Cambridge University Press & Assessment
978-1-108-71600-0 — Numerical Methods
Rajesh Kumar Gupta
Excerpt
More Information

Number Systems

Chapter 1

All the mathematical sciences are founded on relations between physical laws and laws of numbers, so that the aim of exact science is to reduce the problems of nature to the determination of quantities by operations with numbers.

In a few years, all great physical constants will have been approximately estimated, and that the only occupation which will be left to men of science will be to carry these measurements to another place of decimals.

James Clerk Maxwell
(June 13, 1831–November 5, 1879)
He pioneered the classical theory of "Electromagnetism".

1.1 Introduction

In everyday life, we are habituated to doing arithmetic using numbers based on the decimal system. Any number in the decimal number system, say for example, 349.15, can be expressed as a polynomial in the base or radix 10 with integral coefficients 0 to 9.

$$(349.15)_{10} = 3\times10^{2} + 4\times10^{1} + 9\times10^{0} + 1\times10^{-1} + 5\times10^{-2}$$

In number 349.15, 349 is an integral part and .15 is a fractional part. Note that the subscript (10) in the number $(349.15)_{10}$ denotes the base of the number system.

There is no intrinsic reason to use the decimal number system. Computers read electrical signals, and the state of an electrical impulse is either on or off. Hence, binary system, with base 2 and with integer coefficients 0 and 1, is convenient for computers. However, most computer users prefer to work with the familiar decimal system. It is cumbersome to work with the binary number system, as a large number of binary digits are required to represent even a moderate-sized decimal number. Hence the octal and hexadecimal number systems are also used for this purpose. If the base is two, eight or sixteen, the number is called as the binary, octal or hexadecimal number, respectively. Any number $x = (a_n a_{n-1} \ldots a_1 a_0 . b_1 b_2 \ldots)_\beta$ with base β can be represented as follows

$$x = a_n\beta^n + a_{n-1}\beta^{n-1} + \ldots + a_1\beta + a_0\beta^0 + b_1\beta^{-1} + b_2\beta^{-2} \ldots \quad (1.1)$$

The number system with base β contains numbers from 0 to β–1. For examples, decimal number system, with base 10, contains numbers from 0 to 9. Similarly, binary system, with base 2, contains numbers 0 and 1.

Table 1.1 Binary, Octal, Decimal and Hexadecimal Numbers

Binary Base: 2 Digits: 0, 1	Octal Base: 8 Digits: 0, 1, 2, 3, 4, 5, 6, 7	Decimal Base: 10 Digits: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9	Hexadecimal Base: 16 Digits: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F
0000	00	00	0
0001	01	01	1
0010	02	02	2
0011	03	03	3
0100	04	04	4
0101	05	05	5
0110	06	06	6
0111	07	07	7
1000	10	08	8
1001	11	09	9
1010	12	10	A
1011	13	11	B
1100	14	12	C
1101	15	13	D
1110	16	14	E
1111	17	15	F

To work with the computer-preferred binary and the people-preferred decimal, and also with the octal and hexadecimal number systems, it is imperative to have algorithms for conversion from one number system to another number system. In the next two sections, some algorithms are discussed to convert the integral and fractional parts of a number from one number system to another number system.

1.2 Representation of Integers

The arithmetic for various number systems with some examples has been discussed in this section. We will use this for conversion of integers in different number systems.

Example 1.1

Explore the addition and multiplication in the decimal, binary, octal and hexadecimal number systems with some examples.

Decimal Arithmetic (For base 10, digits are 0 … 9)
$(1295)_{10} + (357)_{10} = (1652)_{10}$
$(734)_{10} \times (46)_{10} = (33764)_{10}$

Binary Arithmetic (For base 2, digits are 0 and 1)
$(101011)_2 + (11011)_2 = (1000110)_2$
$(11101)_2 \times (1001)_2 = (100000101)_2$

Octal Arithmetic (For base 8, digits are 0 … 7)
$(1635)_8 + (274)_8 = (2131)_8$
$(752)_8 \times (23)_8 = (22136)_8$

Hexadecimal Arithmetic (For base 16, digits are 0 … 9, A, B, C, D, E, F)
$(5AB7)_{16} + (F63)_{16} = (6A1A)_{16}$
$(A4B)_{16} \times (7A)_{16} = (4E7BE)_{16}$

