THE UNIVERSITY OF TEXAS
AT AUSTIN

CS383C NUMERICAL ANALYSIS

HW05 Numerical Stability

Edited by LaTeX

Department of Computer Science

STUDENT

Jimmy Lin

xl5224

COURSE COORDINATOR

Robert A. van de Geijn

UNIQUE NUMBER

53180

RELEASE DATE

Oct. 08 2014

DUE DATE

Oct. 14 2014

TIME SPENT

10 hours

October 12, 2014

Exercises

Exercise 3.

3.1

Write 1 as floating number.

$$.1\underbrace{00\cdots0}_{t-1}\times 2^1 \tag{1}$$

3.2

Show that $\mathbf{u} = \frac{1}{2}\cdot 2^{1-t}$

Proof. Let $\chi = .\delta_0\delta_1\cdots\delta_{t-1}\delta_t\cdots$ and $\check{\chi}$ to be the value stored in t-digit floating number with rounding mechanism. Then if $\delta_t = 0$, then $\chi = \check{\chi}$ and $|\delta\chi| = 0 \le 2^{e-t-1}$. But if $\delta_t = 1$, due to the rounding mechanism, then $\chi < \check{\chi}$ and

$$|\delta\chi| = |\chi - \check{\chi}| = |.\delta_0\delta_1\cdots\delta_{t-1}\delta_t\cdots\times 2^e - .\delta_0\delta_1\cdots\delta'_{t-1}\times 2^e| \le .\underbrace{00\cdots0}_{t}1\times 2^e = 2^{e-t-1} \tag{2}$$

For χ, since $\delta_0 = 1$ (normalized)

$$|\chi| = |.\delta_0\delta_1 ...\times 2^e| \ge .1\times 2^e \ge 2^{e-1} \tag{3}$$

Thus,

$$\frac{|\delta\chi|}{|\chi|} \le \frac{2^{e-t-1}}{2^{e-1}} = \frac{1}{2}\cdot 2^{1-t} \tag{4}$$

Then,

$$|\delta\chi| \le \frac{1}{2}\cdot 2^{1-t}|\chi| \tag{5}$$

Now, we have

$$\mathbf{u} = \frac{1}{2}2^{1-t} \tag{6}$$

□

Exercise 10.

Show that $|AB| \le |A||B|$.

Proof. Let $C = AB$. And the (i,j) entry of $|C|$ is given by

$$|c_{i,j}| = |\sum_{p=0}^{k} a_{i,k}b_{k,j}| \le \sum_{p=0}^{k}|a_{i,k}b_{k,j}| \le \sum_{p=0}^{k}|a_{i,k}||b_{k,j}| \tag{7}$$

which equals (i,j) entry of $|A||B|$. Hence, we have

$$|AB| \le |A||B| \tag{8}$$

□

Exercise 12.

12.1

Show that if $|A| \leq |B|$, then $||A||_1 \leq ||B||_1$.

Proof. Partition $A_{m\times n} = \left(a_0 \mid a_1 \mid ... \mid a_{n-1} \right)$ where a_j indicates the j-th column of matrix A. Similarly, we partition $B_{m\times n} = \left(b_0 \mid b_1 \mid ... \mid b_{n-1} \right)$ where b_j indicates the j-th column of matrix B. Then we use a_{ij} and b_{ij} to denote i-th element of a_j and b_j respectively.

$$||A||_1 = \max_{0\leq j<n} ||a_j||_1 = \max_{0\leq j<n} \sum_{i=0}^{m-1} |a_{ij}| \leq \max_{0\leq j<n} \sum_{i=0}^{m-1} |b_{ij}| = \max_{0\leq j<n} ||b_j||_1 = ||B||_1 \tag{9}$$

□

Lemma 1. *For arbitrary matrix* $A = \left(a_0 \mid a_1 \mid ... \mid a_{n-1} \right)$, $||A||_1 = \max_{0\leq j<n} ||a_j||_1$.

