Τμημα Επιστημης και Τεχνολογιας Υλικων
Πανεπιστημιο Κρητης

Εισαγωγή στην Αριθμητική Ανάλυση

Σημειώσεις Διαλέξεων και Εργαστηρίων

Ηράκλειο
2021

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

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

Η στοιχειοθεσία έγινε από τον Σ. Σταματιάδη με τη χρήση του XeLaTeX.

Τελευταία τροποποίηση του κειμένου έγινε στις 17 Μαΐου 2021. Η πιο πρόσφατη έκδοση βρίσκεται στο http://www.materials.uoc.gr/el/undergrad/courses/ETY213/notes.pdf

Περιεχόμενα

Κεφάλαιο 1

Σφάλματα

1.1 Εισαγωγή

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

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

1.2 Συστήματα αρίθμησης

1.2.1 Αναπαράσταση ακεραίων

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

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

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

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

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

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

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

ενώ τα επόμενα

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

Η τελευταία τιμή του i είναι αυτή για την οποία ισχύει

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

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

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

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

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

όπου e_i το κρατούμενο για το ψηφίο i. Το e_i ικανοποιεί τη σχέση

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

1.2.2 Αναπαράσταση πραγματικών

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

$$123.4567$$

Ο παραπάνω είναι ισοδύναμος με τους

$$12.34567 \times 10^1 \ , \quad 1.234567 \times 10^2 \ , \quad 0.1234567 \times 10^3 \ , \quad \text{κλπ.}$$

και τους

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

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

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

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

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

Τα ψηφία d_0, d_i, ..., d_n αποτελούν τα *σημαντικά ψηφία (significant digits)* του αριθμού.

1.3 Αναπαράσταση αριθμών στον υπολογιστή

1.3.1 Ακέραιοι

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

$$00000000\ 00000000\ 00000110\ 00100001$$

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

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

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

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

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

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

Ο μεγαλύτερος ακέραιος σε 32 bits είναι ο

$$01111111\ 11111111\ 11111111\ 11111111$$

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

$$10000000\ 00000000\ 00000000\ 00000000$$

δηλαδή ο -2147483648. Προσέξτε ότι ο αντίθετος του συγκεκριμένου αριθμού αναπαρίσταται με την ίδια σειρά. Επίσης, προσέξτε ότι ο αριθμός

$$11111111\ 11111111\ 11111111\ 11111111$$

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

$$00000000\ 00000000\ 00000000\ 00000001$$

δηλαδή, είναι ο -1.

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

1.3.2 Πραγματικοί

Στο δυαδικό σύστημα ο πραγματικός αριθμός έχει τη μορφή

$$\pm 1.d_1 d_2 d_3 \ldots d_n \times 2^e$$

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

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

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

Συγκεκριμένες σειρές των 32 ή 64 bits αντιστοιχούν στο ±infinity και στο NaN (Not A Number): σε αυτές τα 8 (σε απλή ακρίβεια) ή 11 bits (σε διπλή ακρίβεια) για τον εκθέτη είναι όλα 1. Αν τις εξαιρέσουμε, η σειρά των bits που μπορεί να αποθηκευτεί σε απλή ακρίβεια στο τμήμα του εκθέτη, αντιστοιχεί στο δυαδικό αριθμό

$$11111110\ .$$

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

$$00000001$$

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

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

$$111\ 11111110\ .$$

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

$$000\ 00000001$$

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

Παρατηρήσεις:

- Αν ένα πραγματικό αποτέλεσμα πράξης υπερβαίνει κατ᾽ απόλυτη τιμή το μέγιστο αναπαραστάσιμο στον Η/Υ αριθμό, έχουμε *υπερχείλιση (overflow)*. Αντίστοιχα, αν είναι κατ᾽ απόλυτη τιμή μικρότερο από το μικρότερο αναπαραστάσιμο στον Η/Υ αριθμό, τότε έχουμε *υπεκχείλιση (underflow)*. Η τιμή που θα αποκτήσει το αποτέλεσμα και στις δύο περιπτώσεις είναι απροσδιόριστη, ο υπολογισμός όμως μπορεί να συνεχίσει με σχεδόν σίγουρα λάθος αποτέλεσμα. Σε υπολογιστές που υλοποιούν το πρότυπο αναπαράστασης αριθμών IEEE οι τιμές είναι αντίστοιχα $\pm$infinity (το πλησιέστερο «άπειρο») και ± 0.

- Λόγω της πεπερασμένης ακρίβειας αναπαράστασης, το αποτέλεσμα της πράξης μεταξύ πραγματικών $x + (y + z)$ μπορεί να είναι διαφορετικό από το $(x + y) + z$, π.χ. αν το x είναι πολύ μεγαλύτερο κατά μέτρο από τα y, z. Επίσης, η τιμή $x + y + z$ μπορεί να είναι διαφορετική από την $x + z + y$. *Η πρόσθεση (και ο πολλαπλασιασμός) μεταξύ πραγματικών αριθμών στον υπολογιστή δεν έχει την αντιμεταθετική και την προσεταιριστική ιδιότητα, όπως αναμένουμε από τα μαθηματικά.*

- Η σύγκριση για ισότητα δύο πραγματικών αριθμών που προέκυψαν από πράξεις πρέπει να αποφεύγεται.

1.4 Ασκήσεις

1. Να γράψετε κώδικα που να μετατρέπει ένα μη αρνητικό ακέραιο αριθμό από το δεκαδικό στο δυαδικό σύστημα.

2. Γράψτε πρόγραμμα που να τυπώνει στην οθόνη στη μορφή $\pm x.xxxxE \pm yy$ (εκθετική μορφή) τα αποτελέσματα των εκφράσεων $0.1 + 0.2 - 0.3$ και $0.1 - 0.3 + 0.2$. Είναι μηδέν; είναι έστω ίσα;

3. Υπολογίστε με πρόγραμμα το άθροισμα των αριθμών 0.1, 0.2,...,1.9.

   [Απάντηση: 19]

4. Υπολογίστε το έψιλον της μηχανής για πραγματικούς αριθμούς απλής και διπλής ακρίβειας με τους εξής τρόπους:

   (α΄) Εφαρμόστε τον αλγόριθμο:
   Θέτουμε $\varepsilon \leftarrow 1$. Για όσο ισχύει $1 + \varepsilon \neq 1$ θέτουμε $\varepsilon \leftarrow \varepsilon/2$ και επαναλαμβάνουμε.

   (β΄) Χρησιμοποιήστε κατάλληλες συναρτήσεις/τιμές που παρέχει η γλώσσα προγραμματισμού που χρησιμοποιείτε (FLT_EPSILON και DBL_EPSILON στη C, EPSILON() στη Fortran 90, epsilon() από το std::numeric_limits<> της C++, κλπ.).

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

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

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

όταν $a \neq 0$.

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

(α′) Υπολογίστε τα $x_{1,2}$ με απλή και διπλή ακρίβεια. Συγκρίνετέ τα με τις ακριβείς ρίζες ($x1 = -0.001$, $x2 = -3000.0$).

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

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

Τι παρατηρείτε ως προς την ακρίβεια των υπολογισμών σας;

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Κεφάλαιο 2

Επίλυση μη Γραμμικών Εξισώσεων

2.1 Εισαγωγή

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

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

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

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

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

Η επίλυση με αριθμητικές μεθόδους της εξίσωσης (2.1) βασίζεται στην εύρεση μιας ακολουθίας τιμών $x_0, x_1, \ldots, x_n$, που (αν υπάρχει ρίζα) συγκλίνει για $n \to \infty$ σε μία ρίζα της εξίσωσης, $\bar{x}$. Κάθε μία από τις μεθόδους που θα δούμε, παράγει τέτοια ακολουθία με συγκεκριμένη διαδικασία και υπό ορισμένες προϋποθέσεις. Επιπλέον, σε κάθε επανάληψη, μας δίνει μια εκτίμηση του εύρους της περιοχής στην οποία βρίσκεται η ρίζα γύρω από το x_i: η μέθοδος παράγει μία ακολουθία ε_1,

ε_2,..., ε_n για τη μέγιστη ακρίβεια· για κάθε $i = 0, 1, 2, \ldots$ ισχύει $x_i - \varepsilon_i \leq \bar{x} \leq x_i + \varepsilon_i$, με $\varepsilon_i < \varepsilon_{i-1}$.

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

- Το μέγιστο σφάλμα της μεθόδου, ε_k, είναι μικρότερο από το επιθυμητό.
- Η απόλυτη τιμή της συνάρτησης είναι «μικρή»: $|f(x_k)| < \varepsilon$.
- Η σχετική βελτίωση στην προσεγγιστική τιμή είναι «μικρή»: $\left|\dfrac{x_k - x_{k-1}}{x_k}\right| < \varepsilon$ αν $x_k \neq 0$.
- Η απόλυτη βελτίωση στην προσεγγιστική τιμή είναι «μικρή»: $|x_k - x_{k-1}| < \varepsilon$, αν $x_k \approx 0$.

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

2.1.1 Ταχύτητα και τάξη σύγκλισης

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

$$\lim_{n\to\infty} \frac{|x_{n+1} - \bar{x}|}{|x_n - \bar{x}|^\alpha} \equiv \lim_{n\to\infty} \frac{\varepsilon_{n+1}}{{\varepsilon_n}^\alpha} = \lambda \,. \tag{2.2}$$

Ο αριθμός λ αποτελεί την *ταχύτητα σύγκλισης*.

2.1.2 Ευστάθεια

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

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

2.1.3 Εύρεση περισσότερων της μίας ριζών

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

2.1.4 Χρήσιμα θεωρήματα

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

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

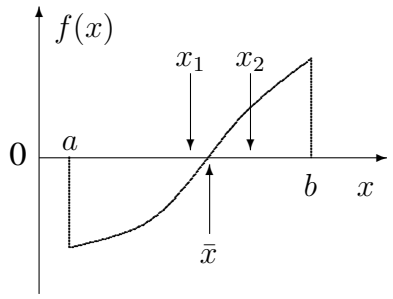
Θεώρημα Taylor Έστω ότι η συνάρτηση $f(x)$, με $x \in [a,b]$, έχει παραγώγους μέχρι τάξης $n+1$ και η $f^{(n+1)}(x)$ είναι συνεχής στο $[a,b]$. Αν $x, x_0 \in [a,b]$, $x \neq x_0$, τότε υπάρχει ξ μεταξύ των x_0, x ώστε

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

2.2 Μέθοδος Διχοτόμησης

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

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



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

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

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

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

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

Παράδειγμα

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

Μετά από 20 επαναλήψεις ισχύει για την ακρίβεια $|x_{20} - \bar{x}| \leq 0.5\,|b_{20} - a_{20}| \approx$

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

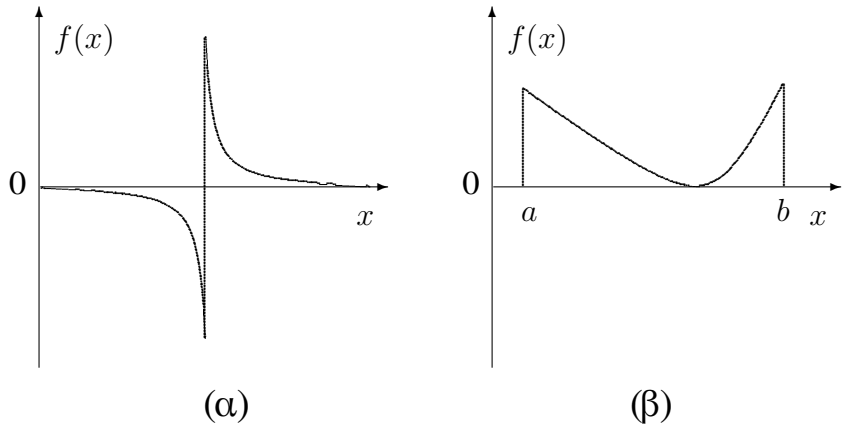
Πίνακας 2.1: Ακολουθίες των διαστημάτων, της προσεγγιστικής ρίζας και της αντίστοιχης τιμής της $f(x) = x^3 + 4x^2 - 10$ κατά την εφαρμογή της μεθόδου διχοτόμησης

0.95×10^{-6}, άρα έχουμε προσδιορίσει σωστά τουλάχιστον μέχρι και το 6 δεκαδικό ψηφίο της ρίζας. Η προσεγγιστική τιμή, στρογγυλεμένη στα 6 δεκαδικά είναι 1.365230 ενώ η ακριβής είναι 1.36523001361638....

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

2.2.1 Ακρίβεια αλγορίθμου διχοτόμησης

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



(α) (β)

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

Απόδειξη:

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

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

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

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

2.2.2 Σύγκλιση αλγορίθμου διχοτόμησης

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

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

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

2.2.3 Αριθμός επαναλήψεων αλγορίθμου διχοτόμησης

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

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

Παράδειγμα

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

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

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

2.3 Μέθοδος ψευδούς σημείου

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

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

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

Επομένως,

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

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

Η μέθοδος ψευδούς σημείου είναι γενικά πιο γρήγορη στην εύρεση ρίζας από τη μέθοδο διχοτόμησης. Όμως, αν κάποιο από τα άκρα του διαστήματος $[a,b]$ δεν μετακινείται σε διαδοχικές επαναλήψεις της μεθόδου ψευδούς σημείου, έχουμε αργή σύγκλιση. Αυτό συμβαίνει σχεδόν πάντα μετά από πολλές επαναλήψεις της.

2.3.1 Μέθοδος Illinois

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

- αν άλλαξε δύο συνεχόμενες φορές το όριο a

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

- αν άλλαξε δύο συνεχόμενες φορές το όριο b

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

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

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

2.4 Μέθοδος τέμνουσας

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

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

Επομένως,

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

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

Παρατηρήστε ότι η μέθοδος τέμνουσας μοιάζει πολύ με τη μέθοδο ψευδούς σημείου (§2.3). Όμως, στη μέθοδο τέμνουσας η ρίζα *δεν είναι απαραίτητα περιορισμένη μεταξύ δύο σημείων*. Αυτό, σε αντίθεση με τη μέθοδο ψευδούς σημείου, μπορεί να οδηγήσει σε μεγάλη απομάκρυνση από τη ρίζα.

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

1. Επιλέγουμε δύο τιμές a, b.
2.
3. Ορίζουμε ως c το
$$c = \frac{bf(a) - af(b)}{f(a) - f(b)} .$$
4. Ελέγχουμε τα κριτήρια σύγκλισης. Αν το c είναι ικανοποιητική προσέγγιση της ρίζας πηγαίνουμε στο βήμα 7.
5. Θέτουμε $a \leftarrow b$, $b \leftarrow c$.
6. Επαναλαμβάνουμε τη διαδικασία από το βήμα 3.
7. Τέλος.

2.4.1 Σύγκλιση της μεθόδου τέμνουσας

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

2.5 Μέθοδος Müller

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

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

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

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

Η μορφή της παραβολής που επιλέχθηκε διευκολύνει τη λύση του συστήματος $y(x_i) = f(x_i)$ με $i = 0, 1, 2$.

Από τις δύο ρίζες της παραβολής,[1]

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

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

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

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

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

2. Ορίζουμε τις ποσότητες

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

3. Η επόμενη προσέγγιση της ρίζας δίνεται από τη σχέση

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

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

4. Αν η νέα προσέγγιση είναι ικανοποιητική, πηγαίνουμε στο βήμα 6.

5. Θέτουμε $x_0 \leftarrow x_1$, $x_1 \leftarrow x_2$, $x_2 \leftarrow x_3$. Επαναλαμβάνουμε τη διαδικασία από το βήμα 2.

6. Τέλος.

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

[1] χρησιμοποιούμε άλλο τύπο από το συνήθη για μεγαλύτερη ακρίβεια.

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

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

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

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

Αν η ακολουθία συγκλίνει σε ένα σημείο $\bar{x}$ και καθώς η $g(x)$ είναι συνεχής[2] έχουμε

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

Άρα

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

2.6.1 Ορισμός–Σχετικά Θεωρήματα

Ορισμός

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

Κριτήριο ύπαρξης σταθερού σημείου

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

Απόδειξη: Ισχύει $g(a) \geq a$, $g(b) \leq b$. Ορίζουμε τη συνεχή συνάρτηση $h(x) = g(x) - x$. Τότε $h(a) \geq 0$, $h(b) \leq 0$. Το Θεώρημα Bolzano εξασφαλίζει ότι υπάρχει τουλάχιστον ένα $\varrho \in [a, b]$ ώστε $h(\varrho) = 0$.

[2] ορισμός συνέχειας της $g(x)$: $\lim_{n\to\infty} g(x_n) = g(\lim_{n\to\infty} x_n)$.

Παράδειγμα

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

Μοναδικότητα σταθερού σημείου

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

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

Παράδειγμα

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

2.6.2 Σύγκλιση της μεθόδου σταθερού σημείου

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

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

Παραδείγματα

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

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

Παρατηρούμε ότι $|g'(x)| < 1$ όταν $-3 < x < 3$. Για αυτά τα x, $5/6 < g(x) < 14/6$, άρα $g(x) \in (-3, 3)$. Επομένως, υπάρχει μοναδικό σταθερό σημείο στο $(-3, 3)$. Οποιαδήποτε αρχική τιμή $|x| < 3$ (αλλά όχι μόνο) δίνει ακολουθία

που συγκλίνει σε αυτό. Ας το εντοπίσουμε: για $x_0 = 2.5$ έχουμε

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

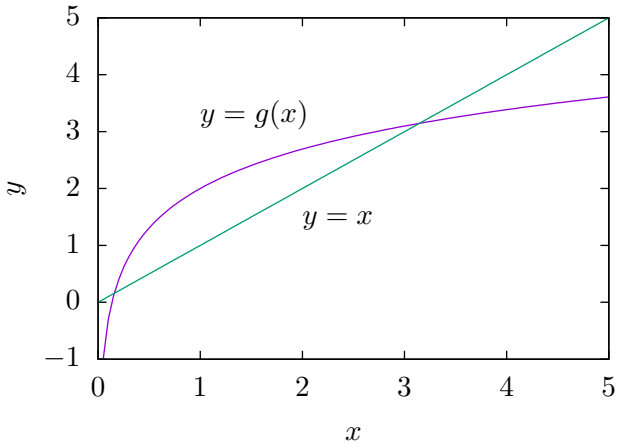
Αν δοκιμάσουμε άλλο αρχικό σημείο στο $(-3, 3)$ θα έχουμε πάλι σύγκλιση στο 1. Άλλες αρχικές τιμές μεγαλύτερες του 5 κατ' απόλυτη τιμή προκαλούν απόκλιση της ακολουθίας στο $+\infty$. Κανένα σημείο εκτός από τα $x_0 = \pm 5.0$ δε δίνει ακολουθία με όριο την άλλη ρίζα.

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

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

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

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



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

2.7 Μέθοδοι Householder

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

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

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

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

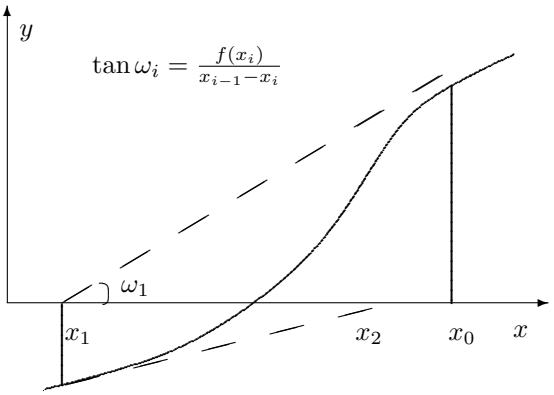
2.7.1 Μέθοδος Newton–Raphson

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

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

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

όπου ξ μεταξύ $\bar{x}, x_0$. Αγνοώντας τον όρο του υπολοίπου, θεωρώντας ότι η από-



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

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

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

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

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \, , \qquad n = 0, 1, 2, \ldots \, . \tag{2.6}$$

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

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

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

Παράδειγμα

Έστω $f(x) = x^2 - 6x + 5$. Έχουμε

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

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

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

Σύγκλιση αλγορίθμου Newton–Raphson

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

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

Λαμβάνοντας υπόψη τη σχέση (2.5) έχουμε

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

Επομένως

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

με ξ μεταξύ των x_n και $\bar{x}$.

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

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

Μέθοδοι Newton–Raphson για πολλαπλές ρίζες

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

Αν γνωρίζουμε την πολλαπλότητα, μπορούμε να αναζητήσουμε τη ρίζα της $f^{(m-1)}(x) = 0$ καθώς σε αυτή το $\bar{x}$ είναι απλή ρίζα. Εναλλακτικά, παρατηρήστε ότι η συνάρτηση $f(x)$ με ρίζα το $\bar{x}$, πολλαπλότητας m, μπορεί να γραφεί στη μορφή $f(x) = (x-\bar{x})^m g(x)$, όπου $g(x)$ συνάρτηση για την οποία το $\bar{x}$ δεν είναι ρίζα. Συνεπώς, η συνάρτηση $h_1(x) = \sqrt[m]{f(x)}$ έχει *απλή* ρίζα το $\bar{x}$. Ο τύπος Newton–Raphson, (2.6), για αυτή τη συνάρτηση αναμένουμε να έχει τετραγωνική σύγκλιση. Η εφαρμογή του δίνει

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

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

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

2.7.2 Μέθοδος Halley

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

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

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

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

Μπορεί να δειχθεί ότι η μέθοδος είναι *τρίτης τάξης* με ταχύτητα σύγκλισης

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

2.8 Ασκήσεις

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

- $f(x) = x^3 + 4x^2 - 10$ στο διάστημα $[1, 2]$,
- $f(x) = \sqrt{x} - \cos x$ στο διάστημα $[0, 1]$.

2. (α΄) Γράψτε ένα πρόγραμμα το οποίο να υλοποιεί τη μέθοδο ψευδούς σημείου.
   (β΄) Εφαρμόστε την για να βρείτε τη ρίζα της
   $$f(x) = -2.0 + 6.2x - 4.0x^2 + 0.7x^3$$
   στο διάστημα $[0.4, 0.6]$.
   (γ΄) Εφαρμόστε τη μέθοδο ψευδούς σημείου και τη μέθοδο διχοτόμησης για να βρείτε τις ρίζες της
   $$f(x) = x^{10} - 0.95$$
   στο διάστημα $[0, 1.4]$. Ποια μέθοδος συγκλίνει πιο γρήγορα με σχετικό σφάλμα $< 10^{-6}$;

3. Βρείτε τη ρίζα της $f(x) = x^2 - (1-x)^5$ στο $[0, 1]$ με ακρίβεια 10^{-9}, εφαρμόζοντας τη μέθοδο διχοτόμησης, τη μέθοδο ψευδούς σημείου και τη τροποποιημένη μέθοδο ψευδούς σημείου (αλγόριθμος Illinois). Πόσες επαναλήψεις και πόσους υπολογισμούς της συνάρτησης χρειαστήκατε σε κάθε μέθοδο;

4. Χρησιμοποιήστε τη μέθοδο τέμνουσας για να βρείτε τη ρίζα της εξίσωσης $g(x) = 3\ln x + 5$ με ακρίβεια 6 σημαντικών ψηφίων.

5. Δείξτε ότι η $g(x) = \ln x + 2$ έχει ένα και μοναδικό σταθερό σημείο στο $[2, 4]$. Υπολογίστε το μέγιστο αριθμό επαναλήψεων ώστε $|x_n - \bar{x}| \leq 10^{-3}$.

6. Γράψτε κώδικα που να υλοποιεί τη γενική επαναληπτική μέθοδο $x = g(x)$. Χρησιμοποιήστε τον για να υπολογίσετε

   - μια ρίζα της $f(x) = x^2 - 6x + 5$,
   - τη ρίζα της $f(x) = x - \cos^3 x$ κοντά στο 0.6.

7. Υπολογίστε το $y = \frac{e^x - 1}{x}$ με ένα ευσταθή αλγόριθμο για μικρό, κατ' απόλυτη τιμή, x. Για μικρό $|x|$ χρησιμοποιούμε το ανάπτυγμα Taylor του e^x ώστε να αποφύγουμε την αλληλοαναίρεση όρων ίδιας τάξης.

8. Υπολογίστε με ευσταθή αλγόριθμο τις λύσεις των εξισώσεων

   (α΄) $1.5x^2 + 13 \times 10^6 x + 0.037 = 0$.
   Οι ακριβείς είναι $x_1 \approx -2.8462 \times 10^{-9}$, $x_2 \approx -8.6667 \times 10^6$.
   (β΄) $1.5x^2 - 37 \times 10^6 x + 0.057 = 0$.
   Οι ακριβείς είναι $x_1 \approx 1.5405 \times 10^{-9}$, $x_2 \approx 2.4667 \times 10^7$.

9. Εφαρμόστε τη μέθοδο Newton–Raphson για να υπολογίσετε τις ρίζες της

   (α΄) $f(x) = \sin x - x^2$,

(β′) $f(x) = 3x\mathrm{e}^x - 1$.

10. Υπολογίστε τις ρίζες της $f(x) = 4\cos x - \mathrm{e}^{-x}$ με ακρίβεια 10^{-8} με τη μέθοδο διχοτόμησης, τη μέθοδο σταθερού σημείου, τη μέθοδο Newton–Raphson και τη μέθοδο τέμνουσας.

11. Βρείτε με 12 ψηφία σωστά το σημείο τομής των καμπυλών e^x, $\tan(2x)$ στο διάστημα $[-1, 1]$. *Συμβουλή:* σχεδιάστε τις καμπύλες.

12. Υλοποιήστε σε κώδικα τον αλγόριθμο Müller. Εφαρμόστε τον για να βρείτε τη μη μηδενική ρίζα της $f(x) = \sin x - x^2$.

13. Υλοποιήστε σε κώδικα τη μέθοδο Newton–Raphson, κατάλληλα τροποποιημένη ώστε να υπολογίζει τις ρίζες πολυωνύμου βαθμού n, $p_n(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_n x^n$, όταν έχουμε ως δεδομένους τους συντελεστές του α_0, α_1, ..., α_n. Το πολυώνυμο και η παράγωγός του να υπολογίζονται με τον αλγόριθμο Horner.

14. Υλοποιήστε σε κώδικα τη μέθοδο τέμνουσας, κατάλληλα τροποποιημένη ώστε να υπολογίζει τις ρίζες πολυωνύμου βαθμού n, $p_n(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_n x^n$, όταν έχουμε ως δεδομένους τους συντελεστές του α_0, α_1, ..., α_n. Το πολυώνυμο και η παράγωγός του να υπολογίζονται με τον αλγόριθμο Horner.

Κεφάλαιο 3

Επίλυση Γραμμικών Συστημάτων και εφαρμογές

3.1 Εισαγωγή

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

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 && \text{(3.1α′)} \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 && \text{(3.1β′)} \\ \vdots \qquad & \vdots \qquad \quad \vdots \qquad & \vdots & \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n\ . && \text{(3.1γ′)} \end{aligned}$$

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

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

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

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

3.1.1 Ευστάθεια γραμμικών συστημάτων

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

Παράδειγμα

$$\begin{bmatrix} 1 & 3 \\ 1 & 3.001 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 4.001 \end{bmatrix}$$

έχει λύση $x_1 = x_2 = 1$.

Το ελαφρά διαφορετικό σύστημα

$$\begin{bmatrix} 1 & 3 \\ 1 & 2.999 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 4.002 \end{bmatrix}$$

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

Ο *δείκτης κατάστασης*, κ, του πίνακα $\boldsymbol{A}$ ως προς τη νόρμα $||\cdot||$ ορίζεται ως

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

Μια νόρμα που μπορούμε να χρησιμοποιήσουμε είναι η «νόρμα αθροίσματος γραμμών»

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

Ο δείκτης κατάστασης είναι ιδιότητα του πίνακα. Ισχύει πάντα $\kappa \geq 1$. Αν $\kappa \gg 1$ το σύστημα είναι ασταθές. Αν ο δείκτης κατάστασης είναι πολύ κοντά στο 1, τότε η λύση που θα υπολογίσουμε, με οποιαδήποτε μέθοδο, θα έχει ακρίβεια παρόμοια με την ακρίβεια των στοιχείων του πίνακα.

3.1.2 Ορισμοί–Βασικές γνώσεις

Επίλυση γραμμικής εξίσωσης μίας μεταβλητής

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

- Αν $a \neq 0$ η εξίσωση έχει μία λύση, την $x = b/a$.
- Αν $a = 0$ εξετάζουμε το b:
  - § Αν $b \neq 0$ η εξίσωση δεν έχει λύση.
  - § Αν $b = 0$ η εξίσωση έχει άπειρες λύσεις (οποιοδήποτε x ικανοποιεί την $0x = 0$).

Στη διαδικασία επίλυσης ενός γραμμικού συστήματος με διάφορους αλγόριθμους, θα χρειαστεί να λύσουμε πρωτοβάθμιες εξισώσεις. Αυτές θα καθορίσουν τη λύση του συστήματος ανάλογα με τις τιμές των συντελεστών τους.

Ιδιοτιμές–Ιδιοδιανύσματα

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

Αν υπάρχει ένας αριθμός λ, εν γένει *μιγαδικός*, και ένα διάνυσμα (πίνακας–στήλη) $\boldsymbol{x}$, διάφορο του $[0, 0, \ldots, 0]^T$ για τα οποία ισχύει

$$\boldsymbol{A} \cdot \boldsymbol{x} = \lambda \boldsymbol{x} \ , \tag{3.3}$$

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

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

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

Ορίζουσα

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

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

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

Συμμετρικός θετικά ορισμένος πίνακας

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

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

- Ισχύει $\boldsymbol{x}^T \cdot \boldsymbol{A} \cdot \boldsymbol{x} > 0$ για κάθε πραγματικό μη μηδενικό διάνυσμα $\boldsymbol{x}$.

- Όλες οι ιδιοτιμές του είναι πραγματικές και θετικές.
- Στην ανάλυση LU του $\boldsymbol{A}$ (§3.2.5), ο πίνακας $\boldsymbol{L}$ έχει θετικά διαγώνια στοιχεία (θεωρούμε ότι κάθε στοιχείο της διαγωνίου του $\boldsymbol{U}$ έχει τιμή 1).
- Υπάρχει πραγματικός αντιστρέψιμος πίνακας B για τον οποίο ισχύει $\boldsymbol{A} = \boldsymbol{B}^T \cdot \boldsymbol{B}$.
- Υπάρχει ένας και μοναδικός πραγματικός κάτω τριγωνικός πίνακας $\boldsymbol{L}$ (ή άνω τριγωνικός πίνακας $\boldsymbol{U}$) με θετικά διαγώνια στοιχεία για τον οποίο ισχύει $\boldsymbol{A} = \boldsymbol{L} \cdot \boldsymbol{L}^T$ (ή $\boldsymbol{A} = \boldsymbol{U}^T \cdot \boldsymbol{U}$) (*ανάλυση Cholesky*).
- Είναι θετικές οι ορίζουσες (§3.1.2) όλων των τετραγωνικών υπο-πινάκων του $\boldsymbol{A}$ με πάνω αριστερό στοιχείο το a_{11} και κάτω δεξιό το a_{ii}, $i = 1, 2, \ldots, n$ (*κριτήριο του Sylvester*[1]).

Μπορεί να δειχθεί ότι για ένα πραγματικό, συμμετρικό, θετικά ορισμένο πίνακα $\boldsymbol{A}$ ισχύουν τα εξής

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

Συνθήκες επιλυσιμότητας

Οι παρακάτω συνθήκες είναι ισοδύναμες:

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

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

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

3.2 Απευθείας μέθοδοι

Οι απευθείας μέθοδοι επίλυσης γραμμικών συστημάτων δίνουν την ακριβή λύση (με κάποιο σφάλμα στρογγύλευσης) σε συγκεκριμένο και εκ των προτέρων υπολογίσιμο αριθμό βημάτων/πράξεων.

