Μάθημα 9ο Ε3. Υπολογιστικά Μαθηματικά Ι 5/11/2018

2. Αριθμητική Γραμμική Άλγεβρα

Έστω $A \in \mathbb{R}^{n\times n}$ αντιστρέψιμος $A=(a_{ij})$, $1\le i,j\le n$, και $b\in\mathbb{R}^n$, $b=(b_1,\dots,b_n)^T$
Σκοπός μας είναι ο υπολογισμός ή η προσέγγιση της λύσης $x=(x_1,\dots,x_n)^T$ του γραμμικού συστήματος

$$Ax = b \quad (1)$$

Υπάρχουν δυο ειδών μέθοδοι:
- Ευθείες μέθοδοι: Η λύση υπολογίζεται ακριβώς (εφόσον οι πράξεις γίνονται ακριβώς. Διαφορετικά υπεισέρχονται σφάλματα στρογγύλευσης)
- Επαναληπτικές μέθοδοι: Η ακριβής λύση του συστήματος προσεγγίζεται από το επαναληπτικό σχήμα της μεθόδου. (οπότε υπάρχει το σφάλμα της μεθόδου, εκτός από τα πιθανά σφάλματα στρογγύλευσης)

2.1 Απαλοιφή Gauss

Έστω $A^{(1)}=A$ και $b^{(1)}=b$. και υποθέτουμε ότι $a_{11}^{(1)}\neq 0$ (αλλιώς με κατάλληλη εναλλαγή γραμμών φέρνουμε το (μη μηδενικό) στοιχείο με το μικρότερο δείκτη γραμμής της πρώτης στήλης στην (1,1) θέση)
Το στοιχείο $a_{11}^{(1)}\neq 0$ ονομάζεται οδηγός του πρώτου βήματος
Στη συνέχεια ορίζουμε τους πολλαπλασιαστές:

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

Πολλαπλασιάζουμε την πρώτη εξίσωση του (1) επί m_{i1} και αφαιρούμε από την i-οστή για $i=2,3,\dots,n$ Έτσι παίρνουμε το ισοδύναμο σύστημα

$$A^{(2)}x = b^{(2)} \quad \text{όπου}$$

$$a_{ij}^{(2)} = \begin{cases} a_{ij}^{(1)} & \text{αν } i=1,\ 1\le j\le n, \\ 0 & \text{αν } j=1,\ 2\le i\le n, \\ a_{ij}^{(1)} - m_{i1}a_{1j}^{(1)} & \text{αν } 2\le i,j\le n \end{cases} \qquad b_i^{(2)} = \begin{cases} b_1^{(1)} & \text{αν } i=1 \\ b_i^{(1)} - m_{i1}b_1^{(1)} & \text{αν } 2\le i\le n \end{cases}$$

Επαναλαμβάνουμε το ίδιο με τον $A^{(2)}$ μετατρέποντας σε μηδενικά τα στοιχεία της δεύτερης στήλης κάτω από την κύρια διαγώνιο
Συνεχίζοντας κατ' αυτόν τον τρόπο, στο k-οστό βήμα παίρνουμε το ισοδύναμο με το (1) σύστημα $A^{(k)}x = b^{(k)}$ (2) όπου

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

Έστω ότι $a_{kk}^{(k)} \neq 0$ (διαφορετικά, με κατάλληλη εναλλαγή γραμμών φέρνουμε στη θέση $a_{kk}^{(k)}$ το (μη μηδενικό) $a_{ik}^{(k)}$ με το μικρότερο δυνατό $i>k$ αποδεικνύεται ότι μπορούμε πάντοτε να το κάνουμε) Ορίζουμε τους πολλαπλασιαστές:

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

Πολλαπλασιάζουμε την κ-οστή εξίσωση του (2) επί m_{ik} και αφαιρούμε από την i-οστή για $i=k+1, k+2, \ldots, n$ οπότε παίρνουμε το ισοδύναμο σύστημα

$$A^{(k+1)} x = b^{(k+1)} \quad \text{όπου}$$

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

Κατ' αυτόν τον τρόπο, μετά από $n-1$ βήματα, καταλήγουμε στο σύστημα

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

ισοδύναμο με το αρχικό, όπου ο $A^{(n)}$ είναι ένας αντιστρέψιμος άνω τριγωνικός πίνακας με στοιχεία $a_{ij}^{(n)}$ είναι

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

ενώ για το διάνυσμα $b^{(n)}$ ισχύει $b_i^{(n)} = b_i^{(i)}$, $1 \leq i \leq n$

Στο σημείο αυτό τελειώνει η πρώτη φάση της απαλοιφής, γνωστή ως τριγωνοποίηση

$$A = \begin{pmatrix} * & * & \cdots & * \\ * & * & \cdots & * \\ \vdots & & & \vdots \\ * & * & \cdots & * \end{pmatrix} \longrightarrow A^{(n)} = \begin{pmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & * \end{pmatrix}$$

ΕΞ. Υπολογιστικά Μαθηματικά Ι — 5/11/2018

Η λύση του συστήματος υπολογίζεται στη δεύτερη φάση, τη λεγόμενη οπισθοδρόμηση

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

Παρατήρηση: Η τριγωνοποίηση μπορεί να εφαρμοστεί ακόμη και αν ο A δεν είναι αντιστρέψιμος. Στην περίπτωση αυτή στο (τυχαίο) k-οστό βήμα της διαδικασίας όλα τα a_{ik}, $k \le i \le n$, θα είναι μηδέν. Παραλείπουμε αυτό το βήμα και συνεχίζουμε όπως πριν. Βεβαίως, στο τέλος της διαδικασίας, ένα τουλάχιστον από τα διαγώνια στοιχεία του $A^{(n)}$ θα είναι μηδέν.

Η φάση της τριγωνοποίησης μπορεί να εκφραστεί και με τη βοήθεια της ανάλυσης LU

Θεώρημα: Για κάθε $A \in \mathbb{R}^{n \times n}$, όχι κατ' ανάγκην αντιστρέψιμος, υπάρχει ένας $n \times n$ πίνακας μεταθέσεως P, τέτοιος ώστε

$$PA = LU$$

όπου ο L είναι ένας $n \times n$, κάτω τριγωνικός πίνακας με μονάδες στην κύρια διαγώνιο, και U ένας $n \times n$ άνω τριγωνικός πίνακας

Απόδειξη:

Ακολουθούμε τον συμβολισμό της απαλοιφής για την οποία, κατ' αρχήν, υποθέτουμε ότι μπορεί να εκτελεστεί χωρίς εναλλαγές γραμμών, δηλαδή για κάθε k στον $A^{(k)}$ είτε $a_{kk}^{(k)} \neq 0$ είτε $a_{ik}^{(k)} = 0$ $k \le i \le n$. Στο πρώτο βήμα, θεωρούμε τον πίνακα

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

αν $a_{11}^{(1)} \neq 0$ και $M_1 = I$ αν $a_{i1}^{(1)} = 0$, $1 \le i \le n$. Είναι εύκολο να δούμε $A^{(2)} = M_1 A^{(1)}$ (ελέγξτε το)

Κατ' αυτόν τον τρόπο, στο k-οστό βήμα έχουμε $A^{(k+1)} = M_k A^{(k)}$ όπου

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

αν $\alpha_{kk}^{(k)} \neq 0$ και $M_k = I$ αν $\alpha_{ik}^{(k)} = 0$, $k \leq i \leq n$. Συνεπώς,

$A^{(n)} = M_{n-1} \cdot A^{(n-1)} = M_{n-1} \cdot M_{n-2} \cdot A^{(n-2)} = \dots = M_{n-1} M_{n-2} \dots M_1 A^{(1)} \Leftrightarrow$

$A^{(n)} = M_{n-1} M_{n-2} \dots M_1 \cdot A$

Οι πίνακες M_k $1 \leq k \leq n-1$, είναι αντιστρέψιμοι και μάλιστα

$$(M_k^{-1})_{ij} = \begin{cases} 1 & \text{αν } i=j \\ m_k & \text{αν } k+1 \leq i \leq n \\ 0 & \text{αλλοιώς} \end{cases}, \quad (\text{ελέγξτε το})$$

οπότε

$$A = M_1^{-1} \cdot M_2^{-1} \dots M_{n-1}^{-1} \cdot A^{(n)}$$

Αν θέσουμε $L = M_1^{-1} M_2^{-1} \dots M_{n-1}^{-1}$, $U = A^{(n)}$, τότε ο L είναι ένας κάτω τριγωνικός πίνακας, που δεν παραλείφθηκε κανένα βήμα έχει τη μορφή

$$L_{ij} = \begin{cases} 1, & \text{αν } i=j \\ m_{ij}, & \text{αν } i>j \\ 0, & \text{αν } i<j \end{cases} \quad (\text{ελέγξτε το})$$

Στη περίπτωση αυτή λοιπόν, που δεν γίνονται εναλλαγές, $P = I$

Μάθημα 10ο ΕΞ. Υπολογιστικά Μαθηματικά Ι 7/11/2018

Συνέχεια απόδειξης:

Τώρα, αν κάνουμε εναλλαγές γραμμών για να βρούμε οδηγούς $a_{kk}^{(k)} \neq 0$, τότε η γραμμή που εναλλάσσεται στο k-οστό βήμα με την προϋπάρχουσα k-οστή γραμμή από εκεί και πέρα μένει αναλλοίωτη. Συνεπώς, όλες οι εναλλαγές θα μπορούσαν να γίνουν εκ των προτέρων, οπότε η τριγωνοποίηση θα μπορούσε να γίνει πλέον χωρίς εναλλαγές γραμμών. Έτσι ο πίνακας που προκύπτει από τον A μετά τις μεταθέσεις, έστω A', γράφεται $A' = LU$

Επιπλέον, υπάρχει πίνακας μεταθέσεως P, που προκύπτει από τον μοναδιαίο αν επιφέρουμε στις γραμμές του τις αλλαγές που επιφέρουμε στον A για να πάρουμε τον A', έτσι ώστε:

$P \cdot A = A'$ (ελέγξτε το)

η οποία αποδεικνύει το θεώρημα.

