Week 1, Tuesday

△ $\boxed{\text{problems}} \to \boxed{\text{models}} \to \boxed{\text{algorithms}} \to \boxed{\text{programming}} \to \boxed{\text{results}}$

feedbacks

$\boxed{\text{modelling error}}$ $\boxed{\text{approximation error}}$ $\boxed{\text{roundoff error}}$

↳ core focus in this course.

△ roundoff error

e.g. $2+2=4$ $\quad 2.1\times 3 = 6.3$ $\quad 7-6.15=0.85$ $\quad 4\div 16 = 0.25$ $\quad (\sqrt{3})^2 = 3$

exact arithmetic! but machine only has finite memory!

e.g. $\sqrt{3}$ (1.73205…) is stored as 1.7321 in a calculator if it can only store 5 digits, in this calculator: $(\sqrt{3})^2 = 1.7321^2 = 3.\underline{0002}$ → roundoff error

roundoff error: produced when perform finite-digit arithmetic

Q: how a number is stored in a machine?

△ binary machine numbers (IEEE 754-2008) 二进制

floating-point number: $(-1)^S\, 2^{c-1023}\,(1+f)$ 浮点数

$$\left.\begin{array}{l} S = \begin{cases} 0 & \text{positive} \\ 1 & \text{negative} \end{cases}, \text{ sign part} \\ c:\ 11 \text{ digits, exponential part} \\ f:\ 52 \text{ digits, mantissa part} \end{array}\right\} 64 \text{ digits (double) 双精度}$$

e.g. $\underbrace{0}_{S}\ \underbrace{10000\ 000011}_{c}\ \underbrace{1011\,001\ 0001\ \overbrace{00\cdots0}^{40}}_{f}$

$S=0$, $\quad c = 2^{10}+2^1+2^0 = 1027$, $\quad f = \left(\frac{1}{2}\right)^1+\left(\frac{1}{2}\right)^3+\left(\frac{1}{2}\right)^4+\left(\frac{1}{2}\right)^5+\left(\frac{1}{2}\right)^8+\left(\frac{1}{2}\right)^{12}$

this number is: $(-1)^0\cdot 2^{1027-1023}\cdot(1+f) = 27.56640625$

Rk: 1. range: $-1.79\times 10^{308} \sim 1.79\times 10^{308}$

Q: what is the smallest positive number and largest one?

2. machine error: $2^{c-1023}\times\left(\frac{1}{2}\right)^{53} \sim 2^{c-1023}\ \cdots^{-16}$

Q: what is the smallest positive number and largest one?

2. machine error : $2^{c-1023} \times \left(\frac{1}{2}\right)^{53} \approx 2^{c-1023} \times 10^{-16}$

only 16 significant digits (有效数字 later) in decimal number

△ Decimal machine number 十进制

normalized decimal floating-point form :

$$\pm 0.d_1 d_2 \cdots d_k \times 10^n, \quad 1 \le d_1 \le 9, \quad 0 \le d_i \le 9, \quad i = 2, 3, \cdots, k.$$

For any positive number $y = 0.d_1 d_2 \cdots d_k d_{k+1} d_{k+2} \cdots \times 10^n$

floating-point form :

① chopping : $fl(y) = 0.d_1 d_2 \cdots d_k \times 10^n$, chop off digits $d_{k+1} d_{k+2} \cdots$

② rounding : $fl(y) = 0.\delta_1 \delta_2 \cdots \delta_k \times 10^n$, add $5 \times 10^{n-(k+1)}$ then chop.

(i) $d_{k+1} \ge 5$ $\tilde{d}_k = d_k + 1$ (ii) $d_{k+1} < 5$ $\tilde{d}_k = d_k$ (四舍五入)

e.g. $\pi = 0.314159265 \cdots \times 10^1$

a) five-digit chopping 0.31415×10^1

b) five-digit rounding 0.31416×10^1

△ Def : suppose p^* is an approximation to p

actual error : $p - p^*$;

absolute error : $|p - p^*|$; relative error : $\frac{|p-p^*|}{|p|}$ if $|p| \ne 0$.

