MATH 4334.001/CS 4334.001
Fall 2023
Paper Homework 3 Solutions

1. Determine the single-precision machine representation in a 32-bit word-length computer of the following decimal numbers:
   (a) 0.125
   (b) -9876.54321

(a) We have

$$(0.125)_{10} = (0.1)_8 = (0.001)_2 = 1.0 \times 2^{-3}$$

Thus, the exponent is -3 which we now rewrite in "excess 127" form:

$$c = -3 + 127 = (124)_{10} = (174)_8 = (001\ 111\ 100)_2$$

Since only 8 bits are allocated for the exponent, this reduces to

$$c = (01\ 111\ 100)_2$$

The mantissa is 0 and hence, the machine representation of 0.125 is

$$\boxed{0\ 01111100\ 00000000000000000000000}$$

(b) Here, we find that

$$\begin{aligned}
(9876)_{10} &= (23224)_8 \\
&= (010\ 011\ 010\ 010\ 100)_2 \\
(0.54321)_{10} &= (0.4260771740)_8 \\
&= (0.100\ 010\ 110\ 000\ 111\ 111\ 001\ 111\ 100\ 000)_2
\end{aligned}$$

and hence,

$$\begin{aligned}
(9876.54321)_{10} &= (010\ 011\ 010\ 010\ 100.100\ 010\ 110\ 000\ 111\ 111\ 001\ 111\ 100\ 000)_2 \\
&= (1.001\ 1010\ 0101\ 0010\ 0010\ 1100) \times 2^{13}
\end{aligned}$$

The exponent is 13 which, in "excess 127" form, is

$$c = 13 + 127 = (140)_{10} = (214)_8 = (010\ 001\ 100)_2 = (10001100)_2 \text{ (8 bits)}$$

Hence, the machine representation of -9876.54321 is

$$\boxed{1\ 10001100\ 00110100101001000101100}$$

2. Identify the floating-point numbers corresponding to the following bit strings:

   (a) $\boxed{0\ 11111111\ 00000000000000000000000}$

   (b) $\boxed{0\ 00000001\ 00000000000000000000000}$

   (a) Here, the floating-point number is

   $$(-1)^0 \times 2^{c-127} \times (1.0)_2$$

   where,

   $$\begin{aligned} c = (11111111)_2 &= 2^7 + 2^6 + 2^5 + 2^4 + 2^3 + 2^2 + 2^1 + 2^0 \\ &= 255 \end{aligned}$$

   However, we know that the value of c is restricted by $0 < c < 255$, with 255 reserved for $\pm\infty$. Thus, the given bit string represents $+\infty$.

   (b) In this case, the floating-point number is

   $$(-1)^0 \times 2^{c-127} \times (1.0)_2$$

   where,

   $$c = (00000001)_2 = 1$$

   so that

   $$c - 127 = -126$$

   Hence, the bit string represents the floating-point number

   $$(1.0)_2 \times 2^{-126} = 2^{-126}$$

3. How many normalized floating-point numbers are available in a binary machine if n bits are allocated to the mantissa and m bits are allocated to the exponent? Assume that two additonal bits are used for signs, as in a 32-bit length computer.

In this case, a typical normalized floating-point number is of the form

$$\pm(1.1\ \underbrace{b_2\, b_3\ \cdots\ b_n}_{n-1 \text{ bits}})_2 \times 2^{\pm k}$$

Thus, we have 2 choices for the sign, 2^{n-1} choices for the mantissa, and 2^m choices for the exponent. Hence, a total of

$$\boxed{2 \times 2^{n-1} \times 2^m = 2^{m+n}}$$

floating-point numbers are available. Note that this includes $\pm\infty$, but excludes ± 0.

4. Show by an example that in computer arithmetic $a + (b + c)$ may differ from $(a + b) + c$.

Consider a 2-decimal machine that properly forms sums before rounding. Then we find that

$$0.22 + (0.92 + 0.44) = 0.22 + 1.4 = 1.6$$
$$(0.22 + 0.92) + 0.44 = 1.1 + 4.4 = 1.5$$