ΑΡΙΘΜΗΤΙΚΗ ΑΝΑΛΥΣΗ

(ΣΗΜΕΙΩΣΕΙΣ ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΜΑΘΗΜΑΤΟΣ)

ΒΑΣΙΛΕΙΟΣ Α. ΔΟΥΓΑΛΗΣ

Μαθηματικό Τμήμα Πανεπιστημίου Αθηνών

ΑΘΗΝΑ 1995

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΡΟΛΟΓΟΣ

Οι σημειώσεις αυτές γράφτηκαν και χρησιμοποιήθηκαν ως κύριο διδακτικό βοήθημα γιά το μεταπτυχιακό μάθημα 350 "Αριθμητική Ανάλυση" που δίδαξα κατά το χειμερινό εξάμηνο τού ακαδημαϊκού έτους 1986-7 στο Μαθηματικό Τμήμα του Πανεπιστημίου Κρήτης σε μεταπτυχιακούς και προχωρημένους προπτυχιακούς φοιτητές. Σκοπός του μαθήματος ήταν να εκθέσει το ακροατήριο, μέσα στη χρονική διάρκεια ενός διδακτικού εξαμήνου, σε αποδείξεις ορισμένων θεωρημάτων κεντρικής σημασίας σε μερικές βασικές περιοχές της κλασικής Αριθμητικής Ανάλυσης, όπως αριθμητική γραμμική και μη γραμμική άλγεβρα, αριθμητική λύση συνήθων διαφορικών εξισώσεων και θεωρία παρεμβολής και προσέγγισης συναρτήσεων. (Μιά εισαγωγή σε θέματα αριθμητικής λύσης μερικών διαφορικών εξισώσεων γίνεται σε άλλο μεταπτυχιακό μάθημα του Μαθηματικού Τμήματος).

Η επιλογή των θεμάτων από την ευρύτατη περιοχή της Αριθμητικής Ανάλυσης, που είναι δυνατόν να διδαχθούν σ' ένα τέτοιο μάθημα μέσα σ' ένα εξάμηνο, είναι προφανές ότι είναι σ' ένα βαθμό αυθαίρετη και αντανακλά πιθανότατα τις προσωπικές προτιμήσεις και τα ενδιαφέροντα του διδάσκοντος. Δεν έχουν λοιπόν οι σημειώσεις αυτές αξιώσεις πληρότητας.

Στο κεφάλαιο 1 (Αριθμητική Γραμμική 'Αλγεβρα) εξετάζεται η αριθμητική λύση συστημάτων γραμμικών εξισώσεων. Στο πρώτο μέρος του το κεντρικό αποτέλεσμα είναι η αντίστροφη ανάλυση του Wilkinson γιά την μελέτη της επιρροής των σφαλμάτων στρογγύλευσης στην απαλοιφή Gauss. Ακολούθησα την απόδειξη που δίνεται στο κλασικό βιβλίο των Forsythe και Moler [1.2] (βλ. Συνοπτική Βιβλιογραφία). Το δεύτερο μέρος του κεφαλαίου αφορά την αριθμητική λύση γραμμικών συστημάτων με πραγματικούς, συμμετρικούς και θετικά ορισμένους πίνακες με ανάλυση Cholesky και, κατά κύριο λόγο, με μεθόδους ελαχιστοποίησης (μεθόδους καθόδου μεγίστης κλίσεως και συζυγών κλίσεων). Ακολούθησα την ανάπτυξη του θέματος που κάνουν οι Golub και Van Loan, [1.4], συμπληρώνοντας τις αποδείξεις σύγκλισης και φράγματος του σφάλματος γιά τις μεθόδους ελαχιστοποίησης. Λόγω ελλείψεως χρόνου δεν εξετάσθηκαν οι κλασικές επαναληπτικές μέθοδοι, καθώς και μέθοδοι γιά το γραμμικό πρόβλημα ελαχίστων τετραγώνων και γιά το πρόβλημα ιδιοτιμών. (Μερικά θέματα που παρέλειψα είτε καλύπτονται σε άλλα προπτυχιακά μαθήματα του προγράμματος σπουδών του Μαθηματικού Τμήματος είτε παρουσιάσθηκαν από τους φοιτητές σε μορφή σεμιναρίων στο τέλος του εξαμήνου).

Στο Κεφάλαιο 2 (Αριθμητική Λύση μη Γραμμικών Συστημάτων) μετά από μία εισαγωγή στον διαφορικό λογισμό πολλών μεταβλητών και στά τοπικά θεωρήματα σύγκλισης, αποδεικνύεται το θεώρημα του Kantorovich γιά την σύγκλιση της μεθόδου του Νεύτωνα. Ακολούθησα το κλασικό βιβλίο [2.3] των Ortega και Rheinboldt καθώς και το βιβλίο [0.5] του Ortega. Δεν υπήρχε δυστυχώς χρόνος σ' αυτό το μάθημα γιά να καλυφθούν σπουδαιότατα θέματα όπως η θεωρία των μεθόδων "του τύπου του Νεύτωνα" ή το μή γραμμικό πρόβλημα ελαχίστων τετραγώνων καθώς και γενικές μέθοδοι βελτιστοποίησης.

Το Κεφάλαιο 3 είναι αφιερωμένο στην αριθμητική λύση του προβλήματος αρχικών τιμών γιά συστήματα συνήθων διαφορικών εξισώσεων πρώτης τάξης με μεθόδους Runge-Kutta και πολυβηματικές μεθόδους. Γιά τις μεθόδους Runge-Kutta, με βάση την διατριβή του Crouzeix (βλ.παρ. 3.2), αποδεικνύεται ένα αποτέλεσμα των Butcher-Crouzeix που δίνει ικανές συνθήκες γιά σύγκλιση με ορισμένη τάξη ακρίβειας στην περίπτωση μη ακάμπτων συστημάτων. Στις πολυβηματικές μεθόδους, γιά το βασικό αποτέλεσμα του Dahlquist, ότι δηλ. η σύγκλιση είναι ισοδύναμη με συνέπεια και ευστάθεια, ακολούθησα το κλασικό βιβλίο [3.4] του Henrici. Στο τέλος του κεφαλαίου γίνεται μιά εισαγωγή στο πρόβλημα της αριθμητικής λύσης ακάμπτων συστημάτων και σε θέματα απόλυτης ευστάθειας και των γενικεύσεών της σε μή γραμμικά συστήματα. Δέν υπήρχε καιρός γιά να εξετασθούν θέματα όπως συνοριακά προβλήματα δύο σημείων, ειδικές μέθοδοι γιά διαφορικές εξισώσεις ανωτέρας τάξης κ.ά.

Από την τεράστια και σημαντική περιοχή της Θεωρίας Προσέγγισης περιορίστηκα, λόγω ελλείψεως χρόνου, σε ένα μονό πρόβλημα δηλ., στην προσέγγιση συναρτήσεων μιάς μεταβλητής με παρεμβολή με πολυώνυμα Lagrange και με τμηματικά πολυωνυμικές συναρτήσεις όπως τμηματικά γραμμικές συναρτήσεις, τμηματικά κυβικές συναρτήσεις Hermite και κυβικές splines. Ακολούθησα, κατά κύριο λόγο, τα βιβλία των De Boor [4.2] και Schultz [4.10]. Από τα πολλά και σημαντικά θέματα της θεωρίας προσέγγισης που δεν έγινε δυνατόν να μελετηθούν, η αριθμητική ολοκλήρωση και η ομοιόμορφη προσέγγιση καλύπτονται σε προπτυχιακό μάθημα του Μαθηματικού Τμήματος.

Κατά το γράψιμο των σημειώσεων αυτών αντιμετώπισα το γνωστό πρόβλημα της μεταγλώττισης στα ελληνικά της ξένης ορολογίας. Ζητώ την επιείκεια του αναγνώστη αν οι μεταφράσεις μου ορισμένων όρων δεν ταιριάζουν στο γλωσσικό του αισθητήριο ή αν από άγνοια δεν υιοθέτησα ήδη καθιερωμένη στά ελληνικά ορολογία.

Θα ήθελα να ευχαριστήσω τους φοιτητές που παρακολούθησαν το μάθημα δίνοντάς μου έτσι την ευκαιρία να το διδάξω και να γράψω αυτές τις σημειώσεις. Ιδιαίτερα ευχαριστώ τον κ. Γεώργιο Ζουράρη που εντόπισε και διόρθωσε πολλά ουσιαστικά και τυπογραφικά λάθη στο κείμενο. Γιά ό,τι σφάλματα απομένουν είμαι φυσικά ο ίδιος υπεύθυνος. Εκφράζω επίσης τις ευχαριστίες μου στο συνάδελφο κ. Γ. Ακρίβη με τον οποίο συζητούσα διαρκώς πολλά θέματα σχετικά με το μάθημα και τις σημειώσεις.

Ιδιαίτερα επίσης ευχαριστώ την δ. Μαρία Σταυρακάκη που με ακρίβεια, ταχύτητα και επιμέλεια έγραψε το μεγαλύτερο μέρος αυτών των σημειώσεων στον μικροϋπολογιστή της. Μέρος του Κεφαλαίου 4 γράφτηκε από την κ. Λιάνα Ζαχαριουδάκη την οποία επίσης ευχαριστώ. Τέλος θα ήθελα να ευχαριστήσω το Ινστιτούτο Υπολογιστικών Μαθηματικών του Ερευνητικού Κέντρου Κρήτης γιά οικονομική υποστήριξη στην δακτυλογράφηση, φωτοτύπηση και έκδοση των σημειώσεων στην παρούσα τους μορφή.

Ηράκλειο, Ιούνιος 1987 Β.Α. Δουγαλής

ΣΥΝΟΠΤΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

0. Γενικά βιβλία

0.1 R. Bulirsch and J. Stoer, "An introduction to numerical analysis", Springer-Verlag, Berlin-Heidelberg-New York 1980.

0.2 S.D. Conte and C. de Boor, "Elementary numerical analysis: an algorithmic approach", 3^{d} ed, McGraw Hill, New York 1980.

0.3 G.E. Forsythe, M.A. Malcolm and C.B.Moler, "Computer methods for mathematical computation", Prentice-Hall, Englewood Cliffs N.J. 1977.

0.4 E. Isaacson and H.B. Keller, "Analysis of numerical methods", Wiley, New York 1966.

0.5 J.M. Ortega, "Numerical analysis: a second course", Academic Press, New York 1972.

0.6 B. Wendroff, "Theoretical numerical analysis", Academic Press, New York 1966.

0.7 Ν. Αποστολάτος, "Αριθμητική Ανάλυσις Ι,ΙΙ", Αθήνα 1971.

0.8 Α. Μπακόπουλος, "Αριθμητική Ανάλυση", Αθήνα 1981.

0.9 Α. Χατζηδήμος, "Αριθμητική Ανάλυση Ι,ΙΙ", Ιωάννινα 1978,1979.

0.10 Η. Χούστης, "Αριθμητικές μέθοδοι, προγραμματισμός και ανάλυση, μέρος 1", Θεσσαλονίκη 1985.

1. Αριθμητική Γραμμική Άλγεβρα

1.1 D.K. Faddeev and V.N. Faddeeva, "Computational methods of linear algebra", Freeman, San Francisco 1963.

1.2 G.E Forsythe and C.B. Moler, "Computer solution of linear algebraic systems", Prentice-Hall, Englewood Cliffs N.J. 1967.

1.3 A. George and J.W. Liu, "Computer solution of large, sparse positive definite systems", Prentice-Hall, Englewood Cliffs N.J. 1981.

1.4 G.H. Golub and C.F. Van Loan, "Matrix computations", Johns Hopkins U. Press, Baltimore 1983.

1.5 C.L. Lawson and R.J. Hanson, "Solving least squares problems", Prentice-Hall, Englewood Cliffs N.J. 1974.

1.6 B.N. Parlett, "The symmetric eigenvalue problem", Prentice-Hall, Englewood Cliffs N.J. 1980.

1.7 G.W Stewart, "Introduction to matrix computations", Academic Press, New York 1973.

1.8 R.S. Varga, "Matrix iterative analysis", Prentice-Hall, Englewood Cliffs N.J. 1962.

1.9 J.H. Wilkinson, "Rounding errors in algebraic processes", Prentice-Hall, Englewood Cliffs N.J. 1963.

1.10 J.H. Wilkinson, "The algebraic eigenvalue problem", Clarendon Press, Oxford 1965.

1.11 D.M. Young, "Iterative solution of large linear systems", Academic Press, New York 1971.

2. Αριθμητική λύση μη γραμμικών συστημάτων

2.1 J.E. Dennis and R.B. Schnabel, "Numerical methods for unconstrained optimization and nonlinear equations", Prentice-Hall, Englewood Cliffs N.J. 1983.

2.2 P.E. Gill, W. Murray and M.H. Wright, "Practical optimization", Academic Press, London 1981.

2.3 J.M. Ortega and W.C. Rheinboldt, "Iterative solution of nonlinear equations in several variables", Academic Press, New York 1970.

2.4 W.C. Rheinboldt, "Methods for solving systems of nonlinear equations", SIAM, Philadelphia 1974.

3. Αριθμητική λύση συνήθων διαφορικών εξισώσεων

3.1 K. Dekker and J.G. Verwer, "Stability of Runge-Kutta methods for stiff nonlinear differential equations", North-Holland, Amsterdam 1984.

3.2 C.W. Gear, "Numerical initial value problems in ordinary differential equations", Prentice-Hall, Englewood Cliffs N.J. 1971.

3.3 R.D. Grigorieff, "Numerik gewohnlicher Differentialgleichungen", Bd.I,II, B.G. Teubner, Stuttgart, Bd.I: 1972, Bd.II: 1977.

3.4 P.Henrici, "Discrete variable methods in ordinary differential equations", Wiley, New York 1962.

3.5 H.B. Keller, "Numerical methods for two-point boundary value problems", Blaisdell, Waltham Mass. 1968.

3.6 J.D. Lambert, "Computational methods in ordinary differential equations", Wiley, London 1973.

3.7 L.F. Shampine and M.K. Gordon, "Computer solution of ordinary differential equations", Freeman, San Francisco 1975.

3.8 M. Crouzeix et A.L. Mignot, "Analyse numerique des equations differentielles", Masson, Paris 1984.

4. Παρεμβολή, προσέγγιση και αριθμητική ολοκλήρωση

4.1 N.I. Achieser, "Theory of approximation", English Translation, Ungar, New York 1956.

4.2 C. de Boor, "A practical guide to splines", Springer-Verlag, New York 1978.

4.3 E.W. Cheney, "Introduction to approximation theory", McGraw-Hill, New York 1966.

4.4 P.J. Davis, "Interpolation and approximation", Blaisdell, Waltham Mass. 1963.

4.5 P.J. Davis and P. Rabinowitz, "Methods of numerical integration", Academic Press, New York 1975.

4.6 I.P. Natanson, "Constructive function theory", vols 1-3, (Transl. from Russian), Ungar, New York 196 4-5.

4.7 M.J.D. Powell, "Approximation theory and methods", Cambridge U.P., Cambridge 1981.

4.8 J.R. Rice, "The approximation of functions", (2 vols), Addison-Wesley, Reading Mass., vol.1 1964, vol.2 1969.

4.9 T.J. Rivlin, "An introduction to the approximation of functions", Blaisdell, Waltham Mass. 1969.

4.10 M.H. Schultz, "Spline analysis", Prentice-Hall, Englewood Cliffs N.J. 1973.

5. Σημειώσεις μαθημάτων Πανεπιστημίου Κρήτης

5.1 Γ.Δ. Ακρίβης, "Θεωρία προσεγγίσεως και αριθμητική ολοκλήρωση", χειρόγραφες σημειώσεις, Χειμ. Εξ. 1985-6.

5.2 Γ.Δ. Ακρίβης, "Εισαγωγή στην Αριθμητική Ανάλυση", δακτυλογραφημένες σημειώσεις, Ηράκλειο 1986.

5.3 Β.Α. Δουγαλής, "Διδακτικές σημειώσεις γιά το μάθημα Αριθμητική Ανάλυση II", χειρόγραφες σημειώσεις, Χειμ. Εξ. 1983-4.

5.4 Β.Α. Δουγαλής, "Αριθμητική λύση γραμμικών συστημάτων στον υπολογιστή", δακτυλογραφημένες σημειώσεις, Ηράκλειο 1986.

5.5 Β.Α. Δουγαλής, "Διδακτικές σημειώσεις γιά το Μάθημα Αριθμητική Ανάλυση I", χειρόγραφες σημειώσεις, Εαρ. Εξ. 1983-4.

5.6 Β.Α. Δουγαλής, Διδακτικές σημειώσεις γιά το μεταπτυχιακό μάθημα 351 "Αριθμητική Λύση Μερικών Διαφορικών Εξισώσεων" (Μέθοδοι Galerkin/πεπερασμένων στοιχείων), χειρόγραφες σημειώσεις, στά Αγγλικά, Εαρ. εξ. 1984-5.

1. ΑΡΙΘΜΗΤΙΚΗ ΓΡΑΜΜΙΚΗ ΑΛΓΕΒΡΑ

1.1 ΑΠΑΛΟΙΦΗ GAUSS

'Εστω $A=(a_{ij})$, $1\le i,j\le n$, πραγματικός nxn πίνακας (θα συμβολίζουμε $A \in R^{n\times n}$); $b=(b_1,\dots,b_n)^T \in R^n$. Υποθέτουμε ότι ο Α είναι αντιστρέψιμος' έστω $x=(x_1,\dots,x_n)^T$ η λύση του γραμμικού συστήματος

(1) $Ax=b$.

θα αναλύσουμε μία βασική μέθοδο γιά την επίλυση του (1), την απαλοιφή Gauss. 'Εστω $A^{(1)}=A$, $b^{(1)}=b$. Στο πρώτο βήμα ας υποθέσουμε ότι $a_{11}^{(1)}\neq 0$. (Αλλοιώς, με εναλλαγή δύο γραμμών μπορούμε να φέρουμε ένα μη μηδενικό στοιχείο - φερ' ειπείν εκείνο με τον μικρότερο δείκτη γραμμής - της πρώτης στήλης στην θέση (1,1)). Το στοιχείο $a_{11}^{(1)}\neq 0$ λέγεται <u>οδηγός</u> του πρώτου βήματος. Ορίζοντας τώρα τους <u>πολλαπλασιαστές</u>

$$m_{i1} = a_{i1}^{(1)} / a_{11}^{(1)}, \quad 2\le i\le n,$$

πολλαπλασιάζοντας την πρώτη εξίσωση του (1) επί m_{i1} και αφαιρώντας από την i-στή, παίρνουμε το ισοδύναμο σύστημα

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

όπου

$$a_{ij}^{(2)} = \begin{cases} a_{1j}^{(1)} & \text{αν } i=1,\ 1\le j\le n \\ 0 & \text{αν } j=1,\ 2\le i\le n \\ a_{ij}^{(1)}-m_{i1}\,a_{1j}^{(1)} & \text{αν } 2\le i,\ j\le n \end{cases}$$

$$b_i^{(2)} = \begin{cases} b_1^{(1)} & \text{αν } i=1 \\ b_i^{(1)} - m_{i1}b_1^{(1)} & \text{αν } 2\le i\le n. \end{cases}$$

Προχωρούμε ανάλογα μετατρέποντας σε μηδενικά τα στοιχεία της δεύτερης στήλης του $A^{(2)}$ κάτω από το διαγώνιο στοιχείο. Στο k-στό βήμα της μεθόδου έχουμε το ισοδύναμο πρός το (1) σύστημα

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

όπου τα $A^{(k)}$, $b^{(k)}$ είναι της μορφής

$$A^{(k)} = \begin{bmatrix} a_{11}^{(k)} & \cdots & & a_{1n}^{(k)} \\ & \ddots & & \vdots \\ & & a_{kk}^{(k)} \cdots & a_{kn}^{(k)} \\ & O & \vdots & \vdots \\ & & a_{nk}^{(k)} \cdots & a_{nn}^{(k)} \end{bmatrix}, \quad b^{(k)} = \begin{bmatrix} b_1^{(k)} \\ \vdots \\ b_k^{(k)} \\ \vdots \\ b_n^{(k)} \end{bmatrix}$$

'Εστω τώρα ότι ο <u>οδηγός</u> $a_{kk}^{(k)}$ είναι διάφορος του μηδενός. (Αν είναι μηδέν, με εναλλαγή της k-στής γραμμής του $A^{(k)}$ με κάποια i-στή γραμμή (i>k) μπορούμε να φέρουμε ένα μη μηδενικό -π.χ. εκείνο με το μικρότερο δυνατό i - στοιχείο στην θέση του οδηγού). Ορίζουμε τους πολλαπλασιαστές:

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

πολλαπλασιάζουμε την k-στή γραμμή επί m_{ik} και αφαιρούμε από την i-στή. Προκύπτει έτσι το σύστημα

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

όπου

$$a_{ij}^{(k+1)} = \begin{cases} a_{ij}^{(k)} & \text{αν } i\leq k \\ 0 & \text{αν } i>k \text{ και } j\leq k \\ a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)} & \text{αν } k+1\leq i,j\leq n, \end{cases}$$

$$b_i^{(k+1)} = \begin{cases} b_i^{(k)} & \text{αν } i\leq k \\ b_i^{(k)} - m_{ik}b_k^{(k)} & \text{αν } k+1\leq i\leq n. \end{cases}$$

Στό τέλος αυτής της διαδικασίας μετά από n-1 βήματα του αλγορίθμου, προκύπτει το σύστημα

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

ισοδύναμο με το (1)· ο πίνακας $A^{(n)}$ είναι ο άνω τριγωνικός αντίστρέψιμος πίνακας με στοιχεία $a_{ij}^{(n)}$, όπου βέβαια

$$a_{ij}^{(n)} = \begin{cases} 0 & i>j \\ a_{ij}^{(i)} & i\leq j \end{cases} .$$

Γιά το διάνυσμα $b^{(n)}$ έχουμε $b_i^{(n)} = b_i^{(i)}$, $1\leq i\leq n$. Με τον υπολογισμό των

$A^{(n)},b^{(n)}$, ολοκληρώνεται η πρώτη φάση της απαλοιφής, η <u>τριγωνοποίηση</u> του (1). Κατά την δεύτερη φάση, την <u>οπισθοδρόμηση</u>, υπολογίζουμε τους αγνώστους με την σειρά $x_n, x_{n-1},\ldots,x_1$ λύνοντας το άνω τριγωνικό σύστημα (2) με τις αναδρομικές σχέσεις:

$$\begin{cases} x_n = b_n^{(n)} / a_{nn}^{(n)} \\ x_i = (b_i^{(n)} - \sum_{j=i+1}^{n} a_{ij}^{(n)} x_j)/a_{ii}^{(n)}, \quad i=n-1,n-2,\ldots,1. \end{cases}$$

(Ας σημειωθεί ότι αν ο πίνακας A δεν είναι αντιστρέψιμος, τότε γιά κάποιο k, όλα τα στοιχεία $a_{ik}^{(k)}$, $k\le i\le n$ θα είναι μηδέν. Παραλείπουμε τότε το k-στό βήμα της απαλοιφής και προχωρούμε στο k+1-στό. Δηλ. η τριγωνοποίηση ενός μη αντιστρέψιμου πίνακα είναι πάντα δυνατή· βέβαια στου $A^{(n)}$ τουλάχιστον ένα διαγώνιο στοιχείο $a_{ii}^{(i)}$ θα είναι μηδέν. Το αν το <u>σύστημα</u> είναι συμβιβαστό ή αδύνατο εξαρτάται φυσικά και από το b).

Η φάση της τριγωνοποίησης του A κατά την απαλοιφή είναι δυνατόν να ερμηνευθεί σε γλώσσα πινάκων με την λεγόμενη "ανάλυση LU", πολύ σημαντική γιά την θεωρία και τις εφαρμογές:

ΘΕΩΡΗΜΑ 1. Γιά κάθε $A\in\mathbb{R}^{n\times n}$ υπάρχει nxn πίνακας μεταθέσεως P, τέτοιος ώστε

$$(3) \qquad PA = LU,$$

όπου ο nxn πίνακας L είναι κάτω τριγωνικός με μονάδες στην διαγώνιό του και ο πίνακας U είναι άνω τριγωνικός.

<u>Απόδειξη</u>: Χρησιμοποιούμε του συμβολισμό και την διαδικασία της απαλοιφής. Υποθέτουμε πρώτα ότι κατά την απαλοιφή δεν έγιναν καθόλου εναλλαγές γραμμών, δηλ. ότι γιά κάθε k στον $A^{(k)}$ έχουμε είτε $a_{kk}^{(k)} \neq 0$ είτε $a_{ik}^{(k)} = 0$, $k \leq i \leq n$. Θεωρούμε τον πίνακα

$$M_1 = \begin{pmatrix} 1 & & & \\ -m_{21} & 1 & & 0 \\ \vdots & 0 & \ddots & \\ -m_{n1} & & & 1 \end{pmatrix}.$$

Από τον ορισμό του $A^{(2)}$ προκύπτει (με $A^{(1)}=A$) ότι

$$A^{(2)} = M_1 A^{(1)}$$

(Αν $a_{i1}^{(1)} = 0$, $1 \leq i \leq n$, θέτουμε $M_1 = I$). Γενικά, στο k-στό βήμα έχουμε

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

όπου ο nxn πίνακας M_k έχει στοιχεία

$$(M_k)_{ij} = \begin{cases} 1 & \text{αν } i=j \\ -m_{ik} & \text{αν } k+1 \leq i \leq n,\ j=k \\ 0 & \text{αλλοιώς} \end{cases},$$

αν $a_{kk}^{(k)} \neq 0$, ή $M_k = I$ αν $a_{ik}^{(k)} = 0$, $k \leq i \leq n$. Συμπεραίνουμε ότι

$$A^{(n)} = M_{n-1} \dots M_1 A,$$

δηλ. ότι

$$(4) \qquad A = LU,$$

όπου οι πίνακες $U = A^{(n)}$ και $L = M_1^{-1} M_2^{-1} \cdots M_{n-1}^{-1}$ έχουν τις επιθυμητές ιδιότητες. Μάλιστα, αν δεν παραλείφθηκε κανένα βήμα της απαλοιφής (δηλ. αν ο πίνακας A είναι αντιστρέψιμος) έχουμε

$$(5) \qquad L_{ij} = \begin{cases} 1 & \text{αν } i=j \\ m'_{ij} & \text{αν } i>j \\ 0 & \text{αν } i<j \, . \end{cases}$$

Αποδείξαμε δηλ. το θεώρημα στην περίπτωση που δεν γίνονται εναλλαγές γραμμών, οπότε P=I. Αν τώρα κάνουμε εναλλαγές γραμμών γιά να βρούμε οδηγούς $a_{kk}^{(k)} \neq 0$, παρατηρούμε ότι η γραμμή που κατά το k-στό βήμα εναλλάσεται με την προϋπάρχουσα k-στή γραμμή του $A^{(k)}$ δεν αλλάζει πιά θέση αλλά και τα στοιχεία της δεν αλλοιώνονται πλέον στην συνέχεια της απαλοιφής. 'Αρα υπάρχει μία <u>μετάθεση</u> των γραμμών του πίνακα A, που να γινόταν εκ των προτέρων θα επέτρεπε να πραγματοποιηθεί η τριγωνοποίηση χωρίς εναλλαγές γραμμών. Συνεπώς, αν A' είναι ο πίνακας που προκύπτει από τον A μετά από αυτήν την κατάλληλη μετάθεση γραμμών, έχουμε δείξει ήδη ότι

$$A' = LU$$

Ο A' προκύπτει από τον A με πολλαπλασιασμό από τα αριστερά με ένα κατάλληλο <u>πίνακα μεταθέσεως</u> P. (Ο πίνακας μεταθέσεως $P \in \mathbb{R}^{n \times n}$ που αντιστοιχεί στην μετάθεση $i \mapsto \pi(i)$, $1 \le i \le n$, των πρώτων n φυσικών προκύπτει από τον n×n μοναδιαίο πίνακα I αν μεταθέσουμε τις γραμμές του κατά την μετάθεση π. Εύκολα βλέπουμε ότι το γινόμενο PA είναι ο πίνακας που προκύπτει από την μετάθεση των γραμμών του A κατά την μετάθεση π). ∎

'Εστω τώρα ότι ο A είναι αντιστρέψιμος και ότι υπολογίσθηκαν οι πίνακες P,L,U με την διαδικασία της απαλοιφής. Γιά να λύσουμε το σύστημα (1) πολλαπλασιάζουμε και τα δύο μέλη με P οπότε προκύπτει το ισοδύναμο σύστημα

(6) $LUx = Pb$,

που λύνεται σε δύο βήματα: πρώτα υπολογίζουμε το ενδιάμεσο διάνυσμα y ως λύση του κάτω τριγωνικού συστήματος

(7) $Ly = Pb$,

με τον προφανή αλγόριθμο που υπολογίζει πρώτα το y_1 και γενικά το y_k συναρτήσει των y_i, $i<k$. Κατόπιν υπολογίζουμε το x λύνοντας το άνω τριγωνικό σύστημα

(8) $Ux = y$

με τον αλγόριθμο της οπισθοδρόμησης.

<u>Παρατηρήσεις</u>

1. Στην πράξη, η λύση του συστήματος (1) με απαλοιφή Gauss γίνεται σε δύο φάσεις, όπως περιγράψαμε πιό πάνω:

(α) Στην φάση της <u>ανάλυσης PA=LU</u>, κατά την οποία εργαζόμαστε μόνο με τον πίνακα A: υπολογίζουμε τα (ενδιαφέροντα) στοιχεία των πινάκων L και U με την διαδικασία της απαλοιφής και καταγράφουμε τις πληροφορίες εναλλαγής γραμμών που πιθανόν θα γίνουν σε ορισμένα βήματα, δηλ. την δράση του P. Η φάση αυτή απαιτεί $n^3/3+O(n)$ πράξεις (πράξη=πολλαπλασιασμός ή διαίρεση). Οι πολλαπλασιαστές m_{ij} (δηλ. τα στοιχεία του L) καταχωρούνται κατά στήλες στις θέσεις μνήμης όπου ήταν αποθηκευμένα τα στοιχεία a_{ij}, $i>j$ του A πού μετατρέπονται σε μηδενικά· τα στοιχεία u_{ij} του $U(=A^{(n)})$ γράφονται κατά γραμμές πάνω στις θέσεις των a_{ij}, $i\leq j$. Οι πληροφορίες εναλλαγής γραμμών καταχωρούνται συνήθως σ' ένα n-διάνυσμα ακεραίων του οποίου η k-στή συνιστώσα είναι ο δείκτης της γραμμής που γίνεται οδηγός στο βήμα k.

(β) Στην φάση του υπολογισμού της λύσης x κατά την οποία, χρησιμοποιώντας τα προϊόντα της πρώτης φάσης, υπολογίζουμε το Pb και κατόπιν την λύση x λύνοντας διαδοχικά τα τριγωνικά συστήματα (7) και (8). Η φάση αυτή είναι φθηνότερη της πρώτης: απαιτεί $n^2+O(n)$ πράξεις και καθόλου επιπλέον μνήμη.

Κατ' αυτόν τον τρόπο αν θέλουμε να λύσουμε πολλά συστήματα με τον ίδιο πίνακα A και διαφορετικά δεύτερα μέλη b,b',b",..., κάνουμε την ανάλυση LU του PA μία φορά μόνο και υπολογίζουμε τις λύσεις x,x',x",...,λύνοντας γιά κάθε δεύτερο μέλος το ζευγάρι των τριγωνικών συστημάτων (7) και (8). Μιά απλή εφαρμογή είναι ο υπολογισμός του αντιστρόφου A^{-1} (αν υπάρχει) του A του οποίου οι στήλες $x^{(i)}$, $1\leq i\leq n$, είναι προφανώς λύσεις των συστημάτων $Ax^{(i)}=e^{(i)}$, όπου $e^{(i)}_j=\delta_{ij}$, $1\leq i,j\leq n$. Ο συνολικός αριθμός πράξεων γιά τον υπολογισμό του αντιστρόφου είναι δηλ. $n^3/3+O(n)+n(n^2+O(n))=4n^3/3+O(n^2)$. (Αν μάλιστα λάβουμε υπ' όψιν μας και τα μηδενικά των $e^{(i)}$ ο αριθμός των πράξεων είναι $n^3+O(n^2)$).

Γιά περισσότερες λεπτομέρειες πάνω στούς αλγόριθμους της απαλοιφής, βλ. τις σημειώσεις [5.4, παρ. 1, 2.1, 2.3]*

2. Υπάρχουν κατηγορίες πινάκων A γιά τους οποίους μπορούμε να αποδείξουμε ότι η ανάλυση LU μπορεί να γίνει χωρίς εναλλαγές γραμμών, δηλ. ότι υπάρχουν πίνακες L,U με τις ιδιότητες της εκφώνησης του Θεωρήματος 1 τέτοιοι ώστε A=LU. Βλ. τις ασκήσεις 5,6 παρακάτω καθώς και την παρ 1.4.

Ασκήσεις 1.1

1. Επαληθεύστε στην απόδειξη του Θεωρήματος 1 ότι αν $L=M_1^{-1}\ldots M_{n-1}^{-1}$, τότε τα στοιχεία του L δίνονται από την (5).

---

* Οι αριθμοί στις βιβλιογραφικές παραπομπές αναφέρονται στην "Συνοπτική Βιβλιογραφία"

2.(α) Υπολογίστε το πλήθος των πολλαπλασιασμών και διαιρέσεων που απαιτούνται γιά την φάση της ανάλυσης $PA=LU$ καθώς και το πλήθος των προσθαφαιρέσεων και του απαιτουμένου χώρου μνήμης.

(β) Υπολογίστε το πλήθος των πράξεων (πράξη=πολλαπλασιασμός ή διαίρεση) που απαιτούνται γιά την επίλυση ενός nxn άνω ή κάτω τριγωνικού συστήματος. Υπολογίστε επίσης το πλήθος των προσθαφαιρέσεων και της απαιτούμενης μνήμης.

(γ) Δείξτε ότι ο υπολογισμός του αντιστρόφου του A (βλ. παρατήρηση 2) απαιτεί $n^3+O(n^2)$ πράξεις και $2n^2+O(n)$ θέσεις μνήμης.

(δ) Στις εφαρμογές <u>σπάνια</u> ενδιαφερόμαστε γιά τον A^{-1} καθεαυτόν. Συνήθως ενδιαφερόμαστε γιά την λύση x ενός συστήματος $Ax=b$ που γράφεται και σαν $x=A^{-1}b$. Γιά τον υπολογισμό της λύσης με απαλοιφή Gauss χρειαζόμαστε όπως είδαμε $n^3/3+O(n^2)$ πράξεις και $n^2+O(n)$ θέσεις μνήμης. Γιά τον υπολογισμό όμως του $A^{-1}b$ χρειαζόμαστε περίπου τον τριπλάσιο αριθμό πράξεων και τον διπλάσιο χώρο μνήμης. Συνεπώς ουδέποτε υπολογίζουμε διανύσματα της μορφής $A^{-1}y$. Θέτουμε $x=A^{-1}y$ και λύνουμε το σύστημα $Ax=y$. Π.χ.: πώς υπολογίζουμε, όσο το δυνατόν φθηνότερα από άποψη πλήθους πράξεων και χώρου μνήμης, τα διανύσματα $A^{-5}y$, $A^{-1}BA^{-1}y$, $ABA^{-1}y$, όπου $A,B\in R^{nxn}$ (A αντιστρέψιμος), $y\in R^n$;

(ε) Δείξτε ότι ο υπολογισμός της ορίζουσας ενός nxn πίνακα απαιτεί $n^3/3+O(n)$ πράξεις και μπορεί να γίνει εύκολα ως παραπροϊόν της ανάλυσης $PA=LU$ (Συγκρίνετε με τον αριθμό των πράξεων που απαιτούνται από τον συνήθη ορισμό της ορίζουσας).

3.(α) Έστω π η μετάθεση $i_k \mapsto k$, $k=1,2,\ldots,n$, των n πρώτων φυσικών και έστω P ο πίνακας μεταθέσεως που αντιστοιχεί στην π. Τότε η k-στή γραμμή του P είναι το διάνυσμα $(e^{(i_k)})^T$ όπου $e_j^{(i)}=\delta_{ij}$, $1\le i,j\le n$.

(β) Δείξτε ότι $PP^T=I$, δηλ. ότι, γιά κάθε πίνακα μεταθέσεως, $P^{-1}=P^T$. Συνεπώς $\det(P)=\pm 1$. Βεβαιωθήτε ότι ο P^T παριστάνει την μετάθεση π^{-1}, δηλ. την $k\mapsto i_k$, $1\le k\le n$.

(γ) Έστω $A\in R^{nxn}$, $b\in R^n$. Βρήτε συναρτήσει των a_{ij}, b_i και της π τα στοιχεία των PA, AP, P^TA, AP^T, PAP^T, Pb, P^Tb.

(δ) 'Εστω P_i οι πίνακες μεταθέσεων που αντιστοιχούν στις μεταθέσεις π_i, $i=1,2$. Δείξτε ότι το γινόμενο P_1P_2 είναι πίνακας μεταθέσεως. Σε ποιά μετάθεση αντιστοιχεί;

4. (α) Υποθέστε ότι ο $A \in R^{n\times n}$ είναι αντιστρέψιμος και ότι η ανάλυση LU του μπορεί να γίνει χωρίς εναλλαγές γραμμών, (δηλ. ότι $P=I$ στην (3)). Τότε το ζευγάρι L,U (με τις γνωστές ιδιότητες) είναι μοναδικό. Μ' άλλα λόγια, έστω ότι A αντιστρέψιμος και υποθέστε ότι

(i) $A=LU$

(ii) L κάτω τριγωνικός με $L_{ii}=1$

(iii) U άνω τριγωνικός.

Δείξτε τότε ότι οι L,U είναι μοναδικοί.

(β) Η μοναδικότητα των L,U (με τις παραπάνω προϋποθέσεις) μας επιτρέπει να τους υπολογίσουμε με κατασκευές διαφορετικές από εκείνη της απόδειξης του Θεωρήματος 1. Π.χ. χρησιμοποιώντας την ισότητα $A=LU$ στοιχείο πρός στοιχείο, δείξτε ότι τα στοιχεία των L,U δίνονται από τους τύπους:

$$L_{ij} = \Big(a_{ij} - \sum_{k=1}^{j-1} L_{ik}U_{kj}\Big)/U_{ii}\ , \quad j<i$$

$$L_{ii} = 1$$

$$U_{ij} = a_{ij} - \sum_{k=1}^{i-1} L_{ik}\, U_{kj}\ , \quad j\geq i,$$

όπου συμβολίζουμε $\sum_{k=p}^{q} =0$ αν $q<p$. Δώστε ένα ακριβή αλγόριθμο γιά του υπολογισμό και αποθήκευση των L,U, γραμμή πρός γραμμή. (Η κατασκευή αυτή των L και U λέγεται μέθοδος του Crout).

(γ) Υποθέστε ότι ισχύουν οι υποθέσεις του ερωτήματος (α) και ότι επιπλέον ο A είναι συμμετρικός. Δείξτε ότι $U=DL^T$ όπου D ο διαγώνιος πίνακας με $D_{ii}=U_{ii}$. Δώστε αλγόριθμο ανάλογο με του του (β) γιά την κατασκευή των D,L.

5. Γιά $A\in R^{n\times n}$ ορίζουμε τις <u>κύριες ορίζουσες</u> δ_i ως εξής:

$$\delta_1 = a_{11},\ \delta_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix},\ \dots,\ \delta_i = \begin{vmatrix} a_{11} & \dots & a_{1i} \\ \vdots & & \\ a_{i1} & \dots & a_{ii} \end{vmatrix}$$

'Εστω ότι $\delta_i\neq 0$, $1\leq i\leq n$. Τότε δείξτε ότι το θεώρημα 1 ισχύει με $P=I$, δηλ. ότι υπάρχουν πίνακες L,U με τις ιδιότητες (i), (ii), (iii) της άσκησης 4(α). Επιπλέον το ζευγάρι L,U είναι μοναδικό. (Υπόδειξη: γιά την ύπαρξη δείξτε με επαγωγή ότι $a^{(k)}_{kk}\neq 0$, $1\leq k\leq n$, χωρίς εναλλαγές γραμμών.

6. 'Εστω ότι ο πίνακας $A\in R^{n\times n}$ είναι αντιστρέψιμος και έχει <u>κυριαρχική διαγώνιο, κατά γραμμές</u>, δηλ. ότι

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

(α) Δείξτε ότι υπάρχει (μοναδικό) ζευγάρι L,U με τις ιδιότητες (i), (ii), (iii) της άσκησης 4(α). (Υπόδειξη: από τις υποθέσεις μας $a^{(1)}_{11}=a_{11}\neq 0$. Θεωρούμε τον $(n-1)x(n-1)$ υποπίνακα $\widehat{A}^{(2)}$ του $A^{(2)}$ που προκύπτει αν αφαιρέσουμε την πρώτη γραμμή και την πρώτη στήλη του $A^{(2)}$. Δείξτε ότι ο $\widetilde{A}^{(2)}$ είναι αντιστρέψιμος και έχει κυριαρχική διαγώνιο κατά γραμμές. Συμπέρασμα: $a^{(2)}_{22}\neq 0$, δηλ. δεν χρειάζεται εναλλαγή γραμμών γιά να υπολογίσουμε τον $A^{(3)}$. Η συνέχεια προφανής, με επαγωγή).

(β) Δείξτε ότι τα στοιχεία του U που προκύπτουν από την ανάλυση LU του A του μέρους (α) ικανοποιούν την ανισότητα $|U_{ij}|\leq 2\max_{i,j}|a_{ij}|$.

(γ) Αν ο A είναι αντιστρέψιμος και έχει <u>κυριαρχική διαγώνιο κατά στήλες</u> (προφανής ο ανάλογος ορισμός), τότε ισχύει πάλι η (α) (με ανάλογη υπόδειξη) και η εκτίμηση $|L_{ij}|\leq 1$.

1.2 ΔΕΙΚΤΗΣ ΚΑΤΑΣΤΑΣΗΣ ΠΙΝΑΚΑ

Αν ο $A \in R^{nxn}$ είναι αντιστρέψιμος ο αλγόριθμος της απαλοιφής Gauss της παρ. 1.1 υπολογίζει την λύση του συστήματος Ax=b υπό την προϋπόθεση βέβαια ότι οι αριθμητικές πράξεις γίνονται ακριβώς. Στην πραγματικότητα όμως αυτό δεν είναι δυνατόν' όπως ξέρουμε οι αριθμητικές πράξεις υπόκεινται σε σφάλματα στρογγύλευσης λόγω της πεπερασμένης ακρίβειας της αριθμητικής κάθε υπολογιστή. Τα σφάλματα στρογγύλευσης συσσωρεύονται και αποβαίνουν συχνά καταστρ~~εφι~~επτικά έτσι ώστε η υπολογιστική λύση $\tilde{x}$ να απέχει πολύ από την θεωρητική λύση x. 'Ενας από τους κύριους σκοπούς μας στη συνέχεια θα είναι να αναλύσουμε την επιρροή αυτών των σφαλμάτων στην απαλοιφή.

Θα αρχίσουμε εξετάζοντας σ' αυτήν την παράγραφο το θεωρητικό πρόβλημα της ευαισθησίας της θεωρητικής λύσης x του γραμμικού συστήματος Ax=b σε διαταραχές (μεταβολές) στα δεδομένα A και b του προβλήματος. Λέμε ότι ένα σύστημα έχει κακή κατάσταση (ill-conditioned) - ~~ή ότι είναι "ασταθές"~~ - αν είναι πολύ ευαίσθητο σε διαταραχές, δηλ. αν "μικρές" διαταραχές των στοιχείων των A,b είναι δυνατόν να επιφέρουν "μεγάλες" μεταβολές στη λύση του. Το πρόβλημα αυτό είναι ανεξάρτητο από την μελέτη οποιουδήποτε αλγορίθμου γιά την αριθμητική λύση του συστήματος. Είναι προφανές όμως ότι και μόνο η (γενικά μη ακριβής) παράσταση των στοιχείων a_{ij} και b_i στου υπολογιστή συνιστά μιά "διαταραχή" στα δεδομένα. Επιπλέον, όπως θα δούμε, η υπολογιστική λύση $\tilde{x}$ μπορεί να θεωρηθεί ως ακριβής λύση κάποιου "παραπλήσιου" πρός το Ax=b συστήματος. Η κατάσταση (ευαισθησία σε διαταραχές) ενός γραμμικού συστήματος - που εξαρτάται από το μέγεθος του δείκτη κατάστασης του πίνακα A - είναι λοιπόν ένας σημαντικός παράγοντας που επηρεάζει την ακρίβεια της υπολογιστικής λύσης $\tilde{x}$. 'Ενας άλλος παράγοντας είναι η ευστάθεια (ή η αστάθεια) του συγκεκριμένου αλγορίθμου που υπολογίζει το $\tilde{x}$, πρόβλημα που θα εξετάσουμε στην επόμενη παράγραφο.

Αρχίζουμε με μιά σύντομη επανάληψη πάνω στις νόρμες (στάθμες) διανυσμάτων και πινάκων' γιά περισσότερες λεπτομέρειες και αποδείξεις βλ. π.χ. [5.4, παρ. 3]. Μιά (διανυσματική) νόρμα στον $\mathbb{C}^n$ είναι μία απεικόνιση $\|\cdot\|: \mathbb{C}^n \mapsto [0,\infty)$ με τις ιδιότητες

(i) $\|x\| \geq 0 \;\; \forall x \in \mathbb{C}^n$, $\;\|x\|=0 \Leftrightarrow x=0$.
(ii) $\|\lambda x\| = |\lambda| \, \|x\| \;\; \forall x \in \mathbb{C}^n$, $\lambda \in \mathbb{C}$.
(iii) $\|x+y\| \leq \|x\|+\|y\| \quad \forall x,y \in \mathbb{C}^n$.

Μερικές χρήσιμες νόρμες στην αριθμητική ανάλυση είναι οι

$\|x\|_\infty = \max_{1\leq i\leq n} |x_i|$ (η λεγόμενη l_∞ ή max(imum) νόρμα)

$\|x\|_1 = \sum_{i=1}^{n} |x_i|$ (l_1 νόρμα)

$\|x\|_2 = (\sum_{i=1}^{n} |x_i|^2)^{1/2}$ (l_2 ή Ευκλείδια νόρμα).

Γιά την l_2 νόρμα έχουμε ότι $\|\cdot\|_2 = (\cdot,\cdot)_2^{1/2}$, όπου $(x,y)_2 = \sum_{i=1}^{n} x_i \bar{y}_i$, $x,y \in \mathbb{C}^n$, είναι το Ευκλείδιο εσωτερικό γινόμενο στον $\mathbb{C}^n$. Ένα θεμελιώδες αποτέλεσμα, απόρροια του ότι ο $\mathbb{C}^n$ έχει πεπερασμένη διάσταση, είναι ότι οποιεσδήποτε δύο νόρμες στον $\mathbb{C}^n$ είναι <u>ισοδύναμες</u> (ή <u>συγκρίσιμες</u>). Δηλ. ότι, δεδομένων δύο νορμών $\|\cdot\|_\alpha$ και $\|\cdot\|_\beta$ στον $\mathbb{C}^n$, υπάρχουν θετικές σταθερές c_1, c_2 (που εξαρτώνται γενικά από το n), τέτοιες ώστε να ισχύει

(1) $c_1\|x\|_\alpha \leq \|x\|_\beta \leq c_2\|x\|_\alpha, \quad \forall x \in \mathbb{C}^n$.

Λέμε ότι η ακολουθία $\{x^k\}_{k=1}^{\infty}$ διανυσμάτων του $\mathbb{C}^n$ συγκλίνει στο διάνυσμα $x \in \mathbb{C}^n$ (γράφουμε $x^k \to x$, $k\to\infty$) αν γιά κάποια νόρμα $\|\cdot\|$ του $\mathbb{C}^n$ ισχύει

$$\lim_{k\to\infty} \|x^k - x\| = 0$$

Από την (1) έπεται ότι η σύγκλιση είναι ανεξάρτητη της χρησιμοποιούμενης νόρμας. Επίσης από την ισοδυναμία των νορμών $\|\cdot\|$ και $\|\cdot\|_\infty$ έπεται ότι $x^k \to x$, $k \to \infty \Leftrightarrow \forall i$, $1 \le i \le n$, $x_i^k \to x_i$, $k \to \infty$.

Κάθε νόρμα $\|\cdot\|$ του $\mathbb{C}^n$ <u>παράγει</u> μία αντίστοιχη νόρμα ("φυσική" νόρμα, νόρμα "τελεστού") στον χώρο $\mathbb{C}^{n\times n}$ των nxn μιγαδικών πινάκων: γιά $A \in \mathbb{C}^{n\times n}$ ορίζουμε την νόρμα $\|A\|$ ως

$$(2)\quad \|A\| = \sup_{\substack{x\in\mathbb{C}^n \\ x\neq 0}} \|Ax\|/\|x\| \;\left(= \sup_{\substack{x\in\mathbb{C}^n \\ \|x\|\le 1}} \|Ax\| = \sup_{\substack{x\in\mathbb{C}^n \\ \|x\|=1}} \|Ax\|\right).$$

Εύκολα προκύπτει ότι αν $A, B \in \mathbb{C}^{n\times n}$ τότε

(α) $\|A\| \ge 0$, $\|A\|=0 \Leftrightarrow A=0$,
(β) $\|\lambda A\| = |\lambda|\ \|A\|\quad \forall \lambda\in\mathbb{C}$,
(γ) $\|A+B\| \le \|A\|+\|B\|$,
(δ) $\|AB\| \le \|A\|\ \|B\|$.

Οι νόρμες πινάκων που παράγονται από τις διανυσματικές νόρμες $\|\cdot\|_\infty$, $\|\cdot\|_1$, $\|\cdot\|_2$ δίνονται, αντίστοιχα, συναρτήσει του $A=(a_{ij})\in\mathbb{C}^{n\times n}$ από τους τύπους:

$$\|A\|_\infty = \max_{1\le i\le n} \left(\sum_{j=1}^{n} |a_{ij}|\right),$$

$$\|A\|_1 = \max_{1\le j\le n} \left(\sum_{i=1}^{n} |a_{ij}|\right),$$

$$\|A\|_2 = \max_{1\le i\le n} [\lambda_i(A^*A)]^{1/2},$$

όπου A^* είναι ο ανάστροφος συζυγής του A, δηλ. όπου $(A^*)_{ij}=\bar{a}_{ji}$, και όπου με $\lambda_i(B)$ συμβολίζουμε τις ιδιοτιμές του $B\in\mathbb{C}^{n\times n}$. Αν ο πίνακας A είναι αυτοσυζυγής (ερμιτιανός), δηλ. αν $A=A^*$, τότε ο τελευταίος τύπος απλουστεύεται σε $\|A\|_2 = \max_{1\le i\le n} |\lambda_i(A)|$.

Από τον ορισμό (2) έπεται ότι οποιεσδήποτε δύο νόρμες πινάκων στον $\mathbb{C}^{n\times n}$ είναι ισοδύναμες. Σύγκλιση ακολουθίας πινάκων ορίζουμε εντελώς ανάλογα με την σύγκλιση ακολουθίας διανυσμάτων. Αξιοσημείωτη είναι τέλος η ιδιότητα (Neumann) ότι αν $\|A\|<1$, τότε ο πίνακας $I-A$ είναι αντιστρέψιμος και ικανοποιεί τις ανισότητες

$$(3) \qquad (1+\|A\|)^{-1} \le \|(I-A)^{-1}\| \le (1-\|A\|)^{-1}.$$

Προχωρούμε τώρα στην διερεύνηση της ευαισθησίας της λύσης του συστήματος $Ax=b$, όπου A $n\times n$ αντιστρέψιμος πίνακας σε διαταραχές των A και b. Ας μεταβάλουμε κατ' αρχήν μόνο το δεύτερο μέλος b σε $b+\delta b$. Έστω $x+\delta x$ η λύση του νέου συστήματος

$$(4) \qquad A(x+\delta x) = b+\delta b.$$

Συμπεραίνουμε ότι $\delta x=A^{-1}\delta b$, δηλ. ότι γιά οποιαδήποτε νόρμα $\|\cdot\|$, $\|\delta x\|\le\|A^{-1}\|\,\|\delta b\|$, από την οποία έπεται γιά $b\neq 0$ η

$$(5) \qquad \|\delta x\|/\|x\| \le \|A\|\,\|A^{-1}\|(\|\delta b\|/\|b\|).$$

Μπορούμε εύκολα να κατασκευάσουμε παραδείγματα (δεδομένων των A,b, $\|\cdot\|$) διαταραχών δb γιά τα οποία η (5) ισχύει ως ισότητα (βλ. Ασκ. 3). Συμπεραίνουμε ότι η ποσότητα $\|A\|\,\|A^{-1}\|$ είναι ένας συντελεστής που προσδιορίζει πόσο μεγάλη μπορεί να γίνει η σχετική μεταβολή $\|\delta x\|/\|x\|$ της λύσης του συστήματος όταν η σχετική μεταβολή (ή σχετικό "σφάλμα" αν θεωρήσουμε π.χ. ότι το δb παριστάνει το σφάλμα της προσεγγιστικής παράστασης στοιχείων του b στον υπολογιστή) του δευτέρου μέλους είναι $\|\delta b\|/\|b\|$. Ο αριθμός

$$(6) \qquad \kappa(A) = \|A\|\,\|A^{-1}\|$$

λέγεται <u>δείκτης κατάστασης</u> του A ως πρός την νόρμα $\|\cdot\|$. Αν ο $\kappa(A)$

είναι πολύ μεγάλος τότε λέμε ότι ο πίνακας έχει κακή κατάσταση. (Αν ο $\kappa(A)$ είναι μικρός - πάντα $\kappa(A)\geq 1$ - λέμε ότι ο A έχει "καλή κατάσταση"' σημειώστε ότι η συγκρισιμότητα δύο οποιωνδήποτε νορμών πινάκων οδηγεί σε συγκρισιμότητα των αντιστοίχων δεικτών κατάστασης: Γιά δύο οποιεσδήποτε νόρμες $\|\cdot\|_\alpha$, $\|\cdot\|_\beta$, υπάρχουν σταθερές c_1,c_2 τέτοιες ώστες $c_1\kappa_\alpha(A)\leq\kappa_\beta(A)\leq c_2\kappa_\alpha(A)$ γιά κάθε A αντιστρέψιμο). Αν η κατάσταση ενός πίνακα είναι κακή, είναι δυνατόν μιά μικρή μεταβολή στο b να προκαλέσει μεγάλη μεταβολή στην λύση.

Ανάλογες παρατηρήσεις ισχύουν αν αντί του b μεταβάλουμε τώρα τα στοιχεία του A μόνο, δηλ. θεωρήσουμε το σύστημα

$$(7)\qquad (A+\delta A)(x+\delta x)=b,$$

όπου $x=A^{-1}b$ και όπου υποθέτουμε ότι η διαταραχή δA είναι αρκετά μικρή έτσι ώστε και ο πίνακας $A+\delta A$ να είναι και αυτός αντιστρέψιμος. Υποθέτουμε συγκεκριμένα ότι

$$(8)\qquad \|\delta A\|\,\|A^{-1}\|<1$$

από την οποία έπεται ότι $\|(\delta A)A^{-1}\|<1 \Rightarrow I+(\delta A)A^{-1}$ αντιστρέψιμος $\Rightarrow (I+(\delta A)A^{-1})A=A+\delta A$ αντιστρέψιμος. Η (7) δίνει τώρα ότι $\delta x=-(A+\delta A)^{-1}(\delta A)x=-A^{-1}(I+(\delta A)A^{-1})^{-1}(\delta A)x$. Παίρνοντας νόρμες και χρησιμοποιώντας τις (3),(8) έχουμε $\|\delta x\|\leq\|A^{-1}\|\,\|\delta A\|\,\|x\|/(1-\|\delta A\|\cdot\|A^{-1}\|)$, απ' την οποία γιά $x\neq 0$ προκύπτει (με $\kappa(A)=\|A\|\,\|A^{-1}\|$) η σχέση

$$(9)\qquad \|\delta x\|/\|x\| \leq [\kappa(A)/(1-\|\delta A\|\,\|A^{-1}\|)](\|\delta A\|/\|A\|).$$

Βλέπουμε δηλ. ξανά τον ρόλο του $\kappa(A)$ ως δείκτη επιρροής των σχετικών σφαλμάτων (μεταβολών) $\|\delta A\|/\|A\|$ πάνω στην λύση του $Ax=b$. Στην γενική περίπτωση, όπου και ο A και το b μεταβάλλονται, υποθέτοντας ότι

$$(10)\qquad (A+\delta A)(x+\delta x)=b+\delta b$$

και επίσης ότι ισχύει η (8), μπορούμε να δείξουμε γιά $b\neq 0$ την ανισότητα

(11) $\|\delta x\|/\|x\| \leq [\kappa(A)/(1-\|A^{-1}\|\,\|\delta A\|)]\,[\|\delta A\|/\|A\|+\|\delta b\|/\|b\|]$,

της οποίας οι (5), (9) αποτελούν ειδικές περιπτώσεις.

<u>Παρατηρήσεις</u>

1. Ο δείκτης κατάστασης ενός πίνακα A τείνει στο ∞ αν ο A τείνει να γίνει μη αντιστρέψιμος. Μάλιστα μπορεί να αποδειχθεί (Kahan)

(12) $(\kappa(A))^{-1} = \inf\{\|A-B\|/\|A\|$, B μη αντιστρέψιμος$\}$

δηλ. ότι ο $(\kappa(A))^{-1}$ μετράει την (σχετική ως πρός το μέγεθος του A) <u>απόσταση</u> του A από το σύνολο των μη αντιστρεψίμων πινάκων. (Η απόδειξη του $\leq$ στην (12) είναι εύκολη: Έστω B nxn μη αντιστρέψιμος. Τότε $\exists\ x\neq 0$ τ.ώ. $Bx=0$. Συνεπώς $\|A-B\|\,\|x\| \geq \|(A-B)x\| = \|Ax\| \geq \|x\|/\|A^{-1}\|$. Η απόδειξη του $\geq$ είναι δυσκολώτερη).

2. Το μέγεθος του δείκτη κατάστασης δεν έχει όμως σχέση (γιά $n>2$) με το μέγεθος της ορίζουσας detA, αν ο A είναι αντιστρέψιμος. Π.χ. ο 100x100 διαγώνιος πίνακας D με στοιχεία $d_{ii}=0.1$ έχει $\det D=10^{-100}$ ενώ η κατάστασή του είναι φυσικά η καλύτερη δυνατή ($\kappa(D)=1$ ως πρός οποιαδήποτε νόρμα). Απ' την άλλη μεριά, γιά τον nxn άνω τριγωνικό πίνακα A με στοιχεία $a_{ii}=1$, $1\leq i\leq n$, $a_{ij}=-1$, $i<j$, ισχύει ότι $\det A=1$ αλλά $\kappa_\infty(A)=n2^{n-1}$.

3. Από την σχέση $\|A\|_2=\max_i(\lambda_i(A^*A))^{1/2}$ εύκολα βλέπουμε ότι $\kappa_2(A)=(\mu_{max}/\mu_{min})^{1/2}$, όπου μ_{max} (μ_{min}) είναι η μέγιστη (ελάχιστη) ιδιοτιμή του πίνακα A^*A. Αν ο πίνακας A είναι αυτοσυζυγής, δηλ. αν $A^*=A$, τότε η παραπάνω σχέση απλουστεύεται στην $\kappa_2(A)=|\lambda_{max}/\lambda_{min}|$ όπου λ_{max} (λ_{min}) είναι η μέγιστη (ελάχιστη) ιδιοτιμή του A.

4. Πολύ γνωστοί γιά την κακή κατάστασή τους, ακόμα και γιά μικρό n,

είναι οι λεγόμενοι <u>πίνακες Hilbert</u> H_n, $n=1,2,3,\ldots$ που ορίζονται ως $(H_n)_{ij}=(i+j-1)^{-1}$, $1\leq i,j\leq n$. Είναι πίνακες συμμετρικοί και θετικά ορισμένοι (γιατί;) των οποίων οι δείκτες κατάστασής τους γιά $2\leq n\leq 10$ δίδονται από τον πίνακα

n:	2	3	4	5	6	7	8
$\kappa_2(H_n)$:	$1.9\cdot 10$	$5.2\cdot 10^2$	$1.6\cdot 10^4$	$4.8\cdot 10^5$	$1.5\cdot 10^7$	$4.8\cdot 10^8$	$1.5\cdot 10^{10}$

n:	9	10
$\kappa_2(H_n)$:	$4.9\cdot 10^{11}$	$1.6\cdot 10^{13}$

5. Έστω $x\neq 0$ η ακριβής λύση του συστήματος $Ax=b$ και έστω $\tilde{x}$ προσέγγιση της x. Το <u>υπόλοιπο</u> της $\tilde{x}$ ορίζεται ως $r=A\tilde{x}-b$. Εύκολα βλέπουμε ότι $r=A(\tilde{x}-x)$, δηλ. ότι $\|\tilde{x}-x\|=\|A^{-1}r\|\leq\|A^{-1}\|\,\|r\|\leq\kappa(A)\|r\|\,\|x\|/\|b\|$ Συμπεραίνουμε ότι

$$\|\tilde{x}-x\|/\|x\| \leq \kappa(A)\,\|r\|/\|b\|,$$

δηλ. ότι αν το <u>υπόλοιπο</u> r μιάς προσεγγιστικής λύσης είναι μικρό αυτό δεν σημαίνει αναγκαστικά ότι το (σχετικό) <u>σφάλμα</u> της $\|\tilde{x}-x\|/\|x\|$ θα είναι μικρό αν $\kappa(A)>>1$.

Ασκήσεις 1.2

1. Στη βιβλιογραφία της αριθμητικής γραμμικής άλγεβρας συνηθίζεται να ονομάζεται "νόρμα πινάκων" μιά απεικόνιση $\|\cdot\|:\mathbb{C}^{n\times n}\to[0,\infty)$ που ικανοποιεί τις ιδιότητες (α)-(δ) στην σελίδα 1.2.3 (Ειδική περίπτωση λοιπόν "νόρμας πινάκων" είναι η "νόρμα τελεστών" - "φυσική" νόρμα όπως λέγεται - που παράγεται από μιά διανυσματική νόρμα και που ορίζεται από τον τύπο (2)).

(α) Δείξτε ότι η παράσταση $\|A\|_E = (\sum_{i,j=1}^{n} |a_{ij}|^2)^{1/2}$ - η λεγόμενη νόρμα Frobenius του A - ορίζει μία "νόρμα πινάκων" που δεν είναι "φυσική", δηλ. δεν παράγεται από διανυσματική νόρμα. Δείξτε ότι ισχύει $\|A\|_2 \le \|A\|_E \le n^{1/2}\|A\|_2$ γιά κάθε A.

(β) Δείξτε ότι η παράσταση $\max_{i,j} |a_{ij}|$ δεν είναι νόρμα πινάκων.

2.(α) Αν $U^*U=I$ δείξτε ότι $\|U^*AU\|_2=\|A\|_2 \ \forall A\in C^{n\times n}$.

(β) Δείξτε ότι $\max_i |\lambda_i(A)| \le \|A\|$ γιά κάθε $A\in C^{n\times n}$ και κάθε νόρμα $\|\cdot\|$.

(γ) Βρήτε σταθερές σύγκρισης γιά τα ζευγάρια νορμών από τις $\|A\|_1$, $\|A\|_\infty$, $\|A\|_2$.

3.(α) 'Εστω $A\in R^{n\times n}$ συμμετρικός αντιστρέψιμος πίνακας. Θεωρείστε το πρόβλημα (4). Βρήτε ένα δεύτερο μέλος b και μία διαταραχή δb στον R^n έτσι ώστε να ισχύουν οι ισότητες $\|\delta x\|_2 = \|A^{-1}\|_2 \|\delta b\|_2$ και $\|b\|_2 = \|A\|_2 \|x\|_2$ και κατά συνέπειαν και η (5) σαν ισότητα. (Υπόδειξη: ο A ως συμμετρικός πίνακας έχει n ορθοκανονικά ιδιοδιανύσματα και n πραγματικές ιδιοτιμές. Δηλ. υπάρχει ορθογώνιος nxn πίνακας $Q(Q^TQ=I)$ τέτοιος ώστε $Q^TAQ = D = diag(\lambda_1,\dots,\lambda_n)$ όπου λ_i οι ιδιοτιμές του A τις οποίες διατάσσουμε κατά φθίνουσα απόλυτη τιμή ως $|\lambda_1| \ge |\lambda_2| \dots \ge |\lambda_n| > 0$. Διαλέξτε $b = Q(1,0,0,\dots,0)^T$, $\delta b=Q(0,0,\dots,0,1)^T$).

(β) Θεωρείστε το σύστημα $Ax=b$ όπου

$$A = \begin{pmatrix} 4.1 & 2.8 \\ 9.7 & 6.6 \end{pmatrix} \qquad b = \begin{pmatrix} 4.1 \\ 9.7 \end{pmatrix}$$

Δείξτε ότι αν $\|\cdot\|=\|\cdot\|_1$, $\delta b=(0.01,0)^T$, η (5) ισχύει σαν ισότητα.

4. (Η άσκηση αυτή αναφέρεται σε εφαρμογές της (12) της παρατήρησης 1.)

(α) Χωρίς να βρεθεί ο A^{-1} δείξτε ότι γιά τον πίνακα

$$A = \begin{pmatrix} 1.01 & .99 \\ .99 & 1.01 \end{pmatrix}$$

ισχύει ότι $\kappa_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty \geq 100$. (Υπόδειξη: ο πίνακας $B: b_{ij}=1$ δεν είναι αντιστρέψιμος).

(β) Έστω A αντιστρέψιμος άνω ή κάτω τριγωνικός πίνακας. Δείξτε ότι $\kappa_\infty(A) \geq \|A\|_\infty / \min_i |a_{ii}|$

(γ) Αν ο A αντιστρέψιμος και ο B ικανοποιεί την σχέση $\|A-B\| < 1/\|A^{-1}\|$, τότε ο B είναι αντιστρέψιμος.

5.(α) Να αποδειχθεί η (11).

(β) Αν οι πίνακες A,B είναι αντιστρέψιμοι, τότε
$\|B^{-1}-A^{-1}\|/\|B^{-1}\| \leq \kappa(A) \, \|A-B\|/\|A\|$.

(γ) Να αποδειχθούν οι ισχυρισμοί της παρατήρησης 2.

(δ) Να αποδειχθούν οι ισχυρισμοί της παρατήρησης 3.

6. Έχουμε πάντοτε ότι $\kappa(A)=\|A\| \, \|A^{-1}\| \geq \|AA^{-1}\|=1$. Τι είδους πίνακες έχουν $\kappa(A)=1$; Δείξτε ότι

(α) Αν ο A είναι διαγώνιος με ίσα στοιχεία, τότε $\kappa(A)=1$.

(β) Αν ο A είναι ισομετρία ως πρός $\|\cdot\|$, δηλ. αν $\|Ax\|=\|x\| \;\; \forall x$, τότε $\kappa(A)=1$.

(γ) Αν ο A είναι τέτοιος ώστε $A^*A=I$, τότε $\kappa_2(A)=1$.

7.(α) Αν ο πραγματικός πίνακας $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ είναι αντιστρέψιμος και $\sigma = (a^2+b^2+c^2+d^2)/2|ad-bc|$, δείξτε ότι $\kappa_2(A)=\sigma+(\sigma^2-1)^{1/2}$.

(β) Δείξτε ότι ο πίνακας

$$\begin{pmatrix} 100 & 99 \\ 99 & 98 \end{pmatrix}$$

(καθώς και οι πίνακες που προέρχονται από αυτόν με μεταθέσεις στοιχείων του) έχει τον μεγαλύτερο δείκτη κατάστασης μέσα στο σύνολο των 2x2 αντιστρέψιμων πινάκων που τα στοιχεία τους είναι θετικοί ακέραιοι ≤ 100.

8. Θεωρείστε γιά $\varepsilon\in\mathbb{R}$ την λύση $x(\varepsilon)$ του συστήματος

$$\begin{cases} (A+\varepsilon F)x(\varepsilon) = b+\varepsilon f \ , \ \varepsilon\neq 0 \\ x(0)=x \end{cases}$$

όπου $A,F\in\mathbb{R}^{n\times n}$ δεδομένοι πίνακες (A αντιστρέψιμος) και b,f δεδομένα διανύσματα στον $\mathbb{R}^n$. Δείξτε ότι η $\varepsilon\mapsto x(\varepsilon)$ είναι διαφορίσιμη σε μιά περιοχή του μηδενός. Χρησιμοποιώντας το ανάπτυγμα Taylor της $x(\varepsilon)$ γύρω από το $\varepsilon=0$ δείξτε ότι

$$\|x(\varepsilon)-x\|/\|x\| \leq \varepsilon\kappa(A)[(\|F\|/\|A\|)+(\|f\|/\|b\|)]+O(\varepsilon^2).$$

1.3 ΣΦΑΛΜΑΤΑ ΣΤΡΟΓΓΥΛΕΥΣΗΣ ΣΤΗΝ ΑΠΑΛΟΙΦΗ GAUSS

Προχωρούμε τώρα στην μελέτη του δευτέρου παράγοντα που, όπως αναφέραμε στις εισαγωγικές παρατηρήσεις της προηγουμένης παραγράφου, επηρεάζει την υπολογιστική λύση $\tilde{x}$ του $Ax=b$, $A\in R^{n\times n}$, $b\in R^{n}$. Ο παράγοντας αυτός είναι η ευστάθεια (ή αστάθεια) του αλγορίθμου της απαλοιφής σε (μικρές) διαταραχές που θα προέλθουν βέβαια από τα σφάλματα στρογγύλευσης στις πράξεις. Η ευστάθεια του αλγορίθμου της απαλοιφής είναι ένα θέμα ανεξάρτητο από την κατάσταση του συστήματος. Στο τέλος όμως και οι δύο αυτοί παράγοντες θα συμβάλουν, όπως θα δούμε, στην διαμόρφωση του σφάλματος της $\tilde{x}$.

Κατ' αρχήν πρέπει να παρατηρήσουμε ότι στην πράξη η πρώτη φάση της απαλοιφής, δηλ. η ανάλυση PA=LU, δεν γίνεται όπως στην παρ. 1.1, όπου στο γενικό βήμα k φέρναμε στη θέση του οδηγού απλώς ένα μη μηδενικό στοιχείο από τα $a_{ik}^{(k)}$, $i\geq k$, π.χ. εκείνο με το μικρότερο δείκτη i. Αν ο A είναι αντιστρέψιμος (πράγμα που θα υποθέσουμε από εδώ κι εμπρός) η εύρεση ενός μη μηδενικού οδηγού είναι εξασφαλισμένη σε κάθε βήμα. Είναι προφανές όμως ότι θα πρέπει να περιμένουμε προβλήματα αν κάποιος οδηγός $a_{kk}^{(k)}$ είναι π.χ. πολύ μικρός σε απόλυτη τιμή. Ο λόγος είναι φυσικά η κατά προσέγγιση παράσταση των αριθμών και η πεπερασμένη ακρίβεια με την οποία γίνονται οι πράξεις στον υπολογιστή. Ένας (απόλυτα) μικρός οδηγός $a_{kk}^{(k)}$ π.χ. μπορεί να οδηγήσει σε μεγάλους (σε απόλυτη τιμή) πολλαπλασιαστές στο βήμα k $m_{ik}=a_{ik}^{(k)}/a_{kk}^{(k)}$ και συνεπώς σε μεγάλη απώλεια ακρίβειας κατά τις αφαιρέσεις $a_{ij}^{(k)}-m_{ik}a_{kj}^{(k)}$ στις οποίες είναι δυνατόν - υποθέτοντας ότι $a_{ij}^{(k)}, a_{kj}^{(k)}=O(1)$ - ο "μεγάλος" όρος $-m_{ik}a_{kj}^{(k)}$ να εξαφανίσει τον "μικρό"

$a_{ij}^{(k)}$ μετά από την στρογγύλευση στην ακρίβεια του υπολογιστή. Τέτοιου είδους σφάλματα στρογγύλευσης διαδίδονται στους υπολογισμούς και καταστρέφουν την ακρίβεια της $\tilde{x}$. Για αριθμητικά παραδείγματα τέτοιων φαινομένων βλ. π.χ. [5.4, παρ. 2.2]. Η απλή απαλοιφή λοιπόν, όπως περιγράφτηκε στη παρ. 1.1, δεν είναι ευσταθής αλγόριθμος.

Προσπαθούμε λοιπόν να φέρουμε σε κάθε βήμα στην θέση του οδηγού ~~όσο το~~ δυνατόν μεγαλύτερα σε απόλυτη τιμή στοιχεία. Μιά τέτοια στρατηγική λέγεται <u>οδήγηση</u> (pivoting). Μιά προφανής επιλογή στο k-στό βήμα είναι να βρούμε εκείνο το στοιχείο από τα $a_{ik}^{(k)}$, $i \geq k$, με την μεγαλύτερη απόλυτη τιμή και με εναλλαγή γραμμών να το φέρουμε στην θέση του οδηγού. Αυτή είναι η λεγόμενη <u>μερική οδήγηση</u> (partial pivoting) ή <u>οδήγηση κατά γραμμές</u>. (Η μερική οδήγηση συχνά βελτιώνεται με μιά παραλλαγή της, την <u>μερική οδήγηση με στάθμιση</u>, βλ. [5.4, σελ. 21]). Το κόστος των συγκρίσεων των στοιχείων κατά την μερική οδήγηση είναι $O(n^2)$ και συνεπώς ασυμπτωτικά πολύ μικρότερο από το κόστος της ανάλυσης LU.

Η μερική οδήγηση είναι στην πράξη σχεδόν πάντα ασφαλής. Υπάρχουν όμως παραδείγματα "παθολογικών" συστημάτων όπου δεν βελτιώνει και πολύ την ακρίβεια στις πράξεις γιά μεγάλο n (βλ. παρακάτω). Σε τέτοιες περιπτώσεις καταφεύγουμε στην λεγόμενη <u>ολική οδήγηση</u> (total pivoting) ή <u>οδήγηση κατά γραμμές και στήλες</u> κατά την οποία με εναλλαγές γραμμών <u>και</u> στηλών φέρνουμε στην θέση του οδηγού $a_{kk}^{(k)}$ το στοιχείο εκείνο του υποπίνακα $a_{ij}^{(k)}$, $k \leq i, j \leq n$ με την μέγιστη απόλυτη τιμή. Μπορεί να αποδειχθεί (βλ. παρακάτω) ότι η ολική οδήγηση είναι πάντα ασφαλής, δηλ. ότι ο "συντελεστής μεγέθυνσης" των σφαλμάτων στρογγύλευσης αυξάνεται πολύ αργά με το n (στη θεωρία' πρακτικά δεν αυξάνεται). Πάντως σπανιώτατα χρησιμοποιούμε ολική οδήγηση στην πράξη μιά και το κόστος της διπλασιάζει το κόστος των $n^3/3$ περίπου πράξεων της ανάλυσης LU. Θα αναλύσουμε λοιπόν την ευστάθεια του αλγορίθμου της απαλοιφής Gauss (ανάλυση PA=LU και λύση των δύο τριγωνικών συστημάτων (1.1.7) και (1.1.8)*) με μερική οδήγηση, τις εναλλαγές γραμμών της 'οποίας καταγράφει ο πίνακας P. Αρχίζουμε με μιά σύντομη επανάληψη περί παράστασης πραγματικών αριθμών στον υπολογιστή' γιά περισσότερες λεπτομέρειες βλ. [5.2] ή [5.4, παρ. 2.2].

'Οπως είναι ~~γνωστό~~ σε κάθε υπολογιστή οι ~~πραγματικοί~~ αριθμοί παριστάνονται από ένα πεπερασμένο σύνολο ρητών, τους λεγόμενους <u>αριθμούς της μηχανής</u> (ή αριθμούς κινητής υποδιαστολής, πεπερασμένης ακρίβειας). Οι ρητοί αυτοί εκφράζονται με t ψηφία ("ακρίβεια" της μηχανής) σ' ένα αριθμητικό σύστημα με βάση β. Κάθε αριθμός της μηχανής είναι της μορφής $y=\pm.d_1d_2 \ldots d_t.\beta^e$ (ή $y=0$), όπου οι ακέραιοι d_i, $1\leq i\leq t$ είναι ψηφία στο σύστημα με βάση β, δηλ. $0\leq d_i\leq\beta-1$ με $d_1\neq 0$. Ο εκθέτης e είναι ακέραιος θετικός ή αρνητικός αλλά περιορίζεται σ' ένα πεπερασμένο διάστημα. Είναι φανερό ότι κάθε τέτοιο σύνολο αριθμών είναι πεπερασμένο. Κάθε πραγματικός x που βρίσκεται μέσα στο εύρος των αριθμών της μηχανής προσεγγίζεται είτε με "αποκοπή" είτε με "στρογγύλευση" από ένα κοντινό του αριθμό της μηχανής που συμβολίζουμε με $fl(x)$. Κατά τα γνωστά ισχύει

$$(1) \qquad fl(x)=x(1+\delta), \text{ όπου } |\delta|\leq u= \begin{cases} \beta^{1-t} & \text{γιά αποκοπή} \\ (\beta^{1-t})/2 & \text{γιά στρογγύλευση} \end{cases}$$

Ο (μικρός) αριθμός u λέγεται "μοναδιαίο σφάλμα στρογγύλευσης". Η (1) λέει απλούστατα ότι το σχετικό σφάλμα $|(fl(x)-x)/x|$ είναι μικρότερο ή ίσο του u.

Δεδομένων δύο αριθμών της μηχανής x και y θα συμβολίζουμε με $fl(x@y)$ όπου $@=+,-,\cdot$ ή $/$ το αποτέλεσμα της κατά προσέγγισιν πρόσθεσης, αντίστοιχα αφαίρεσης, πολ/σμού, διαίρεσής τους στην αριθμητική μονάδα της μηχανής με απλή ακρίβεια.

---

* Η αναφορά σε τύπους, θεωρήματα, παρατηρήσεις, ασκήσεις κλπ. (z) μιάς διαφορετικής παραγράφου x.y θα συμβολίζεται με (x.y.z). Η αναφορά σε τύπους, κλπ. (z) της ίδιας παραγράφου θα συμβολίζεται με (z).

Υποθέτουμε (απλουστευτική αλλά αρκετά ρεαλιστική παραδοχή) ότι πρώτα υπολογίζεται ακριβώς ο πραγματικός αριθμός $x@y$ - που βρίσκεται μέσα στο εύρος των αριθμών της μηχανής - και κατόπιν στρογγυλεύεται (ή αποκόπτεται) στον αριθμό μηχανής $fl(x@y)$. Συνεπώς σαν συνέπεια της (1) έχουμε ότι

(2) $fl(x@y) = (x@y)(1+\delta),\ |\delta|\leq u.$

θα μάς είναι χρήσιμη και η εξής εναλλακτική μορφή της (ψευδο)ισότητας (2) - βλ. 'Ασκηση 1:

(2') $fl(x@y) = (x@y)/(1+\delta'),\ |\delta'|\leq u.$

Χρησιμοποιώντας τώρα την (2) ή την (2') μπορούμε να βρούμε ανάλογες εκφράσεις γιά πιό πολύπλοκες πράξεις. Π.χ. <u>ορίζοντας</u>, γιά x,y,z αριθμούς μηχανής, την παράσταση του αθροίσματός τους $fl(x+y+z)$ ως $fl(fl(x+y)+z)$ έχουμε από κατ' επανάληψιν χρήση της (2) ότι $fl(x+y+z)=fl((x+y)(1+\delta_1)+z)=((x+y)(1+\delta_1)+z)(1+\delta_2)$
$= (x+y)(1+\delta_1)(1+\delta_2)+z(1+\delta_2),\ |\delta_i|\leq u.$ Είναι προφανές ότι το $fl(x+y+z)$ είναι διάφορο γενικά του $fl(x+z+y)$ κλπ. (Η προσεγγιστική πρόσθεση δεν είναι προσεταιριστική!) Γιαυτό γράφοντας $fl(\sum_{i=1}^{m} x_i)$ θα εννοούμε ότι πρώτα προσθέτουμε τα x_1 και x_2, στο αποτέλεσμα το x_3 κ.ο.κ, δηλ. ότι υλοποιούμε τον αλγόριθμο

$s \leftarrow x_1$
Γιά $i=2,\ldots,m$
$\quad s \leftarrow s+x_i$

όπου με $\leftarrow$ συμβολίζουμε την εκχώρηση της τιμής της δεξιάς μεταβλητής στη θέση που βρίσκεται αποθηκευμένη η αριστερή. (Υποθέτουμε σιωπηρά ότι όλες οι πράξεις δίνουν αποτελέσματα μετά στο εύρος των αριθμών της μηχανής).

Θα εφαρμόσουμε τώρα τα παραπάνω εν είδει παραδείγματος σ' ένα πρόβλημα που εμφανίζεται επανειλημμένα στην αριθμητική γραμμική άλγεβρα. Θεωρούμε την πράξη $fl(\sum_{i=1}^{n} x_i y_i)$, δηλ. τον υπολογισμό του Ευκλειδίου εσωτερικού γινομένου δύο πραγματικών διανυσμάτων $x, y \in R^n$ με στοιχεία αριθμούς μηχανής. Θα επιχειρήσουμε να βρούμε μιά εύχρηστη έκφραση του $fl(\sum_{i=1}^{n} x_i y_i)$ συναρτήσει των x_i, y_i, n και u. Γιαυτόν τον σκοπό αποδεικνύουμε πρώτα τα εξής βοηθητικά αποτελέσματα:

Λήμμα 1. Έστω $0 \le u \le 1$ και $n \in N$. Τότε

(α) $1-nu \le (1-u)^n$.

(β) Αν $0 < nu \le 0.01$, τότε $(1+u)^n \le 1+1.01nu$.

(γ) Αν $|\delta_i| \le u$, $i=1,2,\ldots,n$ και $0 < nu \le 0.01$, τότε

(3) $\quad 1-nu \le \prod_{i=1}^{n} (1+\delta_i) \le 1+1.01nu.$

<u>Απόδειξη.</u> Το (α) προκύπτει εύκολα είτε με επαγωγή είτε χρησιμοποιώντας το θεώρημα του Taylor γιά την συνάρτηση $f(u)=(1-u)^n$ γύρω από το 0 με δύο όρους. Γιά το (β) χρησιμοποιούμε το γεγονός ότι $1+x \le e^x$ γιά $x \ge 0$ (προφανές) και το ότι $e^x \le 1+1.01x$ γιά $x \in [0,0.01]$, που προκύπτει από την ανισότητα $e^x \le 1+x+x^2$ που είγουρα ισχύει γιά $0 \le x \le 0.01$. Συνεπώς $(1+u)^n \le e^{nu} \le 1+1.01nu$. Τέλος το (γ) είναι συνέπεια των ανισοτήτων (α) και (β) και της $1-u \le 1+\delta_i \le 1+u$. ©

Σημειώστε ότι η (3), γράφεται και στην μορφή

(3') $\prod_{i=1}^{n} (1+\delta_i)=1+1.01n\theta u$ γιά κάποιο $\theta: |\theta| \le 1$.

Προχωρούμε τώρα στην ζητούμενη έκφραση του $fl(\sum_{i=1}^{n} x_i y_i)$. Τονίζουμε ξανά ότι έχει σημασία η σειρά με την οποία γίνονται οι πράξεις στο εσωτερικό γινόμενο. Θα υποθέσουμε ότι ο αριθμός

$s=fl(\sum_{i=1}^{n} x_i y_i)$ υπολογίζεται από τον αλγόριθμο

$$\text{(4)}\qquad \begin{array}{l} s \leftarrow x_1 y_1 \\ \left[\begin{array}{l} \text{Γιά } i=2,\ldots,n \\ \quad s \leftarrow s+(x_i y_i) \end{array}\right. \end{array}$$

Λήμμα 2. Έστω u το μοναδιαίο σφάλμα στρογγύλευσης και x_i, y_i, $1 \leq i \leq n$ αριθμοί μηχανής. Τότε αν $nu \leq 0.01$, έχουμε

$$\text{(5)}\qquad fl(\sum_{i=1}^{n} x_i y_i) = \sum_{i=1}^{n} x_i y_i \,[1+1.01(n+2-i)\theta_i u],$$

όπου $|\theta_i| \leq 1$, $1 \leq i \leq n$.

<u>Απόδειξη</u>: Λόγω του αλγορίθμου (4) έχουμε

$$\text{(6)}\qquad fl(\sum_{i=1}^{n} x_i y_i) = \underbrace{fl(fl(\ldots(fl}_{n-1\ \text{"fl"}}\,(fl(x_1y_1)+fl(x_2y_2))+fl(x_3y_3))+$$

$$+fl(x_4y_4))+\ldots+fl(x_ny_n)) = x_1y_1 \prod_{i=1}^{n}(1+\delta_i^{(1)})+$$

$$x_2y_2 \prod_{i=1}^{n}(1+\delta_i^{(2)})+x_3y_3 \prod_{i=1}^{n-1}(1+\delta_i^{(3)})+$$

$$+\ldots+x_ny_n \prod_{i=1}^{2}(1+\delta_i^{(n)})$$

όπου $|\delta_i^{(j)}| \le u$. Γιά k>1 έχουμε από την (3') και την υπόθεση $nu \le 0.01$ ότι

$$(7) \quad x_k y_k \prod_{i=1}^{n-k+2} (1+\delta_i^{(k)}) = x_k y_k [1+1.01(n+2-k)\theta_k u]$$

γιά κάποιο $\theta_k : |\theta_k| \le 1$. Γιά k=1 έχουμε από την (3') ότι

$$x_1 y_1 \prod_{i=1}^{n} (1+\delta_i^{(1)}) = x_1 y_1 [1+1.01 n\theta_1' u], \ |\theta_1'| \le 1.$$ Θέτουμε $\theta_1 = n\theta_1'/(n+1)$ οπότε $|\theta_1| \le 1$. Γιαυτό το θ_1 λοιπόν ισχύει η (7) και γιά k=1. Οι (6) και (7) δίνουν τώρα την (5). @

Ας αρχίσουμε τώρα την μελέτη της επιρροής των σφαλμάτων στρογγύλευσης στον αλγόριθμο της απαλοιφής με μερική οδήγηση. Χρησιμοποιώντας τις (1), (2) σε κάθε πράξη της απαλοιφής είναι δυνατόν να παρακολουθήσουμε την συσσώρευση των σφαλμάτων στρογγύλευσης. Η λογιστική όμως ενός τέτοιου εγχειρήματος (που λέγεται και "απ' ευθείας" (forward) ανάλυση του σφάλματος) είναι αρκετά πολύπλοκη. Επιπλέον μιά τέτοια στρατηγική συνήθως δίνει πολύ μεγάλα και μη ρεαλιστικά φράγματα σφαλμάτων. Μιά άλλη τεχνική, η λεγόμενη "αντίστροφη" (inverse, backward) ανάλυση του σφάλματος οφείλεται στον άγγλο μαθηματικό J.H. Wilkinson ο οποίος απέδειξε (δεκαετία του '60) ότι η υπολογιστική λύση $\tilde{x}$ που δίνει η απαλοιφή μπορεί να θεωρηθεί ως ακριβής λύση (δηλ. λύση με αριθμητική απεριόριστης ακρίβειας) όχι του συστήματος $Ax=b$ αλλά ενός "παραπλήσιου" συστήματος $(A+\delta A)\tilde{x}=b$. Η ευστάθεια ή μη του αλγορίθμου εκφράζεται ποσοτικά από το αν η σχετική μεταβολή $\|\delta A\|/\|A\|$ είναι μικρή ή όχι. Στη συνέχεια θα εκτιμήσουμε αυτήν την μεταβολή στην $\|\cdot\|_\infty$ νόρμα.

Υποθέτουμε, γιά να απλουστεύσουμε το πρόβλημα ελαφρώς, ότι τα στοιχεία των A,b είναι αριθμοί της μηχανής (βλέπε παρατήρηση 1). Επίσης υποθέτουμε ότι έχουν γίνει ήδη εκ των προτέρων (αν και αυτό φυσικά δεν μπορεί να γίνει στην πράξη) όλες οι εναλλαγές γραμμών που υπαγορεύει η μερική οδήγηση έτσι ώστε γιά κάθε k τα στοιχεία $a_{kk}^{(k)}$

που θα προκύπτουν από τους υπολογισμούς στις θέσεις των οδηγών να είναι μεγαλύτερα ή ίσα απολύτως απ' όλα τα στοιχεία $a_{ik}^{(k)}$ $i>k$. Προφανώς η εναλλαγή γραμμών δεν επηρεάζει την αριθμητική γιά την οποία ενδιαφερόμαστε εδώ. Υποθέτουμε επίσης ότι όλες οι πράξεις οδηγούν σε αριθμούς μέσα στο εύρος των αριθμών της μηχανής έτσι ώστε να μην σταματά ο υπολογισμός λόγω over- ή underflow. (Γιά απλούστευση του συμβολισμού δεν συμβολίζουμε με $\widetilde{L}, \widetilde{U}, \widetilde{m}_{ik}, \widetilde{a}_{ij}^{(k)}$ κλπ. τα μεγέθη που υπολογίζουμε με αριθμητική πεπερασμένης ακρίβειας κατά την απαλοιφή, αλλά χρησιμοποιούμε τα σύμβολα $L, U, m_{ik}, a_{ij}^{(k)}$ κλπ. έχοντας βέβαια υπ' όψιν μας ότι τα μεγέθη αυτά θα είναι εν γένει διαφορετικά από τα αντίστοιχα μεγέθη που εμφανίζονται στην παρ. 1.1 ως αποτελέσματα ακριβών υπολογισμών). Στο k-στό βήμα της απαλοιφής υπολογίζουμε τους πολλαπλασιαστές ως

$$(8) \qquad m_{ik} = fl(a_{ik}^{(k)} / a_{kk}^{(k)}), \; k+1 \le i \le n.$$

Τα στοιχεία $a_{ij}^{(k+1)}$ του πίνακα $A^{(k)}$ υπολογίζονται κατόπιν βάσει των τύπων

$$(9) \qquad a_{ij}^{(k+1)} = \begin{cases} 0 & , \text{αν } j=k, \; k+1 \le i \le n \\ fl(a_{ij}^{(k)} - m_{ik}\, a_{kj}^{(k)}) & , \text{αν } k+1 \le i, j \le n \\ a_{ij}^{(k)} & , \text{αλλοιώς} \end{cases}$$

Στο τέλος αυτής της διαδικασίας ορίζουμε τους nxn πίνακες

$$(10) \qquad U = A^{(n)}$$

$$(11)\qquad L : L_{ij} = \begin{cases} 0\ , & \text{αν } i<j \\ 1\ , & \text{αν } i=j \\ m_{ij}, & \text{αν } i>j\ . \end{cases}$$

Τονίζουμε ότι $A \neq LU$ γιατί τα στοιχεία των L,U έχουν υπολογισθεί με αριθμητική πεπερασμένης ακρίβειας αναδρομικά από τις (8) & (9). Έχουμε όμως

Λήμμα 3. Αν οι πίνακες L,U ορίζονται από τις (10),(11) έχουμε

$$(12)\qquad LU = A+E,$$

όπου

$$(13)\qquad E = \sum_{k=1}^{n-1} E^{(k)}\ ,\ E^{(k)} \in R^{n\times n}.$$

Γιά $1\leq k\leq n-1$, τα στοιχεία $\varepsilon_{ij}^{(k)}$ του $E^{(k)}$ δίδονται από

$$(14)\qquad \varepsilon_{ij}^{(k)} = \begin{cases} a_{ik}^{(k)}\,\delta_{ik} & k+1\leq i\leq n,\ j=k \\ -\,m_{ik}a_{kj}^{(k)}\,\delta'_{ij} - a_{ij}^{(k+1)}\,\delta''_{ij}, & k+1\leq i,j\leq n \\ 0 & \text{αλλοιώς} \end{cases}$$

όπου τα $\delta_{ij},\delta'_{ij},\delta''_{ij}$ εξαρτώνται και από το k και είναι τέτοια ώστε $|\delta_{ij}|,|\delta'_{ij}|,|\delta''_{ij}| \leq u$, όπου u το μοναδιαίο σφάλμα στρογγύλευσης

Απόδειξη: Η (8) δίνει, λόγω της (2), $m_{ik} = (a_{ik}^{(k)} / a_{kk}^{(k)})(1+\delta_{ik})$, $i\geq k+1$, όπου $|\delta_{ik}|\leq u$. Γράφουμε την σχέση αυτή ως

$$(15)\quad a_{kk}^{(k)} m_{ik} - a_{ik}^{(k)} - \varepsilon_{ik}^{(k)} = 0, \quad i\geq k+1,$$

όπου ορίσαμε

$$(16)\quad \varepsilon_{ik}^{(k)} = a_{ik}^{(k)} \delta_{ik}, \quad i\geq k+1.$$

Η (9) δίνει, με χρήση των (2) και (2')

$$a_{ij}^{(k+1)} = fl\,(a_{ij}^{(k)} - m_{ik}\, a_{kj}^{(k)}) = fl(a_{ij}^{(k)} - fl(m_{ik}\, a_{kj}^{(k)})) =$$

$$= (a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)}\,(1+\delta_{ij}'))/(1+\delta_{ij}''), \quad k+1\leq i,j\leq n,$$

όπου $|\delta_{ij}'|, |\delta_{ij}''| \leq u$. Ξαναγράφουμε την σχέση αυτή ως

$$(17)\quad a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik}\, a_{kj}^{(k)} + \varepsilon_{ij}^{(k)}, \quad k+1\leq i,j\leq n,$$

όπου ορίσαμε

$$(18)\quad \varepsilon_{ij}^{(k)} = -m_{ik} a_{kj}^{(k)}\, \delta_{ij}' - a_{ij}^{(k+1)}\, \delta_{ij}'', \quad k+1\leq i,j\leq n.$$

(Σημειώστε ότι τα $\varepsilon_{ij}^{(k)}$ της (16) και της (18) θα ήταν μηδέν αν δεν γίνουνταν σφάλματα στρογγύλευσης κατά το k-στό βήμα). Ορίζουμε τώρα $\varepsilon_{ij}^{(k)}=0$ γιά τα υπόλοιπα i, j (δηλ. συνολικά ορίζουμε τα $\varepsilon_{ij}^{(k)}$ όπως στην (14)) και θεωρούμε τον nxn πίνακα $E^{(k)}$ με στοιχεία $\varepsilon_{ij}^{(k)}$, $1\leq i,j\leq n$. 'Εστω ο nxn πίνακας $L^{(k)}$ όπου

$$L_{ij}^{(k)} = \begin{cases} m_{ik}, & \text{αν } k+1\leq i\leq n,\ j=k \\ 0, & \text{αλλοιώς} \end{cases}$$

Μπορούμε τώρα εύκολα να δείξουμε, χρησιμοποιώντας τις σχέσεις (15),(17) και (9) ότι

$$A^{(k+1)} = A^{(k)} - L^{(k)}A^{(k)}+E^{(k)}.$$

Αθροίζοντας και τα δύο μέλη αυτής της ισότητας ως πρός k από k=1 έως n-1 παίρνουμε.

$$(19)\quad L^{(1)}A^{(1)}+L^{(2)}A^{(2)}+\ldots+L^{(n-1)}A^{(n-1)}+A^{(n)} = A^{(1)}+\sum_{k=1}^{n-1} E^{(k)}.$$

Υπολογίζοντας τα στοιχεία του πίνακα $L^{(1)}A^{(1)}$ βλέπουμε ότι γιά $1\leq i,j\leq n$

$$(L^{(1)}A^{(1)})_{ij} = \begin{cases} 0, & \text{αν } i=1 \\ m_{i1}a_{1j}, & \text{αλλοιώς} \end{cases} = (L^{(1)}A^{(n)})_{ij}\ .$$

Συνεπώς $L^{(1)}A^{(1)} = L^{(1)}A^{(n)}$. Κατά τον ίδιο τρόπο, επειδή η k-στή γραμμή του $A^{(k)}$ συμπίπτει με την k-στή γραμμή του $A^{(n)}$, συμπεραίνουμε ότι ο πίνακας $L^{(k)}A^{(k)}$ (που εξαρτάται, εκτός από τα m_{ik}, μόνο από την k-στή γραμμή του $A^{(k)}$) συμπίπτει με τον $L^{(k)}A^{(n)}$. Συνεπώς η (19) δίνει, λόγω της (13)

$$(L^{(1)}+L^{(2)}+\ldots+L^{(n-1)}+I)A^{(n)} = A+E,$$

από την οποία προκύπτει η (12) λόγω της (10) και του ότι

$$\sum_{k=1}^{n-1} L^{(k)} + I = L,\ \text{βλ. (11)}\ @$$

Στο επόμενο βήμα φράσσουμε τα στοιχεία του πίνακα E καθώς και την νόρμα $\|E\|_\infty$.

Λήμμα 4. Γιά τον πίνακα E που κατασκευάσαμε στο Λήμμα 3 ισχύει

(20) $\quad \|E\|_\infty < n^2 p \, \|A\|_\infty \, u,$

όπου u το μοναδιαίο σφάλμα στρογγύλευσης και όπου

(21) $\quad p = \max_{i,j,k} |a_{ij}^{(k)}| \, / \, \|A\|_\infty,$

όπου τα $a_{ij}^{(k)}$ είναι τα ενδιάμεσα προϊόντα των υπολογισμών της απαλοιφής που ορίζονται αναδρομικά από την (9).

<u>Απόδειξη</u>. Γιά να να φράξουμε τα στοιχεία $E_{ij} = \sum_{k=1}^{n-1} \varepsilon_{ij}^{(k)}$ του E πρέπει να φράξουμε, λόγω της (14), τις ποσότητες m_{ik} και $a_{ij}^{(k)}$. Θυμόμαστε την υπόθεση ότι οι γραμμές του A έχουν διαταχθεί κατάλληλα έτσι ώστε οι οδηγοί που προκύπτουν από τους υπολογισμούς με αριθμητική πεπερασμένης ακρίβειας, δηλ. τα $a_{kk}^{(k)}$, να είναι τα απολύτως μεγαλύτερα, γιά κάθε k, από τα στοιχεία $a_{ik}^{(k)}$, $k \le i \le n$. Επειδή $m_{ik} = fl(a_{ik}^{(k)} / a_{kk}^{(k)})$, ανακαλώντας την παραδοχή μας ότι πρώτα γίνεται η πράξη $a_{ik}^{(k)} / a_{kk}^{(k)}$ ακριβώς και μετά στρογγυλεύεται ή αποκόπτεται το αποτέλεσμα, βλέπουμε ότι αυτό σημαίνει ότι

(22) $\quad |m_{ik}| \le 1$ γιά όλα τα i,k, (i>k).

Ορίζοντας την ποσότητα p από την (21) - θα επανέλθουμε βέβαια γιά να σχολιάσουμε την σημασία της και τις θεωρητικές και πρακτικές εκτιμήσεις της· πρός το παρόν αρκούμαστε να παρατηρήσουμε ότι μπορεί να υπολογισθεί εύκολα σαν παραπροϊόν της απαλοιφής - "λύνουμε" το πρόβλημα της εκτίμησης των $a_{ij}^{(k)}$ γιατί βέβαια εξ ορισμού

$$(23)\qquad |a_{ij}^{(k)}| \le p\,\|A\|_\infty .$$

Από τις (14), (22) και (23) συμπεραίνουμε ότι

$$(24)\qquad |\varepsilon_{ij}^{(k)}| \le p\|A\|_\infty \cdot \begin{cases} u & \text{αν } k+1\le i\le n,\ j=k \\ 2u & \text{αν } k+1\le i,j\le n \\ 0 & \text{αλλοιώς.} \end{cases}$$

Έστω $|B|$ ο πίνακας με στοιχεία τις απόλυτες τιμές $|b_{ij}|$ των στοιχείων ενός πίνακα B. Επίσης αν $B,C\in\mathbb{R}^{n\times n}$ θα γράφουμε $B\le C$ αν και μόνο αν $b_{ij}\le c_{ij}$, $1\le i,j\le n$. Ανακαλώντας τον ορισμό του πίνακα E έχουμε από την (24) ότι

$$(25)\qquad |E| \le p\|A\|_\infty\, u\, C,$$

όπου ο C είναι ο nxn πίνακας που δίνεται από το άθροισμα των nxn πινάκων

$$\begin{pmatrix} 0 & 0 & \dots & 0 \\ 1 & 2 & \dots & 2 \\ 1 & 2 & \dots & 2 \\ \vdots & \vdots & & \vdots \\ 1 & 2 & \dots & 2 \end{pmatrix} + \begin{pmatrix} 0 & 0 & & \dots & 0 \\ 0 & 0 & & \dots & 0 \\ 0 & 1 & 2 & \dots & 2 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 2 & \dots & 2 \end{pmatrix} + \dots + \begin{pmatrix} 0 & \dots & & & 0 \\ \vdots & & & & \vdots \\ 0 & \dots & & & 0 \\ 0 & 0 & \dots & 0 & 1 & 2 \end{pmatrix}$$

Προφανώς ο C είναι ο πίνακας

$$C = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 1 & 2 & 2 & 2 & \dots & 2 & 2 \\ 1 & 3 & 4 & 4 & \dots & 4 & 4 \\ 1 & 3 & 5 & 6 & \dots & 6 & 6 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 3 & 5 & \dots & \dots & 2n-4 & 2n-4 \\ 1 & 3 & 5 & \dots & \dots & 2n-3 & 2n-2 \end{pmatrix}$$

'Αρα η (25) δίνει

$$\|E\|_\infty \le p\|A\|_\infty u \, \|C\|_\infty = p\|A\|_\infty \, u \, \sum_{j=1}^{n} c_{nj} =$$

$$= p\|A\|_\infty \, u \, (\sum_{j=1}^{n-1} (2j-1)+2n-2) = (n^2-1) \, p\|A\|_\infty u \quad \text{ό.έ.δ} \; @$$

Συγχωνεύουμε τώρα τα λήμματα 3 και 4 στο παρακάτω

ΘΕΩΡΗΜΑ 1. Οι πίνακες L,U που υπολογίζουμε κατά την απαλοιφή Gauss με μερική οδήγηση χρησιμοποιώντας αριθμητική πεπερασμένης ακρίβειας με μοναδιαίο σφάλμα στρογγύλευσης u ικανοποιούν ακριβώς την ισότητα

(26) $LU = A + E$,

όπου

(27) $\|E\|_\infty < n^2 p \, \|A\|_\infty \, u$,

και όπου το p ορίζεται από την (21). @

Το πρώτο αυτό θεώρημα μας λέει ότι τα προϊόντα L,U των υπολογισμών της απαλοιφής είναι η (ακριβής) ανάλυση LU ενός πίνακα A+E, κοντινού στον A, εφ' όσον το n δεν είναι πολύ μεγάλο και εφ' όσον η ποσότητα $\max_{i,j,k} |a_{ij}^{(k)}|$ - περί της οποίας περισσότερα παρακάτω - παραμένει φραγμένη. Επαναλαμβάνουμε ότι έχουμε υποθέσει ότι οι εναλλαγές γραμμών έγιναν όλες εκ των προτέρων.

