Theoretical Questions Chapter 4

Ling Siu Hong
3200300602

November 23, 2022

I : We have $1 \le m < 2, \beta = 2, e = \lceil \log_2 477 \rceil = 8$, since

$$477 = 2^8 + 2^7 + 2^6 + 2^4 + 2^3 + 2^2 + 2^0,$$

therefore

$$477 = (1.1101101)_2 \times 2^8.$$

II : We have $1 \le m < 2, \beta = 2, e = \lfloor \log \frac{3}{5} \rfloor = -1$,

$$\begin{aligned} 0.6 \times 2 = 1.2 - -a_1 = 1, \\ 0.2 \times 2 = 0.4 - -a_2 = 0, \\ 0.4 \times 2 = 0.8 - -a_3 = 0, \\ 0.8 \times 2 = 1.6 - -a_4 = 1, \end{aligned}$$

$$477 = (1.1101101)_2 \times 2^8.$$

III: Let $x = (1.\overbrace{0000...00}^{p-1})_\beta$,

$$x_R = (1.\overbrace{0000...01}^{p-1})_\beta \times \beta^e = (1 + \frac{1}{\beta^{p-1}}) \ \times \beta^e,$$

$$\begin{aligned} x_L =& ((\beta-1).\overbrace{(\beta-1)(\beta-1)...(\beta-1)}^{p-1})_\beta \times \beta^e \\ =& (\beta - \frac{1}{\beta^{e-1}}) \\ =& [(\beta-1) + \frac{\beta-1}{\beta} + ... + \frac{\beta-1}{\beta^{p-1}}] \times \beta^{e-1}. \end{aligned}$$

Since we have $x_R - x = \beta^e + \frac{\beta^e}{\beta^{p-1}} - \beta^e = \beta^{e-p-+1}$ and $x - x_L = 1 \times \beta^{1-p} \times \beta^{e-1}$,thus $x_R - x = \beta(x - x_L)$.

IV : Under IEEE754 single-precision, 24 for the significant, $\frac{3}{5} = (1.0011...)_2 \times 2^{-1}$,the two adjacent are

$$x_R = (1.\overbrace{0011...01}^{23})_2 \times 2^{-1}),$$

$$x_L = (1.\overbrace{0011...10}^{23})_2 \times 2^{-1}).$$

Then, we have $x - x_L = \frac{3}{5} \times 2^{-24}$,$x_R - x_L = 1 \times 2^{-24}$,so that $x - x_L > x_R - x$ which means $fl(x) = x_R$. The relative round off error is

$$\epsilon = \frac{|fl(x) - x|}{|x|} = \frac{2}{3} \times 2{-}24$$

V: We have $\epsilon_M = \beta^{1-p}$,under IEEE754, $p = 24$ and $\beta = 2$,then $\epsilon_M = 2^{-23}$,

$$\epsilon_u = (1 - 2^{-23}) \times \epsilon_M \approx 1.19 \times 10^{-9}$$

VI When $x = \frac{1}{4}$, $1 - cosx = 0.031087578$, then $2^{-6} \leq 1 - cos(\frac{1}{4}) \leq 2^{-5}$.Therefore, the subtraction will lost at least 5, at most 6 bits of precision.

VII:We can avoid catastrophic cancellation,

1. By trigonometric identity

$$1 - cosx = 2sin^2\frac{x}{2}.$$

2. By Taylor's expansion

$$\begin{aligned}1 - cosx =& 1 - (1 - \frac{x^2}{2!} + \frac{x^4}{4!} - ...) \\ =& \frac{x^2}{2!} - \frac{x^4}{4!} + \frac{x^6}{6!}.\end{aligned}$$

VIII

1. $f(x) = (x-1)^\alpha$

$$C_f(x) = |\frac{x\alpha(x-1)^{\alpha-1}}{(x-1)^\alpha}|$$

Thus, $C_f \to \infty$ as $x \to 1$

2. $f(x) = \ln x$

$$C_f(x) = |\frac{1}{\ln x}|$$

Thus, $C_f \to \infty$ as $x \to 1$

3. $f(x) = e^x$

$$C_f(x) = |\frac{x \cdot e^x}{e^x}| = |x|$$

Thus, $C_f \to \infty$ as $x \to \infty$

4. $f(x) = \arccos x$

$$C_f(x) = |\frac{x\cdot(-1)}{\sqrt{1-x^2}\arccos x}| = |\frac{x}{\sqrt{1-x^2}\arccos x}|$$

Thus, $C_f \to \infty$ as $x \to \pm 1$

IX

- We have $cond_f(x) = |\frac{-xe^{-x}}{1-e^{-x}}| = |\frac{x}{e^x-1},|$.Let $g(x) = \frac{x}{e^x-1}$,when $x \in (0,1], x < e^x - 1$, thus $g(x) \in (0,1]$. Since $\lim\limits_{x\to 0}|\frac{x}{e^x-1}| = \lim\limits_{x\to 0}\frac{x}{x+o(x)} = 1$, therefore $cond_f(x) \leq 1$ for $x \in [0,1]$.
- Let $f(x) = 1 - e^x$, $cond_f(x) = \frac{x}{e^x-1}$,

$$\begin{aligned} f_A(x) =& fl(1 - fl(e^-x)) \\ =& [1 - e^{-x}(1+\delta_1)](1+\delta_2), where|\delta_1| \leq \epsilon_u, |\delta_2| \leq \epsilon_u \end{aligned}$$

