FDU 数值算法 Homework 03

Due: Oct. 8, 2024
姓名: 雍崔扬
学号: 21307140051

Problem 1

Assume that $b, \delta b, x, \delta x$ satisfy:

$$\begin{cases} Ax = b \\ A(x+\delta x) = b + \delta b \end{cases} \text{where } A = \begin{bmatrix} 610 & 987 \\ 987 & 1597 \end{bmatrix}$$

Construct examples such that:

- ① $\frac{\|\delta b\|_\infty}{\|b\|_\infty}$ is very small while $\frac{\|\delta x\|_\infty}{\|x\|_\infty}$ is very large
- ② $\frac{\|\delta b\|_\infty}{\|b\|_\infty}$ is very large while $\frac{\|\delta x\|_\infty}{\|x\|_\infty}$ is very small

---

TA (存疑): 有比特征向量更好的例子，可以考虑诱导范数的不等式 $\|Ax\|_\infty \leq \|A\|_\infty \|x\|_\infty$ 的取等条件.
(可以参考高等线性代数 Homework 03 Problem 01)

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \leq \|A\|_\infty \|A^{-1}\|_\infty \frac{\|\delta b\|_\infty}{\|b\|_\infty}$$

然后他又说考虑矩阵 (也许有搞头):

$$\begin{bmatrix} 610 & -987 \\ -987 & 1597 \end{bmatrix}$$

Solution:
根据 $\det(A) = 610 \times 1597 - 987^2 = 1$ 可知 A 非常接近于奇异.
根据 $\det(\lambda I - A) = (\lambda - 610)(\lambda - 1597) - (-987)^2 = \lambda^2 - 2207\lambda + 1$ 可看出 $\lambda_1 \approx 2207, \lambda_2 \approx \frac{1}{2207}$
注意到 A 是一个实对称阵 (自然是正规矩阵)，故其存在谱分解
且其奇异值即为特征值，左、右奇异向量即为特征向量，换言之，其奇异值分解即为谱分解.

```
% 定义矩阵 A
A = [610, 987; 987, 1597];

% 计算特征分解
[V, D] = eig(A);
disp('特征分解: ');
disp('特征值 D: ');
disp(D);
disp('特征向量 V: ');
disp(V);
```

运行上述代码得到:
(这里我们设特征值和奇异值都按从大到小排列，尽管一般情况下我们习惯将特征值从小达大排列)

$$A = Q\Lambda Q^T = [q_1, q_2]\text{diag}\{\lambda_1, \lambda_2\}[q_1, q_2]^T = \begin{bmatrix} 0.5257 & -0.8507 \\ 0.8507 & 0.5257 \end{bmatrix} \begin{bmatrix} 2207.0 & \\ & 0.00045 \end{bmatrix} \begin{bmatrix} 0.5257 & -0.8507 \\ 0.8507 & 0.5257 \end{bmatrix}^T$$

以下 Matlab 函数用于在给定 $b, \delta b$ 的条件下计算 $\begin{cases} Ax = b \\ A(x+\delta x) = b + \delta b \end{cases}$ 中 $x, \delta x$，并输出 $\frac{\|\delta b\|_\infty}{\|b\|_\infty}$ 和 $\frac{\|\delta x\|_\infty}{\|x\|_\infty}$:

```
function calculate_linear_system(b, delta_b)
    % 定义矩阵 A
```

```
    A = [610, 987; 987, 1597];

    % 计算 x
    x = A \ b;  % 求解 Ax = b

    % 计算 delta x
    delta_x = A \ (b + delta_b) - x;

    % 计算相对误差
    relative_error_b = norm(delta_b, Inf) / norm(b, Inf);
    relative_error_x = norm(delta_x, Inf) / norm(x, Inf);

    % 打印结果
    disp('解 x: ');
    disp(x);
    disp('扰动 delta x: ');
    disp(delta_x);
    disp('相对误差 ||delta b||_inf / ||b||_inf: ');
    disp(relative_error_b);
    disp('相对误差 ||delta x||_inf / ||x||_inf: ');
    disp(relative_error_x);
end
```

- ① 为构造 $b, \delta b$ 使得 $\frac{\|\delta b\|_\infty}{\|b\|_\infty}$ 很小而 $\frac{\|\delta x\|_\infty}{\|x\|_\infty}$ 很大，
  我们可以利用小特征值 σ_2 的特征向量 q_2 构造扰动，而大特征值 σ_1 的特征向量 q_1 作为原始数据:

$$b = q_1 = \begin{bmatrix} 0.5257 \\ 0.8507 \end{bmatrix}$$

$$\delta b = 10^{-2} \times q_2 = 10^{-2} \times \begin{bmatrix} -0.8507 \\ 0.5257 \end{bmatrix}$$

$$\frac{\|\delta b\|_\infty}{\|b\|_\infty} = 0.01$$

  运行 `calculate_linear_system(b, delta_b)` 得到结果为:

$$x = \begin{bmatrix} -0.0980 \\ 0.0611 \end{bmatrix}$$

$$\delta x = \begin{bmatrix} -1.8774 \\ 1.1774 \end{bmatrix} \times 10^5$$

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \approx 1.9157 \times 10^6$$

- ② 为构造 $b, \delta b$ 使得 $\frac{\|\delta b\|_\infty}{\|b\|_\infty}$ 很大而 $\frac{\|\delta x\|_\infty}{\|x\|_\infty}$ 很小，
  我们可以利用大特征值 σ_1 的特征向量 q_1 构造扰动，而小特征值 σ_2 的特征向量 q_2 作为原始数据:

$$b = q_2 = \begin{bmatrix} -0.8507 \\ 0.5257 \end{bmatrix}$$

$$\delta b = 10^2 \times q_1 = 10^2 \times \begin{bmatrix} 0.5257 \\ 0.8507 \end{bmatrix}$$

$$\frac{\|\delta b\|_\infty}{\|b\|_\infty} = 100$$

  运行 `calculate_linear_system(b, delta_b)` 得到结果为:

$$x = \begin{bmatrix} -1.8774 \\ 1.1603 \end{bmatrix} \times 10^3$$

$$\delta x = \begin{bmatrix} -9.8000 \\ 6.1100 \end{bmatrix} \times 10^{-3}$$

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \approx 5.2199 \times 10^{-7}$$

Problem 2

Let $A = \begin{bmatrix} I_n & Z \\ 0_{n\times n} & I_n \end{bmatrix}$ where $Z \in \mathbb{C}^{n\times n}$.
Find $\kappa_{\mathrm{F}}(A) = \|A\|_{\mathrm{F}}\|A^{-1}\|_{\mathrm{F}}$

Solution:
首先注意到:

$$\begin{bmatrix} I_n & Z \\ & I_n \end{bmatrix}\begin{bmatrix} I_n & -Z \\ & I_n \end{bmatrix} = \begin{bmatrix} I_n & -Z+Z \\ & I_n \end{bmatrix} = \begin{bmatrix} I_n & \\ & I_n \end{bmatrix}$$

因此 $A^{-1} = \begin{bmatrix} I_n & -Z \\ 0_{n\times n} & I_n \end{bmatrix}$

$$\begin{aligned}
\|A\|_{\mathrm{F}}^2 &= \operatorname{tr}\left(A^H A\right) \\
&= \operatorname{tr}\left(\begin{bmatrix} I_n & Z \\ & I_n \end{bmatrix}^H \begin{bmatrix} I_n & Z \\ & I_n \end{bmatrix}\right) \\
&= \operatorname{tr}\left(\begin{bmatrix} I_n & Z \\ Z^H & Z^H Z + I_n \end{bmatrix}\right) \\
&= \operatorname{tr}\left(I_n\right) + \operatorname{tr}\left(Z^H Z + I_n\right) \\
&= n + \operatorname{tr}\left(Z^H Z\right) + n \\
&= \|Z\|_{\mathrm{F}}^2 + 2n \\
\hline
\|A^{-1}\|_{\mathrm{F}}^2 &= \operatorname{tr}\left(\left(A^{-1}\right)^H A^{-1}\right) \\
&= \operatorname{tr}\left(\begin{bmatrix} I_n & -Z \\ & I_n \end{bmatrix}^H \begin{bmatrix} I_n & -Z \\ & I_n \end{bmatrix}\right) \\
&= \operatorname{tr}\left(\begin{bmatrix} I_n & -Z \\ -Z^H & Z^H Z + I_n \end{bmatrix}\right) \\
&= \operatorname{tr}\left(I_n\right) + \operatorname{tr}\left(Z^H Z + I_n\right) \\
&= n + \operatorname{tr}\left(Z^H Z\right) + n \\
&= \|Z\|_{\mathrm{F}}^2 + 2n \\
\hline
\kappa_{\mathrm{F}}(A) &= \|A\|_{\mathrm{F}}\|A^{-1}\|_{\mathrm{F}} \\
&= \sqrt{\|Z\|_{\mathrm{F}}^2 + 2n}\sqrt{\|Z\|_{\mathrm{F}}^2 + 2n} \\
&= \|Z\|_{\mathrm{F}}^2 + 2n
\end{aligned}$$

因此 $\kappa_{\mathrm{F}}(A) = \|Z\|_{\mathrm{F}}^2 + 2n$

Problem 3

It can be shown that Gaussian elimination without pivoting is numerically stable for solving diagonally dominant linear systems, in the sense that the growth factor is bounded.

