ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΑΝΟΙΚΤΑ ακαδημαϊκά μαθήματα ΠΠ

Επιστημονικός Υπολογισμός Ι

Ενότητα **4** : Μοντέλο Αριθμητικής και Σφάλματα Υπολογισμού

Ευστράτιος Γαλλόπουλος
Τμήμα Μηχανικών Η/Υ & Πληροφορικής

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ 2007-2013
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο **«Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Πατρών»** έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σκοπός Ενότητας

- Απώλεια πληροφορίας στον επιστημονικό υπολογισμό.
- Αριθμητικό μοντέλο και πρότυπο αριθμητικής κινητής υποδιαστολής IEEE.
- Σφάλματα στρογγύλευσης και διάδοσή τους.
- Σφάλματα στρογγύλευσης και διάδοσή τους.
- Δείκτες κατάστασης προβλήματος και αλγορίθμου.
- Θεωρία και εργαλεία εκτίμησης σφάλματος και ποιότητας υπολογισμών.

Περιεχόμενα

1. Υπενθυμίσεις και μεθοδολογία πίσω ανάλυσης σφάλματος
2. Παραδείγματα εφαρμογής πίσω ανάλυσης σφάλματος
3. Σφάλματα στη MM Strassen
4. Το πρόβλημα της άθροισης
5. Προς μια ακριβέστερη αριθμητική

Γενική εικόνα

Στόχοι της «προς τα πίσω ανάλυσης σφάλματος» (Hig02)

Ορισμός Ονομάζουμε «προς τα πίσω ανάλυση σφάλματος» τη διαδικασία εύρεσης φράγματος για το πίσω σφάλμα.

Στόχοι:

1. Να ερμηνεύσουμε τα σφάλματα στρογγύλευσης στους υπολογισμούς ως ισοδύναμες διαταραχές στα δεδομένα.
2. Να αναχθεί το πρόβλημα της εύρεσης φράγματος ή της εκτίμησης του εμπρός σφάλματος σε πρόβλημα της μαθηματικής θεωρίας διαταραχών.

Μεθοδολογία

Ισχύει ότι

$$\frac{\|f_{\text{prog}}(x^*) - f(x)\|}{\|f(x)\|} \leq \text{cond}(f; x^*)\text{cond}(f_{\text{prog}})\mathbf{u}\frac{\|f(x^*)\|}{\|f(x)\|} + \text{cond}(f; x)\mathcal{E}$$

Μεθοδολογία

Ισχύει ότι

$$\frac{\|f_{\text{prog}}(x^*) - f(x)\|}{\|f(x)\|} \leq \text{cond}(f; x^*)\text{cond}(f_{\text{prog}})\mathbf{u}\frac{\|f(x^*)\|}{\|f(x)\|} + \text{cond}(f; x)\mathcal{E}$$

Βήματα για να φράξουμε το *εμπρός σφάλμα* $\frac{\|f_{\text{prog}}(x^*)-f(x)\|}{\|f(x)\|}$ χρησιμοποιώντας την πίσω ανάλυση σφάλματος:

Υπολογισμός κατάστασης αλγορίθμου $\text{cond}(f_{\text{prog}})$. από την οποία προκύπτει το πίσω σφάλμα.

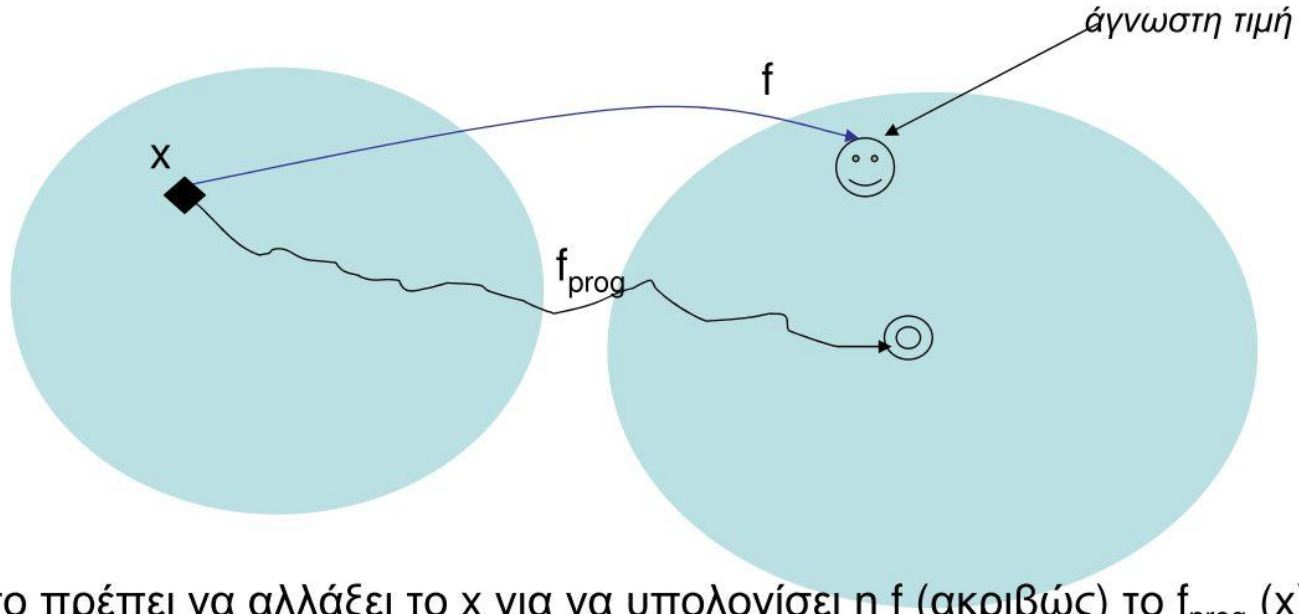
Υπολογισμός κατάστασης προβλήματος $\text{cond}(f; x)$

Εκτίμηση του $\mathcal{E}$ όταν χρειάζεται

Όταν ο παράγοντας $\mathcal{E}$ είναι μικρός και $\frac{\|f(x^*)\|}{\|f(x)\|} \approx 1$:

προς τα εμπρός σφάλμα $<$ δείκτης κατάστασης $\times$ πίσω σφάλμα

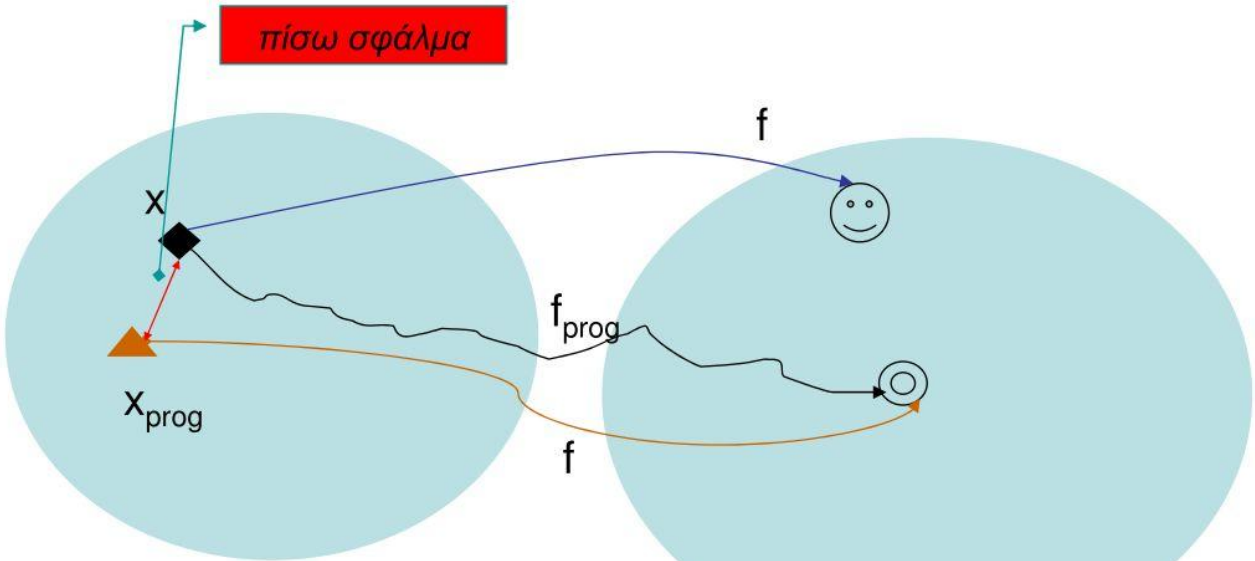
Το πρόβλημα: Να εκτιμήσουμε φράγμα για το εμπρός σφάλμα



Πόσο πρέπει να αλλάξει το x για να υπολογίσει η f (ακριβώς) το f_{prog} (x)?

$$f(x+?) = f_{prog}(x)$$

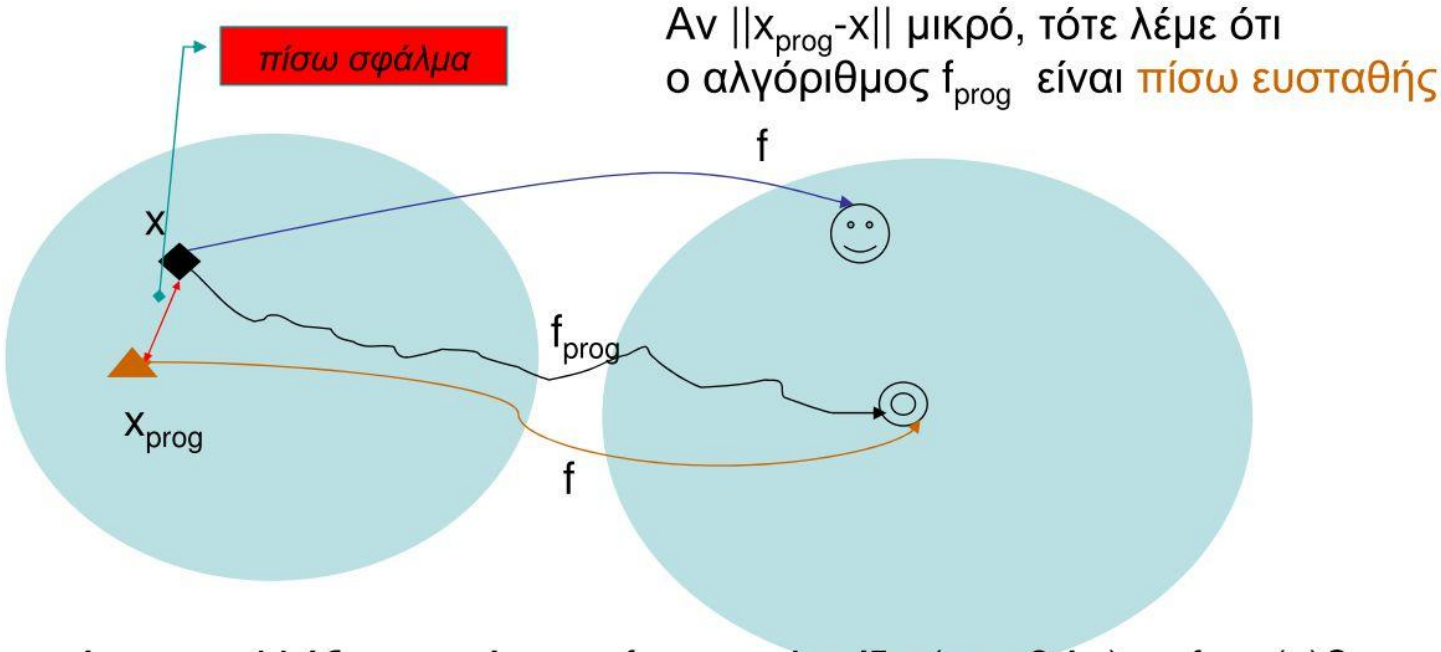
x_{prog}-x



Πόσο πρέπει να αλλάξει το x ώστε η f να υπολογίζει (ακριβώς) το f_{prog} (x)?

