Lecture 2 Computer Arithmetic & Round-off error

binary seq in computer $= \underbrace{\overbrace{\boxed{\ }}^{1}}_{s} \ \underbrace{\overbrace{\boxed{\qquad\qquad}}^{11}}_{c} \ \underbrace{\overbrace{\boxed{\qquad\qquad\qquad\qquad}}^{52}}_{f}$

number that binary seq represents $\Big\}$ $(-1)^s \underbrace{2^{c-1023}}_{\text{sets the "scale"}} \underbrace{(1+f)}_{\text{sets "precision"}}$.

Counting in binary.

10	9	8	⋯	3	2	1	0		
0	0	0	⋯	0	0	0	0	$= 0 =$	$0 \cdot 2^0$
0	0	0	⋯	0	0	0	1	$= 1$	$4 \cdot 2^1$
0	0	0	⋯	0	0	1	0	$= 2 =$	$2^1 + 0 \cdot 2^0$
0	0	0	⋯	0	0	1	1	$= 3$	$2^1 + 1 \cdot 2^0$
0	0	0	⋯	0	1	0	0	$= 4 =$	$2^2 + 0 \cdot 2^1 + 0 \cdot 2^0$
0	0	0	⋯	0	1	0	1	$= 5$	$2^2 + 0 \cdot 2^1 + 1 \cdot 2^0$
0	0	0	⋯	0	1	1	0	$= 6$	$2^2 + 1 \cdot 2^1 + 0 \cdot 2^0$
0	0	0	⋯	0	1	1	1	$= 7$	$2^2 + 1 \cdot 2^1 + 1 \cdot 2^0$
0	0	0	⋯	1	0	0	0	$= 8 =$	$2^3 + 0 \cdot 2^2 + 0 \cdot 2^1 + 0 \cdot 2^0$

$\Rightarrow \max(c) = 2^{11} - 1 = 2047$

$\Rightarrow \max(c-1023) = 1024 \Rightarrow \max(2^{c-1023}) \gg 1$

$\min(c-1023) = -1023 \Rightarrow \min(2^{c-1023}) \ll 1$

$$\min(f) = 0\left(\tfrac{1}{2}\right)^1 + 0\left(\tfrac{1}{2}\right)^2 + \cdots + 0\left(\tfrac{1}{2}\right)^{52} = 0.$$

$$\max(f) = 1\cdot\left(\tfrac{1}{2}\right)^1 + 1\cdot\left(\tfrac{1}{2}\right)^2 + \cdots + 1\cdot\left(\tfrac{1}{2}\right)^{52} = 1 - \left(\tfrac{1}{2}\right)^{52}$$

Example:

$$\frac{1}{10} = \left(\tfrac{1}{2}\right)^4 + \left(\tfrac{1}{2}\right)^5 + \left(\tfrac{1}{2}\right)^8 + \cdots$$

ie. $\underset{\substack{\uparrow \\ \text{base} \\ 10}}{0.1} \approx \underset{\substack{\uparrow \\ \text{base} \\ 2}}{0.00011001\cdots}$