Reals

CS 220 — Applied Discrete Mathematics

March 26, 2025

Representing the Reals

We've talked about how to represent $\mathbb{Z}$ given "hardware limitations":

- Pick a (contiguous) set of m "representatives": $Int \subset \mathbb{Z}$.
- Implement operations using modular arithmetic with modulus m.

How can we represent $\mathbb{R}$?

Representing the Reals

We've talked about how to represent $\mathbb{Z}$ given "hardware limitations":

- Pick a (contiguous) set of m "representatives": $Int \subset \mathbb{Z}$.
- Implement operations using modular arithmetic with modulus m.

How can we represent $\mathbb{R}$?

- Pick a set of "representatives": $Float \subset \mathbb{R}$ (*)
- Implement operations ...somehow.

Two basic strategies:

- fixed-point representation & arithmetic
- floating-point representation & arithmetic

Fixed-Point Representation

Definition (Fixed-Point Representation)

A **fixed-point representation** of real numbers devotes *fixed* amounts of space to the whole part and fractional part.

For example: four (decimal) digits total, two before the decimal point and two after the decimal point. (The decimal “point” is in a “fixed” position.)

Advantages:

- can implement easily on top of integer support
  for example, fractional dollars = integer number of cents
- good for domains that already use given granularity

Disadvantages:

- often can’t represent data with the domain’s natural scale
- poor at handling quantities at different scales; difficult to re-use code

Floating-Point Representation

Definition (Floating-Point Representation)

A **floating-point representation** of real numbers devotes *variable* amounts of space to the whole part and fractional part.

A number is represented as a **significand** multiplied by a scale calculated from an **exponent**, similar to scientific notation.

For example: $Float = \{(s, e) \mid s \in \{0, \dots, 999\}, \; e \in \{-5, \dots, 4\}\}$

- (s, e) represents $s \times 10^e$
- four digits total: three digits of significand, one digit of exponent
- normalization:
  - keep s in range $\{100, \dots, 999\}$ if possible
    for example, 1.0 is represented as 100×10^{-2}
  - pick one exponent for zero: for example, 0.0 is represented as 0×10^{-5}
- IEEE 754 uses sign bit; also adds $+\infty, -\infty, \mathtt{NaN}$

Floating-Point Arithmetic

Like arithmetic using scientific notation (except no significant digits!):

- $s_1 \times 10^{e_1} \boxplus s_2 \times 10^{e_2}$ and $s_1 \times 10^{e_1} \boxminus s_2 \times 10^{e_2}$
  First, put both on the same scale (may temporarily use extra digits).
  Add/subtract, then re-normalize, round to closest element of $Float$.
- $s_1 \times 10^{e_1} \boxtimes s_2 \times 10^{e_2}$
  Multiply significands (may use extra digits!), add exponents,
  then re-normalize, round to closest element of $Float$.

Examples

Compute $123 \times 10^3 \boxplus 456 \times 10^1$:

- Rescale: $123.0 \times 10^3 \boxplus 004.6 \times 10^3$ (one extra temporary digit)
- Add: 127.6×10^3
- Round: 128×10^3

Example: Non-Associative

Consider $123 \times 10^0 \boxplus 246 \times 10^3 \boxminus 246 \times 10^3$:

$$
\begin{aligned}
&(123 \times 10^0 \boxplus 246 \times 10^3) \boxminus 246 \times 10^3 \\
&= 246 \times 10^3 \boxminus 246 \times 10^3 \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
&123 \times 10^0 \boxplus (246 \times 10^3 \boxminus 246 \times 10^3) \\
&= 123 \times 10^0 \boxplus 0 \\
&= 123 \times 10^0 && \neg\text{Associative}
\end{aligned}
$$

Numerical algorithms must be careful to avoid or mitigate such errors.

Keywords: *numerical analysis*, *error analysis*, *numerical stability*, *catastrophic cancellation*

"What Every Computer Scientist Should Know About Floating-Point Arithmetic" (Goldberg 1991)

Example: Pathologically Non-Associative

Consider $500 \times 10^4 \boxplus 500 \times 10^4 \boxplus -500 \times 10^4 \boxplus -500 \times 10^4$:

$$
\begin{aligned}
&((500 \times 10^4 \boxplus 500 \times 10^4) \boxplus -500 \times 10^4) \boxplus -500 \times 10^4 \\
&= (+\infty \boxplus -500 \times 10^4) \boxplus -500 \times 10^4 \\
&= +\infty \boxplus -500 \times 10^4 \\
&= +\infty
\end{aligned}
$$

$$
\begin{aligned}
&500 \times 10^4 \boxplus (500 \times 10^4 \boxplus (-500 \times 10^4 \boxplus -500 \times 10^4)) \\
&= 500 \times 10^4 \boxplus (500 \times 10^4 \boxplus -\infty) \\
&= 500 \times 10^4 \boxplus -\infty \\
&= -\infty
\end{aligned}
$$

$$
\begin{aligned}
&(500 \times 10^4 \boxplus 500 \times 10^4) \boxplus (-500 \times 10^4 \boxplus -500 \times 10^4) \\
&= \infty \boxplus -\infty \\
&= \mathtt{NaN}
\end{aligned}
$$

$$
\begin{aligned}
&500 \times 10^4 \boxplus (500 \times 10^4 \boxplus -500 \times 10^4) \boxplus -500 \times 10^4 \\
&= 500 \times 10^4 \boxplus 0 \boxplus -500 \times 10^4 \\
&= 500 \times 10^4 \boxplus -500 \times 10^4 \\
&= 0
\end{aligned}
$$