

Υπολογιστικά Μαθηματικά

Σημειώσεις Παραδόσεων

Δημήτριος Α. Μητσούδης

Τμήμα Ναυπηγών Μηχανικών

Σχολή Μηχανικών



Νοέμβριος 2023

Περιεχόμενα

Πρόλογος	i
1 Αριθμητική στον υπολογιστή	1
1.1 Συστήματα αρίθμησης	1
1.2 Αριθμοί κινητής υποδιαστολής – Αριθμοί μηχανής	4
1.3 Επιρροή των σφαλμάτων στρογγύλευσης στους υπολογισμούς	10
1.4 Αριθμητική κινητής υποδιαστολής στο MATLAB	14
1.5 Ευστάθεια αλγορίθμων	18
1.6 Συμπληρωματικές παρατηρήσεις	23
2 Επίλυση μη γραμμικών εξισώσεων	27
2.1 Η μέθοδος της διχοτόμησης	27
2.2 Επαναληπτικές μέθοδοι	35
2.3 Η μέθοδος του Νεύτωνα	42
2.4 Η μέθοδος της τέμνουσας	48
2.5 Συνδυασμοί μεθόδων (υβριδικά σχήματα)	49
3 Γραμμικά συστήματα	53
3.1 Η μέθοδος απαλοιφής του Gauss	56
3.1.1 Αλγόριθμος απαλοιφής Gauss.	58
3.1.2 Απαλοιφή Gauss με μερική οδήγηση	62
3.2 Κατάσταση γραμμικών συστημάτων	64
3.3 Νόρμες διανυσμάτων και πινάκων	66
3.4 Δείκτης κατάστασης πίνακα	70
3.5 Τριδιαγώνια συστήματα	73
3.6 Επίλυση γραμμικών συστημάτων στο MATLAB	77
3.7 Επαναληπτικές μέθοδοι	79
3.8 Συμπληρωματικές παρατηρήσεις	84

4	Παρεμβολή	89
4.1	Πολυωνυμική παρεμβολή	90
4.1.1	Παράσταση του πολωνύμου παρεμβολής σε μορφή Lagrange	91
4.1.2	Παράσταση σε μορφή του Νεύτωνα	93
4.2	Σφάλματα στην πολυωνυμική παρεμβολή	95
4.3	Παρεμβολή με splines	96
4.3.1	Γραμμικές splines	97
4.3.2	Κυβικές splines	99
4.4	Παρεμβολή στο MATLAB	100
4.5	Συμπληρωματικές παρατηρήσεις	105
5	Αριθμητική ολοκλήρωση	109
5.1	Ο τύπος του τραπεζίου	111
5.2	Ο τύπος του Simpson	115
5.3	Αριθμητική ολοκλήρωση στο MATLAB	117
6	Προβλήματα αρχικών τιμών για Σ.Δ.Ε.	121
6.1	Στοιχεία Θεωρίας Π.Α.Τ. για Σ.Δ.Ε.	122
6.2	Η μέθοδος του Euler	123
6.2.1	Τρόποι κατασκευής της μεθόδου του Euler	124
6.2.2	Γεωμετρική ερμηνεία της μεθόδου του Euler	128
6.2.3	Σφάλματα της μεθόδου του Euler	129
6.2.4	Ευστάθεια της μεθόδου του Euler	132
6.3	Άλλες απλές μέθοδοι	134
6.3.1	Η πεπλεγμένη μέθοδος του Euler	134
6.3.2	Η μέθοδος του τραπεζίου	137
6.4	Γενικεύσεις της μεθόδου του Euler. Μέθοδοι Runge-Kutta	139
6.4.1	Παραδείγματα μεθόδων Runge-Kutta	141
6.4.2	Η κλασσική μέθοδος των Runge-Kutta	142
6.5	Αυτόματη μεταβολή βήματος	145
6.6	Επίλυση Π.Α.Τ. στο MATLAB	147
6.7	Άκαμπτα προβλήματα	150
6.8	Αριθμητικές μέθοδοι για συστήματα Δ.Ε.	152

A' Επισκόπηση Απειρ. Λογισμού	161
A'.1 Υπολογισμός τιμών πολυωνύμου - Σχήμα Horner	161
A'.2 Ο τύπος του Taylor	162
A'.3 Απόλυτο και σχετικό σφάλμα	163
B' Μ-αρχεία και συναρτήσεις στο MATLAB	165
B'.1 Μ-αρχεία	165
B'.2 Function handles	166
B'.3 Ανώνυμες συναρτήσεις	167

Πρόλογος

Οι σημειώσεις αυτές είναι εξέλιξη σημειώσεων που αρχικά γράφτηκαν για τις ανάγκες ενός μέρους του προπτυχιακού μαθήματος «Μαθηματικά ΙΙΙ» που διδάσκονταν οι φοιτητές του Γ' εξαμήνου του Τμήματος Μηχανικών Ενεργειακής Τεχνολογίας ΤΕ του ΤΕΙ Αθήνας. Στην παρούσα τους μορφή έχουν ως στόχο να καλύψουν τις ανάγκες του μαθήματος «Αριθμητική Ανάλυση» που προσφέρεται στο Γ' εξάμηνο σπουδών του Τμήματος Ναυπηγών Μηχανικών του Πανεπιστημίου Δυτικής Αττικής. Οι σημειώσεις αυτές δεν αποτελούν σε καμία περίπτωση ένα ολοκληρωμένο σύγγραμμα εισαγωγής στην Αριθμητική Ανάλυση. Ακολουθούν, εν γένει, τη δομή του βιβλίου των Γ.Δ. Ακρίβη και Β.Α. Δουγαλή, *Εισαγωγή στην Αριθμητική Ανάλυση*, Πανεπιστημιακές Εκδόσεις Κρήτης, 2015, χωρίς όμως να εστιάζουν στη βαθύτερη μαθηματική ανάλυση των διαφόρων μεθόδων. Για τις περισσότερες από τις μεθόδους που συμπεριλαμβάνονται στις σημειώσεις αυτές προσπαθούμε να δώσουμε τις απαραίτητες πληροφορίες ώστε να μπορεί ο φοιτητής να τις υλοποιήσει στο MATLAB. Θεωρώ ότι ο αναγνώστης έχει να ωφεληθεί σημαντικά, αν εκτελεί στο MATLAB τις διάφορες εντολές και διαδικασίες που περιγράφονται, ταυτόχρονα με την ανάγνωση του κειμένου.

Στο κείμενο ακολουθούνται οι εξής συμβάσεις: Εντολές του MATLAB που δίνονται από το παράθυρο εντολών (command window) περικλείονται με πλαίσιο της ακόλουθης μορφής:

```
>> help eps
```

ενώ, το περιεχόμενο M-αρχείων θα περικλείεται με το ακόλουθο πλαίσιο:

```
triarea.m  
  
b = 5;  
h = 3;  
a = 0.5*(b.*h)
```

Τέλος, επειδή πρόκειται για την πρώτη έκδοση των σημειώσεων αυτών και, αναπόφευκτα, υπάρχουν αρκετές ελλείψεις και αβλεψίες, οποιαδήποτε επισήμανση λαθών είναι εξαιρετικά

ευπρόσδεκτη. Εξίσου ευπρόσδεκτα είναι σχόλια ή παρατηρήσεις που αφορούν το περιεχόμενό τους, καθώς και τον τρόπο παρουσίασης.

Νοέμβριος 2020
Δημήτρης Μητσούδης

Σε αυτή την δεύτερη έκδοση έχουν διορθωθεί κάποιες αβλεψίες και έχουν προστεθεί μερικές ασκήσεις στο τέλος κάθε κεφαλαίου. Οποιαδήποτε επισήμανση λαθών, καθώς και σχόλια ή παρατηρήσεις εξακολουθούν και είναι εξαιρετικά ευπρόσδεκτα.

Νοέμβριος 2023
Δημήτρης Μητσούδης

Κεφάλαιο 1

Αριθμητική στον υπολογιστή

1.1 Συστήματα αρίθμησης

Στην καθημερινή μας ζωή είμαστε εξοικειωμένοι με το δεκαδικό σύστημα και το χρησιμοποιούμε συνεχώς για την παράσταση αριθμών. Στο σύστημα αυτό η βάση είναι το δέκα και τα ψηφία είναι οι αριθμοί $0, 1, 2, \dots, 9$. Για παράδειγμα, ο δεκαδικός αριθμός 314.159 είναι ο

$$3 \cdot 10^2 + 1 \cdot 10^1 + 4 \cdot 10^0 + 1 \cdot 10^{-1} + 5 \cdot 10^{-2} + 9 \cdot 10^{-3}.$$

Γενικά, αν οι αριθμοί $a_n, a_{n-1}, \dots, a_1, a_0, a_{-1}, a_{-2}, \dots$ είναι ψηφία του δεκαδικού συστήματος (δηλαδή ακέραιοι αριθμοί από 0 έως 9), τότε ο αριθμός

$$a_n a_{n-1} \dots a_1 a_0 . a_{-1} a_{-2} \dots$$

στο δεκαδικό σύστημα είναι ο

$$\underbrace{a_n \cdot 10^n + a_{n-1} \cdot 10^{n-1} + \dots + a_1 \cdot 10^1 + a_0 \cdot 10^0}_{\text{ακέραιο μέρος}} + \underbrace{a_{-1} \cdot 10^{-1} + a_{-2} \cdot 10^{-2} + \dots}_{\text{κλασματικό μέρος}} \\ = \sum_{k=0}^n a_k 10^k + \sum_{k=1}^{\infty} a_{-k} 10^{-k}.$$

Κάθε πραγματικός αριθμός έχει μοναδική αναπαράσταση της προηγούμενης μορφής στο δεκαδικό σύστημα (με εξαίρεση όταν έχουμε μια άπειρη ακολουθία από εννιάρια· είναι γνωστό ότι, π.χ., ο δεκαδικός $2.349999 \dots$ είναι ο 2.35^* .)

Βεβαίως, μπορούμε να χρησιμοποιήσουμε ως βάση ενός συστήματος αρίθμησης οποιονδήποτε φυσικό $\beta \geq 2$. Σε ένα σύστημα με βάση β τα ψηφία είναι οι φυσικοί αριθμοί $0, 1, \dots, \beta - 1$.

*Αυτό είναι αποτέλεσμα της σχέσης $9 \cdot \sum_{k=1}^{\infty} 10^{-k} = 1$

(Αν το β είναι μεγαλύτερο του δέκα, τότε για τα ψηφία $10, \dots, \beta - 1$ θα πρέπει να χρησιμοποιήσουμε άλλα σύμβολα.) Έτσι αν τα a_k είναι ψηφία ως προς τη βάση β , το σύμβολο

$$\pm(a_n \dots a_0.a_{-1}a_{-2} \dots)_\beta$$

θα παριστάνει τον αριθμό

$$\begin{aligned} & \pm \sum_{k=-\infty}^n a_k \beta^k \\ & = \pm a_n \cdot \beta^n + \dots + a_1 \cdot \beta^1 + a_0 \cdot \beta^0 + a_{-1} \cdot \beta^{-1} + a_{-2} \cdot \beta^{-2} + \dots \end{aligned} \quad (1.1)$$

Τα συνηθέστερα αριθμητικά συστήματα που χρησιμοποιούνται στον υπολογιστή είναι το δυαδικό, το οκταδικό, και το δεκαεξαδικό. Στον ακόλουθο πίνακα βλέπουμε τα ψηφία των συνηθέστερων αριθμητικών συστημάτων.

Σύστημα	Δεκαδικό	Δυαδικό	Οκταδικό	Δεκαεξαδικό
Βάση (β)	10	2	8	16
Ψηφία	0, 1, ..., 9	0, 1	0, 1, ..., 7	0, 1, ..., 9, A, B, C, D, E, F

A) Μετατροπή αριθμού από ένα σύστημα με βάση β στο δεκαδικό.

Η μετατροπή ενός αριθμού από ένα σύστημα με βάση β στο δεκαδικό είναι πολύ εύκολη· απλά υπολογίζουμε την τιμή της παράστασης (1.1).

Παράδειγμα. Ο αριθμός $(10110.11)_2$ του δυαδικού συστήματος είναι ο αριθμός

$$1 \cdot 2^4 + 0 \cdot 2^3 + 1 \cdot 2^2 + 1 \cdot 2^1 + 0 \cdot 2^0 + 1 \cdot 2^{-1} + 1 \cdot 2^{-2} = (22.75)_{10}.$$

Ο αριθμός $(5137.24)_8$ του οκταδικού συστήματος είναι ο αριθμός

$$5 \cdot 8^3 + 1 \cdot 8^2 + 3 \cdot 8^1 + 7 \cdot 8^0 + 2 \cdot 8^{-1} + 4 \cdot 8^{-2} = 2655 + \frac{5}{16} = (2655.3125)_{10}.$$

Ο αριθμός $(5F9)_{16}$ του δεκαεξαδικού συστήματος είναι ο αριθμός

$$5 \cdot 16^2 + 15 \cdot 16^1 + 9 \cdot 16^0 = (1529)_{10}.$$

□

Η μετατροπή ενός αριθμού από το δεκαδικό σύστημα σε ένα σύστημα με βάση β είναι κάπως πιο πολύπλοκη και γίνεται χωριστά για το ακέραιο μέρος και χωριστά για το κλασματικό.

Β) Μετατροπή ακεραίου από το δεκαδικό σε ένα σύστημα με βάση β .

Η μετατροπή αυτή βασίζεται στον αλγόριθμο της διαίρεσης. Ας δούμε, για παράδειγμα, πώς θα μετατρέψουμε τον αριθμό $(253)_{10}$ στο δυαδικό σύστημα. Για το σκοπό αυτό θα πρέπει να προσδιορίσουμε δυαδικά ψηφία $a_0, a_1, a_2, \dots$, τέτοια ώστε

$$\begin{aligned}(253)_{10} &= (\dots a_2 a_1 a_0)_2 = \dots + a_2 \cdot 2^2 + a_1 \cdot 2^1 + a_0 \cdot 2^0 \\ &= a_0 + 2(a_1 + 2(a_2 + \dots) \dots).\end{aligned}$$

(Σημειώστε ότι περάσαμε από την προτελευταία παράσταση στην τελευταία εισάγοντας παρενθέσεις και εκτελώντας τις πράξεις από «μέσα» προς τα «έξω»: χρησιμοποιούμε δηλαδή το λεγόμενο σχήμα του Horner, βλ. Παράρτημα Α΄.1.) Από την παραπάνω σχέση διαπιστώνουμε ότι το ψηφίο a_0 είναι το υπόλοιπο της διαίρεσης $253 : 2$, δηλαδή $a_0 = 1$. Το πηλίκο αυτής της διαίρεσης είναι ο αριθμός $(126)_{10}$ και ισούται με $a_1 + 2(a_2 + \dots)$. Άρα, τώρα, το a_1 είναι το υπόλοιπο της διαίρεσης $126 : 2$, δηλαδή $a_1 = 0$, και συνεχίζουμε με τον τρόπο αυτό (μέχρι το πηλίκο να γίνει μηδέν) όπως φαίνεται στον ακόλουθο πίνακα:

Αριθμός στο δεκαδικό	Διαίρεση δια β	Πηλίκο	Υπόλοιπο	Ψηφίο στο σύστημα με βάση β
253	$253 : 2$	126	1	$a_0 = 1$
126	$126 : 2$	63	0	$a_1 = 0$
63	$63 : 2$	31	1	$a_2 = 1$
31	$31 : 2$	15	1	$a_3 = 1$
15	$15 : 2$	7	1	$a_4 = 1$
7	$7 : 2$	3	1	$a_5 = 1$
3	$3 : 2$	1	1	$a_6 = 1$
1	$1 : 2$	0	1	$a_7 = 1$

Επομένως $(253)_{10} = (11111101)_2$.

Γ) Μετατροπή κλασματικού από το δεκαδικό σε ένα σύστημα με βάση β .

Για να μετατρέψουμε έναν κλασματικό αριθμό x από το δεκαδικό σε ένα σύστημα με βάση β εργαζόμαστε ως εξής: Έστω

$$x = (0.a_{-1}a_{-2}a_{-3}\dots)_{\beta} = a_{-1} \cdot \beta^{-1} + a_{-2} \cdot \beta^{-2} + a_{-3} \cdot \beta^{-3} \dots$$

Πολλαπλασιάζουμε και τα δύο μέλη επί β , οπότε

$$\beta x = (a_{-1}.a_{-2}a_{-3}\dots)_{\beta}$$

Άρα το a_{-1} είναι το ακέραιο μέρος του βx και το $(0.a_{-2}a_{-3}\dots)_\beta$ το κλασματικό μέρος. Συνεχίζουμε με αυτό τον τρόπο όπως θα δούμε στο επόμενο παράδειγμα. Ας μετατρέψουμε τον αριθμό $(0.325)_{10}$ στο δυαδικό σύστημα. Δηλαδή αναζητούμε ψηφία $a_{-1}, a_{-2}, \dots$, τέτοια ώστε $(0.325)_{10} = (0.a_{-1}a_{-2}\dots)_2$. Τα βήματα που κάνουμε συνοψίζονται στον ακόλουθο πίνακα:

Αριθμός στο δεκαδικό	Πολ/σμός επί β	Ακέραιο μέρος	Κλασματικό μέρος	Ψηφίο στο σύστημα με βάση β
0.325	$0.325 \cdot 2 = 0.65$	0	0.65	$a_{-1} = 0$
0.65	$0.65 \cdot 2 = 1.3$	1	0.3	$a_{-2} = 1$
0.3	$0.3 \cdot 2 = 0.6$	0	0.6	$a_{-3} = 0$
0.6	$0.6 \cdot 2 = 1.2$	1	0.2	$a_{-4} = 1$
0.2	$0.2 \cdot 2 = 0.4$	0	0.4	$a_{-5} = 0$
0.4	$0.4 \cdot 2 = 0.8$	0	0.8	$a_{-6} = 0$
0.8	$0.8 \cdot 2 = 1.6$	1	0.6	$a_{-7} = 1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Επομένως $(0.325)_{10} = (0.0101001\dots)_2$.

Παρατηρούμε ότι προφανώς ένας ακέραιος σε ένα σύστημα αρίθμησης παραμένει ακέραιος αν τον μετατρέψουμε σε ένα σύστημα με άλλη βάση. Ένας κλασματικός παραμένει κλασματικός αλλά το πλήθος των ψηφίων του μπορεί από πεπερασμένο να γίνει άπειρο ή αντίστροφα. Π.χ. αν προσπαθήσετε να μετατρέψετε το κλάσμα $1/10$ στο δυαδικό θα διαπιστώσετε ότι απαιτείται ο υπολογισμός μιας άπειρης σειράς. Συγκεκριμένα, ισχύει ότι

$$\frac{1}{10} = \sum_{k=1}^{\infty} \left(\frac{1}{2^{4k}} + \frac{1}{2^{4k+1}} \right),$$

επομένως

$$(0.1)_{10} = (0.0001\ 1001\ 1001\dots)_2.$$

1.2 Αριθμοί κινητής υποδιαστολής – Αριθμοί μηχανής

Κάθε πραγματικός αριθμός $x \neq 0$ σ' ένα αριθμητικό σύστημα με βάση β μπορεί να γραφτεί στη λεγόμενη (κανονική) μορφή κινητής υποδιαστολής

$$x = \pm(0.d_1d_2\dots) \cdot \beta^e, \quad \text{όπου } 0 \leq d_i \leq \beta - 1 \text{ και } d_1 \neq 0. \quad (1.2)$$

Το e είναι κατάλληλος ακέραιος που λέγεται *εκθέτης*, ενώ ο αριθμός $m = (0.d_1d_2\dots)$ λέγεται *κλασματικό μέρος* (*fractional part* ή *mantissa*).

Π.χ.

$$\begin{aligned} -(0.03726)_{10} &= -0.3726 \cdot 10^{-1} \\ (1011.01)_2 &= 0.101101 \cdot 2^4. \end{aligned}$$

Φυσικά, για την παράσταση ενός πραγματικού αριθμού μπορεί να χρειαστούν άπειρα ψηφία. Σε έναν υπολογιστή όμως, λόγω της πεπερασμένης μνήμης του, το πλήθος των ψηφίων του κλασματικού μέρους πρέπει να είναι πεπερασμένο, και ο εκθέτης πρέπει να παίρνει τιμές σ' ένα πεπερασμένο διάστημα. Δηλαδή, οι αριθμοί που χρησιμοποιεί ένας υπολογιστής είναι ένα σύνολο *ρητών* σε μορφή κινητής υποδιαστολής που ονομάζονται *αριθμοί μηχανής* και θα το συμβολίζουμε με $\mathcal{M}$.

Σε κάθε υπολογιστή το σύνολο των αριθμών μηχανής $\mathcal{M}$ χαρακτηρίζεται από τις εξής παραμέτρους:

- Τη βάση β του αριθμητικού συστήματος.
- Την ακρίβεια t , δηλαδή το πλήθος των ψηφίων του κλάσματος.
- Το κάτω φράγμα L και το άνω φράγμα U του εκθέτη e , όπου L και U είναι ακέραιοι (με $L \approx -U$).

Άρα κάθε αριθμός μηχανής $x \neq 0$ είναι της μορφής:

$$x = \pm(0.d_1d_2 \dots d_t) \cdot \beta^e, \quad (1.3)$$

όπου $0 \leq d_i \leq \beta - 1$, $i = 1, \dots, t$, $d_1 \neq 0$, και e ακέραιος με $L \leq e \leq U$.

Παρατηρήστε ότι το πεπερασμένο πλήθος ψηφίων του κλάσματος μας περιορίζει ως προς την *ακρίβεια* ενός αριθμού, ενώ ο περιορισμός στον εκθέτη περιορίζει το *εύρος* των αριθμών μηχανής.

Σε αυτές τις σημειώσεις δεν θα ασχοληθούμε με τον τρόπο αποθήκευσης των αριθμών μηχανής στον υπολογιστή, δηλαδή με το πώς αποθηκεύεται ο εκθέτης, το κλάσμα, το πρόσημο κ.λ.π. Σημειώστε ότι πριν από μερικές δεκαετίες η κατάσταση ήταν πολύ μπερδεμένη. Κάθε υπολογιστής είχε το δικό του αριθμητικό σύστημα κινητής υποδιαστολής (δεκαδικό, δυαδικό· υπήρχε ακόμα και ένας ρώσικος υπολογιστής με τριαδικό σύστημα, βλ., π.χ.

https://en.wikipedia.org/wiki/Ternary_computer).

Επιπλέον, καθένας είχε διαφορετική ακρίβεια. Το 1985 το Institute of Electrical and Electronics Engineers (IEEE) και το American National Standards Institute (ANSI) υιοθέτησαν το πρότυπο

δυναμικού συστήματος κινητής υποδιαστολής ANSI/IEEE Standard 754-1985, βλ. Πίνακα 1.1 για τις σχετικές παραμέτρους. Η οριστικοποίηση αυτού του προτύπου ήταν το αποτέλεσμα της (σχεδόν δεκαετούς) εργασίας μιας μεγάλης ομάδας ανθρώπων που περιελάμβανε μαθηματικούς, επιστήμονες υπολογιστών, και μηχανικούς, από πανεπιστήμια, κατασκευαστές υπολογιστών, και εταιρείες κατασκευής μικροεπεξεργαστών.

Πίνακας 1.1: ANSI/IEEE Standard 754-1985 for Binary Floating-Point Arithmetic

IEEE	β	t	L	U	β^{1-t}
Απλή ακρίβεια	2	24	-125	128	1.2×10^{-7}
Διπλή ακρίβεια	2	53	-1021	1024	2.2×10^{-16}
Επεκτεταμένη διπλή ακρίβεια	2	64	-16381	16384	1.1×10^{-19}

Όλοι οι υπολογιστές που έχουν σχεδιαστεί από το 1985 και μετά χρησιμοποιούν το δυαδικό σύστημα κινητής υποδιαστολής IEEE. Αυτό δεν σημαίνει ότι όλοι δίνουν ακριβώς τα ίδια αποτελέσματα, αφού υπάρχει κάποια ευελιξία μέσα στο πρότυπο αυτό, αλλά μας δίνει πλέον τη δυνατότητα να μελετήσουμε τη συμπεριφορά ενός συστήματος κινητής υποδιαστολής ανεξάρτητα από το συγκεκριμένο υπολογιστικό σύστημα στο οποίο δουλεύουμε κάθε φορά. Όπως αναφέραμε προηγουμένως, εμάς δεν θα μας απασχολήσει ο ακριβής τρόπος αποθήκευσης των αριθμών μηχανής στον υπολογιστή, αλλά θα εστιάσουμε την προσοχή μας στο ρόλο των παραμέτρων β, t, L, U , και στη λεγόμενη στρογγύλευση των αριθμών. Σε κάθε υπολογιστή (με δεδομένα β, t, L, U) ο μεγαλύτερος κατ' απόλυτο τιμή αριθμός μηχανής είναι ο

$$0. \underbrace{(\beta - 1)(\beta - 1) \dots (\beta - 1)}_{t \text{ ψηφία}} \cdot \beta^U,$$

ενώ ο μικρότερος κατ' απόλυτο τιμή αριθμός μηχανής είναι ο

$$0. \underbrace{10 \dots 0}_{t \text{ ψηφία}} \cdot \beta^L.$$

Εύκολα διαπιστώνει κανείς ότι το $\mathcal{M}$ δεν έχει πολλές από τις συνήθεις ιδιότητες του συνόλου των πραγματικών αριθμών. Για παράδειγμα, γνωρίζουμε ότι το άθροισμα και το γινόμενο δύο πραγματικών αριθμών είναι πραγματικός. Αυτό δεν ισχύει στο σύνολο των αριθμών μηχανής.*

Π.χ., ελέγξτε ότι το γινόμενο του μικρότερου θετικού αριθμού μηχανής, $0.1 \cdot \beta^L$, με τον εαυτό του δεν ανήκει στο $\mathcal{M}$. Επιπλέον, αν θεωρήσουμε έναν υπολογιστή με παραμέτρους $\beta = 10$,

*Στη γλώσσα των μαθηματικών λέμε ότι το $\mathcal{M}$ δεν είναι κλειστό ως προς την πρόσθεση και τον πολλαπλασιασμό· ειδικότερα δεν είναι σώμα ως προς τις πράξεις αυτές.

$t = 5$, $U = -L = 10$, τότε ο $x = 0.1 \cdot 10^1 = 1$ και ο $y = 0.1 \cdot 10^{-4} = 0.00001$ είναι αριθμοί μηχανής, ενώ το άθροισμά τους

$$x + y = 1 + 0.00001 = 1.00001 = 0.\underbrace{100001}_{6 \text{ ψηφία}} \cdot 10^1$$

δεν είναι αριθμός μηχανής, αφού το κλάσμα του έχει έξι ψηφία.

Επιπλέον, κατά τη διάρκεια πράξεων στον υπολογιστή μπορεί να προκύψει ένας αριθμός με εκθέτη μεγαλύτερο από U . Σε αυτή την περίπτωση ο υπολογιστής θα δώσει το μήνυμα «υπερχείλιση» (overflow) και θα διακόψει τον υπολογισμό. Ανάλογα, αν προκύψει αριθμός με εκθέτη μικρότερο του L θα έχουμε «υπεκχείλιση» (underflow). (Σε αυτή την περίπτωση οι περισσότεροι υπολογιστές μηδενίζουν τον αριθμό και συνεχίζουν τις πράξεις.)

Ας δούμε τώρα πώς προσεγγίζουμε πραγματικούς αριθμούς με αριθμούς μηχανής.

Για να παρασταθεί στον υπολογιστή ένας πραγματικός αριθμός x με $0.1 \cdot \beta^L \leq |x| \leq \beta^U$, προσεγγίζεται από έναν αριθμό μηχανής που βρίσκεται «κοντά» στον x , και τον οποίο συμβολίζουμε με $\text{fl}(x)$. Η προσέγγιση του x από τον αριθμό μηχανής $\text{fl}(x)$ μπορεί να γίνει με δύο τρόπους:

α) Με *στρογγύλευση*, όπου επιλέγεται ως $\text{fl}(x)$ ο πλησιέστερος προς τον x αριθμός μηχανής. (Στην πράξη γίνεται στρογγύλευση του t -οστού ψηφίου του x προς τα πάνω ή κάτω.)

β) Με *αποκοπή*, όπου αποκόπτουμε όλα τα ψηφία του κλάσματος μετά το t -οστό.

Παράδειγμα. Έστω ένας υπολογιστής με παραμέτρους $\beta = 10$, $t = 5$, $U = -L = 10$. Τότε ο πραγματικός αριθμός

$$\pi = 3.141592654 \dots = 0.\underbrace{31415}_{5 \text{ ψηφία}} 92654 \dots \times 10^1$$

θα παρασταθεί στον υπολογιστή ως:

$$\begin{aligned} \text{fl}(\pi) &= 0.31416 \cdot 10^1 && \text{με στρογγύλευση,} \\ \text{fl}(\pi) &= 0.31415 \cdot 10^1 && \text{με αποκοπή.} \end{aligned}$$

□

Στη συνέχεια, θα υποθέτουμε ότι αν $x \neq 0$ είναι ένας πραγματικός αριθμός μέσα στο εύρος των αριθμών μηχανής, τότε θα προσεγγίζεται και θα παριστάνεται στον υπολογιστή από τον αριθμό μηχανής $\text{fl}(x)$. Το σχετικό σφάλμα θα είναι

$$\frac{|\text{fl}(x) - x|}{|x|} \leq u. \quad (1.4)$$

Το u λέγεται μοναδιαίο σφάλμα στρογγύλευσης και ισούται με

$$u = \begin{cases} \frac{1}{2}\beta^{1-t} & \text{για στρογγύλευση,} \\ \beta^{1-t} & \text{για αποκοπή.} \end{cases} \quad (1.5)$$

Παρατήρηση 1.1. Η εκτίμηση (1.4) είναι ισοδύναμη με τη σχέση

$$\text{fl}(x) = x(1 + \delta) \quad \text{για κάποιο } \delta = \delta(x), \quad |\delta| \leq u. \quad (1.6)$$

Απόδειξη. Θέτουμε $\delta = \frac{\text{fl}(x) - x}{x}$ και λύνουμε ως προς $\text{fl}(x)$. Επομένως έχουμε $\text{fl}(x) = x(1 + \delta)$ και, λόγω της (1.4), $|\delta| = \left| \frac{\text{fl}(x) - x}{x} \right| \leq u$. $\square$

Φυσικά, στην πράξη, το σφάλμα που γίνεται στην παράσταση ενός πραγματικού αριθμού στον υπολογιστή είναι πολύ μικρό όπως, άλλωστε, υποδεικνύει η τελευταία στήλη του Πίνακα 1.1. Όμως, αν και πολύ μικρά, τέτοια σφάλματα οδηγούν σε σφάλματα και στις αριθμητικές πράξεις, τα οποία συνήθως είναι και αυτά μικρά ώστε να μας ενοχλούν. (Παρόλα αυτά υπάρχουν περιπτώσεις όπου λόγω συσσώρευσης μετά από μεγάλο πλήθος πράξεων έχουν καταστροφικά αποτελέσματα.) Στο σημείο αυτό θα πρέπει να ξεκαθαρίσουμε τι εννοούμε ως *πράξη στον υπολογιστή*. Συγκεκριμένα, θα υποθέσουμε το ακόλουθο, αρκετά ρεαλιστικό, μοντέλο του πραγματικού μηχανισμού των πράξεων:

Αν $*$ είναι μια από τις τέσσερις πράξεις της αριθμητικής ($+$, $-$, $\cdot$, $/$) και x, y είναι πραγματικοί αριθμοί που μπορούν να παρασταθούν κατά προσέγγιση από τους αριθμούς μηχανής $\text{fl}(x)$, $\text{fl}(y)$, αντίστοιχα, τότε θα υποθέσουμε ότι η πράξη αυτή στον υπολογιστή θα γίνει ως

$$x \circledast y = \text{fl}(\text{fl}(x) * \text{fl}(y)). \quad (1.7)$$

Δηλαδή στον υπολογιστή:

- Αρχικά γίνεται η παράσταση των αριθμών x, y από τους αριθμούς μηχανής $\text{fl}(x)$, $\text{fl}(y)$, αντίστοιχα.
- Στη συνέχεια εκτελείται η πράξη $\text{fl}(x) * \text{fl}(y)$ ακριβώς (δηλαδή με άπειρη ακρίβεια).*
- Τέλος, το αποτέλεσμα της πράξης το οποίο δεν είναι, εν γένει, αριθμός μηχανής παριστάνεται στο $\mathcal{M}$ προσεγγίζοντάς το με τη συνάρτηση $\text{fl}(\cdot)$.

Παραδείγματα. Θα θεωρήσουμε ότι εργαζόμαστε σ' έναν υπολογιστή με παραμέτρους $\beta = 10$, $t = 5$, $U = -L = 10$, που κάνει στρογγύλευση.

*Στην πραγματικότητα, η πράξη $\text{fl}(x) * \text{fl}(y)$ συνήθως γίνεται με ακρίβεια $2t$ ψηφίων κλάσματος.

- (1) Έστω $x = 3721.57$, $y = 0.0783251$ και θέλω να υπολογίσω το άθροισμά τους στον παραπάνω υπολογιστή, δηλαδή το $x \oplus y$. Προφανώς, αν οι πράξεις εκτελεστούν με άπειρη ακρίβεια θα βρούμε $x + y = 3721.6483251$.

Αρχικά, γράφουμε τα x, y σε κανονική μορφή κινητής υποδιαστολής:

$$x = 0.\underbrace{372157}_{5 \text{ ψηφία}} \cdot 10^4, \quad y = 0.\underbrace{783251}_{5 \text{ ψηφία}} \cdot 10^{-1}$$

Άρα, κάνοντας στρογγύλευση έχουμε

$$\text{fl}(x) = 0.37216 \cdot 10^4, \quad \text{fl}(y) = 0.78325 \cdot 10^{-1}.$$

Στη συνέχεια μετατρέπουμε τον εκθέτη του $\text{fl}(y)$ σε $+4$ και εκτελούμε την πρόσθεση ακριβώς (με άπειρη ακρίβεια)

$$\begin{array}{r} 0.3721600000 \cdot 10^4 \\ + 0.0000078325 \cdot 10^4 \\ \hline 0.3721678325 \cdot 10^4 \end{array}$$

Επομένως

$$x \oplus y = \text{fl}(0.\underbrace{3721678325}_{5 \text{ ψηφία}} \cdot 10^4) = 0.37217 \cdot 10^4 = 3721.7.$$

Παρατηρήστε ότι το αποτέλεσμα $x \oplus y$ στον υπολογιστή είναι διαφορετικό από το $x + y$, από το $\text{fl}(x + y)$ ($= 0.37216 \cdot 10^4$), και από το $\text{fl}(x) + \text{fl}(y)$.

- (2) Θεωρήστε τους αριθμούς μηχανής $a = 1$, $b = 0.3 \cdot 10^{-4}$, $c = 0.4 \cdot 10^{-4}$. Θέλουμε να υπολογίσουμε με τον παραπάνω υπολογιστή τα αθροίσματα $(a \oplus b) \oplus c$ και $a \oplus (b \oplus c)$. Επαληθεύστε (εφαρμόζοντας την (1.7) δύο φορές) ότι

$$(a \oplus b) \oplus c = 1 \quad \text{ενώ} \quad a \oplus (b \oplus c) = 1.0001.$$

Άρα σε έναν υπολογιστή το αποτέλεσμα ενός αθροίσματος δεν είναι ανεξάρτητο της σειράς με την οποία εκτελούνται οι προσθέσεις (σε αντίθεση με ότι συμβαίνει στους πραγματικούς αριθμούς).

- (3) Στους πραγματικούς αριθμούς η εξίσωση $1 + x = 1$ έχει μοναδική λύση την $x = 0$. Θεωρήστε στον παραπάνω υπολογιστή τον αριθμό μηχανής $x = 0.4 \cdot 10^{-4}$. Τότε

$$1 \oplus x = \text{fl}(0.1 \cdot 10^1 + 0.4 \cdot 10^{-4}) = \text{fl}(0.100004 \cdot 10^1) = 0.1 \cdot 10^1 = 1.$$

(Παρατηρήστε ότι το x είναι πολύ μεγαλύτερο από τον μικρότερο θετικό αριθμό μηχανής στον υπολογιστή μας που είναι ο $0.1 \cdot 10^{-10} = 10^{-11}$.)

Για την ακρίβεια, οποιοσδήποτε $x \in \mathbb{R}$ με $0 < x < 0.5 \cdot 10^{-4}$ είναι λύση της εξίσωσης $1 + x = 1$ στον συγκεκριμένο υπολογιστή.

Γενικότερα, σε έναν υπολογιστή αν προσθέσουμε σε έναν αριθμό της τάξεως του 1 μια ποσότητα που είναι κατ' απόλυτο τιμή μικρότερη του $\frac{1}{2}\beta^{1-t}$ ο αριθμός δεν αλλάζει. Το όριο αυτό λέγεται *μηδέν της μηχανής* ή *έψιλον της μηχανής* και είναι χρήσιμο να το ξέρουμε στους υπολογισμούς μας.

1.3 Επιρροή των σφαλμάτων στρογγύλευσης στους υπολογισμούς

Σε αυτή την παράγραφο θα εξετάσουμε την επιρροή των σφαλμάτων στρογγύλευσης όταν κάνουμε πράξεις στον υπολογιστή και θα δώσουμε ορισμένα παραδείγματα. Θα υποθέσουμε ότι οι πράξεις αυτές ακολουθούν το μοντέλο (1.7). Για μια λεπτομερέστερη μελέτη της επιρροής που έχουν τα σφάλματα στρογγύλευσης στους υπολογισμούς παραπέμπουμε στο βιβλίο των Ακρίβη-Δουγαλή [1, Παράγρ. 1.3, 1.4].

Θα ξεκινήσουμε υποθέτοντας ότι οι αριθμοί x, y και $x * y$ είναι μη μηδενικοί αριθμοί στο εύρος των αριθμών μηχανής και θα προσπαθήσουμε να εκτιμήσουμε την απόλυτη τιμή του σχετικού σφάλματος

$$\frac{(x \circledast y) - (x * y)}{x * y},$$

όπου « $\circledast$ » είναι μια από τις τέσσερις πράξεις της αριθμητικής (+, −, ·, /), και με « $\circledast$ » συμβολίζουμε την εκτέλεση της πράξης « $\circledast$ » στον υπολογιστή σύμφωνα με το μοντέλο (1.7).

Για τη συνέχεια θα μας φανεί χρήσιμη η ακόλουθη παρατήρηση:

Παρατήρηση 1.2. Αν οι αριθμοί δ_i , $1 = 1, \dots, m$, ικανοποιούν τις ανισότητες $|\delta_i| \leq u < 1$, τότε υπάρχει κάποιο δ με $|\delta| \leq u$, τέτοιο ώστε

$$(1 + \delta_1)(1 + \delta_2) \dots (1 + \delta_m) = (1 + \delta)^m.$$

Απόδειξη. Ας συμβολίσουμε $P = (1 + \delta_1)(1 + \delta_2) \dots (1 + \delta_m)$. Προφανώς

$$|\delta_i| \leq u \Leftrightarrow -u \leq \delta_i \leq u \Leftrightarrow \underbrace{1 - u}_{>0} \leq 1 + \delta_i \leq 1 + u.$$

Επομένως, έχουμε ότι $(1 - u)^m \leq P \leq (1 + u)^m$.

Θεωρούμε τώρα τη συνάρτηση $f(x) = (1 + x)^m$ στο διάστημα $[-u, u]$. Είναι πολύ εύκολο να ελέγξουμε ότι η f είναι μια συνεχής, γνησίως αύξουσα συνάρτηση στο διάστημα $[-u, u]$,

και ότι το σύνολο τιμών της είναι το διάστημα $[(1 - u)^m, (1 + u)^m]$. Άρα από το θεώρημα ενδιάμεσης τιμής συμπεραίνουμε ότι υπάρχει κάποιο $\delta \in [-u, u]$ τέτοιο ώστε $f(\delta) = P \Leftrightarrow P = (1 + \delta)^m$. $\square$

Τώρα, είμαστε σε θέση να εκτιμήσουμε το σχετικό σφάλμα $\frac{(x \odot y) - (x * y)}{x * y}$ για κάθε μία από τις βασικές πράξεις. Για το σκοπό αυτό θα υποθέσουμε ότι x, y είναι δύο μη μηδενικοί αριθμοί στο εύρος των αριθμών μηχανής. Από την (1.6) έχουμε ότι

$$\text{fl}(x) = x(1 + \delta_1) \quad \text{και} \quad \text{fl}(y) = y(1 + \delta_2), \quad \text{με} \quad |\delta_1|, |\delta_2| \leq u.$$

Εκτίμηση σχετικού σφάλματος του πολλαπλασιασμού. Θέλουμε να εκτιμήσουμε την ποσότητα

$$\left| \frac{x \odot y - xy}{xy} \right|.$$

Φυσικά,

$$\begin{aligned} x \odot y &= \text{fl}(\text{fl}(x) \text{fl}(y)) = (x(1 + \delta_1) y(1 + \delta_2))(1 + \delta_3) \\ &= xy(1 + \delta_1)(1 + \delta_2)(1 + \delta_3) = xy(1 + \delta)^3, \end{aligned}$$

για κάποιο δ με $|\delta| \leq u$ (όπου η τελευταία ισότητα προκύπτει άμεσα από την Παρατήρηση 1.2). Άρα

$$\begin{aligned} \left| \frac{x \odot y - xy}{xy} \right| &= \left| \frac{xy(1 + \delta)^3 - xy}{xy} \right| = |(1 + \delta)^3 - 1| = |3\delta + 3\delta^2 + \delta^3| \\ &\leq 3|\delta| + 3\delta^2 + |\delta|^3 \leq 3u + 3u^2 + u^3. \end{aligned}$$

Επειδή το u είναι πολύ μικρό, οι όροι u^2, u^3 είναι πολύ μικρότεροι και μπορούμε να τους αγνοήσουμε. Άρα

$$\left| \frac{x \odot y - xy}{xy} \right| \approx 3u.$$

Συνεπώς, στον πολλαπλασιασμό το σχετικό σφάλμα είναι περίπου τριπλάσιο του μοναδιαίου σφάλματος στρογγύλευσης u .

Εκτίμηση σχετικού σφάλματος της διαίρεσης. Εργαζόμενοι ανάλογα μπορούμε να αποδείξουμε ότι

$$\left| \frac{x \oslash y - x/y}{x/y} \right| \approx 3u.$$

Δηλαδή, όταν διαιρούμε δύο αριθμούς στον υπολογιστή το σχετικό σφάλμα είναι περίπου τριπλάσιο του μοναδιαίου σφάλματος στρογγύλευσης u .

Εκτίμηση σχετικού σφάλματος της πρόσθεσης-αφαίρεσης. Θέλουμε να εκτιμήσουμε την ποσότητα

$$\left| \frac{x \oplus y - (x + y)}{x + y} \right|.$$

Φυσικά,

$$\begin{aligned} x \oplus y &= \text{fl}(\text{fl}(x) + \text{fl}(y)) = \text{fl}(x(1 + \delta_1) + y(1 + \delta_2)) \\ &= (x(1 + \delta_1) + y(1 + \delta_2))(1 + \delta_3) \\ &= x(1 + \delta_1)(1 + \delta_3) + y(1 + \delta_2)(1 + \delta_3) \\ &= x(1 + \delta)^2 + y(1 + \varepsilon)^2, \end{aligned}$$

όπου $|\delta|, |\varepsilon| \leq u$. Άρα

$$x \oplus y = (x + y) + 2(\delta x + \varepsilon y) + \delta^2 x + \varepsilon^2 y.$$

Τώρα, αν αγνοήσουμε τους όρους τάξεως u^2 που περιμένουμε να είναι πολύ μικροί, συμπεραίνουμε ότι

$$x \oplus y \approx (x + y) + 2(\delta x + \varepsilon y).$$

Συνεπώς

$$\begin{aligned} \left| \frac{x \oplus y - (x + y)}{x + y} \right| &\approx \left| \frac{(x + y) + 2(\delta x + \varepsilon y) - (x + y)}{x + y} \right| = 2 \left| \frac{\delta x + \varepsilon y}{x + y} \right| \\ &\leq 2 \frac{|\delta||x| + |\varepsilon||y|}{|x + y|} \leq \frac{2u(|x| + |y|)}{|x + y|}. \end{aligned}$$

Άρα

- αν x, y **ομόσημοι**, τότε $|x| + |y| = |x + y|$ και το σχετικό σφάλμα κατά την πρόσθεσή τους είναι και πάλι μικρό, περίπου ίσο με $2u$, ενώ
- αν x, y **ετερόσημοι** (δηλαδή κάνουμε αφαίρεση) και, συγκεκριμένα, αν $x \approx -y$, τότε $x - y \approx 0$ και άρα το φράγμα του σχετικού σφάλματος γίνεται πολύ μεγάλο. Στην περίπτωση αυτή, δηλαδή στην **αφαίρεση σχεδόν ίσων αριθμών**, περιμένουμε καταστροφική επιρροή του σφάλματος στρογγύλευσης.

Συνοψίζοντας, δείξαμε ότι το σχετικό σφάλμα που προκύπτει όταν *πολλαπλασιάζουμε ή διαιρούμε δύο αριθμούς στον υπολογιστή* είναι περίπου τριπλάσιο του μοναδιαίου σφάλματος στρογγύλευσης u . Για το *άθροισμα δύο ομόσημων αριθμών* δείξαμε ότι το σχετικό σφάλμα είναι και πάλι μικρό, περίπου ίσο με $2u$. Άρα η εκτέλεση των πράξεων αυτών στον υπολογιστή δεν παρουσιάζει κάποιο ιδιαίτερο πρόβλημα.

Αντίθετα, αναμένουμε ότι μπορεί να προκύψουν προβλήματα κατά την αφαίρεση περίπου ίσων αριθμών. Ανάλογα προβλήματα προκύπτουν συνήθως και κατά την πρόσθεση μεγάλων κατ' απόλυτο τιμή, ετερόσημων αριθμών όπως θα δούμε στο δεύτερο από τα ακόλουθα παραδείγματα.

Παραδείγματα. Έστω ότι εργαζόμαστε σ' έναν υπολογιστή με παραμέτρους $\beta = 10$, $t = 5$, $U = -L = 10$, που κάνει στρογγύλευση.

- (1) Έστω $x = 0.39134506$ και $y = -0.39117893$. Προφανώς το ακριβές άθροισμα των x και y είναι

$$x + y = 0.16613 \cdot 10^{-3}.$$

Στον παραπάνω υπολογιστή η πρόσθεση αυτή γίνεται ως εξής:

$$x \oplus y = \text{fl}(\text{fl}(x) + \text{fl}(y)) = \text{fl}(0.39135 - 0.39118) = 0.00017 = 0.17 \cdot 10^{-3}.$$

Άρα το σχετικό σφάλμα είναι:

$$\left| \frac{(x \oplus y) - (x + y)}{(x + y)} \right| \approx 233 \cdot 10^{-4}$$

Παρατηρήστε ότι σε αυτόν τον υπολογιστή το μοναδιαίο σφάλμα στρογγύλευσης $u = \frac{1}{2} \cdot 10^{-4}$. Δηλαδή εδώ το σχετικό σφάλμα της πράξης είναι περίπου 466 φορές μεγαλύτερο από το μοναδιαίο σφάλμα στρογγύλευσης, ενώ είπαμε ότι το σχετικό σφάλμα που προκύπτει όταν προσθέτουμε δύο ομόσημους αριθμούς είναι μόλις διπλάσιο του u !

- (2) Θα δούμε τώρα ένα ακραίο παράδειγμα που αφορά πρόσθεση μεγάλων σε απόλυτη τιμή, ετερόσημων αριθμών. Έστω $x = 123454000$ και $y = -123453000$. Προφανώς η ακριβής τιμή του αθροίσματος είναι $x + y = 1000$. Όμως στον υπολογιστή που εργαζόμαστε έχουμε

$$\text{fl}(123454000) = \text{fl}(0.123454000 \cdot 10^9) = 0.12345 \cdot 10^9 = 123450000,$$

$$\text{fl}(-123453000) = \text{fl}(-0.123453000 \cdot 10^9) = -0.12345 \cdot 10^9 = -123450000.$$

Άρα

$$x \oplus y = \text{fl}(\text{fl}(x) + \text{fl}(y)) = \text{fl}(123450000 - 123450000) = 0!$$

Παρατηρήστε ότι αν στα παραδείγματα αυτά χρησιμοποιήσουμε μεγαλύτερη ακρίβεια (π.χ. αν $t = 10$, δηλαδή χρησιμοποιήσουμε διπλή ακρίβεια), τότε οι x, y θα παριστάνονται στη μηχανή χωρίς σφάλμα και τα αποτελέσματα θα ήταν ακριβή. Επομένως, είναι απαραίτητο να χρησιμοποιούμε διπλή ακρίβεια σε περιπτώσεις όπου είναι πιθανόν να έχουμε ακύρωση σημαντικών ψηφίων. Παρόλα αυτά υπάρχουν περιπτώσεις υπολογισμών με πολλές πράξεις, στους οποίους ούτε η διπλή ακρίβεια διορθώνει την κατάσταση.

1.4 Αριθμητική κινητής υποδιαστολής στο MATLAB

Σε αυτή την παράγραφο θα δούμε ορισμένα από τα χαρακτηριστικά της αριθμητικής κινητής υποδιαστολής στο MATLAB. Στο MATLAB όλοι οι υπολογισμοί γίνονται σύμφωνα με το δυαδικό σύστημα κινητής υποδιαστολής IEEE σε διπλή ακρίβεια (double precision). Κάθε αριθμός καταλαμβάνει μια λέξη 64 bits. Οι μη μηδενικοί αριθμοί κατ' απόλυτο τιμή είναι μεγαλύτεροι από περίπου 10^{-308} και μικρότεροι από περίπου 10^{308} , και το μοναδιαίο σφάλμα στρογγύλευσης είναι $2^{-53} \approx 1.11 \cdot 10^{-16}$. Άρα μπορεί κανείς να πει ότι η παράσταση των αριθμών στο MATLAB καθώς και οι βασικές αριθμητικές πράξεις γίνονται με ακρίβεια περίπου 16 σημαντικών δεκαδικών ψηφίων.

Η συνάρτηση `eps` επιστρέφει την τιμή του λεγόμενου *μηδέν ή έψιλον της μηχανής*:

```
>> eps  
ans =  
2.2204e-16
```

Παρατηρήστε ότι αντίθετα απ' ότι συμβαίνει στους πραγματικούς αριθμούς η πρόσθεση του `eps` στο 1 αφήνει τον αριθμό 1 αναλλοίωτο.

```
>> 1+eps  
ans =  
1.0000
```

Οι συναρτήσεις `realmax` και `realmin` επιστρέφουν τον μεγαλύτερο και τον μικρότερο κατ' απόλυτο τιμή αριθμό, αντίστοιχα, που μπορεί να παρασταθεί στο MATLAB. Αν το αποτέλεσμα μιας πράξης είναι μεγαλύτερο από την τιμή που επιστρέφει η `realmax`, τότε έχουμε υπερχείλιση (overflow) και το αποτέλεσμα είναι `Inf` που δηλώνει το ∞ . Αν το αποτέλεσμα μιας πράξης είναι μικρότερο από `-realmax`, τότε το αποτέλεσμα είναι `-Inf` που δηλώνει το $-\infty$. Αν το αποτέλεσμα μιας πράξης είναι μικρότερο κατ' απόλυτο τιμή από `realmin` είτε οδηγεί σε υπερχείλιση (underflow) αν είναι μικρότερο από `eps*realmin`, ή οδηγεί σε έναν μη κανονικό (subnormal) αριθμό (με ηγετικά ψηφία μηδέν στο κλάσμα του).

```
>> realmax
ans =
    1.7977e+308

>> 1.3*realmax
ans =
    Inf

>> -2*realmax
ans =
   -Inf

>> realmin
ans =
    2.2251e-308

>> realmin*eps
ans =
    4.9407e-324

>> realmin*eps/2
ans =
    0
```

Στο MATLAB δίνεται η επιλογή να χρησιμοποιήσει κανείς και απλή ακρίβεια (single precision). Συχνά χρησιμοποιούμε απλή ακρίβεια για να διερευνήσουμε την ακρίβεια ενός αλγορίθμου. Είναι σύνηθες να εκτελεί κανείς έναν αλγόριθμο αρχικά σε απλή ακρίβεια και στη συνέχεια σε διπλή ακρίβεια και να χρησιμοποιεί τη διαφορά των αποτελεσμάτων ως μια εκτίμηση του σφάλματος στη λύση με απλή ακρίβεια. Στην απλή ακρίβεια η παράσταση των αριθμών, καθώς και οι βασικές αριθμητικές πράξεις γίνονται με ακρίβεια περίπου 8 σημαντικών δεκαδικών ψηφίων, ενώ το εύρος των αριθμών είναι σημαντικά μικρότερο με αποτέλεσμα να είναι πιθανότερο να έχουμε φαινόμενα υπερ- ή υπεκ-χείλισης.

```
>> eps('single')
ans =
    1.1921e-07

>> realmax('single')
ans =
    3.4028e+38

>> realmin('single')
ans =
    1.1755e-38
```

Στη συνέχεια θα δούμε με τη βοήθεια του MATLAB ορισμένα παραδείγματα της επίδρασης που έχουν τα σφάλματα στρογγύλευσης σε υπολογισμούς που γίνονται στη μηχανή.

Παράδειγμα 1. (Από το βιβλίο του Moler [10].) Ας εκτελέσουμε τις ακόλουθες εντολές στο MATLAB:

```
>> format long
>> a = 4/3;
>> b = a - 1;
>> c = 3*b;
>> e = 1 - c
e =
    2.220446049250313e-16
```

Φυσικά, αν εκτελέσουμε τις πράξεις ακριβώς έχουμε:

$$a = \frac{4}{3}, \quad b = a - 1 = \frac{4}{3} - 1 = \frac{1}{3}, \quad c = 3b = 3 \cdot \frac{1}{3} = 1, \quad e = 1 - c = 1 - 1 = 0.$$

Προκύπτει ότι το μόνο σφάλμα στρογγύλευσης γίνεται στην πρώτη διαίρεση $4/3$. Το πηλίκιο (στον υπολογιστή) δεν μπορεί να είναι ακριβώς $4/3$ (εκτός αν δουλεύαμε στον ρώσικο υπολογιστή που αναφέρθηκε στην Παράγραφο 1.2). Η αφαίρεση $a - 1$ μας δίνει ένα b που το τελευταίο του ψηφίο (bit) είναι 0. Άρα ο πολλαπλασιασμός $3*b$ γίνεται χωρίς σφάλμα, αλλά η τιμή που αποθηκεύεται στο c δεν είναι ακριβώς 1. Έτσι η τιμή του e δεν είναι ακριβώς 0. (Πριν καθιερωθεί το πρότυπο IEEE, ο κώδικας αυτός αποτελούσε ένα γρήγορο τεστ για την εκτίμηση του επιπέδου του σφάλματος στρογγύλευσης σε διάφορους υπολογιστές.) $\square$

Παράδειγμα 2. Θεωρήστε τους αριθμούς $x = 10\,000\,002$ και $y = 10\,000\,001$. Θέλουμε να υπολογίσουμε με το MATLAB την παράσταση $1/(\sqrt{x} - \sqrt{y})$ χρησιμοποιώντας απλή και διπλή ακρίβεια. Ένας τρόπος να γίνει αυτό είναι ο ακόλουθος:


```
>> xx = 10000002;
>> yy = 10000001;
>> x = single(xx);
>> y = single(yy);
>> r1 = 1/(sqrt(x) - sqrt(y))
r1 =
    4096
>> r2 = 1/(sqrt(xx) - sqrt(yy))
r2 =
  6.324555807119124e+03
```

Δηλαδή εκχωρούμε στις μεταβλητές (διπλής ακρίβειας) xx και yy τις τιμές 10 000 002 και 10 000 001, αντίστοιχα. Στη συνέχεια, εκχωρούμε τις τιμές αυτές και στις μεταβλητές (απλής ακρίβειας) x και y . Το αποτέλεσμα της πράξης $1/(\sqrt{x} - \sqrt{y})$ (με απλή ακρίβεια) εκχωρείται στη μεταβλητή (απλής ακρίβειας) $r1$, και το αντίστοιχο αποτέλεσμα σε διπλή ακρίβεια στην $r2$. Παρατηρούμε ότι τα δύο αποτελέσματα διαφέρουν πολύ· έχουμε $r_1 = 4096$ και $r_2 \approx 6324.56$. Αν θεωρήσουμε ως «σωστό» το αποτέλεσμα με διπλή ακρίβεια r_2 , μπορούμε με την ακόλουθη εντολή να υπολογίσουμε το σχετικό σφάλμα $\left| \frac{r_1 - r_2}{r_2} \right|$,

```
>> err = abs((r1 - r2)/r2)
err =
    0.3523656
```

το οποίο είναι περίπου 35%! Ο λόγος για την μεγάλη απώλεια ακρίβειας εδώ οφείλεται στην αφαίρεση των σχεδόν ίσων αριθμών $\text{sqrt}(x) = 3.1622781\text{e}+03$ και $\text{sqrt}(y) = 3.1622778\text{e}+03$ (όπου ο καθένας από αυτούς έχει υπολογιστεί με ικανοποιητική ακρίβεια). Στην περίπτωση αυτή για να αποφύγουμε την «προβληματική» αφαίρεση σχεδόν ίσων αριθμών μπορούμε να χρησιμοποιήσουμε την ταυτότητα

$$x - y = (\sqrt{x} - \sqrt{y})(\sqrt{x} + \sqrt{y}) \Leftrightarrow \frac{1}{\sqrt{x} - \sqrt{y}} = \frac{\sqrt{x} + \sqrt{y}}{x - y},$$

η οποία για τα δεδομένα μας είναι

$$\frac{1}{\sqrt{10000002} - \sqrt{10000001}} = \frac{\sqrt{10000002} + \sqrt{10000001}}{1}.$$

Ελέγχουμε τα παραπάνω στο MATLAB τόσο σε απλή, όσο και σε διπλή ακρίβεια.

```
>> sqrt(x)+sqrt(y)
ans =
    6.3245557e+03
>> sqrt(xx)+sqrt(yy)
ans =
    6.324555794678388e+03
```

Παράδειγμα 3. (Από το βιβλίο του Moler [10].) Με τη βοήθεια του MATLAB θα σχεδιάσουμε τη γραφική παράσταση του πολυωνύμου 7ου βαθμού

$$p(x) = x^7 - 7x^6 + 21x^5 - 35x^4 + 35x^3 - 21x^2 + 7x - 1,$$

για $x \in (0.988, 1.012)$.

```
>> x = 0.988 : 0.0001 : 1.012;
>> y = x.^7-7*x.^6+21*x.^5-35*x.^4+35*x.^3-21*x.^2+7*x-1;
>> plot(x,y)
```

Το γράφημα της παράστασης αυτής απεικονίζεται στο Σχήμα 1.1, και δεν θυμίζει σε τίποτα γραφική παράσταση ενός πολυωνύμου αφού δεν είναι μια λεία καμπύλη. Αυτό οφείλεται στην επίδραση των σφαλμάτων στρογγύλευσης.

