Week 3.

Rounding : Truncation (Rounding toward zero / Round down)
To the nearest
Toward infinity.

Normalization: $a.bcd \times 10^e$: with only a single digit before decimal point.

Floating-point value < number / location of the radix point.

Significand : the normalized digit part of the value.
In floating-points, the Significand is called Mantissa.

IEEE-754

Range : $1.000\cdots0_2 \times 2^{-e} \sim 1.11\cdots1_2 \times 2^e$.

The significand of an IEEE-754 floating point number is reperesented in sign and magnitude form.

S.EEEEEEEE 1.FFFFFFFFFFFFFFFFFFFFFFF

Sign bit — 8-bit biased exponent — 23-bit fractional significand.

$1 \le E \le 254$.

$x_{10} = (-1)^S \times 2^{(E-B)} \times 1.F$

$E > 0$: normalized

$E = 0$: not normalized (too small to reperesent)

$\Rightarrow x = (-1)^S \times 2^{1-B} \times 0.F$.

$E = 0$ & $F \neq 0$ => Denormalized underflow number.

Rounded:

Truncation = Round to zero.

Convert to decimal:

S = 0 S = +
S = 1 S = −

E = 1111 1111 (255) { F = 0 ∞
F ≠ 0 NaN (not a number)

= 0000 0000 (0) : 2^-126 (1-127).

1 ≤ E ≤ 254 : 2^(E-127).

F: E = 0: 0.------$_{10}$

E ≠ 0: 1.------$_{10}$

Convert to 32-bit IEEE-754 FP

S: + 0, − 1 (-1)^S

E: 2^n: n < -126: too small to be reperesented as a normalized number => reperesent in an un-normalized form => underflow.

=> exponent = -126 => E = 0000 0000

Round the number to 23 bits nearest

* if the rounded number is at the midway, keep the last digit 0

e.g. 000 0000 0000 0000 0000

0001 1000 | 0000 1000

=> 0010 1000 | => 0000 0000

$n > 127$: too big to be reperesented, encoded as +inf.

$\Rightarrow F = 000\ 0000\ 0000\ 0000\ 0000\ 0000$

$E = 1111\ 1111_2$.

$-126 \le n \le 127$: convert to $(n+127)_2$

F : $E = -126$: $0.\text{----}_2$

> -126 : $1.\text{----}_2$