Give a concrete upper bound on the growth factor $\rho = \max\limits_{i,j,k} \frac{|a_{ij}^{(k)}|}{\|A\|_\infty}$

- **一点观察: (存疑)**
  对于严格对角占优阵来说，部分主元可能会破坏严格对角占优性，此时应选择全选主元.
  此时全选主元的代价与部分主元相近，因为我们只需在主对角元中选择主元就可以了.
  特殊地，若严格对角占优阵是对称的，
  那么不选主元的效果就等价于部分主元 (此时在列上也是对角元严格主导的，因此部分主元是可以用的)
- **(数值线性代数, 习题 2.19)**
  若 A^T 是严格对角占优阵，则部分主元的 Gauss 消去法的增长因子 $\rho = \frac{\max_{ij}|a_{ij}^{(n-1)}|}{\max_{ij}|a_{ij}|} \leq 2$

Solution:
记 $A=[a_1,\dots,a_n]\in\mathbb{C}^{n\times n}$，其 (i,j) 位置上的元素记为 a_{ij}
根据 A 的严格对角占优性可知 $a_{11}\neq 0$

现考虑第一步 Gauss 消元:

令 $l_1=\dfrac{1}{a_{11}}a_1-e_1=\dfrac{1}{a_{11}}\begin{bmatrix}0\\a_{21}\\\vdots\\a_{n1}\end{bmatrix}$ 并定义 Gauss 变换矩阵 $L_1=I_n-l_1e_1^T$ (其中 e_1 代表 $\mathbb{C}^n$ 的第 1 个标准单位基向量)

将其作用到 A 上，则我们有:

$$\begin{aligned}L_1A&=(I_n-l_1e_1^T)A\\&=A-l_1e_1^TA\\&=A-\frac{1}{a_{11}}\begin{bmatrix}0\\a_{21}\\\vdots\\a_{n1}\end{bmatrix}[e_1^Ta_1,e_1^Ta_2,\dots,e_1^Ta_n]\\&=A-\frac{1}{a_{11}}\begin{bmatrix}0\\a_{21}\\\vdots\\a_{n1}\end{bmatrix}[a_{11},a_{12},\dots,a_{1n}]\\&=A-\frac{1}{a_{11}}\begin{bmatrix}0&0&\cdots&0\\a_{21}a_{11}&a_{21}a_{12}&\cdots&a_{21}a_{1n}\\\vdots&\vdots&&\vdots\\a_{n1}a_{11}&a_{n1}a_{12}&\cdots&a_{n1}a_{1n}\end{bmatrix}\end{aligned}$$

若记 $A^{(1)}=L_1A=[b_{ij}]_{i,j=1}^n$ 并考虑其右下方的 $(n-1)\times(n-1)$ 子矩阵，则我们有:

$$\begin{aligned}\begin{bmatrix}b_{22}&b_{23}&\cdots&b_{2n}\\b_{32}&b_{33}&\cdots&b_{3n}\\\vdots&\vdots&&\vdots\\b_{n2}&b_{n3}&\cdots&b_{nn}\end{bmatrix}&=\begin{bmatrix}a_{22}&a_{23}&\cdots&a_{2n}\\a_{32}&a_{33}&\cdots&a_{3n}\\\vdots&\vdots&&\vdots\\a_{n2}&a_{n3}&\cdots&a_{nn}\end{bmatrix}-\frac{1}{a_{11}}\begin{bmatrix}a_{21}a_{12}&a_{21}a_{13}&\cdots&a_{21}a_{1n}\\a_{31}a_{12}&a_{31}a_{13}&\cdots&a_{31}a_{1n}\\\vdots&\vdots&&\vdots\\a_{n1}a_{12}&a_{n1}a_{13}&\cdots&a_{n1}a_{1n}\end{bmatrix}\\&=\left[a_{ij}-\frac{1}{a_{11}}a_{i1}a_{1j}\right]_{i,j=2}^n\end{aligned}$$

因此我们有 $b_{ij}=a_{ij}-\frac{1}{a_{11}}a_{i1}a_{1j}\ \ (i,j=2,\dots,n)$
对于任意 $i=2,\dots,n$，考虑对角元 b_{ii} 与非对角元 $b_{ij}\ (j\neq i)$:

$$\begin{aligned}|b_{ii}|-\sum_{\substack{j\neq i\\2\le j\le n}}|b_{ij}|&=\left|a_{ii}-\frac{1}{a_{11}}a_{i1}a_{1i}\right|-\sum_{\substack{j\neq i\\2\le j\le n}}\left|a_{ij}-\frac{1}{a_{11}}a_{i1}a_{1j}\right|\quad\text{(triangle inequality)}\\&\ge|a_{ii}|-|a_{i1}|\frac{|a_{1i}|}{|a_{11}|}-\sum_{\substack{j\neq i\\2\le j\le n}}\left(|a_{ij}|+|a_{i1}|\frac{|a_{1j}|}{|a_{11}|}\right)\\&=|a_{ii}|-\sum_{\substack{j\neq i\\2\le j\le n}}|a_{ij}|-|a_{i1}|\cdot\left\{\frac{1}{|a_{11}|}\sum_{\substack{j\neq i\\1\le j\le n}}|a_{i1}|\right\}\quad\text{(use strict diagonally dominantion)}\\&>|a_{ii}|-\sum_{\substack{j\neq i\\2\le j\le n}}|a_{ij}|-|a_{i1}|\cdot 1\\&=|a_{ii}|-\sum_{\substack{j\neq i\\1\le j\le n}}|a_{ij}|\quad\text{(strict diagonally dominantion)}\\&>0\end{aligned}$$

表明 L_1A 的右下方的 $(n-1)\times(n-1)$ 子矩阵也是严格对角占优的，因而 L_1A 也是严格对角占优的.

可以证明，$A^{(1)}=L_1A=[b_{ij}]_{i,j=1}^n$ 的行和范数小于等于 A 的行和范数:

- ① $A^{(1)}$ 与 A 的第一行相同
- ② 对于任意 $i=2,\dots,n$，我们都有:

$$\begin{aligned}\sum_{j=1}^n|b_{ij}|&=|0|+\sum_{j=2}^n|b_{ij}|\\&=0+\sum_{j=2}^n\left|a_{ij}-\frac{1}{a_{11}}a_{i1}a_{1j}\right|\\&\le\sum_{j=2}^n|a_{ij}|+\frac{\sum_{j=2}^n|a_{1j}|}{|a_{11}|}|a_{i1}|\\&<\sum_{j=2}^n|a_{ij}|+1\cdot|a_{i1}|\\&=\sum_{j=1}^n|a_{ij}|\end{aligned}$$

综合 ①② 我们有:

$$\|A^{(1)}\|_\infty=\max_{1\le i\le n}\sum_{j=1}^n|b_{ij}|<\max_{1\le i\le n}\sum_{j=1}^n|a_{ij}|=\|A\|_\infty$$

以此类推，我们可以知道第 $k=1,\dots,n-1$ 步 Gauss 消元得到的矩阵 $A^{(k)}$ 都是严格对角占优的，
且满足 $\|A^{(n-1)}\|_\infty<\cdots<\|A^{(1)}\|_\infty<\|A\|_\infty$
因此我们有:

$$\rho=\max_{\substack{1\le i,j\le n\\k=1,\dots,n-1}}\frac{|a_{ij}^{(k)}|}{\|A\|_\infty}\le\max_{k=1,\dots,n-1}\frac{\|A^{(k)}\|_\infty}{\|A\|_\infty}<1$$

Problem 4

It can be shown that Gaussian elimination with partial pivoting is numerically stable for solving nonsingular tridiagonal linear systems, in the sense that the growth factor is bounded.

