Κεφάλαιο 1-2

Σκοπος Ε.Υ σχεδιασμος, ανάπτυξη, αξιολόγηση, χρήση υπολογιστικών εργαλείων για την επίλυση μαθηματικών μοντέλων

Υπολογιστικοί πυρήνες τμήματα του προγράμματος/υποπρογράμματα που καταναλώνουν πολύ χρόνο
π.χ. FFT, M*V, random numbers

Κριτήρια αξιολόγησης εργαλείων Ε.Υ. α) Ακρίβεια β) Ταχύτητα c) Κόστος

Πηγές Σφαλμάτων στον Ε.Υ. 1) Δεδομένα 2) Διακριτοποίηση εξισώσεων γ) Διακριτοποίηση $\mathbb{R} \to F$ 4) πεπερασμένος # επαναλήψεων 5) Ενδιάμεσα/ενδιάμεσα αποτελέσματα που δεν είναι προβλέψιμα από την μέθοδο επίλυσης

Φ : μεταφορές Ω : πράξεις $\mu = \frac{\Phi}{\Omega}$ $\frac{\text{μεταφορές}}{\text{πράξεις}}$

Κεφάλαιο 3

A.K.X $y = \pm m * \beta^{e}$

Μέγιστο σχετικό σφάλμα $u = \frac{\beta^{1-t}}{2}$ για $\beta = 2$, $u = 2^{-t}$

Eps $1 \tilde{+} eps = 1$

Συνθήκη ακριβούς στρογγύλευσης $x \tilde{\odot} y = fl(x \odot y)$

FMA $fl(z + x*y) = (z + x*y)(1+\delta)$ αντί για $(z + (x*y)(1+\delta_1))(1+\delta_2)$

Δείκτης Κατάστασης προβλήματος $\frac{\|f'(x)\| * \|x\|}{\|f(x)\|}$

Ανάλυση προς τα εμπρός $\prod^{n}(1+\delta_i) = 1+\theta_n$
$|\theta_n| \le \frac{nu}{1-nu} = \gamma_n$

Πίσω Σφάλμα $\frac{\|x_{proj} - x\|}{\|x\|}$

π.χ. Αν υπολογίζω ως $(x \tilde{*} y \tilde{*} z) = (xyz)(1+\theta_5)$ τότε
$\frac{|x\check{y}z - xyz|}{|xyz|} = |1+\theta_5 - 1| = |\theta_5| \le \gamma_5 = \frac{5u}{1-5u}$

Γιατί εμπρός σφάλμα $\frac{\|f_{proj}(x^*) - f(x)\|}{\|f(x)\|} \le$ Δ.Κ * πίσω σφάλμα αλγορίθμου

* Jacobian
$\left[\frac{df}{dx_1}, \frac{df}{dx_2}, \dots\right]$

Δ.Κ αν αντιστρέψιμο, Δ.Κ $= \|A\| \|A^{-1}\|$

(0.25)

Κεφάλαιο 4

		C^n	Ω	ϕ_{min}
sDOT	$\sigma \leftarrow \sigma + x^T y$		$2n$	$2n+2$
sAXPY	$y \leftarrow y + \alpha x$		$2n$	$3n+1$
ενημέρωση 1ης τάξης	$C \leftarrow C + a b^T$ (R^{n_1}, R^{n_2})		$2n_1 n_2$	$n_2 + n_1 + 2n_1 n_2$
M*V	$y \leftarrow y + Ax$			
	i) με DOT → κατά γραμμές		$2n_1 n_2$	$n_1 n_2 + 2n_1 + n_2$
	ii) με sAXPY → κατά στήλες		$2n_1 n_2$	$2n_1 n_2 + 2n_1$
M*M	$C \leftarrow C + AB$		$2n_1 n_2 n_3$	$2n_1 n_2 + n_2 n_3 + n_1 n_3$

άρα ο DOT υπερτερεί

Κεφάλαιο 5

Τύποι μητρώων: Hessenberg σχεδόν τριγωνικό

Toeplitz $\begin{bmatrix} \alpha & b & b \\ d & \alpha & b \\ e & d & \alpha \end{bmatrix}$ Hankel $\begin{bmatrix} \alpha & b & c \\ b & c & d \\ c & d & e \end{bmatrix}$ Κυκλοτερές $\begin{bmatrix} \alpha & b & c \\ c & \alpha & b \\ b & c & \alpha \end{bmatrix}$

Στοιχειώδη μητρώα $E(u, v; \tau) = I - \tau u v^T$

Ι.στοιχειώδη τριγωνικά $L_k(u) = E(u, e_k; 1)$ ε.ω $u = [0, \dots, 0, u_{k+1}, u_{k+2}, \dots, u_k]$
$e_k = [0, \dots, 1, 0, 0, \dots, 0]$

και άρα είναι στοιχειώδης μετασχηματισμός Gauss

Παραγοντοποίηση Schur: Αν A_{11} αντιστρέψ. τότε $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} I & 0 \\ L_{21} & I \end{pmatrix} \begin{pmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{pmatrix}$

όπου: $U_{11} = A_{11}$ $U_{12} = A_{12}$ $U_{21} = A_{21} A_{11}^{-1}$ $U_{22} = \underbrace{A_{22} - A_{21} A_{11}^{-1} A_{12}}_{\text{συμπλήρωμα Schur}}$

Οδήγηση στην LU:
i) μερική: διαλέγω ως οδηγό το μέγιστο κατ' απόλυτη τιμή της στήλης $[k:n]$
ii) πλήρης: διαλέγω το .... του πίνακα $[k:n, k:n]$

συντελεστής αύξησης $p_n = \frac{\max\{\alpha_0, \alpha_1, \dots \alpha_k\}}{\alpha_0}$ όπου $\alpha_0 \in \{r,u; r,n\}$ το μεγαλύτερο στοιχείο

Σφάλμα:
δείκτης Skeel $\| |A^{-1}| |A| \|_\infty$

πίσω σφάλμα $\eta = \frac{\|r\|}{\|A\| \|\tilde{x}\| + \|b\|}$ όπου $r = b - A\tilde{x}$

→ η υπολογισμένη λύση, δηλ. του $(A + \Delta A)\tilde{x} = (b + \Delta b)$

πίσω ευστάθεια $\| |\hat{L}| |\hat{U}| \|$

Σπυριδόπουλος (0.5)

Κεφάλαιο 8

Εμπρός Euler $\rightarrow u(t+\Delta t) = u(t) + \Delta t \cdot f(t, u(t))$

Αν $u'(t) = A \cdot u(t)$

τότε $u_{k+1} = (I + hA)^{k+1} u_0$

Αν π.χ. $A = \text{diag}[-100, -1, -0.5]$ τότε $u_{k+1} = \begin{pmatrix} (1-100h)^{k+1} & & \\ & (1-h)^{k+1} & \\ & & (1-0.5h)^{k+1} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$

! Γίνεται ευσταθής πρέπει

$$\left. \begin{array}{l} |1-100h| < 1 \\ |1-h| < 1 \\ |1-0.5h| < 1 \end{array} \right\}$$ βρίσκω το μέγιστο h που τις ικανοποιεί

Πίσω Euler $\rightarrow$ $\frac{u(t) - u(t-\Delta t)}{\Delta t} = f(t, u(t))$

Άρα σε κάθε βήμα λύνω το σύστημα

$$(I - \Delta t A(t))\, u(t) = u(t-\Delta t) \quad (\text{μορφή } Ax = b$$

$$\Leftrightarrow u(t) = (I - \Delta t A(t))^{-1} u(t-\Delta t)$$

! Πάντα είναι ευσταθής $\rightarrow$ λύνει

Μέθοδοι Taylor

Προσεγγίζω και τις $u''(t)$ με $u''(t) = \frac{u'(t) - u'(t-h)}{h}$

↳ χρειαζόμαστε τα $u(t), u(t-h)$ $\rightarrow$ πολυβηματική

Μέθοδοι Runge-Kutta

Επιλέγουμε τιμές και σε αυτά σημεία στο $[0, h]$ και χρησιμοποιούμε όλες αυτές τις πληροφορίες για να πάμε από το $t_n \rightarrow t_{n+1}$

