## Please Read the Homework Guidelines Carefully.

1. Let  $x = 0.d_1d_2\cdots d_nd_{n+1}\cdots \beta^b$ ,  $d_1 \neq 0$ ,  $0 \leq d_i \leq \beta - 1$ . We can use the chopping method to express x as a floating number

$$fl_c(x) = 0.d_1 d_2 \cdots d_n \beta^b$$

Find upper bounds of the absolute and relative errors of  $fl_c(x)$  approximating x. Compare the results with the results obtained from the rounding-off approach.

- 2. Let F be a computer number system of 64 bits. Find the following
  - (a) The largest and smallest number.
  - (b) The smallest normalized positive number.
  - (c) The smallest positive number.
  - (d) Give examples of *underflow* and *overflow*.
  - (e) The machine precision.
  - (f) Find upper bounds of the absolute and relative errors of  $fl_c(x)$  approximating x using the rounding approach.

Note that the specifics may differ slightly with different computers and compilers.

3. Assume we use a computer to evaluate the following expressions

(a) 
$$p = xyz$$
, (b)  $s = x + y + z$ ,

where x, y, and z are real numbers. Find upper bounds of absolute and relative errors. Assume all the numbers involved are in the range of the computer number system. Analyze the error bounds.

(**HINT:** You can set  $x_1 = fl(x)$ ,  $y_1 = fl(y)$ ,  $z_1 = fl(z)$ ,  $p_1 = fl(x_1y_1)$ ,  $p_c = fl(p_1z_1)$  is the computed product of x, y, and z. (**Note:** Pay attention to the upper bounds and absolute values, e.g.,  $\delta_5 \leq 5\epsilon$  is wrong, it should be  $|\delta_5| \leq 5\epsilon$ .)

- 4. Design an algorithm (in pseudo-code form) to evaluate the following
  - (a)  $\log(1+x)/x$  in the interval [-0.5, 0.5].
  - (b)  $b \sqrt{b^2 \delta}$ , where b and  $\delta$  are two parameters with  $b^2 \delta \ge 0$ .
  - (c)  $\nabla \phi(x)/|\nabla \phi(x)|$  where  $\phi(x, y)$  is a scalar function of x and y.

You need to consider all possible scenarios.

5. Which of the following two formulas in computing  $\pi$  is better?

$$\pi = 4 \left( 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} + \cdots \right)$$
  
$$\pi = 6 \left( 0.5 + \frac{0.5^3}{2 \cdot 3} + \frac{3(0.5)^5}{2 \cdot 4 \cdot 5} + \frac{3 \cdot 5(0.5)^7}{2 \cdot 4 \cdot 6 \cdot 7} + \cdots \right).$$

How many terms should be chosen such that the error is less than  $10^{-6}$ ? You can write a short Matlab code to compare. Consider both accuracy and speed.

6. We can use the following three formulas to approximate the first derivative of a function f(x) at  $x_0$ .

$$\begin{array}{lll}
f'(x_0) &\approx & \frac{f(x_0+h) - f(x_0)}{h} \\
f'(x_0) &\approx & \frac{f(x_0+h) - f(x_0-h)}{2h} \\
f'(x_0) &\approx & \frac{f(x_0) - f(x_0-h)}{h}
\end{array}$$

When we use computers to find an approximation of a derivative (used in finite difference (FD) method, optimization, and many areas, we need to balance the errors from the algorithm (truncation error) and round-off errors (from computers).

- (a) Which formula is the most accurate in theory? *Hint:* Find the absolute error using the Taylor expansion at  $x = x_0$ :  $f(x_0 \pm h) = f(x_0) \pm f'(x_0)h + f''(x_0)h^2/2 \pm f'''(x_0)h^3/6 + O(h^4)$ .
- (b) Write a program to compute the derivative with
  - $f(x) = x^2, x_0 = 1.8.$
  - $f(x) = e^x \sin x, x_0 = 0.55.$

**Plot** the errors versus h using log-log plot with labels and legends if necessary. In the plot, h should range from 0.1 to the order of machine constant  $(10^{-16})$  with h being cut by half each time (*i.e.*,  $h = 0.1, h = 0.1/2, h = 0.1/2^2, h = 0.1/2^3, \cdots$ , until  $h \le 10^{-16}$ .)

*Hint:* You need to find the true derivative (analytic) values in order to compute and plot the errors.

**Tabulate** the absolute and relative errors corresponding to h = 0.1, 0.1/2, 0.1/4, 0.1/8, and 0.1/16 (that is, difference choices of h compared with that used in the plots). The ratio (should be around 2 or 4) is defined as the quotient of two consecutive errors. **Analyze and explain** your plots and tables. What is the best h for each case with and without round-off errors?

1/h	error (a)	ratio	error (b)	ratio	error (c)	ratio
10		—		—		—
20						
40						
80						
160						

The ratio is defined as, for example

$$ratio = \frac{|\text{error for } n = 10|}{|\text{error for } n = 20|}.$$

7. Mini-project: Choose one of them (a): Find the relation between relative errors and significant digits. (b): Find an example of linear algebra that needs to be solved by computers in your field.