Επιστημονικός 1ο Πακέτο (1)

Μετονεκτήματα Μοντέλων κόστους

Μειονεκτήματα μοντέλων κόστους δεν έχει τρόπο να αποτιμήσει τις επιδόσεις που επιτυγχάνονται κάνοντας χρήση: α) της ιεραρχικής διαμόρφωσης της μνήμης για την επίτευξη μεγαλύτερης επίδοσης β) της παραλληλίας που μπορεί να διαθέτει το σύστημα επεξεργασίας γ) της πολύ-ταχύτητας μείωσης του κόστους των στοιχειωδών πράξεων σε σύγκριση με τη μείωση του κόστους των μεταφορών μεταξύ μνήμης και επεξεργαστή.

- Μέθοδος Horner: για υπολογισμό τιμής ενός $P(x)=a_0+a_1x+\dots+a_nx^n \Rightarrow s=a_n$, for $i=n-1:-1:0$, $s=s\cdot x+a_i$, end

- Υπολογιστικό μοντέλο με ιεραρχία μνήμης: έχει επεξεργαστή, καταχωρητές, κρυφή-κύρια μνήμη. Χαρακτηριστικά: Ο επεξεργαστής προσπελαύνει την μνήμη με εντολές load/store. Χωρητικότητα κρυφής μνήμης K και κύρια $M \gg K$. Χρόνοι εκτέλεσης των $\pm, \times, /$ είναι $\tau_{αρθ}$. Όταν η load αναφέρεται σε στοιχεία της κύριας μνήμης που πάει στους καταχωρητές θέλει χρόνο $\tau_{μετ}$. Αν το στοιχείο είναι στην κρυφή μνήμη τότε $\tau^{(c)}_{μετ} \ll \tau_{μετ}$. Κατά την έναρξη του προγράμματος τα στοιχεία είναι στη κεντρική μνήμη. Η εκτέλεση τελειώνει όταν όλα τα στοιχεία αποθηκευτούν στην κεντρική μνήμη $\tau^{(c)}_{μετ}=$

- Horner: load x, a_n / $s=a_n$ / for $i=n-1:-1:0$ / load a_i / $s=s\cdot x+a_i$ / end / store s

με $\Omega=2n$ $\quad \Phi=n+3$

Ω: αριθμός πράξεων α.κ.υ (flops)

Φ: αριθμός μεταφορών μεταξύ κύριας μνήμης και καταχωρητών ή κρυφής μνήμης.

Φ_{min}: τον ελάχιστο αριθμό μεταφορών Φ που θα απαιτούσε για την υλοποίηση, αν διαθέταμε απεριόριστη μνήμη σε όλα τα επίπεδα.

- $T=T_{αρθ}+T_{μετ}=\tau_{αρθ}\cdot\Omega+\tau_{μετ}\cdot\Phi$

$T=T_{αρθ}\left(1+\mu\dfrac{\tau_{μετ}}{\tau_{αρθ}}\right) \quad \mu=\dfrac{\Phi}{\Omega}$

Διαφορετικές υλοποιήσεις απαιτούν ίδιο αριθμό πράξεων και διαφέρουν στις μεταφορές.

Προμεταφορά: τρόπος για την μείωση ή απόκρυψη της (χρ- καθυστέρησης που οφείλεται στην μεταφορά στοιχείων από την μνήμη. Αφού επιτρέπεται η σύγχρονη εκτέλεση εντολών μεταφοράς και αριθμητικής επεξεργασίας, είναι δυνατόν να μεταφερθούν στοιχεία, χρήσιμα σε μελλοντική αριθμητική επεξεργασία. Γίνεται α) Μέσω συστήματος: μεταφραστής εισάγει στον κώδικα εντολές για την κλήση στοιχείων που πρόκειται να χρειαστούν μελλοντικά. Η προμεταφορά μπορεί να υλοποιηθεί και από το runtime system. β) Μέσω άμεσων εντολών: Ο χρήστης μετατρέπει τον κώδικα εισάγοντας εντολές που καλούν στοιχεία που πρόκειται να χρειαστούν. Για να πετύχει η προμεταφορά, αν τα στοιχεία πρόκειται να χρειαστούν στον κύκλο T, η εντολή μεταφοράς τους στην κρυφή μνήμη πρέπει να δοθεί στον κύκλο $T-\tau$, όπου τ είναι ο μέγιστος αριθμός κύκλων που απαιτούνται για την μεταφορά των στοιχείων από την μνήμη.

Σ/Λ: Αν στο $(0,1)$ ισχύει $-\frac{d^2u}{dx^2}+b\frac{du}{dx}+cu=x$, $b,c>0$ και $u(0)=0$, $u(1)$ ... τότε το A στο τέλος $A=A^T$.

Λάθος: θα πρέπει για να ισχύει $A=A^T$ το A συμμετρικό. Όμως ξέρω ότι το A είναι τριδιαγώνιο

$$A = trid\left[-\frac{1}{h^2}-\frac{b}{2h},\ \frac{1}{h^2}+c,\ -\frac{1}{h^2}+\frac{b}{h^2}\right]$$

Σ/Λ: Το μέγεθος του βήματος Δt για την λύση της $\frac{du}{dt}=f(t,u)$ με μέθοδο Euler δεν μπορεί να γίνει πολύ μεγάλο λόγω αστάθειας.

Λάθος, γιατί η μέθοδος Euler είναι ευσταθής για οποιοδήποτε μέγεθος του Δt.

Σ/Λ Το Δt της μεθόδου Euler καθορίζεται από το μέγεθος του σφάλματος διακριτοποίησης που μπορούμε να δεχτούμε για την λύση.

Λάθος: Μεγάλο μέγεθος Δt οδηγεί σε αστάθεια.

Η απώλεια πληροφορίας οφείλεται κυρίως: α) Στην μαθηματική μοντελοποίηση και στις απλοποιήσεις που γίνονται χάριν επιλυσιμότητας β) Στη διακριτοποίηση και τα σφάλματα αποκοπής που προκύπταν απο αυτή. π.χ. αντικατάσταση παραγώγου με διαιρεμένες διαφορές γ) Στα σφάλματα στρογγύλευσης λόγω χρήσης αριθμητικής κινητής υποδιαστολής (5

$y = \pm m \times b^{e-t}$

Κανονικοποίηση και κρυμμένο bit: Μπορούμε να αναπαραστήσουμε τον ίδιο αριθμό με περισσότερους τρόπους π.χ. $y = 0.d_1 0000 \times b^{e}$ και $y = 0.0 d_1 0000 \times b^{e+1}$. Έτσι έχουμε έλλειψη μοναδικότητας. Γι'αυτό κανονικοποιούμε το σύστημα και χρειάζομε το 1ο ψηφίο του (στη βάση b) να είναι μη μηδενικό δηλ. $y \in RUF \Rightarrow d_1 \neq 0$. Για $b=2$, $d_1 \neq 0 \Rightarrow d_1 = 1$ επομένως μπορούμε να μην το αποθηκεύουμε, αλλά οι αλγόριθμοι που διαχειρίζονται τους αριθμούς αυτούς πρέπει να το λαμβάνουν υπόψη τους. Αυτή είναι η τεχνική του κρυμμένου bit και χρησιμοποιείται ευρύτατα σήμερα γιατί για δεδομένο μήκος του αριθμού κ.υ. δίνει ένα ακόμη bit χωρίς κόστος.

$y = \pm b^{e} \times .d_1 d_2 \ldots d_t$ όπου $0 \leq d_i \leq b-1$, $d_1 \neq 0$

Το ε της μηχανής, που συμβολίζεται με ε_M, είναι η απόσταση από το 1 ως τον αμέσως μεγαλύτερο α.κ.υ. δηλαδή $\boxed{\varepsilon_M = \delta(1, 1^{+}) = 1^{+} - 1}$. Σε κάθε περίπτωση το ε_M δείχνει την ελάχιστη δυνατή ~~διακριτ~~ διακριτότητα του συστήματος α.κ.υ. Αν το σύστημα χρησιμοποιεί $t-1$ δυαδικά ψηφία μετά την υποδιαστολή τότε το επόμενο του 1 είναι το $1 + 2^{1-t} \Rightarrow \varepsilon_M = 2^{1-t}$

Οι α.κ.υ. δεν είναι ισοκατανεμημένοι στο άξονα επομένως οι αποστάσεις δεν είναι ομοιόμορφες αλλά μεγαλώνουν κατά b.

π.χ. $|m b^{e-t} - (m+1) b^{e-t}| = b^{e-t}$

$|m b^{e+1-t} - (m+1) b^{e+1-t}| = b^{e+1-t}$

- Για έναν αφού z οι πλησιέστεροι που περιέχουν το z είναι μοναδικοί: $z_- < z < z_+$ (6)

$$\left.\begin{aligned} z_- &= m \times b^{e-t} \\ z_+ &= (m+1) \times b^{e-t} \end{aligned}\right\} \quad z_+ - z_- = b^{e-t}$$

- Το μέγιστο σφάλμα για το z είναι:

$$\left|\frac{z - fl(z)}{z}\right| \le \frac{b^{e-t}}{2} \Big/ b^{-1} \cdot b^{e} = \frac{b^{1-t}}{2} = u \text{: μονάδα στρογγύλευσης}$$

- Στόχοι τυποποίησης IEEE: α) Συμβατότητα, αποτελεσματικότητα, μεταφερσιμότητα β) αντιμετώπιση ειδικών περιπτώσεων (π.χ. υπερχείλιση, υποχείλιση). γ) υλοποίηση μαθηματικών συναρτήσεων υψηλής ακρίβειας δ) ακριβής περιγραφή της κωδικοποίησης $m = d_0 + d_1 b^{-1} + \dots + d_{t-1} b^{-(t-1)}$

- Βασικές προδιαγραφές προτύπου IEEE: ① Formats: single, double, single extended, double extended. ② 4 είδη στρογγύλευσης προς πλησιέστερο, προς ζυγό, προς $\pm\infty$, προς 0 (αποκοπή) ③ ειδικοί αριθμοί: Συνδυασμοί από bits χρησιμοποιούνται ως "ειδικοί αριθμοί" για κωδικοποίηση αποτελέσματος πράξεων όπως $x/0$, $0/0$, $\sqrt{y}$ με $y<0$. Υιοθετεί την ύπαρξη αριθμών -0, $\pm\infty$, NaN. ④ Επιτρέπει τη χρήση υποκανονικοποιημένων αριθμών, όταν το αποτέλεσμα πράξης είναι μικρότερο του ελάχιστου κανονικοποιημένου αριθμού ⑤ Αριθμητικές εξαιρέσεις: inexact, overflow, underflow, division by 0.

π.χ. $0/0$, $0 \times \infty$, $\sqrt{-1} \rightarrow$ NaN   overflow $\rightarrow \pm$Inf

finite number/0 $\rightarrow \pm$Inf   underflow $\rightarrow$ subnormal numbers

- FMA επιτρέπει την εκτέλεση $z + x \cdot y$ στον ίδιο περίπου χρόνο με την $x * y$ ή $x + y$. Συνεπάγεται μόνο ένα σφάλμα στρογγύλευσης.

δηλ. $fl(z + x \cdot y) = (z + x \cdot y)(1 + \delta)$ $|\delta| \le u$

παρά $fl(z + xy) = (z + xy(1+\delta_1))(1+\delta_2)$

Ποια σχέση του ϵ_M με την u: Η u είναι το μέγιστο (σχετικό σφάλμα που γίνεται μετά από στρογγύλευση ενός α.κ.υ επομένως εξαρτάται από την στρατηγική στρογγύλευσης αν είναι προς τον πλησιέστερο θα είναι το μισό της μέγιστης σχετικής απόστασης δύο διαδοχικών α.κ.υ. Το ϵ_M είναι η απόσταση από το 1 στον αμέσως μεγαλύτερο α.κ.υ και είναι ανεξάρτητο του ονόματος στρογγύλευσης. Στη περίπτωση στρογγύλευσης προς τον πλησιέστερο τότε $\boxed{\epsilon_M = 2u}$

Σ ή Λ; $\forall x \in \mathbb{R}^n$ ισχύει $\frac{1}{n}\|x\|_1 \le \|x\|_\infty \le \|x\|_1$:

μεγαλύτερο

Σωστό: $\|x\|_\infty = \max_{i=1:n}\{|\xi_i|\}$ και $\|x\|_1 = \sum_i |\xi_i| \le n\|x\|_\infty$ επίσης $\|x\|_1 = \sum_i |\xi_i| \ge \|x\|_\infty$ αφού το $\|x\|_\infty$ εμφανίζεται ως ένας από τους μη αρνητικούς όρους του αθροίσματος για το $\|x\|_1$

A = [100, 1; 0, -10] και B = [1, 0; 0, 0] → C = A./B ?

C = [100, Inf; NaN, -Inf] 1/(1/0) → 0 1/1/0 → Inf max(Inf, 4) → Inf

---

Ν.Δ.Ο. η u είναι α.κ.υ: $u = b^{1-t}/2$. Αν $b=2$ όπως στο IEEE τότε $u = 2^{-t}$ και μπορεί να αναπαρασταθεί. Αν $b>2$ τότε $b^{1-t}/2 = (b/2) \times b^{-t}$ συνεπώς αν b άρτιος, η αναπαράσταση είναι εφικτή.

---

temp2 = realmin/2; temp = 2*temp2; boole(temp==realmin)

boole = 1 λόγω ότι το IEEE υποστηρίζει βαθμιαία υποκανονικοποίηση.

---

x = realmin: y = x/2; if (y==0) y = 2·x, else y = y+1; end

Έχουμε υποκανονικοποίηση άρα το $y = \frac{x}{2}$ δεν δίνει μηδέν εκτελείται το else $y = y+1$ επειδή όμως $y = \underline{\text{realmin}}$ είναι πολύ μικρότερο από το ϵ_M τότε $y = 1$

$$\frac{b^{1-t}}{2} = \frac{b}{b^t}\cdot\frac{1}{2} = \frac{b}{2b^t} = \left(\frac{b}{2}\right) b^{-t}$$

Τι σημαίνει αλγόριθμος πίσω ευσταθής; Αν τρέξουμε (ε-
τον αλγόριθμο τότε το υπολογισμένο αποτέλεσμα f με
τα δεδομένα του προβλήματος (και θα περιέχει σφάλματα)
μπορεί να γίνει ίσο με το θεωρητικό αποτέλεσμα και θ
προέκυπτε αν εκτελούσαμε τις πράξεις με αριθμητική ά-
πειρης ακρίβειας χρησιμοποιώντας στοιχεία εισόδου και μπορ
να είναι λίγο διαφορετικά από τα ακριβή. Η απόσταση τ
παραλλαγμένων δεδομένων από τα ακριβή είναι ένα μέτ-
της πίσω ευστάθειας του αλγορίθμου

Παράδειγμα Blas-3. Πλεονέκτημα σε σχέση με Blas-2

Πολλαπλασιασμός μητρώων. Μεγαλύτερη τοπικότητα, αν αξιοποιηθεί από την υλοποίηση.

Η LU φθηνότερη (ταχύτερη, μικρότερο κόστος: $\frac{2n^3}{3} + O(n^2)$)
ενώ QR πίσω ευσταθής ανεξαρτήτως δεδομένων.

$$\parallel \quad \frac{4n^3}{3} + O(n^2)$$

Έστω $A = \begin{bmatrix} 4 & -8 & 1 \\ -6 & 5 & 7 \\ 0 & -10 & -3 \end{bmatrix}$ α) χωρίς οδήγηση το στοιχείο στη θέση (1,1) δηλ. το 4

β) μερική οδήγηση: το μέγιστο κατά απόλυτη τιμή της 1ης στήλης $|-6|$ άρα οδηγός το -6

γ) πλήρης οδήγηση: το μέγιστο κατά απόλυτη τιμή του πίνακα $|-10|$ άρα οδηγός το -10.

Πότε ένα σύστημα α.κ.υ. ικανοποιεί την συνθήκη ακριβούς στρογγ.

Όταν το αποτέλεσμα οποιασδήποτε στοιχειώδους πράξης στο σύστημα και σε δεδομένα x, y που είναι α.κ.υ. είναι ο α.κ.υ. που θα προέκυπτε αν εκτελούσαμε την πράξη με αριθμητική άπειρης ακρίβειας και μετά στρογγυλεύαμε.

$fl(x \odot y) = x \tilde{\odot} y$, $\tilde{\odot}$ υλοποίηση της $\odot$ στο σύστημα α.κ.υ.

Α=LU υπολογίζει τη L,U από το Α. Άρα $\Phi_{min}=2n^2$ για τα load-store. Ακόμα η LU απαιτεί $\Omega=\frac{2n^3}{3}$. άρα $H_{min}=\frac{2n^2}{\frac{2n^3}{3}}=\frac{6n^2}{2n^3}=\frac{3}{n} \Rightarrow O(\frac{1}{n})$ άρα υπάρχει τοπικότητα. Άρα η LU τετραγωνικού μητρώου παρέχει τη δυνατότητα καλής απόδοσης σε συστήματα με ιεραρχική μνήμη.

$\frac{Inf}{Inf}$ = NaN Σ ή Λ? → Σωστό. δεν ορίζεται

$\frac{realmin}{2}==0$ Σ ή Λ? → Λάθος λόγω βαθμιαίας υποκανονικοποίησης.

$\frac{0}{0}$ ==Inf Σ ή Λ? → Λάθος. δεν ορίζεται

---

realmax+1.0==∞ Σ ή Λ? Λάθος: ο εκθέτης του realmax είναι 1023. Για να προστεθεί το y=1.0 πρέπει να το διαιρέσουμε ώστε να γίνει ο εκθέτης του 1023. Κάθε διαίρεση με το 2 αντιστοιχεί σε ολίσθηση του πιο σημαντικού bit της ουράς του y κατά μια θέση προς τα δεξιά. Το μήκος της ουράς στην IEEE διπλής ακριβείας είναι 52. Άρα μετά από 53 ολισθήσεις η ουρά θα μηδενιστεί. Επομένως realmax+1.0 = =realmax.

---

5 σημαντικά προβλήματα της Υ.Γ.Α. (α) Επίλυση γραμμικών συστημάτων (β) υπολογισμός συναρτήσεων μητρώων (γ) επίλυση προβλήματος ιδιαζουσών τιμών (δ) επίλυση προβλημάτων ελαχίστων τετραγώνων (ε) επίλυση προβλ. ιδιοτιμών γενικευμένων.

Αν $A\in R^{n\times n}$ αντιστρέψιμο το Ax=b έχει μοναδική λύση.
συμμετρικό $A=A^T$ Θετικό ορισμένο: $x^TAx>0$
Toeplitz: τα στοιχεία από διαγώνιο είναι ίσα μεταξύ τους.
ορθογώνιο: $AA^T=A^T A=I$
μεταθετικά P.A → ο Α αλλάζει γραμμές A.P → αλλάζουν στήλες
$P\cdot P^T=I$

Για κάθε αντιστρέψιμο τετραγωνικό μητρώο A υπάρχει παραγοντοποίηση LU. Σ ή Λ? Λάθος. Π.χ. το $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ δεν μπορεί ως LU γιατί τότε $l_{11}u_{11} = 0$ άρα $u_{11} = 0$ οπότε το U είναι μη αντιστρέψιμο κάτι που δε γίνεται γιατί τότε και το LU θα ήταν μη αντιστρέψιμο

Το αντίστροφο ενός τυχαίου αντιστρέψιμου κάτω διδιαγώνιου είναι επίσης κάτω διδιαγώνιο. Σ ή Λ? Λάθος: έστω η 1η στήλη $s = [s_1, s_2, s_3]^T$ του αντιστρόφου του $T = \begin{bmatrix} 1 & 0 & 0 \\ t_{21} & 1 & 0 \\ 0 & t_{32} & 1 \end{bmatrix}$ όπου $t_{21} t_{32} \neq 0$. Ισχύει $Ts = e_1$ και αρκεί να δείξουμε ότι $s_3 \neq 0$. Εκ κατασκευής θα πρέπει να ισχύει $s_1 = 1$, $t_{21} + s_2 = 0$, $t_{32} s_2 + s_3 = 0$. Επομένως αν $s_3 = 0$ τότε $s_2 = 0$. Επομένως $t_{21} = 0$ και είναι αδύνατο.

