Περιεχόμενα

ΑΡΙΘΜΗΤΙΚΗ ΓΡΑΜΜΙΚΗ ΑΛΓΕΒΡΑ

(Σημειώσεις)

Ν. Μισυρλής

ΑΘΗΝΑ , 2005

Κεφάλαιο 1

Σφάλματα στους Αριθμητικούς Υπολογισμούς

1.1 Εισαγωγή

Υπάρχουν διάφορες πηγές σφαλμάτων που μπορούν να προκύψουν κατά τους αριθμητικούς υπολογισμούς.

1. *Σφάλματα που προκύπτουν κατά το σχηματισμό του μαθηματικού μοντέλου.* Ένα φυσικό φαινόμενο περιγράφεται από ένα μαθηματικό πρόβλημα, το οποίο προκύπτει από τη μελέτη ενός μοντέλου που προσεγγίζει όσο το δυνατόν καλύτερα το φυσικό φαινόμενο.

2. *Σφάλματα στα δεδομένα.* Κατά τη μέτρηση των δεδομένων ενός προβλήματος υπεισέρχοναι σφάλματα που προέρχονται κυρίως από τα όργανα μέτρησης.

3. *Σφάλματα αποκοπής.* Είναι αυτά που προκύπτουν όταν, για παράδειγμα, επιθυμούμε τον υπολογισμό της τιμής μιας σειράς απείρων όρων. Στην περίπτωση αυτή είμαστε αναγκασμένοι να διατηρήσουμε μόνο ένα συγκεκριμένο πλήθος όρων.

4. *Σφάλματα στρογγύλευσης.* Τα σφάλματα αυτά προκύπτουν λόγω του πεπερασμένου μεγέθους μνήμης που διατίθεται για την αποθήκευση ενός αριθμού στον υπολογιστή.

Υπάρχουν δύο μεγέθη που μετρούν το σφάλμα, το **απόλυτο σφάλμα** και το **απόλυτο σχετικό σφάλμα**.

Ορισμός 1. *Αν $\bar{x}$ είναι μια προσέγγιση του x, το* **απόλυτο σφάλμα** *είναι η ποσότης*

$$|x - \bar{x}|$$

και το **σχετικό σφάλμα** *η ποσότης*

$$\frac{|x - \bar{x}|}{|x|},\ x \neq 0.$$

1.2 Αριθμοί Μηχανής

Οι περισσότεροι υπολογιστές χρησιμοποιούν το δυαδικό σύστημα αρίθμησης. Όπως στο δεκαδικό σύστημα ο αριθμός 432.52 είναι ίσος με

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

κατ' αναλογία ο αριθμός 101.11 του δυαδικού συστήματος είναι ίσος με

$$1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 + 1 \cdot 2^{-1} + 1 \cdot 2^{-2} = 5.75$$

ή

$$(101.11)_2 = (5.75)_{10}$$

όπου ο συμβολισμός $(\cdot)_\beta$ δηλώνει τη βάση του συστήματος αρίθμησης β στο οποίο παριστάνεται ο αριθμός. Η επικοινωνία του υπολογιστή με τον άνθρωπο γίνεται στο δεκαδικό σύστημα. Επειδή όμως η επεξεργασία των δεδομένων από τον υπολογιστή γίνεται στο δυαδικό σύστημα γι' αυτό υπάρχει το ανάλογο λογισμικό για τη μετατροπή των αριθμών από το δεκαδικό στο δυαδικό και αντίστροφα. Κατά τη διαδικασία αυτής της μετατροπής υπεισέρχονται μικρά σφάλματα στρογγύλευσης όπως θα δούμε στη συνέχεια.

Ακέραιοι

Ένας ακέραιος μπορεί να χρησιμοποιήσει όλο το μήκος μιας λέξης για την παράστασή του στη μνήμη εκτός από ένα μόνο δυαδικό ψηφίο (bit) που πρέπει να δεσμευθεί για το πρόσημό του. Αν λοιπόν υποτεθεί ότι διατίθενται n bits για την παράσταση των ψηφίων ενός

ακεραίου αριθμού, τότε ο μεγαλύτερος ακέραιος αριθμός που μπορεί να παρασταθεί στη μνήμη είναι ο:

$$\overbrace{(111\ldots1)}^{n}_2 = 1\cdot 2^{n-1} + 1\cdot 2^{n-2} + \cdots + 1\cdot 2^0 = 2^n - 1.$$

Αν $n = 15$ τότε $2^{15} - 1 = 32.767$. Επομένως, όλοι οι ακέραιοι αριθμοί στην περιοχή $[-(2^n - 1),\ 2^n - 1]$ παριστάνονται με την ορθή (ακριβή) τιμή τους στη μνήμη. Σπάνια όμως υπάρχουν υπολογισμοί που επεξεργάζονται ακεραίους αριθμούς.

Πραγματικοί Αριθμοί - Κινητή Υποδιαστολή

Στο δεκαδικό σύστημα ένας πραγματικός αριθμός μπορεί να παρασταθεί στην **κανονικοποιημένη επιστημονική μορφή**. Αυτό σημαίνει ότι η δεκαδική τελεία μετατοπίζεται έτσι ώστε όλα τα ψηφία του αριθμού να βρίσκονται στα δεξιά της δεκαδικής τελείας και το πρώτο ψηφίο να είναι διάφορο του μηδενός. Για παράδειγμα,

$$15.546 = 0.15564 \cdot 10^2.$$

Έτσι, ένας πραγματικός αριθμός x $(\neq 0)$ μπορεί να παρασταθεί με τη μορφή

$$x = \pm\bar{x} \cdot 10^e \tag{1.1}$$

όπου e είναι ένας ακέραιος, ο οποίος καλείται **εκθέτης** (exponent) και $\bar{x}$ είναι το δεκαδικό τμήμα του αριθμού και καλείται **βάση** (mantissa). Είναι φανερό ότι $0.1 \leq \bar{x} < 1$. Γενικά, ένας αριθμός παριστάνεται σε ένα σύστημα αρίθμησης με βάση β σαν

$$x = \pm\bar{x} \cdot \beta^e, \tag{1.2}$$

όπου

$$\bar{x} = (0.a_1a_2\ldots a_n)_\beta. \tag{1.3}$$

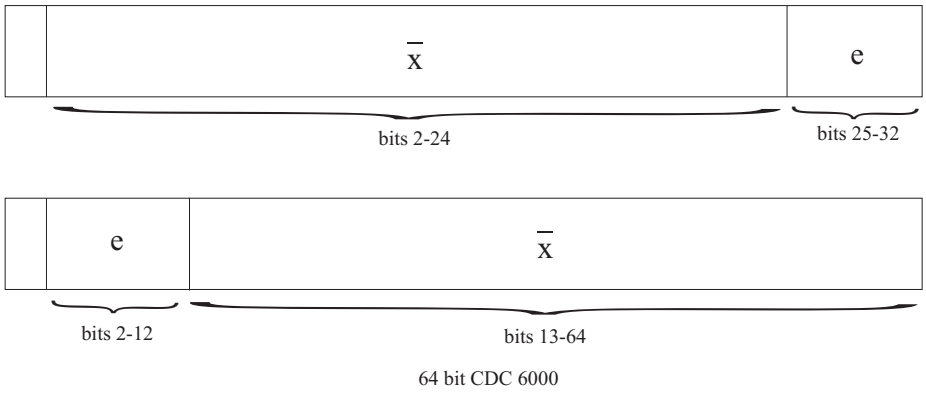
Τα $a_i, i = 1, 2, \ldots n$ είναι όλα ψηφία του συστήματος αρίθμησης με βάση β και ικανοποιούν τις σχέσεις

$$0 \leq a_i \leq \beta - 1,\ a_1 \neq 0. \tag{1.4}$$

Επίσης για τον εκθέτη ισχύει:

$$m \leq e \leq M \tag{1.5}$$

όπου m και M είναι ακέραιοι. Συνήθως $m = -M$ ή $m = -M \pm 1$. Από την (1.3) παρατηρούμε ότι η βάση περιέχει n ψηφία και όλοι οι αριθμοί που έχουν περισσότερα ψηφία πρέπει με κάποιο τρόπο, *να προσεγγισθούν* ώστε να έχουν μόνο n ψηφία. Οι πραγματικοί αριθμοί παριστάνονται στη μνήμη ενός υπολογιστή όπως ακριβώς περιγράφηκε προηγουμένως όπου συνήθως $\beta = 2$. Στο σχήμα 1.1 παριστάνονται πραγματικοί αριθμοί σε υπολογιστές με 64 και 32 bits λέξεις.



Σχήμα 1.1: Παράσταση πραγματικού αριθμού στη μνήμη. Το πρώτο δυαδικό ψηφίο (bit) είναι 0 αν ο αριθμός είναι θετικός και 1 αν είναι αρνητικός.

Αριθμοί Μηχανής

Η παράσταση (1.2) καλείται **κινητής υποδιαστολής** (floating point). Στη συνέχεια θα προσπαθήσουμε να εντοπίσουμε το διάστημα $[s,\ L]$ των πραγματικών αριθμών που μπορούν να παρασταθούν **ορθά** στη μνήμη. Θα υποθέσουμε δηλαδή ότι $\beta = 2$, το x είναι αποθηκευμένο σαν μια ακολουθία από n δυαδικά ψηφία και

$$|e| \leq M. \tag{1.6}$$

Κατ' αρχήν παρατηρούμε ότι η ποσότης $\bar{x}$ φράσσεται ως εξής:

$$\overbrace{(0.10\ldots0)}^{n}_2 \leq \bar{x} \leq \overbrace{(0.11\ldots1)}^{n}_2$$

ή

$$\frac{1}{2} \leq \bar{x} \leq 1 - 2^{-n} \tag{1.7}$$

οπότε

$$s = \frac{1}{2} \cdot 2^e \leq \bar{x} \cdot 2^e = |x| \leq (1 - 2^{-n}) \cdot 2^e = L < 2^e$$

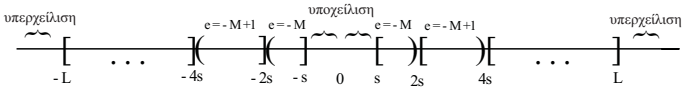
ή

$$2^{e-1} \leq |x| < 2^e. \tag{1.8}$$

Επειδή το $\bar{x}$ είναι της μορφής

$$\bar{x} = (0.1a_2a_3 \ldots a_n)_2$$

συνεπάγεται ότι για κάθε $e = -M, -M+1, \ldots, M-1, M$, υπάρχουν 2^{n-1} κανονικοποιημένα $\bar{x}$. Αυτά τα $\bar{x}$ αντιστοιχούν σε 2^{n-1} ίσης απόστασης αριθμούς x σε κάθε ένα από τα διαστήματα $[2^{e-1}, 2^e)$ και $(-2^e, -2^{e-1}]$ (βλ. Σχήμα 1.2)



Σχήμα 1.2: Παράσταση αριθμών μηχανής.

Όταν αυξάνεται ο εκθέτης κατά 1 διπλασιάζεται το μήκος και των δύο διαστημάτων $[2^{e-1}, 2^e)$ και $(-2^e, -2^{e-1}]$. Συνέπεια του γεγονότος αυτού είναι ότι *οι αριθμοί που παριστάνονται όπως στο σχήμα 1.2 είναι πυκνά κατανεμημένοι πλησίον του μηδενός και αραιά κατανεμημένοι μακρυά του μηδενός*. Οι αριθμοί αυτοί που είναι κινητής υποδιαστολής και για τους οποίους ισχύουν οι (1.6) και (1.7), καλούνται **αριθμοί μηχανής**. Υπάρχει λοιπόν μόνο ένα πεπερασμένο σύνολο πραγματικών αριθμών που μπορεί να παρασταθεί με την ορθή τιμή τους και αυτό βρίσκεται στα δύο διαστήματα $[-L, -s]$ και $[s, L]$. Οποιοσδήποτε αριθμός x για τον οποίο ισχύει $|\bar{x}| > L$ δεν μπορεί να αποθηκευτεί στη μνήμη και το φαινόμενο αυτό είναι γνωστό σαν **υπερχείλιση** (overflow). Όμοια αν $|\bar{x}| < s$ τότε έχουμε το φαινόμενο της **υποχείλισης** (underflow). Ας σημειωθεί ότι τα περισσότερα δυαδικά ψηφία για την παράσταση του $\bar{x}$ αυξάνουν την πυκνότητα των

αριθμών μηχανής, ενώ περισσότερα δυαδικά ψηφία για το e έχει σαν αποτέλεσμα να μεγαλώνει το διάστημα παράστασής τους.

Παράδειγμα

Αν $\beta = 2,\ n = 3,\ m = -1,\ M = 2$, να βρεθούν και να παρασταθούν οι αριθμοί μηχανής.

Λύση

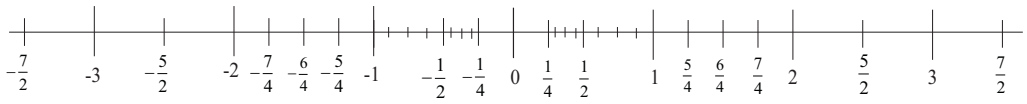
Οι αριθμοί μηχανής έχουν τη μορφή (1.2) - (1.3) ή για τα δεδομένα του παραδείγματος

$$\bar{x} = \pm\bar{x} \cdot 2^e,\ \bar{x} = (0.1a_2a_3)_2$$

με $0 \leq a_i \leq 1,\ i = 2, 3$. Ο μικρότερος $\bar{x}$ είναι ο αριθμός $(0.100)_2$, ενώ οι επόμενοι λαμβάνονται αν κάθε φορά προστίθεται ο αριθμός $(0.100)_2$. Για δεδομένο e έχουμε τους εξής τέσσερις αριθμούς: 0.100, 0.101, 0.110 και 0.111 άρα για $e = -1,\ 0,\ 1,\ 2$ λαμβάνουμε τους ακόλουθους θετικούς αριθμούς μηχανής:

$(0.100)_2 \cdot 2^{-1}$ $(\frac{1}{4})$	$(0.101)_2 \cdot 2^{-1}$ $(\frac{5}{16})$	$(0.110)_2 \cdot 2^{-1}$ $(\frac{6}{16})$	$(0.111)_2 \cdot 2^{-1}$ $(\frac{7}{16})$
$(0.100)_2 \cdot 2^{0}$ $(\frac{1}{2})$	$(0.101)_2 \cdot 2^{0}$ $(\frac{5}{8})$	$(0.110)_2 \cdot 2^{0}$ $(\frac{6}{8})$	$(0.111)_2 \cdot 2^{0}$ $(\frac{7}{8})$
$(0.100)_2 \cdot 2^{1}$ (1)	$(0.101)_2 \cdot 2^{1}$ $(\frac{5}{4})$	$(0.110)_2 \cdot 2^{1}$ $(\frac{6}{4})$	$(0.111)_2 \cdot 2^{1}$ $(\frac{7}{4})$
$(0.100)_2 \cdot 2^{2}$ (2)	$(0.101)_2 \cdot 2^{2}$ $(\frac{5}{2})$	$(0.110)_2 \cdot 2^{2}$ $(\frac{6}{2})$	$(0.111)_2 \cdot 2^{2}$ $(\frac{7}{2})$

Επιπλέον, υπάρχει το μηδέν και το αντίστοιχο σύνολο των αρνητικών αριθμών. Η γραφική παράσταση των αριθμών είναι η ακόλουθη:



Παρατηρούμε ότι δεν υπάρχουν αριθμοί στα διαστήματα $(-\frac{1}{4},\ 0)$ και $(0,\ \frac{1}{4})$. Επίσης, οι αριθμοί δεν είναι αμοιόμορφα κατανεμημένοι. Ωστόσο οι αριθμοί που έχουν κοινό εκθέτη απέχουν ίση απόσταση μεταξύ τους.

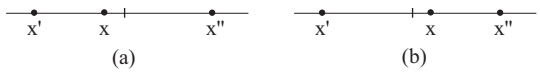
1.3 Ανάλυση σφάλματος των αριθμών κινητής υποδιαστολής

Έστω ο πραγματικός αριθμός κινητής υποδιαστολής

$$x = \pm\,(0.a_1a_2\ldots a_na_{n+1}\ldots)_\beta \cdot \beta^e, a_1 \neq 0 \tag{1.9}$$

όπου χωρίς βλάβη της γενικότητας υποθέτουμε ότι ο β είναι άρτιος (β=2, 8, 10, 16). Εάν υποτεθεί ότι το μέγιστο πλήθος ψηφίων που μπορούν να αποθηκευτούν είναι n, τότε ο αριθμός αυτός δεν μπορεί να παρασταθεί στη μνήμη. Το ερώτημα λοιπόν που τίθεται είναι το εξής: *Ποιός είναι ο πλησιέστερος αριθμός μηχανής προς τον x;* Προκειμένου να δοθεί μια απάντηση στο ερώτημα αυτό ας θεωρήσουμε τους δύο αριθμούς μηχανής μεταξύ των οποίων βρίσκεται ο x (σχήμα 1.3). Όπως παρατηρούμε από το σχήμα 1.3 (α), ο πλησιέστερος αριθμός μηχανής προς τον x είναι ο x' και βρίσκεται αν αποκοπούν τα ψηφία $a_{n+1}\ldots$ από το δεκαδικό τμήμα του (βλ. (1.9)), δηλαδή είναι ο

$$x' = (0.a_1a_2\ldots a_n)_\beta \cdot \beta^e. \tag{1.10}$$



Σχήμα 1.3: Δύο πιθανές θέσεις του x (x', x'' αριθμοί μηχανής).

Εάν όμως έχουμε την περίπτωση του σχήματος 1.3(β), τότε ο x'' είναι πλησιέστερος προς τον x και βρίσκεται αν στον x' προστεθεί η ποσότης $(0.00\ldots01)_\beta$
$= \beta^{-n}$, δηλαδή είναι ο

$$x'' = \left((0.a_1a_2\ldots a_n)_\beta + \beta^{-n}\right) \cdot \beta^e. \tag{1.11}$$

Η τεχνική αυτή καλείται **στρογγύλευση** (rounding up). Χρησιμοποιώντας λοιπόν την τεχνική της στρογγύλευσης ένας πραγματικός αριθμός αντιστοιχεί σε ένα αριθμό μηχανής. Κατ'αυτόν τον τρόπο ένας πραγματικός αριθμός παριστάνεται (προσεγγίζεται) στη μνήμη με τον αντίστοιχο αριθμό μηχανής.

Αν εφαρμόσουμε τη διαδικασία της στρογγύλευσης για την (a) περίπτωση του σχήματος 3.1, έχουμε ότι το απόλυτο σφάλμα φράσσεται ως εξής

$$|x - x'| \leq \frac{1}{2}|x'' - x'| = \left(\frac{1}{2}\beta^{-n}\right) \cdot \beta^e \tag{1.12}$$

ενώ το απόλυτο σχετικό σφάλμα θα έχουμε

$$\frac{|x - x'|}{|x|} \leq \frac{\frac{1}{2}\beta^{-n} \cdot \beta^e}{\overline{x} \cdot \beta^e} = \frac{\frac{1}{2}\beta^{-n}}{\overline{x}} \leq \frac{\frac{1}{2}\beta^{-n}}{\beta^{-1}} = \frac{1}{2}\beta^{-n+1}. \tag{1.13}$$

Εύκολα διαπιστώνεται ότι οι (1.12) και (1.13) ισχύουν και για την περίπτωση β του σχήματος 1.3. Συνήθως ο αριθμός μηχανής κινητής υποδιαστολής που είναι πλησιέστερος προς τον x συμβολίζεται με $fl\,(x)$. Χρησιμοποιώντας το συμβολισμό αυτό οι τύποι (1.12) και (1.13) γράφονται

$$|x - fl\,(x)| \leq \frac{1}{2}\beta^{-n} \cdot \beta^e \tag{1.14}$$

και

$$\frac{|x - fl\,(x)|}{|x|} \leq \frac{1}{2}\beta^{-n+1}, \tag{1.15}$$

αντίστοιχα. Οι τύποι (1.14) και (1.15) παρέχουν άνω φράγματα του σφάλματος στρογγύλευσης. Εάν τεθεί

$$\varepsilon = \frac{fl\,(x) - x}{x},$$

τότε η (1.15) γράφεται

$$fl\,(x) = (1 + \varepsilon)\, x \tag{1.16}$$

με

$$|\varepsilon| \leq \frac{1}{2}\beta^{-n+1}. \tag{1.17}$$

Η ποσότης ε καλείται **μονάδα μηχανής** (machine unit), επειδή δε είναι μικρή, η (1.16) δηλώνει ότι ο $fl\,(x)$ είναι μια ελαφρά διατάραξη του x. Μία άλλη τεχνική απεικόνισης των πραγματικών αριθμών στο σύνολο των αριθμών μηχανής είναι αυτή της **αποκοπής** (chopping). Εάν ακολουθηθεί η τεχνική αυτή, τότε ο πραγματικός αριθμός x προσεγγίζεται πάντα με τον πλησιέστερο από τα αριστερά

του αριθμού μηχανής, δηλαδή με τον x' (βλ. Σχ. 1.3). Στην περίπτωση που χρησιμοποιηθεί η διαδικασία της αποκοπής, τότε με ανάλογο τρόπο εύκολα βρίσκεται ότι

$$|x - fl\,(x)| \leq \beta^{e-n} \tag{1.18}$$

και

$$\frac{|x - fl\,(x)|}{|x|} \leq \beta^{-n+1}. \tag{1.19}$$

Συγκρίνοντας τους τύπους (1.14), (1.15) με τους (1.18) και (1.19) παρατηρούμε ότι:

1. *Το χειρότερο σφάλμα στρογγύλευσης είναι το μισό εκείνου της αποκοπής.*

2. *Το σφάλμα στη στρογγύλευση είναι αρνητικό στις μισές περίπου περιπτώσεις και θετικό στις άλλες μισές με αποτέλεσμα την απαλοιφή του, ενώ στην αποκοπή έχει συνέχεια το ίδιο πρόσημο.*

Η μελέτη του σφάλματος στρογγύλευσης είναι ένα σημαντικό τμήμα της Αριθμητικής Ανάλυσης και αναπτύχθηκε από τον Wilkinson. Η διεξοδική παρουσίασή της ξεφεύγει από τα πλαίσια του παρόντος βιβλίου. Η αξία της στην αξιολόγηση μιας αριθμητικής μεθόδου είναι αναγκαία και σημαντική όπως φαίνεται στο παράδειγμα που ακολουθεί.

1.4 Ανάλυση σφάλματος στο άθροισμα όρων

Στην παράγραφο αυτή θα μελετηθεί το σφάλμα στρογγύλευσης στην πράξη του αθροίσματος ενός μεγάλου πλήθους όρων και θα προταθεί μια μέθοδος για την ελαχιστοποίησή του.

Έστω ότι έχουμε το άθροισμα

$$S = \sum_{i=1}^{n} x_i \tag{1.20}$$

όπου x_i είναι αριθμοί κινητής υποδιαστολής που ήδη έχουν αποθηκευτεί στη μνήμη. Υποθέτουμε ότι προσθέτουμε τους δύο πρώτους, οπότε

$$S_2 = fl\,(x_1 + x_2)$$

και στο αποτέλεσμα προσθέτουμε τον 3ο όρο κ.ο.κ., συνεπώς

$$\begin{aligned} S_3 &= fl\,(x_3 + S_2) \\ S_4 &= fl\,(x_4 + S_3) \\ &\vdots \\ S_n &= fl\,(x_n + S_{n-1}) \end{aligned} \tag{1.21}$$

όπου S_n είναι το αποτέλεσμα του υπολογισμού του S. Λόγω της (1.16), οι ανωτέρω σχέσεις γράφονται

$$\begin{aligned} S_2 &= (x_1 + x_2)\,(1 + \varepsilon_2) \\ S_3 &= (x_3 + S_2)\,(1 + \varepsilon_3) \\ &\vdots \\ S_n &= (x_n + S_{n-1})\,(1 + \varepsilon_n) \end{aligned} \tag{1.22}$$

όπου

$$|\varepsilon_i| \leq \frac{1}{2}\beta^{-n+1}, \quad i = 2, 3, \cdots, n.$$

Αναπτύσσοντας τις πρώτες ποσότητες της (1.22) έχουμε

$$\begin{aligned} S_2 &= (x_1 + x_2) + (x_1 + x_2)\,\varepsilon_2 \\ S_3 &= [(x_1 + x_2 + x_3) + (x_1 + x_2)\,\varepsilon_2]\,(1 + \varepsilon_3) \\ &= (x_1 + x_2 + x_3) + (x_1 + x_2)\,\varepsilon_2 + (x_1 + x_2 + x_3)\,\varepsilon_3 + (x_1 + x_2)\,\varepsilon_2\varepsilon_3 \end{aligned}$$

ή παραλείποντας τον τελευταίο όρο, επειδή $\varepsilon_2\varepsilon_3 \ll \varepsilon_2, \varepsilon_3$, λαμβάνουμε

$$S_3 \simeq (x_1 + x_2 + x_3) + (x_1 + x_2)\,\varepsilon_2 + (x_1 + x_2 + x_3)\,\varepsilon_3.$$

Αναγωγικά βρίσκουμε τελικά ότι

$$S_n \simeq \sum_{i=1}^{n} x_i + (x_1 + x_2)\,\varepsilon_2 + (x_1 + x_2 + x_3)\,\varepsilon_3 + \ldots + (x_1 + x_2 + \ldots + x_n)\,\varepsilon_n$$

ή

$$\begin{aligned} S_n - S \simeq\ & x_1\,(\varepsilon_2 + \varepsilon_3 + \ldots + \varepsilon_n) + x_2\,(\varepsilon_2 + \varepsilon_3 + \ldots + \varepsilon_n) \\ & + x_3\,(\varepsilon_3 + \varepsilon_4 + \ldots + \varepsilon_n) + \ldots + x_n\varepsilon_n \end{aligned}$$

ή

$$|S_n - S| \lesssim |x_1|\left(|\varepsilon_2| + \ldots + |\varepsilon_n|\right) + |x_2|\left(|\varepsilon_2| + \ldots + |\varepsilon_n|\right) + |x_3|\left(|\varepsilon_3| + \ldots + |\varepsilon_n|\right) + \ldots + x_n\,|\varepsilon_n|\,. \tag{1.23}$$

Παρατηρώντας προσεκτικά την (1.23) και προσπαθώντας να ελαχιστοποιήσουμε το απόλυτο σφάλμα $|S - S_n|$ καταλήγουμε στο συμπέρασμα ότι: οι όροι θα πρέπει να διαταχθούν, πριν τον υπολογισμό του αθροίσματός τους, έτσι ώστε

$$|x_1| \leq |x_2| \leq |x_3| \leq \ldots |x_n|\,.$$

Υπό την προυπόθεση αυτή, οι όροι στο δεξί μέλος της (1.23) με το μεγαλύτερο πλήθος σφαλμάτων ε_i θα πολλαπλασιάζονται με τις μικρότερες τιμές μεταξύ των x_i.

1.5 Διαδιδόμενο σφάλμα

Για τον έλεγχο της προσέγγισης του $\overline{x}$ σε σχέση με την τιμή του x χρησιμοποιούμε τα ακόλουθα κριτήρια που προκύπτουν από τις (1.14) και (1.15), αντίστοιχα.

Εάν

$$|\varepsilon_{\overline{x}}| = |x - \overline{x}| < \frac{1}{2}10^{-d}$$

τότε ο $\overline{x}$ προσεγγίζει τον x σε d δεκαδικά ψηφία.

Εάν $x \neq 0$, τότε μπορεί να χρησιμοποιηθεί το σχετικό σφάλμα προκειμένου να εξασφαλισθεί προσέγγιση σε ένα επιθυμητό πλήθος σημαντικών ψηφίων. Σημαντικά ψηφία ενός δεκαδικού αριθμού είναι όλα τα ψηφία του αριθμού, από αριστερά προς τα δεξιά, του πρώτου μη μηδενικού ψηφίου (συμπεριλαμβανομένου).

Εάν

$$|\varrho_{\overline{x}}| = \left|\frac{\varepsilon_{\overline{x}}}{x}\right| \leq \frac{1}{2}10^{-s}$$

τότε ο $\overline{x}$ προσεγγίζει τον x σε s σημαντικά ψηφία.

Ας συμβολίσουμε με $\square$ μια αριθμητική πράξη $(+, -, \times, /)$ και με $\square^*$ την ίδια πράξη που εκτελείται στον υπολογιστή, η οποία περιέχει το σφάλμα στρογγύλευσης. Έστω $\overline{x}$ και $\overline{y}$ οι αριθμοί που χρησιμοποιούνται στους υπολογισμούς και είναι οι προσεγγίσεις των τιμών

$$x = \overline{x} + \varepsilon_{\overline{x}} \quad y = \overline{y} + \varepsilon_{\overline{y}} \tag{1.24}$$

όπου $\varepsilon_{\overline{x}}$ και $\varepsilon_{\overline{y}}$ σφάλματα. Τότε με την εκτέλεση της αριθμητικής πράξης $\square$ ο αριθμός που υπολογίζεται είναι ο $\overline{x}\square^*\overline{y}$ και το συνολικό σφάλμα δίνεται από τον τύπο

$$\begin{aligned} \varepsilon_{\overline{x}\square\overline{y}} &= x\square y - \overline{x}\square^*\overline{y} \\ &= (x\square y - \overline{x}\square\overline{y}) + (\overline{x}\square\overline{y} - \overline{x}\square^*\overline{y})\,. \end{aligned} \tag{1.25}$$

Ο πρώτος όρος στο δεξί μέλος της (1.25) είναι το διαδιδόμενο σφάλμα και ο δεύτερος όρος είναι το σφάλμα στρογγύλευσης κατά τον υπολογισμό του $\overline{x}\square\overline{y}$. Επειδή όμως

$$fl\,(\overline{x}\square\overline{y}) = \overline{x}\square^*\overline{y} \tag{1.26}$$

που σημαίνει ότι το $\overline{x}\square\overline{y}$ υπολογίζεται ακριβώς και στη συνέχεια στρογγυλεύεται. Λόγω των (1.15) και (1.26) έχουμε

$$|\overline{x}\square\overline{y} - \overline{x}\square^*\overline{y}| \leq \frac{1}{2}\beta^{-n+1}\,|\overline{x}\square\overline{y}| \tag{1.27}$$

με την υπόθεση ότι χρησιμοποιείται στρογγύλευση. Υποθέτοντας ότι το σφάλμα στρογγύλευσης είναι μικρό θα μελετηθεί η συμπεριφορά του σφάλματος διάδοσης. Εάν $\square = \pm$ τότε από την (1.25) έχουμε

$$\begin{aligned} \varepsilon_{\overline{x}\pm\overline{y}} &= (x \pm y) - (\overline{x} \pm \overline{y}) \\ &= (x - \overline{x}) \pm (y - \overline{y}) \end{aligned}$$

ή

$$\varepsilon_{\overline{x}\pm\overline{y}} = \varepsilon_{\overline{x}} \pm \varepsilon_{\overline{y}}$$

και τέλος για τα απόλυτα σφάλματα έχουμε τη σχέση

$$|\varepsilon_{\overline{x}\pm\overline{y}}| \leq |\varepsilon_{\overline{x}}| + |\varepsilon_{\overline{y}}|\,. \tag{1.28}$$

Ισχύει δηλαδή το ακόλουθο θεώρημα.

Θεώρημα 1.5.1. *Η μέγιστη τιμή του απολύτου σφάλματος του αθροίσματος ή της διαφοράς δύο αριθμών είναι ίση με το άθροισμα των απολύτων αφαλμάτων των αριθμών αυτών.*

Λόγω του ανωτέρω θεωρήματος, αν $\overline{x}$ και $\overline{y}$ έχουν ακρίβεια τεσσάρων δεκαδικών ψηφίων (δηλαδή, εάν $|\varepsilon_{\overline{x}}|\,, |\varepsilon_{\overline{x}}| < \frac{1}{2}\cdot 10^{-4}$), τότε η ποσότητα $\overline{x} \pm \overline{y}$ μπορεί να διαφέρει από την $x \pm y$ το πολύ κατά 10^{-4}. Επομένως είναι πιθανό η ποσότητα $\overline{x} \pm \overline{y}$ να έχει ένα ψηφίο λάθος

στην τέταρτη δεκαδική θέση. Επιπλέον, αν οι $\overline{x}$ και $\overline{y}$ έχουν διαφορετική ακρίβεια, τότε η ποσότητα $\overline{x} \pm \overline{y}$ στη χειρότερη περίπτωση θα είναι λανθασμένη από εκείνη τη δεκαδική θέση που αντιστοιχεί στο μεγαλύτερο από τα $|\varepsilon_{\overline{x}}|$ και $|\varepsilon_{\overline{y}}|$.

Εάν $\square = \cdot$, τότε

$$\varepsilon_{\overline{xy}} = xy - (x - \varepsilon_{\overline{x}})(y - \varepsilon_{\overline{y}})$$

ή

$$\varepsilon_{\overline{xy}} = y\varepsilon_{\overline{x}} + x\varepsilon_{\overline{y}} + \varepsilon_{\overline{x}}\varepsilon_{\overline{y}} \tag{1.29}$$

Όμοια εάν $\square = /$ και $y, \overline{y} \neq 0$, τότε

$$\varepsilon_{\overline{x}/\overline{y}} = \frac{x}{y} - \frac{\overline{x}}{\overline{y}} = \frac{x}{y} - \frac{x - \varepsilon_{\overline{x}}}{y - \varepsilon_{\overline{y}}} = \frac{y\varepsilon_{\overline{x}} - x\varepsilon_{\overline{y}}}{y^2 + y\varepsilon_{\overline{y}}} \tag{1.30}$$

Διατηρώντας τους υπερέχοντες όρους οι (1.29) και (1.30) δίνουν

$$\varepsilon_{\overline{xy}} \simeq x\varepsilon_{\overline{y}} + y\varepsilon_{\overline{x}} \tag{1.31}$$

και

$$\varepsilon_{\overline{x}/\overline{y}} \simeq \frac{y\varepsilon_{\overline{x}} - x\varepsilon_{\overline{y}}}{y^2} \tag{1.32}$$

αντίστοιχα. *Από την (1.31) παρατηρούμε ότι μεγάλες τιμές του $\overline{x}$ ή του $\overline{y}$ έχουν σαν αποτέλεσμα την αύξηση του σφάλματος στο γινόμενο xy. Όμοια, η (1.32) θα παράγει μεγάλο σφάλμα στη διαίρεση $\overline{x}/\overline{y}$ για μεγάλες τιμές του $\overline{x}$ και/ή μικρές τιμές του $\overline{y}$.*

Το συμπέρασμα είναι ότι τέτοιου είδους πράξεις (δηλ. πολ/σμός με μεγάλους αριθμούς και διαιρέσεις όπου ο διαιρετέος είναι ο μεγάλος αριθμός και/ή ο διαιρέτης είναι μικρός) θα πρέπει να αποφεύγονται με αναδιάταξη των υπολογισμών.

Θεώρημα 1.5.2. *Η μέγιστη τιμή του απόλυτου σχετικού σφάλματος του γινομένου ή του πηλίκου δύο αριθμών είναι κατα προσέγγιση ίση με το άθροισμα των απολύτων σχετικών σφαλμάτων των αριθμών αυτών.*

Απόδειξη. Διαιρώντας την (1.31) δια xy λαμβάνουμε

$$\varrho_{\overline{xy}} = \frac{\varepsilon_{\overline{xy}}}{xy} \backsimeq \frac{\varepsilon_{\overline{x}}}{x} + \frac{\varepsilon_{\overline{y}}}{y} \tag{1.33}$$

ή

$$|\varrho_{\overline{xy}}| \lesssim |\varrho_{\overline{x}}| + |\varrho_{\overline{y}}| \tag{1.34}$$

και αποδείχθηκε το θεώρημα για το γινόμενο δύο αριθμών. Όμοια για το πηλίκο έχουμε διαιρώντας την (1.32) δια ξ/ψ

$$\varrho_{\overline{x}/\overline{y}} \simeq \frac{y\varepsilon_{\overline{x}} - x\varepsilon_{\overline{y}}}{xy} = \frac{\varepsilon_{\overline{x}}}{x} - \frac{\varepsilon_{\overline{y}}}{y} \tag{1.35}$$

ή

$$\varrho_{\overline{x}/\overline{y}} \simeq \varrho_{\overline{x}} - \varrho_{\overline{y}} \tag{1.36}$$

ή

$$\left|\varrho_{\overline{x}/\overline{y}}\right| \lesssim \left|\varrho_{\overline{x}}\right| + \left|\varrho_{\overline{y}}\right|. \tag{1.37}$$

■

Από το παραπάνω θεώρημα προκύπτει ότι εάν x και y έχουν ακρίβεια s σημαντικών ψηφίων, τότε οι ποσότητες xy και x/y θα έχουν περίπου ακρίβεια s σημαντικών ψηφίων. Τέλος, παρατηρούμε ότι

$$\varrho_{\overline{x}\pm\overline{y}} = \frac{\varepsilon_{\overline{x}\pm\overline{y}}}{x \pm y} = \frac{\varepsilon_{\overline{x}}}{x \pm y} \pm \frac{\varepsilon_{\overline{y}}}{x \pm y} = \left(\frac{x}{x \pm y}\right)\varrho_{\overline{x}} \pm \left(\frac{y}{x \pm y}\right)\varrho_{\overline{y}}. \tag{1.38}$$

Η (4.43) δηλώνει ότι αν $x \pm y$ είναι πολύ μικρότερο από το x ή y, τότε οι παράγοντες $x/(x \pm y)$ και $y/(x \pm y)$ θα λάβουν μεγάλες τιμές με συνέπεια να αυξηθεί η τιμή του $\varrho_{\overline{x}\pm\overline{y}}$. *Για το λόγο αυτό θα πρέπει να αποφεύγεται η πρόσθεση ενός πολύ μεγάλου και ενός πολύ μικρού αριθμού ή η αφαίρεση δύο περίπου ίσων αριθμών.*

Κεφάλαιο 2

Άμεσες μέθοδοι για την επίλυση γραμμικών συστημάτων

2.1 Η μέθοδος απαλοιφής του Gauss

Στο παρόν κεφάλαιο θα αναπτύξουμε υπολογιστικές μεθόδους για την αριθμητική επίλυση μεγάλων συστημάτων αλγεβρικών εξισώσεων. Θα περιορισθούμε σε συστήματα των οποίων ο πίνακας των συντελεστών των αγνώστων είναι τετραγωνικοί και υπάρχει μια μοναδική λύση. Η λύση ενός συστήματος 20 εξισώσεων με τη μέθοδο του Gramer δεν είναι καθόλου πρακτική, γιατί ένας υπολογιστής που εκτελεί 2 εκατομμύρια πράξεις το δευτερόλεπτο, θα ήθελε 2 εκατομμύρια χρόνια για να βρεί τη λύση του παραπάνω προβλήματος! Το πλήθος των πράξεων όμως δεν είναι βασικό μόνο για την ελάττωση του υπολογιστικού χρόνου, αλλά είναι εξ'ίσου καθοριστικό για την ακρίβεια των υπολογισμών λόγω συσσώρευσης των σφαλμάτων στρογγύλευσης. Μία μέθοδος για την επίλυση γραμμικών συστημάτων είναι η μέθοδος της απαλοιφής του Gauss. Ας θεωρήσουμε το σύστημα

$$\begin{array}{l} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + \ldots + a_{1n}^{(1)}x_n = b_1^{(1)} \\ a_{21}^{(1)}x_1 + a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)} \\ a_{31}^{(1)}x_1 + a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 + \ldots + a_{3n}^{(1)}x_n = b_3^{(1)} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}^{(1)}x_1 + a_{n2}^{(1)}x_2 + a_{n3}^{(1)}x_3 + \ldots + a_{nn}^{(1)}x_n = b_n^{(1)} \end{array} \qquad (2.1)$$

το οποίο μπορεί να γραφεί και

$$\begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} \\ a_{31}^{(1)} & a_{32}^{(1)} & a_{33}^{(1)} & \cdots & a_{3n}^{(1)} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{n1}^{(1)} & a_{n2}^{(1)} & a_{n3}^{(1)} & \cdots & a_{nn}^{(1)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \\ b_3^{(1)} \\ \vdots \\ b_n^{(1)} \end{bmatrix} \tag{2.2}$$

ή απλά

$$A^{(1)}x = b^{(1)}. \tag{2.3}$$

Υποθέτουμε ότι $\det A^{(1)} \neq 0$ έτσι ώστε να υπάρχει μία και μοναδική λύση. Επίσης υποθέτουμε $b^{(1)} \neq$ του μηδενικού διανύσματος. Το πρώτο βήμα είναι να αντικαταστήσουμε το σύστημα (2.1) με ένα ισοδύναμο σύστημα, το οποίο είναι απλούστερο από το (2.1). Το σύστημα

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + \ldots + a_{1n}^{(1)}x_n &= b_1^{(1)} \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 + \ldots + a_{2n}^{(2)}x_n &= b_2^{(2)} \\ a_{32}^{(2)}x_2 + a_{33}^{(2)}x_3 + \ldots + a_{3n}^{(2)}x_n &= b_3^{(2)} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \quad \vdots \\ a_{n2}^{(2)}x_2 + a_{n3}^{(2)}x_3 + \ldots + a_{nn}^{(2)}x_n &= b_n^{(2)} \end{aligned} \tag{2.4}$$

είναι απλούστερο από το (2.1) γιατί η μεταβλητή x_1 έχει απαλειφθεί από όλες τις εξισώσεις εκτός από την πρώτη. Το σύστημα (2.4) μπορεί να γραφτεί

$$\left[\begin{array}{c|cccc} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \ldots & a_{1n}^{(1)} \\ \hline & a_{22}^{(2)} & a_{23}^{(2)} & \ldots & a_{2n}^{(2)} \\ \mathbf{0} & a_{32}^{(2)} & a_{33}^{(2)} & \ldots & a_{3n}^{(2)} \\ & \vdots & \vdots & \cdots & \vdots \\ & a_{n2}^{(2)} & a_{n3}^{(2)} & \ldots & a_{nn}^{(2)} \end{array}\right] \left[\begin{array}{c} x_1 \\ \hline x_2 \\ x_3 \\ \vdots \\ x_n \end{array}\right] = \left[\begin{array}{c} b_1^{(1)} \\ \hline b_2^{(2)} \\ b_3^{(2)} \\ \vdots \\ b_n^{(2)} \end{array}\right]$$

ή απλούστερα

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

Θεωρώντας το σύστημα που αποτελείται από όλες τις εξισώσεις εκτός από την πρώτη (βλ.(2.4)) επαναλαμβάνουμε την ίδια διαδικασία

απαλείφοντας τώρα τον άγνωστο x_2. Το δεύτερο βήμα είναι το εξής

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 &+ \ldots + a_{1n}^{(1)}x_n &= b_1^{(1)} \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 &+ \ldots + a_{2n}^{(2)}x_n &= b_2^{(2)} \\ a_{33}^{(3)}x_3 &+ \ldots + a_{3n}^{(3)}x_n &= b_3^{(3)} \\ \vdots \quad &\ldots &\vdots \\ a_{n3}^{(3)}x_3 &+ \ldots + a_{nn}^{(3)}x_n &= b_n^{(3)} \end{aligned}$$

ή

$$\left[\begin{array}{ccccc} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \ldots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & a_{23}^{(2)} & \ldots & a_{2n}^{(2)} \\ \hline & & a_{33}^{(3)} & \ldots & a_{3n}^{(3)} \\ & \mathbf{0} & \vdots & \ldots & \vdots \\ & & a_{n3}^{(3)} & \ldots & a_{nn}^{(3)} \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \\ \hline x_3 \\ \vdots \\ x_n \end{array}\right] = \left[\begin{array}{c} b_1^{(1)} \\ b_2^{(2)} \\ \hline b_3^{(3)} \\ \vdots \\ b_n^{(3)} \end{array}\right]$$

ή

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

Μετά από $r-1$ τέτοια βήματα θα έχουμε

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \ldots + a_{1,r-1}^{(1)}x_{r-1} + a_{1,r}^{(1)}x_r + \ldots + a_{1n}^{(1)}x_n &= b_1^{(1)} \\ a_{22}^{(2)}x_2 + \ldots + a_{2,r-1}^{(2)}x_{r-1} + a_{2,r}^{(2)}x_r + \ldots + a_{2n}^{(2)}x_n &= b_2^{(2)} \\ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots &\quad \vdots \\ a_{r-1,r-1}^{(r-1)}x_{r-1} + a_{r-1,r}^{(r-1)}x_r + \ldots + a_{r-1,n}^{(r-1)}x_n &= b_{r-1}^{(r-1)} \\ a_{r,r}^{(r)}x_r + \ldots + a_{rn}^{(r)}x_n &= b_r^{(r-1)} \\ \ldots \quad \ldots \quad \ldots &\quad \vdots \\ a_{n,r}^{(r)}x_r + \ldots + a_{nn}^{(r)}x_n &= b_n^{(r)} \end{aligned}$$

το οποίο μπορεί να γραφεί

$$\left[\begin{array}{cccc|ccc} a_{11}^{(1)} & a_{12}^{(1)} & \ldots & a_{1,r-1}^{(1)} & a_{1r}^{(1)} & \ldots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \ldots & a_{2,r-1}^{(2)} & a_{2r}^{(2)} & \ldots & a_{2n}^{(2)} \\ \mathbf{0} & & \ddots & \vdots & \vdots & \ldots & \vdots \\ & & & a_{r-1,r-1}^{(r-1)} & a_{r-1,r}^{(r-1)} & \ldots & a_{r-1,n}^{(r-1)} \\ \hline & & & & a_{rr}^{(r)} & \ldots & a_{rn}^{(r)} \\ & \mathbf{0} & & & \vdots & \ldots & \vdots \\ & & & & a_{nr}^{(r)} & \ldots & a_{nn}^{(r)} \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_{r-1} \\ \hline x_r \\ \vdots \\ x_n \end{array}\right] = \left[\begin{array}{c} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_{r-1}^{(r-1)} \\ \hline b_r^{(r)} \\ \vdots \\ b_n^{(r)} \end{array}\right]$$

ή

$$A^{(r)}x = b^{(r)}.$$

Τελικά, μετά από $n-1$ τέτοια βήματα το αρχικό σύστημα μετατρέπεται στο ακόλουθο τριγωνικό σύστημα εξισώσεων

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \ldots + a_{1,r-1}^{(1)}x_{r-1} + a_{1,r}^{(1)}x_r + \ldots + a_{1n}^{(1)}x_n &= b_1^{(1)} \\ a_{22}^{(2)}x_2 \ldots + a_{2,r-1}^{(2)}x_{r-1} + a_{2,r}^{(2)}x_r + \ldots + a_{2n}^{(2)}x_n &= b_2^{(2)} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \vdots \\ a_{n-1,n-1}^{(n-1)}x_{n-1} + a_{n-1,n}^{(n-1)}x_n &= b_{n-1}^{(n-1)} \\ a_{nn}^{(n)}x_n &= b_n^{(n)} \end{aligned} \tag{2.5}$$

το οποίο μπορεί να γραφτεί σαν

$$A^{(n)}x = b^{(n)}, \tag{2.6}$$

όπου $A^{(n)}$ είναι ένας άνω τριγωνικός πίνακας. Το σύστημα (2.5) μπορεί να λυθεί πολύ εύκολα με την προς τα πίσω αντικατάσταση από τον τύπο

$$x_n = \frac{b_n^{(n)}}{a_{nn}^{(n)}}$$

και

$$x_i = \frac{b_i^{(i)} - \sum_{j=i+1}^{n} a_{ij}^{(i)}x_j}{a_{ii}^{(i)}}, \quad i = 1(1)n-1. \tag{2.7}$$

Πιο αναλυτικά η μέθοδος απαλοιφής του Gauss δημιουργεί μία ακολουθία από πίνακες $\left\{A^{(k)}\right\}, k = 1(1)n$, όπου $A^{(1)} = A$ και μια ακολουθία από δεξιά μέλη $\left\{b^{(k)}\right\}, k = 1(1)n$ τέτοια ώστε ο $A^{(n)} \equiv U$ είναι ένας άνω τριγωνικός πίνακας. Το πρώτο βήμα της μεθόδου είναι να φυλάξουμε την πρώτη από τις εξισώσεις για να την χρησιμοποιήσουμε αργότερα και να απαλείψουμε τον άγνωστο x_1 από τις υπόλοιπες $n-1$ εξισώσεις. Έτσι άν χρησιμοποιήσουμε τους συμβολισμούς

$$\begin{matrix} a_{ij}^{(1)} = a_{ij} \\ b_i^{(1)} = b_i \end{matrix}, \quad i = 1(1)n, \quad j = 1(1)n,$$

τότε προκειμένου να απαλειφθεί ο x_1 από την i-οστή εξίσωση του (2.1) για $i = 2(1)n$, προσθέτουμε m_{i1} φορές την πρώτη εξίσωση, η

οποία καλείται **οδηγός εξίσωση**, στην i-οστή εξίσωση, όπου

$$m_{i1} = -\frac{a_{i1}^{(1)}}{a_{11}^{(1)}}, \quad i = 2(1)n \tag{2.8}$$

με $a_{11}^{(1)} \neq 0$. Το $a_{11}^{(1)}$ καλείται **οδηγό στοιχείο**. Το αποτέλεσμα αυτής της εργασίας θα είναι το ακόλουθο

$$\left[a_{i2}^{(1)} + m_{i1}a_{12}^{(1)}\right] x_2 + \ldots + \left[a_{in}^{(1)} + m_{i1}a_{1n}^{(1)}\right] x_n = b_1^{(1)} + m_{i1}b_1^{(1)}$$

ή

$$a_{i2}^{(2)}x_2 + \ldots + a_{in}^{(2)}x_n = b_i^{(2)}, \quad i = 2(1)n$$

όπου

$$a_{ij}^{(2)} = a_{ij}^{(1)} + m_{i1}a_{1j}^{(1)}, \quad i = 2(1)n, \quad j = 2(1)n \tag{2.9}$$

και

$$b_i^{(2)} = b_i^{(1)} + m_{i1}b_1^{(1)}, \quad i = 2(1)n.$$

Παρατηρούμε ότι ο άγνωστος x_1 πράγματι απαλείφεται από την i-οστή εξίσωση αφού

$$a_{i1}^{(2)} = a_{i1}^{(1)} + m_{i1}a_{11}^{(1)} = 0.$$

Στη συνέχεια θεωρούμε το σύστημα που αποτελείται από όλες τις εξισώσεις εκτός από την οδηγό, δηλαδή από τις $n-1$ εξισώσεις

$$\begin{aligned} a_{22}^{(2)}x_2 + \ldots + a_{2n}^{(2)}x_n &= b_2^{(2)} \\ &\vdots \\ a_{n2}^{(2)}x_2 + \ldots + a_{nn}^{(2)}x_n &= b_n^{(2)}. \end{aligned}$$

Έτσι αν $a_{22}^{(2)} \neq 0$, τότε ορίζουμε τους πολλαπλασιαστές m_{i2} για την απαλοιφή του x_2 από τις $n-2$ τελευταίες εξισώσεις κ.ο.κ. Κατ' αυτόν τον τρόπο καταλήγουμε στο άνω τριγωνικό σύστημα (2.6). Μπορούμε τώρα εύκολα να παρατηρήσουμε ότι αν $M^{(1)}$ είναι ο πίνακας

$$M^{(1)} = \left[\begin{array}{c|c} 1 & 0 \\ \hline m_{21} & \\ m_{31} & I_{n-1} \\ \vdots & \\ m_{n1} & \end{array}\right] \tag{2.10}$$

τότε το πρώτο βήμα της απαλοιφής του Gauss μπορεί να γίνει πολλαπλασιάζοντας από αριστερά το αρχικό σύστημα επί $M^{(1)}$. Έτσι έχουμε:

$$M^{(1)}A^{(1)}x = M^{(1)}b^{(1)} \tag{2.11}$$

ή

$$A^{(2)}x = b^{(2)} \tag{2.12}$$

όπου

$$A^{(2)} = M^{(1)}A^{(1)} \text{ και } b^{(2)} = M^{(1)}b^{(1)}.$$

Θεώρημα 2.1.1. *Το σύστημα $A^{(1)}x = b^{(1)}$ είναι ισοδύναμο με το $A^{(2)}x = b^{(2)}$.*

Απόδειξη. Υποθέσαμε ότι το (2.1) έχει μια μοναδική λύση, δηλαδή $\det A^{(1)} \neq 0$. Επίσης από την (2.11) και (2.12) έχουμε

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

ή

$$\det A^{(2)} = \left[\det M^{(1)}\right]\left[\det A^{(1)}\right] = \det A^{(1)} \neq 0$$

επειδή $\det M^{(1)} = 1$. Άρα ο $A^{(2)}$ είναι μη ιδιάζων και το σύστημα (2.4) έχει μία μοναδική λύση. Αλλά η (2.11) δείχνει ότι η λύση του (2.1) ικανοποιεί το (2.4), συνεπώς οι μοναδικές λύσεις των (2.1) και (2.4) ταυτίζονται. ■

Παρατηρούμε τώρα ότι το πρώτο βήμα είναι τυπικό καθόσον μετά από $r-1$ απαλοιφές θα έχουμε, αν $a_{rr}^{(r)} \neq 0$, την οδηγό εξίσωση

$$a_{rr}^{(r)}x_r + \ldots + a_{rn}^{(r)}x_n = b_r^{(r)}.$$

Προκειμένου να απαλειφθεί ο x_r από τις υπόλοιπες $n-r$ εξισώσεις προσθέτουμε m_{ir} φορές την οδηγό εξίσωση στην i-οστή για $i = r+1(1)n$. Έτσι ορίζοντας

$$m_{ir} = -\frac{a_{ir}^{(r)}}{a_{rr}^{(r)}}, \quad i = r+1(1)n$$

έχουμε

$$\left[a_{i,r+1}^{(r)} + m_{ir}a_{r,r+1}^{(r)}\right]x_{r+1} + \ldots + \left[a_{in}^{(r)} + m_{ir}a_{rn}^{(r)}\right]x_n = \left[b_i^{(r)} + m_{ir}b_r^{(r)}\right]$$

η οποία μπορεί να γραφεί σαν

$$a_{i,r+1}^{(r+1)}x_{r+1} + \ldots + a_{in}^{(r+1)}x_n = b_i^{(r+1)}, \quad i = r+1(1)n$$

όπου

$$a_{ij}^{(r+1)} = a_{ij}^{(r)} + m_{ir}a_{rj}^{(r)}, \quad i = r+1(1)n, \quad j = r+1(1)n \tag{2.13}$$

και

$$b_i^{(r+1)} = b_i^{(r)} + m_{ir}b_r^{(r)}, \quad i = r+1(1)n.$$

Καταλήγουμε λοιπόν στο σύστημα

$$A^{(r+1)}x = b^{(r+1)}. \tag{2.14}$$

Είναι πάλι εύκολο να διαπιστωθεί ότι αν

$$M^{(r)} = \left[\begin{array}{c|ccccc} I_{r-1} & 0 & & \cdots & & 0 \\ \hline 0 & 1 & & & & \\ & m_{r+1,r} & 1 & & 0 & \\ \vdots & m_{r+2,r} & 0 & 1 & & \\ & \vdots & & & \ddots & \\ 0 & m_{n,r} & 0 & \cdots & & 1 \end{array}\right] \tag{2.15}$$

τότε το r βήμα της απαλοιφής περιγράφεται από την εξίσωση των πινάκων

$$M^{(r)}A^{(r)}x = M^{(r)}b^{(n)}$$

ή

$$A^{(r+1)}x = b^{(r+1)}.$$

Παρατηρούμε λοιπόν ότι η όλη διαδικασία της τριγωνοποίησης μπορεί να περιγραφεί σαν τον πολλαπλασιασμό του αρχικού συστήματος επί τον πίνακα

$$M = M^{(n-1)} \ldots M^{(2)}M^{(1)}. \tag{2.16}$$

Άρα η

$$MA^{(1)}x = Mb^{(1)} \tag{2.17}$$

παράγει την

$$A^{(n)}x = b^{(n)}. \tag{2.18}$$

Θεώρημα 2.1.2. *Το σύστημα $A^{(n)}x = b^{(n)}$ είναι ισοδύναμο με το σύστημα $A^{(1)}x = b^{(1)}$.*

Απόδειξη. Η απόδειξη είναι όμοια με εκείνη του Θεωρήματος 2.1.1 και αφήνεται σαν άσκηση για τον αναγνώστη. ■

2.1.1 Επίλυση των $Ax_k = b_k, k = 1(1)\ell$

Ας υποθέσουμε ότι έχουμε τα ℓ συστήματα

$$Ax_k = b_k, \quad k = 1(1)\ell, \tag{2.19}$$

όπου

$$x_k = [x_{1k}, x_{2k}, \ldots, x_{nk}]^T \quad \text{και} \quad b_k = [b_{1k}, b_{2k}, \ldots, b_{nk}]^T$$

τα οποία μπορούν να γραφούν σαν

$$AX = B$$

όπου X και B δύο $n \times \ell$ πίνακες. Με την προϋπόθεση ότι $\det A \neq 0$, μπορούμε να εφαρμόσουμε τη μέθοδο απαλοιφής του Gauss με τη μόνη διαφορά ότι αντί οι πράξεις της απαλοιφής να εκτελούνται σε μια στήλη του b τώρα εκτελούνται συγχρόνως και στις ℓ στήλες του πίνακα B.

2.1.2 Υπολογισμός του A^{-1}

Στην περίπτωση αυτή έχουμε για επίλυση τα συστήματα

$$AX = I$$

τα οποία είναι μερική περίπτωση της προηγούμενης παραγράφου.

2.1.3 Υπολογισμός της $\det A$

Ο τριγωνικός πίνακας $A^{(n)}$ δίνεται από τη σχέση

$$A^{(n)} = MA^{(1)} \tag{2.20}$$

ή

$$\det A^{(n)} = [\det M] \left[\det A^{(1)}\right].$$

Αλλά λόγω της (2.16) έχουμε ότι

$$\det M = \prod_{r=1}^{n-1} \det M^{(r)}$$

και επειδή οι πίνακες $M^{(r)}, r = 1(1)n-1$ είναι μοναδιαίοι κάτω τριγωνικοί έπεται ότι

$$\det M = 1. \tag{2.21}$$

Επειδή όμως ο $A^{(n)}$ είναι ένας τριγωνικός πίνακας έχουμε

$$\det A^{(1)} = \det A^{(n)} = a_{11}^{(1)} a_{22}^{(2)} \dots a_{nn}^{(n)} \tag{2.22}$$

πράγμα που σημαίνει ότι η τιμή της ορίζουσας ενός πίνακα είναι το γινόμενο των οδηγών στοιχείων στη μέθοδο της απαλοιφής του Gauss.

2.1.4 Ο αλγόριθμος της απαλοιφής του Gauss

Ο παρακάτω αλγόριθμος περιγράφει τη μέθοδο της απαλοιφής του Gauss για τη λύση του $Ax = b$.

1. Διάβασε τα δεδομένα $A = (a_{ij}), b = (b_i)$
2. Για $i = 1(1)n$ εκτελείται $a_{i,n+1} = b_i$
3. Για $r = 1(1)n-1$ εκτελούνται τα βήματα α-γ
   (α) Έστω p ο μικρότερος ακέραιος για τον οποίο $a_{p,r} \neq 0, p = r(1)n$. Εάν δεν υπάρχει ο p τότε τύπωσε (δεν υπάρχει μοναδική λύση). Πήγαινε στο τέλος.
   (β) Εάν $p \neq r$ τότε (εναλλάσονται οι p και r γραμμές).
      i. Για $q = r(1)n$ εκτελούνται οι αντικαταστάσεις

$$\begin{aligned} b_q &= a_{rq} \\ a_{rq} &= a_{pq} \\ a_{pq} &= b_q \end{aligned}$$

   (γ) Για $i = r+1(1)n$ εκτελούνται τα βήματα i-ii
      i. Θέτουμε

$$m_{ir} = -\frac{a_{ir}}{a_{rr}}$$

      ii. Για $j = r+1(1)n+1$ να εκτελεσθεί

$$a_{ij} = a_{ij} + m_{ir} a_{rj}$$

4. Εάν $a_{nn} = 0$ τότε τύπωσε ''δεν υπάρχει μοναδική λύση''. Πήγαινε στο τέλος.

5. Θέτουμε (πίσω αντικατάσταση)

$$x_n = a_{n,n+1}/a_{nn}$$

6. Για $i = n-1(-1)1$ θέτουμε

$$x_i = \frac{\left[a_{i,n+1} - \sum_{j=i+1}^{n} a_{ij}x_j\right]}{a_{ii}}$$

7. Εκτύπωση της λύσης $x_i, i = 1(1)n$. Τέλος.

