(3.33) $(1+\|A\|)^{-1} \leq \|(I-A)^{-1}\| \leq (1-\|A\|)^{-1}$.

ΑΠΟΔΕΙΞΗ Από την εξίσωση $\|A\|<1$ και το γεγονός ότι $\rho(A)\leq\|A\|$ έπεται ότι για κάθε ιδιοτιμή λ_i του A ισχύει ότι $|\lambda_i|<1$. Η' άλλα λόγια για κάθε ιδιοτιμή έχουμε ότι $\lambda_i \neq 1$. Συνεπώς ο πίνακας I-A είναι αντιστρέψιμος γιατί, αν δεν ήταν, θα υπήρχε τουλάχιστον μια μη μηδενική λύση $u \neq 0$ του συστήματος $(I-A)u = 0 \Leftrightarrow Au=1.u$ με $u \neq 0$, δηλ. ο A θα έχει μια ιδιοτιμή ίση με 1, άτοπο. Επίσης, επειδή η $\|\cdot\|$ είναι φυσική νόρμα, έχουμε ότι $\|I\| = 1$ (Άσκηση 3.8). Άρα επειδή υπάρχει ο $(I-A)^{-1}$ έχουμε:

$$1 = \|I\| = \|(I-A)^{-1}(I-A)\| \leq \|(I-A)^{-1}\| \|I-A\| \leq$$
$$\leq \|(I-A)^{-1}\| (\|I\| + \|A\|) = \|(I-A)^{-1}\| (1+\|A\|).$$

Άρα $(1+\|A\|)^{-1} \leq \|(I-A)^{-1}\|$. (αριστερή ανισότητα της (3.33)).

Απ' την άλλη μεριά, επειδή $(I-A)^{-1}(I-A)=I$ έχουμε
$(I-A)^{-1} - A(I-A)^{-1} = I$ δηλ. $(I-A)^{-1} = I + A(I-A)^{-1}$.
Άρα $\|(I-A)^{-1}\| = \|I + A(I-A)^{-1}\| \leq \|I\| + \|A(I-A)^{-1}\|$
$\leq 1 + \|A\| \|(I-A)^{-1}\|$. Συνεπώς $(1- \|A\|) \|(I-A)^{-1}\| \leq 1$ και επειδή $\|A\|<1$ έπεται ότι $\|(I-A)^{-1}\| \leq (1-\|A\|)^{-1}$ (δεξιά ανισότητα της (3.33)).@

4. ΔΕΙΚΤΗΣ ΚΑΤΑΣΤΑΣΗΣ ΠΙΝΑΚΑ

4.1 Δείκτης κατάστασης και ευστάθεια συστημάτων

Με τα εργαλεία του προηγούμενου κεφαλαίου ξαναχυρίζουμε στο πρόβλημα της ευστάθειας γραμμικών συστημάτων και αλγορίθμων για τη λύση τους που θίξαμε στην παρ. 2.5. Εκεί είχαμε ορίσει άτυπα την ευστάθεια (ενός προβλήματος ή ενός αλγορίθμου) σαν την ιδιότητα βάσει της οποίας "μικρά" σφάλματα (στα δεδομένα, στους υπολογισμούς) προκαλούν "μικρή" μεταβολή στο τελικό αποτέλεσμα. Θα δούμε ότι καθοριστικό ρόλο στην ευστάθεια τόσο του συστήματος $Ax=b$ όσο και στην ευστάθεια του αλγορίθμου της απαλοιφής του Gauss για την λύση του παίζει ο λεγόμενος <u>δείκτης κατάστασης</u> του πίνακα.

Αρχίζουμε από το εξής απλό παράδειγμα. Έστω x η λύση του συστήματος

(4.1) $Ax = b,$

όπου A είναι ένας nxn <u>αντιστρέψιμος</u> πίνακας και b ένα nx1 διάνυσμα. Αν μεταβάλουμε σε $b+\delta b$ το δεύτερο μέλος της (4.1), πόσο θα μεταβληθεί η λύση; Έστω $x+\delta x$ η λύση του νέου συστήματος

(4.2) $A(x+\delta x) = b+\delta b.$

Αφαιρώντας κατά μέλη έχουμε ότι $A\delta x=\delta b$ ή

(4.3) $\delta x = A^{-1}\delta b.$

Έστω τώρα $\|\cdot\|$ μια διανυσματική νόρμα και η αντίστοιχη φυσική νόρμα πινάκων που παράγεται από αυτήν. Η (4.3) δίνει ότι

(4.4) $\|\delta x\| = \|A^{-1}\delta b\| \leq \|A^{-1}\| \|\delta b\|.$

Αν υποθέσουμε ότι $b\neq 0$ τότε $x\neq 0$ και συνεπώς $\|\delta x\|/\|x\| \leq \|A^{-1}\| \|\delta b\|/\|x\|$. Αλλά, από την (4.1), $\|A\| \|x\| \geq \|b\|$. Συνεπώς

(4.5) $\|\delta x\|/\|x\| \leq \|A\| \|A^{-1}\| (\|\delta b\|/\|b\|).$

Η ποσότητα $\|\delta b\|/\|b\|$ είναι η <u>σχετική μεταβολή</u> του δευτέρου μέλους της (1.1) και η $\|\delta x\|/\|x\|$ η αντίστοιχη <u>σχετική</u> μεταβολή της λύσης (ή σχετικό "<u>σφάλμα</u>", αν θεωρήσουμε π.χ. ότι το δb παριστάνει ένα μικρό σφάλμα στην παράσταση των δεδομένων του δευτέρου μέλους της (1.1)). Βλέπουμε λοιπόν ότι η ποσότητα $\|A\|\ \|A^{-1}\|$ είναι ένας <u>συντελεστής</u> που προσδιορίζει πόσο μεγάλο μπορεί να γίνει το σχετικό σφάλμα $\|\delta x\|/\|x\|$, όταν το σχετικό σφάλμα του δευτέρου μέλους είναι $\|\delta b\|/\|b\|$.

Ο αριθμός

$$(1.6)\quad \kappa(A) = \|A\|\ \|A^{-1}\|,$$

που εξαρτάται φυσικά από τον A και την χρησιμοποιούμενη νόρμα λέγεται <u>δείκτης κατάστασης</u> του πίνακα A ως πρός την νόρμα $\|\cdot\|$. <u>Αν ο $\kappa(A)$ είναι πολύ μεγάλος τότε περιμένουμε ότι μια μικρή μεταβολή του b μπορεί να προκαλέσει μεγάλη μεταβολή στη λύση</u>.