Proof. This lemma has been proved in Notes on Norms. □

12.2

Show that if $|A| \leq |B|$, then $||A||_\infty \leq ||B||_\infty$.

Proof. Partition $A_{m\times n} = \begin{pmatrix} a_0 \\ \hline a_1 \\ \hline \vdots \\ \hline a_{m-1} \end{pmatrix}$, where a_i indicates the i-th row of matrix A. Similarly, we partition $B_{m\times n} = \begin{pmatrix} b_0 \\ \hline b_1 \\ \hline \vdots \\ \hline b_{m-1} \end{pmatrix}$, where b_i indicates the i-th row of matrix B. Then we use a_{ij} and b_{ij} to denote j-th element of a_i and b_i respectively.

$$||A||_\infty = \max_{0\leq i<m} ||a_i||_1 = \max_{0\leq i<m} \sum_{j=0}^{n-1} |a_{ij}| \leq \max_{0\leq i<m} \sum_{j=}^{n-1} |b_{ij}| = \max_{0\leq i<m} ||b_i||_1 = ||B||_\infty \tag{10}$$

□

Lemma 2. *For arbitrary matrix* $A = \begin{pmatrix} a_0 \\ \hline a_1 \\ \hline \vdots \\ \hline a_{m-1} \end{pmatrix}$, $||A||_\infty = \max_{0\leq i<m} ||a_i||_1$.

Proof. This lemma has been proved in Notes on Norms. □

12.3

Show that if $|A| \leq |B|$, then $||A||_F \leq ||B||_F$. Let $A, B \in \mathbb{R}^{m\times n}$ and a_{ij}, b_{ij} be (i, j) entry of A, B respectively.

Proof.

$$||A||_F = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} |a_{ij}|^2 \leq \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} |b_{ij}|^2 = ||B||_F \tag{11}$$

□

Exercise 13.

$$
\begin{aligned}
\check{\kappa} &= [(\chi_0\psi_0 + \chi_1\psi_1) + \chi_2\psi_2] && (12)\\
&= [[\chi_0\psi_0 + \chi_1\psi_1] + [\chi_2\psi_2]] && (13)\\
&= [[[\chi_0\psi_0] + [\chi_1\psi_1]] + [\chi_2\psi_2]] && (14)\\
&= [[\chi_0\psi_0(1+\epsilon_*^{(0)}) + \chi_1\psi_1(1+\epsilon_*^{(1)})] + \chi_2\psi_2(1+\epsilon_*^{(2)})] && (15)\\
&= [\big(\chi_0\psi_0(1+\epsilon_*^{(0)}) + \chi_1\psi_1(1+\epsilon_*^{(1)})\big)(1+\epsilon_+^{(1)}) + \chi_2\psi_2(1+\epsilon_*^{(2)})] && (16)\\
&= \Big(\big(\chi_0\psi_0(1+\epsilon_*^{(0)}) + \chi_1\psi_1(1+\epsilon_*^{(1)})\big)(1+\epsilon_+^{(1)}) + \chi_2\psi_2(1+\epsilon_*^{(2)})\Big)(1+\epsilon_+^{(2)}) && (17)\\
&= \chi_0\psi_0(1+\epsilon_*^{(0)})(1+\epsilon_+^{(1)})(1+\epsilon_+^{(2)}) + \chi_1\psi_1(1+\epsilon_*^{(1)})(1+\epsilon_+^{(1)})(1+\epsilon_+^{(2)}) + \chi_2\psi_2(1+\epsilon_*^{(2)})(1+\epsilon_+^{(2)}) && (18)\\
&= \begin{pmatrix}\chi_0\\ \chi_1\\ \chi_2\end{pmatrix}^T \begin{pmatrix}\epsilon_0(1+\epsilon_*^{(0)})(1+\epsilon_+^{(1)})(1+\epsilon_+^{(2)})\\ \epsilon_1(1+\epsilon_*^{(1)})(1+\epsilon_+^{(1)})(1+\epsilon_+^{(2)})\\ \epsilon_2(1+\epsilon_*^{(2)})(1+\epsilon_+^{(2)})\end{pmatrix} && (19)\\
&= \begin{pmatrix}\chi_0\\ \chi_1\\ \chi_2\end{pmatrix}^T \left(\begin{array}{c|c|c}(1+\epsilon_*^{(0)})(1+\epsilon_+^{(1)})(1+\epsilon_+^{(2)}) & 0 & 0\\ 0 & (1+\epsilon_*^{(1)})(1+\epsilon_+^{(1)})(1+\epsilon_+^{(2)}) & 0\\ 0 & 0 & (1+\epsilon_*^{(2)})(1+\epsilon_+^{(2)})\end{array}\right) \begin{pmatrix}\epsilon_0\\ \epsilon_1\\ \epsilon_2\end{pmatrix} && (20)\\
&= \begin{pmatrix}\chi_0(1+\epsilon_*^{(0)})(1+\epsilon_+^{(1)})(1+\epsilon_+^{(2)})\\ \chi_1(1+\epsilon_*^{(1)})(1+\epsilon_+^{(1)})(1+\epsilon_+^{(2)})\\ \chi_2(1+\epsilon_*^{(2)})(1+\epsilon_+^{(2)})\end{pmatrix}^T \begin{pmatrix}\epsilon_0\\ \epsilon_1\\ \epsilon_2\end{pmatrix} && (21)
\end{aligned}
$$

