230201003 Tunahan YADİGARBİGÜN
230201057 Furkan Emre YILMAZ

Q1 -

$\frac{7}{8} = \frac{1}{8} + \frac{1}{4} + \frac{1}{2} = 2^{-3} + 2^{-2} + 2^{-1} \longrightarrow (0.111)_2$

$10.5 = 10 + \frac{1}{2}$ $\quad (10)_{10} = (1010)_2$

$(\frac{1}{2})_{10} = (0.1)_2$ $\quad (10.5)_{10} = (1010.1)_2$

$12.8 = 12 + 0.8$ $\quad (12)_{10} = (1100)_2$

$0.8 \times 2 = 0.6 + 1$

$0.6 \times 2 = 0.2 + 1$ $\quad (0.8)_{10} = (\overline{1100})_2$

$0.2 \times 2 = 0.4 + 0$

$0.4 \times 2 = 0.8 + 0$ $\quad (12.8)_{10} = (1100.\overline{1100})_2$

$\frac{2}{3} \times 2 = \frac{1}{3} + 1$ $\quad (\frac{2}{3})_{10} = (0.\overline{10})_2$

$\frac{1}{3} \times 2 = \frac{2}{3} + 0$

$3.2 = 3 + 0.2$ $\quad (3)_{10} = (0011)_2$

$0.2 \times 2 = 0.4 + 0$ $\quad (0.2)_{10} = (\overline{0011})_2$

$0.4 \times 2 = 0.8 + 0$

$0.8 \times 2 = 0.6 + 1$ $\quad (3.2)_{10} = (0011.\overline{0011})_2$

$0.6 \times 2 = 0.2 + 1$

Find the IEEE double precision representation $fl(x)$, and find the exact difference $fl(x) - x$ for the numbers $\frac{1}{3}$, 3.3, and $\frac{9}{7}$

a) First convert $\frac{1}{3}$ to binary

$$\frac{1}{3} \times 2 = \frac{2}{3} + 0$$
$$\frac{2}{3} \times 2 = \frac{1}{3} + 1$$
$$\frac{1}{3} \times 2 = \frac{2}{3} + 0$$
$$\frac{2}{3} \times 2 = \frac{1}{3} + 1$$

$\left(\frac{1}{3}\right)_{10} = (0.\overline{01})_2$ we can check it

$$2^2 x = (01.\overline{01})_2$$
$$- \quad 1.x = (0.\overline{01})_2$$
$$3x = (01)_2$$
$$x = \frac{1}{3}$$

Second step using IEEE double precision $fl(x)$,
we first convert number to normalized format

$$(0.\overline{01})_2 = (01.\overline{01}) \times 2^{-2}$$
$$= (1.\overline{01}) \times 2^{-2}$$

we know that in double precision we have 52 bit for mantissa so, we can write number such as

$$\underbrace{1.0101010101 \text{------} 01}_{52 \text{ bit}},\underbrace{0101\text{---}}_{\hookrightarrow \text{infinite tail}} \times 2^{-2}$$

to represent the number we discard infinite tail

$$2^{-2} \times 2^{-52} (0.\overline{01})_2 \rightarrow \text{discarded}$$

$$fl(1/3) = \frac{1}{3} - \frac{1}{3} \times 2^{-54} \qquad (0,\overline{01})_2 = 1/3$$

$$fl(1/3) - \frac{1}{3} = -\frac{1}{3} \times 2^{-54}$$

Using the same steps as before

b) first convert 3.3 to binary

$$3.3 = 3.0 + 0.3$$

$3/2 = 1.2 + 1$

$1/2 = 0.2 + 1$

$0.3 \times 2 = 0.6 + 0$
$0.6 \times 2 = 0.2 + 1$
$0.2 \times 2 = 0.4 + 0$
$0.4 \times 2 = 0.8 + 0$
$0.8 \times 2 = 0.6 + 1$
$0.6 \times 2 = 0.2 + 1$
$0.2 \times 2 = 0.4 + 0$
$0.4 \times 2 = 0.8 + 0$
$0.8 \times 2 = 0.6 + 1$