2.1.5 Τροποποίηση της μεθόδου απαλοιφής του Gauss

Από τον ανωτέρω αλγόριθμο γίνεται φανερό ότι αν κάποιο οδηγό στοιχείο είναι μηδέν, τότε η μέθοδος απαλοιφής του Gauss σταματά. Για την αποφυγή μιας τέτοιας περίπτωσης ο ανωτέρω αλγόριθμος εξετάζει τους συντελεστές της r-στήλης κάτω από τη κύρια διαγώνιο μέχρις ότου να βρεθεί ένας, ο οποίος είναι διάφορος από το μηδέν, έστω ο $a_{ir}^{(r)}$. Στη συνέχεια εναλλάσει τις i και r εξισώσεις και χρησιμοποιεί τη νέα εξίσωση σαν οδηγό. Η διαδικασία αυτή δεν αλλάζει το μαθηματικό πρόβλημα καθόσον η τυπική λύση ενός συστήματος γραμμικών εξισώσεων είναι ανεξάρτητη από τη διάταξη των εξισώσεων στο σύστημα. Ωστόσο όμως μπορούμε πάντοτε να βρίσκουμε ένα μη μηδενικό συντελεστή $a_{ir}^{(r)}$;

Θεώρημα 2.1.3. *Εάν ο $A^{(1)}$ είναι μη ιδιάζων, τότε υπάρχει ένας μη μηδενικός συντελεστής $a_{ir}^{(r)}$.*

Απόδειξη. Ακολουθώντας το σκεπτικό της απόδειξης του Θεωρήματος 2.1.1 μπορούμε να δείξουμε ότι $\det A^{(r)} = \det A^{(1)}$ άρα $\det A^{(r)} \neq 0$. Αλλά

$$\det A^{(1)} = \det \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{pmatrix} = a_{11}^{(1)} \det \begin{pmatrix} a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & & \vdots \\ a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{pmatrix}$$

$$= a_{11}^{(1)} \det A_{n-1}^{(2)}$$

και γενικά

$$\det A^{(r)} = \left[a_{11}^{(1)} a_{22}^{(2)} \dots a_{r-1,r-1}^{(r-1)}\right] \det A_{n-r+1}^{(r)},$$

όπου

$$A_{n-r+1}^{(r)} = \begin{pmatrix} a_{rr}^{(r)} & \cdots & a_{rn}^{(r)} \\ \vdots & & \vdots \\ a_{nr}^{(r)} & \cdots & a_{nn}^{(n)} \end{pmatrix}.$$

Συνεπώς $\det A_{n-r+1}^{(r)} \neq 0$ το οποίο σημαίνει ότι ο A_{n-r+1} δεν μπορεί να έχει την πρώτη στήλη του μηδέν. Επομένως μπορούμε πάντα να βρούμε ένα οδηγό στοιχείο $\neq 0$ στην πρώτη στήλη του $A_{n-r+1}^{(r)}$. ■

Σε περίπτωση λοιπόν που $a_{rr}^{(r)} = 0$, τότε αναζητείται στην r στήλη, και κάτω από τη διαγώνιο, $a_{ir}^{(r)} \neq 0$ (βλ. Θεώρημα 2.1.3). Στη συνέχεια εναλλάσονται οι r και i εξισώσεις. Στην τροποποιημένη μέθοδο της απαλοιφής χρησιμοποιούμε το συμβολισμό $\mathcal{A}^{(r)}$ και $\beta^{(r)}$ αντί των $A^{(r)}$ και $b^{(r)}$, αντίστοιχα. Η εναλλαγή των εξισώσεων αντιστοιχεί στον πολλαπλασιασμό του

$$\mathcal{A}^{(r)} x = \beta^{(r)} \tag{2.23}$$

με τον πίνακα ($r < i$)

$$I_{ri} = \left[\begin{array}{c|c|c} \begin{matrix} 1 & & \\ & \ddots & \\ & & 1 \end{matrix} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \begin{matrix} 1 & & \\ & \ddots & \\ & & 1 \end{matrix} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \begin{matrix} 1 & & \\ & \ddots & \\ & & 1 \end{matrix} \end{array}\right], \tag{2.24}$$

ο οποίος καλείται μεταθετικός πίνακας και για τον οποίο εύκολα διαπιστώνεται ότι

$$I_{ri}^2 = I. \tag{2.25}$$

Η δημιουργία του r βήματος γίνεται με τον πολλαπλασιασμό της (2.23) με I_{ri} και στη συνέχεια με τον πολλαπλασιασμό της προκύπτουσας εξίσωσης με $M^{(r)}$. Έχουμε λοιπόν

$$\left[M^{(r)}I_{ri}\right]A^{(r)}x = \left[M^{(r)}I_{ri}\right]\beta^{(r)} \tag{2.26}$$

το οποίο δίνει

$$\mathcal{A}^{(r+1)}x = \beta^{(r+1)}. \tag{2.27}$$

Στην περίπτωση όπου δεν χρειάζεται εναλλαγή των γραμμών τότε $i = r$ και

$$I_{ri} = I_{rr} = I.$$

Έτσι η όλη διαδικασία τριγωνοποίησης μπορεί τώρα να περιγραφεί τυπικά από τον πολλαπλασιασμό του

$$\mathcal{A}^{(1)}x = \beta^{(1)} \tag{2.28}$$

με τον πίνακα

$$\mathcal{M} = \left[M^{(n-1)}I_{n-1,i_{n-1}}\right]\left[M^{(n-2)}I_{n-2,i_{n-2}}\right]\ldots\left[M^{(2)}I_{2,i_2}\right]\left[M^{(1)}I_{1,i_1}\right] \tag{2.29}$$

δηλαδή το σύστημα

$$\mathcal{M}\mathcal{A}^{(1)}x = \mathcal{M}\beta^{(1)} \tag{2.30}$$

γίνεται το τριγωνικό σύστημα

$$\mathcal{A}^{(n)}x = \beta^{(n)}. \tag{2.31}$$

Στη συνέχεια αποδεικνύεται η ισοδυναμία του τριγωνικού συστήματος (2.31) και του αρχικού (2.28).

Θεώρημα 2.1.4. *Το σύστημα $\mathcal{A}^{(1)}x = \beta^{(1)}$ είναι ισοδύναμο με το σύστημα $\mathcal{A}^{(n)}x = \beta^{(n)}$.*

Απόδειξη. Έχουμε υποθέσει ότι $\det\mathcal{A}^{(1)} \neq 0$. Επίσης από τις (2.30) και (2.31) παρατηρούμε ότι

$$\mathcal{A}^{(n)} = \mathcal{M}\mathcal{A}^{(1)} \tag{2.32}$$

επίσης

$$\det\mathcal{A}^{(n)} = \left[\det\mathcal{M}\right]\left[\det\mathcal{A}^{(1)}\right]. \tag{2.33}$$

Συνεπώς $\det \mathcal{A}^{(n)} \neq 0$ αν και μόνο αν $\det \mathcal{M} \neq 0$. Αλλά από την (2.29) έχουμε ότι

$$\det \mathcal{M} = \left[\prod_{r=1}^{n-1} \det M^{(r)}\right] \left[\prod_{r=1}^{n-1} \det I_{r,i_r}\right] \tag{2.34}$$

όπου υπενθυμίζεται ότι επειδή οι πίνακες $M^{(r)}, \quad r = 1(1)n-1$ είναι μοναδιαίοι κάτω τριγωνικοί έχουμε

$$\prod_{r=1}^{n-1} \det M^{(r)} = 1. \tag{2.35}$$

Αυτό που απομένει λοιπόν να δειχθεί είναι ότι

$$\det \prod_{r=1}^{n-1} I_{r,i_r} \neq 0 \tag{2.36}$$

πράγμα που ισχύει αφού για $r = 1(1)n-1$ και για $r \neq i_r$

$$\det I_{r,i_r} = -1$$

γιατί ο πίνακας $\det I_{r,i_r}$ προέρχεται από τον I μετά από την εναλλαγή των r και i γραμμών (αν $r = i_r$, τότε $\det I_{rr} = 1$). Έτσι από τις (2.34), (2.35) και (2.36) συνάγεται ότι

$$\det A^{(n)} \neq 0$$

το οποίο σημαίνει ότι η (2.31) έχει μια μοναδική λύση. Η λύση του (2.28) όμως είναι η

$$x = \left[A^{(1)}\right]^{-1} \beta^{(1)}$$

η οποία ικανοποιεί την (2.31) ή την (2.30), γιατί

$$\mathcal{A}^{(n)} x = \mathcal{M}\mathcal{A}^{(1)} \left\{\left[\mathcal{A}^{(1)}\right]^{-1} \beta^{(1)}\right\} = \mathcal{M}\beta^{(1)} = \beta^{(n)}$$

πράγμα που σημαίνει ότι τα συστήματα (2.28) και (2.31) έχουν την ίδια λύση και συνεπώς είναι ισοδύναμα. ■

2.1.6 Αριθμητική αστάθεια

Όταν εφαρμόζεται στην πράξη ο αλγόριθμος της απαλοιφής του Gauss θα πρέπει να εξετασθεί η επιρροή των σφαλμάτων στρογγύλευσης στους υπολογισμούς. Όπως αναφέρεται και στην εισαγωγή, η αριθμητική των υπολογιστών δεν είναι "ακριβής", καθόσον οι αριθμοί που μπορούν να παρασταθούν στη μνήμη δεν αποτελούν σώμα, λόγω των σφαλμάτων στρογγύλευσης. Εάν μπορούσαμε να παραστήσουμε όλους τους πραγματικούς αριθμούς με την ακριβή τιμή τους στη μνήμη και αν μπορούσαμε να εκτελέσουμε αριθμητική με άπειρα ψηφία, τότε το πρόβλημα της αριθμητικής επίλυσης ενός γραμμικού συστήματος θα μπορούσε να γίνει με την τροποποιημένη μέθοδο της απαλοιφής του Gauss δίχως επιπλέον προβληματισμούς. Ωστόσο όμως θα πρέπει να λάβουμε υπόψη τα αποτελέσματα των σφαλμάτων στρογγύλευσης.

Κατά τη μαθηματική διατύπωση της μεθόδου της απαλοιφής μας ενδιέφερε μόνον αν το υποψήφιο οδηγό στοιχείο μηδενίζεται σε κάθε βήμα. Στην πράξη αυτό σημαίνει ότι θα πρέπει να εξεταστεί αν το υποψήφιο οδηγό στοιχείο $a_{rr}^{(r)}$ είναι 'μικρό' συγκρινόμενο με μερικά από τα στοιχεία $a_{rr}^{(r)} (r < i)$. Εάν το $a_{rr}^{(r)}$ είναι μικρό, τότε παρατηρούμε ότι οι αντίστοιχοι πολλαπλασιαστές m_{ir} θα είναι αρκετά μεγάλοι και οι εξισώσεις που θα παράγονται από το βήμα της απαλοιφής θα έχουν συντελεστές, οι οποίοι θα είναι πολύ μεγαλύτεροι από εκείνους του προηγούμενου βήματος. Ο Wilkinson παρατηρεί ότι στη περίπτωση αυτή είναι πιθανό να εμφανισθεί το φαινόμενο της αριθμητικής αστάθειας. Για την αποφυγή της εμφάνισης αυτού του φαινομένου προτείνει την εναλλαγή των γραμμών όταν τα υποψήφια οδηγά στοιχεία είναι μικρότερα από όλους τους συντελεστές στην ίδια στήλη. Πιο συγκεκριμένα επιλέγεται σαν οδηγό στοιχείο εκείνο που έχει τη μεγαλύτερη απόλυτη τιμή στην r στήλη του $A^{(r)}$ και βρίσκεται επί ή κάτω από την r γραμμή, οπότε θα ισχύει $|m_{ir}| \leq 1$, δηλαδή:

$$a_{pr}^{(r)} = \max_i \left| a_{ir}^{(r)} \right|, \quad r \leq i \leq n$$

οπότε και εναλλάσονται οι r και p γραμμές του $A^{(r)}$ πριν από το r-βήμα. Η τεχνική αυτή καλείται **μερική οδήγηση** σε αντιδιαστολή με την **ολική οδήγηση** όπου σε κάθε r βήμα επιλέγεται σαν οδηγό στοιχείο εκείνο που έχει τη μεγαλύτερη απόλυτο τιμή από τις τελευταίες $n - r + 1$ γραμμές και στήλες του $A^{(r)}$ δηλαδή

$$a_{pr}^{(r)} = \max_{ij} \left| a_{ij}^{(r)} \right|, \quad r \leq i, j \leq n$$

οπότε εναλλάσονται οι p και r γραμμές και στήλες. Τέλος, στην περίπτωση όπου ο πίνακας είναι μεγάλης τάξης και αραιός αρκεί το οδηγό στοιχείο να είναι μεγαλύτερο από κάποιο $\varepsilon = 10^{-3}$. Από τα ανωτέρω συμπεραίνουμε ότι ενώ η μερική οδήγηση απαιτεί εναλλαγές γραμμών, η ολική οδήγηση είναι περισσότερο πολύπλοκη αφού απαιτεί την εναλλαγή και των στηλών.

Παράδειγμα

Δίνεται το γραμμικό σύστημα

$$\begin{array}{rcrcrcl} - & x_1 & + & 2x_2 & - & x_3 & = & 0 \\ & 2x_1 & - & x_2 & & & = & 1 \\ & x_1 & + & 7x_2 & - & 3x_3 & = & 5 \end{array}$$

Να υπολογιστεί η λύση του ανωτέρω συστήματος με τη μέθοδο απαλοιφής του Gauss i) χωρίς οδήγηση και ii) με μερική οδήγηση.

Λύση

(i) Κατασκευάζουμε τον επαυξημένο πίνακα $[A{:}b]$, ο οποίος στην προκειμένη περίπτωση είναι ο

$$\left[\begin{array}{rrr|r} -1 & 2 & -1 & 0 \\ 2 & -1 & 0 & 1 \\ 1 & 7 & -3 & 5 \end{array}\right]$$

Στη συνέχεια εφαρμόζουμε τη μέθοδο απαλοιφής του Gauss χωρίς οδήγηση παρουσιάζοντας τους πολλαπλασιαστές από τα αριστερά για κάθε γραμμή. Έτσι έχουμε διαδοχικά τους ακόλουθους υπολογισμούς

$$\begin{array}{r} \\ 2 \\ 1 \end{array} \left[\begin{array}{rrr|r} -1 & 2 & -1 & 0 \\ 2 & -1 & 0 & 1 \\ 1 & 7 & -3 & 5 \end{array}\right]$$

$$\begin{array}{r} \\ -3 \end{array} \left[\begin{array}{rrr|r} -1 & 2 & -1 & 0 \\ 0 & 3 & -2 & 1 \\ 0 & 9 & -4 & 5 \end{array}\right] \text{1ο βήμα}$$

$$\left[\begin{array}{rrr|r} -1 & 2 & -1 & 0 \\ 0 & 3 & -2 & 1 \\ 0 & 0 & 2 & 2 \end{array}\right] \text{2ο βήμα}$$

Λύση του τριγωνικού συστήματος

$$\begin{array}{rcrcrcl} - & x_1 & + & 2x_2 & - & x_3 & = & 0 \\ & & & 3x_2 & - & 2x_3 & = & 1 \\ & & & & & 2x_3 & = & 2 \end{array}$$

Άρα

$$x_3 = 1, \quad x_2 = 1 \quad \text{και} \quad x_1 = 1.$$

(ii) Για την εφαρμογή της μεθόδου απαλοιφής του Gauss με μερική οδήγηση εργαζόμαστε με ανάλογο τρόπο ανταλλάσοντας όπου απαιτείται τις γραμμές του επαυξημένου πίνακα. Έτσι έχουμε

$$\left[\begin{array}{rrr|r} -1 & 2 & -1 & 0 \\ 2 & -1 & 0 & 1 \\ 1 & 7 & -3 & 5 \end{array}\right] \begin{array}{l} (1) \\ (2) \\ (3) \end{array} \qquad \begin{array}{l} \text{Οι αριθμοί των} \\ \text{παρενθέσεων δηλώνουν} \\ \text{τη διάταξη των γραμμών} \end{array}$$

$$\begin{array}{r} \\ 1/2 \\ -1/2 \end{array} \left[\begin{array}{rrr|r} 2 & -1 & 0 & 1 \\ -1 & 2 & -1 & 0 \\ 1 & 7 & -3 & 5 \end{array}\right] \begin{array}{l} (2) \\ (1) \\ (3) \end{array} \qquad \begin{array}{l} \text{Ανταλλαγή των} \\ \text{δύο πρώτων γραμμών} \end{array}$$

$$\left[\begin{array}{rrr|r} 2 & -1 & 0 & 1 \\ 0 & 3/2 & -1 & 1/2 \\ 0 & 15/2 & -3 & 9/2 \end{array}\right] \begin{array}{l} (2) \\ (1) \\ (3) \end{array} \qquad \text{1ο βήμα}$$

$$\begin{array}{r} \\ -1/5 \end{array} \left[\begin{array}{rrr|r} 2 & -1 & 0 & 1 \\ 0 & 15/2 & -3 & 9/2 \\ 0 & 3/2 & -1 & 1/2 \end{array}\right] \begin{array}{l} (2) \\ (3) \\ (1) \end{array} \qquad \begin{array}{l} \text{Ανταλλαγή των} \\ \text{δύο τελευταίων γραμμών} \end{array}$$

$$\left[\begin{array}{rrr|r} 2 & -1 & 0 & 1 \\ 0 & 15/2 & -3 & 9/2 \\ 0 & 0 & -2/5 & -2/5 \end{array}\right] \begin{array}{l} (2) \\ (3) \\ (1) \end{array} \qquad \text{2ο βήμα}$$

Η λύση του τριγωνικού συστήματος

$$\begin{array}{rcrcrcl} 2x_1 & - & x_2 & & & = & 1 \\ & & \frac{15}{2}x_2 & - & 3x_3 & = & \frac{9}{2} \\ & & & - & \frac{2}{5}x_3 & = - & \frac{2}{5} \end{array}$$

είναι η

$$x_3 = 1, \quad x_2 = 1 \quad \text{και} \quad x_1 = 1.$$

2.1.7 Ο αλγόριθμος απαλοιφής του Gauss με μερική οδήγηση

Ο αλγόριθμος της απαλοιφής του Gauss με μερική οδήγηση είναι ο ακόλουθος

1. Διάβασε τα δεδομένα $A = (a_{ij})$, $b = (b_i)$ και n.

2. για $i = 1(1)n$ εκτελείται $a_{i,n+1} = b_i$.

3. Για $i = 1(1)n$ να τεθεί

$$\mathrm{h}(i) = i$$

   (Τοποθέτηση αρχικών τιμών στο δείκτη της γραμμής).

4. για $r = 1(1)n-1$ να εκτελεστούν τα βήματα 4(α′)-4(δ′) (διαδικασία απαλοιφής).

   (α′) Έστω p ο μικρότερος ακέραιος με

$$r \leq p \leq n$$

   και

$$|a(\mathrm{h}(p),r)| = \max_{r \leq j \leq n} |a(\mathrm{h}(j),r)|$$

   (β′) Εάν $a(\mathrm{h}(p), r) = 0$ τότε τύπωσε (δεν υπάρχει μοναδική λύση). Τέλος.

   (γ′) Εάν $\mathrm{h}(r) \neq \mathrm{h}(p)$ τότε

$$\begin{aligned} \mathrm{q} &= \mathrm{h}(r) \\ \mathrm{h}(r) &= \mathrm{h}(p) \\ \mathrm{h}(p) &= \mathrm{q} \end{aligned}$$

   (προσομοίωση της εναλλαγής των γραμμών).

   (δ′) για $i = r + 1(1)n$ να εκτελεστούν τα βήματα i και ii

      i. Να τεθεί

$$m(\mathrm{h}(i),r) = -\frac{a(\mathrm{h}(i),r)}{a(\mathrm{h}(r),r)}$$

ii. Για $j = r + 1(1)n + 1$ να εκτελεσθεί

$$a(\mathrm{h}(i),j) = a(\mathrm{h}(i),j) + m(\mathrm{h}(i),r)a(\mathrm{h}(r),j)$$

Εάν $a(\mathrm{h}(n),n) = 0$ τότε τύπωσε (δεν υπάρχει μοναδική λύση). Πήγαινε στο τέλος.

5. Να τεθεί (πίσω αντικατάσταση)

$$x_n = a(\mathrm{h}(n),n+1)/a(\mathrm{h}(n),n)$$

6. Για $i = n - 1(-1)1$ να υπολογιστούν οι

$$x_i = \frac{a(\mathrm{h}(i),n+1) - \sum_{j=i+1}^{n} a(\mathrm{h}(i),j)x_j}{a(\mathrm{h}(i),i)}$$

7. Εκτύπωση της λύσης $x_i, i = 1(1)n$. Τέλος.

Ενώ για αρκετά γραμμικά συστήματα η μερική οδήγηση παράγει ικανοποιητικά αποτελέσματα, υπάρχουν περιπτώσεις όπου η τεχνική αυτή δεν είναι αρκετή.

Παράδειγμα

Έστω το γραμμικό σύστημα

$$\begin{array}{rcrcl} 30,00x_1 & + & 591.400x_2 & = & 591.700 \\ 5,291 & - & 6,130x_2 & = & 46,78. \end{array}$$

Εάν εφαρμοστεί ο προηγούμενος αλγόριθμος με αριθμητική τεσσάρων ψηφίων θα έχουμε

$$m_{21} = \frac{5,291}{30,00} = 0,1764$$

που οδηγεί στο σύστημα

$$\begin{array}{rcrcrl} 30,00x_1 & + & 591.400x_2 & = & & 591.700 \\ & - & 104.300x_2 & = & - & 104.400 \end{array}$$

το οποίο έχει τις λύσεις $x_2 = 1,001$ και $x_1 = -10,00$.

Ωστόσο οι ακριβείς λύσεις του αρχικού συστήματος είναι οι $x_1 = 10,00$ και $x_2 = 1,000$. Μία διαδικασία μερικής οδήγησης, η οποία θα μπορούσε να αντεπεξέλθει τη δυσκολία αυτή όπως και πολλές

άλλες για τις οποίες η προηγούμενη μέθοδος παρουσιάζει κάποιο πρόβλημα είναι η λεγόμενη κλιμακούμενη scaled μερική οδήγηση. Στην τεχνική αυτή διαιρούνται οι γραμμές από την οδηγό μέχρι την τελευταία με τον εκάστοτε μεγαλύτερο κατά απόλυτο τιμή συντελεστή της κάθε γραμμής. Στη συνέχεια εφαρμόζεται η μερική οδήγηση. Για το παράδειγμα αυτό έχουμε

$$\frac{30,00}{591.400} = 0,00005073 \quad \text{και} \quad \frac{5,291}{6,130} = 0,8631$$

οπότε εναλλάσονται οι δύο γραμμές και το αποτέλεσμα της απαλοιφής δίνει τις ακριβείς λύσεις. Οι δε αλλαγές που διαφέρουν από την προηγούμενη μέθοδο είναι στα βήματα 3-4(α′), τα οποία θα πρέπει να αντικατασταθούν με τα:

3. Για $i = 1(1)n$ να εκτελεσθούν τα βήματα 3(α′)-3(β′)
   - (α′) $s_i = \max_{1 \le j \le n} |a_{ij}|$
   - (β′) Εάν $s_i = 0$ τότε τύπωσε 'δεν υπάρχει μοναδική λύση'
   - (γ′) $\mathrm{h}(i) = i$

4. Για $r = 1(1)n-1$ να εκτελεστούν τα βήματα 4.(α′)-4.(β′)
   Έστω p ο μικρότερος ακέραιος με $r \le p \le n$ και

$$\frac{|a(\mathrm{h}(p),r)|}{s(\mathrm{h}(p))} = \max_{r \le j \le n} \frac{|a(\mathrm{h}(j),r)|}{s(\mathrm{h}(j))}.$$

Η ανωτέρω τεχνική μπορεί επίσης να πραγματοποιηθεί αν πολλαπλασιάσουμε την εξίσωση $Ax = b$ με το διαγώνιο πίνακα D^{-1} του οποίου το i-οστό διαγώνιο στοιχείο είναι το $(s_i)^{-1}$.

2.2 Η μέθοδος απαλοιφής του Jordan

Με την χρησιμοποίηση της μεθόδου της απαλοιφής του Jordan είναι δυνατόν να μετασχηματιστεί ο πίνακας των συντελεστών των αγνώστων σε ένα διαγώνιο πίνακα. Η μορφή αυτή του πίνακα είναι απλούστερη από εκείνη της απαλοιφής του Gauss καθόσον ο υπολογισμός του αγνώστου x_i μπορεί να προκύψει με μία απλή διαίρεση.

Θεωρώντας πάλι το σύστημα (2.1), το πρώτο βήμα της απαλοιφής του Jordan είναι ακριβώς ίδιο με εκείνο της μεθόδου απαλοιφής του Gauss, έτσι έχουμε το σύστημα (2.4) δηλαδή το:

$$\begin{array}{rcrcrcrcl} a_{11}^{(1)}x_1 & + & a_{12}^{(1)}x_2 & + & a_{13}^{(1)}x_3 & + & \ldots & + & a_{1n}^{(1)}x_n & = & b_1^{(1)} \\ & & a_{22}^{(2)}x_2 & + & a_{23}^{(2)}x_3 & + & \ldots & + & a_{2n}^{(2)}x_n & = & b_2^{(2)} \\ & & a_{32}^{(2)}x_2 & + & a_{33}^{(2)}x_3 & + & \ldots & + & a_{3n}^{(2)}x_n & = & b_3^{(2)} \\ & & \vdots & & & & & & & & \\ & & a_{n2}^{(2)}x_2 & + & a_{n3}^{(2)}x_3 & + & \ldots & + & a_{nn}^{(2)}x_n & = & b_n^{(2)} \end{array} \tag{2.37}$$

όπου παρατηρούμε ότι ο x_1 έχει απαλειφθεί από τις $n-1$ τελευταίες εξισώσεις οπότε έχουμε πάλι

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

Ωστόσο στο δεύτερο βήμα της μεθόδου της απαλοιφής του Jordan απαλείφεται ο x_2 όχι μόνο από τις $n-2$ τελευταίες εξισώσεις, αλλά συγχρόνως και από την πρώτη. Έτσι λαμβάνουμε το σύστημα

$$\begin{array}{rcrcrcrcl} a_{11}^{(1)}x_1 & & & & a_{13}^{(3)}x_3 & + & \ldots & + & a_{1n}^{(3)}x_n & = & b_1^{(3)} \\ & & a_{22}^{(2)}x_2 & + & a_{23}^{(3)}x_3 & + & \ldots & + & a_{2n}^{(3)}x_n & = & b_2^{(3)} \\ & & & & a_{33}^{(3)}x_3 & + & \ldots & + & a_{3n}^{(3)}x_n & = & b_3^{(3)} \\ & & & & \vdots & & & & & & \\ & & & & a_{n3}^{(3)}x_3 & + & \ldots & + & a_{nn}^{(3)}x_n & = & b_n^{(3)} \end{array} \tag{2.38}$$

ή

$$A^{(3)}x = b^{(3)} \tag{2.39}$$

αν υποθέσουμε ότι αλλάζουμε τους επάνω δείκτες των συντελεστών της δεύτερης γραμμής εκτός του πρώτου. Μετά από $r-1$ τέτοια βήματα θα έχουμε το σύστημα

$$\begin{array}{rcrcrcrcl} a_{11}^{(1)}x_1 & & & & & & +a_{1r}^{(r)}x_r + \ldots + a_{1n}^{(r)}x_n & = & b_1^{(r)} \\ & a_{22}^{(2)}x_2 & & & & & +a_{2r}^{(r)}x_r + \ldots + a_{2n}^{(r)}x_n & = & b_2^{(r)} \\ & & \ddots & & & & \vdots & & \\ & & & a_{r-1,r-1}^{(r-1)}x_{r-1} & & & +a_{r-1,r}^{(r)}x_r \ldots + a_{r-1,n}^{(r)}x_n & = & b_{r-1}^{(r)} \\ & & & & & & a_{rr}^{(r)}x_r + \ldots + a_{rn}^{(r)}x_n & = & b_r^{(r)} \\ & & & & & & \vdots & & \\ & & & & & & a_{nr}^{(r)}x_r + \ldots + a_{nn}^{(r)}x_n & = & b_n^{(r)} \end{array} \tag{2.40}$$

ή

$$A^{(r)}x = b^{(r)}$$

όπου πάλι για λόγους ομοιομορφίας αλλάξαμε τους επάνω δείκτες των συντελεστών της $r-1$ γραμμής εκτός του πρώτου. Τέλος, μετά από n τέτοια βήματα θα έχουμε το διαγώνιο σύστημα:

$$\begin{array}{rcl} a_{11}^{(1)}x_1 & = & b_1^{(n+1)} \\ a_{22}^{(2)}x_2 & = & b_2^{(n+1)} \\ \ddots & \vdots & \\ a_{nn}^{(n)}x_n & = & b_n^{(n+1)} \end{array} \tag{2.41}$$

το οποίο μπορεί να γραφεί σαν

$$A^{(n)}x = b^{(n+1)} \tag{2.42}$$

όπου ο $A^{(n)}$ τώρα είναι ένας διαγώνιος πίνακας. Η λύση του (2.42) είναι η

$$x_i = \frac{1}{a_{ii}^{(i)}} b_i^{(n+1)}$$

εφόσον $a_{ii}^{(i)} \neq 0, i = 1(1)n$. Αναλυτικότερα χρησιμοποιούμε πάλι τους συμβολισμούς

$$\begin{array}{ll} a_{ij}^{(1)} = a_{ij} & i,j = 1(1)n \\ b_i^{(1)} = b_i & i = 1(1)n \end{array}$$

και αν $a_{11}^{(1)} \neq 0$, τότε ορίζουμε τους πολλαπλασιαστές

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

Τώρα προκειμένου να απαλειφθεί ο x_1 από την i-οστή εξίσωση, προσθέτουμε m_{i1} φορές την πρώτη εξίσωση στην i-οστή, οπότε λαμβάνουμε το σύστημα (2.37), όπου

$$\begin{array}{ll} a_{ij}^{(2)} = a_{ij}^{(1)} + m_{ij}a_{1j}^{(1)}, & i = 1(1)n, i \neq 1 \\ & j = 1(1)n \end{array}$$

και

$$b_i^{(2)} = b_i^{(1)} + m_{i1}b_1^{(1)}, \qquad i = 1(1)n, i \neq 1.$$

Στο σημείο αυτό παρατηρούμε ότι πριν ακολουθήσει το δεύτερο βήμα θα πρέπει για λόγους ομοιομορφίας να κάνουμε τις αντικαταστάσεις

$$\begin{array}{l} a_{1j}^{(2)} = a_{1j}^{(1)}, \quad j = 2(1)n \\ b_1^{(2)} = b_1^{(1)}. \end{array}$$

Στο τελευταίο βήμα τώρα απαλείφεται ο άγνωστος x_2 τόσο από τις τελευταίες $n-2$ εξισώσεις όσο και από την πρώτη, οπότε λαμβάνουμε το σύστημα (2.38) όπου

$$a_{ij}^{(3)} = a_{ij}^{(2)} + m_{i2}a_{1j}^{(2)}, \quad i = 1(1)n, i \neq 2, j = 2(1)n$$

και

$$b_i^{(3)} = b_i^{(2)} + m_{i2}b_2^{(2)}, \quad i = 1(1)n, i \neq 2$$

όπου για λόγους ομοιομορφίας θέτουμε

$$\begin{array}{l} a_{2j}^{(3)} = a_{2j}^{(2)}, \quad j = 3(1)n \\ b_2^{(3)} = b_2^{(2)}. \end{array}$$

Συνεχίζοντας καταλήγουμε στο διαγώνιο σύστημα (2.41). Ο μετασχηματισμός του A σε ένα διαγώνιο πίνακα της ίδιας τάξης αποτελείται από n κύρια βήματα τα οποία αντιστοιχούν σε n πράξεις των γραμμών του πίνακα. Υπό μορφή πινάκων η μέθοδος του Jordan αναζητεί ένα μη ιδιάζων $n \times n$ πίνακα M για τον οποίο να ισχύει

$$MAx = Mb \tag{2.43}$$

με

$$MA = I. \tag{2.44}$$

Συνεπώς

$$M = A^{-1} \tag{2.45}$$

και

$$x = Mb. \tag{2.46}$$

Ας υποθέσουμε ότι το αρχικό σύστημα είναι το

$$A^{(1)}x = b^{(1)} \tag{2.47}$$

τότε μπορούμε να παρατηρήσουμε εύκολα ότι αν $M^{(1)}$ είναι ο πίνακας

$$M^{(1)} = \left[\begin{array}{c|c} \mu_{11} & 0 \\ \hline \mu_{21} & \\ \mu_{31} & I_{n-1} \\ \vdots & \\ \mu_{n1} & \end{array}\right] \tag{2.48}$$

όπου

$$\mu_{i1} = \begin{cases} 1/a_{11}^{(1)}, & i = 1 \\ -a_{i1}^{(1)}/a_{11}^{(1)}, & i \neq 1 \end{cases} \quad \text{με } a_{11}^{(1)} \neq 0, \tag{2.49}$$

τότε το πρώτο βήμα της απαλοιφής του Jordan μπορεί να γίνει με τον πολλαπλασιασμό του (2.47) με τον $M^{(1)}$, δηλαδή:

$$M^{(1)}A^{(1)}x = M^{(1)}b \tag{2.50}$$

ή

$$A^{(2)}x = b^{(2)}. \tag{2.51}$$

Έτσι το r βήμα της απαλοιφής θα περιγράφεται από την εξίσωση:

$$M^{(r)}A^{(r)}x = M^{(r)}b \tag{2.52}$$

όπου

$$M^{(r)} = \left[\begin{array}{ccc|c|ccc} & & & \mu_{1r} & & & \\ & I_{r-1} & & \vdots & & 0 & \\ & & & \mu_{r-1,r} & & & \\ \hline 0 & \cdots & 0 & \mu_{rr} & 0 & \cdots & 0 \\ \hline & & & \mu_{r+1,r} & & & \\ & 0 & & \vdots & & I_{n-r} & \\ & & & \mu_{nr} & & & \end{array}\right] \tag{2.53}$$

και τα μ_{ir} δίνονται από τον τύπο

$$\mu_{i1} = \begin{cases} 1/a_{rr}^{(r)}, & i = r \\ -a_{ir}^{(r)}/a_{rr}^{(r)}, & i \neq r, \quad a_{rr}^{(r)} \neq 0, \quad r = 1(1)n. \end{cases}$$

Από την (2.52) προκύπτει το σύστημα

$$A^{(r+1)}x = b^{(r+1)} \tag{2.54}$$

με

$$A^{(r+1)} = \left[\begin{array}{c|c} I_r & * \\ \hline 0 & * \end{array}\right] \tag{2.55}$$

όπου $*$ συμβολίζει την ύπαρξη στοιχείων. Έτσι η όλη διαδικασία μπορεί να περιγραφεί από τον πολλαπλασιασμό του αρχικού συστήματος με τον πίνακα

$$M = M^{(n)} M^{(n-1)} \dots M^{(2)} M^{(1)} \tag{2.56}$$

όπου $M^{(1)}$ και $M^{(i)}, \quad i = 2(1)n-1$ δίνονται από τις (2.48) και (2.53), αντίστοιχα ενώ

$$M^{(n)} = \left[\begin{array}{ccc|c} & & & \mu_{1,n} \\ & I_{n-1} & & \vdots \\ & & & \mu_{n-1,n} \\ \hline 0 & \cdots & 0 & \mu_{nn} \end{array}\right]. \tag{2.57}$$

Άρα η

$$MA^{(1)}x = Mb^{(1)}$$

δίνει την

$$x = Mb^{(1)}. \tag{2.58}$$

Όπως με τη μέθοδο απαλοιφής του Gauss, μπορεί να αναπτυχθεί η μέθοδος της απαλοιφής του Jordan με μερική ή ολική οδήγηση. (Η ανάπτυξη της σχετικής θεωρίας αφήνεται σαν άσκηση). Στην περίπτωση αυτή η επιλογή του οδηγού στοιχείου γίνεται όπως ακριβώς στη μέθοδο της απαλοιφής του Gauss και δεν επεκτείνεται στην αναζήτηση όλης της στήλης όπως ίσως θα περίμενε κανείς διότι τότε θα καταστρεφόταν η προηγούμενη διαγωνοποίηση του πίνακα (γιατί;). Η μέθοδος της απαλοιφής του Jordan χρησιμοποιείται συνήθως για τον υπολογισμό του αντιστρόφου ενός πίνακα ή για την επίλυση συστημάτων με τον ίδιο πίνακα συντελεστών των αγνώστων.

Παράδειγμα

Να εφαρμοστεί η μέθοδος Jordan με μερική οδήγηση στο σύστημα του προηγούμενου παραδείγματος.

Λύση

$$\left[\begin{array}{ccc|c} -1 & 2 & -1 & 0 \\ 2 & -1 & 0 & 1 \\ 1 & 7 & -3 & 5 \end{array}\right] \begin{array}{l} (1) \\ (2) \\ (3) \end{array}$$

$$\begin{array}{r} \\ 1/2 \\ -1/2 \end{array}\left[\begin{array}{ccc|c} 2 & -1 & 0 & 1 \\ -1 & 2 & -1 & 0 \\ 1 & 7 & -3 & 5 \end{array}\right]\begin{array}{l} (2) \\ (1) \\ (3) \end{array}\quad \begin{array}{c} \text{Ανταλλαγή των δύο} \\ \text{πρώτων γραμμών} \end{array}$$

$$\left[\begin{array}{ccc|c} 2 & -1 & 0 & 1 \\ 0 & 3/2 & -1 & 1/2 \\ 0 & 15/2 & -3 & 9/2 \end{array}\right]\begin{array}{l} (2) \\ (1) \\ (3) \end{array}\quad 1\text{ο βήμα}$$

$$\begin{array}{r} \\ 2/15 \\ -1/5 \end{array}\left[\begin{array}{ccc|c} 2 & -1 & 0 & 1 \\ 0 & 15/2 & -3 & 9/2 \\ 0 & 3/2 & -1 & 1/2 \end{array}\right]\begin{array}{l} (2) \\ (3) \\ (1) \end{array}\quad \begin{array}{c} \text{Ανταλλαγή των δύο} \\ \text{τελευταίων γραμμών} \end{array}$$

$$\left[\begin{array}{ccc|c} 2 & 0 & -2/5 & 8/5 \\ 0 & 15/2 & -3 & 9/2 \\ 0 & 0 & -2/5 & -2/5 \end{array}\right]\begin{array}{l} (2) \\ (3) \\ (1) \end{array}\quad 2\text{ο βήμα}$$

$$\left[\begin{array}{ccc|c} 2 & 0 & 0 & 2 \\ 0 & 15/2 & 0 & 15/2 \\ 0 & 0 & -2/5 & -2/5 \end{array}\right]\begin{array}{l} (2) \\ (3) \\ (1) \end{array}\quad 3\text{ο βήμα}$$

Άρα η λύση είναι η $x_1 = 1, x_2 = 1$ και $x_3 = 1$.

2.2.1 Ο αλγόριθμος απαλοιφής του Jordan με μερική οδήγηση

Ο αλγόριθμος της απαλοιφής του Jordan με μερική οδήγηση για τη λύση του $Ax = b$ είναι ο ακόλουθος:

1. Διάβασε τα δεδομένα $A = (a_{ij}), b = (b_i)$ και n.
2. Για $i = 1(1)n$ να εκτελεσθεί
$$a_{i,n+1} = b_i$$
3. Για $i = 1(1)n$ να τεθεί
$$\mathrm{h}(i) = i$$
4. Για $r = 1(1)n$ να εκτελεσθούν τα βήματα 4(α′)-4(δ′) (διαδικασία απαλοιφής).

(α′) Έστω p ο μικρότερος ακέραιος

$$r \leq p \leq n$$

και

$$|a(\mathrm{h}(p),r))| = \max_{r \leq j \leq n} |a(\mathrm{h}(j),r)|$$

(β′) Εάν $a(\mathrm{h}(p),r)=0$ τότε τύπωσε 'δεν υπάρχει μοναδική λύση'. Τέλος.

(γ′) Εάν $\mathrm{h}(r) \neq \mathrm{h}(p)$ τότε

$$\begin{gathered} \mathrm{q}=\mathrm{h}(r) \\ \mathrm{h}(r)=\mathrm{h}(p) \\ \mathrm{h}(p)=\mathrm{q} \end{gathered}$$

(προσομοίωση της εναλλαγής των γραμμών)

(δ′) Για $i = 1(1)n$ και $i \neq r$ να εκτελεσθούν τα βήματα i και ii

i. Να τεθεί

$$m(\mathrm{h}(i),r) = -\frac{a(\mathrm{h}(i),r)}{a(\mathrm{h}(r),r)}$$

ii. για $j = r+1(1)n+1$ να εκτελεσθεί

$$a(\mathrm{h}(i),j) = a(\mathrm{h}(i),j) + m(\mathrm{h}(i),r)a(\mathrm{h}(r),j)$$

5. Για $i = 1(1)n$ να εκτελεσθεί

Εάν $a(\mathrm{h}(i),i) = 0$ τότε τύπωσε 'όχι μοναδική λύση'. Τέλος.

$$x_i = a(\mathrm{h}(i),n+1)/a(\mathrm{h}(i),i)$$

6. Τύπωσε την λύση $x_i, i = 1(1)n$. Τέλος

2.2.2 Υπολογιστική πολυπλοκότητα

Στη συνέχεια θα συγκρίνουμε το πλήθος και το είδος των πράξεων όπως επίσης και τις απαιτήσεις, όσον αφορά τη μνήμη, για τις μεθόδους απαλοιφής του Gauss και Jordan. Ας υποθέσουμε ότι βρισκόμαστε στο k βήμα της απαλοιφής του Gauss για τη λύση ℓ συστημάτων δηλαδή:

$$
n-k\left\{\left[\begin{array}{ccccc|ccc}
a_{11} & a_{12} & a_{13} & \cdots & a_{1k} & a_{1,k+1} & \cdots & a_{1n} \\
 & \hat{a}_{22} & \hat{a}_{23} & \cdots & \hat{a}_{2k} & \hat{a}_{2,k+1} & \cdots & \hat{a}_{2n} \\
 & & \hat{a}_{33} & \cdots & \hat{a}_{3k} & \hat{a}_{3,k+1} & \cdots & \hat{a}_{3n} \\
 & \mathbf{0} & & \ddots & \vdots & \vdots & & \vdots \\
 & & & & \hat{a}_{kk} & \hat{a}_{k,k+1} & \cdots & \hat{a}_{kn} \\
\hline
 & & & & \hat{a}_{k+1,k} & \hat{a}_{k+1,k+1} & \cdots & \hat{a}_{k+1,n} \\
 & \mathbf{0} & & & \vdots & \vdots & & \vdots \\
 & & & & \hat{a}_{nk} & \hat{a}_{n,k+1} & \cdots & \hat{a}_{nn}
\end{array}\right]\left[\begin{array}{cccc}
x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(\ell)} \\
x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(\ell)} \\
x_3^{(1)} & x_3^{(2)} & \cdots & x_3^{(\ell)} \\
\vdots & \vdots & & \vdots \\
x_k^{(1)} & x_k^{(2)} & \cdots & x_k^{(\ell)} \\
x_{k+1}^{(1)} & x_{k+1}^{(2)} & \cdots & x_{k+1}^{(\ell)} \\
\vdots & \vdots & & \vdots \\
x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(\ell)}
\end{array}\right] =
$$

$$
\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{n-k}
$$

$$
= \left[\begin{array}{cccc}
b_1^{(1)} & b_1^{(2)} & \cdots & b_1^{(\ell)} \\
\hat{b}_2^{(1)} & \hat{b}_2^{(2)} & \cdots & \hat{b}_2^{(\ell)} \\
\hat{b}_3^{(1)} & \hat{b}_3^{(2)} & \cdots & \hat{b}_3^{(\ell)} \\
\vdots & \vdots & & \vdots \\
\hat{b}_k^{(1)} & \hat{b}_k^{(2)} & \cdots & \hat{b}_k^{(\ell)} \\
\hline
\vdots & \vdots & & \vdots \\
\hat{b}_n^{(1)} & \hat{b}_n^{(2)} & \cdots & \hat{b}_n^{(\ell)}
\end{array}\right] \tag{2.59}
$$

όπου το ˆ πάνω από τα στοιχεία συμβολίζει ότι αυτά δεν είναι πλέον τα αρχικά στοιχεία του πίνακα A. Όπως αντιλαμβάνεται κανείς οι πράξεις για το επόμενο βήμα θα γίνουν για εκείνα τα στοιχεία τα οποία βρίσκονται στον κάτω δεξιά υποπίνακα, ο οποίος είναι $(n-k)\times(n-k+1)$. Σε κάθε ένα από τα στοιχεία του πίνακα αυτού θα προστεθεί ένας αριθμός (ο οποίος είναι ένα πολλαπλάσιο ενός στοιχείου της k γραμμής). Άρα απαιτούνται $(n-k)(n-k+1)$ προσθέσεις για το επόμενο βήμα. Παρατηρούμε ότι για τις $n-k$ γραμμές θα πρέπει να ορισθούν οι πολλαπλασιαστές $m_{ik}, i = k+1(1)n$ των οποίων ο υπολογισμός απαιτεί $n-k$ διαιρέσεις. Επιπλέον, κάθε ένας από τους πολλαπλασιαστές αυτούς πολλαπλασιάζεται με τα τελευταία $n-k+1$ στοιχεία της k γραμμής του πίνακα A και των ℓ δεύτερων μελών. Εφόσον η διαδικασία αυτή γίνεται για όλες τις $n-k$ εξισώσεις έπεται ότι απαιτούνται $(n-k)(n-k+1)$ πολλαπλασιασμοί. Επειδή τώρα $k = 1(1)n-1$ το πλήθος και το είδος των πράξεων για την τριγωνοποίηση του συστήματος θα είναι

$$
\sum_{k=1}^{n-1}(n-k) \qquad \text{διαιρέσεις}
$$

$$\sum_{k=1}^{n-1}(n-k)(n-k+\ell) \qquad \text{πολλαπλασιασμοί} \tag{2.60}$$

και

$$\sum_{k=1}^{n-1}(n-k)(n-k+\ell) \qquad \text{προσθαφαιρέσεις.}$$

Χρησιμοποιώντας τους τύπους

$$\sum_{k=1}^{m} k = \frac{m(m+1)}{2} \qquad \text{και} \qquad \sum_{k=1}^{m} k^2 = \frac{m(m+1)(2m+1)}{6}$$

οι (2.60) γράφονται

$$\frac{n(n-1)}{2} \qquad \text{διαιρέσεις}$$

$$\frac{n(n-1)(2n-1+3\ell)}{6} \qquad \text{πολλαπλασιασμοί} \tag{2.61}$$

και

$$\frac{n(n-1)(2n-1+3\ell)}{6} \qquad \text{προσθαφαιρέσεις.}$$

Για τον υπολογισμό των λύσεων $x_k^{(i)}, k = 1(1)n, i = 1(1)\ell$ χρησιμοποιούνται οι τύποι (2.7) από τους οποίους παρατηρούμε ότι για κάποιο i απαιτούνται $n-k$ πολλαπλασιασμοί για τα γινόμενα $a_{ij}x_j, j = k+1(1)n$, $n-k$ προσθαφαιρέσεις και μία διαίρση για $k = 1(1)n-1$, ενώ για τον υπολογισμό του $x_n^{(i)}$ απαιτείται μία διαίρεση μόνον. Συνεπώς για τον υπολογισμό όλων των $x_k^{(i)}$ απαιτούνται

$$\ell\sum_{k=1}^{n} 1 \qquad \text{διαιρέσεις}$$

$$\ell\sum_{k=1}^{n-1}(n-k) \qquad \text{πολλαπλασιασμοί}$$

και

$$\ell\sum_{k=1}^{n-1}(n-k) \qquad \text{προσθαφαιρέσεις}$$

δηλαδή

$$n\ell \qquad \text{διαιρέσεις}$$

$$\frac{n(n-1)\ell}{2} \qquad \text{πολλαπλασιασμοί} \tag{2.62}$$

και

$$\frac{n(n-1)\ell}{2} \qquad \text{προσθαφαιρέσεις.}$$

Συνεπώς το συνολικό πλήθος των πράξεων για την εύρεση της λύσης του (2.59) είναι:

$$\frac{n(n-1+2\ell)}{2} \qquad \text{διαιρέσεις}$$

$$\frac{n(n-1)(2n-1+6\ell)}{6} \qquad \text{πολλαπλασιασμοί} \tag{2.63}$$

και

$$\frac{n(n-1)(2n-1+6\ell)}{6} \qquad \text{προσθαφαιρέσεις.}$$

Έτσι λοιπόν για την επίλυση ενός μόνον γραμμικού συστήματος ($\ell = 1$) η μέθοδος απαλοιφής του Gauss απαιτεί

$$\frac{n^2}{2} + \frac{n}{2} \qquad \text{διαιρέσεις}$$

$$\frac{n^3}{3} + \frac{n^2}{2} - \frac{5n}{6} \qquad \text{πολλαπλασιασμοί} \tag{2.64}$$

$$\frac{n^3}{3} + \frac{n^2}{2} - \frac{5n}{6} \qquad \text{προσθαφαιρέσεις.}$$

Για την εύρεση του αντιστρόφου A^{-1} με τη μέθοδο της απαλοιφής του Gauss ($\ell = n$) απαιτούνται

$$\frac{3n^2}{2} - \frac{n}{2} \qquad \text{διαιρέσεις}$$

$$\frac{4n^3}{3} - \frac{3n^2}{2} - \frac{n}{6} \qquad \text{πολλαπλασιασμοί} \tag{2.65}$$

και

$$\frac{4n^3}{3} - \frac{3n^2}{2} - \frac{n}{6} \qquad \text{προσθαφαιρέσεις.}$$

Συμπεραίνουμε λοιπόν ότι το πλήθος των πράξεων για τη λύση ενός γραμμικού συστήματος με τη μέθοδο της απαλοιφής του Gauss είναι της τάξης $O(n^3/3)$ ενώ για την εύρεση του αντιστρόφου απαιτείται

τετραπλάσιο πλήθος πράξεων. Έτσι πάντοτε αποφεύγουμε να υπολογίσουμε τον αντίστροφο ενός πίνακα (είναι προτιμότερο να λύσουμε το σύστημα) εκτός αν μας ζητείται μόνον ο A^{-1}.

Για τη μέθοδο του Jordan έχουμε

$$\left[\begin{array}{cccc|ccc} a_{11} & & & & a_{1k} & a_{1,k+1} & \cdots & a_{1n} \\ & \hat{a}_{22} & & & \vdots & \vdots & & \vdots \\ & & \ddots & & \vdots & \vdots & & \vdots \\ & & & & \hat{a}_{kk} & \hat{a}_{k,k+1} & \cdots & \hat{a}_{kn} \\ \hline & & & & \hat{a}_{k+1,k} & \hat{a}_{k+1,k+1} & \cdots & \hat{a}_{k+1,n} \\ & & & & \vdots & \vdots & & \vdots \\ & & & & \hat{a}_{nk} & \hat{a}_{n,k+1} & \cdots & \hat{a}_{nn} \end{array}\right] \left[\begin{array}{c} \\ \\ \\ \\ \mathrm{X} \\ \\ \\ \\ \end{array}\right] = \left[\begin{array}{c} \\ \\ \\ \\ \mathrm{B} \\ \\ \\ \\ \end{array}\right]$$

$$\underbrace{\qquad\qquad}_{n-k} \qquad \underbrace{\qquad}_{\ell} \qquad \underbrace{\qquad}_{\ell}$$

Παρατηρούμε λοιπόν ότι για τους πολλαπλασιαστές απαιτούνται $n-1$ διαιρέσεις ενώ απαιτείται ίδιος αριθμός πολλαπλασιασμών και προσθαφαιρέσεων για τα στοιχεία του δεξιού πίνακα $(n-1)\times(n-k+1)$. Πιο συγκεκριμένα απαιτείται ένας πολλαπλασιασμός και μια πρόσθεση για κάθε ένα στοιχείο του υποπίνακα για την εκτέλεση ενός βήματος με τη μέθοδο απαλοιφής του Jordan. Έτσι απαιτούνται $(n-1)(n-k+1)$ πολλαπλασιασμοί και άλλες τόσες προσθέσεις. Συνεπώς για τη διαγωνιοποίηση του A απαιτούνται

$$\sum_{k=1}^{n}(n-1) \qquad \text{διαιρέσεις}$$

$$\sum_{k=1}^{n}(n-1)(n-k+1) \qquad \text{πολλαπλασιασμοί}$$

$$\sum_{k=1}^{n}(n-1)(n-k+1) \qquad \text{προσθαφαιρέσεις}$$

ή τελικά βρίσκουμε ότι χρειάζονται

$$n(n-1) \qquad \text{διαιρέσεις}$$

$$\frac{n(n-1)(n-1+2\ell)}{2} \qquad \text{πολλαπλασιασμοί}$$

$$\frac{n(n-1)(n-1+2\ell)}{2} \qquad \text{προσθαφαιρέσεις.}$$

Αν τώρα προστεθούν και οι n διαιρέσεις για την εύρεση μιας λύσης τότε για τα ℓ συστήματα απαιτούνται $n\ell$ διαιρέσεις. Άρα τελικά για τη λύση ℓ συστημάτων με τη μέθοδο απαλοιφής του Jordan απαιτούνται

$$\begin{aligned} n(n-1+\ell) \qquad & \text{διαιρέσεις} \\ \frac{n(n-1)(n-1+2\ell)}{2} \qquad & \text{πολλαπλασιασμοί} \\ \frac{n(n-1)(n-1+2\ell)}{2} \qquad & \text{προσθαφαιρέσεις.} \end{aligned} \tag{2.66}$$

Για την επίλυση ενός γραμμικού συστήματος οι ανωτέρω τύποι δίνουν ($\ell = 1$)

$$\begin{aligned} n^2 \qquad & \text{διαιρέσεις} \\ \frac{n^3}{2} - \frac{n}{2} \qquad & \text{πολλαπλασιασμοί} \\ \frac{n^3}{2} + \frac{n}{2} \qquad & \text{προσθαφαιρέσεις,} \end{aligned} \tag{2.67}$$

ενώ για τον υπολογισμό του αντιστρόφου ($\ell = n$) δίνουν

$$\begin{aligned} 2n^2 - n \qquad & \text{διαιρέσεις} \\ \frac{3n^3}{2} - 2n^2 + \frac{n}{2} \qquad & \text{πολλαπλασιασμοί} \\ \frac{3n^3}{2} - 2n^2 + \frac{n}{2} \qquad & \text{προσθαφαιρέσεις.} \end{aligned} \tag{2.68}$$

Από τους (2.67) παρατηρούμε ότι για την επίλυση ενός γραμμικού συστήματος με τη μέθοδο του Jordan το πλήθος των πράξεων είναι της τάξης $O(n^3/2)$. Δηλαδή η μέθοδος του Jordan απαιτεί περίπου 50% παραπάνω πράξεις απ΄ ότι η μέθοδος του Gauss για τη λύση ενός γραμμικού συστήματος. Αλλά και για τον υπολογισμό του αντιστρόφου η μέθοδος του Jordan απαιτεί περισσότερες πράξεις. Ωστόσο αν υποτεθεί ότι για τον υπολογισμό του A^{-1} δεν λαμβάνουμε υπόψη τις πράξεις που γίνονται στα δεύτερα μέλη με τα μηδενικά και τις μονάδες, αποδεικνύεται (η απόδειξη αφήνεται για τον αναγνώστη) ότι οι δύο μέθοδοι απαιτούν τον ίδιο ακριβώς πλήθος πράξεων. Γι' αυτό και η μέθοδος του Jordan, αν χρησιμοποιηθεί, θα είναι μόνον για τον υπολογισμό του αντιστρόφου ενός πίνακα.

Είναι φανερό ότι η απαλοιφή του Gauss με μερική οδήγηση θα πρέπει να προτιμάται για την λύση πυκνών γραμμικών συστημάτων.

Η μέθοδος αυτή είναι ευσταθής τουλάχιστον για μία μεγάλη κλάση προβλημάτων όπως αποδεικνύει ο Wilkinson. Επίσης για πίνακες οι οποίοι είναι πραγματικοί συμμετρικοί και θετικά ορισμένοι δεν χρειάζεται η μερική οδήγηση προκειμένου η μέθοδος του Gauss να έχει αριθμητική ευστάθεια.

2.3 Η LU μέθοδος

Ας υποθέσουμε ότι έχουμε την περίπτωση της μεθόδου απαλοιφής του Gauss χωρίς οδήγηση, όπου το σύστημα $A^{(1)}x = b^{(1)}$ μετασχηματίζεται στο τριγωνικό σύστημα

$$MA^{(1)}x = Mb^{(1)}$$

ή στο

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

οπότε παρατηρούμε ότι

$$A^{(1)} = M^{-1}A^{(n)} \tag{2.69}$$

όπου λόγω της (2.56)

$$M^{-1} = \left[M^{(1)}\right]^{-1}\left[M^{(2)}\right]^{-1}\dots\left[M^{(n-1)}\right]^{-1} \tag{2.70}$$

και

$$\left[M^{(r)}\right]^{-1} = \left[\begin{array}{c|ccccc} I_{r-1} & \multicolumn{5}{c}{\mathbf{0}} \\ \hline & 1 & & & & \\ & -m_{r+1,r} & 1 & & \mathbf{0} & \\ \mathbf{0} & -m_{r+2,r} & 0 & 1 & & \\ & \vdots & \vdots & & \ddots & \\ & -m_{nr} & 0 & \cdots & 0 & 1 \end{array}\right]. \tag{2.71}$$

Με απλούς πολλαπλασιασμούς μπορούμε να βρούμε ότι

$$M^{-1} = \begin{bmatrix} 1 & & & & & \\ -m_{21} & 1 & & & \mathbf{0} & \\ -m_{31} & -m_{32} & 1 & & & \\ -m_{41} & -m_{42} & -m_{43} & 1 & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & \\ -m_{n1} & -m_{n2} & -m_{n3} & -m_{n4} & \cdots & 1 \end{bmatrix} \tag{2.72}$$

και συνεπώς ο M^{-1} είναι ένας μοναδιαίος κάτω τριγωνικός πίνακας. Επειδή ο $A^{(n)}$ είναι ένας άνω τριγωνικός πίνακας, η (2.69) δηλώνει ότι αν κανένα από τα οδηγά στοιχεία δεν μηδενίζεται, κατά τη διάρκεια του σχηματισμού του $A^{(n)}$ από τον $A^{(1)}$, τότε ο $A^{(1)}$ μπορεί να διασπασθεί σε ένα γινόμενο ενός μοναδιαίου κάτω τριγωνικού πίνακα M^{-1} και ενός άνω τριγωνικού πίνακα $A^{(n)}$. Το επόμενο θεώρημα δίνει ικανές συνθήκες για την ύπαρξη μιας τέτοιας διάσπασης.

Θεώρημα 2.3.1. *Ένας $n \times n$ πίνακας $A = (a_{ij})$ μπορεί να γραφεί σαν το γινόμενο LU, όπου L είναι κάτω τριγωνικός πίνακας και U άνω τριγωνικός πίνακας, αν*

$$\det [a_{11}] \neq 0, \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \neq 0, \ldots, \det A \neq 0.$$

Επιπλέον, η τριγωνική αυτή διαχώριση είναι μοναδική αν τα διαγώνια στοιχεία είτε του L ή του U είναι ορισμένα.

Απόδειξη. Ας υποθέσουμε ότι ο A μπορεί να γραφεί σαν το γινόμενο

$$A = LU \tag{2.73}$$

όπου ο πίνακας L είναι ένας μοναδιαίος κάτω τριγωνικός πίνακας, τότε θα έχουμε

$$A = \begin{bmatrix} 1 & & & & \\ \ell_{21} & 1 & & \mathbf{0} & \\ \ell_{31} & \ell_{32} & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ & u_{22} & u_{23} & \cdots & u_{2n} \\ & & u_{33} & \cdots & u_{3n} \\ & \mathbf{0} & & \ddots & \vdots \\ & & & & u_{nn} \end{bmatrix}. \tag{2.74}$$

Εάν εκτελέσουμε τον πολλαπλασιασμό των πινάκων στο δεξί μέλος και εξισώσουμε τα αντίστοιχα στοιχεία, θα λάβουμε n^2 μη γραμμικές εξισώσεις με n^2 αγνώστους. Είναι φανερό λοιπόν ότι αν τα διαγώνια στοιχεία του L δεν ήσαν ορισμένα (στη προκειμένη περίπτωση =1), τότε θα είχαμε n επιπλέον αγνώστους και η διαχώριση αυτή δεν θα ήταν μοναδική. Αλλά ας ξεκινήσουμε να προσδιορίσουμε τα στοιχεία των πινάκων L και U. Αν υπολογίσουμε τα στοιχεία της **πρώτης γραμμής** του γινομένου LU και τα εξισώσουμε με τα στοιχεία της

πρώτης γραμμής του A λαμβάνουμε τις εξισώσεις

$$\begin{aligned} u_{11} &= a_{11} \\ u_{12} &= a_{12} \\ &\vdots \\ u_{1n} &= a_{1n} \end{aligned} \tag{2.75}$$

οι οποίες δείχνουν ότι οι **πρώτες γραμμές των A και U ταυτίζονται**. Όμοια αν υπολογίσουμε τα στοιχεία της **πρώτης στήλης** του LU κάτω από την κύρια διαγώνιο, λαμβάνουμε τις εξισώσεις

$$\begin{aligned} \ell_{21}u_{11} &= a_{21} \\ \ell_{31}u_{11} &= a_{31} \\ &\vdots \\ \ell_{n1}u_{11} &= a_{n1}. \end{aligned} \tag{2.76}$$

Επειδή $u_{11} = a_{11}$, αν $a_{11} \neq 0$, τότε οι (2.76) μας δίνουν αμέσως τα στοιχεία της **πρώτης στήλης** του L. Γενικά ας υποθέσουμε ότι θέλουμε να υπολογίσουμε τα στοιχεία της r γραμμής του U και στη συνέχεια τα στοιχεία της r στήλης του L, υποθέτοντας ότι τα στοιχεία των $k = 1(1)r-1$ γραμμών και στηλών του U και L, αντίστοιχα έχουν προηγουμένως προσδιορισθεί. Σχηματικά η όλη κατάσταση μπορεί να περιγραφεί ως ακολούθως

$$\left[\begin{array}{cccccc} 1 & & & & & \\ \ell_{21} & \ddots & & & \mathbf{0} & \\ \vdots & & & & & \\ \ell_{r1} & \cdots & 1 & & & \\ \hline \ell_{r+1,1} & \cdots & \boxed{\ell_{r+1,r}} & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ \ell_{n1} & \cdots & \boxed{\ell_{nr}} & \cdots & 1 \end{array}\right] \left[\begin{array}{cccccc} u_{11} & \cdots & u_{1r} & u_{1,r+1} & \cdots & u_{1n} \\ & \ddots & \vdots & \vdots & & \vdots \\ & & & u_{r-1,r+1} & \cdots & u_{r-1,n} \\ \hline & & \boxed{u_{rr}} & \boxed{u_{r,r+1}} & \cdots & \boxed{u_{rn}} \\ & & & \ddots & & \vdots \\ & \mathbf{0} & & & & u_{nn} \end{array}\right]$$

Εκτελώντας τον πολλαπλασιασμό της r γραμμής του L διαδοχικά με την $r, r+1, \ldots, n$ στήλη του U και εξισώνοντας με τα αντίστοιχα στοιχεία του A, λαμβάνουμε

$$\sum_{j=1}^{r-1} \ell_{rj}u_{jr} + u_{rr} = a_{rr}$$

$$\sum_{j=1}^{r-1} \ell_{rj} u_{j,r+1} + u_{r,r+1} = a_{r,r+1}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\sum_{j=1}^{r-1} \ell_{rj} u_{jn} + u_{rn} = a_{rn}.$$

Οι άγνωστοι στις ανωτέρω εξισώσεις είναι οι $u_{rp}, p = r(1)n$ οπότε έχουμε

$$u_{rp} = a_{rp} - \sum_{j=1}^{r-1} \ell_{rj} u_{jp}, \quad p = r(1)n. \tag{2.77}$$

Όμοια εκτελώντας τον πολλαπλασιασμό της r στήλης του U με τις $r+1, r+2, \ldots, n$ γραμμές του L και εξισώνοντας με τα αντίστοιχα στοιχεία του A λαμβάνουμε

$$\sum_{j=1}^{r-1} \ell_{r+1,j} u_{jr} + \ell_{r+1,r} u_{rr} = a_{r+1,r}$$

$$\sum_{j=1}^{r-1} \ell_{r+2,j} u_{j,r} + \ell_{r+2,r} u_{r,r} = a_{r+2,r}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\sum_{j=1}^{r-1} \ell_{nj} u_{jr} + \ell_{nr} u_{rr} = a_{nr}.$$

Οι άγνωστοι τώρα είναι οι $\ell_{pr}, r+1(1)n$ οπότε από τις παραπάνω εξισώσεις έχουμε, αν $u_{rr} \neq 0$, τότε τα στοιχεία της r στήλης του L δίνονται από τον τύπο

$$\ell_{pr} = \frac{1}{u_{rr}} \left(a_{pr} - \sum_{j=1}^{r-1} \ell_{pj} u_{jr} \right), \quad p = r+1(1)n. \tag{2.78}$$

Επιπλέον, από την (2.77) παρατηρούμε ότι για $p = r$ λαμβάνουμε

$$u_{rr} = a_{rr} - \sum_{j=1}^{r-1} \ell_{rj} u_{jr}, \quad r = 1(1)n$$

και αν $r = 2$ λαμβάνουμε διαδοχικά

$$\begin{aligned} u_{22} &= a_{22} - \ell_{21} u_{12} \\ &= a_{22} - \tfrac{a_{21}}{a_{11}} a_{12} \left(= a_{22}^{(2)} \right) \end{aligned}$$

ή τελικά

$$u_{22} = \frac{1}{a_{11}} \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

Αλλά από την υπόθεση έχουμε ότι $a_{11} \neq 0$ και $\det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \neq 0$ άρα $u_{22} \neq 0$. Εάν λοιπόν ισχύουν οι συνθήκες της υπόθεσης, τότε $u_{rr} \neq 0, r = 1(1)n$. ■

Παρατήρηση. Επειδή η τριγωνική διαχώριση είναι μοναδική όταν ο L είναι μοναδιαίος κάτω τριγωνικός τότε:

$$A^{(1)} = LU$$

με

$$M^{-1} = L \quad \text{και} \quad A^{(n)} = U.$$

Με άλλα λόγια η μέθοδος απαλοιφής του Gauss χωρίς οδήγηση είναι μαθηματικά ίδια με την τριγωνική διαχώριση LU με L μοναδιαίο κάτω τριγωνικό πίνακα. Ο Wilkinson ([1965] σελ. 223) αποδεικνύει ότι οι δύο μέθοδοι δεν δίνουν μόνο τα ίδια μαθηματικά αποτελέσματα αλλά, αν οι υπολογισμοί και στις δύο διαδικασίες γίνουν σε ένα υπολογιστή με την αριθμητική της κινητής υποδιαστολής, τότε ακόμα και τα σφάλματα στρογγύλευσης είναι τα ίδια. Επίσης τυχόν αποτυχία της τριγωνικής διαχώρισης λαμβάνει χώρα στις ίδιες περιπτώσεις με τη μέθοδο απαλοιφής του Gauss χωρίς οδήγηση. Ωστόσο υπάρχει ένα ουσιαστικό πλεονέκτημα της τριγωνικής διαχώρισης σε σχέση με τη μέθοδο απαλοιφής του Gauss για τους υπολογιστές εκείνους που διαθέτουν τη δυνατότητα της καταχώρησης υπολογισμών της μορφής $\sum_{j=1}^{n} a_j b_j$ σε διπλού μήκους λέξεις. Με τη δυνατότητα αυτή δεν απαιτείται επιπλέον χρόνος για να εργασθεί κανείς με διπλού μήκους λέξεις, πράγμα που σημαίνει μεγαλύτερη ακρίβεια στους υπολογισμούς. Επειδή η μέθοδος της απαλοιφής του Gauss δεν παρουσιάζει υπολογισμούς της μορφής αθροισμάτων γινομένων, γιαυτό θα πρέπει να προτιμάται η τριγωνική διαχώριση όπου είναι διαθέσιμη η παραπάνω

δυνατότητα.

Παρατήρηση. Όπως αναφέρθηκε προηγουμένως, η κλάση των πινάκων για τους οποίους δεν χρειάζεται η μερική οδήγηση περιλαμβάνει εκείνους τους πίνακες οι οποίοι είναι πραγματικοί, συμμετρικοί και θετικά ορισμένοι. Περιλαμβάνει επίσης τους μη ιδιάζοντες πίνακες οι οποίοι είναι διαγώνια υπέρτεροι, δηλαδή για τους οποίους

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|, \quad i = 1(1)n. \tag{2.79}$$

Παρατήρηση. Ας υποθέσουμε ότι ο A μπορεί να διαχωριθεί

$$A = LU$$

τότε στη περίπτωση που έχουμε να λύσουμε ένα γραμμικό σύστημα $Ax = b$ λαμβάνουμε

$$LUx = b \tag{2.80}$$

ή ισοδύναμα τα δύο τριγωνικά συστήματα

$$Ly = b \tag{2.81}$$

και

$$Ux = y. \tag{2.82}$$

Η λύση του κάτω τριγωνικού συστήματος (2.81) δίνεται από τους τύπους ($\ell_{ii} = 1, i = 1(1)n$)

$$y_1 = b_1$$

και

$$y_i = b_i - \sum_{j=1}^{i-1} \ell_{ij} y_j, \quad i = 2(1)n \tag{2.83}$$

ενώ η λύση του (2.82) από τους

$$x_n = y_n / u_{nn}$$

και

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

Παρατήρηση. Χρησιμοποιώντας την LU μέθοδο παρατηρούμε ότι τα στοιχεία των L και U υπολογίζονται αμέσως, χωρίς ενδιάμεσους υπολογισμούς όπως αυτό είναι αναγκαίο με την χρησιμοποίηση της μεθόδου απαλοιφής του Gauss. Επίσης τόσο τα στοιχεία του L όσο και του U αποθηκεύονται στα στοιχεία του A χωρίς αυτό να προκαλεί καμία ανωμαλία. Στο τέλος δηλαδή της τριγωνικής διαχώρισης θα έχουμε, αντί του πίνακα $A = (a_{ij})$, τον παρακάτω πίνακα για $n = 4$

$$\begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ \ell_{21} & u_{22} & u_{23} & u_{24} \\ \ell_{31} & \ell_{32} & u_{33} & u_{34} \\ \ell_{41} & \ell_{42} & \ell_{43} & \ell_{44} \end{bmatrix}.$$

Τελειώνοντας τις παρατηρήσεις, ας σημειωθεί ότι η τριγωνική διαχώριση για την οποία ο L είναι ένας μοναδιαίος κάτω τριγωνικός πίνακας είναι γνωστή σαν μέθοδος του Doolittle ενώ στην περίπτωση όπου ο U είναι μοναδιαίος άνω τριγωνικός είναι γνωστή σαν η μέθοδος του Crout.

Παράδειγμα

Να εφαρμοστεί η LU μέθοδος στον πίνακα

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

Λύση

$$A = LU$$

ή

$$\begin{bmatrix} 1 & 1 & -1 \\ 1 & 2 & -2 \\ -2 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \boxed{2}\ell_{21} & 1 & 0 \\ \ell_{31} & \boxed{4}\ell_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12}^{\boxed{1}} & u_{13} \\ 0 & u_{22}^{\boxed{3}} & u_{23} \\ 0 & 0 & u_{33}^{\boxed{5}} \end{bmatrix}$$

Ακολουθώντας την προτεραιότητα υπολογισμού των στοιχείων που δηλώνουν οι αριθμοί στα τετράγωνα έχουμε διαδοχικά

$$
\begin{array}{l}
u_{11} = 1 \quad u_{12} = 1 \quad u_{13} = -1 \\
\ell_{21} u_{11} = 1 \quad \text{ή} \quad \ell_{21} = 1 \\
\ell_{31} u_{11} = -2 \quad \text{ή} \quad \ell_{31} = -2 \\
u_{22} + \ell_{21} u_{12} = 2 \quad \text{ή} \quad u_{22} = 1 \\
\ell_{21} u_{13} + u_{23} = -2 \quad \text{ή} \quad \ell_{23} = -1 \\
\ell_{31} u_{12} + \ell_{32} u_{22} = 1 \quad \text{ή} \quad \ell_{32} = 3 \\
\ell_{31} u_{13} + \ell_{32} u_{23} + u_{33} = 1 \quad \text{ή} \quad u_{33} = 2.
\end{array}
$$

Συνεπώς

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

Ο Αλγόριθμος της LU μεθόδου

Ο παρακάτω αλγόριθμος παραγοντοποιεί ένα $n \times n$ πίνακα A με την LU μέθοδο.

1. Διάβασε τα δεδομένα: την τάξη n, τα στοιχεία $a_{ij}, \quad i, j = 1(1)n$ του A και τα στοιχεία $l_{ii}, \quad i = 1(1)n$ του L ή τα στοιχεία $u_{ii}, \quad i = 1(1)n$ του U.

2. Να επιλεγούν l_{11} και u_{11} τέτοια ώστε να ικανοποιείται η $l_{11}u_{11} = a_{11}$.

   Αν $l_{11}u_{11} = 0$ τότε τύπωσε (παραγοντοποίηση αδύνατος). Τέλος.

3. Για $j = 2(1)n$ να τεθεί

$$
\begin{array}{l}
u_{1j} = a_{1j}/l_{11}(\text{ πρώτη γραμμή του } U) \\
l_{j1} = a_{j1}/u_{11}(\text{ πρώτη γραμμή του } L)
\end{array}
$$

4. Για $r = 2(1)n - 1$ να εκτελεστούν τα βήματα 4.1-4.2

   4.1 Να επιλεγούν l_{rr} και u_{rr} ώστε να ικανοποιείται η

$$l_{rr}\, u_{rr} = a_{rr} - \sum_{j=1}^{r-1} l_{rj}\, u_{jr}$$

Αν $l_{rr}\, u_{rr} = 0$ τότε τύπωσε (παραγοντοποίηση αδύνατος).

4.2 Για $p = r + 1(1)n$ να τεθεί

$$u_{rp} = \left(a_{rp} - \sum_{j=1}^{r-1} l_{rj}\, u_{jp} \right) / \, l_{rr}(\ r \text{ γραμμή του } U)$$

και

$$l_{pr} = \left(a_{pr} - \sum_{j=1}^{r-1} l_{pj}\, u_{jr} \right) / \, u_{rr}(\ r \text{ γραμμή του } L)$$

5. Να επιλεγούν u_{nn} και l_{nn} ώστε να ικανοποιείται η

$$u_{nn}\, l_{nn} = a_{nn} - \sum_{j=1}^{r-1} l_{nj}\, u_{jn}$$

(Παρατηρούμε ότι αν $i_{nn}\, u_{nn} = 0,$ τότε $A = LU$ αλλά ο A θα είναι ένας ιδιάζων πίνακας)

6. Εκτύπωση (l_{ij}, $j = 1(1)i,$ και $i = 1(1)n$)
Εκτύπωση (u_{ij}, $j = 1(1)n,$ και $i = 1(1)n$)
Τέλος

Παρατήρηση. Υπάρχουν όμως και άλλοι τρόποι υπολογισμού των στοιχείων των L και U. Αν φανταστούμε ότι τα στοιχεία των πινάκων αυτών είναι τοποθετημένα σε ένα πίνακα ως εξής:

$$\begin{bmatrix} u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ l_{21} & u_{22} & u_{23} & \dots & u_{2n} \\ l_{31} & l_{32} & u_{33} & \dots & u_{3n} \\ . & & & & \\ . & & & & \\ . & & & & \\ l_{n1} & l_{n2} & l_{n3} & \dots & u_{nn} \end{bmatrix}$$

τότε είναι δυνατόν να υπολογίσουμε τα στοιχεία του παραπάνω πίνακα κατά γραμμές, ξεκινώντας από την πρώτη, δεύτερη, τρίτη κ.ο.κ. Πράγματι, έχουμε ότι

$$u_{1j} = a_{1j}, \quad j = 1(1)n.$$

Αν τώρα υποθέσουμε ότι βρισκόμαστε στον υπολογισμό της r γραμμής του παραπάνω πίνακα, δηλαδή έχουμε

$$\left[\begin{array}{cccccc|c}
1 & & & & & & \\
l_{21} & 1 & & & & & \\
l_{31} & l_{32} & 1 & & & & \\
. & . & . & . & & & \mathbf{0} \\
. & . & . & & & & \\
. & . & . & & . & & \\
\hline
l_{r1} & l_{r2} & l_{r3} & \dots & l_{r,r-1} & 1 & \\
. & . & . & & . & & \\
. & . & . & & & . & \\
. & . & . & & & & . \\
l_{n1} & l_{n2} & l_{n3} & & \dots & & 1
\end{array}\right]
\left[\begin{array}{ccccc}
u_{11} & u_{12} & \dots & u_{1r} & \dots & u_{1n} \\
 & u_{22} & \dots & u_{2r} & \dots & u_{2n} \\
 & . & & . & & . \\
 & & . & . & & . \\
 & & & . & & . \\
 & & & . & & . \\
\hline
 & & & u_{rr} & \dots & u_{rn} \\
 & \mathbf{0} & & & & \\
 & & . & & & . \\
 & & & . & & . \\
 & & & & . & . \\
 & & & & & u_{nn}
\end{array}\right]$$

όπου τα στοιχεία πάνω από την οριζόντιο γραμμή έχουν ήδη υπολογισθεί, τότε τα στοιχεία της r γραμμής του L υπολογίζονται από την λύση του συστήματος

$$\begin{array}{lllllllll}
l_{r1}u_{11} & & & & & & & & = a_{r1} \\
l_{r1}u_{12} & + & l_{r2}u_{22} & & & & & & = a_{r2} \\
l_{r1}u_{13} & + & l_{r2}u_{23} & + & l_{r3}u_{33} & & & & = a_{r3} \\
\vdots & & & & & & & & \vdots \\
l_{r1}u_{1r} & + & l_{r2}u_{2r} & + & l_{r3}u_{3r} & + \dots + & l_{r,r-1}u_{r-1,r-1} & & = a_{r,r-1}
\end{array}$$

ή του

$$\begin{bmatrix} u_{11} & & & & \\ u_{12} & u_{22} & & & \\ u_{13} & u_{23} & u_{33} & & \mathbf{0} \\ . & & . & & \\ . & & & . & \\ . & & & & . \\ u_{1r} & u_{2r} & u_{3r} & \dots & u_{r-1,r-1} \end{bmatrix} \begin{bmatrix} l_{r1} \\ l_{r2} \\ l_{r3} \\ . \\ . \\ . \\ l_{r,r-1} \end{bmatrix} = \begin{bmatrix} a_{r1} \\ a_{r2} \\ a_{r3} \\ . \\ . \\ . \\ a_{r,r-1} \end{bmatrix}$$

ενώ τα στοιχεία της r γραμμής του U υπολογίζονται από τις σχέσεις

$$u_{rp} = a_{rp} - \sum_{j=1}^{r-1} l_{rj}\, u_{jp}, \quad p = r(1)n.$$

Με ανάλογο τρόπο μπορούμε να υπολογίσουμε τα στοιχεία των L και U κατά στήλες. Εργαζόμενοι ανάλογα έχουμε

$$\left[\begin{array}{cccc|c} l & & & & \\ l_{21} & & & & \\ l_{31} & l_{32} & 1 & & \mathbf{0} \\ \vdots & & & \ddots & \\ l_{r1} & l_{r2} & l_{r3} \dots l_{r,r-1} & & 1 \\ l_{r+1,1} & l_{r+1,2} & l_{r+1,3} \dots l_{r+1,r-1} & & l_{r+1,r} \quad 1 \\ \vdots & & & & \\ l_{n1} & & & & l_{n,r} \quad \dots \quad 1 \end{array}\right] \left[\begin{array}{ccc|c} u_{11} & u_{12} \dots & & u_{1r} \dots u_{1n} \\ & u_{22} \dots & & u_{2r} \dots u_{2n} \\ & \ddots & & \\ & & & u_{rr} \dots u_{rn} \\ & \mathbf{0} & \ddots & \vdots \\ & & & u_{nn} \end{array}\right]$$

$$\begin{array}{rcccccccl} u_{1r} & & & & & & & & = a_{1r} \\ l_{21}u_{1r} & + & u_{2r} & & & & & & = a_{2r} \\ \vdots & & & & & & & & \vdots \\ l_{r1}u_{1r} & + & l_{r2}u_{2r} & + & \dots & + & l_{r,r-1}u_{r-1,r} & + & u_{rr} = a_{r,r} \end{array}$$

ή

$$\begin{bmatrix} 1 & & & \\ l_{21} & 1 & & \\ l_{31} & l_{32} & 1 & \mathbf{0} \\ \vdots & & \ddots & \\ l_{r1} & l_{r2} & l_{r3} \dots l_{r,r-1} & 1 \end{bmatrix} \begin{bmatrix} u_{1r} \\ u_{2r} \\ \\ \vdots \\ u_{rr} \end{bmatrix} = \begin{bmatrix} a_{1r} \\ a_{2r} \\ \\ \vdots \\ a_{rr} \end{bmatrix}$$

όπου ο υπολογισμός της r στήλης του L γίνεται από τις σχέσεις

$$l_{pr} = \left(a_{pr} - \sum_{j=1}^{r-1} l_{pj}\, u_{jr} \right) \; / \; u_{rr}, \quad p = r + 1(1)n.$$

2.4 Παραλλαγές της LU μεθόδου

Γενικά η LU παραγοντοποίηση του A δεν είναι μοναδική. Αν $A = LU$ είναι μια LU παραγοντοποίηση του A και D είναι ένας μη ιδιάζων διαγώνιος πίνακας τότε $L' = LD$ είναι ένας κάτω τριγωνικός και $U' = D^{-1}U$ είναι ένας άνω τριγωνικός πίνακας, άρα

$$A = LU = LDD^{-1}U = L'D'$$

και $L'U'$ είναι επίσης μια LU παραγοντοποίηση του A. Το παράδειγμα αυτό δείχνει την πιθανότητα κανονικοποίησης των LU παραγοντοποιήσεων ενός πίνακα με την εισαγωγή ενός διαγωνίου πίνακα. Θα λέμε ότι

$$A = LDU$$

είναι μια LDU παραγοντοποίηση του A αν ο L είναι μοναδιαίος κάτω τριγωνικός, ο D είναι διαγώνιος και ο U είναι μοναδιαίος άνω τριγωνικός.

Θεώρημα 4.1

Ο μη ιδιάζων πίνακας A έχει μια μοναδική LDU παραγοντοποίηση τότε και μόνο τότε αν οι υποπίνακες $A^{[r]}$, $r = 1(1)n-1$ όπου

$$A^{[r]} = \begin{bmatrix} a_{11} & \dots & a_{1r} \\ . & & . \\ . & & . \\ . & & . \\ a_{r1} & \dots & a_{rr} \end{bmatrix}$$

είναι αντιστρέψιμοι.

Απόδειξη

Πρώτα αποδεικνύεται ότι αν ο A έχει μια LDU παραγοντοποίηση τότε αυτή είναι μοναδική. Έστω $A = L_1D_1U_1$ και $A = L_2D_2U_2$.

Επειδή ο A είναι μη ιδιάζων, το ίδιο είναι και οι D_1 και D_2. Από την εξίσωση

$$L_1 D_1 U_1 = L_2 D_2 U_2$$

έχουμε

$$L_2^{-1} L_1 = D_2 U_2 U_1^{-1} D_1^{-1} \tag{2.84}$$

Το αριστερό μέλος της (3.4.3) είναι ένας μοναδιαίος κάτω τριγωνικός πίνακας και το δεξί μέλος είναι ένας άνω τριγωνικός πίνακας. Συνεπώς και τα δύο μέλη είναι ένας ταυτοτικός πίνακας, πράγμα που σημαίνει

$$L_2^{-1} L_1 = I$$

και

$$L_1 = L_2.$$

Όμοια μπορεί να δειχθεί ότι

$$U_1 = U_2.$$

Τέλος επειδή οι L_1 και U_1 είναι μη ιδιάζοντες, η $L_1 D_1 U_1 = L_2 D_2 U_2$ δηλώνει ότι

$$D_1 = D_2.$$

Απομένει λοιπόν να αποδειχθεί ότι ο A έχει μια LDU παραγοντοποίηση αν και μόνο αν οι $A^{[1]} \dots A^{[n-1]}$ είναι μη ιδιάζοντες.

Πρώτα υποθέτουμε ότι $A = LDU$ είναι μια LDU παραγοντοποίηση του A. Τότε οι L, D και U είναι μη ιδιάζοντες. Επειδή οι L και U είναι τριγωνικοί και ο D είναι διαγώνιος έπεται ότι οι $L^{[k]}$, $D^{[k]}$ και $U^{[k]}$ είναι μη ιδιάζοντες. Άλλά $A^{[k]} = L^{[k]} D^{[k]} U^{[k]}$ άρα ο $A^{[k]}$ είναι μη ιδιάζων. Αντίστροφα, ας υποθέσουμε ότι οι $A^{[r]}$, $r = 1(1) - 1$ είναι μη ιδιάζοντες. Τότε από το Θεώρημα 3.1 και την παρατήρηση μπορεί να εφαρμοστεί η μεθοδος της απαλοιφής του Gauss και να λάβουμε $A = LA^{(n)}$ όπου L και $A^{(n)}$ δίνονται από τις (4.28) και (3.103), αντίστοιχα. Τα διαγώνια στοιχεία του $A^{(n)}$ είναι τώρα τα οδηγά στοιχεία $a_{kk}^{(k)}$, $k = 1(1)n$ τα οποία είναι διάφορα του μηδενός. Έστω $D = diag\,(a_{11}^{(1)}, a_{22}^{(2)}, \dots, a_{nn}^{(n)})$ και $U = D^{-1} A^{(n)}$, τότε η

$$A = LDU$$

είναι μια LDU παραγοντοποίηση του A.■

Από την απόδειξη του παραπάνω θεωρήματος είναι φανερό ότι αν υπάρχει μια LDU παραγοντοποίηση του A τότε

$$\begin{aligned} l_{ij} &= -m_{ij}, \quad i = 2(1)n, \quad j = 1(1)i-1 \\ d_i &= a_{ii}^{(i)}, \qquad i = 1(1)n \end{aligned}$$

και

$$u_{ij} = \frac{a_{ij}^{(i)}}{a_{ii}^{(i)}}, \quad i = 1(1)n,\ j = i+1(1)n.$$

Πιο συγκεκριμένα υπάρχουν τρεις σημαντικές παραλλαγές μιας LDU παραγοντοποίησης. Η πρώτη είναι η

$$A = (LD)U = L'U$$

οπότε η διαχώριση είναι η

$$\begin{bmatrix} l'_{11} & & & \\ l'_{21} & l'_{22} & \mathbf{0} & \\ \cdot & \cdot & & \\ \cdot & & \cdot & \\ \cdot & & & \cdot \\ l'_{n1} & l'_{n2} & \dots & l'_{nn} \end{bmatrix} \begin{bmatrix} 1 & u_{11} & \dots & u_{nn} \\ & 1 & \dots & u_{2n} \\ & \cdot & & \cdot \\ & & \cdot & \cdot \\ & \mathbf{0} & & .. \\ & & & 1 \end{bmatrix} = L'U$$

όπου

$$l'_{ii} = d_i,\ i = 1(1)n,\ l'_{ij} = l_{ij}d_i,\ i = 2(1)n,\ j = 1(1)i-1$$

και η οποία είναι η Crout παραγοντοποίηση. Ωστόσο, αν συγκρίνουμε τα στοιχεία στην εξίσωση πινάκων $A = LDU$, τότε εργαζόμενοι ανάλογα όπως στο Θεώρημα 3.1 έχουμε

$$A = \begin{bmatrix} 1 & & & & & & & \\ l_{21} & 1 & & & & & & \\ l_{31} & l_{32} & 1 & & & & & \\ . & . & . & & & & & \\ l_{r1} & l_{r2} & l_{r3} & \dots & l_{r,r-1} & 1 & & \\ l_{r+1,1} & l_{r+1,2} & l_{r+1,3} & \dots & l_{r+1,r-1} & l_{r+1,r} & 1 & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ l_{n1} & l_{n2} & l_{n3} & \dots & l_{n,r-1} & l_{n,r} & \dots & 1 \end{bmatrix} \begin{bmatrix} d_{11} & & & \\ & d_{22} & & \\ & & . & \\ & & & d_{nn} \end{bmatrix}$$

$$\begin{bmatrix} 1 & u_{12} & \dots & u_{1r} & u_{1,r+1} & \dots & u_{1n} \\ & . & & & & & \\ & & . & & & & \\ & & & . & & & \\ & & & 1 & u_{r,r+1} & \dots & u_{rn} \\ & & & & 1 & \dots & \\ & & & & . & & \\ & & & & & . & \\ & & & & & & . \\ & & & & & & 1 \end{bmatrix}$$

ή

$$A = \left[\begin{array}{cccccc} 1 & & & & & \\ l_{21} & 1 & & & & \\ \vdots & & \ddots & & 0 & \\ l_{r1} & l_{r2} & \dots & l_{r,r-1} & 1 & \\ \hline l_{r+1,1} & l_{r+1,2} & \dots & l_{r+1,r-1} & \boxed{l_{r-1,r}} & 1 \\ \vdots & & & & \vdots & \ddots \\ l_{n1} & l_{n2} & \dots & l_{n,r-1} & \boxed{l_{n,r}} & \dots \quad 1 \end{array}\right]$$

$$\begin{bmatrix} d_{11} & d_{11}u_{12} & \dots & d_{11}u_{1r} & d_{11}u_{1,r-1} & \dots & d_{11}u_{1n} \\ & \ddots & & & & & \vdots \\ & & & \boxed{d_{rr}} & d_{rr}u_{r,r+1} & \dots & d_{rr}u_{rn} \\ & & 0 & & \ddots & & \vdots \\ & & & & & & d_{nn} \end{bmatrix}$$

ή τελικά

$$\begin{aligned} d_{11} &= a_{11} \\ d_{11}u_{12} &= a_{12} \\ &\vdots \\ d_{11}u_{1n} &= a_{1n} \end{aligned}$$

οπότε

$$d_{11} = a_{11}$$

και

$$u_{ij} = \frac{a_{ij}}{d_{11}}, \quad j = 2(1)n.$$

Επίσης

$$\begin{aligned} l_{21}d_{11} &= a_{21} \\ &\vdots \\ l_{n1}d_{11} &= a_{n1} \end{aligned}$$

άρα

$$l_{i1} = \frac{a_{i1}}{d_{11}}, \quad i = 2(1)n.$$

Γενικά λοιπόν έχουμε

$$a_{rr} = \sum_{j=1}^{r-1} l_{rj}\, d_{ii}\, u_{jr} + d_{rr}, \quad r = 2(1)n$$

επίσης

$$\begin{aligned} a_{rp} &= \sum_{j=1}^{r-1} l_{rj}\, d_{jj}\, u_{jp} + d_{rr}\, u_{rp}, \qquad p = r+1(1)n \\ a_{pr} &= \sum_{j=1}^{r-1} l_{pr}\, d_{jj}\, u_{jr} + l_{pr}\, d_{rr}, \qquad p = r+1(1)n \end{aligned}$$

Συνοψίζοντας, για τον υπολογισμό των L, D και U έχουμε

$$\left.\begin{aligned} d_{rr} &= a_{rr} - \sum_{\jmath=1}^{r-1} l_{r\jmath}\, d_{\jmath\jmath}\, u_{\jmath r} \\ u_{rp} &= \left(a_{rp} - \sum_{\jmath=1}^{r-1} l_{r\jmath}\, d_{\jmath\jmath}\, u_{\jmath p} \right) / d_{rr} \\ l_{rp} &= \left(a_{rp} - \sum_{\jmath=1}^{r-1} l_{p\jmath}\, d_{\jmath\jmath}\, u_{\jmath r} \right) / d_{rr} \end{aligned} \right\} p = r+1(1)n, \quad r = 1(1)n$$

η δεύτερη παραλλαγή είναι η

$$A = L(DU) = LU'$$

οπότε η διαχώρηση είναι η

$$\begin{bmatrix} 1 & & \\ l_{21} & 1 & \mathbf{0} \\ \vdots & \ddots & \\ l_{n1} & \dots & 1 \end{bmatrix} \begin{bmatrix} u'_{11} & \dots & u'_{1n} \\ & \ddots & \vdots \\ \mathbf{0} & & \\ & & u'_{nn} \end{bmatrix} = LU'$$

όπου

$$u'_{ii} = d_i,\ i = 1(1)n \text{ και } u'_{i\jmath} = d_i\, u_{i\jmath} = a^{(i)}_{i\jmath},\ i = 1(1)n,\ \jmath = 1(1)n$$

η οποία είναι γνωστή σαν η Doolitle παραγοντοποίηση.

Όταν ο A είναι συμμετρικός τότε η LU παραγοντοποίηση καταστρέφει τη συμμετρία ($L \neq U^T$). Στη περίπτωση αυτή θεωρούμε τη μορφή

$$A = LDL^T \tag{2.85}$$

όπου L είναι ένας μοναδιαίος κάτω τριγωνικός πίνακας. Αν τα διαγώνια στοιχεία του D είναι θετικά τότε μπορούμε να σχηματίσουμε τον πίνακα

$$D^{1/2} = diag(d_1^{1/2}, \dots, d_n^{1/2}) \tag{2.86}$$

οπότε ο A μπορεί να γραφεί σαν

$$A = LDL^T = (LD^{1/2})(LD^{1/2})^T = \overline{LL}^T \tag{2.87}$$

όπου

$$\overline{L} = LD^{1/2}. \tag{2.88}$$

Συνεπώς ο A μπορεί να γραφεί σαν

$$A = LL^T.$$

Η παραγοντοποίηση αυτή είναι γνωστή σαν η μέθοδος του Choleski και εφαρμόζεται στους συμμετρικούς και θετικά ορισμένους πίνακες καθόσο για τους πίνακες αυτούς υπάρχει πάντα μια Choleski παραγοντοποίηση. Στη συνέχεια θα παρουσιαστεί αναλυτικότερα η παραγοντοποίηση του Choleski.

Ορισμός. Ένας συμμετρικός πίνακας A είναι θετικά ορισμένος αν για $x \neq 0$ συνεπάγεται

$$x^T Ax > 0$$

και θετικά ημιορισμένος αν

$$x^T Ax \geq 0$$

Παράδειγμα

Έστω ότι ο A είναι ένας $m \times n$ πίνακας, τότε ο πίνακας $B = A^T A$ είναι συμμετρικός, πράγματι

$$B^T = (A^T A)^T = A^T A = B.$$

Επίσης ο B είναι θετικά ημιορισμένος. Πράγματι έστω $x \neq 0$ και $y = Ax$, τότε

$$x^T Bx = x^T A^T Ax = y^T y = \sum_{i=1}^{n} y_i^2 \geq 0.$$

Αν επιπλέον ο βαθμός (rank) του A είναι n, τότε για $x \neq 0$ και $y = Ax \neq 0$, οπότε $x^T Bx > 0$ άρα ο B είναι θετικά ορισμένος.

Θεώρημα 4.2

Αν ο A είναι n $\times$ n θετικά ορισμένος πίνακας τότε ο A είναι μη ιδιάζων.

Απόδειξη

Αν $x \neq 0$ είναι ένα διάνυσμα για το οποίο $Ax = 0$ τότε

$$x^T Ax = 0$$

πράγμα που είναι αντίθετο με την υπόθεση ότι ο A είναι θετικά ορισμένος. Συνεπώς το $Ax = 0$ έχει μόνο τη μηδενική λύση άρα ο πίνακας A είναι μη ιδιάζων.

Θεώρημα 4.3

Ένας κύριος υποπίνακας[1] ενός θετικά ορισμένου πίνακα είναι θετικά ορισμένος.

Απόδειξη

Έστω ένας κύριος υποπίνακας του A ο οποίος σχηματίζεται από την τομή των γραμμών και στηλών $i_1 < i_2 < \cdots < i_r$

Έστω επιπλέον και $\hat{x} \neq 0$ ένα διάνυσμα τάξης r. Κατασκευάζουμε στη συνέχεια το διάνυσμα x τάξης n τέτοιο ώστε

$$x_{i_k} = \hat{x}_k, \quad k = 1(1)r$$

και όλες οι άλλες $n-r$ συντεταγμένες του να είναι μηδέν. Τότε $x \neq 0$ και εύκολά διαπιστώνεται ότι

$$0 < x^T Ax = \hat{x}^T \hat{A} \hat{x}$$

επειδή ο A είναι θετικά ορισμένος. Άρα και ο $\hat{A}$ είναι θετικά ορισμένος.■

Από τα θεωρήματα 4.1, 4.2 και 4.3 συνεπάγεται ότι για ένα θετικά ορισμένο πίνακα A υπάρχει μια LDU παραγοντοποίηση.

Θεώρημα 4.4

Αν ο A είναι ένας $n \times n$ θετικά ορισμένος πίνακας τότε υπάρχει ένας μοναδιαίος κάτω τριγωνικός πίνακας L με θετικά διαγώνια στοιχεία τέτοιος ώστε

[1] Κύριος υποπίνακας ενός πίνακα A είναι εκείνος που σχηματίζεται από τα στοιχεία του A, τα οποία βρίσκονται στις τομές των γραμμών και στηλών $i_1 < i_2 < \cdots < i_r$.

$$A = LL^T \tag{2.89}$$

Απόδειξη

Επειδή ο A είναι θετικά ορισμένος, συνεπάγεται από τα θεωρήματα 4.1, 4.2 και 4.3 ότι υπάρχει μια LU παραγοντοποίηση του A, δηλαδή

$$A = LU$$

όπου ο L είναι ένας κάτω τριγωνικός και ο U ένας άνω τριγωνικός πίνακας. Επειδή ο A είναι θετικά ορισμένος έπεται ότι $a_{ii} > 0, \quad i = 1(1)n$ (γιατί;). Επίσης από τον αλγόριθμο 3.3.4 έχουμε

$$l_{11} = u_{11} = \sqrt{a}_{11}$$

Επειδή τώρα ο A είναι συμμετρικός έχουμε

$$l_{\jmath 1} = \frac{a_{\jmath 1}}{u_{11}} = \frac{a_{1\jmath}}{l_{11}} = u_{1\jmath}, \quad \jmath = 2(1)n \tag{2.90}$$

και τα στοιχεία της πρώτης γραμμής του U είναι τα ίδια με τα αντίστοιχα στοιχεία της πρώτης στήλης του L. Επαγωγικά υποθέτουμε ότι $k < n$ και ότι τα στοιχεία των πρώτων k γραμμών του U είναι τα ίδια με τα αντίστοιχα στοιχεία των πρώτων k στηλών του L. Η απόδειξη του θεωρήματος θα είναι πλήρης αν αποδειχθεί ότι τα στοιχεία της $k+1$ γραμμής του U είναι τα ίδια με τα στοιχεία της $k+1$ στήλης του L. Επειδή ο L είναι κάτω τριγωνικός και ο U είναι άνω τριγωνικός είναι φανερό ότι

$$l_{\jmath,k+1} = 0 = u_{k+1,\jmath}, \quad \jmath = 1, 2, \ldots, k.$$

Στη συνέχεια επειδή ο A είναι συμμετρικός έχουμε

$$l_{k+1,k+1}\, u_{k+1,k+1} = a_{k+1,k+1} - \sum_{\jmath=1}^{k} l_{k+1,\jmath}\, u_{\jmath,k+1}$$

ή

$$l_{k+1,k+1}\, u_{k+1,k+1} = a_{k+1,k+1} - \sum_{\jmath=1}^{k} l_{k+1,\jmath}^2$$

εκλέγουμε (Βλ. βήμα 4.1 του αλγόριθμου 3.3.1)

$$l_{k+1,k+1} = u_{k+1,k+1} = \left(a_{k+1,k+1} - \sum_{j=1}^{k} l_{k+1,j}^2 \right)^{1/2}. \qquad (2.91)$$

Η ποσότητα μέσα στην παρένθεση της (2.91) είναι θετική. Πράγματι ο κύριος υποπίνακας A_{k+1} είναι θετικά ορισμένος λόγω του θεωρήματος 4.3. Η ορίζουσα όμως ενός θετικά ορισμένου υποπίνακα είναι θετική καθόσον ισούται με το γινόμενο των ιδιοτιμών του. Έτσι

$$0 < \det(A_{k+1}) = \det(A_k) l_{k+1,k+1}\, u_{k+1,k+1}$$

ή επειδή $\det(A_k) > 0$,

$$l_{k+1,k+1}\, u_{k+1,k+1} > 0$$

άρα η τιμή του $l_{k+1,k+1}$ που δίνεται από την (2.91) είναι ένας πραγματικός αριθμός και μπορεί να ληφθεί θετική. Επιπλέον (Βλ. βήμα 4.2 του αλγόριθμου 3.3.1)

$$\begin{aligned} l_{p,k+1} &= \left(a_{p,k+1} - \sum_{j=1}^{k} l_{pj}\, u_{j,k+1} \right) / u_{k+1,k+1} \\ &= \left(a_{k+1,p} - \sum_{j=1}^{k} u_{jp}\, l_{k+1,j} \right) / l_{k+1,k+1} \\ &= u_{k+1,p}, \quad p = k + 2(1)n. \blacksquare \end{aligned}$$

2.4.1 Ο αλγόριθμος του Choleski

Για την παραγοντοποίηση ενός θετικά ορισμένου $n \times n$ πίνακα A έχουμε τον παρακάτω αλγόριθμο:

1. Διάβασε την τάξη n και τα στοιχεία $a_{ij}, \quad i, j = 1(1)n$ του πίνακα A.
2. Να τεθεί $l_{11} = \sqrt{a}_{11}$
3. Για $j = 2(1)n$ να τεθεί $l_{j1} = a_{j1}/l_{11}$
4. Για $r = 2(1)n - 1$ να εκτελεστούν τα βήματα 4.1-4.2

4.1 Να τεθεί

$$l_{rr} = \left[a_{rr} - \sum_{j=1}^{r-1} l_{rj}^2 \right]^{1/2}. \tag{2.92}$$

4.2 Για $p = r + 1(1)n$ να υπολογιστούν

$$l_{pr} = \frac{1}{l_{rr}} \left[a_{pr} - \sum_{j=1}^{r-1} l_{pj}\, l_{rj} \right] \tag{2.93}$$

5. Να τεθεί

$$l_{nn} = \left[a_{nn} - \sum_{j=1}^{r-1} l_{nj}^2 \right]^{1/2} \tag{2.94}$$

6. Να εκτυπωθούν τα $l_{rp}, \quad p = 1(1)r, \quad r = 1(1)n$. Τέλος.

Παρατήρηση. Κατά την κωδικοποίηση του αλγόριθμου του Choleski καλό είναι να ελέγχεται το πρόσημο της ποσότητας στην τετραγωνική ρίζα.

2.5 Η LU μέθοδος με μετρική οδήγηση

Στην §3.3 διαπιστώσαμε ότι η τριγωνική διαχώριση και η μέθοδος της απαλοιφής του Gauss χωρίς οδήγηση είναι μαθηματικά ίδιες. Ωστόσο οι μέθοδοι εξακολουθούν να είναι ίδιες αν εφαρμοστεί η τεχνική της μερικής οδήγησης; Η απάντηση στο ερώτημα αυτό δίνεται από το παρακάτω Θεώρημα.

Θεώρημα 5.1

Για ένα $n \times n$ πίνακα A υπάρχει ένας μεταθετικός πίνακας P, ένας μοναδιαίος κάτω τριγωνικός πίνακας L και ένας άνω τριγωνικός πίνακας U τέτοιος ώστε ο πίνακας PA να έχει την τριγωνική διαχώριση

$$PA = LU \tag{2.95}$$

Απόδειξη (βλ. [Young and Gregory] σελ. 127-129).■

Κατά την απόδειξη του ανωτέρω θεωρήματος δείχνεται ότι το αποτέλεσμα της μερικής οδήγησης κατά τη διάρκεια της διαδικασίας της απαλοιφής είναι μαθηματικά ίδιο με την εφαρμογή της απαλοιφής του Gauss χωρίς οδήγηση σε κάποιο πίνακα PA που λαμβάνεται από τον A μεταθέτοντας τις γραμμές του. Συνεπώς η μέθοδος της τριγωνικής διαχώρισης με μερική οδήγηση είναι ίδια με την παραγοντοποίηση του πίνακα PA. Επειδή δε $\det P \neq 0$ τα συστήματα $Ax = b$ και

$$PAx = Pb \tag{2.96}$$

είναι ισοδύναμα και δεν αλλάζει τίποτα αν αντί του A παραγοντοποιήσουμε τον PA. Ο Wilkinson [1967] αποδεικνύει ότι η παρακάτω τεχνική παραγοντοποιεί τον PA χωρίς να είναι ανάγκη να γνωρίζουμε τον P εκ των προτέρων.

Όπως και στη μέθοδο απαλοιφής του Gauss με μερική οδήγηση, εξετάζουμε την πρώτη στήλη του A και εκτελούμε μια εναλλαγή γραμμών αν το a_{11} δεν είναι το στοιχείο με την μεγαλύτερη απόλυτη τιμή από όλα τα στοιχεία της πρώτης στήλης. Στο σημείο αυτό μπορούμε να χρησιμοποιήσουμε τους τύπους (3.81) και (3.82) για τον υπολογισμό της πρώτης γραμμής του U και της πρώτης στήλης του L. Αν τα στοιχεία αυτά αποθηκευθούν στις αντίστοιχες θέσεις των στοιχείων του A θα έχουμε τον επαυξημένο πίνακα

$$\left[\begin{array}{c|cccc|c|c} u_{11} & u_{12} & u_{13} & \dots & u_{1n} & b_1 & 0 \\ \hline l_{21} & a_{22} & a_{23} & \dots & a_{2n} & b_2 & 0 \\ l_{31} & a_{32} & a_{33} & \dots & a_{3n} & b_3 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ l_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} & b_n & 0 \end{array}\right]$$

όπου η στήλη με τα μηδενικά θα εξηγηθεί στη συνέχεια. Για λόγους απλοποίησης του συμβολισμού δεν χρησιμοποιούνται άνω δείκτες για τη δήλωση του εκάστοτε βήματος, ούτε δηλώνονται οι εναλλαγές των γραμμών. Το δεύτερο βήμα είναι ακριβώς το ίδιο με το r βήμα κατά το οποίο υπολογίζονται οι ποσότητες

$$
\begin{aligned}
s_r &= a_{rr} - \sum_{j=1}^{r-1} l_{rj}\, u_{jr} \\
s_{r+1} &= a_{r+1,r} - \sum_{j=1}^{r-1} l_{r+1,j}\, u_{jr} \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
s_n &= a_{nr} - \sum_{j=1}^{r-1} l_{nj}\, u_{jr}
\end{aligned}
\tag{2.97}
$$

και τοποθετούνται στην τελευταία στήλη του επαυξημένου πίνακα. Έτσι αν $r = 2$ έχουμε

$$
\left[\begin{array}{c|cccc|c|c}
u_{11} & u_{12} & u_{13} & \dots & u_{1n} & b_1 & 0 \\
\hline
l_{21} & a_{22} & a_{23} & \dots & a_{2n} & b_2 & s_2 \\
l_{31} & a_{32} & a_{33} & \dots & a_{3n} & b_3 & s_3 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
l_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} & b_n & s_n
\end{array}\right]
$$

Από τις (4.32) παρατηρούμε ότι αν $r = 1$ τότε

$$
s_1 = a_{11}^{(1)} \quad s_2 = a_{21}^{(1)} \dots \; s_n = a_{n1}^{(1)}.
$$

Αν $r = 2$ τότε

$$
\begin{aligned}
s_2 &= a_{22}^{(1)} - l_{21}\, u_{12} \\
&= a_{22}^{(1)} + m_{21}\, a_{12}^{(1)} \\
&= a_{22}^{(2)}
\end{aligned}
$$

όμοια

$$
s_3 = a_{32}^{(2)} \dots \; s_n = a_{n2}^{(2)}
$$

Άρα γενικά έχουμε

$$
s_r = a_{rr}^{(r)} \quad s_{r+1} = a_{r+1,r}^{(r)} \dots \; s_n = a_{nr}^{(r)} \tag{2.98}
$$

οπότε το ψάξιμογια το οδηγό στοιχείο αναφέρεται στις ποσότητες s_i. Πιο συγκεκριμένα ο μικρότερος ακέραιος p για τον οποίο ισχύει

$$|s_p| = \underset{r \le i \le n}{max} |s_i|$$

δηλώνει το οδηγό στοιχείο. Αν λοιπόν το s_r δεν είναι οδηγό στοιχείο τότε εναλλάσσουμε τις r και $p \quad (r < p)$ γραμμές του επαυξημένου πίνακα έτσι ώστε το s_p να κατέχει την θέση του s_r. Στο σημείο αυτό παρατηρούμε ότι

$$u_{rr} = s_r$$

και τα υπόλοιπα στοιχεία της r γραμμής του u υπολογίζονται από τις σχέσεις

$$u_{rp} = a_{r,p} - \sum_{j=1}^{r-1} l_{rj}\, u_{jp}, \quad p = r + 1(1)n$$

Όμοια τα στοιχεία της r στήλης του L δίνονται από τις

$$l_{pr} = s_p / u_{rr}, \quad p = r + 1(1)n$$

Παρατήρηση. Όταν εναλλάσσονται οι γραμμές του επαυξημένου πίνακα, εναλλάσσονται και οι γραμμές στον μερικώς υπολογισθέντα πίνακα L. Αυτό είναι φυσικό να γίνει, γιατί τα στοιχεία l_{ij} συνδέονται με τους πολ/στές της απαλοιφής του Gauss και από την (3.80) παρατηρούμε ότι μια πιθανή εναλλαγή γραμμών επηρεάζει όλα τα προηγούμενα βήματα.

2.5.1 Ο αλγόριθμος της LU μεθόδου με μερική οδήγηση

Ο παρακάτω αλγόριθμος παραγοντοποιεί τον A σε LU με μερική οδήγηση και στη συνέχεια επιλύει τα συστήματα $Lz = b$ και $Ux = z$ όπου τα διαγώνια στοιχεία είτε του L ή του U είναι δεδομένα.

1. Διάβασε την τάξη n, τα στοιχεία $a_{ij},\ i = 1(1)n,\ j = 1(1)n + 1$ της επαυξημένης μορφής του A και τα στοιχεία $l_{ii},\ i = 1(1)n$ του L ή τα στοιχεία $u_{ii},\ i = 1(1)n$ του U.

2. Έστω p ο μικρότερος ακέραιος τέτοιος ώστε $1 \leq p \leq n$ και

$$|a_{p1}| = \max_{1 \leq j \leq n} |a_{j1}|$$

(εύρεση του πρώτου οδηγού στοιχείου)

Αν $|a_{p1}| = 0$ τότε τύπωσε (Δεν υπάρχει μοναδική λύση). Τέλος.

3. Αν $p \neq 1$ τότε να εναλλαχθούν οι γραμμές p και 1 στον A.

4. Να επιλεγούν l_{11} και u_{11} τέτοια ώστε να ικανοποιείται η

$$l_{11}\, u_{11} = a_{11}$$

5. Για $j = 2(1)n$ να τεθεί

$$\begin{array}{ll} u_{1j} & = a_{1j}/l_{11} \quad (\text{πρώτη γραμμή του } U) \\ l_{j1} & = a_{j1}/u_{11} \quad (\text{πρώτη στήλη του } L) \end{array}$$

6. Για $r = 2(1)n - 1$ να εκτελεσθούν τα βήματα 6.1-6.4.

   6.1 Έστω p ο μικρότερος ακέραιος τέτοιος ώστε $r \leq p \leq n$ και

$$\left| a_{pr} - \sum_{j=1}^{r-1} l_{pj}\, u_{jr} \right| = \max_{r \leq k \leq n} \left| a_{kr} - \sum_{j=1}^{r-1} l_{kj}\, u_{jr} \right|$$

   (εύρεση του οδηγού στοιχείου)

   Αν το μέγιστο είναι μηδέν τότε τύπωσε (Δεν υπάρχει μοναδική λύση)

   6.2 Αν $P \neq r$ τότε εναλλαγή των γραμμών p και r και στους δύο πίνακες A και L.

   6.3 Να επιλεγούν l_{rr} και u_{rr} που να ικανοποιούν την

$$l_{rr}\, u_{rr} = a_{rr} - \sum_{j=1}^{r-1} l_{rj}\, u_{jr}$$

6.4 Για $p = r + 1(1)n$ να τεθεί

$$\begin{aligned} u_{rp} &= \left(a_{rp} - \sum_{j=1}^{r-1} l_{rj}\, u_{jp} \right) / l_{rr} \quad (r \text{ γραμμή του } U) \\ l_{pr} &= \left(a_{pr} - \sum_{j=1}^{r-1} l_{pj}\, u_{jr} \right) / u_{rr} \quad (r \text{ στήλη του } L) \end{aligned}$$

7. Να τεθεί

$$hold = a_{nn} - \sum_{j=1}^{r-1} l_{nj}\, u_{jn}$$

Αν $hold = 0$ τότε τύπωσε (Δεν υπάρχει μοναδική λύση). Τέλος.

Να επιλεγούν l_{nn} και u_{nn} τέτοια ώστε να ικανοποιείται η

$$l_{nn}\, u_{nn} = a_{nn} - \sum_{j=1}^{r-1} l_{nj}\, u_{jn}$$

(Τα βήματα 8 και 9 επιλύουν το κάτω τριγωνικό σύστημα $Lz = b$).

8. Να τεθεί

$$z_1 = a_{1,n+1}/l_{11}$$

9. Για $i = 2(1)n$ να τεθεί

$$z_i = \left(a_{i,n+1} - \sum_{j=1}^{i-1} l_{ij}\, z_j \right) / l_{ii}$$

(Τα βήματα 10 και 11 επιλύουν το άνω τριγωνικό σύστημα $Ux = z$)

10. Να τεθεί

$$x_n = z_n/u_{nn}$$

11. Για $i = n-1(-1)1$ να τεθεί

$$x_i = \left(z_i - \sum_{j=i+1}^{n} u_{ij}\, x_j \right) / \, u_{ii}$$

12. Να τυπωθούν τα $x_i,\ i = 1(1)n$. Τέλος.

2.5.2 Λύση τριδιαγώνιου συστήματος

Ας θεωρήσουμε το τριδιαγώνιο σύστημα

$$\begin{aligned} b_1 x_1 + c_1 x_2 &= d_1 \\ a_2 x_1 + b_2 x_2 + c_2 x_3 &= d_2 \\ a_3 x_2 + b_3 x_3 + c_3 x_4 &= d_3 \\ &\cdots\cdots\cdots \\ &\cdots\cdots\cdots \\ a_{n-1} x_{n-2} + b_{n-1} x_{n-1} + c_{n-1} x_n &= d_{n-1} \\ a_n x_{n-1} + b_n x_n &= d_n \end{aligned} \tag{2.99}$$

Το παραπάνω σύστημα μπορεί να επιλυθεί με τη μέθοδο της απαλοιφής του Gauss χωρίς οδήγηση. Είναι φανερό ότι η μέθοδος θα πρέπει να λάβει υπόψη της τη δομή του συστήματος προκειμένου να εξοικονομηθεί ουσιαστική υπολογιστική δουλειά. Παρατηρούμε ότι αν εφαρμοστεί η μέθοδος της απαλοιφής του Gauss τότε θα προκύψει ένα άνω τριγωνικό σύστημα το οποίο θα έχει στοιχεία μόνο στην κύρια διαγώνιο και στη διαγώνιο που βρίσκεται άνω της κυρίας διαγωνίου. Με άλλα λόγια θα απαλειφθεί η διαγώνιος των a_i, δηλαδή σε κάθε βήμα της μεθόδου θα απαλείφεται μόνο ένα στοιχείο. Η δεύτερη παρατήρηση είναι ότι η διαγώνιος των c_i θα παραμείνει αμετάβλητη μετά το τέλος της απαλοιφής. Αν $b_1 \neq 0$ τότε ο x_1 απαλείφεται από την δεύτερη εξίσωση, σχηματίζοντας τον πλο/στή

$$m_1 = -\frac{a_2}{b_1}$$

και παράγοντας την νέα εξίσωση

$$b_2' x_2 + c_2 x_3 = d_2'$$

όπου

$$b_2' = b_2 + m_1 c_1$$

και

$$d_2' = d_2 + m_1 d_1$$

Όμοια αν $b_2' \neq 0$, ο x_2 μπορεί να απαλειφθεί από την τρίτη εξίσωση, αν ορισθεί ο πολ/στής.

$$m_2 = -\frac{a_3}{b_2'}$$

οπότε λαμβάνουμε τη νέα τρίτη εξίσωση

$$b_3' x_3 + c_3 x_4 = d_3'$$

όπου

$$b_3' = b_3 + m_2 c_2$$

και

$$d_3' = d_3 + m_2 d_2$$

Συνεχίζοντας στο i-οστό βήμα, ο x_i θα απαλειφθεί από την $i+1$ εξίσωση, αν ορισθεί ο πολ/στής

$$m_i = -\frac{a_{i+1}}{b_i'} \tag{2.100}$$

οπότε λαμβάνεται η νέα $i+1$ εξίσωση

$$b_{i+1}' x_{i+1} + c_{i+1} x_{i+2} = d_{i+1}' \tag{2.101}$$

όπου

$$b_{i+1}' = b_i' + m_i c_i \tag{2.102}$$

και

$$d_{i+1}' = d_{i+1} + m_i d_i' \tag{2.103}$$

Για $i = 1(1)n-1$ θα λάβουμε τελικά το σύστημα

$$
\begin{aligned}
b'_1 x_1 + c_1 x_2 &= d'_1 \\
b'_2 x_2 + c_2 x_3 &= d'_2 \\
&\cdots\cdots\cdots \\
&\cdots\cdots\cdots \\
b'_{n-1} x_{n-1} + c_{n-1} x_n &= d'_{n-1} \\
b'_n x_n &= d'_n
\end{aligned}
\tag{2.104}
$$

όπου $b'_1 = b_1$ και $d'_1 = d_1$
Η λύση του ανωτέρω συστήματος είναι η

$$
\begin{aligned}
x_n &= \frac{d'_n}{b'_n}, \quad b'_n \neq 0 \\
x_i &= (d'_i - c_i u_{i+1}) / \, b_i, \quad i = n-1(-1)1
\end{aligned}
\tag{2.105}
$$

με $b'_i \neq 0$.
Εύκολα λοιπόν προκύπτει ο αλγόριθμος.

1. Διάβασε την τάξη n του A, τα $b_i, \quad i = 1(1)n, \quad c_i, \quad i = 1(1)n-1$, τα $a_i, \quad i = 2(1)n$ και τα $d_i, \quad i = 1(1)n$

2. Για $i = 1(1)n-1$ να εκτελεστουν τα βήματα 2.1-2.3

   2.1 Να τεθεί

   $$m_i = -\frac{a_{i+1}}{b_i}$$

   2.2 Να τεθεί

   $$b_{i+1} = b_{i+1} + m_i c_i$$

   2.3 Να τεθεί

   $$d_{i+1} = d_{i+1} + m_i d_i$$

   (επίλυση του συστήματος)

3. Αν $b_n = 0$ τότε τύπωσε (Δεν υπάρχει μοναδική λύση). Τέλος.

4. Να τεθεί

$$x_n = \frac{d_n}{b_n}$$

5. Για $n-1(-1)$ να τεθεί

$$x_i = (d_i - c_i x_{i+1})/b_i$$

6. Να τυπωθεί η λύση $x_i, \quad i = 1(1)n$. Τέλος.

