This is a title

Jiuru Lyu

August 28, 2023

Contents

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

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

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:*

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

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

123.456×10^7

12.3456×10^8

$1.23456 \times 10^9 \rightarrow$ normalized

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,

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

□

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

```
1  begin
2      if sign(u) = sign(v) then
3          stop
4      for k=1 to M do
5          e ← e/2
6          c ← a + e
7          w ← f(c)
8          return k, c, w, e
9          if |e| < δ or |w| < ε then
10             stop
11         if sign(u) ≠= sign(v) then
12             b ← c
13             v ← w
14         else
15             a ← c
16             u ← w
```