CYRR 304 **Name:**
Homework 1, Spring 2024

"*The purpose of computing is insight, not numbers.*" RICHARD HAMMING

Homework 1 has questions 1 through 1 with a total of of 8 points. Your recorded score will be scaled to twenty points.

The point value for each question or part of a question is in the box following each question or part of a question. Neatly write your answers on your own paper, being careful to clearly label each part of each question. Digitize your work as a pdf, and turn it into Canvas. This work is due **Saturday 27 January** at 11:59 PM.

1. An IEEE number of the type binary16 has the form

$$(-1)^s(b_0.b_1b_2\cdots b_{10})_2 \times 2^e,$$

where the sign bit $s \in \{0,1\}$, the exponent $e \in \{-14,-13,\cdots,14,15\}$ and each bit $b_k \in \{0,1\}$. There are ten bits in the fractional part of the number, and just like for binary64 numbers, the leading bit b_0 is not stored and defaults to one.

The exponent range of -14 to 15 contains 30 integers. And just like for a binary64 number, there are two values of the exponent reserved for special conditions, including the number zero, overflow, and the like. That makes for a total of 32 distinct values for the exponent.

[1] (a) How many bits must be used to store the exponent in computer memory? Explain.

Solution: We need enough bits to store 32 numbers, so we need choose the number of bits n in the exponent so that $32 \le 2^n$. The least such integer is 5. It would be peculiar to use more bits than need, so indeed there are five bits for the exponent of a binary16 number.

[1] (b) How many bits must be used to store the entire binary16 number in computer memory? Explain.

Solution: We need one bit for the sign, five bits for the exponent, and ten bits for the fractional part, for a total of sixteen bits.

[1] (c) Find the *largest* number of the type binary16.

Solution: To find the largest binary16 number, we set the exponent to its largest value of 15 and we set each bit in the fractional part to one. Thus the largest binary16 number is

$$(1.1111111111)_2 \times 2^{15} = 2^{15} \sum_{k=0}^{10} \frac{1}{2^k} = 65504.$$

[1] (d) Find the *smallest* positive normalized number of the type binary16. Remember that the leading bit of a normalized number is one.

Solution: To find the smallest positive binary16 number, we set the exponent to its smallest value of -14 and we set each bit in the fractional part to zero. Thus the smallest positive binary16 number

$$(1.0000000000)_2 \times 2^{-14} = \frac{1}{16384}.$$

[1] (e) Find the *smallest* positive denormalized number of the type binary16. Remember that the leading bit of a denormalized number is zero. Also, the exponent of a denormalized binary16 is -14.

Solution: To find the smallest positive denormalized binary16 number, we set the exponent to its smallest value of -14, we set the leading bit to zero, and we set each bit in the fractional part to zero, except for the last bit that that is one. Thus the smallest positive denormalized number of the type binary16 is

$$(0.0000000001)_2 \times 2^{-14} = 2^{-24} = \frac{1}{16777216}.$$

[1] (f) Find the *machine epsilon* ε_m for a number of the type binary16.

Solution: There are ten bits in the fractional part of binary16 number, so

$$\varepsilon_m = \frac{1}{2^{11}} = \frac{1}{2048}.$$

[1] (g) We have $\mathrm{Fl}(1/10) = (-1)^0(1.1001100110)_2 \times 2^{-4}$. Express $\mathrm{Fl}(1/10)$ as a rational number (that is, as a quotient of integers).

Solution: We have

$$\begin{aligned}\mathrm{Fl}(1/10) &= (1+1/2+1/16+1/32+1/256+1/512)\times\frac{1}{16},\\ &= \frac{819}{8192},\\ &\approx 0.0999755859375.\end{aligned}$$

1 (h) Show that $|\mathrm{Fl}(1/10)-1/10|\leq\varepsilon_m$.

Solution: We have

$$\begin{aligned}\left[\left|\mathrm{Fl}(1/10)-1/10\right|\leq\frac{1}{2048}\right] &\equiv \left[\left|\frac{819}{8192}-\frac{1}{10}\right|\leq\frac{1}{2048}\right],\\ &\equiv \left[\frac{1}{40960}\leq\frac{1}{2048}\right],\\ &\equiv \text{True}.\end{aligned}$$