Math 125 **Problem Set 1** **Due: Monday 9/19 at 11:59 p.m. EST**

Instruction: Read the assignment policy. For problem 5(c), include a printout your code with your homework submission. You should submit your assignment on Gradescope.

1. For each of the following expressions, indicate for what values catastrophic cancellations could occur. Find an alternative expression that avoids the cancellation.

(a) $f(x) = \dfrac{\sqrt{25+x}-5}{x}$. Show that the alternative expression correctly computes $\lim_{x\to 0} f(x)$.

(b) $f(x) = \dfrac{1-\cos(x)}{x^2}$.

(c) $f(x) = \dfrac{1-\sec(x)}{\tan^2(x)}$.

2. For each of the following expressions, indicate whether underflow, overflow or both could occur. Find an alternative expression that avoids these issues.

(a) Computing the determinant of a matrix $\mathbf{A} \in \mathcal{R}^{n\times n}$.
Remark: Any matrix $\mathbf{A}$ has the following decomposition $\mathbf{A} = \mathbf{LU}$ where $\mathbf{L}$ is lower-triangular and $\mathbf{U}$ is upper-triangular. This is known as the LU decomposition which we cover later in the course. With that, $\det(\mathbf{A}) = \det(\mathbf{L})\det(\mathbf{U})$. Since $\mathbf{L}$ and $\mathbf{U}$ are triangular matrices, the determinants are simply the product of the diagonal entries. Given this, without loss of generality, you can consider computing the determinant of an upper triangular matrix $\mathbf{A}$ and determine when underflow/overflow occurs.

(b) Computing the length of a vector $\mathbf{x} \in \mathcal{R}^n$ using $||\mathbf{x}|| = \sqrt{x_1^2 + x_2^2 + ... + x_n^2}$.

3. This question concerns floating point arithmetic.

(a) The binary representation of a certain number x is 101110. What is x in base 10?

(b) Find the binary representations of 0.375 and 1.25. Consider the sum of the two numbers in their binary representations and show that it agrees with the sum in the standard base 10 representation.

(c) Recall the representation of a number in single precision: $(-1)^s \cdot 2^{e-127} \cdot (1+f)$ where s denotes the 1-bit sign, e denotes the 8-bit exponent and f denotes the 23-bit fraction. Assume that an exponent of all zeros and an exponent of all ones are special and reserved. What is the smallest number that can be represented in single precision? What is the largest number that can be represented in single precision?

(d) Give an example for a number that does not have an exact representation either in single or double precision arithmetic.

(e) Prove that the floating numbers in the range $[2^k, 2^{k+1}]$ are just the numbers in $[1, 2]$ multiplied by 2^k. What does this inform you about the gaps between small numbers as opposed to gaps between large numbers?

4. Consider the normalized representation of a number x in some base B as follows

$$x = \pm B^e \sum_{k=m}^{k=0} d_k B^{-k} \quad d_0 = 1, d_k \in \{0, 1, 2, ..., B-1\} \text{ for } k > 0$$

In the above representation, m denotes the precision. We now consider the case where a number y may not have a finite representation.

$$y = \pm B^e \sum_{k=\infty}^{k=0} d_k B^{-k} \quad d_0 = 1, d_k \in \{0, 1, 2, ..., B-1\} \text{ for } k > 0$$

To obtain a finite representation, we truncate all digits d_{m+1} forward. The rounded representation of y is as follws

$$\text{fl}(y) = \pm B^e \sum_{k=m}^{k=0} d_k B^{-k} \quad d_0 = 1, d_k \in \{0, 1, 2, ..., B-1\} \text{ for } k > 0$$

Prove that $\dfrac{|y - \text{fl}(y)|}{|y|} \leq B^{-m}$.

5. This question concerns polynomial evaluation.

(a) Find an efficient method for evaluating the polynomial $P(x) = 3x^5 + 6x^8 - 2x^{11} + 9x^{14}$. [*Hint*: First do some re-arranging before applying Horner's scheme].

(b) Consider the polynomial $p(x) \equiv a_0 + a_1 x + a_2 x^2 + ... + a_k x^k$. For fixed $x_0 \in \mathcal{R}$, define $c_0, ..., c_k \in \mathcal{R}$ recursively as follows:

$$c_k \equiv a_k$$
$$c_i \equiv a_i + c_{i+1} x_0 \quad \forall i < k.$$

Show $c_0 = p(x_0)$. How many operations are needed to compute $p(x_0)$?

(c) Implement Horner's method and test your function on the following polynomial: $p(x) = 7x^3 - 11x^2 + 12x + 5$ evaluated at $x = 100$.

(d) Consider the polynomial $p(x) = (x+1)^8$. What is the most efficient way to compute $p(x)$? Does this contradict the optimality of Horner's scheme? Explain.

6. For each of the following problems, state whether or not the computation is backward stable. Assume that the computation is done with floating point addition $\oplus$, floating point subtraction $\ominus$ and floating point multiplication $\otimes$. Justify your answer. [*Hint*: Use the definition of backward stability and the fundamental axiom of floating point arithmetic].

(a) Substraction of two numbers: $\tilde{f}(x, y) = \text{fl}(x) \ominus \text{fl}(y)$.

(b) Outer product of two vectors: $\tilde{f}(\mathbf{x}, \mathbf{y}) = \mathbf{x}\mathbf{y}^T$.

7. Extra Credit: During the Gulf War, an American Patriot Missile battery in Saudi Arabia, failed to intercept an incoming Iraqi Scud missile. The report of the software problem that led to the system failure can be found here `https://www.gao.gov/assets/imtec-92-26.pdf`. The main problem was attributed to an internal clock which recorded time in tenths of a second which is then stored as an integer i.e. if the clock reads 10, it means $10 \times (1/10\text{sec}) = 1\text{sec}$ has elapsed. The calculations were done using a 24-bit arithmetic.

(a) The 24-bit binary approximation of 1/10 is 0.000110011001100110011001100. Convert this binary number to a base 10 representation. Let y denote the result. What is $|y - \frac{1}{10}|$?

(b) What is the time error, in units of seconds, after 100 hours of operation?

(c) The Scud missile traveled at approximately 3750 miles per hour. Find the distance that a Scud missile would travel during the time error.