

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ  
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

# Εισαγωγή στην Αριθμητική Ανάλυση

Σημειώσεις Διαλέξεων και Εργαστηρίων

Ηράκλειο  
2019

Copyright © 2005 – 2019

Το έργο αυτό αδειοδοτείται από την άδεια “Creative Commons Αναφορά Δημιουργού - Μη Εμπορική Χρήση - Παρόμοια Διανομή 4.0 Διεθνές” (CC-BY-NC-SA 4.0). Για να δείτε ένα αντίγραφο της άδειας αυτής, επισκεφτείτε το <http://creativecommons.org/licenses/by-nc-sa/4.0/deed.el>.

Στη συγγραφή συνεισέφεραν οι Μ. Γραμματικάκης, Θ. Καλαμπούκης, Γ. Κοπιδάκης, Ν. Παπαδάκης, Σ. Σταματιάδης.

Υπεύθυνος σημειώσεων: Σ. Σταματιάδης ([stamatis@materials.uoc.gr](mailto:stamatis@materials.uoc.gr)).

Η στοιχειοθεσία έγινε από τον Σ. Σταματιάδη με τη χρήση του  $\text{\LaTeX}$ .

Τελευταία τροποποίηση του κειμένου έγινε στις 6 Φεβρουαρίου 2019. Η πιο πρόσφατη έκδοση βρίσκεται στο <http://www.materials.uoc.gr/el/undergrad/courses/ETY213>

# Περιεχόμενα

Περιεχόμενα	i
<b>1 Σφάλματα</b>	<b>1</b>
1.1 Εισαγωγή	1
1.2 Συστήματα αρίθμησης	2
1.2.1 Αναπαράσταση ακεραίων	2
1.2.2 Αναπαράσταση πραγματικών	3
1.3 Αναπαράσταση αριθμών στον ΗΥ	3
1.3.1 Ακέραιοι	3
1.3.2 Πραγματικοί	4
1.4 Ασκήσεις	6
<b>2 Επίλυση μη Γραμμικών Εξισώσεων</b>	<b>9</b>
2.1 Εισαγωγή	9
2.1.1 Ταχύτητα σύγκλισης	10
2.1.2 Ευστάθεια	10
2.1.3 Εύρεση περισσότερων της μίας ριζών	11
2.1.4 Χρήσιμα θεωρήματα	11
2.2 Μέθοδος Διχοτόμησης	11
2.2.1 Ακρίβεια αλγορίθμου διχοτόμησης	13
2.2.2 Σύγκλιση αλγορίθμου διχοτόμησης	14
2.2.3 Αριθμός επαναλήψεων αλγορίθμου διχοτόμησης	15
2.3 Μέθοδος ψευδούς σημείου	15
2.4 Μέθοδος τέμνουσας	16
2.4.1 Σύγκλιση της μεθόδου τέμνουσας	17
2.5 Μέθοδος Müller	17
2.6 $x = g(x)$	19
2.6.1 Ορισμός-Σχετικά Θεωρήματα	19
2.6.2 Σύγκλιση της μεθόδου σταθερού σημείου	20
2.7 Μέθοδοι Householder	22
2.7.1 Μέθοδος Newton-Raphson	22
2.7.2 Μέθοδος Halley	25
2.8 Ασκήσεις	25

<b>3</b>	<b>Επίλυση Γραμμικών Συστημάτων</b>	<b>29</b>
3.1	Εισαγωγή	29
3.1.1	Ευστάθεια γραμμικών συστημάτων	29
3.1.2	Ορισμοί–Βασικές γνώσεις	30
3.2	Απευθείας μέθοδοι	33
3.2.1	Κανόνας Cramer	33
3.2.2	Απαλοιφή Gauss	34
3.2.3	Μέθοδος Gauss–Jordan	41
3.2.4	Ανάλυση LU	41
3.3	Επαναληπτικές Μέθοδοι	44
3.3.1	Στατικές μέθοδοι	45
3.3.2	Μέθοδοι προβολής	47
3.4	Εφαρμογές	47
3.4.1	Υπολογισμός του αντίστροφου πίνακα	47
3.4.2	Υπολογισμός ορίζουσας	49
3.4.3	Εύρεση ιδιοτιμών και ιδιοδιανυσμάτων	50
3.4.4	Επίλυση συστήματος μη γραμμικών εξισώσεων	52
3.5	Ασκήσεις	55
<b>4</b>	<b>Προσέγγιση Συναρτήσεων</b>	<b>57</b>
4.1	Προσέγγιση με πολυώνυμο	57
4.1.1	Μετατροπή	59
4.1.2	Σφάλμα προσέγγισης με πολυώνυμο	60
4.2	Προσέγγιση με λόγο πολυωνύμων	62
4.3	Προσέγγιση κατά τμήματα	62
4.4	Προσέγγιση με spline	63
4.5	Προσέγγιση παραγώγων	64
4.6	Ελάχιστα τετράγωνα	67
4.6.1	Ευθεία ελάχιστων τετραγώνων	67
4.6.2	Πολυώνυμο ελάχιστων τετραγώνων	69
4.6.3	Καμπύλη ελάχιστων τετραγώνων $f(y) = \alpha g(x) + \beta$	70
4.7	Ασκήσεις	71
<b>5</b>	<b>Αριθμητική Ολοκλήρωση</b>	<b>75</b>
5.1	Εισαγωγή	75
5.1.1	Ολοκληρώματα με μη πεπερασμένα όρια ολοκλήρωσης	76
5.2	Κανόνας Τραπεζίου	76
5.2.1	Σφάλμα ολοκλήρωσης κανόνα τραπεζίου	77
5.2.2	Εκτεταμένος τύπος τραπεζίου	78
5.2.3	Σφάλμα ολοκλήρωσης εκτεταμένου τύπου τραπεζίου	78
5.3	Κανόνας Simpson	79
5.3.1	Σφάλμα ολοκλήρωσης κανόνα Simpson	80
5.3.2	Εκτεταμένος τύπος Simpson	81
5.3.3	Σφάλμα ολοκλήρωσης εκτεταμένου τύπου Simpson	81

5.4	Κανόνας Simpson των $3/8$	82
5.4.1	Σφάλμα ολοκλήρωσης κανόνα Simpson $3/8$	82
5.4.2	Εκτεταμένος τύπος Simpson των $3/8$	82
5.5	Άλλος υπολογισμός των Newton–Cotes	82
5.5.1	Παρατηρήσεις	83
5.6	Μέθοδοι Gauss	84
5.6.1	Μέθοδος Gauss–Legendre	84
5.6.2	Μέθοδος Gauss–Hermite	87
5.6.3	Μέθοδος Gauss–Laguerre	87
5.6.4	Μέθοδος Gauss–Chebyshev	88
5.6.5	Κατασκευή μεθόδων Gauss	88
5.7	Μέθοδος Clenshaw–Curtis	90
5.8	Ειδική Περίπτωση	91
5.8.1	Ολοκλήρωση σε άνισα τμήματα	91
5.9	Ασκήσεις	91
<b>6</b>	<b>Ανάλυση Fourier</b>	<b>97</b>
6.1	Ορισμοί	97
6.1.1	Συνεχής συνάρτηση	97
6.1.2	Περιοδική συνάρτηση	98
6.1.3	Συνθήκες Dirichlet	98
6.2	Σειρά Fourier	99
6.3	Υπολογισμός συντελεστών της σειράς Fourier	100
6.3.1	Ιδιότητες	101
6.3.2	Παράδειγμα	102
6.3.3	Συντελεστές Fourier συνάρτησης με συμμετρία	103
6.3.4	Παράδειγμα	104
6.4	Φαινόμενο Gibbs	105
6.5	Παραγωγή σειράς Fourier από άλλη	106
6.5.1	Ολοκλήρωση	106
6.5.2	Παραγωγή	107
6.6	Εναλλακτική θεώρηση της σειράς Fourier	107
6.6.1	Ταυτότητα Parseval	108
6.7	Σειρά Fourier για συναρτήσεις σε πεπερασμένο διάστημα	110
6.7.1	Μετατόπιση	110
6.7.2	Κατοπτρισμός ως προς ευθείες	111
6.7.3	Κατοπτρισμός ως προς σημεία	111
6.7.4	Παράδειγμα	112
6.8	Μιγαδική μορφή της σειράς Fourier	115
6.9	Διακριτός μετασχηματισμός Fourier (DFT)	117
6.9.1	Γρήγορος υπολογισμός του DFT — Αλγόριθμος FFT	118
6.10	Μετασχηματισμός Fourier	120
6.10.1	Ιδιότητες	122

6.10.2	Συμμετρία	122
6.11	Συνάρτηση $\delta$	124
6.11.1	Ιδιότητες	126
6.11.2	Παράγωγοι της συνάρτησης $\delta(x)$	127
6.11.3	Μετασχηματισμός Fourier της συνάρτησης $\delta(x)$	128
6.11.4	Εφαρμογές στη Φυσική	128
6.12	Συνέλιξη συναρτήσεων	128
6.12.1	Ιδιότητες συνέλιξης	129
6.12.2	Θεώρημα συνέλιξης	130
6.13	Συσχέτιση συναρτήσεων	130
6.13.1	Θεώρημα Wiener–Khinchin	131
6.13.2	Αυτοσυσχέτιση συνάρτησης	132
6.14	Ασκήσεις	132
<b>7</b>	<b>Διαφορικές Εξισώσεις</b>	<b>135</b>
7.1	Γενικά	135
7.2	Εισαγωγή	135
7.2.1	Διωνυμικό Ανάπτυγμα	136
7.3	Κατηγορίες και Λύσεις Διαφορικών Εξισώσεων	137
7.3.1	Πρωτοβάθμιες ΔΕ	137
7.3.2	Δευτεροβάθμιες ΔΕ	138
7.3.3	Σύστημα πρωτοβάθμιων ΔΕ με σταθερούς συντελεστές	139
7.4	Μέθοδος Σειράς Taylor	140
7.4.1	Μέθοδος Euler	142
7.4.2	Σφάλμα Μεθόδου Taylor	143
7.5	Μέθοδος Runge–Kutta	144
7.5.1	Μέθοδος Runge–Kutta 2ου βαθμού	145
7.5.2	Μέθοδος Runge–Kutta 4ου βαθμού	147
7.5.3	Σχόλια	147
7.6	Τελεστές Διαφορών	148
7.6.1	Ιδιότητες	149
7.6.2	Άλλοι τελεστές	150
7.6.3	Γενικευμένοι τύποι του Newton	152
7.6.4	Εφαρμογή των τελεστών στον υπολογισμό ολοκληρωμάτων	152
7.7	Πολυβηματικές Μέθοδοι	154
7.7.1	Μέθοδος Adams–Bashforth	154
7.7.2	Μέθοδος Adams–Moulton	155
7.7.3	Μέθοδοι Πρόβλεψης–Διόρθωσης (Predictor–Corrector)	156
7.8	Συστήματα Διαφορικών Εξισώσεων	159
7.9	Εξισώσεις Διαφορών	161
7.9.1	Εξίσωση διαφορών πρώτου βαθμού	162
7.9.2	Εξίσωση διαφορών δεύτερου βαθμού	163
7.9.3	Μη ομογενείς εξισώσεις διαφορών	164

7.9.4 Σχόλια . . . . .	165
7.10 Αριθμητική Ευστάθεια . . . . .	165
7.11 Απόλυτη Ευστάθεια . . . . .	170
7.12 Ασκήσεις . . . . .	171
<b>α' Χρήσιμα Ολοκληρώματα</b>	<b>177</b>
<b>Κατάλογος πινάκων</b>	<b>179</b>
<b>Ευρετήριο</b>	<b>181</b>





# Κεφάλαιο 1

## Σφάλματα

### 1.1 Εισαγωγή

Η αριθμητική ανάλυση είναι κλάδος των Μαθηματικών που ασχολείται με την επίλυση προβλημάτων που ανακύπτουν συχνά στους επιστημονικούς υπολογισμούς. Τέτοια προβλήματα είναι, μεταξύ άλλων, η εύρεση ρίζας μιας συνάρτησης, η λύση γραμμικών ή μη γραμμικών συστημάτων, ο υπολογισμός ολοκληρωμάτων κλπ. Η αριθμητική ανάλυση παρέχει αλγόριθμους για τον αριθμητικό υπολογισμό των ζητούμενων ποσοτήτων σε κάθε πρόβλημα. Η τιμή που προκύπτει, στη γενική περίπτωση, προσεγγίζει την (άγνωστη) πραγματική τιμή και μάλιστα, το εύρος της περιοχής γύρω από την προσεγγιστική τιμή στην οποία μπορεί να βρίσκεται η πραγματική τιμή υπολογίζεται, έστω και ως τάξη μεγέθους, από τον εκάστοτε αλγόριθμο. Το εύρος αυτό μπορεί να γίνει όσο μικρό επιθυμούμε αλλά δεν μπορεί γενικά να μηδενιστεί (και να ταυτιστεί η προσεγγιστική τιμή με την πραγματική) σε πεπερασμένο χρόνο. Συνεπώς, οι τιμές των ζητούμενων ποσοτήτων, όπως υπολογίζονται από τους αλγόριθμους της αριθμητικής ανάλυσης, έχουν κάποιο σφάλμα. Περισσότερα για αυτό θα αναφέρουμε στα επόμενα κεφάλαια.

Ο ηλεκτρονικός υπολογιστής είναι απαραίτητος για τη ρεαλιστική εφαρμογή των αλγόριθμων της αριθμητικής ανάλυσης. Εξαιτίας όμως του μηχανισμού αναπαράστασης των αριθμών στον υπολογιστή, ένα άλλου είδους σφάλμα εμφανίζεται στις προσεγγιστικές τιμές. Με αυτό, θα ασχοληθούμε στο παρόν κεφάλαιο: θα παρουσιάσουμε τα συστήματα αρίθμησης, πώς χρησιμοποιούνται για την αναπαράσταση ακέραιων και πραγματικών αριθμών, και πώς εφαρμόζεται το δυαδικό σύστημα στον υπολογιστή. Το πεπερασμένο μέγεθος της μνήμης του υπολογιστή είναι η «πηγή» του σφάλματος αυτού.

## 1.2 Συστήματα αρίθμησης

### 1.2.1 Αναπαράσταση ακεραίων

Ένας ακέραιος αριθμός αναπαρίσταται σε αριθμητικό σύστημα με βάση  $B$  ως μια σειρά ψηφίων

$$\pm d_n d_{n-1} \dots d_1 d_0 ,$$

με  $d_n \neq 0$ . Τα ψηφία  $d_i$  ικανοποιούν τη σχέση  $0 \leq d_i \leq B - 1$ . Αν δεν επαρκούν τα ψηφία 0–9 για τα  $d_i$ , χρησιμοποιούνται γράμματα του λατινικού αλφάβητου. Έτσι, τα ψηφία στο δεκαεξαδικό σύστημα ( $B = 16$ ) είναι τα 0–9, A–F.

Η τιμή  $K$  του ακεραίου αριθμού που δίνεται από την παραπάνω σειρά είναι

$$K = \pm \sum_{i=0}^n d_i B^i .$$

Αντίστροφα, μπορούμε να προσδιορίσουμε το ψηφίο  $d_i$  ενός ακεραίου αριθμού με απόλυτη τιμή  $|K|$  σε κάποια βάση  $B$  ως εξής: Το ψηφίο  $d_0$  είναι το

$$d_0 = |K| \bmod B$$

ενώ τα επόμενα

$$d_i = \frac{|K| - \sum_{j=0}^{i-1} d_j B^j}{B^i} \bmod B = (|K| \operatorname{div} B^i) \bmod B , \quad i = 1, 2, \dots .$$

Η τελευταία τιμή του  $i$  είναι αυτή για την οποία ισχύει

$$|K| - \sum_{j=0}^i d_j B^j = 0 .$$

Με βάση τα παραπάνω, ο ακέραιος 64206 του δεκαδικού συστήματος γράφεται στα πιο χρησιμοποιούμενα συστήματα ως εξής

- 11111010 11001110 στο δυαδικό σύστημα καθώς  $1 \times 2^1 + 1 \times 2^2 + 1 \times 2^3 + 1 \times 2^6 + 1 \times 2^7 + 1 \times 2^9 + 1 \times 2^{11} + 1 \times 2^{12} + 1 \times 2^{13} + 1 \times 2^{14} + 1 \times 2^{15} = 64206$ .
- 175316 στο οκταδικό σύστημα καθώς  $6 \times 8^0 + 1 \times 8^1 + 3 \times 8^2 + 5 \times 8^3 + 7 \times 8^4 + 1 \times 8^5 = 64206$ .
- FACE στο δεκαεξαδικό σύστημα καθώς  $14 \times 16^0 + 12 \times 16^1 + 10 \times 16^2 + 15 \times 16^3 = 64206$ .

Η πρόσθεση δύο ακέραιων αριθμών στην ίδια βάση  $B$  γίνεται με τους κανόνες που γνωρίζουμε από το δεκαδικό σύστημα. Το άθροισμα δύο ακέραιων αριθμών με ψηφία  $a_i$  και  $b_i$  στη βάση  $B$  είναι η σειρά ψηφίων  $c_i$  για τα οποία ισχύει

$$c_i = (a_i + b_i + e_i) \bmod B, \quad i \geq 0,$$

όπου  $e_i$  το κρατούμενο για το bit  $i$ . Το  $e_i$  ικανοποιεί τη σχέση

$$e_i = \begin{cases} 0, & i = 0, \\ (a_{i-1} + b_{i-1} + e_{i-1}) \operatorname{div} B, & i > 0. \end{cases}$$

### 1.2.2 Αναπαράσταση πραγματικών

Ένας πραγματικός αριθμός σε κάποια βάση  $B$  είναι μια σειρά ψηφίων που χωρίζονται με τελεία (υποδιαστολή). Π.χ. στο δεκαδικό σύστημα μπορούμε να γράψουμε τον αριθμό

$$123.4567$$

Ο παραπάνω είναι ισοδύναμος με τους

$$123.4567 \times 10^1, \quad 1.234567 \times 10^2, \quad 0.1234567 \times 10^3, \quad \text{κλπ.}$$

και τους

$$1234.567 \times 10^{-1}, \quad 12345.67 \times 10^{-2}, \quad 123456.7 \times 10^{-3}, \quad \text{κλπ.}$$

Γενικότερα, ένας πραγματικός αριθμός μπορεί να γραφεί σε βάση  $B$  στη μορφή

$$\pm d_0.d_1d_2d_3 \dots d_n \times B^e$$

με  $d_0 \neq 0$  και με ακέραιο εκθέτη  $e$ . Τα ψηφία  $d_i$ , όπως και στους ακέραιους, ικανοποιούν τη σχέση  $0 \leq d_i \leq B - 1$ . Η τιμή του αριθμού στην παραπάνω μορφή είναι

$$\pm \left( \sum_{i=0}^n d_i B^{-i} \right) \times B^e.$$

Τα ψηφία  $d_0, d_1, \dots, d_n$  αποτελούν τα *σημαντικά ψηφία* (*significant digits*) του αριθμού.

## 1.3 Αναπαράσταση αριθμών στον υπολογιστή

### 1.3.1 Ακέραιοι

Ένας ηλεκτρονικός υπολογιστής χρησιμοποιεί το δυαδικό σύστημα για την αναπαράσταση των αριθμών, ακέραιων ή πραγματικών. Για τους ακέραιους αφιερώνει συνήθως 32 bits. Έτσι, ο αριθμός π.χ. 1569 αντιπροσωπεύεται από τη σειρά

$$00000000 \ 00000000 \ 00000110 \ 00100001$$

καθώς  $1 \times 2^0 + 1 \times 2^5 + 1 \times 2^9 + 1 \times 2^{10} = 1569$ .

Οι αρνητικοί αριθμοί αναπαριστώνται συνήθως ως εξής: αν  $K$  είναι θετικός αριθμός, ο αριθμός  $-K$  είναι αυτός που ικανοποιεί τη σχέση

$$K + (-K) = 0,$$

δηλαδή είναι ο αριθμός που αν προστεθεί στον  $K$  δίνει αποτέλεσμα 0. Στην πρόσθεση κρατάμε μόνο τα πρώτα 32 bits. Έτσι, ο αριθμός  $-1569$  είναι αυτός που αν προστεθεί στο 1569 δίνει 0, η σειρά, δηλαδή,

$$11111111 \ 11111111 \ 11111001 \ 11011111.$$

Μπορούμε εύκολα να βρούμε τον αντίθετο ενός αριθμού στο δυαδικό σύστημα με συγκεκριμένο πλήθος bits αν εντοπίσουμε το δεξιότερο 1 και αντιστρέψουμε όλα τα bits στα αριστερά του (δηλαδή, μετατρέψουμε τα 1 σε 0 και τα 0 σε 1).

Ο μεγαλύτερος ακέραιος σε 32 bits είναι ο

$$01111111 \ 11111111 \ 11111111 \ 11111111$$

δηλαδή, ο 2147483647 του δεκαδικού, ενώ ο μικρότερος είναι ο

$$10000000 \ 00000000 \ 00000000 \ 00000000$$

δηλαδή ο  $-2147483648$ . Προσέξτε ότι ο αντίθετος του συγκεκριμένου αριθμού είναι ο ίδιος αριθμός, καθώς δεν μπορούμε να αναπαραστήσουμε τον 2147483648 σε 32 bits. Επίσης, προσέξτε ότι ο αριθμός

$$11111111 \ 11111111 \ 11111111 \ 11111111$$

που θα περίμενε κανείς να είναι ο μέγιστος που μπορεί να αναπαρασταθεί, έχει αντίθετο τον

$$00000000 \ 00000000 \ 00000000 \ 00000001$$

δηλαδή, είναι ο  $-1$ .

**Παρατήρηση:** Καθώς υπάρχει όριο στους ακέραιους αριθμούς που μπορούν να αναπαρασταθούν σε ένα πρόγραμμα για ηλεκτρονικό υπολογιστή, πρέπει να είμαστε ιδιαίτερα προσεκτικοί όταν χρησιμοποιούμε ακραίους που μπορούν να λάβουν μεγάλη τιμή. Π.χ. το παραγοντικό ακέραιου μεγαλύτερου από το 12 ή το πλήθος των στοιχείων τετραγωνικού πίνακα με διαστάσεις πάνω από  $46340 \times 46340$  δεν είναι αναπαραστήσιμο σε ακέραιο των 32 bits.

### 1.3.2 Πραγματικοί

Στο δυαδικό σύστημα ο πραγματικός αριθμός έχει τη μορφή

$$\pm 1.d_1d_2d_3 \dots d_n \times 2^e$$

και αποθηκεύεται σε 32 bits (για απλή ακρίβεια) ή σε 64 bits (για διπλή ακρίβεια), ως εξής, σύμφωνα με το πρότυπο IEEE 754:

- Το πρώτο bit αναπαριστά το πρόσημο του αριθμού: αν είναι 0 το πρόσημο είναι + και αν είναι 1 το πρόσημο είναι −.
- Τα επόμενα 8 (σε απλή ακρίβεια) ή 11 bits (σε διπλή ακρίβεια) αποθηκεύουν τον ακέραιο εκθέτη,  $e$ , αφού του προσθέσουν τον αριθμό  $2^{8-1} - 1 = 127$  (σε απλή ακρίβεια) ή  $2^{11-1} - 1 = 1023$  (σε διπλή ακρίβεια). Με αυτό τον τρόπο, οι εκθέτες είναι πάντα θετικοί και δεν χρειάζεται bit για το πρόσημό τους.
- Στα τελευταία 23 (σε απλή ακρίβεια) ή 52 bits (σε διπλή ακρίβεια) αποθηκεύονται ισάριθμα δυαδικά ψηφία  $d_1, d_2, \dots$ . Το  $d_0$  που είναι πάντα 1 δεν αποθηκεύεται.

Στην παραπάνω μορφή, η αναπαράσταση πραγματικών αριθμών δεν είναι πάντα δυνατή με απόλυτη ακρίβεια λόγω του πεπερασμένου αριθμού bits. Λόγω της αναγκαστικής αποκοπής ή της στρογγύλευσης των bits μετά το 23ο ή 52ο, έχουμε σφάλμα στην αναπαράσταση των περισσότερων πραγματικών αριθμών. Το σφάλμα αυτό έχει μέγιστη τιμή  $\varepsilon = 2^{-23} \approx 1.19 \times 10^{-7}$  (για απλή ακρίβεια) ή  $\varepsilon = 2^{-52} \approx 2.22 \times 10^{-16}$  (για διπλή ακρίβεια). Η μέγιστη τιμή αποκαλείται *έψιλον της μηχανής*. Για κάθε πραγματικό  $x$  με  $|x| < \varepsilon$  ισχύει  $1 + x = 1$ . Αυτό σημαίνει ότι υπάρχει ένα όριο κάτω από το οποίο οι αριθμοί συμπεριφέρονται ως μηδέν σε προσθέσεις ή αφαιρέσεις με αριθμούς της τάξης του 1.

Συγκεκριμένες σειρές των 32 ή 64 bits αντιστοιχούν στο  $\pm\infty$  και στο NaN (Not A Number). Αν τις εξαιρέσουμε, ο μεγαλύτερος εκθέτης που μπορεί να αποθηκευτεί σε απλή ακρίβεια αντιστοιχεί στο δυαδικό αριθμό

$$11111110.$$

Ο συγκεκριμένος αριθμός είναι ο 254 στο δεκαδικό, άρα ο εκθέτης είναι  $254 - 127 = 127$ . Ο μικρότερος εκθέτης σε απλή ακρίβεια αντιστοιχεί στο δυαδικό αριθμό

$$00000001$$

και είναι ο  $1 - 127 = -126$ . Επομένως, ο μεγαλύτερος πραγματικός αριθμός που μπορεί να αποθηκευτεί σε απλή ακρίβεια είναι της τάξης του  $2^{127} \approx 10^{38}$  ενώ ο κατά μέτρο μικρότερος είναι της τάξης του  $2^{126} \approx 10^{-38}$ .

Ανάλογα, ο μεγαλύτερος εκθέτης που μπορεί να αποθηκευτεί σε διπλή ακρίβεια αντιστοιχεί στο δυαδικό αριθμό

$$111\ 11111110.$$

Ο συγκεκριμένος αριθμός είναι ο 2046 στο δεκαδικό, άρα ο εκθέτης είναι  $2046 - 1023 = 1023$ . Ο μικρότερος εκθέτης σε διπλή ακρίβεια αντιστοιχεί στο δυαδικό αριθμό

$$000\ 00000001$$

και είναι ο  $1 - 1023 = -1022$ . Επομένως, ο μεγαλύτερος πραγματικός αριθμός που μπορεί να αποθηκευτεί σε διπλή ακρίβεια είναι της τάξης του  $2^{1023} \approx 10^{308}$  ενώ ο κατά μέτρο μικρότερος είναι της τάξης του  $2^{-1022} \approx 10^{-308}$ .

**Παρατηρήσεις:**

- Αν ένα πραγματικό αποτέλεσμα πράξης υπερβαίνει κατ' απόλυτη τιμή το μέγιστο αναπαραστάσιμο στον Η/Υ αριθμό, έχουμε *υπερχείλιση (overflow)*. Αντίστοιχα, αν είναι κατ' απόλυτη τιμή μικρότερο από το μικρότερο αναπαραστάσιμο στον Η/Υ αριθμό, τότε έχουμε *υπεκχείλιση (underflow)*. Η τιμή που θα αποκτήσει το αποτέλεσμα και στις δύο περιπτώσεις είναι απροσδιόριστη, ο υπολογισμός όμως μπορεί να συνεχίσει με σχεδόν σίγουρα λάθος αποτελέσματα. Σε υπολογιστές που υλοποιούν το πρότυπο αναπαράστασης αριθμών IEEE οι τιμές είναι αντίστοιχα  $\pm\infty$  (το πλησιέστερο «άπειρο») και  $\pm 0$ .
- Λόγω της πεπερασμένης ακρίβειας αναπαράστασης, το αποτέλεσμα της πράξης μεταξύ πραγματικών  $x + (y + z)$  μπορεί να είναι διαφορετικό από το  $(x + y) + z$ , π.χ. αν το  $x$  είναι πολύ μεγαλύτερο κατά μέτρο από τα  $y, z$ . Επίσης, η τιμή  $x + y + z$  μπορεί να είναι διαφορετική από την  $x + z + y$ . Η πρόσθεση (και ο πολλαπλασιασμός) μεταξύ πραγματικών αριθμών στον υπολογιστή δεν έχει την αντιμεταθετική και την προσεταιριστική ιδιότητα, όπως αναμένουμε από τα μαθηματικά.
- Η σύγκριση για ισότητα δύο πραγματικών αριθμών που προέκυψαν από πράξεις πρέπει να αποφεύγεται.

**1.4 Ασκήσεις**

1. Να γράψετε κώδικα που να μετατρέπει ένα μη αρνητικό ακέραιο αριθμό από το δεκαδικό στο δυαδικό σύστημα.
2. Γράψτε πρόγραμμα που να τυπώνει στην οθόνη στη μορφή  $\pm x.xxxx E \pm yy$  (εκθετική μορφή) τα αποτελέσματα των εκφράσεων  $0.1 + 0.2 - 0.3$  και  $0.1 - 0.3 + 0.2$ . Είναι μηδέν; είναι έστω ίσα;
3. Υπολογίστε με πρόγραμμα το άθροισμα των αριθμών 0.1, 0.2,...,1.9.  
[Απάντηση: 19]
4. Υπολογίστε το έψιλον της μηχανής για πραγματικούς αριθμούς απλής και διπλής ακρίβειας με τους εξής τρόπους:
  - (α') Εφαρμόστε τον αλγόριθμο:  
Θέτουμε  $\varepsilon \leftarrow 1$ . Για όσο ισχύει  $1 + \varepsilon \neq 1$  θέτουμε  $\varepsilon \leftarrow \varepsilon/2$  και επαναλαμβάνουμε.
  - (β') Χρησιμοποιήστε κατάλληλες συναρτήσεις/τιμές που παρέχει η γλώσσα προγραμματισμού που χρησιμοποιείτε (FLT\_EPSILON και DBL\_EPSILON στη C, EPSILON() στη Fortran 90, epsilon() από το std::numeric\_limits<> της C++, κλπ.).

(γ') Καλέστε τις ρουτίνες SLAMCH() και DLAMCH() της συλλογής ρουτινών LAPACK (που υπολογίζουν το μισό του έψιλον μηχανής, σύμφωνα με τον ορισμό που έχουμε χρησιμοποιήσει).

5. Οι ρίζες του τριωνύμου  $ax^2 + bx + c$  δίνονται ως

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

όταν  $a \neq 0$ .

Έστω  $a = 1$ ,  $b = 3000.001$ ,  $c = 3$ .

(α') Υπολογίστε τα  $x_{1,2}$  με απλή και διπλή ακρίβεια. Συγκρίνετέ τα με τις ακριβείς ρίζες ( $x_1 = -0.001$ ,  $x_2 = -3000.0$ ).

(β') Επαναλάβετε τους υπολογισμούς του προηγούμενου σκέλους εφαρμόζοντας τον αλγεβρικά ισοδύναμο τύπο

$$x_{1,2} = \frac{2c}{-b \mp \sqrt{b^2 - 4ac}}.$$

Τι παρατηρείτε ως προς την ακρίβεια των υπολογισμών σας;

6. Γράψτε κώδικα ώστε να υπολογίσετε την τιμή του  $e^1$  εφαρμόζοντας τη σχέση

$$e^x = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n.$$

Βρείτε και τυπώστε, δηλαδή, την τιμή του  $(1 + 1/n)^n$  για  $n = 1, 2, 3, \dots$ . Τι παρατηρείτε ως προς την ταχύτητα σύγκλισης στην πραγματική τιμή του (2.718281828459045...);

7. Γράψτε κώδικα που να υπολογίζει το  $e^x$  εφαρμόζοντας τη σχέση

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}.$$

Για τη διευκόλυνσή σας παρατηρήστε ότι ο  $n$ οστός όρος στο άθροισμα προκύπτει από τον αμέσως προηγούμενο αν αυτός πολλαπλασιαστεί με το  $x/n$ .

Στην πρόσθεση κρατήστε όσους όρους έχουν συνεισφορά (δηλαδή μεταβάλλουν το άθροισμα).

8. Γράψτε κώδικα που να υπολογίζει το  $\sin x$  εφαρμόζοντας τη σχέση

$$\sin x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!}.$$

Για τη διευκόλυνσή σας παρατηρήστε ότι ο  $k$  όρος στο άθροισμα προκύπτει από τον αμέσως προηγούμενο αν αυτός πολλαπλασιαστεί με το  $-\frac{x^2}{2k(2k+1)}$ .

Στην πρόσθεση κρατήστε όσους όρους έχουν συνεισφορά (δηλαδή μεταβάλλουν το άθροισμα).

9. Γράψτε κώδικα που να υπολογίζει το  $\cos x$  εφαρμόζοντας τη σχέση

$$\cos x = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!} .$$

Στην πρόσθεση κρατήστε όσους όρους έχουν συνεισφορά (δηλαδή μεταβάλλουν το άθροισμα).



## Κεφάλαιο 2

# Επίλυση μη Γραμμικών Εξισώσεων

### 2.1 Εισαγωγή

Στο κεφάλαιο αυτό θα παρουσιάσουμε κάποιους αλγόριθμους (μεθόδους) εύρεσης των λύσεων μιας εξίσωσης με ένα άγνωστο. Η εξίσωση έχει γενικά τη μορφή

$$f(x) = 0, \quad x \in \mathcal{R}. \quad (2.1)$$

Οι λύσεις της, τα συγκεκριμένα σημεία  $x$  που την ικανοποιούν, λέγονται και *ρίζες* της συνάρτησης  $f(x)$ . Ορισμένοι αλγόριθμοι από αυτούς που θα παρουσιάσουμε μπορούν να υπολογίσουν και μιγαδικές ρίζες.

Στην εξίσωση  $f(x) = 0$  ανάγεται εύκολα η εξίσωση  $g(x) = c$  με  $c \neq 0$  ή γενικότερα η  $g(x) = h(x)$  με την επιλογή  $f(x) = g(x) - c$  ή  $f(x) = g(x) - h(x)$ . Επομένως, εκτός από την εύρεση ρίζας, οι αλγόριθμοι που θα παρουσιάσουμε εφαρμόζονται για την εύρεση τιμής της αντίστροφης συνάρτησης  $g^{-1}(c)$  ή σημείου τομής συναρτήσεων.

Στην περίπτωση που η συνάρτηση  $f(x)$  είναι γραμμική (δηλαδή, της μορφής  $f(x) = ax + b$ ) η εύρεση της ρίζας είναι τετριμμένη. Οι δυσκολίες εμφανίζονται στην αντίθετη περίπτωση και γι' αυτό θα επικεντρωθούμε στην επίλυση μη γραμμικών εξισώσεων. Όταν η  $f(x)$  είναι γενικό πολυώνυμο μέχρι και 4<sup>ου</sup> βαθμού, υπάρχουν αναλυτικοί τύποι που υπολογίζουν τις ρίζες της. Ήδη, όμως, από τον 3<sup>ο</sup> βαθμό είναι αρκετά δύσκολοι. Στη γενική περίπτωση που δεν είναι πολυώνυμο, η εύρεση των ριζών (ή και η απόδειξη της ύπαρξής τους) γενικά δεν είναι δυνατή με αναλυτικούς τύπους.

Η επίλυση με αριθμητικές μεθόδους της εξίσωσης (2.1) βασίζεται στην εύρεση μιας ακολουθίας τιμών  $x_0, x_1, \dots, x_n$ , που συγκλίνουν για  $n \rightarrow \infty$  σε μία ρίζα της εξίσωσης,  $\bar{x}$ . Κάθε μία από τις μεθόδους που θα δούμε, παράγει τέτοια ακολουθία με συγκεκριμένη διαδικασία και υπό ορισμένες προϋποθέσεις. Επιπλέον, σε κάθε επανάληψη, μας δίνει μία εκτίμηση του εύρους της περιοχής στην οποία βρίσκεται η ρίζα γύρω από το  $x_n$ : η μέθοδος παράγει μία ακολουθία  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$  για την

ακρίβεια· ισχύει  $x_n - \varepsilon_n \leq \bar{x} \leq x_n + \varepsilon_n$ , με  $\varepsilon_n < \varepsilon_{n-1}$ .

Στην πράξη, η διαδικασία που παράγει τις διαδοχικές προσεγγίσεις της ρίζας δεν επαναλαμβάνεται επ' άπειρον αλλά διακόπτεται όταν φτάσουμε στην «κατάλληλη» προσέγγιση της ρίζας. «Κατάλληλη» θεωρείται η προσέγγιση  $x_k$  όταν ικανοποιούνται μία ή περισσότερες από τις ακόλουθες γενικές συνθήκες (με  $\varepsilon$  συμβολίζουμε την επιθυμητή ακρίβεια)

- Η ακρίβεια της μεθόδου  $\varepsilon_k$  είναι μικρότερη από την επιθυμητή.
- Η απόλυτη τιμή της συνάρτησης να είναι «μικρή»:  $|f(x_k)| < \varepsilon$ . Αν η συνάρτηση είναι συνεχής, από το  $f(x_k) \approx f(\bar{x}) = 0$  προκύπτει ότι  $x_k \approx \bar{x}$ .
- Η απόλυτη βελτίωση να είναι «μικρή»:  $|x_k - x_{k-1}| < \varepsilon$ .
- Η σχετική βελτίωση να είναι «μικρή»:  $\left| \frac{x_k - x_{k-1}}{x_k} \right| < \varepsilon$  αν  $x_k \neq 0$ .

Στις δύο τελευταίες συνθήκες πρέπει να ελέγχουμε αν τελικά η τιμή  $x_k$  ικανοποιεί την  $f(x_k) \approx 0$ .

### 2.1.1 Ταχύτητα σύγκλισης

Μια μέθοδος επίλυσης της εξίσωσης  $f(x) = 0$  παράγει την ακολουθία προσεγγιστικών λύσεων  $x_0, x_1, \dots, x_k \dots$ , η οποία συγκλίνει στη ρίζα  $\bar{x}$  με μέγιστη ακρίβεια  $\varepsilon_k \equiv |x_k - \bar{x}|$ . Η μέθοδος χαρακτηρίζεται ως  $\alpha$  τάξης όσον αφορά στη σύγκλιση, αν υπάρχουν  $\alpha, \lambda > 0$  ώστε

$$\lim_{n \rightarrow \infty} \frac{|x_{n+1} - \bar{x}|}{|x_n - \bar{x}|^\alpha} \equiv \lim_{n \rightarrow \infty} \frac{\varepsilon_{n+1}}{\varepsilon_n^\alpha} = \lambda. \quad (2.2)$$

Ο αριθμός  $\lambda$  αποτελεί την *ταχύτητα (ή ρυθμό) σύγκλισης*.

### 2.1.2 Ευστάθεια

Όπως θα δούμε, οι περισσότερες μέθοδοι εύρεσης ρίζας χρειάζονται μια αρχική προσέγγιση της λύσης (ή και περισσότερες), την οποία βελτιώνουν σε κάθε στάδιο της επίλυσης. Η αριθμητική τους ευστάθεια προσδιορίζεται από τη συμπεριφορά τους σε μεταβολές αυτής της αρχικής τιμής. Μια μέθοδος είναι *ευσταθής* αν οποιαδήποτε κατάλληλα μικρή μεταβολή της αρχικής τιμής δεν επηρεάζει την εύρεση της ρίζας, ενώ είναι *ασταθής* αν μια μικρή μεταβολή της αρχικής προσέγγισης οδηγεί μακριά από τη ρίζα.

Γενικά, όσο υψηλότερη είναι η τάξη σύγκλισης μίας μεθόδου, τόσο λιγότερο ευσταθής είναι αυτή.

### 2.1.3 Εύρεση περισσότερων της μίας ριζών

Αν επιθυμούμε να εντοπίσουμε πολλές ρίζες μίας συνάρτησης  $f(x)$  μπορούμε να εφαρμόσουμε τη μέθοδο της επιλογής μας με διαφορετικές αρχικές προσεγγίσεις, ελπίζοντας ότι θα καταλήξουμε σε διαφορετικές ρίζες. Μια συστηματική αντιμετώπιση του προβλήματος βασίζεται στην ακόλουθη παρατήρηση: αν η συνάρτηση  $f(x)$  έχει ρίζα το  $\bar{x}$  με πολλαπλότητα  $m$  (δηλαδή, ισχύει ότι  $f'(\bar{x}) = f''(\bar{x}) = \dots = f^{(m-1)}(\bar{x}) = 0$ ), τότε η συνάρτηση  $g(x) = f(x)/(x - \bar{x})^m$  έχει ως ρίζες της όλες τις ρίζες της  $f(x)$  εκτός από το  $\bar{x}$ . Επομένως, εφαρμόζουμε μία μέθοδο εύρεσης ρίζας της επιλογής μας για να υπολογίσουμε μία ρίζα,  $x_1$ . Κατόπιν, αναζητούμε τη ρίζα της  $g_1(x) = f(x)/(x - x_1)$  ώστε να βρούμε άλλη ρίζα  $x_2$ . Στο επόμενο στάδιο σχηματίζουμε την  $g_2(x) = g_1(x)/(x - x_2)$  και προσπαθούμε να την μηδενίσουμε. Η διαδικασία αυτή επαναλαμβάνεται έως ότου βρούμε όσες ρίζες αναζητούμε.

### 2.1.4 Χρήσιμα θεωρήματα

**Θεώρημα Ενδιάμεσης Τιμής (ΘΕΤ)** Έστω  $f(x)$  συνεχής συνάρτηση στο κλειστό διάστημα  $[a, b]$ . Αν  $\lambda$  είναι ένας οποιοσδήποτε πραγματικός αριθμός μεταξύ των  $f(a), f(b)$  (συμπεριλαμβανομένων και αυτών), τότε υπάρχει ένα τουλάχιστον  $c \in [a, b]$  ώστε  $f(c) = \lambda$ .

**Θεώρημα Μέσης Τιμής** Έστω  $f(x)$  συνεχής συνάρτηση για  $x \in [a, b]$ , διαφορίσιμη στο  $(a, b)$ , με παράγωγο  $f'(x)$ . Τότε υπάρχει ένα τουλάχιστον  $c \in [a, b]$  ώστε  $f(b) - f(a) = f'(c)(b - a)$ . Αν επιπλέον ισχύει  $f(a) = f(b)$  τότε σε κάποιο  $c \in [a, b]$  έχουμε  $f'(c) = 0$  (**Θεώρημα Rolle**).

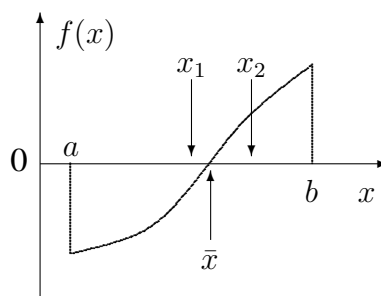
**Θεώρημα Taylor** Έστω ότι η συνάρτηση  $f(x)$ ,  $x \in [a, b]$ , έχει παράγωγο τάξης  $n+1$  και η  $f^{n+1}(x)$  είναι συνεχής στο  $[a, b]$ . Αν  $x, x_0 \in [a, b]$ ,  $x \neq x_0$ , τότε υπάρχει  $\xi$  μεταξύ των  $x_0, x$  ώστε

$$\begin{aligned} f(x) = & f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \dots \\ & + \frac{f^n(x_0)}{n!}(x - x_0)^n + \frac{f^{n+1}(\xi)}{(n+1)!}(x - x_0)^{n+1}. \end{aligned} \quad (2.3)$$

## 2.2 Μέθοδος Διχοτόμησης

Η μέθοδος βασίζεται στο Θεώρημα Ενδιάμεσης Τιμής. Αν η  $f(x)$  είναι συνεχής στο  $[a, b]$  και έχουμε  $f(a)f(b) < 0$ , τότε, από το θεώρημα, υπάρχει  $c = \bar{x} \in (a, b)$  ώστε  $f(\bar{x}) = 0$ . Άρα υπάρχει τουλάχιστον μία ρίζα της  $f(x)$  στο  $(a, b)$ . Το συμπέρασμα αυτό αποτελεί το **θεώρημα Bolzano**.

Η διαδικασία που ακολουθεί η μέθοδος διχοτομεί το διάστημα  $[a, b]$ , εντοπίζει τη ρίζα σε ένα από τα δύο υποδιαστήματα και επαναλαμβάνεται στο επιλεγμένο υποδιάστημα. Παράγεται έτσι μια ακολουθία διαστημάτων  $[a_1, b_1], [a_2, b_2], \dots, [a_n, b_n]$



Σχήμα 2.1: Σχηματική αναπαράσταση της Μεθόδου Διχοτόμησης για την εύρεση ρίζας

και μια ακολουθία προσεγγίσεων της ρίζας  $x_1 = (a_1 + b_1)/2$ ,  $x_2 = (a_2 + b_2)/2, \dots, x_n = (a_n + b_n)/2$ . Όπως θα αναφέρουμε παρακάτω, η περιοχή γύρω από το  $x_n$  στην οποία υπάρχει η αναζητούμενη ρίζα έχει εύρος  $2\varepsilon_n = |b_n - a_n|/2^{n-1}$ .

**Αλγόριθμος:** Επίλυση της  $f(x) = 0$  με τη μέθοδο διχοτόμησης:

1. Επιλέγουμε δύο τιμές  $a, b$ , με  $a < b$  έτσι ώστε η  $f(x)$  να είναι συνεχής στο  $[a, b]$  και να ισχύει  $f(a)f(b) < 0$ .
2. Θέτουμε  $x \leftarrow \frac{a+b}{2}$ .
3. Αν το  $x$  είναι ικανοποιητική προσέγγιση της ρίζας πηγαίνουμε στο βήμα 6.
4. Αν ισχύει ότι  $f(a)f(x) > 0$  τότε θέτουμε  $a \leftarrow x$ . Αλλιώς, θέτουμε  $b \leftarrow x$ .
5. Επαναλαμβάνουμε τη διαδικασία από το βήμα 2.
6. Τέλος.

Παρατηρήστε ότι σε κάθε επανάληψη χρειαζόμαστε ένα νέο υπολογισμό της τιμής της συνάρτησης.

### Παράδειγμα

Έστω η συνάρτηση  $f(x) = x^3 + 4x^2 - 10$ , η οποία είναι συνεχής σε όλο το διάστημα ορισμού της,  $(-\infty, \infty)$ . Παρατηρούμε ότι  $f(1) = -5$  και  $f(2) = 14$ , δηλαδή  $f(1)f(2) < 0$ . Επομένως, υπάρχει μία τουλάχιστον ρίζα της στο  $[1, 2]$ . Παρατηρούμε ακόμα ότι  $f'(x) = 3x^2 + 8x > 0$  για κάθε  $x$  στο συγκεκριμένο διάστημα. Επομένως, η  $f(x)$  είναι αύξουσα σε αυτό και άρα έχει μοναδική ρίζα στο  $[1, 2]$ . Εφαρμόζουμε τη μέθοδο διχοτόμησης για την εύρεσή της και προκύπτουν οι ακολουθίες του Πίνακα 2.1.

Μετά από 20 επαναλήψεις ισχύει για την ακρίβεια  $|x_{20} - \bar{x}| \leq 0.5 |b_{20} - a_{20}| \approx 0.95 \times 10^{-6}$ , άρα έχουμε προσδιορίσει σωστά τουλάχιστον μέχρι και το 6 δεκαδικό

$n$	$a_n$	$b_n$	$x_n \equiv (a_n + b_n)/2$	$f(x_n)$
1	1.00000000	2.00000000	1.50000000	2.3750
2	1.00000000	1.50000000	1.25000000	-1.7969
3	1.25000000	1.50000000	1.37500000	0.16211
4	1.25000000	1.37500000	1.31250000	-0.84839
5	1.31250000	1.37500000	1.34375000	-0.35098
6	1.34375000	1.37500000	1.35937500	$-0.96409 \times 10^{-1}$
7	1.35937500	1.37500000	1.36718750	$0.32356 \times 10^{-1}$
8	1.35937500	1.36718750	1.36328125	$-0.32150 \times 10^{-1}$
9	1.36328125	1.36718750	1.36523438	$0.72025 \times 10^{-4}$
10	1.36328125	1.36523438	1.36425781	$-0.16047 \times 10^{-1}$
11	1.36425781	1.36523438	1.36474609	$-0.79893 \times 10^{-2}$
12	1.36474609	1.36523438	1.36499023	$-0.39591 \times 10^{-2}$
13	1.36499023	1.36523438	1.36511230	$-0.19437 \times 10^{-2}$
14	1.36511230	1.36523438	1.36517334	$-0.93585 \times 10^{-3}$
15	1.36517334	1.36523438	1.36520386	$-0.43192 \times 10^{-3}$
16	1.36520386	1.36523438	1.36521912	$-0.17995 \times 10^{-3}$
17	1.36521912	1.36523438	1.36522675	$-0.53963 \times 10^{-4}$
18	1.36522675	1.36523438	1.36523056	$0.90310 \times 10^{-5}$
19	1.36522675	1.36523056	1.36522865	$-0.22466 \times 10^{-4}$
20	1.36522865	1.36523056	1.36522961	$-0.67174 \times 10^{-5}$

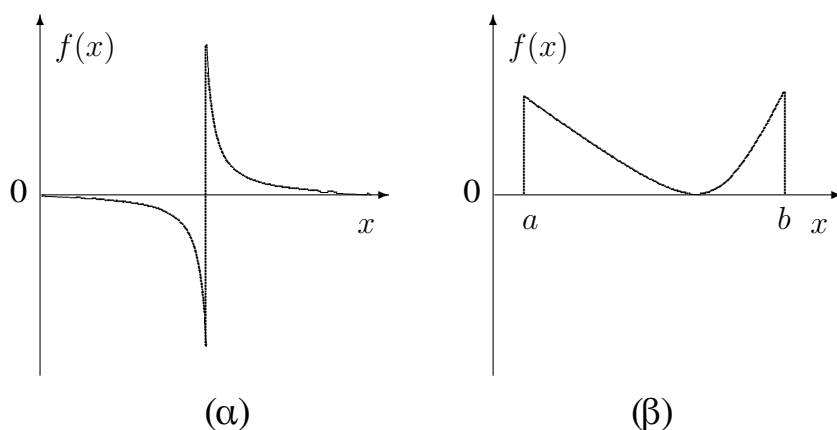
Πίνακας 2.1: Ακολουθίες των διαστημάτων, της προσεγγιστικής ρίζας και της αντίστοιχης τιμής της  $f(x) = x^3 + 4x^2 - 10$  κατά την εφαρμογή της μεθόδου διχοτόμησης

ψηφίο της ρίζας. Η προσεγγιστική τιμή, στρογγυλεμένη στα 6 δεκαδικά είναι 1.365230 ενώ η ακριβής είναι 1.36523001361638....

**Παρατήρηση:** Η μέθοδος διχοτόμησης αποτυγχάνει όταν δεν πληρούνται οι προϋποθέσεις του Θεωρήματος Ενδιάμεσης Τιμής. Π.χ. όταν η συνάρτηση δεν είναι συνεχής, Σχήμα 2.2α, η μέθοδος εντοπίζει για ρίζα το σημείο ασυνέχειας. Αντίστροφα, αν δεν μπορούμε να εντοπίσουμε δύο σημεία στα οποία η συνάρτηση έχει ετερόσημες τιμές, δε σημαίνει ότι δεν έχει ρίζα (Σχήμα 2.2β).

### 2.2.1 Ακρίβεια αλγορίθμου διχοτόμησης

Η μέθοδος διχοτόμησης για την εύρεση της ρίζας,  $\bar{x}$ , της  $f(x)$ , παράγει μια ακολουθία  $x_1, x_2, \dots$  με την ιδιότητα  $|x_n - \bar{x}| \leq \frac{1}{2^n}(b - a)$ ,  $n \geq 1$ .



Σχήμα 2.2: Σχηματικές αναπαραστάσεις συναρτήσεων για τις οποίες η μέθοδος διχοτόμησης (α) εντοπίζει μη υπαρκτή ρίζα, (β) αποτυγχάνει να εντοπίσει ρίζα στο προσδιοριζόμενο διάστημα

**Απόδειξη:**

$$\begin{aligned}
 \Theta\text{ΕΤ} \Rightarrow \quad b_1 - a_1 &= b - a, & \bar{x} &\in (a_1, b_1) \\
 b_2 - a_2 &= \frac{1}{2}(b_1 - a_1) = \frac{1}{2}(b - a), & \bar{x} &\in (a_2, b_2) \\
 b_3 - a_3 &= \frac{1}{2}(b_2 - a_2) = \frac{1}{2^2}(b - a), & \bar{x} &\in (a_3, b_3) \\
 \vdots & & \vdots & \\
 b_n - a_n &= \frac{1}{2^{n-1}}(b - a), & \bar{x} &\in (a_n, b_n)
 \end{aligned}$$

Καθώς  $x_n = \frac{1}{2}(a_n + b_n)$  και είτε  $x_n \leq \bar{x} \leq b_n$  είτε  $a_n \leq \bar{x} \leq x_n$ , έχουμε:

$$|\bar{x} - x_n| = \left| \bar{x} - \frac{1}{2}(a_n + b_n) \right| \leq \frac{1}{2}(b_n - a_n) = \frac{1}{2^n}(b - a).$$

Επομένως,  $\lim_{n \rightarrow \infty} x_n = \bar{x}$  καθώς  $\lim_{n \rightarrow \infty} \frac{1}{2^n}(b - a) = 0$ . Συμπεραίνουμε ότι με τον συγκεκριμένο αλγόριθμο, οι τιμές  $x_n$  είναι διαδοχικές προσεγγίσεις της ρίζας,  $\bar{x}$ . Σε άπειρες επαναλήψεις καταλήγουν σε αυτή.

### 2.2.2 Σύγκλιση αλγορίθμου διχοτόμησης

Για την ακρίβεια  $\varepsilon_n \equiv |x_n - \bar{x}|$  της μεθόδου έχουμε

$$\varepsilon_{n+1} = \frac{b - a}{2^{n+1}} = \frac{1}{2}\varepsilon_n.$$

Επομένως στον τύπο (2.2) έχουμε  $\alpha = 1$  και  $\lambda = 0.5$ , δηλαδή η σύγκλιση είναι πρώτης τάξης και αρκετά αργή.

### 2.2.3 Αριθμός επαναλήψεων αλγορίθμου διχοτόμησης

Ο αριθμός απαιτούμενων επαναλήψεων της μεθόδου διχοτόμησης για να επιτύχουμε μια συγκεκριμένη ακρίβεια  $\varepsilon$  (ή λιγότερο) προκύπτει ως εξής

$$\varepsilon_n \leq \varepsilon \Rightarrow \frac{b-a}{2^n} \leq \varepsilon \Rightarrow 2^n \geq \frac{b-a}{\varepsilon} \Rightarrow n \geq \log_2 \left( \frac{b-a}{\varepsilon} \right).$$

#### Παράδειγμα

Έστω η συνάρτηση  $f(x) = x^3 + 4x^2 - 10$ , συνεχής με μία ρίζα στο  $[1, 2]$ . Ο αριθμός απαιτούμενων επαναλήψεων της μεθόδου διχοτόμησης ώστε  $|x_n - \bar{x}| \leq \varepsilon = 10^{-5}$  είναι

$$n \geq \log_2 \left( \frac{2-1}{10^{-5}} \right) = \log_2 10^5 = 5 \log_2 10 \approx 16.61.$$

Επομένως, αρκούν 17 επαναλήψεις για να έχουμε  $|x_n - \bar{x}| \leq 10^{-5}$ .

## 2.3 Μέθοδος ψευδούς σημείου

Παρά το γεγονός ότι η μέθοδος διχοτόμησης είναι μια απολύτως αποδεκτή μέθοδος για τον προσδιορισμό των ριζών συναρτήσεων μιας μεταβλητής, η μέθοδος είναι σχετικά αναποτελεσματική. Ένα μειονέκτημα της μεθόδου διχοτόμησης είναι ότι με τον χωρισμό του διαστήματος από  $a$  σε  $b$  σε ίσα μισά, δε λαμβάνεται υπόψη η πληροφορία για το μέγεθος των  $f(a)$  και  $f(b)$ .

Η μέθοδος ψευδούς σημείου είναι μια τροποποίηση της μεθόδου διχοτόμησης ώστε η νέα προσέγγιση της ρίζας να εξαρτάται από τις τιμές των  $f(a)$  και  $f(b)$ . Στη νέα μέθοδο υπολογίζουμε την ευθεία που περνά από τα σημεία  $(a, f(a))$  και  $(b, f(b))$  σε κάθε επανάληψη, και ως νέα προσέγγιση ορίζουμε την τομή αυτής με τον άξονα των  $x$  (αντί για το μέσο του  $[a, b]$  της μεθόδου διχοτόμησης). Εύκολα μπορεί ναδειχθεί ότι η ευθεία είναι η

$$y = f(a) + \frac{f(a) - f(b)}{a - b}(x - a).$$

Επομένως,

$$x = a - \frac{f(a)}{f(a) - f(b)}(a - b) = \frac{bf(a) - af(b)}{f(a) - f(b)}.$$

Όπως και στη μέθοδο διχοτόμησης, μετακινούμε σε κάθε επανάληψη το ένα από τα δύο άκρα στο  $x$  ώστε η ρίζα να περικλείεται πάντα. Προσέξτε όμως ότι σε αυτή τη μέθοδο, το μήκος των διαδοχικών διαστημάτων  $[a, b]$  δεν είναι απαραίτητο να τείνει στο 0.

Η μέθοδος ψευδούς σημείου είναι γενικά πιο γρήγορη από τη μέθοδο διχοτόμησης· έχει τάξη σύγκλισης  $\alpha > 1$ . Υπάρχουν όμως περιπτώσεις συναρτήσεων που

η σύγκλιση σε ρίζα τους με αυτή τη μέθοδο είναι γραμμική ή και πιο αργή από τη μέθοδο διχοτόμησης.

Στην περίπτωση που κάποιο από τα άκρα του διαστήματος  $[a, b]$  δεν μετακινείται σε διαδοχικές επαναλήψεις, έχουμε αργή σύγκλιση. Μπορούμε να βελτιώσουμε την τάξη της σύγκλισης της μεθόδου αν κάνουμε την ακόλουθη τροποποίηση στην επιλογή της ρίζας, όποτε συμβαίνει να μην αλλάζει ένα άκρο σε δύο διαδοχικές επαναλήψεις:

- αν άλλαξε δύο συνεχόμενες φορές το όριο  $a$

$$x = \frac{2bf(a) - af(b)}{2f(a) - f(b)}.$$

- αν άλλαξε δύο συνεχόμενες φορές το όριο  $b$

$$x = \frac{bf(a) - 2af(b)}{f(a) - 2f(b)}.$$

Η επιλογή του  $x$  επηρεάζεται μεγαλώνοντας τεχνητά την τιμή της συνάρτησης στο άκρο που έχει μετακινηθεί δύο διαδοχικές φορές.

Η παραπάνω τροποποίηση δίνει τάξη σύγκλισης  $\alpha = \sqrt[3]{3} \approx 1.442$  και είναι γνωστή ως ο αλγόριθμος Illinois.

## 2.4 Μέθοδος τέμνουσας

Σύμφωνα με αυτή τη μέθοδο, προσεγγίζουμε τη συνάρτηση  $f(x)$  με ευθεία που περνά από δύο σημεία  $(x_{n-1}, f(x_{n-1}))$  και  $(x_n, f(x_n))$ . Τα  $x_{n-1}, x_n$  είναι διαδοχικές προσεγγίσεις της ρίζας. Η νέα προσέγγιση,  $x_{n+1}$ , είναι η τομή με τον άξονα  $x$  (η ρίζα) της προσεγγιστικής ευθείας. Η ευθεία  $y = y(x)$  είναι

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

Επομένως,

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

Όπως καταλαβαίνετε, πρέπει να επιλέξουμε δύο αρχικά σημεία,  $x_0$  και  $x_1$ , με  $f(x_0) \neq f(x_1)$ , ώστε να παραγάγουμε την ακολουθία. Από την άλλη, η κάθε επανάληψη χρειάζεται ένα μόνο νέο υπολογισμό τιμής της συνάρτησης, πράγμα σημαντικό όταν ο υπολογισμός είναι σχετικά αργός.

Παρατηρήστε ότι η μέθοδος τέμνουσας μοιάζει πολύ με τη μέθοδο ψευδούς σημείου (§2.3). Όμως, στη μέθοδο τέμνουσας η ρίζα δεν είναι απαραίτητα περιορισμένη μεταξύ δύο σημείων.



**Αλγόριθμος:** Επίλυση της  $f(x) = 0$  με τη μέθοδο της τέμνουσας:

1. Επιλέγουμε δύο τιμές  $a, b$ .
2. Βρίσκουμε την τομή με τον άξονα των  $x$  της ευθείας που περνά από τα σημεία  $(a, f(a)), (b, f(b))$ . Την ονομάζουμε  $c$ .
3. Αν το  $c$  είναι ικανοποιητική προσέγγιση της ρίζας πηγαίνουμε στο βήμα 6.
4. Θέτουμε  $a \leftarrow b, b \leftarrow c$ .
5. Επαναλαμβάνουμε τη διαδικασία από το βήμα 2.
6. Τέλος.

### 2.4.1 Σύγκλιση της μεθόδου τέμνουσας

Μπορεί ναδειχθεί ότι η τάξη της σύγκλισης της μεθόδου τέμνουσας σε απλή ρίζα είναι  $\alpha = (1 + \sqrt{5})/2 \approx 1.618$ . Επομένως, η μέθοδος είναι πιο γρήγορη από άλλες πρώτης τάξης αλλά πιο αργή από μεθόδους δεύτερης τάξης.

## 2.5 Μέθοδος Müller

Η μέθοδος αυτή είναι παρόμοια με τη μέθοδο τέμνουσας αλλά προσεγγίζει τη συνάρτηση με παραβολή (εξίσωση της μορφής  $y = ax^2 + bx + c$ ) και, επομένως, χρειάζεται τρία σημεία για τον προσδιορισμό της. Η νέα προσέγγιση της ρίζας είναι η ρίζα της παραβολής που είναι πιο κοντά στην προηγούμενη προσέγγιση. Επομένως, επιλέγουμε τα σημεία  $x_0, x_1, x_2$  και ορίζουμε την παραβολή

$$y = a(x - x_2)^2 + b(x - x_2) + c.$$

Επιλέγουμε τα  $a, b, c$  ώστε να περνά από τα σημεία  $(x_i, f(x_i))$ ,  $i = 0, 1, 2$ :

$$\begin{aligned} a &= \frac{1}{x_1 - x_0} \left( \frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_2) - f(x_0)}{x_2 - x_0} \right), \\ b &= \frac{1}{x_1 - x_0} \left( (x_2 - x_0) \frac{f(x_2) - f(x_1)}{x_2 - x_1} - (x_2 - x_1) \frac{f(x_2) - f(x_0)}{x_2 - x_0} \right), \\ c &= f(x_2). \end{aligned}$$

Από τις δύο ρίζες της παραβολής<sup>1</sup>

$$x_{\pm} = x_2 - \frac{2c}{b \pm \sqrt{b^2 - 4ac}}$$

<sup>1</sup>χρησιμοποιούμε άλλο τύπο από το συνήθη για μεγαλύτερη ακρίβεια.

επιλέγουμε ως  $x_3$  αυτή που είναι πιο κοντά στη  $x_2$ , δηλαδή αυτή που έχει το μεγαλύτερο παρονομαστή κατ' απόλυτη τιμή. Η  $x_3$  είναι καλύτερη προσέγγιση της ρίζας της  $f(x)$ . Κατόπιν, επαναλαμβάνουμε τη διαδικασία για τα σημεία  $x_1, x_2, x_3$  ώστε να υπολογίσουμε μια ακόμα καλύτερη προσέγγιση (τη  $x_4$ ) κοκ.

Η μέθοδος Müller είναι γενικά πιο γρήγορη από τη μέθοδο τέμνουσας, με τάξη σύγκλισης, σε απλή ρίζα,  $\alpha \approx 1.84$ .

**Αλγόριθμος:** Επίλυση της  $f(x) = 0$  με τη μέθοδο Müller:

1. Επιλέγουμε τρεις διαφορετικές τιμές  $x_0, x_1, x_2$  στην περιοχή της αναζητούμενης ρίζας. Τα σημεία  $(x_i, f(x_i))$  δεν πρέπει να ανήκουν στην ίδια ευθεία.
2. Ορίζουμε τις ποσότητες

$$\begin{aligned} w_0 &= \frac{f(x_2) - f(x_0)}{x_2 - x_0} \\ w_1 &= \frac{f(x_2) - f(x_1)}{x_2 - x_1} \\ a &= \frac{w_1 - w_0}{x_1 - x_0}, \\ b &= w_0 + a(x_2 - x_0), \\ c &= f(x_2). \end{aligned}$$

3. Η επόμενη προσέγγιση της ρίζας δίνεται από τη σχέση

$$x_3 = x_2 - \frac{2c}{d},$$

όπου  $d$  ο, εν γένει μιγαδικός, αριθμός που έχει το μεγαλύτερο μέτρο μεταξύ των  $b + \sqrt{b^2 - 4ac}$ ,  $b - \sqrt{b^2 - 4ac}$ .

4. Αν η νέα προσέγγιση είναι ικανοποιητική πηγαίνουμε στο βήμα 6.
5. Θέτουμε  $x_0 \leftarrow x_1, x_1 \leftarrow x_2, x_2 \leftarrow x_3$ . Επαναλαμβάνουμε τη διαδικασία από το βήμα 2.
6. Τέλος.

Προσέξτε ότι οι διαδοχικές προσεγγίσεις της ρίζας μπορεί να είναι μιγαδικές λόγω της τετραγωνικής ρίζας, οπότε οι ποσότητες  $x_n, w_i, a, b, c, d$  είναι γενικά μιγαδικές. Επομένως, ο συγκεκριμένος αλγόριθμος μπορεί να υπολογίσει μιγαδικές ρίζες μιας συνάρτησης.

## 2.6 Μέθοδος Σταθερού Σημείου $x = g(x)$

Το πρόβλημα εύρεσης (πραγματικής) λύσης της  $f(x) = 0$  είναι ισοδύναμο με την επίλυση της εξίσωσης  $x = g(x)$  όπου  $g(x)$  κατάλληλη συνάρτηση. Ειδικές μορφές της  $g(x)$  δίνουν ευσταθείς και γρήγορους επαναληπτικούς αλγορίθμους για την εύρεση της λύσης.

