COMP 350 Solutions to Assignment 1

1. (2 points)

   No. If a real number has a finite binary representation, then it can be written as a finite sum of powers of two, i.e.

   $$b_m 2^m + b_{m-1}2^{m-1} + ... + b_0 2^0 + b_{-1}2^{-1} + ... + b_{-n+1}2^{-(n-1)} + b_{-n}2^{-n}$$

   $$b_i = 0 \text{ or } 1, \quad i = -n, -n+1, \ldots, 0, 1, 2, \ldots, m$$

   Obviously, the integral part has a finite decimal representation. Now we look at only the fractional part, which has n terms. Since

   $$2^{-j} = 5^j 10^{-j},$$

   each term in the fractional part has a finite decimal representation. Therefore, the sum of the n terms must also has finite decimal representation.

2. (2 points)

   In the 2's complement representation, change 1 to 0 and 0 to 1, and add 1 to the last bit, giving

   $$\boxed{00000000000000000000000000010111}$$

   Thus the number is

   $$-(2^4 + 2^2 + 2^1 + 2^0) = -23.$$

3. (a) (1 point) exponent bias $= 2^{4-1} - 1 = 7$

   (b) (2 points)

   The smallest nonnegative normalized floating point ♯ in this system

   $$(1.00000)_2 \times 2^{-6}$$

   The largest nonnegative normalized floating point ♯ in this system

   $$(1.11111)_2 \times 2^{7}$$

   (c) (2 points) The smallest nonnegative subnormal floating point ♯ in this system

   $$(0.00001)_2 \times 2^{-6}$$

   The largest nonnegative subnormal floating point ♯ in this system

   $$(0.11111)_2 \times 2^{-6}$$

(d) (1 point)
The machine epsilon of this system

$$\varepsilon = 2^{-5}$$

(e) (2 points)
Two floating point number closest to $11 = (1.01100)_2 \times 2^3$ are

$$11_- = \boxed{0}\boxed{110}\boxed{01011} = (1.01011)_2 \times 2^3$$

$$11_+ = \boxed{0}\boxed{110}\boxed{01101} = (1.01101)_2 \times 2^3$$

(f) (2 points)

i. Round down

$$\text{round}(-(1.0110101)_2) = x_- = -(1.01110)_2$$

ii. Round up

$$\text{round}(-(1.0110101)_2) = x_+ = -(1.01101)_2$$

iii. Round towards zero
Same as round up

iv. Round to nearest
Same as round up

4. (a) (2 points)
True. By IEEE rule,

$$x \oplus x = \text{round}(x + x) = \text{round}(2x).$$

Since the base of the floating point system is 2, $\text{round}(2x) = 2x$.

(b) (2 points)
False. Suppose we use the round down mode. For simplicity let the fractional field have only 2 bits. Then

$$1.11 \ominus 1.00 \times 2^{-4} = \text{round}(1.11 - 0.0001) = \text{round}(1.1011) = 1.10,$$

but

$$-(1.00\times 2^{-4} \ominus 1.11) = -\text{round}(0.0001 - 1.11) = -\text{round}(-1.1011) = -(-1.11) = 1.11.$$

5. (2 points)

$$\infty/0 = \infty, \infty/(-\infty) = \text{NaN}, 1^{\text{NaN}} = \text{NaN}, \ -0/(\text{NaN}) = \text{NaN}.$$