$\mathbb{R}$ set of real numbers.

$\mathbb{F} \subseteq \mathbb{R}$

floating point numbers...

These are numbers of the form

$$\pm (1+f) \times 2^n \quad \text{where} \quad f = \sum_{i=1}^{d} b_i 2^{-i}, \qquad b_i \in \{0,1\}$$

precision d, exponent n

Example: $d = 2$ and $n = 0$

two terms in sum

$b_1 = 0$, $b_2 = 0$ $\quad$ $b_1 = 0$, $b_2 = 1$ $\quad$ $b_1 = 1$, $b_2 = 0$ $\quad$ $b_1 = 1$, $b_2 = 1$

$$(1+0)\times 2^0, \quad (1+\tfrac{1}{4})\times 2^0, \quad (1+\tfrac{1}{2})\times 2^0, \quad (1+\tfrac{3}{4})\times 2^0$$

$$\{1, \quad 1+\tfrac{1}{4}, \quad 1+\tfrac{1}{2}, \quad 1+\tfrac{3}{4}\} = \mathbb{F}_2 \cap [1,2)$$

spacing between is $\frac{1}{4} = \frac{2^n}{2^d} = 2^{n-d}$

Example: $d = 2$ and $n = 1$

$b_1 = 0$, $b_2 = 0$ $\quad$ $b_1 = 0$, $b_2 = 1$ $\quad$ $b_1 = 1$, $b_2 = 0$ $\quad$ $b_1 = 1$, $b_2 = 1$

$$(1+0)\times 2^1, \quad (1+\tfrac{1}{4})\times 2^1, \quad (1+\tfrac{1}{2})\times 2^1, \quad (1+\tfrac{3}{4})\times 2^1$$

$$\{2, \quad 2+\tfrac{1}{2}, \quad 2+1, \quad 2+\tfrac{3}{2}\} = \mathbb{F}_2 \cap [2,4)$$

spacing $= \frac{1}{2} = \frac{2^1}{2^2}$

Example: $d = 2$ and $n = -1$

$b_1 = 0$, $b_2 = 0$ $\quad$ $b_1 = 0$, $b_2 = 1$ $\quad$ $b_1 = 1$, $b_2 = 0$ $\quad$ $b_1 = 1$, $b_2 = 1$

$$(1+0)\times 2^{-1}, \quad (1+\tfrac{1}{4})\times 2^{-1}, \quad (1+\tfrac{1}{2})\times 2^{-1}, \quad (1+\tfrac{3}{4})\times 2^{-1}$$

$$\{\tfrac{1}{2}, \quad \tfrac{1}{2}+\tfrac{1}{8}, \quad \tfrac{1}{2}+\tfrac{1}{4}, \quad \tfrac{1}{2}+\tfrac{3}{8}\} = \mathbb{F}_2 \cap [\tfrac{1}{2}, 1)$$

spacing $= \frac{1}{8} = 2^{-1}/2^2$

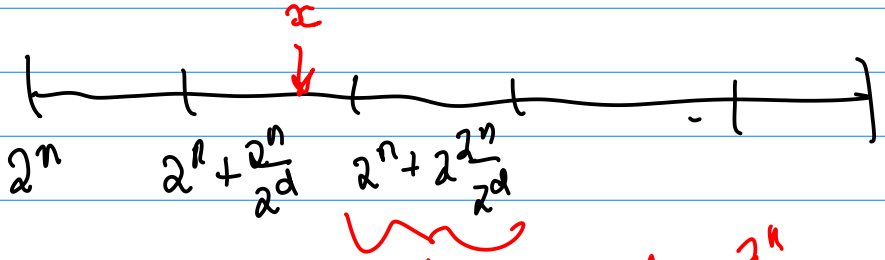
Define Machine epsilon:

Take the smallest element $x \in \mathbb{F}$ that is larger than 1. Then $\varepsilon_{mach} = x - 1$.

For $\mathbb{F}_2$ then $\varepsilon_{mach} = \frac{1}{4} = \frac{1}{2^d}$ when $d = 2$

Define: Rounding operation $fl : \mathbb{R} \to \mathbb{F}$ such that $fl(x)$ is the nearest floating point number to x.

Suppose $x \in [2^n, 2^{n+1})$ the numbers in $\mathbb{F}$ inside this interval are spaced $\frac{2^n}{2^d}$ apart



$$fl(x) = 2^n + 2\frac{2^n}{2^d}$$

$$|fl(x) - x| \leq \frac{1}{2}\frac{2^n}{2^d} = 2^{n-d-1} \quad \text{and} \quad 2^n \leq |x|$$

$$\frac{|fl(x) - x|}{|x|} \leq \frac{2^{n-d-1}}{2^n} = 2^{-d-1} = \frac{1}{2}2^{-d} = \frac{1}{2}\varepsilon_{mach}$$

only depends on d

Relative error in the floating point representation of any real number is uniformly bounded by $\frac{1}{2}\varepsilon_{mach}$.