Number Systems

1. Representation of Integers – We use decimal number system which consists of digits 0, 1, 2, ..., 9.

e.g $1256 = 1\times10^3 + 2\times10^2 + 5\times10^1 + 6\times10^0$

Here 10 is the base of this system.

∴ 1256 is expressible as a polynomial in the base 10 with integral coefficients lies between 0 and 9.

2. Decimals

$$.1256 = 1\times10^{-1} + 2\times10^{-2} + 5\times10^{-3} + 6\times10^{-4}$$

Now $42.965 = 4\times10^1 + 2\times10^0 + 9\times10^{-1} + 6\times10^{-2} + 5\times10^{-3}$

Floating-Point Representation of Numbers

$$\frac{8}{3} = 2.6666\ldots = 2\times10^0 + 6\times10^{-1} + 6\times10^{-2} + 6\times10^{-3} + \ldots$$

Now Number Systems are of the following forms.

Name	Base	Digits
1. Decimal Number System	10	0, 1, 2, ..., 9
2. Binary System	2	0, 1
3. Octal "	8	0, 1, 2, ..., 7
Hexadecimal "	16	0, 1, 2, ..., 9 and A, B, C, D, E, F.

We represent base by β

∴ From above, $42.965 = 4\times\beta^1 + 2\times\beta^0 + 9\times\beta^{-1} + 6\times\beta^{-2} + 5\times\beta^{-3}$

$(111.011)_2 = 1\times2^2 + 1\times2^1 + 1\times2^0 + 0\times2^{-1} + 1\times2^{-2} + 1\times2^{-3}$.

$(127)_8 = 1\times8^2 + 2\times8^1 + 7\times8^0$

Now $2.666.6\ldots = 0.2666\ldots\times10^1$

$= 0.02666\ldots\times10^2$

$= 0.002666\ldots\times10^4$

Defn: A floating point number is a number represented in the form $\cdot d_1 d_2 \ldots d_n \times \beta^e$ where $d_1, d_2, \ldots, d_n$ are integers and satisfy $0 \le d_i < \beta$ and the exponent.

is such that $m \leq e \leq M$.
The fractional part $0.d_1 d_2 \cdots d_n$ is called mantissa and it lies between +1 and -1.

$\therefore fl(x) = \pm (0.a_1 a_2 \cdots a_n)_\beta \times \beta^e$

Defn For each real number x, we associate a floating point representation denoted by $fl(x)$, given by

$$fl(x) = \pm (0.a_1 a_2 \cdots a_n)_\beta \times \beta^e$$

where β based fraction, i.e., $(0.a_1 a_2 \cdots a_n)_\beta$ is called mantissa with all a_i's integers and e is known as exponent. This representation is called **β-based floating point representation of x.**

$$42.965 = 0.42965 \times 10^2$$

$$0 = 0.000\cdots0 \times \beta^e$$

$$-0.00234 = -0.234 \times 10^{-2} \quad \leftarrow \text{(Not unique)}$$

$(\because = -0.0234 \times 10^{-1})$

Defn: (Normal form) - A non-zero floating-point number is in normal form if the values of mantissa lies in $(-1, -\frac{1}{\beta}]$ or $[\frac{1}{\beta}, 1)$

If $\beta = 10$ then $\frac{1}{\beta} = 0.$

e.g $0.4296 \in [0.1, 1)$

$-0.234 \in (-1, -0.1]$

Or A floating-point number $(0.a_1 a_2 \cdots a_n)_\beta \times \beta^e$ is said to be normal if $a_1 \neq 0$

Also the range of exponent is also restricted. There are integers m and M such that $-m \leq e \leq M$.

Rounding and Chopping: Let x be any real number and $fl(x)$ be its machine approximation. There are two ways to do cutting to store a real number.

$$x = \pm (0.a_1 a_2 \cdots a_n a_{n+1} a_{n+2} \cdots)_\beta \times \beta^e, \; a_1 \neq 0$$

(i.e normal form)

Take $fl(x)$ as the approximation upto only n-digits.

