*Lab Assignment 2: Q5*

*Group Number: 14*

*Q5.*

We assume floating point arithmetic followed to be (b, p, Emin, Emax)  
In the IEEE Double Precision System, b=2, p=53, Emin=−1022, Emax=1023

x lies between 0 and 1 throughout the execution of both the programs. Hence, we have p-1 digits after decimal point to represent x’s significand.

Part (i)

Loop terminates at x=1.11022302462516e-16 and k=53.  
The loop brakes when 1 + x <= 1, that is, when 1+x is rounded down to 1 (=1.000..000 \* b0 in floating point arithmetic).

We have p-1 digits to represent x’s significand.

1 + x will be rounded down to 1 when x< b-(p-1).

Thus, the program terminates as soon as x < b-(p-1).

Now, x=2-k

Program terminates when x=2-k < b-(p-1)=2-52

Hence, program terminates at k=53.

Part (ii)

Loop terminates at x=0.00000000000000e+00 and k=1075.  
The loop brakes when x + x <= x.

Since x > 0, x + x <= x only when x is rounded down to 0 by the floating point arithmetic.

So, the program runs till x < bEmin \* b-(p-1).

Program terminates when x=2-k < bEmin-p+1=2-1074

Hence, program terminates at k=1075.

Hence, this program runs longer than the first one. It also gives a value of x lower than the first program.

This exercise demonstrates how subnormal numbers ensure very gradual underflow to 0 in IEEE system.