MA 522 Solution for Homework Assignment #5
(Based on the solutions of Isaac J. Lee)

---

1. *Let $A = LU$ be the LU-factorization of A with $|l_{ij}| \leq 1$. Let $\vec{a_i}^T$ and $\vec{u_i}^T$ denote the i-th row of A and U, respectively. Verify that $\vec{u_i}^T = \vec{a_i}^T - \sum_{j=1}^{i-1} l_{ij}\vec{u_j}^T$. Use it to show that $||U||_{max} \leq 2^{n-1}\,||A||_{max}$ and $||U||_\infty \leq 2^{n-1}\,||A||_\infty$.*

$A = LU$, so $\vec{a_i}^T = \sum_{j=1}^{n} l_{ij}\vec{u_j}^T = \sum_{j=1}^{i} l_{ij}\vec{u_j}^T$, since $l_{ij} = 0$ for all $j > i$. Subtracting $\sum_{j=1}^{i-1} l_{ij}\vec{u_j}^T$ from both sides, we get

$$\vec{a_i}^T - \sum_{j=1}^{i-1} l_{ij}\vec{u_j}^T = l_{ii}\vec{u_i}^T = 1 \cdot \vec{u_i}^T = \vec{u_i}^T.$$

We prove that (1) $||\vec{u_i}||_\infty \leq 2^{i-1}\,||A||_{max}$ by induction, but first we note that $||A||_{max} \geq ||\vec{a_i}||_\infty$ for $i = 1, \cdots, n$ by the definitions of the two norms.

Base: If $i = 1$, then $\vec{u_1} = \vec{a_1}$, so $||\vec{u_1}||_\infty = ||\vec{a_1}||_\infty \leq ||A||_{max} = 2^{1-1}\,||A||_{max}$.

Induction: Assume for some k with $2 \leq k \leq n$ that (1) is true for all $i \leq k-1$. If $i = k$, then $\vec{u_k} = \vec{a_k} - \sum_{j=1}^{k-1} l_{ij}\vec{u_j}$, so we have the following:

$$\begin{aligned}
||\vec{u_k}||_\infty &\leq ||\vec{a_k}||_\infty + \sum_{j=1}^{k-1} ||l_{ij}\vec{u_j}||_\infty && \text{by the Triangle Inequality}\\
&\leq ||\vec{a_k}||_\infty + \sum_{j=1}^{k-1} ||\vec{u_j}||_\infty && ||l_{ij}\vec{u_j}||_\infty = |l_{ij}| \cdot ||\vec{u_j}||_\infty \leq 1 \cdot ||\vec{u_j}||_\infty\\
&\leq ||\vec{a_k}||_\infty + \sum_{j=1}^{k-1} 2^{j-1}\,||A||_{max} && \text{by the inductive assumption } (1 \leq j \leq k-1)\\
&\leq ||A||_{max} + ||A||_{max} \cdot \sum_{j=1}^{k-1} 2^{j-1} && ||\vec{a_k}||_\infty \leq ||A||_{max}\\
&= ||A||_{max}\left(1 + (2^{k-1} - 1)\right) && \text{partial geometric sum}\\
&= 2^{k-1}\,||A||_{max}.
\end{aligned}$$

Therefore, for some $i \in \{1, \cdots, n\}$, we have $||U||_{max} = ||\vec{u_i}||_\infty \leq 2^{i-1}\,||A||_{max} \leq 2^{n-1}\,||A||_{max}$.

We similarly prove that (2) $||\vec{u_i}||_1 \leq 2^{i-1}\,||A||_\infty$ by induction, and we note that $||A||_\infty \geq ||\vec{a_i}||_1$ for $i = 1, \cdots, n$ by the definitions of the two norms.

Base: If $i = 1$, then $\vec{u_1} = \vec{a_1}$, so $||\vec{u_1}||_1 = ||\vec{a_1}||_1 \leq ||A||_\infty = 2^{1-1}\,||A||_\infty$.

Induction: Assume that (2) is true for all $i \leq k-1$. If $i = k$, then $\vec{u_k} = \vec{a_k} - \sum_{j=1}^{k-1} l_{ij}\vec{u_j}$, so we have the following (refer to above for intermediary steps):

$$\begin{aligned}
||\vec{u_k}||_1 &\leq ||\vec{a_k}||_1 + \sum_{j=1}^{k-1} ||l_{ij}\vec{u_j}||_1\\
&\leq ||\vec{a_k}||_1 + \sum_{j=1}^{k-1} 2^{j-1}\,||A||_\infty\\
&\leq ||A||_\infty + ||A||_\infty \cdot \sum_{j=1}^{k-1} 2^{j-1}\\
&= 2^{k-1}\,||A||_\infty.
\end{aligned}$$

Therefore, for some $i \in \{1, \cdots, n\}$, we have $||U||_\infty = ||\vec{u_i}||_1 \leq 2^{i-1}\,||A||_\infty \leq 2^{n-1}\,||A||_\infty$.

2. *Let U be an upper triangular matrix (the diagonals need not be ones). Let $\hat{x}$ be the computed solution to $U\vec{x} = \vec{y}$ using backward substitution. Prove that $(U + \delta U)\hat{x} = \vec{y}$ with $|\delta U| \le (n\epsilon + O(\epsilon^2))\,|U|$.*

Let $\hat{x}_i = fl\left(fl(y_i - fl(\sum_{j=i+1}^{n} u_{ij}\hat{x}_j))/u_{ii}\right) = (y_i - \sum_{j=i+1}^{n} u_{ij}\hat{x}_j(1+\delta_{ij}))(1+\delta_0)/u_{ii} \cdot (1+\delta_0')$ denote the computed $\vec{x}_i$. Then, we have

$$\begin{aligned}
\frac{u_{ii}\hat{x}_i}{(1+\delta_0)(1+\delta_0')} &= y_i - \sum_{j=i+1}^{n} u_{ij}\hat{x}_j(1+\delta_{ij}) && |\delta_{ij}| \le (n-i)\epsilon + O(\epsilon^2);\ |\delta_0|, |\delta_0'| \le \epsilon \\
\Rightarrow u_{ii}(1+\delta_{ii})\hat{x}_i &= y_i - \sum_{j=i+1}^{n} u_{ij}(1+\delta_{ij})\hat{x}_j && \delta_{ii} = \frac{1}{(1+\delta_0)(1+\delta_0')} - 1;\ |\delta_{ii}| \le 2\epsilon + O(\epsilon^2) \\
\Rightarrow \sum_{j=i}^{n} u_{ij}(1+\delta_{ij})\hat{x}_j &= y_i \\
\Rightarrow \hat{U}\hat{x} &= \vec{y} && \hat{U} = [\hat{u}_{ij}] = [u_{ij}(1+\delta_{ij})];\ \hat{x} = [\hat{x}_1 \ \cdots \ \hat{x}_n]^T \\
\Rightarrow (U + \delta U)\hat{x} &= \vec{y}. && \delta U = \hat{U} - U = [\hat{u}_{ij} - u_{ij}] = [\delta_{ij}u_{ij}]
\end{aligned}$$

Furthermore, $|\delta U| = |\delta_{ij}| \cdot |[u_{ij}]| \le \big((n-i+1)\epsilon + O(\epsilon^2)\big) \cdot |U| \le \big(n\epsilon + O(\epsilon^2)\big)\ |U|$.