Indian Institute of Science, Bangalore
Department of Computational and Data Sciences (CDS)

DS284: Numerical Linear Algebra

Assignment 2 [Posted Sept 2, 2024]

Bhikshapathi kaniki
SR No: 06-18-01-10-51-24-1-24489

Problem - 1

How many numbers can this toy system describe?

Ans:- This toy system can describe 24 distinct values.

f = 3bits and e = 2 bits

no of combinations = 2^3 = 8 values

with exponent = 3 combination

So, total distincts values = 8 x 3 = 24

Problem - 2

Create a table with 3 columns. First column should contain normalized binary scientific notation of the form1.fq2 ^ 2 pexponent´1qof all the above numbers.

Index	Normalized binary Scientific notation	Binary representation	Integer Representation
0	$(1.000)_2 \times 2^{0-1}$	0.1000	0.5
1	$(1.001)_2 \times 2^{0-1}$	0.1001	0.5625
2	$(1.010)_2 \times 2^{0-1}$	0.1010	0.625
3	$(1.011)_2 \times 2^{0-1}$	0.1011	0.6875
4	$(1.100)_2 \times 2^{0-1}$	0.1100	0.75
5	$(1.101)_2 \times 2^{0-1}$	0.1101	0.8125
6	$(1.110)_2 \times 2^{0-1}$	0.1110	0.875
7	$(1.111)_2 \times 2^{0-1}$	0.1111	0.9375
8	$(1.000)_2 \times 2^{1-1}$	1.000	1.0
9	$(1.001)_2 \times 2^{1-1}$	1.001	1.125
10	$(1.010)_2 \times 2^{1-1}$	1.010	1.25
11	$(1.011)_2 \times 2^{1-1}$	1.011	1.375
12	$(1.100)_2 \times 2^{1-1}$	1.100	1.5
13	$(1.101)_2 \times 2^{1-1}$	1.101	1.625
14	$(1.110)_2 \times 2^{1-1}$	1.110	1.75
15	$(1.111)_2 \times 2^{1-1}$	1.111	1.875
16	$(1.000)_2 \times 2^{2-1}$	10.00	2.0
17	$(1.001)_2 \times 2^{2-1}$	10.01	2.25
18	$(1.010)_2 \times 2^{2-1}$	10.10	2.5
19	$(1.011)_2 \times 2^{2-1}$	10.11	2.75
20	$(1.100)_2 \times 2^{2-1}$	11.00	3.0
21	$(1.101)_2 \times 2^{2-1}$	11.01	3.25
22	$(1.110)_2 \times 2^{2-1}$	11.10	3.5
23	$(1.111)_2 \times 2^{2-1}$	11.11	3.75

Problem - 3

From the table above, what is the minimum real number and maximum real number you can represent using our toy floating point number system

- Maximum no is 3.75
- Minimum no is 0.5

Problem - 4

What can you say about absolute gaps between the numbers? Are they constant or do they change with the magnitude of the number you are representing?

→ The absolute gaps between the no with change with the magnitude of number being represented

- From the number $(2^{-1} - 2^0)$, the absolute difference is constant $2^{-4} = 0.0625$.
- From the number $(2^0 - 2^1)$, the absolute difference is $2^{-3} = 0.125$.
- From $2^1 - 2^2$, the difference is $2^{-2} = 0.25$.

Problem - 5

What can you say about machine epsilon for our toy floating point system?

Sol:- Pick $x \in R$, there exists $x' \in F$ such that $\frac{|x - x'|}{|x|} \leq \epsilon_{machine}$

→ For toy system, the machine epsilon is $2^{-4} = 0.0625$ or $\frac{1}{16}$

→ due to avaliable of less than of bits, we assign the number to the nearest round off value

→ with som small error we assign to near value.

$fl(x) = x(1 + \epsilon)$ ϵ is small number

$|\epsilon| \leq \epsilon_{mach}$ ϵ_{mach} is machine epsilon

→ 2^{-1} to $2^0 = \frac{0.0625}{1} = 0.0625 = \frac{1}{16}$ ⇒ To get precision $|\epsilon| \leq \epsilon_{machine}$

→ 2^0 to $2^1 = \frac{0.125}{2} = 0.625 = \frac{1}{16}$