Αν τώρα ο A είναι αντιστρέψιμος και έχουμε υπολογίσει τους πίνακες P, L, U έτσι ώστε $PA = LU$, τότε το σύστημα $Ax = b$ λύνεται ως εξής:

$$Ax = b \iff PAx = Pb \iff LUx = Pb \iff L(\underbrace{Ux}_{y}) = Pb \iff \begin{matrix} Ly = Pb \\ Ux = y \end{matrix}$$

Συνεπώς, υπολογίζουμε το διάνυσμα Pb και κατόπιν λύνουμε το σύστημα $Ly = Pb$, υπολογίζουμε το y_1 και κατόπιν τα $y_2, y_3, \dots, y_n$. Στη συνέχεια λύνουμε το σύστημα $Ux = y$ με οπισθοδρόμηση.

Παρατηρήσεις

① Στην πράξη η απαλοιφή του Gauss γίνεται σε δύο φάσεις:

(α) Βρίσκουμε τους πίνακες P, L, U έτσι ώστε $PA = LU$. Αυτό απαιτεί $\frac{n^3}{3} + O(n^2)$ πράξεις (δηλαδή $\frac{n^3}{3} + O(n^2)$ προσθέσεις και $\frac{n^3}{3} + O(n^2)$ πολλαπλασιασμοί/διαιρέσεις)

Τα στοιχεία αποθηκεύονται ως εξής:

$A =$ 

Έτσι δεν απαιτείται επιπλέον χώρος από αυτόν που καταλαμβάνει ο A. Οι πληροφορίες του P καταχωρούνται

συνήθως σε ένα διάνυσμα όπου η κ-οστή συντεταγμένη δείχνει την γραμμή που γίνεται οδηγός στο κ βήμα.

(β) Υπολογίζουμε την λύση λύνοντας τα τριγωνικά συστήματα $Ly = Pb$ και $Ux = y$. Αυτό απαιτεί $n^2 + O(n)$ πράξεις.

Με τον τρόπο αυτό μπορούμε να λύσουμε, με μικρό επιπλέον κόστος, πολλά συστήματα με τον ίδιο πίνακα A, όπως, για παράδειγμα, συμβαίνει κατά τον υπολογισμό του αντιστρόφου ενός πίνακα, οπότε αν $A^{-1} = (x^{(1)}, x^{(2)}, \dots, x^{(n)})$, τότε $Ax^{(i)} = e^{(i)}$, $i = 1, 2, \dots, n$ όπου $e^{(i)} = (0, \dots, 0, \overset{\leftarrow i\text{-θέση}}{1}, 0, \dots 0)^T$ (Αποδείξτε το)

Το επιπλέον κόστος είναι $n(n^2 + O(n)) = n^3 + O(n^2)$ πράξεις.

Το επιπλέον κόστος είναι απαγορευτικό, σε σχέση με την απαλοιφή Gauss, σε κάποιες περιπτώσεις, όπως, για παράδειγμα, στον υπολογισμό του διανύσματος $x = A^{-4}b$ για $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$ ($A^{-4} = A^{-1}A^{-1}A^{-1}A^{-1}$) δεδομένα. (Άσκηση: να βρω τον αποτελεσματικό τρόπο)

(2) Υπάρχουν κατηγορίες πινάκων για τις οποίες μπορούμε να αποδείξουμε εκ των προτέρων ότι η ανάλυση LU μπορεί να γίνει χωρίς εναλλαγές γραμμών.

2.2 Επιρροή των σφαλμάτων στρογγύλευσης στην απαλοιφή

Αν ο $A \in \mathbb{R}^{n \times n}$ είναι αντιστρέψιμος, ο αλγόριθμος της απαλοιφής του Gauss υπολογίζει την λύση του συστήματος $Ax = b$ ακριβώς, υπό την προϋπόθεση ότι οι πράξεις γίνονται ακριβώς. Στην πράξη στους υπολογισμούς υπεισέρχονται σφάλματα στρογγύλευσης. Τα σφάλματα στρογγύλευσης συσσωρεύονται και μπορεί τελικά η υπολογιστική λύση $\tilde{x}$ να απέχει πολύ από την θεωρητική x.

Η επιρροή των σφαλμάτων στρογγύλευσης στην απαλοιφή εξαρτάται από δύο παράγοντες:

(α) Η κατάσταση του συστήματος: Λέμε ότι ένα σύστημα έχει κακή κατάσταση αν μικρές διαταραχές στα δεδομένα επιφέρουν μεγάλες

διαταραχές στη λύση. Οι διαταραχές μπορεί να οφείλονται στη μη ακριβή παράσταση των στοιχείων a_{ij}, $1 \le i,j \le n$ και b_i, $1 \le i \le n$. στον υπολογιστή λόγω των σφαλμάτων στρογγύλευσης
Η σε υπολογιστική λύση $\tilde{x}$ μπορεί να θεωρηθεί ως η ακριβής λύση ενός παραπλήσιου προς το $Ax=b$ συστήματος. Η (καλή ή κακή) κατάσταση ενός συστήματος εξαρτάται (συσιαστικά μετριέται) από ένα μέγεθος γνωστό ως δείκτη κατάστασης του πίνακα A

(β) Την ευστάθεια ή την αστάθεια του αλγορίθμου της απαλοιφής:
Λέμε ότι ο αλγόριθμος της απαλοιφής είναι αστασθής αν μικρές διαταραχές στα δεδομένα, που προέρχονται από τα σφάλματα στρογγύλευσης, επιφέρουν μεγάλες διαταραχές στην λύση του συστήματος
Θα εξετάσουμε την ευστάθεια μιας παραλλαγής του αλγορίθμου της απαλοιφής, της απαλαιφής με μερική οδήγηση.
Οι δύο προηγούμενοι παράγοντες είναι ανεξάρτητοι μεταξύ τους, αλλά τελικά συμβάλλουν και οι δύο στη διαμόρφωση του σφάλματος στη $\tilde{x}$

Παράδειγμα: Να λυθεί με τη μέθοδο της απαλοιφής του Gauss με ολική οδήγηση (ο καλύτερος αλγόριθμος από άποψη ευστάθειας)
Το γραμμικό σύστημα
$0.913 x_1 + 0.659 x_2 = 0.254$
$0.780 x_1 + 0.563 x_2 = 0.217$
σε έναν υπολογιστή με $\beta = 10$, $t = 3$ $U = -L = 10$ και αποκοπή.

Μάθημα 11ο: Ε3. Υπολογιστικά Μαθηματικά Ι 12/11/2018

2.2.1 Κατάσταση του συστήματος - Δείκτης κατάστασης πίνακα

Παράδειγμα. Να λυθεί με τη μέθοδο απαλοιφής του Gauss με ολική οδήγηση (ο καλύτερος αλγόριθμος από άποψη ευστάθειας) το γραμμικό σύστημα:

$$0.913\,x_1 + 0.659\,x_2 = 0.254$$
$$0.780\,x_1 + 0.563\,x_2 = 0.217$$

σε έναν υπολογιστή με $\beta = 10$, $t = 3$, $U = -L = 10$ και ακοκοπή

Η ακριβής λύση του συστήματος είναι $x_1 = 1$, $x_2 = -1$ και η ορίζουσα του πίνακα του συστήματος είναι -10^{-6}.

$$0.913\,x_1 + 0.659\,x_2 = 0.254 \iff x_2 = 1$$
$$0.001\,x_2 = 0.001 \qquad x_1 = -0.443$$

Αλλάζοντας λίγο τα δεδομένα του δεύτερου μέρους σε $\begin{pmatrix} 0.253 \\ 0.218 \end{pmatrix}$ (αλλαγή κατά 10^{-3}) η ακριβής λύση γίνεται $\begin{pmatrix} x_1 = 1.223 \\ x_2 = -1.634 \end{pmatrix}$

Η αντίληψη ότι τα προβλήματα οφείλονται στη μικρή τιμή, -10^{-6}, της ορίζουσας (πίνακας "σχεδόν" μη αντιστρέψιμος) δεν ευσταθεί. Το σύστημα

$$0.913\,10^6 x_1 + 0.659\,10^6 x_2 = 0.254\,10^6$$
$$0.780\,x_1 + 0.563\,x_2 = 0.217$$

έχει ορίζουσα του πίνακα του συστήματος -1, ακριβής λύση $\begin{pmatrix} x_1 = 1 \\ x_2 = -1 \end{pmatrix}$ και στον προηγούμενο υπολογιστή βρίσκουμε $\begin{pmatrix} x_1 = -0.443 \\ x_2 = 1 \end{pmatrix}$

Θα μελετήσουμε την ευαισθησία της λύσης του συστήματος $Ax = b$, $A \in \mathbb{R}^{n\times n}$ αντιστρέψιμος και $b \in \mathbb{R}^n$, $b \neq 0$, σε διαταραχές των A και b

Έστω $\|\cdot\|$ του $\mathbb{R}^n$ και η επαγώμενη φυσική νόρμα στον $\mathbb{R}^{n\times n}$.

