Numerical Analysis
Number Representation and Errors

Hung-Jui Chang

CYCU Applied Mathematics

September 26, 2022

Representation of Number in Different Base

- Base: the basis of the representation system
  - $\beta - base$: Using numbers from 0 to $\beta - 1$
  - 10-base: 100_{10}
  - 2-base: 1100100_2 (64+32+4=100)
  - The number in the bottom right denotes the base
- Some numbers need to be represented by *infinite* many digits
  - $\sqrt{2} = 1.4142135623730950488 0......$
  - $e = 2.718281828459045235 36...$
  - $\pi = 3.1415926535897932384 6...$
  - $\ln 2 = 0.6931471805599453094 1...$
  - $\frac{1}{3} = 0.3333333333333333333....$
- A real number is separated into the *integer part* and the *fractional part*
  - $(a_n a_{n-1} a_{n-2} \ldots a_1 a_0 . b_1 b_2 b_3 \ldots)_{10} = \sum_{k=0}^{n} a_k 10^k + \sum_{k=1}^{\infty} b_k 10^{-k}$
- Q: Is it possible that a number uses finite digits to be represented in base$-\alpha$ but needs infinite many digits to be represented in base$-\beta$?

Base-β Number

- The base-8 system (Octal system)
  - $(a_n a_{n-1} a_{n-2} \dots a_1 a_0 . b_1 b_2 b_3 \dots)_8 = \sum_{k=0}^{n} a_k 8^k + \sum_{k=1}^{\infty} b_k 8^{-k}$
  - $(21467)_8 = 7 \times 8^0 + 6 \times 8^1 + 4 \times 8^2 + 1 \times 8^3 + 2 \times 8^4$ $= 7 + 8(6 + 8(4 + 8(1 + 8(2))))$
  - $(0.36207)_8 = 3 \times 8^{-1} + 6 \times 8^{-2} + 2 \times 8^{-3} + 0 \times 8^{-4} + 7 \times 8^{-5}$ $= 8^{-5}(7 + 8(0 + 8(2 + 8(6 + 8(3)))))$
- $(a_n a_{n-1} a_{n-2} \dots a_1 a_0 . b_1 b_2 b_3 \dots)_\beta = \sum_{k=0}^{n} a_k \beta^k + \sum_{k=1}^{\infty} b_k \beta^{-k}$
- Note that:
  - the integer part starts from a_0 and ends at a_n
  - the fractional part starts from b_1
    - and don't have a fix ending point.

Conversion of the Integer Part

- $N = (c_n c_{n-1} \ldots c_1 c_0)_\beta = c_0 + \beta(c_1 + \beta(c_2 + \ldots))$
  - c_0 is the remainder of N/β
  - c_1 is the remainder of $((N - c_0)/\beta)/\beta$
  - c_2 is the remainder of $(((N - c_0)/\beta - c_1)/\beta)/\beta$
  - $\vdots$

Quotient	Remainders
$2\overline{)3781}$	
$2\overline{)1890}$	$c_0 = 1$
$2\overline{)945}$	$c_1 = 0$
$2\overline{)472}$	$c_2 = 1$
$2\overline{)236}$	$c_3 = 0$
$2\overline{)118}$	$c_4 = 0$
$2\overline{)59}$	$c_5 = 0$
$2\overline{)29}$	$c_6 = 1$
$2\overline{)14}$	$c_7 = 1$
$2\overline{)7}$	$c_8 = 0$
$2\overline{)3}$	$c_9 = 1$
$2\overline{)1}$	$c_{10} = 1$
0	$c_{11} = 1$

- $(3781)_{10} = (111011000101)_2$

Conversion of the Fractional Part

- Assume $0 < x < 1$
- $x = \sum_{k=1}^{\infty} c_k \beta^{-k} = (0.c_1 c_2 \ldots)_\beta = \frac{1}{\beta}(c_1 + \frac{1}{\beta}(c_2 + \ldots))$
  - F: fractional part, I: Integer part

$$
\begin{array}{ll}
d_0 = x & \\
d_1 = F(\beta d_0) & c_1 = I(\beta d_0) \\
d_2 = F(\beta d_1) & c_2 = I(\beta d_1) \\
d_3 = F(\beta d_2) & c_3 = I(\beta d_2) \\
\vdots & \vdots \\
d_0 = 0.372 & \\
d_1 = F(2 \times 0.372) & c_1 = I(2 \times 0.372) = 0 \\
d_2 = F(2 \times 0.744) & c_2 = I(2 \times 0.744) = 1 \\
d_3 = F(2 \times 0.488) & c_3 = I(2 \times 0.488) = 0
\end{array}
$$

