

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 part of the question) from the top of a new page.
- Write legibly and in orderly fashion maintaining all mathematical norms and rules.
- Start working right away based on whatever you know. **Do not wait for the last moment and ask for time extension.**

1. In the classes, there are three forms of floating number representation,

$$\begin{aligned} \text{Lecture Note Form} & : F = \pm(0.d_1d_2d_3 \cdots d_m)_\beta \beta^e, \\ \text{Normalized Form} & : F = \pm(1.d_1d_2d_3 \cdots d_m)_\beta \beta^e, \\ \text{Denormalized Form} & : F = \pm(0.1d_1d_2d_3 \cdots d_m)_\beta \beta^e, \end{aligned}$$

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, $\beta = 2$, $m = 4$ and $e \in \{-2, -1, 0, 1, 2\}$. Based on these, answer the following:

- (3 marks) How many numbers in total can be represented by this system? Find this separately for each of the three forms above. Ignore negative numbers.
- (3 marks) For each of the three forms, find the smallest, positive number and the largest number representable by the system.
- (2 marks) For the IEEE standard (1985) for double-precision (64-bit) arithmetic, find the smallest, positive number and the largest number representable by a system that follows this standard. Do not find their decimal values, but simply represent the numbers in the following format:

$$\pm(0.1d_1 \dots d_m)_\beta \cdot \beta^{e - \text{exponentBias}}$$

Be mindful of the conditions for representing $\pm \text{inf}$ and ± 0 in this IEEE standard.

- (2 marks) In the above IEEE standard, if the exponent bias were to be altered to $\text{exponentBias} = 500$, what would the smallest, positive number and the largest number be? Write your answers in the same format as in part (c). Note that the conditions for representing $\pm \infty$ and ± 0 are still maintained as before.

2. Given a system parameterized by $\beta = 2$, $m = 3$, and $e_{\min} = -1 \leq e \leq e_{\max} = 2$ where $e \in \mathbb{Z}$. For this system,

- (3 marks) find the floating-point representation of the numbers $(6.25)_{10}$ and $(6.875)_{10}$ in the Normalized Form. That is, find $\text{fl}[6.25]$ and $\text{fl}[6.875]$.
- (2 marks) what are the rounding errors δ_1, δ_2 in part (a)?
- (2 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?
- (3 marks) find the upper bound of the rounding error for Lecture Note, Normalized and Denormalized Forms.

3. The following nodes come from the function $f(x) = \ln(5x + 9)$:

x	$f(x)$
-0.5	1.87
0	2.20
0.5	2.44

- (4 marks) Using Newton's divided difference method, find the equation of a second degree polynomial which fits the above data points.
- (5 marks) Expand the function $f(x) = \ln(5x + 9)$ using Taylor Series, centered at 0. Include till the x^2 term of the Taylor series.
- (1 mark) Should the equation which you found in part (a) and part (b) match? Comment on why, or why not.

4. Consider the following nodes:

x	$f(x)$
0	5
3	9.5
6	5

- (a) (1 mark) If an equation of a polynomial which fits through the above nodes is found using both the Vandermonde Matrix approach and the Lagrange approach, will both the equations match?
- (b) (7 marks) Find the equation of a polynomial which fits through the above nodes using the Vandermonde matrix approach.
- (c) (7 marks) Find the equation of a polynomial which fits through the above nodes using the Lagrange approach.

5. Consider the following data set:

x	$f(x)$	$f'(x)$
0.1	-0.620	3.585
0.2	-0.283	3.140

Answer the following based on the above data:

- (a) (8 marks) Compute the Hermite bases: $h_0(x)$, $h_1(x)$, $\hat{h}_0(x)$ and $\hat{h}_1(x)$.
 - (b) (2 marks) Write the Hermite polynomial and find the value at $x = 0.15$.
6. (5 marks) During the class, the derivation of Eq.(2.17) for a_1 (which is the Example in the lecture notes on page-19) is shown in detail. However the derivation of Eq.(2.18) for a_2 has some missing steps (the dotted part in Eq.-2.18 in page-19 of the lecture note). Now, you are asked show the detail derivation of the following

$$a_2 \equiv f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} .$$