Έστω ότι ο πίνακας του συστήματος μεταβάλλεται σε $A + \delta A$, όπου $\|A^{-1}\|\,\|\delta A\| < 1$, το δεύτερο μέλος μεταβάλλεται σε $b + \delta b$, και η λύση του νέου συστήματος σε $x + \delta x$, $(A + \delta A)(x + \delta x) = b + \delta b$. (3)

Κατ' αρχήν, εφόσον $\|A^{-1}\|\,\|\delta A\| < 1$, από γνωστό πόρισμα (πρώτο μάθημα) ο πίνακας $A+\delta A$ είναι αντιστρέψιμος και επιπλέον

$$\|(A+\delta A)^{-1}\| \le \frac{\|A^{-1}\|}{1-\|A^{-1}\|\cdot\|\delta A\|} \quad (4)$$

Επομένως, το σύστημα (3) έχει ακριβώς μια λύση. Έχουμε

$$Ax + (\delta A)x + (A+\delta A)\delta x = b + \delta b \overset{Ax=b}{\iff} (A+\delta A)\delta x = -(\delta A)x + \delta b \overset{\exists (A+\delta A)^{-1}}{\iff}$$

$$\delta x = (A+\delta A)^{-1}(-(\delta A)x + \delta b) \iff \|\delta x\| \le \|(A+\delta A)^{-1}\|\,(\|\delta A\|\,\|x\| + \|\delta b\|)$$

Και λόγω της (4)

$$\frac{\|\delta x\|}{\|x\|} \le \frac{\|A^{-1}\|}{1-\|A^{-1}\|\,\|\delta A\|}\left(\|\delta A\| + \frac{\|\delta b\|}{\|x\|}\right) \quad (5)$$

Επίσης,

$$b = Ax \Rightarrow \|b\| \le \|A\|\,\|x\| \Rightarrow \frac{1}{\|A\|\,\|x\|} \le \frac{1}{\|b\|} \quad (6)$$

οπότε, πολλαπλασιάζοντας και διαιρώντας το δεύτερο μέλος της (5) με $\|A\|$ και χρησιμοποιώντας την (6) έχουμε

$$\frac{\|\delta x\|}{\|x\|} \le \frac{\|A\|\,\|A^{-1}\|}{1-\|A^{-1}\|\cdot\|\delta A\|}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\|\,\|x\|}\right)$$

$$\Rightarrow \frac{\|\delta x\|}{\|x\|} \le \frac{\|A\|\,\|A^{-1}\|}{1-\|A^{-1}\|\cdot\|\delta A\|}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right)$$

Συνεπώς, η ποσότητα $\|A\|\,\|A^{-1}\|$ προσδιορίζει πόσο μεγάλη μπορεί να γίνει η σχετική μεταβολή $\frac{\|\delta x\|}{\|x\|}$ της λύσης του συστήματος όταν η σχετική μεταβολή του πίνακα του συστήματος είναι $\frac{\|\delta A\|}{\|A\|}$ και του δεύτερου μέλους $\frac{\|\delta b\|}{\|b\|}$

Ο αριθμός αυτός συμβολίζεται με $K(A) = \|A\|\,\|A^{-1}\|$ και ονομάζεται δείκτης κατάστασης του A. Αν ο $K(A)$ είναι μικρός λέμε ότι ο A έχει καλή κατάσταση, ενώ αν ο $K(A)$ είναι μεγάλος λέμε ότι ο A έχει κακή κατάσταση. Σ' αυτή τη δεύτερη περίπτωση, μια μικρή μεταβολή στα δεδομένα (πίνακα ή/και στο δεύτερο μέλος) μπορεί να

Ε3. Υπολογιστικά Μαθηματικά Ι — 12/11/2018

προκαλέσει μεγάλη μεταβολή στη λύση. Αν η μεταβολή είναι μόνο στα στοιχεία του A (οπότε $\delta b = 0$) και $\|A^{-1}\|\,\|\delta A\| < 1$, η εκτίμηση παίρνει τη μορφή:

$$\frac{\|\delta x\|}{\|x\|} \le \frac{K(A)}{1 - \|A^{-1}\|\,\|\delta A\|} \cdot \frac{\|\delta A\|}{\|A\|} \qquad ((A+\delta A)(x+\delta x) = b)$$

Αν η μεταβολή είναι μόνο στα στοιχεία του δεύτερου μέλους b (οπότε $\delta A = 0$), η εκτίμηση παίρνει τη μορφή:

$$\frac{\|\delta x\|}{\|x\|} \le K(A)\,\frac{\|\delta b\|}{\|b\|} \qquad (A(x+\delta x) = b+\delta b)$$

Παρατηρήσεις:

(1) Η συγκριτικότητα δυο οποιονδήποτε νορμών πινάκων οδηγεί σε συγκριτικότητα των αντιστοίχων δεικτών κατάστασης. Για δυο οποιεσδήποτε νόρμες $\|\cdot\|_\alpha$ και $\|\cdot\|_\beta$ υπάρχουν σταθερές c_1 και c_2 τέτοιες ώστε $c_1 K_\alpha(A) \le K_\beta(A) \le c_2 K_\alpha(A)$. Επομένως ο δείκτης κατάστασης δεν εξαρτάται από τη νόρμα που επιλέγουμε (διαφορετικές νόρμες δίνουν τιμές διαφορετικές αλλά συγκριτικές). Έτσι (για οποιδήποτε) νόρμα $\|\cdot\|$ έχουμε $K(A) = \|A\|\,\|A^{-1}\| \ge \|A\cdot A^{-1}\| = \|I_n\| = 1 \Rightarrow K(A) \ge 1$ (το "=" είναι εφικτό). Στην πράξη, αν $1 \le K(A) \le 100$, λέμε ότι ο A έχει καλή κατάσταση, ενώ στην αντίθετη περίπτωση έχει κακή.

(2) Στο γραμμικό σύστημα
$$0.913\,x_1 + 0.659\,x_2 = 0.254$$
$$0.780\,x_1 + 0.563\,x_2 = 0.217$$
που είδαμε νωρίτερα στο πρώτο παράδειγμά μας, ο πίνακας $A = \begin{pmatrix} 0.913 & 0.659 \\ 0.780 & 0.563 \end{pmatrix}$ έχει δείκτη κατάστασης, ως προς τη $\|\cdot\|_1$, $K_1(A) \approx 2.7 \cdot 10^6$.

(3) Πολλοί γνωστοί για την κακή τους κατάσταση είναι οι πίνακες του Hilbert H_n, $n = 1, 2, \ldots$, που ορίζονται ως εξής:
$(H_n)_{ij} = (i+j-1)^{-1}$, $1 \le i, j \le n$
για παράδειγμα,

$$H_4 = \begin{pmatrix} 1 & 1/2 & 1/3 & 1/4 \\ 1/2 & 1/3 & 1/4 & 1/5 \\ 1/3 & 1/4 & 1/5 & 1/6 \\ 1/4 & 1/5 & 1/6 & 1/7 \end{pmatrix}$$

n	$K_2(H_n)$
2	$1.9 \cdot 10$
3	$5.2 \cdot 10^2$
4	$1.6 \cdot 10^4$
5	$4.8 \cdot 10^5$
⋮	
9	$4.9 \cdot 10^{11}$
10	$1.6 \cdot 10^{13}$

Μάθημα 12ο Ε3. Υπολογιστικά Μαθηματικά Ι 19/11/2018

④ Το μέγεθος του δείκτη κατάστασης δεν έχει σχέση με το μέγεθος της ορίζουσας του A, αν ο A είναι αντιστρέψιμος. Για παράδειγμα:

(α) Ο διαγώνιος 100×100 πίνακας D με στοιχεία διαγωνίου $d_{ii} = 10^{-1}$ έχει $\det D = 10^{-100}$ και $K(D) = 1$ (ως προς οποιαδήποτε νόρμα)

(β) Ο $n \times n$ άνω τριγωνικός πίνακας A με στοιχεία $a_{ii} = 1$, $1 \le i \le n$, $a_{ij} = -1$ $i < j$, έχει $\det A = 1$ αλλά $K_\infty(A) = n\, 2^{n-1}$

⑤ Αν $A = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix}$, $A^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1/\varepsilon \end{pmatrix}$, $0 < \varepsilon < 1$.

Τότε $K_1(A) = \|A\|_1 \|A^{-1}\|_1 = 1 \cdot 1/\varepsilon = 1/\varepsilon$

Καθώς $\varepsilon \to 0$, ο A τείνει να γίνει μη αντιστρέψιμος και $K_1(A) \to \infty$

Γενικά, για μια νόρμα $\| \ \|$, ισχύει ότι:

$$\frac{1}{K(A)} \le \inf \left\{ \frac{\|A-B\|}{\|A\|}, \ B \text{ μη αντιστρέψιμος} \right\}$$ για μια οποιαδήποτε νόρμα πίνακα.

Συνεπώς, ο $1/K(A)$ μετράει τη (σχετική ως προς το μέγεθος του A) απόσταση του A από το σύνολο των μη αντιστρέψιμων πινάκων.

