COMP 350 Numerical Computing

Assignment #1: Floating Point Computing

Date Given: Monday, September 11. Date Due: 5pm, Friday, September 22, 2017

(To be marked by Sitao Luan and Yangchao Yi (sitao.luan, yangchao.yi@mail.mcgill.ca))

You can submit either an electronic copy (in PDF format) of your assignment through myCourses or a hard copy, which should be placed in the marked COMP 350 boxes in a cabinet located in Trottier building on the 2nd foor near the elevator.

1. (2 points) Suppose the 2's complement representation of a number is (a 32-bit word is used)

   $$\boxed{11111111111111111111111110101010}$$

   What is the number? Give your answer in decimal representation.

2. (6 points) Suppose in the IEEE single format, the width of the exponent field is 4, not 8, and the width of the fraction field is 4, not 23.

   (a) (1 point) What should the exponent bias be?

   (b) (1 point) What is the largest normalized floating point numbers in this system?

   (c) (1 point) What is the largest subnormal floating point numbers in this system?

   (d) (1 point) What is the machine epsilon of this system.

   (e) (2 point) Given number $a = -(10.01101)_2$. Round it using the four rounding modes. Give the answers as normalized floating point numbers, in the form **binary-significand**$\times 2^E$, where E is decimal.

3. (8 points) The pigeon-hole principle. Suppose we use the IEEE single format.

   (a) (2 points) How many floating point numbers x satisfy $1 \le x < 2$? How many of these satisfy $1 \le x < 3/2$ and how many satisfy $3/2 \le x < 2$?

   (b) (3 points) How many floating point numbers y satisfy $1/2 < y \le 1$? How many of these satisfy $1/2 < x \le 2/3$ and how many satisfy $2/3 < x \le 1$?

   (c) (3 points) Does it follows that there must exist two different floating point numbers x_1 and x_2 between 1 and 2 for which the computed reciprocals $1 \oslash x_1$ and $1 \oslash x_2$ are the same (rounded to the same format)? Is this true regardless of the rounding mode?

4. (2 points) Is it true that $x \ominus y = -(y \ominus x)$ for all rounding modes? Either prove it or give a counterexample.

5. (2 points) What are the values of the expressions $-\infty/0$, $\infty/(-\infty)$, 1^{-NaN}, and $-0/NaN$?