**Assignment-3**

**1. Yes, components of floating-point representation are responsible for precision of the number.**

The component is the fractional part because precision is obtained through the fractional part of the decimal number.

For example-

Consider a decimal number +152.62587.

The equivalent binary representation of number is

= 10011000.1010000000111001

=1.00110001010000000111001 x 27

=1.00110001010000000111001 x 2134-127

=1.00110001010000000111001 x 210000110-01111111

Thus, IEEE 754 representation is

0 10000110 00110001010000000111001

Now consider decimal number +152.625

The equivalent binary representation of number is

= 10011000.101

=1.0011000101x 27

=1.0011000101x 2134-127

=1.0011000101x 210000110-01111111

Thus, IEEE 754 representation is

0 10000110 00110001010000000000000

**2**.

**A single precision normal value is represented as** –

Consider the sign bit [MSB] s, the exponent [8 bits after MSB] e and the mantissa (significand or fraction) [lower 23 bits] m. The valid range of the exponents is 1 to 254 (when e is an unsigned number).

The actual exponent is biased by 127 to get e i.e. the actual value of the exponent is e − 127. This gives the range: 2 1 −127 = 2 −126 to 2 254 −127 = 2127.

The normalized significand is 1.m. The binary point is before bit-22 and the 1 (one) is not present explicitly.

The sign bit s = 1 for a negative number is zero (0) for a positive number.

The value of a normalized number is-

( -1) s x 1.m x 2e-127

The smallest magnitude of a normalized number in single precession is

x 0000 0001 000 0000 0000 0000 0000 0000

whose value is 1.0 × 2 −126.

The largest magnitude of a normalized number in single precession is

x 1111 1110 111 1111 1111 1111 1111 1111

whose value is 1.99999988 × 2127 ≈ 3.403 × 1038

**A subnormal value is represented as –**

 Any non-zero number with [magnitude](https://en.wikipedia.org/wiki/Magnitude_(mathematics)#Numbers) smaller than the smallest [normal number](https://en.wikipedia.org/wiki/Normal_number_(computing)) is **subnormal**.

The interpretation of a subnormal number is different. The content of the exponent part (e) is zero and the significand part (m) is non-zero.

There is no implicit one in the significand.

The value of a subnormal number is-

( -1) s x 0.m x 2-126

The smallest magnitude of a subnormal number in single precession is

± 0000 0000 000 0000 0000 0000 0000 0001

whose value is 2 −126+( −23) = 2 −149.

The largest magnitude of a subnormal number in single precession is

± 0000 0000 111 1111 1111 1111 1111 1111

whose value is 0.99999988 × 2 −126.

Representation on number line is-

1. Magnitude wise

Subnormal range Normal range

**- 2 −149 1.0 × 2 −126 3.403 ×1038**

1. Sign wise

**-2-126 -2-1492-1492-126**

**- 3.403 ×1038 0 3.403 ×10**

Normal Range Subnormal Range

1. **Five methods of rounding floating point number are:**
2. roundTiesToEven

It rounds the floating-point number to the nearest floating-point number. If number is exactly the middle of two consecutive floating-point numbers, it is rounded to the floating-point number whose least significand digit is even. This is the default rounding mode.

Example-

1. Consider result of an arithmetic operation is = 1.2376865

Then the rounded off value will be 1.237686

1. Consider result of an arithmetic operation is = 3.4879673.

Then the rounded off value will be 3.487968

1. roundTiesToAway

It rounds the floating-point number to the nearest floating-point number. If number is exactly the middle of two consecutive floating-point numbers, it is rounded to the floating-point number with the larger fractional part.

Example-

1. Consider result of an arithmetic operation is = 1.2376875

Then the rounded off value will be 1.237688

1. Consider result of an arithmetic operation is = -1.2376875

Then the rounded off value will be -1.237687

1. roundTowardPositive

It rounds the number x to the nearest floating-point number that is greater than x.

Example-

1. Consider result of an arithmetic operation is = 1.2376843.

Then the rounded off value will be 1.237685.

1. Consider result of an arithmetic operation is = -3.86723.

Then the rounded off value will be -3.8672.

1. roundTowardNegative

It rounds the number x to the nearest floating-point number that is less than x.

Example-

1. Consider result of an arithmetic operation is = 1.2376843.

Then the rounded off value will be 1.237684.

1. Consider result of an arithmetic operation is = -3.86723.

Then the rounded off value will be -3.8673.

1. roundTowardZero

it rounds the number x to the nearest floating-point number that is no greater in magnitude than x.

Example-

1. Consider result of an arithmetic operation is = 1.2376843.

Then the rounded off value will be 1.237684.

1. Consider result of an arithmetic operation is = -3.86723.

Then the rounded off value will be -3.8672.