π.χ. $K_1 = f(t_n, u_n)$

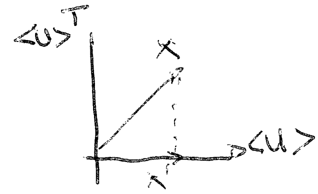
$K_2 = f(t_{n+1}, u_n + hK_1)$

$u_{n+1} = u_n + \frac{h}{2}(K_1 + K_2) \ldots$

(0.75)

<u>[illegible]</u>

Κεφάλαιο 6



το $P = \dfrac{uu^T}{u^Tu}$ το διάνυσμα προβολής

$x' = Px$

[illegible]

Χρονική Πολυπλοκότητα T(n): πόσες πράξεις χρειάζονται για την περάτωση ενός αλγορίθμου

Παραδοχές και απλό Μοντέλο

- Η Load, όταν τα δεδομένα είναι στην κύρια μνήμη χρειάζεται χρόνο τ_{mem}

  M θέσεις >> K θέσεις

  Όταν όμως τα δεδομένα βρίσκονται στην κρυφή μνήμη, τότε χρειάζεται χρόνο $\tau_{mem}^{(0)} \neq 0$

  ισχύει $\tau_{mem}^{(0)} \ll \tau_{mem}$

- Η εκτέλεση των πράξεων +, -, *, / χρειάζεται $\tau_{αρθ}$
- Η εκτέλεση τερματίζεται όταν αποθηκευτούν στη μνήμη όλα τα αποτελέσματα

Μέτρηση Ανάλυσης Επίδοσης

Ω: ο αριθμός πράξεων (αριθμητικών, + - * / κτλ.)

Φ: ο αριθμός μεταφορών μεταξύ μνήμης - καταχωρητών

Φ_{min}: ο ελάχιστος αριθμός μεταφορών που θα χρειάζονταν αν είχαμε άπειρη μνήμη σε όλα τα επίπεδα.

Θεώρημα: Για τα δεδομένα [illegible] ο αλγόριθμος [illegible], είναι $\Phi_{min} \geq$ [illegible]

Απ: η μεταφορά των δεδομένων [illegible]

Κόστος

Άρα το κόστος εκτέλεσης T είναι

$$T = T_{αριθ} + T_{μεσ}$$

$$= τ_{αριθ}\, C + τ_{μεσ}\, Φ$$

$$= T_{αριθ}\left(1 + μ \frac{τ_{μεσ}}{τ_{αριθ}}\right)$$

* $μ = \frac{Φ}{C}$, δηλαδή ο αριθμός [illegible]

$μ_{min} = \frac{Φ_{min}}{C}$

Συμβολισμός

$A[10:14] = [A_{10}, A_{11}, A_{12}, A_{13}, A_{14}]$

$A[10:2:14] = [A_{10}, A_{12}, A_{14}]$

$$A[\underbrace{2:3}_{\text{τιμές γραμμές}}, \underbrace{2:2:4}_{\text{τιμές στήλες}}] = \begin{bmatrix} A_{2,2} & A_{2,4} \\ A_{3,2} & A_{3,4} \end{bmatrix}$$

Βασικές πράξεις γραμμικής άλγεβρας 1ου επιπέδου (BLAS-1)

[illegible]

SDOT ("εσωτερικό")

$$\sigma \leftarrow \sigma + x^T y$$

* Εσωτερικό γινόμενο

Αλγόριθμος

```
LOAD x, y
for i=1:n
    σ = σ + xi yi
end
STORE σ
```

για $x, y \in \mathbb{C}^n$ [illegible]

$\underline{\Theta} = 2n$ [illegible]

$\Phi = \begin{cases} \min\ \underline{2n+2} & \text{[illegible]} \\ \underline{3n+1} & \text{[illegible]} \end{cases}$

SAXPY

$$y \leftarrow y + \overset{\text{βαθμωτό}}{\alpha} x \quad , \quad y, x \in \mathbb{R}^n$$

Αλγόριθμος

```
LOAD α, x, y
for i=1:n
    yi = yi + α xi
end
STORE y
```

$\underline{\Theta} = 2n$

$\Phi = \min\ \underline{3n+1}$ [illegible]

! Και οι 2 παραπάνω υλοποιήσεις έχουν απαίτηση για κρυφή μνήμη $O(n)$.
Παρακάτω υπάρχουν υλοποιήσεις που απαιτούν κρυφή μνήμη $O(1)$

DOT

```
LOAD σ
for i=1:n
    LOAD xi, yi
    σ = σ + xi yi
end
STORE σ
```

SAXPY

```
LOAD α
for i=1:n
    LOAD xi, yi
    yi = yi + α xi
    STORE yi
end
```

Ανανέωση 1ης τάξης (GER) ← BLAS-2

$$C \leftarrow C + \alpha b^T$$

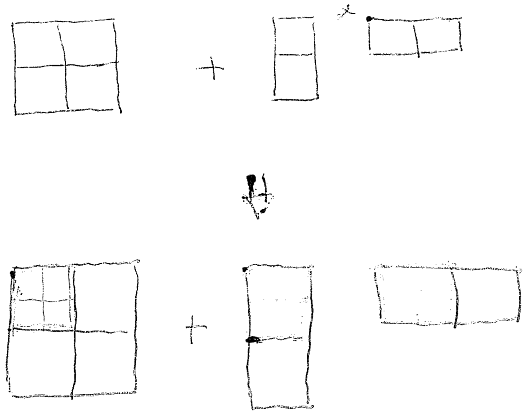
Αλγόριθμος

```
for j = 1 : n2
  LOAD bj
  for i = 1 : n1
    LOAD cij, αi
    cij = cij + αi * bj
    STORE cij
  end
end
```

! Υπάρχουν πολλές εναλλακτικές ώστε να βελτιστοποιήσουμε είτε τον χώρο, είτε τον χρόνο (tradeoff)

→ Ομαδοποίηση ως λύση όταν δεν έχουμε αρκετό χώρο στην κρυφή μνήμη. Ουσιαστικά "σπάμε" τον υπολογισμό μας σε μικρότερους υπολογισμούς που ο καθένας γίνεται με βέλτιστο μ, και συνεπώς όλος ο υπολογισμός θα γίνει με βέλτιστο μ.

π.χ



$2^{ου}$ Επιπέδου (BLAS - 2)

Πολλαπλασιασμός Μητρώο × Διάνυσμα (MV)

$$y \leftarrow y + Ax \quad , \text{ για } A \in \mathbb{R}^{n_1 \times n_2}$$

i) Using DOT, κατά γραμμές

τότε $\quad \eta_i = \eta_i + \sum_{k=1}^{n_2} \alpha_{ik} \xi_k \quad , \quad i = 1, \ldots, n_1$

Αλγόριθμος

```
for i = 1:n1
  for j = 1:n2
     ηi = ηi + αij * ξj
  end
end
```

Ανάλυση

$\Omega = 2 n_1 n_2$

$\Phi^{\text{DOT}} = n_1 n_2 + 2 n_1 + n_2$

ii) Using SAXPY, κατά στήλες

τότε $\quad y = y + \sum_{k=1}^{n_2} \alpha_{:,k} \xi_k$

Αλγόριθμος

```
for j = 1:n2
  for i = 1:n1
     ηi = ηi + αij * ξj
  end
end
```

Ανάλυση

$\Omega = 2 n_1 n_2$

$\Phi^{\text{SAXPY}} = 3 n_1 n_2 + n_2$

[illegible]

$3^{ου}$ Επιπέδου (BLAS-3)

Πολλαπλασιασμός Μητρώο × Μητρώο (ΜΜ)

$$C \leftarrow C + AB \quad , \text{[illegible]}$$

πολλές διαφορετικές υλοποιήσεις, σελ. 124 βιβλίο

Γενικά είναι $\quad \underline{c} = 2 n_1 n_2 n_3$

$$\phi_{min} = 2n_1n_2 + n_2n_3 + n_1n_3 = O(n^2) \quad \text{[illegible]}$$

Further Notes on Chapter 1

→ Loop Unrolling

π.χ.

```
for i=1, i<n, i=i+1
{
    x(i) = α*x(i) + y(i)
}
```

unroll με βήμα 2 ⟹

```
for i=1, i<n, i=i+2
{
    x(i) = α*x(i) + y(i)
    x(i+1) = α*x(i+1) + y(i+1)
}
```

