Numerical Analysis homework # 4

Chen Shuo 12231064 *

(Electronic Science and Technology), Zhejiang University

Submitted time: November 25, 2024

I

$477 = (111011101)_2 = (1.1101101)_2 \times 2^7$

II

$\frac{3}{5} = (0.\overline{1001})_2 = (1.\overline{0011})_2 \times 2^{-1}$

III

$$x = \beta^e = 1 \times \beta^e \Longrightarrow x_R = 2 \times \beta^e, x_R - x = \beta^e$$
$$x = \beta^e = \beta \times \beta^{e-1} \Longrightarrow x_L = (\beta - 1) \times \beta^{e-1}, x - x_L = \beta^{e-1}$$
$$\therefore x_R - x = \beta(x - x_L)$$

IV

According to the result of II, $\frac{3}{5} = (1.\overline{0011})_2 \times 2^{-1}$. In IEEE 754 single-precision protocal, there are 23 bits for the significand, so:

$$\frac{3}{5} = (1.0011001\cdots)_2 \times 2^{-1}$$
$$x_L = (1.0011001\cdots001)_2 \times 2^{-1}$$
$$x_R = (1.0011001\cdots010)_2 \times 2^{-1}$$
$$x - x_L = \frac{3}{5} \times 2^{-24}$$
$$x_R - x = \frac{2}{5} \times 2^{-24}$$

Therefore $\mathrm{fl}(x) = x_R$, roundoff error is $|\mathrm{fl}(x) - x| = \frac{2}{5} \times 2^{-24}$.

V

If the excess bits are simply dropped, then the $(0, 1) \times 2^{-23}$ part would be eliminated. The unit roundoff is $\sup((0, 1) \times 2^{-23}) = 2^{-23}$.

VI

According to Theorem 4.49, $\beta^{-t} \leq 1 - \frac{y}{x} \leq \beta{-}s$. In this case, $\beta = 2$, $y = \cos\frac{1}{4}$, $x = 1$. And $1 - \cos\frac{1}{4} \approx 0.0311$, $2^{-5} \leq 0.0311 \leq 2^{-4}$. Therefore it would lose 4 or 5 bits of precision.

*Email address: shuo_chen@zju.edu.cn

VII

- By using Taylor series: $1-\cos x = 1-(1-\frac{x^2}{2}+\frac{x^4}{4}-\frac{x^6}{6}+\cdots) = \frac{x^2}{2}-\frac{x^4}{4}+\frac{x^6}{6}+\cdots$
- $1-\cos x = \frac{(1-\cos x)(1+\cos x)}{1+\cos x} = \frac{1-\cos^2 x}{1+\cos x} = \frac{\sin^2 x}{1+\cos x}$

VIII

According to the definition, $C_f(x) = |\frac{xf'(x)}{f(x)}|$.

- $f(x) = (x-1)^\alpha$, $f'(x) = \alpha(x-1)^{\alpha-1}$, $C_f(x) = |\frac{\alpha x}{x-1}|$, $C_f(x)$ are large when $x \to 1$.
- $f(x) = \ln x$, $f'(x) = \frac{1}{x}$, $C_f(x) = |\frac{1}{\ln x}|$, $C_f(x)$ are large when $x \to 1$.
- $f(x) = e^x$, $f'(x) = e^x$, $C_f(x) = |x|$, $C_f(x)$ is large when $|x|$ is large.
- $f(x) = \arccos x$, $f'(x) = -\frac{1}{\sqrt{1-x^2}}$, $C_f(x) = |\frac{x}{\sqrt{1-x^2}\arccos x}|$, $C_f(x)$ are large when $x \to \pm 1$ or $x \to \frac{\pi}{2}$.

