Worksheet

10/4/2019

Name: ______________________________

Below is a toy model of IEEE double-precision floating-point format to demonstrate how machine arithmetic is done.
A number is represented by a sequence of 8 bits:

$$\underbrace{c_0}_{\text{sign}} \quad \underbrace{c_1 \quad c_2 \quad c_3 \quad c_4}_{E} \quad \underbrace{a_1 \quad a_2 \quad a_3}_{\bar{x}}$$

This sequence represents the number $x = \sigma \cdot \bar{x} \cdot 2^e$ where

- If $1 \le E \le 14$ then

$$\begin{aligned} \sigma &= \begin{cases} 1 & \text{if} \quad c_0 = 0, \\ -1 & \text{if} \quad c_0 = 1, \end{cases} \\ e &= E - 7, \\ \bar{x} &= (1.a_1a_2a_3)_2 \end{aligned}$$

- If $E = 0$ then $e = -6$ and $\bar{x} = (0.a_1a_2a_3)_2$.
- If $E = 15$ then the bit sequence represents $\pm\infty$ (depending on the sign σ).

(a) What is the largest number (not counting ∞) that can be represented by this format?

$$(1.111)_2 \times 2^7 = \left(1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8}\right) \times 128 = 240$$

(b) What is the smallest positive number that can be represented by this format?

$$(0.001)_2 \times 2^{-6} = 2^{-3} \times 2^{-6} = 2^{-9} = 0.001953125$$

(c) What is the dynamic range of this number format?

$$\text{dynamic range} = \frac{\text{largest number}}{\text{smallest positive}} = \frac{240}{2^{-9}} = 122880$$

(d) What number does the bit sequence $\underbrace{0}\underbrace{1\,1\,0\,1}\underbrace{0\,0\,1}$ represent?

$$\left.\begin{aligned} c_0 &= 0 \rightsquigarrow \sigma = 1 \\ E &= (1101)_2 = 2^3 + 2^2 + 2^0 = 13 \\ \bar{x} &= (1.001)_2 = 2^0 + 2^{-3} = 1.125 \end{aligned}\right\} \quad x = \sigma\,\bar{x}\,2^e = 1 \times 1.125 \times 2^{13-7} = 72$$

(e) If the same bit sequence represented number $(c_0c_1c_2c_3c_4 \, . \, a_1a_2a_3)_2$, what would be the answers to Part (a), (b), (c)?

Largest number $= (11111.111)_2 = 2^4 + 2^3 + 2^2 + 2^1 + 2^0 + 2^{-1} + 2^{-2} + 2^{-3} = 31.875$

Smallest positive number $= (00000.001)_2 = 2^{-3} = 0.125$.

Dynamic range $= \dfrac{31.875}{0.125} = 255$ (much smaller than 122880)

(f) Perform the below floating-point arithmetic operations by following the procedure:

1. Rewrite the smaller number such that its exponent matches with the exponent of the larger number.
2. Add the significands.
3. Normalize the result by moving the floating point.
4. Round the result.

(f1) $(1.101)_2 \times 2^2 + (1.111)_2 \times 2^2 = (11.100)_2 \times 2^2$

$$\begin{array}{r} 1.101 \\ +\ 1.111 \\ \hline 11.100 \end{array}$$

$= (1.1100)_2 \times 2^3$

"=" $(1.110)_2 \times 2^3$ (rounding off)

(f2) $(1.011)_2 \times 2^2 - (0.111)_2 \times 2^2 = \underbrace{(0.100)_2 \times 2^2}_{\text{not yet in the format described on the previous page.}} = (1.000) \times 2^1$

$$\begin{array}{r} 1.011 \\ -0.111 \\ \hline 0.100 \end{array}$$

(f3) $-(1.010)_2 \times 2^1 + \underbrace{(1.011)_2 \times 2^{-1}}_{=(0.01011)_2 \times 2^1}$

$$\begin{array}{r} 1.010 \\ -0.01011 \\ \hline 0.11101 \end{array}$$

$= -[(1.010)_2 \times 2^1 - (0.01011)_2 \times 2^1] = -(0.11101)_2 \times 2^1$

$= -(1.1101) \times 2^0$ "=" $-(1.111)_2 \times 2^0$ (rounding off)