# Αριθμητική Γραμμική Αλγεβρα

Μ. Μητρουλη

# Περιεχόμενα


**1 Βασική Αριθμητική Υπολογιστή** **1**

1.1 Ψηφιακοί Υπολογισμοί . . . . . 1

1.1.1 Υπολογισμοί Σταθερής Υποδιαστολής . . . . . 2

1.1.2 Υπολογισμοί Κινητής Υποδιαστολής . . . . . 5

1.2 Αριθμοί Μηχανής . . . . . 7

1.2.1 Αποθήκευση των floating point αριθμών . . . . . 7

1.2.2 Παράσταση των αριθμών του $\mathcal{M}$ . . . . . 12

1.2.3 Εκτέλεση Πράξεων σε Αριθμητική Κινητής Υποδιαστολής . . . 19

1.3 Αριθμητικά Αποτελεσματικοί Αλγόριθμοι . . . . . 20

1.3.1 Ανάλυση αποτελεσματικότητας και μνήμης για ορισμένους βασικούς αλγόριθμους . . . . . 22

1.4 ΑΣΚΗΣΕΙΣ . . . . . 25

**2 Θεωρία Ανάλυσης Σφάλματος** **33**

2.1 Μορφές της Ανάλυσης Σφάλματος . . . . . 33

2.2 Σφάλματα Στρογγύλευσης σε Υπολογισμούς . . . . . 35

2.2.1 Υπολογισμός Γινομένου . . . . . 36

2.2.2 Υπολογισμός αθροίσματος . . . . . 38

2.2.3 Υπολογισμός εσωτερικού γινομένου . . . . . 39

2.3 Σφάλματα Στρογγύλευσης σε Υπολογισμούς με Πίνακες . . . . . 50

2.3.1 Ανάλυση σφάλματος απλών πράξεων με πίνακες . . . . . 50

2.3.2 Ορθογώνιοι Πίνακες . . . . . 54

2.4 Ευστάθεια και Κατάσταση Προβλημάτων . . . . . 60

2.4.1 Εφαρμογή Ευσταθούς Αλγορίθμου σε well-conditioned Πρόβλημα . . . . . 61

2.4.2 Εφαρμογή Ευσταθούς Αλγορίθμου σε ill-conditioned Πρόβλημα 62

**3 Μετασχηματισμοί Gauss** **65**

3.1 Παραγοντοποίηση LU . . . . . 65

3.2 Μέθοδος απαλοιφής του Gauss χωρίς οδήγηση . . . . . 66

3.2.1 Τριγωνοποίηση πίνακα με απαλοιφή Gauss χωρίς οδήγηση . . . 72

3.2.2 Διανυσματική μορφή Αλγορίθμου *LU* . . . . . 72

3.3 Μετασχηματισμοί Gauss-Jordan . . . . . 72

3.3.1 Μέθοδος απαλοιφής του Gauss-Jordan χωρίς οδήγηση . . . . . 74
3.3.2 Προσδιορισμός του πίνακα $A^{-1}$ . . . . . 76
3.4 Η Τεχνικής της Οδήγησης . . . . . 78
3.4.1 Μεταθετικοί Πίνακες . . . . . 80
3.5 Μέθοδος Gauss με Οδήγηση . . . . . 81
3.6 Μέθοδος Gauss με Μερική Οδήγηση . . . . . 82
3.7 Μέθοδος Gauss με Ολική Οδήγηση . . . . . 90
3.8 Τριγωνική Διαχώριση . . . . . 95
3.8.1 Σχέση μεταξύ Απαλοιφής Gauss και Τριγωνικής Διαχώρισης . 96

**4 Ανάλυση Σφάλματος Αριθμητικών Μεθόδων 99**
4.1 Μέθοδος απαλοιφής του Gauss . . . . . 99
4.2 Μελέτη του Growth Factor . . . . . 104
4.3 Μέθοδος της τριγωνικής διαχώρισης ($LU$) . . . . . 108
4.3.1 Υπολογισμός ορίζουσας . . . . . 110
4.4 Περιορισμός Μεγέθους των Στοιχείων ενός Πίνακα (Scaling) . . . . . 111
4.4.1 Αλγόριθμος B-SCALE . . . . . 112
4.5 Ειδικές Μορφές Πινάκων . . . . . 113
4.5.1 $LU$ παραγοντοποίηση σε ορθογώνιο πίνακα . . . . . 113
4.5.2 Θετικά Ορισμένα Συστήματα-Ανάλυση Cholesky . . . . . 116
4.5.3 Εφαρμογή στην Επίλυση Γραμμικού Συστήματος . . . . . 118
4.6 Προσέγγιση της Αποτελεσματικότητας Αλγορίθμων . . . . . 120

**5 Μετασχηματισμοί Householder 123**
5.1 Παραγοντοποίση $QR$ . . . . . 123
5.2 Μετασχηματισμοί Householder για ορθογώνιο πίνακα . . . . . 132

**6 Αριθμητική Επίλυση Γραμμικών Συστημάτων 135**
6.1 'μεσοι Μέθοδοι . . . . . 136
6.1.1 Επίλυση ενός άνω τριγωνικού συστήματος . . . . . 136
6.1.2 Επίλυση ενός κάτω τριγωνικού συστήματος . . . . . 136
6.1.3 Επίλυση Γραμμικού Συστήματος με πολλαπλό Δεξιό μέλος . . 147
6.2 Ανάλυση Ευαισθησίας Γραμμικού Συστήματος . . . . . 148

**7 Ελάχιστα Τετράγωνα 155**
7.1 Εισαγωγή . . . . . 155
7.2 Πρόβλημα των γραμμικών ελαχίστων τετραγώνων . . . . . 156
7.3 Υπολογιστικές μέθοδοι για την επίλυση overdetermined προβλημάτων ελαχίστων τετραγώνων . . . . . 158
7.3.1 Η μέθοδος των κανονικών εξισώσεων . . . . . 158

**8 Πίνακες Hessenberg (Σχεδόν τριγωνικοί) 161**
8.1 Αλγόριθμος Αναγωγή Householder - Hessenberg . . . . . 166
8.2 Τριδιαγώνια Αναγωγή . . . . . 167

**9 Αριθμητικός Υπολογισμός Ιδιοτιμών** **169**
9.0.1 Η βασική $QR$ επανάληψη . . . . . 171

**10 Ανάλυση Ιδιαζουσών Τιμών** **175**
10.1 Το Θεώρημα Ανάλυσης Ιδιαζουσών Τιμών . . . . . 175
10.2 Η SVD και η δομή ενός πίνακα . . . . . 177
10.2.1 Προσδιορισμός ορθοκανονικών βάσεων . . . . . 178
10.2.2 Numerical rank πίνακα . . . . . 180
10.3 Υπολογισμός λύσης ελαχίστων τετραγώνων . . . . . 180
10.3.1 Αλγόριθμος SVD- ελάχιστα τετράγωνα . . . . . 181

# Κεφάλαιο 1

# Βασική Αριθμητική Υπολογιστή

## 1.1 Ψηφιακοί Υπολογισμοί

Τα δεδομένα σ' έναν υπολογιστή χωρίζονται σε λέξεις (words) αποτελούμενες από 0 και 1. Κάθε 0 ή 1 ονομάζεται bit. Οι αριθμοί αποθηκεύονται σαν μία ακολουθία από bytes όπου 1 byte =8 bits. Ο αριθμός των bits σε μία λέξη ποικίλλει από υπολογιστή σε υπολογιστή αλλά συνήθως τείνει να είναι 32 bits (4 bytes) ή 64 bits (8 bytes). Ο αριθμός αυτός των bits που προσφέρεται για την καταχώρηση των δεδομένων αποτελεί τα διαθέσιμα από τον υπολογιστή ψηφία για την επεξεργασία τους. Επομένως θα περιγράψουμε υπολογισμούς που εκτελούνται από ψηφιακούς υπολογιστές (digital computers) που χαρακτηρίζονται από τον αριθμό των ψηφίων $t$ που διαθέτουν στο χρήστη προς επεξεργασία των δεδομένων του. Κυρίως οι υπολογιστές δουλεύουν με αριθμούς εκφρασμένους σε σύστημα αρίθμησης με βάση $\beta$. Κάθε αριθμός z με n ακέραια και m δεκαδικά ψηφία παριστάνεται σε σύστημα αρίθμησης με βάση $\beta$ στη μορφή

$$z_{(\beta)} = \sigma \cdot d_{n-1}d_{n-2}\ldots d_1 d_0 d_{-1}d_{-2}\ldots d_{-m}$$

όπου $\sigma$ το πρόσημο του. Κάθε ψηφίο $d_i$ παίρνει τιμές από 0 έως $\beta - 1$. Η τιμή του αριθμού αυτού στο δεκαδικό σύστημα δίνεται από την πολυωνυμική έκφραση

$$z_{(\beta)} = \sigma \cdot \sum_{i=-m}^{n-1} d_i \beta^i$$

Συνήθως $\beta = 2$ και σπανιότερα $\beta = 8$ και $\beta = 16$. Σχεδόν ποτέ δε χρησιμοποιείται το δεκαδικό σύστημα. Οι ακέραιοι αποθηκεύονται με διαφορετικό τρόπο από ότι οι πραγματικοί αριθμοί. Το bit pattern που αντιστοιχεί στον πραγματικό αριθμό 1195.0 θα είναι διαφορετικό από το bit pattern που αντιστοιχεί στον ακέραιο 1195 παρόλο που έχουν το ίδιο μέγεθος.

Οι υπολογιστές εκτελούν δύο τύπους αριθμητικής: Ακέραια Αριθμητική και Πραγματική Αριθμητική. Κατά συνέπεια, χρησιμοποιούνται αντίστοιχα και οι ακόλουθοι δύο τρόποι (modes) εκτέλεσης πράξεων:

- Υπολογισμοί Σταθερής Υποδιαστολής (**fixed-point** computations)
- Υπολογισμοί Κινητής Υποδιαστολής (**floating-point** computations)

### 1.1.1 Υπολογισμοί Σταθερής Υποδιαστολής

Χρησιμοποιούνται για τη διαχείριση πράξεων με ακεραίους. Σε κάθε αριθμό $x$ διατίθεται ένας καθορισμένος αριθμός από $t$ ψηφία για την αναπαράστασή του (με βάση $\beta$) και κατά συνέπεια ο υπολογιστής δουλεύει με **λέξεις** των $t$ **ψηφίων**. Κάθε αριθμός $x$ θα πρέπει να ικανοποιεί ανισότητες της μορφής

$$-1 \leq x \leq 1$$

(Πολλές φορές τα όρια εξαιρούνται)

Για μεγαλύτερη ακρίβεια δουλεύουμε με αριθμούς που παριστάνονται με ένα πολλαπλάσιο των $t$ ψηφίων και αναφερόμαστε σε **multiple-precision** υπολογισμούς, για να διαφέρουν από τους καθιερωμένους **single-precision** υπολογισμούς. Οι multiple precision συνήθως απαιτούν την εκτέλεση ειδικών ρουτινών και κάθε αριθμητική πράξη απαιτεί περισσότερο χρόνο από την αντίστοιχη σε single-presicion. Αυτού του τύπου οι υπολογισμοί χρησιμοποιούνται κυρίως σε ακεραίους.

**Παράδειγμα 1.1.1** *Να προστεθούν οι αριθμοί* $a = 0.1131$, $b = 0.8432$ *χρησιμοποιώντας fixed point αριθμητική με* $t = 4$ *(ψηφία μετά την υποδιαστολή* $\beta = 10$*). Στη συνέχεια να πολλαπλασιασθούν.*

$$f_i(a + b) = c \equiv a + b = 0.9563$$

Δεν εμφανίζεται σφάλμα στρογγύλευσης

$$f_i(a * b) = c \equiv a * b = 0.09536592$$

Επειδή διατίθενται μόνο 4 θέσεις προσθέτουμε το $\frac{1}{2} * 10^{-4}$ ή το $10^{-4}$ οπότε έχουμε

$$0.09536592 + 0.00005 = 0.09541592$$

και λαμβάνουμε σα fixed point προσέγγιση τον αριθμό 0.0954.
Τελικά

$$f_i(a \cdot b) = 0.0954 = c \equiv a * b + e$$

όπου

$$e = 0.00003408, \;\; |e| \leq \frac{1}{2} * 10^{-4}$$

□

**Παράδειγμα 1.1.2** *Να διαιρεθούν οι αριθμοί a=0.0635, b=0.6673 χρησιμοποιώντας fixed point αριθμητική με* $t = 4$ *και* $\beta = 10$.

$$f_i(a/b) = c \equiv a/b = 0.095159\ldots$$

Το ακριβές πηλίκο είναι απεριόριστο Επειδή διατίθενται μόνο 4 θέσεις κρατούμε τα 6 πρώτα ψηφία μετά την υποδιαστολή και προσθέτουμε την ποσότητα $\frac{1}{2} * 10^{-4}$ οπότε έχουμε

$$0.095159 + 0.00005 = 0.09520$$

Ο αριθμός 0.0952 είναι το τελικό αποτέλεσμα.
Τελικά

$$f_i(a/b) = c \equiv a/b + \varepsilon$$

όπου

$$\varepsilon = 0.000041\ldots, \ \ |\varepsilon| \leq \frac{1}{2} * 10^{-4}$$

□

**Παρατήρηση:** Συνήθως στην πρόσθεση και στην αφαίρεση δεν παρατηρείται σφάλμα στρογγύλευσης εκτός εάν το αποτέλεσμα προκύψει εκτός των επιτρεπτών ορίων. Αντίθετα στον πολλαπλασιασμό και στη διαίρεση η εμφάνιση σφάλματος είναι συνηθέστερη.

## Αποθήκευση fixed point αριθμού

Η αποθήκευση των προσημασμένων fixed point αριθμών γίνεται με τη μέθοδο της απεικόνισης συμπληρώματος ως προς 2. Η γενική περίπτωση μιας τέτοιας απεικόνισης σε λέξη $k$ ψηφίων χαρακτηρίζεται από τα εξής:

- Απεικονίζονται το πολύ $2^k$ ακέραιοι αριθμοί (integers)
- Στο πρώτο ψηφίο, καταχωρείται το πρόσημο (0 αν είναι θετικός και 1 αν είναι αρνητικός)
- Υπάρχει μία μόνο απεικόνιση του μηδενός
- Υπάρχουν $2^{k-1} - 1$ δυνατοί συνδιασμοί 0 και 1 που έχουν το MSB μηδέν και απεικονίζουν τους θετικούς ακέραιους αριθμούς ($> 0$) από το 1 μέχρι το $2^{k-1} - 1$.
- Υπάρχουν $2^{k-1}$ δυνατοί συνδιασμοί 0 και 1 που έχουν το MSB ένα και απεικονίζουν τους αρνητικούς ακέραιους αριθμούς από το -1 μέχρι το $-2^{k-1}$.
- Ένας παραπάνω αρνητικός αριθμός ο $-2^{k-1}$, πού δεν έχει θετικό συμπλήρωμα (τον $2^{k-1}$).

Για παράδειγμα αν $k = 4$ τότε ο μέγιστος θετικός ακέραιος είναι ο

$$\boxed{0}\boxed{1}\boxed{1}\boxed{1} = 0 * 2^3 + 1 * 2^2 + 1 * 2 + 1 * 2^0 = 7 = 2^3 - 1$$

ενώ ο μικρότερος ακέραιος είναι ο

$$\boxed{1}\boxed{0}\boxed{0}\boxed{0} = -2^3 + 0 * 2^2 + 0 * 2^1 + 0 * 2^0 = -8 = -2^3$$

## Παράσταση Λέξης Υπολογιστή

Ας υποθέσουμε τώρα ότι διαθέτουμε έναν υπολογιστή με $\beta = 2$ και μήκος λέξης 32 bits. Αυτά έχουν την ακόλουθη κατανομή

- πρόσημο αριθμού 1 bit (0 θετικό, 1 αρνητικό)
- ψηφία αριθμού 31 bits

| 0 | 1 | 2 | 3 | ... | 28 | 29 | 30 | 31 |
|---|---|---|---|---|---|---|---|---|
| $\sigma$ | | | | ... | | | | |

31 integer binary representation

Κατά συνέπεια ο υπολογιστής αυτός έχει το ακόλουθο εύρος παράστασης αριθμών:

$$x = \sigma \cdot (a_1 a_2 \dots a_{31})_2 = \sigma \sum_{i=1}^{31} a_i 2^i$$

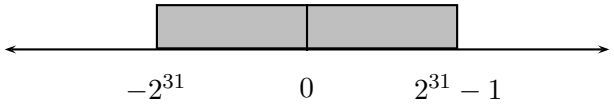
Άρα απεικονίζονται το πολύ $2^{32}$ ακέραιοι αριθμοί, ενώ υπάρχει μια μόνο απεικόνιση του 0.

Ο μέγιστος παραστήσιμος θετικός αριθμός είναι:

$$\overbrace{(11\dots1)}^{31\psi\eta\phi\acute{\iota}\alpha}{}_2 = 2^{31} - 1 = 2.147483647$$

ενώ ο μικρότερος αρνητικός είναι:

$$\overbrace{(10\dots0)}^{31\psi\eta\phi\acute{\iota}\alpha}{}_2 = -2^{31} = -2.147483648$$




Σχήμα 1.1: Εύρος παράστασης ακεραίων

Από τα παραπάνω συμπεραίνουμε ότι με την αριθμητική σταθερής υποδιαστολής μπορούμε να παραστήσουμε ένα εύρος θετικών και αρνητικών ακεραίων κατανεμημένο

γύρω από το 0. Υποθέτοντας επίσης την ύπαρξη σταθερού δεκαδικού σημείου (fixed point), μας επιτρέπεται η παράσταση αριθμών με κλασματικό μέρος. Συγκεκριμένα, δεχόμαστε ότι στα πρώτα 1-23 bits της λέξης καταχωρείται το ακέραιο μέρος του αριθμού και στα υπόλοιπα 8 bits το κλασματικό, ενώ το σταθερό δεκαδικό σημείο (fixed point) τοποθετείται νοητά ανάμεσα στα bits 7 και 8. Τότε ο μέγιστος παραστήσιμος θετικός κλασματικός αριθμός είναι:

$$\overbrace{(11\ldots1)}^{23}\overbrace{(.11\ldots1)}^{8}{}_2 = 1677215.998046875$$

ενώ ο μικρότερος παραστήσιμος θετικός κλασματικός αριθμός είναι:

$$\overbrace{(00\ldots0)}^{23}\overbrace{(.00\ldots1)}^{8}{}_2 = 0.00390625$$

Αυτή η προσέγγιση όμως έχει περιορισμούς· πολύ μεγάλοι αριθμοί δεν μπορούν να παρασταθούν όπως επίσης και πολύ μικρά κλάσματα. Επιπλέον το κλασματικό μέρος του πηλίκου της διαίρεσης δύο πολύ μεγάλων αριθμών μπορεί να χαθεί.

### 1.1.2 Υπολογισμοί Κινητής Υποδιαστολής

Δεδομένου ότι οποιοσδήποτε πραγματικός αριθμός θα πρέπει να παρασταθεί σε πεπερασμένη ψηφιακή μορφή, για να αυξηθεί το εύρος αναπαράστασης χρησιμοποιείται η εκθετική αναπαράσταση. Εάν για παράδειγμα παραστήσουμε τους ακόλουθους ακεραίους σε εκθετική μορφή

$$976000000000000 \rightarrow 9.76 * 10^{14}$$

$$0.000000000000976 \rightarrow 9.76 * 10^{-14}$$

η διαφορά στα απαιτούμενα προς αποθήκευση ψηφία είναι τεράστια. Μετακινώντας έτσι δυναμικά το δεκαδικό σημείο σε κατάλληλη θέση και χρησιμοποιώντας τον εκθέτη του 10 ώστε να αποθηκεύεται η μετακίνηση αυτή, αυξάνουμε το εύρος παράστασης των αριθμών και πετυχαίνουμε τη δυνατότητα παράστασης πολύ μεγάλων και πολύ μικρών αριθμών με μόνο λίγα ψηφία. Εισάγουμε έτσι την έννοια των υπολογισμών κινητής υποδιαστολής, που μπορούν να ορισθούν ως εξής:

Ενας διάφορος του μηδενός δεκαδικός αριθμός παριστάνεται ως **κανονικοποιημένος floating point** αριθμός εάν εκφρασθεί στη μορφή

$$x = \sigma \cdot (.a_1 a_2 ... a_t)_\beta \cdot \beta^e$$

όπου $\sigma$ το πρόσημο του αριθμού, $e$ ο εκθέτης και $a_i$ τα ψηφία της χαρακτηριστικής (mantissa), έτσι ώστε $0 \leq a_i \leq \beta - 1, a_1 \neq 0$. Επίσης $m \leq e \leq M$, και γενικά $m = -M$ ή $m = -M + 1$.

**Ορισμός:** Το σύνολο $\mathcal{M} \subset \mathbb{R}$ όλων των υπό κανονικοποιημένη παράσταση αριθμών $x = \sigma \cdot \overline{x} \cdot \beta^e$,

$$\overline{x} = (0.a_1a_2...a_t)_\beta = \sum_{k=1}^{t} a_k \cdot \beta^{-k}$$

όπου $\sigma$ υποδηλώνει το πρόσημο, $t \in \mathbb{N}$ , $1 \leq a_1 \leq \beta - 1$, $0 \leq a_k \leq \beta - 1$, $k = 2, 3, \ldots, t$, $m \leq e \leq M$, $m \leq 0$, $M \geq 1$, και $e, m, M \in \mathbb{Z}$ καλείται **σύστημα κινητής υποδιαστολής** και συμβολίζεται με $\mathcal{M}(\beta, t, m, M)$ ή $\mathcal{M}$.
Κάθε $x \in \mathcal{M}$ ονομάζεται και **αριθμός μηχανής**.
Το $0 \in \mathcal{M}$(από τον ορισμό) και είναι ο αριθμός $0 := +0.00\ldots0\beta^m$

**Παρατηρήσεις:**

- $\mathcal{M} \subset \mathbb{R}$
- κάθε $x \in \mathcal{M}$ παριστάνει έναν πραγματικό αριθμό
- $1 \in \mathcal{M}$ και είναι ο αριθμός $1 := +1.00\ldots0\beta^0$
- $\forall x \in \mathcal{M} \Rightarrow -x \in \mathcal{M}$
- Η αριθμητική του υπολογιστή είναι διακριτή και όχι συνεχής.

**Λήμμα 1.1.1** *Το σύνολο των κανονικοποιημένων floating point αριθμών που ανήκουν στο $\mathcal{M}$ ισούται με $2(\beta - 1)\beta^{t-1}(M - m + 1) + 1$.*

**Απόδειξη:** Για κάθε τιμή του εκθέτη, η επιλογή του πρώτου ψηφίου της χαρακτηριστικής μπορεί να γίνει με $(\beta - 1)$ τρόπους, ενώ τα επόμενα ψηφία μπορούν να επιλεγούν με $\beta^{t-1}$ τρόπους. Δεδομένου ότι το σύνολο των τιμών του εκθέτη είναι $M - m + 1$ οι δυνατές χαρακτηριστικές είναι $(\beta - 1)\beta^{t-1}1)(M - m + 1)$. Αυτές θα πολλαπλασισθούν επί 2, δεδομένου ότι έχουμε και τους αρνητικούς αριθμούς. Τέλος, δεδομένου ότι το $0 \in \mathcal{M}$ επαυξάνουμε το σύνολο των αριθμών με την μονάδα που αντιστοιχεί στην αναπαράσταση του μηδενός. □

**Παρατήρηση:**

- Το σύνολο των αριθμών μηχανής $\mathcal{M}$ είναι διακριτό και όχι συνεχές. Στους ψηφιακούς υπολογισμούς, το άπειρο και συνεχές σύνολο των πραγματικών αριθμών διακριτοποιείται σε ένα πεπερασμένο διακριτό υποσύνολό του που αποτελεί τους αριθμούς μηχανής.

**Ιδιότητες πράξεων στο $\mathcal{M}$:**

1. Ισχύει η Αντιμεταθετικότητα
   Εάν $a, b \in \mathcal{M}$ τότε
   $a + b = b + a$
   $ab = ba$

2. Δεν ισχύει η Προσεταιριστικότητα
   Εάν $a, b, c \in \mathcal{M}$ τότε
   $a + (b + c)$ μπορεί να διαφέρει από το $(a + b) + c$
   $a(bc)$ μπορεί να διαφέρει από το $(ab)c$
3. Δεν ισχύει η Επιμεριστική ιδιότητα
   Εάν $a, b, c \in \mathcal{M}$ τότε
   $a(b + c)$ μπορεί να διαφέρει από το $ab + ac$
4. Το σύνολο $\mathcal{M}$ δεν είναι κλειστό στις βασικές αριθμητικές πράξεις της πρόσθεσης, αφαίρεσης, πολ/μού, διαίρεσης.

Από τα παραπάνω παρατηρούμε ότι τα αποτελέσματα στην floating point αριθμητική πολλές φορές μπορεί να εξαρτώνται από τη σειρά με την οποία εκτελούνται οι υπολογισμοί.

**Παράδειγμα 1.1.3** *Έστω ότι σε ένα σύστημα κινητής υποδιαστολής έχουμε* $\beta = 10,\ t = 3, m = -1, M = 2$. *Αναπαραστήστε στο σύστημα αυτό τους αριθμούς* $a = 11.2,\ b = 1.13,\ c = a \cdot b$, *και εξετάστε εάν είναι κλειστό ως προς το γινόμενο.*

Έχουμε, $a = 11.2 = 0.112 \cdot 10^2$,
$b = 1.13 = 0.113 \cdot 10^1$
$c = a \cdot b = 12.656 = 0.12656 \cdot 10^2$
Άρα ο αριθμός $c$ δεν ανήκει στο $\mathcal{M}(10, 3, -1, 2)$. Επομένως το σύστημα κινητής υποδιαστολής $\mathcal{M}(10, 3, -1, 2)$ δεν είναι κλειστό ως προσ τον πολλαπλασιασμό. □

Η mantissa περιέχει $t$ ψηφία της εκάστοτε βάσης $\beta$ και όλοι οι αριθμοί που είναι μεγαλύτεροι πρέπει να αποκοπούν (shortened) σ' αυτό το δεδομένο μήκος με άμεση συνέπεια να περιορίζεται η ακρίβεια κάθε υπολογισμού. Επομένως έχουμε **$t$-ψηφίων αριθμητική κινητής υποδιαστολής**.

## 1.2 Αριθμοί Μηχανής

Οποιοσδήποτε αριθμητικός υπολογισμός θα υλοποιηθεί με χρήση των αριθμών μηχανής του συγκεκριμένου υπολογιστή. Είναι λοιπόν πολύ σημαντικό να μελετήσουμε τον τρόπο αποθήκευσης αυτών των αριθμών και να δούμε και τον τρόπο παράστασής τους σ'ένα σύστημα αριθμών κινητής υποδιαστολής. Ετσι θα κατανοήσουμε πως γίνεται η αντιστοίχιση ενός $x \in \mathbb{R}$ με έναν αριθμό μηχανής $fl(x) \in \mathcal{M}$.

### 1.2.1 Αποθήκευση των floating point αριθμών

Στην αποθήκευση των floating point αριθμών, τα bits της λέξης του υπολογιστή κατανέμονται για την αποθήκευση του προσήμου της χαρακτηριστικής, καθώς και των

τιμών του εκθέτη και της mantissa. Εάν για την αποθήκευση των αριθμών χρησιμοποιηθούν ψηφία από δύο χαρακτηριστικές (δηλαδή δεσμευθούν δύο λέξεις) τότε μιλάμε για υπολογισμούς διπλής ακριβείας (**double precision**) σε αντίθεση με τους υπολογισμούς απλής ακριβείας (**single precision**) για τους οποίους χρησιμοποιείται μία μόνο λέξη υπολογιστή. Οι υπολογισμοί διπλής ακριβείας, όπως φαίνεται από την ονομασία τους, δίνουν διπλή ακρίβεια αλλά είναι κάπως αργότεροι στην εκτέλεση.

## Παράσταση λέξης υπολογιστή σε IEEE αριθμητική

Οι αριθμοί κινητής υποδιαστολής αποτελούν τον πυρήνα κάθε αριθμητικού υπολογισμού, είναι λοιπόν απαραίτητο να δούμε πως τους αποθηκεύει και τους επεξεργάζεται ο υπολογιστής. Επειδή η αποθήκευση είναι στενά συνδεδεμένη με το hardware του υπολογιστή, το Institute of Electrical and Electronic Engineers (IEEE) καθιέρωσε το 1985 τις βασικές αρχές παράστασης και επεξεργασίας των αριθμών κινητής υποδιαστολής. Η IEEE δυαδική αριθμητική standard binary arithmetic, χρησιμοποιείται σήμερα από όλους τους κατασκευαστές υπολογιστών στο σχεδιασμό των μονάδων επεξεργασίας της αριθμητικής κινητής υποδιαστολής, εξασφαλίζοντας έτσι τη συμβατότητα μεταξύ όλων των προγραμμάτων, ανεξαρτήτως της μηχανής που χρησιμοποιείται για την εκτέλεση.

Η IEEE περιλαμβάνει δύο κατηγορίες αριθμών κινητής υποδιαστολής˙ απλής ακρίβειας (single precision) με λέξεις των 32 bits και διπλής ακρίβειας (double precision) αριθμούς, με λέξεις των 64 bits. Συνήθως $\beta = 2$.

## IEEE Floating Point Standard 754-1985

## IEEE Απλή Ακρίβεια - word 32 bits

**Format απλής ακρίβειας** $\mathcal{M}(2, 23, -127, 128)$
Η κατανομή των 32 bits της λέξης γίνεται ως εξής.

- πρόσημο mantissa 1 bit
- εκθέτης 8 bits
- mantissa (επονομάζεται και significand) 23 bits

| 1 bit | 8 bits | 23 bits |
|---|---|---|
| πρόσημο | biased exponent[1] | mantissa |

single format

Κατά συνέπεια ο υπολογιστής αυτός έχει το ακόλουθο εύρος παράστασης αριθμών:

$$x = \sigma * (.a_1 a_2 \dots a_{23})_2 * 2^e$$

[1] biased representation είναι μια σταθερή τιμή που ονομάζεται the bias και αφαιρείται από το πεδίο για να μας δώσει την αληθή τιμή του εκθέτη

$$|e| \leq (11111111)_2 = 2^8 - 1 = 255$$

Το εύρος παράστασης είναι από 0 έως 255, με την παραδοχή ότι οι τιμές από 0 έως 126 αντιστοιχούν σε αρνητικούς εκθέτες , το 127 σε εκθέτη 0 και οι τιμές από 128 έως 255 σε θετικούς εκθέτες. Επειδή λοιπόν ο εκθέτης παίρνει και θετικές και αρνητικές τιμές, αντί να χρησιμοποιείται ένα ξεχωριστό bit για το πρόσημο του εκθέτη, το IEEE Floating Point Standard 754-1985 χρησιμοποιεί τη biased representation με τιμή για το bias το 127. Από το εύρος λοιπόν του εκθέτη αφαιρούμε το 127 και το τελικό εύρος είναι από -127 έως 128. Οι εκθέτες -127 (όλα 0) και +128 (όλα 1) δεσμεύονται για την αναπαράσταση ειδικών αριθμών.
Για να αυξηθεί περισσότερο η ακρίβεια στη mantissa, το IEEE 754 standard χρησιμοποιεί κανονικοποιημένη mantissa. Η πρώτη μονάδα, που πάντα θα εμφανίζεται λόγω κανονικοποίησης, δε θα αποθηκεύεται και έτσι έχουμε ότι το πεδίο των 23-bits χρησιμοποιείται για την αποθήκευση σε mantissa 24-bit με τιμή ανάμεσα στο 0.5 και 1. Θεωρούμε δε ότι αυτή η μονάδα θα εμφανίζεται αριστερά της νοητής υποδιαστολής. Έτσι ένας κανονικοποιημένος floating point αριθμός στο IEEE Floating Point Standard 754-1985 θα παριστάνεται ως εξής:

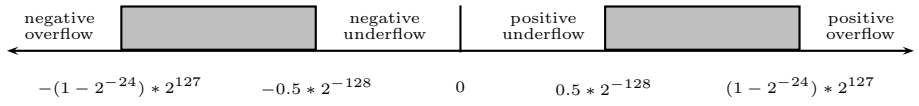
$$x = (-1)^s * (1.a_1a_2 \ldots a_{23})_2 * 2^{(exponent-127)}$$

όπου s είναι το ““sign bit”” με τιμή 0 για θετικό αριθμό και 1 για αρνητικό.

Έτσι λοιπόν ο μικρότερος θετικός παραστήσιμος αριθμός είναι:
$(1.00...0)_2 * 2^{-126} = 1 * 2^{-126} \approx 1.1755 * 10^{-38}$

και ο μέγιστος θετικός παραστήσιμος αριθμός είναι:
$(1.11...1)_2 * 2^{127} = 1 + (1 - 2^{-23}) * 2^{127} \approx 3.403 * 10^{38}$

Αξίζει να σημειωθεί ότι σε fixed point αριθμητική, ο μέγιστος παραστήσιμος αριθμός σε λέξη των 32 bits είναι: $2^{31} - 1 = 2147483647$.




Σχήμα 1.2: Εύρος παράστασης αριθμών σε format 32 bits

**Το ‘‘μηδέν της μηχανής’’**
Στο IEEE Floating Point Standard 754-1985 έχει σημασία να καθορίσουμε τον ελάχιστο αριθμό που είναι το λεγόμενο ‘‘μηδέν της μηχανής’’ (machine epsilon or machine precision). Αυτός λοιπόν προσδιορίζεται σαν ό ελάχιστος αριθμός ο οποίος προστιθέμενος στη μονάδα αυξάνει την τιμή της. Σύμφωνα λοιπόν με τον τρόπο παράστασης των αριθμών, για την απλή ακρίβεια το ‘‘μηδέν της μηχανής’’ είναι το $2^{-23}$ επειδή ισχύει $1 + 2^{-23} = 1.000 \ldots 1 > 1$. Συνήθως το ‘‘μηδέν της μηχανής’’ συμβολίζεται $\epsilon_{\text{mach}}$. Δεδομένου λοιπόν ότι μεταξύ του 1 και του επόμενου μεγαλύτερου αριθμού που μπορεί να παρασταθεί η διαφορά είναι $2^{-23} = 1.19 * 10^{-7}$, η ακρίβεια των

δεκαδικών αριθμών που αποθηκεύονται με αριθμητική απλής ακριβείας εξασφαλίζει 7 δεκαδικά ψηφία ακριβείας στο αποτέλεσμα.

**Παράδειγμα 1.2.1** *Να αποθηκευθεί η μονάδα σε IEEE 754-32 bit floating point format.*

Από τα όσα έχουμε αναφέρει μέχρι τώρα μπορούμε εύκολα να δούμε ότι η αποθήκευση θα γίνει ως εξής:

Κατάρχήν η μονάδα παριστάνεται στο δυαδικό σύστημα ως εξής.
$1 = 1.0 * 2^0$
Για το εκθέτη έχουμε $(exponent - 127) = 0 \quad \rightarrow$ ο εκθέτης θα είναι 127. Εχουμε ότι $127_{(10)} = 1111111_{(2)}$

Τελικά ακολουθώντας το IEEE format η αναπαράσταση των bits θα είναι ως εξής:

| 1 bit | 8 bits | 23 bits |
|---|---|---|
| 0 | 01111111 | 000 . . . 0 |

□

**Παράδειγμα 1.2.2** *Να αποθηκευθεί ο αριθμός* 52.21875 *σε IEEE 754-32 bit floating point format.*

Από τα όσα έχουμε αναφέρει μέχρι τώρα μπορούμε εύκολα να δούμε ότι η αποθήκευση θα γίνει ως εξής:

Κατάρχήν ο πραγματικός αριθμός θα μετατραπεί στον ισοδύναμό του στο δυαδικό σύστημα.
$52.21875 = 110100.00111 \qquad = 1.1010000111 * 2^5$
Κανονικοποιημένη mantissa=.1010000111
Για το εκθέτη έχουμε $(exponent - 127) = 5 \quad \rightarrow$ και άρα ο εκθέτης θα είναι 132. Εχουμε ότι
$132 = 10000100$

Τελικά ακολουθώντας το IEEE format η αναπαράσταση των bits θα είναι ως εξής:

| 1 bit | 8 bits | 23 bits |
|---|---|---|
| 0 | 10000100 | 1010000111. . . 0 |

Παρατηρήστε ότι το πρώτο μηδέν αντστοιχεί στο πρόσημο, καθώς επίσης ότι δεν αποθηκεύεται η πρώτη μονάδα του κανονικοποιημένου αριθμού. Έτσι το πεδίο των 23 bits χρησιμοποιείται για την αποθήκευση μιας mantissa 24 bits. □

## IEEE Διπλή Ακρίβεια - word 64 bits

**Format διπλής ακρίβειας** $\mathcal{M}(2, 52, -1023, 1024)$

Η κατανομή των 64 bits της λέξης γίνεται ως εξής.

- πρόσημο mantissa 1 bit
- εκθέτης 11 bits
- mantissa (επονομάζεται και significand) 52 bits

| 1 bit | 11 bits | 52 bits |
|---|---|---|
| πρόσημο | biased exponent | mantissa |

single format

Κατά συνέπεια ο υπολογιστής αυτός θα έχει το ακόλουθο εύρος παράστασης αριθμών:

$$x = \sigma \cdot (.a_1 a_2 \dots a_{52})_2 \cdot 2^e$$

$$|e| \leq (11111111111)_2 = 2^{11} - 1 = 2047$$

Ετσι ένας κανονικοποιημένος floating point αριθμός στο IEEE Floating Point Standard 754-1985 θα παριστάνεται ως εξής:

$$x = (-1)^s * (1.a_1 a_2 \dots a_{53})_2 * 2^{(exponent-1023)}$$

όπου s είναι το ""sign bit"" με τιμή 0 για θετικό αριθμό και 1 για αρνητικό.
Ο μικρότερος παραστήσιμος αριθμός:$(0.10\dots0)_2 \cdot 2^{-1023} \approx 10^{-308}$
Ο μέγιστος παραστήσιμος αριθμός:$(0.11\dots1)_2 \cdot 2^{1024} \approx 10^{308}$

Το τελευταίο ψηφίο στη mantissa των 53 ψηφίων δίνει ακρίβεια $2^{-53} \cong 10^{-16}$. Αυτό εξασφαλιζει ακρίβεια περίπου 16 δεκαδικών ψηφίων στο αποτέλεσμα.

Ο παρακάτω πίνακας συνοψίζει τις τιμές που χαρακτηρίζουν το format απλής και διπλής ακρίβειας

| | single format | double format |
|---|---|---|
| word | 32 | 64 |
| exponent | 8 | 11 |
| bias | 127 | 1023 |
| max exp | 127 | 1023 |
| min exp | -126 | -1022 |
| number range | $10^{-38} - 10^{38}$ | $10^{-308} - 10^{308}$ |
| mantissa | 23 | 52 |
| mumber of exponents | 254 | 2046 |
| number of mantissa | $2^{23}$ | $2^{52}$ |

Η IEEE αριθμητική περιλαμβάνει τα σύμβολα $\pm\infty$ και NaN (Not a Number). Το $\pm\infty$ εμφανίζεται ανάλογα με τους εξής κανόνες:

$$\frac{x}{\pm\infty} = 0, \ \forall x \in \mathcal{M}$$

$$\frac{x}{0} = \pm\infty, \ \forall x \in \mathcal{M}, \ x \neq 0$$

$$+\infty + \infty = +\infty$$

Το σύμβολο NaN εμφανίζεται όταν εκτελούνται πράξεις χωρίς καλά καθορισμένο ή άπειρο αποτέλεσμα όπως:

$$\infty - \infty,\ \frac{\infty}{\infty},\ \frac{0}{0},\ \sqrt{-1},\ \mathrm{NaN} \cdot x$$

Ορίζουμε λοιπόν:

| | sign | biased exponent | fraction | value |
|---|---|---|---|---|
| $+\infty$ | 0 | $255(\acute{o}\lambda\alpha\, 1)$ | 0 | $\infty$ |
| $-\infty$ | 1 | $255(\acute{o}\lambda\alpha\, 1)$ | 0 | $-\infty$ |
| 0 | 0 | 0 | 0 | 0 |
| -0 | 1 | 0 | 0 | 0 |

Γι' αυτό το εύρος στο format απλής ακρίβειας είναι από 1 έως 254.

### 1.2.2 Παράσταση των αριθμών του $\mathcal{M}$

Ενας αριθμός καλείται παραστάσιμος εάν μπορεί να παρασταθεί μέσα στον υπολογιστή. Εστω ο αριθμός

$$x = \sigma \cdot \overline{x} \cdot \beta^e$$

όπου η mantissa $\overline{x}$ έχει $t$ ψηφία και $m \le e \le M$

Τότε

$$\underbrace{(.\ \overbrace{10...0}^{t-\psi\eta\phi\acute{\iota}\alpha}\ )_\beta}_{\mu\iota\kappa\rho\acute{o}\tau\varepsilon\rho\eta\ \text{mantissa}} \le \overline{x} \le \underbrace{(0.\overbrace{(\beta-1)\cdot(\beta-1)...(\beta-1)}^{t-\psi\eta\phi\acute{\iota}\alpha})_\beta}_{\mu\varepsilon\gamma\alpha\lambda\acute{\upsilon}\tau\varepsilon\rho\eta\ \text{mantissa}} \Rightarrow$$

$$\frac{1}{\beta} \le \overline{x} \le (\beta-1)\cdot\underbrace{(\frac{1}{\beta}+\frac{1}{\beta^2}+...+\frac{1}{\beta^t})}_{\frac{\beta^{-1}\cdot(\beta^{-t}-1)}{\beta^{-1}-1}} = 1-\beta^{-t} \Rightarrow$$

$$\beta^{e-1} \le \overline{x}\cdot\beta^e \le \beta^e - \beta^{e-t} < \beta^e$$

Έτσι

$$\beta^{e-1} \le |x| < \beta^e$$

Εάν η mantissa είναι της μορφής:

$$\overline{x} = (.a_1 \overbrace{a_2 a_3 ... a_t}^{t-1})_\beta$$

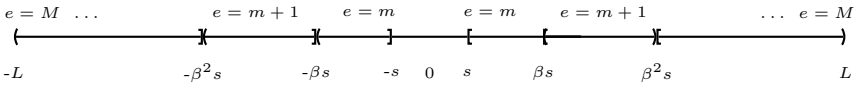
για κάθε $e = m, m+1, ..., M-1, M$ υπάρχουν $(\beta-1)\cdot\beta^{t-1}$ κανονικοποιημένες mantissa's επειδή το $a_1$ λαμβάνει $(\beta-1)$ διαφορετικές τιμές και τα υπόλοιπα $a_i$,

πλήθους $(t-1)$, λαμβάνουν $\beta$ διαφορετικές τιμές. Αυτές αντιστοιχούν σε $(\beta-1)\cdot\beta^{t-1}$ $x$ ισοκατανεμημένα σε κάθε ένα από τα διαστήματα $[\beta^{e-1},\ \beta^e), (-\beta^e,\ -\beta^{e-1}]$.

**Παρατήρηση:** Στο διάστημα $[\beta^{e-1}, \beta^e)$ κατανέμονται $\beta^{t-1}\cdot(\beta-1)$ αριθμοί $x$. Το μήκος του διαστήματος αυτού δια του αριθμού των δυνατών $x$ μας δίνει το βήμα κατανομής. Συγκεκριμένα:

$$\frac{\beta^{\mathrm{e}}-\beta^{\mathrm{e}-1}}{\beta^{\mathrm{t}-1}\cdot(\beta-1)}=\frac{\beta^{\mathrm{e}}\cdot(1-\frac{1}{\beta})}{\beta^{\mathrm{t}}\cdot\beta^{-1}\cdot(\beta-1)}=\frac{\beta^{\mathrm{e}}\cdot\frac{\beta-1}{\beta}}{\beta^{\mathrm{t}}\cdot\frac{\beta-1}{\beta}}=\beta^{e-t}$$

Οι αριθμοί $x$ είναι ισοκατανεμημένοι στο διάστημα $[\beta^{e-1}, \beta^e)$ με βήμα κατανομής $\beta^{e-t}$. Εάν θέσουμε $s=\beta^{m-1}$ προκύπτει η ακόλουθη παράσταση των αριθμών μηχανής.




Σχήμα 1.3: Ευθεία αριθμών μηχανής

Το εύρος αναπαράστασης είναι: $[s, L), (-L, -s]$.

Εάν $|x| > L$ έχουμε overflow ενώ εάν $|x| < s$ έχουμε underflow.
Αυξάνοντας το $e$ κατά 1 το μήκος στα διαστήματα $[\beta^{e-1}, \beta^e)$ και $(-\beta^e, -\beta^{e-1}]$ πολλαπλασιάζεται κατά $\beta$.
Ολα όμως τα διαστήματα που προκύπτουν περιέχουν τον ίδιο αριθμό από δυνατές mantissa's δηλαδή $(\beta-1)\cdot\beta^{t-1}$ δυνατά $x$.

**Συμπέρασμα**: Τα παραστήσιμα $x$ είναι κατανεμημένα πυκνά γύρω από το 0 και αραιότερα όσο απομακρυνόμαστε. Ετσι έχουμε καλλίτερη ακρίβεια στους υπολογισμούς μας όταν δουλεύουμε με αριθμούς που είναι σχετικά μικροί. π.χ. φράσσονται από τη μονάδα. Γενικά πιο πολλά bits στη mantissa αυξάνουν την **πυκνότητα** στα παραστάσιμα $x$ ενώ πιο πολλά bits στον εκθέτη αυξάνουν το **εύρος αναπαράστασης** των $x$.

**Παράδειγμα 1.2.3** *Δίνεται το σύστημα $\mathcal{M}(2,3,-1,3)$. Να προσδιοριστούν οι μή αρνητικοί αριθμοί μηχανής που ανήκουν σ' αυτό και να παρασταθούν γραφικά.*

**Απόδειξη:** Σύμφωνα με τα παραπάνω δεδομένου ότι $m=-1,\ \ M=3$ έχουμε:

$$s=\beta^{-1-1}=2^{-2}=\frac{1}{4}=0.25,$$

Για κάθε $e$ υπάρχουν $(\beta-1)\beta^{t-1}=4$ αριθμοί μέσα σε διαστήματα της μορφής $[\beta^{e-1}, \beta^e)$, ενώ το βήμα κατανομής θα είναι $\beta^{e-t}$ που στο παράδειγμά μας για την πρώτη τιμή του εκθέτη θα είναι $2^{-1-3}=\frac{1}{16}=0.0625$. Και έτσι για τις διάφορες τιμές του $e$ προκύπτουν τα ακόλουθα διαστήματα:

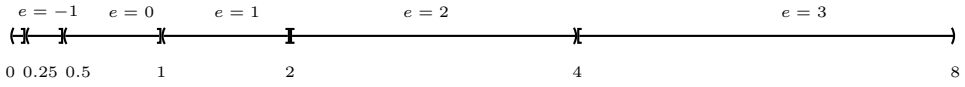
$$e=-1\rightarrow[s\ \ \beta s)=[0.25\ \ 0.5)$$

$$e = 0 \to [\beta s \;\; \beta^2 s) = [0.5 \;\; 1)$$
$$e = 1 \to [\beta^2 s \;\; \beta^3 s) = [1 \;\; 2)$$
$$e = 2 \to [\beta^3 s \;\; \beta^4 s) = [2 \;\; 4)$$
$$e = 3 \to [\beta^4 s \;\; \beta^5 s) = [4 \;\; 8)$$

Σε κάθε διάστημα βρίσκονται κατανεμημένοι οι ακόλουθοι αριθμοί μηχανής (εκπεφρασμένοι σε δεκαδική μορφή).

| $e$ | διάστημα | βήμα κατανομής $\beta^{e-t}$ | αριθμοί μηχανής |
|---|---|---|---|
| $-1$ | $[s \;\; \beta s) = [0.25 \;\; 0.5)$ | $\frac{1}{16} = 0.0625$ | $0.25$<br>$0.25 + \frac{1}{16} = 0.3125$<br>$0.3125 + \frac{1}{16} = 0.375$<br>$0.375 + \frac{1}{16} = 0.4375$ |
| $0$ | $[\beta s \;\; \beta^2 s) = [0.5 \;\; 1)$ | $\frac{1}{8} = 0.125$ | $0.5$<br>$0.5 + \frac{1}{8} = 0.625$<br>$0.625 + \frac{1}{8} = 0.75$<br>$0.75 + \frac{1}{8} = 0.875$ |
| $1$ | $[\beta^2 s \;\; \beta^3 s) = [1 \;\; 2)$ | $\frac{1}{4} = 0.25$ | $1$<br>$1 + \frac{1}{4} = 1.25$<br>$1.25 + \frac{1}{4} = 1.5$<br>$1.5 + \frac{1}{4} = 1.75$ |
| $2$ | $[\beta^3 s \;\; \beta^4 s) = [2 \;\; 4)$ | $\frac{1}{2} = 0.5$ | $2$<br>$2 + \frac{1}{2} = 2.5$<br>$2.5 + \frac{1}{2} = 3$<br>$3. + \frac{1}{2} = 3.5$ |
| $3$ | $[\beta^4 s \;\; \beta^5 s) = [4 \;\; 8)$ | $\frac{1}{1} = 1$ | $4$<br>$4 + 1 = 5$<br>$5 + 1 = 6$<br>$6 + 1 = 7$ |

Έτσι αυτοί οι αριθμοί μπορούν να παρασταθούν ως εξής




Σχήμα 1.4: Ευθεία αριθμών μηχανής

Παρατηρούμε λοιπόν ότι τα διαστήματα διπλασιάζονται σε μήκους όμως σε κάθε ένα από αυτά παραμένουν 4 αριθμοί μηχανής.

□

Ετσι για την παράσταση στον υπολογιστή ενός οποιουδήποτε αριθμού $x \in \mathbb{R}$ θα επιλέγεται ο αριθμός μηχανής $fl(x) \in \mathcal{M}$ που θα βρίσκεται πλησιέστερα σ' αυτόν. Ο μετασχηματισμός $x \to fl(x)$ ονομάζεται στρογγύλευση. Προξενείται υπερχείλισε

όταν $|fl(x)| > \max\{|y| : \ y \in \mathcal{M}\}$. Υποχείλιση έχουμε όταν $0 < |fl(x)| < \min\{|y| : 0 \neq y \in \mathcal{M}\}$

**Παραδείγματα Overflow - Underflow**

Πιο σημαντικό πρόβλημα εμφανίζεται στην υπερχείλιση (overflow) όπου δίνεται σαν τελικό αποτέλεσμα η τιμή $\pm\infty$.
Σαν αποτέλεσμα της υποχείλισης δίνεται η τιμή 0 ή $\pm\beta^m$ ή κάποιος μη κανονικοποιημένος αριθμός.

**Υπερχείλιση**

$\beta = 10,\ t = 3,\ m = -4,\ M = 4$
$a = 0.111 \cdot 10^4$
$b = 0.120 \cdot 10^4$
$c = a \cdot b = 0.133 \cdot 10^7$

**Υποχείλιση**

$\beta = 10,\ t = 3,\ m = -3,\ M = 4$
$a = 0.1 \cdot 10^{-3}$
$b = 0.2 \cdot 10^{-3}$
$c = a \cdot b = 0.2 \cdot 10^{-7}$

Απλές μαθηματικές πράξεις μπορεί να οδηγήσουν σε υπερχείλιση. Με κατάλληλη οργάνωση μπορεί αυτό να αποφευχθεί.

**Παράδειγμα 1.2.4** *Υπολογίστε τη* $\|x\|_2^2$, $x \in \mathbb{R}^n$ *με τον καλύτερο δυνατό τρόπο.*

Αυτός ο υπολογισμός μπορεί να εκτελεσθεί με έναν εκ των δύο αλγορίθμων.

**Αλγόριθμος 1**
$\|x\|_2^2 = x_1^2 + x_2^2 + ... + x_n^2$

**Αλγόριθμος 2**
$m = \max(|x_1|, ..., |x_n|)$
$y_i = \frac{\mathrm{x_i}}{\mathrm{m}}, i = 1, \ldots, n$
$\|x\|_2 = m \cdot \sqrt{y_1^2 + y_2^2 + ... + y_n^2}$

Ενώ ο αλγόριθμος 1 μπορεί να οδηγήσει σε υπερχείλιση εάν εφαρμοσθεί η τεχνική του αλγορίθμου 2 τα αποτελέσματα ελέγχονται. Εν γένει, εάν δουλεύουμε με αριθμούς φραγμένους (ιδανική περίπτωση εάν όλοι είναι μικρότεροι της μονάδας) ελέγχεται το εύρος των αναμενόμενων τιμών. □

## Στρογγύλευση και αποκοπή

Εάν ο πραγματικός αριθμός δεν έχει mantissa $\overline{x}$ που να ταιριάζει στη διαθέσιμη ακρίβεια των $t$ ψηφίων της μαντισσα θα mantissa θα πρέπει να περικοπεί. Ετσι εισάγονται τα λεγόμενα στρογγύλευσης (rounding errors). Υπάρχουν οι τεχνικές της αποκοπής (όσα ψηφία είναι παραπάνω από $t$ αποκόπτονται) και της στρογγύλευσης.

Εστω $\hat{x}$ προσέγγιση της τιμής $x$. Το σφάλμα μπορεί να εκτιμηθεί με τις ποσότητες:

Απόλυτο σφάλμα $= |\hat{x} - x|$

Σχετικό σφάλμα $= \frac{|\hat{x}-x|}{|x|}, x \neq 0$

Το σχετικό σφάλμα δίνει πάντα καλλίτερη εκτίμηση απ΄ότι το απόλυτο και είναι ανεξάρτητο από το scaling. Συγκεκριμένα, το σχετικό σφάλμα των ποσοτήτων $x$ και $\hat{x}$ είναι ίδιο με το σχετικό σφάλμα των ποσοτήτων $ax$ και $a\hat{x}$.

**Παράδειγμα 1.2.5** *Έστω $x_1 = 1.31$, $\hat{x_1} = 1.30$ και $x_2 = 0.12$, $\hat{x_2} = 0.11$. Ποια εκ των δύο προσεγγίσεων είναι καλλίτερη;*

$$|\hat{x_1} - x_1| = |\hat{x_2} - x_2| = 0.01$$

$$|\frac{\hat{x_1} - x_1}{x_1}| = 0.0076$$

$$|\frac{\hat{x_2} - x_2}{x_2}| = 0.0833$$

Αρα το $\hat{x_1}$ είναι κοντίτερα στο $x_1$ απ΄ότι το $\hat{x_2}$ στο $x_2$. □

Στη μελέτη του απόλυτου και σχετικού σφάλματος υπεισέρχεται και η έννοια των σημαντικών ψηφίων (significant digits). Εάν υποθέσουμε ότι έχουμε τις ακόλουθες τιμές

1. Θεωρητική τιμή : $d_1 d_2 \ldots d_n d_{n+1} \ldots d_p, \;\; d_1 \neq 0$
2. Προσεγγιστική τιμή : $d_1 d_2 \ldots d_n e_{n+1} \ldots e_p, \;\; d_1 \neq 0$

οι οποίες διαφέρουν από το $(n+1)$ ψηφίο και μετά. Ορίζουμε ότι οι τιμές 1 και 2 συμφωνούν σε $n$ σημαντικά ψηφία εάν $|d_{n+1} - e_{n+1}| < 5$. Διαφορετικά λέμε ότι συμφωνούν σε $n-1$ ψηφία.

Για παράδειγμα, έστω ότι

1. Θεωρητική τιμή : $\frac{10}{3}$
2. Προσεγγιστική τιμή : 3.333

Τότε προκύπτει ότι το απόλυτο σφάλμα είναι ίσο με : $0.000333\ldots = \frac{1}{3000}$
ενώ το σχετικό σφάλμα είναι ίσο με :

$$\frac{1}{3000} / \frac{10}{3} = \frac{1}{10000} = 1 \cdot 10^{-4}$$

Ο αριθμός των σημαντικών ψηφίων εκφράζεται από τον εκθέτη $d$ του 10 εάν θέσουμε το σχετικό σφάλμα στη μορφή $1.0 \cdot 10^{-d}$, δηλαδή το 4 στο προηγουμένο παράδειγμα.

### Στρογγύλευση

**Θεώρημα 1.2.1** *Έστω $fl(x)$ η floating point αναπαράσταση ενός $x \in \mathbb{R}$. Τότε το σχετικό σφάλμα που προκύπτει από την αναπαράσταση αυτή δίνεται από τη σχέση:*

$$|\frac{\mathrm{fl(x)-x}}{\mathrm{x}}| \leq \mu = \begin{cases} \frac{1}{2} \cdot \beta^{1-t} & στρογγύλευση \\ \beta^{1-t} & αποκοπή \end{cases}$$

**Απόδειξη:**

Στη στρογγύλευση έχουμε ότι ο αριθμός

$$x = \sigma \cdot \overline{x} \cdot \beta^e = \sigma \cdot (.a_1 a_2 \ldots a_t a_{t+1} \ldots)_\beta \cdot \beta^e$$

θα στρογγυλευθεί σε μαντισσα με $t$ ψηφία οπότε

$$fl(x) = \begin{cases} x' = \sigma(0.a_1 a_2 ... a_t)_\beta \cdot \beta^e & , 0 \leq a_{t+1} < \frac{\beta}{2} \\ x'' = \sigma[(0.a_1 a_2 ... a_t)_\beta + \cdot \beta^{-t}] \cdot \beta^e & , \frac{\beta}{2} \leq a_{t+1} < \beta \end{cases}$$

Προφανώς $x \in (x', x'')$. Εστω ότι είναι πιο κοντά στο $x'$

$$|x - x'| \leq \frac{1}{2}|x' - x''| = \frac{1}{2} \cdot \beta^{e-t}$$

Το σχετικό σφάλμα είναι

$$|\frac{\mathrm{x - x'}}{\mathrm{x}}| \leq \frac{1}{2} \cdot \frac{\beta^{-\mathrm{t}}}{(0.\mathrm{a_1 a_2 \ldots a_t \ldots})_\beta} \leq \frac{1}{2} \cdot \frac{\beta^{-\mathrm{t}}}{\frac{1}{\beta}} = \frac{1}{2} \cdot \beta^{1-t}$$

Στην αποκοπή έχουμε ότι ο αριθμός

$$x = \sigma \cdot \overline{x} \cdot \beta^e = \sigma \cdot (.a_1 a_2 \ldots a_t a_{t+1} \ldots)_\beta \cdot \beta^e$$

αποκόπτεται σε mantissa με $t$ ψηφία οπότε,

$$|x - x'| \leq |x' - x''| = \beta^{e-t}$$

$$|\frac{\mathrm{x - x'}}{\mathrm{x}}| \leq \frac{\beta^{\mathrm{e-t}}}{(0.\mathrm{a_1 a_2 \ldots a_t \ldots})_\beta \cdot \beta^{\mathrm{e}}} \leq \frac{\beta^{-\mathrm{t}}}{\frac{1}{\beta}} = \beta^{1-t}$$

□

**Παρατηρήσεις:**

- Η ακρίβεια που πετυχαίνουμε με την τεχνική της στρογγύλευσης είναι πολύ καλλίτερη από εκείνη της αποκοπής αφού το φράγμα του σχετικού σφάλματος είναι το ήμισυ από εκείνο της αποκοπής.

- Το σφάλμα στην τεχνική της στρογγύλευσης αλλάζει πρόσημο (πότε είναι θετικό και πότε είναι αρνητικό) με αποτέλεσμα την πιθανή απαλοιφή του, ενώ στην τεχνική της αποκοπής το πρόσημο στο σφάλμα είναι πάντα θετικό.