**Αλγόριθμος:** Έστω η αρχική λύση (προσέγγιση)  $x_0$ . Κατασκευάζουμε την ακολουθία  $x_0, x_1, x_2, \dots, x_n$  ως εξής:

$$x_1 = g(x_0), \quad x_2 = g(x_1), \quad x_3 = g(x_2), \quad \dots, \quad x_n = g(x_{n-1}).$$

Αν η ακολουθία συγκλίνει σε ένα σημείο  $\bar{x}$  και καθώς η  $g(x)$  είναι συνεχής<sup>2</sup> έχουμε

$$\bar{x} \equiv \lim_{n \rightarrow \infty} x_n = \lim_{n \rightarrow \infty} g(x_{n-1}) = g(\lim_{n \rightarrow \infty} x_{n-1}) \equiv g(\bar{x}).$$

Άρα

1. Θέτουμε στο  $x$  την αρχική προσέγγιση.
2. Ελέγχουμε αν ικανοποιείται το κριτήριο τερματισμού (όποιο έχουμε επιλέξει).  
Αν ναι, πηγαίνουμε στο βήμα 4.
3. Θέτουμε  $x \leftarrow g(x)$  και επαναλαμβάνουμε από το βήμα 2.
4. Τέλος.

### 2.6.1 Ορισμός–Σχετικά Θεωρήματα

#### Ορισμός

Η συνάρτηση  $g(x)$  έχει σταθερό σημείο στο  $[a, b]$  αν υπάρχει  $\varrho \in [a, b]$  ώστε  $g(\varrho) = \varrho$ .

#### Κριτήριο ύπαρξης σταθερού σημείου

Έστω  $g(x)$  συνεχής συνάρτηση στο  $[a, b]$ , με  $a \leq g(x) \leq b, \forall x \in [a, b]$ . Τότε η  $g(x)$  έχει τουλάχιστον ένα σταθερό σημείο στο  $[a, b]$ .

**Απόδειξη:** Ισχύει  $g(a) \geq a, g(b) \leq b$ . Ορίζουμε τη συνεχή συνάρτηση  $h(x) = g(x) - x$ . Τότε  $h(a) \geq 0, h(b) \leq 0$ . Το ΘΕΤ εξασφαλίζει ότι υπάρχει  $\bar{x}$  ώστε  $h(\bar{x}) = 0$ .

#### Παράδειγμα

Έστω  $g(x) = 3^{-x}, x \in [0, 1]$ . Έχουμε  $g(0) = 1, g(1) = 1/3$  και  $g'(x) = -3^{-x} \ln 3 < 0 \forall x \in [0, 1]$ . Η  $g(x)$  είναι φθίνουσα και  $0 < 1/3 \leq g(x) \leq 1 \forall x \in [0, 1]$ . Από το κριτήριο ύπαρξης προκύπτει ότι η  $g(x)$  έχει τουλάχιστον ένα σταθερό σημείο

<sup>2</sup>ορισμός συνέχειας της  $g(x)$ :  $\lim_{n \rightarrow \infty} g(x_n) = g(\lim_{n \rightarrow \infty} x_n)$ .

(μοναδικό καθώς είναι φθίνουσα).

### Μοναδικότητα σταθερού σημείου

Έστω  $g(x)$  συνεχής και διαφορίσιμη συνάρτηση στο  $[a, b]$ , με  $a \leq g(x) \leq b$  και  $|g'(x)| < 1 \ \forall x \in [a, b]$ . Τότε η  $g(x)$  έχει μοναδικό σταθερό σημείο στο  $[a, b]$ .

**Απόδειξη:** Έστω  $p, r$  δύο σταθερά σημεία στο  $[a, b]$  με  $p \neq r$ . Θα έχουμε τότε  $p - r = g(p) - g(r)$ . Από το Θεώρημα Μέσης Τιμής προβλέπεται ότι υπάρχει  $\xi \in [a, b]$  ώστε  $g(p) - g(r) = g'(\xi)(p - r)$ . Επομένως, στο συγκεκριμένο  $\xi$  έχουμε  $g'(\xi) = 1$ , αντίθετα με την αρχική υπόθεση.

### Παράδειγμα

Η  $g(x) = \frac{x^2-1}{3}$  έχει μοναδικό σταθερό σημείο στο  $[-1, 1]$  καθώς, όταν  $|x| \leq 1$ , ισχύει α)  $-1/3 \leq g(x) \leq 0$  και κατ' επέκταση,  $-1 < g(x) < 1$ , και β)  $|g'(x)| = |2x/3| < 1$ .

### 2.6.2 Σύγκλιση της μεθόδου σταθερού σημείου

Έστω  $g(x)$  συνεχής και διαφορίσιμη συνάρτηση στο  $[a, b]$ , με  $a \leq g(x) \leq b$  και  $|g'(x)| \leq k < 1 \ \forall x \in [a, b]$ . Τότε, αν  $x_0 \in [a, b]$ , η ακολουθία  $x_{n+1} = g(x_n)$ ,  $n = 0, 1, \dots$  συγκλίνει στο μοναδικό σταθερό σημείο,  $\bar{x}$ , της  $g(x)$  στο  $[a, b]$ . Η ακρίβεια είναι  $|x_n - \bar{x}| \leq k^n \max(x_0 - a, b - x_0)$ ,  $n \geq 1$ .

Η γενική επαναληπτική μέθοδος  $x_{n+1} = g(x_n)$ ,  $n = 0, 1, \dots$  είναι *πρώτης τάξης* αν  $g'(x) \neq 0$ , *δεύτερης τάξης* αν  $g'(x) = 0$  και η  $g''(x)$  είναι συνεχής σε διάστημα που περικλείει τη ρίζα, κλπ.

### Παραδείγματα

1. Έστω η συνάρτηση  $f(x) = x^2 - 6x + 5$  με ρίζες 1.0, 5.0. Ας δοκιμάσουμε να τις εντοπίσουμε με την επαναληπτική σχέση

$$g(x) = \frac{x^2 + 5}{6} = x.$$

Για  $x_0 = 2.5$  έχουμε

$$\begin{aligned}x_1 = g(x_0) &= 1.8750 \\x_2 = g(x_1) &\approx 1.4193 \\x_3 = g(x_2) &\approx 1.1691 \\x_4 = g(x_3) &\approx 1.0611 \\x_5 = g(x_4) &\approx 1.0210 \\x_6 = g(x_5) &\approx 1.0078 \\x_7 = g(x_6) &\approx 1.0024 \\x_8 = g(x_7) &\approx 1.0008 \\x_9 = g(x_8) &\approx 1.0003 \\x_{10} = g(x_9) &\approx 1.0001 \\x_{11} = g(x_{10}) &\approx 1.0000\end{aligned}$$

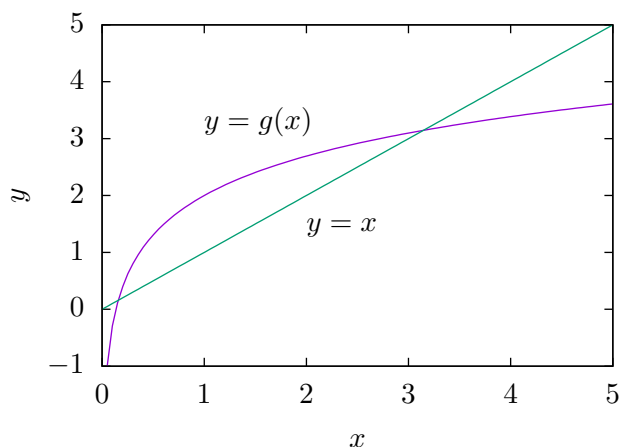
Αν δοκιμάσουμε άλλο αρχικό σημείο θα έχουμε πάλι σύγκλιση στο 1 ή απόκλιση στο  $+\infty$ . Μπορεί να αποδειχθεί ότι κανένα σημείο εκτός από το  $x_0 = 5.0$  δε δίνει ακολουθία με όριο την άλλη ρίζα.

2. Ας υπολογίσουμε τις ρίζες της  $f(x) = \ln x - x + 2$ ,  $x > 0$ . Γράφουμε  $g(x) = \ln x + 2 = x$ . Καθώς η  $g(x)$  είναι αύξουσα και  $g(1) = 2$ , υπάρχει ρίζα στο  $[0, 1]$ . Από το γράφημα Σχήμα 2.3) παρατηρούμε ότι η άλλη ρίζα είναι  $\bar{x} \approx 3.1$ .

Αν δοκιμάσουμε με αρχική προσέγγιση  $x_0 \in \{0.5, 1.0, 1.5, 2.0, 4.0, \dots\}$ , έχουμε σύγκλιση στη ρίζα  $\bar{x} = 3.146193\dots$ . Αντίθετα, δεν μπορούμε να βρούμε αρχικό σημείο για να εντοπίσουμε την άλλη ρίζα. Παρατηρήστε ότι για  $x_0 \leq e^{-2}$  ή  $x_0 \leq e^{e^{-2}-2}, \dots, x_0 \leq 0.158594339563$  δεν ορίζεται ακολουθία. (Η τιμή 0.158594339563 είναι η άλλη ρίζα· μπορείτε να την εντοπίσετε έχοντας ως  $g(x) = e^{x-2}$ ).

Εξετάστε τη σύγκλιση με διάφορα αρχικά  $x$  για την  $g(x) = x^{\frac{\ln x + 1}{x-1}}$ . Παρατηρήστε ότι διαφορετική επιλογή της  $g(x)$  και της αρχικής προσέγγισης μας δίνει διαφορετική ταχύτητα σύγκλισης (διαφορετικό αριθμό επαναλήψεων).

3. Η  $f(x) = x^3 + 4x^2 - 10 = 0$  έχει μία ρίζα στο  $[1, 1.5]$ . Η μέθοδος  $x = g(x)$  έχει διαφορετική ταχύτητα σύγκλισης ανάλογα με την επιλογή της  $g(x)$ , π.χ.  $g(x) = x - x^3 - 4x^2 + 10$ ,  $g(x) = \sqrt{\frac{10}{x}} - 4x$ ,  $g(x) = \sqrt{\frac{10}{4+x}}$ ,  $g(x) = \frac{1}{2}\sqrt{10 - x^3}$ , κλπ.

Σχήμα 2.3: Εκτίμηση των σταθερών σημείων της  $g(x) = \ln x + 2$ 

## 2.7 Μέθοδοι Householder

Η οικογένεια μεθόδων Householder αποτελείται από επαναληπτικές μεθόδους για την εύρεση ρίζας μιας συνάρτησης με συνεχείς παραγώγους τουλάχιστον μέχρι την τάξη  $d + 1$ . Η γενική σχέση που παράγει την ακολουθία  $x_0, x_1, x_2, \dots$  είναι

$$x_{n+1} = x_n + d \frac{(1/f)^{(d-1)}(x_n)}{(1/f)^{(d)}(x_n)}, \quad (2.4)$$

και για να ξεκινήσει χρειάζεται μία αρχική προσέγγιση  $x_0$ . Η τάξη της σύγκλισης είναι  $d + 1$ .

Παρακάτω θα δούμε αναλυτικά την μέθοδο για  $d = 1$ , που έχει την ειδική ονομασία «Newton-Raphson» και θα αναφέρουμε την μέθοδο για  $d = 2$  με την ειδική ονομασία «Halley».

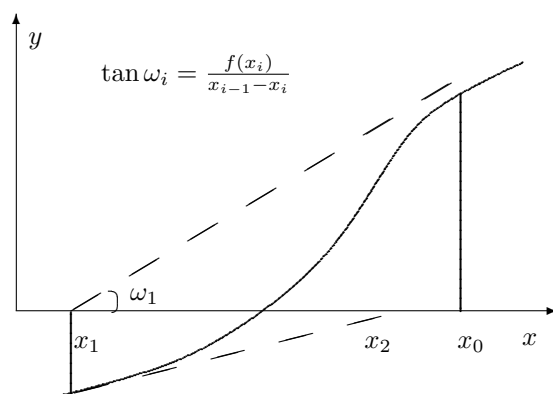
### 2.7.1 Μέθοδος Newton-Raphson

Η μέθοδος Newton-Raphson είναι επαναληπτική μέθοδος της μορφής  $x = g(x)$ . Η επιλογή της  $g(x)$  γίνεται ως εξής:

Έστω ότι αναζητούμε τη ρίζα της συνεχούς και διαφορίσιμης, σε διάστημα  $[a, b]$ , συνάρτησης  $f(x)$ . Αν γνωρίζουμε την τιμή αυτής και των παραγώγων της σε κάποιο σημείο  $x_0 \in [a, b]$ , το θεώρημα Taylor ((2.3)) μας εξασφαλίζει ότι στη ρίζα,  $\bar{x} \in [a, b]$ , ισχύει

$$f(\bar{x}) = f(x_0) + f'(x_0)(\bar{x} - x_0) + \frac{f''(\xi)}{2!}(\bar{x} - x_0)^2, \quad (2.5)$$

όπου  $\xi \in (\bar{x}, x_0)$ . Αγνοώντας τον όρο του υπολοίπου, θεωρώντας ότι η απόσταση



Σχήμα 2.4: Σχηματική εύρεση ρίζας με τη μέθοδο Newton–Raphson

$|\bar{x} - x_0|$  είναι μικρή, και καθώς ισχύει ότι  $f(\bar{x}) = 0$ , έχουμε

$$0 \approx f(x_0) + f'(x_0)(\bar{x} - x_0) \Rightarrow \bar{x} \approx x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Επομένως, η συνάρτηση  $g(x) = x - \frac{f(x)}{f'(x)}$  μπορεί να παραγάγει με τη μέθοδο σταθερού σημείου την ακολουθία διαδοχικών προσεγγίσεων στη ρίζα αρκεί να έχουμε  $f'(x_n) \neq 0$ :

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \quad (2.6)$$

Παρατηρήστε ότι σε κάθε επανάληψη πρέπει να υπολογίσουμε τις τιμές δύο συναρτήσεων  $(f(x), f'(x))$ .

Εύκολα δείχνεται ότι ο τύπος της μεθόδου αυτής μπορεί να προκύψει από τον γενικό τύπο των μεθόδων Householder ((2.4)) για  $d = 1$ . Επίσης, αν η παράγωγος δεν είναι γνωστή αναλυτικά, μπορεί να προσεγγιστεί με τους τύπους που παρουσιάζονται στο §4.5. Η προσέγγιση με τον τύπο (4.8α') μετατρέπει τη μέθοδο Newton–Raphson στη μέθοδο τέμνουσας.

**Θεώρημα (χωρίς απόδειξη):** Έστω ότι η  $f(x)$  είναι συνεχής και τουλάχιστον δύο φορές παραγωγίσιμη στο  $[a, b]$ , με συνεχή τη δεύτερη παράγωγό της. Αν  $\bar{x}$  ρίζα της  $f(x)$  στο  $[a, b]$  (δηλαδή  $f(\bar{x}) = 0$ ) και  $f'(x) \neq 0$  τότε υπάρχει  $\delta > 0$  ώστε η ακολουθία  $\{x_n\}$  που ορίζεται με τη μέθοδο Newton–Raphson συγκλίνει στο  $\bar{x}$ ,  $\forall x_0 \in [\bar{x} - \delta, \bar{x} + \delta]$ .

### Παράδειγμα

Έστω  $f(x) = x^2 - 6x + 5$ . Έχουμε

$$x_{n+1} = x_n - \frac{x_n^2 - 6x_n + 5}{2x_n - 6}, \quad n = 0, 1, 2, \dots$$

Οι διαδοχικές προσεγγίσεις των ριζών 1.0, 5.0 με αρχικά σημεία 2.0, 6.0 είναι οι εξής

$n$	$x_n^{(1)}$	$x_n^{(2)}$
0	2.0	6.0
1	0.5	5.16666666666667
2	0.95	5.00641025641026
3	0.999390243902439	5.00001024002622
4	0.999999907077705	5.00000000002621
5	0.999999999999998	5.0
6	1.0	

### Σύγκλιση αλγορίθμου Newton–Raphson

Ας υπολογίσουμε την ακρίβεια  $\varepsilon_n \equiv |x_n - \bar{x}|$  της μεθόδου. Από τον τύπο (2.6) έχουμε

$$\begin{aligned} x_{n+1} - \bar{x} &= x_n - \frac{f(x_n)}{f'(x_n)} - \bar{x} = \frac{f'(x_n)(x_n - \bar{x}) - f(x_n)}{f'(x_n)} \\ &= -\frac{1}{f'(x_n)} (f(x_n) + f'(x_n)(\bar{x} - x_n)) . \end{aligned}$$

Λαμβάνοντας υπόψη τη σχέση (2.5) έχουμε

$$x_{n+1} - \bar{x} = -\frac{1}{f'(x_n)} \left( f(\bar{x}) - \frac{f''(\xi)}{2} (\bar{x} - x_n)^2 \right) = \frac{f''(\xi)}{2f'(x_n)} (\bar{x} - x_n)^2 .$$

Επομένως

$$\varepsilon_{n+1} = \left| \frac{f''(\xi)}{2f'(x_n)} \right| \varepsilon_n^2 ,$$

με  $\xi$  μεταξύ των  $x_n$  και  $\bar{x}$ .

Συμπεραίνουμε ότι η μέθοδος είναι δεύτερης τάξης, παρουσιάζει δηλαδή *τετραγωνική σύγκλιση*. Αρκούν λίγα βήματα για να έχουμε πολύ ικανοποιητική προσέγγιση της ρίζας, με την προϋπόθεση ότι θα ξεκινήσουμε από σημείο όχι μακριά από αυτή. Από την άλλη, αν  $f'(\bar{x}) \approx 0$  έχουμε πολύ αργή σύγκλιση.

Η μέθοδος αυτή μπορεί να χρησιμοποιηθεί για την εύρεση μιγαδικής ρίζας πραγματικής ή μιγαδικής συνάρτησης. Σε αυτή την περίπτωση παίζει πολύ σημαντικό ρόλο η κατάλληλη επιλογή της αρχικής (μιγαδικής) τιμής ώστε να έχουμε σύγκλιση.

### Μέθοδοι Newton–Raphson για πολλαπλές ρίζες

Αν η ρίζα  $\bar{x}$  είναι πολλαπλή με πολλαπλότητα  $m$ , δηλαδή ισχύει  $f(\bar{x}) = f'(\bar{x}) = \dots = f^{(m-1)}(\bar{x}) = 0$ , με  $f^{(m)}(\bar{x}) \neq 0$ , μπορεί ναδειχθεί ότι ο τύπος Newton–Raphson συγκλίνει γραμμικά. Χρειάζεται τροποποίηση αν θέλουμε να διατηρήσει την τετραγωνική σύγκλιση.



Παρατηρήστε ότι η συνάρτηση  $f(x)$  με ρίζα το  $\bar{x}$ , πολλαπλότητας  $m$ , μπορεί να γραφεί στη μορφή  $f(x) = (x - \bar{x})^m g(x)$ , όπου  $g(x)$  συνάρτηση για την οποία το  $\bar{x}$  δεν είναι ρίζα. Συνεπώς, η συνάρτηση  $h_1(x) = \sqrt[m]{f(x)}$  έχει απλή ρίζα το  $\bar{x}$ . Ο τύπος Newton–Raphson, (2.6), για αυτή τη συνάρτηση αναμένουμε να έχει τετραγωνική σύγκλιση. Η εφαρμογή του δίνει

$$\begin{aligned} x_{n+1} &= x_n - \frac{h_1(x_n)}{h_1'(x_n)} = x_n - \frac{\sqrt[m]{f(x)}}{\frac{\sqrt[m]{f(x)}}{mf(x)} f'(x)} \Rightarrow \\ x_{n+1} &= x_n - m \frac{f(x_n)}{f'(x_n)}. \end{aligned}$$

Εύκολα δείχνεται ότι και η συνάρτηση  $h_2(x) = f(x)/f'(x)$  έχει απλή ρίζα το  $\bar{x}$ . Η εφαρμογή του τύπου Newton–Raphson σε αυτή δίνει άλλον ένα τύπο με τετραγωνική σύγκλιση:

$$\begin{aligned} x_{n+1} &= x_n - \frac{h_2(x_n)}{h_2'(x_n)} \Rightarrow \\ x_{n+1} &= x_n - \frac{f(x_n)f'(x_n)}{[f'(x_n)]^2 - f(x_n)f''(x_n)}. \end{aligned}$$

### 2.7.2 Μέθοδος Halley

Έστω ότι η συνάρτηση  $f(x)$  έχει απλές ρίζες σε κάποιο διάστημα, δεν μηδενίζονται δηλαδή ταυτόχρονα οι  $f(x)$ ,  $f'(x)$ . Τότε οι συναρτήσεις  $f(x)$  και  $g(x) = f(x)/\sqrt{|f'(x)|}$  έχουν τις ίδιες ρίζες.

Η εφαρμογή της μεθόδου Newton–Raphson για την εύρεση ρίζας της  $g(x)$  δίνει

$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)} = x_n - \frac{2f(x_n)f'(x_n)}{2[f'(x_n)]^2 - f(x_n)f''(x_n)}.$$

Ο τύπος της μεθόδου αυτής μπορεί να προκύψει από τον γενικό τύπο των μεθόδων Householder, (2.4), για  $d = 2$ , και μπορεί να χρησιμοποιηθεί για την εύρεση και μιγαδικών ριζών.

Μπορεί ναδειχθεί ότι η μέθοδος είναι τρίτης τάξης με ταχύτητα σύγκλισης

$$\lambda = \frac{3[f''(\bar{x})]^2 - 2f'(\bar{x})f'''(\bar{x})}{12[f'(\bar{x})]^2}.$$

## 2.8 Ασκήσεις

1. Υλοποιήστε τον αλγόριθμο διχοτόμησης σε κώδικα. Χρησιμοποιήστε τον για να εντοπίσετε τη ρίζα της

- $f(x) = x^3 + 4x^2 - 10$  στο διάστημα  $[1, 2]$ ,

- $f(x) = \sqrt{x} - \cos x$  στο διάστημα  $[0, 1]$ .

- (α') Γράψτε ένα πρόγραμμα το οποίο να υλοποιεί τη μέθοδο ψευδούς σημείου.  
(β') Εφαρμόστε την για να βρείτε τη ρίζα της

$$f(x) = -2.0 + 6.2x - 4.0x^2 + 0.7x^3$$

στο διάστημα  $[0.4, 0.6]$ .

- (γ') Εφαρμόστε τη μέθοδο ψευδούς σημείου και τη μέθοδο διχοτόμησης για να βρείτε τις ρίζες της

$$f(x) = x^{10} - 0.95$$

στο διάστημα  $[0, 1.4]$ . Ποια μέθοδος συγκλίνει πιο γρήγορα με σχετικό σφάλμα  $< 10^{-6}$ ;

- Βρείτε τη ρίζα της  $f(x) = x^2 - (1-x)^5$  στο  $[0, 1]$  με ακρίβεια  $10^{-9}$ , εφαρμόζοντας τη μέθοδο διχοτόμησης, τη μέθοδο ψευδούς σημείου και τη τροποποιημένη μέθοδο ψευδούς σημείου (αλγόριθμος Illinois). Πόσες επαναλήψεις και πόσους υπολογισμούς της συνάρτησης χρειαστήκατε σε κάθε μέθοδο;
- Χρησιμοποιήστε τη μέθοδο τέμνουσας για να βρείτε τη ρίζα της εξίσωσης  $g(x) = 3 \ln x + 5$  με ακρίβεια 6 σημαντικών ψηφίων.
- Δείξτε ότι η  $g(x) = \ln x + 2$  έχει ένα και μοναδικό σταθερό σημείο στο  $[2, 4]$ . Υπολογίστε το μέγιστο αριθμό επαναλήψεων ώστε  $|x_n - \bar{x}| \leq 10^{-3}$ .
- Γράψτε κώδικα που να υλοποιεί τη γενική επαναληπτική μέθοδο  $x = g(x)$ . Χρησιμοποιήστε τον για να υπολογίσετε
  - μια ρίζα της  $f(x) = x^2 - 6x + 5$ ,
  - τη ρίζα της  $f(x) = x - \cos^3 x$  κοντά στο 0.6.
- Υπολογίστε το  $y = \frac{e^x - 1}{x}$  με ένα ευσταθή αλγόριθμο για μικρό, κατ' απόλυτη τιμή,  $x$ . Για μικρό  $|x|$  χρησιμοποιούμε το ανάπτυγμα Taylor του  $e^x$  ώστε να αποφύγουμε την αλληλοαναίρεση όρων ίδιας τάξης.
- Υπολογίστε με ευσταθή αλγόριθμο τις λύσεις των εξισώσεων
  - $1.5x^2 + 13 \times 10^6 x + 0.037 = 0$ .  
Οι ακριβείς είναι  $x_1 \approx -2.8462 \times 10^{-9}$ ,  $x_2 \approx -8.6667 \times 10^6$ .
  - $1.5x^2 - 37 \times 10^6 x + 0.057 = 0$ .  
Οι ακριβείς είναι  $x_1 \approx 1.5405 \times 10^{-9}$ ,  $x_2 \approx 2.4667 \times 10^7$ .
- Εφαρμόστε τη μέθοδο Newton-Raphson για να υπολογίσετε τις ρίζες της
  - $f(x) = \sin x - x^2$ ,
  - $f(x) = 3xe^x - 1$ .

10. Υπολογίστε τις ρίζες της  $f(x) = 4 \cos x - e^{-x}$  με ακρίβεια  $10^{-8}$  με τη μέθοδο διχοτόμησης, τη μέθοδο σταθερού σημείου, τη μέθοδο Newton–Raphson και τη μέθοδο τέμνουσας.
11. Βρείτε με 12 ψηφία σωστά το σημείο τομής των καμπυλών  $e^x$ ,  $\tan(2x)$  στο διάστημα  $[-1, 1]$ . Συμβουλή: σχεδιάστε τις καμπύλες.
12. Υλοποιήστε σε κώδικα τον αλγόριθμο Müller. Εφαρμόστε τον για να βρείτε τη μη μηδενική ρίζα της  $f(x) = \sin x - x^2$ .
13. Υλοποιήστε σε κώδικα τη μέθοδο Newton–Raphson, κατάλληλα τροποποιημένη ώστε να υπολογίζει τις ρίζες πολυωνύμου βαθμού  $n$ ,  $p_n(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \dots + \alpha_n x^n$ , όταν έχουμε ως δεδομένους τους συντελεστές του  $\alpha_0, \alpha_1, \dots, \alpha_n$ . Το πολυώνυμο και η παράγωγός του να υπολογίζονται με τον αλγόριθμο Horner.
14. Υλοποιήστε σε κώδικα τη μέθοδο τέμνουσας, κατάλληλα τροποποιημένη ώστε να υπολογίζει τις ρίζες πολυωνύμου βαθμού  $n$ ,  $p_n(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \dots + \alpha_n x^n$ , όταν έχουμε ως δεδομένους τους συντελεστές του  $\alpha_0, \alpha_1, \dots, \alpha_n$ . Το πολυώνυμο και η παράγωγός του να υπολογίζονται με τον αλγόριθμο Horner.



## Κεφάλαιο 3

# Επίλυση Γραμμικών Συστημάτων και εφαρμογές

### 3.1 Εισαγωγή

Στο κεφάλαιο αυτό θα παρουσιάσουμε μεθόδους για την εύρεση της λύσης γενικών γραμμικών συστημάτων  $n \times n$ :

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \quad (3.1\alpha')$$

$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \quad (3.1\beta')$$

$$\vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots$$

$$a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n . \quad (3.1\gamma')$$

Οι συντελεστές  $a_{ij}$  και οι σταθεροί όροι  $b_i$  είναι γνωστοί, ενώ τα  $n$   $x_i$  είναι άγνωστα και προς εύρεση.

Το σύστημα μπορεί να εκφραστεί με την βοήθεια των πινάκων και διανυσμάτων  $\mathbf{A}_{n \times n} = [a_{ij}]$ ,  $\mathbf{x}_{n \times 1} = [x_i]$  και  $\mathbf{b}_{n \times 1} = [b_i]$  ως εξής

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} . \quad (3.2)$$

Αν όλα τα  $b_i$  είναι 0, το σύστημα χαρακτηρίζεται ως *ομογενές*.

#### 3.1.1 Ευστάθεια γραμμικών συστημάτων

Το σύστημα  $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$  χαρακτηρίζεται ως *ασταθές* αν έχουμε μεγάλη απόκλιση στη λύση για μικρές αλλαγές στα  $\mathbf{A}$ ,  $\mathbf{b}$ .

**Παράδειγμα**

$$\begin{bmatrix} 1 & 3 \\ 1 & 3.01 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 4.01 \end{bmatrix}$$

έχει λύση  $x_1 = x_2 = 1$ .

Το ελαφρά διαφορετικό σύστημα

$$\begin{bmatrix} 1 & 3 \\ 1 & 2.99 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 4.02 \end{bmatrix}$$

έχει λύση  $x_1 = 10$ ,  $x_2 = -2$ , τελείως διαφορετική.

Ο δείκτης κατάστασης,  $\kappa$ , του πίνακα  $A$  ως προς τη νόρμα  $\|\cdot\|$  ορίζεται ως

$$\kappa = \|A\| \cdot \|A^{-1}\|.$$

Π.χ. μία νόρμα είναι η «νόρμα αθροίσματος γραμμών»

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

Αν  $\kappa \gg 1$  το σύστημα είναι ασταθές.

**3.1.2 Ορισμοί—Βασικές γνώσεις****Επίλυση γραμμικής εξίσωσης μίας μεταβλητής**

Προτού δούμε τις μεθόδους λύσης γραμμικών συστημάτων, ας θυμηθούμε πώς επιλύεται μία γραμμική εξίσωση μίας μεταβλητής,  $ax = b$ :

- Αν  $a \neq 0$  η εξίσωση έχει μία λύση, την  $x = b/a$ .
- Αν  $a = 0$  εξετάζουμε το  $b$ :

§ Αν  $b \neq 0$  η εξίσωση δεν έχει λύση.

§ Αν  $b = 0$  η εξίσωση έχει άπειρες λύσεις (κάθε  $x$  ικανοποιεί την  $0x = 0$ ).

Στη διαδικασία επίλυσης ενός γραμμικού συστήματος με τη μέθοδο Gauss (§3.2.2), θα χρειαστεί να λύσουμε πρωτοβάθμιες εξισώσεις. Αυτές θα καθορίσουν τη λύση του συστήματος ανάλογα με τις τιμές των συντελεστών τους.

**Ιδιοτιμές—Ιδιοδιανύσματα**

Ας θυμίσουμε τον ορισμό των εννοιών της ιδιοτιμής και του ιδιοδιανύσματος ενός πίνακα  $A$ .

Αν υπάρχει ένας αριθμός  $\lambda$ , εν γένει μιγαδικός, και ένα διάνυσμα (πίνακας-στήλη)  $\mathbf{x}$ , διάφορο του  $[0, 0, \dots, 0]^T$  για τα οποία ισχύει

$$\mathbf{A} \cdot \mathbf{x} = \lambda \mathbf{x} , \quad (3.3)$$

τότε το  $\mathbf{x}$  λέγεται *ιδιοδιάνυσμα* του  $\mathbf{A}$  ενώ το  $\lambda$  είναι η αντίστοιχη *ιδιοτιμή*. Παρατηρήστε ότι το  $\mathbf{x}$  δεν είναι μοναδικό καθώς οποιοδήποτε πολλαπλάσιό του αποτελεί επίσης λύση του συστήματος (3.3) για την ίδια ιδιοτιμή. Συνήθως επιλέγουμε για ιδιοδιάνυσμα που αντιστοιχεί σε μία ιδιοτιμή αυτό που έχει μέτρο 1: επιλέγουμε δηλαδή την πολλαπλασιαστική σταθερά  $c$  στο διάνυσμα  $c\mathbf{x}$  να είναι τέτοια ώστε

$$(c\mathbf{x})^\dagger \cdot (c\mathbf{x}) = 1 \Rightarrow |c|^2 = \frac{1}{\mathbf{x}^\dagger \cdot \mathbf{x}} .$$

Τη φάση της γενικά μιγαδικής ποσότητας  $c$  μπορούμε να την πάρουμε αυθαίρετα ίση με 0, καταλήγοντας σε πραγματική  $c$ . Η διαδικασία αυτή λέγεται κανονικοποίηση.

### Ορίζουσα

Η ορίζουσα είναι ένας αριθμός που σχετίζεται με ένα τετραγωνικό πίνακα. Μπορεί να οριστεί με πολλούς ισοδύναμους τρόπους. Ένας ορισμός είναι το *ανάπτυγμα Laplace*: η ορίζουσα δίνεται ως ανάπτυγμα κατά κάποια στήλη  $j$  της επιλογής μας με την αναδρομική σχέση

$$\det(\mathbf{A}) = \sum_{i=1}^n (-1)^{i+j} a_{ij} \det(\mathbf{A}_{ij}) , \quad (3.4)$$

όπου  $\mathbf{A}_{ij}$  είναι ο πίνακας διαστάσεων  $(n-1) \times (n-1)$  που προκύπτει από τον  $\mathbf{A}$  διαγράφοντας τη γραμμή  $i$  και τη στήλη  $j$ . Ο τύπος αυτός ισχύει για  $n > 1$  και είναι ανεξάρτητος από την επιλογή του  $j$ . Αντίστοιχος τύπος προκύπτει με ανάπτυξη κατά γραμμή. Επιπλέον, η ορίζουσα ενός πίνακα  $1 \times 1$  είναι το μοναδικό στοιχείο του.

### Συμμετρικός θετικά ορισμένος πίνακας

Ένας πραγματικός τετραγωνικός πίνακας  $\mathbf{A}$  είναι *συμμετρικός* αν είναι ίσος με τον *ανάστροφό του*,  $\mathbf{A} = \mathbf{A}^T$ . Ο *ανάστροφος* πίνακας,  $\mathbf{A}^T$ , έχει στοιχεία  $a_{ij}^T = a_{ji}$ .

Ένας πραγματικός συμμετρικός πίνακας  $\mathbf{A}$  χαρακτηρίζεται ως *θετικά ορισμένος* αν ισχύουν (μεταξύ άλλων) τα ισοδύναμα κριτήρια:

- Ισχύει  $\mathbf{x}^T \cdot \mathbf{A} \cdot \mathbf{x} > 0$  για κάθε πραγματικό μη μηδενικό διάνυσμα  $\mathbf{x}$ .
- Όλες οι ιδιοτιμές του είναι πραγματικές και θετικές.
- Στην ανάλυση  $LU$  του  $\mathbf{A}$  (§3.2.4), ο πίνακας  $\mathbf{L}$  έχει θετικά διαγώνια στοιχεία (θεωρούμε ότι κάθε στοιχείο της διαγωνίου του  $\mathbf{U}$  έχει τιμή 1).

- Υπάρχει πραγματικός αντιστρέψιμος πίνακας  $B$  για τον οποίο ισχύει  $A = B^T \cdot B$ .
- Υπάρχει ένας και μοναδικός πραγματικός κάτω τριγωνικός πίνακας  $L$  (ή άνω τριγωνικός πίνακας  $U$ ) με θετικά διαγώνια στοιχεία για τον οποίο ισχύει  $A = L \cdot L^T$  (ή  $A = U^T \cdot U$ ) (ανάλυση Cholesky).
- Είναι θετικές οι ορίζουσες (§3.1.2) όλων των τετραγωνικών υπο-πινάκων του  $A$  με πάνω αριστερό στοιχείο το  $a_{11}$  και κάτω δεξιό το  $a_{ii}$ ,  $i = 1, 2, \dots, n$  (κριτήριο του Sylvester<sup>1</sup>).

Μπορεί ναδειχθεί ότι για ένα πραγματικό, συμμετρικό, θετικά ορισμένο πίνακα  $A$  ισχύουν τα εξής

- τα διαγώνια στοιχεία  $a_{ii}$  είναι θετικά.
- η ορίζουσα είναι θετική και μικρότερη ή ίση από το γινόμενο των διαγώνιων στοιχείων του.
- Σε κάθε γραμμή, το διαγώνιο στοιχείο είναι μεγαλύτερο ή ίσο από τις απόλυτες τιμές των υπόλοιπων στοιχείων της γραμμής.

### Συνθήκες επιλυσιμότητας

Οι παρακάτω συνθήκες είναι ισοδύναμες:

- Για οποιοδήποτε δεύτερο μέλος  $b$ , το σύστημα  $A \cdot x = b$  έχει μοναδική λύση.
- Ο πίνακας  $A$  έχει αντίστροφο.
- Η ορίζουσα του  $A$  είναι μη μηδενική.
- Το ομογενές σύστημα  $A \cdot x = 0$  έχει μοναδική λύση την  $x = 0$ .
- Οι στήλες ή οι γραμμές του  $A$  είναι γραμμικά ανεξάρτητες.

Τις βασικές μεθόδους επίλυσης γραμμικών συστημάτων τις διακρίνουμε σε απευθείας (direct) και επαναληπτικές (iterative).

<sup>1</sup>Η εφαρμογή του κριτηρίου του Sylvester είναι ένας εύκολος τρόπος για να ελέγξουμε αν ένας συμμετρικός πίνακας είναι θετικά ορισμένος. Συγκεκριμένα, τον τριγωνοποιούμε (§3.4.2) κάνοντας άρτιο πλήθος εναλλαγών γραμμών (ή 0) ώστε να διατηρηθεί το πρόσημο των ορίζουσών των υπο-πινάκων. Αν και μόνο αν τα διαγώνια στοιχεία του τριγωνικού πίνακα είναι θετικά, ο πίνακας είναι θετικά ορισμένος.



## 3.2 Απευθείας μέθοδοι

Οι απευθείας μέθοδοι επίλυσης γραμμικών συστημάτων δίνουν την ακριβή λύση (με κάποιο σφάλμα στρογγύλευσης) σε συγκεκριμένο και εκ των προτέρων υπολογίσιμο αριθμό βημάτων/πράξεων. Μια εύκολη στην αντίληψη αλλά χρονοβόρα στην υλοποίηση μέθοδος απαιτεί τον υπολογισμό του αντίστροφου πίνακα του  $A$  (αρκεί αυτός να υπάρχει) ώστε η λύση να είναι  $x = A^{-1} \cdot b$ . Αν και θα παρουσιάσουμε μέθοδο εύρεσης του αντίστροφου πίνακα δεν θα τη χρησιμοποιούμε για επίλυση συστήματος καθώς υπάρχουν πιο γρήγορες μέθοδοι.

### 3.2.1 Κανόνας Cramer

Ο κανόνας Cramer προσδιορίζει τη λύση του γραμμικού συστήματος  $A \cdot x = b$  ως εξής:

$$x_j = \frac{\det(B_j)}{\det(A)}, \quad j = 1, 2, \dots, n,$$

όπου ο πίνακας  $B_j$  προκύπτει από τον  $A$  αν αντικαταστήσουμε την στήλη  $j$  του  $A$  με το διάνυσμα  $b$ .

Ας αποδείξουμε την πρώτη από τις σχέσεις. Θυμηθείτε ότι η ορίζουσα πίνακα που μια στήλη του γράφεται ως άθροισμα προσθετέων, ισούται με το άθροισμα των ορίζουσών που προκύπτουν από την αρχική, η κάθε μία με ένα όρο στη στήλη. Έτσι, αν στην ορίζουσα του  $B_1$  αντικαταστήσουμε τα  $b_i$  με τις αριστερά μέλη των εξισώσεων (3.1) έχουμε

$$\begin{aligned} \det(B_1) &= \begin{vmatrix} b_1 & a_{12} & \cdots & a_{1n} \\ b_2 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_n & a_{n2} & \cdots & a_{nn} \end{vmatrix} = \begin{vmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n & a_{12} & \cdots & a_{1n} \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n & a_{n2} & \cdots & a_{nn} \end{vmatrix} \\ &= \begin{vmatrix} a_{11}x_1 & a_{12} & \cdots & a_{1n} \\ a_{21}x_1 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}x_1 & a_{n2} & \cdots & a_{nn} \end{vmatrix} + \begin{vmatrix} a_{12}x_2 & a_{12} & \cdots & a_{1n} \\ a_{22}x_2 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n2}x_2 & a_{n2} & \cdots & a_{nn} \end{vmatrix} \\ &\quad + \cdots + \begin{vmatrix} a_{1n}x_n & a_{12} & \cdots & a_{1n} \\ a_{2n}x_n & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{nn}x_n & a_{n2} & \cdots & a_{nn} \end{vmatrix} \end{aligned}$$

Όλες οι ορίζουσες στις οποίες καταλήξαμε, εκτός από την πρώτη, είναι μηδέν καθώς έχουν δύο στήλες ανάλογες. Από την πρώτη μπορούμε να βγάλουμε κοινό παράγοντα το  $x_1$  οπότε απομένει η ορίζουσα του  $A$ . Καταλήγουμε στη σχέση που θέλαμε να αποδείξουμε:

$$\det(B_1) = x_1 \det(A),$$

Ο υπολογισμός των οριζουσών μπορεί να γίνει με τον ορισμό (3.4) ή με τις μεθόδους που παρουσιάζονται στην §3.4.2.

Η λύση ενός γενικού γραμμικού συστήματος με τον κανόνα Cramer και απλοϊκό υπολογισμό των οριζουσών, απαιτεί αριθμό πράξεων της τάξης του  $(n+1)!$  και γι' αυτό δεν εφαρμόζεται στην πράξη για  $n \geq 4$ . Επιπλέον, η μέθοδος είναι αριθμητικά ασταθής για πίνακα  $A$  με ορίζουσα πολύ κοντά στο 0 καθώς ο υπολογισμός της ακυρώνει τα σημαντικά ψηφία των συντελεστών του πίνακα. Ας αναφέρουμε ότι έχει αναπτυχθεί πολύπλοκη μέθοδος<sup>2</sup> που υπολογίζει τις ορίζουσες σε λιγότερες πράξεις, κατεβάζοντας το συνολικό αριθμό απαιτούμενων πράξεων σε ανάλογο του  $n^3$ .

### 3.2.2 Απαλοιφή Gauss

Μια απλή μέθοδος επίλυσης είναι η μέθοδος αντικατάστασης: λύνουμε την πρώτη εξίσωση ως προς την πρώτη μεταβλητή και την αντικαθιστούμε στις επόμενες. Κατόπιν λύνουμε τη δεύτερη εξίσωση ως προς τη δεύτερη μεταβλητή και την αντικαθιστούμε στις επόμενες, κοκ. Η συστηματική εφαρμογή της αποτελεί ουσιαστικά τη μέθοδο απαλοιφής Gauss.

Η μέθοδος της απαλοιφής Gauss αποτελείται από δύο στάδια:

1. Μετατρέπουμε, με κατάλληλους μετασχηματισμούς, το γενικό γραμμικό σύστημα (3.1) σε άνω τριγωνικό:

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots + a_{1n}x_n = b_1 \quad (3.5\alpha')$$

$$a_{22}x_2 + a_{23}x_3 + \dots + a_{2n}x_n = b_2 \quad (3.5\beta')$$

$$a_{33}x_3 + \dots + a_{3n}x_n = b_3 \quad (3.5\gamma')$$

$$\vdots \quad \quad \quad \vdots \quad \quad \quad \vdots$$

$$a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n = b_{n-1} \quad (3.5\delta')$$

$$a_{nn}x_n = b_n \quad (3.5\epsilon')$$

Οι μετασχηματισμοί είναι τέτοιοι ώστε να διατηρούν τη λύση.

2. Επιλύουμε το άνω τριγωνικό σύστημα. Η λύση τριγωνικών συστημάτων εκφράζεται με «κλειστούς» τύπους.

#### Τριγωνοποίηση

Σε ένα γραμμικό σύστημα μπορούμε να εκτελέσουμε τους παρακάτω στοιχειώδεις μετασχηματισμούς χωρίς να επηρεαστεί η λύση του:

- Εναλλαγή της σειράς δύο εξισώσεων,
- Πρόσθεση σε μία εξίσωση μιας άλλης,

<sup>2</sup><http://dx.doi.org/10.1016/j.jda.2011.06.007>

- Πολλαπλασιασμός μιας εξίσωσης με ένα μη μηδενικό αριθμό.

Οι δύο τελευταίοι μετασχηματισμοί έχουν ως συνέπεια ότι μπορούμε να προσθέσουμε στην εξίσωση  $p$  το πολλαπλάσιο της εξίσωσης  $q$  χωρίς να αλλάξει η λύση. Ας συμβολίσουμε αυτό το μετασχηματισμό με  $[p] \leftarrow [p] + \lambda[q]$ .

**Πρώτη στήλη.** Ας δούμε με ποιους μετασχηματισμούς μπορούμε να μηδενίσουμε τους όρους κάτω από τη διαγώνιο στην πρώτη στήλη: για να είμαστε συστηματικοί, επιλέγουμε την πρώτη εξίσωση και την προσθέτουμε σε κάθε επόμενη, πολλαπλασιασμένη με κατάλληλους αριθμούς. Έτσι έχουμε

$$\begin{aligned} [2] &\leftarrow [2] + \lambda_2[1], \\ [3] &\leftarrow [3] + \lambda_3[1], \\ &\vdots \\ [n] &\leftarrow [n] + \lambda_n[1]. \end{aligned}$$

Ο μετασχηματισμός σε κάθε εξίσωση  $i = 2, 3, \dots, n$  δίνει

$$\begin{aligned} a_{ij} &\leftarrow a_{ij} + \lambda_i a_{1j}, \quad j = 1, 2, \dots, n \\ b_i &\leftarrow b_i + \lambda_i b_1. \end{aligned}$$

Καθώς θέλουμε να έχουμε μετά το μετασχηματισμό  $a_{i1} = 0$ , πρέπει να ισχύει  $\lambda_i = -a_{i1}/a_{11}$ . Θεωρούμε ότι  $a_{11} \neq 0$ . Θα εξετάσουμε παρακάτω τι πρέπει να κάνουμε αν δεν ισχύει αυτό.

Συνοψίζοντας, μηδενίζουμε τους συντελεστές της πρώτης στήλης κάτω από τη διαγώνιο με τις εξής πράξεις:

$$\lambda_i = -a_{i1}/a_{11} \quad (3.6\alpha')$$

$$a_{ij} \leftarrow a_{ij} + \lambda_i a_{1j}, \quad j = 1, 2, \dots, n \quad (3.6\beta')$$

$$b_i \leftarrow b_i + \lambda_i b_1, \quad (3.6\gamma')$$

για  $i = 2, 3, \dots, n$ .

Το σύστημα (3.1) θα γίνει

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

**Δεύτερη στήλη.** Ας δούμε πώς μηδενίζουμε τα στοιχεία της δεύτερης στήλης, κάτω από τη διαγώνιο. Επιλέγουμε τη *δεύτερη* γραμμή και την προσθέτουμε σε

κάθε επόμενη, πολλαπλασιασμένη με κατάλληλους αριθμούς. Επομένως

$$\begin{aligned} [3] &\leftarrow [3] + \lambda_3[2], \\ [4] &\leftarrow [4] + \lambda_4[2], \\ &\vdots \\ [n] &\leftarrow [n] + \lambda_n[2]. \end{aligned}$$

Ο μετασχηματισμός σε κάθε εξίσωση  $i = 3, 4, \dots, n$  δίνει

$$\begin{aligned} a_{ij} &\leftarrow a_{ij} + \lambda_i a_{2j}, \quad j = 2, 3, \dots, n \\ b_i &\leftarrow b_i + \lambda_i b_2. \end{aligned}$$

Προσέξτε ότι ο δείκτης  $j$  ξεκινά από το 2 (είναι περιττό να ξεκινήσουμε από το 1 καθώς οι συντελεστές  $a_{i1}$  κάθε γραμμής  $i$  με  $i = 3, 4, \dots, n$  είναι 0).

Καθώς θέλουμε να έχουμε μετά το μετασχηματισμό  $a_{i2} = 0$ , προκύπτει ότι πρέπει να ισχύει  $\lambda_i = -a_{i2}/a_{22}$  με  $i = 3, 4, \dots, n$ .

Συνοψίζοντας, μηδενίζουμε τους συντελεστές της δεύτερης στήλης κάτω από τη διαγώνιο με τις εξής πράξεις:

$$\lambda_i = -a_{i2}/a_{22} \quad (3.7\alpha')$$

$$a_{ij} \leftarrow a_{ij} + \lambda_i a_{2j}, \quad j = 2, 3, \dots, n \quad (3.7\beta')$$

$$b_i \leftarrow b_i + \lambda_i b_2, \quad (3.7\gamma')$$

για  $i = 3, 4, \dots, n$ .

**Γενικοί Τύποι.** Από τους τύπους που βγάλαμε για την πρώτη και δεύτερη στήλη, μπορούμε να εξαγάγουμε τους γενικούς τύπους για κάθε στήλη, δηλαδή τον αλγόριθμο που μετατρέπει ένα γενικό γραμμικό σύστημα σε άνω τριγωνικό. Έτσι, αν ο δείκτης που είναι 1 στις εξισώσεις (3.6) γίνεται 2 στις (3.7), συμπεραίνουμε ότι θα γίνεται  $k$  για την στήλη  $k$ :

$$\lambda_i = -a_{ik}/a_{kk} \quad (3.8\alpha')$$

$$a_{ij} \leftarrow a_{ij} + \lambda_i a_{kj}, \quad j = k, k+1, \dots, n \quad (3.8\beta')$$

$$b_i \leftarrow b_i + \lambda_i b_k, \quad (3.8\gamma')$$

με  $i = k+1, \dots, n$  (ο δείκτης  $i$  χρησιμοποιείται για να διατρέξουμε τις επόμενες εξισώσεις από την  $k$ ).

Τις εξισώσεις (3.8) θα τις εκτελέσουμε διαδοχικά για  $k = 1, 2, \dots, n-1$  (η στήλη  $k = n$  δεν έχει στοιχεία κάτω από τη διαγώνιο). Στο τέλος της διαδικασίας, το γενικό γραμμικό σύστημα θα έχει μετατραπεί σε άνω τριγωνικό.

**Παρατήρηση:** Στην περίπτωση που κάποιος συντελεστής  $a_{KK}$  είναι ή γίνει κατά την εφαρμογή του αλγορίθμου ίσος με 0, δεν μπορούμε να εφαρμόσουμε τις εξισώσεις (3.8) για την εξίσωση  $K$  ως έχει. Πρέπει να εναλλάξουμε την επίμαχη εξίσωση  $K$  με κάποια από τις επόμενες της ώστε να έρθει στη διαγώνιο ένας μη μηδενικός συντελεστής. Θα αναφέρουμε στην §3.2.2 πώς μπορούμε να επιλέξουμε την καταλληλότερη εξίσωση. Κατόπιν, μπορούμε να συνεχίσουμε τη διαδικασία.

Αν δεν μπορούμε να βρούμε μη μηδενικό συντελεστή στη στήλη  $K$ , στις επόμενες του  $K$  γραμμές, προχωρούμε τη διαδικασία κανονικά στο επόμενο  $k$ . Το τριγωνικό σύστημα που θα προκύψει, όπως θα δούμε παρακάτω, δεν θα έχει μοναδική λύση.

### Επίλυση άνω τριγωνικού συστήματος

Η εύρεση της λύσης ενός άνω τριγωνικού συστήματος, (3.5), γίνεται με τη μέθοδο οπισθοδρόμησης, από την τελευταία προς την πρώτη εξίσωση. Έχουμε διαδοχικά για την τελευταία, προτελευταία, κλπ. πρώτη εξίσωση

$$\begin{aligned} x_n &= \frac{1}{a_{nn}} b_n, \\ x_{n-1} &= \frac{1}{a_{n-1,n-1}} (b_{n-1} - a_{n-1,n} x_n), \\ &\vdots \\ x_1 &= \frac{1}{a_{11}} (b_1 - a_{12} x_2 - a_{13} x_3 - \cdots - a_{1n} x_n). \end{aligned}$$

Ο γενικός τύπος είναι

$$x_i = \frac{1}{a_{ii}} \left( b_i - \sum_{j=i+1}^n a_{ij} x_j \right), \quad i = n, n-1, \dots, 1. \quad (3.9)$$

Στον υπολογισμό του αθροίσματος χρησιμοποιούμε την ακόλουθη σύμβαση: όταν το κάτω όριο του δείκτη άθροισης είναι μεγαλύτερο από το άνω (επομένως, στην περίπτωσή μας, όταν  $i = n$ ), το άθροισμα είναι 0.

**Παρατήρηση:** Άς θυμηθούμε πώς επιλύεται μια γραμμική εξίσωση μίας μεταβλητής,  $ax = b$ :

- Αν  $a \neq 0$  η εξίσωση έχει μία λύση, την  $x = b/a$ .
- Αν  $a = 0$  εξετάζουμε το  $b$ :

§ Αν  $b \neq 0$  η εξίσωση δεν έχει λύση.

§ Αν  $b = 0$  η εξίσωση έχει άπειρες λύσεις (κάθε  $x$  ικανοποιεί την  $0x = 0$ ).

Σύμφωνα με τα παραπάνω, αν κάποιος συντελεστής  $a_{II}$  είναι 0, εξετάζουμε τον αριθμητή στη σχέση (3.9):

- αν

$$b_I - \sum_{j=I+1}^n a_{Ij}x_j = 0$$

το σύστημα έχει άπειρες λύσεις. Τα  $x_i$  με  $i < I$  θα εκφράζονται ως συναρτήσεις του  $x_I$ , δεν θα μπορούν να πάρουν συγκεκριμένη αριθμητική τιμή. Το  $x_I$  θα είναι ελεύθερη ποσότητα που θα μπορεί να πάρει οποιαδήποτε τιμή θέλουμε.

- αν

$$b_I - \sum_{j=I+1}^n a_{Ij}x_j \neq 0$$

το σύστημα δεν έχει λύση.

### Παράδειγμα

Το σύστημα

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

επιλύεται ως εξής:

1. Καθώς  $a_{11} = 0$  και  $a_{21} \neq 0$  εναλλάσσουμε τις δύο πρώτες εξισώσεις

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

2. Η δεύτερη εξίσωση έχει ήδη  $a_{21} = 0$ , όπως επιδιώκουμε. Πολλαπλασιάζουμε την πρώτη εξίσωση με  $-2/5$  και την προσθέτουμε στην τρίτη, ώστε να μηδενιστεί και το νέο  $a_{31}$ :

$$\begin{bmatrix} 5 & 3 & 1 \\ 0 & 1 & 2 \\ 0 & -3.2 & 0.6 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 4 \\ 3 \\ 4.4 \end{bmatrix}.$$

3. Συνεχίζουμε με τη δεύτερη στήλη: Πολλαπλασιάζουμε τη δεύτερη εξίσωση με 3.2 και την προσθέτουμε στην τρίτη ώστε να μηδενιστεί και το νέο  $a_{32}$ :

$$\begin{bmatrix} 5 & 3 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 7 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 4 \\ 3 \\ 14 \end{bmatrix}.$$

4. Ο πίνακας έχει έρθει σε άνω τριγωνική μορφή. Με οπισθοδρόμηση έχουμε

$$\begin{aligned} 7x_3 &= 14 \Rightarrow x_3 = 2, \\ x_2 + 2x_3 &= 3 \Rightarrow x_2 = 3 - 2x_3 = -1, \\ 5x_1 + 3x_2 + x_3 &= 4 \Rightarrow x_1 = (4 - x_3 - 3x_2)/5 = 1. \end{aligned}$$

## Παρατηρήσεις

**Απαιτήσεις μνήμης και χρόνου (πράξεων)** Ο γενικός πίνακας  $A$  χρειάζεται  $n^2$  θέσεις μνήμης για πραγματικούς ή μιγαδικούς (όποιου τύπου είναι τα στοιχεία του). Επιπλέον  $n$  θέσεις απαιτεί ο  $b$ . Παρατηρήστε ότι ο  $b$  μπορεί να χρησιμοποιηθεί για την αποθήκευση του διανύσματος  $x$ .

Όπως παρατηρούμε από τους γενικούς τύπους της, (3.8), η τριγωνοποίηση ενός γενικού πίνακα απαιτεί

$$\begin{aligned} \sum_{k=1}^{n-1} \sum_{i=k+1}^n 1 &= \frac{n(n-1)}{2} \text{ διαιρέσεις,} \\ \sum_{k=1}^{n-1} \sum_{i=k+1}^n \sum_{j=k+1}^{n+1} 1 &= \frac{n(n-1)(n+1)}{3} \text{ πολλαπλασιασμούς,} \\ \sum_{k=1}^{n-1} \sum_{i=k+1}^n \sum_{j=k+1}^{n+1} 1 &= \frac{n(n-1)(n+1)}{3} \text{ αφαιρέσεις.} \end{aligned}$$

Στην εξίσωση (3.8β') δεν έχουμε συνυπολογίσει στις πράξεις την εφαρμογή της για  $j = k$ , καθώς αυτή εκ κατασκευής μας μηδενίζει τους συντελεστές της στήλης  $k$  κάτω από τη διαγώνιο. Μπορούμε να τους θέσουμε απευθείας 0.

Από τους γενικούς τύπους, (3.9), της επίλυσης ενός άνω τριγωνικού πίνακα, προκύπτει ότι χρειαζόμαστε

$$\begin{aligned} \sum_{i=1}^n 1 &= n \text{ διαιρέσεις,} \\ \sum_{i=1}^{n-1} \sum_{j=i+1}^n 1 &= \frac{n(n-1)}{2} \text{ πολλαπλασιασμούς,} \\ \sum_{i=1}^{n-1} \sum_{j=i+1}^n 1 &= \frac{n(n-1)}{2} \text{ αφαιρέσεις.} \end{aligned}$$

Επομένως, η μέθοδος Gauss χρειάζεται, στη γενική περίπτωση,  $n(n+1)/2$  διαιρέσεις,  $n(n-1)(2n+5)/6$  πολλαπλασιασμούς και  $n(n-1)(2n+5)/6$  αφαιρέσεις. Συνολικά, περίπου  $2n^3/3$  πράξεις, πολύ λιγότερες από τις  $(n+1)!$  που απαιτεί η μέθοδος Cramer.

**Πολλαπλά δεξιά μέλη,  $b = B_{n \times m}$**  Όταν θέλουμε να επιλύσουμε πολλές φορές το σύστημα με ίδιο πίνακα  $A$  αλλά  $m$  διαφορετικά δεξιά μέλη  $b$ , είναι προτιμότερο να εκτελέσουμε συγχρόνως τη διαδικασία για όλα τα  $b$ , δηλαδή, να σχηματίσουμε ένα πίνακα  $B$  με  $m$  στήλες και να επεκτείνουμε τις πράξεις που υπαγορεύει ο αλγόριθμος για το  $b$  σε όλες τις στήλες του.

**Μερική οδήγηση κατά γραμμές** Για να ελαχιστοποιήσουμε τα αριθμητικά σφάλματα κατά την τριγωνοποίηση, είναι σημαντικό να επιλέγουμε κάθε φορά το διαγώνιο συντελεστή  $a_{kk}$  (ο οποίος διαιρεί την  $k$  εξίσωση) ώστε να είναι αρκετά μεγάλος κατ' απόλυτη τιμή. Μπορούμε να κάνουμε κατάλληλη εναλλαγή γραμμών (της  $k$  με κάποια από τις επόμενες, με  $i > k$ ) ώστε να μεταφερθεί στη διαγώνιο το μεγαλύτερο κατ' απόλυτη τιμή στοιχείο από τα  $a_{ik}$ ,  $i \geq k$ . Η συγκεκριμένη πράξη δεν αυξάνει ιδιαίτερα το υπολογιστικό κόστος του αλγόριθμου, ειδικά αν δεν γίνει στην πραγματικότητα η εναλλαγή στοιχείων αλλά τροποποιηθούν οι δείκτες με τους οποίους διατρέχουμε τις εξισώσεις.

Παρατηρήστε ότι οποιοδήποτε στοιχείο σε κάποια γραμμή μπορεί να γίνει όσο μεγάλο θέλουμε αν πολλαπλασιάσουμε την εξίσωση στην οποία ανήκει με κατάλληλο αριθμό. Γι' αυτό, καλό είναι να λαμβάνουμε υπόψη τις σχετικές τιμές των συντελεστών ως προς το μεγαλύτερο συντελεστή της εξίσωσης στην οποία ανήκουν. Σε αυτή την παραλλαγή της μερικής οδήγησης, υπολογίζουμε κάθε φορά το μέγιστο στοιχείο των γραμμών με  $i \geq k$ ,  $M_i = \max_j |a_{ij}|$  με  $j = 1, \dots, n$ . Κατόπιν, κάνουμε σύγκριση των απόλυτων τιμών του  $a_{kk}/M_k$  με τα  $a_{ik}/M_i$ ,  $i > k$ . Το υπολογιστικό κόστος αυξάνει αλλά ο αλγόριθμος γίνεται πιο ευσταθής.

**Παράδειγμα:** Το σύστημα

$$\begin{bmatrix} 0.0003 & 1.566 \\ 0.3454 & -2.436 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1.569 \\ 1.018 \end{bmatrix}$$

έχει λύση  $x_1 = 10$ ,  $x_2 = 1$ . Όμως, αν υποθέσουμε H/Y με αναπαράσταση αριθμών

$$\pm 0.f_1 f_2 \dots f_n \times 10^{\pm |s|}, \quad |s| \leq 10, \quad n = 5,$$

η απλή απαλοιφή Gauss δίνει προσεγγιστικά, μετά την τριγωνοποίηση,

$$\begin{bmatrix} 0.3 \times 10^{-3} & 0.1566 \times 10^1 \\ 0 & -0.1804 \times 10^1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.1569 \times 10^1 \\ -0.1805 \times 10^0 \end{bmatrix}$$

και τότε  $x_1 = 6.868$ ,  $x_2 = 1.0006$ . Η οδήγηση με εναλλαγή γραμμών είναι απαραίτητη για να βρούμε τα ακριβή  $x_1$ ,  $x_2$ . Έτσι, αν εναλλάξουμε την πρώτη με τη δεύτερη εξίσωση, αν, δηλαδή, ξεκινήσουμε με το σύστημα

$$\begin{bmatrix} 0.3454 & -2.436 \\ 0.0003 & 1.566 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1.018 \\ 1.569 \end{bmatrix}$$

η τριγωνοποίηση δίνει

$$\begin{bmatrix} 0.3454 \times 10^0 & -0.2436 \times 10^1 \\ 0 & 0.1568 \times 10^1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.1018 \times 10^1 \\ 0.1568 \times 10^1 \end{bmatrix}.$$

Συνεπώς,  $x_1 = 10$  και  $x_2 = 1$ .



**Ολική οδήγηση (κατά γραμμές και στήλες).** Αν χρειάζεται, μπορούμε να φέρνουμε, σε κάθε επανάληψη, με κατάλληλη εναλλαγή γραμμών και στηλών, το μεγαλύτερο κατ' απόλυτη τιμή στοιχείο όλου του πίνακα στη θέση του  $a_{kk}$ . Προσέξτε ότι η εναλλαγή στηλών απαιτεί και εναλλαγή στοιχείων στο διάνυσμα  $x$ .

### 3.2.3 Μέθοδος Gauss-Jordan

Μια άλλη μέθοδος επίλυσης γραμμικών συστημάτων, παραλλαγή της μεθόδου Gauss, είναι η μέθοδος Gauss-Jordan. Σε αυτή, η διαδικασία της απαλοιφής των συντελεστών κάθε στήλης δεν περιορίζεται στις γραμμές κάτω από τη διαγώνιο αλλά εφαρμόζεται και πάνω από αυτή. Επομένως, με αυτή τη διαδικασία, ένα σύστημα της μορφής

$$A \cdot x = B$$

γίνεται

$$A' \cdot x = B',$$

όπου ο  $A'$  είναι διαγώνιος πίνακας. Με πολύ απλό μετασχηματισμό μπορεί να γίνει ο ταυτοτικός, οπότε

$$I \cdot x = B''.$$

Η μέθοδος αυτή παράγει απευθείας τη λύση του συστήματος, απαιτεί όμως περίπου 50% περισσότερες πράξεις από την τριγωνοποίηση σε συνδυασμό με την οπισθοδρόμηση, και γι' αυτό δεν θα πρέπει να χρησιμοποιείται.

### 3.2.4 Ανάλυση LU

Ας υποθέσουμε ότι ο πίνακας  $A$  στην (3.2) μπορεί να γραφεί ως εξής

$$A = L \cdot U,$$

όπου  $L$  ένας κάτω τριγωνικός πίνακας (έχει δηλαδή μη μηδενικά στοιχεία στη διαγώνιο και κάτω από αυτή) και  $U$  ένας άνω τριγωνικός πίνακας (έχει δηλαδή μη μηδενικά στοιχεία στη διαγώνιο και πάνω από αυτή):

$$L = \begin{bmatrix} \ell_{11} & 0 & 0 & \cdots & 0 \\ \ell_{21} & \ell_{22} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \cdots & \ell_{nn} \end{bmatrix}$$

και

$$U = \begin{bmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & u_{nn} \end{bmatrix}.$$

Η εξίσωση (3.2) μπορεί να γραφεί

$$A \cdot x = b \Rightarrow L \cdot (U \cdot x) = b .$$

Το διάνυσμα  $y = U \cdot x$  μπορεί να προσδιοριστεί λύνοντας την εξίσωση  $L \cdot y = b$ . Καθώς ο πίνακας  $L$  είναι κάτω τριγωνικός, η λύση εύκολα δείχνεται ότι είναι

$$y_i = \frac{1}{\ell_{ii}} \left( b_i - \sum_{j=1}^{i-1} \ell_{ij} y_j \right) , \quad i = 1, 2, \dots, n .$$

Στον υπολογισμό του αθροίσματος χρησιμοποιούμε την ακόλουθη σύμβαση: όταν το κάτω όριο του δείκτη άθροισης είναι μεγαλύτερο από το άνω (επομένως, στην περίπτωση μας, όταν  $i = 1$ ), το άθροισμα είναι 0.

Αφού προσδιορίσουμε το διάνυσμα  $y$ , η επίλυση του άνω τριγωνικού συστήματος  $U \cdot x = y$  σύμφωνα με την §3.2.2 θα μας δώσει τη λύση του αρχικού. Επομένως

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

Νοείται ότι όταν  $i = n$  το άθροισμα είναι 0.

Κάθε αντιστρέψιμος πίνακας μπορεί να αναλυθεί σε γινόμενο δύο τριγωνικών πινάκων  $L, U$  αρκεί να είναι μη μηδενικές οι ορίζουσες όλων των τετραγωνικών υπο-πινάκων του με πάνω αριστερό στοιχείο το  $(1, 1)$  και κάτω δεξιό το  $(i, i)$ , για κάθε  $i = 1, 2, \dots, n$ . Αν δεν ισχύει κάτι τέτοιο, μπορούμε πάντα να εναλλάξουμε τις γραμμές του πίνακα ώστε οι ορίζουσες να γίνουν μη μηδενικές. Σε αυτή την περίπτωση η ανάλυση  $LU$  θα αφορά τον τροποποιημένο πίνακα και θα πρέπει να «μεταφέρουμε» τις εναλλαγές και στο διάνυσμα  $b$ .