Exercise 15.

Now we complete the missing part for the Inductive Step Case 1 of Lemma 14.

Proof. Case 1: $\prod_{i=0}^{n}(1+\epsilon)^{\pm 1} = \prod_{i=0}^{n-1}(1+\epsilon)^{\pm 1}(1+\epsilon_n)$.
By the inductive hypothesis, there exists a θ_n such that

$$(1+\theta_n) = \prod_{i=0}^{n-1}(1+\epsilon_i)^{\pm 1} \text{ and } |\theta_n| \leq n\mathbf{u}/(1-n\mathbf{u}) \tag{22}$$

Then

$$\prod_{i=0}^{n}(1+\epsilon)^{\pm 1} = \left(\prod_{i=0}^{n-1}(1+\epsilon)^{\pm 1}\right)(1+\epsilon_n) = (1+\theta_n)(1+\epsilon_n) = 1 + \underbrace{\theta_n + \epsilon_n + \theta_n \cdot \epsilon_n}_{\theta_{n+1}} \tag{23}$$

which tells us how to pick up θ_{n+1}. Then

$$
\begin{aligned}
|\theta_{n+1}| &= |\theta_n + \epsilon_n + \theta_n \cdot \epsilon_n| && (24)\\
&\leq |\theta_n| + |\epsilon_n| + |\theta_n| \cdot |\epsilon_n| && (25)\\
&= \frac{n\mathbf{u}}{1-n\mathbf{u}} + \mathbf{u} + \frac{n\mathbf{u}}{1-n\mathbf{u}} \cdot \mathbf{u} && (26)\\
&= \frac{n\mathbf{u} + \mathbf{u} - n\mathbf{u}^2 + n\mathbf{u}^2}{1-n\mathbf{u}} && (27)\\
&= \frac{(n+1)\mathbf{u}}{1-n\mathbf{u}} && (28)\\
&\leq \frac{(n+1)\mathbf{u}}{1-(n+1)\mathbf{u}} && (29)
\end{aligned}
$$

□

Exercise 18.

18.1

Show that if $n, b \geq 1$, then $\gamma_n \leq \gamma_{n+b}$.