$$f(x+?) = f_{prog}(x)$$

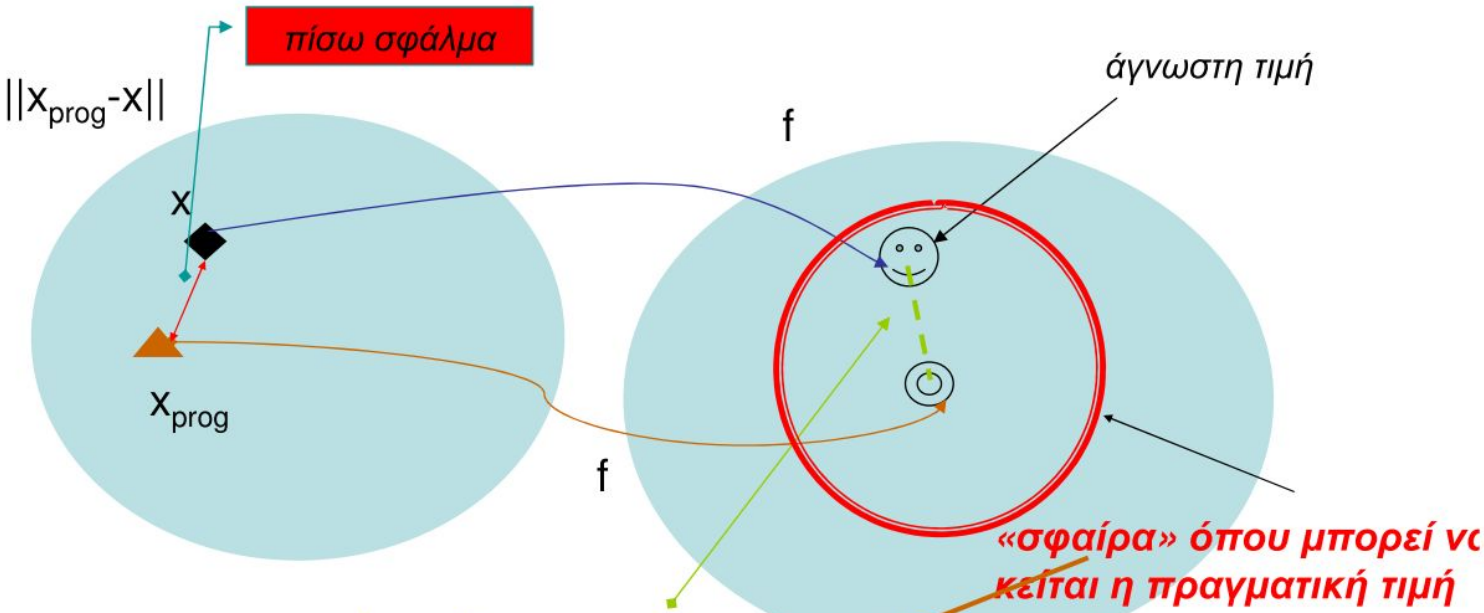
x_{prog}-x



Πόσο πρέπει να αλλάξει το x ώστε η f να υπολογίζει (ακριβώς) το f_{prog} (x)?

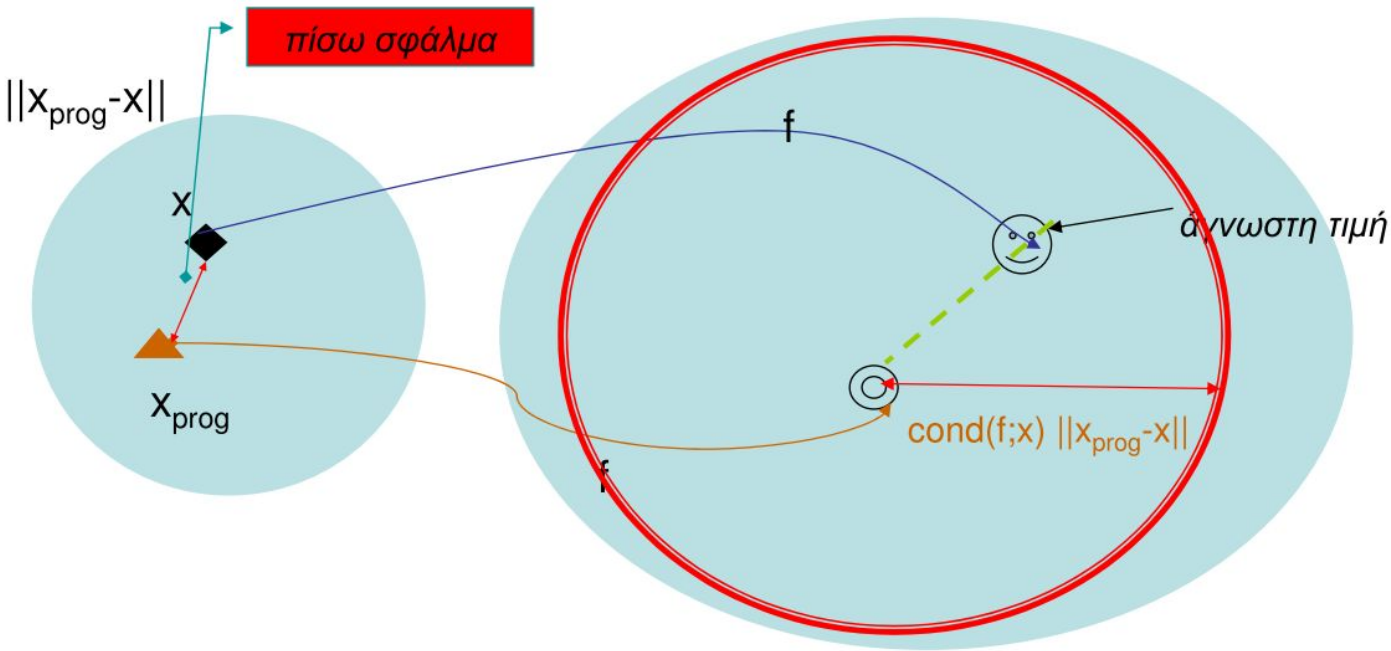
$f(x+?) = f_{prog}(x)$

$x_{prog}-x$

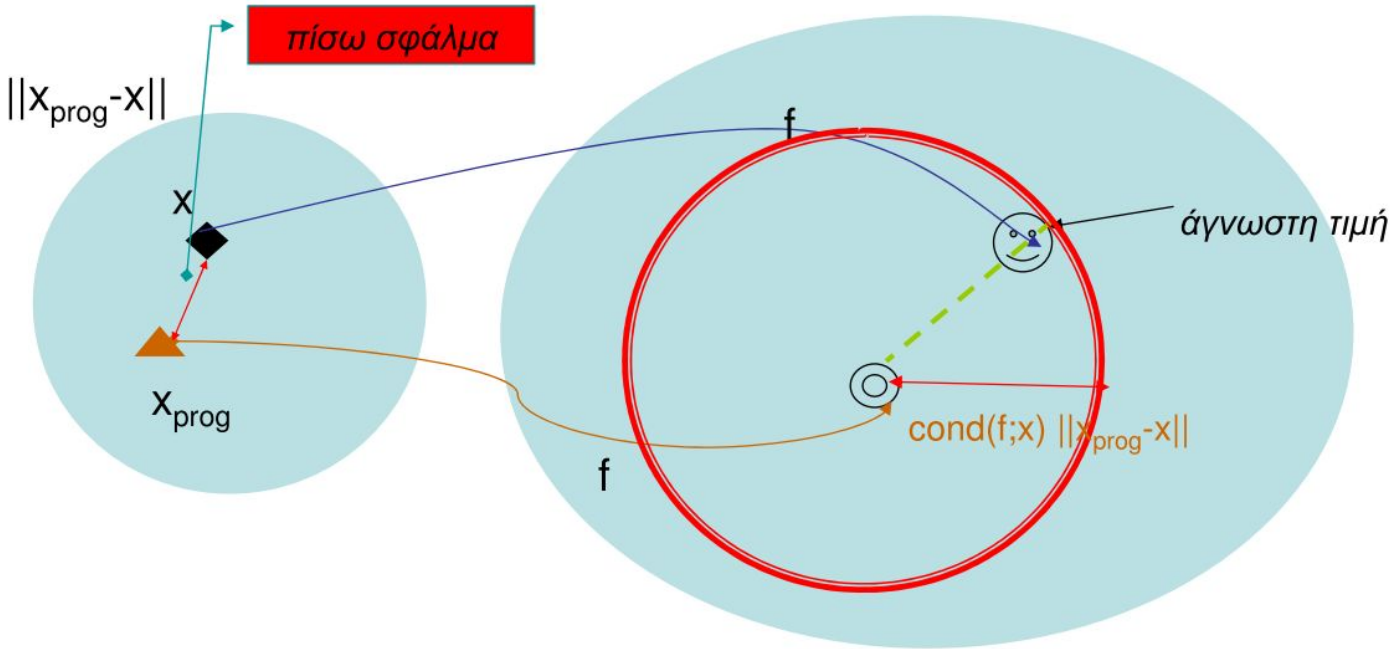


Η εκτίμηση του εμπρός σφάλματος γίνεται «μαθηματικά» βάσει
•του «δείκτη κατάστασης του προβλήματος» (δηλ. της f)
•χρησιμοποιώντας το πίσω σφάλμα.

$||f_{prog}(x) - f(x)|| \leq \mathbf{cond(f;x)}\ ||x_{prog}-x||$



Αν η συνάρτηση είναι πολύ ευαίσθητη, το cond(f;x) είναι μεγάλο (π.χ. αν $\frac{df}{dx}(x)$ μεγάλο) επομένως το f(x) μπορεί να απέχει πολύ από το $f(x_{prog})$



Για δεδομένο cond(f;x), η $||f(x_{prog}-f(x)||$ θα κυμαίνεται ανάλογα με το πίσω σφάλμα $||x_{prog}-x||$

Πίσω ανάλυση σφάλματος στο DOT I

Αν $s_n = x^\top y$ τότε

$$\begin{aligned}
\tilde{s}_1 &= \mathrm{fl}(\xi_1\psi_1) = \xi_1\psi_1(1+\delta_1) \\
\tilde{s}_2 &= \mathrm{fl}(\tilde{s}_1 + \mathrm{fl}(\xi_2\psi_2)) \\
&= (\xi_1\psi_1(1+\delta_1) + \xi_2\psi_2(1+\delta_2))(1+\delta_3) \\
&= \xi_1\psi_1(1+\delta_1)(1+\delta_3) + \xi_2\psi_2(1+\delta_2)(1+\delta_3)
\end{aligned}$$

όπου $|\delta_i| \leq u$.

Πίσω ανάλυση σφάλματος στο DOT II

$$\tilde{s}_n = \xi_1\psi_1 \prod_{\substack{j=1 \\ j\neq 2}}^{n+1}(1+\delta_j) + \xi_2\psi_2\prod_{j=2}^{n+1}(1+\delta_j) + \ldots \xi_3\psi_3\prod_{j=3}^{n+1}(1+\delta_j) + \cdots + \xi_n\psi_n\prod_{j=n}^{n+1}(1+\delta_j).$$

Από το Λήμμα:

$$\tilde{s}_n = \xi_1\psi_1(1+\theta_n) + \xi_2\psi_2(1+\hat{\theta}_n) + \ldots + \xi_3\psi_3(1+\theta_{n-1}) + \cdots + \xi_n\psi_n(1+\theta_2).$$

Το DOT που υπολογίστηκε στην α.κ.υ. είναι η ακριβής τιμή του εσωτερικού γινομένου για τα διανύσματα $x = (\xi_1(1+\theta_n), \xi_2(1+\hat{\theta}_n)..., \xi_n(1+\theta_2))^\top$, και $(\psi_1, \psi_2, ..., \psi_n)^\top$, όπου για τα θ_j ισχύει το φράγμα

$$|\theta_j| \leq \frac{ju}{1-ju} = \gamma_j.$$

Παρατήρηση: Παρόμοια ανάλυση ισχύει χρησιμοποιώντας x και κάποιο $y + \Delta y$

Πίσω ανάλυση σφάλματος στο DOT III

Αποδείξαμε $\mathrm{fl}(x^\top y) = (x + \Delta x)^\top y =$ όπου $|\Delta x| \leq \gamma_n |x|, |\Delta y| \leq \gamma_n |y|$
Δηλαδή το υπολογισμένο DOT είναι ίδιο με το ακριβές εσωτερικό γινόμενο με στοιχεία εισόδου $x + \Delta x, y$, όπου

$$|\Delta x| \leq \gamma_n |x| \Rightarrow \|\Delta x\|_1 \leq \gamma_n \|x\|_1.$$

Άρα το σχετικό πίσω σφάλμα φράσσεται ως εξής

$$\frac{\|\Delta x\|_1}{\|x\|_1} \leq \gamma_n$$

και μπορούμε να λάβουμε ως δείκτη κατάστασης του αλγορίθμου $\mathrm{cond}(f_{\mathrm{prog}}) = \frac{\gamma_n}{\mathbf{u}}$.

ο απλός αλγόριθμος για το εσωτερικό γινόμενο (πολλαπλασιασμός στοιχείο προς στοιχείο και άθροιση από τα αριστερά προς τα δεξιά) είναι προς τα πίσω ευσταθής

ΠΡΟΣΟΧΗ Το ακριβές φράγμα μπορεί να εξαρτάται από τη σειρά υπολογισμού. Π.χ. το παραπάνω αντιστοιχεί στην άθροιση (από τα αριστερά προς τα δεξιά).