Σ ή Λ? Το αντίστροφο σ.θ.ο τριδιαγώνιου μητρώου είναι τριδιαγώνιο: Λάθος. Αν είναι σ.θ.ο μπορεί να χρησιμοποιηθεί Cholesky, $A = LL^T$ όπου L κάτω τριγωνικό. Το A είναι τριδιαγώνιο άρα και το L θα είναι κάτω διδιαγώνιο. Επομένως $A^{-1} = L^{-T} \cdot L^{-1}$. Το L^{-1} είναι κάτω τριγωνικό αλλά όχι κάτω διδιαγώνιο. Ομοίως και το L^{-T} θα είναι άνω τριγωνικό αλλά όχι διδιαγώνιο. Επομένως A^{-1} είναι το γινόμενο ενός άνω και ενός κάτω τριγωνικού μητρώου, τα οποία θα είναι πλήρη, στην γενική περίπτωση.

Γιατί δε χρειάζεται οδήγηση για λύση με τα L, U?
Επειδή το L έχει το μέγιστο στο μέτρο στοιχείο του στη διαγώνιο, επομένως η μερική οδήγηση δε θα είχε αποτέλεσμα. Το U είναι άνω τριγωνικό άρα το μέγιστο στο μέτρο στοιχείο από εκείνα που βρίσκονται στις θέσεις j ως n της στήλης j θα είναι πάντα στη διαγώνιο, άρα η μερική οδήγηση δε θα είχε αποτέλεσμα.

IEEE διπλής ακρίβειας μέγιστο σφάλμα αποκοπής. Ποιο τού?

$$= \max_x \frac{|x - fl(x)|}{|x|} = \frac{|0.0\ldots1|}{1.0} = 2^{-t} + 2^{-t-1} + \ldots = 2^{-t}(1 + 2^{-1} + \ldots)$$

$$= 2^{-t} \quad 1 \quad 2^{-t+1} = 2^{-53+1} \quad 2^{-52}$$

Έστω $p_n(x) = \prod_{j=1}^{n}(x - r_j)$ Αλγόριθμος υπολογισμού των συντελεστών της δυναμοσειράς του $P_n(x)$. (13)

```
=> a(2)=1; a(1) = -r(1); a(3:n+1)=0;
   for j=2:n
       t(2:j+1) = a(1:j), t(1)=0
       a(1:j+1) = t(1:j+1) - r(j)*a(1:j+1)
   end
```

$y \leftarrow y + Ax \quad A \in R^{n\times n} \quad x, y \in R^n \qquad \Phi, \Omega = ?$

$\Rightarrow \Phi = \Phi_L + \Phi_S. \qquad \Phi_L = n^2 + 2n$

$\underset{n\times n}{[\ ]}\underset{n\times 1}{[\ ]} = \underset{n\times 1}{[\ ]} \qquad (n+n-1)n = 2n^2 - n$

$\underset{n\times 1}{[\ ]} + \underset{n\times 1}{[\ ]} = \underset{n\times 1}{[\ ]} \qquad \Omega = n$

$\Phi_A = n^2 + 3n$

$\Omega = 2n^2$

$\Phi_{STORE} = n$

$A \leftarrow A + xy^T$

$\underset{\substack{n\times 1\\ x}}{[\ ]}\underset{\substack{1\times n\\ y^T}}{[\ ]} = \underset{n\times n}{[\ ]} \qquad n^2$

$\Phi_L = n^2 + 2n$

$\underset{n\times n}{[\ ]} + \underset{n\times n}{[\ ]} = n^2 \qquad \Phi_S = n^2$

$\Phi_A = 2n^2 + 2n$

$\Omega = 2n^2$

$y = (A^2 - I)x$ χωρίς LOAD

1. $\underset{n\times n}{[\ ]}\cdot\underset{n\times n}{[\ ]} = 2n^3 - n^2 \qquad [n\times n] - [n\times n] = n^2 \qquad [n\times n][n\times 1]$
$= 2n^2 - n.$

$\Omega = 2n^3 + 2n^2 - n \qquad \Phi = n^2 + 2n + O(n^2)$

$O(n^3) \xleftarrow{\text{πράξεις}}$

ii) $(A^2 - I)x \quad (A \cdot A - I)x = A \cdot (Ax) - Ix$, δηλαδή ...

DOT: $\sigma = x^T y$ SAXPY: $y \leftarrow y + ax$

ανανέωση 1ης τάξης: $C \leftarrow C + ab^T$ MV: $y \leftarrow y + Ax$

$A = (I - uv^T)x$ (?)

$A = Ix - uv^T x = Ix - \underbrace{(v^T x)}_{\text{αριθμός.}} u$

A,B $\in \mathbb{R}^{n\times n}$ κάτω τριγωνικά: Αλγόριθμος για το $C = A \cdot B$ και ακριβές κόστος;

$C(i,j) = A(i,:) * B(:,j) = A(i,1:i) * B(1:i,j)$

```
for i=1:n
    for j=1:i
        C(i,j) = A(i,j:1) *B(1:i,j);
    end
end
```

$\Theta = \sum_{i=1}^{n} \sum_{j=1}^{i} [2(i-j+1)-1] = 2\sum_{i=1}^{n} i^2 - \sum_{i=1}^{n} i = \frac{2n^3}{3} + \frac{n^2}{2} - \frac{n}{6}$

$B = B + (x \cdot y^T)^p$: Ν.Δ.Ο flops ανεξάρτητο του p.

Για το xy^T θέλω n^2 πράξεις.

Για το $(xy^T)^p$ θέλω $(p-1)(2n^3 - n^2)$ πράξεις.

Άρα $\Theta = (p-1)(2n^3 - n^2) + n^2 + n^2$

$\Phi_{\min} = 2n + 2n^2 + 1$

μπορώ όμως: $B = B + (\underbrace{y^T x}_{2n-1+1})^{p-1} \cdot x y^T$

n

n^2

n^2 $\Theta = 2n^2 + 3n$

15

$z = y + A^k x$

$z = y + A \cdots A x$

$\Phi_{min} = n^2 + 3n + 1$

$\Theta = (2n^2 - n)k + n$

```
Load A, x, y, k
    for i = 1 : k
        x = A * k;
    end.
    z = y + x
    store z;
```

υλοποίηση.

$y = (A + uv^T)x$

$Ax + uv^T x \Leftrightarrow Ax + (v^T x)u$ αριθμός

$2n^2 - n$

$n + n - 1$

$+ n \quad + n$

$= 2n^2 + 3n - 1$

άρα $\Theta = 2n^2 + 3n - 1$

$\Phi_{min} = 4n + n^2$

$\mu_{min} = \dfrac{\Phi_{min}}{\Theta}$

υλοποίηση με $O(n)$ μνήμη

```
Load x, u, v
    y = (v^T x) u
    for i = 1 : n
        LOAD A(i,:)
        y = yi + A(i,:) * x
    end
    store y
end.
```

φέρνει γραμμή-γραμμή

$B = B + (xy^T)^P$

$\Phi_{min} = 2n^2 + 2n + 1$

το $(xy^T)^P = x y^T x y^T \cdots x y^T$

$2n^2 - n$

για το $y^T x$: $2n - 1 = \psi$

άρα $(\psi^{(P-1)} x) = 1 + n$

$B + (\psi^{P-1} x) y^T$

$\Theta = 2n^2 + 3n$

$\mu_{min} = \dfrac{\Phi}{\Theta} = \dfrac{2n^2 + 2n + 1}{2n^2 + 3n}$

$\Delta = \Delta + X^T (XY^T)^r$

$\Delta = \Delta + X^T(XY^T XY^T \ldots XY^T) = \Delta + X^T X(Y^T X Y^T \ldots X) Y$

$X^T X : \underline{2n-1}$ [·]

$Y^T X : \underline{2n-1}$ [·]

$X^T X \cdot (Y^T X)^{p-1} = \underline{2} \cdot Y^T = \underline{n}$

$\Delta + \underset{\sim\sim}{n} \longrightarrow n$ άρα $\Delta = n$

$\phi_{min} = 4n + 1 \rightarrow$ για το p $\underline{o} = 6n$

$\mu_{min} = \frac{2}{3} + O\left(\frac{1}{n}\right)$

$A \in R^{10\times n}$ και $b \in R^n$ και $y \leftarrow Ab$

$$A = \begin{bmatrix} 1 & \cdots & n \\ \vdots & & \\ 10 & & n \end{bmatrix} \cdot b = \begin{bmatrix} 1 \\ \vdots \\ n \end{bmatrix} = \begin{bmatrix} \ \end{bmatrix}$$

$10\times n \qquad n\times 1 \qquad 10\times 1$

$\phi_L = 10n + n$

$\phi_S = 10$

$\phi_{CA} = $ [illegible]$n + 1$[illegible]

Σε μνήμη $O(1)$.

```
LOAD y
for j = 1:n
    LOAD b(j)
    for i = 1:10
      LOAD A(i,j)
      y(i) = y(i) + A(i,j)·b(j)
    end
  end
Store y.
```

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

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

με $AU = F$. A τριδιαγώνιο με στοιχεία γ_j, a_j, b_j στις θέσεις $j-1, j, j+1$ της γραμμής j

$A = trid_n[\gamma_j, a_j, b_j]$

$a_j = \left(\frac{2}{h^2} + c_j\right) \quad b_j = \left(-\frac{1}{h^2} + \frac{b_j}{2h}\right) \quad \gamma_j = \left(-\frac{1}{h^2} - \frac{b_j}{2h}\right)$

---

② Έχω την $-x^2 \frac{d^2u(x)}{dx^2} + 2\frac{du(x)}{dx} + x^2 u(x) = x$

$x \in [1,2]$, $u(1) = 0$ $u(2) = 1$ $n = 4$ εσωτερικά

$U_0 \; U_1 \; U_2 \; U_3 \; U_4 \; U_5$ (1 … 2)

$h = \frac{2-1}{n+1} = \frac{1}{5}$ $\quad U_0 = u(1) = 0$

$U_5 = u(2) = 1$

$-x^2 u'' + 2u' + x^2 u = x$

$$-x\left(\frac{U_{j-1} + U_{j+1} - 2U_j}{h^2}\right) + 2\left(\frac{U_{j+1} - U_{j-1}}{2h}\right) + x^2 U_j = x$$

$x_i = 1 + ih \iff x_0 = 1 \quad x_1 = 1,2 \quad x_2 = 1,4 \quad x_3 = 1,6$

$x_4 = 1,8 \quad x_5 = 2,0$

για $i = 1$ ...
για $i = 2$ ... } αλλάζει και το x_i προσοχή

$$\begin{array}{cccc} \begin{bmatrix} - & - & 0 & 0 \\ - & - & - & 0 \\ 0 & 0 & - & - \end{bmatrix} \\ U_1 \; U_2 \; U_3 \; U_4 \end{array} \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{bmatrix} = \begin{bmatrix} \; \end{bmatrix} \leftarrow \text{δεύτερα μέλη εξισώσεων}$$

$U_1 \; U_2 \; U_3 \; U_4$ } → συντελεστές αυτών

Πρέπει να είναι αντιστρέψιμο το ~~A~~ για να μπορεί (18) να λυθεί με LU ή Cholesky. Βρίσκω ιδιοτιμές του A και βλέπω αν μπορεί να περιέχονται στην ένωση δίσκων

$$|\lambda - 0| \leq 2 \quad , \quad |\lambda - 0| \leq 1$$

$$-2 \leq \lambda - 0 \leq 2 \qquad -1 \leq \lambda - 0 \leq 1$$

$$\lambda > 0$$

άρα ιδιοτιμές θετικές αφού είναι συμμετρικό ($A = A^T$). τότε είναι σ.θ.ο. άρα χρησιμοποιώ cholesky που είναι πιο φτηνή από LU.

Euler ΕΜΠΡΟΣ

$\Longrightarrow$ $u_1(0) = 2 \quad u_2(0) = 1 \qquad \frac{du(t)}{dt} = -Au(t) \quad A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$

$\Delta = \frac{1}{2} = h$ βρείτε το $T = 2.0$ $\qquad u = \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix}$

$$u(t+h) - u(t) = -hAu(t) \Leftrightarrow u(t+h) = u(t) - hAu(t) \Leftrightarrow$$

$$u(t+h) = (I - hA)\, u(t)$$

← υπάρχει ο μοναδιαίος πίνακας I. Προσοχή το u(t) το βάζω δεύτερο. Αν το βάλω μπροστά ΚΑΗΚΑ!!!

$$(I - hA) = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix} \qquad \begin{matrix} u_1(0) \\ u_2(0) \end{matrix}$$

Ξεκινάω πάντα από αυτή τη χρον. στιγμή

$t=0$ $\quad U(0,5) = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix} \cdot U(0) = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1 \end{bmatrix}$

$t=1/2$ $\quad U(1) = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix} \cdot U(0,5) = \begin{bmatrix} 1/2 \\ 1/4 \end{bmatrix}$

$t=1$ $\quad U(1,5) = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix} U(1) = \begin{bmatrix} 1/4 \\ 1/8 \end{bmatrix}$

$\cancel{t=2}$ $1,5$ $\quad U(2) = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix} U(1,5) = \begin{bmatrix} 1/8 \\ 1/16 \end{bmatrix}$

τ 3 χρον. στιγμή

(19)

Πίσω Euler

$T=2$? $\Delta t = 2 = h$ $A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ $\left.\begin{matrix} u_1(0)=2 \\ u_2(0)=1 \end{matrix}\right| V_0$

$u(t+h) - u(t) = -hAu(t+h) \Leftrightarrow$

$-u(t) = -hAu(t+h) - u(t+h) \Leftrightarrow$

$u(t) = [I + hA]\, u(t+h)$

Αν είχε άλλο Δt

$[I + hA] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \cdot 2 = \begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix}$

Για t=0

$U(0) = \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix} \begin{bmatrix} U_1(2) \\ U_2(2) \end{bmatrix} \Leftrightarrow$

$2 = 5u_1(2) - 2u_2(2)$ | $u_1(2) = \ldots$

$1 = -2u_1(2) + 5u_2(2)$ | $u_2(2) = \ldots$

Για να τείνει στο 0 πρέπει οι απόλυτες τιμές των ιδιοτιμών του μητρώου να είναι <1. δηλαδή η φασματική ακτίνα του μητρώου να είναι φραγμένη αυστηρά από το 1. Οι ιδιοτιμές του $I - hA$ θα είναι $1 - h\lambda(A)$ πρέπει $|1 - h\lambda(A)| < 1$ και επειδή οι ιδιοτιμές είναι πραγματικές (συμμετρικό μητρώο) θα πρέπει:

$-1 < 1 - h\lambda(A) < 1 \Leftrightarrow -2 < -h\lambda(A) < 0 \Leftrightarrow$

$2 < h\lambda(A) < 0 \Leftrightarrow$

$2 > \frac{h}{\lambda(A)} > 0$ άρα $h < 2\lambda$.

βρίσκω ιδιοτιμές του $I - hA$ $\left|\begin{matrix} \lambda_1 = 1/2 \\ \lambda_2 = 5/2 \end{matrix}\right|$ $h < 1$

Neumann

Μας δίνει την τιμή της παραγώγου σε κάποιο άκρο τότε

U_{-1} ... $U_0 \quad U_1 \quad U_2 \quad U_3 \quad U_4 \quad U_5$

U_{-1} → το προσθέτω

$$U_j' = \frac{U_{j+1} - U_{j-1}}{2h} \Rightarrow U_0' = \frac{U_1 - U_{-1}}{2h}$$

γνωστά

$$\Rightarrow U_1 - U_{-1} = 2hU_0' \Rightarrow$$

$$U_{-1} = U_1 - 2hU_0'$$

το αντικαθιστώ στις εξισώσεις σχέσεις

$$z = \frac{x}{\sqrt{x^2+y^2}}$$

$t_1 \leftarrow fl(x * x) \qquad x^2(1+\delta_1)$

$t_2 \leftarrow fl(y * y) \qquad y^2(1+\delta_2)$

$t_3 \leftarrow fl(t_1 + t_2) \qquad (t_1+t_2)(1+\delta_3)$

$t_3 \leftarrow fl(\sqrt{t_3}) \qquad \sqrt{t_3}(1+\delta_4)$

$t_1 \leftarrow fl\left(\frac{x}{t_3}\right) \qquad \frac{x}{t_3}(1+\delta_5)$

$$y = \frac{x}{t_3}(1+\delta_5) = \frac{x \,(1+\delta_5)}{\sqrt{(x^2(1+\delta_1) + y^2(1+\delta_2))(1+\delta_3)}\,(1+\delta_4)}$$

ισχυρισμός;

$|\delta_j| \leq u$

και πως τα $(1+\delta_n)$ είναι ένα δ;

ΝΑΙ

Ανάλυση σφαλμάτων με FMA!. (2

s=0; for j=1:4; s=s+x(j)·y(j); end;

εμπρός σφάλμα?

$$\hat{s} = [[[(x_1y_1)(1+\delta_1) + x_2y_2](1+\delta_2) + x_3y_3]](1+\delta_3) + (x_4y_4)\cdot$$

$$\cdot(1+\delta_4) = x_1y_1(1+\delta_1)(1+\delta_2)(1+\delta_3)(1+\delta_4) + (x_2y_2)(1+\delta_2)\cdot$$

$$\cdot(1+\delta_3)(1+\delta_4) + (x_3y_3)(1+\delta_3)(1+\delta_4) + (x_4y_4)(1+\delta_4) =$$

$$= (x_1y_1)(1+\theta_4) + (x_2y_2)(1+\theta_3) + (x_3y_3)(1+\theta_2) + (x_4y_4)(1+\theta_1)$$

$$|\hat{s}-s| = |x_1y_1\theta_4 + x_2y_2\theta_3 + x_3y_3\theta_2 + x_4y_4\theta_1| \le$$

$$\le |x_1y_1||\theta_4| + |x_2y_2||\theta_4| + |x_3y_3||\theta_4| + |x_4y_4||\theta_4| \le$$

$$\le \frac{4u|s|}{1-4u} = \gamma_4 |s|$$

Αίσθ. ευστάθεια Horner?

```
sn = an.
for k = n-1:-1:0
  sk = x s(k+1) + ak
end
```

$$\hat{s}_{n-1} = (x s_n\langle 1\rangle + a_{n-1})\langle 1\rangle = x a_n\langle 2\rangle + a_{n-1}\langle 1\rangle$$

$$\hat{s}_{n-2} = (x s_{n-1}\langle 1\rangle + a_{n-2})\langle 1\rangle$$

$$\hat{s}_0 = a_0\langle 1\rangle + a_1 x\langle 3\rangle + \ldots a_{n-1}x^{n-1}\langle 2n-1\rangle + a_n x^n\langle 2n\rangle = (1+\theta_1)a_0 + \ldots + (1+\theta_{2n})a_n x^n$$

άρα $\hat{s}_0 = fprog(a_0,\ldots,a_n,x) = f(a_0(1+\theta_1),\ldots,a_n(1+\theta_{2n}),x)$, άρα μικρός προς τα πίσω σφάλμα. Το θ_{2n} χαρακτηρίζει την μέγιστη ασάφεια στους συντελεστές → δεν καθε συντελ. Δε είναι σταθερώς γνωστός αλλά $|a_i-\hat{a}_i| \le \gamma_{2n}|a_i|$ τότε το σφάλμα από την ασάφεια αυτή θα είναι μεγαλύτερο απ οποιοδήποτε σφάλμα των πράξεων α.κ.υ του αλγορίθμου

Άρα πίσω σταθερός!

