Solution Notes

P346
MROD

Numerical Analysis I – Question A 2004

Context: IEEE arithmetic; absolute and relative errors; machine epsilon.

(a) A binary floating point number can be expressed as

$$\pm\, d_0 . d_1 d_2 \ldots d_{p-1} \times 2^e$$

where e is the exponent, $d_0 . d_1 d_2 \ldots d_{p-1}$ is the binary significand, and p is the precision.

The sign bit is the sign of the significand and determines whether the number is positive (0) or negative (1).

If $d_0 = 1$ then the number is normalised. In IEEE Single Precision normalised numbers have an exponent in the range $e_{min} \leq e \leq e_{max}$.

If $e = e_{min}$ and $d_0 = 0$ then the number is said to be denormal; by convention this is denoted by the exponent $e_{min} - 1$ (rather than e_{min}). [6 marks]

(b) For normalised or denormal numbers d_0 can be predicted from the exponent so is not stored; this is called the hidden bit. The hidden bit is 1 for normalised numbers, 0 for denormal numbers.

The exponent is stored as $e + e_{max}$ so that the exponent does not need a sign bit.

sign	exponent	significand
1	8	23

bits

[4 marks]

over.

(c) Let x^* denote the floating point representation of x.

Absolute error ϵ_x is defined by

$$x^* = x + \epsilon_x .$$

Relative error δ_x is defined by

$$x^* = x(1+\delta_x) = x + x\delta_x$$

$$\therefore \epsilon_x = x\delta_x .$$

Machine epsilon ϵ_m is the smallest $\epsilon_m > 0$ for which

$$(1+\epsilon_m)^* > 1 .$$

[3 marks]

(d) Add relative errors on multiplying

$$\delta_{xy} = |\delta_x| + |\delta_y| = 2\epsilon_m .$$

$$\epsilon_{xy} = |xy|\,\delta_{xy} = 2|xy|\epsilon_m .$$

Add absolute errors when subtracting

$$\epsilon_w = |\epsilon_z| + |\epsilon_{xy}| = |z|\epsilon_m + 2|xy|\epsilon_m$$

$$= (|z| + 2|xy|)\epsilon_m$$

If $w \neq 0$

$$\delta_w = \frac{\epsilon_w}{|w|} = \frac{|z| + 2|xy|}{|z - xy|}\epsilon_m .$$

[4 marks]

(e)

$$\begin{array}{r} 2.018 \\ 2.008\ \times \\ \hline 4036 \\ 16 \\ \hline 4.052 \end{array}$$

$$w^* = 4.058 - 4.052 = 0.006$$

$$\delta_w \simeq \frac{4 + 2\times 4}{0.006}\epsilon_m = 2\times 10^3 \times 0.5\times 10^{-3}$$

$$\therefore \delta_w \simeq 1$$

$\Rightarrow$ No significant decimal digits can be relied on.

[3 marks]