20210906 455 HW3 Shi Qin

2. Convert the following base 10 numbers to binary and express each as a floating point number fl(x) by using the Rounding to Nearest Rule: (a) 9.5 (b) 9.6 (c) 100.2 (d) 44/7

a. $fl(9.5)$

$9.5 = 1001_2 + 0.5_{10}$
$= 1001.1$

$0.5 \mid \times 2$
1
0.1

$x = (-1)^s \times 2^{(-1023} \times (1.f)_2$

$s = 0$

$9.5_{10} = 1001.1_2$
$= 1.0011 \times 2^3$

$x = (-1)^0 \times 2^{1026-1023} \times (1.0011)$

b. $fl(9.6)$

$9.6 = 1001.\overline{1001}$

$0.6 \mid \times 2$
1.2
0.4
0.8
1.6
1.2
0

$fl(9.6) = (-1)^0 \times 1.001\,1001 \cdots 0011 \times 2^{1026-1023}$

c. $fl(100.2)$

$100.2 = 1100100.00110011\cdots$

$fl(100.2) = (-1)^0 \times 1.100100 \cdots 1101 \times 2^{1029-1023}$

d $fl(44/7)$

$$6\frac{2}{7} = 110.\overline{001}$$

$$\begin{array}{l|l}
0 \ \frac{2}{7} & \times 2 \\
0 \ \frac{4}{7} & \\
1 \ \frac{1}{7} & \\
0 \ \frac{2}{7} & \\
0 & \\
1 &
\end{array}$$

$$fl\left(\frac{44}{7}\right) = (-1)^0 \times 1.1001\cdots001 \times 2^{1025-1023}$$

7. Write each of the given numbers in MATLAB's `format hex`. Show your work. Then check your answers with MATLAB. (a) 8 [redacted] (c) 1/8 [redacted] (e) fl (2/3) [redacted] (g) fl (−0.1) [redacted]

a. 8

$8_{10} = 1000_2$

$$fl(8) = (-1)^0 \times 2^{1026-1023} \times (1.00\cdots0)$$

s	11	52
0	10000000010	00000 ⋯ 0

$\underbrace{0100}_{4}\ \underbrace{0000}_{0}\ \underbrace{0010}_{2}\ \underbrace{0000}_{0}\ \underbrace{0000}_{0} \cdots \underbrace{0000}_{0}$

4020000000000000

13个0

c. $\frac{1}{8}$

0	$\frac{1}{8}$	×2
0	$\frac{1}{4}$	
0	$\frac{1}{2}$	
1		

0.001_2

$$fl\left(\frac{1}{8}\right) = (-1)^0 \times 2^{1020-1023} \times (1.00\cdots0)$$

0	01111111100	000 ⋯ 000

$\underbrace{0011}_{3}\ \underbrace{1111}_{F}\ \underbrace{1100}_{C}\ \underbrace{0000}_{0}\ 0 \cdots \underbrace{0000}_{0}$

3FC0000000000000

13个0

e). $fl(\frac{2}{3})$

$$\frac{2}{3}_{10} = 0.\overline{10}_2$$

0	$\frac{2}{3}$	×2
1	$\frac{1}{3}$	
0	$\frac{2}{3}$	
1	$\frac{1}{3}$	
0	$\frac{2}{3}$	

$$fl(\tfrac{2}{3}) = (-1)^0 \times 2^{1022-1023} \times (1.01010\cdots10)$$

0	01111111110	010 ..... 1010 10101

$\underbrace{0011}_{3}\ \underbrace{1111}_{F}\ \underbrace{1110}_{E}\ \underbrace{0101}_{5} \cdots \underbrace{0101}_{5} \cdots \underbrace{0101}_{5}$

