάγκη με το βέλτιστο τρόπο, ώστε τα τμήματά του να παρακολουθούν κατά βήμα την προηγηθείσα θεωρία. Θα μπορούσε λ.χ. ο προγραμματιστής να εκμεταλλευθεί το γεγονός ότι το μητρώο συντελεστών του συστήματος 4.80 είναι συμμετρικό, τόσο κατά τη συμπλήρωσή του όσο και κατά την αντιστροφή του. Εδώ όμως προτιμήθηκε να χρησιμοποιηθεί η κλασική μέθοδος απαλοιφής κατά Gauss.

Το κυρίως πρόγραμμα διαβάζει τα δεδομένα που πινακοποιήθηκαν προηγούμενα από το αρχείο data, μετρώντας συγχρόνως και το πλήθος τους. Τα υποπρογράμματα fill_lhs και fill_rhs συμπληρώνουν τα στοιχεία του μητρώου συντελεστών και το διάνυσμα του δεξιού μέλους της εξίσωσης 4.80.

Το μοναδικό δεδομένο που απαιτείται να πληκτρολογηθεί από το χρήστη είναι ο βαθμός του πολυωνύμου. Για τις δύο περιπτώσεις που θα αναλυθούν, οι τιμές είναι αντίστοιχα 1 και 2. Τα αποτελέσματα του προγράμματος καταγράφονται στο αρχείο approx, με τις δεδομένες τιμές εισόδου x_i, την προσέγγιση $g(x_i)$ που υπολογίζεται και το επί τοις εκατό σχετικό σφάλμα, οριζόμενο ως

$$error = 100\frac{g(x_i) - y(x_i)}{y(x_i)}\% \tag{4.87}$$

```
c*****************************************
      program least_squares
c*****************************************
      implicit double precision(a-h,o-z)
      parameter (ndata=100,npol=10)
      dimension x(ndata),y(ndata)
      dimension a(npol,npol),b(npol),coef(npol)
c
      open(1,file='data')
      do i=1,ndata
        read(1,*,end=10)x(i),y(i)
      enddo
      stop     'Increase NDATA'
  10  npoints=i-1
      close(1)
c
      write(*,*)' Enter polynomial degree '
      read(*,*)mpol
      if(mpol.ge.npol) stop 'Increase NPOL'
c
      do i=1,mpol+1
      do j=i,mpol+1
        call fill_lhs(ndata,npoints,x,i,j,res)
        a(i,j)=res
        a(j,i)=res
      enddo
        call fill_rhs(ndata,npoints,x,y,i,b(i))
      enddo
c
      call gauss(npol,mpol+1,a,b,coef)
c
      open(1,file='approx')
      do i=1,npoints
        app=0.d0
      do j=1,mpol+1
```

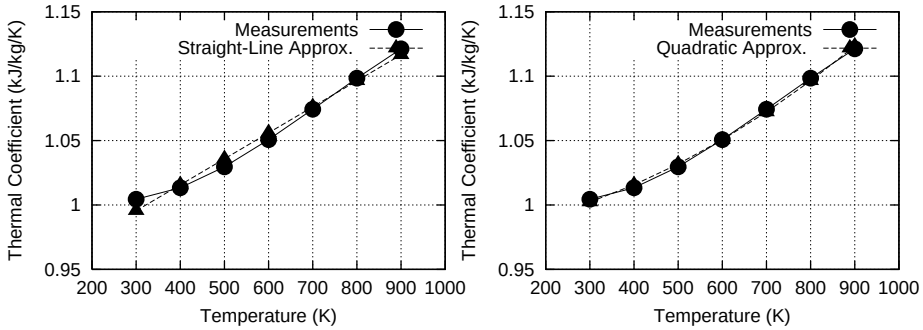
```
        app=app+coef(j)*x(i)**(j-1)
      enddo
      error=100.*(app-y(i))/y(i)
      write(1,'(2x,f7.0,2x,f12.5,2x,f10.5)')x(i),app,error
      enddo
      close(1)
c
      end
c
c
c*********************************************************************
      subroutine fill_lhs(ndata,npoints,x,i,j,res)
c*********************************************************************
      implicit double precision(a-h,o-z)
      dimension x(ndata)
      res=0.d0
      expo=dfloat(i+j-2)
      do m=1,npoints
        res = res + x(m)**expo
      enddo
      return
      end
c
c*********************************************************************
      subroutine fill_rhs(ndata,npoints,x,y,i,res)
c*********************************************************************
      implicit double precision(a-h,o-z)
      dimension x(ndata),y(ndata)
      res=0.d0
      do m=1,npoints
        res = res + y(m)*x(m)**dfloat(i-1)
      enddo
      return
      end
c
c
c*********************************************************************
      subroutine gauss(kdim,n,a,b,x)
c*********************************************************************
      implicit double precision (a-h,o-z)
      dimension a(kdim,kdim),b(kdim),x(kdim)
      eps=1.d-8
      do k=1,n-1
      if (dabs(a(k,k)).lt.eps)   stop   'Divide by Zero !!!!'
```

```
      do i=k+1,n
      factor=a(i,k)/a(k,k)
      do j=k+1,n
      a(i,j)=a(i,j)-factor*a(k,j)
      enddo
      b(i)  =b(i)  -factor*b(k)
      enddo
      enddo
c
      x(n)=b(n)/a(n,n)
      do i=n-1,1,-1
      sum=0.d0
      do j=i+1,n
      sum=sum+a(i,j)*x(j)
      enddo
      x(i)=(b(i)-sum)/a(i,i)
      enddo
c
      return
      end
```

Χρησιμοποιώντας το παραπάνω λογισμικό προκύπτουν τα αποτελέσματα για τη γραμμική προσέγγιση (στη μορφή του αρχείου approx) που πινακοποιούμε στη συνέχεια:

T	$C_p(approx)$	$error(\%)$
300.	0.99550	-0.89597
400.	1.01567	0.22414
500.	1.03584	0.60634
600.	1.05601	0.50579
700.	1.07619	0.17553
800.	1.09636	-0.18598
900.	1.11653	-0.41665

ενώ η τετραγωνική προσέγγιση δίνει:



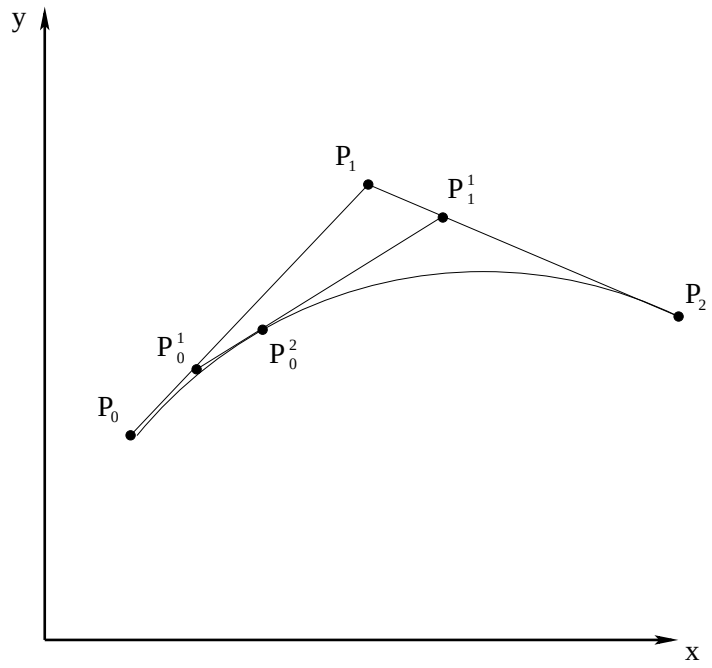
Σχήμα 4.10: Αποτελέσματα από τη γραμμική προσέγγιση (αριστερά) και την τετραγωνική προσέγγιση (δεξιά) με τη μέθοδο ελαχίστων τετραγώνων. Σύγκριση με τις πραγματικές τιμές.

T	$C_p(approx)$	$error(\%)$
300.	1.00229	-0.22044
400.	1.01567	0.22414
500.	1.03177	0.21090
600.	1.05059	-0.01088
700.	1.07211	-0.20345
800.	1.09636	-0.18598
900.	1.12331	0.18857

Μιά γραφική απεικόνιση των αποτελεσμάτων δίνεται στο Σχήμα 4.10

4.4.3 Προσέγγιση Καμπυλών μέσω Πολυωνύμων Bezier–Bernstein

Οι καμπύλες Bezier–Bernstein (ή απλά καμπύλες Bezier όπως επίσης αποκαλούνται λόγω συντομίας) αποτελούν έναν πολύ απλό και ευέλικτο τρόπο να προσεγγιστεί μια γεωμετρική μορφή χρησιμοποιώντας έναν περιορισμένο αριθμό σημείων ελέγχου. Βασική ιδιότητα της καμπύλης Bezier που θα προκύψει από τα $N+1$ σημεία $(x_0, y_0), \ldots, (x_N, y_N)$ είναι το ότι η καμπύλη ξεκινά από το πρώτο και τερματίζει στο τελευταίο σημείο της αλληλουχίας και το ότι το ζεύγος των δύο πρώτων και το ζεύγος των δύο τελευταίων σημείων καθορίζει την κλίση της καμπύλης στο πρώτο και στο τελευταίο σημείο της (όπως ακριβώς ισχύει και για την προσέγγιση με κυβικές B-splines). Το γιατί ισχύουν τα παραπάνω αλλά και άλλες ιδιότητες των καμπύλων Bezier θα αναλυθούν στη συνέχεια. Πρώτα όμως θα παρουσιαστεί ο αναδρομικός αλγόριθμος του de Casteljau, που αποτελεί τη βάση για τη δημιουργία των καμπυλών Bezier.



Σχήμα 4.11: Εφαρμογή του αλγόριθμου de Casteljau για τρία δεδομένα σημεία.

Αλγόριθμος de Casteljau

Ο αλγόριθμος de Casteljau, παρά την εκπληκτική απλότητα του, αποτελεί ένα βασικό εργαλείο για τη σχεδίαση καμπυλών (στο διδιάστατο ή τριδιάστατο χώρο) ή και επιφανειών (στον τριδιάστατο χώρο).

Πριν την παρουσίαση του γενικευμένου αλγορίθμου του de Casteljau, ας παρουσιάσουμε την ειδική εφαρμογή του για τρία σημεία (τα $\overrightarrow{r}_0$, $\overrightarrow{r}_1$, $\overrightarrow{r}_2$) όπως στο σχήμα 4.11.

Έστω t μια παράμετρος $t \in [0,1]$, με τη βοήθεια της οποίας και με αφετηρία τα τρία σημεία $\overrightarrow{r}_0$, $\overrightarrow{r}_1$ και $\overrightarrow{r}_2$ δημιουργούμε για κάθε τιμή της t δύο νέα σημεία ως εξής :

$$\begin{aligned} \overrightarrow{r}^1_0(t) &= (1-t)\ \overrightarrow{r}_0 + t\overrightarrow{r}_1 \\ \overrightarrow{r}^1_1(t) &= (1-t)\ \overrightarrow{r}_1 + t\overrightarrow{r}_2 \end{aligned} \tag{4.88}$$

Από τον τρόπο ορισμού του, το σημείο $r^1_0(t)$ ανήκει στο ευθύγραμμο τμήμα P_0P_1 και το σημείο $r^1_1(t)$ στο τμήμα P_1P_2. Με αφετηρία τα δύο νέα σημεία $\overrightarrow{r}^1_0$ και $\overrightarrow{r}^1_1$ που έστω προέκυψαν για μια τιμή της παραμέτρου t δημιουργούμε ένα τρίτο νέο σημείο με βάση τη σχέση

$$\overrightarrow{r}^2_0(t) = (1-t)\ \overrightarrow{r}^1_0 + t\overrightarrow{r}^1_1 \tag{4.89}$$

(για την ίδια τιμή της t) που, με τη σειρά του θα ανήκει στο ευθύγραμμο τμήμα $P^1_0P^1_1$. Με την εισαγωγή των εξισώσεων 4.88 στην 4.89 προκύπτει ότι

$$\overrightarrow{r}_0^2(t) \;=\; (1-t)^2\;\overrightarrow{r}_0 + 2t(t-1)\overrightarrow{r}_1 + t^2\overrightarrow{r}_2 \tag{4.90}$$

δηλαδή μια τετραγωνική έκφραση ως προς την παράμετρο t, γεγονός που αντικατοπτρίζεται στον άνω δείκτη του διανύσματος θέσης $\overrightarrow{r}_0^2$. Καθώς το t μεταβάλλεται γενικά στο $(-\infty, +\infty)$, ή ειδικά όπως εδώ ορίστηκε στο $[0,1]$ δημιουργείται και σχεδιάζεται μια παραβολή. Η παραβολή αυτή, όπως φάνηκε παραπάνω, σχηματίσθηκε από την επαναληπτική (αναδρομική) εφαρμογή σχημάτων γραμμικής παρεμβολής.

Με κατανοητό το απλό αυτό εισαγωγικό παράδειγμα, παρουσιάζουμε στη συνέχεια τον αναδρομικό τύπο του de Casteljau, για $N+1$ σημεία στο διδιάστατο ή τριδιάστατο χώρο, τα $\overrightarrow{r}_0, \ldots, \overrightarrow{r}_N$. Μαθηματικά εκφράζεται με μια και μόνη σχέση ως

$$\overrightarrow{r}_i^{\alpha}(t) \;=\; (1-t)\;\overrightarrow{r}_i^{\alpha-1}(t) \;+\; t\;\overrightarrow{r}_{i+1}^{\alpha-1}(t) \tag{4.91}$$

όπου θέτουμε ότι $\overrightarrow{r}_i^0(t) = \overrightarrow{r}_i\ ,\ i = 0, \ldots, N$ (δηλαδή τα δεδομένα $N+1$ σημεία που θα ονομάζονται πλέον *σημεία ελέγχου* ή σημεία *Bezier*). Αντίστοιχα, το πολυώνυμο που ορίζεται από την αλληλουχία σημείων $\overrightarrow{r}_0, \ldots, \overrightarrow{r}_N$ θα λέγεται *πολύγωνο Bezier*. Η σχέση 4.91, ως αναδρομικός τύπος προγραμματιζόμενος σε οποιαδήποτε γλώσσα προγραμματισμού αποτελείται από δύο βρόχους τον ένα μέσα στον άλλο. Ο εξωτερικός βρόχος έχει ως μεταβλητή την α ($\alpha = 1, \ldots, N$) και ο εσωτερικός έχει ως μεταβλητή την i και μεταβλητά όρια ($i = 0, \ldots, N-\alpha$). Η εκτέλεση των πράξεων που αντιστοιχούν στη σχέση 4.91 καταλήγουν στην καμπύλη $\overrightarrow{r}_0^N(t)$ που είναι η ζητούμενη καμπύλη Bezier για τα σημεία ελέγχου που προαναφέραμε.

Ο διπλός βρόχος της εξίσωσης 4.91 παριστάνεται εποπτικά με το τριγωνικό σχηματισμό (κάτω τριγωνικό μητρώο)

$$\begin{bmatrix} \overrightarrow{r}_0 & & & & \\ \overrightarrow{r}_1 & \overrightarrow{r}_0^1 & & & \\ \overrightarrow{r}_2 & \overrightarrow{r}_1^1 & \overrightarrow{r}_0^2 & & \\ & & \vdots & & \\ \overrightarrow{r}_N & \overrightarrow{r}_{N-1}^1 & \overrightarrow{r}_{N-2}^2 & \ldots & \overrightarrow{r}_0^N \end{bmatrix} \tag{4.92}$$

που ονομάζεται και σχηματισμός του de Casteljau. Δίνει δε την εποπτεία ότι, για μια τιμή του $t \in [0,1]$ αναδρομικά, η κάθε στήλη δημιουργεί την επομένη στα δεξιά της με ένα στοιχείο λιγότερο, καταλήγοντας στο $\overrightarrow{r}_0^N(t)$ που είναι και η ζητουμένη καμπύλη Bezier.

Μέχρι τώρα παρουσιάσαμε εποπτικά έναν αλγόριθμο, τον αλγόριθμο του de Casteljau που εύκολα και αναδρομικά παράγει μια καμπύλη Bezier, με αφετηρία $N+1$ διαθέσιμα σημεία ελέγχου. Κατά την υλοποίηση του αλγορίθμου de Casteljau σε οποιαδήποτε γλώσσα προγραμματισμού, θα προκύψει μια διδιάστατη ή τριδιάστατη καμπύλη (αν αντίστοιχα τα σημεία ελέγχου είναι στο επίπεδο ή στο χώρο) με τόσα σημεία $\overrightarrow{r}_0^N(t)$ όσες διακριτές τιμές της παραμέτρου $t \in [0,1]$ επιλεγούν.

Παρά όμως την απλότητα ενός τέτοιου αναδρομικού τύπου, θα ήταν επιθυμητό να υπάρχει διαθέσιμη και μια αναλυτική έκφραση για την καμπύλη Bezier, ώστε αυτή να

προκύπτει σημείο–προσ–σημείο με ευθεία αριθμητική αντικατάσταση αντί του αναδρομικού αλγορίθμου. Η μαθηματική αυτή έκφραση υπάρχει και χρησιμοποιεί τα λεγόμενα Bernstein πολυώνυμα. Ετσι η καμπύλη Bezier, ή σωστότερα η καμπύλη Bezier- Bernstein δίνεται από τη σχέση

$$\overrightarrow{r}_0^N(t) \;=\; \sum_{i=0}^{N} \overrightarrow{r}_i \; B_i^N(t) \tag{4.93}$$

ως γραμμικός συνδυασμός των διανυσμάτων θέσης των σημείων ελέγχου και με συντελεστές οι οποίοι προσδιορίζονται από τα πολυώνυμα Bernstein

$$B_i^N(t) \;=\; \left(\begin{array}{c} N \\ i \end{array} \right) \; t^i \; (1-t)^{N-i} \tag{4.94}$$

Οι σχέσεις 4.93 και 4.94 είναι έτοιμες να εφαρμοστούν - προγραμματιστούν για το σχηματισμό της καμπύλης Bezier. Υπενθυμίζεται ότι

$$\left(\begin{array}{c} \alpha \\ i \end{array} \right) \;=\; \frac{\alpha!}{i!(\alpha-i)!} \tag{4.95}$$

Η σχέση 4.94 είναι γενική, ισχύει δηλαδή για οποιαδήποτε τιμή του άνω δείκτη $\alpha = 0, \ldots, N$. Γράφεται δηλαδή ως

$$B_i^\alpha(t) \;=\; \left(\begin{array}{c} \alpha \\ i \end{array} \right) \; t^i \; (1-t)^{\alpha-i} \tag{4.96}$$

ενώ εύκολα αποδεικνύεται ότι ισχύει ο αναδρομικός τύπος

$$B_i^\alpha(t) \;=\; (1-t)B_i^{\alpha-1}(t) \;+ tB_{i-1}^{\alpha-1}(t) \tag{4.97}$$

με

$$B_0^0(t) \;\equiv\; 1 \tag{4.98}$$

και

$$B_j^\alpha(t) \;\equiv\; 0 \qquad , \qquad j \notin \{0, \ldots, N\} \tag{4.99}$$

Ο αναδρομικός τύπος 4.97 ουσιαστικά παραπέμπει στον αλγόριθμο de Casteljau. Η απόδειξη του 4.97 είναι σύντομη και παρατίθεται αμέσως παρακάτω

$$\begin{aligned} B_i^\alpha(t) &= \left(\begin{array}{c} \alpha \\ i \end{array} \right) t^i \, (1-t)^{\alpha-i} \; = \\ &= \left(\begin{array}{c} \alpha-1 \\ i \end{array} \right) t^i \, (1-t)^{\alpha-i} \; + \left(\begin{array}{c} \alpha-1 \\ i-1 \end{array} \right) t^i \, (1-t)^{\alpha-i} \\ &= (1-t) B_i^{\alpha-1}(t) \; + \; t B_{i-1}^{\alpha-1}(t) \end{aligned}$$

Βασική επίσης ιδιότητα των πολυωνύμων Bernstein είναι ότι

$$\sum_{i=0}^{N} B_i^N(t) \; \equiv \; 1 \tag{4.100}$$

Με τον ίδιο τρόπο που ο αλγόριθμος de Casteljau δημιουργεί το τριγωνικό σχηματισμό 4.92, οι συντελεστές–πολυώνυμα Bernstein (που λέγονται και *ενδιάμεσα πολυώνυμα Bernstein*) εντάσσονται και αυτοί σε ένα τριγωνικό μητρώο ως εξής :

$$\left[\begin{array}{ccccc} B_0^0(t) & B_0^1(t) & B_0^2(t) & \ldots & B_0^N(t) \\ & B_1^1(t) & B_1^2(t) & \ldots & B_1^N(t) \\ & & B_2^2(t) & \ldots & B_2^N(t) \\ & & \vdots & & \\ & & & & B_N^N(t) \end{array} \right] \tag{4.101}$$

με κάθε μη-μηδενικό συντελεστή να υπολογίζεται εφαρμόζοντας την αναδρομική σχέση 4.97.

Έστω για παράδειγμα η δημιουργία της καμπύλης Bezier που καθορίζουν τα τέσσερα σημεία ελέγχου P_0,P_1,P_2,P_3 ($N = 3$).

Η καμπύλη Bezier θα δίνεται τελικά από την εξίσωση 4.93 που εδώ ξαναγράφεται ως

$$\overrightarrow{r}_0^3(t) \; = \; \sum_{i=0}^{3} \overrightarrow{r}_i \; B_i^3(t)$$

και φυσικά η εφαρμογή της απαιτεί την εύρεση των τεσσάρων συντελεστών $B_0^3(t)$, $B_1^3(t)$, $B_2^3(t)$, $B_3^3(t)$. Με οδηγό το τριγωνικό μητρώο 4.101 και αφετηρία για τους υπολογισμούς τη σχέση 4.98 που καθορίζει μοναδιαία τιμή στο πάνω–αριστερό στοιχείο του έχουμε τους παρακάτω υπολογισμούς ανά στήλη:

- Δεύτερη στήλη:

$$\begin{aligned} B_0^1(t) &= \left(\begin{array}{c} 1 \\ 0 \end{array} \right) t^0 \, (1-t)^1 \; = \; 1-t \\ B_1^1(t) &= \left(\begin{array}{c} 1 \\ 1 \end{array} \right) t^1 \, (1-t)^0 \; = \; t \end{aligned}$$

με τιμές που ήταν αναμενόμενες αν ως βάση πάρουμε τον αλγόριθμο de Casteljau.

- Τρίτη στήλη:

$$\begin{aligned} B_0^2(t) &= \begin{pmatrix} 2 \\ 0 \end{pmatrix} t^0 (1-t)^2 = (1-t)^2 \\ B_1^2(t) &= \begin{pmatrix} 2 \\ 1 \end{pmatrix} t^1 (1-t)^1 = 2t(1-t) \\ B_2^2(t) &= \begin{pmatrix} 2 \\ 2 \end{pmatrix} t^2 (1-t)^0 = t^2 \end{aligned}$$

τα στοιχεία της οποίας συμβαδίζουν με τους συντελεστές της εξίσωσης 4.90

- Τέταρτη στήλη:

$$\begin{aligned} B_0^3(t) &= \begin{pmatrix} 3 \\ 0 \end{pmatrix} t^0 (1-t)^3 = (1-t)^3 \\ B_1^3(t) &= \begin{pmatrix} 3 \\ 1 \end{pmatrix} t^1 (1-t)^2 = 3t(1-t)^2 \\ B_2^3(t) &= \begin{pmatrix} 3 \\ 2 \end{pmatrix} t^2 (1-t)^1 = 3t^2(1-t) \\ B_3^3(t) &= \begin{pmatrix} 3 \\ 3 \end{pmatrix} t^3 (1-t)^0 = t^3 \end{aligned}$$

Με βάση τα παραπάνω, το μητρώο 4.101 παίρνει τη μορφή (τα στοιχεία που παραλείπονται είναι όλα μηδενικά)

$$\begin{bmatrix} 1 & (1-t) & (1-t)^2 & (1-t)^3 & (1-t)^4 & \dots \\ & t & 2t(1-t) & 3t(1-t)^2 & 4t(1-t)^3 & \dots \\ & & t^2 & 3t^2(1-t) & 6t^2(1-t)^2 & \dots \\ & & & t^3 & 4t^3(1-t) & \dots \\ & & & & t^4 & \dots \\ & & & & & \vdots \end{bmatrix} \tag{4.102}$$

Οι μέχρι τώρα σχέσεις που δόθηκαν για την καμπύλη Bezier είναι άμεσα υλοποιήσιμες σε γλώσσα προγραμματισμού. Για $N+1$ σημεία ελέγχου $\overrightarrow{r}_0, \dots, \overrightarrow{r}_N$, ο αναδρομικός τύπος 4.97 υπολογίζει τελικά τους $N+1$ συντελεστές $B_0^N(t), \dots, B_N^N(t)$ για οποιαδήποτε τιμή του $t \in [0,1]$ (για κάθε τιμή του t υπολογίζεται ένα διακριτό σημείο της καμπύλης, άρα βασικό αρχικό βήμα είναι να καθοριστεί μια αλληλουχία τιμών της παραμέτρου t για την οποία θα υπολογιστούν σημεία της καμπύλης Bezier – προφανώς το πλήθος αυτών είναι επίσης μια παράμετρος ελεύθερη να επιλεγεί από το χρήστη) και τελικά τα σημεία της καμπύλης υπολογίζονται από τη σχέση 4.93

συναρτήσει των συντεταγμένων των σημείων ελέγχου. Είναι σημαντικό να τονισθεί η 'ανεξαρτησία' και 'ομοιότητα' των εκφράσεων που παράγουν λ.χ. τις τρείς συντεταγμένες κάθε σημείου της καμπύλης Bezier. Είναι

$$\begin{aligned} x^N(t) &= x(t) = \sum_{i=0}^{N} B_i^N(t) x_i \\ y^N(t) &= y(t) = \sum_{i=0}^{N} B_i^N(t) y_i \\ z^N(t) &= z(t) = \sum_{i=0}^{N} B_i^N(t) z_i \end{aligned} \tag{4.103}$$

όπου $(x_i, y_i, z_i)\ ,\ i = 0, \ldots, N$ είναι οι τρείς συντεταγμένες κάθε σημείου Bezier, έστω στον τριδιάστατο χώρο.

Η Μητρωική Γραφή μιας Καμπύλης Bezier

Πολλοί συγγραφείς προτιμούν να ενσωματώνουν τα προηγούμενα σε μια βολική μητρωική γραφή η οποία καταλήγει να δίνει κάθε σημείο $\overrightarrow{r}_N$ της καμπύλης Bezier ως

$$\overrightarrow{r}_N(t) \equiv \overrightarrow{r}(t) = \sum_{i=0}^{N} \overrightarrow{r}_i\ C_i(t) \tag{4.104}$$

δηλαδή

$$\begin{aligned} x(t) &= \sum_{i=0}^{N} x_i C_i(t) \\ y(t) &= \sum_{i=0}^{N} y_i C_i(t) \\ z(t) &= \sum_{i=0}^{N} z_i C_i(t) \end{aligned} \tag{4.105}$$

όπου

$$\begin{bmatrix} C_0(t) \\ C_1(t) \\ \vdots \\ C_N(t) \end{bmatrix} = \begin{bmatrix} m_{0,0} & m_{0,1} & \ldots & m_{0,N} \\ m_{1,0} & m_{1,1} & \ldots & m_{1,N} \\ & & \vdots & \\ m_{N,0} & m_{N,1} & \ldots & m_{N,N} \end{bmatrix} \cdot \begin{bmatrix} t^0 \\ t^1 \\ \vdots \\ t^N \end{bmatrix} \tag{4.106}$$

με στοιχεία που ορίζονται ως

$$m_{i,j} \;=\; (-1)^{j-i} \left(\begin{array}{c} N \\ j \end{array} \right) \left(\begin{array}{c} j \\ i \end{array} \right) \tag{4.107}$$

Για παράδειγμα, για $N = 3$, το μητρώο έχει τη μορφή

$$\left[\begin{array}{rrrr} 1 & -3 & 3 & -1 \\ 0 & 3 & -6 & 3 \\ 0 & 0 & 3 & -3 \\ 0 & 0 & 0 & 1 \end{array} \right] \tag{4.108}$$

Έστω ότι χρησιμοποιώντας $N+1$ σημεία ελέγχου δημιουργούμε, κατά τα γνωστά πλέον, την αντίστοιχη καμπύλη Bezier. Η μητρωική γραφή που παρουσιάσαμε παραπάνω δείχνει εύκολα ότι

$$\overrightarrow{r}_N(t=0) \;=\; \overrightarrow{r}(0) \;=\; \overrightarrow{r}_0 \tag{4.109}$$

και

$$\overrightarrow{r}_N(t=1) \;=\; \overrightarrow{r}(1) \;=\; \overrightarrow{r}_N \tag{4.110}$$

δηλαδή ότι η καμπύλη Bezier ξεκινά από το πρώτο και καταλήγει στο τελευταίο σημείο ελέγχου.

Με οδηγό το μητρώο 4.102 και για $N+1$ σημεία ελέγχου δίνουμε την έκφραση της καμπύλης Bezier ως

$$\overrightarrow{r}_N(t) \;=\; \overrightarrow{r}(t) \;=\; (1-t)^N \overrightarrow{r}_0 \;+\; Nt(1-t)^{N-1} \overrightarrow{r}_1 \;+\; O(t^2) \tag{4.111}$$

οπότε, παραγωγίζοντας ως προς t, έχουμε

$$\begin{aligned} \frac{d\overrightarrow{r}(t)}{dt} \;&=\; -N(1-t)^{N-1} \overrightarrow{r}_0 + \\ &+\; N\left[(1-t)^{N-1} - (N-1)t(1-t)^{N-2}\right] \overrightarrow{r}_1 + O(t) \end{aligned} \tag{4.112}$$

από την οποία προκύπτει ότι, για $t = 0$, είναι

$$\frac{d\overrightarrow{r}(0)}{dt} \;=\; N\,(\overrightarrow{r}_0 \;-\; \overrightarrow{r}_1) \tag{4.113}$$

ενώ, για $t = 1$, είναι

$$\frac{d\overrightarrow{r}(1)}{dt} = N\,(\overrightarrow{r}_N - \overrightarrow{r}_{N-1}) \tag{4.114}$$

Οι σχέσεις 4.113 και 4.114 είναι χρήσιμες αφού αποδεικνύουν, ότι σε πρώτης τάξης ακρίβεια, το ευθύγραμμο τμήμα που συνδέει τα δύο πρώτα σημεία ελέγχου καθορίζει την κλίση της καμπύλης Bezier στην αφετηρία της ($t = 0$) και το ευθύγραμμο τμήμα που συνδέει τα δύο τελευταία σημεία ελέγχου καθορίζει την κλίση της στο τέλος της ($t = 1$).

Πέραν των μαθηματικών διατυπώσεων, τα σημεία ελέγχου μιας καμπύλης Bezier πρέπει να γίνουν κατανοητά ως *πόλοι έλξης* της καμπύλης. Μετακινώντας ένα οποιοδήποτε σημείο ελέγχου υπάρχει επίδραση σε ολόκληρη την καμπύλη ενώ η τάση είναι να παρατηρούμε την καμπύλη να μετακινείται (κυρίως τοπικά) προς τη νέα θέση του σημείου ελέγχου που μετακινήθηκε. Η καθολική επίδραση που έχει η μετακίνηση έστω και ενός σημείου ελέγχου στην καμπύλη Bezier αποτελεί συγχρόνως πλεονέκτημα και μειονέκτημα, ανάλογα πάντοτε με το σκοπό που χρησιμοποιεί κάποιος τις καμπύλες Bezier. Για παράδειγμα, αν το πολύγωνο Bezier έχει μια 'λογική λειότητα' με εξαίρεση ένα εσωτερικό του σημείο, το πλεονέκτημα είναι ότι το σημείο αυτό δεν θα επηρεάσει καθόλου τη λειότητα της καμπύλης Bezier. Ως μειονέκτημα (πάντοτε σε αναλογία με την αποσκοπούμενη χρήση) αναφέρεται το γεγονός ότι επιθυμώντας να τροποποιήσουμε κατά κάποιο τρόπο την καμπύλη Bezier αυτό αυτόματα υπαγορεύει την ανάγκη να μετακινήσουμε πολλά σημεία ελέγχου, όχι πάντοτε με προφανή τρόπο.

Εφαρμογή

Έστω μιά καμπύλη Bezier η οποία κατασκευάζεται με τέσσερα σημεία ελέγχου που ανήκουν στο ίδιο επίπεδο.

(α) Βρείτε τη σχετική θέση που αυτά πρέπει να έχουν ώστε η καμπύλη Bezier να είναι ευθεία.

(β) Απαντήστε ξανά στο προηγούμενο ερώτημα αν επιπλέον στόχος είναι η παραμετροποίηση της ευθείας που προκύπτει (μέσω της παραμέτρου t) να συμβαδίζει με την προφανή παραμετροποίηση της (μέσω ποσοστώσεων μήκους τόξου, μετρούμενου).

Λύση:

(α) Αφού η καμπύλη Bezier περνά από τα δύο ακραία σημεία ελέγχου (έστω $\overrightarrow{r}_0$ και $\overrightarrow{r}_3$) και η κλίση της στην αρχή καθορίζεται από το $\overrightarrow{r}_1 - \overrightarrow{r}_0$ ενώ στο τέλος από το $\overrightarrow{r}_3 - \overrightarrow{r}_2$, εύκολα γίνεται αντιληπτό ότι για να είναι ευθεία γραμμή η καμπύλη Bezier πρέπει τα τέσσερα σημεία ελέγχου να είναι συνευθειακά. Ας είναι

$$\begin{aligned}\overrightarrow{r}_1 &= \overrightarrow{r}_0 + \alpha(\overrightarrow{r}_3 - \overrightarrow{r}_0) = (1-\alpha)\overrightarrow{r}_0 + \alpha\overrightarrow{r}_3\\ \overrightarrow{r}_2 &= \overrightarrow{r}_0 + \beta(\overrightarrow{r}_3 - \overrightarrow{r}_0) = (1-\beta)\overrightarrow{r}_0 + \beta\overrightarrow{r}_3\end{aligned}$$

(β) Το ερώτημα αυτό ζητά ουσιαστικά τον υπολογισμό των α και β ώστε να προκύπτει η προφανής παραμετροποίηση του τμήματος $\overrightarrow{r}_3 - \overrightarrow{r}_0$. Έχοντας ήδη παράγει την μητρωική γραφή για $N = 3$ στην εξίσωση 4.108, έχουμε

$$\begin{aligned}\overrightarrow{r} &= (1 - 3t + 3t^2 - t^3)\overrightarrow{r}_0 +\\ &+ (3t - 6t^2 + 3t^3)\overrightarrow{r}_1 +\\ &+ (3t^2 - 3t^3)\overrightarrow{r}_2 +\\ &+ t^3\overrightarrow{r}_3\end{aligned}$$

που με αντικατάσταση δίνει

$$\begin{aligned}\overrightarrow{r} &= (1 - 3t + 3t^2 - t^3)\overrightarrow{r}_0 +\\ &+ (1-\alpha)(3t - 6t^2 + 3t^3)\overrightarrow{r}_0 + \alpha(3t - 6t^2 + 3t^3)\overrightarrow{r}_3 +\\ &+ (1-\beta)(3t^2 - 3t^3)\overrightarrow{r}_0 + \beta(3t^2 - 3t^3)\overrightarrow{r}_3 +\\ &+ t^3\overrightarrow{r}_3\\ &= \left[1 - 3\alpha t + 3(2\alpha - \beta)t^2 + (3\beta - 3\alpha - 1)t^3\right]\overrightarrow{r}_0 +\\ &+ \left[3\alpha t + 3(\beta - 2\alpha)t^2 + (3\alpha - 3\beta + 1)\right]\overrightarrow{r}_3\end{aligned}$$

Η προφανής παραμετροποίηση της ευθείασ–καμπύλης Bezier είναι η

$$\overrightarrow{r}(t) = (1-t)\overrightarrow{r}_0 + t\overrightarrow{r}_3$$

άρα πρέπει να ισχύουν οι σχέσεις

$$\begin{aligned}3\alpha &= 1\\ 2\alpha &= \beta\\ 3\beta &= 1 + 3\alpha\end{aligned}$$

που επιδέχονται τις λύσεις

$$\alpha = \frac{1}{3} \quad , \quad \beta = \frac{2}{3}$$

Η λύση καθορίζει την αναμενόμενη θέση των σημείων ελέγχου ως

$$\overrightarrow{r}_0 \;,\; \overrightarrow{r}_0 + \frac{1}{3}(\overrightarrow{r}_3 - \overrightarrow{r}_0) \;,\; \overrightarrow{r}_0 + \frac{2}{3}(\overrightarrow{r}_3 - \overrightarrow{r}_0) \;,\; \overrightarrow{r}_3$$

Πολυώνυμα Bezier–Bernstein – Προγραμματισμός – Παραδείγματα – Συζήτηση

Η ενότητα αυτή έχει στόχο να παρουσιάσει έναν κώδικα προγραμματισμένο σε Fortran 77 ο οποίος εφαρμόζει όσα παρουσιάστηκαν προηγούμενα σε σχέση με τις ιδιότητες και τον τρόπο χρήσης των πολυωνύμων Bezier–Bernstein ως εργαλεία αριθμητικής παρεμβολής.

Προτιμήθηκε ο προγραμματισμός να γίνει απλά (και όχι 'βέλτιστα') ώστε να συμβαδίζει με τη θεωρία που προηγήθηκε.

Το κυρίως πρόγραμμα διαβάζει έναν πλήθος σημείων ελέγχου καταχωρημένων στο αρχείο bezier.dat. Κάθε γραμμή του αρχείου περιέχει, σε ελεύθερη διαμόρφωση, τις δύο συντεταγμένες ενός σημείου ελέγχου. Το πλήθος των σημείων ελέγχου 'μετράται' κατά την ανάγνωση του αρχείου δεδομένων, όπως φαίνεται από τη σύνταξη της εντολής read. Η σειρά με την οποία περιγράφονται τα σημεία στο αρχείο είναι και η σειρά τους στο πολύγωνο Bezier.

Το πρόγραμμα ακολουθεί τη μητρωική γραφή της καμπύλης Bezier εφαρμόζοντας όσα περιγράφονται στις σχέσεις 4.104, 4.105, 4.106 και 4.107. αποτελείται δε από δύο διακριτά τμήματα που αντιστοιχούν στα υποπρογράμματα inibezier και usebezier.

Στο υποπρόγραμμα inibezier υπολογίζονται τα στοιχεία του τετραγωνικού μητρώου της σχέσης 4.106, εφαρμόζοντας τις σχέσεις 4.107. Ας σημειωθεί ότι αυτός ο υπολογισμός εξαρτάται μόνο από την τάξη του πολυωνύμου Bezier, δηλαδή το αποτέλεσμά του είναι το ίδιο πάντα μητρώο για το ίδιο πλήθος σημείων ελέγχου.

