ΕΠΙΣΤΗΜΟΝΙΚΟΣ ΥΠΟΛΟΓΙΣΜΟΣ Ι *ΑΠΑΝΤΗΣΕΙΣ Εξεταστική Σεπτεμβρίου 2006*

I.

1. Να αναφέρετε ένα παράδειγμα πράξης BLAS-3. Τι πλεονέκτημα έχουμε όταν υλοποιούμε αλγορίθμους με πράξεις BLAS-3 αντί με εναλλακτικές υλοποιήσεις με άλλες πράξεις BLAS;

   Απάντηση. Πολλαπλασιασμός μητρώων. Μεγαλύτερη τοπικότητα, εφόσον βέβαια αξιοποιηθεί από την υλοποποίηση. □

2. Για κάθε έναν από τους παρακάτω υπολογισμούς να αναφέρετε τον εξέχοντα λόγο σφάλματος κατά τη διάρκεια των παρακάτω υπολογισμών (Οι επιλογές σας είναι μεταξύ σφάλματος στρογγύλευσης και σφάλματος διακριτοποίησης): α) Το σφάλμα κατά τον υπολογισμό του αθροίσματος 100 αριθμών κινητής υποδιαστολής. β) Το σφάλμα που υπεισέρχεται όταν προσεγγίζουμε την παράγωγο $\frac{df}{dx}$ της συνάρτησης $f(x)$ στο ξ ως $f(\xi+h)-f(\xi))/h$ (μπορείτε να υποθέσετε ότι τα h και h/ξ είναι μικρά και ότι οι τιμές $h, \xi+h, f(\xi+h), f(\xi)$ δίδονται ακριβώς ως αριθμοί κινητής υποδιαστολής). γ) Το σφάλμα που υπεισέρχεται όταν αναπαριστούμε το π με τον πλησιέστερο αριθμό κινητής υποδιαστολής.

   Απάντηση. α) ΣΤΡΟΓΓΥΛΕΥΣΗΣ, προφανώς δεν διακριτοποιούμε καμια συνεχή συνάρτηση, τελεστή, κ.λπ. β) ΔΙΑΚΡΙΤΟΠΟΪΗΣΗΣ, μια και πρόκειται για σφάλμα που προέρχεται σχεδόν αποκλειστκά από την προσέγγιση της παραγώγου. γ) ΣΤΡΟΓΓΥΛΕΥΣΗΣ, μια αναφερόμαστε σε σφάλμα που προέρχεται αποκλειστικά από την απεικόνιση ενός πραγματικού αριθμού επί του συστήματος των α.κ.υ. □

3. Έστω το μητρώο $A = [4, -8, 1; 6, 5, 7; 0, -10, -3]$. Ποιό θα είναι το στοιχείο «οδηγός» στο πρώτο βήμα α) αν δεν χρησιμοποιηθεί οδήγηση; β) Αν χρησιμοποιηθεί μερική οδήγηση; γ) Αν χρησιμοποιηθεί πλήρης οδήγηση;

   Απάντηση. α) 4 (το στοιχείο στη θέση (1,1)), β) 6 (το μέγιστο σε απόλυτη τιμή στοιχείο της 1ης στήλης), γ) -10 (το μέγιστο σε απόλυτη τιμή στοιχείο όλου του μητρώου). □

4. Για κάθε ένα από τα παρακάτω μητρώα να εξηγήσετε αν έχουν καλό ή κακό δείκτη κατάστασης: α) $[10^{10}, 0; 0, 10^{-10}]$. β) $[10^{10}, 0; 0, 10^{10}]$. γ) $[10^{-10}, 0; 0, 10^{-10}]$. δ) $[1, 2; 2, 4]$.

   Απάντηση. Δείκτης κατάστασης $\kappa(A) := \|A\|\|A^{-1}\|$ για όποια νόρμα διαλέξουμε. α) ΚΑΚΟΣ: γιατί $A^{-1} = [10^{-10}, 0; 0, 10^{10}]$ και σε όλες τις νόρμες (1, 2 ∞) έχουμε $\kappa(A) = 10^{10}/10^{-10} = 10^{20}$. β) ΑΡΙΣΤΟΣ. Καλό (τέλειο) γιατί $A = 10^{10}I$ όπου I ο ταυτοτικό μητρώο και επομένως $\kappa(A) = 1$. γ) ΑΡΙΣΤΟΣ: Ομοίως με πριν. δ) ΚΑΚΙΣΤΟΣ: μητρώο συμμετρικό και ο δείκτης κατάστασης είναι η μέγιστη προς ελάχιστη ιδιοτιμή. Προφανώς το μητρώο είναι μη αντιστρέψιμο, και $\lambda_{\min} = 0$, επομένως $\kappa(A) = \infty$). □

