Chapter 4 Computer Arithmetic

Linghui Ngoo

16th November 2022

Problem 1

$$\begin{aligned}477 =&(11011101)_2\\ =&(1.11011101)_2 \times 2^8\end{aligned}$$

Problem 2

$$\begin{aligned}\frac{3}{5} =&0.6\\ =&(0.10011001.....)_2\\ =&(1.0011001......)_2 \times 2^{-1}\end{aligned}$$

Problem 3 We know that $x = 1.000... \times \beta^e$,assume that the significand digits is p,thus we have

$$\begin{aligned}x_L =&0.(\beta-1)(\beta-1)(\beta-1)...(\beta-1) \times \beta^e\\ =&(\beta-1).(\beta-1)(\beta-1)...(\beta-1) \times \beta^{e-1}\\ x_R =&1.000...001 \times \beta^e\end{aligned}$$

hence,

$$\begin{aligned}x_R - x =&1.00... \times \beta^{e-p}\\ x - x_L =&1.00.. \times \beta^{e-1-p}\\ x_R - x =&\beta(x - x_L)\end{aligned}$$

Problem 4 from Promblem 2 , we know that

$$\frac{3}{5} = (1.0011001......)_2 \times 2^{-1}$$

two adjacent x_L and x_R are

$$\begin{aligned} x_R &= (1.0011...1010)_2 \times 2^{-1} \\ x_L &= (1.0011...1001)_2 \times 2^{-1} \end{aligned}$$

Follow that

$$\begin{aligned} x - x_L &= \frac{3}{5} \times 2^{-24} \\ x_R - x_L &= 1 \times 2^{-24} \\ x_R - x &= (x_R - x_L) - (x - x_L) \\ &= 2^{-24} - \frac{3}{5} \times 2^{-24} \\ &= \frac{2}{5} \times 2^{-24} \end{aligned}$$

thus,

$$fl(x) = x_R$$

Relative roundoff error is

$$\begin{aligned} E_{rel}(x) &= \frac{fl(x) - x}{x} \\ &= \frac{2}{3} \times 2^{-24} \end{aligned}$$

Problem 5 We know that,

$$\epsilon_M = \beta^{1-p}$$

for IEEE 754 single-precision , p=24 , assume that $\beta = 2$,then,

$$\begin{aligned} \epsilon_M &= 2^{-23} \\ \epsilon_u &= (1 - 2^{-23}) \times \epsilon_M \\ &= (1 - 2^{-23}) \times 2^{-23} \end{aligned}$$

Problem 6 For $x = \frac{1}{4}$,we know that $1 > cos(x)$,compute $d = 1 - cos(x)$,when $x = \frac{1}{4}$,we have $d = 0.031087578...$,which is between 2^{-5} and 2^{-6}.

Therefor,we lose at most 6 and at least 5 significant bits.

Problem 7 solution 1:rewrite the expression

$1 - cos(x) = 2sin^2(\frac{x}{2})$

solution 2:use Taylor's expression

$1 - cos(x) = \frac{x^2}{2!} - \frac{x^4}{4!} + \frac{x^6}{6!} + ...$

Problem 8 (1)

$$\begin{aligned} C_f(x) &= |\frac{x\alpha(x-1)^{\alpha-1}}{(x-1)^{\alpha}}| \\ &= |\frac{x\alpha}{x-1}| \end{aligned}$$

Hence,$C_f(x) \to +\infty$, as $x \to 1$

(2)

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

Hence,$C_f(x) \to +\infty$, as $x \to 0$

(3)

$$C_f(x) = |x|$$

Hence,$C_f(x) \to +\infty$, as $x \to +\infty$

(3)

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

Hence,$C_f(x) \to +\infty$, as $x \to \pm 1$

Problem 9

$$\begin{aligned} C_f(x) =& |\frac{xe^{-x}}{1-e^{-x}}| \\ =& |\frac{x}{e^x - 1}| \end{aligned}$$

(a)$C_f(x)$ did not have max point or minimum point for $x \in (0,1)$.When $x = 0$, $C_f(x) = 0$, when $x = 1$, $C_f(x) = \frac{1}{e-1} < 1$, thus $C_f(x) < 0$ for $x \in [0,1]$
(b)For $|\delta_i| \leq \epsilon_u$

$$\begin{aligned} f_A(x) =& (1 - e^{-x}(1+\delta_1))(1+\delta_2) \\ =& (1-e^{-x})(1+\delta_2 - \frac{\delta_1(1+\delta_2)}{e^x - 1}) \end{aligned}$$

thus,

$$\begin{aligned} |\delta(x)| =& |\delta_2 - \frac{\delta_1(1+\delta_2)}{e^x - 1}| \\ \leq & |\epsilon_u| + |\frac{\epsilon_u(1+\epsilon_u)}{e^x - 1}| \\ \leq & \epsilon_u |\frac{e^x}{e^x - 1}| \\ =& \epsilon_u \varphi(x) \end{aligned}$$

