Number representation

Shubham Raghuvanshi
Harish Chandra Research Institute

April 3, 2021

Representation of real numbers

- A number representation is a bijective map from set of real numbers to set of mathematical symbols called digits.
- In a positional number system with base b, a number is represented by an ordered set of integers $d_n d_{n-1},,,, d_0.d_{-1},,,, d_{-r+1} d_{-r}$ such that $0 \leq d_i \leq b-1$ whose magnitude is given by

$$\sum_{i=-r}^{n} d_i b^i$$

- This is like representing locations on a matrix by locations on one dimentional array e.g. a location on a chess board can be either given as a two digit number (d_2, d_1) $0 \leq d_i \leq 7$, or as a point on the real number line whose location is $p = i + 8 * j$.
- With n+m+1 digits in base b one can represent b^{n+m+1} real numbers. However the magnitude of the number depends on the location of the radix point specified by the offset 'r'.

Representation of real numbers

- A number in base b is equal to a number in base B if both of them have same representation on real line i.e. $\sum_{i=-m}^{n} d_i b^i = \sum_{i=-m}^{n} D_i B^i$
- The above equation can be used to convert a number from one base b to B. To do this we can calculate $\sum_{i=-m}^{n} d_i b^i$ but do the arithematic in base B for example

$$(10101.11)_2 = 1 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 + 1 \times 2^{-1} + 1 \times 2^{-2}$$

$$= (21.75)_{10}$$

Arithematic of real numbers in general base

- In order to do calculations in general base 'b', we need arithematic in that base. For this purpose we define a class "number", the objects of which are supposed to be numbers in base b.
- Each number is represented by vector of unsigned integers $0 \leq d_i \leq b-1$. In this system a digit can be represented by one of the possible b symbols. i.e. for base 16 the digits will be denoted by the symbols $|0|, |1|......|10|, |12|, |13|, |14|, |15|$. This generalization is helpful when we want to work in higher base.
- Two "numbers" are added digit wise in and the carry is forwarded. If the result of addition is greater than b-1 the carry is set to 1.
- In order to subtract n2 from n1, the complement of n2 is added to n1 and if the final carry is 1 i.e. there is overflow, the result is positive otherwise negative in the later case the result of addition is complemented to get final answer. e.g. in base 10 $(3-7) \rightarrow 3+(7)_c = 6 \rightarrow (6)_c \rightarrow -4$

High bases

- Although it is much simlper to do calculations in system with low base e.g. binary. The number of digits required to represent a number are also larger even for small numbers.
- Representation of a rational number in binary will terminate only if the denominator is a power of 2. Similarly in decimal the representation will terminate only when the denominator can be factorized in terms of 2 and 5.
- Number represented in high base can be used to store high precision values and do arithematic with high precision. As an application of which we attemps to calculate value of π accurate upto few hundred digits in base 10^5

Estimation of Pi

- We use the following expression

$$\pi = 16 \tan^{-1}(1/5) - 4 \tan^{-1}(1/239)$$

and do all the calculations in base 10^5 to get accurate value of $\tan^{-1}$ given by the Maclaurin series expansion

$$\tan^{-1}(1/x) = \sum_{i=0}^{\infty} (-1)^i \frac{1}{(2i+1)x^{2i+1}}$$