[illegible]

Άσκηση Να υπολογιστεί το $y = y + A^s x$

Απ: Χαζός τρόπος $\Big[\underbrace{\big(\big(\big((A * A) * A\big) * \dots\big) * A}_{s-1 \text{ φορές}}\Big] * x$

↓

πιο έξυπνος τρόπος αντί για $s-1$ πολλαπλασιασμούς, κάνω το εξής:

$$\begin{array}{ccc} A * A & & A * A \\ A^4 & & A^4 \\ & A^{16} & \end{array}$$

Άρα χρειάζομαι $\log s$ πολλαπλασιασμούς πινάκων

↓

γενικός τρόπος $\big(A * A * \dots * (A * (A * x))\big)$

~~αριθμός~~ διάνυσμα

άρα πολ/μο πίνακα * διάνυσμα

αντί για πίνακα * πίνακα

Φ min το λιγότερο που μπορώ να κάνω είναι να χρησιμοποιήσω όλα τα δεδομένα ακριβώς μία φορά και να γράψω μία φορά στη μνήμη

(υποθέτω άπειρη κρυφή μνήμη k)

Άσκηση: Να υπολογιστεί το $B \leftarrow A + (x y^T)^s$

Απ: Είναι $(x y^T)^s = \underbrace{x\, y^T x}_{\text{αριθμός } \alpha}\, y^T x \cdots y^T x\, y^T$

Άρα $(x y^T)^s = x\,(y^T x)^{s-1}\, y^T = x\, \alpha^{s-1}\, y^T$

Υπολογισμός κόστους ϱ:

$y^T x \rightarrow 2n$ πράξεις

$\alpha^{s-1} \rightarrow \log(s-1)$ πράξεις

$x y^T \rightarrow 2n^2$ πράξεις

$\alpha^{s-1}(x y^T) \rightarrow n$ πράξεις

$A + \ldots \rightarrow n^2$ πράξεις

Πως; γιατί xy^T και όχι $(x a^{s-1})$ / $(x a^{s-1} y^T)$

Άρα $\varrho = 3n^2 + \log(s-1) + 2n = O(n^2)$

Υπολογισμός ϕ_{min}: Είναι όπως και στην προηγούμενη άσκηση

$\phi_{min} = 2n^2 + 2n = O(n^2)$

Νόρμες (Required Knowledge)

→ Πινάκων ($\mathbb{R}^{n \times n}$)

- $\|A\|_1 = \max_{j=1:n} \sum_{i=1}^{n} |\alpha_{ij}|$
- $\|A\|_2 = \sqrt{\lambda_{max}(A^* A)}$ — Πως βρίσκω $\lambda_{max}(A^* A)$
- $\|A\|_\infty = \max_{i=1:n} \sum_{j=1}^{n} |\alpha_{i,j}|$

→ Διανυσμάτων

- $\|x\|_1 = \sum_{i=1} |x_i|$
- $\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$
- $\|x\|_\infty = \max |x_i|$

* → Άλλες Νόρμες

- Frobenius $\|A\|_F = \sqrt{\sum_{i,j}^{n} |\alpha_{ij}|^2}$

Απόλυτο και σχετικό Σφάλμα

Έστω ότι προσεγγίζω ποσότητα x με ποσότητα $\hat{x}$. Τότε:

απόλυτο σφάλμα: $E_{abs}(\hat{x}) = |x - \hat{x}|$

σχετικό σφάλμα: $E_{rel}(\hat{x}) = \dfrac{|x - \hat{x}|}{|x|}$

$\|p_2\|_2 = \sqrt{10^3 \, 10^{-6}} = \sqrt{10^{-3}}$ ✗

Αν $R^{1000.000}$ τότε $= 1$

Σύστημα Α.Κ.Υ

$y = $ [illegible]

- b: [illegible]
- e: [illegible]
- m: [illegible]

Έστω F το σύνολο όλων των αριθμών που μπορούν να αναπαρασταθούν με τον παραπάνω τρόπο. Το F χαρακτηρίζεται από ιδιότητες:

i) Πεπερασμένο πλήθος αριθμών, όχι συνεχές
ii) Οι Α.Κ.Υ είναι ρητοί εκ κατασκευής
iii) Οι Α.Κ.Υ δεν είναι ισαπέχοντες

- M: ο Α.Κ.Υ [illegible]
- μ: ο Α.Κ.Υ [illegible]

Απεικόνιση $fl: \mathbb{R} \to F$

Έστω $G \supset F$ τέτοιο ώστε $G: \{x \in \mathbb{R} : \mu \leq |x| \leq M\} \cup \{0\}$

Έστω πως η συνάρτηση απεικόνισης $fl(x) : \mathbb{R} \to F$ γίνεται με κάποιον τρόπο

Διακρίνω τις περιπτώσεις:

- $x \in F$ τότε $fl(x) = x$
- $x \in G$ και $x \notin F$ τότε $fl(x) \neq x$. Άρα θα χρειαστεί κάποια μέθοδος στρογγυλοποίησης και θα υπάρχει κάποιο σφάλμα!
- $x \notin G$ τότε ή $|x| > M$ άρα έχουμε υπερχείλιση
  ή $|x| < M$ άρα έχουμε υποχείλιση

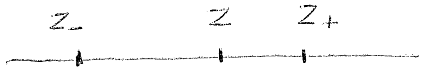
Για αυτές τις περιπτώσεις χρησιμοποιούμε ειδικά σύμβολα

Κανονικοποίηση και Hidden Bit

Αναγκάζουμε τον αριθμό $fl(x) \in F$ να έχει πάντα το πρώτο bit της mantissa ίσο με 1. Αυτή η παραδοχή μας επιτρέπει να υπονοούμε πάντα το bit αυτό (hidden bit), και έτσι να κερδίζουμε άλλο 1 bit για άλλες χρήσεις.

Μονάδα στρογγύλευσης u !

Έστω αριθμός $z \in G$ και $z_+, z_- \in F$ οι κοντινότεροι στον z αριθμοί που ανήκουν στο F.



Όμως αφού πρέπει z_-, z_+ διαδοχικοί
και άρα $z_- = m * \beta^{e-t}$ και $z_+ = (m+1) * \beta^{e-t}$

πως προκύπτει γιατί $u = \frac{\beta^{1-t}}{2}$

Άρα στη χειρότερη περίπτωση θα είναι $|z - fl(z)| = \frac{\beta^{e-t}}{2}$

Άρα $\epsilon abs_{max} = \left|\frac{z - fl(z)}{z}\right| \leq \frac{\beta^{1-t}}{2} = u$? Για $\beta = 2$, $u = 2^{-t}$

! Για $z \in G$, είναι $fl(z) = z(1+\delta)$ για $|\delta| \leq u$

eps

Ο ~~μικρότερος~~ μεγαλύτερος αριθμός τέτοιος ώστε $1 \mp eps = 1$

ή αντίστοιχα ο ελάχιστος αριθμός τ.ω. $1 + eps \simeq 1$

eps: απόσταση του 1 από αμέσως επόμενο αρ.