Στη συνέχεια ας υποθέσουμε ότι επιθυμούμε να επιλύσουμε ένα τριδιαγώνιο σύστημα με την LU μέθοδο.Παρατηρούμε ότι ο A έχει μόνο $(3n-2)$ μη μηδενικά στοιχεία, πράγμα που σημαίνει ότι το σύνολο των στοιχείων τόσο του L όσο και του U θα πρέπει να είναι επίσης $(3n-2)$. Αν υποθέσουμε ότι

$$L = \begin{bmatrix} l_1 & & & & \\ k_2 & l_2 & & \mathbf{0} & \\ & k_3 & l_3 & & \\ \mathbf{0} & & \ddots & \ddots & \\ & & & & k_n \; l_n \end{bmatrix} \text{ και } U = \begin{bmatrix} 1 & u_1 & & & \\ & 1 & u_2 & & \mathbf{0} \\ & & 1 & u_3 & \\ & & & \ddots & \ddots \\ \mathbf{0} & & & & 1 \; u_{n-1} \\ & & & & 1 \end{bmatrix} \tag{2.106}$$

Τότε υπάρχουν $(2n-1)$ άγνωστα στοιχεία του L και $(n-1)$ άγνωστα στοιχεία του U, δηλαδή έχουμε $(3n-2)$ άγνωστα στοιχεία τα οποία μπορούν να προσδιοριστούν από την $A = LU$, η οποία δίνει τις εξισώσεις

$$\begin{array}{rrrr} l_1 = b_1, & l_1 u_1 = c_1 & k_2 = a_2 & \\ k_2\, u_1 + l_2 = b_2, & l_2 u_2 = c_2 & k_3 = a_3 & \\ \ldots\ldots\ldots\ldots & \ldots\ldots & \ldots\ldots\ldots\ldots & (2.107) \\ k_n\, u_{n-1} + l_n = b_n & & k_n = a_n & \end{array}$$

ή

$$
\begin{aligned}
l_1 &= b_1 \\
u_1 &= c_1/l_1 \\
k_i &= a_i, \quad i = 2(1)n \\
l_i &= b_i - k_i\, u_{i-1}, \quad i = 2(1)n \\
u_i &= c_i/l_i, \quad i = 2(1)n-1
\end{aligned} \tag{2.108}
$$

ή τελικά

$$
\begin{aligned}
l_1 &= b_1 \\
u_1 &= c_1/l_1 \\
l_i &= b_i - a_i\, u_{i-1}, \quad i = 2(1)n \\
u_i &= c_i/l_i, \quad i = 2(1)n-1
\end{aligned} \tag{2.109}
$$

Το σύστημα (4.34) τώρα γράφεται

$$LUx = d$$

ή

$$Lz = d \tag{2.110}$$

και

$$Ux = z \tag{2.111}$$

Η λύση του (2.110) δίνεται από τους τύπους

$$
\begin{aligned}
z_1 &= d_1/l_1 \\
z_i &= (d_i - a_i\, z_{i-1})/l_i, \quad i = 2(1)n
\end{aligned} \tag{2.112}
$$

ενώ η λύση του (2.111) είναι η

$$
\begin{aligned}
x_n &= z_n \\
x_i &= z_i - u_i\, x_{i+1}, \quad i = n-1(-1)1
\end{aligned} \tag{2.113}
$$

Επομένως ο αλγόριθμος της μεθόδου του Grout για τη λύση του τριδιαγώνιου συστήματος (4.34) είναι ο ακόλουθος

1. Διάβασε την τάξη n του A, τα $b_i, \quad i = 1(1)n$, τα $a_i, \quad i = 2(1)n$, τα $c_i, \quad i = 1(1)n-1$ και τα $d_i, \quad i = 1(1)n$

2. Να τεθεί

$$l_1 = b_1$$

και

$$u_1 = c_1/l_1$$

3. Για $i = 2(1)n-1$ να τεθεί

$$\begin{aligned} l_i &= b_i - a_i\, u_{i-1} \\ u_i &= c_i/l_i \end{aligned}$$

4. Να τεθεί

$$l_n = b_n - a_n\, u_{n-1}$$

(Τα βήματα 5 και 6 επιλύουν το $Lz = d$).

5. Να τεθεί

$$z_1 = d_1/l_1$$

6. Για $i = 2(1)n$ να τεθεί

$$z_i = (d_i - a_i\, z_{i-1})\,/\,l_i$$

(Τα βήματα 7 και 8 επιλύουν το $Ux = z$).

7. Να τεθεί

$$x_n = z_n$$

8. Για $i = n-1(-1)1$ να τεθεί

$$x_i = z_i - u_i\, x_{i+1}$$

9. Να τυπωθεί η λύση $x_i, \quad i = 1(1)n$. Τέλος.

Εφαρμογή

Έστω το τριδιαγώνιο σύστημα των εξισώσεων

$$\begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Ακολουθώντας τα βήματα του αλγορίθμου του Grout έχουμε

1. $n = 4, \quad b_1 = b_2 = b_3 = b_4 = 2, \quad a_2 = a_3 = a_4 = -1$
   $c_1 = c_2 = c_3 = -1$ και $d_1 = 1$, $d_2 = 0$, $d_3 = 0$ και $d_4 = 1$

2. $l_1 = 2, \quad u_1 = -1/2$

3. $i = 2$

   $l_2 = b_2 - a_2\, u_1 = 2 - (-1)\left(-\frac{1}{2}\right) = \frac{3}{2}$

   $u_2 = c_2/l_2 = -1/\left(\frac{3}{2}\right) = -\frac{2}{3}$

   $i = 3$

   $l_3 = b_3 - a_3\, u_2 = 2 - (-1)\left(-\frac{2}{3}\right) = \frac{4}{3}$

   $u_3 = c_3/l_3 = -1/\left(\frac{4}{3}\right) = -\frac{3}{4}$

4. $l_4 = b_4 - a_4\, u_3 = 2 - (-1)\left(-\frac{3}{4}\right) = \frac{5}{4}$

5. $z_1 = d_1/l_1 = 1/2$

6. $i = 2$

   $z_2 = (d_2 - a_2\, z_1)/l_2 = \frac{0-(-1)\left(\frac{1}{2}\right)}{\left(\frac{3}{2}\right)} = \frac{1}{3}$

   $i = 3$

   $x_3 = (d_3 - a_3\, z_2)/l_3 = \frac{0-(-1)\left(\frac{1}{3}\right)}{\left(\frac{4}{3}\right)} = \frac{1}{4}$

   $i = 4$

   $z_4 = (d_4 - a_4\, z_3)/l_4 = \frac{1-(-1)\left(\frac{1}{4}\right)}{\left(\frac{5}{4}\right)} = 1$

7. $x_4 = z_4 = 1$

8. $i = 3$

$$x_3 = z_3 - u_3\, x_4 = \tfrac{1}{4} - \left(-\tfrac{3}{4}\right) \cdot 1 = 1$$

$i = 2$

$$x_2 = z_2 - u_2\, x_3 = \tfrac{1}{3} - \left(-\tfrac{2}{3}\right) \cdot 1 = 1$$

$i = 1$

$$x_1 = z_1 - u_1\, x_2 = \tfrac{1}{2} - \left(-\tfrac{1}{2}\right) \cdot 1 = 1$$

Παρατηρούμε ο αλγόριθμος έδωσε την ακριβή λύση $x_1 = x_2 = x_3 = x_4 = 1$

Θεώρημα 5.2

Έστω A ο τριδιαγώνιος πίνακας του συστήματος (2.110) με $a_i\, c_i \neq 0$ για $i = 2(1)n - 1$. Αν $|b_1| > |c_1|,\ |b_i| \geq |a_i| + |c_i|$ για $i = 2(1)n - 1$ και $|b_n| > |a_n|$, τότε $detA \neq 0$ και i)$|u_i| < 1,$ ii)$|c_i| < |l_i| < |b_i| + |a_i|$.

Απόδειξη

Από τις (4.43) έχουμε

$$|u_1| = \frac{|c_1|}{|l_1|} = \frac{|c_1|}{|b_1|} < 1$$

Υποθέτοντας τώρα ότι $|u_{i-1}| < 1$ αρκεί να δειχθεί ότι $|u_i| < 1$. Πράγματι πάλι από τις (4.43) έχουμε

$$u_i = \frac{c_i}{l_i} = \frac{c_i}{b_i - a_i\, u_{i-1}}$$

ή

$$|u_i| \leq \frac{|c_i|}{||b_i| - |a_i||u_{i-1}||} < \frac{|c_i|}{|b_i| - |a_i|},\ |u_{i-1}| < 1$$

λόγω της υπόθεσης όμως

$$|u_i| < \frac{|c_i|}{|b_i| - |a_i|} < \frac{|c_i|}{|c_i|} = 1$$

Για την απόδειξη του ii) έχουμε από τις (4.43) ότι

$$|l_i| = |b_i - a_i\, u_{i-1}|.$$

Αλλά

$$|b_i - a_i\, u_{i-1}| \leq |b_i| + |a_i\, u_{i-1}| < |b_i| + |a_i|.$$

Επίσης

$$|b_i - a_i\, u_{i-1}| \geq ||b_i| - |a_i\, u_{i-1}|| > ||b_i| - |a_i|| \geq |c_i|.$$

Λόγω όμως των δύο παραπάνω σχέσεων ισχύει το (ii) του θεωρήματος. Επιπλέον

$$detA = [detL][detU] = \prod_{i=1}^{n} l_i$$

επειδή δε λόγω του (ii) $l_i \neq 0,\ i = 1(1)n$ έπεται ότι και $detA \neq 0$.■

2.5.3 Υπολογιστική πολυπλοκότης της LU μεθόδου

Για τον υπολογισμό του πλήθους των πράξεων της LU μεθόδου θα πρέπει να ανατρέξουμε στον αλγόριθμο της μεθόδου (βλ.§3.3.1). Έτσι αν υποθέσουμε ότι θεωρούμε την μέθοδο του Doolitle τότε

$$u_{1j} = a_{1j}$$

και

$$l_{j1} = a_{j1}/u_{11}, \qquad j = 2(1)n$$

οπότε για τον υπολογισμό των l_{j1} απαιτούνται $n-1$ διαιρέσεις.

Επίσης για τον υπολογισμό του

$$u_{rr} = a_{rr} - \sum_{j=1}^{r-1} l_{rj}\, u_{jr}$$

απαιτούνται $r-1$ πολ/μοί και $r-1$ προσθαφαιρέσεις ή συνολικά για όλα τα $u_{rr}, \quad r = 2(1)n$

$$\sum_{r=2}^{n} (r-1) \qquad \text{πολ/μοί}$$

και

$$\sum_{r=2}^{n} (r-1) \qquad \text{προσθαφαιρέσεις.}$$

Για τον υπολογισμό των

$$u_{rp} = a_{rp} - \sum_{j=1}^{r-1} l_{rj}\, u_{jp}, \quad p = r+1(1)n$$

απαιτούνται

$$(n-r)(r-1) \quad \text{πολ/μοί}$$

και

$$(n-r)(r-1) \qquad \text{προσθαφαιρέσεις.}$$

ή συνολικά

$$\sum_{r=2}^{n-1} (n-r)(r-1) \qquad \text{πολ/μοί}$$

και

$$\sum_{r=2}^{n-1} (n-r)(r-1) \qquad \text{προσθαφαιρέσεις.}$$

Ο υπολογισμός των

$$l_{pr} = \left(a_{pr} - \sum_{j=1}^{r-1} l_{pj}\, u_{jr} \right) / \, u_{rr}, \qquad p = r+1(1)n$$

απαιτεί τον ίδιο αριθμό πολ/μών και προσθαφαιρέσεων για τα u_{rp} και επιπλέον $(n-r)$ διαιρέσεις ή συνολικά

$$\sum_{r=2}^{n-1} (n-r) \qquad \text{διαιρέσεις}$$

Συνοψίζοντας η παραγοντοποίηση του A απαιτεί

$$2\sum_{r=2}^{n-1}(n-r)(r-1)+\sum_{r=2}^{n}(r-1) \qquad \text{πολ/μούς}$$

$$2\sum_{r=2}^{n-1}(n-r)(r-1)+\sum_{r=2}^{n}(r-1) \qquad \text{προσθαφαιρέσεις} \qquad (2.114)$$

και

$$\sum_{r=2}^{n-1}(n-r)+n-1 \qquad \text{διαιρέσεις}$$

Εύκολα βρίσκουμε ότι

$$\sum_{r=2}^{n}(r-1)=\frac{n^2}{2}-\frac{n}{2}$$

$$\sum_{r=2}^{n-1}(n-r)(r-1)=\frac{n^3}{6}-\frac{n^2}{2}+\frac{n}{3}$$

και

$$\sum_{r=2}^{n-1}(n-r)=\frac{n^2}{2}-\frac{3n}{2}+1$$

Άρα οι τύποι (2.114) γίνονται

$$\begin{aligned}&\frac{n^3}{3}-\frac{n^2}{2}+\frac{n}{6} && \text{πολ/μούς}\\ &\frac{n^3}{3}-\frac{n^2}{2}+\frac{n}{6} && \text{προσθαφαιρέσεις}\end{aligned} \qquad (2.115)$$

και

$$\frac{n^2}{2}-\frac{n}{2} \qquad \text{διαιρέσεις}$$

Τέλος για τον υπολογισμό της λύσης του $Ly=b$ που δίνεται από τους τύπους

$$y_1 = b_1$$

και

$$y_i = b_i - \sum_{j=1}^{i-1} l_{ij}\, y_j, \qquad i = 2(1)n$$

απαιτούνται

$$\sum_{i=2}^{n} (i-1) \qquad \text{πολ/μοί}$$

και

$$\sum_{i=2}^{n} (i-1) \qquad \text{προσθαφαιρέσεις}$$

ή

$$\frac{n^2}{n} - \frac{n}{2} \qquad \text{πολ/μοί}$$

και

$$\frac{n^2}{n} - \frac{n}{2} \qquad \text{προσθαφαιρέσεις} \tag{2.116}$$

Για τη λύση του $Ux = y$ που δίνεται από τους τύπους

$$\begin{aligned} x_n &= y_n / x_{nn} \\ x_i &= \left(y_i - \sum_{j=i+1}^{n} x_{ij}\, x_j \right) \,/\, u_{ii}, \qquad i = n-1(-1)1 \end{aligned}$$

απαιτούνται

$$\begin{aligned}
&\sum_{i=1}^{n-1}(n-i) && \text{πολ/μοί}\\
&\sum_{i=1}^{n-1}(n-i) && \text{προσθαφαιρέσεις}\\
&\qquad \text{και} &&\\
&\qquad n && \text{διαιρέσεις}
\end{aligned}$$

ή

$$\begin{aligned}
&\frac{n^2}{2}-\frac{n}{2} && \text{πολ/μοί}\\
&\frac{n^2}{2}-\frac{n}{2} && \text{προσθαφαιρέσεις}\\
&\qquad n && \text{διαιρέσεις}
\end{aligned} \tag{2.117}$$

Λαμβάνοντας υπόψη τις (2.115), 2.116) και (2.117) έχουμε ότι η επίλυση ενός συστήματος με την μέθοδο του Doolitle απαιτεί

$$\begin{aligned}
&\frac{n^3}{3}+\frac{n^2}{2}-\frac{5n}{6} && \text{πολ/μούς}\\
&\frac{n^3}{3}+\frac{n^2}{2}-\frac{5n}{6} && \text{προσθαφαιρέσεις}\\
&\qquad \text{και} &&\\
&\qquad \frac{n^2}{2}+\frac{n}{2} && \text{διαιρέσεις}
\end{aligned} \tag{2.118}$$

δηλαδή ακριβώς το ίδιο πλήθος και είδος πράξεωνμε τη μέθοδο της απαλοιφής του Gauss.

2.6 Norms διανυσμάτων και πινάκων

Για την ποσοτική ανάλυση του σφάλματος της λύσης του $Ax = b$ είναι αναγκαίο να έχουμε κάποιο μέτρο του μεγέθους του. Με άλλα λόγια επιθυμούμε να ορίσουμε ένα μη αρνητικό πραγματικό αριθμό (norm) για κάθε διάνυσμα όπως έχουμε την απόλυτη τιμή για κάθε αριθμό. Ας θεωρήσουμε πρώτα τις (n$orms$) διανύσματος. Μια (norm) διανύσματος $\|.\|$ είναι μια συνάρτηση με πραγματικές και μη αρνητικές τιμές στο διανυσματικό χώρο $\mathbb{C}^n$ με τις παρακάτω ιδιότητες:

$$\begin{aligned} &\text{i)} \|x\| > 0 \text{ αν } x \neq 0, \quad \|x\| = 0 \text{ αν } x = 0, \quad \text{για κάθε } x \in \mathbb{C}^n \\ &\text{ii)} \|cx\| = |c| \cdot |x| \quad \text{για κάθε } c \in \mathbb{C} \text{ και } x \in \mathbb{C}^n \\ &\text{iii)} \|x + y\| \leq \|x\| + \|y\| \text{ για } x, y, \in \mathbb{C}^n \text{ (τριγωνική ανισότης)} \end{aligned} \tag{2.119}$$

Θεώρημα 6.1

Για κάθε $x, y \in \mathbb{C}^n$ ισχύει

$$|\|x\| - \|y\|| \leq \|x - y\| \tag{2.120}$$

Απόδειξη

Από την ιδιότητα (iii) των n$orms$ έχουμε

$$\|x\| = \|(x - y) + y\| \leq \|x - y\| + \|y\|$$

ή

$$\|x\| - \|y\| \leq \|x - y\|$$

Όμοια λαμβάνουμε

$$\|y\| - \|x\| \leq \|y - x\|$$

λόγω όμως της ιδιότητας (ii) με $c = -1$ τα δεύτερα μέλη των δύο τελευταίων ανισοτήτων είναι ίσα, άρα η (4.45) ισχύει.■

Είναι φανερό ότι υπάρχουν άπειρες διανυσματικές n$orms$ από τις οποίες θα θεωρήσουμε μόνο τις $l_p -$ norms (Hölder norms) που δίδονται από τον παρακάτω γενικό τύπο

$$\|x\|_p = \begin{cases} \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, & p = 1, 2, 3, \ldots \\ \max_i |x_i|, & p = \infty \end{cases} \tag{2.121}$$

Οι περισσότερο χρησιμοποιούμενες n*orms* είναι οι l_1, l_2 και $l_\infty - norm$. Συνεπώς θα περιοριστούμε μόνο στις ακόλουθες n*orms*

$$\begin{aligned} \|x\|_1 &= \sum_{i=1}^{n} |x_i| \qquad \text{αθροιστική norm} \\ \|x\|_2 &= \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2} \qquad \text{Ευκλείδια norm} \\ \|x\|_\infty &= \max_i |x_i| \qquad \text{Μεγίστη norm} \end{aligned} \tag{2.122}$$

Θεώρημα 6.2

i)$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$
ii)$\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty$
iii)$\|x\|_\infty \leq \|x\|_2 \leq n^{1/2}\|x\|_\infty$
iv)$n^{-1/2}\|x\|_1 \leq \|x\|_2 \leq \|x\|_1$

Απόδειξη

Αφήνεται σαν άσκηση για τον αναγνώστη

Θεώρημα 6.3

Να αποδειχθεί ότι οι $l_p -$ norms για $p = 1, 2, \infty$ είναι norms διανυσμάτων.

Απόδειξη

Αφήνεται σαν άσκηση για τον αναγνώστη. Υπόδειξη: για $p = 2$ να χρησιμοποιηθεί η ανισότητα $\left[\sum_{i=1}^{n} x_i y_i\right]^2 \leq \sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2$.

Είναι αναγκαίο όμως να έχουμε επίσης και ένα μέτρο του μεγέθους ενός πίνακα. Έστω λοιπόν $\mathbb{C}^{nn}$ το σύνολο των τετραγωνικών $n \times n$ πινάκων, τότε μια norm πίνακα είναι μια συνάρτηση με πραγματικές τιμές $\|\cdot\|$ ορισμένη στο $\mathbb{C}^{nn}$ που έχει τις ιδιότητες:

$$\begin{aligned} &\text{i)} \|A\| > 0, \text{ εκτός αν } A = 0 \text{ οπότε } \|A\| = 0 \\ &\text{ii)} \|cA\| = |c| \, \|A\| \text{ όπου } c \text{ μαθματώτο μέγεθος} \\ &\text{iii)} \|A + B\| \leq \|A\| + \|B\| \\ &\text{iv)} \|AB\| \leq \|A\| \, \|B\| \end{aligned} \tag{2.123}$$

Οι norms πινάκων που αντιστοιχούν στις διανυσματικές norms ορίζονται από τον τύπο

$$\|A\| = \max_{\|x\|=1} \{\|Ax\|\}$$

από τον οποίο προκύπτει η σχέση

$$\|Ax\| \leq \|A\| \, \|x\| \tag{2.124}$$

Θεώρημα 6.4

Οι τρεις norms πινάκων που αντιστοιχούν στις διανυσματικές norms δίνονται από τους τύπους

$$\begin{aligned} \|A\|_1 &= \max_j \sum_{i=1}^{n} |a_{ij}| \\ \|A\|_2 &= [S(A^H A)]^{1/2} \end{aligned} \tag{2.125}$$

και

$$\|A\|_\infty = \max_i \sum_{j=1}^{n} |a_{ij}|$$

όπου $S(A) = \max_{1\le i\le n} |\lambda_i|$ με λ_i ιδιότητες του A και A^H είναι ο συζυγής ανάστροφος του A.

Απόδειξη

Αφήνεται σαν άσκηση για τον αναγνώστη.

Τόσο οι διανυσματικές όσο και οι n$orms$ πινάκων συνδέονται άμεσα με την σύγκλιση ακολουθιών διανυσμάτων και πινάκων. Μια ακολουθία διανυσμάτων $\{X^{(\kappa)}\}, \quad \kappa = 0, 1, 2, \ldots$ με συνιστώσες $\{X_i^{(\kappa)}\}, \quad i = 1(1)n$ έχει όριο το μηδενικό διάνυσμα ή συγκλίνει στο μηδενικό διάνυσμα τότε και μόνον τότε αν οι ακολουθίες $\{X_i^{(\kappa)}\},\ \kappa = 0, 1, 2, \ldots$ για $i = 1(1)n$ έχουν όριο το μηδέν, γράφουμε δε τότε

$$\lim_{k\to\infty} X^{(\kappa)} = 0 \qquad \text{ή } X^{(\kappa)} \underset{\kappa\to\infty}{\to} 0$$

Θεώρημα 6.5

Ικανή και αναγκαία συνθήκη για να ισχύει η

$$\lim_{k\to\infty} x^{(\kappa)} = x$$

είναι η

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

Απόδειξη

Αφήνεται σαν άσκηση για τον αναγνώστη.

Όμοια μια ακολουθία πινάκων $\{A^{(\kappa)}\},\ \kappa = 0, 1, 2, \ldots$ στοιχείων $\{a_{ij}^{(\kappa)}\},\ i, j = 1(1)n$ έχει όριο το μηδενικό πίνακα ή συγκλίνει στο μηδενικό πίνακα τότε και μόνον τότε αν οι n^2 ακολουθίες $\{a_{ij}^{(\kappa)},\ \kappa = 0, 1, 2, \ldots$ για $i, j = 1(1)n$ έχουν όριο το μηδέν, γράφουμε δε τότε

$$\lim_{k\to\infty} A^{(\kappa)} = 0 \quad \text{ή} \quad A^{(\kappa)} \underset{\kappa\to\infty}{\to} 0.$$

Θεώρημα 6.6

Ικανή και αναγκαία συνθήκη για να ισχύει η

$$\lim_{k\to\infty} A^{(\kappa)} = A$$

είναι η

$$\lim_{k\to\infty} \|A^{(\kappa)} - A\| = 0$$

Απόδειξη

Αφήνεται σαν άσκηση

Θεώρημα 6.7

Ικανή και αναγκαία συνθήκη για να συγκλίνει η ακολουθία των διαδοχικών δυνάμεων ενός πίνακα A τάξης n, στο μηδενικό πίνακα είναι η

$$S(A) < 1 \tag{2.126}$$

Απόδειξη (βλ. [Golub and Van Loan])

Θεώρημα 6.8

Για κάθε πίνακα τάξης n ισχύει

$$\|A^\kappa\| \leq \|A\|^\kappa, \quad \kappa = 0, 1, 2, \ldots$$

Απόδειξη

Για $\kappa = 0,\ 1$ ισχύει η ισότητα. Για $\kappa > 1$ έχουμε

$$\|A^\kappa\| = \|A^{\kappa-1}A\| \leq \|A^{\kappa-1}\|\,\|A\| = \|A^{\kappa-2}A\|\,\|A\| \leq$$
$$\leq \|A^{\kappa-2}\|\,\|A\|^2 \leq \cdots \leq \|A\|^\kappa$$

Θεώρημα 6.9

Ικανή συνθήκη για τη σύγκλιση της ακολουθίας των διαδοχικών δυνάμεων ενός πίνακα A τάξης n στο μηδενικό πίνακα είναι η

$$\|A\| < 1$$

Απόδειξη

Από το Θεώρημα (6.6) έχουμε ότι $\lim\limits_{\kappa\to 0} A^\kappa = 0$ αν $\lim\limits_{\kappa\to\infty}\|A^\kappa\| = 0$ αλλά λόγω του θεωρήματος 6.9 αρκεί $\lim\limits_{\kappa\to\infty}\|A\|^\kappa = 0$ η οποία για να ισχύει θα πρέπει $\|A\| < 1$.■

Θεώρημα 6.10

Για κάθε πίνακα A τάξης n ισχύει

$$S(A) \leq \|A\| \tag{2.127}$$

Απόδειξη

Αν λ είναι μια ιδιοτιμή του A και x το αντίστοιχο ιδιοδιάνυσμα τότε

$$Ax = \lambda x$$

οπότε λαμβάνοντας τις norms των δύο μελών έχουμε

$$\|Ax\| = \|\lambda x\|$$

ή εφαρμόζοντας ιδιότητες των norms

$$|\lambda| \, \|x\| \, \leq \|A\| \, \|x\|$$

ή

$$|\lambda| \leq \|A\|$$

συνεπώς

$$S(A) = \max |\lambda| \, \leq \|A\|.\blacksquare$$

Θεώρημα 6.11

Η σειρά

$$\sum_{m=0}^{\infty} A^m = I + A + A^2 + \cdots + A^m + \ldots$$

συγκλίνει τότε και μόνον τότε αν $\lim\limits_{m\to\infty} A^m = 0$. Στη περίπτωση σύγκλισης της σειράς έχουμε

$$\sum_{m=0}^{\infty} A^m = (I - A)^{-1}.$$

Απόδειξη

(i) Ας υποθέσουμε ότι $\lim\limits_{m\to\infty} A^m = 0$ τότε από το Θεώρημα 6.7 έχουμε ότι $S(A) < 1$ και $\lambda = 1$ δεν είναι ιδιοτιμή του A. Αυτό σημαίνει ότι $det(I - A) \neq 0$ και ο $(I - A)^{-1}$ υπάρχει. Θεωρούμε στη συνέχεια την ταυτότητα

$$(I - A)(I + A + A^2 + \cdots + A^m) \equiv I - A^{m+1}$$

από την οποία έχουμε

$$I + A + A^2 + \cdots + A^m \equiv (I - A)^{-1} - (I - A)^{-1}A^{m+1}.$$

Ο δεύτερος όρος του δεύτερου μέλους τείνει στο μηδενικό πίνακα άρα

$$I + A + A^2 + \cdots + A^m \underset{m\to\infty}{\to} (I - A)^{-1}.$$

(ii) Ας υποθέσουμε ότι

$$S^{(m)} = I + A + A^2 + \cdots + A^m \underset{m\to\infty}{\to} (I - A)^{-1}$$

πράγμα που σημαίνει ότι

$$S_{ij}^{(m)} \to (I - A)_{ij}^{-1}, \quad m \to \infty$$

δηλαδή οι n^2 άπειρες σειρές $S_{ij}^{(m)}$ πραγματικών ή μιγαδικών αριθμών συγκλίνουν. Μια αναγκαία συνθήκη για τη σύγκλιση μιας τέτοιας σειράς είναι το m-ιοστό στοιχείο στο άθροισμα να τείνει στο μηδέν. Συνεπώς για όλα τα i και j έχουμε $(A^m)_{ij} \to 0$ για $m \to \infty$ που σημαίνει $\lim\limits_{m\to\infty} A^m = 0$.

Πόρισμα 6.12

$$\sum_{m=0}^{\infty} A^m = (I - A)^{-1} \quad \text{τότε και μόνον τότε αν}$$

$$S(A) < 1$$

Απόδειξη

Το αποτέλεσμα είναι άμεση συνέπεια των θεωρημάτων 6.7 και 6.11.■

Πόρισμα 6.13

Αν υπάρχει τουλάχιστον μια norm πίνακα για την οποία $\|A\| < 1$ τότε

$$\sum_{m=0}^{\infty} A^m = (I - A)^{-1}$$

Απόδειξη

Το αποτέλεσμα είναι άμεση συνέπεια των θεωρημάτων 6.9 και 6.11.■

Θεώρημα 6.14

Αν $\|A\| < 1$ για κάποια norm πίνακα τότε οι πίνακες $I-A$ και $I+A$ είναι μη ιδιάζοντες και ισχύουν οι σχέσεις

$$\frac{1}{1 + \|A\|} \leq \|(I - A)^{-1}\| \leq \frac{1}{1 - \|A\|} \tag{2.128}$$

και

$$\frac{1}{1 + \|A\|} \leq \|(I + A)^{-1}\| \leq \frac{1}{1 - \|A\|} \tag{2.129}$$

Απόδειξη

Από το Θεώρημα 6.10 έχουμε $S(A) \leq \|A\| < 1$ άρα η $\lambda = \pm 1$ δεν είναι ιδιοτιμή του A συνεπώς $det(I \pm A) \neq 0$ και οι πίνακες $I \pm A$ είναι μη ιδιάζοντες. Για την απόδειξη της (2.128) ξεκινούμε από τη γνωστή σχέση

$$I = (I - A)(I - A)^{-1}.$$

Λαμβάνοντας τις norms των δύο μελών έχουμε διαδοχικά

$$1 = \|I\| = \|(I-A)(I-A)^{-1}\| \leq \|I-A\| \, \|(I-A)^{-1}\| \leq$$
$$\leq (\|I\| + \|-A\|) \, \|(I-A)^{-1}\| = (1 + \|A\|) \, \|(I-A)^{-1}\|$$

Από το πρώτο και τελευταίο μέλος των ανωτέρω σχέσεων προκύπτει η αριστερή σχέση των (2.128). Επίσης

$$(I-A)^{-1} = (I-A+A)(I-A)^{-1} = I + A(I-A)^{-1}$$

και λαμβάνοντας τις norms έχουμε διαδοχικά

$$\|(I-A)^{-1}\| = \|I + A(I-A)^{-1}\| \leq \|I\| + \|A(I-A)^{-1}\| \leq$$
$$\leq 1 + \|A\| \, \|(I-A)^{-1}\|$$

Από το πρώτο και τελευταίο μέλος λαμβάνουμε, έχοντας υπόψη ότι $\|A\| < 1$, τη δεξιά σχέση των (2.128). Αν τώρα στις (2.128) θέσουμε όπου A το $-A$ τότε λαμβάνουμε τις (2.129).■

2.7 Ασταθή συστήματα

Όπως αναφέρθηκε προηγούμενα, με τη μερική οδήγηση αποφεύγουμε το πρόβλημα της συσσώρευσης σφαλμάτων στρογγύλευσης. Υπάρχει όμως ακόμη το πρόβλημα των σφαλμάτων στην περίπτωση που το σύστημα

$$Ax = b$$

είναι ασταθές, δηλαδή όταν η λύση του επηρεάζεται δραστικά από μικρές διαταραχές στα στοιχεία του επαυξημένου πίνακα

$$B = [A, b].$$

Αν ένα γραμμικό σύστημα είναι ασταθές, τότε είναι αναπόφευκτη μια προοδευτική απώλεια σημαντικών ψηφίων κατά την διάρκεια των υπολογισμών με αποτέλεσμα την απώλεια της ακρίβειας της λύσης. Μπορούμε όμως στην περίπτωση αυτή να δεχτούμε ότι το αποτέλεσμα που υπολογίσαμε είναι η ακριβής λύση ενός ελαφρά διαταραγμένου συστήματος. Έτσι λοιπόν υποθέτουμε ότι στο σύστημα που δίνεται υπάρχουν αρχικά σφάλματα τόσο στον πίνακα A όσο και στο διάνυσμα b. Έστω δA και δb οι διαταράξεις στον A και b, αντίστοιχα. Τότε, με την προϋπόθεση ότι δεν εισχωρούν νέα σφάλματα κατά την επίλυση, αντί για την ακριβή τιμή του διανύσματος x, θα βρίσκουμε ένα διάνυσμα που θα περιέχει μία διατάραξη δx. Έτσι θα έχουμε το διαταραγμένο σύστημα

$$(A + \delta A)(x + \delta x) = b + \delta b \tag{2.130}$$

και είναι δυνατόν να βρούμε το ακόλουθο φράγμα για την σχετική αλλαγή στη λύση.

Θεώρημα 7.1

Έστω ο μη ιδιάζων πίνακας A με

$$\|A^{-1}\| \, \|\delta A\| < 1 \tag{2.131}$$

τότε

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\kappa}{1 - \|\delta A\| \, \|A^{-1}\|} \left[\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right] \tag{2.132}$$

όπου

$$\kappa = \kappa(A) = \|A^{-1}\| \, \|A\| \tag{2.133}$$

Απόδειξη

Από την (4.49) έχουμε

$$(A + \delta A) \, \delta x = \delta b - \delta A x$$

Για να λύσουμε ως προς δx θα πρέπει να αποδείξουμε πρώτα ότι ο πίνακας $A + \delta A = A(I + A^{-1} \, \delta A)$ είναι μη ιδιάζων. Έχουμε όμως ότι

$$S(A^{-1}\delta A) \leq \|A^{-1}\delta A\| < 1$$

άρα ο $I + A^{-1}\delta A$ είναι μη ιδιάζων. Συνεπώς

$$\delta x = (I + A^{-1}\delta A)^{-1} A^{-1}(\delta b - \delta A x)$$

και λαμβάνοντας τις norms των δύο μελών έχουμε

$$\begin{aligned} \|\delta x\| &= \|(I + A^{-1}\delta A)^{-1} A^{-1}(\delta b - \delta A x)\| \\ &\leq \|(I + A^{-1}\delta A)^{-1}\| \, \|A^{-1}\| \, \|(\delta b - \delta A x)\| \end{aligned}$$

Από το Θεώρημα 6.14 όμως, επειδή $\|A^{-1}\delta A\| \leq \|A^{-1}\| \, \|\delta A\| < 1$, έχουμε ότι

$$\|(I + A^{-1}\delta A)^{-1}\| \leq \frac{1}{1 - \|A^{-1}\delta A\|}$$
$$\leq \frac{1}{1 - \|A^{-1}\| \, \|\delta A\|}$$

Άρα

$$\|\delta x\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \, \|\delta A\|}[\|\delta b\| + \|\delta A\| \, \|x\|]$$

ή

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|A\| \, \|A^{-1}\|}{1 - \|A^{-1}\| \, \|\delta A\|} \left[\frac{\|\delta b\|}{\|A\| \, \|x\|} + \frac{\|\delta A\|}{\|A\|}\right].$$

Αλλά $Ax = b$ και $\|b\| \leq \|A\| \, \|x\|$ άρα

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|A\| \, \|A^{-1}\|}{1 - \|A^{-1}\| \, \|\delta A\|} \left[\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|}\right]. \blacksquare$$

Πόρισμα 7.2
Αν $\delta A = 0$ τότε

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\| \, \|A^{-1}\| \frac{\|\delta b\|}{\|b\|}$$

Πόρισμα 7.3
Αν $\delta b = 0$ τότε

$$\frac{\|\delta x\|}{\|x\|} = \frac{\|A\| \, \|A^{-1}\|}{1 - \|A^{-1}\| \, \|\delta A\|} \cdot \frac{\|\delta A\|}{\|A\|}.$$

Παρατηρήσεις

(i) Αν η διαταραχή δA είναι πολύ μικρή, τότε από το Πόρισμα 7.2 (όπου $\delta A = 0$), η σχετική αλλαγή στη λύση είναι φραγμένη από την ποσότητα $\kappa(A) = \|A^{-1}\| \, \|A\|$.

(ii) Αν $\kappa(A)$ είναι μικρό, τότε μια μικρή διαταραχή του A ή μια μικρή διαταραχή του b ή μικρές διαταραχές των A και b δεν επιτρέπουν μεγάλες αλλαγές στη λύση x.

(iii) $\kappa = \|A^{-1}\| \, \|A\| \geq \|A^{-1} \, A\| = \|I\| = 1$

Ορισμός

Αν ο A είναι μη ιδιάζων, τότε

$$\kappa(A) = \|A\| \cdot \|A^{-1}\| \tag{2.134}$$

είναι ο αριθμός συνθήκης για το σύστημα $Ax = b$

Αν $\kappa(A)$ είναι ένας μεγάλος αριθμός, τότε μικρές διαταραχές του A ή b είναι δυνατόν να προκαλέσουν μεγάλες διαταραχές στη λύση x του συστήματος. Σε αυτή την περίπτωση λέμε ότι το σύστημα είναι ασταθές (ill-conditioned).

Θεώρημα 7.4

Αν ο A είναι πραγματικός και συμμετρικός, τότε

$$\kappa(A) = \left| \frac{\lambda_1}{\lambda_n} \right| \tag{2.135}$$

όπου λ_1 και λ_n είναι η μεγαλύτερη και η μικρότερη κατά απόλυτο τιμή ιδιοτιμές του A, αντίστοιχα.

Απόδειξη

Αφήνεται σαν άσκηση.

Κεφάλαιο 3

Επαναληπτικές Μέθοδοι για την Επίλυση Γραμμικών Συστημάτων

3.1 Γενικά

Στο παρόν κεφάλαιο θα ασχοληθούμε με τις επαναληπτικές μεθόδους για τη λύση του γραμμικού συστήματος

$$Au = b \tag{3.1}$$

όπου ο A είναι ένας παραγματικός ν × ν πίνακας και u, b είναι δανύσματα n τάξης. Οι επαναπληπτικές μέθοδοι χρησιμοποιούνται όταν ο πίνακας A του συστήματος είναι μεγάλης τάξης ($10^3 - 10^6$), αραιός και κάποιας συγκεκριμένης δομής. Συστήματα αυτού του τύπου προκύπτουν από την αριθμητική επίλυση μερικών διαφορικών εξισώσεων χρησιμοποιώντας τη μέθοδο των πεπερασμένων διαφορών ή εκείνη των πεπερασμένων στοιχείων. Γενικά, οι επαναληπτικές μέθοδοι χρειάζονται λιγότερη μνήμη και αριθμητικές πράξεις από τις αμέσους μεθόδους για αρκετά μεγάλα συστήματα. Στη συνέχεια θα ασχοληθούμε με τις γραμμικές στατικές επαναληπτικές μεθόδους πρώτου βαθμού. Η μορφή των μεθόδων αυτών είναι

$$u^{(n+1)} = Gu^{(n)} + k, \quad n = 0, 1, 2, \ldots \tag{3.2}$$

όπου G είναι ο επαναληπτικός πίνακας και k ένα σταθερό διάνυσμα. Επιπλέον, η επαναληπτική μέθοδος (3.100) θα πρέπει να ικανοποιεί τις παρακάτω απαιτήσεις:

1. Αν σε κάποια επανάληψη βρεθεί η λύση του (3.99), τότε οι επόμενες επαναλήψεις να παραμείνουν αμετάβλητες (συμβατότητα).

2. Αν η ακολουθία των διανυσμάτων που ορίζεται απο την (3.100) συγκλίνει, τότε συγκλίνει στη λύση του (3.99)(αμοιβαία συμβατότητα).

Στη συνέχεια παρουσιάζονται οι συνθήκες κάτω από τις οποίες η (3.100) παράγει μια συγκλίνουσα ακολουθία διανυσμάτων για κάποιο αυθαίρετο αρχικό διάνυσμα $x^{(0)}$.

Θεώρημα 3.1.1. *Η επαναληπτική μέθοδος (3.100) συγκλίνει αν και μόνον αν*

$$S(G) < 1. \tag{3.3}$$

Απόδειξη. Ας υποθέσουμε ότι u είναι το όριο της ακολουθίας $u^{(n)}, n = 0, 1, 2, \ldots$ Έστω επιπλέον ότι $\epsilon^{(n)} = u^{(n)} - u$, τότε

$$\epsilon^{(n+1)} = G\epsilon^{(n)} \tag{3.4}$$

αφού το u ικανοποιεί την

$$u = Gu + k. \tag{3.5}$$

Από την (3.102) εύκολα προκύπτει ότι

$$\epsilon^{(n)} = G^n\epsilon^{(0)}. \tag{3.6}$$

Άρα $\lim_{n\to\infty} u^{(n)} = u$ αν και μόνο αν $\lim_{n\to\infty} \epsilon^{(n)} = 0$ ή λόγω της (3.104) αν $\lim_{n\to\infty}(G^{(n)}\epsilon^{(0)}) = 0$ για κάθε αυθαίρετο $\epsilon^{(0)}$. Συνεπώς από το θεώρημα 3.6.7 έχουμε ότι ικανή και αναγκαία συνθήκη για να ισχύει $\lim_{n\to\infty} G^n = 0$ είναι η (3.101).

Αν τώρα υποθέσουμε ότι $S(G) < 1$, τότε ο $I - G$ είναι μη ιδιάζων και το σύστημα $(I - G)u = k$ έχει μία και μοναδική λύση. Αν όμως $S(G) < 1$ τότε $\lim_{n\to\infty} G^n = 0$ ή $\lim_{n\to\infty} \| G^n \| = 0$. Επειδή $\| G^n\epsilon^{(0)} \| \leq \| G^n \| \| \epsilon^{(0)} \|$ συνεπάγεται ότι $\lim_{n\to\infty} G^{(n)}\epsilon^{(0)} = 0$ οπότε από την (3.104) προκύπτει ότι $\lim_{n\to\infty} \epsilon^{(n)} = 0$ ή $\lim_{n\to\infty} u^{(n)} = u$, δηλαδή ότι η (3.100) συγκλίνει. ■

Επειδή η εύρεση της φασματικής ακτίνας του πίνακα G είναι επίπονη εργασία για αυτό στην πράξη εξετάζουμε καταρχήν αν ισχύει η ικανή συνθήκη

$$\| G \|_\alpha \leq 1 \tag{3.7}$$

με $\alpha = 1$ ή ∞ οπότε λόγω της $S(G) \leq \| G \|$ θα ισχύει η ικανή και συνθήκη για τη σύγκλιση. Εφόσον η επαναληπτική μέθοδος (3.100) συγκλίνει, τότε εκλέγουμε ένα αυθαίρετο αρχικό διάνυσμα $u^{(0)}$(συνήθως εκλέγεται το μηδενικό διάνυσμα) και εφαρμόζουμε την (3.100) για $n = 0, 1, 2, \ldots$. Υπάρχουν πολλά κριτήρια διακοπής της σύγκλισης της ανωτέρω ακολουθίας, δύο από τα πιο απλά είναι:

$$\| u^{(n+1)} - u^{(n)} \| \leq \epsilon$$

και

$$\frac{\| u^{(n+1)} - u^{(n)} \|_\alpha}{\| u^{(n+1)} \|_\alpha} \leq \epsilon$$

όπου $\alpha = 1$ ή 2 ή ∞. Στην πράξη εκτός από την εξασφάλιση της σύγκλισης της (3.100), μας ενδιαφέρει η ταχύτητα με την οποία συγκλίνει η μέθοδος που χρησιμοποιούμε. Με άλλα λόγια επιθυμούμε να μελετήσουμε την ταχύτητα με την οποία $\epsilon^{(0)} \to 0$ για $n \to \infty$. Από την (3.104) έχουμε ότι αν $u^{(0)} \neq u$

$$\| \epsilon^{(n)} \| \; / \; \| \epsilon^{(0)} \| \leq \| G^n \| \; . \tag{3.8}$$

Έτσι η $\| G^n \|$ δίνει το μέγεθος με το οποίο η $norm$ του σφάλματος έχει ελαττωθεί σε ένα κλάσμα έστω ϱ της $\| \epsilon^{(0)} \|$. Η ελάττωση αυτή μπορεί να επιτευχθεί αν διαλέξουμε το n έτσι ώστε

$$\| G^n \| \leq \varrho. \tag{3.9}$$

Για όλα λοιπόν τα αρκετά μεγάλα n ώστε

$$\| G^n \| \leq 1$$

η παραπάνω ανισότητα είναι ισοδύναμη με την

$$n \geq \lceil -\log \varrho / \left(-\frac{1}{n} \log \|G^n\| \right) \rceil \tag{3.10}$$

όπου $\lceil \xi \rceil$ συμβολίζει τον ελάχιστο ακέραιο μεγαλύτερο του ξ. Η (3.108) δίνει τον ελάχιστο αριθμό επαναλήψεων για την σύγκλιση της (3.100). Παρατηρούμε δε ότι ο αριθμός αυτός είναι αντιστρόφως ανάλογος προς την ποσότητα $(-\frac{1}{n} log \|G^n\|)$. Έτσι οδηγούμαστε στον ορισμό της μέσης ταχύτητας σύγκλισης που είναι η ποσότητα

$$R_n(G) = -\frac{1}{n} log \; \| \; G^{(n)} \| \; . \tag{3.11}$$

Επίσης ορίζουμε σαν ασυμπτωτική ταχύτητα σύγκλισης ή ταχύτητα σύγκλισης, την ποσότητα

$$R(G) = \lim_{n\to\infty} R_n(G) = -logS(G) \tag{3.12}$$

καθόσον μπορεί να αποδειχθεί ότι (βλ. [])

$$S(G) = \lim_{n\to\infty} (\| G^n \|^{\frac{1}{n}}).$$

Για να έχουμε μια (όχι και τόσο καλή) προσέγγιση του αριθμού των επαναλήψεων που χρειάζεται η (3.100) για να συγκλίνει χρησιμοποιούμε τον τύπο

$$n \simeq \frac{-log\rho}{R(G)}. \tag{3.13}$$

Από τον ανωτέρω τύπο και την (3.110) συμπεραίνουμε ότι όσο μικρότερη είναι η φασματική ακτίνα του επαναληπτικού πίνακα G τόσο ταχύτερα θα συγκλίνει ασυμπτωτικά η επαναληπτική μέθοδος. Ωστόσο για να εκτιμήσουμε την αποτελεσματικότητα μιας επαναληπτικής μεθόδου θα πρέπει να λαβουμε υπόψη τόσο την ταχύτητα σύγκλισής της όσο και την υπολογιστική πολυπλοκότητα που απαιτεί η κάθε επανάληψη.

3.2 Βασικές επαναληπτικές μέθοδοι

Στη συνέχεια θα παράγουμε όλες τις γνωστές επαναληπτικές μεθόδους της μορφής (3.100). Παρατηρούμε ότι αν το σύστημα (3.99) το πολ/σουμε από αριστερά με τον A^{-1}, τότε βρίσκουμε αμέσως τη λύση του. Επειδή όμως δεν είναι δυνατόν να υπολογισθεί ο A^{-1} για πίνακες μεγάλης τάξης, για αυτό ας θεωρήσουμε ότι το σύστημα (3.99) πολ/ζεται από αριστερά με έναν πίνακα R^{-1}, οπότε έχουμε

$$R^{-1}Au = R^{-1}b \tag{3.14}$$

όπου απαιτούμε ο R να είναι ένας μη ιδιάζων πίνακας και ο αντίστροφος του να υπολογίζεται εύκολα (με άλλα λόγια να είναι 'εύκολο' να λυθεί το σύστημα $Rs = t$). Στη συνέχεια από την (4.1) μπορούμε να ορίσουμε την επαναληπτική μέθοδο

$$u^{(n+1)} = u^{(n)} + \tau R^{-1}(b - Au^{(n)}), \quad n = 0, 1, \ldots \tag{3.15}$$

όπου $\tau \neq 0$ είναι μια πραγματική παράμετρος της οποίας ο ρόλος είναι να ελαχιστοποιήσει τη φασματική ακτίνα του επαναληπτικού πίνακα της (4.2). Για την εύρεση του επαναληπτικού πίνακα γράφουμε την (4.2) υπό τη μορφή

$$u^{(n+1)} = G_\tau u^{(n)} + k_\tau, \tag{3.16}$$

όπου

$$G_\tau = I - \tau R^{-1}A \quad \kappa\alpha\iota \quad k_\tau = \tau R^{-1}b. \tag{3.17}$$

Παρατηρούμε ότι αν $\tau = 1$, τότε οι (4.3) και (4.4) δίνουν

$$u^{(n+1)} = Gu^{(n)} + k \tag{3.18}$$

όπου

$$G = I - R^{-1}A \quad \kappa\alpha\iota \quad k = R^{-1}b. \tag{3.19}$$

Θεώρημα 3.2.1. *Αν οι ιδιοτιμές $r_i, i = 1(1)\nu$ του πίνακα $R^{-1}A$ είναι πραγματικές το επαναληπτικό σχήμα (2.2) συγκλίνει αν και μόνο αν*

$$r_1 > 0 \quad \kappa\alpha\iota \quad 0 < \tau < 2/r_\nu \tag{3.20}$$

$$r_\nu < 0 \quad \kappa\alpha\iota \quad 2/r_1 < \tau < 0. \tag{3.21}$$

Απόδειξη. Λόγω της (4.4) οι ιδιοτιμές $\lambda_i, i = 1(1)\nu$ του G_τ και εκείνες του $R^{-1}A$ ικανοποιούν τη σχέση

$$\lambda_i = 1 - \tau r_i, \quad i = 1(1)\nu. \tag{3.22}$$

Ικανή και αναγκαία συνθήκη για τη σύγκλιση της (4.3) είναι η

$$S(G_\tau) < 1 \tag{3.23}$$

ή

$$\max_{1\leq i\leq \nu} |\lambda_i| = \max_{1\leq i\leq \nu} |1 - \tau r_i| < 1 \tag{3.24}$$

από την οποία εύκολα προκύπτουν οι (4.7) και (4.8). ■

Στο σημείο αυτό παρατηρούμε ότι ενώ το επαναληπτικό σχήμα (4.3) συγκλίνει αν ισχύουν μία από τις (4.7), (4.8), η βασική μέθοδος (4.5) μπορεί να αποκλίνει καθόσον είναι δυνατό $S(G) > 1$.

Θεώρημα 3.2.2. *Αν οι ιδιοτιμές του πίνακα $R^{-1}A$ είναι πραγματικές και το επαναληπτικό σχήμα (4.3) συγκλίνει, τότε για*

$$\tau = \tau_0 = \frac{2}{r_1 + r_\nu} \tag{3.25}$$

η $S(G_\tau)$ γίνεται ελάχιστη και η αντίστοιχη τιμή της δίνεται από τον τύπο

$$S(G_{\tau_0}) = \frac{|1 - k(R^{-1}A)|}{1 + k(R^{-1}A)} \tag{3.26}$$

όπου $k(R^{-1}A) = r_\nu / r_1$.

Απόδειξη. Αφήνεται σαν άσκηση. ■

Από την (4.13) παρατηρούμε ότι ο πίνακας R θα πρέπει να εκλεγεί τέτοιος ώστε $k(R^{-1}A) \leq k(A)$ καθόσον η ποσότητα $S(G_{\tau_0})$ είναι μια αύξουσα συνάρτηση του $k(R^{-1}A)$ όταν $r_1 > 0$ (ανάλογη παρατήρηση ισχύει αν $r_\nu < 0$). Με άλλα λόγια για επιπλέον ελαχιστοποίηση της $S(G_{\tau_0})$ θα πρέπει να ελαχιστοποιηθεί η ποσότητα $k(R^{-1}A)$.

Θεώρημα 3.2.3. *Αν οι ιδιοτιμές $r_i, i = 1(1)\nu$ του πίνακα $R^{-1}A$ είναι πραγματίκες το επαναληπτικό σχήμα (4.5) συγκλίνει αν και μόνο αν*

$$0 < r_1 \quad \kappa\alpha\iota \quad r_\nu < 2. \tag{3.27}$$

Επιπλέον,

$$S(G) = \begin{cases} 1 - r_1, & \alpha\nu \quad r_\nu \leq 1 \\ r_\nu - 1, & \alpha\nu \quad r_1 \geq 1. \end{cases} \tag{3.28}$$

Απόδειξη. Προκύπτει εύκολα από εφαρμογή του Θεωρήματος 1.2.1.

Παρατήρηση. Για τη σύγκλιση της (4.5) θα πρέπει να ισχύει η επιπλέον συνθήκη $r_\nu < 2$ σε σχέση με τη σύγκλιση της (4.3). Επίσης είναι εύκολο να αποδειχθεί ότι

$$S(G_{\tau_0}) \leq S(G) \tag{3.29}$$

πράγμα που σημαίνει ότι η (4.3) θα έχει μεγαλύτερη ταχύτητα σύγκλισης από την (4.5) γιαυτό και θα αναφέρεται σαν η επιταχυντική μορφή της (4.5). Στη συνέχεια ο πίνακας R θα λάβει διάφορες μορφές και θα σχηματίσουμε από την (4.3) τις αντίστοιχες επαναληπτικές μεθόδους. Ο πίνακας A αναλύεται σαν

$$A = D - C_L - C_U \tag{3.30}$$

όπου ο D είναι ένας διαγώνιος πίνακας του οποίου τα στοιχεία είναι τα ίδια με τα διαγώνια στοιχεία του A και οι πίνακες C_L, C_U είναι τα αυστηρά κάτω και άνω τριγωνικά μέρη του A, αντίστοιχα. Τότε αν διαλέξουμε τον R έτσι ώστε

$$R = D \tag{3.31}$$

η (4.2) δίνει

$$u^{(n+1)} = u^{(n)} + \tau D^{-1}(b - Au^{(n)}), \quad n = 0, 1, 2, \ldots$$

ή

$$u^{(n+1)} = B_\tau u^{(n)} + \tau c, \tag{3.32}$$

όπου

$$B_\tau = I - \tau D^{-1}A \quad \kappa\alpha\iota \quad c = D^{-1}b. \tag{3.33}$$

Αναλύοντας περισσότερο την (4.19) λαμβάνουμε

$$u^{(n+1)} = [(1-\tau)I + \tau B]u^{(n)} + \tau c, \tag{3.34}$$

όπου

$$B = L + U, \quad L = D^{-1}C_L \quad \kappa\alpha\iota \quad U = D^{-1}C_U. \tag{3.35}$$

Είναι φανέρο ότι για την ύπαρξη της παραπάνω μεθόδου θα πρέπει να υπάρχει ο D^{-1} πράγμα που σημαίνει ότι $detD \neq 0$. Επειδή $detD = \prod_{i=1}^{\nu} a_{ii}$ συνεπάγεται ότι όλα τα διαγώνια στοιχεία του A θα πρέπει να είναι διάφορα του μηδενός. Η μέθοδος (4.21) είναι γνωστή με το όνομα *Jacobi Overrelaxation*(JOR) μέθοδος. Παρατηρούμε ότι αν $\tau = 1$ τότε η (4.21) δίνει

$$u^{(n+1)} = Bu^{(n)} + c, \quad n = 0, 1, 2, \ldots \tag{3.36}$$

η οποία είναι γνωστή σαν η μέθοδος του *Jacobi* (J).

Για την υλοποίηση των παραπάνω μεθόδων χρειαζόμαστε τις σχέσεις που συνδέουν τις συντεταγμένες των διανυσμάτων $u^{(n+1)}$ και $u^{(n)}$. Γράφοντας το σύστημα (3.99) υπό μορφή συντεταγμένων έχουμε

$$\sum_{j=1}^{\nu} a_{ij}u_j = b_i, \quad i = 1(1)\nu$$

ή

$$a_{ii}u_i + \sum_{j=1}^{i-1} a_{ij}u_j + \sum_{j=i+1}^{\nu} a_{ij}u_j = b_i, \quad i = 1(1)\nu. \tag{3.37}$$

Η παραπάνω μορφή είναι συμβιβαστή με τη διάσπαση (4.17) καθόσον υπό μορφή συντεταγμένων οι ποσότητες Du, $-C_L u$ και $-C_U u$ εκφράζονται αντίστοιχα από τον πρώτο, δεύτερο και τρίτο όρο του αθροίσματος της (4.24). Η μέθοδος του *Jacobi* (4.23) μπορεί να γραφτεί υπό μορφή συντεταγμένων σαν

$$u_i^{(n+1)} = \sum_{\substack{j=1 \\ j\neq i}}^{\nu} \hat{a}_{ij}u_j^{(n)} + \hat{b}_i, \quad i = 1(1)\nu, \tag{3.38}$$

όπου

$$\hat{a}_{ij} = -\frac{a_{ij}}{a_{ii}} \quad i, j = 1(1)\nu, \quad i \neq j$$

και

$$\hat{b}_i = \frac{b_i}{a_{ii}} \quad i = 1(1)\nu. \tag{3.39}$$

Επίσης η *JOR* δίνεται από τους τύπους

$$u_i^{(n+1)} = (1-\tau)u_i^{(n)} + \tau \sum_{\substack{j=1 \\ j\neq i}}^{\nu} \hat{a}_{ij}u_j^{(n)}, \quad i = 1(1)\nu, \tag{3.40}$$

Παρατηρούμε ότι η ταχύτητα σύγκλισης των δύο μεθόδων(JOR και J) είναι ανεξάρτητη από τη διάταξη των εξισώσεων του συστήματος.

Παράδειγμα. Δίνεται το γραμμικό σύστημα

$$\begin{aligned} 2x_1 - x_2 &= 1 \\ -x_1 + 2x_2 - x_3 &= 0 \\ -x_2 + 2x_3 &= 1 \end{aligned}$$

Να δειχθεί ότι η μέθοδος του *Jacobi* συγκλίνει και να βρεθούν οι τρείς πρώτες επαναλήψεις, αν $x^{(0)} = (1, 0, 1)$.

Λύση. Ο επαναληπτικός πίνακας της μεθόδου *Jacobi* είναι ο

$$B = I - D^{-1}A = \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 \end{bmatrix}$$

Επίσης $\| B \|_1 = \| B \|_\infty = 1$. Γιά τον λόγο αυτό θα υπολογίσουμε την $S(B)$. Οι ιδιοτιμές του B είναι 0, $\frac{\sqrt{2}}{2}$ και $-\frac{\sqrt{2}}{2}$, επομένως $S(B) = \frac{\sqrt{2}}{2} < 1$ που δηλώνει ότι η μέθοδος συγκλίνει. Από το ανωτέρω γραμμικό σύστημα έχουμε ότι η μέθοδος του *Jacobi* είναι η

$$\begin{aligned} x_1^{(n+1)} &= \frac{1}{2}(1 + x_2^{(n)}) \\ x_2^{(n+1)} &= \frac{1}{2}(x_1^{(n)} + x_3^{(n)}), \qquad\qquad n = 0, 1, 2 \ldots \\ x_3^{(n+1)} &= \frac{1}{2}(1 + x_2^{(n)}). \end{aligned}$$

Για $x^{(0)} = (1, 0, 1)^T$, δηλαδή για $x_1^{(0)} = 1$, $x_2^{(0)} = 0$ και $x_3^{(0)} = 1$ έχουμε

$n = 0$

$$\begin{aligned} x_1^{(1)} &= \frac{1}{2}(1 + x_2^{(0)}) = \frac{1}{2}(1 + 0) = \frac{1}{2} \\ x_2^{(1)} &= \frac{1}{2}(x_1^{(0)} + x_3^{(0)}) = \frac{1}{2}(1 + 1) = 1 \\ x_3^{(1)} &= \frac{1}{2}(1 + x_2^{(0)}) = \frac{1}{2}(1 + 0) = \frac{1}{2} \end{aligned}$$

$n = 1$

$$\begin{aligned} x_1^{(2)} &= \frac{1}{2}(1 + x_2^{(1)}) = \frac{1}{2}(1 + 1) = 1 \\ x_2^{(2)} &= \frac{1}{2}(x_1^{(1)} + x_3^{(1)}) = \frac{1}{2}(\frac{1}{2} + \frac{1}{2}) = \frac{1}{2} \\ x_3^{(2)} &= \frac{1}{2}(1 + x_2^{(1)}) = \frac{1}{2}(1 + 1) = 1 \end{aligned}$$

$n = 2$

$$\begin{aligned} x_1^{(3)} &= \frac{1}{2}(1 + x_2^{(2)}) = \frac{1}{2}(1 + \frac{1}{2}) = \frac{3}{4} \\ x_2^{(3)} &= \frac{1}{2}(x_1^{(2)} + x_3^{(2)}) = \frac{1}{2}(1 + 1) = 1 \\ x_3^{(3)} &= \frac{1}{2}(1 + x_2^{(2)}) = \frac{1}{2}(1 + \frac{1}{2}) = \frac{3}{4}. \end{aligned}$$

Στη συνέχεια θέτουμε

$$R = D - C_L \tag{3.41}$$

όπου η μορφή αυτή του R προσεγγίζει καλύτερα τον A από την προηγούμενη. Από την (4.2) έχουμε

$$u^{(n+1)} = u^{(n)} + \tau(1-L)^{-1}D^{-1}(b - Au^{(n)}) \tag{3.42}$$

ή

$$u^{(n+1)} = L_{\tau,1}u^{(n)} + \tau(1-L)^{-1}c \tag{3.43}$$

όπου

$$L_{\tau,1} = I - \tau(1-L)^{-1}D^{-1}A. \tag{3.44}$$

Εκφράζοντας τον $L_{\tau,1}$ σε όρους των L και U η (3.43) γράφεται ως εξής

$$u^{(n+1)} = (1-\tau)u^{(n)} + Lu^{(n+1)} + (\tau-1)Lu^{(n)} + \tau Uu^{(n)} + \tau c. \tag{3.45}$$

Η ανωτέρω μέθοδος καλείται Επιταχυντική $Gauss - Seidel(EGS)$ γιατί για $\tau = 1$ προκύπτει η γνωστή μέθοδος $Gauss - Seidel(GS)$. Πράγματι, για $\tau = 1$ η (3.45) δίνει

$$u^{(n+1)} = Lu^{(n+1)} + Uu^{(n)} + c. \tag{3.46}$$

Μετατρέποντας τώρα την EGS και GS υπό μορφή συνιστωσών λαμβάνουμε αντίστοιχα τους τύπους

$$u_i^{(n+1)} = \sum_{j=1}^{i-1} \hat{a}_{ij}u_j^{(n+1)} + (1-\tau)u_i^{(n)} + (\tau-1)(\sum_{j=1}^{i-1} \hat{a}_{ij}u_j^{(n)}) + \tau(\sum_{j=i+1}^{\nu} \hat{a}_{ij}u_j^{(n)}) + \tau\hat{b}_i, \quad i = 1(1)\nu \tag{3.47}$$

και θέτοντας $\tau = 1$ για την GS λαμβάνουμε ότι

$$u_i^{(n+1)} = \sum_{j=1}^{i-1} \hat{a}_{ij}u_j^{(n+1)} + \sum_{j=i+1}^{\nu} \hat{a}_{ij}u_j^{(n)} + \hat{b}_i, \quad i = 1(1)\nu. \tag{3.48}$$

Για την ύπαρξη των δύο ανωτέρω μεθόδων θα πρέπει να υπάρχει ο $(D - C_L)^{-1}$ ή $det(D - C_L) = detD \neq 0$ πράγμα που ισχύει αν όλα τα διαγώνια στοιχεία του A είναι διάφορα του μηδενός. *Παρατηρούμε τώρα από τις (3.47) και (3.48), ότι οι αριθμητικές πράξεις επηρεάζονται αν εναλλάξουμε τη σειρά των εξισώσεων του συστήματός μας.*

Παράδειγμα. Να επαναληφθεί η ίδια εργασία, όπως στο προηγούμενο παράδειγμα, για τη μέθοδο $Gauss - Seidel$.