ΑΣΚΗΣΗ 1.1 Να δειχθεί ότι ο δείκτης κατάστασης ως πρός την νόρμα $\|.\|_1$ του συστήματος (2.28) είναι περίπου 2.6×10^6. @

Ο δείκτης κατάστασης ενός πίνακα ως πρός οποιαδήποτε <u>φυσική</u> νόρμα $\|\cdot\|$ είναι πάντα μεγαλύτερος ή ίσος της μονάδας, γιατί $1 = \|I\| = \|AA^{-1}\| \le \|A\|\ \|A^{-1}\| = \kappa(A)$. Λέμε ότι ο πίνακας έχει <u>κακή κατάσταση</u> αν $\kappa(A)>>1$ και <u>καλή κατάσταση</u> αν ο $\kappa(A)$ είναι μικρός, π.χ. αν $1\le \kappa(A)\le 100$. Εκτός από μέτρο της "ευαισθησίας" της λύσης x σε μεταβολές του b, μπορούμε ν' αποδείξουμε ότι ο $\kappa(A)$ είναι και (αντίστροφο) μέτρο του <u>πόσο κοντά βρίσκεται ο A στο σύνολο των μη αντιστρέψιμων πινάκων</u>. Πράγματι έστω B ένας οποιοσδήποτε μη αντιστρέψιμος πχπ πίνακας. Τότε υπάρχει $x\neq0$ τέτοιο ώστε $Bx=0$. Συνεπώς $\|A-B\|\ \|x\| \ge \|(A-B)x\| = \|Ax-Bx\| = \|Ax\|$. Επίσης $\|x\| = \|A^{-1}Ax\| \le \|A^{-1}\|\ \|Ax\| \Rightarrow \|Ax\| \ge \|x\|/\|A^{-1}\|$. Συνεπώς έχουμε, επειδή $x\neq0$, ότι $\|A^{-1}\|^{-1} \le \|A-B\|$, για οποιοδήποτε μη αντιστρέψιμο πίνακα B, δηλ. ότι

$$(\kappa(A))^{-1} \le \inf\{\|A-B\|/\|A\|,\ B \text{ μη αντιστρέψιμος}\}.$$

Μπορεί ν' αποδειχθεί ακριβέστερα ότι η παραπάνω ανισότητα $\le$ είναι στην πραγματικότητα ισότητα, δηλ. ότι ο αντίστροφος $(\kappa(A))^{-1}$ του δείκτη κατάστασης του A μετράει την (σχετική ως πρός το μέγεθος του $\|A\|$) <u>απόσταση</u> του πίνακα A από το σύνολο των πχπ μη αντιστρέψιμων πινάκων.

ΑΣΚΗΣΗ 1.2 Χωρίς να βρεθεί ο A^{-1}, δείξτε ότι για τον πίνακα

$$A = \begin{pmatrix} 1.01 & .99 \\ .99 & 1.01 \end{pmatrix}$$

ισχύει ότι $\kappa_\infty(A) = \|A\|_\infty\ \|A^{-1}\|_\infty \ge 100$ (Υπόδειξη: ο πίνακας $B:b_{ij}=1$ δεν είναι αντιστρέψιμος). @

ΑΣΚΗΣΗ 1.3 Έστω A αντιστρέψιμος άνω ή κάτω τριγωνικός πίνακας. Να δειχθεί ότι

$$\kappa_\infty(A) \ge \|A\|_\infty / \min_i|a_{ii}|. \ @$$

ΑΣΚΗΣΗ 1.4 Αν ο A είναι αντιστρέψιμος και ο πίνακας B ικανοποιεί την σχέση

$$\|A-B\| < 1/\|A^{-1}\|,$$

τότε ο B είναι αντιστρέψιμος. @

Επιστρέφουμε τώρα στην (1.5). Θα έλεγε κανείς ότι ο δείκτης κατάστασης $\kappa(A)$ είναι μόνο ένα άνω όριο του λόγου $(\|\delta x\|/\|x\|)/(\|\delta b\|/\|b\|)$. Μπορεί όμως να δειχθεί ότι <u>για κάθε πίνακα A</u> μπορούμε να βρούμε ένα b και κατάλληλη μεταβολή δb τέτοια ώστε η (1.5) να ισχύει με το σημείο της ισότητας, δηλ. ότι πάντα υπάρχει ένα δεύτερο μέλος b και μια κατάλληλη μεταβολή δb που να μεταβάλλουν την λύση x κατά το μέγιστο δυνατό. Έστω π.χ.

$$A = \begin{pmatrix} 4.1 & 2.8 \\ 9.7 & 6.6 \end{pmatrix},\quad b = \begin{pmatrix} 4.1 \\ 9.7 \end{pmatrix},\ \text{οπότε}\quad x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

και $\|b\|_1=13.8$, $\|x\|_1 = 1$. Αν το δεύτερο μέλος υποστεί την μεταβολή $\delta b = (0.01,0)^T$, τότε η λύση $x+\delta x$ του νέου συστήματος $A(x+\delta x) = b+\delta b$ είναι $x+\delta x = (0.34,0.97)^T$. Συνεπώς $\|\delta b\|_1 = .01$, $\|\delta x\|_1 = 1.63$ και $\|\delta b\|_1/\|b\|_1 = .0007246$, $\|\delta x\|_1/\|x\|_1 = 1.63$. Συνεπώς τα σχετικά σφάλματα έχουν λόγο $1.63/.0007246 = 2249.4$. Απ' την άλλη μεριά μπορούμε να υπολογίσουμε για τον πίνακα A ότι $\kappa_1(A)=\|A\|_1\|A^{-1}\|_1=2249.4$. Δηλ. η (1.5) ισχύει εδώ ισότητα.

