LSP – Challenge 1

Mietze Tang (312562)

[0.] What is the machine precision for the U-96 arithmetic?

The machine precision has different definitions, one of them is given by

$$\frac{1}{2} * \beta^{1-t} = \frac{1}{2} * 2^{-6} = 2^{-7}$$

[1.] How many different numbers are in S?

Representation for e

Let k be the number of bits in the representation of e that are 1. Then there are $k * (10 - k)$ possibilities to switch a bit that is 1 with a bit that is 0.

$$\Rightarrow |S| = k * (10 - k)$$

So we need to find the representation for e in the U-96 arithmetic first:
(1) $+1.010111 * 2^1 = 1 + \frac{1}{4} + \frac{1}{16} + \frac{1}{32} + \frac{1}{64} = 2.71875 > e$
$\rightsquigarrow$ consider next smaller number:
(2) $+1.010110 * 2^1 = 1 + \frac{1}{4} + \frac{1}{16} + \frac{1}{32} = 2.6875$

$|e - 2.71875| \approx 0.00046817$
$|e - 2.6875| \approx 0.0307818$

So, the representation in (1) is the closest.

$$\Rightarrow e_{U96} = 1.010111 * 2^1 = 2.71875$$

The exponent is 1, so its binary representation is 100 (= 4, mapped to 1) Under the assumption that the +-sign is represented by a 0, the representation of e contains 5 zeros and 5 ones (sign = 0, mantissa = 010111, exponent = 100).

$$\Rightarrow |S| = 5 * 5 = 25$$

So, S contains 25 different numbers.

[2.] What is the largest number in S?

Considering $e_{U96} = +1.010111 * 2^{100}$, we will probably achieve the largest number by increasing the exponent and switching that bit with the least significant bit in the mantissa. So we get

$$s_{max} = +1.010110 * 2^{110} = (1 + \frac{1}{4} + \frac{1}{16} + \frac{1}{32}) * 2^3 = 10.75$$

[3.] What is the smallest number in S?

For the smallest number, we can switch the bit for the sign with another bit to get a negative number. As we want the rest of the number to be big, we should switch it with the least significant bit in the mantissa which is $\neq 0$. Then we get:

$$s_{min} = -1.010110 * 2^{100} = -(1 + \frac{1}{4} + \frac{1}{16} + \frac{1}{32}) * 2^1 = -2.6875$$

[4.] Among the numbers in S, what is the best representation for e? What is the corresponding relative error?

Changing bits in the exponent or the sign would result in a large error, so we will probably achieve the smallest error by having the switched bits in the mantissa:
The smallest 0-bit that we can switch is d_3. As this adds $\frac{1}{4}$ to the number, we want to switch a bit that is close to it: Switching d_3 and d_4 will result in the smallest error:

$$s_1 = +1.011011 * 2^1 = (1 + \frac{1}{4} + \frac{1}{8} + \frac{1}{32} + \frac{1}{64}) * 2 = 2.84375$$

$$\Rightarrow Err_{rel} = \frac{|e_{U96} - 2.84375|}{|e_{U96}|} = \frac{|2.71875 - 2.84375|}{|2.71875|} = \frac{4}{87} \approx 4.6\%$$

So s_1 is the best representation for e in S with a relative Error of approx. 4.6%.

[5.] Among the numbers in S, which two are the closest to e?

The first number is s_1 from [4.]. The second one is the result of switching d_3 and d_5:

$$s_2 = 1.011101 * 2^1 = (1 + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{64}) * 2 = 2.90625$$