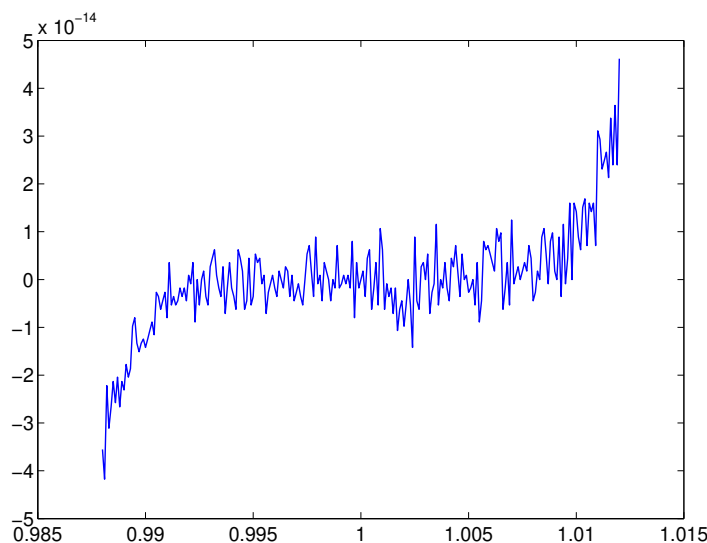
Παρατηρήστε ότι η κλίμακα στον κατακόρυφο άξονα των y είναι της τάξης του 10^{-14} . Οι πολύ μικρές αυτές τιμές έχουν υπολογιστεί παίρνοντας αθροίσματα και διαφορές αριθμών τόσο μεγάλων όσο ο $35 \cdot 1.012^4$, με αποτέλεσμα να υπάρχει εκτεταμένη ακύρωση σημαντικών ψηφίων. Όμως, το πολυώνυμο αυτό προκύπτει ως ανάπτυγμα του διωνύμου $(x - 1)^7$. Ας υπολογίσουμε με αυτόν τον τρόπο τις τιμές του πολυωνύμου και ας ξανακάνουμε το σχήμα.

```
>> z = (x - 1).^7;
>> plot(x,y,x,z, 'r')
```

Όπως βλέπουμε στο Σχήμα 1.2, η κόκκινη γραμμή που αντιστοιχεί στο γράφημα της $(x - 1)^7$ είναι μια λεία καμπύλη. □

1.5 Ευστάθεια αλγορίθμων

Αλγόριθμο ονομάζουμε μια σειρά εντολών που υλοποιούν κάποια αριθμητική μέθοδο. Είδαμε, ήδη, κάποια παραδείγματα όπου αλγόριθμοι είναι αρκετά «ευαίσθητοι» σε σφάλματα στρογ-



Σχήμα 1.1: Μοιάζει αυτή με γραφική παράσταση πολυωνύμου;

γύλευσης, υπό την έννοια ότι μικρές διαταραχές (σφάλματα) που εμφανίζονται κατά την παράσταση των αριθμών και τις πράξεις στον υπολογιστή επιφέρουν μεγάλες μεταβολές στο αποτέλεσμα του. Όταν συμβαίνει αυτό λέμε ότι ο αλγόριθμος είναι (αριθμητικά) *ασταθής*. Στην αντίθετη περίπτωση, όπου το τελικό αποτέλεσμα δεν επηρεάζεται σημαντικά από μικρές διαταραχές που επιφέρουν τα σφάλματα στρογγύλευσης σε κάθε βήμα του, ο αλγόριθμος λέγεται (αριθμητικά) *ευσταθής*. Στο επόμενο παράδειγμα που είναι από το βιβλίο των Forsythe, Malcolm και Moler [6] θα δούμε έναν ασταθή και έναν ευσταθή αλγόριθμο για τον υπολογισμό ενός ολοκληρώματος.

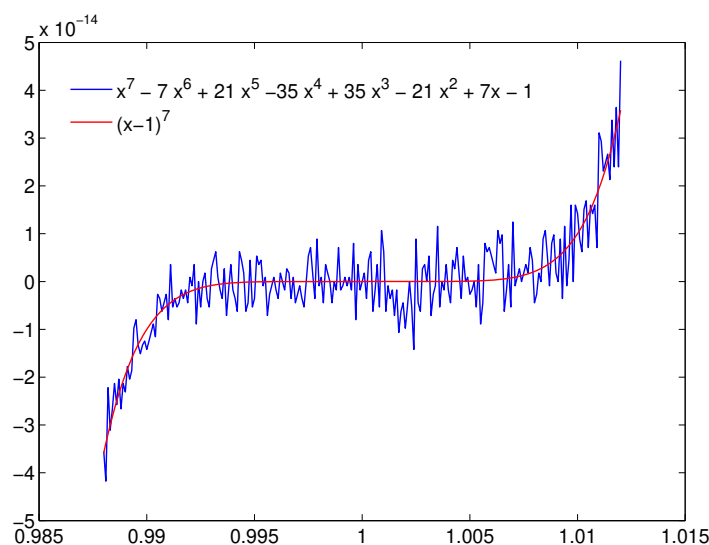
Παράδειγμα. Θέλουμε να υπολογίσουμε το ολοκλήρωμα

$$I_n = \int_0^1 x^n e^{x-1} dx, \quad n = 1, 2, 3, \dots, \quad (1.8)$$

για n αρκετά μεγάλο. Καταρχάς, διαπιστώνουμε άμεσα ότι $I_n \geq 0$, για κάθε $n \geq 1$, και $I_{n+1} < I_n$, αφού $x^{n+1} < x^n$ για κάθε $x \in (0, 1)$. Επιπλέον

$$I_n \leq \int_0^1 x^n dx = \frac{1}{n+1} \rightarrow 0, \quad \text{καθώς } n \rightarrow \infty. \quad (1.9)$$

Άρα η ακολουθία των ολοκληρωμάτων αυτών είναι γνησίως φθίνουσα και τείνει στο μηδέν καθώς το $n \rightarrow \infty$. Είναι εύκολο να υπολογίσουμε ακριβώς την τιμή του ολοκληρώματος για



Σχήμα 1.2: Οι γραφικές παραστάσεις των πολωνύμων $x^7 - 7x^6 + 21x^5 - 35x^4 + 35x^3 - 21x^2 + 7x - 1$ (μπλε γραμμή) και $(x - 1)^7$ (κόκκινη γραμμή).

$n = 1$ κάνοντας μια ολοκλήρωση κατά μέρη. Έχουμε

$$I_1 = \int_0^1 x e^{x-1} dx = \left[x e^{x-1} \right]_0^1 - \int_0^1 e^{x-1} dx = 1 - \left[e^{x-1} \right]_0^1 = \frac{1}{e}.$$

Επίσης, αν ολοκληρώσουμε κατά μέρη στην (1.8) παίρνουμε

$$I_n = \int_0^1 x^n e^{x-1} dx = 1 - n \int_0^1 x^{n-1} e^{x-1} dx = 1 - n I_{n-1} \\ \Rightarrow I_n = 1 - n I_{n-1}$$

Έτσι καταλήγουμε στον εξής αναδρομικό τύπο για τον υπολογισμό των I_n :

$$I_1 = \frac{1}{e}, \quad I_n = 1 - n I_{n-1}, \quad n = 2, 3, \dots \quad (1.10)$$

Η (1.10) αποτελεί έναν απλό αλγόριθμο για τον υπολογισμό των I_n , όπου σε κάθε βήμα του απαιτούνται ένας πολλαπλασιασμός και μια αφαίρεση.

Θα υλοποιήσουμε τον αλγόριθμό αυτό σε MATLAB. Για το σκοπό αυτό θα γράψουμε σ' ένα αρχείο (τύπου script) τις κατάλληλες εντολές του MATLAB. Το αρχείο θα πρέπει να έχει την επέκταση .m, π.χ. `integ1.m`, και για να το εκτελέσουμε απλώς γράφουμε στη γραμμή εντολών το όνομα του αρχείου χωρίς την επέκταση, δηλαδή

```
>> integ1
```

Στο πίνακα που ακολουθεί δίνουμε τις εντολές του MATLAB που θα πρέπει να περιέχονται στο αρχείο `integ1.m` για να υλοποιηθεί ο αλγόριθμος (1.10).

`integ1.m` (Υλοποίηση του αλγορίθμου (1.10) σε MATLAB)

```
clear all;
N = 20;
x = 1/exp(1);
fid = fopen('res1.txt','w');
fprintf(fid,'%3d %15.6f\n',1,x);
for n = 2:N
    x = 1 - n*x;
    fprintf(fid,'%3d %15.6f\n',n,x);
end
fclose(fid);
```

Ας δούμε, εν συντομία, κάποιες από τις (μη προφανείς) εντολές.

- Η εντολή `clear all` καθαρίζει τη μνήμη από τις ενεργές μεταβλητές.
- Η εντολή `fid = fopen('res1.txt','w')` δημιουργεί (ή ανοίγει αν υπάρχει ήδη) το αρχείο με όνομα `res1.txt` για γράψιμο ('w') (με διαγραφή του αρχικού περιεχομένου).
- Ο βρόχος `for` στο MATLAB έχει τη δομή:

```
for index = initial value : step : final value
    statements
end
```

όπου ο μετρητής `index` παίρνει τις τιμές από την αρχική τιμή `initial value` μέχρι την τελική τιμή `final value` με βήμα `step`. Οι εντολές `statements` εκτελούνται για όλες τις τιμές του δείκτη `index`. Αν, όπως στη δική μας περίπτωση, παραλείψουμε το βήμα, τότε το MATLAB χρησιμοποιεί ως βήμα την τιμή 1.
- Η εντολή `fprintf` είναι η αντίστοιχη της C, και χρησιμοποιείται για την εκτύπωση των μεταβλητών `n` και `x` στο αρχείο `res1.txt`.
- Τέλος, η εντολή `fclose` κλείνει το αρχείο που ανοίξαμε.

Εκτελώντας, λοιπόν, το αρχείο `integ1.m` παίρνουμε τα ακόλουθα αποτελέσματα (που έχουν αποθηκευτεί στο αρχείο `res1.txt`).

n	$\tilde{I}_n$
1	0.367879
2	0.264241
$\vdots$	$\vdots$
12	0.071773
13	0.066947
14	0.062736
15	0.058961
16	0.056621
17	0.037447
18	; 0.325949
19	; -5.193040
20	;! 104.860796

Παρατηρούμε ότι μέχρι και για $n = 17$ τα αποτελέσματα μοιάζουν λογικά καθώς η τιμή $\tilde{I}_n$ που υπολογίζουμε με το MATLAB φθίνει καθώς το n αυξάνει. Για $n = 18, 19$ και 20 είναι προφανές ότι τα αποτελέσματα είναι λανθασμένα. Αυτό που συμβαίνει είναι ότι ακόμα και αν υποθέσουμε ότι το μοναδικό σφάλμα στρογγύλευσης έγινε κατά την προσέγγιση του $1/e$, το σφάλμα αυτό σε κάθε βήμα n πολλαπλασιάζεται επί n . Άρα, π.χ. στο εικοστό βήμα πολλαπλασιάζεται επί τον παράγοντα $20! (= 2\,432\,902\,008\,176\,640\,000)$. Άρα ο αλγόριθμος είναι προφανώς ασταθής, ως αποτέλεσμα του πολλαπλασιασμού $n \cdot I_{n-1}$ κατά τον υπολογισμό του I_n .

Η παρατήρηση αυτή μας οδηγεί να βρούμε έναν ευσταθή αλγόριθμο για το ίδιο πρόβλημα: Ας δοκιμάσουμε να λύσουμε την αναδρομική σχέση (1.10) ως προς I_{n-1} . Δηλαδή

$$I_n = 1 - nI_{n-1} \Rightarrow I_{n-1} = \frac{1 - I_n}{n}.$$

Επομένως, αν γνωρίζαμε κάποια μεταγενέστερη τιμή I_m θα μπορούσαμε να υπολογίσουμε την τιμή I_k , με $k < m$, υπολογίζοντας διαδοχικούς όρους της ακολουθίας $I_{m-1}, I_{m-2}, \dots$ με βάση τον τύπο αυτό. Εμείς γνωρίζουμε από την (1.9) ότι, π.χ., $0 < I_{30} < \frac{1}{31}$. Άρα μπορούμε να επιλέξουμε ως αρχική τιμή του αλγορίθμου μας το $\tilde{I}_{30} = 0$. (Σίγουρα μια καλύτερη επιλογή θα ήταν να επιλέξουμε το μέσον του διαστήματος δηλ. $\tilde{I}_{30} = 1/62$.) Έτσι καταλήγουμε στον ακόλουθο αλγόριθμο:

$$I_{30} = 0, \quad I_{n-1} = \frac{1 - I_n}{n}, \quad n = 30, 29, \dots \quad (1.11)$$

Ο αλγόριθμος (1.11) υλοποιείται στο MATLAB στο αρχείο `integ2.m` που ακολουθεί.

`integ2.m` (Υλοποίηση του αλγορίθμου (1.11) σε MATLAB)

```
clear all;
fid = fopen('res2.txt','w');
x = 0;
nend = 30;
ninit = 13;
for n = nend:-1:ninit
    x = (1-x)/n;
    fprintf(fid,'%3d %15.6f\n',n-1,x);
end
fclose(fid);
```

Εκτελώντας το αρχείο `integ2.m` παίρνουμε τα ακόλουθα αποτελέσματα (που έχουν αποθηκευτεί στο αρχείο `res2.txt`).

n	$\tilde{I}_n$
30	0.
29	0.033333
$\vdots$	$\vdots$
19	0.047723
18	0.050120
17	0.052771
16	0.055719
15	0.059018
14	0.062732
13	0.066948
12	0.071773

Ο αλγόριθμος (1.11) είναι ευσταθής διότι το αρχικό σφάλμα δεν μεγαλώνεται αλλά, αντίθετα, καταστέλλεται αφού σε κάθε βήμα n το σφάλμα μειώνεται κατά $1/n$. Για μια λεπτομερέστερη μελέτη των σφαλμάτων αυτών μπορείτε να ανατρέξετε στο [1, Παράγρ. 1.5].

1.6 Συμπληρωματικές παρατηρήσεις

Για ένα άξιο λόγου παράδειγμα ενός ευσταθούς και ενός ασταθούς αλγορίθμου για τον υπολογισμό του π με τη μέθοδο του Αρχιμήδη, ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει

στο βιβλίο των Ακρίβη-Δουγαλή [1, Παράδ. 1.2, Παράγρ. 1.5]. Πάρα πολλά στοιχεία σχετικά με την αριθμητική κινητής υποδιαστολής και τα σφάλματα στρογγύλευσης μπορεί κανείς να βρει στο βιβλίο του Higham [8], και σε αυτό του Overton [11].

Ένα πολύ ενδιαφέρον θέμα, που δεν θίξαμε καθόλου, αφορά την ευαισθησία πολλών υπολογιστικών προβλημάτων σε μεταβολές των δεδομένων τους. Αυτό είναι ανεξάρτητο των αριθμών κινητής υποδιαστολής και των αλγορίθμων που χρησιμοποιούμε για την επίλυσή τους. Για μια σύντομη σχετική παρουσίαση παραπέμπουμε στο [1, Παράγρ. 1.6] και στο [6, Παράγρ. 2.5, 2.6].

Υπάρχουν κάποια πολύ γνωστά παραδείγματα της καταστροφικής επιρροής των σφαλμάτων στρογγύλευσης και της «απρόσεκτης» χρήσης αριθμητικών αλγορίθμων στην «πραγματική ζωή». Μεταξύ αυτών, η αποτυχία ενός πυραύλου Patriot να εξουδετερώσει έναν πύραυλο Scud στον πρώτο πόλεμο του κόλπου το 1991, και η καταστροφή ενός μη επανδρωμένου ευρωπαϊκού πυραύλου Ariane 5 το 1996. Για περισσότερες πληροφορίες βλ., π.χ.,

<http://www.math.umn.edu/~arnold/disasters/>

Ασκήσεις

1. Πώς θα υπολογίζατε με ακρίβεια τις τιμές της συνάρτησης $f(x) = \sqrt{x+4} - 2$, για πολύ μικρές τιμές του x ; (Υποθέστε ότι οι πράξεις γίνονται με αριθμητική κινητής υποδιαστολής και πεπερασμένη ακρίβεια.)
2. Θεωρήστε έναν υπολογιστή με σύνολο αριθμών μηχανής με παραμέτρους $\beta = 10, t = 3, -L = U = 10$ και αποκοπή. Υπολογίστε την παράσταση

$$A = (345 + 0.7) - (151.499 + 193.701)$$

κάνοντας πράξεις στον παραπάνω υπολογιστή. Πόσο είναι το σχετικό σφάλμα;

3. Για κάποιες τιμές του x οι τιμές της συνάρτησης $f(x) = \sqrt{x^2 + 1} - x$ δεν μπορούν να υπολογιστούν με ακρίβεια μέσω του τύπου αυτού. Ποια είναι τα αίτια; Προτείνετε ένα τρόπο για να ξεπεραστεί αυτή η δυσκολία. (Υποθέστε ότι οι πράξεις γίνονται με αριθμητική κινητής υποδιαστολής και πεπερασμένη ακρίβεια.)
4. Βρείτε έναν κατάλληλο τρόπο υπολογισμού της παράστασης

$$\sin(\alpha + x) - \sin(\alpha),$$

για μικρό $|x|$, έτσι ώστε να μην χάνεται ακρίβεια, όταν οι πράξεις γίνονται με αριθμητική κινητής υποδιαστολής και πεπερασμένη ακρίβεια.

5. Θεωρήστε τη δευτεροβάθμια εξίσωση

$$x^2 - 2ax + b = 0,$$

όπου $a, b > 0$ και $a^2 \gg b$ (το a^2 είναι πολύ μεγαλύτερο του b). Προτείνετε έναν ευσταθή αλγόριθμο για τον υπολογισμό των ριζών της.

6. Θεωρήστε έναν υπολογιστή με σύνολο αριθμών μηχανής με παραμέτρους $\beta = 10, t = 3, -L = U = 10$ και *στρογγύλευση*. Πώς θα υπολογίζατε την παράσταση $\sqrt{9.01} - 3.00$ με όσο το δυνατόν μεγαλύτερη ακρίβεια;
7. Θεωρήστε x, y δύο *αριθμούς μηχανής* με $x \approx y$. Εκτιμήστε το σχετικό σφάλμα κατά την αφαίρεση $x - y$.
8. Θεωρήστε έναν υπολογιστή με σύνολο αριθμών μηχανής με παραμέτρους $\beta = 10, t = 8, -L = U = 100$ και *στρογγύλευση*. Αν x, y, z είναι θετικοί *αριθμοί μηχανής* και με $\oplus$ (όπου $*$ = $+, -, \cdot$ ή $/$) συμβολίσουμε την πράξη $*$ σ' αυτόν τον υπολογιστή, εκτιμήστε το σχετικό σφάλμα για τον υπολογισμό του $x \odot (y \oplus z)$.
9. Εκτιμήστε το σχετικό σφάλμα κατά τον υπολογισμό της ορίζουσας $\Delta = \begin{vmatrix} a & b \\ c & d \end{vmatrix} \neq 0$, όπου a, b, c, d *αριθμοί μηχανής*.
Δώστε παράδειγμα για την οποία το σχετικό σφάλμα μπορεί να γίνει πολύ μεγάλο.
10. Πώς θα υπολογίζατε την ποσότητα $x^3 - y^3$, όπου x, y *αριθμοί μηχανής* με $x \simeq y$; Αιτιολογήστε την απάντησή σας.

Κεφάλαιο 2

Επίλυση μη γραμμικών εξισώσεων

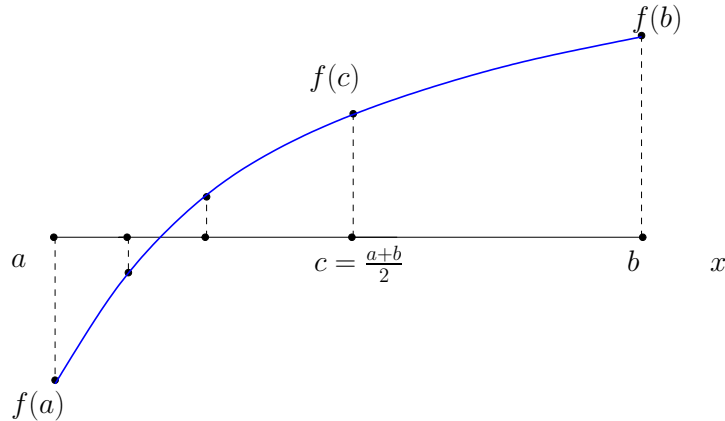
Έστω f μια πραγματική συνάρτηση μιας πραγματικής μεταβλητής. Θα λέμε ότι ένας πραγματικός αριθμός ρ είναι *ρίζα* της εξίσωσης $f(x) = 0$, αν την επαληθεύει, δηλαδή αν $f(\rho) = 0$. Π.χ., όπως εύκολα μπορείτε να επαληθεύσετε η εξίσωση $8x^2 - 10x + 3 = 0$ έχει τις ρίζες $\rho_1 = \frac{1}{2}$ και $\rho_2 = \frac{3}{4}$. Στο κεφάλαιο αυτό θα ασχοληθούμε με αριθμητικές μεθόδους για την εύρεση ριζών της εξίσωσης $f(x) = 0$. Φυσικά, σε κάποιες περιπτώσεις είναι δυνατόν να υπολογίσουμε τις ρίζες μιας εξίσωσης αναλυτικά, για παράδειγμα, υπάρχουν τύποι για τον υπολογισμό ριζών πολωνύμου μέχρι και τετάρτου βαθμού. Στην περίπτωση όμως που πρέπει να λύσουμε μια πολύπλοκη εξίσωση, όπως η $2^{x^2} - 10x + 1 = 0$ ή η $e^x - \sin x - 2 = 0$, δεν είναι δυνατόν να αναμένουμε ότι θα υπάρχουν τύποι για την εύρεση των ριζών της.

Επομένως, χρειαζόμαστε κάποιες γενικές αριθμητικές μεθόδους για τον εντοπισμό ριζών εξισώσεων της μορφής $f(x) = 0$. Ο όρος «γενικές» χρησιμοποιείται για να υποδηλώσει ότι αναζητούμε μεθόδους που θα λαμβάνουν υπ' όψιν γενικές ιδιότητες των συναρτήσεων, όπως η συνέχεια ή η παραγωγισιμότητα, και δεν θα εστιάζουν στη συγκεκριμένη μορφή που θα έχει κάθε φορά η f . Υπάρχουν πολλές μέθοδοι για την αριθμητική επίλυση μη γραμμικών εξισώσεων. Εμείς σε αυτό το κεφάλαιο θα εξετάσουμε αρχικά τη λεγόμενη μέθοδο της διχοτόμησης, στη συνέχεια μια γενική κατηγορία μεθόδων που λέγονται επαναληπτικές μέθοδοι, και, τέλος, τη μέθοδο του Νεύτωνα και τη μέθοδο της τέμνουσας.

2.1 Η μέθοδος της διχοτόμησης

Η μέθοδος της διχοτόμησης είναι η πιο απλή αριθμητική μέθοδος για να προσεγγίσουμε μια ρίζα ρ της εξίσωσης $f(x) = 0$. Βασίζεται στην εξής παρατήρηση: Αν η f είναι μια συνεχής συνάρτηση στο κλειστό διάστημα $[a, b]$ και $f(a)f(b) < 0$, τότε πρέπει να υπάρχει τουλάχιστον ένα $\rho \in (a, b)$ τέτοιο ώστε $f(\rho) = 0$. Με άλλα λόγια, η $f(x) = 0$ πρέπει να έχει

τουλάχιστον μία ρίζα $\rho \in (a, b)$. Αυτό είναι συνέπεια του *Θεωρήματος Ενδιάμεσης Τιμής για Συνεχείς Συναρτήσεις* το οποίο μας εξασφαλίζει ότι αν η f είναι μια συνεχής συνάρτηση στο κλειστό διάστημα $[a, b]$, και αν η είναι ένας πραγματικός αριθμός που βρίσκεται μεταξύ των $f(a)$ και $f(b)$, τότε υπάρχει τουλάχιστον ένα $\xi \in (a, b)$ τέτοιο ώστε $f(\xi) = \eta$.



Σχήμα 2.1: Η μέθοδος της διχοτόμησης.

Η μέθοδος της διχοτόμησης χρησιμοποιεί την ιδέα αυτή ως εξής: Αν f είναι μια συνεχής συνάρτηση στο κλειστό διάστημα $[a, b]$ και $f(a)f(b) < 0$, δηλαδή οι τιμές της f είναι ετερόσημες στα άκρα του διαστήματος, υπολογίζουμε την τιμή της f στο μέσον του διαστήματος $c = \frac{1}{2}(a + b)$, και ελέγχουμε αν ισχύει ότι $f(a)f(c) < 0$. Αν ναι, τότε η $f(x) = 0$ θα έχει μια ρίζα στο διάστημα $[a, c]$, οπότε ονομάζουμε το c ως b και επαναλαμβάνουμε την ίδια διαδικασία με το καινούργιο $[a, b]$, που έχει το μισό μήκος του αρχικού διαστήματος, βλ. Σχήμα 2.1. Αν το $f(a)f(c) > 0$, τότε αναγκαστικά $f(c)f(b) < 0$, και σε αυτή την περίπτωση ονομάζουμε το c ως a . Σε κάθε περίπτωση έχουμε εντοπίσει ένα καινούργιο διάστημα που περιέχει τη ρίζα, και μπορούμε να επαναλάβουμε την ίδια διαδικασία. Φυσικά, στην (σπάνια) περίπτωση που $f(a)f(c) = 0$, τότε $f(c) = 0$ και βρήκαμε μια ρίζα. Σημειώστε ότι όταν οι πράξεις γίνονται στον υπολογιστή είναι κάπως απίθανο λόγω των σφαλμάτων στρογγύλευσης να εντοπίσουμε με αυτό τον τρόπο ένα c ώστε $f(c) = 0$. Επομένως, το κριτήριο τερματισμού της μεθόδου δεν θα πρέπει να είναι κατά πόσον $f(c) = 0$, αλλά, καλύτερα, αν το μήκος του διαστήματος $b - a \leq \text{Tol}$, όπου το Tol είναι μια προκαθορισμένη μικρή σταθερά. Ο αλγόριθμος της διχοτόμησης περιγράφεται στον ακόλουθο ψευδοκώδικα:

```

Είσοδος:  $a, b, \text{Tol}, N_{\max}, f$ 
υπολόγισε  $f(a), f(b)$ 
Αν  $f(a) \cdot f(b) > 0$  τότε έξοδος
μετρητής  $\leftarrow 0$ 
while  $((b - a)/2 > \text{Tol})$ 
    υπολόγισε  $c \leftarrow (a + b)/2, f(c)$ 
    αν  $f(c) = 0$ 
        ρίζα =  $c$ , έξοδος
    διαφορετικά (δηλ. αν  $f(c) \neq 0$ )
        αν  $f(a) \cdot f(c) < 0$ 
             $b \leftarrow c, f(b) \leftarrow f(c)$ 
        διαφορετικά
             $a \leftarrow c, f(a) \leftarrow f(c)$ 
    αν (μετρητής  $\geq N_{\max}$ )
        τύπωσε 'Δεν επετεύχθη σύγκλιση σε  $N_{\max}$  επαναλήψεις',
        έξοδος
    μετρητής  $\leftarrow$  μετρητής + 1
    ρίζα  $\leftarrow (a + b)/2$ , τύπωσε ρίζα

```

Αν ο αλγόριθμος ολοκληρωθεί χωρίς να έχει συμπληρωθεί ο μέγιστος αριθμός επαναλήψεων θα έχει υπολογίσει είτε έναν αριθμό c , που είναι ρίζα της εξίσωσης, ή μια προσέγγιση της ρίζας που θα είναι το μέσον $\frac{a+b}{2}$ του τελικού διαστήματος $[a, b]$. Στην δεύτερη περίπτωση, (αν η f είναι συνεχής) υπάρχει ρίζα ρ που ανήκει στο τελικό διάστημα $[a, b]$, και η προσέγγισή της $\frac{a+b}{2}$ έχει σφάλμα

$$\left| \rho - \frac{a+b}{2} \right| \leq \frac{1}{2}(b-a) \leq \text{Tol}.$$

Ας κάνουμε ορισμένες παρατηρήσεις ώστε να διασαφηνίσουμε κάποια σημεία του ψευδοκώδικα αυτού:

- (1) Με το $\leftarrow$ συμβολίζουμε τον υπολογισμό της παράστασης στο δεξί μέλος και την εκχώρηση της τιμής της στη μεταβλητή που εμφανίζεται στο αριστερό μέλος, όπως π.χ. η ισότητα στην C ή στο MATLAB.
- (2) Αντί να υπολογίζουμε το μέσον του διαστήματος ως $c \leftarrow (a + b)/2$ είναι καλύτερα να το υπολογίζουμε ως $c \leftarrow a + (b - a)/2$. (Γενικά, θα πρέπει να θυμόμαστε ότι στους αριθμητικούς μας υπολογισμούς είναι καλύτερο να υπολογίζουμε μια ποσότητα προσθέτοντας έναν μικρό όρο διόρθωσης στην προηγούμενή μας προσέγγιση.) Μπορείτε

να ανατρέξετε στο βιβλίο των Forsythe, Malcolm και Moler [6, Παράγρ. 7.2] για ένα παράδειγμα στο οποίο το μέσο του διαστήματος $[a, b]$ βρίσκεται εκτός του $[a, b]$, όταν το υπολογίζουμε με τον τύπο $(a + b)/2$ σε μια μηχανή με περιορισμένη ακρίβεια!

- (3) Ο υπολογισμός του γινομένου $f(a) \cdot f(c)$ κατά τον έλεγχο $f(a) \cdot f(c) < 0$ είναι δυνατόν να οδηγήσει σε υπεκχείλιση (underflow) όταν οι αριθμοί $f(a)$ και $f(c)$ είναι πολύ μικροί. (Αυτό είναι κάτι που αναμένουμε να συμβεί όταν πλησιάζουμε κοντά στη ρίζα.) Επομένως είναι καλύτερα είτε να ελέγχουμε αν $\text{sign}(a) \neq \text{sign}(b)$, όπου $\text{sign}(\cdot)$ είναι η συνάρτηση προσήμου ή, εναλλακτικά, αν

$$f(a) \cdot \frac{f(c)}{|f(c)|} < 0.$$

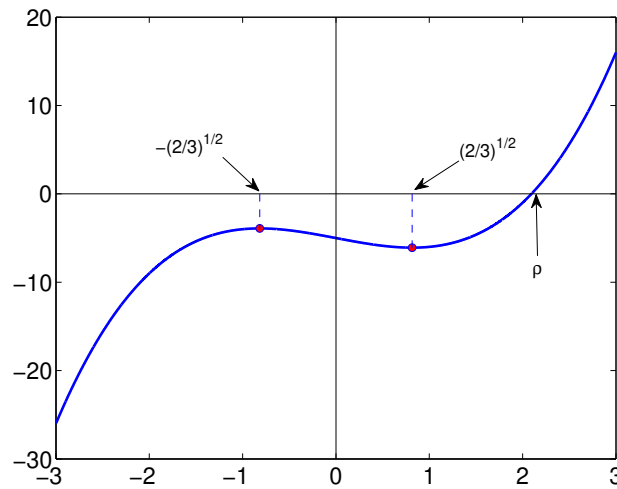
- (4) Στην πράξη, ο έλεγχος $f(c) = 0$ σπάνια θα απαντηθεί θετικά λόγω των σφαλμάτων στρογγύλευσης. Είναι καλύτερο να γίνεται ο έλεγχος $|f(c)| < \varepsilon$, όπου ως ε μπορούμε να πάρουμε το *έψιλον της μηχανής*, βλ. Παράγρ. 1.2.
- (5) Παρατηρήστε ότι στον αλγόριθμο δίνονται τρία κριτήρια τερματισμού. Καταρχάς το $N_{\max}$ είναι το μέγιστο πλήθος βημάτων που θα εκτελεσθούν. Πάντα θα πρέπει να εισάγουμε μια τέτοια δικλείδα ασφαλείας στα προγράμματά μας ώστε να αποφύγουμε την πιθανότητα να οδηγηθούμε σε έναν φαύλο κύκλο. Δεύτερον, ο υπολογισμός θα σταματήσει αν το μήκος του διαστήματος, και επομένως το σφάλμα, γίνει μικρό (μικρότερο του Tol). Το τρίτο κριτήριο τερματισμού ελέγχει αν η τιμή του $f(c)$ γίνει μηδέν. (Όπως είπαμε και στο (4) στην πράξη θα ελέγχουμε αν το $f(c)$ γίνει κατ' απόλυτο τιμή πολύ μικρό.)

Παράδειγμα 2.1. Αναζητούμε ρίζες της εξίσωσης $f(x) = x^3 - 2x - 5 = 0$.*

Στο Σχήμα 2.2 δίνουμε τη γραφική παράσταση της f . Επειδή $f(2) = -1 < 0$ και $f(3) = 16 > 0$, συμπεραίνουμε ότι υπάρχει τουλάχιστον μια ρίζα $\rho \in (2, 3)$. Εύκολα μπορούμε να ελέγξουμε ότι η $f'(x) > 0$, $x \in [2, 3]$, και άρα η ρ είναι η μοναδική ρίζα στο διάστημα $(2, 3)$. Επιπλέον, κάνοντας τη μελέτη της f δεν είναι δύσκολο να δούμε ότι δεν έχει άλλες πραγματικές ρίζες.

Με αρχικό διάστημα το $[2, 3]$ ως κάνουμε μερικές επαναλήψεις της μεθόδου της διχοτόμησης:

*Από το βιβλίο των Forsythe, Malcolm και Moler [6, Παράγρ. 7.2]: «Στο βιβλίο *The Calculus of Observations* των Whittaker και Robinson (1924) υπάρχει το εξής απόσπασμα ενός γράμματος του de Morgan προς τον Whewell (1861): Ο λόγος που ονομάζω την $x^3 - 2x - 5 = 0$ περίφημη εξίσωση είναι γιατί σ' αυτήν έτυχε ο Wallis να εφαρμόσει την μέθοδο του Νεύτωνα όταν την πρωτοδημοσίευσε. Γι' αυτόν τον λόγο, κάθε ένας που υπολογίζει με αριθμητικές μεθόδους αισθάνεται υποχρεωμένος να την συμπεριλάβει στα παραδείγματά του ...»



Σχήμα 2.2: Γραφική παράσταση της $f(x) = x^3 - 2x - 5$.

- 1η προσέγγιση: Το μέσον του διαστήματος είναι το $x_1 = \frac{2+3}{2} = 2.5$. Έχουμε $f(x_1) = f(2.5) = 5.625 > 0$. Αφού $f(2) < 0$ το καινούργιο μας διάστημα είναι το $[2, 2.5]$.
- 2η προσέγγιση: Το μέσον του διαστήματος είναι το $x_2 = \frac{2+2.5}{2} = 2.25$. Έχουμε $f(x_2) = f(2.25) = 1.890625 > 0$. Αφού $f(2) < 0$ το καινούργιο μας διάστημα είναι το $[2, 2.25]$.
- 3η προσέγγιση: Το μέσον του διαστήματος είναι το $x_3 = \frac{2+2.25}{2} = 2.125$. Έχουμε $f(x_3) = f(2.125) = 0.3457031 > 0$. Αφού $f(2) < 0$ το καινούργιο μας διάστημα είναι το $[2, 2.125]$, κ.ο.κ.

Στον ακόλουθο πίνακα δίνονται τα αποτελέσματα της μεθόδου της διχοτόμησης με δεδομένα $a = 2$, $b = 3$, $\text{Tol} = 10^{-6}$. (Η υλοποίηση έγινε με ένα πρόγραμμα MATLAB.)

Επανάλ.	a	b	c	$b - a$	$f(c)$
1	2.0000000	3.0000000	2.5000000	1.0000000	5.6250000
2	2.0000000	2.5000000	2.2500000	0.5000000	1.8906250
3	2.0000000	2.2500000	2.1250000	0.2500000	0.3457031
4	2.0000000	2.1250000	2.0625000	0.1250000	-0.3513184
5	2.0625000	2.1250000	2.0937500	0.0625000	-0.0089417
6	2.0937500	2.1250000	2.1093750	0.0312500	0.1668358
7	2.0937500	2.1093750	2.1015625	0.0156250	0.0785623
⋮	⋮	⋮	⋮	⋮	⋮
16	2.0945435	2.0945740	2.0945587	0.0000305	0.0000807
17	2.0945435	2.0945587	2.0945511	0.0000153	-0.0000044
18	2.0945511	2.0945587	2.0945549	0.0000076	0.0000382
19	2.0945511	2.0945549	2.0945530	0.0000038	0.0000169
20	2.0945511	2.0945530	2.0945520	0.0000019	0.0000062

Παρατηρήστε ότι ο αλγόριθμος σταματά όταν το $(b - a)/2 \leq \text{Tol} = 10^{-6}$, και αυτό συμβαίνει μετά την ολοκλήρωση του εικοστού βήματος, όπου θα έχουμε

$$\frac{1}{2}(b - a) = \frac{1}{2} \cdot 0.0000019 < 10^{-6}.$$

Στο 20ο βήμα η ρίζα έχει εντοπιστεί στο διάστημα $[a, b] = [2.0945511, 2.0945530]$. Άρα η προσέγγισή της $(a + b)/2 = 2.0945520$ ικανοποιεί την ανισότητα

$$\left| \rho - \frac{a + b}{2} \right| \leq \frac{1}{2}(b - a) = \frac{1}{2} \cdot 1.9 \cdot 10^{-6} < \text{Tol} = 10^{-6}.$$

(Η αληθής τιμή της ρίζας είναι: $\rho = 2.0945514815423 \dots$) □

Στη συνέχεια θα εξετάσουμε την ακρίβεια με την οποία η μέθοδος της διχοτόμησης προσεγγίζει μια ρίζα της εξίσωσης $f(x) = 0$.

Πρόταση 2.1. Αν f είναι μια συνεχής συνάρτηση στο $[a, b]$ με $f(a) \cdot f(b) < 0$ και $(c_n)_{n \in \mathbb{N}}$ είναι η ακολουθία των προσεγγίσεων (δηλαδή των μέσων των διαδοχικών διαστημάτων) που παράγει η μέθοδος της διχοτόμησης, τότε είτε $c_N = \rho$ για κάποιο N , ή $c_n \rightarrow \rho$, καθώς $n \rightarrow \infty$, όπου το ρ είναι μια ρίζα της εξίσωσης $f(x) = 0$. Επιπλέον, ισχύει η ακόλουθη εκτίμηση του σφάλματος:

$$|\rho - c_n| \leq \frac{b - a}{2^n}, \quad n = 1, 2, \dots \quad (2.1)$$

Απόδειξη. Θέτουμε $a_1 = a$, $b_1 = b$, και συμβολίζουμε $[a_i, b_i]$, $i = 1, 2, \dots$, τα διαδοχικά διαστήματα που κατασκευάζει η μέθοδος της διχοτόμησης. Ας συμβολίσουμε με c_i το μέσον του

$[a_i, b_i]$, $i = 1, 2, \dots$. Προφανώς κάθε επόμενο διάστημα $[a_{i+1}, b_{i+1}]$ περιέχεται στο προηγούμενο $[a_i, b_i]$. Σε κάθε $[a_i, b_i]$ υπάρχει μια ρίζα, και αν δεν τύχει να συμβεί $c_N = \rho$ για κάποιο N , τότε θα υπάρχουν άπειρα τέτοια διαστήματα που θα περιέχουν τη ρίζα ρ . Βεβαίως ισχύει ότι

$$b_n - a_n = \frac{b_{n-1} - a_{n-1}}{2} = \frac{b_{n-2} - a_{n-2}}{2^2} = \dots = \frac{b_1 - a_1}{2^{n-1}}$$

$$\Rightarrow b_n - a_n = \frac{b - a}{2^{n-1}}.$$

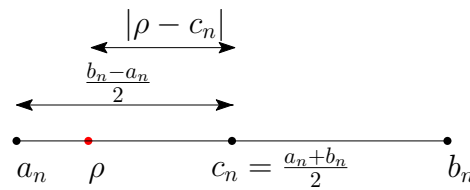
Επίσης

$$\rho \in [a_n, b_n] \quad \text{και} \quad c_n = \frac{1}{2}(a_n + b_n),$$

επομένως, είναι προφανές (βλ. Σχήμα 2.3) ότι

$$|\rho - c_n| \leq \frac{b_n - a_n}{2} = \frac{b - a}{2^n}.$$

□



Σχήμα 2.3: Η μέθοδος της διχοτόμησης: Εκτίμηση σφάλματος.

Η σχέση (2.1) λέμε ότι αποτελεί μια *εκ των προτέρων εκτίμηση του σφάλματος* της μεθόδου, και η Πρόταση 2.1 είναι ένα αποτέλεσμα *σύγκλισης* της μεθόδου.

Η (2.1) μας επιτρέπει να προσδιορίσουμε εκ των προτέρων το πλήθος των βημάτων που απαιτούνται ώστε να προσεγγίσουμε με δεδομένη ακρίβεια μια ρίζα, με τη μέθοδο της διχοτόμησης.

Παράδειγμα 2.2. Έστω ότι στο πρόβλημα του Παραδείγματος 2.1, της εξίσωσης $f(x) = x^3 - 2x - 5 = 0$, με $[a, b] = [2, 3]$, θέλουμε να υπολογίσουμε το πλήθος των βημάτων που απαιτούνται ώστε να προσεγγίσουμε τη ρίζα με σφάλμα μικρότερο του 10^{-5} . Για το σκοπό αυτό θα χρησιμοποιήσουμε την (2.1). Συγκεκριμένα, έχουμε:

$$|\rho - c_n| \leq \frac{3 - 2}{2^n} < 10^{-5} \Leftrightarrow 2^n > 10^5 \Leftrightarrow \log_{10} 2^n > \log_{10} 10^5$$

$$\Leftrightarrow n \log_{10} 2 > 5 \Leftrightarrow n > \frac{5}{\log_{10} 2} \approx 16.61.$$

Δηλαδή αρκεί να εκτελέσουμε 17 βήματα της μεθόδου της διχοτόμησης. (Συγκρίνετε το αποτέλεσμα αυτό με τις τιμές του πίνακα του Παραδείγματος 2.1.) □

Στο σημείο αυτό θα θέλαμε να επιστήσουμε την προσοχή σας στο εξής: Το να εντοπίσει κανείς ένα διάστημα στα άκρα του οποίου μια συνεχής συνάρτηση θα παίρνει ετερόσημες τιμές δεν είναι πάντα μια εύκολη υπόθεση.

Παράδειγμα 2.3. (Από το βιβλίο των Süli και Mayers [12].) Θεωρήστε τη συνάρτηση

$$f(x) = \frac{1}{2} - \frac{1}{1 + M|x - 1.05|}.$$

Προφανώς η εξίσωση $f(x) = 0$ έχει δύο ρίζες, τις $\rho_{1,2} = 1.05 \pm \frac{1}{M}$. Παρατηρήστε ότι η απόσταση των δύο ριζών είναι $2/M$, οπότε αν το M είναι μεγάλο η απόσταση αυτή γίνεται πολύ μικρή.

Με τις ακόλουθες εντολές στο MATLAB μπορούμε να σχεδιάσουμε τη γραφική της παράσταση στο διάστημα $[0.8, 1.8]$ για $M = 200$, βλ. Σχήμα 2.4.

```
>> x = 0.8:0.001:1.8;
>> y = 0.5 - 1./(1 + 200*abs(x-1.05));
>> plot(x,y)
>> grid on
>> xlabel('x')
>> ylabel('y')
```

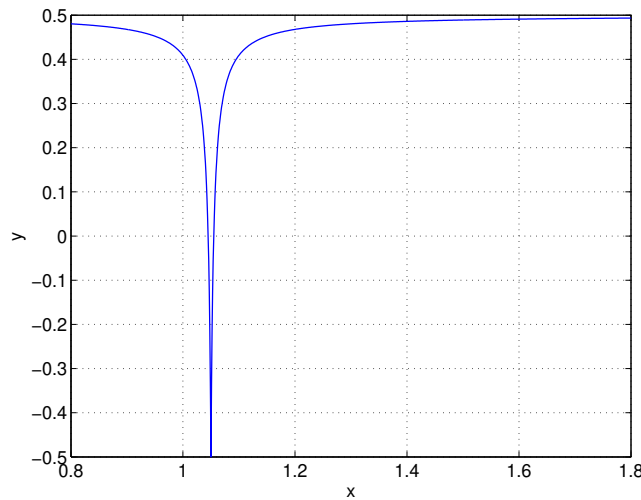
Όπως παρατηρούμε στο Σχήμα 2.4 ο εντοπισμός ενός διαστήματος $[a, b]$ στα άκρα του οποίου οι τιμές της f να είναι ετερόσημες δεν είναι αριθμητικά εύκολος. (Πειραματιστείτε με το MATLAB και προσπαθήστε να κάνετε τις γραφικές παραστάσεις της συνάρτησης αυτής για $M = 2000, 20000, 200000$.)

Κλείνοντας αυτή την παράγραφο ας συνοψίσουμε τα υπέρ και τα κατά της μεθόδου της διχοτόμησης:

Πλεονεκτήματα

- Είναι μια απλή μέθοδος που μπορεί να χρησιμοποιηθεί υπό πολύ γενικές συνθήκες για την f . Απαιτείται μόνο η συνέχεια της f και αλλαγή προσήμου σε μια περιοχή μιας ρίζας.
- Συγκλίνει πάντα και απαιτεί μόνον έναν υπολογισμό της f ανά βήμα.
- Μπορούμε να προσδιορίσουμε εκ των προτέρων το πλήθος των βημάτων που απαιτείται για να προσεγγίσουμε τη ρίζα με δεδομένη ακρίβεια.

Μειονεκτήματα



Σχήμα 2.4: Γραφική παράσταση της $f(x) = \frac{1}{2} - \frac{1}{1+M|x-10.5|}$ για $M = 200$.

- Η μέθοδος συγκλίνει πολύ αργά σε σύγκριση με άλλες μεθόδους. Αυτό άλλωστε φαίνεται από την εκτίμηση (2.1), η οποία υποδεικνύει ότι το σφάλμα σε κάθε βήμα απλώς υποδιπλασιάζεται.
- Έτσι το συνολικό κόστος της μεθόδου, που είναι ανάλογο του αριθμού των υπολογισμών της f είναι σχετικά υψηλό.

2.2 Επαναληπτικές μέθοδοι

Μέχρι στιγμής είδαμε τη μέθοδο της διχοτόμησης για τον υπολογισμό ριζών της εξίσωσης $f(x) = 0$. Όμως κάθε τέτοια εξίσωση μπορεί να αναδιατυπωθεί ως ένα ισοδύναμο πρόβλημα της μορφής $x = \varphi(x)$ κατά πολλούς τρόπους. Σε τέτοιου τύπου αναδιατυπώσεις του προβλήματος βασίζονται οι λεγόμενες επαναληπτικές μέθοδοι.

Ορισμός 2.1. Ένα σημείο x^* του πεδίου ορισμού μιας συνάρτησης φ λέγεται *σταθερό σημείο* της φ , αν $\varphi(x^*) = x^*$.

Στην παράγραφο αυτή θα μελετήσουμε το πρόβλημα εύρεσης σταθερών σημείων μιας συνάρτησης και θα δούμε πώς το πρόβλημα αυτό σχετίζεται με το πρόβλημα εύρεσης ριζών που μας ενδιαφέρει.

Π.χ., αν ενδιαφερόμαστε να εντοπίσουμε μια ρίζα ρ της εξίσωσης $f(x) = 0$ μπορούμε, ισοδύναμα, να γράψουμε $x + f(x) = x$, οπότε αν θέσουμε $\varphi(x) = x + f(x)$, τότε το πρόβλημα

«υπολόγισε μια ρίζα της $f(x) = 0$ » είναι ισοδύναμο με το πρόβλημα «βρες ένα σταθερό σημείο για την φ », αφού

$$\varphi(x) = x \Leftrightarrow x + f(x) = x \Leftrightarrow f(x) = 0.$$

Βεβαίως, κάποιος άλλος θα μπορούσε να επιλέξει τη συνάρτηση $\psi(x) = x - 2f(x)$, οπότε και πάλι $\psi(x) = x \Leftrightarrow x - 2f(x) = x \Leftrightarrow f(x) = 0$.

Γεωμετρικά, το πρόβλημα του εντοπισμού των σημείων τομής του γραφήματος της f με τον άξονα των x αναδιατυπώνεται ως πρόβλημα εντοπισμού των σημείων τομής του γραφήματος της φ με την ευθεία $y = x$ (δηλ. τη διχοτόμο του πρώτου και τρίτου τεταρτημορίου).

Καταρχάς, θα προσπαθήσουμε να εξοικειωθούμε λίγο με την έννοια του σταθερού σημείου και θα δούμε συνθήκες που μας εξασφαλίζουν την ύπαρξη σταθερού σημείου για μια συνάρτηση. Στη συνέχεια θα περιγράψουμε πώς μπορούμε να προσεγγίσουμε σταθερά σημεία με κάποια δεδομένη ακρίβεια.

Παράδειγμα 2.4. Θα βρούμε τα σταθερά σημεία της συνάρτησης $\varphi(x) = x^2 - 6$.

Ένα σταθερό σημείο x^* της φ έχει την ιδιότητα:

$$\varphi(x^*) = x^* \Leftrightarrow (x^*)^2 - 6 = x^* \Leftrightarrow (x^*)^2 - x^* - 6 = 0 \Leftrightarrow (x^* - 3)(x^* + 2) = 0.$$

Επομένως τα σταθερά σημεία της φ είναι $x^* = -2$ και $x^* = 3$, και αυτά είναι οι τετμημένες των σημείων στα οποία το γράφημα της $y = \varphi(x)$ τέμνει την ευθεία $y = x$, βλ. Σχήμα 2.5. $\square$

Στο ακόλουθο θεώρημα δίνονται ικανές συνθήκες για την ύπαρξη και τη μοναδικότητα ενός σταθερού σημείου.

Θεώρημα 2.1. (i) Αν η φ είναι μια συνεχής συνάρτηση στο $[a, b]$ και $\varphi(x) \in [a, b]$ για κάθε $x \in [a, b]$, τότε η φ έχει τουλάχιστον ένα σταθερό σημείο στο $[a, b]$.

(ii) Αν, επιπλέον, η φ είναι παραγωγίσιμη στο (a, b) και υπάρχει μια θετική σταθερά $L < 1$ τέτοια ώστε

$$|\varphi'(x)| \leq L, \quad \text{για κάθε } x \in (a, b),$$

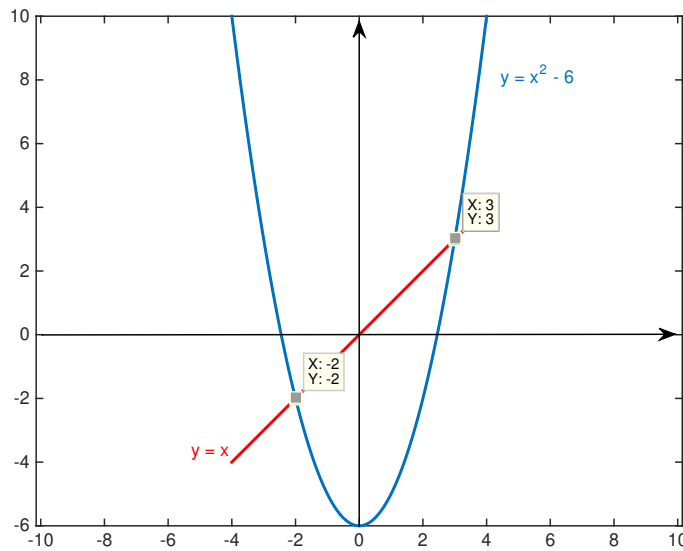
τότε υπάρχει ακριβώς ένα σταθερό σημείο της φ στο $[a, b]$.

Απόδειξη. (i) Αν $\varphi(a) = a$ ή $\varphi(b) = b$, τότε η φ έχει σταθερό σημείο σε ένα από τα άκρα του διαστήματος. Αν όχι, τότε αναγκαστικά $\varphi(a) > a$ και $\varphi(b) < b$. Η συνάρτηση $g(x) = \varphi(x) - x$ είναι συνεχής στο $[a, b]$, με

$$g(a) = \varphi(a) - a > 0 \quad \text{και} \quad g(b) = \varphi(b) - b < 0.$$

Άρα από το θεώρημα ενδιάμεσης τιμής θα υπάρχει τουλάχιστον ένα $x^* \in (a, b)$ τέτοιο ώστε $g(x^*) = 0$. Το σημείο αυτό είναι ένα σταθερό σημείο της φ αφού

$$0 = g(x^*) = \varphi(x^*) - x^* \Rightarrow \varphi(x^*) = x^*.$$



Σχήμα 2.5: Τα σταθερά σημεία της $\varphi(x) = x^2 - 6$.

(ii) Ας υποθέσουμε, επιπλέον, ότι $|\varphi'(x)| \leq L < 1$ και ότι στο $[a, b]$ υπάρχουν δύο σταθερά σημεία x^*, z^* της φ . Αν $x^* \neq z^*$, τότε από το Θεώρημα Μέσης Τιμής έχουμε ότι υπάρχει τουλάχιστον ένα ξ μεταξύ των x^* και z^* , τέτοιο ώστε

$$\frac{\varphi(x^*) - \varphi(z^*)}{x^* - z^*} = \varphi'(\xi).$$

Επομένως,

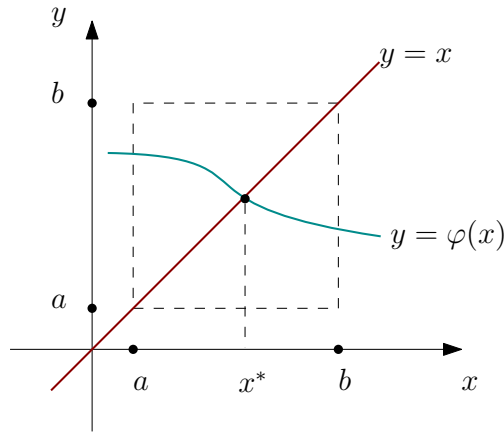
$$|x^* - z^*| = |\varphi(x^*) - \varphi(z^*)| = |\varphi'(\xi)| |x^* - z^*| \leq L |x^* - z^*| < |x^* - z^*|,$$

και έτσι καταλήξαμε σε άτοπο. Το άτοπο αυτό οφείλεται στη μόνη υπόθεση που κάναμε ότι $x^* \neq z^*$. Άρα, αναγκαστικά $x^* = z^*$. □

Παράδειγμα 2.5. Δείξτε ότι η συνάρτηση $\varphi(x) = (x^2 - 4)/5$ έχει μοναδικό σταθερό σημείο στο διάστημα $[-2, 2]$.

Λύση. Η μέγιστη και η ελάχιστη τιμή της $\varphi(x)$, για $x \in [-2, 2]$, θα λαμβάνεται είτε στα άκρα του διαστήματος ή στα σημεία μηδενισμού της παραγώγου της. Προφανώς $\varphi'(x) = 2x/5$, και η φ είναι συνεχής και παραγωγίσιμη στο $[-2, 2]$. Άρα η μέγιστη και η ελάχιστη τιμή της $\varphi(x)$, για $x \in [-2, 2]$, θα λαμβάνεται ή στο $x = -2$ ή στο $x = 0$ είτε στο $x = 2$. Όμως

$$\varphi(-2) = 0, \quad \varphi(0) = -\frac{4}{5}, \quad \varphi(2) = 0.$$



Σχήμα 2.6: Γεωμετρική ερμηνεία του Θεωρήματος 2.1.

Επομένως η ελάχιστη τιμή της $\varphi(x)$, για $x \in [-2, 2]$ είναι $-4/5$ και η μέγιστη τιμή της είναι 0. Δηλαδή, για κάθε $x \in [-2, 2]$: $-\frac{4}{5} \leq \varphi(x) \leq 0$.

Επιπλέον,

$$|\varphi'(x)| = \left| \frac{2x}{5} \right| \leq \frac{4}{5}, \quad \text{για κάθε } x \in (-2, 2).$$

Άρα η φ ικανοποιεί όλες τις υποθέσεις του Θεωρήματος 2.1, επομένως έχει ένα μοναδικό σταθερό σημείο στο $[-2, 2]$. □

Παρατήρηση 2.1. Για τη συνάρτηση του Παραδείγματος 2.5 είναι εύκολο να υπολογίσουμε το σταθερό σημείο της φ αναλυτικά. Πράγματι, αν

$$x^* = \varphi(x^*) = \frac{(x^*)^2 - 4}{5} \Leftrightarrow (x^*)^2 - 5x^* - 4 = 0.$$

Άρα

$$x^* = \frac{5 \pm \sqrt{41}}{2},$$

και έτσι το σταθερό σημείο της φ που ανήκει στο $[-2, 2]$ είναι το $x^* = \frac{5 - \sqrt{41}}{2} \approx -0.702$.

Βεβαίως η φ έχει και μοναδικό σταθερό σημείο στο διάστημα $[5, 6]$ το $x^* = \frac{5 + \sqrt{41}}{2} \approx 5.702$.

Όμως το $\varphi(6) = 6.4 > 6$ και $\varphi'(6) = 2.4 > 1$, επομένως η φ δεν ικανοποιεί τις υποθέσεις του Θεωρήματος 2.1 στο διάστημα $[5, 6]$. Αυτό επιβεβαιώνει ότι οι υποθέσεις του Θεωρήματος 2.1 είναι ικανές, αλλά όχι όμως και αναγκαίες, για την ύπαρξη μοναδικού σταθερού σημείου.

Στις περιπτώσεις που δεν μπορούμε να υπολογίσουμε αναλυτικά το σταθερό σημείο μιας συνάρτησης (πάρτε, π.χ., τη συνάρτηση $\varphi(x) = 2^{-x}$) μπορούμε να υπολογίσουμε μια προσέγγισή του εργαζόμενοι ως εξής:

Ξεκινάμε με μια αρχική προσέγγιση x_0 και κατασκευάζουμε μια ακολουθία προσεγγίσεων $(x_n)_{n=1}^{\infty}$ από τη σχέση:

$$x_n = \varphi(x_{n-1}), \quad n = 1, 2, \dots$$

Αν αυτή η ακολουθία συγκλίνει στο x^* και η φ είναι συνεχής, τότε

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

επομένως το όριο της ακολουθίας αυτής είναι ένα σταθερό σημείο της φ . (Σημειώστε ότι η προτελευταία ισότητα στην παραπάνω σχέση ισχύει επειδή η φ είναι συνεχής.)

Όπως ήδη είπαμε μπορεί κανείς να αναδιατυπώσει το πρόβλημα εύρεσης μιας ρίζας για την f σε πρόβλημα εύρεσης σταθερού σημείου για μια άλλη συνάρτηση με πολλούς τρόπους:

Παράδειγμα 2.6.* Να λυθεί η εξίσωση $x^3 + 4x^2 - 10 = 0$. Η εξίσωση αυτή έχει μια μοναδική ρίζα στο διάστημα $[1, 2]$. Ελέγξτε ότι το πρόβλημα αυτό μπορεί να διατυπωθεί ως πρόβλημα εύρεσης σταθερού σημείου για καθεμιά από τις ακόλουθες συναρτήσεις:

$$(a) \quad x = \varphi_1(x) = x - x^3 - 4x^2 + 10$$

$$(b) \quad x = \varphi_2(x) = \sqrt{\frac{10}{x} - 4x}$$

$$(c) \quad x = \varphi_3(x) = \frac{1}{2}\sqrt{10 - x^3}$$

$$(d) \quad x = \varphi_4(x) = \sqrt{\frac{10}{4 + x}}$$

$$(e) \quad x = \varphi_5(x) = x - \frac{x^3 + 4x^2 - 10}{3x^2 + 8x}$$

Στο βιβλίο των Burden και Faires [3], δίνεται ο ακόλουθος πίνακας με τα αποτελέσματα της επαναληπτικής μεθόδου για καθεμιά από τις παραπάνω συναρτήσεις με αρχική τιμή $x_0 = 1.5$.

