**Problem-1**

In Floating point representation we have three components

1. The Sign Bit

2. Exponent

3. Fractional Part.

Q1. Precession is one the prime attribute of any Floating point Representation. Does any of the above three components play a role in the defining the Precession of the number ? If so which are the component or Components  which play the  role in defining precession  and how ? Explain this with example in your own words.

A1. Precision of a floating point number is defined by number of bits, not any of above. Sign bit, Exponent and fraction determines the value of the number. According to IEEE 754 standard, precision ranges are,

half-Precision 16-bit : 1 Sign bit + 5 exponent bit + 10 Fraction bit

Single-Precision 32-bit : 1 Sign bit + 8 exponent bit + 23 Fraction bit

Double-Precision 64-bit : 1 Sign bit + 11 exponent bit + 52 Fraction bit

Q2.What is Normal and de-normal  values as per IEEE 754  standards  explain this  with the  help of number line.

Here, examples are demonstrated using single precision format.

Fraction (23-bit)

Exponent (8-bit)

S

Normalized Form:

Ex. For The sequence 1 01111010 101 0000 0000 0000 0000 0000,

S = 1

E = 0111 1010 = 122

F = 1 + 2-1 + 2-3 = 1.725

Decimal = -2(122-127) x 1.725 = -0.05390625

The problem with normalized form is, we can not represent zero, as the fraction is always add to 1.So, de-normalized form is introduced where E = 0 and

Ex. For The sequence 0 0000000 000 1000 0000 0000 0000 0000,

S = 1

E = 0 🡺 denormalized form.

F = 0 + 2-4 = 0.0625

Decimal = 2(-126) x 0.0625 ~= 0

Denormalized

-2-126 0 2-126