HW 6

1) Let $A = \begin{bmatrix} 1 & 1 \\ -2 & 3 \end{bmatrix}$. Will the Jacobi iterative method converge on A, regardless of the starting vector $x^{(0)}$?

For convergence, we need all eigenvalues λ_i of $(I - D^{-1}A)$ to be less than 1 in absolute value, ie

$$|\lambda_i| < 1 \qquad \text{for } i \in \{1, 2, \ldots, n\}$$

We have $A = \begin{bmatrix} 1 & 1 \\ -2 & 3 \end{bmatrix}$, $D = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$

$$D^{-1} = \left[\begin{array}{cc|cc} 1 & 0 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{array}\right] \sim \left[\begin{array}{cc|cc} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1/3 \end{array}\right] = \begin{bmatrix} 1 & 0 \\ 0 & 1/3 \end{bmatrix}$$

$$I - D^{-1}A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1/3 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ -2 & 3 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ -2/3 & 1 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 2/3 & 0 \end{bmatrix}$$

$$\det(\lambda I - (I - D^{-1}A)) = \det \begin{bmatrix} \lambda & -1 \\ 2/3 & \lambda \end{bmatrix} = \lambda^2 + 2/3 = 0$$

$$\Rightarrow \lambda^2 = -2/3 \Rightarrow \lambda_1 = \sqrt{2/3}\, i \;, \; \lambda_2 = -\sqrt{2/3}\, i$$

$$|\lambda_1| = \sqrt{2/3} \quad |\lambda_2| = \sqrt{2/3} \;, \; |\lambda_1| = |\lambda_2| < 1$$

Yes, the method will converge on A regardless of $x^{(0)}$.

2) let $B = \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix}$ and $b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Take 5 steps of Gauss-Seidel to attempt to solve $Bx = b$, starting with $x^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$; write down $x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, x^{(5)}$. Based on what you see, do you think Gauss-Seidel will converge?

let $x^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. We know that

$$b_{11} x_1 + b_{12} x_2 = 1$$

$$b_{21} x_1 + b_{22} x_2 = 1$$

$$\Rightarrow x_1^{(k+1)} = \frac{1}{b_{11}} \left(1 - b_{12} x_2^{(k)}\right)$$

$$x_2^{(k+1)} = \frac{1}{b_{22}} \left(1 - b_{21} x_1^{(k+1)}\right).$$

let's plug in our intial guess for $k = 0$:

$$x_1^{(1)} = \frac{1}{b_{11}} \left(1 - b_{12} x_2^{(0)}\right)$$

$$= \frac{1}{1} \left(1 - (1)(0)\right) = 1 - 0 = 1$$

$$x_2^{(1)} = \frac{1}{b_{22}} \left(1 - b_{21} x_1^{(1)}\right)$$

$$= \frac{1}{3} \left(1 - (1)(1)\right) = \frac{1}{3}(1 - 1) = 0$$

$$\Rightarrow x^{(1)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

And repeat:

$$x_1^{(2)} = \frac{1}{b_{11}}\left(1 - b_{12}\, x_2^{(1)}\right)$$

$$= \frac{1}{1}\left(1 - (1)(0)\right) = 1(1-0) = 1$$

$$x_2^{(2)} = \frac{1}{b_{22}}\left(1 - b_{21}\, x_1^{(2)}\right)$$

$$= \frac{1}{3}\left(1 - (1)(1)\right) = \frac{1}{3}(1-1) = 0$$

$$\Rightarrow x^{(2)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$x_1^{(3)} = \frac{1}{b_{11}}\left(1 - b_{12}\, x_2^{(2)}\right)$$

$$= \frac{1}{1}\left(1 - (1)(0)\right) = 1(1-0) = 1$$

$$x_2^{(3)} = \frac{1}{b_{22}}\left(1 - b_{21}\, x_1^{(3)}\right)$$

$$= \frac{1}{3}\left(1 - (1)(1)\right) = \frac{1}{3}(1-1) = 0$$

$$\Rightarrow x^{(3)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$x_1^{(4)} = \frac{1}{b_{11}}\left(1 - b_{12}\, x_2^{(3)}\right)$$

$$= \frac{1}{1}\left(1 - (1)(0)\right) = 1$$

$$x_2^{(4)} = \frac{1}{b_{22}}\left(1 - b_{21}\, x_1^{(4)}\right)$$

$$= \frac{1}{3}\left(1 - (1)(1)\right) = \frac{1}{3}(1-1) = 0$$

$$\Rightarrow x^{(4)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$x_1^{(5)} = \frac{1}{b_{11}}\left(1 - b_{12}\, x_2^{(4)}\right)$$

$$= \frac{1}{1}\left(1 - (1)(0)\right) = 1(1-0) = 1$$

$$x_2^{(5)} = \frac{1}{b_{22}}\left(1 - b_{21}\, x_1^{(5)}\right)$$

$$= \frac{1}{3}\left(1 - (1)(1)\right) = 0$$

$$\Rightarrow x^{(5)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Yes, Gauss-Seidel will converge, since we have found the same result for all 5 iterations.

3) A matrix $A \in \mathbb{C}^{n\times n}$ is called skew-Hermitian if $A^* = -A$. Show that skew-Hermitian matrices have only purely imaginary eigenvalues, i.e., any eigenvalue of a skew-Hermitian matrix has the form $\lambda = ai$ with $a \in \mathbb{R}$.

Hint: Show that iA is Hermitian, then use what we know about Hermitian matrices.

We know that $A^* = (\bar{A})^T = -A$ (if A is skew-Hermitian), that $B^* = B$ if B is Hermitian, and for any $c \in \mathbb{C}$, $c = a+ib$, $\bar{c} = a-ib$.

Let $B = iA$. Then

$$B^* = (\bar{B})^T = (i\bar{A})^T = (\bar{i}\bar{A})^T = (-i\bar{A})^T = -i(\bar{A})^T$$

$$= -i(A^*) = -i(-A) = iA = B$$

Hence, $B = iA$ is Hermitian, so all eigenvalues of iA are real. Now,

$$Av = \lambda v$$

$$\Rightarrow iAv = i(\lambda v)$$

$$\Rightarrow (iA)v = (i\lambda)v$$

$\Rightarrow i\lambda$ is a real eigenvalue of iA

Suppose $\lambda = a$, $a \in \mathbb{R}$. Since $i\lambda$ is real, λ must have some imaginary component sum that multiplying i by λ results in a real value. Hence, $\lambda = ai$ for some $a \in \mathbb{R}$.

4) Using one of the two methods of computing the SVD using eigendecompositions outlined in class, compute by hand the reduced SVD of the matrix $A = \begin{bmatrix} -1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$.

$$A = U \Sigma V^T$$

U (3×3)
V (2×2)
Σ (3×2)

$$A^TA = V\Sigma^T\Sigma V^T \Rightarrow A^TAV = V\Sigma^T\Sigma$$

$\Rightarrow (\delta_1, v_1), (\delta_1, v_2) \ldots (\delta_r, v_r), (0, v_{r+1}) \ldots (0, v_n)$

are eigen pairs of A^TA

$$A^TA = \begin{bmatrix} -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}\begin{bmatrix} -1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

$$\det(A^TA - \lambda I) = \det\begin{bmatrix} 2-\lambda & -1 \\ -1 & 2-\lambda \end{bmatrix} = (2-\lambda)(2-\lambda) - (-1)(-1)$$
$$= \lambda^2 - 4\lambda + 3 = 0$$

$$\Rightarrow (\lambda-3)(\lambda-1) = 0 \quad \Rightarrow \lambda_1 = 3,\ \lambda_2 = 1$$

$$\Rightarrow \sigma_1 = \sqrt{3},\ \sigma_2 = 1 \quad \Rightarrow \Sigma = \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \end{bmatrix}$$

$$(A^TA - 3I)x = 0$$

$$\Rightarrow \begin{bmatrix} 2-3 & -1 \\ -1 & 2-3 \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ -1 & -1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Rightarrow -x_1 - x_2 = 0 \quad \Rightarrow x_2 = -x_1 \quad \Rightarrow x_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$\|x_1\|_2 = \sqrt{1^2 + (-1)^2} = \sqrt{2} \quad \Rightarrow v_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$(A^TA - I)x = 0$$

$$\Rightarrow \begin{bmatrix} 2-1 & -1 \\ -1 & 2-1 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Rightarrow \begin{matrix} x_1 - x_2 = 0 \\ -x_1 + x_2 = 0 \end{matrix} \quad \begin{matrix} \Rightarrow x_1 = x_2 \\ \Rightarrow x_1 = x_2 \end{matrix} \quad \Rightarrow x_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow v_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\Rightarrow V = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

we now have $\Sigma = \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \end{bmatrix}$, $V = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$

$$A = U\Sigma V^T$$

$$\Rightarrow AV = U\Sigma \Rightarrow U = AV\Sigma^{-1}$$

$$\Sigma^{-1} = \left[\begin{array}{cc|cc} \sqrt{3} & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{array}\right] \sim \left[\begin{array}{cc|cc} 1 & 0 & 1/\sqrt{3} & 0 \\ 0 & 1 & 0 & 1 \end{array}\right] = \begin{bmatrix} \sqrt{3}/3 & 0 \\ 0 & 1 \end{bmatrix}$$

$$U = AV\Sigma^{-1} = \begin{bmatrix} -1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{3}/3 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -\sqrt{6}/3 & 0 \\ \sqrt{6}/6 & \sqrt{2}/2 \\ -\sqrt{6}/6 & \sqrt{2}/2 \end{bmatrix}$$

$\Rightarrow$ The reduced SVD of A is

$$A = \begin{bmatrix} -\sqrt{6}/3 & 0 \\ \sqrt{6}/6 & \sqrt{2}/2 \\ -\sqrt{6}/6 & \sqrt{2}/2 \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix}$$

$$U \qquad \Sigma \qquad V^T$$

5) Run the sample_inverse_PM.m code provided. Below is an explanation of what the code does.

The first line resets the pseudo-random number generator to default

Lines 2-4 construct a random matrix A w/ prescribed eigenvalues (given in line 2)

Line 5 constructs a matrix B which shifts the eigenvalues of A by the same amount, while leaving the eigenvectors unchanged

Line 6 inverts B and names the inverse C.

The remainder of the code applies the power method to the matrix C, starting from the vector of all ones, using 8 iteration

Run the code and examine it closely, then answer the following.

(a) What are the eigenvalues of B and C? Start with the eigenvalues of A and see how they get changed when we transform A into B and then C.

The eigenvalues of A are 2, -1, -0.25, 0.5, 2, and -5. For matrix B, subtract 0.25 from each eigenvalue, so the eigenvalues of B are 1.75, -1.25, -0.5, 0.25, 1.75, and -5.25. Now, C is the inverse of B, so the eigenvalues of C are the reciprocal of each of the eigenvalues of B, so the eigenvalues are $1/1.75$, $-1/1.25$, -2, 4, $1/1.75$, $-1/5.25$.

(b) Use reasoning, not MATLAB, to answer this question, and show your reasoning. By the end of the code, q is a very good approximation for an eigenvector of A. What is the corresponding eigenvalue λ?

The vector q is formed by applying the power method to the matrix C, which is used to find an approximate dominant eigenvalue and associated eigenvector. Hence, the corresponding eigenvalue to q is the largest eigenvalue (in magnitude) of the matrix C, which is the

inverse of the matrix B. So, the largest eigenvalue of C corresponds to the smallest eigenvalue of the matrix B, which has the same eigenvalues as A, but subtracting 0.25. Therefore, the corresponding eigenvalue to the matrix C is 4, to matrix B is 0.25, and to matrix A is 0.5. Hence, $\lambda = 0.5$.