Give a concrete upper bound on the growth factor $\rho=\max\limits_{i,j,k}\frac{|a_{ij}^{(k)}|}{\|A\|_\infty}$

- **TA:** 简单的分析可知，其误差传递最多跨越两行就断了，因此我们猜测 ρ 的上界是 2
- **(数值线性代数, 习题 2.18)**

  若 A 是三对角阵，则部分主元的 Gauss 消去法的增长因子 $\rho=\frac{\max_{ij}|a_{ij}^{(k)}|}{\max_{ij}|a_{ij}|}\le 2$

Solution:
设 $A\in\mathbb{R}^{n\times n}$ 为如下形式的非奇异三对角阵:

$$A=\begin{bmatrix}\alpha_1&\gamma_1&&&\\\beta_1&\alpha_2&\gamma_2&&\\&\beta_2&\alpha_3&\ddots&\\&&\ddots&\ddots&\gamma_{n-1}\\&&&\beta_{n-1}&\alpha_n\end{bmatrix}=[a_{ij}^{(0)}]_{i,j=1}^n$$

记 $\|A\|_{\max} := \max_{1\le i,j\le n} |a_{ij}|$ (注意当 $n \ge 2$ 时，$\|\cdot\|_{\max}$ 并不是一个相容的矩阵范数)
记第 $k = 1, \dots, n-1$ 步部分主元的 Gauss 消元产生的矩阵为 $A^{(1)}, \dots, A^{(n-1)}$ (我们额外定义 $A^{(0)} = A$)
可以归纳证明以下命题:
对于任意 $k = 1, \dots, n-1$ 都有:

- ① 从某种意义上讲，$A^{(k)}$ 相对 $A^{(k-1)}$ 只改变了第 $k, k+1$ 行的第 $k, k+1, k+2$ 列上的 6 个元素.
  因此可知 $A^{(k)}$ 仅在第 1 条次对角线和第 $1, 2$ 条超对角线上可能存在非零元素.
  并且 $A^{(k)}$ 的右下 $(n-k) \times (n-k)$ 分块是三对角的.
- ② 第 $k = 1, \dots, n-2$ 步部分主元的 Gauss 消元产生的新元素满足:

$$|a^{(k)}_{k+1,k+1}| \le 2\|A\|_{\max}$$
$$|a^{(k)}_{k+1,k+2}| \le \|A\|_{\max}$$

  特殊地，第 $k = n-1$ 步部分主元的 Gauss 消元产生的新元素满足:

$$|a^{(n-1)}_{n,n}| \le 2\|A\|_{\max}$$

只要上述命题得证，我们就能知道对于任意 $k = 1, \dots, n-1$ 都有 $\|A^{(k)}\|_{\max} \le 2\|A\|_{\max} \le 2\|A\|_\infty$
于是得到增长因子的上界为:

$$\rho = \max_{\substack{1\le i,j\le n \\ k=1,\dots,n-1}} \frac{|a^{(k)}_{ij}|}{\|A\|_\infty} = \max_{k=1,\dots,n-1} \frac{\|A^{(k)}\|_{\max}}{\|A\|_\infty} \le 2$$

下面我们就用数学归纳法证明上述命题.

首先考虑第一步部分主元的 Gauss 消元:

- ① 若 $|\alpha_1| \ge |\beta_1|$，则无需更换主元.

$$l_1 = \frac{\beta_1}{\alpha_1} e_2 = \begin{bmatrix} 0 \\ \frac{\beta_1}{\alpha_1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$L_1 = I_n - l_1 e_1^T = \begin{bmatrix} 1 & & & \\ -\frac{\beta_1}{\alpha_1} & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}$$

$$A^{(1)} = L_1 A = \begin{bmatrix} \alpha_1 & \gamma_1 & & & \\ 0 & \alpha_2 - \frac{\beta_1}{\alpha_1}\gamma_1 & \gamma_2 & & \\ & \beta_2 & \alpha_3 & \ddots & \\ & & \ddots & \ddots & \gamma_{n-1} \\ & & & \beta_{n-1} & \alpha_n \end{bmatrix}$$

  我们发现 $A^{(1)}$ 的右下 $(n-1) \times (n-1)$ 分块仍是三对角的，且有:

$$\begin{aligned} |a^{(1)}_{22}| &= |\alpha_2 - \frac{\beta_1}{\alpha_1}\gamma_1| \\ &\le |\alpha_2| + \frac{|\beta_1|}{|\alpha_1|}|\gamma_1| \quad (\text{note that } |\alpha_1| \ge |\beta_1|) \\ &\le |\alpha_2| + 1 \cdot |\gamma_1| \\ &\le 2\|A\|_{\max} \end{aligned}$$

$$\begin{aligned} |a^{(1)}_{23}| &= |\gamma_2| \\ &\le \|A\|_{\max} \end{aligned}$$

- ② 若 $|\alpha_1| < |\beta_1|$，则需要更换主元.

$$P_1 = I_n - (e_1e_1^T + e_2e_2^T) + (e_1e_2^T + e_2e_1^T) = \begin{bmatrix} 0 & 1 & & & \\ 1 & 0 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix}$$

$$P_1A = \begin{bmatrix} \beta_1 & \alpha_2 & \gamma_2 & & & \\ \alpha_1 & \gamma_1 & 0 & & & \\ & \beta_2 & \alpha_3 & \gamma_3 & & \\ & & \beta_3 & a_4 & \ddots & \\ & & & \ddots & \ddots & \gamma_{n-1} \\ & & & & \beta_{n-1} & \alpha_n \end{bmatrix}$$

$$l_1 = \frac{\alpha_1}{\beta_1}e_2$$

$$L_1 = I_n - l_1e_1^T$$

$$L_1(P_1A) = \begin{bmatrix} \beta_1 & \alpha_2 & \gamma_2 & & & \\ 0 & \gamma_1 - \frac{\alpha_1}{\beta_1}\alpha_2 & -\frac{\alpha_1}{\beta_1}\gamma_2 & & & \\ & \beta_2 & \alpha_3 & \gamma_3 & & \\ & & \beta_3 & a_4 & \ddots & \\ & & & \ddots & \ddots & \gamma_{n-1} \\ & & & & \beta_{n-1} & \alpha_n \end{bmatrix}$$

我们发现 $A^{(1)}$ 的右下 $(n-1)\times(n-1)$ 分块仍是三对角的，且有:

$$\begin{aligned} |a_{22}^{(1)}| &= |\gamma_1 - \frac{\alpha_1}{\beta_1}\alpha_2| \\ &\le |\gamma_1| + \frac{|\alpha_1|}{|\beta_1|}|\alpha_2| \quad (\text{note that } |\alpha_1| < |\beta_1|) \\ &< |\gamma_1| + 1\cdot|\alpha_2| \\ &\le 2\|A\|_{\max} \end{aligned}$$

$$\begin{aligned} |a_{23}^{(1)}| &= |-\frac{\alpha_1}{\beta_1}\gamma_2| \\ &= \frac{|\alpha_1|}{|\beta_1|}|\gamma_2| \quad (\text{note that } |\alpha_1| < |\beta_1|) \\ &< 1\cdot|\gamma_2| \\ &\le \|A\|_{\max} \end{aligned}$$

综合 ①② 可知 $A^{(1)}$ 满足 $\begin{cases} |a_{22}^{(1)}| \le 2\|A\|_{\max} \\ |a_{23}^{(1)}| \le \|A\|_{\max} \end{cases}$，因此满足我们要证明的命题.