(i) Chopping: $fl(x) = \pm(0.a_1 a_2 \cdots a_n)_\beta \times \beta^e$

(ii) Rounding:

$$fl(x) = \begin{cases} \pm(0.a_1 a_2 \cdots a_n)_\beta \times \beta^e, & 0 \le a_{n+1} < \beta/2 \quad \text{(rounding down)} \\ \pm(0.a_1 a_2 \cdots a_n + 1)_\beta \times \beta^e, & \frac{\beta}{2} \le a_{n+1} < \beta \quad \text{(rounding up)} \end{cases}$$

$$\downarrow$$

$$\pm\left[(0.a_1 a_2 \cdots a_n)_\beta + (0.000\cdots01)_\beta\right] \times \beta^e$$

Exp $x = 6/7 = 0.85714 \times 10^0$

Chopping and rounding off upto two decimal places,

$$fl(x) = \begin{cases} 0.85 \times 10^0 & \text{(chopping)} \\ 0.86 \times 10^0 & \text{(rounding)} \end{cases}$$

Rules for rounding off numbers

(1) If digit to be dropped is greater than 5, the last retained digit is increased by one. For example, 12.6 is rounded to 13.

(2) If the digit to be dropped is less than 5, the last retained digit is left as it is. For example, 12.4 is rounded to 12.

(3) If the digit to be dropped is 5 and any digit following it is non-zero, the last remaining digit is increased by one. For example, 12.51 is rounded to 13.

(4) If the digit to be dropped is 5 and is followed by only zeros, the last remaining digit is increased by one if it is odd, but left as it is if even. For example, 11.5 is rounded to 12 and 12.5 is rounded to 12.

Errors: Let x be any real number and $fl(x)$ is its approximation

(1) Absolute Error $= |x - fl(x)|$

(2) Relative Error $= \dfrac{|x - fl(x)|}{|x|}$

$x = 0.33$, $fl(x) = 0.3$

$AE = |x - fl(x)| = 0.03$

$RE = \dfrac{0.03}{0.33} = 0.0909\ldots \approx 0.1 \approx 10^{-1}$ or 10%

$x = 0.33 \times 10^{-5}$, $fl(x) = 0.3 \times 10^{-5}$

$AE = 0.03 \times 10^{-5}$, $R.E = 0.1$ or 10^{-1}

As a measure of accuracy, the absolute error may be misleading and relative error is more meaningful.

Theorem (Error bound theorems in case of chopping and Rounding off)

Let x be any real number in decimal system and $fl(x)$ be its approximation in computer. Then find the error bounds for Absolute and relative errors in case of (i) Chopping and (ii) Rounding off.

Proof (i) **Chopping**

Let $x = (0.a_1 a_2 \cdots a_n a_{n+1} \cdots) \times 10^e$, $a_1 \neq 0$

$$= \left(\sum_{i=1}^{\infty} \frac{a_i}{10^i}\right) \times 10^e, \quad a_1 \neq 0$$

$$fl(x) = (0.a_1 a_2 \cdots a_n) \times 10^e$$

$$= \left(\sum_{i=1}^{n} \frac{a_i}{10^i}\right) \times 10^e$$

$\therefore$ Absolute Error is given by,

$$A.E = |x - fl(x)| = \left(\sum_{i=n+1}^{\infty} \frac{a_i}{10^i}\right) \times 10^e$$

Since each ~~$a_i \leq 9$ in chopping~~ $a_i < 10$ or $a_i \leq (10-1) = 9$

$$\therefore |x - fl(x)| \leq \left(\sum_{i=n+1}^{\infty} \frac{9(10-1)}{10^i}\right) \times 10^e$$

$$= \left(\sum_{i=n+1}^{\infty} \frac{9}{10^i}\right) \times 10^e = 9\left(\sum_{i=n+1}^{\infty} \frac{1}{10^i}\right) \times 10^e$$

$$= 9 \times \left(\frac{1}{10^{n+1}} + \frac{1}{10^{n+2}} + \cdots\right) \times 10^e$$

$$= 9 \times \frac{\frac{1}{10^{n+1}}}{1 - \frac{1}{10}} \times 10^e$$

(Geometric Series with $a = \frac{1}{10^{n+1}}$ & $r = \frac{1}{10}$, Since $|r| = |\frac{1}{10}| < 1$ $\therefore$ Converge)

$$= 9 \times \frac{\frac{1}{10^{n+1}}}{\frac{9}{10}} \times 10^e = \frac{1}{10^n} \times 10^e$$

$$= 10^{e-n}$$

Now $|x| = (0.a_1a_2 \dots a_n a_{n+1} \dots)_{10} \times 10^e$

$\geq \frac{1}{10}(0.1) \times 10^e = \frac{1}{10} \times 10^e = 10^{e-1}$

$\Rightarrow \frac{1}{|x|} \leq \frac{1}{10^{e-1}}$

∴ Relative error,

$$R.E = \frac{|x - fl(x)|}{|x|} \leq \frac{10^{e-n}}{10^{e-1}} = 10^{1-n}$$

(ii) Rounding off

$$fl(x) = \begin{cases} (0.a_1a_2 \dots a_n)_{10} \times 10^e = \sum_{i=1}^{n} \frac{a_i}{10^i} \times 10^e, & 0 \leq a_{n+1} < 5 \\ (0.a_1a_2 \dots [a_n+1])_{10} \times 10^e, & a_{n+1} \geq 5 \end{cases}$$

Also $a_i \leq 9$ $\forall$ $i = 1, 3, \dots$ $i \neq n+1$

Case I For $a_{n+1} < 5$ i.e $a_{n+1} \leq 4$

$$fl(x) = (0.a_1a_2 \dots a_n)_{10} \times 10^e = \sum_{i=1}^{n} \frac{a_i}{10^i} \times 10^e$$

Now $|x - fl(x)| = \left(\sum_{i=n+1}^{\infty} \frac{a_i}{10^i}\right) \times 10^e = \left[\frac{a_{n+1}}{10^{n+1}} + \sum_{i=n+2}^{\infty} \frac{a_i}{10^i}\right] \times 10^e$

$$\leq \left(\frac{4}{10^{n+1}} + \sum_{i=n+2}^{\infty} \frac{9}{10^i}\right) \times 10^e$$

$$= \left(\frac{4}{10^{n+1}} + 9\left[\frac{1}{10^{n+2}} + \frac{1}{10^{n+3}} + \dots\right]\right) \times 10^e$$

$$= \frac{4}{10^{n+1}} + 9 \cdot \left(\frac{\frac{1}{10^{n+2}}}{1 - \frac{1}{10}}\right) \times 10^e$$

$$= \left(\frac{4}{10^{n+1}} + \frac{1}{10^{n+1}}\right) \times 10^e = \left(\frac{5}{10^{n+1}}\right) \times 10^e$$

$$= \frac{5}{10} \times 10^{e-n} = \frac{1}{2} \times 10^{e-n}$$

Case II For $a_{n+1} \geq 5$

$$fl(x) = (0.a_1a_2 \dots [a_n+1])_{10} \times 10^e$$

$$= \left[(0.a_1a_2 \dots a_n)_{10} + (0.0000 \dots 01)_{10}\right] \times 10^e$$

$$= \left(\sum_{i=1}^{n} \frac{a_i}{10^i} + \frac{1}{10^n} \right) \times 10^e$$

Now $|x - fl(x)| = \left| \left(\sum_{i=n+1}^{\infty} \frac{a_i}{10^i} - \frac{1}{10^n} \right) \right| \times 10^e$

$$= \left| \frac{1}{10^n} - \sum_{i=n+1}^{\infty} \frac{a_i}{10^i} \right| \times 10^e$$

$$= \left| \frac{1}{10^n} - \frac{a_{n+1}}{10^{n+1}} - \sum_{i=n+2}^{\infty} \frac{a_i}{10^i} \right| \times 10^e$$

$$\leq \left| \frac{1}{10^n} - \frac{a_{n+1}}{10^{n+1}} \right| \times 10^e \qquad \left(\begin{array}{l} a_{n+1} \geq 5 \\ -a_{n+1} \leq -5 \end{array} \right)$$

$$\leq \left| \frac{1}{10^n} - \frac{5}{10^{n+1}} \right| \times 10^e$$

$$= \left| \frac{1}{10^n} - \frac{1}{2} \frac{1}{10^n} \right| \times 10^e$$

$$= \frac{1}{2} \cdot \frac{1}{10^n} \times 10^e = \frac{1}{2} \times 10^{e-n}$$

Therefore for both Cases I & II,

$$|x - fl(x)| \leq \frac{1}{2} \times 10^{e-n}$$

Now $\dfrac{|x - f(x)|}{|x|} \leq \dfrac{1}{2} \dfrac{10^{e-n}}{10^{e-1}} = \dfrac{1}{2} 10^{1-n}$

Significant figures: Rules for deciding the number of significant figures in measured quantity.

(1) All nonzero digits are significant. e.g 1.234 has 4 significant figures, 1.2 has 2 significant figures.

(2) Zeros between nonzero digits are significant. e.g 1002 has 4 significant figures.

(3) Leading zeros to the left of the first nonzero digit are not significant. e.g 0.001 → 1.01×10^{-1} or 0.1×10^{-2} has only 1 significant figure.

(4) Trailing zeros that are also to the right of a decimal point in a number are significant. e.g: 0.0230 has 3 significant figures. 0.230×10^{-1}, 23.0×10^{-3}

(5) When a number ends in zeros that are not to the right of a decimal point, the zeros are not necessarily significant. e.g 190 may be 2 or 3 significant figures, 50600 may be 3 or 4 or 5 significant figures.

0.50600×10^5 or 0.5060×10^5 or 0.506×10^5

Exact numbers have an infinite number of significant figures. e.g 23 students, 5 km, 500 m

* Significant figures: All measurements are approximations. No measuring device can give perfect measurements without experimental uncertainty. The number of significant figures in a result/calculation is simply the number of figures that are known with some degree of reliability

*(3) Here leading zeros only help to fix the position of the decimal point. Hence these are not significant.

Significant figures

Take x in normalized floating point number. Then

$$x = (0.a_1 a_2 \cdots a_n) \times 10^q$$

No. of digits in mantissa of normal form are the no. of significant digits in that number.

Example.

25.923

$= 0.25923 \times 10^2$

5 digits are significant.

209.3050

$= 0.2093050 \times 10^3$

7 digits are significant.

For exact no.

256

$= 0.256 \times 10^3$

$= 0.2560 \times 10^3$

} At least three significant digits.

~~$x = (0.a_1 a_2 \cdots a_n a_{n+1} \cdots) \times 10$~~

$x \to$ exact value, $fl(x) \to$ Approximation.

$x^* \longrightarrow$ Approximation.

$$\frac{|x - x^*|}{|x|} \leq \frac{1}{2} 10^{1-t} \quad \text{or} \quad \frac{1}{2} 10 \times 10^{-t} = 5 \times 10^{-t}$$

The number x^* is said to approximate x to t-significant digits if t is the largest nonnegative integer for which

$$\frac{|x - x^*|}{|x|} \leq 5 \times 10^{-t}$$

$p = 0.54617$, $q = 0.54601$
Use 4-digit Arithmetic
find $r = p - q$ and find significant digit in approximately
$r = p - q = 0.00016$ (exact value)
$r^* = p^* - q^* = 0.5462 - 0.5460$
$= 0.0002$
R.E $= \frac{|r - r^*|}{|r|}$
$= 0.25 \leq 0.5$
$= 5 \times 10^{-1}$
1-Significant digit.

Example $x = 25.923$, $x^* = 25.92$

$$\frac{|x - x^*|}{|x|} = \frac{0.003}{25.923} = 0.000115727 \leq 0.0005$$

$$= \frac{5}{10^4} = 5 \times 10^{-4}$$

$\Rightarrow x^* = 25.92$ is approximating x to 4-significant digits.

The number x^* is said to approximate x correct upto ~~the~~ to n decimal place if

$$|x - x^*| \leq \frac{1}{2} 10^{-n}$$

~~2005~~

e.g $x = 25.9235$ $\quad x^* = 25.92$

$$|x - x^*| = 0.0035 \leq 0.005 = 5 \times 10^{-3} = \frac{5}{10} \times 10^{-2} = \frac{1}{2} 10^{-2}$$

$\Rightarrow x^*$ is approximating x ~~correct~~ upto 2 decimal places.

$x = 25.9263$, $x^* = 25.93$

$|x - x^*| = 0.0037 \leq 0.005 = \frac{1}{2} \times 10^{-2}$

Arithmetic Operations

(Finite-digit Arithmetic)

(Do all operations)

① **Addition.** Let x and y be two numbers.

$$x \oplus y = fl(fl(x) + fl(y))$$

② **Subtraction,** $x \ominus y = fl(fl(x) - fl(y))$

③ **Multiplication** $x \otimes y = fl(fl(x) \times fl(y))$

④ **Division** $x \oslash y = fl(fl(x) \div fl(y))$

Use 4-digit rounding arithmetic to find $x+y$, $x-y$, $x \times y$, $x \div y$

$x = \frac{5}{7} = 0.714285714285\ldots$, $fl(x) = 0.7143$

$y = \frac{1}{3} = 0.333333\ldots$, $fl(y) = 0.3333$

$x + y = fl(0.7143 + 0.3333) = fl(1.0476)$

$= 1.048$

$x - y = fl(0.7143 - 0.3333) = fl(0.3810) = 0.3810$

$x \times y = fl(0.7143 \times 0.3333) = fl(0.2380\ldots)$

$= 0.2381$

$x \div y = fl(0.7143 \div 0.3333) = fl(2.1431143)$

$= 2.143$