Floating Point Representation

Fractional Binary Numbers
IEEE floating-point standard
Floating-point operations and rounding
Lessons for programmers

Many more details we will skip (it’s a 58-page standard...) See CSAPP 2.4 for more detail.

https://cs.wellesley.edu/~cs240/

Fractional Binary Numbers

Value Representation
5 and 3/4
2 and 7/8
47/64

Observations
Shift left =
Shift right =
Numbers of the form 0.111111... are...?

Limitations:
Exact representation possible when?

1/3 = 0.333333... = 0.

Fixed-Point Representation

Implied binary point.

\[ b_7 b_6 b_5 b_4 b_3 [.] b_2 b_1 b_0 \]
\[ b_7 b_6 b_5 b_4 b_3 b_2 b_1 b_0 [.] \]

range: difference between largest and smallest representable numbers
precision: smallest difference between any two representable numbers

fixed point = fixed range, fixed precision
IEEE Floating Point Standard 754

IEEE = Institute of Electrical and Electronics Engineers

Numerical form:
\[ V_{10} = (-1)^s \cdot M \cdot 2^E \]

- **Sign bit** \( s \) determines whether number is negative or positive
- **Significand (mantissa)** \( M \) usually a fractional value in range \([1.0, 2.0)\)
- **Exponent** \( E \) weights value by a \((-/+)\) power of two

Analogous to scientific notation

Representation:
- MSB \( s \) = sign bit \( s \)
- \( \text{exp} \) field encodes \( E \) (but is *not equal* to \( E \))
- \( \text{frac} \) field encodes \( M \) (but is *not equal* to \( M \))

Finite representation of infinite range...

Three kinds of values

\[ V = (-1)^s \cdot M \cdot 2^E \]

1. **Normalized**: \( M = 1.xxxxx... \)
   - As in scientific notation: \( 0.011 \times 2^5 \) = \( 1.1 \times 2^3 \)
   - Representation advantage?

2. **Denormalized**, near zero: \( M = 0.xxxxx... \), smallest \( E \)
   - Evenly space near zero.

3. **Special values**:
   - **0.0**: \( s = 0 \) \( \text{exp} = 00...0 \) \( \text{frac} = 00...0 \)
   - **+inf, -inf**: \( \text{exp} = 11...1 \) \( \text{frac} = 00...0 \)
   - **NaN** (“Not a Number”): \( \text{exp} = 11...1 \) \( \text{frac} \neq 00...0 \)
   - \( \sqrt(-1), \infty - \infty, \infty / 0, \) etc.

Precisions

**Single precision** (float): 32 bits

\[ \text{MSB} \quad s \quad \text{exp} \quad \text{frac} \]
1 bit 8 bits 23 bits

**Double precision** (double): 64 bits

\[ \text{MSB} \quad s \quad \text{exp} \quad \text{frac} \]
1 bit 11 bits 52 bits

Value distribution

\[ -\infty \quad -\text{Normalized} \quad -\text{Denormalized} \quad -\text{Denormalized} \quad +\text{Normalized} \quad +\infty \]

\[ +\infty \quad +0.0 \quad 0.0 \quad -0.0 \quad +\text{NaN} \quad +\text{NaN} \]
Normalized values, with float example

\[ V = (-1)^s \cdot M \cdot 2^E \]

Value: \( f = 12345.0 \);
\[ 12345_{10} = 1100000111001_2 \]
\[ = 1.100000111001 \times 2^{13} \quad \text{(normalized form)} \]

Significand:

\[ M = 1.100000111001 \]
\[ \text{frac} = 100000000000000000000000 \]

Significand: leading zero

\[ M = 0.\text{xxx...x}_2 \]
\[ \text{frac} = \text{xxx...x} \]

Cases:

- \( \exp = 000...0, \text{frac} = 000...0 \)
- \( \exp = 000...0, \text{frac} \neq 000...0 \)

Try to represent 3.14, 6-bit example

6-bit IEEE-like format

Bias = \( 2^{3-1} - 1 = 3 \)

Full Range

\[ s=1, \exp=101 \]
\[ E = 5-3 = 2 \]

Zoom in to 0

\[ s=0, \exp=110 \]
\[ E = 6-3 = 3 \]

Value distribution example

6-bit IEEE-like format

Bias = \( 2^{3-1} - 1 = 3 \)

Denormalized Values: near zero

"Near zero": \( \exp = 000...0 \)

Exponent:

\[ E = 1 + \exp - \text{Bias} = 1 - \text{Bias} \quad \text{not:} \quad \exp - \text{Bias} \]

Significand: leading zero

\[ M = 0.\text{xxx...x}_2 \]
\[ \text{frac} = \text{xxx...x} \]

Cases:

- \( \exp = 000...0, \text{frac} = 000...0 \)
- \( \exp = 000...0, \text{frac} \neq 000...0 \)

Floating Point
Floating Point Arithmetic*

\[ V = (-1)^s \cdot M \cdot 2^E \]

1. Compute exact result.
2. Fix/Round, roughly:
   - Adjust \( M \) to fit in \([1.0, 2.0)\):
     - If \( M \geq 2.0 \): shift \( M \) right, increment \( E \)
     - If \( M < 1.0 \): shift \( M \) left by \( k \), decrement \( E \) by \( k \)
   - Overflow to infinity if \( E \) is too wide for \text{exp}
   - Round* \( M \) if too wide for \text{frac}
   - Underflow if nearest representable value is 0.

Lessons for programmers

\[ V = (-1)^s \cdot M \cdot 2^E \]

float ≠ real number ≠ double
Rounding breaks associativity and other properties.

double a = ... , b = ...

\*complicated...

if (a == b) ...

if (abs(a - b) < epsilon) ...