Τα δεδομένα του προβλήματος (1.1) δεν είναι μόνο τα στοιχεία του b αλλά και τα στοιχεία του A. Ο δείκτης κατάστασης του πίνακα καθορίζει επίσης και το πώς μεταβολές του A επηρεάζουν την λύση του (1.1). Πραγματικά, αν όταν μεταβάλουμε τον πίνακα A σε A+δA (όπου δA ένας "μικρός" nxn πίνακας) η λύση x του (1.1) γίνει x+δx, έχουμε ότι

(1.7) $(A+\delta A)(x+\delta x) = b.$

Εξ' υποθέσεως έχουμε ότι ο A είναι αντιστρέψιμος, δηλ. ότι το σύστημα (1.1) έχει λύση. Πρέπει να υποθέσουμε επίσης ότι και το (1.7) έχει λύση, δηλ. ότι και ο A+δA είναι αντιστρέψιμος. Μια συνθήκη η οποία το εξασφαλίζει αυτό είναι η

(1.8) $\|\delta A\|\,\|A^{-1}\| < 1,$

για οποιαδήποτε φυσική νόρμα πινάκων $\|\cdot\|$. Πραγματικά, η (1.8) συνεπάγεται ότι $\|\delta A\,A^{-1}\|<1$. Η πρόταση 3.1 δίνει τώρα ότι ο πίνακας $I+(\delta A)A^{-1}$ είναι αντιστρέψιμος => ο πίνακας $(I+(\delta A)A^{-1})A=A+\delta A$ είναι αντιστρέψιμος. Δηλ. η (1.8), που μπορεί να γραφτεί σαν $\|\delta A\| < 1/\|A^{-1}\|$, είναι μια συνθήκη που εξασφαλίζει ότι η μεταβολή δA είναι αρκετά μικρή ώστε όχι μόνο ο A αλλά και ο A+δA να είναι αντιστρέψιμος. (Πρβλ. άσκηση 1.1 με B=A+δA). Υπό την προϋπόθεση ότι η (1.8) ισχύει έχουμε, αφαιρώντας κατά μέλη τις (1.1) και (1.7) ότι $(A+\delta A)\delta x + (\delta A)x = 0 \Rightarrow \delta x = -(A+\delta A)^{-1}(\delta A)x = -[(I+(\delta A)A^{-1})A]^{-1}(\delta A)x = -A^{-1}(I+(\delta A)A^{-1})^{-1}(\delta A)x$.

Άρα, παίρνοντας νόρμες,

(1.9) $\|\delta x\| = \|A^{-1}(I+(\delta A)A^{-1})^{-1}(\delta A)x\| \le \|A^{-1}\|\,\|(I+(\delta A)A^{-1})^{-1}\|.\|\delta A\|\,\|x\|.$

Η (3.32) δίνει τώρα λόγω της (1.8) ότι

(1.10) $\|(I+(\delta A)A^{-1})^{-1}\| \le 1/(1-\|\delta A\|\,\|A^{-1}\|).$

Άρα, από τις (1.9) και (1.10) έχουμε

$\|\delta x\| \le \|A^{-1}\|\,\|\delta A\|\,\|x\|/(1-\|\delta A\|\,\|A^{-1}\|),$

δηλ. για $x\neq 0$

(1.11) $\|\delta x\|/\|x\| \le (\kappa(A)/1-\|A^{-1}\|\,\|\delta A\|).(\|\delta A\|/\|A\|).$

Ξανά δηλ. βλέπουμε τον καθοριστικό ρόλο του κ(A) ως "συντελεστή μεγέθυνσης" σφαλμάτων ή μεταβολών των στοιχείων του πίνακα A, δηλ. ως μέτρου του πόσο το σχετικό σφάλμα $\|\delta A\|/\|A\|$ επηρεάζει το $\|\delta x\|/\|x\|$.

Το γενικό βέβαια πρόβλημα είναι το

(1.12) $(A+\delta A)(x+\delta x) = (b+\delta b),$

η περίπτωση δηλ. που και ο A και το b μεταβάλλονται. Δεν είναι δύσκολο να δείξουμε ότι στην περίπτωση (1.12), αν ισχύει η (1.8) πάλι, έχουμε ότι

(1.13) $\|\delta x\|/\|x\| \le \kappa(A)/(1-\|\delta A\|\,\|A^{-1}\|)\,(\|\delta b\|/\|b\|+\|\delta A\|/\|A\|),$

δηλ. μια ανισότητα της οποίας οι (1.5) και (1.11) αποτελούν ειδικές περιπτώσεις, αν δA=0 και δb=0, αντίστοιχα.

ΑΣΚΗΣΗ 1.5 Αν οι πίνακες A,B είναι αντιστρέψιμοι, να δειχθεί ότι $\|B^{-1}-A^{-1}\|/\|B^{-1}\| \le \kappa(A)\,(\|A-B\|/\|A\|)$. @

ΑΣΚΗΣΗ 1.6 Αν ο A είναι αντιστρέψιμος αυτοσυζυγής πίνακας (δηλ. αν $A = A^*$) να δειχθεί ότι ο δείκτης κατάστασης του $\kappa_2(A) = \|A\|_2\,\|A^{-1}\|_2$ δίνεται από τον τύπο

(1.14) $\kappa_2(A) = \max_i|\lambda_i|/\min_i|\lambda_i|,$

όπου λ_i, $1\le i\le n$ οι ιδιοτιμές του A. @

ΑΣΚΗΣΗ 1.7 Δίνεται ο πίνακας A και ο A^{-1}:

$$A = \begin{pmatrix} 6 & 13 & -17 \\ 13 & 29 & -38 \\ -17 & -38 & 50 \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} 6 & -4 & -1 \\ -4 & 11 & 7 \\ -1 & 7 & 5 \end{pmatrix}.$$

Ο πίνακας A έχει ιδιοτιμές λ_1=.0568, λ_2=.2007, λ_3=84.74. Χρησιμοποιώντας τις σχέσεις (3.26), (4.14) να βρείτε τα $\|A\|_2$, $\|A^{-1}\|_2$, $\kappa_2(A)=\|A\|_2\,\|A^{-1}\|_2$. Βρείτε επίσης τους $\kappa_\infty(A)$, $\kappa_1(A)$ @

ΑΣΚΗΣΗ 1.8 Έστω $x\neq 0$ η (ακριβής) λύση του συστήματος $Ax=b$, όπου A αντιστρέψιμος. Έστω $\tilde{x}$ μια προσέγγιση της x. Ορίζουμε όπως στην άσκηση 2.16, το <u>υπόλοιπο</u> r της $\tilde{x}$ σαν $r=A\tilde{x}-b$. Να δειχθεί ότι για κάθε διανυσματική νόρμα (και αντίστοιχη φυσική νόρμα πινάκων) $\|\cdot\|$:

(4.15) $\|\tilde{x}-x\|/\|x\| \leq \kappa(A)\,(\|r\|/\|b\|)$, $\kappa(A)=\|A\|\,\|A^{-1}\|$.

