Lecture Notes

Section 1.4
Floating Point Arithmetic

I. Floating Pt Arithmetic

$$\boxed{x \circledast_{fl} y = fl(fl(x) \circledast fl(y))}$$

① $\circledast$ is a generic term for $+, -, *, \div$

② $fl(x)$ means round x to maximum allowable digits

a) Example: for $F(10, 2, -2, 2)$ find

$$1/7 +_{fl} 1/23 = fl\left(fl\left(\tfrac{1}{7}\right) + fl\left(1/23\right)\right)$$

$$= fl(.14 + .043)$$
$$= fl(.183) = .18$$

⇒ Note: IF I DID THIS ON CALC I'D GET .186335...

b) ARITHMETIC PROPERTIES DO NOT WORK FOR FP OPERATIONS

Example $F(10, 4, -2, 2)$

(a) $(.1329 +_{fl} 1.543) +_{fl} 23.21$
$= fl(1.6693) +_{fl} 23.21$
$= 1.669 +_{fl} 23.21 = fl(24.879) = \boxed{24.89}$

(b) $.1329 +_{fl} (1.543 +_{fl} 23.21) = \boxed{24.88}$

IE FP ARITHMETIC IS NOT DISTRIBUTIVE!!

II) Accumulation of Round-off Error

a) absolute deviation $\delta_x = fl(x) - x$

b) relative deviation $\epsilon_x = \frac{fl(x)-x}{x}$ (or $\frac{\delta x}{x}$)

$\therefore$ $\boxed{fl(x) = x + \delta_x}$ or $\boxed{fl(x) = x(1+\epsilon_x)}$

c) Multiplication

$$fl(x) * fl(y) = x(1+\epsilon_x)\, y(1+\epsilon_y)$$
$$= xy(1+\epsilon_x+\epsilon_y+\epsilon_{xy})$$

↑ this type of error will not "blow up" since ϵ_x, ϵ_y are small

d) division

$$\frac{fl(x)}{fl(y)} = \frac{x(1+\epsilon_x)}{y(1+\epsilon_y)} = \frac{x}{y}\frac{(1+\epsilon_x)}{(1+\epsilon_y)}$$

$\Rightarrow$ $\frac{1}{1+\epsilon_y} = 1 - \epsilon_y + \epsilon_y^2 - \epsilon_y^3 \ldots$

$\therefore$ $\frac{fl(x)}{fl(y)} = \frac{x}{y}(1+\epsilon_x)(1-\epsilon_y+\epsilon_y^2-\epsilon_y^3 \ldots)$

again these terms will not "blow up"

e) addition/subtraction

$$fl(x) \pm fl(y) = (x+\delta_x) \pm (y+\delta_y)$$

$$= (x \pm y + \delta_x \pm \delta_y)$$

$$= (x \pm y)\left(1 + \frac{\delta_x \pm \delta_y}{x \pm y}\right)$$

↗ if $x \pm y$ close to zero this will "blow up" ("cancelation error")

Algorithms should avoid subtraction of nearly equal numbers when possible

Example: Quadratic Equation

$$ax^2 + bx + c = 0 \Rightarrow x = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$$

① what if $b \gg 4ac \Rightarrow$ then $\sqrt{b^2-4ac} \approx b$

② if $b > 0$, then wish to avoid the + operation

$$\frac{-b+\sqrt{b^2-4ac}}{2a}$$

continues ⟹

Strategy

WHY IS THIS NOT A PROBLEM

$$\frac{-b+\sqrt{b^2-4ac}}{2a}\quad\frac{(-b-\sqrt{b^2-4ac})}{(-b-\sqrt{b^2-4ac})}$$

$$= \frac{b^2-(b^2-4ac)}{2a(-b+\sqrt{b^2-4ac})} = \frac{4ac}{2a(-b-\sqrt{b^2-4ac})} = \frac{2c}{-b\pm\sqrt{b^2-4ac}}$$

We have avoided the operation

SEE Example 1.10
Page 45

III) Computer Demo Example 1.12, p 48