Δείκτης κατάστασης προβλήματος «εσωτερικού γινομένου»

$$f([x;y]) := \sum_{j=1}^{n} \xi_j \psi_j = x^\top y, \quad x, y \in \mathbb{R}^n$$

Θέτουμε $X := [x; y] \in \mathbb{R}^{2n}$ και υπολογίζουμε το δείκτη κατάστασης με βάση την «ευαισθησία» του αποτελέσματος ως προς κάθε στοιχειο:

$$\begin{aligned} K &= \frac{1}{|x^\top y|}\left[\left|\xi_1 \frac{\partial f}{\partial \xi_1}\right|, \ldots, \left|\xi_n \frac{\partial f}{\partial \xi_n}\right|, \left|\psi_1 \frac{\partial f}{\partial \psi_1}\right|, \ldots, \left|\psi_n \frac{\partial f}{\partial \psi_n}\right|\right] \\ &= \frac{1}{|x^\top y|}\left[|\xi_1\psi_1|, \ldots, |\xi_n\psi_n|, |\xi_1\psi_1|, \ldots, |\xi_n\psi_n|\right] \end{aligned}$$

επομένως

$$\|K\|_1 = \frac{2\sum_{j=1}^{n} |\xi_j||\psi_j|}{|\sum_{j=1}^{n} \xi_j\psi_j|}$$

Προς τα εμπρός σφάλμα

Προς τα πίσω ανάλυση $\frac{|x^T y - \mathrm{fl}(x^T y)|}{|x^T y|} \leq \gamma_n \frac{2\sum_{j=1}^{n} |\xi_j||\psi_j|}{|\sum_{j=1}^{n} \xi_j \psi_j|}$

Προς τα εμπρός ανάλυση Χρησιμοποιώντας προς τα εμπρός ανάλυση συμπεραίνουμε ότι

$$\begin{aligned} |x^T y - \mathrm{fl}(x^T y)| &\leq \gamma_n \sum_{i=1}^{n} |\xi_i||\psi_i| = \gamma_n |x|^T |y| \\ &\leq nu|x|^T|y| + O(u^2) \end{aligned}$$

Παράδειγμα: πίσω ανάλυση σφάλματος της μεθόδου Horner

$s_n = \alpha_n$
for $k = n-1 : -1 : 0$
  $s_k = x s_{k+1} + \alpha_k$
end

Χρησιμοποιώντας το γνωστό λήμμα:

$$\begin{aligned}
\hat{s}_{n-1} &= (x s_n(1+\delta_1) + \alpha_{n-1})(1+\delta_2) = x\alpha_n(1+\theta_2) + \alpha_{n-1}(1+\delta_2) \\
\hat{s}_{n-2} &= (x\hat{s}_{n-1}(1+\delta_3) + \alpha_{n-2})(1+\delta_4) \\
\hat{s}_0 &= \alpha_0(1+\delta) + \alpha_1 x(1+\theta_3) + \cdots \alpha_{n-1}x^{n-1}(1+\theta_{2n-1}) + \alpha_n x^n(1+\theta_{2n}) \\
&= f_{\text{prog}}(\alpha_0, ..., \alpha_n, x) = f(\alpha_0(1+\theta_1), ..., \alpha_n(1+\theta_{2n}), x)
\end{aligned}$$

για κάποια θ_j φραγμένα σε απόλυτη τιμή από τα αντίστοιχα γ_j.

Επομένως, το προς τα πίσω σφάλμα της μεθόδου Horner είναι μικρό και η μέθοδος είναι πίσω ευσταθής.

Ειδικότερα:

Αφού $|\theta_{2n}| \leq \gamma_{2n}$ η μέγιστη ασάφεια στους συντελεστές θα είναι γ_{2n}. Από τα παραπάνω προκύπτει ότι αν κάθε συντελεστής α_j δεν είναι γνωστός ακριβώς αλλά γνωρίζουμε προσέγγισή του τ.ώ. $|\alpha_j - \hat{\alpha}_j| \leq \gamma_{2n}|\alpha_j|$, το σφάλμα στο αποτέλεσμα από την ασάφεια αυτή θα είναι μεγαλύτερο ή ίσο από το τελικό σφάλμα που θα προέκυπτε αν οι πράξεις του Horner εκτελούνταν με α.κ.υ.

Η πίσω ευστάθεια του Horner από μόνη της δεν αρκεί για να εγγυηθούμε ότι το τελικό (σχετικό) προς τα εμπρός σφάλμα στην υπολογισμένη τιμή θα είναι μικρό. Επιτρέπει όμως να το μελετήσουμε ανεξάρτητα από τον αλγόριθμο.

Εύκολα αποδεικνύεται η σχέση

$$\frac{|p(x) - \hat{s}_0|}{|p(x)|} \leq \gamma_{2n} \frac{\sum_{k=0}^{n} |\alpha_k||x|^k}{|p(x)|},$$

που συνεπάγεται ότι δεν μπορούμε να εγγυηθούμε μικρό προς τα εμπρός σχετικό σφάλμα. Το σημαντικό είναι όμως ότι αν το εμπρός σφάλμα είναι μεγάλο, αυτό δεν θα οφείλεται στον αλγόριθμο.

Γιατί τόσο ενδιαφέρον για πολυώνυμα

Φαίνονται «εύκολα» αλλά επιπλέον:

- Τα πολυώνυμα είναι από τις πιο σημαντικά «αλγεβρικά αντικείμενα» και η ορθή χρήση τους είναι σημαντικό θέμα.
- Το πρόβλημα του υπολογισμού τιμών πολυωνύμου εμφανίζεται σε πολλές εφαρμογές.
- Για παράδειγμα, αποτελεί υπολογιστικό πυρήνα για την εύρεση των ριζών.
- Προσοχή: Αν και τα πολυώνυμα φαίνονται εύκολα στη διαχείρισή τους, χρειάζονται προσοχή. Ο Wilkinson τα χαρακτήρισε perfidious (ύπουλα, δόλια) στο βραβευμένο άρθρο (Wil84).

Μερικά εμπόδια

Η πίσω ανάλυση σφάλματος:

- Μπορεί να είναι δύσκολη
- Μπορεί να είναι ανέφικτη.

Διαστατικό επιχείρημα: Γενικά αν η διάσταση του διανύσματος εξόδου είναι μεγαλύτερη από τη διάσταση της εισόδου, τότε θα έχουμε δυσκολία να αποδείξουμε πίσω ευστάθεια.

Μερικά εμπόδια

Η πίσω ανάλυση σφάλματος:

- Μπορεί να είναι δύσκολη
- Μπορεί να είναι ανέφικτη.

