Instructions for preparing the solution script:

- Write your name, ID#, and Section number clearly in the very front page.
- Write all answers sequentially.
- Start answering a question (not the pat of the question) from the top of a new page.
- Write legibly and in orderly fashion maintaining all mathematical norms and rules. Prepare a single solution file.
- Start working right away. There is no late submission form. If you miss the deadline, you need to use the make-up assignment to cover up the marks.

---

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

$$\begin{aligned} \text{Lecture Note Form} &: \quad F = \pm(0.d_1 d_2 d_3 \cdots d_m)_\beta \, \beta^e \ , && (1) \\ \text{Normalized Form} &: \quad F = \pm(1.d_1 d_2 d_3 \cdots d_m)_\beta \, \beta^e \ , && (2) \\ \text{Denormalized Form} &: \quad F = \pm(0.1 d_1 d_2 d_3 \cdots d_m)_\beta \, \beta^e \ , , && (3) \end{aligned}$$

where $d_i, \beta, e \in \mathbb{Z}$, $0 \leq d_i \leq \beta - 1$ and $e_{\min} \leq e \leq e_{\max}$. Now, let's take a system where, $\beta = 2$, $m = 4$ and $-3 \leq e \leq 6$. Based on these, answer the following:

   (a) (3 marks) Find how many non-negative numbers in total can be represented by this system?
   Find this separately for each of the three forms above.

   (b) (3 marks) What are the largest/maximum numbers that can be stored in the system for each of the three forms defined above?

   (c) (3 marks) What are the non-negative smallest/minimum numbers that can be stored in the system for each of the three forms defined above?

   (d) (4 marks) Using Eq.(1), find all the decimal numbers for $e = -1$, plot them on a real line and show if the number line is equally spaced or not.

2. Given a system parameterized by $\beta = 2$, fraction = 3 bit, exponent = 4 bit. Note that $e_{\min}$ and $e_{\max}$ are reserved respectively for zero and inf. Answer the following questions:

   (a) (4 marks) Compute the minimum and maximum of $|x|$ for denormalized and normalized form.

   (b) (2 marks) Compute the Machine Epsilon value for the normalized form.

   (c) (2 marks) Compute the maximum delta value for the Lecture Note form/General convention.

3. Given a system parameterized by $\beta = 2$, $m = 3$, $e_{\min} = -1$ and $e_{\max} = 2$. For this system, answer the following questions:

   (a) (3 marks) Find the floating-point representation of the numbers $(6.25)_{10}$ and $(6.875)_{10}$ in the Normalized Form. That means, find fl[6.25] and fl[6.875].

   (b) (2 marks) What are the rounding errors $\delta 1$, $\delta 2$ in part (a)?

   (c) (3 marks) Can the values $(6.25)_{10}$ and $(6.875)_{10}$ be represented in the Denormalized Form? If so, find the floating-point representations. If not, then concisely explain why?

   (d) (3 marks) Find the upper bound of the rounding error(Machine epsilon) for Lecture Note, Normalized and Denormalized Forms.

Motto: Mathematics is NOT difficult, but what is difficult is to believe that mathematics is NOT difficult.