Λύση. Ο επαναληπτικός πίνακας της μεθόδου $Gauss - Seidel$ είναι ο

$$L_1 = (I-L)^{-1}U = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ 0 & -\frac{1}{2} & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 \end{bmatrix}.$$

Από τη σχέση $(I-L)X = I$ υπολογίζεται εύκολα ο $(I-L)^{-1}$. Επομένως

$$L_1 = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{4} & \frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{4} & \frac{1}{2} \\ 0 & \frac{1}{8} & \frac{1}{4} \end{bmatrix}.$$

Οι ιδιοτιμές του L_1 είναι οι $0, 1/2, -1/2$ επομένως $S(L_1) = 1/2 < 1$ που αποδεικνύει ότι η GS συγκλίνει. Παρατηρούμε ότι $S(L_1) = S(B)^2$ για το παρόν παράδειγμα. Το αποτέλεσμα αυτό ισχύει για όλους τους ($block$) τριδιαγώνιους πίνακες. Η μέθοδος $Gauss-Seidel$ δίνεται από το ακόλουθο επαναληπτικό σχήμα (βλ. (3.48))

$$\begin{aligned} x_1^{(n+1)} &= \frac{1}{2}(1 + x_2^{(n)}) \\ x_2^{(n+1)} &= \frac{1}{2}(x_1^{(n+1)} + x_3^{(n)}), \qquad\qquad n = 0, 1, \ldots \\ x_3^{(n+1)} &= \frac{1}{2}(1 + x_2^{(n+1)}). \end{aligned}$$

Για $x^{(0)} = (1, 0, 1)^T$, δηλαδή για $x_1^{(0)} = 1$, $x_2^{(0)} = 0$ και $x_3^{(0)} = 1$, έχουμε

$n = 0$

$$\begin{aligned} x_1^{(1)} &= \frac{1}{2}(1 + x_2^{(0)}) = \frac{1}{2}(1+0) = \frac{1}{2} \\ x_2^{(1)} &= \frac{1}{2}(x_1^{(1)} + x_3^{(0)}) = \frac{1}{2}(\frac{1}{2} + 1) = \frac{3}{4} \\ x_3^{(1)} &= \frac{1}{2}(1 + x_2^{(1)}) = \frac{1}{2}(1 + \frac{3}{4}) = \frac{7}{8} \end{aligned}$$

$n = 1$

$$x_1^{(2)} = \frac{1}{2}(1 + x_2^{(1)}) = \frac{1}{2}(1 + \frac{3}{4}) = \frac{7}{8}$$
$$x_2^{(2)} = \frac{1}{2}(x_1^{(2)} + x_3^{(1)}) = \frac{1}{2}(\frac{7}{8} + \frac{7}{8}) = \frac{7}{8}$$
$$x_3^{(2)} = \frac{1}{2}(1 + x_2^{(2)}) = \frac{1}{2}(1 + \frac{7}{8}) = \frac{15}{16}$$

$n = 2$

$$x_1^{(3)} = \frac{1}{2}(1 + x_2^{(2)}) = \frac{1}{2}(1 + \frac{7}{8}) = \frac{15}{16}$$
$$x_2^{(3)} = \frac{1}{2}(x_1^{(3)} + x_3^{(2)}) = \frac{1}{2}(\frac{15}{16} + \frac{15}{16}) = \frac{15}{16}$$
$$x_3^{(3)} = \frac{1}{2}(1 + x_2^{(3)}) = \frac{1}{2}(1 + \frac{15}{16}) = \frac{31}{32}.$$

Παρατηρούμε ότι η μέθοδος GS συγκλίνει πολύ γρηγορότερα από τη μέθοδο *Jacobi* προς την ακριβή λύση $(1, 1, 1)^T$ του συστήματος. Είναι δυνατόν να βρεθούν δύο άλλες μέθοδοι αν εισάγουμε μία παράμετρο στη μορφή του R. Έτσι αν θέσουμε

$$R = D - \omega C_L \tag{3.49}$$

όπου ω είναι ένας πραγματικός αριθμός του οποίου ο ρόλος στη φάση αυτή, είναι να διαταράξει τον R έτσι ώστε να προσεγγίζει καλύτερα τον A από τη μορφή (3.41), τότε από την (4.3) έχουμε

$$u^{(n+1)} = u^{(n)} + \tau(I - \omega L)^{-1}D^{-1}(b - Au^{(n)}), \quad n = 0, 1, 2 \dots \tag{3.50}$$

$$u^{(n+1)} = L_{\tau,\omega}u^{(n)} + \tau(I - \omega L)^{-1}c \tag{3.51}$$

όπου

$$L_{\tau,\omega} = I - \tau(I - \omega L)^{-1}D^{-1}A. \tag{3.52}$$

Το ανωτέρω επαναληπτικό σχήμα είναι γνωστό σαν η Επιταχυντική *Successive Overrelaxation*$(ESOR)$ μέθοδος. Προκειμένου να βρούμε την εξίσωση των συνιστωσών η (3.50) μπορεί να γραφτεί σαν

$$u^{(n+1)} = (1-\tau)u^{(n)} + \omega Lu^{(n+1)} + (\tau - \omega)Lu^{(n)} + \tau Uu^{(n)} + \tau c \tag{3.53}$$

οπότε

$$u_i^{(n+1)} = (1-\tau)u_i^{(n)} + \omega \sum_{j=1}^{i-1} \hat{a}_{ij}u_j^{(n+1)} + (\tau - \omega) \sum_{j=1}^{i-1} \hat{a}_{ij}u_j^{(n)} + \tau \sum_{j=i+1}^{\nu} \hat{a}_{ij}u_j^{(n)} + \tau \hat{b}_i, \quad i = 1(1)\nu. \tag{3.54}$$

Κατά την υλοποίηση της $ESOR$ είναι δυνατόν να γίνει εξοικονόμηση των υπολογισμών αν αποθηκευτεί η ποσότης $Lu^{(n)}$ προκειμένου να χρησιμοποιηθεί στην επόμενη επανάληψη. Αν θέσουμε $\tau = \omega$ στην $ESOR$ λαμβάνουμε τη δημοφιλή $Successive\ Overrelaxation(SOR)$ μέθοδο, η οποία δίνεται διαδοχικά από τους τύπους

$$u^{(n+1)} = u^{(n)} + \omega(I - \omega L)^{-1}D^{-1}(b - Au^{(n)}) \tag{3.55}$$

$$u^{(n+1)} = L_{\omega,\omega}u^{(n)} + \omega(I - \omega L)^{-1}c \tag{3.56}$$

όπου

$$L_{\omega,\omega} = I - \omega(I - \omega L)^{-1}D^{-1}A. \tag{3.57}$$

Επίσης η SOR γράφεται και σαν

$$u^{(n+1)} = (I - \omega L)^{-1}[(1 - \omega)I + \omega U]u^{(n)} + \omega(I - \omega L)^{-1}c \tag{3.58}$$

ή

$$u^{(n+1)} = (1 - \omega)u^{(n)} + \omega[Lu^{(n+1)} + Uu^{(n)} + c] \tag{3.59}$$

ή

$$u^{(n+1)} = (1 - \omega)u^{(n)} + \omega u_{GS}^{(n+1)} \tag{3.60}$$

όπου $u_{GS}^{(n+1)}$ συμβολίζει την $n + 1$ επανάληψη της GS μεθόδου. Η (3.60) υπήρξε η αφετηρία της ανακάλυψης της SOR μεθόδου. *Ωστόσο η στρατηγική που ακολουθήθηκε μέχρι τώρα, έχοντας δηλαδή σαν αφετηρία γέννεσης των επαναληπτικών μεθόδων την (4.2), είχε σαν αποτέλεσμα την εύρεση μιας γενικότερης και αποτελεσματικότερης μεθόδου της $ESOR$.* Τέλος, υπό μορφή συνιστωσών η SOR δίνεται από τους τύπους

$$u_i^{(n+1)} = (1-\omega)u_i^{(n)} + \omega\sum_{j=1}^{i-1}\hat{a}_{ij}u_j^{(n+1)} + \omega\sum_{j=i+1}^{\nu}\hat{a}_{ij}u_j^{(n)} + \omega\hat{b}_i, \quad i = 1(1)\nu. \tag{3.61}$$

Παράδειγμα. Να εφαρμοστεί η SOR μέθοδος στο προηγούμενο παράδειγμα.

Λύση. Η SOR μέθοδος δίνεται από το ακόλουθο σχήμα (βλ. (3.60))

$$x^{(n+1)} = (1 - \omega)u^{(n)} + \omega u_{GS}^{(n+1)},$$

όπου $u_{GS}^{(n+1)}$ είναι το επαναληπτικό διάνυσμα που προκύπτει από την εφαρμογή της $Gauss-Seidel$ μεθόδου. Επομένως, για το συγκεκριμένο τριδιαγώνιο σύστημα του προηγούμενου παραδείγματος, η SOR παράγει το επαναληπτικό σχήμα

$$\begin{aligned} x_1^{(n+1)} &= (1-\omega)x_1^{(n)} + \omega(1 + x_2^{(n)}) \\ x_2^{(n+1)} &= (1-\omega)x_2^{(n)} + \omega(x_1^{(n+1)} + x_3^{(n)}), \qquad n = 0, 1, 2 \dots \\ x_3^{(n+1)} &= (1-\omega)x_3^{(n)} + \omega(1 + x_2^{(n+1)}). \end{aligned}$$

Η βέλτιστη τιμή του ω δίνεται από τον τύπο (Θεώρημα 1.4.16)

$$\omega_b = \frac{2}{1 + \sqrt{1 - S(B)^2}}$$

ή

$$\omega_b = \frac{2}{1 + \sqrt{1 - \frac{1}{2}}} = \frac{4}{2 + \sqrt{2}} \simeq 1.1716.$$

ενώ

$$S(\omega_b) = \omega_b - 1 \simeq 0.1716.$$

Λαμβάνοντας, όπως και στα προηγούμενα παραδείγματα, $x^{(0)} = (1, 0, 1)^T$, δηλαδή, $x_1^{(0)} = 1$, $x_2^{(0)} = 0$ και $x_3^{(0)} = 1$ έχουμε για $n = 0$

$$\begin{aligned} x_1^{(1)} &= (1 - \frac{4}{2+\sqrt{2}}) \cdot 1 + \frac{4}{2+\sqrt{2}}(1 + 0) = 1 \\ x_2^{(1)} &= (1 - \frac{4}{2+\sqrt{2}}) \cdot 0 + \frac{4}{2+\sqrt{2}}(1 + 1) = \frac{8}{2+\sqrt{2}} \\ x_3^{(1)} &= (1 - \frac{4}{2+\sqrt{2}}) \cdot 1 + \frac{4}{2+\sqrt{2}}(1 + \frac{8}{2+\sqrt{2}}) = 1 + \frac{32}{(2+\sqrt{2})^2} \end{aligned}$$

κ.ο.κ. Στη συνέχεια θα ασχοληθούμε με μία άλλη κατηγορία επαναληπτικών μεθόδων, οι οποίες ονομάζονται επαναληπτικά σχήματα δύο επιπέδων. Μια κλασική τέτοια μέθοδος προκύπτει σαν παραλλαγή της SOR και καλείται $Symmetric\ SOR(SSOR)$. Κάθε επανάληψη της $SSOR$ αποτελείται από δύο ημιεπαναλήψεις. Η πρώτη είναι μία προς τα εμπρός SOR που υπολογίζει τους αγνώστους με τη διάταξη $u_i, i = 1(1)n$ και η άλλη είναι μία προς τα πίσω SOR υπολογίζοντας τους αγνώστους με αντίθετη διάταξη δηλ. $u_i, i = n(-1)1$. Από τα παραπάνω προκύπτει ότι η $SSOR$ είναι το επαναληπτικό σχήμα

$$u^{(n+\frac{1}{2})} = (1-\omega)u^{(n)} + \omega(Lu^{(n+\frac{1}{2})} + Uu^{(n)} + c)$$

και

$$u^{(n+1)} = (1-\omega)u^{(n+\frac{1}{2})} + \omega(Lu^{(n+\frac{1}{2})} + Uu^{(n+1)} + c) \tag{3.62}$$

όπου $\omega \neq 0$ είναι μία πραγματική παράμετρος και $u^{(n+\frac{1}{2})}$ είναι μια ενδιάμεση προσέγγιση της λύσης του συστήματος. Υπό μορφή συντεταγμένων η $SSOR$ γράφεται

$$u_i^{(n+\frac{1}{2})} = (1-\omega)u_i^{(n)} + \omega(\sum_{j=1}^{i-1} \hat{a}_{ij}u_j^{(n+\frac{1}{2})} + \sum_{j=i+1}^{\nu} \hat{a}_{ij}u_j^{(n)} + \hat{b}_i), \quad i = 1(1)\nu,$$

και

$$u_i^{(n+1)} = (1-\omega)u_i^{(n+\frac{1}{2})} + \omega(\sum_{j=1}^{i-1} \hat{a}_{ij}u_j^{(n+\frac{1}{2})} + \sum_{j=i+1}^{\nu} \hat{a}_{ij}u_j^{(n+\frac{1}{2})} + \hat{b}_i), \quad i = \nu(-1)1.$$

Από τις (3.62) έχουμε ότι

$$u^{(n+\frac{1}{2})} = L_\omega u^{(n)} + \omega(I-\omega L)^{-1}c$$

και

$$u^{(n+1)} = U_\omega u^{(n+\frac{1}{2})} + \omega(I-\omega U)^{-1}c \tag{3.63}$$

όπου

$$L_\omega = (I-\omega L)^{-1}[(1-\omega)I + \omega U] \tag{3.64}$$

και

$$U_\omega = (I-\omega U)^{-1}[(1-\omega)I + \omega L].$$

Απαλείφοντας τον όρο $u^{(n+\frac{1}{2})}$ από τις (3.63) η $SSOR$ γίνεται

$$u^{(n+1)} = F_\omega u^{(n)} + k_\omega \tag{3.65}$$

όπου

$$F_\omega = U_\omega L_\omega = I - \omega(2-\omega)(I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}A \tag{3.66}$$

και

$$k_\omega = \omega(2-\omega)(I-\omega U)^{-1}(I-\omega L)^{-1}c. \tag{3.67}$$

Από τις (3.66) και (3.67) παρατηρούμε ότι $\omega \neq 0, 2$. Επειδή η $SSOR$ είναι ο συνδυασμός δύο SOR επαναλήψεων, οι υπολογισμοί θα εξαρτώνται από τη διάταξη των εξισώσεων. Επίσης η $SSOR$ απαιτεί διπλάσια υπολογιστική εργασία από την SOR. Ωστόσο είναι δυνατόν να μειωθεί η υπολογιστική εργασία για κάθε $SSOR$ επανάληψη με τη διάθεση ενός επιπλέον ν-διάστατου διανύσματος. Η τεχνική αυτή είναι δυνατή καθόσον ορισμένα διανύσματα επαναλαμβάνονται μεταξύ δύο διαδοχικών επαναλήψεων και δε χρειάζεται να υπολογιστούν πάλι όπως υποδηλώνεται παρακάτω

$$\begin{cases} u^{(n+\frac{1}{2})} = \omega(\mathbf{Lu^{(n+\frac{1}{2})}} + Uu^{(n)} + c) + (1-\omega)u^{(n)} \\ u^{(n+1)} = \omega(\mathbf{Lu^{(n+\frac{1}{2})}} + \mathbf{Uu^{(n+1)}} + c + (1-\omega)u^{(n+\frac{1}{2})} \end{cases}$$

$$\begin{cases} u^{(n+\frac{3}{2})} = \omega(\mathbf{Lu^{(n+\frac{1}{2})}}) + \mathbf{Uu^{(n+1)}} + c + (1-\omega)u^{(n+1)} \\ u^{(n+2)} = \omega(Lu^{(n+\frac{3}{2})}) + Uu^{(n+2)} + c + (1-\omega)u^{(n+\frac{3}{2})}. \end{cases}$$

Η τεχνική αυτή είναι γνωστή σαν το σχήμα του $Niethammer$. Τέλος, αν ο R λάβει τη μορφή

$$R = (D - \omega C_L)D^{-1}(D - \omega C_U) \tag{3.68}$$

όπου ω πραγματική παράμετρος, τότε από την (4.2) έχουμε

$$u^{(n+1)} = u^{(n)} + \tau(I - \omega U)^{-1}(I - \omega L)^{-1}D^{-1}(b - Au^{(n)}). \tag{3.69}$$

Η ανωτέρω μέθοδος είναι γνωστή ως η $Preconditioned\ Simultanous\ Displacement(PSD)$ μέθοδος. Η PSD είναι μια γενίκευση της $SSOR$ αφού η τελευταία λαμβάνεται για $\tau = \omega(2-\omega)$. Από την (3.69) έχουμε

$$u^{(n+1)} - \omega U u^{(n+1)} + \omega U u^{(n)} = u^{(n)} + \tau(I - \omega L)^{-1}D^{-1}(b - Au^{(n)}),$$

η οποία μπορεί να διασπασθεί στο ακόλουθο επαναληπτικό σχήμα δύο επιπέδων

$$\begin{aligned} & u^{(n+\frac{1}{2})} = u^{(n)} + \tau(I - \omega L)^{-1}D^{-1}(b - Au^{(n)}) \\ & \text{και} \\ & u^{(n+1)} = u^{(n+\frac{1}{2})} + \omega U u^{(n+1)} - \omega U u^{(n)}. \end{aligned} \tag{3.70}$$

Είναι φανερό ότι η πρώτη των (3.70) είναι η $ESOR$ μέθοδος άρα έχουμε

$$u^{(n+\frac{1}{2})} = (1-\tau)u^{(n)} + \omega L u^{(n+\frac{1}{2})} + (\tau-\omega)Lu^{(n)} + \tau(Uu^{(n)}+c)$$

και

$$u^{(n+1)} = u^{(n+\frac{1}{2})} + \omega U u^{(n+1)} - \omega U u^{(n)}. \tag{3.71}$$

Από την (3.71) παρατηρούμε ότι το διάνυσμα $Uu^{(n)}$ επαναλαμβάνεται και στη δεύτερη ημιεπανάληψη οπότε είναι δυνατόν να εφαρμοστεί και εδώ το σχήμα του $Niethammer$. Επίσης η PSD υλοποείται με ανάλογο τρόπο όπως και η $SSOR$. Έτσι οι εξισώσεις των συντεταγμένων για την PSD μέθοδο είναι οι ακόλουθες

$$u_i^{(n+\frac{1}{2})} = (1-\tau)u_i^{(n)} + \omega\sum_{j=1}^{i-1}\hat{a}_{ij}u_j^{(n+\frac{1}{2})} + (\tau-\omega)\sum_{j=1}^{i-1}\hat{a}_{ij}u_j^{(n)} + \tau(\sum_{j=i+1}^{\nu}\hat{a}_{ij}u_j^{(n)} + \hat{b}_i), \quad i = 1(1)\nu \tag{3.72}$$

και

$$u_i^{(n+1)} = u_i^{(n+\frac{1}{2})} + \omega(\sum_{j=i+1}^{\nu}\hat{a}_{ij}u_j^{(n+1)} - \sum_{j=i+1}^{\nu}\hat{a}_{ij}u_j^{(n)}), \quad i = \nu(-1)1.$$

Τέλος, μία άλλη μορφή της PSD μεθόδου είναι η

$$u^{(n+1)} = \Delta_{\tau,\omega}u^{(n)} + \delta_{\tau,\omega}, \tag{3.73}$$

όπου

$$\Delta_{\tau,\omega} = I - \tau\Gamma_\omega, \quad \Gamma_\omega = (I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}A \tag{3.74}$$

και

$$\delta_{\tau,\omega} = \tau(I-\omega U)^{-1}(I-\omega L)^{-1}D^{-1}b.$$

3.3 Αλγόριθμοι των βασικών επαναληπτικών μεθόδων

Αλγόριθμος της μεθόδου *Jacobi*

1. Διάβασε τη διάταξη του πίνακα ν, τα στοιχεία $a_{ij}, i, j = 1(1)\nu$ του A, το δεύτερο μέλος $b_i, \quad i = 1(1)\nu$, το αρχικό διάνυσμα $x0_i, \quad i = 1(1)\nu$, την ανεκτικότητα tol και τον μέγιστο αριθμό επαναλήψεων $maxits$.

2. Να τεθεί
$$itcount = 0$$

3. Όσο ισχύει $itcount \leq maxits$ να εκτελούνται τα βήματα (α′)-(δ′)

   (α′) Για $i = 1(1)\nu$ να τεθεί
$$x1_i = (-\sum_{\substack{j=1 \\ j \neq i}}^{\nu} a_{ij} x0_j + b_i)/a_{ii}$$

   (β′) Να τεθεί
$$itcount = itcount + 1$$

   (γ′) Αν
$$\| x1 - x0 \|_\infty < tol$$
   τότε τύπωσε $x1_i, i = 1(1)\nu$. Τέλος.

   (δ′) Για $i = 1(1)\nu$ να τεθεί
$$x0_i = x1_i$$

4. Τύπωσε(‘ Όχι σύγκλιση μετά από $maxits$ επαναλήψεις ’).

5. Τέλος.

Αλγόριθμος της μεθόδου των $Gauss - Seidel$

1. Διάβασε την τάξη του πίνακα ν, τα στοιχεία $a_{ij}, \quad i, j = 1(1)\nu$ του A, το δεύτερο μέλος $b_i, \quad i = 1(1)\nu$, το αρχικό διάνυσμα $x0_i, \quad i = 1(1)\nu$, την ανεκτικότητα tol και το μέγιστο αριθμό επαναλήψεων $maxits$.

2. Να τεθεί
$$itcount = 0$$

3. Όσο ισχύει $itcount \leq maxits$ να εκτελούνται τα βήματα (α′) - (δ′)

   (α′) Για $i = 1(1)\nu$ να τεθεί

   $$x1_i = (-\sum_{j=1}^{i-1} a_{ij}x1_j - \sum_{j=i+1}^{\nu} a_{ij}x0_j + b_i)/a_{ii}$$

   (β′) Να τεθεί

   $$itcount = itcount + 1$$

   (γ′) Αν

   $$\| x1 - x0 \|_\infty < tol$$

   τότε τύπωσε $x1_i, i = 1(1)\nu$. Τέλος.

   (δ′) Για $i = 1(1)\nu$ να τεθεί

   $$x0_i = x1_i$$

4. Τύπωσε(''Οχι σύγκλιση μετά από $maxits$ επαναλήψεις ').

5. Τέλος.

3.4 Σύγκλιση των βασικών επαναληπτικών μεθόδων

Ένα βασικό ερώτημα που αφορά όλες τις επαναληπτικές μεθόδους είναι: Ποιές είναι οι συνθήκες εκείνες που εξασφαλίζουν τη σύγκλιση μίας επαναληπτικής μεθόδου; Στη συνέχεια παρουσιάζονται ορισμένα θεωρήματα σύγκλισης (βλ. $[Young]$, $[Varga]$) για τις επαναληπτικές μεθόδους της παραγράφου 1.2. Καταρχήν όμως ορίζουμε δύο μεγάλες κλάσεις πινάκων στις οποίες αν ανήκει κάποιος πίνακας, τότε ο πίνακας αυτός είναι μη ιδιάζων, πράγμα που εξασφαλίζει τη μοναδικότητα της λύσης του συστήματος μας, αποφεύγοντας έτσι το κριτήριο της διακρίνουσας.

Ορισμός. Ένας πίνακας $A = (a_{ij})$ τάξης ν είναι αδιάσπαστος (*irreducible*) αν $\nu = 1$ ή αν $\nu > 1$ και για οποιαδήποτε δεδομένα μη κενά και ξένα μεταξύ τους υποσύνολα S και T του συνόλου W των πρώτων ν θετικών ακεραίων αριθμών τέτοια ώστε $S \cup T = W$, υπάρχει $i \epsilon S$ και $j \epsilon T$ τέτοια ώστε $a_{ij} \neq 0$.

Θεώρημα 3.4.1. *Ο πίνακας A είναι αδιάσπαστος αν και μόνον αν δεν υπάρχει ένας μεταθετικός πίνακας P τέτοιος ώστε ο $P^{-1}AP$ να έχει τη μορφή*

$$P^{-1}AP = \begin{bmatrix} F & O \\ G & H \end{bmatrix} \tag{3.75}$$

όπου F και G είναι τετραγωνικοί πίνακες και O είναι ο μηδενικός πίνακας.

Είναι φανερό ότι στην περίπτωση όπου υπάρχει ο P, τότε ο A καλείται διασπάσιμος (reducible) και το αρχικό σύστημα μπορεί να μετασχηματιστεί στο $Au = k$, όπου ο $A = P^{-1}AP$ είναι της μορφής (4.25). *Στην περίπτωση αυτή έχουμε*

$$Fu_1 = k_1$$
$$Gu_1 + Hu_2 = k_2$$

οπότε υποβιβάστηκε το αρχικό σύστημα σε δύο μικρότερης τάξης συστήματα.

Μία χρήσιμη μέθοδος για να διαπιστωθεί αν ένας πίνακας είναι αδιάσπαστος στην πράξη δίνεται από το παρακάτω θεώρημα.

Θεώρημα 3.4.2. *Ένας πίνακας $A = (a_{ij})$ τάξης ν είναι αδιάσπαστος αν και μόνον αν ν=1 ή δοθέντων δύο οποιονδήποτε διακεκριμένων ακεραίων i και j με $1 \leq i \leq \nu$, $1 \leq j \leq \nu$, τότε $a_{ij} \neq 0$ ή υπάρχουν φυσικοί $i_1, i_2, \ldots, i_s$ τέτοιοι ώστε $a_{ii_1}, a_{i_1 i_2} \ldots a_{i_s j} \neq 0$.*

Απόδειξη. Αν δοθεί ο πίνακας A τάξης ν, τότε θεωρούμε τα διακεκριμένα σημεία $P_1, P_2 \ldots, P_\nu$ και κατασκευάζουμε το κατευθυνόμενο γράφημα (directed graph) του A, σχεδιάζοντας ένα τόξο από το P_i προς το P_j για κάθε $a_{ij} \neq 0$. Αν $a_{ii} \neq 0$, τότε απλά σχεδιάζουμε ένα κυκλικό τόξο, το οποίο περιέχει το P_i. Ο πίνακας είναι αδιάσπαστος αν και μόνο αν $\nu = 1$ ή υπάρχει μιά αλυσίδα από τόξα από το P_i στο P_{i_1}, από το P_{i_1} στο P_{i_2}...από το P_{i_2} στο P_j(συνδεδεμένο γράφημα). ■

Παράδειγμα Έστω ο τριδιαγώνιος πίνακας

$$A = \begin{bmatrix} a_{11} & a_{12} & & & & \\ a_{21} & a_{22} & a_{23} & & \mathbf{0} & \\ & & & & & \\ & & & & & \\ & \mathbf{0} & & a_{\nu-1,\nu-2} & a_{\nu-1,\nu-1} & a_{\nu-1,\nu} \\ & & & & a_{\nu,\nu-1} & a_{\nu\nu} \end{bmatrix}$$

Το κατευθυνόμενο γράφημα του πίνακα δίνεται στο σχήμα 3.4.3. Από το σχήμα 3.4.3 εύκολα παρατηρούμε ότι το γράφημα είναι συνδεδεμένο συνεπώς οι τριδιαγώνιοι πίνακες είναι αδιάσπαστοι. Η άλλη σημαντική κλάση πινάκων που είναι εξίσου χρήσιμη περιλαμβάνει εκείνους τους πίνακες που είναι διαγώνια υπέρτεροι.

Ορισμός Ένας πίνακας $A = (a_{ij})$ τάξης ν καλείται ασθενά διαγώνιος υπέρτερος αν

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j \neq i}}^{\nu} |a_{ij}|, \quad i = 1(1)\nu \tag{3.76}$$

και για τουλάχιστον ένα i

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|. \tag{3.77}$$

Στην περίπτωση όπου η (4.27) ισχύει για όλα τα $i = 1(1)\nu$ τότε ο A καλείται αυστηρά διαγώνιος υπέρτερος. Επειδή είναι αρκετά δύσκολο να διαπιστωθεί αν $detA \neq 0$ προκειμένου να αποδείξουμε ότι ο A είναι μη ιδιάζων, γιαυτό συνήθως καταφεύγουμε στο κριτήριο που δίνεται από το ακόλουθο θεώρημα.

Θεώρημα 3.4.3. *Αν ο A είναι αδιάσπαστος πίνακας και ασθενά διαγώνιος υπέρτερος, τότε $detA \neq 0$ και όλα τα διαγώνια στοιχεία του A είναι διάφορα του μηδενός.*

Απόδειξη. Βλ.$[Varga, Young]$ ■

Πόρισμα 3.4.1. *Αν ο A είναι αυστηρά διαγώνια υπέρτερος, τότε $detA \neq 0$.*

Στη συνέχεια δίνεται μία ικανή συνθήκη για να είναι θετικά ορισμένος ένας Ερμειτιανός πίνακας, χρησιμοποιώντας τις παραπάνω ιδιότητες (αδιάσπαστος και ασθενά διαγώνιος υπέρτερος).

Θεώρημα 3.4.4. *Αν ο A είναι ένας Ερμειτιανός πίνακας με μη αρνητικά διαγώνια στοιχεία και επιπλέον είναι ασθενά διαγώνια υπέρτερος, τότε ο A είναι μη αρνητικά ορισμένος. Αν ο A είναι και αδιάσπαστος ή μη ιδιάζων, τότε ο A είναι θετικά ορισμένος.*

Απόδειξη. Είναι γνωστό ότι όλες οι ιδιοτιμές ενός Ερμειτιανού πίνακα είναι πραγματικοί αριθμοί. Αν λ είναι μία ιδιοτιμή του A τότε

$$det(A - \lambda I) = 0. \tag{3.78}$$

Ας υποθέσουμε τώρα ότι $\lambda < 0$, τότε ο πίνακας $A - \lambda I$ είναι αυστηρά διαγώνια υπέρτερος και λόγω του Πορίσματος 1.4.1, $det(A - \lambda I) \neq 0$, το οποίο έρχεται σε αντίθεση με την (4.28). Συνεπώς όλες οι ιδιοτιμές του A είναι μη αρνητικές και ο A είναι μη αρνητικά ορισμένος. Αν ο A είναι και αδιάσπαστος, τότε λόγω του Θεωρήματος 1.4.3, το μηδέν δεν είναι ιδιοτιμή του. Άρα όλες οι ιδιοτιμές του είναι θετικές και ο A είναι θετικά ορισμένος. ■

Στη συνέχεια παρουσιάζονται ορισμένα θεωρήματα, συχνά χωρίς απόδειξη, τα οποία εξασφαλίζουν τη σύγκλιση των βασικών επαναληπτικών μεθόδων κάτω από συγκεκριμένες συνθήκες.

Θεώρημα 3.4.5. *Αν η μέθοδος του Jacobi συγκλίνει, τότε η JOR μέθοδος συγκλίνει για* $0 < \tau \leq 1$.

Απόδειξη. Από την (4.21) οι ιδιοτιμές λ_i του πίνακα B_τ δίνονται από τη σχέση

$$\lambda_i = \tau\mu_i + 1 - \tau \tag{3.79}$$

όπου μ_i είναι οι ιδιοτιμές του B. Αν $\mu_i = re^{i\theta}, \quad r < 1$ και $0 < \tau \leq 1$, τότε η (4.29) δίνει διαδοχικά

$$|\lambda_i|^2 = \tau^2 r^2 + 2\tau r(1-\tau)\cos\theta + (1-\tau)^2 \leq (\tau r + 1 - \tau)^2 < 1.$$

Από την οποία έχουμε ότι $S(B_\tau) < 1$. ■

Θεώρημα 3.4.6. *(Kahan 1958)*

$$S(L_\omega) \geq |\omega - 1|. \tag{3.80}$$

Επίσης αν η SOR συγκλίνει, τότε

$$0 < \omega < 2. \tag{3.81}$$

Απόδειξη. Για τον υπολογισμό των ιδιοτιμών λ_i του L_ω θεωρούμε το χαρακτηριστικό πολυώνυμο του L_ω, $\phi(\lambda) = det(\lambda I - L_\omega)$. Λόγω της (3.57) η προηγούμενη σχέση γράφεται

$$\begin{aligned}\phi(\lambda) &= \det(I - \omega L)^{-1} det[(\lambda + \omega - 1)I - \omega\lambda L - \omega U] \\ &= \det[(\lambda + \omega - 1)I - \omega\lambda L - \omega U].\end{aligned}$$

Ο σταθερός όρος σ του $\phi(\lambda)$, ο οποίος είναι ίσος με το γινόμενο των ιδιοτιμών του L_ω, λαμβάνεται αν τεθεί $\lambda = 0$ στην έκφραση της $\phi(\lambda)$. Συνεπώς

$$\sigma = \prod_{i=1}^{\nu}(-\lambda_i) = det[(\omega-1)I - \omega U] = (\omega-1)^\nu$$

και

$$S(L_\omega) = \max_i |\lambda i| \geq (|\omega-1|^\nu)^{1/\nu} = |\omega-1|.$$

Τέλος, αν $S(L_\omega) < 1$, τότε και $|\omega - 1| < 1$, οπότε λαμβάνεται η (3.81). ■

Αν στη συνέχεια απαιτήσουμε ο πίνακας A να είναι μη ιδιάζων, θα πρέπει λόγω του Θεωρήματος 1.4.3 να υποθέσουμε ότι ο A είναι αδιάσπαστος και ασθενά διαγώνια υπέρτερος. Κάτω από αυτές τις συνθήκες έχουμε το παρακάτω θεώρημα.

Θεώρημα 3.4.7. *Έστω ότι ο A είναι ένας αδιάσπαστος και ασθενά διαγώνιος υπέρτερος πίνακας. Τότε*

1. *Η μέθοδος του Jacobi συγκλίνει και η JOR συγκλίνει για* $0 < \tau \leq 1$.

2. *Η μέθοδος GS συγκλίνει και η EGS συγκλίνει για* $0 < \tau \leq 1$.

3. *Η μέθοδος SOR συγκλίνει για* $0 < \omega \leq 1$ *και η ESOR συγκλίνει για* $0 \leq \omega \leq \tau \leq 1$, $\tau \neq 0$.

Απόδειξη. Αν $S(B) \geq 1$, τότε υπάρχει μία ιδιοτιμή μ του B τέτοια ώστε $|\mu| \geq 1$. Αλλά $det(B - \mu I) = 0$ και $det(I - \mu^{-1}B) = 0$. Είναι φανερό ότι αφού ο A είναι αδιάσπαστος θα είναι και ο $Q = I - \mu^{-1}B$, ο οποίος είναι και ασθενά διαγώνια υπέρτερος επειδή $|\mu^{-1}| \leq 1$. Αλλά από το Θεώρημα 1.4.3 συνεπάγεται ότι $detQ \neq 0$ και έχουμε αντίφαση. Συνεπώς $S(B) < 1$ και η μέθοδος του *Jacobi* συγκλίνει. Έτσι λογω του Θεωρήματος 1.4.5 η (α) ισχύει. Στη συνέχεια ας υποθέσουμε ότι $S(L_\omega) \geq 1$, τότε για κάποια ιδιοτιμή λ του $L_{\tau,\omega}$ έχουμε $|\lambda| \geq 1$. Επιπλέον, $det(L_{\tau,\omega} - \lambda I) = det(H) = 0$, όπου

$$H = I - (\frac{\tau - \omega + \omega\lambda}{\lambda + \tau - 1})L - \frac{\tau}{\lambda + \tau - 1}U.$$

Θέτοντας $\lambda^{-1} = qe^{i\theta}$ με q και θ πραγματικούς έχουμε

$$|\frac{\tau-\omega+\omega\lambda}{\lambda+\tau-1}| = [\frac{(\tau-\omega)^2q^2+2\omega q(\tau-\omega)\cos\theta+\omega^2}{1-2q(1-\tau)\cos\theta+q^2(1-\tau)^2}]^{1/2} \leq \frac{\omega+q(\tau-\omega)}{1-q(1-\tau)}$$

καθόσον $q \leq 1$, $0 \leq \omega \leq \tau$ και $\tau \leq 1$. Αλλά

$$1-\frac{\omega+q(\tau-\omega)}{1-q(1-\tau)} = \frac{(1-q)(1-\omega)}{1-q(1-\tau)} \geq 0$$

συνεπώς

$$|\frac{\tau}{\lambda+\tau-1}| \leq |\frac{\tau+\omega(\lambda-1)}{\lambda+\tau-1}| \leq 1.$$

Το παραπάνω αποτέλεσμα δείχνει ότι ο H είναι ασθενά διαγώνια υπέρτερος. Αλλά αφού ο A είναι αδιάσπαστος, άρα $detH \neq 0$ και έχουμε αντίφαση. Συνεπώς, $S(L_{\tau,\omega}) < 1$ οπότε ισχύει η (c) και για $\tau = \omega$ εύκολα λαμβάνεται η (b). ■

Πόρισμα 3.4.2. *Αν ο A είναι αυστηρά διαγώνια υπέρτερος, τότε ισχύουν τα (a), (β), και (γ) του προηγούμενου θεωρήματος.*

Απόδειξη. Προκύπτει εύκολα από το Πόρισμα 1.4.1. ■

Στην περίπτωση όπου ο A έχει διάφορες ιδιότητες π.χ. θετικά ορισμένος κ.ά. είναι δυνατόν να βρεθούν θεωρήματα σύγκλισης (βλ.[$Young$][$Varga$]) των βασικών επαναληπτικών μεθόδων. Ωστόσο θα περιοριστούμε σε δύο μόνο θεωρήματα, τα οποία είναι χαρακτηριστικά για τις SOR και $SSOR$ μεθόδους.

Θεώρημα 3.4.8. *Αν ο A είναι ένας συμμετρικός πίνακας με θετικά διαγώνια στοιχεία, τότε η SOR μέθοδος συγκλίνει αν και μόνο αν ο A είναι θετικά ορισμένος και $0 < \omega < 2$.*

Απόδειξη. (βλ.[$Young$] σελ. 113-114). ■

Θεώρημα 3.4.9. *Αν ο A είναι συμμετρικός με θετικά διαγώνια στοιχεία, τότε οι ιδιοτιμές του L_ω είναι πραγματικές και μη αρνητικές για κάθε πραγματικό ω. Επίσης, αν ο A είναι θετικά ορισμένος και αν $0 < \omega < 2$, τότε*

$$S(L_\omega) < 1.$$

Απόδειξη. (βλ.[$Young$] σελ. 463). ■

Στη συνέχεια θα προσπαθήσουμε να προσδιορίσουμε τις βέλτιστες τιμές των παραμέτρων ω, τ που υπεισέρχονται στις επαναληπτικές μεθόδους προκειμένου να αυξηθεί η ταχύτητα σύγκλισής τους.

Θεώρημα 3.4.10. *Αν ο A είναι ένας συμμετρικός και θετικά ορισμένος πίνακας, τότε η JOR μέθοδος συγκλίνει αν και μόνον αν*

$$0 < \tau < \frac{2}{M(\hat{A})}. \tag{3.82}$$

Επιπλέον, η $S(B_\tau)$ ελαχιστοποιείται για

$$\tau_0 = \frac{2}{M(\hat{A}) + m(\hat{A})} \tag{3.83}$$

και η αντίστοιχη τιμή της δίνεται από τον τύπο

$$S(B_{\tau_0}) = \frac{P(\hat{A}) - 1}{P(\hat{A}) + 1} \tag{3.84}$$

όπου

$$\hat{A} = D^{-1/2} A D^{1/2}, \quad P(\hat{A}) = \frac{M(\hat{A})}{m(\hat{A})}$$

και $M(\hat{A})$, $m(\hat{A})$ συμβολίζουν τη μεγαλύτερη και τη μικρότερη ιδιοτιμή του $\hat{A}$, αντίστοιχα.

Απόδειξη. Λόγω της υπόθεσης, ο A έχει πραγματικές και θετικές ιδιοτιμές, επιπλέον τα διαγώνια στοιχεία του είναι θετικά και συνεπώς υπάρχει ο $D^{1/2}$. Είναι φανερό τώρα ότι ο $D^{-1}A$ είναι όμοιος με τον $\hat{A} = D^{1/2} A D^{-1/2}$ πράγμα που σημαίνει ότι έχουν τις ίδιες ιδιοτιμές. Επίσης ο $\hat{A}$ είναι συμμετρικός και θετικά ορισμένος (γιατί;). Άρα λόγω του Θεωρήματος 1.2.1 λαμβάνουμε την (3.82) σαν ικανή και αναγκαία συνθήκη σύγκλισης, ενώ εφαρμόζοντας το Θεώρημα 1.2.2 λαμβάνουμε τις (3.83) και (3.84). ■

Από τον ορισμό της ταχύτητας σύγκλισης έχουμε ότι

$$R(B_{\tau_0}) = -\log S(B_{\tau_0}) = -\log \frac{P(\hat{A}) - 1}{P(\hat{A}) + 1} \simeq \frac{2}{P(\hat{A})} \quad \gamma\iota\alpha \quad P(\hat{A}) \gg 1. \tag{3.85}$$

Παρατηρούμε λοιπόν ότι ο αριθμός των επαναλήψεων για τη σύγκλιση της JOR είναι ανάλογος του αριθμού συνθήκης του $\hat{A}$. Το αποτέλεσμα αυτό δείχνει το βαθμό εξάρτησης της αποτελεσματικότητας της JOR από τον αριθμό συνθήκης του πίνακα $\hat{A}$. Προκειμένου όμως να μελετήσουμε την ταχύτητα σύγκλισης των υπολοίπων επαναληπτικών μεθόδων θα πρέπει προηγουμένα να ορίσουμε την κλάση των διατεταγμένων πινάκων με συνέπεια. Για τους πίνακες αυτούς η ταχύτητα σύγκλισης των επαναληπτικών μεθόδων $E(GS)$ και $E(SOR)$ παραμένει σταθερή. Αν ένας πίνακας έχει την κατά ομάδες ($block$) τριδιαγώνια μορφή

$$\begin{bmatrix} D_1 & F_1 & & & & \\ E_2 & D_2 & F_2 & & 0 & \\ & & & & & \\ & & & & & \\ & 0 & & E_{m-1} & D_{m-1} & F_{m-1} \\ & & & & E_m & D_m \end{bmatrix}$$

όπου D_i είναι τετραγωνικοί πίνακες, τότε ο πίνακας αυτός είναι διατεταγμένος με συνέπεια ($consistently \quad ordered$). Είναι δυνατόν όμως να ορίσουμε και μία γενικότερη κλάση πινάκων από την προηγούμενη. Οι πίνακες που ανήκουν στην κλάση αυτή θα λέμε ότι έχουν την 'Ιδιότητα A'.

Θεώρημα 3.4.11. *Ένας πίνακας A έχει την Ιδιότητα A αν και μόνον αν ο A είναι ένας διαγώνιος πίνακας ή υπάρχει ένας μεταθετικός πίνακας P τέτοιος ώστε $P^{-1}AP$ να έχει τη μορφή*

$$A' = P^{-1}AP = \begin{bmatrix} D_1 & H \\ K & D_2 \end{bmatrix}. \tag{3.86}$$

όπου οι D_1 και D_2 είναι τετραγωνικοί διαγώνιοι πίνακες.

Θεώρημα 3.4.12. *Αν ο πίνακας A είναι διατεταγμένος με συνέπεια, ο A έχει την Ιδιότητα A.*

Θεώρημα 3.4.13. *Ένας πίνακας A έχει την Ιδιότητα A αν και μόνον αν υπάρχει μεταθετικός πίνακας P τέτοιος ώστε ο $A' = P^{-1}AP$ να είναι διατεταγμένος με συνέπεια.*

Θεώρημα 3.4.14. *Αν ο πίνακας A είναι διατεταγμένος με συνέπεια, τότε*

1.

$$(\lambda + \omega - 1)^2 = \omega^2 \mu^2 \lambda \tag{3.87}$$

όπου λ είναι μιά ιδιοτιμή του L_ω και μ μία ιδιοτιμή του B και

2. Αν μ είναι μία ιδιοτιμή του B με πολλαπλότητα p, τότε η $-\mu$ είναι επίσης ιδιοτιμή του B με πολλαπλότητα p.

Πόρισμα 3.4.3. *Αν ο πίνακας A είναι διατεταγμένος με συνέπεια, τότε*

$$\lambda = \mu^2, \tag{3.88}$$

όπου λ είναι μία ιδιοτιμή του L. Επιπλέον,

$$S(L) = [S(B)]^2 \quad \kappa\alpha\iota \quad R(L) = 2R(B). \tag{3.89}$$

Από το Πόρισμα 1.4.3 παρατηρούμε ότι αν ο πίνακας A είναι διατεταγμένος με συνέπεια, τότε η ταχύτητα σύγκλισης της μεθόδου $Gauss - Seidel$ είναι διπλάσια από εκείνη της μεθόδου $Jacobi$. Στη συνέχεια δίνονται δύο θεωρήματα τα οποία αποτελούν τη βάση της θεωρίας της SOR μεθόδου.

Θεώρημα 3.4.15. *Αν ο A είναι ένας διατεταγμένος με συνέπεια πίνακας με μη μηδενικά διαγώνια στοιχεία, τέτοιος ώστε ο B να έχει πραγματικές ιδιοτιμές, τότε*

$$S(L_\omega) < 1 \tag{3.90}$$

αν και μόνον αν $0 < \omega < 2$ και $S(B) < 1$.

Απόδειξη. (βλ.[$Young$],[$Varga$]). ■

Ο προσδιορισμός της βέλτιστης τιμής της ω στην SOR μέθοδο δίνεται από το παρακάτω θεώρημα.

Θεώρημα 3.4.16. *Έστω ότι ο A είναι ένας διατεταγμένος με συνέπεια πίνακας με μη μηδενικά διαγώνια στοιχεία τέτοιος ώστε ο πίνακας B να έχει πραγματικές ιδιοτιμές και $\mu = S(B) < 1$. Αν ω_b ορίζεται από τον τύπο*

$$\omega_b = \frac{2}{1 + \sqrt{1 - \bar{\mu}^2}}, \tag{3.91}$$

τότε

$$S(L_{\omega_b}) = \omega_b - 1. \tag{3.92}$$

Απόδειξη. (βλ[*Young*]). ■

Η μελέτη της $S(L_\omega)$ σαν συνάρτηση της ω φαίνεται στο σχήμα (4.2). Παρατηρούμε λοιπόν από το Σχήμα 4.2, ότι όσο η ω αυξάνει από το 0 στο 1, η $S(L_\omega)$ ελαττώνεται γρηγορότερα μέχρις ότου η ω είναι κοντά στο ω_b. Στο σημείο αυτό η ελάττωση είναι πάρα πολύ απότομη. Καθώς η ω συνεχίζει να αυξάνει, η $S(L_\omega)$ αυξάνει γραμμικά στην τιμή 1 για $\omega = 2$. Είναι προτιμότερο λοιπόν να υπερεκτιμούμε παρά να υποεκτιμούμε τη βέλτιστη τιμή της παραμέτρου ω. Όσο αφορά την ταχύτητα σύγκλισης της μεθόδου, αυτή είναι ταχύτερη κατά μία τάξη μεγέθους από εκείνη της JOR για τη βέλτιστη τιμή της ω. Πράγματι, αποδεικνύεται ότι (με τις υποθέσεις του Θεωρήματος 1.4.16)

$$R(L_{\omega_b}) \simeq \frac{4}{\sqrt{P(\hat{A})}}. \quad (\text{γιατί;}) \tag{3.93}$$

Από την (3.93) συνεπάγεται ότι ο αριθμός των επαναλήψεων για την σύγκλιση της SOR είναι ανάλογος προς την τετραγωνική ρίζα του αριθμού συνθήκης του πίνακα A. Η SOR λοιπόν επιτυγχάνει μία σημαντική βελτίωση με την κατάλληλη εκλογή της παραμέτρου ω σε σχέση με τις μεθόδους J,JOR και GS. Όσο αφορά τις μεθόδους $SSOR$ και PSD αποδεικνύεται ότι (i) η ταχύτητά τους είναι ανάλογη με εκείνη της SOR για μία κατηγορία πινάκων και (ii) $R(\Delta_{\tau_0,\omega_0}) \simeq 2R(F_{\omega_0})$, δηλαδή ότι η ταχύτητα σύγκλισης της PSD είναι διπλάσια της $SSOR$.

3.5 Υπολογιστική πολυπολοκότητα της μεθόδου του *Jacobi*

Στην παρούσα ενότητα θα υπολογίσουμε τον αριθμό των πράξεων που χρειάζονται σε μία επανάληψη για τη μέθοδο του $Jacobi$. Ανακαλώντας τον τύπο της μεθόδου του $Jacobi$ έχουμε

$$u_i^{(n+1)} = \hat{b}_i + \sum_{\substack{j=1 \\ j\neq i}}^{\nu} \hat{a}_{ij} u_j^{(n)}, \quad i = 1(1)\nu, \tag{3.94}$$

όπου

$$\hat{b}_i = \frac{b_i}{a_{ii}}, \quad i = 1(1)\nu \quad \kappa\alpha\iota \quad \hat{a}_{ij} = \frac{a_{ij}}{a_{ii}}, \quad i = 1(1)\nu, \quad j = 1(1)\nu, \quad j \neq i.$$

Συνεπώς ο υπολογισμός των συντελεστών b_i και a_{ij} απαιτεί

$$\nu^2 \quad \text{διαιρέσεις.}$$

Στη συνέχεια παρατηρούμε από τις (4.30) ότι για κάθε επανάληψη απαιτούνται $\nu - 1$ πολλαπλασιασμοί και $\nu - 1$ προσθαφαιρέσεις για κάθε συνιστώσα (για κάθε i). Άρα για τις ν συνιστώσες του $u_i^{(n+1)}$ έχουμε

$$\nu(\nu-1) \quad \text{προσθαφαιρέσεις} \quad \text{και} \quad \nu(\nu-1) \quad \text{πολ/μούς.}$$

Αν υποθέσουμε ότι απαιτούνται κ επαναλήψεις για τη σύγκλιση της μεθόδου του *Jacobi*, τότε έχουμε συνολικά

$$\kappa\nu(\nu-1) = \kappa\nu^2 - \kappa\nu \quad \text{προσθαφαιρέσεις}$$

$$\kappa\nu(\nu-1) + \nu^2 = (\kappa+1)^2 - \kappa\nu \quad \text{πολ/μούς.}$$

Αγνοώντας τον όρο $-\kappa\nu$ στους ανωτέρω τύπους παρατηρούμε ότι η μέθοδος του *Jacobi* θα απαιτεί λιγότερες πράξεις από τη μέθοδο απαλοιφής του *Gauss* αν

$$\kappa + 1 \leq \frac{\nu}{3}. \tag{3.95}$$

Πράγματι, αν ικανοποιείται η (4.32), η μέθοδος του *Jacobi* απαιτεί λιγότερες από (βλ.(4.32))

$$(\frac{\nu}{3} - 1)\nu^2 = \frac{\nu^3}{3} - \nu^2 \quad \text{προσθαφαιρέσεις}$$

και

$$\frac{\nu}{3} \cdot \nu^2 = \frac{\nu^3}{3} \quad \text{πολ/μούς.}$$

Στους ανωτέρω υπολογισμούς υποθέσαμε ότι όλα τα στοιχεία a_{ij} και b_i είναι διάφορα του μηδενός. Ωστόσο όμως στην πράξη παρουσιάζεται το φαινόμενο όπου ο A είναι αραιός πίνακας. Ενώ όταν χρησιμοποιούμε άμεσες μεθόδους είναι δύσκολο να λάβουμε υπόψη τη δομή του πίνακα, με τις επαναληπτικές μεθόδους αυτό είναι εφικτό (βλ.π.χ.(4.30)). Έτσι όλοι οι συντελεστές a_{ij} που είναι αρχικά μηδέν διατηρούνται σε όλη τη διάρκεια των επαναλήψεων. Αν τώρα p είναι ο αριθμός των μη μηδενικών στοιχείων του A, τότε η μέθοδος του *Jacobi* χρειάζεται αρχικά p διαιρέσεις για τον υπολογισμό των a_{ij} και

b_i. Επίσης, αν υποθέσουμε ότι οι μηδενικοί συντελεστές είναι ομοιόμορφα κατανεμημένοι σε κάθε εξίσωση, τότε για κάθε μία συνιστώσα απαιτούνται

$$\frac{p}{\nu} - 1 \quad \text{προσθαφαιρέσεις} \tag{3.96}$$

και

$$\frac{p}{\nu} - 1 \quad \text{πολ/μοί.}$$

Άρα για τον υπολογισμό των ν συνιστωσών έχουμε

$$p - \nu \quad \text{προσθαφαιρέσεις} \quad \text{και} \quad p - \nu \quad \text{πολ/μούς.}$$

Τέλος, μετά από κ επαναλήψεις θα έχουμε ένα σύνολο από

$$\kappa(p - \nu) = \kappa p - \kappa\nu \quad \text{προσθαφαιρέσεις} \tag{3.97}$$

και

$$\kappa(p - \nu) + p = \kappa p + p - \kappa\nu \quad \text{πολ/μούς} \quad \text{και} \quad \text{διαιρέσεις.}$$

Ο αριθμός $a = p/\nu^2 \leq 1$ εκφράζει την πυκνότητα του A. Αν αγνοήσουμε πάλι τις αρχικές p διαιρέσεις και τον όρο $-\kappa\nu$ στις (4.34) έχουμε ότι η μέθοδος του *Jacobi* απαιτεί περίπου

$$\kappa p = \kappa a \nu^2 \quad \text{προσθαφαιρέσεις}$$

και

$$\kappa p = \kappa a \nu^2 \quad \text{πολ/μούς.}$$

Συνεπώς, η μέθοδος του *Jacobi* χρειάζεται λιγότερες πράξεις από τη μέθοδο απαλοιφής του *Gauss* αν

$$\kappa \leq \frac{\nu}{3a}, \quad 0 < a < 1. \tag{3.98}$$

Από τον τύπο (4.35) παρατηρούμε ότι όσο η πυκνότητα του A είναι μικρή τόσο η πιθανότητα να απαιτούνται λιγότερες πράξεις στη μέθοδο του *Jacobi*. Ανάλογο συμπέρασμα ισχύει και για τις υπόλοιπες επαναληπτικές μεθόδους.

3.6 Ημι-Επαναληπτικές Μέθοδοι

Ας θεωρήσουμε πάλι την ακολουθία των διανυσμάτων που προκύπτουν από τη βασική επαναληπτική μέθοδο

$$u^{(n+1)} = Gu^{(n)} + k, \quad n = 0, 1, 2, \ldots \tag{3.99}$$

Είναι γνωστό από τη θεωρία του αθροίσματος των ακολουθιών ότι συχνά είναι δυνατόν να αναπτυχθεί μιά νέα ακολουθία διανυσμάτων $v^{(0)}, v^{(1)}, \ldots$ τέτοια ώστε η νέα ακολουθία να συγκλίνει σε περίπτωση που δεν συγκλίνει η αρχική ή η νέα ακολουθία να συγκλίνει ταχύτερα. Η νέα ακολουθία μπορεί να οριστεί σαω ένας γραμμικός συδυασμός της αρχικής ακολουθίας. Έτσι ορίζουμε την ακολουθία

$$v^{(n)} = \sum_{k=1}^{n} a_k(n) u^k, \quad n = 0, 1, 2, \ldots \tag{3.100}$$

την οποία καλούμε ημι-επαναληπτική (Semi-Iterative (SI)) μέθοδο σε σχέση με την (3.99). Είναι ημι-επαναληπτική γιατί πρώτα εκτελούμε την επανάληψη με την (3.99) και στη συνέχεια συνδυάζουμε αλγεβρικά αυτές τις επαναλήψεις χρησιμοποιώντας τη (3.100). Μπορούμε π.χ. να θέσουμε

$$a_k(n) = \frac{1}{n+1}$$

οπότε τα διανύσματα $v^{(n)}$ είναι οι μέσες τιμές των $u^{(n)}$, πράγμα που αντιστοιχεί ακριβώς στο άθροισμα του *Cesaro* $(C, 1)$. Ο αντικειμενικός σκοπός μας είναι να προσδιορίσουμε τους σταθερούς συντελεστές $a_k(n)$ της (3.100) έτσι ώστε η ταχύτητα σύγκλισης της νέας επαναληπτικής μεθόδου να είναι μεγαλύτερη από εκείνη της (3.99). Καταρχήν αν υποτεθεί ότι το αρχικό διάνυσμα $u^{(0)}$ της (3.99) συμπίπτει με την αληθή λύση u του συστήματος τότε θα έχουμε

$$u^{(n)} = u, \quad n = 0, 1, 2, \ldots$$

Στην περίπτωση αυτή απαιτούμε και

$$v^{(n)} = u, \quad n = 0, 1, 2, \ldots$$

οπότε από την (3.100) λαμβάνουμε

$$\sum_{k=1}^{n} a_k(n) = 1, \quad n = 0, 1, 2, \ldots \tag{3.101}$$

Προκειμένου να αναλύσουμε την ταχύτητα σύγκλισης μιας ημι-επαναληπτικής μεθόδου όριζουμε τα διάνύσματα

$$\epsilon^{(n)} = u^{(n)} - u \quad \text{και} \quad \eta^{(n)} = \upsilon^{(n)} - u \tag{3.102}$$

Από τις (3.100) και (3.101) έχουμε

$$\eta^{(n)} = \sum_{k=0}^{n} a_k(n)u^k - \sum_{k=0}^{n} a_k(n)u = \sum_{k=0}^{n} a_k(n)\epsilon^{(n)}. \tag{3.103}$$

Επειδή όμως

$$\epsilon^{(k)} = G^k\epsilon^{(0)}$$

η (3.103) δίνει

$$\eta^{(n)} = \Big(\sum_{k=0}^{n} a_k(n)G^k\Big)\epsilon^{(0)} \tag{3.104}$$

Ορίζοντας το πολυώνυμο $P_n(x)$ από τη σχέση

$$P_n(x) = \sum_{k=0}^{n} a_k(n)x^k, \quad n = 0, 1, 2, \ldots \tag{3.105}$$

η (3.104) γράφεται σαν

$$\eta^{(n)} = P_n(G)\epsilon^{(0)} = P_n(G)\eta^{(0)} \tag{3.106}$$

όπου $P_n(G)$ είναι ένα πολυώνυμο του πίνακα G. Ο μοναδικός περιορισμός για τα πολυώνυμα $P_n(x)$ είναι $P_n(1) = 1$ λόγω της (3.101). Επίσης από την (3.106) έχουμε

$$||\eta^{(n)}||_2 = ||P_n(G)\epsilon^{(0)}||_2 \leq ||P_n(G)||_2||\eta^{(0)}||_2. \tag{3.107}$$

Από την (3.107) παρατηρούμε ότι ελαττώνοντας την $||P_n(G)||$, ελαττώνεται η $||\eta^{(n)}||$, δηλαδή τα σφάλματα της ημιεπαναληπτικής μεθόδου, έτσι οδηγούμαστε στο πρόβλημα της ελαχιστοποίησης της ποσότητας $||P_n(G)||$ όπου $P_n(G) = \alpha_0(n)I + \alpha_1(n)G + \ldots + \alpha_n(n)(G)^n$ με $P_n(1) = 1$. Στο σημείο αυτό υποθέτουμε ότι ο πίνακας G έχει πραγματικές ιδιοτιμές λ_i τέτοιες ώστε

$$\alpha \leq \lambda_i \leq \beta < 1 \tag{3.108}$$

επομένως

$$||P_n(G)||_2 = S(P_n(G)) = \max_{\lambda_i} |P_n(\lambda_i)| \leq \max_{\alpha \leq \lambda \leq \beta} |P_n(\lambda)|. \tag{3.109}$$

Είναι φανερό ότι έχουμε το ακόλουθο πρόβλημα ελαχιστοποίησης

$$\min_{P_n(1)=1} \{ \max_{\alpha \le \lambda \le \beta < 1} |P_n(\lambda)| \} \tag{3.110}$$

όπου $P_n(\lambda)$ είναι ένα πραγματικό πολυώνυμο. Η λύση του παραπάνω προβλήματος είναι γνωστή και δίνεται με τη χρήση των πολυωνύμων του Chebyshev. Πριν όμως προχωρήσουμε στη λύση του παραπάνω προβλήματος απεικονίζουμε το διάστημα $\alpha \le \lambda \le \beta$ στο διάστημα $-1 \le \gamma \le 1$ διαμέσου του μετασχηματισμού

$$\gamma = \frac{2\lambda - (\beta + \alpha)}{\beta - \alpha} \tag{3.111}$$

Ορίζοντας το πολυώνυμο $Q_n(\gamma)$ από τη σχέση

$$Q_n(\gamma) = P_n\Big(\frac{(\beta - \alpha)\gamma + \beta + \alpha}{\beta - \alpha}\Big)$$

έχουμε ότι

$$Pn(\lambda) = Q_n\Big(\frac{2\lambda - (\beta + \alpha)}{\beta - \alpha}\Big) \tag{3.112}$$

άρα

$$\max_{\alpha \le \lambda \le \beta} |P_n(\lambda)| = \max_{-1 \le \gamma \le 1} |Q_n(\gamma)| \tag{3.113}$$

με $P_n(1) = Q_n(z) = 1$, $z = \gamma(1) = \frac{2-(\alpha+\beta)}{\beta-\alpha} > 1$ αφού $2-(\alpha+\beta) > 0$, $\beta - \alpha > 0$ και $[2 - (\alpha + \beta)] - (\beta - \alpha) = 2 - 2\beta > 0$. Το πρόβλημα μας δηλαδή τώρα είναι να βρεθεί το πολυώνυμο $Q_n(\gamma)$ ν-στού βαθμού ή μικροτέρου τέτοιο ώστε $Q_n(z) = 1$ και η ποσότητα

$$\max_{-1 \le \gamma \le 1} |Q_n(\gamma)| \tag{3.114}$$

να γίνει ελάχιστη. Η λύση στο παραπάνω πρόβλημα δίνεται από το ακόλουθο θεώρημα.

