CSE 330 Assignment-1

Total Marks: 20
Deadline: 5 November 2025

Floating-point Representations

Form	Representation and Description
Standard Form	$F = \pm 0.d_1 d_2 d_3 \ldots d_m \times \beta^e, d_1 = 1$
IEEE Normalized Form	$F = \pm 0.1 d_1 d_2 d_3 \ldots d_m \times \beta^e$
IEEE Denormalized Form	$F = \pm 1.d_1 d_2 d_3 \ldots d_m \times \beta^e$

Question 1. (5 Marks)

Consider a system with base $\beta = 2$, mantissa length $m = 4$, and exponent range $-2 \leq e \leq 2$.

(a) Find the maximum and minimum numbers this system can store with and without negative support. Express the numbers both in binary and decimal for all three forms (Standard, IEEE Normalized, and IEEE Denormalized).

(b) Determine how many numbers this system can represent or store in all these forms.

(c) Using **Standard Form**, find all the (positive) decimal numbers representable (i.e., without negative support).

Question 2. (5 Marks)

Consider the quadratic equation:

$$x^2 - 10x + 3 = 0$$

(a) Compute the roots of the quadratic equation while keeping to four significant figures.

(b) Explain how loss of significance occurs in this case due to subtraction of nearly equal numbers.

(c) Discuss an alternative approach to computing the roots to avoid loss of significance (for example, using a numerically stable formula) and determine the correct roots using that approach.

Question 3. (5 Marks)

If $x = \frac{3}{8}$ and $y = \frac{7}{8}$, find $\text{fl}(x \cdot y)$ where $m = 4$ in standard format. Also, check whether $\text{fl}(x \cdot y)$ can be stored in this format or not when $-2 < e < 2$.

Question 4. (5 Marks)

Consider the real number $x = (6.325)_{10}$.

(a) Convert the decimal number x into binary format.

(b) Find $\text{fl}(x)$ if you store it in a system with $m = 3$ using the **Standard Form** of floating-point representation.

(c) Convert back to decimal form the stored value you obtained in the previous part, and calculate the rounding error and the machine epsilon.