1. (a) Let λ be an eigenvalue of A. Show that $|\lambda - a_{ii}| \le \sum_{j \ne i} |a_{ij}|$ for some index i. (This is Gershgorin's theorem).

Proof. Let $Ax = \lambda x$, $x \ne 0$, pick index k s.t. $|x_k| = \max_i |x_i|$.
Consider the k-th row of the $Ax = \lambda x$:

$$\sum_{j=1}^{n} a_{kj} x_j = \lambda x_k \implies (\lambda - a_{kk}) x_k = \sum_{j \ne k} a_{kj} x_j$$

Taking absolute values and using the triangle inequality, we have:

$$|\lambda - a_{kk}||x_k| = \left| \sum_{j \ne k} a_{kj} x_j \right| \le \sum_{j \ne k} |a_{kj}||x_j| \le \sum_{j \ne k} |a_{kj}||x_k|$$

$$|\lambda - a_{kk}| \le \sum_{j \ne k} |a_{kj}|$$

□

(b) Show that if A is strictly diagonally dominant (that is, $|a_{ii}| > \sum_{j \ne i} |a_{ij}|$ for all i), then A is invertible.

Proof. If $Ax = 0, x \ne 0$, pick index k s.t. $|x_k| = \max_i |x_i|$. Then,

$$|a_{kk}||x_k| \le \sum_{j \ne k} |a_{kj}||x_j| \le \Big(\sum_{j \ne k} |a_{kj}| \Big) |x_k|$$

$$\implies |a_{kk}| \le \sum_{j \ne k} |a_{kj}|$$

contradicting the strictly diagonally dominant condition. Thus, $Ax = 0 \Rightarrow x = 0$, so A is invertible. □

2. Consider the matrix

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}$$

(a) What does Gershgorin's theorem say about the e-values of A ?
For the first row, we have $|\lambda - 2| \le 1$, i.e. $\lambda \in [1, 3]$.
For the second row, we have $|\lambda - 2| \le 2$, i.e. $\lambda \in [0, 4]$.
So, the union of all Gershgorin discs is $[0, 4]$. Thus, all eigenvalues of A lie in $[0, 4]$.

(b) Let $\boldsymbol{e}_i$ be the standard unit vector in the i th direction. Calculate the Rayleigh quotients of $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_1+\boldsymbol{e}_3, \boldsymbol{e}_1-\boldsymbol{e}_2+\boldsymbol{e}_3-\boldsymbol{e}_4, \boldsymbol{e}_1+\boldsymbol{e}_2+\boldsymbol{e}_3+\boldsymbol{e}_4$. Use this information to estimate λ_1 and λ_4, the smallest and the largest e-values of A.

$$\begin{aligned}
R_A(\boldsymbol{e}_1) &= \frac{\boldsymbol{e}_1^T A \boldsymbol{e}_1}{\boldsymbol{e}_1^T \boldsymbol{e}_1} = 2 \\
R_A(\boldsymbol{e}_2) &= \frac{\boldsymbol{e}_2^T A \boldsymbol{e}_2}{\boldsymbol{e}_2^T \boldsymbol{e}_2} = 2 \\
R_A(\boldsymbol{e}_1+\boldsymbol{e}_3) &= \frac{(\boldsymbol{e}_1+\boldsymbol{e}_3)^T A (\boldsymbol{e}_1+\boldsymbol{e}_3)}{(\boldsymbol{e}_1+\boldsymbol{e}_3)^T (\boldsymbol{e}_1+\boldsymbol{e}_3)} = \frac{4}{2} = 2 \\
R_A(\boldsymbol{e}_1-\boldsymbol{e}_2+\boldsymbol{e}_3-\boldsymbol{e}_4) &= \frac{(\boldsymbol{e}_1-\boldsymbol{e}_2+\boldsymbol{e}_3-\boldsymbol{e}_4)^T A (\boldsymbol{e}_1-\boldsymbol{e}_2+\boldsymbol{e}_3-\boldsymbol{e}_4)}{(\boldsymbol{e}_1-\boldsymbol{e}_2+\boldsymbol{e}_3-\boldsymbol{e}_4)^T (\boldsymbol{e}_1-\boldsymbol{e}_2+\boldsymbol{e}_3-\boldsymbol{e}_4)} = \frac{14}{4} = 3.5 \\
R_A(\boldsymbol{e}_1+\boldsymbol{e}_2+\boldsymbol{e}_3+\boldsymbol{e}_4) &= \frac{(\boldsymbol{e}_1+\boldsymbol{e}_2+\boldsymbol{e}_3+\boldsymbol{e}_4)^T A (\boldsymbol{e}_1+\boldsymbol{e}_2+\boldsymbol{e}_3+\boldsymbol{e}_4)}{(\boldsymbol{e}_1+\boldsymbol{e}_2+\boldsymbol{e}_3+\boldsymbol{e}_4)^T (\boldsymbol{e}_1+\boldsymbol{e}_2+\boldsymbol{e}_3+\boldsymbol{e}_4)} = \frac{2}{4} = 0.5
\end{aligned}$$