*Από το βιβλίο των Burden και Faires, [3].

n	(a)	(b)	(c)	(d)	(e)
0	1.5	1.5	1.5	1.5	1.5
1	-0.875	0.8165	1.286953768	1.348399725	1.373333333
2	6.732	2.9969	1.402540804	1.367376372	1.365262015
3	-469.7	$(-8.65)^{1/2}$	1.345458374	1.364957015	1.365230014
4	1.03×10^8		1.375170253	1.365264748	1.365230013
5			1.360094193	1.365225594	
6			1.367846968	1.365230576	
7			1.363887004	1.365229942	
8			1.365916734	1.365230022	
9			1.364878217	1.365230012	
10			1.365410062	1.365230014	
15			1.365223680	1.365230013	
20			1.365230236		
25			1.365230006		
30			1.365230013		

Η πραγματική ρίζα της εξίσωσης αυτής είναι ίση με 1.365230013. Παρατηρήστε ότι οι επιλογές (c), (d), (e), δίνουν εξαιρετικές προσεγγίσεις, με την (e) να συγκλίνει σε μόλις 4 επαναλήψεις. Η επιλογή (a) οδηγεί σε μια ακολουθία που αποκλίνει, ενώ η επιλογή (b) οδηγεί σε μια ποσότητα που δεν ορίζεται αφού έχουμε την τετραγωνική ρίζα ενός αρνητικού αριθμού. $\square$

Το Παράδειγμα 2.6 μας οδηγεί στο ακόλουθο φυσιολογικό ερώτημα: Πώς μπορούμε να κατασκευάσουμε μια ακολουθία η οποία να συγκλίνει «αξιόπιστα» και όσο το δυνατόν ταχύτερα στη λύση ενός δεδομένου προβλήματος εύρεσης ρίζας μιας συνάρτησης;

Το παρακάτω θεώρημα μας δίνει μια ιδέα για τα βήματα που πρέπει να ακολουθούμε.

Θεώρημα 2.2. Έστω φ μια συνεχής συνάρτηση στο $[a, b]$ τέτοια ώστε $\varphi(x) \in [a, b]$ για κάθε $x \in [a, b]$. Έστω, επιπλέον, ότι η φ είναι παραγωγίσιμη στο (a, b) και ότι υπάρχει μια θετική σταθερά $L < 1$ τέτοια ώστε

$$|\varphi'(x)| \leq L, \quad \text{για κάθε } x \in (a, b).$$

Τότε για κάθε $x_0 \in [a, b]$ η ακολουθία $(x_n)_{n=0}^{\infty}$ με

$$x_n = \varphi(x_{n-1}), \quad n = 1, 2, \dots,$$

συγκλίνει στο μοναδικό σταθερό σημείο x^* της φ στο $[a, b]$.

Επιπλέον, για τα σφάλματα $x_n - x^*$ ισχύουν οι εκτιμήσεις

$$|x_n - x^*| \leq L^n |x_0 - x^*| \leq L^n \max\{x_0 - a, b - x_0\}, \quad (2.2)$$

$$|x_n - x^*| \leq \frac{L^n}{1 - L} |x_1 - x_0|, \quad (2.3)$$

$$|x_n - x^*| \leq \frac{L}{1 - L} |x_n - x_{n-1}|. \quad (2.4)$$

Απόδειξη. Από το Θεώρημα 2.1 έχουμε άμεσα ότι υπάρχει ένα μοναδικό $x^* \in [a, b]$ τέτοιο ώστε $\varphi(x^*) = x^*$. Δεδομένου ότι η φ απεικονίζει το διάστημα $[a, b]$ στον εαυτό του, η ακολουθία $(x_n)_{n=0}^\infty$ θα ορίζεται για κάθε $n \geq 1$, και τα $x_n \in [a, b]$ για όλα τα n .

Στη συνέχεια θα χρησιμοποιήσουμε την υπόθεση ότι $|\varphi'(x)| \leq L$ και το Θεώρημα Μέσης Τιμής για να δείξουμε ότι για κάθε $n \geq 1$ ισχύει:

$$|x_n - x^*| = |\varphi(x_{n-1}) - \varphi(x^*)| = |\varphi'(\xi_n)| |x_{n-1} - x^*| \leq L |x_{n-1} - x^*|,$$

όπου το $\xi_n \in (a, b)$.

Εφαρμόζουμε την ανισότητα αυτή επαγωγικά και έχουμε

$$\begin{aligned} |x_n - x^*| &\leq L |x_{n-1} - x^*| \leq L^2 |x_{n-2} - x^*| \leq \cdots \leq L^n |x_0 - x^*| \\ \Rightarrow |x_n - x^*| &\leq L^n |x_0 - x^*|. \end{aligned} \quad (2.5)$$

Τώρα επειδή $0 < L < 1$, έχουμε ότι η ακολουθία L^n τείνει στο 0, δηλαδή $\lim_{n \rightarrow \infty} L^n = 0$. Επομένως παίρνοντας όρια στην (2.5) έχουμε ότι

$$\lim_{n \rightarrow \infty} |x_n - x^*| \leq \lim_{n \rightarrow \infty} L^n |x_0 - x^*| = 0.$$

Άρα δείξαμε ότι η ακολουθία $(x_n)_{n=0}^\infty$ συγκλίνει στο x^* .

Επειδή το $x^* \in [a, b]$ η εκτίμηση (2.2) προκύπτει άμεσα από την ανισότητα (2.5):

$$|x_n - x^*| \leq L^n |x_0 - x^*| \leq L^n \max\{x_0 - a, b - x_0\}.$$

Επιπλέον, αν ακολουθήσουμε τα ίδια βήματα με αυτά που κάναμε κατά την πορεία της απόδειξης της (2.5) έχουμε ότι

$$|x_{n+1} - x_n| = |\varphi(x_n) - \varphi(x_{n-1})| \leq L |x_n - x_{n-1}| \leq \cdots \leq L^n |x_1 - x_0|.$$

Επομένως, αν $m > n \geq 1$,

$$\begin{aligned} |x_m - x_n| &= |x_m - x_{m-1} + x_{m-1} - x_{m-2} + \cdots + x_{n+1} - x_n| \\ &\leq |x_m - x_{m-1}| + |x_{m-1} - x_{m-2}| + \cdots + |x_{n+1} - x_n| \\ &\leq L^{m-1} |x_1 - x_0| + L^{m-2} |x_1 - x_0| + \cdots + L^n |x_1 - x_0| \\ &= L^n (1 + L + L^2 + \cdots + L^{m-n-1}) |x_1 - x_0|. \end{aligned}$$

Στο πρώτο μέρος της απόδειξης δείξαμε ότι $\lim_{m \rightarrow \infty} x_m = x^*$, συνεπώς

$$\begin{aligned} |x^* - x_n| &= \lim_{m \rightarrow \infty} |x_m - x_n| \leq \lim_{m \rightarrow \infty} \left(L^n (1 + L + L^2 + \cdots + L^{m-n-1}) |x_1 - x_0| \right) \\ &\leq L^n \left(\lim_{m \rightarrow \infty} \frac{1 - L^{m-n}}{1 - L} \right) |x_1 - x_0| \\ &\stackrel{L \leq 1}{=} L^n \frac{1}{1 - L} |x_1 - x_0| = \frac{L^n}{1 - L} |x_1 - x_0|. \end{aligned}$$

Τέλος, για να αποδείξουμε την (2.4) θέτουμε $y_0 := x_{n-1}$ και $y_1 := \varphi(x_{n-1}) = x_n$. Τότε από την (2.3) έχουμε

$$|y_1 - x^*| \leq \frac{L^1}{1-L} |y_1 - y_0|,$$

η οποία αποδεικνύει την (2.4). □

Παρατήρηση 2.2. (i) Η εκτίμηση (2.2) δεν έχει μεγάλη πρακτική σημασία, αφού περιέχει το σταθερό σημείο x^* το οποίο μας είναι άγνωστο.

(ii) Η φύση των εκτιμήσεων (2.3) και (2.4) είναι αρκετά διαφορετική. Η (2.3) μας δίνει πληροφορία για την προσέγγιση x_n χωρίς να χρησιμοποιεί την ίδια τη x_n και για το λόγο αυτό λέγεται εκ των προτέρων (a priori) εκτίμηση. Αντίθετα, στην (2.4) χρησιμοποιείται η προσέγγιση x_n για την οποία δίνει πληροφορία και λέγεται εκ των υστέρων (a posteriori) εκτίμηση.

(iii) Σε όλες τις εκτιμήσεις (2.2)–(2.4) η ταχύτητα με την οποία η ακολουθία $(x_n)_{n=0}^\infty$ συγκλίνει στο σταθερό σημείο x^* εξαρτάται από τον παράγοντα L^n . Όσο μικρότερη είναι η τιμή του L τόσο μεγαλύτερη θα είναι και η ταχύτητα σύγκλισης.

(iv) Αν στο Θεώρημα 2.2 η σταθερά L είναι ίση με 1, τότε η ακολουθία $(x_n)_{n=0}^\infty$, $x_n = \varphi(x_{n-1})$, $n = 1, 2, \dots$, $x_0 \in [a, b]$ δεν συγκλίνει κατ' ανάγκην.

Π.χ., θεωρήστε την $\varphi(x) = -x$, για $x \in [-1, 1]$. (Προφανώς $|\varphi'(x)| = 1$.) Για οποιοδήποτε x_0 πάρουμε στο $[-1, 1]$ με $x_0 \neq 0$, θα προκύψει η ακολουθία $x_0, -x_0, x_0, -x_0, \dots$, η οποία προφανώς δεν συγκλίνει.

Κλείνουμε αυτή την παράγραφο επανερχόμενοι στο ερώτημα που τέθηκε μετά το Παράδειγμα 2.6:

Πώς μπορούμε να κατασκευάσουμε μια ακολουθία η οποία να συγκλίνει «αξιόπιστα» και όσο το δυνατόν ταχύτερα στη λύση ενός δεδομένου προβλήματος εύρεσης ρίζας μιας συνάρτησης; Τώρα είμαστε σε θέση να πούμε ότι μια πιθανή απάντηση στο ερώτημα αυτό είναι η εξής: Μετατρέπουμε το πρόβλημα εύρεσης ρίζας μιας συνάρτησης σε ένα ισοδύναμο πρόβλημα εύρεσης σταθερού σημείου μιας άλλης συνάρτησης που ικανοποιεί τις υποθέσεις του Θεωρήματος 2.2, και έχει παράγωγο που κατ' απόλυτη τιμή είναι όσο το δυνατόν μικρότερη κοντά στο σταθερό σημείο.

2.3 Η μέθοδος του Νεύτωνα

Η μέθοδος του Νεύτωνα (Newton–Raphson) είναι η πιο γνωστή μέθοδος για την επίλυση μη γραμμικών εξισώσεων της μορφής $f(x) = 0$, και προϋποθέτει ότι η συνάρτηση f είναι παραγωγίσιμη. Η μέθοδος του Νεύτωνα στηρίζεται στην εξής (γεωμετρική) ιδέα: Θεωρούμε ένα σημείο $(x_0, f(x_0))$ του γραφήματος της f . Η εφαπτομένη του γραφήματος στο σημείο αυτό

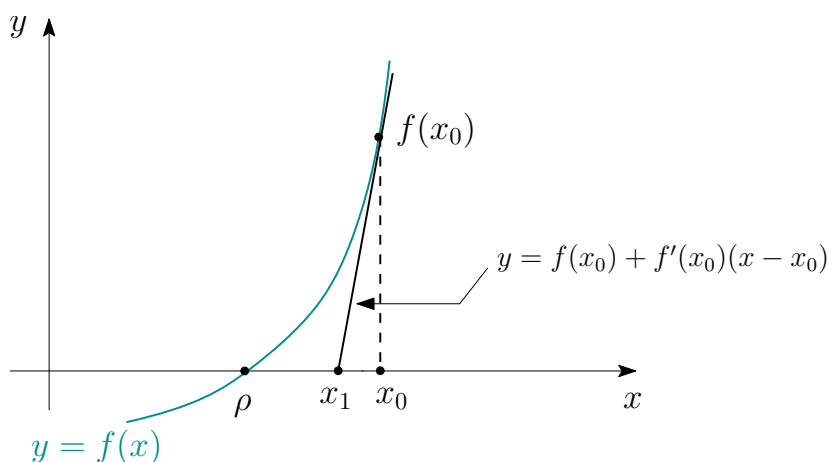
είναι μια αρκετά καλή προσέγγιση της καμπύλης κοντά στο σημείο. Με άλλα λόγια, η γραμμική συνάρτηση

$$t(x) = f(x_0) + f'(x_0)(x - x_0)$$

«μοιάζει» αρκετά με την δεδομένη f για x «κοντά» στο x_0 . (Βεβαίως στο x_0 οι δύο συναρτήσεις παίρνουν την ίδια τιμή $f(x_0)$.) Ας θεωρήσουμε λοιπόν τη ρίζα της εξίσωσης $t(x) = 0$ ως μια προσέγγιση της ρίζας της $f(x) = 0$. Εύκολα βλέπουμε ότι η ρίζα της $t(x) = 0$, δηλαδή το σημείο τομής της εφαπτομένης με τον άξονα των x , είναι ίση με

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)},$$

βλ. Σχήμα 2.7.



Σχήμα 2.7: Η μέθοδος του Νεύτωνα.

Επομένως, ξεκινώντας με ένα αρχικό σημείο x_0 υπολογίζουμε με τον παραπάνω τύπο ένα νέο σημείο x_1 . Επαναλαμβάνοντας τη διαδικασία αυτή υπολογίζουμε μια ακολουθία σημείων

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}, \quad x_3 = x_2 - \frac{f(x_2)}{f'(x_2)}, \quad \dots$$

Άρα:

Η μέθοδος του Νεύτωνα είναι η επαναληπτική μέθοδος που δίνεται από την αναδρομική σχέση

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, 2, \dots, \quad (2.6)$$

υπό την προϋπόθεση ότι $f'(x_n) \neq 0$, για κάθε n , και ότι η x_0 είναι μια δεδομένη αρχική προσέγγιση της ρίζας ρ .

Παρατήρηση 2.3. Η μέθοδος του Νεύτωνα (2.6) είναι μια επαναληπτική μέθοδος με συνάρτηση επανάληψης

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

Παράδειγμα 2.7. Θεωρούμε την εξίσωση $f(x) = x^3 - 2x - 5 = 0$ του Παραδείγματος 2.1. Ας υπολογίσουμε αρχικά τις δύο πρώτες προσεγγίσεις που θα μας δώσει η μέθοδος του Νεύτωνα (2.6), με αρχική τιμή $x_0 = 3$, αν κάνουμε τις πράξεις ακριβώς. Έχουμε ότι $f'(x) = 3x^2 - 2$. Επομένως,

$$\begin{aligned} \text{για } x_0 = 3 : \quad x_1 &= x_0 - \frac{f(x_0)}{f'(x_0)} = 3 - \frac{f(3)}{f'(3)} = 3 - \frac{16}{25} = 2.36, \\ \text{για } x_1 = 2.36 : \quad x_2 &= x_1 - \frac{f(x_1)}{f'(x_1)} = 2.36 - \frac{f(2.36)}{f'(2.36)} \\ &= 2.36 - \frac{3.424256}{14.7088} = 2.12719678. \end{aligned}$$

Ας δούμε τώρα τις προσεγγίσεις της ρίζας $\rho = 2.0945514815423 \dots$ που παράγει η μέθοδος του Νεύτωνα. Η υλοποίηση της μεθόδου έχει γίνει σε MATLAB και σταματάμε τις επαναλήψεις την πρώτη φορά που η διαφορά δύο διαδοχικών προσεγγίσεων $|x_{n+1} - x_n|$ γίνει μικρότερη από $\text{Tol} = 10^{-13}$. Με αρχική τιμή $x_0 = 3$ η μέθοδος δίνει τις προσεγγίσεις που παρατίθενται στη δεύτερη στήλη του Πίνακα 2.1, και συγκλίνει σε μόλις έξι επαναλήψεις δίνοντας την προσέγγιση 2.09455148. Αν θεωρήσουμε ως αρχική τιμή την $x_0 = -2$, παίρνουμε τις προσεγγίσεις που παρατίθενται στην τρίτη στήλη του ίδιου πίνακα. Τώρα, απαιτούνται εννέα επαναλήψεις. Τέλος, αν $x_0 = 0$ παίρνουμε τις προσεγγίσεις της τέταρτης στήλης του Πίνακα 2.1, και παρατηρούμε ότι απαιτούνται αρκετά περισσότερες επαναλήψεις ($= 20$) ώστε η μέθοδος να συγκλίνει. (Αν ανακαλέσετε τη γραφική παράσταση της f στο Σχήμα 2.2 και την συνδυάσετε με τη γεωμετρική ερμηνεία της μεθόδου του Νεύτωνα θα δείτε, όπως άλλωστε υποδεικνύουν οι αριθμοί της τέταρτης στήλης του Πίνακα 2.1, ότι με $x_0 = 0$ η ακολουθία παγιδεύεται, για αρκετά n , γύρω από το σημείο τοπικού μεγίστου της f που είναι το $x = -\sqrt{2/3} \approx -0.8165$.) $\square$

Πίνακας 2.1: Αποτελέσματα της μεθόδου του Νεύτωνα για το Παράδειγμα 2.7 με $x_0 = 3$ (δεύτερη στήλη), $x_0 = -2$ (τρίτη στήλη), και $x_0 = 0$ (τέταρτη στήλη).

Αρχ. τιμή	$x_0 = 3$	$x_0 = -2$	$x_0 = 0$
n	x_n	x_n	x_n
0	3.00000000	-2.00000000	0.00000000
1	2.36000000	-1.10000000	-2.50000000
2	2.12719678	1.43435583	-1.56716418
3	2.09513604	2.61305869	-0.50259245
4	2.09455167	2.20102928	-3.82070647
5	2.09455148	2.10042813	-2.54939339
6	2.09455148	2.09457083	-1.60811150
7		2.09455148	-0.57610043
8		2.09455148	-4.59770958
9		2.09455148	-3.08354315
10			-2.02219426
11			-1.12376411
12			1.20865161
13			3.58079004
⋮			⋮
17			2.09458358
18			2.09455148
19			2.09455148
20			2.09455148

Επομένως, η ενδιαφέρουσα ερώτηση είναι κατά πόσον η ακολουθία της μεθόδου του Νεύτωνα, που ορίζεται μέσω της (2.6), συγκλίνει στη ρίζα ρ της $f(x) = 0$; Η απάντηση δίνεται στο ακόλουθο θεώρημα.

Θεώρημα 2.3. Έστω ότι οι f, f', f'' είναι συνεχείς και ρ είναι μια απλή ρίζα της εξίσωσης $f(x) = 0$ (δηλαδή $f(\rho) = 0$ και $f'(\rho) \neq 0$). Τότε υπάρχει ένα κλειστό διάστημα με μέσον το ρ , τέτοιο ώστε για κάθε αρχική τιμή x_0 στο διάστημα I η ακολουθία (x_n) της μεθόδου του Νεύτωνα συγκλίνει στο ρ . Επιπλέον, ισχύει ότι

$$|x_{n+1} - \rho| \leq C(x_n - \rho)^2. \quad (2.7)$$

Για την απόδειξη βλ. [9, σελ. 66–68].

Όταν ισχύει η (2.7) λέμε ότι η ακολουθία (x_n) συγκλίνει στο ρ (τουλάχιστον) τετραγωνικά. Γενικά, για μια ακολουθία που συγκλίνει στο ρ δίνουμε τον εξής ορισμό:

Ορισμός 2.2. Λέμε ότι η ακολουθία (x_n) συγκλίνει στο ρ (τουλάχιστον) γραμμικά ή ότι η τάξη σύγκλισής της είναι (τουλάχιστον) ένα, αν υπάρχει σταθερά $C < 1$ τέτοια ώστε

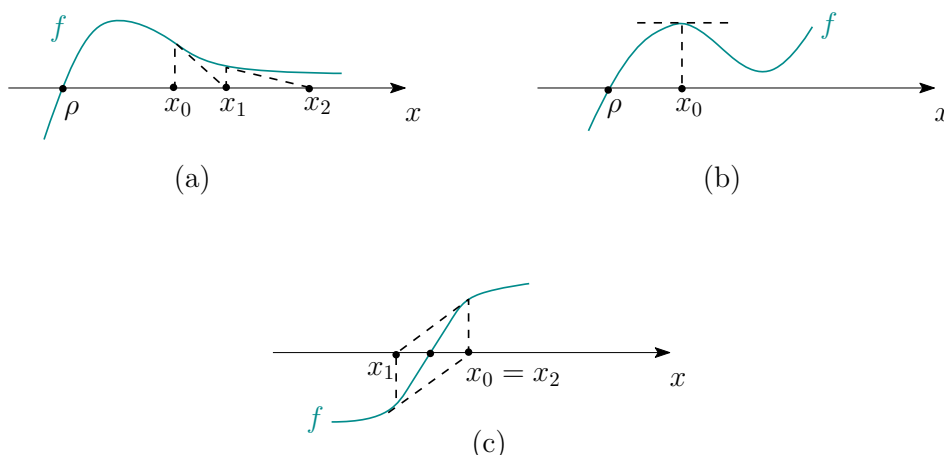
$$|x_{n+1} - \rho| \leq C|x_n - \rho|.$$

Λέμε ότι η τάξη σύγκλισης είναι (τουλάχιστον) p , αν υπάρχει σταθερά $C > 0$ τέτοια ώστε

$$|x_{n+1} - \rho| \leq C|x_n - \rho|^p.$$

Ειδικότερα, για $p = 2, 3$ λέμε ότι η σύγκλιση είναι τετραγωνική, αντίστοιχα, κυβική.

Σημειώστε ότι όταν χρησιμοποιούμε τη μέθοδο του Νεύτωνα απαιτείται ιδιαίτερη προσοχή στην κατάλληλη επιλογή της αρχικής τιμής. Σε κάποιες περιπτώσεις αρκεί απλώς να σχεδιάσουμε μια πρόχειρη γραφική παράσταση της f , αλλά υπάρχουν περιπτώσεις όπου χρειάζεται πολύ προσεκτική ανάλυση της f για να εντοπίσουμε ένα κατάλληλο αρχικό σημείο. Γενικά, η σύγκλιση της μεθόδου εξαρτάται από προϋποθέσεις που δεν είναι εύκολο να ελεγχθούν εκ των προτέρων. Π.χ., στο Σχήμα 2.8 (a), βλέπουμε ότι η εφαπτομένη στο γράφημα της f στο σημείο x_0 τέμνει τον άξονα των x στο σημείο x_1 που βρίσκεται μακρύτερα από τη ρίζα ρ , και επόμενες προσεγγίσεις απομακρύνονται από το ρ αντί να το προσεγγίζουν. Στο Σχήμα 2.8 (b), η συγκεκριμένη επιλογή του x_0 οδηγεί σε μηδενισμό της παραγώγου f' αφού είναι σημείο τοπικού μεγίστου. Άρα το $x_1 = \pm\infty$. Τέλος, στο Σχήμα 2.8 (c), θα έχουμε έναν αέναο κύκλο προσεγγίσεων επειδή $x_2 = x_0$. (Στην πράξη, δηλαδή στον υπολογιστή, τα σφάλματα στρογγύλευσης και η πεπερασμένη ακρίβεια θα επιφέρουν κάποια διαταραχή ώστε είτε οι προσεγγίσεις θα μετατοπιστούν προς τα έξω οπότε η μέθοδος θα αποκλίνει ή προς τα μέσα οπότε τελικά θα συγκλίνει.)



Σχήμα 2.8: Κάποιες περιπτώσεις αποτυχίας της μεθόδου του Νεύτωνα λόγω κακής αρχικής επιλογής (x_0).

Από την άλλη μεριά, υπάρχουν ειδικές περιπτώσεις όπου οι γεωμετρικές ιδιότητες της f (μονοτονία, κυρτότητα) είναι δυνατόν να εξασφαλίσουν τη σύγκλιση της ακολουθίας ακόμα και για x που δεν είναι καθόλου κοντά στη ρίζα. Στην παρακάτω πρόταση (της οποίας την απόδειξη μπορείτε να βρείτε στο [1, Πρωτ. 2.3]) δίνεται ένα τέτοιο παράδειγμα.

Πρόταση 2.2. Έστω ότι οι f, f', f'' είναι συνεχείς για $x \geq a$, και ότι $f'(x) > 0, f''(x) > 0$ για κάθε $x \geq a$. Έστω $f(a) < 0$. Τότε η f έχει μοναδική πραγματική ρίζα ρ στο διάστημα $[a, \infty)$, και η ακολουθία της μεθόδου του Νεύτωνα συγκλίνει στη ρ για οποιοδήποτε $x_0 \geq a$.

Οι υποθέσεις της Πρότασης 2.2 υποδεικνύουν ότι η f είναι γνησίως αύξουσα και κυρτή στο $[a, \infty)$, επομένως το γράφημα της θα συμπεριφέρεται όπως στο Σχήμα 2.7. Με βάση τη γεωμετρική ερμηνεία της μεθόδου του Νεύτωνα επιβεβαιώστε το αποτέλεσμα της πρότασης.

Κλείνουμε την παράγραφο συνοψίζοντας τα υπέρ και τα κατά της μεθόδου του Νεύτωνα:

Πλεονεκτήματα

- Η μέθοδος του Νεύτωνα συγκλίνει ταχύτατα, υπό τις προϋποθέσεις ότι η αρχική προσέγγιση είναι αρκετά κοντά στη ρίζα ρ , ότι η ρίζα είναι απλή (δηλ. $f(\rho) = 0, f'(\rho) \neq 0$), και ότι η f είναι δύο φορές συνεχώς παραγωγίσιμη σε μια περιοχή της ρ .

Μειονεκτήματα

- Δεν είναι πάντα εύκολο να εντοπίσουμε μια καλή αρχική προσέγγιση x_0 .
- Απαιτείται η γνώση τόσο της f όσο και της f' , και υπάρχουν περιπτώσεις όπου η f' δεν είναι γνωστή ή είναι δύσκολο να υπολογιστεί.

- Σε κάθε βήμα της μεθόδου απαιτείται ο υπολογισμός των τιμών δύο συναρτήσεων: της f και της f' .

2.4 Η μέθοδος της τέμνουσας

Ένα από τα μειονεκτήματα της μεθόδου του Νεύτωνα για την επίλυση της $f(x) = 0$, όπως είδαμε στην προηγούμενη παράγραφο, είναι ότι σε κάθε βήμα απαιτείται ο υπολογισμός της παραγώγου f' . Ένας τρόπος για να το ξεπεράσουμε είναι να προσεγγίσουμε την παράγωγο f' που εμφανίζεται στην (2.6) με ένα πηλίκο διαφορών. Συγκεκριμένα,

$$f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}. \quad (2.8)$$

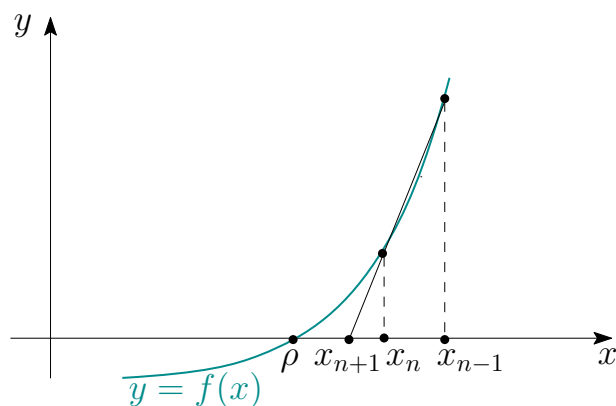
Η ιδέα της προσέγγισης που εμφανίζεται στην (2.8) οφείλεται στον ορισμό της παραγώγου ως όριο, δηλαδή στην

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

Κάνοντας λοιπόν την αντικατάσταση (2.8) στην (2.6) καταλήγουμε στη λεγόμενη μέθοδο της τέμνουσας:

$$x_{n+1} = x_n - \frac{f(x_n)}{\frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}}, \quad n = 1, 2, \dots, \quad (2.9)$$

η οποία προφανώς απαιτεί ως δεδομένα δύο αρχικές τιμές x_0 και x_1 . Παρόλα αυτά, για τον υπολογισμό κάθε νέου x_{n+1} απαιτείται μόνο ένας καινούργιος υπολογισμός της f .



Σχήμα 2.9: Η μέθοδος της τέμνουσας.

Η γεωμετρική ερμηνεία της μεθόδου της τέμνουσας είναι παρόμοια με της μεθόδου του Νεύτωνα. Όπως βλέπουμε στο Σχήμα 2.9, η εφαπτομένη στην καμπύλη αντικαθίσταται από την τέμνουσα.

Η μέθοδος της τέμνουσας είναι λίγο πιο αργή από τη μέθοδο του Νεύτωνα. Μπορεί να αποδειχθεί ότι η μέθοδος της τέμνουσας συγκλίνει σε (απλές) ρίζες με τάξη σύγκλισης $(1 + \sqrt{5})/2 \approx 1.62$ (βλ., π.χ., [1, Θεώρ. 2.3]).

2.5 Συνδυασμοί μεθόδων (υβριδικά σχήματα)

Πολλές φορές για να βρούμε έναν αξιόπιστο και αποτελεσματικό αλγόριθμο για τον υπολογισμό ριζών της εξίσωσης $f(x) = 0$ χρησιμοποιούμε ένα συνδυασμό από αριθμητικές μεθόδους, καταφεύγουμε δηλαδή στις λεγόμενες υβριδικές μεθόδους. Π.χ., μπορούμε να χρησιμοποιήσουμε αρχικά τη μέθοδο της διχοτόμησης κάνοντας μερικές επαναλήψεις για να εντοπίσουμε «χονδρικά» τη ρίζα, και στη συνέχεια να χρησιμοποιήσουμε μια πολύ ταχύτερη μέθοδο, όπως η μέθοδος του Νεύτωνα ή της τέμνουσας ώστε να προσεγγίσουμε ταχύτερα τη ρίζα.

Ένας από τους πιο γνωστούς υβριδικούς αλγορίθμους για των υπολογισμό ριζών λέγεται *zeroin*, βλ. [6, Κεφ. 7] και χρησιμοποιεί ένα συνδυασμό τριών μεθόδων: διχοτόμησης, τέμνουσας και αντίστροφης τετραγωνικής παρεμβολής. (Η τελευταία μέθοδος έχει τάξη σύγκλισης περίπου 1.84, και απαιτεί τρεις αρχικές τιμές.) Ο αλγόριθμος *zeroin* υλοποιείται στο MATLAB μέσω της συνάρτησης *fzero*. Η *fzero* έχει πολλές δυνατότητες: Μπορεί να λάβει ως είσοδο μια τιμή και να αναζητήσει ένα διάστημα στα άκρα του οποίου η f έχει τιμές ετερόσημες. Οι τιμές που επιστρέφει κάθε φορά η f ελέγχονται για NaN,* απειρισμούς και μιγαδικές τιμές. Ο χρήστης μπορεί να αλλάξει το προκαθορισμένο επίπεδο ακρίβειας στην προσέγγιση της ρίζας, καθώς και να πληροφορηθεί κατά την έξοδο για το πλήθος των υπολογισμών των τιμών της συνάρτησης, τον αριθμό των επαναλήψεων, κ.ο.κ.

Η συνάρτηση *fzero* (στην πιο απλή της κλήση) παίρνει δύο ορίσματα: Το πρώτο είναι η συνάρτηση f της οποίας ψάχνουμε ρίζα, και το δεύτερο είναι είτε ένα σημείο κοντά στο οποίο αναζητούμε τη ρίζα ή ένα διάστημα μέσα στο οποίο αναζητούμε τη ρίζα. Σημειώστε ότι μια συνάρτηση περνάει ως όρισμα μιας άλλης συνάρτησης στο MATLAB με δύο τρόπους:

(α) Ως function handle, ή

(β) ως ανώνυμη συνάρτηση (anonymous function),

βλ. Παράρτημα Β'. Εδώ δίνουμε μερικά απλά παραδείγματα χρήσης της *fzero*.

Π.χ., μπορούμε να υπολογίσουμε τη ρίζα της $f(x) = x^3 - 2x - 5 = 0$ (βλ. Παράδειγμα 2.1), με αρχική εκτίμηση το 3, με τον ακόλουθο τρόπο:

*NaN σημαίνει Not a Number και είναι η τιμή που προκύπτει ως αποτέλεσμα μη επιτρεπτών πράξεων, π.χ., $0/0$, $(+\infty) + (-\infty)$, κ.ο.κ.

```
>> format long;
>> f = @(x) x.^3 - 2*x - 5;
>> y = fzero(f,3)
y =
    2.094551481542327
```

Υπολογίζουμε μια ρίζα της $\cos(x)$ στο διάστημα $[1, 2]$ γραφοντας:

```
>> format long;
>> x = fzero(@cos, [1 2])
x =
    1.570796326794897
```

Ασκήσεις

1. Θεωρήστε την εξίσωση $f(x) := x^2 + 3x - 4 = 0$ που έχει ρίζες (προφανώς) $\rho_1 = -4$ και $\rho_2 = 1$. Με αρχικό διάστημα $[a, b] = [-3, 3]$ βρείτε την τρίτη προσέγγιση της ρ_2 , που δίνει η μέθοδος της διχοτόμησης.
2. Αποδείξτε ότι η εξίσωση $f(x) = x^3 - x - 1$ έχει μοναδική πραγματική ρίζα ρ , η οποία μάλιστα βρίσκεται στο διάστημα $[1, 2]$. Με αρχικό διάστημα $[a, b] = [1, 2]$ υπολογίστε τη δεύτερη προσέγγιση της ρ , που δίνει η μέθοδος της διχοτόμησης. Πόσα βήματα της μεθόδου απαιτούνται για τον υπολογισμό μιας προσέγγισης, που απέχει 10^{-6} το πολύ από τη ρ ;

3. Θεωρήστε τη συνάρτηση

$$f(x) = \begin{cases} -1, & 0 \leq x < 0.3, \\ 1, & 0.3 \leq x \leq 1. \end{cases}$$

Συγκλίνει η μέθοδος της διχοτόμησης με $[a, b] = [0, 1]$; Αν ναι, τι δίνει ως «ρίζα» της f ;

4. Βρείτε όλα τα σταθερά σημεία των ακόλουθων συναρτήσεων:

$$(\alpha) f(x) = \frac{8 + 2x}{2 + x^2}, \quad (\beta) g(x) = x^5, \quad (\gamma) h(x) = \frac{6 + 6x^2 - x^3}{11}.$$

5. Για ποιες από τις ακόλουθες συναρτήσεις το $x^* = \sqrt{3}$ είναι σταθερό σημείο;

$$(\alpha) f(x) = \frac{x}{\sqrt{3}}, \quad (\beta) g(x) = \frac{2x}{3} + \frac{1}{x}, \quad (\gamma) h(x) = x^2 - x, \quad (\delta) k(x) = 1 + \frac{2}{x+1}.$$

6. α) Χρησιμοποιήστε το Θεώρημα 2.2 για να αποδείξετε ότι η συνάρτηση

$$g(x) = \frac{1}{3}(x^2 - 1)$$

έχει μοναδικό σταθερό σημείο στο $[-1, 1]$. Ποιο είναι το σταθερό σημείο αυτό;

- β) Χρησιμοποιήστε την εκτίμηση (2.3) του Θεωρήματος 2.2 για να εκτιμήσετε το πλήθος των επαναλήψεων που απαιτούνται ώστε να υπολογίσετε μια προσέγγιση του σταθερού σημείου με την επαναληπτική μέθοδο $x_n = g(x_{n-1})$, $n = 1, 2, \dots$, με ακρίβεια τουλάχιστον 10^{-6} αν ξεκινήσετε με αρχική προσέγγιση $x_0 = 0$.

7. Έστω $x_0 \in [0, 1]$. Να δείξετε ότι η ακολουθία (x_n) που ορίζεται από τον αναδρομικό τύπο

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

συγκλίνει και το όριό της είναι στο $[0, 1]$.

8. Θεωρήστε τη συνάρτηση $g(x) = x - x^3$.

- (α) Δείξτε ότι $x^* = 0$ είναι το μοναδικό σταθερό σημείο της g .
 (β) Θεωρήστε την ακολουθία που ορίζει η επαναληπτική μέθοδος $x_{n+1} = g(x_n)$, $n = 0, 1, 2, \dots$. Δείξτε ότι αν $0 < x_0 < 1$, τότε $x_0 > x_1 > x_2 > \dots > 0$.
 (γ) Δείξτε ότι η ακολουθία (x_n) συγκλίνει στο $x^* = 0$, παρόλο που $g'(0) = 1$. (Υπόδειξη: Χρησιμοποιήστε ότι κάθε μονότονη και φραγμένη ακολουθία συγκλίνει.)

9. Θεωρήστε την εξίσωση $f(x) := x - \cos x = 0$.

- (α) Μελετήστε τη συνάρτηση f για $x \in [0, \pi/3]$ και κάντε τη γραφική της παράσταση. Δείξτε ότι η εξίσωση $f(x) = 0$ έχει μοναδική πραγματική ρίζα $\rho \in (0, \pi/3)$. Δείξτε ότι η $f(x) = 0$ δεν έχει άλλη πραγματική ρίζα.
 (β) Θεωρήστε την ακολουθία (x_n) που ορίζεται από τον αναδρομικό τύπο $x_{n+1} = \cos(x_n)$, $n = 0, 1, 2, \dots$. Δείξτε ότι $x_n \rightarrow \rho$, για κάθε $x_0 \in [0, \pi/3]$, και ότι $\frac{x_{n+1} - \rho}{x_n - \rho} \rightarrow -\sin \rho$.

10. Θεωρήστε την εξίσωση $f(x) := \frac{1}{2}x^2 + \frac{1}{2}x - 1 = 0$ που έχει ρίζες (προφανώς) $\rho_1 = -2$ και $\rho_2 = 1$. Ποια είναι η δεύτερη προσέγγιση που δίνει η μέθοδος του Νεύτωνα για τον υπολογισμό ρίζας της $f(x) = 0$ με $x_0 = 4$;

11. Επαληθεύστε ότι όταν χρησιμοποιούμε τη μέθοδο του Νεύτωνα για την προσέγγιση της $\sqrt{\alpha}$ (θεωρώντας την ως λύση της εξίσωσης $x^2 = \alpha$), η ακολουθία των προσεγγίσεων δίνεται από την επαναληπτική σχέση:

$$x_{n+1} = \frac{1}{2} \left(x_n + \frac{\alpha}{x_n} \right).$$

12. Θεωρήστε την εξίσωση $ax + b = 0$, $a \neq 0$. Αποδείξτε ότι η μέθοδος του Νεύτωνα συγκλίνει στη ρίζα της εξίσωσης αυτής σε ακριβώς μία επανάληψη, για οποιαδήποτε αρχική τιμή x_0 .
13. Αν εφαρμόσουμε τη μέθοδο του Νεύτωνα στην $f(x) = x^5 - x^3 + 3$ και αν $x_n = 1$, βρείτε την επόμενη προσέγγιση x_{n+1} .
14. Ο αντίστροφος ενός αριθμού $a \neq 0$ μπορεί να υπολογιστεί χωρίς διαίρεση από τον αλγόριθμο

$$x_{n+1} = x_n(2 - ax_n), \quad n = 0, 1, 2, \dots$$

- (α) Δείξτε ότι ο αλγόριθμος αυτός προκύπτει από την εφαρμογή της μεθόδου του Νεύτωνα στην εξίσωση

$$\frac{1}{x} - a = 0.$$

- (β) Με $x_0 = 0.3$, υπολογίστε μια προσέγγιση της τιμής του $1/\pi$. Πόσο είναι το σχετικό σφάλμα της προσέγγισης αυτής;

15. Αν εφαρμόσουμε τη μέθοδο της τέμνουσας στην $f(x) = x^3 - 2x + 2$ με αρχικές τιμές $x_0 = 0$ και $x_1 = 1$, τότε ποια είναι η επόμενη προσέγγιση x_2 ;
16. Αν εφαρμόσουμε τη μέθοδο της τέμνουσας στην $f(x) = x^5 + x^3 + 3$ και αν $x_{n-2} = 0$ και $x_{n-1} = 1$, τότε ποια είναι η επόμενη προσέγγιση x_n ;

Κεφάλαιο 3

Γραμμικά συστήματα

Η επίλυση συστημάτων n γραμμικών εξισώσεων, συνήθως με το ίδιο πλήθος αγνώστων, είναι ένα πρόβλημα που συναντάμε πολύ συχνά στις εφαρμογές. Γραμμικά συστήματα τέτοιου τύπου προκύπτουν κατά τη διακριτοποίηση συνεχών προβλημάτων, π.χ., κατά τη διαδικασία επίλυσης διαφορικών και ολοκληρωτικών εξισώσεων, κατά την προσέγγιση συναρτήσεων, κ.α.. Έτσι, οι αλγόριθμοι για την επίλυση γραμμικών συστημάτων αποτελούν βασικό δομικό λίθο πιο σύνθετων εφαρμογών των Υπολογιστικών Μαθηματικών.

Γενικά, ένα σύστημα n γραμμικών εξισώσεων με n αγνώστους γράφεται στη μορφή:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases} \quad (3.1)$$

όπου τα a_{ij} , και b_i , $1 \leq i, j \leq n$, είναι δεδομένοι αριθμοί, και x_i είναι οι άγνωστοι που πρέπει να προσδιορίσουμε ώστε να ικανοποιούνται όλες οι εξισώσεις που εμφανίζονται στην (3.1). Συνήθως τα a_{ij} λέγονται *συντελεστές του συστήματος*, ενώ τα b_i *σταθεροί όροι*.

Το σύστημα (3.1) μπορούμε να το γράψουμε, ισοδύναμα, και στην ακόλουθη μορφή πινάκων

$$Ax = b, \quad (3.2)$$

όπου A είναι ο πίνακας των συντελεστών του συστήματος

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{pmatrix},$$

και x , b τα διανύσματα των αγνώστων και των σταθερών όρων, αντίστοιχα,

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}.$$

Στη συνέχεια δίνουμε κάποιους ορισμούς:

- Θα λέμε *λύση* του συστήματος (3.1) οποιαδήποτε n -αδα αριθμών $(x_1, x_2, \dots, x_n)$ που ικανοποιεί τις εξισώσεις (3.1).
- Ένα σύστημα που έχει τουλάχιστον μία λύση λέγεται *συμβιβαστό*. Αν δεν έχει καμία λύση λέγεται *αδύνατο*.
- Δύο συστήματα που έχουν το ίδιο σύνολο λύσεων λέγονται *ισοδύναμα*.
- Ένα σύστημα της μορφής $Ax = 0$ λέγεται *ομογενές*. (Δηλαδή ομογενές λέγεται το σύστημα που όλοι οι σταθεροί του όροι είναι ίσοι με μηδέν.) Προφανώς ένα ομογενές σύστημα είναι πάντα συμβιβαστό αφού έχει μια τουλάχιστον λύση: τη μηδενική.

Τα συστήματα με τα οποία θα ασχοληθούμε θα έχουν, κατά κανόνα, ακριβώς μία λύση. Ξέρουμε από τη Γραμμική Άλγεβρα ότι οι ακόλουθες είναι ισοδύναμες (ικανές και αναγκαίες) συνθήκες ώστε το γραμμικό σύστημα (3.1) (ή το (3.2)) να έχει ακριβώς μία λύση:

- Ο πίνακας A είναι αντιστρέψιμος (δηλ. υπάρχει ο A^{-1}).
- Η ορίζουσα του A , $\det(A) \neq 0$.
- Το αντίστοιχο ομογενές σύστημα $Ax = 0$, έχει μόνο τη μηδενική λύση ($x = 0$).

Εμείς γνωρίζουμε ήδη δύο τρόπους για την επίλυση του γραμμικού συστήματος (3.2):

(α) Τον κανόνα του Cramer, σύμφωνα με τον οποίο η λύση δίνεται από τον τύπο

$$x_1 = \frac{\det(A^{(1)})}{\det(A)}, \quad x_2 = \frac{\det(A^{(2)})}{\det(A)}, \quad \dots, \quad x_n = \frac{\det(A^{(n)})}{\det(A)},$$

όπου $A^{(j)}$ είναι ο πίνακας που προκύπτει αν αντικαταστήσουμε την j -στήλη του πίνακα A με το διάνυσμα b .

(β) Μπορούμε αρχικά να υπολογίσουμε τον αντίστροφο του πίνακα A και να τον πολλαπλασιάσουμε επί το διάνυσμα b . Τότε $x = A^{-1}b$.

Το πρόβλημα είναι ότι και οι δύο αυτοί τρόποι είναι μάλλον αδύνατον να χρησιμοποιηθούν στην πράξη. Ας υποθέσουμε ότι θέλουμε να λύσουμε ένα σύστημα με $n = 100$ εξισώσεις και αγνώστους, και έστω ότι μας δίνουν τα 10000 στοιχεία του πίνακα A και τα 100 στοιχεία του διανύσματος b , και ότι όλα είναι μη μηδενικά. Πόσο χρόνο πιστεύετε ότι χρειάζεται ένας υπολογιστής για να λύσει το πρόβλημα αυτό με τη μέθοδο του Cramer; Μια ώρα; Ένα μήνα ή ένα χρόνο;

Με μια προσεκτική ανάλυση, βλ. π.χ. [12, §2.1], μπορεί κανείς να διαπιστώσει ότι για τον υπολογισμό των 101 οριζουσών, όπου η καθεμία είναι 100×100 , που απαιτεί ο κανόνας του Cramer, χρειάζεται να εκτελέσουμε περίπου $101!$ αριθμητικές πράξεις (πολλαπλασιασμούς και προσθέσεις). Σημειώστε ότι $101! \approx 2.56 \cdot 10^{160}$. Σήμερα, οι γρηγορότεροι παράλληλοι υπολογιστές έχουν ταχύτητες της τάξης του teraflop/s, δηλαδή μπορούν να εκτελέσουν 10^{12} flops το δευτερόλεπτο. Άρα για να υπολογίσουμε τη λύση του συστήματός μας σ' έναν τέτοιο υπολογιστή θα απαιτηθούν περίπου $2.56 \cdot 10^{160} / 10^{12} = 2.56 \cdot 10^{148}$ δευτερόλεπτα, ή $8.11 \cdot 10^{140}$ χρόνια. (Ως ένα μέτρο σύγκρισης, σημειώστε ότι σύμφωνα με την επικρατούσα θεωρία το σύμπαν δημιουργήθηκε με μια βίαιη έκρηξη, γνωστή ως Big Bang, πριν από περίπου $13.7(\pm 0.2) \cdot 10^9$ χρόνια. Επομένως, είναι μάλλον σαφές ότι χρειαζόμαστε μια πιο αποτελεσματική μέθοδο για να λύσουμε ένα τέτοιο σύστημα.

Βέβαια, η αλήθεια είναι ότι υπάρχουν πολύ αποτελεσματικότεροι τρόποι για τον υπολογισμό οριζουσών, που απαιτούν πολύ λιγότερες πράξεις. Παρόλα αυτά, ο κανόνας του Cramer είναι και πάλι ασύμφορος διότι απαιτεί τον υπολογισμό $n + 1$ οριζουσών.

Επίσης, στην πλειοψηφία των πρακτικών υπολογιστικών προβλημάτων σπάνια χρειαζόμαστε να υπολογίζουμε τον αντίστροφο πίνακα A^{-1} . Ένα ακραίο αλλά διαφωτιστικό παράδειγμα δίνεται στο βιβλίο των Forsythe, Malcolm και Moler [6, §3.0]: Θεωρήστε το σύστημα που αποτελείται από μια μόνο εξίσωση, την

$$7x = 21.$$

Φυσικά, ο καλύτερος τρόπος να λύσουμε το σύστημα είναι άμεσα με μια διαίρεση:

$$x = \frac{21}{7} = 3.$$

Η λύση με χρήση του «αντιστρόφου» πίνακα είναι (με ακρίβεια 6 δεκαδικών ψηφίων):

$$x = (7^{-1})(21) = (0.142857)(21) = 2.999997.$$

Παρατηρήστε ότι η λύση με χρήση του αντιστρόφου πίνακα απαιτεί περισσότερες πράξεις (μια διαίρεση και έναν πολλαπλασιασμό) και οδηγεί σε μια λιγότερο ακριβή λύση. Ανάλογα αποτελέσματα ισχύουν και για μεγαλύτερα συστήματα.

Στο σημείο αυτό είναι σημαντικό να διακρίνουμε δύο κατηγορίες πινάκων:

- (α) Τους πυκνούς ή αποθηκεύσιμους πίνακες, που έχουν μη μηδενικά, εν γένει, στοιχεία, και για την αποθήκευσή τους χρησιμοποιούμε n^2 θέσεις μνήμης, και
- (β) τους αραιούς πίνακες, όπου τα περισσότερα στοιχεία τους είναι μηδενικά. Σε πολλές εφαρμογές οι αραιοί πίνακες έχουν τέτοια δομή ώστε μπορούμε να αποθηκεύσουμε τα μη μηδενικά τους στοιχεία σε λίγες σχετικά θέσεις μνήμης με κατάλληλες δομές δεδομένων.

Επίσης, ανάλογα με την διάσταση του πίνακα A , ένα σύστημα θα λέμε ότι είναι: (α) μικρού μεγέθους αν $n < 100$, (β) μεσαίου μεγέθους αν $100 < n < 1000$, και (γ) μεγάλου μεγέθους αν $n > 1000$. (Οι αριθμοί 100 και 1000 είναι απολύτως ενδεικτικοί, και δεν πρέπει να εκλαμβάνονται ως αυστηρά όρια.)

Στο κεφάλαιο αυτό θα ασχοληθούμε κυρίως με τον αλγόριθμο της απαλοιφής του Gauss, που είναι η πιο γνωστή *άμεση μέθοδος** για την επίλυση γραμμικών συστημάτων, καθώς και με δύο κλασσικές *επαναληπτικές μεθόδους*** επίλυσης γραμμικών συστημάτων.

3.1 Η μέθοδος απαλοιφής του Gauss

Σε αυτή την παράγραφο θα δούμε την απλούστερη εκδοχή της μεθόδου απαλοιφής του Gauss για την επίλυση του γραμμικού συστήματος αλγεβρικών εξισώσεων

$$Ax = b,$$

όπου θα υποθέτουμε ότι ο A είναι ένας $n \times n$ αντιστρέψιμος πίνακας. Γενικά, η απαλοιφή Gauss έχει δύο φάσεις:

- Τη φάση της τριγωνποίησης, όπου μετατρέπουμε μέσω συγκεκριμένων πράξεων τον πίνακα A σε *άνω τριγωνικό*.
- Τη φάση της «οπισθοδρόμησης» (ή προς τα πίσω αντικατάστασης).

Θα παρουσιάσουμε τον αλγόριθμο της απαλοιφής με ένα παράδειγμα τάξης 3.

Θεωρούμε το σύστημα

$$\begin{cases} 10x_1 - 7x_2 &= 7 \\ -3x_1 + 2x_2 + 6x_3 &= 4 \\ 5x_1 - x_2 + 5x_3 &= 6 \end{cases} \quad (3.3)$$

*Μια μέθοδος λέγεται άμεση όταν υπολογίζει τη λύση με ένα πεπερασμένο πλήθος πράξεων.

**Στις επαναληπτικές μεθόδους ξεκινάμε με μια αρχική προσέγγιση $x^{(0)}$ της λύσης και υπολογίζουμε μια ακολουθία διανυσμάτων $x^{(1)}, x^{(2)}, \dots, x^{(k)}, \dots$ που (καλώς εχόντων των πραγμάτων) συγκλίνει στη λύση x καθώς $k \rightarrow \infty$.

Ισοδύναμα, σε μορφή πινάκων, έχουμε

$$\begin{pmatrix} 10 & -7 & 0 \\ -3 & 2 & 6 \\ 5 & -1 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 4 \\ 6 \end{pmatrix} \quad (3.4)$$

Πρώτη φάση: Τριγωνοποίηση.

1ο βήμα: Θα χρησιμοποιήσουμε την πρώτη εξίσωση για να απαλείψουμε το x_1 από τη δεύτερη και την τρίτη εξίσωση. Δηλαδή, θέλουμε να μηδενίσουμε τα στοιχεία του πίνακα που βρίσκονται στην πρώτη στήλη κάτω από το στοιχείο $(1, 1)$, δηλ. κάτω από το 10. Για το σκοπό αυτό:

- Πολλαπλασιάζουμε την πρώτη εξίσωση επί $m_{21} = \frac{-3}{10} = -0.3$ και την αφαιρούμε από τη δεύτερη.
- Πολλαπλασιάζουμε την πρώτη εξίσωση επί $m_{31} = \frac{5}{10} = 0.5$ και την αφαιρούμε από την τρίτη.

(Οι αριθμοί m_{21}, m_{31} λέγονται *πολλαπλασιαστές*.) Έτσι οδηγούμαστε στο ισοδύναμο σύστημα:

$$\begin{pmatrix} 10 & -7 & 0 \\ 0 & -0.1 & 6 \\ 0 & 2.5 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 6.1 \\ 2.5 \end{pmatrix}$$

2ο βήμα: Θα χρησιμοποιήσουμε τη δεύτερη εξίσωση για να απαλείψουμε το x_2 από την τρίτη. Δηλαδή, θέλουμε να μηδενίσουμε τα στοιχεία του πίνακα που βρίσκονται στη δεύτερη στήλη κάτω από το στοιχείο $(2, 2)$, δηλ. κάτω από το -0.1 . Επομένως:

- Πολλαπλασιάζουμε τη δεύτερη εξίσωση επί $m_{32} = \frac{2.5}{-0.1} = -25$ και την αφαιρούμε από την τρίτη.

Έτσι

$$\begin{pmatrix} 10 & -7 & 0 \\ 0 & -0.1 & 6 \\ 0 & 0 & 155 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 6.1 \\ 155 \end{pmatrix}$$

Παρατηρήστε ότι ο τελευταίος πίνακας είναι άνω τριγωνικός. Άρα, ολοκληρώθηκε η πρώτη φάση της απαλοιφής Gauss· η φάση της τριγωνοποίησης.

Δεύτερη φάση: Οπισθοδρόμηση.

- Λύνουμε την τελευταία εξίσωση ως προς x_3 :

$$155x_3 = 155 \Rightarrow x_3 = 1.$$

- Αντικαθιστούμε στη δεύτερη, και λύνουμε ως προς x_2 :

$$-0.1x_2 + 6 \cdot 1 = 6.1 \Rightarrow x_2 = -1.$$

- Τέλος, αντικαθιστούμε στην πρώτη, και λύνουμε ως προς x_1 :

$$10x_1 - 7(-1) + 0 \cdot 1 = 7 \Rightarrow 10x_1 = 0 \Rightarrow x_1 = 0.$$

3.1.1 Αλγόριθμος απαλοιφής Gauss.

Θεωρούμε το γενικό $n \times n$ γραμμικό σύστημα $Ax = b$, για το οποίο υποθέτουμε ότι έχει μοναδική λύση (δηλ. $\det(A) \neq 0$). Συμβολικά, θέτουμε $A^{(1)} = A$ και $b^{(1)} = b$.

Πρώτη φάση: Τριγωνοποίηση.

1ο βήμα: Έστω ότι $a_{11}^{(1)} \neq 0$. (Αν όχι, κάνουμε μια κατάλληλη εναλλαγή γραμμών.) Ορίζουμε τους *πολλαπλασιαστές* m_{i1} , $i = 2, \dots, n$, από τη σχέση

$$m_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}, \quad i = 2, 3, \dots, n,$$

και στη συνέχεια *πολλαπλασιάζουμε* την πρώτη εξίσωση του συστήματος επί m_{i1} και την αφαιρούμε από την i -οστή εξίσωση για όλα τα $i = 2, 3, \dots, n$. Έτσι, μετά το πρώτο βήμα παίρνουμε το (ισοδύναμο με το αρχικό) σύστημα

$$A^{(2)}x = b^{(2)},$$

όπου

$$A^{(2)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \dots & a_{nn}^{(2)} \end{pmatrix} \quad \text{και} \quad b^{(2)} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_n^{(2)} \end{pmatrix}.$$

Σημειώστε ότι:

$$\begin{aligned} a_{ij}^{(2)} &= a_{ij}^{(1)} - m_{i1} a_{1j}^{(1)}, \quad i, j = 1, 2, \dots, n, \\ b_i^{(2)} &= b_i^{(1)} - m_{i1} b_1^{(1)}, \quad i = 1, 2, \dots, n. \end{aligned}$$

2ο βήμα: Στο δεύτερο βήμα απαλείφουμε τα στοιχεία της δεύτερης στήλης του $A^{(2)}$ που βρίσκονται κάτω από το στοιχείο $a_{22}^{(2)}$ (υπό την προϋπόθεση ότι το $a_{22}^{(2)} \neq 0$). Δουλεύοντας όπως πριν, ορίζουμε τους πολλαπλασιαστές $m_{i2}, i = 3, \dots, n$, από τη σχέση

$$m_{i2} = \frac{a_{i2}^{(2)}}{a_{22}^{(2)}}, \quad i = 3, \dots, n,$$

και στη συνέχεια πολλαπλασιάζουμε τη δεύτερη εξίσωση του συστήματος επί m_{i2} και την αφαιρούμε από την i -οστή εξίσωση για όλα τα $i = 3, \dots, n$. Έτσι, μετά το δεύτερο βήμα παίρνουμε το σύστημα

$$A^{(3)}x = b^{(3)},$$

όπου

$$A^{(3)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \dots & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \dots & a_{3n}^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & a_{n3}^{(3)} & \dots & a_{nn}^{(3)} \end{pmatrix} \quad \text{και} \quad b^{(3)} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(3)} \\ \vdots \\ b_n^{(3)} \end{pmatrix}.$$

k -οστό βήμα: Τώρα το σύστημά μας θα είναι της μορφής

$$A^{(k)}x = b^{(k)},$$

όπου

$$A^{(k)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \dots & \dots & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \dots & \dots & \dots & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \dots & \dots & \dots & a_{3n}^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_{kk}^{(k)} & \dots & a_{kn}^{(k)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & a_{ik}^{(k)} & \dots & a_{in}^{(k)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & a_{nk}^{(k)} & \dots & a_{nn}^{(k)} \end{pmatrix} \quad \text{και} \quad b^{(k)} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(3)} \\ \vdots \\ b_k^{(k)} \\ \vdots \\ b_i^{(k)} \\ \vdots \\ b_n^{(k)} \end{pmatrix}.$$

Ορίζουμε και πάλι τους πολλαπλασιαστές

$$m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}, \quad i = k+1, \dots, n,$$

και στη συνέχεια πολλαπλασιάζουμε την k -οστή εξίσωση του συστήματος επί m_{ik} και την αφαιρούμε από την i -οστή εξίσωση για όλα τα $i = k + 1, \dots, n$.

Έτσι καταλήγουμε στο νέο (πάντα ισοδύναμο με το αρχικό) σύστημα

$$A^{(k+1)}x = b^{(k+1)},$$

όπου στον πίνακα $A^{(k+1)}$ έχουμε μηδενίσει τα στοιχεία τα στοιχεία της k -στήλης του $A^{(k)}$ που βρίσκονται κάτω από το στοιχείο $a_{kk}^{(k)}$, και

$$\begin{aligned} a_{ij}^{(k+1)} &= a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)}, & i, j &= k + 1, k + 2, \dots, n, \\ b_i^{(k+1)} &= b_i^{(k)} - m_{ik} b_k^{(k)}, & i &= k + 1, k + 2, \dots, n. \end{aligned}$$

Προχωρώντας ανάλογα μετά το βήμα με $k = n - 1$, καταλήγουμε στο σύστημα

$$A^{(n)}x = b^{(n)},$$

όπου ο $A^{(n)}$ είναι άνω τριγωνικός πίνακας με μη μηδενικά διαγώνια στοιχεία. Δηλαδή έχουμε ολοκληρώσει τη φάση της τριγωνοποίησης του αρχικού συστήματος.