$(3)_{10} = (11)_2$

$(0.0\overline{1001})_2 = (0.3)_{10}$

$(3.3)_{10} = (11.0\overline{1001})_2$

Normalized form can be written as

$$(1.10\overline{1001}) \times 2^1$$

can be represented in computer as

$$\underbrace{1.1010011001\ldots\ldots 100110}_{52\text{ bit}}\ \underbrace{\overline{0110011001}\ldots}_{\text{infinite tail}} \times 2^1$$

to represent the number we discard infinite tail

$$2^1 \times 2^{-52} \times (0.\overline{0110})_2$$

$fl(3.3) = 3.3 - 2^{-51} \times (0.4)$

$fl(3.3) - 3.3 = -2^{-51} \times (0.4)$

$(0.\overline{0110})_2 \Rightarrow 2^4 x = (110.\overline{0110})_2$

$- \quad x = (0.\overline{0110})_2$

$15x = (110)_2$

$x = \frac{6}{15} = \frac{2}{5} = 0.4$

c) first convert $\frac{9}{7}$ to binary

$\frac{9}{7} = 1 + \frac{2}{7}$

$\left(\frac{9}{7}\right)_{10} = (1.0\overline{100})_2$

$\frac{2}{7} \times 2 = \frac{4}{7} + 0$

$\frac{4}{7} \times 2 = \frac{1}{7} + 1$

$\frac{1}{7} \times 2 = \frac{2}{7} + 0$

$\frac{2}{7} \times 2 = \frac{4}{7} + 0$

$\frac{4}{7} \times 2 = \frac{1}{7} + 1$

$\vdots$

$\left(\frac{2}{7}\right)_{10} = (0.0\overline{100})$

This number already in normalized form so our multiplier 2^0

$(1.0\overline{100})_2 \times 2^0$

We can represent in $fl(x)$ as

$\underbrace{1.0100100 \ldots 100}_{52 \text{ bit}} \; \underbrace{\overline{100}100100 \ldots}_{\text{infinite tail}} \times 2^0$

we need to round up

if we round up our number can be represented as

$\underbrace{1.0100100 \ldots 101}_{52 \text{ bit}}$

At the end lets consider what we added what we substracted

we added

$2^0 . 2^{-52}$

we substracted or discarded

$2^0 . 2^{-52} \underbrace{(0.\overline{100})_2}$ $\longrightarrow (0.\overline{100})_2 = \left(\frac{4}{7}\right)_{10}$

Therefore result is

$fl\left(\frac{9}{7}\right) - \frac{9}{7} + 2^{-52} \times 2^0 - \frac{4}{7} \times 2^{-52} \times 2^0$

$fl\left(\frac{9}{7}\right) - \frac{9}{7} = 2^{-52} - \frac{4}{7} \times 2^{-52}$

Do the following computation by hand in IEEE double precision computer arithmetic using Round the nearest rule: $4.3 - 3.3$

first convert the numbers binary

$4/2 = 2.2+0$
$2/2 = 1.2+0$
$1/2 = 0.2+1$

$(100)_2 = 4$
$(0.0\overline{1001})_2 = (0.3)_{10}$

$(100.0\overline{1001})_2$

we already know binary equivalent of 3.3 is

$$(3.3)_{10} = (11.0\overline{1001})_2$$

Second represent numbers in $fl(x)$

$$(100.01001)_2 = (1.000\overline{1001})_2 \times 2^2$$

$$\underbrace{1.000100110011001\ldots\ldots 1}_{52 \text{ bit}}\ \underbrace{\overline{0011001100}}_{\text{infinite tail}} \times 2^2$$

49

to represent the number we discarded infinite tail

$$= 2^2 \cdot 2^{-52} (0.\overline{0011})_2$$

$$= 2^{50} \times (0.2)$$

$(0.\overline{0011})_2 \Rightarrow 2^4 x = (11.\overline{0011})_2$
$- \quad x = (0.\overline{0011})_2$
$15x = (11)_2$
$x = \frac{3}{15} = \frac{1}{5} = 0.2$

this number represent our error. from question two we know the error of 3.3 is $2^{-51} \times 0.4$.