Proof.

$$\gamma_n = \frac{n\mathbf{u}}{1 - n\mathbf{u}} \leq \frac{n\mathbf{u}}{1 - (n+b)\mathbf{u}} \leq \frac{(n+b)\mathbf{u}}{1 - (n+b)\mathbf{u}} = \gamma_{n+b} \tag{30}$$

Note that since $\mathbf{u}$ is extremely small, then $1 - n\mathbf{u} > 0$ and $1 - (n+b)\mathbf{u} > 0$. □

18.2

Show that if $n, b \geq 1$, then $\gamma_n + \gamma_b + \gamma_n\gamma_b \leq \gamma_{n+b}$.

Proof.

$$\begin{aligned}
& \gamma_n + \gamma_b + \gamma_n\gamma_b && (31)\\
=& \frac{n\mathbf{u}}{1 - n\mathbf{u}} + \frac{b\mathbf{u}}{1 - b\mathbf{u}} + \frac{n\mathbf{u}}{1 - n\mathbf{u}} \cdot \frac{b\mathbf{u}}{1 - b\mathbf{u}} && (32)\\
=& \frac{n\mathbf{u} - nb\mathbf{u}^2 + b\mathbf{u} - nb\mathbf{u}^2 + nb\mathbf{u}^2}{(1 - n\mathbf{u})(1 - b\mathbf{u})} && (33)\\
=& \frac{n\mathbf{u} - nb\mathbf{u}^2 + b\mathbf{u}}{1 - n\mathbf{u} - b\mathbf{u} + nb\mathbf{u}^2} && (34)\\
\leq& \frac{n\mathbf{u} + b\mathbf{u}}{1 - n\mathbf{u} - b\mathbf{u} + nb\mathbf{u}^2} && (35)\\
\leq& \frac{n\mathbf{u} + b\mathbf{u}}{1 - n\mathbf{u} - b\mathbf{u}} && (36)\\
=& \frac{(n+b)\mathbf{u}}{1 - (n+b)\mathbf{u}} && (37)\\
=& \gamma_{n+b} && (38)
\end{aligned}$$

Note that since $\mathbf{u}$ is extremely small, then $1 - n\mathbf{u} > 0$ and $1 - (n+b)\mathbf{u} > 0$. Also, note that $nb\mathbf{u}^2 \geq 0$. □

Exercise 19.

19.1 $k = 0$

Show that $\left(\begin{array}{c|c} I + \Sigma^{(k)} & 0 \\ \hline 0 & (1+\epsilon_1) \end{array}\right)(1+\epsilon_2) = I + \Sigma^{(k+1)}$

Proof. Given that if $k = 0$, then $\epsilon_1 = 0$ and Σ^0 is 0×0 matrix, we have

$$(1+0)\cdot(1+\epsilon_2) = \underbrace{1}_{I} + \underbrace{\epsilon_2}_{\Sigma^{(1)}} \tag{39}$$

Thus, $\left(\begin{array}{c|c} I + \Sigma^{(k)} & 0 \\ \hline 0 & (1+\epsilon_1) \end{array}\right)(1+\epsilon_2) = I + \Sigma^{(k+1)}$ holds for $k = 0$. □

19.2 $k > 0$

Proof. For arbitrary $k > 0$,

$$\left(\begin{array}{c|c} I + \Sigma^{(k)} & 0 \\ \hline 0 & (1+\epsilon_1) \end{array}\right)(1+\epsilon_2) = \left(\begin{array}{c|c} (I + \Sigma^{(k)})(1+\epsilon_2) & 0 \\ \hline 0 & (1+\epsilon_1)(1+\epsilon_2) \end{array}\right) \tag{40}$$

$$= \left(\begin{array}{c|c} I + \epsilon_2 I + \Sigma^{(k)} + \epsilon_2\Sigma^{(k)} & 0 \\ \hline 0 & 1+\epsilon_1+\epsilon_2+\epsilon_1\epsilon_2 \end{array}\right) \tag{41}$$

$$= \underbrace{\left(\begin{array}{c|c} I & 0 \\ \hline 0 & 1 \end{array}\right)}_{I} + \underbrace{\left(\begin{array}{c|c} \epsilon_2 I + \Sigma^{(k)} + \epsilon_2\Sigma^{(k)} & 0 \\ \hline 0 & \epsilon_1+\epsilon_2+\epsilon_1\epsilon_2 \end{array}\right)}_{\Sigma^{(k+1)}} \tag{42}$$

