Floating point numbers.

Note Title 23-04-2021

The fact that infinitely many real nos. have to be stored in finite amount of space gives rise to 2 limitations :

1. the represented numbers cannot be arbitrarily large or small
2. there will have to be gaps between them.

Since any real number has to rounded off to the closest represented number, this introduces rounding errors.

Several diff. representations have been proposed but by far the most widely used is the floating point representation.

The floating point number system is a subset $\mathbb{F} \subseteq \mathbb{R}$ determined by a base β (an integer ≥ 2) and an integer $p \geq 1$, known as precision.

The elements of $\mathbb{F}$ are 0 along with all numbers of the form $\pm \frac{m}{\beta^p} \times \beta^e$, where m is an integer $1 \leq m \leq \beta^p$,

e is an arbitrary integer, called exponent.

A floating point number is represented as -

$$\underbrace{\pm}_{\text{sign}} \underbrace{d.d \cdots d}_{\substack{\text{significand} \\ \text{(has } p \text{ digits)}}} \times \underbrace{\beta}_{\text{base}}{}^{e \to \text{exponent}}$$

More precisely, the number $\pm\left(d_0 + d_1\beta^{-1} + \cdots + d_{p-1}\beta^{-(p-1)}\right) \times \beta^e$ (each $0 \leq d_i < \beta$)

is stored as -

$$d_0 . d_1 \cdots d_{p-1} \times \beta^e .$$

eg ① $\beta = 10$, $p = 3$: 0.1 is represented as 1.00×10^{-1}

(2) if $\beta = 2$, $p = 24$, then 0.1 cannot be represented exactly. It is approximated by the nearest floating point number

$$1.10011001100110011001101 \times 2^{-4}.$$

$\left(0.1 = \underset{\text{(decimal)}}{\frac{1}{10}} = \underset{\text{(binary)}}{\frac{1}{1010}} \right.$, long division in binary $\left. 1010 \overline{)\,1} \right)$

(3) Let $\beta = 2$, $p = 3$, $e_{min} = -1$, $e_{max} = 2$.

There are 16 normalized floating point numbers.

$\_\ \_\ \_$, allowed digits : $0, 1$ (since $0 \le d_t < \beta$).

$\left(d_0 . d_1 d_2 \times 2^e \right)$ — Floating pt. representation

$\left(d_0 + d_1 \times 2^{-1} + d_2 \times 2^{-2} \right) \times 2^e$ — real number being represented.

Floating pt. representation	real number
1.00×2^{-1}	0.5
1.01×2^{-1}	0.625
1.10×2^{-1}	0.75
1.11×2^{-1}	0.875
1.00×2^{0}	1
1.01×2^{0}	1.25
1.10×2^{0}	1.5
1.11×2^{0}	1.75
1.00×2^{1}	2
1.01×2^{1}	2.5
1.10×2^{1}	3
1.11×2^{1}	3.5
1.00×2^{2}	4
1.01×2^{2}	5
1.10×2^{2}	6
1.11×2^{2}	7

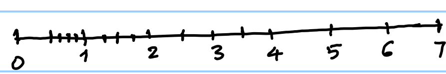
$1.00 \leftrightarrow 1 \times 2^{-1} = \frac{1}{2} = 0.5$

$1.01 \leftrightarrow (1 + 1 \times 2^{-2}) \times 2^{-1} = \left(1 + \frac{1}{4}\right) \times 2^{-1} = \frac{5}{4} \times \frac{1}{2} = \frac{5}{8} = 0.625.$

& so on.

Notice that the floating point numbers are not equally spaced.

The numbers represented are:



and their negatives.

So in this case, $|\mathbb{F}| = 33$ (0, positives & negatives).