# AMATH 583: HW 1

Minho Choi

April 7th, 2023

## Problem 1

The values of **j** (precision) are:

- Single Precision: 24
- Double Precision: 53

## Problem 2

(1) For Single Precision, the largest number that we can get is:

$$\left(1+\frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{2^{23}}\right)\cdot 2^{127}=\left(2-\frac{1}{2^{23}}\right)\cdot 2^{127}\approx 3.40282\times 10^{38}$$

For the smallest number, it is simply the negative value of the above value:

$$-\left(2-\frac{1}{2^{23}}\right)\cdot 2^{127}\approx -3.40282\times 10^{38}$$

The smallest absolute value number that we can get is:

$$2^{-126}=1.17549\times 10^{-38}$$

(2) For Double Precision, the largest number that we can get is:

$$\left(1+\frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{2^{52}}\right)\cdot 2^{1023}=\left(2-\frac{1}{2^{52}}\right)\cdot 2^{1023}\approx 1.79769\times 10^{308}$$

For the smallest number, it is simply the negative value of the above value:

$$-\left(2-\frac{1}{2^{52}}\right)\cdot 2^{1023}\approx -1.79769\times 10^{308}$$

The smallest absolute value number that we can get is:

$$2^{-1022}=2.22507\times 10^{-308}$$

## Problem 3

When we define the product of integers to **int** variable, we get -884901888 with a warning about the *overflow*. When we use cast the product to **int64_t**, we get the desired answer 12000000000. When the *overflow* happens the number goes around to the smallest number. Hence, the largest representable number by **int** variable is 2147483647, and so if we get 2147483648, then the result we get is -2147483648 which is the smallest representable number by **int** variable. In our example, since we have:

$$12000000000\div 2^{31}\approx 5.5879$$

Hence, we get:

$$12000000000-6\times 2^{31}=-884901888$$

For definitions:

- The *underflow* happens when the number is smaller than the smallest representable number in magnitude (by either SP or DP). In other words, the number is closer to zero than the closest number that is representable by either SP or DP.
- The *overflow* happens when the number is larger than the largest representable number in magnitude (by either SP or DP). Also, the *overflow* occurs when the number is smaller than the smallest representable number.

For SP, we have *underflow* when the absolute value of the number is smaller than $1.17549 \times 10^{-38}$ and we have *overflow* when the absolute value of the number is larger than $3.40282 \times 10^{38}$.
For DP, we have *underflow* when the absolute value of the number is smaller than $2.22507 \times 10^{-308}$ and we have *overflow* when the absolute value of the number is larger than $1.79769 \times 10^{308}$.

# Problem 4

For SP, we have:

- 2 possibilities for the sign bit.
- The range of the exponent is $[-126, 127]$, which gives $127 + 126 + 1 = 254$ possibilities for the exponent bits. (Also, we can think as there are $2^8$ possibilities since there are 8 exponent bits, but 2 possibilities, one where all bits are zero and another where all bits are one, are excluded giving $2^8 - 2 = 254$)
- $2^{23}$ possibilities for the mantissa bits.

Therefore, in total, we have:

$$2 \times 254 \times 2^{23} = 4261412864$$

Similarly, for DP, we have:

- 2 possibilities for the sign bit.
- The range of the exponent is $[-1022, 1023]$, which gives $1023 + 1022 + 1 = 2046$ possibilities for the exponent bits.
- $2^{52}$ possibilities for the mantissa bits.

Therefore, in total, we have:

$$2 \times 2046 \times 2^{52} \approx 1.842873 \times 10^{19}$$

Thus, the general formula is:

$$2^s \cdot (2^k - 2) \cdot 2^n$$

where $s$ is number of sign bits (it is either 0 or 1), $k$ is number of exponent bits, and $n$ is number of mantissa bits.

# Problem 5

For normalized case, we have:

- For $M_{00} = 1$, we have:

  $$M_{00} \cdot 2^{E_{001}} = 1 \cdot 2^{-2} = \frac{1}{4}, \quad M_{00} \cdot 2^{E_{010}} = 1 \cdot 2^{-1} = \frac{1}{2}, \quad M_{00} \cdot 2^{E_{011}} = 1 \cdot 2^{0} = 1$$
  $$M_{00} \cdot 2^{E_{100}} = 1 \cdot 2^{1} = 2, \quad M_{00} \cdot 2^{E_{101}} = 1 \cdot 2^{2} = 4, \quad M_{00} \cdot 2^{E_{110}} = 1 \cdot 2^{3} = 8$$

  Hence, we get:

  $$s = 1, v < 0 : \left\{ -\frac{1}{4}, -\frac{1}{2}, -1, -2, -4, -8 \right\}$$
  $$s = 0, v > 0 : \left\{ \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8 \right\}$$

