Introduction to Verified Numerical Computation (Verified Computing)

TANAKA, Kazuaki
田中 一成
tanaka@ims.sci.waseda.ac.jp
Institute for Mathematical Science, Waseda University
LAB: Building No.60, Room 307B

Floating-Point Numbers

Generally, a nonzero real number can be represented in the following form:

$$\pm\left(\frac{d_0}{\beta^0}+\frac{d_1}{\beta^1}+\frac{d_2}{\beta^2}+\cdots\right)\cdot\beta^e$$

where $\beta \geq 2$, $0 \leq d_i \leq \beta - 1$, and e is an integer.

Example)

$$7.375 = +\left(\frac{7}{10^0}+\frac{3}{10^1}+\frac{7}{10^2}+\frac{5}{10^3}\right)\cdot 10^0 \quad \underbrace{(\beta = 10)}_{\text{Base}}$$

$$7.375 = +\left(\frac{1}{2^0}+\frac{1}{2^1}+\frac{1}{2^2}+\frac{0}{2^3}+\frac{1}{2^4}+\frac{1}{2^5}\right)\cdot 2^2 \quad (\beta = 2)$$

Example)

$$0.2 = +\left(\frac{2}{10^0}\right)\cdot 10^{-1} \quad (\beta = 10)$$

$$0.2 = +\underbrace{\left(\frac{1}{2^0}+\frac{1}{2^1}+\frac{0}{2^2}+\frac{0}{2^3}+\frac{1}{2^4}+\frac{1}{2^5}+\frac{0}{2^6}+\frac{0}{2^7}+\cdots\right)}_{\text{infinite}}\cdot 2^{-3} \quad (\beta = 2)$$

$$\pi = +\left(\frac{3}{10^0}+\frac{1}{10^1}+\frac{4}{10^2}+\frac{1}{10^3}+\frac{5}{10^4}+\frac{9}{10^5}+\cdots\right)\cdot 10^0 (\beta = 10)$$

infinite

$$\pi = +\left(\frac{1}{2^0}+\frac{1}{2^1}+\frac{0}{2^2}+\frac{0}{2^3}+\frac{1}{2^4}+\frac{0}{2^5}+\frac{0}{2^6}+\frac{1}{2^7}+\cdots\right)\cdot 2^1$$

infinite $(\beta = 2)$

Floating-Point Number in computers

Since computers cannot possess infinite digits, in them, numbers are represented in the following (finite) form:

$$\pm \underbrace{\left(\frac{d_0}{\beta^0} + \frac{d_1}{\beta^1} + \frac{d_2}{\beta^2} + \cdots + \frac{d_{p-1}}{\beta^{p-1}}\right)}_{\text{“significant”}} \cdot \beta^e$$

β: “base”

e: “exponent”, $\mathrm{E_{min}} \leq e \leq \mathrm{E_{max}}$

p: “precision”

significant digits

IEEE 754 binary64 (double)

Most computers employ, IEEE754 standard binary64 (so called double) , i.e., $\beta = 2$, $p = 53$, $\mathrm{E}_{\min} = -1022$, $\mathrm{E}_{\max} = 1023$.

$$\pm \underbrace{\left(\frac{d_0}{2^0} + \frac{d_1}{2^1} + \frac{d_2}{2^2} + \cdots + \frac{d_{52}}{2^{52}} \right)}_{53} \cdot 2^e \qquad (-1022 \leq e \leq 1023)$$

It is called “normalized” when $d_0 = 1$.
Normalized numbers 52 significant digits.

It is called “denormalized” when $d_0 = 0$.
The significant digits of denormalized numbers are less than 52.