IX

IX-a

$$\mathrm{cond}_f(x) = |\frac{xf'(x)}{f(x)}| = \frac{xe^{-x}}{1-e^{-x}}, x \in (0,1]$$

$$\mathrm{cond}_f(0) = \lim_{x\to 0^+} |\frac{xe^{-x}}{1-e^{-x}}| = 1$$

We have $\frac{xe^{-x}}{1-e^{-x}} = \frac{x}{e^x-1}$ and $g(x) < 1$ since $0 < x < e^x - 1$, so $\mathrm{cond}_f(x) \le 1$ for $x \in [0,1]$.

IX-b

Apply Theorem 4.78 to analyze the conditioning of the algorithm:

$$f_A = \mathrm{fl}(1-\mathrm{fl}(e^{-x}))$$

By neglecting the quadratic terms of $O(\delta_i^2)$:

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

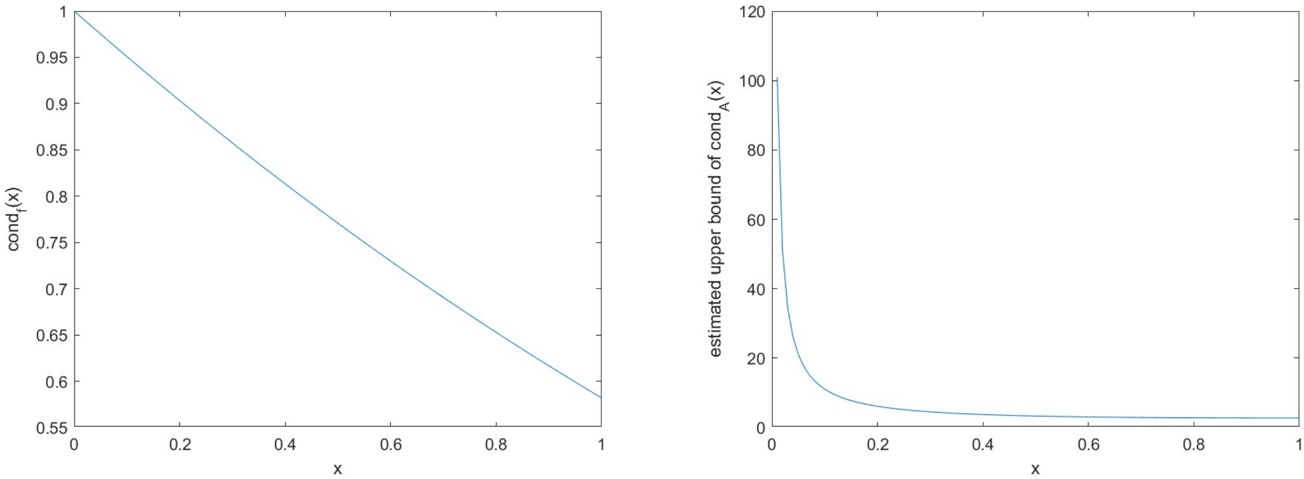
where $|\delta_i| \le \epsilon_u$ for $i=1,2$. Hence we have $\phi(x) = 1+\frac{e^{-x}}{1-e^{-x}} = \frac{1}{1-e^{-x}}$ and

$$\mathrm{cond}_A(x) \le \frac{1-e^{-x}}{xe^{-x}} \cdot \frac{1}{1-e^{-x}} = \frac{e^x}{x}$$

IX-c

$\mathrm{cond}_f(x)$ is plotted as Figure 1.

By IX-b, $\mathrm{cond}_A(x)$ may be unbounded at $x=0$. On the other hand, $\mathrm{cond}_A(x)$ is controlled by e as $x \to \frac{\pi}{2}$.