Σφάλμα υπολογισμού εξωτερικού γινομένου $A = xy^T$ (22

$A = xy^T$ άρα κάθε στοιχείο του A υπολογίζεται ως:

$$fl(a_{ij}) = a_{ij}(1+\delta_{ij}) = x_i y_j (1+\delta_{ij}), \quad |\delta_{ij}| \le u.$$

Επομένως: $fl(A) = xy^T + E$, όπου $|E| \le |x||y|^T u$ άρα έχουμε προς τα εμπρός σταθερότητα. Όμως δεν είναι πίσω σταθερή. Αν ήταν θα υπήρχαν $\Delta x, \Delta y$ ώστε

$$fl(A) = xy^T + E = (x+\Delta x)(y+\Delta y)^T.$$

Αυτή όμως η ισότητα δεν ισχύει γιατί θα είχαμε: $E = (x+\Delta x)(y+\Delta y)^T - xy^T$ και το δεξιό μέλος μπορεί να γραφεί:

$$E = [x, x+\Delta x]\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{bmatrix} (y+\Delta y)^T \\ y \end{bmatrix}$$

το μητρώο στα δεξιά έχει τάξη ≤ 2 ενώ το αριστερό έχει τάξη n. Επομένως δεν έχουμε ισότητα. Δεν είχαμε αρκετά στοιχεία εισόδου σε σχέση με τα εξόδου για να ρίξουμε το σφάλμα του αλγορίθμου.

ΝΔΟ ο πολ/σμός 2×2 κάτω τριγωνικών μητρώων $A, B \in \mathbb{R}^{2\times 2}$ είναι πίσω σταθερός.

$$A = \begin{pmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{pmatrix} \times B = \begin{pmatrix} b_{11} & 0 \\ b_{21} & b_{22} \end{pmatrix} =$$

$$= \begin{pmatrix} a_{11}b_{11}(1+\delta_1) & 0 \\ \left((a_{21}b_{11})(1+\delta_2) + (a_{22}b_{21})(1+\delta_3)\right)(1+\delta_4) & a_{22}b_{22}(1+\delta_5) \end{pmatrix}$$

$$\tilde{A} = \begin{pmatrix} a_{11}(1+\delta_1) & 0 \\ a_{21}(1+\delta_2)(1+\delta_4) & a_{22} \end{pmatrix} \quad \tilde{B} = \begin{pmatrix} b_{11} & 0 \\ b_{21}(1+\delta_3)(1+\delta_4) & b_{22}(1+\delta_5) \end{pmatrix}$$

είναι πίσω σταθερός εφόσον

$$\left|\frac{\tilde{a}_{ij} - a_{ij}}{a_{ij}}\right| \le |\theta_2| = \frac{2u}{1-2u} = \gamma_2 \quad \text{και} \quad \left|\frac{\tilde{b}_{ij} - b_{ij}}{b_{ij}}\right| \le |\theta_2| \le \gamma_2 = \frac{2u}{1-2u}$$

ΝΔΟ ο υπολογισμός της ορίζουσας ενός A 2×2 είναι πίσω σταθερός:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad fl(|A|) = \left[(a\cdot d)(1+\delta_1) - (c\cdot b)(1+\delta_2) \right](1+\delta_3) =$$

$$= ad(1+\delta_1)(1+\delta_3) - cb(1+\delta_2)(1+\delta_3).$$

πίσω σταθερός αφού:

$$fl(|A|) = |\tilde{A}| = \begin{vmatrix} a(1+\theta_1) & b(1+\theta_1') \\ c & d \end{vmatrix} \text{ και } |\theta_2|, |\theta_2'| \leq h$$

Μέθοδος LU

1. Υπολογίζω τους τριγωνικούς LU.
2. "=>" λύνω $Lz = b$ ως προς z.
3. "<=" λύνω $Ux = z$ ως προς x.

→ L = κάτω τριγωνικό με μονάδες στη διαγώνιο.

→ U = άνω τριγωνικό

Μερική οδήγηση: $L_{n-1}P_{n-1} \cdots L_1 P_1 A = U \Leftrightarrow A = (L_{n-1}P_{n-1} \cdots L_1 P_1)^{-1} U$.

$\Rightarrow P_{n-1} \cdots P_1 A = P_{n-1} \cdots P_1 (L_{n-1}P_{n-1} \cdots L_1 P_1)^{-1} U \Leftrightarrow$

$$PA = LU.$$

Έστω μετά από LU έχω

$$\begin{matrix} 1 & 2 & 3 \\ 0,5 & 1 & 2 \\ -1 & 1 & 1 \end{matrix} \rightarrow U, \quad \downarrow L$$

1) $A = ?$

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 0,5 & 1 & 0 \\ -1 & 1 & 1 \end{bmatrix} \quad U = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix} \quad LU = A = \begin{pmatrix} 1 & 2 & 3 \\ 0,5 & 2 & 3,5 \\ -1 & -1 & 0 \end{pmatrix}$$

2) $Ax = b = \begin{bmatrix} 11 \\ 10 \\ -4 \end{bmatrix} = ?$

$LUx = b \Leftrightarrow L^{-1} L \cdot Ux = L^{-1} b \Leftrightarrow Ux = L^{-1} b$

$$L^{-1} = L / I \text{ επαυξημένος } \left[\begin{matrix} 1 & 0 & 0 \\ 0,5 & 1 & 0 \\ -1 & 1 & 1 \end{matrix} \middle| \begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{matrix}\right] \cdots$$

Στοιχειώδης προβολή

$$Px = \frac{uu^T}{u^Tu}\cdot x = \frac{u}{\|u\|}\frac{u^Tx}{\|u\|} = \frac{u}{\|u\|}\cdot\|x\|\cdot\cos(x,u).$$

π.χ. $u = \begin{bmatrix}1\\2\\3\end{bmatrix}$ $\quad P = \dfrac{[1,2,3]\begin{bmatrix}1\\2\\3\end{bmatrix}}{\begin{bmatrix}1\\2\\3\end{bmatrix}\cdot[1,2,3]} = \dots$

Ερμιτιανά

$$H = I - \frac{2uu^T}{u^Tu}. \quad Hx = \left(I - \frac{2}{u^Tu}uu^T\right)x = x - Px - Px.$$

$H^T = H$ ισχύει $H^T\cdot H = \cancel{HH} = H^2 = I$

$$HA = A - \frac{2}{u^Tu}u(A^T\cdot u)^T \quad AH = A - \frac{2}{u^Tu}(Au)u^T$$

QR

Μετά από QR : $\begin{bmatrix}3 & -2 & 1\\2 & 1 & 4\\1 & -1 & 2\end{bmatrix} \to R$

① $A = ?$

$R = \begin{bmatrix}3 & -2 & 1\\0 & 1 & 4\\0 & 0 & 2\end{bmatrix}$ φτιάχνω κάτω τριγωνικό με 0 στη διαγώνιο

$\left\{\begin{bmatrix}1\\2\\1\end{bmatrix}\begin{bmatrix}0\\1\\-1\end{bmatrix}\begin{bmatrix}0\\0\\1\end{bmatrix}\right\}$

$u_1 \quad u_2$

$Q = H_1H_2\dots H_n \to H_1\cdot H_2$

$$H_1 = I - \frac{2u_1\cdot u_1^T}{u_1^Tu_1} = \begin{bmatrix}2/3 & -2/3 & -1/3\\-2/3 & -1/3 & -2/3\\-1/3 & -2/3 & 2/3\end{bmatrix}$$

$H_2 = \begin{bmatrix}1 & 0 & 0\\0 & 0 & 1\\0 & 1 & 0\end{bmatrix}$ $\Big\}$ $Q = H_1\cdot H_2$ γνωστό.

$A = QR$: γνωστό

② λύστε $Ax = b = \frac{1}{3}\begin{bmatrix}-5\\16\\4\end{bmatrix}$

$H_2\cdot H_1\cdot Ax = R\cdot x = H_2\cdot H_1\cdot b = \begin{bmatrix}2\\5\\2\end{bmatrix}$ άρα

$\begin{bmatrix}3 & -2 & 1\\0 & 1 & 4\\0 & 0 & 2\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix} = \begin{bmatrix}2\\5\\2\end{bmatrix}$ $\quad \left.\begin{matrix}x_1\\x_2\\x_3\end{matrix}\right\}$ γνωστά

Blas-1: Όταν ακριβώς μια από τις διαστάσεις n_j είναι μεγαλύτερη της μονάδας, δηλ. πράξεις του τύπου διανυσμα/διανυσμα π.χ το DOT ($n_3>1$) και οι SAXPY ($n_2>1$ ή $n_3>1$). Ο αριθμός πράξεων και μεταφορών τότε είναι $O(n)$ και επομένως $\mu_{min} = O(1)$, άρα έχουμε μικρή τοπικότητα. (25)

Blas-2: Όταν ακριβώς δυο διαστάσεις είναι μεγαλύτερες της μονάδας, δηλ. πράξεις τύπου μητρώο/διάνυσμα. Λαμβάνοντας τις δύο αυτές διαστάσεις ίσες με n έχουμε πως ο αριθμός των πράξεων και των μεταφορών είναι $O(n^2)$, άρα καλύτερη τοπικότητα χώρου και χρόνου από την Blas-1.

Blas-3: Όταν $n_j>1$ για $j=1,2,3$ δηλ. πράξεις του τύπου μητρώο/μητρώο, οπότε έχουμε και αυξημένη τοπικότητα π.χ. αν $n_1=n_2=n_3=n$ τότε έχουμε $O(n^2)$ μεταφορές με $O(n^3)$ πράξεις.

Υλοποίηση Blas: $C = C + AB$ με $n_1=n_2=n_3=n$ $\quad K \ll 3n^2 \ll M$

```
for j=1:N
  (*LOAD Bj, Cj στην κρυφή*)
  for k=1:n
    (*LOAD A:,k στην κρυφή και
       ανανέωση 1ης τάξης του Cj*)
  end
  (*STORE Cj*)
end.
```

```
for I=1:N
for J=1:N
  (*LOAD CIJ στην κρυφή μνήμη
  for K=1:N
  (*LOAD AIK, BKJ στην κρυφή και
   ανανέωση CIJ*)
  end.
  (*STORE CIJ*)
end
end.
```

$$\phi_1 = \sum_{j=1}^{N} \left(n^2 + 3n(n/N)\right) = 2n^2\left(\frac{3}{2} + \frac{n^2}{K}\right)$$

$$\phi_2 = 2n^2(1+N) \approx \frac{n^3\, 2\sqrt{3}}{\sqrt{K}}$$

Οι πίνακες B,C είναι ομοιόμορφα τεμαχισμένοι σε N ομάδες από στήλες και $n = Nm$ για κάποιον ακέραιο m. Το N επιλέγεται ώστε να αποθηκεύονται 2 ομάδες και 1 στήλη του A

Τεμαχίζουμε τους πίνακες σε $N \times N$ τετραγωνικούς υποπίνακες μεγέθους m, ώστε $3m^2 \approx K$ και χρησιμοποιούμε την DC σε μορφή ijk σε επίπεδο υποπινάκων.

Ανάλυση LU

Έστω έχουν οι μεταθέσεις και $A=LU$.

$A+E=\hat{L}\hat{U}$ όπου $|E|\le \gamma_n|\hat{L}||\hat{U}|$, $\||E|\|\le \gamma_n \||\hat{L}|\|\,\||\hat{U}|\|$

λύνοντας $\hat{L}y=b$ έχουμε $(\hat{L}+\Delta L)\hat{y}=b$ όπου $|\Delta L|\le \gamma_n|\hat{L}|$ και το ίδιο για το $\hat{U}$. Επίσης:

$b=(\hat{L}+\Delta L)(\hat{U}+\Delta U)\hat{x}$ άρα $b=(\hat{L}\hat{U}+\hat{L}\Delta U+\Delta L\hat{U}+\Delta L\Delta U)\hat{x}$

$=(A+\underbrace{E+\hat{L}\Delta U+\Delta L\hat{U}+\Delta L\Delta U}_{\Delta A})\hat{x}$ επομένως το υπολογισμένο x ικανοποιεί το $b=(A+\Delta A)x$ όπου $\Delta A=E+\hat{L}\Delta U+\Delta L\hat{U}+\Delta L\Delta U$. Η πίσω ευστάθεια εξαρτάται από το μέγεθος του $\frac{\|\Delta A\|}{\|A\|}$. Για να φράξω το ΔA:

$$|\Delta A|\le |E|+|\hat{L}||\Delta U|+|\Delta L||\hat{U}|+|\Delta L||\Delta U|\le 3\gamma_n|\hat{L}||\hat{U}|+O(u^2)$$

$$\|\Delta A\|\le 3\gamma_n\||\hat{L}|\|\cdot\||\hat{U}|\|.$$

έχουμε πίσω ευστάθεια όταν $\frac{\|\Delta A\|}{\|A\|}\approx O(u)$.

Αν $3\gamma_n\||\hat{L}|\|\cdot\||\hat{U}|\|\approx O(u)\cdot\|A\|$ τότε

$\frac{\|\Delta A\|}{\|A\|}\approx O(u)$. Η πίσω ευστάθεια καθορίζεται από το μέγεθος του $\||\hat{L}|\|\cdot\||\hat{U}|\|$

⇒ Χωρίς οδήγηση ο παράγοντας μπορεί να είναι πολύ μεγάλος, γιατί δεν έχουμε τρόπο να φράξουμε τους όρους του $\hat{L}$, τότε εκτός από ειδικές περιπτώσεις δεν έχουμε πίσω ευστάθ.

⇒ μερική ή πλήρης οδήγηση: ο $\|\hat{L}\|$ δεν μπορεί να γίνει πολύ μεγάλος γιατί το μέγιστο στοιχείο κάθε στήλης είναι 1.

Να ορίσετε με σαφήνεια τι σημαίνει πίσω ευστάθεια μιας $f : R^n \to R^m$ και εξηγήστε πως αν αποδείξαμε την πίσω ευστάθεια μπορούμε να αναγάγουμε την εύρεση της απόστασης $f(x)$ από $f_{prog}(x)$ σε πρόβλημα διαταραχών. (2=

Αν ο υπολογισμός είναι πίσω ευσταθής τότε μπορούμε να βρούμε με. διάνυσμα, έστω $x_{prog} \in R^n$, που είναι κοντά στο x, τέτοιο ώστε $f(x_{prog}) \equiv f_{prog}(x)$, δηλ. ο υπολογισμός της συνάρτησης με το συγκεκριμένο αλγόριθμο σε α.κ.υ. είναι το ίδιο ακριβώς αποτέλεσμα με τον υπολογισμό της συνάρτησης στο x_{prog} με αριθμητική άπειρης ακρίβειας. Αν δείξουμε κάτι τέτοιο, ισχύει ότι $\|f(x) - f_{prog}(x)\| = \|f(x) - f(x_{prog})\|$ επομένως η απόσταση (το απόλυτο εμπρός σφάλμα) των δυο θα είναι ίση με την απόσταση της τιμής του f στο x από την τιμή της f σε ένα κοντινό σημείο x_{prog}, και η μελέτη της, εφόσον το x είναι κοντά στο x_{prog}, αντιστοιχεί σε πρόβλημα διαταραχών.

$A \in C^{n\times n}$ μιγαδικό και $B \in C^{n\times k}$. $A \cdot B = ?$ Ω, Φ, Blas.

$$C = A \cdot B = (A_R B_R - A_I B_I) + i(A_R B_I + A_I B_R)$$ όπου

$A = A_R + iA_I$, $B = B_R + iB_I$, $A_R, A_I \in R^{n\times n}$ και $B_R, B_I \in R^{n\times k}$

$nk(2n-1)$ άρα $\Omega = 4nk(2n-1) + 2nk$

$$\mu_{min} = \frac{2n^2+4nk}{8n^2k-2nk} \approx O(1/k)$$ και αφού $k >> 1$. Blas-3

$$\Phi = 2(n^2+2nk) = 2n^2+4nk.$$

Rigal-Gaches: Ε.κ. των υστέρων υπολογισμός άνω σφάλματος.

Έστω $\hat{x}$ λύση του $Ax=b$ και ότι $(A+\Delta A)\hat{x}=b+\Delta b$

~~και~~ επί $r=b-A\hat{x}$

$$bw=\inf\{\omega \mid (A+\Delta A)\hat{x}=b+\Delta b, \|\Delta A\|\le \omega\|A\|, \|\Delta b\|\le \omega\|b\|\}$$

$$\hookrightarrow bw=\frac{\|r\|}{\|A\|\|\hat{x}\|+\|b\|} \quad \cdot \frac{\|\Delta x\|}{\|x\|}\le \frac{k(A)}{1-k(A)\cdot\frac{\|\Delta A\|}{\|A\|}}\left(\frac{\|\Delta A\|}{\|A\|}+\frac{\|\Delta b\|}{\|b\|}\right)$$

Το μοντέλο με ιεραρχία μνήμης είναι πιο αξιόπιστο από το RAM, αλλά: ① το κόστος πρόσβασης στην κρυφή μνήμη και στους καταχωρητές θεωρείται αμελητέο συγκριτικά με το κόστος πρόσβασης στην κεντρική μνήμη ② ο χρήστης δεν έχει έλεγχο στα στοιχεία που μεταφέρονται στην κρυφή μνήμη ③ είναι δυνατό να υπάρχουν περισσότερα επίπεδα μνήμης ④ οι αριθμητικές πράξεις και οι μεταφορές που εκτελούνται εξαρτώνται από το μεταφραστή και δεν περιγράφονται πλήρως στο επίπεδο της υλοποίησης με ένα κλασσικό μοντέλο προγραμματισμού ⑤ Σύγχρονα συστήματα επιτρέπουν ρυθμούς εκτέλεσης που υπερβαίνουν την μια εντολή και την μια μεταφορά ανά κύκλο.

$A\in\mathbb{R}^{n\times n}$, $x\in\mathbb{R}^n$, $\omega\in\mathbb{R}$ φ μνήμη, $\underline{o}$, υλοποίηση με $O(n)$ ~~χρόνο~~

$y=(A-\omega I)x = Ax-\omega Ix = Ax-\omega x$

$\phi_{min}=n^2+n+n+1$. Σε κάθε βήμα φέρνω την γραμμή του

```
LOAD w,x
for i=1:n
    LOAD A(i,:)
        yi = A(i,:) x - wx
end
STORE y
```

$\underline{o}=2n^2+n$.

$$\mu_{min}=\frac{n^2+2n+1}{2n^2+n}$$

Γενικά $C\in\mathbb{R}^{m\times n}$, $D\in\mathbb{R}^{n\times k}$

$\underline{o}(C\cdot D)=m(2n-1)k = 2mnk-mk$ πράξεις.

