ACCURACY AND SPEED

i) Floating-point representation

a) In ~~plate~~ python: IEEE 754 double precision representation; binary, 64 bits

$$x = (-1)^{sign} \left(1.\, b_{51}\, b_{50} \ldots b_0\right) \times 2^{e-1023}$$

(fraction; exponent)

Equivalently

$$x = (-1)^{sign} \left(1 + \sum_{i=1}^{52} b_{52-i} 2^{-i}\right) \times 2^{e-1023}$$

where:

a) Sign is either 0 or 1 (1 bit)

b) Fraction is the fraction part of a binary number (written in the form 1. fraction) and it is represented with 52 bits; the "1." in front is not explicitly represented to save space

c) The exponent [e] is an 11 bit unsigned integer. The exponent is written in binary representation and can therefore take all integer values between 0 and 2047. Note that the actual final exponent is:

$$\boxed{e - bias} \Rightarrow \text{actual exponent}$$

where, in double precision, we set $\boxed{bias = 1023}$. Therefore we can represent, in principle, numbers that range from ~~com~~ →

=> from 2^{-1023} to 2^{+1024}. The actual range is instead from 2^{-1022} to 2^{+1023}. The reason is that the exponent with all zeros is used to represent "exact 0" (it would be 2^{-1023} otherwise), whereas the exponent with all ones (that would be 2^{1024}) is used to define "infinity"

•) Example. Consider the number $x = 85.125$ and write it in IEEE 754, 64 bit representation:

a) Binary $\begin{cases} 85 = 1010101 \\ 0.125 = 0.001 \end{cases}$ => $85.125 = 1010101.001$

b) Need to write as 1. fraction =>

=> $1010101.001 = 1.\boxed{010101001} \times 2^{6}$ -> unbiased exponent

fraction

c) To get the biased exponent just remember $6 = e - 1023$

Hence $e = 1029$. Need to write e in base 2

$$\boxed{e = 1029 = 10000000101}$$

d) The number is positive, hence $\boxed{\text{sign} = 0}$

Therefore we get the double precision representation of 85.125 as:

85.125 -> 0 10000000101 0101010010.....0

43 zeros

•) The two main things we learn are:

a) The double precision representation allows to store numbers between $\simeq 10^{-308}$ and $\simeq 10^{308}$

b) The number is rounded off with a precision of 2^{-53}, corresponding to $\simeq 16$ decimal digits