CS 370 Fall 2008: Assignment 1

Instructor: Professor Keith Geddes

Lectures: MWF 3:30-4:20 MC 2017

Web Site: UW-ACE

Due: Thu Sep 25, 2008, 5:00 pm, in the Assignment Boxes, 3rd Floor MC

1. Consider the floating point number system $F(2, 5, -10, 10)$ as defined in the course notes. (I.e., base 2, precision 5.)
   (a) What is the smallest positive normalized floating point number?
   (b) What is the largest positive normalized floating point number?
   (c) What is the largest value $a \in F(2, 5, -10, 10)$ such that e^a (equivalently, $\exp(a)$) can be represented (by rounding appropriately) without resulting in overflow? Show your work.

2. Consider a fictitious floating point number system composed of the following numbers:

$$S \;=\{\quad \pm\, d_1.d_2d_3 \times 2^{\pm y} :\; d_2, d_3, y = 0 \text{ or } 1, \\ \text{and } d_1 = 1 \;\text{ unless } d_1 = d_2 = d_3 = 0 \;\}\,.$$

   I.e. each number is normalized unless it is the number zero.
   (a) Plot the elements of S on the real axis. Note, in particular, that successive numbers in S are not always equally spaced.
   (b) Indicate on your plot the regions of OFL (overflow) and UFL (underflow).
   (c) How many elements are contained in S?
   (d) What is the value of ϵ (machine epsilon)?

3. Carry out a roundoff error analysis to show that, in a floating point number system, if $ab+c \neq 0$ then

$$\frac{|(ab + c) - ((a \otimes b) \oplus c)|}{|ab + c|} \leq \frac{|ab|}{|ab + c|}\epsilon\,(1 + \epsilon) + \epsilon$$

   where ϵ denotes machine epsilon. Justify each inequality that you introduce.

4. Carry out (by hand, with the aid of a calculator) the following computations by simulating the 5-significant-digit rounded arithmetic of the floating point number system $F(10, 5, -10, 10)$.

(a) The two roots r_1 and r_2 of the quadratic equation $ax^2+bx+c = 0$ are given by the following well-known formulas:

$$r_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \; ; \quad r_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \; .$$

Calculate the roots r_1 and r_2 using arithmetic in $F(10, 5, -10, 10)$ for the quadratic equation

$$x^2 + 111.11x + 1.2121 = 0 \; . \tag{1}$$

Compare the computed results with the true roots (to 5 significant digits) which you may calculate on a computer using 10 or more digits of precision. Specifically, what is the relative error in r_1 and in r_2?

(b) Note that a *cancellation problem* arises when applying the above formulas for any quadratic equation having the property that

$$|b| \approx \sqrt{b^2 - 4ac} \; .$$

For an equation with this property, if $b > 0$ then the above formula for r_1 will exhibit cancellation and if $b < 0$ then r_2 will exhibit cancellation.
Show that a mathematically equivalent formula for r_1 is

$$r_1 = \frac{2c}{-b - \sqrt{b^2 - 4ac}} \; .$$

Hint: Rationalize the numerator (i.e., multiply numerator and denominator of the original formula for r_1 by an appropriate quantity).

(c) The formula for r_2 can be manipulated in a similar manner. Deduce a better algorithm for calculating the roots of a quadratic equation and present it in the following form.

Algorithm R.

if $b > 0$ then

$$r_2 = (-b - \sqrt{b^2 - 4ac}) \; / \; (2 \; a)$$
$$r_1 = c \; / \; (a \; r_2)$$

else

$r_1 =$

$r_2 =$

(d) Redo the calculation of the roots of equation (1) by applying Algorithm R, using arithmetic in the same number system $F(10, 5, -10, 10)$. Compare the computed results with the true roots. Specifically, what is the relative error in r_1 and in r_2? How do the computed results of part (d) compare with the computed results of part (a)?