5. Έστω ότι για ένα μητρώο A γνωρίζετε ότι οι παραγοντοποιήσεις LU και QR είναι και οι δυο εφικτές. Με βάση τα κριτήρια του επιστημονικού υπολογισμού, να αναφέρετε έναν λόγο για τον οποίον συνήθως προτιμάται η LU για επίλυση τετραγωνικού συστήματος $Ax = b$ και έναν λόγο που θα μπορούσε να καταστήσει τη χρήση της QR πιο επιθυμητή (αναφερόμαστε πάντα σε τετραγωνικό σύστημα).

   Απάντηση. Η LU είναι φθηνότερη (εκτελείται ταχύτερα, μικρότερο κόστος) ενώ η QR (π.χ. με Householder) είναι πίσω ευσταθής ανεξαρτήτως των δεδομένων. □

6. Έστω ο βρόχος `for i=1:n, z(i) = a*x(i)+y(i); end`. Να τον ξαναγράψετε χρησιμοποιώντας ξετύλιγμα μήκους 3 και που να λειτουργεί σωστά ανεξάρτητα από την τιμή του n (θεωρούμε ότι είναι πάντα θετικός ακέραιος).

   Απάντηση. Το θέμα είναι να λειτουργεί σωστά για τιμές του n μικρότερες ή όχι κατ' ανάγκη πολλαπλάσια του 3. Ένας τρόπος είναι να υπολογίσουμε το υπόλοιπο της διαίρεσης του n με το 3, π.χ. με τη συνάρτηση `rem` της MATLAB και να γράψουμε:

```
m = rem(n,3); mp1 = m+1;
for i=1:m, z(i) = a*x(i)+y(i); end
for i = mp1:3:n /* προσέξτε, το n − m είναι πολλαπλάσιο του 3
      z(i) = a*x(i)+y(i);
      z(i+1) = a*x(i+1)+y(i+1);
      z(i+2) = a*x(i+2)+y(i+2);
end
```

□

7. Έστω ότι γνωρίζετε ότι σας δίδεται ένα πρόγραμμα (π.χ. η συνάρτηση MATLAB `myfun`), για το οποίο γνωρίζετε ότι για n δεδομένα εισόδου, έχει πολυπλοκότητα $O(n^2)$ αριθμητικών πράξεων κινητής υποδιαστολής αλλά δεν γνωρίζετε τι ακριβώς υπολογισμούς εκτελεί. Μπορείτε όμως να τρέξετε το πρόγραμμα και να χρησιμοποιήσετε χρονομετρητές (π.χ. `tic, toc`) για να μετρήσετε τον χρόνο που αναλώνει. Με βάση τα παραπάνω, ποιός είναι ο ελάχιστος αριθμός μετρήσεων που χρειάζεται για να εκτιμήσετε την πολυπλοκότητά του (δηλ. να εκτιμήσετε τους παράγοντες $\alpha_2, \alpha_1, \alpha_0$ στην έκφραση πολυπλοκότητας $\Omega = \alpha_2 n^2 + \alpha_1 n + \alpha_0$). Να περιγράψετε *συνοπτικά αλλά ξεκάθαρα* πώς θα ενεργούσατε για να εκτιμήσετε την πολυπλοκότητά του (π.χ. σε μορφή $\alpha_2 n^2 + \alpha_1 n + \alpha_0$ με γνωστούς παράγοντες $\alpha_2, \alpha_1, \alpha_0$).

   Απάντηση. Για να εκτιμήσω τους 3 παράγοντες χρειάζομαι τουλάχιστον 3 μετρήσεις με το χρονομετρητή. Συνήθως όμως χρειάζονται πολύ περισσότερες για να έχω μια καλή εκτίμηση. Ο τρόπος είναι να προσπαθήσω να υπολογίσω τους παράγοντες χρησιμοποιώντας ελάχιστα τετράγωνα και λύνοντας ένα πρόβλημα του τύπου $Va = b$, όπου κάθε γραμμή του $V \in \mathbb{R}^{n\times 3}$ περιέχει στοιχεία $[1, \nu_i, \nu_i^2]$ όπου ν_i είναι κάποια τιμή για το n και η αντίστοιχη θέση του b έχει τη χρονομέτρηση για της συνάρτησης για την τιμή $n = \nu_i$. Το διάνυσμα $a = [\alpha_0, \alpha_1, \alpha_2]^\top$. □

II. Δίνονται διανύματα $x, y \in \mathbb{R}^4$. Υποθέτουμε ότι όλα τα στοιχεία τους είναι μη αρνητικοί αριθμοί κινητής υποδιαστολής. Έστω ο υπολογισμός

```
s=0; for i=1:4, s=s+x(i)*y(i); end
```

1. Να δείξετε ότι ο αλγόριθμος υπολογισμού είναι πίσω σταθερός.