Note: Arithmetic for numbers with base β:
Consider the addition of two numbers $(1635)_8$ and $(274)_8$ in the octal number system with the base $\beta = 8$. Note that, the addition of numbers 5 and 4 will produce number 9. For $\beta = 8$, we have 1 carry, and the remaining number is 1. Similarly, other calculations give the following result

$$\begin{array}{r l} 1\ 1\ 1 & \text{Carry} \\ (1\ 6\ 3\ 5)_8 & \\ +\ (2\ 7\ 4)_8 & \\ \hline (2\ 1\ 3\ 1)_8 & \end{array}$$

$\Rightarrow (1635)_8 + (274)_8 = (2131)_8$

Similarly, consider the multiplication of two numbers. For example, multiplication of numbers 7 and 5 will produce number 35. In octal system (base $\beta = 8$), for number 32, we have 4 carry; and remaining is 3. So, final result is $(7)_8 \times (5)_8 = (43)_8$.

1.2.1 Conversion from Any Number System to the Decimal Number System

Conversion from any number system to the decimal form may be obtained directly from the definition (1.1)

$$x = a_n\beta^n + a_{n-1}\beta^{n-1} + \ldots + a_1\beta + a_0\beta^0 + b_1\beta^{-1} + b_2\beta^{-2} \ldots$$

Some of the examples are as follows

Example 1.2

$$(1101.101)2 = 1\times 2^3 + 1\times 2^2 + 0\times 2^1 + 1\times 2^0 + 1\times 2^{-1} + 0\times 2^{-2} + 1\times 2^{-3} = (13.625)_{10}$$

$$(347.623)8 = 3\times 8^2 + 4\times 8^1 + 7\times 8^0 + 6\times 8^{-1} + 2\times 8^{-2} + 3\times 8^{-3} = (231.787109375)_{10}$$

$$\begin{aligned}(\text{A5F.B42})16 &= 10\times 16^2 + 5\times 16^1 + 15\times 16^0 + 11\times 16^{-1} + 4\times 16^{-2} + 2\times 16^{-3}\\ &= (2655.70361328125)_{10}\end{aligned}$$

1.2.2 Conversion between Binary, Octal and Hexadecimal Number Systems

Conversion in the binary, octal and hexadecimal can be accomplished easily since four/three binary digits make one hexadecimal/octal digit, respectively. To convert from the binary to the octal/hexadecimal, we have to partition the binary digits in groups of three/four (starting from right in an integral part and from left in fractional part) and then replace each group by its equivalent octal/hexadecimal digit. To convert from octal and hexadecimal, we have to replace all digits by their binary equivalents.

Example 1.3

$$(1101.101)_2 = (001\ 101.\ 101) = (\underbrace{001}_{1}\ \underbrace{101}_{5}.\ \underbrace{101}_{5}) = (15.5)_8$$

$$(1101.101)_2 = (1101.\ 1010) = (\underbrace{1101}_{D}.\ \underbrace{1010}_{A}) = (\text{D.A})_{16}$$

$$(347.623)_8 = (\underbrace{011}_{3}\ \underbrace{100}_{4}\ \underbrace{111}_{7}.\ \underbrace{110}_{6}\ \underbrace{010}_{2}\ \underbrace{011}_{3}) = (11100111.110010011)_2$$

$$(\text{A5F.B42})_{16} = (\underbrace{1010}_{A}\ \underbrace{0101}_{5}\ \underbrace{1111}_{F}.\ \underbrace{1011}_{B}\ \underbrace{0100}_{4}\ \underbrace{0010}_{2}) = (101001011111.101101000010)_2$$

1.2.3 Conversion from Decimal Number System to Any Other Number System

The conversion of the integer N in decimal number system to another number system can be easily obtained in a systematic manner described as follows. Let there be a number N with base β

$$N = a_n\beta^n + a_{n-1}\beta^{n-1} + \ldots + a_1\beta + a_0$$

Division by the base β will give

$$\frac{N}{\beta} = a_n\beta^{n-1} + a_{n-1}\beta^{n-2} + \ldots + a_1 + \frac{a_0}{\beta}$$