Ο αλγόριθμος της τριγωνοποίησης συνοψίζεται στον ακόλουθο ψευδοκώδικα:

Αλγόριθμος 3.1

Είσοδος: $a_{ij}, i, j = 1, 2, \dots, n, \quad b_i, i = 1, 2, \dots, n$

for $k = 1, \dots, n - 1$ **do**

for $i = k + 1, \dots, n$ **do**

for $j = k, \dots, n$ **do**

$a_{ij} \leftarrow a_{ij} - (a_{ik}/a_{kk}) a_{kj}$

end for

$b_i \leftarrow b_i - (a_{ik}/a_{kk}) b_k$

end for

end for

Δεύτερη φάση: Οπισθοδρόμηση. Από την προηγούμενη φάση είδαμε ότι καταλήξαμε με ένα γραμμικό σύστημα της μορφής

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

όπου, πλέον, τα a_{ij} και τα b_i δεν είναι τα αρχικά, αλλά αυτά που έχουν προκύψει κατά τη διαδικασία της τριγωνποίησης.*

Αυτό το σύστημα λύνεται πολύ εύκολα με οπισθοδρόμηση, δηλαδή από κάτω προς τα πάνω. Λύνουμε την τελευταία εξίσωση και βρίσκουμε το x_n , το αντικαθιστούμε στην προτελευταία και υπολογίζουμε το x_{n-1} , κ.ο.κ. Στο τέλος, αντικαθιστούμε τα $x_2, \dots, x_n$ στην πρώτη εξίσωση και βρίσκουμε το x_1 .

Επομένως, ο αλγόριθμος της οπισθοδρόμησης είναι ο εξής:

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

Για $k = n - 1, n - 2, \dots, 1$:

$$x_k = \frac{1}{a_{kk}} \left(b_k - \sum_{j=k+1}^n a_{kj} x_j \right)$$

Ο αλγόριθμος της οπισθοδρόμησης συνοψίζεται στον ακόλουθο ψευδοκώδικα:

Αλγόριθμος 3.2

Είσοδος: a_{ij} , b_i (που έχουν υπολογισθεί μετά την τριγωνποίηση)

$x_n \leftarrow b_n / a_{nn}$

for $k = n - 1, n - 2, \dots, 1$ **do**

$sum \leftarrow b_k$

for $j = k + 1, \dots, n$ **do**

$sum \leftarrow sum - a_{kj} x_j$

end for

$x_k \leftarrow sum / a_{kk}$

end for

Παρατηρήσεις. Δεν είναι ιδιαίτερα δύσκολο να μετρήσει κανείς το πλήθος των πράξεων που απαιτούνται στην απαλοιφή Gauss. Συγκεκριμένα, αν μετρήσει κανείς ως πράξεις τους πολλαπλασιασμούς και τις διαιρέσεις που εκτελούνται στα στάδια της τριγωνποίησης και της οπισθοδρόμησης μπορεί να διαπιστώσει ότι απαιτούνται

$$\frac{n^3}{3} + O(n^2)$$

πράξεις (βλ. [1, Παράγρ. 3.2.1]). Άρα για να λύσουμε με απαλοιφή Gauss ένα σύστημα 100 εξισώσεων με 100 αγνώστους, απαιτούνται περίπου $\frac{1}{3}100^3 = \frac{1}{3}10^6$ πολλαπλασιασμοί. Συνεπώς

*Άλλωστε, όπως φαίνεται και από τον ψευδοκώδικα, τα $a_{ij}^{(k)}$ και τα $b_i^{(k)}$, που προκύπτουν μετά το $(k-1)$ -οστό βήμα της απαλοιφής τα αποθηκεύουμε στις θέσεις των παλιών στοιχείων.

ένας ταχύτατος υπολογιστής που εκτελεί 10^{12} πολλαπλασιασμούς το δευτερόλεπτο θα μας δώσει τη λύση σε $\frac{1}{3}10^{-6}$ δευτερόλεπτα. Αξίζει να συγκρίνετε αυτή την τιμή με τον αντίστοιχο χρόνο, που αναφέραμε στην εισαγωγή του κεφαλαίου, ότι χρειάζεται ο ίδιος υπολογιστής για να λύσει ένα τέτοιο σύστημα με τον τύπο του Cramer.

3.1.2 Απαλοιφή Gauss με μερική οδήγηση

Τα διαγώνια στοιχεία $a_{kk}^{(k)}$, $k = 1, 2, \dots, n$, του τελικού προϊόντος $A^{(n)}$ της τριγωνοποίησης λέγονται *οδηγοί*. Παρατηρήστε ότι οι οδηγοί εμφανίζονται ως παρονομαστές τόσο στους πολλαπλασιαστές όσο και κατά την οπισθοδρόμηση στον υπολογισμό των x_i . Είναι επομένως λογικό να αναμένουμε προβλήματα (λόγω της πεπερασμένης ακρίβειας στον Η/Υ) όταν κάποιος οδηγός είναι σχεδόν μηδέν, δηλαδή πολύ μικρός κατ' απόλυτη τιμή.

Παράδειγμα. Θεωρήστε το σύστημα

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

Το σύστημα αυτό έχει μοναδική λύση:

$$x_1 = \frac{10000}{9999} = 1.0001\dots, \quad x_2 = \frac{9998}{9999} = 0.9998\dots$$

Το λύνουμε λοιπόν με απαλοιφή Gauss σε έναν Η/Υ με $\beta = 10$, $t = 3$, $U = -L = 20$ και στρογγύλευση. Έχουμε

$$\begin{aligned} m_{21} &= 1/10^{-4} = 10^4, \\ a_{22}^{(2)} &= \text{fl}(1 - m_{21} \cdot 1) = \text{fl}(1 - 10^4) = \text{fl}(-0.9999 \cdot 10^4) = -10^4, \\ b_2^{(2)} &= \text{fl}(2 - m_{21} \cdot 1) = \text{fl}(2 - 10^4) = \text{fl}(-0.9998 \cdot 10^4) = -10^4. \end{aligned}$$

Καταλήξαμε δηλαδή στο σύστημα

$$\begin{cases} 10^{-4}x_1 + x_2 = 1 \\ -10^4x_2 = -10^4 \end{cases}$$

που έχει ως μοναδική λύση $x_1 = 0$ και $x_2 = 1$, η οποία είναι φυσικά μια πολύ κακή προσέγγιση της πραγματικής λύσης.

Εναλλακτικά, προσπαθούμε να κάνουμε τους υπολογισμούς με μικρό πολλαπλασιαστή m_{21} , δηλαδή με όσο το δυνατόν μεγαλύτερο οδηγό.

Επομένως, στην αρχή κάνουμε εναλλαγή γραμμών, και θεωρούμε το σύστημα:

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

Κάνοντας απαλοιφή Gauss στον ίδιο H/Y (όπως πριν) παίρνουμε

$$\begin{cases} x_1 + x_2 = 2 \\ 1 \cdot x_2 = 1 \end{cases}$$

που οδηγεί στην προσεγγιστική λύση $x_1 = 1, x_2 = 1$, η οποία δεδομένης της ακρίβειας που χρησιμοποιήσαμε αποτελεί μια καλή προσέγγιση της πραγματικής λύσης του συστήματος.

Στην πράξη λοιπόν επιδιώκουμε να έχουμε όσο το δυνατόν μεγαλύτερους κατ' απόλυτο τιμή οδηγούς. Αυτό γίνεται σε κάθε βήμα της τριγωνοποίησης και οδηγεί στη λεγόμενη *απαλοιφή Gauss με μερική οδήγηση ή οδήγηση κατά γραμμές*:

Στο k στάδιο της τριγωνοποίησης εξετάζουμε τα στοιχεία $a_{ik}^{(k)}, i = k, k+1, \dots, n$, δηλαδή αυτά που βρίσκονται στην k στήλη, από το διαγώνιο στοιχείο $a_{kk}^{(k)}$ και κάτω, βρίσκουμε ένα από αυτά που έχει τη μεγαλύτερη απόλυτη τιμή και με εναλλαγή γραμμών το φέρνουμε στη θέση του οδηγού.

Παράδειγμα απαλοιφής Gauss με μερική οδήγηση. Θεωρήστε και πάλι το σύστημα

$$\underbrace{\begin{pmatrix} 10 & -7 & 0 \\ -3 & 2 & 6 \\ 5 & -1 & 5 \end{pmatrix}}_{A^{(1)}} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}}_{b^{(1)}} = \underbrace{\begin{pmatrix} 7 \\ 4 \\ 6 \end{pmatrix}}_{b^{(1)}}$$

Ελέγχουμε την πρώτη στήλη. Παρατηρούμε ότι το μεγαλύτερο κατ' απόλυτο τιμή στοιχείο είναι το 10 και βρίσκεται στη θέση $(1, 1)$, άρα προχωρούμε χωρίς εναλλαγές γραμμών, και καταλήγουμε (όπως πριν) στο σύστημα:

$$\underbrace{\begin{pmatrix} 10 & -7 & 0 \\ 0 & -0.1 & 6 \\ 0 & 2.5 & 5 \end{pmatrix}}_{A^{(2)}} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}}_{b^{(2)}} = \underbrace{\begin{pmatrix} 7 \\ 6.1 \\ 2.5 \end{pmatrix}}_{b^{(2)}}$$

Στη συνέχεια, ελέγχουμε τα στοιχεία $a_{22}^{(2)} = -0.1$ και $a_{23}^{(2)} = 2.5$. Το μεγαλύτερο κατ' απόλυτο τιμή είναι το 2.5, οπότε εναλλάσσουμε τις γραμμές 2 και 3.

$$\begin{pmatrix} 10 & -7 & 0 \\ 0 & 2.5 & 5 \\ 0 & -0.1 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 2.5 \\ 6.1 \end{pmatrix}$$

Τώρα θέλουμε να μηδενίσουμε το στοιχείο του πίνακα που βρίσκεται κάτω από το στοιχείο $(2, 2)$, δηλαδή κάτω από το 2.5. Για το σκοπό αυτό πολλαπλασιάζουμε τη δεύτερη εξίσωση επί $m_{32} = \frac{-0.1}{2.5} = -0.04$ και την αφαιρούμε από την τρίτη.

$$\underset{A^{(3)}}{\begin{pmatrix} 10 & -7 & 0 \\ 0 & 2.5 & 5 \\ 0 & 0 & 6.2 \end{pmatrix}} \underset{b^{(3)}}{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}} = \underset{b^{(3)}}{\begin{pmatrix} 7 \\ 2.5 \\ 6.2 \end{pmatrix}}.$$

Άρα ολοκληρώσαμε τη φάση της τριγωνοποίησης και περνάμε στη φάση της οπισθοδρόμησης:

- Λύνουμε την τελευταία εξίσωση ως προς x_3 :

$$6.2 x_3 = 6.2 \Rightarrow x_3 = 1.$$

- Αντικαθιστούμε στη δεύτερη, και λύνουμε ως προς x_2 :

$$2.5 x_2 + 5 \cdot 1 = 2.5 \Rightarrow x_2 = -1.$$

- Αντικαθιστούμε στην πρώτη, και λύνουμε ως προς x_1 :

$$10 x_1 - 7(-1) + 0 \cdot 1 = 7 \Rightarrow 10 x_1 = 0 \Rightarrow x_1 = 0.$$

3.2 Κατάσταση γραμμικών συστημάτων

Κατά την επίλυση ενός γραμμικού συστήματος της μορφής $Ax = b$ (με A αντιστρέψιμο πίνακα) στον υπολογιστή είναι συχνά αναπόφευκτο, λόγω των σφαλμάτων στρογγύλευσης, να υπολογίζουμε μια λύση – που την συμβολίζουμε με $\tilde{x}$ – η οποία θα διαφέρει κάπως από την ακριβή λύση $x = A^{-1}b$. Για παράδειγμα, αν οι συνιστώσες της λύσης x δεν είναι αριθμοί μηχανής, τότε το $\tilde{x}$ δεν μπορεί να είναι ίσο με το x . Για να εκτιμήσουμε τη διαφορά αυτή χρησιμοποιούμε δύο ποσότητες:

α) το σφάλμα: $e = x - \tilde{x}$, και

β) το υπόλοιπο: $r = b - A\tilde{x}$.

Δεν είναι δύσκολο να αποδείξουμε ότι, αφού ο A είναι αντιστρέψιμος, αν το σφάλμα είναι μηδέν, τότε και το υπόλοιπο θα είναι μηδέν, και αντίστροφα. Από την άλλη μεριά όμως δεν είναι απαραίτητα και τα δύο ταυτοχρόνως «μικρά», όπως θα δούμε με το ακόλουθο παράδειγμα. Θεωρήστε το γραμμικό σύστημα

$$\begin{pmatrix} 0.913 & 0.659 \\ 0.780 & 0.563 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0.254 \\ 0.217 \end{pmatrix}.$$

Το σύστημα αυτό έχει μοναδική λύση $x = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

Ας το λύσουμε με απαλοιφή Gauss με μερική οδήγηση σε έναν υποθετικό υπολογιστή με $\beta = 10$, $t = 3$, $U = -L = 20$ και αποκοπή. Έχουμε

$$\begin{aligned} m_{21} &= \frac{0.780}{0.913} = 0.854, \\ a_{22}^{(2)} &= \text{fl}(0.780 - \text{fl}(m_{21} \cdot 0.913)) = \text{fl}(0.780 - 0.779) = 0.001, \\ b_2^{(2)} &= \text{fl}(0.217 - \text{fl}(m_{21} \cdot 0.254)) = \text{fl}(0.217 - 0.216) = 0.001. \end{aligned}$$

Καταλήξαμε δηλαδή στο σύστημα

$$\begin{pmatrix} 0.913 & 0.659 \\ 0 & 0.001 \end{pmatrix} \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix} = \begin{pmatrix} 0.254 \\ 0.001 \end{pmatrix}.$$

Λύνουμε με οπισθοδρόμηση και έχουμε

$$\begin{aligned} \tilde{x}_2 &= \frac{0.001}{0.001} = 1 \\ \tilde{x}_1 &= \text{fl}\left(\frac{0.254 - \text{fl}(0.659x_2)}{0.913}\right) = -0.443. \end{aligned}$$

Δηλαδή χρησιμοποιήσαμε έναν ευσταθή αλγόριθμο και καταλήξαμε σε μια εντελώς ανακριβή προσεγγιστική λύση $\tilde{x}$. (Δεν είναι δύσκολο να ελέγξουμε ότι αν δουλεύαμε σε έναν υπολογιστή με μεγαλύτερη ακρίβεια, $t = 6$, θα παίρναμε την ακριβή λύση $x = (1, -1)^T$.)

Επίσης, αν προσπαθούσαμε να εκτιμήσουμε την ακρίβεια της προσεγγιστικής λύσης, χωρίς να γνωρίζουμε την ακριβή λύση, θα υπολογίζαμε το υπόλοιπο:

$$\begin{aligned} r = b - A\tilde{x} &= \begin{pmatrix} 0.254 - (0.913)(-0.443) - (0.659)(1.00) \\ 0.217 - (0.780)(-0.443) - (0.563)(1.00) \end{pmatrix} \\ &= \begin{pmatrix} -0.000541 \\ -0.000460 \end{pmatrix}, \end{aligned}$$

όπου εδώ εκτελέσαμε τις πράξεις ακριβώς.

Παρατηρήστε ότι οι συνιστώσες του υπολοίπου είναι μικρές, για την ακρίβεια, μικρότερες από 10^{-3} . Αν όμως μετρήσουμε το σφάλμα, έχουμε:

$$e = x - \tilde{x} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} - \begin{pmatrix} -0.443 \\ 1.000 \end{pmatrix} = \begin{pmatrix} 1.443 \\ -2.000 \end{pmatrix},$$

δηλαδή το σφάλμα είναι μεγαλύτερο από την ίδια τη λύση.

Για να δούμε τι συνέβη θα διερευνήσουμε το σύστημά μας λίγο πιο προσεκτικά. Κρατάμε σταθερό τον πίνακα του συστήματος και μεταβάλλουμε λίγο το δεύτερο μέλος. Συγκεκριμένα, θεωρούμε το ελαφρά διαταραγμένο σύστημα

$$\begin{pmatrix} 0.913 & 0.659 \\ 0.780 & 0.563 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0.253 \\ 0.218 \end{pmatrix}.$$

Η ακριβής λύση αυτού του συστήματος είναι $x = \begin{pmatrix} 1223 \\ -1694 \end{pmatrix}$.

Βλέπουμε δηλαδή ότι μια «μικρή» μεταβολή στα δεδομένα του προβλήματος προκαλεί μια «μεγάλη» μεταβολή στη λύση του προβλήματος. Άρα το πρόβλημα αυτό είναι πολύ ευαίσθητο σε μικρές μεταβολές των δεδομένων του, και για το λόγο αυτό λέμε ότι έχει *κακή κατάσταση*.

Ένα ακόμη θέμα στο οποίο πρέπει να σταθούμε λίγο παραπάνω είναι το εξής: Όπως είδαμε στο προηγούμενο παράδειγμα το υπόλοιπο r που υπολογίσαμε ήταν πολύ μικρό αν και η προσεγγιστική λύση ήταν πολύ ανακριβής. Αυτό οφείλεται σε σύμπτωση ή είναι κάτι γενικότερο; Η απάντηση, όπως μπορεί κανείς να αποδείξει, είναι ότι η απαλοιφή Gauss με μερική οδήγηση παράγει πάντα προσεγγιστικές λύσεις με μικρό υπόλοιπο. Όπως όμως είδαμε, ένα μικρό υπόλοιπο r δεν μας εξασφαλίζει ότι και το σφάλμα e είναι μικρό. Η σχέση μεταξύ του υπολοίπου και του σφάλματος καθορίζεται από μια ποσότητα που λέγεται *δείκτης κατάστασης* του πίνακα, όπως θα δούμε στη συνέχεια.

3.3 Νόρμες διανυσμάτων και πινάκων

Για να μελετήσουμε την κατάσταση γραμμικών συστημάτων πρέπει να μπορούμε να μετράμε αποστάσεις μεταξύ διανυσμάτων και πινάκων. Για το σκοπό αυτό εισάγουμε, καταρχάς, την έννοια της νόρμας ενός διανύσματος, που είναι ένας αριθμός που εκφράζει (εν γένει) το μέγεθος των στοιχείων ενός διανύσματος. Γενικότερα, μπορούμε να ορίσουμε νόρμες σε οποιονδήποτε διανυσματικό χώρο, αλλά εμείς θα περιοριστούμε κυρίως στον $\mathbb{R}^n$. (Εντελώς ανάλογα ορίζονται νόρμες στον $\mathbb{C}^n$.)

Ορισμός 3.1. Μια διανυσματική νόρμα είναι μια απεικόνιση από τον $\mathbb{R}^n$ στο $\mathbb{R}$, η οποία ικανοποιεί τις ακόλουθες τρεις ιδιότητες:

- (1) $\|x\| = 0 \Leftrightarrow x = 0$,
- (2) $\|\lambda x\| = |\lambda| \|x\|$, για κάθε αριθμό λ ,
- (3) $\|x + y\| \leq \|x\| + \|y\|$, (τριγωνική ανισότητα),

όπου $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ είναι διανύσματα του $\mathbb{R}^n$ και $\lambda \in \mathbb{R}$.

Παραδείγματα νορμών: Έστω $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$. Τότε οι ακόλουθες είναι νόρμες

$$\begin{aligned} \|x\|_1 &= \sum_{i=1}^n |x_i|, & \ell_1\text{-νόρμα}, \\ \|x\|_\infty &= \max_{1 \leq i \leq n} |x_i|, & \ell_\infty\text{-νόρμα ή νόρμα μεγίστου}, \\ \|x\|_2 &= \left(\sum_{i=1}^n |x_i|^2 \right)^{1/2} & \ell_2\text{-νόρμα ή Ευκλείδεια νόρμα}. \end{aligned} \tag{3.5}$$

Παράδειγμα. Έστω $x = \begin{pmatrix} -3 \\ 1 \\ -2 \end{pmatrix}$. Τότε

$$\begin{aligned} \|x\|_1 &= \sum_{i=1}^3 |x_i| = |-3| + |1| + |-2| = 3 + 1 + 2 = 6. \\ \|x\|_\infty &= \max_{1 \leq i \leq 3} |x_i| = \max\{|-3|, |1|, |-2|\} = 3. \\ \|x\|_2 &= \left(\sum_{i=1}^3 |x_i|^2 \right)^{1/2} = \sqrt{(-3)^2 + 1^2 + (-2)^2} = \sqrt{14} \approx 3.742. \end{aligned}$$

□

Παρατήρηση 3.1. Δύο άμεσες συνέπειες του Ορισμού 3.1 είναι οι ακόλουθες: Αν $x, y \in \mathbb{R}^n$,

τότε

$$\|x\| \geq 0, \quad (3.6)$$

$$|\|x\| - \|y\|| \leq \|x \pm y\|. \quad (3.7)$$

Η ιδιότητα (3.6) εκφράζει το γεγονός ότι η νόρμα ενός διανύσματος είναι πάντα ένας μη αρνητικός αριθμός, ενώ η (3.7) συμπληρώνει την τριγωνική ανισότητα «προς τα κάτω».

Στη συνέχεια θα εισάγουμε νόρμες πινάκων. Έστω $\mathbb{R}^{n \times n}$ το σύνολο των τετραγωνικών $n \times n$ πινάκων με στοιχεία πραγματικούς αριθμούς.

Ορισμός 3.2. Μια νόρμα πινάκων είναι μια απεικόνιση από τον $\mathbb{R}^{n \times n}$ στο $\mathbb{R}$, η οποία ικανοποιεί τις ακόλουθες τρεις ιδιότητες:

- (1) $\|A\| = 0 \Leftrightarrow A = 0$,
- (2) $\|\lambda A\| = |\lambda| \|A\|$, για κάθε αριθμό λ ,
- (3) $\|A + B\| \leq \|A\| + \|B\|$, (τριγωνική ανισότητα),

και, επιπλέον, πληροί την ιδιότητα:

$$(4) \quad \|AB\| \leq \|A\| \|B\|,$$

όπου $A, B \in \mathbb{R}^{n \times n}$ και $\lambda \in \mathbb{R}$.

Συνήθως χρησιμοποιούμε μια ειδική κατηγορία νορμών πινάκων που λέγονται *φυσικές νόρμες* και συνδέονται με μια διανυσματική νόρμα.

Ορισμός 3.3. Έστω $\|\cdot\|$ μια νόρμα στον $\mathbb{R}^n$. Τότε η αντίστοιχη *φυσική νόρμα* πινάκων ή νόρμα πινάκων που παράγεται από τη νόρμα $\|\cdot\|$ του $\mathbb{R}^n$ ορίζεται ως η απεικόνιση

$$\|\cdot\| : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}, \quad \|A\| = \sup_{0 \neq x \in \mathbb{R}^n} \frac{\|Ax\|}{\|x\|}. \quad (3.8)$$

Παρατήρηση 3.2. Έστω X ένα υποσύνολο του $\mathbb{R}$. Τότε ως *supremum* του συνόλου X ορίζουμε το ελάχιστο άνω φράγμα του X , και το συμβολίζουμε ως $\sup X$. Συγκεκριμένα, ένας αριθμός $\beta = \sup X$ αν για κάθε $x \in X$ ισχύει $x \leq \beta$ και για κάθε $\beta' < \beta$ υπάρχει ένα $x' \in X$ τέτοιο ώστε $x' > \beta'$.

Π.χ., $\sup[a, b] = b$, $\sup(a, b) = b$, $\sup\{x \in \mathbb{Q} : x^2 < 2\} = \sqrt{2}$.

Πρόταση 3.1. Αν $\|\cdot\|$ είναι μια φυσική νόρμα πίνακα, τότε ισχύουν οι ακόλουθες ιδιότητες:

$$\|I\| = 1, \text{ όπου } I \text{ είναι ο μοναδιαίος } n \times n \text{ πίνακας.} \quad (3.9)$$

$$\|Ax\| \leq \|A\| \|x\|, \text{ για κάθε } x \in \mathbb{R}^n. \quad (3.10)$$

Παρατηρήστε ότι στην (3.10) το σύμβολο $\|\cdot\|$ χρησιμοποιείται με δύο διαφορετικές έννοιες, τόσο για να συμβολίσουμε νόρμα διανύσματος όσο και για να συμβολίσουμε νόρμα πίνακα.

Παραδείγματα φυσικών νορμών πινάκων:

1. Θεωρούμε στον $\mathbb{R}^n$ τη νόρμα μεγίστου $\|\cdot\|_\infty$. Η φυσική νόρμα στον $\mathbb{R}^{n \times n}$ που παράγεται από την $\|\cdot\|_\infty$ είναι η λεγόμενη «νόρμα του αθροίσματος γραμμών» και για $A \in \mathbb{R}^{n \times n}$ ορίζεται ως

$$\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|. \quad (3.11)$$

2. Θεωρούμε στον $\mathbb{R}^n$ την ℓ_1 -νόρμα $\|\cdot\|_1$. Η φυσική νόρμα στον $\mathbb{R}^{n \times n}$ που παράγεται από την $\|\cdot\|_1$ είναι η λεγόμενη «νόρμα του αθροίσματος στηλών» και για $A \in \mathbb{R}^{n \times n}$ ορίζεται ως

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}|. \quad (3.12)$$

3. Θεωρούμε στον $\mathbb{R}^n$ την Ευκλείδεια νόρμα $\|\cdot\|_2$. Η φυσική νόρμα στον $\mathbb{R}^{n \times n}$ που παράγεται από την $\|\cdot\|_2$ είναι η λεγόμενη «φασματική νόρμα» και για $A \in \mathbb{R}^{n \times n}$ ορίζεται ως

$$\|A\|_2 = \sqrt{\rho(A^T A)}, \quad (3.13)$$

όπου για έναν πίνακα $B \in \mathbb{R}^{n \times n}$ συμβολίζουμε με $\rho(B)$ την λεγόμενη φασματική ακτίνα του B , που ορίζεται ως το μέγιστο των απολύτων τιμών των ιδιοτιμών του πίνακα B .

Παράδειγμα. Θεωρούμε τον πίνακα $A = \begin{pmatrix} 1 & -1 \\ -3 & 2 \end{pmatrix}$. Τότε

$$\|A\|_\infty = \max\{|1| + |-1|, |-3| + |2|\} = \max\{2, 5\} = 5.$$

$$\|A\|_1 = \max\{|1| + |-3|, |-1| + |2|\} = \max\{4, 3\} = 4.$$

Για να βρούμε την $\|A\|_2$, αρχικά υπολογίζουμε τον πίνακα $A^T A$.

$$A^T A = \begin{pmatrix} 1 & -3 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ -3 & 2 \end{pmatrix} = \begin{pmatrix} 10 & -7 \\ -7 & 5 \end{pmatrix}.$$

Στη συνέχεια πρέπει να υπολογίσουμε τις ιδιοτιμές του πίνακα $A^T A$, οπότε θεωρούμε τη χαρακτηριστική εξίσωση:

$$\begin{aligned} |(A^T A) - \lambda I| = 0 &\Leftrightarrow \begin{vmatrix} 10 - \lambda & -7 \\ -7 & 5 - \lambda \end{vmatrix} = 0 \Leftrightarrow (10 - \lambda)(5 - \lambda) - 49 = 0 \\ &\Leftrightarrow \lambda^2 - 15\lambda + 1 = 0 \Leftrightarrow \lambda = \frac{15 + \sqrt{221}}{2} \text{ ή } \lambda = \frac{15 - \sqrt{221}}{2}. \end{aligned}$$

Άρα οι ιδιοτιμές του πίνακα $A^T A$ είναι οι $\lambda_1 = \frac{15+\sqrt{221}}{2}$ και $\lambda_2 = \frac{15-\sqrt{221}}{2}$, επομένως

$$\rho(A^T A) = \frac{15 + \sqrt{221}}{2} \simeq 14.933.$$

Συνεπώς

$$\|A\|_2 = \sqrt{\rho(A^T A)} = \sqrt{\frac{15 + \sqrt{221}}{2}} \simeq 3.864.$$

□

3.4 Δείκτης κατάστασης πίνακα

Ας επιστρέψουμε τώρα στο πρόβλημα της κατάστασης των γραμμικών συστημάτων. Μια σημαντική ποσότητα που αποτελεί έναν δείκτη της «ευαισθησίας» της λύσης x του συστήματος $Ax = b$ σε διαταραχές των δεδομένων $A \in \mathbb{R}^{n \times n}$ και $b \in \mathbb{R}^n$ είναι ο λεγόμενος δείκτης κατάστασης του πίνακα A (ο ορισμός του οποίου οφείλεται στον A. Turing^{*}).

Ορισμός 3.4. Ορίζουμε ως *δείκτη κατάστασης του πίνακα A* , και συμβολίζουμε με $\kappa(A)$, την ποσότητα:

$$\kappa(A) = \|A\| \|A^{-1}\|. \quad (3.14)$$

Φυσικά, η ακριβής τιμή του $\kappa(A)$ εξαρτάται από τη συγκεκριμένη νόρμα που επιλέγουμε κάθε φορά. Συνήθως, όμως, ενδιαφερόμαστε για μια εκτίμηση του μεγέθους του δείκτη κατάστασης, οπότε η επιλογή κάποιας συγκεκριμένης νόρμας δεν παίζει ιδιαίτερο ρόλο. Επιπλέον, στην πράξη δεν είναι αναγκαίο να υπολογίσουμε τον αντίστροφο ενός πίνακα A για να βρούμε μια εκτίμηση του δείκτη κατάστασης.

Παρατηρήστε ότι από τον ορισμό του δείκτη κατάστασης συμπεραίνουμε εύκολα ότι

$$\kappa(A) \geq 1.$$

Πράγματι, έχουμε:

$$1 = \|I\| = \|AA^{-1}\| \leq \|A\| \|A^{-1}\| = \kappa(A).$$

Ας εξετάσουμε τώρα την απλούστερη περίπτωση στο πρόβλημα της κατάστασης των γραμμικών συστημάτων. Θεωρούμε το σύστημα

$$Ax = b,$$

^{*}A.M. Turing (1912 – 1954): Άγγλος μαθηματικός, καθηγητής της λογικής, κρυπτογράφος και θεωρητικός βιολόγος.

(όπου ο A είναι ένας αντιστρέψιμος πίνακας), και το σύστημα που προκύπτει αν διαταράξουμε το δεύτερο μέλος:

$$A(x + \delta x) = b + \delta b.$$

Αφαιρώντας κατά μέλη καταλήγουμε στο σύστημα:

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

Επομένως, αν συμβολίσουμε $\|\cdot\|$ μια νόρμα στον $\mathbb{R}^n$, καθώς και την αντίστοιχη φυσική νόρμα πίνακα που παράγει, τότε έχουμε

$$Ax = b \Rightarrow \|b\| = \|Ax\| \stackrel{(3.10)}{\leq} \|A\| \|x\| \Rightarrow \frac{1}{\|x\|} \leq \|A\| \frac{1}{\|b\|},$$

$$\text{και } \delta x = A^{-1} \delta b \Rightarrow \|\delta x\| = \|A^{-1} \delta b\| \stackrel{(3.10)}{\leq} \|A^{-1}\| \|\delta b\| \Rightarrow \|\delta x\| \leq \|A^{-1}\| \|\delta b\|.$$

Πολλαπλασιάζουμε κατά μέλη τις δύο αυτές σχέσεις και έχουμε:

$$\begin{aligned} \frac{\|\delta x\|}{\|x\|} &\leq \|A\| \|A^{-1}\| \frac{\|\delta b\|}{\|b\|} \\ \Leftrightarrow \frac{\|\delta x\|}{\|x\|} &\leq \kappa(A) \frac{\|\delta b\|}{\|b\|}. \end{aligned}$$

Το κλάσμα $\|\delta b\|/\|b\|$ είναι η σχετική διαταραχή που επιφέραμε στο δεύτερο μέλος, και το κλάσμα $\|\delta x\|/\|x\|$ είναι το σχετικό σφάλμα εξαιτίας της διαταραχής αυτής. Παρατηρήστε ότι ο δείκτης κατάστασης είναι ένας συντελεστής μεγέθυνσης της σχετικής διαταραχής. (Το ίδιο συμβαίνει και αν διαταράξουμε τα στοιχεία του πίνακα.)

Επιπλέον, ο δείκτης κατάστασης ενός πίνακα είναι ένα μέτρο του πόσο κοντά βρίσκεται ο πίνακας αυτός στο σύνολο των μη αντιστρέψιμων πινάκων. Γενικά, μπορούμε να σκεφτόμαστε τον αντίστροφο του δείκτη κατάστασης $1/\kappa(A)$ ως την απόσταση του πίνακα A από το σύνολο των μη αντιστρέψιμων πινάκων. Επομένως, αν ο $\kappa(A)$ είναι μεγάλος, τότε ο A «τείνει» να γίνει μη αντιστρέψιμος, και τότε είναι λογικό να αναμένουμε δυσκολίες κατά την επίλυση του γραμμικού συστήματος $Ax = b$.

Κλείνουμε την παράγραφο αυτή με ένα ακόμη παράδειγμα.

Παράδειγμα. Θεωρήστε το γραμμικό σύστημα $Ax = b$, με

$$A = \begin{pmatrix} 4.1 & 2.8 \\ 9.7 & 6.6 \end{pmatrix} \quad \text{και} \quad b = \begin{pmatrix} 4.1 \\ 9.7 \end{pmatrix}.$$

Το σύστημα αυτό έχει μοναδική λύση $x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. Θα χρησιμοποιήσουμε τη νόρμα $\|\cdot\|_1$, βλ.

(3.5). Έχουμε

$$\|b\|_1 = 13.8 \quad \text{και} \quad \|x\|_1 = 1.$$

Αν μεταβάλλουμε λίγο το δεύτερο μέλος του συστήματος σε

$$\tilde{b} = \begin{pmatrix} 4.11 \\ 9.70 \end{pmatrix},$$

τότε η ακριβής λύση του συστήματος είναι

$$\tilde{x} = \begin{pmatrix} 0.34 \\ 0.97 \end{pmatrix}.$$

Θέτουμε

$$\delta b = b - \tilde{b} = \begin{pmatrix} -0.01 \\ 0 \end{pmatrix} \quad \text{και} \quad \delta x = x - \tilde{x} = \begin{pmatrix} 0.66 \\ -0.97 \end{pmatrix}.$$

Τότε

$$\|\delta b\|_1 = 0.01 \quad \text{και} \quad \|\delta x\|_1 = 1.63.$$

Αν και μεταβάλαμε ελάχιστα το b βλέπουμε ότι το x αλλάζει τελείως. Συγκεκριμένα, οι σχετικές μεταβολές είναι

$$\frac{\|\delta b\|_1}{\|b\|_1} = 0.0007246 \quad \text{και} \quad \frac{\|\delta x\|_1}{\|x\|_1} = 1.63.$$

Επειδή το $\kappa_1(A)$ είναι ο μέγιστος συντελεστής μεγέθυνσης έχουμε ότι

$$\kappa_1(A) \geq \frac{1.63}{0.0007246} = 2249.4,$$

όπου ο δείκτης 1, στο $\kappa_1(A)$, δηλώνει ότι υπολογίζουμε το δείκτη κατάστασης ως προς τη νόρμα $\|\cdot\|_1$. (Στο συγκεκριμένο παράδειγμα επιλέξαμε τα b και δb που δίνουν το μέγιστο, επομένως εδώ έχουμε ότι $\kappa_1(A) = 2249.4$.) □

Συνοψίζοντας, όταν θέλουμε να λύσουμε γραμμικά συστήματα, ο δείκτης κατάστασης μας προσφέρει ένα μέτρο της ευαισθησίας του συστήματος σε μικρές διαταραχές των δεδομένων. Όταν ο δείκτης κατάστασης ενός πίνακα είναι πολύ μεγάλος, τότε μπορεί η υπολογιστική λύση να είναι λανθασμένη και, σίγουρα, απαιτείται περαιτέρω έλεγχος προτού αποδεχθούμε την υπολογιστική λύση ως μια καλή προσέγγιση της ακριβούς λύσης. Θα μπορούσαμε να πούμε ότι ένας γενικός στόχος στη μελέτη αριθμητικών μεθόδων είναι να κατακτήσουμε έναν βαθμό επίγνωσης του κατά πόσον ένα αριθμητικό αποτέλεσμα είναι αξιόπιστο ή «ύποπτο» (και ως εκ τούτου χρήζει περαιτέρω διερεύνησης). Στην περίπτωση της αριθμητικής επίλυσης γραμμικών συστημάτων σε αυτό μας βοηθάει ο δείκτης κατάστασης. Εκτιμήσεις του δείκτη κατάστασης για έναν πίνακα μπορεί κανείς να υπολογίσει με τη βοήθεια εξελιγμένου μαθηματικού λογισμικού (όπως, π.χ., το MATLAB) και στη συνέχεια να αξιολογήσει το κατά πόσον θα πρέπει να εμπιστευθεί ή όχι τα αποτελέσματά του.

3.5 Τριδιαγώνια συστήματα

Σε πολλές εφαρμογές απαιτείται να επιλύσουμε γραμμικά συστήματα όπου ο πίνακας του συστήματος είναι *αραιός*, δηλαδή τα περισσότερα στοιχεία του είναι μηδέν. Συχνά, τα μη μηδενικά στοιχεία του βρίσκονται πάνω στην κύρια διαγώνιο και σε κάποιες από τις υπερ- και υποδιαγώνιους του, δηλαδή ο πίνακας έχει *δομή ζώνης*. Τέτοιοι πίνακες εμφανίζονται κυρίως κατά την επίλυση συνήθων και μερικών διαφορικών εξισώσεων. Σε αυτές τις περιπτώσεις είναι καλό να αναπτύξουμε αλγόριθμους ειδικά σχεδιασμένους για τέτοια γραμμικά συστήματα επειδή έτσι μπορούμε να μειώσουμε σημαντικά το πλήθος των πράξεων και τις απαραίτητες θέσεις μνήμης.

Ειδικότερα, στην παράγραφο αυτή θα μελετήσουμε τριδιαγώνια συστήματα, δηλαδή συστήματα όπου τα μη μηδενικά στοιχεία του πίνακα του συστήματος βρίσκονται στη διαγώνιο και στην υπερ- και υποδιαγώνιο του. Έστω ότι θέλουμε να λύσουμε το σύστημα

$$Ax = f, \quad f \in \mathbb{R}^n, \quad (3.15)$$

όπου A είναι ο τριδιαγώνιος $n \times n$ πίνακας

$$A = \begin{pmatrix} a_1 & b_1 & & & 0 \\ c_2 & a_2 & b_2 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{n-1} & a_{n-1} & b_{n-1} \\ 0 & & & c_n & a_n \end{pmatrix}, \quad (3.16)$$

με $a_i, b_i, c_i \in \mathbb{R}$. Υποθέτουμε, επιπλέον, ότι

$$a_i, b_i, c_i \neq 0, \quad |a_1| > |b_1|, \quad |a_n| > |c_n|, \quad \text{και} \quad |a_i| \geq |b_i| + |c_i|, \quad 2 \leq i \leq n-1. \quad (3.17)$$

Τότε θα δείξουμε ότι ο A αντιστρέφεται και μπορεί να αναλυθεί στη μορφή

$$A = LU, \quad (3.18)$$

όπου L και U είναι οι ακόλουθοι διδιαγώνιοι πίνακες (L κάτω τριγωνικός και U άνω τριγωνικός)

$$L = \begin{pmatrix} d_1 & & & & 0 \\ c_2 & d_2 & & & \\ & \ddots & \ddots & & \\ & & c_{n-1} & d_{n-1} & \\ 0 & & & c_n & d_n \end{pmatrix}, \quad U = \begin{pmatrix} 1 & e_1 & & & 0 \\ & 1 & e_2 & & \\ & & \ddots & \ddots & \\ & & & 1 & e_{n-1} \\ 0 & & & & 1 \end{pmatrix}.$$

Επιπλέον, η ανάλυση αυτή είναι μοναδική.

Θα ξεκινήσουμε υπολογίζοντας τα άγνωστα στοιχεία $d_i, i = 1, \dots, n$ και $e_i, i = 1, \dots, n-1$, των πινάκων L και U , αντίστοιχα. Για να το πετύχουμε αυτό εξισώνουμε τα αντίστοιχα μη μηδενικά στοιχεία που βρίσκονται στη διαγώνιο και στην υπερδιαγώνιο του αριστερού και δεξιού μέλους της ισότητας (3.18). Έτσι έχουμε:

$$\begin{aligned} a_1 = d_1 &\Rightarrow d_1 = a_1 & b_1 = d_1 e_1 &\Rightarrow e_1 = b_1/d_1 \\ a_2 = c_2 e_1 + d_2 &\Rightarrow d_2 = a_2 - c_2 e_1 & b_2 = d_2 e_2 &\Rightarrow e_2 = b_2/d_2 \\ & & \vdots & \end{aligned}$$

οπότε μπορούμε να γράψουμε τον ακόλουθο αλγόριθμο:

Αλγόριθμος 3.3

Είσοδος: $a_i, i = 1, 2, \dots, n, \quad b_i, i = 1, 2, \dots, n-1, \quad c_i, i = 2, 3, \dots, n$

$d_1 \leftarrow a_1$

$e_1 \leftarrow b_1/d_1$

for $k = 2, \dots, n-1$ **do**

$d_k \leftarrow a_k - c_k \cdot e_{k-1}$

$e_k \leftarrow b_k/d_k$

end for

$d_n \leftarrow a_n - c_n \cdot e_{n-1}$

Παρατηρούμε ότι οι πράξεις που εκτελούνται στον αλγόριθμο (3.3) είναι $n-1$ διαιρέσεις, $n-1$ πολλαπλασιασμοί και $n-1$ προσθέσεις. Άρα, αν μετρήσουμε μαζί πολλαπλασιασμούς και διαιρέσεις έχουμε:

Πλήθος πράξεων Αλγ. (3.3) = $(2n-2)$ πολ/σμοι-διαιρέσεις και $(n-1)$ προσθέσεις.

Φυσικά, ο αλγόριθμος (3.3) θα είναι καλά ορισμένος αν $d_i \neq 0$ για κάθε $i = 1, 2, \dots, n$. Αν το αποδείξουμε αυτό, τότε ο L θα είναι αντιστρέψιμος και επομένως ο A θα είναι και αυτός αντιστρέψιμος ως γινόμενο των αντιστρέψιμων πινάκων L και U .

Πρόταση 3.2. Έστω A, L, U οι πίνακες που εμφανίζονται στην (3.18), έτσι ώστε τα στοιχεία του τριδιαγώνιου πίνακα A να ικανοποιούν τις υποθέσεις (3.17). Τότε τα $d_i, i = 1, 2, \dots, n$, που εμφανίζονται στον αλγόριθμο (3.3) είναι μη μηδενικά και, επομένως ο αλγόριθμος (3.3) είναι καλά ορισμένος. Επιπλέον, $|e_i| < 1$ για κάθε $i = 1, \dots, n-1$.

Απόδειξη. Η απόδειξη θα γίνει επαγωγικά. Αρχικά παρατηρούμε ότι:

$$d_1 = a_1 \text{ και } a_1 \neq 0. \quad \text{Άρα } d_1 \neq 0.$$

Επίσης, λόγω της υπόθεσης $|a_1| > |b_1|$ έχουμε ότι

$$|e_1| = \left| \frac{b_1}{d_1} \right| = \left| \frac{b_1}{a_1} \right| < 1.$$

Έστω τώρα ότι

$$d_{k-1} \neq 0 \text{ και } |e_{k-1}| < 1 \text{ για κάποιο } k, \quad k = 2, \dots, n.$$

Θα αποδείξουμε ότι $d_k \neq 0$ και $|e_k| < 1$. Πράγματι, έχουμε

$$|d_k| = |a_k - c_k e_{k-1}| \geq |a_k| - |c_k| \underbrace{|e_{k-1}|}_{<1} > |a_k| - |c_k| \geq |b_k| > 0,$$

όπου η προτελευταία ανισότητα προκύπτει από την υπόθεση $|a_k| \geq |b_k| + |c_k|$, και η τελευταία από την υπόθεση $b_k \neq 0$.

Άρα

$$|d_k| > |b_k| \Rightarrow \frac{|b_k|}{|d_k|} < 1 \Rightarrow |e_k| < 1.$$

□

(Μπορεί να αποδειχθεί ακόμη ότι κάτω από τις υποθέσεις (3.17) για τα στοιχεία του A ο αλγόριθμος (3.3) είναι ευσταθής όταν οι πράξεις γίνονται με αριθμητική κινητής υποδιαστολής και υπάρχουν σφάλματα στρογγύλευσης.)

Τώρα, για να λύσουμε το σύστημα (3.15) εργαζόμαστε ως εξής:

$$Ax = f \Leftrightarrow LUx = f \Leftrightarrow L(\underbrace{Ux}_{:=y}) = f.$$

Άρα, λύνουμε αρχικά το σύστημα

$$Ly = f,$$

όπου υπολογίζουμε το διάνυσμα y (από την αρχή προς το τέλος) με τον αλγόριθμο (3.4).

Αλγόριθμος 3.4

Είσοδος: $c_i, i = 2, 3, \dots, n, \quad d_i, i = 1, 2, \dots, n, \quad f_i, i = 1, 2, \dots, n$

$y_1 \leftarrow f_1/d_1$

for $k = 2, \dots, n$ **do**

$y_k \leftarrow (f_k - c_k \cdot y_{k-1})/d_k$

end for

Στη συνέχεια, λύνουμε το σύστημα

$$Ux = y$$

Αλγόριθμος 3.5**Είσοδος:** $y_i, i = 1, 2, \dots, n, \quad e_i, i = 1, 2, \dots, n - 1$ $x_n \leftarrow y_n$ **for** $k = n - 1, n - 2, \dots, 1$ **do** $x_k \leftarrow y_k - e_k \cdot x_{k+1}$ **end for**

όπου υπολογίζουμε το διάνυσμα x (από το τέλος προς την αρχή) με τον αλγόριθμο (3.5).

Αν μετρήσουμε όπως και πριν το πλήθος των πράξεων που απαιτούνται για τους αλγορίθμους (3.4) και (3.5) βρίσκουμε:

Πλήθος πράξεων Αλγ. (3.4) = $(2n - 1)$ πολ/σμοι-διαιρέσεις και $(n - 1)$ προσθέσεις.

και

Πλήθος πράξεων Αλγ. (3.5) = $(n - 1)$ πολ/σμοι και $(n - 1)$ προσθέσεις.

Άρα, συνολικά ο αλγόριθμος (3.3), (3.4), (3.5), απαιτεί $5n - 4$ πολλαπλασιασμούς-διαιρέσεις και $3n - 3$ προσθέσεις, δηλαδή πράξεις $O(n)$. Ανακαλέστε ότι η απαλοιφή Gauss για έναν πλήρη $n \times n$ πίνακα απαιτεί $O(n^3)$ πράξεις.

Όσον αφορά την απαιτούμενη μνήμη χρειαζόμαστε περίπου $4n$ θέσεις μνήμης για την αποθήκευση των διανυσμάτων a, b, c, f . Για τον υπολογισμό των d_i, e_i , στην πραγματικότητα δεν χρειαζόμαστε επιπλέον θέσεις μνήμης μιας και μπορούμε να εκχωρήσουμε τα d_i στις θέσεις των a_i , και τα e_i στις θέσεις των b_i . Στη συνέχεια θα γράψουμε το διάνυσμα y πάνω στο f , και το x πάνω στο y . Άρα συνολικά οι απαιτούμενες θέσεις μνήμης είναι μόνο $4n$.

3.6 Επίλυση γραμμικών συστημάτων στο MATLAB

Για να λύσουμε στο MATLAB το γραμμικό σύστημα $Ax = b$ χρησιμοποιούμε τον τελεστή της «αριστερής διαίρεσης» $\backslash$.

Παράδειγμα. Θεωρήστε το γραμμικό σύστημα $Ax = b$, με

$$A = \begin{pmatrix} 10 & -7 & 0 \\ -3 & 2 & 6 \\ 5 & -1 & 5 \end{pmatrix} \quad \text{και} \quad b = \begin{pmatrix} 7 \\ 4 \\ 6 \end{pmatrix}.$$

Γράφουμε στο παράθυρο εντολών του MATLAB τις εξής εντολές:

```
>> A = [10 -7 0; -3 2 6 ; 5 -1 5]
A =
    10    -7     0
    -3     2     6
     5    -1     5
>> b = [7 ; 4; 6]
b =
     7
     4
     6
>> x=A\b
x =
     0
    -1
     1
```

Κλείνουμε τη σύντομη αυτή παράγραφο με ένα παράδειγμα που αφορά την επίλυση ενός γραμμικού συστήματος της μορφής $Hx = b$, όπου ο πίνακας H είναι ένας πίνακας Hilbert διάστασης 12×12 , με στοιχεία $H(i, j) = \frac{1}{i+j-1}$, $1 \leq i, j \leq 12$. Οι πίνακες Hilbert είναι ένα κλασσικό παράδειγμα πινάκων με πολύ κακή κατάσταση.

Η ιδέα είναι η εξής: Αρχικά, θα κατασκευάσουμε με τη βοήθεια του MATLAB τον 12×12 πίνακα Hilbert H , θα ορίσουμε ως x το διάνυσμα στήλη που όλα του τα στοιχεία είναι ίσα με 1, δηλαδή $x = (1, 1, \dots, 1)^T$, και θα υπολογίσουμε με το MATLAB το διάνυσμα $b = Hx$. Με άλλα λόγια, το διάνυσμα $x = (1, 1, \dots, 1)^T$ είναι η ακριβής λύση του συστήματος $Hx = b$. Στη συνέχεια θα βρούμε με το MATLAB την προσεγγιστική λύση $\tilde{x}$ του γραμμικού συστήματος και θα την συγκρίνουμε με την ακριβή λύση x , υπολογίζοντας την Ευκλείδεια νόρμα του

σχετικού σφάλματος, $\|e\|_2 = \|x - \tilde{x}\|_2 / \|x\|_2$, καθώς και την Ευκλείδεια νόρμα του υπολοίπου $\|r\|_2 = \|b - A\tilde{x}\|_2$. Τέλος, θα εκτιμήσουμε τον δείκτη κατάστασης του πίνακα H (ως προς τη νόρμα $\|\cdot\|_2$).

```
>> H = hilb(12);
>> x = ones(12,1);
>> b = H*x;
>> xapp = H\b
xapp =
    1.0000
    1.0000
    0.9997
    1.0041
    0.9700
    1.1324
    0.6296
    1.6729
    0.2088
    1.5810
    0.7579
    1.0437

>> norm(x - xapp)/norm(x)
ans =
    0.3688

>> norm(b - H*xapp)
ans =
    7.2802e-16

>> cond(H)
ans =
    1.6315e+16
```

Ας δούμε αναλυτικά τις εντολές του MATLAB που χρησιμοποιήσαμε παραπάνω.

- Με την εντολή `H=hilb(12)` κατασκευάζουμε στο MATLAB τον 12×12 πίνακα Hilbert και τον καταχωρούμε στη μεταβλητή H .
- Κατασκευάζουμε το διάνυσμα στήλη x που όλα του τα στοιχεία είναι ίσα με 1 την

εντολή $\mathbf{x} = \text{ones}(12, 1)$.

- Εκχωρούμε στο διάνυσμα $\mathbf{b}$ το αποτέλεσμα του πολλαπλασιασμού $H\mathbf{x}$ γράφοντας $\mathbf{b} = H * \mathbf{x}$.
- Λύνουμε το γραμμικό σύστημα με την εντολή $\mathbf{xapp} = H \setminus \mathbf{b}$.
- Υπολογίζουμε την Ευκλείδεια νόρμα του σχετικού σφάλματος με την εντολή $\text{norm}(\mathbf{x} - \mathbf{xapp}) / \text{norm}(\mathbf{x})$, την Ευκλείδεια νόρμα του υπολοίπου με την εντολή $\text{norm}(\mathbf{b} - H * \mathbf{xapp})$ και, τέλος, το δείκτη κατάστασης του πίνακα H (ως προς την Ευκλείδεια νόρμα) με την εντολή $\text{cond}(H)$.

Παρατηρήστε ότι το σχετικό σφάλμα $\|\mathbf{x} - \tilde{\mathbf{x}}\|_2 / \|\mathbf{x}\|_2$ που υπολογίσαμε είναι περίπου 37%, ενώ το υπόλοιπο είναι πάρα πολύ μικρό ($\approx 7.3 \cdot 10^{-16}$). Τα αποτελέσματα αυτά είναι σε συμφωνία με ότι συζητήσαμε στις Παραγράφους 3.3, 3.4, και οφείλονται στην κακή κατάσταση του πίνακα H όπως, άλλωστε, υποδεικνύει ο μεγάλος δείκτης κατάστασης $\kappa_2(H) \approx 1.6 \cdot 10^{16}$.

3.7 Επαναληπτικές μέθοδοι

Στην παράγραφο αυτή θα παρουσιάσουμε δύο κλασσικές επαναληπτικές μεθόδους για την επίλυση του γραμμικού συστήματος $A\mathbf{x} = \mathbf{b}$, όπου ο A είναι ένας αντιστρέψιμος πίνακας. Οι επαναληπτικές μέθοδοι δεν αλλάζουν τον πίνακα A (όπως, π.χ., συμβαίνει κατά την απαλοιφή Gauss), αλλά ξεκινώντας με μια αρχική προσέγγιση $\mathbf{x}^{(0)}$ της λύσης $\mathbf{x}$ κατασκευάζουν μια ακολουθία διανυσμάτων $\mathbf{x}^{(m)}$, η οποία, υπό ορισμένες προϋποθέσεις, συγκλίνει στη λύση $\mathbf{x}$, καθώς $m \rightarrow \infty$.

Θα ξεκινήσουμε με την παρουσίαση της λεγόμενης μεθόδου του *Jacobi*. Θεωρούμε λοιπόν, και πάλι, το σύστημα:

$$A\mathbf{x} = \mathbf{b},$$

όπου

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{pmatrix}, \quad \text{με διαγώνια στοιχεία } a_{ii} \neq 0,$$

και

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}.$$

«Σπάμε» τον πίνακα A ως εξής:

$$A = L + D + U,$$

όπου

$$L = \begin{pmatrix} 0 & & & & \\ a_{21} & 0 & & & \\ a_{31} & a_{32} & 0 & & \\ \vdots & \vdots & & \ddots & \\ a_{n1} & a_{n2} & \dots & a_{n,n-1} & 0 \end{pmatrix}, \quad D = \begin{pmatrix} a_{11} & & & & \\ & a_{22} & & & \\ & & a_{33} & & \\ & & & \ddots & \\ & 0 & & & a_{nn} \end{pmatrix}$$

$$U = \begin{pmatrix} 0 & a_{12} & a_{13} & \dots & a_{1n} \\ & 0 & a_{23} & \dots & a_{2n} \\ & & 0 & & \vdots \\ & 0 & & \ddots & a_{n-1,n} \\ & & & & 0 \end{pmatrix}$$

Επομένως,

$$\begin{aligned} Ax = b &\Leftrightarrow (L + D + U)x = b \Leftrightarrow Dx = -(L + U)x + b \\ &\Leftrightarrow x = -D^{-1}(L + U)x + D^{-1}b. \end{aligned}$$

Τότε για κάθε $m \in \mathbb{N}$ θα υπολογίζουμε προσεγγίσεις με τη μέθοδο Jacobi μέσω του τύπου:

$$x^{(m+1)} = -D^{-1}(L + U)x^{(m)} + D^{-1}b,$$

όπου $x^{(0)}$ είναι ένα δεδομένο αρχικό διάνυσμα.

Αν γράψουμε την ίδια σχέση κατά συνιστώσα θα έχουμε:

$$x_i^{(m+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(m)} - \sum_{j=i+1}^n a_{ij}x_j^{(m)} \right), \quad i = 1, \dots, n. \quad (3.19)$$

Αν, τώρα, για τον υπολογισμό του $x_i^{(m+1)}$ ($i > 1$) αντί για τα $x_1^{(m)}, \dots, x_{i-1}^{(m)}$ χρησιμοποιήσουμε τα $x_1^{(m+1)}, \dots, x_{i-1}^{(m+1)}$, που έχουν ήδη υπολογισθεί, τότε προκύπτει η μέθοδος των Gauss-Seidel:

Για κάθε $m \in \mathbb{N}$ θα υπολογίζουμε προσεγγίσεις με τη μέθοδο Gauss–Seidel μέσω του τύπου:

$$x_i^{(m+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(m+1)} - \sum_{j=i+1}^n a_{ij} x_j^{(m)} \right), \quad i = 1, \dots, n. \quad (3.20)$$

όπου $x^{(0)}$ είναι ένα δεδομένο αρχικό διάνυσμα.

Διαισθητικά περιμένουμε η μέθοδος Gauss–Seidel να είναι καλύτερη από την Jacobi, αφού είναι λογικό να αναμένουμε ότι τα $x_k^{(m+1)}$, $k = 1, \dots, i-1$, είναι καλύτερες προσεγγίσεις των x_k απ' ό,τι τα $x_k^{(m)}$. Αυτό όμως γενικά δεν είναι σωστό, αφού μπορεί κανείς να βρεί παραδείγματα στα οποία η μία από αυτές τις μεθόδους συγκλίνει ενώ η άλλη όχι.

Ορισμός 3.5. Θα λέμε ότι ο πίνακας A έχει αυστηρά κυριαρχική διαγώνιο, αν

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

Αποδεικνύεται ότι αν ένας πίνακας A έχει αυστηρά κυριαρχική διαγώνιο, τότε τόσο η μέθοδος Jacobi όσο και η μέθοδος Gauss–Seidel συγκλίνουν.

Παράδειγμα. Να λυθεί με τη μέθοδο Jacobi το σύστημα

$$\begin{cases} 3x_1 + 6x_2 + 2x_3 = 0 \\ 3x_1 - x_2 + x_3 = 1 \\ 3x_1 + 3x_2 + 7x_3 = 4 \end{cases}.$$

Παρατηρήστε ότι ο πίνακας του συστήματος είναι:

$$\begin{pmatrix} 3 & 6 & 2 \\ 3 & -1 & 1 \\ 3 & 3 & 7 \end{pmatrix}$$

και δεν έχει αυστηρά κυριαρχική διαγώνιο.

Αν όμως εναλλάξουμε τις δύο πρώτες εξισώσεις έχουμε:

$$\begin{cases} 3x_1 - x_2 + x_3 = 1 \\ 3x_1 + 6x_2 + 2x_3 = 0 \\ 3x_1 + 3x_2 + 7x_3 = 4 \end{cases},$$

αλλά τώρα ο πίνακας του συστήματος είναι:

$$\begin{pmatrix} 3 & -1 & 1 \\ 3 & 6 & 2 \\ 3 & 3 & 7 \end{pmatrix}$$

και έχει αυστηρά κυριαρχική διαγώνιο, αφού

$$|3| > |-1| + |1|$$

$$|6| > |3| + |2|$$

$$|7| > |3| + |3|$$

Θα εφαρμόσουμε τη μέθοδο Jacobi στο σύστημα

$$\begin{cases} 3x_1 - x_2 + x_3 = 1 \\ 3x_1 + 6x_2 + 2x_3 = 0 \\ 3x_1 + 3x_2 + 7x_3 = 4 \end{cases} \Leftrightarrow \begin{cases} x_1 = \frac{1}{3}(1 + x_2 - x_3) \\ x_2 = \frac{1}{6}(-3x_1 - 2x_3) \\ x_3 = \frac{1}{7}(4 - 3x_1 - 3x_2) \end{cases}$$

Επομένως

$$\begin{aligned} x_1^{(m+1)} &= \frac{1}{3} \left(1 + x_2^{(m)} - x_3^{(m)} \right) \\ x_2^{(m+1)} &= \frac{1}{6} \left(-3x_1^{(m)} - 2x_3^{(m)} \right) \\ x_3^{(m+1)} &= \frac{1}{7} \left(4 - 3x_1^{(m)} - 3x_2^{(m)} \right) \end{aligned}$$

Θεωρούμε ως αρχική προσέγγιση το διάνυσμα $x^{(0)} = (0, 0, 0)^T$, και θα υπολογίσουμε τις επόμενες δύο προσεγγίσεις.

