Chapter 2

Numbers and Errors

2.1. Find the binary representations of the numbers (a) $a = 32170$; (b) $b = 7/16$; (c) $c = 75/128$.

Solution: (a) $(32170)_{10} = (111110110101010)_2$; (b) $\frac{7}{16} = (.0111)_2$; (c) $\frac{75}{128} = (.1001011)_2$. Details:

(a)

$32170 = 2{\cdot}16085 + 0$
$16085 = 2{\cdot}8042 + 1$
$8042 = 2{\cdot}4021 + 0$
$4021 = 2{\cdot}2010 + 1$
$2010 = 2{\cdot}1005 + 0$
$1005 = 2{\cdot}502 + 1$
$502 = 2{\cdot}251 + 0$
$251 = 2{\cdot}125 + 1$
$125 = 2{\cdot}62 + 1$
$62 = 2{\cdot}31 + 0$
$31 = 2{\cdot}15 + 1$
$15 = 2{\cdot}7 + 1$
$7 = 2{\cdot}3 + 1$
$3 = 2{\cdot}1 + 1$
$1 = 2{\cdot}0 + 1$

(b)

$\frac{7}{16} = 0.\mathrm{a}_{-1}\mathrm{a}_{-2}\ldots$

$2{\cdot}\frac{7}{16} = \frac{7}{8}$, a_{-1}=0;
$2{\cdot}\frac{7}{8} = 1\frac{3}{4}$, a_{-2}=1;
$2{\cdot}\frac{3}{4} = 1\frac{1}{2}$, a_{-3}=1;
$2{\cdot}\frac{1}{2} = 1$, a_{-4}=1;
$2{\cdot}0 = 0$

$\frac{7}{16} = (.0111)_2$

(c)

$\frac{75}{128} = 0.\mathrm{a}_{-1}\mathrm{a}_{-2}\ldots$

$2{\cdot}\frac{75}{128} = 1\frac{11}{64}$, a_{-1}=1;
$2{\cdot}\frac{11}{64} = \frac{11}{32}$, a_{-2}=0;
$2{\cdot}\frac{11}{32} = \frac{11}{16}$, a_{-3}=0;
$2{\cdot}\frac{11}{16} = 1\frac{3}{8}$, a_{-4}=1;
$2{\cdot}\frac{3}{8} = \frac{3}{4}$, a_{-5}=0;
$2{\cdot}\frac{3}{4} = 1\frac{1}{2}$, a_{-6}=1;
$2{\cdot}\frac{1}{2} = 1$, a_{-7}=1;
$2{\cdot}0 = 0$

$\frac{75}{128} = (.1001011)_2$

2.2. Find infinite repeating binary representations of the following numbers: (a) $a = 1/3$; (b) $b = 1/10$; (c) $c = 1/7$.

Solution: (a) $\frac{1}{3} = (.0101\ldots)_2$; (b) $\frac{1}{10} = (.000110011\ldots)_2$; (c) $\frac{1}{7} = (.001001\ldots)_2$. Details:

(a)

$2{\cdot}\frac{1}{3} = \frac{2}{3}$, a_{-1}=0;
$2{\cdot}\frac{2}{3} = 1\frac{1}{3}$, a_{-2}=1;
$2{\cdot}\frac{1}{3} = \frac{2}{3}$, a_{-3}=0;
a cycle occurs.

$\frac{1}{3} = (.0101\ldots)_2$

(b)

$2{\cdot}\frac{1}{10} = \frac{1}{5}$, a_{-1}=0;
$2{\cdot}\frac{1}{5} = \frac{2}{5}$, a_{-2}=0;
$2{\cdot}\frac{2}{5} = \frac{4}{5}$, a_{-3}=0;
$2{\cdot}\frac{4}{5} = 1\frac{3}{5}$, a_{-4}=1;
$2{\cdot}\frac{3}{5} = 1\frac{1}{5}$, a_{-5}=1;
$2{\cdot}\frac{1}{5} = \frac{2}{5}$, a_{-6}=0;
a cycle occurs

$\frac{1}{10} = (.000110011\ldots)_2$

(c)

$2{\cdot}\frac{1}{7} = \frac{2}{7}$, a_{-1}=0;
$2{\cdot}\frac{2}{7} = \frac{4}{7}$, a_{-2}=0;
$2{\cdot}\frac{4}{7} = 1\frac{1}{7}$, a_{-3}=1;
$2{\cdot}\frac{1}{7} = \frac{2}{7}$, a_{-4}=0;
a cycle occurs.