- $(0.372)_{10} = (0.010\ldots)_2$

Base Conversion: $8 \leftrightarrow 2$

Binary(2)	000	001	010	011	100	101	110	111
Octal(8)	0	1	2	3	4	5	6	7

- $x = (a_n a_{n-1} \dots a_2 a_1 a_0)_2 =$
  $a_n 2^n + a_{n-1} 2^{n-1} \dots a_5 2^5 + a_4 2^4 + a_3 2^3 + a_2 2^2 + a_1 2^1 + a_0 2^0$
  $= \dots + (4a_5 + 2a_4 + a_3) 2^3 + (4a_2 + 2a_1 + a_0)$
- $x = (0.b_1 b_2 b_3 \dots)_2 =$
  $b_1 2^{-1} + b_2 2^{-2} + b_3 2^{-3} + b_4 2^{-4} + b_5 2^{-5} + b_6 2^{-6} + \dots =$
  $(4b_1 + 2b_2 + b_3) 2^{-3} + (4b_4 + 2b_5 + b_6) 2^{-6} + \dots$
- $(101\ 101\ 001.110\ 010\ 100)_2 = (551.624)_8$
  - Beard of the $\cdot$ in between the integer and the fractional part.

Base Conversion: 16↔2

Hexadecimal(16)	0	1	2	3	4	5	6	7
Binary(2)	0000	0001	0010	0011	0100	0101	0110	0111
Hexadecimal(16)	8	9	A	B	C	D	E	F
Binary(2)	1000	1001	1010	1011	1100	1101	1110	1111

- Most computer system use base 16 system to represent
- In base-16 system:
  - A:10, B:11, C:12, D:13, E:14, F:15
- $(0111\ 1010\ 1111\ 0010.1100\ 1001\ 1110)_2 = (7AF2.C9E)_{16}$
  - Fill 0 in the front of the integer part or in the end of the fractional if needs.

Programming Exercise

- Write a function with 3 parameters: x, b, n
- Show the decimal number x representing in the base b with n fractional digits
  - b may consider only 2, 8 and 16.
- Example: $x = 10.125,\ b = 2,\ n = 3 \Rightarrow (1010.001)_2$
- You may consider the integer part and fractional part separately.

Normalized Floating Point Number Representation

- $x = \pm 0.d_1 d_2 d_3 \ldots \times 10^n$
  - $123.456 = 1.23456 \times 10^2$
  - $0.002271828 = 2.271828 \times 10^{-3}$
- In the computer system, we use the base 2 system
  - $\pm q \times 2^m = (-1)^s \times 2^{c-127} \times (1.f)_2$
  - q: Normalized mantissa
  - m: Exponent

IEEE Standard Floating-point (IEEE-754)

- Single-precision IEEE Standard Floating-point
- $(-1)^s \times 2^{c-127} \times (1.f)_2$
  - s: sign bit (1) $(+/-)$
  - c: exponent bits (8)
    - use as a base-2 integer (special usage for 0 and 255)
  - f: mantissa bits (23) $(000\ldots000$ to $111\ldots111)$
- $[45DE4000]_{16}$ to floating number
  - $(45DE4000)_{16} = 0100\ 0101\ 1101\ 1110\ 0100\ 0000\ 0000\ 0000$ $\Rightarrow$ $(-1)^0 \times 2^{(10001011)_2-127} \times (1.101\ 1110\ 0100\ 0000\ 0000\ 0000)_2$ $= 2^{12} \times (1.101\ 1110\ 01)_2 = (1\ 101\ 111\ 001\ 000.)_2 =$ $(15710)_8 = 7112$
- Double: $(-1)^2 \times 2^{c-1023} \times (1.f)_2$: 1, 11, 52

Computer Errors

- Let x = q $\times 2^m$ $(\frac{1}{2} \leq q < 1, -126 \leq m \leq 127)$
- x $= (0.1b_2b_3b_4\ldots)_2 \times 2^m$
  - $x_- = (0.1b_2b_3b_4\ldots b_{24})_2 \times 2^m$
  - $x_+ = [(0.1b_2b_3b_4\ldots b_{24})_2 + 2^{-24}] \times 2^m$
  - $x_- \leq x < x_+$
- Either $|x - x_-| \leq \frac{1}{2}|x_+ - x_-| = 2^{-25+m}$ or $|x - x_+| \leq \frac{1}{2}|x_+ - x_-| = 2^{-25+m}$.
- $|\frac{x-x_-}{x}| \leq 2^{-24} = u$ or $|\frac{x-x_+}{x}| \leq 2^{-24} = u$
  - u is the *Unit roundoff error*

Machine Epsilon

- Denote: fl(x) as the *floating point machine number*
  - The corresponding number stored in a machine regarding to x
  - A machine with 5-digit precission, fl(0.37218 71422) $=$ 0.37219
- $\frac{|x - fl(x)|}{|x|} \leq u = 2^{-24}$ for 32-bit floating number
- $fl(x) = x(1+\delta)$, $|\delta| \leq 2^{-24}$
  - If $\epsilon \geq 2^{-23}$, $fl(1+\epsilon) > 1$
  - If $\epsilon < 2^{-23}$, $fl(1+\epsilon) = 1$
- Machine epsilon is the *smallest number* such that $fl(1+\epsilon) > 1$