Το υποπρόγραμμα usebezier δημιουργεί μια κατανομή τιμών της παραμέτρου t στο διάστημα $[0, 1]$ (εδώ, η κατανομή είναι πολύ απλή και αποτελείται από *ideg* ισαπέχοντα σημεία, την τιμή του *ideg* δίνει ο χρήστης διαλογικά). Προφανώς με κατάλληλη παρέμβαση του χρήστη στην αρχή της υπορουτίνας μπορεί να αντικατασταθεί με οποιαδήποτε άλλη κατανομή του t.

```
c*******************************************************
      program test_bezier
c*******************************************************
      implicit double precision (a-h,o-z)
      dimension val(1000),x(1000),y(1000)
      dimension xb(25),yb(25)
      common /bez1/ bezm(25,25)
c
      open(1,file='bezier.dat')
      do i=1,26
  read(1,*,end=10)xb(i),yb(i)
      enddo
      stop  'Increase dimension (xb.yb)'
  10  nb=i-1
      close(1)
c
```

```
      call inibezier(nb)
c
      write(*,*)' Type the number of points along the final curve'
      read(*,*)ideg
      call usebezier(xb,yb,nb,x,y,ideg-1)
c
      open(1,file='curve')
      do i=1,ideg
      write(1,'(2(3x,f12.7))')x(i),y(i)
      enddo
      close(1)
c
      stop
      end
c
c
c ***********************************************************************
      subroutine inibezier(nco)
c ***********************************************************************
c     NCO control points
c
      implicit double precision (a-h,o-z)
      common /bez1/ bezm(25,25)
c
      do 1 mi=0,nco-1      !  for the control points
        b=0.d0
        c=0.d0
          do 1 i=0,nco-1
            call paragon (nco-1,i ,kres1)
            call paragon (i    ,mi,kres2)
            kres3=(-1)**(i-mi)
            coeffi = dfloat(kres1*kres2*kres3)
            if(mi.gt.i) coeffi=0.d0
            bezm(mi+1,i+1) = coeffi
  1   continue
c
      return
      end
c
c
c ***********************************************************************
      subroutine paragon (n,i,k)  ! n=UP, i=LOW, k=RESULT
c ***********************************************************************
      implicit double precision (a-h,o-z)
```

```
c
      ks=max(i,n-i)+1
      kp=min(i,n-i)
      k=1
      do iii=ks,n
        k=k*iii
      enddo
      do iii=1,kp
        k=k/iii
      enddo
      return
      end
c
c
c ***********************************************************************
      subroutine usebezier   (xco,yco,nco,x,y,ideg)
c ***********************************************************************
c     NCO control points (xco,yco)
c     IDEG+1 final points with coordinates (X,Y)
c
      implicit double precision (a-h,o-z)
      common /bez1/ bezm(25,25)
      dimension x(1),y(1),xco(1),yco(1)
c
      aa1  =  0.5d0
      dd   =  0.1
      do k=1,ideg+1
      y(k)=dfloat(k-1)/dfloat(ideg)
      enddo
c
      do 10 kpoi=0,ideg
      kpoi1 = kpoi+1
      tlocal = y(kpoi1)
      x  (kpoi1) = 0.d0
      y  (kpoi1) = 0.d0
        do mi=0,nco-1
        b=0.d0
          do  i=0,nco-1
            b = b + bezm(mi+1,i+1) * tlocal**i
          enddo  ! i
        x(kpoi1)   = x  (kpoi1) + b*xco(mi+1)
        y(kpoi1)   = y  (kpoi1) + b*yco(mi+1)
      enddo  !  mi
  10  continue
```

```
c
      return
      end
c
c
```

Χρησιμοποιώντας το παραπάνω λογισμικό και διάφορα αρχεία δεδομένων μπορούμε να παρακολουθήσουμε και να κατανοήσουμε τη συμπεριφορά των πολυωνύμων Bezier:

1. Με το αρχείο δεδομένων (αρχείο bezier.dat) των 8 σημείων ελέγχου που δίνεται παρακάτω

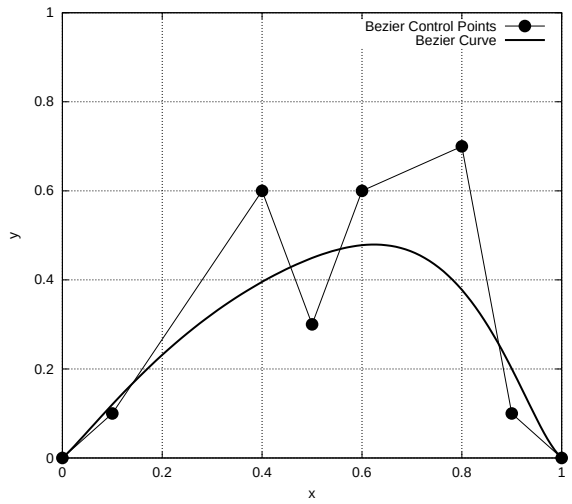
0.0	0.0
0.1	0.1
0.4	0.6
0.5	0.3
0.6	0.6
0.8	0.7
0.9	0.1
1.0	0.0

   δημιουργείται μια καμπύλη Bezier με 101 σημεία, για τιμές του t που ισαπέχουν. Στο Σχήμα 4.12 φαίνονται τα σημεία ελέγχου (παρουσιάζονται ως διακριτά σημεία ενωμένα με λεπτή γραμμή για να φανεί το πολύγωνο Bezier. Όπως και σε κάθε επόμενο σχήμα, μπορεί να γίνει εύκολα αντιληπτό ότι την κλίση της καμπύλης στην αρχή και το τέλος την καθορίζουν το πρώτο–δεύτερο και το τελευταίο–προτελευταίο σημείο ελέγχου. Παρατηρούμε ότι το τέταρτο σημείο, παρότι χαμηλότερα από τα υπόλοιπα, δεν μπορεί να επηρεάσει αισθητά τη μορφή της καμπύλης.

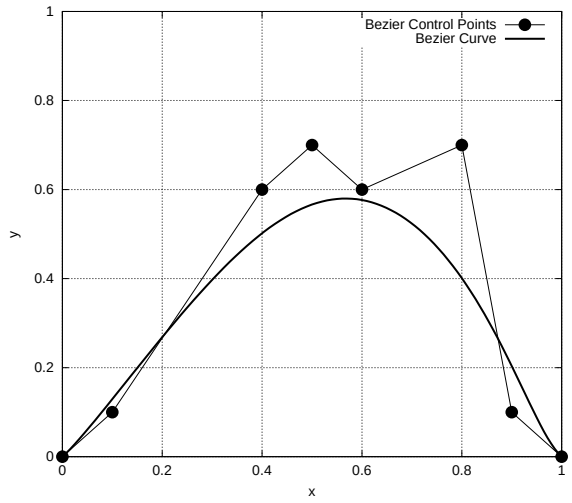
2. Ας επιβεβαιώσουμε την τελευταία παρατήρηση αυξάνοντας την τιμή της τεταγμένης του τέταρτου σημείου. Το αρχείο δεδομένων γίνεται

0.0	0.0
0.1	0.1
0.4	0.6
0.5	0.7
0.6	0.6
0.8	0.7
0.9	0.1
1.0	0.0

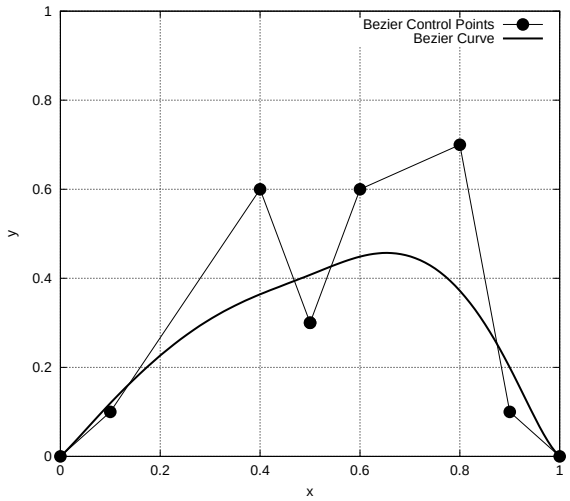
   και το αποτέλεσμα απεικονίζεται στο Σχήμα 4.13. Παρά την αισθητή μεταβολή στη θέση του τέταρτου σημείου, παρατηρούμε τη σχετική μεταβολή στο σχήμα της καμπύλης Bezier.



Σχήμα 4.12: Παράδειγμα προσέγγισης με πολυώνυμα Bezier.



Σχήμα 4.13: Παράδειγμα προσέγγισης με πολυώνυμα Bezier.



Σχήμα 4.14: Παράδειγμα προσέγγισης με πολυώνυμα Bezier.

3. Επαναφέρουμε την τεταγμένη του τέταρτου σημείου στην πρότερη τιμή του, αλλά προσθέτουμε ακόμα ένα σημείο ελέγχου, ίδιων συντεταγμένων με το τέταρτο. Το αρχείο δεδομένων γίνεται

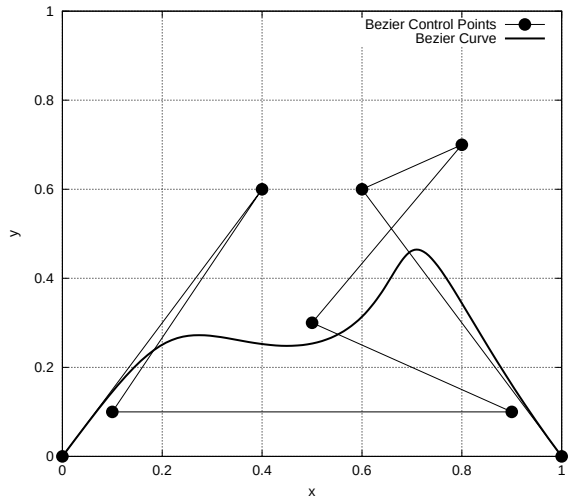
0.0	0.0
0.1	0.1
0.4	0.6
0.5	0.3
0.5	0.3
0.6	0.6
0.8	0.7
0.9	0.1
1.0	0.0

και το αποτέλεσμα απεικονίζεται στο Στο Σχήμα 4.14. Το διπλό σημείο φαίνεται να έλκει την καμπύλη Bezier προς αυτό.

4. Στη συνέχεια, δημιουργούμε ένα αρχείο δεδομένων αλλάζοντας τυχαία τη σειρά των σημείων ελέγχου. Τότε, με το αρχείο

0.0	0.0
0.4	0.6
0.1	0.1
0.9	0.1
0.5	0.3
0.8	0.7
0.6	0.6
1.0	0.0

το αποτέλεσμα απεικονίζεται στο Σχήμα 4.15.

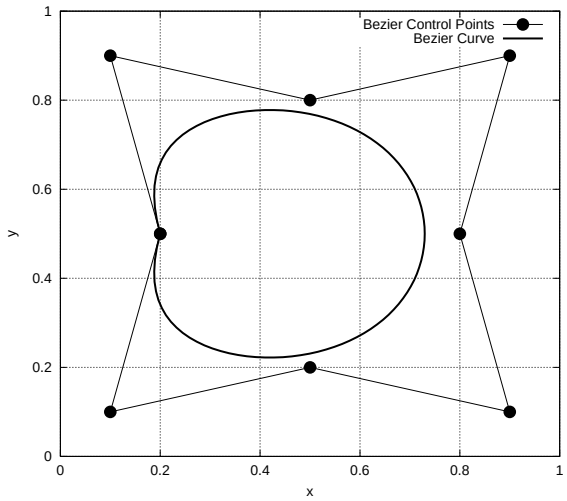


Σχήμα 4.15: Παράδειγμα προσέγγισης με πολυώνυμα Bezier.

5. Τα επόμενα δύο παραδείγματα αφορούν κλειστές καμπύλες που σχηματίζονται από πολυώνυμα Bezier. Ο τρόπος είναι απλός, αρκεί να επαναλάβουμε το πρώτο σημείο ελέγχου και ως τελευταίο. Αφού μια καμπύλη Bezier διέρχεται από το πρώτο και το τελευταίο σημείο, το αποτέλεσμα θα είναι μια καμπύλη με ταυτιζόμενα τα σημεία αρχής–τέλους, άρα μια κλειστή καμπύλη. Με βάση το αρχείο δεδομένων που ακολουθεί,

0.2	0.5
0.1	0.9
0.5	0.8
0.9	0.9
0.8	0.5
0.9	0.1
0.5	0.2
0.1	0.1
0.2	0.5

το πολύγωνο Bezier είναι συμμετρικό ως προς το σημείο $(0.5, 0.5)$. Όπως όμως φαίνεται και στο Στο 4.16, η καμπύλη που τελικά προκύπτει δεν είναι συμμετρική ως προς το ίδιο σημείο. Ο λόγος είναι το ότι η καμπύλη πρέπει να περάσει από τα δύο ακραία σημεία. Είναι δε ιδιαίτερα σημαντικό να παρατηρηθεί ότι στο σημείο ένωσης, δηλαδή στο $(0.2, 0.5)$, η καμπύλη παρουσιάζει ασυνέχεια πρώτης παραγώγου. Αυτό ήταν αναμενόμενο, αφού στο ίδιο σημείο η κλίση από τη μια πλευρά (στην αφετηρία της καμπύλης) καθορίζεται από το ευθύγραμμο τμήμα που ενώνει τα σημεία $(0.2, 0.5)$ και $(0.1, 0.9)$ ενώ από την άλλη πλευρά (στον τερματισμό της καμπύλης) αυτή καθορίζεται από το ευθύγραμμο τμήμα που ενώνει τα σημεία $(0.2, 0.5)$ και $(0.1, 0.1)$. Τα δύο αυτά τμήματα δεν είναι συνευθειακά, άρα οι κλίσεις τους είναι διαφορετικές.

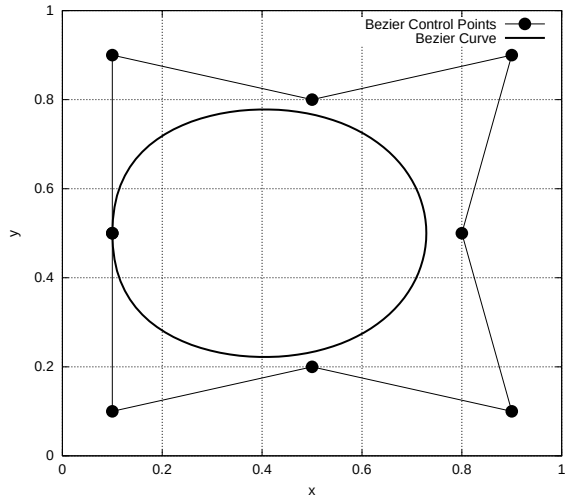


Σχήμα 4.16: Παράδειγμα προσέγγισης με πολυώνυμα Bezier.

6. Τέλος προκαλούμε τη συνέχεια πρώτης παραγώγου στο πρώτο σημείο, αλλάζοντας τις συντεταγμένες του (την τετμημένη του). Το νέο αρχείο δεδομένων είναι το

0.1	0.5
0.1	0.9
0.5	0.8
0.9	0.9
0.8	0.5
0.9	0.1
0.5	0.2
0.1	0.1
0.1	0.5

και όπως φαίνεται και στο Σχήμα 4.17, αφού τα σημεία $(0.1, 0.5)$, $(0.1, 0.9)$ και $(0.1, 0.1)$ είναι συνευθειακά, το αποτέλεσμα είναι μια καμπύλη που έχει σε κάθε σημείο της συνέχεια πρώτης παραγώγου.



Σχήμα 4.17: Παράδειγμα προσέγγισης με πολυώνυμα Bezier.

Κεφάλαιο 5

Αριθμητική Ολοκλήρωση και Παραγώγιση

Η ανάγκη ολοκλήρωσης και παραγώγισης συναρτήσεων εμφανίζεται σε πάρα πολλές μηχανολογικές εφαρμογές και προβλήματα. Η συνηθέστερη χρήση του ολοκληρώματος γίνεται για τον υπολογισμό της συνολικής ποσότητας ή της μέσης τιμής μιας συνάρτησης σε μια γραμμή, μια επιφάνεια ή έναν όγκο. Για παράδειγμα, η εύρεση του κέντρου βάρους ενός σώματος, του σημείου εφαρμογής της συνισταμένης δύναμης λόγω αεροδυναμικής αντίστασης σε ένα σώμα, της ροής μάζας ή θερμότητας μέσω μιας επιφάνειας, της απόστασης που διανύει ένα σώμα σε συγκεκριμένο χρόνο κλπ., μπορεί να γίνει ολοκληρώνοντας τη μαθηματική σχέση που εκφράζει την αντίστοιχη ιδιότητα ή φαινόμενο ως συνάρτηση μιας η περισσότερων ανεξάρτητων μεταβλητών (απόσταση, χρόνος).

Από την άλλη μεριά, ο ρυθμός μεταβολής ενός μεγέθους στον χώρο ή στον χρόνο, που εκφράζεται με την παράγωγό του, εμπεριέχεται σε πλείστους φυσικούς νόμους και θεωρήματα. Για παράδειγμα, η ταχύτητα και η επιτάχυνση ενός σώματος εκφράζουν αντίστοιχα τη μεταβολή της θέσης και της ταχύτητάς του στη μονάδα του χρόνου, ενώ η μεταφορά θερμότητας από ένα θερμότερο προς ένα ψυχρότερο σημείο ενός μέσου είναι ανάλογη της παραγώγου (ή της κλίσης) της θερμοκρασίας μεταξύ των σημείων αυτών.

Όταν μια συνάρτηση εκφράζεται με απλή μαθηματική σχέση, τότε η ολοκλήρωση ή παραγώγισή της μπορεί συνήθως να γίνει αναλυτικά. Σε αντίθετη περίπτωση είναι απαραίτητη η εφαρμογή αριθμητικών μεθόδων, η περιγραφή των οποίων αποτελεί το αντικείμενο του παρόντος κεφαλαίου.

5.1. Αριθμητική Ολοκλήρωση

Η πράξη ορισμένης ολοκλήρωσης (definite integration) μιας συνάρτησης f της ανεξάρτητης μεταβλητής x εκφράζεται μαθηματικά ως εξής:

$$I = \int_a^b f(x)dx \tag{5.1}$$

και αποτελεί το άθροισμα των τιμών $f(x)dx$ στην περιοχή τιμών του x από a έως b. Το ορισμένο ολοκλήρωμα αντιστοιχεί γραφικά στο εμβαδόν μεταξύ της συνάρτησης και του άξονα των x καθώς και των γραμμών $x = a$ και $x = b$, όπως φαίνεται στο παράδειγμα του Σχήματος 5.1α.

Το διάστημα $[a, b]$ μπορεί να χωρισθεί σε n υποδιαστήματα χρησιμοποιώντας n+1 σημεία (x_0, x_1, x_2, … x_n), όπως στο παράδειγμα του Σχήματος 5.1β, οπότε το ορισμένο ολοκλήρωμα παίρνει τη διακριτή μορφή

$$I = \sum_{i=0}^{n-1} \left(\int_{x_i}^{x_{i+1}} f(x)dx \right) \tag{5.2}$$

Όταν το ολοκλήρωμα μιας συνάρτησης *f(x)* μπορεί να εκφρασθεί αναλυτικά με μια άλλη συνάρτηση *F(x),* τότε το ορισμένο ολοκλήρωμά της υπολογίζεται άμεσα, με βάση τη γενική σχέση ολοκλήρωσης

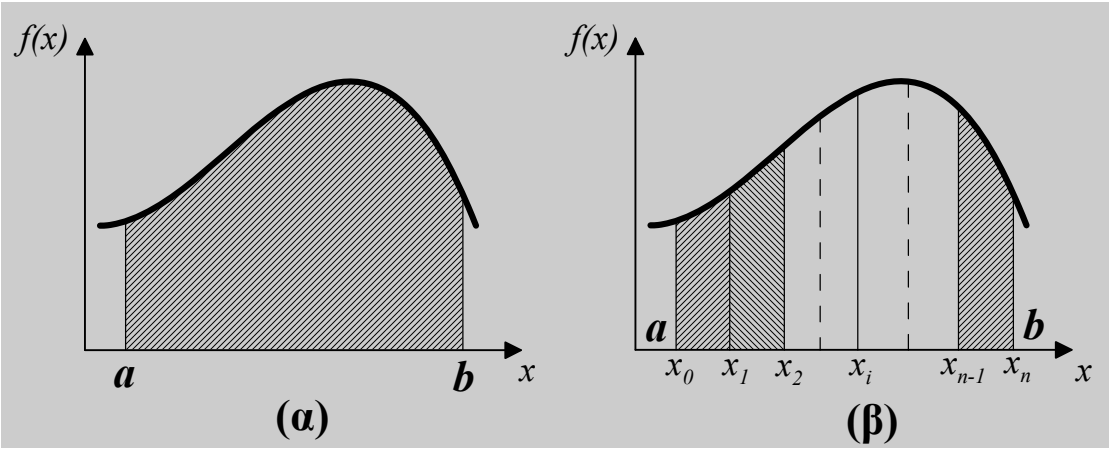
$$I = \int_a^b f(x)dx = F(x)\Big|_a^b = F(b) - F(a) \tag{5.3}$$

Για παράδειγμα, το ολοκλήρωμα της πολυωνυμικής συνάρτησης

$$f(x) = -160x^5 + 365x^4 - 270x^3 + 60x^2 + 5x + 1 \tag{5.4}$$

στο διάστημα [0, 1] είναι

$$I = \left(-\frac{160}{6}x^6 + \frac{365}{5}x^5 - \frac{270}{4}x^4 + \frac{60}{3}x^3 + \frac{5}{2}x^2 + 1x \right)\Big|_0^1 = 2.3333333$$



Σχήμα 5.1. Γράφημα του ορισμένου ολοκληρώματος μιας συνάρτησης: α) ενιαίο εμβαδόν και β) άθροισμα εμβαδών *n* υποδιαστημάτων.

Πολύ συχνά όμως προκύπτουν σε πρακτικές εφαρμογές συναρτήσεις που δεν έχουν αναλυτικό ολοκλήρωμα ή είναι τόσο πολύπλοκες ώστε η εύρεση της αναλυτικής λύσης να είναι πολύ δύσκολη. Σε τέτοιες περιπτώσεις ο υπολογισμός του ορισμένου ολοκληρώματος μπορεί να γίνει μόνο με αριθμητικές μεθόδους. Το ίδιο ισχύει και για την περίπτωση που μια συνάρτηση δεν έχει αναλυτική έκφραση, αλλά αποτελείται από διακριτά ζεύγη τιμών *x* και *f(x)*, όπως είναι συνήθως τα δεδομένα μετρήσεων (πινακοποιημένη συνάρτηση).

Η βασική ιδέα σε όλες τις τεχνικές αριθμητικής ολοκλήρωσης που θα περιγραφούν στη συνέχεια, είναι η προσέγγιση μιας συνάρτησης ή μιας σειράς διακριτών δεδομένων με χρήση απλών πολυωνύμων, συνήθως έως τρίτου βαθμού, των οποίων το ολοκλήρωμα μπορεί να υπολογισθεί άμεσα. Τα πολυώνυμα αυτά ορίζονται έτσι ώστε να διέρχονται από σημεία (*x, f*(*x*)), στα όρια του διαστήματος [*a*, *b*] ή/και σε ενδιάμεσες θέσεις, αναλόγως του επιθυμητού βαθμού (π.χ. ένα πολυώνυμο $2^{ου}$ βαθμού ορίζεται από τρία σημεία). Επομένως, *ακόμα και οι αναλυτικές συναρτήσεις μετατρέπονται πρώτα σε διακριτά δεδομένα*.

Στις περισσότερες τεχνικές είναι απαραίτητο τα διακριτά αυτά δεδομένα να διατίθενται σε ισαπέχοντα σημεία κατά *x*. Αυτό δεν δημιουργεί βέβαια δυσκολία στην

περίπτωση αναλυτικών συναρτήσεων, όπου μπορεί να βρεθεί η τιμή *f(x)* σε οποιοδήποτε σημείο. Όταν όμως η συνάρτηση δεν είναι αναλυτική και τα διαθέσιμα δεδομένα δεν ισαπέχουν, οι περισσότερες μέθοδοι δεν είναι εφαρμόσιμες.

Οι τεχνικές αριθμητικής ολοκλήρωσης που θα αναλυθούν στη συνέχεια χωρίζονται σε δύο κατηγορίες. Η πρώτη περιλαμβάνει τις γενικές μεθόδους, που χρησιμοποιούνται σε αναλυτικές και μη συναρτήσεις. Στη δεύτερη κατηγορία ανήκουν μερικές πιο εξελιγμένες τεχνικές, οι οποίες είναι εφαρμόσιμες μόνο σε εξισώσεις, επειδή παράγουν και χρησιμοποιούν διάφορες τιμές της συνάρτησης για αύξηση της ακρίβειας.

5.1.1. Γενικές Μέθοδοι Ολοκλήρωσης

5.1.1.1. Περιγραφή

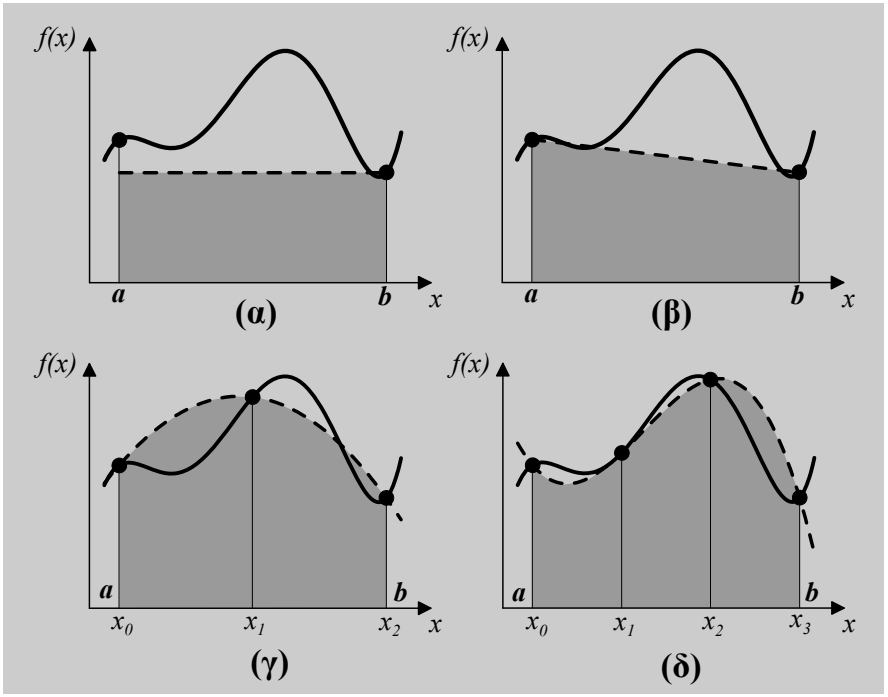
Στην κατηγορία αυτή ανήκουν οι ευρέως χρησιμοποιούμενες μέθοδοι του *Τραπεζίου* και του *Simpson*. Όπως και στην αριθμητική προσέγγιση (βλ. Κεφ. 4.1), η αναλυτική ή πινακοποιημένη συνάρτηση $f(x)$ προσεγγίζεται με μια απλή, αναλυτική συνάρτηση, που εδώ είναι ένα πολυώνυμο $p_m(x)$ βαθμού m, οπότε ισχύει

$$I = \int_a^b f(x)dx \cong \int_a^b p_m(x)dx \tag{5.5}$$

Έτσι προκύπτουν οι ακόλουθες περιπτώσεις:

α) Για $m = 0$, το πολυώνυμο είναι μία οριζόντια γραμμή που διέρχεται από το σημείο $f(a)$ ή το $f(b)$ (Σχ. 5.2α). Στη μορφή αυτή, που αναφέρεται και ως *μέθοδος των ορθογωνίων τμημάτων*, το ορισμένο ολοκλήρωμα του πολυωνύμου είναι

$$I = (b-a)\ f(a) \qquad \text{ή} \qquad I = (b-a)\ f(b) \tag{5.6}$$



Σχήμα 5.2. Παράδειγμα προσέγγισης συνάρτησης *f(x)* με πολυώνυμο: α) μηδενικού βαθμού, β) πρώτου βαθμού, γ) δεύτερου βαθμού και δ) τρίτου βαθμού.

β) Για $m = 1$, το πολυώνυμο είναι μια ευθεία γραμμή που συνδέει τα σημεία $f(a)$ και $f(b)$ (Σχ. 5.2β) και έχει την ακόλουθη μαθηματική έκφραση:

$$p_1(x) = \frac{f(b) - f(a)}{b - a}\,(x - a) + f(a) \tag{5.7}$$

Το ορισμένο ολοκλήρωμα του $p_1(x)$ θα είναι

$$I = (b - a)\,\frac{f(a) + f(b)}{2} \tag{5.8}$$

Το ολοκλήρωμα αυτό ισούται με το εμβαδόν της επιφάνειας κάτω από το γράφημα του πολυωνύμου (Σχ. 5.2β), που είναι ένα τραπέζιο με βάσεις $f(a)$ και $f(b)$. Έτσι, αυτή η μορφή αποτελεί τη *μέθοδο* (ή *κανόνα*) *του Τραπεζίου*.

γ) Για $m = 2$ απαιτείται ένα ακόμη σημείο για να ορισθεί το πολυώνυμο (παραβολή), επομένως η περιοχή ολοκλήρωσης πρέπει να περιέχει δύο ίσα υποδιαστήματα (Σχ. 5.2γ). Το ορισμένο ολοκλήρωμα του 2βάθμιου αυτού πολυωνύμου, το οποίο διέρχεται από τα σημεία $x_0 \equiv a$, $x_1 = (a + b)/2$ και $x_2 \equiv b$ προκύπτει αναλυτικά (Εξ. 5.3) ως εξής:

$$I = \int_a^b p_2(x)dx = \int_{x_0}^{x_2} (\alpha_2 x^2 + \alpha_1 x + \alpha_0)dx = \left(\frac{1}{3}\alpha_2 x^3 + \frac{1}{2}\alpha_1 x^2 + \alpha_0 x\right)\Bigg|_{x_0}^{x_2} =$$

$$= \frac{(x_2 - x_0)}{6}\left[2\alpha_2(x_2^2 + x_0 x_2 + x_0^2) + 3\alpha_1(x_2 + x_0) + 6a_0\right] = \frac{(b-a)}{6}\Big[(\alpha_2 x_0^2 + a_1 x_0 + a_0)$$

$$+ (a_2 x_2^2 + a_1 x_2 + a_0) + \alpha_2(x_0^2 + x_2^2 + 2x_0 x_2) + 2a_1(x_0 + x_2) + 4a_0\Big] =$$

$$= \frac{(b-a)}{6}\left[f(x_0) + f(x_2) + 4a_2\left(\frac{x_0 + x_2}{2}\right)^2 + 4a_1\left(\frac{x_0 + x_2}{2}\right) + 4a_0\right] \Rightarrow$$

$$I = (b - a)\,\frac{f(x_0) + 4f(x_1) + f(x_2)}{6} \tag{5.9}$$

και ισοδυναμεί με το εμβαδόν της επιφάνειας στο Σχήμα 5.2γ. Τα σημεία $f(a)$ και $f(b)$ θα αναφέρονται στη συνέχεια και ως $f(x_0)$ και $f(x_n)$, όπου n+1 ο συνολικός αριθμός των διακριτών δεδομένων. Αυτή είναι η *μέθοδος Simpson 1/3*, όπου το 1/3 είναι ο αριθμητικός συντελεστής που προκύπτει εάν τεθεί $(b - a) = 2h$, με h το πλάτος ενός υποδιαστήματος.

δ) Για $m = 3$ το πολυώνυμο είναι τρίτου βαθμού, συνεπώς απαιτούνται τέσσερα ισαπέχοντα σημεία και τρία υποδιαστήματα (Σχ. 5.2δ). Η αναλυτική έκφραση που προκύπτει τελικά για το ορισμένο ολοκλήρωμα του $p_3(x)$ είναι:

$$I = (b - a)\,\frac{f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)}{8} \tag{5.10}$$

και αποτελεί τη *μέθοδο Simpson 3/8* (αφού $(b - a) = 3h$).

Όπως φαίνεται και στα παραδείγματα του Σχήματος 5.2, όσο αυξάνει ο βαθμός του πολυωνύμου, τόσο καλλίτερα προσεγγίζει την αρχική συνάρτηση, επομένως το σφάλμα του αριθμητικού υπολογισμού μειώνεται.

5.1.1.2. Εκτίμηση αριθμητικού σφάλματος

Το αριθμητικό σφάλμα αποκοπής εκφράζει τη διαφορά μεταξύ της πραγματικής και της αριθμητικής τιμής του ολοκληρώματος:

$$E_t = \int_a^b f(x)dx \; - \; \int_a^b p_m(x)dx \tag{5.11}$$

και μπορεί να εκτιμηθεί κατά προσέγγιση, με βάση το ανάπτυγμα της συνάρτησης σε σειρά Taylor ή ολοκληρώνοντας κάποιο πολυώνυμο παρεμβολής της. Έτσι, για την μέθοδο του Τραπεζίου χρησιμοποιείται ένα πολυώνυμο πρόσω παρεμβολής Newton, η έκφραση πρώτης τάξης του οποίου για μια συνάρτηση $f(x)$ είναι

$$f(x) = f(x_0) + \Delta f(x_0)\cdot\delta + \frac{f''(\xi)}{2!}h^2\,\delta(\delta-1)$$

όπου $\Delta f(x_0) = f(x) - f(x_0)$, $\delta = (x - x_0)/h$, και h είναι η απόσταση των διακριτών δεδομένων. Εάν θεωρηθεί κατά προσέγγιση ο όρος $f''(\xi)$ σταθερός στο διάστημα $[a, b]$, τότε η παραπάνω έκφραση μπορεί να ολοκληρωθεί αναλυτικά, λαμβάνοντας υπόψη ότι $h = a - b$, καθώς και $dx = h\cdot d\delta$ με $\delta \in [0, 1]$ όταν $x \in [a, b]$. Έτσι προκύπτει

$$I = \int_a^b f(x)dx \cong h\cdot\int_0^1 \left[f(a) + \Delta f(a)\cdot\delta + \frac{f''(\xi)}{2!}h^2\,\delta(\delta-1)\right]d\delta =$$

$$= h\left[\delta\cdot f(a) + \Delta f(a)\cdot\frac{\delta^2}{2} + \frac{f''(\xi)}{2!}h^2\left(\frac{\delta^3}{3} - \frac{\delta^2}{2}\right)\right]\Bigg|_0^1 = h\left[f(a) + \frac{\Delta f(a)}{2}\right] - \frac{f''(\xi)}{12}h^3 \;\Rightarrow$$

$$I = (b-a)\frac{f(a)+f(b)}{2} - \frac{1}{12}(b-a)^3 f''(\xi)$$

Συγκρίνοντας την παραπάνω έκφραση με την Εξ. (5.8), προκύπτει από την Εξ. (5.11) ότι το απόλυτο σφάλμα της μεθόδου του Τραπεζίου είναι

$$\text{Μέθοδος Τραπεζίου:} \quad E = -\frac{1}{12}(b-a)^3 f''(\xi) \tag{5.12}$$

Με ανάλογη διαδικασία προκύπτει και το σφάλμα των άλλων μεθόδων, ως εξής:

$$\text{Μέθοδος των ορθογωνίων:} \quad E = \pm\frac{1}{2}(b-a)^2 f'(\xi) \tag{5.13}$$

$$\text{Μέθοδος Simpson 1/3:} \quad E = -\frac{1}{2880}(b-a)^5 f^{(4)}(\xi) \tag{5.14}$$

$$\text{Μέθοδος Simpson 3/8:} \quad E = -\frac{1}{6480}(b-a)^5 f^{(4)}(\xi) \tag{5.15}$$

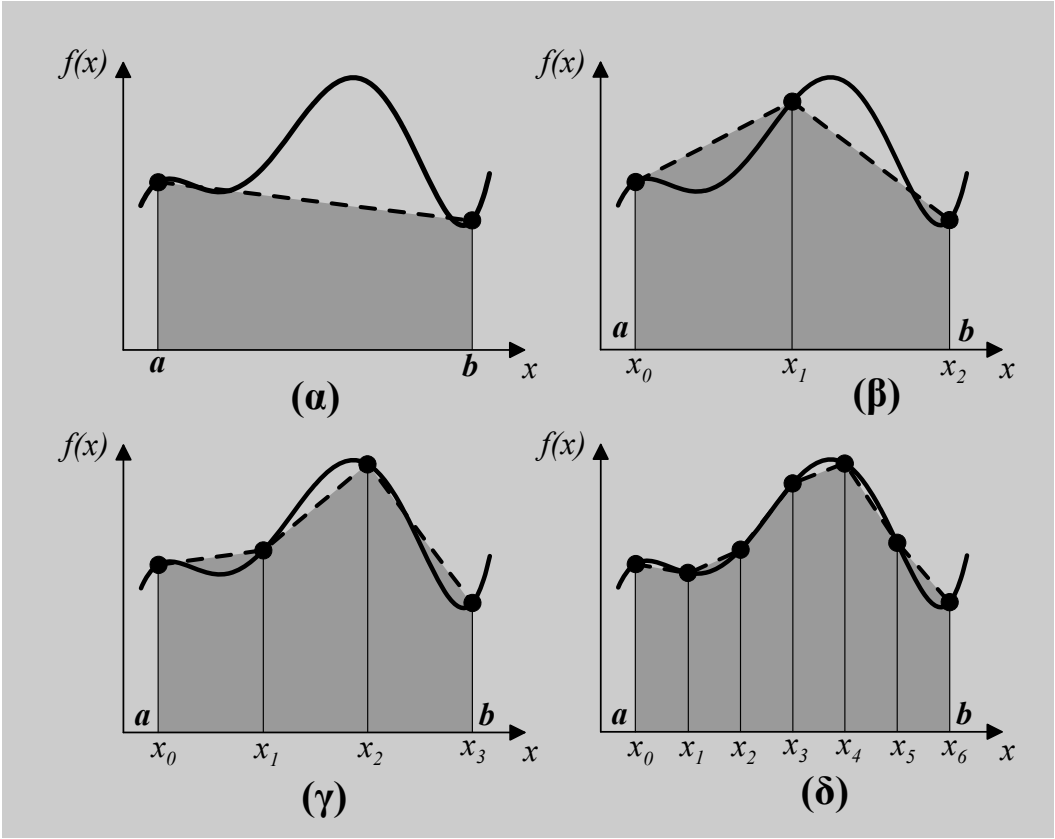
όπου ξ μια τιμή στο διάστημα $[a, b]$.

Είναι φανερό ότι το σφάλμα της μεθόδου των ορθογωνίων είναι σημαντικό, ακόμη και για μια γραμμική συνάρτηση, γι' αυτό συνήθως δεν χρησιμοποιείται. Η μέθοδος του Τραπεζίου πετυχαίνει αισθητή βελτίωση (ακρίβεια πρώτης τάξης). Η μέθοδος Simpson 1/3 εμφανίζει πολύ μικρό σφάλμα, μικρότερο και του αναμενόμενου, αφού είναι ανάλογο της τέταρτης παραγώγου της συνάρτησης, όπως δηλαδή και στη μέθοδο 3/8, που όμως χρησιμοποιεί πολυώνυμο μεγαλύτερου βαθμού. Επομένως, η μέθοδος Simpson 1/3 έχει την

ίδια ακρίβεια με την 3/8 (τρίτης τάξης ακρίβεια), άρα είναι σαφώς προτιμητέα, αφού μπορεί με ένα σημείο λιγότερο να υπολογίζει με ακρίβεια το εμβαδόν ενός πολυωνύμου μέχρι και τρίτου βαθμού.