Εάν ονομάσουμε $\varepsilon = \frac{\mathrm{fl(x)-x}}{\mathrm{x}} \Rightarrow x \cdot \varepsilon = fl(x) - x \Rightarrow x \cdot \varepsilon + x = fl(x) \Rightarrow$

$$\boxed{fl(x) = x \cdot (1+\varepsilon),\ |e| \leq \tfrac{1}{2} \cdot \beta^{1-t}}$$

Ονομάζουμε την ποσότητα $\frac{1}{2} \cdot \beta^{1-t} = u$ και είναι το μοναδιαίο σφάλμα στρογγύλευσης (unit roundoff error) που χαρακτηρίζει κάθε υπολογιστή.

**Συμπέρασμα**: Η στρογγυλευθείσα τιμή $fl(x)$ είναι μία ελαφριά διατάραξη της θεωρητικής $x$.

**Παραδείγματα:**

(ι) Για τον **Cyber 170-730** $\beta = 8,\ t = 64\ \Rightarrow u = \frac{1}{2} \cdot 8^{-63} \simeq 0.635 \cdot 10^{-57}$
(ιι) Για τον **IBM 360** $\beta = 16,\ t = 7$ (ή $t = 14$) (απλή ή διπλή ακρίβεια) $\Rightarrow u = \frac{1}{2} \cdot 16^{-6} = 2.9 \cdot 10^{-8}$ □

**Σημείωση:** Εχουμε δείξει ότι $|fl(x) - x| \leq \frac{1}{2} \cdot \beta^{e-t}$
Εάν εξετάσουμε την ποσότητα:

$$|\frac{\mathrm{x - fl(x)}}{\mathrm{fl(x)}}| \leq \frac{\frac{1}{2} \cdot \beta^{\mathrm{e-t}}}{\sigma \cdot (0.\mathrm{a_1 \ldots a_t})_\beta \cdot \beta^{\mathrm{e}}} \leq \frac{\frac{1}{2} \cdot \beta^{\mathrm{-t}}}{(0.1 \ldots 0)_\beta} = \frac{1}{2} \cdot \beta^{1-t}$$

Θέτουμε

$$\frac{\mathrm{x - fl(x)}}{\mathrm{fl(x)}} = \delta \Rightarrow x - fl(x) = fl(x) \cdot \delta \Rightarrow fl(x) + fl(x) \cdot \delta = x \Rightarrow$$

$$\boxed{fl(x) = \tfrac{\mathrm{x}}{1+\delta},\ |\delta| \leq u}$$

## Εκτίμηση του unit roundoff($u$)

Το μοναδιαίο σφάλμα στρογγύλευσης (unit roundoff $u$) είναι χαρακτηρηστικό για κάθε μηχανή και ονομάζεται επίσης και έψιλον της μηχανής (machine epsilon). Σε πολλές εφαρμογές μας είναι χρήσιμο να γνωρίζουμε την τιμή του $u$ της μηχανής καθώς και τα άλλα χαρακτηριστικά της. Υπάρχουν βιβλιοθήκες προγραμμάτων που τα υπολογίζουν αλλά μπορούμε να υπολογίζουμε μια καλή προσέγγιση του $u$ ως εξής:
Το $u$ είναι κάτι μεγαλύτερο από το μεγαλύτερο αριθμό $x$ για τον οποίο η υπολογιζόμενη τιμή $1 + x$ ισούται με 1.
Για παράδειγμα αν $\beta = 2,\ t = 6$ για $x = 2^{-7}$ τότε $fl(x + 1) = 1$.

Η τιμή $x = 2^{-7}$ μπορεί να αποτελέσει προσέγγιση της τιμής του $u$ και αποτελεί το λεγόμενο έψιλον της μηχανής $\epsilon_M$.
Έτσι μπορούμε να υπολογίσουμε μια προσέγγιση του $\epsilon_M$ αρχίζοντας με μια μεγάλη τιμή του $x$ και ελαττώνοντάς την έως το $1+x$ να υπολογίζεται σαν 1. Αυτό υλοποιείται με τον παρακάτω αλγόριθμο.

$x = 1$
while$(1 + x \neq 1)$
    $x = x/2$

### 1.2.3 Εκτέλεση Πράξεων σε Αριθμητική Κινητής Υποδιαστολής

Ο παρακάτω πίνακας συνοψίζει τις βασικές πράξεις για αριθμητική κινητής υποδιαστολής για δύο floating point αριθμούς $x = x_s\beta^{x_E},\ y = y_s\beta^{y_E},\ x_E \leq y_E$.

| Αριθμητικές Πράξεις |
|---|
| $x + y = (x_s\beta^{x_E - y_E} + y_s)\beta^{y_E}$ |
| $x - y = (x_s\beta^{x_E - y_E} - y_s)\beta^{y_E}$ |
| $x \cdot y = (x_s \cdot y_s)\beta^{x_E + y_E}$ |
| $x/y = (x_s/y_s)\beta^{x_E - y_E}$ |

Παρατηρούμε ότι για την πρόσθεση και την αφαίρεση πρέπει να ευθυγραμμίσουμε τους εκθέτες. Ο πολλαπλασιασμός και η διαίρεση γίνονται άμεσα.

#### Πρόσθεση και αφαίρεση

Ο αλγόριθμος της πρόσθεσης και της αφαίρεσης έχει τις ακόλουθες τέσσερις φάσεις.

1. Έλεγχος για μηδενικά (πχ εάν $x + 0$ γυρίζει σαν αποτέλεσμα το $x$)
2. Ευθυγράμμιση των εκθετών
3. Πρόσθεση ή αφαίρεση των χαρακτηριστικών
4. Κανονικοποίηση του αποτελέσματος (πρώτα κανονικοποίηση και μετά στρογγύλευση)

Για την εκτέλεση της πράξης χρησιμοποιούνται καταχωρητές της ALU. Οι εκθέτες τησ mantissa αποθηκέυονται σε ξεχωριστούς καταχωρητές και ξαναενωποιούνται μετά την εύρεση του αποτελέσματος.
Για παράδειγμα

$$123 \cdot 10^0 + 456 \cdot 10^{-2} = (456 \cdot 10^{-2} + 123) \cdot 10^0 = 127.56 \cdot 10^0$$

Επειδή κατά την ευθυγράμμιση μπορεί να συμβεί απώλεια ψηφίων πάντοτε ευθυγραμμίζουμε το μικρότερο αριθμό, οπότε εάν χαθούν κάποια ψηφία αυτά δε θα είναι τόσο σημαντικά.

### Guard bits

Ο καταχωρητής της ALU που αποθηκεύει τη mantissa περιέχει επιπρόσθετα bits που ονομάζονται guard bits και χρησιμοποιούνται για συμπλήρωση του δεξιού άκρου της mantissa με μηδέν.

**Παράδειγμα 1.2.6** *Αν $t = 24$ να αφαιρεθούν οι αριθμοί $x = 1.00\ldots0 \cdot 2^1$ και $y = 1.11\ldots1 \cdot 2^0$. Η mantissa των 24 bits περιέχει και ένα bit αριστερά από το δεκαδικό σημείο.*

Πρώτα ευθυγραμμίζουμε τους εκθέτες

$$x = 1.00\ldots00 \times 2^1$$
$$y = 0.11\ldots11 \times 2^1$$

και έχουμε

$$x - y = z = 0.00\ldots01 \times 2^1$$

και μετά από την κανονικοποίηση παίρνουμε

$$z = 1.00\ldots \times 2^{-23}$$

Παρατηρούμε ότι κατα την ευθυγράμμιση των εκθετών το $y$ χάνει 1 bit ακριβείας. Αν χρησιμοποιήσουμε Guard digit θα έχουμε

$$x = 1.00\ldots00\ 0000 \times 2^1$$
$$y = 0.11\ldots11\ 1000 \times 2^1$$

και έτσι

$$x - y = z = 0.00\ldots001000 \times 2^1$$

και μετά από την κανονικοποίηση πέρνουμε $z = 1.00\ldots00 \times 2^{-24}$. Κατά την ευθυγράμμιση των εκθετών το τελευταίο σημαντικό ψηφίο δε χάνεται αλλά καταχωρείται στο guard bit. Παρατηρούμε ότι υπάρχει μιαα διαφορά ενός παράγοντα 2 ($\beta = 2$) μεταξύ των δύο αποτελεσμάτων. Σε δεκαεξαδικό σύστημα η απώλεια της ακρίβειας μπορεί να είναι μεγαλύτερη και συγκεκριμένα θα είναι ένας παράγοντας 16.

## 1.3 Αριθμητικά Αποτελεσματικοί Αλγόριθμοι

Ενα υψηλής ποιότητας μαθηματικό software πρέπει να απαρτίζεται από αριθμητικά αποτελεσματικούς αλγορίθμους (numerically effective). Ενας αλγόριθμος πινάκων (matrix algorithm) είναι αριθμητικά αποτελεσματικός εάν έχει τα ακόλουθα χαρακτηριστικά:

1. **Γενικού σκοπού**(General Purpose): Ο αλγόριθμος πρέπει να δουλεύει για μια ευρεία κλάση πινάκων.

2. **Αξιόπιστος**(Reliable): Ο αλγόριθμος πρέπει να δίνει διαγνωστικά μηνύματα αφένός μεν, όποτε υπάρχει ο κίνδυνος εμφάνισης προβλημάτων εξαιτίας των σφαλμάτων στρογγύλευσης αφέτέρου δε, όταν δεν ικανοποιούνται κάποια καθορισμένα κριτήρια σύγκλισης. Υπάρχουν αλγόριθμοι που παράγουν τελείως εσφαλμένα αποτελέσματα χωρίς να εμφανίζουν κανένα διαγνωστικό. Η μέθοδος απαλοιφής του Gauss χωρίς οδήγηση είναι ένας τέτοιος μη αξιόπιστος αλγόριθμος.

3. **Ευσταθής**: Η τελική λύση του προβλήματος, λόγω των συνολικών σφαλμάτων στρογγύλευσης που συσσωρεύτηκαν στον αλγόριθμο, αποτελεί την ακριβή λύση ενός ελαφρά διαταραγμένου αρχικού προβλήματος.

4. **Αποτελεσματικός**: Η αποτελεσματικότητα του αλγορίθμου μετράται από το χρόνο που χρειάστηκε ο υπολογιστής για την εφαρμογή του. Θεωρητικά ο αριθμός των floating-point πράξεων που χρειάζονται για την εφαρμογή του αλγορίθμου δείχνει την αποτελεσματικότητα.

   **Ορισμός:** Μία πράξη κινητής υποδιαστολής (floating point operation ή flop) είναι ο χρόνος που απαιτείται για την εκτέλεση της εντολής

   $$A(i,j) = A(i,j) + t * A(i,j)$$

   Ενα flop περιλαμβάνει ένα πολλαπλασιασμό, μία πρόσθεση και ορισμένους διαχειρισμούς δεικτών. Παρόμοια μία διαίρεση μαζί με μία πρόσθεση ή αφαίρεση θα μετρούνται σαν ένα flop. Αυτός ο ορισμός του flop χρησιμοποιήθηκε στο πολύ δημοφιλές πακέτο λογισμικού LINPACK. □

   **Ορισμός:** Ενας αλγόριθμος πινάκων που περιλαμβάνει υπολογισμούς με πίνακες τάξης $n$ θα καλείται αποτελεσματικός (efficient) εάν δε χρειάζεται περισσότερα από $O(n^3)$ flops. (Ο ιστορικός κανόνας του Cramer για την επίλυση ενός γραμμικού συστήματος δεν είναι αποτελεσματικός αφού απαιτούνται $O(n!)$ flops για την εκτέλεσή του).

   Υπάρχει ένα σημείο που πρέπει να τονίσουμε εδώ. Ενας αλγόριθμος μπορεί να είναι αποτελεσματικός αλλά να είναι ασταθής (unstable).

   Για παράδειγμα, η μέθοδος απαλοιφής του Gauss χωρίς οδήγηση απαιτεί $\frac{n^3}{3}$ flops για έναν $n \times n$ πίνακα. Κατά συνέπεια παρόλο που είναι αποτελεσματική είναι αναξιόπιστη και ασταθής για έναν αυθαίρετο πίνακα.

5. **Οικονομικός στη χρήση μνήμης**: Ενας πυκνός πίνακας τάξης $n$ χρειάζεται $n^2$ θέσεις μνήμης για την αποθήκευση των στοιχείων του. Κατά συνέπεια εάν ένας αλγόριθμος κατά τη διάρκεια της εκτέλεσής του απαιτεί αποθήκευση διαφόρων πινάκων, θα χρειασθεί μεγάλος αριθμός θέσεων μνήμης ακόμα και αν το $n$ μεταβάλλεται. Ετσι είναι σημαντικό να δίνουμε ιδιαίτερη προσοχή κατά το

σχεδιασμό ενός αλγορίθμου στην οικονομική διαχείριση της μνήμης. Με ανακατατάξεις στον αλγόριθμο μπορούμε να πετύχουμε ελάττωση στις αποθηκευτικές απαιτήσεις. Γενικά, εάν ένας πίνακας που παράγεται κατά την εκτέλεση του αλγορίθμου δεν πρόκειται να χρησιμοποιηθεί μελλοντικά θα πρέπει να επικαλυφθεί από άλλα υπολογιζόμενα στοιχεία.

Ενας $n \times n$ πίνακας απαιτεί για την αποθήκευσή του $n^2$ θέσεις μνήμης. Στην Αριθμητική Γραμμική Αλγεβρα συχνά ασχολούμεθα με άνω ή κάτω τριγωνικούς πίνακες οι οποίοι απαιτούν $\frac{n\cdot(n+1)}{2}$ αντί για $n^2$. Στην περίπτωση αυτή οι επιπλέον θέσεις μπορούν να χρηριμοποιηθούν για αποθήκευση άλλων στοιχείων. Ενας $n \times n$ τριδιαγώνιος πίνακας απαιτεί μόνο $3n-2$ θέσεις μνήμης για την καταχώρηση των στοιχείων του.

Ο συμβολισμός $a \equiv b$ θα σημαίνει ότι το $b$ επικαλύπτει το $a$.

### 1.3.1 Ανάλυση αποτελεσματικότητας και μνήμης για ορισμένους βασικούς αλγόριθμους

Παράδειγμα 1: **Υπολογισμός εσωτερικού γινομένου**
Εστω $x, y \epsilon \Re^n$, το εσωτερικό τους γινόμενο

$$z = x^t \cdot y = \sum_{i=1}^{n} x_i \cdot y_i$$

υπολογίζεται σύμφωνα με τον παρακάτω αλγόριθμο:

$z = 0$
**for** $i = 1, 2, \ldots, n$
    $z = z + x_i \cdot y_i$

Για κάθε $i$ απαιτείται 1 flop, έτσι ο αλγόριθμος για την εκτέλεσή του χρειάζεται $n$ flops.

| | | |
|---|---|---|
| Αποτελεσματικότητα: | $O(n)$ | flops |
| Μνήμη : | $2n+1$ | θέσεις |

□

Παράδειγμα 2: **Υπολογισμός γινομένου άνω τριγωνικών πινάκων**.
Εστω $U = (u_{ij})$ και $V = (v_{ij})$ δύο άνω τριγωνικοί πίνακες τάξης $n$.
Ο ακόλουθος αλγόριθμος υπολογίζει το γινόμενο $C = U \cdot V$, επικαλύπτοντας τον πίνακα $U$ με το γινόμενο $U \cdot V$.

**for** $i = 1, 2, ..., n$
    **for** $j = i, i+1, ..., n$
        $u_{ij} \equiv c_{ij} = \sum_{k=i}^{j} u_{ik} \cdot v_{kj}$

Υπολογισμός των flops:

1. Ο υπολογισμός των $c_{ij}$ απαιτεί $j - i + 1$ flops
2. Αφού το $j$ τρέχει από το $i$ έως $n$ και το $i$ από 1 έως $n$, ο συνολικός αριθμός των flops είναι:

$$\sum_{i=1}^{n}\sum_{j=i}^{n}(j-i+1) = \sum_{i=1}^{n}(1+2+\ldots+(n-i+1)) =$$

$$= \sum_{i=1}^{n}\frac{(\mathrm{n}-\mathrm{i}+1)\cdot(\mathrm{n}-\mathrm{i}+2)}{2} \cong \frac{\mathrm{n}^3}{6}$$

για μεγάλα $n$.

Να σημειώσουμε ότι $1+2+3+\ldots+r = \frac{\mathrm{r}\cdot(\mathrm{r}+1)}{2}$ και

$$1^2+2^2+3^2+\ldots+r^2 = \frac{\mathrm{r}\cdot(\mathrm{r}+1)\cdot(2\cdot\mathrm{r}+1)}{6}$$

Επομένως για μεγάλα $n$, το γινόμενο δύο τριγωνικών πινάκων απαιτεί περίπου $\frac{\mathrm{n}^3}{6}$ flops.

| | | |
|---|---|---|
| Αποτελεσματικότητα: | $\mathrm{O}(\frac{\mathrm{n}^3}{6})$ | flops |
| Μνήμη : | $n(n+1)$ | θέσεις |

□

Παράδειγμα 3: **Υπολογισμός γινομένου πινάκων**.
Εστω $A$ ένας $m \times n$ πίνακας και $B$ ένας $n \times p$ πίνακας. Ο ακόλουθος αλγόριθμος υπολογίζει το γινόμενο $C = A \cdot B$.

**for** $i = 1, 2, \ldots, m$
  **for** $j = 1, 2, \ldots, p$
    $c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$

Για τον υπολογισμό κάθε $c_{ij}$ απαιτούνται $n$ πολλαπλασιασμοί. Αφού το $j$ τρέχει από 1 έως $p$ και το $i$ από 1 έως $m$, ο συνολικός αριθμός πολλαπλασιασμών που απαιτούνται είναι $mnp$.

| | | |
|---|---|---|
| Αποτελεσματικότητα: | $O(mnp)$ | flops |
| Μνήμη : | $mn + np + mp$ | θέσεις |

Εάν έχουμε δύο τετραγωνικούς πίνακες τάξης $n$ τότε απαιτούνται συνολικά $n^3$ flops. Οσον αφορά τις απαιτούμενες θέσεις μνήμης, κάθε πίνακας χρειάζεται $n^2$ θέσεις. Για οικονομία στις απαιτούμενες θέσεις μνήμης μπορούμε να επικαλύπτουμε το $B$ με το γινόμενο $AB$, υποθέτοντας ότι μία επιπλέον στήλη έχει προστεθεί στο $B$ και η οποία θα χρησιμεύσει σαν προσωρινό διάνυσμα εργασίας. Ετσι ο παρακάτω αλγόριθμος μπορεί να χρησιμοποιηθεί για τον υπολογισμό του γινομένου δύο $n \times n$ πινάκων με οικονομία στην απαιτούμενη μνήμη.

**for** $j = 1, 2, \ldots, n$
  $h_i = \sum_{k=1}^{n} a_{ik} b_{kj},\ \ i = 1, 2, \ldots, n$
  $b_{ij} \equiv h_i,\ \ i = 1, 2, \ldots, n$

| | | |
|---|---|---|
| Αποτελεσματικότητα: | $O(n^3)$ | flops |
| Μνήμη : | $2n^2 + n$ | θέσεις |

□

Παράδειγμα 4: **Υπολογισμός του πίνακα** $H = (I - \frac{2 \cdot \underline{u} \cdot \underline{u}^{\mathrm{T}}}{\underline{u}^{\mathrm{T}} \cdot \underline{u}}) \cdot A$
Πολύ συχνά στην Αριθμητική Γραμμική Αλγεβρα χρειάζεται να υπολογίσουμε τον πίνακα $(I - \frac{2\underline{u} \cdot \underline{u}^{\mathrm{T}}}{\underline{u}^{\mathrm{T}} \cdot \underline{u}}) \cdot A$ , όπου I ένας $m \times m$ μοναδιαίος πίνακας, $\underline{u} \epsilon \mathbb{R}^m$ και $A$ ένας $m \times n$ πίνακας.

1ος τρόπος: Εάν υπολογίσουμε απ' ευθείας τον $m \times m$ πίνακα $H = (I - \frac{2\underline{u} \cdot \underline{u}^{\mathrm{T}}}{\underline{u}^{\mathrm{T}} \cdot \underline{u}})$ από το διάνυσμα $\underline{u}$ και πάρουμε το γινόμενο του με τον πίνακα $A$ τότε απαιτούνται $O(m^2 \cdot n)$ flops.

| | | |
|---|---|---|
| Αποτελεσματικότητα: | $O(m^2 n)$ | flops |
| Μνήμη : | $m^2 + mn + m$ | θέσεις |

2ος τρόπος: Ο υπολογισμός μπορεί να γίνει με πολύ λιγότερα flops εάν δεν υπολογίσουμε αναλυτικά τον $H$. Συγκεκριμένα μπορούμε να χρησιμοποιήσουμε την ακόλουθη διαδικασία. Εάν ονομάσουμε $\beta = \frac{2}{\underline{u}^{\mathrm{T}} \cdot \underline{u}}$, τότε η $(i, j)$ είσοδος του πίνακα $(A - \beta \cdot \underline{u} \cdot \underline{u}^T \cdot A)$ ισούται με $a_{ij} - \beta \cdot (u_1 \cdot a_{1j} + u_2 \cdot a_{2j} + \ldots + u_m \cdot a_{mj}) \cdot u_i$.
Ετσι προκύπτει ο επόμενος αλγόριθμος:

$\beta = \frac{2}{\underline{u}^{\mathrm{T}} \cdot \underline{u}}$
**for** $j = 1, 2, \ldots, n$
  $a = u_1 \cdot a_{1,j} + u_2 \cdot a_{2,j} + \ldots + u_m \cdot a_{m,j}$
  $a \equiv \beta \cdot a$
  **for** $i = 1, 2, \ldots, m$
    $a_{ij} \equiv a_{ij} - a \cdot u_i$

Για τον υπολογισμό της ποσότητας $\beta$ απαιτούνται $m + 1$ flops ($m$ flops για τον υπολογισμό του εσωτερικού γινομένου και 1 flop για τη διαίρεση του 2 με το εσωτερικό γινόμενο). Αφού υπολογίζουμε $n$ το πλήθος $a$ και καθένα χρειάζεται $(m + 1)$ flops, τελικά απαιτούνται $n \cdot (m+1)$ flops για τον υπολογισμό όλων των $a$. Τέλος κάθε είσοδος $a_{ij}$ του πίνακα απαιτεί 1 flop οπότε χρειάζονται άλλα $m \cdot n$ flops για τις εισόδους του πίνακα. Συνολικά απαιτούνται $(m+1) + n \cdot (m+1) + n \cdot m = (m+1) + 2m \cdot n + n$ flops. Ιδιαίτερα εάν $m = n$ τότε χρειάζονται περίπου $2 \cdot n^2$ flops αισθητά λιγότερα από τα $n^3$ που απαιτεί ο πρώτος τρόπος.

# 1.4 ΑΣΚΗΣΕΙΣ

**1.** Η δυαδική παράσταση του αριθμού $x = \frac{2}{3}$ είναι η:

$$x = \frac{2}{3} = (.101010\ldots)_2$$

Ποιοί είναι οι δύο κοντινότεροι αριθμοί $x', x''$ για ένα συγκεκριμένο τύπο υπολογιστή με μήκος λέξης 32 bits, $\beta = 2$, mantissa 23 bits.
Ποιον επιλέγουμε για $fl(x)$· Εξετάστε το απόλυτο και σχετικό σφάλμα.

**Απόδειξη:** Οι δύο κοντινότεροι με 24 bits είναι:

$$x' = (.101010\ldots1010)_2$$

$$x'' = (.101010\ldots1011)_2.$$

Το $x'$ **ευρίσκεται με αποκοπή** ενώ το $x''$ **με στρογγύλευση**. Για να καθορίσουμε ποιος είναι κοντύτερα υπολογίζουμε τα σφάλματα:

$$x - x' = (.101010...)_2 \cdot 2^{-24} = \frac{2}{3} \cdot 2^{-24}$$

$$x'' - x = (x'' - x') - (x - x') = 2^{-24} - \frac{2}{3} \cdot 2^{-24} = \frac{1}{3} * 2^{-24}$$

Επομένως επιλέγουμε $fl(x) = x''$

Απόλυτο σφάλμα: $|fl(x) - x| = \frac{1}{3} \cdot 2^{-24}$.

Σχετικό Σφάλμα: $|\frac{\mathrm{fl(x)-x}}{\mathrm{x}}| = \frac{\frac{1}{3} \cdot 2^{-24}}{\frac{2}{3}} = 2^{-25} < 2^{-24} = u$

□

**2.** Προσδιορίστε το απόλυτο και σχετικό σφάλμα για τις ακόλουθες περιπτώσεις:

(i) $x = 3.141592,\ \hat{x} = 3.14$

(ii) $x = 1000000,\ \hat{x} = 999996$

(iii) $x = 0.000012,\ \hat{x} = 0.000009$

**Απόδειξη:**

(i) Έχουμε

$$|\hat{x} - x| = 0.001592$$

Και

$$\frac{|\hat{x} - x|}{|x|} \cong 0.000507 \approx \frac{1}{2}10^{-3} \Rightarrow 3 significant digits$$

Σ' αυτή τη περίπτωση τα σφάλματα δε διαφέρουν σημαντικά.

(ii) Έχουμε

$$|\hat{x} - x| = 4$$

Και

$$\frac{|\hat{x} - x|}{|x|} = 0.000004 < \frac{1}{2}10^{-5} \Rightarrow 5 significant digits$$

Η τιμή του σχετικού σφάλματος είναι μικρή ενώ η τιμή του απόλυτου σφάλματος είναι μεγάλη. Το $\hat{x}$ είναι καλή προσέγγιση του $x$.

(ii) Έχουμε

$$|\hat{x} - x| = 0.000003$$

Και

$$\frac{|\hat{x} - x|}{|x|} = 0.25 < \frac{1}{2}10^{0} \Rightarrow 0 significant digits$$

Παρόλο που το απόλυτο σφάλμα είναι μικρό το σχετικό σφάλμα είναι είναι αρκετά μεγάλο με αποτέλεσμα το $\hat{x}$ να μην αποτελεί καλή προσέγγιση του $x$.

□

**Παρατήρηση:** Όσο το $x$ απομακρύνεται από το 1 (είτε προς τα πάνω είτε προς τα κάτω) το σχετικό σφάλμα αποτελεί καλλίτερο κριτήριο για την ακρίβεια της τελικής προσέγγγισης.

**3.** Να υπολογιστεί η ποσότητα $c = \sqrt{a^2 + b^2}$ όταν $a = 10^{60},\ b = 1$ και $t = 4$.

**Απόδειξη:** Έχουμε

$$a^2 = (10^{60})^2 = 10^{120} = 0.1 \cdot 10^{121}$$

$$b^2 = 1^2 = 1 = 0.0 \ldots 01 \cdot 10^{121}$$

Άρα

$$a^2 + b^2 = 0.1 \cdot 10^{121} = 10^{120}$$

(Η πρόσθεση της μονάδας έχει αγνοηθει).

Εάν ο εκθέτης είχε *μόνο δύο* ψηφία τότε θα είχαμε πρόκληση υπερχείλισης $(a^2 = (10^{60})^2)$.

Μπορούμε να υπολογίσουμε το $c$ εναλλακτικά κάνοντας scaling ως εξής:

$$c = s\sqrt{\left(\frac{a}{s}\right)^2 + \left(\frac{b}{s}\right)^2}$$

όπου $s = \max\{|a|, |b|\}$.

Έτσι

$$s = 10^{60}$$

$$c = 10^{60}\sqrt{1^2 + \left(\frac{1}{10^{60}}\right)^2}$$

Το $\left(\frac{1}{10^{60}}\right)^2$ προκαλεί υποχείλιση και τίθεται ίσο με μηδέν. Αυτό δεν πειράζει και τόσο γιατί το $10^{-120}$ που θα αγνοηθεί είναι πολύ μικρό και όχι τόσο σημαντικό συγκριτικά με τη μονάδα που προστίθεται. Τελικά $c = 10^{60}$.

□

**4.**Έστω $\beta = 10, t = 3, u = 0.005$
$x = 0.101 * 10^2, y = -0.994 * 10^1$ (**Θεωρητικά:** $x + y = 0.16$)
Να υπολογιστεί το $fl(x+y)$ χρησιμοποιώντας επεξεργαστή διπλής ακριβείας (double precision accumulator).

**Απόδειξη:**

**Βήμα 1:**

$$\text{Ευθυγράμμιση στους εκθέτες} \begin{cases} x = 0.101000 * 10^2 \\ y = -0.099400 * 10^2 \end{cases}$$

**Βήμα 2:** Εκτελούμε την πρόσθεση (εσωτερικά κρατάμε ψηφία 6 θέσεων)

$$x + y = 0.001600 * 10^2$$

**Βήμα 3:** Κανονικοποιούμε

$$fl(x + y) = 0.160 * 10^0$$

Τελικά $fl(x + y) = (x + y)(1 + \delta),\ \ \delta = 0$

Εάν δε χρησιμοποιηθεί επεξεργαστής διπλής ακριβείας η πράξη θα εκτελεσθεί ως εξής:

**Βήμα 1:** Υποθέτουμε ότι τα επιπλέον ψηφία που δε χωράνε στη mantissa, απλώς διαγράφονται.

$$\begin{cases} x = 0.101 * 10^2 \\ y = -0.099 * 10^2 \end{cases}$$

**Βήμα 2:**

$$fl(x + y) = 0.002 * 10^2$$

**Βήμα 3:** Κανονικοποίηση $fl(x + y) = 0.2$
Τελικά

$$fl(x + y) = (x + y)(1 + \sigma)\ \ \text{όπου}\ \sigma = 0.25 = 50u$$

**Συμπέρασμα:** Πάντα χρησιμοποιείται επεξεργαστής διπλής ακριβείας. □

**5.** Εστω $t = 5$ ψηφία και θεωρούμε δύο αριθμούς μηχανής $x = .31426 * 10^3$, $y = .92577 * 10^5$. Να υπολογισθούν οι ποσότητες $fl(x+y)$, $fl(x-y)$, $fl(xy)$, $fl(x/y)$.

**Απόδειξη:** Χρησιμοποιώντας **επεξεργαστή διπλής ακριβείας** για τα ενδιάμεσα αποτελέσματα έχουμε:

$$x * y = .2909324802 * 10^8$$

$$x + y = .9289126 * 10^5$$

$$x - y = -.92262740 * 10^5$$

$$x/y = .3394579647 * 10^{-2}$$

Μετά τη στρογγύλευση στις 5 θέσεις έχουμε:

$$fl(xy) = .29093 * 10^8, \;\; Rel = 8.5 * 10^{-6}$$

$$fl(x+y) = .92891 * 10^5, \;\; Rel = 2.3 * 10^{-6}$$

$$fl(x-y) = -.92263 * 10^5, \;\; Rel = 2.8 * 10^{-6}$$

$$fl(x/y) = .33946 * 10^{-2}, \;\; Rel = 6.0 * 10^{-6}$$

(Τα Rel είναι όλα μικρότερα από $10^{-5}$, $u = \frac{1}{2}10^{1-5} = .5 * 10^{-4}$) □

**Άσκηση 6**

**α)** Προσδιορίστε παραδείγματα αριθμών μηχανής για τους οποίους δεν ισχύει η προσεταιριστική και επιμεριστική ιδιότητα.

-Προσεταιριστική ιδιότητα της πρόσθεσης:
Έστω το σύστημα $M(10, 4, m, M)$
(γενικά, σε αυτά τα παραδείγματα, τον μεγαλύτερο ρόλο τον παίζει το $t$ της μαντισσα, οπότε απλά θεωρούμε $m$ και $M$ τέτοια ώστε να περιλαμβάνονται στο σύστημα οι εκθέτες που εμφανίζονται στις πράξεις)
Από τις σημειώσεις του μαθήματος, υπάρχουν τα παραδείγματα αριθμών μηχανής του Ωιλκινσον για τους οποίους δεν ισχύει η προσεταιριστική ιδιότητα της πρόσθεσης:
$x_1 = 0.4462 \times 10^{-3}, x_2 = 0.6412 \times 10^{-3}, x_3 = 0.2413 \times 10^{-3}, x_4 = 0.1234 \times 10^0$ για τους οποίους ισχύει
$fl(((x_1 + x_2) + x_3) + x_4) = 0.1247 \times 10^0$ και $fl(((x_4 + x_3) + x_2) + x_1) = 0.1246 \times 10^0$

και
$x_1 = 0.1025 \times 10^4, x_2 = 0.9123 \times 10^3, x_3 = 0.9663 \times 10^2, x_4 = 0.9315 \times 10$ για τους οποίους ισχύει
$fl(((x_1 + x_2) + x_3) + x_4) = 0.6755 \times 10$ και $fl(((x_4 + x_3) + x_2) + x_1) = 0.7 \times 10$

Άλλο παράδειγμα αριθμών για τους οποίους δεν ισχύει η προσεταιριστική ιδιότητα της πρόσθεσης:
$x_1 = 0.9412 \times 10^{-3}, x_2 = 0.9325 \times 10^{-3}, x_3 = 0.8167 \times 10^0$
-$fl((x_1 + x_2) + x_3)$ :

$$\begin{array}{rl} 0.9412 & \times 10^{-3} \\ +\ 0.9325 & \times 10^{-3} \\ \hline 1.8737 & \times 10^{-3} \end{array}$$

$\rightarrow$ (κανονικοποίηση) $0.1873|7 \times 10^{-2} \rightarrow$ (στρογγύλευση) $0.1874 \times 10^{-2}$

ευθυγράμμιση εκθετών: $0.1874 \times 10^{-2} \rightarrow 0.001874 \times 10^0$

$$\begin{array}{rl} 0.001874 & \times 10^0 \\ +\ 0.816700 & \times 10^0 \\ \hline 0.8185|74 & \times 10^0 \end{array}$$

$\rightarrow$ (στρογγύλευση) $0.8186 \times 10^0$

Άρα $fl((x_1 + x_2) + x_3) = 0.8186 \times 10^0$

-$fl(x_1 + (x_2 + x_3))$ :
ευθυγράμμιση εκθετών: $0.9325 \times 10^{-3} \rightarrow 0.0009325 \times 10^0$

$$\begin{array}{rl} 0.0009325 & \times 10^0 \\ +\ 0.8167000 & \times 10^0 \\ \hline 0.8176|325 & \times 10^0 \end{array}$$

$\rightarrow$ (στρογγύλευση) $0.8176 \times 10^0$
ευθυγράμμιση εκθετών: $0.9412 \times 10^{-3} \rightarrow 0.0009412 \times 10^0$

$$\begin{array}{rl} 0.8176000 & \times 10^0 \\ +\ 0.0009412 & \times 10^0 \\ \hline 0.8185|412 & \times 10^0 \end{array}$$

$\rightarrow$ (στρογγύλευση) $0.8185 \times 10^0$

Άρα $fl(x_1 + (x_2 + x_3)) = 0.8185 \times 10^0$
και βλέπουμε ότι τα δύο αποτελέσματα είναι διαφορετικά.

Στην ίδια λογική με το προηγούμενο παράδειγμα, για $M(b, t, m, M)$ τέτοιο ώστε $|m| \geq t - 1$ παίρνουμε κατάλληλους αριθμούς:
$x_1 = 0.a_1a_2 \ldots a_t \times 10^{-(t-1)}$

$x_2 = 0.b_1b_2 \dots b_t \times 10^{-(t-1)}$
$x_3 = 0.c_1c_2 \dots c_t \times 10^0$
έτσι ώστε να μην ισχύει η προσεταιριστική ιδιότητα, π.χ. θα επιλέγουμε τα $a_1 \dots a_t, b_1 \dots b_t$;μεγάλα; ώστε:
-στον υπολογισμό $fl((x_1 + x_2) + x_3)$ στην πρόσθεση $fl(x_1 + x_2)$ οι εκθέτες θα είναι ίδιοι αλλά θα χρειάζεται να γίνει στρογγύλευση και έπειτα στην $fl(fl(x_1 + x_2) + x_3)$ θα πρέπει να γίνει ευθυγράμμιση εκθετών και θα χρειαστεί πάλι στρογγύλευση λόγω του ότι ο ένας εκθέτης είναι -3 και ο άλλος 0 και τα ψηφία της μαντισσα 4
-στον υπολογισμό $fl(x_1 + (x_2 + x_3))$ στην πρόσθεση $fl(x_2 + x_3)$ θα χάνεται λόγω στρογγύλευσης μεγάλο μέρος του αριθμού, λόγω της διαφοράς των εκθετών -3 και 0 (θα υπάρχουν λεαδινγ ζερος), και έπειτα στην $fl(x_1 + fl(x_2 + x_3))$ θα γίνεται πάλι το ίδιο, οπότε θα χάνουμε αρκετά μεγάλο μέρος του αριθμού που θέλουμε να υπολογίσουμε. Δηλαδή στον υπολογισμό $(x_1 + (x_2 + x_3))$ χάνουμε εν γένει μεγαλύτερο μέρος του αποτελέσματος λόγω στρογγυλεύσεων (έχουμε πιο κακή προσέγγιση του αποτελέσματος) απ' ότι στον $((x_1 + x_2) + x_3)$, άρα είναι πολύ πιθανό τα δύο αποτελέσματα να βγουν διαφορετικά

-Προσεταιριστική ιδιότητα του πολλαπλασιασμού:
Για το $M(10, 3, m, M)$ παίρνοντας τους αριθμούς μηχανής:
$x_1 = 0.222 \times 10^0, x_2 = 0.333 \times 10^0, x_3 = 0.444 \times 10^0$, κάνουμε τους υπολογισμούς
-$fl((x_1x_2)x_3)$ :

$$\begin{array}{rl} 0.222 & \times 10^0 \\ \times\ 0.333 & \times 10^0 \\ \hline 0.073926 & \times 10^0 \end{array}$$

$\rightarrow$ (κανονικοποίηση) $0.739|26 \times 10^{-1} \rightarrow$ (στρογγύλευση) $0.739 \times 10^{-1}$

$$\begin{array}{rl} 0.739 & \times 10^{-1} \\ \times\ 0.444 & \times 10^0 \\ \hline 0.328|116 & \times 10^{-1} \end{array}$$

$\rightarrow$ (στρογγύλευση) $0.328 \times 10^{-1}$

Άρα $fl((x_1x_2)x_3) = 0.328 \times 10^{-1}$

-$fl(x_1(x_2x_3))$ :

$$\begin{array}{rl} 0.333 & \times 10^0 \\ \times\ 0.444 & \times 10^0 \\ \hline 0.147|852 & \times 10^0 \end{array}$$

$\rightarrow$ (στρογγύλευση) $0.148 \times 10^0$

$$\begin{array}{rl} 0.148 & \times 10^0 \\ \times\ 0.222 & \times 10^0 \\ \hline 0.032856 & \times 10^0 \end{array}$$

$\rightarrow$ (κανονικοποίηση) $0.328|56 \times 10^{-1} \rightarrow$ (στρογγύλευση) $0.329 \times 10^{-1}$

Άρα $fl(x_1(x_2x_3)) = 0.329 \times 10^{-1}$
και τα δύο αποτελέσματα είναι διαφορετικά.

-Επιμεριστική ιδιότητα:
Για το $M(10, 3, m, M)$
παίρνοντας τους αριθμούς μηχανής:
$x_1 = 0.55 \times 10^0, x_2 = 0.55 \times 10^0, x_3 = 0.45 \times 10^0$, κάνουμε τους υπολογισμούς
-$fl(x_1(x_2 + x_3))$ :

$$fl(x_2 + x_3): \begin{array}{rll} 0.55 & \times 10^0 & \\ +\ 0.45 & \times 10^0 & \\ \hline 1.00 & \times 10^0 & \end{array} \rightarrow \text{(κανονικοποίηση) } 0.1 \times 10^1$$

$$fl(x_1 fl(x_2 + x_3)): \begin{array}{rll} 0.1 & \times 10^1 & \\ \times\ 0.55 & \times 10^0 & \\ \hline 0.055 & \times 10^1 & \end{array} \rightarrow \text{(κανονικοποίηση) } 0.55 \times 10^0$$

Άρα $fl(x_1(x_2 + x_3)) = 0.55 \times 10^0$

-$fl(x_1x_2 + x_1x_3)$ :

$$fl(x_1x_2): \begin{array}{rll} 0.55 & \times 10^0 & \\ \times\ 0.55 & \times 10^0 & \\ \hline 0.302|5 & \times 10^0 & \end{array} \rightarrow \text{(στρογγύλευση) } 0.303 \times 10^0$$

$$fl(x_1x_3): \begin{array}{rll} 0.55 & \times 10^0 & \\ \times\ 0.45 & \times 10^0 & \\ \hline 0.247|5 & \times 10^0 & \end{array} \rightarrow \text{(στρογγύλευση) } 0.248 \times 10^0$$

$$fl(fl(x_1x_2) + fl(x_1x_3)): \begin{array}{rll} 0.303 & \times 10^0 & \\ +\ 0.248 & \times 10^0 & \\ \hline 0.551 & \times 10^0 & \end{array}$$

Άρα $fl(fl(x_1x_2) + fl(x_1x_3)) = 0.551 \times 10^0$
Βλέπουμε ότι τα δύο αποτελέσματα διαφέρουν, άρα δεν ισχύει η επιμεριστική ιδιότητα.

Στην ίδια λογική, για σύστημα $M(b, t, m, M)$ μπορούμε να βρούμε κατάλληλα
$x_1 = 0.a_1a_2 \ldots a_t \times 10^0$
$x_2 = x_1$
$x_3 = 0.c_1c_2 \ldots c_t \times 10^0$ ώστε $fl(x_2 + x_3) = 1$
τέτοια ώστε:
-στον υπολογισμό $x_1(x_2 + x_3)$ να έχουμε αποτέλεσμα $x_1$
-στον υπολογισμό $x_1x_2 + x_1x_3$: -στον πολλαπλασιασμό $x_1x_2$ ή/και στον $x_1x_3$ να προκύπτουν περισσότερα από τ ψηφία στην μαντισσα (οπότε θα έχουμε στρογγύλευση) άρα μεγαλύτερη πιθανότητα το αποτέλεσμα να βγει διαφορετικό.

Άλλο παράδειγμα αριθμών για τους οποίους δεν ισχύει η επιμεριστική:

Οι αριθμοί $0.9412 \times 10^{-3}, 0.9325 \times 10^{-3}, 0.8167 \times 10^{0}$ (από το παράδειγμα της προσεταιριστικής ιδιότητας στην πρόσθεση) στο $M(10, 4, m, M)$:
Κάνοντας τους υπολογισμούς έχουμε
$fl(x_1(x_2 + x_3)) = 0.7695 \times 10^{-3}$ και
$fl(x_1x_2 + x_1x_3) = 0.7696 \times 10^{-3}$
Επομένως εδώ έχουμε αριθμούς για τους οποίους δεν ισχύει ούτε η προσεταιριστική ιδιότητα στην πρόσθεση, ούτε η επιμεριστική ιδιότητα.

**β)** Δίνεται το σύστημα $M(10, 4, -4, 4)$. Να προσδιοριστούν οι αρνητικοί αριθμοί μηχανής που ανήκουν σ' αυτό και να παρασταθούν γραφικά.

Οι αρνητικοί αριθμοί του συστήματος είναι $(b-1)b^{t-1}(M?m+1) = 9 \cdot 10^3 \cdot 9 = 81000$ το πλήθος και θα είναι της μορφής $s(0.a_1a_2a_3a_4)10^e$ όπου:
$s =$ το πρόσημο (εδώ αρνητικό)
$a_1 = 1, \ldots, 9$
$a_2, a_3, a_4 = 0, \ldots, 9$
$e = -4, \ldots, 4$
Για κάθε $e = -4, \ldots, 4$, υπάρχουν $(b-1)b^{t-1} = 9 \cdot 10^3 = 9000$ αριθμοί μηχανής ισοκατανεμημένοι στο διάστημα $(-b^e, -b^{e-1}] = (-10^e, -10^{e?1}]$ και το βήμα κατανομής είναι $b^{e-t} = 10^{e-4}$. Επίσης $s = b^{-4-1} = 10^{-5}$. Οπότε έχουμε:

| $e$ | διάστημα | βήμα κατανομής | αριθμοί μηχανής (τύπος) |
|---|---|---|---|
| -4 | $(-10^{-4}, -10^{-5}]$ | $10^{-8}$ | $-10^{-5} - l \cdot 10^{-8}$, $l = 0, \ldots, 8999$ |
| -3 | $(-10^{-3}, -10^{-4}]$ | $10^{-7}$ | $-10^{-4} - l \cdot 10^{-7}$, $l = 0, \ldots, 8999$ |
| -2 | $(-10^{-2}, -10^{-3}]$ | $10^{-6}$ | $-10^{-3} - l \cdot 10^{-6}$, $l = 0, \ldots, 8999$ |
| -1 | $(-10^{-1}, -10^{-2}]$ | $10^{-5}$ | $-10^{-2} - l \cdot 10^{-5}$, $l = 0, \ldots, 8999$ |
| 0 | $(-1, -10^{-1}]$ | $10^{-4}$ | $-10^{-1} - l \cdot 10^{-4}$, $l = 0, \ldots, 8999$ |
| 1 | $(-10, -1]$ | $10^{-3}$ | $-1 - l \cdot 10^{-3}$, $l = 0, \ldots, 8999$ |
| 2 | $(-10^2, -10]$ | $10^{-2}$ | $-10 - l \cdot 10^{-2}$, $l = 0, \ldots, 8999$ |
| 3 | $(-10^3, -10^2]$ | $10^{-1}$ | $-10^2 - l \cdot 10^{-1}$, $l = 0, \ldots, 8999$ |
| 4 | $(-10^4, -10^3]$ | 1 | $-10^3 - l$, $l = 0, \ldots, 8999$ |

# Κεφάλαιο 2

# Θεωρία Ανάλυσης Σφάλματος

Είναι πολύ σημαντικό εκτός από τον αλγόριθμο κάθε αριθμητικής μεθόδου να εξετάζουμε και την ανάλυση του σφάλματος, που περιέχεται στο αποτέλεσμα που υπολογίζουμε. Μία μέθοδος στην οποία δε γίνεται καμμία ανάλυση σχετικά με το σφάλμα που δίνει, δε θεωρείται καθόλου ικανοποιητική.
Η θεωρία ανάλυσης σφάλματος βασικό σκοπό έχει, όχι μόνο να δώσει κάποια «καλά» φράγματα για το σφάλμα αλλά και να μας εφοδιάσει με χρήσιμα συμπεράσματα τα οποία μπορούν να μας βοηθήσουν στην όλη βελτίωση του αλγορίθμου.
Θεμελιωτής της θεωρίας ανάλυσης σφάλματος (error analysis) είναι ο J. Wilkinson, του οποίου η προσφορά σ' αυτόν τον κλάδο είναι τεράστια. Μέχρι δε σήμερα το βιβλίο του 'Rounding Errors in Algebraic Processes' παραμένει κλασσικό και αναντικατάστατο.
Στις επόμενες παραγράφους, θα παρουσιάσουμε ορισμένα βασικά κομμάτια από αυτόν τον πολύ ενδιαφέροντα κλάδο της ανάλυσης σφάλματος.

## 2.1 Μορφές της Ανάλυσης Σφάλματος

Δύο βασικές μορφές ανάλυσης σφάλματος χρησιμοποιούνται και αναφέρονται σαν **forward error analysis** και **backward error analysis**.
Ας αναπτύξουμε πρώτα την forward error analysis. Μπορούμε να θεωρήσουμε ότι κάθε υπολογισμός περιγράφεται από έναν αριθμό μαθηματικών εξισώσεων. Σε κάθε εξίσωση, κάποια καινούργια ποσότητα $x$ ορίζεται συναρτήσει των ποσοτήτων $a_1, a_2, \ldots, a_n$ που έχουν ήδη προηγούμενα υπολογισθεί και ορισμένες από αυτές μπορεί να είναι και αρχικά δεδομένα. Η μαθηματική εξίσωση μπορεί να γραφεί:

$$x = g(a_1, a_2, \ldots, a_n) \tag{2.1}$$

Ο υπολογισμός του $x$ από τα $a_i$ πρέπει να περιέχει μόνο τις βασικές αριθμητικές πράξεις. (Δε λαμβάνουμε υπόψη μας τα truncation errors. Αλλωστε στην Αριθμητική Γραμμική Αλγεβρα η κύρια πηγή σφαλμάτων είναι τα rounding errors) Λόγω της ύπαρξης των σφαλμάτων στρογγύλευσης (rounding errors) η τιμή του $x$ που θα υπολογιστεί θα διαφέρει από αυτή που θα προέκυπτε εάν υπολογιζόταν ακριβώς η ποσότητα $g(a_1, a_2, \ldots, a_n)$.
Στην forward error analysis συμβολίζουμε την υπολογιζόμενη τιμή με $\overline{x}$ και προσπαθούμε να βρούμε ένα φράγμα για την ποσότητα $|\overline{x} - g(a_1, \ldots, a_n)|$.
Στην backward error analysis, δεν ελέγχουμε σε κάθε βήμα υπολογισμών τις διαφορές μεταξύ των τιμών που υπολογίσαμε και των πραγματικών τιμών, αλλά στο τέλος των υπολογισμών αποδεικνύουμε ότι το $x$ που υπολογίσαμε και θεωρητικά δίνεται από την (2.1), στην πραγματικότητα είναι ίσο με $g(a_1 + \varepsilon_1, a_2 + \varepsilon_2, \ldots, a_n + \varepsilon_n)$ για κάποιες τιμές των $\varepsilon_i$ και αυτά τα $\varepsilon_i$ φράσσονται κατάλληλα. Προφανώς αυτά τα $\varepsilon_i$ δεν είναι μοναδικά.
Επειδή στην backward error analysis δεν μας ενδιαφέρει η σύγκριση των $x$ και $\overline{x}$ και ουδέποτε αναφερόμαστε στην πραγματική τιμή του $x$, δεν υπάρχει ανάγκη να χρησιμοποιήσουμε ξεχωριστό σύμβολο για να παραστήσουμε την υπολογιζόμενη τιμή γι' αυτό, αντί του συμβόλου $\overline{x}$ χρησιμοποιούμε ξανά το $x$ για να συμβολίσουμε την υπολογιζόμενη τιμή. Ετσι στη μαθηματική εξίσωση (2.1) που ορίζει το $x$, αντιστοιχεί η εξής 'υπολογιστική' εξίσωση που ορίζει το $x$ που πρόκειται να υπολογιστεί

$$x \equiv g(a_1 + \varepsilon_1, a_2 + \varepsilon_2, \ldots, a_n + \varepsilon_n) \tag{2.2}$$

Η εξίσωση αυτή θα ακολουθείται από ανισότητες που θα ικανοποιούν τα $\varepsilon_i$. Για να τονίσουμε ότι δεν είναι μαθηματική εξίσωση αντί του συμβόλου '$=$' της ισότητας χρησιμοποιούμε το σύμβολο '$\equiv$' της ισοδυναμίας.

**Παράδειγμα 2.1.1** *Έστω το γραμμικό σύστημα $Ax = b$, $A \in \mathbb{R}^{n\times n}$ μη ιδιάζων και έστω και το διαταραγμένο σύστημα $\tilde{A}\underline{\tilde{x}} = \underline{\tilde{b}}$. Η forward error analysis υπολογίζει εκτίμηση του σφάλματος $x - \tilde{x}$ συναρτήσει των ποσοτήτων $A - \tilde{A}$ και $b - \tilde{b}$. Η backward error analysis υπολογίζει τις μεταβολές $\delta A = (\delta_{ij})$ και $\delta\underline{b} = (\delta b_i)$ έτσι ώστε η υπολογιζόμενη λύση του συστήματος $\hat{x}$ να ικανοποιεί το σύστημα*

$$(A + \delta A)\hat{x} = \underline{b} + \delta\underline{b}.$$

□

Στην πράξη βρέθηκε ότι η backward error analysis είναι πολλές φορές πιο εύκολη στη χρήση ειδικά όταν σχετίζεται με floating-point υπολογισμούς. Γιάυτό χρησιμοποιείται περισσότερο απ' ότι η forward error analysis.

Στη συνέχεια θα ασχοληθούμε αναλυτικότερα με την ανάλυση σφάλματος στους floating-point υπολογισμους, χρησιμοποιώντας backward error analysis. Θα χρησιμοποιούμε δε, τύπους της μορφής (2.2).

## 2.2 Σφάλματα Στρογγύλευσης σε Υπολογισμούς

Εστω $a, b$ δύο floating-point αριθμοί, το σύμβολο $\Box$ παριστάνει οποιαδήποτε από τις τέσσερις αριθμητικές πράξεις $\{+, -, *, /\}$. Η πράξη $a\Box b$ χρησιμοποιώντας floating point αριθμητική συμβολίζεται με $fl(a\Box b)$
Με την παραδοχή ότι ο υπολογιστής έχει επεξεργαστή διπλής ακριβείας (double precision accumulator) που μπορεί να αποθηκεύει αριθμούς ψηφίων $2 \cdot t$, διατυπώνουμε εκ νέου το παρακάτω θεώρημα που είναι το πιο θεμελιώδες στη θεωρία σφαλμάτων στρογγύλευσης σε floating-point υπολογισμούς και μας δίνει μια έκφραση του σχετικού σφάλματος που προξενείται από την αντικατάσταση του $x$ με $fl(x)$.

**Θεώρημα 2.2.1** *Εάν $x$ είναι ένας πραγματικός αριθμός μέσα στο εύρος των floating-point αριθμών, τότε εάν χρησιμοποιηθεί αριθμητική στρογγύλευσης.*

$$fl(x) = x \cdot (1 + \varepsilon),\ |\varepsilon| \leq \frac{1}{2} \cdot \beta^{1-t}$$

*όπου $\beta$ είναι η βάση της αριθμητικής που έχει ο υπολογιστής και $t$ ο αριθμός των ψηφίων της λέξης που έχει ο υπολογιστής.*

**Απόδειξη:** Εστω $x > 0$ (η περίπτωση $x < 0$ αποδεικνύεται παρόμοια και η περίπτωση $x = 0$ είναι τετριμμένη).

Εστω $e$ ο μοναδικός ακέραιος για τον οποίο

$$\beta^{e-1} \leq x \leq \beta^e$$

Στο διάστημα $[\beta^{e-1}, \beta^e]$ οι floating-point αριθμοί είναι ομοιόμορφα κατανεμημένοι με βήμα $\beta^{e-t}$. Ο πιο κοντινός στον $x$ είναι ο $fl(x)$ και πρέπει να βρίσκεται σε απόσταση μικρότερη ή ίση από $\frac{1}{2} \cdot \beta^{e-t}$ από τον $x$. Ετσι

$$|fl(x) - x| \leq \frac{1}{2} \cdot \beta^{e-t}$$

Επειδή $\beta^{e-1} \leq x$ προκύπτει

$$\frac{|fl(x) - x|}{|x|} \leq \frac{\frac{1}{2} \cdot \beta^{e-t}}{\beta^{e-1}} = \frac{1}{2} \cdot \beta^{1-t}.$$

Ομως $\varepsilon = \frac{fl(x)-x}{x}$ και συνεπώς το θεώρημα αποδείχτηκε. □

'μεσα παίρνουμε το παρακάτω

**Πόρισμα 2.2.1** *Εάν χρησιμοποιηθεί αριθμητική αποκοπός τότε*

$$fl(x) = x \cdot (1 + \varepsilon),\ |\varepsilon| \leq \beta^{1-t}$$

□

**Ορισμός:** Εστω $u$ η μονάδα του σφάλματος στρογγύλευσης (unit round off) που ορίζεται ως εξής:

$$u = \begin{cases} \frac{1}{2} \cdot \beta^{1-t} & \text{εάν χρησιμοποιείται αριθμητική στρογγύλευσης} \\ \beta^{1-t} & \text{εάν χρησιμοποιείται αριθμητική αποκοπής} \end{cases}$$

Από τα προηγούμενα, το επόμενο θεώρημα είναι προφανές.

**Θεώρημα 2.2.2** *$x_1, x_2$ δύο τυχόντες floating-point αριθμοί, τότε:*

$$fl(x_1 \Box x_2) = (x_1 \Box x_2) * (1 + \varepsilon), |\varepsilon| \leq u$$

**Σημείωση:** Το παραπάνω αποτέλεσμα μπορεί να ερμηνευθεί ως εξής:

- Το **υπολογιζόμενο άθροισμα** δύο αριθμών $x_1, x_2$ είναι το **ακριβές** άθροισμα δύο αριθμών $x_1 \cdot (1+e)$ και $x_2 \cdot (1+\varepsilon)$, για κάποια τιμή του $\varepsilon$ που ικανοποιεί μία ανισότητα της μορφής $|\varepsilon| \leq u$.

  (για την αφαίρεση δύο αριθμών ισχύει παρόμοιο αποτέλεσμα)

- Το **υπολογιζόμενο γινόμενο** δύο αριθμών $x_1, x_2$ είναι το **ακριβές** γινόμενο των αριθμών $x_1 \cdot (1+\varepsilon)$ και $x_2$ ή των αριθμών $x_1$ και $x_2 \cdot (1+\varepsilon)$ ή των αριθμών $x_1 \cdot (1+\varepsilon)^{1/2}$ και $x_2 \cdot (1+\varepsilon)^{1/2}$, όπου $|\varepsilon| \leq u$.

- Το **υπολογιζόμενο πηλίκο** δύο αριθμών $x_1, x_2$ είναι το **ακριβές** πηλίκο των αριθμών $x_1 \cdot (1+\varepsilon)$ και $x_2$ ή των αριθμών $x_1$ και $\frac{x_2}{(1+\varepsilon)}$, όπου $|\varepsilon| \leq u$.

Από το προηγούμενο θεώρημα προκύπτει:

$$\frac{|fl(a \Box b) - (a \Box b)|}{|a \Box b|} \leq u, \;\; a \Box b \neq 0$$

που μας δείχνει ότι το σχετικό σφάλμα (relative error) που συνδέεται με κάθε μία θεμελιώδη αριθμητική πράξη είναι μικρό.

Στη συνέχεια θα εξετάσουμε τις περιπτώσεις όπου έχουμε ακολουθίες αριθμητικών πράξεων στις οποίες παίζει σημαντικό ρόλο η διάδοση των σφαλμάτων (propagation of errors).

### 2.2.1 Υπολογισμός Γινομένου

Θέλουμε να υπολογίσουμε το γινόμενο $p_n$ που ορίζεται ως εξής:

$$p_n = fl(x_1 x_2 \ldots x_n)$$

όπου $x_1, x_2, \ldots, x_n$ είναι floating-point αριθμοί και οι πολ/σμοί εκτελούνται από αριστερά προς δεξιά. Οι ποσότητες $p_r$ ορίζονται αναγωγικά από τις σχέσεις:

$$p_1 = x_1$$

$$p_r = fl(p_{r-1}x_r) \equiv p_{r-1}x_r(1+\varepsilon_r),\ |\varepsilon_r| \leq u$$

Ετσι έχουμε το εξής αποτέλεσμα:

$$p_n \equiv x_1x_2\ldots x_n(1+\varepsilon_2)(1+\varepsilon_3)\ldots(1+\varepsilon_n)$$

όπου $1-u \leq 1+\varepsilon_r \leq 1+u,\ r=2,3,\ldots,n$

Τελικά:

$$fl(x_1x_2\ldots x_n) \equiv x_1x_2...x_n\prod_{i=2}^{n}(1+\varepsilon_i)$$

και εάν θέσουμε $\prod_{i=2}^{n}(1+\varepsilon_i) = 1+E$ τότε

$$fl(x_1x_2\ldots x_n) \equiv x_1x_2\ldots x_n(1+E) \tag{2.1.1}$$

όπου

$$(1-u)^{n-1} \leq 1+E \leq (1+u)^{n-1}$$

**Παρατήρηση:** Ενώ στον υπολογισμό του γινομένου παίζει ρόλο η σειρά με την οποία εκτελούνται οι πολ/σμοί παρατηρούμε ότι τα φράγματα του σφάλματος που δίνονται από την (2.1.1) είναι **ανεξάρτητα** της σειράς με την οποία εκτελούνται οι πολ/σμοί.