现假设前 k $(k \le n-2)$ 步消元得到的 $A^{(1)}, \dots, A^{(k)}$ 均满足我们要证明的命题
下面证明 $A^{(k+1)}$ 也满足我们要证明的命题:

- **(1) 当 $k \le n-3$ 时**
  注意到第 $k+1$ 次部分主元的 Gauss 消元中，无论是选主元还是 Gauss 消元，其视角都局限在以下位置:

$$\begin{bmatrix} a_{k+1,k+1}^{(k)} & a_{k+1,k+2}^{(k)} & 0 \\ \beta_{k+1} & \alpha_{k+2} & \gamma_{k+2} \end{bmatrix}$$

  - ① 若 $|\alpha_{k+1,k+1}^{(k)}| \ge |\beta_{k+1}|$，则无需更换主元，消元得到:

$$\begin{bmatrix} a_{k+1,k+1}^{(k)} & a_{k+1,k+2}^{(k)} & 0 \\ 0 & \alpha_{k+2}-\frac{\beta_{k+1}}{a_{k+1,k+1}^{(k)}}a_{k+1,k+2}^{(k)} & \gamma_{k+2} \end{bmatrix}$$

可以看出 $A^{(k+1)}$ 的右下 $(n-k-1)\times(n-k-1)$ 分块仍是三对角的，且有:

$$\begin{aligned} |a_{k+2,k+2}^{(k+1)}| &= \left|\alpha_{k+2}-\frac{\beta_{k+1}}{a_{k+1,k+1}^{(k)}}a_{k+1,k+2}^{(k)}\right| \\ &\le |\alpha_{k+2}|+\frac{|\beta_{k+1}|}{|a_{k+1,k+1}^{(k)}|}|a_{k+1,k+2}^{(k)}| \quad (\text{note that } |\alpha_{k+1,k+1}^{(k)}|\ge|\beta_{k+1}|) \\ &\le |\alpha_{k+2}|+1\cdot|a_{k+1,k+2}^{(k)}| \quad (\text{use inductive assumption } |a_{k+1,k+2}^{(k)}|\le\|A\|_{\max}) \\ &\le \|A\|_{\max}+\|A\|_{\max} \\ &\le 2\|A\|_{\max} \end{aligned}$$

$$\begin{aligned} |a_{k+2,k+3}^{(k+1)}| &= |\gamma_{k+2}| \\ &\le \|A\|_{\max} \end{aligned}$$

- ② 若 $|\alpha_{k+1,k+1}^{(k)}|<|\beta_{k+1}|$，则需要更换主元，置换第 $k+1,k+2$ 行，得到:

$$\begin{bmatrix} \beta_{k+1} & \alpha_{k+2} & \gamma_{k+2} \\ a_{k+1,k+1}^{(k)} & a_{k+1,k+2}^{(k)} & 0 \end{bmatrix}$$

消元得到:

$$\begin{bmatrix} \beta_{k+1} & \alpha_{k+2} & \gamma_{k+2} \\ 0 & a_{k+1,k+2}^{(k)}-\frac{\beta_{k+1}}{a_{k+1,k+1}^{(k)}}\alpha_{k+2} & -\frac{\beta_{k+1}}{a_{k+1,k+1}^{(k)}}\gamma_{k+2} \end{bmatrix}$$

可以看出 $A^{(k+1)}$ 的右下 $(n-k-1)\times(n-k-1)$ 分块仍是三对角的，且有:

$$\begin{aligned} |a_{k+2,k+2}^{(k+1)}| &= |a_{k+1,k+2}^{(k)}-\frac{\beta_{k+1}}{a_{k+1,k+1}^{(k)}}\alpha_{k+2}| \\ &\le |a_{k+1,k+2}^{(k)}|+\frac{|\beta_{k+1}|}{|a_{k+1,k+1}^{(k)}|}|\alpha_{k+2}| \quad (\text{note that } |a_{k+1,k+1}^{(k)}|<|\beta_{k+1}|) \\ &< |a_{k+1,k+2}^{(k)}|+1\cdot|\alpha_{k+2}| \quad (\text{use inductive assumption } |a_{k+1,k+2}^{(k)}|\le\|A\|_{\max}) \\ &\le \|A\|_{\max}+\|A\|_{\max} \\ &\le 2\|A\|_{\infty} \end{aligned}$$

$$\begin{aligned} |a_{k+2,k+3}^{(k+1)}| &= |-\frac{\beta_{k+1}}{a_{k+1,k+1}^{(k)}}\gamma_{k+2}| \\ &= \frac{|\beta_{k+1}|}{|a_{k+1,k+1}^{(k)}|}|\gamma_{k+2}| \quad (\text{note that } |a_{k+1,k+1}^{(k)}|<|\beta_{k+1}|) \\ &< 1\cdot|\gamma_{k+2}| \\ &\le \|A\|_{\max} \end{aligned}$$

综合 ①② 可知，当 $k\le n-3$ 时，$A^{(k+1)}$ 满足我们要证明的命题.

- **(2) 当 $k=n-2$ 时**
注意到第 $k+1=n-1$ 次部分主元的 Gauss 消元中，无论是选主元还是 Gauss 消元，其视角都局限在以下位置:

$$\begin{bmatrix} a_{n-1,n-1}^{(n-2)} & a_{n-1,n}^{(n-2)} \\ \beta_{n-1} & \alpha_n \end{bmatrix}$$

- ① 若 $|\alpha_{n-1,n-1}^{(n-2)}|\ge|\beta_{n-1}|$，则无需更换主元，消元得到:

$$\begin{bmatrix} a_{n-1,n-1}^{(n-2)} & a_{n-1,n}^{(n-2)} \\ 0 & \alpha_n-\frac{\beta_{n-1}}{a_{n-1,n-1}^{(n-2)}}a_{n-1,n}^{(n-2)} \end{bmatrix}$$

可以看出 $A^{(n-1)}$ 的右下 1×1 分块仍是三对角的，且有:

$$\begin{aligned}
|a_{n,n}^{(n-1)}| &= \left|\alpha_n - \frac{\beta_{n-1}}{a_{n-1,n-1}^{(n-2)}} a_{n-1,n}^{(n-2)}\right| \\
&\le |\alpha_n| + \frac{|\beta_{n-1}|}{|a_{n-1,n-1}^{(n-2)}|}|a_{n-1,n}^{(n-2)}| \quad (\text{note that } |\alpha_{n-1,n-1}^{(n-2)}| \ge |\beta_{n-1}|) \\
&\le |\alpha_n| + 1\cdot|a_{n-1,n}^{(n-2)}| \quad (\text{use inductive assumption } |a_{n-1,n}^{(n-2)}| \le \|A\|_{\max}) \\
&\le \|A\|_{\max} + \|A\|_{\max} \\
&\le 2\|A\|_{\max}
\end{aligned}$$

- ② 若 $|\alpha_{n-1,n-1}^{(n-2)}| < |\beta_{n-1}|$，则需要更换主元，置换第 $n-1, n$ 行，得到:

$$\begin{bmatrix} \beta_{n-1} & \alpha_n \\ a_{n-1,n-1}^{(n-2)} & a_{n-1,n}^{(n-2)} \end{bmatrix}$$

消元得到:

$$\begin{bmatrix} \beta_{n-1} & \alpha_n \\ & a_{n-1,n}^{(n-2)} - \frac{a_{n-1,n-1}^{(n-2)}}{\beta_{n-1}}\alpha_n \end{bmatrix}$$

可以看出 $A^{(n-1)}$ 的右下 1×1 分块仍是三对角的，且有:

$$\begin{aligned}
|a_{n,n}^{(n-1)}| &= \left|a_{n-1,n}^{(n-2)} - \frac{a_{n-1,n-1}^{(n-2)}}{\beta_{n-1}}\alpha_n\right| \\
&\le |a_{n-1,n}^{(n-2)}| + \frac{|a_{n-1,n-1}^{(n-2)}|}{|\beta_{n-1}|}|\alpha_n| \quad (\text{note that } |\alpha_{n-1,n-1}^{(n-2)}| < |\beta_{n-1}|) \\
&\le |a_{n-1,n}^{(n-2)}| + 1\cdot|\alpha_n| \quad (\text{use inductive assumption } |a_{n-1,n}^{(n-2)}| \le \|A\|_{\max}) \\
&\le \|A\|_{\max} + \|A\|_{\max} \\
&\le 2\|A\|_{\max}
\end{aligned}$$

综合 ①② 可知，当 $k = n-2$ 时，$A^{(k+1)} = A^{(n-1)}$ 也满足我们要证明的命题.

