Chapter 2 Computer Arithmetic

§2.1 Floating-Point Numbers and Roundoff Errors

decimal system

$$427.325 = 4\times10^{2}+2\times10^{1}+7\times10^{0}+3\times10^{-1}+2\times10^{-2}+5\times10^{-3}$$

binary system

$$(1001.11101)_2 = 1\times2^{3}+0\times2^{2}+\cdots+1\times2^{-3}+0\times2^{-4}+1\times2^{-5}$$
$$= (9.90625)_{10}$$

	x	$\tilde{x}$	
rounding	0.1735499	$\longrightarrow 0.1735$	$\lvert x-\tilde{x}\rvert \le \frac{1}{2}\times10^{-4}$
	0.4321609	$\longrightarrow 0.4322$	
		$\hat{x}$	
chopped/truncated		$\longrightarrow 0.1735$	$\lvert x-\hat{x}\rvert < 10^{-4}$
		$\longrightarrow 0.4321$	

Normalized Scientific Notation

decimal system

$$732.5051 = 0.732505\,1\times10^{3}$$
$$-0.005612 = -0.5612\times10^{-2}$$

$$x = \pm r\times10^{n}$$

(r: mantissa, n: exponent)

$$\frac{1}{10} \le r < 1$$

binary system

$$x = \pm q\times2^{m} \quad \text{where} \quad \frac{1}{2} \le q < 1$$

Hypothetical Computer Marc-32

a word length : 32 bits (binary digits)

a nonzero number $x = \pm f \times 2^m$

$= (-1)^s f \times 2^m$ normalized floating-point form

s: 1-bit (pos. 0, neg 1)

$f = (1.f)_2$ and $m = e - 127$ — e: 8-bit integer, 8-bit biased exponent

f: 23-bit fractional part of real number x

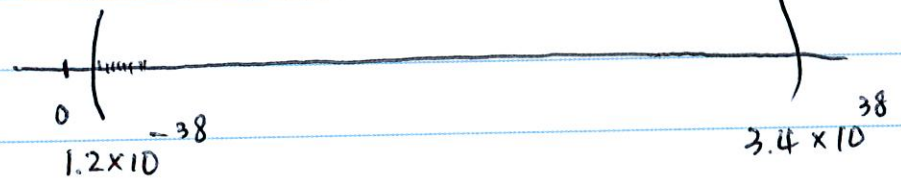
$0 < e < (11111111)_2 = 2^8 - 1 = 255$

$\Rightarrow -126 \le m \le 127$

floating-point range of numbers

$$\left(2^{-126},\ (2-2^{-23})2^{127}\right) \approx \left(1.2\times10^{-38},\ 3.4\times10^{38}\right)$$

distribution: nonuniform



Nearby Machine Number (Floating-point number)

x — real number

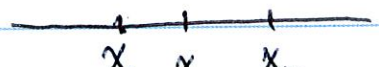
$fl(x) = x^*$ — machine number closest to x satisfies $|\delta| = \left|\frac{x - x^*}{x}\right| \le 2^{-24}$

2^{-24}: unit roundoff error

floating-point number base β (decimal $_{10}$ or binary $_{2}$)

$$fl(x) = x(1+\delta) \text{ with } |\delta| \le \varepsilon = \begin{cases} \frac{1}{2}\beta^{1-n} & \text{rounding} \\ \beta^{1-n} & \text{chopping} \end{cases}$$

(ε: unit roundoff error)



x — any machine number

x_l, x_r — two machine numbers on its left and right sides

$$2^{-24} < \frac{x_r - x}{x} = \frac{x - x_l}{x} \le 2^{-23} \quad \text{on Marc-32}$$

