Name : Haowen Chang UTORid : changh31

Student # : 1006205394 Tutorial Section : TUT0012

I declare that this test was written by the person whose name and student # appear above

Signature : Haowen Chang

Question 1). $\mathbb{R}_6(7,2)$. with exponent range $-(55)_6 \le e \le (55)_6$

a). $(.0000001)_6 \cdot 6^{-(55)_6}$ if not normalized

$(.1000000)_6 \cdot 6^{-(55)_6}$ if normalized

b). $(.5555555)_6 \cdot 6^{(55)_6}$

c) $(407)_{10}$

Numerator	Denominator	Quotient	Reminder
407	6	67	5
67	6	11	1
11	6	1	5
1	6	0	1

$(407)_{10} = (1515)_6$

In $\mathbb{R}_6(7,2)$: $+(.1515000)_6 \cdot 6^{(4)_6}$

d). $(0.9)_{10}$

multiplier	base	product	integer	fraction
0.9	6	5.4	5	0.4
0.4	6	2.4	2	0.4
0.4	6	2.4	2	0.4
		⋮		

$(0.9)_{10} = (0.5\dot{2})_6$ In $\mathbb{R}_6(7,2)$: $+(.5222222)_6 \cdot 6^{(0)_6}$

e). $(1515)_6 + (0.5222222)_6 = (1515.5222222)_6$

$fl(1515.5222222) = +(.1515522)_6 \cdot 6^{(4)_6}$

f) $|\delta| < \frac{1}{2} b^{(1-t)} = (\frac{1}{2})(6^{1-7}) = (\frac{1}{2})(6^{-6}) = \frac{6^{-6}}{2}$ b/c $b=6$, $t=7$ by definition of relative error.

Question 2:

a). $(2-(2-x))$

There will be subtractive cancellation occurred when x close to 2 and x close to 0.

Alternative: $(2-(2-x)) = 2-2+x = x$

use x instead of $(2-(2-x))$

b). $\sqrt{1+x} - \sqrt{1-x}$

There will be subtractive cancellation when x close to 0

But not subtractive cancellation when x close to 1. since x should be ≤ 1 to find a real number result

and $\sqrt{1-x}$ will be evaluated to some value close to 0.

Alternative: $(\sqrt{1+x} - \sqrt{1-x}) \dfrac{(\sqrt{1+x} + \sqrt{1-x})}{(\sqrt{1+x} + \sqrt{1-x})}$

$$= \frac{(1+x)-(1-x)}{\sqrt{1+x}+\sqrt{1-x}} = \frac{1+x-1+x}{\sqrt{1+x}+\sqrt{1-x}} = \frac{2x}{\sqrt{1+x}+\sqrt{1-x}}$$

c). $1-\sin(x)$

when x close to $2k\pi + \frac{\pi}{2}$, $k \in [0, \dots \infty)$, $k \in \mathbb{Z}$ b/c $\sin(2k\pi + \pi/2) = 1$

Alternative: $1-\sin(x) = 1-\sin(x) \cdot \dfrac{1+\sin(x)}{1+\sin(x)}$

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

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

d) $e^x - 1$

when x close to 0 will have subtractive cancellation. b/c $e^0 = 1$

Alternative: $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$

$$\Rightarrow e^x - 1 = \sum_{i=0}^{\infty} \frac{x^i}{i!} - 1$$

$$= \sum_{i=1}^{\infty} \frac{x^i}{i!}$$

Question 3).

$$A = \begin{bmatrix} 2 & 6 & 6 \\ 1 & 7 & 6 \\ 4 & 12 & 12 \end{bmatrix} \qquad b = \begin{bmatrix} 20 \\ 25 \\ 40 \end{bmatrix}$$

a) $PA = LU$.

$$P_1 = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad P_1 A = \begin{bmatrix} 4 & 12 & 12 \\ 2 & 6 & 6 \\ 1 & 7 & 6 \end{bmatrix} \quad L_1 = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ -\frac{1}{4} & 0 & 1 \end{bmatrix} \quad L_1 P_1 A = \begin{bmatrix} 4 & 12 & 12 \\ 0 & 0 & 0 \\ 0 & 4 & 3 \end{bmatrix}$$

$$P_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad P_2 L_1 P_1 A = \begin{bmatrix} 4 & 12 & 12 \\ 0 & 4 & 3 \\ 0 & 0 & 0 \end{bmatrix} \quad L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad U = L_2 P_2 L_1 P_1 A = \begin{bmatrix} 4 & 12 & 12 \\ 0 & 4 & 3 \\ 0 & 0 & 0 \end{bmatrix}$$

$$P = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad L = (T_1^{-1} L_2^{-1}) = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{4} & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{bmatrix} \quad U = \begin{bmatrix} 4 & 12 & 12 \\ 0 & 4 & 3 \\ 0 & 0 & 0 \end{bmatrix}$$