Για φυσικές νόρμες πινάκων, μπορεί να αποδειχθεί ότι ισχύει η ισότητα με min στη θέση του inf

⑥ Έστω $x \neq 0$ η ακριβής λύση του συστήματος $Ax = b$, A αντιστρέψιμος και έστω $\tilde{x}$ μια προσεγγιστική λύση. Το υπόλοιπο της $\tilde{x}$ ορίζεται ως $r = A\tilde{x} - b$. Ισχύει

$$\frac{\|\tilde{x} - x\|}{\|x\|} \le K(A) \frac{\|r\|}{\|b\|} \quad \text{(Άσκηση)}$$

Σημαδιά, αν το υπόλοιπο r της προσεγγιστικής λύσης $\tilde{x}$ είναι μικρό αυτό δεν σημαίνει αναγκαστικά ότι το $\frac{\|\tilde{x} - x\|}{\|x\|}$ θα είναι μικρό αφού το $K(A)$ μπορεί να να είναι $\gg 1$

Συνεπώς, το υπόλοιπο, από μόνο του, δεν είναι ασφαλής δείκτης για την ακρίβεια της προσεγγιστικής λύσης

2.2.2 Ευστάθεια του αλγορίθμου της απαλοιφής

Παράδειγμα: Να λυθεί με τη μέθοδο της απαλοιφής του Gauss το γραμμικό σύστημα:

$$10^{-4}x_1 + x_2 = 1$$
$$x_1 + x_2 = 2$$

σε έναν υπολογιστή με $\beta=10$, $t=3$, $U=-L=10$ και στρογγύλευση.
Η ακριβής λύση του συστήματος είναι $x_1 = 1.0001$, $x_2 = 0.9999$
και η ορίζουσα του συστήματος είναι $10^{-4}-1$

$$0.1\,10^{-3}x_1 + 0.1\,10^{1}x_2 = 0.1\,10^{1}$$
$$0.1\,10^{1}x_1 + 0.1\,10^{1}x_2 = 0.2\,10^{1}$$

$\Longleftrightarrow$

$$0.1\,10^{-3}x_1 + 0.1\,10^{1}x_2 = 0.1\,10^{1} \quad \Longleftrightarrow \quad x_2 = 1 \quad \text{(καλή προσέγγιση)}$$
$$-0.1\,10^{5}x_2 = -0.1\,10^{5} \qquad x_1 = 0 \quad \text{(κακή προσέγγιση)}$$

$$m_{21} = \frac{0.1\,10^{1}}{0.1\,10^{-3}} = 0.1\cdot 10^{5} \quad \text{(αριθμός μηχανής)}$$

υπολογίζεται ακριβώς

$$a_{22}^{(2)} = 0.1\,10^{1} - 0.1\cdot 10^{5}\cdot 0.1\,10^{1} = -0.9999\,10^{4} \xrightarrow{\text{στρογγύλευση}} -0.1\,10^{5}$$
$$b_2^{(2)} = 0.2\cdot 10^{1} - 0.1\cdot 10^{5}\cdot 0.1\,10^{1} = -0.9998\,10^{4} \xrightarrow{\text{στρογγύλευση}} -0.1\,10^{5}$$

Το πρόβλημα δημιουργείται διότι, λόγω του πολύ μεγάλου μεγέθους του πολλαπλασιαστή, η επίδραση των παλαιών στοιχείων $a_{22}^{(1)}$, $b_2^{(1)}$ στον υπολογισμό των καινούργιων $a_{22}^{(2)}$, $b_2^{(2)}$ χάνεται

Το προηγούμενο παράδειγμα δείχνει ότι δεν είναι αρκετό, στο τυχαίο βήμα κ της μεθόδου απαλοιφής του Gauss να φέρουμε στη θέση του οδηγού $a_{kk}^{(k)}$ απλά ένα μη μηδενικό. Αν' αυτού φέρνουμε, με κατάλληλη εναλλαγή γραμμών, το στοιχείο $a_{kk}^{(k)}$ τέτοιο ώστε

$$|a_{kk}^{(k)}| = \max_{k\le i\le n} |a_{ik}^{(k)}|$$

Η στρατηγική αυτή ονομάζεται μερική οδήγηση. Αν στο προηγούμενο παράδειγμα, εφαρμόσαμε μερική οδήγηση, η προσεγγιστική λύση θα είναι πολύ κοντά στην ακριβή.

Ε3. Υπολογιστικά Μαθηματικά Ι

13/11/2018

Το κόστος των συγκρίσεων των στοιχείων στη μερική οδήγηση είναι $O(n^2)$, πολύ μικρότερο από το κόστος της απαλοιφής.
Θα αναλύσουμε την ευστάθεια του αλγορίθμου της απαλοιφής του Gauss με οδήγηση.
Η τεχνική που ακολουθούμε ονομάζεται αντίστροφη ανάλυση του σφάλματος (backward error analysis) και οφείλεται στον Άγγλο Μαθηματικό Wilkinson, ο οποίος απέδειξε ότι η $\tilde{x}$ είναι η ακριβής λύση του συστήματος $(A+\delta A)\tilde{x} = b$
Η ευστάθεια ή η αστάθεια του αλγορίθμου μετριέται από τη σχετική μεταβολή $\frac{\|\delta A\|}{\|A\|}$

Ο αλγόριθμος είναι ευσταθής αν αυτή η σχετική μεταβολή είναι μικρή

Μάθημα 13ο ΕΞ. Υπολογιστικά Μαθηματικά Ι 21/11/2018

Ξεκινάμε την ανάλυση μας υποθέτοντας (για την απλούστευση του προβλήματος):

(1) Τα στοιχεία των A, b είναι αριθμοί μηχανής

(2) Οι εναλλαγές γραμμών που υπαγορεύει η μερική οδήγηση έχουν γίνει εκ των προτέρων. Ασφαλώς, οι εναλλαγές αυτές δεν επηρεάζουν την αριθμητική

(3) Όλες οι πράξεις γίνονται μέσα στο εύρος των αριθμών μηχανής, οπότε δεν έχουμε over- ή underflow

Στη συνέχεια, για απλούστευση του συμβολισμού, αντί να συμβολίσαμε με $\tilde{L}, \tilde{U}, \tilde{m}_{ik}, \tilde{a}_{ij}^{(k)}$, κλπ τα μεγέθη που υπολογίζουμε με αριθμητική πεπερασμένης ακρίβειας, χρησιμοποιούμε το συνήθη συμβολισμό $L, U, m_{ik}, a_{ij}^{(k)}$ κλπ. Έτσι, στο k-οστό βήμα της απαλοιφής υπολογίζουμε τους πολλαπλασιαστές ως εξής:

$$m_{ik} = fl\left(\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}\right), \quad k+1 \le i \le n. \qquad (7)$$

και τα στοιχεία $a_{ij}^{(k+1)}$ του πίνακα $A^{(k)}$ βάσει των τύπων

$$a_{ij}^{(k+1)} = \begin{cases} 0 & , \text{αν } j=k,\ k+1 \le i \le n \\ fl\left(a_{ij}^{(k)} - m_{ik}\, a_{kj}^{(k)}\right) & , \text{αν } k+1 \le i,j \le n \\ a_{ij}^{(k)} & , \text{αλλοιώς} \end{cases} \qquad (8)$$

Στο τέλος της διαδικασίας ορίζουμε τους πίνακες

$$U = A^{(n)} \qquad (9)$$

$$L = (l_{ij})_{i,j=1}^{n}, \quad l_{ij} = \begin{cases} 0 & , \text{αν } i<j \\ 1 & , \text{αν } i=j \\ m_{ij} & , \text{αν } i>j \end{cases} \qquad (10)$$

Προφανώς, $A \neq LU$, αφού οι L, U έχουν υπολογισθεί με αριθμητική πεπερασμένης ακρίβειας. Το επόμενο θεώρημα δίνει την ακριβή σχέση μεταξύ LU και A.

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

$$LU = A + E$$

όπου $\|E\|_\infty \le n^2 \rho \|A\|_\infty u$, (11)

$$\rho = \frac{\max_{1 \le i,j,k \le n} |a_{ij}^{(k)}|}{\|A\|_\infty}, \quad (12)$$

και τα $a_{ij}^{(k)}$ ορίζονται από την (8)

Απόδειξη: Σημειώσεις Δουγαλή

Μετά την κατασκευή των L, U η προσεγγιστική λύση $\tilde{x}$ του συστήματος $Ax = b$ θα βρεθεί λύνοντας τα συστήματα $Ly = b$ και $U\tilde{x} = y$. Σφάλματα στρογγύλευσης θα υπεισέλθουν αφαλώς και κατά την επίλυση αυτών των συστημάτων. Έτσι, η y ικανοποιεί ακριβώς το σύστημα $(L + \delta L)y = b$ (αντί του $Ly = b$) και η $\tilde{x}$ το σύστημα $(U + \delta U)\tilde{x} = y$ (αντί του $U\tilde{x} = y$) με κατάλληλα δL, δU. Αυτό μας οδηγεί στο ακόλουθο λήμμα.

Λήμμα: Η προσεγγιστική λύση $\tilde{x}$ του συστήματος $Ax = b$ που παίρνουμε με απαλοιφή Gauss με μερική οδήγηση χρησιμοποιώντας αριθμητική περιορισμένης ακρίβειας με μοναδιαίο σφάλμα στρογγύλευσης u είναι η ακριβής λύση του συστήματος

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

όπου $\|\delta L\|_\infty \le 1.01 \frac{n(n+1)}{2} u$, (13)

$$\|\delta U\|_\infty \le 1.01 \frac{n(n+1)}{2} \rho \|A\|_\infty u, \quad (14)$$

Απόδειξη:

Σημειώσεις Δουγαλή

Το επόμενο θεώρημα περιγράφει την ευστάθεια της απαλοιφής του Gauss με μερική οδήγηση.