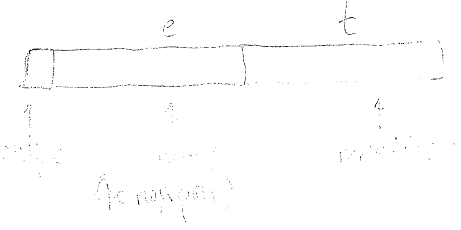
Πρότυπο Α.Κ.Υ της IEEE 754

βάση ψηφία mantissa εύρος εκθέτη

Μονή ακρίβεια $F(2, 24, -125, 128)$

Διπλή ακρίβεια $F(2, 53, -1021, 1024)$

→ ΣΤΗΝ ΠΡΑΞΗ



❗ Κανονικά χρησιμοποιούμε 23 και 52 bits αντίστοιχα για την ουρά. Απλά υπονοείται hidden bit $= 1$ πάντα, και κανονικοποιούμε

- Είναι $u_{single} = 2^{-24}$ και $u_{double} = 2^{-53}$

Ειδικά σύμβολα

- $\pm inf$ → εκθέτης 11…1 και mantissa $b_1 = \dots = b_n = 0$
- -0
- $+0$
- NaN → εκθέτης 11…1 και mantissa $\exists$ bit $\neq 0$

Βαθμιαία Υποχείλιση (Υποκανονικοποίηση)

- Δεν κανονικοποιούμε αριθμούς οι οποίοι είναι μικρότεροι από τον ελάχιστο κανονικοποιημένο $\beta^{e_{min}-1}$

→ ΣΤΗΝ ΠΡΑΞΗ

format long;

realmin

0

Πρόσθεση στο IEEE 754

[illegible]

[illegible]

[illegible]

[illegible]

Αν [illegible] 2^{t}, [illegible] !

e και u στο IEEE 754

είδαμε πριν πως $u_{single} = 2^{-t} = 2^{-24}$

$u_{double} = 2^{-t} = 2^{-53}$

το $e = 2u$ Άρα $e_{single} = 2^{-23}$

$e_{double} = 2^{-52}$

! Φυλλάδιο Προσοχής σελ. 59 Βιβλίο

Συνθήκη Ακριβούς Στρογγύλευσης

Αν με $\tilde{\odot}$ συμβολίζεται η υλοποίηση της πράξης $\odot$ δοθέντων $x, y \in F$, τότε ισχύει ότι:

$$x \tilde{\odot} y = fl(x \odot y) \in F$$

Sticky Bit

Βασικοί Κανόνες Διάδοσης Σφάλματος

- $\dfrac{|fl(x \odot y) - (x \odot y)|}{|x \odot y|} \leq u$ για $\odot \in \{+, -, *, /\}$

- $fl(x) = x(1+\delta), \quad |\delta| \leq u$
- $fl(x \odot y) = (x \odot y)(1+\delta), \quad |\delta| \leq u$
- $fl(x \odot y) = \dfrac{(x \odot y)}{1+\delta}, \quad |\delta| \leq u$

για $x, y \in F$

$x, y \in F$
$fl(x \odot y) = (x \odot y)(1+\delta)$
το σφάλμα οφείλεται στην πράξη μεταξύ $x, y \in F$ επειδή εκτελείται σε σύστημα αυτό.

Σφάλμα Πρόσθεσης — "Πρωτόγονη" Μελέτη

Έστω η πράξη $x + y + z$

Ξεκινάμε την πράξη παρακολουθώντας σταδιακά το σφάλμα

$$x \tilde{+} y \tilde{+} z = \left[(x+y)(1+\delta_1) + z\right](1+\delta_2)$$

Άρα $\left|(x \tilde{+} y \tilde{+} z) - (x+y+z)\right| = \left|x(\delta_1+\delta_2+\delta_1\delta_2) + y(\delta_1+\delta_2+\delta_1\delta_2) + z\,\delta_2\right|$

Από το θεώρημα πριν έχουμε $|\delta_i| \le u$

Άρα $\le |x|(2u+u^2) + |y|(2u+u^2) + |z|u$

! $u^2 < u$ αφού $u < 1$

Αυτή η διαδικασία είναι σωστή αλλά επίπονη.
Εφαρμόσαμε προς τα εμπρός ανάλυση σφάλματος!

Ανάλυση Σφάλματος "Προς τα Εμπρός"

Λήμμα 1 για $|\delta_i| \le u$, $\rho_i = \pm 1$, $nu < 1$ ισχύει $\prod_{i=1}^{n}(1+\delta_i)^{\rho_i} = (1+\delta_1)(1+\delta_2)\dots(1+\delta_n) = 1 + \theta_n$

όπου $|\theta_n| \le \dfrac{nu}{1-nu} = \gamma_n$

Λήμμα 2 Ισχύει $\gamma_n = \dfrac{nu}{1-nu} \le nu(1 + nu + (nu)^2 + \dots) = nu + O(u^2)$

Λήμμα 3 $\prod_{i=1}^{n}(1+\delta_i) \le (1+u)^n < e^{nu}$

$$\big((x\cdot y)(1+\delta_1)\cdot z\big)(1+\delta_2)$$
~~x~~ $x\cdot y(1+\delta_1)(1+\delta_2)+z(1+\delta_2)$ που το $(1+\delta_5)$;

Συμβολισμοί

$\underline{x}$: τα πραγματικά δεδομένα $\in \mathbb{R}$

$\hat{x}$ ή x^* : τα στοιχεία εισόδου που τελικά χρησιμοποιούμε $\in F$
συνήθως $x^* = fl(x)$

$f(x)$: η πραγματική συνάρτηση υπολογισμού, δεν έχει σφάλματα

f_{prog} : η υλοποίηση της f στον υπολογιστή

$f_{prog} = fl$;
↳ χωρίς άπειρη ακρίβεια;
με ~~πράξεις~~ από αριθμούς ;

$$\text{Άρα τελικά το σχετικό σφάλμα} = \frac{|f_{prog}(x^*) - f(x)|}{|f(x)|}$$

Παράδειγμα Έστω η πρόσθεση $x \tilde{+} y = (x+y)(1+\delta)$

$$= x(1+\delta) + y(1+\delta)$$

$$= \hat{x} + \hat{y}$$

Αυτό όμως σημαίνει πως:

$$f_{prog}(x) = f(x_{prog})$$

$f_{prog}(x)$: σφάλμα στις πράξεις

$f(x_{prog})$: σφάλμα στα δεδομένα

Όμως είναι $x_{prog} = x + \Delta x$

Άρα $|f_{prog}(x) - f(x)| = |f(x_{prog}) - f(x)| =$

$$= f(x+\Delta x) - f(x) \overset{Taylor}{=} \frac{f^{(1)}(x)}{1!}\Delta x + \frac{f^{(2)}(x)}{2!}(\Delta x)^2 + \dots$$

$$\Longleftrightarrow \frac{\|f(x_{prog}) - f(x)\|}{\|f(x)\|} \le \underbrace{\frac{\|f'(x)\|\,\|x\|}{\|f(x)\|}} \; \underbrace{\frac{\|x_{prog} - x\|}{\|x\|}}$$

σχ. προς σφάλμα

$$\frac{\|f_{prog}(x) - f(x)\|}{\|f(x)\|} \; ;$$

Αριθμητικά ευσταθής Αλγόριθμος

αν για κάθε x υπάρχει x^* τ.ω

x^* κοντά στο x και $f(x^*)$ κοντά στο $f_{prog}(x)$

$f(x_{prog}) = f(x^*)$ ΚΑΙ

Πίσω ευσταθής Αλγόριθμος

ισχυρά, αν

$\exists$ x^* κοντά στο x ΚΑΙ $f(x^*) = f_{prog}(x)$

Remember από συνθήκη ακρίβειας στρογγύλευσης;

Ανάλυση σφάλματος "προς τα πίσω"

Αντί για να μετρήσω το $\|f_{prog}(x) - f(x)\|$ μετράω το $\|f(x^*) - f(x)\|$

Υπό ποιες προϋποθέσεις ισχύει? $x, y \in F$. το remember

Όμως το $x^* = x + \Delta x$

· Άρα το πρόβλημα ανάγεται στο πόσο ευαίσθητη είναι η f για μικρές μεταβολές στις εισόδους της (διαταραχές/pertubations)

SOS

ικανοποιώ αν πίσω ευστα. θάς πρέπει: πίσω σφάλμα = σχετικό σφάλμα κάθε εισόδου << από; και $|f_{prog}(x_1 x_2 x_3) - f(x)| = |f(\tilde{x}) - f(x)|$

! Παράδειγμα:

με $\tilde{x}_1 = x_1(1+\delta_1)(1+\delta_2)$ δεν το εκβιάζω να βγει $f_{prog} = f(x_{prog})$.

Κατάσταση κ' Μετρητές Αξιοπιστίας

→ Κατάσταση του Προβλήματος: Η ευαισθησία των αποτελεσμάτων λόγω διαταραχών στα στοιχεία εισόδου

→ Κατάσταση του Αλγορίθμου: Η επίδραση των Α.Κ.Υ. στον υλοποίηση και άρα στον υπολογισμό

→ Δείκτης Κατάστασης Υλοποίησης $cond(f_{prog})$

η ελάχιστη τιμή $cond(f_{prog})$ για την οποία ισχύει

$$\frac{\|x - x_{prog}\|}{\|x\|} \le cond(f_{prog})\, u$$

- δείχνει το πόσο ευάλωτη !
-

Ιακωβιανό μητρώο

$\frac{df}{dx_1}$

$f(x_1, x_2, x_3, x_4) = x_1 x_3 + x_2 x_4 \rightarrow [x_3, x_4, x_1, x_2]$

$$f(x_1, x_2) = \begin{pmatrix} a_{11} x_1 + a_{12} x_2 \\ a_{21} x_1 + a_{22} x_2 \end{pmatrix} \rightarrow \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

Δείκτης κατάστασης μητρώου

[illegible] το $\|A\| \, \|A^{-1}\|$

what's going on here!

πρακτικά αυτό σημαίνει πως τα σφάλματα είναι φραγμένα από μονάδα στρογγύλευσης u!

Ασκ. 3.19, 16
3.9, 17
+ Εφαρμογές

Σφάλματα – Summary – Notes

§1-§2 Δίνεται $A\in R^{n\times n}$, $b\in R^{n}$. Να βρεθεί $x\in R^{n}$ τ.ω $Ax=b$

Θεώρημα Έστω το $Ax=b$. Οι εξής προτάσεις είναι ισοδύναμες

i) Το σύστημα έχει λύση

ii) Η τάξη του $[A, b]$ (επαυξημένο) είναι ίση με του A, δηλ. $\text{rank}([A,b])=\text{rank}(A)$

iii) $\forall\, y\in R^{m}$, αν $y^T A=0$ τότε $y^T b=0$

Πόρισμα Αν $A\in \mathbb{R}^{n\times n}$ αντιστρέψιμο, τότε το $Ax=b$ έχει μοναδική λύση

[illegible]

Κανόνας Cramer Έστω το $x=\begin{bmatrix}\xi_1\\ \vdots\\ \xi_n\end{bmatrix}$ για το οποίο $Ax=b$. Τότε $\xi_i=\dfrac{\det(A(i|b))}{\det(A)}$, $i=1,\dots,n$

↓

ανάγαμε την επίλυση $Ax=b$ σε υπολογισμό οριζουσών → πολλές πράξεις

Κατηγορίες Μεθόδων Επίλυσης

i) Άμεσες: οι γνωστές, χρησιμοποιούνται για την επίλυση πυκνών συστημάτων μεσαίου μεγέθους

ii) Επαναληπτικές: Η λύση προσεγγίζεται. Το όριο της προσέγγισης πρέπει να είναι η ακριβής λύση

Κριτήρια επιλογής μεθόδου επίλυσης

i) Μέγεθος μητρώου: ↑ μέγεθος → ↓ άμεσες (απαγορευτικό κόστος)

ii) Δομή μητρώου: χαρακτηριστικά που μπορούμε να εκμεταλλευτούμε

§3 Τύποι Μητρώων (n×n)

Πυκνό n×n πλήρης αποθήκευση απαιτεί n^2 θέσεις

Πυκνό ειδικής μορφής χαρακτηρίζεται από λιγότερα από n^2 στοιχεία

→ Συμμετρικό ισχύει $A^T = A \Rightarrow a_{ij} = a_{ji}$. Χρειάζονται $n(n+1)/2$ λιγότερα στοιχεία

→ Ερμιτιανό αν έχει μιγαδικά στοιχεία και $A = A^* \Rightarrow a_{ij} = \bar{a}_{ji}$

→ Τριγωνικό i) Άνω τριγωνικό ii) Κάτω τριγωνικό

→ Hessenberg "σχεδόν" τριγωνικό

i) Άνω Hessenberg για $i > j+1$ ισχύει $a_{i,j} = 0$

ii) Κάτω " για $j > i+1$ " "

→ Toeplitz τα στοιχεία που ανήκουν στην ίδια διαγώνιο είναι ίσα $\begin{bmatrix} a & b & c \\ d & a & b \\ e & d & a \end{bmatrix}$ n άρα αποθηκεύονται με $n + n - 1 = 2n-1$ στοιχεία ($n-1$ το a)

→ Hankel τα στοιχεία που ανήκουν στην ίδια αντιδιαγώνιο είναι ίσα

$$\begin{bmatrix} a & b & c & d \\ b & c & d & e \\ c & d & e & f \\ d & e & f & g \end{bmatrix}$$

. Όμοίως χρειαζόμαστε $2n-1$ στοιχεία για αποθήκευση

→ Κυκλοτερές (Circulant) κάθε γραμμή προκύπτει από shift της προηγούμενης

δηλαδή $\begin{bmatrix} a & b & c & d \\ d & a & b & c \\ c & d & a & b \end{bmatrix}$ }shift r }shift r . Χρειαζόμαστε n στοιχεία για αποθήκευση

Έχουν σταθερές διαγωνίους!

→ Vandermonde Έστω $\{f_0, f_1, \ldots, f_{n-1}\}$. Τότε το $V(\{f_0, \ldots, f_{n-1}\})$ έχει στοιχεία τέτοια ώστε $V_{ij} = f_{j-1}^{i-1}$. π.χ $V(f_0, \ldots, f_{n-1}) = \begin{bmatrix} f_0^0 & \cdots & f_{n-1}^0 \\ f_0^1 & \cdots & f_{n-1}^1 \\ f_0^{n-1} & & f_{n-1}^{n-1} \end{bmatrix}$

Αραιά αν έχουν λίγα μη μηδενικά στοιχεία. Ο αριθμός των μη μηδενικών στοιχείων συμβολίζεται με nnz. Συνήθως αν nnz = O(n), θεωρούμε το μητρώο αραιό.

Αραιά Δομημένα

→ Ταινιακά $\exists\, m \ll n$ έτσι ώστε $a_{ij} = 0$ όταν $|i-j| > m$

π.χ. στα τριδιαγώνια $|i-j| > 2 \to a_{ij} = 0$

→ Διαγώνια κατά μπλοκ

Ορθογώνιο Αν $A^T A = A A^T = I$

Κανονικό Αν $A^T A = A A^T$

Διαγώνια κυρίαρχο κατά γραμμές αν $|a_{ii}| \geq \sum_{j=1 \text{ και } j\neq i}^{n} |a_{ij}|$, $i = 1,\dots,n$

κατά στήλες ανάλογα

Θετικά ορισμένο αν $x^T A x > 0$, $\forall x \neq 0 \in \mathbb{R}^n$

→ **Στοιχειώδη Μητρώα**

το $E(u, v; \tau) = I - \tau u v^T$, με $\tau \in \mathbb{R}$, $u, v \in \mathbb{R}^n$

→ **Στοιχειώδη Τριγωνικά μητρώα**

είναι τα $L_k(u) = E(u, e_k; 1)$ τ.ω. $u = [0, \dots 0, n_{k+1}, \dots, n_n]$

και $e_k = [0, \dots, 0, 1, 0, \dots 0]$ (θέση k)

τότε $\mathbf{E}(u, e_k; 1)$ είναι στοιχειώδης μετασχηματισμός Gauss ‼

$$\underbrace{\begin{array}{c|c} & k \\ \left[\begin{matrix} 1 & & & & & \\ & 1 & & & & \\ & & \ddots & & & \\ & & & 1 & & \\ & & & -n_{k+1} & \ddots & \\ & & & -n_{k+2} & & \\ & & & -n_{k+2} & & 1 \end{matrix}\right] & k \end{array}}_{E(u, e_k; 1)} = \underbrace{\begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}}_{I} - \underbrace{\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ n_{k+1} \\ \vdots \\ n_n \end{bmatrix} [0 \dots 0\ 1\ 0 \dots 0]}_{u_k \cdot e_k^T}$$