3.2.1 Μέθοδος αντίστροφου πίνακα

Μια εύκολη στην αντίληψη αλλά χρονοβόρα στην υλοποίηση μέθοδος απαιτεί τον υπολογισμό του αντίστροφου πίνακα του $\boldsymbol{A}$ (αρκεί αυτός να υπάρχει) ώστε η λύση να είναι $\boldsymbol{x} = \boldsymbol{A}^{-1} \cdot \boldsymbol{b}$. Αν και θα παρουσιάσουμε μέθοδο εύρεσης του αντίστροφου πίνακα δεν θα τη χρησιμοποιούμε για επίλυση συστήματος καθώς υπάρχουν πιο γρήγορες μέθοδοι.

3.2.2 Κανόνας Cramer

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

$$x_j = \frac{\det(\boldsymbol{B}_j)}{\det(\boldsymbol{A})} \,, \quad j = 1, 2, \ldots, n \,,$$

όπου ο πίνακας $\boldsymbol{B}_j$ προκύπτει από τον $\boldsymbol{A}$ αν αντικαταστήσουμε την στήλη j του $\boldsymbol{A}$ με το διάνυσμα $\boldsymbol{b}$.

Ας αποδείξουμε την πρώτη από τις σχέσεις. Θυμηθείτε ότι η ορίζουσα πίνακα που μια στήλη του γράφεται ως άθροισμα προσθετέων, ισούται με το άθροισμα των οριζουσών που προκύπτουν από την αρχική, η κάθε μία με ένα όρο στη στήλη. Έτσι, αν στην ορίζουσα του $\boldsymbol{B}_1$ αντικαταστήσουμε τα b_i με τις αριστερά μέλη των εξισώσεων (3.1) έχουμε

$$\begin{aligned}
\det(\boldsymbol{B}_1) &= \begin{vmatrix} b_1 & a_{12} & \cdots & a_{1n} \\ b_2 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_n & a_{n2} & \cdots & a_{nn} \end{vmatrix} = \begin{vmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n & a_{12} & \cdots & a_{1n} \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n & a_{n2} & \cdots & a_{nn} \end{vmatrix} \\
&= \begin{vmatrix} a_{11}x_1 & a_{12} & \cdots & a_{1n} \\ a_{21}x_1 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}x_1 & a_{n2} & \cdots & a_{nn} \end{vmatrix} + \begin{vmatrix} a_{12}x_2 & a_{12} & \cdots & a_{1n} \\ a_{22}x_2 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n2}x_2 & a_{n2} & \cdots & a_{nn} \end{vmatrix} \\
&\quad + \cdots + \begin{vmatrix} a_{1n}x_n & a_{12} & \cdots & a_{1n} \\ a_{2n}x_n & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{nn}x_n & a_{n2} & \cdots & a_{nn} \end{vmatrix}
\end{aligned}$$

Όλες οι ορίζουσες στις οποίες καταλήξαμε, εκτός από την πρώτη, είναι μηδέν καθώς έχουν δύο στήλες ανάλογες. Από την πρώτη μπορούμε να βγάλουμε κοινό

παράγοντα το x_1 οπότε απομένει η ορίζουσα του $\boldsymbol{A}$. Καταλήγουμε στη σχέση που θέλαμε να αποδείξουμε:

$$\det(\boldsymbol{B}_1) = x_1 \det(\boldsymbol{A}) \ ,$$

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

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

3.2.3 Απαλοιφή Gauss

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

Η μέθοδος της απαλοιφής Gauss αποτελείται από δύο στάδια:

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

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots\cdots\cdots\cdots + a_{1n}x_n &= b_1 && \text{(3.5α′)}\\
a_{22}x_2 + a_{23}x_3 + \cdots\cdots\cdots\cdots + a_{2n}x_n &= b_2 && \text{(3.5β′)}\\
a_{33}x_3 + \cdots\cdots\cdots\cdots + a_{3n}x_n &= b_3 && \text{(3.5γ′)}\\
\ddots \qquad\qquad \vdots \qquad &\ \ \vdots &&\\
a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n &= b_{n-1} && \text{(3.5δ′)}\\
a_{nn}x_n &= b_n \ . && \text{(3.5ε′)}
\end{aligned}$$

   Οι μετασχηματισμοί είναι τέτοιοι ώστε να διατηρούν τη λύση.

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

Τριγωνοποίηση

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

[2] http://dx.doi.org/10.1016/j.jda.2011.06.007

- Εναλλαγή της σειράς δύο εξισώσεων,
- Πρόσθεση σε μία εξίσωση μιας άλλης,
- Πολλαπλασιασμός μιας εξίσωσης με ένα μη μηδενικό αριθμό.

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

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

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

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

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

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

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

$$\lambda_i = -a_{i1}/a_{11} \tag{3.6α′}$$

$$a_{ij} \leftarrow a_{ij} + \lambda_i a_{1j} \,, \quad j = 1, 2, \ldots, n \tag{3.6β′}$$

$$b_i \leftarrow b_i + \lambda_i b_1 \,, \tag{3.6γ′}$$

για $i = 2, 3, \ldots, n$.

Το σύστημα (3.1) θα γίνει

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

Δεύτερη στήλη. Ας δούμε πώς μηδενίζουμε τα στοιχεία της δεύτερης στήλης, κάτω από τη διαγώνιο. Επιλέγουμε τη *δεύτερη* γραμμή και την προσθέτουμε σε κάθε *επόμενη*, πολλαπλασιασμένη με κατάλληλους αριθμούς. Επομένως

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

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

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

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

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

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

$$\lambda_i = -a_{i2}/a_{22} \tag{3.7α′}$$

$$a_{ij} \leftarrow a_{ij} + \lambda_i a_{2j} \,, \quad j = 2, 3, \ldots, n \tag{3.7β′}$$

$$b_i \leftarrow b_i + \lambda_i b_2 \,, \tag{3.7γ′}$$

για $i = 3, 4, \ldots, n$.

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

$$\lambda_i = -a_{ik}/a_{kk} \tag{3.8α′}$$

$$a_{ij} \leftarrow a_{ij} + \lambda_i a_{kj} \,, \quad j = k, k+1, \ldots, n \tag{3.8β′}$$

$$b_i \leftarrow b_i + \lambda_i b_k \,, \tag{3.8γ′}$$

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

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

Στη δεύτερη ομάδα εξισώσεων μπορούμε να παραλείψουμε την εκτέλεση για $j = k$ καθώς εκ κατασκευής θα μηδενίζουν απλά το a_{ik}.

Παρατήρηση: Στην περίπτωση που κάποιος συντελεστής a_{KK} είναι ή γίνει κατά την εφαρμογή του αλγορίθμου ίσος με 0, δεν μπορούμε να εφαρμόσουμε τις εξισώσεις (3.8) για την εξίσωση K ως έχει. Πρέπει να εναλλάξουμε την επίμαχη εξίσωση K με κάποια από τις *επόμενές* της ώστε να έρθει στη διαγώνιο ένας μη μηδενικός συντελεστής. Θα αναφέρουμε στην §3.2.3 πώς μπορούμε να επιλέξουμε την καταλληλότερη εξίσωση. Κατόπιν, μπορούμε να συνεχίσουμε τη διαδικασία.

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

Επίλυση άνω τριγωνικού συστήματος

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

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

Ο γενικός τύπος είναι

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

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

Παρατήρηση: Σύμφωνα με όσα αναφέραμε για τη λύση πρωτοβάθμιας εξίσωσης (§3.1.2), αν κάποιος συντελεστής a_{II} είναι 0, εξετάζουμε τον αριθμητή στη σχέση (3.9):

- αν
$$b_I - \sum_{j=I+1}^{n} a_{Ij} x_j = 0$$
το σύστημα έχει άπειρες λύσεις. Τα x_i με $i < I$ θα εκφράζονται ως συναρτήσεις του x_I, δεν θα μπορούν να πάρουν συγκεκριμένη αριθμητική τιμή. Το x_I θα είναι ελεύθερη ποσότητα που θα μπορεί να πάρει οποιαδήποτε τιμή θέλουμε.

- αν

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

το σύστημα δεν έχει λύση.

Παράδειγμα

Το σύστημα

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

επιλύεται ως εξής:

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

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

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

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

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

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

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

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

Παρατηρήσεις

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

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

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

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

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

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

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

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

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

Παρατηρήστε ότι οποιοδήποτε στοιχείο σε κάποια γραμμή μπορεί να γίνει όσο μεγάλο θέλουμε αν πολλαπλασιάσουμε την εξίσωση στην οποία ανήκει με κατάλληλο αριθμό. Γι᾽ αυτό, καλό είναι να λαμβάνουμε υπόψη τις σχετικές τιμές των συντελεστών ως προς το μεγαλύτερο συντελεστή της εξίσωσης στην οποία ανήκουν. Σε αυτή την παραλλαγή της μερικής οδήγησης, υπολογίζουμε κάθε φορά το μέγιστο στοιχείο των γραμμών με $i \geq k$, $M_i = \max_j |a_{ij}|$ με $j = k, \ldots, n$. Κατόπιν, διαιρούμε όλη την εξίσωση με M_i και συγκρίνουμε τα $|\bar{a}_{ik}| = |a_{ik}|/M_i$, με $i \geq k$, ώστε να βρούμε το μεγαλύτερο. Το υπολογιστικό κόστος αυξάνει αλλά ο αλγόριθμος γίνεται πιο ευσταθής.

Παράδειγμα: Το σύστημα

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

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

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

η απλή απαλοιφή Gauss δίνει προσεγγιστικά, μετά την τριγωνοποίηση,

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

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

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

η τριγωνοποίηση δίνει

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

Συνεπώς, $x_1 = 10$ και $x_2 = 1$.

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

3.2.4 Μέθοδος Gauss–Jordan

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

$$\boldsymbol{A} \cdot \boldsymbol{x} = \boldsymbol{B}$$

γίνεται

$$\boldsymbol{A'} \cdot \boldsymbol{x} = \boldsymbol{B'} \ ,$$

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

$$\boldsymbol{I} \cdot \boldsymbol{x} = \boldsymbol{B''} \ .$$

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

3.2.5 Ανάλυση LU

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

$$\boldsymbol{A} = \boldsymbol{L} \cdot \boldsymbol{U} \ ,$$

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

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

και

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

Η εξίσωση (3.2) μπορεί να γραφεί

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

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

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

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

Αφού προσδιορίσουμε το διάνυσμα y, η επίλυση του άνω τριγωνικού συστήματος $\boldsymbol{U}\cdot\boldsymbol{x}=\boldsymbol{y}$ σύμφωνα με την §3.2.3 θα μας δώσει τη λύση του αρχικού. Επομένως

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

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

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

Αλγόριθμος του Crout για τον προσδιορισμό των $\boldsymbol{L}$, $\boldsymbol{U}$

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

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

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

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

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

1. Ορίζουμε $u_{ii} = 1$ για $i = 1, 2, \ldots, n$.
2. Για κάθε $j = 1, 2, \ldots, n$ υπολογίζουμε

   (α΄) τα ℓ_{ij} διαδοχικά με $i = j, j+1 \ldots, n$ από τη σχέση

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Το γινόμενο παραμένει $\boldsymbol{A}$.

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

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

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

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

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

3.3 Επαναληπτικές Μέθοδοι

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

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

3.3.1 Στατικές μέθοδοι

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

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

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

και για ένα τουλάχιστον i ισχύει η αυστηρή ανισότητα,

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

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

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

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

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

Παρατηρήστε ότι για να υπολογίσουμε το x_i χρειαζόμαστε τις τιμές όλων των x_j με $j \neq i$.

Κατόπιν, εφαρμόζουμε μία από τις ακόλουθες παραλλαγές:

Jacobi

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

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

Gauss–Seidel

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

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

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

Παρατήρηση: Οι απαιτούμενες πράξεις για τον υπολογισμό του $\boldsymbol{x}^{(k+1)}$ στις επαναληπτικές μεθόδους που παρουσιάσαμε, είναι n^2, όπου n η διάσταση του $\boldsymbol{x}$.

Η μέθοδος Gauss–Seidel μπορεί να εφαρμοστεί και να συγκλίνει οπωσδήποτε, εκτός από τα γραμμικά συστήματα με κυρίαρχη διαγώνιο, και σε συστήματα στα οποία ο πίνακας των συντελεστών είναι συμμετρικός θετικά ορισμένος (§3.1.2).

Παρατηρήστε ότι αν $\boldsymbol{A}$ είναι ένας γενικός αντιστρέψιμος πραγματικός πίνακας, ο πίνακας $\boldsymbol{A}^T \cdot \boldsymbol{A}$ είναι συμμετρικός θετικά ορισμένος: ισχύει $(\boldsymbol{A}^T \cdot \boldsymbol{A})^T = \boldsymbol{A}^T \cdot \boldsymbol{A}$ και

$$\boldsymbol{x}^T \cdot \boldsymbol{A}^T \cdot \boldsymbol{A} \cdot \boldsymbol{x} = (\boldsymbol{A} \cdot \boldsymbol{x})^T (\boldsymbol{A} \cdot \boldsymbol{x}) = \|\boldsymbol{A} \cdot \boldsymbol{x}\|^2 > 0 \; , \forall \boldsymbol{x} \neq 0 \; .$$

Επομένως, ένα γενικό σύστημα $\boldsymbol{A} \cdot \boldsymbol{x} = \boldsymbol{b}$ μπορεί να μετατραπεί στο $(\boldsymbol{A}^T \cdot \boldsymbol{A}) \cdot \boldsymbol{x} = \boldsymbol{A}^T \cdot \boldsymbol{b}$ και να επιλυθεί με τη μέθοδο Gauss–Seidel (με συνολικά περισσότερες πράξεις από την απαλοιφή Gauss καθώς μόνο ο πολλαπλασιασμός των $\boldsymbol{A}^T$, $\boldsymbol{A}$ απαιτεί n^3 πράξεις). Στην περίπτωση βέβαια που οι πολλαπλασιασμοί του $\boldsymbol{A}^T$ με τα $\boldsymbol{A}$, $\boldsymbol{x}$ μπορούν να γίνουν με λιγότερες πράξεις (π.χ. όταν ο πίνακας $\boldsymbol{A}$ είναι *αραιός* (sparse) ή έχει ειδική μορφή), η μέθοδος Gauss–Seidel μπορεί να είναι πιο γρήγορη από την απαλοιφή Gauss (που δεν λαμβάνει υπόψη τη δομή του πίνακα $\boldsymbol{A}$).

Successive overrelaxation (SOR)

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

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

Αν $\omega = 1$ η μέθοδος SOR καταλήγει στη μέθοδο Gauss–Seidel. Μπορεί να δειχθεί ότι ο συντελεστής ω πρέπει να είναι στο διάστημα $(0, 2)$ για να υπάρχει δυνατότητα σύγκλισης. Αν ο πίνακας των συντελεστών είναι συμμετρικός θετικά ορισμένος τότε

η μέθοδος SOR συγκλίνει με οποιαδήποτε τιμή του ω στο $(0, 2)$ (αλλά με διαφορετική ταχύτητα σύγκλισης). Γενικά, μια τιμή $\omega > 1$ δίνει στην μέθοδο μεγαλύτερη ταχύτητα σύγκλισης από την Gauss–Seidel, ενώ, αν η Gauss–Seidel δεν συγκλίνει, μπορεί η μέθοδος SOR να δώσει λύση με κάποιο $\omega < 1$.

3.3.2 Μέθοδοι προβολής

Kaczmarz

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

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

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

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

3.4 Εφαρμογές

Η διαδικασία τριγωνοποίησης με τη μέθοδο απαλοιφής Gauss (§3.2.3) και γενικότερα, η επίλυση γραμμικών συστημάτων, βρίσκει εφαρμογή και σε άλλα προβλήματα γραμμικής άλγεβρας.

3.4.1 Υπολογισμός του αντίστροφου πίνακα

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

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

Συνεπώς, αν επιλέξουμε για διάνυσμα $\boldsymbol{b}$ διαδοχικά τα n διανύσματα

$$\boldsymbol{b}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} , \boldsymbol{b}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} , \boldsymbol{b}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} , \cdots , \boldsymbol{b}_n = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} ,$$

θα έχουμε ως λύσεις τις αντίστοιχες στήλες του πίνακα $\boldsymbol{A}^{-1}$.

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

Παράδειγμα

Ο αντίστροφος του πίνακα

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

βρίσκεται ως εξής:

Συμπληρώνουμε τον πίνακα $\boldsymbol{A}$ με τις στήλες του ταυτοτικού πίνακα:

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

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

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

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

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

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

3. Προσθέτουμε τη δεύτερη στην τέταρτη γραμμή:

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

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

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

Η άνω τριγωνοποίηση του αριστερού τμήματος ολοκληρώθηκε.

Αν θέσουμε

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

έχουμε καταλήξει σε 4 συστήματα με διαφορετικά δεξιά μέλη:

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

Προχωρούμε στην οπισθοδρόμηση (§3.2.3) για κάθε σύστημα χωριστά. Οι λύσεις τους είναι

$$\boldsymbol{X}_1 = \begin{bmatrix} 2 \\ -1 \\ -1 \\ 1 \end{bmatrix}, \quad \boldsymbol{X}_2 = \begin{bmatrix} 2 \\ -2 \\ -1 \\ 2 \end{bmatrix}, \quad \boldsymbol{X}_3 = \begin{bmatrix} -2 \\ 2 \\ 1 \\ -1 \end{bmatrix}, \quad \boldsymbol{X}_4 = \begin{bmatrix} -1 \\ 1 \\ 1 \\ -1 \end{bmatrix}.$$

Τα $\boldsymbol{X}_i$ είναι οι στήλες του αντίστροφου πίνακα του $\boldsymbol{A}$, επομένως

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

3.4.2 Υπολογισμός ορίζουσας

Μια μέθοδος για να υπολογίσουμε την ορίζουσα, εκτός από την εφαρμογή του τύπου (3.4), ξεκινά με τη μετατροπή του αρχικού πίνακα σε ένα τριγωνικό (άνω ή κάτω). Η πρόσθεση σε μία γραμμή ενός (τετραγωνικού) πίνακα του πολλαπλάσιου μίας άλλης είναι διαδικασία που διατηρεί την ορίζουσα. Ο άνω τριγωνικός πίνακας που παράγεται με την απαλοιφή Gauss, αν η τριγωνοποίηση περιοριστεί μόνο σε τέτοιες μεταβολές, έχει ίδια ορίζουσα με τον αρχικό. Προσέξτε ότι σε περίπτωση που εφαρμόσουμε οδήγηση (δηλαδή εναλλαγή γραμμών ή στηλών), πρέπει να λάβουμε υπόψη ότι κάθε τέτοια μεταβολή αλλάζει το πρόσημο της ορίζουσας.

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

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

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

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

Αν έχουμε ήδη υπολογίσει την ανάλυση του $\boldsymbol{A}$ σε $\boldsymbol{L}$, $\boldsymbol{U}$ (§3.2.5), η ορίζουσα υπολογίζεται εύκολα

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

3.4.3 Εύρεση ιδιοτιμών και ιδιοδιανυσμάτων

Η (3.3) μπορεί να γραφεί ως εξής

$$\boldsymbol{A} \cdot \boldsymbol{x} = \lambda \boldsymbol{x} \Rightarrow \boldsymbol{A} \cdot \boldsymbol{x} = \lambda \boldsymbol{I} \cdot \boldsymbol{x} \Rightarrow (\boldsymbol{A} - \lambda \boldsymbol{I}) \cdot \boldsymbol{x} = 0 \, .$$

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

Αφού προσδιοριστούν οι ιδιοτιμές, η επίλυση του γραμμικού συστήματος $(\boldsymbol{A} - \lambda \boldsymbol{I}) \cdot \boldsymbol{x} = 0$ μπορεί να γίνει με την απαλοιφή Gauss. Προσέξτε ότι το σύστημα έχει άπειρες λύσεις, οπότε τουλάχιστον μία από τις συνιστώσες του διανύσματος

$\boldsymbol{x}$ είναι «ελεύθερη». Κατά την τριγωνοποίηση του πίνακα $\boldsymbol{A} - \lambda \boldsymbol{I}$, κάποια γραμμή, συνήθως η τελευταία, αν γίνει κάτω τριγωνικός, θα έχει όλα τα στοιχεία της, και το διαγώνιο, ίσα με 0. Καθώς το σταθερό διάνυσμα στο δεξί μέλος του συστήματος είναι 0, η αντίστοιχη συνιστώσα του $\boldsymbol{x}$ δεν μπορεί να προσδιοριστεί. Η συγκεκριμένη συνιστώσα μπορεί να τεθεί αυθαίρετα 1 και να προχωρήσουμε στην επίλυση του συστήματος. Εύκολα μπορούμε να δούμε ότι όλες οι συνιστώσες του $\boldsymbol{x}$ θα είναι ανάλογες αυτής της συνιστώσας.

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

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

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

Θεώρημα κύκλων του Gershgorin

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

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

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

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

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

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

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

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

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

Θεώρημα Perron–Frobenius

Σύμφωνα με το θεώρημα Perron–Frobenius, ένας πραγματικός πίνακας με θετικά όλα τα στοιχεία του έχει μία θετική (πραγματική) ιδιοτιμή λ_1 με πολλαπλότητα 1 και όλες τις υπόλοιπες, γενικά μιγαδικές, με μέτρο μικρότερο από λ_1.

Το θεώρημα Ostrowski προσδιορίζει με μεγαλύτερη ακρίβεια το μέγιστο μέτρο: αν M είναι το μέγιστο στοιχείο ενός τέτοιου πίνακα και m το ελάχιστο, ισχύει για τις υπόλοιπες ιδιοτιμές ότι

$$|\lambda_i| \leq \lambda_1 \frac{M^2 - m^2}{M^2 + m^2} \ , \qquad i = 2, 3, \ldots, n \ .$$

Ανισότητες του Schur

Οι ιδιοτιμές λ_i ενός πραγματικού ή μιγαδικού πίνακα $\boldsymbol{A}_{n \times n}$ με στοιχεία a_{ij} ικανοποιούν τις σχέσεις

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

3.4.4 Επίλυση συστήματος μη γραμμικών εξισώσεων

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

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

μπορεί να αναχθεί σε επίλυση πολλών γραμμικών συστημάτων.

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

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

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

Η ανάπτυξη των συναρτήσεων f_i στην (3.13) δίνει

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

Ας ορίσουμε τον πίνακα

$$\boldsymbol{A} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \cdots & \dfrac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{\partial f_n}{\partial x_1} & \dfrac{\partial f_n}{\partial x_2} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{bmatrix} ,$$

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

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

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

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

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

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

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

Αλγόριθμος:

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

5. Επαναλαμβάνουμε από το βήμα 2.

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

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

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

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

3.5 Ασκήσεις

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

   Υπόδειξη: Δημιουργήστε ένα πίνακα 4×4 με τυχαία στοιχεία για να ελέγξετε το πρόγραμμά σας.

2. Υλοποιήστε την απαλοιφή Gauss με μερική οδήγηση.

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

3. Υλοποιήστε τη μέθοδο επίλυσης Gauss–Jordan.

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

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

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

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

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

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

   Χρησιμοποιήστε το για να υπολογίσετε την ορίζουσα του

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

7. Υλοποιήστε σε πρόγραμμα τη μέθοδο Cramer.

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

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

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

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

   με τη μέθοδο της ανάλυσης LU.

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

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

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

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

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

$$\begin{aligned} x + y + z &= 3 \\ x^2y + y^2z + z^2x &= 4 \\ x^2 + y^2 + z^2 &= 5 \end{aligned}$$

Κεφάλαιο 4

Προσέγγιση Συναρτήσεων

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

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

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

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

4.1 Προσέγγιση με πολυώνυμο

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

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

$$\sum_{j=0}^{n} a_j x_i^j = f_i \ , \qquad i = 0, 1, \ldots, n \ . \tag{4.1}$$

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

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

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

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

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

$$q_i(x) = \begin{cases} 1 , & i = 0 , \\ \prod_{j=0}^{i-1}(x-x_j) , & i = 1,2,\ldots,n . \end{cases} \tag{4.2β΄}$$

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

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

ή, γενικά,

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

Πλεονέκτημα αυτής της επιλογής για την κατασκευή του πολυωνύμου είναι ότι μπορεί να συμπεριληφθεί (ή να διαγραφεί) πολύ εύκολα κάποιο επιπλέον σημείο.

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

$$p(x) = \sum_{i=0}^{n} \ell_i(x) f_i \text{ , όπου} \tag{4.3α΄}$$

$$\ell_i(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} , \quad i = 0,1,2,\ldots,n . \tag{4.3β΄}$$

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

Παράδειγμα

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

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

όπου

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

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

4.1.1 Μετατροπή

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

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

ενώ οι υπόλοιποι μπορούν να υπολογιστούν από τους τύπους Vieta:

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

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

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

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

4.1.2 Σφάλμα προσέγγισης με πολυώνυμο

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

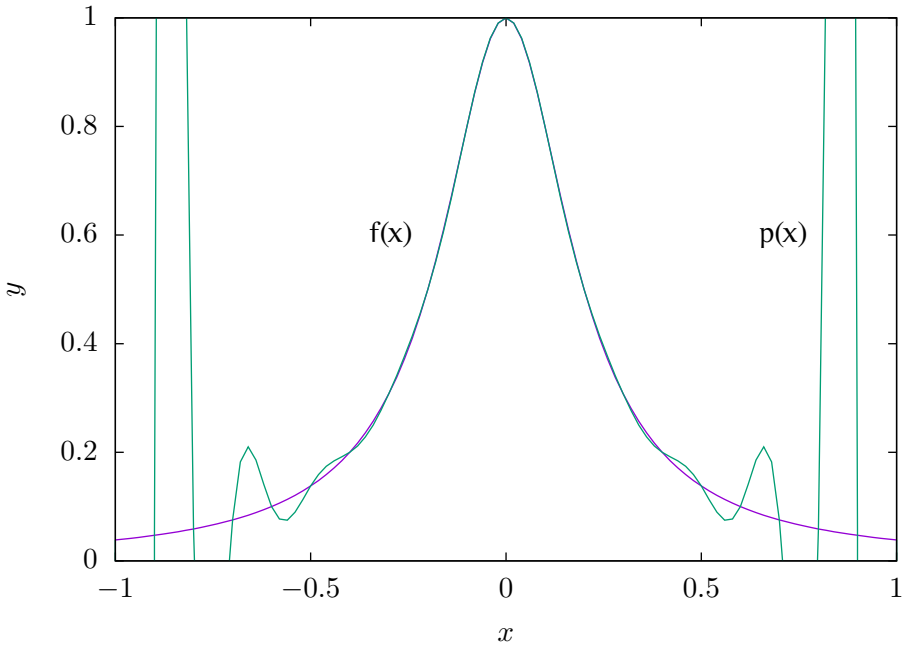
$$|f(x) - p(x)| = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x - x_i) \ , \quad \forall x \in [a, b] \ , \tag{4.4}$$

όπου ξ ένα σημείο στο (a, b).

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

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

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



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

Παράδειγμα

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

Η εφαρμογή του τύπου (4.3) δίνει το πολυώνυμο

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

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

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

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

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

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

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

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

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

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

4.2 Προσέγγιση με λόγο πολυωνύμων

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

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

όπου

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

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

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

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

Το μειονέκτημα αυτής της προσέγγισης είναι ότι δεν γνωρίζουμε πώς να επιλέξουμε τα M, N (παρά μόνο το άθροισμά τους). Κακή επιλογή αυτών θα δώσει κακή προσεγγιστική συνάρτηση.

4.2.1 Προσέγγιση Padé

Αν γνωρίζουμε μια πολύπλοκη συνάρτηση και θέλουμε να την απλοποιήσουμε, μπορούμε εναλλακτικά να απαιτήσουμε η $R(x) = P(x)/Q(x)$, με $P(x)$, $Q(x)$ πολυώνυμα M και N βαθμού αντίστοιχα, και η $f(x)$ να έχουν σε ένα σημείο x_0, ίδιες

τιμές και ίδιες παραγώγους μέχρι και τάξης $M+N$, δηλαδή να ισχύει

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

Η λύση αυτού του μη γραμμικού συστήματος προσδιορίζει τους συντελεστές των $P(x)$, $Q(x)$ (ένας αυθαίρετα μπορεί να οριστεί ίσος με 1) και σχηματίζει την προσεγγιστική συνάρτηση $R(x)$ του Padé. Το σφάλμα της προσέγγισης, $|f(x)-R(x)|$, είναι ανάλογο του $(x-x_0)^{M+N+1}$. Η συνάρτηση Padé συνήθως προσεγγίζει καλύτερα την f παρά το ανάπτυγμα Taylor της f με όρους μέχρι βαθμού $M+N$.

4.3 Προσέγγιση κατά τμήματα με πολυώνυμα ελάχιστου βαθμού

Για να αποφύγουμε τις έντονες ταλαντώσεις στα άκρα του διαστήματος όταν χρησιμοποιούμε ένα πολυώνυμο μεγάλου βαθμού, μπορούμε να προσεγγίσουμε την άγνωστη συνάρτηση κατά τμήματα, χρησιμοποιώντας πολυώνυμα του ελάχιστου δυνατού βαθμού. Αυτό σημαίνει ότι μπορούμε, π.χ., να χωρίσουμε το συνολικό διάστημα $[x_0, x_n]$ σε τμήματα που ορίζονται από δύο διαδοχικά σημεία: $[x_0, x_1], [x_1, x_2], \ldots, [x_{n-1}, x_n]$. Σε καθένα από αυτά τα n διαστήματα, προσαρμόζουμε ένα πολυώνυμο με δύο άγνωστους συντελεστές: $p_i(x) = a_i x + b_i$, $i = 0, 1, \ldots, n-1$. Προσδιορίζουμε δηλαδή τα ευθύγραμμα τμήματα με αρχή τα x_i και τέλος τα x_{i+1}.

