**ARM Assignment 3 - Research Assignment**

1) Precession of the number

IEEE754 defines a technical standard for floating point computation. Any number has three parts in its representation i.e.

* Sign
* Exponent
* Mantissa

Sign Exponent Mantissa

Representation of a floating point number in IEEE754

Precision is defined as a measure of how accurately a number can be represented. Range is a quantifier of how wide a bunch of numbers can be represented.

Precision is directly dependent on the number of bits the mantissa gets. If there are more bits designated to the mantissa, the representation of the number is more accurate/precise. If there are fewer bits, then two numbers which are close to each other in value might end up with the same representation.

IEEE 754 defines two kinds of precision: single and double. Single has 23 bits dedicated to the mantissa while double has 52.

Let’s consider an example:

The value of e upto 10 decimal places is 2.718281828

Representation in single precision

0 10000000 010 1101 1111 1000 0101 0100

Representation in double precision

0 10000000000 0101 1011 1111 0000 1010 1000 1011 0000 0100 1001 0001 1001 1011

## On converting these back to decimal values,

## Single precision= 2.718 280 792

## Double precision= 2.718 281 827

It can be seen that there is less accuracy while storing e in single precision format whereas double precision has more accuracy. This is due to the number of bits allocated to the mantissa.

2) Normal and subnormal values

According to IEEE754, there is a range of numbers that can be represented. By definition, normal values are those which fall into the prescribed range of a floating point format. In contrast, subnormal values are those whose exponent is zero. Subnormal values are those numbers whose value is lesser than the smallest normal number.

Purpose of subnormal values

According to IEEE754, the magnitude of the smallest possible number is *bemin* where emin is the minimum value the exponent can take.

Suppose this value is 2.This means that the smallest number that can be represented is

-2^2 which is 0.25. Now, if we had to subtract 0.125 from 0.25, the result would be zero because these numbers are represented by the same value, if we had only normal numbers. Such conditions are termed underflows. So for this subnormal numbers are introduced. These numbers are permitted to have a leading zero in their mantissa. By doing this, the gap between zero and the smallest number that can be represented is brought down.

3) Rounding methods

IEEE754 defines 5 rounding methods.

* Round to nearest, ties to even

This method rounds to the nearest value. If the number falls exactly midway, then the value with LSB 0(even) is chosen.

Ex: 7.5 -> 8 6.5 -> 6

* Round to the nearest ties away from zero

This method also rounds to the nearest value; but in case of a tie, positive numbers are rounded off to the closest number above whereas negative numbers are rounded to the closest number below.

Ex: 6.5 -> 6 -7.5 -> -8

* Round up, or round toward plus infinity (ceiling)

Numbers are rounds to a larger number

Ex: 7.5 -> 8 -7.5 -> -7

* Round down, or round toward minus infinity (floor)

Numbers are always rounded to a smaller value.

Ex:7.5 -> 7 -7.5 -> -8

* Round toward zero, or chop, or truncate (truncation)

The output chosen is always closer to 0.

Ex.7.5 -> 7 -.6.5 -> -6