Numerical Analysis Assignment 1 — IEEE Arithmetic

Question #1

Determine the maximum and minimum number that can be stored in:

(a) A 5-byte unsigned integer

(b) A 5-byte signed integer

Question #2

Consider a real number stored with 5 bytes. Bit #1 is reserved for the sign, while bits #2 to #10 are reserved for the biased exponent, and bits #11 to #40 are related to the significand. Do the following:

(a) Find the minimum and maximum possible exponent p

(b) Find the smallest possible positive number

(c) Find the largest possible number

(d) Find the smallest possible positive subnormal number

Question #3

Consider a real number stored with 6 bytes. Bit #1 is reserved for the sign, while bits #2 to #13 are reserved for the biased exponent, and bits #14 to #48 are related to the significand. Find the machine precision ϵ_{mach}.

Question #4

Say that

$$x = -g + \sqrt{g^2 + 1}$$

Say that both x and g are stored in memory using single precision numbers with a relative error due to machine accuracy of $\epsilon_{\text{mach}} = 2 \times 10^{-10}$. Do the following:

(a) Find the relative error on x given $g = 1000.0$

(b) Recast the equation for x in difference form to reduce its relative error

(c) Find the relative error on x given $g = 1000.0$ for the recast equation outlined in (b)

Question #5

It is desired to minimize the number of bits that can store a certain range of numbers. The range lower limit is 3×10^{-65}, and the range upper limit is 10^{32}. Find the number of bits needed to store the exponent and the significand. Then find the total number of bits needed.

Question #6

(a) Consider a number of real type. Knowing that the machine accuracy (non-denormal) is of $\epsilon_{\text{mach}} = 9.5367 \times 10^{-7}$ and that the maximum positive number must be at least as high as 10^{23}, do the following:
(i) find the minimum number of bits for the exponent;
(ii) find the minimum number of bits for the significand.

(b) Consider the number 9.5367×10^{-4} stored in memory as a real type. Knowing that the exponent of the real type has 4 bits what is the minimum number of bits that the significand should have if the relative error on the number is less than 0.01?

Answers

2. $255, -254, 3.454 \times 10^{-77}, 1.15792089 \times 10^{77}, 3.217 \times 10^{-86}$.
5. 9, 1, 11.
6. 19, 8, 11.

Due on Wednesday September 19th at 16:30. Do questions #2, #5, #6 only.