### Αλγόριθμος του Crout για τον προσδιορισμό των $L, U$

Η εξίσωση  $A = L \cdot U$  καταλήγει στις ακόλουθες σχέσεις που συνδέουν τα στοιχεία των  $A, L, U$ :

$$\begin{aligned} i < j & : a_{ij} = \ell_{i1}u_{1j} + \ell_{i2}u_{2j} + \dots + \ell_{ii}u_{ij} , \\ i = j & : a_{ij} = \ell_{i1}u_{1j} + \ell_{i2}u_{2j} + \dots + \ell_{ii}u_{jj} , \\ i > j & : a_{ij} = \ell_{i1}u_{1j} + \ell_{i2}u_{2j} + \dots + \ell_{ij}u_{jj} . \end{aligned}$$

Συνολικά έχουμε  $n^2$  μη γραμμικές εξισώσεις με  $n^2 + n$  αγνώστους. Μπορούμε αυθαίρετα να δώσουμε μη μηδενικές τιμές σε  $n$  από τους αγνώστους και να λύσουμε το σύστημα για να προσδιορίσουμε τους υπόλοιπους. Ας ορίσουμε ότι τα διαγώνια στοιχεία του  $U$  είναι ίσα με 1. Μπορούμε να αποθηκεύσουμε τα υπόλοιπα στοιχεία των  $L, U$  μαζί, σε ένα πίνακα με τη μορφή

$$\begin{bmatrix} \ell_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ \ell_{21} & \ell_{22} & u_{23} & \cdots & u_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \cdots & \ell_{nn} \end{bmatrix} .$$

Ο αλγόριθμος του *Crout* προσδιορίζει σταδιακά τις γραμμές και στήλες αυτού του πίνακα ως εξής:

1. Ορίζουμε  $u_{ii} = 1$  για  $i = 1, 2, \dots, n$ .

2. Για κάθε  $j = 1, 2, \dots, n$  υπολογίζουμε

(α') τα  $\ell_{ij}$  διαδοχικά με  $i = j, j+1, \dots, n$  από τη σχέση

$$\ell_{ij} = a_{ij} - \sum_{k=1}^{j-1} \ell_{ik} u_{kj} ,$$

(β') τα  $u_{ji}$  με  $i = j+1, \dots, n$  από τη σχέση

$$u_{ji} = \frac{1}{\ell_{jj}} \left( a_{ji} - \sum_{k=1}^{j-1} \ell_{jk} u_{ki} \right) .$$

Στην περίπτωση που κάποιο από τα  $\ell_{jj}$  είναι 0, χρειάζεται να γίνει εναλλαγή γραμμών. Εναλλαγή πρέπει επίσης να γίνει αν επιθυμούμε να έχουμε υψηλή ακρίβεια: πρέπει να φέρουμε στη διαγώνιο τη μεγαλύτερη ποσότητα (κατ' απόλυτη τιμή). Δεν θα αναφερθούμε περισσότερο στο πώς γίνεται αυτό.

Οι πράξεις που απαιτούνται για την εύρεση της λύσης ενός γραμμικού συστήματος  $n \times n$  μέσω της ανάλυσης  $LU$  είναι περίπου  $2n^3/3$ , όσες περίπου και στη μέθοδο απαλοιφής Gauss, ενώ οι απαιτήσεις μνήμης είναι  $n^2$  πραγματικοί (ή μιγαδικοί) αριθμοί και  $n$  ακέραιοι που θα καταγράφουν τις εναλλαγές γραμμών.

Τα πλεονεκτήματα της μεθόδου  $LU$  είναι ότι δεν τροποποιεί τους πίνακες  $A$ ,  $b$  και αφού προσδιοριστεί η ανάλυση του πίνακα  $A$  σε  $L$ ,  $U$ , μπορεί να εφαρμοστεί για να επιλυθεί γρήγορα το γραμμικό σύστημα  $A \cdot x = b$  με πολλαπλά δεξιά μέλη, να βρεθεί εύκολα ο αντίστροφος  $A^{-1}$ , να υπολογιστεί η ορίζουσα του  $A$ , κλπ. όπως θα δούμε παρακάτω στις εφαρμογές. Συνεπώς, η ανάλυση ενός πίνακα σε γινόμενο  $L$ ,  $U$  είναι προτιμότερη και χρησιμοποιείται περισσότερο από τη διαδικασία απαλοιφής Gauss.

**Σχέση με την απαλοιφή Gauss** Μπορεί ναδειχθεί ότι η ανάλυση ενός πίνακα σε  $L$ ,  $U$  με τον αλγόριθμο του Crout μπορεί να προκύψει κατά την απαλοιφή Gauss. Αν π.χ. θέλουμε να αναλύσουμε τον πίνακα

$$A = \begin{bmatrix} 4 & 0 & 1 \\ 2 & 1 & 0 \\ 2 & 2 & 3 \end{bmatrix} ,$$

τον γράφουμε ως γινόμενο δύο πινάκων, με τον ένα από αυτούς να είναι αρχικά ο ταυτοτικός:

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 4 & 0 & 1 \\ 2 & 1 & 0 \\ 2 & 2 & 3 \end{bmatrix} .$$

Εκτελούμε διαδοχικά κατάλληλους μετασχηματισμούς στον δεξιό πίνακα ώστε να γίνει άνω τριγωνικός (εφαρμόζουμε δηλαδή την απαλοιφή Gauss):

Διαιρούμε την πρώτη γραμμή με το 4 (ώστε να έχουμε 1 στη διαγώνιο) και θέτουμε αυτήν την τιμή στη θέση (1,1) του αριστερού πίνακα. Το γινόμενο των πινάκων παραμένει  $A$ :

$$A = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 1/4 \\ 2 & 1 & 0 \\ 2 & 2 & 3 \end{bmatrix}.$$

Για να μηδενίσουμε το δεύτερο στοιχείο της πρώτης στήλης του δεξιού πίνακα πολλαπλασιάζουμε την πρώτη γραμμή του με το 2 και την αφαιρούμε από τη δεύτερη. Συγχρόνως, τοποθετούμε τον παράγοντα 2 στο δεύτερο στοιχείο της πρώτης στήλης του αριστερού. Το ίδιο κάνουμε και για να μηδενίσουμε το στοιχείο (3,1) του δεξιού πίνακα. Καταλήγουμε στη σχέση

$$A = \begin{bmatrix} 4 & 0 & 0 \\ 2 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 1/4 \\ 0 & 1 & -1/2 \\ 0 & 2 & 5/2 \end{bmatrix}.$$

Το γινόμενο παραμένει  $A$ .

Το επόμενο βήμα για να εμφανιστεί 1 στο στοιχείο (2,2) είναι η διαίρεση της δεύτερης γραμμής με το στοιχείο στη θέση (2,2), εδώ 1. Τοποθετούμε αυτόν τον παράγοντα στη θέση (2,2) του αριστερού πίνακα. Κατόπιν, πολλαπλασιάζουμε τη δεύτερη γραμμή του δεξιού με το 2 και την αφαιρούμε από την τρίτη. Συγχρόνως, γράφουμε τον παράγοντα αυτόν στη θέση (3,2) του αριστερού:

$$A = \begin{bmatrix} 4 & 0 & 0 \\ 2 & 1 & 0 \\ 2 & 2 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 1/4 \\ 0 & 1 & -1/2 \\ 0 & 0 & 7/2 \end{bmatrix}.$$

Τελικά, διαιρούμε την τρίτη γραμμή του δεξιού πίνακα με το στοιχείο στη θέση (3,3) και τοποθετούμε τον παράγοντα αυτόν (7/2) στη θέση (3,3) του αριστερού:

$$A = \begin{bmatrix} 4 & 0 & 0 \\ 2 & 1 & 0 \\ 2 & 2 & 7/2 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 1/4 \\ 0 & 1 & -1/2 \\ 0 & 0 & 1 \end{bmatrix}.$$

Το γινόμενο σε όλα τα στάδια είναι  $A$  και οι πίνακες, αριστερός και δεξιός, που καταλήξαμε είναι κάτω και άνω τριγωνικός αντίστοιχα.

### 3.3 Επαναληπτικές Μέθοδοι

Σε αυτή την κατηγορία μεθόδων ξεκινάμε από μια αρχική προσέγγιση της λύσης,  $x^{(0)}$ , και παράγουμε μια ακολουθία καλύτερων προσεγγίσεων  $x^{(1)}, x^{(2)}, \dots$  η οποία συγκλίνει στη λύση σε άπειρες επαναλήψεις. Στην πράξη, μια προσέγγιση  $x^{(k)}$  είναι ικανοποιητική όταν

- το διάνυσμα  $Ax^{(k)} - b$  έχει «μικρό» μέτρο ή «μικρά» (κατ' απόλυτη τιμή) στοιχεία.
- Η διαφορά (ή η σχετική διαφορά) των  $x^{(k+1)}$  και  $x^{(k)}$  έχει «μικρό» μέτρο ή «μικρά» (κατ' απόλυτη τιμή) στοιχεία.

### 3.3.1 Στατικές μέθοδοι

Οι επαναληπτικές μέθοδοι στις οποίες ο υπολογισμός της προσέγγισης  $x^{(k)}$  γίνεται με τον ίδιο ακριβώς τρόπο ανεξάρτητα από το  $k$ , χαρακτηρίζονται ως *στατικές*. Θα παρουσιάσουμε κάποιες από αυτές.

Ένα σύστημα  $n$  γραμμικών εξισώσεων, (3.1), για το οποίο ισχύει ότι

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

και για ένα τουλάχιστον  $i$  ισχύει η αυστηρή ανισότητα,

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

έχει, δηλαδή, «κυρίαρχη» διαγώνιο, μπορεί να επιλυθεί χωρίς να τροποποιηθεί, ως εξής:

Καταρχάς, λύνοντας προς  $x_i$  φέρνουμε το σύστημα (3.1) στη μορφή

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

Νοείται ότι όταν  $i = 1$  το πρώτο άθροισμα είναι 0 και όταν  $i = n$  το δεύτερο άθροισμα είναι 0.

Παρατηρήστε ότι για να υπολογίσουμε το  $x_i$  χρειαζόμαστε τις τιμές όλων των  $x_j$  με  $j \neq i$ . Κατόπιν, εφαρμόζουμε μία από τις ακόλουθες παραλλαγές:

#### Jacobi

Σε αυτήν την παραλλαγή, οι «παλαιές» τιμές για τα  $x_i$  (δηλαδή της προηγούμενης επανάληψης,  $x_i^{(k)}$ ), χρησιμοποιούνται για να υπολογιστούν οι «νέες»,  $x_i^{(k+1)}$ :

$$\begin{aligned} x_i^{(k+1)} &= \frac{1}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)} \right) \\ &= x_i^{(k)} + \frac{1}{a_{ii}} \left( b_i - \sum_{j=1}^n a_{ij}x_j^{(k)} \right), \quad i = 1, \dots, n. \end{aligned} \quad (3.10)$$

**Gauss–Seidel**

Στη δεύτερη παραλλαγή, οι «νέες» τιμές των  $x_i$ ,  $x_i^{(k+1)}$ , χρησιμοποιούνται στον τύπο αμέσως μόλις υπολογιστούν:

$$\begin{aligned} x_i^{(k+1)} &= \frac{1}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij} x_j^{(k)} \right) \\ &= x_i^{(k)} + \frac{1}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i}^n a_{ij} x_j^{(k)} \right), \quad i = 1, \dots, n. \end{aligned} \quad (3.11)$$

Προσέξτε ότι ο υπολογισμός του  $x_i^{(k+1)}$  χρειάζεται τις «νέες» τιμές  $x_j^{(k+1)}$  για  $j < i$  και τις «παλαιές» τιμές  $x_j^{(k)}$  για  $j > i$ .

**Παρατήρηση:** Η μέθοδος Gauss–Seidel μπορεί να εφαρμοστεί (και να συγκλίνει οπωσδήποτε), εκτός από τα γραμμικά συστήματα με κυρίαρχη διαγώνιο, και σε συστήματα στα οποία ο πίνακας των συντελεστών είναι συμμετρικός θετικά ορισμένος (§3.1.2). Παρατηρήστε ότι αν  $A$  είναι ένας γενικός αντιστρέψιμος πίνακας, ο πίνακας  $A^T \cdot A$  είναι συμμετρικός θετικά ορισμένος: ισχύει  $(A^T \cdot A)^T = A^T \cdot A$  και

$$x^T \cdot A^T \cdot A \cdot x = (A \cdot x)^T (A \cdot x) = \|A \cdot x\|^2 > 0.$$

Επομένως, ένα γενικό σύστημα  $A \cdot x = b$  μπορεί να μετατραπεί στο  $(A^T \cdot A) \cdot x = A^T \cdot b$  και να επιλυθεί με τη μέθοδο Gauss–Seidel (με συνολικά περισσότερες πράξεις από την απαλοιφή Gauss).

**Successive overrelaxation (SOR)**

Στη μέθοδο αυτή, υπολογίζουμε σε κάθε επανάληψη τη νέα προσέγγιση με τη μέθοδο Gauss–Seidel,  $\bar{x}_i^{(k+1)}$ , αλλά η βελτίωση που κάνουμε τελικά είναι ένα ποσοστό της βελτίωσης που προβλέπει η Gauss–Seidel:

$$x_i^{(k+1)} = x_i^{(k)} + \omega \left( \bar{x}_i^{(k+1)} - x_i^{(k)} \right). \quad (3.12)$$

Αν  $\omega = 1$  η μέθοδος SOR καταλήγει στη μέθοδο Gauss–Seidel. Μπορεί ναδειχθεί ότι ο συντελεστής  $\omega$  πρέπει να είναι στο διάστημα  $(0, 2)$  για να υπάρχει δυνατότητα σύγκλισης. Αν ο πίνακας των συντελεστών είναι συμμετρικός θετικά ορισμένος τότε η μέθοδος SOR συγκλίνει με οποιαδήποτε τιμή του  $\omega$  στο  $(0, 2)$  (αλλά με διαφορετική ταχύτητα σύγκλισης). Γενικά, μια τιμή  $\omega > 1$  δίνει στην μέθοδο μεγαλύτερη ταχύτητα σύγκλισης από την Gauss–Seidel, ενώ, αν η Gauss–Seidel δεν συγκλίνει, μπορεί η μέθοδος SOR να δώσει λύση με κάποιο  $\omega < 1$ .

### 3.3.2 Μέθοδοι προβολής

#### Kaczmarz

Μια επαναληπτική μέθοδος που βασίζεται στην προβολή ενός σημείου του  $n$ -διάστατου χώρου σε ακολουθία «επιπέδων» του χώρου αυτού είναι η μέθοδος Kaczmarz. Σύμφωνα με αυτή, η ορθογώνια προβολή μιας προσεγγιστικής λύσης του συστήματος,  $x^{(k)}$ , πάνω στο υπερ-επίπεδο  $\sum_{j=1}^n a_{pj}x_j = b_p$  παράγει τη διόρθωση στην προσέγγιση  $x^{(k+1)}$ . Επομένως, η συνιστώσα  $i$  της νέας προσέγγισης δίνεται από τον τύπο

$$x_i^{(k+1)} = x_i^{(k)} + \frac{b_p - \sum_{j=1}^n a_{pj}x_j^{(k)}}{\sum_{j=1}^n a_{pj}^2} a_{pi}, \quad i = 1, \dots, n.$$

Το  $p$  υπερ-επίπεδο συνήθως επιλέγεται να είναι διαδοχικά το πρώτο, το δεύτερο, κλπ., συνεπώς  $p = (k \bmod n) + 1$ . Η επανάληψη σταματά στην τιμή του  $k$  για την οποία τα κριτήρια σύγκλισης ικανοποιούνται.

Η μέθοδος Kaczmarz μπορεί να εφαρμοστεί σε οποιοδήποτε σύστημα έχει λύση αλλά η σύγκλιση στη λύση είναι πολύ αργή.

## 3.4 Εφαρμογές

Η διαδικασία τριγωνοποίησης με τη μέθοδο απαλοιφής Gauss (§3.2.2) και γενικότερα, η επίλυση γραμμικών συστημάτων, βρίσκει εφαρμογή και σε άλλα προβλήματα γραμμικής άλγεβρας.

### 3.4.1 Υπολογισμός του αντίστροφου πίνακα

Κάθε μέθοδος επίλυσης γραμμικού συστήματος της μορφής  $A \cdot x = b$  παράγει τελικά το

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

Συνεπώς, αν επιλέξουμε για διάνυσμα  $b$  διαδοχικά τα  $n$  διανύσματα

$$b_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, b_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, b_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \dots, b_n = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix},$$

θα έχουμε ως λύσεις τις αντίστοιχες στήλες του πίνακα  $A^{-1}$ . Η μέθοδος αυτή για την εύρεση του αντιστρόφου ενός (τετραγωνικού) πίνακα  $A_{n \times n}$  απαιτεί την επίλυση  $n$  γραμμικών συστημάτων με διαφορετικά δεξιά μέλη. Αν επιλέξουμε για την επίλυσή τους τη διαδικασία της τριγωνοποίησης, οποιαδήποτε μεταβολή των συστημάτων καθορίζεται αποκλειστικά από τα στοιχεία του  $A$  και, συνεπώς, μπορούν να επιλυθούν ταυτόχρονα.

**Παράδειγμα**

Ο αντίστροφος του πίνακα

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 2 & 0 \end{bmatrix}$$

βρίσκεται ως εξής:

Συμπληρώνουμε τον πίνακα  $A$  με τις στήλες του ταυτοτικού πίνακα:

$$\left[ \begin{array}{cccc|cccc} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 2 & 0 & 0 & 0 & 0 & 1 \end{array} \right].$$

Εκτελούμε την τριγωνοποίηση στο αριστερό μέρος και, συγχρόνως, εφαρμόζουμε τις εναλλαγές, προσθέσεις και πολλαπλασιασμούς γραμμών που υπαγορεύονται από την τριγωνοποίηση, στο δεξί μέρος:

1. Επιδιώκουμε να μηδενίσουμε τους συντελεστές της πρώτης στήλης κάτω από τη διαγώνιο. Στο συγκεκριμένο πίνακα τα  $a_{21}$  και  $a_{31}$  είναι ήδη 0. Αφαιρούμε την πρώτη από την τέταρτη γραμμή:

$$\left[ \begin{array}{cccc|cccc} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & -1 & 2 & 0 & -1 & 0 & 0 & 1 \end{array} \right].$$

2. Προχωρούμε στη δεύτερη στήλη. Εναλλάσσουμε τη δεύτερη γραμμή με την τρίτη ώστε να έρθει στη διαγώνιο μη μηδενικό στοιχείο:

$$\left[ \begin{array}{cccc|cccc} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 2 & 0 & -1 & 0 & 0 & 1 \end{array} \right].$$

3. Προσθέτουμε τη δεύτερη στην τέταρτη γραμμή:

$$\left[ \begin{array}{cccc|cccc} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 1 & -1 & 0 & 1 & 1 \end{array} \right].$$



4. Προχωρούμε στην τρίτη στήλη: Πολλαπλασιάζουμε την τρίτη γραμμή με 2 και την αφαιρούμε από την τέταρτη:

$$\left[ \begin{array}{cccc|cccc} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & -2 & 1 & 1 \end{array} \right] .$$

Η άνω τριγωνοποίηση του αριστερού τμήματος ολοκληρώθηκε.

Αν θέσουμε

$$A' = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & -1 \end{bmatrix} ,$$

έχουμε καταλήξει σε 4 συστήματα με διαφορετικά δεξιά μέλη:

$$A' \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \end{bmatrix} , \quad A' \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ -2 \end{bmatrix} , \quad A' \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} , \quad A' \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} .$$

Προχωρούμε στην οπισθοδρόμηση (§3.2.2) για κάθε σύστημα χωριστά. Οι λύσεις τους είναι

$$X_1 = \begin{bmatrix} 2 \\ -1 \\ -1 \\ 1 \end{bmatrix} , \quad X_2 = \begin{bmatrix} 2 \\ -2 \\ -1 \\ 2 \end{bmatrix} , \quad X_3 = \begin{bmatrix} -2 \\ 2 \\ 1 \\ -1 \end{bmatrix} , \quad X_4 = \begin{bmatrix} -1 \\ 1 \\ 1 \\ -1 \end{bmatrix} .$$

Τα  $X_i$  είναι οι στήλες του αντίστροφου πίνακα του  $A$ , επομένως

$$A^{-1} = \begin{bmatrix} 2 & 2 & -2 & -1 \\ -1 & -2 & 2 & 1 \\ -1 & -1 & 1 & 1 \\ 1 & 2 & -1 & -1 \end{bmatrix} .$$

### 3.4.2 Υπολογισμός ορίζουσας

Μια μέθοδος για να υπολογίσουμε την ορίζουσα, εκτός από την εφαρμογή του τύπου (3.4), ξεκινά με τη μετατροπή του αρχικού πίνακα σε ένα τριγωνικό (άνω ή κάτω). Η πρόσθεση σε μία γραμμή ενός (τετραγωνικού) πίνακα του πολλαπλασίου μίας άλλης είναι διαδικασία που διατηρεί την ορίζουσα. Ο άνω τριγωνικός πίνακας που παράγεται με την απαλοιφή Gauss, αν η τριγωνοποίηση περιοριστεί μόνο σε τέτοιες μεταβολές, έχει ίδια ορίζουσα με τον αρχικό. Προσέξτε ότι σε περι-

πτωση που εφαρμόσουμε οδήγηση (δηλαδή εναλλαγή γραμμών ή στηλών), πρέπει να λάβουμε υπόψη ότι κάθε τέτοια μεταβολή αλλάζει το πρόσημο της ορίζουσας.

Η ορίζουσα ενός άνω ή κάτω τριγωνικού πίνακα υπολογίζεται πολύ εύκολα. Η εφαρμογή της σχέσης (3.4), με ανάπτυξη κατά την πρώτη στήλη, δίνει ως ορίζουσα το γινόμενο των διαγωνίων στοιχείων του:

$$\det A = \prod_{i=1}^n a'_{ii}.$$

Προσέξτε ότι τα στοιχεία  $a'_{ii}$  είναι τα διαγώνια στοιχεία του τριγωνικού πίνακα.

Επομένως, η ορίζουσα  $\det A$  μπορεί να υπολογιστεί ως το γινόμενο των στοιχείων της διαγωνίου του τελικού πίνακα (μετά την άνω ή κάτω τριγωνοποίηση),  $A'$ , επί  $(-1)^s$  όπου  $s$  είναι ο συνολικός αριθμός εναλλαγών γραμμών (ή στηλών) που έγιναν κατά την απαλοιφή.

Αν έχουμε ήδη υπολογίσει την ανάλυση του  $A$  σε  $L$ ,  $U$  (§3.2.4), η ορίζουσα υπολογίζεται εύκολα

$$\det A = \det L \det U = \left( \prod_{i=1}^n \ell_{ii} \right) \left( \prod_{i=1}^n u_{ii} \right).$$

### 3.4.3 Εύρεση ιδιοτιμών και ιδιοδιανυσμάτων

Η (3.3) μπορεί να γραφεί ως εξής

$$A \cdot x = \lambda x \Rightarrow A \cdot x = \lambda I \cdot x \Rightarrow (A - \lambda I) \cdot x = 0.$$

Το σύστημα έχει μοναδική λύση, την  $x = [0, 0, \dots, 0]^T$ , αν και μόνο αν ο πίνακας  $A - \lambda I$  αντιστρέφεται. Καθώς δεν ενδιαφερόμαστε για τη μηδενική λύση, οδηγούμαστε στην απαίτηση να ισχύει  $\det(A - \lambda I) = 0$ . Παρατηρήστε ότι η έκφραση  $\det(A - \lambda I)$  είναι ένα πολυώνυμο βαθμού  $n$  ως προς  $\lambda$ , όπου  $n$  είναι η διάσταση του πίνακα  $A$ , και ονομάζεται *χαρακτηριστικό πολυώνυμο* του  $A$ . Η εύρεση των  $n$  (γενικά μιγαδικών) ριζών του μπορεί να γίνει αναλυτικά (για  $n < 5$ ) ή, γενικότερα, αριθμητικά με τις μεθόδους που περιγράψαμε στο Κεφάλαιο 2.

Αφού προσδιοριστούν οι ιδιοτιμές, η επίλυση του γραμμικού συστήματος  $\det(A - \lambda I)x = 0$  μπορεί να γίνει με την απαλοιφή Gauss. Προσέξτε ότι το σύστημα έχει άπειρες λύσεις, οπότε τουλάχιστον μία από τις συνιστώσες του διανύσματος  $x$  είναι «ελεύθερη». Κατά την τριγωνοποίηση του πίνακα  $A - \lambda I$  η τελευταία γραμμή (αν γίνει κάτω τριγωνικός) θα έχει όλα τα στοιχεία της, και το διαγώνιο, ίσα με 0. Καθώς το σταθερό διάνυσμα στο δεξιό μέλος του συστήματος είναι 0, η τελευταία συνιστώσα του  $x$  δεν μπορεί να προσδιοριστεί. Η συγκεκριμένη συνιστώσα μπορεί να τεθεί αυθαίρετα 1 και να προχωρήσουμε στην επίλυση του συστήματος. Εύκολα μπορούμε να δούμε ότι όλες οι συνιστώσες του  $x$  θα προκύψουν ανάλογες της τελευταίας συνιστώσας του.

Αφού προσδιοριστεί το διάνυσμα  $x$  με την αυθαίρετη επιλογή της μίας συνιστώσας του, μπορούμε να το κανονικοποιήσουμε, να το διαιρέσουμε δηλαδή, με το μέτρο του. Το κανονικοποιημένο διάνυσμα αποτελεί τη βάση του χώρου των ιδιοδιανυσμάτων που αντιστοιχούν στη συγκεκριμένη ιδιοτιμή: κάθε πολλαπλάσιό του αποτελεί ιδιοδιάνυσμα του πίνακα  $A$ .

Στην περίπτωση που εμφανιστούν στον πίνακα  $A - \lambda I$ , μετά την τριγωνοποίησή του, δύο (ή περισσότερες) γραμμές με όλα τα στοιχεία τους ίσα με 0, θα έχουμε διπλή (ή πολλαπλή) ελευθερία στην επιλογή των συνιστωσών του  $x$ . Μπορούμε να προσδιορίσουμε τα δύο (ή περισσότερα) διανύσματα βάσης του χώρου των ιδιοδιανυσμάτων που αντιστοιχούν στην ιδιοτιμή  $\lambda$  ως εξής: επιλέγουμε τις ελεύθερες συνιστώσες ώστε, διαδοχικά, μία από αυτές να είναι 1 και οι υπόλοιπες 0. Ο προσδιορισμός των υπόλοιπων συνιστωσών του  $x$  θα μας δώσει διαδοχικά τα διανύσματα βάσης (τα οποία θα πρέπει να κανονικοποιηθούν). Οποιοσδήποτε γραμμικός συνδυασμός αυτών αποτελεί ιδιοδιάνυσμα του  $A$ .

Κατά την αναζήτηση των ιδιοτιμών μπορεί να φανούν χρήσιμα τα παρακάτω θεωρήματα:

### Θεώρημα κύκλων του Gershgorin

Για ένα πίνακα  $A_{n \times n}$ , ορίζουμε ως *κυκλικό δίσκο του Gershgorin* την περιοχή, στο μιγαδικό επίπεδο, εντός του κύκλου με κέντρο ένα διαγώνιο στοιχείο  $a_{ii}$  και ακτίνα  $R_i$  το άθροισμα των απόλυτων τιμών των στοιχείων της γραμμής  $i$ , εκτός του διαγώνιου:

$$R_i = \sum_{\substack{j=1 \\ i \neq j}}^n |a_{ij}| .$$

Σύμφωνα με το Θεώρημα κύκλων του Gershgorin, κάθε ιδιοτιμή  $\lambda$  του πίνακα  $A$  βρίσκεται εντός τουλάχιστον ενός κύκλου Gershgorin, δηλαδή ικανοποιεί τουλάχιστον μία από τις σχέσεις

$$|\lambda - a_{ii}| \leq R_i , \quad i = 1, \dots, n .$$

Προσέξτε ότι το θεώρημα δεν εξασφαλίζει ότι *κάθε* τέτοιος κυκλικός δίσκος περιέχει μία ιδιοτιμή.

Μία βελτίωση του θεωρήματος εξασφαλίζει το εξής: αν  $m$  δίσκοι Gerschgorin επικαλύπτονται μεταξύ τους και είναι απομονωμένοι από τους υπόλοιπους δίσκους, περιέχουν ακριβώς  $m$  ιδιοτιμές (κάπου στην επιφάνεια που καλύπτουν).

Καθώς ο ανάστροφος πίνακας του  $A$  έχει τις ίδιες ιδιοτιμές με τον  $A$ , το θεώρημα εξασφαλίζει ότι κάθε ιδιοτιμή θα ικανοποιεί και μία τουλάχιστον από τις σχέσεις

$$|\lambda - a_{jj}| \leq \sum_{\substack{i=1 \\ i \neq j}}^n |a_{ij}| , \quad j = 1, \dots, n .$$

Παρατηρήστε ότι για πραγματικό πίνακα με κυρίαρχη διαγώνιο και θετικά διαγώνια στοιχεία οι κύκλοι Gershgorin βρίσκονται εξ ολοκλήρου στο μιγαδικό ημιεπίπεδο με θετικό πραγματικό μέρος. Αν επιπλέον ο πίνακας είναι συμμετρικός, οι ιδιοτιμές του είναι θετικοί πραγματικοί αριθμοί. Αντίστοιχα ισχύουν αν τα διαγώνια στοιχεία είναι αρνητικά.

### Θεώρημα Perron–Frobenius

Σύμφωνα με το θεώρημα Perron–Frobenius, ένας πραγματικός πίνακας με θετικά όλα τα στοιχεία του έχει μία θετική (πραγματική) ιδιοτιμή  $\lambda_1$  με πολλαπλότητα 1 και όλες τις υπόλοιπες, γενικά μιγαδικές, με μέτρο μικρότερο από  $\lambda_1$ .

### Ανισότητες του Schur

Οι ιδιοτιμές  $\lambda_i$  ενός πραγματικού ή μιγαδικού πίνακα  $A_{n \times n}$  με στοιχεία  $a_{ij}$  ικανοποιούν τις σχέσεις

$$\begin{aligned} \sum_{i=1}^n |\lambda_i|^2 &\leq \sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^2, \\ \sum_{i=1}^n |\operatorname{Re}(\lambda_i)|^2 &\leq \sum_{i=1}^n \sum_{j=1}^n |a_{ij} + a_{ji}^*|^2, \\ \sum_{i=1}^n |\operatorname{Im}(\lambda_i)|^2 &\leq \sum_{i=1}^n \sum_{j=1}^n |a_{ij} - a_{ji}^*|^2. \end{aligned}$$

#### 3.4.4 Επίλυση συστήματος μη γραμμικών εξισώσεων

Το γενικό πρόβλημα της εύρεσης των τιμών για τα  $x_1, x_2, \dots, x_n$  που ικανοποιούν ταυτόχρονα τις εξισώσεις

$$\begin{aligned} f_1(x_1, x_2, \dots, x_n) &= 0 \\ f_2(x_1, x_2, \dots, x_n) &= 0 \\ &\vdots \\ f_n(x_1, x_2, \dots, x_n) &= 0 \end{aligned} \tag{3.13}$$

μπορεί να αναχθεί σε επίλυση πολλών γραμμικών συστημάτων.

Σύμφωνα με το ανάπτυγμα Taylor για μια συνάρτηση πολλών μεταβλητών, αν γνωρίζουμε την τιμή της συνάρτησης  $f$  και των παραγώγων της σε ένα σημείο  $\mathbf{a} \equiv (a_1, a_2, \dots, a_n)$ , μπορούμε να υπολογίσουμε την τιμή της σε άλλο σημείο  $\mathbf{x} \equiv (x_1, x_2, \dots, x_n)$  (αρκεί η  $f$  να είναι συνεχής και παραγωγίσιμη σε πεδίο που

περιλαμβάνει τα  $\mathbf{a}$  και  $\mathbf{x}$ ) ως εξής

$$\begin{aligned} f(x_1, x_2, \dots, x_n) &= f(a_1, a_2, \dots, a_n) \\ &+ \sum_{i=1}^n \frac{\partial f}{\partial x_i} \Big|_{\mathbf{x}=\mathbf{a}} (x_i - a_i) \\ &+ \frac{1}{2!} \sum_{i=1}^n \sum_{j=1}^n \frac{\partial^2 f}{\partial x_i \partial x_j} \Big|_{\mathbf{x}=\mathbf{a}} (x_i - a_i)(x_j - a_j) + \dots \end{aligned}$$

Αν υποθέσουμε ότι τα  $\mathbf{x}$  και  $\mathbf{a}$  απέχουν «λίγο», μπορούμε να παραλείψουμε τους όρους δεύτερης τάξης και πάνω.

Η ανάπτυξη των συναρτήσεων  $f_i$  στην (3.13) δίνει

$$\begin{aligned} f_1(a_1, a_2, \dots, a_n) + \sum_{i=1}^n \frac{\partial f_1}{\partial x_i} \Big|_{\mathbf{x}=\mathbf{a}} (x_i - a_i) &\approx 0 \\ f_2(a_1, a_2, \dots, a_n) + \sum_{i=1}^n \frac{\partial f_2}{\partial x_i} \Big|_{\mathbf{x}=\mathbf{a}} (x_i - a_i) &\approx 0 \\ &\vdots \\ f_n(a_1, a_2, \dots, a_n) + \sum_{i=1}^n \frac{\partial f_n}{\partial x_i} \Big|_{\mathbf{x}=\mathbf{a}} (x_i - a_i) &\approx 0 \end{aligned} \quad (3.14)$$

Ας ορίσουμε τον πίνακα

$$\mathbf{A} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \dots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \dots & \frac{\partial f_n}{\partial x_n} \end{bmatrix},$$

με όλες τις παραγώγους να υπολογίζονται στο  $(a_1, a_2, \dots, a_n)$ , και το διάνυσμα

$$\mathbf{b} = \begin{bmatrix} f_1(\mathbf{a}) \\ f_2(\mathbf{a}) \\ \vdots \\ f_n(\mathbf{a}) \end{bmatrix}.$$

Υπενθυμίζουμε ότι το  $\mathbf{x}$  είναι το (άγνωστο) σημείο που ικανοποιεί τις εξισώσεις (3.13) και το  $\mathbf{a}$  ένα γειτονικό σημείο σε αυτό. Οι εξισώσεις (3.14) γίνονται

$$\mathbf{A} \cdot (\mathbf{a} - \mathbf{x}) \approx \mathbf{b} \Rightarrow \mathbf{x} \approx \mathbf{a} - \mathbf{A}^{-1} \cdot \mathbf{b}.$$

Η τελευταία σχέση είναι αυτή που επαναληπτικά μπορεί να μας υπολογίσει το  $\mathbf{x}$ : αν θέσουμε στο  $\mathbf{a}$  την  $k$ -οστή προσέγγιση της ρίζας,  $\mathbf{x}^{(k)}$ , με  $k = 0, 1, \dots$ , η επόμενη, πιθανόν καλύτερη, προσέγγιση  $\mathbf{x}^{(k+1)}$  είναι

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{A}^{-1} \cdot \mathbf{b}.$$

Επομένως, για να βρούμε το  $\mathbf{x}$  μπορούμε να εφαρμόσουμε την ακόλουθη μέθοδο (μέθοδος Newton–Raphson για σύστημα μη γραμμικών εξισώσεων):

#### Αλγόριθμος:

1. Επιλέγουμε μια αρχική προσέγγιση της ρίζας,  $\mathbf{x}^{(0)}$ , κοντά στην (άγνωστη) λύση.
2. Ελέγχουμε αν η τρέχουσα προσέγγιση είναι αποδεκτή ως λύση. Αν όχι, συνεχίζουμε στο επόμενο βήμα.
3. Υπολογίζουμε στην τρέχουσα προσέγγιση  $\mathbf{x}^{(k)}$  ( $k = 0, 1, \dots$ ) τον πίνακα  $\mathbf{A}$  και το διάνυσμα  $\mathbf{b}$ .
4. Αν ο πίνακας  $\mathbf{A}$  είναι αντιστρέψιμος, επιλύουμε το γραμμικό σύστημα  $\mathbf{A} \cdot \mathbf{y} = \mathbf{b}$  ως προς  $\mathbf{y}$ . Η νέα προσέγγιση είναι  $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{y}$ .
5. Επαναλαμβάνουμε από το βήμα 2.

Αποδεκτή είναι μια λύση όταν ικανοποιεί ένα τουλάχιστο από τα κριτήρια που έχουμε δει στο Κεφάλαιο 2, προσαρμοσμένα σε πολλές διαστάσεις και πολλές συναρτήσεις:

- Οι απόλυτες τιμές των συναρτήσεων  $f_i$  (τα στοιχεία δηλαδή του  $\mathbf{b}$ ) να είναι «μικρές»:  $|f_i(\mathbf{x}^{(k)})| < \varepsilon_i$ ,  $\forall i$ .
- Το μέτρο του  $\mathbf{b}^{(k)}$  να είναι «μικρό».
- Η απόλυτη βελτίωση στα  $x_i$  να είναι «μικρή» κατά μέτρο:  $|x_i^{(k)} - x_i^{(k-1)}| < \varepsilon_i$ .
- Η σχετική βελτίωση στα  $x_i$  να είναι «μικρή» κατά μέτρο:  $\left| \frac{x_i^{(k)} - x_i^{(k-1)}}{x_i^{(k)}} \right| < \varepsilon_i$  αν  $x_i^{(k)} \neq 0$ .

Στις δύο τελευταίες συνθήκες πρέπει να ελέγχουμε αν τελικά η τιμή  $\mathbf{x}^{(k)}$  ικανοποιεί το σύστημα.

Αν οι παράγωγοι των συναρτήσεων  $f_i(\mathbf{x})$  δεν είναι γνωστές αναλυτικά, μπορούμε να τις υπολογίσουμε με τους τύπους και τις τεχνικές που παρουσιάζονται στο §4.5.

### 3.5 Ασκήσεις

1. Υλοποιήστε σε πρόγραμμα την απαλοιφή Gauss. Θεωρήστε ότι τα διαγώνια στοιχεία του πίνακα είναι και παραμένουν σε όλη τη διαδικασία μη μηδενικά.

*Υπόδειξη:* Δημιουργήστε ένα πίνακα  $4 \times 4$  με τυχαία στοιχεία για να ελέγξετε το πρόγραμμά σας.

2. Υλοποιήστε την απαλοιφή Gauss με μερική οδήγηση.

*Υπόδειξη:* Τροποποιήστε το υποπρόγραμμα της απλής απαλοιφής που γράψατε στην προηγούμενη άσκηση.

3. Υλοποιήστε τη μέθοδο επίλυσης Gauss–Jordan.

4. Να γράψετε δύο υποπρογράμματα που να υλοποιούν τους αλγορίθμους Jacobi και Gauss–Seidel. Να τα χρησιμοποιήσετε για την εύρεση της λύσης του συστήματος  $A \cdot x = b$  όπου

$$A = \begin{bmatrix} 12.1 & 3.9 & 0.3 & -4.1 \\ 4.3 & -11.3 & 0.8 & 1.5 \\ 1.0 & -2.8 & 14.3 & -8.1 \\ 2.4 & 6.1 & -1.1 & 12.5 \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}, \quad b = \begin{bmatrix} 1.2 \\ 2.3 \\ 3.4 \\ 4.5 \end{bmatrix}.$$

Να χρησιμοποιήσετε το ακόλουθο κριτήριο τερματισμού των επαναλήψεων: κάθε στοιχείο του διανύσματος  $Ax - b$  να είναι κατ' απόλυτη τιμή μικρότερο του  $10^{-7}$ .

5. Υλοποιήστε σε υποπρόγραμμα τον αλγόριθμο αντιστροφής πίνακα που περιγράφηκε. Να τη χρησιμοποιήσετε για να βρείτε τον αντίστροφο του

$$\begin{bmatrix} 2.1 & 3.9 & 0.3 & -4.1 \\ 4.3 & -1.3 & 0.8 & 1.5 \\ 1.0 & -2.8 & 4.3 & -8.1 \\ 2.4 & 6.1 & -1.1 & 12.5 \end{bmatrix}.$$

6. Να γράψετε υποπρόγραμμα που να υπολογίζει την ορίζουσα ενός πίνακα. Θα δέχεται ως ορίσματα τον πίνακα και, αν σας χρειάζεται, την τάξη του και θα επιστρέφει την ορίζουσα.

Χρησιμοποιήστε το για να υπολογίσετε την ορίζουσα του

$$\begin{bmatrix} 2.1 & 3.9 & 0.3 & -4.1 \\ 4.3 & -1.3 & 0.8 & 1.5 \\ 1.0 & -2.8 & 4.3 & -8.1 \\ 2.4 & 6.1 & -1.1 & 12.5 \end{bmatrix}.$$

7. Υλοποιήστε σε πρόγραμμα τη μέθοδο Cramer.

8. Να γράψετε πρόγραμμα που να υλοποιεί τον αλγόριθμο εύρεσης ιδιοτιμών ενός τετραγωνικού πίνακα. Να βρείτε μία ιδιοτιμή του πίνακα

$$\begin{bmatrix} 2.1 & 3.9 & 0.3 & -4.1 \\ 4.3 & -1.3 & 0.8 & 1.5 \\ 1.0 & -2.8 & 4.3 & -8.1 \\ 2.4 & 6.1 & -1.1 & 12.5 \end{bmatrix}.$$

9. Λύστε το σύστημα  $A \cdot x = b$  όπου

$$A = \begin{bmatrix} 12.1 & 3.9 & 0.3 & -4.1 \\ 4.3 & -11.3 & 0.8 & 1.5 \\ 1.0 & -2.8 & 14.3 & -8.1 \\ 2.4 & 6.1 & -1.1 & 12.5 \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}, \quad b = \begin{bmatrix} 1.2 \\ 2.3 \\ 3.4 \\ 4.5 \end{bmatrix}$$

με τη μέθοδο της ανάλυσης  $LU$ .

10. Να προσδιορίσετε μία ιδιοτιμή και το αντίστοιχο ιδιοδιάνυσμα του πίνακα

$$\begin{bmatrix} 3 & 2 & -1 \\ 0 & 2 & 5 \\ 1 & 7 & 2 \end{bmatrix}.$$

11. Να βρείτε μία λύση του ακόλουθου μη γραμμικού συστήματος εξισώσεων:

$$\begin{aligned} 4x^2 - y^3 &= -28 \\ 3x^3 + 4y^2 &= 145 \end{aligned}$$



## Κεφάλαιο 4

# Προσέγγιση Συναρτήσεων

Έστω ότι γνωρίζουμε τις τιμές μιας συνάρτησης  $f(x)$ ,  $f_0, f_1, \dots, f_n$  σε σημεία  $x_0, x_1, \dots, x_n$  (με  $x_0 < x_1 < x_2 < \dots < x_n$ ), και ζητάμε να υπολογίσουμε π.χ.

- την τιμή  $f(\bar{x})$  ή την τιμή της παραγώγου  $f'(\bar{x})$  σε ένα σημείο  $\bar{x}$  ανάμεσα στα  $x_0$  και  $x_n$ ,
- το ολοκλήρωμα της  $f(x)$  σε κάποιο διάστημα (μέσα στο  $[x_0, x_n]$ ),
- τη ρίζα της  $f(x)$  στο  $[x_0, x_n]$ .

Μπορούμε να προσεγγίσουμε την άγνωστη συνάρτηση  $f(x)$  με άλλες συναρτήσεις ώστε να μπορούμε να έχουμε μια εκτίμηση για τις ζητούμενες τιμές.

Με τις μεθόδους που θα αναπτύξουμε παρακάτω μπορούμε επίσης να χειριστούμε την περίπτωση που η συνάρτηση είναι γνωστή αλλά εξαιρετικά πολύπλοκη. Έτσι, μπορούμε να υπολογίσουμε μια προσέγγιση για, π.χ., την παράγωγό της, παραγωγίζοντας την (πιο απλή) προσεγγιστική συνάρτηση.

### 4.1 Προσέγγιση με πολυώνυμο

Υπάρχει ένα και μοναδικό πολυώνυμο βαθμού  $n$ ,  $p(x)$ , που περνά από τα  $n+1$  σημεία  $(x_i, f_i)$ , δηλαδή ικανοποιεί τις σχέσεις  $p(x_i) = f_i$ . Μπορούμε να το προσδιορίσουμε σχηματίζοντας ένα πολυώνυμο με τόσους άγνωστους συντελεστές όσες οι σχέσεις που θέλουμε να ικανοποιεί. Ανάλογα με τη μορφή που θα επιλέξουμε για το πολυώνυμο, η επίλυση των σχέσεων  $p(x_i) = f_i$  μπορεί να είναι πολύπλοκη ή πολύ απλή. Έτσι:

- Αν το πολυώνυμο έχει τη γενική μορφή  $p(x) = \sum_{j=0}^n a_j x^j$ , αναπτύσσεται, δηλαδή, στα μονώνυμα  $x^j$ , προκύπτει το ακόλουθο γραμμικό σύστημα εξισώσεων με άγνωστες ποσότητες τους συντελεστές  $a_i$ :

$$\sum_{j=0}^n a_j x_i^j = f_i, \quad i = 0, 1, \dots, n. \quad (4.1)$$

Ο πίνακας  $V$  των συντελεστών των άγνωστων ποσοτήτων έχει διάσταση  $n + 1$  και στοιχεία τα  $V_{ij} = x_{i-1}^{j-1}$  με  $i, j = 0, 1, \dots, n$  και είναι ο πίνακας Vandermonde. Το σύστημα έχει μοναδική λύση και μπορεί να λυθεί με τις μεθόδους που παρουσιάστηκαν στο Κεφάλαιο 3 (π.χ. απαλοιφή Gauss), ή ειδικές μεθόδους που εκμεταλλεύονται την ειδική μορφή του (π.χ. αλγόριθμος Björck-Pereyra) και είναι πιο γρήγορες. Η λύση του μας δίνει τους συντελεστές του πολυωνύμου.

- αν το πολυώνυμο έχει τη μορφή του τύπου του Newton,

$$p(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \dots,$$

δηλαδή αναπτύσσεται στα πολυώνυμα της βάσης Newton, (4.2β'),

$$p(x) = \sum_{i=0}^n a_i q_i(x), \quad \text{όπου} \quad (4.2\alpha')$$

$$q_i(x) = \begin{cases} 1, & i = 0, \\ \prod_{j=0}^{i-1} (x - x_j), & i = 1, 2, \dots, n. \end{cases} \quad (4.2\beta')$$

προκύπτει ένα (κάτω) τριγωνικό σύστημα για τους άγνωστους συντελεστές  $a_i$ , το οποίο λύνεται εύκολα, σταδιακά:

$$\begin{aligned} p(x_0) = f_0 &\Rightarrow a_0 = f_0, \\ p(x_1) = f_1 &\Rightarrow a_1 = \frac{1}{x_1 - x_0}(f_1 - a_0), \\ &\vdots \quad \quad \quad \vdots \end{aligned}$$

ή, γενικά,

$$a_j = \begin{cases} f_0, & j = 0, \\ \frac{1}{q_j(x_j)} \left( f_j - \sum_{i=0}^{j-1} a_i q_i(x_j) \right), & j = 1, 2, \dots, n. \end{cases}$$

- αν το πολυώνυμο εκφράζεται στη βάση Lagrange, (4.3β'), η μορφή του προκύπτει απευθείας από τον τύπο Lagrange:

$$p(x) = \sum_{i=0}^n \ell_i(x) f_i, \quad \text{όπου} \quad (4.3\alpha')$$

$$\ell_i(x) = \prod_{j=0, j \neq i}^n \frac{x - x_j}{x_i - x_j}, \quad i = 0, 1, 2, \dots, n. \quad (4.3\beta')$$

Παρατηρήστε ότι  $\ell_i(x_j) = \delta_{ij}$  οπότε οι συντελεστές των πολυωνύμων της βάσης Lagrange στο ανάπτυγμα (4.3α') είναι οι τιμές  $f_i$ .

**Παράδειγμα**

Το πολυώνυμο παρεμβολής για τη συνάρτηση  $f(x) = \frac{1}{x}$  στα σημεία παρεμβολής  $x_0 = 2.0$ ,  $x_1 = 2.5$ ,  $x_2 = 4.0$ , υπολογιζόμενο από τον τύπο Lagrange, (4.3), είναι

$$\begin{aligned} p(x) &= \ell_0(x)f_0 + \ell_1(x)f_1 + \ell_2(x)f_2 \\ &= 0.5\ell_0(x) + 0.4\ell_1(x) + 0.25\ell_2(x) , \end{aligned}$$

όπου

$$\begin{aligned} \ell_0(x) &= \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{(x-2.5)(x-4)}{(2-2.5)(2-4)} = x^2 - 6.5x + 10 , \\ \ell_1(x) &= \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{(x-2)(x-4)}{(2.5-2)(2.5-4)} = -4(x^2 - 6x + 8)/3 , \\ \ell_2(x) &= \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{(x-2)(x-2.5)}{(4-2)(4-2.5)} = (x^2 - 4.5x + 5)/3 . \end{aligned}$$

Επομένως, το  $p(x) = 0.05x^2 - 0.425x + 1.15$  αποτελεί καλή προσέγγιση της  $f(x) = \frac{1}{x}$  στο διάστημα  $[2, 4]$ .

**4.1.1 Μετατροπή**

Μπορούμε να φέρουμε στη μορφή (4.1) ένα πολυώνυμο που εκφράζεται στη βάση Lagrange. Παρατηρήστε ότι κάθε πολυώνυμο  $\ell_i(x)$  της βάσης Lagrange είναι παραγοντοποιημένο: οι ρίζες του,  $\rho_k$ ,  $k = 1, \dots, n$  είναι τα  $x_j$  με  $j = 0, \dots, n$  εκτός από το  $x_i$ . Ο συντελεστής του όρου  $x^n$  στο πολυώνυμο  $\ell_i(x)$  προκύπτει εύκολα ότι είναι ο

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

ενώ οι υπόλοιποι μπορούν να υπολογιστούν από τους τύπους Vieta:

$$\begin{aligned} a_{n-1}^{(i)} &= -a_n \sum_{k_1=1}^n \rho_{k_1} , \\ a_{n-2}^{(i)} &= +a_n \sum_{k_1=1}^{n-1} \left( \rho_{k_1} \sum_{k_2=k_1+1}^n \rho_{k_2} \right) , \\ a_{n-3}^{(i)} &= -a_n \sum_{k_1=1}^{n-2} \left( \rho_{k_1} \sum_{k_2=k_1+1}^{n-1} \rho_{k_2} \left( \sum_{k_3=k_2+1}^n \rho_{k_3} \right) \right) , \\ &\vdots \\ a_0^{(i)} &= (-1)^n a_n \rho_1 \rho_2 \dots \rho_n . \end{aligned}$$

Τα αθροίσματα που εμφανίζονται, νοείται ότι έχουν τιμή 0 όταν το κάτω όριο είναι μεγαλύτερο από το πάνω όριο.

Αφού υπολογιστούν όλοι οι συντελεστές για κάθε  $i$  μπορούμε να γράψουμε την ακόλουθη σχέση για τους συντελεστές του πολυωνύμου:

$$a_k = \sum_{i=0}^n a_k^{(i)} f_i, \quad k = 0, 1, \dots, n.$$

#### 4.1.2 Σφάλμα προσέγγισης με πολυώνυμο

**Θεώρημα:** Έστω μία συνεχής συνάρτηση  $f(x)$  με  $n+1$  συνεχείς παραγώγους στο  $[a, b]$ . Έστω ακόμα ότι  $x_0 \equiv a, x_1, x_2, \dots, x_n \equiv b$  είναι  $n+1$  διαφορετικά σημεία στο διάστημα  $[a, b]$  και  $p(x)$  το πολυώνυμο παρεμβολής για τη συγκεκριμένη συνάρτηση. Τότε

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^n (x - x_i), \quad \forall x \in [a, b], \quad (4.4)$$

όπου  $\xi$  ένα σημείο στο  $(a, b)$ .

Παρατηρούμε ότι αν το διάστημα  $[a, b]$  είναι μικρό και χρησιμοποιούμε μικρού βαθμού πολυώνυμα έχουμε καλή προσέγγιση. Αντίθετα, πολυώνυμα μεγάλου βαθμού τείνουν να έχουν έντονη ταλαντωτική συμπεριφορά στα ακραία διαστήματα μεταξύ σημείων που ισαπέχουν (Φαινόμενο Runge), Σχήμα 4.1, ή εκτός του διαστήματος.

Μπορούμε να αντιμετωπίσουμε με επιτυχία το φαινόμενο Runge στην περίπτωση που έχουμε τη δυνατότητα επιλογής των σημείων παρεμβολής. Αν διαλέξουμε τα σημεία παρεμβολής ώστε να είναι πυκνά κατανομημένα στα άκρα του διαστήματος  $[a, b]$  και αραιά προς το κέντρο, το φαινόμενο ατονεί· ελαχιστοποιείται αν επιλέξουμε την κατανομή Chebyshev για τα σημεία μας:

$$x_i = \frac{b-a}{2} \cos\left(\frac{i+0.5}{n+1}\pi\right) + \frac{b+a}{2}, \quad i = 0, 1, \dots, n.$$

#### Παράδειγμα

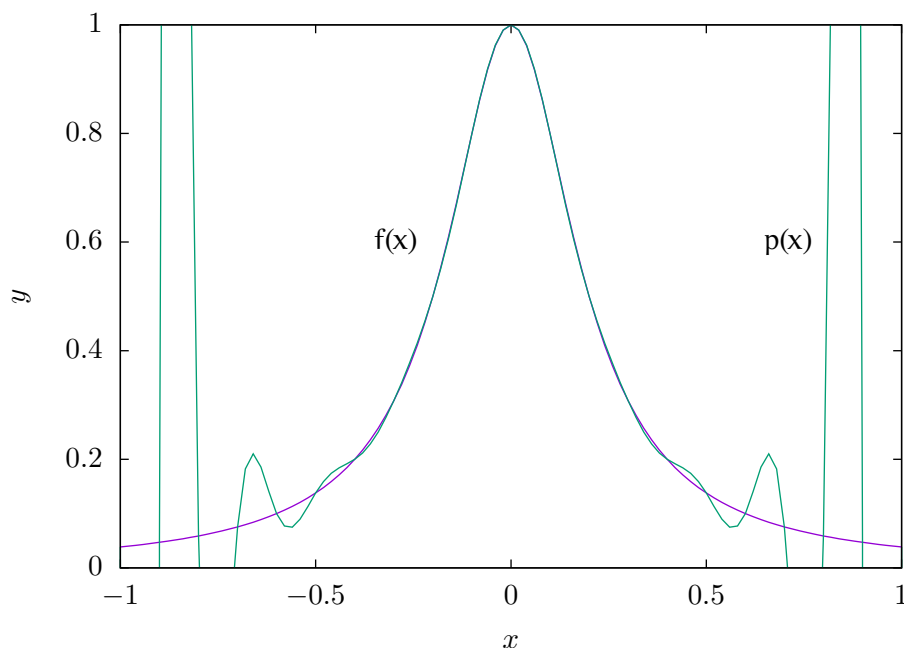
Έστω  $f(x) = e^x$ . Ζητείται να υπολογιστεί το πολυώνυμο και το σφάλμα προσέγγισης με τον τύπο Lagrange στα σημεία  $x_0, x_1$ .

Η εφαρμογή του τύπου (4.3) δίνει το πολυώνυμο

$$p(x) = e^{x_0} \frac{x - x_1}{x_0 - x_1} + e^{x_1} \frac{x - x_0}{x_1 - x_0} = x \frac{e^{x_1} - e^{x_0}}{x_1 - x_0} + \frac{x_1 e^{x_0} - x_0 e^{x_1}}{x_1 - x_0}.$$

Καθώς  $n = 1$  και  $f''(x) = e^x$ , ο τύπος (4.4) γίνεται

$$f(x) - p(x) = \frac{e^\xi}{2!} (x - x_0)(x - x_1).$$



Σχήμα 4.1: Προσέγγιση της συνάρτησης  $f(x) = 1/(1 + 25x^2)$  στο  $[-1, 1]$  με πολυώνυμο  $p(x)$  20<sup>ου</sup> βαθμού σε ισαπέχοντα σημεία. Παρατηρήστε τις έντονες ταλαντώσεις στα άκρα του διαστήματος

Η απόκλιση κατ' απόλυτη τιμή,  $|f(x) - p(x)|$ , είναι

$$|f(x) - p(x)| = \left| \frac{e^x}{2!} (x - x_0)(x - x_1) \right| = \frac{e^x}{2!} (x - x_0)(x_1 - x) .$$

Το  $e^x$  στο διάστημα  $[x_0, x_1]$  είναι παντού μικρότερο από  $e^{x_1}$ , ενώ ο όρος  $(x - x_0)(x_1 - x)$  παρουσιάζει μέγιστο στο  $x = (x_0 + x_1)/2$ , με τιμή  $(x_1 - x_0)^2/4$ . Συνεπώς

$$|f(x) - p(x)| < e^{x_1} \frac{(x_1 - x_0)^2}{8} .$$

Επομένως, για συγκεκριμένο  $x \in [x_0, x_1]$  μπορούμε να υπολογίσουμε το σφάλμα  $\varepsilon$ , ή, αντίστροφα, για συγκεκριμένο επιθυμητό σφάλμα μπορούμε να βρούμε το  $\Delta x = x_1 - x_0$  ως εξής: πρέπει

$$e^{x_1} \frac{(x_1 - x_0)^2}{8} < \varepsilon \Rightarrow (\Delta x)^2 < \frac{8\varepsilon}{e^{x_1}} .$$

Αν  $x_1 = 1.0$ ,  $\varepsilon = 10^{-6}$ , τότε  $\Delta x < 1.716 \times 10^{-3}$ .

## 4.2 Προσέγγιση με λόγο πολυωνύμων

Μία άγνωστη συνάρτηση μιας μεταβλητής,  $f(x)$ , μπορεί να προσεγγιστεί όχι μόνο από πολυώνυμο αλλά (πιο γενικά και με μεγαλύτερη ακρίβεια) και από λόγο πολυωνύμων,  $R(x)$ ,

$$R(x) = \frac{P(x)}{Q(x)}$$

όπου

$$P(x) = a_0 + \sum_{k=1}^M a_k x^k, \quad Q(x) = b_0 + \sum_{k=1}^N b_k x^k.$$

Συνολικά έχουμε  $M+N+2$  άγνωστους συντελεστές  $a_k, b_k$ . Ένας από αυτούς μπορεί αυθαίρετα να οριστεί ίσος με 1, έστω ο  $b_0$ . Μπορούμε να επιλέξουμε τους βαθμούς των πολυωνύμων  $M, N$  ώστε το πλήθος των υπόλοιπων αγνώστων,  $M+N+1$ , να είναι ίσο με το πλήθος των σημείων  $(x_i, f_i)$  στα οποία γνωρίζουμε τη συνάρτηση. Η απαίτηση να περνά η  $R(x)$  από αυτά τα σημεία δίνει ένα γραμμικό σύστημα εξισώσεων

$$f_i = R(x_i) \Rightarrow f_i Q(x_i) = P(x_i), \quad i = 1, \dots, M+N+1,$$

με άγνωστους τους  $M+N+1$  συντελεστές.

Αν γνωρίζουμε μια πολύπλοκη συνάρτηση και θέλουμε να την απλοποιήσουμε, μπορούμε εναλλακτικά να απαιτήσουμε η  $R(x) = P(x)/Q(x)$  και η  $f(x)$  να έχουν σε ένα σημείο  $x_0$ , ίδιες τιμές και ίδιες παραγώγους μέχρι και τάξης  $M+N$ , δηλαδή να ισχύει

$$\begin{aligned} R(x_0) &= f(x_0), \\ R'(x_0) &= f'(x_0), \\ &\vdots \\ R^{(M+N)}(x_0) &= f^{(M+N)}(x_0). \end{aligned}$$

Η λύση αυτού του μη γραμμικού συστήματος προσδιορίζει τους συντελεστές  $a_k, b_k$  και σχηματίζει την προσεγγιστική συνάρτηση  $R(x)$  του Padé.

## 4.3 Προσέγγιση κατά τμήματα με πολυώνυμο ελάχιστου βαθμού

Για να αποφύγουμε τις έντονες ταλαντώσεις στα άκρα του διαστήματος όταν χρησιμοποιούμε ένα πολυώνυμο, μπορούμε να προσεγγίσουμε την άγνωστη συνάρτηση κατά τμήματα, χρησιμοποιώντας πολυώνυμο του ελάχιστου δυνατού βαθμού. Αυτό σημαίνει ότι μπορούμε, π.χ., να χωρίσουμε το συνολικό διάστημα  $[x_0, x_n]$  σε τμήματα που ορίζονται από δύο διαδοχικά σημεία:  $[x_0, x_1], [x_1, x_2], \dots, [x_{n-1}, x_n]$ . Σε καθένα από αυτά τα  $n-1$  διαστήματα, προσαρμόζουμε ένα πολυώνυμο με δύο

άγνωστους συντελεστές:  $p_i(x) = a_i x + b_i$ ,  $i = 0, 1, \dots, n-1$ . Προσδιορίζουμε δηλαδή τα ευθύγραμμα τμήματα με αρχή τα  $x_i$  και τέλος τα  $x_{i+1}$ .

Με αυτό τον τρόπο, έχουμε προσδιορίσει μια συνεχή καμπύλη που προσεγγίζει την άγνωστη συνάρτηση. Όμως, οι παράγωγοι αυτής της καμπύλης είναι ασυνεχείς στα «εσωτερικά» σημεία  $x_i$ .

## 4.4 Προσέγγιση με spline

Η καμπύλη spline στα μαθηματικά είναι ένα πολυώνυμο που ορίζεται τμηματικά από πολυώνυμα χαμηλού βαθμού, αλλά όχι του ελάχιστου δυνατού, με συνεχείς παραγώγους στα σημεία (κόμβοι) που αυτά ενώνονται. Η συγκεκριμένη καμπύλη αποφεύγει να χρησιμοποιήσει πολυώνυμα μεγάλου βαθμού (άρα δεν εμφανίζονται αφύσικες ταλαντώσεις στα διαστήματα μεταξύ των σημείων). Επιπλέον, η προσέγγιση δεν γίνεται με τα πολυώνυμα ελάχιστου βαθμού και έτσι αποφεύγονται οι ασυνέχειες στις παραγώγους στους κόμβους.

Η «φυσική» κυβική spline είναι η συχνότερα χρησιμοποιούμενη καμπύλη. Την κατασκευάζουμε ως εξής:

Έστω ένα σύνολο  $n+1$  σημείων  $(x_i, f_i)$  με  $i = 0, 1, \dots, n$  και  $x_i < x_{i+1}$ . Ανά δύο διαδοχικά σημεία, δηλαδή από τα σημεία με δείκτη  $i = (0, 1), (1, 2), (2, 3), \dots$ , περνούμε πολυώνυμα τρίτου βαθμού (γι' αυτό χαρακτηρίζεται ως «κυβική» η spline) της μορφής

$$p_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i.$$

Το κάθε πολυώνυμο  $p_i(x)$  ορίζεται στο διάστημα  $[x_i, x_{i+1}]$ . Το πλήθος τους είναι  $n$ , όσα τα ζεύγη σημείων (ή τα διαστήματα). Επομένως  $i = 0, 1, 2, \dots, n-1$ .

Η απαίτηση να περνούν τα πολυώνυμα από τα σημεία  $(x_i, f_i)$ , δηλαδή, να ισχύει για  $i = 0, 1, 2, \dots, n-1$

$$\begin{aligned} p_i(x_i) &= f_i, \\ p_i(x_{i+1}) &= f_{i+1}, \end{aligned}$$

παράγει δύο γραμμικές εξισώσεις για τους συντελεστές  $a_i, b_i, c_i, d_i$  κάθε πολυωνύμου. Έτσι έχουμε συνολικά τις ακόλουθες  $2n$  εξισώσεις:

$$d_i = f_i, \quad (4.5\alpha')$$

$$a_i(x_{i+1} - x_i)^3 + b_i(x_{i+1} - x_i)^2 + c_i(x_{i+1} - x_i) + d_i = f_{i+1}. \quad (4.5\beta')$$

Στα σημεία που «ενώνονται» τα πολυώνυμα, δηλαδή στα  $n-1$  «εσωτερικά» σημεία  $x_1, x_2, \dots, x_{n-1}$ , απαιτούμε να έχουν ίσες πρώτες και δεύτερες παραγώγους. Επομένως

$$p'_{i-1}(x_i) = p'_i(x_i) \Rightarrow 3a_{i-1}(x_i - x_{i-1})^2 + 2b_{i-1}(x_i - x_{i-1}) + c_{i-1} = c_i, \quad (4.6\alpha')$$

$$p''_{i-1}(x_i) = p''_i(x_i) \Rightarrow 6a_{i-1}(x_i - x_{i-1}) + 2b_{i-1} = 2b_i, \quad (4.6\beta')$$

για  $i = 1, 2, \dots, n-1$ . Αυτές είναι άλλες  $2(n-1)$  γραμμικές εξισώσεις.

Επιπλέον, απαιτούμε οι δεύτερες παράγωγοι του  $p_0(x)$  στο άκρο  $x_0$  και του  $p_{n-1}(x)$  στο άλλο άκρο  $x_n$  να είναι ίσες με 0 (με τη συγκεκριμένη επιλογή παράγουμε τη «φυσική» spline):

$$p_0''(x_0) = 0 \Rightarrow 2b_0 = 0, \quad (4.7\alpha')$$

$$p_{n-1}''(x_n) = 0 \Rightarrow 6a_{n-1}(x_n - x_{n-1}) + 2b_{n-1} = 0. \quad (4.7\beta')$$

Οι γραμμικές εξισώσεις (4.5), (4.6), (4.7) είναι συνολικά  $4n$ , όσοι και οι άγνωστοι συντελεστές των  $n$  πολυωνύμων. Με την επίλυση του συστήματος γνωρίζουμε πλήρως τους συντελεστές των πολυωνυμικών τμημάτων της spline που περνά από τα δεδομένα σημεία.