- 1η προσέγγιση:

$$\begin{aligned} x_1^{(1)} &= \frac{1}{3} \left(1 + x_2^{(0)} - x_3^{(0)} \right) = \frac{1}{3} (1 + 0 - 0) = \frac{1}{3} \\ x_2^{(1)} &= \frac{1}{6} \left(-3x_1^{(0)} - 2x_3^{(0)} \right) = \frac{1}{6} (-3 \cdot 0 - 2 \cdot 0) = 0 \\ x_3^{(1)} &= \frac{1}{7} \left(4 - 3x_1^{(0)} - 3x_2^{(0)} \right) = \frac{1}{7} (4 - 3 \cdot 0 - 3 \cdot 0) = \frac{4}{7} \end{aligned}$$

- 2η προσέγγιση:

$$\begin{aligned} x_1^{(2)} &= \frac{1}{3} \left(1 + x_2^{(1)} - x_3^{(1)} \right) = \frac{1}{3} \left(1 + 0 - \frac{4}{7} \right) = \frac{1}{7} \approx 0.1429 \\ x_2^{(2)} &= \frac{1}{6} \left(-3x_1^{(1)} - 2x_3^{(1)} \right) = \frac{1}{6} \left(-3 \cdot \frac{1}{3} - 2 \cdot \frac{4}{7} \right) = -\frac{15}{42} \approx -0.3571 \\ x_3^{(2)} &= \frac{1}{7} \left(4 - 3x_1^{(1)} - 3x_2^{(1)} \right) = \frac{1}{7} \left(4 - 3 \cdot \frac{1}{3} - 3 \cdot 0 \right) = \frac{3}{7} \approx 0.4286 \end{aligned}$$

Άρα η δευτερη προσέγγιση της λύσης με τη μέθοδο Jacobi είναι το διάνυσμα:

$$\left(\frac{1}{7}, -\frac{15}{42}, \frac{3}{7}\right)^T \approx (0.1429, -0.3571, 0.4286)^T.$$

□

Παράδειγμα. Να λυθεί το προηγούμενο παράδειγμα με τη μέθοδο Gauss–Seidel.

Τώρα θα εφαρμόσουμε τη μέθοδο Gauss–Seidel στο σύστημα:

$$\begin{cases} 3x_1 - x_2 + x_3 = 1 \\ 3x_1 + 6x_2 + 2x_3 = 0 \\ 3x_1 + 3x_2 + 7x_3 = 4 \end{cases} \Leftrightarrow \begin{cases} x_1 = \frac{1}{3}(1 + x_2 - x_3) \\ x_2 = \frac{1}{6}(-3x_1 - 2x_3) \\ x_3 = \frac{1}{7}(4 - 3x_1 - 3x_2) \end{cases}$$

Επομένως

$$\begin{aligned} x_1^{(m+1)} &= \frac{1}{3} \left(1 + x_2^{(m)} - x_3^{(m)} \right) \\ x_2^{(m+1)} &= \frac{1}{6} \left(-3x_1^{(m+1)} - 2x_3^{(m)} \right) \\ x_3^{(m+1)} &= \frac{1}{7} \left(4 - 3x_1^{(m+1)} - 3x_2^{(m+1)} \right) \end{aligned}$$

Θεωρούμε ως αρχική προσέγγιση το διάνυσμα $x^{(0)} = (0, 0, 0)^T$, και θα υπολογίσουμε τις επόμενες δύο προσεγγίσεις.

- 1η προσέγγιση:

$$\begin{aligned} x_1^{(1)} &= \frac{1}{3} \left(1 + x_2^{(0)} - x_3^{(0)} \right) = \frac{1}{3} (1 + 0 - 0) = \frac{1}{3} \\ x_2^{(1)} &= \frac{1}{6} \left(-3x_1^{(1)} - 2x_3^{(0)} \right) = \frac{1}{6} \left(-3 \cdot \frac{1}{3} - 2 \cdot 0 \right) = -\frac{1}{6} \\ x_3^{(1)} &= \frac{1}{7} \left(4 - 3x_1^{(1)} - 3x_2^{(1)} \right) = \frac{1}{7} \left(4 - 3 \cdot \frac{1}{3} - 3 \cdot \left(-\frac{1}{6}\right) \right) = \frac{1}{2} \end{aligned}$$

- 2η προσέγγιση:

$$\begin{aligned} x_1^{(2)} &= \frac{1}{3} \left(1 + x_2^{(1)} - x_3^{(1)} \right) = \frac{1}{3} \left(1 + \left(-\frac{1}{6}\right) - \frac{1}{2} \right) = \frac{1}{9} \approx 0.1111 \\ x_2^{(2)} &= \frac{1}{6} \left(-3x_1^{(2)} - 2x_3^{(1)} \right) = \frac{1}{6} \left(-3 \cdot \frac{1}{9} - 2 \cdot \frac{1}{2} \right) = -\frac{2}{9} \approx -0.2222 \\ x_3^{(2)} &= \frac{1}{7} \left(4 - 3x_1^{(2)} - 3x_2^{(2)} \right) = \frac{1}{7} \left(4 - 3 \cdot \frac{1}{9} - 3 \cdot \left(-\frac{2}{9}\right) \right) = \frac{13}{21} \approx 0.6190 \end{aligned}$$

Άρα η δευτερη προσέγγιση της λύσης με τη μέθοδο Gauss–Seidel είναι το διάνυσμα:

$$\left(\frac{1}{9}, -\frac{2}{9}, \frac{13}{21}\right)^T \approx (0.1111, -0.2222, 0.6190)^T.$$

Σημειώστε ότι η ακριβής λύση αυτού του συστήματος είναι: $(0.0351, -0.2368, 0.6579)^T$.

□

3.8 Συμπληρωματικές παρατηρήσεις

1. Συμμετρικοί, θετικά ορισμένοι πίνακες. Πολύ συχνά στις εφαρμογές εμφανίζονται συστήματα στα οποία ο πίνακας του συστήματος A είναι ένας $n \times n$ πραγματικός συμμετρικός (δηλ. $A = A^T$) και θετικά ορισμένος πίνακας. Σημειώστε ότι *θετικά ορισμένος* λέγεται ένας πίνακας A για τον οποίο ισχύει

$$x^T A x > 0 \text{ για κάθε } x \in \mathbb{R}^n \text{ με } x \neq 0.$$

(Είναι πολύ εύκολο να αποδειχθεί, π.χ., ότι ένας θετικά ορισμένος πίνακας είναι αντιστρέψιμος και έχει θετικά διαγώνια στοιχεία.)

Μια σημαντική ιδιότητα των συστημάτων με συμμετρικούς, θετικά ορισμένους πίνακες είναι ότι η επίλυσή τους μπορεί να γίνει με απλή απαλοιφή Gauss χωρίς μερική οδήγηση. Επιπλέον, υπάρχει μια παραλλαγή της ανάλυσης LU που λέγεται *ανάλυση Cholesky* κατά την οποία βρίσκουμε έναν μοναδικό κάτω τριγωνικό πίνακα L έτσι ώστε να γράψουμε τον πίνακα A στη μορφή $A = LL^T$. Η ανάλυση αυτή απαιτεί $n^3/6 + O(n^2)$ πράξεις και μπορεί να αποδειχθεί ότι η μέθοδος αυτή είναι ευσταθής. Για περισσότερες λεπτομέρειες σχετικά με τη μέθοδο Cholesky παραπέμπουμε στο βιβλίο των Ακρίβη και Δουγαλή, [1, Ενότητα 3.5].

2. Αραιοί πίνακες. Σε πολλά προβλήματα εφαρμογών και κυρίως όταν λύνουμε αριθμητικά μερικές διαφορικές εξισώσεις εμφανίζεται η ανάγκη επίλυσης συστημάτων με μεγάλους αραιούς πίνακες. Φυσικά, όταν αποθηκεύουμε και κάνουμε πράξεις με αραιούς πίνακες σ' έναν υπολογιστή είναι χρήσιμο και πολύ συχνά απαραίτητο να χρησιμοποιήσουμε εξειδικευμένους αλγόριθμους και δομές δεδομένων που εκμεταλλεύονται την αραιή δομή του πίνακα. Για παράδειγμα, το μειονέκτημα της απαλοιφής Gauss σε συστήματα με μεγάλους αραιούς πίνακες είναι ότι κατά τη διαδικασία της απαλοιφής τα μη μηδενικά στοιχεία του πίνακα μπορεί να «καταστραφούν» και να οδηγηθούμε σε ένα «γέμισμα» (fill-in) του πίνακα. Προφανώς έτσι αυξάνεται σημαντικά η απαιτούμενη μνήμη καθώς και το πλήθος των πράξεων. Παρόλα αυτά είναι δυνατόν για γενικούς αραιούς πίνακες να ελαττώσουμε το «γέμισμα» με κατάλληλες εναλλαγές γραμμών και στηλών.

3. Προγράμματα. Η πλέον καθιερωμένη σήμερα βιβλιοθήκη λογισμικού Αριθμητικής Γραμμικής Άλγεβρας είναι η LAPACK (Linear Algebra Package) (<http://www.netlib.org/lapack/lug/>).

Η LAPACK σχεδιάστηκε ως ο διάδοχος των LINPACK και EISPACK, δύο μεγάλων πακέτων προγραμμάτων Αριθμητικής Γραμμικής Άλγεβρας που γράφτηκαν σε FORTRAN 77 στις δεκαετίες του 1970 και 1980. Η LAPACK περιλαμβάνει προγράμματα για την επίλυση συστημάτων γραμμικών εξισώσεων καθώς και για τον υπολογισμό ιδιοτιμών και ιδιοδιανυσμάτων

πινάκων. Τα προγράμματα αυτά αφορούν τόσο πραγματικούς όσο και μιγαδικούς πίνακες και είναι γραμμένα σε απλή και διπλή ακρίβεια. Επιπλέον, η LAPACK έχει μια εξαιρετική οργάνωση πολύ προσεκτικά γραμμένων BLAS (Basic Linear Algebra Subroutines), δηλαδή προγραμμάτων βασικών πράξεων της γραμμικής άλγεβρας (πράξεις μεταξύ διανυσμάτων, πράξεις μεταξύ διανυσμάτων και πινάκων). Πολλά από τα προγράμματα αυτά ενσωματώθηκαν σε περιβάλλοντα όπως το MATLAB κ.α. Επεκτάσεις της LAPACK για συστήματα κατανεμημένης μνήμης έχουν πραγματοποιηθεί σε μεταγενέστερα πακέτα όπως τα ScaLAPACK και PLAPACK.

Ασκήσεις

1. Θεωρήστε το γραμμικό σύστημα

$$3x + 2y - z = -1$$

$$6x + 2y + 3z = -3$$

$$x + y - z = 0.$$

Λύστε το με τη μέθοδο της απαλοιφής Gauss με μερική οδήγηση.

2. Λύστε το ακόλουθο σύστημα εξισώσεων θεωρώντας ότι κάνετε τις πράξεις σ' έναν Η/Υ με $\beta = 10$, $t = 4$, $-L = U = 10$, που κάνει αποκοπή. Συγκρίνετε το αποτέλεσμα που θα βρείτε με το αντίστοιχο αν κάνετε τις πράξεις στον υπολογιστή αυτό αλλά με αυξημένη ακρίβεια $t = 8$.

$$0.1036x_1 + 0.2122x_2 = 0.7381$$

$$0.2081x_1 + 0.4247x_2 = 0.9327$$

3. Έστω $A \in \mathbb{R}^{n \times n}$ αντιστρέψιμος, $B \in \mathbb{R}^{n \times n}$, και $b \in \mathbb{R}^n$. Πώς θα υπολογίζατε οικονομικότερα, από άποψη πράξεων και μνήμης, τα διανύσματα $A^{-3}b$ και $A^{-1}BA^{-1}b$;
4. Θεωρήστε τον διδιαγώνιο πίνακα $A \in \mathbb{R}^{n \times n}$,

$$A = \begin{pmatrix} a_1 & b_1 & & & \\ & a_2 & b_2 & & \mathbf{0} \\ & & \ddots & \ddots & \\ \mathbf{0} & & & a_{n-1} & b_{n-1} \\ & & & & a_n \end{pmatrix}, \quad \text{όπου } a_i \neq 0, \quad i = 1, 2, \dots, n.$$

- (α) Διατυπώστε έναν οικονομικό αλγόριθμο για τον υπολογισμό του A^{-1} και εκτιμήστε πόσες πράξεις και θέσεις μνήμης απαιτεί.

(β) Πώς θα υπολογίζατε οικονομικότερα, από άποψη πράξεων και μνήμης, το διάνυσμα $x = A^{-2}y$ για δεδομένο $y \in \mathbb{R}^n$;

5. Αποδείξτε τις ιδιότητες (3.6) και (3.7) της Παραγράφου 3.3.

6. Βρείτε και σχεδιάστε τα παρακάτω σύνολα («μοναδιαίες σφαίρες») στον $\mathbb{R}^2$:

$$\alpha) \{x \in \mathbb{R}^2 : \|x\|_2 = 1\}, \quad \beta) \{x \in \mathbb{R}^2 : \|x\|_1 = 1\}, \quad \gamma) \{x \in \mathbb{R}^2 : \|x\|_\infty = 1\}.$$

7. Αν $x \in \mathbb{R}^n$ να δείξετε ότι:

$$\begin{aligned} \|x\|_2 &\leq \|x\|_1 \leq \sqrt{n}\|x\|_2, \\ \|x\|_\infty &\leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty, \\ \|x\|_\infty &\leq \|x\|_1 \leq n\|x\|_\infty. \end{aligned}$$

Επαληθεύστε ότι οι σταθερές στις ανισότητες αυτές είναι βέλτιστες. (Για τον σκοπό αυτό, αρκεί με απλά παραδείγματα να δείξετε ότι οι ανισότητες ισχύουν και ως ισότητες.)

(Υπόδειξη: Για την απόδειξη κάποιας από τις παραπάνω ανισότητες μπορεί να χρειαστείτε την ανισότητα Cauchy-Schwarz: $|\sum_{i=1}^n x_i y_i| \leq (\sum_{i=1}^n x_i^2)^{1/2} (\sum_{i=1}^n y_i^2)^{1/2}$.)

8. Ορίζουν οι παρακάτω ποσότητες φυσικές νόρμες πινάκων $A \in \mathbb{R}^{n \times n}$;

$$\alpha) \|A\|_{\max} = \max_{1 \leq i, j \leq n} |a_{ij}|, \quad \beta) \|A\|_E = \left(\sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^2 \right)^{1/2} \quad (\text{Frobenius-νόρμα του } A).$$

9. Ας συμβολίσουμε $\|\cdot\|$ μια νόρμα στον $\mathbb{R}^n$, καθώς και την αντίστοιχη φυσική νόρμα πίνακα που παράγει. Αν P είναι ένας αντιστρέψιμος $n \times n$ πίνακας, να δείξετε ότι η έκφραση

$$\|x\|_P = \|Px\|,$$

ορίζει μια νόρμα στον $\mathbb{R}^n$ και ότι η παραγόμενη από αυτή φυσική νόρμα πίνακα ορίζεται από τη σχέση $\|A\|_P = \|PAP^{-1}\|$.

10. Έστω $\|\cdot\|$ μια νόρμα στον $\mathbb{R}^n$ και η αντίστοιχη φυσική νόρμα στον $\mathbb{R}^{n \times n}$ που παράγεται από αυτήν.

α) Αν $A \in \mathbb{R}^{n \times n}$ με $\|A\| < 1$, να αποδείξετε ότι ο πίνακας $I - A$ αντιστρέφεται, και ότι ισχύει

$$\frac{1}{1 + \|A\|} \leq \|(I - A)^{-1}\| \leq \frac{1}{1 - \|A\|}.$$

β) Χρησιμοποιήστε το (α) για να αποδείξετε ότι ο τριδιαγώνιος πίνακας

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

αντιστρέφεται και ότι ο δείκτης κατάστασής του ως προς τη νόρμα μεγίστου $\kappa_\infty(B) = \|B\|_\infty \|B^{-1}\|_\infty \leq 3$.

11. Έστω $A, B \in \mathbb{R}^{n \times n}$, A αντιστρέψιμος, και $\|\cdot\|$ είναι μια φυσική νόρμα πινάκων.

α) Αν $\|A - B\| < \frac{1}{\|A^{-1}\|}$, να δείξετε ότι ο B αντιστρέφεται.

β) Αν ο $B \in \mathbb{R}^{n \times n}$ είναι μη αντιστρέψιμος, τότε $\frac{1}{\kappa(A)} \leq \frac{\|A-B\|}{\|A\|}$.

12. Έστω $0 \neq x \in \mathbb{R}^n$ η ακριβής λύση του συστήματος $Ax = b$, όπου $A \in \mathbb{R}^{n \times n}$ αντιστρέψιμος και $b \in \mathbb{R}^n$. Αν $\tilde{x}$ είναι μια προσέγγιση της x (π.χ. με απαλοιφή Gauss με αριθμητική κινητής υποδιαστολής πεπερασμένης ακρίβειας) και $r = A\tilde{x} - b$ είναι το υπόλοιπο της $\tilde{x}$, τότε αποδείξτε ότι για κάθε νόρμα $\|\cdot\|$ στον $\mathbb{R}^n$ έχουμε:

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq \kappa(A) \frac{\|r\|}{\|b\|}, \quad \text{όπου } \kappa(A) = \|A\| \|A^{-1}\|.$$

Πώς ερμηνεύετε την ανισότητα αυτή;

13. Έστω $A \in \mathbb{R}^{n \times n}$ αντιστρέψιμος πίνακας, $b \in \mathbb{R}^n$ και $x \in \mathbb{R}^n$ η ακριβής λύση του συστήματος $Ax = b$. Έστω $\delta A \in \mathbb{R}^{n \times n}$ μια διαταραχή του A και $x + \delta x$ μια μη μηδενική λύση του συστήματος $(A + \delta A)(x + \delta x) = b$. Αν $\|\cdot\|$ είναι μια νόρμα στον $\mathbb{R}^n$ και η αντίστοιχη φυσική νόρμα στον $\mathbb{R}^{n \times n}$ που παράγεται από αυτήν, να δείξετε ότι

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq \kappa(A) \frac{\|\delta A\|}{\|A\|},$$

όπου $\kappa(A) = \|A\| \|A^{-1}\|$.

14. Έστω $0 < \epsilon \ll 1$. Υπολογίστε τον δείκτη κατάστασης ως προς τη νόρμα $\|\cdot\|_\infty$ των πινάκων

$$A_1 = \begin{pmatrix} \epsilon & -1 \\ 1 & 1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} \epsilon & -\epsilon \\ 0 & \epsilon \end{pmatrix}, \quad A_3 = \begin{pmatrix} \epsilon & 1 \\ 0 & \epsilon \end{pmatrix}.$$

Πώς συμπεριφέρεται ο δείκτης κατάστασης των πινάκων αυτών καθώς $\epsilon \rightarrow 0$; Τι μπορείτε να συμπεράνετε για τα αντίστοιχα 2×2 γραμμικά συστήματα, όταν το ϵ είναι μικρό;

15. Θεωρήστε ένα 4×4 σύστημα $Ax = b$ με $b = (0, 1, 0, 1)^T$. Η ακριβής λύση είναι $x = (1, 1, 1, 1)^T$, ενώ η απαλοιφή Gauss με αριθμητική πεπερασμένης ακρίβειας δίνει την προσεγγιστική λύση $\tilde{x} = (1.02, 0.99, 1.01, 0.98)^T$, της οποίας το υπόλοιπο $r = A\tilde{x} - b$ είναι ίσο με το διάνυσμα $10^{-6}(0.22, -0.38, 0.62, 0.78)^T$. Πόσος τουλάχιστον είναι ο δείκτης κατάστασης $\kappa_1(A)$ του πίνακα A ως προς τη νόρμα $\|\cdot\|_1$; (Υπόδειξη: Χρησιμοποιήστε την ανισότητα της Άσκησης 12.)
16. Έστω $A = \begin{pmatrix} 3 & -1 & 1 \\ 3 & 6 & 2 \\ 3 & 3 & 7 \end{pmatrix}$ και $b = \begin{pmatrix} 1 \\ 0 \\ 4 \end{pmatrix}$. Ποια είναι η προσέγγιση της λύσης του γραμμικού συστήματος $Ax = b$ που θα βρείτε κάνοντας δύο επαναλήψεις με τη μέθοδο Jacobi με αρχική προσέγγιση $x^{(0)} = (0, 0, 0)^T$;
17. Έστω $A = \begin{pmatrix} 3 & -1 & 1 \\ 3 & 6 & 2 \\ 3 & 3 & 7 \end{pmatrix}$ και $b = \begin{pmatrix} 1 \\ 0 \\ 4 \end{pmatrix}$. Ποια είναι η προσέγγιση της λύσης του γραμμικού συστήματος $Ax = b$ που θα βρείτε κάνοντας δύο επαναλήψεις με τη μέθοδο Gauss-Seidel με αρχική προσέγγιση $x^{(0)} = (0, 0, 0)^T$;

Κεφάλαιο 4

Παρεμβολή

Ας υποθέσουμε ότι μας δίνεται ένας πίνακας με τις τιμές μιας συνάρτησης:

x	x_0	x_1	$\dots$	x_n
y	y_0	y_1	$\dots$	y_n

όπου τα σημεία x_i , $i = 0, 1, \dots, n$, είναι ανα δύο διαφορετικά μεταξύ τους. Άρα μπορούμε να απεικονίσουμε τα $n + 1$ ζεύγη ως σημεία στο καρτεσιανό επίπεδο. Ένα ερώτημα που τίθεται είναι κατά πόσον μπορούμε να βρούμε μια άλλη απλή συνάρτηση η οποία να διέρχεται ακριβώς από τα σημεία αυτά;

Ένα δεύτερο, παρόμοιο, πρόβλημα είναι να μας δίνεται μια συνάρτηση f της οποίας όμως ο υπολογισμός (π.χ. στον υπολογιστή) είναι πολύπλοκος και δαπανηρός. Μπορούμε άραγε να βρούμε μια άλλη συνάρτηση g που ο υπολογισμός της να είναι απλούστερος και να μας δίνει μια ικανοποιητική προσέγγιση της f ;

Ως ένα τρίτο πρόβλημα, ας θεωρήσουμε ότι οι αριθμητικές τιμές του παραπάνω πίνακα είναι αποτέλεσμα μετρήσεων κάποιου πειράματος και ως εκ τούτου εμπεριέχουν σφάλματα. Είναι δυνατόν να βρούμε μια συνάρτηση που να αναπαριστά τα δεδομένα αυτά (κατά προσέγγιση) και αν είναι δυνατόν να «φιλτράρει» τα σφάλματα;

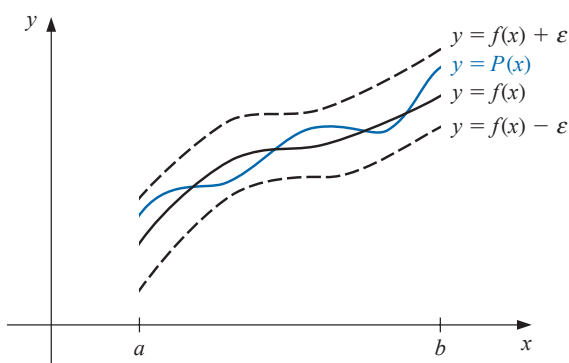
Σε όλα τα παραπάνω προβλήματα ουσιαστικά αναζητούμε μια απλή συνάρτηση p η οποία να αναπαριστά ή να προσεγγίζει τις τιμές του δοσμένου πίνακα ή τις τιμές μιας συνάρτησης f . Από τη στιγμή που υπολογίζουμε μια τέτοια συνάρτηση p μπορούμε να την χρησιμοποιήσουμε αντί για την f σε διάφορες περιπτώσεις. Π.χ., μπορεί να είναι ευκολότερο να υπολογίσουμε το ολοκλήρωμα της p αντί της f .

Στο κεφάλαιο αυτό θα ασχοληθούμε με την προσέγγιση συναρτήσεων με παρεμβολή με πολυωνμικές και κατά τμήματα πολυωνμικές συναρτήσεις. Τα πολυώνυμα χρησιμοποιούνται ευρέως για την προσέγγιση συναρτήσεων επειδή υπολογίζονται πολύ εύκολα με ένα πεπερασμένο πλήθος πράξεων (προσθαιρέσεων και πολλαπλασιασμών). Επιπλέον, μπορούμε πολύ

εύκολα να υπολογίσουμε την παράγωγό τους και το ολοκλήρωμά τους. Τέλος, το θεώρημα προσέγγισης του Weierstrass μας εξασφαλίζει ότι έχουν καλές προσεγγιστικές ιδιότητες (βλ. και Σχήμα 4.1).

Θεώρημα 4.1 (Weierstrass). *Αν η f είναι συνεχής συνάρτηση στο κλειστό και φραγμένο διάστημα $[a, b]$, τότε για κάθε $\varepsilon > 0$ υπάρχει πολυώνυμο P , τέτοιο ώστε*

$$\max_{a \leq x \leq b} |f(x) - P(x)| < \varepsilon.$$



Σχήμα 4.1: Θεώρημα Weierstrass.

Από την άλλη μεριά μπορεί η παρεμβολή με πολυώνυμα να αποτύχει για διάφορους λόγους, όπως θα δούμε στη συνέχεια. Σε τέτοιες περιπτώσεις χρησιμοποιούμε τμηματικά πολυωνυμικές συναρτήσεις, τις λεγόμενες *splines*.

4.1 Πολυωνυμική παρεμβολή

Θα ξεκινήσουμε θεωρώντας και πάλι των πίνακα τιμών:

x	x_0	x_1	$\dots$	x_n
y	y_0	y_1	$\dots$	y_n

όπου $x_0, x_1, \dots, x_n \in \mathbb{R}$, είναι $n+1$ ανά δύο διαφορετικά μεταξύ τους σημεία και $y_0, y_1, \dots, y_n \in \mathbb{R}$. Επομένως, αναζητούμε ένα πολυώνυμο p που να ορίζεται για όλα τα $x \in \mathbb{R}$, και στα x_i να παίρνει την τιμή y_i , δηλαδή $p(x_i) = y_i$, για όλα τα $i = 0, 1, \dots, n$. Αν τα y_i είναι τιμές κάποιας συνάρτησης f , δηλαδή $y_i = f(x_i)$, τότε λέμε ότι το πολυώνυμο p *παρεμβάλλεται στην f στα σημεία $x_0, x_1, \dots, x_n$* .

Ας ξεκινήσουμε υπολογίζοντας ένα πολυώνυμο που παρεμβάλλεται στις τιμές μιας συνάρτησης f σε δύο σημεία a, b ($a < b$), δηλαδή διέρχεται από τα σημεία $(a, f(a))$, $(b, f(b))$. Μιας και

από δύο σημεία διέρχεται μία και μόνο ευθεία, ουσιαστικά, ψάχνουμε ένα πολυώνυμο πρώτου βαθμού. Εύκολα διαπιστώνουμε ότι το πολυώνυμο

$$p_1(x) = \frac{f(b) - f(a)}{b - a}(x - a) + f(a),$$

που είναι βαθμού (το πολύ) ένα, διέρχεται από τα ζητούμενα σημεία. Πράγματι,

$$p_1(a) = \frac{f(b) - f(a)}{b - a}(a - a) + f(a) = 0 + f(a) = f(a).$$

$$p_1(b) = \frac{f(b) - f(a)}{b - a}(b - a) + f(a) = f(b) - f(a) + f(a) = f(b).$$

Τώρα, γενικότερα, μπορεί κανείς να αποδείξει (βλ., π.χ., [1, Θεώρ. 4.1]) ότι:

Θεώρημα 4.2. Έστω $x_0, x_1, \dots, x_n \in \mathbb{R}$, $n + 1$ ανά δύο διαφορετικά μεταξύ τους σημεία και $y_0, y_1, \dots, y_n \in \mathbb{R}$. Τότε υπάρχει ακριβώς ένα πολυώνυμο p βαθμού (το πολύ) n , τέτοιο ώστε

$$p(x_i) = y_i, \quad i = 0, 1, \dots, n.$$

4.1.1 Παράσταση του πολυωνύμου παρεμβολής σε μορφή Lagrange

Έστω $x_0, x_1, \dots, x_n \in \mathbb{R}$ ανά δύο διαφορετικά μεταξύ τους σημεία. Ονομάζουμε πολυώνυμο Lagrange τα πολυώνυμα $L_i(x)$, $i = 0, \dots, n$, βαθμού n , με την ιδιότητα

$$L_i(x_j) = \begin{cases} 1, & \text{αν } i = j \\ 0, & \text{αν } i \neq j. \end{cases} \quad (4.1)$$

Εύκολα μπορούμε να δούμε ότι τα πολυώνυμα

$$L_i(x) = \frac{(x - x_0)(x - x_1) \dots (x - x_{i-1})(x - x_{i+1}) \dots (x - x_n)}{(x_i - x_0)(x_i - x_1) \dots (x_i - x_{i-1})(x_i - x_{i+1}) \dots (x_i - x_n)}, \quad (4.2)$$

πληρούν τις συνθήκες (4.1).

Τότε το πολυώνυμο παρεμβολής p_n , βαθμού το πολύ n , που παρεμβάλλεται στις τιμές της f στα σημεία $x_0, x_1, \dots, x_n$ γράφεται στη μορφή

$$p_n(x) = \sum_{i=0}^n f(x_i) L_i(x) = f(x_0) L_0(x) + f(x_1) L_1(x) + \dots + f(x_n) L_n(x). \quad (4.3)$$

Παράδειγμα. Να προσδιορίσετε το πολυώνυμο παρεμβολής p_2 (βαθμού το πολύ 2) που παρεμβάλλεται στις τιμές της $f(x) = \frac{1}{x}$ στα σημεία $x_0 = 2, x_1 = 2.5, x_2 = 4$.

Λύση. Οι τιμές που μας δίνονται συνοψίζονται στον πίνακα:

i	0	1	2
x_i	2	2.5	4
$f(x_i)$	1/2	2/5	1/4

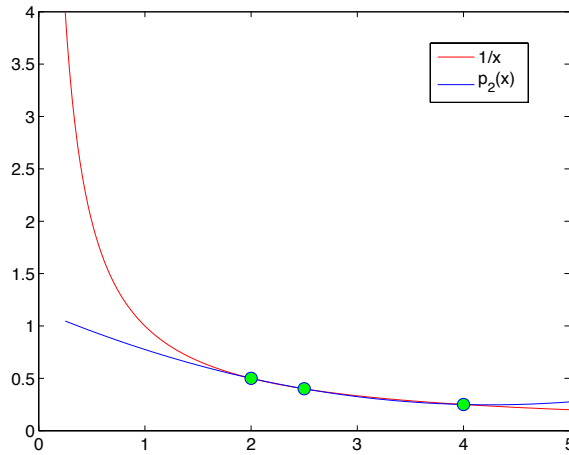
Αρχικά, υπολογίζουμε τα πολυώνυμα Lagrange

$$\begin{aligned}
 L_0(x) &= \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} = \frac{(x - 2.5)(x - 4)}{(2 - 2.5)(2 - 4)} = x^2 - 6.5x + 10, \\
 L_1(x) &= \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} = \frac{(x - 2)(x - 4)}{(2.5 - 2)(2.5 - 4)} = -\frac{4}{3}x^2 + 8x - \frac{32}{3}, \\
 L_2(x) &= \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} = \frac{(x - 2)(x - 2.5)}{(4 - 2)(4 - 2.5)} = \frac{1}{3}x^2 - \frac{3}{2}x + \frac{5}{3}.
 \end{aligned}$$

Άρα

$$\begin{aligned}
 p_2(x) &= \sum_{i=0}^2 f(x_i)L_i(x) = f(x_0)L_0(x) + f(x_1)L_1(x) + f(x_2)L_2(x) \\
 &= \frac{1}{2}(x^2 - 6.5x + 10) + \frac{2}{5}\left(-\frac{4}{3}x^2 + 8x - \frac{32}{3}\right) + \frac{1}{4}\left(\frac{1}{3}x^2 - \frac{3}{2}x + \frac{5}{3}\right) \\
 &= 0.05x^2 - 0.425x + 1.15.
 \end{aligned}$$

Στο Σχήμα 4.2 έχουμε σχεδιάσει τις γραφικές παραστάσεις της $f(x) = 1/x$ και του πολυωνύμου παρεμβολής $p_2(x) = 0.05x^2 - 0.425x + 1.15$ □



Σχήμα 4.2: Το γράφημα της $f(x) = 1/x$ και του πολυωνύμου παρεμβολής $p_2(x) = 0.05x^2 - 0.425x + 1.15$.

4.1.2 Παράσταση σε μορφή του Νεύτωνα

Θεωρούμε και πάλι $x_0, x_1, \dots, x_n \in \mathbb{R}$ ανά δύο διαφορετικά μεταξύ τους σημεία και $y_0, y_1, \dots, y_n \in \mathbb{R}$. Τότε αν γράψουμε το πολυώνυμο p βαθμού το πολύ n για το οποίο ισχύει $p(x_i) = y_i$, $i = 0, \dots, n$, στη μορφή

$$p(x) = \alpha_0 + \alpha_1(x - x_0) + \alpha_2(x - x_0)(x - x_1) + \dots + \alpha_n(x - x_0) \dots (x - x_{n-1}), \quad (4.4)$$

μπορούμε εύκολα να υπολογίσουμε τους συντελεστές $\alpha_0, \alpha_1, \dots, \alpha_n$ διαδοχικά:

$$\begin{aligned} p(x_0) = y_0 &\rightarrow \alpha_0 = y_0 \\ p(x_1) = y_1 &\rightarrow \alpha_1 = \frac{y_1 - \alpha_0}{x_1 - x_0} \\ p(x_2) = y_2 &\rightarrow \alpha_2 = \left(\frac{y_2 - \alpha_0}{x_2 - x_0} - \alpha_1 \right) / (x_2 - x_1) \\ &\vdots \end{aligned}$$

Η μορφή του Νεύτωνα έχει δύο βασικά πλεονεκτήματα:

- α) Έστω ότι έχουμε υπολογίσει τους συντελεστές α_i . Αν παρατηρήσουμε ότι η (4.4) μπορεί να γραφεί στη μορφή:

$$p(x) = \alpha_0 + (x - x_0) \left(\alpha_1 + (x - x_1) (\alpha_2 + \dots + (x - x_{n-1}) \alpha_n) \dots \right),$$

τότε μπορούμε να υπολογίσουμε την τιμή του πολυωνύμου παρεμβολής σε ένα δεδομένο σημείο x με βάση το σχήμα του Horner. Συγκεκριμένα, υπολογίζουμε διαδοχικά τις ποσότητες:

$$\begin{aligned} v_0 &= \alpha_n \\ v_1 &= \alpha_{n-1} + v_0(x - x_{n-1}) \\ v_2 &= \alpha_{n-2} + v_1(x - x_{n-2}) \\ &\vdots \\ v_n &= \alpha_0 + v_{n-1}(x - x_0) \end{aligned}$$

Τελικά, $v_n = p(x)$, έχοντας εκτελέσει $O(n)$ πολλαπλασιασμούς και προσθέσεις. Οι πράξεις αυτές μπορούν να υλοποιηθούν με τον αλγόριθμο που περιγράφεται στον ακόλουθο ψευδοκώδικα:

Αλγόριθμος 4.1**Είσοδος:** $\alpha_i, i = 0, 1, \dots, n, \quad x_i, i = 0, 1, \dots, n, \quad x$ $v \leftarrow \alpha_n$ **for** $i = n - 1, \dots, 0$ **do** $v \leftarrow \alpha_i + v(x - x_i)$ **end for**

β) Από την (4.4) βλέπουμε άμεσα ότι το πολυώνυμο παρεμβολής που αντιστοιχεί στα πρώτα k σημεία $(x_0, y_0), \dots, (x_{k-1}, y_{k-1})$, με $0 < k < n$, παράγεται από τους $k + 1$ πρώτους όρους του p , δηλαδή είναι το

$$q(x) = \alpha_0 + \alpha_1(x - x_0) + \dots + \alpha_k(x - x_0) \dots (x - x_{k-1}).$$

Επομένως μπορούμε να υπολογίσουμε με οικονομικό τρόπο τα πολυώνυμα παρεμβολής που προκύπτουν αν αφαιρέσουμε ή προσθέσουμε σημεία παρεμβολής. Παρατηρήστε ότι με τη μορφή Lagrange, στην αντίστοιχη περίπτωση, θα έπρεπε να υπολογίσουμε όλα τα πολυώνυμα Lagrange L_i εξ αρχής.

Παράδειγμα. Να προσδιορίσετε το πολυώνυμο παρεμβολής p_2 (βαθμού το πολύ 2) που παρεμβάλλεται στις τιμές της $f(x) = \frac{1}{x}$ στα σημεία $x_0 = 2, x_1 = 2.5, x_2 = 4$, κάνοντας χρήση της μορφής του Νεύτωνα.

Λύση. Το πολυώνυμο θα είναι της μορφής:

$$\begin{aligned} p(x) &= \alpha_0 + \alpha_1(x - x_0) + \alpha_2(x - x_0)(x - x_1) \\ &= \alpha_0 + \alpha_1(x - 2) + \alpha_2(x - 2)(x - 2.5). \end{aligned}$$

Επομένως

$$\begin{aligned} p(2) &= \frac{1}{2} &\Rightarrow & \alpha_0 = \frac{1}{2} \\ p(2.5) &= \frac{2}{5} &\Rightarrow & \frac{1}{2} + \alpha_1(2.5 - 2) = \frac{2}{5} \Rightarrow \alpha_1 = -\frac{1}{5} \\ p(4) &= \frac{1}{4} &\Rightarrow & \frac{1}{2} - \frac{1}{5}(4 - 2) + \alpha_2(4 - 2)(4 - 2.5) = \frac{1}{4} \Rightarrow \alpha_2 = \frac{1}{20}. \end{aligned}$$

Άρα το ζητούμενο πολυώνυμο παρεμβολής στη μορφή του Νεύτωνα είναι το:

$$p(x) = 0.5 - 0.2(x - 2) + 0.05(x - 2)(x - 2.5).$$

□

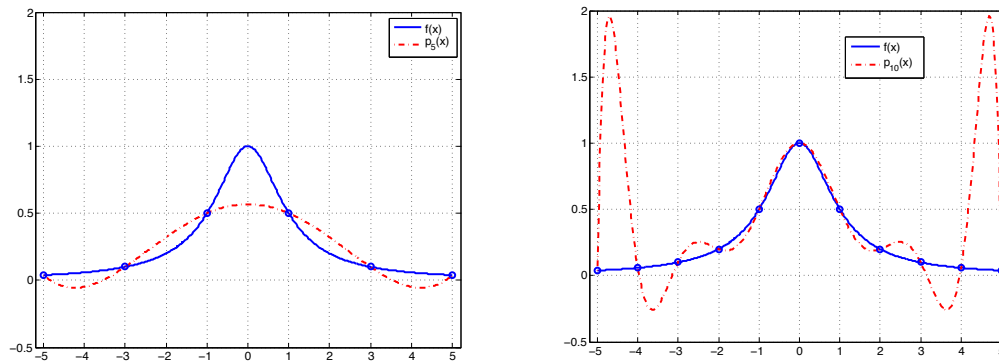
4.2 Σφάλματα στην πολυωνυμική παρεμβολή

Θεωρούμε μια συνάρτηση f σε ένα διάστημα $[a, b]$ και το πολυώνυμο p που παρεμβάλλεται στις τιμές της f σε $n + 1$ σημεία του διαστήματος $[a, b]$. Αρχικά ας υποθέσουμε ότι τα σημεία αυτά αποτελούν έναν ομοιόμορφο διαμερισμό του $[a, b]$, δηλαδή χωρίζουν το $[a, b]$ σε n ισομήκη διαστήματα που το καθένα έχει μήκος $h = (b - a)/n$. Προφανώς οι τιμές της f συμπίπτουν με τις τιμές του p στα σημεία αυτά. Διαισθητικά θα περίμενε κανείς ότι το p θα προσεγγίζει καλά την f και σε όλα τα ενδιάμεσα σημεία και, μάλιστα, ότι καθώς αυξάνουμε το πλήθος των σημείων παρεμβολής αυτή η προσέγγιση θα γίνεται ολοένα και καλύτερη. Παρόλα αυτά η προσδοκία αυτή αποδεικνύεται ότι είναι εντελώς αβάσιμη. Φυσικά, αν η f δεν είναι συνεχής, τότε είναι λογικό να μην υπάρχει συμφωνία μεταξύ των τιμών της f και του p εκτός από τα σημεία παρεμβολής. Εκείνο όμως που είναι πραγματικά αξιοπρόσεκτο είναι ότι το ίδιο μπορεί να συμβεί ακόμη και αν η f είναι μια πολύ ομαλή συνάρτηση με συνεχείς παραγώγους κάθε τάξης.

Ένα συγκεκριμένο παράδειγμα οφείλεται στον Runge που θεώρησε τη συνάρτηση

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

στο διάστημα $[-5, 5]$, η οποία είναι άπειρες φορές συνεχώς παραγωγίσιμη στο διάστημα αυτό. Στο Σχήμα 4.3 έχουμε σχεδιάσει με μπλέ συνεχή γραμμή τη συνάρτηση του Runge, ενώ με κόκκινη διακεκομμένη γραμμή το πολυώνυμο παρεμβολής p_5 που παρεμβάλλεται στις τιμές της f σε 6 ισαπέχοντα σημεία του $[-5, 5]$ (αριστερά), και το πολυώνυμο παρεμβολής p_{10} που παρεμβάλλεται στις τιμές της f σε 11 ισαπέχοντα σημεία του $[-5, 5]$ (δεξιά).



Σχήμα 4.3: Η συνάρτηση Runge και το πολυώνυμο παρεμβολής p_n σε $n + 1$ ισαπέχοντα σημεία του διαστήματος $[-5, 5]$. Αριστερά: $n = 5$. Δεξιά: $n = 10$.

Όπως βλέπετε στο σχήμα οι καμπύλες που υπολογίζουμε με το πολυώνυμο παίρνουν ακόμη και αρνητικές τιμές που φυσικά δεν έχει η f . Αν συνεχίσουμε να παίρνουμε ολοένα και περισ-

σότερα σημεία, δηλαδή ολοένα και μεγαλύτερου βαθμού πολυώνυμα, το μόνο που θα καταφέρουμε είναι να κάνουμε τα πράγματα ολοένα και χειρότερα: οι ταλαντώσεις θα γίνονται όλο και μεγαλύτερου πλάτους στα σημεία παρεμβολής που είναι κοντά στα άκρα του διαστήματος. Για την ακρίβεια, ισχύει ότι

$$\lim_{n \rightarrow \infty} \max_{-5 \leq x \leq 5} |f(x) - p_n(x)| = \infty.$$

Η αιτία για αυτή την προβληματική συμπεριφορά, κρύβεται στο ότι χρησιμοποιήσαμε ομοιόμορφο διαμερισμό. Υπάρχουν καλύτερες επιλογές για τα σημεία παρεμβολής και θα προτρέπαμε τον ενδιαφερόμενο αναγνώστη να ανατρέξει για περισσότερες πληροφορίες σε πιο προχωρημένα βιβλία Αριθμητικής Ανάλυσης, π.χ., [1, §4.3].

Το μήνυμα που θα πρέπει να κρατήσουμε από την παράγραφο αυτή είναι το εξής: Η παρεμβολή με πολυώνυμα μεγάλου βαθμού σε μεγάλο πλήθος σημείων παρεμβολής είναι μια «ριψοκίνδυνη» διαδικασία. Η προσέγγιση μπορεί να είναι πολύ κακή εκτός εάν επιλέξουμε τα σημεία παρεμβολής με πολύ μεγάλη προσοχή. Ακόμη όμως και αν το σφάλμα ήταν μικρό για μεγάλο n , θα είχαμε προβλήματα στον υπολογισμό των τιμών ενός πολυωνύμου μεγάλου βαθμού λόγω των σφαλμάτων στρογγύλευσης στον υπολογιστή. Σκεφτείτε ότι αν, π.χ., μας δίνανε 1000 σημεία θα έπρεπε κάθε φορά να υπολογίζουμε την τιμή ενός πολυωνύμου βαθμού 999! Βεβαίως η παρεμβολή με πολυώνυμα έχει πολλές άλλες εφαρμογές, όπως είναι ο προσδιορισμός τύπων αριθμητικής ολοκλήρωσης. Στην πράξη όμως η προσέγγιση συναρτήσεων γίνεται, σχεδόν πάντα, με τμηματικά πολυωνυμικές συναρτήσεις (splines), όπως θα δούμε στην επόμενη παράγραφο.

4.3 Παρεμβολή με splines

Η ιστορία των splines ξεκινάει από πολύ παλιά, από την ανάγκη των σχεδιαστών που συχνά χρειαζόταν να σχεδιάζουν καμπύλες που να διέρχονται από συγκεκριμένα σημεία. Για να το πετύχουν αυτό χρησιμοποιούσαν μηχανικές splines, που ήταν εύκαμπτες ταινίες από ελαστικό υλικό, συνήθως ξύλο, τις οποίες σταθεροποιούσαν στα σημεία παρεμβολής με βάρη (που στα αγγλικά λέγονται ducks λόγω του σχήματός τους), βλ. Σχήμα 4.4. (Η ορολογία αυτή χρονολογείται τουλάχιστον από το 1891.) Η ελαστική φύση αυτών των ξύλινων ταινιών τους επέτρεπε να λυγίζουν λίγο ώστε να περνάνε από τα προδιαγεγραμμένα σημεία, παίρνοντας ένα σχήμα που ελαχιστοποιούσε τη δυναμική ενέργεια. Η μαθηματική θεωρία των splines άρχισε να αναπτύσσεται από τα μέσα της δεκαετίας του 1940 και μετά. Το πρώτο βιβλίο στο οποίο υπάρχει μια συστηματική παρουσίαση των splines είναι το βιβλίο των Ahlberg, Nilson και Walsh, *The Theory of Splines and their Applications*, Academic Press, 1967.



Σχήμα 4.4: Μηχανική spline.

Στην παράγραφο αυτή θα περιγράψουμε εν συντομία τις splines από μαθηματική άποψη. Θεωρούμε ένα διαμερισμό του διαστήματος $[a, b]$:

$$a = x_0 < x_1 < x_2 < \dots < x_n = b.$$

Τα x_i λέγονται *κόμβοι* του διαμερισμού. Splines ως προς το διαμερισμό αυτό λέγονται γενικά συναρτήσεις, οι οποίες σε κάθε υποδιάστημα $[x_i, x_{i+1}]$ έχουν κάποια συγκεκριμένη μορφή, είναι, π.χ., πολυώνυμα βαθμού το πολύ m , ή τριγωνομετρικά πολυώνυμα κ.ο.κ. Εμείς θα ασχοληθούμε με την ακόλουθη συγκεκριμένη περίπτωση:

Θα λέμε ότι μια συνάρτηση s είναι μια *spline* βαθμού k αν:

1. Οι $s, s', s'', \dots, s^{(k-1)}$ είναι συνεχείς συναρτήσεις στο $[a, b]$.
2. Η s είναι πολυώνυμο βαθμού το πολύ k σε κάθε υποδιάστημα $[x_i, x_{i+1}]$.

Ειδικότερα, θα περιγράψουμε τις γραμμικές (βαθμού 1) splines, και τις κυβικές (βαθμού 3) splines.

4.3.1 Γραμμικές splines

Έστω $[a, b]$ ένα κλειστό και φραγμένο διάστημα, και $a = x_0 < x_1 < \dots < x_n = b$ ένας διαμερισμός του $[a, b]$. Μια συνάρτηση s θα λέγεται *γραμμική spline* αν

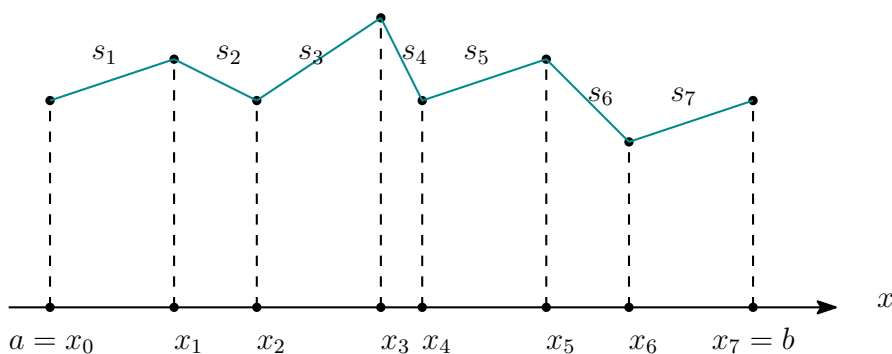
- είναι συνεχής σε ολόκληρο το $[a, b]$, και
- είναι πολυώνυμο πρώτου βαθμού σε κάθε υποδιάστημα $[x_i, x_{i+1}]$, $i = 0, 1, \dots, n-1$.

Προφανώς, μια γραμμική spline σε ένα διαμερισμό με οκτώ κόμβους θα μοιάζει όπως η τεθλασμένη γραμμή του Σχήματος 4.5. Μια τέτοια συνάρτηση θα έχει τύπο της μορφής:

$$s(x) = \begin{cases} s_1(x) & , \quad x \in [x_0, x_1] \\ s_2(x) & , \quad x \in [x_1, x_2] \\ \vdots & \vdots \\ s_n(x) & , \quad x \in [x_{n-1}, x_n] \end{cases}$$

όπου

$$s_i(x) = a_i x + b_i, \quad i = 1, 2, \dots, n.$$



Σχήμα 4.5: Γραμμική spline.

Για να υπολογίσουμε την τιμή της spline σε κάποιο ενδιαμέσο σημείο t , θα πρέπει πρώτα να εντοπίσουμε το υποδιάστημα στο οποίο ανήκει και μετά να υπολογίσουμε την τιμή της αντίστοιχης γραμμικής συνάρτησης.

Παράδειγμα. Ελέγξτε κατά πόσον η ακόλουθη συνάρτηση είναι γραμμική spline:

$$s(x) = \begin{cases} x + 1 & , \quad x \in [-1, 0] \\ -x & , \quad x \in (0, 1) \\ 2x - 3 & , \quad x \in [1, 2] \end{cases}$$

Λύση. Η s είναι προφανώς κατά τμήματα γραμμική αλλά είναι ασυνεχής στο $x = 0$. Πράγματι,

$$\lim_{x \rightarrow 0^+} s(x) = \lim_{x \rightarrow 0^+} (-x) = 0 \quad \text{ενώ} \quad \lim_{x \rightarrow 0^-} s(x) = \lim_{x \rightarrow 0^-} (x + 1) = 1.$$

Επομένως η s δεν είναι συνεχής σ' ολόκληρο το $[-1, 2]$ και άρα δεν είναι γραμμική spline. $\square$

Οι γραμμικές splines χρησιμοποιούνται ευρέως στις εφαρμογές αλλά έχουν ένα αρκετά εμφανές μειονέκτημα. Οι παράγωγοί τους είναι ασυνεχείς συναρτήσεις (κατά τμήματα σταθερές),

μιας και η κλίση της spline μπορεί να αλλάζει τιμή στους κόμβους. Επίσης, για να προσεγγίσουμε με ικανοποιητική ακρίβεια μια συνάρτηση με μια γραμμική spline μπορεί να χρειαστούμε έναν αρκετά λεπτό διαμερισμό, δηλαδή πολλούς κόμβους. Για τους λόγους αυτούς χρειάζεται συχνά να χρησιμοποιήσουμε splines μεγαλύτερου βαθμού.

4.3.2 Κυβικές splines

Σε μια κυβική spline «κολλάμε» μεταξύ τους κυβικά πολυώνυμα έτσι ώστε η συνάρτηση που προκύπτει να είναι συνεχής και να έχει συνεχείς πρώτες και δεύτερες παραγώγους σ' ολόκληρο το διάστημα. Επομένως σε κάθε εσωτερικό κόμβο επιβάλλουμε τρεις συνθήκες συνέχειας. Συγκεκριμένα:

Έστω $[a, b]$ ένα κλειστό και φραγμένο διάστημα, και $a = x_0 < x_1 < \dots < x_n = b$ ένας διαμερισμός του $[a, b]$. Μια συνάρτηση $s(x)$ θα λέγεται *κυβική spline* αν

- είναι δύο φορές συνεχώς παραγωγίσιμη σε ολόκληρο το $[a, b]$ (δηλαδή υπάρχουν οι s, s', s'' και είναι συνεχείς στο $[a, b]$), και
- είναι πολυώνυμο τρίτου βαθμού σε κάθε υποδιάστημα $[x_{i-1}, x_i], i = 1, \dots, n$.

Ας συμβολίζουμε το κυβικό πολυώνυμο του υποδιαστήματος $[x_{i-1}, x_i]$, ως $s_i(x), i = 1, \dots, n$. Π.χ.,

$$s_i(x) = a_i(x - x_{i-1})^3 + b_i(x - x_{i-1})^2 + c_i(x - x_{i-1}) + d_i.$$

Άρα σε καθένα από τα n υποδιαστήματα πρέπει να υπολογίσουμε 4 σταθερές (τα a_i, b_i, c_i, d_i), δηλ. συνολικά $4n$ σταθερές. Ας δούμε τις συνθήκες που έχουμε στη διάθεσή μας:

- Παρεμβολικές συνθήκες: $s(x_i) = f(x_i), i = 0, 1, \dots, n$.
($n + 1$ συνθήκες.)
- Συνέχεια της s στους εσωτερικούς κόμβους: $s_i(x_i) = s_{i+1}(x_i), i = 1, \dots, n - 1$. ($n - 1$ συνθήκες.)
- Συνέχεια της s' στους εσωτερικούς κόμβους: $s'_i(x_i) = s'_{i+1}(x_i), i = 1, \dots, n - 1$. ($n - 1$ συνθήκες.)
- Συνέχεια της s'' στους εσωτερικούς κόμβους: $s''_i(x_i) = s''_{i+1}(x_i), i = 1, \dots, n - 1$. ($n - 1$ συνθήκες.)

Άρα έχουμε συνολικά $4n - 2$ συνθήκες, και πρέπει να υπολογίσουμε $4n$ αγνώστους. Άρα χρειαζόμαστε ακόμα δύο συνθήκες που είναι, συνήθως, διαφορών τύπων συνοριακές συνθήκες για την s στα άκρα του διαστήματος $x_0 = a$, $x_n = b$. Π.χ.

- $s''(x_0) = 0$, $s''(x_n) = 0$. Στην περίπτωση αυτή λέμε ότι έχουμε μια **φυσική κυβική spline**.
- Συνθήκες πρώτης παραγώγου: $s'(x_0) = f'(x_0)$, $s'(x_n) = f'(x_n)$.
- Συνθήκες δεύτερης παραγώγου: $s''(x_0) = f''(x_0)$, $s''(x_n) = f''(x_n)$.

Παράδειγμα. Είναι η συνάρτηση $s(x) = \begin{cases} x^3 + x - 1 & , \quad -1 \leq x \leq 0 \\ x^3 - x - 1 & , \quad 0 < x \leq 1 \end{cases}$ μια κυβική spline;

Λύση. Για να είναι η s κυβική spline θα πρέπει να είναι πολυώνυμο βαθμού το πολύ 3 σε κάθε υποδιάστημα (κάτι που σε αυτή την περίπτωση είναι προφανές) και, επιπλέον, να είναι δύο φορές συνεχώς παραγωγίσιμη σε ολόκληρο το διάστημα (δηλαδή οι s , s' και s'' πρέπει να είναι συνεχείς σε ολόκληρο το $[-1, 1]$). Είναι προφανές ότι οι s , s' και s'' είναι συνεχείς παντού εκτός (πιθανώς) από το σημείο $x = 0$, άρα πρέπει να ελέγξουμε τι συμβαίνει στο $x = 0$.

$$\lim_{x \rightarrow 0^-} s(x) = \lim_{x \rightarrow 0^-} (x^3 + x - 1) = -1 \quad \text{και} \quad \lim_{x \rightarrow 0^+} s(x) = \lim_{x \rightarrow 0^+} (x^3 - x - 1) = -1$$

άρα η s είναι συνεχής στο $x = 0$. Επίσης

$$s'(x) = \begin{cases} 3x^2 + 1 & , \quad -1 \leq x \leq 0 \\ 3x^2 - 1 & , \quad 0 < x \leq 1, \end{cases}$$

επομένως

$$\lim_{x \rightarrow 0^-} s'(x) = \lim_{x \rightarrow 0^-} (3x^2 + 1) = 1 \quad \text{ενώ} \quad \lim_{x \rightarrow 0^+} s'(x) = \lim_{x \rightarrow 0^+} (3x^2 - 1) = -1$$

και άρα η s' δεν είναι συνεχής στο $x = 0$, οπότε η s **δεν** είναι κυβική spline. □

4.4 Παρεμβολή στο MATLAB

Στο MATLAB ένα πολυώνυμο

$$p(x) = a_1 x^n + a_2 x^{n-1} + \cdots + a_n x + a_{n-1},$$

αναπαριστάται από ένα διάνυσμα γραμμή που περιέχει τους συντελεστές του πολυωνύμου, $p = [p(1) \ p(2) \ \dots \ p(n+1)]$. (Παρατηρήστε ότι το πρώτο στοιχείο του p είναι ο συντελεστής του μεγιστοβάθμιου όρου, ενώ το τελευταίο ο σταθερός όρος.)

Τα ακόλουθα τρία προβλήματα αφορούν πολυώνυμα:

Υπολογισμός τιμής: Αν γνωρίζουμε τους συντελεστές του πολυωνύμου να υπολογίσουμε την τιμή του σε ένα ή περισσότερα σημεία.

Εύρεση ριζών: Αν γνωρίζουμε τους συντελεστές του πολυωνύμου να βρούμε τις ρίζες του (τα σημεία στα οποία η τιμή του πολυωνύμου είναι ίση με μηδέν).

Παρεμβολή: Αν μας δίνονται m σημεία (x_i, y_i) , $i = 1, \dots, m$ να υπολογίσουμε ένα πολυώνυμο που να παρεμβάλλεται στα δεδομένα αυτά.

