MATH 125 Lecture 2

Most computers store data in binary format.

For example, to represent the number 729 in binary format, we proceed as follows:

2^0	2^1	2^2	2^3	2^4	2^5	2^6	2^7	2^8	2^9	2^{10}
1	2	4	8	16	32	64	128	256	512	1024

$$
\begin{aligned}
\Rightarrow \quad 729 &= 2^9 + 217 \\
&= 2^9 + 2^7 + 89 \\
&= 2^9 + 2^7 + 2^6 + 25 \\
&= 2^9 + 2^7 + 2^6 + 2^4 + 9 \\
&= 2^9 + 2^7 + 2^6 + 2^4 + 2^3 + 1 \\
&= 2^9 + 2^7 + 2^6 + 2^4 + 2^3 + 2^0
\end{aligned}
$$

Therefore, 729 has the following binary representation

1	0	1	1	0	1	1	0	0	1
2^9	2^8	2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0

Exercise What is the binary representation of 729.25?

Solution

1	0	1	1	0	1	1	0	0	1	0	1
2^9	2^8	2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0	2^{-1}	2^{-2}

Exercise What is the representation in binary of $1/3$?

Solution You can check that $\frac{1}{3} = 0.01010101\ldots$
That is $\frac{1}{3}$ can not be represented exactly.

We also can conclude that all irrational numbers have infinitely long binary representations.

Fixed point representations

- uses a fixed decimal point
- store using 0 or 1 coefficients in front of powers of two
- Range from 2^{-k} to 2^{ℓ} $k, \ell \in \mathbb{Z}$

Example To represent all numbers between 0 and 127.75 in increments of 1/4, we set $k = 2$ and $l = 7$.

Advantage
Many operations can use same machinery for integers

Let a and b be in fixed format

$$a + b = \underbrace{(a \cdot 2^k + b \cdot 2^k)}_{\text{integers}} \cdot 2^{-k}$$

This allows use of preexisting integer arithmetic hardware

Disadvantage
could suffer from precision issues

Example 1 decimal precision

$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$ $\quad 0.1_2 \times 0.1_2 = 0.01_2 \longrightarrow$ Truncated to zero

However, fixed point representation can be useful where precision is not so important e.g. low end GPU

Floating Point representations

Many quantities of interest in science are over different scales

Example Mass of electron 9.11×10^{-31}
Avogadro constant 6.022×10^{23}

Need a representation that is flexible.

Question Why did we write Avogadro's constant as 6.022×10^{23}?

Answer 1 That is the constant.
Answer 2 Scientific notation
$6.022 \times 10^{23} \equiv$ we know the constant to three decimal places

Namely, scientific notation is representation of the form $a \times 10^{e}$ $a \sim 1$ and $e \in \mathbb{Z}$

Decimal point is not fixed. It "floats" so that a is on reasonable scale.

$a \equiv$ significand $\qquad e \equiv$ exponent

$0.10101_2 \times 2^3 \Rightarrow$ Normalized

$0.010101_2 \times 2^4 \Rightarrow$ Not normalized

Normalized representation $\qquad (-1)^s \, d_1.d_2 d_3 \ldots d_t \; 2^e$

- Unique representation
- Extra bits for storage

$t \equiv$ precision $\qquad d_1 = 1$

$(-1)^s \equiv$ sign

$d_i = 0$ or $1 \qquad 2 \le i \le t$

Important parameters
- precision t
- Minimum and maximum exponents L and U

IEEE standard has recommendations

IEEE single precision

1	8	23
Sign	Exponent	Fraction

32 bits

$(-1)^s \cdot 2^{e-127} (1+f)$

IEEE double precision

1	11	52
Sign	Exponent	Fraction

64 bits

$(-1)^s \cdot 2^{e-1023} (1+f)$

(IEEE 754 standard)

Underflow threshold	2^{-126}	single precision
overflow threshold	$2^{127} \cdot (2 - 2^{-23}) \approx 2^{128}$	

$\approx (10^{-38}$ to $10^{38})$

Underflow threshold	2^{-1022}	Double precision
overflow threshold	$2^{1023} \cdot (2 - 2^{-52}) \approx 2^{1024}$	

$\approx (10^{-308}$ to $10^{308})$

<u>Gaps in floating numbers</u>

Let's consider IEEE double precision
What is the number next to 1?
$(1.\square\square\square\square\cdots\square)\times 2^0$