5.1.1.3. Χρήση πολλαπλών υποδιαστημάτων

Στην πρακτική εφαρμογή των παραπάνω μεθόδων, το διάστημα ολοκλήρωσης [*a*, *b*] διαιρείται εξ αρχής σε έναν αριθμό *n* ίσων υποδιαστημάτων, είτε υπολογίζοντας την τιμή μιας αναλυτικής συνάρτησης σε *n*+1 ισαπέχοντα σημεία είτε χρησιμοποιώντας απευθείας τα *n*+1 διακριτά δεδομένα μιας πινακοποιημένης συνάρτησης (βλ. Σχ. 5.1β). Εφαρμόζοντας στη συνέχεια τη μέθοδο ολοκλήρωσης διαδοχικά στα υποδιαστήματα αυτά και αθροίζοντας τα επιμέρους ολοκληρώματα, υπολογίζεται το συνολικό ολοκλήρωμα με αυξημένη ακρίβεια. Το Σχήμα 5.3 δείχνει ένα τέτοιο παράδειγμα πολλαπλής εφαρμογής της μεθόδου του Τραπεζίου, όπου είναι φανερή η σημαντική βελτίωση που επιτυγχάνεται αυξάνοντας τον αριθμό των υποδιαστημάτων. Βέβαια, από την άλλη μεριά αυξάνει αντίστοιχα και ο υπολογιστικός χρόνος που απαιτείται.



Σχήμα 5.3. Παράδειγμα χρήσης πολλαπλών υποδιαστημάτων με τη μέθοδο του τραπεζίου: α) ένα, β) δύο γ) τρία και δ) έξι υποδιαστήματα.

Επειδή η εφαρμογή της μεθόδου Simpson 1/3 απαιτεί τρία σημεία (άρα δύο γειτονικά υποδιαστήματα), ο αριθμός *n* πρέπει να είναι άρτιος. Αντίστοιχα, για τη μέθοδο Simpson 3/8 πρέπει να είναι πολλαπλάσιο του 3. Έτσι, μερικές φορές μπορεί να χρειαστεί συνδυασμός μεθόδων. Για παράδειγμα, όταν ο *n* είναι περιττός και επιθυμείται η μέθοδος

Simpson 1/3, μπορεί στο τελευταίο υποδιάστημα να εφαρμοσθεί η μέθοδος του Τραπεζίου ή στα τρία τελευταία η μέθοδος Simpson 3/8. Η μέθοδος του Τραπεζίου εφαρμόζεται σε ένα υποδιάστημα κάθε φορά, γεγονός που δεν παρέχει μεν μεγάλη ακρίβεια, αλλά προσφέρει στη μέθοδο ένα σημαντικό πλεονέκτημα έναντι των μεθόδων Simpson: τα υποδιαστήματα *δεν* απαιτείται να είναι ίσα.

Οι αντίστοιχες εκφράσεις του ορισμένου ολοκληρώματος όταν το διάστημα [*a*, *b*] διαιρείται σε *n* ίσα υποδιαστήματα είναι οι ακόλουθες:

Μέθοδος του Τραπεζίου:
$$I \cong (b-a)\,\frac{f(x_0)+f(x_n)+2\sum_{i=1}^{n-1} f(x_i)}{2n} \tag{5.16}$$

Μέθοδος Simpson 1/3:
$$I \cong (b-a)\,\frac{f(x_0)+f(x_n)+4\sum_{i=1,3,5}^{n-1} f(x_i)+2\sum_{i=2,4,6}^{n-2} f(x_i)}{3n} \tag{5.17}$$

ενώ το αριθμητικό σφάλμα αποκοπής μπορεί να εκτιμηθεί από τις εκφράσεις:

Μέθοδος του Τραπεζίου:
$$E \cong -\frac{1}{12n^2}\,(b-a)^3\,\overline{f}'' = -\frac{h^3}{12}\,n\,\overline{f}'' \tag{5.18}$$

Μέθοδος Simpson 1/3:
$$E \cong -\frac{1}{180n^4}\,(b-a)^5\,\bar{f}^{(4)} = -\frac{h^5}{180}\,n\,\bar{f}^{(4)} \tag{5.19}$$

όπου εισάγεται η μέση τιμή της παραγώγου στο διάστημα [*a*, *b*], με τη σχέση

$$\bar{f}^{(m)} \cong \frac{1}{n}\sum_{i=1}^{n} f^{(m)}(\xi_i) \qquad \xi_i \in [x_{i-1},\, x_i] \tag{5.20}$$

5.1.1.4. Υπολογιστικοί αλγόριθμοι

Δύο απλοί αλγόριθμοι σε FORTRAN 90 για την εφαρμογή των μεθόδων του Τραπεζίου και του Simpson 1/3 δίνονται στη συνέχεια. Τα δεδομένα εισόδου είναι οι *n*+1 διακριτές τιμές μιας συνάρτησης *f(x)* μεταξύ των ορίων *a* και *b* του διαστήματος ολοκλήρωσης, καθώς και ο επιθυμητός αριθμός των υποδιαστημάτων *n* και το πλάτος κάθε

Κώδικας 5.1. Μέθοδος του Τραπεζίου

```
SUBROUTINE TRAPEZ (f, h, n, res)
total = f(0) + f(n)
DO  i = 1, n–1
      total = total + 2. * f(i)
END DO
res = h * total / 2.
RETURN
END
```

Κώδικας 5.2. Μέθοδος Simpson 1/3

```
SUBROUTINE SIM13 (f, h, n, res)
total = f(0) + f(n)
DO  i = 1, n–1, 2
      total = total + 4. * f(i)
END DO
DO  i = 2, n–2, 2
      total = total + 2. * f(i)
END DO
res = h * total / 3.
RETURN
END
```

ενός, $h=(b-a)/n$. Σημειώνεται ότι ο Κώδικας 5.1 εφαρμόζεται για ίσα υποδιαστήματα, όπως και ο Κώδικας 5.2, στον οποίο επιπλέον ο αριθμός *n* πρέπει να είναι άρτιος. Το αποτέλεσμα της αριθμητικής ολοκλήρωσης δίνεται στη μεταβλητή *res*.

Εφαρμογή 5.1.

Να υπολογισθεί η τιμή του ορισμένου ολοκληρώματος του πολυωνυμικής Εξ. (5.4) στην περιοχή [0, 1], με τις γενικές μεθόδους ολοκλήρωσης Τραπεζίου και Simpson, και να συγκριθεί η ακρίβειά τους.

Τα αποτελέσματα αριθμητικών υπολογισμών που προέκυψαν με χρήση 6 έως και $6\cdot10^5$ υποδιαστημάτων συνοψίζονται στον Πίνακα 5.1. Σε όλες τις μεθόδους το αριθμητικό σφάλμα μειώνεται αυξάνοντας τον αριθμό των υποδιαστημάτων, όχι όμως με τον ίδιο ρυθμό. Η υπεροχή της μεθόδου Simpson 1/3 είναι φανερή, καθώς επιτυγχάνει ακρίβεια έβδομου σημαντικού ψηφίου με 60 μόνο υποδιαστήματα. Είναι αξιοσημείωτο ότι για πολύ μεγάλο αριθμό υποδιαστημάτων το αριθμητικό σφάλμα αυξάνει και πάλι, επειδή το σφάλμα στρογγυλοποίησης μεγαλώνει όσο αυξάνει ο αριθμός των αριθμητικών πράξεων. Το σφάλμα αυτό περιορίζεται εάν χρησιμοποιηθεί διπλή ακρίβεια στους υπολογισμούς. Έτσι προκύπτουν τα αποτελέσματα του Πίνακα 5.2, στα οποία το αριθμητικό σφάλμα όλων των μεθόδων εξακολουθεί να μειώνεται και για μεγάλο αριθμό υποδιαστημάτων.

Πίνακας 5.1. Αποτελέσματα αριθμητικής ολοκλήρωσης της Εξ. (5.4), με μεταβλητές απλής ακρίβειας.

	Μέθοδος Τραπεζίου		**Μέθοδος Simpson 1/3**		**Μέθοδος Simpson 3/8**	
n	***I***	***E_r (%)***	***I***	***E_r (%)***	***I***	***E_r (%)***
6	2.2647890	$0.29\cdot10^{1}$	2.3297326	$0.15\cdot10^{0}$	2.3252316	$0.35\cdot10^{0}$
12	2.3160285	$0.74\cdot10^{0}$	2.3331084	$0.96\cdot10^{-2}$	2.3328269	$0.22\cdot10^{-1}$
18	2.3256281	$0.33\cdot10^{0}$	2.3332886	$0.19\cdot10^{-2}$	2.3332332	$0.43\cdot10^{-2}$
30	2.3305571	$0.12\cdot10^{0}$	2.3333273	$0.26\cdot10^{-3}$	2.3333205	$0.55\cdot10^{-3}$
60	2.3326389	$0.30\cdot10^{-1}$	2.3333334	$-0.93\cdot10^{-6}$	2.3333322	$0.49\cdot10^{-4}$
120	2.3331598	$0.74\cdot10^{-2}$	2.3333332	$0.63\cdot10^{-5}$	2.3333339	$-0.24\cdot10^{-4}$
180	2.3332571	$0.33\cdot10^{-2}$	2.3333335	$-0.54\cdot10^{-5}$	2.3333334	$-0.24\cdot10^{-5}$
300	2.3333049	$0.12\cdot10^{-2}$	2.3333332	$0.62\cdot10^{-5}$	2.3333335	$-0.68\cdot10^{-5}$
600	2.3333257	$0.33\cdot10^{-3}$	2.3333340	$-0.29\cdot10^{-4}$	2.3333333	$-0.24\cdot10^{-6}$
1800	2.3333301	$0.14\cdot10^{-3}$	2.3333351	$-0.75\cdot10^{-4}$	2.3333352	$-0.81\cdot10^{-4}$
6000	2.3333325	$0.36\cdot10^{-4}$	2.3333329	$0.17\cdot10^{-4}$	2.3333330	$0.15\cdot10^{-4}$
60000	2.3333042	$0.12\cdot10^{-2}$	2.3333822	$-0.21\cdot10^{-2}$	2.3333619	$-0.12\cdot10^{-2}$
600000	2.3348700	$-0.66\cdot10^{-1}$	2.3314562	$0.80\cdot10^{-1}$	2.3322642	$0.46\cdot10^{-1}$

Πίνακας 5.2. Ενδεικτικά αποτελέσματα αριθμητικής ολοκλήρωσης της Εξ. (5.4), με μεταβλητές διπλής ακρίβειας.

	Μέθοδος Τραπεζίου		**Μέθοδος Simpson 1/3**		**Μέθοδος Simpson 3/8**	
n	***I***	***E_r (%)***	***I***	***E_r (%)***	***I***	***E_r (%)***
6	2.2647891	$0.29\cdot10^{1}$	2.3297325	$0.15\cdot10^{0}$	2.3252315	$0.35\cdot10^{0}$
60	2.3326390	$0.30\cdot10^{-1}$	2.3333330	$0.15\cdot10^{-4}$	2.3333325	$0.35\cdot10^{-4}$
600	2.3333264	$0.30\cdot10^{-3}$	2.3333333	$0.15\cdot10^{-8}$	2.3333333	$0.35\cdot10^{-8}$
6000	2.3333333	$0.30\cdot10^{-5}$	2.3333333	$0.15\cdot10^{-12}$	2.3333333	$0.17\cdot10^{-12}$
60000	2.3333333	$0.30\cdot10^{-7}$	2.3333333	$-0.89\cdot10^{-12}$	2.3333333	$-0.44\cdot10^{-12}$

5.1.1.5. Χρήση πολυωνύμων μεγαλύτερου βαθμού

Για ακριβέστερη προσέγγιση των δεδομένων είναι δυνατή η χρήση πολυωνύμων μεγαλύτερου βαθμού, με αντίστοιχη αύξηση των σημείων. Για παράδειγμα, το πολυώνυμο $p_4(x)$ που μπορεί να ορισθεί από 5 σημεία (δύο οριακά και τρία εσωτερικά), έχει θεωρητικά ακρίβεια πέμπτης τάξης. Όμως η έκφραση του ολοκληρώματος είναι πιο πολύπλοκη και απαιτεί περισσότερες πράξεις. Αυτό έχει ως αποτέλεσμα να μεγαλώνει και το σφάλμα στρογγυλοποίησης. Επί πλέον, υπάρχουν συναρτήσεις στις οποίες η τιμή της παραγώγου αυξάνει συνεχώς όσο μεγαλώνει η τάξη της παραγώγισης. Επομένως, είναι πιθανό η χρήση πολυωνύμων μεγάλου βαθμού να δώσει χειρότερα αποτελέσματα από εκείνη με μικρό βαθμό. Γι' αυτό στην πράξη προτιμούνται οι απλούστερες μέθοδοι (Τραπεζίου ή Simpson), και η επίτευξη της επιθυμητής ακρίβειας επιτυγχάνεται με αύξηση του αριθμού των υποδιαστημάτων.

5.1.1.6. Άνισα υποδιαστήματα

Όπως ήδη αναφέρθηκε, η μέθοδος του Τραπεζίου είναι δυνατόν να χρησιμοποιηθεί ακόμη και όταν τα διακριτά δεδομένα δεν είναι ισοκατανεμημένα στον άξονα x, επειδή εφαρμόζεται σε ένα μόνο υποδιάστημα κάθε φορά. Στην περίπτωση βέβαια αυτή δεν ισχύει η Εξ. (5.16) και απαιτούνται αρκετά περισσότερες πράξεις. Για n υποδιαστήματα θα είναι

$$I \cong \sum_{i=0}^{n-1}\left[(x_{i+1}-x_i)\frac{f(x_i)+f(x_{i+1})}{2}\right] = \frac{1}{2}\sum_{i=1}^{n-2}(x_{i+1}-x_i)f(x_i) + \frac{x_1-x_0}{2}f(x_0) + \frac{x_n-x_{n-1}}{2}f(x_n) \tag{5.21}$$

Εάν στα δεδομένα περιέχονται και ζεύγη γειτονικών υποδιαστημάτων ίσου πλάτους, τότε μπορεί για μεγαλύτερη ακρίβεια να εφαρμόζεται σ' αυτά η μέθοδος Simpson 1/3. Για να γίνει αυτό θα πρέπει ο υπολογιστικός αλγόριθμος να εντοπίζει τέτοια ζεύγη και να εφαρμόζει την ανάλογη κάθε φορά μέθοδο.

Τέλος, ένας άλλος τρόπος χειρισμού άνισων υποδιαστημάτων, τα οποία προκύπτουν πολλές φορές στα αποτελέσματα πειραματικών μετρήσεων, είναι η προσαρμογή καμπυλών προσέγγισης ή παρεμβολής (π.χ. καμπύλη ελαχίστων τετραγώνων ή κυβικά splines), οι εξισώσεις των οποίων μπορούν να ολοκληρωθούν εύκολα με κάποια αριθμητική μέθοδο.

5.1.2. Η Μέθοδος Romberg

Οι αριθμητικές μέθοδοι που θα παρουσιαστούν στη συνέχεια είναι εφαρμόσιμες μόνο σε συναρτήσεις που εκφράζονται με αναλυτικές εξισώσεις, ώστε να είναι δυνατή η εύρεση και χρησιμοποίηση της τιμής της συνάρτησης σε οποιοδήποτε σημείο ενός διαστήματος $[a, b]$. Οι μέθοδοι αυτές είναι συνθετότερες από τις γενικές μεθόδους, αλλά ο υπολογιστικός χρόνος εκτέλεσής τους είναι μικρότερος, για ίδια ακρίβεια αποτελεσμάτων.

5.1.2.1. Περιγραφή

Η μέθοδος Romberg χρησιμοποιεί τον κανόνα του Τραπεζίου για τον υπολογισμό του ορισμένου ολοκληρώματος της συνάρτησης, αλλά ο υπολογισμός γίνεται δύο ή περισσότερες φορές, για διαφορετικό αριθμό υποδιαστημάτων. Στη συνέχεια εφαρμόζεται η γενική μέθοδος εκτίμησης σφάλματος του Richardson (Richardson's extrapolation), η οποία εδώ χρησιμοποιεί δύο εκτιμήσεις του ολοκληρώματος για τον υπολογισμό μιας τρίτης, που έχει ακρίβεια μεγαλύτερης τάξης, όπως αναλύεται στη συνέχεια.

Έστω ότι το ακριβές ολοκλήρωμα I υπολογίζεται δύο φορές, με χρήση n_1 και n_2 υποδιαστημάτων, παράγοντας τις τιμές I_{n1} και I_{n2}, με σφάλματα αποκοπής E_{n1} και E_{n2} αντιστοίχως. Ισχύει επομένως

$$I = I_{n1} + E_{n1} = I_{n2} + E_{n2} \tag{5.22}$$

Το εκτιμώμενο σφάλμα της μεθόδου του Τραπεζίου για n υποδιαστήματα είναι (Εξ. 5.18)

$$E_n \cong -\frac{1}{12n^2}(b-a)^3 \overline{f}'' \tag{5.23}$$

Έτσι, εάν θεωρηθεί η μέση τιμή της δεύτερης παραγώγου της συνάρτησης σχεδόν σταθερή, ανεξαρτήτως του αριθμού n, τότε προκύπτει η συσχέτιση

$$E_{n1} \cong E_{n2}\left(\frac{n_2}{n_1}\right)^2 \tag{5.24}$$

Αντικατάσταση της έκφρασης αυτής στην Εξ. (5.22) και επίλυση της δεύτερης ισότητάς της ως προς E_{n2}, δίνει

$$E_{n2} \cong \frac{I_{n2} - I_{n1}}{(n_2/n_1)^2 - 1} \tag{5.25}$$

Εισάγοντας τέλος την προηγούμενη έκφραση στην (5.22), λαμβάνουμε μια ακριβέστερη εκτίμηση του ολοκληρώματος:

$$I = I_{n2} + E_{n2} \cong I_{n2} + \frac{I_{n2} - I_{n1}}{(n_2/n_1)^2 - 1} \tag{5.26}$$

η οποία μπορεί να αποδειχτεί ότι έχει σφάλμα $O(h^4)$, όπου $h=(b-a)/n$, ενώ οι δύο προηγούμενες εκτιμήσεις με τον κανόνα του Τραπεζίου έχουν σφάλμα $O(h^2)$. Τέλος, στην περίπτωση που ο αριθμός n_2 είναι διπλάσιος του n_1, προκύπτει η απλούστερη έκφραση:

$$I \cong \frac{4}{3}I_{n2} - \frac{1}{3}I_{n1} \tag{5.27}$$

Εάν τώρα ληφθεί και μία τρίτη εκτίμηση του ολοκληρώματος χρησιμοποιώντας n_3 υποδιαστήματα, τότε μπορούν να γίνουν, σύμφωνα με τα παραπάνω, δύο διαφορετικές εκτιμήσεις του I, συνδυάζοντας είτε τις τιμές $I_{n1} - I_{n2}$ είτε τις $I_{n2} - I_{n3}$ αντιστοίχως, οι οποίες θα έχουν σφάλμα $O(h^4)$. Εάν στη συνέχεια αυτές οι δύο νέες εκτιμήσεις συνδυαστούν μεταξύ τους, τότε θα προκύψει μια ακόμη καλλίτερη εκτίμηση, με σφάλμα $O(h^6)$. Γενικεύοντας την παραπάνω διαδικασία, ο Romberg κατέληξε στην παρακάτω έκφραση:

$$I_{i,j} \cong \frac{4^{j-1} I_{i+1,j-1} - I_{i,j-1}}{4^{j-1} - 1} \tag{5.28}$$

η οποία ισχύει στην περίπτωση που ο αριθμός n διπλασιάζεται σε κάθε νέα αρχική εκτίμηση με τον κανόνα του Τραπεζίου. Ο δείκτης i αναφέρεται στις εκτιμήσεις που έχουν ίδια τάξη ακρίβειας, ενώ ο j στην τάξη ή στο επίπεδο ακρίβειας. Για παράδειγμα, η τιμή της Εξ. (5.27) προκύπτει από δύο αρχικές εκτιμήσεις στο πρώτο επίπεδο, άρα $i+1 = 2$, και $j-1 = 1$.

Είναι φανερό ότι οι νέες εκτιμήσεις που προκύπτουν στο επίπεδο ακρίβειας j από την Εξ. (5.28), θα είναι κατά μία λιγότερες από εκείνες στο προηγούμενο επίπεδο $j-1$. Επομένως, εάν θέλουμε η τελική εκτίμηση του ολοκληρώματος να βρίσκεται στο j επίπεδο ακρίβειας, πρέπει να γίνουν ισάριθμες αρχικές εκτιμήσεις με τον κανόνα του Τραπεζίου. Ο

παρακάτω υπολογιστικός αλγόριθμος κατασκευάστηκε σε αυτή τη λογική, θεωρώντας ως δεδομένο εισόδου τον αριθμό του τελικού επιπέδου ακρίβειας *lev*. Στο πρώτο μέρος του αλγορίθμου περιλαμβάνεται και ο υπολογισμός των n+1 τιμών της συνάρτησης *function*, που απαιτούνται στην εκάστοτε κλήση της μεθόδου του Τραπεζίου, ενώ στο δεύτερο γίνεται χρήση της Εξ. (5.28). Οι διάφορες εκτιμήσεις του ολοκληρώματος αποθηκεύονται στη διδιάστατη μεταβλητή *resr*, ενώ η τελική εκτίμηση του επιπέδου *lev* δίνεται στην *resf*.

Κώδικας 5.3. Μέθοδος του Romberg

```
SUBROUTINE ROMB (a, b, lev, resf)
DO i = 1, lev
     n = 2 ** (i –1)
     h = (b – a) / FLOAT (n)
     x = a
     f (0) = function (x)
     DO k = 1, n
          x = x + h
          f (k) = function (x)
     END DO
     CALL TRAPEZ (f, h, n, res)
     resr (i, 1) = res
END DO

DO j = 2, lev
     iavail = lev + 1 – j
     DO i = 1, iavail
     resr (i, j) = [4.** (j–1) * resr (i+1, j–1) – resr (i, j–1)]
                   / (4.** (j–1) – 1.)
     END DO
END DO
resf = resr (1, lev)
RETURN
END
```

Εφαρμογή 5.2.

Να υπολογισθεί η τιμή του ορισμένου ολοκληρώματος του πολυωνυμικής Εξ. (5.4) στην περιοχή [0, 1], με τη μέθοδο του Romberg.

Χρησιμοποιείται ο Κώδικας 5.3 για τέσσερα επίπεδα ακρίβειας. Ο Πίνακας 5.3 δείχνει τη σειρά με την οποία λαμβάνονται οι διαδοχικές εκτιμήσεις του ολοκληρώματος, καθώς και τα αντίστοιχα σφάλματα % (ακριβής λύση $I = 7/3$). Η πρώτη στήλη περιέχει τις εκτιμήσεις από την εφαρμογή της μεθόδου του Τραπεζίου, για n = 1, 2, 4 και 8 υποδιαστήματα, οι οποίες έχουν σφάλμα $O(h^2)$. Συνδυάζοντας ανά δύο τις τιμές αυτές με την Εξ. (5.28), προκύπτει η δεύτερη στήλη (δεύτερο επίπεδο), με σφάλμα $O(h^2)$. Ομοίως, από τις τρεις τιμές της δεύτερης στήλης προκύπτουν οι δύο τιμές του τρίτου επιπέδου και από αυτές η τελική τιμή στο τέταρτο επίπεδο, που έχει (θεωρητικά) σφάλμα $O(h^8)$.

Όπως είναι αναμενόμενο, το σφάλμα των εκτιμήσεων του πρώτου επιπέδου (μέθοδος Τραπεζίου) μειώνεται όσο αυξάνει ο αριθμός των υποδιαστημάτων. Γι' αυτό και οι εκτιμήσεις των επόμενων επιπέδων γίνονται ακριβέστερες από γραμμή σε γραμμή του Πίνακα 5.3 (δηλ. αυξανομένου του i). Αναμενόμενη είναι επίσης η μεγάλη μείωση του σφάλματος από επίπεδο σε επίπεδο j, με αποτέλεσμα ήδη στο τρίτο επίπεδο να επιτυγχάνεται ακριβής λύση. Αυτό συμβαίνει επειδή η Εξ. (5.4) είναι πολυώνυμο $5^{ου}$ βαθμού, μικρότερου δηλαδή από τη θεωρητική τάξη ακρίβειας του τρίτου επιπέδου.

Επομένως, για επίτευξη ακριβούς λύσης απαιτείται η εύρεση της τιμής της συνάρτησης σε μόλις 15 υποδιαστήματα συνολικά (1 + 2 + 4 + 8), ενώ οι γενικές μέθοδοι ολοκλήρωσης απαιτούν πολύ περισσότερα (βλ. Εφαρμογή 5.1). Επιπλέον, επειδή οι αριθμητικές πράξεις που γίνονται είναι πολύ λίγες, δεν προκαλείται σφάλμα στρογγυλοποίησης, γι' αυτό η μέθοδος μπορεί να δώσει ακρίβεια πολλών σημαντικών ψηφίων χωρίς να χρειάζονται υπολογισμοί με μεταβλητές διπλής ακρίβειας.

Πίνακας 5.3. Αποτελέσματα αριθμητικής ολοκλήρωσης της Εξ. (5.4) με τη μέθοδο Romberg.

$j = 1,\ O(h^2)$	$j = 2,\ \ O(h^4)$	$j = 3,\ \ O(h^6)$	$j = 4,\ \ O(h^8)$
1.0000000			
$-0.57\cdot10^{2}$	**2.0416667**		
1.7812500	$-0.13\cdot10^{2}$	**2.3333333**	
$-0.24\cdot10^{2}$	**2.3151042**	$-0.19\cdot10^{-13}$	**2.3333333**
2.1816406	$-0.78\cdot10^{0}$	**2.3333333**	$-0.19\cdot10^{-13}$
$-0.65\cdot10^{1}$	**2.3321940**	$-0.19\cdot10^{-13}$	
2.2945557	$-0.49\cdot10^{-1}$		
$-0.17\cdot10^{1}$			

5.1.2.2. Κριτήριο τερματισμού

Επειδή στην πρακτική εφαρμογή μιας αριθμητικής μεθόδου ολοκλήρωσης δεν είναι γνωστή εκ των προτέρων η ακριβής λύση, χρειάζεται να ορισθεί κάποιο κριτήριο, ώστε να τερματίζεται η διαδικασία όταν η ακρίβεια του αποτελέσματος θεωρείται επαρκής. Ένα συνηθισμένο κριτήριο είναι η σχετική διαφορά μεταξύ δύο διαδοχικών προσεγγίσεων να γίνει μικρότερη μιας προκαθορισμένης τιμής ε_r. Στη μέθοδο Romberg η σύγκριση γίνεται μεταξύ των πρώτων γραμμών δύο διαδοχικών επιπέδων:

$$\left|\varepsilon_a\right| = \left|\frac{I_{1,j} - I_{1,j-1}}{I_{1,j}}\right| \le \varepsilon_r \tag{5.29}$$

Στα αποτελέσματα της Εφαρμογής 5.2, η σχετική διαφορά είναι 0.51 στο δεύτερο επίπεδο, 0.125 στο τρίτο και 0.0 στο τέταρτο. Έτσι, εάν το κριτήριο σύγκλισης είναι π.χ. $\varepsilon_r = 10^{-5}$, η

επίλυση θα σταματήσει στο τέταρτο επίπεδο. Για μια τέτοια λειτουργία, ο αλγόριθμος του Κώδικα 5.3 μπορεί τροποποιηθεί ώστε να γίνει επαναληπτικός, δηλαδή ο αριθμός των επιπέδων να μην προκαθορίζεται, αλλά να προστίθεται κάθε φορά ένα νέο επίπεδο ολοκλήρωσης, μέχρι να ικανοποιηθεί το κριτήριο σύγκλισης.

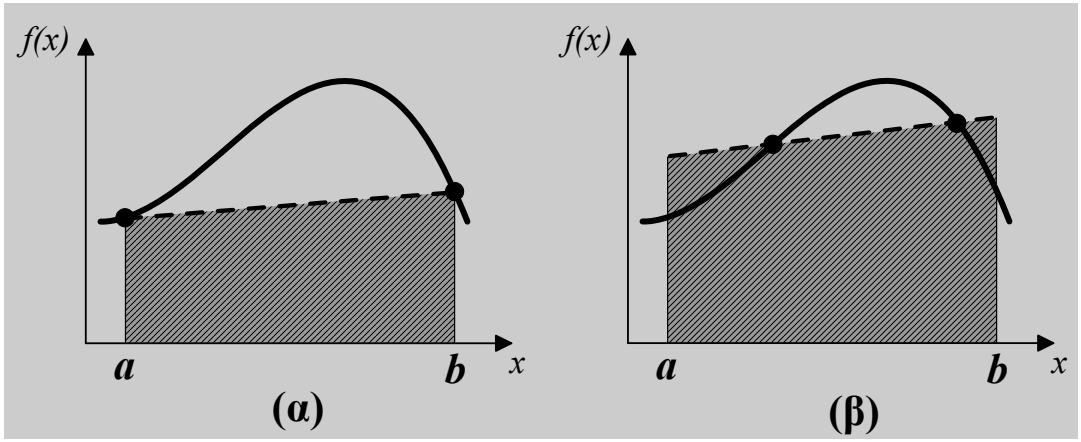
Αντίστοιχο κριτήριο μπορεί να εφαρμοσθεί και στις γενικές μεθόδους ολοκλήρωσης, με την παρατήρηση ότι οι δύο προσεγγίσεις που συγκρίνονται θα πρέπει να έχουν ληφθεί σε αρκετά διαφορετικό αριθμό υποδιαστημάτων, συνήθως 2:1. Έτσι, για την τιμή $\varepsilon_r = 10^{-5}$, σύμφωνα με τα αποτελέσματα της Εφαρμογής 5.1 (Πίνακας 5.1), η μέθοδος του Τραπεζίου θα σταματήσει στα 600 υποδιαστήματα, ενώ του Simpson 1/3 στα 60, δίνοντας προσεγγίσεις με ακρίβεια τουλάχιστον 5 σημαντικών ψηφίων.

5.1.3. Ολοκλήρωση κατά Gauss

Σε όλες τις προηγούμενες μεθόδους χρησιμοποιούνται ισαπέχοντα σημεία στον άξονα των x για τον ορισμό των διακριτών δεδομένων, και το ορισμένο ολοκλήρωμα μιας συνάρτησης $f(x)$ υπολογίζεται τελικά με την ίδια γενική σχέση (βλ. Εξ. 5.8, 5.9, 5.10, 5.28):

$$I = \int_a^b f(x)dx = \sum_{i=0}^{n} w_i\, f(x_i) + E_t \cong \sum_{i=0}^{n} w_i\, f(x_i) \tag{5.30}$$

όπου οι n+1 τιμές w_i είναι αριθμητικοί συντελεστές (συντελεστές βαρύτητας) των τιμών της συνάρτησης στις αντίστοιχες θέσεις x_i και E_t είναι το αριθμητικό σφάλμα αποκοπής.



Σχήμα 5.4. Ολοκλήρωμα συνάρτησης με χρήση δύο σημείων: α) στα όρια του διαστήματος ολοκλήρωσης και β) σε επιλεγμένες ενδιάμεσες θέσεις.

Όταν οι θέσεις x_i ισαπέχουν, τότε ο βαθμός του πολυωνύμου προσέγγισης κάθε μεθόδου είναι ίσος με n (π.χ. στη μέθοδο του Τραπεζίου, με δύο σημεία στις θέσεις a και b, το πολυώνυμο είναι πρώτου βαθμού), με αντίστοιχη τάξη ακρίβειας. Δηλαδή, η μέθοδος του Τραπεζίου υπολογίζει με ακρίβεια το ολοκλήρωμα πολυωνύμων μέχρι και πρώτου βαθμού. Εάν όμως δεν τεθεί περιορισμός στην επιλογή των θέσεων x_i, τότε προκύπτουν περισσότεροι βαθμοί ελευθερίας στην Εξ. (5.30): οι n+1 θέσεις x_i *και* οι ισάριθμοι συντελεστές w_i. Επομένως, οι τιμές όλων αυτών των μεγεθών μπορεί να εκλεγούν έτσι ώστε να προκύψει μικρότερο σφάλμα E_t. Όπως φαίνεται και στο παράδειγμα του Σχήματος

5.4, λαμβάνοντας τα δύο σημεία της μεθόδου του Τραπεζίου σε κατάλληλες ενδιάμεσες θέσεις αντί στα όρια του διαστήματος [a, b], η τιμή του ορισμένου ολοκληρώματος μιας συνάρτησης μπορεί να υπολογισθεί με πολύ μεγαλύτερη ακρίβεια, καθώς αλληλο-εξουδετερώνονται τα θετικά με τα αρνητικά σφάλματα.

Στη γενική περίπτωση προκύπτει ότι για n+1 διακριτά σημεία, οι τιμές x_i μπορούν να εκλεγούν έτσι ώστε, σε συνδυασμό με τους κατάλληλους συντελεστές βαρύτητας w_i, ο αθροιστής της Εξ. (5.30) να δίνει την ακριβή τιμή του ολοκληρώματος κάθε πολυωνυμικής εξίσωσης μέχρι και 2n+1 βαθμού. Για n = 1 για παράδειγμα, μπορεί να επιτευχθεί ακρίβεια για πολυώνυμα μέχρι και τρίτου βαθμού, ενώ η απλή μέθοδος του Τραπεζίου είναι ακριβής μόνο για πρώτου βαθμού. Όπως αποδεικνύεται, οι ζητούμενες n+1 θέσεις x_i αποτελούν τις ρίζες ενός ορθογωνίου πολυωνύμου, βαθμού n+1.

5.1.3.1. Ορθογώνια πολυώνυμα

Μία σειρά πολυωνύμων $p_i(x)$ βαθμού i = 0, 1, 2, …n, ονομάζονται ορθογώνια στο διάστημα [a, b], με συνάρτηση βαρύτητας $w(x)$, όταν ισχύει

$$\int_a^b w(x)\, p_k(x)\, p_m(x)\, dx = \left| \begin{array}{ll} 0, & k \neq m \\ c(m) \neq 0, & k = m \end{array} \right| \tag{5.31}$$

Μία τέτοια οικογένεια αποτελούν τα πολυώνυμα Legendre, που είναι ορθογώνια στο διάστημα [–1, 1], με συνάρτηση βαρύτητας $w(x)$ = 1. Τα πρώτα πολυώνυμα της σειράς αυτής είναι:

$$\begin{aligned} p_0(x) &= 1 \\ p_1(x) &= x \\ p_2(x) &= 0.5 \cdot (3x^2 - 1) \\ p_3(x) &= 0.5 \cdot (5x^3 - 3x) \\ p_4(x) &= 0.125 \cdot (35x^4 - 30x^2 + 3) \end{aligned} \tag{5.32}$$

Τα ορθογώνια πολυώνυμα χρησιμοποιούνται σε αρκετά θέματα μαθηματικής και αριθμητικής ανάλυσης, επειδή έχουν μερικές ενδιαφέρουσες ιδιότητες. Έτσι, κάθε πολυώνυμο $P(x)$ βαθμού n μπορεί να γραφεί ως ένας γραμμικός συνδυασμός μιας οικογένειας ορθογωνίων πολυωνύμων $p_i(x)$. Επίσης, κάθε ορθογώνιο πολυώνυμο p_i έχει i απλές, πραγματικές και διαφορετικές μεταξύ τους ρίζες, που βρίσκονται όλες μέσα στο διάστημα ορισμού της οικογένειας που ανήκει. Τέλος, οι ρίζες αυτές αποτελούν τη βέλτιστη δυνατή κατανομή σημείων x_i, η οποία ελαχιστοποιεί το σφάλμα E_t της Εξ. (5.30), επιτυγχάνοντας βαθμό ακρίβειας 2i+1.

5.1.3.2. Η μέθοδος ολοκλήρωσης Gauss-Legendre

Η ολοκλήρωση κατά Gauss χρησιμοποιεί τα πολυώνυμα παρεμβολής Lagrange ($L_i(x)$, βλ. Κεφ. 4.2.3) για να προσεγγίσει τη συνάρτηση $f(x)$. Εάν η παρεμβολή γίνει σε n+1 σημεία στο διάστημα [a, b], τότε το ορισμένο ολοκλήρωμα της συνάρτησης θα είναι:

$$\begin{aligned} I &= \int_a^b f(x)dx = \int_a^b \sum_{i=0}^{n} L_i(x) f(x_i)\, dx + E_t = \sum_{i=0}^{n} w_i\, f(x_i) + E_t \\ w_i &= \int_a^b L_i(x)\, dx \end{aligned} \tag{5.33}$$

όπου οι τιμές $f(x_i)$ γράφονται έξω από το ολοκλήρωμα επειδή είναι σταθερές (για δεδομένα x_i). Επομένως προκύπτει και πάλι η γενική σχέση (5.30).

Εάν το διάστημα ολοκλήρωσης είναι το [–1, 1], τότε το σφάλμα της παραπάνω έκφρασης ελαχιστοποιείται, όπως αναφέρθηκε προηγουμένως, όταν τα x_i αποτελούν τις ρίζες ενός ορθογώνιου πολυωνύμου Legendre $p_i(x)$. Έτσι, εάν η παρεμβολή γίνει σε δύο σημεία της συνάρτησης (n=1), οι τιμές x_i θα είναι οι ρίζες του $p_2(x)$, δηλαδή $\pm(1/3)^{0.5}$ (Εξ. 5.32), ενώ για τρία σημεία θα είναι οι ρίζες του $p_3(x)$, δηλαδή 0 και $\pm(3/5)^{0.5}$. Για κάθε τιμή x_i υπολογίζεται η αντίστοιχη τιμή της συνάρτησης $f(x_i)$, ενώ οι συντελεστές βαρύτητας w_i προκύπτουν από τη σχέση

$$w_i = \int_{-1}^{1} L_i(x)dx = \int_{-1}^{1} \prod_{\substack{j=0 \\ j\neq i}}^{n} \left(\frac{x - x_j}{x_i - x_j} \right) dx \tag{5.34}$$

Για παράδειγμα, ο συντελεστής w_0 για n=1 θα είναι:

$$w_0 = \int_{-1}^{1} \frac{x - x_1}{x_0 - x_1} dx = \left. \frac{0.5\,x^2 - x_1\,x}{x_0 - x_1} \right|_{-1}^{1} = \frac{-2x_1}{x_0 - x_1} = \frac{-2/3^{0.5}}{-1/3^{0.5} - 1/3^{0.5}} = 1.0$$

Στον Πίνακα 5.4 περιλαμβάνονται οι τιμές των x_i και w_i της μεθόδου για χρήση 2 έως 5 σημείων, οι οποίες μπορούν να βρεθούν σχετικά εύκολα και για περισσότερα σημεία.

Πίνακας 5.4. Ορίσματα της συνάρτησης $f(x)$ και συντελεστές βαρύτητας της μεθόδου Gauss-Legendre.