   Απάντηση. Σύμφωνα με τα στοιχεία που γνωρίζουμε για την διάδοση σφάλματος θα έχουμε:

$$\begin{aligned}
\mathrm{fl}(s) &= (((x(1)y(1)(1+\delta_1) + x(2)y(2)(1+\delta_2))(1+\delta_3) \\
&\quad + x(3)y(3)(1+\delta_4))(1+\delta_5) + x(4)y(4)(1+\delta_6))(1+\delta_7) \\
&= x(1)y(1)(1+\delta_1)(1+\delta_3)(1+\delta_5)(1+\delta_7) \\
&\qquad + x(2)y(2)(1+\delta_2)(1+\delta_3)(1+\delta_5)(1+\delta_7) \\
&\qquad + x(3)y(3)(1+\delta_4)(1+\delta_5)(1+\delta_7) + x(4)y(4)(1+\delta_6)(1+\delta_7) \\
&= x(1)y(1)(1+\theta_4) + x(2)y(2)(1+\hat{\theta}_4) + x(3)y(3)(1+\theta_3) + x(4)y(4)(1+\theta_2)
\end{aligned}$$

   όπου ως συνήθως $|\delta_j| \le u$ και $|\theta_j| \le \gamma_j := ju/(1-ju)$ όπου u είναι η μονάδα στρογγύλευσης. Επομένως (1η συνθήκη πίσω ευστάθειας) θα μπορούσαμε να έχουμε τα ίδια αποτελέσματα χρησιμοποιώντας για είσοδο τα διανύσματα x και

$$\tilde{y} := [y(1)(1+\theta_4), y(2)(1+\hat{\theta}_4), y(3)(1+\theta_3), y(4)(1+\theta_2)]$$

   και επειδή $|\theta_j| \le \gamma_j := ju/(1-ju)$ θα έχουμε ότι $\tilde{y}$ είναι κοντά στο y (2η συνθήκη πίσω ευστάθειας). Άρα ο αλγόριθμος πληροί και τις δυο συνθήκες για πίσω ευστάθεια. ΠΡΟΣΟΧΗ: Μερικοί (λανθασμένα) δεν χρησιμοποίησαν απόλυτες τιμές, π.χ. έφραζαν από τα δεξιά χρησιμοποιώντας μόνον το θ, κ.λπ. □

2. Να δείξετε ότι το ΣΧΕΤΙΚΟ εμπρός σφάλμα θα είναι μικρό.

   Απάντηση. Συνεχίζοντας τα παραπάνω, έχουμε ότι το σχετικό εμπρός σφάλμα είναι φραγμένο ως εξής:

$$\begin{aligned}
\frac{|\mathrm{fl}(s) - s|}{|s|} &= \frac{|x(1)y(1)(1+\theta_4) + x(2)y(2)(1+\hat{\theta}_4) + x(3)y(3)(1+\theta_3) + x(4)y(4)(1+\theta_2) - s|}{|s|} \\
&= \frac{|x(1)y(1)\theta_4 + x(2)y(2)\hat{\theta}_4 + x(3)y(3)\theta_3 + x(4)y(4)\theta_2|}{|s|} \\
&\le \frac{|x(1)y(1)||\theta_4| + |x(2)y(2)||\hat{\theta}_4| + |x(3)y(3)||\theta_3| + |x(4)y(4)||\theta_2||}{|s|} \\
&\le \gamma_4 \frac{|x(1)y(1)| + |x(2)y(2)| + |x(3)y(3)| + |x(4)y(4)|}{|s|}
\end{aligned}$$

αλλά επειδή είναι όλα μη αρνητικά θα έχουμε ότι

$$|s| = s = |x(1)y(1)| + |x(2)y(2)| + |x(3)y(3)| + |x(4)y(4)|$$

$$\frac{|\mathrm{fl}(s) - s|}{|s|} \quad \leq \quad \gamma_4,$$

το οποίο είναι πολύ μικρό. □

3. Να σχολιάσετε τον ισχυρισμό: «Αν δεν ισχύει η προϋπόθεση ότι όλα τα στοιχεία των x, y είναι μη αρνητικά, δεν μπορούμε να είμαστε βέβαιοι για κάποιο από τα προηγούμενα (δηλ. την πίσω ευστάθεια, το μικρό εμπρός σφάλμα, ή και τα δύο) ».

