Assignment3

-G.DHARMIK(IMT2015016)

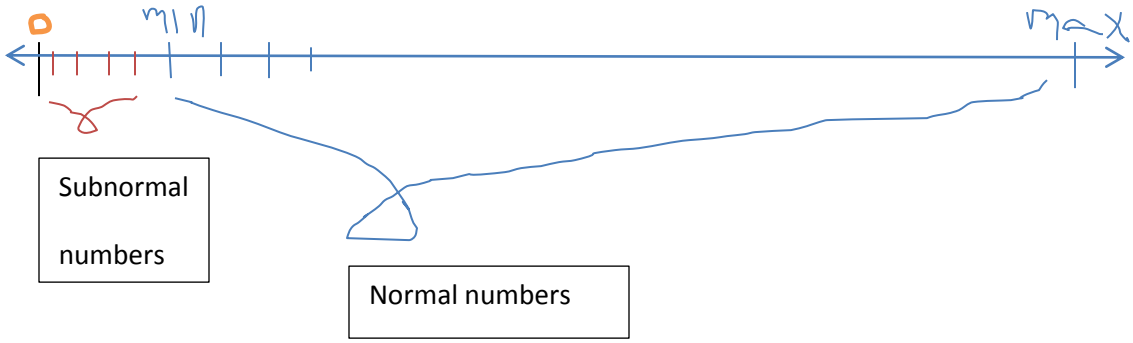
1)Yes,the precision depends on the fraction part and exponent bit. For example 1.13461111111111111116666666661 can be represented more better if we have 23 bits fraction part than 9 bits.It even depends on the exponent bits.If we have more exponent bits we could represent even the smaller numbers i.e underflow and overflow can be covered avoiding rounding off.

2)Normal:

Any non-zero floating point number which is between the minimum number and maximum number for a format is called Normal number.

Subnormal:

Any non-zero floating point number whose magnitude is less than the magnitude of minimum number of that format is subnormal number.subnormal numbers fill the gap of underflow.



For example:

Assume 32-bit FPU,Here 23 bits are for the decimals and 8 bits are for the exponent part.exponent is stored by adding the power with 128.So that we can accomadate for negative power.So the minimum positive number that can be represented is 2^-127.The numbers like 2^-130 cannot be represented using this method in the 32 bit FPU.As exponent will be -2 which cannot represented.These type of numbers are called subnormal.

3.Rounding off:

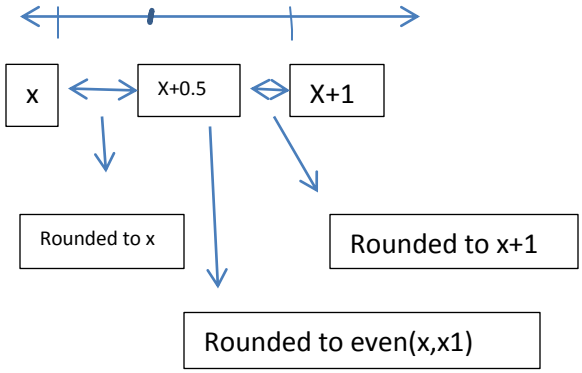
roundToIntegralTiesToEven:

In this method the number is rounded to the nearest integer .Consider a integer x,and number a.

If x<=a and a<x+0.5 then a is rounded to x.

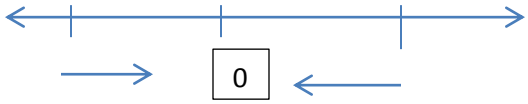
Else if a>x+0.5 then it is rounded to x+1.

If a=x+0.5 then it is rounded to the number x or x+1 which is even.



roundToIntegralTowardZero:

The number will be rounded off to the integer which is towards zero.



For example numbers like -2.5 will be rounded to -2 and numbers like 3.4 will be rounded to 3.

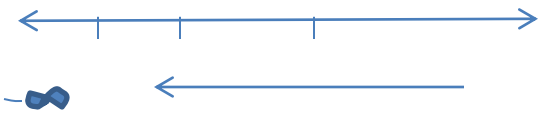
roundToIntegralTowardPositive:

The number will be rounded to the integer towards positive infinity.

Example 2.4 will be rounded to 3 whereas -2.4 will be rounded to -2.

roundToIntegralTowardNegative:

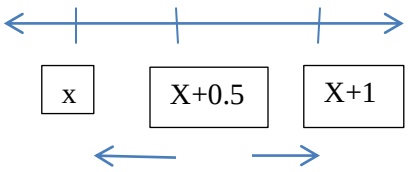
The number will be rounded to integer towards negative infinity.



Example 16.4 will be rounded off to 16 whereas -16.47 is rounded to -17.

roundToIntegralTiesToAway:

The number will be rounded off to the nearest integer.It is similar to first method except x+0.5 is rounded away to integer away from zero.



Example numbers like 2.4 rounds to 2.

2.5 to 2.

-2.4 to -2

-2.5 to -3.