1. The floating point representation can be expressed in any of the following forms:

Standard Form	:	$F = (\pm\ 0.d_1d_2d_3\cdots d_m)_\beta \times \beta^e;\quad d_1 \neq 0.$
IEEE Normalized Form	:	$F = (\pm\ 0.1d_1d_2d_3\cdots d_m)_\beta \times \beta^e.$
IEEE Denormalized Form	:	$F = (\pm\ 1.d_1d_2d_3\cdots d_m)_\beta \times \beta^e.$

a) Consider a system with $\beta = 2$, $m = 4$, and $-3 \leq e \leq 4$. Find out the **maximum** and **minimum** numbers this system can store with and without **negative support**. Express the numbers both in binary and decimal digits for all three forms.

b) How many numbers can this system represent or store in all these forms?

c) Using **Standard Form**, find all the decimal numbers without n**egative support**, plot them on a real line, and show if the number line is **equally spaced** or not.

d) For the **IEEE standard** 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:

$$(\pm\ 0.1d_1d_2d_3\cdots d_m)_\beta \times \beta^{e-exponentBias}.$$

Be mindful of the conditions for representing ±∞ and ±0 in this IEEE standard.

e) In the above IEEE standard, if the exponent bias were to be altered to $exponetBias = 500$, what would the **smallest positive number** and the **largest number** be? Write your answers in the same format as in part (d). Note that the conditions for representing ±∞ and ±0 are still maintained as before.

2. If $x = 3/8$ and $y = 5/8$, find $fl(x \times y)$ where $m = 4$. Also check whether $x \times y = fl(x \times y)$. If not, find the **rounding error** of the product of these two numbers.
3. Consider the quadratic equation, $x^2 - 60x + 1 = 0$. Working to **6 significant figures**, compute the **roots** of the quadratic equation and check that there is a **loss of significance**. Find the **correct roots** such that loss of significance does not occur.
4. Given $\beta = 2$, $m = 5$, $-100 \leq e \leq 100$. Using the IEEE **Normalized form**, answer the following:

a) Compute the Machine Epsilon (ϵ_M).

b) Compute the minimum of |x|.

c) How many non-negative numbers can you represent using this system?

5. Consider the quadratic equation $x^2 - 16x + 3 = 0$. Explain how the loss of significance occurs in finding the roots of the quadratic equation if we restrict to **4 significant figures**. Discuss how to avoid this and find the roots.

6. Given a system parameterized by β = 2, m = 3, and e_{min} = −1 ≤ e ≤ e_{max} = 2, where $e \in Z$. For this system answer the following:

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

   (b) What are the rounding errors δ_1,δ_2 in part (a)?

   (c) 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) Find the rounding error for Standard Form, Normalized and Denormalized Form.

7. Consider the **real number x =** $(8.235)_{10}$

   (a) First convert the decimal number x in binary format at least up to 8 binary places.

   (b) What will be the binary value of x [Find fl(x)] if you store it in a system with **m = 6** using the **Denormalized** form of floating point representation.

   (c) Now convert back to decimal form the stored values you obtained in the previous part, and calculate the **rounding error of both numbers**.

8. Consider the quadratic equation:

$$x^2 - 12x + 5 = 0$$

   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 the subtraction of nearly equal numbers.

   c) Discuss an alternative approach to computing the roots to **avoid loss of significance**, and use this method to determine the correct roots.

9. Consider a computing system with base β =2, m = 3, and $e_{\min}$ = −3 ≤ e ≤ $e_{\max}$ = 2

   a) In the Standard form of this system, determine the total number of representable values including support for negative numbers. Also, compute the maximum value of delta.

   b) Express the floating-point representations (binary format) for the numbers $x = 4/8$ and $y = 7/8$ in this system.

   c) Compute $fl(x \times y)$ and determine whether this value can be stored within the given floating-point system.