CS 357 / MATH 357
Fall 2013

Numerical Methods – Homework 1
Instructor: B. David Semeraro
University of Illinois at Urbana-Champaign

Due: 09/10/2013

Note: Homework is due **5pm** on the due date. Please submit your homework through the dropbox in the Siebel Center basement. Make sure to include your name and **netid** in your homework.

Problem 1 [5pt] Consider IEEE double-precision floating-point arithmetic, using round to nearest. Let a, b, and c be normalized double-precision floating-point numbers, and let $\oplus$ denote correctly rounded floating-point addition.

(a) It is necessarily true that $a \oplus b = b \oplus a$? Explain why or give an example where this does not hold.

(b) It is necessarily true that $(a \oplus b) \oplus c = a \oplus (b \oplus c)$? Explain why or give an example where this does not hold.

Problem 2 [5pt] True/False questions

(a) **(True/False)** $2 + \epsilon_m = 2$?

(b) **(True/False)** IEEE double-precision numbers can represent twice as many numbers than IEEE single-precision numbers.

(c) **(True/False)** The distance between IEEE single-precision numbers are spaced equally.

Problem 3 [10pt] Consider the normalized floating point system where base is 2, precision 4, and the exponent satisfies $-3 \leq e \leq 3$. For each of the following questions give the answer as a binary number. For the binary numbers use the format $\pm(b_0.b_1b_2b_3b_4)_2 \times 2^e$ where $-3 \leq e \leq 3$ is the exponent. For example you would write $(0.5)_{10}$ as $(1.0000)_2 \times 2^{-1}$ in binary in this notation.

(a) What is the value of the machine epsilon ?

(b) What is the next larger number after $(4)_{10}$? Give the answer in base 10 for this question.

(c) What is the largest representable number ?

(d) What is the smallest (in magnitude) representable number if subnormal numbers are *not* allowed ?

(e) What is the smallest (in magnitude) representable number if subnormal numbers *are* allowed ?