Name:	Aneesh Panchal	**Homework No:**	Assignment 2
SR No:	06-18-01-10-12-24-1-25223	**Course Code:**	DS284
Email ID:	aneeshp@iisc.ac.in	**Course Name:**	Numerical Linear Algebra
Date:	September 6, 2024	**Term:**	AUG 2024

Solution 1

Solution 1 (a)

Assumption: Considering exponent takes only 3 values which are $\{-1, 0, 1\}$ and total 5 bits are given out of which 2 which makes up for $\{00, 01, 10\}$ for $exponent$ and 3 for f.
Hence, total number of floating point numbers that can be described using given Toy System are,
$2^3 \times 3 = 8 \times 3 = \mathbf{24}$ floating point numbers.

Solution 1 (b)

Assumption to be consider is that, exponent takes only 3 values i.e. $\{-1, 0, 1\}$. Let the representation be follow this pattern,

$$(1.f)_2 \times 2^{exponent-1} \quad \equiv \quad \underbrace{00}_{exponent}\ \underbrace{000}_{f}$$

According to this, Table 1 represents all the floating points numbers that can be represented by the Toy System described.

Order	Toy System Representation (5 bits)	Binary Representation	Decimal Representation
01.	00000	0.1000	0.5000
02.	00001	0.1001	0.5625
03.	00010	0.1010	0.6250
04.	00011	0.1011	0.6875
05.	00100	0.1100	0.7500
06.	00101	0.1101	0.8125
07.	00110	0.1110	0.8750
08.	00111	0.1111	0.9375
09.	01000	1.0000	1.0000
10.	01001	1.0010	1.1250
11.	01010	1.0100	1.2500
12.	01011	1.0110	1.3750
13.	01100	1.1000	1.5000
14.	01101	1.1010	1.6250
15.	01110	1.1100	1.7500
16.	01111	1.1110	1.8750
17.	10000	10.000	2.0000
18.	10001	10.010	2.2500
19.	10010	10.100	2.5000
20.	10011	10.110	2.7500
21.	10100	11.000	3.0000
22.	10101	11.010	3.2500
23.	10110	11.100	3.5000
24.	10111	11.110	3.7500

Table 1: Floating Point Numbers represented by the Toy System described in increasing order.

Solution 1 (c)

Minimum Floating Point Number represented by Toy System is given by 00000 which is equivalent to,

$$(1.000)_2 \times 2^{(00)_2-1} = (1.000)_2 \times 2^{-1} = (0.1000)_2 = \mathbf{0.5}$$

Maximum Floating Point Number represented by Toy System is given by 10111 which is equivalent to,

$$(1.111)_2 \times 2^{(10)_2-1} = (1.111)_2 \times 2^{1} = (11.11)_2 = \mathbf{3.75}$$

Name: Aneesh Panchal
SR No: 06-18-01-10-12-24-1-25223
Email ID: aneeshp@iisc.ac.in
Date: September 6, 2024

Homework No: Assignment 2
Course Code: DS284
Course Name: Numerical Linear Algebra
Term: AUG 2024

Solution 1 (d)

Absolute gaps between any 2 consecutive numbers in $\mathbb{F}$ in $[x, y)$ are as follows,

1. For $x = 2^{-1} = 0.5$ and $y = 2^0 = 1$, absolute gap is 0.0625.
2. For $x = 2^0 = 1$ and $y = 2^1 = 2$, absolute gap is 0.125.
3. For $x = 2^1 = 2$ and $y = 2^2 = 4$, absolute gap is 0.25.

Hence, the **Absolute Gap** between any 2 consecutive numbers in $\left[2^j, 2^{j+1}\right)$ is $\mathbf{2^{j-3}}$.
That is, absolute gap between any 2 consecutive numbers in every set $\left[2^j, 2^{j+1}\right)$ is constant but changes for different sets.
And the gap in **Relative Sense** between any 2 consecutive numbers is constant for all values and is equal to $\frac{2^{j-3}}{2^j} = 2^{-3} = \mathbf{0.125}$.

Solution 1 (e)

Let us assume range $\left[2^j, 2^{j+1}\right)$. As we know, maximum absolute gap between any 2 consecutive numbers in $\mathbb{F}$ is 2^{j-3}.
Then for maximum absolute gap between $x \in \mathbb{R}$ and $x' \in \mathbb{F}$, assume $x' = x_0$ and $x = x_0 + 2^{j-4}$, then we have,

$$\frac{|x - x'|}{|x|} = \frac{x_0 + 2^{j-4} - x_0}{x_0 + 2^{j-4}} = \frac{2^{j-4}}{x_0 + 2^{j-4}}, \qquad \text{where, } x_0 = 2^j + k(2^{j-3}), \quad k \in \{1, 2, \ldots, 8\}$$

$$\equiv \frac{|x - x'|}{|x|} = \frac{2^{j-4}}{x_0 + 2^{j-4}} \leq \frac{2^{-4}}{1 + k(2^{-3}) + 2^{-4}} \leq 2^{-4} = 0.0625 = \epsilon_{machine} \text{ (say)}$$

Hence, machine epsilon for the given toy floating point system can be $\mathbf{0.0625}$.