(Η (4.15) συμπληρώνει την άσκηση 2.16: Έστω ότι το υπόλοιπο $\|r\|$ είναι <u>μικρό</u>. Αν όμως ο A έχει <u>κακή κατάσταση</u>, δηλ. αν $\kappa(A)>>1$, τότε η (4.15) δείχνει ότι <u>σχετικό σφάλμα</u> $\|\tilde{x}-x\|/\|x\|$ μπορεί να είναι μεγάλο. Άρα μικρό υπόλοιπο $\not\Rightarrow$ μικρό σφάλμα). @

ΑΣΚΗΣΗ 1.9 Για τον πίνακα A της άσκησης 1.7, υποθέστε ότι έχουμε μια προσέγγιση $\tilde{x}$ στη λύση x του συστήματος $Ax=b$, $b=(1,1)^T$, που ικανοποιεί: $\|A\tilde{x}-b\|_2 \leq 10^{-2}$. Βρήτε άνω φράγματα

για το απόλυτο ($\|\tilde{x}-x\|_2$) και το σχετικό σφάλμα ($\|\tilde{x}-x\|_2/\|x\|_2$). της $\tilde{x}$: @

ΑΣΚΗΣΗ 1.10 Να δείξετε ότι για οποιαδήποτε νόρμα πινάκων και οποιοδήποτε μιγαδικό αριθμό $\lambda\neq 0$

(4.16) $\kappa(\lambda A) = \kappa(A)$.

Επίσης, αν ο D είναι <u>διαγώνιος πίνακας</u> με στοιχεία $d_{ii}\neq 0$, $1\leq i\leq n$, να δείξετε ότι

(4.17) $\kappa_\infty(D) = \kappa_1(D) = \kappa_2(D) = \max_i|d_{ii}|/\min_i|d_{ii}|$. @

ΑΣΚΗΣΗ 1.11 Θα έλεγε κανείς ότι το μέγεθος της ορίζουσας ενός πίνακα καθορίζει την ευστάθεια του προβλήματος $Ax=b$ σύμφωνα με τον συλλογισμό: μικρή ορίζουσα (δηλ. κοντά στο 0) => A "σχεδόν" μη αντιστρέψιμος => σύστημα ασταθές.

Αυτό όμως δεν είναι αλήθεια: Θεωρήστε τον 100x100 διαγώνιο πίνακα με d_{ii}=0.1, $1\leq i\leq 100$. Βρείτε τη ορίζουσα det(D) και τους δείκτες κατάστασης $\kappa_\infty(D)$, $\kappa_1(D)$, $\kappa_2(D)$. (Προφανώς το σύστημα $Dx=b$ δεν είναι καθόλου ασταθές. Λύνεται αμέσως: $x=10b$ και για μια μεταβολή δb έχουμε ότι αν $D(x+\delta x)=b+\delta b$ => $\delta x=10\delta b$ => $\|\delta x\|/\|x\| = \|\delta b\|/\|b\|$). Συνεπώς το μέγεθος της ορίζουσας δεν έχει καμμιά σημασία για την ευστάθεια του συστήματος. Το σύστημα αυτό είναι όσο ευσταθές θα μπορούσε να γίνει. @

ΑΣΚΗΣΗ 1.12 Είδαμε ότι πάντα $\kappa(A)\geq 1$. Τι είδους πίνακες έχουν μικρό $\kappa(A)$; Να αποδειχθεί ότι:

(i) Αν ο A είναι διαγώνιος με ίσα στοιχεία ($\neq 0$) τότε $\kappa_\infty(A)=\kappa_1(A)=\kappa_2(A)=1$.

(ii) Αν ο A ικανοποιεί $\|Ax\|=\|x\|$ για κάθε x, να αποδειχθεί ότι $\kappa(A)=\|A\|\,\|A^{-1}\|=1$.

(iii) Για ένα πίνακα U τέτοιον ώστε $UU^*=I$, να δειχθεί ότι $\kappa_2(U)=1$. @

1.2 Επιρροή του δείκτη κατάστασης στην απαλοιφή.

Ο δείκτης κατάστασης παίζει μεγάλο ρόλο και στη μελέτη της διάδοσης των σφαλμάτων στρογγύλευσης κατά την απαλοιφή Gauss (ή την ανάλυση PA=LU) με αριθμητική πεπερασμένης ακρίβειας. Ας θεωρήσουμε πρώτα την "ιδανική" περίπτωση όπου σφάλματα στρογγύλευσης συμβαίνουν μόνο κατά την παράσταση των στοιχείων του A και του b στον υπολογιστή και όπου οι υπόλοιπες αριθμητικές πράξεις της απαλοιφής γίνονται ακριβώς. Με τον συμβολισμό της παρ. 2.2 (βλ. και Σημειώσεις Αριθμητικής Ανάλυσης I) κάθε πραγματικός α (μέσα στο εύρος των αριθμών της μηχανής) παριστάνεται στον υπολογιστή από τον $fl(a)$ όπου

$$(1.18)\quad fl(a) = a(1+\varepsilon),\ |\varepsilon|\le u,$$

και όπου η "μονάδα του σφάλματος στρογγύλευσης" u είναι b^{1-t} για αποκοπή και $b^{1-t}/2$ για στρογγύλευση. Στον υπολογιστή επομένως το δεύτερο μέλος b θα παρασταθεί από ένα διάνυσμα $\tilde{b}$ τέτοιο ώστε $\tilde{b}_j = b_j(1+\varepsilon_j)$, $|\varepsilon_j|\le u$, $1\le i\le n$. Συνεπώς,

$$(1.19)\quad \tilde{b}=b+\delta b,\quad \|\delta b\|_\infty \le u\|b\|_\infty .$$

(Η χρήση της νόρμας $\|\cdot\|_\infty$ είναι ευδεδειγμένη για την εκτίμηση των σφαλμάτων στρογγύλευσης). Αναλόγα, ο A θα παρασταθεί από ένα πίνακα $\tilde{A}$, όπου

$$(1.20)\quad \tilde{A}=A+\delta A,\quad \|\delta A\|_\infty \le u\|A\|_\infty .$$

Έστω x η θεωρητική λύση του συστήματος $Ax=b$. Βάσει των υποθέσεων μας η υπολογιστική λύση $\tilde{x}$ είναι η ακριβής λύση του συστήματος $\tilde{A}\tilde{x}=\tilde{b}$, δηλ. του συστήματος

$$(1.21)\quad (A+\delta A)\tilde{x} = (b+\delta b).$$

Για να βρούμε μια εκτίμηση του σχετικού σφάλματος $\|\tilde{x}-x\|_\infty / \|x\|_\infty$ ανατρέχουμε τώρα στη θεωρία της προηγούμενης παραγράφου, δηλ. στην περίπτωση (4.12), όπου $x+\delta x=\tilde{x}$. Ας υποθέσουμε ότι ισχύει η συνθήκη

$$(1.22)\quad u\,\kappa_\infty(A) = r<1.$$

Συνεπώς ισχύει η (1.8), δηλ. ο πίνακας $A+\delta A$ είναι αντιστρέψιμος και εύκολα, η (1.13) δίνει

$$\|\delta x\|_\infty / \|x\|_\infty = \|x-\tilde{x}\|_\infty / \|x\|_\infty \le (\kappa_\infty(A)/(1-r))(\|\delta b\|_\infty/\|b\| + \|\delta A\|_\infty/\|A\|_\infty)$$