$\frac{1}{7} = (.001001\ldots)_2$

2.3. Find the binary representation of 10^6 and $\frac{4}{3}$.
Solution: $10^6 = (11110100001001000000)_2$;
$\frac{4}{3} = (1.010101\ldots)_2$.
Details:

$10^6 = 5^6 \cdot 2^6;\ 2^6 = (1000000)_2$
$5^6 = 7812 \cdot 2 + 1$
$7812 = 3906 \cdot 2 + 0$
$3906 = 1953 \cdot 2 + 0$
$1953 = 976 \cdot 2 + 1$
$976 = 488 \cdot 2 + 0$
$488 = 244 \cdot 2 + 0$
$244 = 122 \cdot 2 + 0$
$122 = 61 \cdot 2 + 0$
$61 = 30 \cdot 2 + 1$
$30 = 15 \cdot 2 + 0$
$15 = 7 \cdot 2 + 1$
$7 = 3 \cdot 2 + 1$
$3 = 1 \cdot 2 + 1$
$1 = 0 \cdot 2 + 1$
$10^6 = (11110100001001)_2 \cdot (1000000)_2$
$= (11110100001001000000)_2$.

$\frac{4}{3} = 1 + \frac{1}{3}$
$\frac{1}{3} = (0.1_{-1}a_2\ldots)_2$
$2 \cdot \frac{1}{3} = \frac{2}{3}, a_{-1} = 0$
$2 \cdot \frac{2}{3} = 1\frac{1}{3}, a_{-2} = 1$
$2 \cdot \frac{1}{3} = \frac{2}{3}, a_{-3} = 0$
a cycle occurs

$\frac{4}{3} = (1.010101\ldots)_2$

2.4. Convert $(473)_{10}$ into a number with the base (a) 2; (b) 6; (c) 8.
Solution: (a) $(473)_{10} = (111011001)_2$; (b) $(473)_{10} = (2105)_6$; (c) $(473)_{10} = (731)_8$.
Details:

(a)	(b)	(c)
$473 = 2{\cdot}236 + 1$	$473 = 6{\cdot}78 + 5$	$473 = 8{\cdot}59 + 1$
$236 = 2{\cdot}118 + 0$	$78 = 6{\cdot}13 + 0$	$59 = 8{\cdot}7 + 3$
$118 = 2{\cdot}59 + 0$	$13 = 6{\cdot}2 + 1$	$7 = 8{\cdot}0{+}7$
$59 = 2{\cdot}29 + 1$	$2 = 6{\cdot}0 + 2$	
$29 = 2{\cdot}14 + 1$		
$14 = 2{\cdot}7 + 0$	$(473)_{10} = (2105)_6$	$(473)_{10} = (731)_8$
$7 = 2{\cdot}3 + 1$		
$3 = 2{\cdot}1 + 1$		
$1 = 2{\cdot}0 + 1$		
$(473)_{10} = (111011001)_2$		

2.5. Find the decimal representations of the following numbers: (a) $a = (101011010101)_2$; (b) $b = (16341)_8$; (c) $c = (4523)_6$.
Solution: (a) $(101011010101)_2 = (2773)_{10}$. $(1{\cdot}2^{11} + 1{\cdot}2^9 + 1{\cdot}2^7 + 1{\cdot}2^6 + 1{\cdot}2^4 + 1{\cdot}2^2 + 1{\cdot}2^0 = 2773)$
(b) $(16341)_8 = (7393)_{10}$. $(1{\cdot}8^4 + 6{\cdot}8^3 + 3{\cdot}8^2 + 4{\cdot}8^1 + 1{\cdot}8^0 = 7393)$
(c) $(4523)_6 = (1059)_6$. $(4{\cdot}6^3 + 5{\cdot}6^2 + 2{\cdot}6^1 + 3{\cdot}6^0 = 1059)$

2.6. Prove that any number 2^{-n}, where n is a positive integer, can be represented as a n-digit decimal number $0.a_1a_2\ldots a_n$.

Solution: We use mathematical induction for the proof. The induction assumption: $f_n = \frac{1}{2^n}$ can be represented as n-digit decimal number with the last digit equal to 5: $f_n = 0.a_1a_2\ldots a_{n-1}5$.

1. For $n = 1$ we have $f_1 = 1/2 = 0.5$ and the assumption is valid.

2. Suppose that the induction assumption is valid for some integer $n = k$: $f_k = 0.a_1a_2\ldots a_{k-1}5$.

3. Prove that the assumption is valid for $n = k + 1$.

Clearly, $f_{k+1} = f_k/2$. Since $f_k = 0.a_1a_2\ldots a_{k-1}5$ we have

$$f_{k+1} = \frac{0.a_1a_2\ldots a_{k-1}5}{2} = 0.a_1a_2\ldots a_{k-1}\cdot\frac{1}{2} + 0.\underbrace{00\ldots0}_{k-1}5\cdot\frac{1}{2}.$$

Note that $0.a_1a_2\ldots a_{k-1}\cdot\frac{1}{2}$ can be represented as a decimal number with no more than k digits $0.a'_1a'_2\ldots a'_{k-1}a'_k$ (this is true because when multiplied by 10^k this number becomes integer). Since $0.\underbrace{00\ldots0}_{k-1}5\cdot\frac{1}{2} = 0.\underbrace{00\ldots0}_{k-1}25$, we have

$$f_{k+1} = 0.a'_1a'_2\ldots a'_{k-1}a'_k + 0.\underbrace{00\ldots0}_{k-1}25 = 0.b_1b_2\ldots b_k5,$$

and the induction assumption is correct for $n = k+1$. Therefore, by the principle of mathematical induction, the statement is correct for any positive integer n.

