HW Assignment 2 (Intro to Numerical Mathematics 3620/5620)

Reading assignment: Chapter 1. The following are problems that will be graded.

1. Consider the double precision arithmetic. (a) What is the largest possible value m_l of an integer m such that 2^m is a machine number? (b) What is the smallest value m_s of m such that 2^m is a machine number? (c) What are the machine numbers immediately to the right and left of 2^m, where m ranges from m_s to m_l? How far is each from 2^m?

2. Generally, when a list of floating-point numbers is added, less roundoff error will occur if the numbers are added in order of increasing magnitude. (a) Give an example to illustrate this principle. (b) However, also show that the principle is not universally valid. In particular, consider a decimal machine with two decimal digits allocated to the mantissa. Show that the four numbers $0.25, 0.0034, 0.00051, 0.061$ can be added with less roundoff error if not added in ascending order.

3. How many machine numbers are there in the single-precision arithmetic?

4. Show by an example that in computer arithmetic $a + (b + c)$ may differ from $(a + b) + c$.

5. Find a good way to compute $\sin x + \cos x - 1$ for x near zero (and explain/justify your suggestion). No need to program this.

6. Program the following pseudocode:

   $x = 1.0$
   while $(1.0 + x) > 1.0$
   $x = x/2.0$
   end
   $y = 2.0 \times x$
   output y

   What is the purpose of this code? (i.e. how can you interpret the result y?) Run the program and print the output. Is the output value consistent with what we talked about in class when we discussed the floating-point representation of numbers?

7. Suppose that we wish to evaluate the function $f(x) = (x - \sin x)/x^3$ for values of x close to zero. (a) Write a routine for this function. Evaluate $f(x)$ a total of 16 times. Namely, initially, let $x = 1$, and then let $x \leftarrow x/10$ fifteen times. Explain the results. Note: L'Hopital's Rule indicates that $f(x)$ should tend to $1/6$. (b) Write a function procedure that produces more accurate values of $f(x)$ for all values of $x \in [-1, 1]$. (submit the printouts of both codes along with the output values, arranged so that one could compare the accuracy of both codes.) Include some discussion comparing the results. In particular, for both methods, make sure to print out the results for all values of x (i.e. for $x = 1, 0.1, 0.01, 0.001, ...$), along with corresponding **absolute errors**[1] (to be able to see how the error changes with varying values of x). Also, if you are going to use a truncated Taylor series expansion to produce more accurate results, include an error analysis of the truncation error, justifying your choice of the number of terms of the truncated series you are using.

[1] i.e. $|f(x) - 1/6|$.