Θεώρημα 3.6.1. *Έστω n ένας σταθερός μη αρνητικός ακέραιος και z οποιοσδήποτε σταθερός πραγματικός αριθμός τέτοιος ώστε $z > 1$. Αν θέσουμε*

$$P_n(x) = T_n(x)/T_n(z) \tag{3.115}$$

όπου $T_n(x)$ είναι το πολυώνυμο Chebyshev n-στού βαθμού που δίνεται από την

$$\begin{aligned} T_n(x) &= \cos(ncos^{-1}x) = \frac{1}{2}\Big[(x+\sqrt{x^2-1})^n + (x+\sqrt{x^2-1})^{-n})\Big] \\ &= \cosh(ncosh^{-1}x) \end{aligned}$$

τότε

$$P_n(z) = 1$$

$$\max_{-1\le x\le 1} |P_n(x)| = \frac{1}{T_n(z)}. \tag{3.116}$$

Επιπλέον, αν $Q(x)$ είναι οποιοδήποτε πολυώνυμο βαθμού n ή μικροτέρου τέτοιο ώστε $Q(z) = 1$ και

$$\max_{-1\le x\le 1} |Q(x)| \le \max_{-1\le x\le 1} |P_n(x)|$$

τότε

$$Q(x) \equiv P(x)$$

Απόδειξη. (βλ. [Young] σελ. 303) ■

Από το προηγούμενο θεώρημα έχουμε ότι το πολυώνυμο $Q_n(\gamma)$ που ελαχιστοποιεί την ποσότητα (3.114) δίνεται απο την

$$Q_n(\gamma) = \frac{T_n(\gamma)}{T_n(z)} \tag{3.117}$$

επίσης

$$\max_{-1\le \gamma\le 1} |Q_n(\gamma)| = \frac{1}{T_n(z)} = 1\Big/T_n\Big(\frac{2-(\beta+\alpha)}{\beta-\alpha}\Big). \tag{3.118}$$

Επομένως έχουμε

$$P_n(\lambda) = Q_n\Big(\frac{2\lambda-(\beta+\alpha)}{\beta-\alpha}\Big) = T_n\Big(\frac{2\lambda-(\beta+\alpha)}{\beta-\alpha}\Big)\Big/T_n\Big(\frac{2-(\beta+\alpha)}{\beta-\alpha}\Big) \tag{3.119}$$

Επειδή $z > 1$ και $T_n(z) > 1$ έπεται ότι η ημιεπαναληπτική μέθοδος συγκλίνει ακόμη και αν η βασική μέθοδος (3.99) δεν συγκλίνει, αφού

$$||P_n(G)||_2 = 1/T_n(z) < 1 \tag{3.120}$$

Έστω $T_0(x) = 1$, $T_1(x) = x$, $T_2(x) = 2x^2 - 1$ έχουμε

$$P_0(\mu) = \frac{T_0(\gamma)}{T_0(z)} = 1$$

άρα $a_0(0) = 1$, επίσης

$$P_1(\mu) = \frac{T_1(\gamma)}{T_1(z)} = \frac{2\mu - (\alpha + \beta)}{2 - (\alpha + \beta)} = a_1(1)\mu + a_0(1)$$

άρα

$$a_0(1) = -[\beta + \alpha]/[2 - (\beta + \alpha)]$$

και

$$a_1(1) = 2/[2 - (\beta + \alpha)].$$

Ανάλογα έχουμε

$$\begin{aligned} P_2(\mu) &= \Big[2\Big(\frac{2\mu - (\beta + \alpha)}{\beta - \alpha}\Big)^2 - 1\Big] \Big/ \Big[2\Big(\frac{2 - (\beta + \alpha)}{\beta - \alpha}\Big)^2 - 1\Big] \\ &= a_2(2)\mu^2 + a_1(2)\mu + a_0(2) \end{aligned}$$

οπότε βρίσκουμε

$$\begin{aligned} a_2(0) &= \frac{(\alpha + \beta)^2 + 4\alpha\beta}{(\alpha + \beta)^2 + 8(1 - \alpha - \beta) + 4\alpha\beta} \\ a_2(1) &= \frac{-8(\alpha + \beta)^2}{(\alpha + \beta)^2 + 8(1 - \alpha - \beta) + 4\alpha\beta} \\ a_2(2) &= \frac{8}{(\alpha + \beta)^2 + 8(1 - \alpha - \beta) + 4\alpha\beta} \end{aligned}$$

Συνεχίζοντας θα μπορούσαμε να προσδιορίσουμε τα $a_k(n)$ για οποιδήποτε n. Ωστόσο υπάρχει ένας άλλος τρόπος ο οποίος αναπτύσει μιά σχέση μεταξύ των $v^{(n+1)}, v^{(n)}$ και $v^{(n-1)}$ έτσι ώστε να είναι δυνατόν να υπολογιστεί το $v^{(n+1)}$χωρίς να απαιτείται ο υπολογισμός του $u^{(n+1)}$. Ο τρόπος αυτός χρησιμοποιεί την ακόλουθη αναδρομική σχέση των πολυωνύμων *Chebyshev*

$$T_{n+1}(x) = 2T_n(x) - T_{n-1}(x), \quad n \geq 1 \tag{3.121}$$

Από την (3.119) έχουμε

$$P_n(G) = T_n(\frac{2G - (\beta + \alpha)I}{\beta - \alpha})/T_n(z) \tag{3.122}$$

συνεπώς από την (3.106) έχουμε

$$\eta^{(n+1)} = P_{n+1}(G)\epsilon^{(0)} = \Big[T_{n+1}\Big(\frac{2G-(\beta+\alpha)I}{\beta-\alpha}\Big)\Big/T_{n+1}(z)\Big]\epsilon^{(0)} \tag{3.123}$$

ή λόγω της (3.121)

$$\eta^{(n+1)} = \Big[2\Big(\frac{2G-(\beta+\alpha)I}{\beta-\alpha}\Big)T_n\Big(\frac{2G-(\beta+\alpha)I}{\beta-\alpha}\Big) - T_{n-1}\Big(\frac{2G-(\beta+\alpha)I}{\beta-\alpha}\Big)\Big]\Big/T_{n-1}(z)\epsilon^{(0)}. \tag{3.124}$$

Λόγω όμως της (3.123) έχουμε

$$\eta^{(n)} = \Big[T_n\Big(\frac{2G-(\beta+\alpha)I}{\beta-\alpha}\Big)\Big/T_n(z)\Big]\epsilon^{(0)}$$

και

$$\eta^{(n-1)} = \Big[T_{n-1}\Big(\frac{2G-(\beta+\alpha)I}{\beta-\alpha}\Big)\Big/T_{n-1}(z)\Big]\epsilon^{(0)}$$

έτσι η (3.124) γράφεται

$$\eta^{(n+1)} = 2\Big[\frac{2G-(\beta+\alpha)I}{\beta-\alpha}\Big]\frac{T_n(z)}{T_{n+1}(z)}\eta^{(n)} - \frac{T_{n-1}(z)}{T_{n+1}(z)}\eta^{(n-1)}. \tag{3.125}$$

Επομένως λόγω της (3.102) έχουμε

$$\begin{aligned} v^{(n+1)} =& 2\Big(\frac{2G-(\beta-\alpha)I}{\beta-\alpha}\Big)\frac{T_n(z)}{T_{n+1}(z)}v^{(n)} - \frac{T_{n-1}(z)}{T_{n+1}(z)}v^{(n-1)} \\ &+ \Big[I - 2\Big(\frac{2G-(\beta-\alpha)I}{\beta-\alpha}\Big)\frac{T_n(z)}{T_{n+1}(z)} + \frac{T_{n-1}(z)}{T_{n+1}(z)}\Big]u \end{aligned} \tag{3.126}$$

Αλλά ο τελευταίος όρος μέσα στις παρενθέσεις γράφεται διαδοχικά

$$2\Big[zI-\Big(\frac{2G-(\beta-\alpha)I}{\beta-\alpha}\Big)\Big]\frac{T_n(z)}{T_{n+1}(z)}u = \frac{4(I-G)}{\beta-\alpha}\frac{T_n(z)}{T_{n+1}(z)} = \frac{4}{\beta-\alpha}\frac{T_n(z)}{T_{n+1}(z)}k \tag{3.127}$$

επειδή $u = Gu + k$. Επομένως για $n \geq 1$ έχουμε

$$v^{(n+1)} = 2\Big[\frac{2}{\beta-\alpha}G - \frac{\beta+\alpha}{\beta-\alpha}I\Big]\frac{T_n(z)}{T_{n+1}(z)}v^{(n)} - \frac{T_{n-1}(z)}{T_{n+1}(z)}v^{(n-1)} + \frac{4}{\beta-\alpha}\frac{T_n(z)}{T_{n+1}(z)}k \tag{3.128}$$

ή

$$v^{(n+1)} = \hat{\rho}_{n+1}\Big[\bar{\rho}(Gv^{(n)}+k) + (1-\bar{\rho})v^{(n)}\Big] + (1-\hat{\rho}_{n+1})v^{(n-1)} \tag{3.129}$$

όπου

$$\bar{\rho} = \frac{2}{2-(\beta-\alpha)}$$

και

$$\hat{\rho} = \frac{2zT_n(z)}{T_{n+1}(z)}, \quad n \geq 1. \tag{3.130}$$

Από την (3.122) έχουμε

$$P_1(G) = \frac{2G-(\beta+\alpha)I}{2-(\beta+\alpha)}$$

επειδή $T_1(x) = x$, ενώ από την (3.128)

$$v^{(1)} - u = \frac{2G-(\beta+\alpha)I}{2-(\beta+\alpha)}(u^{(0)} - u)$$

Συνεπώς

$$v^{(1)} = \frac{2G-(\beta+\alpha)I}{2-(\beta+\alpha)}u^{(0)} + \Big[I - \frac{2G-(\beta+\alpha)I}{2-(\beta+\alpha)}\Big]u.$$

ή

$$v^{(1)} = \frac{2G-(\beta+\alpha)I}{2-(\beta+\alpha)}u^{(0)} + \frac{2(I-G)}{2-(\beta+\alpha)}u$$

ή

$$v^{(1)} = \bar{\rho}(Gu^{(0)} + k) + (1-\bar{\rho})u^{(0)}. \tag{3.131}$$

Συνοψίζοντας έχουμε ότι οι (3.129) και (3.131) δίνονται από τον τύπο

$$v^{(n+1)} = \hat{\rho}\Big[\bar{\rho}(Gu^{(n)} + k) + (1-\bar{\rho})u^{(n)}\Big] + (1-\hat{\rho}_{n+1})v^{(n-1)} \tag{3.132}$$

όπου

$$\bar{\rho} = \frac{2}{2-(\alpha+\beta)} \tag{3.133}$$

$$\hat{\rho}_1 = 1$$

και

$$\hat{\rho}_n = \frac{2zT_{n-1}(z)}{T_n(z)}, \quad n = 2, 3, \cdots \tag{3.134}$$

αφού $v^{(0)} = u^{(0)}$. Επειδή όμως $T_{n+1}(z) = 2zT_n(z) - T_{n-1}(z)$ άρα

$$\hat{\rho}_{n+1} = (1 - \frac{1}{4z^2}\hat{\rho}_n)^{-1}, \quad n = 2, 3, \cdots \tag{3.135}$$

επίσης

$$\hat{\rho}_2 = 2zT_1(z)/T_2(z) = 2z^2(2z^2-1)^{-1}.$$

Από τα παραπάνω έχουμε ότι η ακολουθία των παραμέτρων $\{\hat{\rho}_n\}, \quad n = 1, 2, \cdots$ μπορεί να γραφεί διαφορετικά, ως εξής:

$$v^{(n+1)} = \hat{\rho}_{n+1}[\bar{\rho}(Gv^{(n)} + k) + (1-\bar{\rho})v^{(n)}] + (1-\rho_{n+1})v^{(n-1)} \quad (3.136)$$

όπου

$$\bar{\rho} = \frac{2}{2-(\alpha+\beta)} \quad (3.137)$$

$$\rho_1 = 1, \quad \rho_2 = (1-\frac{1}{2}\sigma^2)^{-1}$$

$$\rho_{n+1} = (1-\frac{1}{4}\sigma^2)^{-1}, \quad n \geq 2 \quad (3.138)$$

και

$$\sigma = \frac{\beta-\alpha}{2-(\beta+\alpha)} \quad (3.139)$$

Τέλος αποδεικνύεται ότι (βλ. $Varga$, 1962)

$$\lim_{n\to\infty} \rho_n = \rho_\infty = \frac{2}{1+\sqrt{1-\sigma^2}} \quad (3.140)$$

Στη συνέχεια θα εξετάσουμε τη σύγκλιση της ημιεπαναληπτικής μεθόδου (3.136). Η φασματική ακτίνα της ημιεπαναληπτικής μεθόδου είναι

$$S(P_n(G)) = \max_{\alpha\leq\lambda\leq\beta} P_n(\lambda) = 1/T_n(z). \quad (3.141)$$

Επειδή όμως

$$T_n(x) = \cos(n\cos^{-1}x) = \frac{1}{2}[(x+\sqrt{x^2-1})^n + (x-\sqrt{x^2-1})^{-n}]$$

έπεται ότι

$$T_n(z) = \frac{1+\tau^n}{2\tau^{n/2}}$$

όπου

$$\tau \equiv \frac{1-\sqrt{1-\sigma^2}}{1+\sqrt{1+\sigma^2}} \quad (3.142)$$

Άρα

$$S(P_n(G)) = \frac{2\tau^{n/2}}{1+\tau^n} \quad (3.143)$$

Συνεπώς η μέση ταχύτητα σύγκλισης της ημιεπαναληπτικής μεθόδου είναι:

$$R_n(P_n(G)) = -\frac{1}{n}\log S(P_n(G)) = -\frac{1}{2}\log\tau - \frac{1}{n}\log(\frac{2}{1+\tau^n}) \tag{3.144}$$

και η ασυμπτωτική ταχύτητα σύγκλισης είναι

$$R_\infty(P_n(G)) = -\frac{1}{2}\log\tau. \tag{3.145}$$

Από τις (3.140) και (3.142) έχουμε ότι $\tau = \rho_\infty - 1$ άρα

$$\lim_{n\to\infty}[S(P_n(G))]^{1/n} = (\rho_\infty - 1)^{1/2} \quad \text{και} \quad R_\infty(P_n(G)) = -\frac{1}{2}\log(\rho_\infty - 1)$$

Στη συνέχεια συγκρίνουμε την ημιεπαναληπτική μέθοδο (3.136) με την επιταχυντική μορφή της βασικής επαναληπτικής, δηλαδή την

$$u^{(n+1)} = \bar{\rho}(Gu^{(n)} + k) + (1-\bar{\rho})u^{(n)} \tag{3.146}$$

όπου

$$\bar{\rho} = \frac{2}{2-(\beta+\alpha)} \tag{3.147}$$

και

$$S(G_{\bar{\rho}}) = \frac{\beta-\alpha}{2-\beta-\alpha)} = \sigma \text{ (γιατί;)} \tag{3.148}$$

οπότε

$$R_\infty(G_{\bar{\rho}}) = -\log\sigma \tag{3.149}$$

Αλλά

$$-\log\sigma \sim 1-\sigma, \quad \sigma \to 1^-$$

και

$$R_\infty(P_n(G)) = -\frac{1}{2}\log\tau \sim \sqrt{1-\sigma^2} \sim \sqrt{2}\sqrt{1-\sigma} \tag{3.150}$$

για $\sigma \to 1^-$. Συνδυάζοντας τις (3.150) και (3.149) βρίσκουμε

$$R_\infty(P_n(G)) \sim \sqrt{2}\sqrt{R_\infty(G_{\bar{\rho}})} \tag{3.151}$$

Συνεπώς για σ κοντά στη μονάδα η ημιεπαναληπτική μέθοδος είναι κατά μία τάξη μεγέθους ταχύτερη από την (3.146). Σαν παράδειγμα ας θεωρήσουμε την περίπτωση όπου $\beta = -\alpha = 0.99$, τότε $\sigma = 0.99$ και $\tau \simeq 0.753$. Ο αριθμός των επαναλήψεων για την ελάττωση της $norm$ του διανύσματος σφάλματος μικρότερης του 10^{-6} χρησιμοποιώντας την (3.146) είναι περίπου 1375, ενώ είναι 98 για την ημιεπαναληπτική μέθοδο.

3.6.1 Μέθοδοι Μεταβλητής Παρεκτροπής (Variable Extrapolation)

Στην προηγούμενη παράγραφο έχει περιγραφεί πως είναι δυνατόν να κατασκευασθεί μια πλέον αποτελεσματική επαναληπτική μέθοδος με τη χρήση των ημιεπαναληπτικών τεχνικών. Επίσης αναφέρουμε ότι στη νέα μέθοδο κάθε διάνυσμα $v^{(n+1)}$ απαιτεί τον υπολογισμό των δύο προηγουμένων διανυματων $v^{(n)}$ και $v^{(n-1)}$. Αν ο υπολογιστής είναι περιορισμένης χωρητικότητας μνήμης, τότε μπορούμε να θεωρήσουμε έναν άλλο τύπο για την επιτάχυνση της βασικής επαναληπτικής μεθόδου (1). Αυτό μπορεί να επιτευχθεί επιτρέποντας να μεταβάλεται η παράμετρος $\overline{\rho}$ στην επιταχυντική μορφή της βασικής επαναληπτικής μεθόδου, δηλαδή την

$$u^{(n+1)} = \overline{\rho}(Gu^{(n)} + k) + (1 - \overline{\rho})u^{(n)} \tag{3.152}$$

Έτσι λοιπόν οδηγούμαστε στην νέα επιταχυντική μορφή

$$u^{(n+1)} = \theta_{n+1}(Gu^{(n)} + k) + (1 - \theta_{n+1})u^{(n)}, \tag{3.153}$$

όπου $\theta_1, \theta_2, \cdots$ είναι οι παράμετροι επανάληψης.

Η ιδέα αυτή έχει παρουσιαστεί από τον Richardson[1910] και έχει εφαρμοστεί στη συγκεκριμένη μορφή (3.152). Οι παράμετροι επανάληψης θ_n επιλέγονται με την κυκλική σειρά
$\theta_1, \theta_2, \cdots \theta_m, \theta_1, \theta_2, \cdots$ όπου m είναι ακέραιος. Από την (3.153) για δεδομένα $\theta_1, \theta_2, \cdots, \theta_m$ έχουμε

$$u^{(m)} = P_m(G)u^{(0)} + k_m, \tag{3.154}$$

για ένα κατάλληλο διάνυσμα k_m και το $P_m(G)$ είναι το πολυώνυμο

$$P_m(G) = \prod_{k=1}^{m}(\theta_k G + (1 - \theta_k)I), \tag{3.155}$$

Ακολουθώντας την ανάλυση της προηγούμενης παραγράφου συμπεραίνεται ότι το ελαχιστοποιημένο πολυώνυμο $P_m(m)$ δίνεται από τον τύπο

$$P_m(\mu) = \frac{T_m\left(\frac{2\mu-(\beta+\alpha)}{\beta-\alpha}\right)}{T_m\left(\frac{2-(\beta+\alpha)}{\beta-\alpha}\right)}. \tag{3.156}$$

Οι παράμετροι επανάληψης θ_k μπορούν να προσδιορισθούν εξισώνοντας τις ρίζες των (3.155) και (3.156). Ετσι λοιπόν οι τιμές των παραμέτρων θ_k προκύπτουν από τον τύπο

$$\theta_k = \frac{2}{2 - (\beta - \alpha)\ \text{συν}\frac{(2k-1)\pi}{2m} - (\beta + \alpha)}, \quad k = 1, 2, \cdots, m \tag{3.157}$$

Η ιδεατή φασματική ακτίνα της (3.153) μπορεί να επαληθευθεί από την (3.156) ότι είναι

$$\overline{S}(P_{lm}(G)) = \left(\frac{2r^{m/2}}{1 + r^m}\right)^l \tag{3.158}$$

όπου l είναι ένας ακέραιος που προσδιορίζει τον αριθμό των κύκλων. Από τις (3.158) και (45) μπορούμε να διαπιστωθεί, ότι καθώς το m αυξάνει, η ταχύτητα σύγκλισης τείνει σε αυτήν της ημιεπαναληπτικής μεθόδου.

Ομως, τα πειραματικά αποτελέσματα ($Young$[1954α,1956], $Young$ και $Warlick$[1953]) δείχνουν ότι για μεγάλα m ενδέχεται να παρατηρηθεί αριθμητική αστάθεια. Επίσης, δεν είναι επιθυμητό να επιλεγεί το m πολύ μεγάλο επειδή η σύγκλιση αναμένεται μετά από lm επαναλήψεις.

3.6.2 Μέθοδοι Δευτέρου Βαθμού (Second-Degree) (SD)

Μιά επιταχυντική επαναληπτική μέθοδος παρόμοια με την (38) μπορεί να προκύψει αν θεωρήσουμε σταθερές παραμέτρους επανάληψης. Πιό συγκεκριμένα, αν εκφράσουμε την (38) στην ισοδύναμη μορφή (όπου v θέτουμε u)

$$u^{(n+1)} = u^{(n)} + (\rho_{n+1} - 1)\left(u^{(n)} - u^{(n-1)}\right) + \frac{2\rho_{n+1}}{2 - (\alpha + \beta)}\left(Gu^{(n)} + k - u^{(n)}\right) \tag{3.159}$$

και θέσουμε $\xi = \rho_{n+1} - 1$ και $\eta = \dfrac{2\rho_{n+1}}{2 - (\alpha + \beta)}$, τότε προκύπτει η επαναληπτική μέθοδος **δευτέρου βαθμού**(**Second − Degree**)

$$u^{(n+1)} = u^{(n)} + \xi\left(u^{(n)} - u^{(n-1)}\right) + \eta\left(Gu^{(n)} + k - u^{(n)}\right) \tag{3.160}$$

Η μορφή της (3.160) είναι μια ειδική περίπτωση της γραμμικής στατικής επαναληπτικής μεθόδου δευτέρου βαθμού που δίνεται ως εξής

$$u^{(n+1)} = G_1 u^{(n)} + H_1 u^{(n-1)} + k_1 \tag{3.161}$$

Στη συνέχεια (βλ. *Golub* και $Varga$ [1961]) η (3.161) μπορεί να γραφεί ως εξής

$$\begin{bmatrix} u^{(n)} \\ \\ u^{(n+1)} \end{bmatrix} = \begin{bmatrix} O & I \\ \\ H_1 & G_1 \end{bmatrix} \cdot \begin{bmatrix} u^{(n-1)} \\ \\ u^{(n)} \end{bmatrix} + \begin{bmatrix} O \\ \\ k_1 \end{bmatrix} \tag{3.162}$$

Η επαναληπτική μέθοδος (3.162) συγκλίνει αν και μόνο άν

$$S(M) < 1 \tag{3.163}$$

όπου

$$M = \begin{bmatrix} O & I \\ \\ H_1 & G_1 \end{bmatrix} \tag{3.164}$$

Έτσι λοιπόν, αν l είναι μια ιδιοτιμή του M, τότε οι ρίζες της εξίσωσης

$$det(\lambda^2 I - \lambda G_1 - H_1) = 0 \tag{3.165}$$

πρέπει να είναι μικρότερες της μονάδας , έτσι ώστε να ισχύει η (3.163). Στην περίπτωση της (3.160) η (3.165) γίνεται

$$det\left(\lambda^2 I - \lambda\left(\eta G + (1 - \eta + \xi)I\right) + \xi I\right) = 0 \ , \tag{3.166}$$

οπότε αν μ είναι μια ιδιοτιμή του G, τότε ισχύει

$$\lambda^2 - \lambda(\eta\mu + 1 - \eta + \xi) + \xi = 0 \ \cdot \tag{3.167}$$

Για σταθερό ξ η φασματική ακτίνα, που είναι η $\max\limits_{\mu} |\lambda|$, ελαχιστοποιείται όταν

$$(\eta\mu + 1 - \eta + \xi)^2 = 4\xi \tag{3.168}$$

Έτσι (από (7.9)) έχουμε ότι

$$\eta(\beta - 1) + 1 + \xi = 2\xi^{\frac{1}{2}} \tag{3.169}$$

και

$$\eta(\alpha-1)+1+\xi=-2\xi^{\frac{1}{2}} \quad \cdot \tag{3.170}$$

Συνεπώς, με πρόσθεση των (3.169) και (3.170) μπορεί να προσδιοριστεί το h από την σχέση

$$\eta=\frac{2(1+\xi)}{2-(\beta+\alpha)} \cdot \tag{3.171}$$

Επιπλέον, από την (3.171) και μιά από τις (3.169) , (3.170) η καλύτερη εκλογή για το ξ είναι η ακόλουθη

$$\xi_O=\hat{\omega}_O-1 \tag{3.172}$$

όπου

$$\hat{\omega}_O=\frac{2}{1+\sqrt{1-\sigma^2}} \cdot \tag{3.173}$$

και σ όπως ορίζεται στην (41).

Τελικά, από τις (3.172) και (3.173) η καλύτερη τιμή του η προκύπτει από την έκφραση

$$\eta_O=\frac{2\hat{\omega}_O}{2-(\beta+\alpha)} \cdot \tag{3.174}$$

Από τις (3.172), (3.173), (3.167) και (45) η φασματική ακτίνα του M δίνεται ως εξής

$$S(M)=(\hat{\omega}_O-1)^{\frac{1}{2}}=r^{\frac{1}{2}} \ , \tag{3.175}$$

οπότε η ταχύτητα σύγκλισης είναι

$$R(M)=-\frac{1}{2}logr \tag{3.176}$$

η οποία είναι συγκρίσιμη με εκείνη που προκύπτει για τις ημιεπαναληπτικές τεχνικές. Επίσης, από τις (3.175) και (46) συμπεραίνουμε ότι η ταχύτητα σύγκλισης της ημιεπαναληπτικής και της μεθόδου δευτέρου βαθμού εξαρτάται από την ποσότητα r. Εξάλλου, έχει αποδειχθεί (βλ. *Young* και *Kincaid* [1969]) ότι η ημιεπαναληπτική μέθοδος παρουσιάζει μεγαλύτερη επιτάχυνση από την μέθοδο δευτέρου βαθμού.

Αυτό άλλωστε αναμένεται, διότι οι συντελεστές στην μέθοδο δευτέρου βαθμού είναι σταθεροί, ενώ στην ημιεπαναληπτική μέθοδο είναι μεταβλητές. Ομως, στις ημιεπαναληπτικές μεθόδους, χρειάζεται να αποθηκεύονται δύο διανύσματα στην κάθε επανάληψη και συνεπώς η απαίτηση για αποθήκευση μπορεί να είναι κρίσιμη για μεγάλα συστήματα εξισώσεων ή σε υπολογιστές με περιορισμένη χωρητικότητα μνήμης.

3.6.3 Μέθοδος των Συζυγών Κατευθύνσεων (Conjugate Gradient)

Στην παράγραφο αυτή μελετάται σύντομα η μέθοδος **ConjugateGradient**(**CG**) η οποία έχει προταθεί αρχικά από τους $Hestnenes$ και $Stiefel$ [1952], $Stiefel$[1952] ως μια επαναληπτική μέθοδος για την επίλυση μεγάλων αραιών γραμμικών συστημάτων ($Reid$[1971]).
Ας θεωρήσουμε πάλι το γραμμικό σύστημα

$$Au = b \tag{3.177}$$

όπου A είναι ένας $N \times N$ συμμετρικός και θετικά ορισμένος πίνακας. Η μέθοδος στηρίζεται στο γνωστό αποτέλεσμα βελτιστοποίησης του $Luenberger$(1973).

Θεώρημα : Αν ο πίνακας A είναι πραγματικός συμμετρικός και θετικά ορισμένος τότε η επίλυση του γραμμικού συστήματος $Au = b$ είναι ισοδύναμη με την ελαχιστοποίηση της τετραγωνικής συνάρτησης

$$Q(u) = \frac{1}{2}(u, Au) - (u, b) = const \tag{3.178}$$

Επίσης η συνάρτηση $Q(u)$ για $u = A^{-1}b$ έχει ελάχιστη τιμή $\frac{1}{2}(b, A^{-1}b)$.

Η τετραγωνική συνάρτηση (3.178) ορίζει μια οικογένεια ομοίων ελλειψοειδών στον N-διάστατο Ευκλείδιο χώρο, τα οποία έχουν κοινό κέντρο το $A^{-1}b$, σημείο στο οποίο η $Q(u)$ παίρνει την ελάχιστη τιμή της. Για ένα αυθαίρετο διάνυσμα $u^{(n)}$, το υπόλοιπο $r^{(n)}$ δίνεται από

τον τύπο

$$r^{(n)} = b - Au^{(n)} = -\left[Grad\ Q(u)\right]^{1} u^{(n)} \tag{3.179}$$

και αυτό είναι πάντοτε κάθετο πρός την επιφάνεια του ελλιψοειδούς που ορίζεται από την (3.178). Ετσι λοιπόν, επιδιώκουμε να φθάσουμε στη λύση $A^{-1}b$, δηλαδή στο κέντρο του ελλειψοειδούς, με μια ακολουθία διανυσμάτων μετατόπισης της μορφής

$$u^{(n+1)} = u^{(n)} + \epsilon_n p^{(n)} \tag{3.180}$$

όπου $p^{(n)}$ είναι μια αυθαίρετη διεύθυνση και ϵ_n είναι μια αυθαίρετη σταθερά.

Το πρόβλημα λοιπόν ανάγεται στο να προσδιοριστούν τα ϵ_n έτσι ώστε η τετραγωνική συνάρτηση $Q(u^{(n+1)})$ να ελαχιστοποιείται για μια δοθείσα διεύθυνση $p^{(n)}$. Από τις (3.178) και (3.180) έχουμε ότι η $Q(u^{(n+1)})$ δίνεται από τον τύπο

$$Q(u^{(n+1)}) = \frac{1}{2}\left((u^{(n)} + \epsilon_n p^{(n)}), A(u^{(n)} + \epsilon_n p^{(n)})\right) - \left((u^{(n)} + \epsilon_n p^{(n)}), b\right) \tag{3.181}$$

όπου

$$\begin{aligned} \frac{\vartheta Q(u^{(n+1)})}{\vartheta \epsilon_n} &= \left(p^{(n)}, A(u^{(n)} + \epsilon_n p^{(n)})\right) - (p^{(n)}, b) \\ &= -(p^{(n)}, r^{(n)}) + (\epsilon_n p^{(n)}, Ap^{(n)}) \cdot \end{aligned} \tag{3.182}$$

Η βέλτιστη τιμή του e_n προκύπτει θέτοντας την έκφραση (3.182) ίση με μηδέν, από την οποία άμεσα προκύπτει

$$\epsilon_n = \frac{(p^{(n)}, r^{(n)})}{(p^{(n)}, Ap^{(n)})} \quad . \tag{3.183}$$

Επίσης, με τη χρήση του ορισμού του $u^{(n+1)}$ στην (3.180) και της τιμής που προέκυψε για το e_n, έχουμε

$$(p^{(n)}, r^{(n+1)}) = \left(p^{(n)}, (b - Au^{(n+1)})\right) = \left(p^{(n)}, (r^{(n)} - \epsilon_n Ap^{(n)})\right) = 0 \tag{3.184}$$

[1] όπου $[Grad\ Q(u)]\, u^{(n)}$ αναπαριστά ένα διάνυσμα με συνιστώσες $\dfrac{\vartheta Q(u^{(n)})}{\vartheta u_i}$, $i = 1, 2, \cdots, n$

που σημαίνει ότι η διεύθυνση $p^{(n)}$ και το υπόλοιπο $r^{(n+1)}$ είναι ορθογώνια.

Η εκλογή του διανύσματος διεύθυνσης $p^{(n)}$ διαφοροποιεί πολλές μεθόδους, οι οποίες συγκλίνουν για ένα δοθέν $p^{(n)}$. Αν θέλουμε να επιλέξουμε το $p^{(n)}$ να βρίσκεται κατά μήκος της γραμμής της *steepest descent*, τότε απλά παίρνουμε $p^{(n)} = r^{(n)}$ και από τις (3.180) και (3.183) άμεσα ορίζουμε την γνωστή μέθοδο *Steepest Descent* η οποία όμως σε πολλές περιπτώσεις παρουσιάζει μια πολύ αργή σύγκλιση.

Μια καλύτερη στρατηγική για την εκλογή της διεύθυνσης $p^{(n)}$ βασίζεται στο ότι το κέντρο του ελλιψοειδούς βρίσκεται πάνω στο επίπεδο το συζυγές ως προς μια δοθείσα χορδή. Ετσι, αν εκλέξουμε τα διανύσματα $p^{(0)}, p^{(1)}, \cdots, p^{(N-1)}$ να είναι ανά δύο συζυγή, δηλαδή να ισχύει

$$(p^{(i)}, Ap^{(j)})) = 0 \tag{3.185}$$

για $i \neq j$, τότε με τον προσδιορισμό του $p^{(n+1)}$ από τον τύπο

$$p^{(n)} = r^{(n)} + a_{n-1}p^{(n-1)} \tag{3.186}$$

και με συνδυασμό των (3.185) και (3.186) προκύπτει

$$(p^{(n)}, Ap^{(n-1)}) = (r^{(n)}, Ap^{(n-1)}) + (a_{n-1}p^{(n-1)}, Ap^{(n-1)}) = 0 \tag{3.187}$$

και τελικά

$$a_{n-1} = \frac{(r^{(n)}, Ap^{(n-1)})}{(p^{(n-1)}, Ap^{(n-1)})} \tag{3.188}$$

Η εκλογή αυτή του $p^{(n)}$ μας οδηγεί στο επαναληπτικό σχήμα της μεθόδου **Conjugate Gradient**(**CG**), που ορίζεται ως ακολούθως

$$u^{(n+1)} = u^{(n)} + \epsilon_n p^{(n)}, \quad n = 0, 1, 2, \cdots, m-1 \tag{3.189}$$

$$r^{(n)} = b - Au^{(n)}, \quad n = 0, 1, 2, \cdots, m \tag{3.190}$$

$$p^{(n)} = \begin{cases} 0 & , \quad n = -1 \\ \\ r^{(n)} + a_{n-1}p^{(n-1)} & , \quad n = 0, 1, 2, \cdots, m-1 \end{cases} \tag{3.191}$$

$$a_{n-1} = \begin{cases} 0 & , \quad n = 0 \\ -\dfrac{(r^{(n)}, Ap^{(n-1)})}{(p^{(n-1)}, Ap^{(n-1)})} & , \quad n = 0, 1, 2, \cdots, m-1 \end{cases} \tag{3.192}$$

όπου m είναι ο μικρότερος ακέραιος τέτοιος ώστε να ισχύει

$$r^{(m)} = 0 \quad . \tag{3.193}$$

Στη συνέχεια συνοψίζουμε ορισμένες βασικές ιδιότητες της μεθόδου CG (βλ. *Beckmann* [1960])

$$\begin{aligned} &(r^{(i)}, r^{(j)}) = 0 && i \neq j \quad , \quad i, j = 0, 1, 2, \cdots, m-1 \\ &(p^{(i)}, Ap^{(j)}) = 0 && i \neq j \quad , \quad i, j = 0, 1, 2, \cdots, m-1 \\ &p^{(i)} \neq 0 && , \quad i = 0, 1, 2, \cdots, m-1 \\ &m \leq N \\ &kai \\ &a_{n-1} = \frac{(r^{(n)}, r^{(n)})}{(r^{(n-1)}, r^{(n-1)})} && , \quad n = 1, 2, \cdots, m-1 \end{aligned} \tag{3.194}$$

Από την (3.193) συμπεραίνεται ότι η επαναληπτική μέθοδος CG συγκλίνει σε N το πολύ επαναλήψεις, όπου N είναι η τάξη του πίνακα A. Αν και η μέθοδος CG θεωρητικά δίνει μια ακριβή λύση σε N επαναλήψεις, αυτό δεν συμβαίνει πραγματικά στην πράξη, όπου το σφάλμα στρογγύλευσης επηρεάζει δραστικά την ορθογωνιότητα των υπολοίπων. Για το λόγο αυτό έχουν γίνει στην μέθοδο CG κάποιες τροποποιήσεις και βελτιώσεις (βλ. *Rutishauer* [1959], *Daniel* [1967], *Reid* [1971], *Axelson*[1974] και *Evans*[1973]).

Μια ενδιαφέρουσα τροποποίηση της μεθόδου CG είναι η διατύπωσή της ως μιας μεθόδου δευτέρου βαθμού, στην οποία ο όρος $u^{(n+1)}$ εκφράζεται συναρτήσει των δύο προηγουμένων του $u^{(n)}$ και $u^{(n-1)}$. Αντικαθιστώντας όπου n το $n-1$ στην (3.189) έχουμε

$$u^{(n)} = u^{(n-1)} + \epsilon_{n-1} p^{(n-1)} \tag{3.195}$$

ή

$$\frac{a_{n-1}}{\epsilon_{n-1}}\epsilon_n u^{(n)} = \frac{a_{n-1}}{\epsilon_{n-1}}\epsilon_n u^{(n-1)} + \epsilon_n a_{n-1} p^{(n-1)} \tag{3.196}$$

οπότε με απαλοιφή του $p^{(n-1)}$ χρησιμοποιώντας την (3.191) προκύπτει

$$\frac{a_{n-1}}{\epsilon_{n-1}}\epsilon_n u^{(n)} = \frac{a_{n-1}}{\epsilon_{n-1}}\epsilon_n u^{(n-1)} + \epsilon_n (p^{(n)} - r^{(n)}) \tag{3.197}$$

και τελικά με απαλοιφή του $p^{(n)}$ στην (3.189) προκύπτει

$$u^{(n+1)} = \left(1 + \frac{\epsilon_n}{\epsilon_{n-1}}\alpha_{n-1}\right) u^{(n)} - \frac{\epsilon_n}{\epsilon_{n-1}}\alpha_{n-1} u^{(n-1)} + \epsilon_n r^{(n)} \tag{3.198}$$

η οποία μπορεί να γραφεί στη πιό συνεπτυγμένη μορφή

$$u^{(n+1)} = \rho_{n+1}(u^{(n)} + \gamma_{n+1} r^{(n)}) + (1 - \rho_{n+1}) u^{(n-1)} \tag{3.199}$$

όπου

$$r_{n+1} = 1 + \frac{\epsilon_n}{\epsilon_{n-1}}\alpha_{n-1} \tag{3.200}$$

και

$$\gamma_{n+1} = \frac{\epsilon_n}{\rho_{n+1}} \quad . \tag{3.201}$$

Στη συνέχεια για την απλοποίηση των εκφράσεων των ρ_{n+1} και γ_{n+1} εκφράζουμε αυτές συναρτήσει ορισμένων εσωτερικών γινομένων. Πιό συγκεκριμένα, εκφράζουμε την (3.199) συναρτήσει των υπολοίπων με χρήση της (3.190), οπότε προκύπτει

$$r^{(n+1)} = \rho_{n+1}(r^{(n)} - \gamma_{n+1} A r^{(n)}) + (1 - \rho_{n+1}) r^{(n-1)} \tag{3.202}$$

Αν τώρα πάρουμε το εσωτερικό γινόμενο των δύο μελών της (3.202) με το $r^{(n)}$, τότε από την (3.194) προκύπτει

$$0 = \rho_{n+1}\left((r^{(n)}, r^{(n)}) - \gamma_{n+1}(r^{(n)}, A r^{(n)})\right) \tag{3.203}$$

και επειδή $\rho_{n+1} \neq 0$ προκύπτει

$$\gamma_{n+1} = \frac{(r^{(n)}, r^{(n)})}{(r^{(n)}, A r^{(n)})} \tag{3.204}$$

Εξάλλου, αν πάρουμε το εσωτερικό γινόμενο των δύο μελών της (3.202) με το $r^{(n-1)}$ προκύπτει

$$0 = \rho_{n+1}\left(-\gamma_{n+1}(r^{(n-1)}, Ar^{(n)})\right) + (1-\rho_{n+1})\left(r^{(n-1)}, r^{(n-1)}\right) \quad (3.205)$$

ή

$$rho_{n+1} = \left[1 + \frac{(r^{(n-1)}, Ar^{(n)})}{(r^{(n-1)}, r^{(n-1)})} \cdot \gamma_{n+1}\right]^{-1} \quad . \quad (3.206)$$

Επιπλέον, αντικαθιστώντας όπου n το $n-1$ στην (3.202) έχουμε

$$r^{(n)} = \rho_n(r^{(n-1)} - \gamma_n A r^{(n-1)}) + (1-\rho_n) r^{(n-2)} \quad (3.207)$$

και αν πάρουμε το εσωτερικό γινόμενο των δύο μελών με το $r^{(n)}$ προκύπτει

$$(r^{(n-1)}, Ar^{(n)}) = -\frac{(r^{(n)}, r^{(n)})}{\gamma_n \rho_n} \quad (3.208)$$

έτσι λοιπόν η (3.206) γίνεται

$$\rho_{n+1} = \left[1 - \frac{\gamma_{n+1}}{\gamma_n} \cdot \frac{(r^{(n)}, r^{(n)})}{(r^{(n-1)}, r^{(n-1)})} \cdot \frac{1}{\rho_n}\right]^{-1} \quad . \quad (3.209)$$

Συνοψίζοντας τα ανωτέρω η μέθοδος CG μπορεί επίσης να ορισθεί ως εξής

$$u^{(n+1)} = \rho_{n+1}(u^{(n)} + \gamma_{n+1} r^{(n)}) + (1-\rho_{n+1}) u^{(n-1)} \quad (3.210)$$

όπου

$$\rho_1 = 1$$

$$\rho_{n+1} = \left[1 - \frac{\gamma_{n+1}}{\gamma_n} \cdot \frac{(r^{(n)}, r^{(n)})}{(r^{(n-1)}, r^{(n-1)})} \cdot \frac{1}{\rho_n}\right]^{-1} \quad , \qquad n = 1, 2, \cdots \quad (3.211)$$

και

$$\gamma_{n+1} = \frac{(r^{(n)}, r^{(n)})}{(r^{(n)}, Ar^{(n)})} \quad (3.212)$$

Από την (3.210) παρατηρούμε ότι η μέθοδος CG έχει την ίδια μορφή με την μέθοδο SI με τη μόνη διαφορά ότι εδώ οι παράμετροι είναι μεταβλητές (ενώ στην SI είναι $\gamma_1 = \gamma_2 = \cdots = \overline{\rho}$) , που εκλέγονται έτσι ώστε να ελαχιστοποιούν την τετραγωνική συνάρτηση $Q(u)$.

Πράγματι, αναμένεται η μέθοδος CG να παρουσιάζει μια καλύτερη ταχύτητα σύγκλισης συγκριτικά με την εφαρμογή των ημιεπαναληπτικών τεχνικών (SI), εφόσον εμπεριέχει μια επιπλέον παράμετρο g_{n+1} η οποία είναι μεταβλητή (ενώ είναι σταθερά στην SI). Παρατηρούμε ότι η μέθοδος CG απαιτεί περισσότερους υπολογισμούς ανά επανάληψη αλλά όμως δεν απαιτεί την εκτίμηση της μεγαλύτερης και μικρότερης ιδιοτιμής του πίνακα A. Επίσης, αποδεικνύεται$(Young$ [1975]) ότι για κάθε n ισχύει

$$||\hat{u}^{(n)} - \overline{u}||_{A^{\frac{1}{2}}} \leq ||u^{(n)} - \overline{u}||_{A^{\frac{1}{2}}} \tag{3.213}$$

όπου $\overline{u}$ είναι η ακριβής λύση του συστήματος (3.177), $\hat{u}^{(n)}$ είναι η προσεγγιστική λύση που προκύπτει με την μέθοδο CG και $u^{(n)}$ είναι η προσεγγιστική λύση που προκύπτει με την μέθοδο SI.

Η σχέση (3.213) δείχνει ένα ουσιαστικό πλεονέκτημα της μεθόδου CG έναντι της μεθόδου SI διότι χρησιμοποιεί μόνο το άνω και κάτω φράγμα για τις ιδιοτιμές του πίνακα A, ενώ η μέθοδος SI έχει το πλεονέκτημα της διανομής των ιδιοτιμών του G. Τελικά, παρατηρούμε ότι από την (3.213) είναι φανερό ότι η μέθοδος CG είναι καλύτερη από μια γραμμική μη-στατική μέθοδο δεύτερου βαθμού, λόγω της ελαχιστοποίησης της $A^{\frac{1}{2}}-norm$ του διανύσματος σφάλματος. Εφόσον μπορούμε να έχουμε εκτιμήσεις για την ταχύτητα σύγκλισης των μεθόδων SI βρίσκουμε ένα κάτω φράγμα της ταχύτητας σύγκλισης της μεθόδου CG. Συνεπώς, από την (3.213) και από το ότι για την μέθοδο SI ισχύει

$$||\hat{u}^{(n)} - \overline{u}||_{A^{\frac{1}{2}}} \leq \frac{2r^{n/2}}{1+r^n}||u^{(0)} - \overline{u}||_{A^{\frac{1}{2}}} \tag{3.214}$$

προκύπτει άμεσα ότι

$$||\hat{u}^{(n)} - \overline{u}||_{A^{\frac{1}{2}}} \leq \frac{2r^{n/2}}{1+r^n}||\hat{u}^{(0)} - \overline{u}||_{A^{\frac{1}{2}}} \tag{3.215}$$

υποθέτοντας ότι $\hat{u}^{(0)} = u^{(0)}$, όπου $r \equiv \tau$ που δίνεται στην (44).

Κεφάλαιο 4

Αριθμητικός Υπολογισμός Ιδιοτιμών και Ιδιοδιανυσμάτων

4.1 Γενικά

Η λύση πολλών προβλημάτων της Φυσικής απαιτεί τον υπολογισμό των ιδιοτιμών και των αντίστοιχων ιδιοδιανυσμάτων ενός πίνακα A τάξης n. Είναι γνωστό ότι ένας $n \times n$ πίνακας έχει ακριβώς n, όχι αναγκαία διάφορες μεταξύ τους, ιδιοτιμές οι οποίες είναι οι ρίζες του πολυώνυμου $p(\lambda) = det(A - \lambda I)$. Θεωρητικά οι ιδιοτιμές του A προσδιορίζονται από την εύρεση των ριζών του $p(\lambda)$ και στη συνέχεια επιλύεται το σύστημα $(A - \lambda I)x = 0$ για τον προσδιορισμό των αντίστοιχων ιδιοδιανυσμάτων. Στην πράξη είναι δύσκολο να προσδιορίσουμε τη μορφή του $p(\lambda)$ και εκτός αυτού είναι αρκετά δύσκολο να υπολογίσουμε τις ρίζες ενός n βαθμού πολυωνύμου για μεγάλες τιμές του n. Για τον λόγο αυτό χρησιμοποιούμε προσεγγιστικές μεθόδους για τον υπολογισμό των ιδιοτιμών και των αντίστοιχων ιδιοδιανυσμάτων. Στη συνέχεια θα παρουσιαστούν μερικές μέθοδοι οι οποίες εφαρμόζονται όταν ο πίνακας A έχει ορισμένες ιδιότητες.

4.2 Η μέθοδος των δυνάμεων

Η μέθοδος αυτή υπολογίζει τη μεγαλύτερη κατά μέτρο ιδιοτιμή ενός τετραγωνικού πίνακα $A \in C^{nn}$ και το αντίστοιχο ιδιοδιάνυσμα. Έ-

στω ότι ο πίνακας A έχει τις $\lambda_{\mathrm{i}}, \mathrm{i} = 1(1)\mathrm{n}$ ιδιοτιμές και τα αντίστοιχα ιδιοδιανύσματα είναι τα $x^{(i)}, \mathrm{i} = 1(1)\mathrm{n}$. Για την εφαρμογή της μεθόδου υποθέτουμε ότι ο A έχει n γραμμικά ανεξάρτητα ιδιοδιανύσματα. Επιπλέον υποθέτουμε ότι

$$|\lambda_1| > |\lambda_j|, \qquad j = 2(1)n. \tag{4.1}$$

Αν $y^{(0)} \neq 0$ είναι ένα διάνυσμα του C^n, τότε επειδή τα ιδιοδιανύσματα του A αποτελούν μία βάση, το $y^{(0)}$ μπορεί να γραφεί σαν ένας γραμμικός συνδυασμός των $x^{(i)}$, συνεπώς

$$y^{(0)} = \sum_{i=1}^{n} \quad a_i x^{(i)} \tag{4.2}$$

όπου α_i είναι βαθμωτά μεγέθη και όχι όλα μηδέν. Στη συνέχεια θεωρούμε την ακολουθία των διανυσμάτων που ορίζονται από το επαναληπτικό σχήμα

$$y^{(m+1)} = Ay^{(m)}, \quad m = 0, 1, 2, \ldots \tag{4.3}$$

όπου το $y^{(0)}$ είναι ένα αυθαίρετο διάνυσμα. Η εφαρμογή του σχήματος (4.3) για $m = 0$ δίνει διαδοχικά

$$y^{(1)} = Ay^{(0)} = A\sum_{i=1}^{n} a_i x^{(i)} = \sum_{i=1}^{n} a_i A x^{(i)} = \sum_{i=1}^{n} a_i \lambda_i x^{(i)}$$

επίσης για $m = 1$ έχουμε

$$y^{(2)} = Ay^{(1)} = \sum_{i=1}^{n} a_i \lambda_i^2 x^{(i)}$$

Γενικά

$$y^{(m)} = \sum_{i=1}^{n} a_i \lambda_i^m x^{(i)} \tag{4.4}$$

η οποία γράφεται

$$y^{(m)} = \lambda_1^m \left[a_1 x^{(1)} + \sum_{i=2}^{n} a_i \left(\frac{\lambda_i}{\lambda_1} \right)^m x^{(i)} \right]$$
$$= \lambda_1^m [a_1 x^{(1)} + \varepsilon^{(m)}] \tag{4.5}$$

όπου

$$\varepsilon^{(m)} = \sum_{i=2}^{n} a_i \left(\frac{\lambda_i}{\lambda_1} \right)^m x^{(i)}$$

επειδή όμως $|\lambda_i/\lambda_1| < 1, \quad i = 2(1)n$, από την (4.5) προκύπτει ότι

$$\lim_{m\to\infty} y^{(m)} = \lambda_1^m a_1 x^{(1)} \tag{4.6}$$

υποθέτοντας βέβαια ότι $a_1 \neq 0$. Αν όπου m θέσουμε $m-1$ στην (4.6) έχουμε

$$\lim_{m\to\infty} y^{(m-1)} = \lambda_1^{m-1} a_1 x^{(1)} \tag{4.7}$$

Διαιρώντας τις αντίστοιχες συνιστώσες των $y^{(m)}$ και $y^{(m-1)}$ λαμβάνουμε από τις (4.6) και (4.7) ότι

$$\lim_{m\to\infty} \frac{y_j^{(m)}}{y_j^{(m-1)}} = \lambda_1 \quad j = 1(1)n \tag{4.8}$$

Έχοντας υπολογίσει την λ_1 βρίσκουμε από την (4.6) ότι

$$\lim_{m\to\infty} \frac{y^{(m)}}{\lambda_1^{(m)}} = \alpha_1 x^{(1)} \tag{4.9}$$

δηλαδή το ίδιο ιδιοδιάνυσμα που αντιστοιχεί στην λ_1 την μεγαλύτερη κατά μέτρο ιδιοτιμή. Με άλλα λόγια δημιουργούμε την ακολουθία των διανυσμάτων $y^{(0)}, y^{(1)}, \ldots, y^{(m)}$ μέχρις ότου οι λόγοι των αντίστοιχων συνιστωσών δύο διαδοχικών διανυσμάτων τείνουν προς την ίδια σταθερή τιμή η οποία είναι μία προσέγγιση της ιδιοτιμής λ_1. Το διάνυσμα $y^{(m)}$ είναι μία μη κανονικοποιημένη προσέγγιση του αντίστοιχου ιδιοδιανύσματος. Η ταχύτητα σύγκλισης της μεθόδου εξαρτάται από τις σταθερές α_i και τους λόγους $|\lambda_2/\lambda_1|, |\lambda_3/\lambda_1|, \ldots, |\lambda_n/\lambda_1|$. Έτσι όσο μικρότεροι είναι αυτοί οι λόγοι τόσο ταχύτερη είναι η προσέγγιση της μεθόδου. Ιδιαίτερα αν $|\lambda_2| / |\lambda_1|$ είναι κοντά στη μονάδα, τότε η

σύγκλιση της μεθόδου είναι πιθανό να είναι πάρα πολύ αργή. Θεωρητικά, αν τύχει και η εκλογή του $y^{(0)}$ είναι τέτοια ώστε $\alpha_1 = 0$ και $|\lambda_2| > |\lambda_j|\,, \quad j \geq 3$, τότε η μέθοδος θα συγκλίνει στην λ_2 και σε ένα πολλαπλάσιο του $x^{(2)}$. Στην πράξη όμως δεν έχουμε δυσκολίες αν $\alpha_1 = 0$, γιατί τα σφάλματα στρογγύλευσης δημιουργούν μία μικρή (που με την αύξηση του αριθμού των επαναλήψεων μεγαλώνει) τιμή του α_1 αρκετά ικανοποιητική, ώστε να έχουμε σύγκλιση τελικά στην λ_1.

