DOI: https://doi.org/10.3126/cognition.v6i1.64445

Navigating Number Systems for Computational Precision: An Analytical Exploration

Ravindra Mishra[1]

Abstract:-

The main objective of this article is to study all numbers that have occurred so far in the theoretical discussions were real (or complex) numbers in the strict mathematical sense. That is, they were to be conceived as infinite decimal fractions, or as Dedekind cuts. For the purposes of computation such numbers have to be approximated by real numbers of a rather special type, such as terminating decimal fractions, or other rational numbers. The present article is devoted to a study of the number systems that can be used for the purpose of computation.

Keywords:- Infinite decimal fraction, Integers, number system, binary system, decimal conversion, rational number.

Introduction:-

1.1 Representation of the integers:

Let us take a look at our conventional number system. What do we mean by a symbol such as 247? Evidently

$$247 = 2 \cdot 100 + 4 \cdot 10 + 7$$
$$= 2 \cdot 10^2 + 4 \cdot 10^1 + 7 \cdot 10^0.$$

Here's the breakdown of representation of 247:

$$247 = 2 \cdot 100 + 4 \cdot 10 + 7$$

This expression breaks down 247 into its hundreds, tens, and ones place. In the decimal system, the place values increase by powers of 10 from right to left: ones, tens, hundreds, thousands, and so on.

$$2 \cdot 100 + 4 \cdot 10 + 7$$

This line further simplifies the expression by showing that 100 is equivalent to 10×10, which is why it's written as 10^2. Similarly, 10 is just 10^1 and 7 is 10^0.

So, 247 is represented as a polynomial in the base 10, where the coefficients (2, 4, and 7) are integral (whole) numbers ranging from 0 to 9 (since the digits in the decimal system range from 0 to 9), which can be expressed as $10 - 1$.

This representation helps understand how the number is composed of multiples of powers of 10, reflecting its place value system in the decimal system.

1. Mathematics Teacher in Baneshwor Champus & Patan Multiple Campus, TU, Email:ravindramishra004@gamil.com

Received on Oct. 28, 2023 *Accepted on Dec. 19, 2023* *Published on Jan. 31, 2024*

There is no intrinsic reason why 10 should be used as a base; the number of fingers may have to do with it. There is evidence that in cultures different from ours other number systems have been used. The French word quatre-vingts for the number 80 indicates a system with base 20. (Maybe the French counted with their toes as well as with their fingers.) In New Zealand, words for 11^2 and 11^3 have been found. The Babylonian astronomers used a sexagesimal system, i.e., a system with the base 60. A trace of this can be found in our dividing the circumference of the circle into 360 degrees. Also mixed systems, although mathematically much less satisfying, are in use, such as the Anglo-Saxon system for measuring length, and the English monetary system. In electronic computation, the digits of an integer are represented by various states of the physical quantity such as electric current. The technically simplest situation arises when there are only two states to be represented, such as the state "no current" and the state "a unit current." For this reason, modern electronic computers work internally almost exclusively with the base 2. The resulting number system is called the binary system. In this system, only the digits 0 and 1 occur. In order to distinguish them from decimal digits, we shall underline them. Thus, if a given nonnegative integer N is in the binary system represented in the form

$$N = a_n 2^n + a_{n-1} 2^{n-1} + \ldots\ldots\ldots + a_1 2^1 + a_0 2^0 \quad (1)$$

Where the a_1 are either zero or one, it will be written in the form

$$a_n a_{n-1} \ldots a_1 a_0$$

Sample:-

1=1, 2=10, 3=11, 101=5, 8=1000, 1010=10.

If we wish to communicate with a computer working in the binary system, we (or the computer) must be able to convert a number from the decimal to binary system and conversely. To convert from binary to decimal, we regard the number N given by (1) as the value of the polynomial

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots\ldots a_1 x + a_0$$

For x = 2. To evaluate P(2),we may use algorithm.(Note that the coefficients are differently numbered now.) It follows that if we calculated the numbers b_x recursively by

(2) …….. $b_0 = a_n$, $b_k = a_{n-k} + 2b_{k-1}$ (for K = 1, 2, …, n),

Then $b_n = P(2) = N$.

2. To express the number N = 11111001111 in decimal. The Horner

Scheme yields

K	0	1	2	3	4	5	6	7	8	9	10
a_{n-k}	1	1	1	1	1	0	0	1	1	1	1
b_k	1	3	7	15	31	62	124	249	499	999	1999

It follows that N = 1999.

To convert a given integer from decimal to binary, we make use of the fact that the last binary digit a_o of an integer N is zero if and only if N is even. The second binary digit a_1 is zero if and only if $(n - a_0)/2$ is even, and so on. This leads to the following scheme:

Algorithm (a) To find the binary representation (1) of a given positive integer N we let

(3)….. $N_0 = N,$ $N_{k+1} = \frac{N_k - a_k}{2},$ K = 0, 1, 2,...

Where

(4)…… $a_k = \begin{matrix} 1, & if\ N_k\ is\ odd, \\ 0, & if\ N_k\ is\ even. \end{matrix}$

Continue until $N_k = 0$.

3. To express N = 1999 in binary form. Algorithm (a) yields the scheme

K	0	1	2	3	4	5	6	7	8	9	10
N_k	1999	999	499	249	124	62	31	15	7	3	1
a_k	1	1	1	1	0	0	1	1	1	1	1

It follows that 1999 = 11111001111. (Note that the least significant digits are obtained first) The scheme is an exact reversal of the scheme of the 2.

1.2 Binary Fraction:

A binary fraction is a series of the form

(5)……. $z = \sum_{k=1}^{\infty} a_{-k} 2^{-k},$

Where the coefficients a_{-1}, a_{-2},… are either zero or one, the series (5) always converges, because it is majorized by the geometric series

$$\sum_{k=1}^{\infty} 2^{-k} = \frac{1}{2}\frac{1}{1-\frac{1}{2}} = 1.$$

The sum z of (5) will also be denoted by

$$z = 0.a_{-1}a_{-2}\,a_{-3}\,\ldots$$

The binary fraction (5) is said to terminate if , for some integer n, $a_k = 0$, k > n.

The following theorem is fundamental, but will not be proved:-

Statement: *Any real number z, 0 < z ≤ 1 can be represented in a unique manner by a non terminating binary fraction.*

If we drop the condition that the binary fraction shall not terminate, then the representation may not be unique; for instance, the binary fractions

0.1 and 0.01111…..

both represent the number 0.5.

A termination binary fraction $z = 0.a_{-1}a_{-2}\ldots a_{-n}$ *can be regarded as the value of the polynomial*

$$P(x) = a_{-1}x + a_{-2}x^2 + \ldots a_{-n}x^n$$

At $x = \frac{1}{2}$ and thus can be evaluated by algorithm (Horner's scheme) as follows: Let

$$b_0 = a_{-n}, \qquad b_k = a_{-n+k} + \frac{1}{2} b_{k-1}, \qquad k = 1, 2, \ldots, n,$$

where $a_0 = 0$, Then $b_n = z$.

Here is an example,

To express z = 0.00110011 in decimal. Horner's scheme yields

K	a_{n-k}	b_k
0	1	1
1	1	1.5
2	0	0.75
3	0	0.375
4	1	1.1875
5	1	1.59375
6	0	0.796875
7	0	0.3984375
8	0	0.17721875

It follows that z = 0.19921875.

Another method for converting a terminating binary fraction consists in converting the integer

$2^n z = a_{-1}2^{n-2} + a_{-z}2^{n-z} + a_{-z}2^{n-z} + \ldots + a_{-n}$; and dividing the result by 2^n.

Except in special circumstances, non-terminating binary fractions cannot be converted into terminating decimal fractions. To get an approximate decimal representation, we truncate an infinite binary fraction after the n^{th} digit and convert the resulting terminating fraction. The error in this approximation will be less than 2^{-n}.

The inverse problem of converting a given (decimal) fraction into binary fraction is solved by the following algorithm:-

Algorithm: For a real number z such that $0 \leq z \leq 1$, calculate the sequence (z_k) and (a_{-k}) recursively by the relations

$$z1 = z$$

$$z_1 = z$$

(6)……. $$a_{-k} = \begin{cases} 1, & if\ 2z_k > 1, \\ 0, & if\ 2z_k \leq 1, \end{cases}$$

$$z_{k+1} = 2z_k - a_{-k}, \qquad K = 1, 2, \ldots\ldots$$

Reference:-

1. Aiken, A.C. [1926]: On Bernoulli's numerical solution of algebric equations, Proc. Roy. Soc. Edinburgh. 49, 289-305.
2. Rareiss, E. H. [1960]: Resultant procedure and the mechanization of the Graeffe process, J. Assoc. Comp. Mach., 7, 346-386.
3. Birkhoff, G., and S. MaceLane [1953]: *A survey of modera algebra*, rev. ed., Macmillan, New York.
4. Brown, K. M., and P. Henrici [1962]: Sign wave analysis in matrix eigenvalue problems, Math of Comput., 16, 291-300.
5. Comrie, L. J. [1961]: Chamberi's shorter six-figure mathematical tables, W. R. Chambers Ltd., Edinburgh and London.
6. Hildebrand, F.B.[1956]: *Introduction to numerical analysis*, McGraw-Hill, New York, Toronto, London.
7. Huskey, H. D.[1949]: On the precision of a certain procedure of numerical integration, with an appendix by Douglas R. Hartree, J. Res. Nat. Bur. Stand., 42, 57 – 62
8. Jahnke, E., and F. Emde[1945]: Tables of functions with formulae and curves, Dover, New York.