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

4.4 Προσέγγιση με spline

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

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

Έστω ένα σύνολο $n+1$ σημείων (x_i, f_i) με $i = 0, 1, \ldots, n$ και $x_i < x_{i+1}$. Ανά δύο διαδοχικά σημεία, δηλαδή από τα σημεία με δείκτη i = $(0,1)$, $(1,2)$, $(2,3)$, ...,

περνούμε πολυώνυμα τρίτου βαθμού (γι᾽ αυτό χαρακτηρίζεται ως «κυβική» η spline) της μορφής

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

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

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

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

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

$$d_i = f_i \ , \tag{4.5α′}$$

$$a_i(x_{i+1} - x_i)^3 + b_i(x_{i+1} - x_i)^2 + c_i(x_{i+1} - x_i) + d_i = f_{i+1} \ . \tag{4.5β′}$$

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

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

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

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

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

$$p''_0(x_0) = 0 \ \Rightarrow \ 2b_0 = 0 \ , \tag{4.7α′}$$

$$p''_{n-1}(x_n) = 0 \ \Rightarrow \ 6a_{n-1}(x_n - x_{n-1}) + 2b_{n-1} = 0 \ . \tag{4.7β′}$$

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

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

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

Σχηματίζουμε τις εξισώσεις

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

Η λύση του γραμμικού συστήματος υπολογίζει τους συντελεστές a_i, b_i, c_i με $i = 0, 1, \ldots, n-1$. Η spline είναι η κλαδική συνάρτηση

$$\text{spline}(x) = \begin{cases} p_0(x)\ , & \text{αν } x \in [x_0, x_1) \\ p_1(x)\ , & \text{αν } x \in [x_1, x_2) \\ \vdots & \vdots \\ p_{n-1}(x)\ , & \text{αν } x \in [x_{n-1}, x_n] \end{cases}$$

4.5 Προσέγγιση παραγώγων

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

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

Έστω ότι θέλουμε να υπολογίσουμε την παράγωγο τής $f(x)$ στο σημείο x_0. Τα πλησιέστερα σημεία στα οποία γνωρίζουμε τη συνάρτηση, εκατέρωθεν του x_0 είναι τα x_1, x_{-1}. Επομένως,

$$f'(x_0) \approx \frac{f(x_1) - f(x_0)}{x_1 - x_0}\ , \tag{4.8α΄}$$

$$f'(x_0) \approx \frac{f(x_{-1}) - f(x_0)}{x_{-1} - x_0}\ . \tag{4.8β΄}$$

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

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

Λύνοντας ως προς $f'(x_0)$ έχουμε

$$f'(x_0) = \frac{f(x_1) - f(x_0)}{x_1 - x_0} - \frac{x_1 - x_0}{2!}f''(x_0) - \frac{(x_1-x_0)^2}{3!}f'''(x_0) - \cdots\ . \tag{4.9}$$

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

$$f'(x_0) = \frac{f(x_{-1}) - f(x_0)}{x_{-1} - x_0} - \frac{x_{-1} - x_0}{2!} f''(x_0) - \frac{(x_{-1} - x_0)^2}{3!} f'''(x_0) - \cdots . \tag{4.10}$$

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

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

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

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

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

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

4.5.1 Συστηματική παραγωγή τύπων προσέγγισης παραγώγου

Η παράγωγος της $f(x)$ οποιασδήποτε τάξης m (ακόμα και μηδενικής), σε κάποιο σημείο $\bar{x}$ στο πεδίο ορισμού της συνάρτησης, μπορεί να γραφεί ως γραμμικός συνδυασμός γνωστών τιμών της συνάρτησης σε n σημεία x_i, με $i = 0, \ldots, n-1$ και $n > m$:

$$f^{(m)}(\bar{x}) \approx \sum_{i=0}^{n-1} w_i f(x_i) . \tag{4.12}$$

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

[1] Ο πίνακας είναι ο ανάστροφος του πίνακα Vandermonde.

Παράδειγμα

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

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

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

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

Η λύση του γραμμικού συστήματος δίνει

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

Αν $\bar{x} \equiv x_0$ έχουμε $a = -1/2h$, $b = 0$, $c = 1/2h$· παράγουμε δηλαδή τον τύπο (4.11). Αν $\bar{x} \equiv x_0 + h$ έχουμε $a = 1/2h$, $b = -2/h$, $c = 3/2h$. Επομένως

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

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

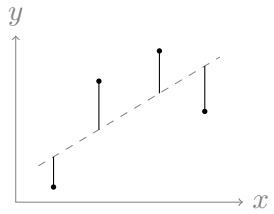
$$w_i = \left. \frac{\mathrm{d}^m \ell_i(x)}{\mathrm{d}x^m} \right|_{x=\bar{x}} \; .$$

4.6 Προσέγγιση με τη μέθοδο ελάχιστων τετραγώνων

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

προκύπτει από την απαίτηση να περνά από αυτά τα σημεία. Η απλούστερη συνάρτηση, εν γνώσει μας, δεν μπορεί να περνά από τα ίδια σημεία με την πολύπλοκη, αλλά επιδιώκουμε να πλησιάζει όσο περισσότερο γίνεται.

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



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

4.6.1 Ευθεία ελάχιστων τετραγώνων

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

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

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

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

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

Η εφαρμογή της μεθόδου Cramer (§3.2.2) δίνει αμέσως ότι

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

$$\beta = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \overline{y} - \alpha\overline{x} , \tag{4.13β′}$$

όπου

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

η μέση τιμή ενός μεγέθους w.

Η ποσότητα

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

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

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

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

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

είναι συμμετρικός θετικά ορισμένος[3].

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

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

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

4.6.2 Πολυώνυμο ελάχιστων τετραγώνων

Έστω ότι γνωρίζουμε από τη θεωρία ότι η συνάρτηση $f(x)$ που έδωσε τις μετρήσεις (x_i, y_i) είναι πολυωνυμική βαθμού m ως προς x.

Σχηματίζουμε τη συνάρτηση $p(x) = \sum_{i=0}^{m} \alpha_i x^i$. Η ελαχιστοποίηση της ποσότητας

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

ως προς τα α_k δίνει τις εξισώσεις

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

Οι $m+1$ εξισώσεις αυτές αποτελούν ένα γραμμικό σύστημα για τους $m+1$ άγνωστους α_j. Ο πίνακας των συντελεστών έχει στοιχεία

$$A_{kj} = \sum_{i=1}^{n} x_i^{k+j} \qquad \text{με } k = 0, 1, \ldots, m \ , \quad j = 0, 1, \ldots, m \ .$$

Οι σταθεροί όροι είναι

$$b_k = \sum_{i=1}^{n} x_i^k y_i \ .$$

[3] Έλεγχος με το κριτήριο του Sylvester, §3.1.2:

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

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

Παράδειγμα

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

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

Επομένως,

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

Η ευθεία ελάχιστων τετραγώνων είναι η $p(x) = 0.89968 + 1.70784x$. Όπως παρατηρούμε, οι συντελεστές της δεν έχουν σχέση με τους δυο πρώτους συντελεστές του δευτεροβάθμιου πολυωνύμου.

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

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

Παραδείγματα

Έστω ότι η θεωρητική σχέση είναι $y = a + b\mathrm{e}^x$. Αν ορίσουμε

$$\tilde{y} = y \ , \qquad \tilde{x} = \mathrm{e}^x \ , \qquad \tilde{\alpha} = b \ , \qquad \tilde{\beta} = a \ ,$$

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

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

$$\tilde{y} = \ln y \ , \qquad \tilde{x} = \ln x \ , \qquad \tilde{\alpha} = b \ , \qquad \tilde{\beta} = \ln a \ ,$$

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

4.7 Ασκήσεις

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

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

   (α′) τον τύπο Lagrange.

   (β′) τον τύπο Newton.

   (γ′) την απαλοιφή Gauss.

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

2. Τροποποιήστε τον κώδικα που γράψατε για την προηγούμενη άσκηση ώστε να υπολογίζει και να επιστρέφει, εκτός από τις προσεγγιστικές τιμές της συνάρτησης (δηλαδή, τις τιμές του πολυωνύμου παρεμβολής, $p(x)$), και τις αντίστοιχες τιμές της *πρώτης παραγώγου*, $p'(x)$.

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

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

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

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

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

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

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

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

   (α΄) $y = ax + b$ (γραμμική),

   (β΄) $y = ax^b$ (δύναμη),

   (γ΄) $y = a + b\mathrm{e}^x$ (εκθετική),

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

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

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

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

$$P = \sigma A T^4 \,,$$

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

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

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

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

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

Υπολογίστε την επιτάχυνση της βαρύτητας από τις ακόλουθες πειραματικές μετρήσεις

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

8. Δημιουργήστε ένα αρχείο με όνομα «points2.dat», που να περιέχει τα ζεύγη $(x_i, f(x_i))$ μιας γνωστής συνάρτησης $f(x)$, π.χ. $\sin^2 x$. Τα x_i ας είναι 15 ισαπέχοντα σημεία στο διάστημα $[2, 4]$. Το αρχείο θα έχει στην πρώτη γραμμή το πλήθος των σημείων και θα ακολουθούν τα ζεύγη (x_i, y_i).

Υπολογίστε την πρώτη και τη δεύτερη παράγωγο στο $x = 2.5$, της συνάρτησης που δίνεται από τα σημεία στο αρχείο «points2.dat» Για τον υπολογισμό των παραγώγων χρησιμοποιήστε όλα τα σημεία.

Κεφάλαιο 5

Αριθμητική Ολοκλήρωση

5.1 Εισαγωγή

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

$$\int_a^b f(x)\,\mathrm{d}x\ ,$$

με $f(x)$ μια συνεχή συνάρτηση στο $[a,b]$, με πεπερασμένο πλήθος απομονωμένων σημείων ασυνέχειας. Για τη συντριπτική πλειονότητα των συναρτήσεων, δεν υπάρχει ή είναι πολύ δύσχρηστος ο τύπος της αντιπαραγώγου της $f(x)$, δηλαδή της $F(x)$ που ικανοποιεί τη σχέση $F'(x) = f(x)$, ώστε να υπολογιστεί ακριβώς το ολοκλήρωμα από τον τύπο

$$\int_a^b f(x)\,\mathrm{d}x = F(b) - F(a)\ .$$

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

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

$$\int_a^b f(x)\,\mathrm{d}x \approx \sum_{i=1}^{n} w_i f_i\ , \tag{5.1}$$

όπου $f_i \equiv f(x_i)$.

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

5.1.1 Ολοκληρώματα με μη πεπερασμένα όρια ολοκλήρωσης

Συχνά υπάρχει η ανάγκη να υπολογιστούν ολοκληρώματα με ένα τουλάχιστον όριο $\pm\infty$. Η πρώτη κατηγορία μεθόδων που θα δούμε παρακάτω, εφαρμόζεται για ολοκληρώματα με πεπερασμένα όρια. Αν η ολοκληρωτέα συνάρτηση έχει κατάλληλη μορφή, υπάρχουν αλγόριθμοι στη δεύτερη κατηγορία μεθόδων που μπορούν να υπολογίσουν το ολοκλήρωμά της με όρια τα $\pm\infty$. Αλλιώς, και με την προϋπόθεση ότι η συνάρτηση $f(x)$ μειώνεται τείνοντας στο άπειρο όριο και μάλιστα ισχύει $xf(x) \to 0$ όταν το x τείνει στο άπειρο όριο (ή στα άπειρα όρια), μπορούμε να κάνουμε μια κατάλληλη αλλαγή μεταβλητής, π.χ. $x = 1/t$, ώστε να προκύψουν ολοκληρώματα με πεπερασμένα όρια. Έτσι:

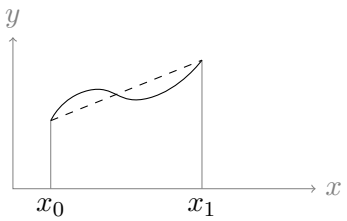
$$\int_a^b f(x)\,\mathrm{d}x = \int_{1/b}^{1/a} \frac{1}{t^2} f(\frac{1}{t})\,\mathrm{d}t$$

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

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

$$\int_{-3}^{+\infty} f(x)\,\mathrm{d}x = \int_{-3}^{c} f(x)\,\mathrm{d}x + \int_{c}^{+\infty} f(x)\,\mathrm{d}x = \int_{-3}^{c} f(x)\,\mathrm{d}x + \int_{0}^{1/c} \frac{1}{t^2} f(\frac{1}{t})\,\mathrm{d}t\ .$$

5.2 Κανόνας Τραπεζίου



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

Μια προσεγγιστική τιμή του ολοκληρώματος

$$\int_{x_0}^{x_1} f(x)\,\mathrm{d}x$$

μπορεί να υπολογιστεί ολοκληρώνοντας το πολυώνυμο παρεμβολής που διέρχεται από τα σημεία $(x_0, f(x_0))$, $(x_1, f(x_1))$, Σχήμα 5.1. Το πολυώνυμο αυτό είναι φυσικά

πρωτοβάθμιο, δηλαδή ευθεία, και δίνεται από τον τύπο

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

Η ολοκλήρωσή του με όρια x_0, x_1 δίνει

$$\int_{x_0}^{x_1} f(x)\,\mathrm{d}x \approx \int_{x_0}^{x_1} p_1(x)\,\mathrm{d}x = \frac{x_1 - x_0}{2}\left[f(x_0) + f(x_1)\right] \ . \tag{5.2}$$

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

5.2.1 Σφάλμα ολοκλήρωσης κανόνα τραπεζίου

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

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

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

Στο ανάπτυγμα έχουμε παραλείψει όρους υψηλότερης τάξης από 2.

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

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

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

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

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

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

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

Με ακριβή μαθηματική αντιμετώπιση καταλήγουμε ότι το σφάλμα ε του απλού τύπου τραπεζίου είναι

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

χωρίς επιπλέον όρους.

5.2.2 Σύνθετος τύπος τραπεζίου

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

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

όπου $f_i \equiv f(x_i)$.

5.2.3 Σφάλμα ολοκλήρωσης σύνθετου τύπου τραπεζίου

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

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

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

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

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

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

και

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

Παράδειγμα

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

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

Επομένως,

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

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

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

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

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

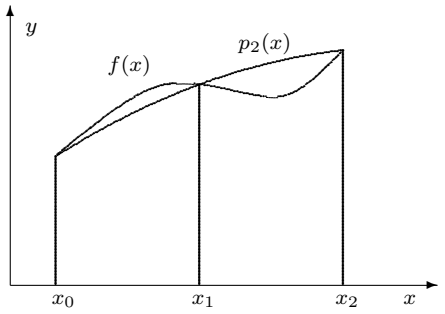
$$|E| \leq \frac{\pi}{12} h^2 = \frac{\pi}{12} \frac{\pi^2}{n^2} \leq 10^{-4} \Rightarrow n \geq 161.$$

5.3 Κανόνας Simpson

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

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

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



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

Lagrange, (4.3), ότι είναι

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

Επομένως,

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

Άρα, ο (απλός) *τύπος του Simpson* είναι

$$\int_{x_0}^{x_2} f(x)\,\mathrm{d}x \approx \frac{h}{3}(f_0 + 4f_1 + f_2) \, . \tag{5.8}$$

5.3.1 Σφάλμα ολοκλήρωσης κανόνα Simpson

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

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

5.3.2 Σύνθετος τύπος Simpson

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

ότι

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

5.3.3 Σφάλμα ολοκλήρωσης σύνθετου τύπου Simpson

Αν η τέταρτη παράγωγος της $f(x)$ είναι φραγμένη στο $[a, b]$,

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

το σφάλμα E του σύνθετου τύπου Simpson είναι

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

Παράδειγμα

Ο υπολογισμός του

$$I = \int_0^{\pi} \sin x\,\mathrm{d}x$$

με τον σύνθετο τύπο Simpson δίνει

n	I_n	$E = I - I_n$
2	2.0943951	−0.0943951
4	2.0045598	−0.0045598
6	2.0008632	−0.0008632
8	2.0002692	−0.0002692
12	2.0000526	−0.0000526
16	2.0000166	−0.0000166
20	2.0000068	−0.0000068

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

$$|E| \leq \frac{\pi^5}{90n^4} .$$

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

5.4 Κανόνας Simpson των 3/8

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

Για 4 δεδομένα ισαπέχοντα σημεία x_0, x_1, x_2, x_3 μπορεί να αποδειχθεί ότι το ολοκλήρωμα της $f(x)$ στο $[x_0, x_3]$ δίνεται προσεγγιστικά από τον τύπο:

$$\int_{x_0}^{x_3} f(x)\,\mathrm{d}x \approx \frac{3h}{8}\left[f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)\right] , \tag{5.11}$$

όπου $h = (x_3 - x_0)/3$.

5.4.1 Σφάλμα ολοκλήρωσης κανόνα Simpson $^3/_8$

Μπορεί να δειχθεί ότι ο τύπος του Simpson $^3/_8$ έχει σφάλμα:

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

Ο τύπος των $^3/_8$ είναι παρόμοιας ακρίβειας με τον τύπο του Simpson με το $^1/_3$, παρά το γεγονός ότι χρησιμοποιεί ένα παραπάνω σημείο. Η μεγάλη χρησιμότητα του τύπου του $^3/_8$ είναι ότι ο αριθμός των διαστημάτων που χρειάζεται για τον υπολογισμό είναι περιττός (3) και συνεπώς μπορεί να χρησιμοποιηθεί όταν θέλουμε υψηλή ακρίβεια σε περιττό πλήθος διαστημάτων συνδυάζοντάς τον με τον τύπο Simpson $^1/_3$: στα πρώτα τρία διαστήματα μπορούμε να εφαρμόσουμε τον τύπο $^3/_8$ και στα υπόλοιπα (που είναι άρτια στο πλήθος) τον τύπο $^1/_3$.

5.4.2 Σύνθετος τύπος Simpson των $^3/_8$

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

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

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

5.5 Εναλλακτικός υπολογισμός των τύπων Newton–Cotes

Οι (απλοί) τύποι τραπεζίου, Simpson, κλπ., έχουν τη γενική ονομασία *τύποι Newton–Cotes*. Όπως είδαμε κατά την εξαγωγή των απλών τύπων τραπεζίου και Simpson, οι συντελεστές w_i στην εξίσωση (5.1) της γενικής μορφής των τύπων Newton–Cotes μπορούν να προκύψουν από την ολοκλήρωση του πολυωνύμου παρεμβολής στη μορφή Lagrange· είναι τα ολοκληρώματα στο $[a, b]$ των συναρτήσεων της βάσης Lagrange, (4.3β′):

$$w_i = \int_{x_0}^{x_n} \ell_i(x)\,\mathrm{d}x .$$

Εναλλακτικά, μπορούν να υπολογιστούν και με άλλο τρόπο. Βασιζόμαστε στη (5.1) αλλά απαιτούμε να είναι *ακριβής* όταν η $f(x)$ είναι διαδοχικά 1, x, x^2, ...,

x^n, όπου $n+1$ το πλήθος των σημείων. Προκύπτει έτσι ένα γραμμικό σύστημα εξισώσεων με άγνωστους τους συντελεστές w_i, το οποίο έχει μοναδική λύση.

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

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

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

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

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

5.5.1 Ανοιχτοί, ημι-ανοιχτοί, κλειστοί τύποι

Οι απλοί τύποι τραπεζίου και Simpson που παρουσιάσαμε είναι *κλειστοί* τύποι. Χαρακτηρίζονται έτσι επειδή περιλαμβάνουν τις τιμές της ολοκληρωτέας συνάρτησης και στα δύο άκρα ολοκλήρωσης. Λιγότερο ακριβείς αλλά κάποιες φορές χρήσιμοι είναι οι *ανοιχτοί* τύποι Newton–Cotes, δηλαδή αυτοί που δεν περιλαμβάνουν στα σημεία υπολογισμού τα όρια ολοκλήρωσης. Έτσι το ολοκλήρωμα

$$\int_a^b f(x)\,\mathrm{d}x$$

υπολογίζεται προσεγγιστικά από το άθροισμα

$$\sum_{i=1}^{n-1} w_i f_i$$

ως ένας γραμμικός συνδυασμός των τιμών της συνάρτησης, f_i, στα σημεία $x_i = a + ih$, όπου $h = (b-a)/n$ και $i = 1, 2, \ldots, n-1$ (χωρίς δηλαδή να περιλαμβάνουμε τα σημεία $x_0 \equiv a$ και $x_n \equiv b$). Εύκολα προκύπτουν οι ανοιχτοί τύποι, διαδοχικά για $n = 2, 3, 4 \ldots$

$$\begin{aligned} &2h f_1 + \mathcal{O}(h^3) \\ &\frac{3}{2} h\,(f_1 + f_2) + \mathcal{O}(h^3) \\ &\frac{4}{3} h\,(2f1 - f_2 + 2f_3) + \mathcal{O}(h^5) \\ &\vdots \end{aligned}$$

Χρησιμότητα έχουν επίσης οι *ημι-ανοιχτοί* τύποι Newton–Cotes που περιλαμβάνουν στα σημεία υπολογισμού το ένα όριο ολοκλήρωσης.

5.5.2 Παρατηρήσεις

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

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

5.6 Μέθοδοι Gauss

5.6.1 Μέθοδος Gauss–Legendre

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

$$\begin{aligned} x &= \frac{b-a}{2}t + \frac{b+a}{2}\,, \\ \mathrm{d}x &= \frac{b-a}{2}\,\mathrm{d}t\,. \end{aligned}$$

Συνεπώς, μπορούμε πάντα να μετασχηματίσουμε ένα ολοκλήρωμα στο διάστημα $[a, b]$ σε άλλο στο διάστημα $[-1, 1]$ με τον τύπο

$$\int_a^b f(x)\,\mathrm{d}x = \frac{b-a}{2}\int_{-1}^{1} f\left(\frac{b-a}{2}t + \frac{b+a}{2}\right)\,\mathrm{d}t\,.$$

Ας ξαναδούμε τον βασικό τύπο ολοκλήρωσης (5.1), γραμμένο όμως τώρα για το διάστημα $[-1, 1]$:

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \approx \sum_{i=1}^{m} w_i f(x_i)\,, \tag{5.12}$$

όπου x_i σταθερά σημεία στο $[-1, 1]$ και w_i συντελεστές. Ο τύπος έχει γενική μορφή και περιλαμβάνει

- τον απλό κανόνα τραπεζίου: έχουμε $w_1 = w_2 = 1$, $x_1 = -1$, $x_2 = 1$.
- τον απλό κανόνα Simpson: έχουμε $w_1 = w_3 = 1/3$, $w_2 = 4/3$, $x_1 = -1$, $x_2 = 0$, $x_3 = 1$.

Το ερώτημα είναι: για δεδομένο αριθμό σημείων m, ποια είναι τα w_i, x_i, $i = 1, 2, \ldots, m$ ώστε ο κανόνας (5.12) να έχει τη μέγιστη δυνατή ακρίβεια; Προσέξτε ότι, σε αντίθεση με τους τύπους που παρουσιάσαμε μέχρι τώρα, έχουμε τη δυνατότητα επιλογής των x_i.

Έστω ότι η μέγιστη δυνατή ακρίβεια σημαίνει πως ο κανόνας δίνει το ακριβές αποτέλεσμα στην ολοκλήρωση των συνολικά $2m$ μονωνύμων (όσα και οι άγνωστοι) 1, x, x^2, ..., x^{2m-1} (και επομένως, οποιουδήποτε γραμμικού συνδυασμού τους). Αυτή η συνθήκη οδηγεί στους *κανόνες ολοκλήρωσης Gauss*.

Κανόνας Gauss με $m = 1$

Έχουμε

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \approx w_1 f(x_1)\ .$$

Καθώς πρέπει να προκύπτει το ακριβές αποτέλεσμα για $f(x) = 1$ και $f(x) = x$ έχουμε

$$\begin{aligned} f(x) = 1 \quad &\Rightarrow \quad w_1 1 = \int_{-1}^{1} 1\,\mathrm{d}x = 2 \Rightarrow w_1 = 2\ , \\ f(x) = x \quad &\Rightarrow \quad w_1 x_1 = \int_{-1}^{1} x\,\mathrm{d}x = 0 \Rightarrow x_1 = 0\ . \end{aligned}$$

Επομένως, ο κανόνας Gauss με ένα σημείο είναι

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \approx 2f(0)\ .$$

Ο κανόνας αυτός ολοκληρώνει ακριβώς τα $1, x$ αλλά όχι το x^2.

Κανόνας Gauss με $m = 2$

Ζητώντας να παράγεται το ακριβές αποτέλεσμα για $f(x) = 1$, $f(x) = x$, $f(x) = x^2$, $f(x) = x^3$, έχουμε

$$\begin{aligned} w_1 + w_2 &= 2\ , \\ w_1 x_1 + w_2 x_2 &= 0\ , \\ w_1 {x_1}^2 + w_2 {x_2}^2 &= \frac{2}{3}\ , \\ w_1 {x_1}^3 + w_2 {x_2}^3 &= 0\ . \end{aligned}$$

Το παραπάνω σύστημα λύνεται αναλυτικά επιλύοντας διαδοχικά τις εξισώσεις ως προς κάποιον από τους αγνώστους και αντικαθιστώντας στις επόμενες. Η λύση του συστήματος είναι μοναδική (πέρα από την αλλαγή $w_1 \leftrightarrow w_2$, $x_1 \leftrightarrow x_2$) και είναι η εξής

$$\begin{aligned} w_1 &= 1 \,, & x_1 &= -\frac{1}{\sqrt{3}} \,, \\ w_2 &= 1 \,, & x_2 &= \frac{1}{\sqrt{3}} \,. \end{aligned}$$

Επομένως,

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \approx f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right) \,.$$

Κανόνας Gauss με $m = 3$

Η απαίτηση για ακριβές αποτέλεσμα όταν $f(x) = x^k$, $k = 0, \ldots, 5$ σχηματίζει το μη γραμμικό σύστημα

$$\begin{aligned} w_1 + w_2 + w_3 &= 2 \,, \\ w_1 x_1 + w_2 x_2 + w_3 x_3 &= 0 \,, \\ w_1 x_1^2 + w_2 x_2^2 + w_3 x_3^2 &= \frac{2}{3} \,, \\ w_1 x_1^3 + w_2 x_2^3 + w_3 x_3^3 &= 0 \,, \\ w_1 x_1^4 + w_2 x_2^4 + w_3 x_3^4 &= \frac{2}{5} \,, \\ w_1 x_1^5 + w_2 x_2^5 + w_3 x_3^5 &= 0 \,. \end{aligned}$$

Η λύση του είναι

$$\begin{aligned} w_1 &= \frac{5}{9} \,, & x_1 &= -\sqrt{0.6} \,, \\ w_2 &= \frac{8}{9} \,, & x_2 &= 0 \,, \\ w_3 &= \frac{5}{9} \,, & x_3 &= \sqrt{0.6} \,. \end{aligned}$$

Κανόνας Gauss με οποιοδήποτε m

Η μέθοδος ολοκλήρωσης που παρουσιάστηκε σε αυτή την παράγραφο λέγεται *μέθοδος Gauss–Legendre*. Ονομάζεται έτσι γιατί, στη γενική περίπτωση, τα σημεία x_i, με $i = 1, \ldots, m$, είναι οι ρίζες του πολυωνύμου Legendre m τάξης, $P_m(x)$, και μπορούν να υπολογιστούν εύκολα, χωρίς τη λύση των μη γραμμικών συστημάτων. Οι συντελεστές w_i δίνονται από τη σχέση

$$w_i = \frac{2}{(1 - {x_i}^2)[P'_m(x_i)]^2} \,.$$

Τα δύο πρώτα πολυώνυμα Legendre είναι τα

$$\begin{aligned} P_0(x) &= 1 \,, \\ P_1(x) &= x \,, \end{aligned}$$

ενώ τα υπόλοιπα μπορούν να παραχθούν από την αναδρομική σχέση

$$P_{n+1}(x) = \frac{1}{n+1}((2n+1)xP_n(x) - nP_{n-1}(x)) \,, \qquad n > 0 \,.$$

Για το σφάλμα ε_m στον υπολογισμό του ολοκληρώματος με τη μέθοδο Gauss–Legendre m σημείων ισχύει

$$|\varepsilon_m| \leq \frac{2^{2m+1}(m!)^4}{(2m+1)[(2m)!]^3} f^{(2m)}(\xi) \,, \quad \text{για κάποιο } \xi \in (-1,1) \,.$$

Στην πράξη, μπορούμε να υπολογίσουμε την προσεγγιστική τιμή για $m = 1, 2, 3, \ldots$ και να επιλέξουμε το μικρότερο m που θα μας δώσει ικανοποιητική προσέγγιση.

Η μέθοδος ολοκλήρωσης Gauss μπορεί να επεκταθεί και στον υπολογισμό ολοκληρωμάτων ειδικής μορφής. Έτσι έχουμε τις ακόλουθες μεθόδους:

5.6.2 Μέθοδος Gauss–Hermite

Σύμφωνα με αυτή τη μέθοδο

$$\int_{-\infty}^{\infty} \mathrm{e}^{-x^2} f(x)\,\mathrm{d}x \approx \sum_{i=1}^{m} w_i f(x_i) \,, \tag{5.13}$$

όπου $m > 0$, x_i οι ρίζες του πολυωνύμου Hermite τάξης m, $H_m(x)$, και w_i τα αντίστοιχα βάρη, τα οποία είναι τα

$$w_i = \frac{2^{m+1} m! \sqrt{\pi}}{[H'_m(x_i)]^2} \,.$$

Καθώς ισχύει $H'_m(x) = 2mH_{m-1}(x)$, ο τύπος για το βάρος w_i μπορεί να γραφεί

$$w_i = \frac{2^{m-1}(m-1)!\sqrt{\pi}}{m[H_{m-1}(x_i)]^2} \,.$$

Τα δύο πρώτα πολυώνυμα Hermite είναι τα

$$\begin{aligned} H_0(x) &= 1 \,, \\ H_1(x) &= 2x \,, \end{aligned}$$

ενώ τα υπόλοιπα μπορούν να παραχθούν από την αναδρομική σχέση

$$H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x) \,, \qquad n > 0 \,.$$

5.6.3 Μέθοδος Gauss–Laguerre

Σύμφωνα με αυτή τη μέθοδο

$$\int_0^\infty \mathrm{e}^{-x} f(x)\,\mathrm{d}x \approx \sum_{i=1}^{m} w_i f(x_i)\ , \tag{5.14}$$

όπου $m > 0$, x_i είναι οι ρίζες του πολυωνύμου Laguerre τάξης m, $L_m(x)$, και w_i τα αντίστοιχα βάρη, τα οποία είναι τα

$$w_i = \frac{x_i}{[(m+1)L_{m+1}(x_i)]^2}\ .$$

Τα δύο πρώτα πολυώνυμα Laguerre είναι τα

$$\begin{aligned} L_0(x) &= 1\ , \\ L_1(x) &= -x+1\ , \end{aligned}$$

ενώ τα υπόλοιπα μπορούν να παραχθούν από την αναδρομική σχέση

$$L_{n+1}(x) = \frac{1}{n+1}((2n+1-x)L_n(x) - nL_{n-1}(x))\ , \qquad n > 0\ .$$

5.6.4 Μέθοδος Gauss–Chebyshev

Σύμφωνα με αυτή τη μέθοδο

$$\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} f(x)\,\mathrm{d}x \approx \sum_{i=1}^{m} w_i f(x_i)\ , \tag{5.15}$$

$$\int_{-1}^{1} \sqrt{1-x^2} f(x)\,\mathrm{d}x \approx \sum_{i=1}^{m} c_i f(\rho_i)\ , \tag{5.16}$$

όπου x_i είναι οι ρίζες του πολυωνύμου Chebyshev *πρώτου* είδους, τάξης m, $T_m(x)$, και ρ_i είναι οι ρίζες του πολυωνύμου Chebyshev *δεύτερου* είδους, τάξης m, $U_m(x)$. Οι ρίζες των δύο πολυωνύμων μπορούν να υπολογιστούν σε κλειστή μορφή:

$$x_i = \cos\left(\frac{2i-1}{2m}\pi\right)\ , \qquad \rho_i = \cos\left(\frac{i\pi}{m+1}\right)\ .$$

Τα αντίστοιχα βάρη w_i, c_i είναι:

$$w_i = \frac{\pi}{m}\ , \qquad c_i = \frac{\pi}{m+1}(1-\rho_i^2)\ .$$

5.6.5 Κατασκευή μεθόδων Gauss

Επιθυμούμε να υπολογίσουμε προσεγγιστικά ένα ολοκλήρωμα

$$\int_a^b f(x)W(x)\,\mathrm{d}x$$

όπου $W(x)$ μια μη αρνητική συνάρτηση στο διάστημα $[a,b]$.

Το ολοκλήρωμα αυτό μπορεί να γραφεί ως άθροισμα των τιμών της $f(x)$ σε συγκεκριμένα σημεία $x_i \in (a,b)$ με κατάλληλα βάρη w_i:

$$\int_a^b f(x)W(x)\,\mathrm{d}x \approx \sum_{i=1}^{n} w_i f(x_i)\ . \tag{5.17}$$

Υπάρχει η δυνατότητα να βρούμε[1] μια οικογένεια ορθογώνιων πολυωνύμων $P_i(x)$, βαθμού $i = 0, 1, \ldots$, που ορίζονται στο διάστημα $[a,b]$ και έχουν συνάρτηση βάρους $W(x)$, ικανοποιούν δηλαδή τη σχέση

$$\int_a^b P_i(x)P_j(x)W(x)\,\mathrm{d}x = \delta_{ij}\ .$$

Οι ρίζες του πολυωνύμου $P_n(x)$ είναι τα ζητούμενα σημεία x_i στον τύπο (5.17).