→ Στοιχειώδη (Αντιμεταθετικά) Μητρώα

βλ. βιβλίο σελ. 163

→ Υλοποίηση πράξεων με στοιχειώδη μητρώα

Για να αποθηκεύσουμε το $E(u,v;\tau)$ χρειάζονται $n+n+1 = 2n+1$ στοιχεία

- Κόστος Πολλαπλασιασμού

$$E(u;v;\tau)\,x = (I-\tau u v^T)x = \underbrace{x - \tau u \overbrace{\underbrace{v^T x}_{dot}}^{\text{βαθμωτό}}}_{sAXPY}$$

Άρα χρειάζονται $\underline{O} = 4n$ πράξεις αντί για $\underline{O} = n(2n-1)$ του κανονικού

→ Σφάλμα πράξεων με στοιχειώδεις μετασχηματισμούς

? σελ 165

5.4 Επίλυση γραμμικών πυκνών συστημάτων

Αποθήκευση πυκνών Μητρώων

Στον προγραμματισμό τα $R^{n\times n}$ πίνακες αποθηκεύονται σαν μονοδιάστατοι ως εξής

i) Κατά στήλες $A=\begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix}$ → Μνήμη: | — a_1 — | — a_2 — | ... | — a_n —

π.χ. Fortran

ii) Κατά γραμμές $A=\begin{bmatrix} — a_1 — \\ — a_2 — \\ \vdots \\ — a_n — \end{bmatrix}$ → Μνήμη | — a_1 → | — a_2 → | ... | — a_n —

Τότε ορίζουμε ως βήμα πρόσβασης την απόσταση στη μνήμη των διαδοχικών διευθύνσεων που θέλουμε στο πρόγραμμα μας να προσπελάσουμε. Έτσι π.χ.

- Αν θέλουμε να προσπελάσουμε τα $A(1,1)$, $A(1,2)$, ..., $A(1,n)$ όπου $A\in R^{n\times n}$
  - στη C έχουμε βήμα διασκελισμού 1
  - στη Fortran έχουμε βήμα διασκελισμού n

Συστήματα με τριγωνικά Μητρώα

π.χ. αν ο A είναι κάτω τριγωνικός τότε

$$Ax=b \iff \begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \end{bmatrix}$$

τότε $a_{11}\xi_1 = b_1 \iff \xi_1 = \dfrac{b_1}{a_{11}}$

$a_{21}\xi_1 + a_{22}\xi_2 = b_2 \iff \xi_2 = \dfrac{b_2 - a_{21}\xi_1}{a_{22}}$

$a_{31}\xi_1 + a_{32}\xi_2 + a_{33}\xi_3 = b_3 \iff \xi_3 = \dfrac{b_3 - a_{31}\xi_1 - a_{32}\xi_2}{a_{33}}$

Γενικά $\xi_j = \left(b_j - \sum_{k=1}^{j-1} a_{jk}\xi_k\right)/a_{jj}$

$= \left(b_j - DOT(a_{j,1:j-1}, \xi_{1:j-1})\right)/a_{jj}$ για $j=1,\dots,n$

Κόστος: Άρα κοστίζει $\sum_{j=1}^{n}(1+(2j-2)) = n^2$ πράξεις. Αντίστοιχα κατά γραμμές

Μεταφ: Πρέπει να μεταφερθούν από τη μνήμη ο A και το b και να γραφτεί το x. Άρα $\phi = \dfrac{n(n+1)}{2} + 2n$

μ: Άρα $\mu_{min} = \dfrac{1}{2} + \dfrac{5}{2n}$

! Είναι πίσω σταθερό → βλ. σελ. 170-171

5.5 Γενικά Συστήματα με LU

Έστω ότι θέλουμε να βρούμε x στο $Ax=b$

i) Υπολογίζω τους L, U στο $LU = A$

ii) Λύνω την $Lz=b$ ως προς z

iii) Λύνω την $Ux=z$ ως προς x

2^{η} αναφορά ... $\approx \frac{2n^3}{3}$

Λύση τριγωνικού ... $\sim 2n^2$

Συνολικό κόστος $O(n^3)+O(n^2)+O(n)$

όπου ο κυρίαρχος όρος αντιστοιχεί στην τριγωνοποίηση! (i)

! Αρκεί να την κάνουμε 1 φορά την τριγωνοποίηση, οπότε αν έχουμε λύσεις στην $Ax=b$ για πολλά b, το κόστος αναλογικά μειώνεται

5 φορές → κόστος/5

→ Παραγοντοποίηση Schur

Αν ο $A=\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ και ο A_{11} αντιστρέψιμος, τότε

$$A=\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} I & 0 \\ L_{21} & I \end{pmatrix}\begin{pmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{pmatrix}$$

όπου:

$$U_{11} = A_{11}$$
$$U_{12} = A_{12}$$
$$L_{21} = A_{21}A_{11}^{-1} = L_{21}$$
$$U_{22} = A_{22} - A_{21}A_{11}^{-1}A_{12}$$

συμπλήρωμα Schur ως προς A_{11}

5.7 Οδήγηση στη LU

Μπορεί στη LU να συναντήσουμε οδηγούς 0 ή →0 στρογγυλεμένους, μπορεί να διαιρέσω ή να έχω ενίσχυση του σφάλματος. Για αυτό χρησιμοποιώ οδήγηση

→ Μερική Οδήγηση

το βήμα k, Διαλέγω ως οδηγό το μέγιστο σε απόλυτη τιμή στοιχείο, που βρίσκεται στην ίδια στήλη $A(k{:}n, k)$ και μεταθέτω γραμμές → k συγκρίσεις /βήμα

→ Πλήρης Οδήγηση

το βήμα k, Διαλέγω ως οδηγό το μέγιστο του πίνακα $A(k{:}n, k{:}n)$, μεταθέτω γραμμές και στήλες

Άρα χρειάζονται k^2 συγκρίσεις / βήμα

Άρα συνολικά η πλήρης επιβαρύνεται με $O(n^3)$ συγκρίσεις και
η μερική " με $O(n^2)$ "

και πλέον η παραγοντοποίηση είναι της μορφής $PA = LU$

→ Συντελεστής Αύξησης (δείκτης αποτελεσματικότητας της οδήγησης)

Αν $\alpha_k = \max\limits_{i,j \geq k+1} |\alpha_{ij}|$ τότε συντελεστής αύξησης $p_n = \dfrac{\max\{\alpha_0, \alpha_1, \ldots, \alpha_{n-1}\}}{\alpha_0}$

για τη μερική οδήγηση p^{MO} και για την πλήρη $p^{ΠΟ}$

→ Σφάλμα → αν υπολογίσω $A \overset{+E}{=} \hat{L}\hat{U}$ στοιχείο, η ποιότητα του αποτελέσματος καθορίζεται από τα $\|\hat{L}\|, \|\hat{U}\|$

δείκτης $K_{Skeel}(A) = \| |A^{-1}||A| \|_\infty$

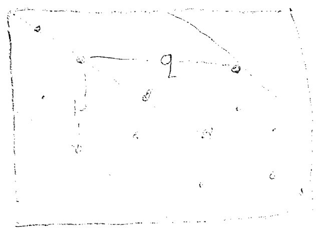
που σφάλμα $e_n = \dfrac{\|r\|}{\|A\|\|\hat{x}\| + \|b\|}$ όπου $r = b - A\hat{x}$
$\hat{x}$ η υπολογισμένη λύση, εάν η λύση του
$(A+\Delta A)\hat{x} = (b+\Delta b)$

(διαταραγμένο σ.κ.λ.)

→ Πολυπλοκότητα το $Ax=b$

Μικρότερη ~~τάξη~~ του n^3 ..

