Numerical Analysis HW4

数学与应用数学 2002 王锦宸

November 2022

1

The normalized FPN of 477 is 1.11011101×2^8.

2

The normalized FPN of $\frac{3}{5}$ is $1.001\ 1001 \cdots \times 2^{-1}$.

3

Let the normalized representation of $x = 1.000\ldots0 \times \beta^e$(there are p digits).
Thus, $x_L = \overline{(\beta-1).(\beta-1)(\beta-1)\ldots(\beta-1)0} \times \beta^{e-1}$, $x_R = 1.000\ldots1 \times \beta^e$.
Then, we have $x_R - x = \beta^{e-p}$ and $x - x_L = \beta^{e-p-1}$, therefore $x_R - x = \beta(x - x_L)$.

4

$x_L = 1.001\ 1001\ 1001\ 1001\ 1001\ 1001 \times 2^{-1}$, $x_R = 1.001\ 1001\ 1001\ 1001\ 1001\ 1010 \times 2^{-1}$
Thus, $x - x_L = \frac{3}{5} \times 2^{-24}$, $x_R - x = \frac{2}{5} \times 2^{-24}$,$fl(x) = x_R$ and $error\ = \frac{2}{3} \times 2^{-24}$

5

It'll be $\epsilon = 2^{-23}$

6

$fl(cos(\frac{1}{4})) = (0.1111100\ldots) \times 2^0 = (1.1111100\ldots) \times 2^{-1}$,
$fl(1) = (1.0000\ldots0) \times 2^0$.
Thus, $fl(1) - fl(cos(\frac{1}{4})) = (0.0000011\ldots) \times 2^0 = (1.1\ldots) \times 2^{-6}$
It loses 6 bits of precision.

7

1.Taylor Expansion 1 - cos(x) $= 1-(1-\frac{x^2}{2!}+\frac{x^4}{4!}-\dots) = \frac{x^2}{2!}-\frac{x^4}{4!}+\dots$
2.Use trigonometric formula $1-cos(x) = 2sin(\frac{x}{2})^2$

8

· $f(x) = (x-1)^\alpha, f'(x) = \alpha(x-1)^{\alpha-1}, C_f(x) = |\frac{\alpha x(x-1)^{\alpha-1}}{(x-1)^\alpha}| = \alpha\frac{x}{x-1}$. Thus, when $\alpha \neq 0$, $C_f(x)$ is large when $x \to \infty$
· $f(x) = \ln(x), f'(x) = \frac{1}{x}, C_f(x) = \left|\frac{1}{\ln(x)}\right|, C_f(x)$is large when $x \to 0$.
· $f(x) = e^x, f'(x) = e^x, C_f(x) = |x|, C_f(x)$ is large when $|x| \to \infty$.
· $f(x) = \arccos(x), C_f(x) = \left|\frac{x}{\sqrt{1-x^2}\arccos(x)}\right|, C_f(x)$ is large when $|x| \to 1$.

9

9.1

$f(x) = 1-e^{-x},\ f'(x) = e^{-x},\ C_f(x) = |\frac{x}{e^x-1}|$
It's monotonically descending in [0,1] and $C_f(x)_{max} = C_f(0) = 1$, thus $C_f(x) \in [0,1]$.

9.2

$\text{cond}_A(x) = \frac{1}{\epsilon_u}\inf_{f(x_A)=f_A(x)}\frac{|x_A-x|}{|x|}$. Because $\forall x \in \mathbf{F}, |f(x)-f_A(x)| = |f(x)-f(x_A)| = |f'(\xi)|\,|\,x-x_A\,| \leqslant e\epsilon_u, \xi \in [x, x_A]$, so $\text{cond}_A(x) \leqslant \frac{e}{|x|}$.

9.3

The following graph depicts $cond_f$ and the upper bound $cond_A$ on $[0,1]$.

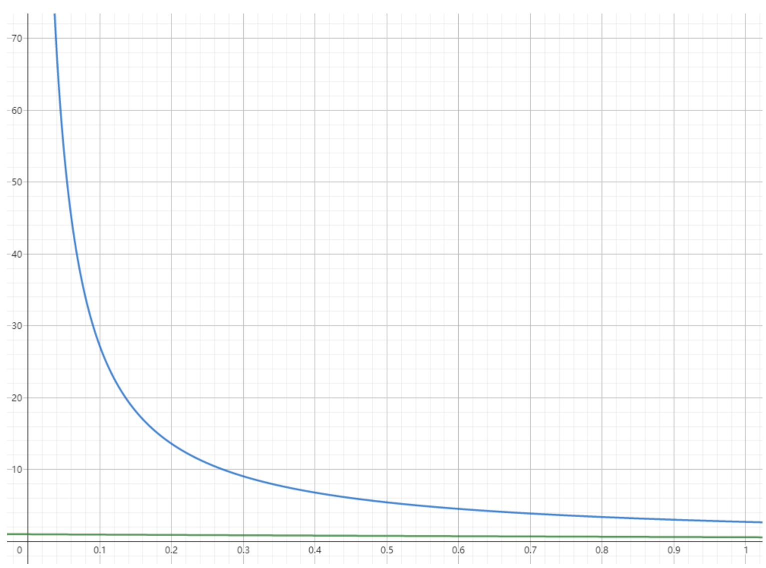

图 1: $cond_f$ and $cond_A$

From the graph, we know that $cond_f$ is small on the whole interval while $cond_A \to \infty$ when $x \to 0$ We notice that f(0) = 0 and $f_A(x) = f(x)(1+\delta(x))$. When $\delta(x) \to \infty$,

