

# Chapter 5

①

Floating point arithmetic.

- Binary representation and base 2 arithmetic.

Addition.

$$\begin{array}{r} \phantom{+} 1010 \\ + 11011 \\ \hline 100101 \end{array} \qquad \begin{array}{r} \phantom{+} 10 \\ + 27 \\ \hline 37 \end{array}$$

Subtraction.

Multiplication:

Shift and add.

Division.

~~Deci~~ Rational Numbers.

$$0.b_1 b_2 b_3 \dots$$

$$= b_1 \times 2^{-1} + b_2 \times 2^{-2} + \dots$$

What is rational in base 10 may not be rational in base 2.

e.g. 0.1 in base 10 is  $0.000\overline{1100}$

# Floating point representation.

$$\pm m \times 2^E \quad 1 \leq m < 2$$

$$10 = 1010_2 = 1.010_2 \times 2^3$$

— .

Three fields.

Sign

mantissa

exponent.

\_\_\_\_\_ .

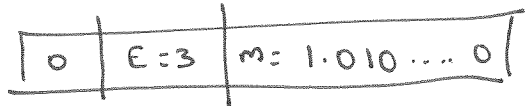
Float: 32 bits.

1 bit for sign.  $\left\{ \begin{array}{l} 0 : \text{positive} \\ 1 : \text{negative.} \end{array} \right.$

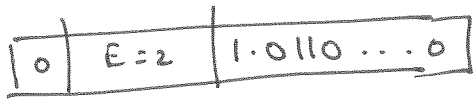
8 bits for exponent

23 bits for mantissa or significand.

eg.  $10 = 1.010_2 \times 2^3$



$$5.5 = 1.011 \times 2^2$$



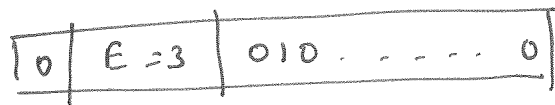
Some optimizations.

$$m \times 2^E$$

$$1 \leq m < 2$$

We always know that the most significant bit is 1. So there is no need to store it.

∴  $10 = 1010 \times 2^3$  is stored as



This is called hidden bit representation.

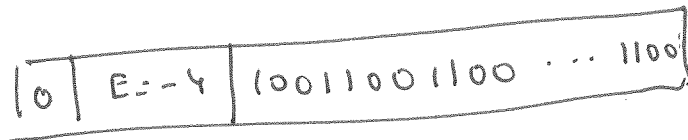
Rounding and approximation.

Consider the number 23 bits.

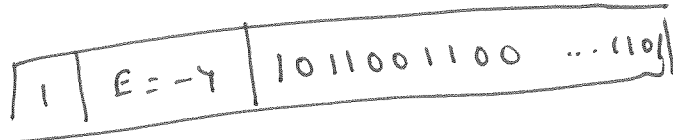
$$\frac{1}{10} = 1. \overbrace{10011001100110011001100}^{23 \text{ bits.}} \times 2^{-4}$$

This is an approximation.

It can be written as



or



Question: How to represent 0?

Machine precision.

(4)

Gap between 1 and next number is called machine precision.

The next number is

$$\boxed{10 \mid E=0 \mid 000 \text{ --- } 11}$$

$$= 1 + 2^{-23}$$

$$\text{Machine precision} = 2^{-23} \approx 1.2 \times 10^{-7}$$

For double precision.

1 bit for sign.

11 bits for exponent.

52 for mantissa.

$\therefore$  double precision machine precision is

$$2^{-52} \approx 2.2 \times 10^{-16}$$

Gap between 0 and smallest non-zero number and next number.

$$0, \quad 1.0 \times 2^{-E}$$

next smallest is  $(1+\epsilon) \times 2^{-E}$

Gap is  $\epsilon \times 2^{-E}$  ( $\epsilon$  is machine precision)

# IEEE Floating Point Arithmetic.

Representation for 0, ±∞ and NaN

- Special bits in the exponent field.
- Also used to represent subnormal numbers.

Three standard precision.

- 32 bit : 1 sign, 8 exponent, 23 significand.
- 64 bit : 1 sign, 11 exponent, 52 significand.
- 80 bit : 1 sign, 15 exponent, 64 significand.

Exponent.

- 00 ... 0 ± . b<sub>1</sub> b<sub>2</sub> ... b<sub>52</sub> × 2<sup>-1022</sup> 0 or Subnormal.
- 00 ... 1 ± 1. b<sub>1</sub> b<sub>2</sub> ... b<sub>52</sub> × 2<sup>-1022</sup>
- 00 ... 10 ± 1. b<sub>1</sub> b<sub>2</sub> ... b<sub>52</sub> × 2<sup>-1021</sup>

Exponent field is actual exponent + 1023

- 1111 ... 1 ± ∞ if b<sub>1</sub> b<sub>2</sub> ... b<sub>52</sub> = 0
- NAN otherwise.

# Smallest Subnormal number.

(6)

$$\frac{0}{s} \quad \frac{00 \dots 0}{E} \quad \frac{00 \dots 1}{m}$$

$$2^{-1022} \times 2^{52} = 2^{-1074}$$

$$0 : \quad 0 \quad \frac{00 \dots 0}{E} \quad \frac{00 \dots 0}{m}$$

## Roundiy :

- Round down
- Round up.
- Round towards 0
- Round to nearest.

Default is round to nearest.

$\frac{1}{10} = 1.1001100 \times 2^{-k}$  is replaced by

$$0 \quad 0111111011 \quad 101100110011 \quad \dots \quad 1001$$

(round to nearest)

$$0 \quad 0111111011 \quad 101100110011 \quad \dots \quad 1010$$

(round down or round towards 0).

# Absolute roundiy error

(7)

$$| \text{round}(x) - x |$$

in double precision, if

$$x = \pm (1. b_1 b_2 \dots b_{53} \dots) \times 2^E$$

$E$  is in the range  $-1022$  to  $1023$ .

then absolute roundiy error  $< 2^{-52} \times 2^E$

for any roundiy mode.

# Relative roundiy error

$$\frac{| \text{round}(x) - x |}{|x|} \leq \epsilon \quad (\text{~~machine precision~~})$$

(Here  $\epsilon$  is machine precision)

$\therefore$  we can write

$$\text{round}(x) = x(1 + \delta) \quad |\delta| < \epsilon$$

(or  $\frac{\epsilon}{2}$  for round to nearest).

$$a \oplus b = \text{round}(a+b) = (a+b)(1 + \delta_1)$$

$$a \ominus b = \text{round}(a-b) = (a-b)(1 + \delta_2)$$

$$a \otimes b = \text{round}(a \times b) = (a \times b)(1 + \delta_3)$$

$$a \oslash b = \text{round}(a/b) = (a/b)(1 + \delta_4)$$

$$|\delta_i| < \epsilon$$