Τα ορθογώνια πολυώνυμα ικανοποιούν τις σχέσεις

$$a_i P_{i-1} + c_i P_{i+1} = (x - b_i)P_i\,, \qquad i > 0\ ,$$

και

$$\begin{aligned} P_0 &= 1\,, \\ c_0 P_1 &= (x - b_0)P_0\ . \end{aligned}$$

Οι αναδρομικές σχέσεις μεταξύ των πολυωνύμων μπορούν να γραφούν με τη μορφή πινάκων:

$$T \cdot \begin{pmatrix} P_0 \\ P_1 \\ P_2 \\ \vdots \\ P_{n-2} \\ P_{n-1} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ P_n \end{pmatrix} = x \begin{pmatrix} P_0 \\ P_1 \\ P_2 \\ \vdots \\ P_{n-2} \\ P_{n-1} \end{pmatrix}\ , \tag{5.18}$$

όπου

$$T = \begin{pmatrix} b_0 & c_0 & 0 & 0 & \cdots & 0 \\ a_1 & b_1 & c_1 & 0 & \cdots & 0 \\ 0 & a_2 & b_2 & c_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & a_{n-2} & b_{n-2} & c_{n-2} \\ 0 & \cdots & 0 & 0 & a_{n-1} & b_{n-1} \end{pmatrix}\ .$$

[1] ή να κατασκευάσουμε με τον αλγόριθμο Gram-Schmidt από τη βάση 1, x, x^2, ...

Αν $P_n(x) = 0$ τότε η (5.18) υποδηλώνει ότι το x είναι ιδιοτιμή του πίνακα T. Με κατάλληλο μετασχηματισμό ομοιότητας ο πίνακας T γίνεται συμμετρικός (και διατηρεί τις ίδιες ιδιοτιμές):

$$J = D^{-1}TD \ ,$$

όπου $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$ με

$$\begin{aligned} d_1 &= 1 \\ d_j &= d_{j-1}\sqrt{\frac{a_j}{c_{j-1}}} \ . \end{aligned}$$

Θεωρούμε ότι $a_j c_{j-1} > 0$, για κάθε $j > 0$.

Από τα παραπάνω συνάγεται ότι οι ρίζες του $P_n(x)$ είναι οι ιδιοτιμές του πίνακα Jacobi, J, ενός συμμετρικού τριδιαγώνιου πίνακα:

$$J = \begin{pmatrix} b_0 & s_1 & 0 & 0 & \cdots & 0 \\ s_1 & b_1 & s_2 & 0 & \cdots & 0 \\ 0 & s_2 & b_2 & s_3 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & s_{n-2} & b_{n-2} & s_{n-1} \\ 0 & \cdots & 0 & 0 & s_{n-1} & b_{n-1} \end{pmatrix} \ ,$$

όπου $s_j = \sqrt{a_j c_{j-1}}$.

Το βάρος w_i που αντιστοιχεί στην ιδιοτιμή x_i στον τύπο (5.17) μπορεί να υπολογιστεί από το αντίστοιχο ιδιοδιάνυσμα του J, $\mathbf{v}^{(i)}$. Αν είναι κανονικοποιημένο ώστε $||\mathbf{v}^{(i)}|| = 1$, τότε

$$w_i = \left(\mathrm{v}_1^{(i)}\right)^2 \int_a^b W(x)\,\mathrm{d}x \ ,$$

όπου $\mathrm{v}_1^{(i)}$ είναι η πρώτη συνιστώσα του $\mathbf{v}^{(i)}$.

Η διαδικασία που περιγράφηκε αποτελεί τον αλγόριθμο Golub–Welsch.

5.7 Μέθοδος Clenshaw–Curtis

Σύμφωνα με τον κανόνα ολοκλήρωσης Clenshaw–Curtis, μπορούμε να υπολογίσουμε ένα ολοκλήρωμα της μορφής

$$\int_{-1}^{1} f(x)\,\mathrm{d}x$$

ως εξής: επιλέγουμε τα $n+1$ (με $n > 1$) μη ισαπέχοντα σημεία

$$x_i = \cos\left(\frac{i\pi}{n}\right) \ , \quad i = 0, \ldots, n$$

στο διάστημα της ολοκλήρωσης. Κατόπιν, βρίσκουμε το πολυώνυμο παρεμβολής που περνά από τα σημεία $(x_i, f(x_i))$, το οποίο ολοκληρώνουμε ακριβώς. Θυμηθείτε ότι το φαινόμενο Runge για τέτοια κατανομή σημείων είναι ελάχιστο.

Μπορεί να δειχθεί ότι στον βασικό τύπο ολοκλήρωσης (5.1),

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \approx \sum_{i=0}^{n} w_i f(x_i)\ ,$$

οι συντελεστές w_i για τη μέθοδο Clenshaw–Curtis είναι

$$w_i = \frac{c_i}{n} \sum_{j=0}^{\lfloor n/2 \rfloor} \frac{b_j}{1-4j^2} \cos\left(\frac{2ij\pi}{n}\right)\ , \quad i = 0, \ldots, n\ ,$$

όπου $\lfloor x \rfloor$ το ακέραιο μέρος του x και

$$b_j = \begin{cases} 1\ , & j = 0 \\ 2\ , & 0 < j < n/2 \\ 1\ , & j = n/2 \end{cases}\ , \qquad c_i = \begin{cases} 1\ , & i = 0 \\ 2\ , & 0 < i < n \\ 1\ , & i = n \end{cases}\ .$$

Οι συντελεστές w_i είναι θετικοί. Μπορεί να αποδειχθεί ότι καθώς ισχύει αυτό, το άθροισμα συγκλίνει στην πραγματική τιμή του ολοκληρώματος όσο αυξάνει το n.

Η μέθοδος Clenshaw–Curtis υπολογίζει το ζητούμενο ολοκλήρωμα με ακρίβεια συγκρίσιμη με τη μέθοδο Gauss–Legendre n σημείων. Έχει πλεονεκτήματα έναντι αυτής ότι

- οι κόμβοι x_i υπολογίζονται εύκολα,
- οι συντελεστές w_i μπορούν να προκύψουν από αλγόριθμους για γρήγορο υπολογισμό του διακριτού μετασχηματισμού Fourier (§6.9),
- οι διαδοχικές εφαρμογές του τύπου για $n, 2n, 4n, \ldots$, που χρειάζονται για την εκτίμηση της ακρίβειάς της, χρησιμοποιούν κοινούς κόμβους.

Παρατήρηση Ορίζουμε το διάνυσμα v, n θέσεων, ως εξής:

$$\begin{aligned} v_k &= \frac{2}{1-4k^2} - \frac{1}{n^2 - 1 + (n \bmod 2)}\ , \text{ για } k = 0, \ldots, \lfloor n/2 \rfloor - 1 \\ v_{\lfloor n/2 \rfloor} &= \frac{n-3}{2\lfloor n/2 \rfloor - 1} - 1 + \frac{1}{n^2 - 1 + (n \bmod 2)}((2 - (n \bmod 2))n - 1) \\ v_{n-k} &= v_k\ , \text{ για } k = 1, \ldots, \lfloor (n-1)/2 \rfloor\ . \end{aligned}$$

Οι συντελεστές w_i, με $i = 0, \ldots, n-1$, προκύπτουν από το διακριτό μετασχηματισμό Fourier (DFT) του διανύσματος v (δείτε την §6.9) και συνεπώς μπορεί να χρησιμοποιηθεί για τον υπολογισμό τους ο αλγόριθμος FFT (§6.9.1). Εύκολα φαίνεται επίσης ότι $w_n = w_0$.

5.8 Εναλλακτικές τεχνικές ολοκλήρωσης

Υπάρχει περίπτωση η ολοκληρωτέα συνάρτηση $f(x)$ να μην είναι δεδομένη ή γνωστή, οπότε δεν μπορούμε να την υπολογίσουμε σε όποια σημεία θέλουμε. Θα έχουμε βέβαια τα σημεία υπολογισμού της, τα οποία όμως μπορεί να μην ισαπέχουν ή να μην έχουν την κατανομή που χρειάζεται για τις μεθόδους Gauss ή Clenshaw–Curtis. Τότε έχουμε διάφορες εναλλακτικές δυνατότητες:

- Χρήση της μεθόδου του τραπεζίου (ή ισοδύναμα, ολοκλήρωση της προσέγγισης με ευθύγραμμα τμήματα, §4.3). Ο (απλός) τύπος του τραπεζίου μπορεί να εφαρμοστεί σε κάθε διάστημα $[x_i, x_{i+1}]$ (με μήκος $h_i = x_{i+1} - x_i$) ώστε να προκύψει ο τύπος:

$$\begin{aligned}\int_{x_0}^{x_N} f(x)\,\mathrm{d}x &= h_0\frac{f(x_0)+f(x_1)}{2} + h_1\frac{f(x_1)+f(x_2)}{2} + \dots \\ &\quad + h_{N-2}\frac{f(x_{N-2})+f(x_{N-1})}{2} + h_{N-1}\frac{f(x_{N-1})+f(x_N)}{2}\,.\end{aligned}$$

  Εάν γειτονικά τμήματα είναι ίσα τότε μπορεί να εφαρμοστεί ένας τύπος Newton–Cotes (π.χ. Simpson) υψηλότερης τάξης.

- Ολοκλήρωση του πολυωνύμου παρεμβολής (§4.1), στο διάστημα ορισμού των δεδομένων. Η μέθοδος αυτή δεν είναι ακριβής για μεγάλο αριθμό σημείων N, εξαιτίας της υψηλής τάξης πολυωνύμου που δημιουργείται, §4.1.2.

- Προσέγγιση της συνάρτησης $f(x)$ με προσαρμογή λόγου πολυωνύμων ή άλλης καμπύλης ή με καμπύλη spline που προσδιορίζεται από τα δεδομένα σημεία, και ολοκλήρωση της προσεγγιστικής συνάρτησης.

- Κατασκευή του τύπου ολοκλήρωσης της μορφής 5.1 με εφαρμογή της μεθόδου που είδαμε στην §5.5 για τα δεδομένα σημεία υπολογισμού της $f(x)$.

5.9 Ασκήσεις

1. (α′) Υλοποιήστε τον αλγόριθμο τραπεζίου σε υποπρόγραμμα. Αυτό θα δέχεται ως ορίσματα τουλάχιστον τα όρια της ολοκλήρωσης και το πλήθος των διαστημάτων. Θα επιστρέφει την προσεγγιστική τιμή του ολοκληρώματος.

   (β′) Χρησιμοποιήστε το υποπρόγραμμα για να υπολογίσετε το ολοκλήρωμα

$$\int_0^{\pi} \sin x\,\mathrm{d}x$$

   διαδοχικά με $N = 2, 4, 8, 16, \dots, 512$ διαστήματα. Το πρόγραμμά σας να τυπώνει για κάθε N την υπολογιζόμενη τιμή και την απόλυτη διαφορά της από την ακριβή τιμή.

2. (α′) Υλοποιήστε τον αλγόριθμο Simpson σε υποπρόγραμμα. Αυτό θα δέχεται ως ορίσματα τουλάχιστον τα όρια της ολοκλήρωσης και το πλήθος των διαστημάτων. Θα επιστρέφει την προσεγγιστική τιμή του ολοκληρώματος.

   (β′) Χρησιμοποιήστε το για να υπολογίσετε το ολοκλήρωμα

$$\int_0^\pi \sin x \, \mathrm{d}x$$

   με όσα διαστήματα χρειάζεται ώστε να έχετε ακρίβεια τουλάχιστον 6 ψηφίων.

   Υπόδειξη: Επιλέξτε κατάλληλα το βήμα (άρα και το πλήθος των διαστημάτων) ώστε το σφάλμα §(5.10) να είναι μικρότερο από 10^{-6}.

3. Υλοποιήστε ένα υποπρόγραμμα που να υπολογίζει ολοκληρώματα ανεξάρτητα με το πλήθος των σημείων στα οποία είναι γνωστή η ολοκληρωτέα συνάρτηση. Αν το πλήθος των διαστημάτων είναι περιττό (και μεγαλύτερο του 3), να χρησιμοποιεί τον τύπο $3/8$ Simpson για τα πρώτα 3 και για τα υπόλοιπα τον τύπο $1/3$ Simpson. Αν είναι άρτιο, να χρησιμοποιεί μόνο τον $1/3$ Simpson.

4. Υπολογίστε προσεγγιστικά με ακρίβεια 10^{-6} τα ολοκληρώματα στο διάστημα $[0, 3]$ των συναρτήσεων

   (α′) $f(x) = 2x + 1$,
   (β′) $f(x) = x^2\sqrt{x}$,
   (γ′) $f(x) = \dfrac{1}{1 + x^2}$,
   (δ′) $f(x) = \dfrac{1}{1 + (x - \pi)^2}$,
   (ε′) $f(x) = \dfrac{1}{2 + \cos x}$,
   (στ′) $f(x) = \cos(4x)\mathrm{e}^x$,
   (ζ′) $f(x) = \mathrm{e}^{\cos x}$,
   (η′) $f(x) = \sqrt{x}$.

5. Έστω

$$f(x) = \begin{cases} -x & -1 \le x \le 0\,, \\ x^2 & 0 \le x \le 1\,, \end{cases}$$

   συνεχής συνάρτηση στο $[-1, 1]$ χωρίς παράγωγο στο $x = 0$. Υπολογίστε το σφάλμα

$$\varepsilon_n = I - I_n = \int_{-1}^{1} f(x)\,\mathrm{d}x - I_n\,,$$

   όπου I_n ο τύπος τραπεζίου με n υποδιαιρέσεις, και δείξτε ότι $|\varepsilon_n| \le Ch$, όπου $h = 2/n$. (Υποθέστε ότι το n είναι άρτιος ή περιττός.)

6. **Παρέκταση Richardson για ολοκληρώματα (Μέθοδος Romberg)**. Μπορεί να δειχθεί ότι ο σύνθετος τύπος τραπεζίου για το ολοκλήρωμα,

$$I_0 = \int_{x_0}^{x_n} f(x)\,\mathrm{d}x \ ,$$

δίνει για την ακριβή τιμή τη σχέση

$$I_0 = I_h + \alpha_2 h^2 + \alpha_4 h^4 + \cdots \ , \tag{5.19}$$

όπου

$$I_h = \frac{h}{2}\left(f_0 + 2f_1 + 2f_2 + \cdots + 2f_{n-1} + f_n\right) \ ,$$

$h = (x_n - x_0)/n$ και α_i οι συντελεστές των όρων h^i του σφάλματος.

Γράψτε τη (5.19) για τρία διαφορετικά βήματα, π.χ. $h, h/2, h/4$. Παρατηρήστε ότι σχηματίζεται ένα σύστημα τριών γραμμικών εξισώσεων με αγνώστους τα I_0, α_2, α_4. Βρείτε τη λύση του συστήματος ως προς I_0· ο τύπος στον οποίο θα καταλήξετε—γραμμικός συνδυασμός των $I_h, I_{h/2}, I_{h/4}$ που έχουν σφάλματα $O(h^2)$—δίνει την ακριβή τιμή του ολοκληρώματος με σφάλμα $O(h^6)$.

[Λύση συστήματος: $I_0 = (I_h - 20I_{h/2} + 64I_{h/4})/45$.]

Υλοποιήστε σε κώδικα τον παραπάνω αλγόριθμο ολοκλήρωσης.

7. Υλοποιήστε σε κώδικα τη μέθοδο ολοκλήρωσης Gauss–Legendre για 2 και για 3 σημεία. Εφαρμόστε τη για να υπολογίσετε το ολοκλήρωμα

$$\int_{2.1}^{5.2} x^3 \mathrm{e}^{-x}\,\mathrm{d}x \ .$$

[Σωστή τιμή ολοκληρώματος: $3.60346\ldots$]

8. Τα πρώτα πολυώνυμα Hermite είναι τα

$$\begin{aligned} H_0(x) = 1 \quad &, \quad H_1(x) = 2x \\ H_2(x) = 4x^2 - 2 \quad &, \quad H_3(x) = 8x^3 - 12x \\ H_4(x) = 16x^4 - 48x^2 + 12 \quad &. \end{aligned}$$

Να γράψετε υποπρόγραμμα που να υλοποιεί τη μέθοδο Gauss–Hermite για $n = 4$. Χρησιμοποιήστε το για να υπολογίσετε το ολοκλήρωμα

$$\int_{-\infty}^{\infty} \mathrm{e}^{-x^2} x^2\,\mathrm{d}x \ .$$

Συγκρίνετε με την ακριβή τιμή $(\sqrt{\pi}/2)$.

9. Τα πρώτα πολυώνυμα Laguerre είναι τα

$$\begin{aligned} L_0(x) &= 1 \\ L_1(x) &= -x+1 \\ L_2(x) &= (x^2-4x+2)/2 \\ L_3(x) &= (-x^3+9x^2-18x+6)/6 \\ L_4(x) &= (x^4-16x^3+72x^2-96x+24)/24 \\ L_5(x) &= (-x^5+25x^4-200x^3+600x^2-600x+120)/120\ . \end{aligned}$$

Να γράψετε υποπρόγραμμα που να υλοποιεί τη μέθοδο Gauss–Laguerre για $n=4$. Χρησιμοποιήστε το για να υπολογίσετε το ολοκλήρωμα

$$\int_0^\infty \mathrm{e}^{-x}(x^6-3\sqrt{x}+2)\,\mathrm{d}x\ .$$

Συγκρίνετε με την ακριβή τιμή $(6!-3\sqrt{\pi}/2+2\times 0!)$.

Υπόδειξη: Το $L_4(x)$ έχει τις 4 ρίζες του πραγματικές στο διάστημα $[0,10]$.

10. Να γράψετε υποπρόγραμμα που να υλοποιεί τη μέθοδο Gauss–Chebyshev για $n=5$. Χρησιμοποιήστε το για να υπολογίσετε το ολοκλήρωμα

$$\int_{-1}^1 \frac{x^2\mathrm{e}^{-x}}{\sqrt{1-x^2}}\,\mathrm{d}x\ .$$

Συγκρίνετε με τη σωστή τιμή $(0.7009067737595233\ldots\times\pi)$.

11. Υπολογίστε με τη μέθοδο Clenshaw–Curtis το ολοκλήρωμα

$$\int_{-2}^2 \frac{1}{1+x^2}\,\mathrm{d}x\ .$$

Πόσα σημεία χρειάζονται για να προσεγγίσετε με 12 ψηφία την ακριβή τιμή $(2\tan^{-1}(2))$;

12. Το ολοκλήρωμα

$$\int_{-1}^1 f(x)\,\mathrm{d}x$$

μπορεί να υπολογιστεί προσεγγιστικά από τύπο της μορφής

$$\int_{-1}^1 f(x)\ \mathrm{d}x \approx \sum_{k=1}^N a_k f(x_k)\ , \tag{5.20}$$

όπου x_1,x_2,...,x_N διακριτά σημεία της επιλογής μας στο διάστημα $[-1,1]$.

Έστω ότι επιλέγουμε να είναι το $N=7$ και τα σημεία x_k τα -0.9, -0.7, -0.4, 0.1, 0.4, 0.8, 0.9. Προσδιορίστε τα a_k $(k=1,\ldots,7)$ ώστε ο τύπος (5.20) να είναι

ακριβής για τις συναρτήσεις $f_0(x) = 1$, $f_1(x) = x$, $f_2(x) = x^2$, ..., $f_6(x) = x^6$. Κατόπιν, χρησιμοποιήστε τον για να υπολογίσετε το ολοκλήρωμα

$$\int_{-1}^{1} x^3 \sin(\pi x)\ \mathrm{d}x\ .$$

Κεφάλαιο 6

Ανάλυση Fourier

Πολλά προβλήματα στη Φυσική αφορούν ταλαντώσεις και κύματα. Ένα ηλεκτρομαγνητικό κύμα (ακτινοβολία), το εναλλασσόμενο ρεύμα σε ένα ηλεκτρικό κύκλωμα, η δόνηση μιας χορδής ή ενός μέσου (ήχος) είναι γενικά μια επαλληλία κυμάτων, το καθένα με συγκεκριμένη συχνότητα. Η ανάλυση Fourier μας δίνει τη δυνατότητα να αναπτύξουμε τέτοιες περιοδικές (αλλά και μη περιοδικές) συναρτήσεις του χρόνου ή κάποιας απόστασης, στα κύματα που τις αποτελούν. Επιπλέον, η ανάλυση Fourier βρίσκει εφαρμογή στη Μαθηματική Φυσική για την επίλυση διαφορικών εξισώσεων.

6.1 Ορισμοί

6.1.1 Συνεχής συνάρτηση

Μια συνάρτηση $f(x)$ είναι *συνεχής* σε ένα σημείο x_0 στο πεδίο ορισμού της αν ικανοποιεί τη σχέση

$$\lim_{x \to x_0} f(x) = f(x_0) \ . \tag{6.1}$$

Σε αυτό τον ορισμό θεωρούμε ότι

- το x_0 δεν είναι απομονωμένο σημείο, έχει δηλαδή, γειτονικά σημεία που ανήκουν στο πεδίο ορισμού, και
- η τιμή του ορίου είναι ανεξάρτητη από την κατεύθυνση από την οποία το σημείο x πλησιάζει το x_0.

Ο συγκεκριμένος ορισμός με απλά λόγια σημαίνει ότι το $f(x)$ πλησιάζει όσο κοντά θέλουμε στο $f(x_0)$ όταν το x πλησιάζει το x_0 από οποιαδήποτε κατεύθυνση. Ισοδύναμα, μια συνάρτηση είναι συνεχής σε σημείο x_0 αν τα δύο όριά της, από μικρότερες και μεγαλύτερες τιμές, ταυτίζονται με την τιμή της στο x_0:

$$\lim_{\varepsilon \to 0} f(x_0 - \varepsilon) = \lim_{\varepsilon \to 0} f(x_0 + \varepsilon) = f(x_0) \ .$$

6.1.2 Περιοδική συνάρτηση

Μια συνεχής συνάρτηση $f(x)$ λέγεται *περιοδική* με (μη μηδενική) περίοδο L, αν ικανοποιεί τη σχέση

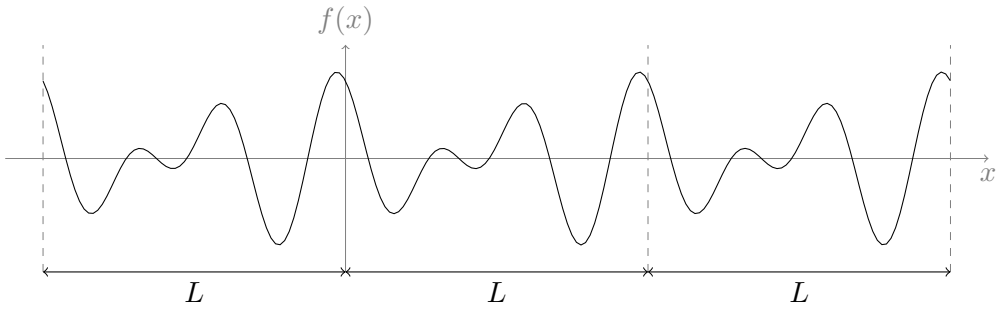
$$f(x+L) = f(x)\ , \tag{6.2}$$

για όλα τα σημεία x που ανήκουν στο πεδίο ορισμού της. Αν το L είναι το μικρότερο διάστημα που ικανοποιεί τη σχέση (6.2), τότε εύκολα δείχνεται ότι κάθε πολλαπλάσιο τού L είναι επίσης περίοδος:

$$f(x+mL) \equiv f(x+(m-1)L+L) = f(x+(m-1)L) = \cdots = f(x)\ ,$$

όπου m θετικός ακέραιος. Παρατηρήστε ακόμα ότι αν μια συνάρτηση έχει περίοδο υποπολλαπλάσιο του L, L/m, τότε το L είναι επίσης περίοδός της.

Παράδειγμα
Η συνάρτηση στο Σχήμα 6.1 είναι περιοδική.



Σχήμα 6.1: Περιοδική συνάρτηση με περίοδο L

6.1.3 Συνθήκες Dirichlet

Μια πραγματική συνάρτηση πραγματικής μεταβλητής $f(x)$, που είναι περιοδική, λέμε ότι ικανοποιεί τις *συνθήκες Dirichlet* αν σε οποιοδήποτε πεπερασμένο διάστημα στο πεδίο ορισμού της:

- Είναι μονότιμη και συνεχής, εκτός ίσως από πεπερασμένο πλήθος διακριτών σημείων στα οποία εμφανίζει ασυνέχεια, χωρίς όμως να απειρίζεται.
- Έχει πεπερασμένο πλήθος μέγιστων και ελάχιστων.
- Ορίζεται και έχει πεπερασμένη τιμή το ολοκλήρωμα τής $|f(x)|$ (όπως λέμε, η $f(x)$ είναι απόλυτα ολοκληρώσιμη).

Οι συνθήκες αυτές είναι πολύ γενικές και οι περιοδικές συναρτήσεις που θα συναντήσουμε σε ρεαλιστικές εφαρμογές τις ικανοποιούν.

Παράδειγμα

Η συνάρτηση στο Σχήμα 6.1 ικανοποιεί τις συνθήκες Dirichlet.

6.2 Σειρά Fourier

Μια περιοδική συνάρτηση $f(x)$ που ικανοποιεί τις συνθήκες Dirichlet μπορεί να αναπαρασταθεί ως άθροισμα άπειρων τριγωνομετρικών συναρτήσεων (ημίτονων και συνημίτονων) με κατάλληλα πλάτη και φάσεις. Το άθροισμα αυτό συγκλίνει στην $f(x)$ σε κάθε σημείο που αυτή είναι συνεχής.

Οι όροι του αθροίσματος είναι της μορφής

$$A_m \cos\left(\frac{2m\pi x}{L}\right) \quad \text{ή} \quad B_m \sin\left(\frac{2m\pi x}{L}\right) ,$$

με m μη αρνητικό ακέραιο[1]. Παρατηρήστε ότι κάθε τέτοιος όρος είναι μια περιοδική συνάρτηση με περίοδο L/m. Το άθροισμα τέτοιων συναρτήσεων, με διάφορα πλάτη A_m, B_m και περιόδους υποπολλαπλάσιες του L (L, $L/2$, $L/3$, ...), αποτελεί τη *σειρά Fourier* για μια συνάρτηση που ικανοποιεί τις συνθήκες Dirichlet και είναι περιοδική με περίοδο L. Δηλαδή

$$\begin{aligned} f(x) &= \frac{A_0}{2} \\ &\quad + A_1 \cos\left(\frac{2\pi x}{L}\right) + B_1 \sin\left(\frac{2\pi x}{L}\right) \\ &\quad + A_2 \cos\left(2\frac{2\pi x}{L}\right) + B_2 \sin\left(2\frac{2\pi x}{L}\right) \\ &\quad + A_3 \cos\left(3\frac{2\pi x}{L}\right) + B_3 \sin\left(3\frac{2\pi x}{L}\right) \\ &\quad \vdots \\ &= \frac{A_0}{2} + \sum_{m=1}^{\infty} A_m \cos\left(\frac{2m\pi x}{L}\right) + \sum_{m=1}^{\infty} B_m \sin\left(\frac{2m\pi x}{L}\right) . \end{aligned} \tag{6.3}$$

Οι συντελεστές A_m, B_m εξαρτώνται από την $f(x)$ και θα υπολογιστούν στην επόμενη παράγραφο. Εκεί θα φανεί και ο λόγος της ιδιαίτερης μορφής του σταθερού όρου, $A_0/2$.

Παρατηρήστε ότι η σειρά Fourier είναι παντού συνεχής ενώ η $f(x)$ μπορεί να έχει σημεία ασυνέχειας. Σε αυτά τα σημεία, η τιμή που παίρνει η σειρά Fourier

[1] Αν η ανεξάρτητη μεταβλητή x είναι μήκος, η περίοδος L λέγεται μήκος κύματος και συμβολίζεται συνήθως με το λ. Η ποσότητα $2\pi/\lambda$ λέγεται (γωνιακός) κυματάριθμος και συμβολίζεται συχνά με το k. Αν η ανεξάρτητη μεταβλητή συμβολίζει χρόνο, η περίοδος L συμβολίζεται με το T. Η ποσότητα $2\pi/T$ λέγεται γωνιακή συχνότητα και συμβολίζεται με το ω.

είναι ο μέσος όρος του δεξιού και του αριστερού ορίου της $f(x)$:

$$\frac{1}{2}\lim_{\varepsilon\to 0}\left(f(x_0-\varepsilon)+f(x_0+\varepsilon)\right)\ .$$

6.3 Υπολογισμός συντελεστών της σειράς Fourier

Οι άγνωστοι συντελεστές $A_0, A_1, \ldots, B_1, \ldots$ στην εξίσωση (6.3) λέγονται *συντελεστές Fourier* και υπολογίζονται ως εξής:

Πολλαπλασιάζουμε τα δύο μέλη της (6.3) με την ποσότητα $\cos(2n\pi x/L)$ και ολοκληρώνουμε σε διάστημα μίας περιόδου:

$$\begin{aligned}\int_0^L \cos\left(\frac{2n\pi x}{L}\right) f(x)\,\mathrm{d}x \;=\; & \frac{A_0}{2}\int_0^L \cos\left(\frac{2n\pi x}{L}\right)\mathrm{d}x \\ & + \sum_{m=1}^{\infty} A_m \int_0^L \cos\left(\frac{2n\pi x}{L}\right)\cos\left(\frac{2m\pi x}{L}\right)\mathrm{d}x \\ & + \sum_{m=1}^{\infty} B_m \int_0^L \cos\left(\frac{2n\pi x}{L}\right)\sin\left(\frac{2m\pi x}{L}\right)\mathrm{d}x\ .\end{aligned}$$

Χρησιμοποιώντας σχέσεις από το τυπολόγιο στο Παράρτημα α΄ προκύπτει ότι

$$\int_0^L \cos\left(\frac{2n\pi x}{L}\right) f(x)\,\mathrm{d}x = A_n\frac{L}{2}\ , \qquad n\geq 0\ .$$

Παρατηρήστε ότι η επιλογή να έχει ο σταθερός όρος τη μορφή $A_0/2$ δίνει ενιαία μορφή στο γενικό τύπο για τα A_n, για κάθε n.