2.7. Prove that $2 = 1.999999\ldots$.

Solution: Assume that $2 \neq 1.999999\ldots$. Then $2 - 1.999999\ldots = \varepsilon > 0$. Since $10^{-n} \to 0$ as $n \to \infty$, there exists a positive integer n such that $\varepsilon \geq 10^{-n}$. We have $1.999999\ldots + \varepsilon \geq 1.999999\ldots + 10^{-n} = 2.\underbrace{000\ldots0}_{n}999\ldots > 2$. This yields $2 - 1.999999\ldots < \varepsilon$, which contradicts to $2 - 1.999999\ldots = \varepsilon$.

2.8. For a decimal integer with m digits, how many bits are needed for its binary representation?

Solution: Let $N > 0$ be an m-digit decimal integer. Then $10^{m-1} \leq N < 10^m$. From equation (2.1) at page 39 we know that

$$\log_2(N+1) \leq n+1 \leq \log_2 N + 1,$$

where n is the number of digits in the binary representation of N. We have:

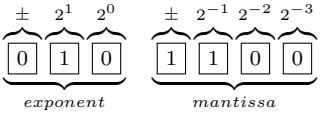
$$\log_2(10^{m-1}+1) \leq \log_2(N+1) \leq n+1 \leq \log_2 N + 1 < \log_2(10^m) + 1.$$

Since $\log_2(10^{m-1}+1) > \log_2(10^{m-1}) = (m-1)\log_2 10$, we have

$$(m-1)\log_2 10 - 1 < n < m\log_2 10.$$

Thus, for large m we have $n \approx m\log_2 10 \approx 3.3219m$.

2.9. Create a hypothetical binary floating-point number set consisting of 7-bit words, in which the first three bits are used for the sign and the magnitude of the exponent, and the last four are used for the sign and magnitude of the mantissa. On the sign place, 0 indicates that the quantity is positive and 1, negative. For example, the number represented in the following figure is $-(.100)_2 \times 2^{(1\times 2^1)} = -(1\times 2^{-1})\times 4 = -2$.



(a) How many numbers are in this set?

(b) Draw a line and show the decimal equivalents of all numbers in the set on this line. What is the distance between two consecutive numbers?

(c) Use your illustration to determine the unit round-off for this system.

Solution:

(a) There are 3 bits for the exponent, each can have two possible values (0 or 1). However, + and - 0 are the same, so there are $2^3 - 1 = 7$ different values for the exponent. Similarly, there are $2^4 - 1 = 15$ different values for the mantissa. One of them is 0, which, when multiplied by any exponent, is still 0. Hence, the set contains $7 * 15 - 6 = 99$ numbers.

(b) Draw a line and show the decimal equivalents of all numbers in the set on this line. What is the distance between two consecutive numbers?

(c) Use your illustration to determine the unit round-off for this system.

2.10. Consider the quadratic equation $ax^2 + bx + c = 0$. Its roots can be found using the formula

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Let $a = 1, b = 100 + 10^{-14}$, and $c = 10^{-12}$. Then the exact roots of the considered equation are $x_1 = -10^{-14}$ and $x_2 = -100$.

(a) Use the above formula to compute the first root on a computer. What is the relative round-off error?

(b) Use the following equivalent formula for the first root:

$$x_1 = \frac{-2c}{b + \sqrt{b^2 - 4ac}}.$$

What do you observe?

Solution:

(a) We observe an error resulting from subtractive cancellation (subtracting b from $\sqrt{b^2 - 4ac}$, which are very close). Computing the first root in MATLAB, we obtain

```
>> x1=(-100-10^(-14)+sqrt((100+10^(-14))^2-4*10^(-12)))/2

x1 =
    -7.1054e-15
```

and the relative round-off error is $\frac{|-10^{-14}+7.1054\cdot 10^{-15}|}{|-10^{-14}|} = 0.2895 = 28.95\%$.

(b) Using the equivalent formula we avoid the subtractive cancellation. We obtain x_1=-1.0000e-14 (exact solution).

```
>> z=-2*10^(-12)/(100+10^(-14)+...
sqrt((100+10^(-14))^2-4*10^(-12)))

x1 =
   -1.0000e-14
```

2.11. Use a computer to perform iterations in the form $x_{k+1} = (0.1x_k - 0.2)30, k \geq 0$ starting with $x_0 = 3$. You could use, e.g., a spreadsheet application. What do you observe after (a) 50 iterations; (b) 100 iterations? Explain.

Solution: Using Microsoft Excel, we obtain $x_{50} = 199256721.8$ and $x_{100} \approx 1.43046 \times 10^{32}$ istead of the expected values $x_k = 3$. This error propagation happens because the error resulting from inexact representation of fractional numbers (such as 0.2) in the binary system is amplified in each iteration.