If we need to speak clearly

$$fl(4.3) + 2^{-50} \times (0.2) = 4.3$$

$$fl(3.3) + 2^{-51} \times (0.4) = 3.3$$

If we subotract this numbers because of error we made for each floating point number is same we will find the exactly 1

$$4.3 - 3.3 = 1$$
$$fl(4.3) - fl(3.3) = 1$$

$\Rightarrow =$

Q4 - Note that $f(x) = x^3 - 9$ $\quad f(2) = -1$ $\quad f(3) = 18$

i	a_i	$f(a_i)$	c_i	$f(c_i)$	b_i	$f(b_i)$
0	2	−	2.5	+	3	+
1	2	−	2.25	+	2.5	+
2	2	−	2.125	+	2.25	+
3	2	−	2.0625	−	2.125	+
4	2.0625	−	2.09375	+	2.125	+
5	2.0625	−	2.078125	−	2.09375	+
⋮	⋮	⋮	2.0859375	⋮	⋮	⋮
			2.08203125			
			2.080078125			
			2.0810546875			
			2.08056640625			
			2.080322265625			
			2.0802001953125			
			2.08013916015625			
			2.080108642578125			
			2.0800933837890625			
			2.0800857543945312			
			2.0800819396972656			
19			2.0800838470458984			
20			2.0800828933715820			
			$\boxed{2.0800833702087402}$			

Since the rule is obvious, no need to indicate all the numbers here

A solution is correct within p decimal places if the error is less than 0.5×10^{-p}

$\frac{1}{2^{n+1}} < 0.5 \times 10^{-6}$ $\qquad n > \frac{6}{\log 2} \cong \frac{6}{0.301} = 19.9$ Therefore, we applied the method until $i = 20$.

Q5 - Note that $f(x) = \cos x - \sin x$ $\quad f(\pi) = -1$ $\quad f(0) = +1$

i	a_i	$f(a_i)$	c_i	$f(c_i)$	b_i	$f(b_i)$
0	π	−	1.570796	−	0	+
1	1.570796	−	0.785398	+	0	+
2	1.570796	−	1.178097	−	0.785398	+
3	1.178097	−	0.981747	−	0.785398	+
4	0.981747	−	0.883572	−	0.785398	+
			0.834485			
			0.809941			
			0.797670			
			0.791534			
			0.788466			
			0.786932			
			0.786165			
			0.785781			
			0.785589			
			0.785494			
			0.785446			
			0.785422			
			0.785410			
			0.785404			
19			0.785401			
20			$\boxed{0.785399}$			

Find each fixed point of $g(x) = x^2 + \frac{1}{2}x - \frac{1}{2}$ and decide whether fixed Point Iteration is locally convergent to it.

ff The real number r is a fixed point of the function g if $g(r)=r$

ff Fixed-Point Iteration is locally convergent if $|g'(r)|<1$

Based on this defimtion lets first find the fixed point of the function

$$x = x^2 + \frac{1}{2}x - \frac{1}{2}$$

This means we are looking for solutions to

$$y = x$$
$$y = x^2 + \frac{1}{2}x - \frac{1}{2}$$

Drawing these functions we can learn whether we have such point or not easily. We will find two intersection point -0.5 and 1 these are our fixed points for this functions.

Lets try to find these points are locally convergent or not

$$g'(x) = 2x + \frac{1}{2}$$

$$g'(-0.5) = -1 + \frac{1}{2}$$
$$= -0.5$$

so $|g'(-0.5)| < 1$ so -0.5 point is locally convergent

$$g'(1) = 2 + \frac{1}{2}$$
$$= \frac{3}{2}$$

so $|g'(1)| < 1$ is not satisfied. It means, it is not locally convergent for 1.

When we apply fixed-point iteration method for different initial point we can see function converges to -0.5 sufficiently near this point. for example;

$x_0 = 0.9$

0.9
0.76
0.4576
-0.06180224....
-0.052708160...
⋮
-0.5

Q8 - Forward error : $|r - x_a| \longrightarrow |0.75 - 0.74| = 0.01$

Backward error : $|f(x_a)| \rightarrow |f(0.74)| = 0.0016$

Q9 -

x_0 = initial guess

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \quad \text{for} \quad i = 0, 1$$

$$f(x) = x^3 + x^2 - 1 \qquad f'(x) = 3x^2 + 2x$$

$$x_0 = 1$$

for $i = 0$,

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 1 - \frac{1}{5} = 0.8$$

for $i = 1$,

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = 0.8 - \frac{0{,}152}{3{,}52} \approx 0{,}7568$$