Μετά την κατασκευή των L και U, η κατά προσέγγισιν λύση $\widetilde{x}$ του συστήματος $Ax=b$ θα βρεθεί με την επίλυση (με αριθμητική πεπερασμένης ακρίβειας) των τριγωνικών συστημάτων $Ly=b$, $U\widetilde{x}=y$. Σφάλματα στρογγύλευσης θα υπεισέλθουν φυσικά και σ' αυτές τις πράξεις. Ας εξετάσουμε λεπτομερώς π.χ. την αριθμητική στο πρώτο σύστημα $Ly=b$. Τα y_i θα βρεθούν από τον αλγόριθμο

$$(28)\quad \begin{cases} y_1 = fl(b_1/L_{11}) \\ y_i = fl((-L_{i1}y_1-L_{i2}y_2 - \ldots - L_{i,i-1}y_{i-1}+b_i)/L_{ii}),\ i=2,\ldots,n \end{cases}$$

Στον (28) αγνοούμε το γεγονός ότι $L_{ii}=1$ και συνεπώς ότι η διαίρεση διά L_{ii} είναι τετριμμένη - και ότι δεν εκτελείται! - έτσι ώστε τα αποτελέσματα να ισχύουν γιά οποιοδήποτε κάτω (και κατ' επέκτασιν και άνω) τριγωνικό σύστημα. Σημαντική είναι η <u>σειρά</u> των πράξεων κατά τον υπολογισμό του y_i. 'Οπως γράφτηκε ο (28), υπονοεί ότι ο αλγόριθμος της κατασκευής του y_i είναι ο

$$(29)\quad \begin{aligned} y_i &\leftarrow -\sum_{j=1}^{i-1} L_{ij}\, y_j \\ y_i &\leftarrow y_i+b_i \\ y_i &\leftarrow y_i/L_{ii} \end{aligned}$$

όπου ο υπολογισμός του εσωτερικού γινομένου $\sum_{j=1}^{i-1} L_{ij}y_j$ γίνεται όπως στην (4). Θα ήταν π.χ. διαφορετικά τα αποτελέσματα αν ο αλγόριθμος ήταν ο

$$y_i \leftarrow b_i$$

$$(30) \quad y_i \leftarrow y_i - \left(\sum_{j=1}^{i-1} L_{ij}\, y_j\right)$$

$$y_i \leftarrow y_i/L_{ii}.$$

Χρησιμοποιώντας τώρα την (2) γιά τις πράξεις του εσωτερικού γινόμενου και την (2') γιά την πρόσθεση του b_i (γιά $i \geq 2$) και την διαίρεση διά L_{ii} έχουμε από την (28) ότι

$$(31) \begin{cases} y_1 = b_1/L_{11}\,(1+\delta_{11}) \\ y_i = [b_i - L_{i1}(1+\delta_{i1})y_1 - \ldots - L_{i,i-1}(1+\delta_{i,i-1})y_{i-1}]/L_{ii}\,(1+\delta_{ii})(1+\delta'_{ii}), \\ \qquad 2 \leq i \leq n \end{cases}$$

όπου (χρησιμοποιώντας την (7) στην απόδειξη του Λήμματος 2) έχουμε

$$(32) \begin{cases} |\delta_{ii}|,\ |\delta'_{ii}| \leq u, \ 1 \leq i \leq n. \\ |\delta_{i1}| \leq (i-1)\ 1.01\ u, \ 2 \leq i \leq n \\ |\delta_{ij}| \leq (i+1-j)\ 1.01\ u, \ 2 \leq i \leq n, \ 2 \leq j \leq i-1. \end{cases}$$

Οι σχέσεις (31) γράφονται τώρα ως εξής:

$$\begin{cases} L_{11}(1+\delta_{11})y_1 = b_1 \\ L_{i1}(1+\delta_{i1})y_1 + \ldots + L_{i,i-1}(1+\delta_{i,i-1})y_{i-1} + L_{ii}(1+\delta_{ii})(1+\delta'_{ii})y_i = b_i, \ 2 \leq i \leq n, \end{cases}$$

δηλ. ορίζοντας τον κάτω τριγωνικό πίνακα $\delta L: (\delta L)_{ij} = L_{ij}\delta_{ij}$, $i > j$, $(\delta L)_{11} = L_{11}\delta_{11}$, $(\delta L)_{ii} = L_{ii}(2.02\ \theta_i u)$, $2 \leq i \leq n$, όπου $|\theta_i| \leq 1$ – χρησιμοποιήσαμε την (3') – παίρνουμε το σύστημα

$$(31') \quad (L+\delta L)y = b$$

το οποίο ικανοποιεί <u>ακριβώς</u> η y (αυτί του $Ly=b$) όταν υπάρχουν σφάλματα στρογγύλευσης. Εύκολα βλέπουμε ότι

$$|\delta L| \le 1.01\, u \begin{pmatrix} |L_{11}| & & & & 0 \\ |L_{21}| & 2|L_{22}| & & & \\ 2|L_{31}| & 2|L_{32}| & 2|L_{33}| & & \\ \vdots & \vdots & \vdots & \ddots & \\ (n-1)|L_{n1}| & (n-1)|L_{n2}| & (n-2)|L_{n3}| & \ldots\ldots & 2|L_{nn}| \end{pmatrix}$$

Συνεπώς

$$(32') \quad \|\delta L\|_\infty \le 1.01\, u \max_{i,j} |L_{ij}|\, n(n+1)/2 \le 1.01\, u\, n(n+1)/2,$$

(ανακαλώντας ότι $\max_{i\neq j}|L_{ij}|\le 1$, βλ. (22), $L_{ii}=1$).

Εντελώς παρόμοια ισχύουν γιά το άνω τριγωνικό σύστημα $U\tilde{x}=y$: Η λύση $\tilde{x}$ ικανοποιεί ακριβώς την εξίσωση

$$(33) \quad (U+\delta U)\tilde{x}=y,$$

όπου, όπως στην (32'),

$$(34) \quad \|\delta U\|_\infty \le 0.01\, u \max_{i,j} |U_{ij}|\, n(n+1)/2 \le 1.01\, u\rho\|A\|_\infty\, n(n+1)/2,$$

επειδή $U=A^{(n)}$ και λόγω της (21). Παίρνουμε λοιπόν το εξής

Λήμμα 5. Η προσεγγιστική λύση $\tilde{x}$ του συστήματος $Ax=b$ που παίρνουμε με απαλοιφή Gauss με μερική οδήγηση είναι ακριβής λύση του συστήματος

$$(35) \quad (L+\delta L)(U+\delta U)\tilde{x}=b,$$

όπου οι ποσότητες $\|\delta L\|_\infty$, $\|\delta U\|_\infty$ φράσσονται όπως στις (32'), (34). @

Τέλος φθάνουμε στο κεντρικό αποτέλεσμα αυτής της παραγράφου:

ΘΕΩΡΗΜΑ 2. Η προσεγγιστική λύση $\tilde{x}$ του $Ax=b$ που δίνει η απαλοιφή Gauss με μερική οδήγηση είναι ακριβής λύση του συστήματος

$$(36)\quad (A+\delta A)\tilde{x}=b,$$

όπου

$$(37)\quad \|\delta A\|_\infty \leq 1.01\ (n^3+3n^2)\ p\|A\|_\infty\ u,$$

όπου το p ορίζεται από την (21), u είναι το μοναδιαίο σφάλμα στρογγύλευσης και όπου υποθέτουμε ότι $n^2u\leq 1$.

<u>Απόδειξη</u>: Η (35) δίνει ότι το $\tilde{x}$ ικανοποιεί την

$$(LU+\delta LU+L\delta U+\delta L\delta U)\tilde{x}=b.$$

Αλλά, από το Θεώρημα 1 έχουμε ότι $LU=A+E$. Συνεπώς η (36) ικανοποιείται με

$$\delta A=E+\delta LU+L\delta U+\delta L\delta U.$$

Άρα $\|\delta A\|_\infty \leq \|E\|_\infty + \|\delta L\|_\infty \|U\|_\infty + \|L\|_\infty \|\delta U\|_\infty + \|\delta L\|_\infty \|\delta U\|_\infty$.

Από την (20) $\|E\|_\infty < n^2p\ \|A\|_\infty\ u$. Επίσης έχουμε ότι $\|U\|_\infty = \max_i \sum_j |U_{ij}| \leq np\ \|A\|_\infty$, $\|L\|_\infty = \max_i \sum_j |L_{ij}| \leq n$. Συνεπώς η (37) έπεται από τα παραπάνω, τις εκτιμήσεις (32), (34) και την υπόθεσή μας ότι $n^2u\leq 1$. @

Συμπεραίνουμε ότι το αν η μέθοδος είναι ευσταθής, δηλ. το αν το $\|\delta A\|_\infty/\|A\|_\infty$ είναι μικρό, θα εξαρτηθεί κυρίως από τον λεγόμενο "συντελεστή μεγέθυνσης" $p=\max_{i,j,k} |a_{ij}^{(k)}| / \|A\|_\infty$. (Ο παράγοντας n^3+3n^2

οφείλεται στην χρήση νορμών πινάκων και σε χονδροειδείς εκτιμήσεις του φράγματος του $\|\delta A\|_\infty$ στην απόδειξη του Θεωρήματος 2 και δεν πρέπει να λαμβάνεται σοβαρά υπ' όψιν - βλέπε και Παρατήρηση 2 -.). Δεν είναι δύσκολο να δούμε, επειδή $|m_{ik}|\leq 1$, ότι

$$\max_{i,j}|a_{ij}^{(k+1)}| = \max_{i,j}|a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)}| \leq 2\max_{i,j}|a_{ij}^{(k)}|$$

και συνεπώς ότι $\max_{i,j,k}|a_{ij}^{(k)}| \leq 2^{n-1}\max_{i,j}|a_{ij}|$. (Μάλιστα ο Wilkinson έχει δώσει παράδειγμα πίνακα γιά τον οποίο ισχύει η ισότητα (!)). Γενικά λοιπόν $p\leq 2^{n-1}$. <u>Στην πράξη όμως τέτοια μεγέθυνση είναι εξαιρετικά σπάνια</u>. (Υπάρχει γενική συμφωνία ότι σχεδόν πάντα γιά μερική οδήγηση (εμπειρικά)

$\max_{i,j,k}|a_{ij}^{(k)}| \lesssim 10$ αν $\max_{i,j}|a_{ij}|\leq 1$)..

Γιά <u>ολική οδήγηση</u> (γιά την οποία ισχύει ένα ανάλογο θεώρημα 2 και γιά την οποία εμφανίζεται πάλι ο παράγων p στην εκτίμηση του $\|\delta A\|_\infty$), ο Wilkinson έδειξε ότι το φράγμα του p είναι της τάξης του $n^{\ln n/4}$ (δηλ. αυξάνεται πολύ αργότερα από το 2^n) αν και δεν είναι γνωστά παραδείγματα όπου $p>n$.

Με βάση πολλά εμπειρικά δεδομένα και αριθμητικά πειράματα με μερική οδήγηση - παρά την ύπαρξη πινάκων γιά τους οποίους ο συντελεστής μεγέθυνσης αυξάνει εκθετικά με τό n - συνήθως υποθέτουμε στην πράξη ότι

$$(38)\quad \frac{\|\delta A\|_\infty}{\|A\|_\infty} \simeq \varepsilon\beta^{-t},$$

όπου συνήθως το ε είναι της τάξεως του β. Επί το συντηρητικώτερον, γιά μεγάλους πίνακες, υποθέτουμε ότι

$$(38')\quad \frac{\|\delta A\|_\infty}{\|A\|_\infty} \simeq nu,$$

που σπανιώτατα δεν συμβαίνει. Στην <u>πράξη</u> λοιπόν δεχόμαστε ότι η απαλοιφή Gauss με μερική οδήγηση είναι συνήθως ((38)) ευσταθής ή το πολύ-πολύ ότι επιτρέπει μιά ασθενή (ανάλογη του n) μεγέθυνση των σφαλμάτων ((38')). Επαναλαμβάνουμε ότι αυτές οι εκτιμήσεις είναι <u>εμπειρικές</u> αλλά συνήθως ρεαλιστικές.

Τέλος, ας κλείσουμε την παράγραφο αυτή συνδιάζοντας τα της ευστάθειας του αλγορίθμου της απαλοιφής με μερική οδήγηση (θεώρημα 2) και τα αποτελέσματα της παρ. 1.2 γιά την ευαισθησία της λύσης του $Ax=b$ σε διαταραχές γιά να εκτιμήσουμε το <u>σφάλμα</u> της προσεγγιστικής λύσης $\tilde{x}$. Ας υποθέσουμε λοιπόν ότι η $\tilde{x}$ ικανοποιεί την (36) και ας θέσουμε

$$(39)\qquad \frac{\|\delta A\|}{\|A\|} = \mu,$$

γιά κάποια νόρμα $\|\cdot\|$. Οι (36), (39), (1.2.8), (1.2.9) δίνουν ($\tilde{x}=x+\delta x$) ότι αν $\kappa=\kappa(A)=\|A\|\,\|A^{-1}\|$, $\|\delta A\|\,\|A^{-1}\| = \mu\kappa <1$, τότε

$$(40)\qquad \|\tilde{x}-x\| \,/\, \|x\| \leq \kappa\mu/(1-\kappa\mu),$$

δηλ. ότι (αν π.χ. μ μικρό) τότε το σχετικό σφάλμα της $\tilde{x}$ (ως πρός $\|x\|$) μπορεί να είναι μεγάλο αν ο δείκτης κατάστασης κ είναι μεγάλος. Σημειώστε ότι γιά το <u>υπόλοιπο</u> $r=A\tilde{x}-b$ έχουμε ότι $r=A\tilde{x}-b = -\delta A\tilde{x}$ από την οποία έπεται ότι

$$(41)\qquad \|r\|/(\|A\|\,\|\tilde{x}\|) \leq \mu,$$

δηλ. ότι το "σχετικό" υπόλοιπο (ως πρός την ποσότητα $\|A\|\,\|\tilde{x}\|$ που μετράει την "κλίμακα" του προβλήματος) είναι πάντα μικρότερο ή ίσο του μ, όπου το μ ορίστηκε από την (39). Ανακαλώντας ότι γιά $\|\cdot\|=\|\cdot\|_\infty$, συνήθως $\mu=O(u)$, βλέπουμε ότι <u>το υπόλοιπο της προσεγγιστικής λύσης που δίνει η απαλοιφή Gauss με μερική οδήγηση είναι σχεδόν πάντα πολύ μικρό, ανεξάρτητα από την κατάσταση του συστήματος</u>. Μάλιστα αυτό ισχύει και γιά μη αντιστρέψιμο A (!) - αρκεί βέβαια να υπάρχει λύση $\tilde{x}$ του (36) -.

Τέλος, παρατηρώντας ότι $\tilde{x}-x = A^{-1}r$, βλέπουμε ότι $\|\tilde{x}-x\| \le \|A^{-1}\|\,\|r\|$ και συνεπώς από την (41) ότι

$$(42)\quad \|x-\tilde{x}\|/\|\tilde{x}\| \le \kappa\mu,$$

που πάλι δείχνει την σημασία του δείκτη κατάστασης στο σχετικό σφάλμα της $\tilde{x}$. Η (42) μάλιστα μας οδηγεί, αν ξέρουμε το κ και το μ, στην εκτίμηση του <u>απολύτου σφάλματος</u> $\|\tilde{x}-x\| \le \|\tilde{x}\|\kappa\mu$ συναρτήσει της υπολογιστικής λύσης $\tilde{x}$ – ένα παράδειγμα εκτίμησης <u>εκ των υστέρων</u> (a posteriori), δηλ. με γνώση του αλγοριθμικού αποτελέσματος $\tilde{x}$, σε αντίθεση με την εκτίμηση του $\|x-\tilde{x}\|$ που δίνει η (40) που είναι εκτίμηση <u>εκ των προτέρων</u> (a priori), δηλ. διατυπώνεται συναρτήσει της ακριβούς λύσης x.

Εκτιμήσεις σφαλμάτων όπως οι ανισότητες (40) και (42) δείχνουν ότι, όσο αφορά σφάλματα της προσεγγιστικής λύσης με απαλοιφή Gauss με μερική οδήγηση, ο δείκτης κατάστασης κ εμφανίζεται σαν όρος του γινομένου κμ όπου το μ είναι της μορφής $c_n u$ και όπου στην πράξη η σταθερά c_n δεν αυξάνεται γρήγορα με το n (βλ. (38),(38')). Έτσι π.χ. γιά ένα σύστημα με "μεγάλο" δείκτη κατάστασης ο αλγόριθμος είναι δυνατόν να δώσει λογικά αποτελέσματα αν κάνουμε τις πράξεις με αριθμητική <u>διπλής ακρίβειας</u>, οπότε $u=O(\beta^{1-2t})$· βεβαίως το κόστος των αριθμητικών πράξεων θα αυξηθεί.

<u>Παρατηρήσεις</u>

1. Αν τα στοιχεία του A και b δεν είναι όλα αριθμοί της μηχανής αλλά πραγματικοί αριθμοί μέσα στο εύρος των αριθμών της μηχανής, η (1) δείχνει ότι τα δεδομένα A,b θα παρασταθούν στον υπολογιστή από τα $\tilde{A},\tilde{b}$ (με στοιχεία αριθμούς μηχανής), όπου $\tilde{b}=b+\delta b$, $\|\delta b\|_\infty \le u\|b\|_\infty$, $\tilde{A}=A+\delta A$, $\|\delta A\|_\infty \le u\|A\|_\infty$. Τέτοια "αρχικά" σφάλματα μπορούν εύκολα να ενσωματωθούν στην ανάλυση που κάναμε σ' αυτήν την παράγραφο. Ας υποθέσουμε όμως, χάριν παιδιάς, ότι δεν γίνονται άλλα σφάλματα στρογγύλευσης κατά τις πράξεις της απαλοιφής πέρα από τα αρχικά αυτά σφάλματα παράστασης. Τότε η υπολογιστική λύση $\tilde{x}$ θα είναι ακριβής

λύση του συστήματος $\tilde{A}\tilde{x}=\tilde{b}$. Χρησιμοποιώντας τώρα τις πιό πάνω εκτιμήσεις γιά τα $\|\delta b\|_\infty$, $\|\delta A\|_\infty$ και την θεωρία της παρ. 1.2 μπορούμε να δούμε (βλ. 'Ασκ. 4) ότι αν η $\kappa_\infty(A)\equiv r<1$, τότε π.χ.

$\|x-\tilde{x}\|_\infty/\|x\|_\infty \leq 2r/(1-r)$, δηλ. ότι αν ο r είναι κοντά στην μονάδα, τότε θα πρέπει να περιμένουμε μεγάλα σφάλματα γιά την $\tilde{x}$ έστω και μόνο λόγω της προσεγγιστικής παράστασης των αριθμών στον υπολογιστή.

2. Μιά έστω και πρόχειρη ματιά στην βιβλιογραφία (π.χ. συγκρίνετε τα αποτελέσματα αυτής της παραγράφου - που στηρίζονται στην ανάλυση του Wilkinson όπως παρουσιάζεται στο βιβλίο [1.2] των Forsythe και Moler - με ανάλογα αποτελέσματα των βιβλίων [0.2], [0.4], [0.5], [1.4], [1.7], [1.9] κλπ.) μας πείθει ότι ακόμη και γιά τον ίδιο ακριβώς αλγόριθμο είναι δυνατόν με διαφορετικές αποδείξεις ή διαφορετικές υποθέσεις γιά την αριθμητική κλπ. να πάρουμε διαφορετικές σταθερές c_n σε εκτιμήσεις π.χ. της μορφής (37) - στην (37) $c_n=1.01(n^3+3n^2)$ - ή ακόμα και διαφορετικού τύπου φράγματα. (Βέβαια, γιά άλλο αλγόριθμο ή διαφορετική σειρά των πράξεων κάτι τέτοιο θα συμβεί σίγουρα). Γιαυτό συνήθως δεν δίνουμε και μεγάλη σημασία π.χ. σε τέτοιες σταθερές - στην πράξη όπως είδαμε είναι συνήθως ασφαλές να υποθέσουμε ότι $c_n\rho=O(n)$, βλ. (38') -. Πάνω σ' αυτό το θέμα ο Wilkinson, βλ. [1.4, σελ. 36], λέει:

> "Εξακολουθεί να υπάρχει μιά τάση να αποδίδεται υπερβολική σημασία στις λεπτομέρειες της μορφής των φραγμάτων που δίνει μιά a priori ανάλυση του σφάλματος. Κατά την γνώμη μου το φράγμα καθεαυτό είναι συνήθως το λιγώτερο σημαντικό κομμάτι μιάς τέτοιας ανάλυσης, ο κύριος στόχος της οποίας είναι να αποκαλύψει πιθανούς μηχανισμούς αστάθειας του αλγορίθμου - αν υπάρχουν - έτσι ώστε ενδεχομένως να μπορέσουμε να βελτιώσουμε τον αλγόριθμο. Συνήθως το φράγμα είναι πιό αδύνατο απ' ότι θα ήταν αν δεν είμαστε υποχρεωμένοι να περιορίσουμε σ' ένα λογικό επίπεδο το πλήθος των λεπτομερειών της απόδειξης και δεν είχαμε τον περιορισμό να εκφράσουμε τα σφάλματα χρησιμοποιώντας νόρμες πινάκων. Τα a priori φράγματα δεν είναι

εν χένει μεχέθη που πρέπει να χρησιμοποιούμε στην πράξη. Φράγματα που έχουν πρακτική άξία συνήθως προκύπτουν από κάποιας μορφής a posteriori ανάλυση του σφάλματος [Σημ. Μετ: δηλ. χρησιμοποιώντας ποσότητες που υπολογίζονται από τον αλγόριθμο]. Μιά τέτοια ανάλυση έχει το πλεονέκτημα ότι παίρνει υπ' όψη της την στατιστική κατανομή των σφαλμάτων στρογγύλευσης και τυχόν ειδικά χαρακτηριστικά του πίνακα όπως π.χ. την δομή των μηδενικών του".

3. Σε ορισμένες περιπτώσεις είναι δυνατόν πολλαπλασιάζοντας, πρίν λύσουμε το σύστημα, από αριστερά και δεξιά τον πίνακα A επί <u>διαγωνίους</u> πίνακες (που αντιστοιχεί σε πολλαπλασιασμό των γραμμών και των στηλών του A με σταθερές, δηλ. σε μιά εκ των προτέρων <u>αλλαγή κλίμακας</u> (scaling) των στοιχείων τῶν A,b,x) να πάρουμε ένα σύστημα με μικρότερο δείκτη κατάστασης, βλ. π.χ. [1.2],[1.4]. Τέτοιοι διαγώνιοι πίνακες μπορούν να βρεθούν γιά ειδικές περιπτώσεις πινάκων A αλλά το γενικό πρόβλημα της κατασκευής τους δεν έχει διελευκανθεί ακόμα αρκετά.

4. Αν ο δείκτης κατάστασης ενός πίνακα δεν είναι πολύ μεγάλος εν συγκρίσει πρός την ακρίβεια των πράξεων, δηλ. αν το γινόμενο $\kappa(A)u$ είναι αρκετά μικρό, τότε είναι δυνατόν, χρησιμοποιώντας τεχνικές <u>επαναληπτικής βελτίωσης</u> (interative improvement) να βελτιώσουμε την ακρίβεια της υπολογιστικής λύσης $\tilde{x}$. Μιά τέτοια τεχνική υπολογίζει το <u>υπόλοιπο</u> r της $\tilde{x}$, λύνει ένα νέο σύστημα $Ay=r$ και βρίσκει μιά "διόρθωση" y, η οποία, προστιθέμενη στην $\tilde{x}$ δίνει μιά νέα προσέγγιση που συχνά είναι ακριβέστερη της $\tilde{x}$, εφ' όσον το υπόλοιπο r έχει υπολογισθεί με μεγαλύτερη ακρίβεια (π.χ. διπλή) απ' ότι οι υπόλοιπες πράξεις. Βλ. π.χ. [0.2], [1.2], [1.4].

5. Στον υπολογισμό στην πράξη του δείκτη κατάστασης $\kappa(A)=\|A\|\,\|A^{-1}\|$ θέλουμε ν' αποφύγουμε τον υπολογισμό του A^{-1} κατά τα γνωστά. Συνήθως λοιπόν δεν υπολογίζουμε ακριβώς τον $\kappa(A)$ αλλά κάνουμε μιά προσεγγιστική του εκτίμηση: αυτό που χρειαζόμαστε είναι ουσιαστικά η τάξη μεγέθους του $\kappa(A)$ ώστε να μπορούμε να εκτιμήσουμε την αξιοπιστία των αποτελεσμάτων του υπολογισμού. Γιά το πώς εκτι-

μούμε τον κ(A) - με υπολογισμούς κόστους $O(n^2)$ - βλ. π.χ. [5.4, σελ. 67].

Ασκήσεις 1.3

1.(α) Θυμηθείτε την απόδειξη της (1) και αποδείξτε με κατάλληλη παραλλαγή της την σχέση

(1') $fl(x) = x/(1+\delta')$, όπου $|\delta'| \leq u$,

από την οποία έπεται η (2') αν κάνουμε τις ίδιες παραδοχές που οδήγησαν από την (1) στην (2).

(β) Αποδείξτε τους ισχυρισμούς στην απόδειξη του Λήμματος 1.

(γ) Υπό τις προϋποθέσεις του Λήμματος 2 δείξτε ότι

$$|fl(\sum_{i=1}^{n} x_i y_i) - \sum_{i=1}^{n} x_i y_i| \leq 1.01\, nu \sum_{i=1}^{n} |x_i||y_i|.$$

Συνεπώς αν π.χ. $|\sum_{i=1}^{n} x_i y_i| << \sum_{i=1}^{n} |x_i||y_i|$, δηλ. αν π.χ. πολλά ζευγάρια όρων $x_i y_i$ έχουν αντίθετα πρόσημα και αλληλοαναιρούνται, τότε είναι δυνατόν το σχετικό σφάλμα του $fl(\sum_{i=1}^{n} x_i y_i)$ να μην είναι μικρό. (Να ένα χρήσιμο συμπέρασμα από μιά a priori ανάλυση του σφάλματος στρογγύλευσης που αναδεικνύει ένα πιθανό μηχανισμό αποσταθεροποίησης του αλγορίθμου γιά τον υπολογισμό του $\sum_{i=1}^{n} x_i y_i$ βλ. παρατήρηση 2).

2.(α). Έστω A,B nxn πίνακες με στοιχεία αριθμούς μηχανής και έστω fl(AB) το υπολογιστικό τους γινόμενο που παίρνουμε με τον συνήθη τρόπο υπολογίζοντας τα εσωτερικά γινόμενα (γραμμή επί στήλη) με τον αλγόριθμο (4). Έστω $E \in R^{n\times n}$ τέτοιος ώστε $fl(AB)=AB+E$. Βρήτε φράγμα γιά τον |E| συναρτήσει των |A|, |B|, n,u. (Παράδειγμα <u>απ' ευθείας</u> ανάλυσης του σφάλματος του πολλαπλασιασμού πινάκων). Πότε είναι δυνατόν το $fl(AB)$ να έχει μεγάλο σχετικό σφάλμα;

(β) (Παράδειγμα αντίστροφης ανάλυσης του σφάλματος): Έστω A,B 2x2 άνω τριγωνικοί πίνακες. Βρήτε άνω τριγωνικούς 2x2 πίνακες $\tilde{A},\tilde{B}$ (με στοιχεία "κοντά" στά αντίστοιχα στοιχεία των A,B) τέτοιους ώστε $fl(AB)=\widetilde{\tilde{A}\tilde{B}}$ (ακριβές γινόμενο!). Εκτιμείστε τους πίνακες $|\delta A|$, αντστχ. $|\delta B|$, όπου $\delta A=\tilde{A}-A$, $\delta B=\tilde{B}-B$, συναρτήσει των n,u και $|A|$, αντστχ. $|B|$.

3. Αν αντί του αλγορίθμου (29) χρησιμοποιηθεί ο (30) γιά την επίλυση του κάτω τριγωνικού συστήματος $Ly=b$, βρήτε το αντίστοιχο δL (ώστε να ισχύει η (31')) και εκτιμείστε τον πίνακα $|\delta L|$ και την νόρμα $\|\delta L\|_\infty$.

4. (α) Αποδείξτε τους ισχυρισμούς στην παρατήρηση 1.

(β) Στην περίπτωση που τα στοιχεία του A είναι αριθμοί μηχανής (δηλ. όταν $\delta A=0$) και το μόνο σφάλμα στρογγύλευσης προέρχεται από την παράσταση του b, δείξτε ότι η εκτίμηση του ερωτήματος (α) απλουστεύεται σε $\|x-\tilde{x}\|_\infty/\|x\|_\infty \le r$. Εφαρμογή: Έστω ότι τα στοιχεία των A,b είναι ακέραιοι με εξαίρεση το b_1 που είναι ίσο με 10^{-1}. Παριστάνουμε τα στοιχεία των A,b στον VAX 11/780 και υποθέτουμε ότι οι υπόλοιπες πράξεις της απαλοιφής γίνονται ακριβώς. Ποιό είναι το αναμενόμενο $\|x-\tilde{x}\|_\infty/\|x\|_\infty$ αν $\kappa_\infty(A)=10^4$; Το ίδιο ερώτημα γιά απλή και διπλή ακρίβεια στον IBM 4361. (Η αριθμητική του 4361 είναι ίδια με την αριθμητική του συστήματος IBM 370. Γιά τις τιμές των β,t βλ. [5.4, σελ. 18]).

1.4 Η ΑΝΑΛΥΣΗ CHOLESKY ΓΙΑ ΣΥΜΜΕΤΡΙΚΟΥΣ, ΘΕΤΙΚΑ ΟΡΙΣΜΕΝΟΥΣ ΠΙΝΑΚΕΣ

Στην παράγραφο αυτή θα ασχοληθούμε με την επίλυση γραμμικών συστημάτων $Ax=b$, $b\in R^n$, όπου ο A είναι ένας nxn πραγματικός, <u>συμμετρικός</u> ($A=A^T$) και <u>θετικά ορισμένος</u> (δηλ. με την ιδιότητα ότι $x^TAx>0$ γιά κάθε $0\neq x\in R^n$) πίνακας. 'Ενας θετικά ορισμένος πίνακας είναι προφανώς αντιστρέψιμος. Τέτοιοι πίνακες εμφανίζονται συχνά στις εφαρμογές. Θα εξετάσουμε μία άμεση μέθοδο γιά την λύση του γραμμικού συστήματος, που μπορεί να θεωρηθεί σαν ειδική μορφή της ανάλυσης LU στην περίπτωσή μας, την λεγόμενη ανάλυση Cholesky.

Παίρνοντας $x=e^i\in R^n$ ($e_i=1$, $e_j=0$, $i\neq j$) βλέπουμε ότι $x^TAx=a_{ii}>0$· γενικώτερα κάθε κύριος υποπίνακας (βλ. Ασκ. 1) ενός συμμετρικού θετικά ορισμένου πίνακα είναι επίσης συμμετρικός και θετικά ορισμένος. Από αυτήν την παρατήρηση (βλ. Ασκ. 3) προκύπτει ότι συστήματα με συμμετρικούς και θετικά ορισμένους πίνακες μπορούν να επιλυθούν με απλή απαλοιφή Gauss χωρίς εναλλαγές γραμμών· μάλιστα οι οδηγοί $a_{ii}^{(i)}$ είναι όλοι θετικοί αριθμοί. Ο αλγόριθμος όμως που θα χρησιμοποιήσουμε εδώ θα είναι μιά παραλλαγή της απαλοιφής και στηρίζεται στο εξής.

ΘΕΩΡΗΜΑ 1 (Ανάλυση Cholesky). 'Εστω $A\in R^{nxn}$ συμμετρικός και θετικά ορισμένος πίνακας. Τότε υπάρχει μοναδικός κάτω τριγωνικός πίνακας L με θετικά διαγώνια στοιχεία (όχι αναγκαστικά μονάδες), τέτοιος ώστε να ισχύει η <u>ανάλυση Cholesky</u>:

$$(1) \quad A = LL^T$$

<u>Απόδειξη</u>: Με επαγωγή. Το θεώρημα προφανώς ισχύει γιά 1x1 θετικά ορισμένους πίνακες: $a_{11}>0$ και $L_{11}=(a_{11})^{1/2}$. 'Εστω ότι το θεώρημα ισχύει γιά (n-1)x(n-1) συμμετρικούς, θετικά ορισμένους πίνακες και έστω A ένας nxn τέτοιος πίνακας. Χωρίζουμε τον A σε υποπίνακες ως εξής:

$$A = \begin{pmatrix} d & u^T \\ u & H \end{pmatrix}$$

όπου $d=a_{11}>0$, u ένα (στηλο)διάνυσμα $(n-1)x1$ και $\widetilde{H}$ $(n-1)x(n-1)$ πίνακας. Μπορούμε να επαληθεύσουμε ότι ο A γράφεται τότε σαν γινόμενο ως εξής:

$$(2)\quad A = \begin{pmatrix} d^{1/2} & 0 \\ u/d^{1/2} & I_{n-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & H \end{pmatrix} \begin{pmatrix} d^{1/2} & u^T/d^{1/2} \\ 0 & I_{n-1} \end{pmatrix} ,$$

όπου $H=\widetilde{H}-uu^T/d$ και I_{n-1} είναι ο $(n-1)x(n-1)$ μοναδιαίος πίνακας. Ο πίνακας H είναι συμμετρικός και θετικά ορισμένος, γιατί γιά κάθε $x\epsilon R^{n-1}$, $x\neq 0$:

$$x^T H x = x^T(\widetilde{H}-uu^T/d)x = y^T \begin{pmatrix} d & u^T \\ u & \widetilde{H} \end{pmatrix} y = y^T A y > 0,$$

όπου το $y\epsilon R^n$ δίνεται από

$$y = \begin{pmatrix} -x^T u/d \\ x \end{pmatrix} .$$

Από την υπόθεση της επαγωγής τώρα ο $(n-1)x(n-1)$ συμμετρικός και θετικά ορισμένος πίνακας H γράφεται στη μορφή $H=L_H L_H^T$, όπου L_H κάτω τριγωνικός με θετικά διαγώνια στοιχεία. Συνεπώς, λόγω της (2) ο A ικανοποιεί

$$A = \begin{pmatrix} d^{1/2} & 0 \\ \\ u/d^{1/2} & I_{n-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ & \\ 0 & L_H \end{pmatrix} \begin{pmatrix} 1 & 0 \\ & T \\ 0 & L_H \end{pmatrix} \begin{pmatrix} d^{1/2} & u^T/d^{1/2} \\ \\ 0 & I_{n-1} \end{pmatrix} =$$

$$= \begin{pmatrix} d^{1/2} & 0 \\ \\ u/d^{1/2} & L_H \end{pmatrix} \begin{pmatrix} d^{1/2} & u^T/d^{1/2} \\ & T \\ 0 & L_H \end{pmatrix} \equiv LL^T$$

Το επαγωγικό βήμα τελείωσε. Γιά να δείξουμε την μοναδικότητα του L υποθέτουμε ότι υπάρχει κάτω τριγωνικός πίνακας M με θετικά διαγώνια στοιχεία τέτοιος ώστε $A=LL^T=MM^T$. Άρα $L^{-1}M=L^T(M^T)^{-1}$. Επειδή L^{-1},M κάτω τριγωνικοί, $L^T,(M^T)^{-1}$ άνω τριγωνικοί, έπεται ότι $L^{-1}M=L^T(M^T)^{-1}=D$ όπου D είναι διαγώνιος πίνακας: Τώρα $L^{-1}M=D \Rightarrow D_{ii}=M_{ii}/L_{ii}$ και $L^T(M^T)^{-1}=D \Rightarrow D_{ii}=L_{ii}/M_{ii}$. Άρα $(L_{ii})^2=(M_{ii})^2 \Rightarrow L_{ii}=M_{ii}$, γιατί $L_{ii},M_{ii}>0$. Συνεπώς $D=I_n \Leftrightarrow L=M$ ό.έ.δ. @

Γιά να βρούμε τώρα έναν αποτελεσματικό <u>αλγόριθμο</u> γιά την κατασκευή των στοιχείων L_{ij} του L συναρτήσει των a_{ij} εξισώνουμε ένα πρός ένα τα στοιχεία του (κάτω τριγώνου του) A με τα αντίστοιχα στοιχεία του γινομένου LL^T. Το στοιχείο (1,1) δίνει $a_{11}=(L_{11})^2 \Rightarrow L_{11}=(a_{11})^{1/2}$. Τα στοιχεία της δεύτερης γραμμής δίνουν $a_{21}=L_{21}.L_{11} \Rightarrow L_{21}=a_{21}/L_{11}$ και $a_{22}=(L_{21})^2+(L_{22})^2 \Rightarrow L_{22}=(a_{22}-(L_{21})^2)^{1/2}$.

Γενικά, εξισώνοντας τα στοιχεία της i-στής γραμμής του A με τα αντίστοιχα του LL^T παίρνουμε τις ισότητες

$$L_{ij} = (a_{ij} - \sum_{k=1}^{j-1} L_{ik}L_{jk})/L_{jj} \text{ γιά } j=1,2,\ldots,i-1$$

και

$$L_{ii} = (a_{ii} - \sum_{k=1}^{i-1} (L_{ik})^2)^{1/2}.$$

Προκύπτει έτσι ο <u>αλγόριθμος Cholesky κατά γραμμές</u> (που υπολογίζει δηλ. τα στοιχεία L_{ij} του L κατά γραμμές, από την πρώτη πρός την n-στή) που εκτελείται ως εξής:

$$(3)\quad \begin{cases} \text{Γιά } i=1 \text{ έως } n: \\ \quad \begin{cases} \text{Γιά } j=1 \text{ έως } i-1: \\ L_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} L_{ik}L_{jk}\right)/L_{jj} \end{cases} \\ L_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} (L_{ik})^2\right)^{1/2}. \end{cases}$$

(Μιά ανακύκλωση από j=1 έως 0 είναι κενή. Παρόμοια ένα άθροισμα $\sum_{k=p}^{q}$ με $q<p$ είναι 0). Απ' του αλγόριθμο (3) βλέπουμε ότι η ανάλυση Cholesky δεν χρειάζεται επιπλέον μνήμη από τις $(n^2+n)/2$ θέσεις που χρησιμεύουν γιά την αποθήκευση των στοιχείων του (κάτω τριγώνου) του A, a_{ij}, $i\geq j$: η τιμή L_{ij} εκχωρείται στη θέση του στοιχείου a_{ij} που δεν χρειάζεται πιά στους υπολογισμούς. Εύκολα βλέπουμε ότι ο (3) απαιτεί $n^3/6+O(n^2)$ πράξεις (= πολλαπλασιασμούς και διαιρέσεις) καθώς και την εξαγωγή n τετραγωνικών ριζών.

Μετά την κατασκευή του κάτω τριγωνικού πίνακα L ο υπολογισμός της λύσης x του συστήματος $Ax=b$, δηλ. του $LL^Tx=b$, γίνεται κατά τα γνωστά με επίλυση των δύο τριγωνικών συστημάτων

$$(4)\quad Ly=b,\ L^Tx=y.$$

Η αριθμητική <u>ευστάθεια</u> του αλγορίθμου του Cholesky έχει μελετηθεί λεπτομερώς πάλι από τον Wilkinson (1968) που απέδειξε ότι η υπολογιστική λύση $\tilde{x}$ του συστήματος, που προκύπτει από την κατασκευή του L από τον αλγόριθμο (3) και την επίλυση των τριγωνικών συστημάτων (4) με αριθμητική πεπερασμένης ακρίβειας, ικανοποιεί το παραπλήσιο πρόβλημα $(A+\delta A)\tilde{x}=b$, όπου $\|\delta A\|_2 \leq c_n u \|A\|_2$, c_n σταθερά που εξαρτάται από το n αλλά αυξάνει αργά με το n, και u το μοναδιαίο σφάλμα στρογγύλευσης. Βλέπουμε δηλ. ότι η ανάλυση Cholesky είναι <u>ευσταθής</u> αλγόριθμος - παρατηρείστε ότι ο "συντελεστής μεγέθυνσης" των σφαλμάτων είναι p=1. (Τα αποτελέσματα αυτά έχουν αποδειχθεί και δεν αποτελούν εμπειρικές εκτιμήσεις). Επιπλέον, αν q_n u $\kappa_2(A)<1$, όπου q_n είναι άλλη μία μικρή σταθερά εξαρτώμενη από το n, ο Wilkinson δείχνει ότι ο αλγόριθμος τερματίζεται ομαλά, δηλ. ότι δεν εμφανίζεται ποτέ στους υπολογισμούς λόγω σφάλματος στρογγύλευσης ένα L_{ii} μηδέν ή φανταστικό. Υπάρχουν δηλ. περιπτώσεις <u>θεωρητικά</u> θετικά ορισμένων πινάκων που είναι "σχεδόν" μη αντιστρέψιμοι, δηλ. που έχουν $\kappa_2(A)>>1$, γιά τους οποίους, αν η ακρίβεια στις πράξεις δεν είναι επαρκής (δηλ. αν το u δεν είναι αρκετά μικρό), είναι δυνατόν ο αλγόριθμος της ανάλυσης του Cholesky να μην ολοκληρωθεί (βλ. Ασκ. 5)

<u>Παρατηρήσεις</u>

1. 'Ενας συμμετρικός και θετικά ορισμένος πίνακας A μπορεί να αναλυθεί και σε γινόμενο παραγόντων της μορφής:

$$A = MDM^T$$

όπου M κάτω τριγωνικός με μονάδες στην διαγώνιο και D διαγώνιος με θετικά στοιχεία. Η ανάλυση αυτή (ανάλυση Crout) προκύπτει π.χ. από την ανάλυση Cholesky αν $M=L\Delta^{-1}$ όπου Δ διαγώνιος πίνακας με $\Delta_{ii}=L_{ii}$ και $D=\Delta\Delta^T$. Προφανώς μπορούμε να βρούμε κατ' ευθείαν αλγόριθμο, ανάλογο του (3) γιά τον υπολογισμό των M_{ij}, $i>j$ και των D_{ii}.

2. Υπάρχουν και διάφοροι άλλοι τρόποι εκτός από του (3) γιά την κατασκευή των στοιχείων του πίνακα L. Π.χ. εξισώνοντας στην $A=LL^T$ τα στοιχεία <u>κατά στήλες</u> παίρνουμε τον αλγόριθμο της "ανάλυσης Cholesky κατά στήλες" κατά τον οποίο πρώτα υπολογίζουμε τα στοιχεία L_{i1}, $i\geq 1$, μετά τα L_{i2}, $i\geq 2$ κ.ο.κ. Μιά άλλη μορφή του αλγορίθμου, ή μορφή <u>γινομένου πινάκων</u>, στηρίζεται σε μία κατασκευαστική μορφή της απόδειξης του Θεωρήματος 1. Η προτίμηση γιά τον ένα ή τον άλλο αλγόριθμο υπαγορεύεται π.χ. από το πώς είναι αποθηκευμένος ο A, από

του A και του L (π.χ. αν είναι αραιοί) κ.τ.λ. Γιά περισσότερες λεπτομέρειες βλ. [5.4, παρ. 5,6].

3. Φυσικά λόγω του Θεωρήματος 1 ύπαρξης-μοναδικότητας του L, είμαστε εκ των προτέρων βέβαιοι ότι στον αλγόριθμο (3) γιά κάθε i, $a_{ii} > \sum_{k=1}^{i-1} (L_{ik})^2$, δηλ. ότι το L_{ii} που θα προκύψει θα είναι πραγματικό και θετικό. (Αυτό γιατί $A = LL^T \Rightarrow a_{ii} = L_{ii}^2 + \sum_{k=1}^{i-1} (L_{ik})^2$ και $L_{ii}>0$). Συνεπώς, αν έχουμε ένα (πραγματικό) συμμετρικό πίνακα A και θέλουμε να ελέγξουμε υπολογιστικά αν είναι θετικά ορισμένος, εκτελούμε τον αλγόριθμο του Cholesky. Αν ο αλγόριθμος ολοκληρωθεί ομαλά, δηλ. αν $L_{ii}>0$, $1\le i\le n$, τότε κατασκευάσαμε κάτω τριγωνικό πίνακα L με $L_{ii}>0$ τέτοιον ώστε $A=LL^T$. Συνεπώς (Ασκ. 2γ) ο A είναι θετικά ορισμένος. (Αν όμως όπως είδαμε παραπάνω ο πίνακας A έχει μεγάλο δείκτη κατάστασης σχετικά με την ακρίβεια των πράξεων τότε είναι δυνατόν, λόγω σφαλμάτων στρογγύλευσης, ο αλγόριθμος στην πράξη να μην τερματισθεί, βλ. Ασκ. 5).

4. Στα κεφάλαια 6-10 των [5.4] - που ακολουθούν πιστά το βιβλίο [1.3] των George και Liu - γίνεται μιά λεπτομερής μελέτη της μεθόδου Cholesky γιά αραιούς, μεγάλους, συμμετρικούς και θετικά ορισμένους πίνακες A. Συστήματα με τέτοιους πίνακες εμφανίζονται συχνότατα στις εφαρμογές. Στην περίπτωση ενός μεγάλου αραιού πίνακα A (του οποίου δηλ. τα περισσότερα στοιχεία είναι μηδέν) μας ενδιαφέρει να αναδιατάξουμε τις γραμμές και τις στήλες του A, να αποθηκεύσουμε με κατάλληλες δομές δεδομένων τα (μη μηδενικά) στοιχεία του και να χρησιμοποιήσουμε μιά κατάλληλη μορφή του αλγορίθμου Cholesky έτσι ώστε να ελαχιστοποιήσουμε κατά το δυνατόν το "γέμισμα" του L (δηλ. το πλήθος των μη μηδενικών στοιχείων L_{ij} του L γιά τα οποία τα αντίστοιχα στοιχεία a_{ij} του A ήταν μηδέν), το κόστος αποθήκευσης του A και L καθώς και το πλήθος των πράξεων. Τα προβλήματα που εμφανίζονται είναι εξαιρετικά ενδιαφέροντα και από μαθηματική άποψη

(χρησιμοποιούνται οι μέθοδοι της θεωρίας γραφημάτων) και από άποψη προγραμματισμού, αριθμητικής ανάλυσης και θεωρίας δομών δεδομένων.

5. Τέλος, μία σημείωση ιστορικού ενδιαφέροντος: ο André-Louis Cholesky (1875-1918) ήταν Γάλλος αξιωματικός του μηχανικού που έκανε γεωδαιτικές και τοπογραφικές μετρήσεις στην ΚΡΗΤΗ και στην Βόρεια Αφρική πρίν από τον Α' Παγκόσμιο Πόλεμο. Ανακάλυψε την ομώνυμη μέθοδο γιά να υπολογίζει λύσεις γραμμικών συστημάτων που προκύπτουν από τις λεγόμενες "κανονικές εξισώσεις" της μεθόδου ελαχίστων τετραγώνων που εφαρμόζεται γιά την προσέγγιση δεδομένων σε γεωδαιτικά προβλήματα. Η μέθοδος του δημοσιεύθηκε μετά θάνατον στο Bulletin Geodesique το 1924.

Ασκήσεις 1.4

1. 'Εστω $A=(a_{ij})$ ένας $n\times n$ συμμετρικός και θετικά ορισμένος πίνακας. Να δειχθεί ότι
   (α) ο A^{-1} είναι επίσης θετικά ορισμένος.
   (β) Κάθε τετραγωνικός υποπίνακας του A του οποίου η κύρια διαγώνιος βρίσκεται πάνω στην κύρια διαγώνιο του A (δηλ. κάθε πίνακας με στοιχεία $k\leq i,j\leq m$, $1\leq k\leq m\leq n$) είναι θετικά ορισμένος.
   (γ) Γιά κάθε k, $1\leq k\leq n$, ισχύει ότι
   $\max_{1\leq i\leq k} a_{ii} = \max_{1\leq i,j\leq k} |a_{ij}|$.

2. (α) Δείξτε ότι ο τριδιαγώνιος συμμετρικός πίνακας A με $a_{ii}=\lambda$, $a_{i,i\pm1}=1$, $\lambda\geq 2$ είναι θετικά ορισμένος.
   (β) Ένας συμμετρικός πραγματικός πίνακας είναι θετικά ορισμένος αν και μόνο αν οι ιδιοτιμές του είναι θετικές.
   (γ) 'Εστω L πραγματικός, κάτω τριγωνικός πίνακας με θετικά διαγώνια στοιχεία. Τότε ο πίνακας $A=LL^T$ είναι συμμετρικός και θετικά ορισμένος.
   (δ) Αν ο A είναι συμμετρικός και θετικά ορισμένος τότε ο πίνακας B^TAB έχει τις ίδιες ιδιότητες αν και μόνο αν ο B είναι αντιστρέψιμος.

3. 'Εστω $A\in R^{n\times n}$, συμμετρικός και θετικά ορισμένος. Δείξτε ότι η τριγωνοποίησή του κατά την απαλοιφή Gauss μπορεί να γίνει χωρίς εναλλαγές γραμμών και μάλιστα ότι όλοι οι οδηγοί $a_{ii}^{(i)}$ είναι θετικοί (Υπόδειξη: Προφανώς $a_{11}^{(1)} = a_{11}>0$. Θεωρείστε τον υποπίνακα $\tilde{A}=(a_{ij}^{(2)})$, $2\le i,j\le n$ που προκύπτει από τον $A^{(2)}$ αν αφαιρέσουμε την πρώτη γραμμή και την πρώτη στήλη του. Δείξτε ότι ο $\tilde{A}$ είναι συμμετρικός και θετικά ορισμένος. Συνεπώς $a_{22}^{(2)}>0$. Η επαγωγική συνέχεια της απόδειξης προφανής).

4. (α) Δείξτε ότι ο αλγόριθμος (3) απαιτεί $(n^3+3n^2-4n)/6$ πράξεις (=πολ/σμούς και διαιρέσεις) και n τετραγωνικές ρίζες.

(β) 'Εστω ότι ο <u>μιγαδικός</u> πίνακας $H=A+iB$ (όπου $A,B\in R^{n\times n}$) είναι αυτοσυζυγής ($H=H^*$) και θετικά ορισμένος ως μιγαδικός πίνακας (δηλ. ισχύει $(z,Hz)_2>0$ γιά κάθε $0\ne z\in C^n$. Δείξτε ότι ο $2n\times 2n$ πραγματικός πίνακας

$$C = \begin{pmatrix} A & -B \\ B & A \end{pmatrix}$$

είναι συμμετρικός και θετικά ορισμένος. Βρήτε έναν αλγόριθμο με $4n^3/3+O(n^2)$ (πραγματικές) πράξεις γιά την λύση του συστήματος $H(x+iy)=(b+ic)$ όπου $x,y,b,c \in R^n$. Πόση μνήμη απαιτείται;

5. Θεωρείστε τον πίνακα

$$A = \begin{pmatrix} 100 & .15 & .01 \\ .15 & 2.3 & .01 \\ .01 & .01 & 1 \end{pmatrix}.$$

Δείξτε ότι με αριθμητική απεριόριστης ακρίβειας, η μέθοδος του Cholesky τερματίζεται κανονικά (= απόδειξη ότι ο A είναι θετικά ορισμένος). Δείξτε όμως ότι αν οι πράξεις γίνουν με $\beta=10$, $t=2$ και στρογγύλευση, τότε οι υπολογιστικές τιμές των στοιχείων του L είναι $\tilde{L}_{11}=10$, $\tilde{L}_{21}=1.5$, $\tilde{L}_{31}=10^{-3}$, $\tilde{L}_{22}=0.0$, δηλ. ότι ο αλγόριθμος σταματά επιχειρώντας να υπολογίσει το $\tilde{L}_{32}$.

1.5 ΜΕΘΟΔΟΙ ΕΛΑΧΙΣΤΟΠΟΙΗΣΗΣ

Μία σημαντική κατηγορία μεθόδων γιά την επίλυση γραμμικών συστημάτων (κυρίως με συμμετρικούς, θετικά ορισμένους πίνακες) είναι μέθοδοι ελαχιστοποίησης καταλλήλων συναρτησιακών $\varphi(x)$, $x\in R^n$. Αν το $\varphi: R^n \to R$ είναι ομαλό, τότε από τον Απειροστικό Λογισμό, ξέρουμε ότι τα τοπικά του ελάχιστα πρέπει να αναζητηθούν μέσα στο σύνολο των κρισίμων σημείων του φ, δηλ. μεταξύ των σημείων x όπου $\nabla\varphi(x)=0$, όπου ως γνωστόν η κλίση $\nabla\varphi$ του φ είναι το διάνυσμα

$$\nabla\varphi = \left(\frac{\partial\varphi}{\partial x_1}, \frac{\partial\varphi}{\partial x_2}, \ldots, \frac{\partial\varphi}{\partial x_n}\right)^T .$$

Επίσης ξέρουμε ότι αν $\nabla\varphi(x)\neq 0$, τότε η συνάρτηση $g(\alpha)=\varphi(x+\alpha u)$, $x,u \in R^n$, $\alpha\in R$, που περιγράφει την συμπεριφορά του φ κοντά στο σημείο x κατά την κατεύθυνση u, ελαττώνεται με μέγιστο ρυθμό ελάττωσης όταν $u=-\nabla\varphi(x)$ λόγω του ότι $g'(0)=(\nabla\varphi(x),u)_2$.

Η τελευταία παρατήρηση μας οδηγεί σε μία προσεγγιστική επαναληπτική μέθοδο γιά τον υπολογισμό των ελαχίστων x^* της $\varphi(x)$, τήν λεγόμενη μέθοδο της "καθόδου μεγίστης κλίσεως" (steepest descent) του Cauchy, η οποία ανάγει το πρόβλημα σε επανειλημμένους υπολογισμούς ελαχίστων συναρτήσεων μιάς μεταβλητής. Η μέθοδος παράγει μιά ακολουθία $\{x^j\}, j\geq 0$ προσεγγίσεων ενός (τοπικού) ελαχίστου x^* της $\varphi(x)$ ως εξής: 'Εστω x^k μία προσέγγιση. Η x^{k+1} ορίζεται ως το πλησιέστερο στο x^k ελάχιστο της $\varphi(x)$ όταν το x περιορίζεται πάνω στην ακτίνα που διέρχεται από το x^k και έχει την διεύθυνση του $-\nabla\varphi(x^k)$. Βρίσκουμε δηλ. (γενικά προσεγγιστικά) τον μικρότερο θετικό αριθμό α_* όπου η συνάρτηση $g(\alpha) = \varphi(x^k-\alpha\nabla\varphi(x^k))$ έχει ελάχιστο και ορίζουμε $x^{k+1} = x^k-\alpha_*\nabla\varphi(x^k)$. Συνεπώς η τακτική μας σε κάθε βήμα k του αλγορίθμου είναι να ελαχιστοποιήσουμε την $\varphi(x)$ κοντά στο x^k πάνω στη διεύθυνση $-\nabla\varphi(x^k)$ κατά την οποία τοπικά η $\varphi(x)$ ελαττώνεται γρηγορώτερα.

Θεωρούμε τώρα το $n\times n$ πραγματικό σύστημα

(1) $Ax=b$

όπου ο A είναι συμμετρικός και θετικά ορισμένος. Συμβολίζοντας με $(\cdot,\cdot)=(\cdot,\cdot)_2$ το ευκλείδιο εσωτερικό γινόμενο στον R^n και με $\|\cdot\|=\|\cdot\|_2$ την αντίστοιχη νόρμα, θεωρούμε το πρόβλημα ελαχιστοποίησης γιά $x\in R^n$ του συναρτησιακού

$$(2) \quad \varphi(x) = (Ax,x)/2 - (b,x).$$

'Εστω $z=A^{-1}b$ η λύση του (1). Τότε επειδή γιά κάθε $y\in R^n$ ισχύει $\varphi(z+y)=\varphi(z)+(Ay,y)/2$, συμπεραίνουμε ότι $\varphi(z+y)>\varphi(z)$ $\forall 0\neq y\in R^n$, δηλ. ότι το $\varphi(x)$ παίρνει την ελάχιστη τιμή του στον R^n (ίση με $-(b,z)/2$) στο μοναδικό σημείο $x=z=A^{-1}b$. Συνεπώς το πρόβλημα της επίλυσης του συστήματος (1) είναι ισοδύναμο με το πρόβλημα <u>ελαχιστοποίησης χωρίς περιορισμούς</u>

$$(3) \quad \min_{x\in R^n} \varphi(x),$$

που έχει μοναδική λύση. Εξ άλλου επειδή η <u>κλίση</u> του φ είναι ίση με

$$(4) \nabla\varphi(x) = Ax-b,$$

το μοναδικό κρίσιμο σημείο του φ είναι το σημείο γιά το οποίο $\nabla\varphi(x)=0$, δηλ. το $A^{-1}b$ στο οποίο το φ έχει ελάχιστο επειδή είναι "αυστηρά κυρτό".

Στο σημείο x^k η συνάρτηση $\varphi(x)$ ελαττώνεται ταχύτερα στην κατεύθυνση $-\nabla\varphi(x^k)=b-Ax^k$, την οποία ταυτίζουμε με το <u>υπόλοιπο</u> r^k του x^k, δηλ. θέτουμε

$$(5) \quad r^k = -\nabla\varphi(x^k)=b-Ax^k.$$

Αν $r^k\neq 0$ (αλλοιώς $x^k=A^{-1}b$), η μέθοδος της καθόδου μεγίστης κλίσεως υπολογίζει συνεπώς το x^{k+1} ελαχιστοποιώντας το <u>διώνυμο</u> ως πρός a

$$\varphi(x^k+ar^k)=(Ar^k,r^k)a^2/2 - (r^k,r^k)a + \varphi(x^k)$$

του οποίου το ελάχιστο λαμβάνεται γιά $a=a_*>0$ όπου

$$a_* = (r^k,r^k)/(Ar^k,r^k).$$

Συνεπώς αν $r^k \neq 0$ θα έχουμε $\varphi(x^k+a_*r^k)<\varphi(x^k)$. Μάλιστα

$$\text{(6)} \quad \varphi(x^k+a_*r^k) = \varphi(x^k)-(r^k,r^k)^2/2\,(Ar^k,r^k)$$

Οδηγούμαστε λοιπόν στον εξής <u>αλγόριθμο</u> της μεθόδου της καθόδου μεγίστης κλίσεως γιά την ελαχιστοποίηση του (2):

$$\text{(7)} \qquad \begin{array}{l} x^0=0 \\ \left[\begin{array}{l} \text{Γιά } k=1,2,\ldots \\ \quad r^{k-1}=b-Ax^{k-1} \\ \quad \text{Αν } r^{k-1}=0 \\ \qquad \text{τότε τέλος, } x=x^{k-1}. \\ \qquad \text{αλλοιώς} \\ \qquad\quad a_k=(r^{k-1},r^{k-1})/(Ar^{k-1},\ r^{k-1}) \\ \qquad\quad x^k=x^{k-1}+a_kr^{k-1} \end{array}\right. \end{array}$$

Ο αλγόριθμος τερματίζεται όταν ικανοποιηθούν ένα ή περισσότερα από τα συνήθη κριτήρια τερματισμού επαναληπτικών μεθόδων (βλ. [5.4, σελ. 159],[1.8],[1.11]). Προφανώς, η αρχική τιμή $x^0=0$ δεν είναι δεσμευτική.

Θα μελετήσουμε την σύγκλιση της μεθόδου στην "φυσική" νόρμα του προβλήματος δηλ. στην νόρμα $x \longmapsto (Ax,x)^{1/2}$ που παράγεται από το εσωτερικό γινόμενο (Ax,y) στον R^n (βλ. Ασκ. 2(α)). Αποδεικνύουμε το εξής θεώρημα σύγκλισης που δίνει επίσης ένα μέτρο του πόσο γρήγορα συγκλίνει η ακολουθία $\{x^j\}$ στην νόρμα $(A\cdot,\cdot)^{1/2}$.

ΘΕΩΡΗΜΑ 1. 'Εστω $\{x^j\}$, $j\geq 0$ η ακολουθία που παράγει ο αλγόριθμος (7) της μεθόδου της καθόδου μεγίστης κλίσεως γιά οποιοδήποτε $x^o \in R^n$ και έστω x η λύση του συστήματος (1). 'Εστω $\kappa=\lambda_{max}/\lambda_{min}$, όπου λ_{max}, αντιστχ. λ_{min}, η μέγιστη, αντιστχ. ελάχιστη, ιδιοτιμή του A. Τότε $x^j \to x$, $j\to\infty$. Μάλιστα, αν $e^j=x-x^j$ είναι το σφάλμα της προσέγγισης x^j, ισχύει ότι

$$(8) \quad (Ae^j,e^j)^{1/2} \leq [(\kappa-1)/(\kappa+1)]^j (Ae^o,e^o)^{1/2}, \; j=0,1,2,\ldots$$

<u>Απόδειξη</u>. 'Εχουμε $Ae^j=Ax-Ax^j=b-Ax^j=r^j$. Επίσης χρησιμοποιώντας τον ορισμό του x^{j+1} έχουμε $e^{j+1}=x-x^{j+1}=x-x^j+x^j-x^{j+1}=e^j-a_{j+1}r^j$. Συνεπώς $Ae^{j+1}=Ae^j-a_{j+1}Ar^j$ από την οποία, χρησιμοποιώντας και τον ορισμό του a_{j+1} από τον (7)

$$(Ae^{j+1},r^j) = (Ae^j,r^j) - a_{j+1}(Ar^j,r^j) = (r^j,r^j) - a_{j+1}(Ar^j,r^j)=0.$$

Οι πιό πάνω σχέσεις δίνουν τώρα γιά οποιοδήποτε $a\in R$

$$(Ae^{j+1},e^{j+1}) = (Ae^{j+1},e^j-a_{j+1}r^j) = (Ae^{j+1},e^j) - a(Ae^{j+1},r^j) =$$
$$= (Ae^{j+1},e^j-ar^j).$$

Χρησιμοποιώντας τώρα την ανισότητα των Cauchy-Schwarz γιά το εσωτερικό γινόμενο $(A\cdot,\cdot)$ έχουμε

$$(Ae^{j+1},e^{j+1}) \leq (Ae^{j+1},e^{j+1})^{1/2} \, (A(e^j-ar^j),e^j-ar^j)^{1/2}, \; \forall a\in R.$$

Τέλος, από την σχέση $Ae^j=r^j$, έχουμε $e^j-ar^j=(I-aA)e^j$, και $A(e^j-ar^j)=$ $=A(I-aA)e^j=(I-aA)Ae^j$. Συμπεραίνουμε ότι

$$(9) \; (Ae^{j+1},e^{j+1}) \leq \inf_{a\in R} ((I-aA)Ae^j, \,(I-aA)e^j), \; j\geq 0.$$

Χρησιμοποιούμε τώρα την λεγόμενη "φασματική παράσταση" του A. Ο A, σαν συμμετρικός πίνακας, έχει η πραγματικές ιδιοτιμές (θετικές γιατί ο A είναι θετικά ορισμένος) $0<\lambda_1\equiv\lambda_{min}\leq\lambda_2\leq\ldots\leq\lambda_n\equiv\lambda_{max}$ και n αντίστοιχα, ορθοκανονικά ως πρός το $(\cdot,\cdot)$, ιδιοδιανύσματα u^j, $1\leq j\leq n$. 'Εχουμε δηλ. $Au^j=\lambda_j u^j$, $1\leq j\leq n$, $(u^i,u^j)=\delta_{ij}$. 'Αρα, γιά κάθε $u\in R^n$ ισχύει

$u=\sum_{j=1}^{n}(u,u^j)u^j$, και συνεπώς $Au=\sum_{j=1}^{n}\lambda_j(u,u^j)u^j$. Γενικώτερα, γιά κάθε πραγματικό πολυώνυμο π, η τελευταία σχέση δίνει

$\pi(A)u=\sum_{j=1}^{n}\pi(\lambda_j)(u,u^j)u^j$, $\forall u\in R^n$. Συνεπώς γιά κάθε $\alpha\in R$ έχουμε

$$((I-\alpha A)Ae^j,\ (I-\alpha A)e^j)=\sum_{i=1}^{n}(1-\alpha\lambda_i)^2\,\lambda_i(e^j,u^i)^2\leq$$

$$\leq \max_{1\leq i\leq n}(1-\alpha\lambda_i)^2\sum_{i=1}^{n}\lambda_i\,(e^j,u^i)^2=\max_{1\leq i\leq n}(1-\alpha\lambda_i)^2\,(Ae^j,e^j).$$

Συμπεραίνουμε, επειδή $\lambda_i\in[\lambda_{min},\lambda_{max}]$, ότι λόγω της (9)

$$(10)\quad (Ae^{j+1},e^{j+1})^{1/2}\leq \inf_{\alpha\in R}(\max_{\lambda\in[\lambda_{min},\lambda_{max}]}|1-\alpha\lambda|)(Ae^j,e^j)^{1/2},\ j\geq 0.$$

Το "min-max" πρόβλημα του υπολογισμού του

$$\sigma=\inf_{\alpha\in R}(\max_{\lambda\in[\lambda_{min},\lambda_{max}]}|1-\alpha\lambda|)$$

μπορεί βέβαια να λυθεί ως ειδική περίπτωση ενός γενικώτερου προβλήματος που θα αναλύσουμε στην παρ. 1.6 χρησιμοποιώντας πολυώνυμα Chebyshev με την παρατήρηση ότι
$\sigma=\inf_{\pi\in\Pi_1}\max_{\lambda\in[\lambda_{min},\lambda_{max}]}|\pi(\lambda)|$ όπου Π_1 το σύνολο των πραγματικών γραμμικών πολυωνύμων $\pi(\lambda)$ τέτοιων ώστε $\pi(0)=1$. Λύνεται όμως και στοιχειωδώς ως εξής: Προφανώς $\sigma=\inf_{\alpha\in R}[\max(|1-\alpha\lambda_{min}|,\ |1-\alpha\lambda_{max}|)]$. Στρέφοντας στο επίπεδο (λ,μ) την ευθεία $\mu=1-\alpha\lambda$ περί το σημείο $(0,1)$ καθώς το α μεταβάλλεται στο R, παρατηρούμε ότι η συνάρτηση $\alpha\mapsto\max(|1-\alpha\lambda_{min}|,|1-\alpha\lambda_{max}|)$ παίρνει την ελάχιστη τιμή της όταν

$1-\alpha\lambda_{min}>0$, $1-\alpha\lambda_{max}<0$ και $|1-\alpha\lambda_{min}|=|1-\alpha\lambda_{max}|$ δηλ. όταν $\alpha=2/(\lambda_{min}+\lambda_{max})$· έχουμε λοιπόν ότι $\sigma=|1-2\lambda_{min}/(\lambda_{min}+\lambda_{max})|=(\kappa-1)/(\kappa+1)$. Η (10) δίνει τότε

$$(8') \quad (Ae^{j+1},e^{j+1})^{1/2} \leq [(\kappa-1)/(\kappa+1)](Ae^{j},e^{j})^{1/2}, \; j\geq 0$$

από την οποία προκύπτει αμέσως η (8) και η σύγκλιση $e^j \to 0, j \to \infty$. @

Η άσκηση 4 δείχνει ότι η ταχύτητα σύγκλισης, δηλ. ο "λόγος" $p=(\kappa-1)/(\kappa+1)$ της "γεωμετρικής" σύγκλισης της ακολουθίας e^j στην νόρμα $(A\cdot,\cdot)^{1/2}$, είναι η καλύτερη δυνατή. Αλλά μπορεί να γίνει πολύ μικρή αν ο λόγος $\kappa=\lambda_{max}/\lambda_{min}$ είναι μεγάλος (δηλ. αν ο δείκτης κατάστασης $\kappa=\kappa_2(A)$ του A είναι μεγάλος - φυσικά!). Ενδιαφέρον έχει εδώ η γεωμετρική ερμηνεία της μεθόδου. (Ασκ. 5). Οι "ισοϋψείς" επιφάνειες $\varphi(x)$=σταθ. του συναρτησιακού φ είναι ελλειψοειδή στον R^n με κέντρο το σημείο $z=A^{-1}b$ και άξονες παράλληλους πρός τα ορθογώνια ιδιοδιανύσματα u^i του A με μήκη αξόνων ανάλογα πρός τους αριθμούς $\lambda_i^{-1/2}$, όπου λ_i οι ιδιοτιμές του A. (Αλλάξτε συντεταγμένες με νέο σύστημα αξόνων κέντρου z και άξονες παράλληλους πρό τα u^i). Από τον ορισμό του $\nabla\varphi(x^k)$, έπεται ότι η προσέγγιση x^{k+1} βρίσκεται πάνω στην κάθετο στο x^k της επιφάνειας του ελλειψοειδούς που διέρχεται από το σημείο x^k. Τα διανύσματα $x^{k+1}-x^k$ και x^k-x^{k-1} είναι εξ άλλου ορθογώνια (Ασκ. 1(γ)). Αν $\lambda_{max}>>\lambda_{min}$ τα ελλειψοειδή είναι υπερβολικά "στενόμακρα" και η κάθοδος πρός το κέντρο μιάς βαθιάς χαράδρας με απότομες πλαγιές γίνεται με ζίγκ-ζάγκ στις διαδοχικές ορθογώνιες κατευθύνσεις $r^k=-\nabla\varphi(x^k)$ μεταξύ των πλαγιών πράγμα που καθυστερεί πολύ την κάθοδο (βλ. και Ασκ. 4). Έτσι η τακτική της ελαχιστοποίησης <u>τοπικά</u> κατά μήκος των διευθύνσεων των υπολοίπων r^k (δηλ. των διευθύνσεων καθόδου μεγίστης κλίσεως) αποδεικνύεται κακή στρατηγική.

Επιχειρούμε λοιπόν τώρα να ελαχιστοποιήσουμε το φ διαδοχικά σε κατευθύνσεις $p^1, p^2, \ldots$ (που δεν συμπίπτουν αναγκαστικά με τα υπόλοιπα $r^0, r^1, r^2, \ldots$) ελπίζοντας ότι η επιλογή καταλλήλων p^i θα βελτιώσει την ταχύτητα σύγκλισης. Προκαταρκτικά, γενικεύοντας λίγο υπολογισμούς που κάναμε προηγουμένως, βλέπουμε ότι, γιά δεδομένα x^{k-1} και $p^k \neq 0$, η συνάρτηση $\varphi(x^{k-1}+\alpha p^k)$ ελαχιστοποιείται όταν

$$(11) \qquad \alpha = \alpha_k = (p^k, r^{k-1})/(Ap^k, p^k),$$

οπότε, αν $x^k = x^{k-1}+\alpha_k p^k$, θα ισχύει $\varphi(x^k) < \varphi(x^{k-1})$ εφ όσον $(p^k, r^{k-1}) \neq 0$. Μάλιστα,

$$(12) \qquad \varphi(x^k) = \varphi(x^{k-1}) - (p^k, r^{k-1})^2/2\,(Ap^k, p^k).$$

(Η μέθοδος της καθόδου μεγίστης κλίσεως επιλέγει $p^k = r^{k-1}$). Ο αλγόριθμος (7) γενικεύεται λοιπόν ως εξής:

(13)

$x^0 = 0$
Γιά $k = 1, 2, \ldots$
  $r^{k-1} = b - Ax^{k-1}$
  Αν $r^{k-1} = 0$
    τότε τέλος, $x = x^{k-1}$
    αλλοιώς
      διάλεξε p^k τέτοιο ώστε $(p^k, r^{k-1}) \neq 0$
      $\alpha_k = (p^k, r^{k-1})/(Ap^k, p^k)$
      $x^k = x^{k-1} + \alpha_k p^k$.

Το ερώτημα βέβαια είναι πως θα διαλέξουμε τα διανύσματα p^i έτσι ώστε ο αλγόριθμος (13) να συγκλίνει και να μην έχει τα μειονεκτήματα της μεθόδου της καθόδου μεγίστης κλίσεως. Παρατηρούμε, λόγω της αναδρομικής σχέσης $x^k = x^{k-1}+\alpha_k p^k$, ότι γιά κάθε k το διάνυσμα x^k που δίνει ο αλγόριθμος (13) δίδεται από ένα γραμμικό συνδυασμό $x^k = a_1 p^1 + \ldots + a_k p^k$ των p^i, $1 \leq i \leq k$. Θα ήταν λοιπόν ιδανικό π.χ. να διαλέγουμε τα διανύσματα p^i έτσι ώστε να είναι <u>γραμμικά ανεξάρτητα</u> και έτσι ώστε το x^k να λύνει το πρόβλημα ελαχιστοποίησης:

$$(14) \quad \min_{x \in \langle p^1, \ldots, p^k \rangle} \varphi(x)$$

όπου με $\langle a^1, \ldots, a^N \rangle$, $a^i \in R^n$, συμβολίζουμε τον υπόχωρο του R^n που παράγεται από τα a^i, $1 \leq i \leq N$, δηλ. το σύνολο των διανυσμάτων που είναι γραμμικοί συνδυασμοί των a^i, $1 \leq i \leq N$. (Το πρόβλημα ελαχιστοποίησης του $\varphi(x)$ γιά $x \in M$, όπου M υπόχωρος του R^n, έχει μοναδική λύση· βλ. Ασκ. 6).

Με μιά τέτοια κατασκευή των x^k, ο αλγόριθμος (13) θα συνέκλινε σε n βήματα στη λύση του συστήματος $Ax=b$! Πράγματι το x^n θα ελαχιστοποιούσε το $\varphi(x)$ γιά $x \in \langle p^1, \ldots, p^n \rangle = R^n \Rightarrow x^n = A^{-1}b$. Βέβαια, από την κατασκευή των a_k, x^k ο αλγόριθμος (13) παράγει γιά $k=1,2,\ldots$ ένα x^k που λύνει κάθε φορά το <u>μονοδιάστατο</u> πρόβλημα ελαχιστοποίησης.

$$(15) \quad \min_{a \in R} \varphi(x^{k-1} + a p^k).$$

Είναι δυνατόν να διαλέξουμε τα p^k έτσι ώστε το x^k, η λύση του (15), να είναι συγχρόνως και λύση του (14);

Η απάντηση είναι θετική. Ας κάνουμε μιά ανάλυση του προβλήματος επαγωγικά. Θεωρούμε τον $n \times j$ πίνακα $P_j = [p^1, \ldots, p^j]$ με στήλες $p^1, \ldots, p^j$. Ενα διάνυσμα $z \in R^n$ ανήκει στο <u>πεδίο τιμών</u> $\mathcal{R}(P_j)$ του πίνακα P_j αν και μόνο άν $z \in \langle p^1, \ldots, p^j \rangle$. Αν $x \in \mathcal{R}(P_k)$, τότε το x μπορεί να γραφτεί στην μορφή $x = P_{k-1}y + a p^k$ γιά κάποιο $y \in R^{k-1}$ και $a \in R$. 'Εχουμε επίσης τότε

$$(16) \quad \varphi(x) = \varphi(P_{k-1}y) + [a^2(Ap^k, p^k)/2 - a(p^k, b)] + a(P_{k-1}y, Ap^k).$$

Αν ο τελευταίος όρος του δευτέρου μέλους της (16) ήταν μηδέν, τότε το πρόβλημα της ελαχιστοποίησης του $\varphi(x)$ γιά $x\in\mathcal{R}(P_k)$ θα αναγόταν: (α) στην ελαχιστοποίηση του $\varphi(P_{k-1}y)$ γιά $y\in R^{k-1}$, δηλ. στην ελαχιστοποίηση του φ πάνω στον υπόχωρο $\langle p^1,\ldots,p^{k-1}\rangle$ - πρόβλημα που η λύση του x^{k-1} υποτίθεται ότι έχει βρεθεί - και (β) στην ελαχιστοποίηση του δευτέρου όρου του δευτέρου μέλους της (16), δηλ. σ' ένα πρόβλημα που λύνεται φυσικά όταν

$$(17)\quad \alpha = \alpha_k \equiv (p^k,b)/(Ap^k,p^k).$$

'Ενας προφανής τρόπος γιά να μηδενίσουμε τον τελευταίο όρο της (16) είναι να υπολογίσουμε το p^k, με δεδομένα τα p^i, $1\le i\le k-1$, έτσι ώστε

$$(18)\quad P_{k-1}^T A p^k = 0.$$

Αν ισχύει η (18), το x^{k-1} ελαχιστοποιεί το φ πάνω στον υπόχωρο $\langle p^1,\ldots,p^{k-1}\rangle$ και το α δίνεται από την (17), συμπεραίνουμε ότι το πρόβλημα (14) λύνεται, γιά $x=x^k\equiv x^{k-1}+\alpha_k p^k$. Επειδή $x^{k-1}\in\langle p^1,\ldots,p^{k-1}\rangle$ έχουμε εξ άλλου ότι $x^{k-1}=P_{k-1}z$ γιά κάποιο $z\in R^{k-1}$. 'Αρα από την (18) προκύπτει ότι $(p^k,Ax^{k-1})=(Ap^k,P_{k-1}z)=z^TP_{k-1}^TAp^k=0$, δηλ. ότι το α_k που δίδεται από την (17) ικανοποιεί και την σχέση

$$(19)\quad \alpha_k = (p^k,b-Ax^{k-1})/(Ap^k,p^k) = (p^k,r^{k-1})/(Ap^k,p^k),$$

δηλ. λύνει και το πρόβλημα της μονοδιάστατης ελαχιστοποίησης (15).

Συνοψίζουμε: αν το p^k επιλεχεί έτσι ώστε να ισχύει η (18), δηλ. έτσι ώστε να ισχύουν οι σχέσεις

$$(20)\quad (Ap^j,p^k) = 0,\ j=1,2,\ldots,k-1,$$

τότε ο αλγόριθμος (13) κατασκευάζει α_k και x^k που λύνουν τα προβλήματα (14) και (15). Αν η (20) ισχύει, τότε λέμε ότι το p^k είναι Α-συζυγές (ή Α-ορθογώνιο) πρός τα $p^1,\dots,p^{k-1}$.

Απομένει να εξετάσουμε τρία ζητήματα:

(α) Είναι τα p^i γραμμικά ανεξάρτητα;

(β) Ισχύει ότι $(p^k,r^{k-1})\neq 0$, έτσι ώστε, βλ. (12), να έχουμε ελάττωση του φ στο βήμα k;

(γ) Πώς υπολογίζουμε στην πράξη τα p^i;

Η απάντηση στο ερώτημα (α) είναι καταφατική: Αν τα διανύσματα $p^i\neq 0$, $1\leq i\leq m$ του R^n είναι Α-συζυγή, δηλ. αν

$$(21)\quad (Ap^i,p^j)=0, \text{ αν } i\neq j,$$

τότε τα p^i, $1\leq i\leq m$ είναι γραμμικά ανεξάρτητα. (Ασκ. 7(β)). Συνεπώς, αν κατασκευάσουμε τα $p^k\neq 0$ γιά $k=1,2,\dots,n$ έτσι ώστε να ισχύει η (20), τότε θα είναι γραμμικά ανεξάρτητα, πράγμα που εγγυάται ότι ο αλγόριθμος δίνει την ακριβή λύση του προβλήματος σε n βήματα, δηλ. ότι $x^n=A^{-1}b$.

Στο ερώτημα (β), αν γιά κάποιο k, $r^{k-1}=0$, τότε ο αλγόριθμος τερματίζει και έχουμε ότι $x^{k-1}=x=A^{-1}b$. Αν $r^{k-1}\neq 0$, τότε υπάρχει p^k που ικανοποιεί την (20) τέτοιο ώστε $(p^k,r^{k-1})\neq 0$. Πράγματι, αν γιά κάθε $p\in R^n$ που είναι Α-συζυγές πρός τα p^i, $1\leq i\leq k-1$, ισχύει $(p,r^{k-1})=0$, θα έχουμε, επειδή $r^{k-1}=b-Ax^{k-1}$ και $x^{k-1}\in\langle p^1,\dots,p^{k-1}\rangle$, ότι $(p,b)-(p,AP_{k-1}z)=0$ γιά κάποιο $z\in R^{k-1}$, δηλ. ότι $(p,b)=0$ γιά κάθε p Α-συζυγές πρός τα p^i, $1\leq i\leq k-1$. (Χρησιμοποιήσαμε την (18)). Συνεπώς $(p,b)=0$ γιά κάθε $p\in\langle Ap^1,Ap^2,\dots,Ap^{k-1}\rangle^{\perp}$ $\Leftrightarrow$ $b\in\langle Ap^1,\dots,Ap^{k-1}\rangle$ $\Leftrightarrow$ $A^{-1}b\in\langle p^1,\dots,p^{k-1}\rangle$ $\Leftrightarrow$ $x^{k-1}=x=A^{-1}b$ $\Leftrightarrow$ $r^{k-1}=0$, άτοπο. Συνεπώς μπορούμε να βρούμε p^k, Α-συζυγές πρός τα p^i, $1\leq i\leq k-1$, τέτοιο ώστε $(p^k,r^{k-1})\neq 0$, αν βέβαια $r^{k-1}\neq 0$.

Όπως παρατηρήσαμε και παραπάνω "p Α-συζυγές πρός τα p^i, $1\leq i\leq k-1$" $\Leftrightarrow$ $p\in\langle Ap^1,\dots,Ap^{k-1}\rangle^{\perp}$. Η λεγόμενη μέθοδος των συζυγών κλίσεων (conjugate gradients) των Hestenes και Stiefel (1952) επιλέγει ως p^k το πλησιέστερο πρός το r^{k-1} διάνυσμα του υπόχωρου

$\langle Ap^1,\ldots,Ap^{k-1}\rangle^{\perp}$, δηλ. κατασκευάζει τα x^k με βάση την εξής εξειδίκευση του γενικού προγράμματος(13) γιά τον υπολογισμό της λύσης x του Ax=b:

$$
\begin{array}{ll}
 & x^0=0 \\
(22) & \left[\begin{array}{l}
\text{Γιά } k=1,2,\ldots n \\
\quad r^{k-1}=b-Ax^{k-1} \\
\quad \text{Αν } r^{k-1}=0 \\
\qquad \text{τότε, εκχώρησε } x=x^{k-1} \text{ και τερμάτισε.} \\
\qquad \text{αλλοιώς, όρισε} \\
\qquad\quad p^k=\begin{cases} r^0, & \text{αν } k=1 \\ \text{ορθή προβολή του } r^{k-1} \text{ στον υπόχωρο} \\ \langle Ap^1,\ldots,Ap^{k-1}\rangle^{\perp}, & \text{αν } k>1 \end{cases} \\
\qquad\quad a_k=(p^k,r^{k-1})/(Ap^k,p^k) \\
\qquad\quad x^k=x^{k-1}+a_kp^k
\end{array}\right. \\
 & x=x^n
\end{array}
$$

Ουσιαστικά όμως δεν απαντήσαμε στο ερώτημα (γ), δηλ. στο πώς υπολογίζουμε τα p^k, k>1 με αποτελεσματικό τρόπο στην πράξη. Με το πρόβλημα αυτό καθώς και με το πρόβλημα της εκτίμησης του σφάλματος $e^j=x-x^j$ γιά κάθε j, θ' ασχοληθούμε στην επόμενη παράγραφο.

Ασκήσεις 1.5

1. (α) Επαληθεύετε την (6).

(β) Βρήτε αναδρομική σχέση μεταξύ των r^k,r^{k-1} που μας επιτρέπει να γράφουμε τον αλγόριθμο (7) έτσι ώστε να απαιτεί ένα μόνο πολλαπλασιασμό πίνακα επί διάνυσμα γιά κάθε k. Πόσες πράξεις απαιτεί ο νέος αλγόριθμος ανά βήμα k;

(γ) Δείξτε ότι δύο διαδοχικά υπόλοιπα της μεθόδου της καθόδου μεγίστης κλίσεως γιά την λύση του $Ax=b$ είναι ορθογώνια, δηλ. ότι $(r^j,r^{j-1})=0$, $j\geq 1$. Επίσης ότι τα διανύσματα $x^{j+1}-x^j$, x^j-x^{j-1}, $j\geq 1$ είναι ορθογώνια.

2. (α) Δείξτε ότι η παράσταση $(A\underline{x},y)$, $x,y \in R^n$ ορίζει ένα εσωτερικό γινόμενο στον R^n και ότι συνεπώς η συνάρτηση $x\mapsto (Ax,x)^{1/2}$ είναι νόρμα. Βρήτε τις (καλύτερες) σταθερές σύγκρισης μεταξύ αυτής της νόρμας και της $\|x\|_2$.

(β) Δείξτε ότι η (8') είναι ισοδύναμη με την
$\varphi(x^{j+1}) + (b,x)/2 \leq ((\kappa-1)/(\kappa+1))^2\ [\varphi(x^j) + (b,x)/2]$.

3. (α) Χρησιμοποιώντας τον συμβολισμό του Θεωρήματος 1, δείξτε ότι γιά $j\geq 0$
$(Ae^{j+1},e^{j+1}) = \{1-[(r^j,r^j)^2/(Ar^j,r^j)(A^{-1}r^j,r^j)]\}\ (Ae^j,e^j)$

(β) Δείξτε την <u>ανισότητα του Kantorovich:</u>
$(Ax,x)(A^{-1}x,x) \leq [(\lambda_{max}+\lambda_{min})^2/4\lambda_{max}\lambda_{min}]\ \|x\|^4$

(γ) Χρησιμοποιώντας την ανισότητα του Kantorovich και την ταυτότητα του ερωτήματος (α) δείξτε την (8') δίνοντας έτσι μία άλλη απόδειξη του Θεωρήματος 1.

4. Θεωρούμε το 2x2 σύστημα $Ax=b$ όπου $A = \begin{pmatrix} 1 & 0 \\ 0 & \lambda \end{pmatrix}$ με $\lambda>0$ και $x=b=0$.

(α) Αν $x^k = (x_1,x_2)^T$, δείξτε ότι η μέθοδος της καθόδου μεγίστης κλίσεως δίνει

$$x^{k+1} = [x_1x_2(\lambda-1)/(x_1^2+\lambda^3x_2^2)] \begin{pmatrix} \lambda^2x_2 \\ \\ -x_1 \end{pmatrix}$$

Σχεδιάστε γιά $\lambda=5$ τις ισοϋψείς καμπύλες του συναρτησιακού $\varphi(x)$ (δηλ. τις καμπύλες $\varphi(x_1,x_2)$=σταθ) και μερικές διαδοχικές προσεγγίσεις x^k, $k=1,2,3,\ldots$ με $x^0=(5,1)^T$.

(β) Δείξτε ότι αν $x^j=c(\lambda,\underline{+}1)^T$, τότε η ανισότητα (8') ισχύει ως ισότητα.

(γ) Αν $\lambda=100$, πόσα βήματα της μεθόδου θα χρειαστούν έτσι ώστε το σφάλμα $(Ae^j,e^j)^{1/2}$ να γίνει μικρότερο από $\varepsilon(Ae^0,e^0)^{1/2}$ όπου $\varepsilon>0$ δεδομένο;

5. Αποδείξτε τους ισχυρισμούς περί γεωμετρικής ερμηνείας της μεθόδου της καθόδου μεγίστης κλίσεως που περιλαμβάνονται στο κείμενο (Σελ 1.5.6) από "Οι "ισοϋψείς" επιφάνειες ..." μέχρι "... του ελλειψοειδούς που διέρχεται από το σημείο x^k".

6. Έστω M ένας υπόχωρος του R^n και έστω $y\in R^n$ ένα δεδομένο διάνυσμα. Δείξτε ότι το πρόβλημα της ελαχιστοποίησης του $\varphi(x)$ γιά $x\in y+M$ έχει μοναδική λύση.

7. (α) Επαληθεύστε τις (11), (12) και δείξτε ότι $(r^j,p^j)=0$, $j\geq 1$.

(β) Εστω ότι $0\neq p^i\in R^n$ $1\leq i\leq m$ και ότι ισχύει η (21), δηλ. ότι τα p^i είναι A-συζυγή. Δείξτε ότι είναι γραμμικά ανεξάρτητα.

1.6 Η ΜΕΘΟΔΟΣ ΤΩΝ ΣΥΖΥΓΩΝ ΚΛΙΣΕΩΝ

Ο σκοπός μας σ' αυτήν την παράγραφο θα είναι να μελετήσουμε λεπτομερώς τις ιδιότητες των x^k, r^k, p^k που κατασκευάζει ο αλγόριθμος (1.5.22), να εκτιμήσουμε τα σφάλματα $e^k = x - x^k$ και να βρούμε πρακτικούς και αποτελεσματικούς τρόπους γιά να υπολογίζουμε σε κάθε βήμα k την νέα κατεύθυνση ελαχιστοποίησης p^k που ορίζεται στον (1.5.22) ως η ορθή προβολή της κλίσεως (υπολοίπου) r^{k-1} πάνω στον υπόχωρο $\langle Ap^1, \dots, Ap^{k-1}\rangle^{\perp}$.

Θα αρχίσουμε αποδεικνύοντας μία σειρά χρησίμων ιδιοτήτων των υπολοίπων r^i και των συζυγών κατευθύνσεων p^i. Λέμε ότι "ο αλγόριθμος (1.5.22) ολοκληρώνει k βήματα" αν $r^i \neq 0$, $0 \leq i \leq k-1$. Τότε τα x^i, $1 \leq i \leq k$ θα έχουν υπολογισθεί επαγωγικά από τις σχέσεις: $x^0 = 0$ και

(1) $x^i = x^{i-1} + a_i p^i$, $i \geq 1$.

Λήμμα 1. Αν ο αλγόριθμος (1.5.22) ολοκληρώνει k βήματα, τότε γιά $i = 1, 2, \dots, k$ έχουμε

(2) $r^i = r^{i-1} - a_i A p^i$,

(3) $P_i^T r^i = 0$,

όπου P_i είναι ο nxi πίνακας $[p^1, \dots, p^i]$ με στήλες p^i.

<u>Απόδειξη</u>: Η (2) είναι προφανής απόρροια της (1). Η (3) μπορεί να προκύψει "αλγεβρικά" από την (1) και το γεγονός ότι τα p^i είναι A-συζυγή (βλ. 'Ασκηση 1.(α)) αλλά και ως εξής: Γνωρίζουμε ότι το πρόβλημα της ελαχιστοποίησης

(4) $$\min_{x \in \langle p^1, \dots, p^i \rangle} \varphi(x).$$

λύνεται μοναδικά γιά $x = x^i$. Εξ άλλου, επειδή $x \in \langle p^1, \dots, p^i \rangle \Leftrightarrow x \in \mathcal{R}(P_i)$, το (4) είναι ισοδύναμο με το πρόβλημα

$$(5) \quad \min_{y \in R^i} \varphi(P_i y),$$

το οποίο, επειδή $\varphi(P_i y)=(AP_i y,P_i y)/2 - (b,P_i y)$, λύνεται (γιατί;) γιά $y=y^i$, όπου y^i είναι η λύση του συστήματος $(P_i^T AP_i)y^i=P_i^T b$. 'Αρα από την σχέση $x^i=P_i y^i$ έπεται η $P_i^T Ax^i=P_i^T b \iff P_i^T r^i=0$. @

Λήμμα 2. Γιά $k \geq 2$, τα διανύσματα p^k που παράγονται από τον αλγόριθμο (1.5.22) (εφ' όσον ολοκληρώνει k βήματα) προσδιορίζονται από τις σχέσεις

$$(6) \quad p^k = r^{k-1}-AP_{k-1}z^{k-1},$$

όπου το $z^{k-1} \in R^{k-1}$ είναι η (μοναδική) λύση του προβλήματος ελαχιστοποίησης

$$(7) \quad \min_{z \in R^{k-1}} \|r^{k-1} - AP_{k-1}z\|.$$

<u>Απόδειξη</u>: 'Εξ ορισμού το p^k είναι η (ορθή) προβολή του r^{k-1} στον υπόχωρο $\mathcal{R}(AP_{k-1})^{\perp}$. Συνεπώς το $r^{k-1}-p^k$ είναι η ορθή προβολή του r^{k-1} στον υπόχωρο $\mathcal{R}(AP_{k-1})$. (Γιά επανάληψη περί προβολών βλ. 'Ασκηση 2). Συνεπώς $r^{k-1}-p^k = AP_{k-1}z^{k-1}$ γιά $z^{k-1} \in R^{k-1}$. Η μοναδικότητα του z^{k-1} είναι τώρα απόρροια της γραμμικής ανεξαρτησίας των p^i. Επιπλέον το z^{k-1} λύνει το πρόβλημα της ελαχιστοποίησης (ελαχίστων τετραγώνων)

$$(\|r^{k-1}-(r^{k-1}-p^k)\| = \|p^k\| =) \|r^{k-1}AP_{k-1}z^{k-1}\| =$$
$$= \min_{u\in \mathcal{R}(AP_{k-1})} \|r^{k-1}-u\| = \min_{z\in R^{k-1}} \|r^{k-1}-AP_{k-1}z\|. \text{ @}$$

Θα χρησιμοποιήσουμε τα παραπάνω δύο λήμματα γιά να αποδείξουμε τα εξής σημαντικό θεωρητικό αποτέλεσμα:

ΠΡΟΤΑΣΗ 1 Αν ο αλγόριθμος (1.5.22) ολοκληρώνει k βήματα, τότε:

(α) Τα υπόλοιπα $r^0,r^1,\ldots,r^{k-1}$ είναι ορθογώνια μεταξύ τους.

(β) Γιά $j=1,2,\ldots,k$ ισχύει

$$(8)\ \langle p^1,\ldots,p^j\rangle = \langle r^0,\ldots,p^{j-1}\rangle = \langle b,Ab,\ldots,A^{j-1}b\rangle.$$

<u>Απόδειξη:</u>

(α) Από την (2) γιά $1\le i\le k-1$ έχουμε ότι $Ap^i=(r_i-r_{i-1})/a_i$ (γιατί $a_i\neq 0$;). Άρα $Ap^i\in\langle r^0,\ldots,r^i\rangle$. Συνεπώς η (6) δίνει ότι $p^j=r^{j-1}-[Ap^1,\ldots,Ap^{j-1}]z^{j-1}\in\langle r^0,\ldots,r^{j-1}\rangle$ γιά $1\le j\le k$. Έπεται λοιπόν ότι γιά $1\le j\le k$ $p^j=c_{jo}r^0+\ldots+c_{j,j-1}r^{j-1}$, όπου $c_{j,j-1}\neq 0$ γιατί τα p^j είναι γραμμικά ανεξάρτητα (ως A-συζυγή). Άρα, εισάγοντας του nxk πίνακα $R_k=[r^0,\ldots,r^{k-1}]$ έχουμε $P_k=R_kC$, όπου C ένας kxk αντιστρέψιμος, άνω τριγωνικός πίνακας. Συνεπώς $R_k=P_kC^{-1}$ και επειδή ο C^{-1} είναι άνω τριγωνικός συμπεραίνουμε ότι γιά $1\le j\le k$, $r^{j-1}\in\langle p^1,\ldots,p^j\rangle$. Από την (3) έχουμε όμως ότι $(p^j,r^i)=0$, $1\le j\le i\le k-1$. Από τις δύο αυτές σχέσεις έχουμε ότι $(r^j,r^i)=0$ γιά $0\le j\le i\le k-1$, ό.έ.δ.

(β) Η (8) αποδεικνύεται με επαγωγή ως πρός j ως εξής: Επειδή $p^1=r^0=b$, ισχύει γιά j=1. Έστω τώρα ότι ισχύει γιά κάποιο j, $1\le j\le k-1$. Τότε η (2) συνεπάγεται ότι $r^j=r^{j-1}-a_jAp^j\in\langle b,Ab,\ldots,A^jb\rangle$.

Εξ άλλου από την (6) έχουμε τότε ότι $p^{j+1}=r^j-A(z_1^jp^1+\ldots+z_j^jp^j)$ $\in\langle b,Ab,\ldots,A^jb\rangle$. Συνεπώς και οι δύο υπόχωροι $\langle r^0,\ldots,r^j\rangle$ και $\langle p^1,\ldots,p^{j+1}\rangle$ (και οι δύο διαστάσεως j+1 - γιατί;) περιέχονται στον υπόχωρο $\langle b,Ab,\ldots,A^jb\rangle$, ο οποίος έχει συνεπώς διάσταση j+1. Άρα η (8) ισχύει γιά j+1. (Χρησιμοποιήσαμε βεβαίως το γεγονός ότι $r^i\neq 0$, $0\le i\le k-1$). @

Τα μέχρι τώρα θεωρητικά αποτελέσματά μας γιά την μέθοδο των συζυγών κλίσεων αρκούν γιά να εκτιμήσουμε το σφάλμα $x-x^k$ σε κάθε βήμα. Κατ' αρχήν μιά παρατήρηση: μπορούμε εύκολα να γενικεύσουμε τηύ μέθοδο γιά οποιαδήποτε αρχική τιμή $x^0 \neq 0$. Γιά οποιοδήποτε $x^0 \in R^n$ ορίζουμε στον αλγόριθμο (1.5.22) $r^0=b-Ax^0$ και τα υπόλοιπα p^k, a^k, x^k, r^k όπως πρίν. Στην γενική περίπτωση $x^0 \neq 0$ όμως, ορισμένα προφανή πράγματα αλλάζουν στις αποδείξεις. Π.χ. το x^j λύνει τώρα το πρόβλημα ελαχιστοποίησης

$$(4') \quad \min_{y \in x^0+\langle p^1,\ldots,p^j\rangle} \varphi(x).$$

Οι σχέσεις (1), (2), (3) εξακολουθούν να ισχύουν όπως επίσης και το συμπέρασμα (α) της Πρότασης 1. Η (8) πρέπει να αντικατασταθεί από την

$$(8') \quad \langle p^1,\ldots,p^j\rangle=\langle r^0,\ldots,r^{j-1}\rangle=\langle r^0,Ar^0,\ldots,A^{j-1}r^0\rangle, \quad 1\leq j\leq k.$$

ΘΕΩΡΗΜΑ 1. 'Εστω $\{x^j\}$, $j\geq 0$ η ακολουθία που παράγει η μέθοδος των συζυγών κλίσεων με οποιαδήποτε αρχική τιμή $x^0 \in R^n$, έστω x η λύση του συστήματος (1.5.1) και $e^j=x-x^j$. 'Εστω $\kappa=\lambda_{max}/\lambda_{min}$ όπου λ_{max}, αντιστχ. λ_{min}, είναι η μέγιστη , αντιστχ. ελάχιστη, ιδιοτιμή του A. Γιά $j\geq 1$ έχουμε τότε ότι

$$(9) \quad (Ae^j,e^j)^{1/2} \leq 2[(\kappa^{1/2}-1)/(\kappa^{1/2}+1)]^j \, (Ae^0,e^0)^{1/2}.$$

<u>Απόδειξη</u>: Κατ' αρχήν μία προκαταρτική παρατήρηση: γιά οποιοδήποτε $y \in R^n$ έχουμε $(A(x-y),x-y)=2\varphi(y)+(Ax,x)$. Συνεπώς το πρόβλημα ελαχιστοποίησης $\min_{y\in S}\varphi(y)$ γιά $S \subset R^n$ είναι ισοδύναμο με το πρόβλημα ελαχιστοποίησης $\min_{y\in S}(A(x-y),x-y)$. Συμπεραίνουμε, επειδή το x^j είναι η (μοναδική) λύση του (4'), ότι

$$(10)\ (Ae^j,e^j) = (A(x-x^j),x-x^j) = \min_{y\in x^0+\langle p^1,\dots,p^j\rangle} (A(x-y),x-y)$$

$$= \min_{z\in\langle p^1,\dots,p^j\rangle} (A(e^0+z),e^0+z).$$

Λόγω της (8') έχουμε ότι $z\in\langle p^1,\dots,p^j\rangle \Leftrightarrow z\in\langle r^0,Ar^0,\dots,A^{j-1}r^0\rangle \Leftrightarrow$ $z = \pi_{j-1}(A)r^0$ γιά κάποιο πολυώνυμο $\pi_{j-1}\in \mathcal{P}_{j-1}$, όπου με $\mathcal{P}_k$ θα συμβολίζουμε τον χώρο των πραγματικών πολυωνύμων βαθμού το πολύ k. Επειδή $r^0=Ae^0$, συμπεραίνουμε από τον (10) ότι

$$(11)\ (Ae^j,e^j) = \min_{\pi\in \mathcal{P}_{j-1}} ((I+A\pi(A))^2\, Ae^0,e^0).$$

(Υποθέσαμε έμμεσα ότι $j\geq 1$ και ότι ο αλγόριθμος της μεθόδου των συζυγών κλίσεων ολοκληρώνει j βήματα. Η σχέση (11) ισχύει και γιά $j=0$, αν ορίσουμε $\mathcal{P}_{-1}=\{0\}$.) Χρησιμοποιώντας τώρα την φασματική παράσταση του A παίρνουμε,με ανάλογες πράξεις με εκείνες της απόδειξης του Θεωρήματος 1.5.1. ότι

$$(Ae^j,e^j) \leq \min_{\pi\in \mathcal{P}_{j-1}} (\max_{\lambda\in[\lambda_{min},\lambda_{max}]} (1+\lambda\pi(\lambda))^2)(Ae^0,e^0).$$

Συνεπώς έχουμε τελικά ότι

$$(12)\quad (Ae^j,e^j)^{1/2} \leq \sigma_j\, (Ae^0,e^0)^{1/2},\ j\geq 0$$

όπου γιά $j\geq 0$ ($\sigma_0=1$)

$$(13)\ \sigma_j = \min_{\pi\in \mathcal{P}_j,\ \pi(0)=1} (\max_{\lambda\in[\lambda_{min},\lambda_{max}]} |\pi(\lambda)|).$$

Θα αποδείξουμε στο Λήμμα 4 πιό κάτω ότι

(14) $\epsilon_j \leq 2[(\kappa^{1/2}-1)/(\kappa^{1/2}+1)]^j, j\geq 1$.