7.1 Μητρώα Ζώνης



Ένα μητρώο ζώνης με άνω εύρος ζώνης q, κάτω εύρος ζώνης p και συνολικό εύρος ζώνης $p+q+1$ είναι ισχύει, $a_{ij}=0$

- Αν $p=q=1$ → τριδιαγώνιο
- Αν $p=1$ $q=0$ → κάτω διδιαγώνιο ($p=0$ $q=1$ άνω)
- Αν $p=1$ → Hessenberg άνω ($q=1$ κάτω Hessenberg)
- Αν $p=0$ → άνω τριγωνικό ($q=0$ άνω τριγωνικό)

Συμβολισμός $A(p|q)$

! Κάθε μητρώο ζώνης $A(p|q)$ έχει το πολύ $nnz = n + (p+q)n - \frac{p(p+1)+q(q+1)}{2}$ μη μηδενικά στοιχεία

↓

Άρα για αποθήκευση χρειάζονται $(p+q+1)$ στήλες n στοιχείων

! Είναι συνήθως αραιά (αφού έχουν $nnz=O(n)$)

↳ Το αντίστροφο αραιού δεν είναι αραιό

Θ1 Έστω $A(p|q) \in \mathbb{R}^{n\times n}$, Αν επιδέχεται παραγοντοποίηση LU τότε $A(p|q)=L(p|0)U(0|$

Θ2 Έστω $A \in \mathbb{R}^{n\times n}$ και $A=A(p|q)$. Χρησιμοποιείται απαλοιφή Gauss με μερική οδήγηση για τον υπολογισμό της $PA=LU$. Τότε:

i) Το U έχει ημιεύρος $p+q$

ii) Το L έχει κατά μήκος $p+1$ μη μηδενικά στοιχεία ανά στήλη

ημιεύρος;

8.1 Διαφορικές Εξισώσεις

Είναι σχέσεις μίας ή περισσοτέρων συναρτήσεων (πιθανών πολλών μεταβλητών) και των παραγώγων τους. Συμβολίζω $\mathcal{L}\left(u(z), \frac{du}{dz_1}(u), \dots, \frac{du}{dz_s}(u), \dots, \frac{d^p u}{dz_1^p}(u), \dots, \frac{d^p u}{dz_s^p}(u)\right) = 0$

για $z \in \Omega \subseteq \mathbb{R}^s$

- Τάξη διαφορικής εξίσωσης η μέγιστη τιμή του p (~~τάξη~~ τάξη της μεγαλύτερης παραγώγου)
- Γραμμική Δ.Ε. αν ο τελεστής $\mathcal{L}$ είναι γραμμικός προς τα ορίσματά του

  Μη γραμμική Δ.Ε αλλιώς

  Ημιγραμμική Δ.Ε αν ο $\mathcal{L}$ είναι γραμμικός μόνο προς τα ορίσματα μέγιστης τάξης

8.2 Διακριτοποίηση κ' Πεπερασμένες Διαφορές

Προσπαθούμε να προσεγγίσουμε τις παραγώγους της u με γραμ. συνδυασμούς των τιμών της σε προκαθορισμένα σημεία του Π.Ο. της. → αντικαθιστώ το Π.Ο. με πλέγμα κόμβων

Τότε η Δ.Ε ανάγεται σε σύστημα εξισώσεων (γραμ. ή όχι)

Προκύπτουν ερωτήματα
i) Πως φτιάχνω το πλέγμα που διακριτοποιεί το Π.Ο.
ii) Πως προσεγγίζω τις παραγώγους
iii) Πως λύνω τις προκύπτουσες εξισώσεις
iv) Πόσο απέχει η λύση των εξισώσεων από την πραγματική λύση της Δ.Ε.

Πεπερασμένες Διαφορές

Προσεγγίζουμε την $u'(x)$ με $\frac{u(x+h)-u(x)}{h}$

Προφανώς αυτό εξαρτάται από το h, αν $h \to 0$ τότε $\frac{u(x+h)-u(x)}{h} \to u'(x)$

Σ.Δ.Ε (συνήθεις) μόνο μία ανεξάρτητη μεταβλητή

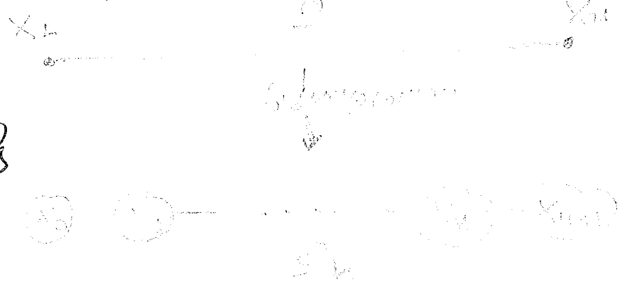
Μ.Δ.Ε (μερικές) >1 ανεξάρτητες μεταβλητές

8.2.1 Επίλυση ΣΔΕ με πεπερασμένες διαφορές

Έστω η $-\frac{d^2u}{dx^2}(x) + b(x)\frac{du}{dx}(x) + c(x)u(x) = d(x)$ αλλιώς γράφεται $L(u,b,c,d,x)=0$

Είναι 2ης τάξης και η $u(x)$ έχει Π.Ο. το $X_L \le x \le X_u$

Έστω ότι ξέρω τις $u(X_L)$, $u(X_u)$ και η u αρκετές φορές παραγωγίσιμη



i) Διακριτοποίηση Χωρίου

Έστω $\Omega_h = \{x_j \mid x_j = X_L + jh \in \Omega_h,\ j=0,\dots,n+1\}$

όπου $h = \frac{X_u - X_L}{n+1}$

- τα $x_0 = X_L$, $x_{n+1} = X_u$ καλούνται ακραία σημεία
- Άρα κατασκευάζω ένα πλέγμα $n+2$ κόμβων στο $\Omega = [X_L, X_u]$

ii) Διακριτοποίηση Δ.Ε.

! Αντιστοιχώ $u(x_j) = u_j$

Χρησιμοποιώ το Θ. Taylor. Έτσι:

$$u_{j+1} = u_j + h u_j^{(1)} + \frac{h^2 u_j^{(2)}}{2!} + \frac{h^3 u_j^{(3)}}{3!} + \dots$$

$$u_{j-1} = \quad - \quad + \quad - \quad + \dots$$

Από αυτές τις εκφράσεις μπορώ να προσεγγίσω τις παραγώγους:

π.χ. $u_{j+1} - u_{j-1} = 2h u_j^{(1)} + \frac{h^3}{3!}(\quad) \dots$

Για να είναι το σφάλμα διακριτοποίησης 2ης τάξης πρέπει η 4η παράγωγος να είναι ομαλής φραγμένη

και προκύπτει ότι

$$\frac{u_{j+1} - u_{j-1}}{2h} = u_j^{(1)} + O(h^2)$$

[illegible]

$$\frac{u_{j-1} + u_{j+1} - 2u_j}{h^2} = u_j^{(2)} + O(h^2)$$

! Η ποιότητα τις προσεγγίσεις εξαρτάται από:

i) το μέγεθος του h

ii) τις παραγώγους μεγαλύτερης τάξης γύρω από το x_j

Πλέον μπορώ να γράψω την $-\frac{d^2u}{dx^2}(x)+b(x)\frac{du}{dx}(x)+c(x)u(x)=d(x)$

Για $x=x_1$: $\frac{-v_0-v_2+2v_1}{h^2}+b_1\frac{u_2-v_0}{2h}+c_1v_1=d_1$

$\vdots$

Για $x=x_j$: $\frac{-v_{j-1}-v_{j+1}+2v_j}{h^2}+b_j\frac{u_{j+1}-v_{j-1}}{2h}+c_jv_j=d_j$, $j=2,\dots,n-1$

$\vdots$

Για $x=x_n$: $\frac{-v_{n-1}-v_{n+1}+2v_n}{h^2}+b_n\frac{v_{n+1}-v_{n-1}}{2h}+c_nv_n=d_n$

Στο σύστημα εξισώσεων. Το γράφουμε σε μορφή που μας βολεύει:

$\frac{-v_2+2v_1}{h^2}+b_1\frac{u_2}{2h}+c_1u_1=d_1+\frac{u_0}{h^2}+b_1\frac{u_0}{2h}$

$\frac{-v_{j-1}+v_{j+1}+2v_j}{h^2}+b_j\frac{u_{j+1}-v_{j-1}}{2h}+c_jv_j=d_j$ για $j=2,\dots,n-2$

$\frac{-u_{n-1}+2v_n}{h^2}+b_n\frac{-v_{n-1}}{2h}+c_nv_n=d_n+\frac{v_{n+1}}{h^2}+b_1\frac{v_0}{2h}$

συνεχίζω

$\left(c_1+\frac{2}{h^2}\right)v_1+\left(-\frac{1}{h^2}+\frac{b_1}{2h}\right)u_2=d_1+\dots$

$-\left(\frac{1}{h^2}+\frac{b_j}{2h}\right)u_{j-1}+\left(c_j+\frac{2}{h^2}\right)u_j+\left(-\frac{1}{h^2}+\frac{b_j}{2h}\right)u_{j+1}=d_j$

$\left(-\frac{1}{h^2}-\frac{b_n}{2h}\right)u_{n-1}+\left(c_n+\frac{2}{h^2}\right)u_n=d_n+\dots$

Αυτός ο A ισχύει μόνο για αυτή την εξίσωση.

