Round-off Errors and Computer Arithmetic

1- Binary Machine Numbers

A 64-bit (binary digit) representation is used for a real number (according to IEEE standards).

$$(-1)^s 2^{c-1023}(1+f)$$

This representation is called floating point representation.

The first bit is a sign indicator, denoted *s*. This is followed by an 11-bit exponent, *c*, called the **characteristic**, and a 52-bit binary fraction, *f*, called the **mantissa**. The base for the exponent is 2.

Example 1: consider the following machine number:

0 10000000011 1011100100010000000000000000000000000000000000000000

Sign ↓ Characteristic ↓ Mantissa ↓

Sign: (0: positive; 1:negative)

Characteristic:

$$c = 1\times 2^0 + 1\times 2^1 + 0\times 2^2 + 0\times 2^3 + 0\times 2^4 + 0\times 2^5 + 0\times 2^6 + 0\times 2^7 + 0\times 2^8 + 0\times 2^9 + 1\times 2^{10}$$

$$= 1 + 2 + 1024 = 1027.$$

The exponential part of the number is, therefore:

$$2^{1027-1023} = 2^4$$

Mantissa The final 52 bits is:

$$f = 1\times\left(\frac{1}{2}\right)^1 + 1\times\left(\frac{1}{2}\right)^3 + 1\times\left(\frac{1}{2}\right)^4 + 1\times\left(\frac{1}{2}\right)^5 + 1\times\left(\frac{1}{2}\right)^8 + 1\times\left(\frac{1}{2}\right)^{12}$$

$$(-1)^s 2^{c-1023}(1+f) = (-1)^0 2^{1027-1023}\left(1+\left(1\times\left(\frac{1}{2}\right)^1+1\times\left(\frac{1}{2}\right)^3+1\times\left(\frac{1}{2}\right)^4+1\times\left(\frac{1}{2}\right)^5+1\times\left(\frac{1}{2}\right)^8+1\times\left(\frac{1}{2}\right)^{12}\right)\right)$$

=27.56640625

2- Decimal Machine Numbers

Machine numbers are represented in the normalized *decimal* floating-point form

$\pm 0.d_1d_2 \ldots d_k \times 10^n, \quad 1 \le d_1 \le 9, \qquad 0 \le d_i \le 9$

Any positive real number within the numerical range of the machine can be normalized to the form:

$$y = 0.d_1d_2 \ldots d_kd_{k+1}d_{k+2} \ldots \times 10^n$$

The floating-point form of *y*, denoted *f l(y)*, is obtained by terminating the mantissa of *y* at *k* decimal digits. This can be performed by using one of two methods:

1. **Chopping:**

$$fl(y) = 0.d_1d_2 \ldots d_k \times 10^n$$

2. **Rounding:**

$$fl(y) = 0.\delta_1\delta_2 \ldots \delta_k \times 10^n$$

<u>Example 2</u>: Convert the following numbers to 4-digit by chopping and rounding:

x = 635894 , y = 0.00218 , z = 584.63

1. **Chopping:**

$x^* = 0.6358 \times 10^6, \quad y^* = 0.2180 \times 10^{-2}, \quad z^* = 0.5486 \times 10^3$

2. **Rounding:**

$x^* = 0.6359 \times 10^6, \quad y^* = 0.2180 \times 10^{-2}, \quad z^* = 0.5486 \times 10^3$

<u>Definition 1</u>: Suppose that p^* is an approximation to p. The **absolute error** is $e_p = |p - p^*|$, and the **relative error** is $\delta_p = \frac{|p-p^*|}{|p|}$ provided that $p \neq 0$.

<u>Example 3</u>: Suppose that $x = \frac{5}{7}$ and $y = \frac{1}{3}$. Use five-digit chopping for calculating $x + y$, $x - y$, $x \times y$, and $x \div y$.

<u>Solution</u>:

$$fl(x) = 0.71428 \times 10^0 , \qquad fl(y) = 0.33333 \times 10^0$$

$x \oplus y = fl(fl(x) + fl(y)) = fl(0.71428 \times 10^0 + 0.33333 \times 10^0)$

$= fl(1.04761 \times 10^0)$

$= 0.10476 \times 10^1$.