(a) $\mathrm{cond}_f(x)$

(b) estimated upper bound of $\mathrm{cond}_A(x)$

Figure 1: ProblemIX

X

Do SVD to A, i.e. $A = U\Sigma V^T$, U, V are orthogonal matrices and $\Sigma = \mathrm{diag}\{\sigma_1, \cdots \sigma_m\}$, σ_i is the singular value of A with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m > 0$. According to the definition of 2-norm:

$$||A||_2 = \sup_x \frac{||Ax||_2}{||x||_2} = \sup_{||x||_2=1} ||Ax||_2 = \sup_{||x||_2=1} ||\Sigma x||_2 = \sup_{||x||_2=1} \sqrt{\sigma_1^2 x_1^2 + \cdots + \sigma_m^2 x_m^2} = \sigma_1 = \sigma_{max}$$

$A^{-1} = V\Sigma^{-1}U^T$, and $\Sigma^{-1} = \mathrm{diag}\{\sigma^{-1}, \cdots \sigma_m^{-1}\}$. Similarly we can get $||A^{-1}||_2 = \sigma_m^{-1} = \sigma_{min}^{-1}$.

Therefore, $\mathrm{cond}_2 A = ||A||_2 ||A^{-1}||_2 = \dfrac{\sigma_{max}}{\sigma_{min}}$.

If A is normal, then $\{\sigma_i\}$ of Σ in SVD are the norm of eigenvalues of A, so $\mathrm{cond}_2 A = \dfrac{|\lambda|_{max}}{|\lambda|_{min}}$.

XI

$$\sum_{j=0}^{n} a_j r^j = 0$$

Consider r as the function of a_j and take derivates with a_j for two sides:

If $j \neq 0$,

$$r^j + a_j j r^{j-1} \frac{\partial r}{\partial a_j} + \sum_{k=1, k\neq j}^{n} a_k k r^{k-1} \frac{\partial r}{\partial a_j} = r^j + \sum_{k=1}^{n} a_k k r^{k-1} \frac{\partial r}{\partial a_j} = 0 \Longrightarrow \frac{\partial r}{\partial a_j} = -\frac{r^j}{\sum\limits_{k=1}^{n} a_k k r^{k-1}}$$

If $j = 0$,

$$1 + \sum_{k=1}^{n} a_k k r^{k-1} = 0 \Longrightarrow \frac{\partial r}{\partial a_j} = -\frac{1}{\sum\limits_{k=1}^{n} a_k k r^{k-1}}$$

Therefore 2 cases can be conbimed into 1, i.e. $\dfrac{\partial r}{\partial a_j} = -\dfrac{r^j}{\sum\limits_{k=1}^{n} a_k k r^{k-1}}$, $j = 0, 1, \cdots, n-1$.

Let $\mathbf{y} = (a_0, \cdots, a_{n-1})$ and according to Definition 4.71, $\mathrm{cond}_f(\mathbf{y}) = ||A(\mathbf{y})||$. In this case $A(\mathbf{y}) \in \mathbb{R}^{1\times n}$, $a_{1j}(\mathbf{y}) = \left|\dfrac{y_j \frac{\partial f}{\partial y_j}}{f(\mathbf{y})}\right| = \dfrac{|a_{j-1} r^{j-1}|}{|\sum\limits_{k=1}^{n} a_k k r^k|}$. Here we use 1-norm and $\mathrm{cond}_f(\mathbf{y}) = ||A(\mathbf{y})||_1 = \sum\limits_{j=1}^{n} |a_{1j}| = \dfrac{\sum\limits_{k=0}^{n-1} |a_k r^k|}{|\sum\limits_{k=1}^{n} a_k k r^k|}$

For the Wilkinson example, $r = 1, 2, \cdots, p$. Let $h(x) = \prod_{k=1}^{p}(x+k)$, and $\sum_{k=0}^{n-1}|a_k r^k| = h(r)$. The denominator term $|\sum_{k=1}^{n} a_k k r^k| = |f'(r)| = |\sum_{j=1}^{p}\prod_{k=1,k\neq j}^{p}(r-k)|$. Let $r = m \in \{1, 2, \cdots, p\}$, $\text{cond}_f(\mathbf{y}) = \frac{h(r)}{|f'(r)|} = \frac{m(m+1)\cdots(m+p)}{\prod_{k=1,k\neq j}^{p}(r-k)} = \frac{(m+p)!}{m!(m-1)!(p-m)!}$. For the root $r = p$, $\text{cond}_f(\mathbf{y}) = \frac{(2p)!}{p!(p-1)!}$.

