

CSE 330 (Summer 2024)

Assignment 1 [CO4]

1. In the class, we discussed three forms of floating number representations as shown below, (1) Standard/General Form, (2) Normalized Form, (3) Denormalized Form.

Now, let's take, $\beta = 2$, $m = 3$ and $-2 \leq e \leq 4$. Based on these, answer the following:

- (a) (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)?
- (b) (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)?
- (c) (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) (4 marks) What will be the binary value of x [**Find fl(x)**] if you store it in a system with $m = 5$ and $m = 6$ 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 of both numbers**.

3. Consider the quadratic equation, $2x^2 - 60x + 3 = 0$. Below calculate **up to 6 significant** figures.

- (a) (4 marks) Find out where the loss of significance occurs when you calculate the roots?
- (b) (3 marks) Show that the roots evaluated in the previous part do not satisfy the fundamental properties of a polynomial.
- (c) (3 marks) Evaluate the correct roots such that loss of significance does not occur.