Math 781 Hw2 Solution

1. Let $x = (1.11\cdots 111000\cdots)_2 \times 2^{16}$, in which the fraction part has 26 1's followed by 0's. For the Marc-32, determine x_-, x_+, $fl(x)$, $x - x_-$, $x_+ - x$, $x_+ - x_-$, and $\frac{x - fl(x)}{x}$.
   Solution. :

$$\begin{aligned}
x_- &= (1.11\cdots1\cdots)_2 \times 2^{16}, \quad 23\ '1' \\
x_+ &= (1.00\cdots0\cdots)_2 \times 2^{17}, \quad 23\ '0' \\
fl(x) &= x_+ = (1.00\cdots0\cdots)_2 \times 2^{17}, \quad 23\ '0' \\
x - x_- &= (1.110\cdots0)_2 \times 2^{-8} \\
x_+ - x &= (1.0\cdots0)_2 \times 2^{-10} \\
x_+ - x_- &= (1.0\cdots0)_2 \times 2^{17} - (1.11\cdots1)_2 \times 2^{16} \\
&= (0.0\cdots01)_2 \times 2^{16}, \quad 23\ '0' \text{ before } 1 \\
&= 2^{-23} \times 2^{16} = 2^{-7}. \\
\left|\frac{x - fl(x)}{x}\right| &= \frac{(1.0\cdots0)_2 \times 2^{-10}}{(1.11\cdots111000\cdots)_2 \times 2^{16}} \\
&= \frac{(1.0\cdots0)_2 \times 2^{-10}}{(2 - 2^{-26}) \times 2^{16}} = \frac{1}{2^{27} - 1} \approx 7.46 \times 10^{-9}.
\end{aligned}$$

2. Which if these is not necessarily true on the Marc-32? (Here x, y, and z are machine numbers and $|\delta| \le 2^{-24}$.

   (a) $fl(xy) = xy(1 + \delta)$
   (b) $fl(x + y) = (x + y)(1 + \delta)$
   (c) $fl(xy) = \frac{xy}{1+\delta}$
   (d) $|fl(xy) - xy)| \le |xy|2^{-24}$
   (e) $fl(x + y + z) = (x + y + z)(1 + \delta)$

   Solution. : (c) and (e).

3. Are these machine numbers in the Marc-32?

   (a) 10^{40}
   (b) $2^{-1} + 2^{-26}$
   (c) $\frac{1}{5}$
   (d) $\frac{1}{3}$
   (e) $\frac{1}{256}$

Solution. : (*e*).

4. Let $x = 2^{16} + 2^{-8} + 2^{-9} + 2^{-10}$. What is $|x - fl(x)|$ in the Marc-32?

   Solution. : Since $x = 2^{16}\left(1 + 2^{-24} + 2^{-25} + 2^{-26}\right)$, $fl(x) = 2^{16}$. $|x - fl(x)| = 2^{-8} + 2^{-9} + 2^{-10} = 7 \times 2^{-10}$.

5. In a typical floating point number system a non-zero number x is stored in the form

$$x = \sigma \cdot (.a_1 a_2 a_3 \cdots a_t)_\beta \cdot \beta^e,$$

   where $\sigma = +1$ or -1, $a_1 \neq 0$, $0 \leq a_i \leq \beta - 1$, $t = 53$, $\beta = 2$, and $-1023 \leq e \leq 1024$.

   (a) Find the greatst and smallest positive numbers and the unit roundoff.

   (b) Which of the following are the numbers in this typical floating point number system?

$$10, \quad 1 + 2^{-53}, \quad 1 - 2^{-53}, \quad 2^{1024}.$$

   Solution. :

   (a) The greatest positive number is

$$(0.1 \cdots 1)_2 \times 2^{1024} \quad = \quad (1 - 2^{-53}) \times 2^{1024}.$$

   The smallest positive number is

$$(0.10 \cdots 0)_2 \times 2^{-1023} \quad = \quad 2^{-1024}.$$

   The machine epsilon

$$\begin{aligned} \epsilon &= (0.10 \cdots 01)_2 2^1 - (0.10 \cdots 0)_2 2^1 \quad (52\ '0'\ \text{for}\ '1') \\ &= 2^{-53} 2^1 = 2^{-52}. \end{aligned}$$

   The unit roundoff $\delta = \frac{1}{2}\epsilon = 2^{-53}$.

   (b)

$$\begin{aligned} 10 &= 2^3 + 2^1 = (0.1010 \cdots 0)_2 2^4 \\ 1 - 2^{-53} &= (0.10 \cdots 0)_2 2^1 - (0.0 \cdots 01)_2 2^1 \\ &= (0.01 \cdots 1)_2 2^1 = (0.1 \cdots 1)_2. \end{aligned}$$

   $1 + 2^{-53}$ is not a machine number since $2^{-53} < \epsilon$.
   $2^{1024} > (1 - 2^{-53}) 2^{1024}$ is not a machine number.