Floating Point Form of Numbers:

In decimal notation, every real number is represented by a finite or an infinite sequence of decimal digits. Most of computers have two ways of representing numbers called Fixed Point and Floating Point.

In fixed point system, all numbers are given with a fixed number of digits after decimal (and sometimes also before decimal).

Standard Form:

IIII.FFFF (Decimal System)

9999.9999 (largest number)

0000.0001 (Smallest number)

For examples, 62.3248, 0.0142, 0.1000 etc.

Fixed point system (representation) is impractical.

In floating point system, we write a number in the form

$$x = \pm m.10^{e} \rightarrow \text{exponent}$$

$\uparrow$ mantisa

where $0.1 \leq |m| < 1$, and $e \in I$.

For examples, $0.6247.10^{3}$, $0.1735.10^{-13}$, $-0.2000.10^{-1}$, etc.

In floating point system, the number of significant digits is kept fixed whereas the decimal point is floating.

Significant Digits (Rules)

(1) Non-zero digits are always significant.

(2) Any zero between two significant digits are significant.

(3) A final zero or trailing zeros in the decimal portion ONLY are significant.

Example:

Number	No. of Significant Digits
406	3
0.00500	3
0.03040	4
136000	6
0.001360	4

Round-Off

An error caused by chopping or rounding is called rounding error or round-off error.

Let $x = 0.d_1 d_2, \ldots, d_k d_{k+1} d_{k+2} \ldots \times 10^e$ be a real number. The floating point form of x, denoted by $fl(x)$, is obtained by terminating the mantissa of x at k decimal digits either by chopping or rounding.

→ The chopping produces

$$fl(x) = 0.d_1 d_2 \ldots d_k \times 10^e.$$

→ The rounding adds $5 \times 10^{e-(k+1)}$ to x and then chops the result to obtain a number of the form

$$fl(x) = 0.\delta_1 \delta_2 \delta_3 \ldots \delta_k \times 10^e.$$

→ For rounding, when $d_{k+1} \geq 5$, we add 1 to d_k, and discards the digits after k^{th} place to obtain $fl(x)$, i.e., we round up.

→ When $d_{k+1} < 5$, we simply chop off all but the first k digits, so we round down.

If we round down, then $\delta_i = d_i$ for $i = 1, 2, \cdots k$. However, if we round up, the digits (and even the exponent) might change.

Example: $\pi = 3.141592655\ldots$

Normalized decimal form: $\pi = 0.31415926555\ldots \times 10^1$

(a) Floating point form of π using 5-digit Chopping is

$$fl(x) = 0.31415 \times 10^1 = 3.1415$$

(b) Floating Point form of π using 5-digit rounding is

$$fl(x) = 0.31416 \times 10^1 = 3.1416$$