Homework 2
Math 166, Fall 2022

Assigned: Friday, September 23, 2022
Due: Friday, September 30, 2022 by 1:30pm on Gradescope

- In your submission, please label all problems (with answers boxed when appropriate), and submit in the order assigned.
- Include printouts of all code (ideally with some comments).

(1) **(Finite precision numbers)** The floating point representation of a real number takes the form $x = \pm(0.d_1 d_2 \dots d_k)_\beta \cdot \beta^e$, where $d_1 \neq 0$, $-m \leq e \leq M$. Suppose that $\beta = 2$, $k = 4$, $m = -5$, and $M = 5$.

   (a) Find the smallest and largest positive numbers that can be represented in this floating point system. Give your answers in decimal form.
   (b) Determine machine precision for this floating point number system.
   (c) Find the floating point number in this system that is closest to $\sqrt{2}$.

(2) **(Size of FPNS)** Consider a general floating point number system $F(\beta, k, m, M)$. Can you determine an expression that gives the number of elements in this number system? Justify your answer.

(3) **(Rounding arithmetic)** Use three-digit, decimal rounding arithmetic (i.e., $\beta = 10$ and $n = 3$) to compute the following sums. Add the numbers by hand in the specified order.

$$\text{(a)} \quad \sum_{k=1}^{6} \frac{1}{3^k} \qquad\qquad \text{(b)} \quad \sum_{k=1}^{6} \frac{1}{3^{7-k}}$$

(4) **(Cancellation Errors)** Near certain values of x, the following functions cannot be accurately computed using the given formula on account of arithmetic cancellations. Identify the values of x where cancellation occurs (e.g., near $x = 0$ or when x is large and positive). Propose a reformulation that removes the problem (e.g., using Taylor series, rationalization, trigonometric identities, etc.).

   (a) $f(x) = 1 + \cos x$
   (b) $f(x) = \mathrm{e}^{-x} + \sin x - 1$
   (c) $f(x) = \ln x - \ln(1/x)$
   (d) $f(x) = \sqrt{x^2+1} - \sqrt{x^2+4}$

(5) **(Cancellation Errors from a class example)** Consider the function $f(x) = x - \sin(x)$. In class, we reformulate this function near $x = 0$ by use of Taylor series to be

$$f(x) \approx x^3/3! - x^5/5! \ .$$

Let $g(x) = x^3/3! - x^5/5!$. Working in double precision, plot the two functions $f(x)$ and $g(x)$:
   (a) on the interval $[-5 \times 10^{-5}, 5 \times 10^{-5}]$ using 1000 uniformly spaced points.
   (b) on the interval $[-5 \times 10^{-8}, 5 \times 10^{-8}]$ using 1000 uniformly spaced points.
   (c) on the interval $[-4, 4]$ using 1000 uniformly spaced points.

Explain how cancellation error plays a role in your results.