The true value of addition:

$$x + y = \frac{5}{7} + \frac{1}{3} = \frac{22}{21}$$

The absolute error is:

$$e_{x+y} = \left|\frac{22}{21} - 0.10476 \times 10^1\right| = 0.190 \times 10^{-4}$$

The relative error is:

$$\delta_{x+y} = \left|\frac{0.190\times10^{-4}}{22/21}\right| = 0.182 \times 10^{-4}$$

$x \ominus y = fl(fl(x) - fl(y)) = fl(0.71428 \times 10^0 - 0.33333 \times 10^0)$

$= 0.38095 \times 10^0$

The true value of subtraction:

$$x - y = \frac{5}{7} - \frac{1}{3} = \frac{8}{21}$$

The absolute error is:

$$e_{x-y} = \left|\frac{8}{21} - 0.38095 \times 10^0\right| = 0.238 \times 10^{-5}$$

The relative error is:

$$\delta_{x-y} = \left|\frac{0.238 \times 10^{-5}}{8/21}\right| = 0.625 \times 10^{-5}$$

Multiplication operation:

$x \otimes y = fl(fl(x) \times fl(y)) = fl(0.71428 \times 10^0 \times 0.33333 \times 10^0)$

$= 0.23809 \times 10^0$

$$x \times y = \frac{5}{21}$$

$$e_{x\times y} = \left|\frac{5}{21} - 0.23809 \times 10^0\right| = 0.524 \times 10^{-5}$$

$$\delta_{x\times y} = \left|\frac{0.524 \times 10^{-5}}{5/21}\right| = 0.220 \times 10^{-5}$$

Division operation:

$x \oslash y = 0.21428 \times 10^1$

$$x \div y = \frac{15}{7}$$

$$e_{x\div y} = 0.571 \times 10^{-4}$$

$$\delta_{x\div y} = 0.267 \times 10^{-4}$$

To calculate the upper bounds for the absolute and relative errors, the number has been partitioned into two parts as follows:

$$x = f_x \times 10^k + g_x \times 10^{k-n}$$

Where:

$$\frac{1}{10} \le |f_x| < 1$$

$$0 \le |g_x| < 1$$

In the case of <u>*chopping*</u> the second part is neglected, so we obtain the following approximate value for the number:

$$x^* = f_x \times 10^k$$

The upper bound for the *absolute error* is therefore:

$$|e_x| = |x - x^*| = g_x \times 10^{k-n} < 10^{k-n}$$

The upper bound for the *relative error* is:

$$|\delta_x| = \frac{|e_x|}{|x|} < \frac{10^{k-n}}{|f_x| \times 10^k} \leq 10^{1-n}$$

In the case of *rounding* the approximated value for the number is:

$$x^* = \begin{cases} f_x \times 10^k, & |g_x| < {}^1\!/_2 \\ f_x \times 10^k + 10^{k-n}, & |g_x| \geq {}^1\!/_2 \end{cases}$$

The value of absolute error is:

$$|e_x| = \begin{cases} |g_x| \times 10^{k-n} & |g_x| < {}^1\!/_2 \\ |1 - g_x| \times 10^{k-n} & |g_x| \geq {}^1\!/_2 \end{cases}$$

The upper bound for the *absolute error* is therefore:

$$|e_x| \leq \frac{1}{2} \times 10^{k-n}$$

The upper bound for the *relative error* is:

$$|\delta_x| = \frac{|e_x|}{x} < \frac{1/2 \times 10^{k-n}}{|f_x| \times 10^k} \leq \frac{1}{2} \times 10^{1-n}$$

The above argument can be generalized to any real number in any numerical system with base b. the number is partitioned into two parts:

$$x = f_x \times b^k + g_x \times b^{k-n}$$

The number is converted to the floating point scheme. In the case of chopping the number would be:

$$x^* = f_x \times b^k$$

In the case of rounding the number would be:

$$x^* = \begin{cases} f_x \times b^k, & |g_x| < 1/2 \\ f_x \times b^k + b^{k-n}, & |g_x| \geq 1/2 \end{cases}$$

The upper bounds for the absolute and relative errors:

1. If x^* is chopped:

   $|e_x| < b^{k-n}$

   $|\delta_x| < b^{1-n}$

2. If x^* is rounded:

   $|e_x| \leq \frac{1}{2} \times b^{k-n}$

   $|\delta_x| < \frac{1}{2} \times b^{1-n}$

<u>HW</u>:

1. Find the decimal numbers equivalent to the following floating point binary numbers:

   0 01111111111 0101001100000000000000000000000000000000000000000000

   1 10000001010 1001001100000000000000000000000000000000000000000000

2. Find the absolute and relative errors for the numbers in example 2.

3. For each of the following:

   a. $\frac{4}{5} + \frac{1}{3}$ b. $\frac{4}{5} \times \frac{1}{3}$

   c. $\left(\frac{1}{3} - \frac{3}{11}\right) + \frac{3}{20}$ d. $\left(\frac{1}{3} + \frac{3}{11}\right) - \frac{3}{20}$

   Find:

   i. The exact value.
   ii. The approximated value for 3-digit chopping floating point number. Find the absolute and relative errors.
   iii. The approximated value for 3-digit rounding floating point number. Find the absolute and relative errors.