Παρατηρήστε ότι οι εξισώσεις (4.5α') μας δίνουν τους σταθερούς όρους των πολυωνύμων. Με αυτό υπόψη, ας ανακεφαλαιώσουμε την κατασκευή της φυσικής κυβικής spline που περνά από τα σημεία  $(x_i, f_i)$  με  $i = 0, 1, \dots, n$ : Στα  $n$  διαστήματα  $[x_i, x_{i+1}]$  με  $i = 0, 1, \dots, n-1$ , ορίζουμε τα πολυώνυμα

$$p_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + f_i.$$

Σχηματίζουμε τις εξισώσεις

$$\begin{aligned} a_i(x_{i+1} - x_i)^2 + b_i(x_{i+1} - x_i) + c_i &= \frac{f_{i+1} - f_i}{x_{i+1} - x_i}, \text{ για } i = 0, \dots, n-1, \\ 3a_i(x_{i+1} - x_i)^2 + 2b_i(x_{i+1} - x_i) + c_i - c_{i+1} &= 0, \text{ για } i = 0, \dots, n-2, \\ 3a_i(x_{i+1} - x_i) + b_i - b_{i+1} &= 0, \text{ για } i = 0, \dots, n-2, \\ b_0 &= 0, \\ 3a_{n-1}(x_n - x_{n-1}) + b_{n-1} &= 0. \end{aligned}$$

Η λύση του γραμμικού συστήματος υπολογίζει τους συντελεστές  $a_i, b_i, c_i$  με  $i = 0, 1, \dots, n-1$ .

## 4.5 Προσέγγιση παραγώγων

Ένας σημαντικός κλάδος της Αριθμητικής Ανάλυσης ασχολείται με τον προσεγγιστικό υπολογισμό των παραγώγων μιας συνάρτησης για την οποία γνωρίζουμε τις τιμές  $\dots, f_{-1}, f_0, f_1, \dots$  σε σημεία (με αύξουσα τιμή)  $\dots, x_{-1}, x_0, x_1, \dots$ . Για να δούμε πώς, ας θυμηθούμε τον ορισμό της πρώτης παραγώγου μιας συνεχούς συνάρτησης  $f(x)$  σε ένα σημείο  $\bar{x}$  στο πεδίο ορισμού της:

$$f'(\bar{x}) = \lim_{x \rightarrow \bar{x}} \frac{f(x) - f(\bar{x})}{x - \bar{x}}.$$

Έστω ότι θέλουμε να υπολογίσουμε την παράγωγο της  $f(x)$  στο σημείο  $x_0$ . Τα πλησιέστερα σημεία στα οποία γνωρίζουμε τη συνάρτηση, εκατέρωθεν του  $x_0$  είναι



τα  $x_1, x_{-1}$ . Επομένως,

$$f'(x_0) \approx \frac{f(x_1) - f(x_0)}{x_1 - x_0}, \quad (4.8\alpha')$$

$$f'(x_0) \approx \frac{f(x_{-1}) - f(x_0)}{x_{-1} - x_0}. \quad (4.8\beta')$$

Ας υπολογίσουμε την ακρίβεια που έχουμε σε αυτούς τους τύπους. Το ανάπτυγμα Taylor στο σημείο  $x_0$  της  $f(x_1)$  είναι

$$f(x_1) = f(x_0) + (x_1 - x_0)f'(x_0) + \frac{(x_1 - x_0)^2}{2!}f''(x_0) + \frac{(x_1 - x_0)^3}{3!}f'''(x_0) + \dots$$

Λύνοντας ως προς  $f'(x_0)$  έχουμε

$$f'(x_0) = \frac{f(x_1) - f(x_0)}{x_1 - x_0} - \frac{x_1 - x_0}{2!}f''(x_0) - \frac{(x_1 - x_0)^2}{3!}f'''(x_0) - \dots \quad (4.9)$$

Αντίστοιχα, από το ανάπτυγμα του  $f(x_{-1})$  στο σημείο  $x_0$  έχουμε

$$f'(x_0) = \frac{f(x_{-1}) - f(x_0)}{x_{-1} - x_0} - \frac{x_{-1} - x_0}{2!}f''(x_0) - \frac{(x_{-1} - x_0)^2}{3!}f'''(x_0) - \dots \quad (4.10)$$

Παρατηρούμε ότι οι τύποι (4.8α'), (4.8β') υπολογίζουν την πρώτη παράγωγο με ακρίβεια ανάλογη των  $x_1 - x_0$  και  $x_{-1} - x_0$  αντίστοιχα.

Ας δούμε πώς απλοποιούνται οι τύποι (4.9), (4.10) όταν  $x_1 - x_0 = x_0 - x_{-1} = h$ . Έχουμε

$$\begin{aligned} f'(x_0) &= \frac{f(x_1) - f(x_0)}{h} - \frac{h}{2!}f''(x_0) - \frac{h^2}{3!}f'''(x_0) - \dots, \\ f'(x_0) &= \frac{f(x_0) - f(x_{-1})}{h} + \frac{h}{2!}f''(x_0) - \frac{h^2}{3!}f'''(x_0) - \dots. \end{aligned}$$

Το άθροισμά τους δίνει έναν άλλο τύπο για την πρώτη παράγωγο στο  $x_0$  (που ισχύει αν έχουμε ισαπέχοντα σημεία  $x_{-1}, x_0, x_1$ ):

$$f'(x_0) = \frac{f(x_1) - f(x_{-1})}{2h} - \frac{h^2}{3!}f'''(x_0) - \dots \quad (4.11)$$

Παρατηρήστε ότι η ακρίβεια είναι καλύτερη από τους τύπους (4.8α'), (4.8β') παρόλο που πάλι δύο μόνο σημεία λαμβάνουμε υπόψη. Ο υπολογισμός της παραγώγου με τιμές σε σημεία που περικλείουν το σημείο υπολογισμού της είναι πιο ακριβής από τον υπολογισμό με τύπο που έχει το σημείο υπολογισμού σε άκρο του.

Γενικότερα, η παράγωγος της  $f(x)$ , οποιασδήποτε τάξης  $m$  (ακόμα και μηδενικής), σε κάποιο σημείο στο πεδίο ορισμού της συνάρτησης  $\bar{x}$ , μπορεί να γραφεί ως γραμμικός συνδυασμός γνωστών τιμών της συνάρτησης σε σημεία  $x_i$ , με  $i = 1, \dots, n$  και  $n > m$ :

$$f^{(m)}(\bar{x}) \approx \sum_{i=1}^n w_i f(x_i). \quad (4.12)$$

Οι συντελεστές  $w_i$  εξαρτώνται από το  $\bar{x}$  και τα  $x_i$  και μπορούν να προκύψουν από την απαίτηση η παραπάνω σχέση να είναι *ακριβής* όταν η  $f(x)$  είναι διαδοχικά οι συναρτήσεις  $g_0(x) = 1$ ,  $g_1(x) = x$ ,  $g_2(x) = x^2$ , ...,  $g_{n-1}(x) = x^{n-1}$ , όπου  $n$  το πλήθος των σημείων. Παράγεται έτσι ένα γραμμικό σύστημα εξισώσεων με άγνωστους τους συντελεστές  $w_i$ , το οποίο έχει μοναδική λύση. Ο πίνακας των συντελεστών σε αυτό το γραμμικό σύστημα εύκολα δείχνεται ότι έχει στοιχεία  $a_{ij} = x_j^{i-1}$  με  $i, j = 1, 2, \dots, n$ .<sup>1</sup> Το διάνυσμα-στήλη των σταθερών όρων στο σύστημα έχει στοιχεία  $b_i = g_{i-1}^{(m)}(\bar{x})$ .

### Παράδειγμα

Ας εφαρμόσουμε αυτή τη μέθοδο για τον υπολογισμό της πρώτης παραγώγου από τις τιμές στα σημεία  $x_0 - h, x_0, x_0 + h$ :

$$f'(\bar{x}) \approx af(x_0 - h) + bf(x_0) + cf(x_0 + h).$$

Όταν η  $f(x)$  είναι διαδοχικά  $1, x, x^2$  έχουμε

$$\begin{aligned} f(x) = 1 &\Rightarrow 0 = a + b + c, \\ f(x) = x &\Rightarrow 1 = a(x_0 - h) + bx_0 + c(x_0 + h), \\ f(x) = x^2 &\Rightarrow 2\bar{x} = a(x_0 - h)^2 + bx_0^2 + c(x_0 + h)^2. \end{aligned}$$

Η λύση του γραμμικού συστήματος δίνει

$$\begin{aligned} a &= -\frac{1}{2h} + \frac{\bar{x} - x_0}{h^2}, \\ b &= -2\frac{\bar{x} - x_0}{h^2}, \\ c &= \frac{1}{2h} + \frac{\bar{x} - x_0}{h^2}. \end{aligned}$$

Αν  $\bar{x} \equiv x_0$  έχουμε  $a = -1/2h$ ,  $b = 0$ ,  $c = 1/2h$ · παράγουμε δηλαδή τον τύπο (4.11).

Αν  $\bar{x} \equiv x_0 + h$  έχουμε  $a = 1/2h$ ,  $b = -2/h$ ,  $c = 3/2h$ . Επομένως

$$f'(x_0 + h) \approx \frac{f(x_0 - h) - 4f(x_0) + 3f(x_0 + h)}{2h}.$$

Παρατηρήστε ότι οι συντελεστές  $w_i$  στην εξίσωση (4.12) μπορούν να προκύψουν (με περισσότερες πράξεις) από την παραγωγή του πολυωνύμου προσέγγισης στη μορφή Lagrange· είναι οι παράγωγοι τάξης  $m$  στο  $\bar{x}$ , των συναρτήσεων της βάσης Lagrange, (4.3β')

$$w_i = \frac{d^m \ell_i(x)}{dx^m} \Big|_{x=\bar{x}}.$$

<sup>1</sup>Ο πίνακας είναι ο ανάστροφος του πίνακα Vandermonde.

## 4.6 Προσέγγιση με τη μέθοδο ελάχιστων τετραγώνων

Πολλές φορές θέλουμε καλή προσέγγιση σε μια συνάρτηση  $f(x)$  όταν έχουμε ένα σύνολο σημείων  $(x_i, y_i)$  (με διαφορετικά  $x_i$ ) αλλά δεν ισχύει απαραίτητα  $y_i = f(x_i)$ . Τέτοια περίπτωση αποτελούν οι μετρήσεις πειραματικών δεδομένων, καθώς περιέχουν σφάλματα. Το ίδιο ισχύει και όταν θέλουμε να «απλοποιήσουμε» μια πολύπλοκη συνάρτηση με κάποια πιο απλή. Η απλή συνάρτηση θα περιέχει κάποιες παραμέτρους που θα πρέπει να αποκτήσουν κατάλληλες τιμές για να προσεγγίζει την πολύπλοκη. Αν οι παράμετροι είναι λιγότερες από τα σημεία στα οποία γνωρίζουμε τη συνάρτηση, δε θα μπορούμε να βρούμε λύση στο σύστημα εξισώσεων που προκύπτει από την απαίτηση να περνά από αυτά τα σημεία. Η απλούστερη συνάρτηση, εν γνώσει μας, δεν μπορεί να περνά από τα ίδια σημεία με την πολύπλοκη, αλλά επιδιώκουμε να πλησιάζει όσο περισσότερο γίνεται.

Με τη μέθοδο ελάχιστων τετραγώνων προσαρμόζουμε στα δεδομένα μας μια συνάρτηση  $f(x)$  προκαθορισμένης μορφής, με παραμέτρους, ώστε το άθροισμα των τετραγώνων των αποκλίσεων από τα  $y_i$ ,  $\sum_{i=1}^n (f(x_i) - y_i)^2$ , να γίνεται ελάχιστο ως προς αυτές τις παραμέτρους.

**Παρατήρηση:** Επιλέγουμε να ελαχιστοποιήσουμε το άθροισμα των τετραγώνων αντί για τις απόλυτες τιμές των αποκλίσεων καθώς θέλουμε να σχηματίσουμε μία συνεχή και παραγωγίσιμη συνάρτηση<sup>2</sup>.

### 4.6.1 Ευθεία ελάχιστων τετραγώνων

Έστω ότι η ζητούμενη συνάρτηση είναι γραμμική (πολυώνυμο βαθμού 1). Τέτοια περίπτωση έχουμε όταν τα σημεία μας αντιστοιχούν σε πειραματικές μετρήσεις και γνωρίζουμε από τη θεωρία ότι η σχέση των  $x, y$  είναι γραμμική ή όταν θέλουμε να προσεγγίσουμε μια πολύπλοκη συνάρτηση με γραμμική σχέση. Σχηματίζουμε τη συνάρτηση  $f(x) = \alpha x + \beta$  με άγνωστους συντελεστές  $\alpha, \beta$ . Αυτοί θα προκύψουν από την απαίτηση να ελαχιστοποιείται το άθροισμα

$$E(\alpha, \beta) = \sum_{i=1}^n (\alpha x_i + \beta - y_i)^2.$$

Η συνάρτηση  $E(\alpha, \beta)$  γίνεται ακρότατη όταν  $\partial E / \partial \alpha = 0$ ,  $\partial E / \partial \beta = 0$ . Οι εξισώσεις οδηγούν στο σύστημα

$$\begin{bmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{bmatrix}.$$

<sup>2</sup>Προσέξτε ότι π.χ. η παράγωγος του  $|x|$  δεν ορίζεται στο 0, ενώ, αντίθετα, η  $x^2$  έχει παραγώγους σε όλο το πεδίο ορισμού της.

Η εφαρμογή της μεθόδου Cramer (§3.2.1) δίνει αμέσως ότι

$$\alpha = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} = \frac{\overline{xy} - \bar{x} \bar{y}}{\overline{x^2} - \bar{x}^2}, \quad (4.13\alpha')$$

$$\beta = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} = \bar{y} - \alpha \bar{x}, \quad (4.13\beta')$$

όπου

$$\bar{w} = \frac{1}{n} \sum_{i=1}^n w_i,$$

η μέση τιμή ενός μεγέθους  $w$ .

Η ποσότητα

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

στον παρονομαστή είναι μη μηδενική, και για την ακρίβεια, θετική: καθώς τα σημεία  $x_i$  είναι διαφορετικά και επομένως, όλα εκτός ίσως από ένα, μη μηδενικά, ισχύει

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})^2 &> 0 \Rightarrow \sum_{i=1}^n (x_i^2 + \bar{x}^2) - 2\bar{x} \sum_{i=1}^n x_i > 0 \Rightarrow \\ \sum_{i=1}^n x_i^2 + n\bar{x}^2 - 2n\bar{x}\bar{x} &> 0 \Rightarrow \sum_{i=1}^n x_i^2 - n\bar{x}^2 > 0 \Rightarrow \\ n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2 &> 0. \end{aligned}$$

Το ακρότατο της  $E(\alpha, \beta)$  στις παραπάνω τιμές των  $\alpha, \beta$  είναι ελάχιστο καθώς ο εσσιανός πίνακας

$$\begin{bmatrix} \frac{\partial^2 E}{\partial \alpha^2} & \frac{\partial^2 E}{\partial \alpha \partial \beta} \\ \frac{\partial^2 E}{\partial \beta \partial \alpha} & \frac{\partial^2 E}{\partial \beta^2} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{bmatrix}$$

είναι συμμετρικός θετικά ορισμένος<sup>3</sup>.

Ο συντελεστής συσχέτισης,  $r^2$ , που προσδιορίζει την ποιότητα της προσέγγισης, είναι

$$r^2 \equiv \frac{\left( n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i \right)^2}{\left( n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2 \right) \left( n \sum_{i=1}^n y_i^2 - \left( \sum_{i=1}^n y_i \right)^2 \right)} = \alpha^2 \frac{\overline{x^2} - \bar{x}^2}{\overline{y^2} - \bar{y}^2}.$$

Ισχύει πάντα ότι  $0 \leq r^2 \leq 1$ . Το  $r^2 = 1$  υποδηλώνει τέλεια προσαρμογή (η ευθεία περνά από όλα τα σημεία), ενώ η τιμή γίνεται μικρότερη από 1 όσο πιο διασκορπισμένα είναι τα σημεία γύρω από την ευθεία.

#### 4.6.2 Πολυώνυμο ελάχιστων τετραγώνων

Έστω  $f(x) \equiv p(x) = \sum_{i=0}^m \alpha_i x^i$ . Τότε, η ελαχιστοποίηση της ποσότητας

$$E(\alpha_0, \alpha_1, \dots, \alpha_m) = \sum_{i=1}^n (p(x_i) - y_i)^2$$

ως προς τα  $\alpha_k$  δίνει τις εξισώσεις

$$\frac{\partial E}{\partial \alpha_k} = 2 \sum_{i=1}^n x_i^k \left( \sum_{j=0}^m \alpha_j x_i^j - y_i \right) = 0, \quad k = 0, 1, \dots, m.$$

Χρησιμοποιώντας πίνακες, οι εξισώσεις γίνονται

$$\begin{bmatrix} S_0 & S_1 & \cdots & S_m \\ S_1 & S_2 & \cdots & S_{m+1} \\ \vdots & \vdots & & \vdots \\ S_m & S_{m+1} & \cdots & S_{2m} \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}$$

<sup>3</sup>Έλεγχος με το κριτήριο του Sylvester, §3.1.2:

$$\sum_{i=1}^n x_i^2 > 0,$$

$$n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2 > 0.$$

όπου

$$S_k = \sum_{i=1}^n x_i^k, \quad k = 0, 1, \dots, 2m,$$

$$\beta_k = \sum_{i=1}^n x_i^k y_i, \quad k = 0, 1, \dots, m.$$

Το σύστημα αυτό έχει μοναδική λύση στο  $\mathcal{R}$  αν  $m < n$ ,  $x_i \neq x_j \forall i \neq j$ . Αν  $n = m + 1$ , η λύση αντιστοιχεί στο πολυώνυμο προσέγγισης  $m$  βαθμού.

### Παράδειγμα

Έστω ότι τα σημεία  $(x_i, y_i)$  είναι  $\{(0.0, 1.0), (0.25, 1.284), (0.5, 1.6487), (0.75, 2.117), (1.0, 2.7183)\}$ ,  $i = 1, \dots, 5$ . Το δευτεροβάθμιο πολυώνυμο που εξάγεται από τη μέθοδο ελάχιστων τετραγώνων προκύπτει ως λύση της

$$\begin{bmatrix} 5.0 & 2.5 & 1.8750 \\ 2.5 & 1.875 & 1.5625 \\ 1.875 & 1.5625 & 1.3828 \end{bmatrix} \cdot \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} 8.768 \\ 5.4514 \\ 4.4015 \end{bmatrix} \Rightarrow \begin{cases} \alpha_0 = 1.0052 \\ \alpha_1 = 0.8641 \\ \alpha_2 = 0.8437 \end{cases}.$$

Επομένως,

$$p(x) = 1.0052 + 0.8641x + 0.8437x^2.$$

Προσδιορίστε το αντίστοιχο πρωτοβάθμιο πολυώνυμο<sup>α</sup>. Τι παρατηρείτε;

$$^a p(x) = 0.89968 + 1.70784x$$

### 4.6.3 Καμπύλη ελάχιστων τετραγώνων $f(y) = \alpha g(x) + \beta$

Έστω ότι έχουμε την περίπτωση πειραματικών σημείων που η θεωρητική τους σχέση είναι (ή μπορεί να γίνει) της μορφής  $f(y) = \alpha g(x) + \beta$ , όπου  $f(y)$  και  $g(x)$  κάποιες συναρτήσεις. Τότε, η εύρεση της καμπύλης ελάχιστων τετραγώνων είναι απλή: Ορίζουμε τα σημεία  $(\tilde{x}_i, \tilde{y}_i)$  με  $\tilde{x}_i = g(x_i)$  και  $\tilde{y}_i = f(y_i)$ , και εφαρμόζουμε για αυτά τις σχέσεις (4.13) (καθώς η σχέση τους είναι γραμμική). Προσέξτε όμως ότι τα σφάλματα στις τιμές  $y_i$  και  $f(y_i)$  διαφέρουν. Η ελαχιστοποίηση των τετραγώνων τους θα δώσει (πιθανόν πολύ) διαφορετικές καμπύλες. Η εναλλακτική διαδικασία (στην οποία δεν θα αναφερθούμε) αντιμετωπίζει το πρόβλημα ως μη γραμμική ελαχιστοποίηση.

### Παραδείγματα

Έστω ότι η θεωρητική σχέση είναι  $y = a + be^x$ . Αν ορίσουμε

$$\begin{aligned}\tilde{y} &= y, \\ \tilde{x} &= e^x, \\ \tilde{\alpha} &= b, \\ \tilde{\beta} &= a\end{aligned}$$

η εξίσωση γίνεται  $\tilde{y} = \tilde{\beta} + \tilde{\alpha}\tilde{x}$ . Η εφαρμογή των τύπων (4.13) υπολογίζει τα  $\tilde{\alpha}$ ,  $\tilde{\beta}$  άρα και τα  $a$ ,  $b$ .

Έστω ότι η θεωρητική σχέση είναι  $y = ax^b$ . Παρατηρούμε ότι η εξίσωση αυτή μπορεί να γραφεί στη μορφή  $\ln y = \ln a + b \ln x$ . Ορίζουμε

$$\begin{aligned}\tilde{y} &= \ln y, \\ \tilde{x} &= \ln x, \\ \tilde{\alpha} &= b, \\ \tilde{\beta} &= \ln a\end{aligned}$$

οπότε η εξίσωση γίνεται  $\tilde{y} = \tilde{\beta} + \tilde{\alpha}\tilde{x}$ . Η εφαρμογή των τύπων (4.13) υπολογίζει τα  $\tilde{\alpha}$ ,  $\tilde{\beta}$  άρα και τα  $a$ ,  $b$ .

## 4.7 Ασκήσεις

Για τις παρακάτω ασκήσεις δημιουργήστε ένα αρχείο με όνομα «points.dat», που να περιέχει τα ζεύγη  $(x_i, f(x_i))$  μιας γνωστής συνάρτησης  $f(x)$ , π.χ.  $\sin x$ . Τα  $x_i$  ως είναι 15 ισαπέχοντα σημεία στο διάστημα  $[2, 4]$ .

1. Γράψτε ένα υποπρόγραμμα που να προσδιορίζει το πολυώνυμο προσέγγισης σε σημεία  $(x_i, y_i)$ . Θα δέχεται ως ορίσματα εισόδου δύο πίνακες  $x, y$  που θα περιέχουν τα ζεύγη σημείων  $(x, y)$  καθώς και την τιμή στην οποία θέλουμε να υπολογίζει το πολυώνυμο παρεμβολής· την τιμή αυτού θα την επιστρέφει. Για τον υπολογισμό του πολυωνύμου παρεμβολής να χρησιμοποιεί

(α') τον τύπο Lagrange.

(β') τον τύπο Newton.

(γ') την απαλοιφή Gauss.

Εφαρμόστε το υποπρόγραμμα που γράψατε για να υπολογίσετε τις προσεγγιστικές τιμές της «άγνωστης» συνάρτησης του αρχείου «points.dat» σε 100 ισαπέχοντα σημεία μεταξύ του ελάχιστου και του μέγιστου από τα  $x_i$ .

2. Τροποποιήστε τον κώδικα που γράψατε για την προηγούμενη άσκηση ώστε να υπολογίζει και να επιστρέφει, εκτός από τις προσεγγιστικές τιμές της

συνάρτησης (δηλαδή, τις τιμές του πολυωνύμου παρεμβολής,  $p(x)$ ), και τις αντίστοιχες τιμές της πρώτης παραγώγου,  $p'(x)$ .

3. Μία άγνωστη συνάρτηση μιας μεταβλητής,  $f(x)$ , μπορεί να προσεγγιστεί όχι μόνο από πολυώνυμο αλλά και από λόγο πολυωνύμων  $R(x)$ ,

$$R(x) = \frac{P(x)}{Q(x)}, \quad P(x) = \sum_{k=0}^M a_k x^k, \quad Q(x) = 1 + \sum_{k=1}^N b_k x^k,$$

με  $M + N + 1$  κατάλληλους συντελεστές  $a_k, b_k$ . Έστω ότι για την  $f(x)$  γνωρίζουμε ότι περνά από τα παρακάτω ζεύγη τιμών

$x$	$y$
0.9	5.607
1.1	4.576
1.5	3.726
2.0	3.354
2.9	3.140
3.5	3.087

Να προσδιορίσετε την  $R(x)$  με  $M = 2, N = 3$  (επομένως, με 6 άγνωστους συντελεστές  $a_k, b_k$ ) ώστε να περνά από τα παραπάνω ζεύγη τιμών, δηλαδή να ικανοποιεί τις σχέσεις  $y_i = R(x_i), \quad i = 1, \dots, 6$ .

4. Γράψτε υποπρόγραμμα που να υπολογίζει τη spline που περνά από  $n + 1$  ζεύγη σημείων  $(x_i, y_i)$ .

Εφαρμόστε το υποπρόγραμμα που γράψατε για να υπολογίσετε τις προσεγγιστικές τιμές της «άγνωστης» συνάρτησης του αρχείου «points.dat» σε 100 ισαπέχοντα σημεία μεταξύ των  $\min\{x_i\}$  και  $\max\{x_i\}$ .

5. Γράψτε πρόγραμμα που να προσεγγίζει άγνωστη συνάρτηση με τη μέθοδο ελάχιστων τετραγώνων. Η συνάρτηση θα δίνεται ως ζεύγη σημείων, σε δύο πίνακες  $x, y$ . Δώστε τη δυνατότητα στο χρήστη του προγράμματος να επιλέγει την προσεγγιστική καμπύλη μεταξύ των

(α')  $y = ax + b$  (γραμμική),

(β')  $y = ax^b$  (δύναμη),

(γ')  $y = a + be^x$  (εκθετική),

(δ')  $y = a + b \ln x$  (λογαριθμική).

Να υπολογίζετε κάθε φορά το συντελεστή  $r^2$  της καμπύλης ελάχιστων τετραγώνων που επιλέγεται.

Χρησιμοποιήστε το αρχείο «points.dat» για να έχετε τα σημεία  $(x_i, y_i)$ .



6. Η συνολική φωτεινή ισχύς,  $P$ , που εκπέμπεται από ένα μέλαν σώμα επιφάνειας  $A$ , δίνεται συναρτήσει της απόλυτης θερμοκρασίας του,  $T$ , από τη σχέση

$$P = \sigma AT^4,$$

όπου  $\sigma$  η σταθερά Stefan–Boltzmann. Πειραματικές μετρήσεις για ένα νήμα ηλεκτρικού λαμπτήρα (που θεωρούμε ότι προσεγγίζει το μέλαν σώμα) σε θερμοκρασίες 300 K–2300 K έδωσαν τις ακόλουθες τιμές

$T(K)$	$P(W)$	$T(K)$	$P(W)$
300	0.0013	1400	1.0031
400	0.0162	1500	1.4193
500	0.0297	1600	1.9052
600	0.0318	1700	2.4026
700	0.0484	1800	2.5031
800	0.0965	1900	3.9072
900	0.1357	2000	4.3156
1000	0.2947	2100	5.5060
1100	0.4563	2200	6.9044
1200	0.5398	2300	7.6370
1300	0.8884		

Αν υποθέσουμε ότι η επιφάνεια του νήματος είναι  $0.05 \text{ cm}^2$ , να επαληθεύσετε από τα δεδομένα το νόμο Stefan–Boltzmann (ότι πράγματι η δύναμη στην οποία υψώνεται το  $T$  είναι 4) και να εκτιμήσετε τη σταθερά  $\sigma$ . Υπολογίστε το συντελεστή  $r^2$  της καμπύλης ελάχιστων τετραγώνων.

7. Η περίοδος,  $T$ , ενός εκκρεμούς σε βαρυτικό πεδίο με επιτάχυνση  $g$ , σχετίζεται με το μήκος του,  $\ell$ , με τη σχέση

$$T = 2\pi\sqrt{\frac{\ell}{g}}.$$

Υπολογίστε την επιτάχυνση της βαρύτητας από τις ακόλουθες πειραματικές μετρήσεις

$\ell(\text{cm})$	$T(\text{s})$
18	0.84958
20	0.89696
22	0.94140
24	0.98530



## Κεφάλαιο 5

# Αριθμητική Ολοκλήρωση

### 5.1 Εισαγωγή

Ένα από τα βασικά προβλήματα στα μαθηματικά είναι ο υπολογισμός του ολοκληρώματος μίας συνάρτησης πραγματικής μεταβλητής,

$$\int_a^b f(x) dx .$$

Για τη συντριπτική πλειοψηφία των συναρτήσεων  $f(x)$  δεν υπάρχουν ή είναι πολύ δύσκολοι οι τύποι της αντιπαράγωγου της  $f(x)$ , δηλαδή της  $F(x)$  που ικανοποιεί τη σχέση  $F'(x) = f(x)$ , ώστε να υπολογιστεί ακριβώς το ολοκλήρωμα από τον τύπο

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

Επίσης, συχνά η  $f(x)$  δεν είναι γνωστή παρά μόνο σε συγκεκριμένα σημεία. Και σε αυτήν την περίπτωση δεν μπορούμε να χρησιμοποιήσουμε «κλειστό» τύπο για τον υπολογισμό του ολοκληρώματός της.

Για τον υπολογισμό ολοκληρωμάτων με συγκεκριμένα όρια έχουν αναπτυχθεί διάφορες αριθμητικές μέθοδοι. Όλες εκφράζουν το ζητούμενο ολοκλήρωμα ως άθροισμα των τιμών της συνάρτησης σε  $n$  συγκεκριμένα σημεία  $x_i$  στο διάστημα ολοκλήρωσης, πολλαπλασιασμένων με κατάλληλες σταθερές  $w_i$ :

$$\int_a^b f(x) dx \approx \sum_{i=1}^n w_i f(x_i) . \quad (5.1)$$

Παρακάτω θα δούμε διάφορες μεθόδους για την επιλογή των σημείων  $x_i$  και τον υπολογισμό των σταθερών  $w_i$ . Αν τα σημεία είναι ισαπέχοντα, οι μέθοδοι παράγουν τύπους στη γενική κατηγορία των τύπων Newton–Cotes. Σε αυτή την κατηγορία ανήκουν οι τύποι τραπεζίου και Simpson. Η ελεύθερη επιλογή των σημείων  $x_i$  οδηγεί σε γενικά πιο ακριβείς τύπους (τύποι Gauss και Clenshaw–Curtis).

### 5.1.1 Ολοκληρώματα με μη πεπερασμένα όρια ολοκλήρωσης

Συχνά υπάρχει η ανάγκη να υπολογιστούν ανώμαλα ολοκληρώματα (που συμπεριλαμβάνουν στο πεδίο τιμών τους το  $+\infty$  ή το  $-\infty$ ). Η πρώτη κατηγορία μεθόδων που θα δούμε παρακάτω, εφαρμόζεται για ολοκληρώματα με πεπερασμένα όρια. Αν η ολοκληρωτέα συνάρτηση έχει κατάλληλη μορφή, υπάρχουν αλγόριθμοι στη δεύτερη κατηγορία μεθόδων που μπορούν να υπολογίσουν το ολοκλήρωμά της με όρια τα  $\pm\infty$ . Αλλιώς, μπορούμε να κάνουμε μια κατάλληλη αλλαγή μεταβλητής, π.χ.  $x = 1/t$ , ώστε να προκύψουν ολοκληρώματα με πεπερασμένα όρια. Έτσι:

$$\int_a^b f(x) dx = \int_{1/b}^{1/a} \frac{1}{t^2} f\left(\frac{1}{t}\right) dt$$

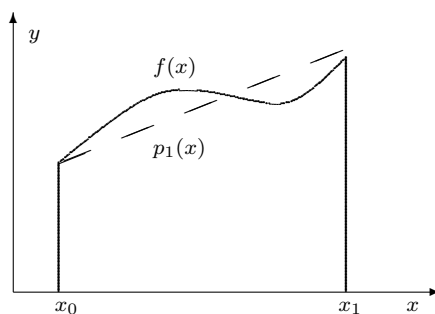
για  $a, b$  ομόσημα. Συνεπώς η συγκεκριμένη αλλαγή μεταβλητής μπορεί να χρησιμοποιηθεί όταν  $a = -\infty$  και  $b < 0$  είτε  $b = +\infty$  και  $a > 0$ .

Εάν τα  $a, b$  είναι ετερόσημα, τότε μπορούμε ορίσουμε ένα σημείο  $c$  μεταξύ των  $a, b$  και ομόσημο με το μη πεπερασμένο όριο. Κατόπιν μπορούμε να χρησιμοποιήσουμε την παραπάνω αλλαγή μεταβλητής. Π.χ. για τον υπολογισμό της  $f(x)$  στο  $[-3, +\infty)$ , μπορούμε να επιλέξουμε ένα θετικό  $c$  και να κάνουμε τα ακόλουθα:

$$\int_{-3}^{+\infty} f(x) dx = \int_{-3}^c f(x) dx + \int_c^{+\infty} f(x) dx = \int_{-3}^c f(x) dx + \int_0^{1/c} \frac{1}{t^2} f\left(\frac{1}{t}\right) dt.$$

Προϋπόθεση για να έχουμε καλή προσέγγιση κάνοντας τέτοια αλλαγή μεταβλητής είναι η συνάρτηση να τείνει στο 0 όταν  $x \rightarrow \infty$ .

## 5.2 Κανόνας Τραπεζίου



Σχήμα 5.1: Γραμμική προσέγγιση συνάρτησης για την εφαρμογή του τύπου ολοκλήρωσης τραπεζίου

Μια προσεγγιστική τιμή του ολοκληρώματος

$$\int_{x_0}^{x_1} f(x) dx$$

μπορεί να υπολογιστεί ολοκληρώνοντας το πολυώνυμο προσέγγισης που διέρχεται από τα σημεία  $(x_0, f(x_0))$ ,  $(x_1, f(x_1))$ , Σχήμα 5.1. Το πολυώνυμο αυτό είναι φυσικά πρωτοβάθμιο, δηλαδή ευθεία, και δίνεται από τον τύπο

$$p_1(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0) .$$

Η ολοκλήρωσή του με όρια  $x_0, x_1$  δίνει

$$\int_{x_0}^{x_1} f(x) dx \approx \int_{x_0}^{x_1} p_1(x) dx = \frac{x_1 - x_0}{2} [f(x_0) + f(x_1)] . \quad (5.2)$$

Η (5.2) αποτελεί τον (απλό) τύπο του τραπεζίου.

### 5.2.1 Σφάλμα ολοκλήρωσης κανόνα τραπεζίου

Το σφάλμα ολοκλήρωσης με τη μέθοδο τραπεζίου μπορεί να υπολογιστεί ως εξής:

Αναπτύσσουμε τη συνάρτηση  $f(x)$  κατά Taylor γύρω από το σημείο  $x_0$ :

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + f''(x_0)\frac{(x - x_0)^2}{2} + \dots . \quad (5.3)$$

Στο ανάπτυγμα έχουμε παραλείψει όρους υψηλότερης τάξης από 2.

Το ολοκλήρωμα της  $f(x)$  χρησιμοποιώντας το ανάπτυγμα είναι

$$\begin{aligned} \int_{x_0}^{x_1} f(x) dx &= \int_{x_0}^{x_1} f(x_0) dx + \int_{x_0}^{x_1} f'(x_0)(x - x_0) dx \\ &\quad + \int_{x_0}^{x_1} f''(x_0)\frac{(x - x_0)^2}{2} dx + \dots \\ &= f(x_0)(x_1 - x_0) + f'(x_0)\frac{(x_1 - x_0)^2}{2} \\ &\quad + f''(x_0)\frac{(x_1 - x_0)^3}{6} + \dots . \end{aligned} \quad (5.4)$$

Ο τύπος του τραπεζίου, (5.2), δίνει για το συγκεκριμένο ολοκλήρωμα

$$\begin{aligned} \int_{x_0}^{x_1} f(x) dx &= \frac{x_1 - x_0}{2} [f(x_0) + f(x_1)] \\ &= \frac{x_1 - x_0}{2} \left[ f(x_0) \right. \\ &\quad \left. + f(x_0) + f'(x_0)(x_1 - x_0) + f''(x_0)\frac{(x_1 - x_0)^2}{2} + \dots \right] . \end{aligned} \quad (5.5)$$

Στον προηγούμενο τύπο η  $f(x_1)$  υπολογίστηκε από το ανάπτυγμα Taylor, (5.3).

Η διαφορά των δύο σχέσεων, (5.4)-(5.5), είναι

$$\varepsilon = -\frac{1}{12}(x_1 - x_0)^3 f''(x_0) + \dots$$

Με ακριβή μαθηματική αντιμετώπιση καταλήγουμε ότι το σφάλμα  $\varepsilon$  της μεθόδου τραπεζίου είναι

$$\varepsilon = -\frac{1}{12}(x_1 - x_0)^3 f''(\xi), \quad \text{για κάποιο } \xi \in (x_0, x_1),$$

χωρίς επιπλέον όρους.

### 5.2.2 Εκτεταμένος τύπος τραπεζίου

Η επανάληψη του τύπου (5.2) για πολλά διαδοχικά διαστήματα, δίνει την προσεγγιστική έκφραση για το ολοκλήρωμα σε εκτεταμένο διάστημα. Έτσι, αν έχουμε χωρίσει το  $[a \equiv x_0, b \equiv x_n]$  σε  $n$  ίσα διαστήματα  $[x_i, x_{i+1}]$  με  $x_i = x_0 + ih$ ,  $i = 0, 1, \dots, n$  και  $h = \frac{b-a}{n}$ , έχουμε

$$\begin{aligned} \int_{x_0}^{x_n} f(x) dx &= \int_{x_0}^{x_1} f(x) dx + \int_{x_1}^{x_2} f(x) dx + \dots + \int_{x_{n-1}}^{x_n} f(x) dx \\ &\approx h \left( \frac{f_0}{2} + f_1 + f_2 + \dots + f_{n-1} + \frac{f_n}{2} \right), \end{aligned} \quad (5.6)$$

όπου  $f_i \equiv f(x_i)$ .

### 5.2.3 Σφάλμα ολοκλήρωσης εκτεταμένου τύπου τραπεζίου

Το σφάλμα ολοκλήρωσης,  $E$ , του εκτεταμένου τύπου τραπεζίου για μία συνάρτηση  $f(x)$ , η οποία είναι συνεχής με δύο συνεχείς παραγώγους στο  $[a, b]$ , μπορεί να εκτιμηθεί ως εξής:

$$\begin{aligned} E &= \int_a^b f(x) dx - \frac{h}{2} \left( f_0 + 2 \sum_{i=1}^{n-1} f_i + f_n \right) \\ &= \sum_{i=0}^{n-1} \left[ \int_{x_i}^{x_{i+1}} f(x) dx - \frac{h}{2} (f_i + f_{i+1}) \right] = \sum_{i=0}^{n-1} \varepsilon_i. \end{aligned}$$

Σε κάθε διάστημα  $[x_i, x_{i+1}]$  έχουμε:

$$\varepsilon_i = -\frac{1}{12}(x_{i+1} - x_i)^3 f''(\xi_i) = -\frac{h^3}{12} f''(\xi_i), \quad \xi_i \in (x_i, x_{i+1}).$$

Έστω ότι υπάρχει αριθμός  $M$  ώστε  $|f''(\xi_i)| \leq M$  για κάθε  $i$ . Επομένως,

$$|\varepsilon_i| \leq \frac{h^3}{12} M, \quad \forall i,$$

και

$$|E| = \sum_{i=0}^{n-1} \varepsilon_i \leq \sum_{i=0}^{n-1} |\varepsilon_i| \leq \frac{nM}{12} h^3 = \frac{(b-a)M}{12} h^2. \quad (5.7)$$

### Παράδειγμα

Ας υπολογίσουμε αριθμητικά το  $I = \int_0^\pi \sin x \, dx$  και να το συγκρίνουμε με την ακριβή του τιμή, 2. Έστω  $n+1$  ισαπέχοντα σημεία στο  $[0, \pi]$ ,  $x_i = i\pi/n$ ,  $i = 0, 1, \dots, n$ . Τότε

$$I \approx I_n = \frac{\pi}{n} \left( \frac{\sin x_0 + \sin x_n}{2} + \sum_{i=1}^{n-1} \sin x_i \right).$$

Επομένως,

$n$	$I_n$	$E = I - I_n$	$n$	$I_n$	$E = I - I_n$
1	0.00000000	2.00000000	11	1.98638699	0.01361301
2	1.57079633	0.42920367	12	1.98856378	0.01143622
3	1.81379936	0.18620064	13	1.99025718	0.00974282
4	1.89611890	0.10388110	14	1.99160043	0.00839957
5	1.93376560	0.06623440	15	1.99268383	0.00731617
6	1.95409723	0.04590277	16	1.99357034	0.00642966
7	1.96631668	0.03368332	17	1.99430494	0.00569506
8	1.97423160	0.02576840	18	1.99492046	0.00507954
9	1.97965081	0.02034919	19	1.99544132	0.00455868
10	1.98352354	0.01647646	20	1.99588597	0.00411403

Παρατηρήστε ότι το σφάλμα  $E$  τείνει στο 0 ανάλογα του  $h^2 \equiv \pi^2/n^2$ .

Το ελάχιστο  $n$  για να έχουμε  $E \leq 10^{-4}$  προσδιορίζεται ως εξής:

$$|f''(x)| = |-\sin x| \leq 1, \quad \forall x \in [0, \pi].$$

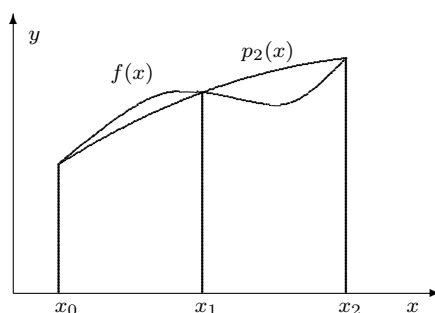
Επομένως, στον τύπο (5.7) έχουμε  $M = 1$  και

$$|E| \leq \frac{\pi}{12} h^2 = \frac{\pi}{12} \frac{\pi^2}{n^2} \leq 10^{-4} \Rightarrow n \geq 161.$$

## 5.3 Κανόνας Simpson

Στη μέθοδο Simpson προσεγγίζουμε την ολοκληρωτέα συνάρτηση χρησιμοποιώντας πολυώνυμο δεύτερου βαθμού που παρεμβάλλεται σε τρία ισαπέχοντα σημεία, Σχήμα 5.2:

$$\int_{x_0}^{x_2} f(x) \, dx \approx \int_{x_0}^{x_2} p_2(x) \, dx.$$



Σχήμα 5.2: Προσέγγιση συνάρτησης με παραβολή για την εφαρμογή του τύπου ολοκλήρωσης Simpson

Για τα σημεία  $x_0, x_1, x_2$  ισχύουν οι σχέσεις  $x_1 = x_0 + h, x_2 = x_1 + h$ . Το πολυώνυμο  $p_2(x)$  που περνά από τα  $(x_0, f_0), (x_1, f_1), (x_2, f_2)$ , προκύπτει από τον τύπο του Lagrange, (4.3), ότι είναι

$$p_2(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} f_0 + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} f_1 + \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} f_2 .$$

Επομένως,

$$\begin{aligned} \int_{x_0}^{x_2} p_2(x) dx &= f_0 \int_{x_0}^{x_2} \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} dx + f_1 \int_{x_0}^{x_2} \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} dx \\ &\quad + f_2 \int_{x_0}^{x_2} \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} dx \\ &= f_0 \frac{h}{3} + f_1 \frac{4h}{3} + f_2 \frac{h}{3} . \end{aligned}$$

Άρα, ο (απλός) τύπος του Simpson είναι

$$\int_{x_0}^{x_2} f(x) dx \approx \frac{h}{3} (f_0 + 4f_1 + f_2) . \quad (5.8)$$

### 5.3.1 Σφάλμα ολοκλήρωσης κανόνα Simpson

Αν η  $f(x)$  έχει συνεχείς τέταρτες παραγώγους στο  $[x_0, x_2]$ , προκύπτει ότι το σφάλμα  $\varepsilon$  είναι:

$$\begin{aligned} \varepsilon &\equiv \int_{x_0}^{x_2} f(x) dx - \frac{h}{3} (f_0 + 4f_1 + f_2) \\ &= -\frac{1}{90} h^5 f^{(4)}(\xi) = -\frac{1}{2880} (x_2 - x_0)^5 f^{(4)}(\xi) , \quad \text{για κάποιο } \xi \in (x_0, x_2) . \end{aligned}$$



### 5.3.2 Εκτεταμένος τύπος Simpson

Παρόμοια με τον τύπο τραπεζίου, μπορούμε να κατασκευάσουμε το σύνθετο τύπο Simpson στο διάστημα  $[a, b]$ , υποθέτοντας ότι  $b - a = 2kh$ . Προκύπτει ότι

$$\int_a^b f(x) dx \approx \frac{h}{3} \left( f_0 + f_{2k} + 4 \sum_{j=1}^k f_{2j-1} + 2 \sum_{j=1}^{k-1} f_{2j} \right). \quad (5.9)$$

### 5.3.3 Σφάλμα ολοκλήρωσης εκτεταμένου τύπου Simpson

Αν η τέταρτη παράγωγος είναι φραγμένη στο  $[a, b]$ ,

$$\max_{x \in [a, b]} |f^{(4)}(x)| \leq M,$$

το σφάλμα  $E$  του σύνθετου τύπου Simpson είναι

$$|E| \leq \frac{b-a}{180} M h^4. \quad (5.10)$$

#### Παράδειγμα

Ο υπολογισμός του

$$I = \int_0^\pi \sin x dx$$

με τον σύνθετο τύπο Simpson δίνει

$n$	$I_n$	$E = I - I_n$
2	2.0943951	-0.0943951
4	2.0045598	-0.0045598
6	2.0008632	-0.0008632
8	2.0002692	-0.0002692
10	2.0001095	-0.0001095
12	2.0000526	-0.0000526
14	2.0000283	-0.0000283
16	2.0000166	-0.0000166
18	2.0000103	-0.0000103
20	2.0000068	-0.0000068

Παρατηρήστε ότι για να επιτύχουμε σφάλμα κάτω από 0.005 χρειαζόμαστε  $4+1$  σημεία· αντίθετα, για ίδιο σφάλμα με τον τύπο τραπεζίου απαιτούνται  $19+1$ . Γενικότερα, καθώς  $|f^{(4)}(x)| = |\sin x| \leq 1 \forall x$ , το σφάλμα συνδέεται με τον αριθμό διαστημάτων  $n$  με τη σχέση

$$|E| \leq \frac{\pi}{90} \left( \frac{\pi}{n} \right)^4.$$

Επομένως, σφάλμα  $< 10^{-6}$  απαιτεί  $n \geq 43$ .

## 5.4 Κανόνας Simpson των $3/8$

Ο κανόνας των  $3/8$  προκύπτει από την ολοκλήρωση ενός πολωνύμου  $3^{\text{ης}}$  τάξης, το οποίο προσεγγίζει την ολοκληρωτέα συνάρτηση.

Για 4 δεδομένα ισαπέχοντα σημεία  $x_0, x_1, x_2, x_3$  μπορεί να αποδειχθεί ότι το ολοκλήρωμα δίνεται προσεγγιστικά από τον τύπο:

$$\int_{x_0}^{x_3} f(x) dx \approx \frac{3h}{8} [f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)] ,$$

όπου  $h = (x_3 - x_0)/3$ .

### 5.4.1 Σφάλμα ολοκλήρωσης κανόνα Simpson $3/8$

Μπορεί ναδειχθεί ότι ο τύπος του Simpson  $3/8$  έχει σφάλμα:

$$\varepsilon = -\frac{3}{80} h^5 f^{(4)}(\xi) = -\frac{1}{6480} (x_3 - x_0)^5 f^{(4)}(\xi) , \quad \text{για κάποιο } \xi \in (x_0, x_3) .$$

Ο τύπος των  $3/8$  είναι παρόμοιας ακρίβειας με τον τύπο του Simpson με το  $1/3$ , παρά το γεγονός ότι χρησιμοποιεί ένα παραπάνω σημείο. Η μεγάλη χρησιμότητα του τύπου των  $3/8$  είναι ότι ο αριθμός των διαστημάτων που χρειάζεται για τον υπολογισμό είναι περιττός (3) και συνεπώς μπορεί να χρησιμοποιηθεί όταν θέλουμε υψηλή ακρίβεια σε περιττό πλήθος διαστημάτων συνδυάζοντάς τον με τον τύπο Simpson  $1/3$ : στα πρώτα τρία διαστήματα μπορούμε να εφαρμόσουμε τον τύπο  $3/8$  και στα υπόλοιπα (που είναι άρτια στο πλήθος) τον τύπο  $1/3$ .

### 5.4.2 Εκτεταμένος τύπος Simpson των $3/8$

Αν και ο τύπος έχει περιορισμένη εφαρμογή, για διαστήματα με πλήθος  $n$  πολυπλασίασιο του 3, μπορεί να αποδειχθεί ότι:

$$I \approx \frac{3h}{8} \left( f_0 + 3 \sum_{i=0}^{k-1} f_{3i+1} + 3 \sum_{i=0}^{k-1} f_{3i+2} + 2 \sum_{i=0}^{k-2} f_{3i+3} + f_n \right) ,$$

όπου  $f_i = f(x_i)$ ,  $x_i \equiv a + ih$ ,  $h \equiv (b - a)/n$  και  $k = n/3$ .

## 5.5 Εναλλακτικός υπολογισμός των τύπων Newton–Cotes

Οι (απλοί) τύποι τραπεζίου, Simpson, κλπ., έχουν τη γενική ονομασία *τύποι Newton–Cotes* και μπορούν να υπολογιστούν και με άλλο τρόπο από αυτόν που είδαμε. Βασιζόμαστε στη (5.1) αλλά απαιτούμε να είναι *ακριβής* όταν  $n$   $f(x)$  είναι

διαδοχικά  $1, x, x^2, \dots, x^{n-1}$ , όπου  $n$  το πλήθος των σημείων. Προκύπτει έτσι ένα γραμμικό σύστημα εξισώσεων με άγνωστους τους συντελεστές  $w_i$ , το οποίο έχει μοναδική λύση.

Ας χρησιμοποιήσουμε αυτόν τον τρόπο για να υπολογίσουμε τον κανόνα Simpson. Ζητούμε να ισχύει

$$\int_{x_0}^{x_2} f(x) dx = \sum_{i=0}^2 w_i f(x_i) ,$$

όπου  $x_i$  (με  $i = 0, 1, 2$ ) τρία ισαπέχοντα σημεία:  $x_1 = x_0 + h$ ,  $x_2 = x_1 + h$ . Έχουμε διαδοχικά

$$\begin{aligned} f(x) = 1 &\Rightarrow x_2 - x_0 = w_0 + w_1 + w_2 , \\ f(x) = x &\Rightarrow \frac{x_2^2 - x_0^2}{2} = w_0 x_0 + w_1 x_1 + w_2 x_2 , \\ f(x) = x^2 &\Rightarrow \frac{x_2^3 - x_0^3}{3} = w_0 x_0^2 + w_1 x_1^2 + w_2 x_2^2 . \end{aligned}$$

Η λύση του γραμμικού συστήματος δίνει  $w_0 = h/3$ ,  $w_1 = 4h/3$ ,  $w_2 = h/3$ . Προκύπτει, επομένως, ο τύπος (5.8).

Όπως είδαμε κατά την εξαγωγή των απλών τύπων τραπεζίου και Simpson, οι συντελεστές  $w_i$  στην εξίσωση (5.1) μπορούν να προκύψουν (με περισσότερες πράξεις) από την ολοκλήρωση του πολυωνύμου προσέγγισης στη μορφή Lagrange· είναι τα ολοκληρώματα στο  $[a, b]$  των συναρτήσεων της βάσης Lagrange, (4.3β):

$$w_i = \int_{x_0}^{x_n} \ell_i(x) dx .$$

### 5.5.1 Παρατηρήσεις

Εξαιτίας του φαινομένου Runge, §4.1.2, η προσέγγιση με τύπο Newton–Cotes υψηλής τάξης δεν παράγει τύπο ολοκλήρωσης με καλή ακρίβεια. Στην πράξη δεν χρησιμοποιούνται τύποι πιο πολύπλοκοι από τον Simpson.

Το σφάλμα κάθε μεθόδου Newton–Cotes έχει τη μορφή  $\varepsilon \propto h^{2k}$  με  $k = 1$  (για τραπέζιο),  $k = 2$  (για Simpson), κλπ. Μπορούμε να θεωρήσουμε τη σταθερά αναλογίας ότι είναι της τάξης του 1 οπότε η απόσταση διαδοχικών σημείων,  $h$ , προκύπτει από την επιθυμητή ακρίβεια  $\epsilon$  ως  $h = \sqrt[2k]{\epsilon}$ . Μια πρώτη εκτίμηση για το αναγκαίο πλήθος των διαστημάτων,  $n$ , είναι το ακέραιο μέρος του  $(b - a)/h$ . Κατόπιν, εφαρμόζουμε την επιλεγμένη μέθοδο με διάφορες τιμές μεγαλύτερες από το συγκεκριμένο  $n$  έως ότου η τιμή που προκύπτει σε διαδοχικές εφαρμογές να μην αλλάζει σημαντικά.

## 5.6 Μέθοδοι Gauss

### 5.6.1 Μέθοδος Gauss–Legendre

Ένα ολοκλήρωμα με πεπερασμένα όρια,  $\int_a^b f(x) dx$ , μπορεί πάντα να μετασχηματιστεί σε ολοκλήρωμα στο διάστημα  $[-1, 1]$  αν επιλέξουμε κατάλληλη αλλαγή μεταβλητής. Έτσι, θέτουμε  $x = \lambda t + \mu$  και ζητούμε να ισχύει  $x = a$  όταν  $t = -1$  και  $x = b$  όταν  $t = 1$ . Τότε

$$\begin{aligned} x &= \frac{b-a}{2}t + \frac{b+a}{2}, \\ dx &= \frac{b-a}{2} dt. \end{aligned}$$

Συνεπώς, μπορούμε πάντα να μετασχηματίσουμε ένα ολοκλήρωμα στο διάστημα  $[a, b]$  σε άλλο στο διάστημα  $[-1, 1]$  με τον τύπο

$$\int_a^b f(x) dx = \frac{b-a}{2} \int_{-1}^1 f\left(\frac{b-a}{2}t + \frac{b+a}{2}\right) dt.$$

Ας ξαναδοούμε τον βασικό τύπο ολοκλήρωσης (5.1), γραμμένο όμως τώρα για το διάστημα  $[-1, 1]$ :

$$\int_{-1}^1 f(x) dx \approx \sum_{i=1}^m w_i f(x_i), \quad (5.11)$$

όπου  $x_i$  σταθερά σημεία στο  $[-1, 1]$  και  $w_i$  συντελεστές. Ο τύπος έχει γενική μορφή και περιλαμβάνει

- τον απλό κανόνα τραπεζίου: έχουμε  $w_1 = w_2 = 1$ ,  $x_1 = -1$ ,  $x_2 = 1$ .
- τον απλό κανόνα Simpson: έχουμε  $w_1 = w_3 = 1/3$ ,  $w_2 = 4/3$ ,  $x_1 = -1$ ,  $x_2 = 0$ ,  $x_3 = 1$ .

Το ερώτημα είναι: για δεδομένο αριθμό σημείων  $m$ , ποια είναι τα  $w_i$ ,  $x_i$ ,  $i = 1, 2, \dots, m$  ώστε ο κανόνας (5.11) να έχει τη μέγιστη δυνατή ακρίβεια; Προσέξτε ότι, σε αντίθεση με τους τύπους που παρουσιάσαμε μέχρι τώρα, έχουμε τη δυνατότητα επιλογής των  $x_i$ .

Έστω ότι η μέγιστη δυνατή ακρίβεια σημαίνει πως ο κανόνας δίνει το ακριβές αποτέλεσμα στην ολοκλήρωση των συνολικά  $2m$  μονωνύμων (όσα και οι άγνωστοι)  $1, x, x^2, \dots, x^{2m-1}$  (και επομένως, οποιουδήποτε γραμμικού συνδυασμού τους). Αυτή η συνθήκη οδηγεί στους κανόνες ολοκλήρωσης Gauss.

**Κανόνας Gauss με  $m = 1$**

Έχουμε

$$\int_{-1}^1 f(x) dx \approx w_1 f(x_1).$$

Καθώς πρέπει να προκύπτει το ακριβές αποτέλεσμα για  $f(x) = 1$  και  $f(x) = x$  έχουμε

$$\begin{aligned} f(x) = 1 &\Rightarrow w_1 1 = \int_{-1}^1 1 \, dx = 2 \Rightarrow w_1 = 2 , \\ f(x) = x &\Rightarrow w_1 x_1 = \int_{-1}^1 x \, dx = 0 \Rightarrow x_1 = 0 . \end{aligned}$$

Επομένως, ο κανόνας Gauss με ένα σημείο είναι

$$\int_{-1}^1 f(x) \, dx \approx 2f(0) .$$

Ο κανόνας αυτός ολοκληρώνει ακριβώς τα  $1, x$  αλλά όχι το  $x^2$ .

**Κανόνας Gauss με  $m = 2$**

Ζητώντας να παράγεται το ακριβές αποτέλεσμα για  $f(x) = 1$ ,  $f(x) = x$ ,  $f(x) = x^2$ ,  $f(x) = x^3$ , έχουμε

$$\begin{aligned} w_1 + w_2 &= 2 , \\ w_1 x_1 + w_2 x_2 &= 0 , \\ w_1 x_1^2 + w_2 x_2^2 &= \frac{2}{3} , \\ w_1 x_1^3 + w_2 x_2^3 &= 0 . \end{aligned}$$

Το παραπάνω σύστημα λύνεται αναλυτικά επιλύοντας διαδοχικά τις εξισώσεις ως προς κάποιον από τους αγνώστους και αντικαθιστώντας στις επόμενες. Η λύση του συστήματος είναι μοναδική (πέρα από την αλλαγή  $w_1 \leftrightarrow w_2$ ,  $x_1 \leftrightarrow x_2$ ) και είναι η εξής

$$\begin{aligned} w_1 &= 1 , & x_1 &= -\frac{1}{\sqrt{3}} , \\ w_2 &= 1 , & x_2 &= \frac{1}{\sqrt{3}} . \end{aligned}$$

Επομένως,

$$\int_{-1}^1 f(x) \, dx \approx f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right) .$$

**Κανόνας Gauss με  $m = 3$** 

Η απαίτηση για ακριβές αποτέλεσμα όταν  $f(x) = x^k$ ,  $k = 0, \dots, 5$  σχηματίζει το μη γραμμικό σύστημα

$$\begin{aligned} w_1 + w_2 + w_3 &= 2, \\ w_1 x_1 + w_2 x_2 + w_3 x_3 &= 0, \\ w_1 x_1^2 + w_2 x_2^2 + w_3 x_3^2 &= \frac{2}{3}, \\ w_1 x_1^3 + w_2 x_2^3 + w_3 x_3^3 &= 0, \\ w_1 x_1^4 + w_2 x_2^4 + w_3 x_3^4 &= \frac{2}{5}, \\ w_1 x_1^5 + w_2 x_2^5 + w_3 x_3^5 &= 0. \end{aligned}$$

Η λύση του είναι

$$\begin{aligned} w_1 &= \frac{5}{9}, & x_1 &= -\sqrt{0.6}, \\ w_2 &= \frac{8}{9}, & x_2 &= 0, \\ w_3 &= \frac{5}{9}, & x_3 &= \sqrt{0.6}. \end{aligned}$$

**Κανόνας Gauss με οποιοδήποτε  $m$** 

Η μέθοδος ολοκλήρωσης που παρουσιάστηκε σε αυτή την παράγραφο λέγεται μέθοδος *Gauss-Legendre*. Ονομάζεται έτσι γιατί, στη γενική περίπτωση, τα σημεία  $x_i$ ,  $i = 1, \dots, m$  είναι οι ρίζες του πολυωνύμου Legendre  $m$  τάξης,  $P_m(x)$ , και μπορούν να υπολογιστούν εύκολα, χωρίς τη λύση των μη γραμμικών συστημάτων. Οι συντελεστές  $w_i$  δίνονται από τη σχέση

$$w_i = \frac{2}{(1 - x_i^2)[P'_m(x_i)]^2}.$$

Τα δύο πρώτα πολυώνυμα Legendre είναι τα

$$\begin{aligned} P_0(x) &= 1, \\ P_1(x) &= x, \end{aligned}$$

ενώ τα υπόλοιπα μπορούν να παραχθούν από την αναδρομική σχέση

$$P_{n+1}(x) = \frac{1}{n+1}((2n+1)xP_n(x) - nP_{n-1}(x)), \quad n > 0.$$

Για το σφάλμα  $\varepsilon_m$  στον υπολογισμό του ολοκληρώματος με τη μέθοδο Gauss-Legendre  $m$  σημείων ισχύει

$$|\varepsilon_m| \leq \frac{2^{2m+1}(m!)^4}{(2m+1)[(2m)!]^3} f^{(2m)}(\xi), \quad \text{για κάποιο } \xi \in (-1, 1).$$

Στην πράξη, μπορούμε να υπολογίσουμε την προσεγγιστική τιμή για  $m = 1, 2, 3, \dots$  και να επιλέξουμε το μικρότερο  $m$  που θα μας δώσει ικανοποιητική προσέγγιση.

Η μέθοδος ολοκλήρωσης Gauss μπορεί να επεκταθεί και στον υπολογισμό ολοκληρωμάτων ειδικής μορφής. Έτσι έχουμε τις ακόλουθες μεθόδους:

### 5.6.2 Μέθοδος Gauss–Hermite

Σύμφωνα με αυτή τη μέθοδο

$$\int_{-\infty}^{\infty} e^{-x^2} f(x) dx \approx \sum_{i=1}^m w_i f(x_i), \quad (5.12)$$

όπου  $x_i$  είναι οι ρίζες του πολυωνύμου Hermite τάξης  $m$ ,  $H_m(x)$ , και  $w_i$  τα αντίστοιχα βάρη, τα οποία είναι τα

$$w_i = \frac{2^{m+1} m! \sqrt{\pi}}{[H'_m(x_i)]^2}.$$

Τα δύο πρώτα πολυώνυμα Hermite είναι τα

$$\begin{aligned} H_0(x) &= 1, \\ H_1(x) &= 2x, \end{aligned}$$

ενώ τα υπόλοιπα μπορούν να παραχθούν από την αναδρομική σχέση

$$H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x), \quad n > 0.$$

### 5.6.3 Μέθοδος Gauss–Laguerre

Σύμφωνα με αυτή τη μέθοδο

$$\int_0^{\infty} e^{-x} f(x) dx \approx \sum_{i=1}^m w_i f(x_i), \quad (5.13)$$

όπου  $x_i$  είναι οι ρίζες του πολυωνύμου Laguerre τάξης  $m$ ,  $L_m(x)$ , και  $w_i$  τα αντίστοιχα βάρη, τα οποία είναι τα

$$w_i = \frac{x_i}{[(m+1)L_{m+1}(x_i)]^2}.$$

Τα δύο πρώτα πολυώνυμα Laguerre είναι τα

$$\begin{aligned} L_0(x) &= 1, \\ L_1(x) &= -x + 1, \end{aligned}$$

ενώ τα υπόλοιπα μπορούν να παραχθούν από την αναδρομική σχέση

$$L_{n+1}(x) = \frac{1}{n+1}((2n+1-x)L_n(x) - nL_{n-1}(x)), \quad n > 0.$$

### 5.6.4 Μέθοδος Gauss–Chebyshev

Σύμφωνα με αυτή τη μέθοδο

$$\int_{-1}^1 \frac{1}{\sqrt{1-x^2}} f(x) dx \approx \sum_{i=1}^m w_i f(x_i), \quad (5.14)$$

$$\int_{-1}^1 \sqrt{1-x^2} f(x) dx \approx \sum_{i=1}^m c_i f(\rho_i), \quad (5.15)$$

όπου  $x_i$  είναι οι ρίζες του πολωνύμου Chebyshev *πρώτου* είδους, τάξης  $m$ ,  $T_m(x)$ , και  $\rho_i$  είναι οι ρίζες του πολωνύμου Chebyshev *δεύτερου* είδους, τάξης  $m$ ,  $U_m(x)$ . Οι ρίζες των δύο πολωνύμων μπορούν να υπολογιστούν σε κλειστή μορφή:

$$x_i = \cos\left(\frac{(2i-1)\pi}{2m}\right), \quad \rho_i = \cos\left(\frac{i\pi}{m+1}\right).$$

Τα αντίστοιχα βάρη  $w_i$ ,  $c_i$  είναι:

$$w_i = \frac{\pi}{m}, \quad c_i = \frac{\pi}{m+1}(1-\rho_i^2).$$

### 5.6.5 Κατασκευή μεθόδων Gauss

Επιθυμούμε να υπολογίσουμε προσεγγιστικά ένα ολοκλήρωμα

$$\int_a^b f(x)W(x) dx$$

όπου  $W(x)$  μια μη αρνητική συνάρτηση στο διάστημα  $[a, b]$ .

Το ολοκλήρωμα αυτό μπορεί να γραφεί ως άθροισμα των τιμών της  $f(x)$  σε συγκεκριμένα σημεία  $x_i \in (a, b)$  με κατάλληλα «βάρη»  $w_i$ :

$$\int_a^b f(x)W(x) dx \approx \sum_{i=1}^n w_i f(x_i). \quad (5.16)$$

Υπάρχει η δυνατότητα να βρούμε<sup>1</sup> μια οικογένεια ορθογώνιων πολωνύμων  $P_i(x)$ , βαθμού  $i = 0, 1, \dots$ , που ορίζονται στο διάστημα  $[a, b]$  και έχουν συνάρτηση βάρους  $W(x)$ , ικανοποιούν δηλαδή τη σχέση

$$\int_a^b P_i(x)P_j(x)W(x) dx = \delta_{ij}.$$

Οι ρίζες του πολωνύμου  $P_n(x)$  είναι τα ζητούμενα σημεία  $x_i$  στον τύπο (5.16).