Σημεία	Ρίζες	Συντελεστές βαρύτητας
2	$x_0 = -0.5773502691$	$w_0 = 1.0$
	$x_1 = 0.5773502691$	$w_1 = 1.0$
3	$x_0 = -0.7745966692$	$w_0 = 0.5555555556$
	$x_1 = 0.0$	$w_1 = 0.8888888889$
	$x_2 = 0.7745966692$	$w_2 = 0.5555555556$
4	$x_0 = -0.8611363115$	$w_0 = 0.3478548451$
	$x_1 = -0.3399810435$	$w_1 = 0.6521451548$
	$x_2 = 0.3399810435$	$w_2 = 0.6521451548$
	$x_3 = 0.8611363115$	$w_3 = 0.3478548451$
5	$x_0 = -0.9061798459$	$w_0 = 0.2369268850$
	$x_1 = -0.5384693101$	$w_1 = 0.4786286704$
	$x_2 = 0.0$	$w_2 = 0.5688888889$
	$x_3 = 0.5384693101$	$w_3 = 0.4786286704$
	$x_4 = 0.9061798459$	$w_4 = 0.2369268850$

Επειδή τα όρια ενός ορισμένου ολοκληρώματος μιας συνάρτησης δεν θα είναι γενικά μεταξύ –1 και 1, όπως απαιτεί η μέθοδος Gauss-Legendre, η βασική έκφρασή της

μπορεί να μετατραπεί, ώστε να είναι εφαρμόσιμη στο επιθυμητό κάθε φορά διάστημα [*a*, *b*], χρησιμοποιώντας τον μετασχηματισμό

$$x_i = \frac{x_i'(b-a)+b+a}{2} \tag{5.35}$$

ώστε οι μετασχηματισμένες τιμές των ριζών να βρίσκονται στο διάστημα [*a*, *b*] (π.χ. $x_i = a \leftrightarrow x_i' = -1, \quad x_i = b \leftrightarrow x_i' = +1$. Έτσι, η τελική, γενική έκφραση γίνεται:

$$I = \int_a^b f(x)\,dx = \frac{b-a}{2} \cdot \int_{-1}^{1} f(x)\,dx' \cong \frac{b-a}{2} \cdot \sum_{i=0}^{n} w_i\, f\left(\frac{x_i'(b-a)+b+a}{2}\right) \tag{5.36}$$

Όπως αναφέρθηκε και προηγουμένως, η μέθοδος Gauss-Legendre με *n*+1 σημεία μπορεί να υπολογίσει ακριβώς το ολοκλήρωμα μιας πολυωνυμικής συνάρτησης βαθμού 2*n*+1 ή μικρότερου. Στη γενική περίπτωση μιας τυχαίας συνάρτησης, το εκτιμώμενο σφάλμα της μεθόδου στο διάστημα [–1, 1] δίνεται από την ακόλουθη έκφραση:

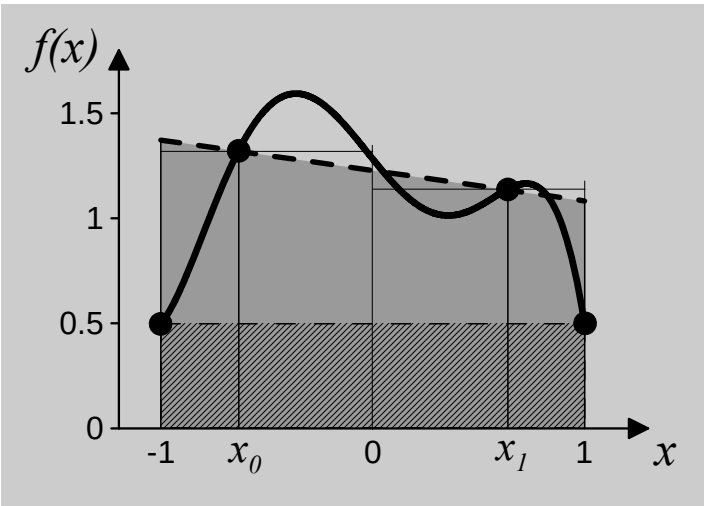
$$E \cong \frac{2^{2n+3}[(n+1)!]^4}{(2n+3)[(2n+2)!]^3}\, f^{(2n+2)}(\xi) \tag{5.37}$$

Εφόσον οι τιμές των παραγώγων μεγάλης τάξης της συνάρτησης *f* μειώνονται ή δεν αυξάνουν σημαντικά, η ακρίβεια της μεθόδου είναι πολύ μεγαλύτερη από όλες τις προηγούμενες. Συγκρίνοντας για παράδειγμα με το αντίστοιχο σφάλμα της μεθόδου Simpson 1/3 (Εξ. 5.14), για τρία σημεία (*n* = 2) και διάστημα [–1, 1], η (5.14) δίνει $(-1/90)\, f^{(4)}(\xi)$, ενώ η (5.37) δίνει $(1/15750) f^{(6)}(\xi)$.

Εφαρμογή 5.3.

Να υπολογισθεί η τιμή του ορισμένου ολοκληρώματος του πολυωνυμικής Εξ. (5.4) στην περιοχή [0, 1], με τη μέθοδο Gauss-Legendre δύο έως και πέντε σημείων.

Χρησιμοποιείται η Εξ. (5.36) και οι τιμές του Πίνακα 5.4, ενώ σύμφωνα με την Εξ. (5.35) είναι $x_i = 0.5\,x_i' + 0.5$, και ισχύει:



Σχήμα 5.5. Ολοκλήρωμα της Εξ. (5.4), τροποποιημένης για το διάστημα [-1, 1], με χρήση της μεθόδου Gauss-Legendre δύο σημείων.

$$\int_0^1 f(x)\,dx = 0.5 \cdot \int_{-1}^{1} [-160(0.5x'+0.5)^5 + 365(0.5x'+0.5)^4 - 270(0.5x'+0.5)^3 + 60(0.5x'+0.5)^2 + 5(0.5x'+0.5) + 1]\,dx'$$

Η μέθοδος δύο σημείων (n = 1) δίνει:

$I = 0.5\,[1.0\,f(-0.5773502691/2 + 0.5) + 1.0\,f(0.5773502691/2 + 0.5)] = 2.84852796$

με σφάλμα 8.333 %.

Στο Σχήμα 5.5 σχεδιάζεται η συνάρτηση $f(x')$, και οι τιμές της στα δύο σημεία της μεθόδου Gauss-Legendre, καθώς και στα όρια του διαστήματος ολοκλήρωσης. Είναι φανερό ότι η γραμμή που συνδέει τα ενδιάμεσα σημεία προσεγγίζει πολύ καλλίτερα το ολοκλήρωμα της συνάρτησης από ότι η γραμμή που συνδέει τα δύο άκρα της.

Αντίστοιχα για τρία, τέσσερα και πέντε σημεία, παίρνουμε:

3 σημεία:	$I = 2.33333336$	$\varepsilon_t = 0.12\cdot10^{-5}$ %
4 σημεία:	$I = 2.33333331$	$\varepsilon_t = -0.62\cdot10^{-6}$ %
5 σημεία:	$I = 2.33333333$	$\varepsilon_t = 0.36\cdot10^{-7}$ %

Επομένως, η μέθοδος τριών σημείων δίνει πρακτικά ακριβές αποτέλεσμα (οκτώ σημαντικά ψηφία), με μόνο τρεις υπολογισμούς της συνάρτησης, ενώ σύμφωνα με τα αποτελέσματα του Πίνακα 5.1, η καλλίτερη των γενικών μεθόδων, η Simpson 1/3, απαιτεί περίπου 60 υπολογισμούς, ενώ η μέθοδος Romberg 15. Βέβαια αυτό ήταν αναμενόμενο, αφού για 3 σημεία είναι n=2 και η μέθοδος εξ ορισμού δίνει ακριβείς λύσεις για πολυώνυμα βαθμού μέχρι και $2n$+1=5, όσος δηλαδή και ο βαθμός της πολυωνυμικής εξίσωσης (5.4).

Εφαρμογή 5.4.

Να βρεθεί με ακρίβεια πέντε σημαντικών ψηφίων το ολοκλήρωμα $\int_{-3}^{3} \frac{2}{1+2x^2}\,dx$

Το ολοκλήρωμα έχει αναλυτική λύση με τιμή I = 3.78816608. Χρησιμοποιείται πρώτα η μέθοδος Gauss-Legendre, η οποία όμως δεν μπορεί να δώσει ικανοποιητική ακρίβεια, ακόμη και με 15 σημεία, όπως φαίνεται στον Πίνακα 5.5. Αυτό συμβαίνει επειδή οι τιμές των παραγώγων μεγάλης τάξης της συνάρτησης αυτής αυξάνουν συνεχώς. Σε τέτοιες περιπτώσεις είναι προτιμότερη η χρήση των γενικών μεθόδων ολοκλήρωσης. Έτσι, για την επίτευξη της επιθυμητής ακρίβειας, η μέθοδος του Τραπεζίου απαιτεί περίπου 60 υποδιαστήματα, ενώ του Simpson 1/3 περίπου 30 (Πίνακας 5.5). Πάντως τέτοιες συναρτήσεις δεν συναντώνται συχνά σε πρακτικά προβλήματα, έτσι η μέθοδος Gauss-Legendre λειτουργεί συνήθως πολύ αποτελεσματικά.

Πίνακας 5.5. Σύγκριση αποτελεσμάτων αριθμητικών μεθόδων.

Μέθοδος Gauss-Legendre			Γενικές Μέθοδοι Ολοκλήρωσης		
Σημεία	I	ε_t (%)	n	Τραπεζίου	Simpson 1/3
2	1.7142858	$-0.54\cdot10^{2}$	6	3.8830409	4.2183236
3	5.8983051	$0.55\cdot10^{2}$	12	3.7866383	3.7545041
4	2.8312862	$-0.25\cdot10^{2}$	18	3.7869520	3.7919489
5	4.5170306	$0.19\cdot10^{2}$	30	3.7877232	3.7882092
6	3.3845136	$-0.11\cdot10^{2}$	60	3.7880553	
10	3.7231652	$-0.17\cdot10^{1}$			
15	3.7945163	$0.17\cdot10^{0}$			

5.1.3.3. Άλλες μέθοδοι ολοκλήρωσης κατά Gauss

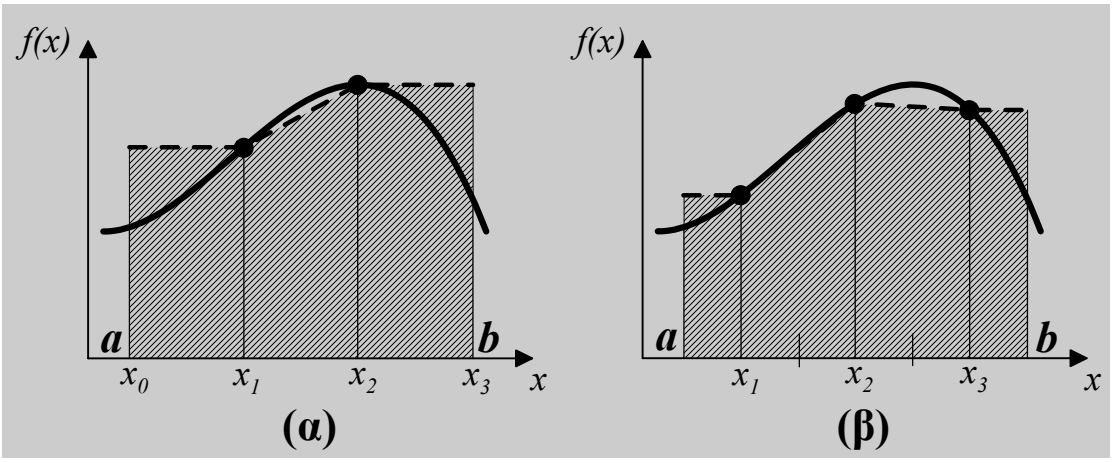
Εκτός από τα πολυώνυμα Legendre, υπάρχουν και άλλες οικογένειες ορθογώνιων πολυωνύμων, όπως τα πολυώνυμα Laguerre, που είναι ορθογώνια στο διάστημα [0, ∞], με συνάρτηση βαρύτητας $w(x) = e^{-x}$, τα πολυώνυμα Chebyshev, στο διάστημα [–1, 1], με $w(x) = 1/\sqrt{1-x^2}$, τα πολυώνυμα Hermite, στο [–∞, ∞], με $w(x) = e^{-x^2}$ κ.ά. Έτσι έχουν προκύψει και οι αντίστοιχες μέθοδοι ολοκλήρωσης κατά Gauss ή τετραγωνισμού κατά Gauss (Gaussian Quadrature), η ανάπτυξη και εφαρμογή των οποίων είναι παρόμοια με την Gauss-Legendre. Κάθε τέτοια μέθοδος είναι αποτελεσματικότερη σε συγκεκριμένο τύπο συναρτήσεων, όπως για παράδειγμα η μέθοδος Gauss-Laguerre σε εκθετικές συναρτήσεις με μη πεπερασμένο άνω όριο ολοκληρώματος ή η μέθοδος Gauss-Chebyshev σε συναρτήσεις που περιέχουν τον όρο $1/\sqrt{1-x^2}$. Τέλος, όλες οι μέθοδοι μπορούν να εφαρμοστούν και τμηματικά, χωρίζοντας το συνολικό διάστημα ολοκλήρωσης σε υποδιαστήματα.

5.1.4. Ειδικά Θέματα

5.1.4.1. Μέθοδοι ολοκλήρωσης ανοικτού τύπου

Μερικές φορές τα όρια του διαστήματος ολοκλήρωσης είναι μεγαλύτερα από την περιοχή των διαθέσιμων δεδομένων. Αυτό μπορεί να συμβεί για παράδειγμα σε μια σειρά πειραματικών μετρήσεων ή σε μια συνάρτηση η οποία δεν ορίζεται σε κάποιο όριο του διαστήματος (ιδιόμορφο σημείο). Στην περίπτωση αυτή οι γενικές μέθοδοι, που είναι κλειστού τύπου, δηλαδή περιλαμβάνουν και τις τιμές της συνάρτησης στα όρια ολοκλήρωσης, δεν είναι εφαρμόσιμες. Αντί αυτών μπορούν να ορισθούν αντίστοιχες μέθοδοι ανοικτού τύπου, στις οποίες δεν χρησιμοποιούνται τα σημεία x_0 και x_n (στις θέσεις a και b), όπως φαίνεται στο παράδειγμα του Σχήματος 5.6α. Η αντίστοιχη έκφραση της μεθόδου του Τραπεζίου (βλ. Εξ. 5.16) θα είναι τότε

$$I \cong (b-a)\,\frac{f(x_1) + f(x_{n-1}) + 2\sum_{i=1}^{n-1} f(x_i)}{2n} \tag{5.38}$$



Σχήμα 5.6. Παράδειγμα ολοκλήρωσης χωρίς τη χρήση των οριακών τιμών *f(a)* και *f(b)*: α) μέθοδος Τραπεζίου, ανοικτού τύπου και β) κανόνας του Μέσου Σημείου.

Όταν πρόκειται για ολοκλήρωση συναρτήσεων είναι προτιμότερη η χρήση του ακριβέστερου κανόνα του Μέσου Σημείου, όπου το πρώτο και το τελευταίο υποδιάστημα έχουν πλάτος $h/2$. (Σχ. 5.6β). Τότε το ολοκλήρωμα εκφράζεται απλά ως

$$I \cong \frac{(b-a)}{n} \sum_{i=1}^{n} f(x_i) \tag{5.39}$$

Επειδή η ακρίβεια των μεθόδων ανοικτού τύπου είναι γενικά μικρότερη των αντίστοιχων κλειστού τύπου, χρησιμοποιούνται συνήθως μόνο στις παραπάνω περιπτώσεις. Επιπλέον, η χρήση τους μπορεί να περιορισθεί μόνο κοντά στην περιοχή του ορίου. Εξαίρεση αποτελούν οι μέθοδοι τετραγωνισμού κατά Gauss, οι οποίες έχουν μεγάλη αποτελεσματικότητα, ενώ μπορούν να χαρακτηριστούν ως ανοικτού τύπου, αφού δεν χρησιμοποιούν τις ακραίες τιμές του διαστήματος ολοκλήρωσης.

5.1.4.2. Ολοκληρώματα με μη πεπερασμένα όρια

Όταν το διάστημα ολοκλήρωσης εκτείνεται μέχρι το -∞ ή/και το +∞, τότε το γενικευμένο αυτό ολοκλήρωμα, εφόσον υπάρχει, είναι δυνατό να υπολογισθεί αριθμητικά με διάφορες τεχνικές. Μία χρήσιμη ιδιότητα είναι η ακόλουθη:

$$\int_a^b f(x)\,dx = \int_{1/b}^{1/a} \frac{1}{z^2} f\left(\frac{1}{z}\right) dz \tag{5.40}$$

η οποία είναι εφαρμόσιμη σε κάθε συνάρτηση που τείνει στο μηδέν όταν το x τείνει στο άπειρο, με ρυθμό τουλάχιστον $1/x^2$ και επιπλέον είναι $ab > 0$. Ισχύει επίσης και η ιδιότητα

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx \qquad c \in [a,b] \tag{5.41}$$

η οποία μπορεί να συνδυασθεί με την προηγούμενη, γράφοντας

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{-a} f(x)\,dx + \int_{-a}^{a} f(x)\,dx + \int_a^{\infty} f(x)\,dx \qquad (a > 0) \tag{5.42}$$

οπότε η Εξ. (5.40) είναι πλέον εφαρμόσιμη στον πρώτο και τρίτο όρο του δεξιού μέλους, αφού $(-a)(-\infty)>0$, $(a)(\infty)>0$, ενώ ο δεύτερος όρος έχει πεπερασμένα όρια.

Εφαρμογή 5.5.

Να υπολογισθεί η συνάρτηση Γάμμα $\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}\,dt$ για $x=5.5$, με κριτήριο σύγκλισης $e_r=10^{-5}$.

Το ολοκλήρωμα έχει αναλυτική λύση μόνο για ακέραια x, ακόμη και τότε όμως η έκφραση είναι αρκετά πολύπλοκη. Για την αριθμητική του επίλυση ακολουθείται η προηγούμενη μεθοδολογία, θέτοντας μια ενδιάμεση τιμή π.χ. $t=10$:

$$\Gamma(5{,}5)=\int_0^{10} t^{4.5}e^{-t}dt+\int_{10}^{\infty} t^{4.5}e^{-t}dt = \int_0^{10} t^{4.5}e^{-t}dt+\int_0^{1/10}\frac{1}{t^{6.5}}e^{-1/t}dt$$

Ο πρώτος όρος μπορεί να υπολογισθεί με οποιαδήποτε μέθοδο. Με την Simpson 1/3, για παράδειγμα, η σχετική διαφορά της λύσης με $n=64$ υποδιαστήματα, ως προς εκείνη με $n=32$, είναι $2.8\cdot10^{-7}$, άρα ικανοποιεί το κριτήριο σύγκλισης. Η τελική τιμή είναι $I_1=49.96952160$.

Ο δεύτερος όρος έχει ένα ιδιόμορφο σημείο στο αριστερό όριο ολοκλήρωσης, το σημείο 0, επομένως πρέπει να χρησιμοποιηθεί μια μέθοδος ανοικτού τύπου. Ο κανόνας του Μέσου Σημείου (Εξ. 5.39) απαιτεί $n=256$ υποδιαστήματα για να φθάσει το σχετικό σφάλμα ως προς τη λύση για $n/2$ υποδιαστήματα κοντά στο κριτήριο σύγκλισης ($4.0\cdot10^{-5}$), δίνοντας τελική τιμή $I_2=2.37324899$. Πολύ ταχύτερη είναι αντίθετα η μέθοδος Gauss-Legendre, που δίνει την επιθυμητή ακρίβεια με 15 σημεία (σχετικό σφάλμα μεταξύ 15 και 10 σημείων: $1.2\cdot10^{-6}$) και τελική τιμή $I_2=2.37325685$. Έτσι η ζητούμενη τιμή της συνάρτησης Γάμμα είναι:

$$\Gamma(5.5)=I_1+I_2=49.96952160+2.37325685=52.34277845.$$

5.1.4.3. Ρυθμιζόμενο πλάτος υποδιαστημάτων

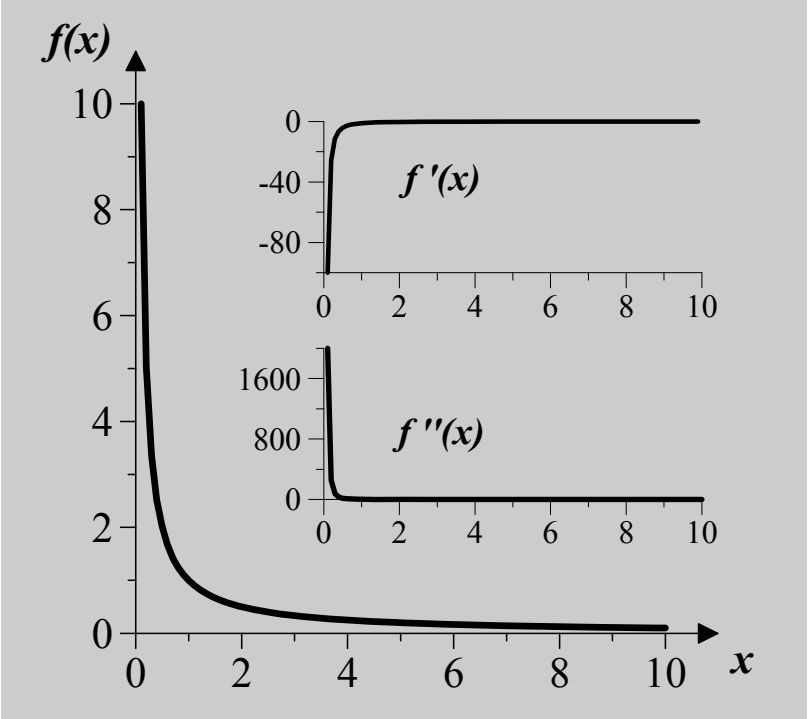
Όσο πιο απότομες είναι οι μεταβολές των τιμών μιας συνάρτησης στο διάστημα ολοκλήρωσης, τόσο μικρότερο πλάτος πρέπει να έχουν τα υποδιαστήματα μιας αριθμητικής μεθόδου για επίτευξη της επιθυμητής ακρίβειας. Πολλές φορές όμως συμβαίνει μια συνάρτηση να έχει τέτοια συμπεριφορά μόνο σε κάποια τμήματα του διαστήματος ολοκλήρωσης, ενώ στο υπόλοιπο να είναι ομαλή. Στην περίπτωση αυτή, η δυνατότητα χρήσης υποδιαστημάτων μεταβλητού πλάτους, δηλαδή μικρότερου στην περιοχή μεγάλων τιμών των παραγώγων της συνάρτησης και μεγαλύτερου στην ομαλή περιοχή, μπορεί να πετύχει σημαντική οικονομία στους υπολογισμούς.

Για την εφαρμογή μιας τέτοιας ρυθμιζόμενης (adaptive) μεθόδου, το διάστημα ολοκλήρωσης χωρίζεται αρχικά σε έναν αριθμό ίσων υποδιαστημάτων. Στη συνέχεια, εφαρμόζεται η μέθοδος ολοκλήρωσης διαδοχικά σε κάθε ένα από αυτά, χωρίζοντάς το σε όλο και περισσότερα υπο-υποδιαστήματα, μέχρι την επίτευξη της επιθυμητής ακρίβειας. Τέλος, αθροίζονται τα επιμέρους ολοκληρώματα όλων των αρχικών υποδιαστημάτων.

Εφαρμογή 5.6.

Να υπολογισθεί η το ολοκλήρωμα $\int_{0.1}^{10} \frac{1}{x} dx$ με απλή και με ρυθμιζόμενη μέθοδο Simpson 1/3, και κριτήριο σύγκλισης $\varepsilon_r = 10^{-5}$.

Όπως φαίνεται στο γράφημα του Σχήματος 5.7, οι παράγωγοι της συνάρτησης 1/*x* έχουν μεγάλες τιμές στην περιοχή [0.1, 1], αλλά πολύ μικρότερες στο υπόλοιπο διάστημα ολοκλήρωσης. Η αναλυτική λύση είναι: ln(10) – ln(0.1) = 4.605170171.



Σχήμα 5.7. Γραφική παράσταση της συνάρτησης f(x) = 1/x, στο [0.1, 10].

Με την απλή μέθοδο Simpson 1/3, η σύγκλιση επιτυγχάνεται για 1024 υποδιαστήματα, δίνοντας αποτέλεσμα $I = 4.605173$. Στη συνέχεια, εφαρμόζεται η ρυθμιζόμενη μέθοδος, χωρίζοντας αρχικά το διάστημα [0.1, 10] σε 20 ίσα υποδιαστήματα. Όπως είναι αναμενόμενο, η σύγκλιση σε κάθε ένα από αυτά απαιτεί διαφορετικό αριθμό υπο-υποδιαστημάτων (π.χ. 64 για το πρώτο και μόνο 4 για το τελευταίο). Έτσι, χρησιμοποιούνται συνολικά 172 υπο-υποδιαστήματα, δίνοντας αποτέλεσμα $I = 4.605172$. Επομένως, η ρυθμιζόμενη μέθοδος είναι σχεδόν έξι φορές ταχύτερη της απλής, για την ίδια ακρίβεια αποτελέσματος.

5.1.4.4. Διπλά και πολλαπλά ολοκληρώματα

Έως τώρα εξετάστηκε η ολοκλήρωση αναλυτικών ή πινακοποιημένων συναρτήσεων μίας ανεξάρτητης μεταβλητής x. Σε πρακτικά προβλήματα όμως, μπορεί να αντιμετωπισθεί η ολοκλήρωση συναρτήσεων περισσότερων ανεξάρτητων μεταβλητών, όπως για παράδειγμα κατά τον υπολογισμό του εμβαδού μιας επιφάνειας ή του όγκου ενός στερεού σώματος. Αν και κάποιες από τις αριθμητικές μεθόδους ολοκλήρωσης που παρουσιάστηκαν μπορούν να αναπτυχθούν και για δύο ή περισσότερες διαστάσεις, οι εκφράσεις θα ήταν πολύπλοκες, αυξάνοντας έτσι τις υπολογιστικές απαιτήσεις. Μια απλή μέθοδος, που εφαρμόζεται όμως εύκολα σε περισσότερες διαστάσεις, είναι ο κανόνας του Μέσου Σημείου (Εξ. 5.39). Ο υπολογισμός για παράδειγμα του ορισμένου ολοκληρώματος μιας συνάρτησης δύο ανεξάρτητων μεταβλητών, μπορεί να γίνει από τη σχέση

$$\int_{x=a}^{x=b}\int_{y=c}^{y=d} f(x,y)\,dx\,dy \cong \frac{(b-a)(d-c)}{n\cdot m}\sum_{i=1}^{n}\sum_{j=1}^{m} f(x_i,y_j) \tag{5.43}$$

όπου n και m ο αριθμός των ίσων υποδιαστημάτων στα οποία χωρίζονται τα διαστήματα ολοκλήρωσης $[a, b]$ και $[c, d]$ αντιστοίχως, και τα σημεία x_i, y_j βρίσκονται στο κέντρο του ορθογωνίου i, j.

Ακολουθεί ένα παράδειγμα αλγορίθμου, που υπολογίζει ένα τέτοιο διπλό ολοκλήρωμα μιας συνάρτησης $f(x, y)$. Είναι φανερό ότι ο Κώδικας 5.4 μπορεί εύκολα να γραφεί και για περισσότερες ανεξάρτητες μεταβλητές.

Κώδικας 5.4. Μέθοδος του Μέσου Σημείου

```
SUBROUTINE MEANP (a, b, c, d, n, m, res)
dx = (b – a) / FLOAT (n)
dy = (d – c) / FLOAT (m)
x = a – dx / 2.
y = c – dy / 2.
total = 0.
DO i = 1, n
DO j = 1, m
      x = x + dx
      y = y + dy
      total = total + f (x, y)
END DO
END DO
res = total * dx * dy
RETURN
END
```

Ο κανόνας του Μέσου Σημείου δεν έχει μεγάλη ακρίβεια, επομένως απαιτεί συνήθως σχετικά μεγάλο αριθμό υποδιαστημάτων για να δώσει την επιθυμητή ακρίβεια αποτελέσματος.

Εφαρμογή 5.7.

Να υπολογισθεί με την απλή μέθοδο Simpson 1/3 (τρία σημεία), καθώς και με τη μέθοδο του Μέσου Σημείου, η τιμή του διπλού ολοκληρώματος

$$\int_0^2 \int_0^2 (x^4 y^3 - 3y^2 + 4x)\, dx\, dy$$

Από την αναλυτική λύση προκύπτει η ακριβής τιμή $I = 25.6$.

Η Εξ. (5.9) της απλής μεθόδου Simpson 1/3 χρησιμοποιείται ως εξής: αρχικά εκφράζεται μόνο για μία ανεξάρτητη μεταβλητή, έστω την x, οπότε ορίζεται μια ενδιάμεση συνάρτηση $g(y)$

$$g(y) = (b-a)\frac{f(x_0, y) + 4f(x_1, y) + f(x_2, y)}{6}$$

Στη συνέχεια, εφαρμόζεται για την ολοκλήρωση της ενδιάμεσης αυτής συνάρτησης, ώστε να προκύψει η τιμή του διπλού ολοκληρώματος

$$I = (d-c)\frac{g(y_0) + 4g(y_1) + g(y_2)}{6}$$

Στο συγκεκριμένο πρόβλημα, με $x_0 = y_0 = 0$, $x_1 = y_1 = 1$, $x_2 = y_2 = 2$, θα είναι:

$g(y) = (20/3)\, y^3 - 6y^2 + 8$ και τελικά $I = 80/3 = 26.6667$.

Δηλαδή η μέθοδος δεν είναι ακριβής, με σφάλμα πάνω από 4%. Βελτίωση της ακρίβειας μπορεί να επιτευχθεί με πολλαπλά υποδιαστήματα και χρήση της συνθετότερης Εξ. (5.17), αντί της (5.9). Για πέντε σημεία ($n = 4$) προκύπτει:

$g(y) = (38.5/6)\, y^3 - 6y^2 + 8$ και $I = 25.6667$, δηλαδή σφάλμα 0.26 %.

Είναι όμως προφανές ότι οι υπολογιστικοί αλγόριθμοι αυτών των περιπτώσεων θα είναι πολύ πιο σύνθετοι από ότι ο Κώδικας 5.4 της μεθόδου του Μέσου Σημείου.

Εφαρμογή του τελευταίου στο ίδιο πρόβλημα, για $n = m$ υποδιαστήματα, δίνει:

$n = m = 2$,	$I = 18.9375$,	$\varepsilon_t = 26$ %
$n = m = 5$,	$I = 24.4176$,	$\varepsilon_t = 4.62$ %
$n = m = 10$,	$I = 25.3010$,	$\varepsilon_t = 1.17$ %
$n = m = 20$,	$I = 25.5247$,	$\varepsilon_t = 0.29$ %
$n = m = 50$,	$I = 25.5879$,	$\varepsilon_t = 0.047$ %
$n = m = 100$,	$I = 25.5970$,	$\varepsilon_t = 0.012$ %
$n = m = 200$,	$I = 25.5993$,	$\varepsilon_t = 0.003$ %

Παρατηρείται δηλαδή ότι η μέθοδος του Μέσου Σημείου χρειάζεται πέντε ÷ έξι και 20 ÷ 25 υποδιαστήματα, ώστε να δώσει αποτέλεσμα ίδιας τάξης ακρίβειας με αυτό της μεθόδου Simpson 1/3 με δύο και τέσσερα υποδιαστήματα αντιστοίχως.

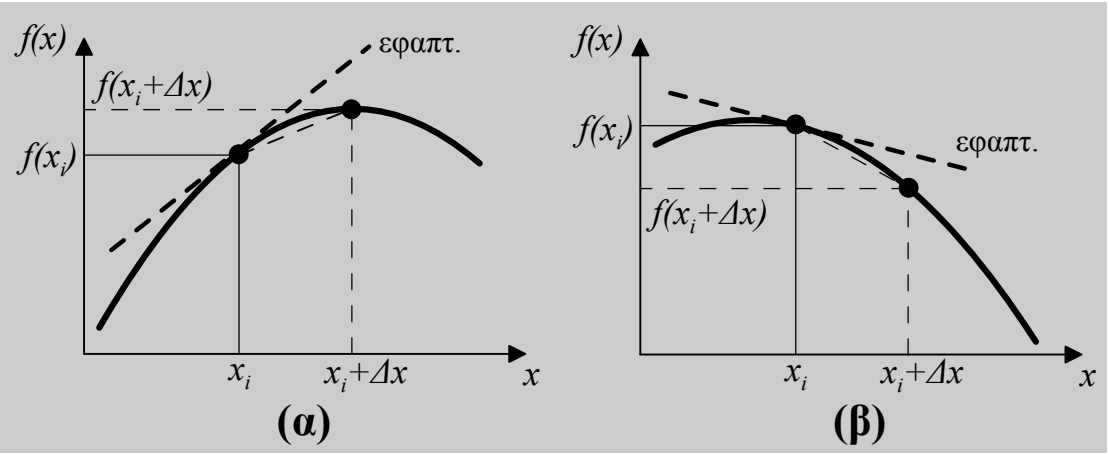
Επίσης, συγκρίνοντας πόσες φορές πρέπει να υπολογισθεί η τιμή της συνάρτησης με κάθε μέθοδο, ώστε το σχετικό σφάλμα να μειωθεί π.χ. κάτω του 0.3%, προκύπτουν οι τιμές: 25 για την Simpson 1/3 και 400 για του Μέσου Σημείου. Εάν επομένως απαιτείται ταχύτητα υπολογισμών, τότε για την αριθμητική ολοκλήρωση μιας σύνθετης συνάρτησης θα πρέπει να προτιμηθεί η πρώτη μέθοδος.

5.2. Αριθμητική Παραγώγιση

Η παράγωγος μιας συνάρτησης $f(x)$ σε ένα σημείο $(x_i, f(x_i))$ συμβολίζεται και ορίζεται μαθηματικά ως εξής:

$$f'(x_i) \equiv \frac{df(x_i)}{dx} \equiv \left(\frac{df}{dx}\right)_{x=x_i} = \lim_{\Delta x \to 0} \frac{f(x_i + \Delta x) - f(x_i)}{\Delta x} \tag{5.44}$$

Η τιμή της παραγώγου αντιπροσωπεύει την τοπική κλίση της καμπύλης της συνάρτησης και ισούται με την τιμή της εφαπτομένης της, όταν η κλίμακα στους άξονες x και $f(x)$ είναι ίδια (Σχήμα 5.8).



Σχήμα 5.8. Γραφική απεικόνιση της παραγώγου μιας συνάρτησης f στο σημείο $x = x_i$: α) θετική παράγωγος και β) αρνητική παράγωγος.

Στη συνέχεια θα περιγραφούν οι υπάρχουσες μέθοδοι για αριθμητικό υπολογισμό της παραγώγου μιας συνάρτησης, όταν δεν είναι δυνατό να βρεθεί αυτή η τιμή αναλυτικά.

5.2.1. Εκφράσεις Πεπερασμένων Διαφορών

Με βάση τον παραπάνω ορισμό, μπορούν να εξαχθούν διάφορες εκφράσεις της παραγώγου μιας συνάρτησης, αναπτύσσοντάς την κατάλληλα σε σειρές Taylor:

$$f(x_{i+1}) = f(x_i) + f'(x_i)h + \frac{f''(x_i)}{2!}h^2 + \ldots + \frac{f^{(n)}(x_i)}{n!}h^n + R_n \tag{5.45}$$

ή $$f(x_{i-1}) = f(x_i) - f'(x_i)h + \frac{f''(x_i)}{2!}h^2 - \ldots \pm \frac{f^{(n)}(x_i)}{n!}h^n + R_n \tag{5.46}$$

όπου έχει τεθεί $h = \Delta x = x_{i+1} - x_i = x_i - x_{i-1}$

Η Εξ. (5.45), όταν περιέχει μόνο τους όρους μηδενικής και πρώτης τάξης, δίνει

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_i)}{h} - \frac{R_1}{h} = \frac{f(x_{i+1}) - f(x_i)}{h} + E_t \tag{5.47}$$

όπου R_1 είναι ο όρος του υπολοίπου της σειράς (Εξ. 1.20) και E_t είναι το άθροισμα των όρων τάξης 2 και άνω, δηλαδή το σφάλμα αποκοπής

$$E_t = -\frac{R_1}{h} = -\frac{1}{h}\frac{f^{(2)}(\xi)}{2!}h^2 = -\frac{f^{(2)}(\xi)}{2!}h \tag{5.48}$$

Επομένως, εάν το διάστημα h είναι σχετικά μικρό και παραλειφθούν οι όροι τάξης δύο και άνω, η παράγωγος της συνάρτησης μπορεί να προσεγγισθεί με σφάλμα τάξης $O(h)$, όσο δηλαδή η εκτιμώμενη τάξη μεγέθους του σφάλματος αποκοπής

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_i)}{h} - O(h) \tag{5.49}$$

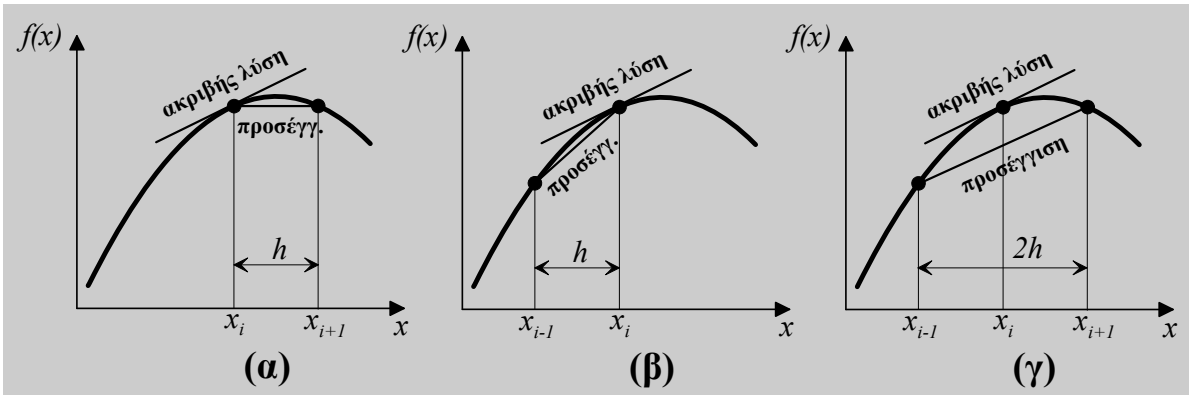
Ανάλογα με την Εξ. (5.48), που αποτελεί πρόσω παραγώγιση (forward difference approximation), προκύπτει και η έκφραση της πίσω παραγώγισης (backward difference approx.)

$$f'(x_i) = \frac{f(x_i) - f(x_{i-1})}{h} + O(h) \tag{5.50}$$

Επίσης, αφαιρώντας κατά μέλη τις Εξ. (5.45) και (5.46), προκύπτει η έκφραση της κεντρικής παραγώγισης (central difference approx.)

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_{i-1})}{2h} - O(h^2) \tag{5.51}$$

η οποία έχει σφάλμα τάξης h^2, είναι δηλαδή ακριβέστερη από τις δύο προηγούμενες. Μια γραφική ερμηνεία αυτής της συμπεριφοράς δίνεται στο Σχήμα 5.9. Υποδιπλασιάζοντας για παράδειγμα το διάστημα h, το σφάλμα της πρόσω και της πίσω παραγώγισης μειώνεται στο μισό, ενώ της κεντρικής στο ένα τέταρτο.



Σχήμα 5.9. Γραφική απεικόνιση των αριθμητικών εκφράσεων της παραγώγου: α) πρόσω παραγώγιση, β) πίσω παραγώγιση και γ) κεντρική παραγώγιση.