Def : Significant digits (figures) $\frac{|p-p^*|}{|p|} \le 5 \times 10^{-t}$, take largest t

we said p^* approximates p to t significant digits.

△ floating-point : relative error $= \left|\frac{y - fl(y)}{y}\right|$

$y = 0.d_1 d_2 \cdots d_k d_{k+1} \cdots \times 10^n$, chopping $fl(y) = 0.d_1 d_2 \cdots d_k \times 10^n$

$$\left|\frac{y - fl(y)}{y}\right| = \frac{0.d_{k+1} d_{k+2} \cdots \times 10^{n-k}}{0.d_1 d_2 \cdots \times 10^n} = \frac{0.d_{k+1} d_{k+2} \cdots}{0.d_1 d_2 \cdots} \times 10^{-k}$$

$$\le \frac{1}{0.1} \times 10^{-k} = 10^{-k+1} \le 5 \times 10^{-(k-1)}.$$

at least $(k-1)$ significant digits

if $d_{k+1} < 5$, then it has k significant digits, it means when $d_{k+1} < 5$, it more reasonable to chop.

△ finite-digit arithmetic: $\oplus$ $\ominus$ $\otimes$ $\oslash$: $\textcircled{op}$

store first, calculate second, store last

$$x \textcircled{op} y = fl(fl(x) \textcircled{op} fl(y))$$

$\boxed{e.g.}$ $x = \frac{5}{7} = 0.\overline{714285}$, $y = \frac{1}{3} = 0.\dot{3}$, five-digit chopping

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

$$x \oplus y = fl(0.71428 \times 10^0 + 0.33333 \times 10^0) = fl(1.04761 \times 10^0)$$
$$= 0.10476 \times 10^0$$

absolute error: $\left|\frac{22}{21} - 0.10476 \times 10^0\right| = 0.190 \times 10^{-4}$

relative error: $\frac{0.190 \times 10^{-4}}{22/21} = 0.182 \times 10^{-4}$.

Try $x \ominus y$, $x \otimes y$, $x \oslash y$

$\boxed{e.g.}$ $u = 0.714251$ $v = 98765.9$ $w = 0.111111 \times 10^{-4}$

$fl(u) = 0.71425 \times 10^0$, $fl(v) = 0.98766 \times 10^5$, $fl(w) = 0.11111 \times 10^{-4}$

$$x \ominus u = fl(0.71428 \times 10^0 - 0.71425 \times 10^0) = fl(0.00003 \times 10^0)$$
$$= 0.3\underline{0000} \times 10^{-4}$$

ab. error $= 0.471 \times 10^{-5}$, re. error $= 0.136$

$$(x \ominus u) \oslash w = fl\left(\frac{0.30000 \times 10^{-4}}{0.11111 \times 10^{-4}}\right) = fl(2.700027) = 0.27\underline{000} \times 10^1$$

ab. error $= 0.424$ re. error $= 0.136$

$$(x \ominus u) \otimes v = fl(0.30000 \times 10^{-4} \times 0.98766 \times 10^5) = fl(0.296298 \times 10^1)$$
$$= 0.29629 \times 10^1$$

ab. error $= 0.465$ re. error $= 0.136$.

△ 4 cases to reduce roundoff error:

△ 4 cases to reduce roundoff error:

① avoid numerator >> denominator

if $fl(z) = z+\delta$, $\varepsilon = 10^{-n}$

$$z \oslash \varepsilon = fl\left(\frac{fl(z)}{fl(\varepsilon)}\right) = fl\left(\frac{z+\delta}{10^{-n}}\right) = (z+\delta)\times 10^{n}$$

e.g. $x = 0.01$ $\quad \frac{e^x - 1}{x} = 1.0050167\cdots \quad e^{0.01} = 1.010050$