Αντίστοιχα, αν πολλαπλασιάσουμε τα δύο μέλη της (6.3) με την ποσότητα $\sin(2n\pi x/L)$ και ολοκληρώσουμε στο διάστημα $[0, L]$, έχουμε

$$\begin{aligned}\int_0^L \sin\left(\frac{2n\pi x}{L}\right) f(x)\,\mathrm{d}x \;=\; & \frac{A_0}{2}\int_0^L \sin\left(\frac{2n\pi x}{L}\right)\mathrm{d}x \\ & + \sum_{m=1}^{\infty} A_m \int_0^L \sin\left(\frac{2n\pi x}{L}\right)\cos\left(\frac{2m\pi x}{L}\right)\mathrm{d}x \\ & + \sum_{m=1}^{\infty} B_m \int_0^L \sin\left(\frac{2n\pi x}{L}\right)\sin\left(\frac{2m\pi x}{L}\right)\mathrm{d}x\ .\end{aligned}$$

Χρησιμοποιώντας σχέσεις από το τυπολόγιο στο Παράρτημα α΄ προκύπτει ότι

$$\int_0^L \sin\left(\frac{2n\pi x}{L}\right) f(x)\,\mathrm{d}x = B_n\frac{L}{2}\ , \qquad n> 0\ .$$

Συγκεντρωτικά, για τους πραγματικούς συντελεστές A_n, B_n έχουμε

$$A_n = \frac{2}{L}\int_0^L \cos\left(\frac{2n\pi x}{L}\right) f(x)\,\mathrm{d}x\,, \qquad n \geq 0\,, \tag{6.4α΄}$$

$$B_n = \frac{2}{L}\int_0^L \sin\left(\frac{2n\pi x}{L}\right) f(x)\,\mathrm{d}x\,, \qquad n > 0\,. \tag{6.4β΄}$$

Παρατήρηση: Οι συναρτήσεις $f(x)$, $\cos(2n\pi x/L)$, $\sin(2n\pi x/L)$ που εμφανίζονται στα ολοκληρώματα είναι περιοδικές με περίοδο L:

$$\begin{aligned} f(x+L) &= f(x)\,, \qquad \text{(εξ ορισμού)} \\ \cos(2n\pi(x+L)/L) &= \cos(2n\pi x/L + 2n\pi) = \cos(2n\pi x/L)\,, \\ \sin(2n\pi(x+L)/L) &= \sin(2n\pi x/L + 2n\pi) = \sin(2n\pi x/L)\,. \end{aligned}$$

Επομένως και οι συναρτήσεις $\cos(2n\pi x/L)f(x)$ και $\sin(2n\pi x/L)f(x)$ είναι περιοδικές με την ίδια περίοδο.

Ένα ολοκλήρωμα μιας περιοδικής συνάρτησης $g(x)$ σε μήκος ίσο με την περίοδό της, L, είναι το ίδιο, ανεξάρτητα από την αρχή (το κάτω όριο) της ολοκλήρωσης: έστω a αυθαίρετο σημείο στο πεδίο ορισμού της συνάρτησης. Τότε

$$\int_a^{a+L} g(x)\,\mathrm{d}x = \int_a^L g(x)\,\mathrm{d}x + \int_L^{a+L} g(x)\,\mathrm{d}x\,.$$

Όμως

$$\int_L^{a+L} g(x)\,\mathrm{d}x \overset{y=x-L}{=} \int_0^a g(y+L)\,\mathrm{d}y = \int_0^a g(y)\,\mathrm{d}y\,.$$

Η τελευταία ισότητα προέκυψε από την περιοδικότητα (εξίσωση (6.2)) της $g(x)$. Συνολικά έχουμε

$$\int_a^{a+L} g(x)\,\mathrm{d}x = \int_a^L g(x)\,\mathrm{d}x + \int_0^a g(y)\,\mathrm{d}y = \int_0^L g(x)\,\mathrm{d}x\,.$$

Επομένως, σε ολοκλήρωμα περιοδικής συνάρτησης σε μία περίοδό της μπορούμε να επιλέξουμε το διάστημα ολοκλήρωσης να είναι $[0, L]$ ή $[-L/2, L/2]$ ή οποιοδήποτε άλλο μας διευκολύνει, αρκεί να έχει μήκος μία περίοδο. Η τιμή του ολοκληρώματος θα είναι η ίδια ανεξάρτητα από την αρχή του διαστήματος.

6.3.1 Ιδιότητες

Εύκολα μπορούμε να δούμε από τους ορισμούς των συντελεστών Fourier, (6.4α΄) και (6.4β΄), ότι:

- Αν η συνάρτηση $f(x)$ ικανοποιεί τις συνθήκες Dirichlet, έχει περίοδο L και συντελεστές Fourier A_m, B_m, η συνάρτηση $\lambda f(x)$ με λ πραγματική σταθερά, ικανοποιεί επίσης τις συνθήκες Dirichlet, έχει περίοδο L και αναπτύσσεται σε σειρά Fourier με συντελεστές λA_m, λB_m.

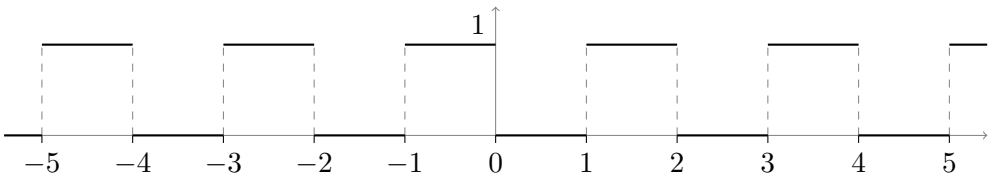
- Αν δύο συναρτήσεις $f_1(x)$ και $f_2(x)$ ικανοποιούν τις συνθήκες Dirichlet, έχουν *ίδια περίοδο* L και συντελεστές Fourier $A_m^{(1)}$, $B_m^{(1)}$ και $A_m^{(2)}$, $B_m^{(2)}$ αντίστοιχα, τότε και το άθροισμά τους, $f_1(x)+f_2(x)$, ικανοποιεί τις συνθήκες Dirichlet, έχει ίδια περίοδο L και συντελεστές Fourier $A_m = A_m^{(1)} + A_m^{(2)}$ και $B_m = B_m^{(1)} + B_m^{(2)}$.

6.3.2 Παράδειγμα

Έστω η συνάρτηση (τετραγωνικός παλμός)

$$f(x) = \begin{cases} 0\,, & 0 \le x < 1\,, \\ 1\,, & 1 \le x < 2\,, \end{cases} \tag{6.5}$$

που επαναλαμβάνεται για $x \ge 2$ και $x < 0$ ώστε $f(x+2m) = f(x)$ με οποιοδήποτε θετικό η αρνητικό ακέραιο m. Η γραφική της παράσταση δίνεται στο Σχήμα 6.2.



Σχήμα 6.2: Τετραγωνικός παλμός, εξίσωση (6.5)

Η συγκεκριμένη συνάρτηση είναι συνεχής με πεπερασμένες ασυνέχειες στα σημεία 0, 1 και είναι περιοδική με περίοδο $L = 2$. Το ολοκλήρωμα της $|f(x)|$ στο $[0, 2)$ ορίζεται και είναι πεπερασμένο:

$$\int_0^2 |f(x)|\,\mathrm{d}x = 1\,.$$

Η $f(x)$ ικανοποιεί τις συνθήκες Dirichlet οπότε μπορεί να αναπτυχθεί σε σειρά Fourier:

$$f(x) = \frac{A_0}{2} + \sum_{m=1}^{\infty} A_m \cos(m\pi x) + \sum_{m=1}^{\infty} B_m \sin(m\pi x)\,,$$

με συντελεστές που υπολογίζονται από τους τύπους (6.4α΄), (6.4β΄):

$$\begin{aligned} A_m &= \int_0^2 \cos(m\pi x) f(x)\,\mathrm{d}x = \int_1^2 \cos(m\pi x)\,\mathrm{d}x = \delta_{m0}\,, \quad m \ge 0\,, \\ B_m &= \int_0^2 \sin(m\pi x) f(x)\,\mathrm{d}x = \int_1^2 \sin(m\pi x)\,\mathrm{d}x = \frac{(-1)^m - 1}{m\pi} \\ &= \begin{cases} 0\,, & m = 2, 4, 6, \ldots\,, \\ -\frac{2}{m\pi}\,, & m = 1, 3, 5, \ldots\,. \end{cases} \end{aligned}$$

Στο A_m χρησιμοποιήθηκε το *δέλτα του Kronecker*, δ_{ij}. Αυτό έχει τιμή 1 αν $i = j$ και 0 αν $i \ne j$.

Επομένως, ο συγκεκριμένος τετραγωνικός παλμός αναπτύσσεται σε σειρά Fourier ως εξής

$$\begin{aligned} f(x) &= \frac{1}{2} + \sum_{m=1}^{\infty} B_m \sin(k\pi x) = \frac{1}{2} - 2\sum_{j=0}^{\infty} \frac{\sin\left((2j+1)\pi x\right)}{(2j+1)\pi} \\ &= \frac{1}{2} - 2\left(\frac{\sin(\pi x)}{\pi} + \frac{\sin(3\pi x)}{3\pi} + \frac{\sin(5\pi x)}{5\pi} + \frac{\sin(7\pi x)}{7\pi} + \ldots\right) . \end{aligned} \tag{6.6}$$

Παρατηρήστε ότι στα σημεία ασυνέχειας της $f(x)$, στα 0 και 1, η σειρά Fourier είναι συνεχής και έχει τιμή $1/2$, όσο το ημιάθροισμα του δεξιού και του αριστερού ορίου της $f(x)$ σε καθένα από τα δύο σημεία.

6.3.3 Συντελεστές Fourier συνάρτησης με συμμετρία

Παρατηρήστε ότι η συνάρτηση $c(x) = \cos(2n\pi x/L)$ στην εξίσωση (6.4α′) είναι συμμετρική ως προς το $x = L/2$: αν $\bar{x}$ ένα οποιοδήποτε μήκος, ισχύει

$$c\left(\frac{L}{2} - \bar{x}\right) = c\left(\frac{L}{2} + \bar{x}\right) .$$

Αντίστοιχα, η συνάρτηση $s(x) = \sin(2n\pi x/L)$ της εξίσωσης (6.4β′) είναι αντισυμμετρική ως προς το $x = L/2$:

$$s\left(\frac{L}{2} - \bar{x}\right) = -s\left(\frac{L}{2} + \bar{x}\right) .$$

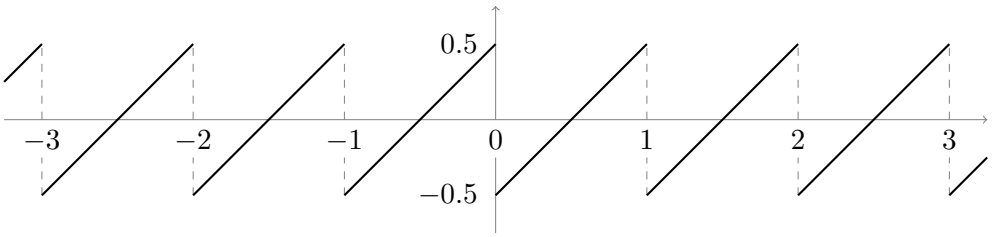
Επομένως, ο υπολογισμός των ολοκληρωμάτων που εκφράζουν τους συντελεστές Fourier απλοποιείται όταν η συνάρτηση $f(x)$ είναι συμμετρική ή αντισυμμετρική ως προς το $x = L/2$ (ή γενικότερα, το μέσο οποιουδήποτε διαστήματος με μήκος L). Αν είναι συμμετρική, τότε $B_n = 0$ και η σειρά Fourier περιέχει μόνο συνημίτονα (και σταθερό όρο), δηλαδή, τους συμμετρικούς όρους. Αν είναι αντισυμμετρική, έχουμε $A_n = 0$ και η σειρά περιέχει μόνο ημίτονα, δηλαδή τους αντισυμμετρικούς όρους της. Στην περίπτωση που η $f(x)$ δεν έχει συγκεκριμένη συμμετρία ως προς το $L/2$, η σειρά περιλαμβάνει γενικά όλους τους όρους.

Με βάση τα παραπάνω, η σειρά Fourier του τετραγωνικού παλμού στην εξίσωση (6.5), που δεν παρουσιάζει κάποια συμμετρία ως προς τη μέση του διαστήματος $[0, 2)$, είναι αναμενόμενο να μην έχει μόνο όρους συγκεκριμένης συμμετρίας. Προσέξτε όμως ότι η συνάρτηση $g(x) = f(x) - 1/2$ είναι αντισυμμετρική σε αυτό το διάστημα, γύρω από το $x = 1$. Το ανάπτυγμα Fourier της $g(x)$ αναμένουμε να έχει μόνο τους όρους των ημιτόνων. Εύκολα επιβεβαιώνεται αυτό από το ανάπτυγμα της $f(x)$.

Γενικά, μια κατάλληλη μετατόπιση της συνάρτησης ή της ανεξάρτητης μεταβλητής κατά ένα σταθερό όρο μπορεί να αναδείξει τη συμμετρία της συνάρτησης, αν υπάρχει, και επομένως να απλοποιήσει τη σειρά Fourier.

6.3.4 Παράδειγμα

Ας δούμε άλλο παράδειγμα με εξαρχής αντισυμμετρική συνάρτηση. Έστω η συνάρτηση στο Σχήμα 6.3 (πριονωτός παλμός)



Σχήμα 6.3: Πριονωτός παλμός, εξίσωση (6.7)

$$f(x) = (x \bmod 1) - \frac{1}{2} \,. \tag{6.7}$$

Η έκφραση $x \bmod 1$ σημαίνει ότι προσθέτουμε ή αφαιρούμε το 1 στο x όσες φορές χρειάζεται ώστε το αποτέλεσμα να είναι στο διάστημα $[0, 1)$.

Η συνάρτηση είναι περιοδική με περίοδο $L = 1$, συνεχής με πεπερασμένες ασυνέχειες στα σημεία $x = m$ ($m = 0, \pm 1, \pm 2, \ldots$). Το ολοκλήρωμα της $|f(x)|$ στο $[0, 1)$ ορίζεται και είναι πεπερασμένο:

$$\int_0^1 |f(x)| \, \mathrm{d}x = \frac{1}{4} \,.$$

Η $f(x)$ ικανοποιεί τις συνθήκες Dirichlet οπότε μπορεί να αναπτυχθεί σε σειρά Fourier. Επιπλέον, είναι αντισυμμετρική ως προς το $x = 1/2$ (το μέσο μιας περιόδου) οπότε η σειρά Fourier έχει μόνο τους όρους των ημιτόνων:

$$f(x) = \sum_{m=1}^{\infty} B_m \sin(2m\pi x) \,.$$

Οι συντελεστές B_m υπολογίζονται από τον τύπο (6.4β′):

$$B_m = 2 \int_0^1 \sin(2m\pi x) \left(x - \frac{1}{2} \right) \mathrm{d}x = -\frac{1}{m\pi} \,, \qquad m > 0 \,. \tag{6.8}$$

Οι συντελεστές A_m είναι 0.

Σύμφωνα με τα παραπάνω, ο συγκεκριμένος πριονωτός παλμός αναπτύσσεται σε σειρά Fourier ως εξής

$$f(x) = -\sum_{m=1}^{\infty} \frac{\sin(2m\pi x)}{m\pi} \,. \tag{6.9}$$

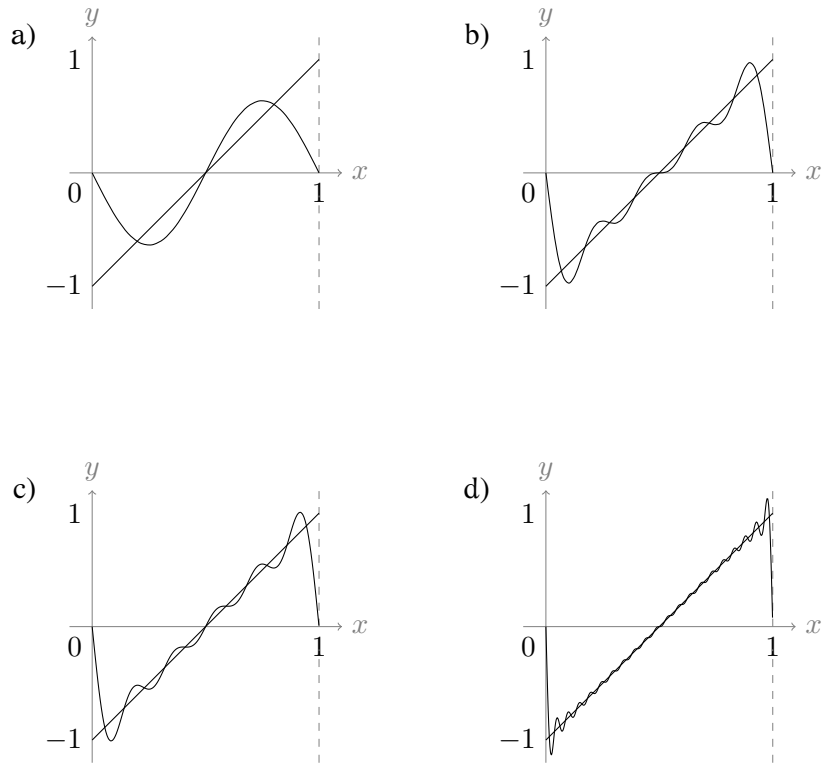
Παρατηρήστε ότι όταν το x παίρνει ακέραιες τιμές, δηλαδή στα σημεία ασυνέχειας της $f(x)$, η σειρά Fourier είναι συνεχής και έχει τιμή 0, όσο το ημιάθροισμα του δεξιού και του αριστερού ορίου της $f(x)$ σε αυτά τα σημεία.

6.4 Φαινόμενο Gibbs

Στο Σχήμα 6.4 βλέπουμε πώς προσεγγίζει τον τριγωνικό παλμό της εξίσωσης (6.7) η σειρά Fourier της εξίσωσης (6.9) με M όρους, δηλαδή το μερικό άθροισμα

$$f_M(x) = -\sum_{m=1}^{M} \frac{\sin(2m\pi x)}{m\pi} \,, \tag{6.10}$$

για διάφορες τιμές του M. Παρατηρήστε ότι η σειρά Fourier παρουσιάζει ταλαντώσεις πλησιάζοντας σε σημείο ασυνέχειας της συνάρτησης, ανεξάρτητα από το πλήθος των όρων που θα λάβουμε υπόψη. Αυτή η συμπεριφορά είναι ανεξάρτητη από τη συνάρτηση που προσεγγίζεται και ονομάζεται φαινόμενο Gibbs. Λόγω των ταλαντώσεων η σειρά μπορεί να ξεπεράσει τη μέγιστη/ελάχιστη τιμή της συνάρτησης και να παρουσιάσει «αφύσικο» ακρότατο.



Σχήμα 6.4: Γραφική παράσταση της συνάρτησης $f_M(x)$ (εξίσωση (6.10)) για a) $M = 1$, b) $M = 4$, c) $M = 5$, d) $M = 20$

6.5 Παραγωγή σειράς Fourier από άλλη

6.5.1 Ολοκλήρωση

Ας υπολογίσουμε το αόριστο ολοκλήρωμα της σειράς Fourier μιας συνάρτησης $f(x)$:

$$\begin{aligned}\int f(x)\,\mathrm{d}x + c &= \int \frac{A_0}{2}\,\mathrm{d}x + \sum_{m=1}^{\infty} A_m \int \cos\left(\frac{2m\pi x}{L}\right)\,\mathrm{d}x + \sum_{m=1}^{\infty} B_m \int \sin\left(\frac{2m\pi x}{L}\right)\,\mathrm{d}x \\ &= \frac{A_0}{2}x + \sum_{m=1}^{\infty} A_m \frac{L}{2m\pi}\sin\left(\frac{2m\pi x}{L}\right) - \sum_{m=1}^{\infty} B_m \frac{L}{2m\pi}\cos\left(\frac{2m\pi x}{L}\right)\,.\end{aligned}$$

Η ποσότητα c είναι η σταθερά ολοκλήρωσης. Παρατηρήστε ότι το δεξί μέλος περιέχει τον όρο $A_0x/2$ που, αν $A_0 \neq 0$, δεν είναι περιοδικός. Η σειρά στο δεξί μέλος δεν αποτελεί σειρά Fourier. Όμως

$$\int f(x)\,\mathrm{d}x - \frac{A_0}{2}x + c = \sum_{m=1}^{\infty} A_m \frac{L}{2m\pi}\sin\left(\frac{2m\pi x}{L}\right) - \sum_{m=1}^{\infty} B_m \frac{L}{2m\pi}\cos\left(\frac{2m\pi x}{L}\right)\,.$$

Από την παραπάνω σχέση προκύπτει η σειρά Fourier της

$$F(x) = \int \left(f(x) - \frac{A_0}{2}\right)\,\mathrm{d}x$$

με προσέγγιση μιας προσθετικής σταθεράς c, η οποία μπορεί να προσδιοριστεί από τις τιμές της $F(x)$ και της σειράς Fourier σε συγκεκριμένο σημείο.

Παράδειγμα

Το αόριστο ολοκλήρωμα του τετραγωνικού παλμού, (6.5), είναι

$$\int f(x)\,\mathrm{d}x = \begin{cases} 1\,, & 0 \le x < 1\,, \\ x\,, & 1 \le x < 2\,, \end{cases}$$

με επανάληψη έξω από το διάστημα $[0, 2)$ ή γενικά

$$\int f(x)\,\mathrm{d}x = \begin{cases} 1\,, & 0 \le x - 2m < 1\,, \\ x - 2m\,, & 1 \le x - 2m < 2\,, \end{cases}$$

για το διάστημα $[2m, 2m + 2)$ με οποιοδήποτε ακέραιο m.

Το αόριστο ολοκλήρωμα της αντίστοιχης σειράς Fourier, (6.6), είναι

$$\int \left[\frac{1}{2} - 2\sum_{j=0}^{\infty} \frac{\sin\left((2j+1)\pi x\right)}{(2j+1)\pi}\right]\,\mathrm{d}x = \frac{x}{2} + 2\sum_{j=0}^{\infty} \frac{\cos\left((2j+1)\pi x\right)}{((2j+1)\pi)^2}\,.$$

Τα δύο ολοκληρώματα διαφέρουν κατά μια προσθετική σταθερά. Η τιμή του πρώτου στο $x = 1/2$ είναι $1\cdot$ η τιμή του δεύτερου είναι $1/4$. Άρα

$$\int f(x)\,\mathrm{d}x - \frac{3}{4} = \frac{x}{2} + 2\sum_{j=0}^{\infty} \frac{\cos\left((2j+1)\pi x\right)}{((2j+1)\pi)^2}\ .$$

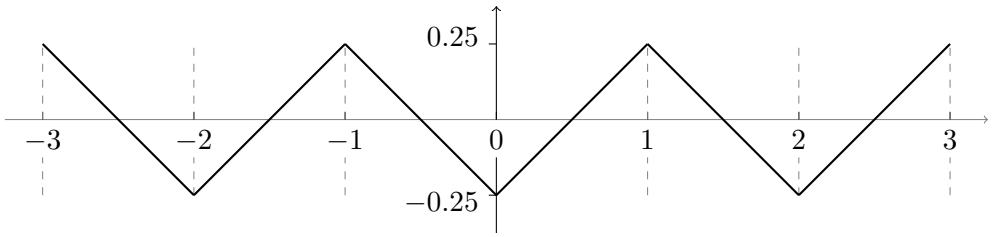
Συνεπώς, η σειρά Fourier της συνάρτησης

$$\int f(x)\,\mathrm{d}x - \frac{3}{4} - \frac{x}{2} = \begin{cases} (1-2x)/4\ , & 0 \le x < 1\ , \\ (2x-3)/4\ , & 1 \le x < 2\ , \end{cases} \tag{6.11}$$

που επαναλαμβάνεται περιοδικά και δίνεται γραφικά στο Σχήμα 6.5, είναι

$$2\sum_{j=0}^{\infty} \frac{\cos\left((2j+1)\pi x\right)}{((2j+1)\pi)^2}\ .$$

Μπορείτε να το επαληθεύσετε υπολογίζοντας τη σειρά Fourier της (6.11);



Σχήμα 6.5: Τριγωνικός παλμός, εξίσωση (6.11)

6.5.2 Παραγώγιση

Αν έχουμε υπολογίσει τη σειρά Fourier, (6.3), για μια περιοδική συνάρτηση $f(x)$ με περίοδο L, μπορούμε να την παραγωγίσουμε ως προς x, όρο–όρο:

$$-\sum_{m=1}^{\infty} A_m \frac{2m\pi}{L} \sin\left(\frac{2m\pi x}{L}\right) + \sum_{m=1}^{\infty} B_m \frac{2m\pi}{L} \cos\left(\frac{2m\pi x}{L}\right)\ .$$

Η σειρά που προκύπτει, περιέχει στη γενική περίπτωση, άπειρους όρους ημιτόνων και συνημιτόνων κατάλληλης μορφής, αποτελεί τη σειρά Fourier της $f'(x)$ και θα συγκλίνει σε αυτή, *με την προϋπόθεση ότι η παράγωγος είναι περιοδική κα ικανοποιεί τις συνθήκες Dirichlet*. Θα δούμε ένα παράδειγμα σε επόμενη παράγραφο.

6.6 Εναλλακτική θεώρηση της σειράς Fourier

Το σύνολο των συναρτήσεων μίας πραγματικής μεταβλητής, που είναι περιοδικές με περίοδο L και ικανοποιούν τις συνθήκες Dirichlet, εφοδιασμένο με τις

γνωστές πράξεις της πρόσθεσης συναρτήσεων και του πολλαπλασιασμού αριθμού με συνάρτηση, αποτελεί ένα διανυσματικό χώρο. Σε αυτόν, μπορούμε να ορίσουμε το εσωτερικό γινόμενο δύο διανυσμάτων $\langle \boldsymbol{f}|\boldsymbol{g}\rangle$ ως εξής:

$$\langle \boldsymbol{f}|\boldsymbol{g}\rangle \equiv \frac{2}{L}\int_0^L f(x)^* g(x)\,\mathrm{d}x\ .$$

Μια βάση του χώρου αυτού, στην περίπτωση που οι συναρτήσεις είναι πραγματικές, είναι το σύνολο

$$\left\{\frac{1}{\sqrt{2}}, \cos\left(\frac{2\pi x}{L}\right), \sin\left(\frac{2\pi x}{L}\right), \cos\left(\frac{2\pi 2x}{L}\right), \sin\left(\frac{2\pi 2x}{L}\right), \ldots,\right.$$
$$\left.\cos\left(\frac{2\pi m x}{L}\right), \sin\left(\frac{2\pi m x}{L}\right), \ldots\right\}$$

με άπειρα διανύσματα που είναι γραμμικώς ανεξάρτητα. Με τη βοήθεια των ολοκληρωμάτων στο Παράρτημα α΄ προκύπτει ότι η βάση είναι ορθοκανονική. Οι συντελεστές του αναπτύγματος οποιουδήποτε μέλους του διανυσματικού χώρου σε αυτή τη βάση δίνονται από τις σχέσεις

$$\begin{aligned} a_0 &= \langle \frac{1}{\sqrt{2}}|\boldsymbol{f}\rangle\ , \\ a_m &= \langle \cos\left(\frac{2m\pi x}{L}\right)|\boldsymbol{f}\rangle\ , \quad m>0\ , \\ b_m &= \langle \sin\left(\frac{2m\pi x}{L}\right)|\boldsymbol{f}\rangle\ , \quad m>0\ . \end{aligned}$$

Το ανάπτυγμα είναι

$$|\boldsymbol{f}\rangle = a_0\frac{1}{\sqrt{2}} + \sum_{m=1}^{\infty} a_m \cos\left(\frac{2m\pi x}{L}\right) + \sum_{m=1}^{\infty} b_m \sin\left(\frac{2m\pi x}{L}\right)\ ,$$

δηλαδή ουσιαστικά η σειρά Fourier: οι συντελεστές της είναι $A_0=\sqrt{2}a_0$, $A_m=a_m$, $B_m=b_m$, $m\geq 1$.

6.6.1 Ταυτότητα Parseval

Το εσωτερικό γινόμενο δύο περιοδικών συναρτήσεων $|\boldsymbol{f^{(1)}}\rangle$ και $|\boldsymbol{f^{(2)}}\rangle$, που είναι μέλη του χώρου, μπορεί να εκφραστεί με τους συντελεστές αυτών στην ορθοκανονική βάση, $\{a_0^{(1)}, a_1^{(1)}, \ldots, b_1^{(1)}, \ldots\}$ και $\{a_0^{(2)}, a_1^{(2)}, \ldots, b_1^{(2)}, \ldots\}$:

$$\begin{aligned} \langle \boldsymbol{f^{(1)}}|\boldsymbol{f^{(2)}}\rangle &= a_0^{(1)}a_0^{(2)} + \sum_{m=1}^{\infty}\left(a_m^{(1)}a_m^{(2)} + b_m^{(1)}b_m^{(2)}\right) \\ &= \frac{1}{2}A_0^{(1)}A_0^{(2)} + \sum_{m=1}^{\infty}\left(A_m^{(1)}A_m^{(2)} + B_m^{(1)}B_m^{(2)}\right)\ . \end{aligned}$$

Στην τελευταία έκφραση εμφανίζονται οι συντελεστές Fourier των δύο συναρτήσεων.

Αν $|\boldsymbol{f^{(1)}}\rangle = |\boldsymbol{f^{(2)}}\rangle = |\boldsymbol{f}\rangle$ η τελευταία σχέση γίνεται

$$\frac{2}{L}\int_0^L f(x)^2\,\mathrm{d}x = \frac{A_0^2}{2} + \sum_{m=1}^{\infty}\left(A_m^2 + B_m^2\right)\ . \tag{6.12}$$

Η εξίσωση αυτή είναι η *ταυτότητα Parseval*.

Η *μέση τιμή* μιας συνάρτησης $f(x)$ σε ένα διάστημα $[a,b]$ συμβολίζεται με $\langle f\rangle$ (ή με $\bar{f}$) και ορίζεται από τη σχέση

$$\langle f\rangle \equiv \frac{\displaystyle\int_a^b f(x)\,\mathrm{d}x}{\displaystyle\int_a^b \mathrm{d}x} = \frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x\ .$$

Σύμφωνα με την ταυτότητα Parseval, η μέση τιμή της συνάρτησης $f(x)^2$ στο $[0,L]$ μπορεί να υπολογιστεί από τους συντελεστές Fourier της $f(x)$:

$$\langle f^2\rangle \equiv \frac{1}{L}\int_0^L f(x)^2\,\mathrm{d}x = \frac{A_0^2}{4} + \frac{1}{2}\sum_{m=1}^{\infty}\left(A_m^2 + B_m^2\right)\ .$$

Παράδειγμα

Ας υπολογίσουμε την ταυτότητα Parseval για τη συνάρτηση $f(x) = x - 1/2$ που ορίζεται στο διάστημα $[0,1)$ και επαναλαμβάνεται περιοδικά έξω από αυτό, Σχήμα 6.3, με περίοδο $L=1$, δηλαδή του πριονωτού παλμού (εξίσωση (6.7)).

Το αριστερό μέλος της εξίσωσης (6.12) δίνει

$$\frac{2}{L}\int_0^L f(x)^2\,\mathrm{d}x = 2\int_0^1\left(x-\frac{1}{2}\right)^2\mathrm{d}x = \frac{1}{6}\ .$$

Βρήκαμε προηγουμένως τους συντελεστές Fourier για τη συγκεκριμένη συνάρτηση (εξίσωση (6.8)). Το άθροισμα στο δεξί μέλος της εξίσωσης (6.12) είναι

$$\frac{A_0^2}{2} + \sum_{m=1}^{\infty}\left(A_m^2 + B_m^2\right) = \sum_{m=1}^{\infty}\frac{1}{m^2\pi^2}\ .$$

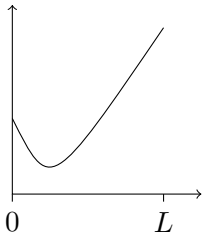
Η ταυτότητα Parseval γίνεται

$$\frac{1}{6} = \frac{1}{\pi^2}\sum_{m=1}^{\infty}\frac{1}{m^2} \Rightarrow \sum_{m=1}^{\infty}\frac{1}{m^2} = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots = \frac{\pi^2}{6}\ .$$

6.7 Σειρά Fourier για συναρτήσεις σε πεπερασμένο διάστημα

Η συνάρτηση $f(x)$ δεν είναι απαραίτητο να είναι περιοδική για να αναπτυχθεί σε σειρά Fourier· μπορεί να ορίζεται και να ικανοποιεί τις συνθήκες Dirichlet σε ένα *πεπερασμένο* διάστημα μήκους L και να την επεκτείνουμε πέρα από αυτό. Κατόπιν, μπορούμε να υπολογίσουμε το ανάπτυγμα Fourier.

