BST 234: Lab - 1

Divy Kangeyan

February 6, 2019

Normalized Floating Point Representation

$$x = \pm m * b^{\pm e}$$

- base $b \in \mathbb{N}$ and $b > 1$
- mantissa $m = m_1 b^{-1} + ... + m_r b^{-r} \in \mathbb{R}$
- exponent $e = e_{s-1} b^{s-1} + ... + e_0 b^0 \in \mathbb{N}$
- digits $m_i, e_i \in 0, ..., b-1$
- significant digits $s \in \mathbb{N}$ and $r \in \mathbb{N}$

Practice: Normalized Floating Point Representation

Express the number $x = 10$ in normalized floating point format for the base $b = 2$:

Practice: Normalized Floating Point Representation

Express the number x = 10 in normalized floating point format for the base b = 2:

- Find k such that $b^k \leq x \leq b^{k+1}$

$$2^3 \leq 10 \leq 2^4$$

Practice: Normalized Floating Point Representation

Express the number x = 10 in normalized floating point format for the base b = 2:

- Find k such that $b^k \leq x \leq b^{k+1}$

$$2^3 \leq 10 \leq 2^4$$

- Factor x into $b^k, b^{k-1}, \ldots$

$$10 = 2^3 \times 1 + 2^2 \times 0 + 2^1 \times 1 + 2^0 \times 0$$

Practice: Normalized Floating Point Representation

Express the number x = 10 in normalized floating point format for the base b = 2:

- Find k such that $b^k \leq x \leq b^{k+1}$

$$2^3 \leq 10 \leq 2^4$$

- Factor x into $b^k, b^{k-1}, ...$

$$10 = 2^3 \times 1 + 2^2 \times 0 + 2^1 \times 1 + 2^0 \times 0$$

- Add terms and factor out b^{k+1}

$$10 = 2^4(2^{-1} \times 1 + 2^{-2} \times 0 + 2^{-3} \times 1 + 2^{-4} \times 0)$$

Practice: Normalized Floating Point Representation

Express the number x = 10 in normalized floating point format for the base b = 2:

- Find k such that $b^k \leq x \leq b^{k+1}$

$$2^3 \leq 10 \leq 2^4$$

- Factor x into $b^k, b^{k-1}, ...$

$$10 = 2^3 \times 1 + 2^2 \times 0 + 2^1 \times 1 + 2^0 \times 0$$

- Add terms and factor out b^{k+1}

$$10 = 2^4(2^{-1} \times 1 + 2^{-2} \times 0 + 2^{-3} \times 1 + 2^{-4} \times 0)$$

- Answer: $x = (.101)_2 * 2^4$

Practice: Normalized Floating Point Representation

Express the number $x = 100$ in normalized floating point format for the base $b = 3$:

Practice: Normalized Floating Point Representation

Express the number x = 100 in normalized floating point format for the base b = 3:

- Find k such that $b^k \leq x \leq b^{k+1}$
- Factor x into $b^k, b^{k-1}, \ldots$
- Add terms and factor out b^{k+1}

Practice: Normalized Floating Point Representation

Express the number x = 100 in normalized floating point format for the base b = 3:

- Find k such that $b^k \leq x \leq b^{k+1}$
- Factor x into $b^k, b^{k-1}, \ldots$
- Add terms and factor out b^{k+1}
- Answer: $x = (.10201)_3 * 3^5$

Machine Precision

Definition:

$$eps := \frac{1}{2} b^{-r+1}$$

For the IEEE-format:

$$eps_{IEEE} \leq \frac{1}{2} 2^{-51}$$

- *Python demonstration*

Definition:

$$eps := \frac{1}{2} b^{-r+1}$$

For the IEEE-format:

$$eps_{IEEE} \leq \frac{1}{2} 2^{-51}$$

- *Python demonstration*

$$\frac{7}{3} = 1.0010101010101010101010101010101010101010101010101011 * 2^1$$

$$\frac{4}{3} = 1.0101010101010101010101010101010101010101010101010101 * 2^0$$

$$\frac{7}{3} - \frac{4}{3} = 1.0000000000000000000000000000000000000000000000000001 * 2^0$$

$$\frac{7}{3} - \frac{4}{3} - 1 = 2^{-52} = \epsilon$$

Floating point arithmetic

- Since floating point arithmetic is inherently approximate and not exact following symbols are used: $\oplus, \ominus, \otimes, \oslash$
- $(x \oplus y) \oplus z \neq x \oplus (y \oplus z)$ (Associative law doesn't hold)
- $(x \oplus y) \otimes z \neq (x \otimes z) \oplus (y \otimes z)$ (Distributive law doesn't hold)
- $x \oplus y = x$ for $|y| \leq \frac{|x|}{b} eps$
- *Python demonstration*

Using Karatsuba algorithm for multiplication

Use Karatsuba algorithm to find the results for 15×24 in binary