$= 1.0000$ (five digit chopping)

but $\frac{e^x - 1}{x} \approx 1 + \frac{1}{2}x = 1.0050$

② avoid two nearly equal numbers $x \approx y$ doing $x \ominus y$

$fl(x) = 0.d_1 d_2 \cdots d_p\, \alpha_{p+1} \cdots \alpha_k \times 10^n$

$fl(y) = 0.d_1 d_2 \cdots d_p\, \beta_{p+1} \cdots \beta_k \times 10^n$

$x \ominus y = 0.\delta_{p+1}\delta_{p+2}\cdots\delta_k \times 10^{n-p}$, $\quad 0.\delta_{p+1}\delta_{p+2}\cdots\delta_k = fl(0.\alpha_{p+1}\cdots\alpha_k - \beta_{p+1}\cdots\beta_k)$

only $k-p$ digits of significance

e.g. $10 \ominus \sqrt{99}$ (3-digit chopping) $\quad 10 - \sqrt{99} \approx 0.0501$

$fl(\sqrt{99}) = 0.994 \times 10^1$, $fl(10) = 0.100 \times 10^2$

$10 \ominus \sqrt{99} = fl(0.100 \times 10^2 - 0.994 \times 10^1) = fl(0.100 \times 10^2 - 0.099 \times 10^2)$

$= fl(0.001 \times 10^2) = 0.100 \times 10^0$ $\quad$ error: 0.0499

but $1 \oslash (10 \oplus \sqrt{99}) = fl\left(\frac{0.100 \times 10^1}{0.199 \times 10^2}\right) = fl(0.5025 \times 10^{-1}) = 0.503 \times 10^{-1}$

error: 0.0002

e.g. $A = 10^7(1 - \cos 2^\circ)$ $\quad$ 4-digit rounding: $A = 0.6092 \times 10^4$

$fl(\cos 2^\circ) = 0.9994 \times 10^0 \quad fl(1) = 0.1000 \times 10^1 \quad 10^7 = 0.1000 \times 10^8$

$$A = fl\left(0.1000 \times 10^8 \times fl\left(0.1000 \times 10^1 - 0.0999 \times 10^1\right)\right) = 0.1000 \times 10^4$$

but: $1 - \cos\theta = 2\sin^2\frac{\theta}{2}$ $\quad fl(\sin 1^\circ) = 0.1745 \times 10^{-1}$

$$fl\left(0.1000 \times 10^8 \times 0.2 \times 10^1 \times (0.1745 \times 10^{-1})^2\right) = 0.6090 \times 10^4$$

③ avoid $x \oplus y$ for $x >> y$

e.g. $A = 52492 + \delta_1 + \delta_2 + \cdots + \delta_{1000}$, $\delta_i = 0.1$ $= 52592 = 0.52592 \times 10^5$

5-digit rounding:

$52492 = 0.52492 \times 10^5$, $0.1 = 0.10000 \times 10^1$

$52492 \oplus 0.1 = fl(0.52492 \times 10^5 + 0.00000 \times 10^5) = 0.52492 \times 10^5$

then $52492 \oplus 0.1 \oplus 0.1 \oplus \cdots \oplus 0.1 = 0.52492 \times 10^5$ error $= 100$

but: $52492 \oplus \underline{(0.1 \oplus 0.1 \oplus \cdots \oplus 0.1)} = fl(0.52492 \times 10^5 + 0.1 \times 10^3)$

$= fl(0.52492 \times 10^5 + 0.00100 \times 10^5) = 0.52592 \times 10^5$ no error

④ reduce steps:

e.g. evaluation of polynomials

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

direct calculation: multiplications: $n + n-1 + \cdots + 1 = \frac{n(n+1)}{2}$

additions: n

(text book)

but with nested arithmetic e.g. 秦九韶算法 (1202年-1261年) 源于《算书九章》(汉代) 起始于

$f(x) = (\cdots(a_n x + a_{n-1})x + a_{n-2})x + \cdots + a_1)x + a_0$

n multiplications + n additions

RK: also called Horner's algorithm ≈1800

Pseudo code: Input: $a_n, a_{n-1}, \cdots, a_0, x$

Output: $f(x)$

Step 1: $S = a_n$

Step 2: For $k = n-1, n-2, \cdots, 1, 0$

Set $S = xS + a_k$

Step 3: set $f(x) = S$. Output $(f(x))$

STOP.

HW1, Sec. 1.2: 1, 3, 5 a c 14

HW1, Sec. 1.2: 1, 3, 5   9c   14

15, 19   20   25   9   28, 29

Only Question (a)