For p = 20, 30, 40, $\text{cond}_f(\mathbf{y})$ is about $2.8 \times 10^{12}, 3.5 \times 10^{18}, 4.3 \times 10^{24}$, respectively. Comparing the result with that in the Wilkinson Example, the problem of root finding for polynomials with very high degrees is hopeless.

XII

By Definitions 4.24, 4.16, and 4.26, the unit roundoff of a register with precision $2p$ is

$$\frac{1}{2}\beta^{1-2p} = \frac{1}{2}\beta^{1-p}\beta^{1-p}\beta^{-1} = \beta^{-1}\epsilon_u\epsilon_M$$

Let $M_a = a = 1, M_b = b = \beta - \epsilon_M$, then:

$$M_{c1} = \frac{M_a}{M_b} + \delta_1, |\delta_1| < \beta^{-1}\epsilon_u\epsilon_M$$

$$M_{c2} = \beta M_{c1} + \delta_2 = \beta\frac{M_a}{M_b}(1 + \frac{\beta\delta_1 + \delta_2}{\beta M_a/M_b}) = \beta\frac{M_a}{M_b}(1 + (\beta - \epsilon_M)\delta_1 + \frac{\beta - \epsilon_M}{\beta}\delta_2), |\delta_2| < \epsilon_u$$

Let $\delta_1 = \beta^{-1}\epsilon_u\epsilon_M - \epsilon_1 < \beta^{-1}\epsilon_u\epsilon_M$, $\delta_2 = \epsilon_u - \epsilon_2 < \epsilon_u$. Choose $\epsilon_1 = \frac{\epsilon_u\epsilon_M^2}{\beta(\beta - \epsilon_M)}$, $\epsilon_2 = \frac{\epsilon_u\epsilon_M^2}{\beta - \epsilon_M}$, then:

$$\begin{aligned}
|\delta| &= |(\beta - \epsilon_M)\delta_1 + \frac{\beta - \epsilon_M}{\beta}\delta_2| \\
&= \frac{(\beta - \epsilon_M)\epsilon_u\epsilon_M}{\beta} - \frac{\epsilon_u\epsilon_M^2}{\beta} + \frac{(\beta - \epsilon_M)\epsilon_u}{\beta} - \frac{\epsilon_u\epsilon_M^2}{\beta} \\
&= \frac{\epsilon_u}{\beta}(\beta + (\beta - 1)\epsilon_M - 3\epsilon_M^2) \\
&> \frac{\epsilon_u}{\beta} \cdot \beta = \epsilon_u
\end{aligned}$$

The result contradicts the conclusion of the model of machine arithmetic, which says $|\delta| < \epsilon_u$.

XIII

In single precision FPNs of IEEE 754, $[128, 129]$ can be represented as $[1, 1 + \frac{1}{2^7}] \times 2^7$(not expanded into binary form for convenience). If we want the absolute accuracy of 10^{-6}, i.e. 2^{-19} or 2^{-20}, it means the significand should be 2^{-26} or 2^{-27}, since the exponent part is 2^7. But this exceed the bits of single precision, which is only 23.

Therefore we cannot compute the root with absolute accuracy $< 10^{-6}$.

Exercise 4.33

a

$a = 1.234 \times 10^4$, $b = 8.769 \times 10^4$

(i) do nothing.
(ii) $m_c \leftarrow 10.003$.
(iii) $m_c \leftarrow 1.0003$; $e_c \leftarrow 5$.
(iv) do nothing.
(v) $m_c \leftarrow 1.000$.
(vi) $c = 1.000 \times 10^5$.

b

$a = 1.234 \times 10^4$, $b = -5.678 \times 10^0$

(i) $b \leftarrow -0.0005678 \times 10^4$; $e_c \leftarrow 4$.
(ii) $m_c \leftarrow 1.2334322$.
(iii) do nothing.
(iv) do nothing.
(v) $m_c \leftarrow 1.233$.
(vi) $c = 1.233 \times 10^4$.

