Introduction to Scientific Computing Languages

Exercises session

Floating point arithmetic — November 4th, 2014

- **[Q1]** Consider the IEEE settings for single precision arithmetic:

$$\beta = 2, \quad t = 24, \quad e_{\min} = -126, \quad e_{\max} = 127$$

  1. What is the smallest floating point number larger than 2?

  2. What is the largest floating point number smaller than 8?

  3. How many floating points in $[1/64, 1/32]$?

  4. Distance between 65,536 and the next floating point number?

  5. What is the first positive integer that cannot be represented exactly?

- **[Q2]** Consider the following ternary arithmetic with normalization:

$$\beta = 3, \quad t = 3, \quad e_{\min} = -2, \quad e_{\max} = 3$$

  (The leading ternary digit of the mantissa must be 1 or 2.)

  1. How is $\pi = 3.1415926536$ represented?

     What is the representation error?

  2. What is the largest floating point number?

  3. First 5 positive integers that cannot be represented?

  4. How does the answer to 1. change if $t = 8$?

5. How does the answer to 3. change if $t = 8$?

- **[Q3]** Consider the following binary arithmetic with normalization:

$$\beta = 2, \quad t = 4, \quad e_{\min} = -2, \quad e_{\max} = 4$$

1. How is π represented?

   What is the representation error?

2. Smallest absolute distance between 2 floating point numbers?

3. Smallest relative distance between 2 floating point numbers?