<sup>1</sup>ή να κατασκευάσουμε με τον αλγόριθμο Gram-Schmidt από τη βάση  $1, x, x^2, \dots$



Τα ορθογώνια πολυώνυμα ικανοποιούν τις σχέσεις

$$a_i P_{i-1} + c_i P_{i+1} = (x - b_i) P_i, \quad i > 0,$$

και

$$\begin{aligned} P_0 &= 1, \\ c_0 P_1 &= (x - b_0) P_0. \end{aligned}$$

Οι αναδρομικές σχέσεις μεταξύ των πολωνύμων μπορούν να γραφούν με τη μορφή πινάκων:

$$T \cdot \begin{pmatrix} P_0 \\ P_1 \\ P_2 \\ \vdots \\ P_{n-2} \\ P_{n-1} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ P_n \end{pmatrix} = x \begin{pmatrix} P_0 \\ P_1 \\ P_2 \\ \vdots \\ P_{n-2} \\ P_{n-1} \end{pmatrix}, \quad (5.17)$$

όπου

$$T = \begin{pmatrix} b_0 & c_0 & 0 & 0 & \cdots & 0 \\ a_1 & b_1 & c_1 & 0 & \cdots & 0 \\ 0 & a_2 & b_2 & c_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{n-2} & b_{n-2} & c_{n-2} \\ 0 & \cdots & 0 & 0 & a_{n-1} & b_{n-1} \end{pmatrix}.$$

Αν  $P_n(x) = 0$  τότε η (5.17) υποδηλώνει ότι το  $x$  είναι ιδιοτιμή του πίνακα  $T$ . Με κατάλληλο μετασχηματισμό ομοιότητας ο πίνακας  $T$  γίνεται συμμετρικός (και διατηρεί τις ίδιες ιδιοτιμές):

$$J = D^{-1} T D,$$

όπου  $D = \text{diag}(d_1, d_2, \dots, d_n)$  με

$$\begin{aligned} d_1 &= 1 \\ d_j &= d_{j-1} \sqrt{\frac{a_j}{c_{j-1}}}. \end{aligned}$$

Θεωρούμε ότι  $a_j c_{j-1} > 0$ , για κάθε  $j > 0$ .

Από τα παραπάνω συνάγεται ότι οι ρίζες του  $P_n(x)$  είναι οι ιδιοτιμές του πίνακα Jacobi,  $J$ , ενός συμμετρικού τριδιαγώνιου πίνακα:

$$J = \begin{pmatrix} b_0 & s_1 & 0 & 0 & \cdots & 0 \\ s_1 & b_1 & s_2 & 0 & \cdots & 0 \\ 0 & s_2 & b_2 & s_3 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & s_{n-2} & b_{n-2} & s_{n-1} \\ 0 & \cdots & 0 & 0 & s_{n-1} & b_{n-1} \end{pmatrix},$$

όπου  $s_j = \sqrt{a_j c_{j-1}}$ .

Το βάρος  $w_i$  που αντιστοιχεί στην ιδιοτιμή  $x_i$  στον τύπο (5.16) μπορεί να υπολογιστεί από το αντίστοιχο ιδιοδιάνυσμα του  $J$ ,  $\mathbf{v}^{(i)}$ . Αν είναι κανονικοποιημένο ώστε  $\|\mathbf{v}^{(i)}\| = 1$ , τότε

$$w_i = \left(v_1^{(i)}\right)^2 \int_a^b W(x) dx ,$$

όπου  $v_1^{(i)}$  είναι η πρώτη συνιστώσα του  $\mathbf{v}^{(i)}$ .

Η διαδικασία που περιγράφηκε αποτελεί τον αλγόριθμο Golub–Welsch.

## 5.7 Μέθοδος Clenshaw–Curtis

Σύμφωνα με τον κανόνα ολοκλήρωσης Clenshaw–Curtis, μπορούμε να υπολογίσουμε ένα ολοκλήρωμα της μορφής

$$\int_{-1}^1 f(x) dx$$

ως εξής: επιλέγουμε τα  $n+1$  (με  $n > 1$ ) μη ισαπέχοντα σημεία

$$x_i = \cos\left(\frac{i\pi}{n}\right) , \quad i = 0, \dots, n$$

στο διάστημα της ολοκλήρωσης. Κατόπιν, βρίσκουμε το πολυώνυμο προσέγγισης που περνά από τα σημεία  $(x_i, f(x_i))$ , το οποίο ολοκληρώνουμε ακριβώς.

Μπορεί να δειχθεί ότι στον βασικό τύπο ολοκλήρωσης (5.1),

$$\int_{-1}^1 f(x) dx \approx \sum_{i=0}^n w_i f(x_i) ,$$

οι συντελεστές  $w_i$  για τη μέθοδο Clenshaw–Curtis είναι

$$w_i = \frac{c_i}{n} \left( 1 - \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{b_j}{4j^2 - 1} \cos\left(\frac{2ij\pi}{n}\right) \right) , \quad i = 0, \dots, n ,$$

όπου  $\lfloor x \rfloor$  το ακέραιο μέρος του  $x$  και

$$\begin{aligned} b_i &= \begin{cases} 1 , & i = n/2 \\ 2 , & i < n/2 \end{cases} , \\ c_i &= \begin{cases} 1 , & i \bmod n = 0 \\ 2 , & i \bmod n \neq 0 \end{cases} . \end{aligned}$$

Η μέθοδος Clenshaw–Curtis υπολογίζει το ζητούμενο ολοκλήρωμα με ακρίβεια συγκρίσιμη με τη μέθοδο Gauss–Legendre  $n$  σημείων. Έχει τα πλεονεκτήματα έναντι αυτής ότι

- οι κόμβοι  $x_i$  υπολογίζονται εύκολα,
- οι συντελεστές  $w_i$  μπορούν να προκύψουν από αλγόριθμους για γρήγορο υπολογισμό του διακριτού μετασχηματισμού Fourier, §6.9.1,
- διαδοχικές εφαρμογές του τύπου για  $n, 2n, 4n, \dots$  που χρειάζονται για την εκτίμηση της ακρίβειάς της, χρησιμοποιούν κοινούς κόμβους.

## 5.8 Ειδική Περίπτωση

### 5.8.1 Ολοκλήρωση σε άνισα τμήματα

Στην περίπτωση που η συνάρτηση  $f(x)$  δεν είναι δεδομένη αλλά τα σημεία της δίνονται με την μορφή πινάκων (π.χ. από πειραματικές μετρήσεις), τότε είναι πιθανό τα σημεία στα οποία ορίζεται η συνάρτησή μας να μην ισαπέχουν ή να μην έχουν την κατανομή που χρειάζεται για τις μεθόδους Gauss. Τότε έχουμε διάφορες εναλλακτικές δυνατότητες:

- Χρήση της μεθόδου του τραπεζίου (ή ισοδύναμα, ολοκλήρωση της προσέγγισης με ευθύγραμμα τμήματα, §4.3). Ο (απλός) τύπος του τραπεζίου μπορεί να εφαρμοστεί σε κάθε διάστημα  $[x_i, x_{i+1}]$  (με μήκος  $h_i = x_{i+1} - x_i$ ) ώστε να προκύψει ο τύπος:

$$\begin{aligned} \int_{x_0}^{x_N} f(x) dx &= h_0 \frac{f(x_0) + f(x_1)}{2} + h_1 \frac{f(x_1) + f(x_2)}{2} + \dots \\ &\quad + h_{N-2} \frac{f(x_{N-2}) + f(x_{N-1})}{2} + h_{N-1} \frac{f(x_{N-1}) + f(x_N)}{2}. \end{aligned}$$

Εάν γειτονικά τμήματα είναι ίσα τότε μπορεί να εφαρμοστεί ένας τύπος Newton–Cotes (π.χ. Simpson) υψηλότερης τάξης.

- Ολοκλήρωση του πολυωνύμου προσέγγισης, §4.1, στο διάστημα ορισμού των δεδομένων. Η μέθοδος αυτή δεν είναι ακριβής για μεγάλο αριθμό σημείων  $N$ , εξαιτίας της υψηλής τάξης πολυωνύμου που δημιουργείται, §4.1.2.
- Ολοκλήρωση της καμπύλης spline, §4.4, που προσδιορίζεται από τα δεδομένα σημεία. Είναι η πιο ακριβής μέθοδος.

## 5.9 Ασκήσεις

1. (α') Υλοποιήστε τον αλγόριθμο τραπεζίου σε υποπρόγραμμα. Αυτό θα δέχεται ως ορίσματα τουλάχιστον τα όρια της ολοκλήρωσης και το πλήθος των διαστημάτων. Θα επιστρέφει την προσεγγιστική τιμή του ολοκληρώματος.

(β') Χρησιμοποιήστε το υποπρόγραμμα για να υπολογίσετε το ολοκλήρωμα

$$\int_0^{\pi} \sin x \, dx$$

διαδοχικά με  $N = 2, 4, 8, 16, \dots, 512$  διαστήματα. Το πρόγραμμά σας να τυπώνει για κάθε  $N$  την υπολογιζόμενη τιμή και την απόλυτη διαφορά της από την ακριβή τιμή.

2. (α') Υλοποιήστε τον αλγόριθμο Simpson σε υποπρόγραμμα. Αυτό θα δέχεται ως ορίσματα τουλάχιστον τα όρια της ολοκλήρωσης και το πλήθος των διαστημάτων. Θα επιστρέφει την προσεγγιστική τιμή του ολοκληρώματος.

(β') Χρησιμοποιήστε το για να υπολογίσετε το ολοκλήρωμα

$$\int_0^{\pi} \sin x \, dx$$

με όσα διαστήματα χρειάζεται ώστε να έχετε ακρίβεια τουλάχιστον 6 ψηφίων.

Υπόδειξη: Επιλέξτε κατάλληλα το βήμα (άρα και το πλήθος των διαστημάτων) ώστε το σφάλμα §(5.10) να είναι μικρότερο από  $10^{-6}$ .

3. Υλοποιήστε ένα υποπρόγραμμα που να υπολογίζει ολοκληρώματα ανεξάρτητα με το πλήθος των σημείων στα οποία είναι γνωστή η ολοκληρωτέα συνάρτηση. Αν το πλήθος των διαστημάτων είναι περιττό (και μεγαλύτερο του 3), να χρησιμοποιεί τον τύπο  $3/8$  Simpson για τα πρώτα 3 και για τα υπόλοιπα τον τύπο  $1/3$  Simpson. Αν είναι άρτιο, να χρησιμοποιεί μόνο τον  $1/3$  Simpson.

4. Υπολογίστε προσεγγιστικά με ακρίβεια  $10^{-6}$  τα ολοκληρώματα στο διάστημα  $[0, 3]$  των συναρτήσεων

(α')  $f(x) = 2x + 1,$

(β')  $f(x) = x^2 \sqrt{x},$

(γ')  $f(x) = \frac{1}{1 + x^2},$

(δ')  $f(x) = \frac{1}{1 + (x - \pi)^2},$

(ε')  $f(x) = \frac{1}{2 + \cos x},$

(στ')  $f(x) = \cos(4x)e^x,$

(ζ')  $f(x) = e^{\cos x},$

(η')  $f(x) = \sqrt{x}.$

5. Έστω

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

συνεχής συνάρτηση στο  $[-1, 1]$  χωρίς παράγωγο στο  $x = 0$ . Υπολογίστε το σφάλμα

$$\varepsilon_n = I - I_n = \int_{-1}^1 f(x) dx - I_n,$$

όπου  $I_n$  ο τύπος τραπεζίου με  $n$  υποδιαίρεσεις, και δείξτε ότι  $|\varepsilon_n| \leq Ch$ , όπου  $h = 2/n$ . (Υποθέστε ότι το  $n$  είναι άρτιος ή περιττός.)

6. **Παρέκταση Richardson για ολοκληρώματα (Μέθοδος Romberg).** Μπορεί ναδειχθεί ότι ο εκτεταμένος τύπος τραπεζίου για το ολοκλήρωμα,

$$I_0 = \int_{x_0}^{x_n} f(x) dx,$$

δίνει για την ακριβή τιμή τη σχέση

$$I_0 = I_h + \alpha_2 h^2 + \alpha_4 h^4 + \dots, \quad (5.18)$$

όπου

$$I_h = \frac{h}{2} (f_0 + 2f_1 + 2f_2 + \dots + 2f_{n-1} + f_n),$$

$h = (x_n - x_0)/n$  και  $\alpha_i$  οι συντελεστές των όρων  $h^i$  του σφάλματος.

Γράψτε τη (5.18) για τρία διαφορετικά βήματα, π.χ.  $h, h/2, h/4$ . Παρατηρήστε ότι σχηματίζεται ένα σύστημα τριών γραμμικών εξισώσεων με αγνώστους τα  $I_0, \alpha_2, \alpha_4$ . Βρείτε τη λύση του συστήματος ως προς  $I_0$ . ο τύπος στον οποίο θα καταλήξετε—γραμμικός συνδυασμός των  $I_h, I_{h/2}, I_{h/4}$  που έχουν σφάλματα  $O(h^2)$ —δίνει την ακριβή τιμή του ολοκληρώματος με σφάλμα  $O(h^6)$ .

[ Λύση συστήματος:  $= (I_h - 20I_{h/2} + 64I_{h/4})/45$ . ]

Υλοποιήστε σε κώδικα τον παραπάνω αλγόριθμο ολοκλήρωσης.

7. Υλοποιήστε σε κώδικα τη μέθοδο ολοκλήρωσης Gauss για 2 και για 3 σημεία. Εφαρμόστε τη για να υπολογίσετε το ολοκλήρωμα

$$\int_{2.1}^{5.2} x^3 e^{-x} dx.$$

[Σωστή τιμή ολοκληρώματος: 3.60346...]

8. Τα πρώτα πολυώνυμα Hermite είναι τα

$$\begin{aligned} H_0(x) &= 1, & H_1(x) &= 2x \\ H_2(x) &= 4x^2 - 2, & H_3(x) &= 8x^3 - 12x \\ H_4(x) &= 16x^4 - 48x^2 + 12. \end{aligned}$$

Να γράψετε υποπρόγραμμα που να υλοποιεί τη μέθοδο Gauss–Hermite για  $n = 4$ . Χρησιμοποιήστε το για να υπολογίσετε το ολοκλήρωμα

$$\int_{-\infty}^{\infty} e^{-x^2} x^2 dx .$$

Συγκρίνετε με την ακριβή τιμή  $(\sqrt{\pi}/2)$ .

9. Τα πρώτα πολυώνυμα Laguerre είναι τα

$$\begin{aligned} L_0(x) &= 1 \\ L_1(x) &= -x + 1 \\ L_2(x) &= (x^2 - 4x + 2)/2 \\ L_3(x) &= (-x^3 + 9x^2 - 18x + 6)/6 \\ L_4(x) &= (x^4 - 16x^3 + 72x^2 - 96x + 24)/24 \\ L_5(x) &= (-x^5 + 25x^4 - 200x^3 + 600x^2 - 600x + 120)/120 . \end{aligned}$$

Να γράψετε υποπρόγραμμα που να υλοποιεί τη μέθοδο Gauss–Laguerre για  $n = 4$ . Χρησιμοποιήστε το για να υπολογίσετε το ολοκλήρωμα

$$\int_0^{\infty} e^{-x} (x^6 - 3\sqrt{x} + 2) dx .$$

Συγκρίνετε με την ακριβή τιμή  $(6! - 3\sqrt{\pi}/2 + 2 \times 0!)$ .

Υπόδειξη: Το  $L_4(x)$  έχει τις 4 ρίζες του πραγματικές στο διάστημα  $[0, 10]$ .

10. Να γράψετε υποπρόγραμμα που να υλοποιεί τη μέθοδο Gauss–Chebyshev για  $n = 5$ . Χρησιμοποιήστε το για να υπολογίσετε το ολοκλήρωμα

$$\int_{-1}^1 \frac{x^2 e^{-x}}{\sqrt{1-x^2}} dx .$$

Συγκρίνετε με τη σωστή τιμή  $(0.7009067737595233 \dots \times \pi)$ .

11. Υπολογίστε με τη μέθοδο Clenshaw–Curtis το ολοκλήρωμα

$$\int_{-2}^2 \frac{1}{1+x^2} dx .$$

Πόσα σημεία χρειάζονται για να προσεγγίσετε με 12 ψηφία την ακριβή τιμή  $(2 \tan^{-1}(2))$ ;

12. Το ολοκλήρωμα

$$\int_{-1}^1 f(x) dx$$

μπορεί να υπολογιστεί προσεγγιστικά από τύπο της μορφής

$$\int_{-1}^1 f(x) \, dx \approx \sum_{k=1}^N a_k f(x_k) , \quad (5.19)$$

όπου  $x_1, x_2, \dots, x_N$  διακριτά σημεία της επιλογής μας στο διάστημα  $[-1, 1]$ .

Έστω ότι επιλέγουμε να είναι το  $N = 7$  και τα σημεία  $x_k$  τα  $-0.9, -0.7, -0.4, 0.1, 0.4, 0.8, 0.9$ . Προσδιορίστε τα  $a_k$  ( $k = 1, \dots, 7$ ) ώστε ο τύπος (5.19) να είναι ακριβής για τις συναρτήσεις  $f_0(x) = 1$ ,  $f_1(x) = x$ ,  $f_2(x) = x^2$ , ...,  $f_6(x) = x^6$ . Κατόπιν, χρησιμοποιήστε τον για να υπολογίσετε το ολοκλήρωμα

$$\int_{-1}^1 x^3 \sin(\pi x) \, dx .$$





## Κεφάλαιο 6

# Ανάλυση Fourier

Πολλά προβλήματα στη Φυσική αφορούν ταλαντώσεις και κύματα. Ένα ηλεκτρομαγνητικό κύμα (ακτινοβολία), το εναλλασσόμενο ρεύμα σε ένα ηλεκτρικό κύκλωμα, η δόνηση μιας χορδής ή ενός μέσου (ήχος) είναι γενικά μια επαλληλία κυμάτων, το καθένα με συγκεκριμένη συχνότητα. Η ανάλυση Fourier μας δίνει τη δυνατότητα να αναπτύξουμε τέτοιες περιοδικές (αλλά και μη περιοδικές) συναρτήσεις του χρόνου ή κάποιας απόστασης, στα κύματα που τις αποτελούν. Επιπλέον, η ανάλυση Fourier βρίσκει εφαρμογή στη Μαθηματική Φυσική για την επίλυση διαφορικών εξισώσεων.

### 6.1 Ορισμοί

#### 6.1.1 Συνεχής συνάρτηση

Μια συνάρτηση  $f(x)$  είναι *συνεχής* σε ένα σημείο  $x_0$  στο πεδίο ορισμού της αν ικανοποιεί τη σχέση

$$\lim_{x \rightarrow x_0} f(x) = f(x_0) . \quad (6.1)$$

Σε αυτό τον ορισμό θεωρούμε ότι

- το  $x_0$  δεν είναι απομονωμένο σημείο, έχει δηλαδή, γειτονικά σημεία που ανήκουν στο πεδίο ορισμού, και
- η τιμή του ορίου είναι ανεξάρτητη από την κατεύθυνση από την οποία το σημείο  $x$  πλησιάζει το  $x_0$ .

Ο συγκεκριμένος ορισμός με απλά λόγια σημαίνει ότι το  $f(x)$  πλησιάζει όσο κοντά θέλουμε στο  $f(x_0)$  όταν το  $x$  πλησιάζει το  $x_0$  από οποιαδήποτε κατεύθυνση. Ισοδύναμα, μια συνάρτηση είναι συνεχής σε σημείο  $x_0$  αν τα δύο όριά της, από μικρότερες και μεγαλύτερες τιμές, ταυτίζονται με την τιμή της στο  $x_0$ :

$$\lim_{\varepsilon \rightarrow 0} f(x_0 - \varepsilon) = \lim_{\varepsilon \rightarrow 0} f(x_0 + \varepsilon) = f(x_0) .$$

### 6.1.2 Περιοδική συνάρτηση

Μια συνεχής συνάρτηση  $f(x)$  λέγεται *περιοδική* με (μη μηδενική) περίοδο  $L$ , αν ικανοποιεί τη σχέση

$$f(x + L) = f(x) , \quad (6.2)$$

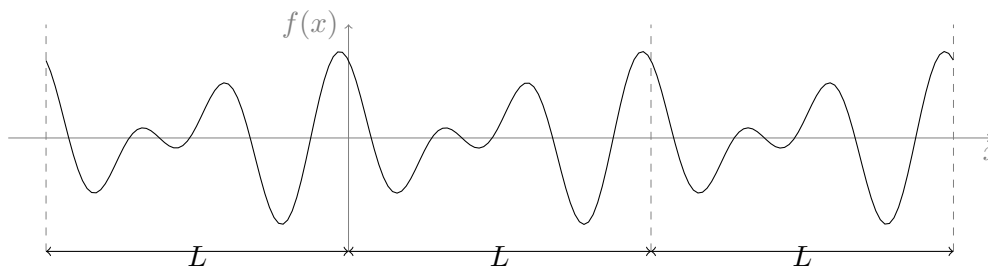
για όλα τα σημεία  $x$  που ανήκουν στο πεδίο ορισμού της. Αν το  $L$  είναι το μικρότερο διάστημα που ικανοποιεί τη σχέση (6.2), τότε εύκολα δείχνεται ότι κάθε πολλαπλάσιο του  $L$  είναι επίσης περίοδος:

$$f(x + mL) \equiv f(x + (m - 1)L + L) = f(x + (m - 1)L) = \cdots = f(x) ,$$

όπου  $m$  θετικός ακέραιος. Παρατηρήστε ακόμα ότι αν μια συνάρτηση έχει περίοδο υποπολλαπλάσιο του  $L$ ,  $L/m$ , τότε το  $L$  είναι επίσης περίοδος της.

#### Παράδειγμα

Η συνάρτηση στο Σχήμα 6.1 είναι περιοδική.



Σχήμα 6.1: Περιοδική συνάρτηση με περίοδο  $L$

### 6.1.3 Συνθήκες Dirichlet

Μια πραγματική συνάρτηση πραγματικής μεταβλητής  $f(x)$ , που είναι περιοδική, λέμε ότι ικανοποιεί τις *συνθήκες Dirichlet* αν σε οποιοδήποτε πεπερασμένο διάστημα στο πεδίο ορισμού της:

- Είναι μονότιμη και συνεχής, εκτός ίσως από πεπερασμένο πλήθος διακριτών σημείων στα οποία εμφανίζει ασυνέχεια, χωρίς όμως να απειρίζεται.
- Έχει πεπερασμένο πλήθος μέγιστων και ελάχιστων.
- Ορίζεται και έχει πεπερασμένη τιμή το ολοκλήρωμα της  $|f(x)|$  (όπως λέμε, η  $f(x)$  είναι απόλυτα ολοκληρώσιμη).

Οι συνθήκες αυτές είναι πολύ γενικές και οι περιοδικές συναρτήσεις που θα συναντήσουμε σε ρεαλιστικές εφαρμογές τις ικανοποιούν.

**Παράδειγμα**

Η συνάρτηση στο Σχήμα 6.1 ικανοποιεί τις συνθήκες Dirichlet.

**6.2 Σειρά Fourier**

Μια περιοδική συνάρτηση  $f(x)$  που ικανοποιεί τις συνθήκες Dirichlet μπορεί να αναπαρασταθεί ως άθροισμα άπειρων τριγωνομετρικών συναρτήσεων (ημίτονων και συνημίτονων) με κατάλληλα πλάτη και φάσεις. Το άθροισμα αυτό συγκλίνει στην  $f(x)$  σε κάθε σημείο που αυτή είναι συνεχής.

Οι όροι του αθροίσματος είναι της μορφής

$$A_m \cos\left(\frac{2m\pi x}{L}\right) \quad \text{ή} \quad B_m \sin\left(\frac{2m\pi x}{L}\right),$$

με  $m$  μη αρνητικό ακέραιο<sup>1</sup>. Παρατηρήστε ότι κάθε τέτοιος όρος είναι μια περιοδική συνάρτηση με περίοδο  $L/m$ . Το άθροισμα τέτοιων συναρτήσεων, με διάφορα πλάτη  $A_m$ ,  $B_m$  και περιόδους υποπολλαπλάσιες του  $L$  ( $L$ ,  $L/2$ ,  $L/3$ , ...), αποτελεί τη *σειρά Fourier* για μια συνάρτηση που ικανοποιεί τις συνθήκες Dirichlet και είναι περιοδική με περίοδο  $L$ . Δηλαδή

$$\begin{aligned} f(x) &= \frac{A_0}{2} \\ &+ A_1 \cos\left(\frac{2\pi x}{L}\right) + B_1 \sin\left(\frac{2\pi x}{L}\right) \\ &+ A_2 \cos\left(2\frac{2\pi x}{L}\right) + B_2 \sin\left(2\frac{2\pi x}{L}\right) \\ &+ A_3 \cos\left(3\frac{2\pi x}{L}\right) + B_3 \sin\left(3\frac{2\pi x}{L}\right) \\ &\vdots \\ &= \frac{A_0}{2} + \sum_{m=1}^{\infty} A_m \cos\left(\frac{2m\pi x}{L}\right) + \sum_{m=1}^{\infty} B_m \sin\left(\frac{2m\pi x}{L}\right). \end{aligned} \quad (6.3)$$

Οι συντελεστές  $A_m$ ,  $B_m$  εξαρτώνται από την  $f(x)$  και θα υπολογιστούν στην επόμενη παράγραφο. Εκεί θα φανεί και ο λόγος της ιδιαίτερης μορφής του σταθερού όρου,  $A_0/2$ .

Παρατηρήστε ότι η σειρά Fourier είναι παντού συνεχής ενώ η  $f(x)$  μπορεί να έχει σημεία ασυνέχειας. Σε αυτά τα σημεία, η τιμή που παίρνει η σειρά Fourier

<sup>1</sup> Αν η ανεξάρτητη μεταβλητή  $x$  είναι μήκος, η περίοδος  $L$  λέγεται μήκος κύματος και συμβολίζεται συνήθως με το  $\lambda$ . Η ποσότητα  $2\pi/\lambda$  λέγεται (γωνιακός) κυματάρηθος και συμβολίζεται συχνά με το  $k$ . Αν η ανεξάρτητη μεταβλητή συμβολίζει χρόνο, η περίοδος  $L$  συμβολίζεται με το  $T$ . Η ποσότητα  $2\pi/T$  λέγεται γωνιακή συχνότητα και συμβολίζεται με το  $\omega$ .

είναι ο μέσος όρος του δεξιού και του αριστερού ορίου της  $f(x)$ :

$$\frac{1}{2} \lim_{\varepsilon \rightarrow 0} (f(x_0 - \varepsilon) + f(x_0 + \varepsilon)) .$$

### 6.3 Υπολογισμός συντελεστών της σειράς Fourier

Οι άγνωστοι συντελεστές  $A_0, A_1, \dots, B_1, \dots$  στην εξίσωση (6.3) λέγονται *συντελεστές Fourier* και υπολογίζονται ως εξής:

Πολλαπλασιάζουμε τα δύο μέλη της (6.3) με την ποσότητα  $\cos(2n\pi x/L)$  και ολοκληρώνουμε σε διάστημα μίας περιόδου:

$$\begin{aligned} \int_0^L \cos\left(\frac{2n\pi x}{L}\right) f(x) dx &= \frac{A_0}{2} \int_0^L \cos\left(\frac{2n\pi x}{L}\right) dx \\ &+ \sum_{m=1}^{\infty} A_m \int_0^L \cos\left(\frac{2n\pi x}{L}\right) \cos\left(\frac{2m\pi x}{L}\right) dx \\ &+ \sum_{m=1}^{\infty} B_m \int_0^L \cos\left(\frac{2n\pi x}{L}\right) \sin\left(\frac{2m\pi x}{L}\right) dx . \end{aligned}$$

Χρησιμοποιώντας σχέσεις από το τυπολόγιο στο Παράρτημα α' προκύπτει ότι

$$\int_0^L \cos\left(\frac{2n\pi x}{L}\right) f(x) dx = A_n \frac{L}{2} , \quad n \geq 0 .$$

Παρατηρήστε ότι η επιλογή να έχει ο σταθερός όρος τη μορφή  $A_0/2$  δίνει ενιαία μορφή στο γενικό τύπο για τα  $A_n$ , για κάθε  $n$ .

Αντίστοιχα, αν πολλαπλασιάσουμε τα δύο μέλη της (6.3) με την ποσότητα  $\sin(2n\pi x/L)$  και ολοκληρώσουμε στο διάστημα  $[0, L]$ , έχουμε

$$\begin{aligned} \int_0^L \sin\left(\frac{2n\pi x}{L}\right) f(x) dx &= \frac{A_0}{2} \int_0^L \sin\left(\frac{2n\pi x}{L}\right) dx \\ &+ \sum_{m=1}^{\infty} A_m \int_0^L \sin\left(\frac{2n\pi x}{L}\right) \cos\left(\frac{2m\pi x}{L}\right) dx \\ &+ \sum_{m=1}^{\infty} B_m \int_0^L \sin\left(\frac{2n\pi x}{L}\right) \sin\left(\frac{2m\pi x}{L}\right) dx . \end{aligned}$$

Χρησιμοποιώντας σχέσεις από το τυπολόγιο στο Παράρτημα α' προκύπτει ότι

$$\int_0^L \sin\left(\frac{2n\pi x}{L}\right) f(x) dx = B_n \frac{L}{2} , \quad n > 0 .$$

Συγκεντρωτικά, για τους πραγματικούς συντελεστές  $A_n, B_n$  έχουμε

$$A_n = \frac{2}{L} \int_0^L \cos\left(\frac{2n\pi x}{L}\right) f(x) dx, \quad n \geq 0, \quad (6.4\alpha')$$

$$B_n = \frac{2}{L} \int_0^L \sin\left(\frac{2n\pi x}{L}\right) f(x) dx, \quad n > 0. \quad (6.4\beta')$$

**Παρατήρηση:** Οι συναρτήσεις  $f(x)$ ,  $\cos(2n\pi x/L)$ ,  $\sin(2n\pi x/L)$  που εμφανίζονται στα ολοκληρώματα είναι περιοδικές με περίοδο  $L$ :

$$\begin{aligned} f(x+L) &= f(x), & (\text{εξ ορισμού}) \\ \cos(2n\pi(x+L)/L) &= \cos(2n\pi x/L + 2n\pi) = \cos(2n\pi x/L), \\ \sin(2n\pi(x+L)/L) &= \sin(2n\pi x/L + 2n\pi) = \sin(2n\pi x/L). \end{aligned}$$

Επομένως και οι συναρτήσεις  $\cos(2n\pi x/L)f(x)$  και  $\sin(2n\pi x/L)f(x)$  είναι περιοδικές με την ίδια περίοδο.

Ένα ολοκλήρωμα μιας περιοδικής συνάρτησης  $g(x)$  σε μήκος ίσο με την περίοδό της,  $L$ , είναι το ίδιο, ανεξάρτητα από την αρχή (το κάτω όριο) της ολοκλήρωσης: έστω  $a$  αυθαίρετο σημείο στο πεδίο ορισμού της συνάρτησης. Τότε

$$\int_a^{a+L} g(x) dx = \int_a^L g(x) dx + \int_L^{a+L} g(x) dx.$$

Όμως

$$\int_L^{a+L} g(x) dx \stackrel{y=x-L}{=} \int_0^a g(y+L) dy = \int_0^a g(y) dy.$$

Η τελευταία ισότητα προέκυψε από την περιοδικότητα (εξίσωση (6.2)) της  $g(x)$ . Συνολικά έχουμε

$$\int_a^{a+L} g(x) dx = \int_a^L g(x) dx + \int_0^a g(y) dy = \int_0^L g(x) dx.$$

Επομένως, σε ολοκλήρωμα περιοδικής συνάρτησης σε μία περίοδό της μπορούμε να επιλέξουμε το διάστημα ολοκλήρωσης να είναι  $[0, L]$  ή  $[-L/2, L/2]$  ή οποιοδήποτε άλλο μας διευκολύνει, αρκεί να έχει μήκος μία περίοδο. Η τιμή του ολοκληρώματος θα είναι η ίδια ανεξάρτητα από την αρχή του διαστήματος.

### 6.3.1 Ιδιότητες

Εύκολα μπορούμε να δούμε από τους ορισμούς των συντελεστών Fourier, (6.4α') και (6.4β'), ότι:

- Αν η συνάρτηση  $f(x)$  ικανοποιεί τις συνθήκες Dirichlet, έχει περίοδο  $L$  και συντελεστές Fourier  $A_m, B_m$ , η συνάρτηση  $\lambda f(x)$  με  $\lambda$  πραγματική σταθερά, ικανοποιεί επίσης τις συνθήκες Dirichlet, έχει περίοδο  $L$  και αναπτύσσεται σε σειρά Fourier με συντελεστές  $\lambda A_m, \lambda B_m$ .

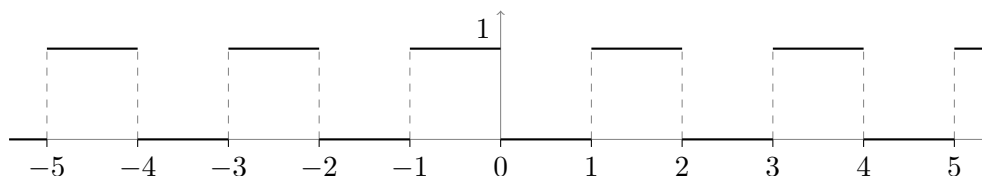
- Αν δύο συναρτήσεις  $f_1(x)$  και  $f_2(x)$  ικανοποιούν τις συνθήκες Dirichlet, έχουν ίδια περίοδο  $L$  και συντελεστές Fourier  $A_m^{(1)}, B_m^{(1)}$  και  $A_m^{(2)}, B_m^{(2)}$  αντίστοιχα, τότε και το άθροισμά τους,  $f_1(x) + f_2(x)$ , ικανοποιεί τις συνθήκες Dirichlet, έχει ίδια περίοδο  $L$  και συντελεστές Fourier  $A_m = A_m^{(1)} + A_m^{(2)}$  και  $B_m = B_m^{(1)} + B_m^{(2)}$ .

### 6.3.2 Παράδειγμα

Έστω η συνάρτηση (τετραγωνικός παλμός)

$$f(x) = \begin{cases} 0, & 0 \leq x < 1, \\ 1, & 1 \leq x < 2, \end{cases} \quad (6.5)$$

που επαναλαμβάνεται για  $x \geq 2$  και  $x < 0$  ώστε  $f(x + 2m) = f(x)$  με οποιοδήποτε θετικό ή αρνητικό ακέραιο  $m$ . Η γραφική της παράσταση δίνεται στο Σχήμα 6.2.



Σχήμα 6.2: Τετραγωνικός παλμός, εξίσωση (6.5)

Η συγκεκριμένη συνάρτηση είναι συνεχής με πεπερασμένες ασυνέχειες στα σημεία 0, 1 και είναι περιοδική με περίοδο  $L = 2$ . Το ολοκλήρωμα της  $|f(x)|$  στο  $[0, 2)$  ορίζεται και είναι πεπερασμένο:

$$\int_0^2 |f(x)| dx = 1.$$

Η  $f(x)$  ικανοποιεί τις συνθήκες Dirichlet οπότε μπορεί να αναπτυχθεί σε σειρά Fourier:

$$f(x) = \frac{A_0}{2} + \sum_{m=1}^{\infty} A_m \cos(m\pi x) + \sum_{m=1}^{\infty} B_m \sin(m\pi x),$$

με συντελεστές που υπολογίζονται από τους τύπους (6.4α'), (6.4β'):

$$\begin{aligned} A_m &= \int_0^2 \cos(m\pi x) f(x) dx = \int_1^2 \cos(m\pi x) dx = \delta_{m0}, \quad m \geq 0, \\ B_m &= \int_0^2 \sin(m\pi x) f(x) dx = \int_1^2 \sin(m\pi x) dx = \frac{(-1)^m - 1}{m\pi} \\ &= \begin{cases} 0, & m = 2, 4, 6, \dots, \\ -\frac{2}{m\pi}, & m = 1, 3, 5, \dots \end{cases} \end{aligned}$$

Στο  $A_m$  χρησιμοποιήθηκε το δέλτα του Kronecker,  $\delta_{ij}$ . Αυτό έχει τιμή 1 αν  $i = j$  και 0 αν  $i \neq j$ .

Επομένως, ο συγκεκριμένος τετραγωνικός παλμός αναπτύσσεται σε σειρά Fourier ως εξής

$$\begin{aligned} f(x) &= \frac{1}{2} + \sum_{m=1}^{\infty} B_m \sin(k\pi x) = \frac{1}{2} - 2 \sum_{j=0}^{\infty} \frac{\sin((2j+1)\pi x)}{(2j+1)\pi} \\ &= \frac{1}{2} - 2 \left( \frac{\sin(\pi x)}{\pi} + \frac{\sin(3\pi x)}{3\pi} + \frac{\sin(5\pi x)}{5\pi} + \frac{\sin(7\pi x)}{7\pi} + \dots \right). \end{aligned} \quad (6.6)$$

Παρατηρήστε ότι στα σημεία ασυνέχειας της  $f(x)$ , στα 0 και 1, η σειρά Fourier είναι συνεχής και έχει τιμή  $1/2$ , όσο το ημίθροισμα του δεξιού και του αριστερού ορίου της  $f(x)$  σε καθένα από τα δύο σημεία.

### 6.3.3 Συντελεστές Fourier συνάρτησης με συμμετρία

Παρατηρήστε ότι η συνάρτηση  $c(x) = \cos(2n\pi x/L)$  στην εξίσωση (6.4α') είναι συμμετρική ως προς το  $x = L/2$ : αν  $\bar{x}$  ένα οποιοδήποτε μήκος, ισχύει

$$c\left(\frac{L}{2} - \bar{x}\right) = c\left(\frac{L}{2} + \bar{x}\right).$$

Αντίστοιχα, η συνάρτηση  $s(x) = \sin(2n\pi x/L)$  της εξίσωσης (6.4β') είναι αντισυμμετρική ως προς το  $x = L/2$ :

$$s\left(\frac{L}{2} - \bar{x}\right) = -s\left(\frac{L}{2} + \bar{x}\right).$$

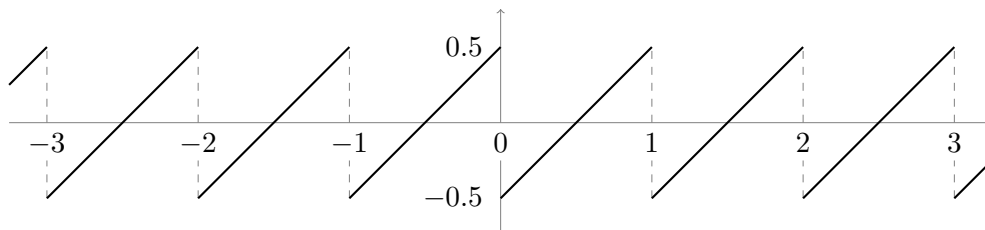
Επομένως, ο υπολογισμός των ολοκληρωμάτων που εκφράζουν τους συντελεστές Fourier απλοποιείται όταν η συνάρτηση  $f(x)$  είναι συμμετρική ή αντισυμμετρική ως προς το  $x = L/2$  (ή γενικότερα, το μέσο οποιουδήποτε διαστήματος με μήκος  $L$ ). Αν είναι συμμετρική, τότε  $B_n = 0$  και η σειρά Fourier περιέχει μόνο συνημίτονα (και σταθερό όρο), δηλαδή, τους συμμετρικούς όρους. Αν είναι αντισυμμετρική, έχουμε  $A_n = 0$  και η σειρά περιέχει μόνο ημίτονα, δηλαδή τους αντισυμμετρικούς όρους της. Στην περίπτωση που η  $f(x)$  δεν έχει συγκεκριμένη συμμετρία ως προς το  $L/2$ , η σειρά περιλαμβάνει γενικά όλους τους όρους.

Με βάση τα παραπάνω, η σειρά Fourier του τετραγωνικού παλμού στην εξίσωση (6.5), που δεν παρουσιάζει κάποια συμμετρία ως προς τη μέση του διαστήματος  $[0, 2)$ , είναι αναμενόμενο να μην έχει μόνο όρους συγκεκριμένης συμμετρίας. Προσέξτε όμως ότι η συνάρτηση  $g(x) = f(x) - 1/2$  είναι αντισυμμετρική σε αυτό το διάστημα, γύρω από το  $x = 1$ . Το ανάπτυγμα Fourier της  $g(x)$  αναμένουμε να έχει μόνο τους όρους των ημιτόνων. Εύκολα επιβεβαιώνεται αυτό από το ανάπτυγμα της  $f(x)$ .

Γενικά, μια κατάλληλη μετατόπιση της συνάρτησης ή της ανεξάρτητης μεταβλητής κατά ένα σταθερό όρο μπορεί να αναδείξει τη συμμετρία της συνάρτησης, αν υπάρχει, και επομένως να απλοποιήσει τη σειρά Fourier.

## 6.3.4 Παράδειγμα

Ας δούμε άλλο παράδειγμα με εξαρχής αντισυμμετρική συνάρτηση. Έστω η συνάρτηση στο Σχήμα 6.3 (πριονωτός παλμός)



Σχήμα 6.3: Πριονωτός παλμός, εξίσωση (6.7)

$$f(x) = (x \bmod 1) - \frac{1}{2}. \quad (6.7)$$

Η έκφραση  $x \bmod 1$  σημαίνει ότι προσθέτουμε ή αφαιρούμε το 1 στο  $x$  όσες φορές χρειάζεται ώστε το αποτέλεσμα να είναι στο διάστημα  $[0, 1)$ .

Η συνάρτηση είναι περιοδική με περίοδο  $L = 1$ , συνεχής με πεπερασμένες ασυνέχειες στα σημεία  $x = m$  ( $m = 0, \pm 1, \pm 2, \dots$ ). Το ολοκλήρωμα της  $|f(x)|$  στο  $[0, 1)$  ορίζεται και είναι πεπερασμένο:

$$\int_0^1 |f(x)| \, dx = \frac{1}{4}.$$

Η  $f(x)$  ικανοποιεί τις συνθήκες Dirichlet οπότε μπορεί να αναπτυχθεί σε σειρά Fourier. Επιπλέον, είναι αντισυμμετρική ως προς το  $x = 1/2$  (το μέσο μιας περιόδου) οπότε η σειρά Fourier έχει μόνο τους όρους των ημιτόνων:

$$f(x) = \sum_{m=1}^{\infty} B_m \sin(2m\pi x).$$

Οι συντελεστές  $B_m$  υπολογίζονται από τον τύπο (6.4β):

$$B_m = 2 \int_0^1 \sin(2m\pi x) \left(x - \frac{1}{2}\right) dx = -\frac{1}{m\pi}, \quad m > 0. \quad (6.8)$$

Οι συντελεστές  $A_m$  είναι 0.

Σύμφωνα με τα παραπάνω, ο συγκεκριμένος πριονωτός παλμός αναπτύσσεται σε σειρά Fourier ως εξής

$$f(x) = - \sum_{m=1}^{\infty} \frac{\sin(2m\pi x)}{m\pi}. \quad (6.9)$$

Παρατηρήστε ότι όταν το  $x$  παίρνει ακέραιες τιμές, δηλαδή στα σημεία ασυνέχειας της  $f(x)$ , η σειρά Fourier είναι συνεχής και έχει τιμή 0, όσο το ημιάθροισμα του δεξιού και του αριστερού ορίου της  $f(x)$  σε αυτά τα σημεία.

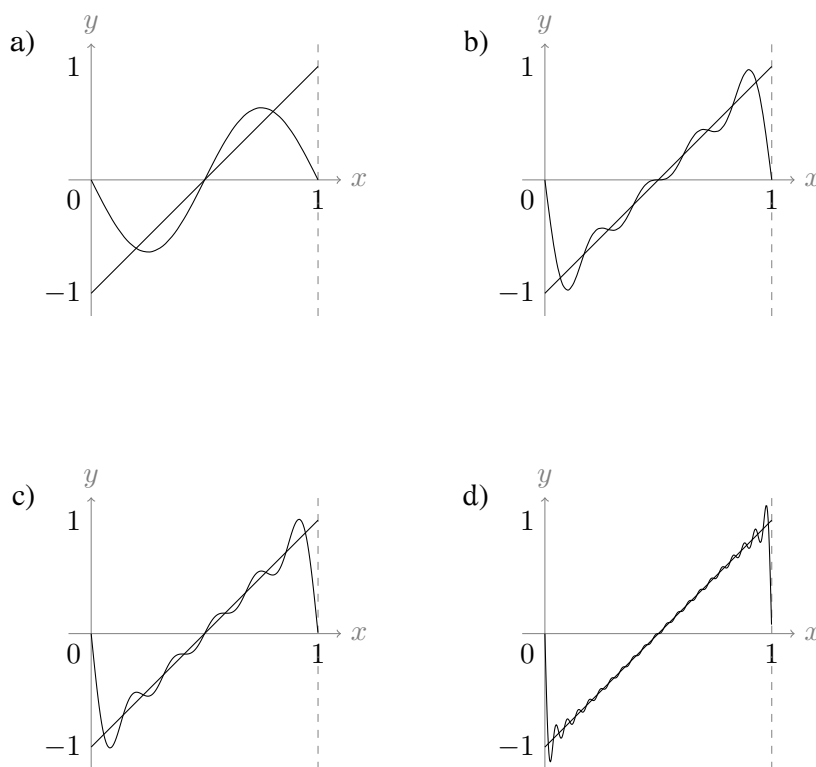


## 6.4 Φαινόμενο Gibbs

Στο Σχήμα 6.4 βλέπουμε πώς προσεγγίζει τον τριγωνικό παλμό της εξίσωσης (6.7) η σειρά Fourier της εξίσωσης (6.9) με  $M$  όρους, δηλαδή το μερικό άθροισμα

$$f_M(x) = - \sum_{m=1}^M \frac{\sin(2m\pi x)}{m\pi}, \quad (6.10)$$

για διάφορες τιμές του  $M$ . Παρατηρήστε ότι η σειρά Fourier παρουσιάζει ταλαντώσεις πλησιάζοντας σε σημείο ασυνέχειας της συνάρτησης, ανεξάρτητα από το πλήθος των όρων που θα λάβουμε υπόψη. Αυτή η συμπεριφορά είναι ανεξάρτητη από τη συνάρτηση που προσεγγίζεται και ονομάζεται φαινόμενο Gibbs. Λόγω των ταλαντώσεων η σειρά μπορεί να ξεπεράσει τη μέγιστη/ελάχιστη τιμή της συνάρτησης και να παρουσιάσει «αφύσικο» ακρότατο.



Σχήμα 6.4: Γραφική παράσταση της συνάρτησης  $f_M(x)$  (εξίσωση (6.10)) για a)  $M = 1$ , b)  $M = 4$ , c)  $M = 5$ , d)  $M = 20$

## 6.5 Παραγωγή σειράς Fourier από άλλη

### 6.5.1 Ολοκλήρωση

Ας υπολογίσουμε το αόριστο ολοκλήρωμα της σειράς Fourier μιας συνάρτησης  $f(x)$ :

$$\begin{aligned}\int f(x) dx + c &= \int \frac{A_0}{2} dx + \sum_{m=1}^{\infty} A_m \int \cos\left(\frac{2m\pi x}{L}\right) dx + \sum_{m=1}^{\infty} B_m \int \sin\left(\frac{2m\pi x}{L}\right) dx \\ &= \frac{A_0}{2} x + \sum_{m=1}^{\infty} A_m \frac{L}{2m\pi} \sin\left(\frac{2m\pi x}{L}\right) - \sum_{m=1}^{\infty} B_m \frac{L}{2m\pi} \cos\left(\frac{2m\pi x}{L}\right).\end{aligned}$$

Η ποσότητα  $c$  είναι η σταθερά ολοκλήρωσης. Παρατηρήστε ότι το δεξί μέλος περιέχει τον όρο  $A_0 x/2$  που, αν  $A_0 \neq 0$ , δεν είναι περιοδικός. Η σειρά στο δεξί μέλος δεν αποτελεί σειρά Fourier. Όμως

$$\int f(x) dx - \frac{A_0}{2} x + c = \sum_{m=1}^{\infty} A_m \frac{L}{2m\pi} \sin\left(\frac{2m\pi x}{L}\right) - \sum_{m=1}^{\infty} B_m \frac{L}{2m\pi} \cos\left(\frac{2m\pi x}{L}\right).$$

Από την παραπάνω σχέση προκύπτει η σειρά Fourier της

$$F(x) = \int \left( f(x) - \frac{A_0}{2} \right) dx$$

με προσέγγιση μιας προσθετικής σταθεράς  $c$ , η οποία μπορεί να προσδιοριστεί από τις τιμές της  $F(x)$  και της σειράς Fourier σε συγκεκριμένο σημείο.

### Παράδειγμα

Το αόριστο ολοκλήρωμα του τετραγωνικού παλμού, (6.5), είναι

$$\int f(x) dx = \begin{cases} 1, & 0 \leq x < 1, \\ x, & 1 \leq x < 2, \end{cases}$$

με επανάληψη έξω από το διάστημα  $[0, 2)$  ή γενικά

$$\int f(x) dx = \begin{cases} 1, & 0 \leq x - 2m < 1, \\ x - 2m, & 1 \leq x - 2m < 2, \end{cases}$$

για το διάστημα  $[2m, 2m + 2)$  με οποιοδήποτε ακέραιο  $m$ .

Το αόριστο ολοκλήρωμα της αντίστοιχης σειράς Fourier, (6.6), είναι

$$\int \left[ \frac{1}{2} - 2 \sum_{j=0}^{\infty} \frac{\sin((2j+1)\pi x)}{(2j+1)\pi} \right] dx = \frac{x}{2} + 2 \sum_{j=0}^{\infty} \frac{\cos((2j+1)\pi x)}{((2j+1)\pi)^2}.$$

Τα δύο ολοκληρώματα διαφέρουν κατά μια προσθετική σταθερά. Η τιμή του πρώτου στο  $x = 1/2$  είναι 1· η τιμή του δεύτερου είναι  $1/4$ . Άρα

$$\int f(x) dx - \frac{3}{4} = \frac{x}{2} + 2 \sum_{j=0}^{\infty} \frac{\cos((2j+1)\pi x)}{((2j+1)\pi)^2}.$$

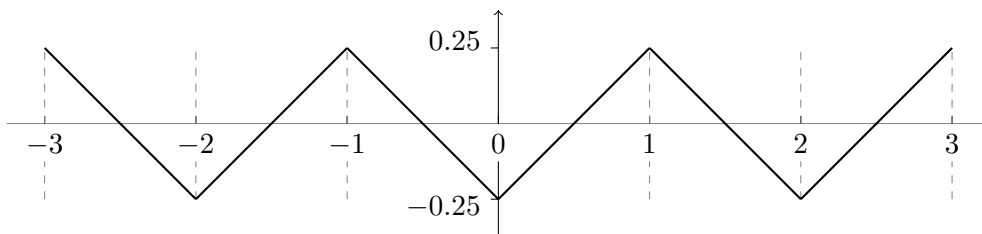
Συνεπώς, η σειρά Fourier της συνάρτησης

$$\int f(x) dx - \frac{3}{4} - \frac{x}{2} = \begin{cases} (1-2x)/4, & 0 \leq x < 1, \\ (2x-3)/4, & 1 \leq x < 2, \end{cases} \quad (6.11)$$

που επαναλαμβάνεται περιοδικά και δίνεται γραφικά στο Σχήμα 6.5, είναι

$$2 \sum_{j=0}^{\infty} \frac{\cos((2j+1)\pi x)}{((2j+1)\pi)^2}.$$

Μπορείτε να το επαληθεύσετε υπολογίζοντας τη σειρά Fourier της (6.11);



Σχήμα 6.5: Τριγωνικός παλμός, εξίσωση (6.11)

### 6.5.2 Παραγωγήιση

Αν έχουμε υπολογίσει τη σειρά Fourier, (6.3), για μια περιοδική συνάρτηση  $f(x)$  με περίοδο  $L$ , μπορούμε να την παραγωγίσουμε ως προς  $x$ , όρο-όρο:

$$- \sum_{m=1}^{\infty} A_m \frac{2m\pi}{L} \sin\left(\frac{2m\pi x}{L}\right) + \sum_{m=1}^{\infty} B_m \frac{2m\pi}{L} \cos\left(\frac{2m\pi x}{L}\right).$$

Η σειρά που προκύπτει, περιέχει στη γενική περίπτωση, άπειρους όρους ημιτόνων και συνημιτόνων κατάλληλης μορφής, αποτελεί τη σειρά Fourier της  $f'(x)$  και θα συγκλίνει σε αυτή, με την προϋπόθεση ότι η παράγωγος είναι περιοδική και ικανοποιεί τις συνθήκες Dirichlet. Θα δούμε ένα παράδειγμα σε επόμενη παράγραφο.

## 6.6 Εναλλακτική θεώρηση της σειράς Fourier

Το σύνολο των συναρτήσεων μίας πραγματικής μεταβλητής, που είναι περιοδικές με περίοδο  $L$  και ικανοποιούν τις συνθήκες Dirichlet, εφοδιασμένο με τις

γνωστές πράξεις της πρόσθεσης συναρτήσεων και του πολλαπλασιασμού αριθμού με συνάρτηση, αποτελεί ένα διανυσματικό χώρο. Σε αυτόν, μπορούμε να ορίσουμε το εσωτερικό γινόμενο δύο διανυσμάτων  $\langle \mathbf{f} | \mathbf{g} \rangle$  ως εξής:

$$\langle \mathbf{f} | \mathbf{g} \rangle \equiv \frac{2}{L} \int_0^L f(x)^* g(x) dx .$$

Μια βάση του χώρου αυτού, στην περίπτωση που οι συναρτήσεις είναι πραγματικές, είναι το σύνολο

$$\left\{ \frac{1}{\sqrt{2}}, \cos\left(\frac{2\pi x}{L}\right), \sin\left(\frac{2\pi x}{L}\right), \cos\left(\frac{2\pi 2x}{L}\right), \sin\left(\frac{2\pi 2x}{L}\right), \dots, \right. \\ \left. \cos\left(\frac{2\pi mx}{L}\right), \sin\left(\frac{2\pi mx}{L}\right), \dots \right\}$$

με άπειρα διανύσματα που είναι γραμμικώς ανεξάρτητα. Με τη βοήθεια των ολοκληρωμάτων στο Παράρτημα **α'** προκύπτει ότι η βάση είναι ορθοκανονική. Οι συντελεστές του αναπτύγματος οποιουδήποτε μέλους του διανυσματικού χώρου σε αυτή τη βάση δίνονται από τις σχέσεις

$$\begin{aligned} a_0 &= \langle \frac{1}{\sqrt{2}} | \mathbf{f} \rangle , \\ a_m &= \langle \cos\left(\frac{2m\pi x}{L}\right) | \mathbf{f} \rangle , \quad m > 0 , \\ b_m &= \langle \sin\left(\frac{2m\pi x}{L}\right) | \mathbf{f} \rangle , \quad m > 0 . \end{aligned}$$

Το ανάπτυγμα είναι

$$|\mathbf{f}\rangle = a_0 \frac{1}{\sqrt{2}} + \sum_{m=1}^{\infty} a_m \cos\left(\frac{2m\pi x}{L}\right) + \sum_{m=1}^{\infty} b_m \sin\left(\frac{2m\pi x}{L}\right) ,$$

δηλαδή ουσιαστικά η σειρά Fourier: οι συντελεστές της είναι  $A_0 = \sqrt{2}a_0$ ,  $A_m = a_m$ ,  $B_m = b_m$ ,  $m \geq 1$ .

### 6.6.1 Ταυτότητα Parseval

Το εσωτερικό γινόμενο δύο περιοδικών συναρτήσεων  $|\mathbf{f}^{(1)}\rangle$  και  $|\mathbf{f}^{(2)}\rangle$ , που είναι μέλη του χώρου, μπορεί να εκφραστεί με τους συντελεστές αυτών στην ορθοκανονική βάση,  $\{a_0^{(1)}, a_1^{(1)}, \dots, b_1^{(1)}, \dots\}$  και  $\{a_0^{(2)}, a_1^{(2)}, \dots, b_1^{(2)}, \dots\}$ :

$$\begin{aligned} \langle \mathbf{f}^{(1)} | \mathbf{f}^{(2)} \rangle &= a_0^{(1)} a_0^{(2)} + \sum_{m=1}^{\infty} (a_m^{(1)} a_m^{(2)} + b_m^{(1)} b_m^{(2)}) \\ &= \frac{1}{2} A_0^{(1)} A_0^{(2)} + \sum_{m=1}^{\infty} (A_m^{(1)} A_m^{(2)} + B_m^{(1)} B_m^{(2)}) . \end{aligned}$$

Στην τελευταία έκφραση εμφανίζονται οι συντελεστές Fourier των δύο συναρτήσεων.

Αν  $|f^{(1)}\rangle = |f^{(2)}\rangle = |f\rangle$  η τελευταία σχέση γίνεται

$$\frac{2}{L} \int_0^L f(x)^2 dx = \frac{A_0^2}{2} + \sum_{m=1}^{\infty} (A_m^2 + B_m^2) . \quad (6.12)$$

Η εξίσωση αυτή είναι η ταυτότητα Parseval.

Η μέση τιμή μιας συνάρτησης  $f(x)$  σε ένα διάστημα  $[a, b]$  συμβολίζεται με  $\langle f \rangle$  (ή με  $\bar{f}$ ) και ορίζεται από τη σχέση

$$\langle f \rangle \equiv \frac{\int_a^b f(x) dx}{\int_a^b dx} = \frac{1}{b-a} \int_a^b f(x) dx .$$

Σύμφωνα με την ταυτότητα Parseval, η μέση τιμή της συνάρτησης  $f(x)^2$  στο  $[0, L]$  μπορεί να υπολογιστεί από τους συντελεστές Fourier της  $f(x)$ :

$$\langle f^2 \rangle \equiv \frac{1}{L} \int_0^L f(x)^2 dx = \frac{A_0^2}{4} + \frac{1}{2} \sum_{m=1}^{\infty} (A_m^2 + B_m^2) .$$

### Παράδειγμα

Ας υπολογίσουμε την ταυτότητα Parseval για τη συνάρτηση  $f(x) = x - 1/2$  που ορίζεται στο διάστημα  $[0, 1]$  και επαναλαμβάνεται περιοδικά έξω από αυτό, Σχήμα 6.3, με περίοδο  $L = 1$ , δηλαδή του πριονωτού παλμού (εξίσωση (6.7)).

Το αριστερό μέλος της εξίσωσης (6.12) δίνει

$$\frac{2}{L} \int_0^L f(x)^2 dx = 2 \int_0^1 \left(x - \frac{1}{2}\right)^2 dx = \frac{1}{6} .$$

Βρήκαμε προηγουμένως τους συντελεστές Fourier για τη συγκεκριμένη συνάρτηση (εξίσωση (6.8)). Το άθροισμα στο δεξί μέλος της εξίσωσης (6.12) είναι

$$\frac{A_0^2}{2} + \sum_{m=1}^{\infty} (A_m^2 + B_m^2) = \sum_{m=1}^{\infty} \frac{1}{m^2 \pi^2} .$$

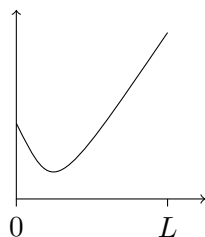
Η ταυτότητα Parseval γίνεται

$$\frac{1}{6} = \frac{1}{\pi^2} \sum_{m=1}^{\infty} \frac{1}{m^2} \Rightarrow \sum_{m=1}^{\infty} \frac{1}{m^2} = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \dots = \frac{\pi^2}{6} .$$

## 6.7 Σειρά Fourier για συναρτήσεις σε πεπερασμένο διάστημα

Η συνάρτηση  $f(x)$  δεν είναι απαραίτητο να είναι περιοδική για να αναπτυχθεί σε σειρά Fourier· μπορεί να ορίζεται και να ικανοποιεί τις συνθήκες Dirichlet σε ένα πεπερασμένο διάστημα μήκους  $L$  και να την επεκτείνουμε πέρα από αυτό. Κατόπιν, μπορούμε να υπολογίσουμε το ανάπτυγμα Fourier.

Ας εξετάσουμε τη συνάρτηση στο Σχήμα 6.6. Θα παρουσιάσουμε τις δυνατότητες

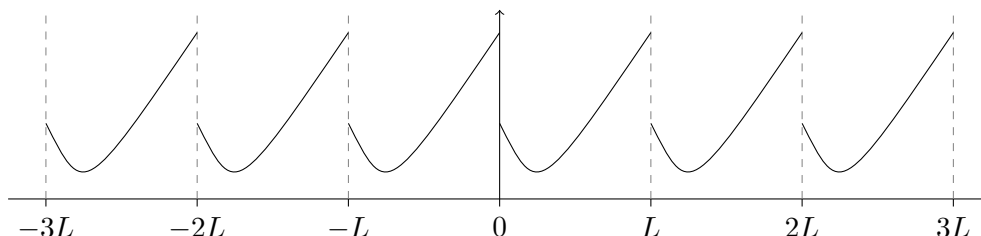


Σχήμα 6.6: Μη περιοδική συνάρτηση στο  $[0, L]$

που έχουμε για την επέκτασή της έξω από το πεδίο ορισμού της.

### 6.7.1 Μετατόπιση

Μπορούμε να την επαναλάβουμε αυτούσια στα διαστήματα  $[L, 2L)$ ,  $[2L, 3L)$ , κλπ. όπως και στα  $[-L, 0)$ ,  $[-2L, -L)$ , κλπ. όπως στο Σχήμα 6.7. Δημιουργούμε επο-

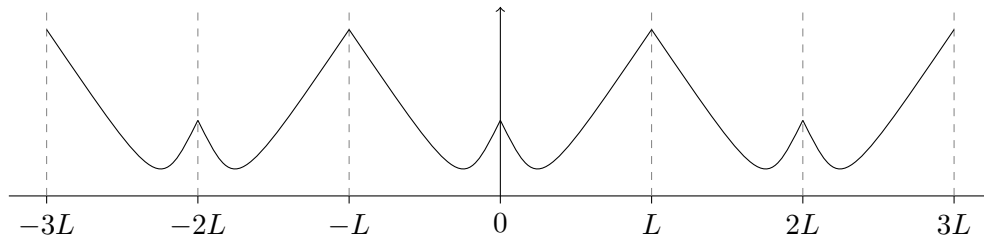


Σχήμα 6.7: Επέκταση μη περιοδικής συνάρτησης με επανάληψη

μένως μια νέα συνάρτηση που είναι περιοδική με περίοδο  $L$ . Προσέξτε ότι αν η μη περιοδική συνάρτηση έχει διαφορετικές τιμές στα άκρα του διαστήματος  $0, L$ , η επέκτασή της με αυτό τον τρόπο δημιουργεί σημεία πεπερασμένης ασυνέχειας (τα  $0, \pm L, \pm 2L \dots$ ). Η σειρά Fourier ορίζεται για τη νέα συνάρτηση και συγκλίνει στην αρχική μας στο πεδίο ορισμού της, το  $[0, L)$ , εκτός από τα σημεία ασυνέχειας σε αυτό ( $0$  και  $L$ ). Επομένως, αν η μη περιοδική συνάρτηση  $f(x)$  που ορίζεται στο  $[0, L)$  ικανοποιεί τη σχέση  $\lim_{x \rightarrow L} f(x) \neq f(0)$ , καλό είναι να αποφεύγουμε αυτό τον τρόπο επέκτασης.

### 6.7.2 Κατοπτρισμός ως προς ευθείες

Δεύτερος τρόπος επέκτασης μιας μη περιοδικής συνάρτησης  $f(x)$  είναι με κατοπτρισμό ως προς τις ευθείες  $x = 0$ ,  $x = \pm L$ ,  $x = \pm 2L$  κλπ., δηλαδή, στο διάστημα  $[-L, 0)$  έχουμε  $f(x) = f(-x)$ , στο διάστημα  $[L, 2L)$  έχουμε  $f(x) = f(2L - x)$ , κλπ., όπως στο Σχήμα 6.8: Με αυτή την επιλογή επέκτασης δημιουργούμε μια νέα συ-

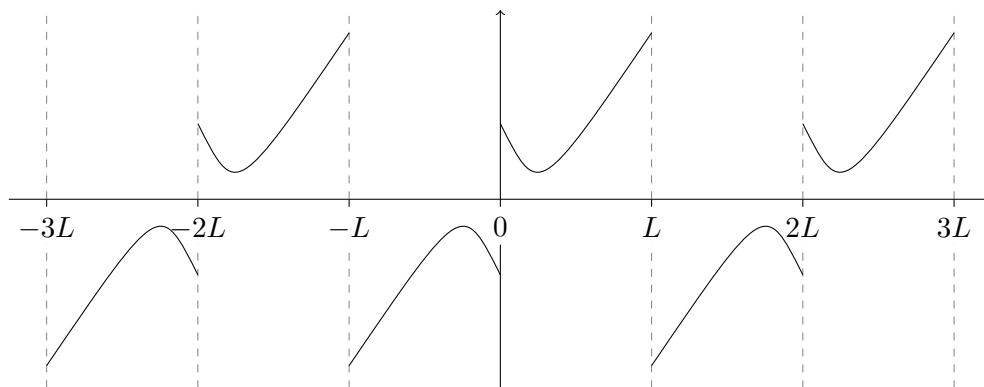


Σχήμα 6.8: Επέκταση μη περιοδικής συνάρτησης με κατοπτρισμό

νάρτηση που είναι περιοδική με περίοδο  $2L$  και δεν εισάγουμε σημεία ασυνέχειας. Επιπλέον, η νέα συνάρτηση είναι συμμετρική, οπότε η σειρά Fourier για αυτή δεν θα περιλαμβάνει όρους με ημίτονα. Το ανάπτυγμα Fourier που θα προκύψει, συγκλίνει στην αρχική μας συνάρτηση στο διάστημα  $[0, L)$ .

### 6.7.3 Κατοπτρισμός ως προς σημεία

Τρίτος τρόπος επέκτασης μιας μη περιοδικής συνάρτησης  $f(x)$  είναι με κατοπτρισμό ως προς τα σημεία  $(0, 0)$ ,  $(\pm L, 0)$ ,  $(\pm 2L, 0)$ , κλπ. Αυτό σημαίνει ότι στο διάστημα  $[-L, 0)$  θέτουμε  $f(x) = -f(-x)$  και σε οποιοδήποτε άλλο σημείο έξω από το  $[-L, L)$  θέτουμε  $f(x + 2L) = f(x)$ . Με αυτή την επιλογή επέκτασης δημιουργούμε μια νέα συνάρτηση που είναι περιοδική με περίοδο  $2L$  (Σχήμα 6.9) και αντισυμμετρική. Παρατηρήστε ότι αν η  $f(x)$  στα άκρα του διαστήματος ορισμού



Σχήμα 6.9: Επέκταση μη περιοδικής συνάρτησης με ανάκλαση ως προς σημεία

της δεν έχει τιμή (ή όριο) το 0, ο συγκεκριμένος τρόπος επέκτασής της δημιουργεί

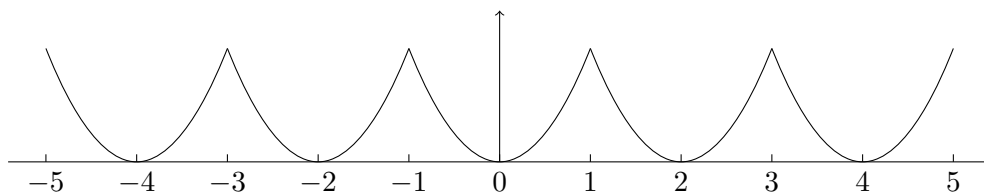
σημεία πεπερασμένης ασυνέχειας. Η σειρά Fourier της περιοδικής επέκτασης θα περιέχει μόνο όρους με ημίτονα και θα συγκλίνει στην  $f(x)$  στο πεδίο ορισμού της εκτός από τα σημεία ασυνέχειας, είτε αυτά που προκαλέσαμε με την επέκταση είτε αυτά που έχει εγγενώς η συνάρτηση.