Με κατάλληλες αλγεβρικές πράξεις προκύπτουν επίσης και άλλες χρήσιμες εκφράσεις για τις παραγώγους πρώτης και ανώτερης τάξης μιας συνάρτησης, όπως είναι οι ακόλουθες, με την αντίστοιχη τάξη μεγέθους σφάλματος αποκοπής:

Πρόσω παραγώγιση:

$$f'(x_i) \cong \frac{-f(x_{i+2}) + 4f(x_{i+1}) - 3f(x_i)}{2h}, \qquad O(h^2) \tag{5.52}$$

$$f''(x_i) \cong \frac{f(x_{i+2}) - 2f(x_{i+1}) + f(x_i)}{h^2}, \qquad O(h) \tag{5.53}$$

Πίσω παραγώγιση:

$$f'(x_i) \cong \frac{3f(x_i) - 4f(x_{i-1}) + f(x_{i-2})}{2h}, \qquad O(h^2) \tag{5.54}$$

$$f''(x_i) \cong \frac{f(x_i) - 2f(x_{i-1}) + f(x_{i-2})}{h^2}, \qquad O(h) \tag{5.55}$$

Κεντρική παραγώγιση:

$$f'(x_i) \cong \frac{-f(x_{i+2}) + 8f(x_{i+1}) - 8f(x_{i-1}) + f(x_{i-2})}{12h}, \qquad O(h^4) \tag{5.56}$$

$$f''(x_i) \cong \frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{h^2}, \qquad O(h^2) \tag{5.57}$$

$$f^{(3)}(x_i) \cong \frac{f(x_{i+2}) - 2f(x_{i+1}) + 2f(x_{i-1}) - f(x_{i-2})}{2h^3}, \qquad O(h^2) \tag{5.58}$$

Πρέπει όμως να τονισθεί ότι οι εκφράσεις αυτές μπορούν να δώσουν αποδεκτό προσεγγιστικό αποτέλεσμα μόνο όταν η απόσταση των σημείων, h, είναι αρκούντως μικρή.

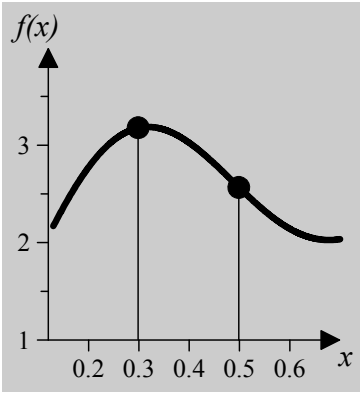
Εφαρμογή 5.8.

Να υπολογισθεί η πρώτη και δεύτερη παράγωγος της Εξ. (5.4) στα σημεία $x = 0.3$ και $x = 0.5$, χρησιμοποιώντας τις παραπάνω εκφράσεις και διαστήματα $h = 0.1$, 0.01, 0.001 και 0.0001.

Στους Πίνακες 5.6 και 5.7 δίνονται οι ακριβείς τιμές των παραγώγων που λαμβάνονται από την αναλυτική λύση, καθώς και τα αριθμητικά αποτελέσματα, όπως προκύπτουν με απλή εφαρμογή των αντίστοιχων εκφράσεων πεπερασμένων διαφορών. Δίνεται επίσης και το σχετικό σφάλμα ε_t %. Η μορφή της Εξ. (5.4) στην περιοχή που ενδιαφέρει σχεδιάζεται στο Σχήμα 5.10.

Πίνακας 5.6. Αριθμητικοί υπολογισμοί της πρώτης παραγώγου της Εξ. (5.4) στα σημεία $x = 0.3$ και $x = 0.5$.

	$x = 0.3$, ακριβείς τιμές: $f'(x) = 1.0400000$, $f''(x) = -58.200000$											
	Απλή προσέγγιση (5.49, 5.50, 5.51)						Διπλή ακρίβεια (5.52, 5.54, 5.56)					
	πρόσω		πίσω		κεντρική		πρόσω		πίσω		κεντρική	
h	f'	$\lvert\varepsilon_t\rvert$ %	f'	$\lvert\varepsilon_t\rvert$ %	f'	$\lvert\varepsilon_t\rvert$ %	f'	$\lvert\varepsilon_t\rvert$ %	f'	$\lvert\varepsilon_t\rvert$ %	f'	$\lvert\varepsilon_t\rvert$ %
0.1	-1.521	246.0	4.049	289.0	1.264	21.5	0.034	96.7	1.534	47.5	1.104	6.15
0.01	0.7515	27.7	1.333	28.2	1.042	0.23	1.034	0.53	1.036	0.39	1.040	6.10^{-4}
0.001	1.011	2.80	1.069	2.8	1.040	$2{\cdot}10^{-3}$	1.040	$5{\cdot}10^{-3}$	1.040	$4{\cdot}10^{-3}$	1.040	6.10^{-8}
	$x = 0.5$, ακριβείς τιμές: $f'(x) = -5.0000000$, $f''(x) = 5.0000000$											
0.1	-4.201	16.0	-4.631	7.4	-4.416	11.7	-5.766	15.3	-6.186	23.7	-4.936	1.28
0.01	-4.969	0.62	-5.019	0.38	-4.994	0.12	-5.012	0.24	-5.012	0.24	-5.000	$1{\cdot}10^{-4}$
0.001	-4.997	0.05	-5.002	0.05	-5.000	$1{\cdot}10^{-3}$	-5.000	$2{\cdot}10^{-3}$	-5.000	$2{\cdot}10^{-3}$	-5.000	$1{\cdot}10^{-8}$



Σχήμα 5.10. Γραφική παράσταση της Εξ. (5.4).

Πίνακας 5.7. Αριθμητικοί υπολογισμοί της $2^{ης}$ παραγώγου της Εξ. (5.4) στο σημείο $x = 0.5$.

	Πρόσω (5.53)		Πίσω (5.55)		Κεντρική (5.57)	
h	f''	$\|\varepsilon_t\|$ %	f''	$\|\varepsilon_t\|$ %	f''	$\|\varepsilon_t\|$ %
0.1	31.30	526.0	-31.1	729.0	4.3	14.0
0.01	8.456	70.9	1.356	72.9	4.993	0.14
0.001	5.360	7.2	4.640	7.2	5.000	$1\cdot10^{-3}$
0.0001	5.036	0.72	4.964	0.72	5.000	$1\cdot10^{-5}$
	Πράξεις απλής ακρίβειας					
0.01	8.545	70.9	1.354	72.9	4.995	0.11
0.001	5.484	9.7	5.007	0.14	4.768	4.63
0.0001	23.84	376.0	23.84	376.0	0.000	100.0

Παρατηρώντας και τα αποτελέσματα της Εφαρμογής 5.8, μπορούν να διατυπωθούν τώρα οι ακόλουθες γενικές παρατηρήσεις σχετικά με τη χρήση των σχέσεων πεπερασμένων διαφορών:

i) Μείωση του διαστήματος από h_1 σε h_2 μειώνει το σφάλμα του αποτελέσματος επί $(h_2/h_1)^n$, όπου n ο εκθέτης του όρου του σφάλματος (π.χ. $n = 1$ για τις Εξ. 5.49, 5.50, 5.53 και 5.55, και $n = 2$ για τις υπόλοιπες, εκτός της Εξ. 5.56 που έχει $n = 4$).
ii) Η κεντρική παραγώγιση είναι ακριβέστερη των δύο άλλων τύπων, για ίδιο h.
iii) Οι εκφράσεις που περιέχουν την τιμή της συνάρτησης σε περισσότερα σημεία έχουν καλλίτερη ακρίβεια από τις αντίστοιχες με δύο σημεία.
iv) Η τάξη μεγέθους του σφάλματος αυξάνει όσο αυξάνει η τάξη της παραγώγισης, για ίδιο h (βλ. και Εξ. 5.52 – 5.53, 5.54 – 5.55, 5.56 – 5.57).

Όταν η τιμή του h είναι σχετικά μεγάλη, τότε οι παραπάνω κανόνες μπορεί να μην ισχύουν πάντοτε. Για παράδειγμα, η πίσω παραγώγιση στο $x = 0.3$ (Πίνακας 5.6) δίνει μικρότερο σφάλμα από την αντίστοιχη κεντρική για $h = 0.1$, ενώ στην ίδια θέση το σφάλμα της πίσω παραγώγισης τριών σημείων είναι επίσης μεγαλύτερο.

Ακόμη, είναι σημαντικό να παρατηρηθεί ότι το μέγεθος του σφάλματος δεν εξαρτάται μόνο από την εξίσωση διαφορών και την τιμή του h, αλλά και από τη θέση του σημείου x, ή αλλιώς από τις τιμές των παραγώγων της συνάρτησης στη θέση x (βλ. Κεφ. 1.3.2). Για παράδειγμα, οι τιμές απλής προσέγγισης της πρώτης παραγώγου στο $x = 0.5$ έχουν πολύ μικρότερο σχετικό σφάλμα από ότι στο $x = 0.3$ (Πίνακας 5.6). Επομένως, η κατάλληλη τιμή του h για την επίτευξη μιας προκαθορισμένης ακρίβειας μπορεί να διαφέρει από σημείο σε σημείο.

Τέλος, στον Πίνακα 5.7 δίνονται και τα αντίστοιχα αποτελέσματα που προκύπτουν με απλή ακρίβεια αριθμών (single precision). Είναι φανερό ότι για πολύ μικρές τιμές του h τα σφάλματα στρογγυλοποίησης –τα οποία οφείλονται εδώ σε προσθέσεις και αφαιρέσεις πολύ κοντινών τιμών της συνάρτησης, που ήδη έχουν υποστεί στρογγυλοποίηση–, μπορεί να γίνουν σημαντικά και να οδηγήσουν σε εντελώς λανθασμένο αποτέλεσμα.

Συνοψίζοντας, πρέπει κατά την αριθμητική παραγώγιση να τηρούνται τα εξής:

- Όταν απαιτείται μεγάλη ακρίβεια του αποτελέσματος, τότε χρειάζεται πολύ μικρό διάστημα h, σε συνδυασμό με σύνθετες εκφράσεις κεντρικών διαφορών και διπλή ακρίβεια μεταβλητών.

- Όταν το διάστημα h είναι εκ των πραγμάτων σχετικά μεγάλο, όπως μπορεί να συμβαίνει σε διακριτά δεδομένα (πινακοποιημένες συναρτήσεις), τότε είναι ασφαλέστερες οι εκφράσεις κεντρικών διαφορών ή, αν πρόκειται για το πρώτο ή το τελευταίο σημείο της σειράς, οι εκφράσεις πρόσω και πίσω παραγώγισης απλής ακρίβειας.

5.2.2. Κριτήριο Σύγκλισης

Επειδή συνήθως η τιμή της παραγώγου σε μια θέση x δεν είναι γνωστή, πρέπει να χρησιμοποιηθεί ένα κριτήριο σύγκλισης, όπως και στην αριθμητική ολοκλήρωση. Ένα απλό κριτήριο είναι η σχετική διαφορά μεταξύ δύο αποτελεσμάτων που λαμβάνονται διαδοχικά για διαστήματα h και $h/2$, να γίνει μικρότερη μιας ορισμένης τιμής ε_r:

$$\left|\varepsilon_a\right| = \left|\frac{f_h^{(m)} - f_{h/2}^{(m)}}{f_{h/2}^{(m)}}\right| \le \varepsilon_r \tag{5.59}$$

Εφαρμογή 5.9.

Να υπολογισθεί η πρώτη παράγωγος της Εξ. (5.4) στο σημείο $x = 0.3$, με χρήση πίσω και κεντρικής παραγώγισης απλής ακρίβειας και με κριτήριο σύγκλισης $\varepsilon_r = 10^{-5}$.

Χρησιμοποιούνται οι Εξ. (5.50) και (5.51) για πίσω και κεντρική παραγώγιση αντιστοίχως. Οι υπολογισμοί αρχίζουν για διάστημα $h = 0.1$ και συνεχίζονται υποδιπλασιάζοντάς το συνεχώς, μέχρι να ικανοποιηθεί η Εξ. (5.59). Τα τελικά αποτελέσματα είναι:

Εξ. (5.50): $f' = 1.0400146$, με $h = 5\cdot10^{-7}$

Εξ. (5.51): $f' = 1.0400060$, με $h = 5\cdot10^{-4}$

Παρατηρείται δηλαδή ότι και τα δύο αποτελέσματα έχουν ακρίβεια τουλάχιστον 5 σημαντικών ψηφίων (ακριβής λύση $f' = 1.0400000$), αλλά η πίσω παραγώγιση χρειάστηκε τρεις τάξεις μεγέθους μικρότερο διάστημα ολοκλήρωσης, επομένως πολύ περισσότερες πράξεις.

5.2.3. Μέθοδος Προεκβολής Richardson

Η τεχνική εκτίμησης σφάλματος του Richardson (Richardson extrapolation) έχει περιγραφεί στο Κεφάλαιο 5.1.2, όπου χρησιμοποιείται στη μέθοδο ολοκλήρωσης του Romberg. Με παρόμοια ανάλυση, μια εξίσωση αντίστοιχη της (5.27) μπορεί να εξαχθεί και για την περίπτωση της παραγώγισης:

$$f^{(m)} \cong \frac{4}{3} f_{h_2}^{(m)} - \frac{1}{3} f_{h_1}^{(m)} \tag{5.60}$$

όπου $h_2 = h_1/2$, f_h είναι οι αντίστοιχες αριθμητικές προσεγγίσεις και f η βελτιωμένη προσέγγιση, η οποία θεωρητικά έχει σφάλμα $O(h^4)$ όταν οι δύο προηγούμενες είναι κεντρικές διαφορές με σφάλμα $O(h^2)$.

Για παράδειγμα, αν εφαρμοσθεί η σχέση κεντρικής παραγώγισης (5.31) στο σημείο $x = 0.3$ της πολυωνυμικής συνάρτησης (5.4), για $h_1 = 0.02$ θα δώσει $f' = 1.0495744$, ενώ για $h_2 = h_1/2 = 0.01$ θα δώσει $f' = 1.0423984$. Και από την Εξ. (5.60) προκύπτει η βελτιωμένη τιμή 1.0400064, η οποία μπορεί να ληφθεί από την Εξ. (5.31) μόνο όταν μειωθεί το διάστημα h σε τιμή περίπου $5\cdot10^{-4}$. Επομένως πράγματι το σφάλμα της μεθόδου είναι $O(h^4)$, αφού $(0.02)^4 \approx (5\cdot10^{-4})^2$.

Τέλος, η Εξ. (5.60) μπορεί να γενικευθεί, όπως και στην περίπτωση της ολοκλήρωσης, δίνοντας μια επαναληπτική σχέση ανάλογη εκείνης του Romberg (Εξ. 5.28).

5.2.4. Παραγώγιση σε Άνισα Διαστήματα

Όλες οι προηγούμενες προσεγγιστικές εκφράσεις των παραγώγων ισχύουν μόνο για ίσα διαστήματα h, δηλαδή όταν τα διαθέσιμα διακριτά δεδομένα ισαπέχουν στον άξονα x. Όμως συχνά τα αποτελέσματα διαφόρων μετρήσεων δεν έχουν αυτή τη μορφή. Σε τέτοιες περιπτώσεις μπορεί να χρησιμοποιηθεί η ακόλουθη μέθοδος: σε κάθε τριάδα γειτονικών σημείων x_{i-1}, x_i και x_{i+1} προσαρμόζεται ένα πολυώνυμο παρεμβολής Lagrange δεύτερου βαθμού (βλ. Κεφ. 4.2.3), το οποίο δεν απαιτεί ίσες αποστάσεις των σημείων. Η παράγωγος του πολυωνύμου αυτού μπορεί να γραφεί αναλυτικά, και μετά από αλγεβρικές πράξεις καταλήγει στην παρακάτω γενική έκφραση:

$$f'(x) \cong f(x_{i+1})\frac{2x - x_i - x_{i-1}}{(x_{i+1} - x_i)(x_{i+1} - x_{i-1})} + f(x_i)\frac{2x - x_{i+1} - x_{i-1}}{(x_i - x_{i+1})(x_i - x_{i-1})} + f(x_{i-1})\frac{2x - x_{i+1} - x_i}{(x_{i-1} - x_{i+1})(x_{i-1} - x_i)} \tag{5.61}$$

Η Εξ. (5.61) έχει την ίδια τάξη σφάλματος με την αντίστοιχη έκφραση κεντρικών διαφορών για ισαπέχοντα σημεία (Εξ. 5.51), είναι όμως γενικότερη αυτής, αφού μπορεί να υπολογίσει την παράγωγο και για μη ισαπέχοντα σημεία, αλλά επίσης και σε κάθε θέση στο διάστημα $[x_{i-1}, x_{i+1}]$. Με άλλα λόγια, η Εξ. (5.51) αποτελεί ειδική περίπτωση της (5.61), όταν η τελευταία εφαρμόζεται στη θέση $x = x_i$ και τα τρία σημεία ισαπέχουν.

Εφαρμογή 5.10.

Ένας αθλητής χρονομετρείται κατά τη διάρκεια αγώνα εκατό μέτρων ως εξής:

x (m)	0	10	20	30	40	50	60	70	80	90	100
t (sec)	0	1.98	3.49	4.61	5.52	6.30	7.04	7.74	8.47	9.23	10.03

Να υπολογισθεί η στιγμιαία ταχύτητα και επιτάχυνσή του ανά δέκα μέτρα, και να σχεδιαστούν τα αντίστοιχα γραφήματα ως προς x.

Για την ταχύτητα u και την επιτάχυνση b ισχύουν οι σχέσεις:

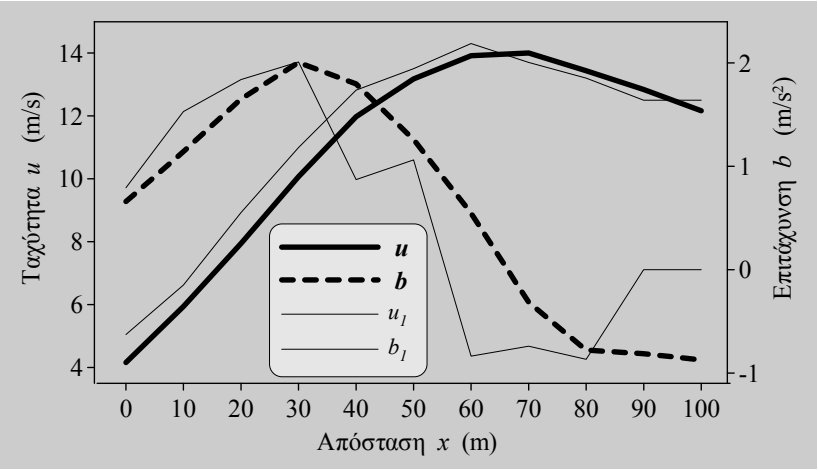
$$u(t) = x'(t) = \frac{dx}{dt}, \qquad b = u'(t) = \frac{du}{dt} = \frac{d^2x}{dt^2}$$

Επειδή τα δεδομένα για την ανεξάρτητη μεταβλητή (που εδώ είναι ο χρόνος) δεν ισαπέχουν, θα εφαρμοσθεί η μέθοδος παραγώγισης σε άνισα υποδιαστήματα,

δηλαδή η Εξ. (5.61). Η τιμή της ταχύτητας λαμβάνεται κάθε φορά στο μεσαίο εκ των τριών γειτονικών χρονικών σημείων, εκτός από τις θέσεις $t = 0$ και $t = 100$, οπότε βρίσκεται στο πρώτο και στο τρίτο σημείο αντιστοίχως. Στη συνέχεια, με βάση τις ταχύτητες, προκύπτουν παρόμοια και οι τιμές της επιτάχυνσης, στις αντίστοιχες χρονικές στιγμές. Τα αποτελέσματα συγκεντρώνονται στον Πίνακα 5.8, από όπου γίνονται και τα γραφήματα $u - x$ και $b - x$, του Σχήματος 5.11.

Πίνακας 5.8. Αριθμητικός υπολογισμός ταχύτητας και επιτάχυνσης ενός αθλητή.

t (sec)	0	1.98	3.49	4.61	5.52	6.30	7.04	7.74	8.47	9.23	10.03
$x\,(t)$ (m)	0	10	20	30	40	50	60	70	80	90	100
$u\,(t)$ (m/s)	4.159	5.942	7.946	10.06	11.97	13.18	13.91	14.00	13.43	12.84	12.16
$b\,(t)$ (m/s^2)	0.659	1.143	1.651	2.006	1.798	1.259	0.547	–0.314	–0.779	–0.813	–0.874



Σχήμα 1.11. Μεταβολή της ταχύτητας και της επιτάχυνσης του αθλητή.

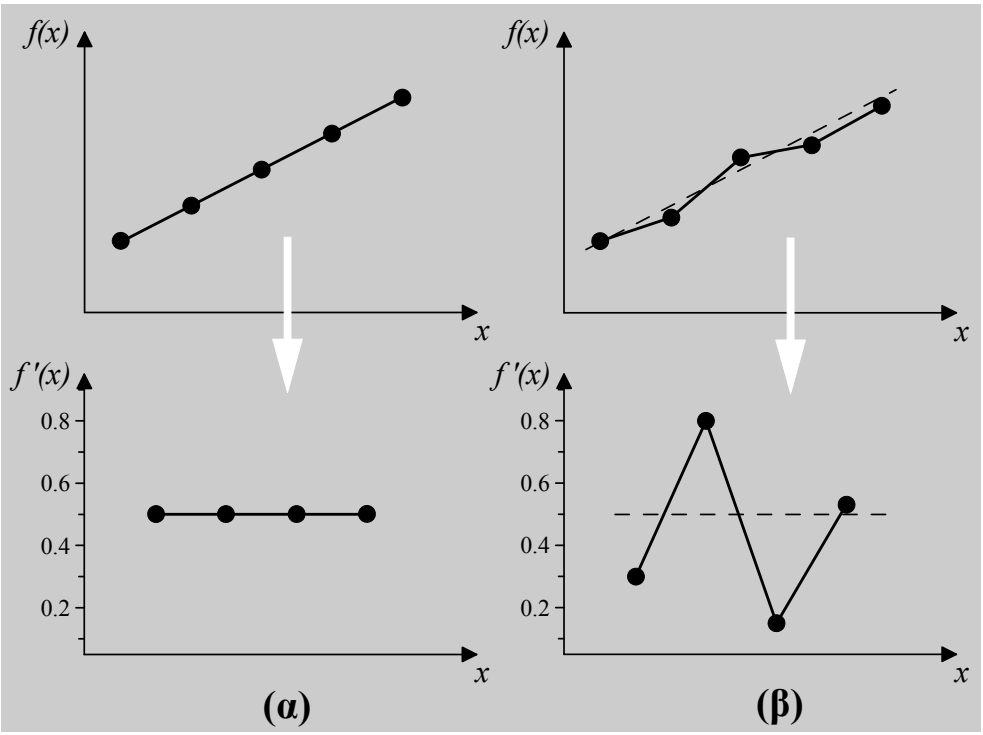
Είναι φανερό ότι η γραφική παράσταση των αποτελεσμάτων μπορεί να δώσει μια πολύ παραστατική εικόνα της συμπεριφοράς και επίδοσης του αθλητή κατά τη διάρκεια του αγωνίσματος.

Τέλος, στο Σχήμα 1.11 σχεδιάζονται επίσης και τα αντίστοιχα γραφήματα για την ταχύτητα u_1 και επιτάχυνση b_1, που προκύπτουν με πρόσω παραγώγιση πρώτης τάξης (Εξ. 5.49), η οποία είναι εφαρμόσιμη, αφού χρησιμοποιεί μόνο δύο σημεία (ένα διάστημα). Όμως η μειωμένη ακρίβεια αυτής της έκφρασης (σφάλμα τάξης h) προκαλεί κάποιο σφάλμα στην ταχύτητα, το οποίοι γίνεται πολύ μεγαλύτερο στην επιτάχυνση, όσο δηλαδή αυξάνει η τάξη της παραγώγισης.

5.2.5. Παραγώγιση σε Διακριτά Δεδομένα με Σφάλματα

Ακόμη κι όταν τα διακριτά δεδομένα κάποιας μέτρησης ισαπέχουν, χρειάζεται προσοχή κατά τον αριθμητικό υπολογισμό της παραγώγου, επειδή τα τυχαία σφάλματα που συνήθως υπάρχουν σε μία μέτρηση μπορεί να μεγεθύνονται πολύ κατά την παραγώγιση, η οποία είναι πράξη αφαίρεσης (βλ. Κεφ. 1.2.3).

Στο Σχήμα 5.12 δίνεται γραφικά η ερμηνεία του μηχανισμού αυτού: όταν τα δεδομένα είναι ακριβή, τότε και η παράγωγος υπολογίζεται σωστά. Αντίθετα, μικρά σφάλματα στα δεδομένα προκαλούν πολύ μεγαλύτερο σφάλμα στην παράγωγο.



Σχήμα 5.11. Παράδειγμα αριθμητικής παραγώγισης πινακοποιημένης συνάρτησης, με κεντρικές διαφορές: α) δεδομένα χωρίς σφάλματα και β) δεδομένα με μικρό σφάλμα.

Για να αποφευχθούν τέτοια προβλήματα πρέπει, όταν τα δεδομένα εμπεριέχουν σφάλματα, να προσαρμόζεται πρώτα σε αυτά μια συνάρτηση, με χρήση κάποιας αριθμητικής μεθόδου προσέγγισης (βλ. Κεφ. 4.4) και στη συνέχεια να υπολογίζεται αναλυτικά ή αριθμητικά η παράγωγός της στο ζητούμενο σημείο. Μια τέτοια πρακτική αντιμετωπίζει ταυτόχρονα και το πρόβλημα ενδεχόμενης ύπαρξης άνισων διαστημάτων.

Στο Σχήμα 5.12 μπορεί επίσης να διαπιστωθεί ότι ακριβώς το αντίθετο συμβαίνει κατά την αριθμητική ολοκλήρωση, η οποία ουσιαστικά είναι πράξη αντίστροφη της παραγώγισης: τα σφάλματα των δεδομένων όχι μόνο δεν αυξάνονται, αλλά εξομαλύνονται σημαντικά.

5.3. Ανακεφαλαίωση

Όλες οι αριθμητικές μέθοδοι ολοκλήρωσης που παρουσιάστηκαν στα προηγούμενα κεφάλαια είναι προσεγγιστικές και η ακρίβειά τους εξαρτάται από το βήμα ολοκλήρωσης ή αλλιώς, από τον αριθμό των ζευγών δεδομένων $(x, f(x))$ που διατίθενται μέσα σε ένα διάστημα ολοκλήρωσης $[a, b]$. Αντίστοιχα, η ακρίβεια των εκφράσεων αριθμητικής παραγώγισης εξαρτάται από το διάστημα Δx, δηλαδή την απόσταση μεταξύ των γειτονικών δεδομένων που χρησιμοποιούνται.

Στον Πίνακα 5.9 συγκρίνονται μερικά βασικά χαρακτηριστικά των μεθόδων ολοκλήρωσης. Η παράμετρος j της μεθόδου Romberg είναι ο αριθμός των επιπέδων ακρίβειας. Επίσης, η μέθοδος Gauss Legendre μπορεί να χρησιμοποιεί ένα ή περισσότερα υποδιαστήματα, με 2 ή περισσότερα σημεία στο κάθε ένα.

Σημειώνεται ότι το πραγματικό σφάλμα κατά την εφαρμογή μιας μεθόδου μπορεί κατά περίπτωση να διαφέρει αρκετά από την εκτιμώμενη τάξη του, επομένως η τελευταία εκφράζει κυρίως τη σχετική ακρίβεια της μεθόδου. Για παράδειγμα, η μέθοδος Romberg δύο επιπέδων έχει ίδια τάξη σφάλματος και ακρίβειας με την Simpson 1/3, ενώ η μέθοδος Gauss-Legendre με δύο μόνο σημεία αντιστοιχεί στη μέθοδο Romberg τριών επιπέδων.

Πίνακας 5.9. Σύγκριση χαρακτηριστικών των μεθόδων αριθμητικής ολοκλήρωσης.

Χαρακτηριστικά	**Τραπεζίου / Μέσου Σημείου**	**Simpson 1/3**	**Simpson 3/8**	**Romberg**	**Gauss-Legendre**
Σημεία για απλή εφαρμογή	2	3	4	3	$2 \div n$
Τάξη σφάλματος	h^2	h^4	h^4	h^{2j}	h^{2n+2}
Δυνατός αριθμός υποδιαστημάτων	κάθε	Άρτιος	πολλαπλ. του 3	πολλαπλ. του $2^{(j-1)}$	κάθε
Πινακοποιημένες συναρτήσεις	ναι / όχι	Ναι	ναι	όχι	όχι

Στην περίπτωση συναρτήσεων που έχουν αναλυτική έκφραση, όλες οι μέθοδοι ολοκλήρωσης μπορούν να δώσουν όση ακρίβεια είναι επιθυμητή, αλλά με διαφορετικές υπολογιστικές απαιτήσεις η κάθε μία. Οι διαφορές μάλιστα αυτές γίνονται εντονότερες όσο πιο πολύπλοκη (και επομένως χρονοβόρα στον υπολογισμό) είναι η μαθηματική έκφραση. Έτσι η επιλογή της καταλληλότερης μεθόδου εξαρτάται από το συγκεκριμένο πρόβλημα.

Όταν η ολοκλήρωση πρόκειται να γίνει μία μόνο φορά ή, γενικότερα, όταν δεν ενδιαφέρει ο υπολογιστικός χρόνος, τότε είναι προτιμότερη η εφαρμογή των γενικών μεθόδων, του Τραπεζίου, του Μέσου Σημείου ή καλλίτερα της ακριβέστερης, Simpson 1/3, με παράλληλη χρήση μεγάλου αριθμού υποδιαστημάτων και αριθμητικής διπλής ακρίβειας. Όλες αυτές οι μέθοδοι προγραμματίζονται πολύ εύκολα και δεν παρουσιάζουν πρόβλημα σύγκλισης.

Αντίθετα, όταν πρόκειται για πολλαπλή εφαρμογή μιας μεθόδου ολοκλήρωσης, τότε ενδιαφέρει κυρίως η ταχύτητα εκτέλεσης των υπολογισμών. Έτσι, η μέθοδος Romberg και η ολοκλήρωση κατά Gauss πρέπει να δοκιμαστούν και να επιλεχθεί η ταχύτερη στο εκάστοτε πρόβλημα. Επαναλαμβάνεται εδώ ότι υπάρχουν διάφορες μέθοδοι ολοκλήρωσης κατά Gauss, κάθε μία από τις οποίες είναι ταχύτερη/ακριβέστερη για συγκεκριμένη κατηγορία συναρτήσεων. Επίσης, ακόμη μεγαλύτερη ταχύτητα μπορεί πολλές φορές να επιτευχθεί με έναν αλγόριθμο ο οποίος ρυθμίζει αυτόματα το πλάτος των υποδιαστημάτων.

Οι γενικές μέθοδοι μπορεί επίσης να μην είναι κατάλληλες και όταν απαιτείται πολύ μεγάλη ακρίβεια αποτελέσματος με χρήση αριθμών απλής ακρίβειας, οπότε το σφάλμα στρογγυλοποίησης μπορεί, λόγω των πολλών αριθμητικών πράξεων, να γίνει σημαντικό.

Στην περίπτωση τώρα των πινακοποιημένων συναρτήσεων, όπου ο αριθμός των δεδομένων είναι συγκεκριμένος, οι επιλογές είναι πιο περιορισμένες, καθώς μόνο οι γενικές μέθοδοι του Τραπεζίου και του Simpson είναι συνήθως εφαρμόσιμοι, και μάλιστα, μόνο η πρώτη εξ αυτών, όταν τα σημεία της ανεξάρτητης μεταβλητής δεν ισαπέχουν. Πρέπει πάντως να σημειωθεί ότι η απαίτηση για ίσα υποδιαστήματα δεν είναι αυστηρή, δηλαδή οι μέθοδοι Simpson μπορούν να χρησιμοποιηθούν και σε άνισα υποδιαστήματα, αλλά με μεγαλύτερο, λόγω αυτού του γεγονότος, σφάλμα. Έτσι, είναι δυνατόν η μέθοδος Simpson 1/3, εφαρμοζόμενη σε παραπλήσια αλλά όχι ίσα υποδιαστήματα, να παραμένει πολύ πιο ακριβής από τον κανόνα του Τραπεζίου.

Η εφαρμογή των ταχύτερων μεθόδων Romberg και Gauss σε διακριτά δεδομένα είναι δυνατή μόνο για συγκεκριμένο αριθμό ή συγκεκριμένη θέση των δεδομένων, αντιστοίχως. Έτσι, κάποιες φορές (π.χ. σε μια διαδικασία μετρήσεων) μπορεί να επιδιωχθεί τα δεδομένα να λαμβάνονται έτσι, ώστε να είναι στη συνέχεια εφαρμόσιμη κάποια από αυτές τις μεθόδους ολοκλήρωσης. Όταν πρόκειται για δεδομένα με σφάλμα ή διακύμανση, τότε μπορεί η ολοκλήρωση να γίνει σε καμπύλες προσέγγισης ή παρεμβολής τους.

Ανάλογα ισχύουν και για την αριθμητική παραγώγιση. Έτσι, στην περίπτωση αναλυτικών συναρτήσεων, τόσο οι απλές, όσο και οι σύνθετες εκφράσεις των παραγώγων μπορούν να δώσουν την επιθυμητή ακρίβεια αποτελέσματος, εάν το μέγεθος του διαστήματος h γίνει αρκούντως μικρό. Όμως οι εκφράσεις των κεντρικών διαφορών εμφανίζουν μικρότερο σφάλμα και συγκλίνουν ταχύτερα, χωρίς να απαιτούν πολύ μικρές τιμές του h, ούτε αριθμητική διπλής ακρίβειας. Το ίδιο ισχύει και για τις άλλες σύνθετες εκφράσεις που δίνονται στο κεφ. 5.2.1, ενώ εκφράσεις με ακόμη περισσότερα σημεία, αν και ακριβέστερες, είναι συνήθως ασύμφορες υπολογιστικά, ιδιαίτερα για πολλαπλή παραγώγιση μιας συνάρτησης με πολύπλοκη μαθηματική έκφραση.

Σε πινακοποιημένες συναρτήσεις υπάρχει, όπως αναφέρθηκε, κίνδυνος μεγέθυνσης τυχόν σφάλματος ή ανακρίβειας των δεδομένων κατά την παραγώγιση. Έτσι, πρέπει είτε να προηγηθεί η προσαρμογή μιας καμπύλης, είτε να προτιμούνται οι απλούστερες εκφράσεις, και μάλιστα κεντρικής διαφοράς όπου αυτό είναι δυνατόν, ιδιαίτερα όταν το διάστημα h είναι σχετικά μεγάλο. Σημειώνεται επίσης ότι το σφάλμα μπορεί να μεγαλώνει σημαντικά όσο ανεβαίνει η τάξη της παραγώγισης.

Τέλος, τονίζεται ότι η ακρίβεια της προσέγγισης κάθε μεθόδου ολοκλήρωσης ή παραγώγισης θα πρέπει σε κάθε περίπτωση να ελέγχεται, επειδή σε πρακτικά προβλήματα δεν είναι γνωστή εκ των προτέρων η ακριβής τιμή του αποτελέσματος. Έτσι, εάν η συνάρτηση είναι αναλυτική, τότε όλες οι μέθοδοι μπορούν να εφαρμοστούν επαναληπτικά, σε προοδευτικά αυξανόμενο (συνήθως διπλάσιο) αριθμό υποδιαστημάτων κατά την ολοκλήρωση, ή μειούμενο (συνήθως στο μισό) διάστημα κατά την παραγώγιση, έως ότου ικανοποιηθεί ένα προκαθορισμένο κριτήριο σύγκλισης. Ειδικά για τη μέθοδο Romberg, αυτό αντιστοιχεί σε προσθήκη κάθε φορά ενός επιπέδου ακρίβειας, ενώ για τη μέθοδο Gauss Legendre, σε πρόσθεση ενός σημείου. Το κριτήριο σύγκλισης πρέπει να είναι συμβατό με την ακρίβεια του εκάστοτε υπολογιστή (βλ. Κεφ. 1.2.2), ώστε να εξασφαλίζεται η σύγκλιση της επαναληπτικής διαδικασίας.

Όταν πρόκειται για πινακοποιημένη συνάρτηση, ο έλεγχος της ακρίβειας μπορεί να γίνει συγκρίνοντας τα αποτελέσματα διαφορετικών μεθόδων ολοκλήρωσης ή εκφράσεων παραγώγισης για την ίδια συνάρτηση, ή χρησιμοποιώντας κάποιο εμπορικό πακέτο (π.χ. Mathcad, Matlab κ.ά.).

Κεφάλαιο 6

Αριθμητική Επίλυση Συνήθων Διαφορικών Εξισώσεων

Πολλές φυσικές διεργασίες μπορούν να μοντελοποιηθούν χρησιμοποιώντας συνήθεις διαφορικές εξισώσεις (*σ.δ.ε.*, ordinary differential equations) . Ανάλογα με τη μορφή που αυτές παίρνουν, άλλες φορές μπορούν να λυθούν αναλυτικά ενώ κάποιες άλλες (που κατά κανόνα είναι και οι περισσότερες) η αναλυτική λύση δεν είναι εφικτή και ο μηχανικός αναγκάζεται να καταφύγει στην αριθμητική τους επίλυση. Τρόπους αριθμητικής επίλυσης *σ.δ.ε.* θα αναλύσουμε στο παρόν κεφάλαιο. Προηγούμενα όμως, θα δώσουμε μερικούς χρήσιμους ορισμούς.

6.1 Χρήσιμοι Ορισμοί

Ονομάζουμε *σ.δ.ε.* n-ιοστής τάξης, κάθε εξίσωση της μορφής

$$F\left(x\ ,\ y\ ,\ \frac{dy}{dx}\ ,\ \frac{d^2y}{dx^2}\ ,\ \ldots \frac{d^ny}{dx^n}\right)\ =\ 0 \tag{6.1}$$

στην οποία η υψηλότερης τάξης παράγωγος που εμφανίζεται είναι η n–ιοστή παράγωγος της εξαρτημένης μεταβλητής (αγνώστου) y ως προς την ανεξάρτητη μεταβλητή x. Στην εξίσωση 6.1 εμπλέκεται μόνο μια ανεξάρτητη μεταβλητή, η x, επομένως η εξίσωση δεν εμπλέκει μερικές παραγώγους και ως εκ τούτου αποτελεί μια *συνήθη διαφορική εξίσωση*.

Η ειδική περίπτωση κατά την οποία η 6.1 μπορεί να διατυπωθεί στη μορφή

$$a_n\frac{d^ny}{dx^n}\ +\ a_{n-1}\frac{d^{n-1}y}{dx^{n-1}}\ +\ \ldots\ +\ a_2\frac{d^2y}{dx^2}\ +\ a_1\frac{dy}{dx}\ +\ a_0\ =\ 0$$

με τους συντελεστές a_i να συναρτώνται μόνο της ανεξάρτητης μεταβλητής x, ονομάζεται *γραμμική σ.δ.ε.*.