综合 $(1)(2)$ 可知命题得证.
根据前文的推理可知增长因子 $\rho \le 2$

Problem 5

Solve the following linear systems, using Gaussian elimination with and without pivoting:

$$\begin{bmatrix}
8 & 1 & & & & & \\
6 & 8 & 1 & & & & \\
 & 6 & 8 & 1 & & & \\
 & & \ddots & \ddots & \ddots & & \\
 & & & 6 & 8 & 1 & \\
 & & & & 6 & 8 & 1 \\
 & & & & & 6 & 8
\end{bmatrix}
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{98} \\ x_{99} \\ x_{100} \end{bmatrix}
=
\begin{bmatrix} 9 \\ 15 \\ 15 \\ \vdots \\ 15 \\ 15 \\ 14 \end{bmatrix}$$

$$\begin{bmatrix}
6 & 1 & & & & & \\
8 & 6 & 1 & & & & \\
 & 8 & 6 & 1 & & & \\
 & & \ddots & \ddots & \ddots & & \\
 & & & 8 & 6 & 1 & \\
 & & & & 8 & 6 & 1 \\
 & & & & & 8 & 6
\end{bmatrix}
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{98} \\ x_{99} \\ x_{100} \end{bmatrix}
=
\begin{bmatrix} 7 \\ 15 \\ 15 \\ \vdots \\ 15 \\ 15 \\ 14 \end{bmatrix}$$

Compare the computed solutions with the exact ones.
What do you find out about the accuracy?

Solution:

(1) Gauss 消去法

不选主元的 Gauss 消去法的算法如下:
(Gauss 消去法, 数值线性代数, 算法 1.1.3)

$$
\begin{aligned}
&\text{function: } [L,U]=\text{GaussianElimination}(A)\\
&\quad n=\dim(A)\\
&\quad \text{for } k=1:n-1\\
&\qquad A(k+1:n,k)\leftarrow A(k+1:n,k)/A(k,k)\\
&\qquad A(k+1:n,k+1:n)\leftarrow A(k+1:n,k+1:n)-A(k+1:n,k)A(k,k+1:n)\\
&\quad \text{end}\\
&\quad L=I_n+A\odot(\text{strictly lower triangular matrix with all ones})\\
&\quad U=A\odot(\text{upper triangular matrix with all ones})\\
&\quad \text{return } [L,U]
\end{aligned}
$$

Matlab 代码如下:

```matlab
function [L, U] = Gaussian_Elimination_without_pivoting(A)
    % Input:
    % A - An n x n matrix
    %
    % Output:
    % L - Lower triangular matrix
    % U - Upper triangular matrix

    % Get the size of the matrix A
    [n, ~] = size(A);

    % Perform Gaussian Elimination
    for k = 1:n-1
        % Update column elements below the diagonal
        A(k+1:n, k) = A(k+1:n, k) / A(k, k);

        % Update the remaining submatrix
        A(k+1:n, k+1:n) = A(k+1:n, k+1:n) - A(k+1:n, k) * A(k, k+1:n);
    end

    % Construct the lower triangular matrix L
    L = eye(n) + tril(A, -1);

    % Construct the upper triangular matrix U
    U = triu(A);

    % Return the results
    return;
end
```

部分主元的 Gauss 消去法的算法如下:
(列主元 Gauss 消去法, 数值线性代数, 算法 1.2.2)

function: $[P, L, U] = \text{GaussianEliminationPartialPivoting}(A)$

$n = \dim(A)$

$P = I_n$

for $k = 1 : n - 1$

$p \in \arg\max_{k \le i \le n} |A(i, k)|$ （确定主元位置）

$A(k, 1 : n) \leftrightarrow A(p, 1 : n)$ （交换第 k, p 行）

$P(k, 1 : n) \leftrightarrow P(p, 1 : n)$ （记录置换矩阵 P_k）

if $A(k, k) \neq 0$

（进行 Gauss 消去）

$A(k + 1 : n, k) = A(k + 1 : n, k + 1 : n)/A(k, k)$

$A(k + 1 : n, k + 1 : n) = A(k + 1 : n, k + 1 : n) - A(k + 1 : n, k)A(k, k + 1 : n)$

else

break （矩阵奇异）

end

$L = I_n + A \odot$ (strictly lower triangular matrix with all ones) ($\odot$ stands for Hadamand product)

$U = A \odot$ (upper triangular matrix with all ones)

return $[P, L, U]$

与不选主元的 Gauss 消去法一样，L 的严格下三角部分和 U 分别存储在 A 的严格下三角部分和上三角部分.
其 Matlab 代码如下:

```
function [P, L, U] = Gaussian_Elimination_Partial_Pivoting(A)
    % 获取矩阵的维度
    [n, m] = size(A);
    if n ~= m
        error('矩阵A必须是方阵');
    end

    % 初始化置换矩阵 P 为单位矩阵
    P = eye(n);

    % 高斯消去过程
    for k = 1:n-1
        % 在第 k 列的 A(k:n, k) 中找到最大值的行索引 p
        [~, p] = max(abs(A(k:n, k)));
        p = p + k - 1; % 调整为在整个矩阵中的行索引

        % 交换第 k 行和第 p 行
        if p ~= k
            A([k, p], :) = A([p, k], :);
            P([k, p], :) = P([p, k], :); % 记录行置换
        end

        % 检查主元是否为零
        if A(k, k) == 0
            error('矩阵是奇异的');
        end

        % Gauss 消去过程：对 A(k+1:n, k) 进行归一化
        A(k+1:n, k) = A(k+1:n, k) / A(k, k);

        % 更新 A(k+1:n, k+1:n)
        A(k+1:n, k+1:n) = A(k+1:n, k+1:n) - A(k+1:n, k) * A(k, k+1:n);
    end

    % 计算 L 和 U 矩阵
    L = tril(A, -1) + eye(n); % L 是单位下三角矩阵
    U = triu(A); % U 是上三角矩阵
end
```

(2) 回代法 & 前代法

根据不选主元的 Gauss 消元法得到 $A = LU$ 后，我们按如下步骤求解线性方程组 $Ax = b$:

- 用前代法求解 $Ly = b$ 得到 y
- 用回代法求解 $Ux = y$ 得到 x

或根据部分主元 Gauss 消去法 $PA = LU$ 后，我们按如下步骤求解线性方程组 $Ax = b$:

- 用前代法求解 $Ly = Pb$ 得到 y
- 用回代法求解 $Ux = y$ 得到 x

Matlab 代码如下:

```
% 求解线性方程组 Ax = b
function x = Solve_Linear_System(A, b, pivot)
    if (pivot == true)
        % 使用部分主元的 Gaussian 消去法计算 PA = LU
        [P, L, U] = Gaussian_Elimination_Partial_Pivoting(A);

        % 使用前代法求解 Ly = Pb
        y = Forward_Sweep(L, P * b);
    else
        % 使用不选主元的 Gaussian 消去法计算 A = LU
        [L, U] = Gaussian_Elimination_Without_Pivoting(A);

        % 使用前代法求解 Ly = b
        y = Forward_Sweep(L, b);
    end

    % 使用回代法求解 Ux = y
    x = Backward_Sweep(U, y);
end
```

(前代法, 数值线性代数, 算法 1.1.1)

$$
\begin{aligned}
&\text{function}: y = \text{ForwardSweep}[L, b] \\
&\quad \text{for } i = 1 : n-1 \\
&\quad\quad b(i) \leftarrow b(i)/L(i,i) \\
&\quad\quad b(i+1:n) \leftarrow b(i+1:n) - b(i)L(i+1:n, i) \\
&\quad \text{end} \\
&\quad b(n) \leftarrow b(n)/L(n,n) \\
&\quad \text{return } b
\end{aligned}
$$

最终 $Ly = b$ 的解 y 存储在 b 中.
Matlab 代码如下:

```
function y = Forward_Sweep(L, b)
    % 前代法求解 Ly = b
    n = length(b);
    for i = 1:n-1
        b(i) = b(i) / L(i, i);  % 对角线归一化
        b(i+1:n) = b(i+1:n) - b(i) * L(i+1:n, i);  % 消去
    end
    b(n) = b(n) / L(n, n);  % 处理最后一行
    y = b;  % 返回结果
end
```