Maximal and minimal numbers

$$M = \underbrace{\left(\frac{1}{2^0} + \frac{1}{2^1} + \cdots + \frac{1}{2^{52}}\right)}_{= \frac{a(1-r^n)}{1-r}} \cdot 2^{1023}$$

$\downarrow$ Inf (overflow)

$$m_n = \left(\frac{1}{2^0} + \frac{0}{2^1} + \cdots + \frac{0}{2^{52}}\right) \cdot 2^{-1022} = 2^{-1022}$$

$$m_d = \left(\frac{0}{2^0} + \frac{0}{2^1} + \cdots + \frac{1}{2^{52}}\right) \cdot 2^{-1022} = 2^{-1074}$$

maxmin_double.cc

Rounding

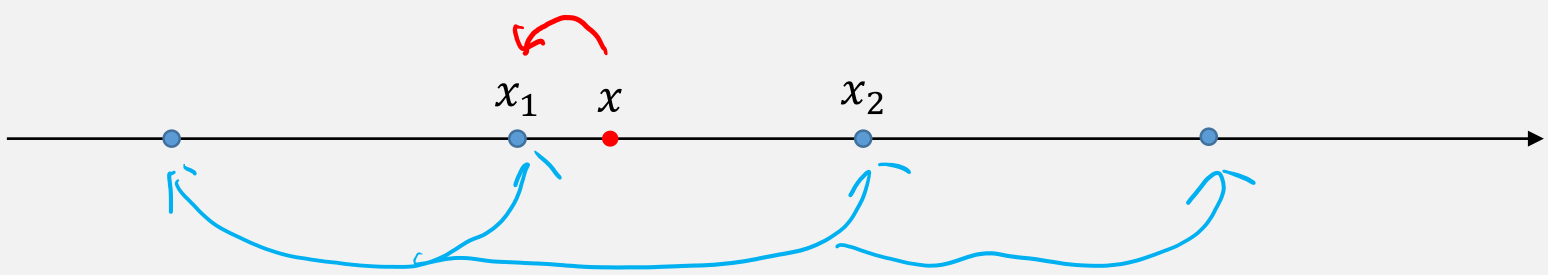
$\mathbb{F}$: the set of floating-point numbers (here binary 64).

For simplicity, we consider a normalized number x satisfying $m \le x \le M$.

Round Nearest

When $x \notin \mathbb{F}$, computers round off x to $\mathrm{RN}(x) \in \mathbb{F}$.

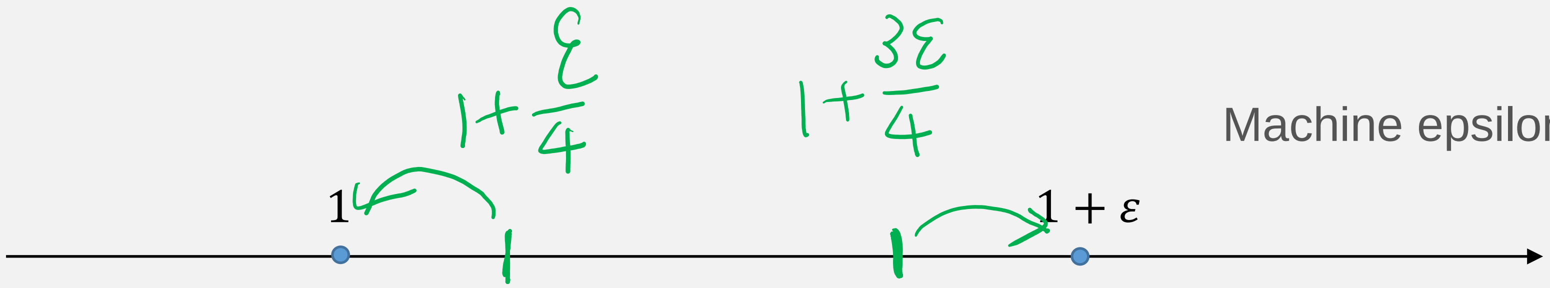
$$|x - \mathrm{RN}(x)| = \min_{y \in \mathbb{F}} |x - y|$$



Floating point numbers

Observe rounding in computers

Machine epsilon



$$1 = \left(\frac{1}{2^0} + \frac{0}{2^1} + \frac{0}{2^2} + \cdots + \frac{0}{2^{52}}\right) \cdot 2^0$$

$$\varepsilon = 2^{-52}$$

observation_rounding_to_nearest.cc
test_rounding.cc