- **Συνθήκη ακριβούς στρογγυλοποίησης**: Αν $\circledast$ συμβολίζει την υλοποίηση της πράξης $\odot$, τότε δοθέντων $x, y \in F$ ισχύει ότι $x \circledast y = fl(x \odot y) \in F$, δηλ. το αποτέλεσμα της πράξης στο σύστημα είναι σαν να εκτελείται η πράξη ακριβώς στον R (δηλ. $x \odot y$) και μετά να στρογγυλοποιείται το αποτέλεσμα.

- **Μοντέλο διάδοσης σφάλματος**: Έστω $x, y \in F$ και $x \odot y$, τότε $fl(x \odot y) = (x \odot y)(1+\delta)$, για κάποιο $|\delta| \leq u$.

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

- **Για την α.κ.υ IEEE μονή και διπλή ακρίβεια μετρήστε την ακρίβεια σε αριθμό δεκαδικών ψηφίων**: Η ουρά της μονής ακρίβειας έχει $24 = 23(+1)$ ψηφία ενώ σε διπλή $53 = 52(+1)$. Και στις δύο περιπτώσεις η αναπαράσταση είναι της μορφής $\pm m \times 2^E$ όπου $1 \leq m < 2$ έχει τη δυαδική μορφή $1.b_s \ldots$. Στην μονή, η δυαδική ακρίβεια δίνεται από τα 24 ψηφία της ουράς, επομένως χρειαζόμαστε d δεκαδικά ψηφία, όπου $10^{-d} \approx 2^{-24}$ ($d = 7 \approx 24 \log_{10} 2$), και στην διπλή $10^{-d} \approx 2^{-53}$ ($d = 16 \approx 53 \log_{10} 2$)

- **LINPACK benchmark**: για την αξιολόγηση της επίδοσης υπολογιστικών συστημάτων, αποτελείται από υπορουτίνες για την επίλυση γραμμικών συστημάτων και είναι από τα σημαντικότερα μετροπρογράμματα για την αξιολόγηση υπολογιστικών συστημάτων υψηλής απόδοσης.

Από τι εξαρτάται η επιλογή της μεθόδου επίλυσης του $Ax=b$? a) το μέγεθος του μητρώου: άν το μητρώο είναι πολύ μεγάλο, το αριθμητικό κόστος και το κόστος αποθήκευσης των άμεσων μεθόδων μπορεί να είναι απαγορευτικό (π.χ. η απαλοιφή Gauss στοιχίζει $O(n^3)$ πράξεις και $O(n^2)$ θέσεις αποθήκευσης) b). τη δομή του μητρώου: το μητρώο έχει χαρακτηριστικά που μπορούμε να εκμεταλλευτούμε και να πετύχουμε αποτελεσματικότερες μεθόδους (π.χ. συστήματα με μητρώα συμμετρικά και θ.ο. επιλύονται μέσω της Cholesky σε μισό κόστος της Gauss). (30)

$fl(a_{ij}) = a_{ij}(1+\delta_{ij})$, $|\delta_{ij}| \leq u$.

$|fl(A) - A| \leq |A| u$.

$fl(B \cdot A) = B \cdot A + E$, $|E| \leq u |BA|$

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

$F(b, t, e_{min}, e_{max})$ $\quad y = \pm m \times b^{e-t}$

t: διακριτότητα - ακρίβεια

e: τάξη μεγέθους

μονή ακρίβεια $F(2, 24, -125, 128)$.

διπλή " $F(2, 53, -1021, 1024)$.

Μετά από το $Ax=b$, $\hat{x}$ η λύση, $r = b - A\hat{x}$ και ξέρω

$$w = \frac{\|r\|}{\|A\|\|\hat{x}\| + \|b\|} = 1,5 \cdot 10^{-13}$$ και $\kappa_2(A) = 10^6$. άνω φράγμα για το εμπρός σχετικό σφάλμα?

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq \frac{2w\kappa(A)}{1 - w\kappa(A)} = 3 \cdot 10^{-7} / (1 - 1,5 \cdot 10^{-7}) \approx 3 \times 10^{-7}$$

Ανάλυση σφάλματος για το DOT $\Rightarrow s_n = x^T y$ (3

$\bar{s}_1 = fl(x_1 y_1) = x_1 y_1 (1+\delta_1)$.

$\bar{s}_2 = fl(\bar{s}_1 + fl(x_2 y_2)) = (x_1 y_1 (1+\delta_1) + x_2 y_2 (1+\delta_2))(1+\delta_3) =$

$= x_1 y_1 (1+\delta_1)(1+\delta_3) + x_2 y_2 (1+\delta_2)(1+\delta_3)$. με $|\delta_i| \leq u$.

$$\bar{s}_n = x_1 y_1 \prod_{\substack{j=1 \\ j\neq 2}}^{n+1} (1+\delta_j) + x_2 y_2 \prod_{j=2}^{n+1} (1+\delta_j) + \dots + x_n y_n \prod_{j=n}^{n+1} (1+\delta_j)$$

$\bar{s}_n = x_1 y_1 (1+\theta_n) + x_2 y_2 (1+\theta_n') + \dots + x_n y_n (1+\theta_2)$.

άρα DOT ακριβές εσωτερικό γινόμενο για τα $x_1, \dots x_n$, $y_1(1+\theta_n), y_2(1+\theta_n'') \dots y_n(1+\theta_2)$ όπου

$|\theta_j| \leq \dfrac{ju}{1-ju}$ Δείξαμε $fl(x^T y) = (x+\Delta x)^T y = x^T(y+\Delta y)$

όπου $|\Delta x| < \gamma_n |x|$, $|\Delta y| < \gamma_n |y|$

$f([x;y]) = x^T y$ θέτω $X = [x;y] \in \mathbb{R}^{2n}$

$\mathrm{cond}(f;X) = \dfrac{\|X\|}{|x^T y|} \left\| \dfrac{\partial f}{\partial X}\Big|_{[x;y]} \right\|$ και $\dfrac{\partial f}{\partial X}\Big|_{[x;y]} = [y;x] \in \mathbb{R}^{1\times}$

άρα $\mathrm{cond}(f;X) = \dfrac{\|[x;y]\|}{|x^T y|} \|[y;x]\|$

θεωρώντας $\|[x;y]\| = \|[y;x]\|$ έχω $\mathrm{cond}(f;X) \leq \dfrac{\|[x;y]\|^2}{|x^T y|}$

σφάλμα προς τα εμπρός < (δείκτης κατάστασης) · πίσω σφάλμα

Έστω ένας αριθμός με μέγιστο σχετικό σφάλμα ~~...~~ (3): 1% και ο άλλος 2% α) Βρείτε το μέγιστο μέγεθος του σφάλματος του γινομένου β). Αν είναι α.κ.ν. και το γινόμενο υπολογίζεται με μονή ακρίβεια βρείτε το μέγιστο μέγεθος του σχετικού σφάλματος στο γινόμενο.

α) Έστω $\hat{x} = x(1+\delta x)$ και $\hat{y} = y(1+\delta y)$ όπου ~~...~~

$|\delta x| \leq 0.01$ και $|\delta y| \leq 0,02$

$\hat{x}\cdot\hat{y} = xy(1+\delta x)(1+\delta y) = $ ~~...~~ $=$

$= xy + xy(\delta x + \delta y + \delta x \delta y)$ επομένως

$$\left|\frac{\hat{x}\hat{y} - xy}{xy}\right| = |\delta x + \delta y + \delta x \delta y| \leq 0.01 + 0,02 + 2\cdot 10^{-4} = 3\cdot 10^{-2} + 2\cdot 10^{-4} = 0,0302$$

β) έχω $\hat{x}\hat{y} = xy(1+\delta x)(1+\delta y)(1+\delta \kappa) =$

$= xy + xy[\delta x + \delta y + \delta\kappa + \delta x\delta y + \delta\kappa\delta x + \delta\kappa\delta y + \delta\kappa\delta x\delta y]$ επειδή έχω μονή ακρίβεια $t=23$ και $|\delta\kappa| \leq u$

$u = 2^{-24} \approx 6\cdot 10^{-8}$ οπότε το παραπάνω άθροισμα:

$0.0302 + 6\cdot 10^{-8} + 6\cdot 10^{-10} + 12\cdot 10^{-10} + 18\cdot 10^{-12}$

Το συμπλήρωμα του Schur:

$S = A_{22} - A_{21}A_{11}^{-1}A_{12}$ συμπλήρωμα του A ως προς A_{11}

$$\begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix}$$

33

Μέθοδος κανονικών εξισώσεων:

Αν: $A \in R^{m \times n}$, $b \in R^n$ και $A^T(b-Ax)=0$ τότε ισχύει

$\|b-Ax\|_2 \leq \|b-Ay\|_2 \quad \forall y \in R^n$

Ο $P = A(A^TA)^{-1} \cdot A^T \Rightarrow$ τελεστής ορθογώνιας προβολής επί του υποχώρου που παράγεται από τις στήλες του A.

η λύση δίνεται από το x τέτοιο ώστε:

$Ax = Pb = A(A^TA)^{-1} \cdot A^Tb$ και θέτοντας $x = (A^TA)^{-1}A^Tb$

άρα το x ικανοποιεί το $A^TAx = A^Tb$

(το A^TA είναι σ.θ.ο και αντιστρέψιμο)

Μέθοδος: • Υπολογισμός κάτω τριγωνικού τμήματος του

$C = A^TA$ και του $d = A^Tb$

• Cholesky $C = G \cdot G^T$

• Επίλυση του $Gy = d$ και του $G^Tx = y$

Μειονεκτήματα: ο A^TA ίσως καταστρέψει ειδική δομή που μπορεί να έχει το A. ο πολ/μος A^TA μπορεί να οδηγήσει σε απώλεια σημαντικών ψηφίων και αλλοίωση αποτελέσματος (ή υπερχείλιση - σπάνιο). Ευαισθησία της μεθόδου σε συσσώρευση αριθμητικών σφαλμάτων. δηλαδή:

$k_2(A^TA) = (k_2(A))^2$ άρα το σχετικό σφάλμα αυξάνει με το τετράγωνο του δείκτη κατάστασης του A και η λύση είναι ευαίσθητη σε αριθμητικά σφάλματα.

QR με Gram-Schmidt

$$q_1 = \frac{A}{\|A\|} \qquad q_2 = \frac{B}{\|B\|} \qquad q_3 = \frac{C}{\|C\|}$$

$\xrightarrow{1^o} a = A \quad \xrightarrow{2^o} B = b - \frac{A^Tb}{A^TA} \cdot A \quad \xrightarrow{3^o} C = c - \frac{A^Tc}{A^TA} A - \frac{B^Tc}{B^TB} \cdot B$

Π-χ $\quad a = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \checkmark \quad b = \begin{bmatrix} 2 \\ 0 \\ -2 \end{bmatrix} \quad c = \begin{bmatrix} 3 \\ -3 \\ 3 \end{bmatrix}$

$A = a$

$A^TA = 2$

άρα $B = b - \frac{A^Tb}{A^TA} \cdot A = b - \frac{2}{2} \cdot A = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix} \checkmark$

$$C = c - \frac{A^Tc}{A^TA} \cdot A - \frac{B^Tc}{B^TB} \cdot B = c - \frac{6}{6} A + \frac{6}{6} B = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

άρα $q_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \quad q_2 = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix} \quad q_3 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

άρα $A = QR \rightarrow \underset{A}{\begin{bmatrix} a & b & c \end{bmatrix}} = \underset{Q}{\begin{bmatrix} q_1 & q_2 & q_3 \end{bmatrix}} \underset{R}{\begin{bmatrix} q_1^Ta & q_1^Tb & q_1^Tc \\ & q_2^Tb & q_2^Tc \\ & & q_3^Tc \end{bmatrix}}$

ο R έχει θετική διαγώνιο και ο A' έχει ανεξάρτ. στήλες (όχι ο μικρός)

ο A^TA ισούται με $R^T \overset{1}{Q^TQ} R = R^TR$ άρα

$$A^TA\hat{x} = A^Tb \Rightarrow R^TR\hat{x} = R^TQ^Tb \text{ ή } R\hat{x} = Q^Tb$$

έτσι το λύνω με πίσω αντικατάσταση.

- Πόσοι αριθμοί κανονικοποιημένοι και πόσοι υποκανονικοποιημένοι μπορούν να αναπαρασταθούν στο $F = \pm m \times b^e$ (35)

κανονικ.: χρησιμοποιεί t ψηφία για την ουρά (m) καθένα παίρνει b τιμές.

ουρά b^{t-1}
εκθέτης $e_{max}-t \to e_{min}-t$
οι τιμές του διαστ. $(e_{max}-e_{min}+1)$
πρόσημο 2

$b^{t-1}(e_{max}-e_{min}+1) \cdot 2(b-$
$+1$ για το μηδέν
τιμές του ... μένου bit σε βάση b.

υποκανον.: $b^{t-1} \to$ ουρά
$2 \to$ πρόσημο
$1 \to$ εκθέτης $e_{min}-t$

$2(b^{t-1}-1)$ δεν υπολογίζω το μηδέν.

κανονικ.	υποκανον.	
$43 \cdot 10^9$	$17 \cdot 10^7$	double
$1,6 \cdot 10^{19}$	$9 \cdot 10^{15}$	single

- Στην IEEE πόσοι αριθμοί διπλής παρεμβάλλονται μεταξύ δύο αριθμών μονής;

$1,\ 1+\epsilon_M = 1+2^{-23}$ $\quad [1, 1+\epsilon_{M,s}]$

$[1,2] = 2^{-52} = 2^{-t}$

$\dfrac{2^{-23}}{2^{-52}} = 2^{29} \to$ αριθμοί διπλής ακρίβειας

$c \quad \left.\begin{matrix} m \cdot 2^e \\ (m+1)2^e \end{matrix}\right\} 2^{-e}$ εύρος του α $\quad$ εύρος του b

Θεώρημα Rigal-Gaches

Έχω $Ax=b$. ορίζω $r=A\hat{x}-b$. θέλω να ελαχιστοποιήσω το r.

$$\to \omega = \frac{\|r\|}{\|A\|\|\hat{x}\|+\|b\|} \qquad \frac{\|x-\hat{x}\|}{\|x\|} \leq \frac{2\omega\, k(A)}{1-\omega\, k(A)}$$

Πχ $\omega = 15\cdot 10^{-13}$ $k_2(A)=10^6$ και ζητείται εκτίμηση για το σχετικό σφάλμα.

$$\frac{\|x-\hat{x}\|}{\|x\|} \leq \frac{2\omega\, k(A)}{1-\omega\, k(A)} = 3\cdot 10^{-7}$$

Παραγοντοποίηση κατά ορθογώνια (LU)

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

$$L_{21} = A_{21}\cdot A_{11}^{-1} \qquad U_{11} = A_{11} \qquad U_{12} = A_{12}$$

$$U_{22} = A_{22} - A_{21}A_{11}^{-1}A_{12} \quad \Big| \text{- Schur}$$

① Στον ΕΥ μας ενδιαφέρουν ο σχεδιασμός, η ανάπτυξη και η χρήση αποδοτικών υπολογιστικών εργαλείων που βοηθούν στην πρακτική επίλυση των μαθηματικών μοντέλων της επιστήμης και της τεχνολογίας. Λαμβάνει υπόψη του τη φυσική και τα μαθηματικά του προβλήματος καθώς και το περιβάλλον υλοποίησης.

② Οι πιο σημαντικές κατηγορίες χρήσεων τεχνικών του ΕΥ: ① Προσομοίωση ② ανάλυση δεδομένων (εικόνας, ήχου) ③ υπολογιστική υποστήριξη γραφικών ④ εφαρμογές που απαιτούν ανταπόκριση σε πραγματικό χρόνο. real-time

③ Δύο κατηγορίες προβλημάτων: ① Αυτά που επιταχύνονται ικανοποιητικά μέσω των βελτιστοποιήσεων που επιτυγχάνει ο μεταφραστής ② αυτά που απαιτούν χρήση πληροφορίας υψηλότερου επιπέδου για την αποτελεσματική του επιτάχυνση.

④ Υπολογιστικοί πυρήνες: παραδείγματα: Διαδικασίες (που μπορεί να είναι υποπρογράμματα ή τμήματα προγράμματος που υλοποιούν συγκεκριμένες μαθηματικές εκφράσεις ή υπολογισμούς) στις οποίες αναλίσκεται σημαντικό ποσοστό του χρόνου εκτέλεσης του προγράμματος εφαρμογής που μελετούμε. Γι' αυτό, σημαντικό βήμα στη μείωση του χρόνου εκτέλεσης ενός προγράμματος εφαρμογής είναι η εύρεση των υπολογιστικών πυρήνων του και η επιτάχυνση της εκτέλεσής τους. π.χ. ① γεννήτρια τυχαίων αριθμών ② ταχύς μετασχηματισμός Fourier FFT ③ πολ/σμός μητρώου με διάνυσμα.

⑤ 5 πηγές σφαλμάτων στον ΕΥ: ① Στα δεδομένα ② Στη διακριτοποίηση των πραγματικών αριθμών και των εξισώσεων ③ Στην διακριτοποίηση [?] ακριβείας αριθμητικές πράξεις με αυτούς και στις περιορισμένης ακρίβειας [?] του πεπερασμένου [?] αριθμούς ④ Στον περιορισμό [?] για την εύρεση αποτελέσματος. Π.χ. τις επαναληπτικές μεθόδους για την εύρεση των ιδιοτιμών ενός μητρώου. Στις μεθόδους αυτές βασιζόμαστε στη σύγκλιση, αλλά στον υπολογιστή είμαστε υποχρεωμένοι να σταματήσουμε τις επαναλήψεις όταν κάποιος δείκτης μέτρησης σφάλματος γίνει αρκετά μικρός ⑤ Σε δεδομένα ή ενδιάμεσα αποτελέσματα τα οποία δεν έχουν προφυλαχθεί από τη λογική της μεθόδου επίλυσης.

⑥ Τα κύρια κριτήρια που χρησιμοποιούνται για την αξιολόγηση των εργαλείων του ΕΥ είναι α) η ακρίβεια των αποτελεσμάτων β) η ταχύτητα των υπολογισμών γ) το κόστος της συνολικής διαδικασίας.

///

πρακτική λύση

Κάποιοι αλγόριθμοι αποφεύγονται ως μη [illegible] (
α) η αριθμητική τους συμπεριφορά είναι κακή β) για να γίνει εκθετική
η υποσχόμενη επιτάχυνση του αλγορίθμου το μέγεθος πρέπει
να γίνει απαγορευτικά μεγάλο γ) η πολυπλοκότητα έχει υπολο-
γισθεί με βάση κάποιο ιδεατό μοντέλο προγραμματισμού
αλλά η υλοποίηση του αλγορίθμου οδηγεί σε σημαντικό επιπλέ-
κόστος
χαρακτηριστικά στις σύγχρονες αρχιτεκτονικές
Σημαντικά χαρακτηριστικά που έχουν επιδράσει στον σχεδιασμό αλγορίθμων είναι α) η αρ-
χιτεκτονική RISC β) η ιεραρχική του συστήματος αποθήκευ-
σης (καταχωρητές, κρυφή - κύρια μνήμη, δίσκος) γ) η ευρύτερη
διάδοση και χρήση παράλληλων συστημάτων και άλλων αρχι-
τεκτονικών δομών για αυξημένη απόδοση.