Ε3. Υπολογιστικά Μαθηματικά Ι — 21/11/2018

Θεώρημα: Η προσεγγιστική λύση $\tilde{x}$ που δίνει η απαλοιφή του Gauss με μερική οδήγηση είναι η ακριβής λύση του συστήματος

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

όπου $\|\delta A\|_\infty \le 1.01\,(n^3+3n^2)\,\rho\cdot\|A\|_\infty u$,

το ρ ορίζεται από την (12), u είναι το μοναδιαίο σφάλμα στρογγύλευσης και όπου υποθέτουμε ότι $n^2\cdot u < 1$

Απόδειξη:

Το $\tilde{x}$ ικανοποιεί την $(L+\delta L)(U+\delta U)\tilde{x} = b$ (λήμμα)

$\Leftrightarrow [LU + L(\delta U) + (\delta L)U + (\delta L)(\delta U)]\tilde{x} = b$

όμως από το προηγούμενο θεώρημα

$$LU = A + E$$

συνεπώς $A + E + L(\delta U) + (\delta L)U + (\delta L)(\delta U) = A + \delta A$

$\Leftrightarrow \delta A = E + L(\delta U) + (\delta L)U + (\delta L)(\delta U)$

$\Rightarrow \|\delta A\|_\infty \le \|E\|_\infty + \|L\|_\infty\|\delta U\|_\infty + \|\delta L\|_\infty\|U\|_\infty + \|\delta L\|_\infty\|\delta U\|_\infty$,

όπου τα $\|E\|_\infty$, $\|\delta L\|_\infty$ και $\|\delta U\|_\infty$ δίνονται από τις (11), (13) και (14) αντίστοιχα, ενώ λόγω των (7)-(10) και (12) έχουμε

$$\|U\|_\infty = \max_{1\le i\le n}\sum_{j=1}^{n}|u_{ij}| \le \max_{1\le i\le n}\sum_{j=1}^{n}\max_{1\le i,j,k\le n}|a_{ij}^{(k)}| = \max_{1\le i\le n}\sum_{j=1}^{n}\rho\|A\|_\infty = n\rho\|A\|_\infty$$

$$\|L\|_\infty = \max_{1\le i\le n}\sum_{j=1}^{n}|l_{ij}| \le \max_{1\le i\le n}\sum_{j=1}^{n}1 = n$$

Άρα, $\|\delta A\|_\infty \le n^2\rho\|A\|_\infty\cdot u + n\cdot 1{,}01\frac{n(n+1)}{2}\rho\|A\|_\infty u + 1{,}01\frac{n(n+1)}{2}u\cdot n\rho\|A\|_\infty +$

$+ 1{,}01\frac{n(n+1)}{2}u\cdot 1{,}01\frac{n(n+1)}{2}\rho\|A\|_\infty\cdot u$

$= n^2\cdot\rho\|A\|_\infty u + 1{,}01\cdot n^2(n+1)\rho\|A\|_\infty u + (1{,}01)^2\frac{n^2(n+1)^2}{4}\rho\|A\|_\infty\cdot u^2$

$\overset{n^2u<1}{<} n^2\rho\|A\|_\infty u + 1{,}01\,n^2(n+1)\rho\|A\|_\infty u + (1{,}01)^2\frac{(n+1)^2}{4}\rho\|A\|_\infty u$

$= \left[n^2 + 1{,}01\cdot n^2(n+1) + (1{,}01)^2\frac{(n+1)^2}{4}\right]\rho\|A\|_\infty u$

Ισχύει ότι: $n^2 + 1{,}01\,n^2(n+1) + (1{,}01)^2\frac{(n+1)^2}{4} \le 1{,}01\,(n^3+3n^2)$, $n>1$ (Άσκηση)

οπότε $\|\delta A\|_\infty \le 1{,}01\,(n^3+3n^2)\,\rho\,\|A\|_\infty\,u$

Ο παράγοντας n^3+3n^2 οφείλεται στην χρήση νορμών πινάκων και σε συνδρομικές εκτιμήσεις του φράγματος του $\|\delta A\|_\infty$ (μη πραγματικές)
Επομένως, το αν το $\frac{\|\delta A\|_\infty}{\|A\|_\infty}$ είναι μικρό εξαρτάται από το αν το ρ είναι μικρό

Για το ρ έχουμε, δεδομένου ότι $|m_{iu}| \le 1$,

$$\max_{1\le i,j\le n} |a_{ij}^{(u+1)}| = \max_{1\le i,j\le n} |a_{ij}^{(u)} - m_{iu}a_{uj}^{(u)}| \le \max_{1\le i,j\le n} \left(|a_{ij}^{(u)}| + \overset{\le 1}{|m_{iu}|}\,|a_{uj}^{(u)}|\right) \le 2 \max_{1\le i,j\le n} |a_{ij}^{(u)}|,$$

οπότε, $\max_{1\le i,j,u\le n} |a_{ij}^{(u)}| \le 2^{n-1} \max_{1\le i,j\le n} |a_{ij}|$

(Ο Wilkinson έδωσε παράδειγμα για το οποίο ισχύει η ισότητα)

Μάθημα 14ο ΕΞ. Υπολογιστικά Μαθηματικά Ι 26/11/2018

Για το ρ έχουμε, δεδομένου ότι $|m_{ik}| \le 1$,

$$\max_{1\le i,j\le n} |\alpha_{ij}^{(k+1)}| = \max_{1\le i,j\le n} |\alpha_{ij}^{(k)} - m_{ik}\alpha_{kj}^{(k)}| \le \max_{1\le i,j\le n} (|\alpha_{ij}^{(k)}| + |m_{ik}||\alpha_{kj}^{(k)}|) \le 2\cdot \max_{1\le i,j\le n} |\alpha_{ij}^{(k)}|$$

οπότε $\max_{1\le i,j,k\le n} |\alpha_{ij}^{(k)}| \le 2^{n-1} \cdot \max_{1\le i,j\le n} |\alpha_{ij}|$

(ο Wilkinson έδωσε παράδειγμα για το οποίο ισχύει η ισότητα)

Έτσι $\rho = \dfrac{\max_{1\le i,j,k} |\alpha_{ij}^{(k)}|}{\|A\|_\infty} = \dfrac{\max_{1\le i,j,k\le n} |\alpha_{ij}^{(k)}|}{\max_{1\le i\le n} \sum_{j=1}^{n} |\alpha_{ij}|} \le \dfrac{\max_{1\le i,j,k\le n} |\alpha_{ij}^{(k)}|}{\max_{1\le i,j\le n} |\alpha_{ij}|} \le 2^{n-1} \Rightarrow \rho \le 2^{n-1}$

Στην πράξη κάτι τέτοιο είναι εξαιρετικά σπάνιο. Εμπειρικά έχει βρεθεί ότι για την μερική οδήγηση ισχύει

$\max_{1\le i,j,k\le n} |\alpha_{ij}^{(k)}| \le 10$ αν $\max_{1\le i,j\le n} |\alpha_{ij}| \le 1 \Rightarrow \rho \le 10$

Αν συνδυάσουμε τώρα τα αποτελέσματα αυτής και της προηγούμενης παραγράφου για να εκτιμήσουμε το σφάλμα της $\tilde{x}$. Ας υποθέσουμε ότι το $\tilde{x}$ ικανοποιεί την $(A+\delta A)\tilde{x} = b$ με $\dfrac{\|\delta A\|}{\|A\|}$ για κάποια νόρμα

Δεδομένου ότι $\tilde{x} = x + \delta x$, $\kappa = \kappa(A) = \|A\|\cdot\|A^{-1}\|$ και αν

$\|\delta A\|\cdot\|A^{-1}\| = \dfrac{\|\delta A\|}{\|A\|}\cdot\|A\|\cdot\|A^{-1}\| = \mu\kappa < 1$, τότε

$$\frac{\|\delta x\|}{\|x\|} \le \frac{\kappa(A)}{1-\|\delta A\|\|A^{-1}\|}\cdot\frac{\|\delta A\|}{\|A\|} \Rightarrow \frac{\|\tilde{x}-x\|}{\|x\|} \le \frac{\kappa\cdot\mu}{1-\kappa\mu} \quad (15)$$

Συλλαδή, αν το μ είναι μικρό, το σχετικό σφάλμα της $\tilde{x}$ ως προς $\|x\|$ εξαρτάται από το $\kappa(A)$.

Αν ορίσουμε το υπόλοιπο της προσεγγιστικής λύσης $\tilde{x}$ με $r = A\tilde{x} - b$ έχουμε $r = A\tilde{x} - b = A\tilde{x} - (A+\delta A)\tilde{x} = -(\delta A)\tilde{x} \Rightarrow$

$$\|r\| \le \|\delta A\|\cdot\|\tilde{x}\| = \frac{\|\delta A\|}{\|A\|}\cdot\|\tilde{x}\| \Rightarrow \frac{\|r\|}{\|\tilde{x}\|} \le \|A\|\cdot\mu \quad (16)$$

Δεδομένου ότι το μ είναι γενικά μικρό (συνήθως $\mu = O(u)$ ή $\mu = O(nu)$)

βλέπουμε το υπόλοιπο της $\tilde{x}$ ως προς την $\|\tilde{x}\|$ που δίνει η απαλοιφή του Gauss με μερική οδήγηση είναι σχεδόν πάντα ανεξάρτητα από την κατάσταση του προβλήματος.