#### 6.7.4 Παράδειγμα

Ας υπολογίσουμε τη σειρά Fourier της συνάρτησης  $f(x) = x^2$  που ορίζεται στο διάστημα  $[0, 1)$ .

**Κατοπτρισμός ως προς την ευθεία  $y = 0$**

Επιλέγουμε καταρχάς να επεκτείνουμε τη συνάρτηση στο διάστημα  $[-1, 0)$  συμμετρικά ώστε να μην εισαχθούν σημεία ασυνέχειας. Εκεί  $f(x) = f(-x) = x^2$ . Κατόπιν, επαναλαμβάνουμε τη συνάρτηση  $g(x) = x^2$  με  $-1 \leq x < 1$  έξω από το διάστημα  $[-1, 1)$ , ώστε να κατασκευάσουμε συνάρτηση περιοδική με περίοδο 2 (Σχήμα 6.10).



Σχήμα 6.10: Επέκταση της  $f(x) = x^2$  συμμετρικά

Η νέα συνάρτηση,  $g(x)$ , είναι συμμετρική σε διάστημα μιας περιόδου,  $[-1, 1)$ , ως προς το μέσο του, οπότε οι συντελεστές  $B_m$  στην εξίσωση (6.4β') είναι 0. Για τους  $A_m$  έχουμε

$$A_m = \int_{-1}^1 \cos(m\pi x) x^2 dx, \quad m \geq 0.$$

Το διάστημα ολοκλήρωσης επελέγη να είναι το  $[-1, 1)$ . Όπως αναφέραμε, αρκεί να έχει μήκος μία περίοδο. Ο υπολογισμός των  $A_m$  δίνει μετά από πράξεις,

$$A_0 = \frac{2}{3}, \quad A_m = (-1)^m \frac{4}{m^2 \pi^2}, \quad m > 0.$$

Επομένως,

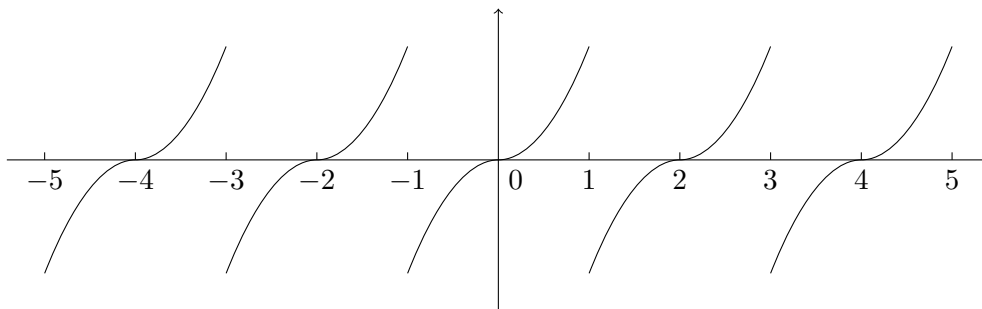
$$f(x) = \frac{1}{3} + 4 \sum_{m=1}^{\infty} \frac{(-1)^m \cos(m\pi x)}{m^2 \pi^2}, \quad 0 \leq x < 1.$$

Ποιες τιμές έχει η σειρά Fourier στα  $x = 0$  και  $x = 1$ ;



### Κατοπτρισμός ως προς την αρχή

Ας επεκτείνουμε τώρα τη συνάρτηση στο διάστημα  $[-1, 0)$  αντισυμμετρικά με κατοπτρισμό ως προς το σημείο  $(0, 0)$ . Στο  $[-1, 0]$  θεωρούμε ότι  $f(x) = -f(-x) = -x^2$ . Κατόπιν, επαναλαμβάνουμε τη νέα συνάρτηση έξω από το διάστημα  $[-1, 1)$ , ώστε να κατασκευάσουμε συνάρτηση περιοδική με περίοδο 2 (Σχήμα 6.11). Καθώς  $f(0) = 0$  το  $x = 0$  και το  $\pm 2, \pm 4, \dots$  δεν είναι σημεία ασυνέχειας. Τα  $x = \pm 1, \pm 3 \dots$  είναι.



Σχήμα 6.11: Επέκταση της  $f(x) = x^2$  αντισυμμετρικά

Η νέα συνάρτηση είναι αντισυμμετρική σε διάστημα μιας περιόδου,  $[-1, 1)$ , ως προς το μέσο του, οπότε οι συντελεστές  $B_m$  είναι 0. Για τους  $B_m$  έχουμε

$$B_m = 2 \int_0^1 \sin(m\pi x) x^2 dx, \quad m > 0.$$

Μετά από πράξεις,

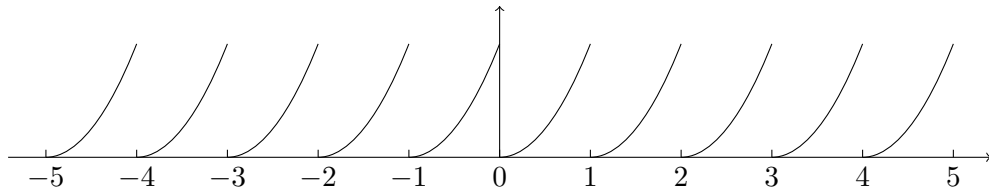
$$\begin{aligned} B_{2k} &= -\frac{1}{k\pi} \quad k > 0, \\ B_{2k+1} &= \frac{1}{(k + 1/2)\pi} - \frac{1}{(k + 1/2)^3 \pi^3} \quad k > 0. \end{aligned}$$

Επομένως,

$$f(x) = \sum_{k=1}^{\infty} B_{2k} \sin(2k\pi x) + \sum_{k=1}^{\infty} B_{2k+1} \sin((2k+1)\pi x), \quad 0 \leq x < 1.$$

### Μετατόπιση

Ας επιλέξουμε να επαναλάβουμε την  $f(x) = x^2$  αυτούσια έξω από το διάστημα  $[0, 1)$  ώστε να δημιουργήσουμε περιοδική συνάρτηση με περίοδο 1. Η συγκεκριμένη επιλογή εισάγει τα σημεία ασυνέχειας  $x = j$  με  $j$  οποιοδήποτε ακέραιο. Σε αυτά, η σειρά Fourier δεν θα συγκλίνει στην περιοδική συνάρτηση.



Σχήμα 6.12: Επέκταση της  $f(x) = x^2$  με επανάληψη

Ο υπολογισμός των συντελεστών  $A_m$  από την εξίσωση (6.4α') δίνει

$$A_0 = \frac{2}{3}, \quad A_m = \frac{1}{m^2\pi^2}, \quad m > 0.$$

Οι συντελεστές  $B_m$  της σχέσης (6.4β') είναι

$$B_m = -\frac{1}{m\pi}, \quad m > 0.$$

Επομένως, η σειρά Fourier είναι

$$\frac{1}{3} + \sum_{m=1}^{\infty} \frac{\cos(2m\pi x)}{m^2\pi^2} - \sum_{m=1}^{\infty} \frac{\sin(2m\pi x)}{m\pi}.$$

και συγκλίνει στην  $f(x) = x^2$  στο διάστημα  $(0, 1)$  (παραλείπουμε το σημείο  $x = 0$  που είναι σημείο ασυνέχειας).

Παρατηρήστε ότι αν παραγωγίσουμε την  $f(x)$  και τη σειρά Fourier έχουμε

$$2x = -2 \sum_{m=1}^{\infty} \frac{\sin(2m\pi x)}{m\pi} - 2 \sum_{m=1}^{\infty} \cos(2m\pi x).$$

Επομένως,

$$x - \frac{1}{2} = - \sum_{m=1}^{\infty} \frac{\sin(2m\pi x)}{m\pi} - \sum_{m=1}^{\infty} \cos(2m\pi x) - \frac{1}{2}.$$

Συγκρίνετε την τελευταία σχέση με τη σειρά Fourier του πριονωτού παλμού, (6.9). Ταυτίζονται αν για οποιοδήποτε σημείο του  $(0, 1)$  ισχύει

$$\begin{aligned} \sum_{m=1}^{\infty} \cos(2m\pi x) &= -\frac{1}{2} \Leftrightarrow 2 \sum_{m=1}^{\infty} \cos(2m\pi x) + 1 = 0 \\ &\Leftrightarrow \sum_{m=1}^{\infty} \cos(2m\pi x) + \sum_{m=-1}^{-\infty} \cos(-2m\pi x) + \cos(2\pi x \cdot 0) = 0 \\ &\Leftrightarrow \sum_{m=0}^{\infty} \cos(2m\pi x) + \sum_{m=-1}^{-\infty} \cos(2m\pi x) = 0 \\ &\Leftrightarrow \sum_{m=-\infty}^{\infty} \cos(2m\pi x) = 0. \end{aligned}$$

Πράγματι ισχύει, αλλά δεν είναι του παρόντος η σχετική απόδειξη<sup>23</sup>

**Παρατήρηση:** Όπως είδαμε στα παραδείγματα, η ίδια μη περιοδική συνάρτηση, στο ίδιο διάστημα, μπορεί να έχει σειρές Fourier με διαφορετική μορφή, ανάλογα με τον τρόπο επέκτασής της. Βέβαια, η τιμή των διάφορων σειρών στο ίδιο σημείο είναι η ίδια.

## 6.8 Μιγαδική μορφή της σειράς Fourier

Η σειρά Fourier μπορεί να γραφεί σε πιο συνοπτική μορφή αν θυμηθούμε ότι

$$e^{i\theta} = \cos \theta + i \sin \theta .$$

Εύκολα προκύπτει ότι

$$\cos \theta = \frac{e^{i\theta} + e^{-i\theta}}{2} , \quad \sin \theta = \frac{e^{i\theta} - e^{-i\theta}}{2i} .$$

Η αντικατάσταση των παραπάνω σχέσεων στην (6.3) δίνει τη σειρά Fourier στην εκθετική μορφή

$$f(x) = \sum_{m=-\infty}^{\infty} C_m \exp \left( i \frac{2m\pi x}{L} \right) , \quad (6.13)$$

όπου οι μιγαδικοί, πλέον, συντελεστές  $C_m$  συνδέονται με τους πραγματικούς  $A_m$ ,  $B_m$  με τις σχέσεις

$$C_0 = \frac{A_0}{2} , \quad (6.14\alpha')$$

$$C_m = \frac{A_m - iB_m}{2} , \quad m > 0 , \quad (6.14\beta')$$

$$C_{-m} = \frac{A_m + iB_m}{2} = C_m^* , \quad m > 0 . \quad (6.14\gamma')$$

<sup>2</sup>σε οποιοδήποτε σημείο  $x$  το τελευταίο άθροισμα είναι  $2\pi\delta(x)$ . Είναι 0 για  $x \neq 0$ . Η συνάρτηση  $\delta(x)$  παρουσιάζεται στο §6.11.

<sup>3</sup> Προσέξτε ότι

$$\sum_{m=1}^{\infty} \cos(2m\pi x) = -\frac{1}{2} \Leftrightarrow 1 + \sum_{m=1}^{\infty} \cos(2m\pi x) = 1 - \frac{1}{2} \Leftrightarrow \sum_{m=0}^{\infty} \cos(2m\pi x) = \frac{1}{2} .$$

Όταν  $x = 1/2$ , η τελευταία έκφραση γίνεται

$$\sum_{m=0}^{\infty} \cos(m\pi) = \frac{1}{2} \Rightarrow \sum_{m=0}^{\infty} (-1)^m = \frac{1}{2} \Rightarrow 1 - 1 + 1 - 1 + 1 - 1 + \dots = \frac{1}{2} .$$

Η σειρά στο αριστερό μέλος είναι γνωστή ως σειρά Grandi, και η τιμή της, με συγκεκριμένο ορισμό για την άθροιση, είναι όντως  $1/2$ , όσο παράδοξο και αν φαίνεται.

Πολύ εύκολα προκύπτει και η αντίστροφη σχέση που προσδιορίζει τα  $A_m$ ,  $B_m$  όταν είναι γνωστά τα  $C_m$ :

$$A_m = C_m + C_{-m}, \quad m \geq 0, \quad (6.15\alpha')$$

$$B_m = i(C_m - C_{-m}), \quad m > 0. \quad (6.15\beta')$$

Συνδυάζοντας τις (6.4α'), (6.4β'), (6.14) προκύπτει ότι

$$C_m = \frac{1}{L} \int_0^L \exp\left(-i \frac{2m\pi x}{L}\right) f(x) dx, \quad m = 0, \pm 1, \pm 2, \dots \quad (6.16)$$

Με τους μιγαδικούς συντελεστές Fourier η ταυτότητα Parseval (εξίσωση (6.12)) γίνεται

$$\begin{aligned} \langle f|f \rangle &\equiv \frac{2}{L} \int_0^L |f(x)|^2 dx = 2C_0^2 + 4 \sum_{m=1}^{\infty} (C_m^* C_m) \Rightarrow \\ \frac{1}{L} \int_0^L |f(x)|^2 dx &= |C_0|^2 + 2 \sum_{m=1}^{\infty} |C_m|^2. \end{aligned}$$

Καθώς  $|C_{-m}| = |C_m|$ , η προηγούμενη σχέση καταλήγει στην ταυτότητα Parseval με τους μιγαδικούς συντελεστές

$$\frac{1}{L} \int_0^L |f(x)|^2 dx = \sum_{m=-\infty}^{\infty} |C_m|^2. \quad (6.17)$$

Άμεση συνέπεια της σχέσης αυτής είναι η έκφραση για τη μέση τιμή της ποσότητας  $|f(x)|^2$ :

$$\langle |f(x)|^2 \rangle = \sum_{m=-\infty}^{\infty} |C_m|^2.$$

### Παράδειγμα

Η μιγαδική μορφή της σειράς Fourier για τον πριονωτό παλμό, (6.7), στο διάστημα  $[0, 1)$ , έχει συντελεστές

$$\begin{aligned} C_0 &= 0, \\ C_m &= \int_0^1 e^{-i2m\pi x} \left(x - \frac{1}{2}\right) dx = \frac{i}{2m\pi}, \quad m = \pm 1, \pm 2, \dots \end{aligned}$$

Μπορείτε εύκολα να επαληθεύσετε ότι  $C_{-m} = C_m^*$  και ότι  $C_m \equiv (A_m - iB_m)/2$  για τους συντελεστές στην (6.8).

Επομένως, η μιγαδική σειρά Fourier που προκύπτει από την (6.13) είναι

$$- \sum_{\substack{m=-\infty \\ m \neq 0}}^{\infty} \frac{e^{i2m\pi x}}{i2m\pi} .$$

Η ταυτότητα Parseval δίνει

$$\int_0^1 f(x)^2 dx = \sum_{m=-\infty}^{\infty} |C_m|^2 = \sum_{\substack{m=-\infty \\ m \neq 0}}^{\infty} \frac{1}{4m^2\pi^2} = \frac{1}{2\pi^2} \sum_{m=1}^{\infty} \frac{1}{m^2} .$$

## 6.9 Διακριτός μετασχηματισμός Fourier (DFT)

Οι συντελεστές  $A_m$ ,  $B_m$  (ή  $C_m$ ) της σειράς Fourier μπορούν να υπολογιστούν από τα αντίστοιχα ολοκληρώματα, είτε ακριβώς (με αναλυτικό υπολογισμό) είτε προσεγγιστικά με τις μεθόδους που είδαμε στο Κεφάλαιο 5. Στην περίπτωση που από τη συνάρτηση  $f(x)$  έχουμε μόνο κάποιες τιμές της σε συγκεκριμένα σημεία μπορούμε να έχουμε μόνο προσεγγιστικό υπολογισμό των συντελεστών.

Ας εφαρμόσουμε τον εκτεταμένο τύπο του τραπεζίου ((5.6)) για να υπολογίσουμε το ολοκλήρωμα στην (6.16): Χωρίζουμε το διάστημα ολοκλήρωσης  $[0, L]$  σε  $N$  ίσα διαστήματα μήκους  $h = L/N$  το καθένα. Τα  $N+1$  σημεία στα οποία θα υπολογίσουμε την ολοκληρωτέα ποσότητα είναι τα  $x_j = jh$ ,  $j = 0, 1, \dots, N$ . Παρατηρήστε ότι η ολοκληρωτέα ποσότητα στα άκρα,  $x_0 = 0$  και  $x_N = L$ , έχει την ίδια τιμή:

$$\exp\left(-i\frac{2m\pi 0}{L}\right) f(0) = \exp\left(-i\frac{2m\pi L}{L}\right) f(L) .$$

Η ισότητα των  $f(0)$  και  $f(L)$  προκύπτει από την περιοδικότητα της συνάρτησης (εξίσωση (6.2)).

Επομένως, η διακριτοποίηση της ολοκληρωτέας ποσότητας δίνει τη σχέση

$$C_m \approx \frac{h}{L} \sum_{j=0}^{N-1} \exp\left(-i\frac{2m\pi jh}{L}\right) f(jh) = \frac{1}{N} \sum_{j=0}^{N-1} \exp\left(-i\frac{2m\pi j}{N}\right) f_j$$

για τους συντελεστές της σειράς Fourier, όπου  $f_j \equiv f(jh)$ . Στην ίδια έκφραση καταλήγουμε και στην περίπτωση που οι τιμές της συνάρτησης είναι γνωστές μόνο σε  $N$  ισαπέχοντα σημεία (π.χ. από πειραματικές μετρήσεις).

Η σχέση

$$\bar{C}_m = \frac{1}{N} \sum_{j=0}^{N-1} \exp\left(-i\frac{2m\pi j}{N}\right) f_j , \quad m = 0, 1, \dots, N-1 , \quad (6.18)$$

αποτελεί το διακριτό μετασχηματισμό Fourier (DFT) της διακριτοποιημένης συνάρτησης  $f(x)$ . Οι συντελεστές  $\bar{C}_m$  που ορίζονται από αυτή τη σχέση προσεγγίζουν τους συντελεστές  $C_m$  στη σειρά Fourier.

Παρατηρήστε ότι η διακριτοποίηση διατηρεί μόνο  $N$  συντελεστές  $\bar{C}_m$  καθώς ισχύει η σχέση

$$\bar{C}_{m+N} \equiv \frac{1}{N} \sum_{j=0}^{N-1} \exp\left(-i\frac{2(m+N)\pi j}{N}\right) f_j = \frac{1}{N} \sum_{j=0}^{N-1} \exp\left(-i\frac{2m\pi j}{N}\right) f_j \equiv \bar{C}_m . \quad (6.19)$$

Ο αντίστροφος διακριτός μετασχηματισμός Fourier ορίζεται ως

$$\bar{f}_j = \sum_{m=0}^{N-1} \exp\left(i\frac{2j\pi m}{N}\right) \bar{C}_m , \quad j = 0, 1, \dots, N-1 . \quad (6.20)$$

και προσεγγίζει τις τιμές  $f_j$  της συνάρτησης.

**Σημείωση:** Ο παράγοντας  $1/N$  που πολλαπλασιάζει το άθροισμα στην (6.18) είναι θέμα σύμβασης. Το γινόμενο των συντελεστών πριν τα αθροίσματα στις εξισώσεις (6.18) και (6.20) πρέπει να είναι  $1/N$ , οι ακριβείς τιμές τους είναι απροσδιόριστες. Για λόγους συμμετρίας των σχέσεων (6.18, 6.20), οι μετασχηματισμοί μπορούν να οριστούν με ένα παράγοντα  $1/\sqrt{N}$  που πολλαπλασιάζει το άθροισμα του καθενός.

Η μέση τιμή της διακριτοποιημένης συνάρτησης  $|f(x)|^2$  είναι προφανώς

$$\langle f^2 \rangle = \frac{1}{N} \sum_{j=0}^{N-1} |f_j|^2 .$$

Η ταυτότητα Parseval γίνεται

$$\frac{1}{N} \sum_{j=0}^{N-1} |f_j|^2 = \sum_{m=0}^{N-1} |\bar{C}_m|^2 . \quad (6.21)$$

Μπορεί να αποδειχθεί αν αντικαταστήσουμε στο δεξί της μέλος την έκφραση για τα  $\bar{C}_m$  από τη σχέση (6.18).

### 6.9.1 Γρήγορος υπολογισμός του DFT – Αλγόριθμος FFT

Υπάρχουν διάφοροι αλγόριθμοι που μπορούν να υπολογίσουν ταυτόχρονα όλους τους συντελεστές Fourier, ιδιαίτερα γρήγορα, εκμεταλλευόμενοι τις συμμετρίες που εμφανίζονται, χωρίς να χρειάζεται να υπολογίσουν κάθε ολοκλήρωμα ξεχωριστά. Παρακάτω θα δούμε τον πιο βασικό.

Ας υποθέσουμε ότι το πλήθος  $N$  των όρων στο άθροισμα της (6.18) είναι δύναμη του 2. Τότε, ο υπολογισμός του μπορεί να γίνει χωρίζοντάς το σε αθροίσματα των όρων με άρτιο και περιττό δείκτη  $j$ :

$$\begin{aligned} & \sum_{j=0}^{N-1} \exp\left(-i\frac{2m\pi j}{N}\right) f_j \\ &= \sum_{r=0}^{N/2-1} \exp\left(-i\frac{2m\pi 2r}{N}\right) f_{2r} + \sum_{r=0}^{N/2-1} \exp\left(-i\frac{2m\pi(2r+1)}{N}\right) f_{2r+1} \\ &= \sum_{r=0}^{N/2-1} \exp\left(-i\frac{2m\pi r}{N/2}\right) f_{2r} + \exp\left(-i\frac{2m\pi}{N}\right) \sum_{r=0}^{N/2-1} \exp\left(-i\frac{2m\pi r}{N/2}\right) f_{2r+1} . \end{aligned}$$

Παρατηρήστε ότι οι όροι

$$\sum_{r=0}^{N/2-1} \exp\left(-i\frac{2m\pi r}{N/2}\right) f_{2r}$$

και

$$\sum_{r=0}^{N/2-1} \exp\left(-i\frac{2m\pi r}{N/2}\right) f_{2r+1}$$

είναι ουσιαστικά οι διακριτοί μετασχηματισμοί Fourier για δύο σύνολα τιμών της διακριτοποιημένης  $f(x)$ : το ένα αποτελείται από τα σημεία  $f_j$  με άρτιο δείκτη και το άλλο από τα σημεία με περιττό δείκτη. Το πλήθος των σημείων σε κάθε σύνολο είναι  $N/2$ .

Ας συμβολίσουμε με  $\bar{C}_m^e$ ,  $\bar{C}_m^o$  τους συντελεστές στους δύο μετασχηματισμούς Fourier, τον «άρτιο» και τον «περιττό» αντίστοιχα. Η προηγούμενη σχέση δίνει

$$\bar{C}_m = \frac{1}{N} \left( \frac{N}{2} \bar{C}_m^e + \frac{N}{2} \exp\left(-i\frac{2m\pi}{N}\right) \bar{C}_m^o \right) = \frac{1}{2} \left( \bar{C}_m^e + e^{-i2m\pi/N} \bar{C}_m^o \right) , \quad (6.22)$$

για  $m = 0, 1, \dots, N-1$ .

Παρατηρήστε ότι, λόγω της (6.19), έχουμε  $\bar{C}_{m+N/2}^{e,o} = \bar{C}_m^{e,o}$ . Επίσης ισχύει ότι

$$\exp\left(-i\frac{2(m+N/2)\pi}{N}\right) = -\exp\left(-i\frac{2m\pi}{N}\right) .$$

Επομένως, η σχέση (6.22) μπορεί να ξαναγραφεί ως εξής

$$\bar{C}_m = \frac{1}{2} \left( \bar{C}_m^e + e^{-i2m\pi/N} \bar{C}_m^o \right) , \quad (6.23\alpha')$$

$$\bar{C}_{m+N/2} = \frac{1}{2} \left( \bar{C}_m^e - e^{-i2m\pi/N} \bar{C}_m^o \right) , \quad (6.23\beta')$$

για  $m = 0, 1, \dots, N/2-1$ .

Η εξίσωση (6.22) (ή, ισοδύναμα, η εξίσωση (6.23)) εκφράζει ότι ο υπολογισμός του DFT  $N$  σημείων απαιτεί τον υπολογισμό δύο DFT των  $N/2$  σημείων ο καθένας. Η συγκεκριμένη ανάλυση μπορεί να χρησιμοποιηθεί για τον υπολογισμό των νέων DFT αναπτύσσοντάς τους σε τέσσερις συνολικά DFT των  $N/4$  σημείων ο καθένας. Η διαδικασία αυτή επαναλαμβάνεται έως ότου καταλήξουμε σε  $N$  DFT του ενός σημείου ο καθένας. Ο υπολογισμός του DFT ενός σημείου είναι πολύ εύκολος: από τη (6.18) προκύπτει ότι ο (μοναδικός) συντελεστής της σειράς Fourier είναι ίσος με την τιμή της συνάρτησης στο σημείο.

Η επαναληπτική διαδικασία που περιγράψαμε είναι η βάση των αλγορίθμων Fast Fourier Transform (FFT). Σε αυτή, ο συντελεστής  $\bar{C}_m$  απαιτεί για τον υπολογισμό του συνολικά  $2 \log_2 N$  μιγαδικούς πολλαπλασιασμούς. Επομένως, οι  $N$  συντελεστές χρειάζονται  $2N \log_2 N$  πράξεις για τον υπολογισμό τους.

Αν επιλέγαμε να υπολογίσουμε το άθροισμα στην (6.18) απευθείας, χρειαζόμαστε  $N$  πολλαπλασιασμούς για τον κάθε συντελεστή· συνολικά, δηλαδή,  $N^2$  πράξεις. Το κέρδος σε ταχύτητα είναι σημαντικό: αν π.χ. έχουμε  $N = 1024$  ο αλγόριθμος FFT χρειάζεται 20480 πράξεις ενώ χωρίς αυτόν θα κάναμε 1048576 πράξεις.

## 6.10 Μετασχηματισμός Fourier

Έχουμε παρουσιάσει μέχρι τώρα την προσέγγιση με σειρά Fourier για περιοδικές συναρτήσεις και για συναρτήσεις που ορίζονται σε πεπερασμένο διάστημα, είναι μη περιοδικές, αλλά επεκτείνονται περιοδικά εκτός αυτού. Θα αναφερθούμε τώρα στην περίπτωση των μη περιοδικών συναρτήσεων που ορίζονται σε άπειρο διάστημα και στο πώς αναλύονται σε επαλληλία περιοδικών κυμάτων. Για τις συγκεκριμένες συναρτήσεις απαιτούμε μόνο να ορίζεται και να είναι πεπερασμένο το ολοκλήρωμα της απόλυτης τιμής τους στο πεδίο ορισμού τους.

Μια μη περιοδική συνάρτηση σε άπειρο διάστημα μπορεί να αντιμετωπιστεί ως περιοδική με περίοδο  $L$  που τείνει στο άπειρο. Ας γράψουμε τη μιγαδική σειρά Fourier για κάποιο  $L$ , (6.13), με αντικατάσταση των συντελεστών από την εξίσωση (6.16) και ας βρούμε το όριο της όταν  $L \rightarrow \infty$ :

$$f(x) = \lim_{L \rightarrow \infty} \sum_{m=-\infty}^{\infty} \left[ \frac{1}{L} \int_{-L/2}^{L/2} \exp\left(-i \frac{2m\pi y}{L}\right) f(y) dy \right] \exp\left(i \frac{2m\pi x}{L}\right).$$

Για διευκόλυνση, θέτουμε  $k_m = m2\pi/L$ . Η απόσταση διαδοχικών σημείων  $k_{m+1} - k_m$  είναι  $\Delta k = 2\pi/L$ . Αναζητούμε το όριο της έκφρασης όταν  $\Delta k \rightarrow 0$ :

$$\begin{aligned} f(x) &= \lim_{\Delta k \rightarrow 0} \sum_{m=-\infty}^{\infty} \frac{\Delta k}{2\pi} \left[ \int_{-\pi/\Delta k}^{\pi/\Delta k} \exp(-iyk_m) f(y) dy \right] \exp(ixk_m) \\ &= \frac{1}{2\pi} \lim_{\Delta k \rightarrow 0} \sum_{m=-\infty}^{\infty} \left[ \exp(ixk_m) \int_{-\pi/\Delta k}^{\pi/\Delta k} \exp(-iyk_m) f(y) dy \right] \Delta k. \end{aligned}$$



Θυμηθείτε τον ορισμό του ολοκληρώματος:

$$\lim_{\Delta k \rightarrow 0} \sum_{m=-\infty}^{\infty} g(k_m) \Delta k = \int_{-\infty}^{\infty} g(k) dk .$$

Επομένως,

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{\infty} e^{-iky} f(y) dy \right] e^{ikx} dk . \quad (6.24)$$

Η ποσότητα

$$\hat{f}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-iky} f(y) dy \quad (6.25)$$

αποτελεί το μετασχηματισμό Fourier της συνάρτησης  $f(x)$ . Είναι συνεχής ποσότητα, αντίστοιχη των συντελεστών Fourier στην περίπτωση περιοδικής συνάρτησης. Εναλλακτικός συμβολισμός της  $\hat{f}(k)$  θα είναι ο  $\mathcal{F}[f(x)]$ .

Η σχέση

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(k) e^{ikx} dk \quad (6.26)$$

αποτελεί τον αντίστροφο μετασχηματισμό Fourier. Είναι αντίστοιχη της σειράς Fourier στην περίπτωση περιοδικής συνάρτησης.

Παρατηρήστε ότι η επιλογή των συντελεστών πριν το ολοκλήρωμα σε καθεμιά από τις εξισώσεις (6.25, 6.26) είναι σχετικά ελεύθερη. Η μόνη υποχρέωση είναι το γινόμενο των συντελεστών να έχει τιμή  $\frac{1}{2\pi}$ . Επιλέξαμε να γίνει συμμετρικός διαχωρισμός και γι' αυτό οι συντελεστές είναι  $\frac{1}{\sqrt{2\pi}}$  σε κάθε εξίσωση.

Όπως θα δούμε στα παραδείγματα του κεφαλαίου, όσο πιο «εντοπισμένη» (περιορισμένη στα  $x$ ) είναι μια συνάρτηση, τόσο πιο «εκτεταμένος» είναι ο μετασχηματισμός Fourier της.

### Παράδειγμα

Έστω η συνάρτηση

$$f(x) = \begin{cases} 0, & x < 0 \\ e^{-\lambda x}, & x \geq 0 \end{cases} \quad (\lambda > 0) .$$

Καταρχάς, το ολοκλήρωμα της απόλυτης τιμής της  $f(x)$  στο πεδίο ορισμού της υπάρχει και είναι πεπερασμένο:

$$\int_{-\infty}^{\infty} |f(x)| dx = \int_0^{\infty} e^{-\lambda x} dx = \frac{1}{\lambda} .$$

Μπορούμε επομένως να υπολογίσουμε το μετασχηματισμό Fourier της: από τον ορισμό (6.25) έχουμε

$$\hat{f}(k) = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-iky} e^{-\lambda y} dy = \frac{1}{\sqrt{2\pi}(\lambda + ik)} .$$

Ο αντίστροφος μετασχηματισμός Fourier είναι

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(k) e^{ikx} dk = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{\lambda + ik} e^{ikx} dk .$$

### 6.10.1 Ιδιότητες

Με εφαρμογή του ορισμού του μετασχηματισμού Fourier μπορούν να αποδειχθούν μεταξύ άλλων οι επόμενες ιδιότητές του:

- $\mathcal{F}[af(x)] = a\hat{f}(k)$ , για οποιοδήποτε μιγαδικό αριθμό  $a$ ,
- $\mathcal{F}[f(x) + g(x)] = \hat{f}(k) + \hat{g}(k)$ ,
- $\mathcal{F}[f'(x)] = ik\hat{f}(k)$ ,
- $\mathcal{F}[f(ax)] = \frac{1}{a}\hat{f}\left(\frac{k}{a}\right)$ ,
- $\mathcal{F}[f(x+a)] = e^{iak}\hat{f}(k)$ ,
- $\mathcal{F}[e^{ax}f(x)] = \hat{f}(k+ia)$ ,
- $\mathcal{F}[f^*(x)] = [\hat{f}(-k)]^*$ .

Επιπλέον, εύκολα δείχνεται ότι το εσωτερικό γινόμενο δύο μη περιοδικών συναρτήσεων, που είναι τετραγωνικά ολοκληρώσιμες, είναι ίσο με το εσωτερικό γινόμενο των μετασχηματισμών Fourier τους (δηλαδή, ο μετασχηματισμός Fourier διατηρεί το εσωτερικό γινόμενο):

$$\int_{-\infty}^{\infty} f^*(x)g(x) dx = \int_{-\infty}^{\infty} [\hat{f}(k)]^* \hat{g}(k) dk . \quad (6.27)$$

Αντίστοιχα, η ταυτότητα Parseval γενικεύεται στο *θεώρημα Plancherel*:

$$\int_{-\infty}^{\infty} |f(x)|^2 dx = \int_{-\infty}^{\infty} |\hat{f}(k)|^2 dk . \quad (6.28)$$

### 6.10.2 Συμμετρία

Ο μετασχηματισμός Fourier απλοποιείται αν η συνάρτηση  $f(x)$  είναι συμμετρική ή αντισυμμετρική ως προς το  $x = 0$ . Συγκεκριμένα, ο μετασχηματισμός Fourier γράφεται

$$\begin{aligned} \hat{f}(k) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-ikx} f(x) dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) (\cos(kx) - i \sin(kx)) dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) \cos(kx) dx - i \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) \sin(kx) dx . \end{aligned}$$

Αν για τη συνάρτηση ισχύει  $f(-x) = f(x)$ , η έκφραση  $f(x) \sin(kx)$  είναι αντισυμμετρική. Το ολοκλήρωμά της σε συμμετρικό διάστημα είναι 0. Ο μετασχηματισμός Fourier είναι

$$\hat{f}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) \cos(kx) dx = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} f(x) \cos(kx) dx .$$

Παρατηρήστε ότι η  $\hat{f}(k)$  είναι και αυτή συμμετρική: ισχύει  $\hat{f}(-k) = \hat{f}(k)$ . Επομένως, ο αντίστροφος μετασχηματισμός Fourier είναι

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(k) e^{ikx} dk = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(k) \cos(kx) dk \\ &= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} \hat{f}(k) \cos(kx) dk . \end{aligned}$$

Ανάλογα ισχύουν αν η  $f(x)$  είναι αντισυμμετρική, δηλαδή ισχύει  $f(-x) = -f(x)$ . Ο μετασχηματισμός Fourier είναι<sup>4</sup>

$$\hat{f}(k) = \sqrt{\frac{2}{\pi}} \int_0^{\infty} f(x) \sin(kx) dx$$

και είναι αντισυμμετρικός. Ο αντίστροφος μετασχηματισμός είναι

$$f(x) = \sqrt{\frac{2}{\pi}} \int_0^{\infty} \hat{f}(k) \sin(kx) dk .$$

### Παράδειγμα

Έστω η συνάρτηση

$$f(x) = \begin{cases} 0, & |x| > \alpha \\ 1, & |x| \leq \alpha \end{cases} .$$

Η  $f(x)$  έχει πεπερασμένη τιμή του ολοκληρώματος της απόλυτης τιμής της:

$$\int_{-\infty}^{\infty} |f(x)| dx = \int_{-\alpha}^{\alpha} dx = 2\alpha .$$

Επομένως, έχει νόημα να υπολογίσουμε το μετασχηματισμό Fourier αυτής. Προσέξτε ότι είναι συμμετρική ως προς το  $x = 0$  άρα

$$\hat{f}(k) = \sqrt{\frac{2}{\pi}} \int_0^{\infty} f(x) \cos(kx) dx = \sqrt{\frac{2}{\pi}} \int_0^{\alpha} \cos(kx) dx = \sqrt{\frac{2}{\pi}} \frac{\sin(k\alpha)}{k} .$$

<sup>4</sup> Αν κάνετε τις πράξεις, εμφανίζεται η πολλαπλασιαστική σταθερά  $-i$  στο μετασχηματισμό Fourier και η  $i$  στον αντίστροφο μετασχηματισμό. Τις παραλείψαμε καθώς το γινόμενο των σταθερών όρων πριν τα ολοκληρώματα έχει σημασία. Χωρίσαμε το γινόμενο, δηλαδή την ποσότητα  $2/\pi$ , σε δύο πραγματικούς όρους.

Η τιμή στο  $k = 0$  της  $\hat{f}(k)$  είναι πεπερασμένη:

$$\hat{f}(0) = \lim_{k \rightarrow 0} \sqrt{\frac{2}{\pi}} \frac{\sin(k\alpha)}{k} = \lim_{k \rightarrow 0} \sqrt{\frac{2}{\pi}} \frac{k\alpha}{k} = \sqrt{\frac{2}{\pi}} \alpha .$$

Ο αντίστροφος μετασχηματισμός Fourier είναι

$$f(x) = \sqrt{\frac{2}{\pi}} \int_0^\infty \hat{f}(k) \cos(kx) dk = \frac{2}{\pi} \int_0^\infty \frac{\sin(k\alpha)}{k} \cos(kx) dk .$$

## 6.11 Συνάρτηση $\delta$

Ας ξαναγράψουμε την εξίσωση (6.24) (ουσιαστικά τον αντίστροφο μετασχηματισμό Fourier μιας συνάρτησης  $f(x)$ ), αναδιατάσσοντας τους όρους της:

$$f(x) = \int_{-\infty}^\infty \left[ \frac{1}{2\pi} \int_{-\infty}^\infty e^{-ik(y-x)} dk \right] f(y) dy .$$

Θέτουμε

$$\delta(x) = \frac{1}{2\pi} \int_{-\infty}^\infty e^{-ikx} dk , \quad (6.29)$$

οπότε, για οποιαδήποτε απόλυτα ολοκληρώσιμη συνάρτηση  $f(x)$ , έχουμε

$$f(x) = \int_{-\infty}^\infty \delta(y-x) f(y) dy .$$

Η παραπάνω σχέση δείχνει με χαρακτηριστικό τρόπο τη συμπεριφορά της έκφρασης που ονομάσαμε συνάρτηση  $\delta$ : διαλέγει σε ένα ολοκλήρωμα μια τιμή της (υπόλοιπης) ολοκληρωτέας ποσότητας. Γενικότερα, η σχέση

$$\int_a^b \delta(x-x_0) f(x) dx = \begin{cases} f(x_0) , & \text{αν } a < x_0 < b , \\ 0 , & \text{σε άλλη περίπτωση.} \end{cases} \quad (6.30)$$

αποτελεί ένα από τους ισοδύναμους ορισμούς της συνάρτησης  $\delta$ . Στην παραπάνω σχέση θεωρήσαμε ότι  $a < b$ .

Δεν είναι μαθηματικά ακριβής περιγραφή αλλά βοηθά να αντιληφθούμε τη συνάρτηση  $\delta(x)$  αν θεωρήσουμε ότι

- απειρίζεται στο  $x = 0$ ,
- μηδενίζεται σε  $x \neq 0$ ,
- το ολοκλήρωμά της στο  $(-\infty, \infty)$  είναι 1.

Καμία συνάρτηση με τη συνήθη έννοια του όρου δεν έχει τις παραπάνω ιδιότητες· γι' αυτό, η συνάρτηση  $\delta(x)$  θεωρείται γενικευμένη συνάρτηση που ορίζεται μέσω της συμπεριφοράς της σε ολοκλήρωμα.

Ένας ορισμός της συνάρτησης  $\delta(x)$ , ισοδύναμος της (6.30), προκύπτει αν χρησιμοποιήσουμε τη συνάρτηση βήματος

$$H(x) = \begin{cases} 1, & x > 0 \\ 0, & x < 0 \end{cases}.$$

Η συνάρτηση  $\delta(x)$  μπορεί να οριστεί ως η παράγωγος της  $H(x)$ :

$$\delta(x) = \frac{d}{dx} H(x).$$

Πράγματι, το  $H'(x)$  συμπεριφέρεται όπως η συνάρτηση  $\delta(x)$  σε ολοκλήρωμα. Καταρχάς,

$$\begin{aligned} \int_a^b f(x) H'(x) dx &= \int_a^b f(x) d[H(x)] \\ &= [f(x)H(x)]_a^b - \int_a^b H(x) f'(x) dx \\ &= f(b)H(b) - f(a)H(a) - \int_a^b H(x) f'(x) dx. \end{aligned}$$

Από τη διερεύνηση της τελευταίας σχέσης για διάφορες τιμές των  $a, b$  (για τα οποία θεωρούμε ότι  $a < b$ ) προκύπτει ότι:

- αν  $0 < a < b$

$$\begin{aligned} \int_a^b f(x) H'(x) dx &= f(b) \cdot 1 - f(a) \cdot 1 - \int_a^b 1 \cdot f'(x) dx \\ &= f(b) - f(a) - (f(b) - f(a)) = 0. \end{aligned}$$

- αν  $a < b < 0$

$$\int_a^b f(x) H'(x) dx = f(b) \cdot 0 - f(a) \cdot 0 - \int_a^b 0 \cdot f'(x) dx = 0.$$

- αν  $a < 0$  και  $b > 0$

$$\begin{aligned} \int_a^b f(x) H'(x) dx &= f(b) \cdot 1 - f(a) \cdot 0 - \int_0^b 1 \cdot f'(x) dx \\ &= f(b) - (f(b) - f(0)) = f(0). \end{aligned}$$

Άρα, για οποιαδήποτε  $f(x)$

$$\int_a^b f(x) H'(x) dx = \begin{cases} f(0), & \text{αν } a < 0 < b, \\ 0, & \text{σε άλλη περίπτωση.} \end{cases}$$

Από τη σύγκριση με την (6.30) προκύπτει ότι η παράγωγος της συνάρτησης βήματος είναι η συνάρτηση  $\delta(x)$ .

## 6.11.1 Ιδιότητες

Αποδεικνύεται εύκολα από τον ορισμό της ότι η συνάρτηση  $\delta(x)$  ικανοποιεί τις ακόλουθες σχέσεις:

- $\delta^*(x) = \delta(x)$ .

Απόδειξη: για οποιαδήποτε  $f(x)$  ισχύει

$$\begin{aligned} \int_a^b \delta^*(x) f(x) dx &= \left( \int_a^b \delta(x) f^*(x) dx \right)^* \\ &= \begin{cases} [f^*(0)]^*, & \text{αν } a < 0 < b, \\ 0, & \text{σε άλλη περίπτωση.} \end{cases} \\ &= \int_a^b \delta(x) f(x) dx. \end{aligned}$$

Χρησιμοποιήσαμε τον ορισμό (6.30). Καθώς η  $f(x)$  είναι αυθαίρετη, διαπιστώνουμε ότι το  $\delta^*(x)$  συμπεριφέρεται όπως το  $\delta(x)$ . Επομένως, η συνάρτηση  $\delta(x)$  είναι πραγματική.

Στο ίδιο συμπέρασμα καταλήγουμε αν επεξεργαστούμε την αναπαράστασή της με σειρά, (6.29), ή, ακόμα πιο απλά, αν παρατηρήσουμε ότι είναι παράγωγος πραγματικής συνάρτησης (της συνάρτησης βήματος).

- $\delta(-x) = \delta(x)$ .

Απόδειξη: για οποιαδήποτε  $f(x)$  ισχύει

$$\begin{aligned} \int_a^b \delta(-x) f(x) dx &= - \int_{-a}^{-b} \delta(y) f(-y) dy = \int_{-b}^{-a} \delta(y) f(-y) dy \\ &= \begin{cases} f(0), & \text{αν } -b < 0 < -a, \\ 0, & \text{σε άλλη περίπτωση.} \end{cases} \\ &= \int_a^b \delta(x) f(x) dx. \end{aligned}$$

Σε συνδυασμό με τον ορισμό (6.30) και καθώς η  $f(x)$  είναι αυθαίρετη, διαπιστώνουμε ότι το  $\delta(-x)$  συμπεριφέρεται όπως το  $\delta(x)$ . Επομένως, η συνάρτηση  $\delta(x)$  είναι συμμετρική.

Στο ίδιο συμπέρασμα καταλήγουμε αν επεξεργαστούμε την αναπαράστασή της με σειρά, (6.29).

- $x^n \delta(x) = 0$ , με  $n \geq 1$ .

Απόδειξη:

$$\int_a^b x^n \delta(x) f(x) dx = \begin{cases} 0 \cdot f(0), & \text{αν } a < 0 < b, \\ 0, & \text{σε άλλη περίπτωση.} \end{cases} = 0.$$

Η παραπάνω σχέση ισχύει για κάθε  $f(x)$  και οποιαδήποτε όρια  $a, b$ . Για να ικανοποιείται πρέπει να ισχύει  $x^n \delta(x) = 0$ .

- $\delta(\lambda x) = \frac{1}{|\lambda|} \delta(x)$  με  $\lambda \neq 0$ .

Απόδειξη: για οποιαδήποτε  $f(x)$  ισχύει

$$\begin{aligned} \int_a^b \delta(\lambda x) f(x) dx &= \int_a^b \delta(\pm |\lambda| x) f(x) dx = \int_a^b \delta(|\lambda| x) f(x) dx \\ &= \frac{1}{|\lambda|} \int_{a|\lambda|}^{b|\lambda|} \delta(y) f(y/|\lambda|) dy \\ &= \frac{1}{|\lambda|} \begin{cases} f(0), & \text{αν } a|\lambda| < 0 < b|\lambda|, \\ 0, & \text{σε άλλη περίπτωση.} \end{cases} \\ &= \frac{1}{|\lambda|} \int_a^b \delta(x) f(x) dx. \end{aligned}$$

Γενικότερα, αν  $x_i$  είναι τα σημεία που μια συνάρτηση  $h(x)$  μηδενίζεται (αλλά δεν μηδενίζεται σε αυτά η παράγωγός της), ισχύει

$$\delta(h(x)) = \sum_i \frac{\delta(x - x_i)}{|h'(x_i)|}.$$

- $\int_0^a \delta(1/x) dx = 0$ , με  $a$  μη αρνητική, πεπερασμένη τιμή.

Απόδειξη:

$$\int_0^a \delta\left(\frac{1}{x}\right) dx = \int_{\infty}^{1/a} \delta(y) \frac{-1}{y^2} dy = 0.$$

### 6.11.2 Παράγωγοι της συνάρτησης $\delta(x)$

Για την πρώτη παράγωγο της συνάρτησης  $\delta(x)$  ισχύει

$$\begin{aligned} \int_{-\infty}^{\infty} f(x) \delta'(x - a) dx &= \int_{-\infty}^{\infty} f(y + a) d[\delta(y)] \\ &= [f(y + a) \delta(y)]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \delta(y) f'(y + a) dy \\ &= -f'(a). \end{aligned}$$

Ανάλογες σχέσεις ισχύουν και για ανώτερες παραγώγους. Μπορεί ναδειχθεί ότι

$$\int_{-\infty}^{\infty} f(x) \delta^{(k)}(x - a) dx = (-1)^k f^{(k)}(a).$$

Εύκολα μπορεί να αποδείξει κανείς ότι

- $\delta'(-x) = -\delta'(x)$ ,
- $x\delta'(x) = -\delta(x)$  και γενικότερα,  $x^n \delta^{(n)}(x) = (-1)^n n! \delta(x)$ ,
- $x^n \delta'(x) = 0$ , με  $n \geq 1$ .

### 6.11.3 Μετασχηματισμός Fourier της συνάρτησης $\delta(x)$

Η συνάρτηση  $\delta(x)$  έχει πεπερασμένο ολοκλήρωμα της απόλυτής της τιμής σε οποιοδήποτε διάστημα. Επομένως, ορίζεται ο μετασχηματισμός Fourier, ο οποίος είναι

$$\hat{\delta}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-iky} \delta(y) dy = \frac{1}{\sqrt{2\pi}} e^{-ik \cdot 0} = \frac{1}{\sqrt{2\pi}}.$$

Ο αντίστροφος μετασχηματισμός Fourier αυτής της σταθερής συνάρτησης δίνεται στην εξίσωση (6.29) και είναι η συνάρτηση  $\delta(x)$ . Παρατηρήστε ότι η συνάρτηση  $\delta(x)$  προκύπτει από την επαλληλία κυμάτων με όλα τα μήκη κύματος και σταθερό πλάτος.

Η συνάρτηση  $\delta(x)$  είναι η πιο εντοπισμένη συνάρτηση που μπορεί να υπάρξει. Προσέξτε ότι μετασχηματισμός Fourier της είναι η πιο εκτεταμένη συνάρτηση: έχει παντού την ίδια τιμή.

### 6.11.4 Εφαρμογές στη Φυσική

Η συνάρτηση  $\delta(x)$  χρησιμοποιείται στη Φυσική για να εκφράσει, για παράδειγμα, την κατανομή στο χώρο μιας σημειακής ποσότητας. Έτσι, μια σημειακή μάζα  $m$  στο σημείο  $x_0$  στον άξονα των  $x$  αντιστοιχεί σε γραμμική κατανομή μάζας

$$\rho(x) \equiv \frac{dm}{dx} = m\delta(x - x_0).$$

Πράγματι, η συνολική μάζα στον άξονα των  $x$  επαληθεύεται ότι είναι  $m$ :

$$\int dm = \int_{-\infty}^{\infty} \rho(x) dx = \int_{-\infty}^{\infty} m\delta(x - x_0) dx = m.$$

Γενικεύεται εύκολα στις τρεις διαστάσεις: μια σημειακή μάζα στο  $\mathbf{r}_0 = (x_0, y_0, z_0)$  έχει πυκνότητα

$$\rho(x, y, z) = m\delta(x - x_0)\delta(y - y_0)\delta(z - z_0).$$

Έτσι, η συνολική μάζα σε ένα όγκο  $V$  είναι

$$\iiint_V \rho(x, y, z) dx dy dz = \begin{cases} m, & \text{αν το } \mathbf{r}_0 \text{ είναι στον όγκο } V, \\ 0, & \text{αλλιώς.} \end{cases}$$

## 6.12 Συνέλιξη συναρτήσεων

Η *συνέλιξη* (*convolution*) δύο συναρτήσεων  $f(x)$  και  $g(x)$  που ορίζονται στο  $(-\infty, \infty)$  είναι μια νέα συνάρτηση που συμβολίζεται με  $f * g$  και ορίζεται από την ακόλουθη σχέση:

$$[f * g](x) \equiv \int_{-\infty}^{\infty} f(y)g(x - y) dy. \quad (6.31)$$

Επομένως, είναι το ολοκλήρωμα της πρώτης συνάρτησης με συνάρτηση βάρους τη δεύτερη, αντεστραμμένη και μετατοπισμένη. Για να ορίζεται το ολοκλήρωμα πρέπει οι δύο συναρτήσεις  $f, g$  να «σβήνουν» αρκετά γρήγορα στο  $\pm\infty$ .



## 6.12.1 Ιδιότητες συνέλιξης

Εύκολα μπορούμε να διαπιστώσουμε ότι η συνέλιξη έχει τις ακόλουθες ιδιότητες:

- αντιμεταθετική:  $f * g = g * f$ .

Απόδειξη:

$$\begin{aligned} [f * g](x) &\equiv \int_{-\infty}^{\infty} f(y)g(x-y) dy \\ &\stackrel{z \leftarrow x-y}{=} \int_{-\infty}^{\infty} f(x-z)g(z) dz = \int_{-\infty}^{\infty} g(z)f(x-z) dz \\ &\equiv [g * f](x) . \end{aligned}$$

- προσεταιριστική:  $f * (g * h) = (f * g) * h$ .

Απόδειξη:

$$\begin{aligned} [f * (g * h)](x) &\equiv \int_{-\infty}^{\infty} f(y) \int_{-\infty}^{\infty} g(z)h(x-y-z) dz dy \\ &\stackrel{w \leftarrow y+z}{=} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y)g(w-y) dy h(x-w) dw \\ &= \int_{-\infty}^{\infty} (f * g)(w)h(x-w) dw \\ &\equiv [(f * g) * h](x) . \end{aligned}$$

- επιμεριστική ως προς την πρόσθεση:  $f * (g + h) = (f * g) + (f * h)$ .

$$\begin{aligned} [f * (g + h)](x) &\equiv \int_{-\infty}^{\infty} f(y)(g(x-y) + h(x-y)) dy \\ &= \int_{-\infty}^{\infty} f(y)g(x-y) dy + \int_{-\infty}^{\infty} f(y)h(x-y) dy \\ &\equiv [f * g](x) + [f * h](x) . \end{aligned}$$

Εύκολα αποδεικνύονται επιπλέον οι ιδιότητες

- $a \cdot (f * g) = (a \cdot f) * g = f * (a \cdot g)$ , με οποιοδήποτε μιγαδικό αριθμό  $a$ ,
- $(f * g)^* = f^* * g^*$ , για τη μιγαδική συζυγή,
- $(f * g)' = f' * g = f * g'$ , για την παράγωγο.

Παρατηρήστε ότι

$$[f * \delta](x) = \int_{-\infty}^{\infty} f(y)\delta(x-y) dy = \int_{-\infty}^{\infty} f(y)\delta(y-x) dy = f(x) .$$

Από το αποτέλεσμα αυτό, σε συνδυασμό με την αντιμεταθετική ιδιότητα, συνάγουμε ότι η συνάρτηση  $\delta(x)$  δρα ως ουδέτερη συνάρτηση στην πράξη της συνέλιξης.

## 6.12.2 Θεώρημα συνέλιξης

Ας υπολογίσουμε το μετασχηματισμό Fourier της συνάρτησης  $h = f * g$ :

$$\hat{h}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-iky} h(y) dy = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-iky} \int_{-\infty}^{\infty} f(z) g(y-z) dz dy .$$

Θέτουμε  $y = w + z$ . Τότε

$$\begin{aligned} \hat{h}(k) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-ik(w+z)} f(z) g(w) dz dw \\ &= \frac{1}{\sqrt{2\pi}} \left[ \int_{-\infty}^{\infty} e^{-ikz} f(z) dz \right] \left[ \int_{-\infty}^{\infty} e^{-ikw} g(w) dw \right] \\ &= \frac{1}{\sqrt{2\pi}} \sqrt{2\pi} \hat{f}(k) \sqrt{2\pi} \hat{g}(k) \\ &= \sqrt{2\pi} \hat{f}(k) \hat{g}(k) . \end{aligned} \quad (6.32)$$

Παρατηρούμε ότι ο μετασχηματισμός Fourier της συνέλιξης δύο συναρτήσεων είναι το γινόμενο των μετασχηματισμών Fourier των συναρτήσεων επί  $\sqrt{2\pi}$ . Το συμπέρασμα αυτό αποτελεί το *θεώρημα συνέλιξης*.

Με εφαρμογή του θεωρήματος συνέλιξης μπορούμε να βρούμε εύκολα το μετασχηματισμό Fourier μιας άγνωστης συνάρτησης  $f(x)$  αν γνωρίζουμε τους μετασχηματισμούς Fourier των συναρτήσεων  $h(x) = (f * g)(x)$  και  $g(x)$ :

$$\hat{f}(k) = \frac{1}{\sqrt{2\pi}} \frac{\hat{h}(k)}{\hat{g}(k)} .$$

Κατόπιν, ο αντίστροφος μετασχηματισμός Fourier (6.26) ανασυνθέτει την  $f(x)$ . Η διαδικασία αυτή, η *αποσυνέλιξη* (*deconvolution*), βρίσκει εφαρμογή κατά την επεξεργασία σημάτων ή εικόνων στη σεισμολογία, αστρονομία, οπτική, κλπ.

## 6.13 Συσχέτιση συναρτήσεων

Η *ετερο-συσχέτιση* (*cross-correlation*) δύο συναρτήσεων  $f(x)$  και  $g(x)$ , γενικά μιγαδικών, που ορίζονται στο  $(-\infty, \infty)$ , είναι μια νέα συνάρτηση που συνήθως συμβολίζεται με  $f \otimes g$  και ορίζεται από την ακόλουθη σχέση:

$$[f \otimes g](z) \equiv \int_{-\infty}^{\infty} f^*(x) g(x+z) dx . \quad (6.33)$$

Η ετερο-συσχέτιση δίνει ένα μέτρο της ομοιότητας που έχουν οι δύο συναρτήσεις όταν η μία είναι μετατοπισμένη ως προς την άλλη. Παίρνει «μεγάλη» τιμή για μια μετατόπιση  $z$  όταν η  $f(x)$  και η  $g(x+z)$  είναι παρόμοιες για κάθε  $x$  ενώ η τιμή της

είναι «μικρή» για μετατόπιση  $z$  που καθιστά τις δύο συναρτήσεις αρκετά διαφορετικές. Εύκολα διαπιστώνουμε ότι η ετερο-συσχέτιση δύο συναρτήσεων συνδέεται με τη συνέλιξή τους με τη σχέση:

$$f \otimes g = f^*(-x) * g.$$

Από το ορισμό μπορούμε να δούμε ότι η ετερο-συσχέτιση έχει

- την προσεταιριστική ιδιότητα:  $f \otimes (g \otimes h) = (f \otimes g) \otimes h$ .
- την επιμεριστική ιδιότητα ως προς την πρόσθεση:  $f \otimes (g + h) = (f \otimes g) + (f \otimes h)$ .

Προσέξτε ότι δεν έχει την αντιμεταθετική ιδιότητα:

$$\begin{aligned} [f \otimes g](z) &= \int_{-\infty}^{\infty} f^*(x)g(x+z) dx \\ &\stackrel{y \leftarrow x+z}{=} \int_{-\infty}^{\infty} f^*(y-z)g(y) dy = \left[ \int_{-\infty}^{\infty} f(y-z)g^*(y) dy \right]^* \\ &= [g \otimes f]^*(-z). \end{aligned}$$

Εύκολα αποδεικνύονται επιπλέον οι ιδιότητες

- $a(f \otimes g) = (a^* f) \otimes g = f \otimes (a g)$ , με οποιοδήποτε μιγαδικό αριθμό  $a$ ,
- $(f \otimes g)^* = f^* \otimes g^*$ , για τη μιγαδική συζυγή,
- $(f \otimes g)' = f \otimes g' = -[g \otimes f']^*$ , για την παράγωγο.
- $(f \otimes g) \otimes (f \otimes g) = (f \otimes f) \otimes (g \otimes g)$ .

### 6.13.1 Θεώρημα Wiener-Khinchin

Ας υπολογίσουμε το μετασχηματισμό Fourier της συνάρτησης  $h = f \otimes g$ :

$$\hat{h}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-iky} h(y) dy = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-iky} \int_{-\infty}^{\infty} f^*(z)g(z+y) dz dy.$$

Θέτουμε  $w = y + z$ . Τότε

$$\begin{aligned} \hat{h}(k) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-ik(w-z)} f^*(z)g(w) dz dw \\ &= \frac{1}{\sqrt{2\pi}} \left[ \int_{-\infty}^{\infty} e^{ikz} f^*(z) dz \right] \left[ \int_{-\infty}^{\infty} e^{-ikw} g(w) dw \right] \\ &= \frac{1}{\sqrt{2\pi}} \sqrt{2\pi} [\hat{f}(k)]^* \sqrt{2\pi} \hat{g}(k) \\ &= \sqrt{2\pi} [\hat{f}(k)]^* \hat{g}(k). \end{aligned} \tag{6.34}$$

Η τελευταία έκφραση για το μετασχηματισμό Fourier της ετερο-συσχέτισης αποτελεί το *θεώρημα Wiener-Khinchin*.

Με ανάλογο τρόπο μπορούμε να αποδείξουμε τη σχέση

$$\mathcal{F}[f^*(x)g(x)] = \frac{1}{\sqrt{2\pi}} \hat{f}(k) \otimes \hat{g}(k).$$

### 6.13.2 Αυτοσυσχέτιση συνάρτησης

Αν στον ορισμό της ετερο-συσχέτισης (6.33) θέσουμε  $g = f$ , παράγουμε τη συνάρτηση αυτοσυσχέτισης (autocorrelation) για τη συνάρτηση  $f(x)$ :

$$[f \otimes f](z) \equiv \int_{-\infty}^{\infty} f^*(x) f(x+z) dx . \quad (6.35)$$

Ο μετασχηματισμός Fourier της συνάρτησης αυτοσυσχέτισης είναι, σύμφωνα με το θεώρημα Wiener–Khinchin (6.34):

$$\mathcal{F}[f \otimes f] = \sqrt{2\pi} \left| \hat{f}(k) \right|^2 .$$

Η ποσότητα  $\left| \hat{f}(k) \right|^2$  αποτελεί τη φασματική πυκνότητα ενέργειας της συνάρτησης  $f(x)$  στον κυματάριθμο  $k$ . Η ποσότητα  $\left| \hat{f}(k) \right|^2 dk$  αντιπροσωπεύει την «ενέργεια» που συνεισφέρουν οι συνιστώσες με κυματάριθμους στο διάστημα  $[k, k + dk]$ . Ως συνολική «ενέργεια» ορίζουμε την έκφραση

$$E = \int_{-\infty}^{\infty} |f(x)|^2 dx .$$

Ο ορισμός της  $\left| \hat{f}(k) \right|^2$  ως πυκνότητας ενέργειας σημαίνει ότι η συνολική ενέργεια μπορεί να υπολογιστεί από το ολοκλήρωμα

$$\int_{-\infty}^{\infty} \left| \hat{f}(k) \right|^2 dk ,$$

δηλαδή καταλήξαμε με άλλο τρόπο στο θεώρημα Plancherel ((6.28)).

## 6.14 Ασκήσεις

1. Ποιες από τις επόμενες συναρτήσεις μπορούν να αναπτυχθούν σε σειρά Fourier; σε ποια σημεία δεν θα συγκλίνει η σειρά στη συνάρτηση;

- $\tanh^{-1} x$ ,
- $\tan x$ ,
- $1/\sqrt{|\sin x|}$ .

2. Βρείτε τη σειρά Fourier που προσεγγίζει την

$$f(x) = \begin{cases} 0, & 0 \leq x < 1/2, \\ 1, & 1/2 \leq x < 3/2, \\ 0, & 3/2 \leq x < 2. \end{cases}$$

Θεωρούμε ότι η συνάρτηση επαναλαμβάνεται για  $x \geq 2$  και  $x < 0$  ώστε  $f(x+2k) = f(x)$  με οποιοδήποτε ακέραιο  $k$ .

3. Βρείτε τη σειρά Fourier της  $f(x) = x$  στο διάστημα  $[-\pi, \pi)$ . Κατόπιν, δείξτε ότι

$$1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots = \frac{\pi}{4}.$$

4. Βρείτε τη σειρά Fourier της συνάρτησης  $f(t) = |\sin(\omega t)|$  με  $-\pi < \omega t < \pi$ . Ποιες συχνότητες έχουν μη μηδενικό πλάτος και πόσο;

5. Βρείτε τη μιγαδική σειρά Fourier της συνάρτησης  $f(x) = |x|$  με  $-\pi < x < \pi$ . Κατόπιν, δείξτε ότι

$$1 + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \cdots = \frac{\pi^2}{8}.$$

6. Επεκτείνετε (α') συμμετρικά (β') αντισυμμετρικά την  $f(x) = 1 - x$  που ορίζεται στο  $[0, 1)$ . Βρείτε τις αντίστοιχες σειρές Fourier.

7. Βρείτε τη σειρά Fourier της  $f(x) = x^3$  στο διάστημα  $[0, 2)$ .

8. Βρείτε τη σειρά Fourier της  $f(x) = x^2$  στο διάστημα  $[-2, 2)$ . Δείξτε ότι

$$1 + \frac{1}{2^4} + \frac{1}{3^4} + \frac{1}{4^4} + \cdots = \frac{\pi^4}{90}.$$

9. Βρείτε τη σειρά Fourier της  $f(x) = e^x$  στο διάστημα  $[-1, 1)$ . Ποια τιμή έχει η σειρά στο  $x = 2$ ;

10. Δείξτε ότι η σειρά Fourier της συνάρτησης  $f(x) = |x|$  στο διάστημα  $[-\pi, \pi)$  είναι

$$|x| = \frac{\pi}{2} - \frac{4}{\pi} \sum_{m=0}^{\infty} \frac{\cos(2m+1)x}{(2m+1)^2}.$$

Ολοκληρώστε τη συγκεκριμένη σειρά Fourier και βρείτε έτσι τη συνάρτηση που έχει σειρά Fourier

$$\frac{4}{\pi} \sum_{m=0}^{\infty} \frac{\sin(2m+1)x}{(2m+1)^3}.$$

στο συγκεκριμένο διάστημα.

11. Μπορείτε να βρείτε από τα αποτελέσματα της άσκησης 10 την τιμή του αθροίσματος

$$1 - \frac{1}{3^3} + \frac{1}{5^3} - \frac{1}{7^3} + \cdots;$$

12. Γράψτε κώδικα που να υλοποιεί τον αλγόριθμο FFT.

13. Βρείτε το μετασχηματισμό Fourier της συνάρτησης  $f(x) = x \exp(-x^2)$  με  $x$  στο διάστημα  $(-\infty, +\infty)$ .

14. Δείξτε ότι

$$\delta(x^2 - a^2) = \frac{1}{2|a|} (\delta(x + a) + \delta(x - a)) .$$

15. Βρείτε τη συνέλιξη των συναρτήσεων  $f(x) = \delta(x + a) + \delta(x - a)$  και  $g(x) = \exp(-x^2)$ .

16. Βρείτε το μετασχηματισμό Fourier της συνέλιξης των συναρτήσεων  $f(x) = \delta(x + a) + \delta(x - a)$  και  $g(x) = \exp(-x^2)$  με και χωρίς τη χρήση του θεωρήματος συνέλιξης.