Για να υπολογίσουμε την τιμή του πολυωνύμου p σε ένα σημείο x χρησιμοποιούμε τη συνάρτηση `polyval`, όπως φαίνεται στο ακόλουθο παράδειγμα που υπολογίζουμε την τιμή του πολυωνύμου $p(x) = x^2 + x - 6$ σε διάφορα σημεία. Για τον υπολογισμό της τιμής του πολυωνύμου σε ένα ή περισσότερα σημεία η `polyval` χρησιμοποιεί το σχήμα του Horner. Επιπλέον, για να βρούμε τις ρίζες του πολυωνύμου χρησιμοποιούμε τη συνάρτηση `roots`.

```
>> p = [1 1 -6]           % Δίνουμε τους συντελεστές του πολυωνύμου
p =
    1    1   -6
>> y = polyval(p,1)       % Υπολογίζουμε την τιμή του πολυωνύμου στο 1
y =
   -4
>> x = -2:2               % Ορίζουμε το διάνυσμα  $x = (-2, -1, 0, 1, 2)$ 
x =
   -2   -1    0    1    2
>> y = polyval(p,x)       % Υπολογίζουμε την τιμή του πολυωνύμου στο  $x$ 
y =
   -4   -6   -6   -4    0
>> z = roots(p)           % Βρίσκουμε τις ρίζες του πολυωνύμου  $p$ 
z =
   -3
    2
>> polyval(p,z)           % Επαληθεύουμε ότι οι ρίζες του  $p$  είναι στο  $z$ 
ans =
    0
    0
```

Επίσης, μπορούμε να πολλαπλασιάσουμε και να διαιρέσουμε πολυώνυμα χρησιμοποιώντας τις εντολές `conv` και `deconv`, αντίστοιχα. Όταν το πολυώνυμο p διαιρείται με το h βρί-

σκουμε ένα πολυώνυμο q (πηλίκο) και ένα πολυώνυμο r (υπόλοιπο), τέτοια ώστε $p(x) = h(x)q(x) + r(x)$ με βαθμό του $r \leq$ βαθμό του h . Στο παράδειγμα που ακολουθεί διαιρούμε το $x^3 - 4x^2 + 5x - 6$ δια το $x - 3$ και βρίσκουμε πηλίκο $x^2 - x + 2$ και υπόλοιπο 0. Στη συνέχεια πολλαπλασιάζουμε το πηλίκο h επί q για να ξαναβρούμε το αρχικό πολυώνυμο p .

```
>> p = [1 -4 5 -6];
>> h = [1 -3];
>> [q,r] = deconv(p,h)
q =
    1    -1     2
r =
    0     0     0     0
>> conv(h,q)
ans =
    1    -4     5    -6
```

Έστω, τώρα, ότι μας δίνονται m ζεύγη δεδομένων (x_i, y_i) , $i = 1, \dots, m$, όπου τα σημεία x_i είναι διαφορετικά ανά δύο. Στο MATLAB η συνάρτηση `polyfit` (σύνταξη: `p = polyfit(x,y,n)`) κάνει προσαρμογή των δεδομένων σε μια καμπύλη, υπολογίζοντας (με την έννοια των ελαχίστων τετραγώνων) το πολυώνυμο βαθμού n που ελαχιστοποιεί την ποσότητα: $\sum_{i=1}^m (p(x_i) - y_i)^2$. Αυτό είναι ένα θέμα που **δεν** αναπτύξαμε στις προηγούμενες παραγράφους. Αν όμως χρησιμοποιήσουμε την `polyfit` με $n = m - 1$, τότε μας επιστρέφει τους συντελεστές του (μοναδικού) πολυωνύμου βαθμού (το πολύ) $m - 1$, που παρεμβάλλεται στις τιμές y_i στα σημεία x_i , $i = 1, \dots, m$. Π.χ. αν θέλουμε να υπολογίσουμε και να σχεδιάσουμε το μοναδικό πολυώνυμο παρεμβολής βαθμού 2, που παρεμβάλλεται στις τιμές της συνάρτησης $f(x) = 1/x$ στα σημεία $x = 2, 2.5, 4$, γράφουμε τις εντολές:

```

>> x = [2 2.5 4]
x =
    2.0000  2.5000  4.0000
>> y = 1./x
y =
    0.5000  0.4000  0.2500
>> p = polyfit(x,y,2) % Υπολογισμός του πολ. παρεμβολής βαθμού 2
p =
    0.0500 -0.4250  1.1500
% Δημιουργία 41 ισαπεχόντων σημείων στο διάστημα [2, 4]
>> xx = linspace(2,4,41);
>> yy = 1./xx;
% Υπολογισμός των τιμών του πολ. παρεμβολής στα σημεία xx
>> z = polyval(p,xx);
>> plot(xx,yy,xx,z,'r',x,y,'k.','MarkerSize',20,'LineWidth',2)
>> grid on
>> legend('1/x','interpolant','interpolation points')

```

Στην Παράγραφο 4.3.2 είδαμε ότι αν έχουμε m ζεύγη (x_i, y_i) , $i = 1, \dots, m$, όπου τα σημεία x_i είναι διαφορετικά ανά δύο, τότε κυβική spline παρεμβολής s είναι μια συνάρτηση με τις ακόλουθες ιδιότητες:

- α) η s είναι δύο φορές συνεχώς παραγωγίσιμη στο $[x_1, x_m]$,
- β) η s είναι κυβικό πολυώνυμο σε κάθε υποδιάστημα $[x_i, x_{i+1}]$, $i = 1, \dots, m-1$, και
- γ) είναι συνάρτηση παρεμβολής, δηλ. $s(x_i) = y_i$, $i = 1, \dots, m$.

(Για τον πλήρη καθορισμό της χρειαζόμαστε δύο επιπλέον συνθήκες.) Στο MATLAB η κυβική spline παρεμβολής υπολογίζεται με τη συνάρτηση `spline(x,y,xx)`, η οποία υπολογίζει τις τιμές της κυβικής spline παρεμβολής στα σημεία `xx`. Σε αυτή τη μορφή η συνάρτηση χρησιμοποιεί ως επιπλέον συνθήκες στα άκρα τις λεγόμενες not-a-knot συνθήκες (που επιβάλλουν τη συνέχεια και της τρίτης παραγώγου στον πρώτο εσωτερικό και στον τελευταίο εσωτερικό κόμβο). Μπορείτε όμως π.χ. να καθορίσετε και την τιμή της 1ης παραγώγου στα άκρα της spline δίνοντας δύο επιπλέον τιμές στο διάνυσμα `y`. Π.χ. για να υπολογίσετε την κυβική spline παρεμβολής στις τιμές της $f(x) = \sin x$ στα σημεία $x = 0, \pi/2, \pi, 3\pi/2, 2\pi$ με συνθήκες not-a-knot γράφετε:

```
>> x = linspace(0,2*pi,5)
x =
    0  1.5708  3.1416  4.7124  6.2832
>> y = sin(x)
y =
    0  1.0000  0.0000 -1.0000 -0.0000
>> xx = linspace(0,2*pi,101);
>> yy = sin(xx);
>> z1 = spline(x,y,xx);
>> plot(x,y,'k.',xx,yy,xx,z1,'r','MarkerSize',20)
>> grid on
>> legend('interp. pts','sin(x)','cubic spline (not-a-knot)')
```

Αν όμως θέλετε να επιβάλλετε στη spline να έχει κλίση στα άκρα ίση με 1, θα πρέπει να δώσετε τις εντολές:

```
>> x = linspace(0,2*pi,5)
x =
    0  1.5708  3.1416  4.7124  6.2832
>> y = [1, sin(x), 1]
y =
    1.0000  0  1.0000  0.0000 -1.0000 -0.0000  1.0000
>> xx = linspace(0,2*pi,101);
>> yy = sin(xx);
>> z2 = spline(x,y,xx);
>> plot(x,y(2:6),'k.',xx,yy,xx,z2,'r','MarkerSize',20)
>> grid on
>> legend('interpolation points','sin(x)','cubic spline')
```

Τέλος, για να υπολογίσουμε τη γραμμική spline παρεμβολής στις τιμές της $f(x) = \sin x$ στα σημεία $x = 0, \pi/2, \pi, 3\pi/2, 2\pi$ μπορούμε να χρησιμοποιήσουμε τη συνάρτηση `interp1` με τον τρόπο που περιγράφεται παρακάτω:


```

>> x = linspace(0,2*pi,5)
x =
    0  1.5708  3.1416  4.7124  6.2832
>> y = sin(x)
y =
    0  1.0000  0.0000 -1.0000 -0.0000
>> xx = linspace(0,2*pi,101);
>> yy = sin(xx);
>> z3 = interp1(x,y,xx);
>> plot(x,y,'k.',xx,yy,xx,z3,'r','MarkerSize',20)
>> grid on
>> legend('interpolation points','sin(x)','linear spline')

```

4.5 Συμπληρωματικές παρατηρήσεις

Θεωρούμε μια συνάρτηση f σε ένα διάστημα $[a, b]$ και το (μοναδικό) πολυώνυμο p , βαθμού $\leq n$, που παρεμβάλλεται στις τιμές της f στα, διαφορετικά ανά δύο, $n + 1$ σημεία του διαστήματος $[a, b]$: $x_0, \dots, x_n$ (βλ. Παράγρ. 4.1). Τότε, αν η f είναι μια αρκετά ομαλή συνάρτηση μπορούμε να βρούμε μια αναλυτική παράσταση του σφάλματος $f(x) - p(x)$, καθώς και μια εκτίμηση του σφάλματος, όπως διατυπώνονται στο ακόλουθο θεώρημα:

Θεώρημα 4.3. Έστω f μια συνάρτηση με συνεχείς παραγώγους μέχρι και τάξης $n + 1$ ($n \geq 0$) στο $[a, b]$, και $x_0, \dots, x_n \in [a, b]$ ανά δύο διαφορετικά μεταξύ τους σημεία. Αν p είναι το (μοναδικό) πολυώνυμο παρεμβολής της f στα σημεία $x_0, \dots, x_n$, τότε ισχύει:

$$\forall x \in [a, b] \quad \exists \xi \in (a, b) : \quad f(x) - p(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^n (x - x_i), \quad (4.5)$$

όπου συμβολίζουμε:

$$\prod_{i=0}^n (x - x_i) = (x - x_0)(x - x_1) \cdots (x - x_n).$$

Επιπλέον,

$$\max_{a \leq x \leq b} |f(x) - p(x)| \leq \max_{a \leq x \leq b} \left| \prod_{i=0}^n (x - x_i) \right| \frac{\max_{a \leq x \leq b} |f^{(n+1)}(x)|}{(n+1)!}. \quad (4.6)$$

Απόδειξη. Αν x είναι κάποιο από τα $x_0, x_1, \dots, x_n$, τότε η (4.5) ισχύει κατά προφανή τρόπο.

Έστω $x \in [a, b]$ και $x \neq x_i$, για $i = 0, 1, \dots, n$. Θέτουμε $\Phi(t) = \prod_{i=0}^n (t - x_i)$ και ορίζουμε τη

βοηθητική συνάρτηση

$$\varphi(t) := f(t) - p(t) - \frac{f(x) - p(x)}{\Phi(x)} \Phi(t).$$

Αφού η f είναι συνάρτηση με συνεχείς παραγώγους μέχρι και τάξης $n + 1$ (συμβολικά γράφουμε $f \in C^{n+1}[a, b]$) και η Φ είναι ένα πολώνυμο βαθμού $n + 1$, συμπεραίνουμε ότι και η συνάρτηση $\varphi \in C^{n+1}[a, b]$. Επίσης

$$\varphi(x_i) = \underbrace{f(x_i) - p(x_i)}_{=0} - \frac{f(x) - p(x)}{\Phi(x)} \underbrace{\Phi(x_i)}_{=0} = 0, \quad i = 0, 1, \dots, n,$$

και

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

Άρα η συνάρτηση φ έχει τουλάχιστον $n + 2$ ρίζες στο $[a, b]$, επομένως από το Θεώρημα του Rolle η φ' έχει τουλάχιστον $n + 1$ ρίζες στο (a, b) , η φ'' έχει τουλάχιστον n ρίζες στο (a, b) , κ.ο.κ. Τέλος, η $\varphi^{(n+1)}$ έχει τουλάχιστον μία ρίζα στο (a, b) , έστω ξ .

Δηλαδή, υπάρχει $\xi \in (a, b)$ έτσι ώστε $\varphi^{(n+1)}(\xi) = 0$.

Επειδή, προφανώς, $\Phi^{(n+1)}(t) = (n + 1)!$, έχουμε ότι:

$$\varphi^{(n+1)}(t) = f^{(n+1)}(t) - \frac{f(x) - p(x)}{\Phi(x)} (n + 1)!, \quad t \in [a, b].$$

Συνεπώς,

$$\begin{aligned} \varphi^{(n+1)}(\xi) = 0 &\Rightarrow f^{(n+1)}(\xi) - \frac{f(x) - p(x)}{\Phi(x)} (n + 1)! = 0 \\ &\Rightarrow f(x) - p(x) = \frac{f^{(n+1)}(\xi)}{(n + 1)!} \Phi(x). \end{aligned}$$

Φυσικά,

$$\begin{aligned} |f(x) - p(x)| &= \frac{|f^{(n+1)}(\xi)|}{(n + 1)!} \left| \prod_{i=0}^n (x - x_i) \right| \\ \Rightarrow |f(x) - p(x)| &\leq \max_{a \leq x \leq b} \left| \prod_{i=0}^n (x - x_i) \right| \frac{\max_{a \leq x \leq b} |f^{(n+1)}(x)|}{(n + 1)!}, \quad \text{για κάθε } x \in [a, b] \\ \Rightarrow \max_{a \leq x \leq b} |f(x) - p(x)| &\leq \max_{a \leq x \leq b} \left| \prod_{i=0}^n (x - x_i) \right| \frac{\max_{a \leq x \leq b} |f^{(n+1)}(x)|}{(n + 1)!}. \end{aligned}$$

□

Παράδειγμα. Ας υπολογίσουμε το μέγιστο σφάλμα που θα έχουμε αν προσεγγίσουμε μια συνάρτηση f που είναι δύο φορές συνεχώς παραγωγίσιμη σ' ένα διάστημα $[a, b]$ με το γραμμικό πολώνυμο που παρεμβάλλεται στις τιμές της f στα σημεία $x = a$ και $x = b$.

Το γραμμικό πολυώνυμο παρεμβολής που παρεμβάλλεται στα σημεία $x = a$ και $x = b$ στις τιμές $f(a)$ και $f(b)$ είναι το:

$$p(x) = \frac{f(b) - f(a)}{b - a} (x - a) + f(a). \quad (\text{γιατί;})$$

Άρα από την εκτίμηση (4.6) έχουμε ότι:

$$\begin{aligned} \max_{a \leq x \leq b} |f(x) - p(x)| &\leq \frac{1}{2!} \underbrace{\max_{a \leq x \leq b} |(x - a)(x - b)|}_{\frac{1}{4}(b-a)^2} \max_{a \leq x \leq b} |f''(x)| \\ &= \frac{(b - a)^2}{8} \max_{a \leq x \leq b} |f''(x)|. \end{aligned}$$

Ασκήσεις

1. Θεωρήστε τη συνάρτηση $f(x) = \sqrt{2}x \cos x$. Υπολογίστε το πολυώνυμο παρεμβολής $p_2(x)$, βαθμού ≤ 2 , που παρεμβάλλεται στην f στα σημεία $x_0 = 0$, $x_1 = \pi/4$ και $x_2 = \pi/2$ του διαστήματος $[0, \pi/2]$.
2. Ελέγξτε ότι τα πολυώνυμα $p(x) = 5x^3 - 27x^2 + 45x - 21$ και $q(x) = x^4 - 5x^3 + 8x^2 - 5x + 3$, παρεμβάλλονται στις τιμές

x	1	2	3	4
y	2	1	6	47

Εξηγήστε γιατί αυτό δεν αντιβαίνει στην μοναδικότητα του πολυωνύμου παρεμβολής που διατυπώνεται στο Θεώρημα 4.2.

3. Ελέγξτε ότι τα πολυώνυμα $p(x) = 4x^2 + 6x - 7$ και $q(x) = 3 + 2(x - 1) + 4(x - 1)(x + 2)$, παρεμβάλλονται στις τιμές

x	1	-2	0
y	3	-3	-7

Εξηγήστε γιατί αυτό δεν αντιβαίνει στην μοναδικότητα του πολυωνύμου παρεμβολής που διατυπώνεται στο Θεώρημα 4.2.

4. Θεωρήστε τα σημεία $x_0 = -1$, $x_1 = 1$, $x_2 = 3$, $x_3 = 4$, και υπολογίστε τα πολυώνυμα Lagrange L_i , $i = 0, 1, 2, 3$, (βλ. (4.2)) ως προς τα σημεία αυτά. Με τη βοήθεια του MATLAB σχεδιάστε τις γραφικές παραστάσεις των πολυωνύμων αυτών.

5. Έστω $x_0, x_1, x_2 \in \mathbb{R}$ ανα δύο διαφορετικά μεταξύ τους σημεία και L_0, L_1, L_2 τα πολυώνυμα Lagrange (δευτέρου το πολύ βαθμού) ως προς τα σημεία αυτά. Αποδείξτε ότι για $i = 0, 1, 2$ ισχύει:

$$x_0^i L_0(x) + x_1^i L_1(x) + x_2^i L_2(x) = x^i, \quad \text{για κάθε } x \in \mathbb{R}.$$

6. Ο πληθυσμός της Ελλάδας για τα έτη 1981, 1991 και 2001, δίνεται στον ακόλουθο πίνακα

x (Έτος)	1981	1991	2001
y (Πληθυσμός)	9, 740, 000	10, 260, 000	10, 960, 000

Εκτιμήστε τον πληθυσμό της Ελλάδας κατά το έτος 1986,

- (α') με το γραμμικό πολυώνυμο παρεμβολής $p_1(x)$ που παρεμβάλλεται στις τιμές του πληθυσμού τα έτη $x_0 = 1981$ και $x_1 = 1991$,
 (β') με το πολυώνυμο παρεμβολής $p_2(x)$ (βαθμού ≤ 2) που παρεμβάλλεται στις τιμές του πληθυσμού τα έτη $x_0 = 1981, x_1 = 1991$ και $x_2 = 2001$.

7. Ελέγξτε αν η

$$s(x) = \begin{cases} x & , \quad -1 \leq x \leq 0.5 \\ 0.5 + 2(x - 0.5) & , \quad 0.5 < x \leq 2 \\ x + 1.5 & , \quad 2 < x \leq 4 \end{cases}$$

είναι μια γραμμική spline;

8. Υπάρχουν $a, b, c, d \in \mathbb{R}$ τέτοια ώστε η συνάρτηση

$$s(x) = \begin{cases} ax^3 + x^2 + cx & , \quad -1 \leq x \leq 0 \\ bx^3 + x^2 + dx & , \quad 0 < x \leq 1 \end{cases}$$

να είναι μια φυσική κυβική spline που παρεμβάλλεται στις τιμές της συνάρτησης $|x|$ στα σημεία $-1, 0, 1$;

9. Υπάρχει κυβική spline s ως προς τον διαμερισμό $\{-1, 0, 1, 2, 3\}$ του διαστήματος $[-1, 3]$, τέτοια ώστε $s(x) = 0$ για $x \in [-1, 0]$, $s(x) = 0$ για $x \in [2, 3]$, και $s(1) = 1$; Αν ναι, κατασκευάστε την.

10. Υπολογίστε τις παραμέτρους $a, b, c, d, e \in \mathbb{R}$ έτσι ώστε η συνάρτηση

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

να είναι μια φυσική κυβική spline.

Κεφάλαιο 5

Αριθμητική ολοκλήρωση

Θα ξεκινήσουμε το κεφάλαιο αυτό θυμίζοντας μερικά βασικά στοιχεία για τα ολοκληρώματα. Διακρίνουμε δύο βασικά είδη ολοκληρωμάτων: το ορισμένο ολοκλήρωμα και το αόριστο ολοκλήρωμα. Το *ορισμένο ολοκλήρωμα* μιας συνάρτησης f σε ένα διάστημα $[a, b]$ συμβολίζεται με $\int_a^b f(x) dx$ και είναι αριθμός (το όριο αθροισμάτων Riemann), ενώ το *αόριστο ολοκλήρωμα*, $\int f(x) dx$, είναι συνάρτηση (για την ακρίβεια άπειρες συναρτήσεις ταυτόχρονα). Τα δύο αυτά ολοκληρώματα συνδέονται μέσω του ακόλουθου θεωρήματος:

Θεώρημα 5.1 (Το Θεμελιώδες Θεώρημα του Απειροστικού Λογισμού). Έστω ότι η $y = f(x)$ είναι ορισμένη στο διάστημα I , συνεχής, και ολοκληρώσιμη σε κάθε κλειστό και φραγμένο υποδιάστημα J του. Παίρνουμε οποιοδήποτε a στο I και θεωρούμε τη συνάρτηση

$$F(x) = \int_a^x f(t) dt.$$

Τότε η $y = F(x)$ είναι παραγωγίσιμη στο I και ισχύει $F'(x) = f(x)$ για κάθε x στο I .

Το αποτέλεσμα αυτό μπορούμε να το γράψουμε και ως εξής:

$$\frac{d}{dx} \int_a^x f(t) dt = f(x),$$

όπου η f είναι συνεχής. Με άλλα λόγια, μπορούμε να πούμε ότι αν, αρχικά, ολοκληρώσουμε την f , και στη συνέχεια παραγωγίσουμε αυτό που θα προκύψει, τότε θα πάρουμε και πάλι την αρχική συνάρτηση f . Δηλαδή, οι πράξεις της παραγωγίσιμης και της ολοκλήρωσης συναρτήσεων είναι, ουσιαστικά, αντίστροφες. Η μια αναιρεί την άλλη.

Η επόμενη πρόταση είναι άμεση συνέπεια του Θεμελιώδους Θεωρήματος και έχει σπουδαία πρακτική σημασία διότι μας επιτρέπει να υπολογίζουμε ορισμένα ολοκληρώματα χωρίς να μπαίνουμε στη διαδικασία να υπολογίζουμε όρια αθροισμάτων Riemann.

Πρόταση 5.1. Έστω ότι η $y = f(x)$ είναι συνεχής στο διάστημα I και έστω $y = F(x)$ μια αντιπαράγωγος της $y = f(x)$ στο I . Τότε το ολοκλήρωμα της $y = f(x)$ σε οποιοδήποτε υποδιάστημα $[a, b]$ του I είναι ίσο με τη διαφορά των τιμών της $y = F(x)$ στα άκρα του $[a, b]$. Δηλαδή,

$$\int_a^b f(x) dx = F(b) - F(a) \quad (a, b \text{ στο } I).$$

Βεβαιώς, ένα πρόβλημα είναι ότι στην πράξη σπάνια μπορούμε να υπολογίζουμε αναλυτικά την αντιπαράγωγο F μιας συνάρτησης f , ώστε στη συνέχεια να χρησιμοποιήσουμε την Πρόταση 5.1 για να υπολογίσουμε το ορισμένο ολοκλήρωμα. Ακόμη όμως και αν γνωρίζουμε την αντιπαράγωγο F είναι δυνατόν μια εύκολα υπολογιζόμενη f να έχει πολύπλοκη παράγουσα F . Π.χ., αν

$$f(x) = 1/(1 + x^4),$$

τότε μια παράγουσα της είναι η

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

Σε αυτές τις περιπτώσεις καταφεύγουμε στους λεγόμενους τύπους αριθμητικής ολοκλήρωσης, ώστε να προσεγγίσουμε την τιμή του ολοκληρώματος.

Γενικά, στην αριθμητική ολοκλήρωση, το ολοκλήρωμα

$$I(f) = \int_a^b f(x) dx,$$

προσεγγίζεται με τύπους ολοκλήρωσης της μορφής:

$$Q_{n+1}(f) = w_0 f(x_0) + w_1 f(x_1) + \dots + w_n f(x_n).$$

- Τα x_i λέγονται **κόμβοι** και (συνήθως) είναι σημεία στο διάστημα ολοκλήρωσης, ενώ
- τα w_i λέγονται **βάρη** και είναι πραγματικοί αριθμοί.

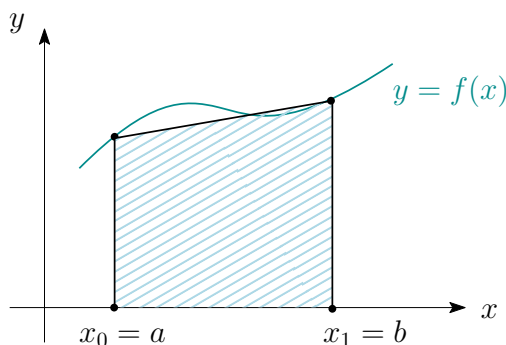
Θα λέμε ότι ένας τύπος αριθμητικής ολοκλήρωσης είναι *ακριβής για πολυώνυμα βαθμού k* αν ολοκληρώνει ακριβώς πολυώνυμα βαθμού το πολύ k ή, ισοδύναμα, αν ολοκληρώνει ακριβώς τα μονώνυμα $1, x, \dots, x^k$.

Εμείς στη συνέχεια θα δούμε δύο βασικούς τύπους αριθμητικής ολοκλήρωσης.

5.1 Ο τύπος του τραπεζίου

Ο ακόλουθος τύπος ολοκλήρωσης με δύο κόμβους είναι ο λεγόμενος τύπος του τραπεζίου:

$$I(f) = \int_a^b f(x) dx \approx Q_2(f) = \frac{b-a}{2} [f(a) + f(b)]. \quad (5.1)$$



Σχήμα 5.1: Ο τύπος του τραπεζίου.

Το όνομα οφείλεται στο ότι προσεγγίζουμε το εμβαδόν του χωρίου που βρίσκεται κάτω από μια καμπύλη με το εμβαδόν ενός τραπεζίου, βλ. Σχήμα 5.1.

Στον τύπο του τραπεζίου μπορούμε να καταλήξουμε αν προσεγγίσουμε τη συνάρτηση f με το γραμμικό πολυώνυμο παρεμβολής, που παρεμβάλλεται στις τιμές της f στα άκρα του διαστήματος a και b , και στη συνέχεια ολοκληρώσουμε το πολυώνυμο αυτό. Μπορούμε άμεσα να ελέγξουμε ότι:

ο τύπος του τραπεζίου ολοκληρώνει ακριβώς πολυώνυμα μέχρι και πρώτου βαθμού.

Θεωρούμε τώρα έναν διαμερισμό του $[a, b]$: $a = x_0 < x_1 < \dots < x_n = b$, σε n υποδιαστήματα. Εφαρμόζουμε σε κάθε υποδιάστημα $[x_i, x_{i+1}]$ τον τύπο του τραπεζίου, οπότε αθροίζοντας οδηγούμαστε στο **σύνθετο τύπο του τραπεζίου**:

$$Q_{n+1}^T(f) = \frac{1}{2} \sum_{i=0}^{n-1} (x_{i+1} - x_i) [f(x_i) + f(x_{i+1})], \quad (5.2)$$

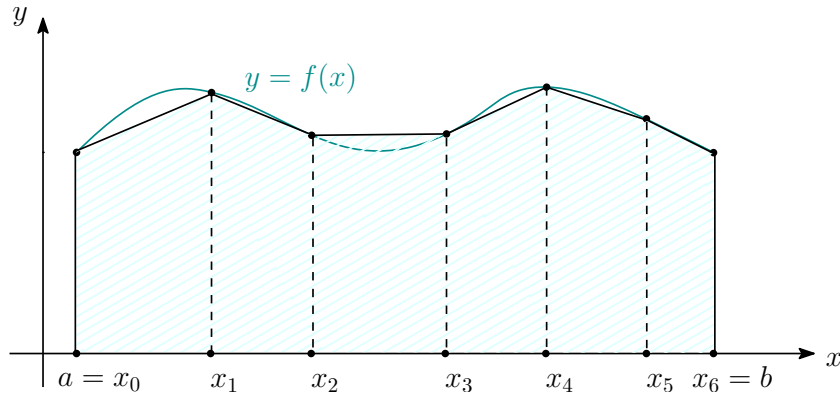
βλ. Σχήμα 5.2.

Αν, επιπλέον, θεωρήσουμε ότι ο διαμερισμός του $[a, b]$ είναι ομοιόμορφος με βήμα h ,

$$h = \frac{b-a}{n}, \quad x_i = a + ih, \quad i = 0, \dots, n.$$

τότε ο **σύνθετος τύπος του τραπεζίου** παίρνει τη μορφή:

$$Q_{n+1}^T(f) = h \left(\frac{1}{2} f(x_0) + f(x_1) + \dots + f(x_{n-1}) + \frac{1}{2} f(x_n) \right). \quad (5.3)$$



Σχήμα 5.2: Ο σύνθετος τύπος του τραπεζίου.

Αποδεικνύεται ότι αν η συνάρτηση f είναι αρκετά ομαλή, συγκεκριμένα, δύο φορές συνεχώς παραγωγίσιμη στο $[a, b]$, τότε μια εκτίμηση του σφάλματος της μεθόδου (5.3) δίνεται από τον τύπο:

$$|I(f) - Q_{n+1}^T(f)| \leq \frac{b-a}{12} h^2 \max_{a \leq x \leq b} |f''(x)|. \quad (5.4)$$

Για το λόγο αυτό λέμε ότι η τάξη του σφάλματος του σύνθετου τύπου του τραπεζίου για ομαλές συναρτήσεις είναι τουλάχιστον 2.

Παράδειγμα. Υπολογίστε μια προσέγγιση του $\int_0^\pi \sin x \, dx$ με το σύνθετο τύπο του τραπεζίου με 4 υποδιαστήματα. Υπολογίστε την απόλυτη τιμή του σφάλματος.

Λύση. Καταρχάς, η ακριβής τιμή του ολοκληρώματος είναι

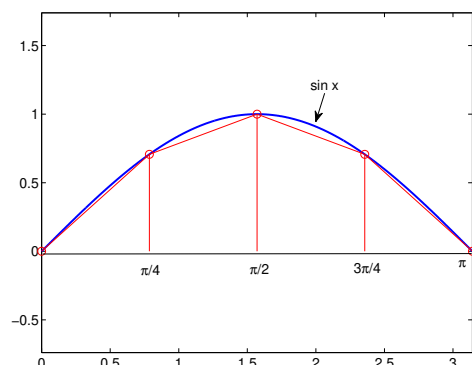
$$\int_0^\pi \sin x \, dx = [-\cos x]_0^\pi = -\cos \pi + \cos 0 = -(-1) + 1 = 2.$$

Χωρίζουμε το διάστημα σε 4 υποδιαστήματα μήκους $h = \frac{\pi-0}{4} = \frac{\pi}{4}$ το καθένα, και ονομάζουμε τα άκρα των υποδιαστημάτων ως εξής:

$$x_0 = 0, \quad x_1 = \frac{\pi}{4}, \quad x_2 = \frac{\pi}{2}, \quad x_3 = \frac{3\pi}{4}, \quad x_4 = \pi.$$

Τότε ο σύνθετος τύπος του τραπεζίου είναι

$$\begin{aligned} \int_0^\pi \sin x \, dx &\approx h \left(\frac{1}{2} f(x_0) + f(x_1) + f(x_2) + f(x_3) + \frac{1}{2} f(x_4) \right) \\ &= \frac{\pi}{4} \left(\frac{1}{2} \sin 0 + \sin \frac{\pi}{4} + \sin \frac{\pi}{2} + \sin \frac{3\pi}{4} + \frac{1}{2} \sin \pi \right) \\ &= \frac{\pi}{4} \left(0 + \frac{\sqrt{2}}{2} + 1 + \frac{\sqrt{2}}{2} + 0 \right) = \frac{(1 + \sqrt{2})\pi}{4} \approx 1.896. \end{aligned}$$



Σχήμα 5.3: Ο σύνθετος τύπος του τραπεζίου για την $f(x) = \sin x$ σε 4 υποδιαστήματα του $[0, \pi]$.

Το απόλυτο σφάλμα είναι

$$\left| 2 - \frac{(1 + \sqrt{2})\pi}{4} \right| = 0.104 \approx 10\%,$$

ενώ το σχετικό σφάλμα είναι (κατ' απόλυτη τιμή) ίσο με

$$\left| \frac{2 - \frac{(1 + \sqrt{2})\pi}{4}}{2} \right| = 0.0519 \approx 5\%.$$

□

Παράδειγμα. Έστω ότι θέλουμε να προσεγγίσουμε το ολοκλήρωμα

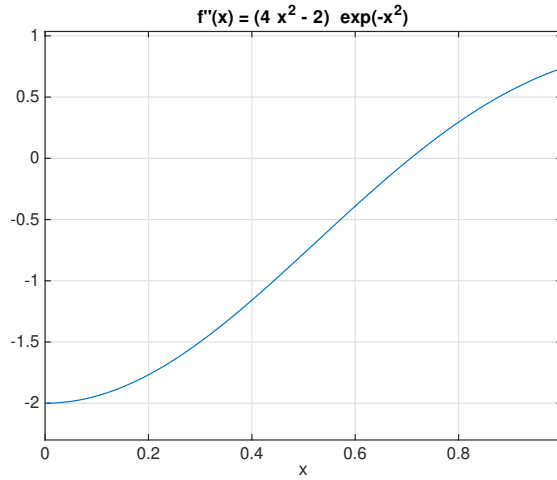
$$\int_0^1 e^{-x^2} dx,$$

με το σύνθετο τύπο του τραπεζίου σε έναν ομοιόμορφο διαμερισμό του διαστήματος $[0, 1]$ με $n + 1$ κόμβους (n υποδιαστήματα). Με τη βοήθεια της εκτίμησης (5.4) μπορούμε να προσδιορίσουμε εκ των προτέρων το πλήθος των κόμβων που απαιτούνται ώστε να προσεγγίσουμε το ολοκλήρωμα αυτό με ακρίβεια τουλάχιστον 10^{-5} . Καταρχάς, έχουμε ότι

$$f(x) = e^{-x^2} \Rightarrow f'(x) = -2xe^{-x^2} \Rightarrow f''(x) = (4x^2 - 2)e^{-x^2}.$$

Δεν είναι δύσκολο να ελέγξουμε (βλ. και Σχήμα 5.4) ότι για κάθε $x \in [0, 1]$ ισχύει:

$$|f''(x)| = |4x^2 - 2|e^{-x^2} \leq 2.$$



Σχήμα 5.4: Γραφική παράσταση της $f''(x) = (4x^2 - 2)e^{-x^2}$, $x \in [0, 1]$.

Άρα από την (5.4) θα έχουμε για $a = 0$ και $b = 1$ ότι το σφάλμα της προσέγγισης θα είναι μικρότερο από

$$\frac{1}{12} h^2 \max_{a \leq x \leq b} |f''(x)| \leq \frac{1}{6} h^2.$$

Επομένως, αν απαιτήσουμε

$$\frac{1}{6} h^2 \leq 10^{-5} \Leftrightarrow \frac{1}{6} \cdot \frac{1}{n^2} \leq 10^{-5} \Leftrightarrow n^2 \geq \frac{10}{6} \cdot 10^4 \Leftrightarrow n \geq 129,$$

συμπεραίνουμε ότι με περισσότερους από 130 ομοιόμορφα κατανεμημένους κόμβους στο διάστημα $[0, 1]$ θα προσεγγίσουμε το ολοκλήρωμα αυτό με ακρίβεια τουλάχιστον 10^{-5} . $\square$

Παρατήρηση. Είδαμε ότι η τάξη της μεθόδου του τραpezίου για ομαλές συναρτήσεις είναι τουλάχιστον 2. Αν όμως η συνάρτηση που θέλουμε να ολοκληρώσουμε δεν είναι ομαλή, τότε μπορεί η τάξη της μεθόδου να είναι μικρότερη του 2.

Γενικά, ένας τρόπος για να προσδιορίσουμε *αριθμητικά* την τάξη της μεθόδου του τραpezίου για μια συγκεκριμένη συνάρτηση f είναι ο ακόλουθος. Έστω $\mathcal{E}(h)$ η απόλυτη τιμή του σφάλματος του σύνθετου τύπου του τραpezίου για έναν ομοιόμορφο διαμερισμό του διαστήματος ολοκλήρωσης με βήμα h , $h = (b - a)/n$.

Ας υποθέσουμε ότι

$$\mathcal{E}(h) \approx Ch^p,$$

όπου η σταθερά C είναι ανεξάρτητη του h αλλά εξαρτάται φυσικά από την f .

(Μια τέτοια σχέση ισχύει μόνο κατά προσέγγιση γιατί, γενικά, εμφανίζονται και όροι του h σε μεγαλύτερη δύναμη, αλλά, αν θεωρήσουμε ότι το h είναι αρκετά μικρό, μας επιτρέπει να υπολογίσουμε αριθμητικά το p .)

Τότε, π.χ., θα έχουμε ότι

$$\mathcal{E}(h/2) \approx C \left(\frac{h}{2}\right)^p,$$

επομένως

$$\frac{\mathcal{E}(h)}{\mathcal{E}(h/2)} \approx \frac{Ch^p}{C \left(\frac{h}{2}\right)^p} = 2^p \Rightarrow p \approx \frac{\log \left(\frac{\mathcal{E}(h)}{\mathcal{E}(h/2)} \right)}{\log 2}. \quad (5.5)$$

Για παράδειγμα, ας προσεγγίσουμε το ολοκλήρωμα

$$\int_0^1 \sqrt{x} dx,$$

με το σύνθετο τύπο του τραπεζίου χρησιμοποιώντας ομοιόμορφους διαμερισμούς του διαστήματος $[0, 1]$ με n υποδιαστήματα μήκους $h = 1/n$ το καθένα. Προφανώς, η συνάρτηση $f(x) = \sqrt{x}$ είναι ολοκληρώσιμη στο $[0, 1]$ αλλά δεν είναι παραγωγίσιμη στο 0. Για να υπολογίσουμε αριθμητικά την τάξη σύγκλισης υπολογίζουμε την απόλυτη τιμή του σφάλματος για $n = 2, 4, 8, \dots, 64$ υποδιαστήματα και χρησιμοποιούμε την (5.5) για να υπολογίσουμε αριθμητικά την τάξη σύγκλισης. Τα αποτελέσματα παρατίθενται στον ακόλουθο πίνακα:

n	h	$\mathcal{E}(h)$	p
2	1/2	$0.63113 \cdot 10^{-1}$	1.4324
4	1/4	$0.23384 \cdot 10^{-1}$	
8	1/8	$0.85364 \cdot 10^{-2}$	1.4538
16	1/16	$0.30855 \cdot 10^{-2}$	
32	1/32	$0.11077 \cdot 10^{-2}$	1.4681
64	1/64	$0.39586 \cdot 10^{-3}$	

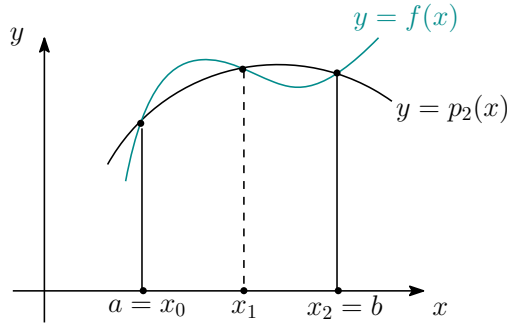
Τα αποτελέσματα αυτά υποδεικνύουν ότι καθώς το h μικραίνει ($h \rightarrow 0$) το p φαίνεται να προσεγγίζει τον αριθμό $3/2$. (Στην περίπτωση αυτή μπορεί κανείς να αποδείξει και θεωρητικά ότι η τάξη σύγκλισης είναι πράγματι $3/2$.)

5.2 Ο τύπος του Simpson

Ο τύπος του Simpson είναι ένας τύπος αριθμητικής ολοκλήρωσης με τρεις κόμβους. Όταν τον εφαρμόζουμε στο διάστημα $[a, b]$ χρησιμοποιούμε ως κόμβους τα άκρα του διαστήματος

$x_0 = a$, $x_2 = b$ και το μέσον του $x_1 = \frac{1}{2}(a + b)$, και παίρνει τη μορφή:

$$I(f) = \int_a^b f(x) dx \approx Q_3(f) = \frac{b-a}{2} \left[\frac{1}{3}f(a) + \frac{4}{3}f\left(\frac{a+b}{2}\right) + \frac{1}{3}f(b) \right]. \quad (5.6)$$



Σχήμα 5.5: Ο τύπος του Simpson.

Στον τύπο του Simpson θα καταλήξουμε αν προσεγγίσουμε τη συνάρτηση f με το πολυώνυμο παρεμβολής δευτέρου βαθμού, που παρεμβάλλεται στις τιμές της f στα άκρα του διαστήματος $x_0 = a$, $x_2 = b$ και στο μέσον του $x_1 = \frac{1}{2}(a + b)$, και στη συνέχεια ολοκληρώσουμε το πολυώνυμο αυτό. Μπορούμε άμεσα να ελέγξουμε ότι:

ο τύπος του Simpson ολοκληρώνει ακριβώς πολυώνυμα μέχρι και τρίτου βαθμού.

Αν τώρα θεωρήσουμε έναν ομοιόμορφο διαμερισμό του $[a, b]$, σε *άρτιο* πλήθος υποδιαστημάτων, με βήμα h ,

$$h = \frac{b-a}{n}, \quad x_i = a + ih, \quad i = 0, \dots, n, \quad n \text{ άρτιος},$$

και εφαρμόσουμε στα υποδιαστήματα $[x_0, x_2]$, $[x_2, x_4]$, $\dots$, $[x_{n-2}, x_n]$ τον τύπο του Simpson, τότε αθροίζοντας οδηγούμαστε στο **σύνθετο τύπο του Simpson**

$$Q_{n+1}^S(f) = \frac{h}{3} \left[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) + \dots + 4f(x_{n-1}) + f(x_n) \right]. \quad (5.7)$$

Για το σύνθετο τύπο του Simpson (5.7) αποδεικνύεται ότι αν η συνάρτηση f είναι τέσσερις φορές συνεχώς παραγωγίσιμη στο $[a, b]$, τότε ισχύει η ακόλουθη εκτίμηση σφάλματος:

$$|I(f) - Q_{n+1}^S(f)| \leq \frac{b-a}{180} h^4 \max_{a \leq x \leq b} |f^{(4)}(x)|. \quad (5.8)$$

5.3 Αριθμητική ολοκλήρωση στο MATLAB

Οι βασικές συναρτήσεις του MATLAB για τον υπολογισμό του ορισμένου ολοκληρώματος $\int_a^b f(x)dx$ είναι οι `quad` και `quadl`. Και στις δύο απαιτείται τα όρια ολοκλήρωσης a , b να είναι πεπερασμένα, ενώ η συνάρτηση f πρέπει να είναι ομαλή στο $[a, b]$. Η συνάρτηση `quad` υλοποιεί έναν αλγόριθμο *αυτόματης αριθμητικής ολοκλήρωσης* που βασίζεται στον τύπο του Simpson. Τέτοιου τύπου αλγόριθμοι «σπάνε» το διάστημα ολοκλήρωσης σε υποδιαστήματα, και σε κάθε υποδιάστημα χρησιμοποιούν κάποιο βασικό τύπο αριθμητικής ολοκλήρωσης. Η επιλογή του μήκους του κάθε υποδιαστήματος γίνεται ανάλογα με την «τοπική» συμπεριφορά της συνάρτησης f . Ουσιαστικά, επιλέγονται μικρότερα υποδιαστήματα εκεί όπου η f παρουσιάζει απότομες μεταβολές. Το πρόγραμμα μας προειδοποιεί σε περίπτωση όπου το μήκος των υποδιαστημάτων γίνεται πολύ μικρό, ή όταν το πλήθος των υπολογισμών της συνάρτησης αυξάνεται υπερβολικά. (Για μια περιγραφή της βασικής ιδέας των αλγορίθμων αυτόματης αριθμητικής ολοκλήρωσης παραπέμπουμε στο [1, Παράγρ. 6.6].) Ανάλογα ισχύουν και για τη συνάρτηση `quadl` που υλοποιεί έναν ακριβέστερο τύπο αυτόματης αριθμητικής ολοκλήρωσης τεσσάρων σημείων.

Ας δούμε ένα παράδειγμα υπολογισμού ολοκληρώματος με τη βοήθεια του MATLAB. Θέλουμε να υπολογίσουμε το ολοκλήρωμα:

$$\int_0^1 h(x) dx, \quad \text{όπου } h(x) = \frac{1}{(x - 0.3)^2 + 0.01} + \frac{1}{(x - 0.9)^2 + 0.04} - 6.$$

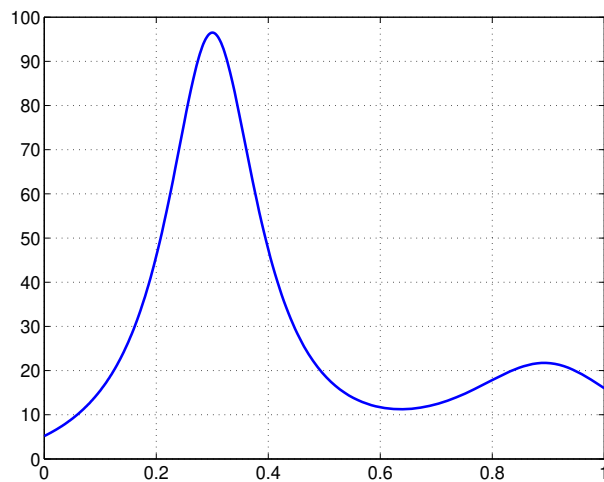
Η h υλοποιείται στο MATLAB με τη συνάρτηση `humps`. Για να δούμε τη γραφική της παράσταση, που δίνεται στο Σχήμα 5.6, γράφουμε στο παράθυρο εντολών του MATLAB:

```
>> fplot(@humps,[0,1])
>> grid on
```

Για να υπολογίσουμε το ολοκλήρωμά της με την `quad` γράφουμε:

```
>> q = quad(@humps,0,1)
q =
    29.8583
```

Στο MATLAB μπορούμε ακόμη να πάρουμε μια προσέγγιση της ακριβούς τιμής του ολοκληρώματος με τη συνάρτηση `trapz`, που υλοποιεί το σύνθετο τύπο του τραpezίου. Στην `trapz`, σε αντίθεση με τις προηγούμενες δύο συναρτήσεις, δίνουμε τους κόμβους x_i και τις τιμές της συνάρτησης $f(x_i)$. Π.χ.



Σχήμα 5.6: Γράφημα της συνάρτησης humps.

```
>> x = 0:0.1:1;           % Ορίζουμε το διάνυσμα x = (0, 0.1, ..., 0.9, 1)
>> y = humps(x);          % Υπολογίζουμε τις τιμές h(x)
>> q = trapz(x,y)
q =
    29.8517
```

Ασκήσεις

- (i) Υπολογίστε τα βάρη w_0, w_1, w_2 στον τύπο αριθμητικής ολοκλήρωσης, έτσι ώστε να ολοκληρώνει ακριβώς πολυώνυμα όσο το δυνατόν μεγαλύτερου βαθμού,

$$\int_0^4 f(x) dx \approx w_0 f(1) + w_1 f(2) + w_2 f(3).$$

Ποιος είναι ο μέγιστος βαθμός πολυωνύμου που ο τύπος αυτός ολοκληρώνει ακριβώς;

(ii) Χρησιμοποιήστε τον τύπο αριθμητικής ολοκλήρωσης του προηγούμενου ερωτήματος για να υπολογίσετε μια προσέγγιση του ολοκληρώματος $\int_0^4 \sqrt{x} dx$. Πόσο είναι το σφάλμα;

- Υπολογίστε τα βάρη w_0, w_1 στον τύπο αριθμητικής ολοκλήρωσης $Q(f) = w_0 f(1/2) + w_1 f(1)$, έτσι ώστε να ολοκληρώνει στο $[-1, 1]$ πολυώνυμα μέχρι και πρώτου βαθμού ακριβώς.

3. Ο ακόλουθος τύπος ολοκλήρωσης λέγεται (αριστερός) τύπος του ορθογωνίου

$$I(f) = \int_a^b f(x) dx \approx Q(f) := (b-a) f(a).$$

- (α') Αποδείξτε ότι ο τύπος Q ολοκληρώνει ακριβώς σταθερές συναρτήσεις (πολυνόμοιο μηδενικού βαθμού).
- (β') Διατυπώστε το σύνθετο (αριστερό) τύπο του ορθογωνίου για έναν ομοιόμορφο διαμερισμό του διαστήματος $[a, b]$ με βήμα $h = \frac{b-a}{n}$, $x_i = a + ih$, $i = 0, \dots, n$.
- (γ') Υπολογίστε μια προσέγγιση του $\int_0^\pi \sin x dx$ με το σύνθετο (αριστερό) τύπο του ορθογωνίου για έναν ομοιόμορφο διαμερισμό του διαστήματος $[0, \pi]$ με βήμα $h = \pi/4$. Υπολογίστε την απόλυτη τιμή του σφάλματος.

4. Ο ακόλουθος τύπος ολοκλήρωσης λέγεται τύπος του μέσου

$$I(f) = \int_a^b f(x) dx \approx Q(f) := (b-a) f\left(\frac{a+b}{2}\right).$$

- (α') Αποδείξτε ότι ο τύπος Q ολοκληρώνει ακριβώς πολυώνυμα μέχρι και πρώτου βαθμού.
- (β') Διατυπώστε το σύνθετο τύπο του μέσου για έναν ομοιόμορφο διαμερισμό του διαστήματος $[a, b]$ με βήμα $h = \frac{b-a}{n}$, $x_i = a + ih$, $i = 0, \dots, n$.
- (γ') Υπολογίστε μια προσέγγιση του $\int_0^\pi \sin x dx$ με το σύνθετο τύπο του μέσου για έναν ομοιόμορφο διαμερισμό του διαστήματος $[0, \pi]$ με βήμα $h = \pi/2$. Υπολογίστε την απόλυτη τιμή του σφάλματος.

5. Βρείτε έναν αναλυτικό τύπο για την προσέγγιση του ολοκληρώματος $\int_0^1 x^2 dx$ με το σύνθετο τύπο του τραπεζίου για έναν ομοιόμορφο διαμερισμό του διαστήματος $[0, 1]$ με $n+1$ (ισαπέχοντα) σημεία. Θα χρειαστεί να χρησιμοποιήσετε τον τύπο:

$$1^2 + 2^2 + \dots + n^2 = \frac{1}{6} n(n+1)(2n+1).$$

- (α') Υπολογίστε την προσέγγιση του ολοκληρώματος για $n = 100$.
- (β') Αποδείξτε ότι καθώς $n \rightarrow \infty$ η προσεγγιστική τιμή του ολοκληρώματος τείνει στη σωστή τιμή που (φυσικά) είναι $1/3$.

6. Υπολογίστε το πλήθος των κόμβων που πρέπει να πάρουμε για να προσεγγίσουμε το ολοκλήρωμα $\int_0^\pi \sin x dx$ με το σύνθετο τύπο του τραπεζίου για έναν ομοιόμορφο διαμερισμό του διαστήματος $[0, \pi]$ με σφάλμα $\leq 10^{-6}$.

7. Βρείτε μια εκτίμηση της απόλυτης τιμής του σφάλματος που θα προκύψει αν προσεγγίσουμε το ολοκλήρωμα $\int_2^5 \sin x \, dx$ με το σύνθετο τύπο του τραπεζίου για έναν ομοιόμορφο διαμερισμό με βήμα $h = 0.01$.
8. Υπολογίστε μια προσεγγίση του ολοκληρώματος $\int_1^2 f(x) \, dx$ με το σύνθετο τύπο του τραπεζίου. Δίνονται οι ακόλουθες τιμές:

x	1	$\frac{5}{4}$	$\frac{3}{2}$	$\frac{7}{4}$	2
$f(x)$	10	8	7	6	5

9. Χρησιμοποιήστε τον τύπο του Simpson με κόμβους $x_0 = 0$, $x_1 = 0.5$ και $x_2 = 1$, για να υπολογίσετε την τιμή του ολοκληρώματος

$$\int_0^1 (1+x^2)^{-1} \, dx.$$

Συγκρίνετε την τιμή αυτή με την ακριβή τιμή του ολοκληρώματος.

10. Δίνονται οι ακόλουθες τιμές μιας συνάρτησης f :

x	1	1.25	1.5	1.75	2
$f(x)$	10	8	7	6	5

Υπολογίστε μια προσεγγίση του ολοκληρώματος $\int_1^2 f(x) \, dx$ με το σύνθετο τύπο του Simpson.

11. Υπολογίστε το πλήθος των κόμβων που πρέπει να πάρουμε για να προσεγγίσουμε το ολοκλήρωμα $\int_0^\pi \sin x \, dx$ με τον σύνθετο τύπο του Simpson για έναν ομοιόμορφο διαμερισμό του διαστήματος $[0, \pi]$ με σφάλμα $\leq 10^{-6}$.
12. Χρησιμοποιήστε τον σύνθετο τύπο του Simpson για να βρείτε μια προσέγγιση του ολοκληρώματος $\int_1^2 x^2 \ln x \, dx$ με σφάλμα μικρότερο από 10^{-4} . (Η ακριβής τιμή του ολοκληρώματος είναι ίση με $(8 \ln 8 - 7)/9 \simeq 1.070615$.)

Κεφάλαιο 6

Προβλήματα Αρχικών Τιμών για Συνήθεις Διαφορικές Εξισώσεις

Στο κεφάλαιο αυτό θα ασχοληθούμε με αριθμητικές μεθόδους για την επίλυση προβλημάτων αρχικών τιμών για συνήθεις διαφορικές εξισώσεις. Συγκεκριμένα, θα ασχοληθούμε με το ακόλουθο πρόβλημα: Δεδομένης μιας συνάρτησης $f(t, y)$ αναζητούμε μια συνάρτηση $y(t)$, $a \leq t \leq b$, η οποία είναι λύση της συνήθους διαφορικής εξίσωσης (Σ.Δ.Ε.)

$$\frac{dy}{dt} = f(t, y), \quad \text{για } a \leq t \leq b, \quad (6.1)$$

και ικανοποιεί την αρχική συνθήκη

$$y(a) = y_0, \quad (6.2)$$

όπου y_0 είναι δοσμένη πραγματική σταθερά.

Η εξίσωση (6.1) μαζί με τη συνθήκη (6.2), λέμε ότι αποτελούν ένα *πρόβλημα αρχικών τιμών* (Π.Α.Τ.). Σε μαθήματα Σ.Δ.Ε. εξετάζεται το πρόβλημα (6.1)–(6.2) και, ειδικότερα, συνθήκες για την f οι οποίες εξασφαλίζουν ύπαρξη ή/και μοναδικότητα της λύσης του Π.Α.Τ..

Προβλήματα της μορφής (6.1)–(6.2) εμφανίζονται πολύ συχνά στις εφαρμογές. Επίσης είναι σύνηθες η παράμετρος t να εκφράζει χρόνο και η Δ.Ε. να είναι η μαθηματική έκφραση ενός φυσικού νόμου που διέπει τις μεταβολές του συστήματος που μελετούμε στο χρόνο. Τα περισσότερα προβλήματα αυτής της μορφής δεν μπορούν να λυθούν αναλυτικά, δηλαδή να βρούμε μαθηματικούς τύπους που να εκφράζουν τη λύση τους, και για το λόγο αυτό χρησιμοποιούμε αριθμητικές μεθόδους για να υπολογίσουμε προσεγγίσεις της λύσης. Μερικές τέτοιες μεθόδους θα περιγράψουμε στις επόμενες παραγράφους.

6.1 Στοιχεία Θεωρίας Προβλημάτων Αρχικών Τιμών για Συνήθεις Διαφορικές Εξισώσεις

Στη σύντομη αυτή εισαγωγική παράγραφο θα παρουσιάσουμε μερικά βασικά στοιχεία από τη θεωρία των συνήθων διαφορικών εξισώσεων που θα μας χρειαστούν στη συνέχεια. Θα ξεκινήσουμε με τον ακόλουθο ορισμό:

Ορισμός 6.1. Μια συνάρτηση $f(t, y)$ λέμε ότι ικανοποιεί τη συνθήκη του Lipschitz ως προς y ομοιόμορφα ως προς t σε ένα σύνολο $D \subset \mathbb{R}^2$, αν υπάρχει σταθερά $L > 0$ τέτοια ώστε

$$|f(t, y_1) - f(t, y_2)| \leq L |y_1 - y_2|, \quad (6.3)$$

για κάθε $(t, y_1), (t, y_2) \in D$. Η L λέγεται σταθερά του Lipschitz για την f .

Παράδειγμα. $D = \{(t, y) : 1 \leq t \leq 2, -3 \leq y \leq 4\}$ και $f(t, y) = t|y|$.

Έστω $(t, y_1), (t, y_2) \in D$. Τότε

$$|f(t, y_1) - f(t, y_2)| = |t|y_1| - t|y_2|| = |t| ||y_1| - |y_2|| \leq 2|y_1 - y_2|.$$

Η $L = 2$ είναι η μικρότερη δυνατή τιμή για τη σταθερά Lipschitz σε αυτό το πρόβλημα μιας και

$$|f(2, 1) - f(2, 0)| = |2 - 0| = 2|1 - 0| = 2.$$

□

Προφανώς, είναι επιθυμητό πριν ξεκινήσουμε να λύσουμε αριθμητικά το Π.Α.Τ. (6.1)–(6.2) να γνωρίζουμε κατά πόσον το πρόβλημα αυτό έχει λύση και αν η λύση αυτή είναι μοναδική. (Γενικά, είναι πολύ δύσκολο να προσεγγίσουμε αριθμητικά λύσεις προβλημάτων που δεν λύνονται μονοσήμαντα.) Το επόμενο θεώρημα μας εξασφαλίζει την ύπαρξη και μοναδικότητα της λύσης του Π.Α.Τ. (6.1)–(6.2) αν, φυσικά, πληρούνται κάποιες προϋποθέσεις.

Θεώρημα 6.1. Έστω $D = \{(t, y) : a \leq t \leq b, y \in \mathbb{R}\}$ και έστω $f(t, y)$ συνεχής στο D . Αν η συνάρτηση f ικανοποιεί τη συνθήκη του Lipschitz ως προς y ομοιόμορφα ως προς t στο D , τότε το Π.Α.Τ.

$$\begin{cases} y'(t) = f(t, y(t)), & a \leq t \leq b, \\ y(a) = y_0 \end{cases}$$

έχει μοναδική λύση $y(t)$ για $t \in [a, b]$.

Παράδειγμα. Θεωρήστε το Π.Α.Τ.:

$$\begin{cases} y'(t) = 1 + t \sin(ty), & 0 \leq t \leq 2, \\ y(0) = 0. \end{cases}$$

Η $f(t, y) = 1 + t \sin(ty)$ είναι συνεχής. Επιπλέον,

$$|f(t, y_1) - f(t, y_2)| = |1 + t \sin(ty_1) - 1 - t \sin(ty_2)| = |t| |\sin(ty_1) - \sin(ty_2)|.$$

Όμως λόγω της απλής τριγωνομετρικής ταυτότητας $\sin A - \sin B = 2 \sin \frac{A-B}{2} \cos \frac{A+B}{2}$, και της κλασσικής ανισότητας $|\sin x| \leq |x|$, για κάθε $x \in \mathbb{R}$, έχουμε ότι:

$$\begin{aligned} \sin(ty_1) - \sin(ty_2) &= 2 \sin \frac{ty_1 - ty_2}{2} \cos \frac{ty_1 + ty_2}{2} \\ \Rightarrow |\sin(ty_1) - \sin(ty_2)| &= 2 \underbrace{\left| \sin \frac{t(y_1 - y_2)}{2} \right|}_{\leq \frac{|t||y_1 - y_2|}{2}} \underbrace{\left| \cos \frac{t(y_1 + y_2)}{2} \right|}_{\leq 1}. \end{aligned}$$

Άρα,

$$|f(t, y_1) - f(t, y_2)| \leq |t| 2 \frac{|t||y_1 - y_2|}{2} = t^2 |y_1 - y_2| \leq 4 |y_1 - y_2|,$$

και επομένως το Π.Α.Τ. έχει μοναδική λύση. □

Παρατήρηση 6.1. Η συνθήκη (6.3) είναι γνωστή ως «ολική» συνθήκη Lipschitz και είναι εν γένει πολύ περιοριστική. Για παράδειγμα, επιβεβαιώστε ότι ακόμη και μια απλή συνάρτηση όπως η $f(t, y) = y^2$ δεν την ικανοποιεί. Συχνά μπορούμε να αντικαταστήσουμε την (6.3) με μια «τοπική» συνθήκη Lipschitz, την οποία ικανοποιούν πολύ περισσότερες συναρτήσεις. Αυτό μας επιτρέπει να αποδείξουμε ύπαρξη και μοναδικότητα λύσης σε ένα υποδιάστημα $[a, b']$ του διαστήματος $[a, b]$, για κατάλληλο b' . Για περισσότερες λεπτομέρειες παραπέμπουμε στο βιβλίο των Ακρίβη-Δουγαλή [1, Θεώρ. 8.2].

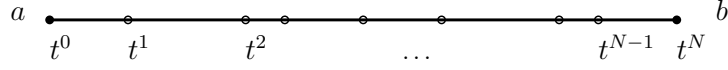
6.2 Η μέθοδος του Euler

Στην παράγραφο αυτή θα παρουσιάσουμε την απλούστερη αριθμητική μέθοδο για την αριθμητική επίλυση Π.Α.Τ. για Σ.Δ.Ε.. Η μέθοδος αυτή χρησιμοποιείται σπάνια στην πράξη, είναι όμως πολύ χρήσιμη για την παρουσίαση και κατανόηση διαφόρων εννοιών λόγω της απλότητάς της.

