COMP 350 Numerical Computing

Assignment #1: Floating Point Arithmetic

Date Given: Tuesday, September 11. Due date: Thursday, September 20, 2018, 11:59pm

Responsible TAs: Mr. Sitao Luan, Mr. Zhilong Chen (sitao.luan, zhilong.chen@mail.mcgill.ca)

TA office hours: Thursday 4:00pm–5:30pm, Trottier 3110.

Please submit your assignment through myCourses.

1. (2 point) Show that a real number cannot have finite binary representation but infinite (or nonterminating) decimal representation.

2. (2 points) Using a 32-bit word, how many different integers can be represented by (a) sign and modulus; (b) 2's complement? Express the answer using powers of 2.

3. Suppose in IEEE single format, the width of the exponent field is 5, not 8, and the width of the fraction field is 5, not 23.

   (a) (.5 point) What should the exponent bias be?

   (b) (.5 point) What is the machine epsilon of this system?

   (c) (2 points) What are the smallest and largest positive normal floating point numbers in this system?

   (d) (2 points) Can any integer number between the smallest and largest positive normal floating point numbers be stored exactly in this floating point system? Either prove it or give a counterexample.

   (e) (2 points) What are the largest and smallest nonnegative subnormal floating point numbers in this system?

   (f) (1 point) What is the largest floating point number smaller than 2?

   (g) (2 points) Given number $-(10.110101)_2$. Round it using the four rounding modes.

4. Are the following statements true or false? If a statement is true, give a proof and if it's false, give a counter example. We assume no overflow occurs in the calculations and the rounding mode used can be any of the four rounding modes.

   (a) (2 points) If x is a nonzero finite floating point number, then $x \oplus x = 2x$.

   (b) (2 points) If x and y are two finite floating point number, then $x \ominus y = -(y \ominus x)$.

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