Επιπλέον

$$r = A\tilde{x} - b = A\tilde{x} - Ax = A(\tilde{x} - x) \Rightarrow \tilde{x} - x = A^{-1}r \Rightarrow \|\tilde{x} - x\| \le \|A^{-1}\|\,\|r\|$$

και από την (16)

$$\frac{\|\tilde{x} - x\|}{\|\tilde{x}\|} \le \|A^{-1}\|\,\frac{\|r\|}{\|\tilde{x}\|} \le \|A^{-1}\|\,\|A\|\,\mu \Rightarrow \frac{\|\tilde{x} - x\|}{\|\tilde{x}\|} \le \kappa(A)\,\mu \quad (17)$$

που πάλι δείχνει την σημασία του δείκτη κατάστασης στο σχετικό σφάλμα της $\tilde{x}$ ως προς $\|\tilde{x}\|$. Η εκτίμηση (15) είναι εκτίμηση a priori (εκ των προτέρων), ενώ η εκτίμηση (17) a posteriori (εκ των υστέρων)

Και στις δύο εμφανίζεται ο δείκτης κατάστασης, που είναι ο κρίσιμος παράγοντας, αφού $\mu = C_n u$ και το C_n δεν αυξάνεται γρήγορα με το n. Έτσι αν το κ είναι μεγάλο, μπορούμε να κάνουμε τις πράξεις με διπλή ακρίβεια, που βέβαια αυξάνει το κόστος

Παρατηρήσεις

① Αν κοιτάξουμε την βιβλιογραφία, θα δούμε ότι για τον ίδιο αλγόριθμο μια διαφορετική απόδειξη ή διαφορετικές υποθέσεις για την αριθμητική οδηγούν σε διαφορετικές σταθερές $C_n = 1{,}01(n^3 + 3n^2)$
Για αυτό δεν δίνουμε σημασία σε αυτές τις σταθερές και, στην πράξη, είναι συνήθως ασφαλές να υποθέσουμε ότι $C_n \rho = O(n)$

② Στον υπολογισμό (εκτίμηση) του $\kappa(A)$ προσπαθούμε να αποφύγουμε τον υπολογισμό του A^{-1}, αυτό μπορεί να γίνει ως ακολούθως:

Αν $Aw = y \Leftrightarrow w = A^{-1}y \Rightarrow \|A\| \le \|A^{-1}\|\,\|y\| \Rightarrow \|A^{-1}\| \ge \frac{\|w\|}{\|y\|} = \frac{\|A^{-1}y\|}{\|y\|}$

Έτσι επιλέγουμε κ-διανύσματα $y_i \neq 0$ $i = 1, 2, \ldots, \kappa$ λύνουμε τα κ-συστήματα $Aw_i = y_i$ $i = 1, 2, \ldots, \kappa$ και κατόπιν παίρνουμε

$$\|A^{-1}\| \cong \max_{1 \le i \le \kappa} \frac{\|w_i\|}{\|y_i\|} = \max_{1 \le i \le \kappa} \frac{\|A^{-1}y_i\|}{\|y_i\|}$$

(Ακριβής τιμή $\|A^{-1}\| = \sup_{\substack{y \in \mathbb{R}^n \\ y \neq 0}} \frac{\|A^{-1}y\|}{\|y\|}$)

Το επιπλέον κόστος είναι $O(kn^2)$. Εμπειρικά έχει αποδειχθεί ότι αν τα y_i, $i=1,2,\dots,k$ επιλέγουν τυχαία, τότε είναι ασφαλές να επιλέξουμε k μικρό συνήθως $k=2$ ή 3

2.3 Επαναληπτικές μέθοδοι για τη λύση γραμμικών συστημάτων

Έστω $A \in \mathbb{R}^{n\times n}$ αντιστρέψιμος. Θέλουμε να κατασκευάσουμε μια επαναληπτική μέθοδο, η οποία προσεγγίζει τη λύση του γραμμικού συστήματος $Ax=b$
Την ιδέα την παίρνουμε από τη μέθοδο του σταθερού σημείου στις μη γραμμικές εξισώσεις

$f(x)=0 \iff x=\varphi(x)$ οπότε

x_0 δεδομένο

$x_{k+1}=\varphi(x_k)$ $k=0,1,2,\dots$

και στη συνέχεια μελετάμε τη σύγκλιση της ακολουθίας $\{x_k\}_{k=0}^{\infty}$ στο σταθερό σημείο x^* της φ
Έτσι και εδώ μπορούμε να κατασκευάσουμε έναν πίνακα επανάληψης T έτσι ώστε:

$Ax=b \iff x=Tx+C$

Στη συνέχεια θα εφαρμόσουμε το επαναληπτικό σχήμα

x_0 δεδομένο

$x^{k+1}=Tx^k+C$ $k=0,1,2,\dots$

Προφανώς, διαφορετικοί πίνακες επανάληψης, δίνουν διαφορετικές μεθόδους.

Η ιδέα για την κατασκευή μεθόδων προέρχεται από τη μέθοδο του σταθερού σημείου, οπότε

$$Ax=b \iff x=Tx+c$$

για κατάλληλο T και c. Έτσι κατασκευάζουμε το επαναληπτικό σχήμα x^0 δεδομένο

$$x^{k+1}=Tx^k+c \quad k=0,1,2,\dots$$

Θα μελετήσουμε τη σύγκλιση της ακολουθίας x^k, $k=0,1,2,$

Θεώρημα: Έστω $\|\cdot\|$ μια νόρμα του $\mathbb{R}^n$ και $\|\cdot\|$ η νόρμα του $\mathbb{R}^{n\times n}$ που παράγεται από αυτήν. Αν $T\in\mathbb{R}^{n\times n}$ με $\|T\|<1$ και $c\in\mathbb{R}^n$, τότε υπάρχει μοναδικό $x\in\mathbb{R}^n$ τέτοιο ώστε $x=Tx+c$. Επιπλέον, για οποιοδήποτε x^0, η ακολουθία x^k, $k=0,1,\dots$, που παράγεται από το επαναληπτικό σχήμα $x^{k+1}=Tx^k+c$, $k=0,1,2,\dots$, συγκλίνει στο x. Μάλιστα ισχύουν οι εκτιμήσεις

$$\|x^k-x\| \le \frac{\|T\|^k}{1-\|T\|}\|x^1-x^0\|, \quad k=1,2,\dots, \qquad (18)$$

$$\|x^k-x\| \le \frac{\|T\|}{1-\|T\|}\|x^k-x^{k-1}\|, \quad k=1,2,\dots \qquad (19)$$

$$\|x^k-x\| \le \|T\|^k\|x^0-x\|, \quad k=1,2,\dots \qquad (20)$$

Απόδειξη:

Θεωρούμε την απεικόνιση $G:\mathbb{R}^n\to\mathbb{R}^n$ με $G(x)=Tx+c$. Για $x,y\in\mathbb{R}^n$ έχουμε:

$$\|G(x)-G(y)\| = \|Tx+c-Ty-c\| = \|T(x-y)\| \le \|T\|\,\|x-y\|$$

και εφόσον $\|T\|<1$, η G είναι συστολή στον $\mathbb{R}^n$ ως προς τη $\|\cdot\|$ με σταθερά $\alpha=\|T\|$. Τώρα το θεώρημά μας αποδεικνύεται από το θεώρημα της συστολής.

Έστω $e^k=x^k-x$. Έχουμε

$$e^k=x^k-x=Tx^{k-1}+c-Tx-c=T(\underbrace{x^{k-1}-x}_{e^{k-1}})=Te^{k-1} \Longrightarrow$$

$$e^k=Te^{k-1}=T^2e^{k-2}=\dots=T^ke^0, \quad k=1,2,\dots$$

Έτσι

$\|e^k\| = \|T^k e^0\| \le \|T^k\| \|e^0\| \le \|T\|^k \|e^0\| \quad k=1,2,\dots$

Επομένως, όσο πιο μικρή είναι η $\|T\|$, τόσο πιο γρήγορα $e^k \to 0$ καθώς $k \to \infty$, δηλαδή τόσο πιο γρήγορα $x^k \to x$ καθώς $k \to \infty$

Θεώρημα: Έστω $A \in \mathbb{R}^{n\times n}$. Τότε $\lim_{k\to\infty} A^k = 0$ αν και μόνο αν $\rho(A) < 1$

Απόδειξη:

($\Rightarrow$) Έστω $\lim_{k\to\infty} A^k = 0$. Αν λ είναι ιδιοτιμή του A, τότε υπάρχει $0 \neq x \in \mathbb{R}^n$ έτσι ώστε $Ax = \lambda x$. Έχουμε $A^k x = \lambda^k x \Rightarrow$

$$0 = x \lim_{k\to\infty} A^k = \lim_{k\to\infty} A^k x = \lim_{k\to\infty} \lambda^k x = x \cdot \lim_{k\to\infty} \lambda^k \overset{x\neq 0}{\Longrightarrow} \lim_{k\to\infty} \lambda^k = 0 \Rightarrow |\lambda| < 1$$

Άρα, $\rho(A) < 1$

($\Leftarrow$) Υποθέτουμε ότι $\rho(A) < 1$. Για $0 < \varepsilon < 1 - \rho(A)$ υπάρχει $\|\cdot\|$ του $\mathbb{R}^{n\times n}$ τέτοια ώστε $\|A\| \le \rho(A) + \varepsilon < 1$

Έτσι για αυτή τη $\|\cdot\|$ $\|A^k\| \le \|A\|^k \to 0$ καθώς $k \to \infty \Rightarrow A^k \to 0$ καθώς $k \to \infty$

Δίνουμε τώρα μια ικανή και αναγκαία συνθήκη ώστε $x^k \to x$ καθώς $k \to \infty$

Θεώρημα: Έστω $T \in \mathbb{R}^{n\times n}$ και $c \in \mathbb{R}^n$. Τότε για οποιοδήποτε $x^0 \in \mathbb{R}^n$, η ακολουθία x^k, $k = 0,1,2,\dots$, που παράγεται από το επαναληπτικό σχήμα $x^{k+1} = Tx^k + c$, $k = 0,1,2,\dots$, συγκλίνει στο μοναδικό x έτσι ώστε $x = Tx + c$ αν και μόνο αν $\rho(T) < 1$

Απόδειξη: ($\Rightarrow$) Έστω ότι η ακολουθία x^k συγκλίνει στη μοναδική λύση του $x = Tx + c$ δηλαδή $x^k \to x$ καθώς $k \to \infty$. Τότε: $\lim_{k\to\infty} x^k = x \Leftrightarrow \lim_{k\to\infty} e^k = 0 \Leftrightarrow \lim_{k\to\infty} T^k e^0 = 0$

$\Leftrightarrow e^0 \lim_{k\to\infty} T^k = 0 \Leftrightarrow \rho(T) < 1$ (από το προηγούμενο θεώρημα)

($\Leftarrow$) Υποθέτουμε ότι $\rho(T) < 1$ τότε υπάρχει $\|\cdot\|$ του $\mathbb{R}^{n\times n}$ τέτοια ώστε $\|T\| < 1$ (βλέπε απόδειξη προηγούμενου θεωρήματος)

Επομένως τα συμπεράσματα έπονται από το πρώτο θεώρημα αυτής της παραγράφου.

Μάθημα 16ο ΕΞ. Υπολογιστικά Μαθηματικά Ι 3/12/2018

2.3.1 Μέθοδοι του Jacobi και των Gauss-Seidel

Ας υποθέσουμε ότι θέλουμε να λύσουμε το γραμμικό σύστημα τριών εξισώσεων με τρεις αγνώστους

$$10x_1 - x_2 + 2x_3 = 13$$
$$-x_1 + 11x_2 - x_3 = -13$$
$$2x_1 - x_2 + 10x_3 = 13$$

Αν λύσουμε τις παραπάνω εξισώσεις ως προς x_1, x_2, x_3 αντίστοιχα

$$x_1 = \frac{x_2}{10} - \frac{2x_3}{10} + \frac{13}{10}$$
$$x_2 = \frac{x_1}{11} + \frac{x_3}{11} - \frac{13}{11}$$
$$x_3 = -\frac{2x_1}{10} + \frac{x_2}{10} + \frac{13}{10}$$

Παρατηρούμε ότι αν ξέρουμε τα x_2 και x_3 μπορούμε να υπολογίσουμε το x_1 και αντίστοιχα για τα x_2 και x_3

Έτσι, θεωρούμε την εξής επαναληπτική μέθοδο:

1) Επιλέγουμε μια αρχική προσέγγιση $x_1^\circ, x_2^\circ, x_3^\circ$
2) Για $k = 0, 1, 2, \ldots$ υπολογίζουμε

$$x_1^{k+1} = \frac{x_2^k}{10} - \frac{2x_3^k}{10} + \frac{13}{10}$$

$$x_2^{k+1} = \frac{x_1^k}{11} + \frac{x_3^k}{11} - \frac{13}{11}$$

$$x_3^{k+1} = -\frac{2x_1^k}{10} + \frac{x_2^k}{10} + \frac{13}{10}$$

Αν ξεκινήσουμε με $x_1^\circ = x_2^\circ = x_3^\circ = 0$, παίρνουμε

k	1	2	3	...	104
x_1^k	1.3	0.922	1.021	...	1.000
x_2^k	-1.1818	-0.945	-1.014	...	-1.000
x_3^k	1.3	0.922	1.021	...	1.000

Αυτή ακριβώς είναι η μέθοδος του Jacobi Το σχήμα της στη γενική περίπτωση είναι:

$$x_i^{k+1} = \frac{\sum_{j=1, j\neq i}^{n} a_{ij} x_j^k + b_i}{a_{ii}} \quad , i = 1, 2, \ldots, n \quad , k = 0, 1, 2, \ldots \qquad (21)$$

για δεδομένα $x_1^0, x_2^0, \ldots, x_n^0$ Οπότε θα πρέπει $a_{ii} \neq 0$, $i = 1, 2, \ldots, n$

Ας προσπαθήσουμε να εκφράσουμε το σχήμα της μεθόδου στη μορφή

$$x^{k+1} = T_J x^k + C_J \quad k = 0, 1, 2, \ldots$$

όπου $T_J \in \mathbb{R}^{n \times n}$ και $C_J \in \mathbb{R}^n$ Καταρχήν γράφουμε τον πίνακα A στη μορφή

$A = D - L - U$, όπου

$D = \text{diag}(d_1, d_2, \ldots, d_n)$ $d_i = a_{ii}$, $1 \leq i \leq n$

$L = (l_{ij})_{i,j=1}^n$ με $l_{ij} = \begin{cases} -a_{ij}, & \text{αν } i > j \\ 0, & \text{αλλιώς} \end{cases}$ (αυστηρά κάτω τριγωνικός)

$U = (u_{ij})_{i,j=1}^n$ με $u_{ij} = \begin{cases} -a_{ij}, & \text{αν } i < j \\ 0, & \text{αλλιώς} \end{cases}$ (αυστηρά άνω τριγωνικός)

$Ax = b \iff (D - L - U)x \iff Dx = (L + U)x + b \underset{i=1,\ldots,n}{\overset{a_{ii} \neq 0}{\iff}} x = \underbrace{D^{-1}(L+U)}_{T_J} x + \underbrace{D^{-1}b}_{C_J}$

$\iff x = T_J x + C_J$

Άρα $x^{k+1} = T_J x^k + C_J$, $T_J = D^{-1}(L+U)$, $C_J = D^{-1}b$, $k = 0, 1, 2, \ldots$ (22)

Τώρα, όταν υπολογίζουμε το x_i^{k+1} ως προς τα $x_1^k, x_2^k, \ldots, x_{i-1}^k$ και τα $x_{i+1}^k, x_{i+2}^k, \ldots, x_n^k$, μπορούμε στην θέση των $x_1^k, x_2^k, \ldots, x_{i-1}^k$ να χρησιμοποιήσουμε τα $x_1^{k+1}, x_2^{k+1}, \ldots, x_{i-1}^{k+1}$ που έχουμε ήδη υπολογίσει και είναι, γενικά, καλύτερες προσεγγίσεις από τα $x_1^k, x_2^k, \ldots, x_{i-1}^k$

Έτσι, στο παράδειγμά μας έχουμε:

1) Επιλέγουμε μια αρχική προσέγγιση x_1^0, x_2^0, x_3^0

2 Για $k = 0, 1, 2, \ldots$ υπολογίζουμε

$$x_1^{k+1} = \frac{x_2^k}{10} - \frac{2x_3^k}{10} + \frac{13}{10}$$

$$x_2^{k+1} = \frac{x_1^{k+1}}{11} + \frac{x_3^k}{11} - \frac{13}{11}$$

$$x_3^{k+1} = -\frac{2x_1^{k+1}}{10} + \frac{x_2^{k+1}}{10} + \frac{13}{10}$$

Αν ξεκινήσουμε με $x_1^0 = x_2^0 = x_3^0 = 0$, παίρνουμε:

k	1	2	3	...
x_1^k	1.3	1.007	1.000	...
x_2^k	-1.063	-1.005	-1.000	...
x_3^k	0.934	0.998	1.000	...

Αυτή ακριβώς είναι η μέθοδος των Gauss-Seidel. Το σχήμα της στη γενική περίπτωση είναι:

$$x_i^{k+1} = \frac{-\sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij} x_j^k + b_i}{a_{ii}}, \quad i=1,2,\dots,n, \ k=0,1,2,\dots \qquad (23)$$

για δεδομένα $x_1^0, x_2^0, \dots, x_n^0$. Οπότε, θα πρέπει $a_{ii} \neq 0$, $i=1,2,\dots,n$

Ας προσπαθήσουμε να εκφράσουμε το σχήμα της μεθόδου στη μορφή

$$x^{k+1} = T_{GS} x^k + c_{GS}, \quad k=0,1,2,\dots$$

όπου $T_{GS} \in \mathbb{R}^{n\times n}$ και $c_{GS} \in \mathbb{R}^n$. Έχουμε

$$Ax=b \Leftrightarrow (D-L-U)x = b \Leftrightarrow (D-L)x = Ux + b \overset{a_{ii}\neq 0,\ i=1,\dots,n}{\Longleftrightarrow}$$

$$x = \underbrace{(D-L)^{-1}U}_{T_{GS}} x + \underbrace{(D-L)^{-1}b}_{c_{GS}}$$

Άρα

$$x^{k+1} = T_{GS} x^k + c_{GS}, \ T_{GS} = (D-L)^{-1}U, \ c_{GS} = (D-L)^{-1}b, \ k=0,1,2, \qquad (24)$$

Στην πράξη χρησιμοποιούνται τα σχήματα (21) και (23) αντί των (22) και (24) αντίστοιχα

Ορισμός: Ο πίνακας $A \in \mathbb{R}^{n\times n}$, $A=(a_{ij})_{i,j=1}^n$ ονομάζεται διαγώνια υπερτερών κατά γραμμές αν:

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

Ο πίνακας είναι αυστηρά διαγώνια υπερτερών κατά γραμμές αν στη παραπάνω ανισότητα ισχύει ">" αντί του "≥"

Μια καλή συνθήκη για τη σύγκλιση των μεθόδων Jacobi και Gauss-Seidel δίνεται στο επόμενο θεώρημα:

Θεώρημα: Έστω ότι ο πίνακας A είναι αυστηρά διαγώνια υπερτερών κατά γραμμές.