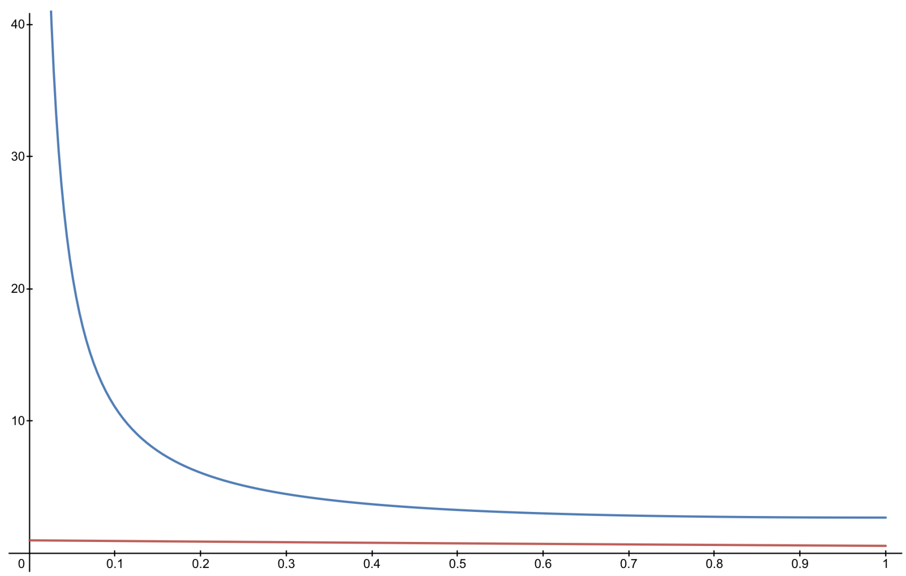
From Theorem 4.76,we have

$$\begin{aligned} cond_A(x) \leq & \frac{\varphi(x)}{C_f(x)} \\ =& |\frac{e^x}{x}| \end{aligned}$$

Hence, $cond_A(x)$ may be unbounded as $x \to 0$
(c)

Problem 10

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



As the root of polynomial q(x) is r and $a_n = 1$, hence $q(r) = 0$, thus we have

$$r^n + a_{n-1}r^{n-1} + ... + a_1 r + a_0 = 0$$

partial derive the polynomial respect to variable a_i,

$$\begin{aligned} nr^{n-1}\frac{\partial r}{\partial a_i} + a_{n-1}(n-1)r^{n-2}\frac{\partial r}{\partial a_i} + ... + a_i i r^{i-1}\frac{\partial r}{\partial a_i} + r^i + ... + a_1\frac{\partial r}{\partial a_i} =& 0 \\ \frac{\partial r}{\partial a_i}[nr^{n-1} + a_{n-1}(n-1)r^{n-2} + ... + 2a_2 r + a_1] + r^i =& 0 \\ \frac{\partial r}{\partial a_i} =& \frac{-r^i}{nr^{n-1} + a_{n-1}(n-1)r^{n-2} + ... + 2a_2 r + a_1} \end{aligned}$$

thus,

$$\begin{aligned} cond_f(x) =& \frac{1}{|r|}|\sum_{i=0}^{n-1}\frac{-a_i r^i}{nr^{n-1} + a_{n-1}(n-1)r^{n-2} + ... + 2a_2 r + a_1} \\ =& \frac{1}{|r|}|\frac{p(r) - r^n}{nr^{n-1} + a_{n-1}(n-1)r^{n-2} + ... + 2a_2 r + a_1}| \\ =& |\frac{r^{n-1}}{nr^{n-1} + a_{n-1}(n-1)r^{n-2} + ... + 2a_2 r + a_1}| \\ =& |\frac{r^{n-1}}{q'(r)}| \end{aligned}$$

In Wilkinson example,assume that r is the foot of f(x)

$$f(x) = \prod_{k=1}^{p}(x-k)$$
$$cond_f(r) = \frac{r^p}{f'(r)}$$

Hence,it has the same result with Wilkinson,a small change of the coefficient of the polynomial would cause a large change of the root.

Problem 11 Assume that $\beta = 2, p = 2, L = -1, U = 1, a = (1.0)_2 \times 2^0, b = (1.1)_2 \times 2^0$.Assume that $c = fl(\frac{a}{b})$is calculated in a register of precision of 2p.Hence,we have $\frac{a}{b} = 0.101$,thus $fl(\frac{a}{b}) = 0.10 = 1.0 \times 2^{-1}$ As $E_{rel}(x) = |fl(\frac{\frac{a}{d}-\frac{a}{b}}{\frac{a}{b}}| = (0.01)_2 = \frac{1}{4} = 2^{-2} = \epsilon_u$,contradict Lemma 4.39

Problem 12

$$\begin{aligned}
128 =& (1.000...)_2 \times 2^7 \\
129 =& (1.00000010...)_2 \times 2^7 \\
E_{abs} =& (0.00...1)_2 \times 2^7 \\
=& 2^{-23} * 2^7 \\
=& 2^{-16} \\
\approx& 1.526 \times 10^{-5} > 10^{-6}
\end{aligned}$$

that's why it cannot compute the root with accuracy 10^{-6}.

Problem 13