Διαστατικό επιχείρημα: Γενικά αν η διάσταση του διανύσματος εξόδου είναι μεγαλύτερη από τη διάσταση της εισόδου, τότε θα έχουμε δυσκολία να αποδείξουμε πίσω ευστάθεια.
Παράδειγμα Ο υπολογισμός

$$\mathrm{fl}(C) = \mathrm{fl}(ab^\top), a, b \in \mathbb{R}^n$$

άρα

$$\mathrm{fl}(\gamma_{ij}) = \alpha_i \beta_j (1 + \delta_{ij})$$

που σημαίνει ότι

$$\begin{aligned} \mathrm{fl}(C) &= ab^\top + E \Rightarrow \\ |C - \mathrm{fl}(C)| &= |E| \leq \mathbf{u} e e^\top \end{aligned}$$

όπου $[E]_{ij} = \alpha_i \beta_j \delta_{ij}$ και $ee^\top$ είναι ίσο με το μητρώο με όλα τα στοιχεία του 1.
Δηλ. το καλύτερο δυνατό άνω φράγμα για το εμπρός σφάλμα!

Για πίσω ευστάθεια θάπρεπε να υπάρχουν $\tilde{a}$, $\tilde{b}$ κοντά στα a, b τ.ώ.

$$\mathrm{fl}(C) = \tilde{a}\tilde{b}^\top = (a + \Delta a)(b + \Delta b)^\top$$

Τότε θα ίσχυε

$$E = \overbrace{a\Delta b^\top + \Delta a(b^\top + \Delta b^\top)}^{\text{τάξη} \leq 2}$$

Αν ο αλγόριθμος υπολογισμού εξωτερικού γινομένου ήταν πίσω ευσταθής θα ίσχυε ότι

$$n = \mathrm{rank}(E) \leq 2$$

πράγμα που είναι γενικά αδύνατο (δεν υπάρχουν αρκετά δεδομένα εισόδου στα οποία να «αναθέσουμε» τα σφάλματα στα αποτελέσματα)

Το εξωτερικό γινόμενο διανυσμάτων δεν είναι πίσω ευσταθής πράξη

ΠΡΟΣΟΧΗ: Αυτό δεν σημαίνει ότι ο αλγόριθμος υπολογισμού έχει πρόβλημα! Μόνον ότι δεν μπορούμε να εφαρμόσουμε πίσω ανάλυση σφάλματος για να φράξουμε το εμπρός σφάλμα ....

Δείκτης κατάστασης προβλήματος

Πολλαπλασιασμός μητρώου με διάνυσμα (πράξη MV)

Έστω $f([A; x]) = Ax$, όπου το $A \in \mathbb{R}^{n\times n}$ είναι αντιστρέψιμο. Διεξάγουμε την μελέτη της ευαισθησίας ως προς διαταραχές του x. Ισχύει ότι

$$\sup_{\|h\|\neq 0} \frac{\|f(x+h) - f(x)\|}{\|h\|} = \sup_{\|h\|\neq 0} \frac{\|Ah\|}{\|h\|} = \|A\|$$

Επομένως

$$\begin{aligned} \mathrm{cond}(f; x) &= \frac{\|x\|}{\|Ax\|}\|A\| \\ \mathrm{cond}(f; x) &= \frac{\|A^{-1}b\|}{\|b\|}\|A\| \\ &\leq \|A\|\|A^{-1}\|. \end{aligned}$$

Διατύπωση σε MATLAB

Το μητρώο `hilb(n)` έχει όρους $\alpha_{i,j} = 1/(i+j-1)$. Έστω $A = \texttt{hilb}(4)$.
Χρησιμοποιούμε ευκλείδεια νόρμα.
Τότε $\|A\|_2 = 1.5$. Θέλουμε να υπολογίσουμε κ ώστε

$$\frac{\|A(x+h) - Ax\|}{\|Ax\|} \quad \leq \quad \kappa \frac{\|h\|}{\|x\|}$$

Αν $A = USV^\top$, $x = U(:,4)$, $h = V(:,4)$. Τότε $\|x\| = \|h\| = 1$. Για τις επιλογές αυτές έχουμε

$$\frac{\|A(x+h) - Ax\|}{\|Ax\|} = 1.5 \times 10^4 \leq \kappa.$$

Δηλαδή η αλλοίωση στο αποτέλεσμα μπορεί να είναι σχετικά μεγάλη.
Αν βέβαια $\|Ax\| \gg 0$ δεν θα παρουσιαζόταν τέτοιο πρόβλημα.

Δείκτης κατάστασης μητρώου

Ορισμός

Ο παράγοντας $\kappa(A) := \|A\|\|A^{-1}\|$ ονομάζεται *δείκτης κατάστασης του A.*

- Όταν ο δείκτης υπολογίζεται, χρησιμοποιείται κάποια από τίς νόρμες. Οι νόρμες είναι ισοδύναμες και συνήθως μας ενδιαφέρει η τάξη μεγέθους και ο ρυθμός αύξησης του δείκτη κατάστασης, και όχι η συγκεκριμένη τιμή.
- Αν γνωρίζουμε τις ιδιάζουσες τιμές του A, $\kappa_2(A) := \sigma_{\text{max}}/\sigma_{\text{min}}$.
- Για μερικά μητρεία, το $\kappa(A)$ μεγαλώνει πολύ γρήγορα με το μέγεθος του A.
- Ο ακριβής υπολογισμός του $\kappa(A)$ είναι δαπανηρός!

Σύνοψη μερικών αποτελεσμάτων για τα πίσω σφάλματα

σε πράξεις BLAS

_AXPY: $\mathrm{fl}(y + \alpha x) = y + \Delta y + \alpha(x + \Delta x)$ όπου $|\Delta y| \leq \mathbf{u}|y|$, $|\Delta x| \leq \gamma_2|x|$.

DOT: $\mathrm{fl}(x^\top y) = (x + \Delta x)^\top y$ όπου $|\Delta x| \leq \gamma_n|x|$

MV: $\mathrm{fl}(Ax) = (A + \Delta A)x$ όπου , $|\Delta A| \leq \gamma_n|A|$.

MM: $\mathrm{fl}(AB) = (A + \Delta A)B$, όπου $|\Delta A| \leq \gamma_n|A||B||B^{-1}|$. Προσοχή: Η ευστάθεια εξαρτάται από το B^{-1}.

Εμπρός σφάλμα Strassen

Θεώρημα (Brent)

Έστω $A, B \in \mathbb{R}^{n\times n}$ με $n = 2^k$ και έστω ότι $C = AB$ υπολογίζεται με την *Strassen* ενώ από την διάσταση $n_0 = 2^{r+1}$ και κάτω χρησιμοποιείται κλασικός πολλαπλασιασμός. Τότε για το υπολογιζόμενο $\hat{C}$ ισχύει:

$$\|\hat{C} - C\| \leq \left[\left(\frac{n}{n_0}\right)^{\log_2 12} (n_0^2 + 5n_0) - 5n \right] u \|A\| \|B\| + O(u^2).$$

όπου η νόρμα $\|A\| := \max_{i,j} |\alpha_{ij}|$.