(i) Οι πίνακες επανάληψης T_J και T_{GS} των μεθόδων Jacobi και Gauss-Seidel ικανοποιούν τις ανισότητες $\|T_J\|_\infty < 1$ και $\|T_{GS}\|_\infty < 1$

(ii) Οι μέθοδοι Jacobi και Gauss-Seidel συγκλίνουν

Απόδειξη:

(i) Έχουμε $T_J = D^{-1}(L+U)$, επομένως $\sum_{j=1}^{n} |(T_J)_{ij}| = \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{|\alpha_{ij}|}{|\alpha_{ii}|} < 1$

λόγω του ότι A είναι αυστηρά διαγώνια υπερτερών κατά γραμμές

Οπότε $\|T_J\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |(T_{ij})| < 1$

Τώρα $T_{GS} = (D-L)^{-1}U$ και για να δείξουμε ότι $\|T_{GS}\|_\infty < 1$, αν $y \in \mathbb{R}^n, y\neq 0$ πρέπει να δείξουμε ότι $\|T_{GS}\, y\|_\infty < \|y\|_\infty$

Θέτουμε $u = T_{GS}\, y = (D-L)^{-1}Uy \Leftrightarrow (D-L)u = Uy$, οπότε λόγω των (23) και (24) έχουμε

$$u_i = \frac{-\sum_{j=1}^{i-1} \alpha_{ij}u_j - \sum_{j=i+1}^{n} \alpha_{ij}y_j}{\alpha_{ii}}, \quad i = 1,2,\dots,n$$

Συνεπώς αρκεί να δείξουμε ότι $|u_i| < \|y\|_\infty$, $i = 1,2,\dots,n$ (25)

Κατ' αρχήν,

$$u_1 = -\frac{\sum_{j=2}^{n} \alpha_{1j}y_j}{\alpha_{11}} \Rightarrow |u_1| \le \frac{\sum_{j=2}^{n} |\alpha_{1j}||y_j|}{|\alpha_{11}|} \le \|y\|_\infty \sum_{j=2}^{n} \frac{|\alpha_{1j}|}{|\alpha_{11}|} < \|y\|_\infty \cdot 1 = \|y\|_\infty$$

εφόσον ο A είναι αυστηρά διαγώνια υπερτερών κατά γραμμές

Έστω $|u_j| < \|y\|_\infty$, $j = 1,2,\dots,i-1$

Τότε, $|u_i| \le \frac{\sum_{j=1}^{i-1} |\alpha_{ij}||u_j| + \sum_{j=i+1}^{n} |\alpha_{ij}||y_j|}{|\alpha_{ii}|} < \|y\|_\infty \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{|\alpha_{ij}|}{|\alpha_{ii}|} < \|y\|_\infty \cdot 1 = \|y\|_\infty$

Εφόσον ο A είναι αυστηρά διαγώνια υπερτερών κατά γραμμές

Άρα η (25) ισχύει

(ii) Άμεση συνέπεια του (i) και του πρώτου θεωρήματος της προηγούμενης παραγράφου.

Μάθημα 17ο ΕΞ. Υπολογιστικά Μαθηματικά Ι 5/12/2018

Ενδιαφέρον παρουσιάζει το ερώτημα ποιά από τις δύο μεθόδους συγκλίνει ταχύτερα. Γενικά, δεν υπάρχει απάντηση στο ερώτημα αυτό, εκτός από ειδικές κατηγορίες πινάκων

Θεώρημα (Stein-Rosenberg): Έστω $A \in \mathbb{R}^{n\times n}$ έτσι ώστε $a_{ij} \le 0$ για $i \neq j$ και $a_{ii} > 0$, $i = 1, 2, \ldots, n$. Τότε, ισχύει ένα και μόνο ένα από τα επόμενα:

α) $0 < \rho(T_{GS}) < \rho(T_J) < 1$

β) $1 < \rho(T_J) < \rho(T_{GS})$

γ) $\rho(T_J) = \rho(T_{GS}) = 0$

δ) $\rho(T_J) = \rho(T_{GS}) = 1$

Έτσι, στην ειδική αυτή περίπτωση που και οι δύο μέθοδοι συγκλίνουν ή αποκλίνουν, η Gauss-Seidel συγκλίνει ή αποκλίνει ταχύτερα

Παρατηρήσεις

(1) Η αυστηρή διαγώνια υπεροχή κατά γραμμές δεν είναι απαραίτητη για να συγκλίνει η μέθοδος του Jacobi ή των Gauss-Seidel. Για παράδειγμα, η μέθοδος των Gauss-Seidel συγκλίνει για τον πίνακα

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 2 & 4 & 2 \\ 2 & 0 & 6 \end{pmatrix}$$

(2) Αν ο πίνακας A είναι συμμετρικός και θετικά ορισμένος, τότε η μέθοδος Gauss-Seidel συγκλίνει άλλα η μέθοδος Jacobi (γενικά) δεν συγκλίνει. Για παράδειγμα, ο πίνακας

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

Εξαίρεση για τη μέθοδο Jacobi αποτελούν οι 2×2 συμμετρικοί, θετικά ορισμένοι πίνακες

(3) Δεν ισχύει ότι όταν η μέθοδος Jacobi συγκλίνει, τότε η μέθοδος των Gauss-Seidel συγκλίνει ταχύτερα. Υπάρχουν πίνακες A για τους οποίους η μέθοδος Jacobi συγκλίνει και η μέθοδος των Gauss-Seidel δεν συγκλίνει. Για παράδειγμα, ο πίνακας.

$$A = \begin{pmatrix} 1 & 2 & -7/2 \\ 1 & 1 & 7/4 \\ 1 & 1 & 1 \end{pmatrix}$$

(4) Οι επαναληπτικές μέθοδοι δεν είναι καλύτερες ή χειρότερες από την απαλοιφή του Gauss για πίνακες με κακή κατάσταση

2.3.2 Μέθοδοι relaxation

Η μέθοδος Gauss-Seidel, είναι γενικά, ελκυστική εκτός από τις περιπτώσεις που το $\rho(T_{GS})$ είναι κοντά στην μονάδα (≤ 1), οπότε η σύγκλιση, η οποία εξαρτάται από την $\|T_{GS}\|^k$ και επομένως από $[\rho(T_{GS})]^k$ είναι πολύ αργή.

Έτσι αν x^k είναι μια γνωστή προσέγγιση της x και $\tilde{x}^{k+1}$ η προσέγγιση Gauss-Seidel που κατασκευάζεται από την x^k, τότε εισάγουμε την παράμετρο $\omega > 0$ έτσι ώστε η προσέγγιση

$$x^{k+1} = \omega \tilde{x}^{k+1} + (1-\omega) x^k$$

να έχει μικρότερο σφάλμα από την $\tilde{x}^{k+1}$.

Κατ' αυτον τον τρόπο κατασκευάζουμε μια καινούργια επαναληπτική μέθοδο, το σχήμα της οποίας είναι

$$x_i^{k+1} = \omega \frac{-\sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij} x_j^k + b_i}{a_{ii}} + (1-\omega) x_i^k \quad , i = 1, 2, \ldots, n \quad , k = 0, 1, 2, \ldots$$

για δεδομένα $x_1^0, x_2^0, \ldots, x_n^0$. Οπότε θα πρέπει $a_{ii} \neq 0 \quad i = 1, 2, \ldots, n$

Σε μορφή πινάκων έχουμε:

$$x^{k+1} = (D - \omega L)^{-1} [(1-\omega) D + \omega U] x^k + \omega (D - \omega L)^{-1} b \quad , k = 0, 1, 2, \ldots$$

Για $0 < \omega < 1$, η μέθοδος ονομάζεται under-relaxation

Για $\omega > 1$, η μέθοδος ονομάζεται over-relaxation (γνωστή και ως OSR)

και χρησιμοποιείται για να επιταχύνουμε τη σύγκλιση συστημάτων για τα οποία η μέθοδος των Gauss-Seidel συγκλίνει (αρχά)
Για $\omega=1$, παίρνουμε τη μέθοδο των Gauss-Seidel.
Ασφαλώς, το κρίσιμο ερώτημα είναι πως επιλέγουμε το ω.
Στο ερώτημα αυτό **δεν** υπάρχει απάντηση που να καλύπτει όλες τις περιπτώσεις. Παρ' όλα αυτά, για συγκεκριμένα συστήματα, η επιλογή του ω είναι δυνατή. Έστω

$$T_\omega = (D-\omega L)^{-1}[(1-\omega)D + \omega U]$$

Το επόμενο θεώρημα δείχνει ότι στις μεθόδους relaxation μόνο τιμές της παραμέτρου ω μεταξύ 0 και 2, στην καλύτερη περίπτωση, οδηγούν σε συγκλίνουσες μεθόδους

Θεώρημα (Kahan): Αν $a_{ii} \neq 0$, $i=1,2,\dots,n$, τότε $\rho(T_\omega) \geq |\omega-1|$

Για πίνακες που ικανοποιούν μια ιδιότητα ή έχουν μια ειδική μορφή, τα ω για τα οποία έχουμε σύγκλιση, ή ακόμα και το βέλτιστο ω, που δίνει την ταχύτερη σύγκλιση μπορούν να υπολογιστούν.

Θεώρημα (Ostrowski-Reich): Αν ο A είναι θετικά ορισμένος, τότε $\rho(T_\omega) < 1$ για όλα $0 < \omega < 2$

Θεώρημα: Αν ο A είναι θετικά ορισμένος και τριδιαγώνιος, τότε $\rho(T_{GS}) = [\rho(T_J)]^2 < 1$ και η βέλτιστη επιλογή για το ω στη μέθοδο relaxation είναι: $\omega = \dfrac{2}{1+\sqrt{1-\rho(T_{GS})}}$

Με αυτή την επιλογή του ω, $\rho(T_\omega) = \omega - 1$

$$\rho(T_\omega) = \omega - 1 = \frac{2}{1+\sqrt{1-\rho(T_{GS})}} - 1 = \dots = \frac{\rho(T_{GS})}{\left(1+\sqrt{1-\rho(T_{GS})}\right)^2}$$