b) $PAx = Pb \Leftrightarrow Ld = Pb$ where $d = Ux$

solve for d: $\begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{4} & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{bmatrix} d = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 20 \\ 25 \\ 40 \end{bmatrix} = \begin{bmatrix} 40 \\ 25 \\ 20 \end{bmatrix} \Rightarrow d = \begin{bmatrix} 40 \\ 15 \\ 0 \end{bmatrix}$

solve for x: $\begin{bmatrix} 4 & 12 & 12 \\ 0 & 4 & 3 \\ 0 & 0 & 0 \end{bmatrix} x = \begin{bmatrix} 40 \\ 15 \\ 0 \end{bmatrix}$ $\quad \begin{cases} 4x_1 + 12x_2 + 12x_3 = 40 \\ 4x_2 + 3x_3 = 15 \end{cases}$ $\quad x = \begin{bmatrix} -2 \\ 3 \\ 1 \end{bmatrix}$

There could be infinite solutions since thepe is a freedom variable x_3, $x = \begin{bmatrix} -2 \\ 3 \\ 1 \end{bmatrix}$ is one of its solutions.

Question 4)

Let $A \in \mathbb{R}^{n\times n}$ be an non-singular matrix. let $x_i, t_i \in \mathbb{R}^n$, $i = 1, \dots, k$.

a) k linear systems: $Ax_1 = t_1$, $Ax_2 = t_2$, ..., $Ax_k = t_k$

According to the lecture, the complexity of LU factorization is: $\frac{n^3}{3} + O(n^2)$ flops.

After we find the L, U, P for matrix A, in $\frac{n^3}{3} + O(n^2)$ flops, we can start to solve k linear systems

Each linear system forward elimination + backward substitution takes $n^2 + O(n)$ flops.

$\Rightarrow k(n^2 + O(n))$ flops

In total $\frac{n^3}{3} + kO(n^2)$

b). Let $x_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ $x_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$ ... $x_i = \begin{cases} 1 & \text{for row } i \\ 0 & \text{otherwise} \end{cases}$ for all $i \in [1, \dots, n]$

After we solved all n linear systems, we put t_is in order from 1 to n to form a matrix which is A^{-1}

$A^{-1} = \begin{bmatrix} | & | & | & & | \\ t_1 & t_2 & t_3 & \cdots & t_n \\ | & | & | & & | \end{bmatrix}$ b/c $[A \mid I] = [I \mid A^{-1}]$

c) No, the Inverse of A then takes $\Rightarrow O(n^3)$ $\frac{n^3}{3} + O(n^2)\cdot n$ to solve, and to solve the system, we would also need $\frac{n^2}{2} + O(n)$ flops for each system

Therefore, it would cost $O(n^3) + k(\frac{n^2}{2} + O(n))$ to solve k linear systems, worse than GE.

Question 5)

Let $\hat{x}$ be a computed solution to $Ax = b$, $A \in \mathbb{R}^{n\times n}$, $x, b \in \mathbb{R}^n$

$$\frac{\|x-\hat{x}\|}{\|x\|} \le \text{cond}(A)\frac{\|r\|}{\|b\|} \quad \text{where } r \in \mathbb{R}^n,\ r = b - A\hat{x}$$

Starting with $Ax = b$ and $A\hat{x} = b - r$, derive a lower bound for $\frac{\|x-\hat{x}\|}{\|x\|}$

$$Ax - A\hat{x} = b - b + r$$

$$\Leftrightarrow A(x-\hat{x}) = r \Rightarrow \|r\| \le \|A\|\,\|x-\hat{x}\| \Rightarrow \|x-\hat{x}\| \ge \frac{\|r\|}{\|A\|} \quad ①$$

$$Ax = b \Leftrightarrow x = A^{-1}b \Rightarrow \|x\| \le \|A^{-1}\|\,\|b\|$$

$$\Rightarrow \frac{1}{\|x\|} \ge \frac{1}{\|A^{-1}\|\,\|b\|} \quad ②$$

Combine ① and ②: $$\frac{\|x-\hat{x}\|}{\|x\|} \ge \frac{1}{\|A\|\,\|A^{-1}\|}\cdot\frac{\|r\|}{\|b\|}$$

$$\Leftrightarrow \frac{\|x-\hat{x}\|}{\|x\|} \ge \frac{1}{\text{cond}(A)}\cdot\frac{\|r\|}{\|b\|}$$ is the lower bound.

Explaination of $\hat{x}$:

for upper bound and lower bound $\text{cond}(A) \ge 1$

$\Rightarrow$ if cond(A) is close to 1 / not too large, the problem is well-conditioned. $\Rightarrow$ small relative residual is a reliable indicator of small relative error and $\hat{x}$, $\hat{x}$ is likely to be more reliable with lower relative residual

else, if cond(A) is large, the problem is poorly conditioned $\Rightarrow$ small relative residual doesn't mean small relative error

$\hat{x}$ with small relative residual is not reliable