Στην παράγραφο αυτή θα θεωρήσουμε το ακόλουθο Π.Α.Τ.

$$\begin{cases} y'(t) = f(t, y(t)), & a \leq t \leq b, \\ y(a) = y_0, \end{cases} \quad (6.4)$$

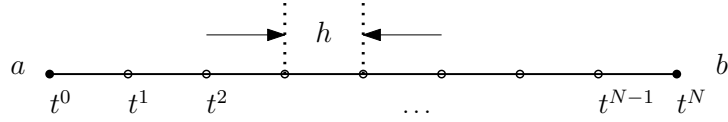
για το οποίο θα υποθέσουμε ότι έχει μοναδική λύση.

Σχήμα 6.1: Διαμερισμός του διαστήματος $[a, b]$.

Θεωρούμε έναν διαμερισμό $a = t^0 < t^1 < \dots < t^{N-1} < t^N = b$ του διαστήματος $[a, b]$, βλ. Σχήμα 6.1.

Ο στόχος μας είναι να κατασκευάσουμε μια αριθμητική μέθοδο για την επίλυση του προβλήματος (6.4) που θα μας επιτρέπει να υπολογίζουμε *προσεγγίσεις* της ακριβούς λύσης στους λεγόμενους *κόμβους* t^n , $n = 0, 1, \dots, N$. Στη συνέχεια, θα συμβολίζουμε με y^n την προσέγγιση της ακριβούς λύσης $y(t^n)$ στον κόμβο t^n . Δηλαδή, $y^n \approx y(t^n)$.

Επιπλέον, για τη συνέχεια, θα θεωρήσουμε έναν *ομοιόμορφο* διαμερισμό του διαστήματος $[a, b]$, δηλαδή θα υποθέσουμε ότι, δεδομένου ενός $N \in \mathbb{N}$, κάθε υποδιάστημα $[t^{n-1}, t^n]$, $n = 1, 2, \dots, N$, έχει μήκος $h = \frac{b-a}{N}$.

Σχήμα 6.2: Ομοιόμορφος διαμερισμός του διαστήματος $[a, b]$.

Άρα, σε αυτή την περίπτωση οι κόμβοι είναι $t^n := a + nh$, $n = 0, 1, 2, \dots, N$, βλ. Σχήμα 6.2.

6.2.1 Τρόποι κατασκευής της μεθόδου του Euler

Ένας τρόπος για την κατασκευή της μεθόδου του Euler είναι ο ακόλουθος: Καταρχάς, γράφουμε τη διαφορική εξίσωση της (6.4) στο σημείο $t = t^n$:

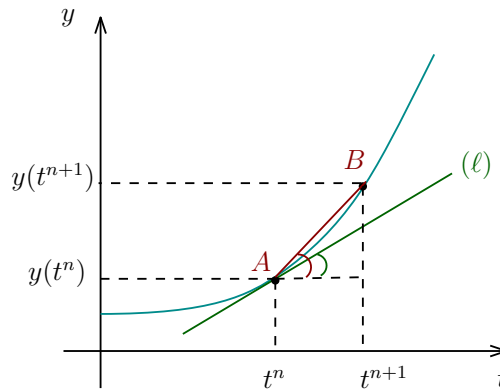
$$y'(t^n) = f(t^n, y(t^n)). \quad (6.5)$$

Στη συνέχεια, αν θεωρήσουμε ότι το h είναι «μικρό» μπορούμε να προσεγγίσουμε την παράγωγο στην εξίσωση (6.5) με το αντίστοιχο πηλίκο διαφορών, βλ. Σχήμα 6.3, καταλήγοντας στη σχέση:

$$\frac{y(t^{n+1}) - y(t^n)}{h} \approx y'(t^n). \quad (6.6)$$

Επομένως, από τις (6.5) και (6.6) έχουμε ότι:

$$\frac{y(t^{n+1}) - y(t^n)}{h} \approx f(t^n, y(t^n)),$$



Σχήμα 6.3: Προσέγγιση της παραγώγου $y'(t^n)$ με το πηλίκο διαφορών $\frac{y(t^{n+1}) - y(t^n)}{h}$.

και, αν υποθέσουμε ότι το y^n είναι μια καλή προσέγγιση του $y(t^n)$, θα προσεγγίσουμε τον αριθμό $y(t^{n+1})$ με τον αριθμό y^{n+1} , όπου

$$\frac{y^{n+1} - y^n}{h} = f(t^n, y^n) \Leftrightarrow y^{n+1} = y^n + h f(t^n, y^n).$$

Έτσι καταλήγουμε στο ακόλουθο αριθμητικό σχήμα:

Μέθοδος του Euler

$$\begin{aligned} y^{n+1} &= y^n + h f(t^n, y^n), \quad n = 0, 1, 2, \dots, N-1, \\ y^0 &= y_0. \end{aligned} \quad (6.7)$$

Δηλαδή, οι προσεγγίσεις $y^1, \dots, y^N$, της μεθόδου του Euler για έναν ομοιόμορφο διαμερισμό με «βήμα» h , υπολογίζονται με τον αναδρομικό τύπο (6.7), με $y^0 = y_0$.

Η μέθοδος του Euler υλοποιείται πολύ εύκολα με τον αλγόριθμο που περιγράφεται στον ακόλουθο ψευδοκώδικα:

Οι πράξεις που απαιτούνται για τη μέθοδο του Euler σε κάθε βήμα είναι ένας πολλαπλασιασμός, μια πρόσθεση και ένας υπολογισμός της f στο σημείο (t^n, y^n) . Συνήθως η f είναι μια πολύπλοκη συνάρτηση και το κόστος υπολογισμού μιας τιμής της σε ένα σημείο είναι κατά κανόνα πολύ υψηλότερο από το κόστος εκτέλεσης ενός πολλαπλασιασμού και μιας πρόσθεσης. Για το λόγο αυτό έχει επικρατήσει ως κόστος ανά βήμα να μετράμε μόνο το πλήθος των υπολογισμών της f . Η μέθοδος του Euler έχει πολύ χαμηλό κόστος, μόλις έναν υπολογισμό της f ανά βήμα.

Αλγόριθμος 6.1

Είσοδος: a, b, y_0, N
Εξωτερική συνάρτηση f
 $h \leftarrow (b - a)/N$
 $t \leftarrow a$
 $y \leftarrow y_0$
for $k = 1, \dots, N$ **do**
 $y \leftarrow y + h f(t, y)$
 $t \leftarrow t + h$
end for

Ένας δεύτερος τρόπος κατασκευής της μεθόδου του Euler (και άλλων αριθμητικών μεθόδων) βασίζεται στη στενή σχέση που υπάρχει μεταξύ της διαδικασίας επίλυσης μιας διαφορικής εξίσωσης και της ολοκλήρωσης. Συγκεκριμένα, θεωρούμε τη Δ.Ε.

$$y'(t) = f(t, y(t)).$$

Ολοκληρώνουμε και τα δύο μέλη από t^n μέχρι t^{n+1} και έχουμε

$$\begin{aligned} \int_{t^n}^{t^{n+1}} y'(t) dt &= \int_{t^n}^{t^{n+1}} f(t, y(t)) dt \\ \Rightarrow y(t^{n+1}) - y(t^n) &= \int_{t^n}^{t^{n+1}} f(t, y(t)) dt. \end{aligned} \quad (6.8)$$

Στη συνέχεια, αν προσεγγίσουμε το ολοκλήρωμα στο δεξί μέλος με κάποιον τύπο αριθμητικής ολοκλήρωσης (σαν αυτούς που είδαμε στο προηγούμενο κεφάλαιο) καταλήγουμε σε κάποιον τύπο για την αριθμητική επίλυση της Δ.Ε.. Συγκεκριμένα, η προσέγγιση του ολοκληρώματος με τον αριστερό τύπο του ορθογωνίου, βλ. Σχήμα 6.4,

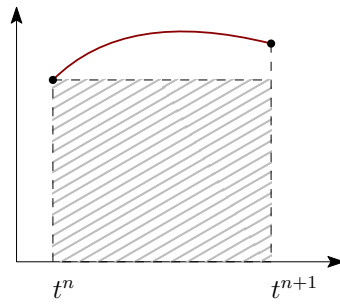
$$\int_{t^n}^{t^{n+1}} f(t, y(t)) dt \approx h f(t^n, y(t^n))$$

μας οδηγεί και πάλι στη σχέση

$$y(t^{n+1}) - y(t^n) \approx h f(t^n, y(t^n)),$$

και, επομένως, στη μέθοδο του Euler (6.7).

Τέλος, ένας τρίτος τρόπος κατασκευής της μεθόδου του Euler βασίζεται στον τύπο του Taylor.



Σχήμα 6.4: Προσέγγιση του ολοκληρώματος $\int_{t^n}^{t^{n+1}} f(t, y(t)) dt$ με τον αριστερό τύπο του ορθογωνίου.

Θεωρούμε ότι η λύση $y(t)$ της Δ.Ε. μπορεί να αναπαρασταθεί με τον τύπο του Taylor ως εξής:

$$\begin{aligned} y(t+h) &= y(t) + y'(t)h + \frac{y''(\eta)}{2}h^2 \\ &= y(t) + h y'(t) + O(h^2) \\ &\stackrel{\Delta.E.}{=} y(t) + h f(t, y(t)) + O(h^2). \end{aligned}$$

Άρα, αν το h είναι «αρκετά μικρό», έχουμε για $t = t^n$ ότι:

$$y(t^{n+1}) = y(t^n + h) \approx y(t^n) + h f(t^n, y(t^n)),$$

οπότε και πάλι καταλήγουμε στη μέθοδο του Euler (6.7).

Παράδειγμα. Θεωρήστε το Π.Α.Τ.

$$\begin{cases} y'(t) = -y + t + 1, & 0 \leq t \leq 1, \\ y(0) = 1. \end{cases}$$

Θεωρούμε έναν ομοιόμορφο διαμερισμό του $[0, 1]$ σε $N = 10$ υποδιαστήματα, μήκους $h = (1-0)/10 = 0.1$ το καθένα. Άρα οι κόμβοι του διαμερισμού είναι $t^n = 0.1n$, $n = 0, 1, \dots, 10$.

Τότε οι προσεγγίσεις της λύσης με τη μέθοδο του Euler δίνονται από τον αναδρομικό τύπο (6.7), δηλαδή

$$y^{n+1} = y^n + h f(t^n, y^n) = y^n + h (-y^n + t^n + 1),$$

όπου $n = 0, 1, \dots, 10$ και $y^0 = y_0 = 1$. Επομένως:

$$\begin{aligned} n = 0 : \quad t^0 &= 0, & y^0 &= y(0) = 1 \\ n = 1 : \quad t^1 &= 0.1, & y^1 &= y^0 + h f(t^0, y^0) = 1 + 0.1(-1 + 0 + 1) = 1 \\ n = 2 : \quad t^2 &= 0.2, & y^2 &= y^1 + h f(t^1, y^1) = 1 + 0.1(-1 + 0.1 + 1) = 1.01 \\ n = 3 : \quad t^3 &= 0.3, & y^3 &= y^2 + h f(t^2, y^2) = 1.01 + 0.1(-1.01 + 0.2 + 1) = 1.029 \\ & & & \vdots \end{aligned}$$

Δεν είναι δύσκολο να ελέγξουμε ότι η ακριβής λύση αυτού του προβλήματος είναι η $y(t) = t + e^{-t}$. Στον ακόλουθο πίνακα δίνονται οι προσεγγίσεις της μεθόδου του Euler, οι ακριβείς τιμές στους κόμβους, καθώς και η απόλυτη τιμή του σφάλματος.

t^n	y^n	$y(t^n)$	σφάλμα $ y^n - y(t^n) $
0.0	1.000000	1.000000	0.000000
0.1	1.000000	1.004837	0.004837
0.2	1.010000	1.018731	0.008731
0.3	1.029000	1.040818	0.011818
0.4	1.056100	1.070320	0.014220
0.5	1.090490	1.106531	0.016041
0.6	1.131441	1.148812	0.017371
0.7	1.178297	1.196585	0.018288
0.8	1.230467	1.249329	0.018862
0.9	1.287420	1.306570	0.019149
1.0	1.348678	1.367879	0.019201

□

6.2.2 Γεωμετρική ερμηνεία της μεθόδου του Euler

Ας δούμε τώρα τη γεωμετρική ερμηνεία της μεθόδου του Euler με τη βοήθεια ενός συγκεκριμένου παραδείγματος. Θεωρούμε το Π.Α.Τ.

$$\begin{cases} y'(t) = 2t, & 0 \leq t \leq 1, \\ y(0) = 0, \end{cases} \quad (6.9)$$

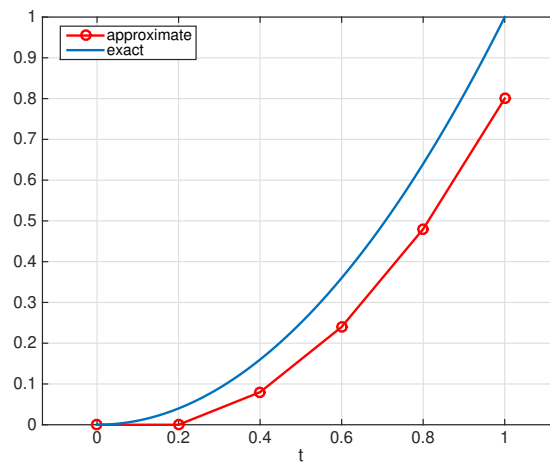
του οποίου η ακριβής λύση είναι προφανώς $y(t) = t^2$. Για το πρόβλημα αυτό οι προσεγγίσεις της μεθόδου του Euler υπολογίζονται από τον αναδρομικό τύπο:

$$y^{n+1} = y^n + h f(t^n, y^n) = y^n + 2 h t^n, \quad n = 0, 1, 2, \dots, N,$$

με $y^0 = 0$.

Θεωρούμε έναν ομοιόμορφο διαμερισμό του $[0, 1]$ σε $N = 5$ υποδιαστήματα, μήκους $h = (1 - 0)/5 = 0.2$ το καθένα. Άρα οι κόμβοι του διαμερισμού είναι $t^n = 0.2n$, $n = 0, 1, \dots, 5$. Στο Σχήμα 6.5 βλέπουμε την ακριβή λύση του Π.Α.Τ. (6.9) στο διάστημα $[0, 1]$ καθώς και τις προσεγγίσεις της μεθόδου του Euler στους κόμβους $t_1 = 0.2$, $t_2 = 0.4$, $t_3 = 0.6$, $t_4 = 0.8$, $t_5 = 1.0$. Παρατηρήστε ότι ξεκινάμε από το σημείο $(t^0, y^0) = (0, 0)$ και βρίσκουμε το σημείο τομής της $t = t^1 = 0.2$, με την ευθεία που διέρχεται από το σημείο $(0, 0)$ και έχει κλίση $f(t^0, y^0) = 0$. Το σημείο αυτό είναι το $(t^1, y^1) = (0.2, 0)$.

Στη συνέχεια, το σημείο (t^2, y^2) είναι το σημείο τομής της $t = t^2 = 0.4$, με την ευθεία που διέρχεται από το σημείο $(0.2, 0)$ και έχει κλίση $f(t^1, y^1) = 2(0.2) = 0.4$. Το σημείο αυτό είναι το $(t^1, y^1) = (0.4, 0.08)$.



Σχήμα 6.5: Η ακριβής και η προσεγγιστική λύση του (6.9) με τη μέθοδο του Euler.

Ανάλογα, το σημείο (t^3, y^3) είναι το σημείο τομής της $t = t^3 = 0.6$, με την ευθεία που διέρχεται από το σημείο $(0.4, 0.08)$ και έχει κλίση $f(t^2, y^2) = 2(0.4) = 0.8$. Το σημείο αυτό είναι το $(t^3, y^3) = (0.6, 0.24)$, κ.ό.κ.. □

6.2.3 Σφάλματα της μεθόδου του Euler

Ιδιαίτερη σημασία σε κάθε αριθμητική μέθοδο έχει η *ακρίβειά* της, δηλαδή η μέθοδος να δίνει μικρά σφάλματα

$$\varepsilon^n := y(t^n) - y^n.$$

Για τη μέθοδο του Euler μπορεί να αποδειχθεί (βλ., π.χ., [1, Θεώρ. 8.4]) το ακόλουθο θεώρημα, το οποίο μας δίνει μια εκτίμηση για το σφάλμα της μεθόδου.

Θεώρημα 6.2. Έστω $D = \{(t, y) : a \leq t \leq b, y \in \mathbb{R}\}$ και έστω $f(t, y)$ συνεχής στο D που ικανοποιεί τη συνθήκη του Lipschitz ως προς y ομοιόμορφα ως προς t στο D . Έστω y δύο φορές συνεχώς παραγωγίσιμη λύση του Π.Α.Τ.

$$\begin{cases} y'(t) = f(t, y(t)), & a \leq t \leq b, \\ y(a) = y_0. \end{cases}$$

Αν $y^0, y^1, \dots, y^N$ είναι οι προσεγγίσεις της μεθόδου του Euler για τον ομοιόμορφο διαμερισμό του $[a, b]$ με βήμα $h = (b - a)/N$, τότε ισχύει

$$\max_{0 \leq n \leq N} |y(t^n) - y^n| \leq \frac{M}{2L} (e^{L(b-a)} - 1) h, \quad (6.10)$$

όπου $M = \max_{a \leq t \leq b} |y''(t)|$ και L η σταθερά Lipschitz.

Το θεώρημα αυτό εξασφαλίζει ότι κάτω από συγκεκριμένες προϋποθέσεις για τις f και y έχουμε μια εκτίμηση για το σφάλμα της μεθόδου του Euler της μορφής

$$\max_{0 \leq n \leq N} |y(t^n) - y^n| \leq C h, \quad h = \frac{b - a}{N},$$

όπου η σταθερά C είναι ανεξάρτητη του βήματος h αλλά εξαρτάται από τα δεδομένα του προβλήματος a, b, y_0 και f . Επειδή ο εκθέτης του h είναι 1, λέμε ότι η μέθοδος του Euler έχει τάξη ακρίβειας τουλάχιστον 1.

Το φράγμα στην εκτίμηση (6.10) δεν μπορεί να βελτιωθεί όση ομαλότητα και αν απαιτήσουμε από τη λύση y , δηλαδή δεν μπορούμε να αντικαταστήσουμε τον όρο Ch από μια συνάρτηση που να συγκλίνει προς το μηδέν ταχύτερα από την Ch .

Για να δείξουμε ότι όσο ομαλή και να θεωρήσουμε τη λύση δεν μπορεί να βελτιωθεί (αυξηθεί) ο εκθέτης του h στο σφάλμα, θεωρούμε το πολύ απλό Π.Α.Τ.

$$\begin{cases} y'(t) = 2t, & 0 \leq t \leq 1, \\ y(0) = 0. \end{cases} \quad (6.11)$$

Το πρόβλημα (6.11) έχει μοναδική λύση: $y(t) = t^2$, $0 \leq t \leq 1$.

Έστω φυσικός N , $h = \frac{1}{N}$, $t^n = nh$, $n = 0, 1, \dots, N$ και $y^0, y^1, \dots, y^N$ οι προσεγγίσεις της μεθόδου του Euler,

$$y^{n+1} = y^n + h2t^n \quad \Rightarrow \quad y^{n+1} = y^n + 2nh^2, \quad n = 0, 1, \dots, N - 1.$$

Συγκεκριμένα, έχουμε:

$$\begin{aligned}
 n = 0 : \quad & y^1 = y^0 \\
 n = 1 : \quad & y^2 = y^1 + 2h^2 = y^0 + 2h^2 \\
 n = 2 : \quad & y^3 = y^2 + 2 \cdot 2h^2 = y^0 + 2h^2 + 2 \cdot 2h^2 = y^0 + 2(1 + 2)h^2 \\
 & \vdots \\
 n : \quad & y^n = y^0 + 2(1 + 2 + \cdots + (n-1))h^2 = y^0 + 2 \frac{n(n-1)}{2} h^2
 \end{aligned}$$

Όμως

$$y^0 = 0 \quad \Rightarrow \quad y^n = n(n-1)h^2, \quad n = 0, 1, \dots, N.$$

Άρα για $n = N$ έχουμε:

$$y^N = N(N-1)h^2 = \underbrace{Nh}_{=1}(\underbrace{Nh}_{=1} - h) = 1 - h.$$

Επομένως, το σφάλμα στο σημείο $t = t^N = 1$, είναι ακριβώς ίσο με:

$$y(1) - y^N = 1 - (1 - h) = h.$$

□

Άρα κάτω από τις υποθέσεις του Θεωρήματος 6.2, η τάξη ακρίβειας της μεθόδου του Euler είναι ακριβώς ένα.

Παρατήρηση 6.2. Αν η λύση του Π.Α.Τ. δεν είναι δύο φορές συνεχώς παραγωγίσιμη αλλά, για παράδειγμα, μόνο μία φορά, τότε αποδεικνύεται ότι η μέθοδος του Euler εξακολουθεί να συγκλίνει αλλά η αντίστοιχη της εκτίμησης (6.10) δεν περιέχει το h στην πρώτη δύναμη, αλλά εμφανίζεται μια συνάρτηση του h η οποία απλώς τείνει στο 0 καθώς $h \rightarrow 0$.

Ορισμός 6.2. Το σφάλμα $\varepsilon^n := y(t^n) - y^n$, $n = 0, 1, 2, \dots, N$ λέγεται σφάλμα διακριτοποίησης ή σφάλμα προσέγγισης της μεθόδου. Συνήθως λέγεται «ολικό σφάλμα».

Τοπικό σφάλμα της μεθόδου του Euler στο σημείο t^n , $n = 0, 1, \dots, N-1$, λέγεται η ποσότητα

$$\delta^n := [y(t^n) + hf(t^n, y(t^n))] - y(t^{n+1}).$$

Δηλαδή, το τοπικό σφάλμα είναι το σφάλμα που κάνουμε αν εκτελέσουμε ένα βήμα της μεθόδου του Euler ξεκινώντας από το σημείο $(t^n, y(t^n))$ (και όχι από το (t^n, y^n) όπως γίνεται στην πράξη).

Στη συνέχεια θα εξετάσουμε τη συμπεριφορά του τοπικού σφάλματος (πάντα υπό τις προϋποθέσεις του Θεωρήματος 6.2).

$$\delta^n := [y(t^n) + hf(t^n, y(t^n))] - y(t^{n+1}) = y(t^n) + hf(t^n, y(t^n)) - y(t^n + h).$$

Αναπτύσσουμε με Taylor τον όρο $y(t^n + h)$ γύρω από το σημείο t^n . Άρα

$$\begin{aligned} \delta^n &= y(t^n) + hf(t^n, y(t^n)) - \left[y(t^n) + h \underbrace{y'(t^n)}_{f(t^n, y(t^n))} + \frac{h^2}{2} y''(\xi_n) \right] \\ &= y(t^n) + hf(t^n, y(t^n)) - y(t^n) - hf(t^n, y(t^n)) - \frac{h^2}{2} y''(\xi_n) \\ &= -\frac{h^2}{2} y''(\xi_n). \end{aligned}$$

Άρα η τάξη του τοπικού σφάλματος της μεθόδου του Euler (δηλαδή η δύναμη του h με την οποία το τοπικό σφάλμα τείνει στο μηδέν όταν το $h \rightarrow 0$) είναι 2.

Παρατήρηση 6.3. Το φαινόμενο αυτό, δηλαδή ότι η τάξη του τοπικού σφάλματος είναι κατά ένα μεγαλύτερη από την τάξη του ολικού σφάλματος, είναι γενικό για όλες τις αριθμητικές μεθόδους για την προσέγγιση λύσεων Π.Α.Τ. για Δ.Ε. πρώτης τάξης.

Διαισθητικά (και πολύ χονδρικά) μπορεί κανείς να πει ότι επειδή κάνουμε $N = \frac{b-a}{h}$ βήματα για να φτάσουμε στο b , τα τοπικά σφάλματα συσσωρεύονται και έτσι η τάξη του σφάλματος διακριτοποίησης αναμένουμε να είναι κατά ένα μικρότερη από την τάξη του τοπικού σφάλματος.

6.2.4 Ευστάθεια της μεθόδου του Euler

Θεωρούμε το Π.Α.Τ. (6.4) και υποθέτουμε ότι η συνάρτηση f πληροί τις υποθέσεις του Θεωρήματος 6.1. Τότε για δεδομένες αρχικές τιμές $y_0, z_0 \in \mathbb{R}$, τα Π.Α.Τ.

$$\begin{cases} y'(t) = f(t, y), & a \leq t \leq b, \\ y(a) = y_0, \end{cases} \quad \text{και} \quad \begin{cases} z'(t) = f(t, z), & a \leq t \leq b, \\ z(a) = z_0, \end{cases}$$

έχουν μοναδικές λύσεις y, z που είναι συνεχώς παραγωγίσιμες συναρτήσεις.

Τότε, δεν είναι ιδιαίτερα δύσκολο να αποδείξει κανείς την ακόλουθη ανισότητα:

$$\max_{a \leq t \leq b} |y(t) - z(t)| \leq e^{L(b-a)} |y_0 - z_0|. \quad (6.12)$$

Η ανισότητα (6.12) εκφράζει, όπως λέμε, τη συνεχή εξάρτηση των λύσεων του Π.Α.Τ. (6.4) από τα αρχικά δεδομένα.

Μπορεί κανείς να αποδείξει ότι και για τις προσεγγίσεις της μεθόδου του Euler ισχύει κάτι ανάλογο. Συγκεκριμένα, αν θεωρήσουμε ως συνήθως έναν ομοιόμορφο διαμερισμό του $[a, b]$ σε N υποδιαστήματα μήκους h , και $y^n, z^n, 0 \leq n \leq N$, είναι οι λύσεις των προβλημάτων

$$\begin{cases} y^{n+1} = y^n + hf(t^n, y^n), & 0 \leq n \leq N-1, \\ y^0 = y_0, \end{cases}$$

και

$$\begin{cases} z^{n+1} = z^n + hf(t^n, z^n), & 0 \leq n \leq N-1, \\ z^0 = z_0, \end{cases}$$

αποδεικνύεται ότι

$$\max_{0 \leq n \leq N} |y^n - z^n| \leq e^{L(b-a)} |y_0 - z_0|. \quad (6.13)$$

Η ανισότητα (6.13) είναι το διακριτό ανάλογο της (6.12) και εκφράζει την *ευστάθεια της αριθμητικής λύσης* σε διαταραχές των αρχικών δεδομένων. Παρόλα αυτά, αυτή η έννοια της «ευστάθειας» της μεθόδου του Euler δεν έχει μεγάλη πρακτική χρησιμότητα, όπως θα δούμε μέσω του επόμενου παραδείγματος.

Θεωρούμε το απλό γραμμικό πρόβλημα αρχικών τιμών:

$$\begin{cases} y' = \lambda y, & t \geq 0, \quad \text{όπου } \lambda < 0, \\ y(0) = 1. \end{cases} \quad (6.14)$$

Η ακριβής λύση του (6.14) είναι $y(t) = e^{\lambda t}$ και, επειδή $\lambda < 0$, προφανώς φθίνει εκθετικά προς το μηδέν καθώς το $t \rightarrow \infty$. Θεωρούμε τους κόμβους $t^n = nh, n = 0, 1, \dots$. Τότε η ακολουθία $(y^n)_{n \geq 0}$ των προσεγγίσεων της μεθόδου του Euler δίνεται από την αναδρομική σχέση

$$\begin{cases} y^{n+1} = y^n + h\lambda y^n = (1 + h\lambda) y^n, & n = 0, 1, 2, \dots, \\ y^0 = 1. \end{cases}$$

Προφανώς

$$\begin{aligned} y^1 &= (1 + h\lambda) y^0 = (1 + h\lambda), \\ y^2 &= (1 + h\lambda) y^1 = (1 + h\lambda)^2, \\ &\vdots \\ y^n &= (1 + h\lambda) y^{n-1} = (1 + h\lambda)^n. \end{aligned}$$

Άρα η ακολουθία (y^n) έχει τύπο

$$y^n = (1 + h\lambda)^n, \quad n = 0, 1, 2, \dots \quad (6.15)$$

Αν σταθεροποιήσουμε ένα $t > 0$, τότε η μέθοδος του Euler δίνει προσεγγίσεις που συγκλίνουν στη λύση $e^{\lambda t}$ του συνεχούς προβλήματος καθώς παίρνουμε ολοένα και μικρότερο «βήμα» h , δηλαδή καθώς $h \rightarrow 0$. Για να το δούμε αυτό ας θεωρήσουμε $nh = t$, έτσι ώστε $n \rightarrow \infty$ καθώς $h \rightarrow 0$. Τότε έχουμε

$$y^n = (1 + h\lambda)^n = \left(1 + \frac{\lambda t}{n}\right)^n \rightarrow e^{\lambda t}, \quad \text{καθώς } n \rightarrow \infty.$$

Στην πράξη όμως οι υπολογισμοί γίνονται με ένα μικρό μεν αλλά θετικό βήμα h . Από την (6.15) έχουμε ότι για $n \geq 0$,

$$|y^n| = |1 + h\lambda|^n.$$

Άρα για σταθερό h , καθώς $n \rightarrow \infty$, δηλαδή καθώς $t^n = nh \rightarrow \infty$, συμπεραίνουμε ότι:

$$\begin{aligned} |y^n| &\rightarrow 0, & \text{όταν } |1 + h\lambda| < 1, \\ |y^n| &= 1, & \text{όταν } |1 + h\lambda| = 1, \\ |y^n| &\rightarrow \infty, & \text{όταν } |1 + h\lambda| > 1. \end{aligned}$$

Άρα, οι προσεγγίσεις y^n που δίνει η μέθοδος του Euler για το πρόβλημα δοκιμής (6.14), για ένα σταθερό $h > 0$, μιμούνται τη συμπεριφορά της ακριβούς λύσης $e^{\lambda t}$ (δηλαδή $y^n \rightarrow 0$, καθώς $n \rightarrow \infty$) μόνον όταν $|1 + h\lambda| < 1$, δηλαδή όταν $h\lambda \in (-2, 0)$. Επίσης, οι προσεγγίσεις ικανοποιούν τη συνθήκη $|y^n| \leq 1$, για κάθε $n \geq 0$, αν και μόνον αν

$$|1 + h\lambda| \leq 1 \quad \Leftrightarrow \quad h\lambda \in [-2, 0].$$

Γενικώς, λέμε ότι μια μέθοδος για την αριθμητική επίλυση του Π.Α.Τ. (6.4) είναι *απόλυτα ευσταθής* για κάποιο $h > 0$, αν, όταν εφαρμοσθεί στο πρόβλημα δοκιμής (6.14), παράγει προσεγγίσεις y^n που παραμένουν φραγμένες καθώς $n \rightarrow \infty$. Το διάστημα I της μορφής $[a, 0]$, με $-\infty \leq a \leq 0$, τέτοιο ώστε η μέθοδος να είναι απόλυτα ευσταθής για $h\lambda \in I$ λέγεται *διάστημα απόλυτης ευστάθειας* της μεθόδου.

Άρα η μέθοδος του Euler έχει διάστημα απόλυτης ευστάθειας το $I = [-2, 0]$ και επομένως είναι απολύτως ευσταθής για $0 \leq h \leq -\frac{2}{\lambda}$. Αυτό σημαίνει ότι αν το $|\lambda|$ είναι μεγάλο, τότε το μέγιστο βήμα h που μπορούμε να πάρουμε πρέπει να είναι αρκετά μικρό ώστε να πάρουμε αποδεκτές προσεγγίσεις για το πρόβλημα (6.14).

6.3 Άλλες απλές μέθοδοι

6.3.1 Η πεπλεγμένη μέθοδος του Euler

Θεωρούμε και πάλι το Π.Α.Τ. (6.4) και έναν ομοιόμορφο διαμερισμό $a = t^0 < t^1 < \dots < t^{N-1} < t^N = b$ του διαστήματος $[a, b]$. Δηλαδή, ορίζουμε ως «βήμα» $h = (b - a)/N$, οπότε

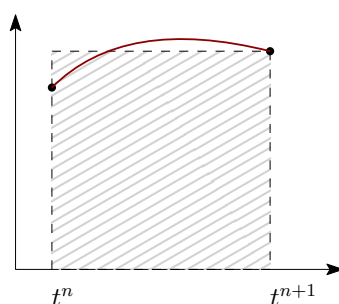
οι κόμβοι $t^n = a + nh$, $n = 0, 1, \dots, N$. Όπως είδαμε στην Παράγραφο 6.2.1 ένας τρόπος κατασκευής της μεθόδου του Euler προκύπτει ολοκληρώνοντας και τα δύο μέλη της διαφορικής εξίσωσης στο διάστημα $[t^n, t^{n+1}]$. Έτσι καταλήγουμε στη σχέση (6.8)

$$y(t^{n+1}) - y(t^n) = \int_{t^n}^{t^{n+1}} f(t, y(t)) dt.$$

Είδαμε λοιπόν ότι αν προσεγγίσουμε το ολοκλήρωμα που εμφανίζεται στο δεξί μέλος με τον αριστερό κανόνα του ορθογωνίου, τότε καταλήγουμε στη μέθοδο του Euler.

Αν τώρα προσεγγίσουμε το ολοκλήρωμα αυτό με τον δεξιό τύπο του ορθογωνίου, βλ. Σχήμα 6.4,

$$\int_{t^n}^{t^{n+1}} f(t, y(t)) dt \approx hf(t^{n+1}, y(t^{n+1}))$$



Σχήμα 6.6: Προσέγγιση του ολοκληρώματος $\int_{t^n}^{t^{n+1}} f(t, y(t)) dt$ με τον δεξιό τύπο του ορθογωνίου.

παίρνουμε τη σχέση

$$y(t^{n+1}) - y(t^n) \approx hf(t^{n+1}, y(t^{n+1})),$$

Επομένως, αν συμβολίσουμε, κατά τα γνωστά, με y^n την προσέγγιση της ακριβούς λύσης $y(t^n)$ στον κόμβο t^n , καταλήγουμε στη λεγόμενη *πεπλεγμένη μέθοδο του Euler*:

Πεπλεγμένη μέθοδος του Euler

$$\begin{aligned} y^{n+1} &= y^n + hf(t^{n+1}, y^{n+1}), \quad n = 0, 1, 2, \dots, N-1, \\ y^0 &= y_0. \end{aligned} \quad (6.16)$$

Παρατηρήστε ότι το y^{n+1} δίνεται στην (6.16) πεπλεγμένα, δηλαδή για να το προσδιορίσουμε πρέπει σε κάθε βήμα να λύνουμε μια μη γραμμική εξίσωση. Αυτό μπορεί φυσικά να γίνει με

χρήση παραδοσιακών μεθόδων όπως η μέθοδος του Νεύτωνα, της τέμνουσας, κ.λ.π., αλλά, εν γένει, αυτή είναι μια δαπανηρή διαδικασία. Συνήθως για τον υπολογισμό του y^{n+1} σε κάθε βήμα χρησιμοποιούμε μια απλή επαναληπτική μέθοδο. Συγκεκριμένα, αν θέσουμε

$$\varphi(x) := y^n + h f(t^{n+1}, x),$$

ο υπολογισμός της y^{n+1} ανάγεται στην εύρεση σταθερού σημείου της φ . Επομένως, δοθείσης μιας αρχικής προσέγγισης y_0^{n+1} της ζητούμενης τιμής y^{n+1} , κατασκευάζουμε μια ακολουθία $y_0^{n+1}, y_1^{n+1}, \dots$, μέσω της αναδρομικής σχέσης

$$y_{j+1}^{n+1} = y^n + h f(t^{n+1}, y_j^{n+1}), \quad j = 0, 1, 2, \dots \quad (6.17)$$

Αποδεικνύεται ότι αν το h είναι αρκούντως μικρό, τότε οι προσεγγίσεις $y_j^{n+1} \rightarrow y^{n+1}$, καθώς $j \rightarrow \infty$. Για να πάρουμε μια ιδέα γιατί το αποτέλεσμα αυτό είναι αληθές, αφαιρούμε κατά μέλη τις (6.16) και (6.17), οπότε

$$\begin{aligned} y^{n+1} - y_{j+1}^{n+1} &= h \left(f(t^{n+1}, y^{n+1}) - f(t^{n+1}, y_j^{n+1}) \right), \\ \Rightarrow y^{n+1} - y_{j+1}^{n+1} &\approx h \frac{\partial f(t^{n+1}, y^{n+1})}{\partial y} (y^{n+1} - y_j^{n+1}). \end{aligned}$$

Η τελευταία σχέση προκύπτει από το Θεώρημα Μέσης Τιμής για την $f(t^{n+1}, x)$, ως συνάρτηση του x . Άρα, αν η αρχική εκτίμηση y_0^{n+1} είναι μια καλή προσέγγιση της y^{n+1} , τότε τα σφάλματα $y^{n+1} - y_{j+1}^{n+1}$ θα συγκλίνουν στο μηδέν, αν ισχύει ότι

$$\left| h \frac{\partial f(t^{n+1}, y^{n+1})}{\partial y} \right| < 1. \quad (6.18)$$

Στην πράξη συχνά επιλέγουμε την αρχική εκτίμηση y_0^{n+1} με τη μέθοδο του Euler

$$y_0^{n+1} = y^n + h f(t^n, y^n),$$

και το βήμα h ώστε η ποσότητα που εμφανίζεται στην (6.18) να είναι αρκετά μικρότερη της μονάδας. Σε αυτή την περίπτωση, συνήθως, απαιτούνται ελάχιστες επαναλήψεις για τον υπολογισμό του y^{n+1} .

Ευστάθεια της πεπλεγμένης μεθόδου του Euler. Θεωρούμε και πάλι το πρόβλημα δοκιμής (6.14)

$$\begin{cases} y' = \lambda y, & t \geq 0, & \text{όπου } \lambda < 0, \\ y(0) = 1, \end{cases}$$

του οποίου η ακριβής λύση είναι $y(t) = e^{\lambda t}$. Θεωρούμε τους κόμβους $t^n = nh$, $n = 0, 1, \dots$, οπότε η ακολουθία $(y^n)_{n \geq 0}$ των προσεγγίσεων της πεπλεγμένης μεθόδου του Euler δίνεται από την αναδρομική σχέση

$$y^{n+1} = y^n + h\lambda y^{n+1} \Rightarrow (1 - h\lambda) y^{n+1} = y^n \Rightarrow y^{n+1} = \frac{1}{1 - h\lambda} y^n,$$

για $n = 0, 1, \dots$, με $y^0 = 1$. Προφανώς

$$\begin{aligned} y^1 &= \frac{1}{1 - h\lambda} y^0 = \frac{1}{1 - h\lambda}, \\ y^2 &= \frac{1}{1 - h\lambda} y^1 = \left(\frac{1}{1 - h\lambda} \right)^2, \\ &\vdots \\ y^n &= \frac{1}{1 - h\lambda} y^{n-1} = \left(\frac{1}{1 - h\lambda} \right)^n. \end{aligned}$$

Άρα η ακολουθία (y^n) έχει τύπο $y^n = \left(\frac{1}{1 - h\lambda} \right)^n$ και θα είναι φραγμένη αν και μόνον αν

$$\left| \frac{1}{1 - h\lambda} \right| \leq 1 \quad \Leftrightarrow \quad |1 - h\lambda| \geq 1.$$

Η τελευταία όμως αληθεύει για κάθε «βήμα» h , άρα η πεπλεγμένη μέθοδος του Euler είναι απόλυτα ευσταθής για κάθε $h \geq 0$.

Αποδεικνύεται ότι κάτω από κάποιες προϋποθέσεις για την f και την y , και η πεπλεγμένη Euler έχει τάξη ακρίβειας 1, δηλαδή υπάρχει μια σταθερά C που είναι ανεξάρτητη του βήματος h τέτοια ώστε

$$\max_{0 \leq n \leq N} |y(t^n) - y^n| \leq C h.$$

6.3.2 Η μέθοδος του τραπεζίου

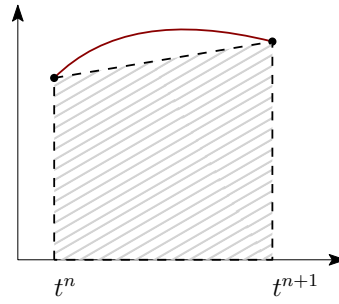
Έστω τώρα ότι προσεγγίζουμε το ολοκλήρωμα που εμφανίζεται στο δεξί μέλος της σχέσης (6.8) με τον κανόνα του τραπεζίου, βλ. Σχήμα 6.7, έχουμε

$$\int_{t^n}^{t^{n+1}} f(t, y(t)) dt \approx \frac{h}{2} [f(t^n, y(t^n)) + f(t^{n+1}, y(t^{n+1}))],$$

οπότε καταλήγουμε στη σχέση

$$y(t^{n+1}) - y(t^n) \approx \frac{h}{2} [f(t^n, y(t^n)) + f(t^{n+1}, y(t^{n+1}))].$$

Έτσι ορίζουμε τη λεγόμενη μέθοδο του τραπεζίου,



Σχήμα 6.7: Προσέγγιση του ολοκληρώματος $\int_{t^n}^{t^{n+1}} f(t, y(t)) dt$ με τον τύπο του τραπεζίου.

Μέθοδος του τραπεζίου

$$y^{n+1} = y^n + \frac{h}{2} [f(t^n, y^n) + f(t^{n+1}, y^{n+1})], \quad n = 0, 1, 2, \dots, N-1, \quad (6.19)$$

$$y^0 = y_0.$$

Αποδεικνύεται ότι κάτω από συγκεκριμένες προϋποθέσεις για την f και την y το τοπικό σφάλμα της μεθόδου είναι $\delta^n = -\frac{1}{12} h^3 y'''(\xi_n)$, και η τάξη ακρίβειας της μεθόδου είναι 2, δηλαδή υπάρχει μια σταθερά C , ανεξάρτητη του βήματος h , τέτοια ώστε

$$\max_{0 \leq n \leq N} |y(t^n) - y^n| \leq C h^2.$$

Επιπλέον, η μέθοδος του τραπεζίου είναι απόλυτα ευσταθής για κάθε $h > 0$.

Σχετικά με την επίλυση της μη γραμμικής εξίσωσης που απαιτείται σε κάθε βήμα, όπως και στην πεπλεγμένη μέθοδο του Euler η καινούργια προσέγγιση y^{n+1} υπολογίζεται μέσω της επαναληπτικής μεθόδου

$$y_{j+1}^{n+1} = y^n + \frac{h}{2} [f(t^n, y^n) + f(t^{n+1}, y_j^{n+1})], \quad j = 0, 1, 2, \dots,$$

όπου ως αρχική προσέγγιση χρησιμοποιούμε είτε αυτήν που μας δίνει η μέθοδος του Euler

$$y_0^{n+1} = y^n + h f(t^n, y^n),$$

ή την

$$y_0^{n+1} = y^n + \frac{h}{2} [3 f(t^n, y^n) - f(t^{n-1}, y^{n-1})].$$

Τέλος, αν στη μέθοδο (6.19) αντικαταστήσουμε την τιμή y^{n+1} στο β' μέλος από την εκτίμηση που μας δίνει η μέθοδος του Euler, προκύπτει η λεγόμενη μέθοδος του Heun:

$$y^{n+1} = y^n + \frac{h}{2} [f(t^n, y^n) + f(t^{n+1}, y^n + h f(t^n, y^n))], \quad n = 0, 1, \dots, N-1, \quad (6.20)$$

$$y^0 = y_0.$$

Η μέθοδος αυτή είναι άμεση, είναι επίσης δεύτερης τάξης, αλλά δεν είναι πλέον απολύτως ευσταθής για οποιαδήποτε τιμή του h .

6.4 Γενικεύσεις της μεθόδου του Euler.

Μέθοδοι Runge-Kutta

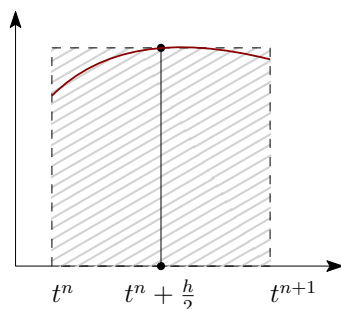
Όπως είδαμε νωρίτερα, μπορούμε να επαναδιατυπώσουμε τη Σ.Δ.Ε. του προβλήματος (6.4) σε κάθε υποδιάστημα ενός διαμερισμού του διαστήματος $[a, b]$ με τη βοήθεια της (ολοκληρωτικής) εξίσωσης (6.8). Έτσι το πρόβλημα της κατασκευής μιας αριθμητικής μεθόδου μπορεί να τεθεί ως ένα πρόβλημα προσέγγισης του ακριβούς προβλήματος

$$y(t^{n+1}) - y(t^n) = \int_{t^n}^{t^{n+1}} f(t, y(t)) dt.$$

Μέχρι τώρα είδαμε διάφορες μεθόδους που προκύπτουν ανάλογα με τον τύπο αριθμητικής ολοκλήρωσης που χρησιμοποιούμε για την προσέγγιση του ολοκληρώματος στο δεξί μέλος της προηγούμενης σχέσης.

Για παράδειγμα, αν θεωρήσουμε έναν ομοιόμορφο διαμερισμό και προσεγγίσουμε το ολοκλήρωμα με τον κανόνα του μέσου, βλ. Σχήμα 6.8, καταλήγουμε στη σχέση

$$y(t^{n+1}) - y(t^n) \approx h f\left(t^n + \frac{h}{2}, y\left(t^n + \frac{h}{2}\right)\right).$$



Σχήμα 6.8: Προσέγγιση του ολοκληρώματος $\int_{t^n}^{t^{n+1}} f(t, y(t)) dt$ με τον τύπο του μέσου.

Αν προσεγγίσουμε την τιμή $y(t^n + \frac{h}{2})$ με τη μέθοδο του Euler, δηλαδή $y(t^n + \frac{h}{2}) \approx y(t^n) + \frac{h}{2} f(t^n, y^n)$, καταλήγουμε στην βελτιωμένη μέθοδο του Euler ή άμεση μέθοδο του μέσου:

$$y^{n+1} = y^n + h f\left(t^n + \frac{h}{2}, y^n + \frac{h}{2} f(t^n, y^n)\right), \quad n = 0, 1, \dots, N-1, \quad (6.21)$$

$$y^0 = y_0.$$

Όλες αυτές οι μέθοδοι εντάσσονται στη γενικότερη κατηγορία των μεθόδων Runge-Kutta.

Μέθοδοι Runge-Kutta

Γενικά, μια μέθοδος Runge-Kutta (RK) με q ενδιάμεσα στάδια υπολογίζει την προσέγγιση y^{n+1} από την y^n ($y^0 = y_0$) μέσω του τύπου:

$$y^{n+1} = y^n + h \sum_{i=1}^q b_i f(t^{n,i}, y^{n,i}), \quad 0 \leq n \leq N-1, \quad (6.22)$$

όπου b_i , $1 \leq i \leq q$, είναι δεδομένες πραγματικές σταθερές.

Οι ενδιάμεσες τιμές $y^{n,i}$, $1 \leq i \leq q$, και τα αντίστοιχα σημεία $t^{n,i}$, $1 \leq i \leq q$, δίνονται από τις σχέσεις:

$$t^{n,i} = t^n + \tau_i h, \quad 1 \leq i \leq q,$$

$$y^{n,i} = y^n + h \sum_{j=1}^q a_{ij} f(t^{n,j}, y^{n,j}), \quad 1 \leq i \leq q. \quad (6.23)$$

Παρατηρήσεις:

- Αν και η σχέση (6.22) δείχνει αρκετά πολύπλοκη, η ιδέα πίσω από αυτήν είναι αρκετά καθαρή. Θυμηθείτε ότι ένας κανόνα αριθμητικής ολοκλήρωσης έχει τη γενική μορφή

$$\int_a^b f(x) dx \approx \sum_{i=1}^q w_i f(x_i),$$

όπου w_i είναι τα βάρη και x_i είναι οι κόμβοι ολοκλήρωσης. Άρα στον τύπο (6.22) μπορούμε να ερμηνεύσουμε τα hb_i ως βάρη, ενώ τα $f(t^{n,i}, y^{n,i})$ είναι προσεγγίσεις της $f(t, y(t))$ σε q κόμβους $t^{n,i}$ στο διάστημα $[t^n, t^{n+1}]$.

- Συνήθως οι παράμετροι μιας μεθόδου RK γράφονται σε μορφή μητρώου Butcher,

$$\begin{array}{ccc|c} a_{11} & \dots & a_{1q} & \tau_1 \\ \vdots & & \vdots & \vdots \\ a_{q1} & \dots & a_{qq} & \tau_q \\ \hline b_1 & \dots & b_q & \end{array}$$

- Αν ο πίνακας a_{ij} είναι αυστηρά κάτω τριγωνικός (δηλαδή τα διαγώνια στοιχεία και όλα τα στοιχεία που βρίσκονται στο «πάνω τρίγωνο» είναι μηδέν), τότε η μέθοδος είναι άμεση, δηλαδή η λύση του συστήματος για τα ενδιαμέσως στάδια $y^{n,i}$ απαιτεί μόνο αντικατάσταση. Αν είναι κάτω τριγωνικός, η μέθοδος είναι ημιπεπλεγμένη, δηλαδή η λύση του συστήματος για τα ενδιαμέσως στάδια απαιτεί τη διαδοχική λύση q (ενδεχομένως μη γραμμικών) εξισώσεων. Στη γενική περίπτωση έχουμε πεπλεγμένες μεθόδους, δηλ. απαιτείται η επίλυση q (ενδεχομένως μη γραμμικών) εξισώσεων.

6.4.1 Παραδείγματα μεθόδων Runge-Kutta

Όλες οι μέθοδοι που έχουμε δει μέχρι τώρα μπορούν να ενταχθούν στο πλαίσιο των RK. Έτσι έχουμε:

Άμεση μέθοδος του Euler (6.7) Είναι άμεση μέθοδος με $q = 1$ στάδιο και τάξη ακρίβειας $p = 1$. Το μητρώο της είναι:

$$\begin{array}{c|c} 0 & 0 \\ \hline 1 & \end{array}$$

Πεπλεγμένη μέθοδος του Euler (6.16): Είναι πεπλεγμένη μέθοδος με $q = 1$ στάδιο και τάξη ακρίβειας $p = 1$. Το μητρώο της είναι:

$$\begin{array}{c|c} 1 & 1 \\ \hline 1 & \end{array}$$

Μέθοδος του τραπεζίου (6.19): Είναι ημιπεπλεγμένη μέθοδος με $q = 2$ στάδια και τάξη ακρίβειας $p = 2$. Το μητρώο της είναι:

$$\begin{array}{cc|c} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 \\ \hline \frac{1}{2} & \frac{1}{2} & \end{array}$$

Βελτιωμένη μέθοδος Euler ή άμεση μέθοδος του μέσου (6.21): Είναι άμεση μέθοδος με $q = 2$ στάδια και τάξη ακρίβειας $p = 2$. Το μητρώο της είναι:

$$\begin{array}{cc|c} 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \hline 0 & 1 & \end{array}$$

6.4.2 Η κλασσική μέθοδος των Runge-Kutta

Μια από τις δημοφιλέστερες μεθόδους RK είναι η λεγόμενη κλασσική μέθοδος των Runge-Kutta με μητρώο:

$$\begin{array}{cccc|c} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 & 1 \\ \hline \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & \end{array}$$

Η μέθοδος αυτή είναι άμεση μέθοδος με $q = 4$ στάδια και τάξη ακρίβειας $p = 4$.

Ας δούμε αναλυτικότερα τους τύπους για τη μέθοδο αυτή. Σύμφωνα με το παραπάνω μητρώο, από τους τύπους (6.22) και (6.23) έχουμε:

$$\begin{aligned} t^{n,1} &= t^n, \\ t^{n,2} &= t^n + \frac{h}{2}, \\ t^{n,3} &= t^n + \frac{h}{2}, \\ t^{n,4} &= t^n + h = t^{n+1}. \end{aligned}$$

Επομένως,

$$\begin{aligned} y^{n,1} &= y^n, \\ y^{n,2} &= y^n + \frac{h}{2} f(t^{n,1}, y^{n,1}) = y^n + \frac{h}{2} f(t^n, y^n), \\ y^{n,3} &= y^n + \frac{h}{2} f(t^{n,2}, y^{n,2}) = y^n + \frac{h}{2} f(t^n + \frac{h}{2}, y^{n,2}), \\ y^{n,4} &= y^n + h f(t^{n,3}, y^{n,3}) = y^n + h f(t^n + \frac{h}{2}, y^{n,3}). \end{aligned}$$

Άρα η επόμενη προσέγγιση υπολογίζεται από τον τύπο

$$\begin{aligned} y^{n+1} &= y^n + \frac{h}{6} \left(f(t^{n,1}, y^{n,1}) + 2f(t^{n,2}, y^{n,2}) + 2f(t^{n,3}, y^{n,3}) + f(t^{n,4}, y^{n,4}) \right) \\ \Leftrightarrow y^{n+1} &= y^n + \frac{h}{6} \left(f(t^n, y^n) + 2f(t^n + \frac{h}{2}, y^{n,2}) + 2f(t^n + \frac{h}{2}, y^{n,3}) + f(t^{n+1}, y^{n,4}) \right). \end{aligned}$$

Η μέθοδος αυτή γράφεται συνήθως συντομότερα ως εξής: Θέτουμε $k_1 = f(t^n, y^n)$, $k_2 = f(t^n + \frac{h}{2}, y^{n,2})$, $k_3 = f(t^n + \frac{h}{2}, y^{n,3})$, και $k_4 = f(t^{n+1}, y^{n,4})$. Τότε δεν είναι δύσκολο να δούμε ότι η μέθοδος αυτή ορίζεται ισοδύναμα από τις σχέσεις:

Η κλασσική μέθοδος των Runge-Kutta (RK4)

$$\begin{aligned}
 k_1 &= f(t^n, y^n) \\
 k_2 &= f\left(t^n + \frac{h}{2}, y^n + \frac{h}{2} k_1\right) \\
 k_3 &= f\left(t^n + \frac{h}{2}, y^n + \frac{h}{2} k_2\right) \\
 k_4 &= f(t^n + h, y^n + h k_3) \\
 y^{n+1} &= y^n + \frac{h}{6} (k_1 + 2k_2 + 2k_3 + k_4).
 \end{aligned} \tag{6.24}$$

Στον ακόλουθο ψευδοκώδικα περιγράφεται ο αλγόριθμος για την υλοποίηση της μεθόδου στον υπολογιστή. Όπως βλέπουμε, η RK4 απαιτεί 4 υπολογισμούς της συνάρτησης f ανά βήμα.

Αλγόριθμος 6.2

Είσοδος: a, b, y_0, N

Εξωτερική συνάρτηση f

$h \leftarrow (b - a)/N$

$t \leftarrow a$

$y \leftarrow y_0$

for $k = 1, \dots, N$ **do**

$k_1 \leftarrow f(t, y)$

$k_2 \leftarrow f\left(t + \frac{h}{2}, y + \frac{h}{2} k_1\right)$

$k_3 \leftarrow f\left(t + \frac{h}{2}, y + \frac{h}{2} k_2\right)$

$k_4 \leftarrow f(t + h, y + h k_3)$

$y \leftarrow y + \frac{h}{6} (k_1 + 2k_2 + 2k_3 + k_4)$

$t \leftarrow t + h$

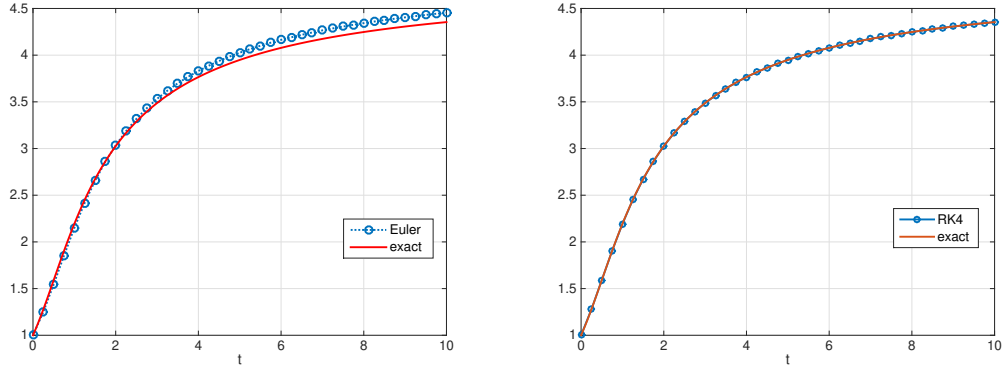
end for

Παράδειγμα. Θεωρήστε το Π.Α.Τ.:

$$\begin{cases} y'(t) = \frac{y}{1+t^2}, & 0 \leq t \leq 10, \\ y(0) = 1. \end{cases} \tag{6.25}$$

Η ακριβής λύση του προβλήματος αυτού είναι $y(t) = e^{\arctan t}$. Θεωρούμε έναν ομοιόμορφο διαμερισμό του $[0, 10]$ σε N υποδιαστήματα, άρα το βήμα είναι $h = 10/N$. Λύνουμε το πρόβλημα αυτό με την άμεση μέθοδο του Euler και με την κλασσική μέθοδο των Runge-Kutta

(RK4). Στο Σχήμα 6.9 σχεδιάζουμε την ακριβή λύση με την κόκκινη συνεχή γραμμή, και τις προσεγγίσεις της με τη μέθοδο του Euler (αριστερά) και με την RK4 (δεξιά), για $N = 40$.



Σχήμα 6.9: Η μέθοδος του Euler (αριστερά) και η κλασσική Runge-Kutta (δεξιά) για το πρόβλημα (6.25) με $N = 40$.

Στον ακόλουθο πίνακα σημειώνουμε τις τιμές του σφάλματος $\max_{0 \leq n \leq N} |y(t^n) - y^n|$ για τις δύο μεθόδους για διάφορες τιμές του N . Παρατηρούμε ότι όταν υποδιαπλασιάζουμε το βήμα (δηλαδή διπλασιάζουμε το N), τότε το σφάλμα της μεθόδου του Euler περίπου υποδιπλασιάζεται, ενώ το σφάλμα της RK4 δείχνει να υποδεκαεξαπλασιάζεται. Αυτό είναι απολύτως αναμενόμενο, αφού όπως έχουμε ήδη αναφέρει, η πρώτη είναι μια μέθοδος τάξης 1, ενώ η δεύτερη τάξης 4.