The digit a_0 is the remainder after the base β divides the number N. Let us consider the above equation in the form

$$\frac{N}{\beta} = N_0 + \frac{a_0}{\beta}, \text{ where } N_0 = a_n\beta^{n-1} + a_{n-1}\beta^{n-2} + \ldots + a_1$$

On dividing N_0 by base β, we get

$$\frac{N_0}{\beta} = a_n\beta^{n-2} + a_{n-1}\beta^{n-3} + \ldots + \frac{a_1}{\beta}$$

The number a_1 is the remainder. We can continue the process till the quotient is 0. Apparently, the conversion from decimal number system to a number system with base β can be achieved by the following algorithm.

$$N = \beta N_0 + a_0$$

$$N_0 = \beta N_1 + a_1$$

$$N_1 = \beta N_2 + a_2$$

$$\vdots$$

till the quotient is 0.

Example 1.4

Convert the decimal number $(231)_{10}$ into its binary equivalent.

Ans.

$231 = 115 \times 2 + 1$	$N_0 = 115$	$a_0 = 1$
$115 = 57 \times 2 + 1$	$N_1 = 57$	$a_1 = 1$
$57 = 28 \times 2 + 1$	$N_2 = 28$	$a_2 = 1$
$28 = 14 \times 2 + 0$	$N_3 = 14$	$a_3 = 0$
$14 = 7 \times 2 + 0$	$N_4 = 7$	$a_4 = 0$
$7 = 3 \times 2 + 1$	$N_5 = 3$	$a_5 = 1$
$3 = 1 \times 2 + 1$	$N_6 = 1$	$a_6 = 1$
$1 = 0 \times 2 + 1$	$N_7 = 0$	$a_7 = 1$

Thus the binary representation of the decimal number $(231)_{10}$ is $(11100111)_2$. It can be computed from the expression $(a_n a_{n-1} \ldots a_1 a_0)_2$.

Example 1.5

Compute the hexadecimal equivalent of the decimal number $(2655)_{10}$.

Ans.

$2655 = 165 \times 16 + 15$	$N_0 = 165$	$a_0 = (15)_{10} = (F)_{16}$
$165 = 10 \times 16 + 5$	$N_1 = 10$	$a_1 = (5)_{10} = (5)_{16}$
$10 = 0 \times 16 + 10$	$N_2 = 0$	$a_2 = (10)_{10} = (A)_{16}$

So, $(A5F)_{16}$ is hexadecimal equivalent of $(2655)_{10}$.

1.2.4 Conversion from One Number System to Any Other Number System

So far, we have discussed the algorithms for conversion of integers in some number systems. The following recursive algorithm can be utilized for conversion of integers in any general number systems.

Algorithm 1.1

Consider a number N with the coefficients $a_n, a_{n-1}, \ldots, a_0$

$$N = a_n\beta^n + a_{n-1}\beta^{n-1} + \ldots + a_1\beta + a_0\beta^0$$

Calculate the following numbers $b_n, b_{n-1}, \ldots, b_0$ recursively using

$$b_n = a_n$$

$$b_i = a_i + b_{i+1}\beta, \; i = n-1, \; n-2, \ldots, 0$$

Then $b_0 = N$.

Example 1.6

Convert the binary number $(110111)_2$ into its decimal equivalent.

Ans.

$(110111)_2 = 1 \times 2^5 + 1 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0$

Since the conversion is from binary to decimal, we will use decimal arithmetic for this conversion. Note that each digit in the following calculation is in decimal number system.

$$b_5 = a_5 = 1$$
$$b_4 = a_4 + b_5\beta = 1 + 1 \times 2 = 3$$
$$b_3 = 0 + 3 \times 2 = 6$$
$$b_2 = 1 + 6 \times 2 = 13$$
$$b_1 = 1 + 13 \times 2 = 27$$
$$b_0 = 1 + 27 \times 2 = 55$$

Example 1.7

Compute the binary equivalent of the decimal number $(231)_{10}$ using recursive algorithm 1.1.

Ans.