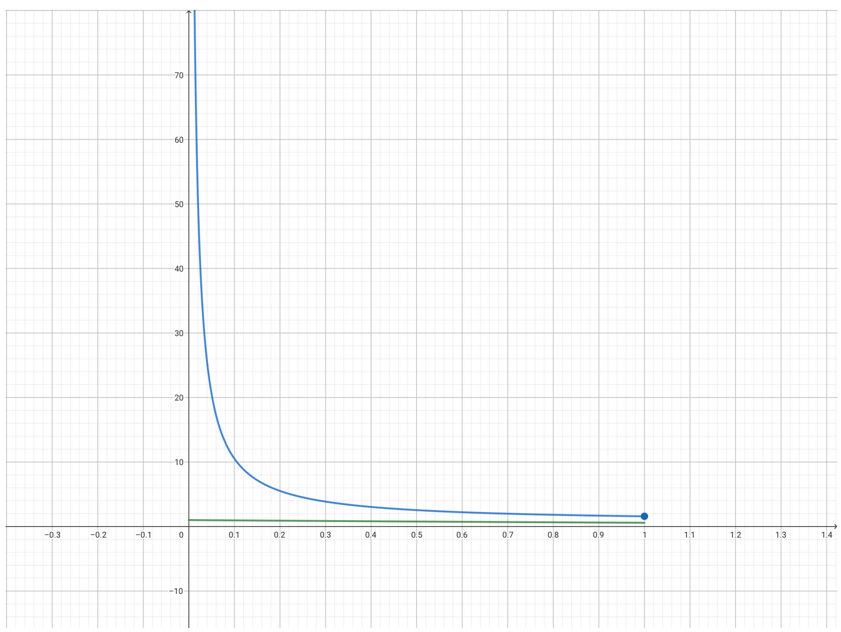
Since $\delta_1\delta_2$ is too small, we ignore it then we get

$$f_A(x) = (1-e^{-x})(1+\delta_2 - \delta_1 \cdot \frac{e^{-x}}{1-e^{-x}})$$

and

$$\phi(x) = 1 + \frac{e^{-x}}{1-e^{-x}} = \frac{1}{1-e^{-x}},$$

By theorem 4.76, $cond_A(x) \leq \frac{e^-x-1}{x}\cdot\frac{1}{1-e^{-x}} = \frac{e^x}{x}$.

- 

$cond_f(x)$: green; $cond_A(x)$: blue

It is clearly that $cond_A(x)$ if greater than $cond_f(x)$ especially $x \to 0$. From $cond_A(x) \to \infty$ as $x \to 0$, the subtraction $1 - e^{-x}$ is not accurate, because $x \to 0$,$1 - e^{-x} \to 0$.

X: Let$r = f(a_0, a_1, ..., a_{n-1})$,then we have

$$cond_1 = |\frac{1}{r}| \sum_{i=0}^{n-1} |a_i \frac{\partial r}{\partial a_i}| = \frac{\sum_{i=0}^{n-1} |a_i r^i|}{r(\sum_{i=1}^{n}(n-i+1)a_i r^{n-i})}$$

Assume that $r = n$, $f(x) = \prod_{i=1}^{n}(x-i)$,thus$cond_1 = \frac{\prod_{i=1}^{n}(n+i) - n^n}{n!}$,then we have $cond_1 \geq \frac{n^n}{n!}$. Comparing with Wilkinson, both are n is larger, $cond_1$ will also become larger.

XI: For instance, in FPN system(2,2,-1,1), $a = (1.0)_2 \times 2^0$, $b = (1.1)_2 \times 2^0$.We calculate it in the register of precision 2p(4), and we have $\frac{a}{b} = (0.101)_2$.However,$E_{rel}(\frac{a}{b}) = (0.01)_2 = \epsilon_u$, contradiction!

XII Since $128 = (1.000...00)_2 \times 2^7$, $129 = (1.0000001...00)_2 \times 2^7$, thus $2^7 \times 2^-23 = 2^-16 > 10^-6$,therefore it cannot compute the root with absolute accuracy $< 10^{-6}$.

XIII Suppose $|x_i, x_i + 1| < \delta$, means that the adjacent knot is too close, the cubic spline can be soled by following equation,

$$\begin{aligned} a_0 &= f(x_i), \\ a_0 + \delta a_1 + \delta^2 a_2 + \delta^3 a_3 &= f(x_{i+1}), \\ a_1 &= f'(x_i), \\ a_1 + 2\delta a_2 + 3\delta a_3 &= f'(x_{i+1}). \end{aligned}$$

When $\delta \to 0$,the condition number of coefficient matrix is too large, thus the result is inaccurate.