$1+2^{-52}$

What is next to $1+2^{-52}$? $1+2^{-51}$
$1+2^{-51} = 1+2\cdot 2^{-52}$

In general,
$1,\ 1+2^{-52},\ 1+2\cdot 2^{-52},\ 1+3\cdot 2^{-52},\ \ldots,\ 2$

For the interval $[2,4]$
$2,\ 2+2^{-51},\ 2+2\cdot 2^{-51},\ \ldots,\ 4$

Dense for small numbers and spread out for large numbers

<u>Remark:</u> IEEE 754 has standard with how to deal with $\pm\infty$, (not a number $\equiv$ NAN)

It is always good to write a code that is cognizant of overflow and underflow

---

<u>Rounding</u>

A simple case of rounding
$(1.\underline{d_1 d_2 \ldots d_m}\, d_{m+1}\ldots)\times 2^e$
do

chopping $\equiv$ Truncate d_{m+1} and forward. Let $fl(x)$ denote the result

<u>Theorem</u> $$\frac{|x - fl(x)|}{|x|} \le 2^{1-m}$$

<u>proof</u> Let $x\in\mathbb{R}$. Then there is some integer e between e_{min} and e_{max} such that

$$x = \pm 2^e \sum_{k=+\infty}^{0} d_k 2^{-k}$$

$$x = \underbrace{\pm 2^e \sum_{k=+m}^{0} d_k 2^{-k}}_{\Downarrow\ fl(x)} \pm 2^e \sum_{k=+\infty}^{m+1} d_k 2^{-k}$$

$$x = fl(x) \pm 2^e \sum_{k=m+1}^{\infty} d_k \, 2^{-k}$$

Geometric series (if we remove d_k)

Note d_k is either 0 or 1.
Therefore $d_k \le 1$

Recall
$$\sum_{k=0}^{n} a r^k = a\left(\frac{1-r^{n+1}}{1-r}\right)$$

It then follows that

$$|x - fl(x)| = \left|\pm 2^e \sum_{k=m+1}^{\infty} d_k \, 2^{-k}\right| \le \left|2^e \sum_{k=m+1}^{\infty} 2^{-k}\right|$$

(Recall $a + ar + ar^2 + ar^3 + \dots = \frac{a}{1-r}$)

$$= 2^e \cdot \frac{2^{-(m+1)}}{1 - 1/2}$$
$$= 2^e \cdot 2^1 \cdot 2^{-(m+1)}$$
$$= 2^e \cdot 2^{-m}$$

Since $d_0 = 1$, note that $|x| \ge 2^e$

$$\frac{|x - fl(x)|}{|x|} \le \frac{2^e \cdot 2^{-m}}{2^e} \le 2^{-m}$$

Fundamental axiom of floating point arithmetic

(i) For all $x \in \mathbb{R}$, there exists ε with $|\varepsilon| \le \varepsilon_{machine}$ such that $fl(x) = x(1+\varepsilon)$

(ii) Let $\otimes$ be the floating point analogue of elementary operation $\times$. $\Rightarrow$ (elementary operations $+, -, \times, \div$)

$$x \otimes y = (x \times y)(1+\varepsilon) \qquad |\varepsilon| \le \varepsilon_{machine}$$

Exercise convert 0.45 to binary

$$0.45 = x_1 \cdot 2^{-1} + x_2 \cdot 2^{-2} + \dots + x_n \cdot 2^{-n}$$
$$0.45 = x_1 \cdot \frac{1}{2} + x_2 \cdot \frac{1}{2^2} + \dots + x_n \cdot \frac{1}{2^n}$$

$$(0.45 \times 2) = x_1 + \underbrace{\left(\frac{1}{2} x_2 + \dots + \frac{1}{2^{n-1}} x_n\right)}_{\text{prove that it is less than 1}}$$

$x_1 = 0$

$$1.8 = x_2 + \underbrace{\left(\frac{1}{2} x_3 + \dots + \frac{1}{2^{n-2}} x_n\right)}_{\text{prove that it is less than 1}}$$

$x_2 = 1$

$$1.6 = x_3 + \left(\frac{1}{2} x_4 + \dots \frac{1}{2^{n-3}} x_n\right)$$

$x_3 = 1$

$$1.2 = x_4 + \left(\frac{1}{2} x_5 + \dots + \frac{1}{2^{n-4}} x_n\right)$$

$x_4 = 1$

continuing this

$0.4 \longrightarrow x_5 = 0$

$0.8 \longrightarrow x_6 = 0$

$1.6 \longrightarrow x_7 = 1$

$1.2 \longrightarrow x_8 = 1$

$0.4 \longrightarrow x_9 = 0$

$0.8 \longrightarrow x_{10} = 0$

$1.6 \longrightarrow x_{11} = 1 \dots$

Binary representation $\equiv 0.0111\underbrace{0011\,0011\,0011}_{\text{Repeating}}$