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

Standard Form : 
$$F = \pm (0.d_1 d_2 d_3 \cdots d_m)_\beta \beta^e$$
,  $(d_1 \neq 0)$  (1)

IEEE Normalized Form : 
$$F = \pm (0.1d_1d_2d_3\cdots d_m)_\beta \beta^e$$
, (2)

IEEE Denormalized Form : 
$$F = \pm (1.d_1 d_2 d_3 \cdots d_m)_{\beta} \beta^e$$
, (3)

where  $d_i, \beta, e \in \mathbb{Z}$ ,  $0 \le d_i \le \beta - 1$  and  $e_{\min} \le e \le e_{\max}$ . Now, let's take,  $\beta = 2$ , m = 5 and  $-2 \le e \le 5$ . Based on these, answer the following:

(a) (6 marks) What are the maximum numbers that can be stored in the system by these three forms defined above (express your answer in decimal values)?

Solution: The maximum numbers that can be stored in these three systems are

$$\begin{array}{lll} \text{Stabdard Form} &=& \left(0.11111\right)_2 \times 2^{e_{\max}} \;, \\ &=& \left(1 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3} + 1 \times 2^{-4} + 1 \times 2^{-5}\right)_2 \times 2^5 \;, \\ &=& \left(2^4 + 2^3 + 2^2 + 2^1 + 2^0\right) = \left(31\right)_{10} \;.\checkmark \\ \text{IEEE Normalized Form} &=& \left(0.111111\right)_2 \times 2^{e_{\max}} \;, \\ &=& \left(1 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3} + 1 \times 2^{-4} + 1 \times 2^{-5} + 1 \times 2^{-6}\right)_2 \times 2^5 \;, \\ &=& \left(2^4 + 2^3 + 2^2 + 2^1 + 2^0 + 2^{-1}\right) = \left(31.5\right)_{10} \;.\checkmark \\ \text{IEEE Denormalized Form} &=& \left(1.11111\right)_2 \times 2^{e_{\max}} \;, \\ &=& \left(1 \times 2^0 + 1 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3} + 1 \times 2^{-4} + 1 \times 2^{-5}\right)_2 \times 2^5 \;. \\ &=& \left(2^5 + 2^4 + 2^3 + 2^2 + 2^1 + 2^0\right) = \left(63\right)_{10} \;.\checkmark \end{array}$$

(b) (6 marks) What are the non-negative minimum numbers that can be stored in the system by the three forms defined above (express your answer in decimal values)?

Solution: The non-negative minimum numbers that can be stored in these three systems are

$$\begin{array}{rcll} \text{Stabdard Form} &=& \left(0.10000\right)_2 \times 2^{e_{\min}} = \left(1 \times 2^{-1}\right)_2 \times 2^{-2} = \left(\frac{1}{2} \times \frac{1}{4}\right) = \left(\frac{1}{8}\right)_{10} = (0.125)_{10} \; . \checkmark \\ \text{IEEE Normalized Form} &=& \left(0.100000\right)_2 \times 2^{e_{\min}} = \left(1 \times 2^{-1}\right)_2 \times 2^{-2} = \left(\frac{1}{2} \times \frac{1}{4}\right) = \left(\frac{1}{8}\right)_{10} = (0.125)_{10} \; . \checkmark \\ \text{IEEE Denormalized Form} &=& \left(1.00000\right)_2 \times 2^{e_{\min}} = \left(1 \times 2^0\right)_2 \times 2^{-2} = \left(\frac{1}{4}\right)_{10} = (0.25)_{10} \; . \checkmark \end{array}$$

(c) (6 marks) Including negative numbers, what range of the floating numbers in these three representations are considered as ZERO and  $\pm \infty$  because of the underflow and overflow respectively.

**Solution**: By definition, the underflow is a phenomena where all values equal and less than  $|x_{\min}|$  are considered to be ZERO, and all value equal to and above  $|x_{\max}|$  are considered to be  $\infty$ . Therefore, we can write,

Stabdard Form : any value  $|\pm 0.125|$  or less is ZERO and any value  $\pm 31|$  or above is  $\infty$ .  $\checkmark$  IEEE Denormalized Form : any value  $|\pm 0.125|$  or less is ZERO and any value  $\pm 31.5|$  or above is  $\infty$ .  $\checkmark$  IEEE Denormalized Form : any value  $|\pm 0.25|$  or less is ZERO and any value  $\pm 63|$  or above is  $\infty$ .  $\checkmark$ 

- 2. Consider the quadratic equation,  $x^2 60x + 1 = 0$ . Below calculate up to 6 significant figures.
  - (a) (4 marks) **Find out** where the loss of significance occur when you calculate the roots?

**Solution**: Let  $x_1$  and  $x_2$  are the roots. Now, the general solution of the above quadratic equation is (up to 6 significant figures)

$$x = \frac{-(-60) \pm \sqrt{(-60)^2 - 4 \times 1 \times 1}}{2 \times 1} = 30 \pm \sqrt{899} = 30 \pm 29.9833$$
.

Hence the roots are:  $x_1 = 30 + 29.9833 = 59.9833$  and  $x_2 = 30 - 29.9833 = 0.0167000$ . The loss of significance occur in evaluating the value of  $x_2$  because we are subtracting two very close numbers.  $\checkmark$ 

(b) (4 marks) **Show that** the roots evaluated in the previous part do not satisfy the fundamental properties of a polynomial.

**Solution**: By the fundamental properties of algebra/polynomial, the sum of the roots must be equal to 60 and the product of the roots must be equal to 1 within 6 significant figures. Here we obtain

$$x_1 + x_2 = 59.9833 + 0.0167000 = 60.0000$$
 .(Satisfied)  
 $x_1 x_2 = 59.9833 \times 0.0167000 = 1.00172 \neq 1 \text{(Not satisfied)} \checkmark$ 

(c) (4 marks) **Evaluate** the correct roots such that loss of significance does not occur.

**Solution**: Since the loss of significance occur in evaluating the value of  $x_2$ , we recalculate  $x_2$  by using the fundamental property. That is,

$$x_1x_2 = 1 \implies x_2 = \frac{1}{59.9833} = 0.0166713$$
 (within 6 significant figures).

Hence  $x_1x_2 = 1$  is automatically satisfied. We also find that  $x_1 + x_2 = 59.9833 + 0.0166713 = 60.0000$  (within 6 significant figures) is also satisfied. Hence the correct roots are:  $x_1 = 59.9833$  and  $x_2 = 0.0166713$ .  $\checkmark$