N	Euler	RK4
10	$3.3384 \cdot 10^{-1}$	$1.9941 \cdot 10^{-2}$
20	$1.8507 \cdot 10^{-1}$	$1.0614 \cdot 10^{-3}$
40	$1.0054 \cdot 10^{-1}$	$6.5959 \cdot 10^{-5}$
80	$5.2790 \cdot 10^{-2}$	$4.0108 \cdot 10^{-6}$
160	$2.7111 \cdot 10^{-2}$	$2.4500 \cdot 10^{-7}$
320	$1.3748 \cdot 10^{-2}$	$1.5100 \cdot 10^{-8}$

Επίσης, παρατηρούμε συγκρίνοντας τα αποτελέσματα της δεύτερης και τρίτης στήλης του πίνακα αυτού, ότι το σφάλμα της RK4 με $N = 10$ βήματα είναι μικρότερο από αυτό της Euler με $N = 160$ βήματα. Από την άλλη μεριά, στα 160 βήματα της Euler χρειαζόμαστε 160 υπολογισμούς τιμών της $f(t, y)$ (έναν ανά βήμα), ενώ στα 10 βήματα της RK4 απαιτούνται 40 υπολογισμοί της f . (Θυμηθείτε ότι ως κόστος ανά βήμα μετράμε μόνο το πλήθος των υπολογισμών της f , μιας και αυτό είναι το πιο δαπανηρό μέρος του υπολογισμού.) Άρα, για το συγκεκριμένο παράδειγμα, ενώ αρχικά βλέποντας το σφάλμα θα έλεγε κανείς ότι η RK4 είναι

16 φορές πιο αποτελεσματική από την Euler, αν συνυπολογίσουμε και το κόστος, είναι πιο ρεαλιστικό να πούμε ότι η RK4 είναι 4 φορές πιο αποτελεσματική από την Euler. $\square$

6.5 Αυτόματη μεταβολή βήματος

Μέχρι στιγμής είδαμε διάφορες μεθόδους στις οποίες (εφ' όσον γνωρίζουμε την τάξη ακρίβειάς τους) μπορούμε να προβλέψουμε τη βελτίωση που επιφέρουμε στο σφάλμα καθώς μικραίνουμε το βήμα h . Καμία όμως από τις μεθόδους αυτές δεν επιτρέπει στο χρήστη να επιλέξει ένα κατάλληλο βήμα h ώστε να λύσει ένα Π.Α.Τ. με μια προεπιλεγμένη ακρίβεια ε και με το λιγότερο δυνατό κόστος. Για παράδειγμα, γνωρίζουμε ότι αν έχουμε μια μέθοδο τάξης $O(h^4)$ και μικρύνουμε το βήμα από h σε $h/2$, τότε το σφάλμα θα μικρύνει περίπου κατά 16 φορές, αλλά αυτό δεν εξασφαλίζει ότι κανένα από τα δύο αυτά βήματα οδηγεί σε σφάλμα μικρότερο από, π.χ., $\varepsilon = 10^{-3}$.

Μια «καλή» επιλογή βήματος θα πρέπει βεβαίως να λαμβάνει υπόψιν τη συμπεριφορά της λύσης του συγκεκριμένου προβλήματος και θα πρέπει να ελέγχεται και πιθανώς να μεταβάλλεται κατά τη διάρκεια της αριθμητικής επίλυσης του προβλήματος. Σκεφτείτε, π.χ., διαστήματα στα οποία η λύση μεταβάλλεται πολύ γρήγορα και άλλα στα οποία μεταβάλλεται αργά. Θα θέλαμε στα μεν η μέθοδός μας να επιλέγει μικρό βήμα, και μεγαλύτερο στα δεύτερα, αφού αν επιλέγαμε ένα μικρό βήμα σε όλο το διάστημα, τότε τόσο το υπολογιστικό κόστος, όσο και η επιρροή των σφαλμάτων στρογγύλευσης, θα αυξάνονταν σημαντικά.

Στη συνέχεια θα περιγράψουμε έναν πρακτικό τρόπο για να πετύχουμε κάτι τέτοιο με άμεσες μεθόδους. Οποιαδήποτε (μονοβηματική) άμεση μέθοδος για την επίλυση του Π.Α.Τ. (6.4), με μεταβλητό βήμα $h_n = t^{n+1} - t^n$, μπορεί να γραφεί συνοπτικά στη μορφή

$$y^{n+1} = y^n + h_n \Phi(t^n, y^n, h_n), \quad \text{για } n = 0, 1, \dots, N-1,$$

για κάποια συνάρτηση Φ . Ας υποθέσουμε ότι έχουμε δύο μεθόδους επίλυσης, μια RK με τάξη ακρίβειας p ,

$$y^{n+1} = y^n + h_n \Phi(t^n, y^n, h_n), \quad \text{για } n = 0, 1, \dots, N-1,$$

και μια μεγαλύτερης τάξης, στην πράξη συνήθως τάξης $p+1$,

$$\tilde{y}^{n+1} = y^n + h_n \tilde{\Phi}(t^n, y^n, h_n), \quad \text{για } n = 0, 1, \dots, N-1.$$

Για τα τοπικά σφάλματα αυτών των μεθόδων (ανατρέξτε στον Ορισμό 6.2) έχουμε

$$\begin{aligned} \delta^n &= [y(t^n) + h_n \Phi(t^n, y(t^n), h_n)] - y(t^{n+1}) = O(h_n^{p+1}) \\ \tilde{\delta}^n &= [y(t^n) + h_n \tilde{\Phi}(t^n, y(t^n), h_n)] - y(t^{n+1}) = O(h_n^{p+2}). \end{aligned}$$

Άρα το τοπικό σφάλμα της μεθόδου τάξης p είναι ίσο με

$$\delta^n = [y(t^n) + h_n \Phi(t^n, y(t^n), h_n)] - [y(t^n) + h_n \tilde{\Phi}(t^n, y(t^n), h_n)] + O(h_n^{p+2}),$$

δηλαδή προσεγγίζεται πολύ καλά από τη διαφορά

$$[y(t^n) + h_n \Phi(t^n, y(t^n), h_n)] - [y(t^n) + h_n \tilde{\Phi}(t^n, y(t^n), h_n)].$$

Τώρα, έστω ότι έχουμε προσδιορίσει τα βήματα h_i μέχρι το σημείο t^n , και έχουμε υπολογίσει τις προσεγγίσεις $y^0, \dots, y^n$ με τη μέθοδο τάξης p . Θέτουμε αρχικά $h_n = h_{n-1}$, και υπολογίζουμε τις προσεγγίσεις

$$\begin{cases} y^{n+1} = y^n + h_n \Phi(t^n, y^n, h_n) \\ \tilde{y}^{n+1} = y^n + h_n \tilde{\Phi}(t^n, y^n, h_n). \end{cases}$$

Δεχόμαστε ότι η διαφορά των δύο τιμών $y^{n+1} - \tilde{y}^{n+1}$ που μόλις υπολογίσαμε είναι μια καλή προσέγγιση της διαφοράς

$$[y(t^n) + h_n \Phi(t^n, y(t^n), h_n)] - [y(t^n) + h_n \tilde{\Phi}(t^n, y(t^n), h_n)],$$

που όπως είδαμε παραπάνω είναι μια καλή εκτίμηση του τοπικού σφάλματος δ^n . Στη συνέχεια, για να προσδιορίσουμε το επόμενο βήμα h_n , συγκρίνουμε, για παράδειγμα, την $|y^{n+1} - \tilde{y}^{n+1}|$ με το γινόμενο εh_n (όπου ε είναι η προκαθορισμένη απαιτούμενη ακρίβεια) («έλεγχος του τοπικού σφάλματος ανά βήμα»). Συγκεκριμένα,

- Αν $|y^{n+1} - \tilde{y}^{n+1}| \leq \varepsilon h_n$, τότε αποδεχόμαστε το βήμα h_n , και συνεχίζουμε έχοντας αποδεχθεί την τιμή y^{n+1} ως προσέγγιση της $y(t^{n+1})$.
(Αν η διαφορά $|y^{n+1} - \tilde{y}^{n+1}|$ προκύψει πολύ μικρότερη από το εh_n θεωρούμε ότι το βήμα είναι μικρότερο απ' ό,τι χρειαζόμαστε, το μεγαλώνουμε, υπολογίζουμε ξανά τα $y^{n+1}, \tilde{y}^{n+1}$ και επαναλαμβάνουμε τον έλεγχο.
- Αν $|y^{n+1} - \tilde{y}^{n+1}| > \varepsilon h_n$, τότε θεωρούμε ότι το τοπικό σφάλμα είναι μεγάλο οπότε μικραίνουμε το βήμα και υπολογίζουμε ξανά τα $y^{n+1}, \tilde{y}^{n+1}$, με την καινούργια τιμή του h_n .

Για να είναι αυτή η διαδικασία αποτελεσματική θα πρέπει βεβαίως η αυτόματη επιλογή βήματος να είναι πολύ οικονομικότερη από τον υπολογισμό του y^{n+1} για ένα εξαρχής πολύ μικρό βήμα h . Για να συμβεί αυτό θα πρέπει, π.χ., το πρόσθετο κόστος για τον υπολογισμό του $\tilde{y}^{n+1}$ να είναι μικρό σε σχέση με τον υπολογισμό του y^{n+1} . Αυτό επιτυγχάνεται στην πράξη αν κατασκευάσουμε ζεύγη μεθόδων RK με τα ίδια ενδιάμεσα στάδια, αλλά με διαφορετικά διανύσματα (b_i) , $(\tilde{b}_i)$. Στην περίπτωση αυτή το μητρώο των μεθόδων θα είναι της μορφής

$$\begin{array}{cccc|c}
 0 & & & 0 & \tau_1 \\
 a_{21} & 0 & & & \tau_2 \\
 \vdots & \vdots & \ddots & & \vdots \\
 a_{q1} & \dots & \dots & 0 & \tau_q \\
 \hline
 b_1 & b_2 & \dots & b_q & \\
 \hline
 \tilde{b}_1 & \tilde{b}_2 & \dots & \tilde{b}_q &
 \end{array}$$

και λέμε ότι η μέθοδος τάξης p (που αντιστοιχεί στα b_i) εμφυτεύεται στη μέθοδο τάξης $p + 1$ (που αντιστοιχεί στα $\tilde{b}_i$).

Για περισσότερες πληροφορίες παραπέμπουμε στο βιβλίο των Ακρίβη-Δουγαλή [2, §3.6].

6.6 Επίλυση Π.Α.Τ. στο MATLAB

Το MATLAB παρέχει ένα μεγάλο πλήθος αλγορίθμων για την επίλυση του Π.Α.Τ. (6.4). Σε αυτή την παράγραφο θα περιγράψουμε τον τρόπο χρήσης της δημοφιλέστερης μεθόδου για την επίλυση τέτοιων προβλημάτων που είναι η λεγόμενη ode45. Η ode45 είναι μια μέθοδος μεταβλητού βήματος και υλοποιεί ένα ζεύγος μεθόδων RK με $q = 7$ στάδια και τάξεις 4 και 5. Η μέθοδος αυτή οφείλεται στους Dormand και Prince (για περισσότερες πληροφορίες για τη μέθοδο αυτή μπορεί, π.χ., κανείς να ανατρέξει στο βιβλίο των Ακρίβη-Δουγαλή [2, §3.6]).

Παράδειγμα. Θεωρούμε το Π.Α.Τ. (6.25), του οποίου η ακριβής λύση είναι $y(t) = e^{\arctan t}$.

Αρχικά, δημιουργούμε το αρχείο `myf.m` για τον υπολογισμό τιμών της συνάρτησης $f(t, y) = \frac{y}{1+t^2}$ (βλ. Παράρτημα Β' για μια σύντομη περιγραφή των M-αρχείων).

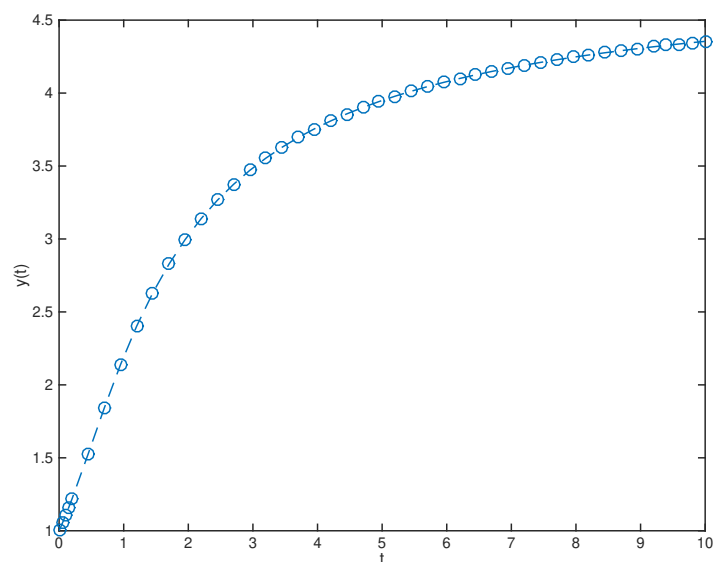
`myf.m`

```
function f = myf(t,y)
%
% Function myf
% Inputs: t, y
% Output: The value f(t,y)
%
f = y./(1 + t.^2);
```

Στη συνέχεια, πληκτρολογούμε στη γραμμή εντολών:

```
>> tspan = [0,10];
>> yzero = 1;
>> [t,y] = ode45(@myf,tspan,yzero);
>> plot(t,y),'o--','MarkerSize',8);
>> xlabel('t')
>> ylabel('y(t)')
```

Μέσω των εντολών αυτών κατασκευάζουμε μια προσέγγιση της ακριβούς λύσης, που αποθηκεύεται στο διάνυσμα (στήλη) y , στα σημεία του διαστήματος $[0, 10]$ που αποθηκεύονται στο διάνυσμα (στήλη) t . Οι προσεγγίσεις αυτές απεικονίζονται γραφικά στο Σχήμα 6.10. Παρατηρήστε ότι τα ορίσματα της συνάρτησης `ode45` είναι η συνάρτηση `myf`, το διάνυσμα `tspan` που καθορίζει το διάστημα στο οποίο παίρνει τιμές η ανεξάρτητη μεταβλητή t , και η αρχική συνθήκη `yzero`.



Σχήμα 6.10: Η μέθοδος `ode45` για το πρόβλημα (6.25).

Αν δεν μας ενδιαφέρουν τα διανύσματα t και y που επιστρέφει η `rk4`, μπορούμε απλώς να γράψουμε στη γραμμή εντολών

```
>> ode45(@myf, tspan, yzero);
```

και το MATLAB θα σχεδιάσει τη λύση ενώ λύνει το Π.Α.Τ., οπότε θα δούμε και πάλι τη γραφική παράσταση του Σχήματος 6.10.

Δεδομένου ότι σε αυτό το παράδειγμα γνωρίζουμε την ακριβή λύση του προβλήματος μπορούμε με την ακόλουθη εντολή να υπολογίσουμε το μέγιστο σφάλμα της προσέγγισης της λύσης με την ode45.

```
>> max(abs(y - exp(atan(t))))
ans =
    4.8982e-04
```

Η συνάρτηση ode45 επιτρέπει στον χρήστη να μεταβάλλει κάποιες παραμέτρους του αλγορίθμου μέσω ενός τέταρτου (προαιρετικού) ορίσματος που λέγεται `options` και συντάσσεται ως εξής:

```
options = odeset('name1', value1, 'name2', value2, ...);
ode45(@myf, tspan, yzero, options);
```

Τότε η συνάρτηση `odeset` αναθέτει στις παραμέτρους `name1`, `name2`, ..., τις τιμές `value1`, `value2`, ..., αντίστοιχα. Κάποιες χρήσιμες παράμετροι είναι οι ακόλουθες:

- `RelTol`. Είναι θετική σταθερά που καθορίζει το σχετικό μέγιστο σφάλμα, και έχει προκαθορισμένη τιμή $1.0e-3$. Η ode45 προσπαθεί να επιλέξει το βήμα έτσι ώστε σε κάθε βήμα το σφάλμα $e(i)$ να ικανοποιεί

$$e(i) \leq \max(\text{RelTol} * \text{abs}(y(i)), \text{AbsTol}(i))$$

Η προκαθορισμένη τιμή $1.0e-3$, αντιστοιχεί χονδρικά σε ακρίβεια 0.1%.

- `AbsTol`. Είναι θετική σταθερά (ή διάνυσμα) με προκαθορισμένη τιμή $1.0e-6$.
- `InitialStep`. Προτεινόμενο αρχικό βήμα.
- `MaxStep`. Μέγιστο επιτρεπτό βήμα. Αν ο χρήστης δεν αλλάξει την τιμή του, τότε ορίζεται ως το $1/10$ του μήκους του διαστήματος ολοκλήρωσης.
- `Stats`. Η προκαθορισμένη τιμή είναι `'off'`. Αν την αλλάξουμε σε `'on'`, τότε με την ολοκλήρωση της επίλυσης δίνονται πληροφορίες για τον αριθμό των βημάτων που εκτελέστηκαν επιτυχώς, για το πλήθος των «αποτυχημένων προσπαθειών», και για το πόσοι υπολογισμοί της συνάρτησης $f(t, y)$ απαιτήθηκαν.

Για τον πλήρη κατάλογο των παραμέτρων της `odeset` ανατρέξτε στο εγχειρίδιο χρήσης του MATLAB ή γράψτε `help odeset` στη γραμμή εντολών.

6.7 Άκαμπτα προβλήματα

Στην παράγραφο αυτή θα δώσουμε μια πολύ αδρή περιγραφή των λεγόμενων άκαμπτων προβλημάτων ακολουθώντας την προσέγγιση της αντίστοιχης παραγράφου του βιβλίου του Moler, [10]. Για μια πιο μαθηματική περιγραφή παραπέμπουμε στο βιβλίο των Ακρίβη-Δουγαλή [2, §2.3]).

Ο χαρακτηρισμός ενός προβλήματος ως άκαμπτου είναι μια εν γένει λεπτή, δύσκολη και ιδιαίτερα σημαντική διαδικασία για την αριθμητική επίλυση μιας συνήθους διαφορικής εξίσωσης. Θα μπορούσαμε να πούμε ότι εξαρτάται από τη διαφορική εξίσωση, τις αρχικές συνθήκες, αλλά και την αριθμητική μέθοδο. Κάπως γενικά, θα μπορούσαμε να πούμε ότι ένα πρόβλημα αρχικών τιμών είναι *άκαμπτο* αν η λύση που αναζητούμε μεταβάλλεται σχετικά αργά, αλλά υπάρχουν άλλες «κοντινές» λύσεις που μεταβάλλονται πολύ γρήγορα έτσι ώστε να αναγκάζομαστε να παίρνουμε πολύ μικρό βήμα στην αριθμητική μας μέθοδο, για να μπορέσουμε να επιτύχουμε ικανοποιητικές προσεγγίσεις.

Ας δούμε το ακόλουθο παράδειγμα. Θεωρήστε το Π.Α.Τ.

$$\begin{cases} y' = y^2 - y^3, & 0 \leq t \leq 2/\delta, \\ y(0) = \delta. \end{cases} \quad (6.26)$$

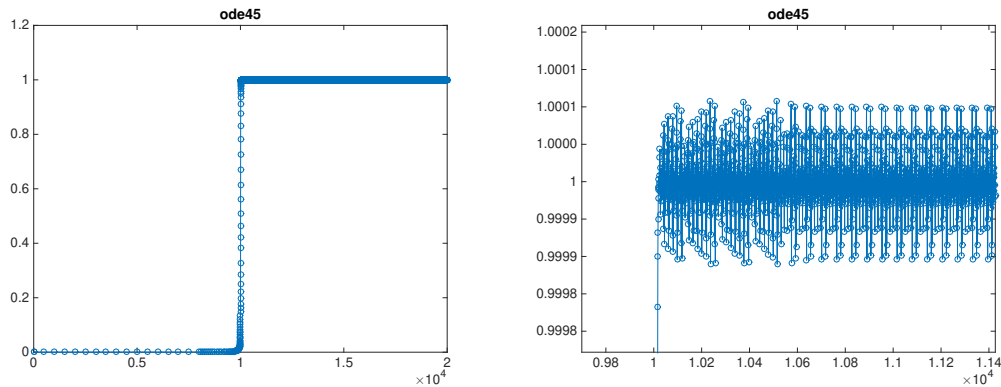
Αυτό το Π.Α.Τ. περιγράφει ένα πρόβλημα ανάφλεξης, όπου όταν ανάβουμε ένα σπέρτο η ακτίνα της «πύρινης μπάλας» μεγαλώνει μέχρις ότου φτάσει σε κάποια κρίσιμη τιμή και μετά παραμένει περίπου σταθερή. Η $y(t)$ περιγράφει την ακτίνα αυτή, κατά τη χρονική στιγμή t , και δ είναι η αρχική ακτίνα που θεωρείται μικρή.

Ξεκινάμε λύνοντας το πρόβλημα αυτό με τη συνάρτηση `ode45` για μια μικρή τιμή της παραμέτρου δ , π.χ. για $\delta = 10^{-4}$. Για να το πετύχουμε αυτό γράφουμε στο MATLAB τις ακόλουθες εντολές:

```
>> delta = 0.0001;
>> tspan = [0 , 2/delta];
>> F = @(t,y) y.^2 - y.^3;
>> options = odeset('RelTol',1.e-4,'Stats','on');
>> ode45(F, tspan, delta, options);
```

```
3040 successful steps
323 failed attempts
20179 function evaluations
```

Τότε η ode45 σχεδιάζει τη λύση του Π.Α.Τ., η οποία απεικονίζεται στη γραφική παράσταση στο αριστερό μέρος του Σχήματος 6.11.



Σχήμα 6.11: Λύση του Π.Α.Τ. (6.26) με την ode45.

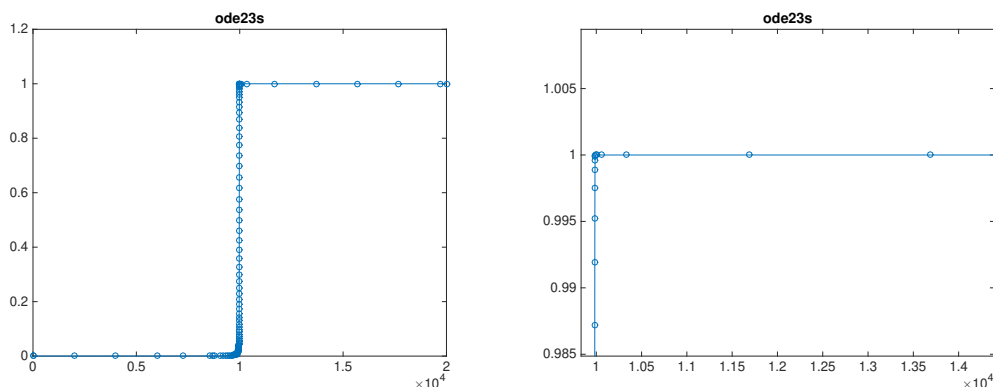
Παρατηρούμε ότι η λύση παραμένει πολύ κοντά στο μηδέν μέχρι τη χρονική στιγμή $t = 10^4$, στη συνέχεια αυξάνεται ταχύτατα μέχρι να πάρει την τιμή περίπου ένα, και μετά παραμένει σχεδόν σταθερή κοντά στο ένα. Διαπιστώνουμε αμέσως ότι η ode45 αναγκάζεται να πάρει πολύ μικρό βήμα αν και η λύση μετά τη χρονική στιγμή $t = 10^4$ παραμένει σχεδόν σταθερή. Αν μεγεθύνουμε τη λύση κοντά στο $t = 10^4$ παίρνουμε την εικόνα στο δεξιό μέρος του Σχήματος 6.11. Βλέπουμε ότι η ode45 φέρνει εις πέρας την αποστολή της μιας και καταφέρνει να κρατήσει τη λύση στο διάστημα 1 ± 10^{-4} , αλλά καταβάλλει πολύ μεγάλη προσπάθεια για να το πετύχει. Η συμπεριφορά της ode45 είναι τυπική του τι συμβαίνει όταν επιχειρούμε να λύσουμε ένα άκαμπτο πρόβλημα με μια μέθοδο μεταβλητού βήματος που δεν είναι σχεδιασμένη ώστε να λαμβάνει υπόψιν την «ακαμψία»: Η μέθοδος υπολογίζει μια προσέγγιση της λύσης που είναι ακριβής, αλλά η συμπεριφορά της δεν είναι «ομαλή» και σίγουρα δεν είναι τόσο αποτελεσματική όσο μια μέθοδος που έχει σχεδιαστεί για να επιλύει άκαμπτα συστήματα.

Μια τέτοια μέθοδος, κατάλληλη για την επίλυση (μετρίως) άκαμπτων προβλημάτων, υλοποιείται στο MATLAB με τη συνάρτηση ode23s. Η ode23s είναι μια μονοβηματική, άμεση μέθοδος μεταβλητού βήματος που υλοποιεί ένα ζεύγος μεθόδων RK με τάξεις 2 και 3 και οφείλεται στους Bogacki και Shampine. Για να την καλέσουμε γράφουμε στη γραμμή εντολών:

```
>> ode23s(F, tspan, delta, options);
```

```
99 successful steps
7 failed attempts
412 function evaluations
99 partial derivatives
106 LU decompositions
318 solutions of linear systems
```

Παρατηρήστε ότι η μέθοδος αυτή χρησιμοποιεί 99 βήματα έναντι 3040 βημάτων της ode45. Η προσέγγιση της λύσης με την ode23s απεικονίζεται στις γραφικές παραστάσεις του Σχήματος 6.12.



Σχήμα 6.12: Λύση του Π.Α.Τ. (6.26) με την ode23s.

6.8 Αριθμητικές μέθοδοι για συστήματα Δ.Ε.

Στην παράγραφο αυτή θα κάνουμε μια πολύ σύντομη παρουσίαση της γενίκευσης των μεθόδων που είδαμε μέχρι τώρα για συστήματα Σ.Δ.Ε. πρώτης τάξης. Γενικά, ένα Π.Α.Τ. για ένα σύστημα Δ.Ε. πρώτης τάξης γράφεται στη μορφή

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}), \quad t \in [a, b], \quad (6.27)$$

$$\mathbf{y}(a) = \mathbf{u}, \quad (6.28)$$

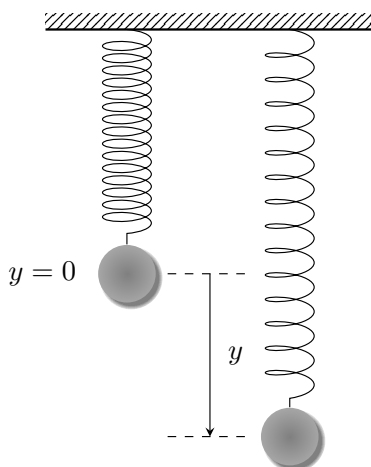
όπου $\mathbf{y}$ είναι ένα διάνυσμα με συνιστώσες τις (άγνωστες) συναρτήσεις $y_1, \dots, y_n$, $\mathbf{f}(t, \mathbf{y})$ είναι επίσης ένα διάνυσμα με συνιστώσες $f_1(t, \mathbf{y}), \dots, f_n(t, \mathbf{y})$, όπου καθεμία από τις f_i είναι, εν

γένει, μια μη γραμμική συνάρτηση όλων των συνιστωσών του διανύσματος $\mathbf{y}$, και $\mathbf{u}$ είναι ένα γνωστό διάνυσμα στον $\mathbb{R}^n$. Χρησιμοποιώντας το συμβολισμό αυτό όλες οι μέθοδοι που είδαμε μέχρι τώρα εξακολουθούν να εφαρμόζονται πρακτικά अपαράλλακτες.

Στη συνέχεια θα παρουσιάσουμε ένα συγκεκριμένο παράδειγμα που αφορά την κίνηση ενός συστήματος μάζας-ελατηρίου. Θα διατυπώσουμε το φυσικό πρόβλημα, θα δούμε τη διαδικασία κατασκευής ενός μαθηματικού «μοντέλου» για το πρόβλημα αυτό και την περιγραφή του ως ένα σύστημα συνήθων διαφορικών εξισώσεων πρώτης τάξης και, τέλος, θα περιγράψουμε τρόπους για την αριθμητική του επίλυση.

Μαθηματική μοντελοποίηση συστήματος μάζας-ελατηρίου

Θεωρούμε ένα ελατήριο που κρέμεται κατακόρυφα από ένα σταθερό σημείο. Στο κάτω μέρος του είναι τοποθετημένη μια μεταλλική μπάλα, βλ. Σχήμα 6.13.



Σχήμα 6.13: Σύστημα μάζας-ελατηρίου.

Έστω ότι η θέση ηρεμίας του συστήματος είναι $y = 0$, και θεωρούμε ως θετική φορά της κίνησης τη φορά προς τα κάτω. Έτσι, δυνάμεις με φορά προς τα κάτω θεωρούνται ως θετικές, ενώ δυνάμεις με φορά προς τα πάνω ως αρνητικές. Αρχικά, τραβάμε τη μπάλα κατακόρυφα προς τα κάτω κατά $y > 0$. Αυτό προκαλεί στο ελατήριο μια δύναμη, που είναι ανάλογη της μετατόπισης y ,

$$F_1 = -ky, \quad (\text{Νόμος του Hooke}) \quad (6.29)$$

όπου k είναι μια σταθερά που εξαρτάται από το ελατήριο (άκαμπτα ελατήρια έχουν μεγάλο k). Το πρόσημο « $-$ » δηλώνει ότι η δύναμη έχει φορά προς τα πάνω, δηλαδή είναι αντίθετη στη μετατόπιση.

Η κίνηση του συστήματος μάζα-ελατήριο καθορίζεται από το δεύτερο νόμο του Νεύτωνα σύμφωνα με τον οποίο

$$\text{Μάζα} \times \text{Επιτάχυνση} = my'' = \text{Δύναμη}, \quad (6.30)$$

όπου ως «Δύναμη» εννοούμε τη συνισταμένη των δυνάμεων που ασκούνται στη μπάλα. Γενικά, σε κάθε σύστημα υπάρχει και απόσβεση, αλλιώς θα συνέχιζε να κινείται για πάντα. Αν όμως θεωρήσουμε ότι η απόσβεση είναι μικρή και μελετάμε την κίνηση του συστήματος για μικρά χρονικά διαστήματα, τότε μπορούμε να αγνοήσουμε την απόσβεση. Τότε από το νόμο του Νεύτωνα (6.30) καταλήγουμε στη Δ.Ε.: $my'' = F_1 = -ky$, επομένως

$$my'' + ky = 0 \quad \Leftrightarrow \quad y'' + \omega_0^2 y = 0, \quad \text{όπου} \quad \omega_0 = \sqrt{\frac{k}{m}}. \quad (6.31)$$

Η (6.31) είναι μια γραμμική ομογενής Δ.Ε. δεύτερης τάξης με σταθερούς συντελεστές. Εύκολα μπορούμε να δούμε ότι η γενική της λύση είναι

$$y(t) = A \cos \omega_0 t + B \sin \omega_0 t. \quad (6.32)$$

Η κίνηση αυτή λέγεται *αρμονική ταλάντωση* και έχει συχνότητα $f = \omega_0/(2\pi)$ Hz διότι το συνημίτονο και το ημίτονο που εμφανίζονται στην (6.32) έχουν περίοδο $2\pi/\omega_0$. Η συχνότητα αυτή αποκαλείται φυσική συχνότητα του συστήματος.

Τη σχέση (6.32) μπορούμε εναλλακτικά να τη γράψουμε στη μορφή

$$y(t) = C \cos(\omega_0 t - \delta), \quad (6.33)$$

όπου $C = \sqrt{A^2 + B^2}$ και γωνίας φάσης δ , με $\tan \delta = B/A$, η οποία αναδεικνύει τα φυσικά χαρακτηριστικά του πλάτους και της φάσης της μετατόπισης της (6.32). (Η (6.33) προκύπτει από την (6.32) αν χρησιμοποιήσουμε την τριγωνομετρική ταυτότητα $\cos(a-b) = \cos a \cos b + \sin a \sin b$, καθώς και τη σχέση $\cos(\arctan x) = \frac{1}{\sqrt{1+x^2}}$.)

Διατύπωση του προβλήματος σε μορφής συστήματος Σ.Δ.Ε. πρώτης τάξης

Θα μετατρέψουμε τη Δ.Ε. (που εδώ είναι δεύτερης τάξης) σ' ένα σύστημα Δ.Ε. πρώτης τάξης. Ας δούμε πως γίνεται αυτό για τη Δ.Ε. (6.31). Καταρχάς, κάνουμε την ακόλουθη αντικατάσταση εισάγοντας τις καινούριες συναρτήσεις:

$$y_1 = y, \quad (6.34)$$

$$y_2 = y'. \quad (6.35)$$

Τώρα, η y_1 συμβολίζει τη μετατόπιση και η y_2 την ταχύτητα. Παραγωγίζουμε τις (6.34), (6.35), και έχουμε

$$\begin{aligned} y_1 = y &\Rightarrow y'_1 = y' \stackrel{(6.35)}{=} y_2, \\ y_2 = y' &\Rightarrow y'_2 = y'' \stackrel{(6.31)}{=} -\omega_0^2 y \stackrel{(6.34)}{=} -\omega_0^2 y_1. \end{aligned}$$

Έτσι καταφέραμε να γράψουμε τη Δ.Ε. (6.31) σε μορφή συστήματος δύο Δ.Ε. πρώτης τάξης,

$$\begin{cases} y'_1 = y_2 \\ y'_2 = -\omega_0^2 y_1. \end{cases} \quad (6.36)$$

Ισοδύναμα, μπορούμε να γράψουμε το σύστημα (6.36) στην ακόλουθη διανυσματική μορφή

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}), \quad (6.37)$$

όπου

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad \text{και} \quad \mathbf{f}(t, \mathbf{y}) = \begin{pmatrix} y_2 \\ -\omega_0^2 y_1 \end{pmatrix}.$$

Τέλος, το σύστημα (6.37) μπορούμε να το γράψουμε και στην ακόλουθη μορφή πινάκων

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}' = \begin{pmatrix} 0 & 1 \\ -\omega_0^2 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \Leftrightarrow \mathbf{y}' = A \mathbf{y}, \quad (6.38)$$

όπου

$$A = \begin{pmatrix} 0 & 1 \\ -\omega_0^2 & 0 \end{pmatrix}.$$

Το Π.Α.Τ. για το σύστημα (6.37)

Έστω τώρα $[a, b] = [0, 4\pi]$ και ας θεωρήσουμε ότι κατά τη χρονική στιγμή $t = 0$ γνωρίζουμε την αρχική μετατόπιση της μάζας, έστω $y_1(0) = 1$, και την αρχική ταχύτητα, έστω $y_2(0) = 0$. Επιπλέον, υποθέτουμε ότι η μάζα $m = 1$ kg και ότι η σταθερά του ελατηρίου $k = 1$ N/m (επομένως $\omega_0^2 = 1$). Τότε καταλήγουμε στο ακόλουθο Π.Α.Τ.

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}), \quad t \in [0, 4\pi], \quad (6.39)$$

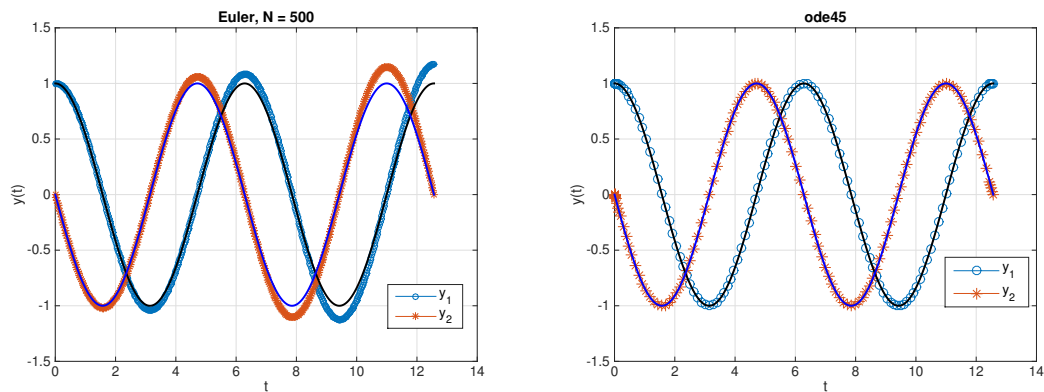
$$\mathbf{y}(0) = \mathbf{u}, \quad (6.40)$$

όπου

$$\mathbf{f}(t, \mathbf{y}) = \begin{pmatrix} y_2 \\ -y_1 \end{pmatrix} \quad \text{και} \quad \mathbf{u} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Για να λύσουμε το Π.Α.Τ. (6.39), (6.40), χρησιμοποιούμε τη μέθοδο του Νεύτωνα για έναν ομοιόμορφο διαμερισμό του διαστήματος $[0, 4\pi]$ με $N = 500$ υποδιαστήματα, και την ode45

του MATLAB. Στο Σχήμα 6.14 σχεδιάζουμε την προσέγγιση της λύσης $(y_1, y_2)^T$ με τη μέθοδο του Euler (αριστερά) και με την ode45 (δεξιά), και τις συγκρίνουμε με την ακριβή λύση $(y_1(t), y_2(t))^T = (\cos t, -\sin t)$. (Οι $y_1(t) = \cos t$ και $y_2(t) = -\sin t$, απεικονίζονται στο Σχήμα 6.14 με συνεχείς γραμμές, μαύρη και μπλε, αντίστοιχα.)



Σχήμα 6.14: Λύση του Π.Α.Τ. (6.39), (6.40), με την μέθοδο του Euler (αριστερά) και με την ode45 (δεξιά).

Αμέσως βλέπουμε ότι το βήμα $h = 4\pi/500$, που πήραμε στη μέθοδο του Euler δεν είναι αρκετά μικρό ώστε να πάρουμε μια ικανοποιητική προσέγγιση. Στο ακόλουθο αρχείο (script) με όνομα `odesystem.m` δίνουμε ένα παράδειγμα χρήσης της ode45 για την επίλυση του Π.Α.Τ. (6.39), (6.40), και στο αρχείο `odefcn.m` ορίζεται η συνάρτηση $f(t, \mathbf{y})$ η οποία αποτελεί ένα από τα ορίσματα της ode45.

odesystem.m

```

a = 0;
b = 4*pi;
tspan = [a,b];
incond = [1,0]';
options = odeset('RelTol',1.e-4,'Stats','on');
[t,y] = ode45(@(t,y) odefcn(t,y), tspan, incond, options);
%
plot(t, y(:,1), 'o-', t, y(:,2), '*-', 'MarkerSize', 9);
axis([0 14 -1.5 1.5]);
xlabel('t');
ylabel('y(t)');
grid on;
%
hold;
%
exact = @(t) [cos(t) ; -sin(t)];
tfine = linspace(a, b, 1001);
exsol = exact(tfine);
%
plot(tfine, exsol(1,:), 'k', tfine, exsol(2,:), 'b')

```

odefcn.m

```

function dydt = odefcn(t,y)
dydt = [0 1; -1 0] * y;
end

```

Παρατηρούμε ότι για την ode45 απαιτούνται 31 βήματα και 199 υπολογισμοί της συνάρτησης, ενώ η Euler με $N = 500$ κάνει 500 υπολογισμούς της συνάρτησης και δίνει σαφώς χειρότερα αποτελέσματα.

Ασκήσεις

1. Ποιές από τις ακόλουθες συναρτήσεις ικανοποιούν τη συνθήκη Lipschitz (6.3) ως προς y ; Για τις συναρτήσεις που την ικανοποιούν να βρείτε τη σταθερά Lipschitz.

(α') $f(t, y) = \sqrt{t^2 + y^2}$, για $t \in [-1, 1]$.

(β') $f(t, y) = |y|$.

(γ') $f(t, y) = |y|/t$, για $t \in [-1, 1]$.

2. Θεωρήστε το Π.Α.Τ.

$$\begin{cases} y'(t) = y(t) + \frac{1}{2}e^t, & 0 \leq t \leq 1, \\ y(0) = 0. \end{cases}$$

(α') Δείξτε ότι το πρόβλημα αυτό έχει μοναδική λύση.

(β') Διατυπώστε την άμεση μέθοδο του Euler για το πρόβλημα αυτό για έναν ομοιόμορφο διαμερισμό του διαστήματος $[0, 1]$.

(γ') Διατυπώστε την πεπλεγμένη μέθοδο του Euler για το πρόβλημα αυτό για έναν ομοιόμορφο διαμερισμό του διαστήματος $[0, 1]$, και εκφράστε το y^{n+1} συναρτήσει του y^n (αν φυσικά αυτό είναι εφικτό).

3. Θεωρήστε το Π.Α.Τ.

$$\begin{cases} y'(t) = 1 + t - y(t), & t \geq 0, \\ y(0) = 0. \end{cases}$$

Αποδείξτε ότι η μέθοδος του Euler με σταθερό βήμα h για το πρόβλημα αυτό γράφεται στη μορφή $y^n = n h$, $n = 0, 1, 2, \dots$

4. Να λύσετε με την κλασσική μέθοδο Runge-Kutta (6.24) το Π.Α.Τ.

$$\begin{cases} y'(t) = y(t) + t, & 0 \leq t \leq 0.1, \\ y(0) = 0, \end{cases}$$

για έναν ομοιόμορφο διαμερισμό με βήμα $h = 0.1$. Η ακριβής λύση για το πρόβλημα αυτό είναι $y(t) = -1 - t + e^t$. Υπολογίστε το σχετικό σφάλμα στο $t = 0.1$.

5. Θεωρήστε τη μέθοδο

$$\begin{aligned} y^{n+1} &= y^n + h f(t^{n+1}, y^n + h f(t^n, y^n)), \quad n = 0, 1, \dots, N-1, \\ y^0 &= y_0. \end{aligned}$$

Ελέγξτε αν είναι απόλυτα ευσταθής για κάθε $h \in \mathbb{R}$.

6. Θεωρήστε τη μέθοδο

$$\begin{aligned}y^{n+1} &= y^n + h f\left(t^{n+1/2}, \frac{1}{2}(y^n + y^{n+1})\right), \quad n = 0, 1, \dots, N-1, \\y^0 &= y_0,\end{aligned}$$

όπου $t^{n+1/2} = (t^n + t^{n+1})/2$. Ελέγξτε αν είναι απόλυτα ευσταθής για κάθε $h \in \mathbb{R}$.

Παράρτημα Α΄

Επισκόπηση αποτελεσμάτων από τον Απειροστικό Λογισμό

Α΄.1 Υπολογισμός τιμών πολυωνύμου - Σχήμα Horner

Έστω ότι θέλουμε να υπολογίσουμε τιμές του πολυωνύμου

$$p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_{n-1}x^{n-1} + a_nx^n. \quad (\text{Α΄.1})$$

Για να υπολογίσουμε τιμές του p αποτελεσματικά, το γράφουμε στη μορφή

$$p(x) = a_0 + x(a_1 + x(a_2 + \cdots + x(a_{n-1} + xa_n))) \cdots), \quad (\text{Α΄.2})$$

και εκτελούμε τις πράξεις από τα μέσα προς τα έξω. Αυτό μπορεί να γίνει πολύ εύκολα με ένα πρόγραμμα που εκτελεί τον αλγόριθμο που δίνεται στον ακόλουθο ψευδοκώδικα:

```
y ← an
για i = n - 1, ..., 0:
    y ← ai + x · y
```

Κατά την έξοδο του αλγορίθμου η μεταβλητή y θα περιέχει την ζητούμενη τιμή $p(x)$. Ο αλγόριθμος αυτός είναι γνωστός ως *σχήμα του Horner*.

Η πράξη $y \leftarrow a_i + x \cdot y$, δηλαδή το να ανασύρουμε από τη μνήμη του υπολογιστή τις σταθερές a_i , x , y , να εκτελέσουμε τον παλλαπλασιασμό και την πρόσθεση στο δεξί μέλος και να αποθηκεύσουμε το αποτέλεσμα ως νέα τιμή στη μεταβλητή y , είναι πολύ συχνή στα υπολογιστικά μαθηματικά. Για το λόγο αυτό έχει καθιερωθεί ως μονάδα μέτρησης των πράξεων στους αλγορίθμους και λέγεται *flop* (floating point operation). Άρα μπορούμε να πούμε ότι

το σχήμα του Horner για τον υπολογισμό της τιμής ενός πολυωνύμου βαθμού n απαιτεί n flops. (Αν επιχειρήσετε να υπολογίσετε την τιμή $p(x)$ απ' ευθείας από την (Α΄.1) θα εκτελέσετε πολύ περισσότερες πράξεις. Μπορείτε να υπολογίσετε, π.χ., το συνολικό πλήθος των πολλαπλασιασμών που απαιτούνται.)

Α΄.2 Ο τύπος του Taylor

Αν μια συνάρτηση f είναι συνεχής σε ένα κλειστό διάστημα $I = [a, b]$, και έχει συνεχείς παραγώγους τάξης $1, 2, \dots, (n+1)$ στο I , τότε για κάθε x και ξ στο I ισχύει

$$f(x) = f(\xi) + f'(\xi)(x - \xi) + \dots + \frac{f^{(n)}(\xi)}{(n)!}(x - \xi)^n + \frac{f^{(n+1)}(\eta)}{(n+1)!}(x - \xi)^{n+1}, \quad (\text{Α΄.3})$$

όπου το η είναι ένας αριθμός μεταξύ των ξ και x (και εξαρτάται και από τους δύο).

Σε συμπαγή μορφή η (Α΄.3) γράφεται ως εξής:

$$f(x) = \sum_{k=0}^n \frac{f^{(k)}(\xi)}{k!} (x - \xi)^k + E_{n+1}, \quad (\text{Α΄.4})$$

όπου ο όρος υπολοίπου E_{n+1} δίνεται από την σχέση

$$E_{n+1} = \frac{f^{(n+1)}(\eta)}{(n+1)!} (x - \xi)^{n+1},$$

με η όπως παραπάνω.

Αν στην (Α΄.3) κάνουμε μια αλλαγή μεταβλητών βάζοντας στη θέση του x το $x + h$ και στη θέση του ξ το x , καταλήγουμε στην ακόλουθη, ισοδύναμη, έκφραση του τύπου του Taylor.

Αν μια συνάρτηση f είναι συνεχής σε ένα κλειστό διάστημα $I = [a, b]$, και έχει συνεχείς παραγώγους τάξης $1, 2, \dots, (n+1)$ στο I , τότε για κάθε x στο I ισχύει

$$f(x+h) = f(x) + f'(x)h + \dots + \frac{f^{(n)}(x)}{(n)!}h^n + \underbrace{\frac{f^{(n+1)}(\eta)}{(n+1)!}h^{n+1}}_{E_{n+1}}, \quad (\text{Α΄.5})$$

όπου το h είναι οποιοσδήποτε αριθμός τέτοιος ώστε $x+h \in I$, και η είναι ένας αριθμός μεταξύ των x και $x+h$.

Ο όρος υπολοίπου E_{n+1} εξαρτάται από το h με δύο τρόπους: α) λόγω της ύπαρξης του όρου h^{n+1} , και β) επειδή το η εξαρτάται εν γένει από το h . Καθώς το h τείνει προς το μηδέν, ο όρος

E_{n+1} συγκλίνει και αυτός προς το μηδέν, ουσιαστικά τόσο γρήγορα όσο γρήγορα το h^{n+1} συγκλίνει στο μηδέν. Για να εκφράσουμε αυτό το ποιοτικό αποτέλεσμα γράφουμε

$$E_{n+1} = O(h^{n+1}) \quad \text{καθώς } h \rightarrow 0.$$

Η γραφή αυτή λέγεται *συμβολισμός του μεγάλου όμικρον* και είναι συντομογραφία της ανίσωσης

$$|E_{n+1}| \leq C |h|^{n+1}, \quad \text{όπου το } C \text{ είναι σταθερά.}$$

(Κάπως χονδρικά, θα μπορούσαμε να πούμε ότι η σχέση $E_{n+1} = O(h^{n+1})$ υποδηλώνει ότι το E_{n+1} έχει παρόμοια συμπεριφορά με την πολύ απλούστερη έκφραση h^{n+1} .)

Σημειώστε ότι, ουσιαστικά, η (Α΄.5) αντιστοιχεί σε μια ακολουθία θεωρημάτων, ένα για κάθε τιμή του n . Π.χ.,

$$\begin{aligned} n = 0 : \quad & f(x+h) = f(x) + f'(\eta_1) h \\ & = f(x) + O(h), \\ n = 1 : \quad & f(x+h) = f(x) + f'(x) h + \frac{f''(\eta_2)}{2} h^2 \\ & = f(x) + f'(x) h + O(h^2), \\ n = 2 : \quad & f(x+h) = f(x) + f'(x) h + \frac{f''(x)}{2} h^2 + \frac{f'''(\eta_3)}{6} h^3 \\ & = f(x) + f'(x) h + \frac{f''(x)}{2} h^2 + O(h^3). \end{aligned}$$

Α΄.3 Απόλυτο και σχετικό σφάλμα

Έστω ότι α και β είναι δύο αριθμοί και ας υποθέσουμε ότι ο β είναι μια προσέγγιση του α . Άρα, μπορούμε να πούμε, ότι ο β προσεγγίζει τον α με *σφάλμα* $\alpha - \beta$. Θα λέμε ότι το *απόλυτο σφάλμα* του β ως προσέγγιση του α είναι ο αριθμός $|\alpha - \beta|$, ενώ το *σχετικό σφάλμα* του β ως προσέγγιση του α είναι ο αριθμός $|\alpha - \beta|/|\alpha|$. Γενικά,

$$\begin{aligned} \text{απόλυτο σφάλμα} &= |\text{ακριβής τιμή} - \text{προσεγγιστική τιμή}| \\ \text{σχετικό σφάλμα} &= \frac{|\text{ακριβής τιμή} - \text{προσεγγιστική τιμή}|}{|\text{ακριβής τιμή}|} \end{aligned}$$

Παρατηρήστε ότι το σχετικό σφάλμα δεν ορίζεται αν η ακριβής τιμή είναι 0.

Για παράδειγμα, θεωρήστε τους αριθμούς $\alpha = 1.500$, $\beta = 1.499$, $\gamma = 0.001$, $\delta = 0.002$, και έστω ότι οι ακριβείς τιμές είναι οι α , γ , ενώ οι β , δ είναι οι προσεγγίσεις τους, αντίστοιχα. Τότε

για το απόλυτο σφάλμα τους έχουμε $|\alpha - \beta| = |\gamma - \delta| = 0.001$. Όμως τα σχετικά σφάλματα είναι

$$\frac{|\alpha - \beta|}{|\alpha|} = \frac{2}{3} \cdot 10^{-3} \approx 0.00067, \quad \text{και} \quad \frac{|\gamma - \delta|}{|\gamma|} = 1!$$

Οι τιμές αυτές υποδεικνύουν ξεκάθαρα ότι το β είναι μια καλή προσέγγιση του α , ενώ το δ είναι μια πολύ κακή προσέγγιση του γ .

Παράρτημα Β΄

M-αρχεία και συναρτήσεις στο MATLAB

Β΄.1 M-αρχεία

Ένα M-αρχείο είναι ένα αρχείο κειμένου που έχει στο όνομά του την επέκταση `.m` και περιέχει εντολές του MATLAB. Υπάρχουν δύο τύποι M-αρχείων:

Αρχεία script (ή αρχεία εντολών) τα οποία δεν έχουν ορίσματα εισόδου ή εξόδου και εκτελούν μια σειρά από εντολές σε μεταβλητές του χώρου εργασίας (workspace).

Αρχεία συναρτήσεων τα οποία περιλαμβάνουν μια γραμμή στην οποία ορίζεται η συνάρτηση, δέχονται ορίσματα εισόδου και επιστρέφουν μεταβλητές εξόδου. Οι εσωτερικές μεταβλητές δρούν τοπικά εκτός αν δηλωθούν ως global (ολικές).

Μπορείτε να δημιουργήσετε M-αρχεία με τον κειμενογράφο (editor) του MATLAB (ή και με οποιοδήποτε άλλο κειμενογράφο, π.χ., το notepad ή το wordpad. Μπορούμε να ανοίξουμε τον editor του MATLAB και από το παράθυρο εντολών πληκτρολογώντας:

```
>> edit
```

Είδαμε δύο παραδείγματα M-αρχείων τύπου script (τα `integ1.m` και `integ2.m`) στην Παράγραφο 1.5.

Θα δούμε τώρα ένα παράδειγμα M-αρχείου συνάρτησης. Έστω ότι θέλουμε να φτιάξουμε ένα αρχείο που να υπολογίζει τιμές της συνάρτησης του Παραδείγματος 2.3

$$f(x) = \frac{1}{2} - \frac{1}{1 + M|x - 1.05|}.$$

Ανοίγουμε με τον editor ένα αρχείο με όνομα, π.χ., `myfun.m` και γράφουμε:

myfun.m

```
function f = myfun(x,M)
%
%myfun : computes values of the function of Example 2.3
%
f = 0.5 - 1./(1 + M*abs(x - 1.05));
```

Παρατηρήστε ότι το όνομα του αρχείου πρέπει να είναι το ίδιο με το όνομα της συνάρτησης (στο παράδειγμά μας `myfun`). Οι γραμμές που ξεκινάνε με το σύμβολο `%` είναι σχόλια.

Προσοχή: Το όνομα μιας νέας συνάρτησης δεν πρέπει να συμπίπτει με όνομα εσωτερικής συνάρτησης του MATLAB. Ένας τρόπος για να το ελέγξετε π.χ., αν το όνομα `exp` χρησιμοποιείται ήδη, είναι μέσω της εντολής `type`. Π.χ. γράψτε:

```
>> type exp
```

```
'exp' is a built-in function.
```

Τώρα, με τις ακόλουθες εντολές στο MATLAB μπορούμε να υπολογίσουμε την τιμή $f(1)$ για $M = 2000$, ή να σχεδιάσουμε, και πάλι, τη γραφική της παράσταση της f στο διάστημα $[0.8, 1.8]$ για $M = 200$ (βλ. Σχήμα 2.4).

```
>>z = myfun(1,2000)
z =
    0.4901

>> x = 0.8:0.001:1.8;
>> y = myfun(x,200);
>> plot(x,y)
```

Παρατηρήστε ότι η `myfun` δέχεται ως όρισμα είτε ένα σημείο και επιστρέφει την τιμή της f στο σημείο αυτό ή ένα διάνυσμα, οπότε επιστρέφει την τιμή της f σε κάθε στοιχείο του διανύσματος αυτού. Αυτό είναι δυνατόν επειδή ορίσαμε την πράξη της διαίρεσης κατά στοιχείο (δηλαδή χρησιμοποιήσαμε `./` αντί του `/`).

B'2 Function handles

Πολλές φορές στο MATLAB χρειάζεται να χρησιμοποιήσουμε μια συνάρτηση ως όρισμα για μια άλλη συνάρτηση. Αυτό μπορεί να γίνει μέσω ενός μηχανισμού που λέγεται `function`

handle. Ένα function handle είναι ένας τύπος δεδομένων του MATLAB που περιέχει όλες τις απαραίτητες πληροφορίες για τον υπολογισμό μιας συνάρτησης. Δημιουργείται βάζοντας το χαρακτήρα @ μπροστά από το όνομα μιας συνάρτησης.

Για παράδειγμα, ας δούμε πώς χρησιμοποιείται η συνάρτηση `ezplot` η οποία σχεδιάζει το γράφημα μιας συνάρτησης $f(x)$ έχοντας ως προκαθορισμένο πεδίο ορισμού το διάστημα $[-2\pi, 2\pi]$. Η `ezplot` δέχεται ως όρισμα το όνομα της συνάρτησης, έστω `fun`, που μπορεί να είναι είτε μια εσωτερική συνάρτηση του MATLAB ή το όνομα μιας συνάρτησης που ορίζεται μέσω ενός M-αρχείου, άρα γράφουμε `ezplot(@fun)`. Π.χ., για να σχεδιάσουμε τη γραφική παράσταση της $\sin(x)$ στο διάστημα $[-2\pi, 2\pi]$ γράφουμε

```
ezplot(@sin)
```

ενώ η εντολή

```
ezplot(@(x) myfun(x,20),[0.8 1.4])
```

σχεδιάζει τη γραφική παράσταση της συνάρτησης `myfun`, που είδαμε στην προηγούμενη παράγραφο, στο διάστημα $[0.8, 1.4]$, για $M = 20$.

B'3 Ανώνυμες συναρτήσεις

Οι ανώνυμες συναρτήσεις (anonymous functions) είναι ένας απλός τύπος συναρτήσεων στο MATLAB οι οποίες ορίζονται είτε στο παράθυρο εντολών ή σ' ένα αρχείο script. Μια ανώνυμη συνάρτηση είναι διαθέσιμη όσο διαρκεί η εργασία μας στο παράθυρο εντολών.

Π.χ., έστω ότι θέλουμε να ορίσουμε τη συνάρτηση $g(x) = \cos(\pi x)$ ως ανώνυμη συνάρτηση. Αυτό γίνεται με την εντολή:

```
>> g = @(x) cos(pi*x)
```

όπου

- το σύμβολο @ δηλώνει ότι η `g` είναι συνάρτηση,
- μετά το σύμβολο @ ακολουθεί η μεταβλητή εισόδου (όρισμα) της συνάρτησης,
- και, τέλος, ορίζεται ο τύπος της συνάρτησης.

Το όνομα της συνάρτησης εμφανίζεται στο παράθυρο των μεταβλητών με τον χαρακτηρισμό `function_handle`. Αν καθαρίσουμε το χώρο εργασίας (με την εντολή `clear`) η ανώνυμη συνάρτηση θα διαγραφεί.

Π.χ., με τις ακόλουθες εντολές μπορούμε να σχεδιάσουμε εύκολα τη γραφική παράσταση της $g(x) = \cos(\pi x)$ στο διάστημα $[-2, 2]$.

```
>> g = @(x) cos(pi*x)
>> ezplot(g, [-2 2])
```

Παρατηρήστε ότι σε αυτή την περίπτωση δεν χρειάζεται να βάλουμε το σύμβολο `@` πριν από την `g` στην `ezplot`.

Εναλλακτικά, θα μπορούσαμε να χρησιμοποιήσουμε την εντολή:

```
>> ezplot(@(x) cos(pi*x), [-2 2])
```

Μια άλλη συνάρτηση που χρησιμοποιείται για το σχεδιασμό γραφικών παραστάσεων συναρτήσεων είναι η `fplot`. Η `fplot` έχει δύο ορίσματα εισόδου: μια συνάρτηση και το διάστημα στο οποίο θα σχεδιάσει το γράφημά της. Π.χ.,

```
>> fplot(@sin, [0 3*pi])
>> fplot(@(x) cos(pi*x), [0 3])
>> fplot(@(x) myfun(x,200), [0.8 1.8])
```


Βιβλιογραφία

- [1] Γ.Δ. Ακρίβης, Β.Α. Δουγαλής, *Εισαγωγή στην Αριθμητική Ανάλυση*, Πανεπιστημιακές Εκδόσεις Κρήτης, Ηράκλειο, δ' έκδοση, 2010, δ' ανατύπωση, 2015.
- [2] Γ.Δ. Ακρίβης, Β.Α. Δουγαλής, *Αριθμητικές Μέθοδοι για Συνήθεις Διαφορικές Εξισώσεις*, Πανεπιστημιακές Εκδόσεις Κρήτης, Ηράκλειο, 2015.
- [3] R.L. Burden, J.L. Faires, *Numerical Analysis*, 9th ed., Brooks/Cole, 2011.
- [4] G. Dahlquist, Å. Björk, *Numerical Methods*, Prentice-Hall, 1974.
- [5] W. Cheney, D. Kincaid, *Numerical Mathematics and Computing*, 6th ed., Brooks/Cole, 2008.
- [6] G.E. Forsythe, M.A. Malcolm, C.B. Moler, *Computer Methods for Mathematical Computations*, Prentice-Hall, 1977.
(Ελληνική μετάφραση: *Αριθμητικές Μέθοδοι και Προγράμματα για Μαθηματικούς Υπολογισμούς*, (Μτφρ. Γ.Δ. Ακρίβης, Β.Α. Δουγαλής), γ' έκδοση, Πανεπιστημιακές Εκδόσεις Κρήτης, Ηράκλειο, 1998.)
- [7] D.J. Higham, N.J. Higham, *MATLAB Guide*, 2nd ed., SIAM, 2005.
- [8] N.J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 2002.
- [9] D. Kincaid, W. Cheney, *Numerical Analysis: Mathematics of Scientific Computing*, Brooks/Cole, 1991.
- [10] C.B. Moler, *Numerical Computing with Matlab*, SIAM 2004.
(Ελληνική μετάφραση: *Αριθμητικές μέθοδοι με το MATLAB*, (Μτφρ. Β. Μεταφτσής), Κλειδάριθμος, 2010.)
- [11] M. Overton, *Numerical Computing with IEEE Floating Point Arithmetic*, SIAM, Philadelphia, 2001.

- [12] E. Süli, D. Mayers, *An Introduction to Numerical Analysis*, Cambridge University Press, 2006.