(c) If you replaced 0.25 with -0.5 in line 5 of the script, which eigenvector of A would q approximate? What would be the corresponding eigenvalue of A? Like in (b), use and show reasoning.

If we replaced 0.25 in line 5 with 0.5, the new eigenvalues of B would be 1.5, -1.5, -0.75, 0, 1.5, -5.5, and so the new eigenvalues of C would be $1/1.5$, $-1/1.5$, $-1/0.75$, 0, $1/1.5$, $1/5.5$. This changes the dominant eigenvalue of C from 4 to $-1/0.75$, which corresponds to -0.75 in B, which finally corresponds to $\lambda = -0.25$ in A.

(d) After running the code with 0.25 in line 5 of the script, show how good an approximation the vector q is to an eigenvector of A by typing in MATLAB `norm((A-λ*eye(6))*q)`, where λ is the number you obtained in (b). This computes the 2-norm of the vector in parentheses; if it is small, the vector q is a very good approximation to an eigenvector with eigenvalue λ.

```
5   A = Q*D*Q';
6   B = A - 0.25*eye(6);
7   C = inv(B);
8   [V,D] = eig(A);
9   disp([V,D])
10
11  q = ones(6, 1); s = 1;
12  for j = 1:12
13      q_old = q;
14      q_new = C*q_old;
15      [~, ind] = max(abs(q_new));
16      s = q_new(ind(1));
17      q = q_new/s;
18  end
19  norm((A - 0.5*eye(6))*q)
20
```

Command Window

```
         0         0         0    2.0000


ans =

   3.5804e-04

fx >>
```

The 2-norm computed is very, very small, so we can conclude that q is a very good approximation to an eigenvector with eigenvalue $\lambda = 0.5$.

HW 5

1) Let A be an $n \times n$ real matrix and denote by $\|A\|_F$ the Frobenius norm of A. Recall that the Frobenius norm has the following equivalent definition:

$$\|A\|_F^2 = \sum_{ij} |a_{ij}|^2 = \text{trace}(A^T A).$$

a) Show that $\|A\|_F = \|UA\|_F$ for any real orthogonal $n \times n$ matrix U.

We have that $\|A\|_F = \text{trace}(A^T A)$, hence,

$$\|UA\|_F = \text{trace}((UA)^T(UA)) \qquad ((AB)^T = B^T A^T)$$

$$= \text{trace}(A^T \underbrace{U^T U}_{I} A), \text{ since } U \text{ is } n \times n \text{ orthogonal}$$

$$= \text{trace}(A^T A) = \|A\|_F$$

b) Show that $\|A\|_F = \|AV\|_F$ for any real orthogonal $n \times n$ matrix V.

Similarly, $\|AV\|_F = \text{trace}((AV)^T(AV))$

$$= \text{trace}(V^T A^T A V)$$

Since V is $n \times n$ orthogonal, for any eigenvalue λ of a matrix $M_{n \times n}$, λ is also an eigenvalue of $V^T M V$. We have that $A^T A$ is $n \times n$. Thus,

$$\text{trace}(V^T A^T A V)$$

$$= \text{trace}(A^T A) = \|A\|_F.$$

c) Conclude that $\|A\|_F = \sqrt{\sum_{i=1}^{n} \sigma_i^2}$, where $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$ are the singular values of A.

We have that $\|A\|_F^2 = \sum_{ij} |a_{ij}|^2$, which is the sum of the eigenvalues of A^TA. Now, since the eigenvalues of A^TA are the squares of the singular values of A ($|a_{ij}|^2 = \sigma_i^2$), we have

$$\|A\|_F^2 = \sum_{i}^{n} \sigma_i^2,$$

where n is the number of columns of A. Hence,

$$\sqrt{\|A\|_F^2} = \sqrt{\sum_i \sigma_i^2} = \|A\|_F.$$

2) Work this problem out by hand. Let

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Carry out the power method with starting vector $q_0 = [a, b]^T$, where $a, b \geq 0$ and $a > b$. Calculate the first few terms; what do you notice?

$$q_0 = \begin{bmatrix} a \\ b \end{bmatrix}, \quad \|q_0\|_\infty = a$$

$$q_1 = \frac{Aq_0}{\|Aq_0\|_\infty} = \frac{1}{a}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{a}\begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} b/a \\ 1 \end{bmatrix}$$

$$q_2 = \frac{Aq_1}{\|Aq_1\|_\infty} = \frac{1}{1}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} b/a \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ b/a \end{bmatrix}$$

$$q_3 = \frac{Aq_2}{\|Aq_2\|_\infty} = \frac{1}{1}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 1 \\ b/a \end{bmatrix} = \begin{bmatrix} b/a \\ 1 \end{bmatrix}$$

$$q_4 = \begin{bmatrix} 1 \\ b/a \end{bmatrix}, \quad q_5 = \begin{bmatrix} b/a \\ 1 \end{bmatrix}, \ldots$$

I notice that after the first term q_0, the terms oscillate back and forth such that $q_i = \begin{bmatrix} b/a \\ 1 \end{bmatrix}$ where i is odd and $q_j = \begin{bmatrix} 1 \\ b/a \end{bmatrix}$ where j is even. The sequence fails to converge.

Explain why the sequence fails to converge. What is the problem with the convergence argument we had in lecture? Which of the conditions is violated?

We have that

$$\det(A - \lambda I) = \begin{bmatrix} 0-\lambda & 1 \\ 1 & 0-\lambda \end{bmatrix} = \lambda^2 - 1 = 0$$

$$\Rightarrow (\lambda - 1)(\lambda + 1) = 0 \Rightarrow \lambda_1 = 1, \lambda_2 = -1$$

$$\Rightarrow |\lambda_1| = |\lambda_2|.$$

The sequence fails to converge because the matrix A does not have a dominant eigenvalue.

3) Let A be an $n \times n$ complex matrix with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ ordered in decreasing order of magnitude, that is, $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$. Let $v_1, \ldots, v_n$ be the corresponding eigenvectors.

Assume that $\lambda_1 = 5$, $\lambda_2 = 3$, $\lambda_3 = 2$, and consider performing the power method on A starting with a vector q which does not depend on the eigenvector v_1:

$$q = v_2 + \sum_{i=3}^{n} c_i v_i \quad \text{, for some } c_i \in \mathbb{C},\ 3 \leq i \leq n.$$

If we start the power method on A with the vector q, will the method converge? Why or why not? Explain.

Let us analyze the convergence of the power method on A with the given q. We have $q = v_2 + \sum_{i=3}^{n} c_i v_i$, and $Av_i = \lambda_i v_i$. Hence,

$$\begin{aligned} Aq &= A\left(v_2 + \sum_{i=3}^{n} c_i v_i\right) \\ &= Av_2 + A\sum_{i=3}^{n} c_i v_i \\ &= \lambda_2 v_2 + \sum_{i=3}^{n} c_i \lambda_i v_i \end{aligned}$$

$$\begin{aligned} A^2 q &= A^2\left(v_2 + \sum_{i=3}^{n} c_i v_i\right) \\ &= A^2 v_2 + A^2 \sum_{i=3}^{n} c_i v_i \\ &= \lambda_2^2 v_2 + \sum_{i=3}^{n} c_i \lambda_i^2 v_i \end{aligned}$$

$$\vdots$$

$$A^j q = \lambda_2^j v_2 + \sum_{i=3}^{n} c_i \lambda_i^j v_i$$

$$= \lambda_2^j v_2 + c_3 \lambda_3^j v_3 + \ldots + c_n \lambda_n^j v_n$$

$$= \lambda_2^j \left(v_2 + c_3 \underbrace{\frac{\lambda_3^j}{\lambda_2^j}} v_3 + \ldots + c_n \frac{\lambda_n^j}{\lambda_2^j} v_n \right)$$

As j tends to ∞, $\frac{\lambda_3^j}{\lambda_2^j}$ tend to 0.

$$\Rightarrow A^j q = \lambda_2^j v_2$$

$$= 3^j v_2 .$$

Yes, the power method on A with vector q will converge, just not to the first eigenvector v_1, but to the second, v_2. This is a consequence of the fact that the initial q does not have a component along v_1, but has components along $v_2, v_3, \ldots, v_n$. Since $|\lambda_2| > |\lambda_i|$ for $i = 3, \ldots, n$, the power method with scaling gives us a vector v aligned with v_2 with corresponding eigenvalue $\lambda_2 = 3$, and $\|v\|_\infty = 1$.

4) Let $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, the 2×2 identity matrix.

(a) Calculate the characteristic polynomial of A and show it has one eigenvalue $\lambda = 1$ with algebraic multiplicity 2.

$$\det(A - \lambda I) = \det\left(\begin{bmatrix} 1-\lambda & 0 \\ 0 & 1-\lambda \end{bmatrix}\right) = (1-\lambda)(1-\lambda) = 0$$

$\Rightarrow \lambda = 1$ is a root of $\det(A - \lambda I) = 0$ twice, hence A has one eigenvalue with algebraic multiplicity 2.

b) We now perturb one coefficient of the characteristic polynomial slightly and consider the equation

$$\lambda^2 - 2\lambda + (1 - \varepsilon) = 0,$$

where $0 < \varepsilon << 1$. Solve the equation for the roots $\hat{\lambda}_1$ and $\hat{\lambda}_2$ in terms of ε.

We use the quadratic equation:

$$\lambda = \frac{2 \pm \sqrt{4 - 4(1-\varepsilon)}}{2} = \frac{2 \pm \sqrt{4\varepsilon}}{2}$$

$$\Rightarrow \quad \hat{\lambda}_1 = \frac{2 + \sqrt{4\varepsilon}}{2} = \frac{2 + 2\sqrt{\varepsilon}}{2} = 1 + \sqrt{\varepsilon}$$

$$\hat{\lambda}_2 = \frac{2 - \sqrt{4\varepsilon}}{2} = \frac{2 - 2\sqrt{\varepsilon}}{2} = 1 - \sqrt{\varepsilon}$$

c) Show that if $\varepsilon = 10^{-12}$, $|\hat{\lambda}_1 - \lambda|$ and $|\hat{\lambda}_2 - \lambda|$ are one million times bigger than ε.

Let $\varepsilon = 10^{-12}$. Then

$$|\hat{\lambda}_1 - \lambda| = |1 + \sqrt{\varepsilon} - 1| = \sqrt{\varepsilon} = 10^{-6} = 10^6 \varepsilon$$

$$|\hat{\lambda}_2 - \lambda| = |1 - \sqrt{\varepsilon} - 1| = \sqrt{\varepsilon} = 10^{-6} = 10^6 \varepsilon$$

d) Conclude that, given a relative change $\varepsilon << 1$ in one of the coefficients of the polynomial, the relative change in the eigenvalues can grow arbitrarily large; therefore, the computation is numerically unstable.

In part (b), we observed that the relative change in one of the coefficients by ε changed $\lambda = 1$ to $\hat{\lambda}_1 = 1 + \sqrt{\varepsilon}$ $\hat{\lambda}_2 = 1 - \sqrt{\varepsilon}$. Then, in part (c), we saw how this change in the coefficient drastically changed the difference between our initial and final eigenvalues, a change from $\lambda = 1$ to $\hat{\lambda}_1 = 1 + 10^{-6}$ and $\hat{\lambda}_2 = 1 - 10^{-6}$. Not only did we change the numerical value of the initial eigenvalue, but we changed the overall number of eigenvalues and their algebraic multiplicity. This can lead to major complications if the eigenvalues have a high multiplicity, or if their values are very close to one another, since the relative change between them will grow immensely. Hence, the computation is quite numerically unstable.

5) Let $A = \begin{bmatrix} 2 & 2 & 2 \\ 4 & 0 & 2 \\ 2 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$.

a) Use the MATLAB function rank to figure the rank of A.

b) Use MATLAB's pinv to compute the pseudoinverse of A.

c) Use MATLAB's svd command to find the singular values of AA^+ and A^+A, and calculate AA^+ and A^+A.