Ως λύση της 6.1 αναζητούμε μια συνάρτηση $y = y(x)$, η οποία προφανώς πρέπει να είναι n φορές διαφορίσιμη και η οποία ικανοποιεί την 6.1. Αναζητώντας αριθμητική

λύση της 6.1, γίνεται κατανοητό ότι η λύση $y(x)$ θα είναι σε διακριτή και όχι σε αναλυτική μορφή. Επειδή η επίλυση μιας *σ.δ.ε.* είναι αλληλένδετη με την ολοκλήρωσή της και επειδή σε κάθε ολοκλήρωση εμπλέκεται και η τιμή της σταθερής ολοκλήρωσης, η ύπαρξη *μοναδικής λύσης* κατά την αριθμητική επίλυση μιας *σ.δ.ε.* απαιτεί τον καθορισμό συμπληρωματικών πληροφοριών για τη συνάρτηση $y(x)$. Οι επιπλέον πληροφορίες αφορούν στον καθορισμό της συνάρτησης y ή/και των τιμών κάποιων παραγώγων της σε συγκεκριμένες τιμές της ανεξάρτητης μεταβλητής x.

Ορίζουμε ως:

1. *πρόβλημα αρχικών τιμών* (initial-value problem), το πρόβλημα της αριθμητικής επίλυσης μιας *σ.δ.ε.* στο οποίο έχουν καθοριστεί τιμές του $y(x)$ ή των παραγώγων του στην ίδια τιμή $x = x_0$ της ανεξάρτητης μεταβλητής. Οι με τον τρόπο αυτό καθοριζόμενες τιμές αποτελούν τις *αρχικές συνθήκες* του προβλήματος.

2. *πρόβλημα οριακών τιμών* (ή *συνοριακών τιμών*, boundary value problem), το πρόβλημα της αριθμητικής επίλυσης μιας *σ.δ.ε.* με καθορισμένες τιμές του y ή/και των παραγώγων του σε περισσότερες της μιας θέσης του x. Οι έτσι καθοριζόμενες τιμές αποτελούν τις *οριακές ή συνοριακές συνθήκες* του προβλήματος.

6.2 Συνήθεις Διαφορικές Εξισώσεις Μεγαλύτερης Τάξης

Κάθε *σ.δ.ε.* n–ιοστής τάξης αναλύεται σε ένα σύστημα n *σ.δ.ε.* πρώτης τάξης, ορίζοντας $n-1$ βοηθητικές μεταβλητές. Η ιδιότητα αυτή είναι σημαντική γιατί επιτρέπει να αντιμετωπίσουμε οποιαδήποτε *σ.δ.ε.* οσοδήποτε υψηλής τάξης, ως ένα σύστημα *σ.δ.ε.* πρώτης τάξης. Κατά συνέπεια, στο κεφάλαιο αυτό θα ασχοληθούμε μόνο με την αριθμητική επίλυση *σ.δ.ε.* πρώτης τάξης. Αυτές θα τις συμβολίζουμε γενικά με δύο τρόπους, είτε ως

$$F\left(x\ ,\ y\ ,\ \frac{dy}{dx}\right)\ =\ 0 \tag{6.2}$$

είτε ως

$$\frac{dy}{dx}\ =\ f\left(x\ ,\ y\ \right) \tag{6.3}$$

Ακολουθεί ένα χαρακτηριστικό παράδειγμα μετατροπής μιας *σ.δ.ε.* δεύτερης τάξης σε ένα σύστημα δύο *σ.δ.ε.* πρώτης τάξης. Έτσι η εξίσωση

$$x^2\ \frac{d^2y}{dx^2}\ +\ x\ \frac{dy}{dx}\ +\ xy\ =\ 0$$

Με τον ορισμό της βοηθητικής ποσότητας

$$\phi \;=\; \frac{dy}{dx}$$

δίνει το σύστημα

$$\begin{aligned} \frac{dy}{dx} \;-\; \phi \;&=\; 0 \\ x^2\,\frac{d\phi}{dx} \;+\; x\phi \;+\; xy \;&=\; 0 \end{aligned}$$

6.3 Αριθμητική Επίλυση Συνήθων Διαφορικών Εξισώσεων και Διακριτοποίηση

Ως αριθμητική επίλυση της 6.3 με προκαθορισμένες αρχικές συνθήκες, όπως λ.χ.

$$y(x_0) \;=\; y_0 \;=\; known \tag{6.4}$$

κατανοούμε την αναζήτηση μιας συνάρτησης $y(x)$ που θα ικανοποιεί τις 6.3 και 6.4, ορισμένης σε ένα κλειστό διάστημα $[\alpha, \beta]$, όπου συνήθως στα προβλήματα αρχικών τιμών είναι $\alpha \equiv x_0$. Αναζητώντας αριθμητικά μια λύση $y(x)$, η οποία είναι αδύνατο να βρεθεί αναλυτικά, δηλαδή στη μορφή συνάρτησης $f(x)$, στόχος μας είναι ουσιαστικά να βρεθούν οι τιμές της $y(x)$ σε διακριτές θέσεισ-σημεία του διαστήματος $[\alpha, \beta]$. Επομένως, υπεισέρχεται άμεσα η έννοια της διακριτοποίησης ή διακριτής διαχείρισης της λύσης. Το διάστημα $[\alpha, \beta]$ διακριτοποιείται σε N υποδιαστήματα χρησιμοποιώντας $N+1$ διακριτά σημεία x_i , $i = 0, \ldots, N$, όπου $x_0 = \alpha$ και $x_N = \beta$. Τα σημεία αυτά δεν είναι υποχρεωτικό να ισαπέχουν, όμως η παραδοχή ότι αυτά ισαπέχουν βοηθά πολύ την αριθμητική επίλυση και ασφαλώς διευκολύνει την παρουσίαση που θα ακολουθήσει. Θα θεωρήσουμε, λοιπόν, για τη συνέχεια ότι

$$x_i \;=\; x_0 \;+\; i\,\Delta x \qquad , \quad i = 0, 1, 2, \ldots, N \tag{6.5}$$

με το

$$\Delta x \;=\; \frac{\beta - \alpha}{N} \tag{6.6}$$

να αποτελεί το (σταθερό) βήμα διακριτοποίησης.

Μέσω της διακριτοποίησης , αναζητούμε τις τιμές της συνάρτησησ–λύσης στα $N+1$ διακριτά σημεία x_i , $i = 0, \ldots, N$. Ας συμβολίσουμε με y_i , $i = 0, \ldots, N$, τις τιμέσ–λύσεις που προκύπτουν από την αριθμητική επίλυση της *σ.δ.ε.* στα $N+1$ αυτά σημεία. Θα ονομάσουμε *σφάλμα διακριτοποίησης* ή *αποκοπής* (discretization ή truncation error) τη διαφορά ανάμεσα στην αριθμητικά υπολογισμένη τιμή y_i της συνάρτησης και στην πραγματική της τιμή (ας συμβολίζεται με $y(x_i)$). Θα είναι δηλαδή

$$\epsilon \;=\; y_i \;-\; y(x_i) \tag{6.7}$$

και, για ευνόητους λόγους, θα αποκαλείται και *τοπικό σφάλμα αποκοπής*. Το σφάλμα αποκοπής εξαρτάται από τη μέθοδο αριθμητικής επίλυσης της *σ.δ.ε.* που επιλέξαμε, και μόνο από αυτή, δεν εξαρτάται δηλαδή από τα χαρακτηριστικά των υπολογιστικών μέσων (υλικό, hardware) που χρησιμοποιούμε και είναι ευνόητη η επιθυμία μας να το ελαχιστοποιήσουμε.

Παρενθετικά, και επειδή προηγούμενα έγινε αναφορά στο υπολογιστικό σύστημα που χρησιμοποιείται για την αριθμητική επίλυση, ας επισημάνουμε το γεγονός ότι το τελευταίο είναι υπεύθυνο για το *σφάλμα στρογγυλοποίησης* (round–off error), που οφείλεται στην στρογγυλοποίηση των αριθμών με περισσότερα σημαντικά δεκαδικά ψηφία από αυτά που μπορεί να αποθηκεύσει ο ηλεκτρονικός υπολογιστής. Στο κεφάλαιο αυτό θα αγνοήσουμε το σφάλμα στρογγυλοποίησης και την επίδραση του στην αριθμητική επίλυση *σ.δ.ε.* και θα επικεντρωθούμε μόνο στο σφάλμα αποκοπής, το σφάλμα δηλαδή το οποίο ο μηχανικός μπορεί να 'συγκρατήσει' σε επιθυμητά επίπεδα, με προσεκτική επιλογή της μεθόδου αριθμητικής επίλυσης ή/και της τιμής του βήματος Δx.

6.4 Ταξινόμηση Μεθόδων Αριθμητικής Επίλυσης *σ.δ.ε.*

Μια πρώτη ταξινόμηση των μεθόδων αριθμητικής ολοκλήρωσης *σ.δ.ε.* ορίζει δύο διαφορετικές κατηγορίες επιλυτών:

1. αυτών που βασίζονται στην άμεση ή έμμεση χρήση των κατά Taylor αναπτυγμάτων της συνάρτησης $y(x)$, και

2. αυτών που χρησιμοποιούν σχήματα αριθμητικής ολοκλήρωσης, όπως αυτά που παρουσιάζονται σε άλλο κεφάλαιο.

Από μια διαφορετική σκοπιά, οι μέθοδοι αριθμητικής ολοκλήρωσης *σ.δ.ε.* διακρίνονται σε

1. μεθόδους *απλού βήματος* (ή *ενός βήματος*, one-step methods) και

2. μεθόδους *πολλών ή πολλαπλών βημάτων* (multi-step methods).

Οι μέθοδοι απλού βήματος (με χαρακτηριστικό αντιπρόσωπο τις μεθόδους Runge-Kutta) επιτρέπουν τον υπολογισμό της τιμής y_{i+1} της λύσης στο σημείο x_{i+1} με πληροφορία η οποία αντλείται μόνο από το διάστημα $[x_i, x_{i+1}]$. Αντίθετα, οι μέθοδοι πολλαπλών βημάτων απαιτούν πληροφορία, είτε για τιμές y_i είτε για τιμές f_i (βλ. εξίσωση 6.3) και έξω από το διάστημα $[x_i, x_{i+1}]$. Χαρακτηριστικός αντιπρόσωπος των μεθόδων πολλαπλών βημάτων είναι οι *μέθοδοι πρόβλεψης - διόρθωσης* (predictor - corrector). Σχόλια, που κυρίως αφορούν συγκριτικά πλεονεκτήματα και μειονεκτήματα των διαφορετικών ομάδων τεχνικών θα δοθούν αργότερα, μετά την παρουσίαση τους.

Η παρουσίαση που ακολουθεί ξεκινά με αναφορά στις τεχνικές εκείνες που χρησιμοποιούν αναπτύγματα κατά Taylor, συνεχίζει με την απλή (μη εφαρμοζόμενη συνήθως

όταν υπάρχουν απαιτήσεις ακριβείας, όμως χρήσιμη για την εξοικείωση του αναγνώστη με τις συναφείς τεχνικές) μέθοδο Euler και ολοκληρώνεται με τεχνικές τύπου Runge-Kutta και πρόβλεψης - διόρθωσης.

6.5 Αριθμητική Επίλυση $\sigma.\delta.\varepsilon.$ με Αναπτύγματα Taylor

Για να προσεγγίσουμε αριθμητικά τη λύση της 6.3 στο διακριτοποιημένο με $N+1$ ισαπέχοντα σημεία (N ισομήκη διαστήματα) διάστημα $[\alpha, \beta]$ την αναπτύσσουμε κατά Taylor με αναφορά ένα σημείο x_0. Είναι

$$y(x_0+\Delta x) \;=\; y(x_0) + \frac{\Delta x}{1!} f\left(x_0, y(x_0)\right) + \frac{\Delta x^2}{2!} f^{'}\left(x_0, y(x_0)\right) + \frac{\Delta x^3}{3!} f^{''}\left(x_0, y(x_0)\right) + \ldots \tag{6.8}$$

όπου μπορεί να επιλεγεί ο όρος στον οποίο θα γίνει αποκοπή, ανάλογα με την επιθυμητή ακρίβεια και το αποδεκτό υπολογιστικό κόστος. Τις παραγώγους της συνάρτησης f, για τις οποίες υιοθετήθηκε ο συμβολισμός

$$\begin{aligned} f^{'}\left(x, y(x)\right) \;&=\; \frac{d}{dx} f\left(x, y(x)\right) \\ f^{''}\left(x, y(x)\right) \;&=\; \frac{d^2}{dx^2} f\left(x, y(x)\right) \end{aligned} \tag{6.9}$$

κ.ο.κ., τις υπολογίζουμε εφαρμόζοντας τον κανόνα της αλυσιδωτής παραγώγισης, δεδομένης της εξάρτησης της f από τα x και y. Για παράδειγμα

$$\frac{df}{dx} \;=\; \frac{d}{dx} f\left(x, y(x)\right) \;=\; \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\,\frac{dy}{dx} \tag{6.10}$$

ενώ αντίστοιχες εκφράσεις προκύπτουν και για τις παραγώγους υψηλότερης τάξης. Διαθέτοντας την αναλυτική έκφραση της $f(x, y(x))$, με διαδοχικές παραγωγίσεις το δεξιό μέλος της 6.8 εκφράζεται μόνο με τις μερικές παραγώγους της f ως προς x και y, αποτελεί ως εκ τούτου γνωστή μαθηματική έκφραση.

Εάν είναι δοσμένη ως αρχική συνθήκη η τιμή $y(x_0)$, τότε είναι άμεσος ο υπολογισμός της τιμής της συνάρτησης $f(x_0, y(x_0))$. Αλυσιδωτές παραγωγίσεις όπως η 6.10 παρέχουν αναλυτικές εκφράσεις για τις παραγώγους $f^{'}\left(x_0, y(x_0)\right)$, $f^{''}\left(x_0, y(x_0)\right)$, κ.ο.κ., οι οποίες, όταν αντικατασταθούν στην εξίσωση 6.8 δίνουν την τελική τιμή που προσεγγίζει την ακριβή λύση στο $x_0 + \Delta x$. Με αναδρομικό τρόπο και με αφετηρία την ήδη υπολογισθείσα λύση στο $x_0 + \Delta x$, με την ίδια μαθηματική έκφραση υπογίζεται η λύση στο επόμενο σημείο $x_0 + 2\Delta x$, κ.ο.κ.

Σχόλια για την ακρίβεια με την οποία η μέθοδος προσεγγίζει τη λύση ακολουθούν στην επόμενη εφαρμογή.

Εφαρμογή

Εφαρμόστε τη μέθοδο επίλυσης *σ.δ.ε.* με αναπτύγματα Taylor για την αριθμητική επίλυση της

$$\frac{dy}{dx} \;=\; x+y \tag{6.11}$$

Δώστε την τελική έκφραση υπολογισμού του $y(x_0+\Delta x)$, σύμφωνα με την εξίσωση 6.8 και σχολιάστε την αποκοπή όρων με την αναλυτική (ακριβή) λύση. Η αρχική συνθήκη είναι η

$$y(x_0) \;=\; y_0 \tag{6.12}$$

με $x_0 = 2$ και $y_0 = -2$.

Λύση:

Σύμφωνα με την 6.11, είναι

$$f(x,y) \;=\; x+y \tag{6.13}$$

Η διαφόριση της 6.13, σύμφωνα με την 6.10, δίνει διαδοχικά ότι

$$\begin{aligned} f^{'}(x,y) \;&=\; 1+1\cdot\frac{dy}{dx} \;=\; 1+x+y \\ f^{''}(x,y) \;&=\; 1+1\cdot\frac{dy}{dx} \;=\; 1+x+y \\ &\vdots \\ f^{(n)}(x,y) \;&=\; 1+1\cdot\frac{dy}{dx} \;=\; 1+x+y \end{aligned} \tag{6.14}$$

Με τα παραπάνω, το ανάπτυγμα 6.8 για το σημείο x_0 δίνει ότι

$$\begin{aligned} y(x_0+\Delta x) \;=\; y(x_0) \;&+\; \frac{\Delta x}{1!}[x_0+y(x_0)] \;+\; \frac{\Delta x^2}{2!}[1+x_0+y(x_0)] \\ &+\; \frac{\Delta x^3}{3!}[1+x_0+y(x_0)] \;+\; \ldots \end{aligned}$$

Στην τελευταία εξίσωση, προσθέτουμε και αφαιρούμε την ποσότητα $(x_0+\Delta x+1)$ στο δεξιό μέλος της και αυτή παίρνει τη μορφή

$$\begin{aligned} y(x_0+\Delta x) &= - x_0-\Delta x-1+[1+x_0+y(x_0)]+\frac{\Delta x}{1!}[1+x_0+y(x_0)] \\ &+ \frac{\Delta x^2}{2!}[1+x_0+y(x_0)]+\frac{\Delta x^3}{3!}[1+x_0+y(x_0)] \ + \ \ldots \\ &= - x_0-\Delta x-1+[1+x_0+y(x_0)]\cdot\left[1+\Delta x+\frac{\Delta x^2}{2!}+\frac{\Delta x^3}{3!}+\ldots\right] \end{aligned} \tag{6.15}$$

Στην εξίσωση 6.15 αναγνωρίζουμε, στην τελευταία αγκύλη, το κατά Taylor ανάπτυγμα του $e^{\Delta x}$. Η παρατήρηση αυτή είναι σημαντική ώστε να συσχετίσουμε την προσεγγιστική λύση της εξίσωσης 6.15 με την αναλυτική λύση που διαθέτει μια απλή *σ.δ.ε.* όπως αυτή της εξίσωσης 6.11. Η αναλυτική λύση προκύπτει εύκολα με ολοκλήρωση και είναι η

$$y(x) \ = \ -(x+1) \ + \ C\ e^x$$

όπου C η σταθερά της ολοκλήρωσης. Για τον προσδιορισμό της έκφρασης του C ικανοποιούμε την αρχική συνθήκη οπότε είναι

$$y_0 \ = \ -(x_0+1)+Ce^{x_0} \ \ \Rightarrow \ C \ = \ e^{-x_0}\,(1+x_0+y_0)$$

Για τις δεδομένες αρχικές τιμές είναι $C = 0.13533528$. Έτσι, η αναλυτική λύση που ικανοποιεί και την αρχική συνθήκη γράφεται τελικά στη μορφή

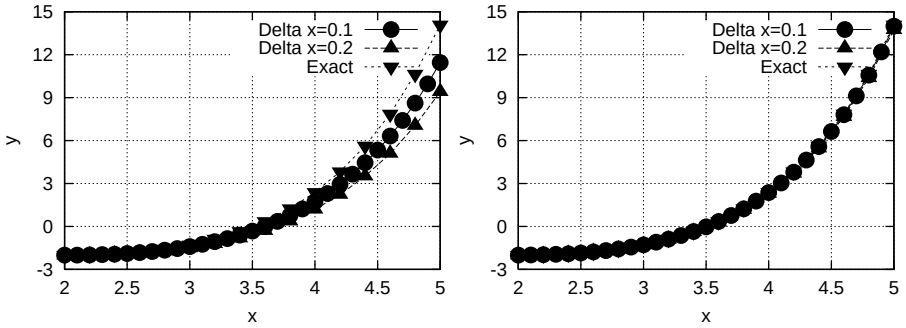
$$y(x) \ = \ -(x+1) \ + \ (1+x_0+y_0)\,e^{x-x_0} \tag{6.16}$$

Η εξίσωση 6.16 καθορίζει και την ακριβή έκφραση της λύσης στο $(x_0+\Delta x)$, ως

$$y(x_0+\Delta x) \ = \ -x_0-\Delta x-1 \ + \ (1+x_0+y(x_0))\,e^{\Delta x} \tag{6.17}$$

Στο σημείο αυτό μπορούμε να συγκρίνουμε την ακριβή λύση, εξ. 6.17, με την προσεγγιστική λύση, εξ. 6.15, όπως η τελευταία προέκυψε με χρήση αναπτυγμάτων κατά Taylor. Η έκφραση 6.15 προσεγγίζει την αναλυτική λύση 6.17, το δε σφάλμα αποκοπής (δηλ. η διαφορά της αριθμητικής από την ακριβή λύση, βλ. ορισμό στην εξ. 6.17), εξαρτάται αποκλειστικά και μόνο από το επίπεδο της αποκοπής (δηλαδή τους όρους που διατηρούμε) στο κατά Taylor ανάπτυγμα του e^x.

Πέραν των προηγουμένων σχολίων, τα αριθμητικά αποτελέσματα της εφαρμογής του σχήματος 6.15 στο πρόβλημα αυτό παρουσιάζονται στο Σχήμα 6.2. Σχεδιάζεται η λύση στο διάστημα $[2,5]$, υπολογισμένη με $\Delta x = 0.1$ και $\Delta x = 0.2$. Οι υπολογισμοί έγιναν με δύο διαφορετικές αποκοπές (αποκοπή όρων ίσης ή μεγαλύτερης δύναμης του Δx^2, την πρώτη φορά, και του Δx^3, στην τελευταία αγκύλη της 6.15). Είναι δε αισθητή η απόκλιση από την αναλυτική λύση στην πρώτη περίπτωση αποκοπής και για τη μεγαλύτερη τιμή του Δx που δοκιμάστηκε.



Σχήμα 6.1: Γραφική αναπαράσταση της αριθμητικής λύσης της *σ.δ.ε.* 6.11, σε σύγκριση με την αναλυτική λύση. Αριστερά παρουσιάζονται τα αποτελέσματα για αποκοπή όρων ίσης ή μεγαλύτερης δύναμης του Δx^2 στην τελευταία αγκύλη της 6.15. Δεξιά παρουσιάζονται τα αποτελέσματα για αποκοπή όρων ίσης ή μεγαλύτερης δύναμης του Δx^3. Σε κάθε περίπτωση, χρησιμοποιήθηκαν δύο βήματα $\Delta x = 0.1$ και $\Delta x = 0.2$.

Για την τελευταία θέση, $x = 5.0$, στην οποία η αναλυτική λύση είναι $y_{exact} = 14.085537$, η αποκοπή όρων ίσης ή μεγαλύτερης δύναμης του Δx^2 υπολόγισε λύσεις $y = 11.449402$ για $\Delta x = 0.1$ και $y = 9.407022$ για $\Delta x = 0.2$. Αντίθετα, η αποκοπή όρων ίσης ή μεγαλύτερης δύναμης του Δx^3 υπολόγισε σαφώς ακριβέστερες λύσεις, δηλαδή $y = 13.992557$ για $\Delta x = 0.1$ και $y = 13.742287$ για $\Delta x = 0.2$.

Σχόλια

Δυστυχώς, και πέραν κάποιων απλών περιπτώσεων όπως αυτής της εφαρμογής που μόλις αναλύθηκε, η διαφόριση της $f(x, y(x))$ παρουσιάζει αρκετή πολυπλοκότητα, ιδίως για τις παραγώγους μεγαλύτερης της πρώτης τάξης. Έτσι, ουσιαστικά, μεθόδους γενικής χρήσης για την αριθμητική επίλυση *σ.δ.ε.* με χρήση αναπτυγμάτων κατά Taylor, θα συναντήσουμε μόνο στη μορφή

$$y(x_0 + \Delta x) \;=\; y(x_0) \;+\; \Delta x f(x_0, y(x_0)) + O(\Delta x^2)$$

ή, ισοδύναμα, στη μορφή

$$y(x_{i+1}) \;=\; y(x_i) \;+\; \Delta x f(x_i, y(x_i)) + O(\Delta x^2) \tag{6.18}$$

Ο όρος $O(\Delta x^2)$ εκφράζει, κατά τα γνωστά, όρους ανάλογους του Δx^2 , ενώ σχήματα όπως η εξίσωση 6.18 μπορούν να χρησιμοποιηθούν μόνο όταν η ακρίβεια της προσέγγισης της λύσης μας είναι επαρκής. Πάντως, η εξέλιξη και η διάδοση λογισμικου διαχείρισης συμβολικών μαθηματικών εκφράσεων, μπορεί να θεωρηθεί ότι καλύπτει, και ίσως αναιρεί, το μειονέκτημα της τεχνικής αυτής που σχετίζεται με την πολυπλοκότητα της εύρεσης των παραγώγων υψηλότερης τάξης.

Η σχέση 6.18 αποτελεί τη βασική σχέση στην οποία στηρίζεται η μέθοδος Euler για την αριθμητική επίλυση *σ.δ.ε.*. Η μέθοδος Euler αναπτύσσεται στη συνέχεια.

6.6 Η Μέθοδος Euler

Η μέθοδος Euler στηρίζεται στη βηματική σχέση 6.18, το τελικό δηλαδή αποτέλεσμα του αναπτύγματος κατά Taylor που προηγούμενα παρουσιάσθηκε. Αν μάλιστα δεχθούμε ως αρχική συνθήκη αυτήν της εξίσωσης 6.12 και μια διαμέριση του πεδίου ορισμού της λύσης σε $N+1$ ισαπέχοντα σημεία $(x_0, x_1, \ldots, x_N)$, τότε η σχέση 6.18 γράφεται και στη μορφή

$$\begin{aligned} y_1 &= y(x_0) + \Delta x\ f(x_0, y(x_0)) \\ y_{i+1} &= y_i + \Delta x\ f(x_i, y_i) \qquad , \quad i \geq 1 \end{aligned} \tag{6.19}$$

με το $y(x_0) = y_0$ να έχει γνωστή δεδομένη τιμή.

Όπως αναφέρθηκε και προηγούμενα, η μέθοδος Euler (που είναι μέθοδος ενός βήματος, μονοβηματική) δεν βρίσκει πλέον συχνή εφαρμογή για λόγους ανεπάρκειας στην ακρίβεια της προσεγγιστικής λύσης που παράγει. Όμως θα συζητηθεί διότι αποτελεί σημαντικό 'εργαλείο' για την παρουσίαση και κατανόηση της ανάλυσης μετάδοσης του σφάλματος.

Η μέθοδος Euler για την επίλυση *σ.δ.ε.* συνοδεύεται από μια απλή γεωμετρική ερμηνεία όσον αφορά στον τρόπο που βηματικά προσεγγίζεται η λύση. Η γεωμετρική ερμηνεία γίνεται κατανοητή από την εφαρμογή των σχέσεων 6.19 στο Σχήμα 6.2. Αν η συνεχής καμπύλη γραμμή απεικονίζει την ακριβή λύση στο διάστημα $[x_0, x_1]$ και η ευθεία γραμμή ορίζεται ως η εφαπτόμενη στην προηγούμενη καμπύλη στο σημείο x_0, τότε η μέθοδος Euler προσεγγίζει τη λύση στο x_1 ως το σημείο (x_1, y_1) του διαγράμματος, αντί της ακριβούς λύσης $(x_1, y(x_1))$. Εφαρμόζοντας διαδοχικά τη μέθοδο Euler σε κάθε διάστημα $[x_0, x_1]$, $[x_1, x_2]$, $[x_2, x_3], \ldots$, η αριθμητική λύση ουσιαστικά ορίζεται από μια πολυγωνική γραμμή, της οποίας κάθε ευθύγραμμο τμήμα έχει την κλίση που καθορίζει η αριθμητική τιμή του f_i , $i = 0, 1, \ldots, N$, σύμφωνα με τη γραφή 6.3.

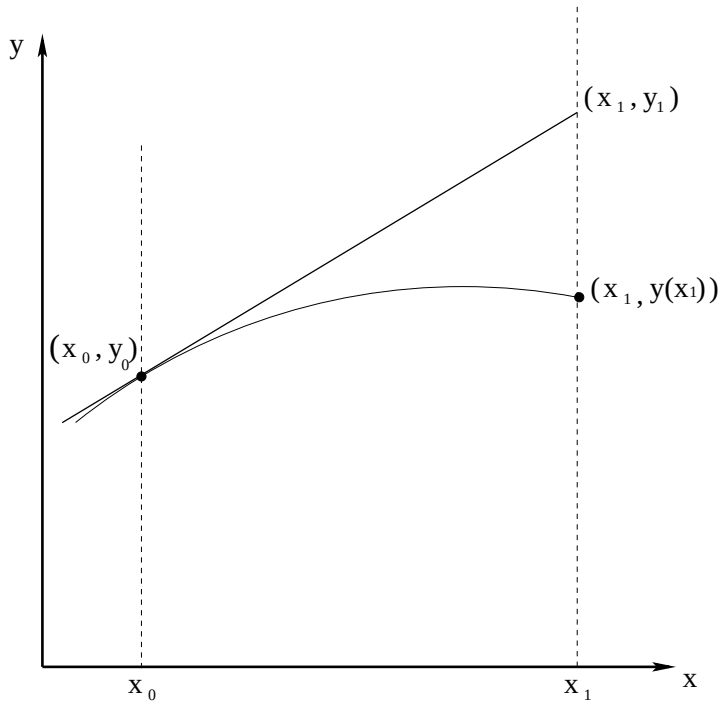
6.6.1 Μετάδοση Σφάλματος στη Μέθοδο Euler

Στην ενότητα αυτή θα δείξουμε ότι η μέθοδος Euler όταν εφαρμόζεται για την αριθμητική επίλυση μιας *σ.δ.ε.*, όπου το πεδίο της λύσης έχει διακριτοποιηθεί με $N+1$ ισαπέχοντα σημεία και βήμα Δx, συνοδεύεται από τοπικό σφάλμα αποκοπής της τάξης του Δx^2 και συνολικό σφάλμα αποκοπής της τάξης του Δx.

Αναφερόμαστε σε μια *σ.δ.ε.* της μορφής 6.3 που πρέπει να επιλυθεί αριθμητικά ως ένα πρόβλημα αρχικών τιμών, σύμφωνα με την αρχική συνθήκη 6.4. Το τοπικό σφάλμα αποκοπής στη θέση x_i έχει ήδη οριστεί στη σχέση 6.7 και, υπολογίζοντας με τη μέθοδο Euler την αριθμητική λύση στο επόμενο σημείο x_{i+1}, το τοπικό σφάλμα στη νέα αυτή θέση θα είναι

$$\epsilon_{i+1} = y_{i+1} - y(x_{i+1})$$

Έτσι το πρόσθετο σφάλμα αποκοπής $\Delta\epsilon_i = \epsilon_{i+1} - \epsilon_i$ που προκαλεί η αριθμητική επίλυση στο διάστημα (x_i, x_{i+1}) θα ισούται με



Σχήμα 6.2: Γεωμετρική ερμηνεία της μεθόδου Euler.

$$\Delta\epsilon_i \;=\; \epsilon_{i+1} - \epsilon_i \;=\; y_{i+1} - y_i - [y(x_{i+1}) - y(x_i)] \tag{6.20}$$

Αφού επιλύεται ένα πρόβλημα αρχικών τιμών και ισχύει η 6.4, το αρχικό σφάλμα αποκοπής θα είναι μηδενικό,

$$\epsilon_0 \;=\; y_0 - y(x_0) \;=\; 0 \tag{6.21}$$

Η εφαρμογή του σχήματος Euler, εξίσωση 6.19, βασίζεται (όπως ήδη έχει δειχθεί) σε ανάπτυγμα κατά Taylor. Γνωρίζουμε την ακριβή σχέση που διέπει το ανάπτυγμα Taylor

$$\begin{aligned} y(x_{i+1}) \;&=\; y(x_i + \Delta x) \;=\; y(x_i) \;+\; \frac{\Delta x}{1!} f\left(x_i, y(x_i)\right) \;+\; \frac{\Delta x^2}{2!} f'\left(x_i, y(x_i)\right) \\ &+\; \cdots + \frac{\Delta x^n}{n!} f^{(n-1)}\left(x_i, y(x_i)\right) \;+\; \frac{\Delta x^{n+1}}{(n+1)!} f^{(n)}\left(\xi, y(\xi)\right) \end{aligned} \tag{6.22}$$

όπου $\xi \in (x_i, x_{i+1})$. Η προσέγγιση οποιασδήποτε μεθόδου βασίζεται σε αναπτύγματα Taylor, αντικαθιστά το $y(x_{i+1})$ με το y_{i+1} και αποκόπτει τον τελευταίο όρο αντικαθιστώντας τον με έναν όρο ακριβείας $O(\Delta x^{n+1})$. Το τοπικό σφάλμα αποκοπής στην θέση x_i θα είναι επομένως ανάλογο του Δx^{n+1}, σύμφωνα με τον τελευταίο όρο της 6.22 και μπορούμε να το θεωρήσουμε φραγμένο σύμφωνα με τη σχέση

$$|\epsilon_i| \;=\; |y_i - y(x_i)| \;\leq\; \frac{\Delta x^{n+1}}{(n+1)!}\, M \tag{6.23}$$

όπου

$$M \geq |f^{(n)}\left(\xi^*, y(\xi^*)\right)|_{max} \tag{6.24}$$

για κάθε τιμή του ξ^* στο διάστημα (x_i, x_{i+1}).

Οι σχέσεις για τη μέθοδο Euler που έχουμε παρουσιάσει προηγούμενα διατυπώνονται ειδικά στις παρακάτω μορφές

$$y(x_{i+1}) = y(x_i) + \Delta x f\left(x_i, y(x_i)\right) + \frac{\Delta x^2}{2!} f'\left(\xi, y(\xi)\right) \ , \ x_i < \xi < x_{i+1} \tag{6.25}$$

$$y_{i+1} = y_i + \Delta x f\left(x_i, y_i\right) + O(\Delta x^2) \tag{6.26}$$

$$|\epsilon_i| = |y_i - y(x_i)| \leq \frac{\Delta x^2}{2} M \quad , \quad M \geq |f'(\xi^*, y(\xi^*)| \tag{6.27}$$

Τότε, η σχέση 6.20 γράφεται

$$\Delta\epsilon_i = \epsilon_{i+1} - \epsilon_i = \Delta x \left[f(x_i, y_i) - f(x_i, y(x_i))\right] + \frac{\Delta x^2}{2!} f'\left(\xi, y(\xi)\right) \tag{6.28}$$

Για την περαιτέρω ανάπτυξη της σχέσης 6.28 θα διατυπώσουμε την παρακάτω πρόταση:

Πρόταση 6.1 *Αν η συνάρτηση $f(x, y)$ και οι πρώτες μερικές παράγωγοί της είναι συνεχείς και φραγμένες στο διάστημα $\alpha \leq x \leq \beta$ και $-\infty \leq y \leq \infty$, τότε υπάρχει σταθερά K για την οποία*

$$|f(x, y^*) - f(x, y)| = \left|\frac{\partial f(x, \alpha)}{\partial y}\right| |y^* - y| \leq K |y^* - y| \tag{6.29}$$

όπου $y^* < \alpha < y$ για τα (x, y) και (x, y^*) στην παραπάνω περιοχή.

Με βάση τις σχέσεις 6.27 και 6.29, η 6.28 δίνει

$$|\Delta\epsilon_i| = |\epsilon_{i+1} - \epsilon_i| \leq \Delta x \ K \ |y_i - y(x_i)| + \frac{M}{2}\Delta x^2 \tag{6.30}$$

ή

$$|\Delta\epsilon_i| = |\epsilon_{i+1} - \epsilon_i| \leq \Delta x \ K \ |\epsilon_i| + \frac{M}{2}\Delta x^2 \tag{6.31}$$

Ομως

$$|\epsilon_{i+1}| \leq |\epsilon_{i+1} - \epsilon_i| + |\epsilon_i|$$

οπότε, χρησιμοποιώντας τη σχέση 6.31 ξαναγράφεται ως

$$|\epsilon_{i+1}| \leq (1 + \Delta x K)\,|\epsilon_i| + \frac{M}{2}\Delta x^2 \qquad , \quad i \geq 0 \tag{6.32}$$

Η σχέση 6.32 είναι ιδιαίτερα σημαντική αφού εκφράζει μαθηματικά τον τρόπο με τον οποίο μεταδίδεται το σφάλμα αποκοπής καθώς εξελίσσεται (αυξάνοντας τις τιμές του δείκτη i) ο αλγόριθμος του Euler. Να υπενθυμίσουμε πάλι ότι σε ένα πρόβλημα αρχικών τιμών, η εφαρμογή της σχέσης 6.32 ξεκινά με $\epsilon_0 = 0$. Για την περίπτωση αυτή, είναι εύκολο να δείξουμε επαγωγικά (η απόδειξη παραλείπεται) ότι η σχέση 6.32 οδηγεί τελικά στην ανισότητα

$$|\epsilon_i| \leq \frac{M\Delta x}{2K}\left[(1 + \Delta x K)^i - 1\right] \qquad , \quad i \geq 0 \tag{6.33}$$

Εκμεταλλευόμενη δε το κατά Taylor ανάπτυγμα μιας εκθετικής συνάρτησης, συγκεκριμένα της $e^{K\Delta x}$, το οποίο με αποκοπή των όρων που περιέχουν δυνάμεις του Δx μεγαλύτερες ή ίσες του τετραγώνου, δίνει ότι

$$(1 + K\Delta x) \;<\; e^{K\Delta x} \tag{6.34}$$

Η σχέση 6.33, μέσω της 6.34, γράφεται

$$|\epsilon_i| \leq \frac{M\Delta x}{2K}\left[e^{i\Delta x K} - 1\right] \;<\; \frac{M\Delta x}{2K}e^{i\Delta x K} \tag{6.35}$$

Ολοκληρώνοντας τα N βήματα υπολογισμού που υπαγορεύει η μέθοδος Euler και φθάνοντας στο τελευταίο $(x_{N-1} \rightarrow x_N)$, η παραπάνω σχέση δίνει

$$|\epsilon_N| < \frac{M\Delta x}{2K}e^{N\Delta x K}$$

όπου όμως $N\Delta x = x_N - x_0 = L$ (το συνολικό μήκος ολοκλήρωσης), άρα

$$|\epsilon_N| < \frac{M\Delta x}{2K}e^{LK} \tag{6.36}$$

Όταν το βήμα Δx τείνει στο μηδέν, το σφάλμα προσεγγίζει επίσης το μηδέν αφού

$$\lim_{\Delta x \to 0} |\epsilon_N| < \lim_{\Delta x \to 0}\left(\frac{M\Delta x}{2K}e^{LK}\right) \tag{6.37}$$

Κάθε αριθμητική διαδικασία για την οποία ισχύει

$$\lim_{\Delta x \to 0} |\epsilon_i| \;=\; 0 \tag{6.38}$$

ονομάζεται *συγκλίνουσα διαδικασία*. Δείξαμε λοιπόν ότι η μέθοδος Euler για την αριθμητική επίλυση *σ.δ.ε.* είναι μια συγκλίνουσα υπολογιστική διαδικασία με συνολικό σφάλμα αποκοπής

$$|\epsilon_i| \;=\; |y_i - y(x_i)| \;=\; O(\Delta x) \tag{6.39}$$

δηλαδή πρώτης τάξης, έστω και αν το τοπικό σφάλμα αποκοπής είναι τάξης $O(\Delta x^2)$.