Από τις (12) και (14) έπεται αμέσως η (9). @

Η λύση του προβλήματος "min-max" (12) είναι κλασική και γίνεται με χρήση πολυωνύμων Chebyshev: Τα πολυώνυμα Chebyshev $T_j(z)$, βαθμού $j\geq 0$, μιάς πραγματική μεταβλητής z, ορίζονται ως γνωστόν από τις αναδρομικές σχέσεις

$$\text{(14)}\quad \begin{aligned} T_o(z) &= 1 \\ T_1(z) &= z \\ T_j(z) &= 2zT_{j-1}(z) - T_{j-2}(z),\ j\geq 2. \end{aligned}$$

Από τον ορισμό αυτό μπορούμε εύκολα να αποδείξουμε επαγωγικά τις εξής ιδιότητες των πολυωνύμων Chebyshev $T_m(z)$, $m\geq 0$:

(α) βαθμ $(T_m(z))=m$, $T_m(1)=1$, $T_m(z)$ περιττή, αντιστχ. άρτια, συνάρτηση του z αν m περιττός, αντιστχ. άρτιος.

$$(\beta)\quad T_m(z) = \begin{cases} \cos(m\cos^{-1}z) & \text{αν } -1\leq z\leq 1 \\ \cosh(m\cosh^{-1}z) & \text{αν } z\geq 1, \end{cases}$$

δηλ.

(β_1) $T_m(z)=\cos(m\vartheta)$, όπου $z=\cos\vartheta$, $\vartheta\in[0,\pi]$, αν $-1\leq z\leq 1$.

(β_2) $T_m(z)=\cosh(mu)$, όπου $z=\cosh u$, $u\geq 0$, αν $z\geq 1$.

(γ) $T_m(z) = [(z+(z^2-1)^{1/2})^m + (z-(z^2-1)^{1/2})^m]/2$.

Τα πολυώνυμα Chebyshev είναι χρήσιμα στην περίπτωσή μας λόγω της εξής σημαντικής ιδιότητάς τους:

Λήμμα 4. 'Εστω $0<\alpha<\beta$. Τότε το πρόβλημα min-max

$$(16)\quad \min_{\pi\in\mathcal{P}_m,\ \pi(0)=1} \left(\max_{\alpha\le z\le\beta} |\pi(z)|\right)$$

λύεται μοναδικά από το πολυώνυμο

$$(17)\quad \tilde{\pi}_m(z)=T_m((\beta+\alpha-2z)/(\beta-\alpha))/T_m((\beta+\alpha)/\beta-\alpha)),$$

γιά το οποίο

$$(18)\quad \max_{\alpha\le z\le\beta} |\tilde{\pi}_m(z)|=1/T_m((\beta+\alpha)/(\beta-\alpha)),$$

Απόδειξη: Από την παράσταση (β_1) των πολυωνύμων Chebyshev γιά $-1\le y\le 1$ έχουμε γιά $m\ge 0$ ότι $\max_{-1\le y\le 1} |T_m(y)|=1$ και ότι η μέγιστη αυτή απόλυτη τιμή λαμβάνεται στά σημεία $y_j=\cos(j\pi/m)$, $j=0,1,\dots,m$, όπου $T_m(y_j)=(-1)^j$, δηλ. όπου το $T_m(y)$ έχει εναλλασσόμενο πρόσημο. Θεωρούμε τώρα τον γραμμικό μετασχηματισμό $z\mapsto y$, $y=(\beta+\alpha-2z)/(\beta-\alpha)$ που απεικονίζει αμφιμονοσήμαντα το διάστημα $\alpha\le z\le\beta$ πάνω στο $-1\le y\le 1$. Γιά $\alpha\le z\le\beta$ θεωρούμε το πολυώνυμο $\tilde{\pi}_m(z)$ που δίνεται από την (17). 'Εχουμε $\tilde{\pi}_m(0)=1$. Επίσης από τα παραπάνω, $\max_{\alpha\le z\le\beta} |\tilde{\pi}_m(z)| =$
$= \max_{-1\le y\le 1} |T_m(y)|/T_m((\beta+\alpha)/(\beta-\alpha))= 1/T_m((\beta+\alpha)/(\beta-\alpha))$, δηλ. ισχύει η (18). (Επειδή $(\beta+\alpha)/(\beta-\alpha)>1$ χρησιμοποιήθηκε η (β_2)). Επίσης, η μέγιστη αυτή τιμή του $|\tilde{\pi}_m(z)|$ γιά $\alpha\le z\le\beta$ λαμβάνεται σε $m+1$ διακριτά σημεία z_j, $0\le j\le m$ του $[a,b]$, (στις προεικόνες των y_j), όπου το $\tilde{\pi}_m$ έχει εναλλασσόμενο πρόσημο. 'Εστω τώρα ότι υπάρχει άλλο πραγματικό πολυώνυμο $q_m(z)$, βαθμού $\le m$, τέτοιο ώστε $q_m(0)=1$ και $\max_{\alpha\le z\le\beta} |q_m(z)| <. \max_{\alpha\le z\le\beta} |\tilde{\pi}_m(z)|$. Θεωρούμε το πολυώνυμο $p_m(z)=\tilde{\pi}_m(z)-q_m(z)$, βαθμού $\le m$. Το $p_m(z)$ έχει εναλλασσόμενο πρόσημο στά σημεία z_j, $0\le j\le m$ και συνεπώς υπάρχουν m διακριτά σημεία s_j, $1\le j\le m$,

τέτοια ώστε $0<\alpha\le z_{j-1}<s_j<z_j\le\beta$, όπου $p_m(s_j)=0$. Επίσης $p_m(0)=\tilde{\pi}_m(0)-q_m(0)=0$, δηλ. το πολυώνυμο $p_m(z)$ έχει $m+1$ διακριτές ρίζες => $p_m(z)=0$, άτοπο. Συνεπώς το πολυώνυμο $\tilde{\pi}_m(z)$ λύνει το πρόβλημα min-max (16). Η μοναδικότητα της λύσης του (16) αφήνεται ως άσκηση). @

Με την βοήθεια του λήμματος αυτού μπορούμε να αποδείξουμε τώρα την ισχύ των (14).

Λήμμα 4. Αν οι αριθμοί ϵ_j ορίζονται από την (13), τότε ισχύει η (14).

<u>Απόδειξη</u>: Από την (13) και το Λήμμα 3 συμπεραίνουμε ότι $\epsilon_j=1/T_j((\lambda_{max}+\lambda_{min})/(\lambda_{max}-\lambda_{min}))$, αν $\lambda_{max}\neq\lambda_{min}$. (Η (13) δίνει ότι $\epsilon_j=0$ αν $\lambda_{min}=\lambda_{max}$, $j\ge 1$). Συνεπώς για $\kappa\neq 1$, $j\ge 1$ έχουμε

$$\epsilon_j = 1/T_j((\kappa+1)/(\kappa-1)).$$

Χρησιμοποιώντας την παράσταση (γ) των πολυωνύμων Chebyshev γιά $z=(\kappa+1)/(\kappa-1)$, παίρνουμε μετά από λίγες πράξεις ότι

$$(20)\quad \epsilon_j = f(\kappa^{-1/2})\,[(\kappa^{1/2}-1)/(\kappa^{1/2}+1)]^j,\ j\ge 1$$

όπου γιά $x\in[0,1]$ η συνάρτηση $f(x)$ ορίζεται ως

$$f(x) = 2(1+x)^{2j}/[(1+x)^{2j}+(1-x)^{2j}],$$

γιά την οποία ισχύει $1=f(0)\le f(x)\le f(1)=2$, $x\in[0,1]$. Συνεπώς, λόγω της (20) και του ότι $\kappa\ge 1$, μιά εκτίμηση του ϵ_j δίνεται από την (14). @

Ας έρθουμε τώρα στο πρόβλημα της αποτελεσματικής υλοποίησης του αλγορίθμου (1.5.22) στην πράξη. Θα στηριχθούμε στα αποτελέσματα των Λημμάτων 1 και 2 και της Πρότασης 1 και στο εξής σημαντικό πόρισμα.

τους: <u>Γιά $k\geq 2$ το διάνυσμα p^k είναι γραμμικός συνδυασμός των p^{k-1} και r^{k-1}</u>. Γιά $k=2$ αυτό είναι προφανές λόγω της (8). Αν $k>2$, υποθέτοντας πάντα ότι $r^{k-1}\neq 0$, θεωρούμε το διάνυσμα $z^{k-1}\in R^{k-1}$ της (6) το οποίο γράφουμε ως $z^{k-1}=(w,\mu)^T$, όπου $w\in R^{k-2}$ και $\mu\in R^1$ ($\mu\neq 0$· γιατί;) Χρησιμοποιώντας τώρα την (6) και την (2) γιά $i=k-1$ έχουμε

$$(21)\quad p^k = (1+(\mu/a_{k-1}))r^{k-1}+s^{k-1},$$

όπου

$$(22)\quad s^{k-1} = -\mu r^{k-2}/a_{k-1} - AP_{k-2}w.$$

Η (22) δίνει τώρα ότι $(s^{k-1}, r^{k-1})=0$, επειδή τα r^i είναι ορθογώνια και επειδή, λόγω της (8), $AP_{k-2}w = A(w_1p^1+\ldots+w_{k-2}p^{k-2}) \in \langle Ab,A^2b,\ldots,A^{k-2}b\rangle \subset \langle r^0,\ldots,r^{k-2}\rangle$. Συνεπώς, το Πυθαγόρειο Θεώρημα και η (21) δίνουν

$$(23)\quad \|p^k\|^2 = (1+(\mu/a_{k-1}))^2\,\|r^{k-1}\|^2 + \|s^{k-1}\|^2.$$

Θυμόμαστε τώρα τον χαρακτηρισμό του $z^{k-1} = (w,\mu)^T$ ως εκείνου του στοιχείου του R^{k-1} που λύνει το πρόβλημα ελαχίστων τετραγώνων (7)· συμπεραίνουμε λοιπόν ότι η κατασκευή του z^{k-1} είναι ισοδύναμη με τον καθορισμό εκείνων των w,μ που ελαχιστοποιούν το $\|p^k\|(=\|r^{k-1}-AP_{k-1}z^{k-1}\|$, βλ. (6)), δηλ. που ελαχιστοποιούν το δεύτερο μέλος της (23). Επειδή, από την (22) η ποσότης

$$\|s^{k-1}\|^2 = (\mu/a_{k-1})^2\,\|r^{k-2} - AP_{k-2}\,w'\|^2,\quad w'=a_{k-1}w/\mu$$

ελαχιστοποιείται ως πρός w' γιά $r^{k-2} - AP_{k-2}w' = p^{k-1}$. (βλ. Λήμμα 2), θα πρέπει το s^{k-1} να είναι πολλαπλάσιο του p^{k-1}. Η (21) τώρα αποδεικνύει τον ισχυρισμό.

Συνεπώς $p^k \in \langle p^{k-1}, r^{k-1}\rangle$ γιά $k \geq 2$, δηλ. η ευθεία $x^{k-1}+\alpha p^k$, $\alpha \in \mathbf{R}$, που διέρχεται από το x^{k-1} και πάνω στην οποία ελαχιστοποιείται το $\varphi(x)$ γιά $x=x^k$, βρίσκεται στο επίπεδο $x^{k-1}+\langle p^{k-1}, r^{k-1}\rangle$ η τομή του οποίου με το ελλειψοειδές $\varphi(x)=\varphi(x^{k-1})$ (δηλ. με την "ισοϋψή" επιφάνεια του φ που διέρχεται από το σημείο x^{k-1}) είναι μία έλλειψη C_k που διέρχεται φυσικά από το x^{k-1}, όπου εφάπτεται στην ευθεία $x^{k-1}+\alpha p^{k-1}$ (αφού το x^{k-1} ήταν ακρότατο του $\varphi(x)$ πάνω στην ευθεία αυτή) και όπου έχει κάθετο παράλληλη πρός το r^{k-1}. Οι ελλείψεις - τομές των ισοϋψών $\varphi(x)$=σταθ. του φ με το επίπεδο $x^{k-1}+\langle p^{k-1}, r^{k-1}\rangle$ είναι ομόκεντρες και ομοιόθετες της C_k. 'Αρα το ελάχιστο του $\varphi(x)$ πάνω σ' αυτό το επίπεδο λαμβάνεται στο σημείο x^k που είναι το κοινό κέντρο των ελλείψεων αυτών. Συνεπώς οι κατευθύνσεις p^k και p^{k-1} είναι "συζυγείς" ως πρός την έλλειψη C_k, εξ ου και η ονομασία τους "A-συζυγείς" (επειδή ικανοποιούν την (1.5.21)) είναι συμβιβαστή με την γνωστή από την επίπεδη γεωμετρία έννοια. Γιά περισσότερα πάνω στην γεωμετρική ερμηνεία της μεθόδου των συζυγών κλίσεων (καθώς και γιά την πληρέστερη ανάλυση των κλασικών ιδιοτήτων των μεθόδων ελαχιστοποίησης), βλ. το βιβλίο του M.R. Hestenes, "Conjugate direction methods in optimization", Springer-Verlag, Berlin 1980.

Επανερχόμαστε στο θέμα της υλοποίησης του αλγορίθμου (1.5.22). Χωρίς περιορισμό της γενικότητας (βλ. Άσκ. 5(γ)) υποθέτουμε ότι

$$(24)\quad p^k = r^{k-1} + \beta_k p^{k-1},\ k \geq 2.$$

Επειδή τα p^k, p^{k-1} είναι A-συζυγή η (24) δίνει

$$(25)\quad \beta_k = -(p^{k-1}, Ar^{k-1})/(Ap^{k-1}, p^{k-1}).$$

Οι τύποι (24) και (25) ορίζουν πλήρως το p^k. Εξ άλλου επειδή $(p^{k-1}, r^{k-1})=0$, η (24) και η (1.5.19) δίνουν ότι

$$(26)\quad \alpha_k = \|r^{k-1}\|^2 / (Ap^k, p^k).$$

Οδηγούμαστε λοιπόν στην εξής υλοποίηση του αλγορίθμου (1.5.22):

(27)

$x^0=0$

Γιά $k=1,2,\ldots,n$

$\quad r^{k-1}=b-Ax^{k-1}$

$\quad$ Αν $r^{k-1}=0$

$\quad\quad$ τότε, εκχώρησε $x=x^{k-1}$ και τερμάτισε.

$\quad\quad$ αλλοιώς,

$\quad\quad\quad$ αν $k=1$

$\quad\quad\quad\quad$ τότε, $p^k=r^0$

$\quad\quad\quad\quad$ αλλοιώς,

$\quad\quad\quad\quad\quad \beta_k=-(p^{k-1},Ar^{k-1})/(Ap^{k-1},p^{k-1})$

$\quad\quad\quad\quad\quad p^k=r^{k-1}+\beta_k p^{k-1}$

$\quad\quad\quad a_k=\|r^{k-1}\|^2/(Ap^k,p^k)$

$\quad\quad\quad x^k=x^{k-1}+a_k p^k$

$x=x^n$

Ο αλγόριθμος απαιτεί δύο πολλαπλασιασμούς του πίνακα A επί διάνυσμα σε κάθε βήμα k. Παρατηρώντας όμως ότι τα υπόλοιπα μπορούν να υπολογισθούν αναδρομικά μέσω της $r^j=r^{j-1}-a_j\ Ap^j$, έχουμε - χρησιμοποιώντας την σχέση αυτή γιά $j=k-1$ και την ορθογωνιότητα των r^{k-1}, r^{k-2} - ότι $\|r^{k-1}\|^2=-a_{k-1}(r^{k-1},\ Ap^{k-1})$ και $\|r^{k-2}\|^2=a_{k-1}(r^{k-2},Ap^{k-1})$ $=a_{k-1}(p^{k-1},Ap^{k-1})$ - λόγω της $p^{k-1}=r^{k-2}+\beta_{k-1}p^{k-2}$ -. Συνεπώς $\beta_k=-(p^{k-1},Ar^{k-1})/(Ap^{k-1},p^{k-1})=\|r^{k-1}\|^2/\|r^{k-2}\|^2$, αποφεύγοντας έτσι τον πολλαπλασιασμό Ar^{k-1}. Καταλήγουμε λοιπόν στην εξής τελική μορφή του αλγορίθμου της μεθόδου των συζυγών κλίσεων όπως ουσιαστικά τον παρουσίασαν οι Hestenes και Stiefel (1952):

$x^0=0$

$r^0=b$

Γιά $k=1,2,\ldots,n$

  Αν $r^{k-1}=0$

    τότε, εκχώρησε $x=x^{k-1}$ τερμάτισε.

    αλλοιώς

      αν $k=1$

(28)   τότε, $p^k=r^0$

        αλλοιώς,

          $\beta_k=\|r^{k-1}\|^2/\|r^{k-2}\|^2$

          $p_k=r^{k-1}+\beta_k p^{k-1}$

      $\alpha_k=\|r^{k-1}\|^2/(Ap^k,p^k)$

      $x^k=x^{k-1}+\alpha_k p^k$

      $r^k=r^{k-1}-\alpha^k Ap^k$

$x=x^n$

Παρατηρήσεις

1. Η απόδειξη της βασικής ιδιότητας της μεθόδου των συζυγών κλίσεων, ότι δηλ. βρίσκει την λύση $x^n=x$ σε n βήματα (ισοδύναμης με την ορθογωνιότητα των υπολοίπων $r^0,r^1,\ldots,r^{n-1}$, οπότε $r^n=0$) έγινε με την προϋπόθεση βέβαια ότι όλοι οι υπολογισμοί, π.χ. στον αλγόριθμο (27), γίνονται *ακριβώς*. Στην πράξη όμως (όπως έγινε γρήγορα σαφές μετά την ανακάλυψη της μεθόδου) τα *σφάλματα στρογγύλευσης* που οφείλονται στην αριθμητική πεπερασμένης ακρίβειας *καταστρέφουν την ορθογωνιότητα των r^j* που υπολογίζουμε, έτσι ώστε τελικά το r^n να μην είναι μηδέν. Μάλιστα, όπως περιμένουμε, όσο αυξάνει ο δείκτης του A, τόσο το φαινόμενο αυτό, δηλ. ότι $(r^i,r^j)\neq 0$, γίνεται πιό έντονο. Εν τούτοις, η μέθοδος των συζυγών κλίσεων, σαν μέθοδος *ελαχιστοποίησης* που είναι, όντως ελαττώνει την τιμή του συναρτησιακού $\varphi(x)$ από βήμα σε βήμα. Συνεπώς μπορεί να χρησιμοποιηθεί (και έτσι χρησιμοποιείται σήμερα στην πράξη) σαν *επαναληπτική μέθοδος* που παράγει προσεγγίσεις

$x^0, x^1, \ldots, x^k, \ldots$ της λύσης x του συστήματος $Ax=b'$ η επανάληψη δεν σταματά στο βήμα n αλλά τερματίζεται όταν ικανοποιηθούν κάποια από τα γνωστά μας κριτήρια τερματισμού. Γιά περισσότερες λεπτομέρειες βλ. [5.4, Παρ.16]. Θα θέλαμε όμως απλώς να παρατηρήσουμε εδώ ότι επειδή η κύρια και επαναλαμβανόμενη σε κάθε βήμα πράξη του αλγορίθμου είναι ο πολλαπλασιασμός του πίνακα A επί διάνυσμα, η μέθοδος των συζυγών κλίσεων είναι ιδιαίτερα αποτελεσματική γιά <u>αραιούς</u> πίνακες A. Γιά τέτοιους πίνακες υπάρχουν δομές δεδομένων γιά την αποθήκευση των μη μηδενικών στοιχείων τους ιδιαίτερα κατάλληλες γιά την πράξη του πολλαπλασιασμού πίνακα επί διάνυσμα, βλ. [5.4, παρ. 16]

2. Στην πράξη λοιπόν χρησιμοποιούμε την μέθοδο συζυγών κλίσεων σαν επαναληπτική. Μας ενδιαφέρει συνεπώς η ταχύτητα σύγκλισης (δηλ. ελάττωσης του σφάλματος), που μας δίνει π.χ. η (9) ιδίως γιά μεγάλο κ. Αν και η ταχύτητα σύγκλισης της μεθόδου των συζυγών κλίσεων είναι μεγαλύτερη της ταχύτητας της μεθόδου της καθόδου μεγίστης κλίσεως (γιατί $(\kappa-1)/(\kappa+1) = 1-2\kappa^{-1}+O(\kappa^{-2})$ ενώ $(\kappa^{1/2}-1)/(\kappa^{1/2}+1) = 1-2\kappa^{-1/2}+O(\kappa^{-1})$ όταν $\kappa\to\infty$, βλ. (1.5.8), (9), εξακολουθεί να είναι πολύ μικρή γιά μεγάλο κ. Μιά σημαντική τεχνική που χρησιμοποιείται γιά την επιτάχυνση της σύγκλισης είναι η λεγόμενη <u>προρύθμιση</u> (preconditioning), γιά την υλοποίηση της οποίας ο αλγόριθμος απαιτεί σε κάθε βήμα του εκτός ενός πολλαπλασιασμού του πίνακα A επί διάνυσμα και την λύση (με την μέθοδο Cholesky π.χ.) ενός nxn γραμμικού συστήματος με ένα συμμετρικό θετικά ορισμένο πίνακα, τον λεγόμενο <u>προρυθμιστή</u> (preconditioner) M. Η επιλογή κατάλληλου προρυθμιστή είναι σημαντικό πρόβλημα: πρέπει να έχει απλή δομή έτσι ώστε η λύση συστημάτων με πίνακα M να μην απαιτεί πολλές πράξεις· επιπλέον πρέπει ο πίνακας $\tilde{A}=M^{-1}A$ να έχει λόγο ιδιοτιμών $\tilde{\lambda}_{max}/\tilde{\lambda}_{min}$ μικρότερο του $\kappa=\lambda_{max}/\lambda_{min}$ ώστε να επιταχύνεται πράγματι η σύγκλιση. Βλέπε [5.4, Παρ. 16] γιά περισσότερα επ' αυτού.

Ασκήσεις 1.6

1. (α) Χρησιμοποιώντας την (1) και τις (1.5.11), (1.5.21), δείξτε ότι $(p^j, r^i)=0$, $j=1,2,\ldots,i$, δηλ. ότι ισχύει η (3), υπό τις προϋποθέσεις του Λήμματος 1.

(β) Απαντήσετε στά "γιατί;" των αποδείξεων του Λήμματος 1 και της Πρότασης 1.

2. Έστω $r \in R^n$ και M υπόχωρος του R^n. Λέμε ότι το $p \in M$ είναι η (ορθή) προβολή του r στον M (ή προβολή ως πρός το Ευκλείδιο εσωτερικό γινόμενο $(\cdot,\cdot)$) αν

$$(r,x) = (p,x), \qquad x \in M.$$

(α) Δείξτε ότι η προβολή p του r στον M υπάρχει, είναι μοναδική και ικανοποιεί $\|p\|^2=\|r\|^2 - \|r-p\|^2$.

(β) Δείξτε ότι το $r-p$ είναι η προβολή του r στον υπόχωρο $M^{\perp}$.

(γ) Δείξτε ότι $\|r-p\|=\min_{x \in M} \|r-x\|$ και ότι η ιδιότητα αυτή χαρακτηρίζει μονοσήμαντα την προβολή p του r.

(δ) Δείξτε ότι γιά κάθε $B \in R^{m \times n}$, $\mathcal{R}(B)^{\perp} = \mathcal{N}(B^T)$, όπου $\mathcal{N}(B^T)=\{x \in R^m : B^T x=0\}$.

(ε) Τι σημαίνει το αποτέλεσμα (δ) στην περίπτωση του Λήμματος 2, όπου $p^k \in \mathcal{R}(AP_{k-1})^{\perp}$;

3. Αναπτύξτε την μέθοδο των συζυγών κλίσεων γιά $x^0 \neq 0$ (το x^j λύνει τώρα το πρόβλημα (4')). Βρήτε τα ανάλογα των Λημμάτων 1 και 2 και της Πρότασης 1.

4. (α). Αποδείξτε τις ιδιότητες (α), (β), (γ) των πολυωνύμων Chebyshev.

(β). Δείξτε ότι το πολυώνυμο $\tilde{\pi}_m(z)$ που ορίζεται από την (17) είναι η <u>μοναδική</u> λύση του προβλήματος min-max (16).

5. (α). Γιατί στην αρχή της απόδειξης του ισχυρισμού ότι το $p^k \epsilon < p^{k-1}, r^{k-1} >$ (σελ. 1.6.9) ισχύει ότι $\mu \neq 0$;

(β) Στην ίδια απόδειξη - στο τελευταίο βήμα της, σελ. 1.6.9 - γιατί αγνοήσαμε το μ στο πρόβλημα της ελαχιστοποίησης του δευτέρου μέλους της (23) και μελετήσαμε μόνο την ελαχιστοποίηση ως πρός ω';

(γ) Δείξτε ότι η υπόθεση (24) δεν αποτελεί περιορισμό της γενικότητας, δηλ. δείξτε ότι ένας γενικός γραμμικός συνδυασμός της μορφής $p^k = \gamma_k r^{k-1} + \beta_k p^{k-1}$ οδηγεί, μέσω π.χ. του αλγορίθμου (27) στον υπολογισμό των ίδιων x^k, όπως και προηγουμένως.

2. ΑΡΙΘΜΗΤΙΚΗ ΛΥΣΗ ΜΗ ΓΡΑΜΜΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

2.1 ΠΑΡΑΓΩΓΙΣΙΜΕΣ ΣΥΝΑΡΤΗΣΕΙΣ ΣΤΟΝ R^n

Στο κεφάλαιο αυτό θα ασχοληθούμε με μεθόδους γιά την αριθμητική επίλυση μη γραμμικών συστημάτων n εξισώσεων με n αγνώστους, δηλ. συστημάτων της μορφής

(1) $f_i(x_1,\ldots,x_n) = 0,\ i=1,2,\ldots,n,$

τα οποία γράφουμε συνήθως στην διανυσματική μορφή

(1') $F(x) = 0,$

όπου F είναι μία (μη γραμμική εν γένει) απεικόνιση ενός υποσυνόλου D του R^n στον R^n (θα γράφουμε $F:D \subset R^n \to R^n$) με συνιστώσες $F(x)=(f_1(x),\ldots,f_n(x))^T$ και όπου $x=(x_1,\ldots,x_n)^T$ κατά τα γνωστά.

Προβλήματα που οδηγούν στην επίλυση μη γραμμικών συστημάτων της μορφής (1) εμφανίζονται πολύ συχνά στις εφαρμογές. Μιά σημαντική πηγή προβλημάτων είναι π.χ. ο υπολογισμός τοπικών ακροτάτων ενός συναρτησιακού $g: R^n \to R^1$, οπότε $F=\nabla g$, υπό την προϋπόθεση βέβαια ότι η κλίση ∇g υπάρχει και μπορεί να υπολογισθεί γιά κάθε x. Θα εξετάσουμε όμως το πρόβλημα στην γενική μορφή (1) χωρίς να υπεισέλθουμε εδώ σε ειδικές μεθόδους γιά προβλήματα βελτιστοποίησης.

Σημαντικό ρόλο τόσο στην θεωρία όσο και στις αριθμητικές μεθόδους γιά την λύση του (1) παίζουν οι ιδιότητες παραγωγισιμότητας της F καθώς και διαφόρου τύπου θεωρήματα "μέσης τιμής"· τα θεωρήματα αυτά θα εξετάσουμε σ' αυτήν την εισαγωγική παράγραφο.

Λέμε ότι η απεικόνιση $F:D \subset R^n \to R^n$ είναι παραγωγίσιμη σ' ένα σημείο $x\in \mathrm{Int}D$ (ακριβέστερα παραγωγίσιμη με την έννοια του Frechet ή F - παραγωγίσιμη) αν υπάρχει γραμμικός τελεστής $A_x:R^n \to R^n$ τέτοιος ώστε γιά κάποια νόρμα $\|\cdot\|$ του R^n να ισχύει

(2) $\lim_{h\to 0} \|F(x+h)-F(x) - A_x h\|/\|h\| = 0$

λόγω της ισοδυναμίας των νορμών στον $\mathbb{R}^n$ η ύπαρξη του A_x είναι ανεξάρτητη της νόρμας $\|\cdot\|$. Είναι επίσης προφανές ότι ο ορισμός (2) γενικεύει την έννοια της παραγωγισιμότητας παραγματικών συναρτήσεων μιάς μεταβλητής.

Αν η F είναι παραγωγίσιμη στο x τότε ο γραμμικός τελεστής A_x είναι <u>μοναδικός</u>. Πράγματι αν υπήρχαν δύο γραμμικοί τελεστές A_1, A_2 τέτοιοι ώστε γιά τον κάθε ένα να ισχύει η (2) θα είχαμε από την τριγωνική ανισότητα ότι γιά κάθε $0 \neq y \in \mathbb{R}^n$, $t \neq 0$

$$\|(A_1 - A_2)y\|/\|y\| \leq (\|F(x+ty) - F(x) - A_1(ty)\|/\|ty\|)$$
$$+ (\|F(x+ty) - F(x) - A_2(ty)\|/\|ty\|),$$

από την οποία, θέτοντας $h = ty$ και παίρνοντας $t \to 0$ έχουμε λόγω της (2) ότι $A_1 y = A_2 y \ \ \forall\, y \in \mathbb{R}^n$, δηλ. ότι $A_1 = A_2$. Αν λοιπόν η F είναι παραγωγίσιμη στο x λέμε ότι ο τελεστής A_x είναι η <u>παράγωγος</u> της F (ακριβέστερα ή "παράγωγος της F με την έννοια του Frechet" ή η "F-παράγωγος της F") στο σημείο x και συμβολίζουμε $A_x = F'(x)$. Γενικά λέμε ότι η $F: D \subset \mathbb{R}^n \to \mathbb{R}^n$ είναι παραγωγίσιμη σ' ένα σύνολο $D_0 \subset D$ αν $D_0 \subset \mathrm{Int}(D)$ και αν η F είναι παραγωγίσιμη $\forall\, x \in D_0$. Τότε η F' μπορεί να θεωρηθεί ως απεικόνιση του D_0 στο σύνολο $L(\mathbb{R}^n)$ των γραμμικών τελεστών από τον $\mathbb{R}^n$ στον εαυτό του. Εύκολα επίσης μπορούμε να αποδείξουμε την γραμμικότητα της πράξης της παραγώγισης: Έστω ότι οι απεικονίσεις $F_1, F_2: D \subset \mathbb{R}^n \to \mathbb{R}^n$ είναι παραγωγίσιμες στο $x \in \mathrm{Int}(D)$. Τότε γιά $\alpha, \beta \in \mathbb{R}$, η $\alpha F_1 + \beta F_2$ είναι παραγωγίσιμη στο x και $(\alpha F_1 + \beta F_2)'(x) = \alpha F_1'(x) + \beta F_2'(x)$.

Αν η παράγωγος $F'(x)$ υπάρχει στο σημείο $x \in \mathrm{Int}(D)$ τότε υπάρχουν όλες οι μερικές παράγωγοι $\partial_j f_i (= \partial f_i / \partial x_j)$, $1 \leq i, j \leq n$ των συνιστωσών f_i της F στο σημείο x, ο δέ $n \times n$ πίνακας που παριστάνει τον γραμμικό τελεστή $F'(x)$ ως πρός την κανονική βάση $\{e^j\}$, $1 \leq j \leq n$, του $\mathbb{R}^n$ $(e^j_i = \delta_{ij})$ είναι ο <u>Ιακωβιανός πίνακας</u> $J(x)$: $J_{ij}(x) = \partial_j f_i(x)$, $1 \leq i, j \leq n$.

Πράγματι, θέτοντας, γιά $j=1,\dots,n$, $h=te^j$ στην (2) (με $\|\cdot\|=\|\cdot\|_2$ π.χ.) και υποθέτοντας ότι στο σημείο x η $F'(x)=A_x$ παριστάνεται ως πρός την βάση $\{e^j\}$ από τον πίνακα (a_{ij}) έχουμε γιά $1\leq i,\ j\leq n$

$$\lim_{t\to 0} |(f_i(x+te^j) - f_i(x))/t - a_{ij}| = 0,$$

δηλ. ότι όντως $a_{ij} = \partial_j f_i(x) = J_{ij}(x)$.

Η ύπαρξη μόνο των μερικών παραγώγων $\partial_j f_i(x)$ δεν εγγυάται όμως ότι η F είναι παραγωγίσιμη στο x. Αυτό φαίνεται αμέσως από την εξής σημαντική συνέπεια της παραγωγισιμότητας:

ΠΡΟΤΑΣΗ 1. 'Εστω ότι η $F : D \subset R^n \to R^n$ είναι παραγωγίσιμη στο σημείο $x\in \mathrm{Int}(D)$. Τότε η F είναι συνεχής στο x.

Απόδειξη: Επειδή $x\in\mathrm{Int}(D)$, $\exists\, \delta_1>0$ τέτοιο ώστε $x+h\in D$ αν $\|h\|<\delta_1$. Η (2) συνεπάγεται τώρα ότι γιά δεδομένο $\epsilon>0$ υπάρχει $\delta>0$, το οποίο μπορούμε να πάρουμε $\leq\delta_1$, τέτοιο ώστε $\|F(x+h)-F(x)-F'(x)h\| \leq \epsilon\|h\|$ αν $\|h\|\leq\delta$, από την οποία έπεται ότι $\|F(x+h) - F(x)\| \leq (\|F'(x)\|+\epsilon)\|h\|$. Σταθεροποιώντας το ϵ λοιπόν συμπεραίνουμε ότι γιά δεδομένο $x\in\mathrm{Int}(D)$ $\exists\, \delta>0$ και $c\geq 0$ τέτοια ώστε $x+h\in D$ και $\|F(x+h)-F(x)\|\leq c\|h\|$ αν $\|h\|\leq\delta$, που είναι μάλιστα ένα συμπέρασμα ισχυρότερο από την συνέχεια της F στο x. ©

Θα διερευνήσουμε σχέσεις μεταξύ της ύπαρξης και της συνέχειας της $F'(x)$ και των ανάλογων ιδιοτήτων του Ιακωβιανού πίνακα $J(x)$ καθώς και μιάς "ασθενέστερης" παραγώγου της F, της λεγόμενης παραγώγου Gateaux, σε μιά σειρά παρατηρήσεων και ασκήσεων στο τέλος της παραγράφου. Πρός το παρόν θα συνεχίσουμε με την μελέτη ορισμένων θεωρημάτων "μέσης τιμής".

Το γνωστό μας θεώρημα μέσης τιμής γιά παραγωγίσιμες συναρτήσεις $f:R^1 \to R^1$, ότι δηλ. $\forall\, x,y\in R^1$, $f(x)-f(y)=f'(z)(x-y)$ γιά κάποιο z μεταξύ των x και y, δεν ισχύει γιά απεικονίσεις $F:R^n\to R^n$ αν $n\geq 2$ (Βλ. Ασκ. 7(β)). Υπάρχουν όμως εναλλακτικά αποτελέσματα του τύπου "μέσης τιμής" πολύ χρήσιμα στην μη γραμμική ανάλυση. Παραδείγματος χάριν πολλές φορές ενδιαφερόμαστε απλώς να φράξουμε την ποσότητα $\|F(x)-F(y)\|$ συναρτήσει της F':

ΠΡΟΤΑΣΗ 2. Υποθέτουμε ότι η $F:D \subset R^n \longrightarrow R^n$ είναι παραγωγίσιμη σ' ένα <u>κυρτό</u> σύνολο $D_0 \subset D$. Τότε αν $x,y \in D_0$

$$(3)\ \|F(x)-F(y)\| \leq \sup_{0\leq t\leq 1} \|F'(x+t(y-x))\|\ \|x-y\|.$$

<u>Απόδειξη</u>: Εξ υποθέσεως $x+t(y-x)\in D_0$ γιά $t\in[0,1]$. 'Εστω ότι $M=\sup_{0\leq t\leq 1} \|F'(x+t(y-x))\|<\infty$. Γιά δεδομένο $\epsilon>0$, έστω Γ_ϵ το σύνολο των $t\in[0,1]$ τέτοιων ώστε

$$(4)\ \|F(x+t(y-x)) - F(x)\| \leq Mt\|y-x\| + \epsilon t\|x-y\|.$$

Προφανώς το Γ_ϵ δεν είναι κενό γιατί $0\in\Gamma_\epsilon$. 'Εστω $\gamma_\epsilon = \sup_{t\in\Gamma_\epsilon} t$. Τότε $0\leq\gamma_\epsilon\leq 1$ και επειδή λόγω της Πρότασης 1 η συνάρτηση $t\mapsto F(x+t(y-x))$ είναι συνεχής στο $[0,1]$, παίρνοντας το όριο στην (4) μιάς ακολουθίας $t_i\to\gamma_\epsilon$ σημείων του Γ_ϵ, έχουμε

$$(5)\ \|F(x+\gamma_\epsilon(y-x))-F(x)\| \leq M\gamma_\epsilon\ \|y-x\| + \epsilon\gamma_\epsilon\ \|x-y\|.$$

Αν γιά κάθε $\epsilon>0$ $\gamma_\epsilon=1$, τότε η (5) δίνει την ζητούμενη ανισότητα (4). Αν γιά κάποιο $\epsilon>0$, $0\leq\gamma_\epsilon<1$ επειδή η F' υπάρχει στο σημείο $x+\gamma_\epsilon(y-x)$ έχουμε, από τον ορισμό της παραγώγου (2), παίρνοντας ως h ένα κατάλληλο (μικρό) πολλαπλάσιο του $y-x$, ότι υπάρχει $\beta_\epsilon\in(\gamma_\epsilon,1)$ τέτοιο ώστε

$$\|F(x+\beta_\epsilon(y-x))-F(x+\gamma_\epsilon(y-x))-F'(x+\gamma_\epsilon(y-x))(\beta_\epsilon-\gamma_\epsilon)(y-x)\|\leq\epsilon(\beta_\epsilon-\gamma_\epsilon)\|y-x\|,$$

από την οποία έπεται ότι

$$(6)\ \|F(x+\beta_\varepsilon(y-x))-F(x+\gamma_\varepsilon(y-x))\| \le M(\beta_\varepsilon-\gamma_\varepsilon)\|y-x\|+\varepsilon(\beta_\varepsilon-\gamma_\varepsilon)\|y-x\|.$$

Οι (5) και (6) δίνουν τότε μέσω της τριγωνικής ανισότητας ότι

$$\|F(x+\beta_\varepsilon(y-x))-F(x)\| \le M\beta_\varepsilon \|y-x\| + \varepsilon\beta_\varepsilon\|y-x\|,$$

δηλ. ότι γιά αυτό το ε η (4) ισχύει γιά κάποιο β_ε: $\gamma_\varepsilon<\beta_\varepsilon<1$, πράγμα που αντιφάσκει στον ορισμό του γ_ε. Συνεπώς $\gamma_\varepsilon=1$ $\forall \varepsilon>0$ και το αποτέλεσμα προκύπτει όπως παραπάνω. @

Ενός άλλου τύπου αποτελέσματα είναι "ολοκληρωτικές" μορφές του θεωρήματος μέσης τιμής. Κατά τα γνωστά, αν η διανυσματική συνάρτηση μιάς μεταβλητής $G:[a,b]\to R^n$ έχει συνιστώσες $G=(g_1,\ldots,g_n)^T$, λέμε ότι η G είναι ολοκληρώσιμη (με την έννοια του Riemann) αν και μόνο αν γιά κάθε i οι συναρτήσεις $g_i:[a,b]\to R^1$ είναι ολοκληρώσιμες κατά Riemann στο [a,b] οπότε και ορίζουμε

$$\int_a^b G(t)dt = \left(\int_a^b g_1(t)dt,\ldots, \int_a^b g_n(t)dt\right)^T.$$

ΠΡΟΤΑΣΗ 3. Υποθέτουμε ότι η $F:D \subset R^n \to R^n$ είναι συνεχώς παραγωγίσιμη σ' ένα κυρτό σύνολο $D_0 \subset D$. Τότε, αν $x,y\in D_0$

$$(7)\ F(y)-F(x) = \int_0^1 F'(x+t(y-x))(y-x)dt.$$

<u>Απόδειξη</u> Επειδή η F' είναι συνεχής στο D_0, τότε, γιά $x,y\in D_0$ η συνάρτηση $t \mapsto F'(x+t(y-x))$ είναι συνεχής στο [0,1]. Συνεπώς οι διανυσματικές συναρτήσεις $\nabla f_i(x+t(y-x))$, $1\le i\le n$ (γραμμές του Ιακωβιανού πίνακα που παριστάνει την $F'(x+t(y-x))$ είναι συνεχείς συναρτήσεις του t γιά $t\in[0,1]$ και συνεπώς ολοκληρώσιμες κατά Riemann στο [0,1]. Επειδή τώρα γιά $x,y\in D_0$

$$df_i(x+t(y-x))/dt = (\nabla f_i\ (x+t(y-x)))^T\ (y-x),$$

έχουμε, ολοκληρώνοντας ως πρός t από 0 έως 1, ότι γιά $1 \leq i \leq n$:

$$f_i(y)-f_i(x) = \int_0^1 \nabla f_i(x+t(y-x))^T\ (y-x)dt,$$

που είναι ακριβώς η (7) γραμμένη κατά συνιστώσες. @

Θα χρησιμοποιήσουμε στην συνέχεια μιά ενδιαφέρουσα εφαρμογή της Πρότασης 3. Πρώτα ένα εύκολο λήμμα:

ΛΗΜΜΑ 1. Έστω ότι η $G:[a,b] \to R^n$ είναι συνεχής στο $[a,b]$. Τότε

$$(8)\quad \left\| \int_a^b G(t)dt \right\| \leq \int_a^b \|G(t)\|dt.$$

<u>Απόδειξη</u>: Η συνάρτηση $t \mapsto \|G(t)\|$ είναι συνεχής στο $[a,b]$ λόγω της υπόθεσής μας και της συνέχειας της $x \mapsto \|x\|$. Συνεπώς η $t \mapsto \|G(t)\|$ είναι ολοκληρώσιμη κατά Riemann στο $[a,b]$. Λόγω της ολοκληρωσιμότητας και της $G(t)$ έχουμε ότι $\forall \varepsilon>0$ υπάρχει διαμερισμός $a \leq t_0 < t_1 < \ldots < t_s \leq b$ τέτοιος ώστε

$$\left\| \int_a^b G(t)dt - \Sigma_{j=1}^{s} G(t_j)\ (t_j-t_{j-1}) \right\| \leq \varepsilon$$

και

$$\left| \int_a^b \|G(t)\|dt - \Sigma_{j=1}^{s} \|G(t_j)\|\ (t_j-t_{j-1}) \right| \leq \varepsilon.$$

Συνεπώς από την τριγωνική ανισότητα και τις δύο αυτές σχέσεις έπεται ότι

$$\left\|\int_a^b G(t)dt\right\| \le \left\|\sum_{j=1}^{s} G(t_j)(t_j-t_{j-1})\right\|+\varepsilon \le \sum_{j=1}^{s} \|G(t_j)\|(t_j-t_{j-1})$$

$$+\ \varepsilon \le \int_a^b \|G(t)\|\,dt + 2\varepsilon.$$

Επειδή το $\varepsilon>0$ ήταν αυθαίρετο, έπεται η (8). @

ΠΡΟΤΑΣΗ 4. Υποθέτουμε ότι η $F:D \subset \mathbb{R}^n \to \mathbb{R}^n$ είναι συνεχώς παραγωγίσιμη σ' ένα κυρτό σύνολο $D_0 \subset D$ και ότι επιπλέον υπάρχουν σταθερές $\alpha, p \ge 0$ τέτοιες ώστε

(9) $\|F'(u) - F'(\upsilon)\| \le \alpha \|u-\upsilon\|^p, \quad u,\upsilon \in D_0.$

Τότε γιά κάθε $x,y \in D_0$

(10) $\|F(y) - F(x) - F'(x)(y-x)\| \le \alpha \|x-y\|^{p+1}/(p+1).$

<u>Απόδειξη</u>: Η F ικανοποιεί τις υποθέσεις της Πρότασης 3. Συνεπώς γιά $x,y \in D_0$

$$F(y)-F(x) = \int_0^1 F'(x+t(y-x))(y-x)dt.$$

Άρα λόγω της (8) και της (9)

$$\|F(y)-F(x)-F'(x)(y-x)\| = \left\|\int_0^1 [F'(x+t(y-x))-F'(x)](y-x)dt\right\|$$

$$\le \|x-y\| \int_0^1 \|F'(x+t(y-x))-F'(x)\|dt \le \alpha\|x-y\|^{p+1} \int_0^1 t^p dt$$

$$= \alpha\|x-y\|^{p+1}/(p+1).\ @$$

<u>Παρατηρήσεις</u>

1. Τα αποτελέσματα αυτής της παραγράφου εύκολα γενικεύονται σε συστήματα m εξισώσεων με n αγνώστους, σε απεικονίσεις δηλ. F της μορφής $F:D \subset R^n \to R^m$, $F = (f_1, f_2, \ldots, f_m)^T$, $x=(x_1,\ldots,x_n)^T$. Ο ορισμός της παραγώγου (2) είναι ο ίδιος (αν με το ίδιο σύμβολο $\|\cdot\|$ παραστήσουμε δύο οποιεσδήποτε νόρμες στον R^n και στον R^m). Η παράγωγος $F'(x)$ είναι γιά κάθε $x\in D$, γραμμικός τελεστής από τον R^n στον R^m (γράφουμε $F'(x) \in L(R^n,R^m)$), και παριστάνεται πάλι - ως πρός τις κανονικές βάσεις του R^n και του R^m - από τον mxn Ιακωβιανό πίνακα $J_{ij}(x) = \partial_j f_i(x)$, $1\leq i\leq m$, $1\leq j\leq n$. Στην ειδική περίπτωση ενός <u>συναρτησιακού</u>, δηλ. μιάς απεικόνισης $g:D \subset R^n \to R^1$, η παράγωγος $g'(x)$ είναι γραμμικός τελεστής από τον R^n στον R^1 και παριστάνεται από το γραμμοδιάνυσμα $(\partial_1 g(x),\ldots,\partial_n g(x))$, δηλ. από το γραμμοδιάνυσμα $\nabla g(x)$. Είναι φανερό ότι οι Προτάσεις 1-4 ισχύουν γιά απεικονίσεις $F:D\subset R^n \quad R^m$, mutatis mutandis.

2. Γιά πολλές εφαρμογές η έννοια της παραγώγου Frechet που ορίσαμε σ' αυτήν την παράγραφο είναι πιό ισχυρή απ' ό,τι χρειάζεται. Μία ασθενέστερη έννοια είναι η λεγόμενη <u>παράγωγος με την έννοια του Gateaux</u> που ορίζεται ως εξής: Λέμε ότι μία απεικόνιση $F:D \subset R^n \to R^m$ είναι παραγωγίσιμη σ' ένα σημείο $x\in \mathrm{Int}(D)$ με την έννοια του Gateaux (ή G-παραγωγίσιμη) αν υπάρχει γραμμικός τελεστής $A_x' \in L(R^n,R^m)$ τέτοιος ώστε γιά κάθε $h\in R^n$ να ισχύει

$$(11) \qquad \lim_{t\to 0} \|F(x+th)-F(x)-tA_x h\|/t=0,$$

δηλ. αν είναι κατά κάποιο τρόπο "παραγωγίσιμη" σε κάθε κατεύθυνση h. Όπως και προηγουμένως, η ύπαρξη του τελεστή A_x είναι ανεξάρτητη της νόρμας $\|\cdot\|$ και ο A_x είναι μοναδικός. Ορίζουμε λοιπόν γιά μία τέτοια συνάρτηση την παραγωγό της $F'(x)$ με την έννοια του Gateaux (ή G-παράγωγο) ως $F'(x)=A_x$. Είναι προφανές ότι η παραγώγιση Gateaux είναι γραμμική πράξη και ότι αν υπάρχει η G-παράγωγος $F'(x)$, τότε υπάρχουν οι μερικές παράγωγοι $\partial_j f_i(x)$, $1 \leq i \leq m$, $1 \leq j \leq n$, η δε $F'(x)$ παριστάνεται από τον Ιακωβιανό πίνακα $J_{ij} = \partial_j f_i(x)$. Είναι επίσης προφανές ότι αν η F είναι παραγωγίσιμη (κατά Frechet) και έχει (F-) παράγωγο $F'(x)$, τότε είναι παραγωγίσιμη κατά Gateaux και η G-παράγωγός της στο x συμπίπτει με την $F'(x)$. Το αντίστροφο όμως δεν είναι αληθινό: Παραγωγισιμότητα κατά Gateaux δεν συνεπάγεται παραγωγισιμότητα κατά Frechet (βλ. Ασκ. 3). Επίσης παραγωγισιμότητα κατά Gateaux σ' ένα σημείο x δεν συνεπάγεται συνέχεια της F στο x σε αντίθεση με ό,τι αποδείξαμε στην Πρόταση 1 για την παραγωγισιμότητα Frechet, (βλ. Ασκ. 3). Πάντως εύκολα βλέπουμε ότι οι Προτάσεις 2,3 και 4 ισχύουν, αν στις υποθέσεις τους η $F'(x)$ είναι η παράγωγος κατά Gateaux.

Μιά σειρά ασκήσεων (Ασκ. 2-6) στο τέλος της παραγράφου διερευνά σχέσεις μεταξύ των εννοιών της παραγώγισης Frechet, Gateaux και της συνηθισμένης (μερικής) παραγώγισης ως πρός x_i.

3. Οι ορισμοί και τα περισσότερα αποτελέσματα αυτής της παραγράφου ισχύουν γενικά και γιά απεικονίσεις μεταξύ χώρων Banach. 'Εστω $F: D \subset X \to Y$ όπου X, Y χώροι Banach. Τότε οι παράγωγοι $F'(x)$ κατά Frechet ή κατά Gateaux ορίζονται ως (φραγμένοι) γραμμικοί τελεστές $A_x: X \to Y$ από τις (2) και (11), αντίστοιχα. Οι Προτάσεις 1-4 εξακολουθούν να ισχύουν' βέβαια το ολοκλήρωμα μιάς συνάρτησης $G:[a,b]\ \ X$ δεν ορίζεται πιά μέσω των συνιστωσών της G.

Ασκήσεις 2.1

1. Με συντομία επιβεβαιώστε τους ισχυρισμούς εκείνους της παρατήρησης 2 σχετικά με την παράγωγο Gateaux γιά τους οποίους δεν υπάρχει ειδική παραπομπή σε άλλη άσκηση.

2. (α) Θεωρείστε την συνάρτηση $f:\mathbb{R}^2 \to \mathbb{R}^1$ δίνεται από

$$f(x_1,x_2) = \begin{cases} x_1 & \text{αν } x_2=0 \\ x_2 & \text{αν } x_1=0 \\ 1 & \text{αλλοιώς} \end{cases} .$$

Δείξτε ότι οι μερικές παράγωγοι $\partial_1 f(0)$ και $\partial_2 f(0)$ υπάρχουν αλλά ότι η f δεν είναι συνεχής και δεν είναι παραγωγίσιμη κατά Gateaux (και συνεπώς ούτε και κατά Frechet) στο 0.

(β) Θεωρείστε την συνάρτηση $f: \mathbb{R}^2 \to \mathbb{R}^1$ που δίνεται από

$$f(x_1,x_2) = \begin{cases} 0 & \text{αν } x=0 \\ x_1x_2^2/(x_1^2+x_2^4) & \text{αν } x\neq 0. \end{cases}$$

Δείξτε ότι το όριο $\lim_{t\to 0} [f(th)-f(0)]/t$ υπάρχει γιά κάθε $h\in\mathbb{R}^2$ αλλά ότι η f δεν είναι παραγωγίσιμη κατά Gateaux (ούτε συνεχής) στο 0.

3. Θεωρείστε την συνάρτηση $f:\mathbb{R}^2 \to \mathbb{R}^1$ που δίνεται από

$$f(x_1,x_2) = \begin{cases} 0 & \text{αν } x_1=0 \\ 2x_2 \exp(-x_1^{-2})/(x_2^2 + \exp(-2x_1^{-2})) & \text{αν } x_1\neq 0. \end{cases}$$

Δείξτε ότι έχει παράγωγο κατά Gateaux στο 0 αλλά ότι δεν είναι συνεχής (και συνεπώς ούτε παραγωγίσιμη κατά Frechet) στό 0.

4. Αν η $F: D \subset \mathbb{R}^n \to \mathbb{R}^m$ είναι παραγωγίσιμη κατά Gateaux στο κυρτό σύνολο $D_0 \subset D$ δείξτε ότι γιά $x,y,z \in D_0$

$$\|F(y)-F(z)-F'(x)(y-z)\| \le \sup_{0\le t\le 1} \|F'(z+t(y-z))-F'(x)\| \, \|y-z\|$$

(Υπόδειξη: Θεωρείστε την απεικόνιση $G(u)=F(u)-F'(x)u$, $u\in D$ και εφαρμόζοντας την Πρόταση 2 - γιά παραγώγους Gateaux και γιά απεικονίσεις από τον R^n στον R^m - στην G, φράξτε το $\|G(y) - G(z)\|$).

5. Υποθέστε ότι η $F:D \subseteq R^n \to R^m$ έχει παράγωγο κατά Gateaux F' σε κάθε σημείο μιάς ανοικτής μπάλας με κέντρο x και ότι η F' είναι συνεχής στο x. Τότε η F είναι παραγωγίσιμη (κατά Frechet) στο x και οι δύο παράγωγοι συμπίπτουν. (Υπόδειξη: χρησιμοποιείστε την Άσκηση 4).

6. Υποθέστε ότι η $F: D \subset R^n \to R^m$ έχει παράγωγο κατά Gateaux F' σε κάθε σημείο μιάς ανοικτής μπάλας με κέντρο x. Τότε η F' είναι συνεχής στο x αν και μόνο αν όλες οι μερικές παράγωγοι $\partial_j f_i$ είναι συνεχείς στο x.

7. (α) Το θεώρημα της μέσης τιμής στη συνηθισμένη του μορφή ισχύει γιά συναρτησιακά: Υποθέστε ότι η $g:D \subset R^n \to R^1$ έχει παράγωγο g' κατά Gateaux σε κάθε σημείο ενός κυρτού συνόλου $D_0 \subset D$. Τότε αν $x,y\in D_0$, υπάρχει $t\in(0,1)$ τέτοιο ώστε $g(y)-g(x)=g'(x+t(y-x))(y-x)$.

(β) Θεωρείστε την απεικόνιση $F:R^2 \to R^2$ με συνιστώσες $f_1(x)=x_1^3$, $f_2(x)=x_2^2$. Αν $x=0$ και $y=(1,1)^T$ δείξτε ότι δεν υπάρχει z της μορφής $x+t(y-x)$, $t\in[0,1]$ τέτοιο ώστε

$$(12)\quad F(y) - F(x) = F'(z)(y-x)$$

(γ) Έστω $F:R^n \to R^n$ μία διαγώνια απεικόνιση (δηλ. μία απεικόνιση $F=(f_1,\dots,f_n)^T$ τέτοια ώστε η f_i να είναι συνάρτηση μόνο της μεταβλητής x_i). Υποθέστε ότι η F είναι παραγωγίσιμη κατά Gateaux στον R^n. Δείξτε ότι ισχύει η (12) αλλά με z όχι αναγκαστικά στο ευθύγραμμο τμήμα $x+t(y-x)$, $0\le t\le 1$.

8. (Σε ασκήσεις επομένων παραγράφων θα αναφερθούμε συχνά στις υποθέσεις αυτής της άσκησης).

Στην θεωρία των μη γραμμικών ταλαντώσεων συναντάμε συχνά το εξής πρόβλημα: Ζητάμε μία πραγματική απεικόνιση $u(x)$, $x\in[0,1]$, δυό φορές συνεχώς παραγωγίσιμη στο $[0,1]$ που να ικανοποιεί το συνοριακό πρόβλημα "δύο σημείων"

$$(*)\begin{cases} u''(x)=g(u(x)), & 0\leq x\leq 1 \\ u(0)=\alpha, \ u(1)=\beta, \end{cases}$$

όπου $g:\mathbb{R}^1\to\mathbb{R}^1$ μία δεδομένη δύο φορές συνεχώς παραγωγίσιμη συνάρτηση και α,β δεδομένες πραγματικές σταθερές. Είναι γνωστό ότι αν η g ικανοποιεί την

$$(13) \quad g'(s)\geq \underline{g} > -\pi^2, \quad \forall s\in\mathbb{R}^1,$$

τότε υπάρχει μοναδική λύση $u(x)$ του (*).

Γιά να επιλύσουμε προσεγγιστικά το πρόβλημα (*) με μιά μέθοδο πεπερασμένων διαφορών, εισάγουμε τον ομοιόμορφο διαμερισμό $x_i=ih$, $0\leq i\leq n+1$ όπου $(n+1)h=1$, $n\in\mathbb{N}$. Ζητάμε να κατασκευάσουμε προσεγγίσεις U_i, $0\leq i\leq n+1$, των τιμών της λύσης $u(x_i)$, $0\leq i\leq n+1$, που ορίζονται ως εξής:

$$(14) \quad \begin{cases} (U_{i-1}-2U_i+U_{i+1})/h^2 = g(U_i), & 1\leq i\leq n, \\ U_0=\alpha, \ U_{n+1}=\beta. \end{cases}$$

Δηλ. το διάνυσμα $U=(U_1,\ldots,U_n)^T$ των αγνώστων ικανοποιεί το $n\times n$ μη γραμμικό σύστημα

$$(15) \quad F(U) = 0,$$

όπου $F(U)=[f_1(U),\ldots,f_n(U)]^T$ και όπου οι συναρτήσεις $f_i:\mathbb{R}^n\to\mathbb{R}^1$ ορίζονται ως

$$(16)\quad \begin{cases} f_1(U) = -\alpha + 2U_1 - U_2 + h^2 g(U_1) \\ f_i(U) = -U_{i-1} + 2U_i - U_{i+1} + h^2 g(U_i),\ 2\le i\le n-1 \\ f_n(U) = -U_{n-1} + 2U_n - \beta + h^2 g(U_n). \end{cases}$$

Εισάγοντας τον τριδιαγώνιο πίνακα

$$(17)\quad A = [-1,2,-1]$$

και την διαγώνια απεικόνιση $\Phi:R^n \longrightarrow R^n$ που ορίζεται γιά $x=(x_1,\dots,x_n)^T \in R^n$ ως

$$(18)\quad \Phi(x)=h^2(g(x_1)-\alpha h^{-2},\ g(x_2),\dots,g(x_{n-1}),g(x_n)-\beta h^{-2})^T,$$

βλέπουμε ότι το σύστημα (15) γράφεται και στην μορφή

$$(19)\quad F(U) \equiv AU + \Phi(U) = 0.$$

<u>Ερώτημα</u>: Υποθέστε ότι η $g':R^1 \to R^1$ είναι συνεχής. Δείξτε ότι η απεικόνιση $F:R^n \longrightarrow R^n$ είναι συνεχώς παραγωγίσιμη (κατά Frechet) στον R^n και ότι

$$(20)\quad F'(x) = A + \Phi'(x),\ x\in R^n$$

όπου η Φ' είναι η παράγωγος (Frechet) της Φ που παριστάνεται από τον <u>διαγώνιο</u> Ιακωβιανό πίνακα

$$(21)\quad J_\phi(x) = h^2 \mathrm{diag}(g'(x_1),\dots,g'(x_n)),\quad x\in R^n.$$

2.2 ΤΟΠΙΚΑ ΘΕΩΡΗΜΑΤΑ ΣΥΓΚΛΙΣΗΣ. ΤΟ ΘΕΩΡΗΜΑ ΤΗΣ ΣΥΣΤΟΛΗΣ

Θα ασχοληθούμε στη συνέχεια με την κατασκευή προσεγγίσεων των λύσεων του μη γραμμικού συστήματος

(1) $F(x)=0$,

όπου $F:D\subset R^n \to R^n$ είναι μία δεδομένη απεικόνιση. Οι αριθμητικές μας μέθοδοι θα είναι επαναληπτικές, θα παράγουν δηλ. μιά ακολουθία $\{x^k\}$, $k\geq 0$, διανυσμάτων του R^n, τέτοια ώστε $x^k \to x^*, k\to \infty$, όπου x^* είναι κάποια λύση του (1). Υπάρχουν βασικά τρία σημαντικά ερωτήματα γιά κάθε τέτοια μέθοδο:

Πρώτα - πρώτα, οι προσεγγίσεις x^k, $k\geq 0$, πρέπει να είναι καλά ορισμένες. Υπάρχει δηλ. το ερώτημα της κατάλληλης επιλογής της αρχικής τιμής x^0 (γενικώτερα πιθανώς των αρχικών τιμών $x^0, x^1, \ldots, x^p$) και της κατασκευής των x^k, $k>0$, έτσι ώστε να βρίσκονται μέσα στα πεδία ορισμού των απεικονίσεων που πρόκειται να υπολογισθούν στά x^k.

Το δεύτερο ερώτημα αφορά την σύγκλιση της ακολουθίας $\{x^k\}$ και το αν το όριό της είναι πράγματι λύση του (1). Εδώ διακρίνουμε διαφόρων τύπων θεωρήματα σύγκλισης. Π.χ. ένα θεώρημα τοπικής σύγκλισης ("τοπικό θεώρημα") τυπικά υποθέτει ότι υπάρχει μία λύση x^* του (1) και διαβεβαιώνει ότι υπάρχει μιά περιοχή S του x^* τέτοια ώστε αν η αρχική τιμή επιλεγεί μέσα στην S, τότε η ακολουθία $\{x^k\}$ είναι καλά ορισμένη και συγκλίνει στο x^*. Ένα θεώρημα περιορισμένης σύγκλισης δεν υποθέτει την ύπαρξη λύσης x^* του (1) αλλά διαβεβαιώνει ότι γιά μία ειδική (περιορισμένη γενικά) επιλογή αρχικών τιμών, η $\{x^k\}$ είναι καλά ορισμένη και η σύγκλιση της εξασφαλισμένη σε κάποιο όριο x^* που είναι λύση της (1). Βέβαια τα πιό επιθυμητά αποτελέσματα είναι εκείνα που γιά ορισμένου τύπου απεικονίσεις F εξασφαλίζουν ότι οποιαδήποτε επιλογή τιμών στον R^n (ή τουλάχιστον σ' ένα μεγάλο υποσύνολό του) δίνει μιά καλά ορισμένη ακολουθία, της οποίας το όριο είναι λύση του (1).

Ένα τρίτο πρόβλημα είναι η ταχύτητα σύγκλισης της ακολουθίας $\{x^k\}$ στο x^*. Σχετικές πληροφορίες μπορεί να μας δώσει μία εκτίμηση του σφάλματος $\|x^k-x^*\|$ αν και συνήθως τέτοιες εκτιμήσεις είναι μάλλον πεσσιμιστικές. Γιαυτό ενδιαφερόμαστε πολλές φορές γιά την "ασυμπτωτική ταχύτητα σύγκλισης" δηλ: την συμπεριφορά του σφάλματος γιά μεγάλο k.

Εκτός από τα παραπάνω θεωρητικά ερωτήματα σημασία στην πράξη έχουν αποτελεσματικοί τρόποι αναζήτησης αρχικών τιμών, εκτέλεσης των βημάτων του αλγορίθμου και τερματισμού του με κατάλληλα κριτήρια.

Θα αναλύσουμε πρώτα μία γενική επαναληπτική μέθοδο γιά την προσεγγιστική επίλυση του (1). Υποθέτουμε ότι ένα $x \in \mathbb{R}^n$ είναι λύση του (1) αν και μόνο αν ικανοποιεί την εξίσωση $x=G(x)$ όπου $G:D \subset \mathbb{R}^n \to \mathbb{R}^n$ είναι μία άλλη απεικόνιση, δηλ. ότι κάθε λύση του (1) είναι ένα σταθερό σημείο της απεικόνισης G. Τότε η γενική επαναληπτική μέθοδος

$$(2) \quad x^{k+1} = G(x^k), \; k=0,1,2,\ldots\ldots, \; x^0 \in \mathbb{R}^n \text{ δεδομένο,}$$

είναι μία προφανής μέθοδος γιά την προσέγγιση των σταθερών σημείων της G. Πράγματι αν $x^k \in D$, $k \geq 0$, αν η G είναι συνεχής και αν $x^k \to x^*$, $k \to \infty$ όπου $x^* \in D$ τότε το x^* είναι σταθερό σημείο της G.

Λέμε ότι ένα σημείο $x^* \in D$ είναι σημείο έλξεως γιά την ακολουθία που παράγεται από την αναδρομική σχέση (επαναληπτική μέθοδο) (2) αν υπάρχει (ανοιχτή) περιοχή $S \subset D$ του x^* τέτοια ώστε γιά κάθε $x^0 \in S$ τα x^k που ορίζονται από την (2) να βρίσκονται στην S και να συγκλίνουν στο x^* όταν $k \to \infty$. Συμβολίζουμε με $S(x,\delta)$ την ανοιχτή μπάλα (γιά κάποια νόρμα) κέντρου x και ακτίνας δ και με $\bar{S}(x,\delta)$ την αντίστοιχη κλειστή μπάλα. Έχουμε το εξής γενικό αποτέλεσμα τοπικής σύγκλισης:

Λήμμα 1. Έστω $G:D \subset \mathbb{R}^n \to \mathbb{R}^n$ και $x^* \in D$. Υποθέτουμε ότι υπάρχει ανοιχτή μπάλα $S=S(x^*,\delta) \subset D$ και σταθερά $\alpha<1$ τέτοια ώστε

$$(3) \quad \|G(x)-x^*\| \leq \alpha \|x-x^*\|, \quad \forall x \in S.$$

Τότε, γιά κάθε $x^0 \in S$, οι όροι x^k της ακολουθίας που ορίζονται από την (2) βρίσκονται στην S και συγκλίνουν στο x^*, δηλ. το x^* είναι σημείο έλξεως της (2).

<u>Απόδειξη</u>: 'Εστω $x^0 \in S$. Η απόδειξη είναι επαγωγική. Γιά k=1 $\|x^1-x^*\|=\|G(x^0)-x^*\| \leq$ (από την (3)) $a\|x^0-x^*\|$. Επειδή $a<1$ και $\|x^0-x^*\|<\delta$ έχουμε ότι $\|x^1-x^*\|<\delta$ δηλ. ότι $x^1 \in S$. Υποθέτοντας τώρα ότι γιά κάποιο k, $x^k \in S$ και $\|x^k-x^*\| \leq a^k\|x^0-x^*\|$, έχουμε ότι $\|x^{k+1}-x^*\|=\|G(x^k)-x^*\| \leq a\|x^k-x^*\| \leq a^{k+1}\|x^0-x^*\|$· επίσης $x^{k+1} \in S$ ό.έ.δ. 'Αρα, $x^0 \in S \Rightarrow x^k \in S$ και $\|x^k-x^*\| \leq a^k\|x^0-x^*\|$ γιά $k \geq 0 \Rightarrow x^k \rightarrow x^*$, $k \rightarrow \infty$, δηλ. το x^* είναι σημείο έλξεως γιά την (2). @

Πρίν χρησιμοποιήσουμε το Λήμμα 1 γιά να διατυπώσουμε μία γενική ικανή συνθήκη γιά τοπική σύγκλιση της (2) θα αποδείξουμε ένα βοηθητικό αποτέλεσμα από την γραμμική άλγεβρα. 'Εστω $A \in L(\mathbb{C}^n)$ και $\lambda_i(A)$, $1 \leq i \leq n$ οι ιδιοτιμές του. Η <u>φασματική ακτίνα</u> $\rho(A)$ του A ορίζεται ως $\rho(A)=\max_i|\lambda_i(A)|$. Είναι προφανές ότι γιά κάθε νόρμα $\|\cdot\|$ ισχύει $\rho(A) \leq \|A\|$. Μας ενδιαφέρει το εξής, σχεδόν αντίστροφο, αποτέλεσμα:

Λήμμα 2. 'Εστω $A \in L(\mathbb{C}^n)$. Τότε γιά κάθε $\varepsilon>0$ υπάρχει νόρμα $\|\cdot\|$ του $\mathbb{C}^n$ τέτοια ώστε

(4) $\|A\| \leq \rho(A)+\varepsilon$.

<u>Απόδειξη</u>: Ταυτίζοντας γραμμικούς τελεστές με τους πίνακες που τους παριστάνουν ως πρός την κανονική βάση $\{e^j\}$ του $\mathbb{C}^n$, έχουμε $P^{-1}AP=J$ όπου $J \in L(\mathbb{C}^n)$ είναι η μορφή Jordan του A, η οποία, ως γνωστόν, παριστάνεται από ένα διαγώνιο πίνακα τετραγωνικών υποπινάκων $J=\mathrm{diag}[J_1,\ldots,J_m]$ όπου κάθε υποπίνακας J_i είναι είτε ο 1x1 πίνακας (λ_i) ή είναι διδιαγώνιος πίνακας της μορφής

$$J_i = \begin{bmatrix} \lambda_i & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_i \end{bmatrix}$$

όπου λ_i οι ιδιοτιμές του A. Θεωρούμε τον διαγώνιο nxn πίνακα

$D = \text{diag}\,(1,\varepsilon,\varepsilon^2,\ldots,\varepsilon^{n-1})$. Εύκολα βλέπουμε ότι ο πίνακας $\tilde{J}=D^{-1}JD$ είναι ο ίδιος με τον J εκτός από το ότι οι μονάδες της πρώτης υπερδιαγωνίου του J αντικαθιστώνται από ε. Συνεπώς $\|\tilde{J}\|_\infty \le \rho(A)+\varepsilon$. Θέτουμε $Q=PD$ και ορίζουμε την νόρμα $\|\cdot\|$ στον $\mathbb{C}^n$ από την σχέση $\|x\|=\|Q^{-1}x\|_\infty$. Τότε

$$\|A\|=\max_{\|x\|=1} \|Ax\| = \max_{\|Q^{-1}x\|_\infty=1} \|Q^{-1}Ax\|_\infty = \max_{\|y\|_\infty=1} \|Q^{-1}AQy\|_\infty =$$

$$= \max_{\|y\|_\infty=1} \|\tilde{J}y\|_\infty = \|\tilde{J}\|_\infty \le \rho(A)+\varepsilon.\ @$$

Το παρακάτω θεώρημα διατυπώνει μιά χρήσιμη ικανή συνθήκη γιά την ισχύ της (3).

ΘΕΩΡΗΜΑ 1. (Ostrowski). Υποθέτουμε ότι η $G:D \subset \mathbb{R}^n \to \mathbb{R}^n$ έχει ένα σταθερό σημείο $x^*\in \text{Int}(D)$ και η G έχει παράγωγο (Frechet) $G'(x^*)$ στο x^*. Τότε, αν $\rho(G'(x^*)) \le \sigma < 1$, το x^* είναι σημείο έλξεως της (2).

<u>Απόδειξη</u>: Από το Λήμμα 2 και την υπόθεσή μας έπεται ότι γιά $\varepsilon>0$ υπάρχει νόρμα $\|\cdot\|$ τέτοια ώστε $\|G'(x^*)\| \le \sigma+\varepsilon$. Επειδή η G είναι παραγωγίσιμη στο x^* έχουμε ότι $\exists \delta=\delta(x^*,\varepsilon)>0$ τέτοιο ώστε $S=S(x^*,\delta)\subset D$ και $\|G(x)-G(x^*)-G'(x^*)(x-x^*)\| \le \varepsilon\|x-x^*\|$, $\forall x\in S$. Από τις δύο αυτές σχέσεις και το γεγονός ότι $G(x^*)=x^*$ παίρνουμε

$$\|G(x)-x^*\| \le \|G(x) - G(x^*) - G'(x^*)(x-x^*)\| + \|G'(x^*)(x-x^*)\|$$
$$\le (\sigma+2\varepsilon)\, \|x-x^*\|.$$

Διαλέγοντας λοιπόν $\varepsilon>0$ τέτοιο ώστε $\alpha\equiv\sigma+2\varepsilon<1$ βλέπουμε ότι υπάρχει νόρμα $\|\cdot\|$ τέτοια ώστε $\alpha<1$ και

$$\|G(x)-x^*\| \le \alpha\|x-x^*\| \quad \forall x\in S(x^*,\delta),$$

δηλ. ότι ικανοποιούνται οι υποθέσεις του Λήμματος 1. Συνεπώς το x^* είναι σημείο έλξεως της (2). @

'Υστερα από αυτό το τυπικό δείγμα θεωρήματος τοπικής σύγκλισης προχωρούμε στην διατύπωση ενός αποτελέσματος του τύπου "περιορισμένης σύγκλισης", του γνωστού μας "θεωρήματος της συστολής". Λέμε ότι η απεικόνιση $G:D \subset R^n \rightarrow R^n$ είναι <u>συστολή</u> σ' ένα σύνολο $D_o \subset D$ αν υπάρχει νόρμα $\|\cdot\|$ ~~του R^n~~ και $a<1$ τέτοια ώστε

$$(5) \quad \|G(x) - G(y)\| \leq a\|x-y\|, \quad \forall\, x,y \in D_o.$$

ΘΕΩΡΗΜΑ 2 (Συστολής). Υποθέτουμε ότι η απεικόνιση $G:D \subset R^n \rightarrow R^n$ είναι συστολή σ' ένα κλειστό σύνολο $D_o \subset D$ και ότι $G(D_o) \subset D_o$. Τότε η G έχει ένα μοναδικό σταθερό σημείο x^* στο D_o. Επιπλέον, γιά οποιοδήποτε $x^o \in D_o$, η ακολουθία $\{x^k\}$, $k \geq 0$, που παράγεται από την (2), συγκλίνει στο x^*. Μάλιστα, γιά $\|\cdot\|$, a όπως στην (5), ισχύουν οι εκτιμήσεις

$$(6) \quad \|x^k-x^*\| \leq (a/(1-a))\, \|x^k-x^{k-1}\|, \quad k=1,2,3,\ldots,$$

$$(7) \quad \|x^k-x^*\| \leq (a^k/(1-a))\, \|G(x^o)-x^o\|, \quad k=1,2,3\ldots.$$

<u>Απόδειξη</u>: 'Εστω $x^o \in D_o$. Επειδή $G(D_o) \subset D_o$, η ακολουθία $x^{k+1}=G(x^k)$, $k=0,1,2,\ldots$, είναι καλά ορισμένη και ανήκει στο D_o. Επιπλέον $\|x^{k+1}-x^k\| = \|G(x^k)-G(x^{k-1})\| \leq a\|x^k-x^{k-1}\|$, $k\geq 1$ από την οποία γιά οποιοδήποτε $m\geq 1$ ακέραιο παίρνουμε γιά $k=1,2,3,\ldots$

$$(8) \quad \|x^{k+m}-x^k\| \leq \sum_{i=1}^{m} \|x^{k+i}-x^{k+i-1}\| \leq (a^{m-1}+\ldots+a+1)\, \|x^{k+1}-x^k\|$$

$$\leq (1-a)^{-1}\, \|x^{k+1}-x^k\| \leq a(1-a)^{-1}\, \|x^k-x^{k-1}\| \leq \ldots$$

$$\leq (a^k/(1-a))\, \|x^1-x^o\|.$$

Συμπεραίνουμε ότι η ακολουθία $\{x^k\}$ είναι Cauchy ως πρός την νόρμα $\|\cdot\|$ στο κλειστό σύνολο D_0. Συνεπώς $\exists\, x^* \in D_0$ τέτοιο ώστε $x^k \to x^*$, $k\to\infty$. Το όριο αυτό είναι σταθερό σημείο της G γιατί

$$\|x^*-G(x^*)\| \le \|x^*-x^{k+1}\| + \|G(x^k)-G(x^*)\| \le \|x^*-x^{k+1}\| + \alpha\,\|x^k-x^*\| \to 0,\ k\to\infty,$$

είναι δε το μοναδικό σταθερό σημείο της G στο D_0 διότι αν υπήρχε και άλλο x^{**} θα είχαμε

$$\|x^*-x^{**}\| = \|G(x^*) - G(x^{**})\| \le \alpha\|x^*-x^{**}\| < \|x^*-x^{**}\|,$$

άτοπο. Οι εκτιμήσεις (6),(7) προκύπτουν τώρα από την (8) παίρνοντας τό όριο $m\to\infty$. @

Παρατηρήσεις

1. Το θεώρημα της συστολής (και η απόδειξή του σ' αυτήν την παράγραφο) ισχύει και γιά απεικονίσεις $G:D \subset X \to Y$, όπου X,Y χώροι Banach. Γενικώτερα, αν ο X πλήρης μετρικός χώρος με μετρική $d(\cdot,\cdot)$ και $G:X\to X$ είναι μία απεικόνιση γιά την οποία υπάρχει $\alpha<1$ τέτοιο ώστε $d(G(x),G(y)) \le \alpha d(x,y)$, $\forall x,y\in X$, τότε η G έχει ένα μοναδικό σταθερό σημείο στου X. Υπάρχει και η εξής ενδιαφέρουσα επέκταση: λέμε ότι μία απεικόνιση $G:D\subset R^n \to R^n$ δεν είναι διαστολή στο σύνολο $D_0 \subset D$ αν ικανοποιεί μία συνθήκη Lipschitz με σταθερά ίση με την μονάδα στο D_0, δηλ. αν

$$(9)\quad \|G(x)-G(y)\| \le \|x-y\|,\quad \forall\, x,y\in D_0.$$

Αν ισχύει η (9), αν το D_0 είναι κλειστό και κυρτό και αν $G(D_0) \subset D_0$, τότε η G έχει σταθερό σημείο x^* στο D_0 αν και μόνο αν η ακολουθία (2) είναι φραγμένη γιά τουλάχιστον ένα $x^0\in D_0$ (Στο x^* συγκλίνει τότε μία υπακολουθία της x^k). Γιά την απόδειξη του θεωρήματος αυτού (που αποτελεί ειδική περίπτωση γενικώτερου θεωρήματος του Brouwer) βλ. [2.3, σελ. 121 και Παρ. 6.3].

2. Οι εκτιμήσεις του σφάλματος (6) και (7) είναι προφανώς πολύ χρήσιμες <u>αν είναι γνωστή η τιμή του α</u>. Τότε π.χ. μπορούμε να υπολογίσουμε ακριβώς εκ των προτέρων τον αριθμό των βημάτων που απαιτούνται έτσι ώστε το σφάλμα να γίνει μικρότερο δεδομένου $\varepsilon>0$. Το πρόβλημα είναι βέβαια στην πράξη ο ακριβής κατά το δυνατόν προσδιορισμός της σταθεράς συστολής α.

3. Ειδική περίπτωση απεικόνισης G είναι και η "αφφινική" απεικόνιση (σ' όλου του R^n) $x \mapsto Hx+d$, όπου $d \in R^n$ σταθερό. Σ' αυτήν την περίπτωση ικανή και αναγκαία συνθήκη γιά την σύγκλιση της ακολουθίας $x^{k+1}=Hx^k+d$ στην (υποτιθέμενη μοναδική) λύση x του συστήματος $x=Hx+d$ γιά οποιοδήποτε $x^0 \in R^n$ είναι να υπάρχει νόρμα $\|\cdot\|$ του R^n τέτοια ώστε $\|H\|<1$, δηλ. να είναι η G συστολή (εδώ $\alpha=\|H\|$). Ισοδύναμη ικανή και αναγκαία συνθήκη είναι (βλ. Λήμμα 2) η $\rho(H)<1$. (Αυτό είναι το βασικό αποτέλεσμα σύγκλισης των κλασικών επαναληπτικών μεθόδων γιά την λύση του γραμμικού συστήματος $Ax=b$ το οποίο γιαυτόν τον σκοπό γράφουμε στην μορφή $Mx=Nx+b$, όπου $A=M-N$, M αντιστρέψιμος, δηλ. στην μορφή $x=Hx+d$ με $H=M^{-1}N$, $d=M^{-1}b$. Βλέπε [5.4, Παρ. 11-13]).

Ασκήσεις 2.2

1. (α) Ν'. αποδειχθούν οι ισχυρισμοί γιά την σύγκλιση της $x^{k+1}=Hx^k+d$ της Παρατήρησης 3.

(β) Δώστε παράδειγμα ενός γραμμικού τελεστή $H \in L(R^2)$ που είναι συστολή ως πρός μία νόρμα του R^2 αλλά δεν είναι ως πρός μία άλλη.

2. (α) Ορίζουμε $f:[0,1] \to R^1$ ως $f(x)=(x/2)+2$. Δείξτε ότι η f είναι συστολή στο [0,1] αλλά δεν έχει σταθερό σημείο στο [0,1]. Τι συμβαίνει;

(β) Υποθέστε ότι η $G:D\in R^n \to R^n$ έχει παράγωγο με την έννοια του Gateaux που ικανοποιεί $\|G'(x)\| \leq \alpha<1$ γιά κάθε x σ' ένα κυρτό σύνολο $D_0 \subset D$. Τότε η G είναι συστολή στο D_0. (Συνήθως έτσι εκτιμούμε στην πράξη την τιμή της σταθεράς συστολής).

(γ) Υποθέστε ότι η $G:D \subset R^n \to R^n$ είναι συνεχώς παραγωγίσιμη (κατά Frechet) σε μία (ανοιχτή) περιοχή S_1 ενός σημείου $x\in \mathrm{Int}(D)$ και ότι $\rho(F'(x))<1$. Δείξτε ότι υπάρχει περιοχή S_2 του x και νόρμα $\|\cdot\|$ του R^n τέτοιες ώστε η G να είναι συστολή στην S_2 ως πρός $\|\cdot\|$.

3. Στο Θεώρημα 1 του Ostrowski δείξτε ότι η συνθήκη $\rho(G'(x^*))<1$ δεν είναι αναγκαία γιά σύγκλιση. (Υπόδειξη: ερευνήστε τις ιδιότητες του σημείου $x^*=0$ γιά τις απεικονίσεις $x\mapsto x\pm x^3$ του R^1 στον R^1).

4. Υποθέστε ότι η $G:R^n \to R^n$ έχει την ιδιότητα ότι γιά κάθε συμπαγές (κλειστό και φραγμένο) σύνολο C του R^n υπάρχει σταθερά $\alpha_c<1$ τέτοια ώστε

$$(10)\quad \|G(x)-G(y)\| \leq \alpha_c \|x-y\|, \quad \forall x,y\in C$$

και υποθέστε ότι η G έχει σταθερό σημείο $x^*\in R^n$. Δείξτε ότι το x^* είναι μοναδικό στον R^n και ότι η ακολουθία (2) συγκλίνει στο x^* γιά κάθε $x^0\in R^n$. Μπορείτε να βρείτε ένα παράδειγμα (π.χ. στον R^1) που να δείχνει ότι μόνο η συνθήκη (10) δεν εξασφαλίζει την ύπαρξη σταθερού σημείου γιά την G;

5. (α) 'Εστω ότι η (9) ισχύει σ' ένα σύνολο $D_0 \subset D$ ως αυστηρή ανισότης. Δείξτε τότε ότι η G έχει το πολύ ένα σταθερό σημείο στον D_0. Μόνο όμως η (9), έστω και αν ισχύει ως αυστηρή ανισότης, δεν είναι ικανή να εξασφαλίσει την <u>ύπαρξη</u> σταθερού σημείου:

εξετάστε την συνάρτηση $g: R^1 \to R^1$

$$g(x) = \begin{cases} x + e^{-x/2}, & αν\ x \geq 0 \\ e^{x/2}, & αν\ x<0, \end{cases}$$

γιά $D_o \subset [0,\infty)$ ή $D_o \subset (-\infty,0]$

(β) Θεωρείστε την συνάρτηση $g(x)=-x$ με πεδίο ορισμού $D_o=[-1,1]$. Δείξτε ότι δεν είναι διαστολή στο D_o, και επιπλέον ότι ικανοποιούνται οι υπόλοιπες συνθήκες του θεωρήματος που αναφέρεται στην παρατήρηση 1. Συνεπώς η ύπαρξη σταθερού σημείου ($x^*=0$) είναι εξασφαλισμένη στο D_o. Τί συμβαίνει αν θεωρήσουμε το πεδίο ορισμού $D_o=[-1,-1/2] \cup [1/2,1]$;

6. Λέμε ότι η απεικόνιση $F:R^n \to R^n$ είναι <u>ομοιομορφισμός</u> στον R^n αν είναι 1 πρός 1 και επί και αν οι F και F^{-1} είναι συνεχείς στον R^n. Αποδείξτε το εξής μη γραμμικό ανάλογο του θεωρήματος του Neumann: 'Εστω $F=I-G$ όπου $G:R^n \to R^n$ είναι συστολή στον R^n και όπου I είναι η ταυτότης στον R^n. Δείξτε ότι η F είναι ομοιόμορφισμός στον R^n.

7. (Η άσκηση αυτή αποτελεί συνέχεια της 'Ασκησης 2.1.8)

(α) Η πιό προφανής μέθοδος γιά την επίλυση του συστήματος (2.1.19) είναι μία <u>γενική επαναληπτική μέθοδος</u>. Κατ' αρχήν δείξτε ότι ο A είναι θετικά ορισμένος. Θεωρείστε την απεικόνισή $P:R^n \to R^n$

$$(11)\quad P(x) = -A^{-1}\phi(x),\ x \in R^n.$$

Προφανώς, κάθε λύση του (2.1.19) είναι σταθερό σημείο της P και αντίστροφα. Δείξτε ότι αν η $g': R^1 \to R^1$ είναι συνεχής, τότε η P είναι συνεχώς παραγωγίσιμη (κατά Frechet) στον R^n και ότι η παράγωγός της δίδεται από $P'=-A^{-1}\phi'$. Υποθέστε επιπλέον ότι η g' ικανοποιεί τις συνθήκες

$$(12)\quad 0 \leq g'(s) \leq b < \infty \quad \forall\, s \in R^1.$$

Υπολογίζοντας ένα κατάλληλο άνω φράγμα της ευκλείδιας νόρμας $\|P'(x)\|_2$ δείξτε ότι αν

(13) $b < h^2$

τότε, γιά h αρκετά μικρό, η P είναι συστολή (ως πρός την νόρμα $\|\cdot\|_2$) στον R^n. Συνεπώς η ακολουθία που ορίζεται από την επανάληψη

(14) $x^{k+1} = P(x^k)$, $k \geq 0$, $x^0 \in R^n$ αυθαίρετο,

συγκλίνει στο μοναδικό (κάτω από αυτές τις συνθήκες) σταθερό σημείο της P στον R^n, δηλ. στη μοναδική λύση του (2.1.19). (Υπόδειξη: βρήτε τις ιδιοτιμές του A και υπολογίστε το $\|A^{-1}\|_2$ συναρτήσει του h).

(β) Η συνθήκη (13) και το γεγονός ότι το h πρέπει να είναι αρκετά μικρό (δηλ. το n αρκετά μεγάλο) γιά να συγκλίνει η (14) κάνει την επιλογή του $P=-A^{-1}\Phi$ ως απεικόνιση επανάληψης όχι επιτυχή. Υπάρχει όμως η εξής εναλλακτική λύση: Γιά $\gamma>0$ σταθερό, θεωρείστε την απεικόνιση $G:R^n \to R^n$

(15) $G(x) = (A+\gamma I)^{-1}(\gamma x-\Phi(x))$, $x \in R^n$.

Προφανώς κάθε σταθερό σημείο της G είναι λύση του (2.1.19) και αντιστρόφως. Υποθέστε ότι η g' είναι συνεχής στον R^1 και ικανοποιεί την (12) γιά οποιοδήποτε $b>0$· διαλέξτε $\gamma=h^2b/2$. Δείξτε τότε ότι η G είναι συστολή (ως πρός την νόρμα $\|\cdot\|_2$) στον R^n. Συνεπώς η επανάληψη $x^{k+1}=G(x^k)$, $k \geq 0$ συγκλίνει γιά κάθε $x^0 \in R^n$ στην μοναδική (κάτω από αυτές τις συνθήκες) λύση του (2.1.19).

(γ) Υποθέστε τώρα ότι αντί της (12) ισχύει απλώς ότι

(16) $0 \leq g'(s)$, $\forall s \in R^1$.

Δείξτε τότε ότι γιά οποιαδήποτε τιμή του γ η G που ορίζεται από την (15) δεν είναι αναγκαστικά συστολή στον R^n.

2.3 ΜΕΘΟΔΟΣ ΤΟΥ ΝΕΥΤΩΝΑ: ΤΟΠΙΚΗ ΣΥΓΚΛΙΣΗ ΚΑΙ ΤΑΧΥΤΗΤΑ ΣΥΓΚΛΙΣΗΣ

Σ' αυτήν την παράγραφο θα μελετήσουμε εξειδικεύσεις της γενικής επαναληπτικής μεθόδου (2.2.2) γιά την λύση του συστήματος (2.2.1) στην περίπτωση που η απεικόνιση G είναι της μορφής

$$(1)\quad G(x) = x - A(x)^{-1} F(x)$$

όπου υποθέτουμε ότι γιά x σε κατάλληλο υποσύνολο του R^n, ο $A(x)\epsilon L(R^n)$ είναι αντιστρέψιμος γραμμικός τελεστής. Θα μελετήσουμε δηλ. επαναλήψεις της μορφής

$$(2)\quad x^{k+1} = x^k - A(x^k)^{-1}F(x^k),\ k\geq 0,\ x^0\epsilon R^n \text{ δεδομένο.}$$

Η σημαντικότερη ειδική περίπτωση των (1), (2) είναι η <u>μέθοδος του Νεύτωνα</u> γιά την οποία $A(x)=F'(x)$, όπου F' είναι η παράγωγος (Frechet) της F. Η μέθοδος αυτή γενικεύει την γνωστή μας από τον R^1 μέθοδο του Νεύτωνα ("μέθοδο των Newton-Raphson") $x^{k+1}=x^k-f(x^k)/f'(x^k)$.

Θα αποδείξουμε πρώτα ένα προκαταρτικό αποτέλεσμα (Θεώρημα 1) γιά επαναληπτικές μεθόδους της μορφής (2) το οποίο θα χρησιμοποιήσουμε γιά να δείξουμε εν συνεχεία ένα θεώρημα τοπικής σύγκλισης γιά την μέθοδο του Νεύτωνα. Πρώτα αναφέρουμε την εξής γενίκευση της πρότασης του Neumann (βλ.(1.2.3)) η οποία προκύπτει αμέσως από την (1.2.3) αν θεωρήσουμε αντί του A, τον πίνακα $-A^{-1}B$:

Λήμμα 1. 'Εστω $A\epsilon L(R^n)$ αντιστρέψιμος και έστω $B\epsilon L(R^n)$ τέτοιος ώστε $\|A^{-1}\|\,\|B\|<1$. Τότε ο $A+B$ είναι αντιστρέψιμος· επίσης έχουμε

$$(3)\quad \|(A+B)^{-1}\| \leq \|A^{-1}\|/(1-\|A^{-1}\|\,\|B\|).\ @$$

ΘΕΩΡΗΜΑ 1 Υποθέτουμε ότι η απεικόνιση $F:D\subset R^n \to R^n$ είναι παραγωγίσιμη σ' ένα σημείο $x^*\epsilon \mathrm{Int}(D)$ όπου $F(x^*)=0$. Γιά $x\epsilon S_0$, όπου

$S_0 \subset D$ είναι μιά (ανοιχτή) περιοχή του x^*, θεωρούμε μία οικογένεια γραμμικών τελεστών $A(x) \in L(R^n)$. Υποθέτουμε ότι η $x \mapsto A(x)$ είναι συνεχής στο x^* και ότι ο $A(x^*)$ είναι αυτιστρέψιμος. Τότε υπάρχει κλειστή μπάλλα $\bar{S}=\bar{S}(x^*,\delta) \subset S_0$, $\delta>0$, στην οποία η απεικόνιση $G:\bar{S}\to R^n$, που ορίζεται γιά $x\in\bar{S}$ από την (1) είναι καλά ορισμένη· επιπλέον η G είναι παραγωγίσιμη στο x^* και η παράγωγός της εκεί είναι

$$(4)\ G'(x^*) = I - A(x^*)^{-1} F'(x^*)$$

<u>Απόδειξη</u>: Η G θα είναι καλά ορισμένη στην $\bar{S}$ αν δείξουμε ότι ο $A(x)$ είναι αντιστρέψιμος γιά $x\in\bar{S}$. Θέτουμε $\beta=\|A(x^*)^{-1}\|$. 'Εστω $\varepsilon>0$ τέτοιο ώστε $0<\varepsilon<1/2\beta$. Από την συνέχεια της $x \mapsto A(x)$ στο x^* συμπεραίνουμε ότι υπάρχει $\delta=\delta(x^*,\varepsilon)>0$ τέτοιο ώστε $\bar{S}=\bar{S}(x^*,\delta) \subset S_0$ και

$$(5)\ \|A(x) - A(x^*)\| \le \varepsilon,\quad \forall x\in\bar{S}.$$

Χρησιμοποιούμε τώρα το Λήμμα 1 με $A=A(x^*)$, $B=A(x)-A(x^*)$: επειδή υπάρχει ο $A(x^*)^{-1}$ και $\|A(x^*)^{-1}\|\ \|A(x)-A(x^*)\| \le \varepsilon\beta<1/2$ γιά $x\in\bar{S}$ συμπεραίνουμε ότι ο $A(x)$ είναι αντιστρέψιμος γιά $x\in\bar{S}$ και ότι

$$(6)\ \|A(x)^{-1}\| = \|(A^*(x) + (A(x) - A(x^*)))^{-1}\| \le$$
$$\le \|A(x^*)^{-1}\|/(1-\|A(x^*)^{-1}\|\ \|A(x)-A(x^*)\|) \le \beta/(1-1/2)=2\beta,\quad \forall x\in\bar{S}$$

'Οντως λοιπόν η G είναι καλά ορισμένη γιά $x\in\bar{S}$. Γιά να δείξουμε τώρα ότι η G είναι παραγωγίσιμη στο x^* και ότι η παράγωγος της εκεί δίνεται από την (4), παρατηρούμε πρώτα ότι επειδή το x^* είναι λύση του $F(x)=0$, θα είναι σταθερό της G, δηλ. θα ισχύει

$$(7)\quad G(x^*) = x^*.$$

Χρησιμοποιώντας τώρα την παραγωγισιμότητα της F στο x^* και υποθέτοντας ότι η ακτίνα δ της μπάλλας $\bar{S}=\bar{S}(x^*,\delta)$ είναι αρκούντως μικρή έχουμε ότι

$$(8)\quad \|F(x)-F(x^*) - F'(x^*)(x-x^*)\| \le \varepsilon\|x-x^*\|, \quad \forall x\in\bar{S}$$

Συνεπώς, γιά $x\in\bar{S}$

$\|G(x)-G(x^*) - (I-A(x^*)^{-1} F'(x))\ (x-x^*)\|$ = (από τις (1), (7)) $\|A(x^*)^{-1} F'(x^*)(x-x^*) - A(x)^{-1}F(x)\|$ = (λόγω της $F(x^*)=0$) $\| -A(x)^{-1} (F(x)-F(x^*) - F'(x^*)(x-x^*)) - A(x)^{-1}(A(x^*) - A(x))\ A(x^*)^{-1} F'(x^*)\ (x-x^*)\| \le \|A(x)^{-1} [F(x)-F(x^*)-F'(x^*)(x-x^*)]\| + \|A(x)^{-1} (A(x^*)-A(x))\ A(x^*)^{-1} F'(x^*)(x-x^*)\| \le$ (λόγω των (6), (8), (5) και του ορισμού του β) $\le 2\beta\varepsilon\|x-x^*\| + 2\beta^2\varepsilon\ \|F'(x^*)\|\,\|x-x^*\| \equiv \varepsilon\gamma\ \|x-x^*\|$, όπου $\gamma = 2\beta+2\beta^2\ \|F'(x^*)\|$. Συμπεραίνουμε ότι η G είναι παραγωγίσιμη (Frechet) στο x^* και ότι όντως η παράγωγός της $G'(x^*)$ δίνεται από την (4). @

Αμέσως προκύπτει τώρα ένα αποτέλεσμα τοπικής σύγκλισης γιά την επαναληπτική μέθοδο (2):

Πόρισμα 1 Έστω ότι ισχύουν οι υποθέσεις του Θεωρήματος 1. Επιπλέον υποθέτουμε ότι

$$(9)\quad \rho\ (G'(x^*)) \equiv \rho\ (I-A(x^*)^{-1}\ F'(x^*)) = \varepsilon<1.$$

Τότε το x^* είναι σημείο έλξεως της (2).

<u>Απόδειξη</u>: Εφαρμόζουμε το Θεώρημα 2.2.1 του Ostrowski, το Θεώρημα 1 και την υπόθεσή μας (9). @

Θεωρούμε τώρα την <u>μέθοδο του Νεύτωνα</u> γιά την οποία $A(x)=F'(x)$. Τότε έχουμε το εξής <u>Θεώρημα τοπικής σύγκλισης</u>:

ΘΕΩΡΗΜΑ 2 Υποθέτουμε ότι η $F:D\subset R^n\to R^n$ είναι παραγωγίσιμη σε μία (ανοιχτή) περιοχή $S_0 \subset D$ ενός σημείου x^* όπου $F(x^*)=0$. Υποθέτουμε επίσης ότι η F' είναι συνεχής στο x^* και ότι ο $F'(x^*)$ είναι αντιστρέψιμος. Τότε το x^* είναι σημείο έλξεως της <u>μεθόδου του Νεύτωνα</u>:

(10) $x^{k+1} = x^k - F'(x^k)^{-1} F(x^k)$, $k \geq 0$

<u>Απόδειξη</u> Χρησιμοποιώντας το Θεώρημα 1 με $A(x)=F'(x)$ γιά $x \in S_0$ (ελέγχουμε εύκολα ότι ισχύουν όλες οι υποθέσεις του), συμπεραίνουμε ότι η απεικόνιση επαναλήψεως $G(x)=x-F'(x)^{-1}F(x)$ της μεθόδου του Νεύτωνα είναι καλά ορισμένη σε μία μπάλλα $\bar{S}=\bar{S}(x^*,\delta) \subset S_0$, $\delta>0$. Επιπλέον τώρα $G'(x^*)=I-F'(x^*)^{-1}F'(x^*)=0$ συνεπώς $\rho(G'(x^*))=\sigma=0$ και το πόρισμα 1 δίνει ότι το x^* είναι σημείο έλξεως της (10), δηλ. ότι αν $x^0 \in S$ τότε $x^k \in S$ και $x^k \to x^*$, $k \to \infty$. @

Μετά από αυτό το σημαντικό αποτέλεσμα θα διερευνήσουμε την <u>ταχύτητα σύγκλισης</u> της ακολουθίας x^k της μεθόδου του Νεύτωνα πρός το σημείο έλξεως x^*. Συγκεκριμένα κάτω από ορισμένες συνθήκες μπορούμε να δείξουμε ότι η σύγκλιση είναι "<u>τετραγωνική</u>".

ΠΡΟΤΑΣΗ 1. Έστω ότι ισχύουν όλες οι υποθέσεις του Θεωρήματος 2. Τότε γιά το σημείο έλξεως x^* ισχύει ότι

(11) $\lim_{k\to\infty} \|x^{k+1}-x^*\| / \|x^k-x^*\| = 0$

Αν επιπλέον γιά κάποια σταθερά c_1 έχουμε

(12) $\|F'(x)-F'(x^*)\| \leq c_1 \|x-x^*\|$

γιά κάθε x σε μιά (ανοιχτή) περιοχή του x^*, τότε υπάρχει φυσικός k_0 και σταθερά c_2 τέτοια ώστε

(13) $\|x^{k+1}-x^*\| \leq c_2 \|x^k-x^*\|^2$ γιά $k \geq k_0$,

δηλ. όπως λέμε, η σύγκλιση είναι "τετραγωνική" γιά k αρκετά μεγάλο.

<u>Απόδειξη</u> Γιά την μέθοδο του Νεύτωνα, θέτουντας $G(x)=x-F'(x)^{-1}F(x)$, έχουμε από το Θεώρημα 2 ότι η G είναι καλά ορισμένη σε μία μπάλλα $\bar{S}=\bar{S}(x^*,\delta),\delta>0$ με κέντρο x^*, όπου $G(x^*)=x^*$ και όπου η G είναι παραγωγίσιμη με $G'(x^*)=0$. Χρησιμοποιώντας αυτά τα δεδομένα έχουμε γιά $x^0\in\bar{S}$ ότι $x^{k+1}-x^* = G(x^k) - G(x^*) = G(x^k)-G(x^*)-G'(x^*)(x^k-x^*)$. Άρα $\lim_{k\to\infty} \|x^{k+1}-x^*\| / \|x^k-x^*\| = \lim_{k\to\infty} \|G(x^k)-G(x^*)-G'(x^*)(x^k-x^*)\| / \|x^k-x^*\|=0$ από την παραγωγισιμότητα της G στο x^*. Συνεπώς αποδείχθη η (11). Έστω τώρα ότι γιά $k\geq k_0$ οι όροι $\{x^k\}$ περιέχονται στην περιοχή του x^* όπου ισχύει η (12). Γιά τέτοιο k θεωρείστε ως κυρτό σύνολο D_0 το ευθύγραμμο τμήμα $[x^k x^*]$ και υποθέστε χωρίς περιορισμό της γενικότητας ότι η (12) - πιθανώς με διαφορετική σταθερά c_1 - ισχύει στο D_0. Εφαρμόζοντας το αποτέλεσμα της πρότασης 2.1.4 και την (12) έχουμε τότε ότι γιά κάποια σταθερά c ανεξάρτητη του k ισχύει

$$(14)\quad \|F(x^k)-F(x^*)-F'(x^*)(x^k-x^*)\| \leq c\|x^*-x^k\|^2.$$

Άρα

$$\begin{aligned}
\|x^{k+1}-x^*\| &= \|x^k-F'(x^k)^{-1}F(x^k)-x^*\| \leq (F(x^*)=0)\\
&\leq \|-F'(x^k)^{-1}\ (F(x^k)-F(x^*)-F'(x^*)(x^k-x^*))\|\\
&+ \|F'(x^k)^{-1}\ (F'(x^k) - F'(x^*))\ (x^k-x^*)\|\\
&\leq (\text{από τις } (14),(12))\ \ \|F'(x^k)^{-1}\|\ c\ \|x^k-x^*\|^2\\
&+ \ \|F'(x^k)^{-1}\|\ c_1'\ \|x^k-x^*\| = \|F'(x^k)^{-1}\|\ (c+c_1')\ \|x^k-x^*\|^2.
\end{aligned}$$

Όπως στην (6) της απόδειξης του Θεωρήματος 1 έχουμε τώρα, επειδή ο $A=F'$ ικανοποιεί όλες τις συνθήκες του Θεωρήματος 1, ότι $\|F'(x^k)^{-1}\| \leq 2\ \|F'(x^*)^{-1}\| = \text{σταθ.}$ Συμπεραίνουμε λοιπόν ότι $\|x^{k+1}-x^*\| \leq c_2\ \|x^k-x^*\|^2$ γιά $k\geq k_0$ και γιά σταθερά c_2 ανεξάρτητη του k. ◎

Η εφαρμογή στην πράξη της μεθόδου του Νεύτωνα (10) σε ένα σύνολο του R^n όπου η $F'(x^k)$ είναι αντιστρέψιμη γίνεται ως εξής:

Γιά δεδομένο x^k, υπολογίζουμε το διάνυσμα $F(x^k)$ και την τρέχουσα τιμή του Ιακωβιανού πίνακα $J(x^k)$ που παριστάνει την $F'(x^k)$, λύνουμε το σύστημα

(15) $J(x^k)\, y^k = F(x^k)$

και μετά υπολογίζουμε το x^{k+1} από την σχέση $x^{k+1}=x^k-y^k$.