Command Window

```
>> A = [2 2 2; 4 0 2; 2 0 1; 0 1 1]

A =

     2     2     2
     4     0     2
     2     0     1
     0     1     1

>> rank(A)

ans =                    (a)

     3

>> A_pinv = pinv(A)

A_pinv =

    0.5000    0.0000    0.0000   -1.0000     (b)
    1.0000   -0.4000   -0.2000   -1.0000
   -1.0000    0.4000    0.2000    2.0000

>> AA_pinv = A*A_pinv

AA_pinv =

    1.0000    0.0000    0.0000   -0.0000
   -0.0000    0.8000    0.4000         0
   -0.0000    0.4000    0.2000         0
    0.0000   -0.0000   -0.0000    1.0000

>> [U,S,V] = svd(AA_pinv)
```

```
U =

   -1.0000   -0.0000    0.0000         0
    0.0000   -0.8698   -0.2084   -0.4472
    0.0000   -0.4349   -0.1042    0.8944
   -0.0000    0.2330   -0.9725         0


S =

    1.0000         0         0         0
         0    1.0000         0         0
         0         0    1.0000         0
         0         0         0         0


V =

   -1.0000         0         0         0
         0   -0.8698   -0.2084    0.4472
         0   -0.4349   -0.1042   -0.8944
         0    0.2330   -0.9725   -0.0000

>> A_pinvA = A_pinv*A

A_pinvA =

    1.0000   -0.0000    0.0000
    0.0000    1.0000         0
   -0.0000         0    1.0000

>> [U,S,V] = svd(A_pinvA)

U =

   -1.0000    0.0000   -0.0000
   -0.0000   -0.6919    0.7220
    0.0000    0.7220    0.6919


S =

    1.0000         0         0
         0    1.0000         0
         0         0    1.0000


V =

   -1.0000         0         0
         0   -0.6919    0.7220
         0    0.7220    0.6919
```

(c)

HW 4

1) Let $A \in \mathbb{R}^{n \times m}$, $n \geq m$, $\text{rank}(A) = m$. Compute the SVD of the matrix $(A^TA)^{-1}A^T$ in terms of the SVD of $A = U\Sigma V^T$, and explain what the parts are.

a) Based on the definitions in class, conceive that the pseudoinverse of A has the format $A^+ = (A^TA)^{-1}A^T$.

We have $A = U\Sigma V^T$, where A is $n \times m$, U is orthogonal and $n \times n$, Σ is diagonal and $n \times m$, and V is orthogonal and $m \times m$. $\text{Rank}(A) = m$, so we have m nonzero singular values in Σ.

Now, $A^T = (U\Sigma V^T)^T = (V^T)^T \Sigma^T U^T$

$= V \Sigma^T U^T$

$A^TA = (V\Sigma^T U^T)(U\Sigma V^T)$ $\quad (U^TU = I)$

$= V\Sigma^T\Sigma V^T$

We know that $A^T = \underset{m \times n}{\left[\quad\right]}$, and $\text{rank}(A) = m$. Hence, A^TA is $m \times m$ with full rank. Thus, $(A^TA)^{-1}$ exists.

$(A^TA)^{-1} = (V\Sigma^T\Sigma V^T)^{-1}$

$= (V^T)^{-1}(\Sigma^T\Sigma)^{-1}(V)^{-1}$ $\quad (V$ is orthogonal; $V^T = V^{-1})$

$= V(\Sigma^T\Sigma)^{-1}V^T$

$\begin{bmatrix} \sigma_1 & \\ & \end{bmatrix}$

Since $n \geq m$, Σ looks like $\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \\ \hline 0 & & \end{bmatrix}$ (more rows than cols).

$$\therefore \Sigma^T\Sigma = \underset{m\times n}{\left[\begin{array}{ccc|c} \sigma_1 & & & 0 \\ & \ddots & & \\ & & \sigma_m & \end{array}\right]} \underset{n\times m}{\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \\ \hline 0 & & \end{bmatrix}} = \underset{m\times m}{\begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_m^2 \end{bmatrix}} \quad \left((\Sigma^T\Sigma)^{-1} \text{ exists, since } \Sigma^T\Sigma \text{ has full rank}\right)$$

$$\therefore (\Sigma^T\Sigma)^{-1} = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_m^2 \end{bmatrix}^{-1} = \begin{bmatrix} \sigma_1^{-2} & & \\ & \ddots & \\ & & \sigma_m^{-2} \end{bmatrix}$$

$$(A^TA)^{-1}A^T = V(\Sigma^T\Sigma)^{-1}V^T(V\Sigma^TU^T) \qquad (V^TV = I)$$

$$= V(\Sigma^T\Sigma)^{-1}\Sigma^TU^T$$

$$(\Sigma^T\Sigma)^{-1}\Sigma^T = \underset{m\times m}{\begin{bmatrix} \sigma_1^{-2} & & \\ & \ddots & \\ & & \sigma_m^{-2} \end{bmatrix}} \underset{m\times n}{\left[\begin{array}{ccc|c} \sigma_1 & & & 0 \\ & \ddots & & \\ & & \sigma_m & \end{array}\right]} = \underset{m\times n}{\left[\begin{array}{ccc|c} \sigma_1^{-1} & & & 0 \\ & \ddots & & \\ & & \sigma_m^{-1} & \end{array}\right]} = \Sigma^+$$

Hence, the SVD of $(A^TA)^{-1}A^T$ is $V\Sigma^+U^T$, where Σ^+ is the pseudo inverse of Σ:

$$\Sigma = \underset{n\times m}{\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \\ \hline 0 & & \end{bmatrix}} \qquad \Sigma^+ = \underset{m\times n}{\left[\begin{array}{ccc|c} \sigma_1^{-1} & & & 0 \\ & \ddots & & \\ & & \sigma_m^{-1} & \end{array}\right]}$$

$$\Sigma\Sigma^{+} = \begin{bmatrix} I_m & 0_{m\times(n-m)} \\ 0_{(n-m)\times m} & 0_{(n-m)\times(n-m)} \end{bmatrix}$$

Now, using the fact that $\Sigma^{+} = (\Sigma^T\Sigma)^{-1}\Sigma^T$, and $(A^TA)^{-1}$ exists, we can conclude that $A^{+} = (A^TA)^{-1}A^T = V\Sigma^{+}U^T$.

b) use (a) to calculate $A^{+}A$.

$$A^{+}A = (V\Sigma^{+}U^T)(U\Sigma V^T) \qquad (U^TU = I_n)$$

$$= V\underbrace{\Sigma^{+}\Sigma}\,V^T$$

$$\begin{bmatrix} I_m & \vdots \\ \cdots & 0 \end{bmatrix}$$

$$= VV^T = I_m$$

c) Use the SVDs of A and A^{+} to calculate AA^{+}, simplifying as much as possible

$$AA^{+} = (U\Sigma V^T)(V\Sigma^{+}U^T)$$

$$= U\underbrace{\Sigma\Sigma^{+}}\,U^T$$

$$\begin{bmatrix} I_m & \vdots \\ \cdots & 0 \end{bmatrix}$$

$$= UU^T$$

2) Let $A = U \cdot \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot V$, $B = P \cdot \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \cdot Q$,

where $U \in \mathbb{R}^{3\times3}$, $V \in \mathbb{R}^{3\times3}$, $P \in \mathbb{R}^{3\times3}$, $Q \in \mathbb{R}^{4\times4}$ are all given orthogonal matrices. Compute $\|A\|_2$, $\|A^{-1}\|_2$, $\kappa_2(A)$, $\|B\|_2$ and the pseudoinverse $B^{\dagger}$ (the answer will be in terms of P and Q).

$\|A\|_2$ = largest singular value of A = σ_1

$\|A^{-1}\|_2$ = largest singular value of A^{-1}

$A^{-1} = V\Sigma^{-1}U^T$, $\Sigma^{-1} = \begin{bmatrix} \sigma_1^{-1} & & \\ & \ddots & \\ & & \sigma_n^{-1} \end{bmatrix}$

$\kappa_2(A) = \frac{\sigma_1}{\sigma_n}$

$\Rightarrow$ $A = U \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} V$, $A^{-1} = V \begin{bmatrix} 1/3 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix} U$

$\|A\|_2 = 3$

$\Rightarrow$ $\|A^{-1}\|_2 = \frac{1}{1} = 1$

$\kappa_2(A) = \frac{3}{1} = 3$

$\Rightarrow$ $B = P \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} Q$, $\|B\|_2 = 4$

$$B^+ = Q\left(\left(\Sigma^T\Sigma\right)^{-1}\Sigma^T\right)P$$

$$\Sigma^T = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \Sigma^T\Sigma = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 16 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\left(\Sigma^T\Sigma\right)^{-1} = \begin{bmatrix} 1/16 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad \left(\Sigma^T\Sigma\right)^{-1}\Sigma^T = \begin{bmatrix} 1/16 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} 4 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$4 \times 4 \qquad 4 \times 3$

$$B^+ = Q\begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}P$$

3) Let $A \in \mathbb{R}^{n \times m}$, $n \geq m$.

a) Use the SVD of A to deduce the SVD of A^TA.

$$A = \underset{n}{\begin{bmatrix} \\ \\ \\ \end{bmatrix}}_{m} = U \Sigma V^T = \underset{n \times n}{\begin{bmatrix} U \end{bmatrix}} \underset{n \times m}{\begin{bmatrix} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ & 0 & & \end{bmatrix}} \underset{m \times m}{\begin{bmatrix} V^T \end{bmatrix}}$$

$$A^T = (U \Sigma V^T)^T = (V^T)^T \Sigma^T U^T = V \Sigma^T U^T$$

$$A^TA = (V \Sigma^T U^T)(U \Sigma V^T)$$

$$= V \Sigma^T \Sigma V^T$$

b) If $m = n$ and A is full rank, use (a) to show that $\|A^TA\|_2 = \|A\|_2^2$ and that $K_2(A^TA) = K_2(A)^2$.

$$A = \underset{n}{\begin{bmatrix} \text{rank}(A) = m \\ \text{or } n \end{bmatrix}}_{\substack{m \\ \text{(square)}}} = U \Sigma V^T = \underset{n \times n}{\begin{bmatrix} U \end{bmatrix}} \underset{n \times m}{\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \end{bmatrix}} \underset{m \times m}{\begin{bmatrix} V \end{bmatrix}}$$

We know that $\|A\|_2$ = largest singular value of $A = \sigma_1$, so $\|A\|_2^2 = \sigma_1^2$. Thus, $\|A^TA\|_2$ = largest singular value of A^TA.

We previously found that $A^TA = V \Sigma^T \Sigma V^T$, where $\Sigma^T \Sigma = \Sigma^2 = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \end{bmatrix}$, since Σ is diagonal. Hence,

$$\begin{bmatrix} & \\ & \sigma_m^2 \end{bmatrix}$$

$\|A^TA\|_2 = \sigma_1^2 = \|A_2\|^2$.

we also know that $K_2(A) = \frac{\sigma_1}{\sigma_m}$, and so $K_2(A)^2 = \frac{\sigma_1^2}{\sigma_m^2}$.

we have that $K_2(A^TA) = \frac{\sigma_1^2}{\sigma_m^2} = K_2(A)^2$.

4) Work this exercise using pencil and paper. You can use MATLAB to check your work. Let A be the following exterior (or outer) product of two vectors:

$$A = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 \end{bmatrix}$$

Note that A is 3×2. Answer the questions below; you can do so <u>without ever forming A explicitly</u>.

a) What is the rank r of A?