   Απάντηση. Τότε δεν μπορούμε να είμαστε βέβαιοι για μικρό εμπρός σχετικό σφάλμα καθώς δεν υπάρχει βεβαιότητα για το μέγεθος του s που βρίσκεται στον παρονομαστή, μπορεί δηλ. το $|s|$ να είναι πολύ μικρό σε σύγκριση με το $|\mathrm{fl}(s) - s|$. □

4. Αν ο αλγόριθμος υλοποιηθεί σε σύστημα που διαθέτει την εντολή FMA, να εξηγήσετε συνοπτικά (χωρίς πλήρη ανάλυση σφάλματος) γιατί η χρήση της FMA μπορεί να επιδράσει ευνοϊκά στην ακρίβεια του παραπάνω υπολογισμού.

   Απάντηση. Η εντολή FMA υλοποιεί την πράξη $s = c + a \times b$ με ένα μόνο σφάλμα στρογγύλευσης, δηλ.

   $$\mathrm{fl}(s) = (c + ab)(1 + \delta) \quad \text{αντί για} \quad \mathrm{fl}(s) = (c + ab(1 + \delta_1))(1 + \delta_2)$$

   επομένως μπορεί να επιφέρει μικρότερο σφάλμα στον παραπάνω βρόχου του οποιου κάθε επανάληψη είναι μια FMA. □

III. Έστω ότι έχουμε εφαρμόσει τον αλγόριθμο παραγοντοποίησης QR στο μητρώο $A \in \mathbb{R}^{3\times 3}$ που επιστρέφει στη θέση του A μητρώο με στοιχεία

$$\begin{pmatrix} 3 & -2 & 1 \\ 2 & 1 & 4 \\ 1 & -1 & 2 \end{pmatrix}$$

1. Να υπολογίσετε το αρχικό μητρώο A.

2. Να χρησιμοποιήσετε τα παραπάνω (απαραίτητο για την πλήρη βαθμολόγηση) για να λύσετε το σύστημα $Ax = b$ όπου $b = \frac{1}{3}[-5, -16, 4]^\top$.

Απάντηση. Ως γνωστό, το R της QR είναι το μητρώο

$$\begin{pmatrix} 3 & -2 & 1 \\ 0 & 1 & 4 \\ 0 & 0 & 2 \end{pmatrix}$$

ενώ τα διανύσματα Householder θα είναι

$$u_1 = [1, 2, 1]^\top, u_2 = [0, 1, -1].$$

Οι αντίστοιχοι μετασχηματισμοί οι

$$\begin{aligned} H_1 &= I - \frac{2u_1u_1^\top}{u_1^\top u_1} = \frac{1}{3}\begin{pmatrix} 2 & -2 & -1 \\ -2 & -1 & -2 \\ -1 & -2 & 2 \end{pmatrix} \\ H_2 &= I - \frac{2u_2u_2^\top}{u_2^\top u_2} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}. \end{aligned}$$

Προσέξτε: Είναι συμμετρικοί και ορθογώνιοι (ανακλαστές) και ισχύει ότι

$$H_2H_1A = R \Rightarrow A = H_1H_2R = \begin{pmatrix} 2 & -5/3 & -2 \\ -2 & 2/3 & -4 \\ -1 & 4/3 & 1 \end{pmatrix}.$$

Τέλος για να λύσουμε το σύστημα $Ax = b$ χρησιμοποιούμε ότι

$$H_2H_1Ax = Rx = H_2H_1b = [2, 5, 2]^\top$$

επομένως λύνουμε ως προς R με πίσω αντικατάσταση,

$$Rx = [2, 5, 2]^\top \Rightarrow x = [1, 1, 1]^\top.$$

Προσέξτε ότι ένας εναλλακτικός τρόπος θα ήταν να πολλαπλασιάσετε με $Q^\top$ και να λύσετε το σύστημα με συντελεστή R: $x = Q^\top b$. Το ενδιαφέρον όμως είναι ότι δεν χρειάζεται να υπολογίσετε το Q. Επίσης, στ' αλήθεια δεν χρειάζεται να υπολογίσετε τα H_1 και H_2, ενώ όλα τα παραπάνω μπορούν να προκύψουν μέσω των διανυσμάτω u_1, u_2 και χρησιμοποιώντας τη δομή των H_j για φθηνό πολλαπλασιασμό H_1H_2A.