<u>Παρατηρήσεις</u>

1. Είδαμε ότι σαν συνέπεια του γεγονότος ότι $G'(x^*)=0$, (βλ. απόδειξη της Πρότασης 1) ισχύει η (11) γιά την μέθοδο του Νεύτωνα. Στην περίπτωση μιάς γενικής επαναληπτικής μεθόδου είδαμε (αν η G είναι συστολή) ότι έχουμε απλώς $\|x^{k+1}-x^*\|/\|x^k-x^*\| \le a < 1$, $k\ge 0$ (βλ. απόδειξη Θεωρήματος 2.2.2). Αν ισχύει η (12) είδαμε επίσης ότι γιά k αρκετά μεγάλο η ακολουθία της μεθόδου του Νεύτωνα ικανοποιεί την $\|x^{k+1}-x^*\| \le c\|x^k-x^*\|^2$ ("τετραγωνική" σύγκλιση) ενώ γιά την γενική επαναληπτική μέθοδο έχουμε "γραμμική" σύγκλιση: $\|x^{k+1}-x^*\| \le a\|x^k-x^*\|$. (Γενικά λέμε ότι μία μέθοδος έχει τάξη σύγκλισης p - σε κάποια ρίζα x^* - αν υπάρχει σταθερά c ανεξάρτητη του k και ακέραιος $k_0\ge 0$ τέτοιος ώστε γιά $k\ge k_0$, $\|x^{k+1}-x^*\| \le c\|x^k-x^*\|^p$). Είναι φανερό ότι όσο μεγαλύτερη είναι η τάξη σύγκλισης τόσο πιό γρήγορη είναι η σύγκλιση της x^k στο x^* (γιά x^k κοντά στο x^*). Ο προσδιορισμός καλών μέτρων <u>σύγκρισης</u> της ταχύτητας σύγκλισης διαφόρων μεθόδων είναι σημαντικό ζήτημα· βλέπε την ανάλυση και τους ορισμούς του κεφ. 9 του [2.3].

2. Είδαμε ότι η εφαρμογή της μεθόδου του Νεύτωνα στην πράξη απαιτεί γιά κάθε k τον υπολογισμό ενός nxn πίνακα $J(x^k)$, ενός nx1 διανύσματος $F(x^k)$ και την επίλυση του nxn γραμμικού συστήματος (15) με $n^3/3+O(n^2)$ γενικά πράξεις. Σε πολλές εφαρμογές (όπου μάλιστα μπορεί να απαιτείται η επίλυση <u>πολλών</u> μη γραμμικών συστημάτων) το κόστος του υπολογισμού των πινάκων $J(x^k)$ και της λύσης των γραμμικών συστημάτων μπορεί να είναι απαγορευτικό. Καταφεύγουμε λοιπόν συχνά σε <u>απλουστεύσεις</u> της μεθόδου του Νεύτωνα, δηλ. σε μεθόδους της

μορφής (2), όπου ο γραμμικός τελεστής $A(x^k)$ αποτελεί γενικά <u>προσέγγιση</u> της $F'(x^k)$, οι οποίες έχουν μέν μικρότερο κόστος ανά βήμα, συγκλίνουν όμως πιά όχι "τετραγωνικά" στη x^*. Πολλές φορές το <u>συνολικό</u> κόστος τέτοιων μεθόδων είναι μικρότερο από το συνολικό κόστος της μεθόδου του Νεύτωνα (ιδίως αν <u>δεν</u> απαιτείται μία πολύ ακριβής λύση).

Μία προφανής τέτοια απλούστευση είναι η λεγόμενη "μέθοδος της χορδής":

$$(16)\ x^{k+1} = x^k - F'(x^0)^{-1} F(x^k),\ k\geq 0,\ x^0 \text{ δεδομένο},$$

η οποία σε κάθε βήμα απαιτεί την λύση γραμμικού συστήματος με <u>σταθερό</u> πίνακα $J(x^0)$, ο οποίος συνεπώς μπορεί να υπολογισθεί και να αναλυθεί σε μορφή LU πρίν αρχίσει η επανάληψη (16). Μπορεί να αποδειχθεί ότι γενικά αυτή η μέθοδος συγκλίνει γραμμικά.

Σε μία άλλη στρατηγική απλούστευσης, παρατηρούμε ότι ο Ιακωβιανός πίνακας $J(x)$ μπορεί πάντα να γραφεί ως άθροισμα

$$(17)\ J(x) = D(x)+L(x)+U(x)$$

ενός διαγωνίου πίνακα $D(x)$ με $D_{ii}=J_{ii}$, ενός αυστηρά κάτω τριγωνικού πίνακα $L(x)$ με $L_{ij}=J_{ij}$, αν $i>j$, $L_{ij}=0$ αν $i\leq j$ και ενός αυστηρά άνω τριγωνικού πίνακα $U(x)$ με $U_{ij}=J_{ij}$ αν $i<j$, $U_{ij}=0$ αν $i\geq j$, σε αναλογία με ότι γίνεται γιά τις κλασικές επαναληπτικές μεθόδους, βλ. [5.4, Κεφ. 11-13]. Θεωρούμε τότε, (αν $J_{ii}\neq 0$), γιά $A(x)$ π.χ. τον πίνακα $D(x)+L(x)$ οπότε προκύπτει η επαναληπτική μέθοδος

$$(18)\ x^{k+1}=x^k-(D(x^k)+L(x^k))^{-1} F(x^k),\ k\geq 0,\ x^0 \text{ δεδομένο},$$

που απαιτεί την λύση ενός τριγωνικού γραμμικού συστήματος γιά κάθε k. (Στην περίπτωση ενός γραμμικού συστήματος η μέθοδος αυτή συμπίπτει με την γνωστή μέθοδο των Gauss-Seidel· ανάλογα λοιπόν η (18) ονομάζεται μέθοδος Newton/Gauss-Seidel). Μία πιό ακραία επιλογή είναι $A(x)=D(x)$ οπότε προκύπτει η λεγόμενη μέθοδος Newton/Jacobi:

(19) $x^{k+1} = x^k - D(x^k)^{-1} F(x^k)$, $k\geq 0$, x^0 δεδομένο.

Γιά τις μεθόδους αυτές μπορούμε να αποδείξουμε τοπικά θεωρήματα σύγκλισης (βλ. Ασκήσεις 2,3 παρακάτω). Βέβαια η σύγκλισή τους είναι γενικά γραμμική.

3. Επιπλέον στην πράξη πολλές φορές δεν είναι γνωστές (ή είναι δύσκολο να υπολογισθούν αναλυτικά) οι μερικές παράγωγοι $\partial_j f_i$ των f_i αλλά μόνο οι f_i. Τότε μπορούμε να καταφύγουμε σε προσεγγίσεις του J_{ij} όπου αντί των $\partial_j f_i(x^k)$ χρησιμοποιούμε <u>πεπερασμένες διαφορές</u> $(f_i(x^k+he^j) - f_i(x^k))/h$. Αν το h είναι αρκετά μικρό (και μάλιστα αν πάρουμε <u>ακολουθία</u> όλο και μικρότερων h_k) μπορούμε να δείξουμε ότι η προσεγγιστική αυτή τεχνική διατηρεί πολλά από τα χαρακτηριστικά της μεθόδου του Νεύτωνα. Μάλιστα μπορεί να αποδειχθεί (βλ. π.χ. [2.1, σελ. 94]) ότι αν η F και η λύση x* ικανοποιούν τις συνθήκες του Θεωρήματος 2 και την (12), τότε υπάρχει ε>0 και περιοχή S του x* τέτοια ώστε αν διαλέξουμε ακολουθία πραγματικών $\{h_k\}$ με $0<|h_k|\leq\epsilon$, $k=0,1,2,\ldots$, τότε η ακολουθία $\{x^k\}$ που παράγει η μέθοδος (1), όπου ο $A(x^k)$ παριστάνεται από τον πίνακα

$$A_{ij}(x^k) = (f_i(x^k+h_k e^j)-f_i(x^k))/h_k,\ 1\leq i,\ j\leq n,\ k\geq 0,$$

είναι καλά ορισμένη γιά $x^0\epsilon S$ και έχει την λύση x* ως σημείο έλξεως. Επιπλέον η σύγκλισή της στο x* είναι γενικά γραμμική. Αν διαλέξουμε ακολουθία $\{h_k\}$ τέτοια ώστε $\lim_{k\to\infty} h_k=0$ τότε η σύγκλιση είναι "υπεργραμμική", δηλ. η μέθοδος έχει τάξη σύγκλισης 1<p<2. Μάλιστα, αν η ακολουθία $\{h_k\}$ τείνει αρκετά γρήγορα στο 0, έτσι ώστε π.χ. $|h_k|\leq c_1\|x_k-x^*\|$ ή $|h_k|\leq c_2\|F(x^k)\|$ γιά σταθερές c_1,c_2, τότε η σύγκλιση είναι τετραγωνική.

Σε μία διάσταση είναι γνωστή μας βέβαια και η λεγόμενη "μέθοδος της τέμνουσας" που αντικαθιστά το πηλίκο $f(x^k)/f'(x^k)$ της μεθόδου του Νεύτωνα με το πηλίκο διαφορών $f(x^k)(x^k-x^{k-1})/(f(x^k)-f(x^{k-1}))$.

Η μέθοδος αυτή έχει υπεργραμμική σύγκλιση, βλ. [5.2]. Γενικεύσεις της μεθόδου αυτής σε συστήματα καθώς και παρεμφερείς προσεγγίσεις $A(x)$ της $F'(x)$ οδηγούν σε μία μεγάλη κλάση μεθόδων, τις λεγόμενες μεθόδους του "τύπου του Νεύτωνα" (Quasi-Newton methods) που αποτελούν τα τελευταία χρόνια αντικείμενο ιδιαίτερου ενδιαφέροντος· βλ. το βιβλίο [2.1].

Ασκήσεις 2.3

1. Θεωρείστε την μέθοδο του Νεύτωνα στον R^1 γιά τις συναρτήσεις $f(x)=x^2$ και $f(x)=x+x^{1+\alpha}$ όπου $0<\alpha<1$. Δείξτε απ' ευθείας ότι και στις δύο περιπτώσεις το σημείο $x^*=0$ είναι σημείο έλξης της ακολουθίας της μεθόδου του Νεύτωνα αλλά ότι η σύγκλιση δεν είναι τετραγωνική.

2. Διατυπώστε και αποδείξτε χρησιμοποιώντας το Θεώρημα 1 και το Πόρισμα 1 (όπως κάναμε δηλ. στο Θεώρημα 2 γιά την μέθοδο του Νεύτωνα) Θεώρημα τοπικής σύγκλισης γιά την μέθοδο της χορδής (16).

3. Ομοίως γιά την μέθοδο Newton/Gauss-Seidel (18) και γιά την μέθοδο Newton/Jacobi (19).

4. (Η άσκηση αυτή αποτελεί συνέχεια των Ασκήσεων 2.1.8, 2.2.7).

(α) Υποθέστε ότι $g'(s)\geq 0$ $\forall s\in R^1$. Δείξτε ότι ο Ιακωβιανός πίνακας που παριστάνει την παράγωγο $F'(x)$ που δίνεται από την (2.1.20) είναι αντιστρέψιμος.

(β) Αποδείξτε ότι η μέθοδος του Νεύτωνα γιά το μη γραμμικό σύστημα (2.1.19) συγκλίνει τοπικά. Δείξτε δηλ. ότι μία λύση U^* του (2.1.19) είναι σημείο έλξης της ακολουθίας της μεθόδου του Νεύτωνα. Υποθέστε ότι η g είναι συνεχώς παραγωγίσιμη και ότι $g'(s)\geq 0$, $\forall s\in R^1$.

(γ) Θεωρείστε την μέθοδο Newton/Gauss-Seidel γιά την επίλυση του (2.1.19). Με τις υποθέσεις του (β) δείξτε ότι μία λύση U^* του (2.1.19) είναι σημείο έλξεως γιά την μέθοδο αυτή.

2.4 ΤΟ ΘΕΩΡΗΜΑ ΤΟΥ KANTOROVICH ΓΙΑ ΤΗΝ ΣΥΓΚΛΙΣΗ ΤΗΣ ΜΕΘΟΔΟΥ ΤΟΥ ΝΕΥΤΩΝΑ

Σ' αυτήν την παράγραφο θα μελετήσουμε την απόδειξη του Kantorovich (1948) γιά την σύγκλιση της μεθόδου του Νεύτωνα. Το αποτέλεσμα αυτό είναι του τύπου "περιορισμένης σύγκλισης", όπως π.χ. ήταν και το Θεώρημα 2.2 της συστολής.

ΘΕΩΡΗΜΑ 1 (Kantorovich). 'Εστω ότι η απεικόνιση $F:D\subset R^n \longrightarrow R^n$ είναι παραγωγίσιμη σ' ένα κυρτό σύνολο $D_0 \subset D$ και έστω ότι γιά κάποια σταθερά γ ισχύει

(1) $\|F'(x) - F'(y)\| \leq \gamma\|x-y\|$, $x,y\in D_0$.

'Εστω ότι $F'(x^0)$ είναι αντιστρέψιμος σ' ένα σημείο $x^0\in D_0$. 'Εστω επίσης ότι γιά κάποιες σταθερές β,ζ έχουμε εκεί ότι

(2) $\|F'(x^0)^{-1}\| \leq \beta$,

(3) $\|F'(x^0)^{-1} F(x^0)\| \leq \zeta$,

όπου

(4) $\alpha \equiv \beta\gamma\zeta \leq 1/2$.

Θέτουμε

(5) $t^* = [1-(1-2\alpha)^{1/2}]/\beta\gamma$, $t^{**} = [1+(1-2\alpha)^{1/2}]/\beta\gamma$

και υποθέτουμε ότι $\bar{S}(x^0,t^*) \subset D_0$. Τότε οι όροι της ακολουθίας της μεθόδου του Νεύτωνα

(6) $x^{k+1} = x^k - F'(x^k)^{-1} F(x^k)$, $k=0,1,2,...$

είναι καλά ορισμένοι, παραμένουν μέσα στην $\bar{S}(x^0,t^*)$ και συγκλίνουν γιά $k \to \infty$ σε μία λύση x^* του $F(x)=0$ (στην μοναδική λύση του $F(x)=0$ στο σύνολο $\bar{S}(x^0,t^*)$). Επιπλέον έχουμε την εξής εκτίμηση του σφάλματος:

$$(7)\quad \|x^k-x^*\| \le (2\alpha)^{2^k} / 2^k\beta\gamma, \quad k=0,1,2,\ldots .$$

Απόδειξη Θα υποθέσουμε ότι η (4) ισχύει ως αυστηρή ανισότης, δηλ. ότι $\alpha<1/2$. Η περίπτωση $\alpha=1/2$ είναι γενικά απλούστερη αλλά απαιτεί ορισμένες απλές αλλαγές στις αποδείξεις (Ασκ. 1)

Κατ' αρχήν δείχνουμε ότι η απεικόνιση επανάληψης $G=I-(F')^{-1}F$ της μεθόδου του Νεύτωνα είναι καλά ορισμένη στην $\bar{S}(x^0,t^*)$. Ορίζουμε $D_1 \equiv S(x^0,(\beta\gamma)^{-1}) \cap D_0$ Τότε, γιά $x \in D_1$ έχουμε απ' την (1) ότι $\|F'(x)-F'(x^0)\| \le \gamma\|x-x^0\| < \beta^{-1}$. Άρα η (2) δίνει ότι $\|F'(x^0)^{-1}\|$ $\|F'(x)-F'(x^0)\| <1$ γιά $x \in D_1$. Συμπεραίνουμε από το Λήμμα 2.3.1 ότι οι γραμμικοί τελεστές $F'(x)$, $x \in D_1$ είναι αντιστρέψιμοι. Επιπλέον από τις (2.3.3), (2) και τα παραπάνω έχουμε γιά $x \in D_1$:

$$\begin{aligned}(8)\quad \|F'(x)^{-1}\| &\le \|F'(x^0)^{-1}\|/(1-\|F'(x^0)^{-1}\|\,\|F'(x)-F'(x^0)\|)\\ &\le \beta/(1-\beta\gamma\|x-x^0\|)\end{aligned}$$

Τώρα, επειδή $\alpha<1/2$, η (5) δίνει ότι $t^*<(\beta\gamma)^{-1}$, δηλ ότι $\bar{S}(x^0,t^*) \subset D_1$, έτσι ώστε η G που δίνεται από

$$(9)\quad G(x) = x-F'(x)^{-1}F(x)$$

είναι καλά ορισμένη στην $\bar{S}(x^0,t^*)$.

Σαν ένα προκαταρτικό βήμα θα κάνουμε τώρα μία εκτίμηση γιά x και $G(x)$ ε $S(x^0,t^*)$ της ποσότητας $\|G^2(x)-G(x)\|$, όπου συμβολίζουμε $G^2(x)=G(G(x))$. Η (9) δίνει

$$(10)\quad \|G^2(x)-G(x)\| = \|F'(G(x))^{-1}\, F(G(x))\| =$$
$$= \|\, F'(G(x))^{-1}\, [F(G(x))-F(x)-F'(x)\,(G(x)-x)]\|$$
$$\le \|F'(G(x))^{-1}\|\, \|F(G(x))-F(x)-F'(x)(G(x)-x)\| \le \text{(από την (8))}$$
$$\le \beta\, \|F(G(x))-F(x)-F'(x)(G(x)-x)\| \,/\, (1-\beta\gamma\, \|G(x)-x^0\|).$$

Τώρα η (1) και η (2.1.10) δίνουν γιά $x, G(x) \in S(x^0,t^*)$

$$(11)\quad \|F(G(x))-F(x)-F'(x)\,(G(x)-x)\| \le \gamma\, \|G(x)-x\|^2/2.$$

Συνεπώς από την (10) και (11) παίρνουμε γιά $x,\ G(x)\in S(x^0,t^*)$ ότι

$$(12)\quad \|G^2(x)-G(x)\| \le \beta\gamma\, \|G(x)-x\|^2 \,/\, 2(1-\beta\gamma\, \|G(x)-x^0\|).$$

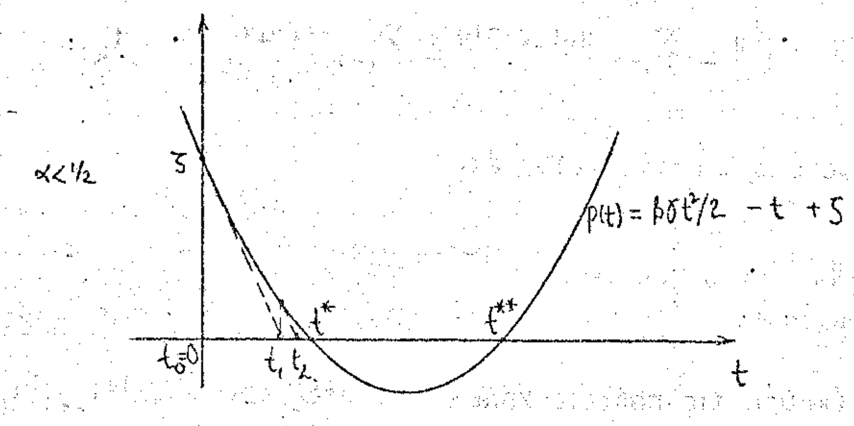
Θα επιστρέψουμε σε λίγο στην εκτίμηση (12). Πρός το παρόν ορίζουμε την ακολουθία των πραγματικών αριθμών $\{t_k\}$, $k\ge 0$ από τις σχέσεις

$$(13)\quad t_{k+1}=t_k - ((\beta\gamma t^2_k/2) - t_k+\zeta)/(\beta\gamma t_k-1),\ k\ge 0,\ t_o=0.$$

Είναι εύκολο να διαπιστώσουμε ότι η ακολουθία $\{t_k\}$, $k\ge 0$ είναι καλά ορισμένη, μονοτονικά αύξουσα και συγκλίνει στον αριθμό t^*. Πράγματι, θεωρούμε το διώνυμο

$$(14)\quad p(t) = (\beta\gamma t^2/2)-t+\zeta.$$

Επειδή $\alpha=\beta\gamma\zeta<1/2$ το $p(t)$ έχει ρίζες τους αριθμούς t^* και t^{**} που δίνονται από την (5). Επιπλέον η (13)



δεν είναι τίποτε άλλο από την ακολουθία που παράγει η μέθοδος του Νεύτωνα γιά την προσέγγιση της ρίζας t^* της εξίσωσης $p(t)=0$ με αρχική τιμή $t_o=0$. Η κυρτότητα και η ομαλότητα τού $p(t)$ δίνουν τώρα ότι $0<t_i<t_{i+1}$ γιά $i\geq 1$ και ότι $t_k\to t^*$, $k\to\infty$.

Ο σκοπός μας είναι να δείξουμε τώρα, χρησιμοποιώντας την (12) και τις ιδιότητες της ακολουθίας $\{t_k\}$, $k\geq 0$, ότι οι όροι $\{x^k\}$, $k\geq 0$ της ακολουθίας (6) της μεθόδου του Νεύτωνα είναι καλά ορισμένοι, ανήκουν στην $\bar{S}(x^o,t^*)$ και ικανοποιούν τις ανισότητες

$$(15)\quad \|x^k-x^{k-1}\|\leq t_k-t_{k-1},\ k\geq 1.$$

Είδαμε ήδη, επειδή η $G(x)$ είναι καλά ορισμένη στην $\bar{S}(x^o,t^*)$, ότι το $x^1=G(x^o)$ υπάρχει. Επιπλέον, από τις (6), (3), (13)

$$\|x^1-x^o\| = \|F'(x^o)^{-1}F(x^o)\|\leq \zeta = t_1 = t_1-t_o.$$

Συνεπώς ισχύει η (15) γιά $k=1$. Επιπλέον, επειδή $\zeta=t_1<t^*$ έχουμε ότι $x^1\in S(x^o,t^*)$. Υποθέτουμε τώρα ότι γιά κάποιο $k\geq 1$ τα $x^o,\dots,x^k$ υπάρχουν στην $\bar{S}(x^o,t^*)$ και ικανοποιούν τις σχέσεις

$$(16)\ \|x^i-x^{i-1}\|\leq t_i - t_{i-1},\ 1\leq i\leq k.$$

Επειδή η G είναι καλά ορισμένη στην $\bar{S}(x^o,t^*)$, το $x^{k+1}=G(x^k)$ υπάρχει. Επιπλέου επειδή η υπόθεση επαγωγής (16) συνεπάγεται ότι

$$(16')\quad \|x^k-x^o\|\leq \sum_{i=1}^{k}\|x^i-x^{i-1}\|\leq \sum_{i=1}^{k}(t_i-t_{i-1}) = t_k,$$

συμπεραίνουμε από τις (12),(16) ότι

$$\begin{aligned}(17)\quad \|x^{k+1}-x^k\| &= \|G(x^k)-G(x^{k-1})\|=\|G^2(x^{k-1})-G(x^{k-1})\|\\ &\leq \beta\gamma\,\|x^k-x^{k-1}\|^2/2(1-\beta\gamma\,\|x^k-x^o\|)\leq \beta\gamma(t_k-t_{k-1})^2/2(1-\beta\gamma t_k)\\ &= (\text{κάντε τις πράξεις χρησιμοποιώντας την (13)})=t_{k+1}-t_k.\end{aligned}$$

Συνεπώς ισχύει η (16) γιά $i=k+1$. Επιπλέον

$$\|x^{k+1}-x^{0}\| \leq \|x^{k}-x^{0}\|+\|x^{k+1}-x^{k}\| \leq t_k+ t_{k+1}-t_k=t_{k+1}<t_*.$$

Συμπεραίνουμε ότι $x^{k+1} \in \bar{S}(x^0,t^*)$. Το επαγωγικό βήμα τελείωσε· έχουμε δείξει ότι όλοι οι όροι της ακολουθίας x^k υπάρχουν και παραμένουν στην $\bar{S}(x^0,t^*)$· επιπλέον δε ικανοποιούν γιά $i\geq 1$ την (16).

Η τελευταία παρατήρηση μας δίνει βέβαια ότι γιά κάθε $k\geq 0$ και $m\geq 1$ έχουμε

$$(18) \quad \|x^{k+m}-x^{k}\| \leq \|x^{k+m}-x^{k+m-1}\| +\ldots+ \|x^{k+1}-x^{k}\| \leq t_{k+m}-t_k,$$

και επειδή $t_k\to t^*$, $k\to\infty$ συμπεραίνουμε ότι η $\{x^k\}$, $k\geq 0$, είναι Cauchy στην $\bar{S}(x^0,t^*)$. Συνεπώς, υπάρχει $x^*\in\bar{S}(x^0,t^*)$ τέτοιο ώστε

$$\lim_{k\to\infty} x_k = x^*.$$

Το x^* είναι όντως λύση του συστήματος $F(x)=0$: Πράγματι, η ύπαρξη της F' στο D_0 συνεπάγεται την συνέχεια της F στο D_0 (Πρόταση 2.1.1). Άρα επειδή

$$\begin{aligned}\|F(x^k)\| &= \|F'(x^k)(x^{k+1}-x^k)\| \leq \|F'(x^k)\|\,\|x^{k+1}-x^k\|\\ &\leq (\|F'(x^k)-F'(x^0)\| + \|F'(x^0)\|)\,\|x^{k+1}-x^k\| \leq \text{(από την (1))}\\ &\leq (\gamma\,\|x^k-x^0\| + \|F'(x^0)\|)\,\|x^{k+1}-x^k\| \leq \text{(από την (16'))}\\ &\leq (\gamma t^* + \|F'(x^0)\|)\,\|x^{k+1}-x^k\| \to 0,\ k\to\infty,\end{aligned}$$

συμπεραίνουμε ότι $F(x^*)=\lim_{k\to\infty} F(x^k)=0$.

Γιά την μοναδικότητα της λύσης x^*, υποθέτουμε ότι υπάρχει και άλλη λύση y^*, $F(y^*)=0$, $y^*\in\bar{S}(x^0,t^*)$. Τότε ισχύει ότι

$$(19) \quad \|y^*-x^k\| \leq t^*-t_k,\quad k=0,1,2,\ldots\ .$$

Πράγματι η (19) ισχύει γιά $k=0$ γιατί $\|y^*-x^o\| \leq t^*=t^*-t_o$. Υποθέτουμε ότι ισχύει γιά $0\leq k\leq i$. Τότε, επειδή $y^*=G(y^*)=G^2(y^*)$ έχουμε από τις (12), (16') ότι

$$\|y^*-x^{i+1}\|=\|G(y^*)-G(x^i)\| = \|y^*-x^i+F'(x^i)^{-1}F(x^i)\|$$
$$= \|F'(x^i)^{-1}(F(x^i)-F(y^*)-F'(x^i)(x^i-y^*))\| \leq$$
$$\leq \beta\gamma\, \|y^*-x^i\|^2/2(1-\beta\gamma\, \|x^i-x^o\|) \leq \beta\gamma\, \|y^*-x^i\|^2/2\, (1-\beta\gamma t_i).$$

Άρα η (19) δίνει ότι

$$(19')\ \|y^*-x^{i+1}\| \leq \beta\gamma(t^*-t^i)^2/2(1-\beta\gamma t_i) = (\text{πράξεις!})=t^*-t_{i+1}.$$

Άρα ισχύει η (19) και γιά $k=i+1$ και συνεπώς γιά κάθε k.
Επειδή $t_k\to t^*$, $k\to\infty$ έχουμε ότι $y^*=\lim_{k\to\infty} x^k = x^*$, δηλ. ότι η λύση x^* είναι η μοναδική λύση του συστήματος $F(x)=0$ στην μπάλα $\bar{S}(x^o,t^*)$.

Τέλος ερχόμαστε στην απόδειξη του φράγματος (7) γιά το σφάλμα της μεθόδου. Κατ' αρχήν δείχνουμε ότι

$$(20)\ t_{k+1}-t_k \leq \zeta 2^{-k},\quad k=0,1,2,\dots\ .$$

Πράγματι η (20) ισχύει γιά $k=0$ επειδή $t_1-t_o=t_1=\zeta$. Υποθέτουντας ότι ισχύει γιά $0\leq k\leq i$ έχουμε από την τελευταία ισότητα της (17) και την επαγωγική υπόθεση ότι

$$(21)\quad t_{i+2}-t_{i+1} = \beta\gamma(t_{i+1}-t_i)^2/2(1-\beta\gamma t_{i+1})$$
$$\leq \beta\gamma\zeta^2 2^{-2i}\ /\ 2(1-\beta\gamma t_{i+1}).$$

Εξ άλλου πάλι απ' την υπόθεση της επαγωγής έχουμε

$$t_{i+1} = \sum_{k=0}^{i} (t_{k+1}-t_k) \leq \zeta \sum_{k=0}^{i} 2^{-k} = 2\zeta(1-2^{-(i+1)}).$$

Άρα στην (21), λόγω της (4)

$$(20') \quad 1-\beta\gamma t_{i+1} = 1-\alpha t_{i+1}/\zeta \geq 1-2\alpha(1-2^{-(i+1)}) \geq 2^{-(i+1)}.$$

Συνεπώς από τις (21), (4), $t_{i+2}-t_{i+1} \leq \beta\gamma\zeta^2 2^{-i} = \zeta\alpha 2^{-i} \leq \zeta 2^{-(i+1)}$, δηλ. ισχύει η (20) γιά κάθε $k\geq 0$, πράγμα που συνεπάγεται την ισχύ της (20') γιά κάθε $i\geq 0$. Τώρα ισχυριζόμαστε ότι έχουμε επίσης

$$(22) \quad t^*-t_k \leq (\beta\gamma 2^k)^{-1} (2\alpha)^{2^k}, \quad k=0,1,2,\dots .$$

Πράγματι, γιά k=0 έχουμε ότι $t^*-t_o=t^*=(1-(1-2\alpha)^{1/2})/\beta\gamma \leq 2\alpha/\beta\gamma$ γιατί $0\leq\alpha<1/2$. Αν η (21) ισχύει γιά $0\leq k\leq i$, έχουμε από την τελευταία ισότητα της (19'), την (20') και την επαγωγική υπόθεση ότι

$$t^*-t_{i+1} = \beta\gamma(t^*-t_i)^2/2(1-\beta\gamma t_i) \leq \beta\gamma(2\alpha)^{2^{i+1}} 2^{i-1}/\beta^2\gamma^2 2^{2i}$$

$$= (2\alpha)^{2^{i+1}} /\beta\gamma 2^{i+1}.$$

Συνεπώς ισχύει η (22) γιά κάθε $k\geq 0$. Η (18) γιά $m\to\infty$ δίνει τώρα $\|x^*-x^k\|\leq t^*-t^k$. Συνεπώς από την (22) έπεται το ζητούμενο φράγμα (7).@

<u>Παρατηρήσεις</u>

1. Το Θεώρημα του Kantorovich (και η απόδειξή του σ' αυτήν την παράγραφο) ισχύει και γιά γενικώτερες απεικονίσεις μεταξύ χώρων Banach. Ένα σημαντικό του πλεονέκτημα είναι ότι δεν υποθέτει εκ των προτέρων την ύπαρξη μιάς λύσης x^* της εξίσωσης $F(x)=0$. Χρησιμεύει λοιπόν και σαν εργαλείο γιά απόδειξη θεωρημάτων ύπαρξης και μοναδικότητας λύσεων συστημάτων μη γραμμικών εξισώσεων αλλά και συναρτησιακών εξισώσεων όπως π.χ. ολοκληρωτικών και διαφορικών εξισώσεων. Ας σημειωθεί ότι μπορεί να αποδειχθεί (βλ. [2.3, Παρ. 12.5.4]) ότι η λύση x^* είναι η μοναδική λύση του συστήματος $F(x)=0$ όχι μόνο στην $\bar{S}(x^0,t^*)$ αλλά και στο μεγαλύτερο, αν $\alpha<1/2$, σύνολο $\bar{S}(x^0,t^{**})\cap D_0$.

2. Γιά την λύση x^* όπως δείχνει η (19) ισχύει η εκτίμηση

$$\|x^*-x^k\| \leq t^*-t_k,\ k=0,1,2,\ldots$$

Συνεπώς γιά k=1 έχουμε χρησιμοποιώντας την συνθήκη $p(t^*)=0$,

$$\|x^*-x^1\| \leq t^*-t^1 = t^*-\zeta = \beta\gamma(t^*)^2/2.$$

Η (5) δίνει τώρα ότι $t^* = [1-(1-2\alpha)^{1/2}]/\beta\gamma \leq 2\alpha/\beta\gamma$. Άρα έχουμε την

$$\|x^*-x^1\| \leq 2\alpha^2/\beta\gamma = 2\beta\gamma\zeta^2,$$

σχέση η οποία ισχύει γιά κάθε $\zeta \geq \|F'(x^0)^{-1}\ F(x^0)\|$, γιά το οποίο ικανοποιείται η (4). Παίρνοντας $\zeta=\|F'(x^0)^{-1}\ F(x^0)\|$, συμπεραίνουμε λοιπόν ότι οι υποθέσεις του θεωρήματος 1 δίνουν επίσης ότι

$$(23)\quad \|x^*-x^1\| \leq 2\beta\gamma\ \|x^1-x^0\|^2.$$

Η ανισότητα (23) είναι πολύ χρήσιμη γιά την <u>a posteriori εκτίμηση του σφάλματος</u> της μεθόδου του Νεύτωνα: Υποθέστε ότι υπολογίζουμε τους όρους της ακολουθίας της μεθόδου $\{x^k\}$, $k=0,1,2,\ldots$ μέχρις ότου ισχύσει γιά πρώτη φορά το "κριτήριο τερματισμού" $\|x^j-x^{j-1}\|\leq\epsilon$ όπου $\epsilon>0$ δεδομένο. Γιά να βρούμε τώρα μία εκτίμηση του σφάλματος της x^j, εφαρμόζουμε το θεώρημα 1 με x^{j-1} αντί του x^0, δηλ. με $\beta\geq\|F'(x^{j-1})^{-1}\|$ (ποσότητα που μπορεί να εκτιμηθεί στην πράξη με την επίλυση δύο-τριών συστημάτων με πίνακα $F'(x^{j-1})$ κατ' αναλογία με την εκτίμηση στην πράξη του δείκτη κατάστασης ενός πίνακα) και με $\zeta=\|F'(x^{j-1})^{-1}\ F(x^{j-1})\| = \|x^j-x^{j-1}\| \leq \epsilon$, δηλ. με $\alpha=\beta\gamma\zeta \leq \beta\gamma\epsilon$. Αν λοιπόν $\beta\gamma\epsilon\leq1/2$ και $\bar{S}(x^0,t^*) \subset D_0$ - και τα δύο αυτά εξασφαλίζονται αν το ϵ είναι αρκετά μικρό - η (23) ερμηνεύεται σαν

$$\|x^*-x^j\| \leq 2\beta\gamma\epsilon^2,$$

που είναι η ζητούμενη a posteriori εκτίμηση του σφάλματος $\|x^*-x^j\|$· ο υπολογισμός του φράγματος απαιτεί βέβαια και κάποια εκτίμηση του γ, δηλ. ενός φράγματος της "δεύτερης" παραγώγου του F.

3. Οι υποθέσεις του θεωρήματος του Kantorovich επισημαίνουν ένα σημαντικό πρόβλημα που εμφανίζεται και στην εφαρμογή της μεθόδου του Νεύτωνα στην πράξη: δηλ. ότι πρέπει να επιλέξουμε τον πρώτο όρο x^0 της ακολουθίας αρκετά κοντά στην (άγνωστη) λύση x^* γιά να έχουμε σύγκλιση της μεθόδου. Ζητάμε να βρούμε λοιπόν τεχνικές γιά τον εντοπισμό καλών x^0, ή παραλλαγές της μεθόδου του Νεύτωνα που έχουν μεγαλύτερες "περιοχές σύγκλισης", δηλ. μεγαλύτερα σύνολα μέσα στα οποία οποιαδήποτε επιλογή του x^0 θα οδηγήσει σε σύγκλιση της $\{x^k\}$.

Μία κατηγορία τεχνικών που χρησιμοποιούνται στην πράξη είναι οι λεγόμενες μέθοδοι "συνέχισης" (continuation) που μπορούν να περιγραφούν ως εξής: Θεωρούμε το σύστημα $F(x)=0$ που πρέπει να λύσουμε σαν το <u>τελευταίο</u> σε μία συνεχή μονοπαραμετρική οικογένεια συστημάτων

$$H(x,t)=0,$$

όπου $x\in\mathbb{R}^n$ και $t\in[0,1]$ παράμετρος. Υποθέτουμε δηλ. ότι $H(x,1)=F(x)$, και ότι η απεικόνιση $H(x,0)$ είναι τέτοια ώστε το σύστημα $H(x,0)=0$ να λύνεται εύκολα. Π.χ. συνήθεις επιλογές είναι, γιά δεδομένο $a\in\mathbb{R}^n$,

$$H(x,t) = tF(x)+(1-t)(F(x)-F(a))$$

ή

$$H(x,t) = tF(x)+(1-t)(x-a).$$

Βρίσκουμε πρώτα την λύση $\xi^{(o)}$ του συστήματος $H(x,0)=0$. Κατόπιν, γιά ένα διαμερισμό $0=t_0<t_1<\ldots<t_N=1$ του $[0,1]$, λύνουμε τα συστήματα $H(x,t_i)=0$ υπολογίζοντας τις λύσεις τους $\xi^{(i)}$ με την μέθοδο του Νεύτωνα και χρησιμοποιώντάς ως αρχική τιμή γιά κάθε i π.χ. την λύση $\xi^{(i-1)}$ του προηγουμένου συστήματος, βασιζόμενοι στο ότι τα προβλήματα $H(x,t^{i-1})=0$ και $H(x,t^i)=0$ είναι αρκετά "κοντά" μεταξύ τους έτσι ώστε η λύση $\xi^{(i-1)}$ του πρώτου να αποτελεί μία καλή πρώτη προσέγγιση στη λύση του δευτέρου.

Πιό συχνά όμως στην πράξη χρησιμοποιούνται κάποιας μορφής <u>συνδυασμοί</u> της μεθόδου του Νεύτωνα (ή των απλουστεύσεων της ή μεθόδων του "τύπου-του Νεύτωνα" βλ. παρατηρήσεις 2.3.2 και 2.3.3) και μιάς <u>μεθόδου ελαχιστοποίησης</u> ενός καταλλήλου συναρτησιακού. Η γενική δομή μιάς τέτοιας μεθόδου είναι η εξής: δεδομένου του x^k υπολογίζουμε με την μέθοδο του Νεύτωνα (ή κάποια παραλλαγή της) την διαφορά $y^k=x^{k+1}-x^k$. Κατόπιν, είτε αποδεχόμαστε το $x^{k+1}=x^k+y^k$ σαν επόμενο όρο της ακολουθίας, είτε το απορρίπτουμε με κριτήριο το αν ελαττώνει ή όχι κάποιο "φυσικό" μη αρνητικό συναρτησιακό του προβλήματος το οποίο μηδενίζει η λύση, όπως, π.χ. την συνάρτηση $f(x)=\|F(x)\|_2^2$. Αν το x^{k+1} απορριφθεί, τότε υπολογίζουμε ένα νέο x^{k+1} είτε της μορφής x^k+ay^k (δηλ. πάνω στην κατεύθυνση y^k από το x^k· μιά τέτοια κατεύθυνση <u>είναι</u> κατεύθυνση τοπικής ελαχιστοποίησης του $f(x)$ στο x^k, βλ. Ασκ. 4), είτε εκτελώντας ένα βήμα κάποιου αλγόριθμου ελαχιστοποίησης, π.χ. μιάς μεθόδου καθόδου μεγίστης κλίσεως ή μιάς μεθόδου συζυγών κλίσεων γιά συναρτησιακά που προέρχονται από μη γραμμικά προβλήματα. Βλέπε το κεφ. 6 του βιβλίου [2.1] γιά μία καλή εισαγωγή στο θέμα.

4. Η γεωμετρική ερμηνεία της μεθόδου του Νεύτωνα στον $\mathbb{R}^1$ μας πείθει αμέσως ότι αν η $f:\mathbb{R}^1\to\mathbb{R}^1$ είναι κυρτή (ή κοίλη), αυστηρά αύξουσα (ή φθίνουσα) και έχει ρίζα x^*, τότε η ακολουθία $\{x^k\}$ της μεθόδου του Νεύτωνα συγκλίνει στην x^* γιά <u>οποιοδήποτε</u> $x^0\in\mathbb{R}^1$. Κάτι ανάλογο συμβαίνει και στον $\mathbb{R}^n$. Λέμε ότι μία απεικόνιση $F:D\subset\mathbb{R}^n\to\mathbb{R}^n$ είναι <u>κυρτή</u> στο κυρτό σύνολο $D_0\subset D$ αν $F(\lambda x+(1-\lambda)y)\leq\lambda F(x)+(1-\lambda)F(y)$, γιά κάθε $x,y\in D_0$ και $\lambda\in[0,1]$, όπου γιά $x,y\in\mathbb{R}^n$ $x\leq y \Leftrightarrow x_i\leq y_i$, $1\leq i\leq n$ (και ανάλογα γιά $A,B\in L(\mathbb{R}^n)$, $A\leq B \Leftrightarrow A_{ij}\leq B_{ij}$, $1\leq i,j\leq n$). Ισχύει το εξής <u>Θεώρημα του Baluev</u>: Έστω ότι η $F:\mathbb{R}^n\to\mathbb{R}^n$ είναι συνεχώς παραγωγίσιμη και κυρτή σ' όλον του $\mathbb{R}^n$. Υποθέτουμε επίσης ότι η $F'(x)^{-1}$ υπάρχει και μάλιστα ότι $F'(x)^{-1}\geq 0$, $\forall x\in\mathbb{R}^n$. Έστω ότι υπάρχει ρίζα x^* του συστήματος $F(x)=0$. Τότε η x^* είναι μοναδική και η ακολουθία $\{x^k\}$ της μεθόδου του Νεύτωνα συγκλίνει στην x^* γιά οποιοδήποτε $x^0\in\mathbb{R}^n$. Επιπλέον $\forall k\geq 1$, $x^*\leq x^{k+1}\leq x^k$. (Βλ. Ασκ. 5).

Ασκήσεις 2.4

1. Αποδείξτε το Θεώρημα 1 του Kantorovich γιά $a=1/2$.

2. Αποδείξτε ότι αν $t_0 < (\beta\gamma)^{-1}$, τότε η ακολουθία $\{t_k\}$ που ορίζεται από την σχέση (13) συγκλίνει, μονοτονικά αυξανόμενη (γιά $k \geq 1$), στην t^*.

3. Υποθέστε ότι η $F: \mathbb{R}^n \to \mathbb{R}^n$ είναι συνεχώς παραγωγίσιμη σε μιά περιοχή μιάς λύσης x^* του $F(x)=0$ και ότι υπάρχει η $F'(x^*)^{-1}$. Δείξτε ότι η απεικόνιση επανάληψης της μεθόδου του Νεύτωνα $G(x)=x-F'(x)^{-1}F(x)$ είναι καλά ορισμένη και είναι <u>συστολή</u> σε μία περιοχή του x^*.

4. Θεωρείστε το συναρτησιακό $f(x) = \|F(x)\|_2^2$, όπου $F: \mathbb{R}^n \to \mathbb{R}^n$ και το διάνυσμα $y^k = x^{k+1} - x^k$, όπου $\{x^k\}$ η ακολουθία της μεθόδου του Νεύτωνα γιά το σύστημα $F(x)=0$, την οποία υποθέτουμε καλά ορισμένη. Δείξτε ότι το $f(x)$ ελαττώνεται τοπικά κατά μήκος της ακτίνας με αρχή x^k και κατεύθυνση y^k.

5. (Η άσκηση αυτή σκοπό έχει να αποδείξει το Θεώρημα του Baluev της Παρατήρησης 4).

(α) Έστω ότι η $F: D \subset \mathbb{R}^n \to \mathbb{R}^n$ είναι παραγωγίσιμη στο κυρτό σύνολο $D_0 \subset D$. Τότε η F είναι κυρτή στο D_0 (βλ. Παρατήρηση 4) αν και μόνο αν

$$F(y)-F(x) \geq F'(x)(y-x), \quad \forall x,y \in D_0.$$

(β) Χρησιμοποιώντας το (α) δείξτε με επαγωγή ότι οι υποθέσεις του Θεωρήματος του Baluev συνεπάγονται ότι $x^* \leq x^k$ και $F(x^k) \geq 0$ γιά $k=1,2,\ldots$ Συμπέρασμα: $x^* \leq x^{k+1} \leq x^k$ γιά $k \geq 1$, από τον ορισμό του x^{k+1} και την υπόθεση ότι $F'(x)^{-1} \geq 0$.

(γ) Δείξτε τώρα ότι η ακολουθία $\{x^k\}$ συγκλίνει όταν $k \to \infty$ σ' ένα διάνυσμα y το οποίο είναι λύση του $F(x)=0$. (Υπόδειξη: Θεωρείστε γιά $1 \leq i \leq n$ την ακολουθία $\{x_i^k\}$, $k \geq 1$ και χρησιμοποιείστε το (β)).

(δ) Χρησιμοποιώντας το (α) δείξτε ότι η ρίζα x^* του $F(x)=0$ είναι μοναδική στον $\mathbb{R}^n$ και συνεπώς ότι $x^k \to x^*$, $k \to \infty$.

(ε) Η υπόθεση της κυρτότητας της F είναι απαραίτητη στην υπόθεση του θεωρήματος του Baluev: Θεωρείστε τις απεικονίσεις $f:\mathbb{R}^1 \to \mathbb{R}^1$, $f(x)=\arctan x$ και $f(x)=2x+\sin x$. Δείξτε ότι ικανοποιούν όλες τις άλλες συνθήκες του θεωρήματος με εξαίρεση την κυρτότητα της f και ότι η μέθοδος του Νεύτωνα δεν συγκλίνει γιά κάθε $x^0 \in \mathbb{R}^1$ στις περιπτώσεις τους.

6. (Η άσκηση αυτή αποτελεί συνέχεια - και τέλος (!) - των ασκήσεων 2.1.8, 2.2.7 και 2.3.4).

(α) Διατυπώστε κατάλληλες υποθέσεις (η ισχύς των οποίων να είναι δυνατόν να ελέγχεται εύκολα, αλλά τέτοιες ώστε να μην οδηγούν σε τετριμμένα g και U^0!) πάνω στά g, U^0, h γιά να ισχύει ένα θεώρημα του Kantorovich γιά την ακολουθία $\{U^k\}$, $k \geq 0$ της μεθόδου του Νεύτωνα γιά το σύστημα (2.1.19). Διατυπώστε σαφώς τις υποθέσεις και τα συμπεράσματα του θεωρήματός σας.

(β) Έστω ότι η $g:\mathbb{R}^1 \to \mathbb{R}^1$ είναι κυρτή, (βλ. Παρατήρηση 4). Δείξτε ότι οι απεικονίσεις Φ και F, που ορίζονται αντίστοιχα από τις (2.1.18) και (2.1.19) είναι τότε κυρτές στον $\mathbb{R}^n$. Υποθέστε επιπλέον ότι η g είναι συνεχώς παραγωγίσιμη με μη αρνητική παράγωγο στον $\mathbb{R}^1$ και ότι το σύστημα (2.1.19) έχει λύση U^*. Δείξτε ότι η U^* είναι η μοναδική λύση του (2.1.19) στην οποία συγκλίνει η μέθοδος του Νεύτωνα γιά κάθε $U^0 \in \mathbb{R}^n$.

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΑΡΙΘΜΗΤΙΚΗ ΑΝΑΛΥΣΗ

(ΣΗΜΕΙΩΣΕΙΣ ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΜΑΘΗΜΑΤΟΣ)
Β΄ ΜΕΡΟΣ (ΚΕΦ. 3,4)

ΒΑΣΙΛΕΙΟΣ Α. ΔΟΥΓΑΛΗΣ
Μαθηματικό Τμήμα Πανεπιστημίου Αθηνών

ΑΘΗΝΑ 1996

3. ΑΡΙΘΜΗΤΙΚΗ ΛΥΣΗ ΣΥΝΗΘΩΝ ΔΙΑΦΟΡΙΚΩΝ ΕΞΙΣΩΣΕΩΝ

3.1 ΠΡΟΒΛΗΜΑ ΑΡΧΙΚΩΝ ΤΙΜΩΝ. Η ΜΕΘΟΔΟΣ ΤΟΥ EULER

Στο κεφάλαιο αυτό θ' ασχοληθούμε με αριθμητικές μεθόδους γιά την προσεγγιστική λύση του προβλήματος αρχικών τιμών γιά συστήματα πρώτης τάξης Συνήθων διαφορικών Εξισώσεων. (Σ.Δ.Ε) 'Εστω $-\infty<a<b<+\infty$. Ζητάμε μιά συνάρτηση $y:[a,b]\to\mathbb{R}^m$, παραγωγίσιμη στο $[a,b]$, που να ικανοποιεί το πρόβλημα αρχικών τιμών

$$(1)\quad \begin{cases} y'(t) = f(t,y(t)), & a\le t\le b, \\ y(a) = y_0, & \end{cases}$$

όπου f δεδομένη συνάρτηση, $f:[a,b]\times\mathbb{R}^m\to\mathbb{R}^m$ και $y_0\in\mathbb{R}^m$ δεδομένη αρχική τιμή. Γράφοντας $y=(y_1,\dots,y_m)^T$ και $f=(f_1,\dots,f_m)^T$, το πρόβλημα (1) διατυπώνεται και ως εξής: Ζητάμε $y_i:[a,b]\to\mathbb{R}^1$, $1\le i\le m$, τέτοια ώστε

$$(1')\quad \begin{cases} y_i'(t) = f_i(t,y_1,\dots,y_m), \ 1\le i\le m, & a\le t\le b, \\ y_i(a) \ = y_{o,i} \quad , \ 1\le i\le m & \end{cases}$$

όπου $f_i:[a,b]\times\mathbb{R}^m\to\mathbb{R}^1$ δεδομένες συναρτήσεις m+1 μεταβλητών.

Από την θεωρία ύπαρξης-μοναδικότητας λύσεων των συνήθων διαφορικών εξισώσεων* μας είναι γνωστό ότι γιά γενικά δεδομένα a,b,y_0,f, το πρόβλημα (1) μπορεί και να μην έχει λύση ή να έχει

---

* Βλ. π.χ. τα βιβλία G. Birkhoff and G.-C.Rota, "Ordinary differential equations", 2nd ed., Wiley, New York 1969 ή E.A Coddington and N.Levinson, "Theory of ordinary differential equations", McGraw-Hill, New York 1955 ή C. Corduneanu, "Principles of differential and integral equations", 2nd ed., Chelsea, New York 1977, κ.ά.

πολλές λύσεις· βλ. τις Παρατηρήσεις και Ασκήσεις αυτής της παραγράφου γιά μιά σύντομη επισκόπηση. Το παρακάτω Θεώρημα δίνει ικανές συνθήκες γιά ύπαρξη και μοναδικότητα λύσεων γιά κάθε $y_0 \in \mathbb{R}^m$.

ΘΕΩΡΗΜΑ 1. Θεωρούμε το πρόβλημα αρχικών τιμών (1) και υποθέτουμε ότι η συνάρτηση f έχει τις εξής ιδιότητες:

(i) Η $f(t,y)$ είναι συνεχής γιά $(t,y) \in [a,b] \times \mathbb{R}^m$.

(ii) Η $f(t,y)$ ικανοποιεί μία συνθήκη Lipschitz ως πρός y γιά $y \in \mathbb{R}^m$, ομοιόμορφα ως πρός $t \in [a,b]$· δηλ. υπάρχει νόρμα $\|\cdot\|$ του $\mathbb{R}^m$ και σταθερά $L \geq 0$ (η "σταθερά Lipschitz" ως πρός $\|\cdot\|$) τέτοιες ώστε

(2) $\|f(t,y) - f(t,y^*)\| \leq L \|y-y^*\|, \ \forall y,y^* \in \mathbb{R}^m, \ t \in [a,b]$.

Τότε, γιά κάθε $y_0 \in \mathbb{R}^m$, το πρόβλημα (1) έχει μοναδική λύση· υπάρχει δηλ. μοναδική συνάρτηση $y:[a,b] \to \mathbb{R}^m$, παραγωγίσιμη στο $[a,b]$ (μάλιστα έχει συνεχή παράγωγο $y'(t)$ στο $[a,b]$) που ικανοποιεί το σύστημα των ΣΔΕ $y'(t)=f(t,y(t))$ γιά $a \leq t \leq b$ και την αρχική συνθήκη $y(a)=y_0$.@

Μία ικανή συνθήκη γιά την ισχύ της (2) είναι να είναι οι παράγωγοι $\partial f_i / \partial y_j$, $1 \leq i,j \leq m$ συνεχείς και φραγμένες συναρτήσεις γιά $(t,y) \in [a,b] \times \mathbb{R}^m$. Τότε, από το θεώρημα μέσης τιμής γιά συναρτησιακά $f_i : [a,b] \times \mathbb{R}^m \to \mathbb{R}^1$ έχουμε ότι γιά κάθε i, $1 \leq i \leq m$, $y,y^* \in \mathbb{R}^m$ και $t \in [a,b]$ υπάρχει $\xi = \xi(i,t,y,y^*) \in \mathbb{R}^m$, πάνω στο ευθύγραμμο τμήμα $[y,y^*]$, τέτοιο ώστε

$$f_i(t,y) - f_i(t,y^*) = \sum_{j=1}^{m} \partial f_i / \partial y_j \ (t,\xi)(y_j - y_j^*), \quad 1 \leq i \leq m.$$

Έπεται ότι ισχύει η (2) γιά κάποια σταθερά $L = L(M, \|\cdot\|)$ όπου $M = \max_{i,j} (\sup_{(t,y) \in [a,b] \times \mathbb{R}^m} |\partial_i f / \partial_j y|) < \infty$.

Οι υποθέσεις του Θεωρήματος 1 είναι αρκετά περιοριστικές, ιδιαίτερα βέβαια η (2) που υποθέσαμε ότι ισχύει $\forall \ y,y^* \in \mathbb{R}^m$. Σάν επακόλουθο όμως παίρνουμε ύπαρξη και μοναδικότητα της λύσης $y(t)$

<u>σ' όλο</u> το διάστημα $[a,b]$ και γιά <u>κάθε</u> αρχική τιμή $y_0 \in R^m$. Μιά <u>τοπική</u> συνθήκη Lipschitz (δηλ. μία συνθήκη της μορφής (2) γιά $(t,y)\in[a,b]\times\bar{S}$ όπου $\bar{S}=\bar{S}(y_0,r)$ π.χ.) και η συνέχεια της f στο $[a,b]\times\bar{S}$ εγγυώνται ύπαρξη-μοναδικότητα της λύσης $y(t)$ μόνο <u>τοπικά</u>, π.χ. σ' ένα διάστημα της μορφής $[a,\delta)$, $\delta>a$ - βλ. τις Παρατηρήσεις και Ασκήσεις αυτής της παραγράφου. Είναι δυνατόν να αναλύσουμε την σύγκλιση προσεγγιστικών μεθόδων γιά την λύση του (1) ακόμα και όταν αντί της (2) πληρούται μόνο μία τοπική συνθήκη Lipschitz (βλ. π.χ. Ασκ. 2) Γιά να αποφύγουμε όμως τεχνικές πολυπλοκότητες θα υποθέσουμε από δώ κι εμπρός σιωπηρά ότι ισχύουν οι υποθέσεις του Θεωρήματος 1 και (συνεπώς) ότι υπάρχει μοναδική λύση $y(t)$ του (1) με συνεχή παράγωγο στο $[a,b]$. Συνήθως θα υποθέτουμε επιπλέον ότι η λύση είναι αρκετά ομαλή (δηλ. έχει ένα απαιτούμενο αριθμό υψηλοτέρων συνεχών παραγώγων $y^{(j)}(t)=d^jy(t)/dt^j$) ώστε να παίρνουμε την επιθυμητή τάξη σύγκλισης των μεθόδων.

Γιά γενικό δεύτερο μέλος $f(t,y)$ (ακόμα και όταν έχουμε ένα γενικό <u>γραμμικό</u> ομογενές σύστημα, δηλ. με $f(t,y)=A(t)y$ όπου $A:[a,b] \rightarrow R^{m\times m}$) είναι γνωστό ότι δεν μπορούμε να λύσουμε το (1) αναλυτικά. Καταφεύγουμε λοιπόν σε αριθμητικές μεθόδους γιά την προσεγγιστική του λύση. Έστω ο διαμερισμός $a=t^0<t^1<\ldots<t^N=b$ του $[a,b]$, τον οποίο υποθέτουμε γιά απλούστευση <u>ομοιόμορφο</u>, δηλ. ότι $t^n=a+nh$, $0\le n\le N$, όπου $h=(b-a)/N$ είναι το (σταθερό) <u>βήμα</u> του διαμερισμού. Μία αριθμητική μέθοδος γιά την λύση του (1) κατασκευάζει διανύσματα $y^n\in R^m$, $0\le n\le N$, τέτοια ώστε $y^n\cong y(t^n)$, $0\le n\le N$ όπου $y(t)$ η λύση του (1).

Η απλούστατη μέθοδος γιά την προσεγγιστική λύση του (1) είναι ίσως η <u>μέθοδος του Euler</u> (ή "πολυγωνική μέθοδος του Cauchy"). Αν το $h=t^{n+1}-t^n$ είναι αρκετά μικρό, τότε η $y'(t^n)$ προσεγγίζεται ικανοποιητικά από το πηλίκο διαφορών $(y(t^{n+1})-y(t^n))/h$ και η Σ.Δ.Ε. στην (1) δίνει γιά $t=t^n$ την προσεγγιστική σχέση $y(t^{n+1}) \cong y(t^n) + h\, f(t^n,y(t^n))$. Η μέθοδος του Euler στηρίζεται ακριβώς σ' αυτήν την παράσταση: Κατασκευάζει $y^n\in R^m$, $0\le n\le N$, προσεγγίσεις των τιμών $y(t^n)$, από τις σχέσεις

$$(3)\begin{cases} y^0 = y_0 \\ y^{n+1} = y^n + hf(t^n, y^n),\ n=0,1,\dots,\ N-1. \end{cases}$$

Η μέθοδος του Euler (3) είναι προφανώς πολύ απλή στην υλοποίησή της. Απαιτεί γιά κάθε βήμα n τον υπολογισμό της (διανυσματικής) συνάρτησης $f(t^n,y^n)$, m πολλαπλασιασμούς και m προσθέσεις. (Στην περιοχή αυτή της Αριθμητικής Ανάλυσης έχει επικρατήσει να εκτιμάται η πολυπλοκότητα των προβλημάτων από τον αριθμό των υπολογισμών της $f(t,y)$ που απαιτούνται ανά βήμα και τον αριθμό των συστημάτων - γραμμικών ή μη - αλγεβρικών εξισώσεων που τυχόν απαιτούνται ανά βήμα). Συνεπώς η μέθοδος του Euler έχει πολύ χαμηλό <u>κόστος</u> (ανά βήμα). Ας προχωρήσουμε να εξετάσουμε την <u>ακρίβειά</u> της, δηλ. ας προσπαθήσουμε να εκτιμήσουμε τα σφάλματα $\|y^n-y(t^n)\|$, $0\le n\le N$.

ΛΗΜΜΑ 1 'Εστω ότι υπάρχουν σταθερές $\delta>0$ και $K\ge 0$, τέτοιες ώστε η ακολουθία των μη αρνητικών αριθμών $d_0,d_1,\dots$ να ικανοποιεί τις σχέσεις

$$(4)\quad d_{j+1} \le d_j(1+\delta)+K,\qquad j=0,1,2,\dots$$

Τότε ισχύει γιά κάθε $n\ge 0$

$$(5)\quad d_n \le d_0e^{n\delta} + K(e^{n\delta}-1)/\delta.$$

<u>Απόδειξη</u>: Η (4) δίνει $d_n \le (1+\delta)^n d_0+K[1+(1+\delta)+\dots+(1+\delta)^{n-1}] = (1+\delta)^n d_0+K[(1+\delta)^n-1]/\delta$ απ' την οποία έπεται η (5) λόγω της $1+\delta\le e^\delta$. @

ΘΕΩΡΗΜΑ 2 'Εστω $y(t)$ η λύση του (1). Υποθέτουμε (επιπλέον των υποθέσεων του Θεωρήματος 1) ότι η $y^{(2)}(t)$ είναι συνεχής στο $[a,b]$. 'Εστω $\{y^n\}$ $0\le n\le N$ η λύση της μεθόδου του Euler (3). Τότε υπάρχει σταθερά C_1 (είναι η σταθερά σύγκρισης $\|x\|\le C_1\|x\|_\infty$ στον R^m) τέτοια ώστε γιά $0\le n\le N$

$$(6)\ \|y^n-y(t^n)\| \le h(C_1(e^{L(t^n-a)}-1)/2L)\ \max_{t\in[a,t^n]} \|y^{(2)}(t)\|_\infty,$$

όπου η L η σταθερά Lipschitz της (2)

<u>Απόδειξη</u>: 'Εστω $e^n=y^n-y(t^n)$, $0\le n\le N$ το σφάλμα της y^n. Εκ κατασκευής $e^0=0$. Γιά $1\le n\le N$ από το Θεώρημα του Taylor έχουμε ότι

$$(7)\ y(t^{j+1}) = y(t^j) + hy^{(1)}(t^j) + p^j,\quad 0\le j\le n-1$$

όπου

$$p_i^j = h^2\, y_i^{(2)}(\xi_i^j)/2,\quad 1\le i\le m$$

γιά κάποια $\xi_i^j \in [t^j,t^{j+1}]$. Συνεπώς

$$(8)\ \|p^j\| \le C\ \max_i \|p_i^j\| \le K_j = C_1h^2\ \max_{t\in[a,t^{j+1}]} \|y^{(2)}(t)\|_\infty/2.$$

Αφαιρούμε τώρα κατά μέλη την (7) από την αναδρομική σχέση της (3). 'Εχουμε, χρησιμοποιώντας την ΣΔΕ (1)

$$e^{j+1} = e^j + h\,[f(t^j,y^j)-f(t^j,y(t^j))] - p^j.$$

Συνεπώς, χρησιμοποιώντας τις (2), (8) έχουμε γιά $0\le j\le n-1$

$$\|e^{j+1}\| \le \|e^j\| + hL\|y^j-y(t^j)\| + \|p^j\| \le (1+hL)\|e^j\|+K_j \le (1+hL)\|e^j\|+K_n$$

Εφαρμόζοντας τώρα το Λήμμα 1 με $d_j=\|e^j\|$, $d_0=\|e^0\|=0$, $\delta=hL$, $K=K_n$ παίρνουμε την (6) χρησιμοποιώντας το γεγονός ότι $nh=t^n-a$. (Σιωπηρά υποθέσαμε ότι $L>0$. Αν $L=0$ τότε στην (6), αντί του όρου $(e^{L(t^n-a)}/L)$ θέτουμε (t^n-a)) @

Το Θεώρημα 2 μας δίνει συνεπώς μία εκτίμηση του σφάλματος της μορφής

(9) $\max_{0\le n\le N} \|y^n-y(t^n)\| \le Ch,$

όπου C μία σταθερά ανεξάρτητη του h (ή του N) που εξαρτάται όμως από τα δεδομένα a,b,f και την λύση y(t) του (1). Παρατηρούμε ότι το φράγμα στην (9) είναι <u>γραμμικό ως πρός h</u>. Η δύναμη αυτή του h δεν μπορεί να αυξηθεί, όσο ομαλή και αν είναι η λύση (υποθέτουμε ότι $y''(t)\neq 0$): Πράγματι, θεωρείστε το πρόβλημα αρχικών τιμών (στον R^1): $y'=2t$, $0\le t\le 1$, $y(0)=0$ το οποίο φυσικά έχει την μοναδική λύση $y(t)=t^2$. Εύκολα βλέπουμε ότι η μέθοδος του Euler με ομοιόμορφο διαμερισμό στο [0,1] δίνει $y^n=n(n-1)h^2$. Συνεπώς γιά $Nh=1$, π.χ., έχουμε το σφάλμα $y(1)-y^N=h$. Η (9) εκφράζει το ότι αν η y(t) είναι αρκετά ομαλή στο [a,b] (τουλάχιστον C^2) τότε η μέθοδος του Euler (3) έχει <u>τάξη ακρίβειας ίση με 1</u>. Αντίθετα, αν η y(t) είναι λιγώτερο ομαλή (αλλά τουλάχιστον C^1), τότε θα έχουμε γενικά μικρότερη δύναμη του h στο δεύτερο μέλος της (9), που εξαρτάται από κατάλληλη εκτίμηση του "μέτρου συνεχείας" της y'(t).

Τέλος, η ανισότητα (6) ή (9) συνεπάγεται, υπό τις προϋποθέσεις π.χ. του θεωρήματος 2, ότι η μέθοδος του Euler "<u>συγκλίνει</u>". Μ' αυτό εννοούμε ακριβώς ότι αν $n\to\infty$ και $h\to 0$ έτσι ώστε $t^n=a+nh\to t$ γιά κάθε σταθερό $t\in(a,b]$ (π.χ. αν $n\to\infty$, $h\to 0$ έτσι ώστε $nh=t-a$), τότε $y^n\to y(t)$, όπου, γιά δεδομένο h, η y^n είναι προσέγγιση που παίρνουμε μετά από n βήματα της μεθόδου του Euler.

<u>Παρατηρήσεις</u>

1. Σε μαθήματα Συνήθων Διαφορικών Εξισώσεων συνήθως αποδεικνύονται και <u>τοπικές</u> μορφές του θεωρήματος 1 ύπαρξης και μοναδικότητας λύσεων του προβλήματος (1). Π.χ. ισχύει το εξής:

ΘΕΩΡΗΜΑ 1'. Υποθέστε ότι

(i) Η f(t,y) είναι συνεχής στο σύνολο $D=\{(t,y)\in R^{m+1}:\ |t-a|\le A,\ \|y-y_0\|\le B\}$. 'Εστω $M=\max_{(t,y)\in D}\|f(t,y)\|$.

(ii) Η f ικανοποιεί την εξής συνθήκη Lipschitz: υπάρχει σταθερά $L \geq 0$ τέτοια ώστε γιά (t,y), $(t,y^*) \epsilon D$

$$(2') \quad \|f(t,y) - f(t,y^*)\| \leq L \|y-y^*\|.$$

Τότε, υπάρχει μοναδική λύση του προβλήματος αρχικών τιμών (1) <u>τοπικά</u>, δηλ. υπάρχει μοναδική συνάρτηση $y(t)$ ορισμένη γιά $|t-a| \leq \delta = \min\{A, B/M\}$, που ικανοποιεί την $y'=f(t,y)$ γιά $|t-a_n| \leq \delta$ και την αρχική συνθήκη $y(a)=y_0$. @

Σημειωτέον ότι η συνθήκη (i) εγγυάται μόνη της την <u>ύπαρξη</u> (αλλά όχι την μοναδικότητα) λύσεων του προβλήματος γιά $|t-a| \leq \delta = \min\{A, B/M\}$. (Αυτό είναι το λεγόμενο Θεώρημα του Peano). Για <u>μοναδικότητα</u> χρειαζόμαστε κάτι παραπάνω από συνέχεια της f' η συνθήκη Lipschitz (2') είναι "πολύ κοντά" στο να είναι αναγκαία γιά μοναδικότητα (βλ. Ασκ. 1) Γενικεύοντας λίγο (Θεώρημα του Osgood) μπορούμε να αποδείξουμε το εξής: Στο Θεώρημα 1' υποθέτουμε ότι ισχύει η (i) και αντικαθιστούμε την (ii) με το αίτημα να ισχύει στο D η

$$(2'') \quad \|f(t,y) - f(t,y^*)\| \leq \Phi(\|y-y^*\|),$$

όπου $\Phi(u)$ μη αρνητική συνεχής αύξουσα συνάρτηση ορισμένη στο διάστημα $[0,2B]$, τέτοια ώστε $\Phi(0)=0$ και

$$\int_0^{2B} du/\Phi(u) = +\infty$$

(Η συνθήκη Lipschitz αντιστοιχεί σε $\Phi(u)=Lu$. Θα είχαμε επίσης μοναδικότητα αν π.χ. $\Phi(u)=Lu|\ln u|$). Τότε ισχύει το συμπέρασμα του θεωρήματος 1'.

2. Λόγω της (6), δηλ. λόγω μιάς εκτίμησης του σφάλματος της μορφής (9), πρέπει να υπολογίζουμε με πολύ μικρό βήμα h στην μέθοδο του Euler γιά να πάρουμε μικρό σφάλμα $\max_n \|y^n-y(t^n)\|$. Αυτό έχει ως αποτέλεσμα μεγάλο αριθμό πράξεων, δηλ. την αύξηση των <u>σφαλμάτων</u>

στρογγύλευσης. Η ακριβής ανάλυση των αποτελεσμάτων της αριθμητικής πεπερασμένης ακρίβειας στις πράξεις δεν είναι δυνατόν να γίνει στο γενικό πρόβλημα (1) λόγω του ότι δεν είναι γνωστό το πώς υπολογίζεται η $f(t,y)$. Μπορούμε όμως να κάνουμε μιά σειρά από "λογικές" παραδοχές Π.χ. έστω ότι αντί των ακριβών τιμών y^n υπολογίζουμε τις τιμές $\tilde{y}^n$ όπου υποθέτουμε ότι

(i) $\tilde{y}^0=y^0+\delta^0$, $\|\delta^0\|\leq\delta$, δ "μικρό",

ότι αντί του $f(t^n,\tilde{y}^n)$ υπολογίζουμε στην πραγματικότητα το διάνυσμα $\tilde{f}(t^n,\tilde{y}^n)$ όπου

(ii) $\tilde{f}(t^n,\tilde{y}^n) = f(t^n,\tilde{y}^n) + \epsilon^n$, $\|\epsilon^n\|\leq \epsilon$, ϵ "μικρό"

και τελικά ότι έχουμε και σφάλμα στρογγύλευσης κατά τον υπολογισμό του $\tilde{y}^{n+1}$, δηλ. ότι

(iii) $\tilde{y}^{n+1} = \tilde{y}^n + h\tilde{f}(t^n,\tilde{y}^n) + \rho^n$, $\|\rho^n\|\leq\rho$, ρ "μικρό".

Μπορεί να δειχθεί τότε (βλ. Ασκ. 4) ότι υπό τις προϋποθέσεις του θεωρήματος 2 υπάρχουν σταθερές C,C_1,C_2 ανεξάρτητες των h,N τέτοιες ώστε

(10) $\max_n \|\tilde{y}^n-y(t^n)\| \leq C\,h + C_1\delta + C_2\,(\epsilon+\rho h^{-1})$,

(όπου η C είναι η ίδια με την C της (9), δηλ. όπου $C=\max_{t\in[a,b]} \|y''(t)\|\ (e^{L(b-a)}-1)/2L$). Η εκτίμηση αυτή δείχνει ότι υπάρχει μία κρίσιμη τιμή h_0 του βήματος (που εξαρτάται από το συγκεκριμένο πρόβλημα και την ακρίβεια της αριθμητικής) τέτοια ώστε γιά $h<h_0$ το σφάλμα όχι μόνο δεν ελαττώνεται αλλά αυξάνεται!

Ασκήσεις 3.1

1. (α) Θεωρείστε γιά $\varepsilon>0$ σταθερό το πρόβλημα (στον $\mathbb{R}^1$)

$$\begin{cases} y' = |y|^{1+\varepsilon}, & 0\leq t\leq b, \\ y(0) = 1 \end{cases}$$

Δείξτε ότι το πρόβλημα δεν έχει λύση αν $b\geq\varepsilon^{-1}$. Γιατί δεν ισχύει το Θεώρημα 1; Ποιό είναι το μέγιστο διάστημα $[0,\delta]$ ύπαρξης λύσης που δίνει το Θεώρημα 1', π.χ. γιά $\varepsilon=1$; είναι η λύση μοναδική εκεί;

(β) Θεωρείστε γιά $0<\varepsilon<1$ σταθερό το πρόβλημα

$$\begin{cases} y' = |y|^{1-\varepsilon}, & 0\leq t\leq b, \\ y(0) = 0 \end{cases}$$

Δείξτε ότι το πρόβλημα δεν έχει μοναδική λύση σε οποιοδήποτε διάστημα $[0,b]$. Γιατί δεν ισχύει το Θεώρημα 1'; Ποιά λύση προσεγγίζει πάντα η μέθοδος του Euler γιά το πρόβλημα; (Δίδαγμα: η συνθήκη Lipschitz $\|f(t,y)-f(t,y^*)\| \leq L\|y-y^*\|$ δεν μπορεί να αντικατασταθεί με συνθήκη της μορφής $\|f(t,y) - f(t,y^*)\| \leq L\|y-y^*\|^{\alpha}$, $\alpha\neq 1$).

2. Υποθέστε ότι το πρόβλημα (1) στον $\mathbb{R}^1$ έχει μοναδική λύση $y(t) \in C^2[a,b]$ και υποθέστε ότι αυτή της (2) ισχύει η συνθήκη Lipschitz

$$|f(t,y)-f(t,y^*)| \leq L\ |y-y^*|\ \forall t\in[a,b],\ y,y^*\in M_\delta,$$

όπου $\delta>0$ και όπου $M_\delta=[m_1-\delta,\ m_2+\delta]$ με $m_1\leq y(t)\leq m_2$ γιά $t\in[a,b]$, δηλ. ότι έχουμε συνθήκη Lipschitz μόνο σε μιά περιοχή M_δ του πεδίου τιμών $M_0=[m_1,m_2]$ της <u>λύσης</u> $y(t)$ του (1) (μία πολύ χρήσιμη και ρεαλιστική συνθήκη). Δείξτε ότι υπάρχει $h_0>0$ τέτοιο ώστε γιά $0<h\leq h_0$, η μέθοδος του Euler γιά το πρόβλημα δίνει προσεγγίσεις $\{y^n\}$ γιά τις οποίες ισχύει το ανάλογο της εκτίμησης (6).

3. Θεωρείστε ένα <u>μη</u> ομοιόμορφο διαμερισμό $a=t_0<t_1<\dots<t_N=b$ και υποθέστε ότι αν $h_n=t^{n+1}-t^n$, $0\le n\le N-1$, είναι το μεταβλητό βήμα, τότε $\min_n h_n \ge \lambda \max_n h_n$ γιά κάποια σταθερά $\lambda>0$ ανεξάρτητη του n. Δείξτε ένα φράγμα του σφάλματος της μεθόδου του Euler ανάλογο του (6) όπου $h=\max_n h_n$.

4. Αποδείξτε την (10) υπο τις προϋποθέσεις της Παρατήρησης 2.

5. Θεωρείστε το 2x2 σύστημα

$$\begin{cases} x' = -y \\ y' = x \qquad t\ge 0 \\ x(0)=1, \quad y(0)=0, \end{cases}$$

που έχει προφανώς την μοναδική λύση $x(t)=\cos t$, $y(t)=\sin t$, που ικανοποιεί την σχέση ("νόμο διατήρησης")

$$(*) \quad x^2(t)+y^2(t) = 1, \quad t\ge 0.$$

(α) Θεωρείστε την μέθοδο του Euler γιά το σύστημα (με σταθερό βήμα $h>0$). Δείξτε ότι δεν ικανοποιεί <u>το διακριτό ανάλογο</u> της (*), δηλ. ότι $(x^n)^2+(y^n)^2 \neq 1$ αν $n>1$. (Μάλιστα δείξτε ότι, γιά σταθερό h, $\lim_{n\to\infty} [(x^n)^2+(y^n)^2]=\infty$!. Σχεδιάστε στο επίπεδο (x,y) τα σημεία (x^n,y^n) γιά μερικές τιμές του n).

(β) Διερευνήστε αν υπάρχει διακριτό ανάλογο της (*) γιά την μέθοδο

$$\begin{cases} (x^{n+1}-x^n)/h = -y^n, \quad n\ge 0 \\ (y^{n+1}-y^n)/h = x^{n+1}, \quad n\ge 0 \\ x^0=1, \quad y^0=0. \end{cases}$$

(γ) 'Ιδιο ερώτημα γιά την μεθοδο ("πεπλεγμένη Euler")

$$\begin{cases} (x^{n+1}-x^n)/h = -y^{n+1}, & n\geq 0 \\ (y^{n+1}-y^n)/h = x^{n+1}, & n\geq 0 \\ x^0=1,\ y^0=0 \end{cases}$$

(δ) 'Ιδιο ερώτημα γιά την μέθοδο ("τραπεζίου")

$$\begin{cases} (x^{n+1}-x^n)/h = -(y^n+y^{n+1})/2 & n\geq 0 \\ (y^{n+1}-y^n)/h = (x^n+x^{n+1})/2 & n\geq 0 \\ x^0=1,\ \ y^0=0 \end{cases}$$

(Παρατηρείστε ότι οι μέθοδοι (γ) και (δ) απαιτούν την επίλυση γραμμικού συστήματος γιά κάθε βήμα n - είναι παραδείγματα "πεπλεγμένων" μεθόδων. Σ' όλα τα ερωτήματα θεωρείστε το $h>0$ σταθερό και διερευνήστε την συμπεριφορά του $(x^n)^2+(y^n)^2$ καθώς αυξάνεται το n).

3.2-ΜΕΘΟΔΟΙ RUNGE-KUTTA

Στο κεφάλαιο αυτό θα ασχοληθούμε με μία κατηγορία αριθμητικών μεθόδων γιά το πρόβλημα αρχικών τιμών (3.1.1), με τις λεγόμενες μεθόδους Runge-Kutta (RK)*. Σκοπός μας θα είναι να κατασκευάσουμε και να αναλύσουμε μεθόδους ανώτερης τάξης ακρίβειας, δηλ. μεθόδους γιά τις οποίες

(1) $\max_{0\leq n\leq N} \|y^n - y(t^n)\| = O(h^p)$, $p>1$.

Ο αριθμός p λέγεται τάξη ακρίβειας της μεθόδου. Γιά την απόδειξη εκτιμήσεων όπως η (1) υποθέτουμε ότι η λύση $y(t)$ του (3.1.1) είναι αρκετά ομαλή και ότι η μέθοδός μας υπολογίζει προσεγγίσεις y^n των τιμών $y(t^n)$ στους κόμβους $t^n=a+nh$, $0\leq n\leq N$ ενός ομοιόμορφου διαμερισμού με βήμα $h=(b-a)/N$. Στην (1) ο συμβολισμός $O(\cdot)$ ερμηνεύεται ως εξής. Έστω $\varepsilon:[0,h_0]\to R^m$. Λέμε ότι $\varepsilon(h)=O(h^a)$ (όταν $h\to 0$) αν υπάρχει σταθερά C ανεξάρτητη του h τέτοια ώστε

(2) $\|\varepsilon(h)\| \leq Ch^a$ γιά $0\leq h\leq h_0$.

Παραδείγματος χάριν γιά το σφάλμα της μεθόδου του Euler δείξαμε ότι ισχύει η (1) με p=1, αν $y\in C^2[a,b]$ (με $C^k[a,b]$ θα συμβολίζουμε τον χώρο των συναρτήσεων $y:[a,b]\to R^m$ με k συνεχείς παραγώγους στο [a,b]) και ότι η τάξη ακρίβειας δεν αυξάνει αν η y είναι περισσότερο ομαλή.

Οι μέθοδοι Runge-Kutta προκύπτουν κατά συστηματικό τρόπο από κατάλληλη εφαρμογή κανόνων αριθμητικής ολοκλήρωσης στην Δ.Ε. του (3.1.1) στο διάστημα $[t^n,t^{n+1}]$. Ολοκληρώνοντας π.χ. την Δ.Ε. του (3.1.1) ως πρός t έχουμε την ακριβή σχέση:

---

* Στους Runge (1985) και Kutta (1901) οφείλεται η λεγόμενη "κλασική" μέθοδος Runge-Kutta (21α). Σήμερα η κατηγορία των μεθόδων που ονομάζουμε RK έχει επεκταθεί σημαντικά.

$$(3)\quad y(t^{n+1})-y(t^n) = \int_{t^n}^{t^{n+1}} f(t,y(t))dt$$

Η αντικατάσταση του ολοκληρώματος στο δεύτερο μέλος της (3) από το "εμβαδόν του ορθογωνίου" πλάτους $h=t^{n+1}-t^n$ και "ύψους" $f(t^n,y(t^n))$ δίνει την προσεγγιστική σχέση $y(t^{n+1})-y(t^n)\cong hf(t^n, y(t^n))$ που οδηγεί στην γνωστή μας μέθοδο του Euler. Αν ως "ύψος" του ορθογωνίου χρησιμοποιηθεί η τιμή $f(t^{n+1},y(t^{n+1}))$ παίρνουμε την λεγόμενη "πεπλεγμένη" (ή "οπισθοδρομική") μέθοδο του Euler

$$(4)\quad y^{n+1} = y^n+hf(t^{n+1},y^{n+1}),\ n\geq 0,$$

(θα παίρνουμε πάντα $y^0=y(a)=y_0$) που σε κάθε βήμα της απαιτεί την λύση ενός mxm μη γραμμικού συστήματος γιά το άγνωστο διάνυσμα y^{n+1}. Το σύστημα αυτό, γιά h αρκετά μικρό, έχει πάντα μοναδική λύση (βλ. Παρατήρηση 1 και Πρόταση 1 παρακάτω). Εφαρμόζοντας τώρα τον κανόνα του <u>τραπεζίου</u> στο β' μέλος της (3) παίρνουμε την (επίσης πεπλεγμένη) <u>μέθοδο του τραπεζίου</u>

$$(5)\quad y^{n+1}= y^n+h[f(t^{n+1}, y^{n+1})+f(t^n,y^n)]/2,\ n\geq 0.$$

Εξ άλλου η εφαρμογή του κανόνα "του μέσου" δίνει στην (3) την προσεγγιστική σχέση

$$y(t^{n+1})-y(t^n) \cong hf(t^n+h/2,\ \ y(t^n+h/2))$$

όπου η τιμή $y(t^n+h/2)$ μπορεί πάλι με κανόνα ορθογωνίου (ή μέθοδο Euler!) στο διάστημα $[t^n,t^n+h/2]$ να προσεγγισθεί ως $y(t^n+h/2)\cong y(t^n)+hf(t^n,y(t^n))/2$. Συνεπώς η εφαρμογή <u>δύο</u> κανόνων αριθμητικής ολοκλήρωσης, του ενός στο $[t^n,t^{n+1}]$ και του άλλου στο $[t^n,t^n+h/2]$ οδηγεί στην λεγόμενη "βελτιωμένη μέθοδο του Euler" ή "<u>άμεση μέθοδο του μέσου</u>"

$$(6)\quad \begin{cases} \tilde{y}^n = y^n + hf(t^n,y^n)/2 \\ y^{n+1} = y^n+hf(t^n+h/2,\tilde{y}^n) \end{cases}$$

που απαιτεί τον υπολογισμό της ενδιάμεσης ποσότητας $\tilde{y}^n$ και δύο υπολογισμούς της f ανά βήμα. Η (6) είναι παράδειγμα μιάς <u>άμεσης</u> μεθόδου, δηλ. μιάς μεθόδου που υπολογίζει το y^{n+1} από το y^n με αντικατάσταση, δηλ. χωρίς να απαιτεί την επίλυση μη γραμμικών συστημάτων.

Δεν είναι δύσκολο να αποδειχθεί (βλ. Ασκ. 2) ότι η (6) έχει τάξη ακρίβειας p=2, δηλ. ότι ισχύει $\max_n \|y^n-y(t^n)\|=O(h^2)$. Θα δούμε αργότερα ότι η πεπλεγμένη μέθοδος του Euler είναι πρώτης τάξης ενώ η μέθοδος του τραπεζίου δευτέρας. Πρός το παρόν προχωρούμε στον ορισμό της γενικής μεθόδου RK.

Έστω τ_i, $1\le i\le q$ q αριθμοί (συνήθως $0\le\tau_i\le1$) και έστω ότι οι <u>κόμβοι</u> τ_i και τα <u>βάρη</u> (συντελεστές) a_{ij}, $1\le i,j\le q$, b_i, $1\le i\le q$, ορίζουν τους q+1 <u>κανόνες αριθμητικής ολοκλήρωσης</u>:

$$(7)\quad \int_0^{\tau_i} \psi(s)ds \cong \sum_{j=1}^{q} a_{ij}\, \psi(\tau_j),\ 1\le i\le q,$$

$$(8)\quad \int_0^{1} \psi(s)ds \cong \sum_{j=1}^{q} b_j\, \psi(\tau_j).$$

Ορίζουμε τώρα γιά το πρόβλημα (3.1.1) $t^n=a+nh$, κατά τα γνωστά, και τα σημεία (γιά $0\le n\le N-1$)

$$(9)\quad t^{n,i} = t^n + \tau_i h,\ 1\le i\le q.$$

Ολοκληρώνοντας την Δ.Ε του (3.1.1) ως πρός t από t^n μέχρι $t^{n,i}$ έχουμε γιά $1\le i\le q$:

$$(10)\ y(t^{n,i})-y(t^n) = \int_{t^n}^{t^n+\tau_i h} f(t,y(t))dt = (\text{με αλλαγή μεταβλητών}$$

$$t \mapsto s,\ t=t^n+sh,\ 0\leq s\leq \tau_i) = h \int_0^{\tau_i} f(t^n+sh,y(t^n+sh))ds \cong (\text{με εφαρμογή}$$

του κανόνα ολοκλήρωσης (7) και χρήση των (9))

$\cong h \sum_{j=1}^{q} a_{ij}\ f(t^{n,j},y(t^{n,j}))$. Εντελώς ανάλογα, ολοκληρώνοντας την δ.Ε. του (3.1.1) από t^n μέχρι t^{n+1} έχουμε (χρησιμοποιώντας τις (8) και (9)) ότι

$$(11)\ y(t^{n+1})-y(t^n) \cong h \sum_{j=1}^{q} b_j f(t^{n,j},y(t^{n,j})).$$

Οι προσεγγιστικές σχέσεις (10), (11) ορίζουν την λεγόμενη <u>γενική μέθοδο Runge-Kutta με q</u> (ενδιάμεσα) <u>στάδια</u> γιά τον υπολογισμό της y^{n+1} αν είναι γνωστή η y^n, $0\leq n\leq N-1$, (με $y^0=y_0$):

$$(12\alpha)\ y^{n,i} = y^n + h \sum_{j=1}^{q} a_{ij}\ f(t^{n,j},y^{n,j}),\ 1\leq i\leq q,$$

$$(12\beta)\ y^{n+1} = y^n + h \sum_{j=1}^{q} b_j\ f(t^{n,j},y^{n,j}),$$

όπου τα $t^{n,i}$ έχουν ορισθεί από τις (9). Τα (ενδιάμεσα) "στάδια" είναι τα q διανύσματα $y^{n,i} \in \mathbb{R}^m$, $1\leq i\leq q$, τα οποία υπολογίζονται ως λύση του (μη γραμμικού γενικά) mq x mq συστήματος που παριστάνουν οι σχέσεις (12α). Στην πράξη συνήθως η μέθοδος γράφεται ισοδύναμα στην εξής μορφή: θέτουμε $k^{n,i}=f(t^{n,i},y^{n,i})$, $1\leq i\leq q$ οπότε οι (12α-β) γίνονται

$$(13)\quad \begin{aligned} k^{n,i} &= f(t^{n,i},y^n+ h \sum_{j=1}^{q} a_{ij}\ k^{n,j}),\quad 1\leq i\leq q,\\ y^{n+1} &= y^n + h \sum_{j=1}^{q} b_j\ k^{n,j}. \end{aligned}$$

Συνεπώς η γενική μέθοδος RK με q στάδια ορίζεται μέσω των q^2+2q σταθερών a_{ij}, b_j, τ_i πού συνήθως διατάσσουμε στο μητρώο*

$$(14) \qquad \begin{array}{ccc|c} a_{11} & \ldots & a_{1q} & \tau_1 \\ \vdots & & \vdots & \vdots \\ a_{q1} & \ldots & a_{qq} & \tau_q \\ \hline b_1 & \ldots & b_q & \end{array}$$

ή σε συντομογραφία στο μητρώο

$$(14') \qquad \begin{array}{c|c} A & \tau \\ \hline b & \end{array}$$

όπου $A=(a_{ij})\epsilon\ R^{q\times q}$, $\tau=(\tau_i)$, $b=(b_i)\ \epsilon\ R^q$. Αν ο πίνακας A είναι (πιθανώς μετά από αναδιάταξη γραμμών) <u>αυστηρά</u> κάτω τριγωνικός, δηλ. αν $a_{ij}=0$ γιά $i\leq j$, τότε η μέθοδος είναι <u>άμεση</u>, δηλ. δεν χρειάζεται την λύση μη γραμμικού συστήματος αφού τότε τα ενδιάμεσα στάδια υπολογίζονται με απλή αντικατάσταση:

$$\begin{aligned} y^{n,1} &= y^n \\ y^{n,2} &= y^n + h\, a_{21}\, f(t^{n,1}, y^{n,1}) \\ &\vdots \\ y^{n,q} &= y^n + h \sum_{j=1}^{q-1} a_{qj}\, f(t^{n,j}, y^{n,j}) \end{aligned}$$

Σε κάθε άλλη περίπτωση η μέθοδος είναι <u>πεπλεγμένη</u> και ο υπολογισμός των $y^{n,i}$ απαιτεί επίλυση μη γραμμικών συστημάτων. Σημαντικές μεταξύ των πεπλεγμένων είναι οι λεγόμενες <u>ημιπεπλεγμένες</u> μέθοδοι γιά τις

---

* Ο φορμαλισμός των μεθόδων οφείλεται στου J.C. Butcher: (Math. Comp. 18(1964), 50-64 και 233-244, ibid. 26 (1972), 79-106 κ.ά.)

οποίες ο πίνακας A είναι κάτω τριγωνικός , δηλ. έχει $a_{ij}=0$, $i<j$. Τότε τα στάδια υπολογίζονται από τις σχέσεις

$$y^{n,1} = y^n + h\, a_{11}\, f(t^{n,1},y^{n,1})$$
$$y^{n,2} = y^n + h\, a_{21}\, f(t^{n,1},y^{n,1}) + h\, a_{22}\, f(t^{n,2},y^{n,2})$$
$$y^{n,q} = y^n + h \sum_{j=1}^{q} a_{qj}\, f(t^{n,j},y^{n,j}).$$

Τα μη χραμμικά συστήματα είναι τώρα <u>αποσυνδεδεμένα</u>: Η πρώτη εξίσωση δίνει το $y^{n,1}$ ως λύση ενός mxm μη χραμμικού συστήματος και γενικά η i-στή δίνει το $y^{n,i}$ ως λύση του mxm μη χραμμικού συστήματος (αν $a_{ii}\neq 0$)

$$y^{n,i} = h\, a_{ii}\, f(t^{n,i},y^{n,i}) + g^{n,i}$$

όπου $g^{n,i}$ ήδη χνωστό διάνυσμα. (Η επίλυση q mxm συστημάτων είναι προτιμότερη από την επίλυση ενός qm x qm)

Ας δούμε τώρα παραδείχματα μεθόδων RK. Κατ' αρχήν τα μητρώα (χιά q=1 στάδιο)

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \quad , \quad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

περιχράφουν, αντίστοιχα, την μέθοδο του Euler (3.1.3) και την πεπλεχμένη μέθοδο του Euler (4). Το μητρώο (q=1)

$$(15) \qquad \begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array}$$

περιχράφει στην μορφή (12) την (πεπλεχμένη) <u>μέθοδο του μέσου</u>

$$(15')\quad y^{n+1} = y^n + hf(t^n+h/2,\ (y^n+y^{n+1})/2),$$

(της οποίας η άμεση μέθοδος του μέσου (6) αποτελεί "χραμμικοποίηση". Η τάξη ακρίβειας της (15) είναι p=2)

Γιά $q=2$ τώρα, παρατηρούμε ότι η μέθοδος του τραπεζίου (5) περιγράφεται από το μητρώο

$$(16)\qquad \begin{array}{cc|c} 0 & 0 & 0 \\ 1/2 & 1/2 & 1 \\ \hline 1/2 & 1/2 & \end{array}$$

ενώ η άμεση μέθοδος του μέσου (6) από το

$$(17)\qquad \begin{array}{cc|c} 0 & 0 & 0 \\ 1/2 & 0 & 1/2 \\ \hline 0 & 1 & \end{array}$$

Μία ενδιαφέρουσα οικογένεια <u>ημιπεπλεγμένων</u> μεθόδων με $q=2$ στάδια δίνεται από το μονοπαραμετρικό μητρώο γιά $\lambda\in\mathbb{R}$

$$(18)\qquad \begin{array}{cc|c} \lambda & 0 & \lambda \\ 1-2\lambda & \lambda & 1-\lambda \\ \hline 1/2 & 1/2 & \end{array}$$

Οι μέθοδοι που περιγράφονται από το (18) είναι γενικά τάξης $p=2$. Οι τιμές $\lambda=(1\pm 3^{-1/2})/2$ δίνουν τις πολύ ενδιαφέρουσες μεθόδους "(2,3) DIRK", τάξης $p=3$. Ιδιαίτερα ενδιαφέρουσα είναι η μέθοδος με $q=2$ στάδια που δίνεται από το μητρώο

$$(19)\qquad \begin{array}{cc|c} 1/4 & 1/4-\mu & 1/2-\mu \\ 1/4+\mu & 1/4 & 1/2+\mu \\ \hline 1/2 & 1/2 & \end{array}$$

όπου $\mu=1/2\sqrt{3}$ είναι η λεγόμενη μέθοδος "Gauss-Legendre δύο σημείων" (Τα $\tau_i=1/2\pm\mu$ της (19) είναι οι κόμβοι - και τα $b_i=1/2$ οι αντίστοιχοι συντελεστές - του κανόνα ολοκλήρωσης Gauss με 2 σημεία με συνάρτηση βάρους $w(x)=1$ στο $[0,1]$). Η (19) είναι η μόνη μέθοδος με $q=2$ στάδια που έχει τάξη ακρίβειας $p=4$. Έχει επίσης πολλές άλλες ενδιαφέρουσες ιδιότητες.

Μερικές κλασικές <u>άμεσες</u> μέθοδοι RK δίνονται από το μητρώα

$$(20\alpha)\quad \begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ -1 & 2 & 0 & 1 \\ \hline 1/6 & 2/3 & 1/6 & \end{array}\ ,\quad (20\beta)\quad \begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/3 \\ 0 & 2/3 & 0 & 2/3 \\ \hline 1/4 & 0 & 3/4 & \end{array}$$

(και λέγονται, αντίστοιχα, μέθοδοι "Kutta τρίτης τάξης" και "Heun τρίτης τάξης" - και οι δύο με q=p=3). Η κλασική <u>μέθοδος Runge-Kutta</u> είναι η άμεση μέθοδος με q=p=4 που δίνεται από το μητρώο (21α) ενώ η μέθοδος (21β) έχει επίσης q=p=4:

$$(21\alpha)\quad \begin{array}{cccc|c} 0 & & O & & 0 \\ 1/2 & 0 & & & 1/2 \\ 0 & 1/2 & 0 & & 1/2 \\ 0 & 0 & 1 & 0 & 1 \\ \hline 1/6 & 1/3 & 1/3 & 1/6 & \end{array}\quad (21\beta)\quad \begin{array}{cccc|c} 0 & & & O & 0 \\ 1/3 & 0 & & & 1/3 \\ -1/3 & 1 & 0 & & 2/3 \\ 1 & -1 & 1 & 0 & 1 \\ \hline 1/8 & 3/8 & 3/8 & 1/8 & \end{array}$$

Προφανώς μας ενδιαφέρει με όσο μικρότερο δυνατό αριθμό βημάτων q να επιτυγχάνουμε όσο το δυνατό μεγαλύτερη τάξη ακρίβειας p. Όπως θα δούμε υπάρχει περιορισμός: η μέγιστη τάξη ακρίβειας μιάς μεθόδου q σταδίων είναι p=2q. (Οι μέθοδοι που έχουν αυτήν την ιδιότητα είναι (πλήρως) πεπλεγμένες και γενικεύουν γιά q>2 την (19)). Γιά άλλα παραδείγματα μεθόδων RK βλ. τα βιβλία [3.1]-[3.4], [3.6], τις εργασίες του Butcher (op.cit) και των συνεργατών του, τις εργασίες του Crouzeix*, και την βιβλιογραφία του [3.1]

Προχωρούμε τώρα στην ανάλυση των μεθόδων RK. Αρχίζουμε με την απόδειξη ότι γιά πεπλεγμένες μεθόδους υπάρχει λύση του μη γραμμικού συστήματος που ορίζει τα $y^{n,i}$ (και συνεπώς και το y^{n+1}) συναρτήσει του y^n. Γιά ό,τι επακολουθεί υποθέτουμε σιωπηρά ότι ισχύουν οι υποθέσεις του Θεωρήματος 3.1.1 ύπαρξης-μοναδικότητας λύσεως του προβλήματος αρχικών τιμών (3.1.1)

---

* Ιδιαίτερα βλ. M. Crouzeix, Thèse, Paris VI, 1975 και Num. Math. 32(1979), 75-82.

ΠΡΟΤΑΣΗ 1 Θεωρούμε μία πεπλεγμένη μέθοδο RK της μορφής (12). 'Εστω $|A|$ ο πίνακας $(|a_{ij}|)$ των απολύτων τιμών των στοιχείων a_{ij} του (14) και έστω $\rho(|A|)$ η φασματική του ακτίνα. Τότε γιά κάθε h τέτοιο ώστε

$$(22)\quad h \le h_0 < (L\,\rho(|A|))^{-1}$$

το σύστημα (12α) έχει μοναδική λύση $(y^{n,i})$, $1\le i\le q$.

