Floating point arithmetic. (Demmel).

Note Title

The fact that infinitely many real nos. have to be stored in a finite amount of space gives rise to 2 limitations:
1) cannot represent arbitrary large or small numbers.
2) there will have to be gaps between the numbers.

So any real no. has to be rounded off to the closest representative — this introduces rounding error.

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

The floating point number system $\mathbb{F}$ (IEEE standard) is the system that is accepted & used in all computing systems now.

$\mathbb{F} \subseteq \mathbb{R}$ determined by a base β & a precision p

$\mathbb{F} = \{0\} \cup \{\text{floating pt. numbers}\}$. ($\beta$ is an integer ≥ 2) (p is an integer ≥ 1)

Elements of $\mathbb{F}$ are called floating point numbers & are represented as follows:

any real number can be considered as

$$\pm \left[d_0 + d_1 \beta^{-1} + d_2 \beta^{-2} + \cdots + d_{p-1} \beta^{-(p-1)} \right] \times \beta^{e}$$

(exponent: e)

where each $0 \leq d_i < \beta$

& is stored as: $\pm \; d_0 \cdot d_1 d_2 \cdots d_{p-1} \times \beta^{e}$ ($e \to$ exponent.)

($\pm$: sign; $d_0 \cdot d_1 d_2 \cdots d_{p-1}$: significand ($p$ digits))

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

$0 \leq d_i \leq 9$.

0.10×10^{0}, 0.01×10^{1} (unnormalized)

In binary.

② $\beta = 2$, $p = 3$: 0.1 is represented as $0.000110011\ldots$

$$(0.1) = \sum_{i=0}^{N} d_i \left(\frac{1}{2^i}\right) = 0\left(\frac{1}{2}\right) + 0\left(\frac{1}{4}\right) + 0\left(\frac{1}{8}\right) + 1\left(\frac{1}{16}\right) + 1\left(\frac{1}{32}\right) + \cdots$$

OR $0.1 = \frac{1}{10} = \frac{1}{1010_{(2)}}$, use long division

Normalised repr. with $p = 3$: 1.10×2^{-4}.

Ex. Let $\beta = 2$, $p = 3$, $e_{min} = -1$, $e_{max} = 2$. $\quad d_0 . d_1 d_2 \quad 2^{-1}$

(normalised). $\quad 0 \le d_i < \beta \quad 2^{-2}$

List all floating point numbers for $\mathbb{F}$ with these parameters

Floating pt. number	Corr. real number.
1.00×2^{-1}	$1.00 \times 2^{-1} = 1 \times \frac{1}{2} = 0.5$
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

Notice that the floating point numbers are not equally spaced.