□

IV. Έστω η διαφορική εξίσωση (αρχικών τιμών) $\frac{du}{dt}(t) = -Au(t)$ όπου $A = [2, -1; -1, 2]$, $u = [u_1(t), u_2(t)]^\top$ και οι συναρτήσεις u_1, u_2 είναι επιλεγμένες ώστε $u_1(0) = 2$, $u_2(0) = 1$.

1. Να χρησιμοποιήσετε την εμπρός μέθοδο Euler με σταθερό βήμα $\Delta t = 0.5$ για να υπολογίσετε αριθμητική προσέγγιση της λύσης στο σημείο $T = 2.0$.

   Απάντηση. Για ευκολία συμβολίζω το Δt με h. Στην εμπρός μέθοδο Euler εφαρμόζουμε τον τύπο

   $$U(t+h) - U(t) = -hAU(t) \Rightarrow u(t+h) = (I - hA)U(t)$$

   Επίσης, $I - hA = \frac{1}{2}[0, 1; 1, 0]$ επομένως εύκολα υπολογίζουμε:

   $$U(2) \quad = \quad (I - hA)((I - hA)((I - hA)((I - hA)U(0)))) = \frac{1}{16}[2, 1]^\top.$$

   ΠΡΟΣΟΧΗ: Ορισμένοι έγραψαν τον τύπο στη μορφή $U(t + h) = U(t)(I - hA)$. Η έκφραση αυτή δεν είναι έγκυρη! □

2. Να χρησιμοποιήσετε την ίδια μέθοδο αλλά με βήμα $\Delta t = 1.0$ για να υπολογίσετε αριθμητική προσέγγιση της λύσης στο σημείο $T = 2.0$.

   Απάντηση. Ομοίως, $I - hA = [-1, 1; 1, -1]$, άρα

   $$U(2) \quad = \quad (I - hA)((I - hA)U(0)) = [2, -2; -2, 2][2, 1]^\top = [2, -2]^\top.$$

   □

3. Είναι γνωστό ότι η ακριβής λύση του παραπάνω συστήματος των ΔΕ τείνει στο 0 καθώς το $t \to \infty$. Με βάση αυτό το στοιχείο, να σχολιάσετε τη συμπεριφορά των προσεγγίσεων που λάβατε χρησιμοποιώντας το IV.1 και το IV.2. (Υπόδειξη: Για μια πλήρη εξήγηση, είναι χρήσιμο να υπολογίσετε τις ιδιοτιμές του A.)

   Απάντηση. Προφανώς, υπάρχει πρόβλημα όταν $h = 1.0$. Ειδικότερα, οι τιμές της λύσης μεγαλώνουν (και μάλιστα αλλάζουν πρόσημο). Επομένως υπάρχει σοβαρή ένδειξη ότι έχουμε αστάθεια. Αυτό εξηγείται από την φασματική ακτίνα του $I - hA$. Για να υπάρχει ευστάθεια πρέπει η φασματική ακτίνα του $I - hA$ να μην είναι μεγαλύτερη του 1. Οι ιδιοτιμές του A είναι $\lambda = \{1, 3\}$, επομένως του $I - hA$ θα είναι $\{1 - h, 1 - 3h\}$. Για δεδομένο h, για ευστάθεια απαιτείται $\max\{|1-h|, |1-3h|\} < 1$. Επομένως, στην περίπτωση του βήματος 0.5 έχουμε ευστάθεια, ενώ όταν $h = 1.0$ άρα έχουμε αστάθεια. □

4. Να εφαρμόσετε ένα βήμα πίσω Euler για να βρείτε τη λύση στο σημείο $T = 2.0$ κατευθείαν με βήμα $\Delta t = 2.0$.

   Απάντηση. Στην πίσω μέθοδο Euler εφαρμόζουμε τον τύπο

   $$\begin{aligned} U(t+h) - U(t) &= -hAU(t+h) \Rightarrow (I + hA)U(t+h) = U(t) \\ \begin{pmatrix} 5 & -2 \\ -2 & 5 \end{pmatrix}\begin{pmatrix} U_1(2) \\ U_2(2) \end{pmatrix} &= \begin{pmatrix} 2 \\ 1 \end{pmatrix} \Rightarrow U(2) = \begin{pmatrix} 0.5714 \\ 0.4286 \end{pmatrix} \end{aligned}$$

   □