## Κεφάλαιο 7

# Διαφορικές Εξισώσεις<sup>1</sup>

### 7.1 Γενικά

Για τη επίλυση των διαφορών προβλημάτων υπάρχουν γενικά δύο τύποι μαθηματικών μοντέλων. 1. *Στατικά μοντέλα*, π.χ. το κυκλοφοριακό σύστημα μιας πόλης, ελαχιστοποίηση κόστους, κλπ. Για να λύσουμε τέτοια προβλήματα χρειαζόμαστε Γραμμικό Προγραμματισμό, Γραμμική Άλγεβρα, κλπ. και 2. *Δυναμικά Μοντέλα*, π.χ. η μετάδοση της θερμότητας, ταλαντώσεις δοκού, οποιαδήποτε μεταβολή ενός μεγέθους συναρτήσει του χρόνου, κλπ. Τέτοια προβλήματα γενικά περιγράφονται με *διαφορικές εξισώσεις (ΔΕ)*, συνήθεις ΔΕ και ΔΕ με μερικές παραγώγους. Για την επίλυσή τους χρειάζονται γνώσεις Γραμμικής Άλγεβρας, πρόβλημα ιδιοτιμών, θεωρία ΔΕ, κλπ.

Στο κεφάλαιο αυτό θα ασχοληθούμε με την αριθμητική λύση ΔΕ. Θα υπέθετε κανείς ότι για όλες τις ΔΕ υπάρχουν αναλυτικοί τύποι που μας δίνουν τη λύση τους. Αυτό όμως δε συμβαίνει στην πράξη· οι περισσότερες ΔΕ δεν επιδέχονται αναλυτική λύση. Εκτός όμως από αυτό, για πολλά προβλήματα δε μας ενδιαφέρει τόσο η αναλυτική λύση όσο οι αριθμητικές τιμές της σε ορισμένα σημεία. Για το λόγο αυτό οι πιο κατάλληλες μέθοδοι για τη λύση ΔΕ είναι οι αριθμητικές.

### 7.2 Εισαγωγή

Θα παραθέσουμε εδώ μερικές βασικές προτάσεις και ορισμούς από τη θεωρία των ΔΕ και από την Ανάλυση που θα χρησιμοποιήσουμε παρακάτω.

**Ορισμός:** Μια εξίσωση που περιγράφει μια σχέση μεταξύ μιας ανεξάρτητης μεταβλητής, μιας εξαρτημένης συνάρτησης και μίας ή περισσότερων παραγώγων της θα τη λέμε ΔΕ, δηλ.

$$y^{(n)} = f\left(x, y(x), y'(x), \dots, y^{(n-1)}(x)\right). \quad (7.1)$$

---

<sup>1</sup>Το κεφάλαιο βασίζεται στις σημειώσεις του Καθηγητή Θ. Καλαμπούκη (1982).

Λύση της ΔΕ θα λέμε μια συνάρτηση  $\phi(x)$  παραγωγίσιμη  $n$  φορές σε ένα διάστημα, η οποία ικανοποιεί την (7.1).

$$\phi^{(n)} = f(x, \phi(x), \phi'(x), \dots, \phi^{(n-1)}(x)) .$$

Μια γενική λύση της (7.1) περιέχει  $n$  αυθαίρετες σταθερές, επομένως υπάρχει μια  $n$ -παραμετρική οικογένεια λύσεων.

Αν  $y(x_0), y'(x_0), \dots, y^{(n-1)}(x_0)$  είναι γνωστά για το σημείο  $x_0$ , τότε λέμε ότι έχουμε ένα πρόβλημα αρχικών τιμών.

**Ορισμός:** Θα λέμε ότι η συνάρτηση  $f(t, y)$ , συνεχής για  $t \in [a, b]$  και  $-\infty < y(t) < \infty$ , ικανοποιεί μια συνθήκη Lipschitz όταν ισχύει

$$|f(t, y_1) - f(t, y_2)| \leq L |y_1 - y_2| \quad (7.2)$$

για κάθε  $t \in [a, b]$  και  $y_1, y_2 \in \mathcal{R}$ .

**Θεώρημα:** Αν  $f(t, y)$  είναι συνεχής για  $t \in [a, b]$  και  $-\infty < y < \infty$  και ικανοποιεί μια συνθήκη Lipschitz τότε το πρόβλημα αρχικών τιμών

$$\begin{aligned} y' &= f(t, y) \\ y(0) &= y_0 \end{aligned}$$

έχει μία μοναδική λύση  $y(t)$ ,  $t \in [a, b]$ .

Στα παρακάτω προϋποθέτουμε ότι η ΔΕ πληροί όλες εκείνες τις συνθήκες που τις εξασφαλίζουν την ύπαρξη και το μονοσήμαντο της λύσης.

### 7.2.1 Διωνυμικό Ανάπτυγμα

$$(1+x)^p = 1 + \frac{p}{1!}x + \frac{p(p-1)}{2!}x^2 + \dots + \binom{p}{n}x^n + \dots, \quad |x| < 1, \quad (7.3)$$

όπου

$$\binom{p}{n} \equiv \frac{p!}{n!(p-n)!} .$$

**Θεώρημα Taylor:** Αν η  $f(x)$  έχει συνεχείς παραγώγους σ' ένα διάστημα, τότε το ανάπτυγμά της στην περιοχή ενός σημείου  $\alpha$  δίνεται από τον τύπο

$$f(\alpha + h) = f(\alpha) + \frac{h}{1!}f'(\alpha) + \frac{h^2}{2!}f''(\alpha) + \dots + \frac{h^{n-1}}{(n-1)!}f^{(n-1)}(\alpha) + R_n, \quad (7.4)$$

όπου

$$R_n = \frac{h^n}{n!}f^{(n)}(\alpha + \xi) \quad \text{για } \xi \in (0, h).$$



Για δύο διαστάσεις:

Αν η  $f(x, y)$  έχει συνεχείς μερικές παραγώγους σ' ένα διάστημα τότε

$$f(\alpha + h, \beta + k) = f(\alpha, \beta) + \frac{1}{1!} df(\alpha, \beta) + \frac{1}{2!} d^2 f(\alpha, \beta) + \cdots + \frac{1}{(n-1)!} d^{n-1} f(\alpha, \beta) + R'_n, \quad (7.5)$$

όπου

$$R'_n = \frac{1}{n!} d^n f(\alpha + \theta h, \beta + \theta k) \quad \text{για } \theta \in (0, 1)$$

και

$$\begin{aligned} df &= h \frac{\partial f}{\partial x} + k \frac{\partial f}{\partial y} \\ d^2 f &= h^2 \frac{\partial^2 f}{\partial x^2} + 2kh \frac{\partial^2 f}{\partial x \partial y} + k^2 \frac{\partial^2 f}{\partial y^2} \\ &\vdots \end{aligned}$$

## 7.3 Κατηγορίες και Λύσεις Διαφορικών Εξισώσεων

### 7.3.1 Πρωτοβάθμιες ΔΕ

Η ΔΕ  $y' = f(x, y)$  είναι

**Διαχωρίσιμος** Όταν η  $f(x, y)$  μπορεί να γραφεί ως  $U(x)V(y)$ . Π.χ.

$$\frac{dy}{dx} = \lambda xy \Rightarrow \int \frac{dy}{y} = \int \lambda x dx \Rightarrow \ln y = \lambda \frac{x^2}{2} + c \Rightarrow y = A \exp\left(\frac{\lambda x^2}{2}\right).$$

Η σταθερά  $A$  υπολογίζεται από τις αρχικές συνθήκες.

**Ομογενής** Όταν ισχύει  $f(\lambda x, \lambda y) = f(x, y)$  για όλα τα κατάλληλα  $\lambda \in \mathcal{R}^*$ . Π.χ.

$$f(x, y) = \frac{x - y}{x + y}.$$

**Γραμμική πρώτου βαθμού** Όταν η  $f(x, y)$  είναι συνάρτηση της μορφής  $-P(x)y + Q(x)$ , γραμμική, δηλαδή, ως προς  $y$ . Τότε

$$\frac{dy}{dx} + P(x)y = Q(x).$$

Ολοκληρώσιμος παράγων είναι το  $\exp\left(\int P(x) dx\right)$ , ποσότητα που δεν είναι εύκολο να υπολογιστεί στην πράξη.

### 7.3.2 Δευτεροβάθμιες ΔΕ

Η διαφορική εξίσωση

$$\alpha y'' + \beta y' + \gamma y = f(x) \quad (7.6)$$

χαρακτηρίζεται ως γραμμική δεύτερου βαθμού με σταθερούς συντελεστές. Αν  $n$   $f(x) \neq 0$  τότε η (7.6) λέγεται μη ομογενής. Η αντίστοιχη ομογενής είναι η

$$\alpha y'' + \beta y' + \gamma y = 0. \quad (7.7)$$

**Θεώρημα:** Αν  $y = G(x)$  είναι η γενική λύση της (7.7) και  $y = y_1(x)$  είναι μια μερική λύση της (7.6), τότε η  $y = G(x) + y_1(x)$  είναι η γενική λύση της (7.6).

**Γενική λύση της (7.7)** Αναζητούμε λύσεις της μορφής  $y = e^{mx}$  με  $m$  σταθερό, οπότε έχουμε

$$\begin{aligned} \alpha m^2 e^{mx} + \beta m e^{mx} + \gamma e^{mx} &= 0, \quad \text{ή} \\ \alpha m^2 + \beta m + \gamma &= 0. \end{aligned}$$

Η τελευταία εξίσωση λέγεται *χαρακτηριστική εξίσωση* της (7.7).

Για τη λύση της (7.6) διακρίνουμε λοιπόν τρεις περιπτώσεις.

1. Η χαρακτηριστική εξίσωση έχει δύο πραγματικές ρίζες  $m_1, m_2$ . Η γενική λύση της (7.7) είναι

$$y = c_1 \exp(m_1 x) + c_2 \exp(m_2 x).$$

Παράδειγμα:

$$\frac{d^2 y}{dx^2} - \lambda^2 y = 0.$$

Για  $y = e^{mx}$  έχουμε  $m^2 - \lambda^2 = 0$  ή  $m = \pm \lambda$  και η λύση θα είναι

$$y = c_1 e^{\lambda x} + c_2 e^{-\lambda x}.$$

2. Η χαρακτηριστική εξίσωση έχει μία διπλή ρίζα  $m_1 = m_2 = m$ . Τότε η  $e^{mx}$  είναι μία λύση της (7.7). Μια άλλη λύση, γραμμικά ανεξάρτητη από την πρώτη, είναι η  $x e^{mx}$ . Η γενική λύση της (7.7) είναι

$$y = c_1 e^{mx} + c_2 x e^{mx}.$$

3. Η χαρακτηριστική εξίσωση έχει μιγαδικές ρίζες  $\alpha \pm i\beta$ . Σε αυτήν την περίπτωση οι δύο λύσεις είναι  $e^{(\alpha \pm i\beta)x}$ . Κατάλληλοι γραμμικοί συνδυασμοί τους δίνουν δύο άλλες, γραμμικά ανεξάρτητες, πραγματικές λύσεις

$$e^{(\alpha+i\beta)x} + e^{(\alpha-i\beta)x} \propto e^{\alpha x} \cos(\beta x)$$

και

$$e^{(\alpha+i\beta)x} - e^{(\alpha-i\beta)x} \propto e^{\alpha x} \sin(\beta x).$$

Η γενική λύση της (7.7) είναι

$$y = c_1 e^{\alpha x} \cos(\beta x) + c_2 e^{\alpha x} \sin(\beta x) .$$

Παράδειγμα:

$$\frac{d^2 y}{dx^2} + \lambda^2 y = 0 .$$

Για  $y = e^{mx}$  έχουμε  $m^2 + \lambda^2 = 0$  ή  $m = \pm i\lambda$ . Η λύση θα είναι

$$y = c_1 \cos(\lambda x) + c_2 \sin(\lambda x) .$$

### 7.3.3 Σύστημα πρωτοβάθμιων ΔΕ με σταθερούς συντελεστές

Ένα σύστημα ΔΕ πρώτου βαθμού με σταθερούς συντελεστές μπορεί να γραφεί σε διανυσματική μορφή ως εξής

$$\dot{q} \equiv \frac{dq}{dt} = Aq , \quad q(t=0) = q_0 , \quad (7.8)$$

όπου  $A$  ένας πραγματικός πίνακας με σταθερά στοιχεία. Η λύση της (7.8) είναι

$$q = q_0 e^{At}$$

όπου

$$e^{At} = I + At + \frac{1}{2!}(At)^2 + \frac{1}{3!}(At)^3 + \dots$$

Θα δώσουμε παρακάτω μια μέθοδο επίλυσης της (7.8) με τη βοήθεια του προβλήματος ιδιοτιμών.

Έστω  $P^{-1}AP = \Lambda$  όπου  $\Lambda$  είναι ο διαγώνιος πίνακας με στοιχεία τις ιδιοτιμές του  $A$  και  $P$  ο πίνακας με στήλες τα ιδιοδιανύσματα. Πολλαπλασιάζουμε από αριστερά την (7.8) με  $P^{-1}$  οπότε έχουμε

$$\frac{dP^{-1}q}{dt} = P^{-1}APP^{-1}q . \quad (7.9)$$

Αν  $z(t) = P^{-1}q(t)$  τότε από την (7.9) έχουμε

$$\dot{z} \equiv \frac{dz}{dt} = \Lambda z ,$$

και η λύση της δίνεται από τον τύπο  $z = z_0 e^{\Lambda t}$  ή  $z_i = (z_0)_i \exp(\lambda_i t)$ . Από τη σχέση  $q = Pz$  βρίσκουμε τη λύση  $q(t)$  της (7.8).

#### Παράδειγμα

Έχουμε το ακόλουθο το σύστημα διαφορικών εξισώσεων

$$\begin{aligned} \frac{dy_1}{dt} &= 2y_1 + 3y_2 , \\ \frac{dy_2}{dt} &= -4y_1 - 5y_2 , \end{aligned}$$

με αρχικές συνθήκες  $y_1(0) = 1, y_2(0) = 0$ .

Το σύστημα αυτό γράφεται στη μορφή  $\dot{y} = Ay$  όπου

$$A = \begin{bmatrix} 2 & 3 \\ -4 & -5 \end{bmatrix}$$

Οι ιδιοτιμές του  $A$  είναι  $-1$  και  $-2$ , με αντίστοιχα ιδιοδιανύσματα  $[1, -1]^T$  και  $[-0.75, 1]^T$ . Οι  $P, P^{-1}$  επομένως είναι

$$P = \begin{bmatrix} 1 & -0.75 \\ -1 & 1 \end{bmatrix}, \quad P^{-1} = \begin{bmatrix} 4 & 3 \\ 4 & 4 \end{bmatrix}.$$

Άρα

$$\dot{z} = P^{-1}y_0 = \begin{bmatrix} 4 & 3 \\ 4 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix},$$

οπότε

$$z = \begin{bmatrix} 4e^{-t} \\ 4e^{-2t} \end{bmatrix},$$

και

$$y = Pz = \begin{bmatrix} 1 & -0.75 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 4e^{-t} \\ 4e^{-2t} \end{bmatrix} = \begin{bmatrix} 4e^{-t} - 3e^{-2t} \\ -4e^{-t} + 4e^{-2t} \end{bmatrix}.$$

## 7.4 Μέθοδος Σειράς Taylor

Έστω το πρόβλημα αρχικών τιμών

$$y' = f(x, y), \quad x \in [a, b] \quad (7.10\alpha')$$

$$y(a) = y_0. \quad (7.10\beta')$$

Το πρώτο βήμα σε μία αριθμητική μέθοδο επίλυσης του (7.10) είναι η διαμέριση του διαστήματος  $[a, b]$  σε πεπερασμένο αριθμό υποδιαστημάτων

$$a \equiv x_0 < x_1 < \cdots < x_n \equiv b.$$

Κατόπιν, ως εύρεση αριθμητικής λύσης του (7.10) θα εννοούμε τον προσδιορισμό των τιμών της  $y(x)$  στα σημεία  $x_i, i = 1, 2, \dots, n$ . Για ευκολία, έστω ότι τα σημεία είναι ισαπέχοντα. Η απόσταση διαδοχικών σημείων είναι  $h = \frac{b-a}{n}$  όπου  $n$  ο αριθμός των υποδιαστημάτων και  $x_k = a + kh, k = 0, 1, \dots, n$ . Στα επόμενα θα συμβολίζουμε με  $y(x_k)$  την ακριβή λύση στο  $x_k$  και με  $y_k$  την αντίστοιχη προσεγγιστική, όπως θα προκύπτει από μία αριθμητική μέθοδο.

Θα υπολογίσουμε τη λύση του (7.10) χρησιμοποιώντας το ανάπτυγμα Taylor της  $y$  στο  $x_r$ . Έστω ότι η τιμή της λύσης στο  $x_r$  είναι γνωστή. Τότε

$$y(x_{r+1}) = y(x_r) + hy'(x_r) + \frac{h^2}{2!}y''(x_r) + \cdots + \frac{h^{m-1}}{(m-1)!}y^{(m-1)}(x_r) + R_m \quad (7.11)$$

όπου

$$R_m = \frac{h^m}{m!} y^{(m)}(\xi) \quad \xi \in (x_r, x_{r+1}) .$$

Ισχύουν τα ακόλουθα

$$\begin{aligned} y' &= f(x, y) \equiv f^{[0]} \\ y'' &= \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} \frac{dy}{dx} = f_x + f_y f \equiv f^{[1]} \\ y''' &= f_{xx} + f_{xy} f + f(f_{yx} + f_{yy} f) + f_y(f_x + f_y f) = f_x^{[1]} + f_y^{[1]} f \equiv f^{[2]} . \end{aligned}$$

Αποδεικνύεται ότι ισχύει ο αναδρομικός τύπος

$$f^{[r+1]} = f_x^{[r]} + f_y^{[r]} f , \quad r = 0, 1, \dots . \quad (7.12)$$

Επομένως

$$y^{(m)} = f^{[m-1]} . \quad (7.13)$$

Από τις δύο προηγούμενες σχέσεις μπορούμε να υπολογίσουμε παραγώγους υψηλού βαθμού. Όταν έχουμε υπολογίσει ένα ικανό αριθμό παραγώγων, τότε μπορούμε να υπολογίσουμε την τιμή της λύσης στο σημείο  $x_r + h \equiv x_{r+1}$  από την (7.11). Με τον ίδιο τρόπο προχωράμε από το  $x_r + h$  στο  $x_r + 2h$  κλπ. μέχρι ότου καλύψουμε όλο το διάστημα  $[a, b]$ .

### Παράδειγμα

Έστω το πρόβλημα αρχικών τιμών

$$y' = \lambda y , \quad y(0) = 1 .$$

Η αναλυτική λύση είναι  $y = e^{\lambda x}$ .

Ισχύουν

$$f^{[0]} = \lambda y , \quad f^{[1]} = \lambda^2 y , \quad f^{[2]} = \lambda^3 y , \dots$$

και

$$y(0 + h) = y(h) = 1 + \lambda h + \frac{1}{2!}(\lambda h)^2 + \frac{1}{3!}(\lambda h)^3 + \dots = e^{\lambda h} .$$

Γενικά, η  $y^{(m)}(x)$  εξαρτάται από όλες τις μερικές παραγώγους της  $f$  βαθμού  $\leq m-1$ . Επομένως, όταν υπολογίζουμε τη σειρά Taylor χρησιμοποιώντας όρους μέχρι και  $m$  βαθμού, θα πρέπει να υπολογίσουμε  $m$  παραγώγους βαθμού  $m-1$ ,  $m-1$  παραγώγους βαθμού  $m-2$  κλπ., χρειάζεται, δηλαδή, ο υπολογισμός  $\frac{m(m+1)}{2}$  συναρτήσεων. Αν η  $f$  είναι πολύπλοκη συνάρτηση, όπως είναι συνήθως, τότε ο υπολογισμός όλων αυτών των παραγώγων είναι μια επίπονη και χρονοβόρα δουλειά. Άρα, αντί να χρησιμοποιήσουμε μια υψηλού βαθμού σειρά Taylor για τον υπολογισμό του  $y(x)$  σε ένα σχετικά μεγάλο διάστημα, είναι προτιμότερο να διαιρέσουμε το διάστημα  $[a, b]$  σε μικρά τμήματα και να χρησιμοποιήσουμε μια σειρά Taylor μικρότερου βαθμού.

**Αλγόριθμος μεθόδου Taylor βαθμού  $m - 1$ :**

1. Εκλογή κατάλληλου βήματος  $h = \frac{b-a}{n}$ .
2.  $y(a) = y_0$ .
3. Για  $r = 0, 1, \dots, n$  θέτουμε

$$y_{r+1} = y_r + hf_r^{[0]} + \dots + \frac{h^{m-1}}{(m-1)!} f_r^{[m-2]}, \quad (7.14)$$

όπου τα  $f_r^{[i]}$  θα υπολογιστούν από την (7.12).

#### 7.4.1 Μέθοδος Euler

Η απλούστερη από τις μεθόδους Taylor είναι η μέθοδος Euler ( $m = 1$ ) που προκύπτει από την (7.11) αν αποκόψουμε τους όρους της σειράς μετά το δεύτερο όρο, δηλαδή,

$$y_{r+1} = y_r + hf(x_r, y_r). \quad (7.15)$$

Η μέθοδος Euler είναι πολύ εύκολο να προγραμματιστεί. Σε κάθε βήμα υπολογίζουμε το  $f(x_r, y_r)$ . το οποίο χρησιμοποιούμε για τον υπολογισμό του  $y_{r+1}$  κλπ.

##### Παράδειγμα

Έστω  $y' = -y$ ,  $y(0) = 1$ ,  $x \in [0, 1]$ , της οποίας η αναλυτική λύση είναι  $y(x) = e^{-x}$ . Η μέθοδος Euler δίνει

$$y_{r+1} = (1 - h)y_r.$$

Για  $h = 0.1$  δίνουμε στον παρακάτω πίνακα, μερικές τιμές της λύσης και τις αντίστοιχες ακριβείς.

$x$	$y$	$y(x)$
0.0	1.0000	1.0000
0.1	0.9000	0.9048
0.2	0.8100	0.8187
0.3	0.7290	0.7408
0.4	0.6561	0.6703
0.5	0.5905	0.6065
0.6	0.5314	0.5488
0.7	0.4783	0.4966
0.8	0.4305	0.4493
0.9	0.3874	0.4066
1.0	0.3486	0.3679

### 7.4.2 Σφάλμα Μεθόδου Taylor

Το σφάλμα στη μέθοδο Taylor προέρχεται από δύο πηγές: 1. *Σφάλμα αποκοπής*, που προέρχεται από την προσέγγιση (7.14) (global discretization error ή global truncation error) και 2. το *σφάλμα στρογγύλευσης*, που οφείλεται στο πεπερασμένο της αναπαράστασης των πραγματικών αριθμών από τον ΗΥ. Στα επόμενα αγνοούμε το σφάλμα στρογγύλευσης και θεωρούμε ότι όλες οι πράξεις γίνονται χωρίς σφάλμα.

Ορίζουμε ως

$$E(h) = \max_{1 \leq r \leq n} |y_r - y(x_r)| , \quad (7.16)$$

το *ολικό σφάλμα αποκοπής* ( $E(h) \rightarrow 0$  όταν  $h \rightarrow 0$ ), και

$$L(x, h) = \frac{1}{h} [y(x+h) - y(x)] - f(x, y(x)) , \quad (7.17)$$

το *τοπικό σφάλμα αποκοπής*, που χαρακτηρίζει τη διαφοροποίηση της προσεγγιστικής παραγώγου  $y'$  από την ακριβή τιμή.

Υποθέτουμε ότι η  $y''$  είναι φραγμένη στο  $[a, b]$ , δηλαδή

$$\max_{a \leq x \leq b} |y''(x)| = M ,$$

και ότι  $y = y(x_r)$  για κάποιο  $r$ . Τότε

$$y(x_{r+1}) - y_{r+1} = y(x_{r+1}) - y(x_r) = hf(x_r, y(x_r)) = hL(x_r, h) ,$$

δηλαδή το σφάλμα σε ένα βήμα στη μέθοδο Euler είναι  $h$  φορές το τοπικό σφάλμα αποκοπής αρχίζοντας από την αρχική λύση.

Ενδιαφερόμαστε για το  $\max L(x, h)$  για κάθε τιμή του  $x$ , συνεπώς ορίζουμε ως *τοπικό σφάλμα αποκοπής για τη μέθοδο Euler* το

$$L(h) = \max_{a \leq x \leq b-h} |L(x, h)| .$$

Από το ανάπτυγμα Taylor, (7.11), και την (7.4.2) έχουμε ότι

$$L(h) \leq \frac{h}{2} M = O(h) .$$

Με  $O(h)$  υποδηλώνουμε ότι η ποσότητα τείνει στο 0 ανάλογα με το  $h$ .

Το πρόβλημά μας είναι να βρούμε μια σχέση μεταξύ του τοπικού και του ολικού σφάλματος αποκοπής. Αν θέσουμε  $z_r = y(x_r) - y_r$  τότε

$$\begin{aligned} z_{r+1} &\equiv y(x_{r+1}) - y_{r+1} = y(x_r) + hf(x_r, y(x_r)) + hL(x_r, h) - y_r - hf(x_r, y_r) \\ &= z_r + h[f(x_r, y(x_r)) - f(x_r, y_r)] + hL(x_r, h) . \end{aligned} \quad (7.18)$$

Έστω ότι

$$\left| \frac{\partial f(x, y)}{\partial y} \right| \leq M_1 \quad x \in [a, b] , \quad |y| < \infty .$$

Τότε, από το Θεώρημα μέσης τιμής για  $0 < \theta < 1$  έχουμε

$$|f(x_r, y(x_r)) - f(x_r, y_r)| = \left| \frac{\partial f}{\partial y}(x_r, \theta y(x_r) + (1 - \theta)y_r)(y(x_r) - y_r) \right| \leq M_1 z_r. \quad (7.19)$$

Από τις (7.19) και (7.18) έχουμε ότι

$$|z_{r+1}| \leq (1 + hM_1) |z_r| + h |L(h)|.$$

Αν  $c = 1 + hM_1$  τότε

$$\begin{aligned} |z_{r+1}| &\leq c |z_r| + hL(h) \\ &\leq c^2 |z_{r-1}| + ch |L(h)| + h |L(h)| \\ &\leq \dots \\ &\leq c^r |z_1| + c^{r-1} h |L(h)| + \dots + ch |L(h)| + h |L(h)|. \end{aligned}$$

Το φράγμα αυτό περιέχει το άθροισμα  $n$  όρων της μορφής  $O(h^2)$  (γιατί;), και καθώς  $n = \frac{b-a}{h}$ , το άθροισμα είναι της μορφής  $O(h)$ . Συνεπώς, αν η συνάρτηση  $f$  έχει φραγμένη μερική παράγωγο ως προς  $y$  και αν η λύση του προβλήματος αρχικών τιμών έχει φραγμένη δεύτερη παράγωγο, τότε το ολικό σφάλμα αποκοπής στη μέθοδο Euler ικανοποιεί το  $E(h) \propto h$  (Η μέθοδος Euler είναι πρώτου βαθμού). Για την ολοκλήρωση της απόδειξης χρειάζεται να δειχτεί ότι το  $c^n \equiv (1 + hM_1)^n$  είναι φραγμένο όταν  $h \rightarrow 0$ .

## 7.5 Μέθοδος Runge-Kutta

Μια επαναληπτική μέθοδος θα λέμε ότι είναι βαθμού  $p$  όταν το τοπικό σφάλμα αποκοπής μπορεί γενικά να γραφεί ως

$$R(x, h) = ch^{p+1}y^{(p+1)}(\xi), \quad \xi \in (x, x + h).$$

Όπως είδαμε, το τοπικό σφάλμα αποκοπής στη μέθοδο του Euler είναι ανάλογο του  $h^2$ , επομένως, για να έχουμε καλή ακρίβεια, θα πρέπει το  $h$  να είναι πολύ μικρό. Αυτό είναι και το βασικότερο μειονέκτημα της μεθόδου του Euler γιατί όταν το  $h$  μικραίνει τότε αυξάνει το σφάλμα στρογγύλευσης. Αντίθετα, στη μέθοδο Euler  $n$  βαθμού, το τοπικό σφάλμα είναι ανάλογο του  $h^{n+1}$ , οπότε μπορούμε να χρησιμοποιήσουμε μεγαλύτερο βήμα  $h$ . Το μειονέκτημα είναι ότι θα πρέπει να υπολογίσουμε υψηλές παραγώγους, βαθμού  $n - 1$ , πράγμα που είναι δύσκολο στην πράξη. Η μέθοδος Runge-Kutta (RK) που θα περιγράψουμε παρακάτω ξεπερνά όλες τις δυσκολίες της μεθόδου Taylor (υψηλές παράγωγοι) και της Euler (μικρό  $h$ ). Η ιδέα της μεθόδου RK είναι να δώσουμε στο  $y_{r+1} = y(x_r + h)$  μια μορφή η οποία να ταυτίζεται με το ανάπτυγμα Taylor για ένα συγκεκριμένο αριθμό όρων της σειράς (βαθμός του  $h$ ) χωρίς, όμως, να υπολογίζουμε στην πράξη τις παραγώγους της  $f(x, y)$  όπως απαιτείται στη μέθοδο Taylor.



Έστω το πρόβλημα αρχικών τιμών

$$y' = f(x, y), \quad (7.20\alpha')$$

$$y(x_0) = y_0. \quad (7.20\beta')$$

Η μέθοδος RK ουσιαστικά αντικαθιστά την προσέγγιση της  $y_{r+1}$  από τη σειρά Taylor

$$y_{r+1} = y_r + hy'_r + \frac{h^2}{2!}y''_r + \frac{h^3}{3!}y'''_r + \dots \quad (7.21)$$

με τη σχέση

$$y_{r+1} = y_r + [\omega_1 f(x_r, y_r) + \omega_2 f(x_r + c_2 h, y_r + d_2 h) + \omega_3 f(x_r + c_3 h, y_r + d_3 h) + \dots + \omega_p f(x_r + c_p h, y_r + d_p h)] . \quad (7.22)$$

Οι σταθερές  $\omega_i$ ,  $c_i$ ,  $d_i$  στην (7.22) υπολογίζονται κατά τέτοιο τρόπο ώστε όταν αναπτύξουμε στο δεύτερο μέλος της εξίσωσης τις συναρτήσεις  $f$  κατά Taylor και εξισώσουμε με την (7.21) τότε οι συντελεστές των ομοιόβαθμων όρων να συμπίπτουν. Για να απλοποιήσουμε τις πράξεις και να συστηματοποιήσουμε τη μέθοδο, εκφράζουμε τα  $d_i$  ως γραμμικούς συνδυασμούς των προηγούμενων τιμών της  $f$ . Δηλαδή, αντί της (7.22) γράφουμε

$$y_{r+1} = y_r + \omega_1 k_1 + \omega_2 k_2 + \dots + \omega_p k_p \quad (7.23)$$

με

$$k_1 = hf(x_r, y_r) \quad (7.24\alpha')$$

$$k_2 = hf(x_r + c_2 h, y_r + a_{21} k_1) \quad (7.24\beta')$$

$$k_3 = hf(x_r + c_3 h, y_r + a_{31} k_1 + a_{32} k_2) \quad (7.24\gamma')$$

$$\vdots \quad \vdots \quad (7.24\delta')$$

$$k_p = hf(x_r + c_p h, y_r + a_{p1} k_1 + a_{p2} k_2 + \dots + a_{p(p-1)} k_{p-1}) , \quad (7.24\epsilon')$$

όπου οι σταθερές  $\omega_i$ ,  $c_i$ ,  $a_{ij}$  θα πρέπει να υπολογιστούν. Παρακάτω θα εξετάσουμε την περίπτωση για  $p = 2$  που είναι η απλούστερη.

### 7.5.1 Μέθοδος Runge–Kutta 2<sup>ου</sup> βαθμού

Για  $p = 2$  οι (7.23, 7.24) γίνονται

$$y_{r+1} = y_r + \omega_1 k_1 + \omega_2 k_2 \quad (7.25\alpha')$$

$$k_1 = hf(x_r, y_r) \quad (7.25\beta')$$

$$k_2 = hf(x_r + c_2 h, y_r + a_{21} k_1) . \quad (7.25\gamma')$$

Αναπτύσσουμε την  $f(x_r + c_2 h, y_r + a_{21} k_1)$  κατά Taylor:

$$k_2 = h \left[ f + c_2 h f_x + a_{21} k_1 f_y + \frac{1}{2!} ((c_2 h)^2 f_{xx} + 2c_2 h a_{21} k_1 f_{xy} + (a_{21} k_1)^2 f_{yy} + \dots) \right] ,$$

όπου  $f \equiv f(x_r, y_r)$ ,  $f_x \equiv \frac{\partial f}{\partial x}$ ,  $f_y \equiv \frac{\partial f}{\partial y}$ ,  $f_{xx} \equiv \frac{\partial^2 f}{\partial x^2}$ ,  $f_{xy} \equiv \frac{\partial^2 f}{\partial x \partial y}$ ,  $f_{yy} \equiv \frac{\partial^2 f}{\partial y^2}$ .

Αντικαθιστούμε τα  $k_1, k_2$  στην (7.25)

$$\begin{aligned} y_{r+1} &= y_r + \omega_1 h f \\ &\quad + \omega_2 h \left[ f + c_2 h f_x + a_{21} k_1 f_y \right. \\ &\quad \left. + \frac{1}{2!} \left( (c_2 h)^2 f_{xx} + 2c_2 h a_{21} k_1 f_{xy} + (a_{21} k_1)^2 f_{yy} + \dots \right) \right] \end{aligned} \quad (7.26)$$

$$\begin{aligned} &= y_r + (\omega_1 + \omega_2) h f + \omega_2 c_2 h^2 f_x + \omega_2 a_{21} h^2 f f_y \\ &\quad + \frac{\omega_2 h^3}{2!} (c_2^2 f_{xx} + 2c_2 a_{21} f f_{xy} + a_{21}^2 f^2 f_{yy}) + \dots \end{aligned} \quad (7.27)$$

Το ανάπτυγμα της  $y$  κατά Taylor στο  $x_r$  δίνεται από την (7.21). Ισχύουν

$$y' = f, \quad (7.28\alpha')$$

$$y'' = f_x + f_y f, \quad (7.28\beta')$$

$$y''' = f_{xx} + 2f f_{xy} + f^2 f_{yy} + f_y (f_x + f f_y). \quad (7.28\gamma')$$

Αντικαθιστώντας τις (7.28) στην (7.21) έχουμε:

$$y_{r+1} = y_r + h f + \frac{h^2}{2!} (f_x + f_y f) + \frac{h^3}{3!} [f_{xx} + 2f f_{xy} + f^2 f_{yy} + f_y (f_x + f f_y)] + \dots \quad (7.29)$$

Αν στις σχέσεις (7.27, 7.29) εξισώσουμε τους ομοιόβαθμους όρους μέχρι και δεύτερου βαθμού ως προς  $h$  έχουμε το σύστημα

$$\omega_1 + \omega_2 = 1,$$

$$\omega_2 c_2 = \frac{1}{2},$$

$$\omega_2 a_{21} = \frac{1}{2}.$$

Το σύστημα αυτό είναι τριών εξισώσεων με τέσσερις αγνώστους, δηλαδή έχουμε ένα βαθμό ελευθερίας (και άρα άπειρες λύσεις). Αν  $\omega_2 = \alpha \neq 0$  με  $\alpha$  σταθερά, τότε  $\omega_1 = 1 - \alpha$ ,  $c_2 = a_{21} = \frac{1}{2\alpha}$ . Για τις τιμές αυτές των  $\omega_1, \omega_2, c_2, a_{21}$  η (7.27) γίνεται

$$y_{r+1} = y_r + h f + \frac{h^2}{2} (f_x + f f_y) + \frac{h^3}{8\alpha} (f_{xx} + 2f f_{xy} + f^2 f_{yy}) + O(h^4). \quad (7.30)$$

Το τοπικό σφάλμα αποκοπής της μεθόδου θα είναι (7.29)–(7.30) δηλαδή

$$\left( \frac{h^3}{6} - \frac{h^3}{8\alpha} \right) (f_{xx} + 2f f_{xy} + f^2 f_{yy}) + \frac{h^3}{6} f_y (f_x + f f_y) + O(h^4). \quad (7.31)$$

Από την (7.31) βλέπουμε ότι δεν μπορούμε να καθορίσουμε την αυθαίρετη σταθερά  $\alpha$  έτσι ώστε το τοπικό σφάλμα αποκοπής να είναι ανάλογο του  $h^4$ . Μια απλή

μορφή στους σχετικούς τύπους της μεθόδου RK δίνει η επιλογή  $\alpha = 1/2$  οπότε  $\omega_1 = \omega_2 = 1/2$ ,  $c_2 = a_{21} = 1$  με αντίστοιχους τύπους δεύτερου βαθμού

$$\begin{aligned}y_{r+1} &= y_r + \frac{1}{2}(k_1 + k_2), \\k_1 &= hf(x_r, y_r) \\k_2 &= hf(x_r + h, y_r + k_1).\end{aligned}$$

Το τοπικό σφάλμα αποκοπής για τη συγκεκριμένη επιλογή γίνεται

$$-\frac{h^3}{12}(f_{xx} + 2ff_{xy} + f^2f_{yy} + 2f_xf_y + 2ff_y^2) + O(h^4).$$

Παρατηρούμε ότι το τοπικό σφάλμα αποκοπής της RK ( $p = 2$ ) είναι ανάλογο του  $h^3$ , σε αντίθεση με τη μέθοδο Euler στην οποία είναι ανάλογο του  $h^2$ . Συνεπώς, μπορούμε να χρησιμοποιήσουμε μεγαλύτερο βήμα  $h$ .

### 7.5.2 Μέθοδος Runge-Kutta 4<sup>ου</sup> βαθμού

Με το ίδιο τρόπο όπως για τη RK δευτέρου βαθμού μπορούμε να κατασκευάσουμε μια RK υψηλότερου βαθμού. Σε μια RK  $n$  βαθμού το τοπικό σφάλμα αποκοπής είναι ανάλογο του  $h^{n+1}$ . Η πιο εύχρηστη είναι η RK τετάρτου βαθμού:

$$\begin{aligned}y_{r+1} &= y_r + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4), \\k_1 &= hf(x_r, y_r), \\k_2 &= hf\left(x_r + \frac{h}{2}, y_r + \frac{k_1}{2}\right), \\k_3 &= hf\left(x_r + \frac{h}{2}, y_r + \frac{k_2}{2}\right), \\k_4 &= hf(x_r + h, y_r + k_3).\end{aligned}$$

Στη περίπτωση αυτή το σφάλμα είναι ανάλογο του  $h^5$  αλλά χρειάζονται τέσσερις υπολογισμοί της  $f$  σε κάθε επανάληψη.

### 7.5.3 Σχόλια

**Πλεονεκτήματα RK** Η μέθοδος γράφεται εύκολα σε κώδικα, είναι οικονομική σε μνήμη, είναι εύκολο να ξεκινήσει (χρειάζεται μία αρχική τιμή) και το μήκος του βήματος  $h$  μπορεί εύκολα να μεταβληθεί από επανάληψη σε επανάληψη χωρίς επιπλέον δουλειά.

**Μειονεκτήματα RK** Είναι δύσκολο να υπολογίσουμε το σφάλμα αποκοπής. Ο υπολογισμός ενός φράγματος για το σφάλμα εξαρτάται από παράγοντες οι οποίοι δεν παρουσιάζονται απ' ευθείας στη μέθοδο και συνεπώς απαιτεί επιπλέον υπολογισμούς. Επίσης, μια RK βαθμού  $n$  απαιτεί  $n$  υπολογισμούς της συνάρτησης  $f$ : όταν αυτή είναι πολύπλοκη, όπως συνήθως, χρειάζεται σεβαστός αριθμός πράξεων σε κάθε επανάληψη.

## 7.6 Τελεστές Διαφορών

Ο σκοπός αυτής της παραγράφου είναι διπλός: πρώτον, για να δούμε τη δύναμη και την απλότητα των μεθόδων με τελεστές στην κατασκευή διαφορών τύπων που είναι χρήσιμοι σε πολλά θέματα της Αριθμητικής Ανάλυσης, όπως στις ΔΕ. Δεύτερον, για να εξαγάγουμε μερικούς από τους τελεστές που θα χρησιμοποιήσουμε παρακάτω.

Αν  $f(x)$  είναι πραγματική συνάρτηση, ορισμένη σε ένα διάστημα  $[a, b]$ , χωρίζουμε το διάστημα αυτό σε  $n$  ίσα τμήματα μήκους  $h$  το καθένα, δηλαδή,  $[a \equiv x_0, x_1], [x_1, x_2], \dots, [x_{n-1}, x_n \equiv b]$ . Τότε  $x_r = x_0 + rh$ . Θεωρούμε ότι είναι γνωστές οι τιμές της  $f$  στα σημεία  $x_i$ ,  $i = 0, 1, \dots, n$ ,  $y_i = f(x_i)$ . Ορίζουμε ως *τελεστή διαφοράς προς τα εμπρός*, και συμβολίζουμε με  $\Delta$  το

$$\Delta y_r = y_{r+1} - y_r, \quad y_{r+1} = f(x_{r+1}) = f(x_r + h).$$

Ισχύει ότι

$$\Delta^2 y_r = \Delta(\Delta y_r) = y_{r+2} - 2y_{r+1} + y_r.$$

Γενικά ισχύει  $\Delta^p y_r = \Delta^{p-1}(\Delta y_r)$ .

### Παράδειγμα

Έστω  $y = x^n$ . Τότε

$$\begin{aligned} \Delta^n x^n &= \Delta^{n-1}(\Delta x^n) = \Delta^{n-1}[(x+h)^n - x^n] = \Delta^{n-1}(nhx^{n-1} + \dots) \\ &= \Delta^{n-2}[nh(n-1)x^{n-2} + \dots] = \dots \\ &= n!h^n. \end{aligned}$$

Επομένως, η  $n$ -ιοστή διαφορά του  $x^n$  είναι ένας σταθερός αριθμός. Γενικά ισχύει ότι η  $n$ -ιοστή διαφορά ενός πολυωνύμου  $n$  βαθμού είναι σταθερός αριθμός.

### Παράδειγμα

Να υπολογιστούν οι διαφορές της  $y = x^2$  στο  $[0, 2]$ , με  $h = 0.5$ .

$r$	$x$	$y = x^2$	$\Delta y$	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
0	0.0	0.0				
			0.25			
1	0.5	0.25		0.5		
			0.75		0.0	
2	1.0	1.0		0.5		0.0
			1.25		0.0	
3	1.5	2.25		0.5		
			1.75			
4	2.0	4.0				

Από το παράδειγμα βλέπουμε ότι οι διαφορές  $\Delta^3 y$  είναι μηδέν. Γενικά ισχύει ότι οι  $n + 1$  διαφορές ενός πολυωνύμου βαθμού  $n$  είναι μηδέν. Αν η  $f(x)$  είναι συνάρτηση και όχι πολυώνυμο, τότε μπορούμε να βρούμε ένα  $h$  τέτοιο ώστε οι διαφορές υψηλότερου βαθμού συνεχώς να ελαττώνονται σε μέγεθος.

### Παράδειγμα

Έστω  $f(x) = e^{kx}$ . Τότε

$$\begin{aligned}\Delta f(x) &= e^{k(x+h)} - e^{kx} = e^{kx} (e^{kh} - 1) \\ \Delta^2 f(x) &= e^{kx} (e^{kh} - 1)^2 \\ &\vdots \\ \Delta^n f(x) &= e^{kx} (e^{kh} - 1)^n\end{aligned}$$

Για να ελαττώνονται οι διαφορές θα πρέπει  $e^{kh} - 1 < 1$  ή  $h < \frac{\ln 2}{k}$ .

#### 7.6.1 Ιδιότητες

Ο τελεστής  $\Delta$ , όπως και οι άλλοι τελεστές που θα δούμε παρακάτω, υπακούουν στους νόμους της Άλγεβρας.

**Ισότητα τελεστών** Δυο τελεστές  $T_1, T_2$ , είναι ίσοι όταν για κάθε  $f(x)$  ισχύει  $T_1 f(x) = T_2 f(x)$ .

**Τελεστής μετατόπισης** Ο τελεστής μετατόπισης  $E$  ορίζεται από τη σχέση  $y_{r+1} = E y_r$ . Εύκολα μπορούμε να δούμε ότι  $E = 1 + \Delta$ , όπου 1 είναι ο ταυτοτικός τελεστής, δηλαδή ο τελεστής που αφήνει μια συνάρτηση  $f$  αμετάβλητη.

Για τον τελεστή  $E$  η δύναμη  $E^\alpha$  για κάθε  $\alpha$  ορίζεται από

$$E^\alpha y_r = y_{r+\alpha}, \quad \text{ή} \quad E^\alpha f(x) = f(x + \alpha h).$$

**Διαφορικός τελεστής** Ο διαφορικός τελεστής  $D$  ορίζεται από τη σχέση  $y'_r = D y_r$ . Συνδέεται με τον τελεστή μετατόπισης με τη σχέση  $E = e^{hD}$ . Πράγματι

$$y_{r+1} = y_r + h D y_r + \frac{h^2}{2!} D^2 y_r + \cdots = \left[ 1 + h D + \frac{1}{2!} (h D)^2 + \cdots \right] y_r = e^{h D} y_r.$$

Συνεπώς,  $E = e^{hD}$  και  $hD = \ln E = \ln(1 + \Delta)$ . Από την τελευταία σχέση μπορούμε να υπολογίσουμε την τιμή της παραγώγου μιας συνάρτησης σε ένα σημείο  $x_r$ , όταν είναι γνωστές οι πεπερασμένες διαφορές της συνάρτησης στα  $x_i$ , από τη σχέση:

$$h y' = h D y = \ln(1 + \Delta) y = \Delta y - \frac{1}{2} \Delta^2 y + \frac{1}{3} \Delta^3 y - \cdots$$

**Παράδειγμα**

Να υπολογιστεί η παράγωγος της  $f(x) = e^x$  στο  $x = 1.0$ . Οι διαφορές της  $e^x$  δίνονται στον Πίνακα 7.1, ( $h = 0.1$ ). Άρα

$$\begin{aligned} 0.1y'_{x=1.0} &\approx 0.285884 - \frac{1}{2} \times 0.030067 + \frac{1}{3} \times 0.003162 \\ &\quad - \frac{1}{4} \times 0.000333 + \frac{1}{5} \times 0.000035 \\ &\approx 0.271828 . \end{aligned}$$

Επομένως,  $y'_{x=1.0} \approx 2.718280$ . Η ακριβής τιμή είναι  $e^1 = 2.718281828 \dots$

Με τον ίδιο τρόπο μπορούμε να υπολογίσουμε την τιμή της δεύτερης παραγώγου της  $y$  σε ένα σημείο. Από τη σχέση  $y''_r = D^2 y_r$  έχουμε

$$h^2 y''_r = h^2 D^2 y_r = [\ln(1 + \Delta)]^2 y_r = \Delta^2 y_r - \Delta^3 y_r + \frac{11}{12} \Delta^4 y_r + \dots$$

**7.6.2 Άλλοι τελεστές**

**Τελεστής διαφοράς προς τα πίσω** Συμβολίζεται με  $\nabla$  και ορίζεται από τη σχέση

$$\nabla y_r = y_r - y_{r-1} .$$

**Τελεστής κεντρικής διαφοράς** Συμβολίζεται με  $\delta$  και ορίζεται από τη σχέση

$$\delta y_r = y_{r+\frac{1}{2}} - y_{r-\frac{1}{2}} .$$

**Τελεστής μέσης τιμής** Συμβολίζεται με  $\mu$  και ορίζεται από τη σχέση

$$\mu y_r = \frac{y_{r+\frac{1}{2}} + y_{r-\frac{1}{2}}}{2} .$$

Ισχύουν οι σχέσεις

$$\begin{aligned} \nabla &= 1 - E^{-1} , \quad \delta = E^{1/2} - E^{-1/2} , \\ \delta &= e^{hD/2} - e^{-hD/2} = 2 \sinh\left(\frac{hD}{2}\right) , \quad \delta = \Delta(1 + \Delta)^{-1/2} . \end{aligned}$$

Από τη σχέση  $\delta = 2 \sinh(hD/2)$  μπορούμε να υπολογίσουμε την τιμή της δεύτερης παραγώγου μιας συνάρτησης όταν είναι γνωστές οι κεντρικές διαφορές της,  $\delta y_r$

$$h^2 y''_r = h^2 D^2 y_r = \left[ 2 \sinh^{-1}\left(\frac{\delta}{2}\right) \right]^2 y_r .$$

Καθώς

$$\sinh^{-1} z = z - \frac{1}{6} z^3 + \frac{3}{40} z^5 - \frac{5}{56} z^7 + \dots ,$$

έχουμε

$$h^2 y''_r = \left( \delta^2 - \frac{1}{2} \delta^4 + \frac{1}{90} \delta^6 + \dots \right) y_r .$$

$x$	$f(x) = e^x$	$\Delta f(x)$	$\Delta^2 f(x)$	$\Delta^3 f(x)$	$\Delta^4 f(x)$	$\Delta^5 f(x)$	$\Delta^6 f(x)$	$\Delta^7 f(x)$
0.0	1.000000	0.105171						
0.1	1.105171	0.116232	0.011061	0.001163				
0.2	1.221403	0.128456	0.012224	0.001286	0.000122	0.000013		
0.3	1.349859	0.141966	0.013510	0.001421	0.000135	0.000014	0.000001	0.000000
0.4	1.491825	0.156897	0.014931	0.001570	0.000149	0.000016	0.000001	0.000000
0.5	1.648721	0.173398	0.016501	0.001735	0.000165	0.000017	0.000002	0.000000
0.6	1.822119	0.191634	0.018236	0.001918	0.000183	0.000019	0.000002	0.000000
0.7	2.013753	0.211788	0.020154	0.002120	0.000202	0.000021	0.000002	0.000000
0.8	2.225541	0.234062	0.022274	0.002343	0.000223	0.000023	0.000002	0.000000
0.9	2.459603	0.258679	0.024617	0.002589	0.000246	0.000026	0.000002	0.000000
1.0	2.718282	0.285884	0.027205	0.002861	0.000272	0.000029	0.000003	0.000000
1.1	3.004166	0.315951	0.030067	0.003162	0.000301	0.000032	0.000003	0.000000
1.2	3.320117	0.349180	0.033229	0.003495	0.000333	0.000035		
1.3	3.669297	0.385903	0.036724	0.003862	0.000368			
1.4	4.055200	0.426489	0.040586					
1.5	4.481689							

Πίνακας 7.1: Πίνακας διαφορών της συνάρτησης  $f(x) = e^x$  με ακρίβεια 6 δεκαδικών ψηφίων

### 7.6.3 Γενικευμένοι τύποι του Newton

Έστω  $y(x)$  μια συνάρτηση η οποία γράφεται  $y(x) = y(x_r + \alpha h) = y_{r+\alpha}$ , με  $\alpha = (x - x_r)/h$ ,  $\alpha$  συνεχής. Τότε

$$\begin{aligned} y(x) &= y(x_r + \alpha h) = y_{r+\alpha} = E^\alpha y_r = (1 + \Delta)^\alpha y_r \\ &= \left( 1 + \alpha \Delta + \frac{\alpha(\alpha-1)}{2!} \Delta^2 + \frac{\alpha(\alpha-1)(\alpha-2)}{3!} \Delta^3 + \dots \right) y_r. \end{aligned} \quad (7.32)$$

Με τον ίδιο τρόπο, από τη σχέση  $E = (1 - \nabla)^{-1}$  έχουμε:

$$\begin{aligned} y_{r+\alpha} &= E^\alpha y_r = (1 - \nabla)^{-\alpha} y_r \\ &= \left( 1 + \alpha \nabla + \frac{\alpha(\alpha+1)}{2!} \nabla^2 + \frac{\alpha(\alpha+1)(\alpha+2)}{3!} \nabla^3 + \dots \right) y_r. \end{aligned} \quad (7.33)$$

### 7.6.4 Εφαρμογή των τελεστών στον υπολογισμό ολοκληρωμάτων

1. Να υπολογιστεί το ολοκλήρωμα

$$I = \int_{x_r}^{x_r+2h} y(x) dx.$$

Για  $x \in (x_r, x_r + 2h)$ ,  $x = x_r + \alpha h$ . Από την (7.32) έχουμε:

$$\begin{aligned} I &= \int_{x_r}^{x_r+2h} \left( y_r + \alpha \Delta y_r + \frac{\alpha(\alpha-1)}{2!} \Delta^2 y_r + \dots \right) dx \\ &= \int_0^2 \left( y_r + \alpha \Delta y_r + \frac{\alpha(\alpha-1)}{2!} \Delta^2 y_r + \dots \right) h d\alpha \\ &= 2h \left( y_r + \Delta y_r + \frac{1}{6} \Delta^2 y_r + 0 \Delta^3 y_r - \frac{1}{180} \Delta^4 y_r + \dots \right). \end{aligned}$$

Όταν στην παραπάνω εξίσωση αποκόψουμε τους όρους τρίτου και μεγαλύτερου βαθμού ως προς  $\Delta$  έχουμε:

$$I = \frac{h}{3} (y_r + 4y_{r+1} + y_{r+2}).$$

Η σχέση αυτή είναι η μέθοδος Simpson, §5.3, και ο κυρίαρχος όρος του σφάλματος αποκοπής θα είναι:

$$R = -\frac{h}{90} \Delta^4 y_r \Rightarrow R = -\frac{h^5}{90} y_r^{(4)}.$$

2. Να υπολογιστεί το ολοκλήρωμα

$$I = \int_{x_r}^{x_r+1} y(x) dx$$



με τη βοήθεια του τελεστή  $\nabla$ . Από την (7.33) έχουμε:

$$\begin{aligned}\int_{x_r}^{x_{r+1}} y(x) dx &= \int_{x_r}^{x_{r+1}} \left( y_r + \alpha \nabla y_r + \frac{\alpha(\alpha+1)}{2!} \nabla^2 y_r + \dots \right) dx \\ &= h \left( y_r + \frac{1}{2} \nabla y_r + \frac{5}{12} \nabla^2 y_r + \frac{3}{8} \nabla^3 y_r + \dots \right). \quad (7.34)\end{aligned}$$

3. Να υπολογιστεί το ολοκλήρωμα

$$I = \int_{x_r}^{x_{r+1}} y(x) dx$$

με τη βοήθεια του τελεστή  $\nabla$  και του  $y_{r+1}$  αντί του  $y_r$ .

$$\begin{aligned}\int_{x_r}^{x_{r+1}} y(x) dx &= \int_{-1}^0 \left( y_{r+1} + \alpha \nabla y_{r+1} + \frac{\alpha(\alpha+1)}{2!} \nabla^2 y_{r+1} + \dots \right) h d\alpha \\ &= h \left( y_{r+1} - \frac{1}{2} \nabla y_{r+1} - \frac{1}{12} \nabla^2 y_{r+1} - \frac{1}{24} \nabla^3 y_{r+1} - \dots \right). \quad (7.35)\end{aligned}$$

Από τις (7.34, 7.35) παρατηρούμε ότι οι συντελεστές του  $\nabla$  μικραίνουν γρηγορότερα στη δεύτερη. Για το λόγο αυτό, όταν αποκόπτουμε σε αυτές τους όρους από ένα συγκεκριμένο όρο και μετά, θα πρέπει να αναμένουμε καλύτερη προσέγγιση του ολοκληρώματος από την (7.35).

Όταν αποκόψουμε από την (7.35) τους όρους δεύτερου και μεγαλύτερου βαθμού, τότε παίρνουμε τον κανόνα τραπεζίου, §5.2,

$$I = \frac{h}{2} (y_{r+1} + y_r),$$

και ο κυρίαρχος όρος του σφάλματος αποκοπής θα είναι  $-\frac{1}{12} h^3 y_{r+1}''$ .

4. Να υπολογιστεί το ολοκλήρωμα

$$I = \int_{x_r}^{x_{r+1}} y(x) dx$$

με τη βοήθεια του τελεστή  $\delta$ .

Ισχύει ότι

$$D = \frac{2}{h} \sinh^{-1} \left( \frac{\delta}{2} \right) = \frac{1}{h} \left( \delta - \frac{1}{24} \delta^3 + \frac{3}{640} \delta^5 - \dots \right). \quad (7.36)$$

Για ένα  $x \in (x_r - h, x_r + h)$ ,  $x = x_r - h + \alpha h$  και

$$y(x) = E^{\alpha-1} y_r = e^{h(\alpha-1)D} y_r.$$

Η παραπάνω σχέση, σε συνδυασμό με την (7.36), δίνει μετά από πράξεις:

$$\begin{aligned} \int_{x_r}^{x_{r+1}} y(x) dx &= \int_0^2 \left[ 1 + (\alpha - 1) \left( \delta - \frac{1}{24} \delta^3 + \dots \right) \right. \\ &\quad \left. + \frac{(\alpha - 1)^2}{2!} \left( \delta - \frac{1}{24} \delta^3 + \dots \right) + \dots \right] h d\alpha \\ &= 2h \left( 1 + \frac{1}{6} \delta^2 - \frac{1}{180} \delta^4 + \frac{1}{1512} \delta^6 - \dots \right) y_r . \end{aligned}$$

## 7.7 Μέθοδοι βασιζόμενες σε αριθμητική ολοκλήρωση. Πολυβηματικές μέθοδοι

Στις μεθόδους που εξετάσαμε για την επίλυση μιας ΔΕ, Taylor, Euler, Runge-Kutta, μέσω εξίσωσης διαφορών, η ιδέα ήταν να αναπτύξουμε τη λύση σε σειρά Taylor. Στις μεθόδους αυτές, για να υπολογίσουμε την τιμή της λύσης  $y_{r+1}$  σε ένα σημείο  $x_{r+1}$ , χρησιμοποιήσαμε την τιμή της λύσης  $y_r$  σε ένα μόνο σημείο  $x_r$ . Γι' αυτό, οι μέθοδοι αυτές λέγονται μονοβηματικές. Μια μέθοδος στην οποία για τον υπολογισμό της  $y_{r+1}$  χρειαζόμαστε περισσότερες από μία τιμές της λύσης σε προηγούμενα σημεία,  $y_r, y_{r-1}, y_{r-2}, \dots$  θα τη λέμε πολυβηματική. Με άλλα λόγια, όταν για μία ΔΕ η αντίστοιχη εξίσωση διαφορών έχει βαθμό  $\geq 2$ , τότε η μέθοδος επίλυσής της είναι πολυβηματική. Στα επόμενα θα εξετάσουμε μεθόδους για την επίλυση μιας ΔΕ χρησιμοποιώντας αριθμητική ολοκλήρωση.

Έστω το πρόβλημα αρχικών τιμών

$$y' = f(x, y) , \quad (7.37\alpha')$$

$$y(x_0) = y_0 . \quad (7.37\beta')$$

Ολοκληρώνουμε τα δύο μέλη της (7.37) μεταξύ  $x_r$  και  $x_{r+1}$  οπότε έχουμε:

$$y_{r+1} - y_r = \int_{x_r}^{x_{r+1}} f(x, y) dx = \int_{x_r}^{x_{r+1}} y'(x) dx \Rightarrow y_{r+1} = y_r + \int_{x_r}^{x_{r+1}} y'(x) dx . \quad (7.38)$$

Η ιδέα εδώ είναι να αντικαταστήσουμε το ολοκλήρωμα με μια κατάλληλη προσέγγιση ή, καλύτερα, να αντικαταστήσουμε την  $f$  με ένα πολυώνυμο παρεμβολής. Υποθέτουμε ότι έχουμε ήδη υπολογίσει προσεγγίσεις  $y_0, y_1, \dots, y_n$ , της  $y(x)$  στα ισαπέχοντα σημεία  $x_r = x_0 + rh$ ,  $r = 0, 1, \dots, n$  με  $f_r$  θα συμβολίζουμε μια προσέγγιση της  $f(x_r, y(x_r))$ ,  $r = 0, 1, \dots, n$ .

### 7.7.1 Μέθοδος Adams-Bashforth

Αν για τον υπολογισμό του ολοκληρώματος (7.38) χρησιμοποιήσουμε την (7.34), παίρνουμε έναν αλγόριθμο για τη λύση του (7.37) ο οποίος λέγεται *αλγόριθμος των*

Adams–Bashforth:

$$y_{r+1} = y_r + \left(1 + \frac{1}{2}\nabla + \frac{5}{12}\nabla^2 + \frac{3}{8}\nabla^3 + \dots\right) y'_r. \quad (7.39)$$

Ο τύπος αυτός λέγεται και *τύπος της πρόβλεψης* (Predictor) ή, ακόμη, *ανοικτός τύπος* (explicit). Αν στον τύπο παρεμβολής της  $f$  χρησιμοποιήσουμε τα  $x_r$  και τα προηγούμενά τους  $x_{r-1}, x_{r-2}, \dots$  τότε η μέθοδος που παίρνουμε λέγεται ανοικτή (explicit) ενώ αν για την προσέγγιση της  $y$  στο  $x_{r+1}$  χρησιμοποιήσουμε το  $x_{r+1}$  και τα προηγούμενά του η μέθοδος λέγεται κλειστή.

Από την (7.39) έχουμε όλους τους τύπους της μεθόδου Adams–Bashforth ανάλογα με τους όρους που αποκόπτουμε κάθε φορά. Έτσι

1. Αν αποκόψουμε τους όρους μετά το σταθερό όρο τότε προκύπτει η μέθοδος Euler, §7.4.1,

$$y_{r+1} = y_r + h y'_r.$$

2. Αν κρατήσουμε και τον όρο  $\nabla$  τότε

$$y_{r+1} = y_r + \frac{h}{2} (3y'_r - y'_{r-1}),$$

με τον κυρίαρχο όρο στο σφάλμα αποκοπής τον  $\frac{5}{12}h^3 y'''(\xi)$ .

3. Αν αποκόψουμε τους όρους μετά το  $\nabla^3$  έχουμε τον τύπο

$$y_{r+1} = y_r + \frac{h}{24} (55y'_r - 59y'_{r-1} + 37y'_{r-2} - 9y'_{r-3}).$$

Όπως βλέπουμε από τα παραδείγματα αυτά, σε μία μέθοδο Adams–Bashforth βαθμού  $p$  χρειαζόμαστε  $p$  τιμές της  $y$  για να ξεκινήσουμε τον αλγόριθμο. Αυτό αποτελεί και το σοβαρότερο μειονέκτημα της μεθόδου. Για το λόγο αυτό, η μέθοδος χρησιμοποιείται σε συνδυασμό με μια άλλη μέθοδο για τον προσδιορισμό των  $y(x_0), \dots, y(x_{p-1})$ . η τελευταία είναι συνήθως μια Runge–Kutta του ίδιου βαθμού. Έχοντας γνωστή την αρχική συνθήκη  $y(x_0) = y_0$ , ξεκινάμε τη λύση της (7.37) με μία Runge–Kutta βαθμού  $p$  και υπολογίζουμε τις τιμές της λύσης στα  $x_0, x_1, \dots, x_{p-1}$ , έστω  $u_0, u_1, \dots, u_{p-1}$ . Κατόπιν, με τη βοήθεια των υπολογισμένων τιμών συνεχίζουμε τη λύση του προβλήματος με μια Adams–Bashforth βαθμού  $p$ . Το πλεονέκτημα της Adams–Bashforth σε αντίθεση με τη Runge–Kutta είναι ότι χρειαζόμαστε μόνο ένα υπολογισμό της  $f$ , ενώ στη Runge–Kutta  $p$  βαθμού χρειαζόμαστε  $p$  υπολογισμούς της  $f$  σε κάθε επανάληψη.

### 7.7.2 Μέθοδος Adams–Moulton

Θα προσεγγίσουμε και πάλι το ολοκλήρωμα στην (7.38) με τη βοήθεια της (7.35). Έχουμε

$$y_{r+1} = y_r + \left(1 - \frac{1}{2}\nabla - \frac{1}{12}\nabla^2 - \frac{1}{24}\nabla^3 - \dots\right) y'_{r+1}. \quad (7.40)$$

Η σχέση αυτή είναι ένας κλειστός τύπος (implicit), με την έννοια ότι χρειαζόμαστε πληροφορίες για την τιμή της λύσης στο  $x_{r+1}$ , (καθώς  $y'_{r+1} = f(x_{r+1}, y_{r+1})$ ). Γενικότερα, μη γραμμικές εξισώσεις προκύπτουν για τον υπολογισμό του  $y_{r+1}$ . Ο τύπος (7.40) είναι γνωστός και ως τύπος της διόρθωσης (Corrector) ή αλγόριθμος των Adams-Moulton.

Όταν στην (7.40) αποκόψουμε τους όρους μετά το  $\nabla$  έχουμε τον κανόνα τραπεζίου

$$y_{r+1} = y_r + \frac{h}{2} (y'_r + y'_{r+1}) . \quad (7.41)$$

Ο κυρίαρχος όρος του σφάλματος αποκοπής είναι  $-\frac{1}{12}h^3 y'''(\xi)$ .

### Παραδείγματα

1. Να λυθεί η

$$y' = xy .$$

Χρησιμοποιώντας τον τύπο (7.41) έχουμε

$$y_{r+1} = \frac{1 + \frac{1}{2}hx_r}{1 - \frac{1}{2}hx_{r+1}} y_r .$$

2. Να λυθεί η

$$y' = xy^2$$

με τον κανόνα τραπεζίου.

$$y_{r+1} = y_r + \frac{h}{2} (x_r y_r^2 + x_{r+1} y_{r+1}^2) .$$

Εδώ έχουμε μια μη γραμμική εξίσωση διαφορών ως προς  $y_{r+1}$  και θα πρέπει σε κάθε επανάληψη, να λύνουμε μια μη γραμμική εξίσωση, πράγμα το οποίο είναι πολύ δύσκολο στην πράξη.

### 7.7.3 Μέθοδοι Πρόβλεψης-Διόρθωσης (Predictor-Corrector)

Με τη μέθοδο πρόβλεψης υπολογίζουμε το  $y_{r+1}$  από τις γνωστές τιμές  $y_r, y_{r-1}, \dots, y_0$ , αλλά οι συντελεστές της σειράς των πεπερασμένων διαφορών (7.39) μικραίνουν πολύ αργά. Δηλαδή, για να έχουμε μια ικανοποιητική ακρίβεια θα πρέπει να πάρουμε ένα σχετικά μεγάλο αριθμό όρων στη (7.39), πράγμα το οποίο σημαίνει ότι σε κάθε επανάληψη θα έχουμε περισσότερη δουλειά, περισσότερες πράξεις, περισσότερα λάθη στρογγύλευσης. Με τη μέθοδο της διόρθωσης, οι συντελεστές της σειράς πεπερασμένων διαφορών (7.40) μπορεί να μικραίνουν πολύ γρήγορα αλλά προκύπτουν μη γραμμικές εξισώσεις.

