CS 231 – Homework #1 – Due March 16, 2022 at the beginning of class

This assignment consists of questions regarding floating-point representations. For each question, it is not sufficient to give just a numerical answer. Explain your reasoning carefully and thoroughly.

1. In class, we saw that the largest single-precision real number is $2^{128} - 2^{104}$. Use the same reasoning to determine the largest double-precision real number. Assume that the exponent occupies 11 bits, and its bias is 1023.

2. We also saw that the smallest positive number that can be represented in single precision is 2^{-149}. Use the same reasoning to derive the value of the smallest positive number in double precision.

For questions 3-5, suppose you write a program that adds 1 plus 1/2 plus 1/3 + 1/4 + 1/5 and so on, and stops when the sum no longer changes. This program would use two floating-point variables: one that keeps track of the sum, and another to keep track of the next number to add in the series. For questions 3 and 4, you might find it useful to write a short program (in any language) to experiment with this summation, but you do not need to turn in this program. You may also find the following approximation useful. Note that γ (gamma) is a transcendental number near 0.5772.

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \cdots + \frac{1}{n} = \gamma + \ln(n)$$

3. What will the final sum be if the floating-point values are represented as single-precision numbers?

4. What will the final sum be if the floating-point values are represented as double-precision numbers?

5. If there is no limit to how much space is used to store the floating-point numbers, the program will never terminate. How long will it take for the sum to exceed 100.0? Assume that we can add 4 million terms of the series each second.