Εφαρμογή

Ποσοτικοποιείστε το ρόλο των σταθερών M και K στον υπολογισμό του συνολικού σφάλματος αποκοπής στη *σ.δ.ε.* 6.11, με αρχική συνθήκη αυτής της εξίσωσης 6.12 και η οποία δέχεται την αναλυτική λύση 6.16. Για λόγους απλούστευσης των αναμενόμενων αποτελεσμάτων, επιβάλουμε διαφορετική αρχική συνθήκη, που είναι ότι για $x_0 = 0$ ισχύει

$$y\left(x_0 = 0\right) \;=\; y_0 \;=\; 0 \tag{6.40}$$

Η λύση θα αφορά το διάστημα $x \in [0, 1]$ με $N = 10$ και $\Delta x = 0.1$. Διατυπώστε συγκεκριμένα τη μορφή που παίρνει η εξίσωση 6.33

Λύση:

Με τις παραπάνω οριακές συνθήκες, η αναλυτική λύση είναι η

$$y(x) \;=\; e^x - x - 1 \tag{6.41}$$

Στη σχέση 6.27, ορίσαμε το M ως ένα άνω φράγμα της ποσότητας $|f^{'}(x, y)|$ στο διάστημα $[x_0, x_i]$ και επειδή

$$|f^{'}(x, y)|_{max} \;=\; |1 + x + y|_{max} \;=\; |e^x|_{max}$$

άρα επιλέγουμε να είναι

$$M \;=\; |e^x|_{max} \quad , \quad x \in [0, 1] \tag{6.42}$$

Αντίστοιχα, το K είναι ένα άνω όριο της μερικής παραγώγου του $f(x, y)$ ως προς το y και επειδή

$$\frac{\partial f}{\partial x} \;=\; \frac{\partial}{\partial x}\left(x + y\right) \;=\; 1$$

εύκολα βρίσκουμε ότι

$$K \;=\; 1 \tag{6.43}$$

Για λύση λοιπόν μέσω της μεθόδου του Euler στο διάστημα $[0, 1]$ με μηδενική αρχική τιμή και βήμα $\Delta x = 0.1$, η σχέση 6.33 γράφεται

$$|\epsilon_i| \leq \frac{0.1\; e^{x_i}}{2}\left[(1 + 0.1)^i - 1\right] \tag{6.44}$$

6.7 Μέθοδοι Runge–Kutta

Σε προηγούμενη ενότητα σχολιάστηκαν επαρκώς οι αδυναμίες κάθε τεχνικής για την αριθμητική επίλυση *σ.δ.ε.* η οποία απαιτεί το απευθείας ανάπτυγμα κατά Taylor της συνάρτησης $f(x, y(x))$, ιδίως όταν η τελευταία εμφανίζει πολύπλοκη μορφή. Ευνόητη, συνεπώς, θα ήταν η αναζήτηση μεθόδων αριθμητικής επίλυσης οι οποίες θα χρειάζονταν μόνο υπολογισμούς της πρώτης παραγώγου της συνάρτησης f, αλλά θα έδιναν αποτελέσματα που θα αντιστοιχούσαν σε ανάπτυγμα Taylor υψηλότερης τάξης. Τέτοιες αριθμητικές μέθοδοι είναι οι μέθοδοι Runge–Kutta. Πρόκειται για μεθόδους απλού βήματος, οι οποίες συνήθως εμφανίζονται ως δεύτερης, τρίτης ή τέταρτης τάξης σχήματα. Οι προαναφερθείσες τάξεις ισοδυναμούν με αναπτύγματα Taylor στα οποία έχουν διατηρηθεί οι όροι που περιέχουν τα Δx^2, Δx^3 και Δx^4, αντίστοιχα. Πρακτικά δε, απαιτούν τον υπολογισμό της πρώτης παραγώγου σε πολλαπλά (δύο, τρία και τέσσερα, αντίστοιχα) σημεία του διαστήματος $x_i \leq x \leq x_{i+1}$. Σε όλη αυτή την ενότητα, θα υποθέσουμε ότι η συνάρτηση f είναι επαρκώς λεία και θα συμβολίσουμε με $y(x_i)$ την πραγματική λύση στο σημείο x_i και με y_i την αριθμητική λύση που προκύπτει από τη μέθοδο Runge–Kutta.

Για εκπαιδευτικούς λόγους, θα παράγουμε τη μέθοδο Runge–Kutta δεύτερης τάξης (δηλαδή με τοπικό σφάλμα αποκοπής τρίτης τάξης), έστω και αν η πιο συνηθισμένη μέθοδος Runge-Kutta που χρησιμοποιείται για την επίλυση αριθμητικών προβλημάτων είναι η τέταρτης τάξης, για ευνόητους λόγους υψηλότερης ακριβείας.

Η βασική ιδέα που συνοδεύει τη μέθοδο Runge–Kutta δεύτερης τάξης είναι να αντικατασταθεί το ανάπτυγμα

$$\begin{aligned} y(x_{i+1}) &= y(x_i + \Delta x) = y(x_i) + \Delta x f(x_i, y(x_i)) \\ &+ \frac{\Delta x^2}{2!}\left[\frac{\partial f(x_i, y(x_i))}{\partial x} + \frac{\partial f(x_i, y(x_i))}{\partial y} f(x_i, y(x_i))\right] + O(\Delta x^3) \end{aligned} \quad (6.45)$$

με μια σχέση της μορφής

$$y(x_{i+1}) = y(x_i + \Delta x) = y(x_i) + w_1 k_1 + w_2 k_2 \quad (6.46)$$

όπου

$$\begin{aligned} k_1 &= \Delta x f(x_i, y(x_i)) \\ k_2 &= \Delta x f(x_i + a\Delta x, y(x_i) + b k_1) \end{aligned} \quad (6.47)$$

ενώ τα w_1, w_2, a, b πρέπει να επιλεγούν με τρόπο ώστε να διατηρούν την επιθυμητή ακρίβεια της προσέγγισης. Η έκφραση 6.47 για το k_2 δίνει το παρακάτω ανάπτυγμα

$$k_2 = \Delta x\left[f(x_i, y(x_i)) + a\Delta x \frac{\partial f}{\partial x}(x_i, y(x_i)) + b k_1 \frac{\partial f}{\partial y}(x_i, y(x_i)) + O(\Delta x^2)\right] \quad (6.48)$$

οπότε με αντικατάσταση της 6.47 για το k_1 και της 6.48 για το k_2 στην εξίσωση 6.46 παίρνουμε

$$\begin{aligned} y(x_{i+1}) &= y(x_i + \Delta x) = y(x_i) + w_1 \Delta x f(x_i, y(x_i)) \\ &+ w_2 \Delta x \left[f(x_i, y(x_i)) + a\Delta x \frac{\partial f}{\partial x}(x_i, y(x_i)) + bk_1 \frac{\partial f}{\partial y}(x_i, y(x_i)) \right] \\ &+ O(\Delta x^3) \end{aligned} \tag{6.49}$$

Αρκεί πλέον να εξισώσουμε τα δεξιά μέλη των εξισώσεων 6.45 και 6.49, που ούτως ή άλλως έχουν και τα δύο ακρίβεια τρίτης τάξης ως προς το Δx, οπότε προκύπτουν οι τρεις παρακάτω σχέσεις (αφορούν αντίστοιχα τους συντελεστές των f, $\partial f/\partial x$ και $\partial f/\partial y$):

$$\begin{aligned} w_1 + w_2 &= 1 \\ w_2 a &= \frac{1}{2} \\ w_2 b &= \frac{1}{2} \end{aligned} \tag{6.50}$$

Το σύστημα 6.50 διαθέτει τρεις εξισώσεις για τους τέσσερις αγνώστους w_1, w_2, a και b, οπότε επιτρέπει να οριστεί ελεύθερα ο ένας από αυτούς. Μια λύση του συστήματος είναι η

$$\begin{aligned} w_1 = w_2 &= \frac{1}{2} \\ a = b &= 1 \end{aligned} \tag{6.51}$$

δίνοντας την τελική μορφή της μεθόδου Runge–Kutta δεύτερης τάξης, ως

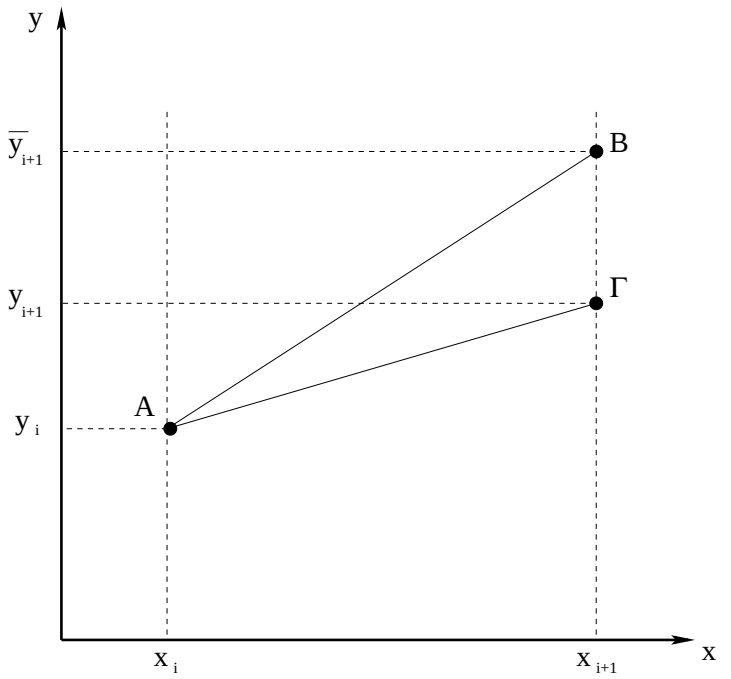
$$\begin{aligned} y_{i+1} &= y_i + \frac{1}{2}k_1 + \frac{1}{2}k_2 \\ k_1 &= \Delta x \; f(x_i, y_i) \\ k_2 &= \Delta x \; f(x_{i+1}, y_i + k_1) \end{aligned} \tag{6.52}$$

Το σχήμα 6.52 έχει τοπικό σφάλμα αποκοπής της τάξης του Δx^3, ίδιο με αυτό που δίνει το ανάπτυγμα Taylor της σχέσης 6.45. Ασφαλώς η σχέση 6.52 γράφεται και στην ενιαία μορφή

$$y_{i+1} = y_i + \frac{\Delta x}{2} \left[f(x_i, y_i) + f(x_{i+1}, y_i + \Delta x f(x_i, y_i)) \right] \tag{6.53}$$

ή και ακόμα ως

$$y_{i+1} = y_i + \frac{\Delta x}{2} \left[f(x_i, y_i) + f(x_{i+1}, \overline{y}_{i+1}) \right] \tag{6.54}$$



Σχήμα 6.3: Γεωμετρική ερμηνεία της μεθόδου Runge–Kutta ως μεθόδου Euler.

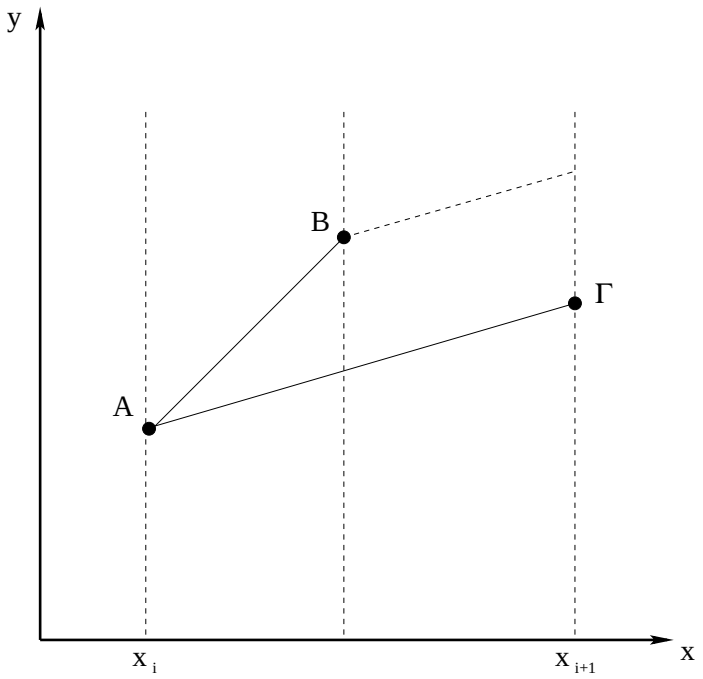
με την ενδιάμεση τιμή

$$\overline{y}_{i+1} \;=\; y_i + \Delta x f(x_i, y_i) \tag{6.55}$$

Χρησιμοποιώντας την τελευταία γραφή των σχέσεων 6.54 και 6.55 μπορούμε να αντιληφθούμε τη μέθοδο Runge–Kutta δεύτερης τάξης ως μια διπλή εφαρμογή της μεθόδου Euler: προχωρώντας από την ήδη γνωστή θέση x_i στη νέα θέση x_{i+1}, υπολογίζεται πρώτα η ενδιάμεση τιμή $\overline{y}_{i+1}$ με τη σχέση 6.55 και ακολουθεί η τελική λύση y_{i+1} από τη σχέση 6.54. Με βάση αυτόν τον αλγόριθμο, μπορεί επιπλέον η μέθοδος Runge–Kutta να γίνει κατανοητή ως μια *μέθοδος πρόβλεψησ–διόρθωσης*, με την έννοια ότι η εξίσωση 6.55 προβλέπει την τιμή του $\overline{y}_{i+1}$ (μια 'όχι αποδεκτής ακρίβειασ' προσέγγιση του $y(x_{i+1})$ και στη συνέχεια η τιμή αυτή διορθώνεται–βελτιώνεται με τη σχέση 6.54. Το Σχήμα 6.3 παρουσιάζει τη γεωμετρική ερμηνεία του διπλού βήματος υπολογισμού από το x_i στο x_{i+1}. Στο πρώτο βήμα (το θεωρούμενο ως πρόβλεψη, σχέση 6.55), υπολογίζεται με αφετηρία το γνωστό σημείο (x_i, y_i) (σημείο A στο σχήμα 6.3) και κλίση ίση με τη γνωστή τιμή $f(x_i, y_i)$, το σημείο $(x_{i+1}, \overline{y}_{i+1})$ (σημείο B στο σχήμα 6.3).

Στη συνέχεια, για τη διόρθωση μέσω της σχέσης 6.54 και με αφετηρία πάλι το σημείο A, εντοπίζεται το σημείο (x_{i+1}, y_{i+1}) ή Γ, φέροντας ευθεία με κλίση το ημιάθροισμα των τιμών της συνάρτησης $f(x, y)$ όπως αυτές υπολογίζονται στα σημεία A και B.

Ο βαθμός ελευθερίας που εμφανίστηκε κατά τον υπολογισμό των 4 αγνώστων από το σύστημα 6.50 με μόνο τρεις εξισώσεις, επιτρέπει τη δημιουργία εναλλακτικών της προηγούμενης γραφών της μεθόδου Runge–Kutta δεύτερης τάξης. Αν, αντί της 6.51,



Σχήμα 6.4: Γεωμετρική ερμηνεία των σχέσεων 6.57 και 6.58.

ως λύση του συστήματος επιλεγεί η :

$$w_1 \;=\; 0 \quad , \quad w_2 \;=\; 1 \quad , \quad a \;=\; b \;=\; \frac{1}{2} \tag{6.56}$$

τότε οι εξισώσεις 6.54 και 6.55 γράφονται

$$y_{i+1} \;=\; y_i + \Delta x f\left(x_i + \frac{\Delta x}{2}, \overline{y}_{i+\frac{1}{2}}\right) \tag{6.57}$$

με την ενδιάμεση τιμή $\overline{y}_{i+\frac{1}{2}}$ να ορίζεται τώρα ως

$$\overline{y}_{i+\frac{1}{2}} \;=\; y_i + \frac{\Delta x}{2} f(x_i, y_i) \tag{6.58}$$

Η γεωμετρική ερμηνεία των σχέσεων 6.57 και 6.58 παρουσιάζεται στο Σχήμα 6.4. Από το σημείο A και με κλίση $f(x_i, y_i)$ υπολογίζεται πρώτα το σημείο B, στη θέση $x_{i+\frac{1}{2}} = x_i + \frac{\Delta x}{2}$ και στη συνέχεια με κλίση την τιμή της f στο $(x_i + \frac{\Delta x}{2}, \overline{y}_{i+\frac{1}{2}})$ και αφετηρία πάλι το A προσδιορίζεται το τελικό σημείο–λύση Γ.

Οι μέθοδοι Runge–Kutta υψηλότερης τάξης παρουσιάζονται στη συνέχεια χωρίς τις αποδείξεις τους, οι οποίες εξάλλου είναι ανάλογες αυτής που προηγήθηκε.

6.7.1 Μέθοδος Runge–Kutta Τρίτης Τάξης

Με αφετηρία μια σχέση της μορφής

$$y_{i+1} \;=\; y_i + w_1 k_1 + w_2 k_2 + w_3 k_3 \tag{6.59}$$

και

$$\begin{aligned} k_1 &= \Delta x\ f(x_i, y_i) \\ k_2 &= \Delta x\ f(x_i + a\Delta x, y_i + bk_1) \\ k_3 &= \Delta x\ f(x_i + c\Delta x, y_i + dk_2 + (c-d)k_1) \end{aligned} \tag{6.60}$$

μπορεί να δειχθεί ότι μια μορφή της μεθόδου θα ήταν η

$$y_{i+1} = y_i + \frac{1}{6}(k_1 + 4k_2 + k_3) \tag{6.61}$$

με

$$\begin{aligned} k_1 &= \Delta x\ f(x_i, y_i) \\ k_2 &= \Delta x\ f(x_i + \frac{\Delta x}{2}, y_i + \frac{k_1}{2}) \\ k_3 &= \Delta x\ f(x_i + \Delta x, y_i + 2k_2 - k_1) \end{aligned} \tag{6.62}$$

6.7.2 Μέθοδος Runge–Kutta Τέταρτης Τάξης

Ένα τυπικό σχήμα με τοπικό σφάλμα αποκοπής $O(\Delta x^5)$ είναι το

$$y_{i+1} = y_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \tag{6.63}$$

όπου

$$\begin{aligned} k_1 &= \Delta x\ f(x_i, y_i) \\ k_2 &= \Delta x\ f(x_i + \frac{\Delta x}{2}, y_i + \frac{k_1}{2}) \\ k_3 &= \Delta x\ f(x_i + \frac{\Delta x}{2}, y_i + \frac{k_2}{2}) \\ k_4 &= \Delta x\ f(x_{i+1}, y_i + k_3) \end{aligned} \tag{6.64}$$

Εναλλακτικά, χρησιμοποιείται και το σχήμα

$$y_{i+1} = y_i + \frac{1}{8}(k_1 + 3k_2 + 3k_3 + k_4) \tag{6.65}$$

με

$$\begin{aligned} k_1 &= \Delta x\ f(x_i, y_i) \\ k_2 &= \Delta x\ f(x_i + \frac{\Delta x}{3}, y_i + \frac{k_1}{3}) \\ k_3 &= \Delta x\ f(x_i + \frac{2\Delta x}{3}, y_i - \frac{k_1}{3} + k_2) \\ k_4 &= \Delta x\ f(x_{i+1}, y_i + k_1 - k_2 + k_3) \end{aligned} \tag{6.66}$$

Ίσως όμως η πιο συχνά χρησιμοποιούμενη μορφή της μεθόδου Runge–Kutta τέταρτης τάξης είναι αυτή που φέρεται και ως μέθοδος Gill και η οποία διατυπώνεται ως

$$y_{i+1} = y_i + \frac{1}{6}\left(k_1 + 2\left(1 - \frac{1}{\sqrt{2}}\right)k_2 + 2\left(1 + \frac{1}{\sqrt{2}}\right)k_3 + k_4\right) \tag{6.67}$$

με

$$\begin{aligned} k_1 &= \Delta x\ f(x_i, y_i) \\ k_2 &= \Delta x\ f(x_i + \frac{\Delta x}{2}, y_i + \frac{k_1}{2}) \\ k_3 &= \Delta x\ f\left(x_i + \frac{\Delta x}{2}, y_i + \left(-\frac{1}{2} + \frac{1}{\sqrt{2}}\right)k_1 + \left(1 - \frac{1}{\sqrt{2}}\right)k_2\right) \\ k_4 &= \Delta x\ f\left(x_{i+1}, y_i + -\frac{1}{\sqrt{2}}k_2 + \left(1 + \frac{1}{\sqrt{2}}\right)k_3\right) \end{aligned} \tag{6.68}$$

Εφαρμογή

Εφαρμόστε τις μεθόδους επίλυσης Runge–Kutta, δεύτερης ως και τέταρτης τάξης, για την αριθμητική επίλυση της εξίσωσης ακτινοβολίας των Stefan–Boltzmann

$$\frac{dT}{dt} = -\alpha\left(T^4 - T_a^4\right) \tag{6.69}$$

Η εξίσωση αυτή διέπει το χρονικό ρυθμό μεταβολής (t είναι ο χρόνος σε seconds) της εσωτερικής θερμοκρασίας (T, σε Kelvin) μιας συγκεντρωμένης μάζας m (σε kg), εξωτερικής επιφάνειας A (σε m^2), αδιάστατου συντελεστή εκπομπής σ και ευρισκόμενης σε περιβάλλον με σταθερή θερμοκρασία T_a (σε Kelvin). Αν C είναι η ειδική θερμοχωρητικότητα του υλικού της μάζας (σε $J/kg/K$) και ϵ ($\epsilon = 5.67{\cdot}10^{-8}\ J/m^2/K^4/sec$) είναι η σταθερά των Stefan–Boltzmann, τότε ο συντελεστής α της εξίσωσης 6.69 θα είναι

$$\alpha = \frac{A\ \epsilon\ \sigma}{m\ C} \tag{6.70}$$

Κατά την επίλυση, θεωρείστε ως αρχική θερμοκρασία της μάζας, κατά τη χρονική στιγμή $t = 0$, την $T_0 = 2500K$ και ότι $T_a = 250K$. Δίνεται ότι $\alpha = 2.0{\cdot}10^{-12} sec^{-1} K^{-3}$. Επιλέξτε χρονικό βήμα $\Delta t = 1\ sec$ και ολοκληρώστε μέχρι $t = 10\ sec$. Υπολογίστε την αναλυτική λύση και συγκρίνετε τα αριθμητικά σας αποτελέσματα με αυτή.

Λύση:

Η αναλυτική λύση της 6.69 δίνεται από τη σχέση

$$tan^{-1}\left(\frac{T}{T_a}\right) \; - \; tan^{-1}\left(\frac{T_0}{T_a}\right) \; + \; \frac{1}{2} ln \frac{(T_0 - T_a)(T + T_a)}{(T - T_a)(T_0 + T_a)} \; = \; 2\alpha T_a^3 t \qquad (6.71)$$

η οποία, όταν επιλυθεί με κάποιον από τους τρόπους που έχετε διδαχθεί στο αντίστοιχο κεφάλαιο του βιβλίου αυτού, δίνει την παρακάτω χρονική κατανομή της θερμοκρασίας του σώματος

t (sec)	T (K)	t (sec)	T (K)
0.0	2500.00000000	6.0	2154.47079576
1.0	2426.43487359	7.0	2113.03386111
2.0	2360.82998846	8.0	2074.61189788
3.0	2301.79075068	9.0	2038.84084646
4.0	2248.24731405	10.0	2005.41636581
5.0	2199.36266993		

Η χρήση των μεθόδων Runge–Kutta δεύτερης (σχέση 6.52), τρίτης (σχέσεις 6.61, 6.62) και τέταρτης τάξης (σχέσεις 6.65, 6.66) έγινε προγραμματίζοντας κώδικες σε Fortran 77. Στη συνέχεια παρατίθεται ένας από τους τρεις αυτούς κώδικες, αυτός που αντιστοιχεί στην μέθοδο τέταρτης τάξης. Με πολύ μικρές επεμβάσεις μπορούν να παραχθούν από αυτόν και οι άλλοι δύο κώδικες. Παρατηρείστε τις τέσσερις κλήσεις στη συνάρτηση $f(x, y)$ (εδώ είναι μόνο $f(y)$, όπου x ο χρόνος και y η θερμοκρασία της μάζας), ανά χρονικό βήμα, γεγονός που καταγράφεται στα μειονεκτήματα της μεθόδου Runge–Kutta τέταρτης τάξης, παρά την αυξημένη ακρίβειά της.

```
      program stefan_runge
      implicit double precision (a-h,o-z)
      fun(temp)=-2.d-12*(temp**4-250.d0**4)
c
      time=0.d0
      deltat=1.0
      temp=2500.d0
      write(*,'(2x,f5.1,3x,f16.8)')time,temp
c
      do i=1,10
      ak1 = deltat*fun(temp)
      ak2 = deltat*fun(temp+ak1/3.d0)
      ak3 = deltat*fun(temp-ak1/3.d0+ak2)
      ak4 = deltat*fun(temp+ak1-ak2+ak3)
      temp=temp+(ak1+3.*ak2+3.*ak3+ak4)/8.d0
      time=time+deltat
```

```
      write(*,'(2x,f5.1,3x,f16.8)')time,temp
      enddo
c
      end
```

Ο πίνακας που ακολουθεί περιλαμβάνει τα αποτελέσματα και με τις τρείς μεθόδους, σε σύγκριση με την αναλυτική λύση:

t (sec)	T(αναλυτική)	T(δεύτερης)	T(τρίτης)	T(τέταρτης)
0.0	2500.00000000	2500.00000000	2500.00000000	2500.00000000
1.0	2426.43487359	2426.54103036	2426.43321644	2426.43483927
2.0	2360.82998846	2361.00512406	2360.82738459	2360.82993707
3.0	2301.79075068	2302.01064691	2301.78761715	2301.79069129
4.0	2248.24731405	2248.49583792	2248.24390225	2248.24725157
5.0	2199.36266993	2199.62888327	2199.35913414	2199.36260705
6.0	2154.47079576	2154.74718220	2154.46723165	2154.47073396
7.0	2113.03386111	2113.31520610	2113.03032806	2113.03380117
8.0	2074.61189788	2074.89456327	2074.60843236	2074.61184019
9.0	2038.84084646	2039.12229016	2038.83747027	2038.84079119
10.0	2005.41636581	2005.69481760	2005.41309127	2005.41631298

ενώ, για μεγαλύτερη εποπτεία, πινακοποιούνται στη συνέχεια και τα σφάλματα (ως προσημασμένες διαφορές της αναλυτικής λύσης από κάθε αριθμητική):

t (sec)	T(δεύτερης) -T(αναλυτική)	T(τρίτης) -T(αναλυτική)	T(τέταρτης) -T(αναλυτική)
0.0	0.00000000	0.00000000	0.00000000
1.0	0.10615677	-0.00165715	-0.00003432
2.0	0.17513560	-0.00260387	-0.00005139
3.0	0.21989623	-0.00313353	-0.00005939
4.0	0.24852387	-0.00341180	-0.00006248
5.0	0.26621334	-0.00353579	-0.00006288
6.0	0.27638644	-0.00356411	-0.00006180
7.0	0.28134499	-0.00353305	-0.00005994
8.0	0.28266539	-0.00346552	-0.00005769
9.0	0.28144370	-0.00337619	-0.00005527
10.0	0.27845179	-0.00327454	-0.00005283

Εφαρμογή

Συμπαγής σφαίρα ακτίνας R από υλικό πυκνότητας ϱ_S βυθίζεται σε μεγάλη δεξαμενή υγρού πυκνότητας $\varrho_F = \varrho_S/4$. Η βύθιση αρχίζει τη στιγμή $t = 0$ με μηδενική αρχική ταχύτητα και τη σφαίρα να βρίσκεται με το κέντρο της σε βάθος $y = R$, όπου το βάθος $y > 0$ αυξάνει από την ελεύθερη επιφάνεια ($y = 0$) του υγρού προς τον πυθμένα. Η εξίσωση κίνησης κατά τη βύθιση της σφαίρας γράφεται

$$m\frac{d^2y}{dt^2} \;=\; mg - Vg\varrho_F - \pi R^2 C_D \varrho_F v^2(t) \tag{6.72}$$

όπου m η μάζα της σφαίρας, V ο όγκος της, g η επιτάχυνση της βαρύτητας και C_D ο συντελεστής αντίστασης της σφαίρας ο οποίος θεωρείται σταθερός και ανεξάρτητος της στιγμιαίας ταχύτητάς της $v(t)$.

1. Σχολιάστε σύντομα τους όρους στο δεύτερο μέλος της παραπάνω εξίσωσης και γράψτε την στη μορφή που προτείνετε ώστε να επιλυθεί αριθμητικά ως προς το $v(t)$.

2. Η συνήθης διαφορική εξίσωση που γράψατε πρόκειται να επιλυθεί με μέθοδο Runge–Kutta δεύτερης τάξης, ώστε να υπολογιστούν οι τιμές της ταχύτητας $v(t_i)$, όπου $t_i = i\Delta t$, $i = 1, 2, 3, \ldots$ με Δt το χρονικό βήμα. Εκτελέστε τις πράξεις και υπολογείστε την τιμή της ταχύτητας $v(t = 1sec)$ με χρονικό βήμα $\Delta t = 0.25sec$. Δίνονται: $R = 5cm$, $C_D = 0.01$, $g = 9.81m/sec^2$.

Λύση:

Οι όροι στο δεξί μέλος της εξίσωσης 6.72 είναι, κατά σειρά, το βάρος της συμπαγούς σφαίρας ($mg \;=\; V\varrho_S g$), η άνωση που αυτή δέχεται και τέλος η αντίσταση που δέχεται κατά τη βύθισή της στο υγρό πυκνότητας ϱ_F. Και οι τρείς αυτές δυνάμεις είναι κατακόρυφες και η συνισταμένη τους καθορίζει την επιτάχυνση d^2y/dt^2 της βυθιζόμενης σφαίρας. Είναι ευνόητο ότι με θετικό πρόσημο συμβολίζεται κάθε δύναμη με φορά προς τα κάτω.

Με σκοπό την αριθμητική της επίλυση, η εξίσωση 6.72 διατυπώνεται με άγνωστη την ταχύτητα βύθισης της σφαίρας $v(t) \;=\; dy(t)/dt$ και, εκτελώντας τις ακόλουθες πράξεις

$$\begin{aligned} m\frac{dv}{dt} &= mg - Vg\varrho_F - \pi R^2 C_D \varrho_F v^2 \;\Rightarrow \\ \frac{dv}{dt} &= g - \frac{g\varrho_F}{\varrho_S} - \frac{R^2 C_D \varrho_F v^2}{\frac{4}{3}R^3\varrho_S} \end{aligned}$$

προκύπτει η τελική μορφή της $\sigma.\delta.\varepsilon.$ που θα επιλυθεί αριθμητικά

$$\frac{dv(t)}{dt} \;=\; g - \frac{\varrho_F}{\varrho_S}\left[g \;+\; \frac{3C_D}{4R}v^2(t)\right] \tag{6.73}$$

Ορίζοντας τέλος ως

$$f(t,v(t)) \;=\; g - \frac{\varrho_F}{\varrho_S}\left[g \;+\; \frac{3C_D}{4R}v^2(t)\right] \tag{6.74}$$

η $\sigma.\delta.\varepsilon.$ γράφεται, στο πρότυπο της βασικής γραφής 6.3, ως

$$\frac{dv(t)}{dt} \;=\; f(t,v(t)) \tag{6.75}$$

Με αριθμητική αντικατάσταση στο σύστημα μονάδων SI, η 6.75 δίνει

$$\frac{dv(t)}{dt} \;=\; 7.3575 \;-\; 0.0375\; v^2(t) \tag{6.76}$$

που επιλύεται με αρχική συνθήκη την $v(t_0 = 0) = 0$, αφού η βύθισή της ξεκινά με μηδενική ταχύτητα.

Η αριθμητική επίλυση με τη μέθοδο Runge–Kutta δεύτερης τάξης υλοποιείται εφαρμόζοντας τον αλγόριθμο της σχέσης 6.52, ο οποίος επαναλαμβάνεται στη συνέχεια με προσαρμογή των μεταβλητών του στα σύμβολα που χρησιμοποιούνται στην παρούσα εφαρμογή

$$\begin{aligned} v_{i+1} &= v_i + \frac{1}{2}k_1 + \frac{1}{2}k_2 \\ k_1 &= \Delta t\; f(t_i, v_i) \\ k_2 &= \Delta t\; f(t_{i+1}, v_i + k_1) \end{aligned} \tag{6.77}$$

όπου $t_i = t_0 + \Delta t$, με $\Delta t = 0.25$. Ακολουθούν τα ενδιάμεσα αριθμητικά αποτελέσματα από τα τέσσερα χρονικά βήματα που πρέπει να εκτελεσθούν ώστε να υπολογισθεί η ζητούμενη τιμή της ταχύτητας της σφαίρας κατά τη χρονική στιγμή $t = 1sec$, με το δεδομένο χρονικό βήμα:

$$\begin{aligned} t_1 &= 0.25 \;\; sec \\ k_1 &= 1.839375019073486 \;\; m/sec \\ k_2 &= 1.807656575993188 \;\; m/sec \\ v(t_1) &= 1.823515797533337 \;\; m/sec \end{aligned}$$

$$\begin{aligned} t_2 &= 0.50 \;\; sec \\ k_1 &= 1.808201175361121 \;\; m/sec \\ k_2 &= 1.715724687554973 \;\; m/sec \\ v(t_2) &= 3.585478728991384 \;\; m/sec \end{aligned}$$

$$\begin{aligned} t_3 &= 0.75 \ sec \\ k_1 &= 1.718853223196415 \ m/sec \\ k_2 &= 1.575600594913904 \ m/sec \\ v(t_3) &= 5.232705638046544 \ m/sec \end{aligned}$$

$$\begin{aligned} t_4 &= 1.00 \ sec \\ k_1 &= 1.582676181112778 \ m/sec \\ k_2 &= 1.403911601698531 \ m/sec \\ v(t_4) &= 6.725999529452197 \ m/sec \end{aligned}$$

Άρα η αριθμητικά υπολογισμένη τιμή της ταχύτητας της σφαίρας κατά τη χρονική στιγμή $t = 1sec$, με μέθοδο Runge–Kutta δεύτερης τάξης και χρονικό βήμα $\Delta t = 0.25sec$, είναι $v = 6.725999529452197m/sec$.

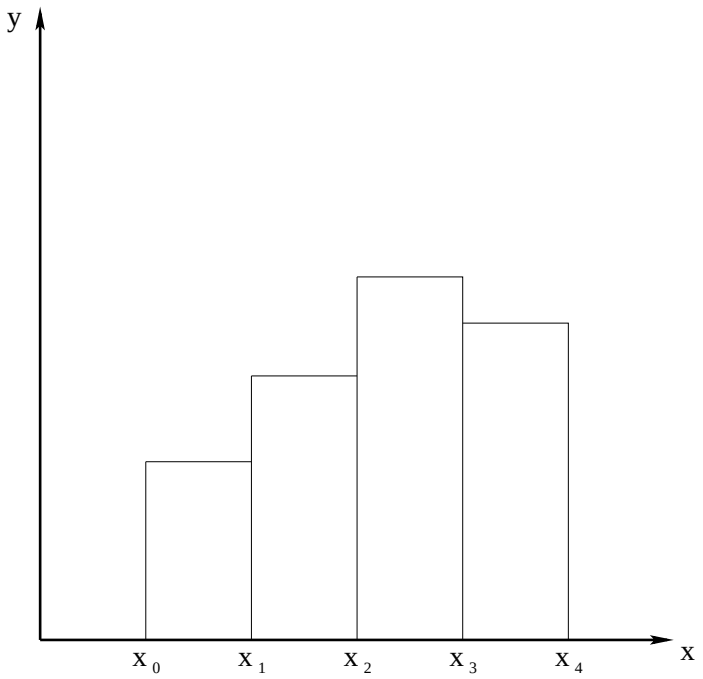
Στην ίδια εφαρμογή, παρουσιάζονται τα αποτελέσματα που προέκυψαν με χρήση κώδικα υπολογιστή για διάφορα χρονικά βήματα. Έτσι, από τον πίνακα που ακολουθεί, ο αναγνώστης μπορεί να αντιληφθεί την έννοια του σφάλματος αποκοπής, συσσωρευτικά στην τελική λύση. Είναι δε χαρακτηριστικό ότι αυξάνοντας τον αριθμό των χρονικών βημάτων, η τελική λύση τείνει ασυμπτωτικά πρός την ίδια τιμή.

Χρονικά βήματα	Δt	$v(1sec)$
4	0.2500	6.725999529452197
10	0.1000	6.744658908029153
50	0.0200	6.747886087859326
100	0.0100	6.747984205916035
1000	0.0010	6.748016434066797
10000	0.0001	6.748016755312032

6.8 Μέθοδοι Πολλών Βημάτων

Κατά την παρουσίαση της μεθόδου Euler για την επίλυση της $\sigma.\delta.\varepsilon.$ 6.3, με αρχική συνθήκη την $y(x_0) = y_0$ καταλήξαμε σε ένα σχήμα το οποίο (σύμφωνα με την εξίσωση 6.19 για τον υπολογισμό της τιμής y_{i+1} στο x_{i+1} απαιτείται η τιμή y_i στο x_i και η τιμή $f(x_i, y_i)$. Η εξίσωση 6.19 μπορεί να γίνει αντιληπτή ως μια ολοκλήρωση στο διάστημα (x_i, x_{i+1}) της μορφής

$$y_{i+1} = y_i + \int_{x_i}^{x_{i+1}} f_i dx \tag{6.78}$$



Σχήμα 6.5: Παράδειγμα μειωμένης ακρίβειας κατά την ολοκλήρωση με το σχήμα της εξίσωσης 6.78.

Η έννοια της μονοβηματικής μεθόδου που χαρακτηρίζει τη μέθοδο Euler είναι πλήρως κατανοητή αφού για τον υπολογισμό του y_{i+1} απαιτούνται πληροφορίες μόνο από ένα βήμα 'πίσω', δηλαδή από τη θέση x_i. Συγχρόνως όμως είναι προφανής και η μειωμένη ακρίβεια ενός τέτοιου σχήματος, βλ. Σχήμα 6.5, λόγω του τρόπου ολοκλήρωσης που χρησιμοποιεί.

Στις μεθόδους πολλών βημάτων που θα παρουσιάσουμε στην τρέχουσα ενότητα, η ιδέα είναι να αντικατασταθεί η σχέση 6.78 με ένα σχήμα ολοκλήρωσης που θα εμπλέκει περισσότερα από ένα βήματα, λ.χ.

$$y_{i+1} \;=\; y_{i-k} + \int_{x_{i-k}}^{x_{i+1}} \psi_i(x)dx \tag{6.79}$$

όπου εδώ εμπλέκονται $k+1$ βήματα, δηλαδή πληροφορία που υπάρχει στο διάστημα από το x_{i-k} μέχρι το x_{i+1}. Η συνάρτηση $\psi_i(x)$ είναι ένα πολυώνυμο παρεμβολής το οποίο διέρχεται από τα σημεία $(x_{i-k}, f_{i-k}), \ldots, (x_i, f_i)$, εξασφαλίζοντας έτσι υψηλότερη ακρίβεια από τον αλγόριθμο της κατά τμήματα σταθερής ολοκλήρωσης. Όταν το πολυώνυμο παρεμβολής διέρχεται από τα προαναφερθέντα σημεία αναφερόμαστε σε σχέσεις *ανοικτής ολοκλήρωσης* (open integration formulas). Όταν επιπλέον το πολυώνυμο διέρχεται και από το (x_{i+1}, f_{i+1}) τότε πρόκειται για σχέσεις *κλειστής ολοκλήρωσης* (closed integration formulas).