Στην περίπτωση όπου $\lambda_1 = \lambda_2$ και $|\lambda_2| > |\lambda_j|\,, \quad j = 3(1)n$, τότε είναι φανερό ότι αντί της (4.5) θα έχουμε τη σχέση

$$y^{(m)} = \lambda_1^m \left[\alpha_1 x^{(1)} + \alpha_2 x^{(2)} + \sum_{i=3}^{n} \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^m x^{(i)} \right] \tag{4.10}$$

Αν υποθέσουμε ότι $|\alpha_1| + |\alpha_2| \neq 0$, τότε εργαζόμενοι όπως και προηγουμένως εύκολα βρίσκουμε ότι

$$\lim_{m-\infty} \frac{y_j^{(m)}}{y_j^{(m-1)}} = \lambda_1 \tag{4.11}$$

και

$$\lim_{m-\infty} \frac{y^{(m)}}{\lambda_1^m} = \alpha_1 x^{(1)} + \alpha_2 x^{(2)} \tag{4.12}$$

δηλαδή ένα ιδιοδιάνυσμα που αντιστοιχεί στην λ_1. (Αν $x^{(1)}, \quad x^{(2)}$ είναι ιδιοδιανύσματα που αντιστοιχούν στην ίδια ιδιοτιμή λ, τότε και $c_1 x^{(1)} + c_2 x^{(2)}$, όπου c_1 και c_2 σταθερές είναι επίσης ιδιοδιάνυσμα που αντιστοιχεί στην ιδιοτιμή λ). Για να βρούμε στην πράξη ένα ιδιοδιάνυσμα μη συγγραμμικό προς το $\alpha_1 x^{(1)} + \alpha_2 x^{(2)}$ που αντιστοιχεί στην ίδια ιδιοτιμή λ_1 αλλάζουμε το αρχικό διάνυσμα $y^{(0)}$. Αν τώρα $\lambda_1 = -\lambda_2$ και $|\lambda_2| > |\lambda_j|\,, \quad j = 3(1)n$ τότε αντί της (4.5) θα έχουμε την

$$y^{(m)} = \lambda_1^{(m)} \left[\alpha_1 x^{(1)} + (-1)^m \alpha_2 x^{(2)} + \sum_{i=3}^{n} \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^m x^{(i)} \right], \quad m = 0, 1, 2, \ldots$$

από την οποία δεν μπορεί να βρεθεί η λ_1. Μπορούμε όμως να σχηματίσουμε τις εξής δύο ακολουθίες

$$y^{(2m)} = \lambda_1^{2m}\left[\alpha_1 x^{(1)} + \alpha_2 x^{(2)} + \sum_{i=3}^{n} \alpha_i \left(\frac{\lambda_i}{\lambda_1}\right)^{2m} x^{(i)}\right], \quad m = 0, 1, 2, \ldots \tag{4.13}$$

και

$$y^{(2m+1)} = \lambda_1^{2m+1}\left[\alpha_1 x^{(1)} - \alpha_2 x^{(2)} + \sum_{i=3}^{n} \alpha_i \left(\frac{\lambda_i}{\lambda_1}\right)^{2m+1} x^{(i)}\right], \quad m = 0, 1, 2, \ldots \tag{4.14}$$

οπότε από την (4.13) λαμβάνουμε

$$\lim_{m\to\infty} \frac{y_j^{(2m+2)}}{y_j^{(2m)}} = \lambda_1^2 \tag{4.15}$$

από την οποία υπολογίζουμε τις δύο ιδιοτιμές $\pm\lambda_1$. Για την εύρεση των αντίστοιχων ιδιοδιανυσμάτων από την (4.13) έχουμε

$$\lim_{m\to\infty} \frac{y^{(2m)}}{\lambda_1^{2m}} = \alpha_1 x^{(1)} + \alpha_2 x^{(2)} \tag{4.16}$$

ενώ από την (4.14) έχουμε

$$\lim_{m\to\infty} \frac{y^{(2m+1)}}{\lambda_1^{2m+1}} = \alpha_1 x^{(1)} - \alpha_2 x^{(2)} \tag{4.17}$$

Με πρόσθεση και αφαίρεση των μελών των (4.16) και (4.17) βρίσκουμε τα ιδιοδιανύσματα $\alpha_1 x^{(1)}$ και $\alpha_2 x^{(2)}$ που αντιστοιχούν στις ιδιοτιμές λ_1 και $\lambda_2(= -\lambda_1)$. Αν υποθέσουμε τέλος ότι έχουμε την περίπτωση όπου $\lambda_1 = r + iq$ και λ_2 είναι ο συζυγής μιγαδικός αριθμός της λ_1 δηλαδή $\lambda_2 = \overline{\lambda_1}$. Είναι φανερό πως αν $x^{(1)}$ είναι το ιδιοδιάνυσμα που αντιστοιχεί στην λ_1, τότε $x^{(2)} = \overline{x}^{(1)}$ εφόσον $A \in R^{nn}$. Επίσης αν $y^{(0)}$ είναι το αρχικό πραγματικό διάνυσμα για την εφαρμογή της μεθόδου των δυνάμεων θα έχουμε $\alpha_2 = \overline{\alpha}_1$. Έτσι μετά από m επαναλήψεις θα έχουμε

$$y^{(m)} = \lambda_1^m \left[\alpha_1 x^{(1)} + \left(\frac{\overline{\lambda}_1}{\lambda_1} \right)^m \overline{\alpha}_1 \overline{x}^{(1)} + \sum_{i=3}^{n} \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^m x^{(i)} \right]$$
$$= \lambda_1^m \left[\alpha_1 x^{(1)} + \left(\frac{\overline{\lambda}_1}{\lambda_1} \right)^m \overline{\alpha}_1 \overline{x}^{(1)} + \varepsilon^{(m)} \right]$$

όπου

$$\varepsilon^{(m)} = \sum_{i=3}^{n} \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^m x^{(i)}$$

ή

$$\lim_{m \to \infty} y^{(m)} = \lambda_1^m \left[\alpha_1 x^{(1)} + \left(\frac{\overline{\lambda}_1}{\lambda_1} \right)^m \overline{\alpha}_1 \overline{x}^{(1)} \right]. \tag{4.18}$$

Αν $\overline{\lambda}_1$ και λ_1 είναι οι ρίζες της εξίσωσης

$$\lambda^2 + b\lambda + c = 0 \qquad b, c \in \mathbb{R} \tag{4.19}$$

τότε λόγω της (4.4) η (4.19) γράφεται διαδοχικά

$$\lambda^{m+2} + b\lambda^{m+1} + c\lambda^m = 0$$

ή

$$y^{(m+2)} + by^{(m+1)} + cy^{(m)} = 0 \tag{4.20}$$

όπου υπετέθει ότι

$$\lim_{m \to \infty} \varepsilon^{(m+i)} = 0, \quad i = 0, 1, 2.$$

Οι σταθερές b, c μπορούν να υπολογισθούν από δύο οποιεσδήποτε εξισώσεις του συνόλου (4.20). Μία καλύτερη διαδικασία είναι η χρησιμοποίηση όλων των n εξισώσεων της (4.20). Πράγματι, ο υπολογισμός των b, και c είναι τέτοιος ώστε η ποσότης

$$\sum_{i=1}^{n} [y_i^{(m+2)} + by_i^{(m+1)} + cy_i^{(m)}]^2 \tag{4.21}$$

να είναι ελάχιστη. Το πρόβλημα αυτό είναι ένα γραμμικό πρόβλημα ελαχίστων τετραγώνων. Από τη στιγμή που προσδιοριστούν τα b, c μπορούμε να υπολογίσουμε από την (4.19) τις λ_1 και $\lambda_2(=\overline{\lambda}_1)$. Για την εύρεση των αντίστοιχων ιδιοδιανυσμάτων $x^{(1)}$ και $x^{(2)}(=\overline{x}^{(1)})$ χρησιμοποιούμε την (4.18) για δύο διαδοχικά διανύσματα y^m και $y^{(m+1)}$. Έχουμε λοιπόν

$$\lim_{m\to\infty}\frac{y^{(m)}}{\lambda_1^m}=\alpha_1x_1+\left(\frac{\overline{\lambda}_1}{\lambda_1}\right)^m\overline{\alpha}_1\overline{x}^{(1)} \tag{4.22}$$

και

$$\lim_{m\to\infty}\frac{y^{(m+1)}}{\lambda_1^{m+1}}=\alpha_1x_1+\left(\frac{\overline{\lambda}_1}{\lambda_1}\right)^{m+1}\overline{\alpha}_1\overline{x}^{(1)} \tag{4.23}$$

Πολ/ζοντας την (4.22) επί $\overline{\lambda}_1/\lambda_1$ και αφαιρώντας κατά μέλη από την (4.23) βρίσκεται το ιδιοδιάνυσμα $\alpha_1x^{(1)}$ που αντιστοιχεί στην ιδιοτιμή λ_1. Για την εύρεση του ιδιοδιανύσματος, που αντιστοιχεί στην λ_2 αρκεί να πάρουμε το συζυγές του $\alpha_1x^{(1)}$.
Η παραπάνω διαδικασία μπορεί να γενικευθεί προκειμένου να υπολογίζει οποιοδήποτε αριθμό άνισων μεγαλύτερων ιδιοτιμών που έχουν το ίδιο μέτρο, ή τις ιδιοτιμές $\lambda_1,\lambda_2,\ldots,\lambda_k$ (πραγματικές ή μιγαδικές) οι οποίες ικανοποιούν τις σχέσεις

$$|\lambda_1|\geq|\lambda_2|\geq\ldots\geq|\lambda_k|>>|\lambda_{k+1}|\geq\ldots\geq|\lambda_n| \tag{4.24}$$

Όπως όμως είναι γνωστό υπάρχουν ουσιαστικά προβλήματα με την μέθοδο των ελαχίστων τετραγώνων για $k\geq 8$.

Παρατήρηση

Για $|\lambda_1|>1$ τότε από την (2.6) έχουμε ότι $\lim_{m\to\infty}y_i^{(m)}=\pm\infty,\quad j=1(1)n$, ενώ για $|\lambda_1|<1,\quad\lim_{m\to\infty}y_j^{(m)}=0$. Έτσι εκτός αν $|\lambda_1|\simeq 1$ θα πρέπει να εκτελούμε πράξεις με απόλυτα πάρα πολύ μεγάλους ή πάρα πολύ μικρούς αριθμούς, πράγμα που σημαίνει αύξηση των σφαλμάτων στρογγύλευσης στους υπολογισμούς. Το πρόβλημα αυτό αποφεύγεται με μία τροποποίηση της μεθόδου των δυνάμεων.

4.3 Τροποποίηση της μεθόδου των δυνάμεων

Η τροποποίηση της μεθόδου των δυνάμεων συνίσταται από τα παρακάτω τρία διαδοχικά βήματα σε κάθε επανάληψη

$$\begin{aligned} y_{j_m}^{(m)} &= \max_j \left| y_j^{(m)} \right| = \| \, y^{(m)} \, \|_\infty, \\ z^{(m)} &= \frac{1}{y_{j_m}^{(m)}} y^{(m)} \\ y^{(m+1)} &= A z^{(m)}, \quad m = 0, 1, 2, \ldots \end{aligned} \tag{4.25}$$

Εργαζόμενοι με ανάλογο τρόπο όπως και στη προηγούμενη παράγραφο, η αντίστοιχη έκφραση της $y^{(m)}$ θα δίνεται από τη σχέση

$$y^{(m)} = \frac{1}{y_{j_0}^{(0)} y_{j_1}^{(1)} \ldots y_{j_{m-1}}^{(m-1)}} \lambda_1^m \left[\alpha_1 x^{(1)} + \sum_{i=2}^{\nu} \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^m x^{(i)} \right] \tag{4.26}$$

Επίσης έχουμε ότι

$$\begin{aligned} z^{(m-1)} &= \frac{1}{y_{j_{m-1}}^{(m-1)}} y^{(m-1)} = \\ &= \frac{1}{y_{j_{m-1}}^{(m-1)}} \left[\frac{1}{y_{j_0}^{(0)} y_{j_1}^{(1)} \ldots y_{j_{m-2}}^{(m-2)}} \lambda_1^{m-1} \left(\alpha_1 x^{(1)} + \sum_{i=2}^{\nu} \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^{m-1} x^{(1)} \right) \right] \end{aligned} \tag{4.27}$$

Από τις (4.26) και (4.27) έχουμε

$$\lim_{m \to \infty} \frac{y_{j_{m-1}}^{(m)}}{z_{j_{m-1}}^{(m-1)}} = \lambda_1$$

αλλά $z_{j_{m-1}}^{(m-1)} = 1$ (το $z^{(n-1)}$ είναι κανονικοποιημένο) και η παραπάνω σχέση γράφεται

$$\lim_{m \to \infty} y_{j_{m-1}}^{(m)} = \lambda_1. \tag{4.28}$$

Η ακολουθία λοιπόν που δημιουργείται από τις συνιστώσες $\jmath_{m-1}$ του διανύσματος $y^{(m)}$ τείνει στη μεγαλύτερη κατά μέτρο ιδιοτιμή. Υπενθυμίζεται ότι η συνιστώσα $\jmath_{m-1}$ του διανύσματος $y^{(m)}$ είναι εκείνη που αντιστοιχεί στην απόλυτα μεγαλύτερη συνιστώσα του προηγούμενου διανύσματος $y^{(m-1)}$. Για την εύρεση του αντίστοιχου ιδιοδιανύσματος $x^{(1)}$ παρατηρούμε ότι, αν ο δείκτης $\jmath_m$ από μία ορισμένη τιμή του m και πέρα παραμένει σταθερός, τότε η αντίστοιχη με την (4.9) σχέση είναι η

$$\lim_{m-\infty} z^{(m)} = cx^{(1)} \tag{4.29}$$

όπου c σταθερά τέτοια ώστε η απόλυτα μεγαλύτερη συνιστώσα του $cx^{(1)}$ να είναι μονάδα. Άρα η ακολουθία των διανυσμάτων $z^{(m)}$ συγκλίνει προς το κανονικοποιημένο ιδιοδιάνυσμα που αντιστοιχεί στην λ_1. Σημειώνεται τέλος ότι η σύγκλιση της μεθόδου είναι γραμμική. Η ταχύτητα σύγκλισης της ακολουθίας $\left\{y^{(m)}_{\jmath_{m-1}}\right\}_{m=1}^{\infty}$ προς την λ_1 προσδιορίζεται, όπως αναφέρθηκε, από τους λόγους $|\lambda_\jmath/\lambda_1|^m$ για $\jmath = 2(1)n$ και ιδιαίτερα από τον λόγο $|\lambda_2/\lambda_1|^m$. Με άλλα λόγια η τάξη σύγκλισης είναι $\bigcirc((\lambda_2/\lambda_1)^m)$. Επομένως για μεγάλα m έχουμε

$$\left|y^{(m)}_{\jmath_{m-1}} - \lambda_1\right| \simeq k \left|\frac{\lambda_2}{\lambda_1}\right|^m$$

όπου k σταθερά, πράγμα που σημαίνει ότι

$$\lim_{m\to\infty} \frac{\left|y^{(m+1)}_{\jmath_m} - \lambda_1\right|}{\left|y^{(m)}_{\jmath_{m-1}} - \lambda_1\right|} \simeq \left|\frac{\lambda_2}{\lambda_1}\right|$$

ή

$$\varepsilon^{(m+1)} \simeq \left|\frac{\lambda_2}{\lambda_1}\right| \varepsilon^{(m)}$$

όπου $\varepsilon^{(m)} = \left|y^{(m)}_{\jmath_{m-1}} - \lambda_1\right|$, δηλαδή η ταχύτητα σύγκλισης είναι γραμμική.

4.4 Ο αλγόριθμος της μεθόδου των δυνάμεων

Σαν αρχικό διάνυσμα $y^{(0)}$ στη μέθοδο των δυνάμεων λαμβάνουμε συνήθως το $y^{(0)} = (1, 1, \ldots, 1)^T$. Ο αλγόριθμος της μεθόδου είναι ο παρακάτω:

1. Διάβασε τη διάσταση n του Πίνακα A, τα στοιχεία $\alpha_{ij}, i, j = 1(1)n$, το αρχικό διάνυσμα $y_i, i = 1(1)n$, την ανεκτικότητα tol και το μέγιστο αριθμό επαναλήψεων M.

2. Να τεθεί

$$k = 0$$
$$\lambda_0 = 0$$

3. Να βρεθεί ένας ακέραιος p τέτοιος ώστε

$$|y_p| = \max_{1 \le i \le n} |y_i|$$

4. Για $i = 1(1)n$ να υπολογιστεί

$$z_i = \frac{1}{y_p} y_i$$

5. Όσο ισχύει $k \le M$ να εκτελούνται τα βήματα 5.1-5.7

   5.1 Για $i = 1(1)n$ να τεθεί

$$y_i = \sum_{j=1}^{n} a_{ij} z_j$$

   5.2 Να βρεθεί ένας ακέραιος p τέτοιος ώστε

$$|y_p| = \max_{1 \le i \le n} |y_i|$$

5.3 Να τεθεί

$$\lambda_1 = y_p$$

5.4 Αν $y_p = 0$ τότε τύπωσε (Ο A έχει ιδιοτιμή 0, επέλεξε νέο αρχικό διάνυσμα και άρχισε πάλι την διαδικασία). Τέλος.

5.5 Για $i = 1(1)n$ να υπολογισθεί

$$z_i = \frac{1}{y_p} y_i$$

5.6 Αν

$$|\lambda_0 - \lambda_1| < tol$$

τότε τύπωσε (λ_1, z). Τέλος.

5.7 Να τεθεί

$$k = k + 1$$
$$\lambda_0 = \lambda_1$$

6. Τύπωσε (Όχι σύγκλιση μετά από M επαναλήψεις). Τέλος.

4.5 Τεχνικές επιτάχυνσης της μεθόδου των δυνάμεων

Επειδή η ταχύτητα σύγκλισης της μεθόδου των δυνάμεων είναι γενικά αργή γιαυτό θα πρέπει να βρεθούν τρόποι επιτάχυνσης της. Στη συνέχεια θα παρουσιάσουμε ορισμένες τέτοιες τεχνικές.

4.5.1 Η μέθοδος του Aitken

Η μέθοδος του Aitken μπορεί να χρησιμοποιηθεί για την επιτάχυνση οποιασδήποτε ακολουθίας που συγκλίνει γραμμικά. Έστω $\{P_n\}_{n=0}^{\infty}$ μία ακολουθία που έχει όριο p, δηλαδή για $\varepsilon_n = p_n - p$ έχουμε

$$\varepsilon_n = K_n \varepsilon_{n-1}, \quad n = 1, 2, \ldots \tag{4.30}$$

και οι σταθερές K_n είναι τέτοιες ώστε $|K_n| < 1$. Αλλά $K_n \to K$ για $n \to \infty$ και έτσι για αρκετά μεγάλες τιμές του n έχουμε

$$\frac{p_{n+2}-p}{p_{n+1}-p} \simeq \frac{p_{n+1}-p}{p_n-p} \tag{4.31}$$

οπότε λύνοντας ως προς p λαμβάνουμε

$$p \approx \frac{p_{n+2}p_n - p_{n+1}^2}{p_{n+2}-2p_{n+1}+p_n}$$

ή

$$p \approx p_n - \frac{(p_{n+1}-p_n)^2}{p_{n+2}-2p_{n+1}+p_n} \tag{4.32}$$

Είναι φανερό λοιπόν ότι η ακολουθία που ορίζεται από την

$$\hat{p}_n = p_n - \frac{(\Delta p_n)^2}{\Delta^2 p_n} \tag{4.33}$$

όπου

$$\Delta p_n = p_{n+1} - p_n \quad \text{και} \quad \Delta^2 p_n = p_{n+2} - 2p_{n+1} + p_n \tag{4.34}$$

έχει όριο το p.

Η μέθοδος αυτή είναι γνωστή σαν η Δ^2- μέθοδος του Aitken. Είναι δυνατόν να αποδειχθεί (αφήνεται σαν άσκηση για τον αναγνώστη) ότι

$$\lim_{n\to\infty} \frac{\hat{p}_n - p}{p_n - p} = 0. \tag{4.35}$$

Επίσης κάτω από επιπρόσθετες υποθέσεις η σύγκλιση της νέας ακολουθίας είναι δεύτερης τάξης. Με βάση τα παραπάνω είναι δυνατόν να τροποποιηθεί ο αλγόριθμος της μεθόδου των δυνάμεων έτσι ώστε να παράγεται η ακολουθία

$$\tilde{\lambda}_1 = \lambda_1 - \frac{(\Delta\lambda_1)^2}{\Delta^2\lambda_1} \tag{4.36}$$

4.5.2 Η μέθοδος των πηλίκων του Rayleigh

Στην περίπτωση που ο πίνακας A είναι πραγματικός και συμμετρικός, τότε είναι δυνατόν να επιταχυνθεί η σύγκλιση προς την μεγαλύτερη κατά μέτρο ιδιοτιμή χρησιμοποιώντας τη μέθοδο των πηλίκων του Rayleigh.

Ορισμός

Για κάθε διάνυσμα $x \neq 0$ η ποσότης

$$\frac{(x, Ax)}{(x, x)}$$

καλείται πηλίκο του Rayleigh που αντιστοιχεί στο x.

Ένα βασικό αποτέλεσμα που δείχνει τη σπουδαιότητα των πηλίκων του Rayleigh είναι το παρακάτω:

Θεώρημα 5.1

Αν ο $A \in \mathbb{R}^{nxn}$ είναι συμμετρικός και $x \neq 0$ είναι ένα αυθαίρετο διάνυσμα, τότε

$$\lambda_1 = \underset{x \neq 0}{max} \frac{(x, Ax)}{(x, x)} = \frac{(x^{(1)}, Ax^{(1)})}{(x^{(1)}, x^{(1)})}$$

και

$$\lambda_n = \underset{x \neq 0}{min} \frac{(x, Ax)}{(x, x)} = \frac{(x^{(n)}, Ax^{(n)})}{(x^{(n)}, x^{(n)})} \tag{4.37}$$

όπου λ_1, λ_n είναι η μεγαλύτερη και η μικρότερη ιδιοτιμή, αντίστοιχα και $x^{(1)}, x^{(n)}$ ιδιοδιανύσματα του A που αντιστοιχούν στις λ_1 και λ_n.

Από το παραπάνω θεώρημα παρατηρούμε ότι ο υπολογισμός της λ_1 είναι ένα πρόβλημα βελτιστοποίησης. Το ενδιαφέρον μας όμως εδώ είναι η χρήση των πηλίκων του Rayleigh για την επιτάχυνση της σύγκλισης της μεθόδου των δυνάμεων. Έτσι ας θεωρήσουμε το βασικό επαναληπτικό σχήμα της μεθόδου των δυνάμεων που είναι το

$$y^{(m+1)} = Ay^{(m)}$$

τότε εύκολα παρατηρούμε, χρησιμοποιώντας την (4.4), ότι

$$(y^{(m)}, y^{(m-1)}) = (y^{(m)}, Ay^{(m)})$$
$$= \sum_{i=1}^{n} \alpha_i^2 \lambda_i^{2m+1} \tag{4.38}$$

καθόσον

$$(x^{(i)}, x^{(j)}) = \delta_{ij} = \begin{cases} cc1, & \text{αν } i = j \\ 0, & \text{αν } i \neq j \end{cases}$$

αφού ο A είναι συμμετρικός. Επίσης

$$(y^{(m)}, y^{(m)}) = \sum_{i=1}^{n} \alpha_i^2 \lambda_i^{2m} \tag{4.39}$$

Από τις (4.38) και (4.39) έχουμε

$$\frac{(y^{(m)}, Ay^{(m)})}{(y^{(m)}, y^{(m)})} = \lambda_1 + \bigcirc((\lambda_i/\lambda_1)^{2m}) \tag{4.40}$$

η οποία συγκρινόμενη με την (4.5) δείχνει ότι το πηλίκο του Rayleigh που αντιστοιχεί στο $y^{(m)}$ γενικά θα συγκλίνει ταχύτερα ($\bigcirc(\lambda_i/\lambda_1)^{2m}$) από την μέθοδο των δυνάμεων ($\bigcirc(\lambda_i/\lambda_1)^{m}$).

4.5.3 Ο αλγόριθμος της μεθόδου των πηλίκων του Rayleigh

1. Διάβασε τη διάσταση n του Πίνακα A, τα στοιχεία $\alpha_{ij}, i, j = 1(1)n$, το αρχικό διάνυσμα $y_i, i = 1(1)n$, την ανεκτικότητα tol και το μέγιστο αριθμό επαναλήψεων M.

2. Να τεθεί

$$k = 0$$
$$\lambda_0 = 0$$

3. Για $i = 1(1)n$ να υπολογιστεί η ποσότης

$$z_i = y_i / \parallel y \parallel_2$$

4. Όσο ισχύει $k \leq M$ να εκτελούνται τα βήματα 4.1 - 4.6

   4.1 Για $i = 1(1)n$ να υπολογισθεί

   $$y_i = \sum_{j=1}^{n} \alpha_{ij} z_j$$

   4.2 Για $i = 1(1)n$ να υπολογισθεί

   $$\lambda = \sum_{i=1}^{n} z_i y_i$$

   4.3 Αν $\parallel y \parallel_2 = 0$ τότε τύπωσε (Ο A έχει ιδιοτιμή 0, επέλεξε νέο αρχικό διάνυσμα και άρχισε πάλι τη διαδικασία). Τέλος.

   4.4 Για $i = 1(1)n$ να υπολογισθεί η ποσότης

   $$z_i = y_i / \parallel y \parallel_2$$

   4.5 Αν

   $$|\lambda - \lambda_0| < \text{tol}$$

   τότε τύπωσε (λ, z). Τέλος.

   4.6 Να τεθεί

   $$k = k + 1$$
   $$\lambda_0 = \lambda$$

5. Τύπωσε (Όχι σύγκλιση μετά από M επαναλήψεις). Τέλος.

4.5.4 Μετατόπιση της αρχής των αξόνων (Shift of Origin)

Η παρακάτω Πρόταση συνδέει τις ιδιοτιμές και τα ιδιοδιανύσματα των πινάκων A και $A - qI$, όπου $A \in \mathbb{C}^{nn}$ και $q \in \mathbb{C}$.

Πρόταση 5.2. Οι πίνακες $A - qI$ και A έχουν τα ίδια ιδιοδιανύσματα και αν λ_i είναι ιδιοτιμή του A τότε η αντίστοιχη ιδιοτιμή του $A - qI$ είναι η $\lambda_i - q$.

Απόδειξη

Αν $Ax^{(i)} = \lambda_i x^{(i)}$ τότε

$$(A - qI)x^{(i)} = Ax^{(i)} - qIx^{(i)} = (\lambda_i - q)x^{(i)}.\blacksquare$$

Αφαιρώντας λοιπόν την ποσότητα q από τα διαγώνια στοιχεία του A έχει σαν αποτέλεσμα την αφαίρεση της q από τις ιδιοτιμές. Μια άλλη ερμηνεία είναι ότι η αρχή των αξόνων έχει μετατοπισθεί κατά q στο μιγαδικό επίπεδο που περιέχει τις ιδιοτιμές. Ας υποθέσουμε τώρα ότι ο $A \in \mathbb{R}^{nn}$ έχει n γραμμικά ανεξάρτητα ιδιοδιανύσματα και όλες οι ιδιοτιμές του είναι πραγματικές και ικανοποιούν τη σχέση

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3 \geq \ldots \geq |\lambda_{n-1}| \geq |\lambda_n| \tag{4.41}$$

Αν αφαιρέσουμε την ποσότητα $q \in \mathbb{R}$ με $q \notin (|\lambda_n|, |\lambda_1|)$ από τα διαγώνια στοιχεία του A, τότε ανεξάρτητα από την τιμή της q, η μεγαλύτερη κατά μέτρο ιδιοτιμή του $A - qI$ θα είναι πάντα η $\lambda_1 - q$ ή η $\lambda_n - q$. Ας υποθέσουμε ότι θέλουμε να προσδιορίσουμε την λ_1. Λόγω της Πρότασης 5.2 οι ιδιοτιμές το $A - qI$ θα είναι οι $\mu_i = \lambda_i - q$. Η ταχύτητα σύγκλισης της μεθόδου των δυνάμεων, χρησιμοποιώντας τώρα τον πίνακα $A - qI$ αντί του A, εξαρτάται από την ποσότητα

$$\max_{i \neq 1} \left| \frac{\lambda_i - q}{\lambda_1 - q} \right| \tag{4.42}$$

Όσο μικρότερη η ανωτέρω ποσότης, τόσο ταχύτερη η σύγκλιση της μεθόδου. Αρκεί δηλαδή να εκλέξουμε το q τέτοιο ώστε να ελαχιστοποιείται η ποσότης

$$\max_{i \neq 1} |\lambda_i - q| \tag{4.43}$$

Λόγω του θεωρήματος 4-2.1, η (4.43) γίνεται ελάχιστη αν $q = 1/2(\lambda_2 + \lambda_n)$. Όμοια εργαζόμενοι βρίσκουμε ότι η μέγιστη ταχύτητα σύγκλισης στην $\lambda_n - q$ επιταχύνεται αν επιλέξουμε $q = 1/2(\lambda_1 + \lambda_{n-1})$. Με τη μέθοδο αυτή μπορούμε να υπολογίσουμε τόσο την λ_1 όσο και την λ_n, ωστόσο όμως χρειαζόμαστε κάποιες εκτιμήσεις των ιδιοτιμών λ_2 και λ_n (ή των λ_1 και λ_{n-1}) πράγμα που απαιτεί επιπλέον υπολογισμούς στην πράξη και είναι ένα μειονέκτημα αυτής της μεθόδου.

4.6 Η αντίστροφη μέθοδος των δυνάμεων

Οι μέθοδοι που αναπτύχθηκαν στις προηγούμενες παραγράφους για τον προσδιορισμό της μεγαλύτερης /μικρότερης κατά μέτρο ιδιοτιμής και του αντίστοιχου ιδιοδιανύσματος έχουν αργή ταχύτητα σύγκλισης. Στη συνέχεια θα αναπτυχθεί η αντίστροφη μέθοδος των δυνάμεων η οποία έχει το πλεονέκτημα να υπολογίζει μια οποιαδήποτε ιδιοτιμή και το αντίστοιχο ιδιοδιάνυσμα και να έχει γρήγορη ταχύτητα σύγκλισης.

Λήμμα 6.1. Οι πίνακες A και A^{-1} έχουν τα ίδια ιδιοδιανύσματα και αν λ_i είναι μια ιδιοτιμή του A τότε η αντίστοιχη ιδιοτιμή του A^{-1} είναι η $1/\lambda_i$.

Απόδειξη

Αν $Ax^{(i)} = \lambda_i x^{(i)}$ τότε πολ/ζοντας από αριστερά με A^{-1} έχουμε $1/\lambda_i x^{(i)} = A^{-1}x^{(i)}$.■

Ας υποθέσουμε πάλι ότι ο $A \in \mathbb{R}^{nn}$, έχει n γραμμικά ανεξάρτητα ιδιοδιανύσματα και όλες οι ιδιοτιμές του είναι πραγματικές. Επίσης αν γνωρίζουμε κάποια ποσότητα $q \in \mathbb{R}$ η οποία βρίσκεται πλησιέστερα στην απλή ιδιοτιμή λ_k του A από οποιαδήποτε άλλη ιδιοτιμή του, τότε θα ισχύει

$$|\lambda_k - q| < |\lambda_i - q|, \qquad i = 1(1)n, \quad i \neq k \tag{4.44}$$

Δηλαδή η ιδιοτιμή $\lambda_k - q$ είναι η μικρότερη κατά απόλυτο τιμή ιδιοτιμή του πίνακα $A - qI$. Λόγω δε και του ανωτέρω Λήμματος 6.1, αν αντί του A χρησιμοποιήσουμε τον πίνακα $(A - qI)^{-1}$ στο βασικό επαναληπτικό σχήμα της μεθόδου των δυνάμεων, τότε είναι δυνατόν

να υπολογισθεί η ποσότης $1/\lambda_k - q$ και από αυτήν η λ_k. Πράγματι, αν αντί της (4.3) χρησιμοποιήσουμε την

$$(A - qI)y^{(m+1)} = y^{(m)}, \qquad m = 0, 1, 2, \ldots \tag{4.45}$$

όπου $y^{(0)} \neq 0$ αυθαίρετο διάνυσμα είναι δυνατόν να υπολογισθεί η απόλυτα μεγαλύτερη ιδιοτιμή του $(A - qI)^{-1}$ δηλαδή η $1/\lambda_k - q$ και το αντίστοιχο ιδιοδιάνυσμα. Είναι φανερό τώρα ότι η ταχύτητα σύγκλισης της μεθόδου εξαρτάται από την ποσότητα

$$\underset{i \neq k}{max} \left| \frac{\lambda_k - q}{\lambda_i - q} \right| \tag{4.46}$$

αφού

$$\begin{aligned} y^{(m)} &= (A - qI)^{-1} y^{(m-1)} = (A - qI)^{-m} y^{(0)} \\ &= \frac{\alpha_1}{(\lambda_1 - q)^m} x^{(1)} + \frac{\alpha_2}{(\lambda_2 - q)^m} x^{(2)} + \ldots + \frac{\alpha_n}{(\lambda_n - q)^m} x^{(n)} \\ &= \frac{1}{(\lambda_k - q)^m} \left[\alpha_1 \left(\frac{\lambda_k - q}{\lambda_1 - q} \right)^m x^{(1)} + \ldots + \alpha_k x^{(k)} + \ldots + \alpha_n \left(\frac{\lambda_k - q}{\lambda_n - q} \right)^m x^{(n)} \right] \end{aligned}$$

Έτσι η επιλογή της ποσότητας q καθορίζει και την ταχύτητα σύγκλισης της μεθόδου. Παρατηρούμε λοιπόν ότι [βλ. (4.46)] όσο πλησιέστερα η q είναι στην ιδιοτιμή λ_k τόσο ταχύτερη θα είναι και η σύγκλιση της μεθόδου. Επειδή η q μπορεί να εκλεγεί αυθαίρετα, μπορούμε να βρούμε μια προσέγγιση σε οποιαδήποτε ιδιοτιμή του A. Ο προσδιορισμός των $y^{(m)}$ γίνεται από την επίλυση των συστημάτων

$$(A - qI)y^{(m)} = y^{(m-1)}, \qquad m = 1, 2, \ldots \tag{4.47}$$

Στην πράξη φυσικά θα πρέπει τα διανύσματα να κανονικοποιούνται, με άλλα λόγια, να χρησιμοποιείται η παραλλαγή της μεθόδου των δυνάμεων. Παρατηρούμε επίσης ότι τα συστήματα που προκύπτουν από την (4.47) έχουν τον ίδιο πίνακα και διαφορετικά δεύτερα μέλη. Γιαυτό και θα πρέπει, αν χρησιμοποιήσουμε άμεσους μεθόδους, να σχηματίσουμε την LU διάσπαση του $A - qI$ μόνο μία φορά. Αν λοιπόν χρησιμοποιήσουμε κανονικοποιημένα διανύσματα και την LU μέθοδο τότε το βασικό επαναληπτικό σχήμα θα είναι το παρακάτω:

$$Lz = z^{(m)}$$
$$Uy^{(m+1)} = z \tag{4.48}$$

όπου

$$LU = A - qI$$

Ο αλγόριθμος της αντίστροφης μεθόδου των δυνάμεων αφήνεται σαν άσκηση για τον αναγνώστη.

4.7 Υπολογισμός των υπερεχουσών ιδιοτιμών

Υπάρχουν πολλές μέθοδοι για τον προσδιορισμό των άλλων υπερεχουσών κατά μέτρο ιδιοτιμών από τη στιγμή που υπολογισθεί η μεγαλύτερη. Στη συνέχεια θα αναφερθούμε σε μία μόνο μέθοδο που βασίζεται σε μετασχηματισμούς ομοιότητας. Ας υποθέσουμε ότι οι ιδιοτιμές ενός πίνακα A ικανοποιούν τη σχέση

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_m| >> |\lambda_{m+1}| \geq \ldots \geq |\lambda_n| \tag{4.49}$$

δηλαδή οι τιμές $|\lambda_1|, |\lambda_2|, \ldots, |\lambda_m|$ απέχουν αρκετά η μία από την άλλη. Τότε η λ_1 μπορεί να υπολογισθεί με τη μέθοδο των δυνάμεων και απομένει ο υπολογισμός των άλλων ιδιοτιμών που υπερέχουν, των $\lambda_2, \lambda_3, \ldots, \lambda_m$. Μερικές από τις πιο χρήσιμες μεθόδους για τον προσδιορισμό αυτών των ιδιοτιμών είναι εκείνες οι οποίες σχηματίζουν ένα νέο πίνακα από τον αρχικό με κάποια μορφή διατάραξης (deflation). Ο νέος πίνακας κατασκευάζεται κατά τέτοιο τρόπο ώστε να έχει σαν ιδιοτιμές μόνο τις υπόλοιπες άγνωστες ιδιοτιμές του αρχικού πίνακα. Η επαναληπτική εφαρμογή της διαδικασίας αυτής θα υπολογίσει όλες τις υπόλοιπες υπερέχουσες ιδιοτιμές και τα αντίστοιχα ιδιοδιανύσματα. Οι πιο εύχρηστες μέθοδοι διαταραχής είναι εκείνες που βασίζονται στους μετασχηματισμούς ομοιότητας. Για την περιγραφή της μεθόδου υποθέτουμε κατ' αρχήν ότι η ιδιοτιμή λ_1 και το αντίστοιχο ιδιοδιάνυσμα $x^{(1)}$ του πίνακα A_1 έχουν υπολογιστεί. Έστω τώρα H_1 ένας μη ιδιάζων πίνακας τέτοιος ώστε

$$H_1 x^{(1)} = k e^{(1)} \tag{4.50}$$

όπου $k \neq 0$ και $e^{(1)} = (1, 0, 0, \ldots, 0)^T$. Αν αναβάλουμε τη διαδικασία εύρεσης του H_1, τότε έχουμε

$$A_1 x^{(1)} = \lambda_1 x^{(1)}$$

από την οποία λαμβάνουμε

$$H_1 A_1 H_1^{-1} (H_1 x^{(1)}) = \lambda_1 H_1 x^{(1)} \tag{4.51}$$

η οποία λόγω της (4.50) γράφεται

$$H_1 A_1 H_1^{-1} e^{(1)} = \lambda_1 e^{(1)} \tag{4.52}$$

που δηλώνει ότι η πρώτη στήλη του πίνακα $H_1 A_1 H_1^{-1}$ πρέπει να είναι η $\lambda_1 e^{(1)}$, άρα μπορούμε να γράψουμε

$$A_2 = H_1 A_1 H_1^{-1} \left[\begin{array}{c|c} \lambda_1 & b^T \\ \hline 0 & B_2 \end{array}\right] \tag{4.53}$$

όπου ο πίνακας B_2 είναι $n-1$ τάξης και το διάνυσμα b έχει $n-1$ στοιχεία. Επειδή ο A_2 έχει τις ίδιες ιδιοτιμές με τον A_1, έπεται ότι ο πίνακας B_2 έχει ιδιοτιμές τις $\lambda_2, \lambda_3, \ldots, \lambda_n$. Μπορούμε λοιπόν να εργαστούμε με τον πίνακα B_2 προκειμένου να προσδιορίσουμε την επόμενη ιδιοτιμή λ_2 και το αντίστοιχο ιδιοδιάνυσμα $y^{(2)}$ του B_2 που ικανοποιούν την

$$B_2 y^{(2)} = \lambda_2 y^{(2)} \tag{4.54}$$

Αυτό που απομένει είναι η εύρεση του ιδιοδιανύσματος $x^{(2)}$ του A_1 που αντιστοιχεί στην λ_2. Έστω $z^{(2)}$ το ιδιοδιάνυσμα του A_2 που αντιστοιχεί στην λ_2, τότε

$$A_2 z^{(2)} = \lambda_2 z^{(2)} \tag{4.55}$$

ή

$$H_1 A_1 H_1^{-1} z^{(2)} = \lambda_2 z^{(2)}$$

ή

$$A_1 (H_1^{-1} z^{(2)}) = \lambda_2 (H_1^{-1} z^{(2)})$$

συνεπώς

$$x^{(2)} = H_1^{-1} z^{(2)} \tag{4.56}$$

αφού $A_1 x^{(2)} = \lambda_2 x^{(2)}$. Αρκεί λοιπόν να υπολογισθεί το $z^{(2)}$ για την εύρεση του $x^{(2)}$. Η (4.55) λόγω της (4.53) γράφεται

$$\left| \begin{array}{c|c} \lambda_1 & b^T \\ \hline 0 & B_2 \end{array} \right| z^{(2)} = \lambda_2 z^{(2)} \tag{4.57}$$

λόγω όμως της (4.54) μπορούμε να λάβουμε

$$z^{(2)} = \left[\begin{array}{c} \alpha \\ y^{(2)} \end{array} \right] \tag{4.58}$$

όπου α ένα βαθμωτό μέγεθος, που προσδιορίζεται από την (4.57) ή την

$$\lambda_1 \alpha + b^T y^{(2)} = \lambda_2 \alpha$$

ή

$$\alpha = \frac{b^T y^{(2)}}{\lambda_2 - \lambda_1} \tag{4.59}$$

Συνοψίζοντας παρατηρούμε ότι τα $\lambda_2, y^{(2)}$ υπολογίζονται με τη μέθοδο των δυνάμεων [βλ. (4.54)], το $z^{(2)}$ υπολογίζεται από την (4.58), όπου το α δίνεται από την (4.59). Έχοντας υπολογίσει το $z^{(2)}$, το $x^{(2)}$ βρίσκεται από την (4.56). Συνεχίζοντας κατ' αυτό τον τρόπο υπολογίζουμε τις υπόλοιπες υπερέχουσες ιδιοτιμές και τα αντίστοιχα ιδιοδιανύσματα του A_1. Είναι φανερό ότι οι διαδοχικές διαταραχές του A_1 θα τον μετασχηματίσουν, στο όριο, σε ένα άνω τριγωνικό πίνακα. Στη συνέχεια θα περιγράψουμε ένα τρόπο για την εκλογή του H_1 έτσι ώστε η διαδικασία της διατάραξης να είναι αριθμητικά ευσταθής. Διαλέγουμε τον H_2 τέτοιον ώστε

$$H_1 = L_1 I_{1,p} \tag{4.60}$$

όπου L_1 είναι ένας στοιχειώδης κάτω τριγωνικός πίνακας και $I_{1,p}$ ένας μεταθετικός πίνακας, όπου p είναι τέτοιο ώστε η $x_p^{(1)}$ είναι η μεγαλύτερη κατά μέτρο συνιστώσα του $x^{(1)}$. Από τις (4.50) και (4.60) έχουμε ότι

$$y = I_{1,p} x^{(1)} \tag{4.61}$$

και

$$L_1 y = k e^{(1)} \tag{4.62}$$

όπου

$$L_1 = \begin{vmatrix} 1 & & & \\ -y_2/y_1 & 1 & & 0 \\ \vdots & & \ddots & \\ -y_n/y_1 & & & 1 \end{vmatrix} \tag{4.63}$$

και $k = y_1 = x_p^{(1)}$.

Η εισαγωγή του μεταθετικού πίνακα $I_{1,p}$ ουσιαστικά ορίζει μία διαδικασία οδήγησης, η οποία απαιτεί τα στοιχεία του H_1 να είναι κατά μέτρο μικρότερα ή ίσα από τη μονάδα, εξασφαλίζοντας έτσι την αριθμητική ευστάθεια όπως και στη μέθοδο της απαλοιφής του Gauss.

Παράδειγμα

Έστω ο πίνακας

$$A_1 = \begin{bmatrix} 2 & 3 & 2 \\ 10 & 3 & 4 \\ 3 & 6 & 1 \end{bmatrix}$$

Με την μέθοδο των δυνάμεων υπολογίζουμε την ιδιοτιμή $\lambda_1 = 11.0$ και το αντίστοιχο ιδιοδιάνυσμα $x^{(1)} = (0.5, 1.0, 0.75)^T$. Παρατηρούμε ότι $k = y_1 = 1.0 = x_2^{(1)}$, άρα $p = 2$ και $y = (1.0, 0.5, 0.75)^T$ έτσι

$$L_1 = \begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ -0.75 & 0 & 1 \end{bmatrix}$$

Άρα

$$
\begin{aligned}
A_2 &= L_1 I_{1,2} A_1 I_{1,2} L_1^{-1} \\
&= \begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ -0.75 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 & 10 & 4 \\ 3 & 2 & 2 \\ 6 & 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0.75 & 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 3 & 10 & 4 \\ 1.5 & -3 & 0 \\ 3.75 & -4.5 & -2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0.75 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 11 & 10 & 4 \\ 0 & -3 & 0 \\ 0 & -4.50 & -2 \end{bmatrix}
\end{aligned}
$$

Παρατηρούμε ότι ο πολ/μός με L_1^{-1} δεν χρειαζόταν αφού γνωρίζουμε ότι η πρώτη στήλη του A_2 είναι ίση με $\lambda_1 e^{(1)}$. Από την τελευταία σχέση έχουμε ότι

$$
B_2 = \begin{bmatrix} -3 & 0 \\ -4.5 & -2 \end{bmatrix}
$$

και οι υπόλοιπες ιδιοτιμές του είναι -3 και -2. Ο υπολογισμός των ιδιοδιανυσμάτων αφήνεται σαν άσκηση για τον αναγνώστη.

Ο αλγόριθμος για τη μέθοδο της διατάραξης αφήνεται σαν άσκηση για τον αναγνώστη.

4.8 Η μέθοδος του Jacobi

Η μέθοδος των δυνάμεων, σε συνδυασμό με τις διάφορες τεχνικές, χρησιμοποιείται συνήθως για τον προσδιορισμό μερικών ιδιοτιμών και των αντίστοιχων ιδιοδιανυσμάτων και αυτό γιατί η διαδοχική εφαρμογή της μεθόδου της διατάραξης έχει σαν αποτέλεσμα την αύξηση των σφαλμάτων στρογγύλευσης. Στη συνέχεια θα στρέψουμε το ενδιαφέρον μας σε εκείνες τις μεθόδους που βρίσκουν όλο το ιδιοσύστημα (ιδιοτιμές και ιδιοδιανύσματα) ταυτόχρονα και όχι μετά από την εφαρμογή κάποιας τεχνικής διατάραξης. Οι μέθοδοι αυτές βασίζονται στους μετασχηματισμούς ομοιότητας τους οποίους χρησιμοποιούν για να μετασχηματίσουν τον αρχικό πίνακα A σε έναν άλλο του οποίου το ιδιοσύστημα είναι εύκολο να υπολογισθεί.

Θεώρημα 8.1

Αν ένας πίνακας A έχει n γραμμικά ανεξάρτητα ιδιοδιανύσματα $x_1, x_2, \ldots, x_n$ που αντιστοιχούν στις ιδιοτιμές $\lambda_1, \lambda_2, \ldots, \lambda_n$ και αν $P = [x_1, x_2, \ldots, x_n]$ τότε ο P είναι μη ιδιάζων και

$$P^{-1}AP = D \tag{4.64}$$

όπου

$$D = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix} \tag{4.65}$$

Απόδειξη

Ο P είναι ένας πίνακας του οποίου οι στήλες είναι γραμμικά ανεξάρτητες, άρα ο P είναι μη ιδιάζων. Επίσης

$$\begin{aligned} AP &= A\,[x_1, x_2, \ldots x_n] \\ &= [Ax_1, Ax_2, \ldots Ax_n] \\ &= [\lambda_1 x_1, \lambda_2 x_2, \ldots \lambda_n x_n] \\ &= [x_1, x_2, \ldots x_n] \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix} \\ &= PD \blacksquare \end{aligned}$$

Αν τώρα περιοριστούμε σε Ερμειτιανούς πίνακες τότε ισχύει το παρακάτω θεώρημα.

Θεώρημα 8.2

Αν ο A είναι ένας Ερμειτιανός πίνακας τάξης n, με ιδιότητες $\lambda_1, \lambda_2 \ldots \lambda_n$ (όχι κατ' ανάγκη διακεκριμένες), τότε υπάρχει ένας μοναδιαίος (unitary) πίνακας P, τέτοιος ώστε

$$P^*AP = D \tag{4.66}$$

όπου D είναι ένας πραγματικός διαγώνιος πίνακας με διαγώνια στοιχεία τα $\lambda_1, \lambda_2, \ldots \lambda_n$.

Από την (4.66) και επειδή ο P είναι μοναδιαίος ($P^* = P^{-1}$) έπεται ότι

$$AP = PD$$

πράγμα που σημαίνει ότι οι στήλες του P είναι ιδιοδιανύσματα του A. Είναι φυσικό λοιπόν να προσπαθήσουμε να δημιουργήσουμε τον πίνακα P χρησιμοποιώντας διαδοχικούς στοιχειώδεις μοναδιαίους μετασχηματισμούς. Πράγματι στη συνέχεια κατασκευάζουμε μία προσέγγιση του P τέτοια ώστε ο πίνακας P^*AP να έχει τα εκτός της διαγωνίου του στοιχεία ασήμαντα σε σύγκριση με τα διαγώνια στοιχεία του A. Έτσι η μέθοδος μας θα πρέπει να έχει σαν αποτέλεσμα την ελάττωση των εκτός της κυρίας διαγωνίου στοιχείων του πίνακα A. Στο σημείο αυτό είναι χρήσιμο να διατυπώσουμε το παρακάτω θεώρημα.

Θεώρημα 8.3

Αν ο πίνακας A είναι Ερμειτιανός, τότε ο πίνακας $B = P^*AP$ όπου P^* μοναδιαίος (unitary) είναι Ερμειτιανός και έχει την ίδια Ευκλείδια norm με εκείνη του A, δηλαδή

$$\| B \|_E^2 = \| A \|_E^2 \tag{4.67}$$

όπου

$$\| A \|_E^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2 \tag{4.68}$$

Απόδειξη

Παρατηρούμε ότι $B^* = B$. Επίσης

$$\begin{aligned} \| A \|_E^2 &= tr(A^* A) \\ &= tr(A^2) \\ &= \sum_{i=1}^{n} \lambda_i^2 \end{aligned}$$

όπου λ_i είναι ιδιοτιμή του A. Επειδή όμως ο $B = P^* AP$ είναι όμοιος με τον A, πράγμα που σημαίνει ότι οι πίνακες A και B έχουν τις ίδιες ιδιοτιμές, συνεπάγεται ότι η (4.67) ισχύει.■

Συνεπώς αν κατασκευάσουμε την ακολουθία των Ερμειτιανών πινάκων

$$\begin{aligned} A_1 &= A \\ A_2 &= P_1^* A_1 P_1 \\ &\vdots \\ A_{s+1} &= P_s^* A_s P_s \\ &\vdots \end{aligned} \tag{4.69}$$

από το θεώρημα 8.3 συνεπάγεται ότι η Ευκλείδια norm του κάθε πίνακα A_s είναι ίση με εκείνη του $A_1 = A$. Αλλά ο σκοπός μας είναι να ελαττωθεί το μέγεθος των εκτός της κυρίας διαγωνίου στοιχείων του A, χρησιμοποιώντας τους μετασχηματισμούς ομοιότητας. Άρα αν ορίσουμε τις ποσότητες

$$E_{(s)} = \sum_{\substack{i,j=1 \\ i \neq j}} \left| a_{ij}^{(s)} \right|^2 \tag{4.70}$$

$$D_{(s)} = \sum_{i=1}^{n} \left| a_{ij}^{(s)} \right|^2 \tag{4.71}$$

και διαλέξουμε την ακολουθία $\left\{ P_{(s)} \right\}, s = 1, 2, \ldots$ στην (4.69)τέτοια ώστε

$$E_{(s+1)} \leq E_{(s)} \tag{4.72}$$

τότε

$$D_{(s+1)} \geq D_{(s)} \tag{4.73}$$

αφού $\| A_{s+1} \|_E = \| A_s \|_E$. Είναι φανερό λοιπόν ότι μεγιστοποιώντας την ποιότητα $D_{(s+1)} - D_{(s)}$ η ταχύτητα σύγκλισης της ακολουθίας $\{A_s\}, s = 1, 2, \ldots$ θα είναι η μεγαλύτερη δυνατή. Στη συνέχεια θα αναπτύξουμε τη μέθοδο του Jacobi (1846) η οποία μετασχηματίζει τον πίνακα A_1 σε διαγώνιο μορφή εκτελώντας μία ακολουθία από επίπεδες περιστροφές. Για λόγους ευκολίας θα υποθέσουμε ότι ο πίνακας A είναι πραγματικός και συμμετρικός οπότε ο P είναι ένας ορθογώνιος πίνακας και η (4.66) γράφεται σαν $P^T A P = D$ ενώ οι (4.69) μετασχηματίζονται στις

$$\begin{aligned} A_1 &= A \\ A_2 &= P_1^T A_1 P_1 \\ A_3 &= P_2^T A_2 P_2 \\ &\vdots \\ A_{s+1} &= P_s^T A_s P_s \\ &\vdots \end{aligned}$$

Αν επιλέξουμε σαν πίνακα P τον

$$P = \begin{array}{c} \\ \\ \\ r \\ \\ \\ q \\ \\ \\ \end{array} \begin{bmatrix} 1 & & & & & & & & & \\ & \ddots & & & & \mathbf{0} & & & & \\ & & 1 & & & & & & & \\ & & & \cos\theta & & & & \sin\theta & & \\ & & & & 1 & & & & & \\ & \mathbf{0} & & \vdots & & \ddots & & \vdots & & \mathbf{0} \\ & & & & & & 1 & & & \\ & & & -\sin\theta & & & & \cos\theta & & \\ & & & & & & & & 1 & \\ & & & & & \mathbf{0} & & & & \ddots \\ & & & & & & & & & 1 \end{bmatrix} \tag{4.74}$$

τότε έχουμε το ακόλουθο θεώρημα.

Θεώρημα 8.4

Αν ο A είναι ένας πραγματικός και συμμετρικός πίνακας τότε: (a) $P^TP = I$ (δηλ. ο P είναι ορθογώνιος) (b) Αν $B = (b_{ij}) = P^TAP$, τότε ο B είναι συμμετρικός και

(i) Για $\quad i \neq r$ και $\quad j \neq r \quad b_{ij} = \mathrm{a}_{ij}$ (4.75)

(ii) Για $\quad i \neq r, q$

$$b_{ir} = b_{ri} = a_{ir}\cos\theta - a_{iq}\sin\theta b_{iq} = b_{qi} = a_{ir}\sin\theta + a_{iq}cos\theta$$

(iii)

$$\begin{aligned} b_{rr} &= a_{rr}\cos^2\theta + a_{qq}\sin^2\theta - 2a_{rq}\sin\theta\cos\theta \\ b_{qq} &= a_{rr}\sin^2\theta + a_{qq}\cos^2\theta + 2a_{rq}\sin\theta\cos\theta \end{aligned} \tag{4.76}$$

(iv)

$$b_{rq} = b_{qr} = \frac{1}{2}(a_{rr} - a_{qq})\sin 2\theta + a_{rq}\cos 2\theta \tag{4.77}$$

Απόδειξη

Έστω $P = [p_1, p_2 \dots p_n]$ τότε για $j \neq r, q$ έχουμε

$$p_j = e_j = (0, 0, \dots, 0, \underset{\substack{\uparrow \\ j}}{1}, 0, \dots 0)^T \quad \text{και}$$

$$P_r = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \cos\theta \\ \vdots \\ -\sin\theta \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \\ \\ \\ \leftarrow \ r \\ \\ \leftarrow \ q \\ \\ \\ \end{matrix} \qquad P_q = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \sin\theta \\ \vdots \\ \cos\theta \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \\ \\ \\ \leftarrow \ r \\ \\ \leftarrow \ q \\ \\ \\ \end{matrix}$$

Επομένως

$$P^T P = \begin{bmatrix} p_1^T \\ p_2^T \\ \vdots \\ p_n^T \end{bmatrix} \qquad [p_1 \ p_2 \dots p_n] = \begin{bmatrix} p_1^T p_1 & p_1^T p_2 & \dots & p_1^T p_n \\ p_2^T p_1 & p_2^T p_2 & \dots & p_2^T p_n \\ \vdots & \vdots & & \vdots \\ p_n^T p_1 & p_n^T p_2 & \dots & p_n^T p_n \end{bmatrix}$$

Το τυχόν σημείο (i, j) του $P^T P$ είναι το $p_i^T p_j$. Επίσης για $i, j \neq r, q$ έχουμε

$$p_i^T p_j = e_i^T e_j = \begin{cases} i & \text{αν } i = j \\ 0 & \text{αν } i \neq j \end{cases}$$

Για $i = r$ και $j \neq r, q$

$$p_i^T p_j = p_i^T e_j = 0$$

Για $i = r$ και $j = r$

$$p_r^T p_r = \cos^2\theta + \sin^2\theta = 1$$

Όμοια για $i = q$ και $\jmath = q$

$$p_q^T p_q = \cos^2\theta + \sin^2\theta = 1$$

Για $i = r$ και $\jmath = q$

$$p_r^T p_q = \cos\theta \sin\theta - \sin\theta \cos\theta = 0$$

Επειδή ο πίνακας $P^T P$ είναι συμμετρικός έχουμε ότι και $p_q^T p_r = 0$. Άρα απεδείχθη ότι για όλα τα i και $\jmath$

$$p_i^T p_\jmath = \begin{cases} 1 & \text{αν } i = \jmath \\ 0 & \text{αν } i \neq \jmath \end{cases}$$

δηλαδή $P^T P = I$ και ο P είναι ένας ορθογώνιος πίνακας. Για την απόδειξη του (b) θεωρούμε τον πίνακα A σαν $A = [a_1\ a_2\ \ldots\ a_n]$ όπου $a_i = [a_{1i}\ a_{2i}\ \ldots\ a_{ni}]^T$. Τότε

$$B = P^T AP = \begin{bmatrix} p_1^T \\ p_2^T \\ \vdots \\ p_n^T \end{bmatrix} \quad A[p_1\ p_2\ \ldots\ p_n] = \begin{bmatrix} p_1^T \\ p_2^T \\ \vdots \\ p_n^T \end{bmatrix}$$

$$[Ap_1\ Ap_2\ \ldots\ Ap_n] = \begin{bmatrix} p_1^T Ap_1 & p_1^T Ap_2 & \ldots & p_1^T Ap_n \\ p_2^T Ap_1 & p_2^T Ap_2 & \ldots & p_2^T Ap_n \\ \vdots & \vdots & \vdots & \vdots \\ p_n^T Ap_1 & p_n^T Ap_2 & \ldots & p_n^T Ap_n \end{bmatrix}$$

Συνεπώς, $b_{i\jmath} = p_i^T Ap_\jmath$ και διακρίνουμε πάλι τις προηγούμενες περιπτώσεις.

Για $i \neq r, q$ και $\jmath \neq r, q$ έχουμε

$$b_{i\jmath} = p_i^T Ap_\jmath = e_i^T Ae_\jmath = e_i^T a_\jmath = a_{i\jmath}.$$

Για $i \neq r, q$ και $\jmath = r$

$$b_{ir} = p_i^T A p_r = e_i^T \begin{bmatrix} a_{11} & \dots & a_{1r} & \dots & a_{1q} & \dots & a_{1n} \\ \vdots & & & & & & \vdots \\ a_{r1} & \dots & a_{rr} & \dots & a_{rq} & \dots & a_{rn} \\ \vdots & & & & & & \vdots \\ a_{q1} & \dots & a_{qr} & \dots & a_{qq} & \dots & a_{qn} \\ \vdots & & & & & & \vdots \\ a_{n1} & \dots & a_{nr} & \dots & a_{nq} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ \cos\theta \\ \vdots \\ -\sin\theta \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \\ \\ \leftarrow r \\ \\ \leftarrow q \\ \\ \\ \end{matrix}$$

$$= e_i^T \begin{bmatrix} a_{1r}\cos\theta - a_{1q}\sin\theta \\ a_{2r}\cos\theta - a_{2q}\sin\theta \\ \vdots \\ a_{nr}\cos\theta - a_{nq}\sin\theta \end{bmatrix} = a_{ir}\cos\theta - a_{iq}\sin\theta.$$

Για $i \neq r, q$ και $\jmath = q$ έχουμε

$$b_{iq} = p_i^T A p_q = e_i^T \begin{bmatrix} a_{1r}\cos\theta + a_{1q}\sin\theta \\ a_{2r}\cos\theta + a_{2q}\sin\theta \\ \vdots \\ a_{nr}\cos\theta + a_{nq}\sin\theta \end{bmatrix} = a_{ir}\cos\theta + a_{iq}\sin\theta.$$

Για $i = \jmath = r$ έχουμε

$$b_{rr} = p_r^T A p_r = [0 \ \dots \ \overset{r\downarrow}{\cos\theta} \ \dots \ \overset{q\downarrow}{-\sin\theta} \ \dots \ 0] \begin{bmatrix} a_{1r}\cos\theta - a_{1q}\sin\theta \\ a_{2r}\cos\theta - a_{2q}\sin\theta \\ \vdots \\ a_{nr}\cos\theta - a_{nq}\sin\theta \end{bmatrix}$$

$$= a_{rr}\cos^2\theta - a_{rq}\cos\theta\sin\theta - a_{qr}\sin\theta\cos\theta + a_{qq}\sin^2\theta.$$

$$= a_{rr}\cos^2\theta - 2a_{rq}\cos\theta\sin\theta + a_{qq}\sin^2\theta \qquad (a_{rq} = a_{qr}).$$

Για $i = \jmath = q$ έχουμε

$$b_{qq} = p_q^T A p_q = [0 \ \ldots \ \overset{r_\downarrow}{\sin\theta} \ \ldots \ \overset{q_\downarrow}{\cos\theta} \ \ldots \ 0] \begin{bmatrix} a_{1r}\sin\theta + a_{1q}\cos\theta \\ a_{2r}\sin\theta + a_{2q}\cos\theta \\ \vdots \\ a_{nr}\sin\theta + a_{nq}\cos\theta \end{bmatrix}$$

$$= a_{rr}\sin^2\theta + a_{rq}\cos\theta\sin\theta + a_{qr}\sin\theta\cos\theta + a_{qq}\cos^2\theta.$$

$$= a_{rr}\sin^2\theta + 2a_{rq}\cos\theta\sin\theta + a_{qq}\cos^2\theta.$$

Για $i = r$ και $\jmath = q$ έχουμε

$$b_{rq} = p_r^T A p_q = [0 \ \ldots \ \overset{r_\downarrow}{\cos\theta} \ \ldots \ \overset{q_\downarrow}{-\sin\theta} \ \ldots \ 0] \begin{bmatrix} a_{1r}\sin\theta + a_{1q}\cos\theta \\ a_{2r}\sin\theta + a_{2q}\cos\theta \\ \vdots \\ a_{nr}\sin\theta + a_{nq}\cos\theta \end{bmatrix}$$

$$= a_{rr}\cos\theta\sin\theta + a_{rq}\cos^2\theta - a_{qr}\sin^2\theta - a_{qq}\cos\theta\sin\theta.$$

$$= 1/2(a_{rr} - a_{qq})\sin^2\theta + a_{rq}\cos^2\theta.$$

Παρατηρούμε τώρα ότι ο B είναι συμμετρικός, και συνεπώς η απόδειξη του (b) είναι πλήρης.■

Ο ορθογώνιος πίνακας P που ορίστηκε από την (4.75) καλείται πίνακας επίπεδης περιστροφής καθόσον περιστρέφει τους άξονες r και q κατά μία γωνία θ. Η κλασική μέθοδος του Jacobi αναζητεί τα στοιχεία του πίνακα A_s της ακολουθίας (4.74) που βρίσκονται πάνω από την κύρια διαγώνιο και προσδιορίζει το στοιχείο $a_{rq}^{(s)}$ με τη μεγαλύτερη απόλυτο τιμή, το οποίο καλείται οδηγό στοιχείο. Η αναζήτηση αυτή γίνεται μόνο στο άνω τριγωνικό μέρος του A_s αφού είναι συμμετρικός. Στη συνέχεια η γωνία περιστροφής θ_s εκλέγεται έτσι ώστε το στοιχείο $a_{rq}^{(s+1)}$ να είναι μηδέν. Αν λοιπόν απαιτήσουμε $b_{rq} = 0$, τότε από την (4.77) έχουμε

$$b_{rq} = \frac{1}{2}(a_{rr} - a_{qq})\sin 2\theta + a_{rq}\cos 2\theta = 0$$

ή

$$\tan 2\theta = \frac{2a_{rq}}{a_{qq} - a_{rr}}, \quad a_{qq} - a_{rr} \neq 0 \tag{4.78}$$

Η δε γωνία θ εκλέγεται έτσι ώστε

$$-\frac{\pi}{4} \leq \theta \leq \frac{\pi}{4}.$$

Αν $a_{qq} - a_{rr} = 0$ τότε $a_{rq} \cos\theta = 0$ και $\theta = \pi/4$. Στην περίπτωση όπου $a_{rq} = 0$ δεν χρειάζεται καμία περιστροφή ενώ ο A_s θα είναι ήδη διαγώνιος. Άρα υπάρχει πάντα μία γωνία θ τέτοια ώστε $b_{rq} = b_{qr} = 0$. Δεν χρειάζεται όμως να λύσουμε την τριγωνομετρική εξίσωση (4.78) προκειμένου να βρούμε την θ καθόσον χρειαζόμαστε μόνο τα $\sin\theta$ και $\cos\theta$. Έτσι αν θέσουμε

$$b = |a_{qq} - a_{rr}|$$

και

$$a = 2a_{rq}\text{sign}(a_{qq} - a_{rr}) \tag{4.79}$$

όπου

$$\text{sign}x = \frac{x}{|x|} = \begin{cases} 1, & \text{αν } x > 0 \\ -1, & \text{αν } x < 0 \end{cases}$$

τότε η (4.77) γράφεται ως εξής

$$\tan 2\theta = \frac{\alpha}{\beta}$$

Για τον υπολογισμό του $\cos\theta$ έχουμε $\sec^2\theta = 1 + \tan^2 2\theta$ ή

$$\sec^2 2\theta = 1 + \frac{\alpha^2}{\beta^2}$$

άρα

$$\cos^2 2\theta = \frac{\beta^2}{\alpha^2 + \beta^2}$$

Επειδή $|\theta| \leq \pi/4$ διαλέγουμε το $\cos 2\theta$ να είναι θετικό, έτσι

$$\cos 2\theta = \frac{\beta}{\sqrt{\alpha^2 + \beta^2}}$$

Επίσης εύκολα βρίσκουμε ότι

$$\sin 2\theta = \frac{\alpha}{\sqrt{\alpha^2 + \beta^2}}$$

Από την $\cos 2\theta = 2\cos^2\theta - 1$ έχουμε τελικά ότι

$$\cos\theta = \left[\frac{1}{2}\left(1 + \frac{\beta}{\sqrt{\alpha^2 + \beta^2}}\right)\right]^{1/2} \tag{4.80}$$

ενώ από την $\sin 2\theta = 2\cos\theta\sin\theta$ λαμβάνουμε

$$\sin\theta = \frac{\alpha}{2\cos\theta\sqrt{\alpha^2 + \beta^2}} \tag{4.81}$$

Συνοπτικά τα βήματα του αλγορίθμου της μεθόδου του Jacobi είναι τα παρακάτω.