<u>Απόδειξη</u>: Θεωρείστε την απεικόνιση $F:(R^m)^q \to (R^m)^q$, που ορίζεται γιά $\zeta=(\zeta_1,\dots,\zeta_q)^T \in (R^m)^q$ (δηλ. $\zeta_i\in R^m$, $1\le i\le q$) από

$$\zeta = \begin{pmatrix}\zeta_1\\ \zeta_2\\ \vdots\\ \zeta_q\end{pmatrix} \mapsto F(\zeta) = \begin{pmatrix} y^n + h\sum_{j=1}^{q} a_{1j} f(t^{n,j},\zeta_j)\\ \vdots\\ y^n + h\sum_{j=1}^{q} a_{qj} f(t^{n,j},\zeta_j)\end{pmatrix}$$

Γιά την i-στη "συνιστώσα" (διάνυσμα στου R^m) του $F(\zeta)$ έχουμε λοιπόν λόγω της συνθήκης Lipschitz στην f ότι γιά κάθε $\zeta,\zeta^*\in(R^m)^q$

$$\|(F(\zeta))_i - (F(\zeta^*))_i\| \le Lh \sum_{j=1}^{q} |a_{ij}|\,\|\zeta_j-\zeta^*_j\|,\quad 1\le i\le q.$$

Γιά $x\in(R^m)^q$ έστω $[x]\in R^q$ το διάνυσμα με συνιστώσες $(\|x_1\|,\dots,\|x_q\|)$ όπου $x=(x_1,\dots,x_q)^T$, $x_i\in R^m$. Τότε έχουμε (γιά διανύσματα $u\le v \Leftrightarrow u_i\le v_i\ \forall i$)

$$[F(\zeta)-F(\zeta^*)] \le Lh|A|\,[\zeta-\zeta^*]\quad \forall \zeta,\zeta^*\in(R^m)^q.$$

'Εστω τώρα $F^2(\zeta)=F(F(\zeta))$ και γενικά $F^\nu(\zeta)=F(F^{\nu-1}(\zeta))$. Τότε η παραπάνω σχέση δίνει εφαρμοζόμενη κατ' επανάληψιν ότι

$$(23)\quad [F^{\nu}(\zeta) - F^{\nu}(\zeta^*)] \leq (Lh|A|)^{\nu}\, [\zeta-\zeta^*] \qquad \forall \zeta,\zeta^*\epsilon(\mathbb{R}^m)^q$$

Η συνθήκη (22) λέει τώρα ότι $hL\rho(|A|)=\rho(hL|A|)<1$. Συνεπώς η ακολουθία πινάκων $(Lh|A|)^{\nu}$ τείνει στο 0 (γιατί;) καθώς $\nu\to\infty$. Άρα η (23) δίνει ότι γιά κάποιο ν_o αρκετά μεγάλο όλα τα στοιχεία του πίνακα $(Lh|A|)^{\nu_o}$ θα γίνουν μικρότερα από α/q όπου $\alpha=\alpha(\nu_o)<1$. Συνεπώς η (23) δίνει ότι $\exists\alpha<1$, $\nu_o\epsilon\mathbb{N}$ τέτοια ώστε

$$\|(F^{\nu_o}(\zeta))_i-(F^{\nu_o}(\zeta^*))_i\| \leq \alpha/q \sum_{i=1}^{q} \|\zeta_i-\zeta^*_i\| \Rightarrow$$

$$\sum_{i=1}^{q} \|(F^{\nu_o}(\zeta))_i-F^{\nu_o}(\zeta^*))_i\| \leq \alpha \sum_{i=1}^{q} \|\zeta_i-\zeta^*_i\|$$

δηλ. ότι η απεικόνιση F^{ν_o} είναι <u>συστολή</u>, ως πρός την νόρμα

$|||\cdot||| : |||\zeta||| = \sum_{i=1}^{q} \|\zeta_i\|$ του $(\mathbb{R}^m)^q$ σ' όλο του $(\mathbb{R}^m)^q$. Έπεται ότι η απεικόνιση $\zeta \mapsto F(\zeta)$ έχει μοναδικό σταθερό σημείο στον $(\mathbb{R}^m)^q$, (γιατί;), δηλ. ότι το σύστημα (12α) έχει μοναδική λύση $\{y^{n,i}\}$, $1\leq i\leq q$. @

Θα αποδείξουμε τώρα ένα βασικό αποτέλεσμα <u>ευστάθειας</u> των μεθόδων RK:

ΠΡΟΤΑΣΗ 2. Έστω ότι ισχύουν οι υποθέσεις της πρότασης 1. Με $y^o\epsilon\mathbb{R}^m$ δεδομένο θεωρούμε τις προσεγγίσεις $\{y^n\}$, $0\leq n\leq N$ που παράγει η μέθοδος RK (12α-b) γιά $n=0,1,2,\ldots,N-1$. Θεωρούμε επίσης τα διανύσματα $z^{n,i}$, $0\leq n\leq N-1$, $1\leq i\leq q$ και z^n, $0\leq n\leq N$ του $\mathbb{R}^m$ που ορίζουν οι εξισώσεις:

$z^o\epsilon\mathbb{R}^m$ δεδομένο. Γιά $n=0,1,\ldots,N-1$:

$$(24\alpha)\quad z^{n,i}=z^n+h \sum_{j=1}^{q} \alpha_{ij}\, f(t^{n,j},z^{n,j}),\quad 1\leq i\leq q,$$

$$(24\beta)\quad z^{n+1}=z^n+h\sum_{j=1}^{q} b_j\, f(t^{n,j},z^{n,j})+\epsilon^n,$$

που προφανώς έχουν επίσης μοναδική λύση γιά $0\leq n\leq N-1$ (Το σύστημα (24α,β) είναι μία "διαταραχή" του συστήματος (12α,β)). Τότε ισχύει ότι

$$(25)\quad \max_{0\leq n\leq N}\|y^n-z^n\|\leq C_1\,\|y^0-z^0\|+C_2h^{-1}\max_{0\leq n\leq N}\|\epsilon^n\|$$

όπου οι σταθερές C_1 και C_2 είναι ανεξάρτητες του h.

<u>Απόδειξη</u>: Χρησιμοποιούμε τον υπολογισμό της απόδειξης της προηγούμενης πρότασης. Αφαιρώντας κατά μέλη τις (24α), (12α) και χρησιμοποιώντας την συνθήκη Lipschitz γιά την f έχουμε

$$\|y^{n,i}-z^{n,i}\|\leq\|y^n-z^n\|+hL\sum_{j=1}^{q}|a_{ij}|\,\|y^{n,j}-z^{n,j}\|,\quad 1\leq i\leq q.$$

'Εστω $Y^n=(y^{n,1},\ldots,y^{n,q})^T$, $Z^n=(z^{n,1},\ldots,z^{n,q})^T$ διανύσματα του $(\mathbb{R}^m)^q$ με "συντεταγμένες" $y^{n,i}$, $z^{n,i}\in\mathbb{R}^m$ αντίστοιχα. Η παραπάνω σχέση (θυμόμαστε ότι το $[Y^n]$ είναι το διάνυσμα $(\|y^{n,1}\|,\ldots,\|y^{n,q}\|)^T$ του $\mathbb{R}^q$) γράφεται ως:

$$(26)\quad [Y^n-Z^n]\leq\|y^n-z^n\|\,u+hL\,|A|\,[Y^n-Z^n],$$

όπου $u=(1,1,\ldots,1)^T\in\mathbb{R}^q$. Η (αυτο-) αναδρομική ανισότητα (26) δίνει τώρα εφαρμοζόμενη κατ' επανάληψιν γιά κάθε $\nu\in\mathbb{N}$, (όπου I_q είναι η ταυτότητα στον $\mathbb{R}^q$):

$$[Y^n-Z^n]\leq\|y^n-z^n\|\,(I_q+hL|A|+\ldots+(hL)^\nu|A|^\nu)\,u+(hL)^{\nu+1}|A|^{\nu+1}[Y^n-Z^n]$$

απ' την οποία, χρησιμοποιώντας την $h\leq h_0$ και παίρνοντας το όριο γιά $\nu\to\infty$ έχουμε (γιατί;)

$$[Y^n-Z^n]\leq\|y^n-z^n\|\,(I_q-h_0L|A|)^{-1}\,u$$

Συμπεραίνουμε ότι υπάρχει σταθερά C ανεξάρτητη των h,n τέτοια ώστε

$$(27)\ \|y^{n,i}-z^{n,i}\| \le C\|y^n-z^n\|,\ 1\le i\le q,\ 0\le n\le N-1.$$

Τώρα, αφαιρώντας τις (12β),(24β) κατά μέλη έχουμε γιά $0\le n\le N-1$

$$(28)\ \|y^{n+1}-z^{n+1}\| \le \|y^n-z^n\| + Lh\sum_{j=1}^{q} |b_j|\,\|y^{n,j}-z^{n,j}\| + \|\epsilon^n\|$$

Αντικαθιστώντας την (27) στην (28) έχουμε με $C'=LC\sum_{j=1}^{q}|b_j|$

$$\|y^{n+1}-z^{n+1}\| \le (1+C'h)\,\|y^n-z^n\| + \|\epsilon^n\|,\ 0\le n\le N-1,$$

από την οποία και το Λήμμα 3.1.1 παίρνουμε ότι

$$\max_n \|y^n-z^n\| \le e^{C'Nh}\|y^0-z^0\| + \max_n \|\epsilon^n\|\,(e^{C'Nh}-1)/C'h,$$

απ' όπου προκύπτει η (25) με $C_1=e^{C'(b-a)}$, $C_2=(e^{C'(b-a)}-1)/c'$ @.

Η Πρόταση 2 ("ευστάθεια") είναι ένα από τα δύο κύρια αποτελέσματα που χρειαζόμαστε γιά την απόδειξη ενός θεωρήματος "σύγκλισης" (και εκτίμησης του σφάλματος) γιά τις μεθόδους RK. Το άλλο αφορά την <u>τάξη (ακρίβειας)</u> της μεθόδου.

Υποθέτουμε ότι το δεύτερο μέλος $f(t,y)$, και συνεπώς και η λύση $y(t)$ του προβλήματος αρχικών τιμών (3.1.1), είναι αρκετά ομαλά. Θεωρούμε την μέθοδο RK (12). Γιά να ορίσουμε την τάξη (ακρίβειάς) της, ορίζουμε πρώτα γιά $0\le n\le N-1$ το $\delta^n\in R^m$ από τις εξισώσεις:

$$(29\alpha)\ \zeta^{n,i} = y(t^n) + h\sum_{j=1}^{q} a_{ij}\, f(t^{n,j},\zeta^{n,j}),\ 1\le i\le q,$$

$$(29\beta)\ \delta^n = y(t^{n+1})-y(t^n) - h\sum_{j=1}^{q} b_j\, f(t^{n,j},\zeta^{n,j}),$$

όπου υποθέτουμε βέβαια ότι $h \le h_0 < (Lp(|A|))^{-1}$ έτσι ώστε (βλ. Πρόταση 1) το σύστημα (29α) να έχει μοναδική λύση $(\zeta^{n,i})$, $1 \le i \le q$. Λέμε ότι η μέθοδος (12) έχει <u>τάξη (ακρίβειας) p</u> αν $\delta^n = O(h^{p+1})$, δηλ. αν υπάρχει σταθερά D, ανεξάρτητη του h ή N (που μπορεί να εξαρτάται όμως από τα δεδομένα και την λύση του προβλήματος (3.1.1) και την συγκεκριμένη μέθοδο RK) τέτοια ώστε γιά h αρκετά μικρό να ισχύει

$$(30)\quad \max_{0 \le n \le N} \|\delta^n\| \le D\, h^{p+1}.$$

Λέμε ότι η μέθοδος (12) είναι <u>συνεπής</u> αν η τάξη της p είναι τουλάχιστον 1. Τονίζουμε το γεγονός ότι οι ποσότητες $y(t^n)$, $y(t^{n+1})$ στις (29α-β) είναι τιμές της <u>λύσης</u> του προβλήματος (3.1.1). Η τάξη λοιπόν εκφράζει το κατά πόσον <u>ένα</u> βήμα της μεθόδου RK (12) με αρχική συνθήκη $y(t^n)$ δίνει μία καλή προσέγγιση της τιμής $y(t^{n+1})$.

Η χρήση της τάξης μιάς μεθόδου (δηλ. μιάς εκτίμησης του τύπου (30)) και η "ευστάθεια" της που αποδείχθη στην Πρόταση 2 μάς εξασφαλίζει ένα φράγμα $O(h^p)$ στο <u>σφάλμα</u> της μεθόδου, όπως μπορούμε εύκολα να διαπιστώσουμε:

ΘΕΩΡΗΜΑ 1 Έστω ότι ισχύουν οι υποθέσεις της Πρότασης 1. Με $y^0 = y_0 \in R^m$ θεωρούμε την μέθοδο RK (12) γιά την οποία υποθέτουμε ότι ισχύει η (30), δηλ. η (12) έχει τάξη p. Τότε, αν η λύση y(t) του (3.1.1) είναι αρκετά ομαλή, έχουμε

$$(31)\quad \max_{0 \le n \le N} \|y^n - y(t^n)\| \le Ch^p,$$

όπου $C = D(e^{c'(b-a)} - 1)/C''$ η σταθερά C' ορίστηκε στην απόδειξη της Πρότασης 2.

<u>Απόδειξη</u>: Οι εξισώσεις (29α,β) γράφονται, γιά $0 \le n \le N-1$:

$$\zeta^{n,i} = y(t^n) + h \sum_{j=1}^{q} a_{ij} f(t^{n,j},\zeta^{n,j}),\ 1\le i\le q$$

$$y(t^{n+1}) = y(t^n) + h \sum_{j=1}^{q} b_j f(t^{n,j},\zeta^{n,j}) + \delta^n,$$

δηλ. είναι της μορφής (24α,β) με $z^n\equiv y(t^n)$, $z^{n,i}\equiv\zeta^{n,i}$, $e^n\equiv\delta^n$. Συνεπώς η (25) και η υπόθεση ότι $y^0=y(t^0)\equiv z^0$ δίνουν

$$\max_{0\le n\le N} \|y^n-y(t^n)\| \le C_2\, h^{-1} \max_{0\le n\le N} \|\delta^n\|.$$

Η (31) προκύπτει τώρα από την παραπάνω ανισότητα, την μορφή της C_2 από την απόδειξη της Πρότασης 2, και την (30). Σημειώνουμε ότι η σταθερά C' εξαρτάται από την μέθοδο RK (δηλ. τις σταθερές a_{ij},b_j), το h_0 και την σταθερά Lipschitz L της f, ενώ η σταθερά D εξαρτάται από την μέθοδο RK και νόρμες παραγώγων της y και της f. @

Το Θεώρημα 1 εκφράζει ποσοτικά την γενική αρχή ότι "ευστάθεια"+"συνέπεια"=>"σύγκλιση". Η προσοχή μας στρέφεται λοιπόν τώρα στην διερεύνηση της τάξης ακρίβειας μιάς δεδομένης μεθόδου, δηλ. στην απόδειξη μιάς ανισότητας της μορφής (30).

Αν η μέθοδος είναι <u>άμεση</u>, τότε, με απλή αντικατάσταση, μπορούμε να απαλείψουμε τα $\zeta^{n,i}$ από τις (29α,β) και να εκφράσουμε το δ^n από την (29β) ως

$$(32)\quad \delta^n = y(t^{n+1})-y(t^n)-h\, \Phi(t^n,y(t^n),h),$$

όπου η Φ θα είναι μιά (πολύπλοκη γενικά) συνάρτηση. Αναπτύσσοντας τώρα κατά Taylor την $y(t^{n+1})-y(t^n)$ γύρω από το σημείο t^n, παίρνουμε μιά σειρά δυνάμεων του h επί παραγώγους $y^{(j)}(t^n)$. Οι όροι αυτοί (μέχρι και τάξεως h^p) θα πρέπει να απαλειφθούν από ανάλογους όρους του αναπτύγματος της συνάρτησης $\Phi(t^n,y(t^n),h)$ γύρω από το σημείο $(t^n,y(t^n))$.

'Ενα τέτοιο ανάπτυγμα της Φ δίνει σειρά δυνάμεων του h οι συντελεστές των οποίων εξαρτώνται από την f και τις μερικές παραγώγους της. Οι συντελεστές αυτοί θα πρέπει να εκφρασθούν ως παράγωγοι της y στο t^n μέσω των σχέσεων

$$y'=f,\ y''=D_t f = f_t + \sum_{i=1}^{m} f_{y_i} f_i \text{ κλπ.}$$

Βλ. π.χ. τις Ασκήσεις 2-5. Γιά πεπλεγμένες μεθόδους και h αρκετά μικρό πάντα μπορούμε βέβαια να γράφουμε την (29β) στην μορφή (32)· ο προσδιορισμός όμως της συνάρτησης Φ απαιτεί την απαλοιφή των $\zeta^{n,i}$ από τις (29α,β) δηλ. την αναλυτική επίλυση του μη γραμμικού συστήματος (29α) ως πρός τους αγνώστους $\zeta^{n,i}$. Αυτό βέβαια είναι αδύνατο γενικά' στην πραγματικότητα όμως δεν χρειαζόμαστε ένα τύπο γιά την $\Phi(t,y,h)$ αλλά μιά δυναμοσειρά της ως πρός h, την οποία μπορούμε να υπολογίσουμε (η δυσκολία αυξάνει εκθετικά με το q!) με διαδοχικές "συνθέσεις" δυναμοσειρών ενδιαμέσων ποσοτήτων (συνήθως των $k^{n,i}$, βλ. (13)). Βλ. π.χ. τις Ασκήσεις 6,7 γιά τις περιπτώσεις <u>απλών</u> Δ.Ε.

Το πρόβλημα της κατασκευής όλων των μεθόδων RK με ορισμένο αριθμό σταδίων q που έχουν δεδομένη τάξη p γίνεται εξαιρετικά πολύπλοκο και δύσκολο όσο αυξάνει το q. Στην 'Ασκηση 5 βρίσκουμε όλες τις <u>άμεσες</u> μεθόδους (γιά απλές Δ.Ε.) με $q \le 3$· τέτοιες μέθοδοι έχουν μέγιστη τάξη ακρίβειας $p=3$. Στην 'Ασκηση 7 λύσαμε το ανάλογο πρόβλημα γιά όλες τις μεθόδους RK με $q=2$ βρίσκοντας διαδοχικά όλο και περισσότερο περιοριστικές συνθήκες ώστε να έχουν τάξη $p=1,2,3,4$, αντίστοιχα. Γενικά, γιά οποιοδήποτε q είναι πολύ δύσκολο να διατυπωθούν εύκολα ελέγξιμες ικανές και αναγκαίες συνθήκες πάνω στους συντελεστές a_{ij}, b_i, τ_i έτσι ώστε η μέθοδος να έχει δεδομένη τάξη ακρίβειας p όταν εφαρμόζεται στο σύστημα (3.1.1) (βλ. όμως τις εργασίες του Butcher). Θα αποδείξουμε όμως τώρα ένα θεώρημα το οποίο δίνει <u>ικανές</u> συνθήκες, πάνω στους συντελεστές a_{ij}, b_i, τ_i της μεθόδου, οι οποίες εξασφαλίζουν ορισμένη τάξη ακρίβειας.

ΘΕΩΡΗΜΑ 2 (Butcher-Crouzeix). Υποθέτουμε ότι υπάρχουν ακέραιοι p, s, r≥0 τέτοιοι ώστε

$$(33)\quad \sum_{i=1}^{q} b_i \tau_i^k = 1/(k+1), \text{ γιά } 0\le k\le p-1$$

$$(34)\quad \sum_{j=1}^{q} a_{ij}\tau_j^k = \tau_i^{k+1}/(k+1), \; 1\le i\le q, \text{ γιά } 0\le k\le s-1.$$

$$(35)\quad \sum_{i=1}^{q} b_i \tau_i^k a_{ij} = b_j(1-\tau_j^{k+1})/k+1, \; 1\le j\le q, \text{ γιά } 0\le k\le r-1.$$

(36) $p\le r+s+1$ και $p\le 2s+2$.

Τότε η μέθοδος RK (12) έχει τάξη p όταν εφαρμοσθεί στο σύστημα (3.1.1) (στο οποίο υποθέτουμε ότι οι $y(t)$, $f(t,y)$ είναι αρκετά ομαλές).

<u>Απόδειξη</u>: Συμβολίζουμε γιά $k=0,1,2,\dots$, $0\le n\le N-1$

$$(37)\quad E_n(k) = \sum_{i=1}^{q} b_i \,(t^{n,i}-t^n)^k\,(f(t^{n,i},y(t^{n,i}))-f(t^{n,i},\zeta^{n,i})),$$

όπου $\{\zeta^{n,i}\}$ $1\le i\le q$ είναι η λύση του μη γραμμικού συστήματος (29α) - υποθέτουμε ότι υπάρχει και είναι μοναδική -. Από τον ορισμό (29β) του δ^n έχουμε λοιπόν ότι

$$(38)\quad \delta^n = y(t^{n+1})-y(t^n) - h\sum_{j=1}^{q} b_j\, f(t^{n,j},y(t^{n,j}))+hE_n(0) =$$

$$= y(t^{n+1})-y(t^n) - h\sum_{j=1}^{q} b_j\, y'(t^{n,j}) + hE_n(0).$$

Η υπόθεση (33) είναι ισοδύναμη με την δήλωση ότι ο κανόνας ολοκλήρωσης (8) ισχύει ως ισότητα γιά τις συναρτήσεις $\psi(t)=t^k$, $0\le k\le p-1$, δηλ. είναι <u>ακριβής</u> γιά πολυώνυμα βαθμού ≤ p-1. Από την θεωρία αριθμητικής ολοκλήρωσης (βλ. π.χ. [5.2]) έχουμε τότε ότι

$$y(t^{n+1})-y(t^n)-h\sum_{j=1}^{q} b_j y'(t^{n,j}) = \int_{t^n}^{t^{n+1}} y'(\tau)d\tau - h\sum_{j=1}^{q} b_j y'(t^{n,j}) =$$
$$= O(h^{p+1}).$$

'Αρα η (38) δίνει ότι

$$\delta^n = hE_n(0) + O(h^{p+1}).$$

Γιά να αποδείξουμε λοιπόν το θεώρημά μας αρκεί να δείξουμε ότι $E_n(0)=O(h^p)$. Το αποτέλεσμα αυτό είναι ειδική περίπτωση της σχέσης

(39) $E_n(k) = O(h^p)$ γιά $k\geq 0$

την οποία αποδεικνύουμε παρακάτω.

Γιά $k\geq r$, η (39) αποδεικνύεται ως εξής: Παρατηρούμε ότι η (34) ισοδυναμεί με την συνθήκη ότι οι κανόνες ολοκλήρωσης (7) είναι όλοι ακριβείς γιά πολυώνυμα βαθμού $\leq s-1$. 'Αρα έχουμε γιά $1\leq i\leq q$

$$y(t^{n,i})-y(t^n) = \int_{t^n}^{t^{n,i}} y'(\tau)d\tau = h\sum_{j=1}^{q} a_{ij}\, y'(t^{n,j}) + O(h^{s+1}),$$

δηλ. ότι

$$y(t^{n,i})=y(t^n) + h\sum_{j=1}^{q} a_{ij}\, f(t^{n,j},y(t^{n,j})) + O(h^{s+1}),\ 1\leq i\leq q$$

Συνεπώς, η (29α) και η παραπάνω σχέση δίνουν

$$y(t^{n,i})-\zeta^{n,i} = h\sum_{j=1}^{q} a_{ij}[f(t^{n,j},y(t^{n,j}))-f(t^{n,j},\zeta^{n,j})]+O(h^{s+1}),\ 1\leq i\leq q$$

Παίρνοντας νόρμες, χρησιμοποιώντας την συνθήκη Lipschitz στην f και χρησιμοποιώντας ανάλογους συλλογισμούς μ' αυτούς που οδήγησαν στην ανισότητα (27) της απόδειξης της Πρότασης 2 (όπου $y^{n,i}$ θέτουμε $y(t^{n,i})$ και όπου y^n-z^n θέτουμε $O(h^{s+1})$) παίρνουμε την σχέση

$$(40)\quad y(t^{n,i}) - \zeta^{n,i} = O(h^{s+1}),\ 1\le i\le q,$$

η οποία, λόγω της (37), δίνει

$$(41)\quad E_n(k) = O(h^{k+s+1}),\ k\ge 0.$$

Άρα ισχύει η (39) γιά $k\ge r$ γιατί υποθέσαμε ((36)) ότι $p\le r+s+1$. 'Εστω τώρα $0\le k\le r$. Παρατηρούμε ότι αν $(\partial_y f)_{ij} = \partial f_i/\partial y_j$, $1\le i,j\le m$, τότε

$$f(t^{n,j},y(t^{n,j}))-f(t^{n,j},\zeta^{n,j}) - \partial_y f\ (t^{n,j},y(t^{n,j}))(y(t^{n,j})-\zeta^{n,j})$$
$$= O(\|y(t^{n,j})-\zeta^{n,j}\|^2) = O(h^{2s+2}).$$

(Υποθέσαμε ότι η f είναι αρκετά ομαλή - και παραγωγίσιμη άρα - και ότι η παράγωγός της είναι Lipschitz - κάτι που εξασφαλίζεται αν η "δεύτερη" παράγωγος π.χ. είναι συνεχής. Χρησιμοποιήσαμε μετά την Πρόταση 2.1.4 με p=1 και τέλος την (40)). Συμπεραίνουμε ότι

$$E_n(k)= \sum_{i=1}^{q} b_i(t^{n,i}-t^n)^k\ \partial_y f(t^{n,i},y(t^{n,i}))(y(t^{n,i})-\zeta^{n,i})+O(h^{2s+2+k}).$$

'Εστω τώρα $\varphi(t)=\partial_y f(t,y(t)) \in R^{m\times m}$ γιά $a\le t\le b$ και έστω

$$C(n,\lambda) = \varphi^{(\lambda)}(t^n)/\lambda!,\ \text{όπου}\ \varphi^{(\lambda)}(t) = D_t^{\lambda}\ \varphi(t).$$

Ο τύπος του Taylor δίνει λοιπόν αν αναπτύξουμε γύρω από το t^n: $(r>k\ge 0)$

$$\partial_y f(t^{n,i},y(t^{n,i})) = \sum_{\lambda=0}^{r-k-1} (t^{n,i}-t^n)^{\lambda}\ C(n,\lambda) + O(h^{r-k})$$

'Αρα, γιά $0\le k\le r-1$ αντικαθιστώντας στον παρακάτω τύπο γιά το $E_n(k)$ έχουμε με χρήση της (40)) ότι

$$(42)\ E_n(k) = \sum_{i=1}^{q} b_i(t^{n,i}-t^n)^k \left(\sum_{\lambda=0}^{r-k-1} (t^{n,i}-t^n)^\lambda C(n,\lambda)\right)(y(t^{n,i}-\zeta^{n,i})$$

$$+ O(h^{r+s+1}) + O(h^{2s+2+k}) =$$

$$= \sum_{\lambda=k}^{r-1} C(n,\lambda-k)\ \left[\sum_{i=1}^{q} b_i(t^{n,i}-t^n)^\lambda (y(t^{n,i})-\zeta^{n,i})\right]$$

$$+ O(h^{r+s+1}) + O(h^{2s+2+k})$$

Τώρα, χρησιμοποιώντας τις σχέσεις (29α), (35) έχουμε γιά $0\le\lambda\le r-1$

$$(43)\quad \sum_{i=1}^{q} b_i\ (t^{n,i}-t^n)^\lambda\ (\zeta^{n,i}-y(t^n))$$

$$= h^{\lambda+1} \sum_{i=1}^{q} b_i\ \tau_i^\lambda\ \left(\sum_{j=1}^{q} a_{ij} f(t^{n,j},\zeta^{n,j})\right)$$

$$= h^{\lambda+1} \sum_{j=1}^{q} f(t^{n,j},\zeta^{n,j})\ \left(\sum_{i=1}^{q} b_i\tau_i^\lambda\ a_{ij}\right)$$

$$= h^{\lambda+1}(\lambda+1)^{-1} \sum_{j=1}^{q} b_j(1-\tau_j^{\lambda+1})\ f(t^{n,j},\zeta^{n,j})$$

Χρησιμοποιώντας εξ άλλου πάλι ότι η (33) σημαίνει ότι ο κανόνας ολοκλήρωσης (8) είναι ακριβής γιά πολυώνυμα βαθμού $\le p-1$, έχουμε

$$(44)\ \sum_{i=1}^{q} b_i\ (t^{n,i}-t^n)^\lambda\ (y(t^{n,i})-y(t^n))$$

$$= h^{-1} \int_{t^n}^{t^{n+1}} (t-t_n)^\lambda\ (y(t)-y(t^n))dt + O(h^p)$$

$$= h^{-1} \int_{t^n}^{t^{n+1}} (t-t_n)^{\lambda} \left(\int_{t^n}^{t} f(x,y(x))dx\right)dt + O(h^p)$$

= (με ολοκλήρωση κατά μέρη)

$$= (h(\lambda+1))^{-1} \int_{t^n}^{t^{n+1}} (h^{\lambda+1}-(t-t^n)^{\lambda+1})\ f(t,y(t))dt+O(h^p).$$

= (χρησιμοποιώντας τον κανόνα ολοκλήρωσης (8))

$$= (\lambda+1)^{-1} \sum_{j=1}^{q} b_j\ (h^{\lambda+1}-\tau_j^{\lambda+1}h^{\lambda+1})\ f(t^{n,j},y(t^{n,j}))+O(h^p)$$

$$= h^{\lambda+1}(1+\lambda)^{-1} \sum_{j=1}^{q} b_j\ (1-\tau_j^{\lambda+1})\ f(t^{n,j},y(t^{n,j})) + O(h^p).$$

Αφαιρώντας κατά μέλη την (43) από την (44) έχουμε λοιπόν γιά $0\le\lambda\le r-1$

$$\sum_{i=1}^{q} b_i\ (t^{n,i}-t^n)^{\lambda}\ (y(t^{n,i})-\zeta^{n,i}) =$$

$$= h^{\lambda+1}(\lambda+1)^{-1} \sum_{j=1}^{q} b_j\ (1-\tau_j^{\lambda+1})(f(t^{n,j},y(t^{n,j}))-f(t^{n,j},\zeta^{n,j}))+O(h^p).$$

Αντικαθιστώντας στην (42) παίρνουμε γιά $0\le k\le r-1$, επειδή $p\le r+s+1$ και $p\le 2s+2$ λόγω των υποθέσεών μας

$$E_n(k) = \sum_{\lambda=k}^{r-1} h^{\lambda+1}(\lambda+1)^{-1}\ C(n,\lambda-k)\ \{\sum_{i=1}^{q} b_i\ (1-\tau_i^{\lambda+1})\ [f(t^{n,i},y(t^{n,i})-$$
$$-f(t^{n,i},\zeta^{n,i})]\} + O(h^p).$$

Ανακαλώντας τώρα τόν ορισμό του $E_n(k)$ από την (37) έχουμε γιά $0\le k\le r-1$

$$E_n(k) = \sum_{\lambda=k}^{r-1} h^{\lambda+1} (\lambda+1)^{-1} C(n,\lambda-k) E_n(0)$$

$$- \sum_{\lambda=k}^{r-1} (\lambda+1)^{-1} C(n,\lambda-k) E_n(\lambda+1)+O(h^p)$$

και επειδή ήδη ξέρουμε ότι $E_n(r)=O(h^p)$ (γιατί ήδη αποδείξαμε ότι ισχύει η (39) γιά $k\geq r$, η παραπάνω σχέση γράφεται

$$(45)\quad E_n(k) = \sum_{\lambda=k}^{r-1} h^{\lambda+1}(\lambda+1)^{-1} C(n,\lambda-k) E_n(0)$$

$$- \sum_{\lambda=k}^{r-2} (\lambda+1)^{-1} C(n,\lambda-k) E_n(\lambda+1)+O(h^p),\ 0\leq k\leq r-1 .$$

Ας θυμηθούμε ότι η (41) ισχύει γιά κάθε $k\geq 0$. Αντικαθιστώντας την στην (45) έχουμε γιά $0\leq k\leq r-1$

$$(46)\quad E_n(k) = O(h^{s+2k+2})+O(h^{s+k+2})+O(h^p)=O(h^{s+k+2})+O(h^p) .$$

Αντικαθιστώντας την σχέση αυτή πάλι στην (45) παίρνουμε γιά $0\leq k\leq r-1$

$$(47)\ E_n(k) = O(h^{s+2k+3})+O(h^{s+k+3})+O(h^p)=O(h^{s+k+3})+O(h^p).$$

Παρατηρούμε δηλ. ότι με κάθε νέα αντικατάσταση η (45) δίνει μία αυξημένη κατά 1 τάξη στον πρώτο όρο του δευτέρου μέλους. Επαναλαμβανοντας αυτήν την διαδικασία όσο χρειάζεται κατορθώνουμε τελικά να δείξουμε ότι γιά κάθε $0\leq k\leq r-1$:

$$(48)\ E_n(k) = O(h^{s+r+1})+O(h^p) = (\text{λόγω της υποθέσεως } (36)) = O(h^p)$$

που είναι το ζητούμενο αποτέλεσμα (39) και γιά $0\leq k\leq r-1$. @

Είδαμε ότι οι πρώτες δύο συνθήκες του θεωρήματος 2 εκφράζουν, αντίστοιχα, ότι οι κανόνες αριθμητικής ολοκλήρωσης (8), αντστχ. (7), είναι ακριβείς γιά πολυώνυμα βαθμού $\leq$ p-1, αντίστχ. $\leq$ s-1. Δηλ. είναι όπως λέμε <u>τάξης</u> p-1, αντίστοιχα s-1. Είναι χρήσιμο να δούμε ένα πόρισμα του θεωρήματος 2 που βρίσκει ικανές συνθήκες (γιά να έχει η μέθοδος RK(12) τάξη p) που διατυπώνονται μόνο μέσω της <u>τάξης</u> των κανόνων ολοκλήρωσης (8) και (7), δηλ. χωρίς χρήση της συνθήκης (35). Γιά (μερική) απόδειξη βλ. τις Ασκήσεις 8,9,10.

ΠΟΡΙΣΜΑ 1 (Butcher-Crouzeix)

(α) Αν ο κανόνας ολοκλήρωσης (8) είναι τάξης p-1 και όλοι οι κανόνες (7) είναι τάξης p-2, τότε η μέθοδος (12) έχει τάξη p.

(β) Εστω q' ο αριθμός των $\{\tau_i\}$ $1\le i\le q$ που είναι διάφορα μεταξύ τους. Αν όλοι οι κανόνες (7) είναι τάξης q'-1 και ο κανόνας (8) είναι τάξης p-1, τότε η μέθοδος (12) έχει τάξη p.

(γ) Υπάρχει μόνο μία μέθοδος RK με q στάδια που έχει τάξη 2q. Είναι η μέθοδος γιά την οποία τα τ_i και b_i είναι, αντίστοιχα, οι κόμβοι και οι συντελεστές της ολοκλήρωσης Gauss-Legendre (δηλ. της ολοκλήρωσης Gauss με βάρος $w(x)=1$) στο διάστημα [0,1]. (Τα a_{ij} κατασκευάζονται έτσι ώστε οι τύποι ολοκλήρωσης (7) να είναι ακριβείς γιά πολυώνυμα βαθμού $\le$ q-1. Αυτό δίνει ένα $q^2 x q^2$ ((34) με s=q) γραμμικό αντιστρέψιμο σύστημα γιά τα a_{ij}). @

Κλείνουμε αυτήν την παράγραφο αναφέροντας μία αξιοσημείωτη οικογένεια μεθόδων, που δεν ανήκουν στην κατηγορία των μεθόδων RK (12) αλλά αποτελούν κατά κάποιο τρόπο ένα ενδιάμεσο βήμα μεταξύ των αμέσων και των πεπλεγμένων μεθόδων RK. Θεωρούμε γιά απλούστευση (βλ. Παρατήρηση 2) το αυτόνομο σύστημα $y'=f(y)$ και υποθέτουμε ότι γιά $y\in R^m$ είναι γνωστός ο mxm Ιακωβιανός πίνακας $J(y)=(\partial_y f)(y)$. Οι λεγόμενες <u>μέθοδοι Rosenbrock</u> είναι της μορφής:

$$y^{n,1} = hf(y^n) + h\gamma_1\, J(y^n)\, y^{n,1}$$

$$y^{n,2} = hf(y^n+a_{21}y^{n,1})+h\gamma_2 J(y^n+c_{21}y^{n,1})y^{n,2}$$

(49α) $\quad\vdots$

$$y^{n,q} = hf(y^n+\sum_{j=1}^{q-1} a_{qj}\, y^{n,j}) + h\gamma_q\, J(y^n+\sum_{j=1}^{q-1} c_{qj}\, y^{n,j})\, y^{n,q}$$

$$(49\beta)\quad y^{n+1} = y^n + \sum_{i=1}^{q} b_i\, y^{n,i},$$

όπου a_{ij}, c_{ij}, $1 \le j \le i-1 \le q-1$, b_i, γ_i, $1 \le i \le q$ δεδομένες σταθερές. Οι μέθοδοι αυτής της μορφής, όπως φαίνεται από την (49α), απαιτούν την λύση q γραμμικών συστημάτων γιά τον υπολογισμό των ενδιαμέσων τιμών $y^{n,i}$, $1 \le i \le q$ και, όπως θα δούμε στην παράγραφο 3.4, έχουν μερικές από τις επιθυμητές ιδιότητες "απόλυτης ευστάθειας" των πεπλεγμένων μεθόδων RK. 'Ενα πολύ γνωστό παράδειγμα μεθόδων Rosenbrock είναι η λεγόμενη μέθοδος του Calahan:

$$(50)\quad \begin{cases} y^{n,1} = hf(y^n) + h\gamma J(y^n)y^{n,1} \\ y^{n,2} = hf(y^n + a_{21}y^{n,1}) + h\gamma J(y^n)y^{n,2} \\ y^{n+1} = y_n + b_1 y^{n,1} + b_2 y^{n,2} \end{cases}$$

όπου $a_{21} = -2/3^{1/2}$, $\gamma = (1+3^{-1/2})/2$, $b_1 = 3/4$, $b_2 = 1/4$ που έχει τάξη ακρίβειας $p=3$ και απαιτεί την λύση δύο γραμμικών συστημάτων (με τον ίδιο πίνακα $I - h\gamma J(y^n)$) σε κάθε βήμα n γιά τον προσδιορισμό των $y^{n,1}$, $y^{n,2}$.

Παρατηρήσεις

1. Στην Πρόταση 1 είδαμε ότι το μη γραμμικό σύστημα (12α) έχει μοναδική λύση $(y^{n,i})$, $1 \le i \le q$, δηλ. το μοναδικό σταθερό σημείο στον $(\mathbb{R}^m)^q$ της απεικόνισης F (βλ. απόδειξη της Πρότασης 1). Η λύση $Y_n = (y^{n,i}) \in (\mathbb{R}^m)^q$ του συστήματος $Y_n = F(Y_n)$ μπορεί π.χ. να προσεγγισθεί με την απλή επαναληπτική μέθοδο $Y_n^{[j+1]} = F(Y_n^{[j]})$, $j = 0,1,2,\ldots$, ή με την μέθοδο του Νεύτωνα ή (συνηθέστατα) με μιά απλουστευμένη μέθοδο του τύπου του Νεύτωνα - π.χ. με την μέθοδο της χορδής -. Ως αρχική τιμή $Y_n^{[0]}$ της ακολουθίας - ανακαλώντας ότι η λύση $(y^{n,i})$ του συστήματος αποτελεί προσέγγιση των τιμών $(y(t^{n,i}))$ - μπορούμε π.χ.

να πάρουμε το διάνυσμα $(y^n,\dots,y^n)^T$ ή ένα διάνυσμα γραμμικών συνδυασμών τιμών y^k, $k\leq n$, γιά τις οποίες οι αντίστοιχοι γραμμικοί συνδυασμοί $y(t^k)$ είναι καλές προσεγγίσεις των $y(t^{n,i})$. Τονίζουμε ότι δεν μας ενδιαφέρει η ακριβής λύση των συστημάτων (49) - μιά και οι τιμές y^n, είναι τελικά προσεγγίσεις των $y(t^n)$ - αλλά μιά προσεγγιστική τους λύση που να κατασκευάζεται σχετικά εύκολα (π.χ. με 2 ή 3 ανακυκλώσεις j το πολύ στην πράξη γιά κάθε n) και που να δίνει σφάλμα αρκετά μικρό έτσι ώστε το συνολικό τοπικό σφάλμα να εξακολουθεί να είναι της τάξης h^{p+1}, όπου p η τάξη της αριθμητικής μεθόδου γιά την λύση της διαφορικής εξίσωσης.

'Οπως αναφέραμε ήδη σ' αυτήν την παράγραφο, οι ημιπεπλεγμένες μέθοδοι RK είναι ιδιαίτερα ενδιαφέρουσες από την άποψη της ευκολίας επίλυσης των μη γραμμικών συστημάτων τους. 'Ενα ενδιαφέρον υποσύνολο των ημιπεπλεγμένων μεθόδων αποτελούν οι λεγόμενες διαγώνια πεπλεγμένες (DIRK) γιά τις οποίες τα διαγώνια στοιχεία του πίνακα A είναι όλα ίσα. Π.χ. οι διαγώνια πεπλεγμένες μέθοδοι με q=2 δίνονται από το μητρώο (18) και είναι γενικά τάξης p=2 εκτός αν το λ είναι μία από τις ρίζες του διωνύμου $\lambda^2-\lambda+1/6$ οπότε p=3. Αξιοσημείωτες είναι οι διαγώνια πεπλεγμένες μέθοδοι με q=3 στάδια που δίνονται από το μητρώο

$$(51)\qquad \begin{array}{ccc|c} \beta & 0 & 0 & \beta \\ 1/2-\beta & \beta & 0 & 1/2 \\ 2\beta & 1-4\beta & \beta & 1-\beta \\ \hline b_1 & b_2 & b_3 & \end{array} \quad ,$$

όπου $b_1=b_3=[6(2\beta-1)^2]^{-1}$ και $b_2=1-2b_1$. Αν το β είναι μία από τις τρείς πραγματικές ρίζες του πολυωνύμου $\beta^3-3\beta^2/2+\beta/2-1/24$, η μέθοδος (51) έχει τάξη p=4.

Ο λόγος γιά τον οποίο οι διαγώνια πεπλεγμένες μέθοδοι είναι ιδιαίτερα ενδιαφέρουσες στην πράξη είναι (εκτός από τις καλές ιδιότητες ευστάθειας τους, βλ. Παρ. 3.4) ο εξής: Θεωρούμε, αντί του (3.1.1) το αυτόνομο σύστημα

$$(52)\quad y'=f(y),$$

δηλ. την περίπτωση που η f δεν εξαρτάται άμεσα από το t. (Σημειώστε ότι το σύστημα $y'=f(t,y)$, $y\in R^m$ μετατρέπεται εύκολα σε αυτόνομο με m+1 εξισώσεις ως πρός την μεταβλητή $z=(y,t)^T\in R^{m+1}$: το σύστημα $z'=f(z)$ είναι το προηγούμενο με την προσθήκη της εξίσωσης $t'=1$, $t(\alpha)=\alpha$, γιά την μεταβλητή $z_{m+1}=t$). Γιά το (50) η επίλυση του μη γραμμικού συστήματος (12α) γιά μία διαγώνια πεπλεγμένη μέθοδο με $a_{ii}=\lambda\neq 0$ ανάγεται στη επίλυση q μη γραμμικών συστημάτων της μορφής

$$(53)\quad y^{n,i} = h\lambda f(y^{n,i})+z^{n,i},\ 1\leq i\leq q,$$

όπου $z^{n,i}$ γνωστά διανύσματα. Αν λύσουμε το (51) π.χ. με την μέθοδο της χορδής θα πρέπει γιά κάθε i να υπολογίσουμε και να αναλύσουμε σε μορφή LU μόνο μία φορά του Ιακωβιανό πίνακα $I-h\lambda\,\partial_y f$. (Συνήθως χρησιμοποιούμε τον ίδιο πίνακα γιά όλα τα στάδια q και τον μεταβάλλουμε κάθε 10 ή 20 π.χ. χρονικά βήματα n). Σημειώνουμε τέλος ότι οι διαγώνια πεπλεγμένες μέθοδοι είναι υποσύνολο των πεπλεγμένων μεθόδων γιά τις οποίες ο πίνακας $A=(a_{ij})$ εχει μιά μόνο ιδιοτιμή λ πολλαπλότητας q' η κατηγορία αυτή των μεθόδων είναι επίσης πολύ αποτελεσματική στην πράξη. (Χρειάζεται κατάλληλη αλλαγή μεταβλητών και χρήση της μορφής Jordan του A).

2. Στην πράξη οι μέθοδοι RK χρησιμοποιούνται με μεταβλητό γενικά βήμα $t^{n+1}-t^n=h_n$ το οποίο είναι επιθυμητό να αυξομειώνεται χωρίς την παρέμβασή μας ανάλογα με το αν η λύση αλλάζει πολύ από βήμα σε βήμα ή αν μεταβάλλεται αργά και ομαλά. (Μας ενδιαφέρει να κρατάμε το σφάλμα μικρό: θυμηθείτε ότι τα φράγματα των σφαλμάτων είναι της μορφής $C(y)h^p$ όπου $C(y)$ κάποια συνάρτηση (ημι)νορμών υψηλών παραγώγων της λύσης $y(t)$). Η τεχνική που χρησιμοποιείται γιά την "αυτόματη" μεταβολή του βήματος στηρίζεται σε μία εκτίμηση κάποιου τοπικού σφάλματος, το οποίο συνήθως ορίζεται ως η διαφορά $u(t^{n+1})-y^{n+1}$, όπου $\{y^n\}$ η αριθμητική λύση και $u(t)$ είναι η λύση, γιά $t^n\leq t\leq t^{n+1}$, του προβλήματος $u'(t)=f(t,u(t))$, $t^n\leq t\leq t^{n+1}$, $u(t^n)=y^n$. Αν κατορθώσουμε να ελέγξουμε το τοπικό σφάλμα, τότε το ολικό σφάλμα $y(t^{n+1})-y^{n+1}=(y(t^{n+1})-u(t^{n+1}))+(u(t^{n+1})-y^{n+1})$ θα είναι επίσης μικρό,

υπό την προϋπόθεση ότι η τιμή y^n ήταν κοντά στην πραγματική τιμή $y(t^n)$. Δέν είναι δύσκολο να δει κανείς (βλ. π.χ. [5.3, παρ. Β4]) ότι το τοπικό σφάλμα <u>προσεγγίζεται</u> αρκετά καλά από μία διαφορά της μορφής $\tilde{y}^{n+1}-y^{n+1}$, όπου $\tilde{y}^{n+1}$ μία προσέγγιση της $y(t^{n+1})$ μεγαλύτερης τάξης ακρίβειας από την y^{n+1}. Αν $|\tilde{y}^{n+1}-y^{n+1}| \leq \epsilon h_n$, όπου ε κάποιο ανεκτό επίπεδο "σφάλματος ανά βήμα", αποδεχόμαστε την προσέγγιση y^{n+1} και συνεχίζουμε τον υπολογισμό με το ίδιο $\Delta t=h_n$. Αν $|\tilde{y}^{n+1}-y^{n+1}|>\epsilon h_n$, την απορρίπτουμε, μειώνουμε το βήμα h_n, βρίσκουμε νέα y^{n+1}, $\tilde{y}^{n+1}$ και επαναλαμβάνουμε τον έλεγχο κ.ο.κ.

Μας ενδιαφέρει συνεπώς να μπορούμε, χωρίς σημαντική αύξηση του αριθμού των πράξεων ανά βήμα, να υπολογίζουμε και ένα $\tilde{y}^{n+1}$ με μία συνοδό-μέθοδο μεγαλύτερης τάξης ακρίβειας από την βασική μας μέθοδο που παράγει το y^{n+1}. Τέτοια ζεύγη μεθόδων είναι π.χ. οι μέθοδοι RKF (Runge-Kutta-Fehlberg) τάξεων (4,5) ή (5,6) κ.λ.π. που χρησιμοποιούνται ευρύτατα στην πράξη γιά την κατασκευή αλγορίθμων με "αυτόματη" επιλογή βήματος και εκτίμηση του τοπικού σφάλματος γιά συστήματα Σ.Δ.Ε. που <u>δέν</u> είναι άκαμπτα. Βλ. [5.3, παρΒ4], το βιβλίο [3.7] καθώς και τα άρθρα των Hull et al. στο SIAM J. Num. Anal., 9(1972), 603-637, και των Shampine et al. στο SIAM Review, 18(1976), 376-411, μεταξύ άλλων, γιά περισσότερες λεπτομέρειες.

Ασκήσεις 3.2

1. Απαντήστε στά δύο "γιατί" της απόδειξης της Πρότασης 1 και στο "γιατί" της απόδειξης της Πρότασης 2 διατυπώνοντας και αποδεικνύοντας ανάλογα θεωρήματα.

2(α) Θεωρείστε την "άμεση μέθοδο του μέσου" (6) (Ο παρακάτω προσδιορισμός της τάξης s ακρίβειάς της αποτελεί υπόδειγμα που μπορεί κανείς ν' ακολουθήσει κανείς γιά άμεσες μεθόδους. Μη χρησιμοποιείστε το Θεώρημα 2 ή την 'Ασκηση 5). Δείξτε ότι η μέθοδος είναι της μορφής

$$y^{n+1}= y^n+h\Phi(t^n,y^n;h),\ n\geq 0$$

όπου

$$\phi(t,y;h) = f(t+h/2,\ y+hf(t,y)/2).$$

Ορίστε την ποσότητα δ^n όπως στην (32), δηλ. ως

$$\delta^n=y(t^{n+1})-y(t^n)-f(t^n+h/2,y(t^n)+hf(t^n,y(t^n))/2),\ 0\le n\le N-1$$

(β) Θεωρείστε το πρόβλημα (3.1.1) στον $\mathbb{R}^1$, δηλ. γιά μία <u>απλή</u> Δ.Ε.. Αναπτύξτε την συνάρτηση $\delta^n=\delta^n(t^n,y(t^n),h)$ σε δυναμοσειρά του h αναπτύσσοντας κατά Taylor την $\phi(t^n,y(t^n),h)$ ως συνάρτηση δύο μεταβλητών $f(t^n+h/2,\ y(t^n)+hk)$ γύρω απ' το σημείο $(t^n,y(t^n))$ και την διαφορά $y(t^{n+1})-y(t^n)$ γύρω από το t^n. Δείξτε ότι αν οι συναρτήσεις $f, f_t, f_y, f_{tt}, f_{yy}, f_{ty}$ είναι φραγμένες και συνεχείς γιά $t\in[a,b]$, τότε

$$\max_{0\le n\le N} |\delta^n| \le Dh^3,$$

και ότι η μέθοδος έχει τάξη ακρίβειας p=2 γιά απλές Δ.Ε. (Δηλ. ότι η δύναμις 3 στο φράγμα δεν μπορεί να αυξηθεί γιά γενικά f και y).

(γ) Επεκτείνετε την παρακάτω ανάλυση στον $\mathbb{R}^m$ και δείξτε ότι γιά $y(t)$, $f(t,y)$ ομαλές υπάρχει σταθερά $D=D(a,b,y,f)$ τ.ώ.

$$\max_n \|\delta^n\| \le Dh^3$$

δηλ. ότι η μέθοδος είναι τάξης p=2 και γιά συστήματα Δ.Ε.

3. (Γιά όσους αγαπούν τις πράξεις). Ακολουθώντας τα βήματα της Άσκησης 1 (δηλ. χωρίς χρήση του Θεωρήματος 2 ή της Άσκησης 5) δείξτε, γιά <u>απλές Δ.Ε</u> (δηλ. στον $\mathbb{R}^1$), ότι η άμεση μέθοδος RK (20β) έχει τάξη ακρίβειας p=3. (Το ίδιο ισχύει και γιά συστήματα αλλά η λογιστική των αναπτυγμάτων Taylor αυξάνει).

4. (Γιά όσους αγαπούν <u>πολύ</u> τις πράξεις). Όπως στις Ασκήσεις 1 και 2 δείξτε ότι η τάξη της κλασικής μεθόδου RK (21α) είναι p=4 γιά απλές Δ.Ε. (Το ίδιο ισχύει και γιά συστήματα).

5. (Βραβείο Runge-Kutta) Σ' αυτήν την άσκηση θα κατασκευάσουμε όλες τις άμεσες μεθόδους RK με αριθμό σταδίων $q\leq 3$ που έχουν την μεγαλύτερη δυνατή τάξη (γιά απλές Δ.Ε). Η γενική άμεση μέθοδος με 3 στάδια γράφεται - στη μορφή (13) με $k^{n,i}=k^i$ - ως:

(i) $y^{n+1} = y^n+h\phi(t^n,y^n,h)$, όπου

(ii) $\Phi(t,y,h) = \sum_{j=1}^{3} b_j\, k^j$ και όπου

(iii) $k^1 = f(t,y)$

$k^2 = f(t+h\tau_2,\ y+ha_{21}k^1)$

$k^3 = f(t+h\tau_3,\ y+h(a_{31}k^1+a_{32}k^2))$

(Παρατηρούμε ότι η γενική άμεση μέθοδος με $q=1$, αντίστοιχα με $q=2$, είναι της μορφής (i)-(iii) αν υποθέσουμε ότι $b_2=b_3=0$, αντίστοιχα $b_3=0$).

(α) Δείξτε ότι γιά να έχει η μέθοδος τάξη ακρίβειας $p\geq 3$ είναι αναγκαίο να έχει η Φ ανάπτυγμα της μορφής

$$\Phi_*(t,y,h) = f+hF/2 + h^2(Ff_y+G)/6 + O(h^3)$$

όπου $F=f_t+ff_y$, $G=f_{tt}+2ff_{ty}+f^2f_{yy}$. (Η f και οι μερικές της παράγωγοι υπολογίζονται στο σημείο (t,y). Φυσικά γιά $p\geq 1$ είναι αναγκαίο $\Phi_*=f+O(h)$. Γιά $p\geq 2$, $\Phi_*=f+hF/2+O(h^2)$).

(β) Αναπτύξτε την k^2 σε σειρά Taylor περί το σημείο (t,y), χρησιμοποιώντας την σχέση $k^1=f$· υπολογίστε τους όρους τάξης μέχρι και $O(h^2)$. Κατόπιν, αντικαθιστώντας στην k^3 το ανάπτυγμα της k^2, βρήτε παρόμοιο γιά την k^3. Αντικαθιστώντας αυτά τα αναπτύγματα στο δεύτερο μέλος της (ii) πάρτε τελικά ένα ανάπτυγμα της μορφής

$$\Phi(t,y,h) = A+Bh+Ch^2+O(h^3)$$

(γ) Εξισώνοντας όρους του δευτέρου μέλος της Φ_* με αντίστοιχους όρους του παραπάνω αναπτύγματος της F δείξτε ότι:

(γ_1) Γιά $q=1$ (δηλ. με $b_2=b_3=0$), η μόνη άμεση μέθοδος με τάξη ακρίβειας $p\geq 1$ είναι η μέθοδος του Euler, δηλ. η μέθοδος με $b_1=1$, $a_{11}=\tau_1=0$, που έχει $p=1$.

(γ_2) Γιά $q=2$ (δηλ. με $b_3=0$) δείξτε ότι υπάρχει μονοπαραμετρική οικογένεια μεθόδων με $p\geq 2$· οι σταθερές τους a_{21},τ_1,b_1,b_2 ικανοποιούν τις σχέσεις:

$$b_1+b_2=1$$

$$a_{21}=\tau_2$$

$$b_2\tau_2=1/2$$

δείξτε ότι όλες αυτές οι μέθοδοι έχουν τάξη $p=2$, (δηλ. ότι δεν υπάρχει άμεση μέθοδος με $q=2$, $p>2$). Σε ποιές τιμές των σταθερών αντιστοιχεί η μέθοδος του μέσου; Σαν τι μπορούν να ερμηνευθούν οι μέθοδοι με $b_2=1$ και με $b_2=1/2$;

(γ_3) Γιά $q=3$ προκύπτουν (αν θέλουμε $p\geq 3$) οι συνθήκες: $b_1+b_2+b_3=1$, $a_{21}=\tau_2$, $a_{31}+a_{32}=\tau_3$, $b_2\tau_2+b_3\tau_3=1/2$, $b_2\tau_2^2+b_3\tau_3^2=1/3$, $b_3\tau_2 a_{32}=1/6$. Συνεπώς, υπάρχει διπαραμετρική οικογένεια αμέσων μεθόδων με $q=3$, $p\geq 3$. (Μπορεί να δειχθεί - με υπολογισμό του όρου $O(h^3)$ - ότι καμμιά από αυτές τις μεθόδους δεν έχει $p>3$, δηλ. ότι γιά όλες $p=3$). Βεβαιωθείτε ότι οι (20α),(20β) ανήκουν σ' αυτήν την κλάση.

6. (Διάλειμμα!) Δείξτε ότι η μόνη μέθοδος RK με 1 στάδιο και τάξη ακρίβειας 2 είναι η μέθοδος του μέσου (15).

7. (Μέγα βραβείο Runge-Kutta). Θεωρούμε την γενική (πεπλεγμένη) μέθοδο RK με $q=2$ την οποία γράφουμε στην μορφή (13) - βλ. και άσκηση 5 - δηλ. ως

$$(i) \quad y^{n+1}-y^n = h\Phi(t^n,y^n,h), \text{ όπου}$$
$$(ii) \quad \Phi(t,y,h) = b_1k^1+b_2k^2, \text{ όπου}$$
$$(iii) \quad \begin{cases} k^1 = f(t+h\tau_1,\ y+h(a_{11}k^1+a_{12}k^2)) \\ k^2 = f(t+h\tau_2,\ y+h(a_{21}k^1+a_{22}k^2)) \end{cases}$$

θα προσδιορίσουμε συνθήκες πάνω στά a_{ij}, b_i, τ_i ώστε η μέθοδος να έχει τάξη ακρίβειας αντίστοιχα $p\geq 1,2,3,4$

(α) Δείξτε ότι η αναγκαία συνθήκη γιά να έχει η μέθοδος τουλάχιστον τάξη $p=4$ γιά απλή ΔΕ είναι να έχει η Φ ανάπτυγμα της μορφής

$$\Phi_*(t,y,h) = f+hF/2 + h^2(Ff_y+G)/6 + h^3[(3f_{ty}+3ff_{yy}+f_y{}^2)F + Gf_y+H]/24+O(h^4)$$

όπου τα F,G ορίστηκαν στην 'Ασκηση 5(α) και όπου

$$H = f_{ttt} + 3ff_{tty} + 3f^2f_{tyy} + f^3f_{yyy}.$$

(β) Αναπτύσσοντας τα k^i σε σειρά Taylor περί το (t,y) δείξτε ότι

$$\begin{aligned}(iv)\ k^i = {} & f+h[\tau_if_t + (a_{i1}k^1 + a_{i2}k^2)f_y] \\ & + h^2[\tau_i^2f_{tt}+2\tau_i(a_{i1}k^1+a_{i2}k^2)f_{ty}+(a_{i1}k^1+a_{i2}k^2)^2f_{yy}]/2 \\ & + h^3[\tau_i^3f_{ttt}+3\tau_i^2(a_{i1}k^1+a_{i2}k^2)f_{tty}+3\tau_i(a_{i1}k^1+a_{i2}k^2)^2f_{tyy}+ \\ & + (a_{i1}k^1+a_{i2}k^2)^3f_{yyy}]/6+O(h^4),\ i=1,2\end{aligned}$$

Επειδή τα αναπτύγματα αυτά είναι "πεπλεγμένα" δεν μπορούμε να προχωρήσουμε με διαδοχικές αντικαταστάσεις όπως στην 'Ασκηση 5. Υποθέστε λοιπόν ότι υπάρχουν τα αναπτύγματα της μορφής

$$(v)\ k^i = A_i + hB_i + h^2C_i + h^3D_i + O(h^4),\ i=1,2,$$

και αντικαθιστώντας τις (v) στις (iv) και εξισώνοντας δυνάμεις του h, βρήτε ένα σύστημα της μορφής

$$
\begin{aligned}
A_i &= f \\
B_i &= B_i\ (A_1,A_2) \\
C_i &= C_i\ (A_1,A_2,B_1,B_2) \\
D_i &= D_i\ (A_1,A_2,B_1,B_2,C_1,C_2),
\end{aligned}
$$

που μπορεί να λυθεί με απλή αντικατάσταση. Λύστε το! Βρήτε τέλος το ανάπτυγμα της Φ μέσω της (ii) (κρατήστε μέχρι και όρους $O(h^3)$).

(γ) Εξισώνοντας όρους ίσων δυνάμεων του h στα αναπτύγματα των Φ και $Φ_*$ βρήτε συνθήκες έτσι ώστε, η μέθοδος (i)-(iii) να έχει, αντίστοιχα, τάξη $p\geq 1,2,3$ και 4.

(δ) Βρήτε την γενική μορφή των ημιπεπλεγμένων μεθόδων με $p\geq 2$ και με ίσα διαγώνια στοιχεία $a_{11}=a_{22}=\lambda$. (βλ. (18)!). Γιά ποιά λ έχουμε $p\geq 3$; Υπάρχει τέτοια μέθοδος με p=4;

(ε) Αποδείξτε ότι υπάρχει μόνο μία μέθοδος με q=2, $p\geq 4$, δηλ. η (19). (Μπορεί να δειχθεί ότι η τάξη της είναι ακριβώς 4, βλ. Πόρισμα 1(γ)).

8. Να αποδειχθεί ο ισχυρισμός (α) του Πορίσματος 1. (Υπόδειξη: αποδείξτε, όπως στο πρώτο μέρος της απόδειξης του Θεωρήματος 2, ότι $E_n(0)=O(h^p)$).

9. Να αποδειχθεί ο ισχυρισμός (β) του Πορίσματος 1. (Υπόδειξη: Δείξτε ότι ισχύουν οι υποθέσεις του Θεωρήματος 2 με s=q και r=p-q. Χρησιμοποιείστε την ταυτότητα

$$\int_0^1 \int_0^x x^k \psi(t)dtdx = [\int_0^1 (1-x^{k+1})\ \psi(x)dx]/(k+1)$$

και δείξτε την (35) παίρνοντας ως ψ(t) τα πολυώνυμα βαθμού $\leq q-1$ τέτοια ώστε γιά δεδομένο j, $\psi(\tau_i)=0$ αν $i\neq j$, $\psi(\tau_j)=1$ και χρησιμοποιώνας το αποτέλεσμα της άσκησης 10(α)).

10. (α) Θεωρείστε τον κανόνα αριθμητικής ολοκλήρωσης

$$\int_0^1 \psi(\tau)d\tau \cong \sum_{j=1}^{q} b_j \, \psi(\tau_j)$$

με q διακριτά σημεία τ_j, $1\leq j\leq q$. Δείξτε ότι η τάξη του δεν μπορεί να υπερβαίνει το 2q-1.

(β) Δείξτε ότι η τάξη μιάς μεθόδου RK με q στάδια της μορφής (12) δεν μπορεί να υπερβαίνει το 2q (Υπόδειξη: Θεωρείστε μία Δ.Ε. της μορφής $y'=f(t)$).

(γ) Δείξτε - ανακαλώντας αποτελέσματα της ολοκλήρωσης Gauss' βλ. [5.2] - ότι οι μέθοδοι που περιγράφονται στο Πόρισμα 1(γ) είναι όντως τάξεως 2q. (Η μοναδικότητά τους γιά κάθε q είναι πιό δύσκολο να αποδειχθεί' βλ. Butcher (1964)).

11. ('Ασκηση μαμούθ). Γιά κάθε μία από τις μεθόδους: Euler, πεπλεγμένη Euler, (15)-(17), (18) (δύο περιπτώσεις: είτε λ ρίζα του $\lambda^2-\lambda+1/6=0$ είτε όχι), (19), (20α), (21α), (51), συγκρίνετε την πραγματική τους τάξη με την τάξη p που δίνουν οι (ικανές) συνθήκες των (α), (β) του Πορίσματος 1 καθώς και οι συνθήκες τού Θεωρήματος 2, παίρνοντας έτσι μιά ιδέα γιά την ισχύ των συνθηκών αυτών γιά συνηθισμένες μεθόδους RK..

12. Το παρακάτω αποτέλεσμα δίνει κάποια βάση στον ισχυρισμό "μικρό τοπικό σφάλμα" => "μικρό ολικό σφάλμα" της Παρατήρησης 2. Θεωρούμε την Δ.Ε.

$$w' = f(t,w), \quad t^n\leq t\leq t^{n+1},$$

και υποθέτουμε ότι η f είναι συνεχής γιά $(t,w)\in[t^n,t^{n+1}]\times R$, και ότι $|f(t,w_1)-f(t,w_2)|\leq L|w_1-w_2|$ γιά κάθε $t\in[t^n,t^{n+1}]$, $w_1,w_2\in R$. Θεωρείστε δύο λύσεις $u(t)$, $y(t)$ της Δ.Ε. γιά $t\in[t^n,t^{n+1}]$ που αντιστοιχούν στις αρχικές τιμές $u(t^n)$, $y(t^n)$. Δείξτε ότι αν $h_n=t^{n+1}-t^n$, $h_nL<1$, τότε

$$\max_{t^n\leq t\leq t^{n+1}} |u(t)-y(t)| \leq |u(t^n)-y(t^n)|/(1-Lh_n)$$

13. Γιά μία αυτόνομη Δ.Ε. $y'=f(y)$ δείξτε ότι η μέθοδος του Calahan (50) έχει <u>τάξη</u> $p=3$. (Ορίστε την τάξη της σε αναλογία με ό,τι κάναμε γιά τις μεθόδους RK. Αναπτύξτε σε δυναμοσειρές του h τις ενδιάμεσες ποσότητες και αντικαταστήστε στην τελευταία γραμμή).

3.3 ΠΟΛΥΒΗΜΑΤΙΚΕΣ ΜΕΘΟΔΟΙ

Σ' αυτήν την παράγραφο θα εξετάσουμε μια δεύτερη μεγάλη κατηγορία μεθόδων για την αριθμητική λύση του προβλήματος αρχικών τιμών (3.1.1), τις λεγόμενες (γραμμικές) πολυβηματικές μεθόδους. Ενα παράδειγμα τέτοιας μεθόδου είναι το εξής: Εστω ότι η λύση $y(t)$ του (3.1.1) είναι C^3 στο $[a,b]$. Τότε, αν t, $t \pm h \in [a,b]$, το θεώρημα του Taylor μας δίνει:

$$y(t+h) = y(t)+hy'(t)+h^2y''(t)/2 \quad +O(h^3)$$
$$y(t-h) = y(t)-hy'(t)+h^2y''(t)/2 \quad +O(h^3).$$

Αφαιρώντας κατά μέλη και χρησιμοποιώντας την Δ.Ε. έχουμε την σχέση

$$y(t+h)-y(t-h) = 2hy'(t)+O(h^3)=2hf(t,y(t))+O(h^3),$$

που δίνει (για τον συμβολισμό του ομοιόμορφου διαμερισμού βλ. παρ.3.1) την μέθοδο:

$$(1) \quad y^{n+1}-y^{n-1}=2hf^n, \quad n=1,2,\dots,N-1,$$

όπου, από δώ και εμπρός, $f^k=f(t^k,y^k)$, $0 \le k \le N$.

Η μέθοδος (1) είναι ένα παράδειγμα πολυβηματικής (διβηματικής) μεθόδου: για τον προσδιορισμό της y^{n+1} απαιτεί γνώση των τιμών y^n, y^{n-1} των δύο προηγουμένων βημάτων. Επίσης η (1) απαιτεί δύο αρχικές τιμές: την y^0, που παίρνουμε από το πρόβλημα αρχικών τιμών (3.1.1), και την y^1 που μπορούμε να υπολογίσουμε με μία "μονοβηματική" μέθοδο ή μια μέθοδο RK. Αλλάζοντας τον δείκτη στην (1) θα γράφουμε συνήθως την μέθοδο στην κανονική της μορφή

$$(1') \quad y^{n+2}-y^n = 2hf^{n+1} \;, \quad 0 \le n \le N-2.$$

Είναι προφανές ότι χρησιμοποιώντας αναπτύγματα Taylor τιμών $y(t+mh)$, $m=-1,-2,\ldots$ γύρω από το t μπορούμε να βρούμε και άλλες πολυβηματικές μεθόδους. Μιά άλλη τεχνική που χρησιμοποιείται για τον ίδιο σκοπό είναι η αριθμητική ολοκλήρωση. Π.χ. από την ταυτότητα

$$y(t^{n+2})-y(t^n) = \int_{t^n}^{t^{n+2}} f(t,y(t))dt,$$

προσεγγίζοντας το ολοκλήρωμα του δευτέρου μέλους με τον κανόνα του Simpson, δηλ. χρησιμοποιώντας την σχέση

$$\int_{t^n}^{t^{n+2}} f(t,y(t))dt \cong$$

$$\cong(t^{n+2}-t^n)[f(t^{n+2},y(t^{n+2}))+4f(t^{n+1},y(t^{n+1}))+f(t^n,y(t^n))]/6,$$

παίρνουμε την (επίσης διβηματική) μέθοδο του Simpson

$$(2) \qquad y^{n+2} - y^n = (h/3)(f^{n+2}+4f^{n+1}+f^n), \qquad 0\leq n\leq N-2,$$

η οποία είναι πεπλεγμένη, διότι απαιτεί την επίλυση του μη γραμμικού συστήματος $y^{n+2} = (h/3)f(t^{n+2},y^{n+2})+g^n$ σε κάθε βήμα n, $0\leq n\leq N-2$. Αντίθετα η μέθοδος (1) είναι άμεση.

Γενικά, μια (γραμμική) k-βηματική ($k\geq 1$) μέθοδος για την λύση του (3.1.1) είναι της μορφής:

$$(3)\begin{cases} y^0,y^1,\ldots,y^{k-1} \text{ δεδομένα. (Συνήθως υπολογίζονται με μεθόδους RK).} \\ \alpha_k y^{n+k}+\alpha_{k-1}y^{n+k-1}+\ldots+\alpha_0 y^n=h(\beta_k f^{n+k}+\beta_{k-1}f^{n+k-1}+\ldots+\beta_0 f^n), \\ (\sum_{j=0}^{k} \alpha_j\, y^{n+j}=h \sum_{j=0}^{k} \beta_j f^{n+j}), \qquad 0\leq n\leq N-k, \end{cases}$$

όπου $\{\alpha_j, \beta_j\}$, $0\leq j\leq k$, δεδομένες πραγματικές σταθερές ανεξάρτητες του h ή του n. Θα υποθέτουμε συνήθως ότι $\alpha_k=1$ και ότι $|\alpha_0|+|\beta_0|>0$ ώστε να έχουμε πράγματι μια k-βηματική μέθοδο. Αν $\beta_k=0$ η μέθοδος θα λέγεται <u>άμεση</u>: ο υπολογισμός του y^{n+k} γίνεται με απλή αντικατάσταση των γνωστών τιμών y^{n+i}, $0\leq i\leq k-1$. Αν $\beta_k\neq 0$ ο προσδιορισμός του y^{n+k} απαιτεί την λύση ενός mxm μη γραμμικού συστήματος της μορφής

$$(4) \qquad y^{n+k} = h\beta_k f(t^{n+k}, y^{n+k}) + g^n,$$

όπου g^n γνωστό διάνυσμα. Το (4) έχει προφανώς μοναδική λύση (από το θεώρημα συστολής) αν $h|\beta_k|\,L < 1$, δηλ. αν πάρουμε αρκετά μικρό h' το L είναι η σταθερά Lipschitz της f ως προς y. Είναι φανερό λοιπόν ότι οι πεπλεγμένες πολυβηματικές μέθοδοι έχουν πολύ μικρότερο κόστος ανά βήμα απ'ότι οι πεπλεγμένες μέθοδοι RK. Οι δε άμεσες πολυβηματικές μέθοδοι απαιτούν σε κάθε βήμα η ένα μόνο υπολογισμό της f (f^{n+k}). Είναι λοιπόν οι πολυβηματικές μέθοδοι πολύ φθηνότερες από τις μεθόδους RK*. Δες όμως την παράγραφο 3.4 για σαφή πλεονεκτήματα των (πεπλεγμένων) μεθόδων RK ως προς την "απόλυτη ευστάθεια" τους.

Η κλάση των μονοβηματικών μεθόδων (k=1) δηλ. οι μέθοδοι

$$(5) \qquad \alpha_1 y^{n+1} + \alpha_0 y^n = h(\beta_1 f^{n+1} + \beta_0 f^n), \qquad 0\leq n\leq N-1$$

με $\alpha_1=1$, $|\alpha_0|+|\beta_0|>0$, αποτελεί την "τομή" των πολυβηματικών μεθόδων με τις μεθόδους RK. Ετσι οι μέθοδοι Euler, πεπλεγμένη Euler και τραπεζίου, μπορούν να μελετηθούν και ως μονοβηματικές μέθοδοι.

---

* Γι' αυτό και χρησιμοποιήθηκαν ευρύτατα, από τον 19ο αιώνα ακόμη, για την αριθμητική λύση προβλημάτων αστρονομίας.

Υπάρχουν πολλοί τρόποι κατασκευής πολυβηματικών μεθόδων. Είδαμε παραδείγματα χρήσης αναπτυγμάτων Taylor και αριθμητικής ολοκλήρωσης. Άλλες τεχνικές που χρησιμοποιούνται είναι χρήση του πολυωνύμου παρεμβολής και αριθμητικής παραγώγισης. Π.χ. θεωρώντας το πολυώνυμο παρεμβολής $P_{k,n}(t)$ βαθμού $\leq k$ που παρεμβάλλεται στα σημεία t^{n+k}, t^{n+k-1}, ..., t^n στις τιμές $y(t^{n+k}),\dots,y(t^n)$, αντίστοιχα, και υπολογίζοντας την παράγωγό του στο σημείο t^{n+k} έχουμε ότι

$$P'_{n,k}\,(t^{n+k}) \cong y'(t^{n+k}) = f(t^{n+k},y(t^{n+k})).$$

Αντικαθιστώντας στην σχέση αυτή τα $y(t^{n+k})$ με τα y^{n+k} παίρνουμε μια ενδιαφέρουσα κατηγορία μεθόδων, τις μεθόδους "<u>οπισθοδρομικών διαφορών με k βήματα</u>"

$$(6)\ \sum_{j=1}^{k} j^{-1}\nabla^j y^{n+k} = hf^{n+k}, \quad 0\leq n\leq N-k,$$

όπου χρησιμοποιούμε τον συμβολισμό $\nabla^1y^n=y^n-y^{n-1}$, $\nabla^2y^n= \nabla^1(\nabla^1y^n)$, ... του λογισμού διαφορών. Οι μέθοδοι (6) σε κανονική μορφή (έτσι ώστε $\alpha_k=1$) γράφονται:

$$(6')\ \sum_{j=0}^{k} \alpha_j y^{n+j} = h\beta_k f^{n+k},$$

όπου

γιά k=1 : $\alpha_1=1$, $\alpha_0=-1$, $\beta_1=1$ (πεπλεγμένη Euler),

γιά k=2: $\alpha_2=1$, $\alpha_1=-4/3$, $\alpha_0=1/3$, $\beta_2=2/3$,

γιά k=3: $\alpha_3=1$, $\alpha_2=-18/11$, $\alpha_1=9/11$, $\alpha_0=-2/11$, $\beta_3=6/11$

κ.ο.κ.

Γιά μία συστηματική μελέτη των τρόπων κατασκευής πολυβηματικών μεθόδων παραπέμπουμε στο βιβλίο του Henrici [3.4]. Συνήθως, k-βηματικές μέθοδοι της μορφής

$$(7)\quad y^{n+k} - y^{n+k-1} = h\sum_{j=0}^{k} \beta_j f^{n+j}$$

λέγονται μέθοδοι <u>Adams</u>. Ειδικώτερα, αν $\beta_k=0$ (άμεσες), είναι γνωστές ως μέθοδοι <u>Adams-Bashforth</u> (1883). Αν $\beta_k\neq 0$ παίρνουμε τις μεθόδους <u>Adams-Moulton</u>. Μέθοδοι της μορφής

$$(8)\quad y^{n+k}-y^{n+k-2} = h\sum_{j=0}^{k} \beta_j f^{n+j}$$

λέγονται συνήθως μέθοδοι του <u>Nyström</u> αν $\beta_k=0$ και μέθοδοι των <u>Milne-Simpson</u> αν $\beta_k\neq 0$.

Προχωρούμε τώρα στην ανάλυση των πολυβηματικών μεθόδων. Για να απλουστεύσουμε τα αποτελέσματα και τον συμβολισμό περιοριζόμαστε στην περίπτωση <u>μιας</u> Δ.Ε., δηλ. θεωρούμε το πρόβλημα (3.1.1) στον $\mathbb{R}^1$, ακολουθώντας την ανάλυση του Henrici [3.4, Κεφ.5]. (Για συστήματα ισχύουν εντελώς ανάλογα· βλ. π.χ., το βιβλίο του Gear [3.2, Κεφ.10].) Ανακαλούμε πρώτα ορισμένα αποτελέσματα από την θεωρία των <u>εξισώσεων διαφορών</u> με σταθερούς συντελεστές, εντελώς ανάλογα παρόμοιων αποτελεσμάτων για διαφορικές εξισώσεις ανωτέρας τάξης με σταθερούς συντελεστές. Θεωρούμε την ομογενή εξίσωση διαφορών

$$(9)\quad \alpha_k y^{n+k} + \alpha_{k-1}y^{n+k-1} + \dots + \alpha_0 y^n=0 \quad , \qquad n\geq 0,$$

όπου α_j, $0\leq j\leq k$ <u>σταθερές</u> με α_k, $\alpha_0\neq 0$. Ζητάμε να βρούμε λύσεις y^n της

(9) για $n\geq 0$, δηλ. γενικά μιγαδικές ακολουθίες $\{y^n\}$, $n\geq 0$ που κάνουν την (9) ταυτότητα. Υποθέτοντας ότι υπάρχουν λύσεις της μορφής $y^n=z^n$ (n-στή δύναμη του z), όπου $z \in \mathbb{C}$, $z\neq 0$ βλέπουμε ότι πρέπει να ισχύει

$$\alpha_k z^{n+k}+\ldots+\alpha_0 z^n=0 \quad , \quad n\geq 0,$$

δηλ. ότι το z πρέπει να είναι ρίζα του πολυωνύμου

$$(10) \qquad p(z) \equiv \alpha_k z^k + \alpha_{k-1} z^{k-1} + \ldots + \alpha_0 .$$

Αντίστροφα, κάθε ρίζα z του (10) ορίζει την λύση $y^n = z^n$ της (9). Ισχύει δε το εξής βασικό αποτέλεσμα, για την (εύκολη) απόδειξη του οποίου βλ. π.χ. [3.4, σελ. 213-4]: Αν το πολυώνυμο p(z) έχει k διακριτές ρίζες $z_1,\ldots,z_k$, τότε η γενική λύση της ομογενούς εξίσωσης διαφορών (9) δίδεται από τον γραμμικό συνδυασμό

$$(11) \qquad y^n = c_1 z_1^{\,n} + c_2 z_2^{\,n} + \ldots c_k z_k^{\,n}$$

για οποιεσδήποτε σταθερές $c_i\in\mathbb{C}$ (οι οποίες μπορούν να προσδιορισθούν μονοσήμαντα π.χ. αν δίνονται οι αρχικές τιμές y^j, $0\leq j\leq k-1$· οδηγούμαστε έτσι σε ειδικές λύσεις της (9)). Αν το πολυώνυμο p(z) έχει $m \leq k$ διακριτές ρίζες $z_1,\ldots,z_m$ με πολλαπλότητες $p_1,\ldots,p_m$ αντίστοιχα $(p_1+\ldots+p_m=k)$, τότε η γενική λύση της (9) δίνεται από τον γραμμικό συνδυασμό

$$(12) \qquad y^n = c_1\zeta_1(n) +\ldots+ c_k\zeta_k(n) \ , \quad c_i \in \mathbb{C}$$

όπου

$$\zeta_1(n) = z_1^n$$

$$\zeta_2(n) = nz_1^n$$

$$\vdots$$

$$\zeta_{p_1}(n) = n(n-1)\ldots(n-p_1+2)z_1^n$$

$$\zeta_{p_1+1}(n) = z_2^n$$

$$\vdots$$

$$\zeta_{p_1+p_2}(n) = n(n-1)\ldots(n-p_2+2)z_2^n$$

$$\vdots$$

$$\zeta_k(n) = n(n-1)\ldots(n-p_m+2)z_m^n.$$

Σημαντικό ρόλο στην ανάλυση της <u>σύγκλισης</u> των πολυβηματικών μεθόδων παίζουν οι έννοιες της <u>ευστάθειας</u> και της <u>συνέπειας</u>. Το ξεκαθάρισμα των εννοιών αυτών και τα θεωρήματα που τις συνδέουν οφείλονται κυρίως στον Dahlquist (1956). Λέμε ότι η μέθοδος (3) είναι <u>συγκλίνουσα</u> (ή ότι <u>συγκλίνει</u>) αν για κάθε πραγματική συνάρτηση $f(t,y)$ - που ικανοποιεί τις συνθήκες του Θεωρήματος 3.1.1 στον R^1 - και κάθε $y_0 \in R^1$, συμβολίζοντας με $y(t)$ την λύση του προβλήματος (3.1.1) στον R^1, έχουμε για κάθε $t \in [a,b]$ ότι ισχύει

(13) $\lim y^n = y(t)$ όταν $h \to 0$, $n \to \infty$, $t^n \equiv \alpha + nh \to t$,

για κάθε λύση y^n της (3) που έχει αρχικές τιμές $y^0, \ldots, y^{k-1}$ τέτοιες ώστε

(14) $\lim_{h \to 0} y^j = y_0$, $\quad 0 \leq j \leq k-1$.

Θα δούμε στην συνέχεια δύο σημαντικές συνέπειες του να είναι μια πολυβηματική μέθοδος συγκλίνουσα. Η πρώτη είναι η ευστάθειά της. Λέμε ότι η μέθοδος (3) είναι ευσταθής αν οι ρίζες z_i του πολυωνύμου $p(z)$ που δίνεται από την (10) ικανοποιούν την λεγόμενη συνθήκη των ριζών:

(15) $|z_i| \leq 1 \ \forall i$ και $|z_i| < 1$ αν η z_i είναι πολλαπλή ρίζα.

ΠΡΟΤΑΣΗ 1 (Dahlquist). Αν η μέθοδος (3) συγκλίνει, τότε είναι ευσταθής.

Απόδειξη: Εστω ότι η μέθοδος (3) συγκλίνει. Τότε θα συγκλίνει και για το πρόβλημα αρχικών τιμών $y'=0$, $y(0)=0$ του οποίου η λύση είναι $y(t) \equiv 0$. Για το πρόβλημα αυτό η (3) έχει την μορφή

(16) $\alpha_k y^{n+k} + \ldots + \alpha_0 y^n = 0$, $\quad n \geq 0$.

Για κάθε λύση της (16) λοιπόν τέτοια ώστε

(17) $\lim\limits_{h \to 0} y^j = 0$, $\ 0 \leq j \leq k-1$,

θα ισχύει, για κάθε $t \geq 0$, ότι

(18) $\lim_{n\to\infty} y^n = 0$, $h = t/n$.

Εστω $\zeta = re^{i\varphi}$, $r \geq 0$, $0 \leq \varphi < 2\pi$, μία ρίζα του πολυωνύμου $p(z)$, (10). Θεωρείστε τους αριθμούς

(19) $y^n = h\, Re(\zeta^n) = hr^n\cos n\varphi$, $n=0,1,2,\ldots$,

οι οποίοι, επειδή $\alpha_j \in R^1$, ικανοποιούν την (16) για $n\geq 0$ με βάση τα όσα αναφέρθηκαν πιο πάνω για την εξίσωση διαφορών (9). Επίσης είναι προφανές ότι ικανοποιούν την (17). Συνεπώς για την ακολουθία (19) πρέπει να ισχύει η (18) για κάθε $t>0$. Αν $\varphi=0$ ή $\varphi=\pi$ τότε από την (18) έχουμε αναγκαστικά ότι $r\leq 1$. Αν τώρα $\varphi\neq 0$ και $\varphi\neq\pi$, παρατηρούμε ότι

$$\begin{aligned}(y^n)^2-y^{n+1}y^{n-1} &= h^2r^{2n}\cos^2 n\varphi - h^2r^{2n}\cos(n+1)\varphi\,\cos(n-1)\varphi \\ &= h^2r^{2n}\sin^2\varphi,\end{aligned}$$

δηλ. ότι

$$[(y^n)^2-y^{n+1}y^{n-1}]/\sin^2\varphi = h^2r^{2n}.$$

Το αριστερό μέλος αυτής της ισότητας τείνει στο μηδέν λόγω της (18) όταν $n \to \infty$. Άρα για κάθε $t>0$ $h^2r^{2n} = t^2(r^n/n)^2 \to 0$, $n \to \infty$. Αναγκαστικά λοιπόν $r\leq 1$. Συμπεραίνουμε ότι "σύγκλιση της (3)" => "κάθε ρίζα ζ του $p(z)$ ικανοποιεί $|\zeta|\leq 1$".

Εστω τώρα $\zeta = re^{i\varphi}$ μια ρίζα του $p(z)$ με πολλαπλότητα μεγαλύτερη της μονάδας. Από την θεωρία των λύσεων της ομογενούς εξίσωσης διαφορών (9) σ' αυτήν την περίπτωση, συμπεραίνουμε ότι η ακολουθία

(20) $y^n = h^{1/2}\, nr^n\cos n\varphi$, $n=0,1,2,\ldots$

αποτελεί λύση της (16) που ικανοποιεί την (17). Συνεπώς, θα ικανοποιεί την (18) για κάθε $t>0$. Αν $\varphi=0$ ή $\varphi=\pi$ συμπεραίνουμε ότι $|y^n|=h^{1/2}nr^n = t^{1/2}n^{1/2}r^n \to 0$, $n \to \infty$ για $t>0$. Συνεπώς $r<1$. Αν $\varphi\neq 0$ ή $\varphi\neq\pi$, εύκολα βλέπουμε ότι αν $z^n = n^{-1}h^{-1/2}y^n$, τότε

$$(21) \qquad ((z^n)^2 - z^{n+1}z^{n-1})/\sin^2\varphi = r^{2n}.$$

Επειδή $z^n = y^n/nh^{1/2} = y^n/(nt)^{1/2} \longrightarrow 0 \quad \forall\, t>0$ όταν $n \longrightarrow \infty$ (λόγω της (18)) η σύγκλιση στο 0 του αριστερού μέλους της (21) για $n \longrightarrow \infty$ δίνει $r<1$ πάλι. Συνεπώς κάθε πολλαπλή ρίζα ζ του $p(z)$ πρέπει να βρίσκεται στον ανοικτό μοναδιαίο δίσκο, δηλ. να ικανοποιεί $|\zeta| < 1$. @

Εξετάζοντας τις ρίζες του αντίστοιχου $p(z)$ βλέπουμε ότι οι μέθοδοι Euler, πεπλεγμένη Euler και τραπεζίου (για τις οποίες $p(z) = z-1$), η μέθοδος (1') και η μέθοδος Simpson (2) (για τις οποίες $p(z)=z^2-1$) ικανοποιούν την συνθήκη των ριζών (15) και συνεπώς είναι ευσταθείς. (Ευσταθείς είναι γενικά οι μέθοδοι Adams (7) για τις οποίες $p(z)=z^{k-1}(z-1)$ και οι μέθοδοι (8) για τις οποίες $p(z) = z^{k-2}(z^2-1)$). Έχει αποδειχθεί (Cryer) ότι οι μέθοδοι οπισθοδρομικών διαφορών (6) είναι ευσταθείς αν και μόνο αν $1 \leq k \leq 6$ (βλ. και Άσκηση 8).

Αν μία μέθοδος <u>δεν είναι ευσταθής</u>, τότε και για την απλούστατη Δ.Ε. $y' = 0$, $y(0)=0$ με $y^j=0$, $0 \leq j \leq k-1$, <u>λόγω σφαλμάτων στρογγύλευσης</u>, η λύση δεν θα είναι $y^n=0$, $n\geq 0$ αλλά θα έχει γενικά μια συνιστώσα - βλ. (11), (12) - που δεν θα μένει φραγμένη καθώς αυξάνει το n. Στην πράξη αν υπολογίσουμε με μια ασταθή μέθοδο βλέπουμε πολύ γρήγορα μια "έκρηξη" της αριθμητικής λύσης y^n καθώς αυξάνει το n. Η χρήση μικρότερου h δεν βελτιώνει (αντίθετα χειροτερεύει) την κατάσταση' βλ. [5.3, σελ 188] για αριθμητικά παραδείγματα.

Θα δούμε αργότερα (Παρ. 3.4) ότι για πολλά ενδιαφέροντα προβλήματα η έννοια της "ευστάθειας" που εισαγάγαμε δεν είναι επαρκής' δηλ. ότι οι μέθοδοι που ικανοποιούν την συνθήκη των ριζών (15) μπορεί να μην συμπεριφέρονται καθόλου ικανοποιητικά στην πράξη σε περιβάλλον σφαλμάτων στρογγύλευσης. Προς το παρόν όμως στρεφόμαστε στην μελέτη της <u>τάξης ακρίβειας</u> και της <u>συνέπειας</u> των πολυβηματικών μεθόδων.

Δεδομένης της k-βηματικής μεθόδου (3) (δηλ. των σταθερών $\{\alpha_j, \beta_j\}$, $0 \leq j \leq k$), θεωρούμε για $a \leq t \leq b$ και γιά μια αρκετά ομαλή συνάρτηση $y(t)$ την ποσότητα

$$(22) \quad L[\,y(t);h] = \sum_{j=0}^{k} \{\alpha_j y(t+jh) - h\beta_j y'(t+jh)\},$$

που παίζει τον ρόλο της ποσότητας $y(t+h)-y(t)-h\Phi(t,y(t),h)$ των μεθόδων RK. Αναπτύσσοντας το δεύτερο μέλος της (22) σε σειρά Taylor γύρω απ' το σημείο t έχουμε

$$(23) \quad L[y(t);h] = C_0 y(t) + C_1 h y'(t) + C_2 h^2 y''(t) + \ldots + C_j h^j y^{(j)}(t) + \ldots \quad ,$$

όπου οι σταθερές C_j είναι ανεξάρτητες των $h, t, y(t)$ και εξαρτώνται μόνο από την μέθοδο (3). Μάλιστα εύκολα αποδεικνύεται ότι

$$C_0 = \alpha_0 + \alpha_1 + \ldots + \alpha_k = \sum_{j=0}^{k} \alpha_j \,,$$

$$(24) \quad C_1 = \alpha_1 + 2\alpha_2 + \ldots + k\alpha_k - (\beta_0 + \ldots + \beta_k) = \sum_{j=1}^{k} j\alpha_j - \sum_{j=0}^{k} \beta_j \,,$$

και

$$C_j = (\alpha_1 + 2^j\alpha_2 + 3^j\alpha_3 + \ldots + k^j\alpha_k)/j! - (\beta_1 + 2^{j-1}\beta_2 + 3^{j-1}\beta_3 + \ldots + k^{j-1}\beta_k)/(j-1)! \quad \text{για } j \geq 2.$$

Λέμε ότι η μέθοδος (3) έχει <u>τάξη ακρίβειας</u> p αν στο ανάπτυγμα (23), $C_0 = C_1 = \ldots = C_p = 0$ αλλά $C_{p+1} \neq 0$, δηλ. αν $L[y(t);h] = C_{p+1} h^{p+1} y^{(p+1)}(t) + \ldots$. Εύκολα βλέπουμε ότι οι τάξεις των μεθόδων Euler, πεπλεγμένης Euler και τραπεζίου είναι αντίστοιχα p=1, 1 και 2. Η τάξη της μεθόδου (1') είναι $p=2$ ενώ της μεθόδου του Simpson (2) είναι $p=4$. Οι μέθοδοι (6) έχουν τάξη ακρίβειας $p=k$.

Λέμε ότι η μέθοδος (3) είναι <u>συνεπής</u> αν έχει τάξη (ακρίβειας) τουλάχιστον 1. Από τις (24) βλέπουμε ότι η (3) είναι συνεπής αν $C_o=C_1=0$ δηλ. αν

$$(25) \quad \sum_{j=0}^{k} \alpha_j=0, \quad \sum_{j=1}^{k} j\alpha_j = \sum_{j=0}^{k} \beta_j .$$

Εισάγοντας, εκτός από το πολυώνυμο $p(z)$ (10), το πολυώνυμο $\sigma(z)$:

$$(26) \quad \sigma(z) = \beta_k z^k + \beta_{k-1} z^{k-1} + \ldots + \beta_o ,$$

βλέπουμε ότι οι συνθήκες (25) της συνέπειας μιάς μεθόδου γράφονται ισοδύναμα ως

$$(25') \quad p(1) = 0, \; p'(1)=\sigma(1).$$

ΠΡΟΤΑΣΗ 2. Αν η μέθοδος (3) συγκλίνει τότε είναι συνεπής.

<u>Απόδειξη</u>: 'Εστω ότι η (3) συγκλίνει. Τότε θα συγκλίνει και γιά το πρόβλημα $y'=0$, $y(0)=1$ του οποίου η λύση είναι $y(t)=1$. Γιά το πρόβλημα αυτό η (3) πάλι έχει την μορφή (16). Επειδή η μέθοδος συγκλίνει, η λύση $\{y^n\}$ της (16) που αντιστοιχεί στις αρχικές τιμές $y^j=1$, $0\le j\le k-1$ (γιά την οποία δηλ. ισχύουν τετριμμένα οι (14)), πρέπει να ικανοποιεί την (13) δηλ. να ισχύει ότι γιά κάθε $t>0$

$$(26) \quad \lim y^n = 1, \; n\to\infty, \; h\to 0, \; nh=t.$$

Λόγω όμως της (16) - που δεν εξαρτάται από το h - και των αρχικών τιμών $y^j=1$, $0\le j\le k-1$ - που δεν εξαρτώνται από το h - η ακολουθία $\{y^n\}$ δεν εξαρτάται από το h ή το t. Συνεπώς η (26) λέει απλώς ότι

$$(26') \quad \lim y^n=1, \; n\to\infty.$$

'Αρα, αν το n τείνει στο ∞ στην (16) παίρνουμε ότι $\alpha_k+\alpha_{k-1}+\ldots+\alpha_o=0$, δηλ. ότι $C_o=0$.

Γιά να αποδείξουμε ότι $C_1=0$ θεωρούμε το πρόβλημα αρχικών τιμών $y'=1$, $y(0)=0$ με λύση $y(t)=t$. Η (3) παίρνει τώρα την μορφή

(27) $\alpha_k y^{n+k}+\dots+\alpha_0 y^n=h(\beta_k+\dots+\beta_0)$.-

Επειδή η (3) συγκλίνει έχουμε ότι κάθε λύση της (27) γιά την οποία

(28) $\lim_{h\to 0} y^j = 0,\ 0\le j\le k-1,$

θα πρέπει να ικανοποιεί, γιά κάθε $t>0$, την

(29) $\lim_{h\to 0} y^n=t,\ nh=t.$

Θεωρούμε την ακολουθία y^n, $n\ge 0$ που ορίζεται ως

(30) $y^n=nhK,\ K\equiv\sigma(1)/\rho'(1),\ n\ge 0.$

(Από την Πρόταση 1 λόγω της ευστάθειας της μεθόδου έχουμε $\rho'(1)\neq 0$). Προφανώς γιά την (30) ισχύει η (28). Έχουμε επίσης ότι

$$\sum_{j=0}^{k} \alpha_j y^{n+j} = hK \sum_{j=0}^{k} (n+j)\alpha_j = (\text{λόγω της } C_0=0) =$$

$$hK \sum_{j=0}^{k} j\alpha_j = hK\rho'(1) = h \sum_{j=0}^{k} \beta_j,$$

δηλ. ότι η (30) είναι λύση της (27). Συνεπώς πρέπει να ικανοποιεί την (29) η οποία μας δίνει ότι γιά $t>0$ $t=\lim_{h\to 0} nhK = \lim_{h\to 0} tK = tK$, δηλ. ότι $K=1 \iff \rho'(1)=\sigma(1) \iff C_1=0$. @

Συνεπώς η συνέπεια μιάς μεθόδου είναι αναγκαία γιά την σύγκλισή της. Στο [5.3, σελ. 189] βλέπουμε ένα τυπικό παράδειγμα υπολογισμού με μιά ευσταθή αλλά μη συνεπή μέθοδο: Το σφάλμα μεγαλώνει καθώς $h\to 0$ αλλά η ακολουθία y^n μένει φραγμένη· πλησιάζει δε την λύση κάποιου άλλου προβλήματος αρχικών τιμών!

'Οπως θα δούμε παρακάτω, (πόρισμα του θεωρήματος 1), συνέπεια+ευστάθεια => σύγκλιση. δηλ. η {συνέπεια και ευστάθεια} είναι ικανή και αναγκαία συνθήκη γιά σύγκλιση. Ας σημειώσουμε όμως το εξής σημαντικό αποτέλεσμα του Dahlquist (1956) το οποίο περιορίζει την δυνατή τάξη ακρίβειας μιάς ευσταθούς μεθόδου. Η απόδειξη μπορεί να βρεθεί στο βιβλίο του Henrici [3.4, σελ. 229]:

ΠΡΟΤΑΣΗ 3. Η μέγιστη τάξη μιάς ευσταθούς k-βηματικής μεθόδου της μορφής (3) είναι p=k+1 αν k περιττός και p=k+2 αν k άρτιος. @

<u>Μέθοδοι βέλτιστης τάξης</u> γιά δεδομένο αριθμό βημάτων k λέγονται λοιπόν οι μέθοδοι με τάξη p=k+1 αν k περιττός και p=k+2 αν k άρτιος. Ο προσδιορισμός τους παρουσιάζει θεωρητικό και πρακτικό προφανώς ενδιαφέρον (βλ. Ασκήσεις 3,6). Προφανώς η μέθοδος του τραπεζίου (k=1, p=2) και η μέθοδος του Simpson (k=2, p=4) είναι μέθοδοι βέλτιστης τάξης γιά k=1,2 αντίστοιχα.

Προχωρούμε τώρα στην απόδειξη του βασικού αποτελέσματος αυτής της παραγράφου, δηλ. στην απόδειξη ενός φράγματος άριστης τάξης σύγκλισης γιά το σφάλμα της μεθόδου (3). Πρώτα δύο προκαταρτικά αποτελέσματα:

Λήμμα 1. 'Εστω ότι το πολυώνυμο $p(z) = \sum_{j=0}^{k} \alpha_j z^j$ ικανοποιεί την συνθήκη των ριζών (15). Ορίζουμε τους συντελεστές γ_j, $j \geq 0$, από το ανάπτυγμα

$$(31) \quad 1/(\alpha_k+\alpha_{k-1}z+\dots+\alpha_0 z^k) = \gamma_0+\gamma_1 z+\gamma_2 z^2+\dots .$$

Τότε τα γ_j ικανοποιούν τις συνθήκες

$$(32) \quad \begin{aligned} &\alpha_k \gamma_0 = 1, \\ &\alpha_k\gamma_j+\alpha_{k-1}\gamma_{j-1}+\dots+\alpha_{k-j}\gamma_0 = 0 \text{ αν } 1\leq j\leq k, \\ &\vdots \\ &\alpha_k\gamma_j+\alpha_{k-1}\gamma_{j-1}+\dots+\alpha_0\gamma_{j-k} = 0 \text{ αν } j>k. \end{aligned}$$

Επιπλέον,

(33) $\Gamma \equiv \sup_{j\geq o} |\gamma_j| < \infty$.

<u>Απόδειξη</u>: Έστω $\hat{p}(z) = \alpha_k+\alpha_{k-1}z+\ldots+\alpha_o z^k=z^k p(z^{-1})$. Επειδή, λόγω ευστάθειας, το πολυώνυμο $p(z)$ δεν έχει ρίζες με $|z|>1$ και επειδή $\alpha_k\neq 0$, συμπεραίνουμε ότι το πολυώνυμο $\hat{p}(z)$ δεν έχει ρίζες z τέτοιες ώστε $|z|<1$. Συνεπώς η συνάρτηση $1/\hat{p}(z)$ είναι αναλυτική στον ανοιχτό δίσκο $D=\{z\in\mathbb{C}:|z|<1\}$. Θεωρούμε το ανάπτυγμά της (31) κατά Taylor γύρω απ' το μηδέν. Οι ταυτότητες (32) προκύπτουν εύκολα από την (31) στην μορφή

$$1=(\alpha_k+\alpha_{k-1}z+\ldots+\alpha_o z^k)(\gamma_o+\gamma_1 z+\gamma_2 z^2+\ldots),$$

μετά από πολλαπλασιασμό των σειρών του δευτέρου μέλους και εξίσωση δυνάμεων του z των δύο μελών. (Επειδή $\alpha_k\neq 0$ οι συνθήκες (32) προσδιορίζουν μονοσήμαντα τα γ_j, $j\geq 0$ συναρτήσει των α_j, $0\leq j\leq k$).

Γιά να αποδείξουμε την εκτίμηση (33) γιά τους συντελεστές της σειράς Taylor του $(\hat{p}(z))^{-1}$, παρατηρούμε πρώτα ότι λόγω της (15) οι ρίζες $z_1,\ldots,z_m$ του $p(z)$ που ικανοποιούν $|z_i|=1$ είναι απλές (αν κάν υπάρχουν). Συμπεραίνουμε ότι οι μόνοι πόλοι της συνάρτησης $(\hat{p}(z))^{-1}$ στην περιφέρεια $|z|=1$ είναι οι <u>απλοί</u> πόλοι στα σημεία z_i^{-1}, $1\leq i\leq m$. Άρα, από γνωστό μας αποτέλεσμα της θεωρίας μιγαδικών συναρτήσεων, υπάρχουν σταθερές A_i, $1\leq i\leq m$ τέτοιες ώστε η συνάρτηση

(34) $f(z)=(\hat{p}(z))^{-1}-A_1/(z-z_1^{-1})-\ldots-A_m/(z-z_m^{-1})$,

να είναι αναλυτική για $|z|\leq 1$. Θεωρούμε την σειρά Taylor της $f(z)$ περί το μηδέν:

$$f(z) = \sum_{n=0}^{\infty} z^n f^{(n)}(0)/n!$$

Από το θεώρημα του Cauchy αν $C=\{z:|z|=1\}$ έχουμε

$$f^{(n)}(0) = n!\ (2\pi i)^{-1} \int_C (f(z)/z^{n+1})dz,$$

από την οποία συμπεραίνουμε ότι $|f^{(n)}(0)|/n! \leq \sup_{z\in C} |f(z)| \equiv M < \infty$, δηλ. ότι οι συντελεστές $f^{(n)}(0)/n!$ της σειράς Taylor της f περί το μηδέν είναι ομοιόμορφα φραγμένοι από κάποιο $M<\infty$. Επειδή εξ άλλου γιά $|z|<1$ επειδή $|z_i|=1$

$$-A_i/(z-z_i^{-1}) = A_i z_i(1+z_i z+(z_i z)^2+\ldots+(z_i z)^n+\ldots),$$

βλέπουμε ότι οι συντελεστές της σειράς Taylor της συνάρτησης $-A_i(z-z_i^{-1})$ περί το $z=0$ φράσσονται από $|A_i|$. Συμπεραίνουμε από την (34) ότι οι συντελεστές γ_i της σειράς Taylor της $1/\hat{p}(z)$ περί το 0 φράσσονται ομοιόμορφα:

$$\Gamma = \sup_{j\geq 0} |\gamma_j| \leq M + \sum_{i=1}^{m} |A_i| < \infty\ @$$

Θα χρησιμοποιήσουμε το παραπάνω λήμμα γιά να εκτιμήσουμε a priori τις λύσεις μιάς μη ομογενούς εξίσωσης διαφορών που σχετίζεται με την πολυβηματική μέθοδο (3):

Λήμμα 2. 'Εστω ότι η k-βηματική μέθοδος (3) είναι ευσταθής. 'Εστω λ^n, $0\leq n\leq N-k$ δεδομένες σταθερές και έστω β_i^n, $0\leq i\leq k$, $0\leq n\leq N-k$ δεδομένοι αριθμοί, τέτοιοι ώστε $|\beta_i^n|\leq B<\infty$ $\forall\, i,n$. Θεωρείστε την εξίσωση διαφορών

$$(35)\quad \alpha_k \psi^{n+k}+\alpha_{k-1}\psi^{n+k-1}+\ldots+\alpha_0\psi^n =$$

$$= h(\beta_k^n \psi^{n+k}+\beta_{k-1}^n\psi^{n+k-1}+\ldots+\beta_0^n\psi^n)+\lambda^n,\ 0\leq n\leq N-k.$$

Τότε υπάρχει $h_0>0$ τέτοιο ώστε γιά $0\leq h\leq h_0$:

$$(36)\quad \max_{0\le n\le N}|\psi^n|\le C(N\max_n|\lambda^n| + \max_{0\le j\le k-1}|\psi^j|),$$

όπου η σταθερά C εξαρτάται από τα b-a, h_0, B και την μέθοδο (3) αλλά όχι από τα h,λ^n,ψ^n,N.