10

For $r = f\left(a_0, a_1, \cdots, a_{n-1}\right) \neq 0, a_i(x) = \left|\frac{a_i \frac{\partial r}{\partial a_i}}{r}\right|$. Since r is the root of p(x), $\sum_{i=0}^{n-1} a_i r^i = 0$, we have $\frac{\partial r}{\partial a_i} = -\frac{r^i}{\sum_{j=1}^{n-1} j a_j r^{j-1}} = -\frac{r^i}{p'(r)} \cdot a_i(x) = \left|\frac{a_i r^{i-1}}{p'(r)}\right|$.Thus $\text{cond}_f(x) = \|A(x)\|_1 = \max_i a_i(x) = \max_i \left|\frac{a_i r^{i-1}}{p'(r)}\right|$.

Put it into Wilkinson example, consider the condition number for $f(x) = \prod_{k=1}^{p}(x-k)$, at point p , we have $\text{cond}_f(x) = \max_i \left|\frac{a_i p^{i-1}}{(p-1)!}\right| \geqslant \frac{\sum_{k=1}^{p} k p^{p-2}}{(p-1)!} = \frac{(p+1)p^{p-1}}{2(p-1)!}$. Thus we know that the difficulty of solving polynomials with high degrees is out of its high condition number.

11

In the FPN system (2,2,-1,1), $a = 1.0 \times 2^0$, $b = 1.1 \times 2^0$. Then $\frac{a}{b} = 0.101$(of precision 4), so $fl(\frac{a}{b}) = 1.0 \times 2^{-1}$ and $error(\frac{a}{b}) = 0.01 = \epsilon_u$, which is contradictory to the model of arithmetic.

12

In IEEE 754, the parameters of single precision FPN is (2,24,-126,127). The root in the interval $[128, 129]$ will be represented as $m \times 2^7$, thus the distance between adjacent floating point is $2^7 \times \epsilon_M = 2^{-16} \approx 1.525 \times 10^{-5} > 10^{-6}$.

13

For $s(x) = ax^3 + bx^2 + cx + d$, we need to know the values of $s(x), s'(x)$ at $x_i, xi+1$.Thus we need to solve the equations with the coefficient matrix,

$$\begin{bmatrix} x_i^3 & x_i^2 & x_i & 1 \\ x_{i+1}^3 & x_{i+1}^2 & x_{i+1} & 1 \\ 3x_i^2 & 2x_i & 1 & 0 \\ 3x_{i+1}^2 & 2x_{i+1} & 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} f\left(x_i\right) \\ f\left(x_{i+1}\right) \\ f'\left(x_i\right) \\ f'\left(x_{i+1}\right) \end{bmatrix}$$

When x_i is close to x_{i+1}, the condition number will be large, thus it will get inaccurate number.

Programming 1

The output of 10 equally spaced points between [0.99, 1.01] is,(It's just for display and we draw the plot with 100 points)

f	g	h
$1.77636e-15$	$-1.11022e-15$	$1e-16$
$3.55271e-15$	$-6.66134e-16$	$1.67772e-17$
$5.32907e-15$	$9.99201e-16$	$1.67962e-18$
0	$-3.55271e-15$	$6.5536e-20$
$1.77636e-15$	$1.0103e-14$	$2.56e-22$
0	0	0
$-1.24345e-14$	$2.22045e-16$	$2.56e-22$
$3.55271e-15$	$-1.77636e-15$	$6.5536e-20$
$5.32907e-15$	$-2.44249e-15$	$1.67962e-18$
$-3.55271e-15$	$2.66454e-15$	$1.67772e-17$
$1.95399e-14$	$8.88178e-16$	$1e-16$

and the figure is,

Although those three function are "similar", we can see big difference from the chart and the

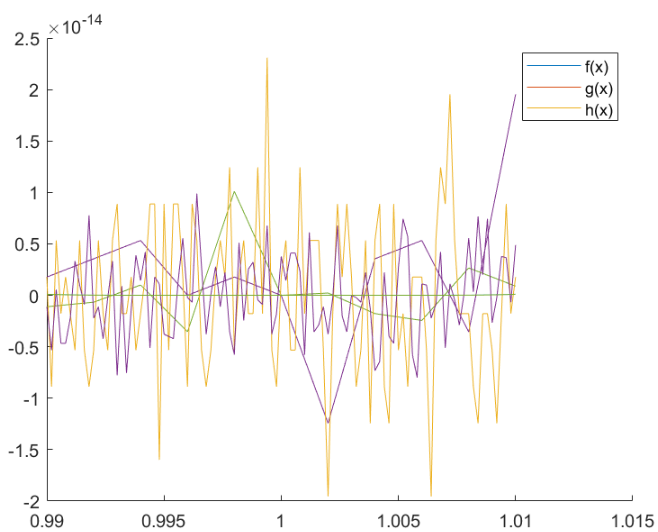

图 2: $f(x)$, $g(x)$ and $h(x)$

plot. The purple line and the yellow line are obviously unstable and have great vibration. That is because the yellow one has the greatest arithmetic operations and has the most accumulated error while the purple one is next to is.