Homework 1, due Thursday, February 7.

1. IEEE standard requires the approximate result of an arithmetic operation (for example, addition $a \,\mathrm{plus}\, b$, where plus denotes addition with machine precision is equal to the precise sum of a and b rounded to the nearest floating-point number. Find an example showing that this rule does not imply that the machine addition is associative, that is, numbers a, b, c, such that $(a\mathrm{plus}b)\mathrm{plus}c \neq a\mathrm{plus}(b\mathrm{plus}c)$

2. The series
$$\sum_{n=1}^{\infty} \frac{1}{n}$$
is divergent. However, if we try to compute the partial sums
$$S_m = \sum_{n=1}^{m} \frac{1}{n}$$
for sequential values of m in floating-point arithmetics, at some point, the numbers will stop growing. (An approximation to these partial sums which will give a more accurate result for very large m is $\gamma + \ln m$, where $\gamma = 0.577215664901532866$).

   Determine the value of m for which the partial sums computed using floating point arithmetics stop growing, explain why. Using the approximate formula for partial sums above, determine for what m the error $|S_m - S_m^{approx}|$ exceeds S_m^{approx} where S_m^{approx} is the floating-point approximation.

3. Modify the program used to plot the absolute error of the forward difference $(f(x+h) - f(x))/h$ for the derivative of sin to plot the relative error of the central difference $(f(x+h) - f(x-h))/(2h)$. Plot the *relative errors* and *absolute* errors for the forward and central differences at $x = 1.2$ (the original plot), $x = \pi/2 - 1e-4$, $x = \pi/2 - 1e-10$. What are the best values of h in each of these cases? What is the minimal absolute and relative error?

4. Plot $(2^m + 1) * x - 2^m * x$ for $m = 48 \ldots 54$, $x = 0 \ldots 1$. Explain the behavior of the plots.

5. How many distinct numbers can be represented in a floating-point system following IEEE 754 standard but with only 6 bits in mantissa and 3 bits in the exponent? Count both normalized and unnormalized numbers. What is the largest and smallest magnitude of numbers that can be represented?

6. Do problem 15 from Chapter 2 of the textbook.