Τα μοντέλα που μας ενδιαφέρουν είναι αλλά και συμβολικά,
κατασκευάζονται χρησιμοποιώντας ~~[illegible]~~ σύμβολα, χρη-
σιμοποιώντας προκαθορισμένους κανόνες σύνθεσης. Τα μοντέλα αυτά
είναι κατά κάποιον τρόπο απλούστερα από κάποιο άλλο σύστημα
του οποίου συνιστούν πρότυπα - απλούστερα αλλά όχι απλοϊκά. Ένα
ένα μοντέλο είναι η ένδειξη μιας απλούστερης και σαφέστερης προσ-
διορισμένης κατάστασης πραγμάτων, με πρόθεση να διευκολυνθεί
η παρατήρηση περαιτέρω συνεπειών που μπορούν να επανεφαρμο-
σθούν δοκιμαστικά στο περισσότερο πολύπλοκο σύστημα της
πραγματικότητας. Περιγράφοντας ένα σύστημα με σαφή αξιώματα
που προδιαγράφουν τις ιδιότητες του μοντέλου και έτσι, κα-
τά κάποιον τρόπο, το δημιουργούν, μπορεί κανείς να συναγάγει
και άλλες συνέπειες από αυτά τα αξιώματα περί του προ-
τύπου, εφαρμόζοντας την μέθοδο της αυστηρής παραγωγής.

Ένα υπολογιστικό μοντέλο περιγράφει μια ιδεατή μηχανή για
την οποία μπορούμε να γράψουμε λογισμικό. Είναι ιδεατή γιατί
κρύβει αρχιτεκτονικές λεπτομέρειες και αλλαγές που οφείλο-
νται σε καθαρά τεχνολογικές εξελίξεις αλλά είναι και αρκετά
συγκεκριμένη ώστε να επιτρέπει την εξαγωγή συμπερασμάτων σχε-
τικών με την απόδοση των προγραμμάτων που γράφονται γι' αυτή.
π.χ. RAM. Σε αυτό, το κόστος και η χρονοπολυπλοκότητα θα
αποτελείται από επεξεργαστή
ανάγονται στον αριθμό πράξεων.
και μια μνήμη.
υπολογιστικό μοντέλο της αριθμητικής, δηλαδή
το [illegible] το μοντέλο.
[illegible] αλγορίθμων που επιτυγχάνουν καλές επιδόσεις
[illegible] αλγορίθμους
[illegible] παρέχουν ακριβή αποτελέσματα.

(α) Κάποιοι αλγόριθμοι αποδεικνύονται ως μη πρακτικοί [illegible] (α) η αριθμητική τους συμπεριφορά είναι κακή (β) για να γίνει εμφανής η υποσχόμενη επιτάχυνση του αλγορίθμου το μέγεθος πρέπει να γίνει απαγορευτικά μεγάλο (γ) η πολυπλοκότητα έχει υπολογισθεί με βάση κάποιο ιδεατό μοντέλο προγραμματισμού αλλά η υλοποίηση του αλγορίθμου οδηγεί σε σημαντικό επιπλέον κόστος

Σημαντικά χαρακτηριστικά στις σύγχρονες αρχιτεκτονικές έχουν επιδράσει στον σχεδιασμό αλγορίθμων είναι α) η αρχιτεκτονική RISC β) η ιεραρχική του συστήματος αποθήκευσης (καταχωρητές, κρυφή - κύρια μνήμη, δίσκος) γ) η ευρύτερη διάδοση και χρήση παράλληλων συστημάτων και άλλων αρχιτεκτονικών δομών για αυξημένη απόδοση.

Τα μοντέλα που μας ενδιαφέρουν είναι αλλά και συμβολικά ~~[illegible]~~ σύμβολα, κατασκευάζονται χρησιμοποιώντας σύμβολα, χρησιμοποιώντας προκαθορισμένους κανόνες σύνθεσης. Τα μοντέλα αυτά είναι κατά κάποιον τρόπο απλούστερα από κάποιο άλλο σύστημα αλλά όχι απόλυτα του οποίου συνιστούν πρότυπα - απλούστερα αλλά όχι [illegible] ένα μοντέλο είναι η ένδειξη μιας απλούστερης και σαφέστερης προ[illegible] διορισμένης κατάστασης πραγμάτων, με πρόθεση να διευκολυνθεί η παραγωγή περαιτέρω συνεπειών που μπορούν να επανεφαρμο[illegible] στο περισσότερο πολύπλοκο σύστημα [illegible] δοκιμαστικά στο σύστημα με σαφή αξιώματα πραγματικότητας. Περιγράφονται ένα σύστημα με σαφή αξιώματα που προδιαγράφουν τις ιδιότητες του μοντέλου και έτσι, κατά κάποιον τρόπο, το δημιουργούν, μπορεί κανείς να συναγάγει και άλλες συνέπειες από αυτά τα αξιώματα περί του [illegible] εφαρμόζοντας την μέθοδο της αυστηρής παραγωγής.

Ένα υπολογιστικό μοντέλο περιγράφει μια ιδεατή μηχανή για την οποία μπορούμε να γράψουμε λογισμικό. Είναι ιδεατή γιατί [illegible] αρχιτεκτονικές λεπτομέρειες και αλλαγές που οφείλονται σε καθαρά τεχνολογικές εξελίξεις αλλά είναι και αρκετά συγκεκριμένη ώστε να επιτρέπει την εξαγωγή συμπερασμάτων σχετικά με την απόδοση των προγραμμάτων που γράφονται γι' αυτήν. Θα [illegible] π.χ. RAM. Σε αυτό, το κόστος και η χρονοπολυπλοκότητα [illegible] στον αριθμό πράξεων. Αποτελείται από επεξεργαστή [illegible] μνήμη.

[illegible] υπολογιστικό μοντέλο της αριθμητικής, δηλαδή το μοντέλο.

[illegible] αλγορίθμων που επιτυγχάνουν καλές επιδόσεις [illegible] παρέχουν ακριβή αποτελέσματα.

Στον ΕΥ μας ενδιαφέρουν ο σχεδιασμός και ... αποδοτικών υπολογιστικών εργαλείων που βοηθούν στην πρακτική επίλυση των μαθηματικών μοντέλων της επιστήμης και της τεχνολογίας. Λαμβάνει υπόψη του τη φυσική και τα μαθηματικά του προβλήματος καθώς και το περιβάλλον υλοποίησης.

Οι πιο σημαντικές κατηγορίες χρήσεων τεχνικών του ΕΥ: ① μοντελοποίηση ② ανάλυση δεδομένων (εικόνας/ήχου) ③ υπολογιστική υποστήριξη γραφικών ④ εφαρμογές που απαιτούν απαντήσεις σε πραγματικό χρόνο.

Δύο κατηγορίες προβλημάτων: ① Αυτά που επιταχύνονται ικανοποιητικά μέσω των βελτιστοποιήσεων που επιτυγχάνει ο μεταφραστής ... ② αυτά που απαιτούν χρήση πληροφορίας υψηλότερου επιπέδου για την αποτελεσματική τους επιτάχυνση.

Υπολογιστικοί πυρήνες; παραδείγματα; Διαδικασίες (που μπορεί να είναι υποπρογράμματα ή τμήματα προγράμματος που συνήθως αντιστοιχούν σε συγκεκριμένες μαθηματικές διεργασίες ή υπολογισμούς) στις οποίες καταναλώνεται σημαντικό ποσοστό του χρόνου εκτέλεσης του προγράμματος εφαρμογής που μελετάμε. Γι' αυτό σημαντικό βήμα στη μείωση του χρόνου εκτέλεσης ενός προγράμματος εφαρμογής είναι η επιτάχυνση της εκτέλεσης των υπολογιστικών πυρήνων του. π.χ. (γεννήτρια τυχαίων αριθμών, ταχύς μετασχηματισμός Fourier, πολλαπλασιασμός μητρώου με διάνυσμα).

Πηγές σφαλμάτων στον ΕΥ: ① στα δεδομένα ② στη διακριτοποίηση των εξισώσεων ③ στην διακριτοποίηση των πραγματικών αριθμών και στις περιορισμένης ακρίβειας αριθμητικές πράξεις με αυτούς ... στον περιορισμό του πεπερασμένου πλήθους αριθμούς ④ στον περιορισμό για την εύρεση αποτελέσματος. π.χ. οι επαναληπτικές μέθοδοι για την εύρεση των ιδιοτιμών ενός μητρώου. Στις μεθόδους αυτές βασιζόμαστε στη σύγκλιση, αλλά στον υπολογιστή είμαστε υποχρεωμένοι να σταματήσουμε τις επαναλήψεις όταν κάποιος δείκτης μέτρησης σφάλματος γίνει αρκετά μικρός ⑤ σε δεδομένα ή ενδιάμεσα αποτελέσματα τα οποία δεν έχουν προσφερθεί από την μέθοδο επίλυσης.

Τα κύρια κριτήρια που χρησιμοποιούνται για την αξιολόγηση των εργαλείων του ΕΥ είναι α) η ακρίβεια των αποτελεσμάτων β) η ταχύτητα των υπολογισμών γ) το κόστος της ανάπτυξης της διαδικασίας.

///

Προμεταφορά: Τρόπος για την μείωση ή απόκρυψη της (ή καθυστέρησης που οφείλεται στην μεταφορά στοιχείων από την μνήμη. Αφού επιτρέπεται η σύγχρονη εκτέλεση εντολών μεταφοράς και αριθμητικής επεξεργασίας, είναι δυνατόν να μεταφερθούν στοιχεία, χρήσιμα σε μελλοντική αριθμητική επεξεργασία. Γίνεται α) Μέσω συστήματος: μεταφραστής εισάγει στον κώδικα εντολές για την κλήση στοιχείων που πρόκειται να χρειαστούν μελλοντικά. Η προμεταφορά μπορεί να υλοποιηθεί και από το runtime system. b) Μέσω άμεσων εντολών: Ο χρήστης μετατρέπει τον κώδικα εισάγοντας εντολές που καλούν στοιχεία που πρόκειται να χρειαστούν. Για να πετύχει η προμεταφορά, αν τα στοιχεία πρόκειται να χρειαστούν στον κύκλο T, η εντολή μεταφοράς τους στην κρυφή μνήμη πρέπει να δοθεί στον κύκλο T-τ, όπου τ είναι ο μέγιστος αριθμός κύκλων που απαιτούνται για την μεταφορά των στοιχείων από την μνήμη.

ΣήΛ: Αν στο (0,1) ισχύει $-\frac{d^2u}{dx^2}+b\frac{du}{dx}+cu=x$, $b,c>0$ και $u(0)=0$, $u(1)$ ... τότε το A στο τέλος $A=A^T$.

Λάθος: θα πρέπει για να ισχύει $A=A^T$ το A συμμετρικό. Όμως ξέρω ότι το A είναι τριδιαγώνιο

$$A = trid\left[-\frac{1}{h^2}-\frac{b}{2h},\ \frac{1}{h^2}+c,\ -\frac{1}{h^2}+\frac{b}{h^2}\right]$$

ΣήΛ: Το μέγεθος του βήματος Δt για την λύση της $\frac{du}{dt}=f(t,u)$ με πίσω Euler δεν μπορεί να γίνει πολύ μεγάλο λόγω αστάθειας.

Λάθος, γιατί η πίσω Euler είναι ευσταθής για οποιοδήποτε μέγεθος του Δt

ΣήΛ: Το Δt της εμπρός Euler καθορίζεται από το μέγιστο του σφάλματος διακριτοποίησης που μπορούμε να δεχτούμε για την λύση.

Λάθος: Μεγάλο μέγεθος Δt οδηγεί σε αστάθεια.

Μαθηματικό μοντέλο κόστους: Δεν έχει τρόπο να αποτιμήσει βάσει τις επιδόσεις που επιτυγχάνονται κάνοντας χρήση: α) της ιεραρχικής διαμόρφωσης της μνήμης για την επίτευξη μεγαλύτερης επίδοσης β) της παραλληλίας που μπορεί να διαθέτει το σύστημα επεξεργασίας γ) της πολύ ταχύτερης μείωσης του κόστους των στοιχειωδών πράξεων σε σύγκριση με τη μείωση του κόστους των μεταφορών μεταξύ μνήμης και επεξεργαστή.

Μέθοδος Horner: για υπολογισμό τιμής ενός $P(x) = a_0 + a_1 x \ldots a_n x^n \Rightarrow s = a_n$, for $i = n-1 : -1 : 0$, $s = s \cdot x + a_i$, end

Υπολογιστικό μοντέλο με ιεραρχία μνήμης: έχει επεξεργαστή, καταχωρητές, κρυφή-κύρια μνήμη. Χαρακτηριστικά: Ο επεξεργαστής προσπελαύνει την μνήμη με εντολές load/store. Χωρητικότητα κρυφής μνήμης K και κύρια $M \gg K$. Χρόνος εκτέλεσης των $\pm, \times, /$ είναι $\tau_{αρ\theta}$. Όταν η load αναφέρεται σε στοιχεία της κύριας μνήμης που πάει στους καταχωρητές θέλει χρόνο $\tau_{μετ}$. Αν το στοιχείο είναι στην κρυφή μνήμη τότε $\tau^{(ο)}_{μετ} \ll \tau_{μετ}$. Πριν την έναρξη του προγράμματος τα στοιχεία είναι στην κεντρική μνήμη. Η εκτέλεση τελειώνει όταν όλα τα στοιχεία αποθηκεύονται στην κεντρική μνήμη $\tau^{(ο)}_{μετ}$.

Horner: load x, a_n / $s = a_n$ / for $i = n-1 : -1 : 0$ / load a_i / $s = s \cdot x + a_i$ / end / store s

Με $\underline{Θ} = 2n$ $φ = n+3$.

$\underline{Θ}$: αριθμός πράξεων α.κ.υ (flops).

φ: αριθμός μεταφορών μεταξύ κύριας μνήμης και καταχωρητών ή κρυφής μνήμης.

$φ_{min}$: τον ελάχιστο αριθμό μεταφορών φ που θα απαιτούσε για την υλοποίηση, αν διαθέταμε απεριόριστη μνήμη σε όλα τα επίπεδα.

$T = T_{αρ\theta} + T_{μετ} = \tau_{αρ\theta} \underline{Θ} + \tau_{μετ} \cdot φ$

$T = T_{αρ\theta}\left(1 + μ \dfrac{\tau_{μετ}}{\tau_{αρ\theta}}\right)$ $μ = \dfrac{φ}{\underline{Θ}}$

Διαφορετικές υλοποιήσεις απαιτούν ίδιο αριθμό πράξεων και διαφέρουν στις μεταφορές.

Για έναν αρ. z οι πλησιέστεροι που περιέχουν το z είναι μοναδικοί: $z_- < z < z_+$ (6)

$$\left.\begin{aligned} z_- &= m \times b^{e-t} \\ z_+ &= (m+1) \times b^{e-t} \end{aligned}\right\} \quad z_+ - z_- = b^{e-t}$$

Το μέγιστο σφάλμα για το z είναι:

$$\left|\frac{z - fl(z)}{z}\right| \le \frac{b^{e-t}}{2} \Big/ b^{-1} \cdot b^{e} = \frac{b^{1-t}}{2} = u$$: μονάδα στρογγύλευσης

Στόχοι τυποποίησης IEEE: α) Συμβατότητα, αποτελεσματικότητα, μεταφερσιμότητα β) αντιμετώπιση ειδικών περιπτώσεων (π.χ. υπερχείλιση, υποχείλιση). γ) υλοποίηση μαθηματικών συναρτήσεων υψηλής ακρίβειας δ) ακριβής περιγραφή της κωδικοποίησης $m = d_0 + d_1 b^{-1} + \dots + d_{t-1} b^{-(t-1)}$

Βασικές προδιαγραφές προτύπου IEEE: ① Formats: single, double, single extended, double extended. ② 4 είδη στρογγύλευσης: προς πλησιέστερο, προς ζυγό, προς $\pm\infty$, προς 0 (αποκοπή) ③ ειδικοί αριθμοί: Συνδυασμοί από bits χρησιμοποιούνται ως "ειδικοί αριθμοί" για κωδικοποίηση αποτελέσματος πράξεων όπως $x/0$, $0/0$, $\sqrt{y}$ με $y<0$. Υιοθετεί την ύπαρξη αριθμών -0, $\pm\infty$, NaN. ④ Επιτρέπει τη χρήση υποκανονικοποιημένων αριθμών, όταν τ' αποτέλεσμα πράξης είναι μικρότερο του ελάχιστου κανονικοποιημένου αριθμού ⑤ Αριθμητικές εξαιρέσεις: inexact, overflow, underflow, division b 0

π.χ. $0/0$, $0\times\infty$, $\sqrt{-1} \to$ NaN   overflow $\to \pm$Inf

finite number/0 $\to \pm$Inf   underflow $\to$ subnormal numbers

FMA επιτρέπει την εκτέλεση $z + x\cdot y$ στον ίδιο περίπου χρόνο με την $x*y$ ή $x+y$. Συνεπάγεται μόνο ένα σφάλμα στρογγύλευσης.

π.χ. $fl(z + x\cdot y) = (z + x\cdot y)(1+\delta)$ $|\delta| \le u$

αντί: $fl(z + xy) = (z + xy(1+\delta_1))(1+\delta_2)$

Η απώλεια πληροφορίας οφείλεται κυρίως: α) στην μαθηματική μοντελοποίηση και στις απλοποιήσεις που γίνται χάριν επιλυσιμότητας β) Στη διακριτοποίηση και τα σφάλματα αποκοπής που προκύπτουν από αυτή. π.χ. αντικατάσταση παραγώγου με διαιρεμένες διαφορές γ) Στα σφάλματα στρογγύλευσης λόγω χρήσης αριθμητικής κινητής υποδιαστολής ~~[illegible]~~ (5

$y = \pm m \times b^{e-t}$

Κανονικοποίηση και κρυμμένο bit: Μπορούμε να αναπαραστήσουμε τον ίδιο αριθμό με περισσότερους τρόπους π.χ. $y = 0.d_1 0000 \times b^e$ και $y = 0.0 d_1 0000 \times b^{e+1}$. Έτσι έχουμε έλλειψη μοναδικότητας. Γι' αυτό κανονικοποιούμε το σύστημα και χρειάζομαστε το 1ο ψηφίο του (στη βάση b) να είναι μη μηδενικό δηλ. $y \in R(F) \Rightarrow d_1 \neq 0$. Για $b = 2$, $d_1 \neq 0 \Rightarrow d_1 = 1$ επομένως μπορούμε να μην το αποθηκεύουμε, αλλά οι αλγόριθμοι που διαχειρίζονται τους αριθμούς αυτούς πρέπει να το λαμβάνουν υπόψη τους. Αυτή είναι η τεχνική του κρυμμένου bit και χρησιμοποιείται ευρύτατα σήμερα γιατί για δεδομένο μήκος του αριθμού κ.υ. δίνει ένα ακόμη bit χωρίς κόστος.

