1.7 Representation of numbers: Floating point numbers

Consider a decimal number $x \neq 0$. It can be written as follows:

$$\begin{aligned} x = S \times \bar{x} \times 10^E \quad &\text{— normalized form of } x, \text{ with} \\ S = \pm 1 \quad &\text{— sign,} \\ \bar{x} \quad &\text{— mantissa (real number), } 1 \leq \bar{x} < 10, \\ E \quad &\text{— exponent (signed integer number).} \end{aligned} \tag{1.7}$$

Examples:

$$\begin{aligned} x &= -0.031_{10} = -3.1 \times 10^{-2} \;\rightarrow\; S = -1,\ \bar{x} = 3.1,\ E = -2; \\ x &= 185.79_{10} = 1.8579 \times 10^{2} \;\rightarrow\; S = 1,\ \bar{x} = 1.8579,\ E = 2. \end{aligned} \tag{1.8}$$

Analogously, consider $x \neq 0$ in the binary system (note the different range of $\bar{x}$):

$$\begin{aligned} x = S \times \bar{x} \times 2^E \quad &\text{— normalized form of } x, \text{ with} \\ S = \pm 1 \quad &\text{— sign,} \\ \bar{x} \quad &\text{— mantissa (real number), } 1 \leq \bar{x} < 2, \\ E \quad &\text{— exponent (signed integer number).} \end{aligned} \tag{1.9}$$

Examples:

$$\begin{aligned} x &= 100.011_2 = 1.00011 \times 2^2 \;\rightarrow\; S = 1,\ \bar{x} = 1.00011,\ E = 2; \\ x &= -0.00101_2 = -1.01 \times 2^{-3} \;\rightarrow\; S = -1,\ \bar{x} = 1.01,\ E = -3. \end{aligned} \tag{1.10}$$