Relative Roundoff Error in Adding

$x_0, x_1, \cdots, x_n$ — positive machine numbers in a computer with machine ε

the relative roundoff error in computing $\sum_{i=0}^{n} x_i$

$$\le (1+\varepsilon)^n - 1 \approx n\varepsilon.$$

§2.2 Absolute and Relative Errors: Loss of Significance

x — a given real number

x^* — an approx to x

absolute error $|x - x^*|$

relative error $\left|\frac{x - x^*}{x}\right|$

example $\left|\frac{x - fl(x)}{x}\right| \leq \varepsilon$

Loss of Significance

a decimal computer have a 5-digit mantissa

example $x = 0.37214\,78693$

$y = 0.37202\,30572$

$x - y = 0.00012\,48121$

$fl(x) = 0.37215$

$fl(y) = 0.37202$

$fl(x) - fl(y) = 0.00013 = 0.13000 \times 10^{-3}$

0.1248121×10^{-3}

$$\left|\frac{(x-y) - (fl(x) - fl(y))}{x - y}\right| = \left|\frac{0.00012\,48121 - 0.00013}{0.00012\,48121}\right| \approx 4\%$$

Subtraction of Nearly Equal Quantities

$$y = \sqrt{x^2+1} - 1 = \frac{x^2}{\sqrt{x^2+1} + 1}$$

Thrm on Loss of Precision

$x > y$ — positive normalized floating-pt binary machine numbers satisfy

$$2^{-q} \le \frac{x-y}{x} \le 2^{-p}$$

$\Rightarrow$ at most q and at least p significant binary bits are lost in the subtraction $x-y$.

Proof on the lower bound

Let $x = r \times 2^n \qquad \frac{1}{2} \le r, s < 1 \qquad$ i.e., $r = (0.1\cdots)_2$

$y = s \times 2^m \qquad \cdots 2^{-1} \; 2^0 \; 2^1 \cdots$ ($2^{-1} = \frac{1}{2}$, $2^0 = 1$)

$$x > y \Rightarrow y = \left(s \times 2^{m-n}\right) \times 2^n$$

$$x - y = \left(r - s \times 2^{m-n}\right) \times 2^n$$

$$\frac{1}{2} \cdot 2^{-q} \le r - s \times 2^{m-n} = r\left(1 - \frac{s \times 2^m}{r \times 2^n}\right) = r\,\frac{x-y}{x} < 2^{-p}$$

$\frac{1}{2} \cdot 2^{-q} = 2^{-(q+1)}$

#

Example $\quad y = x - \sin x$

(a) $\sin x \approx x$

$$y = x - \sin x = \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} - \cdots$$

$$\approx \frac{x^3}{6}\left\{1 - \frac{x^2}{20}\left[1 - \frac{x^2}{42}\left(1 - \frac{x^2}{72}\right)\right]\right\}$$

(b) By the thm, if $1 - \frac{\sin x}{x} \geq \frac{1}{2} = 2^{-1} \Rightarrow |x| \geq 1.9$

$y = x - \sin x$ loss at most of one bit

(c) $|x| < 1.9$

at the worst case $x = 1.9$

$$y = x - \sin x \approx \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} - \frac{x^9}{9!} + \frac{x^{11}}{11!} - \frac{x^{13}}{13!}$$

has an error of at most 10^{-9}. (Prob 2.2.1)

Evaluation of Functions

for large x, for certain functions f

evaluation $f(x)$ could loss significant digits

range reduction

$$\cos(x + 2n\pi) = \cos x$$

$$\Rightarrow \cos 33278.21 = \cos(33278.21 - 5296 \times 2\pi)$$

$$= \cos(2.46) = -0.7765702835$$

§2.3 Stable and Unstable Computations: Conditioning

Numerical Instability

a numerical process is unstable

$\Longleftrightarrow$ small errors made at one stage of the process are magnified in subsequent stages and seriously degrade the accuracy of the overall calculation.

example 1 $\begin{cases} x_0 = 1, \; x_1 = \frac{1}{3} \\ x_{n+1} = \frac{13}{3} x_n - \frac{4}{3} x_{n-1} \quad \text{for } n \geq 1 \end{cases} \Longrightarrow x_n = \left(\frac{1}{3}\right)^n$

on Marc-32 $x_0 = 1, \; x_1 = \cdots, \; x_{15} = 3.657493$ (see p65)

$\Longrightarrow$ This algorithm is unstable.