Ας εξετάσουμε τη συνάρτηση στο Σχήμα 6.6. Θα παρουσιάσουμε τις δυνατότητες

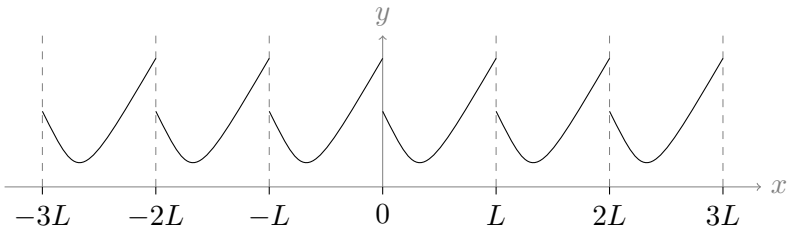


Σχήμα 6.6: Μη περιοδική συνάρτηση στο $[0, L)$

που έχουμε για την επέκτασή της έξω από το πεδίο ορισμού της.

6.7.1 Μετατόπιση

Μπορούμε να την επαναλάβουμε αυτούσια στα διαστήματα $[L, 2L)$, $[2L, 3L)$, κλπ. όπως και στα $[-L, 0)$, $[-2L, L)$, κλπ. όπως στο Σχήμα 6.7. Δημιουργούμε επο-

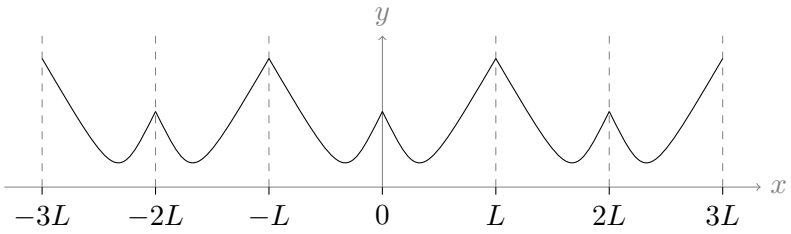


Σχήμα 6.7: Επέκταση μη περιοδικής συνάρτησης με επανάληψη

μένως μια νέα συνάρτηση που είναι περιοδική με περίοδο L. Προσέξτε ότι αν η μη περιοδική συνάρτηση έχει διαφορετικές τιμές στα άκρα του διαστήματος 0, L, η επέκτασή της με αυτό τον τρόπο δημιουργεί σημεία πεπερασμένης ασυνέχειας (τα 0, $\pm L$, $\pm 2L$ ...). Η σειρά Fourier ορίζεται για τη νέα συνάρτηση και συγκλίνει στην αρχική μας στο πεδίο ορισμού της, το $[0, L)$, εκτός από τα σημεία ασυνέχειας σε αυτό (0 και L). Επομένως, αν η μη περιοδική συνάρτηση $f(x)$ που ορίζεται στο $[0, L)$ ικανοποιεί τη σχέση $\lim_{x\to L} f(x) \neq f(0)$, καλό είναι να αποφεύγουμε αυτό τον τρόπο επέκτασης.

6.7.2 Κατοπτρισμός ως προς ευθείες

Δεύτερος τρόπος επέκτασης μιας μη περιοδικής συνάρτησης $f(x)$ είναι με κατοπτρισμό ως προς τις ευθείες $x = 0$, $x = \pm L$, $x = \pm 2L$ κλπ., δηλαδή, στο διάστημα $[-L, 0)$ έχουμε $f(x) = f(-x)$, στο διάστημα $[L, 2L)$ έχουμε $f(x) = f(2L - x)$, κλπ., όπως στο Σχήμα 6.8: Με αυτή την επιλογή επέκτασης δημιουργούμε μια νέα συ-

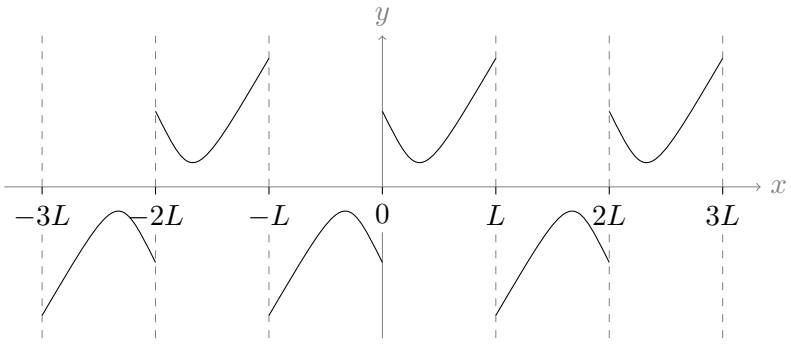


Σχήμα 6.8: Επέκταση μη περιοδικής συνάρτησης με κατοπτρισμό

νάρτηση που είναι περιοδική με περίοδο $2L$ και δεν εισάγουμε σημεία ασυνέχειας. Επιπλέον, η νέα συνάρτηση είναι συμμετρική, οπότε η σειρά Fourier για αυτή δεν θα περιλαμβάνει όρους με ημίτονα. Το ανάπτυγμα Fourier που θα προκύψει, συγκλίνει στην αρχική μας συνάρτηση στο διάστημα $[0, L)$.

6.7.3 Κατοπτρισμός ως προς σημεία

Τρίτος τρόπος επέκτασης μιας μη περιοδικής συνάρτησης $f(x)$ είναι με κατοπτρισμό ως προς τα σημεία $(0, 0)$, $(\pm L, 0)$, $(\pm 2L, 0)$, κλπ. Αυτό σημαίνει ότι στο διάστημα $[-L, 0)$ θέτουμε $f(x) = -f(-x)$ και σε οποιοδήποτε άλλο σημείο έξω από το $[-L, L)$ θέτουμε $f(x + 2L) = f(x)$. Με αυτή την επιλογή επέκτασης δημιουργούμε μια νέα συνάρτηση που είναι περιοδική με περίοδο $2L$ (Σχήμα 6.9) και αντισυμμετρική. Παρατηρήστε ότι αν η $f(x)$ στα άκρα του διαστήματος ορισμού



Σχήμα 6.9: Επέκταση μη περιοδικής συνάρτησης με ανάκλαση ως προς σημεία

της δεν έχει τιμή (ή όριο) το 0, ο συγκεκριμένος τρόπος επέκτασής της δημιουργεί σημεία πεπερασμένης ασυνέχειας. Η σειρά Fourier της περιοδικής επέκτασης θα

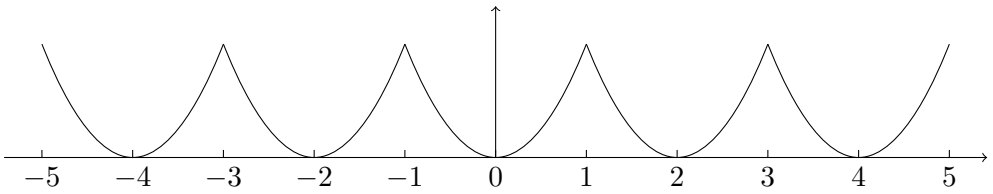
περιέχει μόνο όρους με ημίτονα και θα συγκλίνει στην $f(x)$ στο πεδίο ορισμού της εκτός από τα σημεία ασυνέχειας, είτε αυτά που προκαλέσαμε με την επέκταση είτε αυτά που έχει εγγενώς η συνάρτηση.

6.7.4 Παράδειγμα

Ας υπολογίσουμε τη σειρά Fourier της συνάρτησης $f(x) = x^2$ που ορίζεται στο διάστημα $[0, 1)$.

Κατοπτρισμός ως προς την ευθεία $y = 0$

Επιλέγουμε καταρχάς να επεκτείνουμε τη συνάρτηση στο διάστημα $[-1, 0)$ συμμετρικά ώστε να μην εισαχθούν σημεία ασυνέχειας. Εκεί $f(x) = f(-x) = x^2$. Κατόπιν, επαναλαμβάνουμε τη συνάρτηση $g(x) = x^2$ με $-1 \leq x < 1$ έξω από το διάστημα $[-1, 1)$, ώστε να κατασκευάσουμε συνάρτηση περιοδική με περίοδο 2 (Σχήμα 6.10).



Σχήμα 6.10: Επέκταση της $f(x) = x^2$ συμμετρικά

Η νέα συνάρτηση, $g(x)$, είναι συμμετρική σε διάστημα μιας περιόδου, $[-1, 1)$, ως προς το μέσο του, οπότε οι συντελεστές B_m στην εξίσωση (6.4β΄) είναι 0. Για τους A_m έχουμε

$$A_m = \int_{-1}^{1} \cos(m\pi x) x^2 \, \mathrm{d}x \, , \qquad m \geq 0 \, .$$

Το διάστημα ολοκλήρωσης επελέγη να είναι το $[-1, 1)$. Όπως αναφέραμε, αρκεί να έχει μήκος μία περίοδο. Ο υπολογισμός των A_m δίνει μετά από πράξεις,

$$A_0 = \frac{2}{3} \, , \qquad A_m = (-1)^m \frac{4}{m^2\pi^2} \, , \quad m > 0 \, .$$

Επομένως,

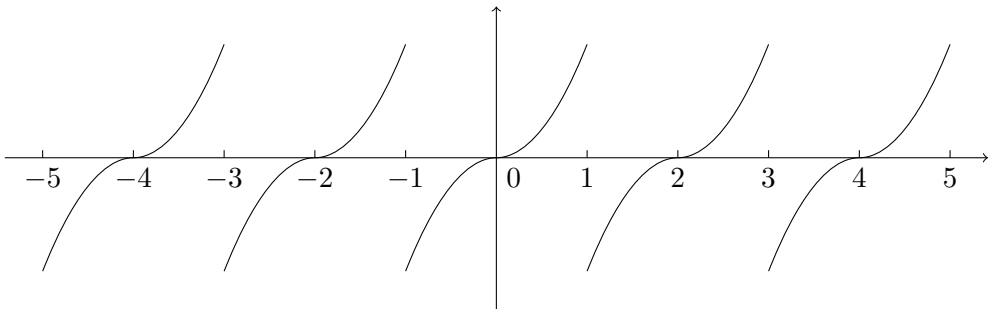
$$f(x) = \frac{1}{3} + 4 \sum_{m=1}^{\infty} \frac{(-1)^m \cos(m\pi x)}{m^2\pi^2} \, , \qquad 0 \leq x < 1 \, .$$

Ποιες τιμές έχει η σειρά Fourier στα $x = 0$ και $x = 1$;

Κατοπτρισμός ως προς την αρχή

Ας επεκτείνουμε τώρα τη συνάρτηση στο διάστημα $[-1, 0)$ αντισυμμετρικά με κατοπτρισμό ως προ το σημείο $(0, 0)$. Στο $[-1, 0]$ θεωρούμε ότι $f(x) = -f(-x) =$

$-x^2$. Κατόπιν, επαναλαμβάνουμε τη νέα συνάρτηση έξω από το διάστημα $[-1, 1)$, ώστε να κατασκευάσουμε συνάρτηση περιοδική με περίοδο 2 (Σχήμα 6.11). Καθώς $f(0) = 0$ το $x = 0$ και το ± 2, ± 4, ... δεν είναι σημεία ασυνέχειας. Τα $x = \pm 1$, ± 3 ... είναι.



Σχήμα 6.11: Επέκταση της $f(x) = x^2$ αντισυμμετρικά

Η νέα συνάρτηση είναι αντισυμμετρική σε διάστημα μιας περιόδου, $[-1, 1)$, ως προς το μέσο του, οπότε οι συντελεστές $_m$ είναι 0. Για τους B_m έχουμε

$$B_m = 2 \int_0^1 \sin(m\pi x) x^2 \, \mathrm{d}x \, , \qquad m > 0 \, .$$

Μετά από πράξεις,

$$\begin{aligned} B_{2k} &= -\frac{1}{k\pi} \quad k > 0 \, , \\ B_{2k+1} &= \frac{1}{(k + 1/2)\pi} - \frac{1}{(k + 1/2)^3 \pi^3} \quad k > 0 \, . \end{aligned}$$

Επομένως,

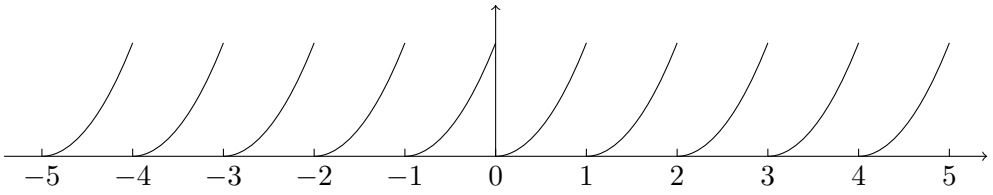
$$f(x) = \sum_{k=1}^{\infty} B_{2k} \sin(2k\pi x) + \sum_{k=1}^{\infty} B_{2k+1} \sin((2k+1)\pi x) \, , \qquad 0 \le x < 1 \, .$$

Μετατόπιση

Ας επιλέξουμε να επαναλάβουμε την $f(x) = x^2$ αυτούσια έξω από το διάστημα $[0, 1)$ ώστε να δημιουργήσουμε περιοδική συνάρτηση με περίοδο 1. Η συγκεκριμένη επιλογή εισάγει τα σημεία ασυνέχειας $x = j$ με j οποιοδήποτε ακέραιο. Σε αυτά, η σειρά Fourier δεν θα συγκλίνει στην περιοδική συνάρτηση.

Ο υπολογισμός των συντελεστών A_m από την εξίσωση (6.4α′) δίνει

$$A_0 = \frac{2}{3} \, , \qquad A_m = \frac{1}{m^2 \pi^2} \, , \quad m > 0 \, .$$



Σχήμα 6.12: Επέκταση της $f(x) = x^2$ με επανάληψη

Οι συντελεστές B_m της σχέσης (6.4β′) είναι

$$B_m = -\frac{1}{m\pi}\ , \quad m > 0\ .$$

Επομένως, η σειρά Fourier είναι

$$\frac{1}{3} + \sum_{m=1}^{\infty} \frac{\cos(2m\pi x)}{m^2\pi^2} - \sum_{m=1}^{\infty} \frac{\sin(2m\pi x)}{m\pi}\ .$$

και συγκλίνει στην $f(x) = x^2$ στο διάστημα $(0, 1)$ (παραλείπουμε το σημείο $x = 0$ που είναι σημείο ασυνέχειας).

Παρατηρήστε ότι αν παραγωγίσουμε την $f(x)$ και τη σειρά Fourier έχουμε

$$2x = -2\sum_{m=1}^{\infty} \frac{\sin(2m\pi x)}{m\pi} - 2\sum_{m=1}^{\infty} \cos(2m\pi x)\ .$$

Επομένως,

$$x - \frac{1}{2} = -\sum_{m=1}^{\infty} \frac{\sin(2m\pi x)}{m\pi} - \sum_{m=1}^{\infty} \cos(2m\pi x) - \frac{1}{2}\ .$$

Συγκρίνετε την τελευταία σχέση με τη σειρά Fourier του πριονωτού παλμού, (6.9). Ταυτίζονται αν για οποιοδήποτε σημείο του $(0, 1)$ ισχύει

$$\begin{aligned}
\sum_{m=1}^{\infty} \cos(2m\pi x) = -\frac{1}{2} \quad &\Leftrightarrow \quad 2\sum_{m=1}^{\infty} \cos(2m\pi x) + 1 = 0 \\
&\Leftrightarrow \quad \sum_{m=1}^{\infty} \cos(2m\pi x) + \sum_{m=-1}^{-\infty} \cos(-2m\pi x) + \cos(2\pi x 0) = 0 \\
&\Leftrightarrow \quad \sum_{m=0}^{\infty} \cos(2m\pi x) + \sum_{m=-1}^{-\infty} \cos(2m\pi x) = 0 \\
&\Leftrightarrow \quad \sum_{m=-\infty}^{\infty} \cos(2m\pi x) = 0\ .
\end{aligned}$$

Πράγματι ισχύει, αλλά δεν είναι του παρόντος η σχετική απόδειξη[23]

Παρατήρηση: Όπως είδαμε στα παραδείγματα, η ίδια μη περιοδική συνάρτηση, στο ίδιο διάστημα, μπορεί να έχει σειρές Fourier με διαφορετική μορφή, ανάλογα με τον τρόπο επέκτασής της. Βέβαια, η τιμή των διάφορων σειρών στο ίδιο σημείο είναι η ίδια.

6.8 Μιγαδική μορφή της σειράς Fourier

Η σειρά Fourier μπορεί να γραφεί σε πιο συνοπτική μορφή αν θυμηθούμε ότι

$$\mathrm{e}^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta \ .$$

Εύκολα προκύπτει ότι

$$\cos\theta = \frac{\mathrm{e}^{\mathrm{i}\theta} + \mathrm{e}^{-\mathrm{i}\theta}}{2} \ , \qquad \sin\theta = \frac{\mathrm{e}^{\mathrm{i}\theta} - \mathrm{e}^{-\mathrm{i}\theta}}{2\mathrm{i}} \ .$$

Η αντικατάσταση των παραπάνω σχέσεων στην (6.3) δίνει τη σειρά Fourier στην εκθετική μορφή

$$f(x) = \sum_{m=-\infty}^{\infty} C_m \exp\left(\mathrm{i}\frac{2m\pi x}{L}\right) \ , \tag{6.13}$$

όπου οι *μιγαδικοί*, πλέον, συντελεστές C_m συνδέονται με τους πραγματικούς A_m, B_m με τις σχέσεις

$$C_0 = \frac{A_0}{2} \ , \tag{6.14α′}$$

$$C_m = \frac{A_m - \mathrm{i}B_m}{2} \ , \qquad m > 0 \ , \tag{6.14β′}$$

$$C_{-m} = \frac{A_m + \mathrm{i}B_m}{2} = C_m^* \ , \qquad m > 0 \ . \tag{6.14γ′}$$

[2] σε οποιοδήποτε σημείο x το τελευταίο άθροισμα είναι $2\pi\delta(x)$. Είναι 0 για $x \neq 0$.

[3] Προσέξτε ότι

$$\sum_{m=1}^{\infty}\cos(2m\pi x) = -\frac{1}{2} \Leftrightarrow 1 + \sum_{m=1}^{\infty}\cos(2m\pi x) = 1 - \frac{1}{2} \Leftrightarrow \sum_{m=0}^{\infty}\cos(2m\pi x) = \frac{1}{2} \ .$$

Όταν $x = 1/2$, η τελευταία έκφραση γίνεται

$$\sum_{m=0}^{\infty}\cos(m\pi) = \frac{1}{2} \Rightarrow \sum_{m=0}^{\infty}(-1)^m = \frac{1}{2} \Rightarrow 1 - 1 + 1 - 1 + 1 - 1 + \cdots = \frac{1}{2} \ .$$

Η σειρά στο αριστερό μέλος είναι γνωστή ως σειρά Grandi, και η τιμή της, με συγκεκριμένο ορισμό για την άθροιση, είναι όντως $1/2$, όσο παράδοξο και αν φαίνεται.

Πολύ εύκολα προκύπτει και η αντίστροφη σχέση που προσδιορίζει τα A_m, B_m όταν είναι γνωστά τα C_m:

$$A_m = C_m + C_{-m} \,, \qquad m \geq 0 \,, \tag{6.15α΄}$$

$$B_m = \mathrm{i}(C_m - C_{-m}) \,, \qquad m > 0 \,. \tag{6.15β΄}$$

Συνδυάζοντας τις (6.4α΄), (6.4β΄), (6.14) προκύπτει ότι

$$C_m = \frac{1}{L}\int_0^L \exp\left(-\mathrm{i}\frac{2m\pi x}{L}\right) f(x)\,\mathrm{d}x \,, \qquad m = 0, \pm 1, \pm 2, \ldots \,. \tag{6.16}$$

Με τους μιγαδικούς συντελεστές Fourier η ταυτότητα Parseval (εξίσωση (6.12)) γίνεται

$$\begin{aligned} \langle \boldsymbol{f}|\boldsymbol{f}\rangle \equiv \frac{2}{L}\int_0^L |f(x)|^2\,\mathrm{d}x &= 2C_0^2 + 4\sum_{m=1}^{\infty}(C_m^* C_m) \Rightarrow \\ \frac{1}{L}\int_0^L |f(x)|^2\,\mathrm{d}x &= |C_0|^2 + 2\sum_{m=1}^{\infty}|C_m|^2 \,. \end{aligned}$$

Καθώς $|C_{-m}| = |C_m|$, η προηγούμενη σχέση καταλήγει στην ταυτότητα Parseval με τους μιγαδικούς συντελεστές

$$\frac{1}{L}\int_0^L |f(x)|^2\,\mathrm{d}x = \sum_{m=-\infty}^{\infty}|C_m|^2 \,. \tag{6.17}$$

Άμεση συνέπεια της σχέσης αυτής είναι η έκφραση για τη μέση τιμή της ποσότητας $|f(x)|^2$:

$$\langle |f(x)|^2 \rangle = \sum_{m=-\infty}^{\infty}|C_m|^2 \,.$$

Παράδειγμα

Η μιγαδική μορφή της σειράς Fourier για τον πριονωτό παλμό, (6.7), στο διάστημα $[0, 1)$, έχει συντελεστές

$$C_0 = \int_0^1 \left(x - \frac{1}{2}\right)\mathrm{d}x = 0 \,, \tag{6.18α΄}$$

$$C_m = \int_0^1 \mathrm{e}^{-\mathrm{i}2m\pi x}\left(x - \frac{1}{2}\right)\mathrm{d}x = \frac{\mathrm{i}}{2m\pi} \,, \qquad m = \pm 1, \pm 2, \ldots \,. \tag{6.18β΄}$$

Μπορείτε εύκολα να επαληθεύσετε ότι $C_{-m} = C_m^*$ και ότι $C_m \equiv (A_m - \mathrm{i}B_m)/2$ για τους συντελεστές στην (6.8).

Επομένως, η μιγαδική σειρά Fourier που προκύπτει από την (6.13) είναι

$$-\sum_{\substack{m=-\infty\\ m\neq 0}}^{\infty} \frac{\mathrm{e}^{\mathrm{i}2m\pi x}}{\mathrm{i}2m\pi} \,.$$

Η ταυτότητα Parseval δίνει

$$\int_0^1 f(x)^2\,\mathrm{d}x = \sum_{m=-\infty}^{\infty} |C_m|^2 = \sum_{\substack{m=-\infty\\ m\neq 0}}^{\infty} \frac{1}{4m^2\pi^2} = \frac{1}{2\pi^2}\sum_{m=1}^{\infty}\frac{1}{m^2} \,.$$

6.9 Διακριτός μετασχηματισμός Fourier (DFT)

Οι συντελεστές A_m, B_m (ή C_m) της σειράς Fourier μπορούν να υπολογιστούν από τα αντίστοιχα ολοκληρώματα, είτε ακριβώς (με αναλυτικό υπολογισμό) είτε προσεγγιστικά με τις μεθόδους που είδαμε στο Κεφάλαιο 5. Στην περίπτωση που από τη συνάρτηση $f(x)$ έχουμε μόνο κάποιες τιμές της σε συγκεκριμένα σημεία μπορούμε να έχουμε μόνο προσεγγιστικό υπολογισμό των συντελεστών.

Ας εφαρμόσουμε τον προσεγγιστικό υπολογισμό ολοκληρώματος με τον εκτεταμένο τύπο του τραπεζίου (5.6) για να υπολογίσουμε το ολοκλήρωμα στην (6.16): Χωρίζουμε το διάστημα ολοκλήρωσης $[0, L]$ σε n ίσα διαστήματα μήκους $h = L/n$ το καθένα. Τα $n+1$ σημεία στα οποία θα υπολογίσουμε την ολοκληρωτέα ποσότητα είναι τα $x_j = jh$, $j = 0, 1, \ldots, n$. Παρατηρήστε ότι η ολοκληρωτέα ποσότητα στα άκρα, $x_0 = 0$ και $x_n = L$, έχει την ίδια τιμή:

$$\exp\left(-\mathrm{i}\frac{2m\pi 0}{L}\right) f(0) = \exp\left(-\mathrm{i}\frac{2m\pi L}{L}\right) f(L) \,.$$

Η ισότητα των $f(0)$ και $f(L)$ προκύπτει από την περιοδικότητα της συνάρτησης (εξίσωση (6.2)). Η διακριτοποίηση της ολοκληρωτέας ποσότητας δίνει τη σχέση

$$C_m \approx \frac{h}{L}\sum_{j=0}^{n-1}\exp\left(-\mathrm{i}\frac{2m\pi jh}{L}\right) f(jh) = \frac{1}{n}\sum_{j=0}^{n-1}\exp\left(-\mathrm{i}\frac{2m\pi j}{n}\right) f_j$$

για τους συντελεστές της σειράς Fourier, όπου $f_j \equiv f(jh)$. Στην ίδια έκφραση καταλήγουμε και στην περίπτωση που οι τιμές της συνάρτησης είναι γνωστές μόνο σε n ισαπέχοντα σημεία (π.χ. από πειραματικές μετρήσεις).

Η σχέση

$$\bar{C}_m = \frac{1}{n}\sum_{j=0}^{n-1}\exp\left(-\mathrm{i}\frac{2m\pi j}{n}\right) f_j \,, \qquad m = 0, 1, \ldots, n-1 \,, \tag{6.19}$$

αποτελεί το *διακριτό μετασχηματισμό Fourier (DFT)* της διακριτοποιημένης συνάρτησης $f(x)$. Οι συντελεστές $\bar{C}_m$ που ορίζονται από αυτή τη σχέση προσεγγίζουν τους συντελεστές C_m στη σειρά Fourier.

Παρατηρήστε ότι η διακριτοποίηση διατηρεί μόνο n συντελεστές $\bar{C}_m$ καθώς ισχύει η σχέση

$$\bar{C}_{m+n} \equiv \frac{1}{n}\sum_{j=0}^{n-1}\exp\left(-\mathrm{i}\frac{2(m+n)\pi j}{n}\right) f_j = \frac{1}{n}\sum_{j=0}^{n-1}\exp\left(-\mathrm{i}\frac{2m\pi j}{n}\right) f_j \equiv \bar{C}_m \ . \tag{6.20}$$

Ο *αντίστροφος διακριτός μετασχηματισμός Fourier* ορίζεται ως

$$\bar{f}_j = \sum_{m=0}^{n-1}\exp\left(\mathrm{i}\frac{2j\pi m}{n}\right)\bar{C}_m \ , \qquad j = 0, 1, \ldots, n-1 \ . \tag{6.21}$$

και προσεγγίζει τις τιμές f_j της συνάρτησης.[4]

Η μέση τιμή της διακριτοποιημένης συνάρτησης $|f(x)|^2$ είναι προφανώς

$$\langle f^2 \rangle = \frac{1}{n}\sum_{j=0}^{n-1}|f_j|^2 \ .$$

Η ταυτότητα Parseval γίνεται

$$\frac{1}{n}\sum_{j=0}^{n-1}|f_j|^2 = \sum_{m=0}^{n-1}\left|\bar{C}_m\right|^2 \ . \tag{6.22}$$

Μπορεί να αποδειχθεί αν αντικαταστήσουμε στο δεξί της μέλος την έκφραση για τα $\bar{C}_m$ από τη σχέση (6.19).

6.9.1 Γρήγορος υπολογισμός του DFT – Αλγόριθμος FFT

Υπάρχουν διάφοροι αλγόριθμοι που μπορούν να υπολογίσουν ταυτόχρονα όλους τους συντελεστές Fourier, ιδιαίτερα γρήγορα, εκμεταλλευόμενοι τις συμμετρίες που εμφανίζονται, χωρίς να χρειάζεται να υπολογίσουν κάθε άθροισμα ξεχωριστά. Παρακάτω θα δούμε τον πιο βασικό.

Ας υποθέσουμε ότι το πλήθος n των όρων στο άθροισμα της (6.19) είναι δύναμη του 2. Τότε, ο υπολογισμός του μπορεί να γίνει χωρίζοντάς το σε αθροίσματα των

[4] Ο παράγοντας $1/n$ που πολλαπλασιάζει το άθροισμα στην (6.19) είναι θέμα σύμβασης. Το γινόμενο των συντελεστών πριν τα αθροίσματα στις εξισώσεις (6.19) και (6.21) πρέπει να είναι $1/n$, οι ακριβείς τιμές τους είναι απροσδιόριστες. Για λόγους συμμετρίας των σχέσεων (6.19, 6.21), οι μετασχηματισμοί μπορούν να οριστούν με ένα παράγοντα $1/\sqrt{n}$ που πολλαπλασιάζει το άθροισμα του καθενός.

όρων με άρτιο και περιττό δείκτη j:

$$\begin{aligned}
&\sum_{j=0}^{n-1} \exp\left(-\mathrm{i}\frac{2m\pi j}{n}\right) f_j \\
&= \sum_{r=0}^{n/2-1} \exp\left(-\mathrm{i}\frac{2m\pi 2r}{n}\right) f_{2r} + \sum_{r=0}^{n/2-1} \exp\left(-\mathrm{i}\frac{2m\pi(2r+1)}{n}\right) f_{2r+1} \\
&= \sum_{r=0}^{n/2-1} \exp\left(-\mathrm{i}\frac{2m\pi r}{n/2}\right) f_{2r} + \exp\left(-\mathrm{i}\frac{2m\pi}{n}\right) \sum_{r=0}^{n/2-1} \exp\left(-\mathrm{i}\frac{2m\pi r}{n/2}\right) f_{2r+1} \,.
\end{aligned}$$

Παρατηρήστε ότι οι όροι

$$\sum_{r=0}^{n/2-1} \exp\left(-\mathrm{i}\frac{2m\pi r}{n/2}\right) f_{2r} \qquad \text{και} \qquad \sum_{r=0}^{n/2-1} \exp\left(-\mathrm{i}\frac{2m\pi r}{n/2}\right) f_{2r+1}$$

είναι ουσιαστικά οι διακριτοί μετασχηματισμοί Fourier για δύο σύνολα τιμών της διακριτοποιημένης $f(x)$· το ένα αποτελείται από τα σημεία f_j με άρτιο δείκτη και το άλλο από τα σημεία με περιττό δείκτη. Το πλήθος των σημείων σε κάθε σύνολο είναι $n/2$.

Ας συμβολίσουμε με $\bar{C}^e_m$, $\bar{C}^o_m$ τους συντελεστές στους δύο μετασχηματισμούς Fourier, τον «άρτιο» και τον «περιττό» αντίστοιχα. Η προηγούμενη σχέση δίνει

$$\bar{C}_m = \frac{1}{n}\left(\frac{n}{2}\bar{C}^e_m + \frac{n}{2}\exp\left(-\mathrm{i}\frac{2m\pi}{n}\right)\bar{C}^o_m\right) = \frac{1}{2}\left(\bar{C}^e_m + \mathrm{e}^{-\mathrm{i}2m\pi/n}\bar{C}^o_m\right) \,, \tag{6.23}$$

για $m = 0, 1, \ldots, n-1$.

