Ανάλυση Σφάλματος στους Αριθμητικούς Υπολογισμούς (floating point)

$x \to \underset{\text{αριθμός μηχανής}}{fl(x)}$, $\left|\frac{x-fl(x)}{x}\right| \le \frac{1}{2}\beta^{1-k\leftarrow \text{ψηφία στη mantisssa}}$ ή $|\varepsilon| \le u$

Παρατήρηση: $(1-u)^n \le 1+E \le (1+u)^n \Leftrightarrow |E| \le nu_1, u_1 = 1.01u$

1. **Υπολογισμός του γινομένου n αριθμών $x_1,\ldots,x_n$ (σε παράσταση με κινητή υποδιαστολή(floating point))**

Ως γνωστό $x \to fl(x) = x(1+\varepsilon)$ (ε μονάδα μηχανής).

Υποθέτουμε ότι οι αριθμοί $x_1, x_2, \ldots, x_n$ είναι αριθμοί μηχανής τότε

το υπολογιζόμενο γινόμενό τους είναι

$$P_n = fl(x_1x_2\ldots x_n) = x_1x_2\ldots x_n(1+E) \overset{(1)}{=} x_1x_2\ldots x_n + \underbrace{x_1x_2\ldots x_nE}_{E'} = x_1x_2\ldots x_n + E'$$

Ορίζουμε αναγωγικά τις ποσότητες:

$P_1 = x_1$

$P_r = fl(P_{r-1}x_r) = P_{r-1}x_r(1+\varepsilon_r),\ \ r = 2,3,\ldots,n,$ όπου $|\varepsilon_r| \le u$

Με διαδοχικές αντικαταστάσεις στον ανωτέρω τύπο

για $r = 2,3,\ldots,n$ προκύπτουν:

$P_2 = fl(P_1x_2) = P_1x_2(1+\varepsilon_2) = x_1x_2(1+\varepsilon_2)$

$P_3 = fl(P_2x_3) = P_2x_3(1+\varepsilon_3) = x_1x_2x_3(1+\varepsilon_2)(1+\varepsilon_3)$

$\vdots$

$P_n = fl(P_{n-1}P_n) = P_{n-1}x_n(1+\varepsilon_n) = x_1x_2\ldots x_n(1+\varepsilon_2)(1+\varepsilon_3)\ldots(1+\varepsilon_n) =$

$= x_1x_2\ldots x_n(1+E),$ όπου $1+\mathrm{E} = (1+\varepsilon_2)(1+\varepsilon_3)\ldots(1+\varepsilon_n)$

Αφού $|\varepsilon_r| \le u \Leftrightarrow 1-u \le 1+\varepsilon_r \le 1+u$

Συνεπώς $(1-u)^{n-1} \le 1+E \le (1+u)^{n-1} \Rightarrow |E| \le (n-1)u_1 = 1.01(n-1)u$

Επομένως το σφάλμα στον υπολογισμό του γινομένου είναι:

$\Sigma\varphi\acute{\alpha}\lambda\mu\alpha \equiv E' = x_1x_2\ldots x_nE,$

$\text{Απόλυτο σφάλμα} \equiv |E'| = |x_1||x_2|\ldots|x_n|\cdot|E| \le |x_1||x_2|\ldots|x_n|1.01(n-1)u$ και

$$\text{Απόλυτο Σχετικό Σφάλμα} \equiv \frac{|E'|}{|x_1||x_2|\cdot\ldots\cdot|x_n|} \le 1.01(n-1)u$$

Άρα στον πολλαπλασιασμό δεν παίζει ρόλο η σειρά των παραγόντων για το σφάλμα.

2. Υπολογισμός εσωτερικού γινομένου δύο n-διάστατων διανυσμάτων

Έστω δύο *n*-διάστατα διανύσματα *a* και *b*

$$\left.\begin{array}{l} a=(a_1,a_2,...,a_n) \\ b=(b_1,b_2,...,b_n) \end{array}\right\} \quad \Rightarrow \qquad ab \equiv < a,b >= a_1b_1+...+a_nb_n$$

Τότε είναι $fl(a_1b_1+...+a_nb_n)=a_1b_1(1+\varepsilon_1)+...+a_2b_2(1+\varepsilon_n)$ (1)

Ορίζουμε αναγωγικά τις ποσότητες ip_r (internal product)$_r$ και t_r ως εξής:

$t_r = fl(a_rb_r)$

$ip_1 = t_1$

$ip_r = fl(ip_{r-1}+t_r),\ \ r=2,3,...,n$

οπότε έχουμε:

$t_r = fl(a_rb_r)=a_rb_r(1+\xi_r),$ όπου $|\xi_r|\le u$ (ξ_r : το σφάλμα του r γινομένου)

$ip_1 = t_1$

$ip_r = fl(ip_{r-1}+t_r)=(ip_{r-1}+t_r)(1+\eta_r),$ όπου $|\eta_r|\le u$ (η_r : το σφάλμα του r αθροίσματος)

Με διαδοχικές αντικαταστάσεις στον ανωτέρω τύπο για $r=2,3,...,n$ προκύπτουν:

$\underline{r=1}$: $t_1 = fl(a_1b_1)=a_1b_1(1+\xi_1)$

$ip_1 = t_1 = a_1b_1(1+\xi_1)$

$\underline{r=2}$: $t_2 = fl(a_2b_2)=a_2b_2(1+\xi_2)$

$$ip_2 = fl(ip_1+t_2)=(ip_1+t_2)(1+n_2)=[a_1b_1(1+\xi_1)+a_2b_2(1+\xi_2)](1+n_2)+$$
$$= a_1b_1(1+\xi_1)(1+n_2)+a_2b_2(1+\xi_2)(1+n_2)$$

$\underline{r=n}$: $t_n = fl(a_nb_n)=a_nb_n(1+\xi_n)$

$$ip_n = fl(ip_{n-1}+t_n)=(ip_{n-1}+t_n)(1+n_n)=a_1b_1(1+\xi_1)(1+n_2)\dots(1+\eta_n)+a_2b_2(1+\xi_2)(1+n_2)\dots(1+\eta_n)$$
$$+\dots+a_nb_n(1+\xi_n)(1+\eta_n) \qquad (2)$$

Εξισώνοντας τις (1) και (2) προκύπτουν

$$1+\varepsilon_1=(1+\xi_1)(1+\eta_2)...(1+\eta_n) \qquad (3\alpha)$$

$$1+\varepsilon_r=(1+\xi_r)(1+\eta_r)...(1+\eta_n)^{(2)},\ \ r=2,3,...,n \qquad (3\beta)$$

Από την (3α): $(1-u)^n \le 1+\varepsilon_1 \le (1+u)^n$ $(3\alpha')$

Από την (3β): $(1-u)^{n-r+2} \le 1+\varepsilon_r \le (1+u)^{n-r+2},\ \ r=2,3,...,n$ $(3\beta')$

$\Rightarrow$ ενοποίηση των (3α) και (3β)(ενιαίο φραγμ

Από την $(3\alpha')$: $(1-u)^{n-1} \le (1-u)^n \le 1+\varepsilon_1 \le (1+u)^n \le (1+u)^{n+1}$ (διότι 1-u<1,1+u>1) $(3\alpha'')$

Τελικά από τις $(3\alpha'')$ και $(3\beta')$ προκύπτει η ενοποιημένη ανισότητα

$$(1-u)^{n-r+2} \le 1+\varepsilon_r \le (1+u)^{n-r+2},\quad r=1,2,3,...,n \qquad (4)$$

Σύμφωνα με την παρατήρηση του λήμματος 3 προκύπτει:

$$|\varepsilon_r| \le (n-r+2)1.01u \qquad (5)$$

Από την (1) έχουμε

$$fl(a_1b_1+...+a_nb_n)=a_1b_1(1+\varepsilon_1)+...+a_2b_2(1+\varepsilon_n)=a_1b_1+...+a_nb_n+\underbrace{a_1b_1\varepsilon_1+...+a_nb_n\varepsilon_n}_{\sigma\varphi\acute{\alpha}\lambda\mu\alpha\ E}$$

Επομένως το σφάλμα στον υπολογισμό του εσωτερικού γινομένου είναι:

$$\text{Απόλυτο Σφάλμα} \equiv |\,E\,| = |a_1b_1\varepsilon_1+...+a_nb_n\varepsilon_n| \le |a_1||b_1||\varepsilon_1|+...+|a_n||b_n||\varepsilon_n| \le$$

$$\overset{(5)}{\le} 1.01u\{(n+1)|a_1||b_1|+n|a_2||b_2|+...+2|a_n||b_n|\} \le$$

$$\le 1.01u(n+1)(|a_1||b_1|+|a_2||b_2|+...+|a_n||b_n|) = 1.01u(n+1)|a|^T|b|$$

(συμβολικά $|ab| = |a|^T|b|$)

$$\text{Απόλυτο σχετικό σφάλμα} \equiv \frac{|fl(a^Tb)-a^Tb|}{|a^Tb|} = \frac{|E|}{|a^Tb|} \le \frac{(n+1)1.01u|a|^T|b|}{|a^Tb|}$$

Αν $|a^Tb| << |a|^T|b|$ το απόλυτο σχετικό σφάλμα μπορεί να μην είναι "μικρό".

Ανάλυση Σφάλματος(Βιβλιογραφία)

Είναι αξιοσημείωτο να αναφερθεί ότι ο θεμελιωτής της θεωρίας σφάλματος είναι ο J. Wilkinson: με το βιβλίο του με τίτλο Rounding Error in Algebraic Processes.

Η ανάλυση σφάλματος διακρίνεται σε 1**) Forward**(εκ των προτέρων) και 2) **Backward**(εκ των υστέρων) ανάλυση σφάλματος.

Αν υποτεθεί ότι x είναι η είσοδος σε μιά συνάρτηση f. Έστω ŷ η προσέγγιση της y=f($x_{\text{είσοδος}}$) με ακρίβεια ε.

1)Forward ανάλυση: Εύρεση άνω φραγματος για το forward σφάλμα $|\Delta y| = |y - \hat{y}|$

2)Backward ανάλυση: Εύρεση άνω φραγματος για το backward σφάλμα Δx: f(x+Δx)=ŷ

Πρόβλημα καλώς ορισμένο(well posed)

1) Υπάρχει η λύση του

2) Είναι μοναδική

3) Εξαρτάται κατά συνεχή τρόπο από τα δεδομένα (ευστάθεια), δηλαδή μικρές μεταβολές στην είσοδο (δεδομέν
δεν παρουσιάζουν μεγάλες μεταβολές στην έξοδο(αποτελέσματα).

Πρόβλημα well-conditioned(καλής κατάστασης)

1) Καλώς ορισμένο

2) Έχει μικρό μεταδιδόμενο σφάλμα

Σχετικός αριθμός συνθήκης (δείκτης κατάστασης(condition number)):

$$cond_f(x) = \frac{\left\| \frac{f(\hat{x}) - f(x)}{\hat{x} - x} \right\|}{\frac{\|\hat{x} - x\|}{\|x\|}} = \frac{\frac{\|\Delta y\|}{\|y\|}}{\frac{\|\Delta x\|}{\|x\|}}$$

- Προβλήματα
  - Καλώς Ορισμένα
    - Well Conditioned ("μικρό" $\frac{\frac{\|\Delta y\|}{\|y\|}}{\frac{\|\Delta x\|}{\|x\|}}$)
    - Ill conditioned
  - Κακώς Ορισμένα
- Αλγόριθμοι
  - Backward ευσταθείς ("μικρό" $\frac{\|\Delta x\|}{\|x\|}$)
  - Ασταθείς

Ακριβής λύση αν : • Well Conditioned πρόβλημα •Backward ευσταθής αλγόριθμος