δηλ., λόγω των (1.19) και (1.20), ότι

$$(1.23)\quad \|x-\tilde{x}\|_\infty / \|x\|_\infty \le 2u\kappa_\infty(A)/(1-r).$$

Είναι λοιπόν φανερό ότι αν το γινόμενο $u.\kappa_\infty(A)$ πλησιάζει την μονάδα, θα πρέπει να περιμένουμε μεγάλο σχετικό σφάλμα για την υπολογιστική λύση $\tilde{x}$.

ΑΣΚΗΣΗ 1.13 Δείξτε ότι στην περίπτωση που τα στοιχεία του πίνακα A παριστάνονται ακριβώς, (δηλ. όταν το σφάλμα στρογγύλευσης προέρχεται μόνο από την παράσταση του b), ισχύει η εκτίμηση $\|x-\tilde{x}\|_\infty / \|x\|_\infty \le u\kappa_\infty(A)$, χωρίς την υπόθεση (1.22). Εφαρμογή: Έστω ότι τα στοιχεία των A και b είναι ακέραιοι με την εξαίρεση του b_1 που είναι ίσο με 10^{-1}. Παριστάνουμε τα στοιχεία του A και b σε απλή ακρίβεια στον UNIX 11 και υποθέτουμε ότι οι υπόλοιπες πράξεις της απαλοιφής γίνονται ακριβώς. Ποιό είναι το αναμενόμενο σχετικό σφάλμα $\|x-\tilde{x}\|_\infty / \|x\|_\infty$ αν $\kappa_\infty(A)=10^4$; Το ίδιο ερώτημα για απλή και διπλή ακρίβεια στον IBM 4361. (Η αριθμητική του IBM 4361 είναι ίδια με την αριθμητική του συστήματος IBM 370). @

Προχωρούμε τώρα στην περίπτωση που η απαλοιφή (με οδήγηση ή χωρίς) γίνεται με αριθμητική πεπερασμένης ακρίβειας και που τα σφάλματα στρογγύλευσης συσσωρεύονται κατά τις πράξεις. Θέλουμε να εκτιμήσουμε το (σχετικό) σφάλμα της υπολογιστικής λύσης x. Χρησιμοποιώντας την σχέση (1.18) και το γεγονός ότι $fl(a \circledast b)=(a \circledast b)(1+\varepsilon)$, $|\varepsilon|\le u$, όπου $\circledast$ οποιαδήποτε αριθμητική πράξη $+,-,.,/$, είναι δυνατόν να παρακολουθήσουμε την επίδραση και την συσσώρευση των σφαλμάτων στρογγύλευσης σε κάθε πράξη της απαλοιφής (δηλ. στον υπολογισμό των πολλαπλασιαστών m_{ij}, των στοιχείων $a_{ij}^{(k)}$ κλπ.) και της οπισθοδρόμησης. Η "λογιστική" βέβαια ενός τέτοιου εγχειρήματος είναι πολύπλοκη και η τεχνική αυτή (που λέγεται "απ' ευθείας" (forward) εκτίμηση του σφάλματος) μάλλον θα δώσει ένα ευτελώς μη ρεαλιστικό (δηλ. ένα πολύ μεγάλο) άνω φράγμα για το σφάλμα $\|x-\tilde{x}\|_\infty / \|x\|_\infty$. Μια άλλη τεχνική, η

λεχόμενη "αντίστροφη" (inverse) ανάλυση του σφάλματος του προβλήματος οφείλεται στον άγγλο μαθηματικό J.H. Wilkinson (δεκαετία '60) που απέδειξε ότι η υπολογιστική (προσεγγιστική) λύση $\tilde{x}$ που τελικά παίρνουμε, μπορεί να θεωρηθεί ως <u>ακριβής λύση</u> (δηλ. ως λύση με αριθμητική απεριόριστης ακρίβειας) όχι του συστήματος $Ax=b$ βέβαια, αλλά του "παραπλήσιου" συστήματος

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

όπου ο πίνακας μεταβολής δA έχει χονικά "μικρά" στοιχεία. Αν έχουμε μια καλή εκτίμηση του $\|\delta A\|_\infty$ π.χ., τότε η θεωρία της παρ. 4.1 (βλ. (4.7).- (4.11)) μας δίνει μια καλή εκτίμηση τού σφάλματος $\|x-\tilde{x}\|_\infty / \|x\|_\infty$.

ΑΣΚΗΣΗ 4.14 (Για φανατικούς). Η "αντίστροφη" ανάλυση του σφάλματος έχει εφαρμογή και σε άλλα παραδείγματα υπολογισμών. Π.χ. θεωρείστε δυο 2x2 άνω τριγωνικούς πίνακες A και B και έστω $fl(AB)$ ο πίνακας που υπολογίζεται από τον υπολογιστή ως το γινόμενο τους (δηλ. με σφάλματα στρογγύλευσης στην παράσταση των A και B και στις πράξεις του πολλαπλασιασμού των δυο πινάκων). Δείξτε ότι υπάρχουν δυο 2x2 πίνακες $\tilde{A},\tilde{B}$, παραπλήσιοι στους A και B αντίστοιχα, τέτοιοι ώστε $fl(AB)=\tilde{A}\tilde{B}$, δηλ. τέτοιοι ώστε το <u>ακριβές</u> γινόμενο τους $\tilde{A}\tilde{B}$ να είναι ίσο με το προσεγγιστικό "γινόμενο" $fl(AB)$. @

Υποθέτοντας για απλούστευση ότι τα στοιχεία των A,b παριστάνονται ακριβώς αρχικά στον υπολογιστή, ο Wilkinson έδειξε (για την απόδειξη, που παραλείπεται, βλ. π.χ. τα βιβλία [A1],[B1],[B3] ή την "βίβλο" [B5] στη βιβλιογραφία του μαθήματος) ότι αν $n^2u<1$, τότε για τον πίνακα δA του (4.24) ισχύει

(4.25) $\|\delta A\|_\infty \le 1.01(n^3+3n^2)\varepsilon\|A\|_\infty u$

όπου

(4.26) $\varepsilon = \max_{i,j,k} |a_{ij}^{(k)}| / \|A\|_\infty$,

και όπου τα $a_{ij}^{(k)}$ είναι οι μετασχηματισμοί των a_{ij} κατά την απαλοιφή.

Η σταθερά $1.01(n^3+3n^2)$ στην (4.25) είναι ένα πολύ "απαισιόδοξο" φράγμα και μπορεί να αχνοηθεί στην πράξη.

Με μια συγκεκριμένη στρατηγική <u>οδήγησης</u> μπορούμε τώρα να <u>ελέγξουμε</u> το μέγεθος του ε. Ο Wilkinson έδειξε ότι για την απαλοιφή Gauss με ολική οδήγηση το ε αυξάνεται πολύ αργά ως συνάρτηση του n ενώ υπάρχουν ακραία παραδείγματα όπου για μερική οδήγηση το ε αυξάνεται εκθετικά με το n. Όπως όμως έχει παρατηρηθεί στην πράξη και όπως ο ίδιος ο Wilkinson είχε υπολογίσει, είναι εξαιρετικά σπάνιο φαινόμενο να πάρουμε, με μερική οδήγηση π.χ., ε μεγαλύτερο του 10.

Για να κάνουμε τώρα ορισμένες ακόμα παρατηρήσεις σχετικά με το σφάλμα της προσεγγιστικής λύσης $\tilde{x}$, υποθέτουμε ότι ισχύει η (4.24) και κάνουμε την υπόθεση εργασίας ότι για κάποια φυσική νόρμα $\|.\|$ ισχύει

(4.25) $\|\delta A\|/\|A\| = p\beta^{-t}$.

(Στην πράξη εκτός από σπανιότατες περιπτώσεις, έχει παρατηρηθεί ότι με μερική οδήγηση, το p είναι της τάξης μεγέθους του β). Από τις (4.24) και (4.25) μπορούμε να εκτιμήσουμε το υπόλοιπο της προσεγγιστικής λύσης $\tilde{x}$ και το σχετικό της σφάλμα. Για το <u>υπόλοιπο</u> r έχουμε, από την (4.24),

$r = A\tilde{x}-b = -(\delta A)\tilde{x}.$

Συνεπώς $\|r\| \le \|\delta A\|\,\|\tilde{x}\|$, οπότε η (4.25) (αν $A,\ \tilde{x}\neq 0$) δίνει

(4.26) $\|r\|/(\|A\|\,\|\tilde{x}\|) \le p\beta^{-t}$.

Μ' άλλα λόγια το "σχετικό" υπόλοιπο (δηλ. το υπόλοιπο δια του γινομένου $\|A\|.\|\tilde{x}\|$, που μετράει την "κλίμακα" του προβλήματος, δηλ. το μέγεθος των δεδομένων A και της υπολογιστικής λύσης $\tilde{x}$) είναι της τάξης $p\beta^{-t}$. Επειδή συνήθως το p είναι της τάξης του β, συμπεραίνουμε ότι το υπόλοιπο της υπολογιστικής (προσεγγιστικής) λύσης που παράγει η απαλοιφή Gauss είναι, <u>ανεξάρτητα απ' την κατάσταση του πίνακα A, μικρό, και μάλιστα της τάξης β^{1-t}, δηλ. του σφάλματος στρογγύλευσης</u>! (Πρβλ. Άσκηση 2.15). Συνεπώς, αν σε κάποια εφαρμογή μας ενδιαφέρει να υπολογίσουμε ένα $\tilde{x}$ που να έχει μικρό υπόλοιπο $r=A\tilde{x}-b$, μπορούμε άφοβα να κάνουμε απαλοιφή με μερική οδήγηση, ανεξάρτητα από την <u>κατάσταση</u> του A. (Μάλιστα, η ανισότητα (4.26) ισχύει και για μη αντιστρέψιμο A -αρκεί βέβαια να υπάρχει λύση $\tilde{x}$ του (4.24)-).

Προχωρούμε τώρα στην μελέτη του εφάλματος της υπολογιστικής λύσης $\tilde{x}$. Αν ο A είναι αντιστρέψιμος, οι (4.24), (4.25) και (4.11) δίνουν

$$(4.27)\quad \|x-\tilde{x}\|/\|x\| \le [\kappa(A)/(1-\kappa(A)p\beta^{-t})]\, p\beta^{-t},$$

υπό την προϋπόθεση ότι ισχύει η $\|\delta A\|\,\|A^{-1}\| = \kappa(A)\, p\beta^{-t}<1$. δηλ. βλέπουμε τώρα ότι το σχετικό σφάλμα της $\tilde{x}$ θα είναι μικρό, αν ο δείκτης κατάστασης $\kappa(A)$ είναι μικρός, αλλά ότι μπορεί να είναι αρκετά μεγάλο, αν $\kappa(A)>>1$.

Η ανισότητα (4.27) δεν είναι και πολύ χρήσιμη στην πράξη γιατί το σχετικό σφάλμα $\|x-\tilde{x}\|/\|x\|$ υπολογίζεται ως πρός το $\|x\|$, δηλ. ως πρός μια άγνωστη ποσότητα στον παρανομαστή, ενώ ο αλγόριθμος υπολογίζει το $\tilde{x}$. Αλλά μπορούμε να δούμε από τον ορισμό του υπολοίπου $r=A\tilde{x}-b$ ότι

$$x-\tilde{x} = A^{-1}b-\tilde{x} = A^{-1}(b-A\tilde{x}) = -A^{-1}r.$$

Συνεπώς, από την (4.26),

$$\|x-\tilde{x}\| \le \|A^{-1}\|\,\|r\| \le \|A^{-1}\|\,\|A\|\,\|\tilde{x}\|\, p\beta^{-t},$$

δηλ. ισχύει η ανισότητα

$$(4.28)\quad \|x-\tilde{x}\|/\|\tilde{x}\| \le \kappa(A)\, p\beta^{-t},$$

που ήταν η ζητούμενη. Επειδή το $\|\tilde{x}\|$ είναι χωστό, η (4.28) δίνει μια εκτίμηση του απολύτου σφάλματος $\|x-\tilde{x}\| \le \|\tilde{x}\|\,\kappa(A)\, p\beta^{-t}$ συναρτήσει της υπολογιστικής λύσης $\tilde{x}$ - ένα παράδειγμα εκτίμησης εκ των υστέρων (a posteriori), δηλ. με χώση του αλγοριθμικού αποτελέσματος $\tilde{x}$, σε αντίθεση με την εκτίμηση του $\|x-\tilde{x}\|$ που δίνει η (4.27) που είναι εκτίμηση εκ των προτέρων (a priori), δηλ. διατυπώνεται συναρτήσει της ακριβούς λύσης x.

Κλείνουμε αυτό το κεφάλαιο με ορισμένες παρατηρήσεις

1. Σε ορισμένες περιπτώσεις είναι δυνατόν, πολλαπλασιάζοντας πρίν λύσουμε το σύστημα από αριστερά και δεξιά του πίνακα A επί διαγωνίους πίνακες (που αντιστοιχεί σε πολλαπλασιασμό των γραμμών και

των στηλών του A με σταθερές, δηλ. σε μια εκ των προτέρων στάθμιση-αλλαχή κλίμακας (scaling) - των στοιχείων του A) να πάρουμε ένα σύστημα με μικρότερο δείκτη κατάστασης, βλ. [B1].[B3]. Τέτοιοι κατάλληλοι διαγώνιοι πίνακες μπορούν να βρεθούν για ειδικές περιπτώσεις πινάκων A αλλά το γενικό πρόβλημα του υπολογισμού τους δεν έχει διελευκανθεί ακόμα αρκετά.

2. Αν ο δείκτης κατάστασης ενός πίνακα δεν είναι πολύ μεγάλος σχετικά με την χρησιμοποιούμενη ακρίβεια, δηλ. αν $\kappa(A)u<<1$, είναι δυνατόν, χρησιμοποιώντας τεχνικές επαναληπτικής βελτίωσης (iterative improvement) να βελτιώσουμε την ακρίβεια της υπολογιστικής λύσης $\tilde{x}$. Μια τέτοια τεχνική υπολογίζει το υπόλοιπο της $\tilde{x}$ και λύνει ένα νέο σύστημα με πίνακα A και δεύτερο μέλος το υπόλοιπο r, υπολογίζοντας μια "διόρθωση", η οποία, προστιθέμενη στη λύση $\tilde{x}$, δίνει μια νέα προσέγγιση, που συχνά είναι ακριβέστερη της $\tilde{x}$ εφ' όσον το υπόλοιπο r έχει υπολογιστεί με μεγαλύτερη ακρίβεια (π.χ. διπλή) απ' ότι οι υπόλοιπες πράξεις. Βλ. π.χ. [A1],[B1],[B3],[B6].

3. Αν γνωρίζουμε τον δείκτη κατάστασης ενός πίνακα, μπορούμε, χρησιμοποιώντας π.χ. την ανισότητα (4.28), να εκτιμήσουμε το σφάλμα $\|x-\tilde{x}\|$ της υπολογιστικής λύσης $\tilde{x}$. Βεβαίως $\kappa(A)=\|A\|\,\|A^{-1}\|$, αλλά θέλουμε ν' αποφύγουμε τον υπολογισμό του A^{-1}, δηλ. μια πράξη κατά πολύ ακριβώτερη από την λύση του συστήματος. Στην πράξη λοιπόν δεν υπολογίζουμε ακριβώς τον δείκτη κατάστασης αλλά κάνουμε μια προσεγγιστική του εκτίμηση αφού εξ' άλλου συνήθως μόνο η γνώση της τάξης του $\kappa(A)$ μας χρειάζεται.

Μια τεχνική για την εκτίμηση του $\kappa(A)$ είναι η εξής: Λύνουμε μερικά συστήματα $Aw^i=y^i$, $i=1,2,\ldots,k$ (k συνήθως 2 ή 3) με τυχαία δεύτερα μέλη y^i. (Υπενθυμίζεται ότι ο A έχει ήδη τριγωνοποιηθεί για τη λύση του συστήματος $Ax=b$. Έτσι, ο υπολογισμός κάθε w^i απαιτεί μόνο $O(n^2)$ πράξεις). Υπολογίζουμε κατόπιν τα πηλίκα $\|w^i\|/\|y^i\|$ και προσεγγίζουμε κατόπιν $\|A^{-1}\| \cong \max_{1\le i\le k} (\|w^i\|/\|y^i\|)$. (Βέβαια, η πραγματική τιμή του $\|A^{-1}\|$ είναι ίση με το $\sup_{y\neq 0} (\|A^{-1}y\|/\|y\|) = \sup_{y\neq 0} (\|w\|/\|y\|,\ w=A^{-1}y \Leftrightarrow$ w λύση συστήματος $Aw=y$)).

Στο υποπρόγραμμα DECOMP του εργαστηρίου του μαθήματος ο δείκτης κατάστασης $\kappa_1(A)$ προσεγγίζεται ως $\kappa_1(A)\cong\|A\|_1\|z\|_1/\|y\|_1$, όπου $\|A^{-1}\|_1\cong\|z\|_1/\|y\|_1$. Τα διανύσματα z,y υπολογίζονται ως λύσεις δυο συστημάτων $A^Ty=e$, $Az=y$, όπου e ένα κατάλληλα επιλεγμένο διάνυσμα του οποίου οι συνιστώσες είναι ίσες με ± 1. Για την θεωρητική απόδειξη του ότι αυτό είναι μια καλή εκτίμηση του $\kappa_1(A)$, βλ.[B3],[B6].

Προχωρούμε τώρα στην μελέτη του <u>σφάλματος</u> της υπολογιστικής λύσης $\tilde{x}$. Αν ο A είναι αντιστρέψιμος, οι (4.24), (4.25) και (4.11) δίνουν

$$(4.27)\quad \|x-\tilde{x}\|/\|x\| \le [\kappa(A)/(1-\kappa(A)p\beta^{-t})]\, p\beta^{-t},$$

υπό την προϋπόθεση ότι ισχύει η $\|\delta A\|\,\|A^{-1}\| = \kappa(A)\, p\beta^{-t}<1$. Δηλ. βλέπουμε τώρα ότι το <u>σχετικό σφάλμα</u> της $\tilde{x}$ θα είναι μικρό, αν ο δείκτης κατάστασης $\kappa(A)$ είναι μικρός, αλλά ότι μπορεί να είναι αρκετά μεγάλο, αν $\kappa(A)>>1$.