**'Ασκηση 2.2.1** *Να υπολογιστεί η εξής παράσταση:*

$$d = fl(x_1x_2\ldots x_m/y_1y_2\ldots y_n)$$

*όπου $x_1x_2\ldots x_m,\ y_1y_2\ldots y_n$ είναι floating point αριθμοί.*
*Τί παρατηρείται για τη σειρά με την οποία εκτελούνται οι πράξεις· Επηρεάζεται το σφάλμα από αυτήν· Προκύπτει διαφορετικό σφάλμα εάν η παράσταση $d$ υπολογιστεί αρχίζοντας την εκτέλεση των πράξεων από δεξιά προς αριστερά·*

**Απόδειξη:** Θέλουμε να υπολογίσουμε το εξής:

$$d = fl(x_1x_2\ldots x_m/y_1y_2\ldots y_n)$$

όπου οι πράξεις εκτελούνται από αριστερά προς τα δεξιά

Παρατηρούμε ότι:

$$d = fl(x_1x_2\ldots x_m/y_1y_2\ldots y_n) = fl[fl(x_1x_2\ldots x_m)/y_1y_2\ldots y_n] =$$

$$fl[fl(x_1x_2\ldots x_m\prod_{i=2}^{m}(1+\varepsilon_i)/y_1)y_2\ldots y_n] =$$

$$fl(\frac{x_1x_2\dots x_m}{y_1}\prod_{i=2}^{m}(1+\varepsilon_i)(1+\varepsilon)y_2\dots y_n) =$$

$$\frac{x_1x_2\dots x_m}{y_1}y_2\dots y_n\prod_{i=2}^{m}(1+\varepsilon_i)(1+\varepsilon)\prod_{i=2}^{n}(1+\varepsilon_i)$$

όπου $(1-u)^{m-1} \le \prod_{i=2}^{m}(1+\varepsilon_i) \le (1+u)^{m-1}$

$$1-u \le 1+\varepsilon \le 1+u$$

$$(1-u)^{n-1} \le \prod_{i=2}^{n}(1+\varepsilon_i) \le (1+u)^{n-1}$$

Εάν θέσουμε $\prod_{i=2}^{m}(1+\varepsilon_i)(1+\varepsilon)\prod_{i=2}^{n}(1+\varepsilon_i) = 1+E$ τότε

$$fl(x_1x_2\dots x_m/y_1y_2\dots y_n) \equiv (x_1x_2\dots x_m/y_1y_2\dots y_n)(1+E)$$

όπου

$$(1-u)^{m+n-1} \le 1+E \le (1+u)^{m+n-1} \tag{2.1.2}$$

Παρατηρούμε ότι η σειρά με την οποία εκτελούνται οι πράξεις δεν επηρεάζει τα φράγματα του σφάλματος όπως φαίνεται από την (2.1.2). Επομένως το ίδιο σφάλμα προκύπτει εάν οι πράξεις εκτελεστούν από δεξιά προς αριστερά. □

### 2.2.2 Υπολογισμός αθροίσματος

Θέλουμε να υπολογίσουμε το άθροισμα $s_n$, που ορίζεται ως εξής:

$$s_n = fl(x_1 + x_2 + \ldots + x_n)$$

όπου $x_1, x_2, \ldots, x_n$ είναι floating point αριθμοί και οι προσθέσεις εκτελούνται από αριστερά προς τα δεξιά. Οι ποσότητες $s_r$ ορίζονται αναγωγικά από τις σχέσεις

$$s_1 = x_1$$

$$s_r = fl(s_{r-1} + x_r) \equiv (s_{r-1} + x_r)(1+\varepsilon_r),\ |\varepsilon_r| \le u$$

Συνδυάζοντας αυτές τις εξισώσεις προκύπτει:

$$s_n \equiv x_1\prod_{i=2}^{n}(1+\varepsilon_i) + \sum_{r=2}^{n}\{x_r\prod_{i=r}^{n}(1+\varepsilon_i)\}$$

όπου $|\varepsilon_i| \le u,\ i = 2, 3, \ldots, n.$

Εστω

$$1 + n_1 = \prod_{i=2}^{n}(1 + \varepsilon_i)$$

$$1 + n_r = \prod_{i=r}^{n}(1 + \varepsilon_i),\ r = 2, 3, \ldots, n$$

Τότε $s_n \equiv x_1(1 + n_1) + x_2(1 + n_2) + \ldots + x_n(1 + n_n)$
όπου

$$(1 - u)^{n-1} \leq 1 + n_1 \leq (1 + u)^{n-1}$$

$$(1 - u)^{n+1-r} \leq 1 + n_r \leq (1 + u)^{n+1-r},\ r = 2, 3, \ldots, n$$

Τις δύο τελευταίες ανισότητες τις συνδυάζουμε σε μία της μορφής:

$$(1 - u)^{n+1-r} \leq 1 + n_r \leq (1 + u)^{n+1-r},\ r = 1, 2, \ldots, n$$

Τελικά

$$fl(x_1 + x_2 + \ldots + x_n) \equiv x_1(1 + n_1) + x_2(1 + n_2) + ... + x_n(1 + n_n) \qquad (2.2.1)$$

όπου $(1 - u)^{n+1-r} \leq 1 + n_r \leq (1 + u)^{n+1-r},\ r = 1, 2, 3, \ldots, n$

**Παρατήρηση:** Παρατηρούμε ότι τα φράγματα των σφαλμάτων όπως δίνονται από την (2.2.1) εξαρτώνται από την σειρά με την οποία εκτελείται η πρόσθεση. Κατά συνέπεια το άνω φράγμα για **το σφάλμα είναι μικρότερο εάν οι προστιθέμενοι αριθμοί είναι διατεταγμένοι κατά αύξουσα σειρά μεγέθους**. Αυτό δικαιολογείται από το γεγονός ότι ο μεγαλύτερος παράγοντας σφάλματος $\prod_{i=2}^{n}(1+\varepsilon_i)$ θα συσχετίζεται με το μικρότερο $x_i$.

### 2.2.3 Υπολογισμός εσωτερικού γινομένου

Θέλουμε να υπολογίσουμε το εσωτερικό γινόμενο $ip_n$, που ορίζεται ως εξής:

$$ip_n = fl(a_1b_1 + a_2b_2 + \ldots + a_nb_n)$$

όπου $a_i$ και $b_i$ είναι floating-point αριθμοί. Πρώτα υπολογίζονται τα γινόμενα και μετά προστίθενται μεταξύ τους από αριστερά προς δεξιά. Οι ποσότητες $ip_r$ και $t_r$ ορίζονται αναγωγικά από τις σχέσεις:

$$t_r = fl(a_rb_r)$$

$$ip_1 = t_1$$

$$ip_r = fl(ip_{r-1} + t_r)$$

Από αυτές τις σχέσεις προκύπτει ότι:

$$t_r \equiv a_r b_r(1+\xi_r),\ |\xi_r| \le u$$

$$ip_r \equiv (ip_{r-1} + t_r)(1+n_r),\ |n_r| \le u$$

Συνεπώς:

$$ip_n \equiv a_1b_1(1+\varepsilon_1) + a_2b_2(1+\varepsilon_2) + \ldots + a_nb_n(1+\varepsilon_n)$$

όπου

$$(1+\varepsilon_1) = (1+\xi_1)(1+n_2)\ldots(1+n_n)$$

$$(1+\varepsilon_r) = (1+\xi_r)(1+n_r)\ldots(1+n_n),\ r = 2,3,\ldots,n$$

και

$$(1-u)^n \le 1+\varepsilon_1 \le (1+u)^n$$

$$(1-u)^{n-r+2} \le 1+\varepsilon_r \le (1+u)^{n-r+2},\ r = 2,3,\ldots,n.$$

Για να μην έχουμε ξεχωριστή ανισότητα για το $1+\varepsilon_1$ ενοποιούμε τις δύο προηγούμενες ανισότητες στην εξής ανισότητα

$$(1-u)^{n-r+2} \le 1+\varepsilon_r \le (1+u)^{n-r+2}, r = 1,2,\ldots,n$$

Επειδή σε όλους τους μέχρι τώρα υπολογισμούς έχουμε συχνή εμφάνιση παραστάσεων της μορφής $\prod_{i=1}^{n}(1+\varepsilon_i),\ |\varepsilon_i| \le u$ , είναι χρήσιμο να μπορούμε να δώσουμε σ΄ αυτές τις παραστάσεις πιο εύχρηστα φράγματα. Σ' αυτό θα μας βοηθήσουν τα ακόλουθα λήμματα από την Ανάλυση.

**Λήμμα 2.2.1** *Εάν* $0 \le u < 1$ *και εάν* $n = 1,2,3\ldots$ *τότε* $1-nu \le (1-u)^n$

**Απόδειξη:** Εστω $f(u) = (1-u)^n$. Θεωρούμε το ανάπτυγμα Taylor $f(u) = f(0) + uf'(0) + \frac{1}{2}f''(ju)u^2,\ \ 0 < j < 1$, $f(u) = 1 - nu + \frac{n(n-1)}{2}(1-ju)^{n-2}u^2$. Επειδή ο τελευταίος όρος στο άθροισμα αυτό είναι θετικός καταλήγουμε ότι : $f(u) \ge 1 - nu$. □

**Λήμμα 2.2.2** *Εάν* $n = 1,2,\ldots$ *και εάν* $0 \le nu < 0.01$ *τότε*

$$(1+u)^n \le 1 + 1.01nu$$

**Απόδειξη:** Γνωρίζουμε ότι:

$$1 + x \leq e^x, \ \forall x \geq 0$$

$$e^x \leq 1 + 1.01x, \ 0 \leq x \leq 0.01$$

Κατά συνέπεια

$$(1 + u)^n \leq e^{nu} \leq 1 + 1.01nu$$

□

**Σημείωση:** Στις εφαρμογές αυτού του λήμματος, $n$ είναι η τάξη ενός πίνακα και $u$ είναι η μονάδα του σφάλματος στρογγύλευσης. Η υπόθεση $nu \leq 0.01$ ικανοποιείται σε όλα τα πρακτικά προβλήματα. Για παράδειγμα, εάν $u = 10^{-15}$ τότε για να μην ισχύει αυτή η ανισότητα θα πρέπει $n > 10^{14}$. Εάν αρχίσουμε να προσθέτουμε αριθμούς σε μία μηχανή με δυνατότητα πρόσθεσης $1\mu sec = 10^{-6}sec$ τότε ο χρόνος που θα χρειασθεί για να προσθέσουμε $10^{14}$ αριθμούς είναι $10^8 sec = 3.2$ χρόνια.

**Λήμμα 2.2.3** *Εάν* $|\varepsilon_i| \leq u$ *για* $i = 1, 2, \ldots, n$ *και εάν* $nu \leq 0.01$ *τότε*

$$1 - nu \leq \prod_{i=1}^{n}(1 + \varepsilon_i) \leq 1 + 1.01nu$$

**Παρατήρηση:** Ορίζουμε το $t_1$ από τη σχέση

$$\frac{1}{2}\beta^{1-t_1} = (1.01)\frac{1}{2}\beta^{1-t}$$

έτσι

$$t_1 = t - log_\beta(1.01)$$

Ορίζουμε

$$u_1 = \frac{1}{2}\beta^{1-t_1}$$

Με τη βοήθεια αυτών των ορισμών μπορούμε να αντικαταστήσουμε ανισότητες της μορφής

$$(1 - u)^n \leq 1 + \varepsilon \leq (1 + u)^n$$

με την απλούστερη ανισότητα

$$|\varepsilon| \leq nu_1$$

Μετά από αυτά τα λήμματα έχουμε το ακόλουθο αποτέλεσμα για τον υπολογισμό του εσωτερικού γινομένου:

$$fl(a_1b_1 + a_2b_2 + \ldots + a_nb_n) \equiv a_1b_1(1 + \varepsilon_1) + a_2b_2(1 + \varepsilon_2) + \ldots + a_nb_n(1 + \varepsilon_n)$$

$$= a_1b_1 + a_2b_2 + \ldots + a_nb_n + a_1b_1\varepsilon_1 + \ldots + a_nb_n\varepsilon_n =$$

$$= a_1b_1 + a_2b_2 + \ldots + a_nb_n + E$$

όπου

$$|E| = |a_1b_1\varepsilon_1 + a_2b_2\varepsilon_2 + \ldots + a_nb_n\varepsilon_n| \leq |a_1||b_1||\varepsilon_1| + |a_2||b_2||\varepsilon_2| + \ldots + |a_n||b_n||\varepsilon_n|$$

Για τα $\varepsilon_r$ ισχύει η γενική ανισότητα

$$(1-u)^{n-r+2} \leq 1 + \varepsilon_r \leq (1+u)^{n-r+2},\ r = 1, 2, \ldots, n$$

η οποία μπορεί να αντικατασταθεί από την ανισότητα

$$|\varepsilon_r| \leq (n-r+2)\cdot u_1,\ r = 1, 2, \ldots, n$$

Ετσι έχουμε τελικά

$$fl(a_1b_1 + a_2b_2 + \ldots + a_nb_n) \equiv a_1b_1 + a_2b_2 + \ldots + a_nb_n + E \tag{2.3.1}$$

όπου

$|E| \leq u_1\{(n+1)|a_1||b_1| + n|a_2||b_2| + (n-1)|a_3||b_3| + \ldots + 3|a_{n-1}||b_{n-1}| + 2|a_n||b_n|\}$

**Παρατήρηση:** Το σχετικό σφάλμα στον υπολογισμό του εσωτερικού γινομένου δίνεται από την σχέση:

$$\frac{|fl(\underline{a}^t\underline{b}) - \underline{a}^t\underline{b}|}{\underline{a}^t\underline{b}} \leq (n+1)\cdot u_1 \cdot \frac{|\underline{a}|^t|\underline{b}|}{|\underline{a}^t\underline{b}|}$$

Παρατηρούμε ότι εάν $|\underline{a}^t\underline{b}| << |\underline{a}|^t|\underline{b}|$, τότε το σχετικό σφάλμα στο $fl(\underline{a}^t\underline{b})$ μπορεί να μην είναι μικρό.

Με τη βοήθεια της ανισότητας Cauchy-Schwartz ($|\langle \underline{x}^t\underline{y}\rangle| \leq \|\underline{x}\|_2\|\underline{y}\|_2$) ισχύει ότι:

$$\frac{|fl(\underline{a}^t\underline{b}) - \underline{a}^t\underline{b}|}{\underline{a}^t\underline{b}} \leq (n+1)\cdot u_1 \cdot \|\underline{a}\|_2\|\underline{b}\|_2.$$

**'Ασκηση 2.2.2** *Να υπολογισθεί το σχετικό σφάλμα που προκύπτει από τον υπολογισμό του γινομένου και του αθροίσματος $n$ αριθμών.*

Για το **γινόμενο** έχουμε τα ακόλουθα από την (2.1.1)

$$fl(x_1x_2\ldots x_n) \equiv x_1x_2\ldots x_n + x_1x_2...x_nE = x_1x_2...x_n + E'$$

όπου

$$|E'| = |x_1||x_2|\ldots|x_n||E|$$

Η ανισότητα $(1-u)^{n-1} \leq 1+E \leq (1+u)^{n-1}$ αντικαθίσταται από την ανισότητα $|E| \leq (n-1)u_1$, οπότε

$$|E'| \leq |x_1||x_2|\ldots|x_n|(n-1)u_1$$

Το σχετικό σφάλμα δίνεται από τον τύπο

$$\frac{|fl(x_1x_2\ldots x_n)-(x_1x_2\ldots x_n)|}{|x_1x_2\ldots x_n|} \leq (n-1)\cdot u_1 \cdot \frac{|x_1||x_2|\ldots|x_n|}{|x_1x_2\ldots x_n|}$$

$$\Rightarrow \frac{|fl(x_1x_2\ldots x_n)-(x_1x_2\ldots x_n)|}{|x_1x_2\ldots x_n|} \leq (n-1)\cdot u_1$$

**Παρατήρηση:** Επειδή πάντοτε το $(n-1)\cdot u_1$ είναι πολύ μικρό, προκύπτει το συμπέρασμα ότι πάντοτε ο υπολογισμός του γινομένου $n$ αριθμών δίνει μικρό σχετικό σφάλμα.

Για το **άθροισμα** έχουμε τα ακόλουθα από την (2.2.1)

$$fl(x_1+x_2+\ldots+x_n) \equiv x_1+x_2+\ldots+x_n+x_1n_1+\ldots+x_nn_n$$

$$= x_1+x_2+\ldots+x_n+E$$

όπου

$$|E| \leq |x_1||n_1|+|x_2||n_2|+\ldots+|x_n||n_n|$$

Η γενική ανισότητα

$$(1-u)^{n+1-r} \leq 1+n_r \leq (1+u)^{n+1-r},\ r=1,2,\ldots,n$$

αντικαθίσταται από την

$$|n_r| \leq (n+1-r)\cdot u_1$$

οπότε

$$|E| \leq u_1(n|x_1|+(n-1)|x_2|+\ldots+2|x_{n-1}|+|x_n|)$$

Συνεπώς το σχετικό σφάλμα δίνεται από τον τύπο:

$$\frac{|fl(x_1+x_2+\ldots+x_n)-(x_1+x_2+\ldots+x_n)|}{|x_1+x_2+\ldots+x_n|} \leq n\cdot u_1 \frac{|x_1|+|x_2|+\ldots+|x_n|}{|x_1+x_2+\ldots+x_n|}$$

**Παρατήρηση 1:** Εάν $|x_1+x_2+\ldots+x_n| << |x_1|+|x_2|+\ldots+|x_n|$ τότε το σχετικό σφάλμα στον υπολογισμό του αθροίσματας $n$ αριθμών μπορεί να μην είναι μικρό.

**Παρατήρηση 2:** Στην πρόσθεση πραγματικών αριθμών αποδείξαμε ότι

$$Rel = \frac{|fl(x_1 + x_2 + \ldots + x_n) - (x_1 + x_2 + \ldots + x_n)|}{|x_1 + x_2 + \ldots + x_n|} \leq$$

$$nu_1 \frac{|x_1| + |x_2| + \ldots + |x_n|}{|x_1 + x_2 + \ldots + x_n|}$$

Η ποσότητα $K = \frac{|x_1|+|x_2|+\ldots+|x_n|}{|x_1+x_2+\ldots+x_n|}$ αποτελεί το δείκτη κατάστασης (condition number) της πρόσθεσης και ανάλογα με την τιμή του το σχετικό σφάλμα μπορεί να παίρνει μικρή ή μεγάλη τιμή.

Όταν όλοι οι όροι του αθροίσματος είναι θετικοί (ή όλοι αρνητικοί) τότε το condition number $k$ είναι μονάδα και ο υπολογισμός του αθροίσματος είναι ευσταθής. Έτσι προκύπτει ότι

$$Rel \leq n \cdot u_1$$

Κατά συνέπεια τα σφάλματα στρογγύλευσης θα συσσωρεύονται αργά καθώς θα προστίθενται οι όροι του αθροίσματος. Σ΄ αυτήν τη περίπτωση το φράγμα $n$ είναι μια υπερεκτίμηση της τιμής που εμφανίζεται στην πράξη αφού τα σφάλματα στρογγύλευσης πότε θα έχουν θετική και πότε αρνητική τιμή οπότε πολλά θα αλληλοαναιρούνται. Στην περίπτωση που η πρόσθεση εκτελείται με αποκοπή τότε τα σφάλματα στρογγύλευσης τείνουν όλα στην ίδια κατεύθυνση (προς τα κάτω) και πιέζει το ένα το άλλο οπότε ο παράγοντας $n$ του φράγματος είναι ρεαλιστικός.

**Παράδειγμα 2.2.1** *Έστω $t = 4$. Να ελεχθούν τα σφάλματα που προκύπτουν κατά την πρόσθεση των παρακάτω αριθμών χρησιμοποιώντας στρογγύλευση και αποκοπή.*

| αριθμός | στρογγύλευση | σφάλμα | αποκοπή | σφάλμα |
|---|---|---|---|---|
| 1374.8 | 1375 | -0.2 | 1374 | 0.8 |
| 3856.4 | 3856 | +0.4 | 3856 | + 0.4 |
| 5231.2 | 5231 | +0.2 | 5230 | +1.2 |

Παρατηρούμε ότι τα σφάλματα που προέκυψαν στην στρογγύλευση έχουν αντίθετα πρόσημα κι έτσι στο τελικό άθροισμα έχουμε το μικρό σφάλμα 0.2. Αντίθετα στην αποκοπή τα σφάλματα έχουν το ίδιο πρόσημο και πιέζουν το ένα το άλλο με αποτέλεσμα να έχουμε το μεγαλύτερο σφάλμα 1.2.

**Συμπέραμα:** Τα σφάλματα από την στρογγύλευση τείνουν να αναιρεί το ένα το άλλο ενώ στην αποκοπή συσσωρεύονται προσθετικά. Γι΄ αυτό πάντοτε προτιμάται η τεχνική της στρογγύλευσης. □

Ολα τα προηγούμενα συνοψίζονται στο παρακάτω πίνακα

| Υπολογισμός | Σχετικό σφάλμα | Παρατηρήσεις |
|---|---|---|
| $fl(x_1 + x_2 + \ldots + x_n)$ | Οχι πάντοτε μικρό εάν $\lvert\sum_{i=1}^{n} x_i\rvert << \sum_{i=1}^{n} \lvert x_i\rvert$ | Μικρότερο σφάλμα εάν οι αριθμοί είναι διατεταγμένοι κατάύξουσα σειρά |
| $fl(x_1 \cdot x_2 \cdot \ldots \cdot x_n)$ | Πάντοτε μικρό | Δεν παίζει ρόλο η σειρά εκτέλεσης των πολ/σμων |
| $fl(\underline{a}^t\underline{b}), \underline{a}, \underline{b}\epsilon\mathbb{R}^n$ | Οχι πάντοτε μικρό εάν $\lvert\underline{a}^t\underline{b}\rvert << \lvert\underline{a}\rvert^t\lvert\underline{b}\rvert$ | Οι πράξεις εκτελούνται ιεραρχικά |

## ΑΣΚΗΣΕΙΣ

βφ 1.(Καταστροφική διαγραφή) Εστω $b = 10, t = 4$ και οι αριθμοί $s_1 = 0.8134 * 10^3$, $s_2 = 0.3547 * 10^3$, $s_3 = -0.1168 * 10^4$. Να υπολογισθεί η ποσότητα $fl(s_1 + s_2 + s_3)$.

**Απόδειξη:** Παρατηρούμε ότι:

$$|s_1 + s_2 + s_3| = 0.1$$

$$|s_1| + |s_2| + |s_3| = 0.23361 * 10^4$$

δηλαδή $0.1 << 0.23361 * 10^4$

Από την ανάλυση σφάλματος στο άθροισμα περιμένουμε την εμφάνιση μεγάλου σχετικού σφάλματος στον υπολογισμό της ποσότητας $fl(s_1 + s_2 + s_3)$.

Πράγματι ο υπολογισμός $fl(s_1 + s_2 + s_3)$ θα εκτελεσθεί ως εξής:

$$fl(s_1 + s_2 + s_3) = fl(fl(s_1 + s_2) + s_3) = 0.0$$

αφού

$$fl(s_1 + s_2) = 0.1168 * 10^4$$

$$(0.8134 * 10^3 + 0.3547 * 10^3 = 1.1681 * 10^3 = 0.11681 * 10^4)$$

Το ακριβές θεωρητικό άθροισμα είναι: $s_1 + s_2 + s_3 = 0.1$

Παρατηρούμε ότι το σχετικό σφάλμα

$$Rel = \frac{|(s_1 + s_2 + s_3) - fl(s_1 + s_2 + s_3)|}{|s_1 + s_2 + s_3|} = 1$$

Το σχετικό σφάλμα είναι μεγάλο με αποτέλεσμα την παραγωγή ανακριβούς αποτελέσμτος. Το φαινόμενο αυτό ονομάζεται καταστροφική διαγραφή (catastrophic cancellation) και εμφανίζεται όταν δύο αριθμοί περίπου ίδιου μεγέθους πρόκειται να αφαιρεθούν.

Σάυτήν την περίπτωση η διαγραφή των πιο αριστερών ψηφίων από τη mantissa ισχυροποιεί τα πιο δεξιά καθιστώντας τα πιο σημαντικά.

**Παρατήρηση:** Το αποτέλεσμα παραμένει το ίδιο ανακριβές ακόμα και αν οι αριθμοί προστεθούν κατά αύξουσα σειρά.

Πράγματι εάν πάρουμε $s_1 = -0.1168*10^4,\ s_2 = 0.3547*10^3,\ s_3 = 0.8134*10^3$ και υπολογίσουμε $fl(s_1+s_2+s_3) = 0.0$ αφού $(0.11680*10^4 - 0.03547*10^4 = 0.08133*10^4$, και $0.08134 - 0.08130 = 0.00004$, για $t = 4$ παίρνουμε 0). □

**2.**(Απώλεια ακριβείας)

**Απόδειξη:** (i) Έστω οι αριθμοί $p = 3.1415926536, q = 3.145957341$ οι οποίοι είναι περίπου ίσοι για $t = 11$. Η διαφορά τους $p - q = -0.0000030805$ έχει μόνο 5 δεκαδικά ψηφία ακριβείας. Τα εμφανιζόμενα μηδενικά στη mantissa αντιστοιχούν σε απώλεια ακρίβειας. Αυτό ονομάζεται loss of significance ή subtructive cancelation.

(ii) Έστω $t = 6$ και θέλουμε να αφαιρέσουμε τους αριθμούς $1 - 0.999999$.

Βήμα 1. Ευθυγράμμιση των εκθετών

$$0.1 * 10^1$$

$$0.0999999 * 10^1$$

(το τελευταίο 9 στην παραπάνω παράσταση εξασφαλίζεται εάν μπορούμε να κρατήσουμε μια θέση για το τελευταίο guard bit.

Βήμα 2. Εκτέλεση πράξης

$$0.1 * 10^1 - 0.0999999 * 10^1 = 0.0000001 * 10^1 = 0.1 * 10^{-5}$$

Εάν η μηχανή δεν εξασφαλίζει το guard bit το τελευταίο 9 θα αποκοπεί και το αποτέλεσμα θα είναι $0.1 * 10^{-4}$ προκύπτοντας έτσι σχετικό σφάλμα της τάξης του 10.

**Συμπέρασμα:** Πολύ σημαντική η ύπαρξη του guard bit στη διεξαγωγή των αριθμητικών υπολογισμών.

**Παρατήρηση:** Η ποσότητα $0.1 * 10^{-5}$ προκύπτει εάν αφαιρέσουμε από το 1 την ποσότητα 0.99999 αντί της 0.999999.

Το σχετικό σφάλμα της 0.99999 σαν προσέγγιση της 0.999999 είναι περίπου $9*10^{-6}$ (της ίδιας τάξης με το $\varepsilon_M$). Αυτό μας δείχνει ότι η υπολογιζόμενη τιμή μπορούσε να είχε προκύψει επακριβώς εάν είχαμε κάνει μια μικρή διαταραχή της τάξης του $10^{-6}$ στα αρχικά δεδομένα μας. Δηλαδή

$$fl(a-b) = a(1+\varepsilon_a) - b(1+\varepsilon_b),\ |\varepsilon_a|,\ |\varepsilon_b| \leq \varepsilon_M$$

(όπου $\varepsilon_M$ έχει προσαρμοσθεί κατάλληλα).

Αυτός είναι στόχος της backward error analysis, που θα εξετάσουμε στο επόμενο κεφάλαιο, όπου προσπαθούμε να εκφράσουμε το τελικό αποτέλεσμα σαν ακριβή τιμή μεταξύ μικρών διαταράξεων των αρχικών δεδομένων.

Οι διαταράξεις αυτές από μόνες τους είναι μικρές και όχι σημαντικές. Εάν το τελικό αποτέλεσμα που θα προκύψει με την εφαρμογή μικρών διαταράξεων στα δεδομένα δεν είναι ικανοποιητικό τότε θα φταίει η κατάσταση του προβλήματος. □

**3.** Εστω $x_1, x_2 \in \mathcal{R}$. Να υπολογισθεί η ποσότητα: $fl(\frac{x_1}{x_1+x_2})$

**Απόδειξη:** Προκύπτουν τα παρακάτω βήματα:

$a = fl(x_1 + x_2)$

$b = fl(\frac{x_1}{a})$

Οι floating point υπολογισμοί μας δίνουν:

$$a = fl(x_1 + x_2) = (x_1 + x_2)(1 + \varepsilon),\ (1 - u) \leq 1 + \varepsilon \leq (1 + u),\ |\varepsilon| \leq u_1$$

$$b = fl(\frac{x_1}{a}) = \frac{x_1}{x_1 + x_2} \cdot \frac{1 + n}{1 + \varepsilon},\ |n| \leq u_1$$

Ετσι τελικά προκύπτει

$$fl(\frac{x_1}{x_1 + x_2}) = \frac{x_1}{x_1 + x_2}(1 + k),\ \text{όπου}\, |k| \leq 2.02u_1$$

$$(|k| = |\frac{1 + n}{1 + \varepsilon} - 1| = \frac{|n - \varepsilon|}{|1 + \varepsilon|} \leq \frac{2u_1}{1 - u} \leq 1.01 \cdot 2u_1 = 2.02u_1$$

(Υποθέσαμε ότι $\frac{1}{1-u} \leq 1.01$). □

**4. (Υπολογισμός τετραγωνικής ρίζας)**

**Απόδειξη:** Οι τετραγωνικές ρίζες είναι απαραίτητες στους διάφορους υπολογισμούς και ιδιαίτερα στις μεθόδους που χρησιμοποιούν στοιχειώδεις ορθογώνιους μετασχηματισμούς. Το σφάλμα που προξενείται κατά την εύρεση μιας τετραγωνικής ρίζας κυρίως εξαρτάται από το είδος του αλγορίθμου που χρησιμοποιήθηκε. Δεν επιθυμούμε να κάνουμε αναλυτική περιγραφή αυτών των αλγόριθμων, αλλά απλά θα υποθέτουμε ότι:

$$fl(\sqrt{x}) = \sqrt{x}(1 + \varepsilon),\ |\varepsilon| < (1.00001)u$$

Στους αλγόριθμους που ασχολούνται με πίνακες ο αριθμός των τετραγωνικών ριζών είναι μικρός συγκρινόμενος με τον απαιτούμενο αριθμό άλλων πράξεων και κατά συνέπεια ακόμα και σφάλματα μεγαλύτερα από αυτά του προηγούμενου τύπου επιφέρουν μικρές διαφορές στο συνολικό φράγμα σφάλματος. □

**5. (Υπολογισμός Ορίζουσας ενός τριδιαγώνιου πίνακα)**

**Απόδειξη:** Ορίζουμε τον τριδιαγώνιο πίνακα ως εξής:

$$c_{ii} = a_i,\ c_{i,i+1} = \beta_{i+1},\ c_{i+1,i} = \gamma_{i+1}$$

$$C = \begin{pmatrix} a_1 & \beta_2 & & & O \\ \gamma_2 & a_2 & \beta_3 & & \\ & \gamma_3 & a_3 & \beta_4 & \\ & & & \ddots & \\ & & & & \ddots \\ & O & & \gamma_n & a_n \end{pmatrix}$$

όπου

$$det \begin{pmatrix} a_1 & \beta_2 & 0 \\ \gamma_2 & a_2 & \beta_3 \\ 0 & \gamma_3 & a_3 \end{pmatrix} = p_3$$

Η ορίζουσα $det(D) = p_n$ βρίσκεται από την εξής ακολουθία:

$$p_0 = 1, p_1 = a_1,$$

$$p_r = a_r p_{r-1} - \beta_r \gamma_r p_{r-2},\ \ r = 2, 3, \ldots, n \tag{1}$$

(Για παράδειγμα $p_3$ είναι η $3 \times 3$ κύρια υποορίζουσα, $p_r$ είναι η $r \times r$ κύρια υποορίζουσα του πίνακα $C$ ). Σύμφωνα με τη backward error-analysis θα δείξουμε ότι $p_r$ είναι η ορίζουσα του πίνακα $C'$ με τροποποιημένα στοιχεία $a_i', \beta_i', \gamma_i'$.

Εστω ότι $p_0, p_1, \ldots, p_{r-1}$ είναι ακριβή αποτελέσματα για ένα πίνακα τάξης $r-1$ με τροποποιημένα στοιχεία $a_i', \beta_i', \gamma_i',\ (i = 1, 2, \ldots, r-1)$

Υπολογίζουμε το $p_r$ από την (1) (Υπολογίζουμε ξεχωριστά τις ποσότητες)

$$fl(a_r p_{r-1}) \equiv a_r p_{r-1}(1 + \varepsilon_1) \equiv A_r,\ \ |\varepsilon_1| \leq u$$

$$fl(\beta_r \gamma_r p_{r-2}) \equiv \beta_r \gamma_r p_{r-2}(1 + \varepsilon_2) \equiv B_r,$$

$$(1-u)^2 \leq 1 + \varepsilon_2 \leq (1+u)^2$$

Το υπολογιζόμενο

$$p_r \equiv fl(A_r - B_r) \equiv A_r(1+\varepsilon_3) - B_r(1+\varepsilon_3), \;\; |\varepsilon_3| \leq u$$

$$\Rightarrow p_r \equiv a_r p_{r-1}(1+\varepsilon_1)(1+\varepsilon_3) - \beta_r\gamma_r p_{r-2}(1+\varepsilon_2)(1+\varepsilon_3)$$

Θέτουμε:

$$a'_r \equiv a_r(1+\varepsilon_1)(1+\varepsilon_3)$$

$$\beta'_r \equiv \beta_r(1+\varepsilon_2)^{1/2}(1+\varepsilon_3)^{1/2}$$
$$\gamma'_r \equiv \gamma_r(1+\varepsilon_2)^{1/2}(1+\varepsilon_3)^{1/2}$$

Ετσι προκύπτει ότι

$$p_r \equiv a'_r p_{r-1} - \beta'_r\gamma'_r p_{r-2}$$

οπότε υπολογίζουμε την ορίζουσα πίνακα

$$C' = \begin{pmatrix} a'_1 & \beta'_2 & \\ \gamma'_2 & a'_2 & \beta'_3 \\ & \ddots & \ddots \end{pmatrix}$$

και πρέπει να δώσουμε τα κατάλληλα φράγματα:
Εστω

$$a'_r \equiv a_r(1+E_1) = a_r + a_r E_1,$$

όπου

$$1 + E_1 = (1+\varepsilon_1)(1+\varepsilon_3)$$

$$\beta'_r \equiv \beta_r(1+E_2) = \beta_r + a_r E_2,$$

όπου

$$1 + E_2 = (1+\varepsilon_2)^{1/2}(1+\varepsilon_3)^{1/2}$$

$$\gamma'_r \equiv \gamma_r(1+E_2) = \gamma_r + \gamma_r E_2,$$

όπου

$$1 + E_2 = (1 + \varepsilon_2)^{1/2}(1 + \varepsilon_3)^{1/2}$$

Από τις προηγούμενες σχέσεις προκύπτει ότι:

$$(1 - u)^2 \leq 1 + E_1 \leq (1 + u)^2$$

ή

$$|E_1| \leq 2u_1$$

$$(1 - u)^{3/2} \leq 1 + E_2 \leq (1 + u)^{3/2}$$

ή

$$|E_2| \leq \frac{3}{2}u_1$$

□

## 2.3 Σφάλματα Στρογγύλευσης σε Υπολογισμούς με Πίνακες

Από πολλούς έχει θεωρηθεί ότι η αυστηρή θεωρία ανάλυσης σφάλματος διαδικασιών που περιέχουν πίνακες είναι κάτι το πολύ δύσκολο και μ' αυτό ασχολούνται μόνο ειδικοί! Αυτό είναι πολύ μακρυά από την πραγματικότητα. Οι δυσκολίες που περικλείονται σ' αυτού του είδους την ανάλυση σφάλματος είναι περισσότερο εμπειρικές παρά μαθηματικές. Ο μη πεπειραμένος μελετητής οδηγείται σε συγκρίσεις μεταξύ των αριθμών που προκύπτουν από τη πρακτική εφαρμογή του αλγορίθμου και εκείνων που προκύπτουν από τους ακριβείς θεωρητικούς υπολογισμούς. Τέτοιες συγκρίσεις σπάνια οδηγούν σε χρήσιμα αποτελέσματα. Συνήθως είναι πολύ δύσκολο να περατωθούν και συχνά καταλήγουν στα συμπεράσματα ότι οι αριθμοί που προέκυψαν από τους υπολογισμούς διαφέρουν αρκετά από τους αντίστοιχους θεωρητικά υπολογιζόμενος. Το πιο σημαντικό για μία επιτυχή ανάλυση σφάλματος είναι η επιλογή μίας «καλής» βάσης σύγκρισης, γι' αυτό χρησιμοποιείται κυρίως η τεχνική της backward error analysis, η οποία εμφανίζει πολλά πλεονεκτήματα.

### 2.3.1 Ανάλυση σφάλματος απλών πράξεων με πίνακες

**(i) Πολ/σμός πίνακα με βαθμωτό**

Εστω $A \epsilon \mathbb{R}^{m \times n}$ και $k \epsilon \mathbb{R}$.
Συμβολίζουμε με $B$ τον πίνακα που προκύπτει από τον floating-point υπολογισμό.

$$b_{ij} \equiv fl(k \cdot a_{ij}) \equiv k \cdot a_{ij} \cdot (1 + \varepsilon_{ij}), \;\; |\varepsilon_{ij}| \leq u$$

$$b_{ij} \equiv fl(k \cdot a_{ij}) \equiv k \cdot a_{ij} + k \cdot a_{ij} \cdot \varepsilon_{ij}, \ \ |\varepsilon_{ij}| \leq u$$

Θέτουμε $k \cdot a_{ij} \cdot \varepsilon_{ij} = \varepsilon'_{ij}$ όπου

$$|\varepsilon'_{ij}| = |k \cdot a_{ij} \cdot \varepsilon_{ij}| \leq u \cdot |k| \cdot |a_{ij}|$$

Εστω
$|E'| = (|\varepsilon'_{ij}|)$ τότε $|E'| \leq u \cdot |k| \cdot |A|$
Τελικά

$$B = k \cdot A + E', \ \ \text{όπου } |E'| \leq u \cdot |k| \cdot |A|$$

Το σχετικό σφάλμα δίνεται από τη σχέση

$$\frac{\|B - k \cdot A\|}{\|k \cdot A\|} = \frac{\|E'\|}{\|k \cdot A\|}, \ \ |E'| \leq u \cdot |k| \cdot |A|$$

Ανάλογα με τη νόρμα που θα επιλέξουμε δίνουμε κατάλληλα φράγματα στη προηγούμενη ανισότητα. Εάν επιλέξουμε την $\|\cdot\|_\infty$ παίρνουμε:

$$\frac{\|B - k \cdot A\|_\infty}{\|k \cdot A\|_\infty} \leq \frac{u \cdot |k| \cdot \|A\|_\infty}{\|k \cdot A\|_\infty} \leq u$$

**Παρατήρηση:** Το σχετικό σφάλμα που προκύπτει από τον πολ/σμό πίνακα με βαθμωτό είναι μικρό.

**Ορισμός:** Εστω $A, B \epsilon \mathbb{R}^{m \times n}$ και ο $B$ προκύπτει με υπολογισμούς από τον $A$. Το **υπόλοιπο (residual)** του $B$ είναι ο πίνακας $A - B$

Εάν $\nu : \mathbb{R}^{m \times m} \to \mathbb{R}$ είναι μία νόρμα, τότε το **απόλυτο σφάλμα** στον υπολογισμό του $B$ ως προς τη νόρμα $\nu$ είναι ο αριθμός: $\nu(A - B)$
Εάν $A \neq 0$ το **σχετικό σφάλμα** στον υπολογισμό του $B$ ως πρός τη νόρμα $\nu$ είναι ο αριθμός

$$\frac{\nu(A - B)}{\nu(A)}$$

**Παράδειγμα 2.3.1** *Εάν $A, B \epsilon \mathbb{R}^{m \times n}$ αποδείξτε ότι:*
*1)$fl(A) = A + E, \ \ |E| \leq u|A|$*
*2)$fl(A + B) = (A + B) + E, \ \ |E| \leq u|A + B|$*
*Δώστε κατάλληλα φράγματα στο σφάλμα χρησιμοποιώντας κάποια νόρμα και ελέγξτε και στις δύο περιπτώσεις το σχετικό σφάλμα. Τι παρατηρείται;*

**Απόδειξη:** Εστω $C \in \mathbb{R}^{mxn}$ ο πίνακας που προκύπτει από τους υπολογισμούς.

$1) c_{ij} \equiv fl(a_{ij}) \equiv a_{ij}(1 + \varepsilon'_{ij}) = a_{ij} + a_{ij} \cdot \varepsilon'_{ij}, \ \ |\varepsilon'_{ij}| \leq u$

Εάν $a_{ij}\varepsilon'_{ij} = \varepsilon_{ij}$ τότε $|\varepsilon_{ij}| \leq u|a_{ij}|$.

Τελικά

$$C \leq A + E, \ |E| \leq u|A|$$

Εάν επιλέξουμε την $\|\cdot\|_\infty$ έχουμε

$$C \leq A + E, \ \|E\|_\infty \leq u\|A\|_\infty$$

Το σχετικό σφάλμα δίνεται από τον τύπο:

$$\frac{\|C - A\|_\infty}{\|A\|_\infty} \leq \frac{\|E\|_\infty}{\|A\|_\infty} \leq \frac{u\|A\|_\infty}{\|A\|_\infty} = u$$

**Παρατήρηση:** Το σχετικό σφάλμα που προκύπτει από τον υπολογισμό ενός πίνακα είναι μικρό.

2)$c_{ij} \equiv fl(a_{ij} + b_{ij}) \equiv (a_{ij} + b_{ij})(1 + \varepsilon'_{ij}) = a_{ij} + b_{ij} + (a_{ij} + b_{ij})\varepsilon'_{ij}, \ |e'_{ij}| \leq u$

Εάν

$$(a_{ij} + b_{ij})\varepsilon'_{ij} = \varepsilon_{ij}, \ |\varepsilon_{ij}| \leq u|a_{ij} + b_{ij}|$$

Τελικά

$$C \leq (A + B) + E, \ |E| \leq u|A + B|$$

Εάν επιλέξουμε την $\|\cdot\|_\infty$ έχουμε

$$C \equiv (A + B) + E, \ \|E\|_\infty \leq u\|A + B\|_\infty \leq u(\|A\|_\infty + \|B\|_\infty)$$

Το σχετικό σφάλμα δίνεται από το τύπο:

$$\frac{\|C - (A + B)\|_\infty}{\|A + B\|_\infty} \leq \frac{\|E\|_\infty}{\|A + B\|_\infty} \leq u\frac{\|A\|_\infty + \|B\|_\infty}{\|A + B\|_\infty}$$

**Παρατήρηση:** Το σχετικό σφάλμα που προκύπτει από τη πρόσθεση δύο πινάκων $A, B$ εάν $||A + B||_\infty << \|A\|_\infty + \|B\|_\infty$ μπορεί να μην είναι μικρό. □

### (ii) Πολλαπλασιασμός πινάκων

Εξετάζουμε πρώτα τον υπολογισμό $A\underline{x}$ όπου $A \epsilon \mathbb{R}^{n\times n}$ και $\underline{x} \epsilon \mathbb{R}^n$

Εάν $\underline{y} = A\underline{x}$ μπορούμε να γράψουμε:

$$y_i = fl(a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n)$$

Σύμφωνα με τον υπολογισμό του εσωτερικού γινομένου έχουμε:

$$y_i = a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n + \varepsilon_i$$

όπου

$$|\varepsilon_i| \leq u_1[n|a_{i1}||x_1| + n|a_{i2}||x_2| + (n-1)|a_{i3}||x_3| + \ldots + 2|a_{in}||x_n|]$$

Τελικά

$$\underline{y} \equiv A\underline{x} + \underline{\varepsilon},\ \ \underline{\varepsilon}\epsilon\mathbb{R}^n$$

με

$$|\underline{\varepsilon}| \leq u_1 D \cdot |A| \cdot |\underline{x}|,\ \ D = \begin{pmatrix} n & 0 & 0 & 0 & \cdots & 0 \\ 0 & n & 0 & 0 & \cdots & \vdots \\ 0 & 0 & n-1 & 0 & \cdots & \vdots \\ 0 & 0 & 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & \vdots & 0 & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 2 \end{pmatrix}$$

**Παρατήρηση:** Εάν θεωρήσουμε τις νόρμες $\|\cdot\|_\infty,\ \|\cdot\|_2$ προκύπτει ότι:

$$\|\underline{\varepsilon}\|_\infty \leq u_1\|D\|_\infty \cdot \|A\|_\infty \cdot \|\underline{x}\|_\infty \leq nu_1 \cdot \|A\|_\infty \cdot \|\underline{x}\|_\infty$$

$$\|\underline{\varepsilon}\|_2 \leq u_1\|D\|_2 \cdot \|A\|_2 \cdot \|\underline{x}\|_2 \leq u_1\sqrt{n}\|D\|_2 \cdot \sqrt{n}\|A\|_2 \cdot \|\underline{x}\|_2 \leq n^2u_1 \cdot \|A\|_2 \cdot \|\underline{x}\|_2$$

**Σημείωση:** Ο πίνακας $|A|$ έχει στοιχεία $|a_{ij}|$.

Το σχετικό σφάλμα δίνεται από τον τύπο:

$$\frac{\|\underline{y} - A\underline{x}\|_\infty}{\|A\underline{x}\|_\infty} = \frac{\|\underline{\varepsilon}\|_\infty}{\|A\underline{x}\|_\infty} \leq \frac{u_1 n\|A\|_\infty\|\underline{x}\|_\infty}{\|A\underline{x}\|_\infty}$$

**Παρατήρηση:** Εάν $\|A\underline{x}\|_\infty \ll \|A\|_\infty\|\underline{x}\|_\infty$ τότε το σχετικό σφάλμα που προκύπτει κατά τον υπολογισμό του $A\underline{x}$ μπορεί να μην είναι μικρό.

Με τη βοήθεια αυτού του υπολογισμού εάν $A, B, C \in \mathbb{R}^{n\times n}$ έχουμε:

$$C \equiv fl(AB) \equiv AB + E$$

όπου

$$\|E\|_\infty \leq nu_1\|A\|_\infty\|B\|_\infty$$

Το σχετικό σφάλμα δίνεται από τον τύπο:

$$\frac{\|C - AB\|_\infty}{\|AB\|_\infty} = \frac{\|E\|_\infty}{\|AB\|_\infty} \leq nu_1 \frac{\|A\|_\infty \|B\|_\infty}{\|AB\|_\infty}$$

**Παρατήρηση:** Εάν $\|AB\|_\infty \ll \|A\|_\infty \|B\|_\infty$ τότε το σχετικό σφάλμα που προκύπτει κατά τον πολ/σμό δύο πινάκων μπορεί να μην είναι μικρό.

Για παράδειγμα έστω $A = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix},\ B = \begin{pmatrix} 1 & 0 \\ -.99 & 0 \end{pmatrix}$

Παρατηρούμε ότι $AB = \begin{pmatrix} .01 & 0 \\ 0 & 0 \end{pmatrix}$
Βλέπουμε ότι $\|A\|_\infty = 1,\ \|B\|_\infty = 1,\ \|AB\|_\infty = .01$

### 2.3.2 Ορθογώνιοι Πίνακες

Ένας τετραγωνικός πίνακας $U$ λέγεται ορθογώνιος εάν

$$U^T \cdot U = U \cdot U^T = I$$

Οι ορθογώνιοι πίνακες παίζουν πολύ σημαντικό ρόλο στους αριθμητικούς υπολογισμούς με πίνακες.
**Βασικές Ιδιότητες**

1. Ο αντίστροφος ενός ορθογώνιου πίνακα είναι ο ανάστροφός του.
2. Το γινόμενο δύο ορθογώνιων πινάκων είναι ορθογώνιος πίνακας.

**Θεώρημα 2.3.1** *Έστω $A$ ορθογώνιος πίνακας. Τότε $\|A\|_2 = 1$.*

**Απόδειξη:** $\|A\|_2 = \sqrt{\rho(A^T A)} = \sqrt{\rho(I)} = 1.$ □

**Θεώρημα 2.3.2** *Έστω $A \in \mathbb{R}^{n\times n}$. Τότε*

1. $\|PA\|_2 = \|A\|_2,\ P \in \mathbb{R}^{n\times n}$ *ορθογώνιος*
2. $\|AQ\|_2 = \|A\|_2,\ Q \in \mathbb{R}^{n\times n}$ *ορθογώνιος*
3. $\|PAQ\|_2 = \|A\|_2,\ P, Q \in \mathbb{R}^{n\times n}$ *ορθογώνιοι*

*Δηλαδή η $\|\cdot\|_2$ είναι ορθογώνια αναλλοίωτη.*

**Απόδειξη:**

1. $\|PA\|_2 = \sqrt{\rho(A^T P^T P A)} = \sqrt{\rho(A^T A)} = \|A\|_2$
2. $\|AQ\|_2 = \sqrt{\rho(Q^T A^T A Q)} = \sqrt{\rho(A^T A)} = \|A\|_2$

3. $\|PAQ\|_2 = \sqrt{\rho(Q^T A^T P^T PAQ)} = \|A\|_2$

□

Η παρακάτω άσκηση μας δείχνει τη χρησιμότητα των ορθογώνιων πινάκων στην ανάλυση σφάλματος.

**'Ασκηση 2.3.1** *Έστω $A \in \mathbb{R}^{n\times n}$, $Q \in \mathbb{R}^{n\times n}$ ορθογώνιος. Να αποδείξετε ότι:*

$$fl(QA) = QA + E, \ \ \|E\|_2 \leq n^2 u_1 \|A\|_2$$

*(forward error analysis)*
*Τι συμπεραίνετε για την ευστάθεια του γινομένου αυτού;*

**Απόδειξη:** Γνωρίζουμε ότι $fl(QA) = QA + E, \ \ \|E\| \leq u_1 \|D\| \cdot \|Q\| \cdot \|A\|$ όπου

$$D = \begin{pmatrix} n & & & & \\ & n & & & \\ & & n-1 & & \\ & & & \ddots & \\ & & & & 2 \end{pmatrix}$$

Ισχύει ότι $\|A\|_2 = \sqrt{\rho(A^T A)}$, οπότε

$$fl(QA) = QA + E, \ \ \|E\|_2 \leq \sqrt{n^2} u_1 \|Q\|_2 \|A\|_2 \leq$$

$$n u_1 \sqrt{n} \|Q\|_2 \sqrt{n} \|A\|_2 = n^2 u_1 \|A\|_2$$

διότι $\|Q\|_2 = 1$. Τελικά $\|E\|_2 \leq n^2 u_1 \|A\|_2$.

Το σχετικό σφάλμα υπολογίζεται από τον τύπο

$$Rel = \frac{\|fl(QA) - QA\|_2}{\|QA\|_2} = \frac{\|E\|_2}{\|QA\|_2} \leq \frac{n^2 u_1 \|A\|_2}{\|QA\|_2}$$

Επειδή $Q$ ορθογώνιος προκύπτει ότι $\|QA\|_2 = \|A\|_2$. Τελικά $Rel \leq n^2 u_1$. Επομένως το γινόμενο τυχαίου πίνακα με ορθογώνιο πίνακα είναι πάντοτε ευσταθές. □

**Παρατήρηση:** Όποτε εμφανίζονται ορθογώνιοι πίνακες δίνουν πάντα ευσταθές γινόμενο όταν πολλαπλασιάζονται με άλλους τυχαίους πίνακες. Γι' αυτό στις μεθόδους παραγοντοποιήσεων πινάκων ενδείκνυται η εφαρμογή ορθογώνιων μετασχηματισμών που ισοδυναμεί με εξ' αριστερών ή εκ' δεξιών πολλαπλασιασμό με κάποιον άλλο ορθογώνιο. Έτσι θα προκύπτει ευσταθής μέθοδος.

**Παρατήρηση:** Γενικά στην ανάλυση σφάλματος δεν είμαστε τόσο αυστηροί στο ακριβές θεωρητικό φράγμα που μπορεί να δοθεί για το εμφανιζόμενο σφάλμα και το οποίο μπορεί να εκφρασθεί με διάφορους τρόπους. Ο κύριος σκοπός της ανάλυσης σφάλματος είναι να ανακαλύπτει πιθανές αστάθειες και να καταλήγει σε συμπεράσματα που πιθανόν οδηγούν σε βελτίωση των αλγορίθμων. Αλλωστε οι εκ των προτέρων εκτιμήσεις στα εμφανιζόμενα σφάλματα στρογγύλευσης απέχουν πολύ από τις εμφανιζόμενες στην πράξη τιμές οι οποίες μπορεί να εκτιμηθούν εκ των υστέρων λαμβάνοντας υπ΄ όψη τη στατιστική κατανομή των σφαλμάτων και τη μορφή των πινάκων που χρησιμοποιούνται.

Ολα τα προηγούμερα συνοψίζονται στον παρακάτω πίνακα

| **Υπολογισμός** | **Σχετικό σφάλμα** |
|---|---|
| $A\epsilon\mathbb{R}^{m\times m}$<br>$fl(A)$ | Πάντοτε μικρό |
| $A\epsilon\mathbb{R}^{m\times n}, k\epsilon\mathbb{R}$<br>$fl(kA)$ | Πάντοτε μικρό |
| $A, B\epsilon\mathbb{R}^{m\times n}$<br>$fl(A+B)$ | Οχι πάντοτε μικρό<br>εάν $\|A+B\| << \|A\|+\|B\|$ |
| $A\epsilon\mathbb{R}^{n\times n}, x\epsilon\mathbb{R}^{n}$<br>$fl(A\cdot x)$ | Όχι πάντοτε μικρό<br>εάν $\|A\cdot x\| << \|A\|\cdot\|x\|$ |
| $A, B\epsilon\mathbb{R}^{n\times n}$<br>$fl(A\cdot B)$ | Οχι πάντοτε μικρό<br>εάν $\|A\cdot B\| << \|A\|\cdot\|B\|$ |

## ΑΣΚΗΣΕΙΣ

**1.** Δίνεται ο πίνακας

$$A = \begin{pmatrix} -20.8571 & -9.06467 & 12.9955 \\ -43.1924 & -18.7528 & 26.8994 \\ -63.5860 & -27.6219 & 39.6100 \end{pmatrix}$$

Εάν υπολογίσουμε $fl(A^3)$ προκύπτει

$$\bar{A}^2 = fl(A * A)$$
$$\bar{A}^3 = fl(\bar{A}^2 * A)$$

Με χρήση floating point αριθμητικής $\beta = 2$, $t = 15$, $2^{-15} \cong 3 * 10^{-5}$

$$\bar{A}^3 = \begin{pmatrix} -0.0570040 & -0.0247397 & 0.0354452 \\ -0.785385 & -0.341064 & 0.489075 \\ -0.154087 & -0.0671120 & 0.0962429 \end{pmatrix}$$

Ακριβής δύναμη μέχρι 6 σημαντικά ψηφία είναι:

$$A^3 = \begin{pmatrix} 0.00906177 & 0.00393613 & -0.00564315 \\ 0.0187545 & 0.00814694 & -0.00116797 \\ 0.027619 & 0.0119938 & -0.0171949 \end{pmatrix}$$

Μεγάλη απόκλιση ακόμα και στο πρόσημο γιατί:

$$\|A^3\|_F \cong 0.043$$

ενώ

$$\|A\|_F * \|A\|_F * \|A\|_F \cong 100 * 100 * 100 = 1.000.000$$

$$(\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}, A\epsilon\mathbb{R}^{m\times n})$$

Παρατηρούμε ότι $\|A\|_F * \|A\|_F * \|A\|_F >> \|A^3\|_F$

**2.** Να υπολογιστεί το γινόμενο δύο άνω τριγωνικών πινάκων

$$A, B\epsilon\mathbb{R}^{2\times 2}.$$

Να υπολογιστεί αναλυτικά ο πίνακας του σφάλματος και να δοθεί κατάλληλο φράγμα. Τι παρατηρείτε για το σχετικό σφάλμα;

**Απόδειξη:** Εστω $A = \begin{pmatrix} a_{11} & a_{12} \\ 0 & a_{22} \end{pmatrix}$, $B = \begin{pmatrix} b_{11} & b_{12} \\ 0 & b_{22} \end{pmatrix}$

$$C \equiv fl(AB) \equiv$$

$$\equiv \begin{pmatrix} a_{11}b_{11}(1+\varepsilon_1) & [a_{11}b_{12}(1+\varepsilon_2) + a_{12}b_{22}(1+\varepsilon_3)](1+\varepsilon_4) \\ 0 & a_{22}b_{22}(1+\varepsilon_5) \end{pmatrix}$$

**i. Forward error analysis**

$$C \equiv \begin{pmatrix} a_{11}b_{11} & a_{11}b_{12} + a_{12}b_{22} \\ 0 & a_{22}b_{22} \end{pmatrix} +$$

$$+ \begin{pmatrix} a_{11}b_{11}\varepsilon_1 & a_{11}b_{12}(\varepsilon_2+\varepsilon_4) + a_{12}b_{22}(\varepsilon_3+\varepsilon_4) + a_{11}b_{12}\varepsilon_2\varepsilon_4 + a_{12}b_{22}\varepsilon_3\varepsilon_4 \\ 0 & a_{22}b_{22}\varepsilon_5 \end{pmatrix}$$

όπου $|\varepsilon_i| \leq u, i = 1, 2, \ldots, 5$
Συνεπώς

$$C \equiv AB + E$$

όπου

$$|E| \leq \begin{pmatrix} u|a_{11}\|b_{11}| & 2u(|a_{11}\|b_{12}| + |a_{12}\|b_{22}| + u^2(|a_{11}\|b_{12}| + |a_{12}\|b_{22}|) \\ 0 & u|a_{22}\|b_{22}| \end{pmatrix}$$

Εάν αγνοήσουμε το σφάλμα τάξης $u^2(\mathcal{O}(u^2))$ τότε

$$|E| \leq 2u|A||B|$$

Το σχετικό σφάλμα δίνεται από τον τύπο:

$$\frac{\|C - AB\|_\infty}{\|AB\|_\infty} = \frac{\|E\|_\infty}{\|AB\|_\infty} \leq 2u \cdot \frac{\|A\|_\infty \|B\|_\infty}{\|AB\|_\infty}$$

**Παρατήρηση:** Εάν $\|AB\|_\infty << \|A\|_\infty \|B\|_\infty$ τότε το σχετικό σφάλμα στο πολ/σμό μπορεί να μην είναι μικρό.

**Σημείωση:** Στο σχετικό σφάλμα μπορούμε να χρησιμοποιήσουμε όποια νόρμα θέλουμε. Συνήθως επιλέγουμε εκείνη τη νόρμα που υπολογίζεται ευκολότερα.