$y = \pm b^e \times .d_1 d_2 \ldots d_t$ όπου $0 \leq d_i \leq b-1$, $d_1 \neq 0$

Το έψιλον της μηχανής, που συμβολίζεται με ϵ_m, είναι η απόσταση από το 1 ως τον αμέσως μεγαλύτερο α.κ.υ. δηλαδή $\epsilon_m = \delta(1, 1^+) = 1^+ - 1$. Σε κάθε περίπτωση το ϵ_m δείχνει την ελάχιστη δυνατή ~~[illegible]~~ διακριτότητα του συστήματος α.κ.υ. Αν το σύστημα χρησιμοποιεί $t-1$ δυαδικά ψηφία μετά την υποδιαστολή τότε το επόμενο του 1 είναι το $1 + 2^{1-t} \Rightarrow \epsilon_m = 2^{1-t}$

Οι α.κ.υ. δεν είναι ισοκατανεμημένοι στον άξονα. Επομένως οι αποστάσεις δεν είναι ομοιόμορφες αλλά μεγαλώνουν κατά b.

π.χ. $|m b^{e-t} - (m+1) b^{e-t}| = b^{e-t}$

$$|m b^{e+1-t} - (m+1) b^{e+1-t}| = b^{e+1-t}$$

Τι σημαίνει αλγόριθμος πίσω σταθερός; Αν τρέξουμε (ε τον αλγόριθμο τότε το υπολογισμένο αποτέλεσμα $\hat{f}$ με τα δεδομένα του προβλήματος (και θα περιέχει σφάλματα) μπορεί να γίνει ίσο με το θεωρητικό αποτέλεσμα και θ προέκυπτε αν εκτελούσαμε τις πράξεις με αριθμητική ά ρης ακρίβειας χρησιμοποιώντας στοιχεία εισόδου και μπορ να είναι λίγο διαφορετικά από τα ακριβή. Η απόσταση τ παραλλαγμένων δεδομένων από τα ακριβή είναι ένα μέτρ της πίσω ευστάθειας του αλγορίθμου

Παράδειγμα Blas-3. Πλεονέκτημα σε σχέση με Blas-2

Πολλαπλασιασμός μητρώων. Μεγαλύτερη τοπικότητα, αν αξιοποιηθεί από την υλοποίηση.

Η LU φθηνότερη (ταχύτερη, μικρότερο κόστος $= \frac{2n^3}{3} + O(n$ ενώ QR πίσω ευσταθής ανεξαρτήτως δεδομένων.
$\shortparallel$
$\frac{4n^3}{3} + O(n^2)$

Έστω $A = \begin{bmatrix} 4 & -8 & 1 \\ -6 & 5 & 7 \\ 0 & -10 & -3 \end{bmatrix}$ α) χωρίς οδήγηση το στοιχείο στη θέση (1,1) δηλ το 4

β) μερική οδήγηση: το μέγιστο κατά απόλυτη τιμή της 1ης στήλης $|-6|$ άρα οδηγός το -6

γ) πλήρης οδήγηση: το μέγιστο κατά απόλυτη τιμή του πίνακα $|-10|$ άρα οδηγός το -10.

Πότε ένα σύστημα α.κ.υ. ικανοποιεί την συνθήκη ακριβ. στρογ.: όταν το αποτέλεσμα οποιασδήποτε στοιχειώδους πράξης στο σύστημα και σε δεδομένα x, y που είναι α.κ.υ είναι ο α.κ.υ που θα προέκυπτε αν εκτελούσαμε την πράξη με αριθμητική άπειρης ακρίβειας και μετά στρογγυλεύαμε.

$fl(x \odot y) = x \tilde{\odot} y$, $\tilde{\odot}$ υλοποίηση της $\odot$ στο σύστημα α.κ.υ

Ποια σχέση έχει το ϵ_M με την u: Η u είναι το μέγιστο σχετικό σφάλμα που γίνεται μετά από στρογγύλευση επόμενως εξαρτάται από την στρατηγική στρογγύλευσης. Αν είναι προς τον πλησιέστερο θα είναι το μισό της μέγιστης σχετικής απόστασης δύο διαδοχικών α.κ.υ. Το ϵ_M είναι η απόσταση από το 1 στον αμέσως μεγαλύτερο α.κ.υ και είναι ανεξάρτητο του συστήματος στρογγύλευσης. Στη περίπτωση στρογγύλευσης προς τον πλησιέστερο τότε $\epsilon_M = 2u$

Σ ή Λ ; $\forall x \in \mathbb{R}^n$ ισχύει $\frac{1}{n}\|x\|_1 \leq \|x\|_\infty \leq \|x_1\|$ — ($N \neq N \leq 2$) μεγαλύτερο

Σωστό: $\|x\|_\infty = \max_{i=1:n}\{|\xi_i|\}$ και $\|x\|_1 = \Sigma_i|\xi_i| \leq n\|x\|_\infty$

επίσης $\|x\|_1 = \Sigma_i|\xi_i| \geq \|x\|_\infty$ αφού το $\|x\|_\infty$ εμφανίζεται ως ένας από τους μη αρνητικούς όρους του αθροίσματος για το $\|x\|_1$

$A = [100, 1; 0, -10]$ και $B = [1, 0; 0, 0]$ → $C = A./B$?

$C = [100, \mathrm{Inf}; \mathrm{NaN}, -\mathrm{Inf}]$ $\quad 1/(1/0) \to 0 \quad 1/1/0 \to \mathrm{Inf}$

$\max(\mathrm{Inf}, 4) \to \mathrm{Inf}$

ΝΔΟ η u είναι α.κ.υ: $u = b^{1-t}/2$. Αν $b=2$ όπως στο IEEE τότε $u = 2^{-t}$ και μπορεί να αναπαρασταθεί. Αν $b > 2$ τότε $b^{1-t}/2 = (b/2) \times b^{-t}$ συνεπώς αν b άρτιος, η αναπαράσταση είναι εφικτή.

temp2 = realmin/2; temp = 2 * temp2; boole(temp==realmin)

boole = 1 λόγω ότι το IEEE υποστηρίζει βαθμιαία υποκανονικοποίηση.

x = realmin; y = $\frac{x}{2}$; if (y==0) y = 2·x, else y = y+1; end

έχουμε υποκανονικοποίηση άρα το $y = \frac{x}{2}$ δεν δίνει μηδέν εκτελείται το else $y = y+1$ επειδή όμως y = realmin είναι πολύ μικρότερο από το ϵ_M τότε $y = 1$

$b^{1-t} - b^1 \cdot \frac{1}{2} = \frac{b}{2} = \left(\frac{b}{2}\right) \cdot b^{-t}$

Για κάθε αντιστρέψιμο τετραγωνικό μητρώο A υπάρχει παραγοντοποίηση LU. Σ ή Λ? Λάθος. Π.χ. το $A=\begin{bmatrix}0&1\\1&0\end{bmatrix}$ δεν μπορεί ως LU γιατί τότε $l_{11}u_{11}=0$ άρα $u_{11}=0$ οπότε το U είναι αντιστρέψιμο και να δε γίνεται γιατί τότε και το LU θα ήταν μη αντιστρέψιμο

Το αντίστροφο ενός τυχαίου αντιστρέψιμου κάτω διδιαγώνιου είναι επίσης κάτω διδιαγώνιο. Σ ή Λ? Λάθος: έστω η 1η στήλη $s=[s_1,s_2,s_3]^T$ του αντιστρόφου του $T=\begin{bmatrix}1&0&0\\t_{21}&1&0\\0&t_{32}&1\end{bmatrix}$ όπου $t_{21}t_{32}\neq 0$. Ισχύει $Ts=e_1$ και αρκεί να δείξουμε ότι $s_3\neq$ Εκ κατασκευής θα πρέπει να ισχύει $s_1=1$, $t_{21}+s_2=0$, $t_{32}s_2+s_3=0$. Επομένως αν $s_3=0$ τότε $s_2=0$. Επομένως $t_{21}=$ που είναι αδύνατο.

Σ ή Λ? Το αντίστροφο σ.θ.ο τριδιαγώνιου μητρώου είναι τριδιαγώνιο: Λάθος. Αν είναι σ.θ.ο μπορεί να χρησιμοποιη- τη Cholesky, $A=LL^T$ όπου L κάτω τριγωνικό. Το A είναι τριδιαγώνιο άρα και το L θα είναι κάτω διδιαγώνιο. Επομέ- ως $A^{-1}=L^{-T}L^{-1}$. Το L^{-1} είναι κάτω τριγωνικό αλλά όχι κάτω διδιαγώνιο. Ομοίως και το L^{-T} θα είναι άνω τριγωνικό αλλά όχι διδιαγώνιο. Επομένως A^{-1} είναι το γινόμενο ενός άνω και ε- διδιαγώνιο. ως κάτω τριγωνικού μητρώου, τα οποία θα είναι πλήρες, στην γενική περίπτωση.

Γιατί δε χρειάζεται οδήγηση για λύση με τα L, U? Επειδή το L έχει το μέγιστο στο μέτρο στοιχείο του στη δια- γώνιο, επομένως η μερική οδήγηση δε θα είχε αποτέλεσμα. Ομοίως, επίσης η μερική... Το U είναι άνω τριγωνικό άρα το μέγιστο στο μέτρο στοι- χείο από εκείνα που βρίσκονται στις θέσεις j ως n της στήλης j θα είναι πάντα στη διαγώνιο, άρα η μερι- κή οδήγηση δε θα είχε αποτέλεσμα.

IEEE απλή ακρίβεια μέγιστο σχετικό σφάλμα αποκοπής. Ποιο το u?

$$u=\max_x \frac{|x-fl(x)|}{|x|} = \frac{|0.0\ldots 1|}{1.0} = 2^{-t}+2^{-t-1}+\ldots = 2^{-t}(1+2^{-1}+\ldots)$$

$$= 2^{-t}\,\frac{1}{\ldots} = 2^{-t+1} = 2^{-53+1} = 2^{-52}$$

A=LU υπολογίζει τη L,U από το A. Άρα $\Phi_{min} = 2n^2$ για τα load-store. Ακόμα η LU απαιτεί $\Omega = \frac{2n^3}{3}$. Άρα

$$\mu_{min} = \frac{2n^2}{\frac{2n^3}{3}} = \frac{6n^2}{2n^3} = \frac{3}{n} \Rightarrow O\left(\frac{1}{n}\right)$$

άρα υπάρχει τοπικότητα. Άρα η LU τετραγωνικού μητρώου παρέχει τη δυνατότητα καλής απόδοσης σε συστήματα με ιεραρχική μνήμη.

$\frac{Inf}{Inf}$ = NaN Σ ή Λ? → Σωστό. Δεν ορίζεται

$\frac{realmin}{2}$ == 0 - Σ ή Λ? → Λάθος λόγω βαθμιαίας υποκανονικοποίησης.

$\frac{0}{0}$ == Inf Σ ή Λ? → Λάθος. δεν ορίζεται

---

realmax+1.0 == ∞ Σ ή Λ? Λάθος: ο εκθέτης του realmax είναι 1023. Για να προστεθεί το y=1.0 πρέπει να το διαιρέσουμε ώστε να γίνει ο εκθέτης του 1023. Κάθε διαίρεση με 2 αντιστοιχεί σε ολίσθηση του μη σημαντικού bit της ουράς του y κατά μια θέση προς τα δεξιά. Το μήκος της ουράς στην IEEE διπλής ακρίβειας είναι 52. Άρα μετά από 53 ολισθήσεις η ουρά θα μηδενιστεί. Επομένως realmax+1.0 = realmax.

---

5 σημαντικά προβλήματα της Υ.Γ.Α. α) Επίλυση γραμμικών συστημάτων β) υπολογισμός συναρτήσεων μητρώων γ) επίλυση προβλημάτων ιδιαζουσών τιμών δ) επίλυση προβλημάτων ελαχίστων τετραγώνων ε) επίλυση προβλ. ιδιοτιμών.

Αν $A \in R^{n \times n}$ αντιστρέψιμο το $Ax=b$ έχει μοναδική λύση.

συμμετρικό $A=A^T$ Θετικά ορισμένο: $x^TAx \geq 0$

Toeplitz: τα στοιχεία ανά διαγώνιο είναι ίσα μεταξύ τους.

Ορθογώνιο: $AA^T = A^TA = I$

Μεταθετικά $P \cdot A$ → ο A αλλάζει γραμμές $A \cdot P$ → αλλάζει στήλες

$P \cdot P^T = I$

Σε ποιο διάστημα περιέχονται περισσότεροι α.κ.υ (1Σ) [1,2] ή [2,3] Η πυκνότητα των ακυ μειώνεται καθώς μεγαλώνουν (για διάστημα δοθέντος μήκους). Γνωρίζουμε ότι η απόλυτη απόσταση δύο διαδοχικών ακυ είναι $2^{1-t} \times 2^{e}$. Επομένως στο [1,2] είναι 2^{1-t}, ενώ στο [2,3] είναι $2 \times 2^{1-t}$. Άρα το [1,2] περιέχει πιο πολλούς. Σε βιβλίο

Τι είναι σωστό για το $1 + eps - \frac{eps}{2}$?

α) 1 β) 1+eps γ) κανένα

Απάντηση: το (α) ισχύει $fl(1+eps) = 1+eps$. Η αφαίρεση μας πάει στο μέσον του 1 με τον επόμενο 1^+. Λόγω στρογγύλευσης προς τον πλησιέστερο ζυγό εάν ισαπέχει στρογγύλευση προς τον πλησιέστερο ζυγό ... θα βρεθούμε κατά μέ από τους περιβάλλοντες α.κ.υ. θα βρεθούμε κατ' ανάγκη στο 1. δηλ $fl\left(1+eps-\frac{eps}{2}\right) = 1$

ξετύλιγμα βρόχου με b=3

έστω for i=1:n , y(i) = a·x(i) + y(i), end

⇒ έστω r=rem(n,m) επιστρέφει το υπόλοιπο της διαίρεσης των θετικών ακεραίων n, m. δηλ. n=p·m+r, $0 \le r \le m-1$ τότε

```
r=rem(n,3)
for i=1:r, y(i) = a·x(i) + y(i); end
for i = r+1 :3 : n
  y(i) = a·x(i) + y(i);
  y(i+1) = a·x(i+1) + y(i+1);
  y(i+2) = a·x(i+2) + y(i+2);
end.
```

ξετύλιγμα b=5

έστω s=0.0; for j=1:n, s = s + x(i)y(j); end.

λύση

```
m = rem(n,5); s=0.0;
for j=1:m, s=s + x(j)*y(j); end
for j=m+1 : 5:n
s=s + x(j)y(j) + x(j+1) y(j+1) + x(j+2) y(j+2) + x(j+3) y(j+3) + x(j+4)·y(j+4);
```

Οι αποκλίσεις του υπολογιστικού μοντέλου σε σχέση με (1)
την πραγματικότητα: (α) Οι πράξεις α.κ.υ εκτελούνται με ίδιο
κόστος ενώ το κόστος των άλλων πράξεων δεν λαμβάνεται υπ
ψην (β) χρησιμοποιεί ένα μόνο επίπεδο κρυφής μνήμης (γ) το κό
στος μεταφοράς της LOAD είναι ίσο με της STORE (δ) Δεν λαμ
νεται υπόψην η επικάλυψη πράξεων μεταφοράς και αριθμη
τικών πράξεων (ε) Ο προγραμματιστής δεν ελέγχει ποια στοι
χεία πηγαίνουν και διατηρούνται στην κρυφή μνήμη (στ) σε
περίπτωση miss, δεν μεταφέρεται μόνο ένα στοιχείο στην
φή μνήμη αλλά ολόκληρη γραμμή κρυφής μνήμης.

Στο IEEE αν στρογγυλεύσουμε προς το πλησιέστερ
αιου και μ είναι η μονάδα στρογγύλευσης, ισχύει $fl(x) =$
$= x(1+\delta)$, $\delta \leq \mu$ Σ ή Λ ; Λάθος: $fl(x) = x(1+\delta)$, όπου $|\delta|$
αλλιώς το δ θα μπορούσε να είναι πάρα πολύ μεγά
σε απόλυτη τιμή

Σ ή Λ ; Αν ένας αλγόριθμος είναι πίσω σταθερός, τότε
μικρές αλλαγές στα στοιχεία εισόδου οδηγούν σε μικρές αλ
λαγές στο υπολογιστικό αποτέλεσμα: Λάθος, η πίσω ευστάθ
εξασφαλίζει μόνο ότι τα υπολογισμένα αποτελέσματα εί
ακριβώς ίδιο με το αποτέλεσμα που θα είχαμε αν χρησι
μοποιούσαμε αριθμητική άπειρης ακρίβειας με στοιχεία εισό
λίγο παραλλαγμένα. Για να εξασφαλιστεί το ζητούμενο τη
ερώτησης χρειάζεται το πρόβλημα (και όχι ο αλγόριθμος
να έχει μικρό δείκτη κατάστασης)

Να εξηγήσετε τι αποκλίσεις κάνουμε στο υπολογιστι
κό μοντέλο όταν λέμε: $T = T_{αρ} + T_{μετ}$. Η απόκλιση
είναι ότι στο υπολογιστικό μοντέλο δεν υπάρχει επικά
ψη πράξεων μεταφοράς και αριθμητικής όπως στις σύγ
χρονες αρχιτεκτονικές.

- **DOT**: $\sigma = x^T y$ $\quad$ SAXPY: $y \leftarrow y + ax$ $\quad$ (12)

ανανέωση 1ης τάξης: $C \leftarrow C + ab^T$ $\quad$ MV: $y \leftarrow y + Ax$

- $A = (I - uv^T)x$;

$A = Ix - uv^Tx = Ix - \underbrace{(v^Tx)}_{\text{αριθμός}}u$

- $A, B \in \mathbb{R}^{n\times n}$ κάτω τριγωνικά: Αλγόριθμος για το $C = A\cdot B$ και ακριβές κόστος;

$C(i,j) = A(i,:) * B(:,j) = A(i,1:i) * B(1:i,j)$

```
for i=1:n
    for j=1:i
        C(i,j) = A(i,j:i) * B(j:i,j);
    end
end
```

$$\Theta = \sum_{i=1}^{n}\sum_{j=1}^{i}\left[2(i-j+1)-1\right] = 2\sum_{i=1}^{n} i^2 - \sum_{i=1}^{n} i = \frac{2n^3}{3} + \frac{n^2}{2} - \frac{n}{6}$$

- $B = B + (x\cdot y^T)^p$: Ν.Δ.Ο. flmin ανεξάρτητο του p.

Για το xy^T θέλω n^2 πράξεις.

Για το $(xy^T)^p$ θέλω $(p-1)(2n^3-n^2)$ πράξεις.

Άρα $\Theta = (p-1)(2n^3-n^2) + n^2 + n^2$

Φflmin $= 2n + 2n^2 + 1$

Μπορώ όμως $B = B + \underbrace{(y^Tx)}_{2n-1+1}{}^{p-1} \cdot x y^T$

n $\quad$ n^2 $\quad$ n^2 $\quad$ $\Theta = 2n^2 + 3n$

13

έστω $r_n(x) = \prod_{j=1}^{n}(x - r_j)$ αλγόριθμος υπολογισμού των συντελεστών της δυνατομορφής του $P_n(x)$.