#### Μέθοδος πρόβλεψης

$$y_{r+1} = y_r + h y'_r .$$

Το σφάλμα αποκοπής είναι

$$T_p = \frac{h^2}{2} y''(\xi) .$$

**Μέθοδος διόρθωσης**

$$y_{r+1} = y_r + \frac{h}{2} (y'_r + y'_{r+1}) .$$

Το σφάλμα αποκοπής είναι

$$T_c = -\frac{h^3}{12} y'''(\xi) .$$

Παρακάτω θα περιγράψουμε έναν αλγόριθμο που λέγεται *αλγόριθμος πρόβλεψης-διόρθωσης*. Αποτελεί ένα συνδυασμό των παραπάνω μεθόδων και ξεπερνά όλες τις δυσκολίες της καθεμίας χωριστά.

**Αλγόριθμος:** Έστω ότι δίνεται η τιμή του  $y$  στο  $x_r$ ,  $y_r = y(x_r)$ .

1. Πρόβλεψε την τιμή  $\bar{y}_{r+1}$  με μια μέθοδο Adams–Bashforth.
2. Υπολόγισε το  $\bar{y}'_{r+1} = f(x_{r+1}, \bar{y}_{r+1})$ .
3. Διόρθωσε, υπολόγισε το  $y_{r+1}$  χρησιμοποιώντας το  $\bar{y}'_{r+1}$  στο δεξί μέλος μιας μεθόδου Adams–Moulton.
4. Υπολόγισε το  $y'_{r+1} = f(x_{r+1}, y_{r+1})$ .
5. Επανάλαβε για το επόμενο διάστημα.

**Υπολογισμός του σφάλματος αποκοπής**

Διαλέγουμε ένα αλγόριθμο Πρόβλεψης–Διόρθωσης του ίδιου βαθμού και έστω  $T_p$  το σφάλμα αποκοπής από την πρόβλεψη και  $T_c$  από τη διόρθωση.

Όταν μια μέθοδος είναι βαθμού  $s$  τότε

$$T_p = A_p h^{s+1} y^{(s+1)}(\xi_p) , \quad \text{και} \quad (7.42)$$

$$T_c = A_c h^{s+1} y^{(s+1)}(\xi_c) , \quad (7.43)$$

όπου τα  $A_p$ ,  $A_c$  είναι σταθεροί συντελεστές. Υποθέτουμε ότι το διάστημα  $h = x_{r+1} - x_r$  είναι τόσο μικρό ώστε

$$y^{(s+1)}(\xi_p) \approx y^{(s+1)}(\xi_c) \approx y^{(s+1)}(\xi) . \quad (7.44)$$

Ισχύουν

$$\bar{y}_{r+1} - Y_{r+1} = T_p = A_p h^{s+1} y^{(s+1)}(\xi_p) , \quad (7.45\alpha')$$

$$y_{r+1} - Y_{r+1} = T_c = A_c h^{s+1} y^{(s+1)}(\xi_c) , \quad (7.45\beta')$$

όπου  $Y_{r+1}$  είναι η πραγματική λύση και  $\bar{y}_{r+1}$  η υπολογισμένη τιμή από την πρόβλεψη και  $y_{r+1}$  η υπολογισμένη από τη διόρθωση. Από τις (7.44, 7.45) έχουμε

$$y_{r+1} - \bar{y}_{r+1} = Ah^{s+1}y^{(s+1)}(\xi),$$

με  $A = A_p - A_c$ . Από τη σχέση αυτή είναι εύκολο να υπολογίσουμε τα  $T_p$ ,  $T_c$ .

### Παράδειγμα

Ως τύπο πρόβλεψης διαλέγουμε μια μέθοδο δεύτερου βαθμού

$$y_{r+1} = y_r + \frac{h}{2} (3y'_r - y'_{r-1}), \quad (7.46\alpha')$$

$$\text{με} \quad T_p = \frac{5}{12} h^3 y'''(\xi_p) \quad (7.46\beta')$$

και ως τύπο διόρθωσης την αντίστοιχη σχέση

$$y_{r+1} = y_r + \frac{h}{2} (y'_r + y'_{r+1}), \quad (7.47\alpha')$$

$$\text{με} \quad T_c = -\frac{1}{12} h^3 y'''(\xi_c). \quad (7.47\beta')$$

Τότε

$$T_p = \frac{5}{6} (y_{r+1} - \bar{y}_{r+1}) \quad (7.48\alpha')$$

$$\text{και} \quad T_c = -\frac{1}{6} (y_{r+1} - \bar{y}_{r+1}). \quad (7.48\beta')$$

### Επιλογή του βήματος $h$

Σκοπός μας είναι να χρησιμοποιήσουμε ένα βήμα  $h$  το οποίο θα κρατά το σφάλμα αποκοπής, το οποίο εξαρτάται από το  $h$ , μέσα σε ορισμένα όρια. Όταν το  $h$  είναι πολύ μεγάλο, τότε η λύση δεν είναι ακριβής (μεγάλο σφάλμα αποκοπής). όταν το  $h$  είναι πολύ μικρό τότε και πάλι η λύση δεν είναι καλή (πολλά λάθη στρογγύλευσης).

Όμως, μπορούμε να ξεκινήσουμε μια μέθοδο Πρόβλεψης-Διόρθωσης με βήμα  $h$  το οποίο θα μεταβάλεται σε κάθε επανάληψη ανάλογα με το σφάλμα αποκοπής. Όταν το σφάλμα είναι μεγάλο θα παίρνουμε μικρότερο βήμα στην επόμενη επανάληψη και αντίστροφα. Θα πρέπει, βέβαια, να καθορίσουμε τι είναι μεγάλο και τι μικρό σφάλμα αποκοπής. Αλλάζοντας την τιμή του  $h$  σε κάθε επανάληψη εμφανίζονται πολλές δυσκολίες· το πρόβλημα απλοποιείται αν πάρουμε ως νέο βήμα το διπλάσιο ή το μισό του παλαιού, ανάλογα με το σφάλμα.

Στην πρώτη περίπτωση, όταν δηλαδή, πάρουμε ως νέο  $h$  το διπλάσιο του προηγούμενου, το πρόβλημα είναι απλό καθώς  $\tilde{x}_{r-1} \equiv x_{r-2}$ ,  $\tilde{x}_{r-2} \equiv x_{r-4}$ , κλπ. και οι τιμές της  $y$  είναι γνωστές στα σημεία  $\tilde{x}_{r-1}$ ,  $\tilde{x}_{r-2}$ ,  $\tilde{x}_{r-3}, \dots$ . Όταν το νέο  $h$  είναι ίσο με

$h/2$  τότε  $\tilde{x}_{r-2} \equiv x_{r-1}$ ,  $\tilde{x}_{r-4} \equiv x_{r-2}, \dots$  το νέο  $\tilde{x}_{r-1}$  είναι το σημείο στο μέσο μεταξύ των  $x_{r-1}$  και  $x_r$ , και αντίστοιχα για τα υπόλοιπα. Στην περίπτωση αυτή, θα πρέπει να υπολογίσουμε με παρεμβολή την τιμή της  $y$  στο  $\tilde{x}_{r-1} \equiv x_{r-\frac{1}{2}}$ ,  $y_{r-\frac{1}{2}} = y(x_{r-\frac{1}{2}})$ , κ.ο.κ.

Όταν στον αλγόριθμο Πρόβλεψης-Διόρθωσης που δώσαμε παραπάνω, ενσωματώσουμε τη μέθοδο υπολογισμού του σφάλματος και τη μέθοδο επιλογής του διαστήματος  $h$ , ο αλγόριθμος γίνεται:

**Αλγόριθμος Πρόβλεψης-Διόρθωσης:** Κατ' αρχήν αποφασίζουμε ότι

$$E_{\min} < |E_r| < E_{\max}.$$

1. Πρόβλεψε το  $\bar{y}_{r+1}$ , υπολόγισε το  $\bar{y}'_{r+1} = f(x_{r+1}, \bar{y}_{r+1})$ .
2. Διόρθωσε βρίσκοντας το  $y_{r+1}$ , υπολόγισε το  $y'_{r+1} = f(x_{r+1}, y_{r+1})$ .
3. Υπολόγισε το σφάλμα από τα  $\bar{y}_{r+1}$ ,  $y_{r+1}$ .
4. Αν  $|E_r| > E_{\max}$  τότε  $h \leftarrow h/2$  και επανάλαβε από το  $x_r$ .
5. Αν  $|E_r| < E_{\min}$  τότε  $h \leftarrow 2h$  και επανάλαβε από το  $x_r$ .
6. Αλλιώς, συνέχισε με το ίδιο βήμα  $h$  ξανά από την πρώτη εντολή.

Η μέθοδος Πρόβλεψης-Διόρθωσης είναι πολυβηματική μέθοδος, επομένως, θα πρέπει να γνωρίζουμε περισσότερες της μίας αρχικές συνθήκες για να ξεκινήσουμε. Ο πιο συνηθισμένος τρόπος είναι να εφαρμόσουμε μια Runge-Kutta ανάλογου βαθμού για το ξεκίνημα του αλγορίθμου και μόλις υπολογιστούν οι απαιτούμενες αρχικές συνθήκες να συνεχίσουμε με μέθοδο Πρόβλεψης-Διόρθωσης. Ένα άλλο σημείο στο οποίο πρέπει να παρεμβάλουμε τη μέθοδο Runge-Kutta είναι το βήμα 4. Στο βήμα αυτό θα υπολογίσουμε το  $y_{r-\frac{1}{2}}$  με Runge-Kutta από τις γνωστές τιμές της  $y$  στα προηγούμενα σημεία.

## 7.8 Συστήματα Διαφορικών Εξισώσεων

Οι μέθοδοι αριθμητικής επίλυσης ΔΕ που εξετάσαμε, μπορούν εύκολα να εφαρμοστούν στην περίπτωση συστημάτων ΔΕ ή μιας ΔΕ υψηλότερου βαθμού ( $\geq 2$ ). Έστω το σύστημα των ΔΕ πρώτου βαθμού

$$y'_1 = f_1(x, y_1, y_2, \dots, y_n) \quad (7.49\alpha')$$

$$y'_2 = f_2(x, y_1, y_2, \dots, y_n) \quad (7.49\beta')$$

$$\vdots \quad \vdots \quad (7.49\gamma')$$

$$y'_n = f_n(x, y_1, y_2, \dots, y_n) \quad (7.49\delta')$$

όπου οι  $f_i$  είναι πραγματικές συναρτήσεις, ορισμένες για  $x \in [a, b]$  και για κάθε πραγματικό  $y_1, y_2, \dots, y_n$ . Οι τιμές των  $y_i$  στο  $(n+1)$  βήμα θα υπολογιστούν από τις τιμές των  $y'_i$  και τις προηγούμενες τιμές των  $y_i, y'_i$ , με τον ίδιο τρόπο όπως στις απλές ΔΕ. Ο υπολογισμός των  $y'_i$  από την (7.49) είναι το μόνο σημείο στο οποίο υπάρχει διαφορά από την απλή περίπτωση καθώς έχουμε  $n$  τιμές  $y_i$  αντί για μία, όπως στις απλές ΔΕ. Γενικά, η λύση του (7.49), αν υπάρχει, δε θα είναι μοναδική, εκτός αν προσδιοριστούν  $n$  αρχικές συνθήκες:

$$y_i(x_0) = s_i, \quad i = 1, \dots, n, \quad (7.50)$$

όπου τα  $s_i$  είναι γνωστά και  $x_0 \in [a, b]$ . Οι (7.49, 7.50) συνιστούν ένα πρόβλημα αρχικών τιμών που σε διανυσματική μορφή γράφεται:

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \quad (7.51\alpha')$$

$$\mathbf{y}(x_0) = \mathbf{s}. \quad (7.51\beta')$$

### Παράδειγμα

Να λυθεί το πρόβλημα αρχικών τιμών

$$\begin{aligned} y'_1 &= xy_1 - y_2 \\ y'_2 &= -y_1 + y_2 \\ y_1(0) &= s_1 \\ y_2(0) &= s_2 \end{aligned}$$

Ορίζουμε  $y_{10} \equiv y_1(0)$  και  $y_{20} \equiv y_2(0)$ . Η σειρά Taylor 2<sup>ου</sup> βαθμού είναι:

$$\begin{aligned} y_{1r+1} &= y_{1r} + h(x_r y_{1r} - y_{2r}) + \frac{h^2}{2} [(x_r^2 + 2)y_{1r} - (1 + x_r)y_{2r}] \\ y_{2r+1} &= y_{2r} + h(-y_{1r} + y_{2r}) + \frac{h^2}{2} [-(1 + x_r)y_{1r} + 2y_{2r}], \end{aligned}$$

με  $r = 0, 1, \dots$

Η Runge-Kutta δεύτερου βαθμού δίνει:

$$\begin{aligned} y_{1r+1} &= y_{1r} + \frac{h}{2} [x_r y_{1r} - y_{2r} + (x_r - h)(y_{1r} + x_r y_{1r} - 2y_{2r}) - 2y_{2r} + y_{1r}] \\ y_{2r+1} &= y_{2r} + \frac{h}{2} (4y_{2r} - 3y_{1r} - x_r y_{1r}). \end{aligned}$$

Η μέθοδος Πρόβλεψης-Διόρθωσης εφαρμόζεται με παρόμοιο τρόπο.

**Παρατήρηση:** Όπως είναι γνωστό, μια ΔΕ βαθμού  $\geq 2$  μπορεί να γραφτεί ως ένα σύστημα ΔΕ πρώτου βαθμού, οπότε για την αριθμητική επίλυση μιας ΔΕ βαθμού ανώτερου του πρώτου, μετατρέπουμε την εξίσωση σε ισοδύναμο σύστημα ΔΕ πρώτου βαθμού και συνεχίζουμε κατά τα γνωστά.



Το (7.51) έχει μία μοναδική λύση αν η  $\mathbf{f}$  ικανοποιεί μια συνθήκη Lipschitz, §7.2, δηλαδή,  $\exists L > 0$  τέτοιο ώστε  $\forall \mathbf{y}, \mathbf{z} \in \mathcal{R}^n$  και για κάθε  $x \in [a, b]$  να ισχύει

$$\|\mathbf{f}(x, \mathbf{y}) - \mathbf{f}(x, \mathbf{z})\|_{\infty} \leq L \|\mathbf{y} - \mathbf{z}\|_{\infty} .$$

Οι διάφορες μέθοδοι επίλυσης ΔΕ πρώτης τάξης γενικεύονται εύκολα για συστήματα ΔΕ πρώτου βαθμού. Η μέθοδος Taylor θα είναι:

$$\mathbf{y}(x_{r+1}) = \mathbf{y}(x_r) + h\mathbf{f}(x_r, \mathbf{y}(x_r)) + \frac{h^2}{2!}\mathbf{f}'(x_r, \mathbf{y}(x_r)) + \cdots + \frac{h^p}{p!}\mathbf{f}^{(p)}(x_r, \mathbf{y}(x_r)) + \mathbf{R}_{p+1} ,$$

όπου

$$\begin{aligned} \mathbf{f}'(x, \mathbf{y}(x)) \equiv \frac{d}{dx}\mathbf{f}(x, \mathbf{y}(x)) &= \frac{\partial \mathbf{f}}{\partial x} + \sum_{j=1}^n \frac{dy_j(x)}{dx} \frac{\partial \mathbf{f}}{\partial y_j} = \frac{\partial \mathbf{f}}{\partial x} + \sum_{j=1}^n f_j \frac{\partial \mathbf{f}}{\partial y_j} , \\ \frac{\partial \mathbf{f}}{\partial y_j} &\equiv \left( \frac{\partial f_1}{\partial y_j}, \frac{\partial f_2}{\partial y_j}, \dots, \frac{\partial f_n}{\partial y_j} \right) . \end{aligned}$$

Η μέθοδος Taylor δεύτερου βαθμού θα είναι:

$$\begin{aligned} \mathbf{y}(x_{r+1}) &= \mathbf{y}(x_r) + h\mathbf{f}(x_r, \mathbf{y}_r) + \frac{h^2}{2!}\mathbf{f}'(x_r, \mathbf{y}_r) , \\ \mathbf{y}_0 &= \mathbf{s} . \end{aligned}$$

Η Runge–Kutta 4<sup>ov</sup> βαθμού θα είναι:

$$\mathbf{y}_{r+1} = \mathbf{y}_r + \frac{h}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4)$$

με

$$\begin{aligned} \mathbf{k}_1 &= \mathbf{f}(x_r, \mathbf{y}_r) , \\ \mathbf{k}_2 &= \mathbf{f}(x_r + h/2, \mathbf{y}_r + h/2\mathbf{k}_1) , \\ \mathbf{k}_3 &= \mathbf{f}(x_r + h/2, \mathbf{y}_r + h/2\mathbf{k}_2) , \\ \mathbf{k}_4 &= \mathbf{f}(x_r + h, \mathbf{y}_r + h\mathbf{k}_3) . \end{aligned}$$

## 7.9 Εξισώσεις Διαφορών

Για να μπορέσουμε στα επόμενα να μιλήσουμε για ακρίβεια και ευστάθεια των προσεγγίσεων των λύσεων που παίρνουμε από διάφορες μεθόδους όπως η Euler, Runge–Kutta κλπ. για την επίλυση ενός προβλήματος αρχικών τιμών θα πρέπει να χρησιμοποιήσουμε ορισμένες ιδιότητες των εξισώσεων διαφορών (ΕΔ). Στην §7.2 δώσαμε τον ορισμό μιας ΔΕ ως μια σχέση μεταξύ μιας ανεξάρτητης μεταβλητής  $x$ , μιας εξαρτημένης συνάρτησης  $y$  και μιας ή περισσότερων παραγώγων της. Ομοίως, ορίζουμε μια ΕΔ ως μια σχέση μεταξύ μιας ανεξάρτητης μεταβλητής  $x$ , μιας εξαρτημένης συνάρτησης  $y$  και μιας ή περισσότερων διαφορών της  $\Delta y, \Delta^2 y, \dots$ . Δηλαδή,

εκεί που μια ΔΕ περιέχει μια συνάρτηση σε ένα διάστημα πραγματικών αριθμών και τις παραγώγους της, μια ΕΔ περιέχει τη συνάρτηση ορισμένη σε ένα διάστημα ακεραίων και τις διαφορές της. Όπως φαίνεται από τον ορισμό παραπάνω, υπάρχει μεγάλη ομοιότητα μεταξύ ΔΕ και ΕΔ, και ό,τι ισχύει στη θεωρία για τις ΔΕ μπορούμε να το μεταφέρουμε στις ΕΔ. Μια ΕΔ βαθμού  $n$  μπορεί να γραφεί ως

$$\Delta^n y_n = f(n, y_n, \Delta y_n, \Delta^2 y_n, \dots, \Delta^{n-1} y_n) . \quad (7.52)$$

Μια ακολουθία  $y_m, y_{m+1}$  που ικανοποιεί την (7.52) για κάθε  $n = m, m+1, \dots$  θα τη λέμε *λύση της ΕΔ*. *Βαθμός μιας ΕΔ* είναι η μεγαλύτερη διαφορά μεταξύ των δεικτών της άγνωστης συνάρτησης. Η πιο απλή ΕΔ βαθμού  $m$  είναι της μορφής

$$\alpha_0 y_n + \alpha_1 y_{n+1} + \dots + \alpha_m y_{n+m} = 0 ,$$

όπου τα  $\alpha_0, \alpha_1, \dots, \alpha_m$  είναι ανεξάρτητα του  $n$ . Η παραπάνω σχέση αποτελεί μια *ομογενή γραμμική ΕΔ βαθμού  $n$* . Παρακάτω θα εξετάσουμε τις ΕΔ πρώτου και δεύτερου βαθμού.

### 7.9.1 Εξίσωση διαφορών πρώτου βαθμού

Έστω η ΕΔ

$$\alpha_0 y_{r+1} + \alpha_1 y_r = 0 .$$

Αναζητούμε λύσεις της μορφής  $\beta^r$ . Αντικαθιστώντας στην ΕΔ έχουμε

$$\alpha_0 \beta^{r+1} + \alpha_1 \beta^r = 0 .$$

Η εξίσωση αυτή λέγεται *χαρακτηριστική εξίσωση της ΕΔ* και η λύση της είναι  $\beta = -\frac{\alpha_1}{\alpha_0}$ . Η λύση της ΕΔ επομένως είναι

$$y_r = c \left( -\frac{\alpha_1}{\alpha_0} \right)^r ,$$

όπου το  $c$  είναι μια αυθαίρετη σταθερά που υπολογίζεται από τις αρχικές συνθήκες.

#### Παράδειγμα

$$\begin{aligned} y_{r+1} + y_r &= 0 , \\ y(0) &= 1 . \end{aligned}$$

Η γενική της λύση είναι  $y_r = c(-1)^r$ . Από την αρχική συνθήκη  $c = 1$ , συνεπώς  $y_r = (-1)^r$ .

## 7.9.2 Εξίσωση διαφορών δεύτερου βαθμού

Έστω η ΕΔ

$$\alpha_0 y_{r+2} + \alpha_1 y_{r+1} + \alpha_2 y_r = 0 . \quad (7.53)$$

Αναζητούμε και πάλι λύσεις της μορφής  $\beta^r$ . Η χαρακτηριστική εξίσωση θα είναι

$$\alpha_0 \beta^2 + \alpha_1 \beta + \alpha_2 = 0 .$$

Αν συμβολίσουμε με  $\Delta$  το  $\alpha_1^2 - 4\alpha_0\alpha_2$ , διακρίνουμε τις περιπτώσεις

1.  $\Delta > 0$ . Τότε

$$\beta_{1,2} = \frac{-\alpha_1 \pm \sqrt{\Delta}}{2\alpha_0}$$

και η γενική λύση της (7.53) είναι

$$y_r = c_1 \beta_1^r + c_2 \beta_2^r ,$$

όπου τα  $c_1, c_2$  θα υπολογιστούν από τις αρχικές συνθήκες.

2.  $\Delta = 0$ . Τότε

$$\beta = -\frac{\alpha_1}{2\alpha_0} \quad (\text{διπλή}) .$$

Μια άλλη λύση της ΕΔ θα είναι η  $r\beta^r$ . Πράγματι

$$\begin{aligned} & \alpha_0(r+2)\beta^{r+2} + \alpha_1(r+1)\beta^{r+1} + \alpha_2\beta^r \\ &= \beta^r [r(\alpha_0\beta^2 + \alpha_1\beta + \alpha_2) + \beta(2\alpha_0\beta + \alpha_1)] = 0 , \end{aligned}$$

και η γενική λύση είναι

$$y_r = c_1 \beta_1^r + c_2 r \beta_2^r .$$

3.  $\Delta < 0$ . Τότε

$$\beta_{1,2} = \frac{-\alpha_1 \pm i\sqrt{|\Delta|}}{2\alpha_0} = \varrho e^{\pm i\theta} ,$$

όπου

$$\varrho = \frac{\sqrt{\alpha_1^2 + \Delta^2}}{2|\alpha_0|} , \quad \theta = \arctan \left( \frac{\sqrt{|\Delta|}}{\alpha_1} \right) .$$

Η γενική λύση θα είναι:

$$\begin{aligned} y_r &= c_1 \beta_1^r + c_2 \beta_2^r = c_1 \varrho^r e^{ir\theta} + c_2 \varrho^r e^{-ir\theta} \\ &= \varrho^r [c_1 (\cos(r\theta) + i \sin(r\theta)) + c_2 (\cos(r\theta) - i \sin(r\theta))] \\ &= \varrho (C_1 \cos(r\theta) + C_2 \sin(r\theta)) \end{aligned}$$

όπου οι σταθερές  $C_1 = c_1 + c_2$  και  $C_2 = i(c_1 - c_2)$  υπολογίζονται από τις αρχικές συνθήκες.

Τα παραπάνω μπορούν να γενικευτούν για ΕΔ βαθμού  $n > 2$ . Όταν μια ρίζα του χαρακτηριστικού πολυωνύμου είναι  $p$  βαθμού τότε

$$y_r = (c_1 + c_2 r + c_3 r^2 + \cdots + c_p r^{p-1}) \beta^r .$$

### Παραδείγματα

1. Έστω η εξίσωση διαφορών

$$y_{r+1} = y_r + y_{r-1} , \quad y(0) = 1 , \quad y(1) = 1 .$$

Η λύση είναι:

$$y_r = \frac{1}{\sqrt{5}} \left( \frac{1 + \sqrt{5}}{2} \right)^r - \frac{1}{\sqrt{5}} \left( \frac{1 - \sqrt{5}}{2} \right)^r .$$

2. Έστω η εξίσωση διαφορών

$$y_{r+2} - 4y_{r+1} + 4y_r = 0 .$$

Η γενική λύση είναι:

$$y_r = c_1 2^r + c_2 r 2^r .$$

3. Έστω η εξίσωση διαφορών

$$y_{r+2} - 2y_{r+1} + 10y_r = 0 .$$

Η γενική λύση είναι:

$$y_r = \sqrt{10}^r (c_1 \cos(r\theta) + c_2 \sin(r\theta)) ,$$

όπου

$$\cos \theta = \frac{1}{\sqrt{10}} , \quad \sin \theta = \frac{3}{\sqrt{10}} .$$

### 7.9.3 Μη ομογενείς εξισώσεις διαφορών

Η μη ομογενής ΕΔ

$$\alpha_0 y_n + \alpha_1 y_{n+1} + \cdots + \alpha_m y_{n+m} = b_n ,$$

με  $b_n \neq 0$  λύνεται υπολογίζοντας πρώτα τη γενική λύση της αντίστοιχης ομογενούς

$$\alpha_0 y_n + \alpha_1 y_{n+1} + \cdots + \alpha_m y_{n+m} = 0 .$$

Κατόπιν, βρίσκουμε μια μερική λύση της μη ομογενούς· η γενική λύση της ισούται με το άθροισμα αυτής της μερικής λύσης και της γενικής λύσης της ομογενούς.

**Παράδειγμα**

$$\Delta y_r + 2y_r = r \rightarrow y_{r+1} + y_r = r .$$

Η γενική λύση της ομογενούς είναι  $y_r = c(-1)^r$ . Για να βρούμε μια λύση της μη ομογενούς αναζητούμε λύσεις της μορφής  $y_r = \alpha r + \beta$ , οπότε έχουμε

$$\alpha r + \alpha + \beta + \alpha r + \beta = r ,$$

άρα  $\alpha = 0.5$ ,  $\beta = -0.25$ . Η γενική λύση της μη ομογενούς ΕΔ είναι

$$y_r = c(-1)^r + 0.5r - 0.25 .$$

Αν το δεξί μέλος της μη ομογενούς ΕΔ είναι μια σταθερά, αναζητούμε λύσεις της μορφής  $y_r = c$ .

**7.9.4 Σχόλια**

Παρακάτω δίνουμε ένα παράδειγμα στο οποίο φαίνεται πόσο καταστροφικό μπορεί να είναι το αποτέλεσμα της λύσης μιας ΕΔ όταν υπεισέρχονται λάθη στρωγύλευσης κατά τη λύση της.

**Παράδειγμα**

Έστω η ΕΔ

$$y_{r+1} - 10.1y_r + y_{r-1} = 0 ,$$

της οποίας η γενική λύση είναι

$$y_r = A_1 10^r + A_2 10^{-r} .$$

Παρατηρούμε ότι στη λύση αυτή κυριαρχεί ο όρος  $10^r$ . Όταν υποθέσουμε ότι ισχύουν οι αρχικές συνθήκες  $y_0 = 10$ ,  $y_1 = 1$  τότε  $A_1 = 0$  και  $A_2 = 10$ . Επομένως, η λύση που ζητάμε είναι  $y_r = 10 \times 10^{-r}$ . Στην πράξη, όμως, το  $A_1$  δε θα είναι ακριβώς 0 αλλά θα είναι π.χ. της τάξης του  $10^{-16}$ , δηλαδή, η λύση που θα υπολογίζουμε πραγματικά θα είναι

$$y_r = 10^{-16} \times 10^r + 10 \times 10^{-r} .$$

Έτσι για  $r = 10$ , έχουμε  $y_r = 10^{-6} + 10^{-9}$ . για  $r = 20$ ,  $y_r = 10^4 + 10^{-19}$ . Οι τιμές αυτές είναι πράγματι πολύ «κακές» προσεγγίσεις των πραγματικών τιμών.

**7.10 Αριθμητική Ευστάθεια**

Η ευστάθεια μιας αριθμητικής μεθόδου αναφέρεται στη συμπεριφορά του σφάλματος μεταξύ της τιμής που υπολογίζουμε και της πραγματικής τιμής της λύσης. Αν

το σφάλμα μεγαλώνει όπως ο αριθμός των βημάτων και τελικά κυριαρχεί της λύσης, η μέθοδος που ακολουθείται είναι ασταθής. Ας θεωρήσουμε εδώ τον υπολογισμό των  $y_r$  από την επαναληπτική μέθοδο που ορίζεται από την ΕΔ

$$y_{n+1} = F(y_n, y_{n-1}, \dots, y_{n-m}), \quad n = m, m+1, \dots, \quad (7.54)$$

όταν δίνονται οι αρχικές τιμές  $y_0, y_1, \dots, y_m$ . Έστω ότι για κάποιο αυθαίρετο  $r$  έχουμε ένα σφάλμα στρογγύλευσης  $\varepsilon$ , δηλαδή υπολογίζουμε το

$$y_r^* = y_r + \varepsilon.$$

Θα έχουμε, τότε, την ακολουθία  $y_0, y_1, \dots, y_{r-1}, y_r^*, y_{r+1}^*, \dots$

Θα λέμε ότι η επαναληπτική μέθοδος (7.54) είναι *αριθμητικά ευσταθής* αν, για  $y_n \neq 0$ , τα σχετικά σφάλματα

$$\left| \frac{y_n - y_n^*}{y_n} \right|$$

είναι φραγμένα όταν  $n \rightarrow \infty$ .

Έστω ότι θέλουμε να υπολογίσουμε τα  $y_n$  από την

$$y_{n+1} = 100.01y_n - y_{n-1},$$

με  $y_0 = 1$  και  $y_1 = 0.01$ . Η γενική λύση της είναι

$$y_n = c_1 100^n + c_2 0.01^n,$$

και η λύση που ικανοποιεί τις αρχικές συνθήκες είναι

$$y_n = 0.01^n, \quad (c_1 = 0, c_2 = 1).$$

Η εισαγωγή λαθών στρογγύλευσης κατά τη διάρκεια των υπολογισμών έχει ως αποτέλεσμα να αλλάξει τη λύση που θέλουμε να υπολογίσουμε. Δηλαδή, το  $c_1$  δε θα είναι ακριβώς 0· επομένως, θα πρέπει να προσθέσουμε τον όρο  $c_1 100^n$  στο  $y_n$ . Αυτό έχει ως αποτέλεσμα το  $c_1 100^n$  πολύ γρήγορα να κυριαρχεί της πραγματικής λύσης (που είναι  $c_2 0.01^n$ ). Έτσι, π.χ., αν  $y_0^* = y_0 + 10^{-6} = 1.000001$  και  $y_1^* = y_1 = 0.01$  τότε

$$\begin{aligned} y_2^* &\approx 0.0000990214, \\ y_3^* &\approx -0.0000968733, \\ y_4^* &\approx -0.0097873247, \\ y_5^* &\approx -0.9787334946, \\ y_6^* &\approx -97.8733515650. \end{aligned}$$

Προφανώς, σε μια προσέγγιση ενός προβλήματος αρχικών τιμών, η συνάρτηση  $F$  στην (7.54) θα εξαρτάται από το  $h$  και την  $f(x, y)$ . Για τον ορισμό της αριθμητικής ευστάθειας δεν ελαττώνουμε το  $h$  όταν  $n \rightarrow \infty$  όπως στην περίπτωση της σειράς

Taylor. Καθώς μιας προσέγγιση της ΔΕ από μία ΕΔ εξαρτάται από τη μορφή της  $f(x, y)$ , είναι πολύ δύσκολο να εξάγουμε γενικά συμπεράσματα όσον αφορά την αριθμητική ευστάθεια. Παρακάτω θα εξετάσουμε λεπτομερώς τη ΔΕ  $y' = -\lambda y$  με  $y(0) = 1$ ,  $\lambda > 0$ , της οποίας η λύση είναι η  $y(x) = e^{-\lambda x}$ . Ο τρόπος διερεύνησης αυτής της ΔΕ και τα αποτελέσματα του μπορούν να εφαρμοστούν και σε άλλα προβλήματα αρχικών τιμών.

Για τη λύση της ΔΕ θα χρησιμοποιήσουμε τη μέθοδο που δίνεται από τον τύπο

$$y_{n+1} = y_n + 2hf_n. \quad (7.55)$$

Καθώς  $f_n = -\lambda y_n$  η παραπάνω σχέση γίνεται

$$y_{n+1} + 2\lambda h y_n - y_{n-1} = 0,$$

της οποίας η γενική λύση είναι

$$y_n = c_1 \beta_1^n + c_2 \beta_2^n,$$

με  $\beta_{1,2} = -\lambda h \pm \sqrt{1 + \lambda^2 h^2}$ .

Από την παραπάνω σχέση βλέπουμε ότι η  $\beta_1^n$  συμπεριφέρεται σαν την  $y(x)$ , όταν  $h \rightarrow 0$  με  $x_n$  σταθερό, ενώ η  $\beta_2^n$  έχει τελείως διαφορετική συμπεριφορά. Πράγματι, όταν αναπτύξουμε την τετραγωνική ρίζα έχουμε

$$\beta_1 = -\lambda h + 1 + \frac{1}{2}(\lambda h)^2 + O(h^4).$$

Συγκρίνοντας με την

$$e^{-\lambda h} = 1 - \lambda h + \frac{1}{2}(\lambda h)^2 + O(h^3),$$

βλέπουμε ότι

$$\beta_1 = e^{-\lambda h} + O(h^3) = e^{-\lambda h} [1 + O(h^3)].$$

Η τελευταία ισότητα προκύπτει από τη σχέση  $e^{-\lambda h} = 1 + O(h)$ . Άρα

$$\beta_1^n = e^{-\lambda n h} [1 + O(h^3)]^n = e^{-\lambda x_n} [1 + nO(h^3)] = e^{-\lambda x_n} [1 + O(h^2)],$$

και  $x_n = nh$ . Όταν  $h \rightarrow 0$  με το  $x_n$  σταθερό τότε

$$\beta_1^n = e^{-\lambda x_n} = y(x_n),$$

δηλαδή, για μικρές τιμές του  $h$  η  $\beta_1^n$  συμπεριφέρεται σαν μια καλή προσέγγιση της  $y(x)$ . Από την άλλη πλευρά, όμως,

$$\begin{aligned} \beta_2 &= -\lambda h - \sqrt{1 + \lambda^2 h^2} = -\lambda h - 1 - \frac{1}{2}(\lambda h)^2 + O(h^4) \\ &= -e^{\lambda h} + O(h^3) = -e^{\lambda h} [1 + O(h^3)] \end{aligned}$$

Άρα

$$\beta_2^n = (-1)^n e^{\lambda x_n} [1 + O(h^2)] .$$

Όταν  $h \rightarrow 0$  με το  $x_n$  σταθερό η  $\beta_2^n$  δεν προσεγγίζει το  $y(x)$ . Γενικά, η λύση της (7.55) είναι

$$y_n = c_1 e^{-\lambda x_n} [1 + O(h^2)] + c_2 (-1)^n e^{\lambda x_n} [1 + O(h^2)] . \quad (7.56)$$

Για να είναι αυτή η τιμή του  $y_n$  καλή προσέγγιση του  $y(x_n)$  θα πρέπει  $c_1 = 1$ ,  $c_2 = 0$ , δηλαδή, πρέπει να υπολογίσουμε τη λύση  $y_n = \beta_1^n$ . Στην πράξη, όμως, επειδή δεν μπορούμε να αποφύγουμε τα λάθη στρογγύλευσης, υπολογίζουμε μία τιμή  $y_n^*$  σαν προσέγγιση του  $y_n$

$$y_n^* = (1 + \delta_1) \beta_1^n + \delta_2 \beta_2^n ,$$

οπότε το σχετικό σφάλμα είναι

$$\left| \frac{y_n - y_n^*}{y_n} \right| = \left| \frac{\delta_1 \beta_1^n + \delta_2 \beta_2^n}{\beta_1^n} \right| = \left| \delta_1 + \delta_2 \left( \frac{\beta_2}{\beta_1} \right)^n \right| .$$

Όταν  $|\beta_2| < |\beta_1|$  το σχετικό σφάλμα είναι φραγμένο και συνεπώς, η αντίστοιχη επαναληπτική μέθοδος είναι ευσταθής. Στην αντίθετη περίπτωση η μέθοδος είναι ασταθής. Αν για το παραπάνω παράδειγμα έχουμε

$$|\beta_2| \equiv \left| \lambda h + \sqrt{1 + \lambda^2 h^2} \right| > \left| \lambda h - \sqrt{1 + \lambda^2 h^2} \right| \equiv |\beta_1| ,$$

για  $h$  σταθερό το  $\left( \frac{\beta_2}{\beta_1} \right)^n$  (και το σχετικό σφάλμα) δεν είναι φραγμένο όταν  $n \rightarrow \infty$ .

Η  $\beta_2^n$  λέγεται παρασιτική ή απορριπτέα λύση της ΕΔ και εμφανίζεται επειδή έχουμε αντικαταστήσει μια ΔΕ πρώτου βαθμού με μια ΕΔ δεύτερου βαθμού. Στην περίπτωση αυτή, η παρασιτική λύση υπερισχύει της ζητούμενης λύσης.

### Παραδείγματα

1. Έστω το πρόβλημα αρχικών τιμών  $y' = -3y$ ,  $y(0) = 1$ , το οποίο έχει λύση την  $y(x) = e^{-3x}$ . Αν για την επίλυση αυτής της ΔΕ χρησιμοποιήσουμε τον τύπο (7.55) με  $h = 0.1$  και  $y_0 = y(0) = 1$ ,  $y_1 = 0.7408$  έχουμε τις τιμές του Πίνακα 7.2. Όπως παρατηρούμε, η μέθοδος είναι ασταθής για  $x_n > 1$  και οι εναλλαγές του προσήμου οφείλονται στον όρο  $(-1)^n$  της (7.56).
2. Θα εξετάσουμε ως προς την ευστάθεια τη μέθοδο Adams–Bashforth δεύτερου βαθμού

$$y_{n+1} = y_n + \frac{h}{2}(3f_n - f_{n-1}) ,$$

όταν εφαρμοστεί για την επίλυση του προβλήματος αρχικών τιμών  $y' = -\lambda y$ ,  $y(0) = 1$ ,  $\lambda > 0$ . Ο τύπος Adams–Bashforth γράφεται

$$y_{n+1} - \left( 1 - \frac{3}{2}\lambda h \right) y_n + \frac{1}{2}\lambda h y_{n-1} = 0 . \quad (7.57)$$



$x_n$	$y_n^*$	$y_n$
0.0	1.0	1.0
0.1	0.7408	0.740818
0.2	0.555520	0.548812
0.3	0.407488	0.406570
0.4	0.311027	0.301194
0.5	0.220872	0.223130
0.6	0.178504	0.165299
0.7	0.113769	0.122456
0.8	0.110243	0.090718
0.9	0.047624	0.067206
1.0	0.081669	0.049787
1.1	-0.001378	0.036883
1.2	0.082495	0.027324
1.3	-0.050875	0.020242
1.4	0.113020	0.014996
1.5	-0.118687	0.011109

Πίνακας 7.2: Επίλυση της ΔΕ  $y' = -3y$ ,  $y(0) = 1$ , με εφαρμογή της (7.55) με  $h = 0.1$

Οι ρίζες του χαρακτηριστικού πολωνύμου του είναι

$$\beta_{1,2} = \frac{1}{2} \left[ \left( 1 - \frac{3}{2}\lambda h \right) \pm \sqrt{1 - \lambda h + \frac{9}{4}\lambda^2 h^2} \right] .$$

Έχουμε

$$\begin{aligned} \beta_1 &= \frac{1}{2} \left[ \left( 1 - \frac{3}{2}\lambda h \right) + 1 + \frac{1}{2} \left( -\lambda h + \frac{9}{4}\lambda^2 h^2 \right) - \frac{1}{8} \left( -\lambda h + \frac{9}{4}\lambda^2 h^2 \right) + O(h^3) \right] \\ &= 1 - \lambda h + \frac{1}{2}\lambda^2 h^2 + O(h^3) = e^{-\lambda h} [1 + O(h^3)] . \end{aligned}$$

Επομένως,

$$\beta_1^n = e^{-\lambda x_n} [1 + O(h^2)] .$$

Από τη σχέση αυτή έχουμε ότι για μικρό  $h$  η  $\beta_1^n$  είναι μια καλή προσέγγιση του  $y(x_n)$ .

Στην πράξη, η λύση της (7.57) είναι της μορφής

$$y_n^* = (1 + \delta_1)\beta_1^n + \delta_2\beta_2^n$$

και

$$\frac{\beta_2}{\beta_1} = \frac{1 - \frac{3}{2}\lambda h - \sqrt{1 - \lambda h + \frac{9}{4}\lambda^2 h^2}}{1 - \frac{3}{2}\lambda h + \sqrt{1 - \lambda h + \frac{9}{4}\lambda^2 h^2}} .$$

Η ποσότητα κάτω από την τετραγωνική ρίζα είναι θετική και συνεπώς για  $1 - 3/2\lambda h \geq 0$  θα ισχύει  $\left| \frac{\beta_2}{\beta_1} \right| \leq 1$ . Το σχετικό σφάλμα θα είναι φραγμένο—και επομένως η μέθοδος είναι ευσταθής—όταν  $h \leq \frac{2}{3\lambda}$ .

Δυσκολίες ευστάθειας δεν υπάρχουν όταν για μια ΔΕ πρώτου βαθμού χρησιμοποιούμε μια ΕΔ πρώτου βαθμού καθώς δεν προκύπτει παρασιτική λύση. Υπάρχει, όμως, ένα άλλου τύπου πρόβλημα ευστάθειας που μπορεί να παρουσιαστεί για ΕΔ πρώτου βαθμού.

## 7.11 Απόλυτη Ευστάθεια

Έστω ότι μια ΔΕ αντικαθίσταται από μια ΕΔ βαθμού  $k$ . Τότε η γενική λύση της ΕΔ είναι

$$y_n = c_1\beta_1^n + \cdots + c_k\beta_k^n.$$

Αν μια λύση της ΕΔ, έστω η  $\beta_1^n$  αντιπροσωπεύει την πραγματική λύση της ΔΕ, τότε οι άλλες  $k - 1$  ρίζες της ΕΔ θα είναι παρασιτικές. Αν  $|\beta_i| > |\beta_1|$  για  $i \neq 1$ , τότε η παρασιτική λύση  $\beta_i^n$  θα κυριαρχεί στη λύση της ΔΕ και θα έχουμε αστάθεια. Για τη λύση της ΔΕ  $y' = \lambda y$ , η ΕΔ είναι

**Απόλυτα ευσταθής** όταν  $|\beta_i| < 1$  για όλα τα  $i$ ,

**Σχετικά ευσταθής** όταν  $|\beta_1| > |\beta_i|$  για  $i = 2, \dots, k$ .

### Παράδειγμα

Έστω  $y' = -\lambda y$ ,  $\lambda > 0$  και ενδιαφερόμαστε για την απόλυτη ευστάθεια της μεθόδου Euler όταν εφαρμοστεί για τη λύση αυτής της ΔΕ. Ο τύπος Euler είναι

$$y_{r+1} = (1 - \lambda h)y_r = \beta y_r, \quad \text{με } \beta = 1 - \lambda h.$$

Για να έχουμε απόλυτη ευστάθεια πρέπει να έχουμε  $|\beta| < 1$  ή  $h < 2/\lambda$ . Η συνθήκη αυτή όταν το  $\lambda$  είναι πολύ μεγάλο είναι πολύ δαπανηρή. Π.χ. αν  $y' = -1000y$ ,  $y(0) = 1$ · για απόλυτη ευστάθεια θα πρέπει να έχουμε  $h < 2/1000 = 0.002$ . Για να ολοκληρώσουμε από  $x = 0$  έως  $x = 100$  χρειαζόμαστε τουλάχιστον 50000 βήματα.

**Παρατήρηση:** Οι ανοικτές (explicit) μέθοδοι δεν είναι ποτέ απόλυτα ευσταθείς για όλες τις τιμές του  $h$ . Εξαίρεση αποτελεί ο κανόνας του τραπεζίου. Όταν τον εφαρμόσουμε στην  $y' = -\lambda y$ ,  $y(0) = 1$ ,  $\lambda > 0$ , βλέπουμε ότι είναι απόλυτα ευσταθής για όλα τα  $h$ .

## 7.12 Ασκήσεις

1. Εφαρμόστε τη μέθοδο Euler για την επίλυση της ΔΕ

$$y' = \cos x - \sin y + x^2$$

στο διάστημα  $[-1, 1]$ , με  $y(-1) = 3.0$ . Τυπώστε τη λύση ανά  $h = 0.01$ .

2. Εφαρμόστε τη μέθοδο Taylor με 5 όρους για την επίλυση της ΔΕ της προηγούμενης άσκησης.
3. Να λυθεί το σύστημα

$$\begin{aligned} y_1' &= 2y_1 - 2y_2 + 3y_3 \\ y_2' &= y_1 + y_2 + y_3 \\ y_3' &= y_1 + 3y_2 - y_3 \end{aligned}$$

με αρχικές συνθήκες (στο  $t = 0$ )  $y_1 = -2$ ,  $y_2 = 30$ ,  $y_3 = 0$ . Δίνεται ότι

$$\begin{bmatrix} 1 & 11 & 1 \\ -1 & 1 & 1 \\ -1 & -14 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1/2 & -5/6 & 1/3 \\ 0 & 1/15 & -1/15 \\ 1/2 & 1/10 & 2/5 \end{bmatrix}.$$

4. Να γράψετε κώδικα για την επίλυση της ΔΕ  $y' = -y$ , με  $y(0) = 1$  στο διάστημα  $[0, 1]$  με τη μέθοδο Euler και βήματα  $h = 0.2, 0.02, 0.002, 0.0002, 0.00002$ .
5. Να δείξετε ότι η μέθοδος Runge–Kutta 3<sup>ου</sup> βαθμού για την επίλυση της ΔΕ  $y' = f(x, y)$ ,  $y(x_0) = y_0$  είναι:

$$\begin{aligned} y_{r+1} &= y_r + \frac{1}{6}(k_1 + 4k_2 + k_3) \\ k_1 &= hf(x_r, y_r) \\ k_2 &= hf(x_r + h/2, y_r + k_1/2) \\ k_3 &= hf(x_r + h, y_r - k_1 + 2k_2) \end{aligned}$$

6. Εφαρμόστε τη μέθοδο Taylor με 4 όρους για την επίλυση του συστήματος ΔΕ

$$\begin{aligned} y' &= y + z^2 - x^3 \\ z' &= z + y^3 + \cos x \end{aligned}$$

με αρχικές συνθήκες (στο  $x = 0$ )  $y = 0.3$ ,  $z = 0.1$ . Τυπώστε τις τιμές των  $y, z$  στο διάστημα  $[0, 1]$  με βήμα 0.1.

7. Να γράψετε κώδικα σε Fortran που να υλοποιεί τη μέθοδο Runge–Kutta 2<sup>ου</sup> βαθμού.

8. Δώστε έναν αλγόριθμο για τον έλεγχο του σφάλματος για τη Runge-Kutta βαθμού  $p$ . Διπλασιάστε το βήμα  $h$  αν το σφάλμα είναι «πολύ μικρό» ή υποδιπλασιάστε το αν το σφάλμα είναι «πολύ μεγάλο».
9. Δείξτε ότι όταν εφαρμοστεί η Runge-Kutta 2<sup>ου</sup> βαθμού σε ΔΕ της μορφής  $y' = \lambda x$  δίνει την ακριβή λύση.
10. Δώστε ένα αλγόριθμο για τη μέθοδο Taylor τέταρτου βαθμού και υλοποιήστε τον σε κώδικα Fortran. Εφαρμόστε τον για τη ΔΕ  $y' = y^3 + x + y$ , στο  $[0, 0.5]$ , με  $y(0) = 1$ ,  $h = 0.1$ .
11. Να λυθεί η ΔΕ  $y' = x^2 + x - y$ ,  $y(0) = 0$  με τη Runge-Kutta 2<sup>ου</sup> βαθμού. Υπολογίστε τη λύση στο  $x = 0.6$  με  $h = 0.02$ .
12. Το πρόβλημα αρχικών τιμών  $y' = ax + b$ ,  $y(0) = 0$ , έχει λύση

$$y(x) = \frac{a}{2}x^2 + bx .$$

Αν εφαρμοστεί η μέθοδος Euler δείξτε ότι η ΕΔ που προκύπτει έχει λύση

$$y_n = \frac{1}{2}(ax_n + 2b - ah)x_n ,$$

όπου  $x_r = rh$ , και, επομένως,

$$y(x_n) - y_n = \frac{a}{2}hx_n .$$

13. Να εφαρμόσετε την Runge-Kutta 2<sup>ου</sup> βαθμού για δύο διαδοχικά διαστήματα  $h$ , για τη ΔΕ  $y' = \lambda y$ ,  $y(0) = 1$ . Να δείξετε ότι ο κυρίαρχος όρος του σφάλματος αποκοπής στο  $y(2h)$  είναι  $(\lambda h)^3/3$ .
14. Χρησιμοποιήστε τη μέθοδο Runge-Kutta 4<sup>ου</sup> βαθμού για να επιλύσετε τη ΔΕ

$$y' = \frac{y}{x} \left( 1 - \frac{y}{x} \right)$$

στο διάστημα  $[1, 3]$ , με  $y(1) = 2$ . Τυπώστε τις τιμές με βήμα  $h = 1/128$ , καθώς και το σφάλμα ως προς την ακριβή λύση

$$y(x) = \frac{x}{0.5 + \ln x} .$$

15. Να εφαρμόσετε την Runge-Kutta 2<sup>ου</sup> βαθμού για την εύρεση της κίνησης σώματος μάζας  $m = 2$  kg, εξαρτώμενου από ελατήριο με δύναμη επαναφοράς  $F(x) = x^2 + 0.01x^3$ . Το σώμα αφήνεται για  $t = 0$  ελεύθερο, χωρίς αρχική ταχύτητα, στη θέση  $x = 2.5$  cm.

16. Να βρείτε την κίνηση εκκρεμούς για το οποίο ισχύει

$$\ddot{\theta} = -\sin \theta ,$$

όπου  $\theta$  η γωνία απομάκρυνσης από την κάθετο. Το εκκρεμές αφήνεται ελεύθερο, χωρίς αρχική ταχύτητα σε γωνία  $\theta = 45^\circ$ .

Για μικρές γωνίες  $\theta$  ισχύει  $\sin \theta \approx \theta$ . Εφαρμόστε την προσέγγιση αυτή και συγκρίνετε τη λύση της νέας διαφορικής εξίσωσης με τη λύση της ακριβούς ΔΕ.

17. Να λύσετε τη ΔΕ  $\psi'' = (x^2 - 5)\psi$  με αρχική συνθήκη  $\psi(0) = -(2\sqrt{\pi})^{-1/2}$ . Τυπώστε 100 ισαπέχουσες τιμές στο διάστημα  $[-2, 2]$ .

**Υπόδειξη** Να λύσετε δύο προβλήματα αρχικών τιμών, τη ΔΕ στα διαστήματα  $[0, 2]$  και  $[-2, 0]$ .

18. Να λυθούν οι ΕΔ

- $y_{n+1} - ay_n = 0$ ,
- $y_{n+2} - 4y_{n+1} + 4y_n = 0$ ,
- $y_{n+2} - 4y_{n+1} + 4y_n = 1$ .

19. Να δειχθούν οι σχέσεις:

- $\Delta = E - 1$ ,
- $E = e^{hD}$ ,
- $\nabla = \frac{E - 1}{E}$ ,
- $\delta = E^{1/2} - E^{-1/2}$ ,
- $\mu = 0.5 (E^{1/2} + E^{-1/2})$ ,
- $\mu = \cosh\left(\frac{hD}{2}\right)$ ,
- $E = 1 + \mu\delta + \frac{\delta^2}{2}$ ,
- $\mu^2 = 1 + \frac{\delta^2}{4}$ .

20. Να λυθούν οι ΕΔ

- $y_{r+2} - 5y_{r+1} + 6y_r = 0, \quad y_0 = 0, y_1 = 1,$
- $y_{r+2} - 4y_{r+1} + 4y_r = 0, \quad y_0 = 1, y_1 = 6,$
- $y_{r+2} + 6y_{r+1} + 25y_r = 0, \quad y_0 = 0, y_1 = 4.$

21. Ναδειχτεί ότι η γενική λύση της ΕΔ

$$y_{r+2} - (2 + h^2) y_{r+1} + y_r = h^2, \quad h > 0,$$

είναι

$$y_n = c_1 \left[ 1 + h + \frac{h^2}{2} + O(h^3) \right]^n + c_2 \left[ 1 - h + \frac{h^2}{2} + O(h^3) \right]^n - 1.$$

22. Ναδειχτεί ότι η γενική λύση της ΕΔ

$$y_{r+2} + 4hy_{r+1} - y_r = 2h, \quad h > 0,$$

είναι

$$y_n = c_1 [1 - 2h + O(h^2)]^n + c_2 (-1)^n [1 + 2h + O(h^2)]^n + \frac{1}{2}.$$

23. Ναδειχτεί ότι

$$\frac{h}{2} (y'_r + y'_{r+1}) = \left( \delta + \frac{1}{12} \delta^3 - \frac{1}{120} \delta^5 + \dots \right) y_{r+\frac{1}{2}}.$$

Δίνεται ότι

$$\sinh^{-1} z = z - \frac{1}{6} z^3 + \frac{3}{4} z^5 + \dots$$

24. Να αποδείξετε ότι το

$$y_0 - y_1 + y_2 - \dots = \sum_{r=0}^{\infty} (-1)^r E^r y_0$$

είναι ίσο με

$$\frac{1}{2} y_0 - \frac{1}{4} \Delta y_0 + \frac{1}{8} \Delta^2 y_0 - \dots$$

25. Να υλοποιήσετε σε κώδικα τον αλγόριθμο Adams–Bashforth 3<sup>ου</sup> βαθμού και το αντίστοιχο σφάλμα αποκοπής.
26. Να υπολογίσετε μια προσέγγιση του  $y(1.0)$  για τη ΔΕ  $y' = 1 - y$ ,  $y(0) = 0$ , χρησιμοποιώντας τον αλγόριθμο Adams–Bashforth 2<sup>ου</sup> βαθμού με  $h = 0.2$  και αρχικές τιμές  $y(0) = 0$ ,  $y_1 = y(0.2) = 0.18127$ . Συγκρίνετε τα αποτελέσματα με την αναλυτική λύση  $y(x) = 1 - e^{-x}$ .
27. Υλοποιήστε σε κώδικα Fortran τον αλγόριθμο Πρόβλεψης–Διόρθωσης στον οποίο το  $h$  να μεταβάλεται αυτόματα υπολογίζοντας κάθε φορά το σφάλμα αποκοπής  $T_c$ .
28. Να εξετάσετε ως προς την ευστάθεια τις παρακάτω μεθόδους για τη ΔΕ  $y' = -\lambda y$ ,  $\lambda > 0$ .

(α') Adams–Moulton 3<sup>ου</sup> βαθμού

$$y_{n+1} = y_n + \frac{h}{12} (5y'_{n+1} + 8y'_n - y'_{n-1}) ,$$

(β') Milne–Simpson

$$y_{n+1} = y_{n-1} + \frac{h}{3} (y'_{n+1} + 4y'_n + y'_{n-1}) ,$$

(γ')

$$y_{n+1} = y_{n-1} + \frac{h}{2} (y'_n + 3y'_{n-1}) .$$

Να δείξετε ότι το απόλυτο σφάλμα για αρκετά μικρό  $h$  συνεχώς μικραίνει παρόλο που η συγκεκριμένη μέθοδος είναι ασταθής.

29. Έστω η ΔΕ  $y' = -\lambda y$ ,  $\lambda > 0$ . Να δείξετε ότι  $y_n \rightarrow 0$  όταν  $n \rightarrow \infty$  και  $h$  σταθερό (δηλαδή είναι απόλυτα ευσταθής), αν

- $|1 - \lambda h + \frac{1}{2}\lambda^2 h^2| < 1$  για τη Runge–Kutta 2<sup>ου</sup> βαθμού,
- $h < 1/\lambda$  για τη Adams–Bashforth 2<sup>ου</sup> βαθμού.

30. Να δείξετε ότι η μέθοδος

$$y_{n+1} = y_{n-1} + \frac{h}{2} (y'_{n+1} + 2y'_n + y'_{n-1})$$

είναι ασταθής για τη ΔΕ  $y' = -\lambda y$  και ευσταθής για την  $y' = -\lambda y + \mu$ , όπου  $\lambda > 0$  και  $\mu \neq 0$ .

31. Να περιγράψετε τις ΕΔ για τη μέθοδο Taylor 2<sup>ου</sup> βαθμού, όταν εφαρμοστεί στο σύστημα

$$\begin{aligned} y'_1 &= x^2 y_1 - y_2 , & y_1(0) &= s_1 , \\ y'_2 &= -y_1 + x y_2 , & y_2(0) &= s_2 . \end{aligned}$$

32. Να λυθεί η ΔΕ

$$y'' + 4xyy' + 2y^2 = 0 , \quad y(0) = 1 , \quad y'(0) = 0 , \quad (7.58)$$

με τη μέθοδο Euler και με  $h = 0.1$ . Να υπολογιστούν τα  $y(0.5)$ ,  $y'(0.5)$ .

33. Να λυθεί η (7.58) με μια μέθοδο πρόβλεψης-διόρθωσης 2<sup>ου</sup> βαθμού. Να γίνει περιγραφή των ΕΔ και του αλγορίθμου μόνο.

34. Έστω η ΔΕ  $y' - 2y$ ,  $y(a) = y_0$ . Να εφαρμοστεί η μέθοδος Euler και να βρεθεί η έκφραση του  $y_k$  ως συνάρτηση του  $y_0$ . Να εξεταστεί η συμπεριφορά της μεθόδου όταν  $h \rightarrow 0$ .

35. Να εφαρμοστεί δευτεροβάθμια μέθοδος πρόβλεψης-διόρθωσης για τη ΔΕ  $y' = xy^2$ , στο  $[0, 0.2]$ , με  $y(0) = 0.25$  και  $h = 0.1$ .
36. Να βρείτε όρια του  $λh$  για μια ευσταθή ολοκλήρωση της ΔΕ  $y' = λy$ ,  $y(0) = 1$  με τη μέθοδο Runge-Kutta 3<sup>ου</sup> βαθμού.
37. Να δείξετε ότι το σφάλμα αποκοπής στη μέθοδο τραπεζίου είναι

$$-\frac{h^3}{12}y'''(\xi), \quad \xi \in (x_r, x_{r+1}).$$

38. Να διερευνήσετε την ευστάθεια της Adams-Moulton 3<sup>ου</sup> βαθμού

$$y_{r+1} = y_r + \frac{h}{12} (5y'_{r+1} + 8y'_r - y'_{r-1}).$$

39. Να δείξετε ότι η προσέγγιση της παραγώγου μιας συνάρτησης από τον τύπο των κεντρικών διαφορών είναι καλύτερη από την προσέγγιση με τον τύπο διαφορών προς τα εμπρός.
40. Να εξεταστεί ως προς την ευστάθεια ο κανόνας τραπεζίου για τη ΔΕ  $y' = λy$ .



## Παράρτημα α΄

# Χρήσιμα Ολοκληρώματα

Γνωρίζουμε ότι ισχύουν οι ακόλουθες τριγωνομετρικές σχέσεις:

$$\begin{aligned}\sin x \sin y &= \frac{\cos(x-y) - \cos(x+y)}{2}, \\ \cos x \cos y &= \frac{\cos(x-y) + \cos(x+y)}{2}, \\ \sin x \cos y &= \frac{\sin(x-y) + \sin(x+y)}{2}.\end{aligned}$$

Επίσης,

$$\int \sin x \, dx = -\cos x + c, \quad \int \cos x \, dx = \sin x + c.$$

Χρησιμοποιώντας αυτές, εύκολα μπορούν ναδειχθούν οι παρακάτω σχέσεις (για ακέραιο  $n \geq 0$ ). Αποδείξτε τις!

$$\int_0^L \sin\left(\frac{2n\pi x}{L}\right) dx = 0, \quad (\alpha'.1)$$

$$\int_0^L \cos\left(\frac{2n\pi x}{L}\right) dx = \begin{cases} L, & n=0 \\ 0, & n>0 \end{cases}, \quad (\alpha'.2)$$

$$\int_0^L \sin\left(\frac{2n\pi x}{L}\right) \cos\left(\frac{2k\pi x}{L}\right) dx = 0, \quad (\alpha'.3)$$

$$\int_0^L \cos\left(\frac{2n\pi x}{L}\right) \cos\left(\frac{2k\pi x}{L}\right) dx = \begin{cases} 0, & n \neq k \\ L, & n = k = 0 \\ L/2, & n = k > 0 \end{cases}, \quad (\alpha'.4)$$

$$\int_0^L \sin\left(\frac{2n\pi x}{L}\right) \sin\left(\frac{2k\pi x}{L}\right) dx = \begin{cases} 0, & n \neq k \\ 0, & n = k = 0 \\ L/2, & n = k > 0 \end{cases}. \quad (\alpha'.5)$$



# Κατάλογος πινάκων

2.1	Ακολουθίες των διαστημάτων, της προσεγγιστικής ρίζας και της αντίστοιχης τιμής της $f(x) = x^3 + 4x^2 - 10$ κατά την εφαρμογή της μεθόδου διχοτόμησης . . . . .	13
7.1	Πίνακας διαφορών της συνάρτησης $f(x) = e^x$ με ακρίβεια 6 δεκαδικών ψηφίων . . . . .	151
7.2	Επίλυση της ΔΕ $y' = -3y$ , $y(0) = 1$ , με εφαρμογή της (7.55) με $h = 0.1$	169



# Ευρετήριο

spline, **63**

Γραμμικό σύστημα εξισώσεων, **29**

Μέθοδος απαλοιφής Gauss, **34**

Μέθοδος απαλοιφής Gauss–Jordan, **41**

Μέθοδος επίλυσης Cramer, **33**

Μέθοδος επίλυσης Gauss–Seidel, **46**

Μέθοδος επίλυσης Jacobi, **45**

Μέθοδος επίλυσης Successive overrelaxation (SOR), **46**

ευστάθεια, **29**

Θεώρημα

Bolzano, **11**

Perron–Frobenius, **52**

Plancherel, **122**

Rolle, **11**

Taylor, **11**

Wiener–Khinchin, **131**

Ενδιάμεσης τιμής, **11**

Μέσης τιμής, **11**

κύκλων του Gershgorin, **51**

συνέλιξης, **130**

Μέθοδοι εύρεσης ρίζας

Müller, **17**

Διχοτόμηση, **11**

ακρίβεια, **13**

σύγκλιση, **14**

Ευστάθεια, **10**

Ταχύτητα σύγκλισης, **10**

Ψευδούς σημείου, **15**

αλγόριθμος Illinois, **16**

σταθερού σημείου, **19**

σύγκλιση, **20**

τέμνουσα, **16**

σύγκλιση, **17**

τύποι Householder, **22**

Halley, **25**

Newton–Raphson, **22**

Μέθοδος ελάχιστων τετραγώνων, **67**

Πίνακας

Ιδιοδιάνυσμα, **31**

Ιδιοτιμή, **31**

Ορίζουσα, **31**

Υπολογισμός, **49**

ανάλυση Cholesky, **32**

ανάλυση LU, **41**

αλγόριθμος Crout, **42**

ανάστροφος, **31**

δείκτης κατάστασης, **30**

θετικά ορισμένος, **31**

κριτήριο του Sylvester, **32**

συμμετρικός, **31**

έψιλον της μηχανής, **5**

αλγόριθμος

Adams–Bashforth, **155**

Adams–Moulton, **156**

Crout, **42**

Gauss–Seidel, **46**

Golub–Welsch, **90**

Jacobi, **45**

Successive overrelaxation (SOR), **46**

πρόβλεψης–διόρθωσης, **157**

αποσυνέλιξη, **130**

αυτοσυσχέτιση, **132**

διαφορική εξίσωση, **135**

γραμμική δευτέρου βαθμού με

σταθερούς συντελεστές, **138**

μη ομογενής, **138**

γραμμική πρώτου βαθμού, **137**

λύση, **136**

μέθοδος Euler, **142, 155**

μέθοδος Runge–Kutta, **144, 155**

ομογενής, **137**

ορισμός, **135**

πρόβλημα αρχικών τιμών, **136**

- χαρακτηριστική εξίσωση, **138**
- εξίσωση διαφορών
  - βαθμός, **162**
  - γραμμική ομογενής, **162**
  - λύση, **162**
  - μέθοδος επίλυσης
    - απόλυτα ευσταθής, **170**
    - αριθμητικά ευσταθής, **166**
    - σχετικά ευσταθής, **170**
  - ορισμός, **161**
  - χαρακτηριστική εξίσωση, **162**
- ετεροσυσχέτιση, **130**
- κανόνας ολοκλήρωσης Clenshaw–Curtis, **90**
- κανόνες ολοκλήρωσης Gauss, **84**
- κυκλικός δίσκος Gershgorin, **51**
- μέθοδος ολοκλήρωσης
  - Simpson, **79, 152**
  - τραπεζίου, **76, 153**
- οπισθοδρόμηση, **37**
- σημαντικά ψηφία, **3**
- σταθερό σημείο συνάρτησης, **19**
  - κριτήριο ύπαρξης, **19**
  - μοναδικότητα, **20**
- συνάρτηση
  - περιοδική, **98**
  - συνεχής, **97**
- συνάρτηση βήματος, **125**
- συνέλιξη, **128**
- συνθήκες Dirichlet, **98**
- συνθήκη Lipschitz, **136, 161**
- σφάλμα
  - αποκοπής, **143**
  - ολικό, **143**
  - τοπικό, **143**
- στρογγύλευσης, **143**
- σύγκλιση
  - τετραγωνική, **24**
- ταυτότητα
  - Parseval, **108**
- τελεστής
  - διαφορικός, **149**
  - μέσης τιμής, **150**
  - μετατόπισης, **149**
- τελεστής διαφοράς
  - κεντρικής, **150**
  - προς τα εμπρός, **148**
  - προς τα πίσω, **150**
- τύποι Vieta, **59**
- τύπος
  - ανοικτός, **155**
  - διόρθωσης, **156**
  - πρόβλεψης, **155**
- υπεκχείλιση, **6**
- υπερχείλιση, **6**
- φαινόμενο
  - Gibbs, **105**
- φασματική πυκνότητα ενέργειας, **132**