Since we are multiplying a 3×1 vector with a 1×2 vector, the result has dimension 3×2. Since the 1×2 vector is [1 1], the result will be a matrix of 2 columns with the same 3 entries in each column. Thus, the result is a matrix with linearly dependent columns. Hence, rank(A) = 1.

b) Think about the "sum of rank one matrices" expression for the SVD of A, then consider the reduced SVD of A: $A = U_r \Sigma_r V_r^T$. What are the sizes of U_r, Σ_r, and V_r?

We have $\underset{3\times 2}{[A]} = \underset{3\times 3}{[U]} \underset{3\times 2}{\begin{bmatrix} \sigma_1 & \\ & \sigma_2 \\ 0 & 0 \end{bmatrix}} \underset{2\times 2}{[V^T]}$,

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T = \text{sum of rank 1 matrices}$$

since rank(A) = r = 1, we have

$A = \sigma_1 u_1 v_1^T \Rightarrow A = U_1 \Sigma_1 V_1^T$

$\Rightarrow \Sigma$ has one singular value, so Σ is 1×1. Hence, U_1 must be 3×1, and V_1^T must be 1×2.

c) Use the fact that the columns of U_r and V_r are orthonormal to figure out U_r, V_r, and Σ_r.

We know that U_r is the left singular vector of A, meaning it is the normalized column space of A. We know that rank(A) = 1, so $col(A) = \left\{\begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}\right\}$. Hence,

$$U_r = \frac{1}{\left\|\begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}\right\|} \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix} = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}.$$

Similarly, V is the right singular vector, so it is the normalized row space of A, where $row(A) = \{[1 \quad 1]\}$. Thus,

$$V_r = \frac{1}{\|[1 \quad 1]\|} [1 \quad 1] = \frac{1}{\sqrt{2}} [1 \quad 1] = [1/\sqrt{2} \quad 1/\sqrt{2}]$$

Now, to find Σ_r, we have

$$A^T A = \begin{bmatrix} 1 \\ 1 \end{bmatrix} [1 \quad 2 \quad -1] \cdot \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix} [1 \quad 1] = \begin{bmatrix} 6 & 6 \\ 6 & 6 \end{bmatrix}$$

$$\sigma_1 = \sqrt{6} \Rightarrow \Sigma_1 = [\sqrt{6}]$$

5) Run the attached low-rank-approximation.m MATLAB code.

a) Explain line by line what the code does (you might need to google some of the commands).

```
% load image

A = imread('street2.jpg');
A = rgb2gray(A);
B = double(A);
```

- Reads image specified by 'street2.jpg'
- A is an RGB image, so this converts the color image to a grayscale image
- converts the type of the entries in A from int to double, stores the matrix w/ double type entries in B

```
% compute SVD

size(B)
r = rank(B)
[U,S,V] = svd(B);
```

- gets size of B
- gets rank of B and stores the number in variable r
- computes the SVD of B, such that $B = U \cdot S \cdot V'$

```
% approximate image

ranks = [1 2 4 8 16 32 64 r];
```

creates a row vector of these specified values, stored in variable "ranks"

```
l = length(ranks);
```

- computes the length (number of elements) of ranks

```
for i = 1:l
```

- for loop, iterates l times (# of elements in "ranks")

```
    % compute rank i approximation

    k = ranks(i);
```

- stores the ith element of "ranks" in variable k

```
    approxB = U(:,1:k)*S(1:k,1:k)*V(:,1:k)';
```

- corresponds to $\delta_k u_k v_k$, meaning for each entry k in ranks, we are computing the rank-k approximation of B.

```
    approxA = uint8(approxB);
```

- converts (approx B) into an unsigned 8-bit int

```
    % plot images

    figure(1)
    subplot(2,4,i)
    imshow(approxA);
    title(sprintf('rank %d approximation',k))
```

- gets figure with number property = 1
- divides current figure into a 2x4 grid, creates axes for each iteration
- displays grayscale image of approx(A) according to figure
- titles each subplot using the print function, %d corresponds to the value of k during the particular iteration.

```
end
```

end loop

b) Explain what the algorithm, as a whole, does.

The algorithm takes an image, converts it to grayscale, and then creates visual approximations of the image with rank 1, 2, ..., 64, and finally the actual rank. Essentially, the algorithm uses Single Value Decomposition to compress/expand an image by using a certain number of singular values.

c) Note that the approximation gets better as we increase k. Even when $k = 32$, the resulting approximation looks reasonable. What is the advantage to use/store the $k=32$ approximation instead of the original image? What is the disadvantage?

We know that the matrices U_{32}, Σ_{32}, and V_{32}^T are much smaller than U_{480}, Σ_{480}, V_{480}^T. Hence, storing the $k=32$ approx. uses much less space than the $k=480$ approx. Also, when using the image, such as for a newspaper, since the $k=32$ approx. has fewer singular values, processing the image is much faster than the original.

The disadvantage is that since we are approximating using a much smaller rank, and although the approximation is reasonable, it will never have as many full details as the original.

HW 3

SABELLA SAULINO

1) Let $w_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$, $w_2 = \begin{bmatrix} 3 \\ 3 \\ -1 \\ -1 \end{bmatrix}$, $w_3 = \begin{bmatrix} 6 \\ 0 \\ 2 \\ 0 \end{bmatrix}$ be vectors in $\mathbb{R}^4$.

Apply the classical GS process to find an orthonormal basis for the subspace spanned by w_1, w_2, w_3.

i) $q_1 = \frac{w_1}{\|w_1\|_2} = \frac{1}{2}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix}$ $\quad r_{11} = 2$

ii) $\tilde{q}_2 = w_2 - \langle q_1, w_2 \rangle q_1$

$$= \begin{bmatrix} 3 \\ 3 \\ -1 \\ -1 \end{bmatrix} - \left((1/2, 1/2, 1/2, 1/2) \cdot (3, 3, -1, -1) \right) \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix}$$

$$= \begin{bmatrix} 3 \\ 3 \\ -1 \\ -1 \end{bmatrix} - (2) \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix} \qquad r_{12} = 2$$

$$= \begin{bmatrix} 3 \\ 3 \\ -1 \\ -1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ -2 \\ -2 \end{bmatrix}$$

$$q_2 = \frac{\tilde{q}_2}{\|q_2\|_2} = \frac{1}{4}\begin{bmatrix} 2 \\ 2 \\ -2 \\ -2 \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1/2 \\ -1/2 \\ -1/2 \end{bmatrix} \qquad r_{22} = 4$$

iii) $\tilde{q}_3 = w_3 - \langle q_1, w_3 \rangle q_1 - \langle q_2, w_3 \rangle q_2$

$$= \begin{bmatrix} 6 \\ 0 \\ 2 \\ 0 \end{bmatrix} - \left((1/2, 1/2, 1/2, 1/2) \cdot (6, 0, 2, 0) \right) \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix}$$

$$= \begin{bmatrix} 6 \\ 0 \\ 2 \\ 0 \end{bmatrix} - (4) \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix} \qquad r_{13} = 4$$

$$= \begin{bmatrix} 6 \\ 0 \\ 2 \\ 0 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 4 \\ -2 \\ 0 \\ -2 \end{bmatrix}$$

$$\begin{bmatrix} 4 \\ -2 \\ 0 \\ -2 \end{bmatrix} - \left((1/2, 1/2, -1/2, -1/2) \cdot (6, 0, 2, 0) \right) \begin{bmatrix} 1/2 \\ 1/2 \\ -1/2 \\ -1/2 \end{bmatrix}$$

$$= \begin{bmatrix} 4 \\ -2 \\ 0 \\ -2 \end{bmatrix} - (2) \begin{bmatrix} 1/2 \\ 1/2 \\ -1/2 \\ -1/2 \end{bmatrix} = \begin{bmatrix} 3 \\ -3 \\ 1 \\ -1 \end{bmatrix} \qquad r_{23} = 2$$

$$q_3 = \frac{\tilde{q}_3}{\|q_3\|_2} = \frac{1}{\sqrt{20}} \begin{bmatrix} 3 \\ -3 \\ 1 \\ -1 \end{bmatrix} \qquad r_{33} = \sqrt{20}$$

orthonormal basis : $\left\{ \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix}, \begin{bmatrix} 1/2 \\ 1/2 \\ -1/2 \\ -1/2 \end{bmatrix}, \frac{1}{\sqrt{20}} \begin{bmatrix} 3 \\ -3 \\ 1 \\ -1 \end{bmatrix} \right\}$

2) For $A = [w_1, w_2]$ with w_1, w_2 as above, find the minimizer to the least squares problem $\min_{x \in \mathbb{R}^2} \|b - Ax\|_2$ for $b = [1, 1, 0, 1]^T$, using the QR method discussed in class.

$$A = \begin{bmatrix} 1 & 3 \\ 1 & 3 \\ 1 & -1 \\ 1 & -1 \end{bmatrix} \quad Q = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \end{bmatrix} \quad R = \begin{bmatrix} r_{11} & r_{12} \\ 2 & 2 \\ 0 & 4 \end{bmatrix}_{r_{22}} \quad b = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

$w_1 \quad w_2$ $\qquad$ $q_1 \quad q_2$

$$\min_{x \in \mathbb{R}^2} \|b - Ax\| \implies x = R^{-1} Q^T b$$

$$R^{-1} = \frac{1}{8} \begin{bmatrix} 4 & -2 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1/2 & -1/4 \\ 0 & 1/4 \end{bmatrix} \quad Q^T = \begin{bmatrix} 1/2 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & -1/2 & -1/2 \end{bmatrix}$$

$$R^{-1} Q^T b = \underbrace{\begin{bmatrix} 1/2 & -1/4 \\ 0 & 1/4 \end{bmatrix}}_{2 \times 2} \underbrace{\begin{bmatrix} 1/2 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & -1/2 & -1/2 \end{bmatrix}}_{2 \times 4} \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1/8 & 1/8 & 3/8 & 3/8 \\ 1/8 & 1/8 & -1/8 & -1/8 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 5/8 \\ 1/8 \end{bmatrix} = x$$

3) Let Q be an orthogonal $n \times n$ matrix ($QQ^T = Q^TQ = I$) and consider the induced 2-norm on $n \times n$ matrices A,

$$\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}$$

with $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$ the usual 2-norm on vectors of length n

(a) Calculate $\|Q\|_2$ and $K_2(Q) = K_{\|\cdot\|_2}(Q)$

Multiplication by orthogonal matrices preserves 2-norm:

$$\begin{aligned} \|Qx\|_2 &= \langle Qx, Qx \rangle \\ &= (Qx)^T(Qx) \\ &= x^T \underbrace{Q^TQ}_{I} x \\ &= \langle x, x \rangle = \|x\|_2 . \end{aligned}$$

Hence, $\|Q\|_2 = \frac{\|Qx\|_2}{\|x\|_2} = \frac{\|x\|_2}{\|x\|_2} = 1$.

Now, $K_2(Q) = K_{\|\cdot\|_2}(Q) = \|Q\|_2 \|Q^{-1}\|_2$.

we have that $QQ^T = Q^TQ = I \Rightarrow Q = Q^{-1}$.

Hence, $\|Q\|_2 = \|Q^{-1}\|_2 = 1$.

Therefore, $K_2(Q) = 1$.

(b) Consider the perturbed system $(Q + \delta Q)\hat{x} = b$, with $\hat{x} = x + \delta x$. What can we say about the relative

error $\|\delta x\|_2 / \|\hat{x}\|_2$ of the solution, in terms of the induced 2-norm $\|\delta Q\|_2$ of the perturbation matrix δQ?

By a previous Theorem, we have

$$\frac{\|\delta x\|_2}{\|\hat{x}\|_2} \leq K_2(Q) \frac{\|\delta Q\|_2}{\|Q\|_2}.$$

We found in part (a) that $K_2(Q) = 1$. Hence,

$$\frac{\|\delta x\|_2}{\|x\|_2} \leq \frac{\|\delta Q\|_2}{\|Q\|_2}.$$

Also, we found that $\|Q\|_2 = 1$. Thus,

$$\frac{\|\delta x\|_2}{\|x\|_2} \leq \|\delta Q\|_2.$$

We can say that the relative error $\frac{\|\delta x\|_2}{\|x\|_2}$ of the solution is bounded by the relative perturbation $\|\delta Q\|_2$ of the matrix Q.

4) Let M be an $n \times m$ matrix, $n \geq m$, and A be an $m \times m$ matrix such that $A = M^T M$.

Also, suppose M is full-rank.

If $M = QR$ is the reduced QR factorization of M, write

down the relationship b/w R and A. Specifically, given the properties of M and R, what is R to A. Justify your answer.

First, let's draw a picture.

$$n\left[\begin{array}{c} M \\ m \end{array}\right] = n\left[\begin{array}{c} Q \\ m \end{array}\right] m\left[\begin{array}{c} R \\ m \end{array}\right] \qquad (n \geq m)$$

$$\Rightarrow \text{rank } M = m$$

$$m\left[\begin{array}{c} M^T \\ n \end{array}\right] n\left[\begin{array}{c} M \\ m \end{array}\right] = m\left[\begin{array}{c} A \\ m \end{array}\right]$$

We know by the Gram-Schmidt process that Q consists of m orthonormal columns and R is an upper triangular matrix. Hence, Q is orthogonal, so $Q^TQ = I_m$. Since M is full rank and R is $m \times m$, R is invertible. Also, since M is full rank, A is symmetric and positive definite.

We have $A = M^TM$, $M = QR$

$$\Rightarrow A = (QR)^T(QR) \qquad ((AB)^T = B^TA^T)$$

$$\Rightarrow A = R^T \underbrace{Q^TQ}_{I_m} R$$

$$\Rightarrow A = R^TR$$

Hence, $A = R^T R$.

Now, we have that R is upper triangular and invertible and A is symmetric and positive definite. Therefore, R is the Cholesky factor of A.

5) Use the MATLAB code for classical and modified Gram-Schmidt Process

Command Window

```
>> A = hilb(8);
>> [Q1, R1] = gs(A);
Unrecognized function or variable 'gs'.

Did you mean:
>> [Q1, R1] = classicalgs(A);

Q =

  Columns 1 through 6

    0.8091   -0.5467    0.2064   -0.0605    0.0144   -0.0028
    0.4046    0.2825   -0.6766    0.4934   -0.2236    0.0714
    0.2697    0.3736   -0.2413   -0.4274    0.6135   -0.3855
    0.2023    0.3636    0.0543   -0.4519   -0.1109    0.5920
    0.1618    0.3354    0.2226   -0.2124   -0.4481    0.0758
    0.1349    0.3059    0.3161    0.0579   -0.3360   -0.4428
    0.1156    0.2792    0.3669    0.2932    0.0292   -0.3318
    0.1011    0.2559    0.3929    0.4813    0.4965    0.4289

  Columns 7 through 8

    0.0001   -0.0001
   -0.0074    0.0068
    0.0958   -0.0926
   -0.3983    0.3931
    0.5803   -0.5805
   -0.0057    0.0097
   -0.6117    0.6135
    0.3482   -0.3513


R =

  Columns 1 through 6
```

```
    1.2359    0.7192    0.5214    0.4130    0.3435    0.2947
         0    0.1499    0.1665    0.1612    0.1512    0.1407
         0         0    0.0118    0.0192    0.0233    0.0254
         0         0         0    0.0007    0.0015    0.0023
         0         0         0         0    0.0000    0.0001
         0         0         0         0         0    0.0000
         0         0         0         0         0         0
         0         0         0         0         0         0

  Columns 7 through 8

    0.2584    0.2303
    0.1308    0.1220
    0.0264    0.0267
    0.0028    0.0032
    0.0002    0.0002
    0.0000    0.0000
    0.0000    0.0000
         0    0.0000

>> [Q2, R2] = modifiedgs(A);
>> checkQ = norm(Q1-Q2)

checkQ =

    1.4138

>> checkR = norm(R1-R2)

checkR =

   2.0354e-06

>> orthoQ1 = norm(Q1'*Q1-eye(8))

orthoQ1 =

    1.0325

>> orthoQ2 = norm(Q2'*Q2-eye(8))

orthoQ2 =

   4.3754e-07

>>
```

HW 2

ISABELLA SAULINO

1) a) verify that $\|x\|_1$ is a norm

i) $\|x\|_1 > 0$, $\forall x \neq 0$; $\|x\|_1 = 0$ when $x = 0$

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| = |x_1| + |x_2| + \ldots + |x_n| .$$

let $x = 0$. Then $\|x\|_1 = |0| = 0$.

let $x \neq 0$. Then $\|x\|_1 = |x_1| + |x_2| + \ldots + |x_n|$ is always positive since we take the absolute value of each entry and add.

ii) $\forall x$, $\|cx\|_1 = |c| \cdot \|x\|_1$

$$\|cx\|_1 = \sum_{i=1}^{n} |cx_i| = |cx_1| + |cx_2| + \ldots + |cx_n|$$

$$\Rightarrow |c| \sum_{i=1}^{n} = |c| (|x_1| + |x_2| + \ldots + |x_n|)$$

$$= |c| \, \|x\|_1 .$$

iii) $\|x+y\|_1 \leq \|x\|_1 + \|y\|_1$

$$\|x+y\|_1 = \sum_{i=1}^{n} |x_i + y_i| = |x_1 + y_1| + |x_2 + y_2| + \ldots + |x_n + y_n|$$

If <u>only</u> x_i or <u>only</u> y_i is negative, $|x_i + y_i| < |x_i| + |y_i|$.

If both x_i and y_i are the same sign, $|x_i + y_i| = |x_i| + |y_i|$.

Hence, $\|x+y\|_1 \leq \|x\| + \|y\|$.

b) verify that $\|x\|_\infty$ is a norm

i) $\forall x \neq 0, \|x\|_\infty > 0$; $\|x\|_\infty = 0$ when $x = 0$

let $x \neq 0$. Then $\|x\|_\infty = \max_{i=1:n} \{|x_i|\}$, where the maximum $|x_i|$ is always positive. Hence, $\|x\|_\infty > 0$ when $x \neq 0$.

let $x = 0$. Then $\|x\|_\infty = \max_{i=1:n} \{|x_i|\} = |0| = 0$.

ii) $\forall x \in \mathbb{R}^n, \|cx\|_\infty = |c| \cdot \|x\|_\infty$

$$\|cx\|_\infty = \max_{i=1:n} \{|cx_i|\} = \max\{|cx_1|, |cx_2|, \ldots, |cx_n|\}$$

$$= \max\{|c|(|x_1|, |x_2|, \ldots, |x_n|)\} = |c| \cdot \max\{|x_1|, \ldots, |x_n|\}$$

$$= |c| \cdot \max_{i=1:n} \{|x_i|\} = |c| \|x\|_\infty$$

- MAX NORM : L_∞ norm

$\|x\|_\infty = \max\{|x_1|, |x_2|, \ldots, |x_n|\}$

$= \max_{i=1:n} \{|x_i|\}$

for vector $x = [x_1, x_2, \cdots, x_n]$

iii) $\|x+y\|_\infty \leq \|x\|_\infty + \|y\|_\infty$

$$\|x+y\|_\infty = \max_{i=1:n} \{|x_i + y_i|\} = \max\{|x_1+y_1| + |x_2+y_2| + \ldots + |x_n+y_n|\}$$

If only x_i or only y_i is positive, $|x_i + y_i| < |x_i| + |y_i|$.

If x_i and y_i are the same sign, $|x_i + y_i| = |x_i| + |y_i|$.

Hence, for $\max_{i=1:n} \{|x_i + y_i|\}$, $\|x+y\|_\infty \leq \|x\|_\infty + \|y\|_\infty$.

2) Let $A = \alpha I$ be a multiple of the $n \times n$ identity matrix, with $\alpha \in \mathbb{R}$, $\alpha \neq 0$, and consider $\|\cdot\|$ to be an induced matrix norm. Calculate $\|A\|$, $\|A^{-1}\|$, $\det(A)$, and $\kappa_{\|\cdot\|}(A)$.

$A = \alpha I$, $\alpha \neq 0 \Rightarrow \|A\| = \|\alpha I\| = |\alpha| \|I\|$

$= |\alpha| \cdot 1 = |\alpha|$

$\|A^{-1}\| = \|(\alpha I)^{-1}\| = \frac{1}{|\alpha|}\|I\| = \frac{1}{|\alpha|} \cdot 1 = \frac{1}{|\alpha|}$

$\det(A) = \det(\alpha I) = \underbrace{\alpha \cdot \alpha \cdot \ldots \cdot \alpha}_{\#\text{ cols/rows}} = \alpha^n$

$K_{\|\cdot\|}(A) = \|A\| \|A^{-1}\| = |\alpha| \cdot \frac{1}{|\alpha|} = 1$

3) let A be a non-singular nxn matrix.

a) Show that, in any norm, $K(A) = K(A^{-1})$

$K(A) = \|A\| \|A^{-1}\|$

$K(A^{-1}) = \|A^{-1}\| \|(A^{-1})^{-1}\|$

$= \|A^{-1}\| \|A\| = \|A\| \|A^{-1}\| = K(A)$

b) By rewriting $Ax = b$ as $A^{-1}b = x$, use the proof template we did in class to show the so-called companion inequality:

$$\frac{\|\delta b\|}{\|b\|} \leq K(A) \frac{\|\delta x\|}{\|x\|}$$

Rewrite $Ax = b$ as $A^{-1}b = x$.

Let us perturb x such that b will also perturb

$$A^{-1}(b + \delta b) = (x + \delta x)$$

$$A^{-1}b + A^{-1}\delta b = x + \delta x \qquad (A^{-1}b = x)$$

$$A^{-1}\delta b = \delta x$$

$$\text{or} \quad \delta b = A\delta x$$

Take a norm on both sides

$$\|\delta b\| = \|A\delta x\| \left(\leq \|A\| \|\delta x\| \right) \quad \ldots (1)$$

↳ we know that $\|AB\| \leq \|A\|\|B\|$

Now, consider $A^{-1}b = x$, take norm on both sides

$$\|x\| = \|A^{-1}b\| \leq \|A^{-1}\| \|b\|$$

$$\Rightarrow \|A^{-1}\|\|b\| \geq \|A^{-1}b\| = \|x\| \qquad \Rightarrow \|b\| \geq \frac{\|x\|}{\|A^{-1}\|}$$

$$\Rightarrow \frac{1}{\|b\|} \leq \frac{\|A^{-1}\|}{\|x\|} \quad \ldots\ldots\ldots\ldots (2)$$

Combine ① and ②:

$$\frac{\|\delta b\|}{\|b\|} \leq \frac{\|A\| \|\delta x\|}{\|b\|} \qquad \left(\frac{1}{\|b\|} \leq \frac{\|A^{-1}\|}{\|x\|} \right)$$

$$\Rightarrow \frac{\|\delta b\|}{\|b\|} \leq \frac{\|A\|\,\|\delta x\|}{\|b\|} \leq \|A\|\,\|\delta x\|\,\frac{\|A^{-1}\|}{\|x\|}$$

$$\leq \underbrace{\|A\|\,\|A^{-1}\|}_{K(A)}\,\frac{\|\delta x\|}{\|x\|}$$

Hence,

$$\frac{\|\delta b\|}{\|b\|} \leq K(A)\,\frac{\|\delta x\|}{\|x\|}$$

4) Let $A = \begin{bmatrix} 3 & 0 & 2 \\ 2 & 0 & -2 \\ 0 & 1 & 1 \end{bmatrix}$ and let $b = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$.

a) Calculate the norms $\|A\|_1$, $\|A\|_\infty$, $\|A\|_F$.

$$\|A\|_1 = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|_1} = \max_{j=1:n} \sum_{i=1}^{n} |a_{ij}| = \max\{5, 1, 5\} = 5$$

$$\|A\|_\infty = \max_{x \neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \max_{i=1\ldots n} \sum_{j=1}^{n} |a_{ij}| = \max\{5, 4, 2\} = 5$$

$$\|A\|_F = \sqrt{\sum_{i,j=1}^{n} a_{ij}} = \sqrt{3^2 + 0^2 + 2^2 + 2^2 + 0^2 + (-2)^2 + 0^2 + 1^2 + 1^2} = \sqrt{23}$$

b) Compute the condition numbers $K_1(A)$, $K_\infty(A)$, $K_F(A)$ with respect to the 1-norm, ∞-norm, Frobenius norm respectively.

$$A^{-1} = \left[\begin{array}{ccc|ccc} 3 & 0 & 2 & 1 & 0 & 0 \\ 2 & 0 & -2 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right] \frac{1}{3}R_1 \sim \left[\begin{array}{ccc|ccc} 1 & 0 & 2/3 & 1/3 & 0 & 0 \\ 2 & 0 & -2 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right] R_2 - 2R_1$$

$$\sim \left[\begin{array}{ccc|ccc} 1 & 0 & 2/3 & 1/3 & 0 & 0 \\ 0 & 0 & -10/3 & -2/3 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right] R_2 \leftrightarrow R_3 \sim \left[\begin{array}{ccc|ccc} 1 & 0 & 2/3 & 1/3 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & -10/3 & -2/3 & 1 & 0 \end{array}\right] -\frac{3}{10}\cdot R_3$$

$$\sim \left[\begin{array}{ccc|ccc} 1 & 0 & 2/3 & 1/3 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1/5 & -3/10 & 0 \end{array}\right] \begin{array}{l} R_2 - R_3 \\ R_1 - \frac{2}{3}R_3 \end{array} \sim \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1/5 & 1/5 & 0 \\ 0 & 1 & 0 & -1/5 & 3/10 & 1 \\ 0 & 0 & 1 & 1/5 & -3/10 & 0 \end{array}\right] \quad A^{-1} = \begin{bmatrix} 1/5 & 1/5 & 0 \\ -1/5 & 3/10 & 1 \\ 1/5 & -3/10 & 0 \end{bmatrix}$$

$$K_1(A) = \|A\|_1 \|A^{-1}\|_1$$

$$\|A\|_1 = \max\{\text{col sum}\} = \max\{1/5, 1/5, 1\} = 1$$

$$K_1(A) = 5 \cdot 1 = 5$$

$$K_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty$$

$$\|A^{-1}\|_\infty = \max\{\text{row sum}\} = \max\{2/5, 11/10, 1/10\} = 11/10$$

$$K_\infty(A) = 5 \cdot \frac{11}{10} = \frac{55}{10} = \frac{11}{2} = 5.5$$

$$K_F(A) = \|A\|_F \|A^{-1}\|_F$$

$$\|A^{-1}\|_F = \sqrt{(1/5)^2 + (1/5)^2 + 0^2 + (-1/5)^2 + (3/10)^2 + 1^2 + (1/5)^2 + (-3/10)^2 + 0^2} = \frac{\sqrt{134}}{10}$$

$$K_F(A) = \sqrt{23} \cdot \frac{\sqrt{134}}{10} = \frac{\sqrt{3082}}{10}$$

c) Let $\delta b = \begin{bmatrix} \epsilon \\ 0 \\ 0 \end{bmatrix}$. Consider the linear systems $Ax = b$,

$A\hat{x} = b + \delta b$ and let $\delta x = \hat{x} - x$ $(x + \delta x = \hat{x})$. Estimate the relative error from above and below through condition numbers, for norms $1, \infty$, and Frobenius. (see Q2 for lower bound)

From a previous theorem, we have

$$\frac{\|\delta x\|}{\|x\|} \leq K_{\|\cdot\|}(A) \frac{\|\delta b\|}{\|b\|}$$

$\frac{\|\delta x\|}{\|x\|}$: relative error in x; $K_{\|\cdot\|}(A)$: condition no. of A; $\frac{\|\delta b\|}{\|b\|}$: relative error in b

we have the system $Ax = b$,

$$\begin{bmatrix} 3 & 0 & 2 \\ 2 & 0 & -2 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \Rightarrow \begin{aligned} 3x_1 + 2x_3 &= 0 \\ 2x_1 - 2x_3 &= 1 \\ x_2 + x_3 &= 1 \end{aligned} \quad \Rightarrow 3x_1 = -2x_3 \quad x_1 = \frac{-2}{3}x_3$$

$$2\left(\frac{-2}{3}x_3\right) - 2x_3 = 1 \Rightarrow \frac{-4}{3}x_3 - \frac{6}{3}x_3 = 1 \Rightarrow \frac{-10}{3}x_3 = 1 \Rightarrow x_3 = \frac{-3}{10}$$

$$x_1 = \frac{-2}{3}\left(\frac{-3}{10}\right) = \frac{6}{30} = 1/5$$

$$x_2 + \left(\frac{-3}{10}\right) = 1 \Rightarrow x_2 = 1 + \frac{3}{10} = \frac{13}{10} \Rightarrow \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1/5 \\ 13/10 \\ -3/10 \end{bmatrix}$$

Now, let us perturb b such that $Ax = b \Rightarrow$

$$A\hat{x} = b + \delta b$$

$$\begin{bmatrix} 3 & 0 & 2 \\ 2 & 0 & -2 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \end{bmatrix} = \begin{bmatrix} \epsilon \\ 1 \\ 1 \end{bmatrix} \Rightarrow \begin{aligned} 3\hat{x}_1 + 2\hat{x}_3 &= \epsilon \\ 2\hat{x}_1 - 2\hat{x}_3 &= 1 \\ \hat{x}_2 + \hat{x}_3 &= 1 \end{aligned} \quad \Rightarrow 3\hat{x}_1 = \epsilon - 2\hat{x}_3 \quad x_1 = \frac{\epsilon}{3} - \frac{2}{3}\hat{x}_3$$

$$\Rightarrow 2\left(\frac{\epsilon}{3} - \frac{2}{3}\hat{x}_3\right) - 2\hat{x}_3 = 1 \Rightarrow \frac{2}{3}\epsilon - \frac{4}{3}\hat{x}_3 - \frac{6}{3}\hat{x}_3 = 1$$

$$\Rightarrow -\frac{10}{3}\hat{x}_3 = 1 - \frac{2}{3}\epsilon \quad \Rightarrow \quad \hat{x}_3 = -\frac{3}{10}\left(1 - \frac{2}{3}\epsilon\right) = \frac{-3}{10} + \frac{1}{5}\epsilon$$

$$\hat{x}_1 = \frac{\epsilon}{3} - \frac{2}{3}\left(\frac{-3}{10} + \frac{1}{5}\epsilon\right) = \frac{\epsilon}{3} + \frac{1}{5} - \frac{2}{15}\epsilon = \frac{1}{5} + \frac{1}{5}\epsilon$$

$$\hat{x}_2 = 1 - \left(\frac{-3}{10} + \frac{1}{5}\epsilon\right) = 1 + \frac{3}{10} - \frac{1}{5}\epsilon = \frac{13}{10} - \frac{1}{5}\epsilon$$

$$\begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \end{bmatrix} = \begin{bmatrix} 1/5 + 1/5\,\epsilon \\ 13/10 - 1/5\,\epsilon \\ -3/10 + 1/5\,\epsilon \end{bmatrix}$$

So, we have

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1/5 \\ 13/10 \\ -3/10 \end{bmatrix}, \quad \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \end{bmatrix} = \begin{bmatrix} x_1 + \partial x_1 \\ x_2 + \partial x_2 \\ x_3 + \partial x_3 \end{bmatrix} = \begin{bmatrix} 1/5 + 1/5\,\epsilon \\ 13/10 - 1/5\,\epsilon \\ -3/10 + 1/5\,\epsilon \end{bmatrix}.$$

Hence, $\partial x = \begin{bmatrix} \partial x_1 \\ \partial x_2 \\ \partial x_3 \end{bmatrix} = \begin{bmatrix} 1/5\,\epsilon \\ -1/5\,\epsilon \\ 1/5\,\epsilon \end{bmatrix}$.

1-norm:

$$\frac{\|\delta x\|_1}{\|x\|_1} \le \kappa_1(A) \frac{\|\delta b\|_1}{\|b\|_1}$$

$$\|\delta x\|_1 = \max\left\{\text{col sum of } \begin{bmatrix} 1/5\,\epsilon \\ -1/5\,\epsilon \\ 1/5\,\epsilon \end{bmatrix}\right\} = \max\{1/5\,\epsilon\} = 1/5\,\epsilon$$

$$\|x\|_1 = \max\left\{\text{col sum of } \begin{bmatrix} 1/5 \\ 13/10 \\ -3/10 \end{bmatrix}\right\} = \max\left\{\frac{6}{5}\right\} = 6/5$$

$$\|\delta b\|_1 = \max\left\{\text{col sum } \begin{bmatrix} \epsilon \\ 0 \\ 0 \end{bmatrix}\right\} = \max\{\epsilon\} = \epsilon$$

$$\|b\|_1 = \max\left\{\text{col sum } \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}\right\} = \max\{2\} = 2$$

$$\Rightarrow \frac{(1/5\,\epsilon)}{(6/5)} \le (5)\frac{\epsilon}{2} \quad \Rightarrow \quad \frac{\epsilon}{6} \le \frac{5\epsilon}{2}$$

relative error in x $\leq$ $K_1(A)$ relative error in b

∞-norm:

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \leq K_\infty(A) \frac{\|\delta b\|_\infty}{\|b\|_\infty}$$

$$\|\delta x\|_\infty = \max\{\text{row sum of } \begin{bmatrix} 1/5\,\epsilon \\ -1/5\,\epsilon \\ 1/5\,\epsilon \end{bmatrix}\} = \max\{1/5\,\epsilon, -1/5\,\epsilon, 1/5\,\epsilon\} = \frac{\epsilon}{5}$$

$$\|x\|_\infty = \max\{\text{row sum of } \begin{bmatrix} 1/5 \\ 13/10 \\ -3/10 \end{bmatrix}\} = \max\{1/5, 13/10, -3/10\} = 13/10$$

$$\|\delta b\|_\infty = \max\{\text{row sum } \begin{bmatrix} \epsilon \\ 0 \\ 0 \end{bmatrix}\} = \max\{\epsilon, 0, 0\} = \epsilon$$

$$\|b\|_\infty = \max\{\text{row sum } \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}\} = \max\{0, 1, 1\} = 1$$

$$\Rightarrow \frac{(\epsilon/5)}{(13/10)} \leq (11/2)\frac{(\epsilon)}{1} \Rightarrow \frac{2\epsilon}{13} \leq \frac{11\epsilon}{2}$$

relative error in x; $K_\infty(A)$; relative error in b

Frobenius norm:

$$\frac{\|\delta x\|_F}{\|x\|_F} \leq K_F(A) \frac{\|\delta b\|_F}{\|b\|_F}$$

$$\|\delta x\|_F = \sqrt{\left(\frac{\epsilon}{5}\right)^2 + \left(\frac{-\epsilon}{5}\right)^2 + \left(\frac{\epsilon}{5}\right)^2} = \sqrt{\frac{\epsilon^2}{25} + \frac{\epsilon^2}{25} + \frac{\epsilon^2}{25}} = \frac{\sqrt{3}\,\epsilon}{5}$$

$$\|x\|_F = \sqrt{(1/5)^2 + (13/10)^2 + (-3/10)^2} = \sqrt{\frac{1}{25} + \frac{169}{100} + \frac{9}{100}} = \frac{\sqrt{182}}{10}$$

$$\|\delta b\|_F = \sqrt{\epsilon^2 + 0^2 + 0^2} = \sqrt{\epsilon^2} = \epsilon$$

$$\|b\|_F = \sqrt{0^2 + 1^2 + 1^2} = \sqrt{2}$$

$$\Rightarrow \frac{\left(\frac{\sqrt{3}\,\epsilon}{5}\right)}{\left(\frac{\sqrt{182}}{10}\right)} \leq \left(\frac{\sqrt{3082}}{10}\right)\frac{(\epsilon)}{\sqrt{2}} \quad \Rightarrow \frac{\sqrt{546}\,\epsilon}{91} \leq \frac{\sqrt{1541}\,\epsilon}{10}$$

relative error in x ; $K_F(A)$; relative error in b

5) Prove the theorem: Let $\|\cdot\|$ denote a norm on $\mathbb{R}^m$, and also the norm on $\mathbb{R}^{m\times m}$ that it induces. Let $I \in \mathbb{R}^{m\times m}$ denote the identity matrix. Then

(a) $\|I\| = 1$

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{m\times m} \quad x = \begin{bmatrix} x_1 \\ \vdots \\ \vdots \\ x_m \end{bmatrix}_{m\times 1} \quad Ix = \begin{bmatrix} x_1 \\ \vdots \\ \vdots \\ x_1 \end{bmatrix}_{m\times 1}$$

(the result Ix is simply x)

$$\|I\| = \max_{x\neq 0} \frac{\|Ix\|}{\|x\|} = \frac{\|x\|}{\|x\|} = 1$$

(b) $\|Ax\| \leq \|A\|\,\|x\|$ for all $A \in \mathbb{R}^{m\times m}$ and $x \in \mathbb{R}^m$

We have

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$ (the definition of an induced norm).

We take the maximum value of $\frac{\|Ax\|}{\|x\|}$ in order to find $\|A\|$.

Hence, when we do not choose the maximum for $\frac{\|Ax\|}{\|x\|}$,

$$\frac{\|Ax\|}{\|x\|} \leq \|A\|.$$

When we multiply by $\|x\|$ on both sides, this implies

$$\|Ax\| \leq \|A\| \cdot \|x\|.$$

6) What is wrong with the following reasoning?

Changing b by 1% means multiplying it by 1.01. If $Ax = b$, then $A(1.01x) = 1.01b$, by linearity. So a 1% change of b, from b to $1.01b$, causes a 1% change in x, from x to $1.01x$. So forget the whole story about condition numbers - a 1% change in b always causes a 1% change in x.

There are a number of things wrong, starting with the fact that we cannot ignore condition numbers. A system with

a small condition number of matrix is well-conditioned, meaning there is a smaller margin between the error in x and the error in b. A system w/ large condition number of matrix is ill-conditioned, meaning the margin of error between x and b is amplified. This is shown by the relationship

$$\frac{\|\delta x\|}{\|x\|} \leq K_{\|\cdot\|}(A) \frac{\|\delta b\|}{\|b\|}$$

relative error of x — condition no. of A — relative error in b.

Hence, a 1% change of b does not necessarily cause a 1% change in x.

7)

Editor - /Users/sabella/q7.m

q7.m

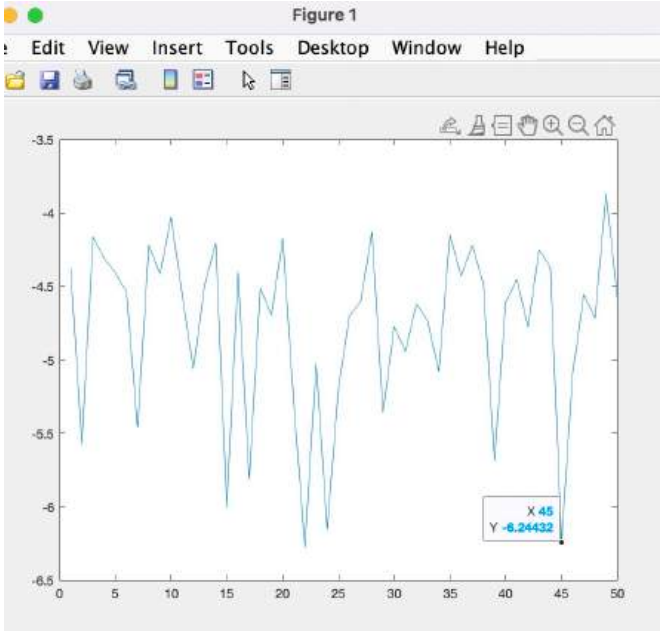
```
1  q = zeros(1, 50);
2  for i = 1:50
3      A = randn(500);
4      b = randn(500, 1);
5      x = A\b;
6      delta_b = 10^(-3)*randn(500, 1);
7      x_tilde = A\(b+delta_b);
8      rel_error_sol = norm(x_tilde-x, 1)/norm(x, 1);
9      rel_error_b = norm(delta_b, 1)/norm(b, 1);
10     q(i) = (rel_error_sol)/(cond(A, 1)*rel_error_b);
11 end
12 y = log10(q);
13 plot(y);
```

Command Window

```
>> clear
>> q7
>>
```

Workspace

Name	Value
A	500x500 double
b	500x1 double
delta_b	500x1 double
i	50
q	1x50 double
rel_error_b	9.9354e-04
rel_error_sol	8.5604e-04
x	500x1 double
x_tilde	500x1 double
y	1x50 double



$$\frac{\|\delta x\|}{\|x\|} \leq K_{\|\cdot\|}(A) \frac{\|\delta b\|}{\|b\|}$$

relative error of x | condition no. of A | relative error in b

$$q = \left(\frac{\|\delta x\|_1}{\|x\|_1}\right) \Big/ \left(K_1(A) * \frac{\|\delta b\|}{\|b\|}\right)$$

In the code, q represents the ratio between the error in x (solution) and the perturbation of b. As shown in Figure 1, no value of $q(i)$ is greater than or equal to 1. Hence, the relative error in x is smaller than the perturbation in b (scaled by the condition number of A with respect to the 1-norm). So, our theory that well-conditioned systems lead to smaller errors in x holds.

HW 1

SABELLA SAULINO

1) Let L_1, L_2, L_3 be the matrices below:

$$L_1 = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ l_{31} & 0 & 1 \end{bmatrix}, \quad L_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & l_{32} & 1 \end{bmatrix}.$$

Calculate the matrix product $L_1 L_2 L_3$ and show its equal to

$$L = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix}.$$

↳ $$L_1 L_2 L_3 = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ l_{31} & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & l_{32} & 1 \end{bmatrix}$$

$$L_1 L_2 = \begin{bmatrix} 1+0+0 & 0+0+0 & 0+0+0 \\ l_{21}+0+0 & 0+1+0 & 0+0+0 \\ 0+0+l_{31} & 0+0+0 & 0+0+1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & 0 & 1 \end{bmatrix}$$

$$L_1 L_2 L_3 = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & l_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1+0+0 & 0+0+0 & 0+0+0 \\ l_{21}+0+0 & 0+1+0 & 0+0+0 \\ l_{31}+0+0 & 0+0+l_{32} & 0+0+1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} = L = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix}$$

2) Does the matrix $A = \begin{bmatrix} 6 & 0 & 6 \\ -3 & 2 & -5 \\ -4 & 1 & -4 \end{bmatrix}$ have an LU factorization? If so, compute it. If not, explain.

↳ A is said to have LU factorization if there exists a lower triangular matrix L, upper triangular matrix U, such that

$$A = LU.$$

$$A = \begin{bmatrix} 6 & 0 & 6 \\ -3 & 2 & -5 \\ -4 & 1 & 4 \end{bmatrix} \begin{matrix} R_1 \\ R_2 \\ R_3 \end{matrix}$$

$R_2 + \frac{1}{2}R_1$, $R_3 + \frac{2}{3}R_1$

$$\begin{bmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ -2/3 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 6 & 0 & 6 \\ 0 & 2 & -2 \\ 0 & 1 & 0 \end{bmatrix}$$

$R_3 - \frac{1}{2}R_2$

$$\begin{bmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ -2/3 & 1/2 & 1 \end{bmatrix} \quad \begin{bmatrix} 6 & 0 & 6 \\ 0 & 2 & -2 \\ 0 & 0 & 1 \end{bmatrix}$$

L U

Yes, A has an LU factorization, as shown above.

3) Let $A = \begin{bmatrix} 0 & 1 & 1 & 2 \\ 2 & 0 & 0 & 0 \\ 1 & 1 & 1 & 3 \\ 1 & 0 & 2 & 4 \end{bmatrix}$. Does A have an LU factorization? If yes, compute. If no, explain why not, and then compute the PLU decomposition of A.

$A = \begin{bmatrix} 0 & 1 & 1 & 2 \\ 2 & 0 & 0 & 0 \\ 1 & 1 & 1 & 3 \\ 1 & 0 & 2 & 4 \end{bmatrix}$ The first element in row 1 is zero.

Hence we cannot use Gaussian Elimination, and so A does not have an LU factorization. We will compute the PLU decomposition of A.

$$\underbrace{\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{P_1} \underbrace{\begin{bmatrix} 0 & 1 & 1 & 2 \\ 2 & 0 & 0 & 0 \\ 1 & 1 & 1 & 3 \\ 1 & 0 & 2 & 4 \end{bmatrix}}_{A} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 2 \\ 1 & 1 & 1 & 3 \\ 1 & 0 & 2 & 4 \end{bmatrix} \begin{matrix} R_3 - \frac{1}{2}R_1 \to R_3 \\ R_4 - \frac{1}{2}R_1 \to R_4 \\ L_{13} = 1/2 \\ L_{14} = 1/2 \end{matrix}$$

$$\sim \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 2 \\ 0 & 1 & 1 & 3 \\ 0 & 0 & 2 & 4 \end{bmatrix} \begin{matrix} R_3 - R_2 \to R_3 \\ L_{23} = 1 \end{matrix} \sim \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 2 & 4 \end{bmatrix} \begin{matrix} \nearrow \\ \text{swap} \\ \swarrow \end{matrix}$$

(Also swap L_{23} and L_{24})

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{P_2} A = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 2 & 4 \\ 0 & 0 & 0 & 1 \end{bmatrix} = U$$

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \boxed{1/2} & 0 & 1 & 0 \\ \boxed{1/2} & \boxed{1} & 0 & 1 \end{bmatrix}$$

Hence, the PLU decomposition of A is:

$$\underbrace{\begin{bmatrix} 0 & 1 & 1 & 2 \\ 2 & 0 & 0 & 0 \\ 1 & 1 & 1 & 3 \\ 1 & 0 & 2 & 4 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{P} \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/2 & 0 & 1 & 0 \\ 1/2 & 1 & 0 & 1 \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 2 & 4 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{U}$$

4) Determine whether or not the following matrices have a Cholesky factorization; if they do, compute (by hand) the Cholesky factor R:

$$A = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 13 & 6 \\ 0 & 6 & 5 \end{bmatrix}, \quad B = \begin{bmatrix} 4 & -4 & 0 \\ -4 & 4 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

In order to have a Cholesky factorization, a matrix A must be positive definite, that is, $A = A^T$ and $\forall x \neq 0$, $x^T A x > 0$. If λ is an eigenvalue of A with a corresponding eigenvector v ($v \neq 0$), $Av = \lambda v$. Hence,

$$v^T A v = v^T \lambda v = \lambda (v^T v) > 0,$$

since $\lambda > 0$ and $v^T v > 0$.

A is positive definite, since $A = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 13 & 6 \\ 0 & 6 & 5 \end{bmatrix} = A^T$,

$\det(A) = 1(65 - 36) + 2(-10 - 0) = 9 > 0$, and all pivots/eigenvalues are positive. Hence we proceed with

Cholesky factorization.

Define $R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{bmatrix}$. $A = R^T R$, so

$$\underset{A}{\begin{bmatrix} 1 & -2 & 0 \\ -2 & 13 & 6 \\ 0 & 6 & 5 \end{bmatrix}} = \underset{R^T}{\begin{bmatrix} r_{11} & 0 & 0 \\ r_{12} & r_{22} & 0 \\ r_{13} & r_{23} & r_{33} \end{bmatrix}} \underset{R}{\begin{bmatrix} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{bmatrix}}$$

$$\begin{bmatrix} 1 & -2 & 0 \\ -2 & 13 & 6 \\ 0 & 6 & 5 \end{bmatrix} = \begin{bmatrix} r_{11}^2 & r_{11}r_{12} & r_{11}r_{13} \\ r_{12}r_{11} & r_{22}^2 + r_{12}^2 & r_{12}r_{13} + r_{22}r_{23} \\ r_{11}r_{13} & r_{12}r_{13} + r_{22}r_{23} & r_{13}^2 + r_{23}^2 + r_{33}^2 \end{bmatrix}$$

Row 1

$r_{11}^2 = 1 \Rightarrow r_1 = 1$

$r_{11}r_{12} = -2 \Rightarrow r_{12} = -2$

$r_{11}r_{13} = 0 \Rightarrow r_{13} = 0$

Row 2

$r_{12}^2 + r_{22}^2 = 13 \Rightarrow (r_{12})^2 = 4 \Rightarrow r_{22}^2 = 13 - 4 = 9$

$\Rightarrow r_{22} = 3$

$r_{12}r_{13} + r_{22}r_{23} = 6 \Rightarrow (-2)(0) + (3)(r_{23}) = 6$

$\Rightarrow r_{23} = 2$

Row 3 $r_{13}^2 + r_{23}^2 + r_{33}^2 = 5 \Rightarrow (0) + (2)^2 + r_{33}^2 = 5$

$\Rightarrow r_{33} = 1$

Hence, $R = \begin{bmatrix} 1 & -2 & 0 \\ 0 & 3 & 2 \\ 0 & 0 & 1 \end{bmatrix}$.

Now, although $B = B^T$, B is not positive definite, since $\det(B) = 5(16-16) = 5(0) = 0 \not> 0$. Hence, B does not have a Cholesky factorization.

5) Let A be an $n \times n$ positive definite and symmetric matrix, and X an $n \times n$ invertible matrix. Show that $B = X^T A X$ is positive definite and symmetric.

A is positive definite, so $A = A^T$ and $x^T A x > 0$.

X is invertible, so X^{-1} exists.

Let $B = X^T A X$. Then

$$(B)^T = (X^T A X)^T$$

$$\Rightarrow \quad B^T = (AX)^T (X^T)^T$$

$$\Rightarrow \quad B^T = A^T X^T X$$

$$\Rightarrow \quad B^T = X^T A^T X \qquad (A = A^T)$$

$$\Rightarrow \quad B^T = X^T A X = B .$$

Hence, B is symmetric.

Now, let $y \in \mathbb{R}^n$ be a vector in $\mathbb{R}^n$, $y \neq 0$. Then

$$\begin{aligned} y^T B y &= y^T (X^T A X) y \\ &= y^T X^T A X y \\ &= (Xy)^T A (Xy) \\ &= \langle Xy , Xy \rangle \end{aligned}$$

Let $z = Xy$, since X is invertible and $Xy = z$ always has a solution. Then

$$\langle z , z \rangle$$

$$= \sum_{i=1}^{n} (z)^2 \qquad > 0 ,$$

since n^2 is always positive.

Hence, B is positive definite and symmetric.

6) Consider the system $Ax = b$, where A is upper triangular. The system above can be solved by backward

substitution, starting with solving for x_n, substituting into the second equation to get x_{n-1}, etc.

(a) Find the formula to compute x_i given $x_{i+1}, \dots, x_n$.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \ddots & & \vdots \\ 0 & 0 & a_{ii} & a_{in} \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{bmatrix} \quad b = \begin{bmatrix} b_1 \\ \vdots \\ b_i \\ \vdots \\ b_n \end{bmatrix}$$

$$a_{nn} \cdot x_n = b_n \Rightarrow x_n = \frac{b_n}{a_{nn}}$$

$$a_{n-1\,n-1} x_{n-1} + a_{n-1\,n} x_n = b_{n-1} \Rightarrow x_{n-1} = \frac{b_{n-1} - a_{n-1\,n} x_n}{a_{n-1\,n-1}}$$

$$\vdots$$

$$a_{ii} \cdot x_i + a_{i(i+1)} x_{i+1} + \dots + a_{in} x_n = b_i$$

$$\Rightarrow x_i = \frac{b_i - \sum_{j=i+1}^{n} a_{ij} x_j}{a_{ii}}$$

(b) Using the model code given for forward substitution, write a MATLAB code that performs backward substitution.

```
function x = backsub(A, b)
    n = length(b);
    x = b;
    for i = n:-1:1
        for j = i+1:n
            x(i) = x(i) - A(i,j) * x(j);
        end
        x(i) = x(i) / A(i,i);
    end
end
```

7) Write a MATLAB code that does Gauss Elimination with partial pivoting (row switching). The input should be

- an $n \times n$ matrix A
- a column vector b of size $n \times 1$

and the output should be the solution to $Ax = b$.

```
function x = Gauss_PartialPivot(A,b)
    n = size(A,1);   % get num rows of A
    Ab = [A, b];
    x = zeros(n, 1);
    for j = 1 : n-1
        [~, maxRow] = max(abs(Ab(j:n, j)));
        maxRow = maxRow + j - 1;
        if maxRow ~= j
            Ab([j, maxRow], :) = Ab([maxRow, j], :);
        end
        for i = j+1 : n
            m = Ab(i,j) / Ab(j,j);
            Ab(i, j:n+1) = Ab(i, j:n+1) - m * Ab(j, j:n+1);
        end
    end
    x(n) = Ab(n, n+1) / Ab(n,n);
```

```
    for i = n-1 : -1 : 1
        x(i) = (Ab(i, n+1) - Ab(i, i+1:n) * x(i+1:n)) / Ab(i, i);
    end
end
```

8) Write a MATLAB function with inputs

- a matrix A of size $m \times n$,
- a matrix B of size $n \times p$,

and which outputs the product matrix $A \times B$ (note that the size of this matrix should be $m \times p$). Your code should check that the sizes are right and then do the multiplication using three nested "for" loops.

```
function X = matrix_Product(A, B);
    [m, n] = size(A);
    [nB, p] = size(B);
    if n ~= nB
        error('incorrect sizes');
    end
    X = zeros(m, p);
    for i = 1 : m
        for j = 1 : p
```

$$\sum_{j=1}^{p} 2n = p^2 + p$$

$$\sum_{i=1}^{m} p^2 + p = \sum_{i=1}^{m} p^2 + \sum_{i=1}^{m} p$$

```
        for k = 1:n
            X(i,j) = X(i,j) + A(i,K) * B(K,j);
        end
    end
end
```

n

2 operations, n times

end

(a) Yes, $A \times B$ and X = matrix_Product(A, B) output the same result (I would include, but it's a lot).

(b) How many Flops does the code use? Find a formula in terms of m, n, p.

inner loop: 2 operations $(+, *)$ n times $= 2n$

$$\sum_{k=1}^{n} 2 = 2n$$

middle loop: iterates p times $\sum_{j=1}^{p} 2n = p^2 + p$

outer loop: iterates m times $\sum_{i=1}^{m} p^2 + p =$

$$\sum_{i=1}^{m} p^2 + \sum_{i=1}^{m} p = \frac{m(m+1)(2m+1)}{6} + \frac{m(m+1)}{2} = \frac{1}{3}m^3 + m^2 + \frac{2}{3}m$$

$$\sum_{i=1}^{m} \left(\sum_{j=1}^{p} \left(\sum_{k=1}^{n} 2 \right) \right) = \frac{1}{3}m^3 + m^2 + \frac{2}{3}m$$

9) Use for and if loops to write two MATLAB functions that take as an input

- $n \times n$ matrix A
- $n \times n$ matrix B
- $n \times 1$ vector x.

Check that the inputs are square matrices of the same size. Then have the first function compute ABx through (AB)x and the second through A(Bx). Then

(a) Take a screenshot of your first function

Editor - /Users/sabella/compute_ABx.m

compute_ABx.m

```
1  function first = compute_ABx(A, B, x)
2      [mA, nA]= size(A);
3      [mB, nB]= size(B);
4      if mA ~= nA || mB ~= nB || mA ~= mB
5          error('matrices must be square and same size');
6      end
7      AB = zeros(mA, nB);
8      for i = 1 : mA
9          for j = 1 : nB
10             for k = 1 : nA
11                 AB(i, j) = AB(i, j) + A(i, k)*B(k, j);
12             end
13         end
14     end
15     first = zeros(mA, 1);
16     for i = 1 : mA
17         for j = 1 : nB
18             first(i) = first(i) + AB(i, j)*x(j);
19         end
20     end
21 end
```

2 op *n] *n *n

$O(n^3)$

2op *n] *n

$O(n^2)$

(b) Take a screenshot of your second function

Editor - /Users/sabella/compute_A_Bx.m

compute_ABx.m | compute_A_Bx.m

```
1  function second = compute_A_Bx(A, B, x)
2      [mA, nA]= size(A);
3      [mB, nB]= size(B);
4      if mA ~= nA || mB ~= nB || mA ~= mB
5          error('matrices must be square and same size');
6      end
7      Bx = zeros(mB, 1);
8      for i = 1 : mB
9          for j = 1 : nB
10             Bx(i) = Bx(i)+ B(i, j)*x(j);
11         end
12     end
13     second = zeros(mA, 1);
14     for i = 1 : mA
15         for j = 1 : nA
16             second(i) = second(i) + A(i, j)*Bx(j);
17         end
18     end
19 end
20
21
```

2 operations * n] * n = $O(n^2)$

2 op * n] * n = $O(n^2)$

(c) Calculate the number of flops for both approaches, theoretically (in terms of n) and compare: which one is better to use?

Function 1: $\sum_{i=1}^{n} \left(\sum_{j=1}^{n} 2n \right)$ (2 operations inside j-loop - the two operations iterate n times)

$= \sum_{i=1}^{n} n^2 + n \qquad = \sum_{i=1}^{n} n^2 + \sum_{i=1}^{n} n$

$= \frac{n(n+1)(2n+1)}{6} + \frac{n(n+1)}{2} = \frac{1}{6}(2n^3+3n^2+n) + \frac{1}{2}(n^2+n)$

$= \frac{1}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{6}n + \frac{1}{2}n^2 + \frac{1}{2}n \qquad = \frac{1}{3}n^3 + n^2 + \frac{2}{3}n$

$= O(n^3)$

(second loop) $\sum_{i=1}^{n} 2i = n^2 + n = O(n^2)$

Total: $\frac{1}{3}n^3 + n^2 + \frac{2}{3}n + n^2 + n = \frac{1}{3}n^3 + 2n^2 + \frac{5}{3}n = O(n^3)$

Function 2: $\sum_{i=1}^{n} 2i = n^2 + n = O(n^2)$

(second loop) $\sum_{i=1}^{n} 2i = n^2 + n = O(n^2)$

Total: $2n^2 + 2n = O(n^2)$

The second function is better to use since it is faster and takes less operations.