Introduction to Scientific Computing Languages
Practice questions

Prof. **Paolo Bientinesi**

pauldj@aices.rwth-aachen.de

Floating Point Arithmetic

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

$$\beta = 2, \quad t = 24, \quad e_{\min} = -125, \quad e_{\max} = 128$$

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 point numbers are in the interval [1/64, 1/32] ?
4. What is the distance between 65536 and the next floating point number?
5. What is the first integer that cannot be represented exactly?

More on Floating Point Arithmetic

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

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

1. How is π represented? What is the representation error?
2. What is the largest floating point number?
3. What are the first 5 positive integers that cannot be represented exactly?

- **[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. What is the smallest absolute distance between two floating point numbers*?
3. What is the smallest relative distance between two floating point numbers*?

*: the arithmetic is normalized. What if this is not the case?