1. Προσδιορισμός των r και q τέτοιων ώστε

$$a_{rq}^{(s)} = \max_{1 \leq i,j \leq n} \left|a_{ij}^{(s)}\right|.$$

2. Υπολογισμός των α και β από την (4.79)

3. Υπολογισμός των $\cos\theta$ και $\sin\theta$ από τις (4.80) και (4.81).

4. Υπολογισμός του A_{s+1} χρησιμοποιώντας τις (4.75), (4.76) και (4.8).Αντί της (4.77) το βήμα 4 ολοκληρώνεται θέτοντας $a_{rq}^{(s+1)} = a_{qr}^{(s+1)} = 0$. Θα πρέπει ωστόσο να χρησιμοποιηθεί η (4.77) σαν υπολογιστικός έλεγχος μαζί με έναν έλεγχο του μεγέθους της ποσότητας $|\sin^2\theta + \cos^2\theta - 1|$

5. Ο σχηματισμός της ακολουθίας των A_s θα σταματήσει όταν $E_{(s)} \leq \epsilon, \quad \epsilon > 0$, γιατί στη φάση αυτή τα διαγώνια στοιχεία του A_s είναι καλές προσεγγίσεις των ιδιοτιμών $\lambda_1, \lambda_2, \ldots, \lambda_n$.

Είναι εύκολο τώρα να αποδειχθεί, χρησιμοποιώντας τις σχέσεις του Θεωρήματος 8.4, ότι

$$E_{(s+1)} = E_{(s)} - 2(a_{rq}^{(s)})^2 \tag{4.82}$$

το οποίο δίνει

$$D_{(s+1)} = D_{(s)} + 2(a_{rq}^{(s)})^2 \tag{4.83}$$

πράγμα που αποδεικνύει ότι η ακολουθία των πινάκων A_s συγκλίνει σε διαγώνιο μορφή και τελικά ο πίνακας A_s θα είναι ουσιαστικά ένας διαγώνιος πίνακας. Θα πρέπει όμως να δειχθεί ότι η ακολουθία τείνει σε ένα σταθερό διαγώνιο πίνακα. Δηλαδή θα πρέπει να δειχθεί ότι

$$|a_{rr}^{(s+1)} - a_{rr}^{(s)}| \to 0$$

για s $\to \infty$. Από την (4.8) έχουμε ότι

$$\begin{aligned} a_{rr}^{(s+1)} - a_{rr}^{(s)} &= -a_{rr}^{(s)}(1 - \cos^2\theta) - 2a_{rq}^{(s)}\cos\theta\sin\theta + a_{qq}^{(s)}\sin^2\theta \\ &= (a_{qq}^{(s)} - a_{rr}^{(s)})\sin^2\theta - 2a_{rq}^{(s)}\cos\theta\sin\theta \end{aligned}$$

Αν $a_{qq}^{(s)} - a_{rr}^{(s)} = 0$ έχουμε αμέσως ότι

$$|a_{rr}^{(s+1)} - a_{rr}^{(s)}| > |a_{rq}^{(s)}|$$

αφού $|\theta| = \pi/4$. Αν $a_{qq}^{(s)} - a_{rr}^{(s)} \neq 0$ τότε χρησιμοποιώντας την (4.78) και απλές τριγωνομετρικές σχέσεις βρίσκουμε

$$|a_{rr}^{(s+1)} - a_{rr}^{(s)}| = |a_{rq}^{(s)}|\,|\tan\theta|.$$

Αλλά $|\theta| \leq \pi/4$ και συνεπώς γενικά

$$|a_{rr}^{(s+1)} - a_{rr}^{(s)}| \leq |a_{rq}^{(s)}|.$$

Αν είμαστε στη φάση όπου $|a_{rq}^{(s+1)}| \leq \varepsilon$ τότε και

$$|a_{rr}^{(s+1)} - a_{rr}^{(s)}| \leq \varepsilon.$$

όμοια

$$|a_{qq}^{(s+1)} - a_{qq}^{(s)}| \leq \varepsilon.$$

Συνεπώς ο αλγόριθμος του Jacobi δημιουργεί μία ακολουθία πινάκων οι οποίοι τείνουν σε ένα σταθερό διαγώνιο πίνακα, ο οποίος είναι όμοιος με τον αρχικό. Αποδεικνύεται ότι η μέθοδος είναι και ευσταθής. Η κωδικοποίηση της μεθόδου αφήνεται σαν άσκηση για τον αναγνώστη.

4.8.1 Παραλλαγές της μεθόδου του Jacobi

Επειδή η αναζήτηση, σε κάθε βήμα του αλγορίθμου, για το στοιχείο a_{rq} με το μέγιστο μέτρο είναι χρονοβόρα, γιαυτό υπάρχουν δύο εναλλακτικές στρατηγικές που είναι περισσότερο χρήσιμες στην πράξη. Η πρώτη είναι εκείνη κατά την οποία δεν γίνεται καμία αναζήτηση. Τα στοιχεία μηδενίζονται σύμφωνα με μία κυκλική σειρά. Η συνηθισμένη σειρά είναι εκείνη των γραμμών, δηλαδή μηδενίζονται τα στοιχεία στις θέσεις $(1,2)$, $(1,3)$, $\ldots$, $(1,n)$, $(2,3)$, $(2,4)$, $\ldots$, $(2,n)$, $(3,4)$, $\ldots$, $(n-1,n)$. Σε κάθε δε βήμα ελέγχεται αν το στοιχείο που πρόκειται να μηδενισθεί είναι μεγάλο σε μέγεθος συγκρινόμενο με το άθροισμα των τετραγώνων των διαγωνίων στοιχείων. Η μέθοδος αυτή είναι γνωστή σαν η **κυκλική** μέθοδος του **Jacobi**. Η άλλη παραλλαγή της μεθόδου είναι εκείνη κατά την οποία ορίζεται μία σταθερά (threshold) κατά την διάρκεια κάθε βήματος. Αν το υποψήφιο για μηδενισμό στοιχείο του πίνακα είναι μικρότερο από την σταθερά αυτή κατά απόλυτο τιμή, τότε αγνοείται (διαφορετικά μηδενίζεται). Φυσικά, η σταθερά αυτή ελαττώνεται σε κάθε βήμα. Επειδή η ποσότης $E_{(s)}$ τείνει στο μηδέν, θα χρειαστεί η σταθερά αυτή να γίνει τελικά ένας αριθμός πολύ κοντά στο μικρότερο αριθμό που μπορεί να παρασταθεί από τον υπολογιστή που χρησιμοποιείται. Στην περίπτωση αυτή μπορούμε να υποθέσουμε ότι η μέθοδος έχει συγκλίνει. Η μέθοδος αυτή είναι γνωστή σαν η **σταθερή μέθοδος του Jacobi** (the threshold Jacobi method). Είναι δυνατόν να αποδειχθεί ότι για τις παραπάνω τρεις μεθόδους του Jacobi η σύγκλιση είναι τετραγωνική με την έννοια ότι

$$E_{(s+N)} \leq K[E_{(s)}]^2$$

όπου το K εξαρτάται από την απόσταση των ιδιοτιμών και την τάξη n του πίνακα A και $N = 1/2n(n-1)$. Αποδεικνύεται επίσης ότι η μεγίστη τιμή της ποσότητας $D_{(s+1)} - D_{(s)}$ λαμβάνεται όταν ισχύει η (4.78) (γιατί;).

4.8.2 Υπολογισμός των ιδιοδιανυσμάτων

Είναι φανερό ότι

$$A_{s+1} = (P_1 P_2 \ldots P_s)^T A_1 (P_1 P_2 \ldots P_s)$$

και στη περίπτωση όπου η μέθοδος έχει συγκλίνει, έχουμε ότι

$$P^T A_1 P = D \tag{4.84}$$

όπου

$$P = P_1 P_2 \dots P_s \tag{4.85}$$

Συνεπώς από την (4.84) έχουμε ότι οι στήλες του P δίνουν τα ιδιοδιανύσματα του πίνακα A. Αξίζει να σημειωθεί ότι τα ιδιοδιανύσματα προσδιορίζονται στη μέθοδο του Jacobi ταυτόχρονα με τις ιδιοτιμές. Επίσης τα ιδιοδιανύσματα που δίνονται από τον P είναι ορθογώνια. Στη περίπτωση όπου οι ιδιοτιμές του A είναι κοντά η μία με την άλλη, υπάρχει πιθανότητα τα ιδιοδιανύσματα να μην υπολογισθούν με μεγάλη ακρίβεια. Το γεγονός ότι, χρησιμοποιώντας την μέθοδο του Jacobi, μπορούμε πάντα να προσδιορίσουμε όλα τα ιδιοδιανύσματα είναι ένα σημαντικό πλεονέκτημα της μεθόδου. Ωστόσο, στη συνέχεια θα παρουσιαστούν ορισμένες ταχύτερες μέθοδοι, ιδιαίτερα όταν οι ιδιοτιμές είναι διακεκριμένες.

Άσκηση

Να εφαρμοστούν οι τρεις μέθοδοι του Jacobi για τον υπολογισμό των ιδιοτιμών και ιδιοδιανυσμάτων του πίνακα

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

Συγκρίνατε και σχολιάσατε τα αποτελέσματά σας.

4.9 Η μέθοδος του Givens

Ένα βασικό μειονέκτημα της μεθόδου του Jacobi είναι ότι τα στοιχεία του πίνακα που μηδενίζονται σε ένα βήμα μπορεί να γίνουν μη μηδενικά (με αρκετά μεγάλη τιμή) στα επόμενα βήματα. Το φαινόμενο αυτό έχει σαν συνέπεια ότι η μέθοδος του Jacobi είναι μη πεπερασμένη. Έτσι όταν απαιτείται μεγάλη ακρίβεια για τις ιδιοτιμές ή όταν τα διαγώνια στοιχεία του A δεν είναι μεγάλα σε μέγεθος, συγκρινόμενα με τα εκτός της διαγωνίου στοιχεία του A, είναι δυνατόν να έχουμε

πάρα πολύ αργή σύγκλιση της μεθόδου. Επίσης, συνήθως επιθυμούμε μόνο τον υπολογισμό των ιδιοτιμών και όχι των ιδιοδιανυσμάτων. Στη συνέχεια θα μελετήσουμε μεθόδους, οι οποίες μετασχηματίζουν ένα Ερμειτιανό πίνακα A σε έναν άλλο, του οποίου το ιδιοσύστημα είναι εύκολο να υπολογιστεί. Πιο συγκεκριμένα ο αρχικός πίνακας θα μετασχηματιστεί σε ένα πραγματικό συμμετρικό τριδιαγώνιο πίνακα με ένα πεπερασμένο αριθμό βημάτων.

Προκειμένου να αποφευχθεί το βασικό μειονέκτημα της μεθόδου του Jacobi, ο Givens (1954) ανέπτυξε μία μέθοδο, η οποία διατηρεί τα μηδενικά στοιχεία στις εκτός της διαγωνίου θέσεις από την στιγμή που θα δημιουργηθούν. Η μέθοδος αυτή ξεκινά με την εκλογή του $a_{23}, a_{24}, \ldots, a_{2n}$ σαν οδηγά στοιχεία αλλά αντί να επιλεγεί η γωνία θ τέτοια ώστε να μηδενίζονται τα στοιχεία αυτά (όπως στη μέθοδο Jacobi), η γωνία θ επιλέγεται έτσι ώστε να μηδενίζονται τα στοιχεία $a_{13}, a_{14}, \ldots, a_{1n}$(και τα συμμετρικά τους). Κατ' αυτό τον τρόπο ένας συμμετρικός πίνακας A μετασχηματίζεται στον συμμετρικό πίνακα της μορφής

x	x	0	0	0	...	0
x	x	x	x	x	...	x
0	x	x	x	x	...	x
0	x	x	x	x	...	x
0	x	x	x	x	...	x
⋮	⋮	⋮	⋮	⋮		⋮
0	x	x	x	x	...	x

ο οποίος έχει $n-2$ μηδενικά στην πρώτη γραμμή και στην πρώτη στήλη. Εκλέγοντας στη συνέχεια σαν οδηγά στοιχεία από την τρίτη γραμμή (στήλη) που βρίσκονται στις θέσεις $(3,4),(3,5),\ldots(3,n)$ και τη γωνία περιστροφής θ έτσι ώστε να μηδενίζονται τα στοιχεία στις θέσεις που έχουν υπογραμμιστεί δηλ. στις $(2,4),(2,5),\ldots,(2,n)$ παράγεται ο πίνακας

$$\left|\begin{array}{cc|ccccc} \mathrm{x} & \mathrm{x} & 0 & 0 & 0 & \dots & 0 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & 0 & 0 & \dots & 0 \\ \hline 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \dots & \mathrm{x} \\ 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \dots & \mathrm{x} \\ 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \dots & \mathrm{x} \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \dots & \mathrm{x} \end{array}\right|$$

ο οποίος έχει $n-3$ μηδενικά στη δεύτερη γραμμή και στη δεύτερη στήλη. Παρατηρούμε ότι τα μηδενικά της πρώτης γραμμής και της πρώτης στήλης θα διατηρηθούν αφού αυτά αντικαθίστανται από γραμμικούς συνδυασμούς μηδενικών [βλ. (4.76)]. Γενικά το r κύριο βήμα της όλης διαδικασίας αποτελείται από $n-r-1$ επί μέρους βήματα, κατά τα οποία δημιουργούνται μηδενικά διαδοχικά στις θέσεις $r+2, r+3, \dots, n$ της r γραμμής και στήλης και το μηδέν στην $(r-1,q)$ θέση παράγεται εκλέγοντας σαν οδηγό στοιχείο εκείνο της θέσης (r,q). Συνεχίζοντας αυτή τη διαδικασία η μέθοδος θα μετασχηματίσει τον αρχικό πίνακα σε ένα συμμετρικό τριδιαγώνιο μετά από $\sum_{i=1}^{n-2}(n-i-1) = (n-1)(n-2)/2$ περιστροφές, πράγμα που δείχνει ότι η μέθοδος Givens απαιτεί πεπερασμένο αριθμό βημάτων. Είναι φανερό λοιπόν ότι οι υπολογισμοί στη μέθοδο του Givens έχουν διαταχθεί κατά τέτοιο τρόπο ώστε να μπορούμε να μηδενίσουμε στοιχεία εκτός της κύριας διαγωνίου, ενώ την ίδια στιγμή να παραμένουν μηδενικά τα προηγούμενα (μηδενικά) στοιχεία του πίνακα. Χρησιμοποιώντας λοιπόν τον πίνακα P_s όπως ορίστηκε στην μέθοδο Jacobi μπορούμε να μηδενίσουμε αντί του $a_{rq}^{(s+1)}$ το στοιχείο $a_{r-1,q}^{(s+1)}$ (και το $a_{q,r-1}^{(s+1)}$). Αυτό μπορεί να γίνει αν θέσουμε $i = r-1$ στη δεύτερη σχέση της (;;) οπότε

$$b_{r-1,q} = b_{q,r-1} = a_{r-1,r}\sin\theta + a_{r-1,q}\cos\theta = 0 \tag{4.86}$$

η οποία ικανοποιείται αν

$$\sin\theta = -\alpha a_{r-1,q} \text{ και } \cos\theta = \alpha a_{r-1,r} \tag{4.87}$$

όπου

$$\alpha = \frac{1}{\sqrt{a_{r-1,r}^2 + a_{r-1,q}^2}} \tag{4.88}$$

Για την παραγωγή ενός τριδιαγώνιου πίνακα οι τιμές των r και q είναι $r = 2(1)n-1$ και $q = r+1(1)n$. Τέλος μπορεί να υπολογιστεί ότι περίπου $4/3n^3$ πολ/μοί χρειάζονται για να γίνει η τριγωνοποίηση σε σύγκριση με $2n^3$ πολ/μούς για ένα βήμα της μεθόδου Jacobi.

Παράδειγμα

Να μετατραπεί ο πίνακας

$$\begin{bmatrix} 1 & 2 & 1 & 2 \\ 2 & 2 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 2 & 1 & 1 & 1 \end{bmatrix}$$

σε διαγώνιο με τη μέθοδο Givens

Λύση

$$\alpha = \frac{1}{\sqrt{a_{1,2}^2 + a_{1,3}^2}} = \frac{1}{\sqrt{5}},\ \sin\theta = -\frac{1}{\sqrt{5}},\ \cos\theta = \frac{2}{\sqrt{5}},\ i = 1$$

$$b_{1,2} = a_{12}\cos\theta - a_{13}\sin\theta = 2\cdot\frac{2}{\sqrt{5}} - 1\left(-1\frac{1}{\sqrt{5}}\right) = \frac{5}{\sqrt{5}} = b_{12},\ i = 2$$

$$b_{2,4} = b_{4,2} = a_{42}\cos\theta - a_{43}\sin\theta = 1\cdot\left(\frac{2}{\sqrt{5}}\right) - 1\left(-\frac{1}{\sqrt{5}}\right) = \frac{3}{\sqrt{5}},\ i = 3$$

$$b_{34} = b_{4,3} = a_{42}\sin\theta + a_{43}\cos\theta = 1\cdot\left(-\frac{1}{\sqrt{5}}\right) + 1\cdot\left(\frac{2}{\sqrt{5}}\right) = \frac{1}{\sqrt{5}}$$

$$b_{22} = a_{22}\cos^2\theta + a_{33}\sin^2\theta - 2a_{23}\sin\theta\cos\theta = 1$$

$$b_{33} = a_{22}\sin^2\theta + a_{33}\cos^2\theta + 2a_{23}\sin\theta\cos\theta = 2$$

$$b_{23} = b_{32} = \frac{1}{2}(a_{22}\cdot a_{33})\sin 2\theta + a_{23}\cos 2\theta =$$

$$= \frac{1}{2}(2-1)\left(-\frac{4}{5}\right) + (-1)\left(\frac{3}{5}\right) = -1.$$

4.10 Η μέθοδος του Householder

Ο Householder (1958) ανέπτυξε μία μέθοδο, η οποία τριδιαγωνοποιεί ένα συμμετρικό πίνακα A χρησιμοποιώντας ακριβώς $n-2$ ορθογώνιους μετασχηματισμούς. Αυτοί οι μετασχηματισμοί είναι πιο πολύπλοκοι από εκείνους της απλής περιστροφής αλλά η πλήρης τριδιαγωνοποίηση του πίνακα απαιτεί μόνο περίπου τους μισούς υπολογισμούς σε σύγκριση με τη μέθοδο του Givens. Ο μετασχηματισμός του Householder χρησιμοποιεί τους πίνακες

$$P \in \mathbb{R}^{nn}$$

της μορφής

$$P = I \cdot 2ww^T \tag{4.89}$$

όπου $w = (w_i) \in \mathbb{R}^n$ είναι ένα διάνυσμα στήλη τέτοιο ώστε

$$w^T \mathrm{w} = 1 \tag{4.90}$$

Θεώρημα 10.1

Αν ο πίνακας P οριστεί από τις (4.89) και (4.90), τότε

(i) $P = P^T$
(ii) $P = P^{-1}$.

Απόδειξη

$$\begin{aligned}
\text{(i)} P^T &= I - 2(ww^T)^T = I - 2w^T \\
\text{(ii)} P^T P &= I - 2ww^T(I - 2ww^T) \\
&= I - 4ww^T + 4w(w^T w)w^T \\
&= I - 4ww^T + 4ww^T \\
&= I. \blacksquare
\end{aligned}$$

Άρα $P = P^T = P^{-1}$ και ο P είναι συμμετρικός και ορθογώνιος. Στη συνέχεια ας θεωρήσουμε $n-2$ μετασχηματισμούς Householder αρχίζοντας με τον συμμετρικό πίνακα $A \in \mathbb{R}^{nn}$. Έχουμε

$$
\begin{aligned}
A_1 &= A \\
A_2 &= P_2 A_1 P_2 \\
A_3 &= P_3 A_2 P_3 \\
&\vdots \\
A_{n-1} &= P_{n-1} A_{n-2} P_{n-1}
\end{aligned}
\tag{4.91}
$$

όπου ορίζουμε τους πίνακες

$$P_r, \quad r = 2(1)n - 1$$

από την

$$P_r = I - 2w^{(r)} w^{(r)^T} \tag{4.92}$$

με

$$
w^{(r)} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ w_r \\ w_{r+1} \\ \vdots \\ w_n \end{bmatrix}
\tag{4.93}
$$

και

$$w_r^2 + w_{r+1}^2 + \cdots + w_n^2 = 1 \tag{4.94}$$

Ο σκοπός μας είναι να μετασχηματίσουμε τον πραγματικό συμμετρικό πίνακα A_1 σε ένα πραγματικό συμμετρικό τριδιαγώνιο πίνακα A_{n-1} χρησιμοποιώντας $n-2$ ορθογώνιους μετασχηματισμούς ομοιότητας.

Προκειμένου να υπολογίσουμε τον

$$A_r = P_r A_{r-1} P_r \tag{4.95}$$

θα πρέπει να υπολογίσουμε πρώτα τα στοιχεία $w_r, w_{r+1}, \ldots, w_n$ του διανύσματος $w^{(r)}$ έτσι ώστε ο πίνακας $P_r A_{r-1} P_r$ να μηδενίζει τα $n-r$ στοιχεία εκτός των τριδιαγωνίων, στη γραμμή $r-1$ και στη στήλη $r-1$ του A_{r-1}. Για την εξαγωγή των τύπων υπολογισμού

των $w_r, w_{r+1}, \ldots, w_n$ ας εξετάσουμε αναλυτικά την περίπτωση όπου $n = 4$ και $r = 2$. Με άλλα λόγια έστω ο συμμετρικός πίνακας

$$A_1 = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

Στη συνέχεια θέλουμε να μηδενίσουμε τα συμμετρικά ζευγάρια δηλ. θέλουμε

$$a_{13} = a_{31} = 0$$

και

$$a_{14} = a_{41} = 0 \tag{4.96}$$

χρησιμοποιώντας τον μετασχηματισμό

$$A_2 = P_2 A_1 P_2 \tag{4.97}$$

όπου

$$P_2 = I - 2w^{(2)} w^{(2)^T} \tag{4.98}$$

και

$$w^{(2)} = \begin{bmatrix} 0 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} \tag{4.99}$$

με

$$w_2^2 + w_3^2 + w_4^2 = 1 \tag{4.100}$$

Σχηματίζοντας τον P_2 έχουμε ότι

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 - 2w_2^2 & -2w_2 \mathrm{w}_3 & -2w_2 \mathrm{w}_4 \\ 0 & -2w_3 w_2 & 1 - 2w_3^2 & -2w_3 \mathrm{w}_4 \\ 0 & -2w_4 w_2 & -2w_4 \mathrm{w}_3 & 1 - 2w_4^2 \end{bmatrix}$$

Επειδή ο A_2 είναι συμμετρικός δείχνουμε μόνο την 1η στήλη. Έτσι

$$A_2 = P_2 A_1 P_2$$

$$= P_2 \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-2\mathrm{w}_2^2 & -2\mathrm{w}_2\mathrm{w}_3 & -2\mathrm{w}_2\mathrm{w}_4 \\ 0 & -2\mathrm{w}_3\mathrm{w}_2 & 1-2\mathrm{w}_3^2 & -2\mathrm{w}_3\mathrm{w}_4 \\ 0 & -2\mathrm{w}_4\mathrm{w}_2 & -2\mathrm{w}_4\mathrm{w}_3 & 1-2\mathrm{w}_4^2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-2\mathrm{w}_2^2 & -2\mathrm{w}_2\mathrm{w}_3 & -2\mathrm{w}_2\mathrm{w}_4 \\ 0 & -2\mathrm{w}_3\mathrm{w}_2 & 1-2\mathrm{w}_3^2 & -2\mathrm{w}_3\mathrm{w}_4 \\ 0 & -2\mathrm{w}_4\mathrm{w}_2 & -2\mathrm{w}_4\mathrm{w}_3 & 1-2\mathrm{w}_4^2 \end{bmatrix} \begin{bmatrix} a_{11} & a'_{12} & a'_{13} & a'_{14} \\ a_{21} & a'_{22} & a'_{23} & a'_{24} \\ a_{31} & a'_{32} & a'_{33} & a'_{34} \\ a_{41} & a'_{42} & a'_{43} & a'_{44} \end{bmatrix}$$

$$= \begin{bmatrix} a''_{11} & a''_{12} & a''_{13} & a''_{14} \\ a_{21}(1-2w_2^2) + a_{31}(-2w_2w_3) + a_{41}(-2w_2\mathrm{w}_4) & a''_{22} & a''_{23} & a''_{24} \\ a_{21}(-2w_3w_2) + a_{31}(1-2w_3^2) + a_{41}(-2w_3\mathrm{w}_4) & a''_{32} & a''_{33} & a''_{34} \\ a_{21}(-2w_4w_2) + a_{31}(-2w_4w_3) + a_{41}(1-2w_4^2) & a''_{42} & a''_{43} & a''_{44} \end{bmatrix}$$

Για την απλότητα των υπολογισμών χρησιμοποιούμε τους τόνους για να διαφοροποιήσουμε τα στοιχεία του A_2 από τα αντίστοιχα του A_1 . Αν τώρα θέσουμε

$$p = w_2 a_{21} + w_3 a_{31} + w_4 a_{41} \tag{4.101}$$

τότε τα στοιχεία της πρώτης στήλης του A_2 γράφονται στην απλούστερη μορφή

$$\begin{aligned} a''_{11} &= a_{11} \\ a''_{21} &= a_{21} - 2w_2 p \\ a''_{31} &= a_{31} - 2w_3 p \\ a''_{41} &= a_{41} - 2w_4 p \end{aligned} \tag{4.102}$$

Έστω τώρα s_k^2 συμβολίζει το άθροισμα των τετραγώνων των στοιχείων κάτω από την κύρια διαγώνιο στην k στήλη. Τότε για τον A_2 έχουμε στην πρώτη στήλη

$$\begin{aligned} s_1^2 &= (a''_{21})^2 + (a''_{31})^2 + (a''_{41})^2 \\ &= (a_{21} - 2w_2 p)^2 + (a_{31} - 2w_3 p)^2 + (a_{41} - 2w_4 p)^2 \\ &= a_{21}^2 + a_{31}^2 + a_{41}^2 - 4p(a_{21}w_2 + a_{31}w_3 + a_{41}w_4) + 4p^2(w_2^2 + w_3^2 + w_4^2) \end{aligned} \tag{4.103}$$

ή

$$s_1^2 = a_{21}^2 + a_{31}^2 + a_{41}^2 - 4p^2 + 4p^2$$
$$= a_{21}^2 + a_{31}^2 + a_{41}^2$$

Συνεπώς η ποσότης s_1^2 είναι η ίδια για τους A_3 και A_2 δηλαδή παραμένει αναλλοίωτος κάτω από τον μετασχηματισμό του Householder.

Θεώρημα 10.2

Το άθροισμα των τετραγώνων των στοιχείων κάτω από την κύρια διαγώνιο της στήλης $r-1$ του πίνακα A_{r-1} παραμένει αναλλοίωτο κάτω από τον μετασχηματισμό του Householder $A_r = P_r A_{r-1} P_r$. Το ίδιο συμβαίνει, λόγω συμμετρίας και για τα στοιχεία δεξιά της κύριας διαγωνίου στη γραμμή $r-1$.

Απόδειξη

Αφήνεται σαν άσκηση για τον αναγνώστη

Ας θέσουμε τώρα

$$a''_{31} = a''_{41} = 0$$

τότε

$$\begin{aligned} a_{21} - 2w_2 p &= \pm s_1 (s_1^2 = (a''_{21})^2) \\ a_{31} - 2w_3 p &= 0 \\ a_{41} - 2w_4 p &= 0 \end{aligned} \tag{4.104}$$

όπου

$$s_1 = \sqrt{a_{21}^2 + a_{31}^2 + a_{41}^2}. \tag{4.105}$$

Αν πολ/σουμε τις παραπάνω εξισώσεις με w_2, w_3 και w_4 αντίστοιχα και προσθέτουμε τότε θα έχουμε

$$w_2 a_{21} + w_3 a_{31} + w_4 a_{41} - 2p(w_2^2 + w_3^2 + w_4^2) = \pm w_2 s_1$$

ή

$$p - 2p = \pm w_2 s_1$$

ή

$$p = \pm w_2 s_1 \tag{4.106}$$

Έτσι λοιπόν αντικαθιστώντας την τιμή του p στις εξισώσεις (4.104) λαμβάνουμε

$$\begin{aligned} a_{21} \pm 2w_2^2 s_1 &= \pm s_1 \\ a_{31} \pm 2w_3^2 s_1 &= 0 \\ a_{41} \pm 2w_4^2 s_1 &= 0 \end{aligned} \tag{4.107}$$

Οπότε έχουμε να λύσουμε τρεις μη-γραμμικές εξισώσεις με τρεις αγνώστους w_2, w_3 και w_4 . Οι παραπάνω εξισώσεις δίνουν

Στις παραπάνω σχέσεις πρέπει να αποφασίσουμε για τη χρήση του πρόσημου. Προκειμένου να αποφύγουμε τον μηδενισμό του αριθμητή στην πρώτη εξίσωση

Στη συνέχεια δεν χρησιμοποιούμε την τετραγωνική ρίζα για τον υπολογισμό του w_2 επειδή όλες οι παραπάνω σχέσεις είναι δευτέρου βαθμού ως προς w_2. Έτσι διαιρώντας την πρώτη εξίσωση με w_2 έχουμε

$$w_2 = \frac{a_{21}(\text{sign} : a_{21}) + 2s_1}{2w_2 s_1}$$

τότε το $w^{(2)}$ έχει την εξής απλή μορφή

$$w^{(2)} = \left(\frac{\text{sign} : a_{21}}{2w_2 s_1} \right) \begin{bmatrix} 0 \\ a21 + s_1(\text{sign} : a_{21}) \\ a_{31} \\ a_{41} \end{bmatrix} \tag{4.108}$$

ή

$$w^{(2)} = \beta_2 v^{(2)} \tag{4.109}$$

όπου

$$\beta_2 = \frac{\text{sign} : a_{21}}{2w_2 s_1} \qquad \text{και} \qquad \upsilon^{(2)} = \begin{bmatrix} 0 \\ a_{21} + s_1(\text{sign} : a_{21}) \\ a_{31} \\ a_{41} \end{bmatrix} \tag{4.110}$$

Εκφράζοντας τον P_2 συναρτήσει του $\upsilon^{(2)}$ αντί του $\mathrm{w}^{(2)}$ έχουμε

$$\begin{aligned} P_2 &= I - 2w^{(2)}w^{(2)^T} \\ &= I - 2\beta_2^{(2)}\upsilon^{(2)}\upsilon^{(2)^T} \\ &= I - 2\alpha_2\upsilon^{(2)}\upsilon^{(2)^T}. \end{aligned} \tag{4.111}$$

όπου

$$\alpha_2 = 2\beta_2^{(2)} - 2\frac{1}{2w_2^2 s_1^2} = \frac{1}{s_1^2 + s_1|a_{21}|} \tag{4.112}$$

λόγω της (4.110) και της πρώτης σχέσης των (;;). Χρησιμοποιώντας τα $\upsilon^{(2)}$ αντί των $w^{(2)}$ αποφεύγουμε τον υπολογισμό μιας τετραγωνικής ρίζας κατά τους υπολογισμούς μας. Έτσι λοιπόν έχουμε το θεώρημα

Θεώρημα 10.3

Ο μετασχηματισμός

$$P_r A_{r-1} P_r \tag{4.113}$$

όπου

$$P_r = I - \alpha_r \upsilon^{(r)}\upsilon^{(r)^T} \tag{4.114}$$

$$\alpha_r = \frac{1}{s_{r-1}^2 + s_{r-1}|a_{r,r-1}|} \tag{4.115}$$

$$\upsilon^{(r)} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \hat{a}_{r,r-1} \\ a_{r+1,r-1} \\ \vdots \\ a_{n,r-1} \end{bmatrix} \tag{4.116}$$

με

$$\hat{a}_{r,r-1} = a_{r,r-1} + s_{r-1}(\text{sign} : a_{r,r-1}) \tag{4.117}$$

και

$$s_{r-1} = \sqrt{a_{r,r-1}^2 + a_{r+1,r-1} + \cdots + a_{n,r-1}^2} \tag{4.118}$$

θα μηδενίσει τα $n-r$ εκτός των τριδιαγωνίων στοιχείων στη γραμμή $r-1$ και στήλη $r-1$ του A_{r-1} (η απόδειξη αφήνεται στην άσκηση).

Δεν χρειάζεται όμως να σχηματίζουμε τον P_r προκειμένου να εφαρμόσουμε τον

$$P_1 A_{r-1} P_r$$

καθόσον

$$\begin{aligned} A' &= PAP \\ &= (I - \alpha v v^T) A (I - \alpha v v^T) \\ &= A - \alpha v v^T A - \alpha A v v^T + \alpha^2 v (v^T A v) v^T \\ &= A - \alpha v v^T A - \alpha A v v^T + \alpha^2 (v^T A v) v v^T \\ &= A - (v q^T + q v^T) \end{aligned} \tag{4.119}$$

όπου

$$\begin{aligned} q &= u - \mu v \\ u &= 2Av \\ \mu &= \frac{\alpha}{2} v^T u \end{aligned} \tag{4.120}$$

Έτσι αν γνωρίζουμε το v τότε υπολογίζουμε το q και μπορούμε να χρησιμοποιήσουμε την (4.119) προκειμένου να εκτελέσουμε το μετασχηματισμό. Παρατηρούμε τέλος ότι μετά το μετασχηματισμό $P_r A_{r-1} P_r$ τα μετασχηματισμένα στοιχεία στη στήλη $r-1$ (με όμοια αποτελέσματα για τη γραμμή $r-1$) είναι τα

$$
\begin{aligned}
a'_{r-1,r-1} &= a_{r-1,r-1} \\
a'_{r,r-1} &= s_{r-1} \qquad (4.121) \\
a'_{r+1,r-1} &= 0 \\
&\vdots \\
a'_{n,r-1} &= 0 \qquad (4.122)
\end{aligned}
$$

Η δεύτερη εξίσωση της (4.121) είναι άμεση συνέπεια του θεωρήματος 10.2 και μπορεί να χρησιμοποιηθεί σαν ένας έλεγχος για την ακρίβεια των υπολογισμών. Τέλος ο αριθμός των πολ/μών για την τριδιαγωνοποίηση είναι ο μισός της μεθόδου του Givens.

4.11 Υπολογισμός του ιδιοσυστήματος ενός συμμετρικού τριδιαγώνιου πίνακα

Οι μέθοδοι του Givens και του Householder μετασχηματίζουν ένα συμμετρικό τριδιαγώνιο. Απομένει λοιπόν ο προσδιορισμός του ιδιοσυστήματος αυτού του πίνακα. Απομένει λοιπόν ο προσδιορισμός του ιδιοσυστήματος αυτού του πίνακα. Στο σημείο αυτό θα πρέπει να αποφασίσουμε αν χρειαζόμαστε το πλήρες ιδιοσύστημα (δηλαδή όλες τις ιδιοτιμές με ή χωρίς τα αντίστοιχα ιδιοδιανύσματα) ή απλά μερικές ιδιοτιμές και τα ισοδύναμα τους. Στη συνέχεια θα ασχοληθούμε με το τελευταίο πρόβλημα, ενώ για το πρώτο πρόβλημα υπάρχει μία περισσότερο αποτελεσματική μέθοδος.

Έστω ο συμμετρικός τριδιαγώνιος πίνακας

$$
\begin{bmatrix}
a_1 & b_1 & & \\
b_1 & a_2 & b_2 & \mathbf{0} \\
 & \ddots & \ddots & \ddots \\
\mathbf{0} & b_{n-2} & a_{n-1} & b_{n-1} \\
 & & b_{n-1} & a_n
\end{bmatrix} \qquad (4.123)
$$

Υποθέτουμε ότι $b_i \neq 0, \quad i = 1(1)n-1$, γιατί αν $b_i = 0$ για κάποιο i ο πίνακας είναι block τριδιαγώνιος και μπορούμε να εργαστούμε με

το κάθε ένα block χωριστά. Στη συνέχεια θα υπολογίσουμε την τιμή του χαρακτηριστικού πολυωνύμου του T χωρίς τον υπολογισμό των συντελεστών του. Αν $P_i(\lambda)$ είναι η ορίζουσα του οδηγού κύριου υποπίνακα, τάξης i του $T - \lambda I$, τότε για $i = 1, 2, 3$ λαμβάνουμε

$$\begin{aligned} P_1(\lambda) &= a_1 - \lambda \\ P_2(\lambda) &= (a_1 - \lambda)P_1(\lambda) - b_1^2 \\ P_3(\lambda) &= (a_3 - \lambda)P_2(\lambda) - b_2^2 P_1(\lambda) \end{aligned} \tag{4.124}$$

πράγμα που δηλώνει μία γενική αναδρομική σχέση η οποία δίνεται από το παρακάτω θεώρημα.

Θεώρημα 11.1

Αν T είναι ένας πραγματικός συμμετρικός τριδιαγώνιος πίνακας και αν $T - \lambda I$ δίνεται από την

$$\begin{bmatrix} a_1 - \lambda & b_1 & & \\ b_1 & a_2 - \lambda & b_2 & \mathbf{0} \\ \ddots & \ddots & \ddots & \ddots \\ \mathbf{0} & b_{n-2} & a_{n-1} - \lambda & b_{n-1} \\ & & b_{n-1} & a_n - \lambda \end{bmatrix} \tag{4.125}$$

τότε για $i = 1(1)n$

$$P_i(\lambda) = (a_i - \lambda)P_{i-1}(\lambda) - b_{i-1}^2 P_{i-2}(\lambda) \tag{4.126}$$

όπου $P_{-1}(\lambda) \equiv 0, \quad P_0(\lambda) \equiv 1$ και $b_0 = 0$.

Απόδειξη

Αν

$$P_i(\lambda) = det \begin{bmatrix} a_1 - \lambda & b_1 & & \\ b_1 & a_2 - \lambda & b_2 & \mathbf{0} \\ \ddots & \ddots & \ddots & \ddots \\ \mathbf{0} & b_{i-2} & a_{i-1} - \lambda & b_{i-1} \\ & & b_{i-1} & a_i - \lambda \end{bmatrix} \tag{4.127}$$

τότε αναπτύσσοντας την ορίζουσα ως προς τα στοιχεία της τελευταίας σειράς έχουμε για $i \geq 3$ την (4.126).■

Είναι φανερό τώρα ότι οι ρίζες κάθε πολυωνυμικής εξίσωσης $P_i(\lambda) = 0$ είναι πραγματικές. Επειδή $P_n(\lambda) = det(T - \lambda I)$είναι το χαρακτηριστικό πολυώνυμο του T, ο σκοπός μας είναι να βρούμε τις ρίζες της $P_n(\lambda) = 0$. Κατ' αρχήν χρειαζόμαστε το παρακάτω αποτέλεσμα.

Θεώρημα 11.2 (Givens)

Αν T είναι ο πραγματικός συμμετρικός τριδιαγώνιος πίνακας που ορίζεται από την (4.123) με $b_i \neq 0,\ i = 1(1)n-1$, τότε οι ρίζες κάθε εξίσωσης $P_i(\lambda) = 0$ είναι διακεκριμένες και χωρίζονται από τις ρίζες της $P_{i-1}(\lambda) = 0$.

Απόδειξη

Επειδή $b_k \neq 0$ για $k = 1(1)n-1$, καμία από τις δύο διαδοχικές εξισώσεις $P_{i-1}(\lambda) = 0$ και $P_i(\lambda) = 0$ δεν είναι δυνατόν να έχουν μία κοινή ρίζα, γιατί αν είχαν, τότε, λόγω της (4.126) και οι $P_{i-2}(\lambda) = 0,\ P_{i-3}(\lambda) = 0,\ \ldots,\ P_i(\lambda) = 0$ θα είχαν την ίδια ρίζα. Αλλά το a_1 είναι η μοναδική ρίζα της $P_1(\lambda) = 0$ και το a_1 δεν είναι ρίζα της$P_2(\lambda) = 0$, αφού $b_1 = 0$ λόγω της υπόθεσης. Ο διαχωρισμός των ριζών αποδεικνύεται με την μέθοδο της επαγωγής. Εύκολα μπορεί να διαπιστωθεί ότι το a_1, η ρίζα της $P_1(\lambda) = 0$, βρίσκεται μεταξύ των διακεκριμένων ριζών της $P_2(\lambda) = 0$. Υποθέτουμε ότι οι ρίζες των $P_{i-2}(\lambda) = 0$ και $P_{i-1}(\lambda) = 0$ είναι διακεκριμένες και οι ρίζες της πρώτης χωρίζουν τις ρίζες της δεύτερης. Έστω $r_1 < r_2 < \cdots < r_{i-1}$ οι ρίζες της $P_{i-1}(\lambda) = 0$. Τότε από την αναδρομική σχέση (4.126) για $k = 1(1)i-1$ έχουμε

$$P_i(r_k) = -b_{i-1}^2 P_{i-2}(r_k)$$

το οποίο σημαίνει ότι οι ποσότητες $P_i(r_k)$ και $P_{i-2}(r_k)$ έχουν αντίθετα πρόσημα. Ωστόσο, λόγω της υπόθεσης της επαγωγής, το $P_{i-2}(\lambda)$ αλλάζει πρόσημο μεταξύ των r_k και $r_{k+1},\ k = 1(1)i-2$ πράγμα που σημαίνει ότι και το$P_i(\lambda)$ αλλάζει πρόσημο. Με άλλα λόγια η $P_i(\lambda) = 0$ έχει μια ρίζα μεταξύ κάθε ζευγαριού διαδοχικών ριζών της $P_{i-1}(\lambda) = 0$. Επειδή δε

$$P_i(\lambda) \to \begin{cases} \infty, & \lambda \to -\infty \\ (-1)^i \infty, & \lambda \to \infty \end{cases}$$

για $i = 1(1)n$ έπεται ότι η $P_i(\lambda) = 0$ έχει μια ρίζα δεξιά της r_{i-1} και μια ρίζα αριστερά της r_1. Αν οι ρίζες της $P_i(\lambda)$ είναι οι $s_1,\ s_2,\ \ldots,\ s_i$ δείξαμε ότι

$$s_1 < r_1 < s_2 < r_2, \cdots < s_{i-1} < r_{i-1} < s_i$$

Συνεπώς οι ρίζες της $P_i(\lambda) = 0$ είναι διακεκριμένες και χωρίζονται από τις ρίζες της $P_{i-1}(\lambda) = 0$.■

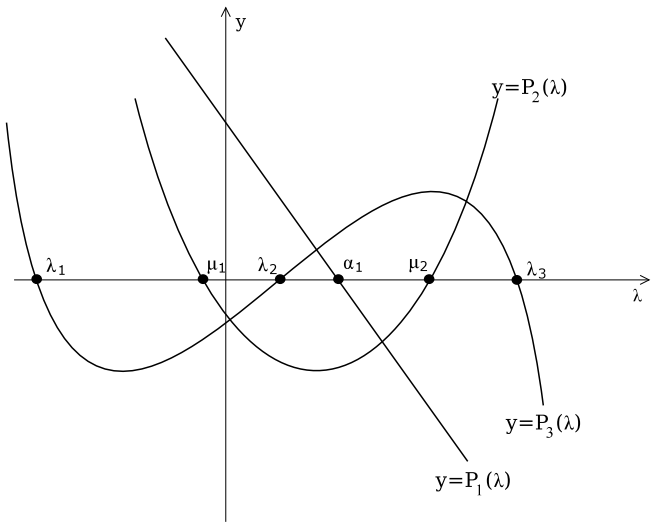
Η ιδιότητα αυτή αποτελεί την βάση μιας αποτελεσματικής τεχνικής για τον προσδιορισμό των θέσεων των ιδιοτήτων του T. Πιο συγκεκριμένα η ακολουθία των πολυωνύμων $\{P_i(\lambda)\},\ i = 1(1)n$ που ορίζεται από την (4.126) αποτελεί μια ακολουθία Sturm στο διάστημα $(-\infty,\ \infty)$ σύμφωνα με τον ακόλουθο ορισμό

Ορισμός (Gantmacher [1960], vol II, p. 175). Ας θεωρήσουμε την ακολουθία των πραγματικών πολυωνύμων $p_0(x), p_1(x), p_2(x), \ldots, p_n(x)$ με τις παρακάτω δύο ιδιότητες, όσον αφορά ένα ανοικτό διάστημα (a, b), όπου a μπορεί να είναι το $-\infty$ και b το ∞:
(i) Για κάθε τιμή $x_0 \in (a, b),\ $ αν $p_k(x_0) = 0$, τότε

$$P_{k-1}(x_0)P_{k+1}(x_0) < 0$$

το οποίο σημαίνει ότι τα $P_{k-1}(x_0)$ και $P_{k+1}(x_0)$ έχουν αντίθετα πρόσημα.
(ii) $p_0(x) \neq 0$ για κάθε $x \in (a, b)$.

Τότε η ακολουθία αυτή των πολυωνύμων καλείται μια ακολουθία Sturm στο διάστημα (a, b). Στο σχήμα 4.11 παρουσιάζεται ο διαχωρισμός των ριζών για $n = 3$, όπου a_1 είναι ρίζα της $P_1(\lambda) = 0,\ \mu_1$ και μ_2 οι ρίζες της $P_2(\lambda) = 0$ και λ_1, λ_2 και λ_3 οι ρίζες της $P_3(\lambda) = 0$. Δεν έχει σημασία αν $\lambda_2 < a_1$ ή $a_1 < \lambda_2$ στο διάγραμμα αρκεί $a_1, \lambda_2 \in (\mu_1, \mu_2)$



Σχήμα 4.11

Ορισμός

Έστω $s(\lambda)$ ο αριθμός των αμετάβλητων προσήμων στην ακολουθία $1, P_1 : (\lambda), : P_2(\lambda), : \ldots, : P_n(\lambda)$. Αν $P_i(\lambda) = 0$ για κάποιο i λαμβάνουμε σαν πρόσημο το πρόσημο του $P_{i-1}(\lambda)$. Για παράδειγμα, για το σχήμα 4.11, παρατηρούμε ότι έχουμε τον παρακάτω πίνακα όταν το λ βρίσκεται στα αντίστοιχα διαστήματα:

Πίνακας 4.1

	$(-\infty, \lambda_1)$	(λ_1, μ_1)	(μ_1, λ_2)	(λ_2, a_1)	(a_1, μ_2)	(μ_2, λ_3)	(λ_3, ∞)
1	+	+	+	+	+	+	+
$P_1(\lambda)$	+	+	+	+	-	-	-
$P_2(\lambda)$	+	+	-	-	-	+	+
$P_3(\lambda)$	+	-	-	+	+	+	-
s_λ	3	2	2	1	1	1	0

Από το σχήμα 4.11 παρατηρούμε ότι ο αριθμός $s(\lambda)$ είναι ίσος με τον αριθμό εκείνων των ριζών της $P_3(\lambda) = 0$, οι οποίες είναι μεγαλύτερες ή ίσες με λ. Γενικά έχουμε:

Θεώρημα 11.3

Η ποσότης $s(\lambda)$ είναι ο αριθμός των ριζών της $P_n(\lambda) = 0$ οι οποίες είναι μεγαλύτερες ή ίσες με λ.

Απόδειξη

Αφήνεται σαν άσκηση για τον αναγνώστη.

Είναι φανερό λοιπόν ότι μία συστηματική αναζήτηση, χρησιμοποιώντας το θεώρημα 11.3 για διάφορες τιμές του λ θα προσδιορίσει προσεγγιστικά διαστήματα τα οποία περικλείουν μία ιδιοτιμή. Επειδή $S(T) \leq \| T \|_\beta$ για κάθε norm πίνακα έχουμε ότι όλες οι ιδιοτιμές του T (ρίζες της $P_n(\lambda)$) βρίσκονται στο κλειστό διάστημα $[- \| T \|_\infty, \| T \|_\infty]$. Χρησιμοποιώντας τώρα τη μέθοδο της διχοτόμησης και το θεώρημα 11.3 μπορούμε να βρούμε ένα διάστημα το οποίο να περιέχει μόνον την ιδιοτιμή που αναζητούμε. Στο σημείο αυτό η μέθοδος της διχοτόμησης μπορεί να συνεχιστεί μέχρις ότου επιτευχθεί η επιθυμητή ακρίβεια για την ιδιοτιμή που έχουμε απομονώσει. Φυσικά μετά από m διχοτομήσεις το μέγεθος του διαστήματος θα είναι $2^{-m}[- \| T \|_\infty, \| T \|_\infty]$.

Παράδειγμα

Δίνεται ο πίνακας

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

Πόσες από τις ιδιοτομές του βρίσκονται στο διάστημα $[-2, 0]$;

Λύση

Αν $\lambda = 0$ τότε από το Θεώρημα 11.3 βρίσκουμε:

$$P_0(0) = 1, \quad P_1(0) = -2, \quad P_2(0) = 3, \quad P_3(0) = -4, \quad \text{και} \quad P_4(0) = 5$$

Η ακολουθία των σημείων της $\{P_i(0)\}, \quad i = 0(1)4$ είναι

$$+ - + - +$$

οπότε $s(0) = 0$ πράγμα που δηλώνει ότι ο T δεν έχει καμία θετική ή μηδέν ιδιοτιμή, δηλαδή ο T είναι ένας αρνητικά ορισμένος πίνακας. Αν $\lambda = -2$ τότε έχουμε την ακολουθία:

$$P_0(-2) = 1, \quad P_1(-2) = 0, \quad P_2(-2) = -1, \quad P_3(-2) = 0, \quad P_4(-2) = 1$$

και η ακολουθία των προσήμων είναι η $++--+$, άρα $s(-\lambda) = 2$ πράγμα που σημαίνει ότι ο T έχει δύο ιδιοτιμές του στο διάστημα $[-2, 0]$. Στη συνέχεια μπορούμε να διχοτομήσουμε το διάστημα $[-2, 0]$ και να επαναλάβουμε την παραπάνω διαδικασία μέχρις ότου βρούμε κάποιο διάστημα που περιέχει μόνο την ιδιοτιμή που επιθυμούμε.

Η παραπάνω τεχνική είναι πολύ αποτελεσματική αν χρειάζεται να βρούμε τις ιδιοτιμές σε ένα συγκεκριμένο διάστημα ή μερικές από τις πρώτες ή μερικές από τις τελευταίες ιδιοτιμές ενός nxn συμμετρικού τριδιαγώνιου πίνακα.

Υποθέτουμε τώρα ότι έχουμε υπολογίσει μία ή περισσότερες ιδιοτιμές ενός συμμετρικού πίνακα $A \in \mathbb{R}^{nn}$ αφού πρώτα τον μετασχηματίσαμε σε ένα συμμετρικό τριδιαγώνιο πίνακα T. Έστω

$$A = P^T T P$$

με $P^T P = I$. Αν λ είναι μία ιδιοτιμή του T με αντίστοιχο ιδιοδιάνυσμα y, τότε $Ty = \lambda y$ και έτσι

$$\begin{aligned} \lambda(P^T y) &= P^T T y \\ &= P^T T P (P^T y) \\ &= A(P^T y) \end{aligned}$$

δηλαδή το αντίστοιχο ιδιοδιάνυσμα στη λ ιδιοτιμή του A είναι το

$$x = P^T y \tag{4.128}$$

Με άλλα λόγια μπορούμε να υπολογίσουμε το ιδιοδιάνυσμα x του A υπολογίζοντας πρώτα το ιδιοδιάνυσμα y του T και μετά να χρησιμοποιήσουμε την (4.128). Για λόγους υπολογιστικής ευστάθειας, ο

Wilkinson [1963, σελ. 142] συνιστά τη χρήση της αντίστροφης μεθόδου των δυνέμεων για τον υπολογισμό του y (βλ. επίσης Ortega [1967], σελ. 98-99).

Παρατήρηση

Αν ο τριδιαγώνιος πίνακας T παράγεται από μία ακολουθία μετασχηματισμών του Householder, τότε από την (4.128) έχουμε

$$\begin{aligned} x &= (P_2 P_3 \dots P_{n-1})^y \\ &= [I - a_2 v^{(2)} v^{(2)^T}] \dots [I - a_{n-1} v^{(n-1)} v^{(n-1)^T}] y \end{aligned}$$

και οι πολ/μοί αυτοί μπορούν να γίνουν πολύ απλά αν παρατηρήσουμε ότι

$$(I - a v v^T) y = y - a(v^T y) v$$

Η απλότητα των υπολογισμών ενισχύεται από το γεγονός ότι, για το $v^{(r)}$, τα πρώτα $r-1$ στοιχεία είναι μηδέν.

Παρατήρηση

Προφανώς η μέθοδος του Householder βρίσκει μία ιδιοτιμή κάθε φορά και αν, όπως συχνά είναι η περίπτωση, επιθυμούμε μόνο μία ιδιοτιμή, τότε ο αλγόριθμος αυτός είναι κατάλληλος. Ακόμα και αν επιθυμούμε όλες τις ιδιοτιμές απαιτεί λιγότερο υπολογιστικό χρόνο από τις παραλλαγές της μεθόδου Jacobi και για τον λόγο αυτό χρησιμοποιείται πάρα πολύ συχνά για πραγματικούς, συμμετρικούς πίνακες. Η μέθοδος του Jacobi από την άλλη πλευρά είναι κατάλληλη για την εύρεση όλων των ιδιοτιμών ταυτόχρονα μαζί με το πλήρες ορθοκανονικό σύνολο των ιδιοδιανυσμάτων (αν η τάξη του πίνακα n δεν είναι τόσο μεγάλη ώστε ο υπολογιστικός χρόνος να είναι λίγος).

Βιβλιογραφία

[1] AHO, A.V., HOPCROFT J.E. and ULLMAN J.D., The Design and Analysis of Computer Algorithms, Addison-Wesley, Reading, Mass., 470 pp. 1974.

[2] ΑΠΟΣΤΟΛΑΤΟΣ Ν.Θ., Αριθμητική Ανάλυση, Τεύχος 1 και 2, 1983.

[3] ATKINSON K.E., An Introduction to Numerical Analysis (second edition), John Wiley & Sons, New York, 1989.

[4] BURDEN, R.L. and FAIRES J.D., Numerical Analysis, Brooks/Cole, Publishing Company, Pacific Grove, Cal., 2001.

[5] COLEMAN, T.F. and C. VAN LOAN, Handbook for Matrix Computations, SIAM Publications, Philadelphia, 1988.

[6] CONTE S.D. and C. de BOOR, Elementary Numerical Analysis, 3rd ed, McGraw-Hill, 1980.

[7] DAHLQUIST G. and BJORCK A., Numerical Methods, Prentice-Hall, Inc., 1974.

[8] FORSYTHE G.E., MALCOLM M.A. and MOLER C.A. Computer Methods for Mathematical Computations, Prentice-Hall, Engle- wood Cliffs, N.J. 1977.

[9] FORSYTHE G.E. and MOLER C.B. Computer solution of linear algebraic systems, Prentice Hall, Englewood Cliffs, NJ, 1967.

[10] FROBERG C.E., Introduction to Numerical Analysis, Second edition, Addison-Wesley, 1969.

[11] GEORGE J.A. and LIU J.W.H., Computer Solution of Large Sparse Positive Definite Systems, Prentice Hall, Englewood Cliffs, N.J. 1981.

[12] GERALD C.F., Applied Numerical analysis, Addison-Wesley, 1970. GOLUB, G.H. and VAN LOAN C.F. Matrix Computations (second edition). Johns Hopkins University Press, Baltimore, 1989.

[13] HAGEMAN L.A. and YOUNG D.M. Applied Iterative Methods, Academic Press, New York, 1981.

[14] HENRICI P. Elements of Numerical Analysis, John Wiley & Sons, New York, 1964.

[15] HILDEBRAND F.B., Introduction to Numerical Analysis, 2d ed., McGraw-Hill, 1974.

[16] HOUSEHOLDER A.S., The theory of Matrices in Numerical Analysis, Blaisdell PubL Company, 1964.

[17] ISSACSON E. and KELLER H.B., Analysis of Numerical Methods, John Wiley & Sons, New York, 1966.

[18] JOHNSON L.W. and RIESS R.D., Numerical Analysis, Addison-Wesley, 1977.

[19] KAHANER D., MOLER C. and NASH S. Numerical Methods and Software. Prentice-Hall, Englewood Cliffs, NJ. 1989.

[20] KINCAlD D. and CHENEY W. Numerical Analysis: Mathematics of Scientific Computing. Brooks/Cole Publishing Company, Pacific Grove, Calif., 1991.

[21] MITCHELL A.R. Computational Methods for Partial-Differential Equations, John Wiley & Sons, London, 1969.

[22] MISSIRLIS N. M., Preconditioned Iterative methods for solving Elliptic Partial Differential Equations, Ph.D, Loughborough Univ., UK, 1978.

[23] ORTEGA J.M., Numerical Analysis: A second course, Academic Press, 1972.

[24] PARLETT B. The Symmetric Eigenvalue Problem, Prentice Hall, Englewood Cliffs, NJ. 1980.

[25] RALSTON A. and RABINOWITZ P. A First Course in Numerical Analysis (second edition), McGraw-Hill, New York, 1978.

[26] STEWART G.W. Introduction to Matrix Computations, Academic Press, New York, 1973.

[27] STOER J. and BULIRSCH R., Introduction to Numerical Analysis, Springer-Verlag, 1980.

[28] STRANG G., Linear Algebra and its Applications, Academic Press, (second edition), 1980.

[29] VARGA R.S. Matrix Iterative Analysis, Prentice Hall, Englewood Cliffs, N.J. 1962.

[30] WILKINSON J.H. Rounding Errors in Algebraic Processes, H.M. Stationery Office, London, 1963.

[31] WILKINSON J.H. The Algebraic Eigenvalue Problem, Clarendon Press, Oxford, 1965.

[32] ΧΑΤΖΗΔΗΜΟΣ Α., Εισαγωγή στην Αριθμητική Ανάλυση, Ιωάννινα, 1977.

[33] ΧΑΤΖΗΔΗΜΟΣ Α., Αριθμητική Ανάλυση I, II Ιωάννινα, 1978,1979.

[34] YOUNG D.M., Iterative solution of Large Linear Systems, Academic Press, New York, 1971.

[35] YOUNG D.M. and GREGORY R.T., A Survey of Numerical Mathematics, Volumes I-II, Addison-Wesley, 1972, 1973.