CSE330 Assignment-1 [Spring-2024]
[CO4]

Instructions for submission: [Handwritten submission]

- **Write your Name, Student_ID, Section No. in the cover page of the assignment.**
- **Mark the answers properly for each corresponding question.**

---

1. In the classes, we discussed three forms of floating number representations as shown below,

Standard/General Form: $F = \pm(0.d1d2d3 \cdots dm)_\beta \beta^e$,

Normalized Form: $F = \pm(1.d1d2d3 \cdots dm)_\beta \beta^e$,

Denormalized Form: $F = \pm(0.1d1d2d3 \cdots dm)_\beta \beta^e$

Now, let's take, **β = 2, m = 3 and −2 ≤ e ≤ 2**. Based on these, answer the following:
(a) (4 marks) How many **numbers in total/ possible combinations** can be represented by this system? Find this separately for each of the three forms above. Ignore negative numbers.
(b) (3 marks) What are the **maximum/largest numbers** that can be stored in the system by these three forms defined above (express your answer in decimal values)?
(c) (3 marks) What are the **non-negative minimum/smallest numbers** that can be stored in the system by the three forms defined above (express your answer in decimal values)?
(d) (4 marks) What are the **maximum/largest and minimum/smallest numbers** that can be stored in the system by the three forms defined above if the system has **negative support**?

2. Consider the **real number x = $(6.235)_{10}$**
(a) (3 marks) First convert the decimal number x in binary format at least up to 9 decimal/binary places.
(b) (2 marks) What will be the binary value of x [**Find fl(x)**] if you store it in a system with **m = 5** using the **general/standard form** of Floating point representation.
(c) (3 marks) Now convert back to the decimal form the stored values you obtained in the previous part, and calculate the **rounding error**.

3. Given a system parameterized by β = 2, m = 3, exponent = 2. Answer the following questions.
(a) (3 marks) Compute the **Machine Epsilon** for Normalized and Denormalized form.
(b) (3 marks) Compute $|x|_{min}$ for Normalized and Denormalized form.
(c) (2 marks) Compute the maximum delta value using the Standard/General Form.