Нечуйвітер О.П., Голіусов В. О. (ДІТ-П14 пр)

ОБЧИСЛЕННЯ ОЦІНОК ПОХИБОК ЗАОКРУГЛЕННЯ ДЛЯ ЗАДАЧ ОБЧИСЛЮВАЛЬНОЇ МАТЕМАТИКИ

На основі зазначених результатів можна одержувати мажорантні оцінки похибок заокруглення для багатьох обчислювальних алгоритмів розв'язку задач обчислювальної і прикладної математики. Розрізняють два режими роботи ЕОМ – з фіксованою комою (Ф.К.) і плаваючою комою (П.К.). При обчисленнях з Ф.К. кожне число x знаходиться в інтервалі $-1 \le x \le 1$, до якого вхідні числа приводяться шляхом масштабування. При обчисленнях з П.К. кожне число x представляється у вигляді $x = 2^b \times a$, де b – ціле додатнє або від'ємне число, яке називається порядком, і a (мантиса) – число, що задовольняє одну з нерівностей: $-1 \le a \le -1/2$ або $1/2 \le a \le 1$. Передбачається, що ЕОМ оперують з числами, які мають у p–ічному (для простоти обмежимося $p = 2$) представленні τ розрядів після коми у випадку Ф.К. і τ розрядів у мантисі – у випадку П.К.; такі числа будемо називати *стандартними.*

Нехай $fl(*)$ - результат обчислення на даній ЕОМ виразу в дужках, тобто рівність виду $z = fl\begin{pmatrix} \pm \\ x \times y \\ : \end{pmatrix}$ означає, що x, y і z – стандартні числа і що z отримано з x і y виконанням відповідної операції в режимі П.К. Нехай обчислення виконуються на ЕОМ у режимі П.К. з 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

$$\left(1+2^{-\tau}\right)^r < 1+1{,}06 \cdot r \cdot 2^{-\tau}, \qquad \left(1-2^{-\tau}\right)^r < 1-1{,}06 \cdot r \cdot 2^{-\tau},$$

$$1-1{,}06 \cdot r \cdot 2^{-\tau} < 1+\varepsilon < 1+1{,}06 \cdot r \cdot 2^{-\tau},$$

звідки $|\varepsilon| < 1{,}06 \cdot r \cdot 2^{-\tau}$.