general solution $x_n = A\left(\frac{1}{3}\right)^n + B\,4^n$

$$\begin{cases} 1 = A + B \\ \frac{1}{3} = \frac{A}{3} + 4B \end{cases} \Longrightarrow A = 1 \text{ and } B = 0$$

$\begin{cases} x_0 = 1, \; x_1 = 4 \\ x_{n+1} = \frac{13}{3} x_n - \frac{4}{3} x_{n-1} \end{cases} \Longrightarrow x_n = 4^n$ this is stable

$x_1 = 4.000006$

$x_{10} = 1.048576 \times 10^6$

$x_{20} = 1.099512 \times 10^{12}$

example 2 $y_n = \int_0^1 x^n e^x \, dx$ $\quad u = x^n,\ v' = e^x$ $\quad u' = n x^{n-1},\ v = e^x$

$$= x^n e^x \Big|_0^1 - n \int_0^1 x^{n-1} e^x \, dx$$

$$\Rightarrow \begin{cases} y_n = e - n y_{n-1} \\ y_0 = e - 1 \end{cases}$$

$y_2 = 0.718281$

$y_{11} = 1.422453$

$y_{15} = 39711.43$

but these cannot be correct, because

$1 = y_1 > y_2 > y_3 > \cdots > 0$ and $\lim_{n \to \infty} y_n = 0$

Conditioning

- evaluation of a function f at a pt x

$$f(x+h) - f(x) = f'(\xi) h \approx h f'(x) \quad \text{for small } h$$

$$\Rightarrow \frac{f(x+h) - f(x)}{f(x)} \approx \frac{h f'(x)}{f(x)} = \left[\frac{x f'(x)}{f(x)} \right] \frac{h}{x}$$

- $\frac{f(x+h) - f(x)}{f(x)}$: relative perturbation in evaluation of f
- $\left[\frac{x f'(x)}{f(x)} \right]$: condition number
- $\frac{h}{x}$: relative perturbation in x

example $f(x) = \arcsin x$

$$\frac{x f'(x)}{f(x)} = \frac{x}{\sqrt{1-x^2} \arcsin x}$$ is large for x near 1.

- **compute a simple root of f**

$f(r)=0$ and $f'(r)\neq 0$

perturbation of f $\quad F \equiv f+\varepsilon g$

root for F: $r+h$ $\quad 0=F(r+h)=f(r+h)+\varepsilon g(r+h)$

$$= \left[f(r)+hf'(r)+\tfrac{1}{2}h^2 f''(\xi)\right]+\varepsilon\left[g(r)+hg'(r)+\tfrac{1}{2}h^2 g''(\eta)\right]$$

$$\approx h f'(r)+\varepsilon g(r)+h g'(r)$$

$$\Rightarrow h \approx -\varepsilon \frac{g(r)}{f'(r)+\varepsilon g'(r)} \approx -\varepsilon\left(\frac{g(r)}{f'(r)}\right) \leftarrow \text{condition number}$$

$h=(r+h)-r$

$\varepsilon=(F-f)/g$

example $\quad f(x)=\prod_{k=1}^{20}(x-k)$ $\quad$ roots $1, 2, \cdots, 20$

$g(x)=x^{20}$ means an approximation error in the coefficient of x^{20}

$\Rightarrow$ for root $r=20$

$$h \approx -\varepsilon \frac{g(20)}{f'(20)} = -\varepsilon \frac{20^{20}}{19!} \approx -\varepsilon 10^{19}$$

Condition Number for $A_{n\times n}$

$$\kappa(A) = \|A\| \cdot \|A^{-1}\|$$

Hilbert matrix $\quad H_n=(h_{ij})_{n\times n}$ $\quad h_{ij}=\dfrac{1}{i+j-1}$, $\quad 1\le i,j\le n$.

$$\kappa_\infty(H_n) = \|H_n\|_\infty \|H_n^{-1}\|_\infty = c e^{3.5n}$$

$$\|A\|_\infty = \max_{i,j} |a_{ij}|$$