Note that the the definition of $\Sigma^{(k+1)}$ are valid because if $\Sigma^{(k)}$ is diagonal, then so as $\Sigma^{(k+1)}$. Hence,

$$\left(\begin{array}{c|c} I + \Sigma^{(k)} & 0 \\ \hline 0 & (1+\epsilon_1) \end{array}\right)(1+\epsilon_2) = I + \Sigma^{(k+1)} \text{ holds for } k > 0. \tag{43}$$

□

Exercise 23.

23.1

Show that $\check{\kappa} = (x + \delta x)^T y$, where $|\delta x| \le \gamma_n |x|$.

Proof. Let $\delta x = \Sigma^{(n)} x$, where $\Sigma^{(n)}$ is as in Theorem 20.

$$|\delta x| = |\Sigma^{(n)} x| = \begin{pmatrix} |\theta_n \chi_0| \\ |\theta_n \chi_1| \\ \vdots \\ |\theta_2 \chi_{n-1}| \end{pmatrix} \le \begin{pmatrix} |\theta_n||\chi_0| \\ |\theta_n||\chi_1| \\ \vdots \\ |\theta_2||\chi_{n-1}| \end{pmatrix} \le |\theta_n| \begin{pmatrix} |\chi_0| \\ |\chi_1| \\ \vdots \\ |\chi_{n-1}| \end{pmatrix} \le \gamma_n \begin{pmatrix} |\chi_0| \\ |\chi_1| \\ \vdots \\ |\chi_{n-1}| \end{pmatrix} = \gamma_n |x| \tag{44}$$

Thus, it can be concluded for the backward analysis that

$$|\delta x| \le \gamma_n |x| \tag{45}$$

□

23.2

Show that $\check{\kappa} = x^T (y + \delta y)$, where $|\delta y| \le \gamma_n |y|$.

Proof. The proof for perturbation on input y is the same as that of perturbation on input x. □

Exercise 25.

Proof. We partition matrix $A \in \mathbb{R}^{m\times n}$ and have

$$A = \begin{pmatrix} a_0^T \\ a_1^T \\ \vdots \\ a_{m-1}^T \end{pmatrix} \tag{46}$$

Then in terms of algorithm in Fig. 4 and R1-B,

$$\check{y} = \begin{pmatrix} \psi_0 \\ \psi_1 \\ \vdots \\ \psi_{m-1} \end{pmatrix} = \begin{pmatrix} a_0^T(x+\delta x_0) \\ a_1^T(x+\delta x_1) \\ \vdots \\ a_{m-1}^T(x+\delta x_{m-1}) \end{pmatrix} = Ax + \begin{pmatrix} a_0^T\delta x_0 \\ a_1^T\delta x_1 \\ \vdots \\ a_{m-1}^T\delta x_{m-1} \end{pmatrix} \tag{47}$$

where $\forall i,\ |\delta x_i| \le \gamma_n |x|$. However, $\begin{pmatrix} a_0^T\delta x_0 \\ a_1^T\delta x_1 \\ \vdots \\ a_{m-1}^T\delta x_{m-1} \end{pmatrix}$ cannot be consolidated into $A\delta x$, then

$$\text{We } \textbf{cannot} \text{ prove that } \check{y} = A(x+\delta x) \text{ where } \delta x \text{ is small.} \tag{48}$$

□

Exercise 27.

By suffering similar problem as the exercise 25 does, we have

$$\text{We cannot prove that } C = (A+\Delta A)(B+\Delta B) \text{ where } \Delta A \text{ and } \Delta A \text{ is small.} \tag{49}$$