Error Analysis

- What is the error bound of $z(x+y)$
  - $fl(z(x+y)) = fl(zfl(x+y))$
    $= zfl(x+y)(1+\delta_2),\ |\delta_2| \le 2^{-24}$
    $= z(x+y)(1+\delta_1)(1+\delta_2),\ |\delta_1| \le 2^{-24}$
    $= z(x+y)(1+\delta_1+\delta_2+\delta_1\delta_2),\ (|\delta_1\delta_2| \le 2^{-48})$
    $\approx z(x+y)(1+\delta_1+\delta_2)$
    $= z(x+y)(1+\delta),\ |\delta| \le 2^{-23}$

Exercise: (1/2)

- Use your computer to construct a table of three functions: f, g and h defined as follows.
  - For each integer n in the range 1 to 50, let $f(n) = 1/n$.
  - Then $g(n)$ is computed by adding $f(n)$ to it self $n - 1$ times.
  - Finally, set $h(n) = nf(n)$.
- We want to see the effects of roundoff error.

Excercise: (2/2)

- The harmonic series $1 + \frac{1}{2} + \frac{1}{3} + \ldots +$ is known to diverge to ∞. The n-th partial sum approaches ∞ at the same rate as $\ln(n)$.
- Euler's constant is defined to be $\gamma = \lim_{n\to\infty}[\sum_{k=1}^{n} \frac{1}{k} - \ln(n)] \approx 0.5772$
- Consider the pseudocode:
  - real s, x
  - x$\leftarrow$1.0
  - s$\leftarrow$1.0
  - repeat
    - x$\leftarrow x + 1.0$
    - s$\leftarrow s + 1.0/x$
  - end repeat
- If the loop repeats n times, calculate the value of $s - \ln(n)$
  - Draw a figure with x as n, y as $s - \ln(x)$

Loss of Significance

- Normalize representation and significant Digits
  - $x = \pm r \times 10^n$ where $\frac{1}{10} \leq r < 1$
  - The digits in the fractional part in r is the *significant digits*
  - Example: 0.3721498×10^{-5}
    - significant digits: 3, 7, 2, 1, 4, 9, 8
- Loss of significance means the significant digit decrease during the operation.

- Consider $y \leftarrow x - \sin(x)$
  - Assume the value of x is $\frac{1}{15}$
  - Assume the number of significant digits is 10

$$
\begin{aligned}
x &\leftarrow 0.66666\ 66667 \times 10^{-1} \\
\sin(x) &\leftarrow 0.66617\ 29492 \times 10^{-1} \\
x - \sin(x) &\leftarrow 0.00049\ 37175 \times 10^{-1} \\
x - \sin(x) &\leftarrow 0.49371\ 75000 \times 10^{-4} \\
\frac{1}{15} - \sin(\frac{1}{15}) &\leftarrow 0.49371\ 74327 \times 10^{-4}
\end{aligned}
$$

- The relative error is $\frac{|0.4937175000 \times 10^{-4} - 0.4937174327 \times 10^{-4}|}{|0.4937174327 \times 10^{-4}|} \approx 0.136312788 \times 10^{-6}$

Theorem

Let x and y be normalized floating-point machine numbers, where $x > y > 0$. If $2^{-p} \leq 1 - (y/x) \leq 2^{-q}$ for some positive integers p and q, then at most p and at least q significant binary bits are lost in the subtraction $x - y$.

Part of the proof.

Let $x = r \times 2^n$ and $y = s \times 2^m$, where $\frac{1}{2} \leq r, s < 1$.
Since $x > y$, $y = (s2^{m-n})2^n$
$x - y = \left(r - s2^{m-n}\right) \times 2^n = r\left(1 - \frac{s2^m}{r2^n}\right) = r\left(1 - \frac{y}{x}\right) < 2^{-q}$
When normalize the result of $x - y$, we need to shift q bits left.
At least q significant bits are lost. $\square$

Example of Loss of Significance

- Consider $37.593621 - 37.584216$
  - $1 - \frac{y}{x} = 0.0002501754$
  - This number lies between 2^{-12} and 2^{-11}
  - At least 11 bits, at most 12 bits are lost.

Avoiding Loss of Significance in Subtraction

- Consider $f(x) = \sqrt{x^2+1} - 1$
- When x is closed to zero $1 - \frac{\sqrt{x^2+1}}{1}$ is closed to 0.
    - A potential of loss of significance.
- Transform $\sqrt{x^2+1} - 1$ to $(\sqrt{x^2+1} - 1)\frac{\sqrt{x^2+1}+1}{\sqrt{x^2+1}+1} = \frac{x^2}{\sqrt{x^2+1}+1}$