This is a title

Jiuru Lyu

August 31, 2023

Contents

1 Floating Point Numbers **2**
1.1 Binary Representation 2
1.2 Integers in Computers 2
1.3 Representation of Floating Point Numbers 3

1 Floating Point Numbers

1.1 Binary Representation

Definition 1.1.1 (Binary). 0 and 1; on and off.

Example 1.1.2 Represent Numbers in Base-2

Consider $13 = 1(10) + 3(1) = 1(10) + 3(10^0)$ in base-10. It can be converted into base-2 by decomposing 13 as $1(2^3) + 1(2^2) + 0(2^1) + 1(2^0)$.

Example 1.1.3 Fractions in Base-2

$$\frac{7}{16} = \frac{1}{16}(7) = \left(2^{-4}\right)\left(2^2 + 2^1 + 2^0\right) = 2^{-2} + 2^{-3} + 2^{-4}.$$

Example 1.1.4 Repeating Fractions in Base-2

$$\begin{aligned} \frac{1}{5} = \frac{1}{8} + \varepsilon_1 \quad &\Longrightarrow \quad \varepsilon_1 = \frac{1}{5} - \frac{1}{8} = \frac{8-5}{(5 \times 8)} = \frac{3}{40} \\ \varepsilon_1 = \frac{3}{3(16)} + \varepsilon_2 \quad &\Longrightarrow \quad \cdots \end{aligned}$$

Repeating the steps above, we would finally get

$$\frac{1}{5} = \frac{1}{8} + \frac{1}{16} + \frac{1}{128} + \frac{1}{256} + \cdots$$

Theorem 1.1.5

Let $n \in \mathbb{Z}$ and $n \geq 1$, then

$$\sum_{k=0}^{n-1} 2^k = 2^{n-1} + 2^{n-2} + \cdots + 2^0 = 2^n - 1.$$

1.2 Integers in Computers

Definition 1.2.1 (Storing Integers). `unit8` stands for unsigned integers and `int8` stands for signed integers.

Remark. *The 8 here represents 8 bits. It is a measure of how much storage (how many 0s or 1s).*

	b_7	b_6	b_5	b_4	b_3	b_2	b_1	b_0
unsigned:	2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
signed:	-2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0

Example 1.2.2

$$\texttt{unit8}(13) = 00001101$$

Since $-13 = 1(-2^7) + 1(2^6) + 1(2^5) + 1(2^4) + 0(2^3) + 0(2^2) + 1(2^1) + 1(2^0)$, we have

$$\texttt{int8}(-13) = 11110011$$

Remark. *Largest and Smallest Integers:*

$$\begin{aligned}
\texttt{uint8}(x_L) &= 11111111 &&\Longrightarrow x_L = 2^7 + 2^6 + \cdots + 2^0 = 2^8 - 1 = 255\\
\texttt{uint8}(x_S) &= 00000000 &&\Longrightarrow x_S = 0(2^7) + 0(2^6) + \cdots + 0(2^0) = 0\\
\texttt{int8}(x_L) &= 01111111 &&\Longrightarrow x_L = 0(-2^7) + 2^6 + \cdots + 2^0 = 2^7 - 1 = 127\\
\texttt{int8}(x_S) &= 100000000 &&\Longrightarrow x_S = 1(-2^7) + 0(2^6) + \cdots + 0(2^0) = -128
\end{aligned}$$

1.3 Representation of Floating Point Numbers

Definition 1.3.1 (Normalized Scientific Notation). Only 1 digit (non-zero) to the left of the decimal point.

Example 1.3.2

$$\begin{aligned}
&123.456 \times 10^7\\
&12.3456 \times 10^8\\
&1.23456 \times 10^9 \to \text{normalized}
\end{aligned}$$

Definition 1.3.3 (Anatomy of Floating Point Numbers). A floating point number, $\texttt{float}(x)$, consists of three parts: $s(x)$ (sign bit), $e(x)$ (exponent bits), and $f(x)$ (fraction bits).
Definition 1.3.4 (Precision). Precision is defined by the number of bits per part:

	$s(x)$	$e(x)$	$f(x)$	total
double precision (DP)	1	11	52	64
single precision (SP)	1	8	23	32
half precision (HP)	1	5	10	16

Remark. *The less bits the float point number has, the less storage it requires and faster computation it performs, but more error introduces.*

Definition 1.3.5 (Floating Point Number).

$$\texttt{float}(x) = (-1)^{s(x)}\left(1 + \frac{f(x)}{2^{\#\text{ of fraction bits}}}\right)2^{E(x)}, \tag{1}$$

where $E(x)$ is called the *unbiased exponent* because it is centered about 0 and is calculated through the $e(x)$, the *biased exponent* because it can only be non-negative integers, by the following formula:

$$E(x) = e(x) - \left(2^{\#\text{ of exponent bits}-1} - 1\right).$$

Remark. *Eq. (1) is in normalized scientific notation because the largest number $f(x)$ can represent is $2^{\#\text{ of fraction bits}} - 1$. Hence,*

$$1 + \frac{f(x)}{2^{\#\text{ of fraction bits}}} < 2,$$

and thus there will be only 1 digit in front of the decimal point.

Example 1.3.6 Formula for a Floating Point Number in Double Precision (DP)

$$\texttt{float}_{\text{DP}}(x) = (-1)^{s(x)}\left(1 + \frac{f(x)}{2^{52}}\right)2^{e(x)-1023}.$$

Example 1.3.7 Converting DP into Decimal

Suppose a DP floating number is stored as $s(x) = 0$, $e(x) = 10000000011$, and $f(x) = 0100100\cdots0$. Find its representation in decimal base-10.

Solution 1.