(回代法, 数值线性代数, 算法 1.1.2)

$$
\begin{aligned}
&\text{function: } x = \text{BackwardSweep}[U, y] \\
&\quad \text{for } i = n : -1 : 2 \\
&\quad\quad y(i) \leftarrow y(i)/U(i,i) \\
&\quad\quad y(1:i-1) \leftarrow y(1:i-1) - y(i)U(1:i-1, i) \\
&\quad \text{end} \\
&\quad y(1) \leftarrow y(1)/U(1,1) \\
&\quad \text{return } y
\end{aligned}
$$

最终 $Ux = y$ 的解 x 存储在 y 中.
Matlab 代码如下:

```
function x = Backward_Sweep(U, y)
    % 回代法求解 Ux = y
    n = length(y);
    for i = n:-1:2
        y(i) = y(i) / U(i, i);  % 对角线归一化
        y(1:i-1) = y(1:i-1) - y(i) * U(1:i-1, i);  % 消去
    end
    y(1) = y(1) / U(1, 1);  % 处理第一行
    x = y;  % 返回结果
end
```

(3) 函数调用

```
% Define the size of the matrix
n = 100;

% Construct the first tridiagonal matrix for System 1
A1 = diag(8 * ones(1, n)) + diag(1 * ones(1, n-1), 1) + diag(6 * ones(1, n-1), -1);

% Condition number of A1
eigA1 = eig(A1' * A1);
kappaA1 = max(abs(eigA1)) / min(abs(eigA1));
disp('Condition number of A1:')
disp(kappaA1)

% Construct the second tridiagonal matrix for System 2
A2 = diag(6 * ones(1, n)) + diag(1 * ones(1, n-1), 1) + diag(8 * ones(1, n-1), -1);

% Condition number of A2
eigA2 = eig(A2' * A2);
kappaA2 = max(abs(eigA2)) / min(abs(eigA2));
disp('Condition number of A2:')
disp(kappaA2)

% Define the right-hand side vectors for both systems
b1 = [9; 15 * ones(n-2, 1); 14];
b2 = [7; 15 * ones(n-2, 1); 14];

% Solve System 1 using Gaussian Elimination without Pivoting
pivot = false;
x1_without_pivot = Solve_Linear_System(A1, b1, pivot);

% Solve System 1 using Gaussian Elimination with Partial Pivoting
pivot = true;
x1_with_pivot = Solve_Linear_System(A1, b1, pivot);

% Solve System 2 using Gaussian Elimination without Pivoting
pivot = false;
x2_without_pivot = Solve_Linear_System(A2, b2, pivot);
```

```
% Solve System 2 using Gaussian Elimination with Partial Pivoting
pivot = true;
x2_with_pivot = Solve_Linear_System(A2, b2, pivot);

% exact_solution of System 1 & 2
exact_solution = ones(n, 1);

% Compare the accuracy in System 1
disp('Difference in System 1 (Without Pivoting vs Exact Solution):');
disp(norm(x1_without_pivot - exact_solution));

disp('Difference in System 1 (With Pivoting vs Exact Solution):');
disp(norm(x1_with_pivot - exact_solution));

% Compare the accuracy in System 2
disp('Difference in System 2 (Without Pivoting vs Exact Solution):');
disp(norm(x2_without_pivot - exact_solution));

disp('Difference in System 2 (With Pivoting vs Exact Solution):');
disp(norm(x2_with_pivot - exact_solution));
```

输出结果:

```
Condition number of A1:
  218.7352

Condition number of A2:
   6.5769e+15

Difference in System 1 (Without Pivoting vs Exact Solution):
   1.6542e-15

Difference in System 1 (With Pivoting vs Exact Solution):
   1.6542e-15

Difference in System 2 (Without Pivoting vs Exact Solution):
   4.7575e+13

Difference in System 2 (With Pivoting vs Exact Solution):
    0.2479
```

(4) 理论分析

考虑如下线性系统:

$$A_1x=\begin{bmatrix}8&1&&&&&&\\6&8&1&&&&&\\&6&8&1&&&&\\&&\ddots&\ddots&\ddots&&&\\&&&&6&8&1&\\&&&&&6&8&1\\&&&&&&6&8\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\\\vdots\\x_{98}\\x_{99}\\x_{100}\end{bmatrix}=\begin{bmatrix}9\\15\\15\\\vdots\\15\\15\\14\end{bmatrix}=b_1$$

$$A_2x=\begin{bmatrix}6&1&&&&&&\\8&6&1&&&&&\\&8&6&1&&&&\\&&\ddots&\ddots&\ddots&&&\\&&&&8&6&1&\\&&&&&8&6&1\\&&&&&&8&6\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\\\vdots\\x_{98}\\x_{99}\\x_{100}\end{bmatrix}=\begin{bmatrix}7\\15\\15\\\vdots\\15\\15\\14\end{bmatrix}=b_2$$

- ① 第一个线性系统的系数矩阵 A_1 是严格对角占优的，而且 A_1^T 也是严格对角占优的，且 $\kappa_2(A_1)\approx 218.7$
  - 根据 **Problem 3** 的结论，不选主元的 Gauss 消去法的增长因子 $\rho=\max\limits_{\substack{1\le i,j\le n\\k=1,\dots,n-1}}\dfrac{|a_{ij}^{(k)}|}{\|A\|_\infty}\le 1$
  - **(数值线性代数, 习题 2.19)**
    若 A^T 是严格对角占优阵，则部分主元的 Gauss 消去法的增长因子 $\rho=\dfrac{\max_{ij}|a_{ij}^{(n-1)}|}{\max_{ij}|a_{ij}|}\le 2$
    (尽管这里的增长因子定义不同，但实际效果是差不多的)

  因此对于第一个线性系统，不选主元和部分主元的 Gauss 消去法都是数值稳定的.
  (我有一个疑问: 这里的不选主元和部分主元是不是等价的?)
- ② 第二个线性系统的系数矩阵 A_2 就不是严格对角占优的了 (A_2^T 同样也不是)，且 $\kappa_2(A_2)\approx 6.577\times 10^{15}$
  - 可以预见，不选主元的 Gauss 消去法中，主对角元序列满足 $\begin{cases}d_1=6\\d_k=6-\frac{8}{d_{k-1}}\quad(k\ge 2)\end{cases}$
    这个递归序列将逐渐收敛于 4 (其另一个可能的极限是 2)
    尽管其增长因子也不会太大，但这个线性系统比较病态，导致解得不准确.
  - 根据 **Problem 4** 的结论，部分主元的 Gauss 消去法的增长因子 $\rho=\max\limits_{\substack{1\le i,j\le n\\k=1,\dots,n-1}}\dfrac{|a_{ij}^{(k)}|}{\|A\|_\infty}\le 2$

  因此对于第二个线性系统，部分主元的 Gauss 消去法是数值稳定的.

Problem 6 (optional)

Provide a rounding error analysis for solving a triangular linear system.

Solution:
可以证明:

- 若下三角阵 $L=[l_{ij}]\in\mathbb{R}^{n\times n}$ 非奇异 (即满秩),
  则使用前代法求解 $Ly=b$ 得到的 $\tilde{y}$ 满足 $\begin{cases}(L+\Delta_L)\tilde{y}=b\\|\Delta_L|\le\gamma_n|L|\end{cases}$
  其中 γ_n 代表 n 层浮点运算的累积相对误差的绝对值.
- 类似地，若上三角阵 $U=[u_{ij}]\in\mathbb{R}^{n\times n}$ 非奇异 (即满秩),
  则使用回代法求解 $Ux=y$ 得到的 $\tilde{x}$ 满足 $\begin{cases}(U+\Delta_U)\tilde{x}=y\\|\Delta_U|\le\gamma_n|U|\end{cases}$
  其中 γ_n 代表 n 层浮点运算的累积相对误差的绝对值.

下面我们使用数学归纳法证明第一个命题:

对 n 使用数学归纳法.
当 $n=1$ 时，命题显然成立.
现假设命题对于所有 $n-1$ 阶下三角方程组都成立，考虑 n 阶的情形.