$\Rightarrow a(2)=1;\ a(1) = -r(1);\ a(3:n+1)=0;$

for $j=2:n$

$t(2:j+1) = a(1:j),\ t(1)=0$

$a(1:j+1) = t(1:j+1) - r(j) * a(1:j+1)$

end

$y \leftarrow y + Ax \quad A \in \mathbb{R}^{n\times n} \quad x, y \in \mathbb{R}^n \quad \Phi, \underline{O} = ?$

$\Rightarrow \Phi = \Phi_L + \Phi_S. \quad \Phi_L = n^2 + 2n$

$\underset{n\times n}{[\ ]}\underset{n\times 1}{[\ ]} = \underset{n\times 1}{[\ ]} \quad (n+n-1)n = 2n^2 - n$

$\Phi_A = n^2 + 3n$

$\underline{O} = 2n^2$

$\underset{n\times 1}{[\ ]} + \underset{n\times 1}{[\ ]} = \underset{n\times 1}{[\ ]} \quad \underline{O} = n$

$\Phi_{STORE} = n$

$A \leftarrow A + xy^T$

$\underset{\substack{n\times 1 \\ x}}{[\ ]}\underset{\substack{1\times n \\ y^T}}{[\ ]} = \underset{n\times n}{[\ ]} \quad n^2$

$\Phi_L = n^2 + 2n$

$\underset{n\times n}{[\ ]} + \underset{n\times n}{[\ ]} = n^2$

$\Phi_S = n^2$

$\Phi_A = 2n^2 + 2n$

$\underline{O} = 2n^2$

$y = (A^2 - I)x$ χωρίς LOAD.

$\underset{n\times n}{[\ ]}\cdot\underset{n\times n}{[\ ]} = 2n^3 - n^2 \quad [n\times n] - [n\times n] = n^2 \quad [n\times n][n\times 1]$

$= 2n^2 - n.$

$\underline{O} = 2n^3 + 2n^2 - n \quad \Phi = n^2 + 2n +$

$O(n^3) \xleftarrow{\text{πράξεις}} \quad O(n^2)$

$A(Ax)$ Τ.χ. διανύσματα βήμ-

(16)

$\Delta = \Delta + x^T (x y^T)^r$

$\Delta = \Delta + x^T (x y^T x y^T \dots x y^T) = \Delta + x^T x (y^T x y^T \dots x) y$

$x^T x : \underline{2n-1}$ [·]

$y^T x : \underline{2n-1}$ [·]

$x^T x \cdot (y^T x)^{p-1} = \underline{2} \cdot y^T = \underline{n}$

$\Delta + \underline{n} \longrightarrow n$ άρα $\Delta = n$

$\phi_{min} = 4n + 1 \rightarrow$ για το p $\quad \underline{o} = 6n$

$\mu_{min} = \frac{2}{3} + O(\frac{1}{n})$

$A \in R^{10 \times n}$ και $b \in R^n$ και $y \leftarrow Ab$

$$A = \underset{10\times n}{\begin{bmatrix} 1 & \cdots & n \\ \vdots & & \\ 10 & & n \end{bmatrix}} \cdot b = \underset{n \times 1}{\begin{bmatrix} 1 \\ \vdots \\ n \end{bmatrix}} = \underset{10\times 1}{\begin{bmatrix} \ \end{bmatrix}}$$

$\phi_L = 10n + n$

$\phi_S = 10$

$\phi_{OA} = n + 1$

Σε μνήμη $O(1)$.

```
LOAD y
 for j = 1:n
     LOAD b(j)
     for i = 1:10
      LOAD A(i,j)
       y(i) = y(i) + A(i,j) · b(j)
     end
  end
 Store y.
```

$z = y + A^k x$

$z = y + A \dots A x$

$\Phi_{min} = n^2 + 3n + 1$

$\Theta = (2n^2 - n)k + n$

```
Load  A, x, y, k
      for i = 1 : k
          x = A * k;
      end.
      z = y + x
      store z;
```

υλοποίηση.

$y = (A + uv^T)x$

$Ax + uv^T x \Leftrightarrow Ax + (v^T x) u$ αριθμός

$2n^2 - n$

$n + n - 1$

$+ n \quad + n$

$= 2n^2 + 3n - 1$

άρα $\Theta = 2n^2 + 3n - 1$

$\Phi_{min} = 4n + n^2$

$\mu_{min} = \frac{\Phi_{min}}{\Theta}$

υλοποίηση με $O(n)$ μνήμη

```
Load  x, u, v
   y = (v^T x) u
   for i = 1 : n
       Load A(i,:)
       y = y_i + A(i,:) · x
   end
   Store y
   end.
```

φέρνει γραμμή-γραμμή

$\Phi_{min} = 2n^2 + 2n + 1$

το $(xy^T)^P = x y^T x y^T \dots x y^T$

$2n^2 - n$

$B = B + (xy^T)^P$

για το $y^T x$: $2n - 1 = \Psi$

άρα $(\Psi^{(P-1)} x) = 1 + n$

$\Theta = 2n^2 + 3n$

$\mu_{min} = \frac{\Phi}{\Theta} = \frac{2n^2 + 2n + 1}{2n^2 + 3n}$

$(\dots)^{P-1} \, y^T$

Πρέπει να είναι αντιστρέψιμο το ~~A~~ για να μπορεί (18) να λυθεί με LU ή Cholesky. Βρίσκω ιδιοτιμές του A και βλέπω αν μπορεί να περιέχονται στην ένωση δίσκων

$$|\lambda - 0| \leq 2 \quad , \quad |\lambda - 0| \leq 1$$

$$-2 \leq \lambda - 0 \leq 2 \qquad -1 \leq \lambda \cancel{\leq} 0 \leq 1$$

$$\lambda > 0$$

άρα ιδιοτιμές θετικές αφού είναι συμμετρικό ($A = A^T$). τότε είναι σ.θ.ο. άρα χρησιμοποιώ cholesky που είναι ~~πιο γρήγορη από LU.~~

Euler ΕΜΠΡΟΣ

$\Longrightarrow$ $u_1(0) = 2 \quad u_2(0) = 1 \qquad \dfrac{du(t)}{dt} = -A\,u(t) \quad A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$

$\Delta = \frac{1}{2} = h$ βρείτε το $T = 2.0$ $\qquad u = \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix}$

$u(t+h) - u(t) = -hAu(t) \Leftrightarrow u(t+h) = u(t) - hAu(t) \Leftrightarrow$

$u(t+h) = (I - hA)\,u(t)$ $\leftarrow$ υπάρχει ο μοναδιαίος πίνακας I
Προσοχή το u(t) το βάζω δεύτερο. Αν το βάλω μπροστά ΚΑΗΚΑ!!!

$(I - hA) = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix}$ $\qquad \begin{bmatrix} u_1(0) \\ u_2(0) \end{bmatrix}$

Ξεκινάω πάντα από αυτή τη χρον. στιγμή

$t=0 \quad U(0,5) = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix} \cdot U(0) = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1 \end{bmatrix}$

$t = 1/2 \quad U(1) = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix} \cdot U(0,5) = \begin{bmatrix} 1/2 \\ 1/4 \end{bmatrix}$

$t = 1 \quad U(1,5) = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix} U(1) = \begin{bmatrix} 1/4 \\ 1/8 \end{bmatrix}$

$t = 1{,}5 \quad U(2) = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix} U(1,5) = \begin{bmatrix} 1/8 \\ 1/16 \end{bmatrix}$

t: χρον. στιγμή

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

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

με $AU = F$. A τριδιαγώνιο με στοιχεία γ_j, a_j, b_j στις θέσεις $j-1$, j, $j+1$ της γραμμής j

$A = \text{trid}_n[\gamma_j, a_j, b_j]$

$$a_j = \left(\frac{2}{h^2} + c_j\right) \quad b_j = \left(-\frac{1}{h^2} + \frac{b_j}{2h}\right) \quad \gamma_j = \left(-\frac{1}{h^2} - \frac{b_j}{2h}\right)$$

Έχω την $-x^2 \frac{d^2u(x)}{dx^2} + 2\frac{du(x)}{dx} + x^2u(x) = x$

$x \in [1,2]$, $u(1) = 0$ $u(2) = 1$ $n = 4$ εσωτερικά

$h = \frac{2-1}{n+1} = \frac{1}{5}$ $U_0 = u(1) = 0$

$U_0 \; U_1 \; U_2 \; U_3 \; U_4 \; U_5$ $U_5 = u(2) = 1$

$$-x^2u'' + 2u' + x^2u = x$$

$$-x\left(\frac{U_{j-1} + U_{j+1} - 2U_j}{h^2}\right) + 2\left(\frac{U_{j+1} - U_{j-1}}{2h}\right) + x^2U_j = x$$

$x_i = \cancel{1} + ih$ $\Leftrightarrow$ $x_0 = 1$ $x_1 = 1{,}2$ $x_2 = 1{,}4$ $x_3 = 1{,}$
$x_4 = 1{,}8$ $x_5 = 2{,}0$

για $i = 1 \ldots$
για $i = 2 \ldots$ } αλλάζει και το x_i προσοχή

$$\begin{bmatrix} - & - & 0 & 0 \\ - & - & - & 0 \\ 0 & 0 & - & - \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{bmatrix} = \begin{bmatrix} \\ \\ \end{bmatrix}$$ ← δευτεροβάθμια εξίσωση

$U_1 \; U_2 \; U_3 \; U_4$ } → συντελεστές αγνώστων

Neumann

Μας δίνει την τιμή της παραγώγου σε κάποιο άκρο τότε

$U_0 \quad U_1 \quad U_2 \quad U_3 \quad U_4 \quad U_5$

U_{-1} → το προσθέτω

$$U_j' = \frac{U_{j+1} - U_{j-1}}{2h} \Rightarrow U_0' = \frac{U_1 - U_{-1}}{2h}$$

γνωστά

$$\Rightarrow U_1 - U_{-1} = 2hU_0' \Rightarrow$$

$$U_{-1} = U_1 - 2hU_0'$$

το αντικαθιστώ

σφάλματα στρογγύλευσης

$$z = \frac{x}{\sqrt{x^2+y^2}}$$

$t_1 \leftarrow fl(x*x) \qquad x^2(1+\delta_1)$

$t_2 \leftarrow fl(y*y) \qquad y^2(1+\delta_2)$

$t_3 \leftarrow fl(t_1+t_2) \qquad (t_1+t_2)(1+\delta_3)$

$t_3 \leftarrow fl(\sqrt{t_3}) \qquad \sqrt{t_3}(1+\delta_4)$

$t_1 \leftarrow fl\left(\frac{x}{t_3}\right) \qquad \frac{x}{t_3}(1+\delta_5)$

$$z = \frac{x}{t_3}(1+\delta_5) = \frac{x\,(1+\delta_5)}{\sqrt{(x^2(1+\delta_1)+y^2(1+\delta_2))(1+\delta_3)}(1+\delta_4)}$$

$|\delta_j| \leq u$

ισχυρισμός?

και πως
τα $(1+\delta_j)$
είναι κάτω?

ΝΑΙ

(19)

Πίσω Euler

$T=2\ ?\quad \Delta t = 2 = h \qquad A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \qquad \left. \begin{matrix} u_1(0)=2 \\ u_2(0)=1 \end{matrix} \right| V_0$

$$u(t+h) - u(t) = -hAu(t+h) \iff$$
$$-u(t) = -hAu(t+h) - u(t+h) \iff$$
$$u(t) = [I + hA]\, u(t+h)$$

Αν είχε άλλο Δt

$$[I + hA] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \cdot 2 = \begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix}$$

Για t=0

$$U(0) = \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix} \begin{bmatrix} U_1(2) \\ U_2(2) \end{bmatrix} \iff$$

$$\left. \begin{matrix} 2 = 5u_1(2) - 2u_2(2) \\ 1 = -2u_1(2) + 5u_2(2) \end{matrix} \right| \begin{matrix} u_1(2) = \dots \\ u_2(2) = \dots \end{matrix}$$

Για να τείνει στο 0 πρέπει οι απόλυτες τιμές των ιδιοτιμών του μητρώου να είναι <1. δηλαδή η φασματική ακτίνα του μητρώου να είναι φραγμένη αυστηρά από το 1. Οι ιδιοτιμές του $I-hA$ θα είναι $1-h\lambda(A)$ πρέπει $|1-h\lambda(A)|<1$ και επειδή οι ιδιοτιμές είναι πραγματικές (συμμετρικό μητρώο) θα πρέπει:

$$-1 < 1 - h\lambda(A) < 1 \iff -2 < -h\lambda(A) < 0 \iff$$
$$2 < h\lambda(A) < 0 \iff$$
$$2 > \frac{h}{\lambda(A)} > 0 \text{ άρα } h < 2\lambda.$$

Βρίσκω ιδιοτιμές του $I-hA$ $\left| \begin{matrix} \lambda_1 = 1/2 \\ \lambda_2 = 5/2 \end{matrix} \right| h < 1$

2

Ανάλυση σφάλματος με FMA !.

s=0; for j=1:4; s=s+x(j)·y(j); end;

εμπρός σφάλμα?

$$\hat{s} = (((x_1y_1)(1+\delta_1) + x_2y_2)(1+\delta_2) + x_3y_3)(1+\delta_3) + (x_4y_4)\cdot(1+\delta_4) = x_1y_1(1+\delta_1)(1+\delta_2)(1+\delta_3)(1+\delta_4) + (x_2y_2)(1+\delta_2)\cdot(1+\delta_3)(1+\delta_4) + (x_3y_3)(1+\delta_3)(1+\delta_4) + (x_4y_4)(1+\delta_4) =$$

$$= (x_1y_1)(1+\theta_4) + (x_2y_2)(1+\theta_3) + (x_3y_3)(1+\theta_2) + (x_4y_4)(1+\theta_1)$$

$$|\hat{s}-s| = |x_1y_1\theta_4 + x_2y_2\theta_3 + x_3y_3\theta_2 + x_4y_4\theta_1| \le$$

$$\le |x_1y_1||\theta_4| + |x_2y_2||\theta_4| + |x_3y_3||\theta_4| + |x_4y_4||\theta_4| \le$$

$$\le \frac{4u|s|}{1-4u} = \gamma_4|s|$$

Πίσω ευστάθεια Horner?

```
s_n = a_n
for k = n-1:-1:0
  s_k = x s_{k+1} + a_k
end
```

$$\hat{s}_{n-1} = (x s_n\langle 1\rangle + a_{n-1})\langle 1\rangle = x a_n\langle 2\rangle + \theta_{n-1}\langle 1\rangle$$

$$\hat{s}_{n-2} = (x s_{n-1}\langle 1\rangle + a_{n-2})\langle 1\rangle$$

$$\hat{s}_0 = a_0\langle 1\rangle + a_1 x\langle 3\rangle + \dots a_{n-1}x^{n-1}\langle 2n-1\rangle + a_n x^n\langle 2n\rangle = (1+\theta_1)a_0 + \dots + (1+\theta_{2n})a_n x^n$$

άρα $\hat{s}_0 = fprog(a_0,\dots,a_n,x) = f(a_0(1+\theta_1),\dots,a_n(1+\theta_{2n}))$

και μικρό προς τα πίσω σφάλμα. Το θ_{2n} την μικρή ακρίβεια στους συντελεστές => δεν είναι ακριβώς γνωστός αλλά $|a_i - \hat{a}_i| \le \gamma_{2n}|$ σφάλμα από την ακρίβεια αυτή θα είναι οποιοδήποτε σφάλμα των πράξεων α.κ.

Άρα πίσω σταθερό!

υπολογίζεται ως:

$A = xy^T$ άρα κάθε στοιχείο του A υπολογίζεται ως: $fl(a_{ij}) = a_{ij}(1+\delta_{ij}) = x_i y_j (1+\delta_{ij})$, $|\delta_{ij}| \le u$. Επομένως: $fl(A) = xy^T + E$, όπου $|E| \le |x||y|^T u$ άρα έχουμε προς τα εμπρός σταθερότητα. Όμως δεν είναι πίσω σταθερή. Αν ήταν θα υπήρχαν $\Delta x, \Delta y$ ώστε $fl(A) = xy^T + E = (x+\Delta x)(y+\Delta y)^T$. Αυτή όμως η ισότητα δεν ισχύει γιατί θα είχαμε: $E = (x+\Delta x)(y+\Delta y)^T - xy^T$ και το δεξιό μέλος μπορεί να γραφεί:

$$E = [x, x+\Delta x]\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{bmatrix} (y+\Delta y)^T \\ y \end{bmatrix}$$

το μητρώο στα δεξιά έχει τάξη ≤ 2 ενώ το αριστερό έχει τάξη n. Επομένως δεν έχουμε ισότητα. Δεν είχαμε αρκετά στοιχεία εισόδου σε σχέση με τα εξόδου για να ρίξουμε το σφάλμα του αλγορίθμου.

ΝΔΟ ο πολ/σμος 2x2 κάτω τριγωνικών μητρώων $A, B \in \mathbb{R}^{2\times2}$ είναι πίσω σταθερός.

$$A = \begin{pmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{pmatrix} \times B = \begin{pmatrix} b_{11} & 0 \\ b_{21} & b_{22} \end{pmatrix} =$$

$$= \begin{pmatrix} a_{11}b_{11}(1+\delta_1) & 0 \\ ((a_{21}b_{11})(1+\delta_2) + (a_{22}b_{21})(1+\delta_3))(1+\delta_4) & a_{22}b_{22}(1+\delta_5) \end{pmatrix}$$

$$\begin{pmatrix} a_{11}(1+\delta_1) & 0 \\ a_{21}(1+\delta_2)(1+\delta_4) & a_{22} \end{pmatrix} \cdot B = \begin{pmatrix} b_{11} & 0 \\ b_{21}(1+\delta_3)(1+\delta_4) & b_{22}(1+ \end{pmatrix}$$

σταθερός εφόσον

$$\le |\theta_2| = \frac{2u}{1-2u} = \gamma_2 \quad \text{και} \quad \left|\frac{\tilde{b}_{ij} - b_{ij}}{b_{ij}}\right| \le |\theta_2| \le \gamma$$

$Px = \frac{uu^T}{u^Tu}\cdot x = \frac{u}{\|u\|}\frac{u^Tx}{\|u\|} = \frac{u}{\|u\|}\cdot\|x\|\cdot\cos(x,u)$

π.χ. $u = \begin{bmatrix}1\\2\\3\end{bmatrix}$

$P = \frac{[1,2,3]\begin{bmatrix}1\\2\\3\end{bmatrix}}{\begin{bmatrix}1\\2\\3\end{bmatrix}\cdot[1,2,3]} = \ldots$

Ερμιτιανά

$H = I - \frac{2uu^T}{u^Tu}$. $Hx = \left(I - \frac{2}{u^Tu}uu^T\right)x = x - Px - Px$

$H^T = H$ ισχύει $H^T\cdot H = H^2 = I$

$HA = A - \frac{2}{u^Tu}u(A^T\cdot u)^T$ $AH = A - \frac{2}{u^Tu}(Au)u^T$

QR

άσκ. QR: $\begin{bmatrix}3 & -2 & 1\\2 & 1 & 4\\1 & -1 & 2\end{bmatrix} \to R$

$= ?$

$= \begin{bmatrix}3 & -2 & 1\\0 & 1 & 4\\0 & 0 & 2\end{bmatrix}$ φτιάχνω κάτω τριγωνικό με H στη διαγ.

$\begin{bmatrix}0\\1\\-1\end{bmatrix}$ $\begin{bmatrix}0\\0\\1\end{bmatrix}$ $Q = H_1H_2\ldots H_n \to H_1\cdot H_2$

u_2 $H_1 = I - \frac{2u_1\cdot u_1^T}{u_1^Tu_1} = \begin{bmatrix}2/3 & -2/3 & -1/3\\-2/3 & -1/3 & -2/3\\-1/3 & -2/3 & 2/3\end{bmatrix}$

$\begin{bmatrix}1 & 0 & 0\\0 & 0 & 1\\0 & 1 & 0\end{bmatrix}$ $Q = H_1\cdot H_2$ γνωστό.