Ισοδύναμα, μια σχέση της μορφής της 6.79 μπορεί να διατυπωθεί και στη μορφή

$$\begin{aligned} y_{i+1} &= a_1 y_i + a_2 y_{i-1} + \cdots + a_{k+1} y_{i-k} \\ &+ \Delta x \left[b_0 f(x_{i+1}, y_{i+1}) + b_1 f(x_i, y_i) + \cdots + b_{k+1} f(x_{i-k}, y_{i-k}) \right] \end{aligned} \quad (6.80)$$

Η εξίσωση 6.80 παριστάνει μια μέθοδο πολλών (συγκεκριμένα $k+1$) βημάτων. Όταν $b_0 = 0$, η μέθοδος λέγεται *ρητή* (explicit) ενώ για οποιαδήποτε άλλη τιμή $b_0 \neq 0$ η μέθοδος ονομάζεται *πεπλεγμένη* (implicit). Στην περίπτωση αυτή, η τιμή του y_{i+1} εμφανίζεται και στα δύο μέλη της εξίσωσης και αυτό πρέπει να ληφθεί υπόψη κατά τη διαχείριση της εξίσωσης. Επισημαίνεται ότι το άθροισμα των συντελεστών a και (χωριστά) αυτό των συντελεστών b πρέπει να ισούται με τη μονάδα.

Μια σημαντική παρατήρηση που αφορά τις μεθόδους πολλών βημάτων είναι το γεγονός ότι δεν είναι αυτο-εκκινούμενες. Για παράδειγμα, μια μέθοδος δύο βημάτων, βασισμένη σε σχέσεις της μορφής 6.79 ή 6.80, θα απαιτούσε τον υπολογισμό του y_1 (πέραν του y_0 που είναι γνωστό ως αρχική συνθήκη) πριν την ενεργοποίησή του. Μέθοδοι ενός βήματος, όπως οι Runge–Kutta οποιασδήποτε τάξης, μπορούν να χρησιμοποιηθούν για την υποστήριξη της εκκίνησης μεθόδων πολλών βημάτων.

6.8.1 Σχέσεις Ανοιχτής Ολοκλήρωσης

Το πολυώνυμο $\psi_i(x)$ της σχέσης 6.79 παρεμβάλει τα σημεία $(x_{i-k}, f_{i-k}), \ldots, (x_i, f_i)$ με όλες τις εμπλεκόμενες τιμές της συνάρτησης f να θεωρούνται γνωστές. Ξαναγράφουμε τη σχέση 6.79 ως

$$y_{i+1} = y_{i-k} + \Delta x \int_{-k}^{1} \psi_i(x_i + a\Delta x) da \quad (6.81)$$

και αναπτύσσουμε το $\psi_i(x_i + a\Delta x)$ σύμφωνα με το γνωστό τύπο των ανάντι πεπερασμένων διαφορών της παρεμβολής κατά Newton.

$$\begin{aligned} \psi_i(x_i + a\Delta x) &= f_i + a\nabla f_i + a(a+1)\frac{\nabla^2 f_i}{2!} + a(a+1)(a+2)\frac{\nabla^3 f_i}{3!} \\ &+ \cdots + a(a+1)(a+2)\ldots(a+r-1)\frac{\nabla^r f_i}{r!} \end{aligned} \quad (6.82)$$

όπου $f_i = f(x_i, y_i)$ και ο αδιάστατος μετρητής a ορίζεται ως

$$a = \frac{x - x_i}{\Delta x} \quad (6.83)$$

και $dx = \Delta x da$. Παραθέτουμε σε συντομία και μερικές από τις σχέσεις ανάντι πεπερασμένων διαφορών, για τη διευκόλυνση του αναγνώστη.

$$\begin{aligned} \nabla f(x) &= f(x) - f(x - \Delta x) \\ \nabla^2 f(x) &= \nabla f(x) - \nabla f(x - \Delta x) \\ \nabla^3 f(x) &= \nabla^2 f(x) - \nabla^2 f(x - \Delta x) \end{aligned} \quad (6.84)$$

Ετσι, το ολοκλήρωμα της σχέσης 6.79 γράφεται

$$\int_{x_{i-k}}^{x_{i+1}} \psi_i(x)dx = \Delta x \left[af_i + \frac{a^2}{2}\nabla f_i + a^2\left(\frac{a}{3}+\frac{1}{2}\right)\frac{\nabla^2 f_i}{2!} + a^2\left(\frac{a^2}{4}+a+1\right)\frac{\nabla^3 f_i}{3!} + a^2\left(\frac{a^3}{5}+\frac{3a^2}{2}+\frac{11a}{3}+3\right)\frac{\nabla^4 f_i}{4!} + \ldots \right]_{a=-k}^{a=1} \quad (6.85)$$

Από τη σχέση 6.85 και για διάφορες τιμές του k μπορούμε να δημιουργήσουμε διαφορετικούς αλγόριθμους. Έτσι, ενδεικτικά

- $k = 0$:

$$y_{i+1} = y_i + \Delta x \left(f_i + \frac{1}{2}\nabla f_i + \frac{5}{12}\nabla^2 f_i + \frac{3}{8}\nabla^3 f_i + \frac{251}{720}\nabla^4 f_i + \ldots \right) \quad (6.86)$$

- $k = 1$:

$$y_{i+1} = y_{i-1} + \Delta x \left(2f_i + \frac{1}{3}\nabla^2 f_i + \frac{1}{3}\nabla^3 f_i + \frac{29}{90}\nabla^4 f_i + \ldots \right) \quad (6.87)$$

Για τις περιττές τιμές του k θα παρατηρήσουμε ότι ο συντελεστής της k-στής ανάντι παραγώγου είναι μηδενικός και, γι αυτό, το συνηθέστερα χρησιμοποιούμενο σχήμα είναι με περιττό k και αποκοπή για $r = k$ (βλ. εξίσωση 6.82). Έτσι, για παράδειγμα, για $r = k = 3$ έχουμε ένα τυπικό σχήμα ανοιχτής ολοκλήρωσης

$$y_{i+1} = y_{i-3} + \Delta x \left(4f_i - 4\nabla f_i + \frac{8}{3}\nabla^2 f_i \right) \quad (6.88)$$

με σφάλμα ίσο με

$$\frac{14}{45}\Delta x^5 f^{(4)}(\xi) \quad (6.89)$$

όπου $x_{i-3} < \xi < x_{i+1}$. Η σχέση 6.89 προκύπτει από τη γενική σχέση του σφάλματος που προκύπτει για αποκοπή μετά τον όρο που περιέχει την r–στή ανάντι παράγωγο της f και ο οποίος δίνεται χωρίς απόδειξη ως

$$\Delta x^{r+2} \int_{-k}^{1} \frac{a(a+1)(a+2)\ldots(a+r)}{(r+1)!}\, \psi^{(r+1)}(\xi)da \quad , \quad x_{i-k} < \xi < x_{i+1} \quad (6.90)$$

Η απόδειξη της είναι εξάλλου προφανής, δεδομένης της σχέσης 6.82.

Σχέσεις όπως οι 6.86, 6.87 και 6.88 γίνονται άμεσα χρησιμοποιήσιμες αν αντικατασταθούν οι ανάντι πεπερασμένες διαφορές ως συνάρτηση των τιμών της f στα εμπλεκόμενα σημεία. Δίνονται στη συνέχεια τρεις τέτοιες εκφράσεις για άμεση χρήση

- $k = 1$, $r = 1$:

$$y_{i+1} = y_{i-1} + 2\Delta x f_i + O(\Delta x^3) \tag{6.91}$$

- $k = 3$, $r = 3$:

$$y_{i+1} = y_{i-3} + \frac{4\Delta x}{3}(2f_i - f_{i-1} + 2f_{i-2}) + O(\Delta x^5) \tag{6.92}$$

- $k = 5$, $r = 5$:

$$y_{i+1} = y_{i-5} + \frac{3\Delta x}{10}(11f_i - 14f_{i-1} + 26f_{i-2} - 14f_{i-3} + 11f_{i-4}) + O(\Delta x^7) \tag{6.93}$$

Η τελευταία λ.χ. σχέση χρησιμοποιεί ήδη υπολογισθείσα πληροφορία στα σημεία $x_{i-5}, x_{i-4}, \ldots, x_i$ και παρεκβάλλει στο διάστημα $[x_i, x_{i+1}]$.

6.8.2 Σχέσεις Κλειστής Ολοκλήρωσης

Η παρουσίαση των σχέσεων κλειστής ολοκλήρωσης μπορεί να γίνει εύκολα όταν έχουν ήδη παρουσιαστεί οι σχέσεις της ανοιχτής ολοκλήρωσης. Η διαφορά, όπως έχει ήδη αναφερθεί, εστιάζεται στο ότι η παρεμβολή καλύπτει και το σημείο (x_{i+1}, f_{i+1}), οπότε η συνάρτηση $\psi_i(x)$ αναπτύσσεται με βάση τις σχέσεις ανάντι πεπερασμένων διαφορών στο x_{i+1} αντί του x_i (που χρησιμοποιήθηκε προηγούμενα). Η αντίστοιχη της σχέσης 6.85 θα είναι η

$$\begin{aligned} y_{i+1} &= y_{i-k} + \Delta x \int_{-k}^{1} [\, f_{i+1} + (a-1)\nabla f_{i+1} + \frac{(a-1)a}{2!}\nabla^2 f_{i+1} + \frac{(a-1)a(a+1)}{3!}\nabla^3 f_{i+1} \\ &+ \cdots + \frac{(a-1)a(a+1)\ldots(a+r-2)}{r!}\nabla^r f_{i+1}\,]\, da \end{aligned} \tag{6.94}$$

οπότε, μετά την ολοκλήρωση, γράφεται

$$\begin{aligned} y_{i+1} &= y_{i-k} + \Delta x\,[\, a f_{i+1} + a\left(\frac{a}{2} - 1\right)\nabla f_{i+1} + \frac{a^2(\frac{a}{3} - \frac{1}{2})}{2!}\nabla^2 f_{i+1} + \frac{a^2(\frac{a^2}{4} - \frac{1}{2})}{3!}\nabla^3 f_{i+1} \\ &+ \frac{(\frac{a^5}{5} + \frac{a^4}{2} - \frac{a^3}{3} - a^2)}{4!}\nabla^4 f_{i+1} + \ldots\,]_{-k}^{1} \end{aligned} \tag{6.95}$$

Πρακτικά, λ.χ., για $k = 5$ προκύπτει

$$\begin{aligned} y_{i+1} = y_{i-5} &+ \Delta x\,(\, 6f_{i+1} - 18\nabla f_{i+1} + 27\nabla^2 f_{i+1} - 24\nabla^3 f_{i+1} \\ &+ \frac{123}{10}\nabla^4 f_{i+1} - \frac{33}{10}\nabla^5 f_{i+1} + \ldots\,) \end{aligned} \tag{6.96}$$

ενώ το σφάλμα που αντιστοιχεί σε αποκοπή μετά τον όρο της r–στής παραγώγου θα είναι ίσο με

$$\Delta x^{r+2} \int_{-k}^{1} \frac{(a-1)a(a+1)\dots(a+r-1)}{(r+1)!}\, \psi^{(r+1)}(\xi)da \quad , \quad x_{i-k} < \xi < x_{i+1} \tag{6.97}$$

Έτσι, για παράδειγμα, δίνονται δύο ενδεικτικές περιπτώσεις - σχήματα :

- $k = 1$, $r = 3$:

$$y_{i+1} \;=\; y_{i-1} + \Delta x \left(2f_{i+1} - 2\nabla f_{i+1} + \frac{1}{3}\nabla^2 f_{i+1} \right) + O(\Delta x^5) \tag{6.98}$$

ή, με ανάπτυξη,

$$y_{i+1} \;=\; y_{i-1} + \frac{\Delta x}{3}\left(f_{i+1} + 4f_i + f_{i-1} \right) + O(\Delta x^5) \tag{6.99}$$

- $k = 3$, $r = 5$:

$$y_{i+1} \;=\; y_{i-3} + \Delta x \left(4f_{i+1} - 8\nabla f_{i+1} + \frac{20}{3}\nabla^2 f_{i+1} - \frac{8}{3}\nabla^3 f_{i+1} + \frac{14}{45}\nabla^4 f_{i+1} \right) + O(\Delta x^7) \tag{6.100}$$

ή, με ανάπτυξη,

$$y_{i+1} \;=\; y_{i-3} + \frac{2\Delta x}{45}\left(7f_{i+1} + 32f_i + 12f_{i-1} + 32f_{i-2} + 7f_{i-3} \right) + O(\Delta x^7) \tag{6.101}$$

Εφαρμογή

Εφαρμόστε τη μέθοδο κλειστής ολοκλήρωσης με $k = 3$, $r = 5$, για την αριθμητική επίλυση της εξίσωσης ακτινοβολίας των Stefan–Boltzmann, με τις ίδιες αρχικές τιμές που χρησιμοποιήσατε και προηγουμένως.

Λύση:

Χρησιμοποιούμε το σχήμα 6.101 ως βασικό σχήμα επίλυσης. Επειδή η μέθοδος δεν είναι αυτο-εκκινούμενη, τα πρώτα τρία βήματα πραγματοποιούνται με την ήδη γνωστή Runge–Kutta τέταρτης τάξης. Η ανάγκη χρήσης της ποσότητας f_{i+1}, στο δεξιό μέλος της εξίσωσης απαιτεί *εσωτερικές επαναλήψεις* για τη σύγκλιση της τιμής του y_{i+1}. Στον κώδικα που ακολουθεί, πραγματοποιούνται 10 εσωτερικές επαναλήψεις σε κάθε χρονικό

βήμα. Δεν χρησιμοποιείται κριτήριο σύγκλισης του επαναληπτικού σχήματος, αν και αυτό μπορεί να προστεθεί εύκολα από τον αναγνώστη. Παρατηρείστε την ανάγκη αποθήκευσης σε μεταβλητές με δείκτη (πίνακες) τόσο της υπό εξέλιξης λύσης όσο και της συνάρτησης f (για την τελευταία, ο λόγος είναι απλά η αποφυγή περιττών πολλαπλών υπολογισμών της ίδιας συνάρτησης). Ο κώδικας ακολουθεί:

```
      program stefan_close35
c     Open Integration Formula (k=3, r=3)
      implicit double precision (a-h,o-z)
      dimension temp(0:10),f(0:10)
      fun(t)=-2.d-12*(t**4-250.d0**4)
c
      time=0.d0
      deltat=1.0
      temp(0)=2500.d0
      f(0)=fun(temp(0))
      write(*,'(2x,f5.1,3x,f16.8)')time,temp(0)
c
      do i=1,10
      if(i.lt.4)then
      ak1 = deltat*fun(temp(i-1))
      ak2 = deltat*fun(temp(i-1)+ak1/3.d0)
      ak3 = deltat*fun(temp(i-1)-ak1/3.d0+ak2)
      ak4 = deltat*fun(temp(i-1)+ak1-ak2+ak3)
      temp(i)=temp(i-1)+(ak1+3.*ak2+3.*ak3+ak4)/8.d0
      else
      temp(i)=temp(i-1)  !   initialization
      do iter=1,10
      temp(i)=temp(i-4)+2.d0*deltat/45.d0*
     :        (7.*f(i)+32.*f(i-1)+12.*f(i-2)
     :        +32.*f(i-3)+7.*f(i-4))
      f(i)=fun(temp(i))
      enddo
      endif
      time=time+deltat
      f(i)=fun(temp(i))
      write(*,'(2x,f5.1,3x,f16.8)')time,temp(i)
      enddo
c
      end
```

Στη συνέχεια, πινακοποιούνται οι λύσεις και τα σφάλματα από την αναλυτική λύση:

t (sec)	Τ(αριθμητική)	Τ(αριθμητική) -Τ(αναλυτική)
0.0	2500.00000000	0.00000000
1.0	2426.43483927	-0.00003432
2.0	2360.82993707	-0.00005139
3.0	2301.79069129	-0.00005939
4.0	2248.24713185	-0.00018220
5.0	2199.36255815	-0.00011178
6.0	2154.47071012	-0.00008564
7.0	2113.03380119	-0.00005992
8.0	2074.61172171	-0.00017617
9.0	2038.84075236	-0.00009410
10.0	2005.41629441	-0.00007140

6.9 Μέθοδοι Πρόβλεψησ–Διόρθωσης

Με κύριο στόχο να φανεί η ανάγκη η οποία υπαγόρευσε την ανάπτυξη των λεγομένων μεθόδων πρόβλεψησ–διόρθωσης, θα επιχειρήσουμε αρχικά μια σύγκριση ανάμεσα στα σχήματα ανοιχτής και κλειστής ολοκλήρωσης. Για τη σύγκριση αυτή, ας πάρουμε δύο σχήματα ίδιας τάξης ακριβείας, λ.χ. $O(\Delta x^5)$. Με το σχήμα ανοιχτής ολοκλήρωσης για $k = 3$ και $r = 3$, που περιγράφουν οι σχέσεις 6.88 ή 6.92, το αντίστοιχο σφάλμα αποκοπής δίνεται από τη σχέση 6.89. Ίδιας τάξης ακρίβεια δίνει το σχήμα κλειστής ολοκλήρωσης 6.98 ή 6.99 με σφάλμα αποκοπής ίσο με

$$-\frac{1}{90}\,\Delta x^5\; f^{(4)}(\xi) \qquad , \qquad x_{i-1} < \xi < x_{i+1} \tag{6.102}$$

(η έκφραση του σφάλματος αποκοπής προκύπτει εύκολα από τη σχέση 6.97 για ($k = 1$, $r = 3$). Σε επίπεδο σφάλματος αποκοπής η σύγκριση αποβαίνει υπέρ του σχήματος κλειστής ολοκλήρωσης, αφού ο συντελεστής (σε απόλυτη τιμή) είναι κατά πολύ μικρότερος, $\frac{1}{90} << \frac{14}{45}$. Όμως, σε επίπεδο απλότητας εφαρμογής υπερτερεί σαφώς το σχήμα της ανοιχτής ολοκλήρωσης, δηλ. το σχήμα 6.92 σε σύγκριση με το 6.99, αφού το πρώτο είναι ρητό και το δεύτερο πεπλεγμένο, εμπλέκει δηλαδή την τιμή του y_{i+1} για τον υπολογισμό του ίδιου του y_{i+1}.

Με βάση το χαμηλότερο σφάλμα αποκοπής, ας εστιάσουμε την προσοχή μας σε τρόπους με τους οποίους θα μπορούσε να λυθεί το πεπλεγμένο σχήμα 6.99, που εποπτικά ξαναγράφεται ως

$$y_{i+1} \;=\; y_{i-1} + \frac{\Delta x}{3}\left(f(x_{i+1}, y_{i+1}) + 4f_i + 4f_{i-1}\right)$$

Ένας τυπικός τρόπος για τη διαχείριση της είναι η *μέθοδος των διαδοχικών αντικαταστάσεων*. Δηλαδή, ο υπολογισμός της τιμής της y_{i+1} πραγματοποιείται μέσω

διαδοχικών επαναλήψεων (δείκτης j), μέχρι τελικής σύγκλισης, με ένα σχήμα της μορφής

$$y_{i+1,j+1} \;=\; y_{i-1} + \frac{\Delta x}{3}\left(f(x_{i+1}, y_{i+1,j}) + 4f_i + 4f_{i-1}\right) \qquad , \quad j = 0, 1, 2, \ldots \tag{6.103}$$

Το σχήμα 6.103 προϋποθέτει την αυθαίρετη θεώρηση μιας αρχικής τιμής του $y_{i+1,0}$. Στις διαδοχικές αντικαταστάσεις για τον υπολογισμό του y_{i+1}, κάθε άλλη ποσότητα (όπως λ.χ. τα y_{i-1}, f_i, f_{i-1}) θεωρείται γνωστή και 'σταθερή 'έτσι ώστε το σχήμα 6.103 να λαμβάνει τη μορφή

$$y_{i+1,j+1} \;=\; F(y_{i+1,j}) \;=\; \frac{\Delta x}{3} f(x_{i+1}, y_{i+1,j}) + C \tag{6.104}$$

όπου C η ποσότητα που προαναφέραμε. Το επαναληπτικό σχήμα 6.104 συγκλίνει υπό την προϋπόθεση ότι

$$\left|F^{'}(y_{i+1})\right| \;=\; \frac{\Delta x}{3}\left|\frac{\partial f(y_{i+1})}{\partial y}\right| \;<\; 1 \tag{6.105}$$

Η ανισότητα 6.105 δίνει ένα άνω φράγμα στην τιμή του βήματος Δx (για δεδομένη μορφή της συνάρτησης f) ώστε να συγκλίνει η επαναληπτική διαδικασία. Γενικά, η προτίμηση στα σχήματα κλειστής ολοκλήρωσης λόγω του μικρότερου σφάλματος αποκοπής, πρέπει να συνοδεύεται με ταχύτατη σύγκλιση της επαναληπτικής διαδικασίας 6.104 αφού κάθε επιπλέον διαδοχική αντικατάσταση απαιτεί έναν επιπλέον υπολογισμό της τιμής της συνάρτησης f. Για τη μείωση του υπολογιστικού κόστους που ενέχει η διαδικασία των διαδοχικών αντικαταστάσεων, ο προγραμματιστής πρέπει να επιλέξει προσεκτικά αφενός μεν την αρχική τιμή $y_{i+1,0}$ αφετέρου δε το βήμα Δx.

Για την καλύτερη πρόβλεψη της τιμής του $y_{i+1,0}$ μπορούμε να χρησιμοποιήσουμε σχήματα ανοιχτής ολοκλήρωσης, όπως αυτά που παρουσιάσαμε σε προηγούμενη ενότητα. Προβλέποντας την τιμή του y_{i+1} (δηλ. υπολογίζοντας το $y_{i+1,0}$) μέσω ενός σχήματος ανοιχτής ολοκλήρωσης και στη συνέχεια διορθώνοντας την τιμή αυτή με ένα επαναληπτικό σχήμα όπως τη σχέση 6.103 της κλειστής ολοκλήρωσης, δημιουργούνται οι *μέθοδοι δύο βημάτων* που ήδη ονομάσαμε *μεθόδους πρόβλεψησ–διόρθωσης* (prediction–corrector). Μερικές από τις πιο συνηθισμένες μεθόδους πρόβλεψησ–διόρθωσης θα παρουσιάσουμε στη συνέχεια:

Μέθοδος Milne τέταρτης τάξης. Περιλαμβάνει τα εξής δύο βήματα:

Πρόβλεψη:

$$y_{i+1} \;=\; y_{i-3} + \frac{4\Delta x}{3}\left(2f_i - f_{i-1} + 2f_{i-2}\right) + O(\Delta x^5) \tag{6.106}$$

Διόρθωση:

$$y_{i+1} \;=\; y_{i-1} + \frac{\Delta x}{3}\left(f_{i+1} + 4f_i + f_{i-1}\right) + O(\Delta x^5) \tag{6.107}$$

(Η μέθοδος Milne τέταρτης τάξης χρησιμοποιεί ως πρόβλεψη τη σχέση 6.92 και ως διόρθωση τη σχέση 6.99 την οποία επιλύει επαναληπτικά με βάση το σχήμα 6.103).

Μέθοδος Milne έκτης τάξης. Περιλαμβάνει τα εξής δύο βήματα:

Πρόβλεψη:

$$y_{i+1} = y_{i-5} + \frac{3\Delta x}{10}\left(11f_i - 14f_{i-1} + 26f_{i-2} - 14f_{i-3} + 11f_{i-4}\right) + O(\Delta x^7) \tag{6.108}$$

Διόρθωση:

$$y_{i+1} = y_{i-3} + \frac{2\Delta x}{45}\left(7f_{i+1} + 32f_i + 12f_{i-1} + 32f_{i-2} + 7f_{i-3}\right) + O(\Delta x^7) \tag{6.109}$$

(Η πρόβλεψη βασίζεται στη σχέση 6.93 και η διόρθωση στη σχέση 6.101, η οποία λύνεται επαναληπτικά σύμφωνα με το σχήμα 6.103).

Τροποποιημένη Μέθοδος Adams ή Μέθοδος Adams–Moulton. Περιλαμβάνει τα εξής δύο βήματα:

Πρόβλεψη:

$$y_{i+1} = y_i + \frac{\Delta x}{24}\left(55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}\right) + O(\Delta x^5) \tag{6.110}$$

Διόρθωση:

$$y_{i+1} = y_i + \frac{\Delta x}{24}\left(9f_{i+1} + 19f_i - 5f_{i-1} + f_{i-2}\right) + O(\Delta x^5) \tag{6.111}$$

6.10 Συστήματα Συνήθων Διαφορικών Εξισώσεων

Το σύστημα M *σ.δ.ε.* πρώτης τάξης επιλύεται αριθμητικά με μεθόδους οι οποίες αποτελούν επέκταση–γενίκευση αυτών που ήδη παρουσιάσθηκαν για απλές *σ.δ.ε.*. Εκτός του ότι πολλά φυσικά προβλήματα καταλήγουν σε συστήματα *σ.δ.ε.*, τονίζεται ότι, όπως αναλύθηκε και σε προηγούμενη ενότητα, η αριθμητική επίλυση *σ.δ.ε.* μεγαλύτερης τάξης οδηγεί επίσης στην επίλυση συστημάτων *σ.δ.ε.*.

Γενική γραφή ενός συστήματος M *σ.δ.ε.* πρώτης τάξης είναι η παρακάτω

$$\begin{aligned}
\tfrac{dy_1}{dx} &= f_1(x, y_1, y_2, \ldots, y_M) \\
\tfrac{dy_2}{dx} &= f_2(x, y_1, y_2, \ldots, y_M) \\
&\vdots \\
\tfrac{dy_M}{dx} &= f_M(x, y_1, y_2, \ldots, y_M)
\end{aligned} \tag{6.112}$$

Η επίλυση του συστήματος 6.112, ως πρόβλημα αρχικών τιμών, διέπεται από τις αρχικές συνθήκες

$$\begin{aligned} y_1(x_0) &= y_{1,0} = known \\ y_2(x_0) &= y_{2,0} = known \\ &\vdots \\ y_M(x_0) &= y_{M,0} = known \end{aligned} \tag{6.113}$$

Όπως φαίνεται και από τις εκφράσεις 6.113, υιοθετείται η χρήση διπλού κάτω δείκτη στις άγνωστες ποσότητες y_m. Έτσι, στα επόμενα, το σύμβολο $y_{m,i}$ θα δηλώνει την τιμή της συνάρτησης y_m στη θέση x_i που υπολογίσθηκε αριθμητικά. Δηλαδή, η $y_{m,i}$ θα προσεγγίζει την $y_m(x_i)$.

Δεδομένης της άμεσης 'ομοιότητασ' των μεθόδων επίλυσης συστημάτων *σ.δ.ε.* με όσα εκτενώς παρουσιάσθηκαν για μια *σ.δ.ε.* θα παρουσιασθεί στη συνέχεια επιλεκτικά η μέθοδος Runge–Kutta τέταρτης τάξης. Εδώ, οι ενδιάμεσες ποσότητες k θα συμβολίζονται με δύο δείκτες, ως $k_{m,j}$, όπου ο πρώτος δείκτης θα δηλώνει τον αύξοντα αριθμό της άγνωστης ποσότητας στην οποία αντιστοιχεί ($m = 1, \ldots, M$) ενώ ο δεύτερος θα δηλώνει το βήμα της μεθόδου Runge–Kutta ($j = 1, \ldots, 4$). Ο αλγόριθμος για τον υπολογισμό της ποσότητας $y_{m,i+1}$, όταν είναι ήδη γνωστή η ποσότητα $y_{m,i}$ (από αντίστοιχη εφαρμογή της αναδρομικής σχέσης ή ως αρχική συνθήκη), έχει ως εξής:

$$\begin{aligned} k_{m,1} &= \Delta x\ f_m(x_i, y_{1,i}, y_{2,i}, \ldots, y_{M,i}) \qquad , m = 1, \ldots, M \\ y_m^{(1)} &= y_{m,i} + \frac{1}{2} k_{m,1} \qquad , m = 1, \ldots, M \\ k_{m,2} &= \Delta x\ f_m(x_i + \frac{\Delta x}{2}, y_1^{(1)}, y_2^{(1)}, \ldots, y_M^{(1)}) \qquad , m = 1, \ldots, M \\ y_m^{(2)} &= y_{m,i} + \frac{1}{2} k_{m,2} \qquad , m = 1, \ldots, M \\ k_{m,3} &= \Delta x\ f_m(x_i + \frac{\Delta x}{2}, y_1^{(2)}, y_2^{(2)}, \ldots, y_M^{(2)}) \qquad , m = 1, \ldots, M \\ y_m^{(3)} &= y_{m,i} + k_{m,3} \qquad , m = 1, \ldots, M \\ k_{m,4} &= \Delta x\ f_m(x_i + \Delta x, y_1^{(3)}, y_2^{(3)}, \ldots, y_M^{(3)}) \qquad , m = 1, \ldots, M \\ y_{m,i+1} &= y_{m,i} + \frac{1}{6}\left(k_{m,1} + 2k_{m,2} + 2k_{m,3} + k_{m,4}\right) \qquad , m = 1, \ldots, M \end{aligned} \tag{6.114}$$

Σε επίπεδο προγραμματισμού, κάθε γραμμή στις σχέσεις 6.114 αντιστοιχεί σε ένα βρόχο που πραγματοποιεί τόσους υπολογισμούς όσες και οι εξισώσεις του συστήματος. Πρόκειται δε για έναν σειριακό αλγόριθμο λόγω της εμπλοκής όλων των μεταβλητών σε κάθε υπολογιστική σχέση. Στην εφαρμογή που ακολουθεί δίνεται σχετικός κώδικας σε Fortran 77 ο οποίος, παρόλο που είναι ειδικά γραμμένος για τη συγκεκριμένη εφαρμογή, εντούτοις εύκολα γενικεύεται.

Εφαρμογή

Σε ένα απομονωμένο νησί ζούν αποκλειστικά λαγοί και αλεπούδες. Λόγω της πυκνής βλάστησης, οι λαγοί βρίσκουν άφθονη τροφή, πρακτικά ανεξάρτητα από τον εκάστοτε πληθυσμό τους. Από την άλλη πλευρά, οι αλεπούδες στηρίζουν τη διατροφή τους και άρα εξαρτούν την ύπαρξή τους από το κυνήγι του λαγού. Αν κατά τη χρονική στιγμή t ο πληθυσμός των λαγών και των αλεπούδων συμβολίζεται αντίστοιχα με $N_R(t)$ και $N_F(t)$, η χρονική εξέλιξη του οικοσυστήματος λαγών–αλεπούδων περιγράφεται από το σύστημα

$$\begin{aligned} \frac{dN_R(t)}{dt} &= \alpha N_R(t) - \beta N_R(t) N_F(t) \\ \frac{dN_F(t)}{dt} &= -\gamma N_F(t) + \delta N_R(t) N_F(t) \end{aligned} \tag{6.115}$$

Η φυσική σημασία των συντελεστών είναι εύλογη: α είναι ο ρυθμός αναπαραγωγής λαγών, β είναι ο ρυθμός εξολόθρευσης λαγών από τις αλεπούδες, γ είναι ο ρυθμός φυσικού θανάτου των αλεπούδων και δ είναι ο συντελεστής επιβίωσης των αλεπούδων.

Με τη βοήθεια υπολογιστικού κώδικα παρουσιάστε γραφήματα (α) $N_R = N_R(t)$, (β) $N_F = N_F(t)$ και (γ) $N_F = N_F(N_R)$, για δεδομένους αρχικούς πληθυσμούς $N_R(t=0)$ και $N_F(t=0)$. Για την αριθμητική εφαρμογή χρησιμοποιείστε τις τιμές: $\alpha = 1.2$, $\beta = 0.6$, $\gamma = 0.8$, $\delta = 0.3$, $N_R(t=0) = 1000$ και $N_F(t=0) = 50$.

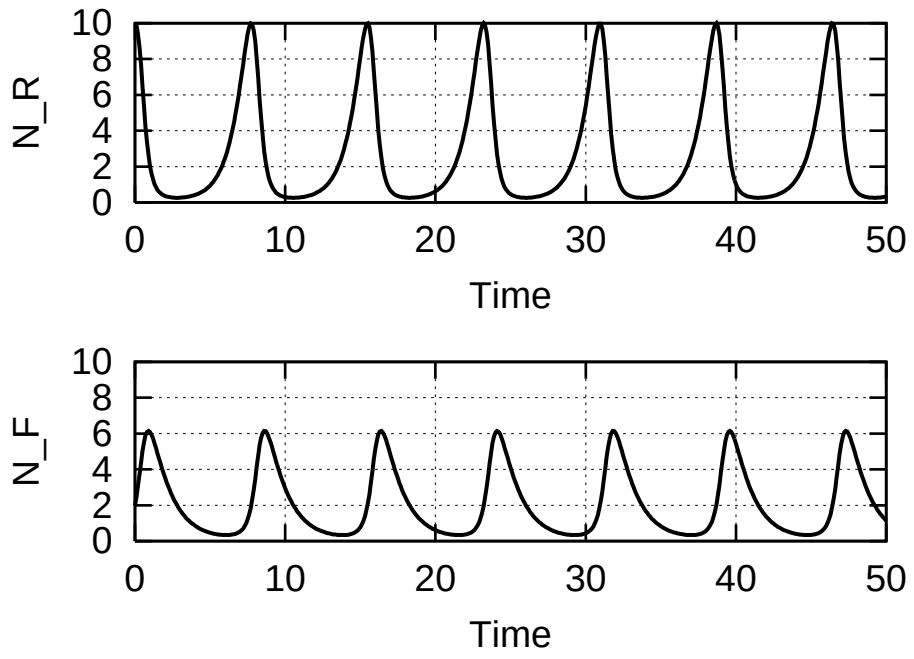
Λύση:

Η χρήση της μεθόδου Runge–Kutta τέταρτης τάξης (σχέση 6.114) υλοποιείται σε κώδικα γραμμένο σε Fortran 77.

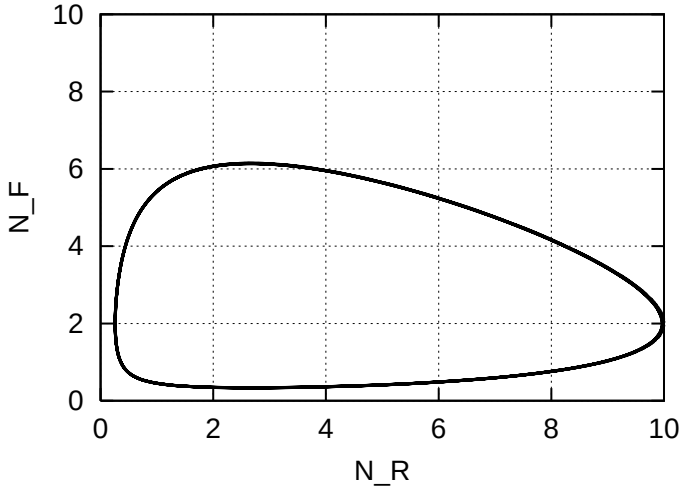
```
      program rabbit_fox
      implicit double precision (a-h,o-z)
      dimension y(2),aux(2),ak(2,4)
      fun1(yrab,yfox)=1.2d0*yrab-0.6d0*yrab*yfox
      fun2(yrab,yfox)=-0.8d0*yfox+0.3d0*yrab*yfox
c
      time=0.d0
      deltat=0.1
c
      neqs=2
      y(1)=10
      y(2)=2
      write(*,'(2x,f5.1,10(1x,f10.5))')time,(y(i),i=1,neqs)
c
      do ktimestep=1,500
```

```
c-step 1:
         ak(1,1) = deltat*fun1(y(1),y(2))
         ak(2,1) = deltat*fun2(y(1),y(2))
         aux(1) = y(1)+0.5d0*ak(1,1)
         aux(2) = y(2)+0.5d0*ak(2,1)
c-step 2:
         ak(1,2) = deltat*fun1(aux(1),aux(2))
         ak(2,2) = deltat*fun2(aux(1),aux(2))
         aux(1) = y(1)+0.5d0*ak(1,2)
         aux(2) = y(2)+0.5d0*ak(2,2)
c-step 3:
         ak(1,3) = deltat*fun1(aux(1),aux(2))
         ak(2,3) = deltat*fun2(aux(1),aux(2))
         aux(1) = y(1)+ak(1,3)
         aux(2) = y(2)+ak(2,3)
c-step 4:
         ak(1,4) = deltat*fun1(aux(1),aux(2))
         ak(2,4) = deltat*fun2(aux(1),aux(2))
c-synthesis:
      y(1)=y(1)+(ak(1,1)+2.d0*ak(1,2)+2.d0*ak(1,3)+ak(1,4))/6.d0
      y(2)=y(2)+(ak(2,1)+2.d0*ak(2,2)+2.d0*ak(2,3)+ak(2,4))/6.d0
      time=time+deltat
      write(*,'(2x,f5.1,10(1x,f10.5))')time,(y(i),i=1,neqs)
      enddo
c
      end
```

Επιλέχθηκε χρονικό βήμα ίσο με 0.1 χρονικές μονάδες (μεταβλητή *deltat* και ο υπολογισμός πραγματοποιήθηκε για 500 χρονικά βήματα. Στο Σχήμα 6.6 παρουσιάζονται γραφικά οι χρονικές εξελίξεις των πληθυσμών $N_R = N_R(t)$ των λαγών και $N_F = N_F(t)$ των αλεπούδων. Είναι εμφανής η περιοδικότητα που παρουσιάζουν, η οποία είναι ενδιαφέρον και εύκολο να εξηγηθεί. Η περιοδικότητα αυτή απεικονίζεται με διαφορετικό τρόπο στο Σχήμα 6.7, στη μορφή $N_F = N_F(N_R)$. Παρατηρούμε έναν κλειστό βρόχο από τον οποίο εντοπίζονται οι μέγιστες και ελάχιστες τιμές των δύο πληθυσμών.



Σχήμα 6.6: Γραφική αναπαράσταση της εξέλιξης των πληθυσμών $N_R = N_R(t)$ των λαγών και $N_F = N_F(t)$ των αλεπούδων ως συνάρτηση του χρόνου. Χρησιμοποιήθηκε Runge–Kutta τέταρτης τάξης με χρονικό βήμα ίσο με 0.1 χρονικές μονάδες.



Σχήμα 6.7: Γραφική αναπαράσταση της εξέλιξης των πληθυσμών λαγών και αλεπούδων στη μορφή διαγράμματος $N_F = N_F(N_R)$.

Βιβλιογραφία

1. Al-Khafaji A.W. and Tooley J.R., Numerical Methods in Engineering Practice, Rinehart and Winston, NY 1986.

2. Carnahan B., Luther H.A. and Wilkes J.O., Applied Numerical Methods, Krieger Publishing Co, 1990.

3. Chapra St. C. and Canale R.P., Numerical Methods for Engineers, with Programming and Software Applications, McGraw-Hill, 1998.

4. Davis P.J. and Rabinowitz P., Methods of Numerical Integration, Academic Press, 1975

5. Ferziger J.H., Numerical Methods for Engineering Application, Wiley, NY 1981.

6. Gerald C.F. and Wheatley P.O., Applied Numerical Analysis, Addison-Wesley, 1997.

7. Hamming R.W., Numerical Methods for Scientists and Engineers, McGraw-Hill, 1973.

8. Heath M.T., Scientific Computing – An Introductory Survey, McGraw-Hill, 2002.

9. Hoffman J.D., Numerical Methods for Engineers and Scientists, McGraw-Hill, 1993

10. Press, Flanney, Teukolsky and Vettering, Numerical Recipes- The Art of Scientific Computing, Cambridge University Press, 1986.

11. Rice J.R., Numerical Methods, Software and Analysis, McGraw-Hill, 1983.

12. Stewart G.W., Afternotes on Numerical Analysis, Siam, 1996.

13. Taylor J.R., An Introduction to Error Analysis, Mill Valley, CA, 1982.