$(231)_{10} = 2 \times 10^2 + 3 \times 10^1 + 1 \times 10^0 = (10)_2 \times (1010)_2^{\ 2} + (11)_2 \times (1010)_2 + (1)_2$

This conversion uses binary arithmetic as follows

$$b_2 = a_2 = (10)_2$$
$$b_1 = a_1 + b_2\beta = (11)_2 + (10)_2 \times (1010)_2 = (10111)_2$$
$$b_0 = a_0 + b_1\beta = (1)_2 + (10111)_2 \times (1010)_2 = (11100111)_2$$

Example 1.8

Compute the octal equivalent of the decimal number $(231)_{10}$.

Ans.

$(231)_{10} = 2 \times 10^2 + 3 \times 10^1 + 1 \times 10^0 = (2)_8 \times (12)_8^{\ 2} + (3)_8 \times (12)_8 + (1)_8$

On using octal arithmetic in the Algorithm 1.1, we have

$$b_2 = a_2 = (2)_8$$
$$b_1 = a_1 + b_2\beta = (3)_8 + (2)_8 \times (12)_8 = (27)_8$$
$$b_0 = a_0 + b_1\beta = (1)_8 + (27)_8 \times (12)_8 = (1)_8 + (346)_8 = (347)_8$$

Example 1.9

Convert the decimal number $(2655)_{10}$ into hexadecimal number.

Ans.

$(2655)_{16} = 2 \times 10^3 + 6 \times 10^2 + 5 \times 10^1 + 5 \times 10^0$

$$= (2)_{16} \times (A)_{16}^3 + (6)_{16} \times (A)_{16}^2 + (5)_{16} \times (A)_{16} + (5)$$

$$b_3 = a_3 = (2)_{16}$$

$$b_2 = a_2 + b_3\beta = (6)_{16} + (2)_{16} \times (A)_{16} = (6)_{16} + (14)_{16} = (1A)_{16}$$

$$b_1 = a_1 + b_2\beta = (5)_{16} + (1A)_{16} \times (A)_{16} = (5)_{16} + (104)_{16} = (109)_{16}$$

$$b_0 = a_0 + b_1\beta = (5)_{16} + (109)_{16} \times (A)_{16} = (5)_{16} + (A5A)_{16} = (A5F)_{16}$$

1.3 Representation of Fractions

In a number system with base β, the fractional part can always be written as follows

$$x_F = \sum_{k=1}^{\infty} b_k \beta^{-k}$$

where b_k is a non-negative integer less than the number β. If $b_k = 0$ for all k greater than a positive integer, then the fractional part is said to be terminating otherwise non-terminating. For example $\frac{1}{4} = 0.25$ is terminating, while $\frac{1}{6} = 0.166666...$ is non-terminating. Conversion of the fractional part from one number system to another number system can be achieved with the help of the following algorithm.

Algorithm 1.2

On multiplying the fraction $x_F = \sum_{k=1}^{\infty} b_k \beta^{-k} = .b_1 b_2 b_3 \ldots$ with base β, we get

$$\beta x_F = \sum_{k=1}^{\infty} b_k \beta^{-k+1} = b_1 + \sum_{k=1}^{\infty} b_{k+1} \beta^{-k}$$

Thus the number b_1 is an integral part of the product βx_F. On repeating the process, we find that b_2 is an integral part of $\beta(\beta x_F)_F$, b_3 is an integral part of $\beta(\beta(\beta x_F)_F)_F$ and so on. One can easily conclude the following algorithm for a general base β from the procedure above.

Let $c_0 = x_F$

$$b_1 = (\beta c_0)_I, \qquad c_1 = (\beta c_0)_F$$
$$b_2 = (\beta c_1)_I, \qquad c_2 = (\beta c_1)_F$$
$$\vdots$$

where subscript I denotes an integral part, while subscript F denotes the fractional part.

Example 1.10

Calculate the binary equivalent of the decimal number $(.3125)_{10}$ using the recursive algorithm 1.2.

Ans.

Let $c_0 = (.3125)_{10}$

$$2(.3125)_{10} = (.6250)_{10} \qquad b_1 = 0 \qquad c_1 = (.6250)_{10}$$
$$2(.6250)_{10} = (1.250)_{10} \qquad b_2 = 1 \qquad c_2 = (.250)_{10}$$
$$2(.250)_{10} = (.50)_{10} \qquad b_3 = 0 \qquad c_3 = (.50)_{10}$$
$$2(.50)_{10} = (1.00)_{10} \qquad b_4 = 1 \qquad c_4 = (0)_{10}$$

The binary equivalent of $(.3125)_{10}$ is $(.b_1b_2b_3b_4)_2 = (.0101)_2$. This example has a terminating binary fraction, but not each terminating decimal fraction will give a terminating binary fraction, and this is true for other number systems also.

