CSE330: Numerical Methods

Topic: Representation of Numbers in Computer System and Human System

Prepared by:

Saad Bin Sohan

BRAC University

Email: sohan.academics@gmail.com
GitHub: https://github.com/saad-bin-sohan

CSE330

Numerical methods

Faculty name : Md Aquib Azmain

Initial : AQU

email : aquib.azmain@bracu.ac.bd

Room : 80615

For Course Curriculum

Quiz- 10% (Best 4 out of 6)

Assignment- 15% (Best 4 out of 6)

Mid - 20%

Final - 30%

Lab - 20%

Attendence - 5%

All resource you need:

Numerical Analysis-II
(Lecture Notes) - Anthony Yeates

Date: 26/sept/23

Week-1, Lecture-1

Topic: Representation of numbers in computer system (and human system)

Floating Point Arithmetic

Fixed point representation (human representation of numbers)

any number, $X = \pm \left(d_1 d_2 \dots d_{k-1} \overset{\text{floating point}}{\cdot} d_k \dots d_n\right)_{\underset{\text{base}}{\beta}}$

where $d_1, d_2, d_3 \ldots, d_n \in \{0, 1, \ldots, \beta - 1\}$

উদাহরণ: $$X = + (10.1)_2$$

Floating point representation (computer representation)

There are 3 conventions আছে এই representation করা হয়।

Convention -1 (or general form)

any number,

$$F = \pm \underbrace{(0.d_1 d_2 d_3 \dots d_m)}_{\text{mantissa}}{}_{\underset{\hookrightarrow \text{base}}{\beta}} \times \beta^{\overset{\text{exponent}}{\overset{\uparrow}{e}}}$$

mantissa = fraction part

এই convention-এ $d_1 = 1$ (always)

$\beta, d_i, e \in \mathbb{Z}$ (always integer হবে)

$0 \leq d_i \leq \beta - 1$

$e_{min} \leq e \leq e_{max}$ (to be confirmed)

d_i হলো single digit

(fixed point থেকে difference টা আসলো β^e এর জন্য)

/// convention-1 এ $d_1 = 1$ হওয়ার কারণ যেন there is a unique representation for each number.

যেমন: $(1001.11)_2 \times 2^2$ কে

$(0.100111)_2 \times 2^6$ এবং

$(0.00100111)_2 \times 2^8$ এই both ভাবে represent logical হলেও computer এতো different representation কনসিডার করতে চায় না।

এজন্য $(1001.11)_2 \times 2^2$ এর convention 1 অনুযায়ী representation হলো

$$(0.100111)_2 \times 2^6$$

Que-1

$\beta = 2$, $e_{min} = -1$

$m = 3$, $e_{max} = 2$

(i) highest possible positive number কত?

(ii) lowest possible non-negative number কত?

(iii) lowest possible negative number কত?

Solution :

(i) $+(0.111)_2 \times 2^2$

(ii) $+(0.100)_2 \times 2^{-1}$

(iii) $-(0.111)_2 \times 2^2$

[★ so far Zero represnt possible না convention-1 ↵]

NB: decimal এর জন্য convention 1 এ

d_1 হবে $1 \le d_1 \le 9$

Convention-2 (or normalized form)

any number,

$$F = \pm (1.d_1 d_2 \cdots d_m)_\beta \times \beta^e$$

যেখানে

$\beta, d_i, e \in \mathbb{Z}$ (always integer হবে)

$0 \le d_i \le \beta - 1$

d_i হলো single digit

$$e_{min} \le e \le e_{max}$$

Que-2 $\beta = 2$, $e_{min} = -1$

$m = 3$, $e_{max} = 2$

convention-2 তে,

(i) highest possible positive number কত?

(ii) lowest possible non-negative number কত?

(iii) lowest possible negative number / lowest possible number considering signed bit কত?

solution :

(i) $+(1.111)_2 \times 2^2$

(ii) $+(1.000)_2 \times 2^{-1}$

(iii) $-(1.111)_2 \times 2^2$

Convention-3 (de-normalized form)

any number,

$$F = \pm (0.1 d_1 d_2 \cdots d_m) \times \beta^e$$

যেখানে

$\beta, d_i, e \in Z$ (always integer হবে)

$0 \leq d_i \leq \beta - 1$

d_i হলো single digit

$\boxed{\text{Que-3}}$ $\beta = 2$, $e_{min} = -1$

$m = 3$, $e_{max} = 2$

convention-3 তে,

(i) highest possible positive number কত?

(ii) lowest possible non-negative number কত?

(iii) lowest possible negative number / lowest possible number considering signed bit কত?

$\boxed{\text{solution}}$:

(i) $+(0.1\underbrace{111}_{m=3}) \times 2^{2}$

(ii) $+(0.1\underbrace{000}_{m=3}) \times 2^{-1}$

(iii) $-(0.1111)_2 \times 2^{2}$

Que-4

$\beta = 2$, $e_{min} = -1$

$m = 3$, $e_{max} = 2$

Total কতগুলো নাম্বার represent করা possible এই কম্পিউটার দিয়ে and number গুলো কী কী?

Solution:

convention- 1

এই চারটা number for $e = -1$

$(0.100)_2 \times 2^{-1} =$

$(0.101)_2 \times 2^{-1}$

$(0.110)_2 \times 2^{-1}$

$(0.111)_2 \times 2^{-1}$

$e = 0, 1, 2$ এর জন্য $4 \times 3 = 12$ টা নাম্বার হবে হবে

• • •

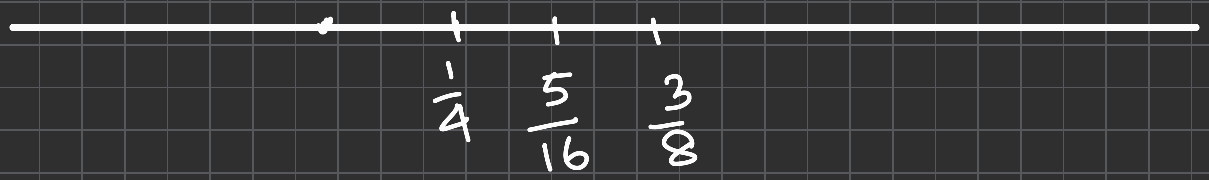
• • •

• • •

এরকম total 16টা positive number হবে হবে। (neg গুলো xm এ ধরবো না)

নাম্বারগুলো বের করার পর সংখ্যাগুলোকে decimal এ represent করতে হবে।

Topic: Decimal number গুলোকে number line এ represent করা



NB: zero এবং যে number গুলো decimal এ convert করা যায় যাইনি, তাদের লেখা যায় না।

IEEE Standards

(How computer Operates is in the denormalised form of numbers. But before that, computer converts every number to normalized form, then to denormalized form. And then the operations are done in de-norm form.)

the IEEE standards are

$\beta = 2$

52 bits for mantissa/fraction part

11 bits for exponents

1 bit for sign

(+)

64 bits in total (thats the standard how computers are designed)

The normalised form (part in IEEE Standards)

$$\pm(1.d_1 d_2 \cdots d_{52}) \times 2^e$$

$e_{min} = 0$ $\qquad$ $e_{max} = 2^{11} - 1$

$$= 2047$$

largest non-negative number,

$$= +(1.\underbrace{111\ldots\ldots111}_{\text{52টি 1}}) \times 2^{2047}$$

smallest non negative number,

$$= +(1.\underbrace{000\ldots\ldots00}_{\text{52টি zero}}) \times 2^{0}$$

topic: exponent biasing

2^{11} এ 2048 টা bit এর zero একটা (যেটা non negative এর মধ্যে), তাই negative -এ আছে-এর অর্ধেক থেকে 1 কম টা

$((2048/2)-1) = 1023$ টা.

positive এ $(2048/2) = 1024$ টা

zero একটা

negative এ $(2048/2)-1 = 1023$ টা

exponent biasing -এ,

$$\Rightarrow (1.d_1 d_2 d_3 \dots d_{52}) \times 2^{\overset{0 \le e \le 2047}{e} - 1023}$$