Thus, we estimate

$$0 < \lambda_1 \le 0.5, \quad 3.5 \le \lambda_4 \le 4$$

.

(c) Find $\|A\|_1, \|A\|_2, \|A\|_\infty, \|A\|_E, \rho(A), \operatorname{tr}(A), \kappa_2(A)$.

$$\begin{aligned}
\|A\|_1 &= \max_{1\le j\le 4} \sum_{i=1}^{4} |a_{ij}| = 4 \\
\|A\|_\infty &= \max_{1\le i\le 4} \sum_{j=1}^{4} |a_{ij}| = 4 \\
\|A\|_E &= \sqrt{\sum_{i,j=1}^{4} |a_{ij}|^2} = \sqrt{22} \\
\operatorname{tr}(A) &= \sum_{i=1}^{4} a_{ii} = 8 \\
\rho(A) &= \max_{1\le i\le 4} |\lambda_i| = \frac{5+\sqrt{5}}{2} \\
\|A\|_2 &= \sqrt{\rho\left(A^T A\right)} = \frac{5+\sqrt{5}}{2} \\
\kappa_2(A) &= \|A\|_2 \left\|A^{-1}\right\|_2 = \frac{\frac{5+\sqrt{5}}{2}}{\frac{3-\sqrt{5}}{2}} = 5+2\sqrt{5}
\end{aligned}$$

3. (a) Suppose $A\boldsymbol{x} = \boldsymbol{b}$. Let $\widetilde{\boldsymbol{x}}$ be an approximation of the exact solution $\boldsymbol{x}$. The error is defined by $\boldsymbol{e} = \boldsymbol{x} - \widetilde{\boldsymbol{x}}$, and the residual is $\boldsymbol{r} = \boldsymbol{b} - A\widetilde{\boldsymbol{x}}$. Show that $A\boldsymbol{e} = \boldsymbol{r}$ and $\frac{\|\boldsymbol{e}\|}{\|\boldsymbol{x}\|} \le \kappa(A)\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{b}\|}$.

Proof.

$$A\tilde{\boldsymbol{x}} = \boldsymbol{b} - \boldsymbol{r}, \qquad A\boldsymbol{x} = \boldsymbol{b} \;\Rightarrow\; A(\boldsymbol{x} - \tilde{\boldsymbol{x}}) = \boldsymbol{r} \;\Rightarrow\; A\boldsymbol{e} = \boldsymbol{r}.$$

$$\|\boldsymbol{e}\| = \|A^{-1}\boldsymbol{r}\| \le \|A^{-1}\|\,\|\boldsymbol{r}\|, \quad \|\boldsymbol{b}\| = \|A\boldsymbol{x}\| \le \|A\|\,\|\boldsymbol{x}\| \Rightarrow \frac{\|\boldsymbol{e}\|}{\|\boldsymbol{x}\|} \le \|A\|\,\|A^{-1}\|\,\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{b}\|} = \kappa(A)\frac{\|\boldsymbol{r}\|}{\|\boldsymbol{b}\|}.$$