Η ανισότητα (4.27) δεν είναι και πολύ χρήσιμη στην πράξη γιατί το σχετικό σφάλμα $\|x-\tilde{x}\|/\|x\|$ υπολογίζεται ως πρός το $\|x\|$, δηλ. ως πρός μια <u>άγνωστη</u> ποσότητα στον παρανομαστή, ενώ ο αλγόριθμος υπολογίζει το $\tilde{x}$. Αλλά μπορούμε να δούμε από τον ορισμό του υπολοίπου $r=A\tilde{x}-b$ ότι

$$x-\tilde{x} = A^{-1}b-\tilde{x} = A^{-1}(b-A\tilde{x}) = -A^{-1}r.$$

Συνεπώς, από την (4.26),

$$\|x-\tilde{x}\| \le \|A^{-1}\|\,\|r\| \le \|A^{-1}\|\,\|A\|\,\|\tilde{x}\|\, p\beta^{-t},$$

δηλ. ισχύει η ανισότητα

$$(4.28)\quad \|x-\tilde{x}\|/\|\tilde{x}\| \le \kappa(A)\, p\beta^{-t},$$

που ήταν η ζητούμενη. Επειδή το $\|\tilde{x}\|$ είναι γνωστό, η (4.28) δίνει μια εκτίμηση του <u>απολύτου σφάλματος</u> $\|x-\tilde{x}\| \le \|\tilde{x}\|\,\kappa(A)\, p\beta^{-t}$ συναρτήσει της υπολογιστικής λύσης $\tilde{x}$ - ένα παράδειγμα εκτίμησης <u>εκ των υστέρων</u> (a posteriori), δηλ. με χρήση του αλγοριθμικού αποτελέσματος $\tilde{x}$, σε αντίθεση με την εκτίμηση του $\|x-\tilde{x}\|$ που δίνει η (4.27) που είναι εκτίμηση <u>εκ των προτέρων</u> (a priori), δηλ. διατυπώνεται συναρτήσει της ακριβούς λύσης x.

Κλείνουμε αυτό το κεφάλαιο με ορισμένες παρατηρήσεις

1. Σε ορισμένες πριπτώσεις είναι δυνατόν, πολλαπλασιάζοντας πρίν λύσουμε το σύστημα από αριστερά και δεξιά του πίνακα A επί <u>διαγωνίους</u> πίνακες (που αντιστοιχεί σε πολλαπλασιασμό των γραμμών και

των στηλών του A με σταθερές, δηλ. σε μια εκ των προτέρων στάθμιση-αλλαχή κλίμακας (scaling) - των στοιχείων του A) να πάρουμε ένα σύστημα με μικρότερο δείκτη κατάστασης, βλ. [B1].[B3]. Τέτοιοι κατάλληλοι διαγώνιοι πίνακες μπορούν να βρεθούν χια ειδικές περιπτώσεις πινάκων A αλλά το γενικό πρόβλημα του υπολογισμού τους δεν έχει διελευκανθεί ακόμα αρκετά.

2. Αν ο δείκτης κατάστασης ενός πίνακα δεν είναι πολύ μεγάλος, σχετικά με την χρησιμοποιούμενη ακρίβεια, δηλ. αν $\kappa(A)u<<1$, είναι δυνατόν, χρησιμοποιώντας τεχνικές <u>επαναληπτικής βελτίωσης</u> (iterative improvement) να βελτιώσουμε την ακρίβεια της υπολογιστικής λύσης $\tilde{x}$. Μια τέτοια τεχνική υπολογίζει το υπόλοιπο της $\tilde{x}$ και λύνει ένα νέο σύστημα με πίνακα A και δεύτερο μέλος το υπόλοιπο r, υπολογίζοντας μια "διόρθωση", η οποία, προστιθέμενη στη λύση $\tilde{x}$, δίνει μια νέα προσέγγιση, που συχνά είναι ακριβέστερη της $\tilde{x}$ εφ' όσον το υπόλοιπο r έχει υπολογιστεί με μεγαλύτερη ακρίβεια (π.χ. διπλή) απ' ότι οι υπόλοιπες πράξεις. Βλ. π.χ. [A1],[B1],[B3],[B6].

3. Αν γνωρίζουμε τον δείκτη κατάστασης ενός πίνακα, μπορούμε, χρησιμοποιώντας π.χ. την ανισότητα (4.28), να εκτιμήσουμε το σφάλμα $\|x-\tilde{x}\|$ της υπολογιστικής λύσης $\tilde{x}$. Βεβαίως $\kappa(A)=\|A\|\,\|A^{-1}\|$, αλλά θέλουμε ν' αποφύγουμε τον υπολογισμό του A^{-1}, δηλ. μια πράξη κατά πολύ ακριβώτερη από την λύση του συστήματος. Στην πράξη λοιπόν δεν υπολογίζουμε ακριβώς τον δείκτη κατάστασης αλλά κάνουμε μια προσεγγιστική του εκτίμηση αφού εξ' άλλου συνήθως μόνο η γνώση της τάξης του $\kappa(A)$ μας χρειάζεται.

Μια τεχνική χια την εκτίμηση του $\kappa(A)$ είναι η εξής: Λύνουμε μερικά συστήματα $Aw^i=y^i$, $i=1,2,\ldots,k$ (k συνήθως 2 ή 3) με <u>τυχαία</u> δεύτερα μέλη y^i. (Υπενθυμίζεται ότι ο A έχει ήδη τριγωνοποιηθεί χια τη λύση του συστήματος $Ax=b$. Έτσι, ο υπολογισμός κάθε w^i απαιτεί μόνο $O(n^2)$ πράξεις). Υπολογίζουμε κατόπιν τα πηλίκα $\|w^i\|/\|y^i\|$ και προσεγγίζουμε κατόπιν $\|A^{-1}\| \cong \max_{1\le i\le k} (\|w^i\|/\|y^i\|)$. (Βέβαια, η <u>πραγματική</u> τιμή του $\|A^{-1}\|$ είναι ίση με το $\sup_{y\neq 0} (\|A^{-1}y\|/\|y\|) = \sup_{y\neq 0} (\|w\|/\|y\|,\ w=A^{-1}y \Leftrightarrow$ w λύση συστήματος $Aw=y$)).

Στο υποπρόγραμμα DECOMP του εργαστηρίου του μαθήματος ο δείκτης κατάστασης $\kappa_1(A)$ προσεγγίζεται ως $\kappa_1(A)\cong\|A\|_1\|z\|_1/\|y\|_1$, όπου $\|A^{-1}\|_1\cong\|z\|_1/\|y\|_1$. Τα διανύσματα z,y υπολογίζονται ως λύσεις δυο συστημάτων $A^Ty=e$, $Az=y$, όπου e ένα κατάλληλα επιλεγμένο διάνυσμα του οποίου οι συνιστώσες είναι ίσες με ± 1. Για την θεωρητική απόδειξη του ότι αυτό είναι μια καλή εκτίμηση του $\kappa_1(A)$, βλ.[B3],[B6].