R : γνωστό

$Ax = b = \frac{1}{3}\begin{bmatrix}-5\\16\\4\end{bmatrix}$

$Ax = R\cdot x = H_2\cdot H_1\cdot b = \begin{bmatrix}2\\3\\2\end{bmatrix}$ άρα

$\begin{bmatrix}3 & -2 & 1\\0 & 1 & 4\\0 & 0 & 2\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix} = \begin{bmatrix}2\\5\\2\end{bmatrix}$ $\left.\begin{matrix}x_1\\x_2\\x_3\end{matrix}\right\}$ γνωστά

ΝΔΟ ο υπολογισμός της ορίζουσας ενός A 2×2 είναι πίσω σταθερός:

$$A=\begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad fl(|A|)=\Big((a\cdot d)(1+\delta_1)-(c\cdot b)(1+\delta_2)\Big)(1+\delta_3)=$$

$$=ad(1+\delta_1)(1+\delta_3)-cb(1+\delta_2)(1+\delta_3).$$

πίσω σταθερός αφού

$$fl(|A|)=|\tilde{A}|=\begin{vmatrix} a(1+\theta_1) & b(1+\theta_1') \\ c & d \end{vmatrix} \quad \text{και } |\theta_1|,|\theta_1'|\le h=$$

Μέθοδος LU

① Υπολογίζω τους τριγωνικούς LU.

② "$\Rightarrow$" λύνω $Lz=b$ ως προς z.

③ "$\Leftarrow$" λύνω $Ux=z$ ως προς x.

L = κάτω τριγωνικό με μονάδες στη διαγώνιο.

U = άνω τριγωνικό

Μερική οδήγηση: $L_{n-1}P_{n-1}\cdots L_1P_1A=U \Leftrightarrow A=(L_{n-1}P_{n-1}\cdots L_1P_1)^{-1}U$

$\Rightarrow P_{n-1}\cdots P_1A=P_{n-1}\cdots P_1(L_{n-1}P_{n-1}\cdots L_1P_1)^{-1}U \Leftrightarrow$

$$PA=LU.$$

πχ μετά από LU έχω $\begin{bmatrix} 1 & 2 & 3 \\ 0,5 & 1 & 2 \\ -1 & 1 & 1 \end{bmatrix} \to U$, $\downarrow L$

$$L=\begin{bmatrix} 1 & 0 & 0 \\ 0,5 & 1 & 0 \\ -1 & 1 & 1 \end{bmatrix} \quad U=\begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix} \quad LU=A=\begin{bmatrix} 1 & 2 & 3 \\ 0,5 & 2 & 3,5 \\ -1 & -1 & 0 \end{bmatrix}$$

$$b=\begin{bmatrix} 11 \\ 10 \\ -4 \end{bmatrix}=?$$

$$\Leftrightarrow L^{-1}LUx=L^{-1}b \Leftrightarrow Ux=L^{-1}b$$

I επαυξημένος $\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0,5 & 1 & 0 & 0 & 1 & 0 \\ -1 & 1 & 1 & 0 & 0 & 1 \end{array}\right] \cdots$

~~Ανάλυση LU~~

Έστω έχουν οι μεταθέσεις και $A = LU$.

Έστω ότι έχουν οι μεταθέσεις $|E| \leq \gamma n |\hat{L}||\hat{U}|$, $\|E\| \leq \gamma n \||\hat{L}|\| \||\hat{U}|\|$

$A + E = \hat{L}\hat{U}$ όπου $|E| \leq \gamma n |\hat{L}||\hat{U}|$. $(\hat{L} + \Delta L)\hat{y} = b$ όπου $|\Delta L| \leq \gamma n |\hat{L}|$ και

λύνοντας $\hat{L}y = b$ έχουμε $(\hat{L} + \Delta L)\hat{y} = b$ όπου $|\Delta L| \leq \gamma n |\hat{L}|$ και

το ίδιο για το $\hat{U}$. Επίσης:

$b = (\hat{L} + \Delta L)(\hat{U} + \Delta U)\hat{x}$ άρα $b = (\hat{L}\hat{U} + \hat{L}\Delta U + \Delta L \hat{U} + \Delta L \Delta U)\hat{x}$ επομένως το υπολο-

$$= (A + \underbrace{E + \hat{L}\Delta U + \Delta L \hat{U} + \Delta L \Delta U}_{\Delta A})\hat{x}$$

$= (A + \Delta A)x$ όπου

γισμένο x ικανοποιεί το $b = (A + \Delta A)x$ όπου

$\Delta A = E + \hat{L}\Delta U + \Delta L \hat{U} + \Delta L \Delta U$. Η πίσω ευστάθεια εξαρτάται

από το μέγεθος του $\frac{\|\Delta A\|}{\|A\|}$. Για να φράξω το ΔA:

$$|\Delta A| \leq |E| + |\hat{L}||\Delta U| + |\Delta L||\hat{U}| + |\Delta L||\Delta U| \leq 3\gamma n |\hat{L}||\hat{U}| + o$$

$$\|\Delta A\| \leq 3\gamma n \||\hat{L}|\| \cdot \||\hat{U}|\|$$

έχουμε πίσω ευστάθεια όταν $\frac{\|\Delta A\|}{\|A\|} \gtrsim O(u)$.

Αν $3\gamma n \||\hat{L}|\| \cdot \||\hat{U}|\| \gtrsim O(u) \cdot \|A\|$ τότε

$\frac{\|\Delta A\|}{\|A\|} \gtrsim O(u)$. Η πίσω ευστάθεια καθορίζεται

από το μέγεθος του $\||\hat{L}|\| \cdot \||\hat{U}|\|$

⇒ <u>Χωρίς οδήγηση</u> ο παράγοντας μπορεί να είναι πολύ μεγά[λος] και δεν έχουμε τρόπο να φράξουμε τους όρους του $\hat{L}$ [...] άρα δεν έχουμε πίσω ευστ[άθεια] τότε εκτός από ειδικές περιπτώσεις δεν έχουμε πίσω ευσ[τάθεια]

⇒ ~~μερική οδήγηση~~ <u>μερική οδήγηση</u>: ο $\|\hat{L}\|$ δεν μπορεί να γίνει πολύ μεγάλος γιατί το μέγιστο στοιχείο κάθε στήλης είναι

Blas-1: Όταν ακριβώς μια από τις διαστάσεις n_j είναι (μεγαλύτερη της μονάδας, δηλ. πράξεις του τύπου διάνυσμα/διάνυσμα π.χ. το DOT ($n_3>1$) και οι SAXPY ($n_2>1$ ή $n_3>1$). Ο αριθμός πράξεων και μεταφορών τότε είναι $O(n)$ και επομένως $\mu_{min}=O(1)$, άρα έχουμε μικρή τοπικότητα.

Blas-2: Όταν ακριβώς δύο διαστάσεις είναι μεγαλύτερες της μονάδας, δηλ. πράξεις τύπου μητρώο/διάνυσμα. Λαμβάνοντας τις αυτές διαστάσεις ίσες με n έχουμε ότι ο αριθμός των πράξεων και των μεταφορών είναι $O(n^2)$, άρα καλύτερη τοπικότητα χώρου και χρόνου από την Blas-1.

Blas-3: Όταν $n_j>1$ για $j=1,2,3$ δηλ πράξεις του τύπου μητρώο/μητρώο, οπότε έχουμε και αυξημένη τοπικότητα π.χ. αν $n_1=n_2=n_3=n$ τότε έχουμε $O(n^2)$ μεταφορές με $O(n^3)$ πράξεις.

Υλοποίηση Blas: $C=C+AB$ με $n_1=n_2=n_3=n$ $K<<3n^2<<\Lambda$

```
for j=1:N
  (*LOAD Bj, Cj στην κρυφή*)
  for k=1:n
    (*LOAD A:,k στην κρυφή και
      ενημέρωση 1ης τάξης του Cj*)
  end
  (*STORE Cj*)
end
```

$$\Phi_1=\sum_{j=1}^{N}(n^2+3n(n/N))=2n^2\left(\frac{3}{2}+\frac{n^2}{K}\right)$$

Οι πίνακες B, C είναι ομοιόμορφα τεμαχισμένοι σε N ομάδες από στήλες και $n=Nm$ για κάποιον ακέραιο m. Το N επιλέγεται ώστε να αποθηκεύονται 2 ομάδες και 1 στήλη του A

```
for I=1:N
  for J=1:N
    (*LOAD CIJ στην κρυφή μνήμη
    for K=1:N
      (*LOAD AIK, BKJ στην κρυφή και
       ενημέρωση CIJ*)
    end
    (*STORE CIJ*)
  end
end
```

$$\Phi_2=2n^2(1+N)\simeq\frac{n^3 2\sqrt{3}}{\sqrt{K}}$$

Τεμαχίζουμε τους πίνακες σε $N\times N$ τετραγωνικούς υποπίνακες μεγέθους m, ώστε $3m^2\simeq K$ και χρησιμοποιούμε την DC σε μορφή ijk σε επίπεδο υποπινάκων.

(3) Να ορίσετε με σαφήνεια τι σημαίνει πίσω ευστάθεια μιας $f : \mathbb{R}^n \to \mathbb{R}^m$ και εξηγήστε πως αν αποδείξατε την πίσω ευστάθεια μπορούμε να αναγάγουμε της εύρεση της απόστασης $f(x)$ από $f_{prog}(x)$ σε πρόβλημα διαταραχών. (2=

Αν ο υπολογισμός είναι πίσω ευσταθής τότε μπορούμε να με διάνυσμα, έστω $x_{prog} \in \mathbb{R}^n$, που είναι κοντά στο x, τέτ ώστε $f(x_{prog}) = f_{prog}(x)$, δηλ. ο υπολογισμός της συνάρτησης με το συγκεκριμένο αλγόριθμο σε α.κ.υ. είναι το ίδιο ακρι αποτέλεσμα με τον υπολογισμό της συνάρτησης στο x_{prog} με αριθμητική άπειρης ακρίβειας. Αν δείξουμε κάτι τέτοι ισχύει ότι $\|f(x) - f_{prog}(x)\| = \|f(x) - f(x_{prog})\|$ επομένως η απόσταση (το απόλυτο εμπρός σφάλμα) των δύο θα είναι με την απόσταση της τιμής του f στο x από την τιμή f σε ένα κοντινό σημείο x_{prog}, και η μελέτη της, εφόσ το x είναι κοντά στο x_{prog}, αντιστοιχεί σε πρόβλημα διαταραχών.

(4) $A \in \mathbb{C}^{n \times n}$ μιγαδικό και $B \in \mathbb{C}^{n \times k}$. $A \cdot B = ?$ Ω, Φ, Blas

$$C = A \cdot B = (A_R B_R - A_I B_I) + i(A_R B_I + A_I B_R) \text{ όπου}$$

$A = A_R + iA_I$, $B = B_R + iB_I$, $A_R, A_I \in \mathbb{R}^{n \times n}$ και $B_R, B_I \in \mathbb{R}$

$nk(2n-1)$ άρα $\Omega = 4nk(2n-1) + 2nk$

$$\mu_{min} = \frac{2n^2 + 4nk}{8n^2k - 2nk} \approx O(1/k) \text{ και αφού } k \gg 1 : \text{Blas-3}$$

$$\Phi = 2(n^2 + 2nk) = 2n^2 + 4nk.$$

Συνθήκη ακριβούς στρογγυλοποίησης: Αν $\circledast$ συμβολίζει την υλοποίηση της πράξης $\odot$, τότε δοθέντων $x, y \in F$ ισχύει ότι $x \circledast y = fl(x \odot y) \in F$, δηλ. το αποτέλεσμα της πράξης στο σύστημα είναι σαν να εκτελείται η πράξη ακριβώς στον $\mathbb{R}$ (δηλ. $x \odot y$) και μετά να στρογγυλοποιείται το αποτέλεσμα.

Μοντέλο διάδοσης σφάλματος: Έστω $x, y \in F$ και $x \odot$ τότε $fl(x \odot y) = (x \odot y)(1+\delta)$, για κάποιο $|\delta| \le u$.

και $$fl(x \odot y) = \frac{x \odot y}{1+\delta}, \quad |\delta| \le u$$

Για την α.κ.υ IEEE μονής και διπλής ακρίβειας μετρήστε την ακρίβεια σε αριθμό δεκαδικών ψηφίων: Η ουρά της μονής ακρίβειας έχει $24 = 23(+1)$ ψηφία ενώ σε διπλή $53 = 52(+1)$. Και στις δύο περιπτώσεις η αναπαράσταση είναι της μορφής $\pm m \times 2^{E}$ όπου $1 \le m < 2$ έχει τη δυαδική μορφή $1.b_s \ldots$. Στην μονή, η δυαδική ακρίβεια δίνεται από τα 24 ψηφία της ουράς, επομένως χρειαζόμαστε d δεκαδικά ψηφία, όπου $10^{-d} \approx 2^{-24}$ $(d = 7 \approx 24\log_{10}2)$, και στην διπλή $10^{-d} \approx 2^{-53}$ $(d = 16 \approx 53\log_{10}2)$

LINPACK benchmark: για την αξιολόγηση της επίδοσης υπολογιστικών συστημάτων, αποτελείται από ρουτίνες για την επίλυση γραμμικών συστημάτων και είναι από τα σημαντικότερα μετροπρογράμματα για την αξιολόγηση υπολογιστικών συστημάτων υψηλής απόδοσης.

Ανάλυση σφάλματος για το DOT $\Rightarrow s_n = x^T y$

$$\bar{s}_1 = fl(x_1 y_1) = x_1 y_1 (1+\delta_1).$$

$$\bar{s}_2 = fl(\bar{s}_1 + fl(x_2 y_2)) = [x_1 y_1 (1+\delta_1) + x_2 y_2 (1+\delta_2)](1+\delta_3) =$$

$$= x_1 y_1 (1+\delta_1)(1+\delta_3) + x_2 y_2 (1+\delta_2)(1+\delta_3). \quad \text{με } |\delta_i| \le u.$$

$$\bar{s}_n = x_1 y_1 \prod_{\substack{j=1 \\ j\neq 2}}^{n+1} (1+\delta_j) + x_2 y_2 \prod_{j=2}^{n+1} (1+\delta_j) + \dots + x_n y_n \prod_{j=n}^{n+1} (1+\delta_j)$$

$$\bar{s}_n = x_1 y_1 (1+\theta_n) + x_2 y_2 (1+\theta_n') + \dots + x_n y_n (1+\theta_2).$$

άρα DOT ακριβές επακριβές αποτέλεσμα για τα $x_1, \dots x_n$, $y_1(1+\theta_n), y_2(1+\theta_n'') \dots y_n(1+\theta_2)$. όπου

$|\theta_j| \le \dfrac{ju}{1-ju}$ Δείξαμε $fl(x^T y) = (x+\Delta x)^T y = x^T(y+\Delta y)$

όπου $|\Delta x| < \gamma_n |x|$, $|\Delta y| < \gamma_n |y|$

$f([x;y]) = x^T y$ θέτω $X = [x;y] \in \mathbb{R}^{2n}$

$$\text{cond}(f;X) = \frac{\|X\|}{\|x^T y\|} \left\| \frac{\partial f}{\partial X}\Big|_{[x;y]} \right\| \quad \text{και} \quad \frac{\partial f}{\partial X}\Big|_{[x;y]} = [y;x] \in \mathbb{R}^{1\times}$$

άρα $\text{cond}(f;X) = \dfrac{\|[x;y]\|}{|x^T y|} \|[y;x]\|$

θεωρώντας $\|[x;y]\| = \|[y;x]\|$ έχω $\text{cond}(f;X) \le \dfrac{\|[x;y]\|^2}{|x^T y|}$

σφάλμα προς τα εμπρός $<$ (δείκτης κατάστασης) $\cdot$ πίσω σφάλμα

Μέθοδος κανονικών εξισώσεων:

Αν: $A \in R^{m\times n}$, $b \in R^n$ και $A^T(b-Ax)=0$ τότε ισχύει

$$\|b-Ax\|_2 \le \|b-Ay\|_2 \quad \forall y \in R^n$$

Ο $P = A(A^TA)^{-1}\cdot A^T \Rightarrow$ τελεστής ορθογώνιας προβολής επί του υποχώρου που παράγεται από τις στήλες του A.

η λύση δίνεται από το x τέτοιο ώστε:

$Ax = Pb = A(A^TA)^{-1}\cdot A^Tb$ και θέτοντας $x = (A^TA)^{-1}A^Tb$

άρα το x ικανοποιεί το $A^TAx = A^Tb$.

(το A^TA είναι σ.θ.ο και αντιστρέψιμο)

Μέθοδος: • Υπολογισμός κάτω τριγωνικού τμήματος του $C = A^TA$ και του $d = A^Tb$

• Cholesky $C = G\cdot G^T$

• Επίλυση του $Gy = d$ και του $G^Tx = y$

Μειονεκτήματα: ο A^TA ίσως καταστρέψει ειδική δομή που μπορεί να έχει το A. Ο πολ/μος A^TA μπορεί να οδηγήσει σε απώλεια σημαντικών ψηφίων και αλλοίωση αποτελέσματος (ή υπερχείλιση = σπάνιο). Ευαισθησία της μεθόδου σε συσσώρευση αριθμητικών σφαλμάτων. Δηλαδή:

$k_2(A^TA) = (k_2(A))^2$ άρα το σχετικό σφάλμα αυξάνει με το τετράγωνο του δείκτη κατάστασης του A και η ευαισθησία σε αριθμητικά σφάλματα.

Λύση είναι

- Πόσοι αριθμοί κανονικοποιημένοι και πόσοι υποκανονικοποιημένοι μπορούν να αναπαρασταθούν στο $F = \pm m \times b^{e}$ 35

κανονικ.: χρησιμοποιεί t ψηφία για την ουρά (m) καθένα παίρνει b τιμές.

ουρά b^{t-1}
εκθέτης $e_{max}-t \to e_{min}-t$
οι τιμές του διαστ. $(e_{max}-e_{min}+1)$
πρόσημο 2

$b^{t-1}(e_{max}-e_{min}+1)\cdot 2(b-1)$ $+1$ για το μηδέν

$(b-1)$ τιμές του 1ου bit σε βάση b.

υποκανον.: $b^{t-1} \to$ ουρά
$2 \to$ πρόσημο
$1 \to$ εκθέτης $e_{min}-t$

$2(b^{t-1}-1)$ δεν υπολογίζω το μηδέν.

κανονικ.	υποκανον.	
$43\cdot 10^{9}$	$17\cdot 10^{7}$	double
$1{,}8\cdot 10^{19}$	$9\cdot 10^{15}$	single

- Στην IEEE πόσοι αριθμοί διπλής ακρίβειας παρεμβάλλονται μεταξύ δύο αριθμών (μονής);

$1,\ 1+\epsilon_M = 1+2^{-23}$ $[1, 1+\epsilon_M]$

$[1,2] = 2^{-52} = 2^{-t}$

$\frac{2^{-23}}{2^{-52}} = 2^{29} \to$ αριθμοί διπλής ακρίβειας

C $\left.\begin{matrix} m\cdot 2^{e} \\ (m+1)2^{e} \end{matrix}\right\} 2^{e}$ εύρος του α   εύρος του b