Name: _____ CSCI 237

Practice Problems :: Floating Point

For all questions, please show your work to receive partial credit. Clearly indicate your final answers.

1. **IEEE 754 FP:** For this question, please consider the following **32-bit floating point** representation of a number **x**:

   x = 0100 0001 0100 1000 0000 0000 0000 0000

   *Notes: Recall that for a **normalized** 32-bit floating point number, the **exp** field has 8 bits. Therefore, the bias is $2^8-1=127$.*

   What is the decimal value of **x**?

Our goal is to convert the bit representation of this number into the form: $(-1)^s \times M \times 2^E$

S (the sign bit) is the leftmost bit:

S = 0

E is derived from the **exp** field, which is the next 8 bits:

exp = 100 0001 0

M is derived from the **frac** field, which is the rest of the number:

frac = 100 1000 0000 0000 0000 0000

Since this is a normalized number, our equation for **E** is:

E = exp - bias

E = (2^7 + 2^1) - 127 = 130 - 127 = 3

Also since this is a normalized number, we prepend an implicit 1 when we convert the **frac** field to **M**:

M = 1.100 1000 0000 0000

Thus, our final number is:

$(-1)^0 \times 1.1001 \times 2^3 = 1100.1 = 1 \times (2^3+2^2+2^{-1}) =$ **12.5**

2^0=1, 2^1=2, 2^2=4, 2^3=8, 2^4=16, 2^5=32, 2^6=64, 2^7=128, 2^8=256, 2^9=512, 2^{10}=1024, 2^{11}=2048, 2^{12}=4096, 2^{13}=8192, 2^{14}=16384, 2^{15}=32678, ...

1. For this question, consider all *non-negative* numbers representable using the IEEE 754 Floating point representation from the previous question (32-bits with a 1-bit `sign` field followed by an 8-bit `exp` field, followed by a 23-bit `frac` field).

   What is the bit representation of the **smallest** *normalized* number?

   x = 0000 0000 1000 0000 0000 0000 0000 0000

   What is the value of that number, expressed in the form $v = (-1)^s \times M \times 2^E$?

   The number is positive, so the sign bit must be `0`.
   `S = 0`

   The smallest unsigned number we can represent in the `frac` field is `0`, but in normalized numbers, there is a leading one when we convert the `frac` field to the Mantissa.
   `M = 1.00000000000000000000000`

   In a normalized number, the value of `E` is the unsigned interpretation of the `exp` field minus the bias (2^8-1 = 127). The smallest value we can represent in the `exp` field before it becomes zero (and is therefore a denormalized number) is 1.
   `E = 1-127 = -126`

   $v = (-1)^0 \times 1.0 * 2^{-126}$

   a. What is the bit representation of the **largest** *denormalized* number?

      x = 0000 0000 0111 1111 1111 1111 1111 1111

      What is the value of that number, expressed in the form $v = (-1)^s \times M \times 2^E$?

      The number is positive, so the sign bit must be `0`.
      `S = 0`

      The largest unsigned number we can represent in the `frac` field is all `1`s, but in denormalized numbers, we add a leading zero when we convert the `frac` field to the Mantissa.
      `M = 0.11111111111111111111111`

      In a denormalized number, the value of `E` is always 1-bias = -126
      `E = 1-127 = -126`

      $v = (-1)^0 \times 1.11111111111111111111111 \times 2^{-127}$