<u>Απόδειξη</u>: Για $k\le m\le N$ θεωρούμε την (35) για $n=0,1,2,\dots,m-k$. Για $n=m-k-j$, $0\le j\le m-k$ πολλαπλασιάζουμε την (35) επί γ_j (που ορίστηκαν στο Λήμμα 1) και προσθέτουμε κατά μέλη τις εξισώσεις που προκύπτουν. Από το αριστερό μέλος παίρνουμε το άθροισμα

$$
\begin{aligned}
S_m \equiv{}& \gamma_0(\alpha_k\psi^m+\alpha_{k-1}\psi^{m-1}+\dots+\alpha_0\psi^{m-k}) + \gamma_1(\alpha_k\psi^{m-1}+\alpha_{k-1}\psi^{m-2}\\
(37)\qquad &+\dots+\alpha_0\psi^{m-k-1})+\dots+\gamma_{m-k}(\alpha_k\psi^k+\alpha_{k-1}\psi^{k-1}+\dots+\alpha_0\psi^0)=\\
&\alpha_k\gamma_0\psi^m+(\alpha_k\gamma_1+\alpha_{k-1}\gamma_0)\psi^{m-1}+\dots+(\alpha_k\gamma_{m-k}+\\
&\alpha_{k-1}\gamma_{m-k-1}+\dots+\alpha_0\gamma_{m-2k})\psi^k+(\alpha_{k-1}\gamma_{m-k}+\dots+\alpha_0\gamma_{m-2k+1})\psi^{k-1}\\
&+\dots+\alpha_0\gamma_{m-k}\psi^0= \text{(χρησιμοποιώντας τις ταυτότητες (32))}\\
&= \psi^m+(\alpha_{k-1}\gamma_{m-k}+\dots+\alpha_0\gamma_{m-2k+1})\psi^{k-1}+\dots+\alpha_0\gamma_{m-k}\psi^0.
\end{aligned}
$$

Από το δεξιό μέλος παίρνουμε μετά από λίγες ανακατατάξεις όρων

$$
\begin{aligned}
(38)\quad S_m ={}& h\{\beta_k^{m-k}\gamma_0\psi^m+(\beta_{k-1}^{m-k}\gamma_0+\beta_k^{m-k-1}\gamma_1)\ \psi^{m-1}+\dots+\\
&(\beta_0^{m-k}\gamma_0+\dots+\beta_k^{m-2k}\gamma_k)\ \psi^{m-k}+\dots+\beta_0^0\gamma_{m-k}\psi^0\}\\
&+\lambda^{m-k}\gamma_0+\lambda^{m-k-1}\gamma_1+\dots+\lambda^0\gamma_{m-k}.
\end{aligned}
$$

Εξισώνοντας τις (37) και (38) και λύνοντας ως πρός ψ^m (χρησιμοποιώντας ότι $\gamma_0=1/\alpha_k=1$ από την (32)) έχουμε:

$$
\begin{aligned}
(1-h\beta_k^{m-k})\psi^m ={}& h\{(\beta_{k-1}^{m-k}\gamma_0+\beta_k^{m-k-1}\gamma_1)\psi^{m-1}+\dots+\beta_0^0\gamma_{m-k}\psi^0\}\\
&-\{(\alpha_{k-1}\gamma_{m-k}+\dots+\alpha_0\gamma_{m-2k+1})\psi^{k-1}+\dots+\alpha_0\gamma_{m-k}\psi^0\}\\
&+\lambda^{m-k}\gamma_0+\lambda^{m-k-1}\gamma_1+\dots+\lambda^0\gamma_{m-k},
\end{aligned}
$$

από την οποία, χρησιμοποιώντας την (33) παίρνουμε

$$|1-h\beta_k^{m-k}|\ |\psi^m| \le C_1\ h \sum_{j=0}^{m-1} |\psi^j| + C_2 \sum_{j=0}^{k-1} |\psi^j| + \Gamma N \max_{0\le j\le m-k} |\lambda^j|,$$

όπου $C_1=(k+1)B\Gamma$, $C_2=\Gamma \sum_{j=0}^{k} |\alpha_j|$. Συνεπώς για $h\le h_0$ όπου $h_0<B^{-1}$, συμπεραίνουμε ότι υπάρχει σταθερά C' τ.ω.

$$(39)\quad |\psi^m| \le C'(h \sum_{j=0}^{m-1} |\psi^j| + N \max_j |\lambda^j| + \sum_{j=0}^{k-1} |\psi^j|),\ k\le m\le N.$$

Θέτουμε $A = C'(N \max_j |\lambda^j| + \sum_{j=0}^{k-1} |\psi^j|)$. Τότε ισχύει (τετριμμένα) ότι

$$(40)\quad |\psi^j| \le A(1+hC')^j,\ 0\le j\le k-1.$$

Θα δείξουμε επαγωγικά ότι η (40) ισχύει γιά $0\le j\le N$. Πράγματι, έστω ότι έχουμε

$$(41)\quad |\psi^j| \le A(1+hC')^j,\ 0\le j\le m-1.$$

Η (39) και η (41) δίνουν τότε

$$|\psi^m| \le C'\ h\ A \sum_{j=0}^{m-1} (1+hC')^j + A = A(C'h \sum_{j=0}^{m-1} (1+hC')^j+1)$$

$$= A\ (C'h\ \{(1+hC')^m-1\}/C'h + 1) = A(1+hC')^m.$$

Άρα ισχύει η (41) γιά $j=m$· το επαγωγικό βήμα τελείωσε. Συνεπώς γιά κάθε m, $0\le m\le N$, έχουμε

$$|\psi^m| \le A(1+hC')^m \le Ae^{hmC'} \le A\ e^{(b-a)C'}.$$

Άρα ισχύει η (36) με $C=e^{(b-a)C'}$. @

Ερχόμαστε τώρα στο κεντρικό αποτέλεσμα αυτής της παραγράφου:

ΘΕΩΡΗΜΑ 1. Υποθέτουμε ότι η k-βηματική μέθοδος (3) είναι ευσταθής και έχει τάξη ακρίβειας $p \geq 1$. Έστω ότι η λύση y του προβλήματος αρχικών τιμών (γιά μία απλή Δ.Ε.) (3.1.1) ανήκει στον χώρο $C^{p+1}[a,b]$. Υπάρχει τότε $h_0>0$ τέτοιο ώστε γιά $0 \leq h \leq h_0$ να ισχύει η εκτίμηση

$$(42)\quad \max_{0\leq n\leq N}|y^n-y(t^n)| \leq C(\max_{0\leq j\leq k-1}|y^j-y(t^j)|+h^p\max_{a\leq t\leq b}|y^{(p+1)}(t)|),$$

όπου C σταθερά ανεξάρτητη των y^n, $y(t)$, h,N.

<u>Απόδειξη</u>: Από την υπόθεσή μας ότι η (3) έχει τάξη p, συμπεραίνουμε ότι γιά τις ποσότητες

$$(42)\quad p^n = \sum_{j=0}^{k} [\alpha_j y(t^n+jh) - h\beta_j y'(t^n+jh)],\ 0\leq n\leq N-k,$$

ισχύει, από το θεώρημα του Taylor με υπόλοιπο, ότι γιά κάποια σταθερά C':

$$(44)\quad \max_{0\leq n\leq N-k}|p^n| \leq C'h^{p+1}\max_{a\leq t\leq b}|y^{(p+1)}(t)|.$$

Θεωρούμε τώρα το σφάλμα $e^n=y^n-y(t^n)$, $0\leq n\leq N$, το οποίο ικανοποιεί την εξίσωση, γιά $0\leq n\leq N-k$,

$$(45)\quad \alpha_k e^{n+k} + \alpha_{k-1}e^{n+k-1}+\ldots+\alpha_0 e^n =$$
$$=(\alpha_k y^{n+k}+\ldots+\alpha_0 y^n)-(\alpha_k y(t^{n+k})+\ldots+\alpha_0 y(t^n)) =$$

$$\text{(από τις (3), (43))},=h(\beta_k f^{n+k}+\ldots+\beta_0 f^n)-h[\beta_k f(t^{n+k},y(t^{n+k}))+\ldots$$
$$\ldots+\beta_0 f(t^n,y(t^n))]-p^n=$$
$$= h\{\beta_k[f^{n+k}-f(t^{n+k},y(t^{n+k}))]+\ldots+\beta_0[f^n-f(t^n,y(t^n))]\}-p^n.$$

Ορίζουμε τώρα g^n, $0\le n\le N$ ως

$$(46)\quad g^n = \begin{cases} [f^n-f(t^n,y(t^n))]/\epsilon^n & \text{αν } \epsilon^n\neq 0 \\ 0 & \text{αν } \epsilon^n=0, \end{cases}$$

οπότε η (45) γράφεται

$$(47)\quad \alpha_k\epsilon^{n+k}+\ldots+\alpha_0\epsilon^n = h\{\beta_k g^{n+k}\epsilon^{n+k}+\ldots+\beta_0 g^n\epsilon^n\}-p^n,\ 0\le n\le N-k.$$

Η (46) δίνει τώρα γιά $\epsilon^n\neq 0$, με χρήση της συνθήκης Lipschitz της f,

$$(48)\quad |g^n| \le L|y^n-y(t^n)|/|\epsilon^n|=L,\ 0\le n\le N.$$

που ισχύει βέβαια και γιά $\epsilon^n=0$. Ορίζουμε τους αριθμούς β_i^n, $0\le i\le k$, $0\le n\le N-k$, από τις σχέσεις $\beta_i^n=\beta_i g^{n+i}$, $0\le i\le k$, $0\le n\le N-k$. Λόγω της (48) έχουμε

$$(49)\quad \max_{i,n} |\beta_i^n| = \max_i |\beta_i| \max_n |g^n| \le \max_i |\beta_i|\cdot L \equiv B < \infty$$

Η (47) γράφεται λοιπόν ώς

$$\alpha_k\epsilon^{n+k}+\ldots+\alpha_0\epsilon^n = h\{\beta_k^n\epsilon^{n+k}+\ldots+\beta_0^n\epsilon^n\}-p^n,\ 0\le n\le N-k,$$

δηλ. ως μία εξίσωση διαφορών γιά τα ϵ^n με μεταβλητούς συντελεστές της μορφής (35) του Λήμματος 2. Εφαρμόζοντας τώρα την (36) παίρνοντας υπ' όψιν την (44) και ότι Nh=(b-a), καταλήγουμε στην (42). @

Το Θεώρημα 1 αποτελεί σχεδόν το αντίστροφο των Προτάσεων 1 και 2' μας διαβεβαιώνει ότι γιά τουλάχιστον $y\in C^2[a,b]$, "ευστάθεια και συνέπεια"=> "σύγκλιση". (Μας δίνει βέβαια και εκτίμηση του σφάλματος και γιαυτό είναι πολύ χρήσιμο). Μπορεί να αποδειχθεί, βλ. Henrici [3.4, Παρ. 5.3-3] ότι "ευστάθεια και συνέπεια" => "σύγκλιση", χωρίς να υποθέσουμε ότι $y \in C^2[a,b]$, δηλ. μόνο με τις συνθήκες στην f του θεωρήματος 3.1.1 που εγγυώνται $y \in C^1[a,b]$. Το αποτέλεσμα αυτό και οι Προτάσεις 1 και 2 συνοψίζονται λοιπόν με την διατύπωση "ευστάθεια και συνέπεια" <=> "σύγκλιση".

Είναι φανερό από την (42) ότι γιά να πάρουμε συνολικό σφάλμα $O(h^p)$ αρκεί $|y^j-y(t^j)|=O(h^p)$, $0<j\leq k-1$. Αρκεί δηλ. οι τιμές y^j, $0\leq j\leq k-1$, $(y^0=y_0)$ να υπολογισθούν με μιά μέθοδο RK <u>τάξης p-1</u> (της οποίας το τοπικό σφάλμα, όντας $O(h^p)$, εξασφαλίζει ότι $|y^j-y(t^j)|=O(h^p)$ γιά μικρό αριθμό βημάτων $0\leq j\leq k-1$.)

<u>Παρατηρήσεις</u>

1. Η συνθήκη των ριζών (15) είναι βέβαια εύκολο να ελεγχθεί π.χ. γιά δευτεροβάθμια πολυώνυμα p(z). Γιά πολυώνυμα βαθμού k>2 θα ήταν χρήσιμο να βρούμε αναλυτικές συνθήκες πάνω στούς συντελεστές του p(z) γιά την ισχύ της. Μιά χρήσιμη σχετική θεωρία είναι η λεγόμενη <u>θεωρία Schur</u>. Λέμε (βλ. π.χ. Miller, J.Inst. Math. Applics, 8 (1971), 397-406) ότι ένα πολυώνυμο με μιγαδικούς συντελεστές a_i, βαθμού k,

$$(50)\quad \pi(z) = a_0+a_1z+\dots+a_kz^k,$$

όπου $a_0\neq 0$, $a_k\neq 0$ είναι <u>(πολυώνυμο) Schur</u> αν όλες οι ρίζες του περιέχονται στον ανοιχτό μοναδιαίο δίσκο $D=\{z\in\mathbb{C}:|z|<1\}$. Λέμε ότι το π είναι <u>απλό (πολυώνυμο) von Neumann</u> αν όλες οι ρίζες του βρίσκονται στον κλειστό δίσκο $\bar{D}$ και αν μόνο απλές ρίζες του βρίσκονται πάνω στην περιφέρεια $|z|=1$. Μ' άλλα λόγια η συνθήκη των ριζών (15) είναι ισοδύναμη με την συνθήκη να είναι το p(z) απλό von Neumann.

Δεδομένου του π θεωρούμε το πολυώνυμο

$$(51)\quad \pi^*(z) \equiv \sum_{j=0}^{k} \bar{a}_{k-j}\, z^j = \bar{a}_k+\bar{a}_{k-1}z+\dots+\bar{a}_0z^k,$$

όπου $\bar{z}$ ο συζυγής του z. Προφανώς έχουμε

$$\pi^*(z)=z^k\bar{\pi}(z^{-1}).$$

Επίσης θεωρούμε το λεγόμενο "ανηγμένο πολυώνυμο"

$$(52)\ \pi_1(z) = (\pi^*(0)\pi(z) - \pi(0)\ \pi^*(z))/z,$$

βαθμού $\leq k-1$, και συμβολίζουμε με $\pi'(z)$ την παράγωγο του π. Τότε ισχύουν τα εξής:

(i) Το πολυώνυμο π είναι <u>Schur</u> αν και μόνο αν $|\pi^*(0)|>|\pi(0)|$ και το π_1 είναι Schur.

(ii) Το πολυώνυμο π είναι <u>απλό von Neumann</u> αν και μόνο αν είτε
(α) $|\pi^*(0)| > |\pi(0)|$ και το π_1 είναι απλό von Neumann
είτε
(β) $\pi_1 \equiv 0$ και το π' είναι Schur.

Με την θεωρία αυτή ανάγουμε λοιπόν το ερώτημα σε ανάλογο ερώτημα γιά ένα πολυώνυμο βαθμού κατά ένα μικρότερο και προχωρούμε κατά τον ίδιο τρόπο. Τα κριτήρια αυτά είναι αρκετά εύχρηστα γιά π.x. k=3 ή 4.

Αν και στην παράγραφο αυτή δεν μελετούμε πολυώνυμα Schur (εκτός π.x αν εφαρμόζουμε το κριτήριο (ii.β)) θα δούμε στην Παρ. 3.4 ότι γιά την λεγόμενη <u>απόλυτη ευστάθεια</u> μιάς πολυβηματικής μεθόδου, ανάλογα πολυώνυμα πρέπει να είναι Schur (και όχι απλά von Neumann). Το εξής κριτήριο <u>(Routh-Hurwicz, βλ. Lambert, [3.6, σελ. 80]) είναι</u> επίσης χρήσιμο (γιά k=2,3,4, κυρίως, στην πράξη). Θεωρούμε το πολυώνυμο π(z), (50), και κάνουμε την αλλαγή των μεταβλητών

$$(53)\ w=(1+z)/(1-z),\ z=(w-1)/(w+1)$$

που απεικονίζει (επί) την περιφέρεια $|w|=1$ στον φανταστικό άξονα $\mathrm{Re}z=0$, τον δίσκο $D=\{w:|w|<1\}$ στο ημιεπίπεδο $\mathrm{Re}z<0$, το σημείο $w=1$ στο

$z=0$, και το $w=-1$ στο επ' άπειρου σημείο $z=\infty$. Βλέπουμε εύκολα τότε ότι το $\eta(z)$ είναι Schur αν και μόνο αν το

$$\tilde{\eta}(z) = (1-z)^k \eta((1+z)/(1-z)) = b_0 z^k + \ldots + b_k$$

έχει όλες τις ρίζες του με <u>αρνητικά πραγματικά μέρη</u>. Γιά να ισχύει αυτό ικανές και αναγκαίες συνθήκες είναι οι συνθήκες των <u>Routh-Hurwicz</u> στους συντελεστές b_i, που γιά k=2,3,4 είναι οι

$k=2$: $b_i>0$, $0\le i\le 2$,

$k=3$: $b_i>0$, $0\le i\le 3$, $b_1b_2-b_3b_0>0$,

$k=4$: $b_i>0$, $0\le i\le 4$, $b_1b_2b_3-b_0b_3^2-b_4b_1^2>0$,

2. Ένας σχετικά απλός τρόπος γιά τον προσδιορισμό της τάξης p και της λεγόμενης "σταθερά σφάλματος" c^* (που θα ορισθεί παρακάτω) μιάς πολυβηματικής μεθόδου είναι ο εξής: Υποθέτουμε κατ' αρχήν ότι τα πολυώνυμα ρ και σ δεν έχουν κοινό παράγοντα και ότι η μέθοδος είναι συνεπής, οπότε $\rho(1)=0$, $\sigma(1)=\rho'(1)$. Συνεπώς $\sigma(1)\neq 0$ (αλλοιώς τα ρ,σ θα είχαν του κοινό παράγοντα $(z-1)$). Έστω ότι η μέθοδος (3) έχει τάξη p. Τότε γιά κάθε ομαλή συνάρτηση $\psi(t)$ έχουμε ((22), (23)), ότι όταν $h\to 0$

$$L[\psi(t);h] \equiv \sum_{j=0}^{k} \{\alpha_j\psi(t+jh)-h\beta_j\psi'(t+jh)\} \sim C_{p+1}h^{p+1}\psi^{(p+1)}(t),$$

όπου με το σύμβολο $\sim$ εννοούμε ότι ο πρώτος μη μηδενικός όρος του αναπτύγματος του $L[\psi(t);h]$ σε δυνάμεις του h γιά $h\to 0$ είναι ο $C_{p+1}h^{p+1}\psi^{(p+1)}(t)$. Επειδή τα C_i,p είναι ανεξάρτητα της $\psi(t)$ διαλέγουμε $\psi(t)=e^t$, οπότε η παραπάνω σχέση γράφεται

$$\sum_{j=0}^{k} (\alpha_j e^{t+jh}-h\beta_j e^{t+jh}) \sim C_{p+1}h^{p+1}e^t,$$

από την οποία, θέτοντας $e^h=\zeta$, δηλ. $h=\ln\zeta$, $\zeta>1$ παίρνουμε

$$\sum_{j=0}^{k}(\alpha_j\zeta^j-\ln\zeta\,\beta_j\,\zeta^j)\sim C_{p+1}(\ln\zeta)^{p+1},\ \zeta\downarrow 1.$$

Αναπτύσσοντας τον $\ln\zeta$ σε δυνάμεις του $\zeta-1$ παίρνουμε

$$\ln\zeta-[\rho(\zeta)/\sigma(\zeta)]\sim C_{p+1}(\zeta-1)^{p+1}/\sigma(\zeta),\ \zeta\downarrow 1,$$

και επειδή, καθώς $\zeta\to 1$, $\sigma(\zeta)=\sigma(1)+O((\zeta-1))$ και $\sigma(1)\neq 0$, έχουμε ότι $(\sigma(\zeta))^{-1}=(\sigma(1))^{-1}(1+O((\zeta-1)))$. Συνεπώς, ορίζοντας την "σταθερά σφάλματος" c^* της μεθόδου ως

$$(54)\quad c^*=C_{p+1}/\sigma(1),$$

παίρνουμε από την παραπάνω σχέση ότι

$$(55)\quad \ln\zeta-[\rho(\zeta)/\sigma(\zeta)]\sim c^*(\zeta-1)^{p+1},\ \zeta\downarrow 1.$$

Άρα, για να βρούμε τα c^*,p, αναπτύσσουμε τον $\ln\zeta$ σε δυνάμεις του $\zeta-1$:

$$\ln\zeta=(\zeta-1)-(\zeta-1)^2/2+(\zeta-1)^3/3+\ldots+(-1)^{n-1}(\zeta-1)^n/n+\ldots$$

καθώς και το πηλίκο $\rho(\zeta)/\sigma(\zeta)$ (παρατηρούμε ότι $\rho(\zeta)/\sigma(\zeta)=(\zeta-1)+O((\zeta-1)^2)$), και βρίσκουμε τον πρώτο μη μηδενικό όρο της διαφοράς των δύο αναπτυγμάτων.

3. Στην περίπτωση μιάς πεπλεγμένης πολυβηματικής μεθόδου πρέπει, όπως είδαμε, να λύσουμε σε κάθε βήμα ένα μη γραμμικό σύστημα για τον υπολογισμό του αγνώστου y^{n+k}. Στην περίπτωση των πολυβηματικών μεθόδων είναι πολύ συνηθισμένο στην πράξη να γίνεται μία αρκετά ακριβής <u>πρόβλεψη</u> y_0^{n+k} του y^{n+k} με μία βοηθητική άμεση πολυβηματική μέθοδο και να χρησιμοποιείται η κύρια, (πεπλεγμένη) μέθοδος για τον υπολογισμό <u>διορθώσεων</u> y_j^{n+k}, $j\geq 1$ με απλή επανάληψη. Προκύπτουν

έτσι ζευγάρια μεθόδων, οι λεγόμενες μέθοδοι <u>πρόβλεψης-διόρθωσης</u> (Π-Δ) που μπορούν να γραφούν γενικά στη μορφή

$$\underline{\text{Πρόβλεψη}}(\Pi):\ y_0^{n+k} + \sum_{j=0}^{k-1} \tilde{\alpha}_j y^{n+j} = h\sum_{j=0}^{k-1} \tilde{\beta}_j\, f^{n+j},\ (\tilde{\alpha}_k=1)$$

$$\underline{\text{Διόρθωση}}(\Delta):\ y_{i+1}^{n+k} + \sum_{j=0}^{k-1} \alpha_j y^{n+j} = h\beta_k f(t^{n+k}, y_i^{n+k}) +$$

$$+ h \sum_{j=0}^{k-1} \beta_j f^{n+j},\quad i=0,1,2,\ldots\quad (\alpha_k=1).$$

Η μέθοδος πρόβλεψης (Π) είναι μία άμεση k-βηματική μέθοδος $\{\tilde{\alpha}_j,\tilde{\beta}_j\}_{j=0}^{k}$, $\tilde{\beta}_k=0$, που παράγει μία <u>αρχική τιμή</u> y_0^{n+k} γιά την μέθοδο διόρθωσης (Δ). Η μέθοδος διόρθωσης χρησιμοποιείται είτε ως απλή επαναληπτική μέθοδος γιά την λύση του μη γραμμικού συστήματος $y^{n+k}=h\beta_k f^{n+k}+g^n$, οπότε η διόρθωση επαναλαμβάνεται μέχρις ότου π.χ. $\|y_{i+1}^{n+k}-y_i^{n+k}\|\leq\varepsilon$ γιά κάποιο δεδομένο μικρό ε της τάξεως του σφάλματος στρογγύλευσης, είτε γιά να παράγει ένα y_i^{n+k} <u>γιά μικρό i</u> (συνήθως (i=1 ή 2) το οποίο ορίζουμε ως y^{n+k} και προχωρούμε στο επόμενο βήμα. Στην δεύτερη περίπτωση το y^{n+k} δεν είναι βέβαια η ακριβής λύση του μη γραμμικού συστήματος που παριστάνει η πεπλεγμένη μέθοδος (Δ) αλλά μιά προσέγγιση, η οποία πολλές φορές (αν το ζευγάρι έχει κατάλληλες ιδιότητες ευστάθειας και ακρίβειας) είναι σχεδόν εξ ίσου ακριβής ως προσέγγιση της $y(t^{n+k})$ όσο και η λύση του μη γραμμικού συστήματος.

Δύο γνωστά ζευγάρια (Π)-(Δ) είναι:

(i) <u>Euler-Τραπεζίου</u>

(Π): $y^{n+1}-y^n=hf^n$

(Δ): $y^{n+1}-y^n=h(f^{n+1}+f^n)/2$

(ii) <u>Μέθοδος του Milne</u>

(Π): $y^{n+4}-y^n=4h(2f^{n+3}-f^{n+2}+2f^{n+1})/3$

(Δ): $y^{n+4}-y^{n+2}=h(f^{n+4}+4f^{n+3}+f^{n+2})/3$

Αν η τάξη ακρίβειας της (Π) είναι $\tilde{p}$ και της (Δ) είναι p, τότε, ανάλογα με τον αριθμό $m\geq 1$ των διορθώσεων (δηλ. της εφαρμογής της (Δ) γιά $i=0,1,2,\ldots,m-1$), μπορούμε να υπολογίσουμε την τάξη ακρίβειας του ζεύγους (που ορίζεται όπως π.χ. ορίζουμε την τάξη ακρίβειας των μεθόδων RK). Π.χ. αν $m\geq 1$ και $\tilde{p}\geq p-1$ τότε η τελική τάξη ακρίβειας είναι p. Αν $\tilde{p}=p-2$ και $m=1$ η τελική τάξη είναι p-1. Αν $\tilde{p}=p-2$ και $m>1$ τότε η τελική τάξη είναι p. Π.χ. η μέθοδος Euler-τραπεζίου με $\tilde{p}=1$, $p=2$ έχει τελική τάξη ακρίβειας 2 γιά οποιοδήποτε αριθμό διορθώσεων $m\geq 1$. Γιά την μέθοδο του Milne $\tilde{p}=p=4$, οπότε και η τελική τάξη είναι 4. Η μέθοδος (Δ) εξ άλλου καθορίζει την ευστάθεια του ζεύγους το οποίο είναι ευσταθές αν και μόνο αν η (Δ) είναι ευσταθής. Μ' άλλα λόγια δεν χρειάζεται και η (Π) να είναι ευσταθής.

4. Όπως γίνεται γιά τις μεθόδους RK, είναι δυνατόν να μεταβάλουμε το βήμα $h_j=t^{j+1}-t^j$ και γιά τις πολυβηματικές μεθόδους. Αυτό αυξάνει την λογιστική των μεθόδων γιατί θα πρέπει (με παρεμβολή π.χ.) κάθε φορά που αλλάζει το βήμα να βρίσκονται και προσεγγίσεις σε διαφορετικές προηγούμενες τιμές του t. Στρατηγικές γιά τον έλεγχο του μεγέθους του h_j στηρίζονται και πάλι σε κάποιου είδους εκτίμηση του τοπικού σφάλματος, η οποία είναι αρκετά απλή αν χρησιμοποιούμε ζεύγη μεθόδων πρόβλεψης-διόρθωσης. Καλά προγράμματα που στηρίζονται σε μεθόδους πρόβλεψης-διόρθωσης εκτός από μεταβλητό βήμα και εκτίμηση του τοπικού σφάλματος έχουν την δυνατότητα να <u>μεταβάλλουν και την τάξη ακρίβειάς τους</u>, δηλ. να καταφεύγουν σε διαφορετικά ζεύγη Π-Δ, μεγαλύτερης (μικρότερης) συνολικής τάξης ακρίβειας, αν η λύση μεταβάλλεται - με δείκτη το μέγεθος του τοπικού σφάλματος - πολύ (λίγο) από βήμα σε βήμα. Χαρακτηριστικά τέτοια προγράμματα είναι το πρόγραμμα DIFSUB του Gear, βλ. [3.2], το πρόγραμμα DVDQ του Krogh κ.ά, που υπάρχουν σε πολλές βιβλιοθήκες αλγορίθμων.

Ασκήσεις 3.3

1. Αν $\sigma(z)=z^2$, βρήτε $\rho(z)$ (με α_k πιθανώς όχι 1) τέτοιο ώστε:

(α). Το $p(z)$ να είναι δευτέρου βαθμού και η τάξη της μεθόδου (p,σ) να είναι 2.

(β) Το $p(z)$ να είναι τρίτου βαθμού και η τάξη της μεθόδου (p,σ) να είναι 3.

~~(γ) Είναι ευσταθείς οι δύο μέθοδοι;~~

2. Αν $p(z)=z^4-1$, βρήτε $\sigma(z)$ βαθμού 4 τέτοιο ώστε η μέθοδος (p,σ) να έχει μέγιστη τάξη. Ποιά είναι η τάξη της και ποιά η σταθερά του σφάλματος;

3. Βρήτε τους συντελεστές $\{\alpha_j,\beta_j\}$ $0\le j\le 2$, $\alpha_2=1$ της γενικής διβηματικής μεθόδου έτσι ώστε η τάξη ακρίβειάς της να είναι $p\ge 2,3$ ή 4. Υπάρχει μέθοδος με $p=4$; Με $p\ge 5$;

4. Βρήτε τους συντελεστές α_j,β_j της τριβηματικής μεθόδου μέγιστης τάξης ακρίβειας. Ποιά είναι η τάξη της; Είναι ευσταθής;

5. Εκφράστε τους συντελεστές της οικογένειας των διβηματικών μεθόδων τάξης ≥ 3 συναρτήσει της παραμέτρου β_0. Γιά ποιές τιμές του β_0 είναι ευσταθείς οι μέθοδοι; Εκφράστε την σταθερά σφάλματος ως συνάρτηση του β_0 (γιά τιμές του β_0 που δίνουν ευσταθείς μεθόδους). Τι παρατηρείτε γιά την μέθοδο του Simpson. ;

6. Βρήτε τις ευσταθείς 4-βηματικές μεθόδους με τάξη ακρίβειας $p=6$.

7. Αποδείξτε ότι οι μέθοδοι οπισθοδρομικών διαφορών με k βήματα (6) έχουν τάξη ακρίβειας k.

8. Αποδείξτε, χρησιμοποιώντας την θεωρία Schur (Παρατήρηση 1) ότι οι k-βηματικές μέθοδοι οπισθοδρομικών διαφορών (6) είναι ευσταθείς γιά $k=1,2,3,4$. (Είναι γνωστό ότι οι μέθοδοι αυτές είναι ευσταθείς αν και μόνο αν $1\le k\le 6$).

9. Σε αναλογία με ό,τι κάναμε στις μεθόδους RK ορίστε την τάξη ακρίβειας και την ευστάθεια της μεθόδου (Π)-(Δ) Euler - τραπεζίου (Βλ. Παρατήρηση 3(i)) και αποδείξτε ότι η τάξη ακρίβειας είναι p=2 και ότι η μέθοδος είναι ευσταθής.

10. Γιά δευτεροβάθμιες Δ.Ε. της μορφής $y''=f(t,y)$ - γιά τις οποίες δηλ. η f δεν είναι συνάρτηση του y' - είναι δυνατόν να αναπτύξουμε ανάλογη θεωρία k-βηματικών μεθόδων της μορφής $\sum_{j=0}^{k} \alpha_j y^{n+j} = h^2 \sum_{j=0}^{k} \beta_j f^{n+j}$. Η συνθήκη ευστάθειας τώρα στο πολυώνυμο $p(z)=\sum_{j=0}^{k} \alpha_j z^j$ είναι ότι όλες οι ρίζες του z_i πρέπει να ικανοποιούν $|z_i|\leq 1$ και ότι η πολλαπλότητα των ριζών με $|z_i|=1$ πρέπει να είναι το πολύ 2. Αν ισχύουν αυτές οι συνθήκες, δείξτε ότι στό ανάπτυγμα

$$[1/(\alpha_k+\alpha_{k-1}z+\ldots+\alpha_0 z^k)] = \gamma_0+\gamma_1 z+\gamma_2 z^2+ \ldots$$

οι συντελεστές γ_i αυξάνουν "γραμμικά", δηλ. ότι υπάρχουν σταθερές c_1 και c_2 τέτοιες ώστε

$$|\gamma_i| \leq c_1 i + c_2, \quad i=0,1,2,\ldots$$

3.4 ΑΚΑΜΠΤΑ ΠΡΟΒΛΗΜΑΤΑ. ΑΠΟΛΥΤΗ ΕΥΣΤΑΘΕΙΑ ΚΑΙ ΓΕΝΙΚΕΥΣΕΙΣ ΤΗΣ

Θεωρούμε το πρόβλημα αρχικών τιμών γιά την απλή Δ.Ε

$$(1)\begin{cases} y' = \lambda y, & t \geq 0, \\ y(0) = 1, \end{cases}$$

και υποθέτουμε ότι η σταθερά λ είναι τέτοια ώστε $\lambda << 0$, δηλ. $\lambda < 0$ με $|\lambda| >> 1$. Προφανώς η λύση $y(t) = e^{\lambda t}$ του (1) τείνει πολύ γρήγορα στό μηδέν καθώς αυξάνεται το t. Ας προσεγγίσουμε το πρόβλημα (1) με την μέθοδο του Euler με κάποιο σταθερό βήμα h. Προφανώς προκύπτει η ακολουθία $\{y^n\}$, $n \geq 0$ όπου $y^{n+1} = (1+h\lambda)y^n$, $n \geq 0$, $y^0 = 1$, δηλ. η ακολουθία

$$(2)\ y^n = (1+h\lambda)^n,\ n \geq 0.$$

'Εστω ότι σε κάποιο βήμα k του αλγορίθμου υπολογίζουμε, π.χ. λόγω σφαλμάτων στρογγύλευσης, τον αριθμό z^k αντί του y^k και ότι γιά $n > k$ υπολογίζουμε τα z^n ακριβώς. Θεωρούμε δηλ. την ακολουθία $\{z^n\}$, $n \geq 0$:

$$z^n = y^n,\ 0 \leq n \leq k-1,\quad z^{n+1} = (1+h\lambda)z^n,\ n \geq k,\ z^k \neq y^k.$$

Συνεπώς έχουμε

$$y^n - z^n = (1+h\lambda)^{n-k}\,(y^k - z^k),\ n \geq k,$$

δηλ. ότι

$$(3)\ |y^n - z^n| = |1+h\lambda|^{n-k}\,|y^k - z^k|,\ n \geq k,$$

από την οποία βλέπουμε ότι αν $|1+h\lambda| < 1$ (δηλ. αν $0 < h < -2/\lambda$) η διαταραχή $|y^k - z^k|$ στο βήμα k δεν θα έχει σοβαρές επιπτώσεις. Πράγματι, τότε, $|1+h\lambda|^{n-k} \to 0$, $n \to \infty$, δηλ. η προκαλούμενη μεταβολή $|y^n - z^n|$ (θεωρούμενη ως διαταραχή της "ακριβούς" λύσης y^n του διακριτού προβλήματος) θα γίνεται όλο και πιό μικρή καθώς αυξάνεται το n. Αν $|1+h\lambda|=1$ το σφάλμα $|y^k-z^k|$ διατηρείται ($|y^n-z^n| = |y^k-z^k|$, $n\geq k$), ενώ αν $|1+h\lambda|>1$, θα έχει καταστροφικές επιπτώσεις γιατί τότε $|y^n-z^n| \to \infty$ εκθετικά με το n. Ας σημειωθεί ότι αν π.χ. υπολογίζουμε στο διάστημα $0\leq t\leq T$, έχουμε, γιά $h=T/N$, ότι όντως ισχύει η εκτίμηση

$$\max_{0\leq n\leq N} |y^n-z^n| \leq |1+h\lambda|^{N-k} |y^k-z^k| \leq (1+h|\lambda|)^N |y^k-z^k|$$
$$\leq e^{|\lambda|T} |y^k-z^k|,$$

(όπως προβλέπεται από την Πρόταση 3.2) που εκφράζει στην απλή μας περίπτωση την "ευστάθεια" της μεθόδου του Euler ως μεθόδου RK, με την έννοια της "ευστάθειας" που εισαγάγαμε στην παράγραφο 3.2. Όμως η σταθερά Lipschitz $L=|\lambda|$ του προβλήματός μας είναι πολύ μεγάλη έτσι ώστε το φράγμα $e^{|\lambda|T}$ να μην έχει πρακτική αξία στην περίπτωσή μας. Το παράδειγμα αυτό μας δείχνει όμως ότι μιά νέα συνθήκη, δηλ. η $|1+h\lambda|<1$, εξασφαλίζει μιά πραγματική ευστάθεια της μεθόδου με την έννοια ότι καταστέλλει τυχόν "σφάλματα" $|y^k-z^k|$ και δεν τους επιτρέπει να επηρεάσουν σημαντικά τους υπολογισμούς γιά $n>k$. <u>Η συνθήκη αυτή αποτελεί ένα σοβαρό περιορισμό στο μέγεθος του βήματος h</u>.

Σημειώστε ότι η συνθήκη $|1+h\lambda|<1$ εξασφαλίζει επίσης ότι η ακολουθία $\{y^n\}$ που παράγει η μέθοδος του Euler έχει την ιδιότητα (βλ. (2)) ότι $y^n \to 0$, $n \to \infty$ (Τονίζουμε ότι παίρνουμε το όριο γιά σταθερό h όταν $n \to \infty$). Δηλ. <u>είναι ικανή και αναγκαία συνθήκη έτσι ώστε η λύση $\{y^n\}$ του διακριτού προβλήματος (γιά κάθε σταθερό h) να μιμείται την συμπεριφορά της λύσης του συνεχούς προβλήματος (1), $y(t)=e^{\lambda t}$, γιά την οποία προφανώς $y(t) \to 0$, $t \to \infty$</u>.

Θεωρείστε τώρα την ακολουθία $\{y^n\}$ που παράγει η <u>πεπλεγμένη</u> μέθοδος του Euler γιά το ίδιο πρόβλημα, δηλ. η μέθοδος $y^{n+1}-y^n=h\lambda y^{n+1}$, $n\geq 0$, $y^0=1$, που γράφεται ως

(4) $y^n = (1-h\lambda)^{-n}$, $n\geq 0$.

Γιά την (4) το ανάλογο της σχέσης (3) είναι

$$(5)\quad |y^n-z^n| = |1-h\lambda|^{-(n-k)}\,|y^k-z^k|,\ n\geq k.$$

Επειδή $\lambda<0$ συμπεραίνουμε ότι $|1-h\lambda|^{-1}<1$ $\forall h>0$, δηλ. ότι η (5) γιά κάθε $h>0$ δίνει $|y^n-z^n|<<|y^k-z^k|$, αν $n>>k$: η πεπλεγμένη μέθοδος του Euler δεν επηρεάζεται σχεδόν καθόλου από σφάλματα $|y^k-z^k|\neq 0$. Ισοδύναμα, $\forall h>0$, η ακολουθία (4) ικανοποιεί $y^n\to 0$, $n\to\infty$ δηλ. το διακριτό ανάλογο της ιδιότητας $y(t)\to 0, t\to\infty$ της λύσης του συνεχούς προβλήματος (1).

Ας θεωρήσουμε τώρα το πρόβλημα

$$(6)\begin{cases} y' = \lambda y + f'(t)-\lambda f(t),\ t\geq 0,\\ y(0) = y_0, \end{cases}$$

του οποίου η λύση είναι προφανώς

$$(7)\quad y(t) = f(t) + e^{\lambda t}(y_0-f(0)),\ t\geq 0.$$

Υποθέτουμε ξανά ότι $\lambda<<0$ και ότι η $f(t)$ είναι μία δεδομένη ομαλή συνάρτηση που δεν μεταβάλλεται γρήγορα με το t. Από τον τύπο (7) βλέπουμε ότι η λύση $y(t)$ εκφράζεται ως άθροισμα του όρου $f(t)$ και του <u>εφήμερου</u> όρου $e^{\lambda t}(y_0-f(0))$ που τείνει στο 0 πολύ γρήγορα (δηλ. που εξαφανίζεται σχεδόν αμέσως και δεν συμβάλλει στην λύση) καθώς αυξάνει το t. Προφανώς ο σημαντικός όρος γιά την λύση γιά $t>0$ είναι ο $f(t)$.

Γράφοντας την (7) γιά $t+h$ αντί t και απαλείφοντας τον αρχικό όρο $y_0-f(0)$ παίρνουμε την ταυτότητα

$$(8)\quad y(t+h) = f(t+h) + e^{\lambda h}(y(t)-f(t)),\ t,h\geq 0.$$

Η μέθοδος του Euler γιά το βήμα $t^n\to t^{n+1}$ γράφεται ως $y^{n+1}=y^n+h(\lambda y^n+f'(t^n)-\lambda f(t^n))$, δηλ. ως

$$(9)\quad y^{n+1} = f(t^n)+hf'(t^n) + (1+h\lambda)(y^n-f(t^n)).$$

Παρατηρούμε ότι $e^{\lambda h}=1+h\lambda+O(h^2\lambda^2)$ και ότι $f(t^n+h)=f(t^n)+hf'(t^n)+O(h^2)$, δηλ. ότι όντως η (9) είναι προσέγγιση της (8) γιά $t=t^n$. Ενώ όμως στην (8) η διαφορά $(y(t)-f(t))$ - που μπορεί να θεωρηθεί ως "διαταραχή" στην τιμή της "θεμελιώδους" λύσης στο $t+h$, $y(t+h)\cong f(t+h)$ - πολλαπλασιάζεται επί $e^{\lambda h}<<1$ (γιά $\lambda h<<0$) και δεν διαταράσσει συνεπώς σχεδόν καθόλου την σχέση $y(t+h)\cong f(t+h)$, στην (9), αν $|1+h\lambda|\geq 1$, η διαφορά $y^n-f(t^n)$ επηρεάζει σημαντικά την σχέση $y^{n+1}\cong f(t^n)+hf'(t^n)$ δηλ. την $y^{n+1}\cong f(t^n+h)$. Βλέπουμε δηλ. πάλι τον σημαντικό ρόλο της συνθήκης $|1+h\lambda|<1$ γιά την μέθοδο του Euler γιά αυτό το πρόβλημα. Η <u>πεπλεγμένη</u> μέθοδος του Euler, όπως μπορούμε εύκολα να δούμε, δίνει την σχέση

$$y^{n+1}=(1-h\lambda)^{-1}[f(t^n)+hf'(t^{n+1})-\lambda h f(t^{n+1})]+(1-h\lambda)^{-1}(y^n-f(t^n)),$$

που μας δείχνει - παρατηρείστε ότι ο πρώτος όρος είναι μία $O(h^2)$ προσέγγιση της $f(t^{n+1})$ - ότι γιά κάθε $h>0$ η "διαταραχή" $y^n-f(t^n)$ εμποδίζεται να διαταράξει την σχέση $y^{n+1}\cong f(t^{n+1})$ επειδή πολλαπλασιάζεται επί του παράγοντα $(1-h\lambda)^{-1}<<1$.

Το πρόβλημα (6) αποτελεί τυπικό δείγμα προβλήματος αρχικών τιμών γιά μιά <u>άκαμπτη</u> Δ.Ε.. Τα χαρακτηριστικά των ακάμπτων προβλημάτων διαφαίνονται ήδη από τύπους όπως οι (7) και (8): Η λύση τους είναι βασικά ομαλή ($y(t)\cong f(t)$) και μεταβάλλεται αργά με τον χρόνο γιά $t>\varepsilon>0$, ε "μικρό", αλλά περιέχει μία συνιστώσα η οποία αποσβένεται εκθετικά καθώς αυξάνει το t, δηλ. που πρακτικά παύει να συνεισφέρει στη λύση γιά $t>\varepsilon$, βλ. (7). Απ' την άλλη μεριά η (8) μας λέει ότι γιά κάθε t (π.χ. μεγάλο) η λύση $y(t+h)$ είναι περίπου ίση με την τιμή $f(t+h)$ του θεμελιώδους "φορέα" της· η σχέση όμως $y(t+h)\cong f(t+h)$ διαταράσσεται από μία μικρή μεταβολή (στο πρόβλημά μας την $e^{h\lambda}(y(t)-f(t))$) που προκύπτει από το ότι $y(t)\neq f(t)$, αλλά που αποσβένεται από τον παράγοντα $e^{\lambda h}$. (Ουσιαστικά ο (8) είναι ο ίδιος τύπος με τον (7) αλλά με αρχικές συνθήκες στο σημείο t). Αν λοιπόν η αριθμητική μέθοδος προσεγγίζει το εκθετικό $e^{\lambda h}$ με όχι καλό τρόπο γιά $h\lambda<<0$, όπως π.χ. συμβαίνει με την μέθοδο του Euler (γιά την οποία η προσέγγιση $1+h\lambda$ είναι μεν αποδεκτή γιά $|h\lambda|<<1$, αλλά ανακριβής και ποιοτικά τελείως λανθασμένη γιά $h\lambda<<0$ οπότε $|1+h\lambda|>>1$), τότε περιμένουμε σοβαρή διαταραχή της λύσης y^n (αστάθεια) που θα προέλθει στο παράδειγμά μας από τον όρο $(1+\lambda h)(y^n-f(t^n))$. Πρέπει δηλ. να επιβάλουμε στο h την περιοριστική συνθήκη $|1+h\lambda|<1 \iff 0<h<-2/\lambda$, η οποία πρέπει να ισχύει σ' όλη την διάρκεια των υπολογισμών παρά το γεγονός ότι ο παράγοντας $e^{\lambda t}(y^0-f(0))$ της λύσης που προκαλεί αυτόν τον περιορισμό στο h δεν συνεισφέρει πρακτικά τίποτε στη λύση γιά $t>\varepsilon>0$! Στις σημειώσεις [5.3], σελ. 163 et seq. γίνεται η εφαρμογή της μεθόδου του Euler στο παράδειγμα (6) με $\lambda=-1000$, $f(t)=t^2$, $y(0)=0$. Παρατηρούμε ότι $y_0-f(0)=0$, δηλ. ότι ο παράγοντας $e^{\lambda t}(y_0-f(0))$ <u>δεν υπάρχει καθόλου</u> στην λύση $y(t)=t^2$· ο τύπος (9) εξακολουθεί βέβαια να ισχύει. Αναπόφευκτα το αριθμητικό πείραμα δείχνει καταστροφικά γρήγορα αυξανόμενη αστάθεια αν $h>.002$ και καλά αποτελέσματα - με $O(h^2)$ βέβαια σφάλμα - γιά $h<.002$ όπως προβλέπει η συνθήκη $|1-1000h|<1$. Αντίθετα η πεπλεγμένη μέθοδος του Euler προσεγγίζει το εκθετικό $e^{\lambda h}$ με την ποιοτικά σωστή γιά $\lambda h<<0$ ρητή προσέγγιση $(1-\lambda h)^{-1}$ (Πάλι τάξης $O(h^2\lambda^2)$ κοντά στο 0 αλλά που μιμείται την ιδιότητα $e^{\lambda h}<<1$ γιά $\lambda h<<0$). Συνεπώς γιά κάθε $h>0$ η πεπλεγμένη μέθοδος δεν θα έχει φαινόμενα αστάθειας προερχόμενα από τον όρο $(1-\lambda h)^{-1}(y^n-f(t^n))$.

Μετά από τις εισαγωγικές αυτές παρατηρήσεις προχωρούμε σ' ένα ορισμό, που οφείλεται (όπως και μεγάλο μέρος της αρχικής έρευνας γιά αυτά τα θέματα) στον Dahlquist (1963). Λέμε ότι μιά μέθοδος (γιά κάποιο ορισμένο βήμα h) είναι <u>απόλυτα ευσταθής</u> (γιά απλές Δ.Ε) αν, όταν εφαρμοσθεί στο πρόβλημα (1) με $\lambda<0$, έχει την ιδιότητα να δίνει προσεγγίσεις y^n, $n\geq 0$ τέτοιες ώστε $y^n\to 0$, $n\to\infty$ (γιά σταθερό h). Οι τιμές του γινομένου λh γιά τις οποίες η μέθοδος έχει αυτή την ιδιότητα αποτελούν την <u>περιοχή απόλυτης ευστάθειας της μεθόδου</u> που είναι, συνεπώς, υποσύνολο (γιά απλές Δ.Ε.) του αρνητικού πραγματικού ημιάξονα. Συνήθως ενδιαφερόμαστε να εντοπίσουμε (αν υπάρχει) κάποιο <u>διάστημα απόλυτης ευστάθειας</u> με μέγιστο μήκος αλλά της μορφής $(r,0)$, $-\infty\leq r<0$. Τότε, αν $h\lambda\in(r,0)$, δηλ. αν επιλέξουμε h τέτοιο ώστε $0<h<r/\lambda$, η μέθοδος θα είναι <u>απόλυτα ευσταθής</u>, εξ ορισμού. Παραδείγματος χάριν είδαμε ότι η μέθοδος του Euler έχει διάστημα απόλυτης ευστάθειας το (-2,0) και η πεπλεγμένη μέθοδος του Euler το $(-\infty,0)$, δηλ. είναι απόλυτα ευσταθής γιά κάθε h. Δεν είναι δύσκολο να δούμε ότι η μέθοδος του τραπεζίου (3.2.5) έχει διάστημα απόλυτης ευστάθειας επίσης ίσο με $(-\infty,0)$ (δηλ. είναι και αυτή ιδιαίτερα κατάλληλη γιά προβλήματα όπως το (1) με $\lambda<<0$). Αντίθετα η μέθοδος του μέσου (άμεση) (3.2.6) έχει (-2,0), οι άμεσες μέθοδοι RK τρίτης τάξης (3.20 α,β) έχουν (-2.51,0), οι κλασικές μέθοδοι RK (3.21a,b) έχουν (-2.78,0) κ.ο.κ. Θα μελετήσουμε πιό συστηματικά την απόλυτη ευστάθεια των μεθόδων RK καί των πολυβηματικών μεθόδων παρακάτω.

Μία παρατήρηση: εξετάζουμε το πρόβλημα $y'=\lambda y$, $\lambda<0$, γιατί το αντίστοιχο με $\lambda>0$ έχει εκθετικά αυξανόμενες λύσεις $e^{\lambda t}$ γιά τις οποίες γιά λ ή $t>>0$ είναι δύσκολο να διακρίνει κανείς την λύση από την ασταθή διαταραχή της. Γεννάται επίσης το εύλογο ερώτημα κατά πόσο η συμπεριφορά μιάς αριθμητικής μεθόδου στην περίπτωση του απλού παραδείγματος (1) μπορεί να μας οδηγήσει σε συμπεράσματα γιά την συμπεριφορά της σε πιό πολύπλοκα, μη γραμμικά προβλήματα. Θα μελετήσουμε λοιπόν αργότερα, επεκτάσεις της έννοιας της απόλυτης ευστάθειας γιά(κατάλληλα) μη γραμμικά προβλήματα. Πρός το παρόν ας θεωρήσουμε την ευσταθή συμπεριφορά της σ' ένα απλό πρόβλημα όπως το (1) ως <u>αναγκαία συνθήκη</u> γιά ευσταθή συμπεριφορά σε πιό πολύπλοκα προβλήματα. Δεν είναι όμως σαφές ότι είναι και <u>ικανή</u> συνθήκη αν καί κανείς θα μπορούσε εύκολα να διατυπώσει την εικασία ότι ευσταθής συμπεριφορά γιά το (1) συνεπάγεται ευσταθή συμπεριφορά γιά προβλήματα με μεγάλες (απολύτως) σταθερές Lipschitz, που όμως έχουν φθίνουσες λύσεις, όπως π.χ. $y'=\lambda(t)y$ με $\lambda(t)<0$, $\max_t|\lambda(t)|>>1$ ή $y'=f(y)$ με $f'(y)<0$ αλλά με $\max_y|f'(y)|>>1$ κ.ο.κ. Αυτό είναι γιά ορισμένες μεθόδους σωστό αλλά, όπως θα δούμε, υπάρχουν απόλυτα ευσταθείς μέθοδοι που δεν είναι ευσταθείς γιά ανάλογα μη γραμμικά προβλήματα ή ακόμα καί γιά προβλήματα με μεταβλητούς συντελεστές $\lambda(t)$.

Μία σχετική απλή επέκταση όμως του ορισμού της απόλυτης ευστάθειας αρκεί γιά την ρεαλιστική μελέτη της ευστάθειας μιάς μεθόδου όταν εφαρμοσθεί σε <u>γραμμικά συστήματα Δ.Ε. με σταθερούς συντελεστές</u>, δηλ. σε συστήματα της μορφής $y'=Ay+F(t)$, όπου $A\in R^{m\times m}$(a_{ij} σταθερές), $y\in R^m$. Θεωρούμε το εξής πρόβλημα αρχικών τιμών γιά το <u>ομογενές</u> γραμμικό σύστημα

$$(10)\quad \begin{cases} y' = Ay, & t\geq 0, \\ y(0) = y_o \end{cases}$$

και υποθέτουμε γιά απλούστευση ότι ο A διαγωνοποιείται, δηλ. ότι υπάρχει αντιστρέψιμος πίνακας S τέτοιος ώστε $S^{-1}AS=\Lambda=diag(\lambda_1,\ldots,\lambda_m)$, όπου λ_i, $1\leq i\leq m$ οι (εν γένει μιγαδικές) ιδιοτιμές του A. Η αλλαγή των μεταβλητών $y=S\psi$ δίνει το ισοδύναμο σύστημα

$$(10')\quad \begin{cases} \psi' = \Lambda\psi, & t\geq 0 \\ \psi(0) = \psi_o = S^{-1}y_o, \end{cases}$$

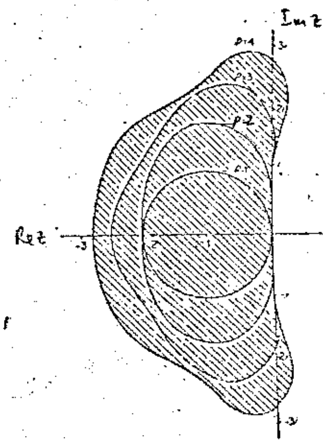
του οποίου η λύση είναι προφανώς $\psi_i=(\psi_o)_i e^{\lambda_i t}$. Συμπεραίνουμε π.χ. ότι η λύση $y(t)$ του (10) έχει την ιδιότητα $y(t)\to 0$, $t\to\infty$ γιά κάθε $y_o\in R^m$ αν και μόνο αν $Re\lambda_i<0$, $1\leq i\leq m$. (Το ίδιο ισχύει, όπως εύκολα μπορούμε να δούμε χρησιμοποιώντας την μορφή Jordan του A, και στην γενική περίπτωση ενός οποιουδήποτε, όχι αναγκαστικά διαγωνοποιήσιμου, πίνακα $A\in R^{m\times m}$). Επειδή στο σύστημα (10') οι Δ.Ε. έχουν αποσυνδεθεί σε m ανεξάρτητες απλές Δ.Ε. της μορφής $\psi_i'=\lambda_i\psi_i$, φαίνεται λογικό ότι γιά να επεκτείνουμε την έννοια της απόλυτης ευστάθειας έτσι ώστε να έχει εφαρμογή σε συστήματα της μορφής (10), θα πρέπει να μελετήσουμε την συμπεριφορά των λύσεων της αριθμητικής μεθόδου όταν εφαρμοσθεί σε απλή Δ.Ε. της μορφής (1) αλλά <u>με μιγαδικό λ με $Re\lambda<0$.</u>

Λέμε λοιπόν γενικά, σύμφωνα με τον Dahlquist (1963), ότι μία αριθμητική μέθοδος γιά την λύση προβλημάτων αρχικών τιμών γιά συστήματα Δ.Ε. είναι <u>απόλυτα ευσταθής</u> (γιά κάποιο ορισμένο βήμα h) αν, όταν εφαρμοσθεί στο πρόβλημα

$$(11)\quad \begin{cases} y' = \lambda y, & t>0,\ \lambda\in C,\ Re\lambda<0 \\ y(0) = 1. \end{cases}$$

δίνει προσεγγίσεις y^n, $n\geq 0$, τέτοιες ώστε $y^n\to 0$, $n\to\infty$(σταθερό h). Οι τιμές του γινομένου $h\lambda$ (υποσύνολο του αριστερού μιγαδικού ανοιχτού ημιεπιπέδου $\mathrm{Re}z<0$) γιά τις οποίες μία μέθοδος έχει αυτήν την ιδιότητα αποτελούν την <u>περιοχή απόλυτης ευστάθειας της μεθόδου</u>. Προφανώς, ο νέος ορισμός της απόλυτης ευστάθειας γενικεύει τον προηγούμενο, που υπέθετε ότι $\lambda\in\mathbb{R}$, $\lambda<0$. Στο εξής θα χρησιμοποιούμε τον νέο ορισμό.

Η μέθοδος του Euler εφαρμοζόμενη στο πρόβλημα (11) δίνει την (μιγαδική) ακολουθία $y^n=(1+h\lambda)^n$ γιά την οποία $y^n\to 0$, $n\to\infty$ $\Leftrightarrow$ $|1+h\lambda|<1$. Θέτοντας $z=h\lambda$ βλέπουμε ότι η περιοχή απόλυτης ευστάθειας της μεθόδου του Euler είναι ο ανοιχτός δίσκος $\{z:|1+z|<1\}$ κέντρου -1 και ακτίνας 1 στο μιγαδικό επίπεδο. Προφανώς το <u>διάστημα</u> (-2,0) απόλυτης ευστάθειας της μεθόδου (γιά απλές Δ.Ε.) είναι η τομή του $|1+z|<1$ με τον πραγματικό άξονα. Η (άμεση) μέθοδος του μέσου (3.2.6) εφαρμοζόμενη στο πρόβλημα (11) δίνει $y^n = (1+h\lambda+(h^2\lambda^2)/2)^n$, δηλ. ότι $y^n\to 0$, $n\to\infty$ $\Leftrightarrow$ $|1+z+z^2/2|<1$, $z=h\lambda$.



Σχήμα 3.4.1

Το χωρίο $\{z\in\mathbb{C}:|1+z+z^2/2|<1\}$ είναι συμμετρικό ως πρός τον πραγματικό άξονα (γιά κάθε πολυώνυμο με πραγματικούς συντελεστές $\overline{\pi(z)}=\pi(\bar{z})$), περιέχεται όλο στο ανοιχτό ημιεπίπεδο $\mathrm{Re}z<0$ και περιέχει τον δίσκο $|1+z|<1$ της μεθόδου του Euler. Το σύνορό του, δηλ. η καμπύλη $1+z+z^2/2=\exp(i\theta)$, μπορεί να σχεδιασθεί εύκολα αν δίνουμε τιμές στο θ και λύνουμε κάθε φορά την εξίσωση ως πρός z. Η καμπύλη (p=2 στο σχήμα 1) που προκύπτει, εφάπτεται του φανταστικού άξονα στο 0 και της περιφέρειας $|1+z|=1$ στο (-2,0).

Η πεπλεγμένη μέθοδος του Euler δίνει την ακολουθία $y^n=(1-h\lambda)^{-n}$, $n\geq 0$, δηλ. ικανοποιεί $y^n \to 0$, $n \to \infty$ αν και μόνο αν

$$|(1-z)^{-1}| < 1, \ z=h\lambda,$$

που ισχύει γιά κάθε z με $\mathrm{Re}z<0$. Συνεπώς η περιοχή απόλυτης ευστάθειας της μεθόδου είναι όλο το ημιεπίπεδο $\mathrm{Re}z<0$. Τέτοιες μέθοδοι λέγονται <u>A-ευσταθείς</u>, (Dahlquist, 1963). Η μέθοδος του τραπεζίου, η οποία δίνει $(1-h\lambda/2)y^{n+1} = (1+h\lambda/2)y^n$, δηλ.

$$y^n = [(1+h\lambda/2)(1-h\lambda/2)^{-1}]^n$$

είναι απόλυτα ευσταθής γιά $|(1+z/2)/(1-z/2)|<1$, $z=h\lambda$, δηλ. γιά κάθε z με $\mathrm{Re}z<0$. Συμπεραίνουμε ότι και η μέθοδος του τραπεζίου είναι A-ευσταθής

Λέμε ότι ένα πρόβλημα της μορφής

$$(12) \qquad \begin{cases} y' = Ay + F(t), \ t\geq 0 \\ y(0)=y_0 \end{cases}$$

όπου οι ιδιοτιμές λ_i του A ικανοποιούν $\mathrm{Re}\lambda_i<0$ $\forall i$, είναι <u>άκαμπτο</u> αν $\max_i|\mathrm{Re}\lambda_i| \gg \min_i|\mathrm{Re}\lambda_i|$. Δεν είναι δύσκολο να δούμε ότι ένα τέτοιο σύστημα έχει τα χαρακτηριστικά του άκαμπτου προβλήματος γιά απλές Δ.Ε. που εξετάσαμε προηγουμένως. Υπάρχουν συνιστώσες της λύσης (αυτές που αντιστοιχούν σε ιδιοτιμές με μεγάλο $|\mathrm{Re}\lambda_i|$) που τείνουν στο μηδέν πολύ γρήγορα καθώς αυξάνει το t, σε σχέση με άλλες συνιστώσες που τείνουν πιό αργά στο 0 ή μεταβάλλονται σε χρονικές κλίμακες που εξαρτώνται από το $F(t)$. Γιά να είναι μία μέθοδος "ρεαλιστικά ευσταθής" γιά το πρόβλημα (12) είναι προφανές ότι πρέπει να έχει μία κατάλληλη μη κενή περιοχή απόλυτης ευστάθειας S στο ημιεπίπεδο $\mathrm{Re}z<0$ και το βήμα h να είναι τέτοιο ώστε $h\lambda_i \in S$, $\forall i$. Αν η περιοχή S είναι "μικρή" (όπως π.χ. συμβαίνει με την μέθοδο του Euler ή την μέθοδο του μέσου), έπεται ότι πρέπει να πάρουμε το h <u>πολύ μικρό</u> έτσι ώστε γιά <u>όλες</u> τις ιδιοτιμές λ_i (ακόμα και εκείνες γιά τις οποίες $|Re\lambda_i|>>1$ και που δεν συμβάλλουν καθόλου σχεδόν στην λύση!) να έχουμε $h\lambda_i\in S$. Άρα οι Α-ευσταθείς μέθοδοι είναι ιδιαίτερα κατάλληλες γιά άκαμπτα προβλήματα που έχουν ιδιοτιμές λ_i οπουδήποτε στο $Rez<0$: γιά τις Α-ευσταθείς μεθόδους έχουμε $S=\{z:Rez<0\}$ και συνεπώς $h\lambda_i\in S$ γιά κάθε $h>0$.

Πολλά άκαμπτα συστήματα προκύπτουν από την αριθμητική επίλυση μερικών διαφορικών εξισώσεων που η λύση τους εξαρτάται από τον χρόνο, όπως π.χ. παραβολικών και υπερβολικών εξισώσεων. Θεωρούμε π.χ. το εξής απλό πρόβλημα αρχικών και συνοριακών τιμών γιά την εξίσωση της θερμότητας: ζητάμε $u(x,t)$ πραγματική συνάρτηση ορισμένη γιά $x\in[0,1]$, $t\in[0,T]$ τέτοια ώστε

$$(13)\quad \begin{cases} u_t = u_{xx}, \ (x,t)\in[0,1]\times[0,T] \\ u(x,0)= u(x), \ x\in[0,1] \\ u(0,t) = u(1,t)=0, \ t\in[0,T] \end{cases}$$

όπου $u(x)$ δεδομένη πραγματική συνάρτηση στο $[0,1]$ με $u(0)=u(1)=0$. Έστω $x_j=j\Delta x$, $j=0,1,2,\ldots,(J+1)$ ένας ομοιόμορφος διαμερισμός του $[0,1]$ με $(J+1)\Delta x=x_{J+1}=1$. Αντικαθιστώντας στην Μ.Δ.Ε. του (13) την παράγωγο $u_{xx}(x,t)$ με την κευτρική διαφορά $(u(x+\Delta x,t) - 2u(x,t) + u(x-\Delta x,t))/(\Delta x)^2$ παίρνουμε το εξής σύστημα Σ.Δ.Ε. γιά τις προσεγγίσεις $u_j(t)$, των τιμών $u(x_j,t)$ της λύσης του (13):

$$(14)\quad \begin{cases} u_j'(t) = (u_{j+1}(t)-2u_j(t)+u_{j-1}(t))/(\Delta x)^2, \ 1\le j\le J, \ t\in[0,T] \\ u_0(t) = u_{J+1}(t) = 0 \ , \ t\in[0,T] \\ u_j(0) = u_j \equiv u(x_j), \ 0\le j\le J+1 \end{cases}$$

Το $J\times J$ σύστημα των Δ.Ε. της (14) γράφεται γιά $u=[u_1,\ldots,u_J]^T$

$$(15)\quad u' = Au$$

όπου A ο συμμετρικός τριδιαγώνιος πίνακας με $a_{ii} = -2/(\Delta x)^2$,

$\alpha_{i,i\pm 1}=(\Delta x)^{-2}$. Δεν είναι δύσκολο να δούμε ότι οι ιδιοτιμές του A είναι όλες αρνητικοί αριθμοί (γιατί ο $-A$ είναι θετικά ορισμένος) και δίνονται από τους τύπους

$$\lambda_j=(\Delta x)^{-2}\,[-2+2\cos(j\pi/(J+1))],\quad 1\le j\le J$$

Είναι φανερό ότι $\lambda_J<\lambda_{J-1}<\ldots<\lambda_1<0$. Μάλιστα $\lambda_1=(-2+2\cos(\pi/(J+1)))/(\Delta x)^2$ $=(-2+2\cos(\pi\Delta x))/(\Delta x)^2=-\pi^2+O((\Delta x)^2)$ ενώ $\lambda_J=(-2+2\cos(J\pi/(J+1)))/(\Delta x)^2$ $=-4/(\Delta x)^2+O(1)$. Βλέπουμε δηλ. ότι γιά Δx μικρό

$$\lambda_J\cong-4(\Delta x)^{-2}<<\lambda_1\cong-\pi^2.$$

Προφανώς πρόκειται περί ακάμπτου συστήματος με πραγματικές ιδιοτιμές.

Επειδή στις εφαρμογές συναντάμε συχνά προβλήματα (όπως το (14) π.χ.) με πραγματικές αρνητικές ιδιοτιμές ή ιδιοτιμές με $\mathrm{Re}\lambda_i<0$ που βρίσκονται όμως σ' ένα <u>κώνο</u> γωνίας $0<\theta<\pi/2$ στο $\mathrm{Re}z<0$, δηλ. στο χωρίο $S_\theta=\{z:z=re^{i\varphi},\ \pi-\theta<\varphi<\pi+\theta,\ r>0\}$, δεν χρειάζεται να καταφεύγουμε πάντα σε μιά A-ευσταθή μέθοδο, αλλά π.χ. σε μία μέθοδο της οποίας η περιοχή απόλυτης ευστάθειας περιλαμβάνει τον αρνητικό πραγματικό ημιάξονα ή το S_θ αντίστοιχα. Λέμε συγκεκριμένα ότι μιά μέθοδος είναι <u>A_0-ευσταθής</u> (Cryer, 1973) αν η περιοχή της απόλυτης ευστάθειάς της περιέχει τον αρνητικό πραγματικό άξονα $\{z=x+iy,\ x<0,\ y=0\}$. Μία μέθοδος λέγεται <u>A(θ) - ευσταθής</u> (Widlund, 1967) αν η περιοχή απόλυτης ευστάθειάς της περιέχει των κώνο S_θ. Προφανώς γιά επιτυχή αριθμητική λύση του συστήματος (14) ((15)) αρκεί η μέθοδος να είναι A_0-ευσταθής οπότε γιά κάθε $h>0$ οι τιμές $h\lambda_i$ θα βρεθούν μέσα στην περιοχή απόλυτης ευστάθειας της μεθόδου.

Θα εξετάσουμε τώρα λεπτομερέστερα θέματα απόλυτης ευστάθειας (γιά το μιγαδικό πρόβλημα (11)) των δύο μεγάλων κατηγοριών μεθόδων που έχουμε ήδη μελετήσει, δηλ. των μεθόδων RK και των (γραμμικών) πολυβηματικών μεθόδων. Αρχίζουμε με τις μεθόδους RK. Θεωρούμε την γενική (πεπλεγμένη) μέθοδο RK με q-στάδια (3.2.12α,b)' την εφαρμόζουμε στο πρόβλημα (11) και παίρνουμε

$$(16\alpha)\quad y^{n,i} = y^n + h\lambda\sum_{j=1}^{q} a_{ij}y^{n,j},\ 1\le i\le q$$

$$(16\beta)\quad y^{n+1} = y^n + h\lambda\sum_{j=1}^{q} b_j\, y^{n,j}$$

λύνοντας ως πρός $y^{n,i}$ την σχέση (16α) (συμβολίζοντας με $A_q=(a_{ij})$ $1\le i,\ j\le q$ του πίνακα των συντελεστών a_{ij} ώστε να μην γίνεται σύγχιση με τον πίνακα A του συστήματος (10)) παίρνουμε, γιά $Y^n=(y^{n,1},\dots,y^{n,q})^T \in \mathbb{C}^q$, $u=(1,\dots,1)^T \in \mathbb{C}^q$, I_q ταυτότητα στον $\mathbb{C}^q$, ότι

$$Y^n = y^n(I_q-h\lambda A_q)^{-1}u,$$

και, αντικαθιστώντας στην (16β), τελικά ότι

$$y^{n+1} = y^n(1+h\lambda\, b^T(I_q-h\lambda A_q)^{-1}u).$$

Θέτοντας $z=h\lambda\in\mathbb{C}$ ($\mathrm{Re}z<0$) μπορούμε να γράψουμε την σχέση αυτή στην μορφή

$$(17)\quad y^{n+1} = r(z)y^n,\ z=h\lambda,\ n\ge 0,$$

όπου r είναι η συνάρτηση

$$(18)\quad r(z) = 1+zb^T(I_q-zA_q)^{-1}u,\ \mathrm{Re}z<0.$$

Υπό την προϋπόθεση ότι η $r(z)$ είναι καλά ορισμένη γιά $\mathrm{Re}z<0$ (ή γιά κατάλληλο υποσύνολο του $\mathrm{Re}z<0$) - δηλ. ότι ο πίνακας I_q-zA_q είναι αντιστρέψιμος - βλέπουμε, γράφοντας τον αντίστροφο $(I_q-zA_q)^{-1}$ συναρτήσει ορίζουσών ότι η $r(z)$ είναι ρητή συνάρτηση του z με βαθμούς αριθμητού και παρονομαστού το πολύ q.

Η σχέση (17) είναι το διακριτό ανάλογο της

$$(19)\quad y(t^{n+1}) = e^{z}y(t^{n}),\quad z=h\lambda,\ n\geq 0,$$

που ικανοποιεί η λύση $y(t)$ του (11). Η συνάρτηση $r(z)$ είναι συνεπώς μία ρητή προσέγγιση του εκθετικού e^{z} γιά $\mathrm{Re}z<0$ και, συνεπώς, οι ιδιότητες απόλυτης ευστάθειας της μεθόδου μεταφράζονται σε κατάλληλες ιδιότητες της $r(z)$. Είναι π.χ. προφανές από την (17), ότι μία μέθοδος RK είναι απόλυτα ευσταθής γιά κάποιο h αν το $z=h\lambda$ είναι τέτοιο ώστε η $r(z)$ να είναι καλά ορισμένη και να ικανοποιεί $|r(z)|<1$. Συνεπώς η περιοχή απόλυτης ευστάθειας S μιάς μεθόδου RK μπορεί να ορισθεί ως το σύνολο των $z=h\lambda$:

$$S = \{z:\mathrm{Re}z<0,\ r(z)\ \text{καλά ορισμένη και}\ |r(z)|<1\}.$$

Μία μέθοδος είναι A – ευσταθής αν $S = \{z:\mathrm{Re}z<0\}$, A_0 – ευσταθής αν $S\supset\{z:\mathrm{Re}z<0,\ \mathrm{Im}z=0\}$, κ.ο.κ.

Από τον ορισμό της τάξης ακρίβειας μιάς μεθόδου RK (βλ. Παρ. 3.2) συμπεραίνουμε ότι αν η μέθοδος (3.2.12α,β) έχει τάξη ακρίβειας p, τότε η ποσότητα $y(t^{n+1})-r(z)y(t^{n}) = (e^{z}-r(z))\,y(t^{n})$ θα πρέπει να είναι τάξης h^{p+1} γιά κάθε λ,n. Άρα θα ικανοποιεί την αναγκαία συνθήκη

$$(20)\quad e^{z}-r(z) = O(z^{p+1})\ \text{γιά}\ z\to 0 \ *$$

(Παρατηρείστε ότι η $r(z)$ είναι αναλυτική σε μία περιοχή του μηδενός).

Οι ρητές συναρτήσεις που αντιστοιχούν στην μέθοδο του Euler, στην πεπλεγμένη Euler και στην μέθοδο του τραπεζίου είναι, αντίστοιχα, $r_1(z)=1+z$, $r_2(z)=(1-z)^{-1}$, $r_3(z)=(1+z/2)/(1-z/2)$.

---

* Δεν είναι όμως γενικά σωστό ότι η (20) συνεπάγεται ότι η τάξη ακρίβειας μιάς μεθόδου RK που αντιστοιχεί στην ρητή συνάστηση $r(z)$ είναι p. Η (20) είναι ικανή συνθήκη έτσι ώστε η τάξη ακρίβειας της μεθόδου γιά γραμμικά προβλήματα με σταθερούς συντελεστές να είναι p.

ικανοποιούν, αντίστοιχα, $|r_1(z)|<1$ γιά $|z+1|<1$ (περιοχή απόλυτης ευστάθειας της μεθόδου του Euler) και $|r_i(z)|<1$ γιά $Re z<0$, αν $i=2,3$ (A-ευσταθείς μέθοδοι). (Παρατηρείστε ότι $e^z=r_i(z)+O(z^2)$ αν $i=1,2$ ενώ $e^z=r_3(z)+O(z^3)$ σε συμφωνία με την (20) και τις γνωστές μας τάξεις των τριών μεθόδων). Είναι φανερό ότι οι <u>άμεσες</u> μέθοδοι RK έχουν <u>πολυωνυμικές</u> συναρτήσεις $r(z)$ με $r(0)=1$. Αν μία άμεση μέθοδος έχει τάξη p η (20) δίνει ότι η αντίστοιχη $r(z)$ της θα είναι αναγκαστικά

$$(21)\quad r(z) = 1+z+z^2/2!+ \dots + z^p/p!$$

Π.χ. <u>όλες</u> οι άμεσες μέθοδοι RK τάξης $p=2$ έχουν $r(z)=1+z+z^2/2$. Είναι φανερό ότι δεν υπάρχει άμεση A-ευσταθής μέθοδος RK! (Οι περιοχές όπου $|r(z)|<1$ γιά άμεσες μεθόδους με $p=1,2,3,4$ είναι τα γραμμοσκιασμένα χωρία του Σχήματος 1)

Από τις μεθόδους που ξεχωρίσαμε στην Παρ. 3.2, η πεπλεγμένη μέθοδος του μέσου (3.2.15') συμπίπτει, γιά το πρόβλημα (11), με την μέθοδο του τραπεζίου και είναι συνεπώς A-ευσταθής με $r(z)=(1+z/2)/(1-z/2)$. (Δύο διαφορετικές μέθοδοι RK μπορεί να αντιστοιχούν στην ίδια ρητή προσέγγιση του εκθετικού). Οι ημιπεπλεγμένες μέθοδοι (3.2.18) με $q=2$, $p=3$, δηλ. με $\lambda=(1\pm 3^{-1/2})/2$ δίνουν την ρητή συνάρτηση

$$(22)\quad r(z) = (1+(1-2\lambda)z+(1/2-2\lambda+\lambda^2)z^2)/(1-\lambda z)^2$$

και είναι A-ευσταθείς. A-ευσταθής είναι επίσης και η ημιπεπλεγμένη μέθοδος (3.2.51) με $q=3$, $p=4$. Οι μέθοδοι "Gauss-Legendre q σημείων" (μέθοδος (3.2.19) αν $q=2$· βλ. Πόρισμα 3.2.1 (γ) γιά τον ορισμό τους γενικά) είναι πολύ ενδιαφέρουσες όπως ξέρουμε γιατί δίνουν την μέγιστη τάξη ακρίβειας ($p=2q$) γιά δεδομένο αριθμό σταδίων q. Η ρητή προσέγγιση που αντιστοιχεί στην (3.2.19) είναι η

$$(23)\quad r(z) = (1+z/2+z^2/12)/(1-z/2+z^2/12)$$

γιά την οποία ισχύει $|r(z)|<1$ γιά κάθε z: $Re z<0$, δηλ. ότι η αντίστοιχη μέθοδος είναι A-ευσταθής. Γενικά η μέθοδος Gauss-Legendre με q σημεία δίνει ως ρητή προσέγγιση του εκθετικού το <u>q-στο διαγώνιο στοιχείο του πίνακα Padé γιά την συνάρτηση e^z</u> (βλ. Παρατήρηση 2). Από γνωστή ιδιότητα του πίνακα Padé (Birkhoff-Varga, 1965) συμπεραίνουμε ότι όλες οι μέθοδοι Gauss-Legendre είναι A-ευσταθείς.

Εκτός από την περιπτωσιολογία υπάρχει φυσικά και πολλή θεωρία γιά τις ιδιότητες των ρητών προσεγγίσεων του εκθετικού e^z και ανάλογη σημαντική θεωρία χαρακτηρισμών ρητών συναρτήσεων $r(z)$ διαφόρων μορφών που οδηγούν σε A-ευσταθείς μεθόδους. Π.χ. ένα απλό αποτέλεσμα είναι το εξής:

ΠΡΟΤΑΣΗ 1 Αν οι ιδιοτιμές μ_i του πίνακα A_q ικανοποιούν $Re\mu_i \geq 0$, $1 \leq i \leq q$, αν η μέθοδος είναι συνεπής καί άν γιά κάθε $y \in R$ έχουμε $|r(iy)| \leq 1$, τότε $|r(z)|<1$ γιά $Re z<0$, δηλ. η αντίστοιχη μέθοδος είναι A-ευσταθής.

<u>Απόδειξη:</u> Επειδή οι ιδιοτιμές μ_i, $1 \leq i \leq q$ τού A_q ικανοποιούν $Re\mu_i \geq 0$ $\forall i$ καί επειδή οι ιδιοτιμές τού πίνακα $I_q - zA_q$ είναι οι $1-z\mu_i$, $1 \leq i \leq q$, συμπεραίνουμε ότι ο $I_q - zA_q$ είναι αντιστρέψιμος γιά κάθε z: $Re z<0$. (Πράγματι άν γιά κάποιο $\mu=\mu_i$, $\mu=1/z$, έχουμε ότι $Re\mu = Re z/|z|^2$, άτοπο). Συνεπώς, η ρητή συνάρτηση $r(z)$, βλ.(18), δέν έχει πόλους στό ημιεπίπεδο $Re z<0$. Από τήν υπόθεσή μας ότι $|r(iy)| \leq 1$, $\forall y \in R$ συμπεραίνουμε ότι δέν έχει πόλους ούτε στόν φανταστικό άξονα. Αρα η $r(z)$ είναι αναλυτική σέ μιά περιοχή τού κλειστού ημιεπιπέδου $Re z \leq 0$. Επειδή η $r(z)$ είναι ρητή καί ισχύει $|r(iy)| \leq 1$ $\forall y \in R$ συμπεραίνουμε ότι $\lim_{|z| \to \infty} |r(z)| \leq 1$. Αρα από τήν αρχή τού μεγίστου γιά αναλυτικές συναρτήσεις συμπεραίνουμε ότι $|r(z)| \leq 1$ $\forall z: Re z \leq 0$. Επειδή η μέθοδος είναι συνεπής ($p \geq 1$) η (20) δίνει ότι $r(z) \neq 1$ γιά $Re z \leq 0$. Συμπεραίνουμε, πάλι απ' την αρχή του μεγίστου, ότι, ακριβέστερα, $|r(z)|<1$ στο εσωτερικό $Re z<0$. @

Το αποτέλεσμα αυτό μας επιτρέπει, αν μπορεί να εφαρμοσθεί, να ελέγχουμε αν $|r(z)|\leq 1$ μόνο γιά $z=iy$, $y\in R$. Επίσης από την απόδειξή του συμπεραίνουμε ότι αν η μέθοδος είναι συνεπής (αρκεί $r(z)\neq 1$) και η $r(z)$ είναι καλά ορισμένη γιά $Rez\leq 0$, τότε $\{|r(z)|<1, Rez<0\}$ <=> $\{|r(z)|\leq 1, Rez\leq 0\}$. Άλλη μία ενδιαφέρουσα ικανή συνθήκη είναι η εξής: Αν οι ιδιοτιμές μ_i του A_q ικανοποιούν $Re\mu_i\geq 0$ $\forall i$, αν η τάξη p της μεθόδου ικανοποιεί την ανισότητα $p\geq 2q-2$ και αν $\lim_{x\to\infty}|r(x)|\leq 1$, τότε η μέθοδος είναι A-ευσταθής.

Όπως αναφέρθηκε προηγουμένως υπάρχει σημαντική θεωρία γιά την μελέτη των ιδιοτήτων (ιδίως ευστάθειας, ακρίβειας, εντοπισμού των πόλων κ.λ.π.) ρητών προσεγγίσεων της e^z που βασίζεται συνήθως σε προχωρημένα θέματα της θεωρίας μιγαδικών συναρτήσεων. Γιά μία γεύση των μεθόδων που χρησιμοποιούνται σήμερα στις αποδείξεις βλ. π.χ. το άρθρο των G. Wanner, E. Hairer και S.P. Nørsett, "Order stars and stability theorems", BIT 18(1978), 475-89.

Προχωρούμε τώρα στην μελέτη της απόλυτης ευστάθειας (γιά την μιγαδική απλή Δ.Ε. (11)) των (γραμμικών) πολυβηματικών μεθόδων - που προηγείται ιστορικά (βλ. Dahlquist, 1963) της ανάλογης μελέτης γιά τις μεθόδους RK -. Εφαρμόζοντας την k-βηματική μέθοδο (3.3.3) στην εξίσωση (11) παίρνουμε την ομογενή εξίσωση διαφορών

$$(24)\quad \sum_{j=0}^{k}(a_j-h\lambda\beta_j)y^{n+j}=0,\qquad n\geq 0$$

Γιά να ισχύει $y^n\to 0$, $n\to\infty$ γιά κάποιο $h>0$ γιά κάθε λύση της (24) με $y^0=1$ πρέπει συνεπώς (βάσει της θεωρίας των λύσεων ομογενών εξισώσεων διαφορών με σταθερούς συντελεστές, βλ. Παρ. 3.3., ιδίως (3.3.1), (3.3.12)) οι ρίζες $\zeta_i=\zeta_i(\lambda h)$, $1\leq i\leq k$, του πολυωνύμου

$$(25)\quad \pi(\zeta,\lambda h)\equiv\rho(\zeta)-\lambda h\sigma(\zeta),\ Re\lambda<0$$

να ικανοποιούν

(26) $|\zeta_i|<1,\ 1\leq i\leq k$.

(Τονίζουμε την <u>αυστηρή</u> ανισότητα στην (26)). Το πρόβλημα συνεπώς του προσδιορισμού της απόλυτης ευστάθειας της μεθόδου (3.3.3) ανάγεται στην μελέτη των ριζών ζ_i του π ως συναρτήσεων της μιγαδικής παραμέτρου $h\lambda$, $\mathrm{Re}\,\lambda<0$ και στην διατύπωση συνθηκών έτσι ώστε να ισχύει η (26).

Ας δούμε μερικά παραδείγματα. Γιά την μέθοδο (3.1.1') έχουμε $\rho(\zeta)=\zeta^2-1$, $\sigma(\zeta)=2\zeta$· συνεπώς $\pi=\zeta^2-2h\lambda\zeta-1$, του οποίου οι ρίζες ζ_1,ζ_2 ικανοποιούν, γιά κάθε $h\lambda\in\mathbb{C}$, την συνθήκη $|\zeta_1||\zeta_2|=1$. Συνεπώς δεν είναι δυνατόν να ισχύει η (26) γιά καμμιά τιμή του $h\lambda$: η μέθοδος (3.1.1') - αν και "ευσταθής" σύμφωνα με τον ορισμό της Παρ. 3.3 - δεν είναι ποτέ απόλυτα ευσταθής. Το ίδιο συμβαίνει και με την μέθοδο του Simpson (3.3.2). 'Εχουμε ήδη δεί ότι οι μέθοδοι πεπλεγμένη Euler και τραπεζίου είναι A-ευσταθείς. Οι μέθοδοι "οπισθοδρομικών διαφορών" με k βήματα (3.3.6), $1\leq k\leq 6$ έχουν τις εξής ιδιότητες: Γιά $k=1$ (πεπλεγμένη Euler) και $k=2$ είναι A-ευσταθείς. Γιά $3\leq k\leq 6$ είναι <u>A(θ)-ευσταθείς</u> με γωνίες θ_k, $3\leq k\leq 6$ τίς εξής: $\theta_3\cong 88^o$, $\theta_4\cong 73^o$, $\theta_5\cong 51^o$, $\theta_6\cong 18^o$. Είναι δηλ. κατάλληλες γιά άκαμπτα συστήματα με ιδιοτιμές που βρίσκονται μέσα στούς αντίστοιχους κώνους $S_{\theta k}$. Θα δούμε στις Ασκήσεις και άλλα παραδείγματα· Βλ. και Παρατήρηση 1.

Υπάρχει σημαντικό σώμα θεωρίας γιά τις ιδιότητες απόλυτης ευστάθειας των (γραμμικών) πολυβηματικών μεθόδων, που αρχίζει με την εργασία του G.D. Dahlquist "A special stability problem for linear multistep methods", BIT 3(1963), 27-43. Ο Dahlquist (αφού δίνει τον ορισμό της απόλυτης ευστάθειας) αποδεικνύει ότι δεν υπάρχουν άμεσες ($\beta_k=0$) A-ευσταθείς πολυβηματικές μέθοδοι και ότι η τάξη ακρίβειας p μιάς A-ευσταθούς πολυβηματικής μεθόδου δεν μπορεί να είναι μεγαλύτερη του 2. Μάλιστα, η μέθοδος του τραπεζίου είναι εκείνη η A-ευσταθής μέθοδος με $p=2$ με την μικρότερη "σταθερά σφάλματος" (Βλ. Παρατήρηση 3.3.2) $c_*=c_3/\sigma(1)=1/12$. 'Ετσι έχουμε ένα σοβαρό περιορισμό στην τάξη ακρίβειας μιάς A-ευσταθούς πολυβηματικής μεθόδου, πράγμα που αναδεικνύει την σημασία των (πεπλεγμένων) μεθόδων RK μεταξύ των οποίων υπάρχουν A-ευσταθείς μέθοδοι οποιασδήποτε τάξης ακρίβειας (π.χ. οι μέθοδοι Gauss-Legendre με q στάδια και p=2q).

Η κατάσταση όμως είναι πολύ καλύτερη όταν δεν επιβάλουμε A-ευστάθεια αλλά αρκεσθούμε σε A(θ)-ευστάθεια γιά $0<\theta<\pi/2$ ή A_0-ευστάθεια. (Γιά την αριθμητική λύση με πλήρως διακριτές μεθόδους π.χ. προβλημάτων παραβολικών ΜΔΕ, συνήθως A_0 - ή A(θ) - ευστάθεια είναι υπεραρκετές. Δεν ισχύει όμως αυτό γιά υπερβολικά π.χ. προβλήματα όπου οι ιδιοτιμές του A είναι φανταστικές βλ. και Ασκ.7). Είναι γνωστό (Widlund, 1967) ότι δεν υπάρχουν άμεσες A(θ)-ευσταθείς μέθοδοι και ότι η μόνη A(θ)-ευσταθής k-βηματική μέθοδος της οποίας η τάξη p υπερβαίνει το k είναι η μέθοδος του τραπεζίου (k=1,p=2). Αλλά είναι γνωστό ότι γιά <u>κάθε</u> $\theta\in(0,\pi/2)$ υπάρχουν A(θ)-ευσταθείς μέθοδοι με k=p=3 και k=p=4 και επιπλέον ότι γιά τουλάχιστον $1\le k\le 6$ υπάρχουν $A(\theta_k)$-ευσταθείς μέθοδοι, $0<\theta_k<\pi/2$, με p=k (οι μέθοδοι "οπισθοδρομικών διαφορών"). Αν περιοριστούμε σε A_0-ευσταθείς μεθόδους, είναι γνωστό (Cryer, 1973) ότι ισχύει πάλι ότι οι A_0-ευσταθείς μέθοδοι είναι πεπλεγμένες και ότι η τάξη τους p δεν μπορεί να υπερβαίνει το k με εξαίρεση την μέθοδο του τραπεζίου. Όμως γιά <u>κάθε</u> αριθμό βημάτων k υπάρχουν A_0-ευσταθείς μέθοδοι τάξης p=k. Γιά μία πολύ καλή ανασκόπηση των παραπάνω (αλλά και πολλών άλλων παρόμοιων) κλασικών αποτελεσμάτων γιά την απόλυτη ευστάθεια των πολυβηματικών μεθόδων και γιά τις αποδείξεις τους με ένα ενοποιημένο και στοιχειώδη τρόπο βλ. την παράγραφο 3.2. του β' τόμου του βιβλίου [3.3] του Grigorieff. Γιά μιά πρόσφατη ανασκόπηση νεωτέρων αποτελεσμάτων με πιό προχωρημένες μεθόδους της θεωρίας μιγαδικών συναρτήσεων, βλ. π.χ. το άρθρο των Jeltsch και Nevanlinna, Numer. Math. 40(1982), 245-296.

Η μελέτη της απόλυτης ευστάθειας των μεθόδων RK και των πολυβηματικών μεθόδων έγινε με βάση την συμπεριφορά αυτών των μεθόδων όταν εφαρμοσθούν στην απλή μιγαδική Δ.Ε. (11). Ας δούμε τώρα την σημασία αυτών των ιδιοτήτων γιά την αριθμητική λύση του mxm πραγματικού <u>συστήματος</u> (10) του οποίου το μη ομογενές ανάλογο γράφουμε ως

$$(27)\begin{cases} y' = My + f(t),\ t\geq 0, \\ y(0) = y_o, \end{cases}$$

και γιά το οποίο, γενικεύοντας λίγο, υποθέτουμε ότι γιά όλες τις ιδιοτιμές λ_i του πίνακα $M\in R^{m\times m}$ ισχύει $Re\lambda_i\leq 0$ και ότι υπάρχουν ιδιοτιμές τέτοιες ώστε π.χ. $|Re\lambda_i|=O(1)$ ενώ γιά άλλες $|Re\lambda_i|>>1$, δηλ. ότι το (27) είναι άκαμπτο. (Γράφουμε M αντί A ώστε να μην υπάρξει σύγχιση μεταξύ των στοιχείων του M και των συντελεστών a_{ij} της μεθόδου RK). Ας εξετάσουμε πρώτα την ευστάθεια των μεθόδων RK γιά το (27). Η μέθοδος (3.2.12a,b) στην περίπτωση του (27) γίνεται (γιά $n\geq 0$, $y^o=y_o$):

$$(28)\quad\begin{cases} y^{n,i} = y^n + h\sum_{j=1}^{q} a_{ij}\ (My^{n,j}+f(t^{n,j})),\ 1\leq i\leq q, \\ y^{n+1} = y^n + h\sum_{j=1}^{q} b_j\ (My^{n,j}+f(t^{n,j})). \end{cases}$$

Υποθέτουμε ότι το γραμμικό σύστημα (ως πρός $y^{n,i}$) που ορίζεται μέσω της (28) έχει μοναδική λύση γιά κάθε n. (Αυτό π.χ. εξασφαλίζεται, χωρίς περιορισμό στο h, αν οι ιδιοτιμές μ_i του πίνακα $A_q=(a_{ij})$ έχουν θετικό πραγματικό μέρος, βλ. Πρόταση 1). Σε αναλογία με ό,τι κάναμε στην Παράγραφο 3.2 (Πρόταση 3.2.2) θεωρούμε την "διαταραχή" του (28) με δεδομένα $z^o\in R^m$, $\epsilon^n\in R^m$, $n\geq 0$:

$$(29)\quad\begin{cases} z^{n,i} = z^n + h\sum_{j=1}^{q} a_{ij}(Mz^{n,j}+f(t^{n,j})),\ 1\leq i\leq q, \\ z^{n+1} = z^n + h\sum_{j=1}^{q} a_{ij}\ (Mz^{n,j}+f(t^{n,j}))+\epsilon^n. \end{cases}$$

Γιά τεχνικούς λόγους θα μελετήσουμε την ευστάθεια της (28) ως πρός την ευκλείδια νόρμα $\|\cdot\|=\|\cdot\|_2$ στου R^m. Σκοπός μας είναι να φράξουμε την διαφορά $\|y^n-z^n\|$ συναρτήσει των $\|y^0-z^0\|$, $\max_{0\le j\le n-1}\|\varepsilon^j\|$ χωρίς όμως να χρησιμοποιήσουμε το αποτέλεσμα (3.2.25) της Πρότασης 3.2.2., το οποίο φυσικά ισχύει και εδώ αλλά με σταθερά Lipschitz $L=\|M\|=\max_{1\le i\le m}|\lambda_i(M^*M)|^{1/2}$ που μπορεί να γίνει πολύ μεγάλη (π.χ. θεωρείστε $M^*=M$ οπότε $\lambda_i(M^*M)=\lambda_i^2$) λόγω της ακαμψίας του (27). Συνεπώς, στην περίπτωσή μας, οι σταθερές C_1 και C_2 της (3.2.25) είναι πολύ μεγάλες, η δε (3.2.25) πρακτικά άχρηστη.

Έστω ότι η μέθοδος RK είναι <u>A-ευσταθής</u>. Τότε ισχύει, γιά κάθε $n\ge 0$ ότι γιά κάθε $h>0$

(30) $\|y^n-z^n\| \le \|y^0-z^0\| + n \max_{0\le j\le n-1} \|\varepsilon^n\|$.

Η απόδειξη της (30) γιά γενικούς πίνακες M με $Re\lambda_i\le 0$ είναι τεχνικά όχι τόσο απλή (Crouzeix 1975). Γιά <u>συμμετρικούς</u> πίνακες M η απόδειξη είναι πολύ ευκολώτερη· έστω λοιπόν M <u>συμμετρικός</u> m×m πίνακας με ιδιοτιμές $\lambda_i\le 0$, $1\le i\le m$. Στην ειδική αυτή περίπτωση θα αποδείξουμε την (30) γιά A_0-ευσταθείς μεθόδους, γιά μεθόδους δηλ. γιά τις οποίες ισχύει ότι $|r(x)|<1$ γιά $x<0$, όπου $r(x)$ η ρητή προσέγγιση του e^x γιά $x<0$ που δίνεται από την (18). Υποθέτοντας ότι η μέθοδος είναι τουλάχιστον συνεπής (ακριβέστερα ότι $r(x)\not\equiv 1$) αυτό είναι ισοδύναμο με την ασθενέστερη συνθήκη ότι

(31) $|r(x)|\le 1$ γιά $x\le 0$,

την οποία και υποθέτουμε ότι ισχύει. Χρησιμοποιούμε τώρα την <u>φασματική παράσταση</u> του πίνακα M, δηλ. ότι ο M έχει m ορθοκανονικά ιδιοδιανύσματα u^i τέτοια ώστε $Mu^i=\lambda_i u^i$. Τότε, γιά κάθε πραγματικό πολυώνυμο π έχουμε $\forall u\in R^m$ $\pi(M)u = \sum_{i=1}^{m} \pi(\lambda_i)(u,u^i)u^i$, όπου $(\cdot,\cdot)$ το ευκλείδιο εσωτερικό γινόμενο στον R^m. Επίσης, αν ο πίνακας $\alpha I+\beta M$

είναι αντιστρέψιμος, έχουμε ότι $(\alpha I+\beta M)^{-1}u = \sum_{i=1}^{m} (\alpha+\beta\lambda_i)^{-1}(u,u^i)u^i$.
Συμπεραίνουμε ότι γιά κάθε ρητή συνάρτηση $r(x)$, καλά ορισμένη γιά $x\leq 0$ μπορούμε να ορίσουμε γιά κάθε $u\in R^m$

$$r(M)u = \sum_{i=1}^{m} r(\lambda_i)(u,u^i)u^i.$$

Αφαιρώντας κατά μέλη τις (28) και (29) έχουμε τώρα ότι

$$y^{n,i}-z^{n,i} = y^n-z^n + h\sum_{j=1}^{q} a_{ij}\, M(y^{n,j}-z^{n,j})$$

$$y^{n+1}-z^{n+1} = y^n-z^n + h \sum_{j=1}^{q} b_j M\,(y^{n,j}-z^{n,j})-\epsilon^n.$$

Εκφράζοντας τα διανύσματα $y^n,y^{n,i},z^n,z^{n,i}$ ως αναπτύγματα ιδιοδιανυσμάτων του M και κάνοντας λίγες πράξεις δεν είναι δύσκολο να δούμε ότι, σε αναλογία με την (17) π.χ., τώρα ισχύει

$$y^{n+1} - z^{n+1} = r(hM)(y^n-z^n) - \epsilon^n,\ n\geq 0,$$

από την οποία προκύπτει με επανάληψη η σχέση

$$(33)\ \|y^n-z^n\| \leq \|r(hM)\|^n\, \|y^0-z^0\|+ \sum_{j=0}^{n-1} \|r(hM)\|^{n-1-j}\, \|\epsilon^j\|,\ n\geq 1.$$

Τώρα, γιά κάθε συνάρτηση $\varphi(x)$ ορισμένη γιά $x=\lambda_i$, $1\leq i\leq m$, θεωρούμε τον αντίστοιχο γραμμικό τελεστή $\varphi(M)$ που ορίζεται γιά $u\in R^m$ ως

$$(34)\ \varphi(M)u = \sum_{i=1}^{m} \varphi(\lambda_i)(u,u^i)u^i.$$

Εχουμε τότε ότι λόγω της ορθογωνιότητας των u^i

$$\|\varphi(M)u\|^2 = \sum_{i=1}^{m} [\varphi(\lambda_i)]^2\, (u,u^i)^2 \leq \max_{1\leq i\leq m} [\varphi(\lambda_i)]^2\, \|u\|^2$$

Συνεπώς ισχύει $\|\varphi(M)u\| \le \max_{1\le i\le m} |\varphi(\lambda_i)|\ \|u\|$ γιά κάθε $u\in R^m$, με ισότητα αν $u=u^s$ όπου $|\varphi(\lambda_s)|=\max_i|\varphi(\lambda_i)|$. Συνεπώς αν η $\varphi(M)$ ορίζεται από την (34) έχουμε

$$(35)\ \|\varphi(M)\| = \max_{1\le i\le m} |\varphi(\lambda_i)|.$$

Συμπεραίνουμε, λόγω της (31), επειδή $\lambda_i\le 0$,

$$(36)\ \|r(hM)\| = \max_{1\le i\le m} |r(h\lambda_i)| \le 1\ \ \forall h>0.$$

Η (36) και η (33) δίνουν συνεπώς την ανισότητα (30), η οποία πραγματικά εκφράζει την ευστάθεια της μεθόδου, κάτω από τις προϋποθέσεις μας, στην περίπτωση τού χραμμικού συστήματος (27).

Τι μπορούμε να πούμε τώρα γιά μεθόδους που δεν είναι A-ευσταθείς, π.χ. γιά μεθόδους, όπως οι άμεσες RK, γιά τις οποίες η συνθήκη $|r(z)|\le 1$ ισχύει μόνο γιά ένα φραγμένο υποσύνολο του ημιεπιπέδου $\mathrm{Re}z\le 0$; Ενα γενικό αποτέλεσμα δίνεται από τον Crouzeix, op. cit. Εδώ ας περιοριστούμε πάλι σε άκαμπτα συστήματα με συμμετρικούς πίνακες M με $\lambda_i\le 0$, $1\le i\le m$. Υποθέτουμε ότι τώρα ισχύει η $|r(x)|\le 1$ <u>μόνο σ' ένα διάστημα</u> δηλ. ότι αντί της (31) έχουμε γιά κάποιο $\alpha<0$ ότι

$$(37)\ |r(x)| \le 1,\ x\in(\alpha,0].$$

(Π.χ. γιά την μέθοδο του Euler ή την μέθοδο του μέσου $\alpha=-2$ κλπ.). Τότε η (36) δεν ισχύει γιά κάθε $h>0$ αλλά γιά περιορισμένα h. Πράγματι, λόγω της (37), μόνο αν $h\lambda_i \in (\alpha,0]$, $1\le i\le m$, δηλ. αν

$$(38)\ h \max_{1\le i\le m} |\lambda_i| < |\alpha|,$$

θα έχουμε ότι

$$(39)\ \|r(hM)\| = \max_{1\le i\le m} |r(h\lambda_i)| \le 1.$$

Συνεπώς αν ισχύει η (37), τότε γιά μικρό h (έτσι ώστε να ικανοποιείται η (38), που γιά άκαμπτο σύστημα αποτελεί συνήθως σοβαρό περιορισμό στο h) προκύπτει πάλι η (30). (Γιαυτόν τον λόγο οι A-, A(θ) - ή A_0- ευσταθείς μέθοδοι λέγονται μερικές φορές και "απεριόριστα ευσταθείς" ενώ μιά μέθοδος γιά την οποία ισχύει η (37) λέγεται και "ευσταθής υπό συνθήκη" - την (38)).

Η σχέση (3.2.25) ήταν το "κλειδί" στην απόδειξη της εκτίμησης (3.2.31) γιά την τάξη του σφάλματος $\|y^n-y(t^n)\|$ των μεθόδων RK υπό την προϋπόθεση ότι είναι γνωστή μία σχέση όπως η (3.2.30) όπου $\delta^n=\varepsilon^n$. Αν υποθέσουμε και εδώ ότι η τάξη ακρίβειας της μεθόδου μας είναι p, τότε ισχύει εξ ορισμού σχεδόν μία σχέση της μορφής (3.2.30). Όμως η σταθερά D θα είναι τυπικά πολυωνυμική συνάρτηση νορμών υψηλών παραγώγων $y^{(j)}$ της λύσης του συστήματος (37), δηλ. υψηλών δυνάμεων του πίνακα M, δηλ. υψηλών δυνάμεων των ιδιοτιμών λ_i του M, οπότε η σταθερά C της (3.2.31) θα είναι πολύ μεγάλη και πρακτικά άχρηστη γιά άκαμπτο σύστημα. Η απόδειξη φραγμάτων με άριστη τάξη ακρίβειας $O(h^p)$ αλλά με "λογικές" σταθερές C είναι πρόβλημα που δεν μπορεί συνήθως να επιλυθεί στο επίπεδο των συνήθων διαφορικών εξισώσεων. Χρειάζεται να εισάγουμε ειδικές νόρμες και να χρησιμοποιήσουμε την ομαλότητα της λύσης του προβλήματος (π.χ. της ΜΔΕ) από το οποίο προέρχεται το σύστημά μας. Σχετικά εύκολα όμως μπορεί κανείς να αποδείξει εκτιμήσεις της μορφής

$$\|y^n-y(t^n)\| \le Ch^p(t^n)^{-p}\,\|y^0\|,\ n\ge 1$$

γιά μία A_0-ευσταθή μέθοδο π.χ. με $|r(x)|<1$ γιά $x<0$, $\lim_{|x|\to\infty}|r(x)|<1$, M συμμετρικό με $\lambda_i\le 0$ και όπου $\|\cdot\|=\|\cdot\|_2$· βλ. π.χ. LeRoux, Math. Comp. 33(1979), 919-931 και Baker, Bramble και Thomée, Math. Comp. 31(1977), 818-847. Εκτιμήσεις τέτοιου τύπου είναι ουσιαστικά εκτιμήσεις του πόσο γρήγορα φθίνει το σφάλμα (εδώ όπως το n^{-p}) γιά μεγάλο n. Το C είναι ανεξάρτητο των h,n και y(t) (και των λ_i).

Στην περίπτωση πολυβηματικών μεθόδων υπάρχει επίσης ανάλογη θεωρία (Zlamal 1975, Crouzeix-Raviart, 1978). Γιά το γενικό πρόβλημα (27) με $Re\lambda_i\le 0$ ισχύει ότι αν η μέθοδος είναι A-ευσταθής και έχει τάξη p ($p\leq 2$!) τότε, γιά κάποια σταθερά C ανεξάρτητη των n,h, y(t), λ_i, έχουμε

$$\|y(t^n)-y^n\| \leq C(\max_{0\leq i\leq k-1}\|y(t^i)-y^i\| + h^p\int_0^{t^n}\|y^{(p+1)}(t)\|\,dt).$$

Παρόμοια εκτίμηση ισχύει και γιά A_0-ευσταθείς μεθόδους όταν ο M είναι συμμετρικός με $\lambda_i\leq 0$ οπότε μπορούμε να επιτύχουμε και τάξη ακρίβειας p>2. Βέβαια πάλι $\|y^{(p+1)}(t)\|=O(\max_i|\lambda_i|^{p+1})$.

Τα τελευταία 10-15 χρόνια οι προσπάθειες στην περιοχή της απόλυτης ευστάθειας έχουν στραφεί σε γενικεύσεις της έννοιας γιά εφαρμογή σε άκαμπτα μη γραμμικά προβλήματα ή σε γραμμικά άκαμπτα προβλήματα με μεταβλητούς συντελεστές. Τα προβλήματα αυτά έχουν γενικά λύσεις που φθίνουν καθώς αυξάνει το t αλλά έχουν πολύ μεγάλες σταθερές Lipschitz έτσι ώστε η θεωρία ευστάθειας των Παρ. 3.2. και 3.3 να μην είναι πρακτικά εφαρμόσιμη. Η έρευνα έχει αποδειχθεί πολύ καρποφόρα' έχουν ήδη εντοπισθεί και χαρακτηρισθεί οικογένειες αποτελεσματικών και πρακτικά ευσταθών μεθόδων ικανοποιητικής ακρίβειας ενώ έχουν κατανοηθεί σ' ένα βαθμό μηχανισμοί "αποσταθεροποίησης" ακόμα και A-ευσταθών (γιά γραμμικά προβλήματα) μεθόδων σε άκαμπτα μη γραμμικά (ή γραμμικά με μεταβλητούς συντελεστές) προβλήματα. Η έρευνα συνεχίζεται' μιά, κάπως ασύνδετη, εικόνα της κατάστασης γιά μεθόδους RK μέχρι το 1984 δίνει το βιβλίο [3.1]. 'Εδώ θα περιοριστούμε σε μερικές βασικές παρατηρήσεις, μόνο γιά μεθόδους RK*.

Θα παραλείψουμε το ενδιάμεσο στάδιο των γραμμικών προβλημάτων με μεταβλητούς συντελεστές. Θεωρούμε το μιγαδικό σύστημα των μη γραμμικών Δ.Ε.

$$(40)\qquad \begin{cases} y' = f(t,y), & t\geq 0,\\ y(0) = y_0, \end{cases}$$

---

* Βλ. π.χ. εργασίες των Butcher, BIT 15(1975) 358-361, Crouzeix, Num. Math. 32(1979), 75-82, Burrage-Butcher SIAM J.N.A. 16(1979), 46-57.

όπου η $y(t)$ και η f έχουν τιμές στον $\mathbb{C}^m$. Υποθέτουμε ότι το (40) έχει μοναδική λύση (τουλάχιστον τοπικά, σ' ένα διάστημα $[0,T]$) και ότι ισχύει η σχέση

$$(41)\quad \mathrm{Re}\{(f(t,y)-f(t,z),y-z)\}\leq 0\quad \forall t\geq 0,\ y,z\in\mathbb{C}^m,$$

όπου τώρα με (x,y) συμβολίζουμε το ευκλείδιο εσωτερικό γινόμενο $\sum_{i=1}^{m} x_i\bar{y}_i$ στον $\mathbb{C}^m$ (και με $\|\cdot\|$ την αντίστοιχη νόρμα). Γιά δύο λύσεις $y(t)$, $z(t)$ του (40) σ' ένα διάστημα $[t',t'']$ έχουμε

$$y'(t)-z'(t) = f(t,y(t)) - f(t,z(t)),\quad t'\leq t\leq t''.$$

Συνεπώς, παίρνοντας το εσωτερικό γινόμενο με την $y(t)-z(t)$ έχουμε, επειδή

$$(y'-z',y-z) = (d/dt)\|y-z\|^2 - (y-z,y'-z'),$$

δηλ. επειδή

$$2\ \mathrm{Re}\{(y'-z',y-z)\} = (d/dt)\ \|y-z\|^2,$$

ότι λόγω της (41),

$$(d/dt)\ \|y-z\|^2 = 2\ \mathrm{Re}\{(f(t,y)-f(t,z),y-z)\}\leq 0,\quad t'\leq t\leq t''$$

και συνεπώς ότι

$$(42)\quad \|y(t'')-z(t'')\|\leq\|y(t')-z(t')\|,\quad t'\leq t''.$$

Σημειώστε ότι η σταθερά Lipschitz της f (αν η f είναι Lipschitz) δεν εμφανίζεται στην ανισότητα!

Λέμε ότι η μέθοδος RK της μορφής (3.2.12a,b) είναι <u>B-ευσταθής</u> (Butcher, Crouzeix), αν, όταν εφαρμοσθεί στο πρόβλημα (40) (με την ιδιότητα (41)), ικανοποιεί το διακριτό ανάλογο της (42), δηλ. αν γιά οποιεσδήποτε δύο λύσεις y^n, z^n, $y^{n,i}, z^{n,i}$ των (3.2.12a,b) ισχύει

$$(43)\quad \|y^{n+1}-z^{n+1}\| \le \|y^n-z^n\|,\ n\ge 0,$$

οπότε πραγματικά έχουμε ευστάθεια των λύσεων του (3.2.12a,b). Σημειώστε ότι μιά B-ευσταθής μέθοδος εφαρμοζόμενη σε μία μιγαδική εξίσωση της μορφής $y'=\lambda y$, $Re\lambda\le 0$, δίνει, επειδή ισχύει η (41), $|y^{n+1}-z^{n+1}|\le|y^n-z^n|$, δηλ. (με $z^n=0$), $|y^{n+1}|\le|y^n|$ => $|r(h\lambda)|\le 1$ $\forall h>0$, $\lambda: Re\lambda\le 0$, δηλ. ότι η μέθοδος - υποθέτουμε ότι $r(z)\not\equiv 1$ - είναι A-ευσταθής. Άρα B-ευστάθεια => A-ευστάθεια.

Προχωρούμε τώρα σε ικανές και αναγκαίες συνθήκες ώστε η μέθοδος RK που αντιστοιχεί στο μητρώο (3.2.14) να είναι B-ευσταθής. Θεωρούμε τον qxq συμμετρικό πίνακα $M=(m_{ij})$ που ορίζεται από

$$(44)\quad m_{ij} = b_i a_{ij}+b_j a_{ji}-b_i b_j,\ 1\le i,j\le q.$$

ΠΡΟΤΑΣΗ 2. (Butcher, Crouzeix). Υποθέτουμε ότι οι αριθμοί τ_i, $1\le i\le q$ είναι διάφοροι μεταξύ τους. Τότε η μέθοδος RK (3.2.12a,b) είναι B-ευσταθής αν και μόνο αν $b_i\ge 0$ και ο πίνακας M ικανοποιεί $z^*Mz\ge 0$ $\forall z\in\mathbb{C}^q$, $z^*=\bar{z}^T$. (Οι συνθήκες αυτές στα b_i και στον M λέγονται και συνθήκες "αλγεβρικής ευστάθειας" της μεθόδου). Αν τα τ_i δεν είναι όλα διαφορετικά, τότε οι συνθήκες $b_i\ge 0$ και $z^*Mz\ge 0$, $\forall z\in\mathbb{C}^q$, είναι μόνο ικανές γιά B-ευστάθεια.

<u>Απόδειξη</u>: Θα αποδείξουμε μόνο το ικανό γιά οποιαδήποτε τ_i, $1\le i\le q$. Γιά απλούστευση της απόδειξης (έτσι ώστε να μην ανησυχούμε γιά συμμετρία εσωτερικών γινομένων και γιά να μην παίρνουμε όλο πραγματικά μέρη) υποθέτουμε ότι το πρόβλημα (40) είναι πραγματικό, ότι ισχύει η (41), ότι τα $(\cdot,\cdot)$, $\|\cdot\|$ είναι το ευκλείδιο εσωτ. γινόμενο, αντιστχ. νόρμα, στον $\mathbb{R}^m$ και ότι $x^TMx\ge 0$ $\forall x\in\mathbb{R}^q$. Θεωρούμε δύο λύσεις $y^n, z^n, y^{n,i}, z^{n,i}$ των (3.2.12a,b). Αφαιρώντας έχουμε

$$(45)\quad y^{n,i}-z^{n,i} = y^n-z^n + \sum_{j=1}^{q} a_{ij}\,\varphi^j,\ 1\le i\le q$$

$$(46)\ y^{n+1}-z^{n+1} = y^n-z^n + \sum_{i=1}^{q} b_i\ \varphi^i,$$

όπου θέσαμε

$$(47)\ \varphi^j = h(f(t^{n,j},y^{n,j})-f(t^{n,j},z^{n,j})) \in R^m,\ 1\le j\le q.$$

Η (46) δίνει

$$(48)\ \|y^{n+1}-z^{n+1}\|^2 = \|y^n-z^n\|^2 + 2\sum_{i=1}^{q} b_i(\varphi^i,y^n-z^n)+ \|\sum_{i=1}^{q} b_i\varphi^i\|^2.$$

Χρησιμοποιώντας τώρα τις σχέσεις (45) έχουμε

$$\sum_{i=1}^{q} b_i(\varphi^i,y^n-z^n) = \sum_{i=1}^{q} b_i\ (\varphi^i,y^{n,i}-z^{n,i}) - \sum_{i,j=1}^{q} b_i a_{ij}\ (\varphi^i,\varphi^j).$$

Συνεπώς, η (48) γίνεται

$$(49)\ \|y^{n+1}-z^{n+1}\|^2 = \|y^n-z^n\|^2 + 2\ \sum_{i=1}^{q} b_i(\varphi^i,y^{n,i}-z^{n,i})$$

$$+ \sum_{i,j=1}^{q} b_i b_j(\varphi^i,\varphi^j) - 2\ \sum_{i,j=1}^{q} b_i a_{ij}(\varphi^i,\varphi^j).$$

Τώρα

$$(\varphi^i,y^{n,i}-z^{n,i}) = h(f(t^{n,i},y^{n,i}) - f(t^{n,i},z^{n,i}),\ y^{n,i}-z^{n,i})\le 0$$

λόγω της (41). Επίσης λόγω της συμμετρίας του $(\cdot,\cdot)$ έχουμε

$$2\ \sum_{i,j=1}^{q} b_i a_{ij}(\varphi^i,\varphi^j) = \sum_{i,j=1}^{q} b_i\ a_{ij}(\varphi^i,\varphi^j) + \sum_{i,j=1}^{q} b_j a_{ji}(\varphi^i,\varphi^j).$$

Άρα η (49), λόγω της υπόθεσης $b_i\ge 0$ δίνει

$$(50)\ \|y^{n+1}-z^{n+1}\|^2 \le \|y^n-z^n\|^2 - \sum_{i,j=1}^{q} (b_i a_{ij}+b_j a_{ji}-b_i b_j)/(\varphi^i,\varphi^j)$$

$$= \|y^n-z^n\|^2 - \sum_{i,j=1}^{q} m_{ij}(\varphi^i,\varphi^j).$$

'Εστω $\{\xi^k\}$, $0\le k\le m$ μιά ορθοκανονική βάση του R^m ως πρός το $(\cdot,\cdot)$.

Τότε, αν $\varphi^i=\sum_{k=1}^{m} c_k^{(i)}\xi^k$, $\varphi^j=\sum_{k=1}^{m} c_k^{(j)}\xi^k$ έχουμε $(\varphi^i,\varphi^j)=\sum_{k=1}^{m} c_k^{(i)}c_k^{(j)}$.

'Αρα

$$\sum_{i,j=1}^{q} m_{ij}(\varphi^i,\varphi^j) = \sum_{i,j=1}^{q} m_{ij} \left(\sum_{k=1}^{m} c_k^{(i)}c_k^{(j)}\right) =$$

$$= \sum_{k=1}^{m} \left(\sum_{i,j=1}^{q} c^{(i)}\, m_{ij}\, c^{(j)}\right) \ge 0,$$

επειδή $x^T M x \ge 0 \quad \forall\, x\in R^q$. Συνεπώς η (50) δίνει την (43). @

Η μέθοδος RK

λ	λ
1	

με $q=1$ είναι προφανώς B-ευσταθής αν $\lambda\ge 1/2$. Συνεπώς και η πεπλεγμένη μέθοδος του Euler ($\lambda=1$) αλλά και η (πεπλεγμένη) μέθοδος του μέσου (3.2.15) ($\lambda=1/2$, $p=2$) είναι B-ευσταθείς. Οι διαγώνια πεπλεγμένες μέθοδοι (3.2.18) με $q=2$ είναι B-ευσταθείς γιά $\lambda\ge 1/4$. (Συνεπώς η μέθοδος με $\lambda=(1+3^{-1/2})/2$ είναι B-ευσταθής, και έχει τάξη $p=3$). Η διαγώνια πεπλεγμένη μέθοδος με $q=3$ (3.2.51) είναι B-ευσταθής στην ενδιαφέρουσα περίπτωση που το β είναι η μεγαλύτερη ρίζα του πολυωνύμου $\beta^3-3\beta^2/2+\beta/2-1/24=0$, δηλ. όταν $\beta=2\cos(\pi/18)/\sqrt{3}$, οπότε $p=4$. 'Ολες οι μέθοδοι Gauss-Legendre με q σημεία (τάξης $p=2q$) είναι B-ευσταθείς! (Μάλιστα γιά αυτές τις μεθόδους $M=0$.) Συνεπώς όλες αυτές οι μέθοδοι είναι κατάλληλες γιά μη γραμμικά άκαμπτα συστήματα όπως τα (40)-(41).

Απ' την άλλη μεριά η A-ευσταθής μέθοδος του τραπεζίου, που ως γνωστόν δίνεται από το μητρώο (βλ. 3.2.16))

$$\begin{array}{cc|c} 0 & 0 & 0 \\ 1/2 & 1/2 & 1 \\ \hline 1/2 & 1/2 & \end{array} ,$$

δεν είναι B-ευσταθής' πράγματι έχουμε

$$M = 1/4\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$$

γιά τον οποίο φυσικά $\exists\, x \neq 0$ τέτοιο ώστε $x^T M x < 0$. Πολλά αριθμητικά πειράματα γιά κατάλληλα μη γραμμικά άκαμπτα προβλήματα έχουν όντως δείξει ότι η μέθοδος του τραπεζίου πάσχει από συσσώρευση σφαλμάτων στρογγύλευσης γιά μεγάλο n.

Παρατηρήσεις

1. Γιά να βρούμε την περιοχή απόλυτης ευστάθειας μιάς πολυβηματικής μεθόδου στο μιγαδικό επίπεδο, δηλ. γιά να προσδιορίσουμε την περιοχή $S \subset \{z : Re z < 0\}$ γιά την οποία αν $z = h\lambda \in S$ τότε οι ρίζες ζ_i του πολυωνύμου

$$\pi(\zeta, z) \equiv \rho(\zeta) - z\,\sigma(\zeta)$$

ικανοποιούν $|\zeta_i| < 1$, μπορούμε να χρησιμοποιήσουμε τα κριτήρια του Schur ή των Routh-Hurwicz (Παρατήρηση 3.3.1). Στην πράξη χρησιμοποιείται και η εξής μέθοδος: 'Εστω ∂S το σύνορο της S. Επειδή οι ρίζες ζ_i είναι συνεχείς συναρτήσεις της παραμέτρου $z = h\lambda$, και $z \in S \Leftrightarrow |\zeta_i| < 1$, τότε $z \in \partial S$ όταν κάποια ρίζα ζ_i βρεθεί στην μοναδιαία περιφέρεια, δηλ. είναι της μορφής $e^{i\theta}$ γιά κάποιο $0\leq\theta<2\pi$. Συνεπώς το σύνορο ∂S δίνεται από την καμπύλη $z(\theta)= p(e^{i\theta})/\sigma(e^{i\theta})$, $0\leq\theta<2\pi$ την οποία μπορούμε να σχεδιάσουμε στο ημιεπίπεδο $\{z:\mathrm{Re}z<0\}$. Το διάστημα απόλυτης ευστάθειας θα ορισθεί από την τομή της $z(\theta)$ με τον (πραγματικό) αρνητικό ημιάξονα.

2. Ξέρουμε ότι το πολυώνυμο Taylor βαθμού $\leq m$ μιάς συνάρτησης $f(z)$ αναλυτικής σε μία περιοχή του μηδενός είναι το (μοναδικό) πολυώνυμο βαθμού $\leq m$ που ικανοποιεί την συνθήκη $|f(z)-p(z)|=O(|z|^{\nu})$, καθώς $|z|\to 0$, με ν όσο το δυνατόν μεγαλύτερο γιά αυθαίρετη f (το ν που προκύπτει είναι $\nu=m+1$). Γενικεύοντας, αναζητούμε ρητή προσέγγιση της f της μορφής

$$r(z) = p(z)/q(z),$$

αναλυτική σε μία περιοχή του μηδενός, με αριθμητή $p(z)$ βαθμού $\leq m$ και παρονομαστή $q(z)$, βαθμού $\leq n$, που, γιά δεδομένα m,n, να ικανοποιεί

$$|f(z)-r(z)| = O(|z|^{\nu}),$$

με ν όσο το δυνατόν μεγαλύτερο γιά γενική $f(z)$. Εύκολα βλέπουμε ότι υπάρχει μοναδική τέτοια συνάρτηση $r(z)$, η λεγόμενη (m,n) προσέγγιση Padé της f γιά την οποία το άριστο γενικό ν είναι $\nu=n+m+1$. Οι συντελεστές p_i,q_i των πολυωνύμων p και q ($p(z) = \sum_{i=0}^{m} p_i z^i$, $q(z) = \sum_{i=0}^{n} q_i z^i$) μπορούν να βρεθούν αναπτύσσοντας κατά Taylor την συνάρτηση $f(z)q(z)-p(z)$ γύρω απ' το μηδέν και ζητώντας να μηδενίζονται όσο περισσότεροι όροι της σειράς είναι δυνατόν.

Γιά την συνάρτηση $f(z)=e^z$ είναι γνωστό (βλ. π.χ. [1.8]) ότι οι (m,n) προσεγγίσεις Padé δίνονται από τους τύπους

$$r_{m,n}(z) = p(z)/q(z) = \sum_{k=0}^{m} p_k z^k \Big/ \sum_{k=0}^{n} q_k z^k,$$

όπου

$$p_k = p_k(m,n) = (m+n-k)!m!/(m+n)!k!(m-k)!$$

και

$$q_k = q_k(m,n) = (-1)^k(m+n-k)!n!/(m+n)!k!(n-k)!$$

και διατάσσονται συνήθως ως στοιχεία ενός (άπειρου) πίνακα, του λεγόμενου <u>πίνακα Padé</u> γιά την e^z:

$m \backslash n$	0	1	2	...
0	1	$\dfrac{1}{1-z}$	$\dfrac{1}{1-z+z^2/2}$	...
1	$1+z$	$\dfrac{1+z/2}{1-z/2}$	$\dfrac{1+z/3}{1-2z/3+z^2/6}$	...
2	$1+z+z^2/2$	$\dfrac{1+\frac{2}{3}z+\frac{z^2}{6}}{1-z/3}$	$\dfrac{1+z/2+z^2/12}{1-z/2+z^2/12}$	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	

Πολλές ρητές προσεγγίσεις του εκθετικού που προέρχονται από μεθόδους RK είναι στοιχεία του πίνακα Padé. Π.χ. οι μέθοδοι (με q σημεία (τάξης p=2q) είναι (για q≥1) τα στοιχεία) Gauss-Legendre (q,q) της διαγωνίου του πίνακα Padé. Η πεπλεγμένη μέθοδος του Euler δίνει $r(z)=1/(1-z)$ που είναι στοιχείο της $1^{ης}$ <u>υπερδιαγωνίου</u> του πίνακα Padé, τα στοιχεία της οποίας δίνουν ρητές προσεγγίσεις που αντιστοιχούν σε A-ευσταθείς μεθόδους. Γενικά είναι γνωστό (βλ. το άρθρο των Wanner, Hairer και Nørsett που αναφέραμε προηγουμένως) ότι γιά τα στοιχεία του πίνακα Padé

$$|r_{m,n}(z)|<1 \text{ γιά } \mathrm{Re}\,z<0 \iff n-2 \le m \le n,$$

δηλ. ότι A-ευσταθείς προσεγγίσεις δίνουν μόνο τα στοιχεία της διαγωνίου και των δύο πρώτων υπερδιαγωνίων του πίνακα.

3. Μεταξύ των A-ευσταθών μεθόδων RK, δηλ. των μεθόδων που ικανοποιούν την σχέση

(51) $|r(z)|<1$ για κάθε $z\in\mathbb{C}$ με $\mathrm{Re}z<0$,

διακρίνουμε υπο-κατηγορίες ανάλογα με την συμπεριφορά του $|r(z)|$ καθώς $|z|\to\infty$, $\mathrm{Re}z<0$, δηλ. της τιμής $|r(\infty)|$. Π.χ. για την μέθοδο του τραπεζίου ισχύει η (51) και η $|r(\infty)|=1$. Αυτό μας κάνει να περιμένουμε ότι η μέθοδος του τραπεζίου δεν θα αποσβένει (γιά οποιοδήποτε $h>0$) αρκετά καλά συνιστώσες της λύσης ενός άκαμπτου συστήματος που αντιστοιχούν σε πάρα πολύ μεγάλα $|\mathrm{Re}\lambda_i|$ - και όντως αυτό φαίνεται στην πράξη -. Αντίθετα, η μέθοδος του Euler (πεπλεγμένη), με $r(z)=1/(1-z)$, ικανοποιεί την (51) αλλά και την

(52) $|r(\infty)|=0$,

που πραγματικά μιμείται την ανάλογη ιδιότητα του εκθετικου $|e^z|=e^{\mathrm{Re}z}\to 0$ όταν $\mathrm{Re}z\to-\infty$. Οι A-ευσταθείς μέθοδοι για τις οποίες ισχύει η (52) (λέγονται και "L-ευσταθείς"), είναι ιδιαίτερα κατάλληλες λοιπόν για την αριθμητική λύση πολύ ακάμπτων συστημάτων. Κάπου ενδιάμεσα βρίσκονται A-ευσταθείς μέθοδοι για τις οποίες ισχύει

(53) $|r(\infty)|<1$,

όπως π.χ. η μέθοδος (3.12.18) με $\lambda=(1+3^{-1/2})/2$ με $q=2$, $p=3$. Τέτοιες μέθοδοι λέγονται και "ισχυρά A-ευσταθείς". (Ανάλογα, στον πραγματικό άξονα ορίζουμε L_0 - και ισχυρά A_0-ευσταθείς μεθόδους). Για ισχυρά A-ευσταθείς μεθόδους μπορούμε να βελτιώσουμε την εκτίμηση ευστάθειας (30): π.χ. με $e^n=0$ ισχύει η $\|y^n-z^n\|\le e^{-|\mathrm{Re}\lambda_s|t}\|y^0-z^0\|$, όπου λ_s η ιδιοτιμή του M με το μεγαλύτερο πραγματικό μέρος $\mathrm{Re}\lambda_s$.

4. Προγράμματα γενικής χρήσης για την αριθμητική λύση ακάμπτων συστημάτων χρησιμοποιούν συνήθως μιά <u>οικογένεια</u> μεθόδων (όπως π.χ. τις A(ϑ)-ευσταθείς μεθόδους οπισθοδρομικών διαφορών) μεταβλητής τάξης καθώς και μεταβλητό βήμα που ελέγχεται με κάποια στρατηγική εκτίμησης του τοπικού σφάλματος. Προγράμματα τέτοια όπως η "μέθοδος του Gear" ή το πρόγραμμα EPISODE των Byrne et al. υπάρχουν σε πολλές βιβλιοθήκες αλγορίθμων. Δέν μπορούμε όμως να πούμε ακόμα ότι γράφτηκε <u>το</u> πρόγραμμα που θα ολοκληρώνει εγγυημένα οποιοδήποτε άκαμπτο πρόβλημα, έστω και "μεσαίου" μεγέθους.

5. Για ανάλογη με την B-ευστάθεια θεωρία για πολυβηματικές μεθόδους βλ. π.χ. την εργασία του Dahlquist στα πρακτικά συνεδρίου Springer LNM v.506 (1976), καθώς και εργασίες των Butcher (SIAM JNA 18 (1981), 37-44), Nevanlinna-Liniger, π.χ. BIT 19 (1979), 53-72) κ.α. Η σημασία της ανάλυσης των μεθόδων για μη γραμμικές εξισώσεις είχε αναγνωρισθεί από τόν Dahlquist ήδη στην σημαντική εργασία του του 1963, όπου εισήγαγε την A-ευστάθεια.

Ασκήσεις 3.4

1. Βρείτε τα διαστήματα απόλυτης ευστάθειας για άμεσες μεθόδους RK τάξης p=2, 3 και 4.

2. Για την λύση της απλής Δ.Ε. $y'=f(t,y)$ θεωρείστε την πεπλεγμένη μέθοδο (που δεν ανήκει στην κατηγορία των RK ή των γραμμικών πολυβηματικών μεθόδων)

$$y^{n+1}= y^n+\frac{h}{2}(f^n+f^{n+1}) + \frac{h^2}{12}\left((D_t f)^n + (D_t f)^{n+1}\right),$$

όπου $D_t f = \partial_t f+(\partial_y f)f$. Ορίστε την τάξη ακρίβειας της μεθόδου και υπολογίστε την. Βρήτε την περιοχή απόλυτης ευστάθειάς της. (Μέθοδοι όπως η παραπάνω ανήκουν στην κατηγορία των "μεθόδων Obrechkoff" ή των διβηματικών "πολυπαραγωγικών" μεθόδων).

3. Βρήτε την περιοχή απόλυτης ευστάθειας της μεθόδου Adams-Moulton

$$y^{n+3}-y^{n+2} = h(9f^{n+3}+19f^{n+2}-5f^{n+1}+f^{n})/24$$

(Δείξτε ότι η περιοχή είναι συμμετρική ως πρός τον πραγματικό άξονα, ότι το διάστημα απόλυτης ευστάθειας είναι το $(-3,0)$ και σχεδιάστε την περιοχή στο μιγαδικό επίπεδο. Τι συμβαίνει κοντά στον φανταστικό άξονα;)

4. Δείξτε γιά την μέθοδο οπισθοδρομικών διαφορών με k βήματα (3.3.6) ότι

(α) Γιά $k=2$ είναι A-ευσταθής.

(β) Γιά $k=3$ είναι A_0-ευσταθής αλλά όχι A-ευσταθής.

5(α). Θεωρείστε τις διαγώνια πεπλεγμένες μεθόδους (3.2.18) με παράμετρο $\lambda\in R$. Γιά ποιές τιμές του λ είναι οι μέθοδοι A-ευσταθείς; Εφαρμόστε τις στο γραμμικό σύστημα (27). Τι παρατηρείτε σχετικά με τα γραμμικά συστήματα που πρέπει να λυθούν σε κάθε στάδιο και σε κάθε βήμα;

(β) Θεωρείστε την μέθοδο του Calahan (3.2.50). Ποιά είναι η περιοχή απόλυτης ευστάθειάς της;

6. Βρήτε την περιοχή της απόλυτης ευστάθειας των μεθόδων πρόβλεψης - διόρθωσης (i) και (ii) της Παρατήρησης 3.3.3 γιά m=1 και m=2 διορθώσεις.

7. Ως πρότυπο γιά συστήματα γιά τα οποία οι ιδιοτιμές του πίνακα M είναι <u>φανταστικές</u>, (τέτοια συστήματα προέρχονται π.χ. από υπερβολικές Μ.Δ.Ε.) θεωρούμε το "συντηρητικό" πρόβλημα

$$(*)\quad \begin{cases} y' = i\lambda y, \ t\geq 0, \ \lambda\in R, \\ y(0) = 1, \end{cases}$$

του οποίου η λύση $y(t)=e^{i\lambda t}$ ικανοποιεί $|y(t)|=1$, $t\geq 0$. Λέμε ότι μία μέθοδος είναι <u>I-ευσταθής</u> αν, όταν εφαρμοσθεί στο (*), δίνει, γιά κάθε $h>0$, ακολουθία $\{y^n\}$ τέτοια ώστε $|y^n|\leq 1$, $n\geq 0$. Γιά μεθόδους RK αυτό σημαίνει ότι $|r(iy)|\leq 1$ $\forall y\in R$. Συνεπώς, κάθε A-ευσταθής μέθοδος είναι <u>I-ευσταθής</u>. Από τις I-ευσταθείς μεθόδους ξεχωρίζουμε τις λεγόμενες <u>συντηρητικές</u> μεθόδους, γιά τις οποίες $|y^n|=1$, $n\geq 0$, δηλ. γιά τις οποίες ισχύει ακριβώς το διακριτό ανάλογο της $|y(t)|=1$, $t\geq 0$.

(α). Δείξτε ότι όλες οι μέθοδοι RK που έχουν ρητή συνάρτηση $r(z)$ που δίνεται από οποιοδήποτε στοιχείο της διαγωνίου του πίνακα Padé γιά την e^z είναι I-ευσταθείς και μάλιστα συντηρητικές.

(β) Χρησιμοποιώντας το (α) και την Πρόταση 1 δείξτε ότι η μέθοδος Gauss-Legendre με 2 σημεία, (23), είναι A-ευσταθής.

(γ) Συγκρίνετε τις δύο μεθόδους Euler και την μέθοδο του τραπεζίου ως πρός την καταλληλότητά τους γιά την αριθμητική ολοκλήρωση του (*) γιά μεγάλο $|\lambda|$.

8.(α) Να αποδειχθεί ότι οι διαγώνια πεπλεγμένες μέθοδοι (3.2.18) είναι B-ευσταθείς γιά $\lambda\geq 1/4$.

(β) Να αποδειχθεί ότι η μέθοδος Gauss-Legendre με q=2 σημεία (3.2.19) είναι B-ευσταθής.

(γ) Να αποδειχθεί ότι η μέθοδος RK

$$\begin{array}{cc|c} 1/8 & 1/8 & 1/4 \\ 3/8 & 3/8 & 3/4 \\ \hline 1/2 & 1/2 & \end{array}$$

είναι A-ευσταθής αλλά όχι B-ευσταθής.

9. (Dahlquist) (α) Δείξτε ότι μία k-βηματική μέθοδος είναι A-ευσταθής αν και μόνο αν η συνάρτηση $\rho(z)/\sigma(z)$ είναι αναλυτική και έχει μη αρνητικό πραγματικό μέρος γιά $|z|>1$.

(β) Χρησιμοποιώντας το μέρος (α) δείξτε ότι μία άμεση k-βηματική μέθοδος δεν μπορεί να είναι A-ευσταθής.

4. ΠΑΡΕΜΒΟΛΗ ΚΑΙ ΠΡΟΣΕΓΓΙΣΗ

4.1 ΠΑΡΕΜΒΟΛΗ ΜΕ ΠΟΛΥΩΝΥΜΟ LAGRANGE

Στο κεφάλαιο αυτό θα κάνουμε μία σύντομη εισαγωγή στην προσέγγιση πραγματικών συνεχών συναρτήσεων μιάς μεταβλητής με παρεμβολή με πολυώνυμα ή τμηματικά πολυωνυμικές συναρτήσεις. Ο λόγος που τα πολυώνυμα χρησιμοποιούνται κατ' εξοχήν γιά την προσέγγιση συναρτήσεων είναι βέβαια το γεγονός ότι μπορούν να υπολογισθούν (αλλά και να παραγωγισθούν ή να ολοκληρωθούν) εύκολα με ένα πεπερασμένο πλήθος προσθαφαιρέσεων και πολλαπλασιασμών αλλά και το γεγονός ότι έχουν καλές προσεγγιστικές ιδιότητες, όπως υποδηλώνει το θεώρημα του Weierstrass. Θα συμβολίζουμε με P_m τον διανυσματικό χώρο όλων των πραγματικών πολυωνύμων βαθμού $\leq m$.

Αρχίζουμε εξετάζοντας σύντομα στην παράγραφο αυτή διάφορα ερωτήματα που σχετίζονται με την παρεμβολή Lagrange. Το θέμα μας είναι γνωστό από το μάθημα της Εισαγωγής στην Αριθμητική Ανάλυση' γιά λεπτομέρειες ή αποδείξεις που θα παραλειφθούν βλ. π.χ. [5.2].

Έστω $\tau \equiv \{x_i\}_{i=1}^{n}$ μιά ακολουθία $n \geq 1$ διακριτών σημείων. Το πολυώνυμο βαθμού n-1

$$(1)\ L_i(x) = \prod_{i\neq j=1}^{n} (x-x_j)/(x_i-x_j),$$

λέγεται το i-στό πολυώνυμο Lagrange γιά την τ. Προφανώς

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

Συνεπώς γιά μιά οποιαδήποτε συνάρτηση f(x) (ή γενικά γιά οποιαδήποτε δεδομένα $f(x_i)$), το πολυώνυμο βαθμού $\leq n-1$

$$(2)\ p(x) = \sum_{i=1}^{n} f(x_i)\, L_i(x)$$

ικανοποιεί $p(x_i)=f(x_i)$, $1\leq i\leq n$, δηλ. παρεμβάλλεται στις τιμές της συνάρτησης (των δεδομένων $f(x_i)$) στά σημεία της τ. (Λέγεται και "παρεμβάλλει την f στά x_i"). Προφανώς υπάρχει μόνο ένα τέτοιο $p\in P_{n-1}$ γιατί αν υπήρχε και άλλο $q\in P_{n-1}$ με την ίδια ιδιότητα, η διαφορά τους $r=p-q$ θα ήταν ένα στοιχείο του P_{n-1} με n διακριτές ρίζες, δηλ. θα ήταν το μηδενικό πολυώνυμο. Θα ονομάζουμε λοιπόν το πολυώνυμο $p(x)$, (2), <u>πολυώνυμο παρεμβολής Lagrange</u> γιά την συνάρτηση f στά σημεία της τ.

Η παράσταση (2) του πολυωνύμου παρεμβολής δεν είναι η πιό κατάλληλη γιά τίς εφαρμογές. Ίσως η πιό εύχρηστη είναι η παράστασή του στη <u>μορφή Newton</u>, που μπορεί να οριστεί με την βοήθεια <u>διαιρεμένων διαφορών</u>.

Η <u>k-στή διαιρεμένη διαφορά μιάς συνάρτησης f στά σημεία $x_i,\ldots,x_{i+k}$</u> είναι ο συντελεστής του μονωνύμου μεγίστου βαθμού (δηλ. του x^k) του πολυωνύμου βαθμού $\leq k$ που συμφωνεί με τις τιμές της f στα σημεία $x_i,\ldots,x_{i+k}$. Την συμβολίζουμε με

$$f[x_i,\ldots,x_{i+k}].$$

Έχουμε δηλ.

$$f[x_1] = f(x_1),$$

$$f[x_1,x_2]=(f(x_2)-f(x_1))/(x_2-x_1),\ \text{γιά } x_1\neq x_2$$

και γενικά (βλ. Ασκ. 1α)

$$(3)\quad f[x_i,\ldots,x_{i+k}]=(f[x_{i+1},\ldots,x_{i+k}]-f[x_i,\ldots,x_{i+k-1}])/(x_{i+k}-x_i),$$

ή χρησιμοποιώντας την μορφή (2) του πολυωνύμου παρεμβολής,

$$(3')\quad f[x_i,\ldots,x_{i+k}] = \sum_{j=i}^{i+k} f(x_j) \Big/ \Big(\prod_{j\neq\mu=i}^{i+k} (x_j-x_\mu)\Big).$$

Ο ορισμός της k-στής διαιρεμένης διαφοράς μας δίνει το εξής αποτέλεσμα: Αν τα πολυώνυμα $p_{i-1}\in P_{i-1}$ συμφωνούν με την f στά σημεία $x_1,\dots,x_i$ γιά i=k και k+1, αντίστοιχα, τότε

$$(4)\quad p_k(x) = p_{k-1}(x)+(x-x_1)\dots(x-x_k)f[x_1,\dots,x_{k+1}].$$

Πράγματι, το πολυώνυμο p_k-p_{k-1} είναι βαθμού $\leq k$ και μηδενίζεται στά σημεία $x_1,\dots,x_k$. Εξ άλλου το μονώνυμό του μεγίστου βαθμού (δηλ. το αντίστοιχο μονώνυμο του p_k) έχει εξ ορισμού συντελεστή $f[x_1,\dots,x_{k+1}]$. Άρα πρέπει να είναι της μορφής $p_k(x)-p_{k-1}(x)=f[x_1,\dots,x_{k+1}](x-x_1)\dots(x-x_k)$· δηλ. ισχύει η (4).

Η σχέση (4) μας δίνει την δυνατότητα να κατασκευάσουμε το πολυώνυμο παρεμβολής $p(x)\in P_{n-1}$ στα σημεία $x_1,\dots,x_{n-1}$ βήμα πρός βήμα, προσθέτοντας κάθε φορά ένα νέο σημείο παρεμβολής, πράγμα πολύ χρήσιμο στις εφαρμογές. Με τον παραπάνω συμβολισμό των p_i, $0\leq i\leq n-1$, έχουμε από την (4)

$$\begin{aligned} p(x)=p_{n-1}(x) &= p_0(x)+(p_1(x)-p_0(x))+(p_2(x)-p_i(x))+\\ &\quad+\dots+(p_{n-1}(x)-p_{n-2}(x))=\\ (4')\qquad &= f[x_1]+(x-x_1)f[x_1,x_2]+(x-x_1)(x-x_2)f[x_1,x_2,x_3]\\ &\quad+\dots+(x-x_1)\dots(x-x_{n-1})f[x_1,\dots,x_n], \end{aligned}$$

η οποία είναι η <u>μορφή Newton</u> του πολυωνύμου παρεμβολής, πολύ χρήσιμη στην θεωρία και στις εφαρμογές. (Ο υπολογισμός του p(x) από την (41) γίνεται π.χ. εύκολα με τον κανόνα του Horner).

Μία σημαντική ιδιότητα των διαιρεμένων διαφορών (Άσκηση 1β) είναι η εξής γενίκευση του θεωρήματος μέσης τιμής: Αν η συνάρτηση f, ορισμένη στο [α,b], είναι k φορές παραγωγίσιμη στο (α,b) και τα $x_1,\dots,x_{k+1}$ είναι διακριτά σημεία του [α,b], τότε υπάρχει $\xi\in(\alpha,b)$, τέτοιο ώστε

(5) $f[x_1,\ldots,x_{k+1}] = f^{(k)}(\xi)/k!$

Η (5) δίνει τώρα την γνωστή μας έκφραση του σφάλματος του πολυωνύμου παρεμβολής. 'Εστω $p=p_{n-1}$ το πολυώνυμο παρεμβολής Lagrange γιά την f, βαθμού $\leq n-1$ στά διακριτά σημεία της $\tau=\{x_i\}_{i=1}^{n}\subset[a,b]$ και έστω $e_{n-1}(x)=f(x)-p_{n-1}(x)$ το σφάλμα της παρεμβολής. 'Εστω σημείο $\bar{x}\in[a,b]$, διάφορο των x_i, $1\leq i\leq n$. Αν p_n είναι το πολυώνυμο που παρεμβάλλεται στις τιμές της f στά n+1 σημεία x_i, $1\leq i\leq n$, και $\bar{x}$, τότε $p_n(\bar{x})=f(\bar{x})$ και λόγω της (4)

$$p_n(x) = p_{n-1}(x)+(x-x_1)\ldots(x-x_n)f[x_1,\ldots,x_n,\bar{x}].$$

'Αρα

$$f(\bar{x})=p_n(\bar{x})=p_{n-1}(\bar{x})+f[x_1,\ldots,x_n,\bar{x}]\ \prod_{i=1}^{n}(\bar{x}-x_i).$$

Δηλ. γιά κάθε $\bar{x}\notin\tau$

(6) $$e(\bar{x}) \equiv f(\bar{x}) - p_{n-1}(\bar{x}) = f[x_1,\ldots,x_n,\bar{x}]\ \prod_{i=1}^{n}(\bar{x}-x_i).$$

Οι (5) και (6) δίνουν λοιπόν το

ΘΕΩΡΗΜΑ 1. 'Εστω ότι η $f:[a,b]\to\mathbb{R}^1$ είναι n φορές παραγωγίσιμη στο (a,b). Αν $p_{n-1}(x)$ είναι το πολυώνυμο παρεμβολής Lagrange στά n διακριτά σημεία $x_1,\ldots,x_n\in[a,b]$, τότε γιά κάθε $\bar{x}\in[a,b]$ υπάρχει $\xi=\xi(\bar{x})\in(a,b)$ τέτοιο ώστε

(7) $$e_{n-1}(\bar{x}) \equiv f(\bar{x})-p_{n-1}(\bar{x}) = \left(\prod_{i=1}^{n}(\bar{x}-x_i)\right) f^{(n)}(\xi)/n!$$

Θα εξετάσουμε τώρα ορισμένες προβληματικές πλευρές της προσέγγισης μιάς συνάρτησης από το πολυώνυμο παρεμβολής της. Είναι γνωστό π.χ. ότι γιά <u>ισαπέχοντα σημεία</u>, δηλ. ακολουθίες τ με $x_i - x_{i-1}=$ σταθ., μπορούμε, ακόμα και γιά ομαλότατες συναρτήσεις, να οδηγηθούμε σε πολυώνυμα παρεμβολής $p_{n-1}(x)$ των οποίων το μέγιστο σφάλμα αυξάνει απεριόριστα καθώς αυξάνει το n. Το κλασσικό παράδειγμα είναι το λεγόμενο <u>παράδειγμα του Runge</u>, το οποίο αναφέρεται στην συνάρτηση

(8) $f(x)=1 / (1+25x^2), \ x \in [-1,1]$,

($f \in C^{\infty}(R^1)$!) και στην παρεμβολή της με πολυώνυμα Lagrange στις ακολουθίες τ_n, $n\geq 2$, ισαπεχόντων σημείων στο $[-1,1]$ όπου $\tau_n=\{x_k^n\}$, $k=1,2,3,\ldots n$, με $x_1^n=-1$ και $x_k^n = x_{k-1}^n + h_n$, $k= 2,3,\ldots,n$, $h_n=2/(n-1)$.

Κατ'αρχήν μία προκαταρκτική παρατήρηση. 'Εστω $g(x)=1/(ax+b)$. Τότε γιά y_i, $1\leq i\leq m$, διακριτά σημεία έχουμε

(9) $$g[y_1,\ldots,y_m] = (-a)^{m-1} \prod_{i=1}^{m} (ay_i+b)^{-1}.$$

θα αποδείξουμε την (9) με επαγωγή. Για m=1 ισχύει προφανώς. Παρατηρούμε επί τη ευκαιρία ότι η γενική διαιρεμένη διαφορά $g[x_i,\ldots,x_{i+k}]$ είναι <u>συμμετρική συνάρτηση</u> των $x_i,\ldots,x_{i+k}$, δηλ. εξαρτάται μόνο από τους αριθμούς $x_i,\ldots,x_{i+k}$ και όχι από την σειρά με την οποία εμφανίζονται μέσα στις αγκύλες. Αυτό είναι προφανές από τον ορισμό της επειδή το πολυώνυμο παρεμβολής στα σημεία $x_i,\ldots,x_{i+k}$ εξαρτάται μόνο από τα σημεία και όχι από την σειρά τους. Εστω λοιπόν τώρα ότι η (9) ισχύει για m-1. Εχουμε, από την (3) και την συμμετρία της g ως πρός τις μεταβλητές y_i,

$$g[y_1,\ldots,y_m]=g[y_{m-1},y_1,y_2,\ldots,y_{m-2},y_m]=$$

$$= (g[y_1,\ldots,y_{m-2},y_m] - g[y_{m-1},y_1,\ldots,y_{m-2}]) / (y_m-y_{m-1})$$

$$= (g[y_1,\ldots,y_{m-2},y_m] - g[y_1,\ldots,y_{m-1}]) / (y_m-y_{m-1})$$

$$= \{(-a)^{m-2}(\prod_{i=1}^{m-2}(ay_i+b)^{-1})(ay_m+b)^{-1}-(-a)^{m-2}(\prod_{i=1}^{m-1}(ay_i+b)^{-1})\}/(y_m-y_{m-1})$$

$$= (-a)^{m-2}\prod_{i=1}^{m-2}(ay_i+b)^{-1}[1/((ay_m+b)(y_m-y_{m-1}))-1/((ay_{m-1}+b)(y_m-y_{m-1}))]$$

$$= (-a)^{m-1}\prod_{i=1}^{m}(ay_i+b)^{-1},$$

δηλ. ότι η (9) ισχύει και γιά m. Εφαρμόζοντας την (9) γιά τα διακριτά σημεία $x_1^n,\dots,x_n^n,x$ και παραλείποντας γιά ευκολία τους άνω δείκτες (n) στά σημεία x_k^n παίρνουμε γιά $g(x)=(ax+b)^{-1}$

$$(9')\quad g[x_1,\dots,x_n,x] = ((-a)^n/(ax+b))\prod_{i=1}^{n}(ax_i+b)^{-1}.$$

Γράφοντας τώρα την συνάρτηση $f(x) = (1+25x^2)^{-1}$ ως

$$(10)\quad f(x) = (i/2)\{(5x+i)^{-1} - (5x-i)^{-1}\}$$

και παρατηρώντας (π.χ. από την (3')) ότι γενικά η διαιρεμένη διαφορά $g[x_i,\dots,x_{i+k}]$ είναι γραμμική ως πρός g, δηλ. ότι

$$(\lambda g_1+\mu g_2)[x_i,\dots,x_{i+k}]=\lambda g_1[x_i,\dots,x_{i+k}]+\mu g_2[x_i,\dots,x_{i+k}],$$

έχουμε, από τις (9'), (10) ότι

$$(11)\quad f[x_1,\dots,x_n,x] =$$

$$=(-5)^n(i/2)\{(5x+i)^{-1}\prod_{k=1}^{n}(5x_k+i)^{-1}-(5x-i)^{-1}\prod_{k=1}^{n}(5x_k-i)^{-1}\}.$$

Τώρα, $x_{n-k+1} = -1+(n-k)h_n = -1+(n-1)h_n-(k-1)h_n=1-(k-1)h_n=-x_k$

Συνεπώς, <u>γιά n άρτιο</u>, έχουμε

$$\prod_{k=1}^{n} (5x_k \pm i)^{-1} = \prod_{k=1}^{n/2} (5x_k \pm i)^{-1}(-5x_k \pm i)^{-1} = -\prod_{k=1}^{n/2} (25x_k^2+1)^{-1}$$

Υποθέτοντας από δω και και πέρα ότι ο n είναι άρτιος έχουμε από την (11) ότι

$$f[x_1,\ldots,x_n,x]=(-5)^n (i/2) \{\prod_{k=1}^{n/2} (25x_k^2+1)^{-1} ((5x+i)^{-1}-(5x-i)^{-1})$$

$$(12) \qquad = 5^n(-1)^{n/2}(25x^2+1)^{-1} \prod_{k=1}^{n/2} (25x_k^2+1)^{-1}.$$

Από την (6) μπορούμε να γράψουμε το σφάλμα $e_{n-1}(x)=f(x)-p_{n-1}(x)$ της παρεμβολής Lagrange στά σημεία $x_1,\ldots,x_n$ ως

$$e_{n-1}(x)= (x-x_1)\ldots(x-x_n)\ f[x_1,\ldots,x_n,x].$$

Συνεπώς η (12) δίνει

$$(13)\quad |(1+25x^2)\ e_{n-1}(x)| = |\prod_{i=1}^{n} (x-x_i)|\ (\prod_{k=1}^{n/2} 25/(25x_k^2+1)).$$

Θέτουμε τώρα

$$(14)\quad r_n(x) \equiv 2(\log |(1+25x^2)\ e_{n-1}(x)|)/(n-1).$$

Λόγω της (13), γιά $x \neq x_k = x_k^n$

$$r_n(x)= 2(n-1)^{-1} \sum_{i=1}^{n} \log |x-x_i| - 2(n-1)^{-1} \sum_{k=1}^{n/2} \log (x_k^2+(1/25)).$$

Άρα

$$(15)\quad \lim_{\substack{n\to\infty \\ n\ \text{άρτιος}}} r_n(x) = \int_{-1}^{x} \log (x-\xi)d\xi + \int_{x}^{1} \log(\xi-x)d\xi$$

$$- \int_{-1}^{0} \log (\xi^2+(1/25))d\xi \equiv r(x).$$

Η $r(x)$ είναι συνεχής στο $[-1,1]$ και $r(1)>0$ ('Άσκηση 2). Συνεπώς υπάρχει ανοιχτό διάστημα στο $[-1,1]$ στο οποίο η $r(t)$ είναι θετική. 'Άρα υπάρχει $x_0 \neq x_k^n$ στο $[-1,1]$ γιά κάθε n και $k \leq n$ γιά το οποίο $r(x_0)>0$. 'Άρα από την (14)

$$\lim_{\substack{n\to\infty \\ n \text{ άρτιος}}} (\log |(1+25x_0^2)\, e_{n-1}(x_0)|) = +\infty,$$

δηλ.

$$\lim_{\substack{n\to\infty \\ n \text{ άρτιος}}} \log|e_{n-1}(x_0)| = +\infty,$$

δηλ.

$$\lim_{\substack{n\to\infty \\ n \text{ άρτιος}}} |e_{n-1}(x_0)| = +\infty,$$

που δείχνει ότι υπάρχει ακολουθία ομοιόμορφων διαμερισμών του $[-1,1]$ γιά την οποία η αντίστοιχη ακολουθία των μεγίστων σφαλμάτων του πολυωνύμου παρεμβολής τείνει στο άπειρο.

Ας κοιτάξουμε το πρόβλημα λίγο γενικότερα. 'Εστω $P_n f \in P_{n-1}$ το πολυώνυμο παρεμβολής μιάς συνεχούς συνάρτησης f στά (διακριτά) σημεία x_i, $1 \leq i \leq n$, της πεπερασμένης ακολουθίας τ, η οποία υποθέτουμε ότι περιέχεται σ' ένα διάστημα $[a,b]$. Γιά $f \in C[a,b]$ θεωρούμε την "maximum" νόρμα της ομοιόμορφης σύγκλισης

$$(16)\quad \|f\|_\infty = \max_{a \leq x \leq b} |f(x)|.$$

(Θεωρούμε γνωστά από την Συναρτησιακή Ανάλυση τα βασικά περί χώρων συναρτήσεων με νόρμα). Από την (2)

$$|(P_n f)(x)| \leq \sum_{i=1}^{n} |f(x_i)|\, |L_i(x)| \leq \max_i |f(x_i)| \sum_{i=1}^{n} |L_i(x)|.$$

Εισάγοντας την λεγόμενη <u>συνάρτηση του Lebesgue</u>

(17) $\lambda_n(x) = \sum_{i=1}^{n} |L_i(x)|$,

έχουμε τελικά ότι

(18) $\|P_n f\|_\infty \leq \|\lambda_n\|_\infty \|f\|_\infty$.

Δεν είναι δύσκολο να δούμε (Άσκ 3) ότι υπάρχει, γιά κάθε τ με n σημεία, συνεχής συνάρτηση $f \neq 0$ τέτοια ώστε να ισχύει η (18) ως ισότης. (Δηλ. η νόρμα του φραγμένου γραμμικού τελεστή $P_n: X \longrightarrow X$, $X=(C[a,b], \|\cdot\|_\infty)$ είναι ίση με $\|\lambda_n\|_\infty$. Είναι γνωστό εξ άλλου ότι υπάρχουν σταθερές $c_1>0$, $c_2 \in \mathbb{R}$, ανεξάρτητες των n, τ τέτοιες ώστε γιά την λεγόμενη "σταθερά Lebesgue" $\|\lambda_n\|_\infty$ να ισχύει

(19) $\|\lambda_n\|_\infty \geq c_1 \log n + c_2$.

(Βλ. π.χ. το βιβλίο του Rivlin [4.9, σελ. 90-91] γιά μία απόδειξη με $c_1=4/\pi^2$, $c_2=-1$. Είναι γνωστό (Erdös) ότι η καλύτερη σταθερά c_1 είναι $c_1=2/\pi$). Το αποτέλεσμα αυτό μας οδηγεί στο περίφημο θεώρημα του Faber, βάσει του οποίου γιά κάθε δεδομένο "τριγωνικό σύστημα παρεμβολής", δηλ. γιά κάθε ακολουθία $\tau_1, \tau_2, \ldots, \tau_n, \ldots$, $\tau_n = \{x_n^1, \ldots, x_n^n\}$, πεπερασμένων ακολουθιών σημείων παρεμβολής $x_n^i \in [a,b]$, υπάρχει συνάρτηση $f^* \in C[a,b]$ τέτοια ώστε

$$\overline{\lim}_{n \to \infty} \|P_n f^* - f^*\|_\infty = \infty,$$

βλ. π.χ. [4.9, σελ. 92-3].

Γιά ισαπέχοντα σημεία μάλιστα είναι γνωστό ότι η σταθερά Lebesgue αυξάνεται εκθετικά· υπάρχουν σταθερές $c_1, c_2 > 0$, $\alpha_1 > \alpha_2 > 1$, ανεξάρτητες των $n, h_n=(b-a)/(n-1)$, τέτοιες ώστε γιά κάθε $n>1$

$$c_2(\alpha_2)^n \leq \|\lambda_n\|_\infty \leq c_1(\alpha_1)^n.$$

Βλ. π.χ. [4.9, σελ. 99] γιά μία απόδειξη με n άρτιο, $\alpha_2=(1.5)^{1/2}$, $\alpha_1=2^{1/2}e$. Συνεπώς το παράδειγμα του Runge δεν πρέπει να μας εκπλήσσει και πολύ.

Απ' την άλλη μεριά είναι ενθαρρυντικό (βλ. Ασκ. 4) ότι <u>γιά κάθε</u> $f\in C[a,b]$ υπάρχει τριγωνικό σύστημα παρεμβολής τέτοιο ώστε η ακολουθία πολυωνύμων $P_n f$ να συγκλίνει ομοιόμορφα στην f στο [a,b] καθώς $n\to\infty$. Φυσικά, το τριγωνικό αυτό σύστημα (ακολουθία τ_n n διακριτών σημείων) μπορεί να μην είναι εύκολο να κατασκευασθεί και βέβαια θα διαφέρει από συνάρτηση σε συνάρτηση. Τίθεται το ερώτημα: υπάρχει σύστημα παρεμβολής, που να μπορεί να βρεθεί εύκολα, τέτοιο ώστε η $\|\lambda_n\|_\infty$ να αυξάνεται αργά (λογαριθμικά είναι φυσικά ο αργότερος ασυμπτωτικά δυνατός τρόπος, βλ. (19)) καθώς $n\to\infty$; Η απάντηση είναι θετική και η λύση δίνεται από τις ρίζες των (πανταχού παρόντων) πολυωνύμων Chebyshev! Μπορεί ν' αποδειχθεί, βλ. π.χ. [4.9, σελ. 93 et seq.], ότι αν τα σημεία $x_1,\dots,x_n$ της τ είναι οι ρίζες του πολυωνύμου Chebyshev $n^{ου}$ βαθμού γιά το [a,b], δηλ. αν

$$(20)\quad x_i=x_i^c \equiv \{(a+b)-(a-b)\cos((2i-1)\pi/2n)\}/2,\ 1\leq i\leq n$$

(απεικονίστε το διάστημα [-1,1] πάνω στο [a,b] και χρησιμοποιήστε ότι $T_n(z)=\cos(n\cos^{-1}z)$, $-1\leq z\leq 1$, βλ. Παρ. 1.6), τότε γιά την αντίστοιχη σταθερά του Lebesgue $\|\lambda_n\|_\infty \equiv \|\lambda_n^c\|_\infty$ έχουμε

$$(21)\quad \|\lambda_n^c\|_\infty \leq (2/\pi)\log n + 4,$$

που, σε συνδυασμό με την (19), δείχνει ότι οι ρίζες των πολυωνύμων Chebyshev δίνουν ένα πολύ καλό τριγωνικό σύστημα παρεμβολής. Το πείραμα δείχνει (βλ. παρακάτω γιατί) ότι η συνάρτηση Runge $f(x)=1/(1+25x^2)$ στο [-1,1] δίνει ακολουθία πολυωνύμων παρεμβολής στά x_i^c που συγκλίνει γιά $n\to\infty$ ομοιόμορφα στην f.

Η (21) μας δίνει επίσης ότι το πολυώνυμο παρεμβολής $P_n f$ στά σημεία Chebyshev x_i^c, $1 \leq i \leq n$ είναι "σχεδόν" (modulo λογαριθμικό παράγοντα) η καλύτερη προσέγγιση της f στον $(\mathbf{P}_{n-1}, \|\cdot\|_\infty)$. Πράγματι, έστω $p^* \in \mathbf{P}_{n-1}$ η βέλτιστη προσέγγιση της $f \in C[a,b]$ από στοιχεία του $\mathbf{P}_{n-1}$ ως πρός την νόρμα $\|\cdot\|_\infty$, δηλ. έστω ότι

$$(22) \quad \|f-p^*\|_\infty = \min_{p \in \mathbf{P}_{n-1}} \|f-p\|_\infty$$

(Γιά την ύπαρξη, μοναδικότητα, ιδιότητες και κατασκευή του p^* βλ. π.χ. [5.1]). Επειδή ο τελεστής P_n της παρεμβολής στά σημεία x_i, $1 \leq i \leq n$ μιάς οποιασδήποτε τ είναι ταυτότητα στου $\mathbf{P}_{n-1}$, έχουμε

$$f-P_n f = f-p-P_n(f-p) \quad \forall p \in \mathbf{P}_{n-1},$$

οπότε από την (18) έχουμε $\forall p \in \mathbf{P}_{n-1}$

$$\|f-P_n f\|_\infty \leq \|f-p\|_\infty + \|P_n(f-p)\|_\infty \leq \|f-p\|_\infty + \|\lambda_n\|_\infty \|f-p\|_\infty.$$

Συμπεραίνουμε, παίρνοντας $p=p^*$ στην παραπάνω ανισότητα ότι

$$(23) \quad \min_{p \in \mathbf{P}_{n-1}} \|f-p\|_\infty \leq \|f-P_n f\|_\infty \leq (1+\|\lambda_n\|_\infty) \min_{p \in \mathbf{P}_{n-1}} \|f-p\|_\infty .$$

Γιά παρεμβολή στις ρίζες των πολυωνύμων Chebyshev, $\|\lambda_n\|_\infty = \|\lambda_n^c\|_\infty$, που αυξάνεται πολύ αργά (βλ. (21)) με το n. Π.χ. γιά $n \leq 20$ είναι γνωστό ότι $1+\|\lambda_n^c\|_\infty \leq 4$. Αυτό σημαίνει, λόγω της (23), ότι η καλύτερη δυνατή προσέγγιση της f στον $(\mathbf{P}_{n-1}, \|\cdot\|_\infty)$ (γιά $n \leq 20$) δίνει σφάλμα το οποίο είναι <u>το πολύ</u> 4 φορές μεγαλύτερο από το σφάλμα του πολυωνύμου παρεμβολής στά σημεία Chebyshev x_i^c. Φυσικά το πολυώνυμο παρεμβολής είναι πολύ απλό να κατασκευασθεί σε αντίθεση με το p^*!

Η (23) μας δίνει ένα πολύ χρήσιμο άνω φράγμα του σφάλματος $\|f-P_nf\|_\infty$ της παρεμβολής συναρτήσει της σταθεράς Lebesgue (που εξαρτάται μόνο από την τ) και του σφάλματος $\min\limits_{p\in P_{n-1}} \|f-p\|_\infty = \|f-p^*\|_\infty$ της βέλτιστης προσέγγισης p^* της f στον $(P_{n-1}\ \|\cdot\|_\infty)$, (που εξαρτάται μόνο από την f, το διάστημα [a,b] και το n). Γιά το τελευταίο έχουμε πολύ ακριβείς εκτιμήσεις από την Θεωρία Προσέγγισης γιά διάφορες κλάσεις συναρτήσεων f (Θεωρήματα Jackson). Συνήθως υποθέτουμε ότι γιά κάποιον ακέραιο $r\geq 0$, $f\in C^r[a,b]$, δηλ. ότι η f έχει r συνεχείς παραγώγους στο [a,b] ($C^0[a,b] = C[a,b]$), εκφράζουμε δε τα φράγματα των σφαλμάτων συναρτήσει του <u>μέτρου συνέχειας</u> της $f^{(r)}$.

Γιά μία συνάρτηση $g\in C[a,b]$, το <u>μέτρο συνέχειάς της</u> είναι μία συνάρτηση $\omega(g;h)$, που ορίζεται γιά $h\geq 0$ από

$$(24)\quad \omega(g;h) = \max\{|g(x)-g(y)| : x,y\in[a,b],\ |x-y|\leq h\}.$$

Είναι προφανές ότι $0\leq h_1\leq h_2 \Rightarrow \omega(g;h_1) \leq \omega(g;h_2)$ και ότι γιά $h_1,h_2\geq 0$ $\omega(g;h_1+h_2) \leq \omega(g;h_1)+\omega(g;h_2)$. Είναι επίσης προφανές ότι επειδή $g\in C[a,b]$,

$$(25)\quad \lim_{h\downarrow 0} \omega(g;h) = 0.$$

Ο <u>ρυθμός</u> όμως (σάν συνάρτηση του h) με τον οποίο η $\omega(g;h)$ τείνει στο μηδέν καθώς $h\to 0$, μεταβάλλεται όταν η g διατρέχει τις συνεχείς συναρτήσεις στο [a,b]. Δέν είναι δύσκολο να δούμε (Άσκηση 5) ότι ο γρηγορώτερος τρόπος με τον οποίο μπορεί το μέτρο συνέχειας $\omega(g;h)$ μιάς $g\in C[a,b]$ να τείνει στο μηδέν (αν $g\not\equiv$ σταθ.) είναι <u>γραμμικός ως πρός h</u>, δηλ. όταν υπάρχει σταθερά c, ανεξάρτητη του h, τέτοια ώστε

$\omega(g;h) \leq ch$. Αυτό επιτυγχάνεται από τις συναρτήσεις $g\in C^1[a,b]$ γιά τις οποίες $\omega(g;h) \leq \|g'\|_\infty h$ και, γενικώτερα, από τις συναρτήσεις που ικανοποιούν συνθήκη Lipschitz με σταθερά L στο $[a,b]$ γιά τις οποίες $\omega(g;h)\leq Lh$. Μιά μεγαλύτερη κατηγορία συνεχών συναρτήσεων είναι οι συναρτήσεις που ικανοποιούν μιά συνθήκη Hölder με εκθέτη $\alpha\in(0,1)$ στο $[a,b]$ δηλ. οι συναρτήσεις $g\in C[a,b]$ γιά τις οποίες υπάρχει σταθερά K και $\alpha\in(0,1)$ τέτοιες ώστε

$$(26)\quad \omega(g;h) \leq Kh^\alpha \text{ γιά } h\geq 0.$$

Π.χ. η συνάρτηση $g(x)=x^\alpha$, $0<\alpha<1$ στο διάστημα $[0,1]$ έχει $\omega(g;h)=h^\alpha$. Γιά την συνάρτηση $g(x)=|x|^{1/2}$ στο $[-1,1]$ έχουμε γιά $|x-y|\leq h$, $x,y\in[-1,1]$ ότι $\max|g(x)-g(y)|\equiv\omega(g;h)=|g(0)-g(h)|=\sqrt{h}$.

Ένα από τα θεωρήματα του Jackson (βλ. π.χ. [4.9, σελ. 23]) μας λέει ότι αν $f\in C^r[a,b]$ και $n>r+1$, τότε

$$(26')\quad \min_{p\in P_{n-1}} \|f-p\|_\infty \leq c(r)((b-a)/(n-1))^r\ \omega(f^{(r)};(b-a)/2(n-1-r)),$$

όπου $c(r)=6(3e)^r/(r+1)$. Π.χ. γιά την συνάρτηση $f(x)=\sqrt{x}$ στο $[0,1]$ γιά την οποία $r=0$, $\omega(f;h)=h^{1/2}$, έχουμε ότι

$$\min_{p\in P_{n-1}} \|\sqrt{x}-p\|_\infty = \|\sqrt{x}-p^*\|_\infty \leq 6((b-a)/2(n-1))^{1/2},$$

δηλ. ότι το σφάλμα της βέλτιστης προσέγγισης τείνει στο 0 όπως το $n^{-1/2}$ όταν $n\to\infty$. Από την (23) βλέπουμε λοιπόν ότι το σφάλμα της παρεμβολής Lagrange γιά την $f(x)=\sqrt{x}$ στο $[0,1]$ στά σημεία Chebyshev x_i^c, $1\leq i\leq n$, τείνει στο μηδέν τουλάχιστον όσο γρήγορα τείνει η ακολουθία $\log n\ n^{-1/2}$ όταν $n\to\infty$. Γιά μία οποιαδήποτε συνάρτηση του $C^k[a,b]$ (σταθερό $k\geq 1$) - π.χ. γιά την συνάρτηση του Runge στο $[-1,1]$ - η παρεμβολή στα σημεία Chebyshev δίνει σφάλμα που τείνει στο μηδέν

- βλ. (26) με $r=k-1$, $\omega(f^{(k-1)};h)\leq \|f^{(k)}\|_\infty h$ - τουλάχιστον όσο γρήγορα και η ακολουθία $(\log n)/n^k$ όταν $n\to\infty$.

Από την (26$'$) βλέπουμε ότι το σφάλμα $\|f-p^*\|_\infty$ είναι μικρό αν ο λόγος $(b-a)/(n-1)$ γίνει μικρός. Αυτό μπορεί να επιτευχθεί είτε αυξάνοντας τον βαθμό του πολυωνύμου $n-1$ είτε υποδιαιρώντας το διάστημα $[a,b]$ σε μικρότερα διαστήματα σε κάθε ένα απ' τα οποία προσεγγίζουμε την f με κατάλληλα πολυώνυμα σταθερού βαθμού, συνολικά δηλ. προσεγγίζοντας την f με μία τμηματικά πολυωνυμική συνάρτηση. 'Εστω ότι υποδιαιρούμε το διάστημα $[a,b]$ σε k υποδιαστήματα (k μεγάλο) στο κάθε ένα από τα οποία προσεγγίζουμε την συνάρτησή μας με ένα πολυώνυμο βαθμού $n-1$ ($2\leq n\leq 6$ στην πράξη). Αυτό ισοδυναμεί, όσο αφορά το μέγεθος του φράγματος του σφάλματος, περίπου με την χρήση ενός πολυωνύμου βαθμού $k(n-1)$ στο $[a,b]$. Ο συνολικός αριθμός των σταθερών ("βαθμών ελευθερίας"), δηλ. των συντελεστών των πολυωνύμων που πρέπει να υπολογίσουμε, παραμένει επίσης περίπου ο ίδιος ($\cong kn$) και γιά τις δύο διαδικασίες. Υπάρχουν όμως στην πράξη σοβαρές διαφορές ανάμεσά τους:

(α) Ο υπολογισμός ενός πολυωνύμου μεγάλου βαθμού χρειάζεται (π.χ. ας πούμε γιά βαθμό ≥ 20) πολλή προσοχή στις πράξεις (π.χ. έκφραση του πολυωνύμου συναρτήσει βάσης καταλλήλων ορθογωνίων πολυωνύμων κλπ.), γιατί πολύ εύκολα μπορούμε, λόγω σφάλματος στρογγύλευσης, να καταλήξουμε σε overflow (αστάθεια στούς υπολογισμούς) ή underflow.

(β) Στην πράξη, σε πολλούς υπολογισμούς, εκφράζουμε τις προσεγγίσεις μας σαν γραμμικούς συνδυασμούς καταλλήλων συναρτήσεων βάσης. Γιά να υπολογίσουμε την τιμή ενός πολυωνύμου βαθμού $k(n-1)$ στο $[a,b]$ πρέπει να υπολογίσουμε $k(n-1)+1$ συντελεστές και τις τιμές $k(n-1)+1$ συναρτήσεων βάσης που είναι πολυώνυμα, δηλ. έχουν φορέα όλο το $[a,b]$. Αντίθετα, γιά τον υπολογισμό σ' ένα σημείο μιάς τμηματικά πολυωνυμικής συνάρτησης, υπολογίζουμε το πολύ ένα (μικρό) γραμμικό συνδυσμό (της τάξεως του n) συναρτήσεων βάσης που είναι τμηματικά πολυωνυμικές συναρτήσεις βαθμού $\leq n-1$ με μικρό φορέα (της τάξης n υποδιαστημάτων) ανεξάρτητο του αριθμού των υποδιαστημάτων k που μπορεί να γίνει όσο μεγάλος θέλουμε. Οι συντελεστές, γιά πολλές εφαρμογές, είναι λύσεις γραμμικών συστημάτων (μεγέθους $O(kn)\times O(kn)$)

τα οποία, γιά προσέγγιση με τμηματικά πολυωνυμικές συναρτήσεις, έχουν πίνακες μεγάλους αλλά <u>αραιούς</u> (συνήθως πίνακες ζώνης με πλάτη ζώνης $O(n)$ ενώ γιά προσέγγιση με πολυώνυμα στο $[a,b]$ έχουν τυπικά μεγάλους αλλά πυκνούς πίνακες.

(γ) Πολλές φορές η προσέγγιση γίνεται μέσω <u>παρεμβολής</u>. (Ακριβέστερα, το σφάλμα μιάς προσέγγισης πολλές φορές είναι δυνατόν να εκτιμηθεί εύκολα συναρτήσει του σφάλματος της παρεμβολής). Είδαμε τα προβλήματα της πολυωνυμικής παρεμβολής γιά ομοιόμορφο διαμερισμό με βήμα $h=(b-a)/k$, $k>>1$. Αντίθετα, η παρεμβολή σ' ένα διάστημα μικρού πλάτους $h=b-a$ με πολυώνυμα μικρού (σταθερού) βαθμού n, δηλ. η παρεμβολή Lagrange <u>τοπικά</u>, είναι πολύ <u>επιτυχής</u>: η (7) μας δίνει γιά το σφάλμα $e_{n-1}(x)$ γιά σταθερό n και $f\in C^n[a,b]$ ότι

$$(27)\quad |e_{n-1}(x)| \leq \prod_{i=1}^{n} |x-x_i|\ |f^{(n)}(\xi)/n!| \leq C_n h^n \max_{a\leq x\leq b} |f^{(n)}(x)|$$

<u>Παρατηρήσεις</u>

1. Είναι δυνατόν (βλ. π.χ. [4.7]) να βελτιώσουμε λίγο περισσότερο την σταθερά του Lebesgue που δίνουν τα σημεία Chebyshev (20) αν θεωρήσουμε γιά κάθε n την λεγόμενη <u>επέκταση</u> των σημείων Chebyshev. Τα αντίστοιχα σημεία x_i, $1\leq i\leq n$ της τ δίδουται τότε από τους τύπους

$$x_i=x_i^e = [(a+b)-(a-b)\cos((2i-1)\pi/2n)/\cos(\pi/2n)]/2,\quad 1\leq i\leq n.$$

Γιά την αντίστοιχη σταθερά του Lebesgue ισχύει ότι

$$(2/\pi)\log n+0.5 \leq \|\lambda_n^e\|_\infty \leq (2/\pi)\log n + .73.$$

Μπορεί επίσης να αποδειχθεί ότι <u>γιά κάθε n</u> η τιμή της $\|\lambda_n^e\|_\infty$ απέχει το πολύ .201 από την μικρότερη δυνατή τιμή $\|\lambda_n\|_\infty$ που μπορεί να επιτευχθεί γιά κάθε n με κατάλληλη επιλογή της τ.

2. Από τον τύπο του σφάλματος της παρεμβολής (7) έχουμε γιά $f \in C^n[a,b]$ ότι

(28) $\|f-P_n f\|_\infty \leq \|w\|_\infty \|f^{(n)}\|_\infty / n!$

όπου $w(x) = \prod_{i=1}^{n}(x-x_i)$, $x \in [a,b]$.

Είναι γνωστό (βλ. Ασκ. 6) ότι

(29) $$\min_{x_i \in [a,b]} \|w\|_\infty = 2(b-a)^n/4^n$$

και ότι η ελάχιστη αυτή τιμή της $\|w\|_\infty$ συμβαίνει όταν τα σημεία παρεμβολής x_i συμπίπτουν με τις ρίζες x_i^c, $1 \leq i \leq n$ (20), του πολυωνύμου Chebyshev $n^{ου}$ βαθμού! Ξαναβλέπουμε δηλ. την σημασία των x_i^c ως κόμβων στην παρεμβολή Lagrange. Συνεπώς η αποτυχία της παρεμβολής Lagrange πρέπει να οφείλεται στον όρο $\|f^{(n)}\|_\infty/n!$ ο οποίος γιά πολλές C^∞ συναρτήσεις δεν είναι φραγμένος καθώς $n \to \infty$. Υπάρχουν όμως βέβαια κλάσεις συναρτήσεων (π.χ. εκείνες γιά τις οποίες $\|f^{(n)}\|_\infty \leq M^n$, $n=1,2,\ldots$ γιά κάποιο $0 \leq M < \infty$) γιά τις οποίες η ακολουθία των πολυωνύμων παρεμβολής συγκλίνει ομοιόμορφα στην f καθώς $n \to \infty$ γιά οποιαδήποτε τριγωνικό σύστημα παρεμβολής, Ασκ. 7)

Ασκήσεις 4.1

1. (α) Αποδείξτε την (3).
   (β) Αποδείξτε την (5).

2. Αποδείξτε γιά το παράδειγμα του Runge ότι το όριο $\lim r_n(x)$, n άρτιος $\to \infty$, υπάρχει και ορίζει μία συνεχή συνάρτηση $r(x)$ στο $[-1,1]$ με $r(1)>0$.

3. Δείξτε ότι γιά κάθε τ με n σημεία υπάρχει $f\in C[a,b]$, $f\neq 0$ τέτοια ώστε να ισχύει η (18) ως ισότης. (Υπόδειξη: έστω $\hat{x}\in[a,b]$ τέτοιο ώστε $\lambda_n(\hat{x})=\|\lambda_n\|_\infty$. Ορίστε $\sigma_i=\text{sign}\ (L_i(x))$, $1\leq i\leq n$ και κατασκευάστε συνεχή συνάρτηση g στο [a,b] τέτοια ώστε $g(x_i)=\sigma_i$, $1\leq i\leq n$ και $\|g\|_\infty=1$ - π.χ. την τμηματικά γραμμική συνεχή συνάρτηση με $g(x_i)=\sigma_i$ με επέκταση με σταθερές έξω απ' το διάστημα $[\min_i x_i, \max_i x_i]$.
- Δείξτε τότε ότι $\|P_n g\|_\infty = \|\lambda_n\|_\infty \|g\|_\infty$.

4. (α) Δείξτε ότι γιά κάθε $f\in C[a,b]$ υπάρχει τριγωνικό σύστημα παρεμβολής τέτοιο ώστε $\|f-P_nf\|_\infty \to 0$, $n\to\infty$. (Χρησιμοποιείστε το γεγονός ότι το πολυώνυμο $p^*\in P_{n-1}$ της βέλτιστης προσέγγισης της f από στοιχεία του P_{n-1} έχει την εξής ιδιότητα: Υπάρχουν n+1 διακριτά σημεία x_i στο [a,b] στά οποία το σφάλμα $f-p^*$ έχει ίσες απολύτως και εναλλασσόμενες στο πρόσημο τιμές. Μάλιστα $|(f-p^*)(x_i)| = \|f-p^*\|_\infty$. Συνεπώς η διαφορά $f-p^*$ θα μηδενίζεται σε n διακριτά σημεία $y_i\in[a,b]$, δηλ. το p^* θα είναι το πολυώνυμο παρεμβολής Lagrange της f στά σημεία y_i, $1\leq i\leq n$. Χρησιμοποιώντας τώρα το θεώρημα του Weierstrass αποδείξτε το ζητούμενο).

(β) Χρησιμοποιώντας την ιδιότητα του p^* του μέρους (α) δείξτε γιά δεδομένο n και $f\in C^n[a,b]$

$$\min_{p\in P_{n-1}} \|f-p\|_\infty = c(y_1,\ldots,y_n)\, f^{(n)}(\xi)$$

γιά κάποιο $\xi\in(a,b)$, όπου $c(y_1,\ldots,y_n)$ σταθερά που εξαρτάται μόνο από τα σημεία y_i. Από την παρατήρηση 2 συμπεράνετε ότι

$$\min_{p\in P_{n-1}} \|f-p\|_\infty \geq 2((b-a)^n/4^n n!)\, \min_{a\leq x\leq b} |f^{(n)}(x)|$$

5.(α) Δείξτε ότι η $\omega(g;h)$ είναι μονοτονική και υπο-αθροιστική, δηλ. ότι $\omega(g;h) \le \omega(g;h+k) \le \omega(g;h) + \omega(g;k)$ γιά $h,k \ge 0$, $g \in C[a,b]$. Συμπεράνετε ότι $\omega(g;\lambda h) \le N_\lambda\, \omega(g;h)$ όπου γιά $\lambda \ge 0$, ο N_λ είναι ο ελάχιστος ακέραιος τέτοιος ώστε $\lambda \le N_\lambda$.

(β) Αν $\omega(g;h)/h \to 0$ όταν $h \downarrow 0$, δείξτε ότι η g είναι σταθερά.

6. Αποδείξτε τις (28) και (29) και ότι η ελάχιστη αυτή τιμή της $\|\omega\|_\infty$ συμβαίνει όταν $x_i = x_i^c$, $1 \le i \le n$. (Θυμηθείτε την παράσταση των πολυωνύμων Chebyshev στο $[-1,1]$ και υπολογίστε τον συντελεστή του μεγιστοβάθμιου όρου του T_n)

7. Αποδείξτε ότι αν $f \in C^\infty[a;b]$ τέτοια ώστε $\|f^{(n)}\|_\infty \le M^n$, $n=0,1,2,\ldots$ γιά κάποιο $0 \le M < \infty$, τότε $\|f - P_n f\|_\infty \to 0$, $n \to \infty$ γιά οποιοδήποτε τριγωνικό σύστημα παρεμβολής.

8. Στο Θεώρημα 1 δείξτε ότι η $f^{(n)}(\xi(\bar{x}))$ είναι συνεχής συνάρτηση του $\bar{x} \in [a,b]$.

9. Δείξτε ότι γιά $-1 \le x_1 < x_2 < x_3 \le 1$ η ελάχιστη τιμή της σταθεράς του Lebesgue $\|\lambda_3\|_\infty$ ($[a,b]=[-1,1]$) είναι 5/4 και ότι η τιμή αυτή λαμβάνεται γιά $x_2=0$, $-x_1=x_3>2\sqrt{2}/3$. Δείξτε ότι $\|\lambda_3^c\|_\infty=5/3$. Συνεπώς οι ρίζες των πολυωνύμων Chebyshev δεν δίνουν γιά κάθε n την ελάχιστη τιμή της σταθεράς του Lebesgue.

10. <u>Στο μιγαδικό επίπεδο</u> έστω p_n το μιγαδικό πολυώνυμο βαθμού $\le n$ που παρεμβάλλει την συνάρτηση $f(z)=1/z$ στις n n-στές ρίζες της μονάδος. Δείξτε ότι $p_n(z)=z^{n-1}$ και ότι

$$\max_{|z|=1} |p_n(z) - f(z)| \not\to 0, \quad n \to \infty$$

(11) (α) (<u>Παρεμβολή Hermite</u>). Δείξτε ότι υπάρχει μοναδικό πολυώνυμο $q(x)$ βαθμού $\leq 2n-1$ που ικανοποιεί τις σχέσεις $q(x_i)=f(x_i)$, $q'(x_i)=f'(x_i)$, $1\leq i\leq n$ γιά δεδομένη $\tau=\{x_i\}$, $1\leq i\leq n$ και συναρτήση $f\in C^1[a,b]$. Το $q(x)$ λέγεται πολυώνυμο παρεμβολής Hermite γιά την f στα σημεία x_i.

(β) Δείξτε ότι

$$q(x) = \sum_{i=1}^{n} [f(x_i)\, A_i(x) + f'(x_i)\, B_i(x)]$$

όπου τα $A_i, B_i \in P_{2n-1}$ είναι τα (μοναδικά) πολυώνυμα βαθμού $\leq 2n-1$ τέτοια ώστε $A_i(x_j)=\delta_{ij}$, $A_i'(x_j)=0$, $B_i(x_j)=0$, $B_i'(x_j)=\delta_{ij}$, $1\leq i,\ j\leq n$ και ότι δίνονται από τους τύπους

$$A_i(x) = (1-2(x-x_i)\, L_i'(x_i))\, L_i^2(x)$$
$$B_i(x) = (x-x_i)\, L_i^2(x).$$

(γ) Αν $\tau \subset [a,b]$ και $f\in C^{2n}[a,b]$, ισχύει το εξής αποτέλεσμα γιά το σφάλμα της παρεμβολής Hermite: γιά $x\in[a,b]$ υπάρχει $\xi=\xi_x\in(a,b)$ τέτοιο ώστε

$$f(x)-q(x) = \{\prod_{i=1}^{n} (x-x_i)^2\}\, f^{(2n)}(\xi_x)/(2n)!$$

4.2 ΠΑΡΕΜΒΟΛΗ ΚΑΙ ΠΡΟΣΕΓΓΙΣΗ ΜΕ ΤΜΗΜΑΤΙΚΑ ΓΡΑΜΜΙΚΕΣ ΣΥΝΑΡΤΗΣΕΙΣ

Αρχίζουμε την μελέτη της προσέγγισης με τμηματικά πολυωνυμικές συναρτήσεις με την απλούστερη περίπτωση, δηλ. με τις τμηματικά γραμμικές συναρτήσεις, που βέβαια δεν έχουν την πρακτική σημασία π.χ. των τμηματικά κυβικών συναρτήσεων (κυβικών splines). Η μελέτη μας όμως θα εστιασθεί σε προβλήματα (απλής μορφής εδώ) σημαντικά και γιά την προσέγγιση με τμηματικά πολυωνυμικές συναρτήσεις οποιουδήποτε βαθμού.

'Εστω ο διαμερισμός τ: $a=x_1<x_2<\dots x_N=b$ του $[a,b]$ και $f\in C[a,b]$. Προφανώς υπάρχει μόνο μία συνάρτηση, γραμμική σε κάθε διάστημα $[x_i,x_{i+1}]$ που παίρνει τις τιμές $f(x_i)$ γιά $i=1,2,\dots,N$. Η "τεθλασμένη αυτή γραμμή" λέγεται συνάρτηση παρεμβολής της f στον χώρο των τμηματικά γραμμικών συναρτήσεων (που ορίζει η τ) και θα παριστάνεται με I_2f. Προφανώς η I_2f κατασκευάζεται πολύ εύκολα και δίνεται από

$$(1)\quad (I_2f)(x) = f(x_i)+(x-x_i)f[x_i,x_{i+1}] \text{ γιά } x_i\le x\le x_{i+1},\ 1\le i\le N-1.$$

Γιά να εκτιμήσουμε το σφάλμα $\varepsilon(x)=f(x)-(I_2f)(x)$ παρατηρούμε ότι στο διάστημα $[x_i,x_{i+1}]$, η I_2f είναι το γραμμικό πολυώνυμο παρεμβολής Lagrange της $f(x)$ γιά τα σημεία x_i,x_{i+1}. Συνεπώς, αν $f\in C^2[a,b]$ η (4.1.7) δίνει ότι γιά $x\in[x_i,x_{i+1}]$, υπάρχει $\xi=\xi(x)\in(x_i,x_{i+1})$ τέτοιο ώστε

$$(2)\quad \varepsilon(x) = (x-x_i)(x-x_{i+1})\, f^{(2)}(\xi)/2,$$

από την οποία έπεται ότι γιά $x\in[x_i,x_{i+1}]$

$$|\varepsilon(x)| \le \max_{x_i\le x\le x_{i+1}} (x-x_i)(x_{i+1}-x) \max_{x_i\le x\le x_{i+1}} |f''(x)|/2.$$

Η μέγιστη τιμή της συνάρτησης $(x-x_i)(x_{i+1}-x)$ γιά $x\in[x_i,x_{i+1}]$ λαμβάνεται στο σημείο $x=(x_i+x_{i+1})/2$ και είναι ίση με $(x_{i+1}-x_i)^2/4$. Συνεπώς, η παραπάνω σχέση δίνει γιά $f\in C^2[x_i,x_{i+1}]$

$$(3)\ \max_{x_i\le x\le x_{i+1}} |(f-I_2f)(x)|\le(x_{i+1}-x_i)^2 \max_{x_i\le x\le x_{i+1}} |f''(x)|/8,\ 1\le i\le N-1$$

που μάλιστα ισχύει ως ισότητα γιά την συνάρτηση $f(x)=(x-x_i)(x_{i+1}-x)$ γιά την οποία $(I_2f)(x)\equiv 0$ στο $[x_i,x_{i+1}]$. Από την (3), γιά κάθε $f\in C^2[a,b]$ έχουμε, με $\|g\|_\infty=\max_{a\le x\le b}|g(x)|$, ότι

$$(4)\ \|f-I_2f\|_\infty \le h^2\|f''\|_\infty/8,$$

όπου

$$(5)\ h=\max_{1\le i\le N-1}(x_{i+1}-x_i)$$

Συνεπώς μπορούμε να κάνουμε το σφάλμα όσο μικρό θέλουμε παίρνοντας όλο και λεπτότερους διαμερισμούς τ του $[a,b]$. Το πρόβλημα δεν γίνεται πιό πολύπλοκο καθώς $h\to 0$ γιατί πάντα σε κάθε υποδιάστημα $[x_i,x_{i+1}]$ η I_2f είναι ευθύγραμμο τμήμα.

Γιά δεδομένο διαμερισμό $\tau: a=x_1<x_2<\ldots<x_N=b$ του $[a,b]$ ορίζουμε τον διανυσματικό χώρο των (συνεχών) τεθλασμένων γραμμών (χώρο των <u>γραμμικών splines</u>)

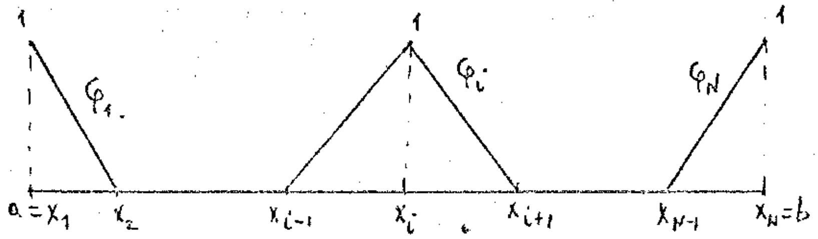
$$S_\tau^2=\{\varphi:\ \varphi\in C[a,b],\ \varphi\in P_1\ (x_i,x_{i+1}),\ 1\le i\le N-1\},$$

όπου με $P_k(I)$ εννοούμε τον χώρο των πολυωνύμων βαθμού $\le k$ στο διάστημα I. Προφανώς $\forall f\in C[a,b]$ ορίζεται η I_2f ως στοιχείο του S_τ^2. Σε πολλές εφαρμογές είναι απαραίτητο να εκφράζουμε τα στοιχεία του

S_τ^2 συναρτήσει μιάς εύχρηστης <u>βάσης</u> του διανυσματικού αυτού χώρου.

Λήμμα 1. Ο S_τ^2 είναι N-διάστατος υπόχωρος του $C[a,b]$. Οι συναρτήσεις $\varphi_i \in S_\tau^2$, $1 \leq i \leq N$ που ορίζονται γιά $x \in [a,b]$ από τις σχέσεις $\varphi_i(x_j)=\delta_{ij}$, αποτελούν βάση του S_τ^2.

<u>Απόδειξη</u>: Οι συναρτήσεις φ_i δίνονται από



γιά $x \in [a,b]$ από τους τύπους

$$\varphi_1(x)=\begin{cases}(x_2-x)/(x_2-x_1), & \text{αν } x\in[x_1,x_2]\\ 0 & \text{αλλοιώς,}\end{cases}$$

$$\varphi_N(x)=\begin{cases}(x-x_{N-1})/(x_N-x_{N-1}) & \text{αν } x\in[x_{N-1},x_N]\\ 0 & \text{αλλοιώς}\end{cases}$$

(6)

$$\text{και γιά } 2\leq i\leq N-1: \varphi_i(x)=\begin{cases}(x-x_{i-1})/(x_i-x_{i-1}) & \text{αν } x\in[x_{i-1},x_i]\\ (x_{i+1}-x)/(x_{i+1}-x_i) & \text{αν } x\in[x_i,x_{i+1}]\\ 0 & \text{αλλοιώς}\end{cases}$$

Προφανώς $\varphi_i \in S_\tau^2$, $1\leq i\leq N$, $\varphi_i(x_j)=\delta_{ij}$, $1\leq i,j\leq N$. (Θα αποδειχθεί σημαντικό αργότερα ότι οι φ_i έχουν μικρό φορέα: $|\mathrm{supp}(\varphi_i)|\leq 2h$). Είναι προφανές ότι οι φ_i είναι γραμμικά ανεξάρτητες. Πράγματι, αν $\sum_{i=1}^{N} c_i\varphi_i(x)=0$

$\forall x \in [a,b]$, τότε θέτοντας $x=x_j$ γιά $1 \le j \le N$ παίρνουμε $c_j=0$. Είναι επίσης προφανές ότι παράγουν τον S_τ^2, γιατί για κάθε $\psi \in S_\tau^2$ έχουμε την παράσταση

(7) $\psi(x) = \sum_{i=1}^{N} \psi(x_i)\, \varphi_i(x)$.

(Το δεύτερο μέλος της (7) ανήκει σον S_τ^2 και γιά $x=x_j$ έχει την τιμή $\psi(x_j)$. Συνεπώς οι δύο τεθλασμένες γραμμές $\psi(x)$, $\sum_{i=1}^{N} \psi(x_i)\varphi_i(x)$ συμπίπτουν παντού στο $[a,b]$). @

Συνεπώς, η συνάρτηση παρεμβολής $I_2 f$ της $f \in C[a,b]$ στον S_τ^2 είναι το (μοναδικό) στοιχείο του S_τ^2 με παράσταση

(8) $(I_2 f)(x) = \sum_{i=1}^{N} f(x_i)\varphi_i(x)$, $1 \le i \le N$,

όπου οι συναρτήσεις βάσης φ_i δίνονται από τις σχέσεις (6).

Ας συγκρίνουμε τώρα το σφάλμα $\|f-I_2 f\|_\infty$ της συνάρτησης παρεμβολής $I_2 f \in S_\tau^2$ μιάς συνάρτησης $f \in C[a,b]$ με το ελάχιστο σφάλμα $\min\{\|f-\varphi\|_\infty : \varphi \in S_\tau^2\}$ που μπορούμε να πετύχουμε προσεγγίζοντας την f με στοιχεία του S_τ^2 (ως προς την νόρμα $\|.\|_\infty$), δηλ. με το σφάλμα μιάς <u>βέλτιστης προσέγγισης</u> της f στον $(S_\tau^2, \|.\|_\infty)$ ('Οτι υπάρχει βέλτιστη προσέγγιση είναι απόρροια της <u>πεπερασμένης διάστασης</u> του S_τ^2, βλ. π.χ. [4.9 σελ. 1]). Παρατηρούμε ότι ο τελεστής της παρεμβολής $I_2 : C[a,b] \to S_\tau^2$ είναι γραμμικός και συμπίπτει με την ταυτότητα στον S_τ^2, δηλ. ότι

(9) $I_2\varphi = \varphi \quad \forall \varphi \in S_\tau^2$.

Επιπλέον, είναι φραγμένος στον $C[a,b]$ και μάλιστα

$$(10)\ \|I_2 g\|_\infty = \max_{1\le i\le N}|(I_2 g)(x_i)| = \max_{1\le i\le N}|g(x_i)| \le \|g\|_\infty \quad \forall g\in C[a,b].$$

Από τις (9) και (10) έχουμε για $f\in C[a,b]$ ότι

$$\|f-I_2 f\|_\infty = \|(f-\varphi)-I_2(f-\varphi)\|_\infty \le \|f-\varphi\|_\infty + \|I_2(f-\varphi)\|_\infty$$

$$\le \|f-\varphi\|_\infty + \|f-\varphi\|_\infty = 2\|f-\varphi\|_\infty \ \forall \varphi\in S^2_\tau.$$

Συνεπώς, έχουμε τελικά

$$(11)\ \min_{\varphi\in S^2_\tau}\|f-\varphi\|_\infty \le \|f-I_2 \varphi\|_\infty \le 2\min_{\varphi\in S^2_\tau}\|f-\varphi\|_\infty,$$

δηλ. ότι η $I_2 f$ είναι "<u>σχεδόν βέλτιστη</u>"- με την έννοια ότι $\|f-I_2 f\|_\infty \le c \min_{\varphi\in S^2_\tau} 2\|f-\varphi\|_\infty$, c σταθερά ανεξάρτητη των τ, N-προσέγγιση: Αν βρίσκαμε μία βέλτιστη προσέγγιση της f στον $(S^2_\tau, \|.\|_\infty)$, το πολύ-πολύ να υποδιπλασιάζαμε το σφάλμα της $I_2 f$.

Προχωρούμε τώρα σε μιά πιό λεπτομερή μελέτη του σφάλματος της παρεμβολής στον χώρο S^2_τ. Θα χρησιμοποιήσουμε το λεγόμενο <u>θεώρημα του πυρήνα Peano</u> σε μιά απλή μορφή του. Γενικεύοντας λίγο τον χώρο $C^k[a,b]$, θα θεωρήσουμε, για $k\ge 1$ ακέραιο, τον χώρο $PC^k[a,b]$, που ορίζεται ως ο χώρος εκείνων των συναρτήσεων f του $C^{k-1}[a,b]$ των οποίων η k-στή παράγωγος υπάρχει σε όλα τα σημεία του [a,b], εκτός πιθανώς από ένα πεπερασμένο σύνολο σημείων στο [a,b] (ανάμεσα στα οποία και στα a,b είναι συνεχής), είναι φραγμένη στο [a,b] και ορίζεται σε κάθε σημείο του [a,b] μονοσήμαντα σαν όριο τιμών της από δεξιά ή αριστερά, λόγω συνέχειας.

ΘΕΩΡΗΜΑ 1 (Πυρήνα Peano). Εστω $E:PC^{n+1}[a,b]\to R^1$, $n\geq 0$ ένα <u>γραμμικό</u> συναρτησιακό (δηλ. έστω ότι $E(\lambda f+\mu g)=\lambda E(f)+\mu E(g)$ $\forall\, f,g\in PC^{n+1}[a,b]$, $\lambda,\mu\in R^1$), τέτοιο ώστε $E(p)=0$ $\forall\, p\in P_n$. Τότε για κάθε $f\in PC^{n+1}[a,b]$ έχουμε την παράσταση

$$(12)\quad E(f)=(n!)^{-1}\,E_x\Big[\int_a^b (x-t)_+^n f^{(n+1)}(t)\,dt\Big],$$

όπου
$$(x-t)_+^n=\begin{cases}(x-t)^n & x\geq t\\ 0 & x<t\end{cases}$$

και όπου με E_x τονίζουμε το γεγονός ότι το E δρα στον όρο μέσα στις αγκύλες θεωρούμενο ως συνάρτηση του x.

<u>Απόδειξη</u>: Η συνάρτηση

$$G_n(x)\equiv\int_a^b (x-t)_+^n f^{(n+1)}(t)\,dt=\int_a^x (x-t)^n f^{(n+1)}(t)\,dt,$$

για $f\in PC^{n+1}[a,b]$, ανήκει στον χώρο $PC^{n+1}[a,b]$ όπως εύκολα μπορούμε να δούμε παραγωγίζοντας το δεύτερο μέλος ως προς x. Από το Θεώρημα του Taylor με την ολοκληρωτική μορφή υπολοίπου, έχουμε για $x\in[a,b]$ επειδή $f\in PC^{n+1}[a,b]$

$$f(x)=f(a)+(x-a)f'(a)+\ldots+(x-a)^n f^{(n)}(a)/n!+G_n(x)/n!$$

Η (12) τώρα προκύπτει αμέσως επειδή το E είναι γραμμικό συναρτησιακό και επειδή λόγω της υπόθεσής μας, $E_x((x-a)^k)=0$ για $0\leq k\leq n$.@

ΘΕΩΡΗΜΑ 2. Έστω $f\in C^2[a,b]$ και έστω $e(x)=f(x)-(I_2 f)(x)$ το σφάλμα της παρεμβολής στον S_τ^2. Για $x_i\leq x\leq x_{i+1}$ με $h_i=(x_{i+1}-x_i)$ έχουμε

$$(13)\quad \varepsilon(x) = \int_{x_i}^{x_{i+1}} K(x,t) f^{(2)}(t)dt,$$

όπου

$$(14) K(x,t)= \begin{cases} -(x_{i+1}-x)(t-x_i)/h_i, & x_i \le t \le x \le x_{i+1}, \\ \\ -(x-x_i)(x_{i+1}-t)/h_i, & x_i \le x < t \le x_{i+1}. \end{cases}$$

και

$$(15)\quad \varepsilon'(x)= \int_{x_i}^{x_{i+1}} \Lambda(x,t) f^{(2)}(t)dt,$$

όπου

$$(16)\quad \Lambda(x,t)= \begin{cases} (t-x_i)/h_i, & x_i \le t \le x \le x_{i+1}, \\ \\ -(x_{i+1}-t)/h_i, & x_i \le x < t \le x_{i+1}. \end{cases}$$

<u>Απόδειξη</u>: Σταθεροποιούμε ένα x στο $[x_i, x_{i+1}]$. Τότε το $\varepsilon(x)=\varepsilon_x(f)=f(x)-I_2 f(x)$ είναι ένα συναρτησιακό στου χώρο $C^2[x_i, x_{i+1}]$ για το οποίο $\varepsilon_x(p)=0 \quad \forall p \in P_1[x_i, x_{i+1}]$. Εφαρμόζοντας λοιπόν το Θεώρημα του Πυρήνα του Peano στο $[x_i, x_{i+1}]$ για $n=1$ έχουμε

$$(17)\ \varepsilon(x)=\varepsilon_x\Big[\int_{x_i}^{x_{i+1}}(x-t)_+^1 f^{(2)}(t)dt\Big]=\int_{x_i}^{x_{i+1}}(x-t)_+^1 f^{(2)}(t)dt-$$

$$I_{2,x}\Big[\int_{x_i}^{x}(x-t)f^{(2)}(t)dt\Big],\qquad x\varepsilon[x_i,x_{i+1}]$$

Εξ ορισμού της συνάρτησης παρεμβολής I_2g στο $[x_i,x_{i+1}]$ (είναι η γραμμική συνάρτηση στο $[x_i,x_{i+1}]$ με $(I_2g)(x_k)=g(x_k),k=i,i+1)$, βλέπουμε ότι για $x\varepsilon[x_i,x_{i+1}]$

$$I_{2,x}\Big[\int_{x_i}^{x}(x-t)f^{(2)}(t)dt\Big]=((x-x_i)/h_i)\int_{x_i}^{x_{i+1}}(x_{i+1}-t)f^{(2)}(t)dt.$$

Η (17) και η παραπάνω σχέση δίνουν τις (13)-(14). Γράφοντας τώρα

$$\varepsilon(x)=\int_{x_i}^{x}K(x,t)f^{(2)}(t)dt+\int_{x}^{x_{i+1}}K(x,t)f^{(2)}(t)dt,$$

παρατηρώντας ότι λόγω της συνέχειας της $K(x,t)$ για $x=t$

$$(d/dx)\Big(\int_{x_i}^{x}K(x,t)f^{(2)}(t)dt\Big)=\int_{x_i}^{x}K_x(x,t)f^{(2)}(t)dt+K(x,x)f^{(2)}(x-)$$

$$(d/dx)\Big(\int_{x}^{x_{i+1}}K(x,t)f^{(2)}(t)dt\Big)=\int_{x}^{x_{i+1}}K_x(x,t)f^{(2)}(t)dt-K(x,x)f^{(2)}(x^+)$$

και χρησιμοποιώντας ότι $f\varepsilon C^2[a,b]$ παίρνουμε τις (15)-(16) γιατί η συνάρτηση $\Lambda(x,t)$ είναι ίση με $K_x(x,t)$ τμηματικά, για $x_i<t<x$ και $x<t<x_{i+1}$.@

Με βάση τις παραστάσεις (13),(15) μπορούμε να βρούμε εκτιμήσεις του σφάλματος της παρεμβολής και της παραγώγου του στην νόρμα $\|.\|_\infty$ αλλά και σε άλλες νόρμες. Ιδιαίτερα μας ενδιαφέρει η L^2-νόρμα, που για τετραγωνικά ολοκληρώσιμες πραγματικές συναρτήσεις στο $[a,b]$ θα συμβολίζεται ως

(18) $\|f\|=(\int_a^b f^2(x)dx)^{1/2}$

ΠΟΡΙΣΜΑ 1. Εστω $f\in C^2[a,b]$ και $h=\max(x_{i+1}-x_i)$, $1\le i\le N-1$. Για το σφάλμα $e(x)=f(x)-I_2f(x)$ της παρεμβολής στον S_τ^2 έχουμε

(19) $\|e\|_\infty \le h^2\|f''\|_\infty/8$

(20) $\|e'\|_\infty \le h\|f''\|_\infty/2$

(21) $\|e\| \le h^2\|f''\|/3\sqrt{10}$

(22) $\|e'\| \le h\,\|f''\|/\sqrt{6}$

<u>Απόδειξη</u>:

Οι (13)-(14) δίνουν, για $x\in[x_i,x_{i+1}]$, $h_i=x_{i+1}-x_i$

$$|e(x)|\le(\int_{x_i}^{x_{i+1}} |K(x,t)|dt)\|f''\|_\infty=\|f''\|_\infty h_i^{-1}[\int_{x_i}^{x}(t-x_i)(x_{i+1}-x)dt$$

$$+\int_x^{x_{i+1}} (x-x_i)(x_{i+1}-t)dt=$$

$$=\|f''\|_\infty h_i^{-1}[(x_{i+1}-x)(x-x_i)^2+(x-x_i)(x-x_{i+1})^2]/2$$

$$= \|f''\|_\infty (x_{i+1}-x)(x-x_i)/2 \leq \|f''\|_\infty h_i^2/8$$

όπως προηγουμένως. Συνεπώς ισχύει η (19) την οποία είχαμε αποδείξει και προηγουμένως χρησιμοποιώντας την "σημειακή" παράσταση του σφάλματος.

Οι (15)-(16) δίνουν τώρα για $x\in[x_i,x_{i+1}]$

$$|\varepsilon'(x)| \leq \|f''\|_\infty \int_{x_i}^{x_{i+1}} |\Lambda(x,t)|dx=$$

$$=\|f''\|_\infty h_i^{-1}[\int_{x_i}^{x} (t-x_i)dt+\int_{x}^{x_{i+1}} (x_{i+1}-t)dt]$$

$$=\|f''\|_\infty h_i^{-1}[(x-x_i)^2+(x_{i+1}-x)^2]/2 \leq \|f''\|_\infty h_i/2$$

από την οποία προκύπτει η (20) με προφανή επέκταση της νόρμας $\|.\|_\infty$ στις συναρτήσεις του χώρου $PC^0[a,b]$ των τμηματικά συνεχών φραγμένων συναρτήσεων στον οποίο ανήκει η $\varepsilon'(x)$.

Από τις (13)-(14) τώρα, με χρήση της ανισότητας Cauchy-Schwarz έχουμε, για $x\in[x_i,x_{i+1}]$ ότι

$$(23)\quad \varepsilon^2(x) \leq \left(\int_{x_i}^{x_{i+1}} K^2(x,t)dt\right)\left(\int_{x_i}^{x_{i+1}} (f''(t))^2dt\right).$$

Εύκολα βλέπουμε ότι

$$\int_{x_i}^{x_{i+1}} K^2(x,t)dt = h_i^{-2}[(x_{i+1}-x)^2\int_{x_i}^{x}(t-x_i)^2dt+(x-x_i)^2\int_{x}^{x_{i+1}} (x_{i+1}-t)^2dt]$$

$$= (3h_i)^{-1}(x_{i+1}-x)^2(x-x_i)^2$$

Άρα, ολοκληρώνοντας και τα δύο μέλη της (23) ως προς x από x_i έως x_{i+1} έχουμε

$$\int_{x_i}^{x_{i+1}} \epsilon^2(x)dt \le (3h_i)^{-1}\Big[\int_{x_i}^{x_{i+1}} (x_{i+1}-x)^2(x-x_i)^2dx\Big]\int_{x_i}^{x_{i+1}} (f''(t))^2dt$$

$$=(h_i^4/90)\int_{x_i}^{x_{i+1}} (f''(t))^2dt$$

Συνεπώς,

$$\|\epsilon\|^2=\int_a^b \epsilon^2(x)dx=\sum_{i=1}^{N-1}\int_{x_i}^{x_{i+1}} \epsilon^2(x)dx$$

$$\le(h^4/90)\sum_{i=1}^{N-1}\int_{x_i}^{x_{i+1}} (f''(t))^2dt= (h^4/90)\|f''\|^2$$

από την οποία έπεται η (21).

Τέλος, οι (15)-(16) δίνουν για $x\epsilon[x_i,x_{i+1}]$

$$(\epsilon'(x))^2\le\Big(\int_{x_i}^{x_{i+1}} A^2(x,t)dt\Big)\Big(\int_{x_i}^{x_{i+1}} (f''(t))^2dt\Big),$$

και επειδή

$$\int_{x_i}^{x_{i+1}} A^2(x,t)dt = h_i^{-2}\Big[\int_{x_i}^{x} (t-x_i)^2dt+\int_{x}^{x_{i+1}} (t-x_{i+1})^2dt\Big]$$

$$= [(x-x_i)^3+(x_{i+1}-x)^3]/3h_i^2$$

έχουμε ότι

$$\int_{x_i}^{x_{i+1}} (e'(x))^2 dx \le \left(3^{-1}h_i^{-2}\int_{x_i}^{x_{i+1}} [(x-x_i)^3+(x_{i+1}-x)^3]dx\right)\left(\int_{x_i}^{x_{i+1}} (f''(t))^2 dt\right)$$

$$\le (h_i^2/6)\int_{x_i}^{x_{i+1}} (f''(t))^2 dt$$

από την οποία προκύπτει η (22) αν αθροίσουμε ως πρός i και πάρουμε την τετραγωνική ρίζα και των δύο μελών.@

Αν η συνάρτηση f είναι λιγότερο ομαλή τότε, γενικά, δεν περιμένουμε τάξη ακρίβειας π.χ. $O(h^2)$ στην (19) αλλά μικρότερη. Θα εξετάσουμε μερικές τέτοιες περιπτώσεις στις Παρατηρήσεις και στις Ασκήσεις. Ας προχωρήσουμε όμως τώρα στην μελέτη ενός άλλου προβλήματος, δηλ. στην προσέγγιση μιάς συνεχούς συνάρτησης f από στοιχεία του S_τ^2 με την έννοια των <u>ελαχίστων τετραγώνων</u>.

Θεωρούμε στον διανυσματικό χώρο C[a,b] το <u>εσωτερικό γινόμενο</u>

$$(24)\quad (f,g) = \int_a^b f(x)g(x)dx,$$

το οποίο βέβαια παράγει την L^2-νόρμα $\|.\|$, (18). Επειδή ο χώρος S_τ^2 είναι υπόχωρος του {C[a,b], (.,.)} πεπερασμένης διαστάσεως, είναι γνωστό (βλ. Ασκηση 1α) ότι το πρόβλημα του προσδιορισμού ενός $f_h \in S_\tau^2$ που να ικανοποιεί

$$(25)\quad \|f-f_h\| = \min_{\varphi\in S_\tau^2} \|f-\varphi\|,$$

δηλ. το πρόβλημα της <u>βέλτιστης προσέγγισης</u> της f από στοιχεία του S_τ^2 για την νόρμα $\|.\|$ (της λεγόμενης <u>προσέγγισης ελαχίστων τετραγώνων</u>

της f από στοιχεία του S_τ^2), έχει μοναδική λύση f_h που είναι επίσης και η μοναδική λύση του προβλήματος των κανονικών εξισώσεων

(26) $(f_h,\varphi) = (f,\varphi) \quad \forall\varphi\in S_\tau^2$.

Από την (26) π.χ. προκύπτει ότι υπάρχει γραμμικός τελεστής $P:C[a,b]\to S_\tau^2$, ο λεγόμενος τελεστής της ορθής προβολής (ως προς το εσωτερικό γινόμενο (24)) ή της L^2-προβολής στον S_τ^2 τέτοιος ώστε

(27) $f_h=Pf$.

Η (26) γράφεται και ως $(f-Pf,\varphi)=0 \quad \forall\varphi\in S_\tau^2$. Δηλ. το σφάλμα $f-Pf$ της L^2- προβολής της f στον S_τ^2 είναι ορθογώνιο προς τον S_τ^2. Από την ορθογωνιότητα αυτή προκύπτει το Πυθαγόρειο Θεώρημα, δηλ. ότι $\|Pf-f\|^2=\|f\|^2-\|Pf\|^2$. Άρα ισχύει

(28) $\|Pf\| \le \|f\| \quad \forall f\in C[a,b]$.

Για να εκτιμήσουμε το σφάλμα της L^2-προβολής ως πρός την νόρμα $\|.\|$, έχουμε από την (25) ότι

(29) $\|f-Pf\|=\min_{\varphi\in S_\tau^2}\|f-\varphi\| \le \|f-I_2f\|$,

από την οποία, χρησιμοποιώντας τα γνωστά μας φράγματα για το σφάλμα της παρεμβολής στην νόρμα $\|.\|$, (π.χ. για $f\in C^2[a,b]$ έχουμε από την (21) ότι $\|f-I_2f\| \le Ch^2\|f''\|$) βρίσκουμε ανάλογα φράγματα για το σφάλμα $\|f-Pf\|$.

Ο υπολογισμός της Pf γίνεται από τις κανονικές εξισώσεις (26) που προφανώς είναι ισοδύναμες με τις

(30) $(Pf,\ \varphi_i) = (f,\varphi_i)$, $1\le i\le N$,

όπου, φ_i, $1 \le i \le N$, οι συναρτήσεις βάσης που κατασκευάστηκαν στο Λήμμα 1. Υποθέτοντας ότι

$$(31) \quad Pf = \sum_{i=1}^{N} c_i \varphi_i,$$

βλέπουμε ότι οι συντελεστές c_i, $1 \le i \le N$, της Pf ικανοποιούν το $N \times N$ γραμμικό σύστημα

$$(32) \quad G\, c = \beta,$$

όπου $G \in R^{N \times N}$, ο λεγόμενος <u>πίνακας Gram</u> ή <u>πίνακας μάζας</u> της βάσης $\{\varphi_i\}$, ορίζεται ως

$$(33) \quad G_{ij} = (\varphi_i, \varphi_j) \quad 1 \le i, j \le N,$$

και είναι προφανώς συμμετρικός και θετικά ορισμένος. Στην (32) $c = [c_1, \ldots, c_N]^T$ και $\beta = [\beta_1, \ldots, \beta_N]^T$, όπου

$$(34) \quad \beta_i = (f, \varphi_i), \quad 1 \le i \le N.$$

Η σημασία του μικρού φορέα των συναρτήσεων βάσης φ_i είναι τώρα προφανής: Από την (33) βλέπουμε ότι επειδή $(\varphi_i, \varphi_j) = 0$ για $|i-j| > 1$, ο πίνακας G είναι <u>αραιός</u> και μάλιστα <u>τριδιαγώνιος</u>. Μερικοί απλοί υπολογισμοί των ολοκληρωμάτων

$\int_a^b \varphi_i^2 dx$, $\int_a^b \varphi_i \, \varphi_{i+1} dx$ δίνουν, για $h_j = x_{j+1} - x_j$, $1 \le j \le N-1$,

$$(35)\quad G= 1/6 \begin{bmatrix} 2h_1 & h_1 & & & O \\ h_1 & 2(h_1+h_2) & h_2 & & \\ & \ddots & \ddots & \ddots & \\ & & h_{N-2} & 2(h_{N-2}+h_{N-1}) & h_{N-1} \\ O & & & h_{N-1} & 2h_{N-1} \end{bmatrix}$$

Ο G έχει επίσης αυστηρά κυριαρχική διαγώνιο (άλλη μιά καλή ιδιότητα που οφείλεται και αυτή στην συγκεκριμένη μορφή των φ_i). Συνεπώς ο τριδιαγώνιος αλγόριθμος (είτε η ανάλυση Cholesky του G που οδηγεί σε <u>διδιαγώνιους</u> L,L^T) είναι ευσταθής. Το σύστημα (32) μπορεί να λυθεί συνεπώς με $O(N)$ πράξεις, αν είναι γνωστό το δεύτερο μέλος β.

Προχωρούμε τώρα στην διερεύνηση του σφάλματος της L^2-προβολής στην maximum νόρμα $\|.\|_\infty$. Σημαντικό στην κατεύθυνση αυτή είναι το εξής αποτέλεσμα:

Λήμμα 2. ("Ευστάθεια" της L^2-προβολής στην νόρμα $\|.\|_\infty$)

(36) $\|Pf\|_\infty \le 3\|f\|_\infty,\ \forall f\in C[a,b]$.

<u>Απόδειξη</u>: Για ευκολία εισάγουμε τον συμβολισμό $h_0=h_{N+1}=c_0=c_{N+1}=0$. Τότε οι σχέσεις (32) γράφονται (λόγω των (34), (35)) ως

(37) $(h_{i-1}/(h_i+h_{i-1}))c_{i-1}+2c_i+(h_i/(h_i+h_{i-1}))\,c_{i+1}=$

$= 6\beta_i/(h_i+h_{i-1}) \equiv \hat{\beta}_i,\ 1\le i\le N.$

Έστω $j(1\leq j\leq N)$ δείκτης για τον οποίο $|c_j|=\max_{0\leq i\leq N+1}|c_i|$. Τότε η (37) για $i=j$ δίνει

$$2|c_j|=|\hat{\beta}_j-(h_j+h_{j-1})^{-1}(h_{j-1}c_{j-1}+h_jc_{j+1})|\leq|\hat{\beta}_j|+$$

$$+(h_j+h_{j-1})^{-1}(h_{j-1}+h_j)\max(|c_{j-1}|,|c_{j+1}|)\leq|\hat{\beta}_j|+|c_j|,$$

δηλ. ότι

$$(38)\quad |c_j|=\max_i|c_i|\leq|\hat{\beta}_j|=6(h_j+h_{j-1})^{-1}\left|\int_a^b f\,\varphi_j dx\right|$$

$$\leq 6(h_j+h_{j-1})^{-1}\|f\|_\infty\int_a^b\varphi_j dx=3\|f\|_\infty.$$

Επειδή $Pf\in S_\tau^2$, $\|Pf\|_\infty=\max_{1\leq i\leq N}|(Pf)(x_i)|=\max_{1\leq i\leq N}|c_i|=|c_j|$.

Από την (38) προκύπτει συνεπώς η (36).@

Όπως έχουμε δει και προηγουμένως (βλ. και Άσκηση 5), η "ευστάθεια" (36) της L^2-προβολής στον S_τ^2 ως προς την νόρμα $\|.\|_\infty$ και το γεγονός ότι $P\varphi=\varphi$ για $\varphi\in S_\tau^2$ μας οδηγεί στην εξής εκτίμηση για κάθε $f\in C[a,b]$ και $\varphi\in S_\tau^2$:

$$\|Pf-f\|_\infty\leq\|Pf-\varphi\|_\infty+\|\varphi-f\|_\infty=\|P(f-\varphi)\|_\infty+\|f-\varphi\|_\infty\leq 4\|f-\varphi\|_\infty,$$

από την οποία τελικά έχουμε

$$(39)\quad \|Pf-f\|_\infty\leq 4\min_{\varphi\in S_\tau^2}\|f-\varphi\|_\infty(\text{π.χ. }\leq 4\|f-I_2f\|_\infty),$$

Έτσι, αν $f\in C^2[a,b]$ π.χ., παίρνουμε, λόγω της (19), ότι $\|Pf-f\|_\infty\leq ch^2\|f''\|_\infty$.

Παρατηρήσεις

1. Εστω ότι $f\in PC^2[a,b]$. Από το Θεώρημα του πυρήνα του Peano και την απόδειξη του Θεωρήματος 2 βλέπουμε ότι εξακολουθεί να ισχύει η παράσταση (13) για το σφάλμα $e(x)=f(x)-(I_2f)(x)$ της παρεμβολής, όπου ο πυρήνας $K(x,t)$ δίνεται πάλι από την (14). Συνεπώς ερμηνεύοντας με επέκταση κατά τον προφανή τρόπο τις νόρμες $\|f''\|_\infty$, $\|f''\|$ για $f\in PC^2[a,b]$ - και ακολουθώντας τα βήματα της απόδειξης του Πορίσματος 1, βλέπουμε ότι ισχύουν οι (19) και (21) και για $f\in PC^2[a,b]$. Για να αποδείξουμε φράγματα για το σφάλμα της παραγώγου, βλέπουμε τώρα από την απόδειξη του Θεωρήματος 1 ότι για $f\in PC^2[a,b]$, $x\in[x_i,x_{i+1}]$

$$(40)\quad e'(x)=\int_{x_i}^{x_{i+1}} \Lambda(x,t)f^{(2)}(t)dt+K(x,x)[f^{(2)}(x^-)-f^{(2)}(x^+)],$$

όπου ο πυρήνας Λ δίνεται από την (16) και ο K από την (14). (Ο $K(x,t)$ είναι συνεχής για $x\neq t$). Χρησιμοποιώντας την (40) εύκολα μπορούμε να δούμε ότι για $x\in[x_i,x_{i+1}]$

$$|e'(x)|\leq \max(\|f''\|_{L^\infty(x_i,x_{i+1})},|f^{(2)}(x^+)-f^{(2)}(x^-)|)h_i/2$$

από την οποία παίρνουμε π.χ. ότι $\|e'\|_\infty\leq h\|f''\|_\infty$. Τέλος, μιά ανισότητα της μορφής $\|e'\|\leq Ch\|f''\|$ (δηλ. το ανάλογο της (22) μπορεί να αποδειχθεί όπως στην Ασκηση 10β με $C=1$.

2. Ας εξετάσουμε τώρα περιπτώσεις και λιγότερο ομαλών f, υποθέτοντας απλώς ότι $f\in C[a,b]$. Από τον τοπικό ορισμό της συνάρτησης παρεμβολής $I_2 f$ στο $[x_i,x_{i+1}]$ εύκολα προκύπτει η παράσταση για $x\in[x_i,x_{i+1}]$

$$\varepsilon(x)=f(x)-(I_2 f)(x)=h_i^{-1}[(x_{i+1}-x)(f(x)-f(x_i))+(x-x_i)(f(x)-f(x_{i+1}))].$$

Συμπεραίνουμε λοιπόν ότι για $x\in[x_i,x_{i+1}]$, $h=\max_i h_i$

$$(41)\ |\varepsilon(x)|\leq|h_i^{-1}[(x_{i+1}-x)+(x-x_i)]|\max_{x\in[x_i,x_{i+1}]}(|f(x)-f(x_i)|,|f(x_{i+1})-f(x)|)$$

$$\leq\omega(f;h_i)\leq\omega(f;h).$$

Συνεπώς επειδή $f\in C[a,b]$, $\|\varepsilon\|_\infty\leq\omega(f;h)\to 0$, $h\to 0$, και μάλιστα $\|\varepsilon\|_\infty=O(h)$ αν η f είναι Lipschitz με ειδική περίπτωση το φράγμα $\|\varepsilon\|_\infty\leq\|f'\|_\infty h$ για $f\in C^1[a,b]$. (Το θεώρημα του Πυρήνα του Peano δίνει όμως το καλύτερο φράγμα $\|\varepsilon\|_\infty\leq\|f'\|_\infty h/2$, $f\in PC^1[a,b]$, βλ. Ασκ. 4). Εξ άλλου, το θεώρημα της μέσης τιμής δίνει π.χ. $\|\varepsilon'\|_\infty\to 0$, $h\to 0$ αν $f\in C^1[a,b]$. Η τάξη σύγκλισης όμως μπορεί μερικές φορές να βελτιωθεί με κατάλληλη επιλογή των x_i (πύκνωση κοντά στις ιδιομορφίες της f). Βλ. π.χ. [4.2, σελ. 46].

3. Οι σταθερές (1/8 και 1/2) και οι εκθέτες του h (2 και 1) στις σχέσεις (19) και (20) είναι οι καλύτεροι δυνατοί για $f\in C^2[a,b]$ π.χ.. Αντίθετα οι σταθερές στα φράγματα (21) και (22) (αλλά όχι όμως οι εκθέτες!) μπορούν να βελτιωθούν: Οι καλύτερες σταθερές στα φράγματα $\|\varepsilon^{(k)}\|\leq C_k h^{2-k}\|f''\|$, $k=0,1$, $f\in C^2[a,b]$ μπορεί να δειχθεί ότι είναι $C_0=\pi^{-2}$, $C_1=\pi^{-1}$, βλ. π.χ. [4.10, παρ. 2.3].

Ασκήσεις 4.2

1(α) Επιβεβαιώστε τους ισχυρισμούς περί της L^2-προβολής στον S_τ^2 στο κείμενο και τους τύπους (25)-(28).

(β) Από την ανισότητα $\max_i|c_i|\leq\max_i|\hat{\beta}_i|$ που προκύπτει από την (38), συμπεράνετε ότι το σύστημα (32) έχει μοναδική λύση.

2. Αποδείξτε τους ισχυρισμούς της Παρατήρησης 1.

3. Αποδείξτε ότι η συνάρτηση παρεμβολής $I_2 f$ μιάς συνεχούς συνάρτησης f διατηρεί ιδιότητες μονοτονίας ή κυρτότητας της f.

4. Χρησιμοποιώντας το Θεώρημα του πυρήνα του Peano δείξτε ότι

$$\|f-I_2f\|_\infty\leq h\|f'\|_\infty/2 \text{ για } f\in PC^1[a,b].$$

5. Για g∈X, όπου X κάποιος διανυσματικός χώρος συναρτήσεων, έστω Πg προσέγγιση του g σ'ένα υπόχωρο S του X με τις ιδιότητες:

(i) $\Pi\varphi=\varphi,\ \forall\varphi\in S$
(ii) $\Pi(f+g)=\Pi f+\Pi g,\ \forall f,g\in X$
(iii) $|||\Pi|||<\infty$, όπου $|||\Pi||| = \sup_{0\neq g\in X} |||\Pi g|||/|||g|||$,

και όπου $|||\cdot|||$ κάποια νόρμα του X.

Δείξτε τότε ότι για κάθε g∈X

$$|||g-\Pi g|||\leq(1+|||\Pi|||)\inf_{\varphi\in S}|||g-\varphi|||.$$

6. Εστω S_τ^1 ο διανυσματικός χώρος των τμηματικά σταθερών συναρτήσεων στο [a,b] ως προς τον διαμερισμό $\tau: a=x_1<x_2<\ldots<x_N=b$, δηλ. έστω

$S_\tau^1=\{\varphi:\varphi(x)=c_i \quad$ για $x\in[x_i,x_{i+1}],\ 1\le i\le N-1\}$.

(Για να αποφύγουμε ορισμούς των $\varphi\in S^1$ στους κόμβους ας συμβολίσουμε $\|\varphi\|_\infty=\max\{|\varphi(x)|:\ x\in[a,b]\setminus\{x_2,\dots,x_{N-1}\}\}$)

(α) Δείξτε ότι για $g\in C[a,b]$

$$\min_{\varphi\in S_\tau^1}\|g-\varphi\|\le\omega(g;h/2),$$

όπου κατά τα γνωστά $h=\max_i h_i$.

(β) Δείξτε ότι $\|g'-(I_2g)'\|_\infty\le 2\min_{\varphi\in S_\tau}\|g'-\varphi\|$

για κάθε $g\in C^1[a,b]$ και μάλιστα ότι $\|g'-(I_2g')\|_\infty=O(h)$ για Lipschitz g'.

(γ) ("Υπερσύγκλιση" στον S_τ^1). Εστω $f\in PC^1[a,b]$ και θεωρείστε την συνάρτηση $I_1f\in S_\tau^1$ που ορίζεται ως

$(I_1f)(x)=f((x_i+x_{i+1})/2)$ για $x\in[x_i,x_{i+1}]$, $1\le i\le N-1$. Δείξτε ότι

$$\|f-I_1f\|_\infty\le h\|f'\|_\infty/2.$$

7. ("Υπερσύγκλιση"" για παραγώγους στον S_τ^2). Υποθέστε ότι $f\in C^3[a,b]$. Ξεκινώντας από την παράσταση (15), (16) του σφάλματος $\varepsilon'(x)=f'(x)-(I_2f)'(x)$, $\quad x\in[x_i,x_{i+1}]$ και χρησιμοποιώντας ολοκλήρωση κατά μέρη δείξτε ότι

$$|\epsilon'((x_i+x_{i+1})/2)| \le h^2\|f'''\|_\infty/6, \qquad 1\le i\le N-1,$$

δηλ. ότι υπάρχουν σημεία (τα μέσα των διαστημάτων $[x_i,x_{i+1}]$) στο $[a,b]$ όπου το σημειακό σφάλμα της παραγώγου $(I_2f)'$ είναι της τάξης $O(h^2)$, (για $f\in C^3[a,b]$), ενώ, όπως ξέρουμε, ισχύει γενικά ότι

$\|\epsilon'\|_\infty \le h\|f''\|_\infty/2$ έστω και για $f\in C^\infty[a,b]$.

8. Δείξτε ότι για κάθε $f\in PC^1[a,b]$ ισχύει ότι

(α) $((I_2f)',\varphi') = (f',\varphi') \qquad \forall\varphi\in S_h$.

(β) $\|(I_2f)'\|^2 + \|(I_2f-f)'\|^2 = \|f'\|^2$

(γ) Αν $f\in PC^2[a,b]$ δείξτε ότι $\|(f-I_2f)'\|^2=(f-I_2f,f'')$

9. (Ανισότητα των Poincaré-Friedrichs). Δείξτε ότι για κάθε $f\in PC^2[c,d]$ τέτοια ώστε $f(c)=0$ ή $f(d)=0$ ισχύει

$$\int_c^d f^2(x)dx \le (d-c)^2 \int_c^d (f'(x))^2\,dx.$$

10(α). Χρησιμοποιώντας τα αποτελέσματα των ασκήσεων 8 και 9 δείξτε ότι αν $f\in PC^1[a,b]$, τότε

$$\|(f-I_2f)'\| \le \|f'\|$$

$$\|(f-I_2f)\| \le h\|f'\|$$

(β) Αν $PC^2[a,b]$ δείξτε, χρησιμοποιώντας τα παραπάνω ότι

$$\|(f-\overset{*}{I}_2 f)'\| \leq h\|f''\| \qquad \text{και}$$

$$\|f-I_2 f\| \leq h^2 \|f''\|$$

(Στην άσκηση αυτή μην χρησιμοποιείσετε το Θεώρημα του Πυρήνα του Peano).

1.3 ΠΑΡΕΜΒΟΛΗ ΜΕ ΤΜΗΜΑΤΙΚΑ ΚΥΒΙΚΕΣ ΣΥΝΑΡΤΗΣΕΙΣ HERMITE

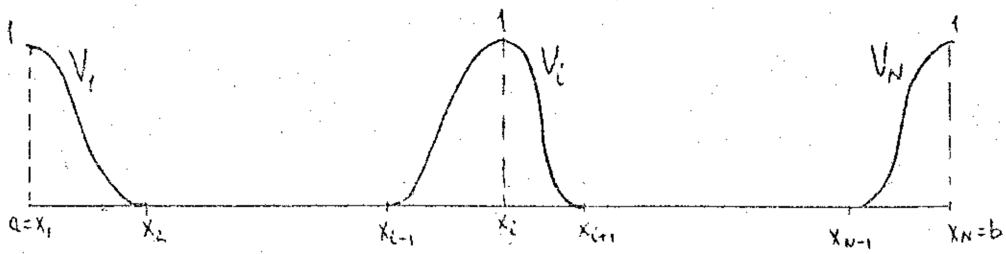
Στην παράγραφο αυτή και στην επόμενη θα ασχοληθούμε με παρεμβολή σε χώρους τμηματικά __κυβικών__ συναρτήσεων. Υπάρχουν αρκετοί τρόποι με τους οποίους μπορεί να γίνει παρεμβολή σε τέτοιους χώρους. Θα αρχίσουμε εδώ με την λεγόμενη παρεμβολή __Hermite__ (που χρησιμοποιεί δηλ. εκτός από τις τιμές $f(x_i)$ της συνάρτησης στους κόμβους και τις τιμές της παραγώγου της $f'(x_i)$) σε χώρους τμηματικά κυβικών συναρτήσεων που ανήκουν στον χώρο $C^1[a,b]$.

Για δεδομένο διαμερισμό τ: $a = x_1<x_2<\ldots<x_N= b$ του $[a,b]$ θεωρούμε τον διανυσματικό χώρο των (τμηματικά) __κυβικών συναρτήσεων Hermite__.

$$H_\tau=\{\varphi:\ \varphi\in C^1[a,b],\ \varphi\in P_3(x_i,x_{i+1}),\ 1\le i\le N-1\}$$

Λήμμα 1. Ο H_τ είναι $2N$-διάστατος υπόχωρος του $C^1[a,b]$. Οι συναρτήσεις $\{V_i(x),\ S_i(x)\}$, $1\le i\le N$, που δίνονται από τους τύπους (1), (2), παρακάτω, αποτελούν βάση του H_τ.

__Απόδειξη.__ Προφανώς ο H_τ είναι διανυσματικός χώρος, υπόχωρος του $C^1[a,b]$. Συμβολίζοντας πάλι $h_i=x_{i+1}-x_i$, $1\le i\le N-1$, θεωρούμε τις συναρτήσεις $V_i, S_i\in H_\tau$, $1\le i\le N$, που δίνονται από

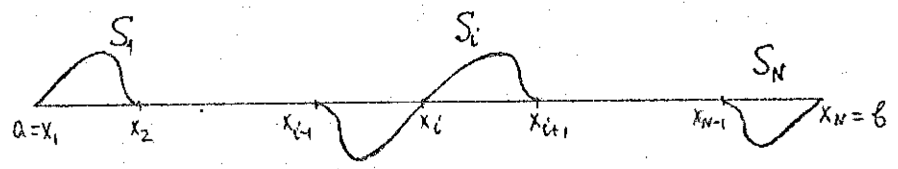


$$(1)\quad V_1(x) = \begin{cases} 2h_1^{-3}(x-x_1)^3-3h_1^{-2}(x-x_1)^2+1 & \text{αν } x\in[x_1,x_2] \\ 0 & \text{αλλοιώς} \end{cases}$$

Για $2\leq i\leq N-1$

$$U_i(x)=\begin{cases}-2h_{i-1}^{-3}(x-x_{i-1})^3+3h_{i-1}^{-2}(x-x_{i-1})^3 & x\in[x_{i-1},x_i]\\ 2h_i^{-3}(x-x_i)^3-3h_i^{-2}(x-x_i)^2+1 & x\in[x_i,x_{i+1}]\\ 0 & \text{αλλοιώς}\end{cases}$$

$$U_N(x)=\begin{cases}-2h_{N-1}^{-3}(x-x_{N-1})^3+3h_{N-1}^{-2}(x-x_{N-1})^2 & x\in[x_{N-1},x_N]\\ \\ 0 & \text{αλλοιώς}\end{cases}$$



και

$$(2)\quad S_1(x)=\begin{cases}h_1^{-2}(x-x_1)(x_2-x)^2 & x\in[x_1,x_2]\\ \\ 0 & \text{αλλοιώς}\end{cases}$$

Για $2\leq i\leq N-1$

$$S_i(x)=\begin{cases}h_{i-1}^{-2}(x-x_{i-1})^2\,(x-x_i) & x\in[x_{i-1},x_i]\\ h_i^{-2}(x-x_i)\,(x_{i+1}-x)^2 & x\in[x_i,x_{i+1}]\\ 0 & \text{αλλοιώς}\end{cases}$$

$$S_N(x)=\begin{cases}h_{N-1}^{-2}(x-x_{N-1})^2(x-x_N) & x\in[x_{N-1},x_N]\\ \\ 0 & \text{αλλοιώς.}\end{cases}$$

Είναι προφανές ότι $S_i, U_i \in H_\tau$, $1\le i\le N$ και ότι ικανοποιούν τις σχέσεις

$$(3)\quad \begin{cases} U_i(x_j)=\delta_{ij}, & U_i'(x_j)=0 \\ S_i(x_j)=0, & S_i'(x_j)=\delta_{ij} \end{cases} \quad ,1\le i,j\le N.$$

Επιπλέον οι S_i, U_i είναι οι συναρτήσεις του H_τ με τον μικρότερο δυνατό φορέα που ικανοποιούν τις (3). Επίσης είναι <u>γραμμικά ανεξάρτητες</u>. Πράγματι, αν $\sum_{j=1}^{N} c_j U_j(x) + d_j S_j(x) = 0$ $\forall x \in [a,b]$ θα έχουμε και $\sum_{j=1}^{N} c_j U_j'(x) + d_j S_j'(x) = 0$ $\forall x \in [a,b]$.

Θέτοντας $x=x_i$, $1\le i\le N$ στην πρώτη σχέση παίρνουμε $c_i=0$, $1\le i\le N$, ενώ από την δεύτερη $d_i=0$, $1\le i\le N$. Είναι εύκολο επίσης να δούμε ότι το σύνολο $\{U_i, S_i\}$ <u>παράγει</u> τον H_τ. Πράγματι, για $\varphi \in H_\tau$ ισχύει

$$(4)\quad \varphi(x) = \sum_{i=1}^{N} \varphi(x_i)\, U_i(x) + \varphi'(x_i) S_i(x),$$

όπως εύκολα μπορούμε να δούμε. Για $x=x_k$, $1\le k\le N$, οι τιμές και οι τιμές των παραγώγων των δύο μελών της (4) προφανώς συμπίπτουν. Σε οποιοδήποτε διάστημα $[x_k, x_{k+1}]$, $1\le k\le N-1$, και τα δύο μέλη της (4) είναι λοιπόν κυβικά πολυώνυμα των οποίων οι τιμές και οι τιμές των παραγώγων συμπίπτουν στα άκρα x_k, x_{k+1} του διαστήματος. Συνεπώς θα συμπίπτουν και για κάθε $x \in [x_k, x_{k+1}]$ επειδή κάθε κυβικό πολυώνυμο ορίζεται μονοσήμαντα από τις τιμές του και τις τιμές της παραγώγου του σε δύο διακριτά σημεία. @

Η σχέση (4) μας λέει ότι μπορούμε να ορίσουμε μονοσήμαντα οποιαδήποτε συνάρτηση $\varphi \in H_\tau$ αν δίδονται οι τιμές $\varphi(x_i), \varphi'(x_i)$. Για οποιαδήποτε λοιπόν συνάρτηση $f \in C^1[a,b]$ π.χ., μπορούμε να κατασκευάσουμε την <u>συνάρτηση παρεμβολής</u> της $I_H f$ στον χώρο H_τ ως το (μοναδικό) στοιχείο του H_τ πού ικανοποιεί τις σχέσεις

(5) $(I_H f)(x_i) = f(x_i)$, $(I_H f)'(x_i) = f'(x_i)$, $1 \le i \le N$.

Η κατασκευή της $I_H f$ απαιτεί γνώση των τιμών $f(x_i)$, $f'(x_i)$· μάλιστα, λόγω της (4) η $I_H f$ δίνεται από

$$(6)\quad (I_H f)(x) = \sum_{i=1}^{N} f(x_i) V_i(x) + f_i'(x) S_i(x)$$

Χρησιμοποιώντας το θεώρημα του Peano είναι δυνατόν να εκτιμήσουμε το σφάλμα της προσέγγισης $f - I_H f$ και των παραγώγων της ως πρός διάφορες νόρμες συναρτήσεων στο $[a,b]$. Θα αποδείξουμε εδώ μέρος του παρακάτω θεωρήματος:

ΘΕΩΡΗΜΑ 1. Εστω $f \in C^4[a,b]$ και $h = \max_i h_i$. Τότε

$$(7)\quad \|f - I_H f\|_\infty \le h^4 \|f^{(4)}\|_\infty / 384$$

$$(8)\quad \|(f - I_H f)'\|_\infty \le (\sqrt{3}/216)\, h^3 \|f^{(4)}\|_\infty$$

$$(9)\quad \|(f - I_H f)^{(2)}\|_\infty \le h^2 \|f^{(4)}\|_\infty / 12$$

$$(10)\quad \|(f - I_H f)^{(3)}\|_\infty \le h \|f^{(4)}\|_\infty / 2$$

όπου η δεύτερη και η τρίτη "παράγωγος" της $I_H f$ έχουν υπολογισθεί κατά τμήματα, δηλ. όπου στις (9) και (10) εννοούμε ότι για $k=2,3$:

$$\|(f - I_H f)^{(k)}\|_\infty = \max_{1 \le i \le N-1} \Big(\max_{x_i \le x \le x_{i+1}} |(f - I_H f)^{(k)}(x)|\Big)$$

<u>Απόδειξη.</u> Θα αποδείξουμε την (7) και θα σχολιάσουμε απλώς τις (8)-(10) παραπέμποντας για την απόδειξη στην βιβλιογραφία.

Σταθεροποιούμε ένα $x \in (x_i, x_{i+1})$, $1 \le i \le N-1$. Το σφάλμα τότε της παρεμβολής Hermite $\varepsilon(x) = \varepsilon_f(x) = f(x) - I_H f(x)$ είναι γραμμικό συναρτησιακό στον χώρο $C^4[a,b]$. Επιπλέον, επειδή $p = I_H p$ $\forall p \in P_3(x_i, x_{i+1})$, από το θεώρημα του πυρήνα του Peano 4.2.1 παίρνουμε την παράσταση

$$(11) \quad \varepsilon(x) = (3!)^{-1} \varepsilon_x \left[\int_{x_i}^{x_{i+1}} (x-t)_+^3 f^{(4)}(t)\, dt \right]$$

όπου ανακαλούμε ότι

$$(x-t)_+^3 = \begin{cases} (x-t)^3 & \text{αν } x \ge t \\ 0 & \text{αν } x < t \end{cases} .$$

Από τον ορισμό του τελεστού I_H της παρεμβολής βλέπουμε ότι οι πράξεις ε_x και $\int_{x_i}^{x_{i+1}} dt$ αντιμετατίθενται, δηλ. ότι

$$(12) \quad \varepsilon(x) = 6^{-1} \int_{x_i}^{x_{i+1}} \varepsilon_x[(x-t)_+^3] f^{(4)}(t)\, dt = 6^{-1} \int_{x_i}^{x_{i+1}} K(x,t) f^{(4)}(t)\, dt,$$

όπου

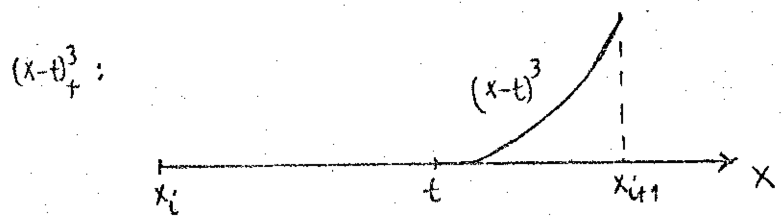
$$(13) \quad K(x,t) = (x-t)_+^3 - I_{H,x}[(x-t)_+^3], \quad x_i \le x, t \le x_{i+1}.$$

Από τις (12),(13) συμπεραίνουμε ότι

$$(14) \quad |\varepsilon(x)| \le \| f^{(4)} \|_\infty \int_{x_i}^{x_{i+1}} |K(x,t)|\, dt \,/6.$$

Συνεπώς πρέπει να βρούμε μιά όσο το δυνατό καλύτερη εκτίμηση του ολοκληρώματος στο δεύτερο μέλος της (14).

Για να υπολογίσουμε τον πυρήνα $K(x,t)$ από την (13) πρέπει να βρούμε την συνάρτηση $I_{H,x}[(x-t)_+^3]$, για $x\in[x_i,x_{i+1}]$, δηλ. το κυβικό πολυώνυμο $\pi(x)$ του οποίου η τιμή και η τιμή της παραγώγου συμπίπτουν με τις αντίστοιχες τιμές της συνάρτησης $(x-t)_+^3$, θεωρούμενης ως συνάρτησης του x, στα σημεία $x=x_i$ και $x=x_{i+1}$,



δηλ. το κυβικό πολυώνυμο $\pi(x)$ που ικανοποιεί $\pi(x_i)=\pi'(x_i)=0$, $\pi(x_{i+1})=(x_{i+1}-t)^3$, $\pi'(x_{i+1})=3(x_{i+1}-t)^2$. Εύκολα βρίσκουμε ότι

$$(15)\quad \pi(x)=I_{H,x}[(x-t)_+^3]\equiv\Lambda(x,t)=h_i^{-3}(x-x_i)^2(x_{i+1}-t)^2[x(x_{i+1}-3x_i+2t) +tx_i-3tx_{i+1}+2x_ix_{i+1}],\quad x_i\le x,t\le x_{i+1}.$$

Άρα ο πυρήνας $K(x,t)$, λόγω των (13), (15) γίνεται

$$(16)\quad K(x,t)=\begin{cases}(x-t)^3-\Lambda(x,t), & x_i\le t\le x\le x_{i+1}\\ \\ -\Lambda(x,t), & x_i\le x\le t\le x_{i+1}.\end{cases}$$

Από τις (15),(16) βλέπουμε ότι $K(x,t)\ge 0\ \ \forall x,t\in[x_i,x_{i+1}]$: Ας εξετάσουμε πρώτα την περίπτωση $x_i<x<t<x_{i+1}$. Για τέτοια x,t έχουμε, λόγω των (15),(16) ότι $\operatorname{sgn} K(x,t)=-\operatorname{sgn}\Lambda(x,t)=-\operatorname{sgn}M(x,t)$, όπου $M(x,t)=x(x_{i+1}-3x_i+2t)+tx_i-3tx_{i+1}+2x_ix_{i+1}$.

Η $M(x,t)$ είναι γραμμική συνάρτηση του t και ικανοποιεί $M(x,x)=-2h_i(x-x_i)<0$ και $M(x,x_{i+1})=-3(x_{i+1}-x)h_i<0$. Άρα $M(x,t)<0$ για $x\leq t\leq x_{i+1}$. Συμπεραίνουμε ότι $K(x,t)\geq 0$ για $x_i\leq x\leq t\leq x_{i+1}$. Μιά σειρά από απλούς υπολογισμούς δίνει επίσης ότι $K(x,t)\geq 0$ για $x_i\leq x\leq t\leq x_{i+1}$. Συνεπώς έχουμε

$$\int_{x_i}^{x_{i+1}} |K(x,t)|dt=\int_{x_i}^{x_{i+1}} K(x,t)dt=\int_{x_i}^{x} [(x-t)^3-\Lambda(x,t)]dt+\int_{x}^{x_{i+1}} [-\Lambda(x,t)]dt$$

$$=\int_{x_i}^{x} (x-t)^3dt-\int_{x_i}^{x_{i+1}} \Lambda(x,t)dt = (x-x_i)^4/4+(x-x_i)^2\,[4(xx_{i+1}-3xx_i+2x_ix_{i+1})$$

$$+(2x+x_i-3x_{i+1})(x_{i+1}+3x_i)]/12 = (x-x_i)^2(x_{i+1}-x)^2/4.$$

Άρα η (14) δίνει

$$(17)\quad |\varepsilon(x)| \leq \|f^{(4)}\|_\infty (x-x_i)^2(x_{i+1}-x)^2/24, \quad x \in [x_i,x_{i+1}].$$

Το μέγιστο της συνάρτησης $(x-x_i)^2\,(x_{i+1}-x)^2$, $x\in[x_i,x_{i+1}]$, προφανώς λαμβάνεται για $x=(x_i+x_{i+1})/2$ και είναι ίσο με $h_i^4/16$. Συνεπώς η (7) έπεται από την (17). Οπως δείχνει η συνάρτηση $f(x)=(x-x_i)^2(x_{i+1}-x)^2$ (γιά την οποία $I_Hf=0$ για $x\in[x_i,x_{i+1}]$, $f^{(4)}(x)=24$), η (17) - και συνεπώς και η (7) - δέν είναι δυνατόν να βελτιωθεί.

Για να αποδείξουμε τις (8)-(10) παρατηρούμε πρώτα (Ασκ. 1β) ότι για $x \in (x_i, x_{i+1})$

$$(18)\quad e^{(k)}(x) = 6^{-1} \int_{x_i}^{x_{i+1}} \partial_x^k [K(x,t)]\, f^{(4)}(t)dt,\ k=1,2,3,$$

όπου με $\partial_x^k[K(x,t)]$, $1 \le k \le 3$, συμβολίζουμε την μερική παράγωγο k τάξεως ως προς x της συνάρτησης $K(x,t)$, ορισμένη τμηματικά λόγω της (16) για $x<t$ και $x>t$. Δεν ισχύει όμως πιά ότι οι συναρτήσεις $\partial_x^k[K(x,t)]$ έχουν

σταθερό πρόσημο για $x,t \in (x_i, x_{i+1})$. Οι Birkhoff και Priver (J.Math. and Phys. 46 (1967), 440-447), βρίσκουν τις ρίζες καθώς και τα σημεία όπου οι $\partial_x^k[K(x,t)]$ αλλάζουν ασυνεχώς πρόσημο (ως συναρτήσεις του t). Τα ολοκληρώματα $\int_{x_i}^{x_{i+1}} |\partial_x^k[K(x,t)]|dt$ γράφονται κατόπιν ως αθροίσματα ολοκληρωμάτων πάνω σε διαστήματα όπου οι $\partial_x^k[K(x,t)]$ έχουν σταθερό πρόσημο και υπολογίζονται ακριβώς. Προκύπτουν οι (8)-(10). Μπορεί να αποδειχθεί ότι οι σταθερές και οι δυνάμεις του h στις (7)-(10) είναι οι καλύτερες δυνατές για $f \in C^k[a,b]$, $k \ge 4$.@

<u>Παρατηρήσεις</u>

1. Μπορούμε να δείξουμε βέβαια και φράγματα σφαλμάτων στην L^2-νόρμα $\|f\| = (\int_a^b f^2(x)dx)^{1/2}$. Π.χ. ισχύει (βλ. Ασκ. 4)

$$\|(f-I_H f)^{(k)}\| \le C_k h^{2-k} \|f''\|,\ k=0,1,2,\ \text{αν } f \in PC^2[a,b]$$

και

$$\|(f-I_H f)^{(k)}\| \le C_k^* h^{4-k} \|f^{(4)}\|, k=0,1,2,\ \text{αν } f \in PC^4[a,b].$$

Οι καλύτερες σταθερές είναι $C_0=\pi^{-2}, C_1=\pi^{-1}, C_2=1, C_0^*=\pi^{-4}, C_1^*=\pi^{-3}, C_2^*=\pi^{-2}$.

Ανάλογες εκτιμήσεις ισχύουν αν π.χ. $f \in PC^3[a,b]$ και φυσικά και στην νόρμα $\|\cdot\|_\infty$ αν η f είναι λιγώτερο ομαλή από C^4.

2. <u>Κυβική παρεμβολή Bessel</u>. Μία παραλλαγή της παρεμβολής με κυβικές συναρτήσεις Hermite είναι η λεγόμενη <u>κυβική παρεμβολή Bessel</u>, την οποία μπορούμε να χρησιμοποιήσουμε αν γνωρίζουμε μόνο τις τιμές της συνάρτησης f στα x_i, $1\le i\le N$ και τις τιμές της παραγώγου f' μόνο στα a και στα b. Αντί $f'(x_i)$, $2\le i\le N-1$ δηλ., στον τύπο (6) χρησιμοποιούμε αριθμούς s_i που είναι οι κλίσεις στα σημεία x_i του πολυωνύμου παρεμβολής βαθμού ≤ 2 που συμφωνεί με τις τιμές της f στα σημεία x_{i-1},x_i,x_{i+1}. Εύκολα βλέπουμε ότι οι κλίσεις αυτές δίνονται από τους τύπους

$$s_i = (h_i f[x_{i-1},x_i] + h_{i-1}f[x_i,x_{i+1}])/(h_i+h_{i-1}),\ 2\le i\le N-1.$$

Η κυβική παρεμβολή Bessel είναι και αυτή <u>τοπική</u> μέθοδος, όπως και η παρεμβολή Hermite. Δηλ. για τον προσδιορισμό της τιμής της συνάρτησης παρεμβολής $(If)(x)$ σ' ένα σημείο $x\in[x_i,x_{i+1}]$ χρησιμοποιούνται πληροφορίες για τις τιμές της f (ή και της f') σε λίγα σημεία x_k κοντά στο x. Οπως αποδεικνύεται όμως (Ασκ. 2) η κυβική παρεμβολή Bessel έχει γενικά σφάλμα της τάξης $O(h^3)$ και όχι $O(h^4)$ όπως η παρεμβολή Hermite.

Ασκήσεις 4.3

1(α) Δείξτε ότι ο πυρήνας $K(x,t)$ που ορίζεται από την (16) είναι μη αρνητικός και για $x_i\le t\le x\le x_{i+1}$.

(β) Δείξτε ότι ισχύει η (18) για $f\in C^4[a,b]$.

2. Εστω f μία αρκετά ομαλή συνάρτηση στο $[a,b]$ και έστω f_h οποιαδήποτε τμηματικά (ως προς οποιοδήποτε διαμερισμό τ του $[a,b]$) κυβική συνάρτηση παρεμβολής της f, δηλ. έστω $f_h \in P_3(x_i,x_{i+1})$, $1 \leq i \leq N-1$ με $f_h(x_i)=f(x_i)$, $1 \leq i \leq N$.

(α) Χρησιμοποιώντας το γεγονός ότι ένα κυβικό πολυώνυμο στο $[x_i,x_{i+1}]$ προσδιορίζεται μονοσήμαντα από τις τιμές του και τις τιμές της παραγώγου του στα άκρα x_i, x_{i+1} του διαστήματος, δείξτε ότι το σφάλμα της f_h γράφεται ως

$$e = f - f_h = f - I_H f + \epsilon$$

όπου $I_H f$ η συνάρτηση παρεμβολής Hermite της f και όπου

$$\epsilon(x) = h_i^{-2}[e'(x_i)(x_{i+1}-x) - e'(x_{i+1})(x-x_i)](x-x_i)(x_{i+1}-x),$$

$$x_i \leq x \leq x_{i+1}, \quad 1 \leq i \leq N-1.$$

(β) Αποδείξτε ότι για αρκετά ομαλή f έχουμε για $h = \max_i (x_{i+1}-x_i)$

$$\|f-f_h\|_\infty \leq \|f-I_H f\|_\infty + h \max_i |e'(x_i)|/4, \text{ δηλ. ότι}$$

$$\|f-f_h\|_\infty = O(h^4) + \max_i |e'(x_i)| O(h).$$

(γ) Εστω f_h η κυβική συνάρτηση παρεμβολής Bessel για την f με $f_h'(a)=f'(a)$, $f_h'(b)=f'(b)$ (β λ. παρατήρηση 2). Δείξτε ότι $\|f-f_h\|_\infty = O(h^3)$ και ότι $\|f-f_h\|_\infty = O(h^4)$ για <u>ομοιόμορφο</u> διαμερισμό, υποθέτοντας ότι η f είναι όσο ομαλή χρειάζεται.

3.(α) Αν $f \in PC^2[a,b]$ δείξτε ότι

$$((I_H f-f)'', \varphi'') = 0, \quad \forall \varphi \in H_\tau .$$

(β) Αν $f \in PC^2[a,b]$ τότε

$$\|(I_H f)''\|^2 + \|(I_H f-f)''\|^2 = \|f''\|^2$$

(γ) Αν $f \in PC^4[a,b]$ τότε $\|(f-I_H f)''\|^2 = (f-I_H f, f^{(4)})$.

4. Χρησιμοποιώντας τις "ολοκληρωτικές ταυτότητες" της Άσκησης 3 και την ανισότητα του Poincaré (Άσκηση 4.2.9) δείξτε,

$$\|(f-I_H f)^{(k)}\| \le h^{2-k}\|f^{(2)}\|, \; k=0,1,2 \qquad \text{αν } f \in PC^2[a,b]$$

και

$$\|(f-I_H f)^{(k)}\| \le h^{4-k}\|f^{(4)}\|, \; k=0,1,2 \qquad \text{αν } f \in PC^4[a,b].$$

4.4 ΠΑΡΕΜΒΟΛΗ ΜΕ ΚΥΒΙΚΕΣ SPLINES

Οι τμηματικά κυβικές συναρτήσεις Hermite που εξετάσαμε στην προηγούμενη παράγραφο αποτελούν διανυσματικό χώρο διαστάσεως 2N για δεδομένο διαμερισμό τ του $[a,b]$ με N διακριτά σημεία x_i. Συνεπώς ο αριθμός των παραμέτρων που ορίζουν μία συνάρτηση του H_τ είναι αρκετά μεγάλος σε σχέση π.χ. με τις N παραμέτρους που προσδιορίζουν μιά συνάρτηση του S_τ^2. (Οι δυσκολίες αυξάνουν αν θεωρήσουμε τα πολυδιάστατα ανάλογα αυτών των χώρων για προσέγγιση συναρτήσεων πολλών μεταβλητών). Επιπλέον η κατασκευή της συνάρτησης παρεμβολής $I_H f$ απαιτεί, εκτός των $f(x_i)$, και τις κλίσεις $f'(x_i)$ που πολλές φορές είναι δύσκολο ή αδύνατο να υπολογισθούν ακριβώς σε πρακτικά προβλήματα.

Τίθεται λοιπόν το ερώτημα αν μπορούμε να κατασκευάσουμε ένα χώρο τμηματικά κυβικών συναρτήσεων μέ όσο το δυνατόν μικρότερη διάσταση στον οποίο να μπορούμε να λύσουμε αποτελεσματικά το πρόβλημα της παρεμβολής χρησιμοποιώντας εί δυνατόν μόνο τιμές της συνάρτησης $f(x_i)$ και διατηρώντας σφάλματα παρεμβολής τάξης $O(h^4)$ για ομαλές f. Η απάντηση είναι καταφατική και δίνεται από τον χώρο των λεγόμενων κυβικών splines που ορίζεται, για τον διαμερισμό τ: $a=x_1<x_2<\dots<x_N=b$, ως το σύνολο

$$S_\tau^4 = \{\varphi:\ \varphi\in C^2[a,b],\ \varphi\in P_3(x_i,x_{i+1}),\ 1\le i\le N-1\},$$

δηλ. ως το σύνολο $S_\tau^4=H_\tau\cap C^2[a,b]$.

Αρχίζουμε μελετώντας το πρόβλημα της παρεμβολής στον S_τ^4. Εστω λοιπόν μία συνάρτηση $f\in C[a,b]$, έστω $N>2$ και ας δεχθούμε ότι υπάρχει συνάρτηση $s(x)=(I_4 f)(x)\in S_\tau^4$ τέτοια ώστε $s(x_i)=f(x_i)$, $1\le i\le N$. Θα δείξουμε ότι τότε ισχύουν οι σχέσεις

(1) $h_j s'(x_{j-1})+2(h_j+h_{j-1})s'(x_j)+h_{j-1}s'(x_{j+1})=$

$$=3[h_{j-1}(f(x_{j+1})-f(x_j))/h_j+h_j(f(x_j)-f(x_{j-1}))/h_{j-1}],\ 2\le j\le N-1,$$

όπου, όπως και προηγουμένως, χρησιμοποιούμε τον συμβολισμό $h_j=x_{j+1}-x_j$, $1\le j\le N-1$.

Πράγματι, όπως εύκολα βλέπουμε, το μοναδικό πολυώνυμο $p\in P_3[\lambda,\mu]$ που για δεδομένα u_1,u_2,u_1',u_2' ικανοποιεί τις συνθήκες $p(\lambda)=u_1, p(\mu)=u_2$, $p'(\lambda)=u_1', p'(\mu)=u_2'$, δίνεται, αν $h=\mu-\lambda$, από τον τύπο

$$p(x)=u_1[h^{-2}(x-\mu)^2+2h^{-3}(x-\lambda)(x-\mu)^2]+u_2[h^{-2}(x-\lambda)^2-2h^{-3}(x-\lambda)^2(x-\mu)]$$

$$+u_1'h^{-2}(x-\lambda)(x-\mu)^2+u_2'h^{-2}(x-\lambda)^2(x-\mu).$$

(Το $p(x)$ είναι το κυβικό πολυώνυμο Hermite με δύο σημεία στο διάστημα $[\lambda,\mu]$). Συνεπώς, αν γνωρίζαμε τις τιμές $s'(x_j), s'(x_{j+1})$ θα μπορούσαμε να κατασκευάσουμε την $s(x)$ από τον τύπο

(2) $s(x)=f(x_j)[h_j^{-2}(x-x_{j+1})^2+2h_j^{-3}(x-x_j)(x-x_{j+1})^2]+f(x_{j+1})[h_j^{-2}(x-x_j)^2$

$$-2h_j^{-3}(x-x_{j+1})(x-x_j)^2]+s'(x_j)h_j^{-2}(x-x_j)(x-x_{j+1})^2+$$

$$s'(x_{j+1})h_j^{-2}(x-x_j)^2(x-x_{j+1}),\quad x\in[x_j,x_{j+1}].$$

Παραγωγίζοντας στην (2) παίρνουμε

$$s''(x_j^+)=2[3h_j^{-2}(f(x_{j+1})-f(x_j))-h_j^{-1}(s'(x_{j+1})+2s'(x_j))],\ 1\le j\le N-1.$$

Για το διάστημα $[x_{j-1},x_j]$ υπολογίζουμε την $s(x)$ από την (2) θέτοντας $j-1$ αντί του j. Βλέπουμε ότι

$$s''(x_j^-)= 2[-3h_{j-1}^{-2}(f(x_j)-f(x_{j-1}))+h_{j-1}^{-1}(s'(x_{j-1})+2s'(x_j))],\ 2\leq j\leq N,$$

Αλλά $s\in S_\tau^4$, δηλ. $s\in C^2[a,b]$. Από την ισότητα $s''(x_j^-)=s''(x_j^+)$, $2\leq j\leq N-1$ και τις παραπάνω σχέσεις προκύπτει η (1).

Η (1) μας οδηγεί στην κατασκευή της λεγόμενης (πλήρους) συνάρτησης παρεμβολής I_4f μιας παραγωγίσιμης συνάρτησης f στον χώρο S_τ^4 που παρεμβάλλεται στις τιμές $f(x_i)$ στα σημεία x_i, $1\leq i\leq N$, και της οποίας η παράγωγος $(I_4f)'$ συμφωνεί με την f' στα άκρα x_1,x_N του διαστήματος $[a,b]$:

ΠΡΟΤΑΣΗ 1. Δίδεται $f\in C^1[a,b]$ και διαμερισμός τ: $a=x_1<x_2<\ldots x_N=b$. Υπάρχει μοναδική κυβική spline $I_4f\in S_\tau^4$, τέτοια ώστε $(I_4f)(x_i)=f(x_i)$, $1\leq i\leq N$, $(I_4f)'(x_1)=f'(x_1)$, $(I_4f')'(x_N)=f'(x_N)$.

Απόδειξη: Οι σχέσεις (1) και οι εξισώσεις $s'(x_1)=f'(x_1)$, $s'(x_N)=f'(x_N)$ ορίζουν ένα $(N-2)x(N-2)$ γραμμικό σύστημα για τους αγνώστους $s'(x_i)$, $2\leq j\leq N-1$. Ο πίνακας του συστήματος αυτού είναι ο τριδιαγώνιος πίνακας

$$(3)\ B=\begin{pmatrix} 2(h_2+h_1) & h_1 & & & O \\ h_3 & 2(h_3+h_2) & h_2 & & \\ & \ddots & \ddots & \ddots & \\ & & h_{N-2} & 2(h_{N-2}+h_{N-3}) & h_{N-3} \\ O & & & h_{N-1} & 2(h_{N-1}+h_{N-2}) \end{pmatrix}$$

ο οποίος έχει αυστηρά κυριαρχική διαγώνιο,, δηλ. ικανοποιεί τις σχέσεις $|b_{ii}|>\Sigma_{j,j\neq i}|b_{ij}|$, $1\leq i\leq N-2$. Τέτοιοι πίνακες είναι ως γνωστόν αντιστρέψιμοι (από το θεώρημα του Gerschgorin, βλ. π.χ. [5.4]). Συνεπώς, οι (1) και οι συνοριακές συνθήκες $s'(x_1)=f'(x_1)$, $s'(x_N)=f'(x_N)$ ορίζουν μονοσήμαντα τις κλίσεις $s'(x_j)$, $1\leq j\leq N$. Η συνάρτηση $s(x)=(I_4f)(x)$, που ορίζεται τώρα σε κάθε διάστημα $[x_j,x_{j+1}]$ από την (2), ικανοποιεί λοιπόν τις συνθήκες $(I_4f)(x_i)=f(x_i)$, $1\leq i\leq N$, $(I_4f)'(x_k)=f'(x_k)$, $k=1,N$. Αν υπήρχε και άλλη συνάρτηση $\psi(x)$ με τις ίδιες ιδιότητες στον S_t^4 οι κλίσεις της $\psi'(x_j)$ θα ικανοποιούσαν τό ίδιο γραμμικό σύστημα με τις $s'(x_j)$ και συνεπώς θα είχαμε $\psi'(x_j)=s'(x_j)$, $1\leq j\leq N$. Επειδή $\psi(x_j)=f(x_j)$, $1\leq j\leq N$, η παράσταση (2) για την $\psi(x)$ θα ήταν ταυτόσημη με την παράσταση της $s(x)$, δηλ. $\psi=s$. @

Δεν είναι δύσκολο να δούμε ότι η κατασκευή του πίνακα B και του δεύτερου μέλους -βλ. (1)- του συστήματος για τον προσδιορισμό των $s'(x_j)$ καθώς και η επίλυση του τριδιαγώνιου συστήματος (με τον γνωστό μας "τριδιαγώνιο αλγόριθμο" που είναι ευσταθής στην περίπτωσή μας γιατί ο B έχει αυστηρά κυριαρχική διαγώνιο) απαιτεί $15N+O(1)$ περίπου πράξεις. Η τιμή $s(x)$ της κυβικής spline βρίσκεται κατόπιν από την (2) για δεδομένο x. Βλέπουμε δηλ. ότι ο προσδιορισμός της κυβικής spline παρεμβολής I_4f απαιτεί λύση γραμμικού συστήματος, δηλ. ότι η τιμή της $(I_4f)(x)$ εξαρτάται τελικά από <u>όλες</u> τις τιμές $f(x_j)$, $1\leq j\leq N$, $f'(x_1)$, $f'(x_N)$ των δεδομένων, σε αντίθεση π.χ. με τις συναρτήσεις παρεμβολής I_2f, I_Hf που για $x\in[x_j,x_{j+1}]$ εξαρτώνται μόνο <u>τοπικά</u> από την συνάρτηση f.(Η $(I_2f)(x)$ από τις τιμές $f(x_j)$, $f(x_{j+1})$ και η $(I_Hf)(x)$ από τις $f(x_j)$, $f'(x_j)$, $f(x_{j+1})$, $f'(x_{j+1})$).

Προχωρούμε τώρα στην εκτίμηση του σφάλματος της παρεμβολής $e(x)=f(x)-(I_4f)(x)$. Η εκτίμηση στηρίζεται στο εξής αποτέλεσμα:

Λήμμα 1. Εστω $f\in C^4[a,b]$ και $h=\max_j h_j=\max_j(x_{j+1}-x_j)$. Τότε, αν $e(x)=f(x)-(I_4f)(x)$, έχουμε

(4) $\max_{1\le i\le N}|e'(x_i)|\le h^3\|f^{(4)}\|_\infty/24.$

Απόδειξη: Εκ κατασκευής της I_4f έχουμε $e'(x_1)=e'(x_N)=0$. Στον R^{N-2} θεωρούμε τα διανύσματα $e'=(e'_2,\ldots,e'_{N-1})^T$, όπου $e'_i=e'(x_i),1\le i\le N$, και $r=(r_2,\ldots,r_{N-1})^T$ που ορίζεται από την εξίσωση

(5) $B\ e'=r,$

όπου B ο $(N-2)\times(N-2)$ πίνακας (3). Χρησιμοποιώντας τις σχέσεις (1) βλέπουμε ότι για $2\le j\le N-1$, επειδή $e'(x_1)=e'(x_N)=0$

$$(6)\ r_j=r_j(f)=h_je'(x_{j-1})+2(h_j+h_{j-1})e'(x_j)+h_{j-1}e'(x_{j+1})$$

$$=h_jf'(x_{j-1})+2(h_j+h_{j-1})f'(x_j)+h_{j-1}f'(x_{j+1})$$

$$-3[h_{j-1}(f(x_{j+1})-f(x_j))/h_j+h_j(f(x_j)-f(x_{j-1}))/h_{j-1}],$$

όπου γράψαμε $r_j=r_j(f)$ για να συμβολίσουμε το γεγονός ότι για κάθε j, $2\le j\le N-1$, το r_j είναι γραμμικό συναρτησιακό στον $C^4[a,b]$. Επιπλέον, αν $p\in P_3$ έχουμε $r_j(p)=0$. Συνεπώς, το θεώρημα του πυρήνα του Peano 4.2.1 μας δίνει·

$$(7)\ r_j(f)=(1/6)\int_{x_{j-1}}^{x_{j+1}} r_{j,x}[(x-t)_+^3]\ f^{(4)}(t)\ dt$$

γιατί η μορφή του συναρτησιακού (6) επιτρέπει την εναλλαγή των πράξεων $r_{j,x}$ και της ολοκλήρωσης ως πρός t.

Θα δείξουμε τώρα, χρησιμοποιώντας τις (6),(7), ότι

(8) $|r_j(f)|\le(1/24)\|f^{(4)}\|_\infty\,(h_jh_{j-1}^3+h_{j-1}h_j^3)$, $2\le j\le N-1$.

Μία σειρά απλών πράξεων δίνει, κατ'αρχήν, λόγω της (6), ότι για $2\le j\le N-1$

$$(9)\ r_{j,x}[(x-t)_+^3]=\begin{cases}6(h_j+h_{j-1})(x_j-t)^2+3h_{j-1}(x_{j+1}-t)^2-3[h_jh_{j-1}^{-1}(x_j-t)^3\\ +h_{j-1}h_j^{-1}((x_{j+1}-t)^3-(x_j-t)^3)],\ \text{αν}\ x_{j-1}\le t\le x_j,\\ 3h_{j-1}(x_{j+1}-t)^2-3[h_{j-1}h_j^{-1}(x_{j+1}-t)^3],\text{αν}\ x_j\le t\le x_{j+1}.\end{cases}$$

Επίσης μία σειρά απλών πράξεων δίνει αμέσως ότι

(10) $r_{j,x}[(x-t)_+^3]\ge 0\ \forall t\in[x_j,x_{j+1}]$ και $r_{j,x}[(x-t)_+^3]\le 0\ \ \forall t\in[x_{j-1},x_j]$.

(Η απόδειξη της (10) για $t\in[x_j,x_{j+1}]$ προκύπτει αμέσως από το δεύτερο σκέλος της (9). Για $t\in[x_{j-1},x_j]$ το πρώτο σκέλος της (9) δίνει ότι η $q(t)=r_{j,x}[(x-t)_+^3]$ είναι κυβικό πολυώνυμο στο $[x_{j-1},x_j]$ με ιδιότητες $q(x_j)=0$, $q(x_{j-1})=0$, $q'(x_j)=-6h_{j-1}h_j+9h_{j-1}h_j^{-1}h_j^2=3h_{j-1}h_j>0$, $q'(x_{j-1})=0$. Συνεπώς το κυβικό πολυώνυμο $q(t)$ έχει διπλή ρίζα στο σημείο x_{j-1}, απλή στο x_j και θετική παράγωγο στο x_j. Αναγκαστικά λοιπόν $q(t)\le 0$ στο διάστημα $[x_{j-1},x_j]$.

Χρησιμοποιώντας τις (10) στην (7) παίρνουμε

$$|r_j(f)|\le(\|f^{(4)}\|_\infty/6)\ \Big(\int_{x_{j-1}}^{x_j}-r_{j,x}[(x-t)_+^3]dt+\int_{x_j}^{x_{j+1}}r_{j,x}[(x-t)_+^3]dt\Big)$$

Χρήση τώρα της (9) και απλές ολοκληρώσεις δίνουν την (8).

Το πρόβλημα είναι τώρα να εκτιμήσουμε την λύση $e' \in R^{N-2}$ του συστήματος (5) χρησιμοποιώντας τις εκτιμήσεις (8) των συνιστωσών του δευτέρου μέλους. Ας συμβολίσουμε με $\|\cdot\|_\infty$ και την διανυσματική maximum νόρμα στον R^{N-2} καθώς και την παραγόμενη νόρμα πινάκων στον $R^{(N-2)\times(N-2)}$. Πολλαπλασιάζοντας και τα δύο μέλη του συστήματος (5) επί τον $(N-2)\times(N-2)$ διαγώνιο πίνακα $D=2\ \mathrm{diag}(h_2+h_1, h_3+h_2, \ldots h_{N-1}+h_{N-2})$ παίρνουμε

$$(11)\quad D^{-1}\ B\ e' \equiv (I+M)e' = D^{-1}r.$$

Αμέσως βλέπουμε ότι

$$\|M\|_\infty = \max_i \sum_j |m_{ij}| = \max_{2\le i\le N-1}[(h_i/2(h_i+h_{i-1})) + (h_{i-1}/2(h_i+h_{i-1}))] = 1/2.$$

Συνεπώς από το θεώρημα του Neumann έχουμε ότι ο $I+M$ είναι αντιστρέψιμος και ότι $\|(I+M)^{-1}\|_\infty \le 1/(1-\|M\|_\infty) \le 2$. Η (11) λοιπόν δίνει

$$(12)\quad \max_{2\le i\le N-1}|e'(x_i)| = \|e'\|_\infty \le 2\|D^{-1}r\|_\infty = \max_{2\le j\le N-1}\{|r_j|/(h_j+h_{j-1})\}.$$

Εξ άλλου από την (8) έχουμε για $2\le j\le N-1$

$$(13)\quad |r_j|\ /\ (h_j+h_{j-1}) \le (\|f^{(4)}\|_\infty/24)(h_j h_{j-1}^3 + h_{j-1}h_j^3)/(h_j+h_{j-1})$$

$$\le \max(h_{j-1}^3, h_j^3)\|f^{(4)}\|_\infty/24 \le h^3\|f^{(4)}\|_\infty/24.$$

(Οι δύο τελευταίες ανισότητες δεν μπορούν να βελτιωθούν όπως δείχνει η περίπτωση του ομοιόμορφου διαμερισμού). Οι (12) και (13) δίνουν λοιπόν την (4).@

Με το αποτέλεσμα του Λήμματος 1 μπορούμε να αποδείξουμε το βασικό αποτέλεσμα αυτής της παραγράφου:

ΘΕΩΡΗΜΑ 1. Αν $f\in C^4[a,b]$ και $h=\max_j h_j$ έχουμε

$$(14)\quad \|f-I_4 f\|_\infty \le (5/384)h^4\|f^{(4)}\|_\infty.$$

<u>Απόδειξη</u>: Εστω $I_H f$ η συνάρτηση παρεμβολής Hermite της f για τον ίδιο διαμερισμό τ. Η τριγωνική ανισότητα μας δίνει

$$(15)\quad \|f-I_4 f\|_\infty \le \|f-I_H f\|_\infty+\|I_H f-I_4 f\|_\infty.$$

Για τον πρώτο όρο του δεύτερου μέλους έχουμε ήδη την εκτίμηση (4.3.7)· απομένει να εκτιμήσουμε την διαφορά $I_H f-I_4 f$ η οποία, ως τμηματικά κυβική συνάρτηση και στοιχείο του $C^1[a,b]$, ανήκει στον χώρο H_τ των κυβικών συναρτήσεων Hermite. Εξ ορισμού των I_H, I_4 έχουμε εξ άλλου ότι $(I_H f-I_4 f)(x_i)=0$, $1\le i\le N$ και $(I_H f-I_4 f)'(x_k)=0$, $k=1,N$. Συνεπώς ο τύπος (4.3.4) για $\varphi=I_H f-I_4 f$ δίνει

$$(I_H f-I_4 f)(x)=\sum_{i=2}^{N-1}[(I_H f)'(x_i)-(I_4 f)'(x_i)]S_i(x)$$

$$=\sum_{i=2}^{N-1}(f-I_4 f)'(x_i)S(x_i)=\sum_{i=2}^{N-1}e'(x_i)S_i(x),$$

όπου χρησιμοποιήσαμε τον συμβολισμό $e=f-I_4 f$ του Λήμματος 1. Συνεπώς, η (4) δίνει

$$(16)\quad \|I_H f-I_4 f\|_\infty \le \max_i|e'(x_i)|\max_{a\le x\le b}\left(\sum_{i=2}^{N-1}|S_i(x)|\right)$$

$$\le(h^3\|f^{(4)}\|_\infty/24)\max_{a\le x\le b}\left(\sum_{i=2}^{N-1}|S_i(x)|\right).$$

Απομένει λοιπόν να εκτιμήσουμε τον τελευταίο όρο του δεύτερου μέλους της (16). Σε κάθε υποδιάστημα $[x_j,x_{j+1}]$, $1\le j\le N-1$ μόνο οι συναρτήσεις S_j και S_{j+1} (βλ. (4.3.2)) έχουν μη μηδενικές τιμές. Για $x\in[x_j,x_{j+1}]$ έχουμε λοιπόν από τις (4.3.2) ότι

$$\sigma_j(x)=|S_j(x)|+|S_{j+1}(x)|=h_j^{-2}(x-x_j)(x_{j+1}-x)^2+h_j^{-2}(x-x_j)^2(x_{j+1}-x)$$

$$=h_j^{-1}(x-x_j)(x_{j+1}-x)\le h_j/4.$$

Άρα, αν η συνάρτηση $\sum_{i=2}^{N-1}|S_i(x)|$ παίρνει το μέγιστό της σ' ένα σημείο $\bar{x}\in[x_k,x_{k+1}]$ για κάποιο k, $1\le k\le N-1$, έχουμε

$$(17)\quad \max_{a\le x\le b}\sum_{i=2}^{N-1}|S_i(x)|=\sigma_k(\bar{x})\le h_k/4\le h/4.$$

Οι (16) και (17) δίνουν συνεπώς

$$(18)\quad \|I_Hf-I_4f\|_\infty\le h^4\|f^{(4)}\|_\infty/96.$$

Οι (4.3.7), (18) και (15) δίνουν τώρα την (14). @

Μπορεί να αποδειχθεί (βλ. Hall και Meyer, J.Approx. Theory, 16(1976), 105-122) ότι η σταθερά 5/384 είναι η καλύτερη δυνατή. Στην ίδια εργασία αποδεικνύονται επίσης και φράγματα με άριστες σταθερές για τα σφάλματα των παραγώγων $(f-I_4f)^{(k)}$. Ισχύει το εξής αποτέλεσμα:

ΘΕΩΡΗΜΑ 2 (Hall και Meyer). Εστω $f\in C^4[a,b]$, $h=\max_j h_j$. Τότε έχουμε

$$(19)\quad \|(f-I_4 f)^{(k)}\|_\infty \le c_k\, h^{4-k}\, \|f^{(4)}\|_\infty,\ k=0,1,2,3,$$

όπου $c_0=5/384$, $c_1=1/24$, $c_2=3/8$ και $c_3=(\lambda+\lambda^{-1})/2$ όπου $\lambda=h/(\min_j h_j)$.
Επιπλέον οι σταθερές c_0, c_1 είναι οι καλύτερες δυνατές με την έννοια ότι

$$c_k=\sup_{f\in C^4[a,b],\tau} \{\|(f-I_4 f)^{(k)}\|/h^{4-k}\|f^{(4)}\|_\infty\},\ k=0,1.@$$

Προχωρούμε τώρα στην διερεύνηση των άλλων ερωτημάτων που θέσαμε στην εισαγωγή αυτής της παραγράφου, δηλ. στο να βρούμε την διάσταση του χώρου S_τ^4 και να κατασκευάσουμε μία αποτελεσματική για τις εφαρμογές βάση του. Η διάσταση που περιμένουμε να έχει ο χώρος S_τ^4 είναι $N+2$. Πράγματι ο προσδιορισμός ενός στοιχείου του S_τ^4(δηλ. μιας κυβικής spline) απαιτεί τον προσδιορισμό $4(N-1)$ σταθερών, δηλ. των τεσσάρων συντελεστών ενός κυβικού πολυωνύμου σε κάθε διάστημα $[x_j,x_{j+1}]$, $1\le j\le N-1$. Επιβάλλοντας τον περιορισμό $S_\tau^4 \subset C^2[a,b]$, δηλ. επιβάλλοντας C^2 ομαλότητα στους εσωτερικούς κόμβους x_j, $2\le j\le N-1$ παίρνουμε $3(N-2)$ συνθήκες. Άρα οι "βαθμοί ελευθερίας" (συνολικός αριθμός ελευθέρων παραμέτρων = πλήθος "συντεταγμένων" μιάς κυβικής spline = διάσταση του χώρου S_τ^4) είναι $4(N-1)-3(N-2)=N+2$. Θα αποδείξουμε αυτό το αποτέλεσμα κατασκευάζοντας συγχρόνως και μια χρήσιμη βάση του S_τ^4, τις B-splines.

Μια επιθυμητή ιδιότητα για τις συναρτήσεις βάσης είναι να έχουν μικρό φορέα. Είναι εύκολο να δεί κανείς ότι δεν υπάρχει άλλη κυβική spline (δηλ. C^2) με φορέα $[x_{j-1},x_{j+1}]$, δηλ. δύο διαστήματα, εκτός από την τετριμμένη μηδενική συνάρτηση· πράγματι πρέπει να επιβάλλουμε 9 συνθήκες (τρείς σε κάθε κόμβο x_{j-1},x_j,x_{j+1}) για να προσδιορίσουμε 8 σταθερές. Δεν υπάρχει επίσης μη τετριμμένη κυβική spline με φορέα τριών διαστημάτων. Ο μικρότερος δυνατός φορέας αποτελείται από 4 διαστήματα, είναι δηλ. για $3\le j\le N-2$ της μορφής $[x_{j-2},x_{j+2}]$.

Στην περίπτωση αυτή πρέπει να προσδιορίσουμε 16 παραμέτρους από 15 εξισώσεις που επιτρέπουν το υπολογισμό ενός τμηματικά κυβικού πολυωνύμου στον χώρο $C^2[a,b]$ με φορέα $[x_{j-2},x_{j+2}]$· όλα τα πολλαπλάσιά του επί σταθερά θα έχουν προφανώς τις ίδιες ιδιότητες.

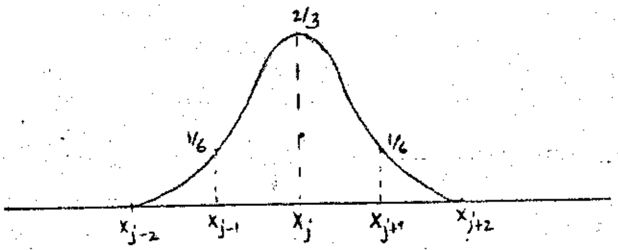
Η συνάρτηση $s_j(x)$ που προκύπτει λέγεται <u>B-spline</u> (με κέντρο τόν κόμβο x_j). Χρησιμοποιώντας διαιρεμένες διαφορές (βλ. π.χ. το βιβλίο του De Boor [4.2, κεφ. 9]) μπορούμε να την κατασκευάσουμε αρκετά εύκολα χωρίς να λύσουμε το γραμμικό σύστημα. Η συνάρτηση που προκύπτει είναι, για $3\le j\le N-2$ η

$$(20)\quad s_j(x) = (x_{j+2}-x_{j-2}) \sum_{k=-2}^{2}\Big[(x_{j+k}-x)_+^3 \prod_{\substack{s=-2\\ s\neq k}}^{2} (x_{j+k}-x_{j+s})^{-1}\Big].$$

Στην περίπτωση του <u>ομοιόμορφου διαμερισμού</u> με $h=x_{j+1}-x_j=\text{σταθ}$ η B-spline s_j με κέντρο x_j δίνεται για $3\le j\le N-2$ από

$$(21)\quad s_j(x)=\begin{cases} (x-x_{j-2})^3/6h^3, & x_{j-2}\le x\le x_{j-1}\\ (2/3)-(x-x_j)^2/h^2-(x-x_j)^3/2h^3, & x_{j-1}\le x\le x_j\\ (1/6)+(x_{j+1}-x)/2h+(x_{j+1}-x)^2/2h^2-(x_{j+1}-x)^3/2h^3, & x_j\le x\le x_{j+1}\\ (x_{j+2}-x)^3/6h^3\ , & x_{j+1}\le x\le x_{j+2}\\ 0\ , & x\notin[x_{j-2},x_{j+2}] \end{cases}$$

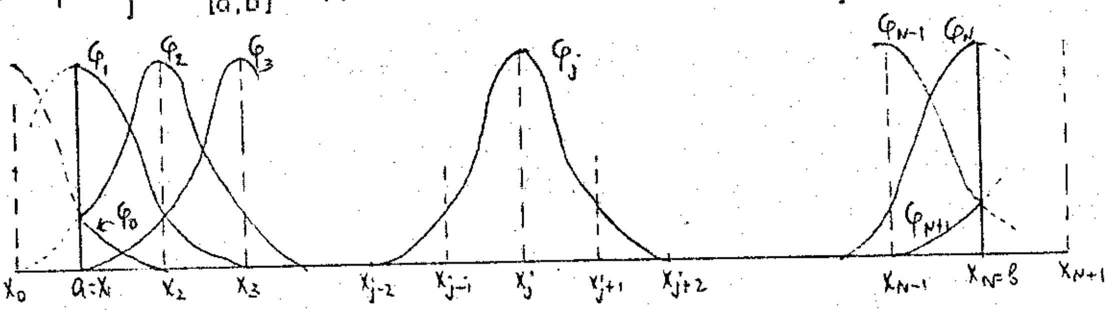
$s_j(x)$:



Μ' αυτόν τόν τρόπο κατασκευάζουμε N-4 B-splines s_j, $3\le j\le N-2$ με κέντρα στα αντίστοιχα σημεία x_j. Η βάση του χώρου S_τ^4 προσδιορίζεται τώρα ως εξής: Θεωρούμε δύο επιπλέον κόμβους $x_0<a$ και $x_{N+1}>b$ - στην περίπτωση ομοιόμορφου διαμερισμού $x_0=a-h$, $x_{N+1}=b+h$ - και ορίζουμε N+2 συναρτήσεις φ_j ως

$$(22)\quad \varphi_j(x) = s_j(x)|_{[a,b]}, \qquad 0\le j\le N+1,$$

όπου με $s_j(x)|_{[a,b]}$ συμβολίζουμε τον περιορισμό της s_j στο $[a,b]$.



Οι συναρτήσεις φ_j, $0\le j\le N+1$, αποτελούν <u>βάση</u> του S_τ^4, ο οποίος συνεπώς έχει διάσταση N+2. Ας αποδείξουμε το αποτέλεσμα αυτό, για να απλουστεύσουμε τα πράγματα, στην περίπτωση του ομοιόμορφου διαμερισμού με $h=x_{j+1}-x_j$, $0\le j\le N$. Κατ'αρχήν οι φ_j, $0\le j\le N+1$ είναι γραμμικά ανεξάρτητες: Ας υποθέσουμε ότι $\sum_{j=0}^{N+1} c_j\varphi_j(x)=0 \quad \forall x\in[a,b]$.

Τότε $\sum_{j=0}^{N+1} c_j\varphi_j(x_i)=0$, $1\le i\le N$ και $\sum_{j=0}^{N+1} c_j\varphi'_j(x_k)=0$, $k=1,N$. Από την (21) βλέπουμε ότι οι παραπάνω εξισώσεις οδηγούν στις συνθήκες

(23) $A\,c = 0,\ c=[c_1,\dots,c_N]^T,\ c_0=c_2, c_{N+1}=c_{N-1},$

όπου ο $N\times N$ τριδιαγώνιος πίνακας A δίνεται από

$$(24)\ A = \begin{pmatrix} 4 & 2 & & 0 \\ 1 & 4 & 1 & \\ \ddots & \ddots & \ddots & \\ 0 & 1 & 4 & 1 \\ & & 2 & 4 \end{pmatrix}$$

και είναι αντιστρέψιμος επειδή έχει αυστηρά κυριαρχική διαγώνιο. Από τις (23) συμπεραίνουμε ότι $c_j=0$, $0\le j\le N+1$, δηλ. ότι οι φ_j, $0\le j\le N+1$ είναι γραμμικά ανεξάρτητες. Για να δείξουμε ότι παράγουν τον S_τ^4, θεωρούμε οποιοδήποτε διάστημα $I_j=[x_j,x_{j+1}]$ για κάποιο $1\le j\le N-2$ και τον τετραδιάστατο χώρο $P_3(I_j)$. Επειδή τα κυβικά πολυώνυμα στο I_j $\varphi_{j-1},\varphi_j,\varphi_{j+1},\varphi_{j+2}$ (εκείνες δηλ. οι φ_j των οποίων ο φορέας περιέχει το I_j) είναι γραμμικά ανεξάρτητα, θα παράγουν τον $P_3(I_j)$. Έστω $\varphi(x)$ τυχόν στοιχείο του S_τ^4. Επειδή $\varphi|_{I_j}\in P_3(I_j)$ θα έχουμε για $x\in I_j$

$$(25)\ \varphi(x) = c_{j-1}^{(j)}\varphi_{j-1}(x)+c_j^{(j)}\varphi_j(x)-c_{j+1}^{(j)}\varphi_{j+1}(x)+c_{j+2}^{(j)}\varphi_{j+2}(x),$$

όπου τονίζουμε με τον άνω δείκτη (j) των $c_{j+k}^{(j)}$, $k=-1,0,1,2$ το ότι η τετράδα των συντελεστών αυτών θα εξαρτάται κατ'αρχήν από το διάστημα I_j. Αν δείξουμε ότι είναι ανεξάρτητοι του j, δηλ. ότι αν π.χ. για $x\in I_{j+1}$

$$(26)\quad \varphi(x)=c_j^{(j+1)}\varphi_j(x)+c_{j+1}^{(j+1)}\varphi_{j+1}(x)+c_{j+2}^{(j+1)}\varphi_{j+2}(x)+c_{j+3}^{(j+1)}\varphi_{j+3}(x),$$

συνεπάγεται $c^{(j)}_{j+k}=c^{(j+1)}_{j+k}$, k=0,1,2, τότε είναι φανερό ότι κάθε $\varphi\in S^4_\tau$ μπορεί να εκφρασθεί για $x\in[a,b]$ ως γραμμικός συνδυασμός των φ_j, $0\leq j\leq N+1$. Πράγματι, οι σχέσεις (25) και (26) και η C^2 συνέχεια της φ στο $x=x_{j+1}$ δίνουν, λόγω της (21),

$$\varphi(x^-_{j+1})=\varphi(x^+_{j+1}):\ c^{(j)}_j/6+2c^{(j)}_{j+1}/3+c^{(j)}_{j+2}/6=c^{(j+1)}_j/6+2c^{(j+1)}_{j+1}/3+c^{(j+1)}_{j+2}/6$$

$$\varphi'(x^-_{j+1})=\varphi'(x^+_{j+1}):-c^{(j)}_j/2h+c^{(j)}_{j+2}/2h=-c^{(j+1)}_j/2h+c^{(j+1)}_{j+2}/2h$$

$$\varphi''(x^-_{j+1})=\varphi''(x^+_{j+1}):c^{(j)}_j/h^2-2c^{(j)}_{j+1}/h^2+c^{(j)}_{j+2}/h^2=c^{(j+1)}_j/h^2$$

$$-2c^{(j+1)}_{j+1}h^2+c^{(j+1)}_{j+2}/h^2.$$

Συνεπώς οι αριθμοί $d_k=c^{(j)}_{j+k}-c^{(j+1)}_{j+k}$, k=0,1,2, ικανοποιούν το 3x3 ομογενές σύστημα

$$\begin{pmatrix} 1 & 4 & 1 \\ -1 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix}\begin{pmatrix} d_0 \\ d_1 \\ d_2 \end{pmatrix}=\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

που έχει την μοναδική λύση $d_k=0$, k=0,1,2.

Η χρήση της βάσης $\{\varphi_j\}$ $0\leq j\leq N-1$ του S^4_τ μας επιτρέπει να λύνουμε <u>αποτελεσματικά</u>, δηλ. με λύση αραιών γραμμικών συστημάτων με καλούς δείκτες κατάστασης, προβλήματα όπως π.χ. του υπολογισμού της L^2-προβολής μιάς συνάρτησης f στον S^4_τ (βλ. παρ. 4.2) της αριθμητικής λύσης διαφορικών εξισώσεων με μεθόδους τύπου Galerkin κ.ά. Σαν παράδειγμα αναφέρουμε το πρόβλημα της <u>παρεμβολής</u> στον S^4_τ: Γράφοντας για $f\in C^1[a,b]$

(27) $(I_4f)(x)=\sum_{j=0}^{N+1}c_j\varphi_j(x)$,

και χρησιμοποιώντας τις συνθήκες παρεμβολής $(I_4f)(x_i)=f(x_i)$, $1\le i\le N$, $(I_4f)'(x_k)=f'(x_k)$, $k=1,N$ οδηγούμαστε στις εξισώσεις (π.χ. για ομοιόμορφο διαμερισμό):

$$\sum_{j=0}^{N+1} c_j\varphi_j(x_i)=f(x_i),\ 1\le i\le N,\ c_0\varphi'_0(x_1)+c_2\varphi'_2(x_1)=f'(x_1),$$

$$c_{N-1}\varphi'_{N-1}(x_N)+c_{N+1}\varphi'_{N+1}(x_N)=f'(x_N)$$

που δίνουν το NxN γραμμικό σύστημα

(28) $A\ c = b$,

όπου $c=[c_1,\dots,c_N]^T$, A ο τριδιαγώνιος πίνακας (24) και $b=[b_1,\dots,b_N]^T$, όπου $b_1=6f(x_1)+2hf'(x_1)$, $b_i=6f(x_i)$, $2\le i\le N-1$, $b_N=6f(x_N)-2hf'(x_N)$. Το (28) προφανώς έχει μοναδική λύση. Τέλος προσδιορίζοντας τα c_0,c_{N+1} από τις εξισώσεις $c_0=c_2-2hf'(x_1)$, $c_{N+1}=c_{N-1}+2hf'(x_N)$ υπολογίζουμε την συνάρτηση παρεμβολής I_4f από την (27). Η λύση του (28) βρίσκεται με 5N περίπου πράξεις χρησιμοποιώντας τον (ευσταθή στην περίπτωσή μας) τριδιαγώνιο αλγόριθμο.

Παρατηρήσεις

1. Η πρώτη απόδειξη ότι $\|f-I_4f\|_\infty=O(h^4)$ για αρκετά ομαλή f οφείλεται στούς Birkhoff και De Boor (J. Math. Mech., 13(1964), 827-836) για διαμερισμούς για τους οποίους ο λόγος $\lambda=\max_j h_j/\min_j h_j$ μένει φραγμένος. Η απόδειξη του θεωρήματος 1 χωρίς περιορισμό στον διαμερισμό οφείλεται στον Hall (J. Approx. Theory 1(1968), 209-218) ενώ του θεωρήματος 2 στους Hall και Meyer (op. cit.).

2. Μία σημαντική ιδιότητα της κυβικής spline παρεμβολής $I_4 f$ (βλ. Ασκ. 8γ) είναι ότι ελαχιστοποιεί το ολοκλήρωμα $\int_a^b (g''(t))^2 dt$ μεταξύ όλων των C^2- συναρτήσεων g που ικανοποιούν τις σχέσεις παρεμβολής $g(x_i)=f(x_i)$, $1\le i\le N$. Η ιδιότητα αυτή έχει κάποια σημασία στην μηχανική (θεωρία εύκαμπτων λεπτών δοκών) και σ'αυτήν οφείλεται η ονομασία "spline", που είναι το σχήμα που παίρνει μία εύκαμπτη λεπτή οδηγός καμπύλη όταν διέρχεται από δεδομένα σημεία (x_i,y_i), $1\le i\le N$ στο επίπεδο. Τέτοιες καμπύλες χρησιμοποιούνται για την σχεδίαση πλοίων, αυτοκινήτων κ.λ.π.

3. Θα δούμε στην Άσκηση 10 ανάλογα αποτελέσματα του θεωρήματος 2 αλλά ως πρός την L^2-νόρμα $\|.\|$. Προφανώς υπάρχουν και τα αναμενόμενα ανάλογα αποτελέσματα σ'όλες τις συνήθεις νόρμες για φράγματα σφαλμάτων μικρότερης τάξης ακρίβειας αν η f δεν είναι C^4. Βλ. πχ. Ασκ.2,6,9.

4. Η κυβική spline παρεμβολής $f_h=I_4 f$ που μελετήσαμε είναι η λεγόμενη <u>πλήρης</u> συνάρτηση παρεμβολής της f στον χώρο S_τ^4 και, όπως είδαμε, αντιστοιχεί σε συνοριακές συνθήκες δεδομένης κλίσεως στα άκρα, δηλ. στις συνθήκες $(I_4 f)'(a)=f'(a)$, $(I_4 f)'(b)=f'(b)$. Είναι βέβαια δυνατόν να επιβάλουμε, πέρα από τις συνθήκες παρεμβολής στις τιμές της f στα σημεία x_i, $1\le i\le N$, και άλλα ζευγάρια βοηθητικών επιπλέον συνθηκών, π.χ. άλλου τύπου συνοριακές συνθήκες. Αν π.χ. οι τιμές $f''(a)$, $f''(b)$ είναι γνωστές, μπορούμε να επιβάλουμε τις συνθήκες $f''_h(a)=f''(a), f''_h(b)=f''(b)$ που οδηγούν στις εξισώσεις:

$$2s'(x_1)+s'(x_2)= 3h_1^{-1}(f(x_2)-f(x_1))-h_1 f''(x_1)/2$$

και

$$s'(x_{N-1})+2s'(x_N)=3h_{N-1}^{-1}(f(x_N)-f(x_{N-1}))+h_{N-1}f''(x_N)/2,$$

οι οποίες μαζί με τις N-2 εξισώσεις (1) επιτρέπουν τον υπολογισμό των κλίσεων $s'(x_i)$, $1\le i\le N$ της spline f_h.

Αν τώρα είναι γνωστές μόνο οι τιμές $f(x_i)$, $1\le i\le N$, μπορούμε να προσεγγίσουμε π.χ. τις τιμές $f'(x_1), f'(x_N)$ που απαιτούνται για τον προσδιορισμό του I_4f π.χ. αντικαθιστώντας την $f'(x_1)$ με $p'(x_1)$ όπου p το κυβικό πολυώνυμο παρεμβολής Lagrange της f για τα σημεία x_1, x_2, x_3, x_4 κ.ο.κ..

5. Αξιοσημείωτη είναι η ιδιότητα "<u>υπερσύγκλισης</u>" (που ισχύει για <u>ομοιόμορφο</u> διαμερισμό και $f\in C^5[a,b]$):

$$(29)\quad \max_{1\le i\le N}|(f-I_4f)'(x_i)| \le h^4\|f^{(5)}\|_\infty/60$$

(βλ. Ασκ. 3) και που εξηγεί την μεγάλη ακρίβεια των τιμών και της παραγώγου $(I_4f)'$ στους κόμβους για ομοιόμορφο διαμερισμό.

6. Παραπέμπουμε τον αναγνώστη π.χ. στο βιβλίο [4.2] του De Boor για την μελέτη τμηματικά πολυωνυμικών συναρτήσεων ("splines") βαθμού 2 καθώς και οποιουδήποτε βαθμού k>3.

Ασκήσεις 4.4

1. Μία άλλη κατασκευή της συνάρτησης I_4f είναι η εξής: Υπολογίστε το κυβικό πολυώνυμο $s_i(x)=(I_4f)(x)|_{[x_i,x_{i+1}]}$, $1\le i\le N-1$, συναρτήσει των τιμών του στα σημεία x_i και x_{i+1} και των τιμών m_i, m_{i+1} των <u>δευτέρων</u> παραγώγων του στα x_i, x_{i+1}. Οι σταθερές m_i, $1\le i\le N$ προσδιορίζονται κατόπιν από τις συνθήκες $s'_{i-1}(m_i)=s'_i(x_i)$, $2\le i\le N-1$ (C^1 στους εσωτερικούς κόμβους) και $s'_1(x_1)=f'(x_1)$, $s'_{N-1}(x_N)=f'(x_N)$. Βρείτε το γραμμικό σύστημα που προκύπτει για τα m_i και αποδείξτε ότι έχει μοναδική λύση.

2. (Σφάλμα της I_4f όταν $f\in C^1[a,b]$.)

(α) Χρησιμοποιώντας τις τεχνικές της απόδειξης του Λήμματος 4.2.2, δείξτε ότι η λύση $s'(x_j)$, $2\le j\le N-1$ του συστήματος με πίνακα B, (3), στην απόδειξη της πρότασης 1 ικανοποιεί, για δεδομένα $s'(x_1)$, $s'(x_N)$.

$$\max_i |s'(x_i)| \le 3\max(|s'(x_1)|, \max_{2\le j\le N-1}|b_j|, |s'(x_N)|),$$

όπου b_j, $2\le j\le N-1$ είναι το δεύτερο μέλος του συστήματος, δηλ. το δεύτερο μέλος των εξισώσεων (1).

(β). Συμπεράνετε ότι $\|(I_4 f)'\|_\infty \le 4 \max_{0\le i\le N}|h_i^{-1}(f(x_{i+1})-f(x_i))|$ όπου ορίζουμε $h_0=h_N=0$. (Υπόδειξη: Δείξτε ότι

$(I_4 f)'((x_i+x_{i+1})/2)=3h_i^{-1}(f(x_{i+1})-f(x_i))/2-(s'(x_i)+s'(x_{i+1}))/4$

και κατόπιν χρησιμοποιείστε την ανισότητα

$$\max_{a\le x\le b}|p(x)| \le (5/4)\max_{a\le x\le b}(|p(a)|, |p((a+b)/2)|, |p(b)|),$$

που ισχύει για κάθε $p\in P_3$).

(γ) Συμπεράνετε από το (β) και την άσκηση 4.2.5 ότι

$$\|f'-(I_4 f)'\|_\infty \le 5 \inf_{\varphi\in H_\tau} \|f'-\varphi\|_\infty$$

(δ) Τέλος αποδείξτε ότι

$$\|f'-(I_4 f)'\|_\infty \le 5h\|f'\|_\infty/2.$$

3. Υποθέστε ότι έχουμε ομοιόμορφο διαμερισμό και $f\in C^5[a,b]$. Τότε, από την (6) έχουμε

$$r_j(f)/3=h(f'(x_{j-1})+2f'(x_j)+f'(x_{j+1}))/3-\int_{x_{j-1}}^{x_{j+1}} f'(x)dx, \quad 2\le j\le N-1$$

(α) Χρησιμοποιώντας ένα γνωστό από την Εισαγωγή στην Αριθμητική Ανάλυση αποτέλεσμα για το σφάλμα του κανόνα του Simpson για αριθμητική ολοκλήρωση δείξτε ότι

$$|r_j(f)| \le h^5\|f^{(5)}\|_\infty/30, \quad 2\le j\le N-1.$$

(β) Συμπεράνετε, όπως στην απόδειξη του Λήμματος 1, ότι ισχύει η (29).

6. Μιμηθείτε την απόδειξη του Θεωρήματος 1 και χρησιμοποιώντας το Θεώρημα του Peano δείξτε ότι

$$\|f-I_4f\|_\infty \le 2h^2\|f''\|_\infty/3 \quad \text{αν } f\in C^2[a,b].$$

7. Με χρήση Peano δείξτε επίσης τα όχι άριστα ως προς τις σταθερές φράγματα (Hall, 1968) για ομοιόμορφο διαμερισμό:

$$\|(f-I_4f)'\|_\infty \le (\sqrt{3}/216 + 1/24)h^3\|f^{(4)}\|_\infty$$

$$\|(f-I_4f)''\|_\infty \le 5h^2\,\|f^{(4)}\|_\infty/12$$

$$\|(f-I_4f)'''\|_\infty \le h\,\|f^{(4)}\|_\infty.$$

8.(α) Αν $f\in PC^2[a,b]$ τότε $((I_4f-f)'',\varphi'')=0,\ \forall\varphi\in S_\tau^4$.

(β) Αν $f\in PC^2[a,b]$, τότε $\|(I_4f)''\|^2+\|(I_4f-f)''\|^2=\|f''\|^2$

(γ) Αποδείξτε τον ισχυρισμό της Παρατήρησης 2.

(δ) Αν $f \in PC^4[a,b]$, τότε $\|(f-I_4f)''\|^2=(f-I_4f,f^{(4)})$

9. Χωρίς χρήση του Θεωρήματος του Peano αλλά με χρήση των "ολοκληρωτικών ταυτοτήτων" της Άσκησης 8 και της ανισότητας του Poincaré (Ασκ. 4.2.9) δείξτε ότι αν $f \in PC^2[a,b]$ τότε

$$\|(f-I_4f)^{(k)}\| \le C_k h^{2-k}\|f''\|,\ k=0,1,2 \text{ με } C_0=2, C_1=2, C_2=1.$$

10. Με παρόμοιο τρόπο δείξτε, αν $f \in PC^4[a,b]$ ότι

$$\|(f-I_4f)^{(k)}\| \le C_k h^{4-k}\|f^{(4)}\|,\ k=0,1,2, \text{ με } C_0=4,\ C_1=4, C_2=2.$$