Παρατηρήστε ότι, λόγω της (6.20), έχουμε $\bar{C}^{e,o}_{m+n/2} = \bar{C}^{e,o}_m$. Επίσης ισχύει ότι

$$\exp\left(-\mathrm{i}\frac{2(m+n/2)\pi}{n}\right) = -\exp\left(-\mathrm{i}\frac{2m\pi}{n}\right) \,.$$

Επομένως, η σχέση (6.23) μπορεί να ξαναγραφεί ως εξής

$$\bar{C}_m = \frac{1}{2}\left(\bar{C}^e_m + \mathrm{e}^{-\mathrm{i}2m\pi/n}\bar{C}^o_m\right) \,, \tag{6.24α′}$$

$$\bar{C}_{m+n/2} = \frac{1}{2}\left(\bar{C}^e_m - \mathrm{e}^{-\mathrm{i}2m\pi/n}\bar{C}^o_m\right) \,, \tag{6.24β′}$$

για $m = 0, 1, \ldots, n/2-1$.

Η εξίσωση (6.23) (ή, ισοδύναμα, η εξίσωση (6.24)) εκφράζει ότι ο υπολογισμός του DFT n σημείων απαιτεί τον υπολογισμό δύο DFT των $n/2$ σημείων ο καθένας. Η συγκεκριμένη ανάλυση μπορεί να χρησιμοποιηθεί για τον υπολογισμό των νέων DFT αναπτύσσοντάς τους σε τέσσερις συνολικά DFT των $n/4$ σημείων ο καθένας. Η διαδικασία αυτή επαναλαμβάνεται έως ότου καταλήξουμε σε n DFT του ενός

σημείου ο καθένας. Ο υπολογισμός του DFT ενός σημείου είναι πολύ εύκολος: από τη (6.19) προκύπτει ότι ο (μοναδικός) συντελεστής της σειράς Fourier είναι ίσος με την τιμή της συνάρτησης στο σημείο.

Η επαναληπτική διαδικασία που περιγράψαμε είναι η βάση των αλγορίθμων Fast Fourier Transform (FFT). Σε αυτή, ο συντελεστής $\bar{C}_m$ απαιτεί για τον υπολογισμό του συνολικά $2\log_2 n$ μιγαδικούς πολλαπλασιασμούς. Επομένως, οι n συντελεστές χρειάζονται $2n\log_2 n$ πράξεις για τον υπολογισμό τους.

Αν επιλέγαμε να υπολογίσουμε το άθροισμα στην (6.19) απευθείας, χρειαζόμαστε n πολλαπλασιασμούς για τον κάθε συντελεστή· συνολικά, δηλαδή, n^2 πράξεις. Το κέρδος σε ταχύτητα είναι σημαντικό: αν π.χ. έχουμε $n = 1024$ ο αλγόριθμος FFT χρειάζεται 20480 πράξεις ενώ χωρίς αυτόν θα κάναμε 1048576 πράξεις.

6.10 Ασκήσεις

1. Ποιες από τις επόμενες συναρτήσεις μπορούν να αναπτυχθούν σε σειρά Fourier; σε ποια σημεία δεν θα συγκλίνει η σειρά στη συνάρτηση;

   - $\tanh^{-1} x$,
   - $\tan x$,
   - $1/\sqrt{|\sin x|}$.

2. Βρείτε τη σειρά Fourier που προσεγγίζει την

$$f(x) = \begin{cases} 0\,, & 0 \le x < 1/2\,, \\ 1\,, & 1/2 \le x < 3/2\,, \\ 0\,, & 3/2 \le x < 2\,. \end{cases}$$

   Θεωρούμε ότι η συνάρτηση επαναλαμβάνεται για $x \ge 2$ και $x < 0$ ώστε $f(x + 2k) = f(x)$ με οποιοδήποτε ακέραιο k.

3. Βρείτε τη σειρά Fourier της $f(x) = x$ στο διάστημα $[-\pi, \pi)$. Κατόπιν, δείξτε ότι

$$1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots = \frac{\pi}{4}\,.$$

4. Βρείτε τη σειρά Fourier της συνάρτησης $f(t) = |\sin(\omega t)|$ με $-\pi < \omega t < \pi$. Ποιες συχνότητες έχουν μη μηδενικό πλάτος και πόσο;

5. Βρείτε τη μιγαδική σειρά Fourier της συνάρτησης $f(x) = |x|$ με $-\pi < x < \pi$. Κατόπιν, δείξτε ότι

$$1 + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \cdots = \frac{\pi^2}{8}\,.$$

6. Επεκτείνετε (α′) συμμετρικά (β′) αντισυμμετρικά την $f(x) = 1 - x$ που ορίζεται στο $[0, 1)$. Βρείτε τις αντίστοιχες σειρές Fourier.

7. Βρείτε τη σειρά Fourier της $f(x) = x^3$ στο διάστημα $[0, 2)$.

8. Βρείτε τη σειρά Fourier της $f(x) = x^2$ στο διάστημα $[-2, 2)$. Δείξτε ότι
$$1 + \frac{1}{2^4} + \frac{1}{3^4} + \frac{1}{4^4} + \cdots = \frac{\pi^4}{90} .$$

9. Βρείτε τη σειρά Fourier της $f(x) = \mathrm{e}^x$ στο διάστημα $[-1, 1)$. Ποια τιμή έχει η σειρά στο $x = 2$;

10. Δείξτε ότι η σειρά Fourier της συνάρτησης $f(x) = |x|$ στο διάστημα $[-\pi, \pi)$ είναι
$$|x| = \frac{\pi}{2} - \frac{4}{\pi} \sum_{m=0}^{\infty} \frac{\cos(2m+1)x}{(2m+1)^2} .$$
Ολοκληρώστε τη συγκεκριμένη σειρά Fourier και βρείτε έτσι τη συνάρτηση που έχει σειρά Fourier
$$\frac{4}{\pi} \sum_{m=0}^{\infty} \frac{\sin(2m+1)x}{(2m+1)^3} .$$
στο συγκεκριμένο διάστημα.

11. Μπορείτε να βρείτε από τα αποτελέσματα της άσκησης 10 την τιμή του αθροίσματος
$$1 - \frac{1}{3^3} + \frac{1}{5^3} - \frac{1}{7^3} + \cdots ;$$

12. Γράψτε κώδικα που να υλοποιεί τον αλγόριθμο FFT. Θα σας διευκολύνει να χρησιμοποιήσετε αναδρομική συνάρτηση. Μπορείτε να τη γράψετε χωρίς να χρησιμοποιεί νέα διανύσματα;

    Εφαρμόστε τον κώδικά σας για να υπολογίσετε το διακριτό μετασχηματισμό Fourier για τον πριονωτό παλμό: $f(x) = x - 0.5$ για $0 \le x \le 1$. Θεωρούμε ότι η συνάρτηση επαναλαμβάνεται περιοδικά με μετατόπιση. Επιλέξτε 1024 ισαπέχοντα σημεία στο $[0, 1)$ (προσέξτε ότι δεν περιλαμβάνεται το δεξί άκρο) και υπολογίστε σε αυτά τη συνάρτηση.

    Η ακριβής λύση είναι, σύμφωνα με την (6.18): $C_0 = 0$, $C_m = \frac{\mathrm{i}}{2m\pi}$ με $m = \pm 1, \pm 2, \ldots$.

13. Χρησιμοποιήστε τον αλγόριθμο FFT για να υπολογίσετε τους συντελεστές της μεθόδου ολοκλήρωσης Clenshaw–Curtis (§5.7).

Κεφάλαιο 7

Διαφορικές Εξισώσεις

7.1 Εισαγωγή

Μια εξίσωση που περιγράφει μια σχέση μεταξύ μιας ανεξάρτητης μεταβλητής, x, μιας εξαρτημένης συνάρτησης, y, και μίας ή περισσότερων παραγώγων τής y λέγεται *συνήθης Διαφορική Εξίσωση* (ΔΕ):

$$y^{(n)}(x) = f\left(x, y(x), y'(x), \ldots, y^{(n-1)}(x)\right) \ . \tag{7.1}$$

Η συγκεκριμένη εξίσωση είναι ΔΕ τάξης n.

Μια συνάρτηση $\phi(x)$, παραγωγίσιμη n φορές σε κάποιο διάστημα, η οποία ικανοποιεί την (7.1), δηλαδή ισχύει

$$\phi^{(n)}(x) = f\left(x, \phi(x), \phi'(x), \ldots, \phi^{(n-1)}(x)\right) \ ,$$

αποτελεί (μία) *λύση* της συγκεκριμένης ΔΕ. Μια γενική λύση της (7.1) περιέχει n αυθαίρετες σταθερές, επομένως υπάρχει μια n-παραμετρική οικογένεια λύσεων.

Ο αναλυτικός υπολογισμός της λύσης $\phi(x)$, ως έκφραση του x δηλαδή, είναι εφικτός για πολύ ειδικές περιπτώσεις διαφορικών εξισώσεων. Στην πράξη οι περισσότερες ΔΕ δεν επιδέχονται αναλυτική λύση. Εκτός όμως από αυτό, για πολλά προβλήματα δε μας ενδιαφέρει τόσο η αναλυτική λύση όσο οι αριθμητικές τιμές της σε ορισμένα σημεία.

Αν τα $y(x_0), y'(x_0), \ldots, y^{(n-1)}(x_0)$ είναι γνωστά για κάποιο σημείο x_0, έχουμε *πρόβλημα αρχικών τιμών*. Με αυτή κατηγορία προβλημάτων θα ασχοληθούμε στο παρόν κεφάλαιο: θα επιδιώκουμε να υπολογίσουμε την *τιμή* της $y(x)$ σε κάποιο σημείο x_1 όταν είναι γνωστές η τιμή και οι παράγωγοί της σε κάποιο σημείο x_0.

Αφού δούμε την αναγκαία συνθήκη για την ύπαρξη λύσης και την έννοια της ευστάθειας, θα παρουσιάσουμε διάφορες μεθόδους για την επίλυση του προβλήματος αρχικών τιμών με διαφορική εξίσωση πρώτης τάξης, δηλαδή της μορφής

$$y'(x) = f(x, y) \ , \tag{7.2α′}$$
$$y(x_0) = y_0 \ . \tag{7.2β′}$$

Στην εξίσωση εμφανίζεται η πρώτη μόνο από τις παραγώγους.

Η επίλυση της (7.2) θα είναι αριθμητική, δηλαδή, γνωρίζοντας της τιμή της συνάρτησης $y(x)$ στο x_0 θα υπολογίζουμε μια *τιμή* y_1 που θα προσεγγίζει την (άγνωστη) τιμή της $y(x)$ σε κάποιο σημείο x_1. Οι μέθοδοι που θα παρουσιάσουμε υπολογίζουν μια προσεγγιστική τιμή, y_1, για το $y(x_1)$. Το σφάλμα που κάνουν, η διαφορά $y_1 - y(x_1)$, αυξάνει όταν αυξάνει το $(x_1 - x_0)$. Ειδικότερα, το σφάλμα είναι ανάλογοτου $(x_1 - x_0)^{p+1}$. Ο ακέραιος p είναι η *τάξη* της μεθόδου. Η απόσταση $(x_1 - x_0)$ πρέπει να είναι κατάλληλα «μικρή» ώστε το σφάλμα να είναι αποδεκτό. Συνήθως όμως γνωρίζουμε την τιμή της $y(x)$ σε σημείο a και επιθυμούμε να την υπολογίσουμε σε κάποιο «μακρινό» σημείο b. Αντί να εφαρμόσουμε μία φορά την επιλεγμένη μέθοδο με υψηλό σφάλμα, είναι προτιμότερο να διαιρέσουμε το διάστημα $[a, b]$ σε μικρά τμήματα μεταξύ των σημείων $x_0 \equiv a$, x_1, x_2, ..., $x_n \equiv b$, και να χρησιμοποιήσουμε επαναληπτικά την επιλεγμένη μέθοδο ώστε από τη γνωστή τιμή στην αρχή του πρώτου τμήματος, στο $x_0 \equiv a$, να υπολογίσουμε την τιμή στο τέλος του πρώτου τμήματος, x_1, (και συνεπώς στην αρχή του δεύτερου), κατόπιν στο τέλος του δεύτερου τμήματος, x_2, κοκ. έως ότου φτάσουμε στο $x_n \equiv b$.

Τα σημεία διαμοιρασμού του διαστήματος $[a, b]$ δεν είναι απαραίτητο να ισαπέχουν, ούτε να είναι γνωστά εκ των προτέρων: οι μέθοδοι μεταβλητού βήματος χρησιμοποιούν τις τιμές των x_0, x_1, ..., x_i και τις αντίστοιχες τιμές του y ώστε να υπολογίσουν το σημείο x_{i+1}.

7.1.1 Επιλυσιμότητα

Μια συνάρτηση $f(x, y)$, συνεχής για $x \in [a, b]$ και $y \in \mathcal{R}$ λέγεται ότι ικανοποιεί μια συνθήκη Lipschitz στο χώρο $[a, b] \times R$ αν για κάποια σταθερή ποσότητα L (σταθερά Lipschitz) έχουμε

$$|f(x, y_1) - f(x, y_2)| \leq L\,|y_1 - y_2|$$

για κάθε $x \in [a, b]$ και $y_1, y_2 \in \mathcal{R}$.

Αν η συνάρτηση $f(x, y)$ και ικανοποιεί μια συνθήκη Lipschitz στο χώρο $[a, b] \times R$ τότε το πρόβλημα αρχικών τιμών (7.2) έχει μία και μοναδική λύση $y(x)$, $x \in [a, b]$.

Στα παρακάτω προϋποθέτουμε ότι η ΔΕ πληροί όλες εκείνες τις συνθήκες που της εξασφαλίζουν την ύπαρξη και το μονοσήμαντο της λύσης.

7.1.2 Αριθμητική Ευστάθεια

Η ευστάθεια μιας μεθόδου επίλυσης διαφορικών εξισώσεων αναφέρεται στη συμπεριφορά της διαφοράς μεταξύ της τιμής που υπολογίζει και της πραγματικής τιμής της λύσης. Αν το σφάλμα μεγαλώνει σε κάθε επανάληψη και τελικά κυριαρχεί της λύσης, η μέθοδος που εφαρμόζεται χαρατηρίζεται ως ασταθής.

Μαθηματικά, θα λέμε ότι μια μέθοδος επίλυσης είναι *αριθμητικά ευσταθής* αν,

για $y(x_i) \neq 0$, τα σχετικά σφάλματα

$$\left| \frac{y_i - y(x_i)}{y(x_i)} \right|$$

είναι φραγμένα (δεν απειρίζονται) όταν $i \to \infty$.

7.1.3 Διάκριση σε explicit/implicit

Οι μέθοδοι επίλυσης συνήθων διαφορικών εξισώσεων διακρίνονται σε δύο βασικές κατηγορίες: άμεσες (explicit) και πεπλεγμένες (implicit). Οι αλγόριθμοι της πρώτης κατηγορίας εκφράζουν την y_1 ως συνάρτηση του y_0 και τιμών της $f(x, y)$ σε διάφορα σημεία, μπορούν δηλαδή να την υπολογίσουν απευθείας με έκφραση της μορφής π.χ.

$$y_1 = G(x_0, x_1, y_0) \ .$$

Οι αλγόριθμοι της κατηγορίας implicit, αντίθετα, προσδιορίζουν μια έκφραση της μορφής

$$G(x_0, x_1, y_0, y_1) = 0 \ ,$$

ή γενικότερα, ένα μη γραμμικό σύστημα εξισώσεων. Σε αυτή την περίπτωση ο υπολογισμός της y_1 απαιτεί τη λύση μιας αλγεβρικής εξίσωσης με μέθοδο εύρεσης ρίζας συνάρτησης, ή τη λύση του μη γραμμικού συστήματος.

Μια μέθοδος implicit είναι πιο χρονοβόρα αλλά ευσταθής σε Διαφορικές Εξισώσεις που οι explicit μέθοδοι αδυνατούν να επιλύσουν σωστά χωρίς να χρειαστεί να κάνουν το βήμα ιδιαίτερα μικρό (άκαμπτες (stiff) εξισώσεις).

7.2 Μέθοδος Σειράς Taylor

Γνωρίζουμε ότι μια συνάρτηση $y(x)$, για την οποία οι τιμές αυτής και των παραγώγων της σε κάποιο σημείο x_0 είναι γνωστές, μπορεί να υπολογιστεί σε κάποιο σημείο x_1 από το ανάπτυγμα Taylor,

$$y(x_1) = y(x_0) + y'(x_0)(x_1 - x_0) + \frac{y''(x_0)}{2!}(x_1 - x_0)^2 + \cdots \ ,$$

αρκεί να είναι συνεχής και παραγωγίσιμη στο διάστημα $[x_0, x_1]$.

Παρατηρήστε ότι στη διαφορική εξίσωση (7.2) η πρώτη παράγωγος της $y(x)$ είναι γνωστή συνάρτηση, συνεχής και παραγωγίσιμη. Έτσι μπορούμε να την υπολογίσουμε στο x_0:

$$y'(x_0) = f(x_0, y(x_0)) = f(x_0, y_0) \ .$$

Επιπλέον, μπορούμε να τη χρησιμοποιήσουμε για να εξαγάγουμε τη δεύτερη παράγωγο:

$$y''(x) = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}x} = f_x + f_y y' \ .$$

Η τρίτη παράγωγος μπορεί να εξαχθεί από την $y''(x)$:

$$\begin{aligned} y'''(x) &= \frac{\mathrm{d}}{\mathrm{d}x}(f_x + f_y y') = (f_{xx} + f_{xy}y') + (f_{yx} + f_{yy}y')y' + f_y y'' \\ &= f_{xx} + 2f_{xy}y' + f_{yy}(y')^2 + f_y y'' \ . \end{aligned}$$

Διαδοχικές παραγωγίσεις μπορούν να παραγάγουν μαθηματικές εκφράσεις για τις παραγώγους οποιασδήποτε τάξης. Μπορούμε έτσι να υπολογίσουμε θεωρητικά απεριόριστους όρους στο ανάπτυγμα Taylor της $y(x_1)$. Στην πράξη μπορούμε να περιλάβουμε τους όρους με παράγωγο έως κάποιας τάξης m. Ανώτερες παράγωγοι θα είναι αρκετά πολύπλοκες για να υπολογιστούν. Επομένως, ο τύπος με τον οποίο η συγκεκριμένη μέθοδος υπολογίζει την προσέγγιση y_1 της $y(x_1)$ είναι

$$\begin{aligned} y_1 \approx\ & y_0 + f(x_0, y_0)(x_1 - x_0) + \frac{(x_1 - x_0)^2}{2!}\,(f_x + f_y f)|_{x_0} \\ & + \frac{(x_1 - x_0)^3}{3!}\,(f_{xx} + 2f_{xy}f + f_{yy}(f)^2 + f_y(f_x + f_y f))\big|_{x_0} + \cdots \ . \end{aligned} \tag{7.3}$$

Η συγκεκριμένη μέθοδος είναι explicit. Μπορούμε όμως να χρησιμοποιήσουμε το ανάπτυγμα Taylor με τέτοιο τρόπο ώστε να καταλήξουμε σε implicit μέθοδο. Ας γράψουμε το ανάπτυγμα της $y(x_0)$ όταν γνωρίζουμε την τιμή και τις παραγώγους στο σημείο x_1:

$$\begin{aligned} y(x_0) &= y(x_1) + y'(x_1)(x_0 - x_1) + \frac{y''(x_1)}{2!}(x_0 - x_1)^2 + \cdots \Rightarrow \\ y_1 &= y_0 + (x_1 - x_0)y'(x_1) - \frac{y''(x_1)}{2!}(x_1 - x_0)^2 + \cdots \Rightarrow \\ y_1 &= y_0 + (x_1 - x_0)f(x_1, y_1) - \frac{(x_1 - x_0)^2}{2!}\,(f_x + f_y f)|_{x_1} + \cdots \ . \end{aligned}$$

Η έκφραση στην οποία καταλήγουμε έχει μοναδικό άγνωστο το y_1 αλλά πρέπει να επιλυθεί ώστε να το υπολογίσουμε.

Σφάλμα μεθόδου Taylor

Μπορεί να δειχθεί ότι το σφάλμα που θα κάνουμε από την αποκοπή όρων ανώτερων του m δίνεται από τη σχέση

$$\varepsilon = \frac{y^{(m+1)}(\xi)}{(m+1)!}(x_1 - x_0)^{m+1} \ , \qquad \xi \in (x_0, x_1) \ .$$

Η απόσταση $x_1 - x_0$ πρέπει να είναι κατάλληλα «μικρή» και η τάξη m κατάλληλα «μεγάλη» ώστε το σφάλμα να είναι στα αποδεκτά όρια.

Όπως αναφέραμε, όταν έχουμε εκτεταμένο διάστημα $[a, b]$ θα πρέπει να το χωρίσουμε σε n διαστήματα μήκους $h = (b - a)/n$ και να εφαρμόσουμε τη μέθοδο

σειράς Taylor στο καθένα από αυτά. Το σφάλμα που θα έχουμε στο διάστημα $[x_i, x_{i+1}]$ (με $x_i = a + ih$) είναι

$$\varepsilon_i = \frac{y^{(m+1)}(\xi)}{(m+1)!}(x_{i+1} - x_i)^{m+1} , \qquad \xi \in (x_i, x_{i+1}) .$$

Έστω ότι υπάρχει αριθμός M ώστε $\left|y^{(m+1)}(\xi_i)\right| \leq M$ για κάθε i. Επομένως,

$$|\varepsilon_i| \leq \frac{M}{(m+1)!}h^{m+1} , \quad \forall i ,$$

και το συνολικό σφάλμα είναι

$$|E| = \sum_{i=0}^{n-1} \varepsilon_i \leq \sum_{i=0}^{n-1} |\varepsilon_i| \leq n\frac{M}{(m+1)!}h^{m+1} = \frac{(b-a)M}{(m+1)!}h^m . \tag{7.4}$$

Παρατηρούμε ότι το ολικό σφάλμα της μεθόδου είναι ανάλογο του h σε κάποια δύναμη και αυτή η δύναμη είναι κατά 1 μικρότερη από τη δύναμη του h στο τοπικό σφάλμα. Το συμπέρασμα αυτό είναι γενικό για όλες τις τις μεθόδους που θα παρουσιάσουμε.

7.2.1 Μέθοδος Euler

Η απλούστερη από τις μεθόδους Taylor είναι η *μέθοδος forward Euler* που προκύπτει από την (7.3) αν αποκόψουμε τους όρους της σειράς μετά το δεύτερο όρο, δηλαδή,

$$y_{i+1} = y_i + (x_{i+1} - x_i)f(x_i, y_i) + \mathcal{O}((x_{i+1} - x_i)^2) . \tag{7.5}$$

Η μέθοδος είναι explicit με περιορισμένη περιοχή ευστάθειας.

Παράδειγμα

Έστω $y' = -y$, $y(0) = 1$, και ζητούμε την τιμή $y(1)$. Η αναλυτική λύση είναι $y(x) = \mathrm{e}^{-x}$.

Χωρίζουμε το διάστημα $[0, 1]$ σε $n = 10$ ίσα διαστήματα, μήκους $h = 1/10$. Τα σημεία που ορίζουν τα τμήματα είναι $x_i = 0 + ih$. Η μέθοδος Euler δίνει

$$y_{i+1} \approx y_i + (x_{i+1} - x_i)y'(x_i) = y_i - hy_i = (1 - h)y_i .$$

Για $h = 0.1$ δίνουμε στον παρακάτω πίνακα τιμές της λύσης και τις αντίστοιχες

ακριβείς.

x	y	e^{-x}
0.0	1.0000	1.0000
0.1	0.9000	0.9048
0.2	0.8100	0.8187
0.3	0.7290	0.7408
0.4	0.6561	0.6703
0.5	0.5905	0.6065
0.6	0.5314	0.5488
0.7	0.4783	0.4966
0.8	0.4305	0.4493
0.9	0.3874	0.4066
1.0	0.3486	0.3679

Παρόλο που η μέθοδος Euler δεν είναι ικανοποιητικής ακρίβειας στο συγκεκριμένο παράδειγμα, ειδικά σε μεγάλες τιμές του x, η λύση που υπολογίζει έχει τη συμπεριφορά της σωστής λύσης.

Η implicit μέθοδος backward Euler μπορεί να παραχθεί από την explicit μέθοδο forward Euler αν θεωρήσουμε γνωστή την τιμή στο x_{i+1} και άγνωστη την τιμή στο x_i. Η (7.5) γίνεται

$$y_i = y_{i+1} + (x_i - x_{i+1}) f(x_{i+1}, y_{i+1}) + \mathcal{O}((x_i - x_{i+1})^2)$$

άρα

$$y_{i+1} = y_i + (x_{i+1} - x_i) f(x_{i+1}, y_{i+1}) + \mathcal{O}((x_{i+1} - x_i)^2)\ . \tag{7.6}$$

Παρατηρήστε ότι πρέπει να επιλύσουμε μια γενικά μη γραμμική έκφραση για να υπολογίσουμε το y_{i+1}. Η backward Euler έχει μεγαλύτερες υπολογιστικές απαιτήσεις από την forward Euler όμως μπορεί να είναι πιο ευσταθής μέθοδος από αυτή.

Το τοπικό σφάλμα των μεθόδων Euler είναι ανάλογο του h^2. Η επαναληπτική εφαρμογή των τύπων και στις δύο μεθόδους Euler οδηγεί σε ολικό σφάλμα που είναι ανάλογο του h, όπως προκύπτει από την (7.4) για $m = 1$.

7.3 Μέθοδοι Runge–Kutta

Οι μέθοδοι της οικογένειας Runge–Kutta (RK) επιλύουν αριθμητικά την (7.2) προσεγγίζοντας το αποτέλεσμα τής σειράς Taylor με γραμμικό συνδυασμό s τιμών της συνάρτησης $f(x, y)$ υπολογισμένων σε διάφορα σημεία.

Η γενική μορφή των explicit μεθόδων Runge–Kutta με s στάδια είναι

$$\begin{aligned} y_{i+1} &= y_i + \sum_{j=1}^{s} b_j k_j \ , \ \text{με} && \text{(7.7α΄)} \\ k_1 &= hf(x_i, y_i) && \text{(7.7β΄)} \\ k_2 &= hf(x_i + c_2 h, y_i + a_{21} k_1) && \text{(7.7γ΄)} \\ k_3 &= hf(x_i + c_3 h, y_i + a_{31} k_1 + a_{32} k_2) && \text{(7.7δ΄)} \\ &\vdots \quad \vdots \\ k_s &= hf(x_i + c_s h, y_i + a_{s1} k_1 + a_{s2} k_2 + \cdots + a_{s,s-1} k_{s-1}) \ , && \text{(7.7ε΄)} \end{aligned}$$

με $h = x_{i+1} - x_i$.

Η γενική μορφή των implicit μεθόδων Runge–Kutta είναι

$$\begin{aligned} y_{i+1} &= y_i + \sum_{j=1}^{s} b_j k_j \ , \ \text{με} && \text{(7.8α΄)} \\ k_1 &= hf(x_i + c_1 h, y_i + a_{11} k_1 + a_{12} k_2 + \cdots + a_{1s} k_s) && \text{(7.8β΄)} \\ k_2 &= hf(x_i + c_2 h, y_i + a_{21} k_1 + a_{22} k_2 + \cdots + a_{2s} k_s) && \text{(7.8γ΄)} \\ k_3 &= hf(x_i + c_3 h, y_i + a_{31} k_1 + a_{32} k_2 + \cdots + a_{3s} k_s) && \text{(7.8δ΄)} \\ &\vdots \quad \vdots \\ k_s &= hf(x_i + c_s h, y_i + a_{s1} k_1 + a_{s2} k_2 + \cdots + a_{ss} k_s) \ . && \text{(7.8ε΄)} \end{aligned}$$

Οι συντελεστές a_{ij}, είναι τα στοιχεία του πίνακα Runge–Kutta, τα b_i λέγονται *βάρη* και τα c_i κόμβοι. Οι τιμές τους για μια μέθοδο Runge–Kutta τάξης p προσδιορίζονται, αν και όχι μονοσήμαντα, από την απαίτηση η τιμή για το y_{i+1} που υπολογίζει η (7.7) ή η (7.8), να διαφέρει κατά ένα όρο $\mathcal{O}(h^{p+1})$, το πολύ, από την τιμή που υπολογίζει η μέθοδος σειράς Taylor κρατώντας μέχρι και την παράγωγο τάξης p.

Οι συντελεστές b_i καιc_i ικανοποιούν υποχρεωτικά τις συνθήκες $\sum_{i=1}^{s} b_i = 1$ και $c_i \leq 1$. Σε όλες σχεδόν τις μεθόδους ισχύει, μεταξύ άλλων σχέσεων, και ότι $c_i = \sum_{j=1}^{s} a_{ij}$.

Το πλήθος s των συντελεστών k για τις μεθόδους explicit σχετίζεται με την τάξη p της μεθόδου με τις σχέσεις $s \geq p$ για $p \leq 4$ και $s > p$ για $p > 4$. Για τις μεθόδους implicit μπορεί να ισχύει $p > s$.

7.3.1 Παραδείγματα

Η κλασική explicit μέθοδος Runge–Kutta δεύτερης τάξης λέγεται και μέθοδος Heun και έχει εξισώσεις

$$\begin{aligned} y_{i+1} &= y_i + \frac{1}{2}(k_1 + k_2) \ , && \text{(7.9α΄)} \\ k_1 &= hf(x_i, y_i) \ , && \text{(7.9β΄)} \\ k_2 &= hf(x_i + h, y_i + k_1) \ . && \text{(7.9γ΄)} \end{aligned}$$

Ένα άλλο σύνολο συντελεστών δίνει τη μέθοδο Ralston, πάλι Runge–Kutta δεύτερης τάξης, η οποία έχει εξισώσεις

$$y_{i+1} = y_i + \frac{1}{4}(k_1 + 3k_2) \,, \tag{7.10α′}$$
$$k_1 = hf(x_i, y_i) \,, \tag{7.10β′}$$
$$k_2 = hf(x_i + 2h/3, y_i + 2k_1/3) \tag{7.10γ′}$$

και βελτιωμένο σφάλμα σε σχέση με την Heun. Και στις δύο βέβαια, το τοπικό σφάλμα είναι ανάλογο του h^3 και το ολικό, ανάλογο του h^2.

Η κλασική explicit μέθοδος Runge–Kutta τέταρτης τάξης (RK4) έχει συντελεστές $c_2 = c_3 = 1/2$, $c_4 = 1$, $b_1 = b_4 = 1/6$, $b_2 = b_3 = 1/3$ και $a_{21} = a_{32} = 1/2$, $a_{43} = 1$, $a_{31} = a_{41} = a_{42} = 0$, δηλαδή οι εξισώσεις της είναι

$$y_{i+1} = y_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \,, \tag{7.11α′}$$
$$k_1 = hf(x_i, y_i) \,, \tag{7.11β′}$$
$$k_2 = hf(x_i + h/2, y_i + k_1/2) \,, \tag{7.11γ′}$$
$$k_3 = hf(x_i + h/2, y_i + k_2/2) \,, \tag{7.11δ′}$$
$$k_4 = hf(x_i + h, y_i + k_3) \,. \tag{7.11ε′}$$

Το τοπικό σφάλμα της συγκεκριμένης μεθόδου είναι ανάλογο του h^5 και επομένως το ολικό είναι ανάλογο του h^4.

Μια τροποποίηση της RK4, πάλι μέθοδος τέταρτης τάξης, η Runge–Kutta $3/8$, είναι η ακόλουθη:

$$y_{i+1} = y_i + \frac{1}{8}(k_1 + 3k_2 + 3k_3 + k_4) \,, \tag{7.12α′}$$
$$k_1 = hf(x_i, y_i) \,, \tag{7.12β′}$$
$$k_2 = hf(x_i + h/3, y_i + k_1/3) \,, \tag{7.12γ′}$$
$$k_3 = hf(x_i + 2h/3, y_i - k_1/3 + k_2) \,, \tag{7.12δ′}$$
$$k_4 = hf(x_i + h, y_i + k_1 - k_2 + k_3) \,. \tag{7.12ε′}$$

Η συγκεκριμένη επιλογή συντελεστών ελαττώνει το τοπικό σφάλμα ως προς την κλασική RK4, διατηρώντας το βέβαια ανάλογο του h^5.