$e(x) = 10000000011 = 2^{10} + 2^1 + 2^0$ and $f(x) = 0100100\cdots0 = 2^{50} + 2^{47}$. Then, the unbiased exponent $E(x) = e(x) - 1023 = 2^{10} + 2^1 + 2^0 - (2^{10} - 1) = 4$. So,

$$\begin{aligned}
\texttt{float}_{\text{DP}}(x) &= (-1)^{s(x)} + \left(1 + \frac{f(x)}{2^{52}}\right)2^{E(x)} \\
&= (-1)^0\left(1 + \frac{2^{50} + 2^{47}}{2^{52}}\right)2^4 \\
&= \left(1 + 2^{-2} + 2^{-5}\right)2^4 \\
&= 2^4 + 2^2 + 2^{-1} \\
&= 16 + 4 + 0.5 = 20.5
\end{aligned}$$

□

Example 1.3.8 Converting Value to DP

Suppose a number in base-10 is -10.75. Find its representation of floating point number under DP.

Solution 2.

We have

$$\begin{aligned}\texttt{value}(x) = -10.75 &= (-1)(10+0.75)\\ &= (-1)\left(2^3+2^1+2^{-1}+2^{-2}\right)\\ &= (-1)\left(1+2^{-2}+2^{-4}+2^{-5}\right)2^3 \quad \left[\text{In normalized scientific notation}\right]\\ &= (-1)^1\left(1+\frac{2^{50}+2^{48}+2^{47}}{2^{52}}\right)2^{1026-1023}\\ &= (-1)^1\left(1+\frac{2^{50}+2^{48}+2^{47}}{2^{52}}\right)2^{2^{10}+2^1-1023}\end{aligned}$$

So, we have $s(x)=1$, $e(X)=10000000010$, and $f(x)=010110\cdots0$. □

Theorem 1.3.9 Some Special Rules

1. The formula
$$\texttt{value}(x) = (-1)^{s(x)} + \left(1+\frac{f(x)}{2^{52}}\right)2^{e(x)-1023}$$
only holds when $0 < e(x) < 2^{11}-1$ or $00\cdots01 < e(x) < 11\cdots10$.

2. If $e(x) = 11\cdots1$, then it encodes special numbers.

3. If $e(x) = 00\cdots0$:
   - If $f(x) = 00\cdots0$, then $\texttt{value}(x) = 0$.
   - If $f(x) > 0$, it encodes a *denormalized floating point number*:
   $$\texttt{value}(x) = (-1)^{s(x)}\left(0+\frac{f(x)}{2^{52}}\right)2^{-1022}.$$
   This denormalized floating point number is more precise when describing really small things.

Definition 1.3.10 (Machine Epsilon/$\varepsilon_{\mathbf{WP}}$). Let "WP" stands for the working precision (DP/SP/HP/etc.). The *machine epsilon*, denoted as ε_{WP}, is the gap between 1 and the next largest floating point number. Equivalently, it can be viewed as the smallest possible non-zero value of $\frac{f(x)}{2^{\text{number of fraction bits}}}$. So, $\varepsilon_{\text{DP}} = 2^{-52}$, $\varepsilon_{\text{SP}} = 2^{-23}$, and $\varepsilon_{\text{HP}} = 2^{-10}$.

Definition 1.3.11 (Special Numbers).

1. ± 0: when $s(x) = \pm 1$ and $e(x) = f(x) = 0$.

2. $\pm$`Inf`

3. `NaN`: not-a-number

Definition 1.3.12 (Floating Point Arithmetic).

1. The set of real numbers, $\mathbb{R}$, is closed under arithmetic operations.

2. The set of all WP floating point numbers, however, is not closed under arithmetic operations. For example, $\texttt{float}_{\text{DP}}(x) = \texttt{float}_{\text{DP}}(y) = 2^{52} + 1$, but $xy = 2^{104} + \varepsilon$ cannot be represented using DP.

3. Suppose x and y are floating point numbers, then $x \oplus y = \texttt{float}(x + y)$ and $x \otimes y = \texttt{float}(xy)$. Consider `float` as a rounding process, we can also define subtraction and division of floating point numbers.

Example 1.3.13

Assume we are only allowed three significant digits (in Base-10) in a computer. Suppose $x = 1.23 \times 10^4$ and $y = 6.54 \times 10^3$. Find $x \oplus y = \texttt{float}(x + y)$.

Solution 3.

$$\begin{aligned}
x \oplus y &= \texttt{float}(x + y) \\
&= \texttt{float}(1.23 \times 10^4 + 6.54 \times 10^3) \\
&= \texttt{float}(1.23 \times 10^4 + 0.654 \times 10^3) \\
&= \texttt{float}(1.884 \times 10^4) \\
&= 1.88 \times 10^4.
\end{aligned}$$

□

Answer.m

```
% Plot function f(x) = 2*x^3 - x - 2
ezplot('2*x^3-x-2', [0, 2])
hold on
plot([0, 2], [0, 0], 'r')
```

Algorithm 1: Bisection Algorithm

Input: $a, b, M, \delta, \varepsilon$
$u \leftarrow f(a)$
$b \leftarrow f(b)$
$e \leftarrow b - a$
Output: output

begin
  if *sign*$(u) =$ *sign*(v) **then**
    stop
  for *k=1* **to** M **do**
    $e \leftarrow e/2$
    $c \leftarrow a + e$
    $w \leftarrow f(c)$
    return k, c, w, e
    if $|e| < \delta$**or** $|w| < \varepsilon$ **then**
      stop
    if *sign*$(u) \neq=$ *sign*(v) **then**
      $b \leftarrow c$
      $v \leftarrow w$
    else
      $a \leftarrow c$
      $u \leftarrow w$