δηλαδή έχουμε ένα σύστημα της μορφής

$$\begin{array}{ccc} A & U & F \\ \begin{bmatrix} \alpha_j & \beta_j & & \\ \gamma_j & \alpha_j & \beta_j & \\ & \gamma_j & \alpha_j & \beta_j \\ & & \ddots & \\ & & \gamma_j & \alpha_j \end{bmatrix} & \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = & \begin{bmatrix} d_1+\dots \\ d_2 \\ \vdots \\ d_n+\dots \end{bmatrix} \end{array}$$

όπου $\alpha_j=\left(\frac{2}{h^2}+c_j\right)$, $\gamma_j=\left(-\frac{1}{h^2}-\frac{b_j}{2h}\right)$, $\beta_j=\left(-\frac{1}{h^2}+\frac{b_j}{2h}\right)$

και $A=\mathrm{tridiag}[\gamma_j,\alpha_j,\beta_j]$

Ισχύει [illegible] LD [illegible]

Αποδείξαμε ότι [illegible] A [illegible] >0 και $R \to$ S.D.D $\to$ Cholesky

ΜΔΕ σε 2-D

βλ. σελ. 104, διαφάνειες

8.2.2 Επίλυση προβλήματος αρχικών τιμών

Έστω μια Δ.Ε 1ης τάξης, η $\frac{du}{dt} = f(t, u(t))$, με $u(t_0) = c$

$\forall t \in I = [0, T]$

όπου $t \in I$ διάστημα στο $\mathbb{R}$

$f : I \times \mathbb{R}^n \to \mathbb{R}^n$ γνωστή συνάρτηση

c γνωστή σταθερά

! η τιμή της παραγώγου $\frac{du}{dt}$ εξαρτάται από το $u(t)$ → δεν μπορώ να ολοκληρώσω

→ **Μέθοδος Euler**

Έστω πως $t_0 = 0$ και γνωρίζω την $u(t_0) = u(0)$

Τότε θα γνωρίζω και την $f(0, u(0)) = u'(0)$

Όμως $u'(t) = \frac{u(t+h) - u(t)}{h}$

1ο Βήμα Για $t=0$, → $u(h) \simeq U_1 = u_0 + h f(0, u(0))$

Προσέγγισα το $u(h)$ με το U_1

2ο Βήμα Ισχύει $u(2h) \simeq U_2 = u(h) + h f(h, u(h))$

Όμως $u(h) \simeq U_1$ — γνωστή, υπολογίσιμη

Άρα $u(2h) \simeq U_2 = U_1 + h f(h, U_1)$

Έτσι μπορώ να υπολογίσω το $u(h), u(2h), \dots, u(nh)$ ως το $T = nh$

! Μέθοδος μορφής $U_{n+1} = U_n + h_n f_n(t_n, U_n)$

Μονοβηματικές σε κάθε βήμα προσεγγίζω την νέα τιμή U_{n+1} της άγνωστης μεταβλητής, χρησιμοποιώντας μόνο την αμέσως προηγούμενη τιμή → U_n

Μετασχηματισμός γραμμικής Σ.Δ.Ε n-τάξης σε ΣΔΕ 1-τάξης

Έστω η Σ.Δ.Ε n-τάξης

$$u^{(n)} = \alpha_0(t)u(t) + \dots + \alpha_{n-1}u^{(n-1)}(t) + d(t)$$

και $u(0) = \gamma_1, \dots, u^{(n-1)} = \gamma_n$

i) Ορίζω καινούργιες μεταβλητές $\{u_1, \dots, u_n\}$ ως:

$$u_1 = u$$
$$u_2 = u^{(1)}$$
$$\vdots$$
$$u_n = u^{(n-1)}$$

ii) Τότε ισχύει:

$$\frac{d}{dt}u_1(t) = u_2(t)$$
$$\frac{d\,u_2(t)}{dt} = u_3(t)$$
$$\vdots$$
$$\frac{d}{dt}u_n(t) = \alpha_0 u_1(t) + \dots + \alpha_{n-1}u_n(t) + d(t)$$

$\rightarrow \frac{d}{dt}u(t) = A\,u(t) + c(t)$

με $u(t) = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}$ $u(0) = \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{bmatrix}$

$c(t) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ d(t) \end{bmatrix}$

και θα είναι $A = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & \dots & 0 \\ & & & \vdots & & & \\ 0 & 0 & \dots & & \dots & & 1 \\ \alpha_0(t) & \alpha_1(t) & \alpha_2(t) & \dots & & & \alpha_{n-1}(t) \end{bmatrix}$

(34)

8.2.3 Μέθοδοι Euler

Εμπρός Euler

Διακριτοποιώ την παράγωγο της u στο t ως $\frac{d}{dt}u(t) = \frac{u(t+\Delta t) - u(t)}{\Delta t} + O(\Delta t)$

και καταλήγω γενικά σε $u(t+\Delta t) = u(t) + \Delta t * f(t, u(t))$

αν έχω αρχική τιμή $u(t_0) = c$.

→ Κόστος

Σε κάθε βήμα πρέπει να i) υπολογίσω την $f(t_s, u_s)$ ii) υπολογίσω την $u_s + \Delta t * f(t_s, u_s)$

↳ το μεγαλύτερο κόστος είναι ο υπολογισμός της f που μπορεί να είναι δύσκολος

→ Σύγκλιση

Συγκλίνει, βλ. σελ. 292 σημ.

πρέπει να ισχύει $\lim_{h \to 0} \max_j \|e_j(h)\| = 0$ όπου $e_j = u(t_j) - u_j$

→ Τάξη σύγκλισης

↳ p αν $\|u(t_j) - u_j(h)\| = O(h^p)$

Αρκεί όμως για να έχω τ.σ. p, πρέπει να έχω συνεχή παράγωγο μέχρι το $u^{(p+2)}$

→ Ευστάθεια

πρέπει $|1 + h\lambda(A)| < 1$ $\forall \lambda(A)$

Απόλυτη ευστάθεια Μαρος

φασματική $I + hA < 1 \Rightarrow$

$|1 + h \cdot \lambda(A)| < 1$

$\lambda(A) = \det(A)$

Πίσω Euler

Διακριτοποιώ την παράγωγο της u στο t ως $\frac{d}{dt}u(t) = \frac{u(t)-u(t-\Delta t)}{\Delta t} + O(\Delta t)$

Άρα γενικά καταλήγω $u(t-\Delta t) = u(t) - \Delta t * f(t,u(t))$

αν έχω αρχική τιμή $u(t_0)=c$

→ Κόστος

Σε κάθε βήμα πρέπει να λύνω ένα σύστημα

→ Ευστάθεια

Απόλυτα ευσταθής για κάθε $h>0$

Άρα το λάθος περιορίζεται μόνο από το σφάλμα διακριτοποίησης

→ Κόστος

σελ-296 book / 1/4 διαφάνειες

Σύγκριση Εμπρός/Πίσω Euler

Εμπρός $U(t+\Delta t) = (I + A\Delta t)\, u(t)$

$U(t+\Delta t) = (I + A\Delta t)^{-1} u(t)$