7.3.2 Butcher tableau

Οι explicit μέθοδοι Runge–Kutta κωδικοποιούνται γράφοντας τους συντελεστές a_{ij}, b_i, c_i στον ακόλουθο πίνακα

$$\begin{array}{c|cccccc} 0 & 0 & & & & & \\ c_2 & a_{21} & & & & & \\ c_3 & a_{31} & a_{32} & & & & \\ \vdots & \vdots & \ddots & & & & \\ c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} & & \\ \hline & b_1 & b_2 & \cdots & b_{s-1} & b_s \end{array} \,.$$

Ο πίνακας αυτός λέγεται *Butcher tableau*.

Για παράδειγμα, η κλασική RK4 έχει το ακόλουθο Butcher tableau:

$$\begin{array}{c|cccc} 0 & 0 & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array} .$$

Η μέθοδος Runge–Kutta $3/8$ έχει το ακόλουθο Butcher tableau:

$$\begin{array}{c|cccc} 0 & 0 & & & \\ 1/3 & 1/3 & & & \\ 2/3 & -1/3 & 1 & & \\ 1 & 1 & -1 & 1 & \\ \hline & 1/8 & 3/8 & 3/8 & 1/8 \end{array} .$$

Για τις implicit μεθόδους Runge–Kutta το Butcher tableau έχει τη μορφή

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1,s-1} & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2,s-1} & a_{2s} \\ c_3 & a_{31} & a_{32} & \cdots & a_{3,s-1} & a_{3s} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_{s-1} & b_s \end{array} .$$

Παρατηρήστε ότι στις μεθόδους που είναι explicit, ο πίνακας των συντελεστών a_{ij} έχει μη μηδενικά στοιχεία μόνο κάτω από τη διαγώνιο ενώ στις implicit είναι πλήρης ή τουλάχιστον έχει μη μηδενική διαγώνιο.

Μια μέθοδος που είναι explicit μπορεί να μετατραπεί σε implicit κάνοντας την αλλαγή $a_{ij} \to -a_{ij} + b_j$, $c_i \to 1 - c_i$. Οι σχέσεις αυτές προκύπτουν από την αλλαγή $x_i \leftrightarrow x_{i+1}$, $y_i \leftrightarrow y_{i+1}$.

Για παράδειγμα, η explicit μέθοδος Heun που έχει Butcher tableau το

$$\begin{array}{c|cc} 0 & 0 & \\ 1 & 1 & \\ \hline & 1/2 & 1/2 \end{array}$$

παράγει μια implicit μέθοδο δεύτερης τάξης με Butcher tableau το

$$\begin{array}{c|cc} 1 & 1/2 & 1/2 \\ 0 & -1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array} .$$

Η μέθοδος είναι η πρώτη της οικογένειας Lobatto IIIC.

Προσέξτε ότι η μέθοδος explicit Euler που παρουσιάσαμε παραπάνω έχει τη γενική μορφή των μεθόδων Runge–Kutta. Το Butcher tableau της forward Euler είναι

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

ενώ της backward

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

7.3.3 Ευστάθεια μεθόδων Runge–Kutta

Έστω ότι η λύση της διαφορικής εξίσωσης (7.2) είναι η συνάρτηση $y^*(x)$. Ας εκφράσουμε τη συνάρτηση $y(x)$ ως μια μικρή απομάκρυνση $\epsilon(x)$ από τη λύση $y^*(x)$. Η διαφορική εξίσωση γίνεται

$$\begin{aligned} (y^*(x)+\epsilon(x))' &= f(x, y^*(x)+\epsilon(x)) \Rightarrow \\ (y^*(x))' + \epsilon'(x) &\approx f(x, y^*(x)) + \epsilon(x) \left.\frac{\partial f(x,y)}{\partial y}\right|_{y^*(x)} \Rightarrow \\ \epsilon'(x) &\approx \epsilon(x) f_y(x, y^*) \ . \end{aligned}$$

Καταλήξαμε στη διαφορική εξίσωση που ισχύει για το σφάλμα όταν ξεκινούμε από το σημείο x. Θεωρήσαμε ότι $f_y(x, y^*) \neq 0$.

Μπορούμε να θέσουμε $\lambda = f_y(x, y^*(x))$ που σε πρώτη προσέγγιση είναι μια σταθερή ποσότητα, με τιμή όση στο σημείο εκκίνησης, $f_y(x, y^*(x)) \approx f_y(x, y(x))$. Προκύπτει έτσι η γραμμική εξίσωση $\epsilon' \approx \lambda\epsilon$.

Όλες οι μέθοδοι Runge–Kutta, είτε implicit είτε explicit, όταν εφαρμοστούν στην γραμμική διαφορική εξίσωση $\epsilon' = \lambda\epsilon$ καταλήγουν στην έκφραση

$$\epsilon_{i+1} = R(h\lambda)\epsilon_i \ .$$

Η μιγαδική συνάρτηση μιγαδικής μεταβλητής

$$R(z) = 1 + z\boldsymbol{e}\cdot(\boldsymbol{I} - z\boldsymbol{A})^{-1}\cdot\boldsymbol{b}^T = \frac{\det(\boldsymbol{I} - z\boldsymbol{A} + z\boldsymbol{e}\cdot\boldsymbol{b}^T)}{\det(\boldsymbol{I} - z\boldsymbol{A})}$$

ονομάζεται *συνάρτηση ευστάθειας* της συγκεκριμένης μεθόδου. Ο πίνακας $\boldsymbol{A}$, διάστασης $s \times s$, και το διάνυσμα $\boldsymbol{b}$, διάστασης s, έχουν στοιχεία τους συντελεστές a_{ij}, b_i μιας μεθόδου s σταδίων από την οικογένεια Runge–Kutta και $\boldsymbol{e}$ είναι το διάνυσμα s στοιχείων με την τιμή 1 σε όλα τα στοιχεία του.

Η συνάρτηση $R(z)$ είναι γενικά ρητή, δηλαδή λόγος πολυωνύμων $P(z)/Q(z)$. Στις μεθόδους με s στάδια που είναι explicit η $R(z)$ είναι πολυώνυμο βαθμού s, δηλαδή έχει $Q(z) = 1$.

Τα σημεία z του μιγαδικού επιπέδου στα οποία ισχύει $|R(z)| \leq 1$ αποτελούν την περιοχή ευστάθειας της μεθόδου. Όταν επιλέξουμε $z \equiv h\lambda$ σε αυτή την περιοχή, η απομάκρυνση ϵ από τη λύση παραμένει σε περιοχή γύρω από τη λύση.

Παράδειγμα

Η συνάρτηση ευστάθειας για τη forward Euler είναι $R(z) = 1 + z$. Η περιοχή ευστάθειάς της είναι τα σημεία $z \equiv h\lambda$ που ικανοποιούν τη σχέση $|1 + z| \leq 1$, δηλαδή τα σημεία κυκλικού δίσκου με κέντρο -1 και ακτίνα 1. Η επιλογή του βήματος h ώστε το $h\lambda$ να ανήκει σε αυτό το δίσκο δίνει ευσταθή μέθοδο forward Euler. Αν το λ είναι πραγματικό θέλουμε να ισχύει $-1 \leq 1 + h\lambda \leq 1$ δηλαδή $-2 \leq h\lambda \leq 0$. Επομένως, ευσταθή forward Euler μπορούμε να έχουμε μόνο όταν $\lambda < 0$ και επιλέξουμε $h < 2/|\lambda|$.

Η συνάρτηση ευστάθειας για την backward Euler είναι $R(z) = 1/(1 - z)$. Η περιοχή ευστάθειας είναι τα σημεία z που ικανοποιούν τη σχέση $|1 - z| > 1$ δηλαδή τα σημεία *έξω* από το δίσκο με κέντρο το 1 και ακτίνα 1. Για να έχουμε ευσταθή backward Euler πρέπει λοιπόν να έχουμε $h\lambda < 0$ ή $h\lambda > 2$. Αν $\lambda < 0$ μπορούμε να επιλέξουμε οποιοδήποτε θετικό h ενώ αλλιώς το h πρέπει να είναι μεγαλύτερο του $2/\lambda$.

7.4 Υπολογισμός με ολοκλήρωμα

Παρατηρήστε ότι η (7.2) μπορεί να λυθεί με ολοκλήρωση:

$$\begin{aligned} y' = f(x,y) \quad &\Rightarrow \quad \mathrm{d}y = f(x,y)\,\mathrm{d}x \Rightarrow \int_{y_i}^{y_{i+1}} \mathrm{d}y = \int_{x_i}^{x_{i+1}} f(x,y)\,\mathrm{d}x \\ &\Rightarrow \quad y_{i+1} - y_i = \int_{x_i}^{x_{i+1}} f(x,y)\,\mathrm{d}x\ . \end{aligned} \tag{7.13}$$

7.4.1 Μέθοδος τραπεζίου/Crank–Nicolson

Ας εφαρμόσουμε τον κανόνα τραπεζίου για τον υπολογισμό του ολοκληρώματος:

$$y_{i+1} \approx y_i + \frac{x_{i+1} - x_i}{2}(f(x_i, y_i) + f(x_{i+1}, y_{i+1}))\ . \tag{7.14}$$

Προέκυψε ο κανόνας τραπεζίου για την επίλυση διαφορικών εξισώσεων, μια μέθοδος που είναι implicit και ουσιαστικά είναι η *μέθοδος Crank–Nicolson* που αναπτύχθηκε για τις διαφορικές εξισώσεις με μερικές παραγώγους. Είναι μέλος της οικογένειας Runge–Kutta με Butcher tableau

$$\begin{array}{c|cc} 0 & 0 & \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}$$

Το τοπικό σφάλμα είναι αυτό που προβλέπεται από τη μέθοδο τραπεζίου, δηλαδή ανάλογο του h^3. Το ολικό σφάλμα μετά από πολλές επαναλήψεις είναι ανάλογο του h^2. Η συνάρτηση ευστάθειας είναι $R(z) = (2 + z)/(2 - z)$ δηλαδή η μέθοδος είναι ευσταθής όταν $Re(z) < 0$.

Παρατηρήστε ότι στον ίδιο τύπο καταλήγουμε όταν προσθέσουμε τους τύπους forward και backward Euler (ή τους αντίστοιχους Taylor). Το τοπικό σφάλμα του τύπου φαινομενικά μόνο προκύπτει ανάλογο του h^2: Ο όρος του σφάλματος, όπως προκύπτει από τους τύπους Taylor, είναι

$$\frac{(x_{i+1}-x_i)^2}{2!}\left((f_x+f_yf)|_{x_i}-(f_x+f_yf)|_{x_{i+1}}\right)\propto(x_{i+1}-x_i)^3\ .$$

Το ανάπτυγμα Taylor της συνάρτησης f_x+f_yf στο σημείο x_i, όταν υπολογιστεί στο x_{i+1}, δικαιολογεί τον επιπλέον συντελεστή $(x_{i+1}-x_i)$.

7.5 Συστηματική κατασκευή μεθόδων Runge–Kutta

Μπορούμε να κατασκευάσουμε πολλές μεθόδους Runge–Kutta αν επιλέξουμε s συντελεστές c_i, $i=1,\ldots,s$ στο διάστημα $[0,1]$, ή ισοδύναμα, s σημεία της μορφής $\bar{x}_i=x_0+c_ih$ για τον υπολογισμό των συντελεστών k_i. Οι συντελεστές a_{ij} και b_i υπολογίζονται από τις ακόλουθες σχέσεις:

$$a_{ij}=\int_0^{c_i}\ell_j(t)\,\mathrm{d}t\quad\text{και}\quad b_i=\int_0^1\ell_i(t)\,\mathrm{d}t\ ,$$

όπου $\ell_i(t)$ είναι το πολυώνυμο της βάσης Lagrange

$$\ell_i(t)=\prod_{j=1,j\neq i}^{s}\frac{t-c_j}{c_i-c_j}\ .$$

Εφαρμογή

Αν επιλέξουμε $c_1=0$, $c_2=1/2$, $c_3=1$ (δηλαδή τα άκρα του διαστήματος και το μέσο του) έχουμε

$$\begin{array}{c|ccc}0&0&0&0\\1/2&5/24&1/3&-1/24\\1&1/6&2/3&1/6\\\hline&1/6&2/3&1/6\end{array}\ ,$$

μια μέθοδο 4ης τάξης της οικογένειας Lobatto IIIA.

Προσέξτε ότι τα σημεία c_i είναι αυτά που χρησιμοποιούμε στον υπολογισμό ολοκληρώματος με τη μέθοδο Simpson. Οι συντελεστές b_i είναι ουσιαστικά οι συντελεστές του απλού τύπου Simpson (5.8), αν λάβουμε υπόψη ότι το h που χρησιμοποιούσαμε στην ολοκλήρωση με τη μέθοδο Simpson είναι το μισό του h που χρησιμοποιούμε εδώ.

Το τοπικό σφάλμα είναι 5ης τάξης, όσο αναμένουμε από τον απλό τύπο Simspon, ενώ το συνολικό σφάλμα είναι ανάλογο του h^4.

7.6 Συστήματα Διαφορικών Εξισώσεων

Οι μέθοδοι αριθμητικής επίλυσης ΔΕ που εξετάσαμε, μπορούν εύκολα να εφαρμοστούν στην περίπτωση συστημάτων ΔΕ ή μιας ΔΕ υψηλότερης τάξης (≥ 2). Έστω το σύστημα των ΔΕ πρώτης τάξης

$$\begin{aligned} y_1' &= f_1(x, y_1, y_2, \ldots, y_n) \,, && \text{(7.15α΄)} \\ y_2' &= f_2(x, y_1, y_2, \ldots, y_n) \,, && \text{(7.15β΄)} \\ &\vdots \quad \vdots \\ y_n' &= f_n(x, y_1, y_2, \ldots, y_n) \,, && \text{(7.15γ΄)} \end{aligned}$$

όπου οι f_i είναι πραγματικές συναρτήσεις, ορισμένες για $x \in [a, b]$ και για κάθε πραγματικό y_1, y_2,..., y_n. Οι τιμές των y_k στο βήμα $(i+1)$ θα υπολογιστούν από τις τιμές των y_k' και τις προηγούμενες τιμές των y_k, y_k', με τον ίδιο τρόπο όπως στις απλές ΔΕ. Ο υπολογισμός των y_k' από την (7.15) είναι το μόνο σημείο στο οποίο υπάρχει διαφορά από την απλή περίπτωση καθώς έχουμε n τιμές y_i αντί για μία, όπως στις απλές ΔΕ. Γενικά, η λύση του (7.15), αν υπάρχει, δε θα είναι μοναδική, εκτός αν προσδιοριστούν n αρχικές συνθήκες:

$$y_k(x_0) = s_k \,, \qquad i = 1, \ldots, n \,, \tag{7.16}$$

όπου τα s_i είναι γνωστά και $x_0 \in [a, b]$. Οι (7.15,7.16) συνιστούν ένα πρόβλημα αρχικών τιμών που σε διανυσματική μορφή γράφεται:

$$\begin{aligned} \mathbf{y}' &= \mathbf{f}(x, \mathbf{y}) \,, && \text{(7.17α΄)} \\ \mathbf{y}(x_0) &= \mathbf{s} \,. && \text{(7.17β΄)} \end{aligned}$$

Οι διάφορες μέθοδοι επίλυσης διαφορικών εξισώσεων πρώτης τάξης γενικεύονται εύκολα για συστήματα διαφορικών εξισώσεων πρώτης τάξης. Η μέθοδος Taylor θα είναι:

$$\mathbf{y}(x_{i+1}) = \mathbf{y}(x_i) + h\mathbf{f}\left(x_i, \mathbf{y}(x_i)\right) + \frac{h^2}{2!}\mathbf{f}'\left(x_i, \mathbf{y}(x_i)\right) + \cdots + \frac{h^p}{p!}\mathbf{f}^{(p)}\left(x_i, \mathbf{y}(x_i)\right) + \mathbf{R}_{p+1} \,,$$

όπου

$$\begin{aligned} \mathbf{f}'\left(x, \mathbf{y}(x)\right) \equiv \frac{\mathrm{d}}{\mathrm{d}x}\mathbf{f}\left(x, \mathbf{y}(x)\right) &= \frac{\partial \mathbf{f}}{\partial x} + \sum_{j=1}^{n} \frac{\mathrm{d}y_j(x)}{\mathrm{d}x}\frac{\partial \mathbf{f}}{\partial y_j} = \frac{\partial \mathbf{f}}{\partial x} + \sum_{j=1}^{n} f_j \frac{\partial \mathbf{f}}{\partial y_j} \,, \\ \frac{\partial \mathbf{f}}{\partial y_j} &\equiv \left(\frac{\partial f_1}{\partial y_j}, \frac{\partial f_2}{\partial y_j}, \ldots, \frac{\partial f_n}{\partial y_j}\right) \,. \end{aligned}$$

Η μέθοδος Taylor δεύτερου βαθμού θα είναι:

$$\begin{aligned} \mathbf{y}(x_{i+1}) &= \mathbf{y}(x_i) + h\mathbf{f}\left(x_i, \mathbf{y}_i\right) + \frac{h^2}{2!}\mathbf{f}'\left(x_i, \mathbf{y}_i\right) \,, \\ \mathbf{y}_0 &= \mathbf{s} \,. \end{aligned}$$

Η κλασική Runge–Kutta τέταρτης τάξης θα είναι:

$$\mathbf{y}_{i+1} = \mathbf{y}_i + \frac{1}{6}\left(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4\right)$$

με

$$\begin{aligned}
\mathbf{k}_1 &= h\mathbf{f}\left(x_i, \mathbf{y}_i\right) \ , \\
\mathbf{k}_2 &= h\mathbf{f}\left(x_i + {}^h\!/_2, \mathbf{y}_i + {}^h\!/_2\mathbf{k}_1\right) \ , \\
\mathbf{k}_3 &= h\mathbf{f}\left(x_i + {}^h\!/_2, \mathbf{y}_i + {}^h\!/_2\mathbf{k}_2\right) \ , \\
\mathbf{k}_4 &= h\mathbf{f}\left(x_i + h, \mathbf{y}_i + h\mathbf{k}_3\right) \ .
\end{aligned}$$

Παράδειγμα

Να λυθεί το πρόβλημα αρχικών τιμών

$$\begin{aligned}
y_1' &= xy_1 - y_2 \\
y_2' &= -y_1 + y_2 \\
y_1(0) &= s_1 \\
y_2(0) &= s_2
\end{aligned}$$

Ορίζουμε $y_{1\,0} \equiv y_1(0)$ και $y_{2\,0} \equiv y_2(0)$. Η σειρά Taylor δεύτερου βαθμού είναι:

$$\begin{aligned}
y_{1\,i+1} &= y_{1\,i} + h\left(x_r y_{1\,i} - y_{2\,i}\right) + \frac{h^2}{2}\left[\left(x_i^2 + 2\right) y_{1\,i} - \left(1 + x_i\right) y_{2\,i}\right] \\
y_{2\,i+1} &= y_{2\,i} + h\left(-y_{1\,i} + y_{2\,i}\right) + \frac{h^2}{2}\left[-\left(1 + x_i\right) y_{1\,i} + 2y_{2\,i}\right] \ ,
\end{aligned}$$

με $i = 0, 1, \ldots$.

Η μέθοδος Heun δίνει:

$$\begin{aligned}
k_1 &= h(x_i y_{1\,i} - y_{2\,i}) \\
\ell_1 &= h(y_{2\,i} - y_{1\,i}) \\
k_2 &= h((x_i + h)(y_{1\,i} + k_1) - (y_{2\,i} + \ell_1)) \\
\ell_2 &= h((y_{2\,i} + \ell_1) - (y_{1\,i} + k_1)) \\
y_{1\,i+1} &= y_{1\,i} + \frac{1}{2}(k_1 + k_2) \\
y_{2\,i+1} &= y_{2\,i} + \frac{1}{2}(\ell_1 + \ell_2)
\end{aligned}$$

7.6.1 Επιλυσιμότητα

Το (7.15) έχει μία μοναδική λύση αν n $\mathbf{f}$ ικανοποιεί μια συνθήκη Lipschitz (§7.1.1), δηλαδή, $\exists L > 0$ τέτοιο ώστε $\forall \mathbf{y}, \mathbf{z} \in \mathcal{R}^n$ και για κάθε $x \in [a, b]$ να ισχύει

$$\left\|\mathbf{f}(x, \mathbf{y}) - \mathbf{f}(x, \mathbf{z})\right\|_\infty \leq L \left\|\mathbf{y} - \mathbf{z}\right\|_\infty \ .$$

7.7 Διαφορικές εξισώσεις ανώτερης τάξης

Μια διαφορική εξίσωση τάξης $n \geq 2$ μπορεί να γραφτεί ως ένα σύστημα n διαφορικών εξισώσεων πρώτης τάξης αν κάθε παράγωγο μικρότερη της ανώτερης την αποδώσουμε σε νέα συνάρτηση. Έτσι π.χ. η διαφορική εξίσωση

$$y'' = f(x, y, y') \ ,$$

με $y(x_0) = y_0$ και $y'(x_0) = d_0$, τροποποιείται ως εξής: θέτουμε $z \equiv y'$ οπότε η διαφορική εξίσωση γίνεται

$$z' = f(x, y, z) \ .$$

Συμπληρώνεται με την εξίσωση $y' = z$ ώστε να δημιουργηθεί το ακόλουθο σύστημα:

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

με $y(x_0) = y_0$ και $z(x_0) = d_0$.

Η λύση του συστήματος υπολογίζει και τη λύση της αρχικής εξίσωσης.

7.8 Ασκήσεις

1. Εφαρμόστε τη μέθοδο forward Euler για την επίλυση της διαφορικής εξίσωσης

   $$y' = \cos x - x \sin x$$

   στο διάστημα $[0, 3]$, με $y(0) = 2.0$. Τυπώστε τη λύση ανά $h = 0.01$. Συγκρίνετε με την ακριβή λύση, $y(x) = 2 + x \cos x$.

2. Χρησιμοποιήστε τη μέθοδο backward Euler για να βρείτε προσεγγιστικά την τιμή της συνάρτησης $y(x)$ στο σημείο $x = 0.2$ αν στο $x = 0$ έχει τιμή 1.0 και ικανοποιεί τη σχέση $0.02y' + y - \cos x = 0$. Συγκρίνετε με τη σωστή λύση $y(x) = (\mathrm{e}^{-50x} + 2500 \cos(x) + 50 \sin(x))/2501$.

3. Να επιλύσετε τη διαφορική εξίσωση $y' = -10y$, με $y(0) = 10$ στο διάστημα $[0, 0.5]$ με τις μέθοδους forward και backward Euler και βήματα $h = 10^{-2}$, $h = 10^{-3}$, $h = 10^{-4}$. Συγκρίνετε με την ακριβή λύση, $y(x) = 10 \exp(-10x)$.

4. Εφαρμόστε τις μεθόδους forward και backward Euler για την επίλυση της διαφορικής εξίσωσης

   $$y' = \cos x - \sin y + x^2$$

   στο διάστημα $[-1, 1]$, με $y(-1) = 3.0$. Τυπώστε τη λύση ανά $h = 0.01$.

5. Εφαρμόστε τη μέθοδο Taylor με 5 όρους για την επίλυση της διαφορικής εξίσωσης της προηγούμενης άσκησης.

6. Δείξτε ότι όταν εφαρμοστεί η μέθοδος Heun σε διαφορική εξίσωση της μορφής $y' = \lambda x$ δίνει την ακριβή λύση.

7. Να βρείτε την τιμή $y(0.6)$ αν η $y(x)$ ικανοποιεί τη διαφορική εξίσωση $y' = x^2 + x - y$ με $y(0) = 0$. Χρησιμοποιήστε τις μεθόδους Heun και Ralston. Συγκρίνετε με την τιμή που υπολογίζεται από τη λύση $y(x) = 1 - \mathrm{e}^{-x} + x^2 - x$.

8. Γράψτε συνάρτηση που να δέχεται ένα Butcher tableau (δηλαδή τους πίνακες $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{c}$) κάποιας explicit μεθόδου Runge–Kutta, το αρχικό σημείο x_0, την τιμή της συνάρτησης y_0 εκεί και το τελικό σημείο x_1 και να υπολογίζει την τιμή y_1 εφαρμόζοντας τη μέθοδο Runge–Kutta που περιγράφεται στο συγκεκριμένο tableau.

   Χρησιμοποιήστε τη για να εφαρμόσετε την κλασική Runge–Kutta τέταρτης τάξης ώστε να βρείτε την τιμή $y(2)$ όταν η συνάρτηση $y(x)$ ικανοποιεί τη διαφορική εξίσωση

$$\begin{aligned} y' &= (y + x)/(y - x) \,, \\ y(0) &= 1 \,. \end{aligned}$$

   Συγκρίνετε με την ακριβή λύση, $y(x) = x + \sqrt{1 + 2x^2}$.

   Μπορείτε να τροποποιήσετε τον κώδικά σας για implicit Runge–Kutta;

9. Χρησιμοποιήστε τις μεθόδους RK4 και RK $3/8$ για να επιλύσετε τη διαφορική εξίσωση

$$y' = \frac{y}{x}\left(1 - \frac{y}{x}\right)$$

   στο διάστημα $[1, 3]$, με $y(1) = 2$. Τυπώστε τις τιμές με βήμα $h = 0.1$, καθώς και το σφάλμα ως προς την ακριβή λύση

$$y(x) = \frac{x}{0.5 + \ln x} \,.$$

   Ποια μέθοδος έχει μικρότερο σφάλμα;

10. Η συνάρτηση $y(x)$ ικανοποιεί τη διαφορική εξίσωση

$$y' = -\frac{xy}{1 - x^2} \,.$$

    Ποια τιμή πρέπει να έχει στο $x = 0$ ώστε στο $x = 0.5$ να έχει τιμή 1;

11. Η συνάρτηση $y(x)$ ικανοποιεί τη διαφορική εξίσωση

$$y' = \frac{x - 2y}{x + 2y} \,.$$

    Ποια τιμή πρέπει να έχει στο $x = 0$ ώστε στο $x = 2$ να έχει την ίδια τιμή;

Υπόδειξη: Σχηματίστε τη συνάρτηση $g(y_{\text{αρχικό}})$. Αυτή θα δέχεται την αρχική τιμή του y, θα λύνει τη διαφορική εξίσωση ώστε να βρεί το $y_{\text{τελικό}}$ και θα επιστρέφει τη διαφορά $y_{\text{τελικό}} - y_{\text{αρχικό}}$. Κατόπιν, βρείτε για ποιο $y_{\text{αρχικό}}$ μηδενίζεται.

12. Εφαρμόστε τη μέθοδο Taylor με 4 όρους για την επίλυση του συστήματος ΔΕ

$$\begin{aligned} y' &= y + z^2 - x^3 \\ z' &= z + y^3 + \cos x \end{aligned}$$

με αρχικές συνθήκες, στο $x = 0$, $y = 0.3$ και $z = 0.1$. Τυπώστε τις τιμές των y, z στο διάστημα $[0, 1]$ με βήμα 0.01.

13. Να εφαρμόσετε μια μέθοδο Runge–Kutta $2^{\text{ης}}$ τάξης για την εύρεση της κίνησης σώματος μάζας $m = 2\,\text{kg}$, εξαρτώμενου από ελατήριο με δύναμη επαναφοράς $F(x) = x - 0.01x^3$. Το σώμα αφήνεται για $t = 0$ ελεύθερο, χωρίς αρχική ταχύτητα, στη θέση $x = 2.5\,\text{cm}$.

14. Να βρείτε την κίνηση εκκρεμούς για το οποίο ισχύει

$$\ddot{\theta} = -\sin\theta\ ,$$

όπου θ η γωνία απομάκρυνσης από την κάθετο. Το εκκρεμές αφήνεται ελεύθερο, χωρίς αρχική ταχύτητα σε γωνία $\theta = 45^\circ$.

Για μικρές γωνίες θ ισχύει $\sin\theta \approx \theta$. Εφαρμόστε την προσέγγιση αυτή και συγκρίνετε τη λύση της νέας ΔΕ με τη λύση της ακριβούς ΔΕ.

15. Να λύσετε τη ΔΕ $\psi'' = (x^2 - 5)\psi$ με αρχική συνθήκη $\psi(0) = -(2\sqrt{\pi})^{-1/2}$, $\psi'(0) = 0$. Τυπώστε 100 ισαπέχουσες τιμές στο διάστημα $[-2, 2]$.

Υπόδειξη Να λύσετε δύο προβλήματα αρχικών τιμών, τη ΔΕ στα διαστήματα $[0, 2]$ και $[-2, 0]$.

16. Να λυθεί το σύστημα

$$\begin{aligned} y_1' &= 2y_1 - 2y_2 + 3y_3 \\ y_2' &= y_1 + y_2 + y_3 \\ y_3' &= y_1 + 3y_2 - y_3 \end{aligned}$$

με αρχικές συνθήκες (στο $t = 0$) $y_1 = -2$, $y_2 = 30$, $y_3 = 0$. Δίνεται ότι

$$\begin{bmatrix} 1 & 11 & 1 \\ -1 & 1 & 1 \\ -1 & -14 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1/2 & -5/6 & 1/3 \\ 0 & 1/15 & -1/15 \\ 1/2 & 1/10 & 2/5 \end{bmatrix}\ .$$

Παράρτημα α΄

Χρήσιμα Ολοκληρώματα

Γνωρίζουμε ότι ισχύουν οι ακόλουθες τριγωνομετρικές σχέσεις:

$$\begin{aligned}
\sin x \sin y &= \frac{\cos(x-y) - \cos(x+y)}{2} \,, \\
\cos x \cos y &= \frac{\cos(x-y) + \cos(x+y)}{2} \,, \\
\sin x \cos y &= \frac{\sin(x-y) + \sin(x+y)}{2} \,.
\end{aligned}$$

Επίσης,

$$\int \sin x \, \mathrm{d}x = -\cos x + c \,, \qquad \int \cos x \, \mathrm{d}x = \sin x + c \,.$$

Χρησιμοποιώντας αυτές, εύκολα μπορούν να δειχθούν οι παρακάτω σχέσεις (για ακέραιο $n \geq 0$). Αποδείξτε τις!

$$\int_0^L \sin\left(\frac{2n\pi x}{L}\right) \mathrm{d}x = 0 \,, \tag{α΄.1}$$

$$\int_0^L \cos\left(\frac{2n\pi x}{L}\right) \mathrm{d}x = \begin{cases} L, & n = 0 \\ 0, & n > 0 \end{cases} \,, \tag{α΄.2}$$

$$\int_0^L \sin\left(\frac{2n\pi x}{L}\right) \cos\left(\frac{2k\pi x}{L}\right) \mathrm{d}x = 0 \,, \tag{α΄.3}$$

$$\int_0^L \cos\left(\frac{2n\pi x}{L}\right) \cos\left(\frac{2k\pi x}{L}\right) \mathrm{d}x = \begin{cases} 0, & n \neq k \\ L, & n = k = 0 \\ L/2, & n = k > 0 \end{cases} \,, \tag{α΄.4}$$

$$\int_0^L \sin\left(\frac{2n\pi x}{L}\right) \sin\left(\frac{2k\pi x}{L}\right) \mathrm{d}x = \begin{cases} 0, & n \neq k \\ 0, & n = k = 0 \\ L/2, & n = k > 0 \end{cases} \,. \tag{α΄.5}$$

Κατάλογος πινάκων

Ευρετήριο