Παρατηρήσεις

- Αν $n_0 = 1$ ο παράγοντας σφάλματος είναι $\approx 6n^{3,58}$.
- Αν $n_0 = n/2$ ο παράγοντας σφάλματος είναι $\approx 3n^2 + 25n$.

Αποδεικνύεται ότι το εμπρός σφάλμα στον κλασικό πολλαπλασιασμό μητρώων $C = AB$ φράσσεται ως εξής:

$$|\mathrm{fl}(C) - C| \leq \gamma_n |A||B| + \mathrm{O}(u^2).$$

Ένα ενδιαφέρον (αρνητικό) αποτέλεσμα (Mil75)

Ένα σημαντικό αποτέλεσμα του Webb Miller λέει ότι οποιοσδήποτε αλγόριθμος πολλαπλασιασμού μητρώων ικανοποιεί ανισότητα τόσο σφικτή όσο η παραπάνω για το εμπρός σφάλμα, θα πρέπει αναγκαστικά να εκτελεί τουλάχιστον n^3 πολλαπλασιασμούς.

Το πρόβλημα της άθροισης (Higham'02)

Δίδεται σύνολο αριθμών $\mathcal{S} = \{\xi_1, ..., \xi_n\}$ και ζητούμε το άθροισμα $\mathrm{sum}(\mathcal{S})$.

- Μια από τις πιο συχνές πράξεις σε επιστημονικούς και εμπορικούς υπολογισμούς
- π.χ. στον υπολογισμό στατιστικών και άλλων συνοπτικών χαρακτηριστικών, π.χ. μέσος όρος, απόκλιση, νόρμα, ...
- Αξίζει τον κόπο μια πιο προσεκτική θεώρηση
- ... σε σχέση με την *ταχύτητα* και την *ακρίβεια*

Αναδρομική άθροιση

$s = 0$
for $i = 1 : n$
  $s = s + \xi_i$
end

Σχετικά με την αναδρομική άθροιση

- Πολύ απλή υλοποίηση
- Ιδιαίτερα χρήσιμη για ομόσημα στοιχεία μετά από ταξινόμηση σε αύξουσα σειρά κατ' απόλυτη τιμή
- Αποφυγή προβλημάτων «απορρόφησης» από μεγάλα στοιχεία,
- ... πρέπει να είναι γνωστά τα στοιχεία και να ταξινομήσουμε πριν την άθροιση.

Πίσω σφάλμα άθροισης

Όπως φαίνεται εύκολα

$$\begin{aligned} \mathrm{fl}(s) &= (\cdots((\xi_1+\xi_2)(1+\delta_1)+\xi_3)(1+\delta_2)\cdots+\xi_n)(1+\delta_{n-1}) \\ &= \xi_1(1+\theta_{n-1})+\xi_2(1+\hat{\theta}_{n-1})+\xi_3(1+\hat{\theta}_{n-2})+\cdots+\xi_n(1+\theta_1) \end{aligned}$$

όπου ως συνήθως

$$|\theta_j| \leq \gamma_j = \frac{j\mathbf{u}}{1-j\mathbf{u}}$$

Επομένως, το πίσω σφάλμα εξαρτάται άμεσα από το n, όπως θα περιμέναμε...

Μερικές ακόμα μέθοδοι

- Αναγωγική άθροιση ανά ζεύγη[1] που χρησιμοποιείται σε παράλληλες υλοποιήσεις άθροισης
- Ενθετική άθροιση
- Αντισταθμισμένη άθροιση και νεώτερες παραλλαγές

Χρήσιμα χαρακτηριστικά

- είναι οι αριθμοί ομόσημοι;
- είναι η ακολουθία διαθέσιμη από την αρχή;
- ποιό είναι το εύρος (μέγιστο, ελάχιστο);
- είναι η ακολουθία διατεταγμένη/ταξινομημένη;

[1] paiwise / cascade / fan-in

Ενθετική άθροιση

1. ταξινόμηση του $\mathcal{S}$ κατ' απόλυτη τιμή σε αύξουσα σειρά $\mathcal{L} := \xi_1 \leq \xi_2 \leq \xi_3 \leq \cdots \leq \xi_n$
2. Υπολογισμός του $\xi_1 + \xi_2$, διαγραφή των ξ_1, ξ_2 από την $\mathcal{L}$, και ένθεση του $\xi_1 + \xi_2$ στην $\mathcal{L}$ στην κατάλληλη θέση ώστε να διατηρηθεί η μονοτονικότητα. Επαναρίθμηση του $\mathcal{L}$ και αν περιέχει 2 ή περισσότερα στοιχεία, επιστροφή στο (2).
3. Δημιουργείται ένα «δυαδικό δένδρο άθροισης» με φύλλα τους αρχικούς αριθμούς $\{\xi_1, \ldots, \xi_n\}$ και ρίζα την άθροιση από την οποία προκύπτει το τελικό άθροισμα. Οι υπόλοιπο κόμβοι αντιστοιχούν στις ενδιάμεσες αθροίσεις.

Παρατήρηση Η διαδικασία είναι παρόμοια με την προθετική κωδικοποίηση Huffman.

(KW00)

Αν τα δεδομένα έχουν μικτό πρόσημο (θετικά και αρνητικά στοιχεία) η εύρεση του αθροίσματος που ελαχιστοποιεί το μέγιστο σφάλμα στρογγύλευσης για όλους τους δυνατούς τρόπους άθροισης είναι NP-hard.

Συστάσεις

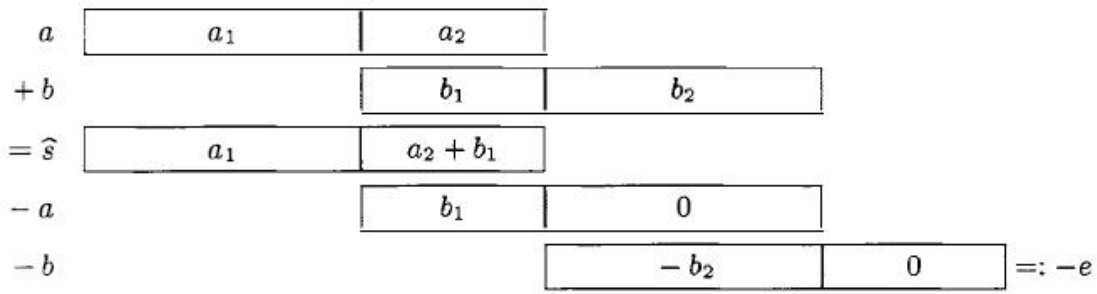
Εφόσον το πρόβλημα κατασκευής του βέλτιστου δένδρου άθροισης είναι πολύ δύσκολο:

- *χαλαρώνουμε* τις απαιτήσεις και αναζητούμε κάτι λιγότερο φιλόδοξο,
- *αξιοποιούμε* ό,τι πληροφορία υπάρχει για το $\mathcal{S}$.