c

$a = 1.234 \times 10^4$, $b = -5.678 \times 10^3$

(i) $b \leftarrow -0.5678 \times 10^4$; $e_c \leftarrow 4$.
(ii) $m_c \leftarrow 0.7662$.
(iii) $m_c \leftarrow 7.662$; $e_c \leftarrow 3$.
(iv) do nothing.
(v) do nothing.
(vi) $c = 7.662 \times 10^3$.

Exercise 4.42

Let $a_0 = 1, a_1 = 2, a_2 = 3$.

(i) Add in the ascending order:

- $s_1 = 3$, $s_2 = 6$,$\delta_1 = \epsilon_0$
- $\delta_2 = \epsilon_1 + \delta_1(1+\epsilon_1)\frac{s_1}{s_2} = \epsilon_1 + \frac{\epsilon_0}{2}(1+\epsilon_1)$

(ii) Add in the descending order:

- $s_1 = 5$, $s_2 = 6$,$\delta_1 = \epsilon_0$
- $\delta_2 = \epsilon_1 + \delta_1(1+\epsilon_1)\frac{s_1}{s_2} = \epsilon_1 + \frac{5\epsilon_0}{6}(1+\epsilon_1)$

Compare 2 results and δ_2 in case (i) is smaller.

Exercise 4.43

By neglecting the terms of $O(\delta_i^2)$, we get:

$$\begin{aligned}
&\mathrm{fl}(a_1b_1 + a_2b_2 + a_3b_3) \\
&= \mathrm{fl}(\mathrm{fl}(\mathrm{fl}(a_1b_1) + \mathrm{fl}(a_2b_2)) + \mathrm{fl}(a_3b_3)) \\
&= ((a_1b_1(1+\delta_1) + a_2b_2(1+\delta_2))(1+\delta_3) + a_3b_3(1+\delta_4))(1+\delta_5) \\
&= (a_1b_1 + a_2b_2 + a_3b_3) \\
&\cdot (1 + \delta_5 + (\delta_1+\delta_3)\frac{a_1b_1}{a_1b_1+a_2b_2+a_3b_3} + (\delta_2+\delta_3)\frac{a_2b_2}{a_1b_1+a_2b_2+a_3b_3} + \delta_4\frac{a_3b_3}{a_1b_1+a_2b_2+a_3b_3}) \\
&< (a_1b_1 + a_2b_2 + a_3b_3)(1+3\epsilon_u)
\end{aligned}$$

Guess that $\mathrm{fl}(\sum_{i=1}^m \prod_{j=1}^n a_{i,j}) = (\sum_{i=1}^m \prod_{j=1}^n a_{i,j})(1+(n-1)m\delta_n)$, where $|\delta_n| < \epsilon_u$.

Exercise 4.80

Similar with Example 4.79:

$$f_A = \mathrm{fl}\left[\frac{\mathrm{fl}(\sin x)}{\mathrm{fl}(1+\mathrm{fl}(\cos x))}\right]$$

$$f_A(x) = \frac{\sin x(1+\delta_3)}{(1+\cos x(1+\delta_1))(1+\delta_2)}(1+\delta_4)$$

Neglecting the terms of $O(\delta_i^2)$, the above equation is equivalent to

$$f_A(x) = \frac{\sin x}{1+\cos x}(1 + \delta_3 + \delta_4 - \delta_2 - \delta_1 \frac{\cos x}{1+\cos x})$$

Hence we have $\phi(x) = 3 + \frac{\cos x}{1+\cos x}$ and

$$\mathrm{cond}_A(x) \leq \frac{\sin x}{x}(3 + \frac{\cos x}{1+\cos x}).$$

Obviously $\mathrm{cond}_A(x)$ is bounded for $x \in (0, \pi/2)$.