□

(b) It follows that if A is invertible, then $\boldsymbol{e} = \boldsymbol{0}$ if and only if $\boldsymbol{r} = \boldsymbol{0}$. However, if A is ill-conditioned, the error $\boldsymbol{e}$ may be large even though the residual $\boldsymbol{r}$ is small. This occurs in the example below. Show that $A\boldsymbol{x} = \boldsymbol{b}$. Find $\|\boldsymbol{e}\|_\infty$ and $\|\boldsymbol{r}\|_\infty$ for the vectors $\widetilde{\boldsymbol{x}}_1, \widetilde{\boldsymbol{x}}_2$. Find $\kappa_\infty(A)$. (Use exact arithmetics). $A = \begin{pmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{pmatrix}$, $\boldsymbol{b} = \binom{0.8642}{0.1440}$, $\boldsymbol{x} = \binom{2}{-2}$, $\widetilde{\boldsymbol{x}}_1 = \binom{0}{1}$, $\widetilde{\boldsymbol{x}}_2 = \binom{0.9911}{-0.4870}$

Solution:

$$A = \begin{pmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} 0.8642 \\ 0.1440 \end{pmatrix}, \quad \boldsymbol{x} = \begin{pmatrix} 2 \\ -2 \end{pmatrix}.$$

$$A\boldsymbol{x} = \begin{pmatrix} 1.2969 \cdot 2 + 0.8648(-2) \\ 0.2161 \cdot 2 + 0.1441(-2) \end{pmatrix} = \begin{pmatrix} 0.8642 \\ 0.1440 \end{pmatrix} = \boldsymbol{b}.$$

For $\tilde{\boldsymbol{x}}_1 = \binom{0}{1}$:

$$\boldsymbol{e}_1 = \begin{pmatrix} 2 \\ -3 \end{pmatrix}, \quad \|\boldsymbol{e}_1\|_\infty = 3, \qquad A\tilde{\boldsymbol{x}}_1 = \begin{pmatrix} 0.8648 \\ 0.1441 \end{pmatrix}, \quad \boldsymbol{r}_1 = \begin{pmatrix} -0.0006 \\ -0.0001 \end{pmatrix}, \quad \|\boldsymbol{r}_1\|_\infty = 6 \times 10^{-4}.$$

For $\tilde{\boldsymbol{x}}_2 = \binom{0.9911}{-0.4870}$:

$$\boldsymbol{e}_2 = \begin{pmatrix} 1.0089 \\ -1.5130 \end{pmatrix}, \quad \|\boldsymbol{e}_2\|_\infty = 1.5130,$$

$$A\tilde{\boldsymbol{x}}_2 = \begin{pmatrix} 0.86419999 \\ 0.14400001 \end{pmatrix}, \quad \boldsymbol{r}_2 = \begin{pmatrix} 10^{-8} \\ -10^{-8} \end{pmatrix}, \quad \|\boldsymbol{r}_2\|_\infty = 10^{-8}.$$

Condition number $\kappa_\infty(A)$.

$$\|A\|_\infty = \max\{1.2969 + 0.8648,\ 0.2161 + 0.1441\} = 2.1617,$$

$$\det(A) = 1.2969 \cdot 0.1441 - 0.8648 \cdot 0.2161 = 10^{-8},$$

$$A^{-1} = \frac{1}{10^{-8}} \begin{pmatrix} 0.1441 & -0.8648 \\ -0.2161 & 1.2969 \end{pmatrix}, \quad \|A^{-1}\|_\infty = \max\{1.0089, 1.5130\} \times 10^8 = 1.5130 \times 10^8,$$

$$\kappa_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty = 2.1617 \cdot 1.5130 \times 10^8 = 3.27065210 \times 10^8.$$

4. Compute the solution of the 2×2 system:

$$\begin{aligned} 10^{-3}x + y &= 5 \\ x - y &= 6 \end{aligned}$$

using standard Gaussian elimination and Gaussian elimination with pivoting. Conduct all computations with two significant digits (decimal). Compare and explain the results.

(A) Standard Gaussian elimination (no pivoting).

Start with (to two sig. figs.)

$$\begin{bmatrix} 1.0 \times 10^{-3} & 1.0 & \mid & 5.0 \end{bmatrix}, \qquad \begin{bmatrix} 1.0 & -1.0 & \mid & 6.0 \end{bmatrix}.$$

Eliminate a_{21} using multiplier $m = \dfrac{1.0}{1.0 \times 10^{-3}} = 1.0 \times 10^3$.

$$(1.0 \times 10^3) \cdot \text{row}_1 = \begin{bmatrix} 1.0 & 1.0 \times 10^3 & \mid & 5.0 \times 10^3 \end{bmatrix}.$$

Row operation (to two sig. figs.):

$$\text{row}_2 \leftarrow \text{row}_2 - (1.0 \times 10^3) \cdot \text{row}_1 \Rightarrow \begin{bmatrix} 0.0 & -1.0 \times 10^3 & \mid & -5.0 \times 10^3 \end{bmatrix}.$$

Thus $-1.0 \times 10^3\, y = -5.0 \times 10^3 \Rightarrow y = 5.0$. Back–substitute into the first equation:

$$1.0 \times 10^{-3} x + 5.0 = 5.0 \ \Rightarrow\ 1.0 \times 10^{-3} x = 0.0 \ \Rightarrow\ x = 0.0.$$

$$\boxed{(x, y)_{\text{no pivot}} = (0.0,\ 5.0)}.$$

(B) Gaussian elimination with partial pivoting.

Swap the rows so the first pivot is 1.0:

$$\begin{bmatrix} 1.0 & -1.0 & \mid & 6.0 \end{bmatrix}, \qquad \begin{bmatrix} 1.0 \times 10^{-3} & 1.0 & \mid & 5.0 \end{bmatrix}.$$

Eliminate a_{21} with $m = \dfrac{1.0 \times 10^{-3}}{1.0} = 1.0 \times 10^{-3}$:

$$(1.0 \times 10^{-3}) \cdot \text{row}_1 = \begin{bmatrix} 1.0 \times 10^{-3} & -1.0 \times 10^{-3} & \mid & 6.0 \times 10^{-3} \end{bmatrix}.$$

Row operation (rounded to two sig. figs.):

$$\text{row}_2 \leftarrow \text{row}_2 - (1.0 \times 10^{-3}) \cdot \text{row}_1 \Rightarrow \begin{bmatrix} 0.0 & 1.0 & \mid & 5.0 \end{bmatrix},$$

so $y = 5.0$. Back–substitute in the first (pivoted) row:

$$x - y = 6.0 \ \Rightarrow\ x = 11.0.$$

$$\boxed{(x, y)_{\text{pivot}} = (11.0,\ 5.0)}.$$

Comparison and explanation. The exact solution is $(x, y) = (10.989\ldots, 4.989\ldots)$. Without pivoting the tiny pivot 10^{-3} produces a huge multiplier 10^3, forcing subtraction of nearly equal large numbers and catastrophic cancellation, yielding $x = 0.0$. Pivoting selects the well–scaled pivot 1.0, avoids amplification of rounding, and returns the correct two–digit result (11.0, 5.0).

5. (a) Show that for any matrix A, there exists a permutation matrix P such that $PA = LU$, where L is a unit lower triangular matrix and U is an upper triangular matrix.

   Hint: P is a composition, that is, a product of all permutation matrices used at every stage of the Gaussian elimination algorithm.

Proof. Let Gaussian elimination with (partial) pivoting act on A. At step k a permutation P_k swaps rows to place a nonzero pivot and a unit lower–triangular elimination matrix E_k zeroes entries below it. After m steps,

$$P_m \cdots P_1 A = LU, \qquad E_m \cdots E_1 P_m \cdots P_1 A = U.$$

Set $P = P_m \cdots P_1$ and $L = (E_m \cdots E_1)^{-1}$ (unit lower triangular). Then $PA = LU$. $\square$

(b) For the following matrix A, find P, L and U. Use the LU factorization to solve $A\boldsymbol{x} = \boldsymbol{b}$.

$$A = \begin{pmatrix} 0 & 2 & -1 \\ 1 & 1 & 1 \\ 2 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix}.$$

For $A = \begin{pmatrix} 0 & 2 & -1 \\ 1 & 1 & 1 \\ 2 & 0 & 1 \end{pmatrix}$, $\boldsymbol{b} = \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix}$, partial pivoting on column 1 then column 2 gives

$$P = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad L = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & -1 \\ 0 & 0 & 1 \end{pmatrix}, \quad \text{so } PA = LU.$$

Solve $A\boldsymbol{x} = \boldsymbol{b}$ via $LU\boldsymbol{x} = P\boldsymbol{b}$:

$$P\boldsymbol{b} = \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix}, \quad L\boldsymbol{y} = P\boldsymbol{b} \Rightarrow \boldsymbol{y} = \begin{pmatrix} 0 \\ -2 \\ 2 \end{pmatrix}, \quad U\boldsymbol{x} = \boldsymbol{y} \Rightarrow \boldsymbol{x} = \begin{pmatrix} -1 \\ 0 \\ 2 \end{pmatrix}.$$

6. Let M be an invertible matrix. Show that $\|\boldsymbol{x}\|_M := \|M\boldsymbol{x}\|_\infty$ defines a vector norm for which the subordinate matrix norm is $\|A\|_M := \|MAM^{-1}\|_\infty$.

Proof. **1) $\|\cdot\|_M$ is a norm.**

- (Positivity) $\|x\|_M = \|Mx\|_\infty \geq 0$, and $\|x\|_M = 0 \iff Mx = 0 \iff x = 0$ (since M is invertible).
- (Absolute homogeneity) $\|\alpha x\|_M = \|M(\alpha x)\|_\infty = |\alpha|\,\|Mx\|_\infty = |\alpha|\,\|x\|_M$.
- (Triangle inequality) $\|x+y\|_M = \|M(x+y)\|_\infty \leq \|Mx\|_\infty + \|My\|_\infty = \|x\|_M + \|y\|_M$, using the triangle inequality of $\|\cdot\|_\infty$.

2) Subordinate matrix norm. Define the operator norm subordinate to $\|\cdot\|_M$ by

$$\|A\|_M := \sup_{x\neq 0} \frac{\|Ax\|_M}{\|x\|_M} = \sup_{x\neq 0} \frac{\|MAx\|_\infty}{\|Mx\|_\infty}.$$

Let $y = Mx$ (bijection since M is invertible). Then

$$\|A\|_M = \sup_{y\neq 0} \frac{\|MAM^{-1}y\|_\infty}{\|y\|_\infty} = \|MAM^{-1}\|_\infty.$$

Hence the subordinate matrix norm to $\|\cdot\|_M$ is exactly $\|A\|_M := \|MAM^{-1}\|_\infty$. $\square$

7. Consider the 2 -point boundary value problem

$$-\phi''(x) + x^2\phi(x) = \left(1 + 4x + 2x^2 - x^4\right) e^x, \quad \phi(0) = 1, \phi(1) = 0.$$

Check that the solution is $\phi(x) = (1 - x^2)\, e^x$. Write a computer program to solve this problem using the second-order finite-difference scheme discussed in class. Solve the tridiagonal system using the LU factorization. Run the code with mesh size $h = 1/2^p$ for $p = 1, \ldots, 14$. Output the results in the following format,

column1: h
column2: $\|u_h - \phi_h\|_\infty$
column3: $\|u_h - \phi_h\|_\infty / h^2$

Describe and explain what happens to the error for small values of h.

Model problem and exact solution.

$$-\phi''(x) + x^2\phi(x) = (1 + 4x + 2x^2 - x^4)e^x, \quad \phi(0) = 1,\ \phi(1) = 0, \quad \phi(x) = (1 - x^2)e^x.$$

Second-order FD scheme (interior nodes $x_i = ih$, $i = 1, \ldots, n$).

$$-\frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} + x_i^2 u_i = f(x_i), \qquad u_0 = 1,\ u_{n+1} = 0,$$

$$A = \mathrm{tridiag}\Big(-\tfrac{1}{h^2},\ \tfrac{2}{h^2} + x_i^2,\ -\tfrac{1}{h^2}\Big), \quad b_i = f(x_i) + \tfrac{1}{h^2}\mathbf{1}_{\{i=1\}}.$$

Solve by LU. Compute $[L, U, P] = \mathrm{lu}(A)$, then $U\, u = L^{-1} P\, b$. Form $u_h = [1;\ u;\ 0]$ and the exact vector $\phi_h = \phi(x)$. Report:

h	$\|u_h - \phi_h\|_\infty$	$\|u_h - \phi_h\|_\infty / h^2$
5.000000e−01	6.472527e−02	2.589011e−01
2.500000e−01	1.651983e−02	2.643173e−01
1.250000e−01	4.154789e−03	2.659065e−01
6.250000e−02	1.056844e−03	2.705520e−01
3.125000e−02	2.643184e−04	2.706620e−01
1.562500e−02	6.608632e−05	2.706896e−01
7.812500e−03	1.652200e−05	2.706965e−01
3.906250e−03	4.130527e−06	2.706982e−01
1.953125e−03	1.032633e−06	2.706986e−01
9.765625e−04	2.581585e−07	2.706988e−01
4.882812e−04	6.454092e−08	2.707042e−01
2.441406e−04	1.613556e−08	2.707097e−01
1.220703e−04	3.081321e−09	2.067840e−01
6.103516e−05	2.814358e−10	7.554734e−02

Behavior for small h: For h down to 2.44×10^{-4}, the ratio $\|u_h - \phi_h\|_\infty / h^2$ stays essentially constant at $\approx 2.707 \times 10^{-1}$, confirming the expected second–order convergence:

$$\|u_h - \phi_h\|_\infty \approx \alpha\, h^2, \qquad \alpha \approx 2.707 \times 10^{-1}.$$

For the last two meshes, $h = 1.22 \times 10^{-4}$ and 6.10×10^{-5}, the ratio *drops* (to 2.07×10^{-1} and 7.55×10^{-2}). This is a finite–precision effect: the total error is well modeled by

$$E(h) \;\approx\; \underbrace{\alpha h^2}_{\text{truncation}} \;+\; \underbrace{\beta\, \varepsilon_{\text{mach}}\, h^{-2}}_{\text{roundoff/backward error}},$$

with $\varepsilon_{\text{mach}} \approx 2.22 \times 10^{-16}$ in double precision. When h is small, the two terms become comparable; if β has opposite sign to α, they partially cancel, producing the observed dip in $E(h)/h^2$. If you refine further, the roundoff term will eventually dominate and $E(h)$ will stagnate and then *increase* (so $E(h)/h^2$ rises), as the LU solve backward error scales like $O(\varepsilon_{\text{mach}}\|A\|) \sim O(\varepsilon_{\text{mach}}\, h^{-2})$.

Summary.

$$\begin{cases} \text{Second–order regime:} & h \gtrsim 2.4 \times 10^{-4}, \quad E(h)/h^2 \approx 0.2707, \\ \text{Roundoff interaction:} & h \lesssim 1.2 \times 10^{-4}, \quad \text{favorable cancellation lowers } E(h)/h^2, \\ \text{Prediction:} & \text{for even smaller } h,\ \text{roundoff dominates } \Rightarrow E(h)\uparrow. \end{cases}$$