将 L, b 和 $\tilde{y}$ 分块为:

$$L=\begin{bmatrix} l_{11} & \\ l_1 & L_1 \end{bmatrix} \qquad \tilde{y}=\begin{bmatrix} \tilde{y}_1 \\ \tilde{x} \end{bmatrix} \qquad b=\begin{bmatrix} b_1 \\ c \end{bmatrix}$$

对于 $\tilde{y}$ 的第 1 个分块，我们有 $\tilde{y}_1=\mathrm{fl}(\frac{b_1}{l_{11}})=\frac{b_1}{l_{11}(1+\delta_1)} \quad (|\delta_1|\le \mathrm{eps})$
对于 $\tilde{y}$ 的第 2 个分块，由归纳假设我们有 $\begin{cases} (L_1+\Delta_1)\tilde{x}=\mathrm{fl}(c-\mathrm{fl}(\tilde{y}_1 l_1)) \\ |\Delta_1|\le \gamma_{n-1}|L_1| \end{cases}$

关于 $\mathrm{fl}(c-\mathrm{fl}(\tilde{y}_1 l_1))$ 我们有:

$$\begin{aligned} \mathrm{fl}(c-\mathrm{fl}(\tilde{y}_1 l_1)) &= \begin{bmatrix} (b_2-\tilde{y}_1 l_{21}(1+\delta_2))\cdot\frac{1}{1+\varepsilon_2} \\ \vdots \\ (b_n-\tilde{y}_1 l_{n1}(1+\delta_n))\cdot\frac{1}{1+\varepsilon_n} \end{bmatrix} \quad (\text{note that } c=\begin{bmatrix} b_2 \\ \vdots \\ b_n \end{bmatrix} \text{ and } l_1=\begin{bmatrix} l_{21} \\ \vdots \\ l_{n1} \end{bmatrix}) \\ &= (I+D_\varepsilon)^{-1}[c-(I+D_\delta)\tilde{y}_1 l_1] \end{aligned}$$

其中 $\begin{cases} D_\varepsilon=\mathrm{diag}(\varepsilon_2,\dots,\varepsilon_n) \\ D_\delta=\mathrm{diag}(\delta_2,\dots,\delta_n) \\ |\varepsilon_i|,|\delta_i|\le \mathrm{eps} \qquad (i=2,\dots,n) \end{cases}$

因此我们有 $(L_1+\Delta_1)\tilde{x}=\mathrm{fl}(c-\mathrm{fl}(\tilde{y}_1 l_1))=(I+D_\varepsilon)^{-1}[c-(I+D_\delta)\tilde{y}_1 l_1]$
于是 $c=(I+D_\varepsilon)(L_1+\Delta_1)\tilde{x}+(I+D_\delta)\tilde{y}_1 l_1$

联立 $\begin{cases} l_{11}(1+\delta_1)\cdot\tilde{y}_1=b_1 \\ (I+D_\delta)l_1\cdot\tilde{y}_1+(I+D_\varepsilon)(L_1+\Delta_1)\cdot\tilde{x}=c \end{cases}$ 可知:

$$(L+\Delta_L)\tilde{y}=\begin{bmatrix} l_{11}(1+\delta_1) & \\ (I+D_\delta)l_1 & (I+D_\varepsilon)(L_1+\Delta_1) \end{bmatrix}\begin{bmatrix} \tilde{y}_1 \\ \tilde{x} \end{bmatrix}=\begin{bmatrix} b_1 \\ c \end{bmatrix}=b$$

其中我们记 $\Delta_L=\begin{bmatrix} \delta_1 l_{11} & \\ D_\delta l_1 & (I+D_\varepsilon)\Delta_1+D_\varepsilon L_1 \end{bmatrix}$
对于 Δ_L 我们有:

$$\begin{aligned} |\Delta_L| &= \begin{bmatrix} |\delta_1 l_{11}| & \\ |D_\delta l_1| & |(I+D_\varepsilon)\Delta_1+D_\varepsilon L_1| \end{bmatrix} \\ &\le \begin{bmatrix} |\delta_1||l_{11}| & \\ |D_\delta||l_1| & |\Delta_1|+|D_\varepsilon|(|\Delta_1|+|L_1|) \end{bmatrix} \\ &\le \begin{bmatrix} \mathrm{eps}|l_{11}| & \\ \mathrm{eps}|I||l_1| & \gamma_{n-1}|L_1|+\mathrm{eps}|I|(\gamma_{n-1}|L_1|+|L_1|) \end{bmatrix} \\ &= \begin{bmatrix} \mathrm{eps}|l_{11}| & \\ \mathrm{eps}|l_1| & [\gamma_{n-1}+\mathrm{eps}(\gamma_{n-1}+1)]|L_1| \end{bmatrix} \quad (\text{note that } \gamma_{n-1}+\mathrm{eps}(\gamma_{n-1}+1)\approx(1+\mathrm{eps})(1+\gamma_{n-1})=\gamma_n) \\ &\approx \begin{bmatrix} \mathrm{eps}|l_{11}| & \\ \mathrm{eps}|l_1| & \gamma_n|L_1| \end{bmatrix} \\ &\le \gamma_n|L| \end{aligned}$$

这样我们就有 $\begin{cases} (L+\Delta_L)\tilde{y}=b \\ |\Delta_L|\le\gamma_n|L| \end{cases}$，命题得证.

Problem 7 (optional)

A Note on Rounding-Error Analysis of Cholesky Factorization
Provide a rounding error analysis for solving a symmetric positive-definite linear system through Cholesky factorization. In addtion to the standard Wilkinson's rounding model, you may assume that the square root of a positive real number can be computed accurately to the machine precision, i.e.,

$$\mathrm{fl}(\sqrt{\alpha})=\sqrt{\alpha}(1+\theta) \text{ where } |\theta|\le\mathrm{eps}$$

Solution: (存疑)

关于 Cholesky 分解，一个简单实用的方法是逐元素比较 $A = LL^T$ 来计算 L.

设 $L = \begin{bmatrix} l_{11} & & & \\ l_{21} & l_{22} & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{bmatrix}$

比较 $A = LL^T$ 两边对应元素，得到 $a_{ij} = \sum\limits_{p=1}^{\min(i,j)} l_{ip}l_{jp}\ (1 \le i, j \le n)$

- 首先由 $a_{11} = l_{11}^2$ 得到 $l_{11} = \sqrt{a_{11}}$
  再由 $a_{i1} = l_{11}l_{i1}\ (1 \le i \le n)$ 得到 $l_{i1} = \frac{1}{l_{11}}a_{i1}\ (1 \le i \le n)$
  这样便得到矩阵 L 的第 1 列元素.
- 假设已经计算出 L 的前 $k-1$ 列元素.
  由 $a_{kk} = \sum\limits_{p=1}^{k} l_{kp}^2$ 得到 $l_{kk} = (a_{kk} - \sum\limits_{p=1}^{k-1} l_{kp}^2)^{\frac{1}{2}}$
  再由 $a_{ik} = \sum\limits_{p=1}^{k} l_{ip}l_{kp} = \sum\limits_{p=1}^{k-1} l_{ip}l_{kp} + l_{ik}l_{kk}\ \ (i = k+1, \dots, n)$
  得到 $l_{ik} = \frac{1}{l_{kk}}(a_{ik} - \sum\limits_{p=1}^{k-1} l_{ip}l_{kp})\ \ (i = k+1, \dots, n)$
  这样便得到矩阵 L 的第 k 列元素.

上述次序可以调整为按行计算.

由于 A 的元素 a_{ij} 被用来计算 l_{ij} 后就不再使用，故我们可将 L 的元素存储在 A 的对应位置上.

(平方根法, 数值线性代数, 算法 $1.3.1$)

$$
\begin{aligned}
&\text{for } k = 1 : n \\
&\quad A(k,k) = \sqrt{A(k,k)} \\
&\quad A(k+1:n,k) = A(k+1:n,k)/A(k,k) \\
&\quad \text{for } j = k+1 : n \\
&\qquad A(j:n,j) = A(j:n,j) - A(j:n,k)A(j,k) \\
&\quad \text{end} \\
&\text{end}
\end{aligned}
$$

可以证明得到的下三角阵 L 满足 $\begin{cases} LL^T = A + \Delta_{\text{Cholesky}} \\ |\Delta_{\text{Cholesky}}| \le \gamma_n |L||L^T| \end{cases}$

当 $n = 1$ 时，命题显然成立.

现假设命题对于所有 $n-1$ 阶对称正定阵都成立，考虑 n 阶的情形.