3FE 5555555555555

13 个 5

g.) $fl(-0.1)$

$$(-1)^1 \times 2^{1019-1023} \times (1.10011)$$

0	.1	×2
0	.2	
0	.4	
0	.8	
1	.6	
1	.2	
0	.4	
0	.8	

$0.0\overline{0011}$

1	01111111011	10011001 ...... 100 1010

$\underbrace{1011}_{B}\ \underbrace{1111}_{F}\ \underbrace{1011}_{B}$

BFB999999999999a

12 个 9 one a

12. Find the IEEE double precision representation $\mathrm{fl}(x)$, and find the exact difference $\mathrm{fl}(x) - x$ for the given real numbers. Check that the relative rounding error is no more than $\epsilon_{\mathrm{mach}}/2$.
(a) $x = 1/3$ (b) $x = 3.3$ (c) $x = 9/7$

a. $x = \frac{1}{3}$

$$fl\left(\tfrac{1}{3}\right) = (-1)^0 \times 2^{1021-1023} \times (1.\overline{01}0101\cdots)$$

$$\begin{array}{c|l} 0 & \frac{1}{3} \quad \times 2 \\ 0 & \frac{2}{3} \\ 1 & \frac{1}{3} \\ 0 & \frac{2}{3} \end{array}$$

$0.\overline{01}0101$

$$fl\left(\tfrac{1}{3}\right) - \tfrac{1}{3} = -0.\overbrace{00\cdots-00}^{52\text{bit}}\overline{01}$$

$$= -0.\overline{01} \times 2^{-54}$$

$$\frac{|-0.\overline{01} \times 2^{-54}|}{|\frac{1}{3}|} = 2^{-54} < \frac{\epsilon_{mach}}{2}$$

b $x = 3.3$

$$fl(3.3) = (-1)^0 \times 2^{1024-1023} \times (1.10\overline{1001}\ 1001 \cdots 1001)$$

$$fl(3.3) - 3.3 = -0.\overbrace{0\cdots\cdots0}^{52\text{ bit}}\overline{0110}$$

$$= -0.\overline{0110} \times 2^{-51}$$

$$\frac{|-0.\overline{0110} \times 2^{-51}|}{|3.3|} < \frac{\epsilon_{mach}}{2}$$

c $x = \frac{9}{7}$

$$\begin{array}{c|l} 1 & \frac{2}{7} \quad \times 2 \\ 0 & \frac{4}{7} \\ 1 & \frac{1}{7} \\ 0 & \frac{2}{7} \\ 0 & \frac{4}{7} \\ 1 & \frac{1}{7} \end{array}$$

$1.0\overline{100}$

$$fl\left(\tfrac{9}{7}\right) = (-1)^0 \times 2^{1023-1023} \times (1.0\overline{100}\ 100 \cdots 100)$$

$$fl\left(\tfrac{9}{7}\right) - \tfrac{9}{7} = -0.0\overline{100} \times 2^{-52} < \frac{\epsilon_{mech}}{2}$$

16. Find the IEEE double precision representation $\mathrm{fl}(x)$, and find the exact difference $\mathrm{fl}(x) - x$ for the given real numbers. Check that the relative rounding error is no more than $\epsilon_{\text{mach}}/2$.
(a) $x = 2.75$ (b) $x = 2.7$ (c) $x = 10/3$

a. $X = 2.75$ $\quad 1.011 \times 2^{-1}$

10.11_2

$\text{error} = 0 < \frac{\epsilon_{mach}}{2}$

b. $X = 2.7$

$10.1\overline{0110}_2$ $\quad fl(2.7) = (-1)^0 \times 2^{1024-1023} \times (1.010\overline{0110} \cdots 1001)$

$fl(2.7) - 2.7 = 0.\underbrace{0 \cdots 0}_{52bit}\overline{1001}$

$= -0.\overline{1001} \times 2^{-5} < \frac{\epsilon_{mach}}{2}$

c. $X = \frac{10}{3}$

$10.\overline{01}_2$ $\quad fl(\frac{10}{3}) = (-1)^0 \times 2^{1024-1023} \times (1.0\overline{01} \cdots 010)$

$3\frac{1}{3}$ | $\times 2$
$0\frac{2}{3}$
$1\frac{1}{3}$

$fl(\frac{10}{3}) - \frac{10}{3} = -0.\underbrace{0 \cdots 0}_{52bit}\overline{10}10 \cdots$

$= 0.\overline{10} \times 2^{-52} < \frac{\epsilon_{mach}}{2}$