- Αν χρειάζεται μεγάλη ακρίβεια, εξετάστε τη χρήση επανορθωμένης άθροισης ή άθροισης με εκτεταμένη ακρίβεια
- Για τις περισσότερες μεθόδους, το σφάλμα είναι στη χειρότερη περίπτωση ανάλογο με το n. Για πολύ μεγάλο n καθίσταται ενδιαφέρουσα η χρήση της επανορθωμένης άθροισης ή της αναγωγικής άθροισης
- Αν υπάρχει πιθανότητα σημαντικής (ως καταστροφικής) απαλοιφής λόγω μεικτών προσήμων, η αναδρομική άθροιση με φθίνουσα διάταξη είναι συνήθως καλύτερη.
- Αν το σύνολο ομόσημο, όλες οι μέθοδοι προσφέρουν σχετικό σφάλμα το πολύ $n\mathbf{u}$ με καλύτερη την επανορθωμένη άθροιση[2] (compensated summation).

[2] Στις επόμενες διαφάνειες.

Επανορθωμένη άθροιση I



Σχήμα: Ιδιοφυής ιδέα (Gill'51, Kahan'65) από (Higham'02)

Για κάθε α.κ.υ. $|a| \geq |b|$ μπορούμε να υπολογίσουμε

$$\hat{s} = \mathrm{fl}(a+b), \quad \hat{e} = \mathrm{fl}((a-\hat{s})+b)$$

χρησιμοποιώντας στρογγύλευση προς το πλησιέστερο. Το υπολογισμένο $\hat{e}$ είναι το ακριβές σφάλμα τής άθροισης! Ειδικότερα,

$$a+b = \hat{s}+\hat{e}$$

Παρόλα αυτά, $\mathrm{fl}(\hat{s}+\hat{e}) = \hat{s}$ γιατί το αποτέλεσμα $\hat{s} = \mathrm{fl}(a+b)$ είναι το καλύτερο δυνατό (λόγω αρχής ακριβούς στρογγύλευσης). Άρα το $\hat{e}$ από μόνο του δεν παρέχει μεγαλύτερη ακρίβεια.

Επανορθωμένη άθροιση II

Ιδέα επανόρθωσης για 3 τιμές: Αν αθροίζονται 3 τιμές, αποθηκεύουμε το ακριβές σφάλμα $\hat{e}_1$ ως την επόμενη πρόσθεση. Τότε αυτό και το ακριβές σφάλμα της επόμενης πρόσθεσης, έστω $\hat{e}_2$, αν προστεθούν σε $e = \hat{e}_1 + \hat{e}_2$ μπορεί η τιμή να είναι τόσο (σχετικά) μεγάλη που να επανορθώνει την τελική τιμή ώστε το υπολογισμένο $(a + b) + c + e$ να είναι ακριβέστερο του υπολογισμένου $(a + b) + c$. Η ιδέα αυτή μπορεί να επεκταθεί με διάφορους τρόπους σε αθροίσματα περισσοτέρων τιμών.
Πλεονεκτήματα επανορθωμένης άθροισης Γενικά πιο ακριβής από την συνηθισμένη άθροιση.

Εμπρός σφάλμα (Knuth, Kahan)

φράσσεται ως

$$|\sum_{j=1}^{n} \xi_j - \mathrm{fl}\left(\sum_{j=1}^{n} \xi_j\right)| \leq (2\mathbf{u} + O(n\mathbf{u}^2)) \sum_{j=1}^{n} |\xi_j|.$$

Πίσω σφάλμα

Ο παράγοντας 1ης τάξης στο φράγμα δεν εξαρτάται από το n:

$$\mathrm{fl}(s) = \sum_{j=1}^{n} (1 + \mu_j)\xi_j, \quad \text{όπου} \quad |\mu_j| \leq 2\mathbf{u} + O(n\mathbf{u}^2)$$

Επανορθωμένη αριθμητική;

A Floating-Point Technique for Extending the Available Precision

T. J. DEKKER*

Received July 26, 1971

Abstract. A technique is described for expressing multilength floating-point arithmetic in terms of singlelength floating point arithmetic, i.e. the arithmetic for an available (say: single or double precision) floating-point number system. The basic algorithms are exact addition and multiplication of two singlelength floating-point numbers, delivering the result as a doublelength floating-point number. A straightforward application of the technique yields a set of algorithms for doublelength arithmetic which are given as ALGOL 60 procedures.

Let x and y be singlelength floating-point numbers and let

$$z = fl(x+y);$$

i.e. z is the result of a singlelength floating-point addition of x and y. Let zz be the correction term exactly satisfying

$$z+zz=x+y.$$

It will be shown that, under various conditions, zz can be obtained by the formula

$$zz=fl((x-z)+y).$$

Βιβλιογραφία I

N.J. Higham.

Accuracy and Stability of Numerical Algorithms.

SIAM, Philadelphia, 2nd edition, 2002.

Ming-Yang Kao and Jie Wang.

Linear-time approximation algorithms for computing numerical summation with provably small errors.

SIAM J. Comput., 29(5):1568--1576, 2000.

W. Miller.

Computational complexity and numerical stability.

SIAM J. Comput., 4(2):97--107, June 1975.

Ε. Γαλλόπουλος.

Επιστημονικός Υπολογισμός Ι.

Πανεπιστήμιο Πατρών, 2008.

J.H. Wilkinson.

The perfidious polynomial.

In G.H. Golub, editor, *Studies in Numerical Analysis*, volume 24, pages 1--28. Mathematical Association of America, 1984.

Χρήση Έργου Τρίτων Ι

1. κωδικοποίηση Huffman (βλ. σελ 32)
2. http://link.springer.com/article/10.1007%2FBF01397083 (βλ. σελ 36)

Σημείωμα Αναφοράς

``Επιστημονικός Υπολογισμός Ι'', Έκδοση: 1.0, Πάτρα 2013-2014.

Διαθέσιμο από τη δικτυακή διεύθυνση: https://eclass.upatras.gr/courses/CEID1096/

Τέλος Ενότητας

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης