Нечуйвітер О.П., Мотря А.І., ДКІ-ПОМ 14 пр+ПОБ14 пр

ПРО ОЦІНКИ ПОХИБОК ЗАОКРУГЛЕННЯ ОСНОВНИХ АРИФМЕТИЧНИХ ОПЕРАЦІЙ

Нехай обчислення виконуються на ЕОМ у режимі з плаваючою комою (П.К.) із заокругленням результатів арифметичних операцій за класичним правилом до τ двійкових розрядів у нормалізованих мантис чисел. Похибки заокруглення в цих операціях мають вигляд

$$z = fl\begin{pmatrix} \pm \\ x \times y \\ : \end{pmatrix}(1+\varepsilon)\ , \qquad (1)$$

де ε – відносна похибка і $|\varepsilon| \le 2^{-\tau}$. При цьому x і y вважаються заданими точно в ЕОМ, звідси випливає, що мантиси x і y мають не більш, ніж τ розрядів. Якщо вхідне число x має нормалізовану мантису з числом цифр, більшим, ніж τ (і порядок числа x може бути записаний в ЕОМ), то

$$x = x_\tau(1+\varepsilon),\ |\varepsilon| \le 2^{-\tau-1}\ , \qquad (2)$$

де x_τ – машинне заокруглене представлення числа x. Усі результати, які мають місце в режимі П.К., є прямим наслідком співвідношень (1), (2), застосування яких приводить до оцінок вигляду $\left(1-2^{-\tau}\right)^r \le 1+\varepsilon \le \left(1+2^{-\tau}\right)^r$,які можна спростити, припустивши, що виконується умова $r \cdot 2^{-\tau} < 0{,}1$ (це цілком виправдано в практичних застосуваннях для будь-якого прийнятного τ). Тоді $\left(1+2^{-\tau}\right)^r < 1+1{,}06 \cdot r \cdot 2^{-\tau}$, $\left(1-2^{-\tau}\right)^r < 1-1{,}06 \cdot r \cdot 2^{-\tau}$, $1-1{,}06 \cdot r \cdot 2^{-\tau} < 1+\varepsilon < 1+1{,}06 \cdot r \cdot 2^{-\tau}$, звідки $|\varepsilon| < 1{,}06 \cdot r \cdot 2^{-\tau}$. Останнє співвідношення використовується у всіх наступних оцінках:

1) $fl(x_1 \cdot x_2 \cdot \ldots \cdot x_N) \equiv \prod_{i=1}^{N} x_i(1+E)$, де $|E| < (N-1) \cdot 1{,}06 \cdot 2^{-\tau}$;

2) $fl(x_1 + x_2 + \ldots + x_N) \equiv x_1(1+\varepsilon_1) + x_2(1+\varepsilon_2) + \ldots + x_N(1+\varepsilon_N)$, де $|\varepsilon_1| < (N-1) \cdot 1{,}06 \cdot 2^{-\tau}$, $|\varepsilon_r| < (N-r+1) \cdot 1{,}06 \cdot 2^{-\tau}$, $r = \overline{2, N}$. Тут передбачалося, що $S_2 = fl(x_1 + x_2)$, $S_r = fl(S_{r-1} + x_r)$, $r = \overline{3, N}$;

3) $fl(x_1 \times y_1 + x_2 \times y_2 + \ldots + x_N \times y_N) \equiv x_1 y_1(1+\varepsilon_1) + x_2 y_2(1+\varepsilon_2) + \ldots + x_N y_N(1+\varepsilon_N)$, де $|\varepsilon_1| < N \cdot 1{,}06 \cdot 2^{-\tau}$, $|\varepsilon_r| < (N-r+2) \cdot 1{,}06 \cdot 2^{-\tau}$, $r = \overline{2, N}$;

4) $fl\left(\frac{x_1 \cdot x_2 \cdot \ldots \cdot x_m}{y_1 \cdot y_2 \cdot \ldots \cdot y_n}\right) \equiv \frac{x_1 \cdot x_2 \cdot \ldots \cdot x_m}{y_1 \cdot y_2 \cdot \ldots \cdot y_n}(1+E)$, де $|E| < (m+n-1) \cdot 1{,}06 \cdot 2^{-\tau}$.

Для операцій додавання і множення, реалізованих на ЕОМ у режимі П.К., справедливі нерівності: $|fl(a \cdot x_1) - fl(a \cdot x_2)| \le 2|a| \cdot |x_1 - x_2|$, $|fl(a + x_1) - fl(a + x_2)| \le s \cdot |x_1 - x_2|$. Постійна s залежить від співвідношення порядків доданків і способу заокруглення, зафіксованого в ЕОМ. Можна підібрати способи запису й заокруглення такі, що $s = 2$.