将 L 和 A 分块为:

$$
L = \begin{bmatrix} L_1 & \\ l_2^T & l_{22} \end{bmatrix} \qquad A = \begin{bmatrix} A_{11} & a_2 \\ a_2^T & a_{22} \end{bmatrix}
$$

$$
LL^T = \begin{bmatrix} L_1 & \\ l_2^T & l_{22} \end{bmatrix} \begin{bmatrix} L_1^T & l_2 \\ & l_{22} \end{bmatrix} = \begin{bmatrix} L_1L_1^T & L_1l_2 \\ l_2^TL_1^T & l_2^Tl_2 + l_{22}l_{22} \end{bmatrix} = A + \Delta_{\text{Cholesky}}
$$

- 对于 $(1,1)$ 位置的分块，根据归纳假设我们有: $\begin{cases} L_1L_1^T = A_{11} + \Delta_1 \\ |\Delta_1| \le \gamma_{n-1}|L_1||L_1^T| \end{cases}$
- 对于 $(1,2)$ 位置的分块，根据**前代法下三角方程组的舍入误差分析**可知: $\begin{cases} (L_1 + \Delta_2)l_2 = a_2 \\ |\Delta_2| \le \gamma_{n-1}|L_1| \end{cases}$
- 对于 $(2,2)$ 位置的分块，我们有 $l_{22} = \text{fl}(\sqrt{\text{fl}(a_{22} - \text{fl}(l_2^Tl_2))})$
  注意到对于 $x, y \in \mathbb{R}^n$ 的内积运算，我们有 $|\text{fl}(x^Ty) - x^Ty| \le \gamma_n |x|^T|y|$
  因此我们有: $\begin{cases} \text{fl}(l_2^Tl_2) = l_2^Tl_2(1+\delta_1) \\ |\delta_1| \le \gamma_{n-1} \end{cases}$
  考虑到减法和开方运算都只产生机器精度级别的相对误差，故我们有:

$$
\begin{aligned}
l_{22} &= \sqrt{a_{22} - \mathrm{fl}(l_2^T l_2)}\frac{1}{1+\delta_2} \quad (\text{where } |\delta_2| \le \gamma_2) \\
&= \sqrt{a_{22} - l_2^T l_2(1+\delta_1)}\frac{1}{1+\delta_2} \quad (\text{where } |\delta_1| \le \gamma_{n-1})
\end{aligned}
$$

联立 $\begin{cases} L_1 L_1^T = A_{11} + \Delta_1 \\ (L_1 + \Delta_2) l_2 = a_2 \\ l_{22}(1+\delta_2) = \sqrt{a_{22} - l_2^T l_2(1+\delta_1)} \end{cases}$ 可知:

$$
\begin{aligned}
A + \Delta_{\text{Cholesky}} &= LL^T \\
&= \begin{bmatrix} L_1 L_1^T & L_1 l_2 \\ l_2^T L_1^T & l_2^T l_2 + l_{22} l_{22} \end{bmatrix} \\
&= \begin{bmatrix} A_{11} + \Delta_1 & a_2 - \Delta_2 l_2 \\ (a_2 - \Delta_2 l_2)^T & a_{22} - l_2^T l_2 \delta_1 - l_{22}^2[(1+\delta_2)^2 - 1] \end{bmatrix} \\
\hline
\Delta_{\text{Cholesky}} &= \begin{bmatrix} \Delta_1 & -\Delta_2 l_2 \\ -(\Delta_2 l_2)^T & -l_2^T l_2 \delta_1 - l_{22}^2[(1+\delta_2)^2 - 1] \end{bmatrix}
\end{aligned}
$$

因此我们有:

$$
\begin{aligned}
|\Delta_{\text{Cholesky}}| &= \begin{bmatrix} |\Delta_1| & |\Delta_2 l_2| \\ |(\Delta_2 l_2)^T| & |l_2^T l_2 \delta_1 + l_{22}^2[(1+\delta_2)^2 - 1]| \end{bmatrix} \\
&\le \begin{bmatrix} |\Delta_1| & |\Delta_2||l_2| \\ |l_2|^T |\Delta_2|^T & \delta_1 |l_2^T||l_2| + [(1+\delta_2)^2 - 1]|l_{22}|^2 \end{bmatrix} \\
&\le \begin{bmatrix} \gamma_{n-1}|L_1||L_1^T| & \gamma_{n-1}|L_1||l_2| \\ \gamma_{n-1}|l_2^T||L_1| & \gamma_{n-1}|l_2^T||l_2| + [(1+\gamma_2)^2 - 1]|l_{22}|^2 \end{bmatrix} \\
&\le \gamma_n \begin{bmatrix} |L_1||L_1^T| & |L_1||l_2| \\ |l_2^T||L_1^T| & |l_2^T||l_2| + |l_{22}|^2 \end{bmatrix} \\
&= \gamma_n \begin{bmatrix} |L_1| & \\ |l_2^T| & |l_{22}| \end{bmatrix} \begin{bmatrix} |L_1^T| & |l_2| \\ & |l_{22}| \end{bmatrix} \\
&= \gamma_n |L||L^T|
\end{aligned}
$$

这样我们就证明了 Cholesky 分解得到的下三角阵 L 满足 $\begin{cases} LL^T = A + \Delta_{\text{Cholesky}} \\ |\Delta_{\text{Cholesky}}| \le \gamma_n |L||L^T| \end{cases}$

(值得注意的是，如果我们归纳的是 $\Delta_{\text{cholesky}} \le \gamma_{n-1}|L||L^T|$，那 $n = 1$ 的情况就不满足了)

当对 A 完成 Cholesky 分解 $\begin{cases} LL^T = A + \Delta_{\text{Cholesky}} \\ |\Delta_{\text{Cholesky}}| \le \gamma_n |L||L^T| \end{cases}$ 后，求解对称正定线性方程组 $Ax = b$ 的问题就归结为:

- 用前代法求解 $Ly = b$ 得到 y
- 用回代法求解 $L^T x = y$ 得到 x

根据**前代法下三角方程组的舍入误差分析** (Problem 6) 可知最后的解 x 满足

$$
\begin{cases} (L + \Delta_{\text{forward}})(L^T + \Delta_{\text{backward}})x = b \\ |\Delta_{\text{forward}}| \le \gamma_n |L| \\ |\Delta_{\text{backward}}| \le \gamma_n |L^T| \end{cases}
$$

联立 $\begin{cases} LL^T = A + \Delta_{\text{Cholesky}} \\ (L + \Delta_{\text{forward}})(L^T + \Delta_{\text{backward}})x = b \end{cases}$ 就得到:

$$
(A + \Delta_{\text{Cholesky}} + L\Delta_{\text{backward}} + \Delta_{\text{forward}} L^T + \Delta_{\text{forward}}\Delta_{\text{backward}})x = b
$$

$$
\Leftrightarrow
$$

$$
(A + \Delta_A)\tilde{x} = b \text{ where } \Delta_A = \Delta_{\text{Cholesky}} + L\Delta_{\text{backward}} + \Delta_{\text{forward}} L^T + \Delta_{\text{forward}}\Delta_{\text{backward}}
$$

关于 Δ_A 我们有:

$$
\begin{aligned}
|\Delta_A| &= |\Delta_{\text{Cholesky}} + L\Delta_{\text{backward}} + \Delta_{\text{forward}}L^T + \Delta_{\text{forward}}\Delta_{\text{backward}}| \\
&\le |\Delta_{\text{Cholesky}}| + |L||\Delta_{\text{backward}}| + |\Delta_{\text{forward}}||L^T| + |\Delta_{\text{forward}}||\Delta_{\text{backward}}| \\
&\le \gamma_n|L||L^T| + |L|\cdot\gamma_n|L^T| + \gamma_n|L|\cdot|L^T| + \gamma_n|L|\cdot\gamma_n|L^T| \\
&= (3\gamma_n + \gamma_n^2)|L||L^T| \\
&\le (3\gamma_n + 3\gamma_n^2 + \gamma_n^3)|L||L^T| \\
&= [(1+\gamma_n)^3 - 1]|L||L^T| \\
&= [(1+\gamma_{3n}) - 1]|L||L^T| \\
&= \gamma_{3n}|L||L^T|
\end{aligned}
$$

因此利用 Cholesky 分解求解对称正定线性方程组 $Ax = b$ 的舍入误差为: $\begin{cases}(A+\Delta A)x = b \\ |\Delta A| \le \gamma_{3n}|L||L^T|\end{cases}$

其中 L 是 Cholesky 分解 $A = LL^T$ 的计算解.