**ii. Backward error analysis**

Εάν θέσουμε

$$\hat{A} = \begin{pmatrix} a_{11} & a_{12}(1+\varepsilon_3)(1+\varepsilon_4) \\ 0 & a_{22}(1+\varepsilon_5) \end{pmatrix}$$

$$\hat{B} = \begin{pmatrix} b_{11}(1+\varepsilon_1) & b_{12}(1+\varepsilon_2)(1+\varepsilon_4) \\ 0 & b_{22} \end{pmatrix}$$

όπου $|\varepsilon_i| \leq u, i = 1, 2, \ldots, 5.$
Παρατηρούμε ότι

$$fl(AB) = \hat{A} \cdot \hat{B} = (A + E)(B + F)$$

Εάν αγνοήσουμε το σφάλμα τάξης $u^2$ $(\mathcal{O}(u^2))$ τότε

$$\|E\| \leq 2u\|A\|, \ \ \|F\| \leq 2u\|B\|$$

□

**3.** Έστω $A \in \mathbb{R}^{n\times n}$, $Q \in \mathbb{R}^{n\times n}$ ορθογώνιος. Να αποδείξετε ότι:

1. $fl(Q \cdot A) = Q \cdot (A + E)$, $\|E\|_2 \leq n^2 u\|A\|_2$
2. Να υπολογισθεί το σχετικό σφάλμα
3. Τί συμπεραίνετε για την ευστάθεια του γινομένου αυτού;
4. Τί συμπεραίνετε για την ευστάθεια του γινομένου $H \cdot A$ όπου $H$ πίνακας Housholder;

Σημείωση: $\|A\|_2 = \sqrt{\text{μέγιστη ιδιοτιμή του } (A^T \cdot A)}$

**Απόδειξη:**

$$fl(Q\underline{a}) = fl(q_{11}a_1 + q_{12}a_2 + \ldots + q_{1n}a_n) =$$
$$q_{11}a_1(1+\varepsilon_1) + q_{12}a_2(1+\varepsilon_2) + \ldots + q_{1n}a_n(1+\varepsilon_n) =$$
$$\underline{q}^t \cdot (\underline{c} + \underline{\varepsilon}), \ |\varepsilon_i| \leq (n-i+1)u_1$$

Αυτό μπορεί να γενικευθεί ως εξής:

$$fl(Q \cdot A) = Q \cdot (A + E)$$

$$E = \begin{pmatrix} a_{11}\varepsilon_{11} & a_{12}\varepsilon_{12} & \ldots & a_{1n}\varepsilon_{1n} \\ a_{2n}\varepsilon_{2n} & a_{2n}\varepsilon_{2n} & \ldots & a_{2n}\varepsilon_{2n} \\ \vdots & \vdots & & \vdots \\ a_{nn}\varepsilon_{nn} & a_{nn}\varepsilon_{nn} & \ldots & a_{nn}\varepsilon_{nn} \end{pmatrix}, \ |\varepsilon_{ij}| \leq (n-i+1)u_1$$

$$\|E\|_2 \leq u_1 \left\| \begin{matrix} n & & & & \\ & n & & & \\ & & n-1 & & \\ & & & \ddots & \\ & & & & 2 \end{matrix} \right\|_2 \cdot \|A\|_2$$

$$Rel = \frac{\|QA - Q(A+E)\|_2}{\|QA\|_2} = \frac{\|QE\|_2}{\|QA\|_2} = \frac{\|E\|_2}{\|A\|_2} \leq n^2 u_1$$

β' τρόπος

$$fl(QA) = QA + E$$

$$\frac{\|fl(QA) - QA\|_2}{\|QA\|_2} = \frac{\|E\|_2}{\|Q\|_2\|A\|_2} \leq \frac{\|E\|_2}{\|A\|_2}$$

□

## 2.4 Ευστάθεια και Κατάσταση Προβλημάτων

Πολλά αριθμητικά προβλήματα μπορεί να περιγραφούν με τη μορφή της συνάρτησης $f : D \subset \mathbb{R}^m \longmapsto \mathbb{R}^n$, όπου οι $m$ συντεταγμένες της παραμέτρου $x$ της $f$ είναι τα δεδομένα που καθορίζουν το πρόβλημα και οι $n$ συντεταγμένες της $f(x)$ είναι τα αποτελέσματα. Το αριθμητικό πρόβλημα ανάγεται στον υπολογισμό μιας προσέγγισης του $f(x)$ δοσμένου ενός $x$.
Η φύση της συνάρτησης $f$ περιορίζει την ακρίβεια που μπορεί να επιτευχθεί από τους υπολογισμούς επειδή χρησιμοποιούνται αριθμοί με $t$ δυαδικά ψηφία.
Ακόμα και αν οι $m$ συντεταγμένες του $x$ είναι γνωστές επακριβώς ίσως να μη μπορούν να παρασταθούν επακριβώς από αριθμούς με $t$ δυαδικά ψηφία.
Εάν λοιπόν περιοριστούμε σε αριθμούς $t$-ψηφίων θα έχουμε προσεγγίσεις $t$-ψηφίων στα αρχικά δεδομένα και έτσι περιορίζεται η ακρίβεια πριν ακόμα αρχίσουν οι υπολογισμοί. Βλέπουμε λοιπόν ότι εκτός από περιπτώσεις όπου τα δεδομένα ορίζονται και παριστάνονται επακριβώς από αριθμούς με $t$-ψηφία αρχίζουμε στην καλύτερη των περιπτώσεων με αυτό που ονομάζουμε **προσέγγιση $t$-ψηφίων** (t-digit approximation).

Εστω $x^*$ μία προσέγγιση στα αρχικά δεδομένα. Στην καλύτερη των περιπτώσεων το αριθμητικό μας πρόβλημα μπορεί να υπολογίσει την $f(x^*)$.
Για ορισμένες κατηγορίες προβλημάτων το $f(x)$ και $f(x^*)$ μπορεί να διαφέρουν σημαντικά. Ενα τέτοιο πρόβλημα λέμε ότι έχει **κακή κατάσταση (ill-conditioned)**. Εάν προσπαθήσουμε να επιλύσουμε ένα ill-conditioned πρόβλημα αρχίζοντας με ανακριβή δεδομένα, η λύση θα είναι ανακριβής ανεξάρτητα του τρόπου με τον οποίο υπολογίσθηκε.

**Παράδειγμα 2.4.1** *Ill-conditioned πρόβλημα: Εύρεση του αριθμού των πραγματικών ριζών ενός πολυωνύμου.*

**Απόδειξη:** Το πολυώνυμο $p(x) = x^4 - x^2(2a-1) + a(a-1)$ δείχνει μια ασυνεχή μεταβολή του αριθμού των πραγματικών ριζών του καθώς το $a$ μεταβάλεται συνεχώς στο πεδίο των πραγματικών αριθμών. Έτσι προκύπτει το εξής:

| Τιμή του $a$ | Αριθμός πραγματικών ριζών |
|---|---|
| $a \geq 1$ | 4 |
| $a \in [0, 1)$ | 2 |
| $a < 0$ | όχι πραγματικές ρίζες |

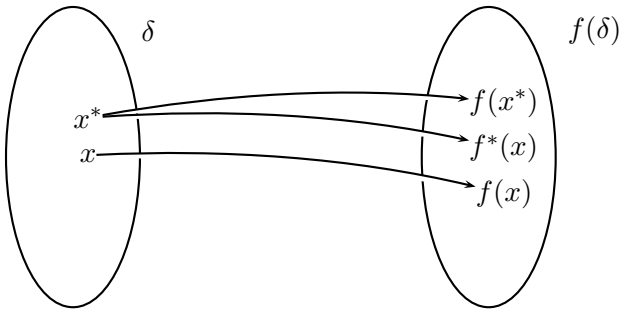
□

Εάν το $f(x)$ και $f(x^*)$ δε διαφέρουν σημαντικά το πρόβλημα έχει **καλή κατάσταση (well-conditioned)**.
Η κατάσταση ενός προβλήματος είναι ιδιότητα που χαρακτηρίζει αποκλειστικά το συγκεκριμένο πρόβλημα και εξετάζει το πόσο θα αλλάζει η λύση του προβλήματος εάν γίνουν διαταράξεις στα αρχικά δεδομένα. Οι διαταράξεις αυτές μπορεί να οφείλονται σε σφάλματα στρογγύλευσης, σε διαφορές μετρήσεων στα αρχικά δεδομένα, σε σφάλματα διακριτοποίησης κ.α.
Έτσι όταν επιλύουμε αριθμητικά ένα πρόβλημα θα πρέπει να εξετάζουμε την επίδραση που επιφέρουν οι διαταράξεις στη λύση του προβλήματος. Μ΄ αυτό το πρόβλημα ασχολείται η ανάλυση διαταραχών(perturbation analysis).
Η εκλογή και η εφαρμογή ενός αλγορίθμου για την επίλυση ενός μαθηματικού προβλήματος που σχετίζεται με την $f$ απαιτεί τον ορισμό μίας νέας συνάρτησης $f^*$ η οποία για δοσμένο $x$ υπολογίζει μία προσεγγιστική λύση $f^*(x)$.
Δεν περιμένουμε η $f^*$ να επιλύσει ill-conditioned προβλήματα με περισσότερη ακρίβεια από αυτήν που εγγυώνται τα δεδομένα, παρόλα αυτά, θα είναι πολύ άσχημο εάν η $f^*$ εισάγει περισσότερες ανακρίβειες από αυτές που ήδη προυπάρχουν.

**Ορισμός:** Ενας αλγόριθμος θα ονομάζεται **ευσταθής (stable)** εάν για κάθε $x \in D$ υπάρχει ένα κοντινό $x^* \in D$ με $d(x, x^*) < \varepsilon$ έτσι ώστε $d(f^*(x), f(x^*)) < \varepsilon$, όπου η συνάρτηση $d(x, y)$ εκφράζει την απόσταση του $x$ από το $y$.
Δηλαδή ο αλγόριθμος δίνει λύση που είναι κοντά στην ακριβή λύση ενός ελαφρά διαταραγμένου αρχικού προβλήματος.

**Ορισμός:** Ένα πρόβλημα έχει καλή(κακή) κατάσταση εάν για κάθε $x \in \mathbb{R}^m$και $x^* \in \mathbb{R}^m$ με $d(x, x^*) < \varepsilon \Rightarrow d(f(x), f(x^*) < (>)\varepsilon$.
Κατά συνέπεια το σχετικό σφάλμα στο αποτέλεσμα που υπολογίζουμε $\left|\frac{f(x)-f(x^*)}{f(x)}\right|$ θα είναι κατά πολύ μεγαλύτερο από το σχετικό σφάλμα στα δεδομένα $\left|\frac{x-x^*}{x}\right|$.
Η κατάσταση ενός προβλήματος δε σχετίζεται με τον αλγόριθμο επίλυσης του προβλήματος.

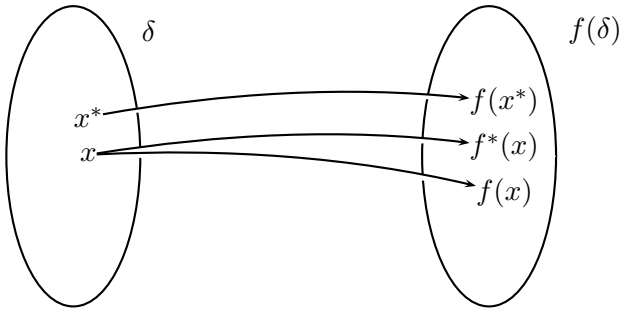
### 2.4.1 Εφαρμογή Ευσταθούς Αλγορίθμου σε well-conditioned Πρόβλημα



Εστω $x^*$ μικρή διατάραξη των αρχικών δεδομένων έτσι ώστε $d(x, x^*) < \varepsilon$ Λόγω της ευστάθειας του αλγορίθμου $d(f^*(x), f(x^*)) < \varepsilon$ Επίσης επειδή το πρόβλημα είναι well conditioned $d(f(x), f(x^*)) < \varepsilon$ Κατά συνέπεια $d(f(x), f^*(x)) < \varepsilon$

**Παρατήρηση:** Εάν ένας ευσταθής αλγόριθμος εφαρμοσθεί σε ένα πρόβλημα well-conditioned η λύση που υπολογίζουμε είναι κοντά στη θεωρητικά αναμενόμενη.

### 2.4.2 Εφαρμογή Ευσταθούς Αλγορίθμου σε ill-conditioned Πρόβλημα




Εστω $x^*$ μία μικρή διατάραξη των αρχικών δεδομένων έτσι ώστε $(x, x^*) < \varepsilon$. Λόγω της ευστάθειας του αλγορίθμου $d(f^*(x), f(x^*)) < \varepsilon$. Όμως επειδή το πρόβλημα είναι ill-conditioned τα $f(x)$ και $f(x^*)$ μπορεί να διαφέρουν σημαντικά με άμεση συνέπεια να μη μπορούμε να εγγυηθούμε ότι τα $f^*(x)$ και $f(x)$ είναι κοντά οπότε η υπολογιζόμενη λύση και η θεωρητικά αναμενόμενη μπορεί να μη συμφωνούν καθόλου.

**Παρατήρηση:** Εάν ένας ευσταθής αλγόριθμος εφαρμοσθεί σε ένα πρόβλημα ill-conditioned η λύση που υπολογίζουμε μπορεί να διαφέρει σημαντικά από τη θεωρητικά αναμενόμενη.

**Παράδειγμα 2.4.2** Η συνάρτηση $f : \mathbb{R}^2 \to \mathbb{R} : f(x) = x_1 + x_2$ σχετίζεται με το πρόβλημα του υπολογισμού του αθροίσματος δύο αριθμών. Ανάλογα με τα δεδομένα το πρόβλημα αυτό μπορεί να είναι η όχι ill-conditioned. Εάν $\underline{x} = (1.947, -1.943)^t$ και $\underline{x}^* = (1.946, -1.944)^t$. Τότε

$$Rel = \frac{|x - x^*|}{|x|} \cong (0.00051, 0.00051)^t$$

$$\frac{\|x - x^*\|_\infty}{\|x\|_\infty} \cong 0.00051$$

Όμως

$$f(x) = 0.004, f(x^*) = 0.003 \text{ και } Rel = \frac{|f(x) - f(x^*)|}{|f(x)|} = 0.25$$

Παρατηρούμε ότι μικρά σχετικά σφάλματα στις συντεταγμένες του $\underline{x}$ επιφέρουν μεγάλα σχετικά σφάλματα στην $f(x)$. Ετσι το πρόβλημα του υπολογισμού του αθροίσματος είναι ill-conditioned για $\underline{x} = (1.947, -1.943)^t$.

Αντίθετα εάν επιλέξουμε $\underline{x} = (2.321, -1.023)^t$ και $\underline{x}^* = (2.320, -1.024)^t$. Τότε

$$Rel = \frac{|x - x^*|}{|x|} \cong (0.00043, 0.0009775)^t$$

και

$$f(x) = 1.298, f(x^*) = 1.296$$

οπότε

$$Rel = |\frac{f(x) - f(x^*)}{f(x)}| = 0.00154$$

Ετσι το πρόβλημα του υπολογισμού του αθροίσματος είναι well conditioned για $\underline{x} = (2.321, -1.023)^t$.
Ας υποθέσουμε τώρα ότι το άθροισμα υπολογίζεται σε floating point αριθμητική. Η συνάρτηση $f^*$ που σχετίζεται μάυτόν τον αλγόριθμο είναι $f^*(x) = fl(x_1 + x_2)$ και

$$f^*(x) = x_1(1 + \varepsilon) + x_2(1 + \varepsilon),\ |\varepsilon| \leq u$$

Ετσι εάν $\underline{x} = (x_1, x_2)^t$ κάποιο δοσμένο διάνυσμα τότε υπάρχει ένα κοντινό διάνυσμα $\underline{x}^* = (x_1(1 + \varepsilon), x_2(1 + \varepsilon))^t$ έτσι ώστε $f^*(x) = f(x^*)$.
Επομένως ο υπολογισμός του αθροίσματος σε floating point αριθμητική είναι ευσταθής σύμφωνα με την τεχνική της backwark error analysis αφού η υπολογιζόμενη λύση του προβλήματος είναι η ακριβής λύση ενός ελαφρά διαταραγμένου αρχικού προβλήματος. □

Για να μπορούμε να διακρίνουμε την κατάσταση ενός προβλήματος συσχετίζουμε το πρόβλημα με έναν αριθμό που ονομάζεται αριθμός κατάστασης (condition number). Το condition number ενός προβλήματος αποτελεί δείκτη της κατάστασής του.
Στην Αριθμητική Γραμμική 'λγεβρα καθορίζονται condition numbers για αρκετά προβλήματα. Πολλές φορές ο υπολογισμός του δείκτη κατάστασης είναι πιο πολύπλοκος και χρονοβόρος από το να επιλύσεις κατ' ευθείαν το πρόβλημα.

Έστω $\Delta x$ η μεταβολή στα δεδομένα έτσι ώστε $x^* = x + \Delta x$. Τότε η μεταβολή στη λύση $y = f(x)$ του υπολογιστικού προβλήματος $x \mapsto f(x)$ είναι $\Delta y = f^*(x) - f(x)$.
Έστω $\delta x = \frac{\|\Delta x\|}{\|x\|}$, $\delta y = \frac{\|\Delta y\|}{\|y\|}$ οι σχετικέςμεταβολές στα αρχικά δεδομένα και στη λύση αντίστοιχα.
Θεωρούμε τον αριθμό

$$c(f, x) = \lim_{d \to 0} \max \left\{ \frac{\delta y}{\delta x} : \|\Delta x\| \leq d \right\}$$

Το υπολογιστικό πρόβλημα είναι καλά ορισμένο (well posed) εάν $c(f, x) < \infty$ και καλά ορισμένο στο $D$ εάν $c(f, x) < \infty\ \forall x \in D$. Ο αριθμός $c(f, x)$ ονομάζεται δείκτης κατάστασης του προβλήματος.

# Κεφάλαιο 3

# Μετασχηματισμοί Gauss

## 3.1 Παραγοντοποίηση LU

Με τη βοήθεια στοιχειωδών πινάκων μπορούμε να πετύχουμε την τριγωνοποίηση ενός πίνακα $A$. Πιο συγκεκριμένα, θα περιγράψουμε το κλασσικό σχήμα απαλοιφής γνωστό σα σχήμα απαλοιφής του Gauss.

Χρειάζεται να προσδιορίσουμε μία διαδικασία που να **μηδενίζει** συγκεκριμένες εισόδους διανυσμάτων. Π.χ. για $n = 2$ εάν $\underline{x} = [x_1, x_2]^t$, $x_1 \neq 0$ και εάν $m = \frac{x_2}{x_1}$ τότε:

$$\begin{pmatrix} 1 & 0 \\ -m & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ 0 \end{pmatrix}$$

**Ορισμός:** Εστω $\underline{x} \in \mathbb{R}^n, x_k \neq 0, \underline{m}^t = [\overbrace{0, \ldots, 0}^{k}, m_{k+1}, \ldots, m_n]$, $m_i = \frac{x_i}{x_k}$, $i = k+1, \ldots, n$ Ο στοιχειώδης κάτω τριγωνικός πίνακας

$$M_k = I - \underline{m} \cdot \underline{e}_k^t = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & \vdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & 0 & \cdots & \vdots \\ \vdots & \vdots & \vdots & -m_{k+1} & 1 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & -m_n & 0 & \cdots & 1 \end{pmatrix}$$

καλείται **μετασχηματισμός Gauss**, τα στοιχεία $m_{k+1}, \ldots, m_n$ καλούνται **πολλαπλασιαστές**. Το διάνυσμα $\underline{m}$ καλείται διάνυσμα Gauss. Το στοιχείο $x_k$ ονομάζεται **οδηγό** στοιχείο.

**Λήμμα 3.1.1** *Εστω* $\underline{a}^t = [a_1, \ldots, a_n], a_1 \neq 0$, *υπάρχει ένας μετασχηματισμός Gauss* $M$ *έτσι ώστε* $M \cdot \underline{a}$ *είναι πολλαπλάσιο του* $\underline{e}_1$.

**Απόδειξη:** Εάν ορίσουμε

$$M = \begin{pmatrix} 1 & 0 & \cdots & \vdots \\ -\frac{a_2}{a_1} & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ -\frac{a_n}{a_1} & 0 & \cdots & 1 \end{pmatrix}$$

τότε

$$M \cdot \underline{a} = \begin{pmatrix} a_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

□

## 3.2 Μέθοδος απαλοιφής του Gauss χωρίς οδήγηση

Εστω $A$ ένας $n \times n$ πίνακας. Χρησιμοποιώντας μετασχηματισμούς Gauss μπορούμε να πετύχουμε την τριγωνοποίησή του ως εξής:

Εστω $A = A^{(0)} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$

**Βήμα 1:** Προσδιορίζουμε κατάλληλο μετασχηματισμό Gauss $M_1$ έτσι ώστε:

$$M_1 \cdot \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} = \begin{pmatrix} a_{11} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$M_1 = \begin{pmatrix} 1 & 0 & & \\ -m_{21} & 1 & & \\ \vdots & \vdots & \ddots & \\ -m_{n1} & 0 & & 1 \end{pmatrix}, \quad m_{i1} = \frac{a_{i1}}{a_{11}}, \quad i = 2, \ldots, n$$

Επιδρούμε μ' αυτόν στον πίνακα $A^{(0)}$ και προκύπτει

$$A^{(1)} = M_1 \cdot A^{(0)} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & & \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{pmatrix}$$

**Βήμα 2:** Προσδιορίζουμε κατάλληλο μετασχηματισμό Gauss $\hat{M}_2$ έτσι ώστε

$$\hat{M}_2 \cdot \begin{pmatrix} a_{22}^{(1)} \\ \vdots \\ \vdots \\ a_{n2}^{(1)} \end{pmatrix} = \begin{pmatrix} a_{22}^{(1)} \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix},$$

$$\hat{M}_2 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -m_{32} & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ -m_{n2} & 0 & & 1 \end{pmatrix}, \quad m_{i2} = \frac{a_{i2}^{(1)}}{a_{22}^{(1)}}, \quad i = 3, \ldots, n$$

Εάν θέσουμε

$$M_2 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \hat{M}_2 & \\ 0 & & & \end{pmatrix}$$

και επιδράσουμε πάνω στον πίνακα $A^{(1)}$ προκύπτει:

$$A^{(2)} = M_2 \cdot A^{(1)} = \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1n} \\ 0 & a_{22}^{(1)} & \cdots & \cdots & a_{2n}^{(1)} \\ \vdots & 0 & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & a_{n3}^{(2)} & \cdots & a_{nn}^{(2)} \end{pmatrix}$$

Παρατηρούμε ότι ο πολλαπλασιασμός εξ αριστερών του $A^{(1)}$ με τον $M_2$ διατηρεί τις προηγούμενες μηδενικές εισόδους.

**Βήμα $k$:** Προσδιορίζουμε κατάλληλο μετασχηματισμό Gauss $\hat{M}_k$ έτσι ώστε:

$$\hat{M}_k \cdot \begin{pmatrix} a_{kk}^{(k-1)} \\ a_{k+1,k}^{(k-1)} \\ \vdots \\ a_{n,k}^{(k-1)} \end{pmatrix} = \begin{pmatrix} a_{kk}^{(k-1)} \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

$$\hat{M}_k = \begin{pmatrix} 1 & 0 & & \\ -m_{k+1,k} & 1 & & \\ \vdots & \vdots & \ddots & \\ -m_{n,k} & 0 & & 1 \end{pmatrix}, \quad m_{i,k} = \frac{a_{i,k}^{(k-1)}}{a_{k,k}^{(k-1)}}, \quad i = k+1, \ldots, n$$

Εάν θέσουμε

$$M_k = \begin{pmatrix} I_{k-1} & O \\ O & \hat{M}_k \end{pmatrix}$$

και επιδράσουμε πάνω στον πίνακα $A^{(k-1)}$ προκύπτει

$$A^{(k)} = M_k \cdot A^{(k-1)} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} & \cdots & \cdots & a_{1n} \\ 0 & a_{22}^{(1)} & \cdots & a_{2k}^{(1)} & \cdots & \cdots & a_{2n}^{(1)} \\ \vdots & 0 & \cdots & a_{kk}^{(k-1)} & \cdots & \cdots & a_{kn}^{(k-1)} \\ \vdots & \vdots & \cdots & 0 & a_{k+1k+1}^{(k)} & \cdots & a_{k+1n}^{(k)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & a_{nk+1}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix}$$

**Βήμα** $n-1$**:** Στο τέλος του $(n-1)$ βήματος ο πίνακας $A^{(n-1)}$ είναι άνω τριγωνικός και έχει τη μορφή:

$$A^{(n-1)} = \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1n} \\ 0 & a_{22}^{(1)} & \cdots & \cdots & a_{2n}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \cdots & a_{nn}^{(n-1)} \end{pmatrix}$$

## Προσδιορισμός των πινάκων $L$ και $U$

Από την εφαρμογή των προηγούμενων βημάτων προκύπτει ότι

$$A^{(n-1)} = M_{n-1} \cdot A^{(n-2)} = M_{n-1} \cdot M_{n-2} \cdot A^{(n-3)} = \ldots =$$

$$M_{n-1} \cdot M_{n-2} \cdot M_{n-3} \ldots \cdot M_2 \cdot M_1 A$$

Θέτουμε

$$U = A^{(n-1)}, L_1 = M_{n-1} M_{n-2} \ldots M_2 M_1$$

Τότε $U = L_1 \cdot A$

Ο πίνακας $L_1$ είναι κάτω τριγωνικός με μονάδες στη διαγώνιο (αφού ορίζεται σα γινόμενο τέτοιων πινάκων) επομένως ορίζεται και υπάρχει ο αντίστροφος $L_1^{-1}$ και μπορεί να εκφρασθεί ως εξής:
$L_1^{-1} = M_1^{-1} M_2^{-1} \ldots M_{n-1}^{-1}$ όπου $M_i^{-1} = I + \underline{m}_i \cdot \underline{e}_i^t$ και

$$\underline{m}_i = [0, \ldots, 0, m_{i+1,i}, \ldots, m_{n,i}]^t$$

Τελικά εάν θέσουμε:

$$L = L_1^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ m_{21} & 1 & 0 & 0 & 0 \\ m_{31} & m_{32} & 1 & \vdots & \vdots \\ \vdots & \vdots & & \ddots & \ddots \\ m_{n1} & m_{n2} & & m_{nn-1} & 1 \end{pmatrix}$$

προκύπτει $A = L \cdot U$.

Η παραγοντοποίηση αυτή του $A$ είναι γνωστή σαν $L \cdot U$ παραγοντοποίηση **(LU factorisation)**

**Θεώρημα 3.2.1 (Θεώρημα LU)**: *Εστω $A$ ένας $n \times n$ πίνακας με όλες τις κύριες υποορίζουσες διάφορες του μηδενός. Τότε ο $A$ έχει μοναδική $LU$ παραγοντοποίηση:*

$$A = L \cdot U$$

*όπου ο $L$ είναι κάτω τριγωνικός με μονάδες στη διαγώνιο και $U$ είναι άνω τριγωνικός. Η ορίζουσα του $A(\det(A))$ δίνεται από τη σχέση*

$$\det(A) = u_{11} \dots u_{nn}$$

**Απόδειξη:**

**(i) Υπαρξη της LU**

Εστω ότι έχουν γίνει $k-1$ βήματα του αλγορίθμου. Στην αρχή του βήματος $k$ έχει υπολογισθεί ο πίνακας $A^{(k-1)} = M_{k-1} \dots M_1 A$. Το στοιχείο $a_{kk}^{(k-1)}$ είναι το $k$-οδηγό στοιχείο. Επειδή οι μετασχηματισμοί Gauss που επέδρασαν στον $A$ είναι μοναδιαίοι κάτω τριγωνικοί, παρατηρώντας το $k \times k$ μέρος του πίνακα $A^{(k-1)}$ προκύπτει ότι εάν $A_k$ είναι ο υποπίνακας του $A$ που αποτελείται από τις πρώτες $k$ γραμμές και στήλες του, τότε $\det(A_k) = a_{11} a_{22}^{(1)} \dots a_{kk}^{(k-1)}$. Εάν $\det(A_k) \neq 0$ τότε το κ-οδηγό στοιχείο είναι διάφορο του μηδενός. Ετσι εάν $\det(A_k) \neq 0, \;\; k = 1, 2, \dots, n$ η LU παραγοντοποίηση υπάρχει πάντα.

**(ii) Μοναδικότητα της LU**

Εάν $A = L_1 U_1$ και $A = L_2 U_2$ τότε $L_2^{-1} \cdot L_1 = U_2 \cdot U_1^{-1}$
Ομως $L_2^{-1} \cdot L_1$ μοναδιαίος κάτω τριγωνικός και $U_2 \cdot U_1^{-1}$ άνω τριγωνικός, επομένως $L_2^{-1} \cdot L_1 = U_2 \cdot U_1^{-1} = I \Rightarrow L_2 = L_1, U_2 = U_1$

**(iii) Ορίζουσα του $A$**
Αφού $A = L \cdot U$ τότε

$$\det(A) = \det(L \cdot U) = \det(L) \cdot \det(U) = \det(U) = u_{11} u_{22} \dots u_{nn}$$

□

**Παρατήρηση:** Εάν ο $n \times n$ πίνακας $A$ είναι μη αντιστρέψιμος (κάποια κύρια υποορίζουσά του είναι 0) η $LU$ παραγοντοποίησή του υπολογίζεται όμως κάποιο διαγώνιο στοιχείο στον άνω τριγωνικό πίνακα $U$ θα ισούται με 0.

**Παράδειγμα 3.2.1** *Να προσδιοριστεί η $LU$ παραγοντοποίηση του πίνακα*

$$A = \begin{pmatrix} 2 & 2 & 3 \\ 4 & 5 & 6 \\ 1 & 2 & 4 \end{pmatrix}$$

**Απόδειξη:** **Βήμα 1:** Υπολογισμός του $M_1$
Οι πολλαπλασιαστές είναι: $m_{21} = 2, m_{31} = \frac{1}{2}$

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{pmatrix}$$

$$A^{(1)} = M_1 \cdot A = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 2 & 2 & 3 \\ 4 & 5 & 6 \\ 1 & 2 & 4 \end{pmatrix} = \begin{pmatrix} 2 & 2 & 3 \\ 0 & 1 & 0 \\ 0 & 1 & \frac{5}{2} \end{pmatrix}$$

**Βήμα 2:** Υπολογισμός του $M_2$
Οι πολλαπλασιαστές είναι: $m_{32} = 1$

$$\hat{M}_2 = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$$

$$M_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \hat{M}_2 \end{pmatrix}$$

$$A^{(2)} = M_2 \cdot A^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 2 & 2 & 3 \\ 0 & 1 & 0 \\ 0 & 1 & \frac{5}{2} \end{pmatrix} = \begin{pmatrix} 2 & 2 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{5}{2} \end{pmatrix}$$

Ετσι $U = \begin{pmatrix} 2 & 2 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{5}{2} \end{pmatrix}$

Υπολογισμός του $L$

$$L_1 = M_2 \cdot M_1 = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -\frac{1}{2} & -1 & 1 \end{pmatrix}$$

$$L = L_1^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & m_{32} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ \frac{1}{2} & 1 & 1 \end{pmatrix}$$

□

### 3.2.1 Τριγωνοποίηση πίνακα με απαλοιφή Gauss χωρίς οδήγηση

Εστω $A$ ένας $n \times n$ πίνακας. Ο παρακάτω αλγόριθμος υπολογίζει την τριγωνοποίηση του $A$ όποτε υπάρχει. Ο αλγόριθμος επικαλύπτει το πάνω τριγωνικό μέρος του $A$ συμπεριλαμβανομένης και της διαγωνίου με το $U$ οι δε είσοδοι του $A$ κάτω από τη διαγώνιο επικαλύπτονται με τους πολλαπλασιαστές που χρειάζονται για τον υπολογισμό του $L$.

**Αλγόριθμος** $LU$**:**

**for** $k = 1, 2, \ldots, n-1$
Βήμα 1: Δημιουργία των πολλαπλασιαστών:
$a_{ik} \equiv m_{ik} = \frac{a_{ik}}{a_{kk}},\ \ i = k+1, \ldots, n$
Βήμα 2: Ενημέρωση των εισόδων του πίνακα
$a_{ij} \equiv a_{ij} - m_{ik}a_{kj},\ \ i = k+1, \ldots, n,\ \ j = k+1, \ldots, n$

**Αποτελεσματικότητα:** Ο αλγόριθμος απαιτεί περίπου $\frac{n^3}{3}$ flops.
Συγκεκριμένα, για $k = 1$ υπολογίζουμε $n-1$ πολλαπλασιαστές και ενημερώνουμε $(n-1)^2$ εισόδους του $A$. Κάθε πολλαπλασιαστής και κάθε ενημέρωση απαιτούν 1 flop. Ετσι για $k = 1$ χρειάζονται $[(n-1)^2 + (n-1)]$ flops.
Παρόμοια για $k = 2$ απαιτούνται $[(n-2)^2 + (n-2)] flops$. Γενικά το $k$ βήμα απαιτεί $[(n-k)^2 + (n-k)]$ flops. Επειδή συνολικά έχουμε (ν-1) βήματα προκύπτει ότι:
Συνολικό αριθμός $flops = \sum_{k=1}^{n-1}(n-k)^2 + \sum_{k=1}^{n-1}(n-k) =$

$$\frac{n(n-1)(2n-1)}{6} + \frac{n(n-1)}{2} \cong [\frac{n^3}{3} + O(n^2)]$$

### 3.2.2 Διανυσματική μορφή Αλγορίθμου $LU$

**for** $k = 1, 2, \ldots, n-1$
$A(k+1:n, k) = \frac{A(k+1:n,k)}{A(k,k)}$
$A(k+1:n, k+1:n) = A(k+1:n, k+1:n) -$
$A(k+1:n, k) \cdot A(k, k+1:n)$

## 3.3 Μετασχηματισμοί Gauss-Jordan

**Ορισμός:** Εστω $\underline{x} \in \mathbb{R}^n, x_k \neq 0, \underline{g}^t = [g_1, g_2, \ldots, g_{k-1}, 0, g_{k+1}, \ldots, g_n]$,
$g_i = \frac{x_i}{x_k}, i = 1, \ldots, n\ i \neq k$
Ο στοιχειώδης πίνακας

$$G_k = I - \underline{g} \cdot \underline{e}_k^t = \begin{pmatrix} 1 & 0 & 0 & \cdots & -g_1 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & \vdots & \cdots & 0 \\ 0 & 0 & 1 & \cdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & -g_{k-1} & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & 1 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & -g_n & \cdots & 1 \end{pmatrix}$$

καλείται **μετασχηματισμός Gauss-Jordan**, τα στοιχεία $g_1, \ldots, g_n$ καλούνται **πολλαπλασιαστές**. Το διάνυσμα $\underline{g}$ καλείται διάνυσμα Gauss-Jordan. Το στοιχείο $x_k$ ονομάζεται **οδηγό στοιχείο**.

**Λήμμα 3.3.1** *Εστω* $\underline{a}^t = [a_1, \ldots, a_n], a_k \neq 0$, *υπάρχει ένας μετασχηματισμός Gauss-Jordan* $G$ *έτσι ώστε* $G_k \cdot \underline{a}$ *είναι πολλαπλάσιο του* $\underline{e}_k$.

**Απόδειξη:** Εάν ορίσουμε

$$G_k = \begin{pmatrix} 1 & 0 & \cdots & -\frac{a_1}{a_k} & \cdots & 0 \\ 0 & 1 & \cdots & -\frac{a_2}{a_k} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & 1 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\frac{a_n}{a_k} & \cdots & 1 \end{pmatrix}$$

τότε

$$G_k \cdot \underline{a} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ a_k \\ \vdots \\ 0 \end{pmatrix}$$

□

### 3.3.1 Μέθοδος απαλοιφής του Gauss-Jordan χωρίς οδήγηση

Εστω $A$ ένας $n \times n$ πίνακας. Χρησιμοποιώντας μετασχηματισμούς Gauss-Jordan μπορούμε να πετύχουμε τη διαγωνοποίησή του ως εξής:

Εστω $A = A^{(0)} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$

**Βήμα 1:** Προσδιορίζουμε κατάλληλο μετασχηματισμό Gauss-Jordan $G_1$ έτσι ώστε:

$$G_1 \cdot \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} = \begin{pmatrix} a_{11} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$G_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -g_{21} & 1 & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ -g_{n1} & 0 & \cdots & 1 \end{pmatrix}, \; g_{i1} = \frac{a_{i1}}{a_{11}}, \;\; i = 2, \ldots, n$$

Επιδρούμε μ' αυτόν στον πίνακα $A^{(0)}$ και προκύπτει

$$A^{(1)} = G_1 \cdot A^{(0)} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{pmatrix}$$

**Βήμα 2:** Προσδιορίζουμε κατάλληλο μετασχηματισμό Gauss-Jordan $G_2$ έτσι ώστε

$$G_2 \cdot \begin{pmatrix} a_{12} \\ a_{22}^{(1)} \\ \vdots \\ \vdots \\ a_{n2}^{(1)} \end{pmatrix} = \begin{pmatrix} 0 \\ a_{22}^{(1)} \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix},$$

$$G_2 = \begin{pmatrix} 1 & -g_{12} & 0 & \cdots & 0 \\ 0 & 1 & 0 & \ddots & \vdots \\ 0 & -g_{32} & 1 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & -g_{n2} & 0 & & 1 \end{pmatrix}, \quad g_{i2} = \frac{a_{i2}^{(1)}}{a_{22}^{(1)}}, \quad i = 1, \ldots, n, \quad i \neq 2$$

Εάν επιδράσουμε πάνω στον πίνακα $A^{(1)}$ προκύπτει:

$$A^{(2)} = G_2 \cdot A^{(1)} = \begin{pmatrix} a_{11} & 0 & \cdots & \cdots & a_{1n} \\ 0 & a_{22}^{(1)} & \cdots & \cdots & a_{2n}^{(1)} \\ \vdots & 0 & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & a_{n3}^{(2)} & \cdots & a_{nn}^{(2)} \end{pmatrix}$$

Παρατηρούμε ότι ο πολλαπλασιασμός εξ αριστερών του $A^{(1)}$ με τον $G_2$ διατηρεί τις προηγούμενες μηδενικές εισόδους.

**Βήμα** $k$: Προσδιορίζουμε κατάλληλο μετασχηματισμό Gauss-Jordan $G_k$ έτσι ώστε:

$$G_k \cdot \begin{pmatrix} a_{1k} \\ \vdots \\ a_{kk}^{(k-1)} \\ \vdots \\ a_{n,k}^{(k-1)} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ a_{kk}^{(k-1)} \\ \vdots \\ 0 \end{pmatrix},$$

$$G_k = \begin{pmatrix} 1 & 0 & \ldots & -g_{1k} & \ldots & 0 \\ 0 & 1 & \ldots & -g_{2k} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \vdots & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & -g_{nk} & \ldots & 1 \end{pmatrix}, \quad g_{ik} = \frac{a_{i,k}^{(k-1)}}{a_{k,k}^{(k-1)}}, \quad i = 1, \ldots, n \quad i \neq k$$

Εάν επιδράσουμε πάνω στον πίνακα $A^{(k-1)}$ προκύπτει

$$A^{(k)} = G_k \cdot A^{(k-1)} = \begin{pmatrix} a_{11} & 0 & \cdots & 0 & \cdots & \cdots & a_{1n} \\ 0 & a_{22}^{(1)} & \cdots & 0 & \cdots & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & a_{kk}^{(k-1)} & \cdots & \cdots & a_{kn}^{(k-1)} \\ \vdots & \vdots & \cdots & 0 & a_{k+1k+1}^{(k)} & \cdots & a_{k+1n}^{(k)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & a_{nk+1}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix}$$

**Βήμα** $n$: Στο τέλος του $n$ βήματος ο πίνακας $A^{(n)}$ είναι διαγώνιος και έχει τη μορφή:

$$A^{(n)} = \begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{22}^{(1)} & 0 & \cdots & 0 \\ 0 & 0 & a_{33}^{(2)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & a_{nn}^{(n-1)} \end{pmatrix}$$

### 3.3.2 Προσδιορισμός του πίνακα $A^{-1}$

Από την εφαρμογή των προηγούμενων βημάτων προκύπτει ότι

$$A^{(n)} = G_n \cdot A^{(n-1)} = G_n \cdot G_{n-1} \cdot A^{(n-2)} = \ldots =$$

$$G_n \cdot G_{n-1} G_{n-2} \ldots G_2 G_1 A$$

Θέτουμε

$$D = A^{(n)}, A_1 = G_n G_{n-1} G_{n-2} \ldots G_2 G_1$$

Τότε $D = A_1 \cdot A$ Από τη σχέση αυτή προκύπτει ότι:
$D^{-1} \cdot D = D^{-1} \cdot A_1 \cdot A$ και κατά συνέπεια $I_n = D^{-1} \cdot A_1 \cdot A$ Επομένως ο πίνακας $D^{-1} \cdot A_1$ μας δίνει τον αντίστροφο του πίνακα $A$.

**Παράδειγμα 3.3.1** *Να προσδιοριστεί η Gauss-Jordan παραγοντοποίηση του πίνακα*

$$A = \begin{pmatrix} 2 & 2 & 3 \\ 4 & 5 & 6 \\ 1 & 2 & 4 \end{pmatrix}$$

**Απόδειξη: Βήμα 1:** Υπολογισμός του $G_1$

Οι πολλαπλασιαστές είναι: $g_{21} = 2, g_{31} = \frac{1}{2}$

$$G_1 = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{pmatrix}$$

$$G^{(1)} = G_1 \cdot A = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 2 & 2 & 3 \\ 4 & 5 & 6 \\ 1 & 2 & 4 \end{pmatrix} = \begin{pmatrix} 2 & 2 & 3 \\ 0 & 1 & 0 \\ 0 & 1 & \frac{5}{2} \end{pmatrix}$$

**Βήμα 2:** Υπολογισμός του $G_2$

Οι πολλαπλασιαστές είναι: $g_{12} = 2, \quad g_{32} = 1$

$$G_2 = \begin{pmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}$$

$$A^{(2)} = G_2 \cdot A^{(1)} = \begin{pmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 2 & 2 & 3 \\ 0 & 1 & 0 \\ 0 & 1 & \frac{5}{2} \end{pmatrix} = \begin{pmatrix} 2 & 0 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{5}{2} \end{pmatrix}$$

**Βήμα 3:** Υπολογισμός του $G_3$

Οι πολλαπλασιαστές είναι: $g_{13} = \frac{6}{5}$

$$G_3 = \begin{pmatrix} 1 & 0 & -\frac{6}{5} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$A^{(3)} = G_3 \cdot A^{(2)} = \begin{pmatrix} 1 & 0 & -\frac{6}{5} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 2 & 0 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{5}{2} \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{5}{2} \end{pmatrix}$$

Ετσι $D = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{5}{2} \end{pmatrix}$

Υπολογισμός του $A^{-1}$

$$A^{-1} = D^{-1} \cdot G_3 \cdot G_2 \cdot G_1 = \begin{pmatrix} \frac{8}{5} & -\frac{2}{5} & -\frac{3}{5} \\ -2 & 1 & 0 \\ \frac{3}{5} & -\frac{2}{5} & \frac{2}{5} \end{pmatrix}$$

□

**Παρατήρηση:** Η ακολουθία των μετασχηματισμών Gauss-Jordan όταν δρα επί του $A$ παράγει τον $I_n$, η ίδια ακολουθία όταν δράσει επί του $[A \mid I_n]$ παράγει τον πίνακα $[I_n \mid A^{-1}]$

**Αλγόριθμος Gauss-Jordan**: Διαγωνοποίηση πίνακα χρησιμοποιώντας απαλοιφή Gauss-Jordan χωρίς οδήγηση.

Εστω $A$ ένας $n \times n$ πίνακας. Ο παρακάτω αλγόριθμος υπολογίζει τη διαγωνοποίησή του όποτε υπάρχει.

**for** $k = 1, 2, \ldots, n$
Βήμα 1: Δημιουργία των πολλαπλασιαστών:
$g_{ik} = \frac{a_{ik}}{a_{kk}}, \;\; i = 1, \ldots, n, \;\; i \neq k$
Βήμα 2: Ενημέρωση των εισόδων του πίνακα
$a_{ij} \equiv a_{ij} - g_{ik} a_{kj}, \;\; i = 1, \ldots, n, \;\; i \neq k \;\; j = k+1, \ldots, n$

**Αποτελεσματικότητα:** Ο αλγόριθμος απαιτεί περίπου $\frac{n^3}{2}$ flops. Συγκεκριμένα σε κάθε βήμα υπολογίζουμε $n-1$ πολλαπλασιαστές και ενημερώνουμε $(n-1) \cdot (n-k)$ εισόδους του $A$. Κάθε πολλαπλασιαστής και κάθε ενημέρωση απαιτούν 1 flop. Ετσι συνολικά χρειάζονται

$$flops = \sum_{k=1}^{n} (n-1) + \sum_{k=1}^{n} (n-k)(n-1) \cong [\frac{n^3}{2} + O(n^2)]$$

## 3.4 Η Τεχνικής της Οδήγησης

**Παρατήρηση:** Η μέθοδος Gauss χωρίς οδήγηση μπορεί να είναι ασταθής

Θεωρούμε το ακόλουθο κλασσικό παράδειγμα.
Εστω ότι εφαρμόζουμε απαλοιφή Gauss χωρίς οδήγηση στον πίνακα

$$A = \begin{pmatrix} 0.0001 & 1 \\ 1 & 1 \end{pmatrix}$$

[Παρατηρούμε ότι $k(A) = \|A\|_\infty \|A^{-1}\|_\infty \simeq 4$. Δηλαδή, ο πίνακας είναι σε καλή κατάσταση]

Απαιτείται ένα μόνο βήμα

$$m_{21} = +\frac{1}{10^{-4}} = +10^4 \quad \text{(πολύ μεγάλη τιμή)}$$

$$U = A^{(1)} = \begin{pmatrix} 0.0001 & 1 \\ 0 & 1 - 10^4 \end{pmatrix}$$

$$L = \begin{pmatrix} 1 & 0 \\ 10^4 & 1 \end{pmatrix}$$

Συγκεκριμένα εάν εφαρμόσουμε αριθμητική κινητής υποδιαστολής υπολογίζουμε τα ακόλουθα στοιχεία

$$fl(l_{21}) = fl(\frac{1}{0.0001}) = (0.1 * 10^{-3})^{-1} = 10^4$$

$$fl(u_{22}) = fl(1 - 10^4) = (1 - 10^4)(1 + e) = (1 - 10^4) + (1 - 10^4)\varepsilon,\ |\varepsilon| \leq u$$

Ετσι έχουμε ότι οι πίνακες $\hat{L} = fl(l_{ij}), \hat{U} = fl(u_{ij})$ που υπολογίσαμε ικανοποιούν τη σχέση

$$\hat{L}\hat{U} = \begin{pmatrix} 1 & 0 \\ 10^4 & 1 \end{pmatrix} \begin{pmatrix} 0.0001 & 1 \\ 0 & (1 - 10^4)(1 + e) \end{pmatrix} = \mathrm{LU} + E = A + E,$$

$$||E||_\infty \leq (1 - 10^4)u$$

Το φράγμα στον πίνακα σφάλματος $E$ παίρνει αρκετά μεγάλη τιμή και επομένως παρατηρούμε ότι η τιμή του γινομένου $\hat{L}\hat{U}$ που υπολογίζουμε μπορεί να διαφέρει από τον $A$ σημαντικά και επομένως η μέθοδος δεν είναι ευσταθής.
(Εάν θεωρήσουμε $t = 3$ τότε $fl(u_{22}) = -10^4$ και

$\hat{L}\hat{U} = \begin{pmatrix} 0.0001 & 1 \\ 1 & 0 \end{pmatrix}$)

Τι φταίει γι΄ αυτό·

Παρατηρούμε ότι η τιμή του οδηγού $a_{11}^{(1)} = 0.0001$ είναι πολύ μικρή και μάλιστα για $t = 3$ μπορεί να θεωρηθεί και ίση με μηδέν. Η μικρή τιμή του οδηγού έδωσε πολύ μεγάλη τιμή στον πολλαλασιαστή με αποτέλεσμα την εισαγωγή μεγάλου σφάλματος στρογγύλευσης που μπορεί να οδηγήσει και στην εξάλειψη μικρών ποσοτήτων (π.χ. για $t = 3$, $fl(1 - 10^4) = -10^4$, παρατηρούμε ότι αγνοείται τελείως η 1)

Το πρόβλημα αυτό μπορεί να ξεπεραστεί εάν κάνουμε κατάλληλες εναλλαγές γραμμών. Εστω ότι εναλλάσσουμε τις δύο γραμμές του πίνακα $A$ παίρνοντας

$$A' = \begin{pmatrix} 1 & 1 \\ 0.0001 & 1 \end{pmatrix}$$

Παρατηρούμε τώρα ότι

$$m_{21} = 10^{-4} \ \ (|m_{21}| < 1)$$

$$U = A^{(1)} = \begin{pmatrix} 1 & 1 \\ 0 & 1 - 10^{-4} \end{pmatrix}$$

$$L = \begin{pmatrix} 1 & 0 \\ 0.0001 & 1 \end{pmatrix}$$

Συγκεκριμένα εάν εφαρμόσουμε αριθμητική κινητής υποδιαστολής υπολογίζουμε τα ακόλουθα στοιχεία

$$fl(l_{21}) = fl(10^{-4}) = 10^{-4}$$

$$fl(u_{22}) = fl(1 - 10^{-4}) = (1 - 10^{-4})(1 + \varepsilon), \ \ |\varepsilon| \leq u$$

Ετσι οι πίνακες $\hat{L} = fl(l_{ij}), \hat{U} = fl(u_{ij})$ που υπολογίζουμε ικανοποιούν τη σχέση

$$\hat{L}\hat{U} = \begin{pmatrix} 1 & 0 \\ 0.0001 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & (1 - 10^{-4})(1 + e) \end{pmatrix} = \mathrm{LU} + E = A + E,$$

$$||E||_\infty \leq (1 - 10^{-4})u$$

Το φράγμα στον πίνακα Ε είναι μικρό και επομένως παρατηρούμε ότι η τιμή του γινομένου $\hat{L}\hat{U}$ που υπολογίσαμε δίνει μία μικρή διατάραξη του πίνακα $A$. (Εάν θεωρήσουμε

$t = 3$, $\hat{L}\hat{U} = \begin{pmatrix} 1 & 1 \\ 0.0001 & 1.0001 \end{pmatrix} = A'$ αφού για $t = 3$, φλ$(1 - 10^{-4})$=1)

### 3.4.1 Μεταθετικοί Πίνακες

Η ευστάθεια της μεθόδου του Gauss περιλαμβάνει μετακινήσεις δεδομένων όπως αλλαγή γραμμών πίνακα. Αυτό πετυχαίνεται με τους λεγόμενους μεταθετικούς πίνακες. Ενας τέτοιος πίνακας είναι ο ταυτοτικός με τις γραμμές του ανακατανεμημένες π.χ.

$$P = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

Ο πίνακας αυτός παριστάνεται με το ακέραιο διάνυσμα , $\underline{p} = (4132)$ που δηλώνει τις θέσεις όπου υπάρχουν μονάδες στον $P$.

Η επίδραση του $P$ σ'έναν πίνακα $A$ επιφέρει το ακόλουθο αποτέλεσμα:

$$PA \rightarrow \text{αλλαγή γραμμών του } A$$

$$AP \rightarrow \text{αλλαγή στηλών του } A$$

Ο πίνακας $P$ ορθογώνιος $\Rightarrow P^{-1} = P^t$
Εστω διάνυσμα $\underline{x}$ το οποίο θέλουμε να επικαλύψουμε από το $P\underline{x}$. Αυτό υλοποιείται με τον ακόλουθο αλγόριθμο:

**for** $k = 1 : n$
    $x(k) = x(p(k))$

**Αλλαγές Γραμμών**

Ο πίνακας

$$P = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

εάν επιδράσει εξάριστερών σ'έναν πίνακα $A$ μεταθέτει την 1η και την 4η γραμμή του. Σε μία αλλαγή γραμμών δεν περιέχεται **σφάλμα στρογγύλευσης**.

## 3.5 Μέθοδος Gauss με Οδήγηση

Από το προηγούμενο παράδειγμα είδαμε ότι εάν κατά τη διάρκεια των μετασχηματισμών Gauss προκύπτει οδηγό στοιχείο του οποίου το μέγεθος είναι πολύ μικρό η μέθοδος μπορεί τελικά να είναι ασταθής. Είναι επομένως απαραίτητο να ελέγχουμε το μέγεθος των οδηγών στοιχείων και να επιλέγουμε ως οδηγό το στοιχείο με το μεγαλύτερο μέγεθος. Ετσι αναζητάμε στον πίνακα το στοιχείο με το μεγαλύτερο μέγεθος το οποίο τοποθετούμε στη θέση του οδηγού.
Εάν για κάθε βήμα της μεθόδου η αναζήτηση αυτή γίνεται ανά στήλη έχουμε την τεχνική της μερικής οδήγησης (partial pivoting) ενώ εάν κάθε φορά γίνεται σε ολόκληρο τον πίνακα έχουμε την τεχνική της ολικής οδήγησης (complete pivoting).

Η εισαγωγή της οδήγησης καθίσταται αναγκαία γιατί έτσι έχοντας πάντα πολλαπλασιαστές με τιμή μικρότερη της μονάδας ελέγχεται το μέγεθος των στοιχείων που εμφανίζονται σε κάθε βήμα της μεθόδου Gauss και αποφεύγονται οι πολύ μεγάλες τιμές που πιθανόν να οδηγήσουν τη μέθοδο σε αστάθεια.

## 3.6 Μέθοδος Gauss με Μερική Οδήγηση

Εστω $A = A^{(0)} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$

**Βήμα 1:**
Εντοπίστε στην πρώτη στήλη του $A$ το μεγαλύτερο σε μέγεθος στοιχείο. Εστω ότι είναι το $a_{r_{1,1}}$

- Κατασκευάστε ένα μεταθετικό πίνακα $P_1$ αλλάζοντας τις γραμμές 1 και $r_1$ του ταυτοτικού πίνακα και αφήνοντας τις άλλες γραμμές του αμετάβλητες.
- Κατασκευάστε τον $P_1$ αλλάζοντας τις γραμμές $r_1$ και 1 του $A$.
- Προσδιορίστε ένα στοιχειώδη κάτω τριγωνικό πίνακα $M_1$ έτσι ώστε ο $A^{(1)} = M_1 P_1 A$ να έχει μηδενικά κάτω από την (1,1) είσοδο της πρώτης στήλης.

Ο πίνακας $M_1$ αρκεί να κατασκευαστεί έτσι ώστε

$$M_1 \begin{pmatrix} a_{11} \\ \vdots \\ \vdots \\ a_{n1} \end{pmatrix} = \begin{pmatrix} * \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Ας σημειώσουμε ότι ο

$$M_1 = \begin{pmatrix} 1 & 0 & 0 & \cdots & \cdots & 0 \\ -m_{21} & 1 & 0 & \cdots & \cdots & 0 \\ -m_{31} & 0 & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & \\ -m_{n1} & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

όπου $m_{i1} = \frac{a_{i1}}{a_{11}},\ i = 2, \ldots, n$, και με $a_{ij}$ αναφερόμαστε στην $(i, j)$ είσοδο του μετατιθεμένου πίνακα $P_1 A$. Αποθηκεύουμε τους πολλαπλασιαστές $m_{i1},\ i = 2, \ldots, n$ και καταγράφουμε τις αλλαγές των γραμμών.

$$A^{(1)} = \begin{pmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & & \vdots \\ 0 & * & \cdots & * \end{pmatrix}$$

**Βήμα 2:** Εντοπίστε στη δεύτερη στήλη του $A^{(1)}$ και κάτω από τη πρώτη γραμμή το μεγαλύτερο σε μέγεθος στοιχείο. Εστω ότι είναι το στοιχείο $a^{(1)}_{r_2,2}$.

- Κατασκευάστε ένα μεταθετικό πίνακα $P_2$ αλλάζοντας τις γραμμές 2 και $r_2$ του ταυτοτικού πίνακα και αφήνοντας τις άλλες γραμμές του αμετάβλητες.
- Κατασκευάστε τον $P_2 A^{(1)}$
- Προσδιορίστε έναν στοιχειώδη κάτω τριγωνικό πίνακα $M_2$ έτσι ώστε ο $A^{(2)} = M_2 P_2 A^{(1)}$ να έχει μηδενικά κάτω από την (2,2) είσοδο της δεύτερης στήλης.

Ο $M_2$ κατασκευάζεται ως εξής. Κατάρχήν κατασκευάζουμε έναν στοιχειώδη πίνακα $\hat{M}_2$ τάξης $(n-1)$ έτσι ώστε

$$\hat{M}_2 \begin{pmatrix} a_{22} \\ a_{32} \\ \vdots \\ a_{n2} \end{pmatrix} = \begin{pmatrix} * \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Στη συνέχεια ορίζουμε

$$M_2 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \hat{M}_2 & \\ 0 & & & \end{pmatrix}$$

Ας σημειώσουμε ότι με $a_{ij}$ αναφερόμαστε στην $(i, j)$ είσοδο του τρέχοντα πίνακα $P_2 A^{(1)}$. Στο τέλος του βήματος 2 έχουμε

$$A^{(2)} = M_2 P_2 A^{(1)} = \begin{pmatrix} * & * & \cdots & \cdots & * \\ 0 & * & \cdots & \cdots & * \\ 0 & 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & * & \cdots & * \end{pmatrix}$$

$$M_2 = \begin{pmatrix} 1 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & -m_{32} & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & & \ddots \\ & & & & \ddots & \\ 0 & -m_{n2} & 0 & \cdots & 0 & 1 \end{pmatrix}$$

όπου $m_{i2} = \frac{a_{i2}}{a_{22}}$, $i = 3, 4, \ldots, n$
Αποθηκεύουμε τους πολλαπλασιαστές $m_{i2}$ και καταγράφουμε τις αλλαγές γραμμών.
**Βήμα** $k$**:** Γενικά, στο $k$ βήμα εντοπίστε στις εισόδους της $k$ στήλης του πίνακα $A^{(k-1)}$ και κάτω από τη $(k-1)$ γραμμή το μέγιστο σε τιμή στοιχείο $a^{(k-1)}_{r_k,k}$

- Κατασκευάστε ένα μεταθετικό πίνακα $P_k$ αλλάζοντας τις γραμμές $k$ και $r_k$ του ταυτοτικού πίνακα και αφήνοντας τις άλλες γραμμές του αμετάβλητες.
- Κατασκευάστε τον $P_k A^{(k-1)}$
- Προσδιορίστε έναν στοιχειώδη κάτω τριγωνικό πίνακα $M_k$ έτσι ώστε ο $A^{(k)} = M_k P_k A^{(k-1)}$ να έχει μηδενικά κάτω από την $(k,k)$ είσοδο της $k$ στήλης.

Ο $M_k$ κατασκευάζεται ως εξής. Κατασκευάζουμε πρώτα τον πίνακα $\hat{M}_k$ τάξης $(n-k+1)$ έτσι ώστε

$$\hat{M}_k \begin{pmatrix} a_{kk} \\ \vdots \\ \vdots \\ a_{nk} \end{pmatrix} = \begin{pmatrix} * \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

και στη συνέχεια ορίζουμε τον

$$M_k = \begin{pmatrix} I_{k-1} & 0 \\ 0 & \hat{M}_k \end{pmatrix}$$

όπου 0 είναι ένας πίνακας μηδενικών. Με $a_{i,k}$ αναφερόμαστε στις $(i,k)$ εισόδους του πίνακα $P_k A^{(k-1)}$.
**Βήμα** $n-1$**:** Στο τέλος του $(n-1)$ βήματος ο πίνακας $A^{(n-1)}$ θα είναι άνω τριγωνικός.

**Δημιουργία του πίνακα** $U$**:**

Θέτουμε

$$A^{(n-1)} = U$$

Τότε

$$U = A^{(n-1)} = M_{n-1} P_{n-1} A^{(n-2)}$$
$$= M_{n-1} P_{n-1} M_{n-2} P_{n-2} A^{(n-3)} = \ldots$$

$$= M_{n-1}P_{n-1}M_{n-2}P_{n-2}\dots M_2P_2M_1P_1A$$

Θέτουμε

$$M_{n-1}P_{n-1}M_{n-2}P_{n-2}\dots M_2P_2M_1P_1 = M$$

Από τα προηγούμενα προκύπτει η ακόλουθη παραγοντοποίηση του $A$:

$$U = MA$$

**Θεώρημα 3.6.1** *(**Παραγοντοποίηση Μερικής Οδήγησης**) Εστω ένας $n\times n$ πίνακας $A$. Οταν εφαρμοσθεί σ' αυτόν απαλοιφή Gauss με μερική οδήγηση προκύπτει ένας άνω τριγωνικός πίνακας $U$ και ένας πίνακας $M$ ο οποίος με κατάλληλες εναλλαγές γίνεται κάτω τριγωνικός, έτσι ώστε*

$$MA = U$$

*όπου*

$$A^{(n-1)} = U,$$

$$M = M_{n-1}P_{n-1}M_{n-2}P_{n-2}\dots M_2P_2M_1P_1$$

Εάν ορίσουμε

$$P = P_{n-1}\dots P_2P_1$$

$$L = P(M_{n-1}P_{n-1}\dots M_1P_1)^{-1}$$

τότε $PA = LU$.

**Πόρισμα 3.6.1** *(LU **Παραγοντοποίηση με μερική οδήγηση**) Εστω ένας $n \times n$ πίνακας $A$. Οταν εφαρμοσθεί σ' αυτόν απαλοιφή Gauss με μερική οδήγηση προκύπτει η εξής LU παραγοντοποίηση :*

$$PA = LU$$

*όπου $P = P_{n-1}\dots P_2P_1$ ένας μεταθετικός πίνακας, $L = P(M_{n-1}P_{n-1}\dots M_1P_1)^{-1}$ ένας μοναδιαίος κάτω τριγωνικός πίνακας, και $U = A^{(n-1)}$ ένας άνω τριγωνικός πίνακας.*

**Παράδειγμα 3.6.1** *Τριγωνοποιείστε τον παρακάτω πίνακα $A$ χρησιμοποιώντας μερική οδήγηση.*

$$A = \begin{pmatrix} 0.0001 & 1 \\ 1 & 1 \end{pmatrix}$$

*Απαιτείται μόνο ένα βήμα. Το οδηγό στοιχείο είναι 1, και $r_1 = 2$.*

$$P_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$P_1 A = \begin{pmatrix} 1 & 1 \\ 0.0001 & 1 \end{pmatrix}$$

$$m_{21} = \frac{0.0001}{1} = 10^{-4}$$

$$M_1 = \begin{pmatrix} 1 & 0 \\ -m_{21} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -10^{-4} & 1 \end{pmatrix}$$

$$M_1 P_1 A = \begin{pmatrix} 1 & 0 \\ -10^{-4} & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0.0001 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 0.999 \end{pmatrix} = U$$

$$M = M_1 P_1 = \begin{pmatrix} 1 & 0 \\ -10^{-4} & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -10^{-4} \end{pmatrix}$$

**Παράδειγμα 3.6.2** *Χρησιμοποιώντας μερική οδήγηση, τριγωνοποιείστε τον παρακάτω πίνακα*

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 1 & 1 \end{pmatrix}$$

*Εκφράστε τον πίνακα $A$ στη μορφή $A = MU$. Προσδιορίστε επίσης πίνακες $P$ και $L$ έτσι ώστε $PA = LU$.*

**Απόδειξη:** **Βήμα 1:** Το οδηγό στοιχείο είναι $a_{21} = 1$, και $r_1 = 2$.

$$P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$P_1 A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}$$

$$A^{(1)} = M_1 P_1 A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 1 \\ 0 & -1 & -2 \end{pmatrix}$$

**Βήμα 2:** Το οδηγό στοιχείο είναι $a_{22} = 1$

$$P_2 A^{(1)} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 1 \\ 0 & -1 & -2 \end{pmatrix}$$

$$P_2 = I_3 \quad (\text{δε χρειάζονται εναλλαγές})$$

$$\hat{M}_2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

$$M_2 = \begin{pmatrix} I_1 & 0 \\ 0 & \hat{M}_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

$$U = A^{(2)} = M_2 P_2 A^{(1)} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 1 \\ 0 & 0 & -1 \end{pmatrix}$$

$$M = M_2 P_2 M_1 P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & -1 & 1 \end{pmatrix}$$

Εύκολα επαληθεύεται ότι $A = MU$.
Διαμόρφωση των πινάκων $L$ και $P$:

$$P = P_2 P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$L = P(M_2 P_2 M_1 P_1)^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix}$$

Αμεσα επαληθεύεται ότι $PA = LU$. □

### Διαμόρφωση του αλγορίθμου της μερικής οδήγησης-Παρατηρήσεις

- Κάθε μεταθετικός πίνακας $P_k$ μπορεί να διαμορφωθεί καταγράφοντας μόνο το δείκτη $r_k$, επειδή ο $P_k$ είναι ο ταυτοτικός πίνακας με αλλαγμένες τις γραμμές $k$ και $r_k$. Επιπρόσθετα δε χρειάζεται να υπολογίζονται αναλυτικά οι πίνακες $P_k$ και $P_kA^{(k-1)}$ αφού ο πίνακας $P_kA^{(k-1)}$ απλώς προκύπτει εναλλάσοντας τις γραμμές $r_k$ και $k$ του $A^{(k-1)}$.
- Κάθε στοιχειώδης πίνακας $M_k$ μπορεί να διαμορφωθεί αναλυτικά εάν έχουν αποθηκευθεί οι $(n-k)$ πολλαπλασιαστές. Επίσης δε χρειάζεται να υπολογίζονται αναλυτικά οι πίνακες $M_kP_kA^{(k-1)} = M_kB$ αφού τα στοιχεία στις πρώτες $k$ γραμμές του πίνακα $M_kB$ είναι ακριβώς τα ίδια με τα στοιχεία στις πρώτες $k$ γραμμές του πίνακα $B$, και τα στοιχεία στις υπόλοιπες $(n-k)$ γραμμές δίνονται από τη σχέση:

$$b_{ij} + m_{ik}b_{kj} \quad (i = k+1, \ldots, n \quad j = k+1, \ldots, n)$$

- Οι πολλαπλασιαστές μπορούν να αποθηκεύονται σε κατάλληλες θέσεις στο κάτω τριγωνικό μέρος του πίνακα $A$.
- Ο τελικός άνω τριγωνικός πίνακας $U = A^{(n-1)}$ αποθηκεύεται στο άνω τριγωνικό μέρος του $A$.
- Οι δείκτες των οδηγών στοιχείων $r_k$ αποθηκεύονται σ'ένα ξεχωριστό μονοδιάστατο ακέραιο πεδίο.
- Κάθε πίνακας $A^{(k)}$ επικαλύπτει τον πίνακα $A$.

Λαμβάνοντας υπ'όψη τα προηγούμενα μπορεί να προκύψει ο ακόλουθος αλγόριθμος για την $LU$ παραγοντοποίηση με μερική οδήγηση.

### Αλγόριθμος: Τριγωνοποίηση πίνακα εφαρμόζοντας απαλοιφή Gauss με μερική οδήγηση

Εστω $A$ ένας $n \times n$ μη ιδιάζων πίνακας. Ο ακόλουθος αλγόριθμος υπολογίζει την τριγωνοποίηση του $A$ με εναλλαγμένες τις γραμμές του χρησιμοποιώντας απαλοιφή Gauss με μερική οδήγηση. Το άνω τριγωνικό μέρος του $U$ αποθηκεύεται στο άνω τριγωνικό μέρος του $A$, συμπεριλαμβανομένης και της διαγωνίου. Στο κάτω τριγωνικό μέρος του $A$ αποθηκεύονται οι πολλαπλασιαστές που χρειάστηκαν για τον υπολογισμό του πίνακα $M$ έτσι ώστε $MA = U$. Οι δείκτες των εναλλαγών $r_k$ αποθηκεύονται σε ξεχωριστό πεδίο.

**For** $k = 1, 2, \ldots n-1$
Βήμα 1: Προσδιορίστε δείκτη $r_k$ : $|a_{(r_k,k)}| = max_{1 \le i \le n}|a_{ik}|$.
Αποθηκεύστε τον $r_k$.
**If** $a_{r_k,k} = 0$ then stop.
Βήμα 2: Εναλλαγή των γραμμών $r_k$ και $k$
$a_{kj} \to a_{r_k,j}$ $(j = k, k+1, \ldots, n)$.

Βήμα 3: Διαμόρφωση των πολλαπλασιαστών
$a_{ik} \equiv m_{ik} = -\frac{a_{ik}}{a_{kk}} (i = k+1, \ldots, n)$.
Βήμα 4: Ενημέρωση των εισόδων
$a_{ij} \equiv a_{ij} + m_{ik} a_{kj} = a_{ij} + a_{ik} a_{kj}$ $(i = k+1, \ldots, n; j = k+1, \ldots, n)$

**Αποτελεσματικότητα:** Ο αλγόριθμος απαιτεί για την εκτέλεσή του περίπου $O(n^3/3)$ flops και χρειάζεται $O(n^2)$ συγκρίσεις επειδή για την αναζήτηση του οδηγού στοιχείου στο $k$ βήμα απαιτούνται $(n-k)$ συγκρίσεις. Συνολικά απαιτούνται $(n-1) + (n-2) + \ldots + 1$, συγκρίσεις.

Ας σημειώσουμε ότι ο παραπάνω αλγόριθμος δεν προσδιορίζει αναλυτικά τους πίνακες $M$ και $P$. Παρ΄όλα αυτά εάν χρειάζονται μπορούν εύκολα να κατασκευασθούν από τους πολλαπλασιαστές και τους δείκτες γραμμών που έχουμε κρατήσει. Ο αλγόριθμος αυτός μπορεί να επεκταθεί και για την περίπτωση $m \times n$ πινάκων.

**Παράδειγμα 3.6.3** *Εστω*

$$A = \begin{pmatrix} 1 & 2 & 4 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

*Βήμα 1:* $k = 1$

1. *Το οδηγό στοιχείο είναι* 7, *και* $r_1 = 3$.

2. *Εναλλαγή γραμμών* 3 *και* 1

$$A \equiv \begin{pmatrix} 7 & 8 & 9 \\ 4 & 5 & 6 \\ 1 & 2 & 4 \end{pmatrix}$$

3. *Διαμόρφωση των πολλαπλασιαστών* :

$$a_{21} \equiv m_{21} = -\frac{4}{7}, \quad a_{31} \equiv m_{31} = -\frac{1}{7}$$

4.*Ενημέρωση* :

$$A = \begin{pmatrix} 7 & 8 & 9 \\ 0 & \frac{3}{7} & \frac{6}{7} \\ 0 & \frac{6}{7} & \frac{19}{7} \end{pmatrix}$$

*Βήμα 2:* $\kappa$*=2*

1. *Το οδηγό στοιχείο είναι* $\frac{6}{7}$, *και* $r_2 = 3$

2. *Εναλλαγή γραμμών* 2 *και* 3

$$A = \begin{pmatrix} 7 & 8 & 9 \\ 0 & \frac{6}{7} & \frac{19}{7} \\ 0 & \frac{3}{7} & \frac{6}{7} \end{pmatrix}$$

3. *Διαμόρφωση των πολλαπλασιαστών* :

$$m_{32} = -\frac{1}{2}$$

4. *Ενημέρωση* :

$$A \equiv \begin{pmatrix} 7 & 8 & 9 \\ 0 & \frac{6}{7} & \frac{19}{7} \\ 0 & 0 & -\frac{1}{2} \end{pmatrix}$$

*Διαμόρφωση του* $M$*:*

$$M = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & -\frac{1}{7} \\ -\frac{1}{2} & 1 & -\frac{1}{2} \end{pmatrix}$$

## 3.7 Μέθοδος Gauss με Ολική Οδήγηση

Στην απαλοιφή Gauss με ολική οδήγηση στο $k$ βήμα η αναζήτηση του οδηγού στοιχείου γίνεται μεταξύ των εισόδων του υποπίνακα κάτω από τις πρώτες $(k-1)$ γραμμές. Ετσι εάν το οδηγό στοιχείο είναι το $a_{rs}$, για να το φέρουμε στην $(k,k)$ θέση πρέπει να εναλλάξουμε τις γραμμές $r$ και $k$ καθώς και τις στήλες $k$ και $s$. Αυτό ισοδυναμεί με εξάριστερών και εκ'δεξιών πολλαπλασιασμό του πίνακα $A^{(k-1)}$ με τους μεταθετικούς πίνακες $P_k$ και $Q_k$ αντίστοιχα. Στη συνέχεια εφαρμόζουμε στον πίνακα $P_k A^{(k-1)} Q_k$ τη συνήθη απαλοιφή Gauss δηλαδή υπολογίζουμε έναν στοιχειώδη κάτω τριγωνικό πίνακα $M_k$ έτσι ώστε ο πίνακας

$$A^{(k)} = M_k P_k A^{(k-1)} Q_k$$

να έχει μηδενικά στη $k$ στήλη κάτω από την $(k,k)$ είσοδο. Στο τέλος του $(n-1)$ βήματος ο πίνακας $A^{(n-1)}$ είναι άνω τριγωνικός.
Θέτουμε

$$A^{(n-1)} = U$$

Τότε

$$U = A^{(n-1)} = M_{n-1} P_{n-1} A^{(n-1)} Q_{n-1}$$
$$= M_{n-1} P_{n-1} M_{n-2} P_{n-2} A^{(n-3)} Q_{n-2} Q_{n-1}$$

$$= \ldots = M_{n-1}P_{n-1}M_{n-2}P_{n-2}\ldots M_1P_1AQ_1Q_2\ldots Q_{n-1}$$

Θέτουμε

$$M_{n-1}P_{n-1}\ldots M_1P_1 = M$$
$$Q_1\ldots Q_{n-1} = Q$$

Τότε έχουμε

$$U = MAQ$$

**Θεώρημα 3.7.1** *(**Παραγοντοποίηση Ολικής Οδήγησης**) Εστω ένας $n\times n$ πίνακας $A$. Οταν εφαρμοσθεί σ' αυτόν απαλοιφή Gauss με μερική οδήγηση προκύπτει ένας άνω τριγωνικός πίνακας $U$, ένας πίνακας $M$ ο οποίος με κατάλληλες εναλλαγές γίνεται κάτω τριγωνικός, και ένας μεταθετικός πίνακας $Q$ έτσι ώστε*

$$MAQ = U$$

*όπου*

$$U = A^{(n-1)}$$
$$M = M_{n-1}P_{n-1}\ldots M_1P_1$$
$$Q = Q_1\ldots Q_{n-1}$$

**Πόρισμα 3.7.1** *(LU* **Παραγοντοποίηση με ολική οδήγηση**) *Εστω ένας $n\times n$ πίνακας $A$. Οταν εφαρμοσθεί σ' αυτόν απαλοιφή Gauss με ολική οδήγηση προκύπτει η εξής LU παραγοντοποίηση :*

$$PAQ = LU$$

*όπου $P$ και $Q$ είναι μεταθετικοί πίνακες που δίνονται από τις σχέσεις*

$$P = P_{n-1}\ldots P_1$$

$$Q = Q_1\ldots Q_{n-1}$$

*και $L$ είναι κάτω τριγωνικός με μονάδες στη διαγώνιο που ισούται με*

$$L = P(M_{n-1}P_{n-1}\ldots M_1P_1)^{-1}$$

**Παράδειγμα 3.7.1** *Να τριγωνοποιηθεί ο πίνακας*

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 1 & 1 \end{pmatrix}$$

*χρησιμοποιώντας ολική οδήγηση*

**Απόδειξη:**

Βήμα 1 : $k = 1$. Το οδηγό στοιχείο είναι $a_{23} = 3$

$$P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$Q_1 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad P_1AQ_1 = \begin{pmatrix} 3 & 2 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{1}{3} & 1 & 0 \\ -\frac{1}{3} & 0 & 1 \end{pmatrix}$$

$$A^{(1)} = M_1P_1AQ_1 = \begin{pmatrix} 3 & 2 & 1 \\ 0 & \frac{1}{3} & -\frac{1}{3} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

Βήμα 2: $k = 2$. Το οδηγό στοιχείο είναι $a_{33}^{(1)} = \frac{2}{3}$.

$$P_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \quad Q_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

$$P_2A^{(1)}Q_2 = \begin{pmatrix} 3 & 1 & 2 \\ 0 & \frac{2}{3} & \frac{1}{3} \\ 0 & -\frac{1}{3} & \frac{1}{3} \end{pmatrix} \quad \hat{M}_2 = \begin{pmatrix} 1 & 0 \\ \frac{1}{2} & 1 \end{pmatrix}$$

$$M_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{2} & 1 \end{pmatrix}$$

$$U = A^{(2)} = M_2P_2A^{(1)}Q_2 = M_2P_2(M_1P_1AQ_1)Q_2 =$$

$$\begin{pmatrix} 3 & 1 & 2 \\ 0 & \frac{2}{3} & \frac{1}{3} \\ 0 & 0 & \frac{1}{2} \end{pmatrix}$$

□

**Αλγόριθμος: Τριγωνοποίηση πίνακα εφαρμόζοντας απαλοιφή Gauss με ολική οδήγηση**

Εστω $A$ ένας $n \times n$ μη ιδιάζων πίνακας. Ο ακόλουθος αλγόριθμος υπολογίζει την τριγωνοποίηση του $A$ με εναλλαγμένες τις γραμμές και στήλες του χρησιμοποιώντας απαλοιφή Gauss με ολική οδήγηση. Το άνω τριγωνικό μέρος του $U$ αποθηκεύεται στο άνω τριγωνικό μέρος του $A$, συμπεριλαμβανομένης και της διαγωνίου. Στο κάτω τριγωνικό μέρος του $A$ αποθηκεύονται οι πολλαπλασιαστές που χρειάστηκαν. Οι δείκτες των εναλλαγών $r_k$, $s_k$ αποθηκεύονται σε ξεχωριστό πεδίο.

**For** $k = 1, 2, \ldots, n-1$

Βήμα 1: Προσδιορίστε δείκτες $r_k$ και $s_k$ έτσι ώστε
$|a_{r_k,s_k}| = \max\{|a_{ij}| : i, j \geq k\}$,
Αποθηκεύστε τους $r_k$ και $s_k$
**If** $a_{r_k,s_k} = 0$, then stop.

Βήμα 2: Εναλλαγή των γραμμών $r_k$ και $k$
$a_{kj} \to a_{r_k,j}$ $(j = k, k+1, \ldots, n)$

Βήμα 3: Εναλλαγή των στηλών $s_k$ και $k$
$a_{ik} \to a_{i,s_k}$ $(i = 1, 2, \ldots n)$

Βήμα 4: Διαμόρφωση των πολλαπλασιαστών
$a_{ik} \equiv m_{ik} = \frac{a_{ik}}{a_{kk}}$ $(i = k+1, \ldots, n)$

Βήμα 5: Ενημέρωση των εισόδων του $A$
$a_{ij} \equiv a_{ij} - m_{ik}a_{kj} = a_{ij} - a_{ik}a_{kj}$
$(i = k+1, \ldots, n; j = k+1, \ldots, n)$

Ο παραπάνω αλγόριθμος δεν υπολογίζει αναλυτικά τους πίνακες $M, P$ και $Q$. Εάν αυτοί χρειάζονται μπορούν άμεσα να διαμορφωθούν από τους πολλαπλασιαστές $m_{ik}$ και τους δείκτες μεταθέσεων $r_k$ και $s_k$. Ο αλγόριθμος μπορεί να επεκταθεί και για την περίπτωση $m \times n$ πινάκων.

**Αποτελεσματικότητα:** Ο αλγόριθμος απαιτεί $O(n^3/3)$ flops και $O(n^3/3)$ συγκρίσεις. Σε κάθε βήμα έχουμε $n^2 + (n-1)^2 + \ldots + 2^2 + 1 \approx O(\frac{2n^3}{6})$ συγκρίσεις.

**Παράδειγμα 3.7.2** *Εστω*

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

*Μόνο ένα βήμα απαιτείται*
*κ=1: Το οδηγό στοιχείο είναι 4.*

$$r_1 = 2$$

$$s_1 = 2$$

*Αντιμεταθέτουμε τη δεύτερη και πρώτη γραμμή και στη συνέχεια αντιμεταθέτουμε τη δεύτερη και πρώτη στήλη έτσι ώστε το οδηγό στοιχείο 4 να βρεθεί στη θέση (1,1) του πίνακα:*

$$A \equiv \begin{pmatrix} 4 & 3 \\ 2 & 1 \end{pmatrix}$$

*Ο παλλαπλασιαστής είναι* $a_{21} \equiv m_{21} = \frac{a_{21}}{a_{11}} = \frac{2}{4} = \frac{1}{2}$. *Τελικά αφού ενημερώσουμε τις εισόδους του πίνακα προκύπτει*

$$A \equiv \begin{pmatrix} 4 & 3 \\ 0 & -\frac{1}{2} \end{pmatrix}$$

$$P_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad Q = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$M = M_1 P_1 = \begin{pmatrix} 1 & 0 \\ -\frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -\frac{1}{2} \end{pmatrix}$$

**Παρατήρηση:** Εάν σ' έναν πίνακα εκτελέσουμε εναλλαγές των γραμμών και των στηλών του τότε η εφαρμογή ολικής οδήγησης σ' αυτόν μπορεί να δίνει διαφορετική δομή των οδηγών στοιχείων που προκύπτουν. Για παράδειγμα εάν θεωρήσουμε τον πίνακα

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 \end{pmatrix}$$

Εάν εφαρμοσθεί Gauss με ολική οδήγηση σ' αυτόν τον πίνακα προκύπτει η ακόλουθη δομή οδηγών στοιχείων:

$$(1, 2, 2, 4, \frac{5}{2})$$

Εάν εκτελέσουμε με τη σειρά τις ακόλουθες εναλλαγές στον πίνακα: αλλαγή 1ης και 5ης στήλης, αλλαγή 2ης και 5ης στήλης, αλλαγή 2ης και 5ης γραμμής, αλλαγή 4ης και 5ης γραμμής, και εφαρμοσθεί Gauss με ολική οδήγηση σ΄ αυτόν τον πίνακα παίρνουμε την ακόλουθη δομή οδηγών στοιχείων:

$$(1, 2, 2, 3, \frac{10}{3})$$

**Παρατηρήση:** Ισοδύναμοι κατά γραμμές πίνακες μπορούν να δώσουν διαφορετική δομή οδηγών στοιχείων.

## 3.8 Τριγωνική Διαχώριση

Έστω $A \in \mathbb{R}^{n\times n}$. Θεωρούμε το πρόβλημα καθορισμού ενός μοναδιαίου κάτω τριγωνικού πίνακα $L$ ενός άνω τριγωνικού πίνακα $U$ έτσι ώστε

$$A = LU \tag{1}$$

Ας υποθέσουμε ότι οι πρώτες $(r-1)$ γραμμές των $L$ και $U$ μπορούν να καθορισθούν εξισώνοντας τα στοιχεία των πρώτων $(r-1)$ γραμμών και των δύο πλευρών της εξίσωσης (1). Τότε εξισώνοντας τα στοιχεία στην $r$ γραμμή έχουμε

$$\left.\begin{array}{lcl} l_{r1}u_{11} & = & a_{r1} \\ l_{r1}u_{12} + l_{r2}u_{22} & = & a_{r2} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & \\ l_{r1}u_{r1} + l_{r2}u_{2r} + \ldots + l_{rr}u_{rr} & = & a_{rr} \end{array}\right\} \tag{2}$$

$$\left.\begin{array}{lllllllll} l_{r1}u_{1,r+1} & + & l_{r2}u_{2,r+1} & + & \ldots & + & l_{rr}u_{r,r+1} & = & a_{r,r+1} \\ \ldots\ldots\ldots\ldots & & \ldots\ldots\ldots\ldots & & \ldots\ldots & & \ldots\ldots\ldots\ldots & & \ldots\ldots \\ l_{r1}u_{1n} & + & l_{r2}u_{2n} & + & \ldots & + & l_{rr}u_{rn} & = & a_{rn} \\ l_{r1}c_1 & + & l_{r2}c_2 & + & \ldots & + & l_{rr}c_r & = & b_r \end{array}\right\} \tag{3}$$

Από τις εξισώσεις 1 έως $(r-1)$ των σχέσεων (2) καθορίζονται μοναδικά τα στοιχεία $l_{r1}, l_{r2}, \ldots, l_{r,r-1}$. Εάν το στοιχείο $l_{rr}$ επιλεγεί αυθαίρετα, η $r$ εξίσωση καθορίζει το $u_{rr}$. Οι υπόλοιπες εξισώσεις καθορίζουν μοναδικά τα στοιχεία $u_{r,r+1}$ έως $u_{rn}$. Το αποτέλεσμα αυτό είναι προφανές για τα στοιχεία της 1ης γραμμής και επομένως είναι αληθές γενικά.

Βλέπομε ότι γινόμενο $l_{rr}u_{rr}$ καθορίζεται μοναδικά αλλά ή το $l_{rr}$ ή το $u_{rr}$ θα πρέπει να επιλεγεί αυθαίρετα. Εάν θέλουμε ο $L$ να είναι και μοναδιαίος κάτω τριγωνικός τότε $l_{rr} = 1$. Στην περίπτωση αυτή, οι μόνες διαιρέσεις που περιέχονται είναι αυτές στις οποίες καθορίζονται τα $l_{ri}$ $(i = 1, \ldots, r-1)$ και οι διαιρέτες είναι τα $u_{ii}$. Η παραγοντοποίηση αυτή δε μπορεί να αποτύχει εάν $u_{ii} \neq 0$, και τότε μοναδική.

Στην παραπάνω διαδικασία τα στοιχεία καθορίζονται με την εξής σειρά:

Πρώτη γραμμή του $U$, δεύτερη γραμμή του $L$
δεύτερη γραμμή του $U$, τρίτη γραμμή του $L$ κ.ο.κ

ή

Πρώτη γραμμή του $U$, δεύτερη στήλη του $L$
δεύτερη γραμμή του $U$, τρίτη στήλη του $L$ κ.ο.κ

### 3.8.1 Σχέση μεταξύ Απαλοιφής Gauss και Τριγωνικής Διαχώρισης

Έχουμε αποδείξει ότι εαν ο πίνακας $A$ είναι μη ιδιάζων τότε η παραγωντοποίση $LU$ του $A$ εάν υπάρχει είναι μοναδική και επομένως οι πίνακες $L$ και $U$ εάν προσδιορισθούν με τη μέθοδο της τριγωνικής διαχώρισης θα είναι ίδιοι με τους πίνακες $L_1^{-1}$ και $A^{n-1}$ που προκύπτουν από τη μέθοδο των μετασχηματισμών Gauss.
Για παράδειγμα ας θεωρήσουμε τον καθορισμό του $l_{r4}$ από τη σχέση

$$l_{r4} = (a_{r4} - l_{r1}u_{14} - l_{r2}u_{24} - l_{r3}u_{34})/u_{44} \tag{4}$$

Μια σύγριση με τους μετασχηματισμούς Gauss μας δείχνει ότι

$$\begin{aligned} a_{r4}^{(0)} &= a_{r4} \\ a_{r4}^{(1)} &= a_{r4} - l_{r1}u_{14} \\ a_{r4}^{(2)} &= a_{r4} - l_{r1}u_{14} - l_{r2}u_{24} \\ a_{r4}^{(3)} &= a_{r4} - l_{r1}u_{14} - l_{r2}u_{24} - l_{r3}u_{34} \end{aligned}$$

και επομένως καθορίζοντας τον αριθμητή της (4) βρίσκουμε με τη σειρά κάθε ένα από τα στοιχεία στις $(r,4)$ θέσεις των πινάκων $A^{(1)}, A^{(2)}, A^{(3)}$. Παρόμοια σχόλια ισχύουν και για τον υπολογισμό των στοιχείων του $U$.

#### Διανυσματική Μορφή του Αλγορίθμου Τριγωνικής Διαχώρισης

**For** $j = 1 : n$
  Επιλύστε $L(1 : j-1, 1 : j-1)U(1 : j-1, j) = A(1 : j-1, j)$
  $u(j : n) = A(j : n, j) - L(j : n, 1 : j-1)U(1 : j-1, j)$
  $U(j, j) = u(j)$
  $L(j+1 : n, j) = u(j+1 : n)/U(j, j)$
**end**

## ΑΣΚΗΣΕΙΣ

**1.** Να αποδειχθεί ότι στο $k$-βήμα της απαλοιφής Gauss με μερική οδήγηση ισχύει $|a_{ij}^{(k)}| \leq 2^{k-1}$ αν $|a_{ij}^{(1)}| \leq 1$

**Απόδειξη:** Τα στοιχεία στη μέθοδο του Gauss παράγονται ως εξής:

$$a_{ij}^{(k+1)} = \begin{cases} 0 & , i \geq k+1,\ j = k \\ a_{ij}^{(k)} - m_{i,k}a_{kj}^{k} & , i \geq k+1,\ j \geq k+1 \\ a_{ij}^{(k)} & , \text{διαφορετικά} \end{cases}$$

για $k=0$ ισχύει $a_{ij}^1 \leq 1$
για $k=1,2$ έχουμε

$$|a_{ij}^{(2)}| = \begin{cases} 0 \\ |a_{ij}^{(1)} - m_{i,1}a_{1j}^1| \leq 1+1 = 2^{1(=k-1)} \\ |a_{ij}^{(1)}| \leq 1 \end{cases}$$

$$|a_{ij}^{(3)}| = \begin{cases} 0 \\ |a_{ij}^{(2)} - m_{i,2}a_{2j}^2| \leq 2+2 = 4 = 2^{2(=k-1)} \\ |a_{ij}^{(2)}| \leq 2 < 2^2 \end{cases}$$

Έστω ότι στο $k$-βήμα ισχύει $|a_{ij}^{(k)}| \leq 2^{k-1}$, τότε $k+1$-βήμα έχουμε

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

□

**2.** Έστω $T$ ένας $n \times n$ άνω τριγωνικός πίνακας. Να δοθεί αλγόριθμος υπολογισμού του αντιστρόφου του. Να εκτιμηθεί η υπολογιστική πολυπλοκότητά του.

**Απόδειξη:** Έστω $S = (\underline{s}_1 \underline{s}_2 \dots \underline{s}_n)$ ο αντίστροφος πίνακας. Επειδή $TS = I_n = (\underline{e}_1 \underline{e}_2 \dots \underline{e}_n)$, χρειάζεται να επιλύσουμε τα συστήματα:

$$T \cdot \underline{s}_i = \underline{e}_i, \ i = 1, \dots n$$

Έστω $\underline{s}_i = (s_{1i}, s_{2i}, \dots, s_{ni})^T$
Για $i=1$

$$T \cdot \underline{s}_i = e_i \Rightarrow s_{11} = \frac{1}{t_{11}}$$

Για $i=2$

$$T\underline{s}_2 = \underline{e}_2 \Rightarrow s_{22} = \frac{1}{t_{22}}, \ s_{12} = -\frac{1}{t_{11}}(t_{12}s_{22})$$

Για $i=k$

$$T\underline{s}_k = \underline{e}_k \Rightarrow s_{kk} = \frac{1}{t_{kk}}, \ s_{ik} = -\frac{1}{t_{ii}}(t_{i,i+1}s_{i+1,k} + \dots + t_{ik}s_{ik})$$
$$i = k-1, k-2, \dots, 1.$$

**Αλγόριθμος**

**For** $k = n, n-1, \dots, 1$
$s_{kk} = \frac{1}{t_{kk}}$
$s_{ik} = -\frac{1}{t_{ii}} \sum_{j=i+1}^{k} t_{ij}s_{jk}$, $i = k-1, k-2, \dots, 1$

**Πολυπλοκότητα**

Για $k = 1$ έχουμε 1 flop

Για $k = 2$ έχουμε 3 flops

Για $k = 3$ έχουμε 6 flops

$\vdots$

Για $k = n$ έχουμε $\frac{n(n+1)}{2}$ flops

Συνολικός αριθμός flops

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

**Παράδειγμα**

$$T = \begin{pmatrix} 5 & 2 & 3 \\ 0 & 2 & 1 \\ 0 & 0 & 4 \end{pmatrix}$$

Για $k = 3$

$$\begin{aligned} s_{33} &= 1/4 \\ s_{23} &= -1/t_{22}(t_{23}s_{33}) = -1/8 \\ s_{13} &= -1/t_{11}(t_{12}s_{23} + t_{13}s_{33}) = -1/10 \end{aligned}$$

Για $k = 2$

$$\begin{aligned} s_{22} &= 1/t_{12} = 1/2 \\ s_{12} &= -1/t_{11}(t_{12}s_{22}) = -1/10 \end{aligned}$$

Για $k = 1$

$$s_1 = 1/t_{11} = 1/5$$

Τελικά

$$T = \begin{pmatrix} 1/5 & -1/5 & -1/10 \\ 0 & 1/2 & -1/8 \\ 0 & 0 & 1/4 \end{pmatrix}$$

□

# Κεφάλαιο 4

# Ανάλυση Σφάλματος Αριθμητικών Μεθόδων

## 4.1 Μέθοδος απαλοιφής του Gauss

Οπως είναι γνωστό το πρώτο και σπουδαιότερο βήμα στη μέθοδο του Gauss είναι η ανάλυση του πίνακα $A$ σε γινόμενο δύο τριγωνικών πινάκων $L$ και $U$.
Υποθέτουμε ότι στον πίνακα $A$ έχει αρχικά εφαρμοσθεί κάποιο scaling στις γραμμές του ώστε δε χρειάζεαι οδήγηση. Στην πράξη, αυτό δε συμβαίνει πάντοτε αλλά επειδή η οδήγηση περιλαμβάνει μόνο μεταθέσεις γραμμών δεν επηρεάζει καθόλου την ανάλυση σφάλματος κι έτσι στα επόμενα θα την αγνοούμε.
Κατά την ανάλυση του πίνακα $A \in \mathbb{R}^{n\times n}$ υπολογίζεται μία ακολουθία πινάκων $A^{(1)} = A, A^{(2)}, \ldots, A^{(n)}$, όπου ο πίνακας $A^{(k)}$ έχει μηδενικά κάτω από τη διαγώνιο στις πρώτες $k-1$ στήλες. Ο πίνακας $A^{(k+1)}$ βρίσκεται από τον $A^{(k)}$ αφαιρώντας ένα πολ/σιο της $k$ γραμμής από όλες τις επόμενές της γραμμές το δε υπόλοιπο κομμάτι του $A^{(k)}$ παραμένει αναλλοίωτο. Οι πολ/στές διαλέγονται κατά τέτοιο τρόπο ώστε αν δεν υπήρχαν σφάλματα στρογγύλευσης (rounding errors) ο πίνακας $A^{(k+1)}$ θα έχει μηδενικά κάτω από τη διαγώνιο στη $k$ στήλη.
**Προσοχή:** Αυτά τα στοιχεία δεν θα τα υπολογίζουμε αλλά θα τα παίρνουμε ίσα με το μηδέν εξόρισμού.

Πιο αναλυτικά, έστω ότι ο πίνακας $A^{(k)}$ έχει στοιχεία $a_{ij}^{(k)}$ Τότε έστω:

$$m_{i,k} = fl(\frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}),\ \ i \le k+1 \tag{4.1}$$

$$a_{ij}^{(k+1)} = \begin{cases} 0 & ,\ i \ge k+1,\ \ j = k \\ fl(a_{i,j}^{(k)} - m_{i,k}\cdot a_{k,j}^{(k)}) & ,\ i \ge k+1,\ \ j \ge k+1 \\ a_{i,j}^{(k)} & \text{διαφορετικά} \end{cases}$$

Τα βήματα αυτά εκτελούνται για $k = 1, 2, \ldots, n-1$

Τελικά έστω $U = A^{(n)}$ και

$$L = \begin{pmatrix} 1 & & & & & \\ m_{21} & 1 & & & & \\ m_{31} & m_{32} & 1 & & & \\ \vdots & \vdots & \vdots & & & \\ m_{n1} & m_{n2} & m_{n3} & \cdots & \cdots & 1 \end{pmatrix} \tag{4.2}$$

Οι πίνακες $L$ και $U$ είναι κάτω και άνω τριγωνικοί αντίστοιχα.

Σύμφωνα με τη τεχνική της backward error analysis πρέπει να αποδείξουμε ότι

**1)** $L \cdot U = A + E$ όπου $E$ είναι ένα πίνακας με μικρά στοιχεία και ο οποίος προκύπτει εξαιτίας των σφαλμάτων στρογγύλευσης.
**2)** Πρέπει να φράξουμε κατάλληλα τον πίνακα $E$.

Ας εξετάσουμε τώρα τα δύο προηγούμενα χωριστά.

**1)** Το σφάλμα στον υπολογισμό των πολ/στών $m_{i,k}$ εκφράζεται ως εξής:

$$m_{i,k} = \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}(1+\varepsilon), \;\; |\varepsilon| \leq u \tag{4.3}$$

ή

$$0 = a_{i,k}^{(k)} - m_{i,k} \cdot a_{k,k}^{(k)} + a_{i,k}^{(k)}\varepsilon \tag{4.4}$$

Επομένως

$$\varepsilon_{i,k}^{(k)} = a_{i,k}^{(k)}\varepsilon, \;\; i \geq k+1, \;\; |\varepsilon| \leq u$$

είναι το σφάλμα που προκύπτει όταν θέτουμε το $a_{i,k}^{(k+1)}$ ίσο με μηδέν.
Για τα άλλα στοιχεία από τον τύπο (4.1) προκύπτει

$$a_{i,j}^{(k+1)} = fl(a_{i,j}^{(k)} - fl(m_{i,k} \cdot a_{k,j}^{(k)})) = fl(a_{i,j}^{(k)} - m_{i,k} \cdot a_{k,j}^{(k)} \cdot (1+\varepsilon_1))$$

=

$$\frac{(a_{i,j}^{(k)} - m_{i,k} \cdot a_{k,j}^{(k)}(1+\varepsilon_1))}{(1+\varepsilon_2)}, \;\; |\varepsilon_1|, |\varepsilon_2| \leq u \tag{4.5}$$

Από την (4.5) προκύπτει:

$$a_{i,j}^{(k+1)}(1+\varepsilon_2) = a_{i,j}^{(k)} - m_{i,k} \cdot a_{k,j}^{(k)} - m_{i,k} \cdot a_{k,j}^{(k)}\varepsilon_1 \Rightarrow$$

$$a_{i,j}^{(k+1)} = a_{i,j}^{(k)} - m_{i,k} \cdot a_{k,j}^{(k)} - m_{i,k} \cdot a_{k,j}^{(k)}\varepsilon_1 - a_{i,j}^{(k+1)}\varepsilon_2 \tag{4.6}$$

Επομένως

$$\varepsilon_{i,j}^{(k)} = -a_{i,j}^{(k+1)}\varepsilon_2 - m_{i,k} \cdot a_{k,j}^{(k)} \cdot \varepsilon_1 \tag{4.7}$$

$$i \geq k+1,\ j \geq k+1,\ |\varepsilon_1|, |\varepsilon_2| \leq u,\quad k = 1, 2, \ldots, n-1$$

Ετσι το $\varepsilon_{i,j}^{(k)}$ που δίνεται από την (4.7) είναι το σφάλμα στον υπολογισμό του $a_{i,j}^{(k+1)}$ για το καινούργιο κομμάτι του $A^{(k+1)}$. Το υπόλοιπο μέρος του $A^{(k+1)}$ παίρνεται απ' ευθείας από τον $A^{(k)}$ κι έτσι δεν υπάρχει σφάλμα.
Συνοπτικά

$$\varepsilon_{i,j}^{(k)} = \begin{cases} a_{i,k}^{(k)}\varepsilon & i \geq k+1, j = k \\ -a_{i,j}^{(k+1)}\varepsilon_2 - m_{i,k} \cdot a_{k,j}^{(k)} \cdot \varepsilon_1 & i \geq k+1, j \geq k+1 \\ 0 & \text{διαφορετικά} \end{cases} \tag{4.8a}$$

Ετσι εάν $E^{(k)}$ είναι πίνακας με στοιχεία $\varepsilon_{i,j}^{(k)}$ και έστω

$$L^{(k)} = \begin{pmatrix} & 0 & \\ & 0 & \\ & \vdots & \\ 0 & 0 & 0 \\ & m_{k+1,k} & \\ & m_{k+2,k} & \\ & \vdots & \\ & m_{n,k} & \end{pmatrix}$$

τότε η εξίσωση

$$A^{(k+1)} = A^{(k)} - L^{(k)}A^{(k)} + E^{(k)} \tag{4.8}$$

περιγράφει πλήρως ένα βήμα της ανάλυσης, περικλείει δε και το σφάλμα στρογγύλευσης.
Εάν αθροίσουμε τις σχέσεις (4.8) για $k = 1, 2, \ldots, n-1$ παίρνουμε:

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

$$A^{(1)} + E^{(1)} + E^{(2)} + \ldots + E^{(n-1)} \tag{4.9}$$

Ο πίνακας $L^{(k)}A^{(k)}$ εξαρτάται μόνο από την $k$-γραμμή του $A^{(k)}$ και αυτή η γραμμή είναι ίση με την $k$ γραμμή του $A^{(n)}$. Ετσι έχουμε

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

Συνεπώς

$$L \cdot U = A^{(1)} + E \tag{4.10}$$

όπου $L, U$ ορίζονται από την (4.2) και

$$E = E^{(1)} + E^{(2)} + \ldots + E^{(n-1)} \tag{4.11}$$

είναι το άθροισμα των σφαλμάτων που παρατηρούνται σε κάθε βήμα.

**2)** Είναι φανερό ότι εάν θέλουμε να επιτύχουμε ένα ικανοποιητικό φράγμα για τα $\varepsilon_{i,j}^{(k)}$, χρειαζόμαστε ικανοποιητικά φράγματα για τα $m_{i,k}$ και τα $a_{i,j}^{(k)}$, τα οποία εμφανίζονται στα $\varepsilon_{i,j}^{(k)}$. Η οδήγηση χρησιμοποιείται για να κραταέι αυτά τα φράγματα μικρά και να επιβεβαιώνει τη συνθήκη

$$|m_{i,k}| \leq 1 \text{ για όλα τα } i,k$$

Αυτό πετυχαίνεται είτε με μερική είτε με ολική οδήγηση.
Συμβολίζουμε το μέγιστο στοιχείο σε κάθε πίνακα $|A^{(r)}|$ με $g$. Δέν υπάρχει βλάβη της γενικότητας εάν υποθέσουμε ότι $|a_{i,j}^{(i)}| \leq 1$. Αυτή η συνθήκη μπορεί να επιτευχθεί χρησιμοποιώντας σσαλινγ (χωρίς σφάλμα στρογγύλευσης). Οποια οδήγηση και εάν χρησιμοποιήσουμε, η σχέση $L \cdot U = A^{(1)} + E$ είναι αληθής, δεδομένου ότι με $A^{(1)}$ συμβολίζεται ο αρχικός πίνακας A με τις γραμμές του (ή και τις στήλες του, εάν χρησιμοποιήσουμε ολική οδήγηση) κατάλληλα αλλαγμένες.
Δίνουμε τώρα φράγματα για το σφάλμα που εκφράζεται από την (4.8α)

$$\begin{aligned} |\varepsilon_{i,j}^{(k)}| &= |-a_{i,j}^{(k+1)}\varepsilon_2 - m_{i,k}a_{k,j}^{(k)}\varepsilon_1| \leq gu + gu \\ &\leq 2gu, \quad i \geq k+1, \quad j \geq k+1 \end{aligned} \tag{4.12}$$

$$|\varepsilon_{i,j}^{(k)}| = |a_{i,j}^{(k)}\varepsilon| \leq gu, \;\; i \geq k+1, j = k$$

Ετσι

$$\varepsilon_{i,j}^{(k)} \leq \begin{cases} gu & i \geq k+1, j = k \\ 2gu & i \geq k+1, j \geq k+1 \\ 0 & \text{διαφορετικά} \end{cases}$$

Εάν $|B|$ ο πίνακας με στοιχεία τις απόλυτες τιμές $|b_{ij}|$ των στοιχείων του $B$, τότε

$$|E| \leq guC, \text{ όπου}$$

$$C = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 1 & 2 & \cdots & 2 \\ 1 & 2 & \cdots & 2 \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 2 & \cdots & 2 \end{pmatrix} +$$

$$+ \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 2 & \cdots & 2 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 2 & \cdots & 2 \end{pmatrix} + \ldots + \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ & & & & \\ & & & & \\ 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 2 \end{pmatrix} =$$

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

$$||C||_\infty = \sum_{j=1}^{n} c_{nj} = \sum_{j=1}^{n-1}(2j-1) + 2n - 2 = n^2 - 1$$

$$||E||_\infty \leq gu(n^2-1)$$

Εάν θέσουμε

$$\rho = \frac{max_{i,j,k}|a_{ij}^{(k)}|}{max_{i,j}|a_{ij}|}$$

ο συντελεστής μεγέθυνσης (growth factor) του πίνακα $A$ και έστω $g = max_{i,j}|a_{ij}|$, τότε $||E||_\infty \leq g_1 u\rho(n^2-1)$
Επειδή $g \leq ||A||_\infty$ μπορούμε να πούμε ότι

$$||E||_\infty \leq (n^2-1)\rho u||A||_\infty \leq n^2\rho u||A||_\infty$$

Παρατηρούμε ότι έτσι υπεισέρχεται στο φράγμα του σφάλματος ο παράγοντας growth factor του πίνακα $A$ και επομένως ο παράγοντας αυτός επηρεάζει τις τιμές του σφάλματος.
Από όλα τα προηγούμενα προκύπτει το ακόλουθο θεώρημα.

**Θεώρημα 4.1.1** *Οι πίνακες $L$ και $U$ που υπολογίζονται από τη μέθοδο απαλοιφής του Gauss με οδήγηση, χρησιμοποιώντας floating-point αριθμητική με μονάδα σφάλματος στρογγύλευσης $u$, ικανοποιούν τη σχέση*

$$L \cdot U = A + E$$

*όπου*

$$||E||_\infty \leq n^2\rho u||A||_\infty$$

*Με άλλα λόγια οι πίνακες $L$ και $U$ αποτελούν τη διάσπαση ενός* **ελαφρά διαταραγμένου αρχικού πίνακα**. *Το πόσο μικρή είναι η διατάραξη αυτή εξαρτάται από τη τιμή του συντελεστή μεγέθυνσης $\rho$.*

## 4.2 Μελέτη του Growth Factor

Από τα προηγούμενα παρατηρούμε ότι ο συντελεστής μεγένθυνσης $\rho$ (growth factor) επηρεάζει την ευστάθεια της μεθόδου απαλοιφής Gauss. Είναι επομένως σημαντικό να μελετήσουμε τις τιμές που αυτός μπορεί να πάρει. Για το λόγο αυτό πρέπει να εκτιμήσουμε τις τιμές των στοιχείων που εμφανίζονται στους πίνακες $A^{(k)}$. Ας σημειώσουμε ότι παρόλο που η οδήγηση κρατάει πάντα μικρούς τους πολλαπλασιαστές (μικρότερους της μονάδας) τα στοιχεία στους πίνακες $A^{(k)}$ μπορεί να μεγαλώνουν αυθαίρετα. Δίνουμε τον ακόλουθο ορισμό.

**Ορισμός:** Ο **Συντελεστής μεγέθυνσης (growth factor)** $\rho$ ισούται με το λόγο του μεγαλύτερου σε μέγεθος στοιχείου των πινάκων $A, A^{(1)}, \ldots, A^{(n-1)}$ προς το μεγαλύτερο σε μέγεθος στοιχείο του πίνακα $A$:

$$\rho = \frac{\max(a, a_1, a_2, \ldots, a_{n-1})}{a}$$

όπου $a = max|a_{ij}|$ και $a_k = max|a_{ij}^{(k)}|$.

**Παράδειγμα 4.2.1** *Εστω*

$$A = \begin{pmatrix} 0.0001 & 1 \\ 1 & 1 \end{pmatrix}$$

*(a) Η απαλοιφή Gauss χωρίς οδήγηση για $t = 3$ δίνει*

$$A^{(1)} = U \equiv \begin{pmatrix} 0.0001 & 1 \\ 0 & -10^4 \end{pmatrix}$$

$$max|a_{ij}^{(1)}| = 10^4$$

$$max|a_{ij}| = 1$$

$$\rho = συντελεστή\ μεγέθυνσης = 10^4$$

*(β) Η απαλοιφή Gauss με μερική οδήγηση δίνει*

$$A^{(1)} = U = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

$$\max|A_{ij}^{(1)}| = 1$$

$$\max|a_{ij}| = 1$$

$$\rho = συντελεστής\ μεγέθυνσης = 1$$

**Ερώτηση:** Πόσο μεγάλο μπορεί να είναι το $\rho$ για έναν αυθαίρετο πίνακα;

**Growth Factor για απαλοιφή Gauss με ολική οδήγηση**
Στην απαλοιφή Gauss με ολική οδήγηση ισχύει ότι

$$\rho \leq \{n * 2^1 * 3^{1/2} * 4^{1/3} \dots n^{1/n-1}\}^{1/2}$$

Η συνάρτηση αυτή μεγαλώνει αργά με το $n$. Επιπλέον στην πράξη αυτό το φράγμα δεν πετυχαίνεται ποτέ. Το 1965 είχε τεθεί από τον Wilkinson η εικασία ότι το growth factor στην ολική οδήγηση φράσσεται από το $n$ για πραγματικούς $n \times n$ πίνακες. Δηλαδή ότι το $\rho$ δε μπορεί να ξεπεράσει τη διάσταση του πίνακα. Το 1991 κατασκευάσθηκε από τον Gould ένας $13 \times 13$ πίνακας για τον οποίο η απαλοιφή Gauss με ολική οδήγηση δίνει growth factor $\rho = 13.0205$. Παρ' όλα αυτά η μέθοδος απαλοιφής του Gauss με ολική οδήγηση είναι ΕΥΣΤΑΘΗΣ αλγόριθμος.

**Growth Factor για απαλοιφή Gauss με μερική οδήγηση**

Στην απαλοιφή Gauss με μερική οδήγηση ισχύει ότι, $\rho \leq 2^{n-1}$, δηλαδή,

$$\rho \text{ μπορεί να είναι τόσο μεγάλο όσο } 2^{n-1}$$

Δυστυχώς μπορεί να κατασκευασθούν πίνακες για τους οποίους αυτό το φράγμα επιτυγχάνεται. Ας θεωρήσουμε τον ακόλουθο πίνακα:

$$A = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 1 \\ -1 & 1 & 0 & \cdots & 0 & 1 \\ \vdots & \ddots & \ddots & & & \vdots \\ \vdots & & & \ddots & \ddots & \\ -1 & \cdots & \cdots & \cdots & -1 & 1 \end{pmatrix}$$

Δηλαδή

$$a_{ij} = \begin{cases} 1 & \text{για } j = i, n \\ -1 & \text{για } j < i \\ 0 & \text{διαφορετικά} \end{cases}$$

Ο Wilkinson το 1965 έδειξε ότι το growth factor $\rho$ για αυτόν τον πίνακα με μερική οδήγηση είναι $2^{n-1}$. Για να το δούμε αυτό ας θεωρήσουμε την ειδική περίπτωση $n = 4$.

$$A = \begin{pmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{pmatrix}$$

$$A^{(1)} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & -1 & 1 & 2 \\ 0 & -1 & -1 & 2 \end{pmatrix}$$

$$A^{(2)} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & -1 & 4 \end{pmatrix}$$

$$A^{(3)} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 8 \end{pmatrix}$$

Ετσι το growth factor είναι

$$\rho = \frac{8}{1} = 2^3 = 24 - 1$$

**Σημείωση:** Αυτός δεν είναι ο μόνος πίνακας για τον οποίο $\rho = 2^{n-1}$.
Οι Higham and Higham το 1989 αναγνώρισαν ένα σύνολο πινάκων για τους οποίους

$$\rho = 2^{n-1}.$$

Ο πίνακας

$$B = \begin{pmatrix} 0.7248 & 0.7510 & 0.5241 & 0.7510 \\ 0.7317 & 0.1889 & 0.0227 & -0.7510 \\ 0.7298 & -0.3756 & 0.1150 & 0.7511 \\ -0.6993 & -0.7444 & 0.6647 & -0.7500 \end{pmatrix}$$

υπόκειται σ' αυτήν την κατηγορία. Πρόσφατα ο Foster (1994) και Wright (1993) έδωσαν παραδείγματα όπου τα στοιχεία των πινάκων $A^{(k)}$ πολύ συχνά συνεχίζουν να ελαττώνονται σε μέγεθος. Ετσι η απαλοιφή Gauss με μερική οδήγηση θεωρητικά είναι ευσταθής υπό συνθήκες, στην πράξη όμως μπορεί να θεωρηθεί ΕΥΣΤΑΘΗΣ αλγόριθμος.

### Growth Factor για απαλοιφή Gauss χωρίς οδήγηση

Στην απαλοιφή Gauss χωρίς οδήγηση το $\rho$ μπορεί να είναι αυθαίρετα μεγάλο εκτός κάποιων ειδικών περιπτώσεων όπως όταν ο πίνακας είναι συμμετρικός θετικά ορισμένος. Γενικά η μέθοδος απαλοιφής του Gauss χωρίς οδήγηση είναι ΑΣΤΑΘΗΣ αλγόριθμος.

Πριν προχωρήσουμε στην ανάλυση σφάλματος άλλης μεθόδου είναι απαραίτητο να αναφέρουμε λίγα στοιχεία για τον υπολογισμό εσωτερικών γινομένων σε διπλή ακρίβεια.

### Υπολογισμός εσωτερικών γινομένων σε διπλή ακρίβεια

Εάν χρησιμοποιήσουμε απλή ακρίβεια για τον υπολογισμό του εσωτερικού γινομένου $x^T y$ εμφανίζονται $(2n-1)$ σφάλματα στρογγύλευσης απλής ακρίβειας (ένα για κάθε πολλαπλασιαμό και πρόσθεση). Γ' αυτό μετατρέπουμε κάθε $x_i$ και $y_i$ σε διπλή ακρίβεια, προσθέτοντας μηδενικά στη mantissa τους, εκτελούμε όλη την πράξη σε διπλή ακρίβεια και δίνουμε το τελικό αποτέλεσμα σε απλή ακρίβεια. Αυτή η διαδικασία είναι γνωστή σαν προσομείωση του εσωτερικού γινομένου σε διπλή ακρίβεια (ή εκτεταμένη ακρίβεια). Ισοδύναμα μπορούμε να πούμε ότι το εσωτερικό γινόμενο υπολογίζεται σε fixed point αριθμητική.

Σχεδόν όλοι οι ψηφιακοί υπολογιστές υπολογίζουν το ακριβές γινόμενο σε $2t$ ψηφία δύο αριθμών $t$ ψηφίων και αυτό το γεγονός το εκμεταλλευόμαστε για να πάρουμε πιο ακριβή αποτελέσματα. Πολύ συχνά χρειάζεται να υπολογίσουμε το εσωτερικό γινόμενο (inner poroduct) $s$, που ορίζεται ως εξής:

$$s = \sum_{i=1}^{n} a_i b_i \tag{4.13}$$

Το εσωτερικό γινόμενο υπολογίζεται σε $2t$ ψηφία είτε το θέλουμε είτε όχι. Μπορούμε να πούμε ότι ένα αποτέλεσμα διπλής ακριβείας παράγεται, παρόλο που ο χρόνος που χρειάζεται για αυτόν τον υπολογισμό είναι ίδιος με το χρόνο που χρειάζεται για να υπολογιστεί ένα αποτέλεσμα απλής ακριβείας.
Χρησιμοποιώντας αυτή τη δυνατότητα του υπολογιστή προκύπτει η ακόλουθη 'υπολογιστική εξίσωση'

$$s \equiv \sum_{i=1}^{n} a_i b_i + \varepsilon, \quad |\varepsilon| \leq u_f \tag{4.14}$$

Το σφάλμα $\varepsilon$ που προκύπτει από την αντικατάσταση ενός $2t$-ψηφίων αριθμού από μία προσέγγιση $t$ ψηφίων προκύπτει εάν προσθέσουμε στον αριθμό το $\frac{1}{2}\beta^{-t}$ και διαγράψουμε τα ψηφία $t+1$ έως $2t$.
Ορίζουμε

$$u_f = \frac{1}{2}\beta^{-t}$$

το μοναδιαίο σφάλμα στρογγύλευσης σε fixed point αριθμητική. Παρατηρούμε ότι $u_f = \frac{u}{\beta}$ συνεπώς $u_f < u$.

Οι υπολογιστές που έχουν την ιδιότητα που αναφέραμε διαθέτουν επεξεργαστή $2t$-ψηφίων και έτσι μπορούν να υπολογίζουν το εσωτερικό γινόμενο με διπλή ακρίβεια. Αυτόν τον υπολογισμό συμβολίζουμε με $fl_2$.

**Παρατήρηση:** Εστω $\underline{a}, \underline{b}, \in \mathbb{R}^n$ δοσμένα διανύσματα. Το σχετικό σφάλμα που προκύπτει κατά τον υπολογισμό του εσωτερικού τους γινομένου χρησιμοποιώντας floating point αριθμητική με διπλή ακρίβεια δίνεται από τον τύπο:

$$\frac{|fl_2(\underline{a}^t \cdot \underline{b}) - \underline{a}^t \cdot \underline{b}|}{|\underline{a}^t \cdot \underline{b}|} \leq | \in | \leq u_f = \frac{u}{\beta}$$

Παρατηρούμε ότι είναι πάντα μικρό ανεξάρτητο του $n$.

Την ύπαρξη ενός τέτοιου επεξεργαστή εκμεταλλευόμαστε στη πράξη της διαίρεσης. Η δυνατότητα της διαίρεσης που διαθέτουν αυτού του είδους οι υπολογιστές είναι η ακόλουθη: Εάν ο διαιρετέος είναι αριθμός $2t$ ψηφίων και ο διαιρέτης είναι αριθμός $t$ ψηφίων τότε ο διαιρέτης μετατρέπεται σε αριθμό $2t$-ψηφίων με την πρόσθεση $t$ μηδενικών ψηφίων.
Εστω ότι θέλουμε να υπολογίσουμε την ποσότητα $d$, που ορίζεται από την εξίσωση

$$d = \frac{\sum_{i=1}^{n} a_i b_i}{c} \tag{4.15}$$

Εάν ο υπολογιστής διαθέτει τις προηγούμενες ιδιότητες που αναφέραμε, τότε η ποσότητα $\sum a_i b_i$ υπολογίζεται ακριβώς και αυτή η ποσότητα διαιρείται με το $a$. Η 'υπολογιστική εξίσωση' που αντιστοιχεί στην (4.15) είναι:

$$d \equiv (\frac{\sum_{i=1}^{n} a_i b_i}{c}) + \varepsilon, \quad |\varepsilon| \leq u_f \tag{4.16}$$

και συνεπώς

$$|cd - \sum_{i=1}^{n} a_i b_i| \equiv |ce| \leq |c| u_f$$

## 4.3 Μέθοδος της τριγωνικής διαχώρισης ($LU$)

Η τριγωνική διαχώριση με μερική οδήγηση αναλύει ένα πίνακα $A \in \mathbb{R}^{n \times n}$ σε ένα γινόμενο δύο πινάκων $L$ και $U$, όπου ο $L$ είναι κάτω τριγωνικός και ο $U$ είναι άνω τριγωνικός.Χρησιμοποιούμε τους παρακάτω τύπους:

$$s_p = a_{pr} - \sum_{j=1}^{r-1} l_{pj} u_{jr}, \quad p = r, r+1, \ldots, n, r = 1, 2, \ldots, n$$

$$u_{rr} = s_r$$

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

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

Λόγω της μερικής οδήγησης πάντοτε

$$|l_{pr}| \leq 1$$

Για να έχουμε κάποιο κέρδος απ' αυτό τον απεύθείας υπολογισμό των $L$ και $U$ θα πρέπει να υπολογιστούν τα εσωτερικά γινόμενα με το καλύτερο δυνατό τρόπο. Εάν χρησιμοποιήσουμε floating-point αριθμητική τότε η τριγωνική διαχώρηση και το σφάλμα στρογγύλευσης που δίνει είναι ακριβώς τα ίδια με την μέθοδο απαλοιφής του Gauss με μερική οδήγηση.
Εάν ο πίνακας $A$ υποστεί κάποιο scaling από την αρχή ώστε όλα τα στοιχεία του $|U|$ να παραμένουν φραγμένα από τη μονάδα, τότε χρησιμοποιώντας fixed-point αριθμητική έχουμε:

$$u_{i1} = a_{1i}, \;\; i = 1, 2, \ldots, n$$

$$u_{rr} \equiv s_r + \varepsilon_{rr}, \;\; r = 2, 3, \ldots, n, \;\; |\varepsilon_{rr}| \leq u_f$$

$$u_{rp} \equiv (a_{rp} - \sum_{j=1}^{r-1} l_{rj} u_{jr}) + \varepsilon_{rp}, \;\; p = r+1, \ldots, n, \;\; |\varepsilon_{rp}| \leq u_f$$

$$l_{pr} \equiv \frac{s_p}{u_{rr}} + \varepsilon_{pr}^{'} \Rightarrow l_{pr} u_{rr} \equiv s_p + \varepsilon_{pr}^{'} u_{rr} \Rightarrow$$

$$l_{pr} \equiv s_p + \varepsilon_{pr}, \;\; p = r+1, \ldots, n, \;\; |\varepsilon_{pr}| \leq u_f |u_{rr}|$$

Έτσι εάν $E = (\varepsilon_{ij})$ είναι ο πίνακας σφάλματος, τότε φράσσεται από τα ακόλουθα φράγματα: $|E| = (|\varepsilon_{ij}|) \leq \begin{cases} 0 & , \quad i = 1, \; j = 1, \ldots, n \\ u_f & , \quad i = 2, \ldots, n, \;\; j = i, \ldots, n \\ u_f |u_{rr}| & , \quad i = 2, \ldots, n, \;\; j = 1, \ldots, i-1 \end{cases}$
Ετσι τελικά προκύπτει

$$|E| \leq u_f \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ |u_{11}| & 1 & 1 & 1 & \cdots & \cdots & 1 \\ |u_{11}| & |u_{22}| & 1 & 1 & \cdots & \cdots & 1 \\ |u_{11}| & |u_{22}| & |u_{33}| & 1 & \cdots & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ |u_{11}| & |u_{22}| & |u_{33}| & \cdots & \cdots & |u_{n-1n-1}| & 1 \end{pmatrix}$$

Από την υποθεσή μας, εάν αντικατασταθεί κάθε $|u_{rr}|$ με τη μονάδα και εάν πάρουμε $\|\cdot\|_\infty$ προκύπτει

$$||E||_\infty \leq n u_f$$

Απ΄όλα τα προηγούμενα προκύπτει το ακόλουθο θεώρημα

**Θεώρημα 4.3.1** *Οι πίνακες $L$ και $U$ που υπολογίζονται από τη μέθοδο της τριγωνικής διαχώρισης με μερική οδήγηση, χρησιμοποιώντας fixed-point αριθμητική με μονάδα σφάλματος στρογγύλευσης $u_f$, ικανοποιούν τη σχέση*

$$LU = A + E$$

*όπου*

$$||E||_\infty \leq n u_f$$

*Με άλλα λόγια οι πίνακες $L$ και $U$ αποτελούν τη διάσπαση ενός* **ελαφρά διαταραγμένου αρχικού πίνακα**.

**Παράδειγμα 4.3.1** *Για $n = 5$ ένα φράγμα για τον πίνακα $E$ ισούται με:*

$$u_f \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ |u_{11}| & 1 & 1 & 1 & 1 \\ |u_{11}| & |u_{22}| & 1 & 1 & 1 \\ |u_{11}| & |u_{22}| & |u_{33}| & 1 & 1 \\ |u_{11}| & |u_{22}| & |u_{33}| & |u_{44}| & 1 \end{pmatrix}$$

**Παρατήρηση**: Εάν συγκρίνουμε το σφάλμα μεταξύ των μεθόδων της απαλοιφής του Gauss και της $LU$ παρατηρούμε ότι:

| **Μέθοδος απαλοιφής Gauss** | **Μέθοδος LU** |
|---|---|
| $LU = A + E$ | $LU = A + E$ |
| $\|\|E\|\|_\infty \leq n^2 pu \|\|A\|\|_\infty$ | $\|\|E\|\|_\infty \leq n u_f$ |
| σε floating-point αριθμητική | σε fixed-point αριθμητική |

Επειδή $u_f < u$ , εάν χρησιμοποιήσουμε την μέθοδο $LU$ σε fixed-point αριθμητική παίρνουμε καλύτερο φράγμα για τον πίνακα σφάλματος που προκύπτει.

### 4.3.1 Υπολογισμός ορίζουσας

Από τη σχέση $L \cdot U = A + E$, όπου $A$ ο αρχικός πίνακας (μπορεί με μετατιθεμένες τις αρχικές του γραμμές) παρατηρούμε ότι:

$$D \equiv \det(A + E) = \prod_{i=1}^{n} u_{ii}$$

Ετσι,

$$D = (1 + \varepsilon) \prod_{i=1}^{n} u_{ii}, \quad |\varepsilon| \leq (n - 1)u_1$$

Επομένως η $D$ είναι η **ακριβής** ορίζουσα του πίνακα $A + E$.

**Σχέση μεταξύ ορίζουσας και ιδιοτιμών**

Εάν $\lambda'_i$ είναι οι ιδιοτιμές του $A + E$ και $\lambda_i$ είναι οι ιδιοτιμές του $A$ τότε

$$\det(A) = \prod_{i=1}^{n} \lambda_i, \quad D = (1 + \varepsilon) \prod_{i=1}^{n} \lambda'_i$$

## 4.4 Περιορισμός Μεγέθους των Στοιχείων ενός Πίνακα (Scaling)

**Γενικό πρόβλημα:** Πολ/σμος ενός πίνακα $A \in \mathbb{R}^{n\times n}$ με κατάλληλο πίνακα $D \in \mathbb{R}^{n\times n}$ από αριστερά, συνήθως $D = diag\{d_i\}$ έτσι ώστε τα στοιχεία του $D \cdot A$ να φράσσονται.

$$\text{Εάν} \quad A = \begin{pmatrix} a_{11} & \cdots & \cdots & a_{1n} \\ a_{21} & \cdots & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & \cdots & \cdots & a_{nn} \end{pmatrix} \in \mathbb{R}^{n\times n} (\ \ \text{ή} \mathbb{R}^{m\times n})$$

$$\text{τότε} \quad D = \begin{pmatrix} d_1 & & & \\ & \ddots & & O \\ & & \ddots & \\ & O & & \ddots \\ & & & & d_n \end{pmatrix} \in \mathbb{R}^{n\times n}$$

Τα στοιχεία $d_i,\ i = 1, 2, \ldots,$ μπορούν να ορισθούν με διάφορους τρόπους

1. $d_i = \frac{1}{max|a_{ij}|} \Rightarrow \frac{a_{ij}}{d_i} \leq 1$
2. $d_i = \frac{1}{||\underline{a_i}^t||_2} = \frac{1}{\sqrt{a_{11}^2+\ldots+a_{1n}^2}}$

τότε κάθε γραμμή του $DA$ έχει $\|\cdot\|_2 = 1$

Ομως αυτά τα scaling επιφέρουν rounding errors και τίθεται το ερώτημα εάν ο υπολογισμός $fl(DA)$ είναι ευσταθής.

**Παράδειγμα 4.4.1** *Αποδείξτε ότι* $fl(DA) = D(A + E)$

**Απόδειξη:**

$$fl(\frac{a_{ij}}{s_i}) = \frac{a_{ij}}{s_i}(1 + \varepsilon_i) = \frac{1}{s_i}(a_{ij} + a_{ij}\varepsilon_i) = D(A + E)$$

όπου $E = (\varepsilon'_{ij})$, $\varepsilon'_{ij} = (a_{ij}\varepsilon_i)$, $|\varepsilon_i| \leq u$ ο πίνακας σφάλματος για τον οποίο ισχύει $\|E\| \leq u\|A\|$. □

Για να αποφύγουμε τα προβλήματα που επιφέρουν τα σφάλματα στρογγύλευσης μπορεί να εφαρμοσθεί το λεγόμενο B-scaling

Το B-scaling χρησιμοποιεί τη βάση αριθμητικής $\beta$ του Η/Υ και δεν επιφέρει σφάλματα στρογγύλευσης. Ο παρακάτω αλγόριθμος εισαγάγει B-scaling στα στοιχεία ενός

πίνακα $A = [\underline{r_1}, \ldots, \underline{r_m}]^t \in \mathbb{R}^{m\times n}$. Μετά την εφαρμογή του τα στοιχεία του $A$ θα ικανοποιούν τη σχέση:

$$\beta^{-1} < max_{1\le j\le n}|a_{ij}| \le 1,\ i = 1, 2, \ldots, m$$

### 4.4.1 Αλγόριθμος B-SCALE

**for** $i = 1, 2, \ldots, m$
$s_i := max_{1\le j\le n}|a_{ij}|$
Καθορίστε το μικρότερο ακέραιο $r_i \in \mathbb{Z}$:
$\beta^{r_i} > s_i$
$d_i := \frac{1}{\beta^{r_i}}$
$\underline{r_i} = \underline{r_i} * d_i$

**Παρατήρηση:** Εάν $x$ είναι ένας floating point αριθμός π.χ. $x = \sigma\bar{x}\beta^e$ τότε εάν πολ/στεί με κατάλληλη δύναμη του εκθέτη π.χ. $\beta^k$ τότε

$$x\beta^k = \sigma\bar{x}\beta^e\beta^k = \sigma\bar{x}\beta^{e+k}$$

Απλώς τροποποιείται ο εκθέτης κι'έτσι δεν υπεισέρχεται σφάλμα στρογγύλευσης. Επομένως το B-scaling δεν επιφέρει σφάλμα στρογγύλευσης.

**Παράδειγμα 4.4.2** *Να επιλυθεί το σύστημα* $A\underline{x} = \underline{b}$ *όπου:*

$$30x_1 + 591400x_2 = 591700$$
$$5.291x_1 - 6.130x_2 = 46.78$$

$$A = \begin{pmatrix} 30 & 591400 \\ 5.291 & -6.13 \end{pmatrix},\quad \underline{b} = [591700\ \ 46.78]^t$$

*Θεωρητική λύση:* $x_1 = 10,\ x_2 = 1.$

Εάν εφαρμοσθεί η τεχνική της μερικής οδήγησης θα έχουμε: (εφαρμόζουμε αριθμητική 4 δεκαδικών ψηφίων)

$$m_{21} = \frac{5.291}{30.00} = 0.1764$$

που οδηγεί στο σύστημα

$$30.00x_1 + 591400x_2 = 591700$$
$$-104300x_2 = -104400$$
$$\Rightarrow x_1 = -10.00,\quad x_2 = 1.001$$

Η έλλειψη ακριβείας οφείλεται στο **μεγάλο** μέγεθος των στοιχείων του πίνακα $A$. Γιάυτό θα πρέπει να εφαρμόζουμε την λεγόμενη **scaled μερική οδήγηση**. Διαιρούμε κάθε γραμμή με το μέγιστο στοιχείο της και στη συνέχεια εφαρμόζουμε την οδήγηση. Στο παραπάνω παράδειγμα:

$$\frac{30}{591400} = 0.00005073x_1 + x_2 = \frac{591700}{591400}$$

$$\frac{5.291}{6.130}x_1 - x_2 = \frac{46.78}{6.130}$$

$$\Rightarrow$$

$$0.8631x_1 - x_2 = 7.6313$$

$$0.00005073x_1 + x_2 = 1.0005$$

$$\Rightarrow \text{ (Με αριθμητική 4 ψηφίων) } \quad x_2 \simeq 1, \quad x_1 \simeq 10.0004$$

**Αλγόριθμος SCALED μερική οδήγηση**

Δίνεται πίνακας $A = [\underline{r_1}, \ldots, \underline{r_n}]^t \in \mathbb{R}^{n\times n}$ και αντικαθίσταται από τον κατάλληλο πίνακα με scaling.

.
.
**φορ** $i = 1 : n$
$\quad s_i = ||\underline{r_i}||_\infty$
ενδ
$D := diag\{s_i\}$
$A := DA$
.
.
.

**Παρατήρηση:** Αυτή η τεχνική μπορεί να εφαρμόζεται και στα ενδιάμεσα βήματα της μεθόδου του Gauss.

## 4.5 Ειδικές Μορφές Πινάκων

### 4.5.1 $LU$ παραγοντοποίηση σε ορθογώνιο πίνακα

Εάν $A \in \mathbb{R}^{m\times n}$ πάλι μπορεί να γίνει τριγωνοποίηση.
Εάν $\underline{m > n}$

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

$\underline{m < n}$

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

Για $A \in \mathbb{R}^{m \times n}$ η LU παραγοντοποίηση υπάρχει εάν $A(1:k, 1:k)$ μη ιδιάζων για $k = 1 : min(m, n)$.

### Αποφυγή της μερικής οδήγησης

Σε ορισμένες κατηγορίες πινάκων δεν χρειάζεται να εφαρμοσθεί η τεχνική της μερικής οδήγησης. Αυτό συμβαίνει όταν οι πίνακες είναι αυστηρά διαγώνια υπερτερόντες.

**Ορισμός:** Εάν $A \in \mathbb{R}^{n \times n}$ είναι αυστηρά διαγώνια υπερτερών εάν

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

**Θεώρημα 4.5.1** *Εάν $A^T$ είναι διαγώνια υπερτερών τότε ο $A$ έχει $LU$ παραγοντοποίηση και $|l_{ij}| \leq 1$.*

**Απόδειξη:** Διαχωρίζουμε τον $A$ ως εξής:

$$A = \begin{pmatrix} a & w^T \\ v & c \end{pmatrix}$$

όπου $a$ είναι $1 \times 1$ και λόγω της αρχικής ιδιότητας $a$ το μέγιστο στοιχείο της $[a \;\; v]^t$. Μετά από το πρώτο βήμα προκύπτει η επόμενη παραγοντοποίηση

$$\begin{pmatrix} a & w^T \\ v & c \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{v}{a} & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & c - \frac{vw^T}{a} \end{pmatrix} \begin{pmatrix} a & w^T \\ 0 & I \end{pmatrix}$$

(Το θεώρημα θα προκύψει από επαγωγή στο $n$)

Εάν, $B^T, B = C - \frac{vw^T}{a}$ είναι διαγώνια υπερτερών τελικά $B = L_1 U_1$ οπότε

$$A = \begin{pmatrix} 1 & 0 \\ \frac{v}{a} & L_1 \end{pmatrix} \begin{pmatrix} a & w^T \\ 0 & v_1 \end{pmatrix} \equiv LU$$

Η απόδειξη ότι $B^T$ είναι διαγώνια υπερτερών είναι άμεση.

Πράγματι

$$\sum_{i=1}^{n-1}|b_{ij}| = \sum_{i=1}^{n-1}|c_{ij} - v_i w_j/a| \leq \sum_{i=1}^{n-1}|c_{ij}| + \frac{|w_j|}{|a|}\sum_{i=1}^{n-1}|v_i|$$

$$\leq (|c_{jj}| - |w_j|) + \frac{|w_j|}{|a|}(|a| - |v_j|)$$

$$\leq |c_{jj} - \frac{w_j v_j}{a}| = |b_{jj}|$$

□

**Θεώρημα 4.5.2** *Εάν όλες οι κύριες υποορίζουσες του $A \in \mathbb{R}^{n\times n}$ είναι μη ιδιάζουσες, τότε υπάρχουν μοναδικοί μοναδιαίοι κάτω και άνω τριγωνικοί πίνακες $L$ και $M$ και ένας μοναδικός διαγώνιος πίνακας $D = diag(d_1, \ldots, d_n)$ έτσι ώστε $A = LDM^T$.*

**Απόδειξη:** Ως γνωστόν ο $A$ έχει $LU$ παραγοντοποίηση, $A = LU$. Θέτουμε $D = diag(d_1, \ldots, d_n)$ με $d_i = u_{ii},\ i = 1, 2, \ldots, n$. Τότε $D$ μη ιδιάζων και ο πίνακας $M^T = D^{-1}U$ είναι μοναδιαίος άνω τριγωνικός.

Ετσι $A = LU = LD(D^{-1}U) = LDM^T$

Η μοναδικότητα προκύπτει από τη μοναδικότητα της $LU$. □

**'Ασκηση 4.5.1** *Να δοθεί αλγόριθμος που να υπολογίζει απευθείας τη διάσπαση ενός πίνακα $A$ σε $LDM^T$.*

**Παράδειγμα 4.5.1**

$$A = \begin{pmatrix} 10 & 10 & 20 \\ 20 & 25 & 40 \\ 30 & 50 & 61 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix} \begin{pmatrix} 10 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

# Επίλυση Συστήματος

$$A\underline{x} = \underline{b}$$
$$A = LDM^T$$
$$\Rightarrow (LDM^T)\underline{x} = \underline{b} \Rightarrow$$

Προκύπτει η ακόλουθη ισοδύναμία:

$$L\underline{y} = \underline{b}$$
$$D\underline{z} = \underline{y}$$
$$M^T\underline{x} = \underline{z}$$

**Θεώρημα 4.5.3** *Εάν $A = LDM^T$, $A$ συμμετρικός τότε $L = M$.*

**Απόδειξη:** Ο πίνακας $M^{-1}AM^{-T} = M^{-1}LD$ είναι συμμετρικός και κάτω τριγωνικός επομένως διαγώνιος. Αφού $D$ μη ιδιάζων τότε $M^{-1}L$ διαγώνιος. Όμως γενικά $M^{-1}L$ είναι μοναδιαίος κάτω τριγωνικός κι'έτσι $M^{-1}L = I$. □

**Παράδειγμα 4.5.2**

$$A = \begin{pmatrix} 10 & 20 & 30 \\ 20 & 45 & 80 \\ 30 & 80 & 17 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix} \begin{pmatrix} 10 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{pmatrix}$$

## 4.5.2 Θετικά Ορισμένα Συστήματα-Ανάλυση Cholesky

Ο πίνακας $A \in \mathbb{R}^{n\times n}$ λέγεται **θετικά ορισμένος** εάν $\underline{x}^T A\underline{x} > 0$ $\forall \underline{x} \in \mathbb{R}^n, \underline{x} \neq \underline{0}$.

### Ο Αλγόριθμος Cholesky

Έστω $A$ συμμετρικός θετικά ορισμένος πίνακας. Τότε ο $A$ γράφεται ως εξής:

$$A = HH^T$$

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} = \begin{pmatrix} h_{11} & 0 & \dots & 0 \\ h_{21} & h_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \dots & h_{nn} \end{pmatrix} \begin{pmatrix} h_{11} & h_{12} & \dots & h_{1n} \\ 0 & h_{22} & \dots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & h_{nn} \end{pmatrix}$$

όπου

$$h_{11} = \sqrt{a_{11}},\ \ h_{i1} = \frac{a_{i1}}{h_{11}},\ i = 1, \ldots, n$$

$$\sum_{k=1}^{i} h_{ik}^2 = a_{ii},\ \ a_{ij} = \sum_{k=1}^{j} h_{ik} h_{jk},\ \ j < i.$$

**Απόδειξη: Ύπαρξη:** Θα χρησιμοποιήσουμε επαγωγή ως προς $n$. Έστω λοιπόν ότι υπάρχει η ανάλυση Cholesky για θετικά ορισμένους πίνακες $n-1 \times n-1$. Τότε ο πίνακας $A$ μπορεί να γραφεί στη μορφή:

$$A = \begin{pmatrix} A_{n-1} & c \\ c^T & a_{nn} \end{pmatrix},\ \ A_{n-1} = G_{n-1}^T \cdot G_{n-1},\ \ G_{n-1} \text{ μοναδικός},\ c \in \mathbb{R}^{n-1}$$

Ορίζουμε τον πίνακα $G$ ως:

$$G = \begin{pmatrix} G_{n-1} & b \\ 0^T & \sqrt{k} \end{pmatrix},\ \ \text{όπου } b = (G_{n-1}^T)^{-1} \text{ και } k = a_{nn} - b^T$$

τότε $k = a_{nn} - c^T \cdot A_{n-1}^{-1} \cdot c > 0$ και ο πίνακας $A$ μπορεί να γραφεί στη μορφή $A = G^T \cdot G$.

**Μοναδικότητα του πίνακα $G$:** Αφού $G = \begin{pmatrix} G_{n-1} & b \\ 0^T & \sqrt{k} \end{pmatrix}$ και $G^T \cdot G = A = \begin{pmatrix} A_{n-1} & c \\ c^T & a_{nn} \end{pmatrix}$ έπεται ότι $G_{n-1}^T \cdot G_{n-1} = A_{n-1}$ όπου $G_{n-1}$ είναι μοναδικός. Επίσης έχουμε

$$G_{n-1}^T \cdot b = c,\ b^T \cdot b + k = a_{nn}$$

και άρα $b$ και $k$ είναι μοναδικά.

Αφού τώρα το θεώρημα ισχύει για $n = 1, 2$ τότε ισχύει για κάθε $n$. □

Από τα παραπάνω προκύπτει ο αλγόριθμος Cholesky κατά γραμμές.

Δοθέντος ενός $n \times n$ συμμετρικό και θετικά ορισμένο πίνακα $A$, ο παρακάτω αλγόριθμος υπολογίζει τον παράγωντα Cholesky $H$ Ο πίνακας $H$ υπολογίζεται γραμμή-γραμμή και αποθηκεύεται στο κάτω τριγωνικό τμήμα του πίνακα $A$.

**For** $k = 1, 2, \ldots, n$ do
  **For** $i = 1, 2, \ldots, k-1$ do
    $a_{ki} = h_{ki} = \frac{1}{h_{ii}} \left( a_{ki} - \sum_{j=1}^{i-1} h_{ij} h_{kj} \right)$
  $a_{kk} = h_{kk} = \sqrt{a_{kk} - \sum_{j=1}^{k-1} h_{kj}^2}$

### Πολυπλοκότητα του Αλγόριθμου Cholesky

Ο συνολικός αριθμός των flops που απαιτούνται για τον αλγόριθμο Cholesky ισούτε με

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

$$=\frac{1}{2}\frac{n(n+1)(2n+1)}{6}\approx O\left(\frac{n^3}{6}\right)\text{ flops}$$

Επιπλέον χρειάζεται ο υπολογισμός $O(n)$ τετραγωνικών ριζών.

## 4.5.3 Εφαρμογή στην Επίλυση Γραμμικού Συστήματος

Θεωρούμε ένα θετικά ορισμένο σύστημα $A\underline{x}=\underline{b}$. Αν είναι γνωστή η ανάλυση Cholesky του πίνακα $A=HH^T$ τότε μπορεί να λυθεί το σύστημα λύνοντας λύνοντας το κάτω τριγωνικό σύστημα $H\underline{y}=\underline{b}$, ακολουθούμενο από το άνω τριγωνικό σύστημα $H^T\underline{x}=\underline{y}$.

### Ο Αλγόριθμος Cholesky για Θετικά Ορισμένο Σύστημα

Βήμα 1. Βρίσκουμε την ανάλυση Cholesky του $A=HH^T$ χρησιμοποιώντας τον παραπάνω αλγόριθμο.

Βήμα 2. Λύνουμε το κάτω τριγωνικό σύστημα για το $\underline{y}$: $H\underline{y}=\underline{b}$

Βήμα 3. Λύνουμε το άνω τριγωνικό σύστημα για το $\underline{x}$: $H^T\underline{x}=\underline{y}$

**Παράδειγμα 4.5.3**

Έστω

$$A=\begin{pmatrix}1 & 1 & 1\\ 1 & 5 & 5\\ 1 & 5 & 14\end{pmatrix},\ b=\begin{pmatrix}3\\ 11\\ 20\end{pmatrix}$$

Βήμα 1. Υπολογισμός της ανάλυσης Cholesky

1η γραμμή($k=1$):

$$h_{11}=1$$

2η γραμμή($k=2$):

$$h_{21}=\frac{a_{21}}{h_{11}}=1$$

$$h_{21}=\sqrt{a_{22}-h_{21}^2}=\pm 2$$

(λόγω του ότι τα διαγώνια στοιχεία του $H$ πρέπει να είναι θετικά, επιλέγουμε το πρόσημο $+$.

3η γραμμή($k=3$):

$$h_{31} = \frac{a_{31}}{h_{11}} = 1$$

$$h_{32} = \frac{1}{h_{22}}(a_{32} - h_{21}h_{31}) = 2$$

$$h_{33} = \sqrt{a_{33} - (h_{31}^2 - h_{32}^2)} = \pm\sqrt{9}$$

δηλαδή

$$h_{33} = +3$$

'ρα

$$H = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 2 & 3 \end{pmatrix}$$

Βήμα 2. Λύνουμε το $H\underline{y} = \underline{b}$:

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 3 \\ 11 \\ 20 \end{pmatrix}$$

$$y_1 = 3, \;\; y_2 = 4, \;\; y_3 = 3$$

Βήμα 3. Λύνουμε το $H^T\underline{x} = \underline{y}$:

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 3 \\ 4 \\ 3 \end{pmatrix}$$

$$x_1 = 3, \;\; x_2 = 1, \;\; x_3 = 1$$

**Υπολογισμός των flops:** Ο αλγόριθμος Cholesky απαιτεί $n^3/6$ flops για τον υπολολογισμό του $H$, τα μισά από ότι απαιτούνται για την $LU$ παραγοντοποίηση. Παρατηρούμε επίσης ότι χρειάζονται επιπλέον $n$ υπολογισμοί τετραγωνικών ριζών. Για την λύση του κάθε τριγωνικού συστήματος χρειάζονται $n^2/2$ flops. 'ρα, για τη λύση ενός θετικά ορισμένου συσήματος με χρήση του αλγόριθμου Cholesky χρείαζονται $n^3/6 + n^2$ flops και $n$ τετραγωνικές ρίζες.

## 4.6 Προσέγγιση της Αποτελεσματικότητας Αλγορίθμων

Όταν προσπαθούμε να μετρήσουμε τα flops που εμπεριέχονται σ' έναν αλγόριθμο μέχρι στιγμής είδαμε ότι εμφανίζονται δικτυωτά αθροίσματα (διπλά ή τριπλά) για τα οποία ενδιαφερόμαστε να υπολλογίσουμε τον κυρίαρχο όρο (τη μεγαλύτερη εμφανιζόμενη δύναμη). Αυτός ο όρος μπορεί να υπολογισθεί εάν αντικαταστήσουμε το άθροισμα με ολοκλήρωμα προσαρμόζοντας κατάλληλα τα όρια ολοκλήρωσης.
Για παράδειγμα

$$\sum_{i=1}^{n}\sum_{j=1}^{i}1=\sum_{i=1}^{n}i=\frac{n(n+1)}{2}\cong\frac{n^2}{2}$$

Ο κυρίαρχος όρος του προηγούμενου αθροίσματος μπορεί να προσεγγισθεί ώς εξής:

$$\int_0^n\int_0^i 1djdi=\int_0^n idi=\frac{n^2}{2}$$

Μετρώντας τις αριθμητικές πράξεις που περιέχονται σ' έναν αλγόριθμο μπορεί να θεωρηθεί και δυνατότητα πρόβλεψης του χρόνου εκτέλεσης του αλγορίθμου. Συγκεκριμένα εάν πολλαπλασιάσουμε τον αριθμό των πράξεων με το χρόνο που χρειάζεται για την εκτέλεση η συγκεκριμένη πράξη έχουμε τη συνολική εκτίμηση του απαιτούμενου χρόνου. Π.χ. εάν $a$ είναι ο απαιτούμενος χρόνος για την εκτέλεση μιας πρόσθεσης ή ενός πολλαπλασιασμού τότε η επίλυση ενός άνω τριγωνικού συστήματος θα απαιτούσε $\frac{n^2}{2}a$ χρόνο.
Στην πράξη όμως αυτό δεν υπολογίζεται. Όταν ο αλγόριθμος αναφέρεται στο $a_{ij}$ στοιχείο ενός πίνακα θα πρέπει να το προσπελάσει από τη μνήμη και να ελέγξει τις τιμές των δεικτών $i$, $j$. Κατά συνέπεια απαιτείται επιπρόσθετος χρόνος από την απλή εκτέλεση μιας πράξης. Έτσι η εκτίμηση του συνολικού αριθμού των flops για την εκτέλεση ενός αλγορίθμου μας επιτρέπει τα εξής:

- Πρόβλεψη της αύξησης του χρόνου εκτέλεσης του αλγορίθμου συναρτήσει της διάστασής του $n$, γραμμικά, εκθετικά κ.τ.π.
- Καθιερώνει ένα μέτρο σύγκρισης μεταξύ διαφορετικών αλγορίθμων. Πάντα θα προτιμάτε ένας αλγόριθμος που έχει κυρίαρχο όρο μικρότερης δύναμης.

**Συνοπτικός Πίνακας**

**Παραγοντοποιήσεις**

| ΠΙΝΑΚΑΣ | ΙΔΙΟΤΗΤΑ | ΔΙΑΣΠΑΣΗ |
|---|---|---|
| $A \in \mathbb{R}^{n\times n}$ | Ολες οι κύριες υποορίζουσες του $\neq 0$ | **μοναδική:** $LU$ διάσπαση $A = LU$<br>$L$:κάτω τριγωνικός με μοναδιαία διαγώνιο<br>$U$:άνω τριγωνικός<br>( διάσπαση Doolitle )<br>---<br>$L$:κάτω τριγωνικός<br>$U$:άνω τριγωνικός με μοναδιαία διαγώνιο<br>( διάσπαση Crout )<br>---<br>**μοναδική:** $LDM^T$ διάσπαση<br>$A = LDM^T$<br>$L$, $M$: κάτω τριγωνικοί με μοναδιαία διαγώνιο<br>$D$: διαγώνιος πίνακας |
| $A \in \mathbb{R}^{n\times n}$ | Συμμετρικός και όλες οι κύριες υποορίζουσές του $\neq 0$ | **μοναδική:** $LDL^T$ διάσπαση<br>$A = LDL^T$<br>$L$: κάτω τριγωνικός με μοναδιαία διαγώνιο<br>$D$: διαγώνιος |
| $A \in \mathbb{R}^{n\times n}$ | Συμμετρικός και θετικά ορισμένος | **μοναδική:** $GG^T$<br>$A = GG^T$<br>$G$ κάτω τριγωνικός με θετικά διαγώνια στοιχεία<br>( ανάλυση CHOLESKY ) |

# Κεφάλαιο 5

# Μετασχηματισμοί Householder

## 5.1 Παραγοντοποίση $QR$

**Ορισμός:** Ενας πίνακας $n \times n$ της μορφής:

$$H = I - \frac{2\underline{u}\ \underline{u}^T}{\underline{u}^T\ \underline{u}},\ \underline{u} \neq \underline{0},\ \underline{u} \in \mathbb{R}^n$$

λέγεται **πίνακας Householder ή μετασχηματισμός Householder**.

Εάν θέσουμε $\underline{u}^T \cdot \underline{u} = \|\underline{u}\|_2^2\ kai\ \underline{v} = \frac{\underline{u}}{\|\underline{u}\|_2}$, τότε ο πίνακας Η μπορεί να πάρει τη μορφή

$$H' = I - 2\frac{\underline{u}\ \underline{u}^T}{\|\underline{u}\|_2\|\underline{u}\|_2} = I - 2\underline{v}\ \underline{v}^T,\ \underline{v} \in \mathbb{R}^n, \|\underline{v}\|_2^2 = 1$$

### Βασικές Ιδιότητες

1) Ο πίνακας Householder είναι συμμετρικός. Πράγματι

$$(H')^T = I - (2\underline{v}\ \underline{v}^T)^T = I - 2\underline{v}\ \underline{v}^T = H'$$

2) Ο πίνακας Householder είναι ορθογώνιος

$$(H')^T(H') = (H')^2 = (I - 2\underline{v}\,\underline{v}^T)(I - 2\underline{v}\ \underline{v}^T) = I - 4\underline{v}\ \underline{v}^T + 4\underline{v}\ \underline{v}^T\underline{v}\ \underline{v}^T =$$

$$I - 4\underline{v}\ \underline{v}^T + 4\underline{v}\ \underline{v}^T = I,\ \text{αφού}\ \underline{v}^T\underline{v} = \|\underline{v}\|_2^2 = 1$$

3) Οι πίνακες Householder λέγονται και στοιχειώδεις πίνακες ανάκλασης

Η σπουδαιότητα των μετασχηματισμών Householder έγκειται στο γεγονός ότι μπορούν να χρησιμοποιηθούν για τη δημιουργία μηδενικών εισόδων σ'ένα διάνυσμα. Συγκεκριμένα ισχύει το παρακάτω Λήμμα:

**Λήμμα 5.1.1** *Εστω ένα δοσμένο διάνυσμα $\underline{x}$ μη μηδενικό, $\underline{x} \neq \underline{e_1}$. Πάντοτε υπάρχει ένας πίνακας Householder $H$ έτσι ώστε $H \cdot \underline{x}$ είναι πολλαπλάσιο του $\underline{e_1}$.*

**Απόδειξη:** Από τον ορισμό του πίνακα Householder έχουμε ότι:

$$H\underline{x} = [I - \frac{2\underline{u}\,\underline{u}^T}{\underline{u}^T\underline{u}}]\underline{x} = \underline{x} - \frac{2\underline{u}\,\underline{u}^T}{\underline{u}^T\underline{u}}\underline{x}$$

Παρατηρούμε ότι για να ανήκει το $H \cdot \underline{x}$ στο χώρο που παράγεται από το $\underline{e_1}$ δηλαδή να ισχύει $H\underline{x} \in \ span\{\underline{e_1}\}$ θα πρέπει $\underline{u} \in \ span\{\underline{x}, \underline{e_1}\}$.
Εάν θέσουμε $\underline{u} = \underline{x} + a\underline{e_1}$ παίρνουμε:

$$\underline{u}^T\underline{x} = \underline{x}^T\underline{x} + ax_1$$

και

$$\underline{u}^T\underline{u} = \underline{x}^T\underline{x} + 2ax_1 + a^2$$

και κατά συνέπεια

$$H\underline{x} = [1 - 2\frac{\underline{x}^T\underline{x} + ax_1}{\underline{x}^T\underline{x} + 2ax_1 + a^2}]\underline{x} - 2a\frac{\underline{u}^T\underline{x}}{\underline{u}^T\underline{u}}\underline{e_1}$$

Για να είναι ο συντελεστής του $\underline{x}$ ίσος με μηδέν αρκεί να θέσουμε $a = \pm\|\underline{x}\|_2$
Έτσι, εάν $\underline{u} = \underline{x} \pm \|\underline{x}\|_2\underline{e_1}$ προκύπτει ότι $H\underline{x} = \pm\|x\|_2\underline{e_1}$

□

**Παρατηρήσεις:**

1. Συνήθως για το πρόσημο του $a$ επιλέγουμε $a = sign(x_1)\|\underline{x}\|_2$ οπότε το $\underline{u} = \underline{x} + sign(x_1)\|\underline{x}\|_2\underline{e_1}$ Αυτός ο απλός ορισμός του $\underline{u}$ καθιστά τον πίνακα Householder ένα πολύ χρήσιμο εργαλείο. Παρατηρούμε επίσης ότι αφού $u = x + sign(x_1)\|x\|_2\underline{e_1} = (x_1 + sign(x_1)\|x\|_2, x_2, \ldots, x_n)^T$, το διάνυσμα $\underline{u}$ μπορεί να αποθηκεύεται πάνω στο $\underline{x}$. Επίσης ισχύει ότι $H\underline{x} = (-sign(x_1)s, 0, \ldots, 0)^t$.
2. Για να εξαλειφθεί οποιαδήποτε πιθανότητα υπερχείλισης ή υποχείλισης κατά τον υπολογισμό της $\|\underline{x}\|_2$ χρησιμοποιούμε κατάλληλο scaling στο διάνυσμα $\underline{x}$. Συγκεκριμένα για τον ορισμό του $\underline{u}$ χρησιμοποιούμε το διάνυσμα $\frac{\underline{x}}{\max\{x_i\}}$ αντί για το $\underline{x}$.

Ο παρακάτω αλγόριθμος εισάγει μηδενικές εισόδους σ'ένα διάνυσμα με χρήση ενός πίνακα Householder.

**Αλγόριθμος**: Μηδενισμός εισόδων διανύσματος χρησιμοποιώντας τον πίνακα Householder.
Έστω $\underline{x}\varepsilon\mathbb{R}^n, \underline{x} \neq \underline{0}$. Ο παρακάτω αλγόριθμος υπολογίζει ένα διάνυσμα $\underline{u}$ και ένα βαθμωτό $\sigma$ έτσι ώστε

$$H\underline{x} = (I - 2\frac{\underline{u}\,\underline{u}^T}{\underline{u}^T\underline{u}})\underline{x} = (\sigma, 0, \ldots, 0)^T$$

Το $\underline{u}$ αποθηκεύεται πάνω στο $\underline{x}$.

**Αλγόριθμος House1**

$m = max|x_i|,\ i = 1, 2, \ldots, n$
$x_i \equiv u_i = \frac{x_i}{m},\ i = 1, 2, \ldots, n$
$\sigma = sign(u_1)\sqrt{u_1^2 + u_2^2 + \ldots + u_n^2}$
$x_1 \equiv u_1 = u_1 + \sigma$
$\sigma = -m\sigma$

Ο αλγόριθμος αυτός για την εκτέλεσή του απαιτεί $2(n+1)$ flops και μία τετραγωνική ρίζα.
Όσον αφορά την ευστάθειά του, αποδεικνύεται ότι

$$\|H - \hat{H}\| \leq 10u$$

Παρατηρούμε ότι το φράγμα που προκύπτει είναι ανεξάρτητο του $n$.

**Παράδειγμα 5.1.1** *Εστω* $\underline{x} = \begin{pmatrix} 0 \\ 4 \\ 1 \end{pmatrix}$. *Να προσδιοριστεί το* $\underline{u}$ *έτσι ώστε*

$$H\underline{x} = (\sigma, 0, 0)^T, \ \text{όπου } H = I - 2\frac{\underline{u}\,\underline{u}^T}{\underline{u}^T\,\underline{u}}$$

$$m = 4,\ u_1 = 1.0308,\ u_2 = 1,\ u_3 = .25$$

*Έτσι* $\underline{u} = (1.0308, 1, 0.25)^T, \sigma = -4 \cdot 1.0308 = -4.1232$
*Εύκολα βλέπουμε ότι*

$$H\underline{x} = (-4.1232, 0, 0)^T$$

Στη συνέχεια αναλύουμε την πολύ σημαντική παραγοντοποίηση $QR$.

**Θεώρημα 5.1.1 Παραγοντοποίηση** $QR$
*Έστω ένας* $n \times n$ *πίνακας* $A$. *Υπάρχει ένας ορθογώνιος πίνακας* $Q$ *και ένας άνω τριγωνικός πίνακας* $R$ *έτσι ώστε:*

$$A = Q \cdot R$$

*Ο πίνακας* $Q$ *μπορεί να γραφεί σαν* $Q = H_1 H_2 \ldots H_{n-1}$, *όπου κάθε πίνακας* $H_1$ *είναι πίνακας Householder. Η παραγοντοποίηση αυτή του* $A$ *ονομάζεται* $QR$ *παραγοντοποίηση του* $A$.

**Απόδειξη:** Το θεώρημα θα αποδειχθεί κατασκευαστικά

**Βήμα 1:** Κατασκευάζουμε πίνακα Householder $H_1$ της μορφής $I_n - 2\frac{\underline{u}_n \underline{u}_n^T}{\underline{u}_n^T \underline{u}_n}$ έτσι ώστε

$$H_1 \cdot \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} = \begin{pmatrix} * \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Στη θέση του $A$ τοποθετείται ο πίνακας $H_1 A$ έτσι έχουμε

$$A \equiv A^{(1)} = H_1 \cdot A = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{pmatrix}$$

**Βήμα 2:** Κατασκευάζουμε πίνακα Householder $\hat{H}_2$ της μορφής $\hat{H}_2 = I_{n-1} - 2\frac{\underline{u}_{n-1} \underline{u}_{n-1}^T}{\underline{u}_{n-1}^T \underline{u}_{n-1}}$, τάξης $n-1$ έτσι ώστε

$$\hat{H}_2 \cdot \begin{pmatrix} a_{22}^{(1)} \\ a_{32}^{(1)} \\ \vdots \\ a_{n2}^{(1)} \end{pmatrix} \begin{pmatrix} * \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Στη συνέχεια ορίζουμε

$$H_2 = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & & & & \\ \vdots & & \hat{H}_2 & & \\ 0 & & & & \end{pmatrix}$$

Στη θέση του $A^{(1)}$ τοποθετείται ο πίνακας $H_2 A^{(1)}$ έτσι ώστε

$$A^{(2)} = H_2 A^{(1)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & & & \\ \vdots & & A_1^{(1)} & \\ 0 & & & \end{pmatrix}$$

**Βήμα** $k$: Κατασκευάζουμε πίνακα Householder $\hat{H}_k$ της μορφής

$$\hat{H}_k = I_{n-k+1} - \frac{2\underline{u}_{n-k+1}\ \underline{u}_{n-k+1}^T}{\underline{u}_{n-k+1}^T \underline{u}_{n-k+1}}$$

τάξης $n-k+1$ έτσι ώστε

$$\hat{H}_k \cdot \begin{pmatrix} a_{kk}^{(k-1)} \\ a_{k+1k}^{(k-1)} \\ \vdots \\ a_{nk}^{(k-1)} \end{pmatrix} = \begin{pmatrix} * \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Στη συνέχεια ορίζουμε

$$H_k = \begin{pmatrix} I_{k-1} & 0 \\ 0 & \hat{H}_k \end{pmatrix}$$

Στη θέση του $A^{(k-1)}$ τοποθετείται ο πίνακας $H_k A^{(k-1)}$ έτσι ώστε

$$A^{(k)} = H_k A^{(k-1)} =$$

$$= \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1k}^{(1)} & \cdots & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2k}^{(2)} & \cdots & \cdots & a_{2n}^{(2)} \\ \vdots & & & & & & \\ 0 & 0 & \cdots & a_{k-1,k}^{(k-1)} & \cdots & \cdots & a_{nk}^{(k-1)} \\ 0 & 0 & \cdots & 0 & & & \\ \vdots & \vdots & & \vdots & & A_{k-1}^{(k-1)} & \\ 0 & 0 & & 0 & & & \end{pmatrix}$$

**Βήμα** $n-1$**:** Στο τέλος του $n-1$ βήματος ο πίνακας $A^{(n-1)}$ θα έχει τη μορφή:

$$A^{(n-1)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & \cdots & a_{2n}^{(2)} \\ & & a_{33}^{(3)} & \cdots & a_{3n}^{(3)} \\ & & & \ddots & \\ & O & & & \\ & & & & a_{nn}^{(n-1)} \end{pmatrix}$$

γίνεται δηλαδή ένας άνω τριγωνικός πίνακας $R$.

Έτσι τελικά $R = A^{(n-1)} = H_{n-1}A^{(n-2)} = H_{n-1}H_{n-2}A^{(n-3)} = \ldots = H_{n-1}H_{n-2}\ldots H_2H_1A$.

Θέτουμε $Q^T = H_{n-1}H_{n-2}\ldots H_2H_1$.

Αφού κάθε πίνακας Householder $H_i$ είναι ορθογώνιος, είναι και το γινόμενο τους ο πίνακας $Q^T$ και ο πίνακας $Q = H_1^T H_2^T \ldots H_{n-1}^T$.

Τελικά προκύπτει ότι

$$R = Q^T A \quad A = QR$$

□

**Παράδειγμα 5.1.2** *Εστω* $A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 1 & 1 \end{pmatrix}$
*Να υπολογισθεί η QR παραγοντοποίησή του. Χρησιμοποιήστε ακρίβεια 4 δεκαδικών ψηφίων.*

**Απόδειξη:** **Βήμα 1:** Υπολογίζουμε τον πίνακα $H_1$ έτσι ώστε:

$$H_1 \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} * \\ 0 \\ 0 \end{pmatrix}$$

$$\text{Θέτουμε } \underline{u}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + \sqrt{2} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \sqrt{2} \\ 1 \\ 1 \end{pmatrix}$$

$$H_1 = I_3 - \frac{2\underline{u}_3\underline{u}_3^T}{\underline{u}_3^T\underline{u}_3} = \begin{pmatrix} 0 & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{\sqrt{2}} & -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

$$A \equiv A^{(1)} = H_1A = \begin{pmatrix} -\sqrt{2} & -\frac{3\sqrt{2}}{2} & 2\sqrt{2} \\ 0 & \frac{1-\sqrt{2}}{2} & \frac{2-\sqrt{2}}{2} \\ 0 & -(1+\sqrt{2}) & -(2+\sqrt{2}) \end{pmatrix}$$

Τελικά

$$A^{(1)} = \begin{pmatrix} -1.4142 & -2.1213 & -2.8284 \\ 0 & -0.2071 & 0.2929 \\ 0 & -1.2071 & -1.7071 \end{pmatrix}$$

**Βήμα 2:** Υπολογίζουμε τον πίνακα $\hat{H}_2$ έτσι ώστε

$$\hat{H}_2 \begin{pmatrix} -0.2071 \\ -1.2071 \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}$$

$$\underline{u}_2 = \begin{pmatrix} -0.2071 \\ -1.2071 \end{pmatrix} - 1.2247 \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -1.4318 \\ -1.2071 \end{pmatrix}$$

$$\hat{H}_2 = \begin{pmatrix} -0.1691 & -0.9856 \\ -0.9856 & 0.1691 \end{pmatrix}$$

$$A^{(2)} = H_2 A^{(1)} = \begin{pmatrix} -1.4142 & -2.1213 & -2.8284 \\ 0 & 1.2247 & 1.6330 \\ 0 & 0 & -0.5774 \end{pmatrix} = R$$

$$Q = H_1 H_2 = \begin{pmatrix} 0 & 0.8165 & 0.5774 \\ -0.7071 & 0.4082 & -0.5774 \\ -0.7071 & -0.4082 & 0.5774 \end{pmatrix}$$

□

## Παρατηρήσεις

1) Κάθε πίνακας Householder $H_k$ ορίζεται μονοσήμαντα από το διάνυσμα $\underline{u}_{n-k+1}$ που καθορίζει τον πίνακα $\hat{H}_k$. Επομένως αρκεί κάθε φορά να αποθηκεύουμε μόνο το διάνυσμα $\underline{u}_{n-k+1}$. Εάν χρειαζόμαστε ολόκληρο τον πίνακα $Q$, αυτός μπορεί να υπολογισθεί από τους πίνακες Householder $H_1$ εώς $H_{n-1}$.

2) Το διάνυσμα $\underline{u}_{n-k+1} = (u_{kk}, \ldots, u_{nk})^T$ έχει $n-k+1$ συνιστώσες και παράγει $(n-k)$ μηδενικά στο $k$ βήμα. Έτσι μπορούμε να αποθηκεύσουμε τη 2η έως τη $(n-k+1)$ συνιστώσα του στις θέσεις $(k+1,k), \ldots, (n,k)$ του πίνακα $A$. Η πρώτη συνιστώσα θα αποθηκεύεται ξεχωριστά σ'ένα μονοδιάστατο array.

3) Για τον υπολογισμό του γινομένου $A^{(k)} = H_k A^{(k-1)}$ αρκεί μόνο υπολογισμός του διανύσματος $\underline{u}_{n-k+1}$ που καθορίζει τον $H_k$. Πράγματι, όπως φαίνεται και από την απόδειξη του θεωρήματος $QR$, ο πίνακας $A^{(k)}$ προκύπτει από τον $A^{(k-1)}$ως εξής:
*i*. Οι πρώτες $(k-1)$ γραμμές παραμένουν αναλλοίωτες.
*ii*. Οι υπόλοιπες (ν-κ+1) γραμμές προκύπτουν από τη σχέση
$(I - \frac{2\underline{u}_{n-k+1}\underline{u}_{n-k+1}^T}{\underline{u}_{n-k+1}^T \underline{u}_{n-k+1}}) A_{k-1}^{(k-1)}$, όπου $A_{k-1}^{(k-1)}$ είναι ο
πίνακας που προκύπτει από τον $A^{(k-1)}$ εάν απομακρυνθούν οι πρώτες $(k-1)$ γραμμές και στήλες του. Από τη σχέση αυτή φαίνεται ότι το γινόμενο μπορεί να υπολογισθεί αποθηκεύοντας μόνο το διάνυσμα $\underline{u}_{n-k+1}$.

### Αλγόριθμος $QR$: Householder $QR$ παραγοντοποίηση

Εστω ένας $n \times n$ πίνακας $A$. Ο παρακάτω αλγόριθμος δημιουργεί τα διανύσματα $\underline{u}_{n-k+1} = (u_{kk}, \ldots, u_{nk})^T,\ k = 1, \ldots, n-1$ που ορίζουν τους πίνακες Householder $H_1, \ldots, H_{n-1}$, και έναν άνω τριγωνικό πίνακα $R$ έτσι ώστε $A = QR$ όπου $Q = H_1 H_2 \ldots H_{n-1}$.
Οι συνιστώσες $u_{k+1,k}, \ldots, u_{nk}$ αποθηκεύονται στις θέσεις $(k+1, k), \ldots, (n, k)$ του $A$.Οι πρώτες συνιστώσες $u_{kk}$ αποθηκεύονται σ΄ ένα μονοδιάστατο array

$$\underline{u} = (u_1, u_2, \ldots, u_n)^T.$$

Οι μη μηδενικές είσοδοι του

$$A^{(k)} = H_k A^{(k-1)},\ k = 1, \ldots, n-1$$

επικαλύπτουν αυτές του $A$, έτσι ο $R$ καταχωρείται στο άνω τριγωνικό μέρος του $A$.

### Αλγόριθμος $QR$

For $k = 1, 2, \ldots, n-1$
**Βήμα 1:** Να προσδιοριστεί διάνυσμα
$\underline{u}_{n-k+1} = (u_{kk}, \ldots, u_{nk})^T$
που ορίζει τον $\hat{H}_k$ έτσι ώστε

$$\hat{H}_k \begin{pmatrix} a_{kk} \\ \vdots \\ \vdots \\ a_{nk} \end{pmatrix} = \begin{pmatrix} \sigma \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$a_{kk} \equiv \sigma$
$a_{ik} \equiv u_{ik},\ i = k+1, \ldots, n$
$u_k \equiv u_{kk}$
$\beta = \frac{2}{\underline{u}^T_{n-k+1} \underline{u}_{n-k+1}}$

**Βήμα 2:** Ενημέρωση των εισόδων του υποπίνακα του $A$ που περιέχει τις γραμμές από $k$ έως $n$ και στήλες από $k+1$ έως $n$
**For** $j = k+1, \ldots, n$
$s = b \sum_{i=k}^{n} u_{ik} a_{ij}$
$a_{ij} = a_{ij} - s u_{ik},\ i = k, k+1, \ldots, n$

**Αποτελεσματικότητα:** Για την κατασκευή κάθε $\hat{H}_k$ απαιτούνται περίπου $2(n-k)$ flops ενώ για την κατασκευή κάθε $A^{(k)}$ από τη σχέση $A^{(k)} = H_k A$ (λαμβάνοντας υπ΄όψη την ειδική δομή των $H_k$) απαιτούνται περίπου $2(n-k)^2$ φλοπς. Έτσι Συνολικός αριθμός flops $= 2 \sum_{k=1}^{n-1} [(n-k)^2 + (n-k)] =$

$$2\frac{n(n-1)(2n-1)}{6} + 2\frac{n(n-1)}{2} \simeq \frac{2n^3}{3}$$

Σάυτόν τον αριθμό των flops δεν περιλαμβάνεται η αναλυτική κατασκευή του $Q$. Στις περισσότερες των περιπτώσεων είναι αρκετό να έχουμε τον $Q$ σε παραγοντοποιημένη μορφή ενώ σε άλλες εφαρμογές ο $Q$ δε χρειάζεται καθόλου. Εάν παρόλα αυτά απαιτείται ο υπολογισμός του $Q$, ο αλγόριθμος χρειάζεται επιπλέον $\frac{2}{3}n^3$ flops.

**Ευστάθεια:** Εχει αποδειχθεί ότι οι πίνακες $\hat{R}$ και $\hat{Q}$ που υπολογίζουμε ικανοποιούν τη σχέση

$$(A+E) = \hat{Q}\hat{R}$$

όπου ο πίνακας σφάλματος $E$ ικανοποιεί το φράγμα

$$\|E\|_F \leq f(n)u\|A\|_F$$

**Παράδειγμα 5.1.3** *Να προσδιοριστεί η $QR$ παραγοντοποίηση του πίνακα*

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 7 \\ 3 & 8 & 9 \end{pmatrix}$$

*χρησιμοποιώντας τη μέθοδο Householder. Να χρησιμοποιηθεί ακρίβεια 4 δεκαδικών ψηφίων.*

**Απόδειξη:** $k = 1$ : Χρησιμοποιώντας τον αλγόριθμο House1 προσδιορίζουμε διάνυσμα $\underline{u}_3 = (u_{11}, u_{21}, u_{31})^T$ έτσι ώστε

$$\hat{H}_1 \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \left( I - \frac{2\underline{u}_3\underline{u}_3^T}{\underline{u}_3^T\underline{u}_3} \right) \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} \sigma \\ 0 \\ 0 \end{pmatrix}$$

$\underline{u}_3 = (1.5806, 0.6667, 1)^T$
$\sigma = -3.7417, H_1 = \hat{H}_1$

$$A^{(1)} \equiv H_1 A = \begin{pmatrix} -3.7417 & -9.6214 & -11.7595 \\ 0 & 0.0982 & 0.7745 \\ 0 & 0.6473 & -0.3382 \end{pmatrix}$$

$k = 2$: Χρησιμοποιώντας τον αλγόριθμο House1 προσδιορίζουμε διάνυσμα $\underline{u}_2 = (u_{22}, u_{32})^T$ έτσι ώστε

$$\hat{H}_2 \begin{pmatrix} 0.0982 \\ 0.6473 \end{pmatrix} = \begin{pmatrix} s \\ 0 \end{pmatrix}$$

$$\underline{u}_2 = [1.1631, 1]^T,\ \sigma = -0.6547,\ H_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \hat{H}_2 & \\ 0 & & \end{pmatrix}$$

$$A^{(2)} = H_2 A^{(1)} = H_2 H_1 A = R = \begin{pmatrix} -3.7417 & -9.6214 & -11.7595 \\ 0 & -0.6547 & 0.2182 \\ 0 & 0 & -0.8165 \end{pmatrix}$$

**Δημιουργία του πίνακα $Q$:**
Ο πίνακας $Q$ μπορεί να κατασκευασθεί από τα διανύσματα $\underline{u}_3$ και $\underline{u}_2$. Συγκεκριμένα,

$$Q^T = H_2 H_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{I - 2\underline{u}_2 \underline{u}_2^T}{\underline{u}_2^T \underline{u}_2} & \\ 0 & & \end{pmatrix} \left( I - \frac{2\underline{u}_3 \underline{u}_3^T}{\underline{u}_3^T \underline{u}_3} \right) =$$

$$\begin{pmatrix} -0.2673 & 0.8729 & 0.4082 \\ -0.5345 & 0.2182 & -0.8165 \\ -0.8018 & -0.4364 & 0.4082 \end{pmatrix}$$

Εύκολα επαληθεύεται ότι: $Q^T A = R$ □

## 5.2 Μετασχηματισμοί Householder για ορθογώνιο πίνακα

Σε πολλές εφαρμογές (όπως σε προβλήματα ελαχίστων τετραγώνων) χρειάζεται η $QR$ παραγοντοποίηση ενός πίνακα $A$ $m \times n$. Η παραπάνω μέθοδος μπορεί να χρησιμοποιηθεί μόνο που τώρα θα απαιτούνται $s = \min\{n, m-1\}$ βήματα. Έτσι θα κατασκευάζονται διαδοχικά πίνακες Householder $H_1, H_2, \ldots, H_s$ που τελικά θα ικανοποιούν τη σχέση

$$H_s H_{s-1} \ldots H_2 H_1 A = Q^T A = \begin{cases} \begin{pmatrix} R_1 \\ 0 \\ \end{pmatrix} & , \text{εάν } m > n \\ \\ \begin{pmatrix} R_1 & S \end{pmatrix} & , \text{εάν } m \leq n \end{cases}$$

**Απαιτούμενα flops:** Σάυτή την περίπτωση τα φλοπς που χρειάζονται για την εκτέλεση του αλγορίθμου είναι:

$$\begin{cases} n^2(m-\frac{n}{3}) & , \; en \; m \geq n \\ m^2(n-\frac{m}{3}) & , \; en \; m \leq n \end{cases}$$

**Ευστάθεια:** Η $QR$ παραγοντοποίηση ενός ορθογώνιου πίνακα με χρήση μετασχηματισμών Householder είναι ευσταθής.

**Παράδειγμα 5.2.1** *Έστω*

$$A = \begin{pmatrix} 1 & 1 \\ 0.0001 & 0 \\ 0 & 0.0001 \end{pmatrix}, \; s = min(2,2) = 2$$

**Βήμα 1:** *Δημιουργία του* $H_1$ *από το διάνυσμα* $\underline{u}_2$

$$\underline{u}_2 = \begin{pmatrix} 1 \\ 0.0001 \\ 0 \end{pmatrix} + \sqrt{1+(0.0001)^2} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 0.0001 \\ 0 \end{pmatrix}$$

$$H_1 = I - \frac{2\underline{u}_2\underline{u}_2^T}{\underline{u}_2^T\underline{u}_2}$$

$$A^{(1)} = H_1 A = \begin{pmatrix} -1 & -1 \\ 0 & -0.0001 \\ 0 & 0.0001 \end{pmatrix}$$

**Βήμα 2:** *Δημιουργία του* $H_2$ *από το διάνυσμα* $\underline{u}_1$

$$\underline{u}_1 = \begin{pmatrix} -0.0001 \\ 0.0001 \end{pmatrix} - \sqrt{(-0.0001)^2+(0.0001)^2} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 10^{-4} \begin{pmatrix} -2.4141 \\ 0.1000 \end{pmatrix}$$

$$\hat{H}_2 = I - \frac{2\underline{u}_1\underline{u}_1^T}{\underline{u}_1^T\underline{u}_1} = \begin{pmatrix} -0.7071 & 0.7071 \\ 0.7071 & 0.7071 \end{pmatrix}$$

$$H_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \hat{H}_2 & \\ 0 & & \end{pmatrix}$$

$$A^{(2)} = H_2 A^{(1)} = H_2 H_1 A = \begin{pmatrix} -1 & -1 \\ 0 & 0.0001 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} R_1 \\ 0 \end{pmatrix}$$

*Τελικά*

$$Q = H_1 H_2 = \begin{pmatrix} -1 & 0.0001 & -0.0001 \\ -0.0001 & -0.7071 & 0.7071 \\ 0 & 0.7071 & 0.7071 \end{pmatrix},$$

$$R_1 = \begin{pmatrix} -1 & -1 \\ 0 & 0.0001 \end{pmatrix}$$

# Κεφάλαιο 6

# Αριθμητική Επίλυση Γραμμικών Συστημάτων

Το πρόβλημα της αριθμητικής επίλυσης του γραμμικού συστήματος

$$Ax = b$$

εμφανίζεται σε ένα ευρύ φάσμα εφαρμογών. Δεν είναι υπερβολή να αναφέρουμε το γεγονός ότι σχεδόν όλα τα προβλήματα των εφαρμοσμένων επιστημών απαιτούν επίλυση ενός προβλήματος γραμμικού συστήματος. Συγκεκριμένα θα παρουσιάσουμε μεθόδους για μη ιδιάζοντα τετραγωνικά συστήματα. Η επίλυση συστημάτων με τη μέθοδο του Cramer έχει αξία μόνο θεωρητικά. Εάν $A$ είναι ένας $n \times n$ μη ιδιάζων πίνακας και $\underline{b}$ ένα $n$-διάστατο διάνυσμα, η λύση $\underline{x}$ του συστήματος $A\underline{x} = \underline{b}$ δίνεται από τη σχέση

$$x_i = \frac{\det(A_i)}{\det(A)},\ i =, 1, 2, \ldots, n$$

όπου $A_i$ ο πίνακας που προκύπτει εάν αντικατασταθεί η $i$ στήλη του $A$ από τπ διάνυσμα $\underline{b}$. Ο κανόνας του Cramer δεν είναι καθόλου πρακτικός από υπολογιστικής πλευράς. Για παράδειγμα, η επίλυση ενός συστήματος $20 \times 20$ με τον κανόνα του Cramer και τη χρήση του απλού ορισμού των οριζουσών, θα απαιτούσε περισσότερο από ένα εκατομμύριο χρόνια ακόμα και σε ένα πολύ γρήγορο υπολογιστή. Για ένα $n \times n$ σύστημα απαιτεί περίπου $O(n!)$ flops. Για την αριθμητική επίλυση γραμμικών συστημάτων συνήθως χρησιμοποιούνται δύο τύποι μεθόδων.

1. Αμεσοι μέθοδοι (Direct methods)
2. Επαναληπτικές μέθοδοι (Iterative methods)

Στη συνέχεια θα παρουσιάσουμε τις πιο σημαντικές από αυτές.

## 6.1 'μεσοι Μέθοδοι

Οι άμεσοι μέθοδοι εκτελούν ένα πεπερασμένο αριθμό βημάτων πριν προσδιορισθεί η λύση. Η βασική ιδέα έγκειται στην μετατροπή του γραμμικού συστήματος σε ένα ισοδύναμο τριγωνικό σύστημα το οποίο επιλύεται εύκολα. Η μετατροπή αυτή υλοποιείται προσδιορίζοντας τους τριγωνικούς παράγοντες του πίνακα $A$ χρησιμοποιώντας τον κατάλληλο αλγόριθμο.

### 6.1.1 Επίλυση ενός άνω τριγωνικού συστήματος

Εστω το σύστημα

$$T\underline{y} = \underline{b}$$

όπου $T = (t_{ij})$ , ένας άνω τριγωνικός πίνακας με $t_{ii} \neq 0,\ i = 1, \ldots, n$.
Ο παρακάτω αλγόριθμος της προς τα πίσω αντικατάστασης υπολογίζει τη λύση του συστήματος αυτού.

$y_n := \frac{b_n}{t_{nn}}$
**For** $i = n-1, \ldots, 3, 2, 1$
$\quad y_i = \frac{1}{t_{ii}}(b_i - \sum_{j=i+1}^{n} t_{ij} y_j)$

Απαιτούμενα flops: Για τον υπολογισμό του $y_n$ απαιτείται 1 flop, 2 flops για το $y_{n-1}$ κ.ο.κ. Τελικά χρειάζονται

$$(1 + 2 + 3 + \ldots + n) = \frac{n(n+1)}{2} \simeq \frac{n^2}{2}\ flops$$

### 6.1.2 Επίλυση ενός κάτω τριγωνικού συστήματος

Εστω το σύστημα

$$L\underline{y} = \underline{b}$$

όπου $L = (l_{ij})$, ένας κάτω τριγωνικός πίνακας με $l_{ii} \neq 0,\ i = 1, \ldots, n$. Ο παρακάτω αλγόριθμος της προς τα μπρος αντικατάστασης υπολογίζει τη λύση του συστήματος αυτού.

$y_1 := \frac{b_1}{l_{11}}$
**For** $i = 2, \ldots, n$
$\quad y_i = \frac{1}{l_{ii}}(b_i - \sum_{j=1}^{i-1} l_{ij} y_j)$

Απαιτούμενα flops: Ο αλγόριθμος απαιτεί περίπου $\frac{n^2}{2}$ flops (1 flop για τον υπολογισμό του $y_1$, 2 flops για τον υπολογισμό του $y_2$, 3 flops για τον υπολογισμό του $y_3$ κ.ο.κ.)

**Ευστάθεια**

## Ανάλυση σφάλματος στην επίλυση ενός κάτω τριγωνικού συστήματος

Θα μελετήσουμε το σφάλμα στη λύση ενός κάτω τριγωνικού συστήματος

$$L\underline{x} = \underline{b}$$

Αναλυτικά γράφουμε τις εξισώσεις στην εξής μορφή:

$$\begin{pmatrix} l_{11}x_1 & & & & & = b_1 \\ l_{21}x_1 & l_{22}x_2 & & & & = b_2 \\ l_{31}x_1 & l_{32}x_2 & l_{33}x_3 & & & = b_3 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ l_{n1}x_1 & l_{n2}x_2 & l_{n3}x_3 & \cdots & l_{nn}x_n & = b_n \end{pmatrix}$$

Οι μεταβλητές $x_1, x_2, \ldots$ ορίζονται διαδοχικά από τις εξισώσεις $1, 2, \ldots$ αντίστοιχα. Τη στιγμή που υπολογίζεται το $x_r$ από την $r$ εξίσωση τα $x_1, x_2, \ldots, x_{r-1}$ έχουν ήδη υπολογισθεί.
Σε floating-point αριθμητική έχουμε:

$$x_r \equiv fl(\frac{-l_{r_1x_1} - l_{r_2x_2} - \ldots - l_{r,r-1}x_{r-1} + b_r}{l_{rr}}) \tag{6.1}$$

**Σημείωση**: Εάν $a, b$ είναι δύο floating-point αριθμοί και $\square$ συμβολίζει μία από τις πράξεις $\{+, -, \cdot, /\}$ τότε έχουμε δείξει ότι ισχύει το ακόλουθο:

$$fl(a\square b) = \frac{a\square b}{1+\varepsilon}, |\varepsilon| \leq u$$

Κατά συνέπεια η σχέση (5.1) γίνεται:

$$x_r \equiv \frac{-l_{r1}x_1(1+E_{r1}) - l_{r2}x_2(1+E_{r2}) - \ldots - l_{rr-1}x_{r-1}(1+E_{rr-1}) + b_r(1+e_r)}{l_{rr}(1+n_r)} \tag{6.2}$$

όπου $|n_r| \leq u_1,\ |\varepsilon_r| \leq u_1\ |E_{ri}| \leq (r-i+2)u_1$.

Αυτό το αποτέλεσμα προέκυψε επειδή όλος ο αριθμητής είναι ένα εσωτερικό γινόμενο με τελευταίο γινόμενο τον όρο $b_r$ μόνο του. Ο παράγοντας $1+n_r$ καλύπτει το σφάλμα από τη διαίρεση με $l_{rr}$.
Εκφράζουμε την (5.2) στην ακόλουθη μορφή.

$$x_r \equiv \frac{-l_{r1}x_1\frac{1+E_{r1}}{1+\varepsilon_r} - l_{r2}x_2\frac{1+E_{r2}}{1+\varepsilon_r} - \ldots - l_{rr-1}x_{r-1}\frac{1+E_{r,r-1}}{1+\varepsilon_r} + b_r}{l_{rr}\frac{1+n_r}{1+\varepsilon_r}} \tag{6.3}$$

Εάν γράψουμε :

$$\frac{1+E_{ri}}{1+\varepsilon_r} = 1 + F_{ri}$$

$$\frac{1+n_r}{1+\varepsilon_r} = 1 + F_{rr} \tag{6.4}$$

η εξίσωση (5.3) παίρνει τη μορφή:

$$l_{r1}x_1(1+F_{r1}) + l_{r2}x_2(1+F_{r2}) + \ldots + l_{rr}x_r(1+F_{rr}) \equiv b_r \Rightarrow$$

$$l_{r1}x_1 + l_{r2}x_2 + \ldots + l_{rr}x_r + l_{r1}F_{r1}x_1 + l_{r2}F_{r2}x_2 + \ldots + l_{rr}F_{rr}x_r \equiv b_r \tag{6.5}$$

Από την (5.4) προκύπτει ότι η λύση που θα υπολογίσουμε είναι η ακριβής λύση του συστήματος.

$$(L + \Delta L)x = b$$

όπου $\Delta l_{ij} = l_{ij}F_{ij}$

Για να δώσουμε τώρα ένα κατάλληλο φράγμα στον πίνακα $\Delta L$, από τις (5.2) και (5.4) παρατηρούμε ότι

$$|F_{rr}| \leq 2u_1$$

επειδή δε ο όρος $(1+\varepsilon_r)$ εμφανίζεται σε κάθε $(1+E_{ri})$

$$|F_{ri}| \leq (r-i+1)u_1$$

Ειδικά για το $\Delta l_{11}$ έχουμε ότι $|\Delta l_{11}| = |l_{11}n_1| \leq u_1|l_{11}|$
Κατά συνέπεια ο πίνακας $\Delta L$ φράσσεται ως ακολούθως:

$$|\Delta L| \leq u_1 \begin{pmatrix} |l_{11}| & 0 & \cdots & \cdots & \cdots & \vdots \\ 2|l_{21}| & 2|l_{22}| & 0 & \cdots & \cdots & \vdots \\ 3|l_{31}| & 2|l_{32}| & 2|l_{33}| & 0 & \cdots & \vdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ n|l_{n1}| & (n-1)|l_{n2}| & \cdots & \cdots & 2|l_{nn-1}| & 2|l_{nn}| \end{pmatrix}$$

Εάν $|l_{ij}| \leq g$ και εάν πάρουμε την $\|\cdot\|_\infty$ τότε

$$\|\Delta L\|_\infty \leq \frac{1}{2}(n^2+n+2)gu_1$$

Γενικά για την $\|\cdot\|_1, \|\cdot\|_\infty$ και $\|\cdot\|_E$ ισχύει

$$\|\Delta L\| \leq nu_1\|L\|$$

**Λήμμα 6.1.1** *Εστω το κάτω τριγωνικό σύστημα* $L\underline{y} = \underline{b}$. *Η λύση* $\underline{y}$ *που προκύπτει με floating point αριθμητική ικανοποιεί το σύστημα*

$$(L + \Delta L)\underline{y} = \underline{b}$$

*όπου*

$$\|\Delta L\|_\infty \leq \frac{n(n+1)+2}{2}u_1$$

**Λήμμα 6.1.2** *Εστω το άνω τριγωνικό σύστημα* $U\underline{x} = \underline{y}$. *Η λύση* $\underline{x}$ *που προκύπτει με floating point αριθμητική ικανοποιεί το σύστημα*

$$(U + \Delta U)\underline{x} = y$$

*όπου*

$$\|\Delta U\|_\infty \leq \frac{n(n+1)+2}{2} u_1 \cdot max|u_{ij}| \leq \frac{n(n+1)+2}{2} u_1 r \|A\|_\infty$$

## Ακρίβεια της υπολογιζόμενης λύσης

Εάν στο πίνακα $L$ και στο διάνυσμα $\underline{b}$ έχει γίνει κατάλληλο σκαλινγ έτσι ώστε όλα τα στοιχεία τους να έχουν μέγεθος μικρότερο της μονάδας μπορούμε να υποθέσουμε ότι $g = 1$ και

$$\underline{b} - L\underline{x} = \Delta L x$$

$$\|\underline{b} - L\underline{x}\|_\infty = \|\Delta L \underline{x}\|_\infty \leq \frac{1}{2}(n^2 + n + 2)\|\underline{x}\|_\infty$$

Η ποσότητα $\|\underline{b} - L\underline{x}\|_\infty$ ονομάζεται διάνυσμα υπόλοιπο (residual vector) και η παραπάνω σχέση μας δίνει ένα φράγμα για το μέγιστο στοιχείο του residual σε σχέση με το μέγιστο στοιχείο του $\underline{x}$. Παρατηρούμε ότι το φράγμα αυτό εξαρτάται άμεσα από το $\|\underline{x}\|_\infty$.
Εάν $\|\underline{x}\|_\infty$ είναι της τάξης της μονάδας το residual είναι κατάνάγκη πολύ μικρό ανεξάρτητα από το αν το $\underline{x}$ είναι ακριβής λύση.
Ο παράγοντας $\frac{1}{2}(n^2 + n + 2)$ είναι πολύ σπάνιο να εμφανιστεί στην πράξη και ακόμα και η αντικατάστασή του με την τετραγωνική του ρίζα αποτελεί ικανοποιητικό φράγμα για το υπόλοιπο που εμφανίζεται.

Στη συνέχεια θα εξετάσουμε αναλυτικά την αριθμητική επίλυση του συστήματος $A\underline{x} = \underline{b}$ χρησιμοποιώντας τις παρακάτω μεθόδους:

## Απαλοιφή Gauss χωρίς οδήγηση

Βήμα 1: Προσδιορίστε την παραγοντοποίηση
$A = L \cdot U$
Βήμα 2: Επιλύστε διαδοχικά τα δύο τριγωνικά συστήματα
$L \cdot \underline{y} = \underline{b}$
$U \cdot \underline{x} = \underline{y}$

Απαιτούμενα flops. Για την παραγοντοποίηση χρειάζονται $\frac{n^3}{3}$ flops ενώ για την επίλυση κάθε τριγωνικού συστήματος απαιτούνται $\frac{n^2}{2}$ flops. Έτσι συνολικά χρειάζονται $\frac{n^3}{3} + n^2$ flops.

## Απαλοιφή Gauss με μερική οδήγηση

Βήμα 1: Προσδιορίστε την παραγοντοποίηση
$M \cdot A = U$
Βήμα 2: Επιλύστε το άνω τριγωνικό σύστημα
$U \cdot \underline{x} = M \cdot \underline{b} = \underline{b}'$

Ο υπολογισμός του διανύσματος $b' = M \cdot \underline{b} = M_{n-1}P_{n-1}M_{n-2}P_{n-2}\dots M_1P_1\underline{b}$ θα γίνεται ως εξής:

Βήμα 1: $\underline{b} = (b_1, b_2, \dots, b_n)^T$
Βήμα 2: **For** $k = 1, 2, \dots, n-1$
$\underline{b} = M_kP_k\underline{b}$

Στην πράξη δε χρειάζεται ο αναλυτικός υπολογισμός των πινάκων $M_k$ και $P_k$. Το διάνυσμα $\underline{b}$ στο $k+1$ βήμα μπορεί να υπολογισθεί αμέσως εάν γνωρίζουμε το δείκτη $r_k$ των αλλαγών γραμμών καθώς και τους πολλαπλασιαστές $m_{ik}$ που καθορίστηκαν στο $k$ βήμα. Αυτό φαίνεται στο ακόλουθο παράδειγμα:

**Παράδειγμα 6.1.1** *Εστω*

$$n = 3,\ P_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},\ M_1 = \begin{pmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & 0 & 1 \end{pmatrix}$$

*και έστω*

$$\underline{b} = M_1P_1\underline{b} = \begin{pmatrix} b_1^{(2)} \\ b_2^{(2)} \\ b_3^{(2)} \end{pmatrix}$$

*Τότε όμως προκύπτει*

$$P_1\underline{b} = \begin{pmatrix} b_1 \\ b_3 \\ b_2 \end{pmatrix}$$

*και οι είσοδοι του $\underline{b}$ στο δεύτερο βήμα δίνονται από τις σχέσεις*

$$b_1^{(2)} = b_1$$

$$b_2^{(2)} = m_{21}b_1 - b_3$$

$$b_3^{(2)} = m_{31}b_1 - b_2$$

## Απαλοιφή Gauss με ολική οδήγηση

Βήμα 1: Προσδιορίστε την παραγοντοποίηση
$M \cdot A \cdot Q = U$
Βήμα 2: Επιλύστε το άνω τριγωνικό σύστημα
$U \cdot \underline{y} = M \cdot \underline{b} = \underline{b'}$
Βήμα 3: Υπολογίστε το $\underline{x}$ από τον τύπο
$\underline{x} = Q \cdot \underline{y}$

Για τον υπολογισμό του διανύσματος $\underline{x}$ θα χρησιμοποιήσουμε τον τύπο $\underline{x} = Q\underline{y} = Q_1 Q_2 \dots Q_{n-1}\underline{y}$, ο οποίος υλοποιείται ως εξής:

Βήμα 1: $\underline{x} = \underline{y}$
Βήμα 2: **For** $k = n-1, \dots, 2, 1$
$\underline{x} = Q_k \underline{x}$

Επειδή ο πίνακας $Q_k$ είναι μεταθετικός, οι είσοδοι του $\underline{x}$ είναι απλά κάθε φορά κατάλληλα μετατιθεμένες.

**Παράδειγμα 6.1.2** *Να επιλυθεί το σύστημα*

$$A\underline{x} = \underline{b}$$

$$\text{όπου } A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 1 & 1 \end{pmatrix},\ \underline{b} = \begin{pmatrix} 2 \\ 6 \\ 3 \end{pmatrix}$$

*(α) χρησιμοποιώντας μερική οδήγηση και (β) χρησιμοποιώντας ολική οδήγηση.*

**Μερική οδήγηση**
Χρησιμοποιώντας τα αποτελέσματα προηγούμενου παραδείγματος έχουμε:

$$P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},\ P_1\underline{b} = \begin{pmatrix} 6 \\ 2 \\ 3 \end{pmatrix}$$

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix},\ M_1 P_1 \underline{b} = \begin{pmatrix} 6 \\ 2 \\ -3 \end{pmatrix}$$

$$P_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},\ P_2 M_1 P_1 \underline{b} = \begin{pmatrix} 6 \\ 2 \\ -3 \end{pmatrix}$$

$$M_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix},\ \underline{b}^{'} = M_2 P_2 M_1 P_1 \underline{b} = \begin{pmatrix} 6 \\ 2 \\ -1 \end{pmatrix}$$

Επιλύουμε το άνω τριγωνικό σύστημα

$$U \cdot \underline{x} = \underline{b}'$$

$$\text{όπου } U = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 1 \\ 0 & 0 & -1 \end{pmatrix},\ \underline{b}' = \begin{pmatrix} 6 \\ 2 \\ -1 \end{pmatrix}$$

Τελικά προκύπτει $\underline{x} = [1, 1, 1]^T$

**Ολική Οδήγηση**

Χρησιμοποιώντας τα αποτελέσματα προηγούμενου παραδείγματος έχουμε:

$$P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},\ P_1 \underline{b} = \begin{pmatrix} 6 \\ 2 \\ 3 \end{pmatrix},\ Q_1 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{1}{3} & 1 & 0 \\ -\frac{1}{3} & 0 & 1 \end{pmatrix},\ M_1 P_1 \underline{b} = \begin{pmatrix} 6 \\ 0 \\ 1 \end{pmatrix}$$

$$P_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},\ P_2 M_1 P_1 \underline{b} = \begin{pmatrix} 6 \\ 1 \\ 0 \end{pmatrix},\ Q_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

$$M_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{2} & 1 \end{pmatrix},\ \underline{b}' = M_2 P_2 M_1 P_1 \underline{b} = \begin{pmatrix} 6 \\ 1 \\ \frac{1}{2} \end{pmatrix}$$

Επιλύουμε το άνω τριγωνικό σύστημα:

$$U \cdot \underline{y} = \underline{b}'$$

$$pou\ U = \begin{pmatrix} 3 & 1 & 2 \\ 0 & \frac{2}{3} & \frac{1}{3} \\ 0 & 0 & \frac{1}{2} \end{pmatrix},\ \underline{b}' = \begin{pmatrix} 6 \\ 1 \\ \frac{1}{2} \end{pmatrix}$$

Η λύση $y = [1, 1, 1]^T$

Η τελική λύση $\underline{x} = Q \cdot \underline{y} = Q_1 Q_2 \underline{y} = [1,1,1]^T$.
Στην προκειμένη περίπτωση αφού όλες οι συνιστώσες του $\underline{y}$ είναι μονάδες, η λύση $\underline{q}$ προκύπτει με απλές αναδιατάξεις των συνιστωσών που δίνουν πάλι μονάδες και έτσι δε χρειάζεται καθόλου ο υπολογισμός του γινομένου $Q_1 Q_2 \underline{y}$.

**Απαιτούμενα flops**:
Η διαδικασία της τριγωνοποίησης απαιτεί $\frac{n^3}{3}$ flops.
Η επίλυση των τριγωνικών συστημάτων $U\underline{x} = \underline{b}'$ $U\underline{y} = \underline{b}'$ απαιτούν $\frac{n^2}{2}$ flops και για τον υπολογισμό του διανύσματος $\underline{b}'$ χρειάζονται $n^2$ flops λαμβάνοντας υπόψη την ειδική δομή των πινάκων $M_k$ και $P_k$. Στην ολική οδήγηση η εύρεση του $\underline{q}$ από τον τύπο $\underline{x} = Q\underline{y}$ δεν απαιτεί flops αφού απλώς οι είσοδοι του $\underline{q}$ προκύπτουν από κατάλληλες αναδιατάξεις των εισόδων του $\underline{y}$. Έτσι η λύση του συστήματος $A\underline{x} = \underline{b}$ με μερική ή ολική οδήγηση απαιτεί $\frac{n^3}{3} + O(n^2)$ flops. Στην περίπτωση της ολικής οδήγησης απαιτούνται $\frac{n^3}{3}$ συγκρίσεις ενώ εάν υλοποιηθεί η μερική οδήγηση χρειάζονται μόνο $O(n^2)$ συγκρίσεις.

**Ευστάθεια**
Το παρακάτω Θεώρημα εξετάζει την ευστάθεια της αριθμητικής επίλυσης ενός συστήματος με την απαλοιφή του Gauss.

**Λήμμα 6.1.3** *Έστω το σύστημα $A\underline{x} = \underline{b}$. Η λύση που προκύπτει με απαλοιφή Gauss με μερική οδήγηση και χρήση floating point αριθμητικής είναι ακριβής λύση του συστήματος*

$$(L + \Delta L)(U + \Delta U)\underline{x} = \underline{b}$$

$$\text{όπου } \|\Delta L\|_\infty \leq \frac{n(n+1)+2}{2} u_1$$

$$\|\Delta U\|_\infty \leq \frac{n(n+1)+2}{2} u_1 p \|A\|_\infty$$

**Θεώρημα 6.1.1** *Εστω το σύστημα $A\underline{x} = \underline{b}$. Η προσεγγιστική λύση $\underline{x}$ που δίνει η απαλοιφή Gauss με μερική οδήγηση σε floating point αριθμητική είναι ακριβής λύση του συστήματος*

$$(A + \Delta A)\underline{x} = \underline{b}$$

$$\text{όπου } \|\Delta A\|_\infty \leq (n^3 + 2n^2 + 2n) u_1 p \|A\|_\infty$$

**Απόδειξη:** Γνωρίζουμε ότι $(L + \Delta L)(U + \Delta U)\underline{x} = \underline{b} \Rightarrow$

$$(LU + L\Delta U + \Delta L U + \Delta L \Delta U)\underline{x} = \underline{b} \Rightarrow$$

$$(A + \Delta A)\underline{x} = \underline{b} \text{ όπου } \Delta A = E + L\Delta U + \Delta L U + \Delta L \Delta U \Rightarrow$$

$$\|\Delta A\|_\infty \leq \|E\|_\infty + \|L\|_\infty \|\Delta U\|_\infty + \|\Delta L\|_\infty \|U\|_\infty + \|\Delta L\ |_\infty \|\Delta U\|_\infty$$

Γνωρίζουμε ότι

$\|E\|_\infty \leq n^2\rho\|A\|_\infty u$
$\|U\|_\infty \leq n\rho\|A\|_\infty$, $\|L\|_\infty \leq n$, $\|\Delta U\| \leq \frac{n(n+1)+2}{2}u_1$, $\|\Delta U\|_\infty \leq \frac{n^2+n+2}{2}u_1\rho\|A\|_\infty$

Τελικά

$$\|\Delta A\|_\infty \leq n^2\rho\|A\|_\infty u + n\frac{n^2+n+2}{2}u_1\rho\|A\|_\infty +$$

$$\frac{n^2+n+2}{2}u_1 n\rho\|A\|_\infty + \frac{1}{4}(n^2+n+2)(n^2+n+2)u_1^2\rho\|A\|_\infty$$

$$\leq u_1\rho\|A\|_\infty(n^2+n(n^2+n+2)+\frac{1}{4}(n^2+n+2)^2u_1^2)$$

Εάν $n \cdot u_1 << 1$ οι όροι της τελευταίας παρένθεσης αγνοούνται οπότε:

$$\|\Delta A\|_\infty \leq (n^3+2n^2+2n)u_1 \cdot \rho \cdot \|A\|_\infty$$

Από το φράγμα αυτό παρατηρούμε ότι εάν το growth είναι μεγάλο μπορεί να αναμένεται πρόβλημα στη λύση του προβλήματος. □

**Παρατήρηση:** Η επίλυση του συστήματος $Ax = \underline{b}$ μπορεί να γίνει επιλύοντας το άνω τριγωνικό σύστημα που προκύπτει από την επεξεργασία του πίνακα $A$ και του διανύσματος $\underline{b}$ ταυτόχρονα. Συγκεκριμένα μπορούμε να τριγωνοποιήσουμε τον επαυξημένο πίνακα $(A, \underline{b})$ και στη συνέχεια να επιλύσουμε το άνω τριγωνικό σύστημα με προς τα πίσω αντικατάσταση.

**Παράδειγμα 6.1.3** *Να επιλυθεί το σύστημα $A\underline{x} = \underline{b}$ χρησιμοποιώντας απαλοιφή Gauss με μερική οδήγηση με χρήση του επαυξημένου πίνακα.*

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 2 & 2 & 3 \\ 4 & 1 & 1 \end{pmatrix},\ \underline{b} = \begin{pmatrix} 2 \\ 6 \\ 3 \end{pmatrix}$$

**Απόδειξη:** $k = 1$ Το οδηγό στοιχείο είναι $a_{31} = 4,\ r_1 = 3$

Αλλάζουμε τις γραμμές 3 και 1 του $A$ και τις αντιστοιχίες εισόδους του $\underline{\beta}$.

$$A = \begin{pmatrix} 4 & 1 & 1 \\ 2 & 2 & 3 \\ 0 & 1 & 1 \end{pmatrix},\ \underline{b} = \begin{pmatrix} 3 \\ 6 \\ 2 \end{pmatrix}$$

$$m_{21} = -\frac{a_{21}}{a_{11}} = -\frac{1}{2}$$

$$A = A^{(1)} = \begin{pmatrix} 4 & 1 & 1 \\ 0 & \frac{3}{2} & \frac{5}{2} \\ 0 & 1 & 1 \end{pmatrix},\ \underline{b} = \underline{b}^{(1)} = \begin{pmatrix} 3 \\ \frac{9}{2} \\ 2 \end{pmatrix}$$

$k = 2$: Το οδηγό στοιχείο είναι $a_{22} = \frac{3}{2}$

$$m_{32} = -\frac{a_{32}}{a_{22}} = -\frac{2}{3}$$

$$A = A^{(2)} = \begin{pmatrix} 4 & 1 & 1 \\ 0 & \frac{3}{2} & \frac{5}{2} \\ 0 & 0 & -\frac{2}{3} \end{pmatrix}, \ \underline{b} = \underline{b}^{(2)} = \begin{pmatrix} 3 \\ \frac{9}{2} \\ -1 \end{pmatrix}$$

Το προκύπτον τριγωνικό σύστημα $A^{(2)}\underline{x} = \underline{b}^{(2)}$ είναι

$$4x_1 + x_2 + x_3 = 3$$
$$\frac{3}{2}x_2 + \frac{5}{2}x_3 = \frac{9}{2}$$
$$-\frac{2}{3}x_3 = -1$$

Η λύση είναι:

$$x_3 = \frac{3}{2}, \ x_2 = \frac{1}{2}, \ x_1 = \frac{1}{4}$$

□

## $QR$ παραγοντοποίηση

Βήμα 1: Προσδιορίζουμε την $QR$ παραγοντοποίηση του $A$
$Q^T \cdot A = R$
Βήμα 2: Διαμορφώνουμε το διάνυσμα
$\underline{b}' = Q^T \cdot \underline{b}$
Βήμα 3: Επιλύουμε το σύστημα
$R\underline{x} = \underline{b}'$

Για τον υπολογισμό του διανύσματος $\underline{b}'$ δε χρειάζεται ο αναλυτικός υπολογισμός του πίνακα $Q$. Συγκεκριμένα αφού $Q^T = H_{n-1}H_{n-2}\dots H_2H_1$ το $\underline{b}'$ υπολογίζεται ως εξής:

**For** $k = 1, 2, \dots, n-1$
$\quad \underline{b} = H_k \cdot \underline{b}$
$\underline{b}' = \underline{b}$

**Παράδειγμα**: Να επιλυθεί με $QR$ το σύστημα $A\underline{x} = \underline{b}$ όπου

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 1 & 1 \end{pmatrix}, \ \underline{b} = \begin{pmatrix} 2 \\ 6 \\ 3 \end{pmatrix}$$

Από προηγούμενο παράδειγμα έχουμε υπολογίσει ότι

$$R = \begin{pmatrix} -1.4142 & -2.1213 & -2.8284 \\ 0 & 1.2247 & 1.6330 \\ 0 & 0 & -0.5774 \end{pmatrix}$$

$$H_1 = \begin{pmatrix} 0 & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{\sqrt{2}} & -\frac{1}{2} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 0 & -0.7071 & -0.7071 \\ -0.7071 & 0.5000 & -0.5000 \\ -0.7071 & -0.5000 & 0.5000 \end{pmatrix}$$

$$H_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -0.1691 & -0.9856 \\ 0 & -0.9856 & -0.1691 \end{pmatrix}$$

Υπολογισμός του $\underline{b'}$

$$\underline{b} = \begin{pmatrix} 2 \\ 6 \\ 3 \end{pmatrix}$$

$$\underline{b} = H_1\underline{b} = \begin{pmatrix} -6.3640 \\ 0.0858 \\ -2.9142 \end{pmatrix}$$

$$\underline{b'} = \underline{b} = H_2\underline{b} = \begin{pmatrix} -6.3640 \\ 2.8560 \\ -0.5774 \end{pmatrix}$$

Παρατηρούμε ότι το $\underline{b'}$ υπολογίστηκε χωρίς να διαμορφώσουμε τον πίνακα $Q$.
Επιλύουμε το σύστημα:

$$R\underline{x} = \underline{b'}$$

$$\underline{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

**Απαιτούμενα flops:** Με τη χρήση της μεθόδου Householder για τον υπολογισμό της $QR$ παραγοντοποίησης χρειάζονται $\frac{2}{3}n^3 + O(n^2)$ flops για την επίλυση του συστήματος $A\underline{x} = \underline{b}$.

**Ευστάθεια:** Έχει δειχθεί ότι η υπολογιζόμενη λύση $\underline{x}$ που βρίσκεται χρησιμοποιώντας τη μέθοδο Householder $QR$ είναι η ακριβής λύση του συστήματος

$$(A+E)\underline{x} = \underline{b} + d\underline{b}$$

όπου

$$\|E\|_F \leq (3n^2+41n)u\|A\|_F + O(u^2)$$
$$\|d\underline{b}\| \leq (3n^2+40n)u\|\underline{b}\| + O(u^2)$$

Παρατηρούμε ότι δεν υπάρχει παράγοντας growth στα υπεισερχόμενα σφάλματα.

### 6.1.3 Επίλυση Γραμμικού Συστήματος με πολλαπλό Δεξιό μέλος

Έστω το σύστημα

$$A \cdot X = B$$
$$\text{όπου } A\epsilon\mathbb{R}^{n\times n},\ X\epsilon\mathbb{R}^{n\times m},\ B\epsilon\mathbb{R}^{n\times m},\ m \leq n$$

Η επίλυση τέτοιων συστημάτων προκύπτει σε πολλά προβλήματα εφαρμογών.
Ο παρακάτω αλγόριθμος μπορεί να χρησιμοποιηθεί για τον προσδιορισμό της λύσης εάν χρησιμοποιηθεί Gauss με μερική οδήγηση για την παραγοντοποίηση του $A$.

Βήμα 1: Παραγοντοποιείστε τον $A$
$M \cdot A = U$
Βήμα 2: Να επιλυθούν τα $m$ άνω τριγωνικά συστήματα
$U\underline{x}_i = \underline{b}'_i = M\underline{b}_i,\ i = 1,\ldots,m$

**Παράδειγμα 6.1.4** *Να επιλυθεί το σύστημα $AX = B$, όπου*

$$A = \begin{pmatrix} 1 & 2 & 4 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix},\ B = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}$$

**Απόδειξη:** Από προηγούμενο παράδειγμα γνωρίζουμε ότι

$$U = \begin{pmatrix} 7 & 8 & 9 \\ 0 & \frac{6}{7} & \frac{19}{7} \\ 0 & 0 & -\frac{1}{2} \end{pmatrix},\ M = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & -\frac{1}{7} \\ -\frac{1}{2} & 1 & -\frac{1}{2} \end{pmatrix}$$

Επιλύουμε τα συστήματα:

$$U\underline{x}_1 = M\underline{b}_1 = \begin{pmatrix} 5 \\ 0.2857 \\ 0 \end{pmatrix}$$

$$\underline{x}_1 = \begin{pmatrix} 0.3333 \\ 0.3333 \\ 0 \end{pmatrix}$$

$$U\underline{x}_2 = M\underline{b}_2 = \begin{pmatrix} 6 \\ 1.1429 \\ 0 \end{pmatrix}, \quad \underline{x}_2 = \begin{pmatrix} -0.6667 \\ 1.3333 \\ 0 \end{pmatrix}$$

Τελικά $X = [\underline{x}_1, \underline{x}_2]$. □

**Εφαρμογή:** Εάν $B = I$, η προηγούμενη μέθοδος υπολογίζει τον αντίστροφο πίνακα.

## 6.2 Ανάλυση Ευαισθησίας Γραμμικού Συστήματος

Για τη μελέτη ευαισθησίας του προβλήματος του γραμμικού συστήματος $A\underline{x} = \underline{b}$ ο δείκτης κατάστασης (condition number) είναι

$$cond(A) = \|A\|\|A^{-1}\|.$$

Έτσι ο υπολογισμός του $cond(A)$ περιλαμβάνει υπολογισμό αντιστρόφου, που όπως είδαμε είναι πολύ πιο επίπονος από τη λύση εξ αρχής του συστήματος. Υπάρχουν όμως μέθοδοι εκτίμησης του $cond(A)$ που αποφεύγουν τον αναλυτικό υπολογισμό του $A^{-1}$. Ο πίνακας

$$H = \begin{pmatrix} 1 & 1/2 & 1/3 & \dots & 1/n \\ 1/2 & 1/3 & 1/4 & \dots & 1/(n+1) \\ \vdots & \vdots & \vdots & & \vdots \\ 1/n & 1/(n+1) & \dots & 1/(2n-1) & \end{pmatrix}$$

ονομάζεται πίνακας Hilbert. Η επίλυση του προβλήματος $H\underline{x} = \underline{b}$ είναι ill-conditioned για οποιοδήποτε $n$. Έστω $n = 5$,

$$b = (2.2833,\ 1.4500,\ 1.0929,\ 0.8845,\ 0.7456).$$

Η ακριβής λύση είναι

$$x = (1,\ 1,\ 1,\ 1,\ 1)^T.$$

Εάν διαταράξουμε την (5,1) είσοδο του $H$ στην πέμπτη θέση ώστε να πάρουμε 0.20001 (από $1/5 = 0.2$) η λύση που προκύπτει από αυτή την πολύ μικρή διατάραξη είναι

$$\underline{x} = (0.9937,\ 1.1252,\ 0.4365,\ 1.8765,\ 0.5618)^T.$$

Παρατηρούμε ότι $cond(H) = O(10^5)$.

Θα μελετήσουμε τις επιδράσεις που επιφέρουν μικρές μεταβολές στα δεδομένα του γραμμικού συστήματος $A\underline{x} = \underline{b}$.

**Θεώρημα 6.2.1** *(Θεώρημα δεξιάς διαταραχής)*
*Εάν $\Delta\underline{x}$ και $\Delta\underline{b}$ είναι διαταραχές στο $\underline{x}$ και $\underline{b}$ αντίστοιχα του γραμμικού συστήματος $A\underline{x} = \underline{b}$, $A$ μη ιδιάζων, $\underline{b} \neq \underline{0}$ τότε*

$$\frac{\|\Delta\underline{x}\|}{\|\underline{x}\|} \leq cond(A)\frac{\|\Delta\underline{b}\|}{\|\underline{b}\|}$$

**Απόδειξη:** Ισχύει ότι $A\underline{x} = \underline{b}$

$$A(\underline{x} + \Delta\underline{x}) = (\underline{b} + \Delta\underline{b}) \Rightarrow A\underline{x} + A\Delta\underline{x} = \underline{b} + \Delta\underline{b} \Rightarrow$$

$$A\Delta\underline{x} = \Delta\underline{b} \text{ αφού } A\underline{x} = \underline{b}$$

Έτσι $\Delta\underline{x} = A^{-1}\Delta\underline{b}$.

Παίρνοντας μία επαγόμενη νόρμα πινάκων έχουμε

$$\|\Delta\underline{x}\| \leq \|A^{-1}\| \cdot \|\Delta\underline{b}\| \tag{6.6}$$

Παίρνοντας και την ίδια νόρμα στο σύστημα $A\underline{x} = \underline{b}$ έχουμε

$$\|\underline{b}\| = \|A\underline{x}\| \leq \|A\| \cdot \|\underline{x}\| \tag{6.7}$$

Συνδυάζοντας τις (6.6) και (6.7) προκύπτει

$$\frac{\|\Delta\underline{x}\|}{\|\underline{x}\|} \leq \|A\| \cdot \|A^{-1}\|\frac{\|\Delta\underline{b}\|}{\|\underline{b}\|}$$

□

**Παρατήρηση** Η αλλαγή στη λύση $\underline{x}$ εξαρτάται από το πόσο θα επηρεάσει η ποσότητα $cond(A)$ το γινόμενο $cond(A)\frac{\|\Delta\underline{b}\|}{\|\underline{b}\|}$.
Έτσι εάν έχουμε μικρή μεταβολη στο $\underline{b}$ και το $cond(A)$ είναι μικρό παραμένει μικρή και η μεταβολή που υπεισέρχεται στο $\underline{x}$. Εάν όμως το $cond(A)$ είναι μεγάλο η μεταβολή στη λύση $\underline{x}$ μπορεί και αυτή να είναι μεγάλη.

**Παράδειγμα 6.2.1** *Εστω το ακόλουθο ill-conditioned πρόβλημα*

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4.0001 & 2.002 \\ 1 & 2.002 & 2.004 \end{pmatrix}, \quad \underline{b} = \begin{pmatrix} 4 \\ 8.0021 \\ 5.006 \end{pmatrix}$$

*Η ακριβής λύση είναι $\underline{x} = [1, 1, 1]^t$*

*Μεταβάλουμε το* $\underline{b}$ *σε* $\underline{b}' = [4, 8.0020, 5.0061]^t$
*Τότε η σχετική αλλαγή στο* $\underline{b}$ *είναι*

$$\frac{\|\underline{b}' - \underline{b}\|}{\|\underline{b}\|} = \frac{\|\Delta\underline{b}\|}{\|\underline{b}\|} = 1.379 * 10^{-5} \ \ (\mu\iota\kappa\rho\acute{\eta})$$

$cond(A) = 3.2222 \cdot 10^5$ *(πολύ μεγάλο)*
*Εάν τώρα επιλύσουμε το σύστημα* $A\underline{x}' = \underline{b}'$ *παίρνουμε* $\underline{x}' = \underline{x} + \Delta\underline{x} = [3.0850 \ -$ $0.0436 \ 1.0022]^t$.
*Το* $\underline{x}'$ *είναι τελείως διαφορετικό από το* $\underline{x}$.
*Παρατηρούμε ότι* $\frac{\|\Delta\underline{x}\|}{\|\underline{x}\|} = 1.3461 \leq cond(A)\frac{\|\Delta\underline{b}\|}{\|\underline{b}\|} = 4.4434$

**Παράδειγμα 6.2.2** *Εστω το ακόλουθο well-conditioned πρόβλημα*

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \ b = \begin{pmatrix} 3 \\ 7 \end{pmatrix}$$

*Η ακριβής λύση είναι* $\underline{x} = [1, 1]^t$.
*Εστω* $\underline{b}' = \underline{b} + \Delta\underline{b} = [3.0001 \ 7.0001]^t$
*Η σχετική αλλαγή στο* $\underline{b}$ *είναι:*

$$\frac{\|\underline{b}' - \underline{b}\|}{\|\underline{b}\|} = 1.875 * 10^{-5} \ \ (\mu\iota\kappa\rho\acute{\eta})$$

$$cond(A) = 14.9330 \quad (\mu\iota\kappa\rho\acute{o})$$

*Ετσι δεν αναμένεται δραστική αλλαγή στη λύση* $\underline{x}$.
*Πράγματι το* $\underline{x}'$ *ικανοποιεί το σύστημα* $A\underline{x}' = \underline{b}'$ *όπου* $\underline{x}' = [0.9999, 1.0001]^t$
*Παρατηρούμε ότι* $\frac{\|\Delta\underline{x}\|}{\|\underline{x}\|} = 10^{-5}$

**Θεώρημα 6.2.2** *(Θεώρημα αριστερής διαταραχής)*
*Εστω* $A$ *μη ιδιάζων και* $\underline{b} \neq \underline{0}$. *Εστω* $\Delta A$ *και* $\Delta\underline{x}$ *αντίστοιχα οι μεταβολές στον* $A$ *και στο* $\underline{x}$ *του γραμμικού συστήματος* $A\underline{x} = \underline{b}$.
*Επιπλέον υποθέτουμε ότι η* $\Delta A$ *είναι τέτοια ώστε*

$$\|\Delta A\| < \frac{1}{\|A^{-1}\|}$$

*Τότε*

$$\frac{\|\Delta\underline{x}\|}{\|\underline{x}\|} \leq \frac{cond(A)}{(1 - cond(A)\frac{\|\Delta A\|}{\|A\|})}\frac{\|\Delta A\|}{\|A\|}$$

**Απόδειξη:** Εχουμε ότι

$$(A+\Delta A)(\underline{x}+\Delta\underline{x})=\underline{b}\Rightarrow(A+\Delta A)\underline{x}+(A+\Delta A)\Delta\underline{x}=\underline{b}$$

Επειδή $A\underline{x}=\underline{b}$ από τη προηγούμενη σχέση προκύπτει

$$(A+\Delta A)\Delta\underline{x}=-\Delta A\underline{x}\Rightarrow\Delta\underline{x}=-A^{-1}\Delta A(\underline{x}+\Delta\underline{x})$$

Παίρνοντας νόρμες έχουμε:

$$\|\Delta\underline{x}\|\leq\|A^{-1}\|\cdot\|\Delta A\|\cdot(\|\underline{x}\|+\|\Delta\underline{x}\|)=$$

$$\frac{\|A^{-1}\|\cdot\|A\|}{\|A\|}\|\Delta A\|\cdot(\|\underline{x}\|+\|\Delta\underline{x}\|)$$

Έτσι

$$(1-\frac{\|A^{-1}\|\|A\|\|\Delta A\|}{\|A\|})\|\Delta\underline{x}\|\leq\frac{\|A\|\|A^{-1}\|\|\Delta A\|}{\|A\|}\|\underline{x}\|$$

Επειδή $\|A^{-1}\|\|\Delta A\|<1$ η έκφραση της παρένθεσης στο αριστερό μέλος είναι θετική κι'έτσι μπορούμε να διαιρέσουμε την ανισότητα μάυτήν χωρίς να αλλάξει. Διαιρούμε επίσης και με την ποσότητα $\|\underline{x}\|$ και τελικά προκύπτει

$$\frac{\|\Delta\underline{x}\|}{\|\underline{x}\|}\leq\frac{\frac{\|A\|\|A^{-1}\|\|\Delta A\|}{\|A\|}}{1-\frac{\|A\|\|A^{-1}\|\|\Delta A\|}{\|A\|}}=\frac{cond(A)}{(1-cond(A)\frac{\|\Delta A\|}{\|A\|})}\frac{\|\Delta A\|}{\|A\|}$$

□

**Παρατήρηση:** Η υπόθεση $\|\Delta A\|<\frac{1}{\|A^{-1}\|}$ είναι ρεαλιστική και εξασφαλίζει ότι ο παρανομαστής του κλάσματος του δεξιού μέλους είναι μικρότερος της μονάδας. Ετσι ακόμα και αν η ποσότητα $\frac{\|\Delta A\|}{\|A\|}$ είναι μικρή μπορεί να επέλθει δραστική αλλαγή εάν η ποσότητα $cond(A)$ είναι μεγάλη.

**Παράδειγμα 6.2.3** *Ας θεωρήσουμε και πάλι το παράδειγμα 1 κι ας μεταβάλουμε το $a_{23}$ σε 2.0021 κρατώντας το $\underline{b}$ σταθερό.*
*Τότε*

$$\Delta A=-10^{-4}\begin{pmatrix}0&0&0\\0&0&1\\0&0&0\end{pmatrix}\quad(\textit{μικρή})$$

*Εάν λύσουμε το σύστημα $(A+\Delta A)\underline{x}'=\underline{b}$ προκύπτει*

$$\underline{x}'=[3.0852\ -0.0437\ 1.0021]^t$$

$$\delta\underline{x}=\underline{x}'-\underline{x}=[2.0852\ -1.0437\ 0.0021]^t$$

*Ετσι $\|\delta\underline{x}\|=1.3463$ (αρκετά μεγάλο).*

**Θεώρημα 6.2.3** *(Θεώρημα Γενικής Διαταραχής)*
*Έστω $A$ μη ιδιάζων, $\underline{b} \neq \underline{0}$ και έστω $\Delta A$, $delta\underline{x}$ και $delta\underline{b}$ οι μεταβολές στον $A$, στο $\underline{x}$ και στο $\underline{b}$ του γραμμικού συστήματος $A\underline{x} = \underline{b}$. Επιπλέον υποθέτουμε ότι η $DeltaA$ είναι τέτοια ώστε $\|\Delta A\| < \frac{1}{\|A^{-1}\|}$. Τότε*

$$\frac{\|\delta\underline{x}\|}{\|\underline{x}\|} \leq \left(\frac{cond(A)}{1 - cond(A)\frac{\|\Delta A\|}{\|A\|}}\right)\left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\delta\underline{b}\|}{\|\underline{b}\|}\right)$$

**Απόδειξη:**

Αφαιρούμε τη σχέση $A\underline{x} = \underline{b}$ από τη σχέση

$$(A + \Delta A)(\underline{x} + \delta\underline{x}) = \underline{b} + \delta\underline{b}\ kai\ prokptei$$

$$(A + \Delta A)(\underline{x} + \delta\underline{x}) - A\underline{x} = \delta\underline{b}$$

Αυτό γράφεται και ως εξής:

$$(A + \Delta A)(\underline{x} + \delta\underline{x}) - (A + \Delta A)\underline{x} + (A + \Delta A)\underline{x} - A\underline{x} = \delta\underline{b} \Rightarrow$$

$$(A + \Delta A)(\delta\underline{x}) + \Delta A\underline{x} = \delta\underline{b} \Rightarrow$$

$$A(I - A^{-1}(-\Delta A))\delta\underline{x} = \delta\underline{b} - \Delta A\underline{x} \tag{6.8}$$

Εστω $A^{-1}(-\Delta A) = F$. Τότε από Θεώρημα ισχύει ότι εάν $\|F\| < 1, I - F$ είναι αντιστρέψιμος και

$$\|(I - F)^{-1}\| \leq \frac{1}{1 - \|F\|}$$

Ετσι η σχέση (5.8) γράφεται

$$\delta\underline{x} = (I - F)^{-1}A^{-1}(\delta\underline{b} - \Delta A\underline{x})$$

Από την οποία προκύπτει

$$\|\delta\underline{x}\| \leq \frac{\|A^{-1}\|}{(1 - \|F\|)}(\|\delta\underline{b}\| + \|\Delta A\|\|\underline{x}\|)$$

$$\Rightarrow \frac{\|\delta\underline{x}\|}{\|\underline{x}\|} \leq \frac{\|A^{-1}\|}{(1 - \|F\|)}\left(\frac{\|\delta\underline{b}\|}{\|\underline{x}\|} + \|\Delta A\|\right)$$

Επειδή

$$\frac{1}{\|\underline{x}\|} \leq \frac{\|A\|}{\|\underline{b}\|}$$

$$\frac{\|\delta\underline{x}\|}{\|\underline{x}\|} \leq \frac{\|A^{-1}\|}{(1 - \|F\|)}\left(\frac{\|\delta\underline{b}\|\|A\|}{\|\underline{b}\|} + \|\Delta A\|\right)$$

προκύπτει

$$\frac{\|\delta\underline{x}\|}{\|\underline{x}\|} \leq \frac{\|A^{-1}\|\|A\|}{(1 - \|F\|)}\left(\frac{\|\delta\underline{b}\|}{\|\underline{b}\|} + \frac{\|\Delta A\|}{\|A\|}\right) \tag{6.9}$$

Επίσης

$$\|F\| = \|A^{-1}(-\Delta A)\| \leq \|A^{-1}\|\|\Delta A\| = \frac{\|A^{-1}\|\|A\|}{\|A\|}\|\Delta A\| \tag{6.10}$$

$$\|F\| < 1 \tag{6.11}$$

Συνδυάζοντας τις (5.9),(5.10) και (5.11) προκύπτει το εξής:

$$\frac{\|\delta\underline{x}\|}{\|\underline{x}\|} \leq (\frac{\|A^{-1}\|\|A\|}{(1-\frac{\|A^{-1}\|\|A\|}{\|A\|})\|\Delta A\|})(\frac{\|\delta\underline{b}\|}{\|\underline{b}\|} + \frac{\|\Delta A\|}{\|A\|})$$

$$= \frac{cond(A)}{(1-cond(A)\frac{\|\Delta A\|}{\|A\|})}(\frac{\|\delta\underline{b}\|}{\|\underline{b}\|} + \frac{\|\Delta A\|}{\|A\|})$$

□

**Παρατήρηση:** Ακόμα και αν οι μεταβλητές $\frac{\|\delta\underline{b}\|}{\|\underline{b}\|}$ και $\frac{\|\Delta A\|}{\|A\|}$ είναι μικρές μπορεί να υπάρξει δραστική αλλαγή στη λύση εάν το $cond(A)$ είναι μεγάλο. Έτσι το $cond(A)$ παίζει πολύ σημαντικό ρόλο στην ευαισθησία της λύσης.

## Επίδραση των σφαλμάτων στρογγύλευσης στην ανάγνωση δεδομένων

Σύμφωνα με το θεώρημα της αριστερής διαταραχής εάν ο πίνακας A του συστήματος διαταραχθεί κατά ένα μικρό πίνακα $E$ (ή $\delta A$) και η λύση $\underline{x}$ κατά μία ποσότητα $\underline{h}$ (ή $\delta\underline{x}$) (πάρουμε δηλαδή αντί του συστήματος $A\underline{x} = \underline{b}$ το ελαφρά διαταραγμένο $(A+E)(\underline{x}+\underline{h}) = \underline{b}$), τότε

$$\frac{\|\underline{h}\|}{\|\underline{x}\|} \leq \frac{\|A^{-1}\| \cdot \|E\|}{1 - \|A^{-1}\| \cdot \|E\|} \tag{6.12}$$

με τη προυπόθεση ότι $\|A^{-1}\| \cdot \|E\| < 1$

Χρησιμοποιώντας floating-point αριθμητική, ως γνωστόν

$$fl(A) = A + E,\ |\varepsilon_{ij}| \leq u|a_{ij}|$$

και

$$\|E\|_E \leq u\|A\|_E \tag{6.13}$$

$$\text{όπου } \|A\|_E = (\sum_i \sum_j |a_{ij}|^2)^{1/2})$$

Λόγω της (6.13) η (6.12) γίνεται

$$\frac{\|\underline{h}\|_2}{\|\underline{x}\|_2} \leq \frac{u\|A\|_E \cdot \|A^{-1}\|_E}{1 - u\|A\|_E \cdot \|A^{-1}\|_E} \tag{6.14}$$

Για να είναι το σχετικό σφάλμα $\frac{\|\underline{h}\|_2}{\|\underline{x}\|_2}$ μικρό θα πρέπει

$$u\|A\|_E \cdot \|A^{-1}\|_E \ll 1$$

Γνωρίζουμε ότι ισχύει η εξής σχέση:

$$\|A\|_2 \leq \|A\|_E \leq n^{1/2}\|A\|_2$$

οπότε η σχέση (6.14) γίνεται:

$$\frac{\|\underline{h}\|_2}{\|\underline{x}\|_2} \leq \frac{u \cdot n^{1/2}\|A\|_2 \cdot \|A^{-1}\|_2}{1 - u \cdot n^{1/2}\|A\|_2 \cdot \|A^{-1}\|_2} \tag{6.15}$$

Ομως $\|A\|_2\|A^{-1}\|_2 = k(A)$ είναι το δείκτη κατάστασης (condition number) του πίνακα $A$. Συνεπώς η σχέση (6.15) γίνεται:

$$\frac{\|\underline{h}\|_2}{\|\underline{x}\|_2} \leq \frac{u \cdot n^{1/2}k(A)}{1 - u \cdot n^{1/2}k(A)}$$

**Παρατήρηση:** Για να είναι το σχετικό σφάλμα στη λύση $\underline{x}$ μικρό, θα πρέπει $u \cdot n^{1/2}k(A) \ll 1$. Κατά συνέπεια, χρησιμοποιώντας floating-point αριθμητική δεν μπορούμε να βρούμε ούτε καν προσεγγιστική λύση σε ένα σύστημα $A\underline{x} = \underline{b}$, εάν ισχύει $k(A) \gg n^{-\frac{1}{2}}u^{-1}$.
Βλέπουμε εδώ πόσο σημαντικό ρόλο παίζει η εξέταση του condition number $k(A)$ ενός πίνακα πριν ξεκινήσουμε την εφαρμογή κάποιας αριθμητικής μεθόδου.

# Κεφάλαιο 7

# Ελάχιστα Τετράγωνα

## 7.1 Εισαγωγή

Σε προηγούμενο κεφάλαιο αναπτύξαμε διάφορες μεθόδους για την επίλυση του γραμμικού συστήματος

$$A\underline{x} = \underline{b}$$

όπου ο πίνακας υποθέσαμε ότι είναι τετραγωνικός και μη ιδιάζων.

Παρόλα αυτά στις περισσότερες εφαρμογές και κυρίως όσες προέρχονται από τη Στατιστική και την Επεξεργασία σήματος χρειάζεται η επίλυση ενός συστήματος όπου ο πίνακας Α δεν είναι τετραγωνικός και ενδεχόμενα ιδιάζων. Σάυτές τις περιπτώσεις μπορεί είτε να μην υπάρχουν λύσεις ή να υπάρχουν άπειρες το πλήθος. Για παράδειγμα, όταν ο $A$ είναι $m \times n$ και $m > n$ έχουμε ένα οερδετερμινεδ σύστημα το οποίο τυπικά δεν έχει λύση. Αντίθετα ένα υνδερδετερμινεδ σύστημα ($m < n$) τυπικά έχει άπειρο αριθμό από λύσεις.

Σάυτές τις περιπτώσεις το καλλίτερο που μπορεί να συμβεί είναι να βρούμε ένα διάνυσμα $\underline{\xi}$ το οποίο κάνει το $A\underline{x}$ να βρίσκεται όσο το δυνατόν πιο κοντά στο $\underline{\beta}.Mellalgiayqnoumenadinusma\underline{\xi}$ τέτοιο ώστε να ελαχιστοποιεί την ποσότητα $||r(x)|| = ||A\underline{x} - b||$

Όταν χρησιμοποιείται η Ευκλείδια νόρμα η λύση αυτή αναφέρεται σα λύση ελαχίστων τετραγώνων για το σύστημα $A\underline{x} = \underline{b}$. Ο όρος λύση ελαχίστων τετραγώνων προκύπτει από το γεγονός ότι η λύση αυτή ελαχιστοποιεί την Ευκλείδια νόρμα του διανύσματος του υπολοίπου, το τετράγωνο της οποίας από τον ορισμό της είναι ακριβώς το άθροισμα των τετραγώνων των συνιστωσών του διανύσματος. Το πρόβλημα του προσδιορισμού λύσεων ελαχίστων τετραγώνων για το γραμμικό σύστημα $A\underline{x} = \underline{b}$ είναι γνωστό σα το **Γραμμικό Πρόβλημα Ελαχίστων Τετραγώνων (ΓΠΕΤ) (Linear Least-squares problem LSP)**.

Το ΓΠΕΤ τυπικά ορίζεται ως εξής:

## 7.2 Πρόβλημα των γραμμικών ελαχίστων τετραγώνων

Εστω $A \in \mathbb{R}^{m\times n}$ ένας δοσμένος πίνακας και $\underline{b} \in \mathbb{R}^n$. Ζητείται να προσδιοριστεί διάνυσμα $\underline{\xi}$ έτσι ώστε η συνάρτηση $||r(x)|| = ||Ax - b||_2$ να ελαχιστοποιείται.

Εάν το πρόβλημα έχει περισσότερες από μία λύση αυτή που έχει την ελάχιστη Ευκλείδια νόρμα ονομάζεται λύση ελάχιστου μήκους ή λύση ελάχιστης νόρμας.

### Υπαρξη και μοναδικότητα των λύσεων

Οπως και στην περίπτωση επίλυσης ενός γραμμικού συστήματος, προκύπτουν οι ακόλουθες ερωτήσεις.
1. Υπάρχει πάντοτε μία λύση ελαχίστων τετραγώνων για το $A\underline{x} = \underline{b}$ ·
2. Είναι η λύση αυτή μοναδική·
3. Πως μπορούμε να προσδιορίσουμε τέτοιες λύσεις·

Υποθέτουμε ότι $A \in \mathbb{R}^{m\times n},\ m \geq n$ δηλαδή το σύστημα είναι οερδετερμινεδ ή τετραγωνικό.

**Λήμμα 7.2.1** *Το $\underline{x}$ είναι λύση ελαχίστων τετραγώνων του συστήματος $A\underline{x} = \underline{b}$ αν και μόνο αν το $\underline{x}$ ικανοποιεί το σύστημα*

$$A^T A\underline{x} = A^T\underline{b}$$

**Απόδειξη:** Εστω $r(\underline{x}) = b - A\underline{x}$ και $y \in \mathbb{R}^n$. Τότε $r(\underline{y}) = b - A\underline{y} = r(\underline{x}) + A\underline{x} - A\underline{y} = r(\underline{x}) + A(\underline{x} - \underline{y})$.
Έτσι

$$||r(\underline{y})||_2^2 = (r(\underline{x}) + A(\underline{x} - \underline{y}))^T(r(\underline{x}) + A(\underline{x} - \underline{y})) =$$
$$||r(\underline{x})||_2^2 + 2(\underline{x} - \underline{y})^T A^T r(\underline{x}) + ||A(\underline{x} - \underline{y})||_2^2$$

Αρχικά υποθέτουμε ότι το $\underline{x}$ ικανοποιεί το σύστημα

$$A^T A\underline{x} = A^T\underline{b}$$

έτσι $A^T r(\underline{x}) = 0$. Τότε όμως

$$||r(\underline{y})||_2^2 = ||r(\underline{x})||_2^2 + ||A(\underline{x} - \underline{y}||_2^2 \geq ||r(\underline{x})||_2^2$$

και επομένως η $\underline{x}$ είναι λύση ελαχίστων τετραγώνων.

Στη συνέχεια υποθέτουμε ότι $A^T r(\underline{x}) \neq 0$. Θέτουμε $A^T r(\underline{x}) = \underline{z} \neq \underline{0}$

Ορίζουμε ένα διάνυσμα $\underline{\psi}$ έτσι ώστε $\underline{y} = \underline{x} + c\underline{z}$. Τότε

$$r(\underline{y}) = r(\underline{x}) + A(\underline{x} - \underline{y}) = r(\underline{x}) - cA\underline{z}$$

$$||r(\underline{y})||_2^2 = (r(\underline{x}) - cA\underline{z})^T(r(\underline{x}) - cA\underline{z}) =$$

$$||r(\underline{x})||_2^2 + c^2||A\underline{z}||_2^2 - 2c\underline{z}^T A^T r(\underline{x}) =$$

$$||r(\underline{x})||_2^2 + c^2||A\underline{z}||_2^2 - 2c||\underline{z}||_2^2 < ||r(\underline{x}||_2^2$$

Αυτό ισχύει $\forall c > 0$ *en* $A\underline{z} = \underline{0}$ και για

$$0 < c < 2\frac{||\underline{z}||_2^2}{||\mathrm{A}\underline{z}||_2^2} \; en \; A\underline{z} \neq \underline{0}$$

Ετσι το $\underline{\xi}$ δεν είναι λύση ελαχίστων τετραγώνων. □

## Θεώρημα Ύπαρξης και Μοναδικότητας Ελαχίστων Τετραγώνων

Το ΓΠΕΤ πάντοτε έχει λύση. Η λύση αυτή είναι μοναδική. Εάν και μόνο αν $rank(A) = n$ (ο $A$ είναι full rank)

**Απόδειξη:** Από το προηγούμενο Λήμμα έχουμε ότι το $\underline{x}$ είναι λύση ελαχίστων τετραγώνων του $Ax = b$ εάν και μόνο αν το $\underline{x}$ ικανοποιεί το σύστημα

$$A^T A\underline{x} = A^T\underline{b}$$

Θα δείξουμε ότι η λύση αυτή είναι μοναδική εάν και μόνο αν $rank(A) = n$.

Η λύση του συστήματος αυτού είναι μοναδική αν και μόνο αν $A^T A$ είναι μη ιδιάζων. Όμως ισχύει το ακόλουθο αποτέλεσμα $A^T A$ είναι θετικά ορισμένος και επομένως μη ιδιάζων αν και μόνο αν $rank(A) = n$. Πράγματι εάν ο $m \times n$ πίνακας Α έχει ρανκ(Α)=ν $m \geq n$, τότε εάν $\underline{x} \neq \underline{0}$ θα προκύπτει ότι $\underline{y} = A\underline{x} \neq \underline{0}$. Έτσι $\underline{x}^T A^T A\underline{x} = \underline{y}^T\underline{y} > 0$ και κατά συνέπεια ο Α είναι θετικά ορισμένος. Από την άλλη πλευρά εάν ο Α δεν έχει ρανκ(Α)=ν, τότε για κάποιο $\underline{x} \neq \underline{0}$ θα έχουμε $A\underline{x} = 0$ και κατά συνέπεια $\underline{x}^T A^T A\underline{x} = \underline{0}$ οπότε ο Α δεν είναι θετικά ορισμένος. □

**Ορισμός:** Το σύστημα των εξισώσεων

$$A^T A\underline{x} = A^T\underline{b}$$

λέγονται κανονικές εξισώσεις (normal equations)

**Ορισμός:** Ο πίνακας $A^+ = (A^T A)^{-1}A^T$, όταν $A$ $m \times n$ $(m \geq n)$ και ρανκ(Α)=ν ονομάζεται **ψευδοαντίστροφος (pseudoinverse)** του $A$. Ο ψευδοαντίστροφος αναφέρεται επίσης σαν Moore-Penrose γενικευμένος αντίστροφος του $A$.

Από τα προηγούμενα φαίνεται ότι η μοναδική λύση ελαχίστων τετραγώνων για το οερδετερμινεδ σύστημα $A\underline{x} = \underline{b}$ δίνεται από:

$$\underline{x} = (A^T A)^{-1} A^T \underline{b} = A^+ \underline{b}$$

Ο ορισμός του γενικευμένου αντιστρόφου γενικεύει τον συνήθη ορισμό του αντιστρόφου ενός τετραγωνικού πίνακα. Όταν ο πίνακας $A$ είναι τετραγωνικός και αντιστρέψιμος τότε

$$A^+ = (A^T A)^{-1} A^T = A^{-1} (A^T)^{-1} A^T = A^{-1}$$

**Ορισμός:** Εστω $A$ $m \times n$ πίνακας και $rank(A) = n$, τότε

$$cond(A) = ||A|| \cdot ||A^+||$$

**Παράδειγμα:**

$$\begin{pmatrix} 1 & 2 \\ 2 & 3 \\ 4 & 5 \end{pmatrix}, \quad rank(A) = 2$$

Ετσι ο Α έχει φυλλ ρανκ

$$A^+ = (A^T A)^{-1} A^T = \begin{pmatrix} -1.2857 & -0.5714 & 0.8577 \\ 1 & 0.5000 & -0.5000 \end{pmatrix}$$

$$Cond_2(A) = ||A||_2 ||A||_2^+ = 7.6656 \cdot 2.0487 = 15.7047$$

□

## 7.3 Υπολογιστικές μέθοδοι για την επίλυση overdetermined προβλημάτων ελαχίστων τετραγώνων

### 7.3.1 Η μέθοδος των κανονικών εξισώσεων

Μία από τις πιο διαδεδομένους μεθόδους (ειδικά στη Στατιστική) για τον υπολογισμό της λύσης ελαχίστων τετραγώνων είναι η μέθοδος των κανονικών εξισώσεων. Η μέθοδος αυτή στηρίζεται στη λύση του συστήματος των κανονικών εξισώσεων.

$$A^T A\underline{x} = A^T \underline{b}$$

Εχουμε υποθέσει ότι $A$ είναι $m \times n$ $(m > n)$ και έχει full rank. Επειδή σάυτήν την περίπτωση ο πίνακας $A^T A$ είναι συμμετρικός και θετικά ορισμένος και επομένως μπορούμε να πάρουμε την ανάλυση Cholesky

$$A^T A = HH^T$$

## Αλγόριθμος Κανονικών Εξισώσεων

Εστω $A \quad m \times n, \quad m > n$, $rank(A) = n$. Ο παρακάτω αλγόριθμος υπολογίζει τη λύση $\underline{x}$ ελαχίστων τετραγώνων από τις κανονικές εξισώσεις χρησιμοποιώντας την παραγοντοποίηση Cholesky.

Βήμα 1: Διαμορφώστε $\underline{c} = A^T \underline{b}$
Βήμα 2: Υπολογίστε τον παράγοντα ἠολεσκψ $H$ του $A^T A$.
Βήμα 3: Επιλύστε τα τριγωνικά συστήματα
$Hy = \underline{x}$
$H^T \underline{x} = \underline{y}$

**Πολυπλοκότητα:** Για τον υπολογισμό των ποσοτήτων $A^T A$ και $A^T \underline{b}$ απαιτούνται $mn^2/2$ φλοπς, για τον υπολογισμό της ανάλυσης Cholesky απαιτούνται $n^3/6$ flops. Ετσι συνολικά έχουμε: $mn^2/2 + n^3/6$ flops οπότε η μέθοδος είναι αρκετά αποτελεσματική.

# Κεφάλαιο 8

# Πίνακες Hessenberg (Σχεδόν τριγωνικοί)

**Ορισμός:** Ενας τετραγωνικός πίνακας $A$ είναι άνω (κάτω) Hessenberg (upper Hessenberg) εάν $a_{ij} = 0$ για $i > j + 1$ ($a_{ij} = 0$ για $j > i + 1$)

$$\begin{pmatrix} * & * & \cdots & \cdots & 0 \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ * & \vdots & & & * \\ * & * & \cdots & \cdots & * \end{pmatrix} \qquad \begin{pmatrix} * & \cdots & \cdots & * & * \\ * & \cdots & \cdots & * & * \\ \vdots & \ddots & & \vdots & \vdots \\ & & \ddots & \vdots & \vdots \\ 0 & & & * & * \end{pmatrix}$$

### Κάτω Hessenberg Ανω Hessenberg

Ενας άνω (κάτω) Hessenberg πίνακας $A = (a_{ij})$ είναι unreduced εάν $a_{i,i-1} \neq 0,\ i = 2, 3, \ldots, n$ ($a_{i,i+1} \neq 0,\ i = 1, 2, \ldots, n-1$)

**Παράδειγμα 8.0.1**

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 3 & 4 \\ 1 & 1 & 1 \end{pmatrix} \text{ unreduced κάτω Hessenberg}$$

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 2 & 3 \end{pmatrix} \text{ unreduced άνω Hessenberg}$$

**Παρατήρηση:** Η αναγωγή ενός πίνακα στην άνω Hessenberg μορφή είναι πολύ σημαντική στους υπολογισμούς ιδιοτίμων. Επίσης η μορφή Hessenberg αποτελεί χρήσιμο εργαλείο σε πολλές άλλες εφαρμογές όπως στη Θεωρία Ελέγχου και στην επεξεργασία σήματος.

**Θεώρημα 8.0.1** *(Αναγωγή Hessenberg)*
*Εστω An × n πίνακας. Πάντοτε υπάρχει ορθογώνιος πίνακας P έτσι ώστε*

$$PAP^T = H_u$$

*όπου $H_u$ άνω Hessenberg πίνακας.*

**Απόδειξη:** Η απόδειξη θα γίνει κατασκευαστικά. Ο πίνακας $P$ θα κατασκευαστεί σαν το γινόμενο $(n-2)$ πινάκων Householder $P_1$ έως $P_{n-2}$. Ο $P_1$ κατασκευάζεται έτσι ώστε να εισάγει μηδενικά στην πρώτη στήλη του $A$ κάτω από την είσοδο (2,1). Ο $P_2$ κατασκευάζεται έτσι ώστε να δημιουργεί μηδενικά κάτω από την είσοδο (3,2) στη δεύτερη στήλη του πίνακα $P_1AP_1^T$ κ.ο.κ. Η όλη διαδικασία χρειάζεται $(n-2)$ βήματα. (Ας σημειώσουμε ότι ένας $n \times n$ Hessenberg πίνακας περιέχει τουλάχιστον $\frac{(n-2)(n-1)}{2}$ μηδενικά).
Συγκεκριμένα έχουμε τα εξής βήματα:

**Βήμα 1:** Προσδιορίστε ένα Householder πίνακα $\hat{P}_1$ τάξης $n-1$ έτσι ώστε

$$\hat{P}_1 \begin{pmatrix} a_{21} \\ a_{31} \\ \vdots \\ a_{n1} \end{pmatrix} = \begin{pmatrix} * \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Ορίζουμε τον πίνακα $P_1 = \begin{pmatrix} I_1 & O \\ O & \hat{P}_1 \end{pmatrix}$ και υπολογίζουμε τον πίνακα

$$A \equiv A^{(1)} = P_1AP_1^T = \begin{pmatrix} * & * & \cdots & \cdots & * \\ * & * & \cdots & \cdots & * \\ 0 & * & \cdots & \cdots & * \\ \vdots & \vdots & & & \vdots \\ 0 & * & \cdots & \cdots & * \end{pmatrix}$$

**Βήμα 2:** Προσδιορίστε ένα Householder πίνακα $\hat{P}_2$ τάξης $(n-2)$ έτσι ώστε

$$\hat{P}_2 \begin{pmatrix} a_{32} \\ \vdots \\ \vdots \\ a_{n2} \end{pmatrix} = \begin{pmatrix} * \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Ορίζουμε τον πίνακα $P_2 = \begin{pmatrix} I_2 & O \\ O & \hat{P}_2 \end{pmatrix}$ και υπολογίζουμε τον πίνακα

$$A \equiv A^{(2)} = P_2 A^{(1)} P_2^T = \begin{pmatrix} * & * & * & \cdots & \cdots & * \\ * & * & * & \cdots & \cdots & * \\ 0 & * & * & \cdots & \cdots & * \\ 0 & 0 & * & \cdots & \cdots & * \\ \vdots & \vdots & \vdots & & & \\ \vdots & \vdots & \vdots & & & \\ 0 & 0 & * & \cdots & \cdots & * \end{pmatrix}$$

Στο τέλος των $(n-2)$ βημάτων, ο πίνακας $A^{(n-2)}$ είναι έως άνω Hessenberg πίνακας $H_u$. Ο πίνακας $H_u$ είναι ορθογώνια όμοιος με τον $A$ όπως φαίνεται παρακάτω

$$H_k = A^{(n-2)} = P_{n-2} A^{(n-3)} P_{n-2}^T = P_{n-2}(P_{n-3} A^{(n-4)} P_{n-3}^T) P_{n-2}^T = \dots$$
$$= (P_{n-2} P_{n-3} \dots P_1) A (P_1^T P_2^T \dots P_{n-3}^T P_{n-2}^T)$$

Θέτουμε $P = P_{n-2} P_{n-3} ... P_1$ και προκύπτει

$$H_u = P A P^T$$

□

(Ο πίνακας $P$ είναι ορθογώνιος σα γινόμενο ορθογώνιων πινάκων).

**Παρατήρηση:** Λόγω της μορφής που έχει κάθε πίνακας $P_k$

$$P_k = \begin{pmatrix} I_k & O \\ O & \hat{P}_k \end{pmatrix}$$

ο εκ δεξιών πολ/σμός του $P_k A$ με $P_k^T$ δε χαλάει τη δομή των μηδενικών που ήδη έχει δημιουργήσει το γινόμενο $P_k A$.
Για παράδειγμα εάν $n = 4$ και $k = 1$, τότε

$$P_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{pmatrix}$$

$$P_1 A = \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{pmatrix}$$

και

$$P_1 A P_1^T = \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{pmatrix} = \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{pmatrix}$$

**Παράδειγμα 8.0.2**

Εστω

$$A = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 2 & 3 \\ 1 & 1 & 1 \end{pmatrix}$$

Επειδή $n = 3$ έχουμε να εκτελέσουμε ένα μόνο βήμα.
Διαμορφώνουμε τον $\hat{P}_1$ έτσι ώστε:

$$\hat{P}_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}$$

$$u_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \sqrt{2} e_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \sqrt{2} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1+\sqrt{2} \\ 1 \end{pmatrix}$$

$$\hat{P}_1 = I_2 - \frac{2 \mathrm{u}_2 \mathrm{u}_2^{\mathrm{T}}}{\mathrm{u}_2^{\mathrm{T}} \mathrm{u}_2} \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - 0.2929 \begin{pmatrix} 5.8284 & 2.4142 \\ 2.4142 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} -0.7071 & -0.7071 \\ -0.7071 & 0.7071 \end{pmatrix}$$

Διαμορφώνουμε τον $P_1$

$$P_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & & \\ 0 & & \hat{P}_1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -0.7071 & -0.7071 \\ 0 & -0.7071 & 0.7071 \end{pmatrix}$$

$$A \equiv A^{(1)} = P_1 A P_1^T = \begin{pmatrix} 0 & -2.1213 & 0.7071 \\ -1.4142 & 3.5000 & -0.5000 \\ 0 & 1.5000 & -0.5000 \end{pmatrix} = H_u$$

(Στους υπολογισμούς χρησιμοποιήσαμε αριθμητική τεσσάρων ψηφίων).

**Παρατηρήσεις**

1. Κάθε Householder πίνακας $P_k$ ορίζεται μονοσήμαντα από το διάνυσμα $u_{n-k}$ που καθορίζει τον πίνακα $\hat{P}_k$. Έτσι κάθε φορά θα αποθηκεύουμε το διάνυσμα $u_{n-k}$. Εάν χρειάζεται αναλυτικά ο πίνακας $P$ αυτός μπορεί να υπολογισθεί από τους πίνακες Householder $P_1$ έως $P_{n-2}$.

2. Το διάνυσμα $u_{n-k} = (u_{k+1,k}, ..., u_{n\ k})^T$ έχει $n-k$ συνιστώσες, ο δε αριθμός των μηδενικών που παράγει στο κ βήμα είναι (ν-κ-1).
Ετσι οι συνιστώσες 2 έως ν-κ του $u_{n-k}$ μπορούν να αποθηκευθούν στις θέσεις $(k+2, k)$ έως $(n, k)$ του $A$, η δε πρώτη συνιστώσα θα αποθηκεύεται ξεχωριστά.

3. Ο πίνακας $A^{(k)} = P_k A^{(k-1)} P_k^T$ μπορεί να υπολογισθεί χωρίς ποτέ να δημιουργούμε αναλυτικά τον πίνακα $P_k$. Συγκεκριμένα,

$$A^{(k)} = P_k A^{(k-1)} P_k^T = \begin{pmatrix} I_k & O \\ O & \hat{P}_k \end{pmatrix} \begin{pmatrix} B & & C \\ O & \underline{b} & D \end{pmatrix} \begin{pmatrix} I_k & O \\ O & \hat{P}_k{}^T \end{pmatrix}$$

όπου $B$ είναι $k \times k$ Hessenberg πίνακας, $\hat{P}_k$ είναι Householder πίνακας τάξης $(n-k)$ και $\underline{b}$ διάνυσμα. Το διάνυσμα $\hat{P}_k b$ είναι πολλαπλάσιο του $\underline{e}_1$. Ο πίνακας $\hat{P}_k D \hat{P}_k{}^T$ υπολογίζεται χωρίς να δημιουργήσουμε αναλυτικά τον πίνακα $P_k$.

4. Οι μη μηδενικές είσοδοι του $A^{(k)}$ μπορούν να αποθηκευθούν πάνω στον $A$. Ετσι στο τέλος του $(n-2)$ βήματος θα έχουμε:

$$A \equiv A^{(n-1)} = \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & \cdots & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & \cdots & \cdots & \cdots & a_{2n} \\ u_{31} & a_{32} & a_{33} & \cdots & \cdots & \cdots & a_{3n} \\ \vdots & u_{42} & & & & & \vdots \\ \vdots & \vdots & & & & & \vdots \\ u_{n1} & u_{n2} & \cdots & \cdots & u_{n\ n-2} & a_{n\ n-1} & a_{nn} \end{pmatrix}$$

Το μονοδιάστατο array $v$ που περιέχει τις πρώτες συντεταγμένες των διανυσμάτων $u_{n-k}$ δίνεται από:

$$v = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} u_{21} \\ u_{32} \\ u_{43} \\ \vdots \\ \vdots \\ u_{n-1\ n-2} \end{pmatrix}$$

Ετσι όλες οι πληροφορίες που χρειάζονται για τον υπολογισμό του $P$ αποθηκεύονται στο κάτω τριγωνικό μέρος του $A$ (κάτω από τη διαγώνιο) και στο μονοδιάστατο array $v$.

## 8.1 Αλγόριθμος Αναγωγή Householder - Hessenberg

Εστω ένας $n \times n$ πίνακας $A$, ο παρακάτω αλγόριθμος καθορίζει τα διανύσματα $u_{n-k} = (u_{k+1,k}, \ldots, u_{nk})^T,\ k = 1, \ldots, n-1$ που ορίζουν τους πίνακες Householder $\hat{P}_k,\ k = 1, 2, \ldots, n-2$ (και κατά συνέπεια και τους $P_k$) έτσι ώστε $P = P_{n-2}P_{n-3}\ldots P_2P_1,\ H_u = PAP^T$ είναι άνω Hessenberg πίνακας.
Οι συνιστώσες $u_{k+2,k}$ έως $u_{n,k}$ του διανύσματος $u_{n-k}$ αποθηκεύονται στις αντίστοιχες θέσεις $(k+2)$ έως $(n,k)$ του $A$. Οι πρώτες συντεταγμένες $u_{k+1,k}$ αποθηκεύονται στο ξεχωριστό array $v = (v_1, v_2, \ldots, v_n)^T$.
Οι μη μηδενικές είσοδοι του $A^{(k)} = P_kAP_k^T$ υπερκαλύπτουν αυτές τον $A$.

**For** $k = 1, 2, \ldots, n-2$
Βήμα 1: Καθορίστε το διάνυσμα $u_{n-k} = (u_{k+1,k}, \ldots, u_{nk})^T$ που ορίζει πίνακα Householder.
$\hat{P}_k = I_{n-k} - 2\frac{u_{n-k}u_{n-k}^T}{(u_{n-k}^T u_{n-k})}$
τάξης $(n-k)$ έτσι ώστε

$$\hat{P}_k \begin{pmatrix} a_{k+1,k} \\ \vdots \\ \vdots \\ a_{n,k} \end{pmatrix} = \begin{pmatrix} \sigma \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Βήμα 2: Υπερκάλυψη του $a_{k+1,k}$ από το $\sigma$
$a_{k+1,k} \equiv \sigma$
Βήμα 3: Αποθήκευση του διανύσματος $u_{n-k}$ ως εξής:
$v_k \equiv u_{k+1,k}$
$a_{k+i,k} \equiv u_{k+i,k},\ \ i = 2, \ldots, n-k$
Βήμα 4: Υπολογίστε την ποσότητα
$b = \frac{2}{u_{n-k}^T u_{n-k}}$
Βήμα 5: Ενημέρωση των εισόδων του $A$ στις στήλες $k+1$ έως και στις γραμμές $k+1$ έως $n$ πολλαπλασιάζοντας εξ' αριστερών
**For** $j = k+1, \ldots, n$
$s = b\sum_{i=k+1}^{n} u_{ik}a_{ij}$
$a_{ij} \equiv a_{ij} - su_{ik},\ \ i = k+1, \ldots, n$
Ενημέρωση των εισόδων του $A$ στις στήλες $k+1$ έως $n$ και στις γραμμές 1 έως $n$ πολλαπλασιάζοντας εκ δεξιών
**For** $i = 1, 2, \ldots, n$
$s = b\sum_{j=k+1}^{n} a_{ij}u_{jk}$
$a_{ij} = a_{ij} - su_{jk},\ \ j = k+1, \ldots, n$

**Υπολογιστική πολυπλοκότητα:** Ο αλγόριθμος χρειάζεται $\frac{5}{3}n^3$ *flops* για να υπολογίσει τον $H_u$. Εάν χρειάζεται ο αναλυτικός υπολογισμός του πίνακα $P$ θα

χρειασθούν επιπλέον $\frac{2}{3}n^3$ $flops$. Για μεγάλα $n$ δημιουργούνται σοβαρά αποθηκευτικά προβλήματα εάν χρειάζεται αναλυτικά ο πίνακας $P$.

**Ευστάθεια:** Εχει αποδειχθεί ότι ο πίνακας $H_u$ που υπολογίζεται ικανοποιεί τη σχέση

$$P^T H_u P = A + E, \quad \text{όπου} \quad ||E||_f \leq cn^2 u||A||_F$$

**Παράδειγμα 8.1.1**

Να μετασχηματισθεί ο πίνακας

$$A = \begin{pmatrix} 1 & 2 & 5 \\ 3 & 7 & 9 \\ 2 & 5 & 3 \end{pmatrix}$$

σε άνω Hessenberg μορφή και να προσδιοριστεί ο πίνακας $P$.

Απαιτείται ένα μόνο βήμα. Υπολογίζουμε το διάνυσμα $u_2 = (u_{21}, u_{31})^T$ έτσι ώστε

$$\hat{P}_1 \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \left( I - \frac{2\mathrm{u}_2\mathrm{u}_2^{\mathrm{T}}}{\mathrm{u}_2^{\mathrm{T}}\mathrm{u}_2} \right) \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} \sigma \\ 0 \end{pmatrix}$$

$$u_2 = (2.2019, 0.6667)^T, \quad \sigma = -3.6056$$

Ενημερώνουμε τις εισόδους του $A$ και προκύπτει

$$A \equiv H_u = \begin{pmatrix} 1 & -4.4376 & 3.0509 \\ -3.6056 & 12.2308 & -2.8462 \\ 0.6667 & 1.1538 & -2.2308 \end{pmatrix}$$

$$u = u_{21} = 2.2019$$

## 8.2 Τριδιαγώνια Αναγωγή

Εάν ο πίνακας $A$ είναι συμμετρικός τότε από τη σχέση $PAP^T = H_u$ προκύπτει ότι ο άνω Hessenberg πίνακας $H_u$ είναι επίσης συμμετρικός και κατά συνέπεια τριδιαγώνιος. Έτσι εάν ο παραπάνω αλγόριθμος εφαρμοσθεί σ'έναν συμμετρικό πίνακα $A$ ο προκύπτων πίνακας $H_u$ θα είναι ένας συμμετρικός τριδιαγώνιος πίνακας $T$. Λαμβάνοντας υπ' όψη τη συμμετρία απαιτείται λιγότερη μνήμη για αποθήκευση και μόνο $\frac{2}{3}n^3$ για τον υπολογισμό του $T$ αντί των $\frac{5}{3}n^3$ που χρειάζονται για τον υπολογισμό του $H_u$.

**Παράδειγμα 8.2.1**

Εστω

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

Αφού το $n = 3$, απαιτείται ένα μόνο βήμα. Διαμορφώνουμε τον $\hat{P}_1$:

$$\hat{P}_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}$$

$$u_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \sqrt{2}e_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \sqrt{2} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1+\sqrt{2} \\ 1 \end{pmatrix}$$

$$\hat{P}_1 = I_2 - \frac{2\mathrm{u}_2\mathrm{u}_2^{\mathrm{T}}}{\mathrm{u}_2^{\mathrm{T}}\mathrm{u}_2} \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - .2929 \begin{pmatrix} 5.8284 & 2.4142 \\ 2.4142 & 1.0000 \end{pmatrix}$$

$$= \begin{pmatrix} -0.7071 & -0.7071 \\ -0.7071 & 0.7071 \end{pmatrix}$$

Διαμορφώνουμε τον $P_1$

$$P_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & & \\ 0 & \hat{P}_1 & \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -0.7071 & -0.7071 \\ 0 & -0.7071 & 0.7071 \end{pmatrix}$$

$$Etsi, \qquad H_u = P_1 A P_1^T = \begin{pmatrix} 0 & -1.4142 & 0 \\ -1.4142 & 2.5000 & 0.5000 \\ 0 & 0.5000 & 0.5000 \end{pmatrix}$$

(ο $H_u$ είναι συμμετρικός τριδιαγώνιος)

# Κεφάλαιο 9

# Αριθμητικός Υπολογισμός Ιδιοτιμών

Η κεντρική ιδέα έγκειται στο μετασχηματισμό του πίνακα $A$ μέσω μετασχηματισμών ομοιότητας καλής κατάστασης (well conditioned) σε μία κατάλληλη κανονική μορφή. Ένας τέλειος πίνακας καλής κατάστασης είναι ένας ορθογώνιος πίνακας ο αριθμός κατάστασης (condition number) του οποίου (ως προς τη 2 νόρμα ή την $F$) είναι 1. Πράγματι εάν είναι πίνακας $A$ μετασχηματίζεται σ' έναν πίνακα $B$ με χρήση ορθογώνιων μετασχηματισμών ομοιότητας, τότε μία μεταβολή στον $A$ επιφέρει μεταβολή ίδιου μεγέθους στο $B$, συγκεκριμένα εάν

$$B = U^T A U,$$

$U$ ορθογώνιος και

$$B + \Delta B = U^T (A + \Delta A) U$$

τότε

$$||\Delta B||_2 = ||\Delta A||_2$$

**Παράδειγμα:**

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \\ 6 & 7 & 8 \end{pmatrix}, \quad U = \begin{pmatrix} -0.5774 & -0.5774 & -0.5774 \\ -0.5774 & 0.7887 & -0.2113 \\ -0.5774 & -0.2113 & 0.7887 \end{pmatrix}$$

$$B = \begin{pmatrix} 13 & -0.6340 & -2.3660 \\ -0.9019 & 0 & 0 \\ -6.0981 & 0 & 0 \end{pmatrix}, \quad \Delta A = 10^{-5} \cdot I_3$$

$$A_1 = A + \Delta A = \begin{pmatrix} 1.00001 & 2 & 3 \\ 3 & 4.00001 & 5 \\ 6 & 7 & 8.00001 \end{pmatrix}$$

$$B_1 = U^*(A + \Delta A)U = \begin{pmatrix} 13.00001 & -0.633974 & -2.3660 \\ -0.9019 & 0.00001 & 0 \\ -6.0981 & 0 & 0.00001 \end{pmatrix}$$

$$\Delta B = B_1 - B = 10^{-5} \cdot I_3, \quad ||\Delta A|| = ||\Delta B|| = 10^{-5}$$

Η καλλίτερη μορφή πίνακα που εμφανίζει τις ιδιοτιμές είναι μία τριγωνική μορφή (οι διαγώνιοι είσοδοι είναι οι ιδιοτιμές).
Στη συνέχεια αναφέρουμε ένα αποτέλεσμα που αφορά τη δημιουργία ενός σχεδόν τριγωνικού πίνακα $R$, γνωστού σαν την πραγματική κανονική μορφή (real schur form, RSF) ενός πίνακα $A$.

**Θεώρημα 9.0.1** *(Η πραγματική τριγωνοποίηση του Schur)*
*Εστω $A$ ένας $n \times n$ πραγματικός πίνακας. Τότε υπάρχει ένας $n \times n$ ορθογώνιος πίνακας $Q$ έτσι ώστε*

$$Q^T A Q = R = \begin{pmatrix} R_{11} & R_{12} & \cdots & R_{1k} \\ \vdots & R_{22} & & R_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & R_{kk} \end{pmatrix}$$

*όπου κάθε $R_{ii}$ είναι είτε βαθμωτό είτε $2 \times 2$ πίνακας. Οι βαθμωτές ορθογώνιες είσοδοι αντιστοιχούν σε πραγματικές ιδιοτιμές ενώ οι $2 \times 2$ πίνακες στις διαγώνιες εισόδους αντιστοιχούν σε συζυγείς μιγαδικές ιδιοτιμές.*

βφ Ορισμός: Ο πίνακας $R$ του παραπάνω θεωρήματος είναι γνωστός σαν η Schur πραγματική μορφή (RSF) του $A$.

**Παρατηρήσεις:**
(ι) Οι $2 \times 2$ πίνακες στη διαγώνιο του $R$ αναφέρονται σα 'bumps'.
(ιι) Η RSF δεν είναι μοναδική.

Στη συνέχεια παρουσιάζουμε τη μέθοδο της $QR$ επανάληψης για τον υπολογισμό της RSF ενός πίνακα $A$. Η μέθοδος $QR$ εφαρμοζόμενη κατάλληλα χρησιμοποιείται σήμερα ευρέως για τον υπολογισμό των ιδιοτιμών ενός αυθαίρετου πίνακα $A$. Όπως λέει και το όνομά της η μέθοδος στηρίζεται στην παραγοντοποίηση $QR$ και η φύση της είναι επαναληπτική. Επειδή οι ιδιοτιμές ενός πίνακα $A$ είναι οι ρίζες του χαρακτηριστικού πολυωνύμου και είναι γνωστό ότι οι ρίζες ενός αυθαίρετου πολυωνύμου βαθμού μεγαλύτερου από 4 δε μπορούν να προσδιορισθούν σ'ένα πεπερασμένο αριθμό βημάτων συνεπάγεται ότι κάθε αριθμητική μέθοδος υπολογισμού ιδιοτιμών θα έχει φύση επαναληπτική.

### 9.0.1 Η βασική $QR$ επανάληψη

Αρχικά παρουσιάζουμε τη βασική $QR$ επαναληπτική μέθοδο.

Θέτουμε $A_0 = A$. Στη συνέχεια χρησιμοποιώντας $QR$ παραγοντοποίηση δημιουργούμε την ακολουθία πινάκων $\{A_k\}$ που ορίζεται ως εξής:

$$A_0 = Q_0 R_0$$

$$A_1 = R_0 Q_0 = Q_1 R_1$$
$$A_2 = R_1 Q_1 = Q_2 R_2$$

$$\vdots$$

Γενικά παίρνουμε

$$A_k = Q_k R_k = R_{k-1} Q_{k-1}, \quad k = 1, 2, \ldots$$

Για τους πίνακες της ακολουθίας $\{A_k\}$ ισχύει η εξής σημαντική ιδιότητα: Κάθε πίνακας της ακολουθίας είναι ορθογώνια όμοιος με τον προηγούμενό του και κατά συνέπεια είναι ορθογώνια όμοιος με τον αρχικό πίνακα. Αυτό αποδεικνύεται εύκολα. Για παράδειγμα

$$A_1 = R_0 Q_0 = Q_0^T A_0 Q_0 \quad (\text{επειδή } Q_0^T A_0 = R_0)$$
$$A_2 = R_1 Q_1 = Q_1^T A_1 Q_1$$

Ετσι ο $A_1$ είναι ορθογώνια όμοιος με τον Α και ο $A_2$ είναι ορθογώνια όμοιος με τον $A_1$. Κατά συνέπεια και ο $A_2$ είναι ορθογώνια όμοιος με τον Α όπως φαίνεται παρακάτω:

$$A_2 = Q_1^T A Q_1 = Q_1^T (Q_0^T A_0 Q_0) Q_1 = (Q_0 Q_1)^T A_0 (Q_0 Q_1)$$

Επειδή κάθε πίνακας είναι ορθογώνια όμοιος με τον $A$ έχει τις ίδιες μάυτόν ιδιοτιμές κι έτσι εάν η ακολουθία πινάκων $\{A_k\}$ συγκλίνει σε μία τριγωνική ή σχεδόν τριγωνική μορφή μπορούμε άμεσα να προσδιορίσουμε τις ιδιοτιμές. Το παρακάτω θεώρημα εξασφαλίζει τις συνθήκες σύγκλισης της ακολουθίας $\{A_k\}$.

**Θεώρημα 9.0.2** *(Θεώρημα Σύγκλισης για τη βασική $QR$-επανάληψη)*
*Εστω οι ιδιοτιμές $\lambda_1, \lambda_2, \ldots, \lambda_n$ του $A$ να είναι τέτοιες ώστε $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$ και έστω ο πίνακας των αριστερών ιδιοδιανυσμάτων για τον οποίο ισχύει ότι οι κύριες υποορίζουσες είναι μη μηδενικές. Τότε η ακολουθία $\{A_k\}$ συγκλίνει σε μία άνω τριγωνική μορφή γνωστή σαν πραγματική Schur μορφή (RSF).*

Στην πραγματικότητα μπορεί να αποδειχθεί ότι εάν ισχύουν οι προηγούμενες συνθήκες η πρώτη στήλη του $A_k$ προσεγγίζει ένα πολλαπλάσιο του $\underline{e}_1$. Έτσι για αρκετά μεγάλο $k$ παίρνουμε:

$$A_k = \begin{pmatrix} l_1 & u \\ 0 & \hat{A}_k \end{pmatrix}$$

Εφαρμόζουμε ξανά $QR$ επανάληξη στον $\hat{A}_k$ και η διαδικασία συνεχίζεται μέχρι τη σύγκλιση της ακολουθίας σε μία άνω τριγωνική μορφή.

**Παράδειγμα:** Εστω

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

με ιδιοτιμές $\lambda_1 = 5.3723$ και $\lambda_2 = -0.3723$ για τις οποίες ισχύει $|\lambda_1| > |\lambda_2|$.
Να εφαρμοσθεί $QR$-επανάληψη για τον προσδιορισμό αυτών των ιδιοτιμών $k = 0$

$$A_0 = A = Q_0 R_0$$

$$Q_0 = \begin{pmatrix} -0.3162 & -0.9487 \\ -0.9487 & 0.3162 \end{pmatrix}$$

$$R_0 = \begin{pmatrix} -3.1623 & -4.4272 \\ 0 & -0.6325 \end{pmatrix}$$

$k = 1$

$$A_1 = R_0 Q_0 = \begin{pmatrix} 5.2 & 1.6 \\ 0.6 & -0.2 \end{pmatrix} = Q_1 R_1$$

$$Q_1 = \begin{pmatrix} -0.9934 & -0.1146 \\ -0.1146 & -0.9934 \end{pmatrix}$$

$$R_1 = \begin{pmatrix} -5.2345 & -1.5665 \\ 0 & -0.3821 \end{pmatrix}$$

$k = 2$

$$A_2 = R_1 Q_1 = \begin{pmatrix} 5.3796 & -0.9562 \\ 0.0438 & -0.3796 \end{pmatrix} = Q_2 R_2$$

Ηδη έχει δημιουργηθεί πρόοδος στον προσδιορισμό των ιδιοτιμών.

$$Q_2 = \begin{pmatrix} -1 & -0.0082 \\ -0.0081 & 1 \end{pmatrix}$$

$$R_2 = \begin{pmatrix} -5.3797 & 0.9593 \\ 0 & -0.3718 \end{pmatrix}$$

$k = 3$

$$A_3 = R_2 Q_2 = \begin{pmatrix} 5.3718 & 1.0030 \\ 0.0030 & -0.3718 \end{pmatrix} = Q_3 R_3$$

$$Q_3 = \begin{pmatrix} 1 & -0.0006 \\ -0.0006 & 1 \end{pmatrix}$$

$$R_3 = \begin{pmatrix} -5.3718 & -1.0028 \\ 0 & -0.3723 \end{pmatrix}$$

$k = 4$

$$A_4 = R_3 Q_3 = \begin{pmatrix} 5.3723 & -0.9998 \\ 0.0002 & -0.3723 \end{pmatrix}$$

□

# Κεφάλαιο 10

# Ανάλυση Ιδιαζουσών Τιμών

Τα τελευταία χρόνια η ανάλυση ιδιαζουσών τιμών έχει γίνει απαραίτητο υπολογιστικό εργαλείο για την επίλυση ενός μεγάλου εύρους προβλημάτων που προέρχονται από πολλές πρακτικές εφαρμογές. Το κεντρικό σημείο για τη χρήση της ανάλυσης ιδιαζουσών τιμών (Singular Value Decomposition - SVD) έγκειται στο γεγονός ότι αυτές οι εφαρμογές απαιτούν γνώση της τάξης (rank) ενός πίνακα, προσεγγίσεων ενός πίνακα χρησιμοποιώντας πίνακες χαμηλότερης τάξης, ορθοκανονικές βάσεις για τους χώρους που παράγονται από τις γραμμές και τις στήλες του πίνακα καθώς επίσης και των ορθογώνιων συμπληρωμάτων τους και προβολών πάνω σ' αυτούς τους υπόχωρους. Στις περισσότερες περιπτώσεις αυτοί οι υπολογισμοί γίνονται με την παρουσία διαταράξεων στα δεδομένα (που ονομάζονται 'θόρυβοι') και όπως θα δούμε η SVD είναι πολύ αποτελεσματική γι' αυτούς τους υπολογισμούς.

## 10.1 Το Θεώρημα Ανάλυσης Ιδιαζουσών Τιμών

**Θεώρημα 10.1.1** *(SVD)*
*Εστω $A$ ένας πραγματικός $m \times n$ πίνακας. Τότε υπάρχουν ορθογώνιοι πίνακες $U$ και $V$ έτσι ώστε*

$$U^T A V = \begin{pmatrix} S_1 & 0 \\ 0 & 0 \end{pmatrix} = \Sigma$$

*όπου $\Sigma_1$ ένας μη ιδιάζων διαγώνιος πίνακας, με θετικά διαγώνια στοιχεία διατεταγμένα και αύξουσα σειρά. Ο αριθμός των μη μηδενικών διαγώνιων στοιχείων του $\Sigma$ ισούται με την τάξη του πίνακα $A$.*

Η ανάλυση $A = U\Sigma V^T$ είναι γνωστή σαν ανάλυση ιδιαζουσών τιμών (Singular Value Decomposition - SVD) του πίνακα $A$.

**Ορισμός:** Οι διαγώνιοι είσοδοι του πίνακα $\Sigma$ ονομάζονται **ιδιάζουσες τιμές** (singular values) του $A$. Οι αριθμοί $\sigma_1, \sigma_2, \ldots, \sigma_r$ είναι οι θετικές ιδιάζουσες τιμές του $A$.
Οι στήλες του $U$ ονομάζονται **αριστερά** ιδιάζοντα διανύσματα και αυτές του $V$ ονομάζονται **δεξιά** ιδιάζοντα διανύσματα.

## Μοναδικότητα της SVD

Υπάρχουν $k = min(m, n)$ ιδιάζουσες τιμές του $A$. Έστω $r$ η τάξη του $A$. Τότε υπάρχουν $r$ θετικές ιδιάζουσες τιμές. Αυτές είναι οι θετικές τετραγωνικές ρίζες των μη μηδενικών ιδιοτιμών του $A^TA$ (ή του $AA^T$). Οι υπόλοιπες $k - r$, εάν $r < k$, ιδιάζουσες τιμές είναι μηδέν.
Έτσι οι ιδιάζουσες τιμές είναι μοναδικές. Παρ' όλα αυτά τα ιδιάζοντα διανύσματα μπορεί να μην είναι μοναδικά. Π.χ. εάν ο $A$ έχει μία ιδιάζουσα τιμή $\sigma > 0$ με πολλαπλότητα, τότε οι αντίστοιχες στήλες του πίνακα $V$ μπορεί να επιλεγούν σαν οποιαδήποτε ορθοκανονική βάση του χώρου που παράγεται από τα ιδιοδιανύσματα που σχετίζονται με την πολλαπλή ιδιοτιμή $\lambda = \sigma^2$ του $A^TA$.

**Παραδοχή:** Από εδώ και στο εξής θα υποθέτουμε χωρίς βλάβη της γενικότητας ότι $m \geq n$ γιατί εάν $m < n$ θεωρούμε την SVD του $A^T$ και εάν η SVD του $A^T$ είναι $U\Sigma V^T$ τότε η SVD του $A$ είναι $V\Sigma U^T$.
Υποθέτουμε επίσης ότι οι ιδιάζουσες τιμές εμφανίζονται με τη σειρά $\sigma_1 = \sigma_{max} \geq \sigma_2 \geq \ldots \geq \sigma_n = \sigma_{min}$. Το σύμβολο $\sigma(A)$ συμβολίζει το σύνολο των ιδιαζουσών τιμών του $A$.

Το ακόλουθο θεώρημα συσχετίζει τις ιδιάζουσες τιμές και τις ιδιοτιμές ενός πίνακα.

**Θεώρημα 10.1.2** *Εστω $A = U\Sigma V^T$ η ανάλυση ιδιαζουσών τιμών ενός $m \times n$ πίνακα $A$ $(m \geq n)$. Εστω $r$ η τάξη του πίνακα $A$. Τότε*

1. $V^T(A^TA)V = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2, 0, \ldots, 0)_{n\times n}$
2. $U^T(AA^T)U = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2, 0, \ldots, 0)_{m\times m}$

**Απόδειξη:**

$A^TA = (U\Sigma V^T)^T(U\Sigma V^T) = V\Sigma^TU^TU\Sigma V^T = V\Sigma^T\Sigma V^T = V\Sigma' V^T$, όπου $\Sigma'$ είναι ένας $n \times n$ διαγώνιος πίνακας με $\sigma_1^2, \ldots, \sigma_r^2, 0, \ldots, 0$ σαν διαγώνια στοιχεία. Έτσι

$$V^TA^TAV = \Sigma' = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2, 0, \ldots, 0)_{n\times n}$$

Παρόμοια

$$U^TAA^TU = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2, 0, \ldots, )_{m\times m}$$

□

**Παρατήρηση:** Από το προηγούμενο θεώρημα παρατηρούμε ότι

1. Τα δεξιά ιδιάζοντα διανύσματα $v_1, v_2, \ldots, v_n$ είναι τα ιδιοδιανύσματα του πίνακα $A^T A$.
2. Τα αριστερά ιδιάζοντα διανύσματα $u_1, u_2, \ldots, u_m$ είναι τα ιδιοδιανύσματα του $AA^T$.
3. $\sigma_1^2, \ldots, \sigma_r^2$ είναι οι μη μηδενικές ιδιοτιμές των $A^T A$ και $AA^T$.

**Πόρισμα 10.1.1** *Εστω $A$ ένας συμμετρικός πίνακας με ιδιοτιμές $\lambda_1, \ldots, \lambda_n$. Τότε οι ιδιάζουσες τιμές του $A$ είναι $|\lambda_i|,\ i = 1, \ldots, n$.*

**Απόδειξη:** Επειδή $A$ είναι συμμετρικός ισχύει ότι $A = A^T$. Επομένως $A^T A = A^2$. Οι $n$ ιδιάζουσες τιμές του $A$ είναι οι μη αρνητικές τετραγωνικές ρίζες των $n$ ιδιοτιμών του $A^2$ που έχουν τιμή $\lambda_i^2,\ i = 1, \ldots, n$. □

**Πόρισμα 10.1.2** *Ενας $n \times n$ πίνακας $A$ είναι μη ιδιάζων αν και μόνο αν όλες οι ιδιάζουσες τιμές είναι διάφορες του μηδέν.*

**Απόδειξη:** Γνωρίζουμε ότι $det(A^T A) = (det(A))^2$. Επομένως ο $A$ είναι μη ιδιάζων αν και μόνο αν ο $A^T A$ είναι μη ιδιάζων. Το αποτέλεσμα ισχύει αφού ένας πίνακας είναι μη ιδιάζων αν και μόνο αν όλες οι ιδιοτιμές είναι διάφορες του μηδενός. □

## 10.2 Η SVD και η δομή ενός πίνακα

Η SVD μπορεί να χρησιμοποιηθεί αποτελεσματικά για τον υπολογισμό συγκεκριμένων σπουδαίων ιδιοτήτων που σχετίζονται με τη δομή ενός πίνακα όπως την τάξη, την Ευκλείδια νόρμα, τον αριθμό καράστασης και ορθοκανονικές βάσεις για το null space και το range του πίνακα.

**Θεώρημα 10.2.1** *Εστω $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$ οι $n$ ιδιάζουσες τιμές ενός $m \times n$ πίνακα $A$. Τότε*

1. $||A||_2 = \sigma_1 = \sigma_{max}$
2. $||A||_F = (\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2)^{1/2}$
3. $||A^{-1}||_2 = \frac{1}{\delta_n} = \frac{1}{\delta_{min}}$ , *όπου ο $A$ $n \times n$ μη ιδιάζων πίνακας.*
4. *Εάν $A$ είναι $n \times n$ και μη ιδιάζων τότε*
   $cond_2(A) = ||A||_2 ||A^{-1}||_2 = \frac{\sigma_1}{\sigma_n} = \frac{\sigma_{max}}{\sigma_{min}}$

**Απόδειξη:** 1.$||A||_2 = ||U\Sigma V^T||_2 = ||\Sigma||_2 = \sqrt{max \text{ ιδιοτιμή}(\Sigma^T\Sigma)} = max(\sigma_i)$. Η $||\cdot||_2$ είναι αναλλοίωτη σε ορθογώνια γινόμενα πινάκων.

2.$||A||_F = ||U\Sigma V^T||_F = ||\Sigma||_F = (s_1^2 + s_2^2 + \ldots + s_n^2)^{1/2}$ Επίσης η $||\cdot||_F$ είναι αναλλοίωτη σε ορθογώνιους μετασχηματισμούς

3. Εάν πάρουμε την SVD του $A^{-1}$ παρατηρούμε ότι η μεγαλύτερη ιδιάζουσα τιμή του $A^{-1}$ είναι $\frac{1}{\sigma_n}$

4. Από τον ορισμό του αριθμού κατάστασης και από το 1 και 3 προκύπτει ότι

$$cond_2(A) = \frac{\sigma_{max}}{\sigma_{min}}$$

□

**Παρατήρηση:** Για έναν $m \times n$ πίνακα $A$, $m \geq n$ όταν $rank(A) < n$ τότε $\sigma_{min} = 0$ και λέμε ότι το $cond(A)$ απειρίζεται

**Παράδειγμα 10.2.1**

Εστω $A = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \\ 6 & 7 & 7.999 \end{pmatrix}$

Οι ιδιάζουσες τιμές του $A$ είναι $\sigma_1 = 14.5570,\ \sigma_2 = 1.0375,\ \sigma_3 = 0.0001$

1. $||A||_2 = \sigma_1 = 14.5570$
2. $||A||_F = \sqrt{\sigma_1^2 + \sigma_2^2 + \sigma_3^2} = 14.5940$
3. $cond_2(A) = \frac{\sigma_1}{\sigma_3} = 1.0993 \times 10^5$

### 10.2.1 Προσδιορισμός ορθοκανονικών βάσεων

Εστω $A = U\Sigma V^T$ η SVD ενός $m \times n$ πίνακα. Έστω $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$ οι θετικές ιδιάζουσες τιμές του $A$. Από το βασικό θεώρημα της SVD έχουμε ότι

$$Av_i = \sigma_i u_i,\ i = 1, \ldots, r$$

$$Av_i = 0,\ i = r+1, \ldots, n$$

Παρόμοια εάν πάρουμε την SVD του $A^T = V\Sigma U^T$ προκύπτει:

$$A^T u_i = \sigma_i v_i,\ i = 1, \ldots, r$$

$$A^T u_i = 0$$

Επομένως

$$R(A) = span\{u_1, \ldots, u_r\}$$

$$N(A) = span\{v_{r+1}, \ldots, v_n\}$$
$$R(A^T) = span\{v_1, \ldots, v_r\}$$
$$N(A^T) = span\{u_{r+1}, \ldots, u_m\}$$

Ετσι προσδιορίζονται οι ακόλουθες βάσεις:

Ορθοκανονική βάση για το $R(A)$: Προσδιορίζεται από τις στήλες του $U$ που αντιστοιχούν σε μη μηδενικές ιδιάζουσες τιμές του $A$.
Ορθοκανονική βάση για το $N(A)$: Προσδιορίζεται από τις στήλες του $V$ που αντιστοιχούν σε μηδενικές ιδιάζουσες τιμές του $A$

Αντίστοιχα αποτελέσματα ισχύουν για τον $A^T$.

**Παράδειγμα 10.2.2**

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \\ 6 & 7 & 8 \end{pmatrix}$$

$$\sigma_1 = 14.5576,\ \sigma_2 = 1.03672,\ \sigma_3 = 0$$

$$U = \begin{pmatrix} 0.2500 & 0.8371 & 0.4867 \\ 0.4852 & 0.3267 & -0.8111 \\ 0.8378 & -0.4379 & 0.3244 \end{pmatrix}$$

$$V = \begin{pmatrix} 0.4625 & -0.7870 & 0.4082 \\ 0.5706 & -0.0882 & -0.8165 \\ 0.6786 & 0.6106 & 0.4082 \end{pmatrix}$$

Μια ορθοκανονική βάση για το null space του $A$ είναι

$$V_3 = \begin{pmatrix} 0.4082 \\ -0.8165 \\ 0.4082 \end{pmatrix}$$

Μια ορθοκανονική βάση για το ρανγε του $A$ είναι:

$$U_1 = \begin{pmatrix} 0.2500 & 0.8371 \\ 0.4852 & 0.3267 \\ 0.8370 & -0.4379 \end{pmatrix}$$

### 10.2.2 Numerical rank πίνακα

Η πιο αξιόπιστη μέθοδος αριθμητικού υπολογισμού της τάξης πίνακα στηρίζεται στη χρήση ιδιαζουσών τιμών. Έχοντας καθορίσει κάποια ακρίβεια $\varepsilon$ σύμφωνα με την οποία τιμές που είναι μικρότερες από αυτήν θεωρούνται μηδέν, μπορούμε να κάνουμε την εξής παραδοχή.

Ο πίνακας $A$ έχει 'numerical rank' $r$ εάν οι υπολογιζόμενες ιδιάζουσες τιμές

$$\sigma_1, \sigma_2, \ldots, \sigma_n$$

ικανοποιούν τη σχέση

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \varepsilon \geq \sigma_{r+1} \geq \ldots \geq \sigma_n$$

Συνήθως η ακρίβεια $\varepsilon$ μπορεί να έχει την τιμή $10^{-t}||A||_\infty$.
Ο αριθμός λοιπόν των ιδιαζουσών τιμών που είναι μεγαλύτερος από την καθορισμένη ακρίβεια $\varepsilon$, προσδιορίζει το numerical rank ενός πίνακα.

## 10.3 Υπολογισμός λύσης ελαχίστων τετραγώνων

Εστω ότι θέλουμε να προσδιορίσουμε διάνυσμα $\underline{x}$ έτσι ώστε να ελαχιστοποιείται η ποσότητα $||r||_2 = ||A\underline{x} - b||_2$, $Am \times n$ πίνακας, $m \geq n$.
Έστω $A = U\Sigma V^T$ η SVD του $A$. Τότε προκύπτει το εξής:

$$||r||_2 = ||U\Sigma V^T \underline{x} - \underline{b}||_2 = ||U(\Sigma V^T \underline{x} - U^T \underline{b}||_2 = ||\Sigma y - b'||_2$$

όπου έχουμε θέσει $V^T\underline{x} = \underline{y}$ και $U^T\underline{b} = \underline{b}'$.
Έτσι με τη χρήση των ιδιαζουσών τιμών το πρόβλημα ελαχίστων τετραγώνων παίρνει την ακόλουθη μορφή:

Να προσδιοριστεί $\underline{y}$ έτσι ώστε να ελαχιστοποιείται η ποσότητα:

$$||\Sigma\underline{y} - b'||_2$$

Εάν υποθέσουμε ότι $k$ είναι οι μη μηδενικές ιδιάζουσες τιμές τότε έχουμε:

$$||\Sigma\underline{y} - \underline{b}'||_2 = \sum_{i=1}^{k} |s_i y_i - b'_i|^2 + \sum_{i=k+1}^{m} |b'_i|^2$$

Ετσι το διάνυσμα $\underline{y} = [y_1, \ldots, y_n]^T$ θα δίνεται από τη σχέση

$$y_i = \begin{cases} \frac{b'_i}{\sigma_i} & , \text{ εάν } \sigma_i \neq 0 \\ \text{αυθαίρετο} & , \text{ εάν } \sigma_i = 0 \end{cases}$$

Εάν $k = n$ τότε η λύση ελαχίστων τετραγώνων είναι μοναδική.
Εάν $k < n$ (ο πίνακα $A$ είναι rank deficient), οι συντεταγμένες $y_{k+1}$ εώς $y_n$ δεν εμφανίζονται στο υπόλοιπο $||r||_2$ και επομένως δεν το επηρεάζουν καθόλου. Στην περίπτωση αυτή οι συντεταγμένες $y_{k+1}$ έως $y_n$ παίρνουν αυθαίρετες τιμές και έτσι προκύπτουν άπειρες το πλήθος λύσεις ελαχίστων τετραγώνων.
Υπάρχουν περιπτώσεις όπου αυτή η πλούσια οικογένεια λύσεων είναι πραγματικά επιθυμητή και μπορεί να χρησιμοποιηθεί για βελτιστοποίηση διαφόρων πλευρών του αρχικού προβλήματος.

Ο παρακάτω αλγόριθμος μπορεί να χρησιμοποιηθεί για τον υπολογισμό λύσεων ελαχίστων τετραγώνων με τη βοήθεια της SVD.

### 10.3.1 Αλγόριθμος SVD- ελάχιστα τετράγωνα

Βήμα 1: Να προσδιοριστεί η SVD του $A$
$A = U\Sigma V^T$
Βήμα 2: Δημιουργία του διανύσματος
$\underline{b}' = U^T\underline{b} = [b'_1, b'_2, \ldots, b'_m]^T$
Βήμα 3: Υπολογισμός λύσης $\underline{y}$ που προκύπτει από τις σχέσεις

$$y_i = \begin{cases} \frac{b'_i}{\sigma_i} & , \text{ εάν } \sigma_i \neq 0 \\ \text{αυθαίρετο} & , \text{ εάν } \sigma_i = 0 \end{cases}$$

Βήμα 4: Υπολογισμός της οικογένειας των λύσεων ελαχίστων τετραγώνων
$\underline{x} = V\underline{y}$

**Αποτελεσματικότητα:** Χρησιμοποιώντας τον αλγόριθμο Golub-Kahan-Reinsch για τον υπολογισμό της SVD ενός πίνακα $Am \times n, \quad m \geq n$ προκύπτει ότι ο παραπάνω αλγόριθμος απαιτεί $2mn^2 + 4n^3$ flops. Στον υπολογισμό του διανύσματος $\underline{b}'$ λαμβάνονται υπόψη μόνο οι συντεταγμένες του που σχετίζονται με τις στήλες του $U$ που αντιστοιχούν σε μη μηδενικές ιδιάζουσες τιμές.
Στην περίπτωση που ο πίνακας $A$ είναι rank deficient από την άπειρη οικογένεια των λύσεων μασ ενδιαφέρει εκείνη που δίνει την ελάχιστη ως προς τη 2-νόρμα λύση. Αυτή προκύπτει θέτοντας $y_i = 0$ οπότε $\sigma_i = 0$.
Ετσι προκύπτει η ακόλουθη έκφραση για την ελάχιστη ως προς τη 2-νόρμα λύση:

$$\underline{x} = \sum_{i=1}^{k} \frac{u_i^T b_i}{\sigma_i} v_i, \; k = rank(A) < n$$

**Παράδειγμα 10.3.1**

Εστω $A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 1 & 2 & 3 \end{pmatrix}$ , $\underline{b} = \begin{pmatrix} 6 \\ 9 \\ 6 \end{pmatrix}$

Βήμα 1: $\sigma_1 = 7.5358,\ \sigma_2 = 0.4957,\ \sigma_3 = 0$
Ο $A$ είναι rank deficient

$$U = \begin{pmatrix} 0.4956 & 0.5044 & 0.7071 \\ 0.7133 & -0.7008 & -0.0000 \\ 0.4956 & 0.5044 & -0.7071 \end{pmatrix}$$

$$V = \begin{pmatrix} 0.3208 & -0.8546 & 0.4082 \\ 0.5470 & -0.1847 & -0.8165 \\ 0.7732 & 0.4853 & 0.4082 \end{pmatrix}$$

Βήμα 2: $\underline{b}' = U^T \underline{b} = [12.3667, -0.2547, 0]^T$

Βήμα 3: $\underline{y} = [1.6411, -0.5541, 0]^T$

Βήμα 4: Η ελάχιστη ως προς τη 2-νόρμα λύση ελαχίστων τετραγώνων δίνεται από
$\underline{x} = V\underline{y} = [1, 1, 1]^T$