Example 1.11

Find the binary equivalent of the decimal number $(0.3)_{10}$.

Ans.

Let $c_0 = (.3)_{10}$

$$2(.3)_{10} = (.6)_{10} \qquad b_1 = 0 \qquad c_1 = (.6)_{10}$$
$$2(.6)_{10} = (1.2)_{10} \qquad b_2 = 1 \qquad c_2 = (.2)_{10}$$
$$2(.2)_{10} = (.4)_{10} \qquad b_3 = 0 \qquad c_3 = (.4)_{10}$$
$$2(.4)_{10} = (.8)_{10} \qquad b_4 = 0 \qquad c_4 = (.8)_{10}$$
$$2(.8)_{10} = (1.6)_{10} \qquad b_5 = 1 \qquad c_5 = (.6)_{10}$$
$$\vdots$$

Since the digits are repeating, we can conclude that the binary equivalent of $(.3)_{10}$ is a non-terminating fraction $(.0\ 1001\ 1001\ 1001\ \ldots)_2$ (or) $(.0\overline{1001})$

Cambridge University Press & Assessment
978-1-108-71600-0 — Numerical Methods
Rajesh Kumar Gupta
Excerpt
More Information

Example 1.12

Find the decimal representation of the binary number $(.0101)_2$.

Ans.

Using the algorithm 1.2 and binary arithmetic, we get

$c_0 = (.0101)_2$

$(1010)_2(.0101)_2 = (11.0010)_2$ $\quad b_1 = (11)_2 = (3)_{10}$ $\quad c_1 = (.0010)_2$

$(1010)_2(.0010)_2 = (1.010)_2$ $\quad b_2 = (1)_2 = (1)_{10}$ $\quad c_2 = (.010)_2$

$(1010)_2(.010)_2 = (10.10)_2$ $\quad b_3 = (10)_2 = (2)_{10}$ $\quad c_3 = (.10)_2$

$(1010)_2(.10)_2 = (101.0)_2$ $\quad b_4 = (101)_2 = (5)_{10}$ $\quad c_4 = (0)_2$

Hence $(.3125)_{10}$ is decimal equivalent of the binary fraction $(.0101)_2$.

Example 1.13

Convert the octal fraction $(.71)_8$ to its equivalent decimal representation.

Ans.

Let $c_0 = (.71)_8$

$(12)_8(.71)_8 = (10.72)_8$ $\quad b_1 = (10)_8 = (8)_{10}$ $\quad c_1 = (.72)_8$

$(12)_8(.72)_8 = (11.04)_8$ $\quad b_2 = (11)_8 = (9)_{10}$ $\quad c_2 = (.04)_8$

$(12)_8(.04)_8 = (0.5)_8$ $\quad b_3 = (0)_8 = (0)_{10}$ $\quad c_3 = (.5)_8$

$(12)_8(.5)_8 = (6.2)_8$ $\quad b_4 = (6)_8 = (6)_{10}$ $\quad c_4 = (.2)_8$

$(12)_8(.2)_8 = (2.4)_8$ $\quad b_5 = (2)_8 = (2)_{10}$ $\quad c_5 = (.4)_8$

$(12)_8(.4)_8 = (5.0)_8$ $\quad b_6 = (5)_8 = (5)_{10}$ $\quad c_6 = (0)_8$

The decimal representation is $(.890625)_{10}$.

Example 1.14

Convert the hexadecimal fraction $(.B4)_{16}$ to its equivalent decimal representation.

Ans.

Let $c_0 = (.B4)_{16}$

$(A)_{16}(.B4)_{16} = (7.08)_{16}$ $\quad b_1 = (7)_{16} = (7)_{10}$ $\quad c_1 = (.08)_{16}$

$(A)_{16}(.08)_{16} = (0.5)_{16}$ $\quad b_2 = (0)_{10}$ $\quad c_2 = (.5)_{16}$