- For $M_{01} = \frac{5}{4}$, we have:

$$M_{01} \cdot 2^{E_{001}} = \frac{5}{4} \cdot 2^{-2} = \frac{5}{16}, \quad M_{01} \cdot 2^{E_{010}} = \frac{5}{4} \cdot 2^{-1} = \frac{5}{8}, \quad M_{01} \cdot 2^{E_{011}} = \frac{5}{4} \cdot 2^{0} = \frac{5}{4}$$
$$M_{01} \cdot 2^{E_{100}} = \frac{5}{4} \cdot 2^{1} = \frac{5}{2}, \quad M_{01} \cdot 2^{E_{101}} = \frac{5}{4} \cdot 2^{2} = 5, \quad M_{01} \cdot 2^{E_{110}} = \frac{5}{4} \cdot 2^{3} = 10$$

  Hence, we get:

$$s = 1, v < 0 : \left\{ -\frac{5}{16}, -\frac{5}{8}, -\frac{5}{4}, -\frac{5}{2}, -5, -10 \right\}$$
$$s = 0, v > 0 : \left\{ \frac{5}{16}, \frac{5}{8}, \frac{5}{4}, \frac{5}{2}, 5, 10 \right\}$$

- For $M_{10} = \frac{3}{2}$, we have:

$$M_{10} \cdot 2^{E_{001}} = \frac{3}{2} \cdot 2^{-2} = \frac{3}{8}, \quad M_{10} \cdot 2^{E_{010}} = \frac{3}{2} \cdot 2^{-1} = \frac{3}{4}, \quad M_{10} \cdot 2^{E_{011}} = \frac{3}{2} \cdot 2^{0} = \frac{3}{2}$$
$$M_{10} \cdot 2^{E_{100}} = \frac{3}{2} \cdot 2^{1} = 3, \quad M_{10} \cdot 2^{E_{101}} = \frac{3}{2} \cdot 2^{2} = 6, \quad M_{10} \cdot 2^{E_{110}} = \frac{3}{2} \cdot 2^{3} = 12$$

  Hence, we get:

$$s = 1, v < 0 : \left\{ -\frac{3}{8}, -\frac{3}{4}, -\frac{3}{2}, -3, -6, -12 \right\}$$
$$s = 0, v > 0 : \left\{ \frac{3}{8}, \frac{3}{4}, \frac{3}{2}, 3, 6, 12 \right\}$$

- For $M_{11} = \frac{7}{4}$, we have:

$$M_{11} \cdot 2^{E_{001}} = \frac{7}{4} \cdot 2^{-2} = \frac{7}{16}, \quad M_{11} \cdot 2^{E_{010}} = \frac{7}{4} \cdot 2^{-1} = \frac{7}{8}, \quad M_{11} \cdot 2^{E_{011}} = \frac{7}{4} \cdot 2^{0} = \frac{7}{4}$$
$$M_{11} \cdot 2^{E_{100}} = \frac{7}{4} \cdot 2^{1} = \frac{7}{2}, \quad M_{11} \cdot 2^{E_{101}} = \frac{7}{4} \cdot 2^{2} = 7, \quad M_{11} \cdot 2^{E_{110}} = \frac{7}{4} \cdot 2^{3} = 14$$

  Hence, we get:

$$s = 1, v < 0 : \left\{ -\frac{7}{16}, -\frac{7}{8}, -\frac{7}{4}, -\frac{7}{2}, -7, -14 \right\}$$
$$s = 0, v > 0 : \left\{ \frac{7}{16}, \frac{7}{8}, \frac{7}{4}, \frac{7}{2}, 7, 14 \right\}$$

For denormalized case, we have:

$$M_{00} \cdot 2^{E_{000}} = 0 \cdot 2^{-2} = 0, \quad M_{01} \cdot 2^{E_{000}} = \frac{1}{4} \cdot 2^{-2} = \frac{1}{16},$$
$$M_{10} \cdot 2^{E_{000}} = \frac{1}{2} \cdot 2^{-2} = \frac{1}{8}, \quad M_{11} \cdot 2^{E_{000}} = \frac{3}{4} \cdot 2^{-2} = \frac{3}{16}$$

Hence, we get:

$$s = 1, v < 0 : \left\{ -0, -\frac{1}{16}, -\frac{1}{8}, -\frac{3}{16} \right\}$$
$$s = 0, v > 0 : \left\{ +0, \frac{1}{16}, \frac{1}{8}, \frac{3}{16} \right\}$$

The distribution plot of all the numbers obtained is the following: