Homework Assignment for Week 01

Goal: Review basic programming skills. Understand source and amplification of errors resulted from floating point arithmetic and how to avoid it, if possible.

1. Section 1.2: Problems 19, 22, 28.

   For your convenience, a scanned copy for section 1.2 problems is available on the course homepage, for this week only.

   In problem 22, use standard double-precision calculation, which is default in most modern languages including matlab (i.e. need not round or chop to 3-digits). In addition of summing to $i = 9$, experiment with summing to $i = 10$, $i = 11$, etc. and check the result. It helps you to explain why.

   If you choose to use matlab, don't forget to turn on 'format long'. Also, you can trust the result of 'exp(-5)' as the true answer.

   Hint for problem 28: Study these examples and estimate $|\frac{x_i - fl(x_i)}{x_i}|$ for each case:

   (a) $k = 4, x_1 = 0.99995, fl(x_1) = 1.0000$

   (b) $k = 4, x_2 = 0.10005, fl(x_2) = 0.1000$

   (c) $k = 4, x_3 = 0.99994999, fl(x_3) = 0.9999$

   (d) $k = 4, x_4 = 0.10004999, fl(x_4) = 0.1000$

2. We showed in class that relative error resulted from multiplication is roughly $3\varepsilon_M$. Derive similar bounds for adding two positive numbers and for division.

3. Find both roots of $x^2 + 10^4 x + 10^{-4} = 0$ as accurate as possible with standard double-precision calculation.

4. Download plotf.m from the course homepage, read it and figure out what it is about.

5. Find the first 3 digits of the mantissa (excluding the leading '1') for the binary expression of 0.8.

6. In half-precision arithmetic, a floating number is stored in 16 bits (half the amount of a single precision floating number). Among which, 1 bit is used to store sign, 5 bits are used to store the exponent. What is the distance between 1.0 and the next larger floating number (the half-precision machine ε_M)?

7. Let $A = \{(x - 100)^2 + y^2 < (40\pi)^2\}$, and $B = \{(x + 100)^2 + y^2 < (100\sqrt{2})^2\}$. Study the 'for' loop, 'if', 'and', 'or' commands in your programming language (preferrably C, C++ or popular ones in the C-family) and find the numbers of grid points (i, j) (i, j are integers) in the sets $A \cup B$ and $A \cap B$, respectively.