

Homework 1

Caitlin Tang 505683651

Due: Monday 1/20/2025

I. Linear Algebra Refresher (25 points)

(a) (12 points) Let Q be a real orthogonal matrix.

- Show that Q^T and Q^{-1} are also orthogonal.
- Show that Q has eigenvalues with norm 1.
- Show that the determinant of Q is either +1 or -1.
- Show that Q defines a length preserving transformation.

Q is defined as a real orthogonal matrix \rightarrow a $n \times n$ matrix A with real values and the transpose A^T is equal to the inverse (A^{-1})

\therefore if $A^T = A^{-1}$ then $A^T A = I$ (identity matrix)

Part (a) (i) Show that Q^T and Q^{-1} are also orthogonal

Matrix Q is orthogonal if $Q^T Q = I$

$$\begin{aligned} &\text{Assume } Q^T Q = I \quad \text{take transpose} \\ &(Q^T Q)^T = I^T \\ &(Q^T)(Q^T)^T = I \quad \text{use } (AB)^T = B^T A \text{ and } I^T = I \\ &Q^T Q = I \quad (Q^T)^T = Q \end{aligned}$$

Showing
 Q^T is
orthogonal

Alternatively, given property $A^T A = I$, set $A = Q^T$

$$\begin{aligned} &A^T A = I \quad \text{substitute } A = Q^T \\ &(Q^T)^T Q^T = I \\ &Q Q^T = I \quad \text{take the transpose} \\ &(Q Q^T)^T = I^T \quad \text{use the property } (AB)^T = B^T A^T \text{ and } I^T = I \\ &Q^T Q = I \end{aligned}$$

Therefore, Q^T satisfies the condition $A^T A = I$ and is orthogonal

We know that $Q^T = Q^{-1}$ and $Q Q^{-1} = I$

Showing
 Q^{-1} is
orthogonal

$$\begin{aligned} &A^T A = I \quad \text{substitute } A = Q^T \\ &(Q^T)^T Q^T = I \\ &Q Q^T = I \quad \text{substitute } Q^T = Q^{-1} \\ &(Q^{-1})(Q^{-1})^T = I \\ &[(Q^{-1})(Q^{-1})^T] = I^T \quad \text{Apply the } (AB)^T = B^T A^T \text{ property and } I^T = I \\ &(Q^{-1})^T (Q^{-1}) = I \end{aligned}$$

Since Q^{-1} is as $(Q^{-1})^T (Q^{-1}) = I$ which is of the form $A^T A = I$ then Q^{-1} is also orthogonal

Therefore Q^T is orthogonal and Q^{-1} is orthogonal

Part (a) (ii) Show that Q has eigenvalues with norm 1

Review: The **norm** of a matrix is a real number which is a measure of the magnitude of the matrix. The following properties hold true for defining the norm of a square matrix A (norm is a non-negative real number denoted $\|A\|$)

1. $\|A\| \geq 0$ for any square matrix A
2. $\|A\| = 0$ iff matrix $A = 0$
3. $\|kA\| = |k| \|A\|$, for any scalar k
4. $\|A + B\| \leq \|A\| + \|B\|$
5. $\|AB\| \leq \|A\| \|B\|$

For the real orthogonal matrix Q ...

if λ is an eigenvalue of Q with its corresponding vector \vec{x} then

$$Q\vec{x} = \lambda\vec{x} \quad \text{where } \vec{x} \neq 0$$

since Q is orthogonal $Q^T Q = I$ and $Q^T = Q^{-1}$

The **Euclidean norm** of $\vec{x} \in \mathbb{R}^n$ is defined as $\|\vec{x}\|_2 = \sqrt{\vec{x}^T \vec{x}}$

When our orthogonal matrix Q is added it transforms the vector to become $Q\vec{x}$

The Euclidean norm of $Q\vec{x}$ is thus $\|Q\vec{x}\| = \sqrt{(Q\vec{x})^T (Q\vec{x})}$

$$\begin{aligned} \|Q\vec{x}\| &= \sqrt{(\vec{x}^T)(Q^T)(Q)(\vec{x})} \\ Q^T Q &= I \quad \leftarrow \\ \|Q\vec{x}\| &= \sqrt{(\vec{x})^T I (\vec{x})} \\ \|Q\vec{x}\| &= \sqrt{\vec{x}^T \vec{x}} = \|\vec{x}\| \end{aligned}$$

$$(Q\vec{x})^T = \vec{x}^T Q^T$$

from the property
 $(AB)^T = B^T A^T$

$$\|Q\vec{x}\| = \|\lambda\vec{x}\|$$

we know that $\|Q\vec{x}\| = \|\vec{x}\| \quad \therefore \|\lambda\vec{x}\| = \|\vec{x}\| \quad \rightarrow$ separate based on the scalar property
 $|\lambda| \|\vec{x}\| = \|\vec{x}\| \quad \rightarrow \quad |\lambda| \|\vec{x}\| = \|k\vec{x}\| = |k| \|\vec{x}\| \text{ for some scalar } k$

since $\vec{x} \neq 0$, $\|\vec{x}\| \neq 0$ and $|\lambda| = 1$. Therefore, the eigenvalues of Q will have a magnitude of 1

NOTE: Eigenvalues are scalars so we can alternatively prove it using the property $\|kA\| = |k| \|A\|$

Assume we have a vector \vec{x}

$$(\lambda\vec{x})^T (\lambda\vec{x}) = (Q\vec{x})^T (Q\vec{x}) \quad \rightarrow \text{Apply the property } (AB)^T = B^T A^T$$

$$\vec{x}^T \lambda^T \lambda \vec{x} = \vec{x}^T Q^T Q \vec{x}$$

$$\lambda^2 \vec{x}^T \lambda \vec{x} = \vec{x}^T I \vec{x}$$

move the λ^2 because it's the scalar

$$\lambda^2 \vec{x}^T \vec{x} = \vec{x}^T I \vec{x}$$

$$\lambda^2 = 1$$

$$\lambda = \pm 1$$

↳ this proves eigen value = ± 1 and norm of $|\pm 1| = 1$

magnitude (absolute value)

(a)(iii) Show that the determinant of Q is either +1 or -1

Review: Properties of Determinants

1. Determinant of Identity Matrix $\det(I) = 1$
2. Determinant of Transposed Matrix A $\det(A^T) = \det(A)$
3. Determinant of Product of Two Matrices $\det(AB) = \det(A)\det(B)$
4. Determinant of Inverse: $\det(A^{-1}) = \frac{1}{\det(A)}$

5. If you exchange 2 rows of a matrix, you reverse the sign of the determinant

6. If you multiply a row or column by a scalar k , determinant also scales!

NOTE: determinant of a matrix is product of its eigenvalues $\det(A) = \lambda_1 \cdot \lambda_2 \cdots \lambda_n$

We want to show $\det(Q) \in \{+1, -1\}$

For an orthogonal matrix Q , we know that $Q^T Q = I$

$$\begin{aligned} \det(Q^T Q) &= \det(I) \quad \text{use property } \det(I) = 1 \\ \det(Q^T Q) &= 1 \\ \det(Q^T) \det(Q) &= 1 \quad \text{use property } \det(AB) = \det(A)\det(B) \\ \det(Q) \cdot \det(Q) &= 1 \quad \text{use property } \det(Q^T) = \det(Q) \end{aligned}$$

$$\begin{aligned} \sqrt{(\det(Q))^2} &= 1 \quad \text{take the square root} \\ \det(Q) &= \pm 1 \end{aligned}$$

Therefore, for the real, orthogonal matrix Q has a determinant of 1 or -1

(a)(iv) Show that Q defines a length preserving transformation \rightarrow show that $\|Qv\| = \|v\|$

For Q , we know that

$$Q^T Q = I$$

For any vector \vec{v} , the transformation $Q\vec{v}$ should preserve the Euclidean length of \vec{v}

The Euclidean norm of v is $\|\vec{v}\| = \sqrt{\vec{v}^T \vec{v}}$

The Euclidean norm of transformed vector $Q\vec{v}$ is $\|Q\vec{v}\| = \sqrt{(Q\vec{v})^T (Q\vec{v})}$

$$\begin{aligned} \|Q\vec{v}\| &= \sqrt{(Q\vec{v})^T (Q\vec{v})} \\ \|Q\vec{v}\| &= \sqrt{\vec{v}^T Q^T Q \vec{v}} \\ \|Q\vec{v}\| &= \sqrt{\vec{v}^T I \vec{v}} \\ \|Q\vec{v}\| &= \sqrt{\vec{v}^T \vec{v}} = \|\vec{v}\| \end{aligned}$$

} Apply the property $(AB)^T = B^T A^T$

$$Q^T Q = I$$

Therefore the transformed vector $Q\vec{v}$ has a Euclidean norm $\|Q\vec{v}\|$ equal to $\|\vec{v}\|$

Problem #1 Part b

(b) (8 points) Let A be a matrix.

- (4 points) What is the relationship between the singular vectors of A and the eigenvectors of AA^T ? What about A^TA ?
- (4 points) What is the relationship between the singular values of A and the eigenvalues of AA^T ? What about A^TA ?

(b)(i) What is the relationship between the singular vectors of A and the eigenvectors of AA^T ? What about A^TA ?

For this problem, we assume that A is a $m \times n$ dimensional matrix

- AA^T will be a square symmetric matrix size $m \times m$
- A^TA will be a square symmetric matrix size $n \times n$

Singular Value Decomposition (SVD) Review

The SVD of $m \times n$ matrix A is $A = U\Sigma V^T$ $\rightarrow U = m \times m$ orthogonal matrix of left singular vectors of A
 $\rightarrow V = n \times n$ orthogonal matrix of right singular vectors of A
 $\rightarrow \Sigma = m \times n$ diagonal matrix of singular values of A ($\sigma_1, \dots, \sigma_m$)

Visual example $m=3, n=4$

$$3 \begin{array}{|c|} \hline \end{array} = 3 \begin{array}{|c|c|c|} \hline & & 3 \\ \hline \end{array} \cdot 3 \begin{array}{|c|c|c|c|} \hline & & & 4 \\ \hline 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & & \\ \vdots & & \ddots & \\ 0 & \cdots & \sigma_3 & 0 \\ \hline \end{array} \cdot 4 \begin{array}{|c|c|c|c|} \hline & & & 4 \\ \hline \end{array} \rightarrow \begin{array}{l} \text{right singular} \\ \text{vectors} \\ \text{rows of } n \times n \text{ matrix } V \end{array}$$

matrix A left singular vectors
 columns of the $m \times n$ matrix Σ $m \times m$ matrix U

Solving for AA^T

$$AA^T = (U\Sigma V^T)(U\Sigma V^T)^T = U\Sigma V^T \underbrace{V\Sigma^T}_{I_n} U^T = U\Sigma I_n \Sigma^T U^T = U\Sigma \Sigma^T U^T = U\Sigma^2 U^T$$

NOTE: $\Sigma \Sigma^T \neq \Sigma^T \Sigma$ $\Rightarrow \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & \sigma_m^2 \end{pmatrix} = \Sigma^2$
 $(m \times n)(n \times m)(n \times m)(m \times n)$

Therefore, AA^T has eigenvectors given by the singular left vectors of A (columns of U) and its corresponding eigenvalues are the squares of the singular values of A ($\sigma_1^2, \dots, \sigma_m^2$)

Solving for A^TA

$$A^TA = (U\Sigma V^T)^T (U\Sigma V^T) = V\Sigma^T \underbrace{U^T U \Sigma}_{{I_m}} V^T = V\Sigma^T I_m \Sigma V^T = V\Sigma^T \Sigma V^T = V\Sigma^2 V^T$$

NOTE: $\Sigma \Sigma^T = \Sigma^T \Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & \sigma_m^2 \end{pmatrix} = \Sigma^2$ (n \times n matrix)

Therefore A^TA has eigenvectors given by the right singular vectors of A (rows of V)

Summary

- AA^T and A^TA share the same nonzero eigenvalues ($\sigma_1^2, \dots, \sigma_m^2$)
- The eigenvectors of AA^T are the left singular vectors of A
- The eigenvectors of A^TA are the right singular vectors of A

(b) (ii) What is the relationship between the singular values of A and the eigenvalues of AA^T ? What about A^TA ?

Again using the Singular Value Decomposition (SVD)

where $A = U\bar{Z}V^T \rightarrow U = m \times m$ orthogonal matrix of left singular vectors of A

$\rightarrow V = n \times n$ orthogonal matrix of right singular vectors of A

$\rightarrow \bar{Z} = m \times n$ diagonal matrix of singular values of A ($\sigma_1, \dots, \sigma_m$)

Solving AA^T

$$\begin{aligned} AA^T &= (U\bar{Z}V^T)(U\bar{Z}V^T)^T \\ &= U\bar{Z}V^TV\bar{Z}^TU^T \\ &= U\bar{Z}I_m\bar{Z}^TU^T \\ &= U\bar{Z}^2U \end{aligned}$$

Solving A^TA

$$\begin{aligned} A^TA &= (U\bar{Z}V^T)^T(U\bar{Z}V^T) \\ &= V\bar{Z}^TU^TU\bar{Z}V^T \\ &= V\bar{Z}^T I_n \bar{Z}V^T \\ &= V\bar{Z}^2V^T \end{aligned}$$

in both cases, \bar{Z}^2 is a diagonal matrix with σ_i^2 on the diagonal

\rightarrow with AA^T , $\bar{Z}\bar{Z}^T$ is a $m \times m$ matrix with σ_i^2 on the diagonal
 \rightarrow with A^TA , $\bar{Z}^T\bar{Z}$ is a $n \times n$ matrix with σ_i^2 on the diagonal

The eigenvalues of both AA^T and A^TA are the squares of the singular values of A .

Therefore, the singular values of A ($\sigma_1, \dots, \sigma_m$) are the square roots of the eigenvalues of AA^T and A^TA

Problem #1 Part (c) TRUE / FALSE

(c) (5 points) True or False. Partial credit on an incorrect solution may be awarded if you justify your answer.

- i. Every linear operator in an n -dimensional vector space has n distinct eigenvalues.
- ii. A non-zero sum of two eigenvectors of a matrix A is an eigenvector.
- iii. If a matrix A has the positive semidefinite property, i.e., $x^T Ax \geq 0$ for all x , then its eigenvalues must be non-negative.
- iv. The rank of a matrix can exceed the number of distinct non-zero eigenvalues.
- v. A non-zero sum of two eigenvectors of a matrix A corresponding to the same eigenvalue λ is always an eigenvector.

(c)(i) Every linear operator in an n -dimensional vector space has n distinct eigenvalues

Answer: FALSE

Counterexample: The identity matrix I_n is a $n \times n$ matrix with n eigenvalues $I = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & & \dots \\ \vdots & & \ddots & \\ 0 & \dots & \dots & 1 \end{pmatrix}$ of 1 which are not distinct (aka only 1 distinct eigenvalue)

Counterexample 2: eigenvalues can have multiplicities so if there is an eigenvalue with $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ m multiplicities (or multiple eigenvalues have multiplicities) then the total number of distinct eigenvalues will be less than n

Therefore, not every linear operator has n distinct eigenvalues

(c)(ii) A non-zero sum of two eigenvectors of a matrix A is an eigenvector

Answer: FALSE

Any sum of eigenvectors is an eigenvector if they share the same eigenvalue

Let \vec{x}_1 and \vec{x}_2 be eigenvectors of A corresponding to two different eigenvalues λ_1 and λ_2

$$A\vec{x}_1 = \lambda_1 \vec{x}_1 ; A\vec{x}_2 = \lambda_2 \vec{x}_2$$

Considering the sum $\vec{x}_1 + \vec{x}_2$. Applying the matrix A to the sum

$$A(\vec{x}_1 + \vec{x}_2) = A\vec{x}_1 + A\vec{x}_2 = \lambda_1 \vec{x}_1 + \lambda_2 \vec{x}_2 \quad \text{unless } \lambda_1 = \lambda_2, \text{ it is not a scalar multiple and therefore, } \vec{x}_1 + \vec{x}_2 \text{ are generally not an eigenvector}$$

if the eigenvalue is shared $\lambda = \lambda_1 = \lambda_2$ then the sum will be an eigenvector

$$A(\vec{x}_1 + \vec{x}_2) = \lambda \vec{x}_1 + \lambda \vec{x}_2 = \lambda(\vec{x}_1 + \vec{x}_2)$$

Therefore, the statement is not generally true and only applies when the two eigenvectors share the same eigenvalue

(c)(iii) If a matrix A has the positive semidefinite property (i.e. $\vec{x}^T A \vec{x} \geq 0$ for all \vec{x} , then its eigenvalues must be non-negative

Answer: True

For any eigenvector \vec{x} of matrix A that corresponds to an eigenvalue λ

$$\vec{x}^T A \vec{x} = \lambda \underbrace{\|\vec{x}\|^2}_{> 0 \text{ because } \vec{x} \text{ is non-zero}} \text{ which means that } \lambda \geq 0$$

Therefore, all eigenvalues of A are non-negative when A is positive semidefinite

(c)(iv) The rank of a matrix can exceed the number of distinct non-zero eigenvalues

Answer: True

Since eigenvalues can have multiplicities, the number of distinct non-zero eigenvalues

Example

$$D = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} \text{ we have one distinct eigenvalue of 2 when the rank of the matrix is 3 } \therefore \# \text{ eigenvalues} < \text{rank}$$

1 < 3

(c)(v) A non-zero sum of two eigenvectors of a matrix A corresponding to the same eigenvalue λ is always an eigenvector

Answer: True

if \vec{x}_1 and \vec{x}_2 are eigenvectors of A corresponding to the same eigenvalue λ

$$A\vec{x}_1 = \lambda \vec{x}_1 \text{ and } A\vec{x}_2 = \lambda \vec{x}_2$$

$$\vec{x}_1 + \vec{x}_2 \text{ and apply } A \rightarrow A(\vec{x}_1 + \vec{x}_2) = A\vec{x}_1 + A\vec{x}_2 = \lambda \vec{x}_1 + \lambda \vec{x}_2 = \lambda(\vec{x}_1 + \vec{x}_2)$$

This means it is a scalar multiple and $\vec{x}_1 + \vec{x}_2$ is also an eigenvector corresponding to λ

Problem #2: Probability Refresher (25 Points)

(a) (10 points) A and B are involved in a duel. The rules of the duel are that they are to pick up their guns and shoot at each other simultaneously. If one or both are hit, then the duel is over. If both shots miss, then they repeat the process. Suppose that the results of the shots are independent and that each shot of A will hit B with probability p_A and each shot of B will hit A with probability p_B . What is:

- (2 points) the probability that A is not hit?
- (2 points) the probability that both duelists are hit?
- (2 points) the probability that the duel ends after the n^{th} round of shots?
- (2 points) the conditional probability that the duel ends after the n^{th} round of shots given that A is not hit?
- (2 points) the conditional probability that the duel ends after the n^{th} round of shots given that both duelists are hit?

(a)(i) Probability A is not hit by B $\rightarrow P(A \text{ not hit by } B) = 1 - p_B$

(a)(ii) Probability that both duelists are hit?

The shots are independent and therefore $P(\text{both are hit}) = p_A \cdot p_B$
↳ mult

(a)(iii) Probability that the duel ends after the n^{th} round of shots?

In order for the duel to end in the n^{th} round, both duelists have to miss for all $n-1$ rounds prior to the n^{th} round

Probability that both duelists miss in a single round: $(1-p_A)(1-p_B)$

Probability that both duelists miss $n-1$ rounds: $[(1-p_A)(1-p_B)]^{n-1}$

Probability that at least one duelist hits in the n^{th} round: $1 - [(1-p_A)(1-p_B)]$

$P(\text{duel ends after } n \text{ rounds}) = [(1-p_A)(1-p_B)]^{n-1} \cdot [1 - (1-p_A)(1-p_B)]$

(a)(iv) The conditional probability that the duel ends after the n^{th} round of shots given that A is not hit?

Probability that neither duelist is hit for $(n-1)$ rounds: $[(1-p_A)(1-p_B)]^{(n-1)}$

Probability that A hits B aka A is not hit: p_A

Both events have happened independently \therefore multiplied together

$P(\text{The duel ends after } n^{th} \text{ round} | A \text{ is not hit}) = [(1-p_A)(1-p_B)]^{n-1} p_A$

(a)(v) The conditional probability that the duel ends after the n^{th} round of shots given that both duelists are hit?

Probability that neither duelist is hit after $n-1$ rounds $[(1-p_A)(1-p_B)]^{(n-1)}$

Probability that both duelists are hit: $p_A \cdot p_B$

$P(\text{the duel ends after } n^{th} \text{ round} | \text{Both duelists hit}) = [(1-p_A)(1-p_B)]^{(n-1)} [p_A \cdot p_B]$

NOTE:

p_A : Probability A hits B

p_B : Probability B hits A

Problem 2 Part b

(b) (5 points) Let X be a binary signal, such that $P(X = +1) = P(X = -1) = 0.5$. Suppose X is sent across a noisy channel, with noise N modeled by a zero-mean Gaussian distribution with variance σ^2 , where the noise is independent of the signal that was sent. The received signal is $Y = X + N$.

- (2 points) Find the conditional PDFs of Y given both $\{X = +1\}$ and $\{X = -1\}$.
- (2 points) Suppose a detector compares the received signal Y to a fixed threshold γ to decide which signal was sent. Specifically, the detector decides that $X = +1$ was sent if $Y \geq \gamma$ and that $X = -1$ was sent otherwise. For a given threshold γ , express the probability of error in terms of the Φ function, γ , and σ . Recall, if $Z \sim \mathcal{N}(0, 1)$, then

$$P(Z \leq z) = \Phi(z)$$

- (1 point) What is the optimal value of γ to minimize the probability of error? For that optimal value of γ , what is the probability of error? Leave your answer for the optimal value of γ as a real number and the probability of error in terms of the Φ function and σ .

(2)(b)(i) Find conditional PDFs of Y

Given $Y = X + N$ and

- if $X = +1$: Y becomes $Y = 1 + N$, so $Y \sim N(1, \sigma^2)$

$$\text{conditional pdf: } f_{Y|X=+1}(y) = (2\pi\sigma)^{-\frac{1}{2}} e^{-\frac{(y-1)^2}{2\sigma^2}}$$

- if $X = -1$: Y becomes $Y = -1 + N \therefore Y \sim N(-1, \sigma^2)$

$$\text{conditional pdf } f_{Y|X=-1}(y) = (2\pi\sigma)^{-\frac{1}{2}} e^{-\frac{(y+1)^2}{2\sigma^2}}$$

(2)(b)(ii) Express the probability of error in terms of Φ , γ , and σ

$$\begin{cases} X = +1 & \text{if } Y \geq \gamma \\ X = -1 & \text{otherwise} \end{cases} \rightarrow \text{probability of error} \begin{cases} X = +1 \text{ but } Y < \gamma \\ X = -1 \text{ but } Y \geq \gamma \end{cases}$$

happens when

Case 1: $X = +1$ but $Y < \gamma \quad P(\text{error} | X = +1) = P(Y < \gamma | X = +1) = \Phi\left(\frac{\gamma-1}{\sigma}\right)$

Case 2: $X = -1$ but $Y \geq \gamma \quad P(\text{error} | X = -1) = P(Y \geq \gamma | X = -1) = 1 - \Phi\left(\frac{\gamma+1}{\sigma}\right)$

Total probability: $P(\text{error}) = 0.5 \cdot \Phi\left(\frac{\gamma-1}{\sigma}\right) + 0.5 \cdot \left[1 - \Phi\left(\frac{\gamma+1}{\sigma}\right)\right]$

(2)(b)(iii) What is optimal value of γ to minimize probability of error? What is the probability of error for the optimal value of γ ?

$$\frac{\gamma-1}{\sigma} = \frac{\gamma+1}{\sigma} \rightarrow \gamma - 1 = \gamma + 1 \rightarrow \gamma = 0$$

Substitute $\gamma = 0$ into the probability of error

$$P(\text{error}) = 0.5 \Phi\left(-\frac{\gamma-1}{\sigma}\right) + 0.5 \left[1 - \Phi\left(\frac{\gamma+1}{\sigma}\right)\right] = 0.5 \Phi\left(-\frac{1}{\sigma}\right) + 0.5 \left[1 - \Phi\left(\frac{1}{\sigma}\right)\right]$$

$$\text{Use } \Phi(-z) = 1 - \Phi(z) \therefore P(\text{error}) = 0.5 \left[1 - \Phi\left(\frac{1}{\sigma}\right)\right] + 0.5 \left[1 - \Phi\left(\frac{1}{\sigma}\right)\right] = 1 - \Phi\left(\frac{1}{\sigma}\right)$$

OR $P(\text{error}) = 0.5 \Phi\left(-\frac{1}{\sigma}\right) + 0.5 \Phi\left(-\frac{1}{\sigma}\right) = \Phi\left(-\frac{1}{\sigma}\right)$

$$\therefore P(\text{error}) = \Phi\left(-\frac{1}{\sigma}\right) = 1 - \Phi\left(\frac{1}{\sigma}\right)$$

Problem 2 Part c

(c) (5 points) There is a screening test for lung cancer that looks at the level of LSA (lung specific antigen) in the blood. There are a number of reasons besides lung cancer that a man can have elevated LSA levels. In addition, many types of lung cancer develop so slowly that they are never a problem. Unfortunately, there is currently no test to distinguish the different types and using the test is controversial because it's hard to quantify the accuracy rates and the harm done by false positives. For this problem, we will call a positive test a true positive if it catches a dangerous type of lung cancer. Also, we will assume the following numbers:

- Rate of dangerous type of lung cancer among men over 30 = 0.0005
- True positive rate for the test = 0.9
- False positive rate for the test = 0.01

Suppose you randomly select a man over 30 and perform a screening test.

- (3 points) What is the probability that the man has a dangerous type of the disease given that he had a positive test?
- (2 points) What is the probability that the man has a dangerous type of the disease given that he had a negative test?

$$P(\text{Dangerous}) = 0.0005$$

$$P(\text{Not Dangerous}) = 1 - P(\text{Dangerous}) = 0.9995$$

$$P(\text{Positive} \mid \text{Dangerous}) = 0.9 \quad (\text{The true positive rate})$$

$$P(\text{Positive} \mid \text{Not Dangerous}) = 0.01 \quad (\text{The false positive rate})$$

$$P(\text{Negative} \mid \text{Not Dangerous}) = 0.99 \quad (\text{The negative rate})$$

$$P(\text{Negative} \mid \text{Dangerous}) = 0.1 \quad (\text{The false negative rate})$$

Bayes Theorem

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Part (i): $P(\text{Dangerous} \mid \text{Positive})$

$$P(\text{Dangerous} \mid \text{Positive}) = \frac{P(\text{Positive} \mid \text{Dangerous}) \cdot P(\text{Dangerous})}{P(\text{Positive})} \quad \text{using Baye's Theorem}$$

$$\begin{aligned} P(\text{Positive}) &= P(\text{Positive} \mid \text{Dangerous}) \cdot P(\text{Dangerous}) + P(\text{Positive} \mid \text{Not Dangerous}) \cdot P(\text{Not Dangerous}) \\ &= (0.9)(0.0005) + (0.01)(0.9995) = 0.00045 + 0.009995 = 0.010445 \end{aligned}$$

$$P(\text{Dangerous} \mid \text{Positive}) = \frac{(0.9)(0.0005)}{0.010445} = \frac{0.00045}{0.010445} \approx 0.0431 \quad P(\text{Dangerous} \mid \text{Positive}) = 0.0431$$

Part (ii): $P(\text{Dangerous} \mid \text{Negative})$

$$P(\text{Dangerous} \mid \text{Negative}) = \frac{P(\text{Negative} \mid \text{Dangerous}) \cdot P(\text{Dangerous})}{P(\text{Negative})} \quad \text{using Baye's Theorem}$$

$$\begin{aligned} P(\text{Negative}) &= P(\text{Negative} \mid \text{Dangerous}) \cdot P(\text{Dangerous}) + P(\text{Negative} \mid \text{Not Dangerous}) \cdot P(\text{Not Dangerous}) \\ &= (0.1)(0.0005) + (0.99)(0.9995) = 0.00005 + 0.989505 = 0.989505 \end{aligned}$$

$$P(\text{Dangerous} \mid \text{Negative}) = \frac{0.00005}{0.989505} \approx 0.0000505 \quad P(\text{Dangerous} \mid \text{Negative}) = 0.0000505$$

Problem 2 Part d

(d) (5 points) A family with three daughters and three sons needs to go to the grocery store. Besides the father, who is driving the car, exactly three of the children can come along to the grocery store with him. Suppose that the three children to join the father are chosen randomly, and all such choices are equally likely. Let X denote the number of daughters who accompany the father to the grocery. Let $X_i = 1$ if the i^{th} child who joins the father is a girl, and $X_i = 0$ otherwise.

- (1 point) Find $\text{Cov}(X_i, X_i) = E(X_i X_i) - E(X_i)E(X_i)$, for $1 \leq i \leq 3$. Leave your answer as a fraction.
- (2 points) Find $\text{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$, for $1 \leq i, j \leq 3$ and $i \neq j$. Leave your answer as a fraction.
- (2 points) Find $\text{Var}(X)$. Leave your answer as a fraction.

(i) $\text{Cov}(X_i, X_i)$

$$\text{Cov}(X_i) = \text{var}(X_i) = E[X_i^2] - E[X_i]E[X_i] = E[X_i^2] - (E[X_i])^2$$

$$E[X_i] = P(X_i=1) = \frac{3}{6} = \frac{1}{2}$$

$$\therefore \text{Cov}(X_i) = E[X_i^2] - E[X_i]^2 = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

Σ : variance covariance matrix

$$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2 & \dots & \sigma_1\sigma_n \\ \sigma_2\sigma_1 & \sigma_2^2 & & \sigma_2\sigma_n \\ \vdots & \ddots & \ddots & \vdots \\ \sigma_n\sigma_1 & \dots & \dots & \sigma_n^2 \end{pmatrix}$$

(ii) Find $\text{Cov}(X_i, X_j)$

$$\text{Cov}(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j] \quad E[X_i] = P(X_i=1) = \frac{1}{2}, E[X_j] = P(X_j=1) = \frac{1}{2}$$

$X_i X_j = 1$ if both i^{th} and j^{th} children are both daughters

choosing $\frac{\binom{3}{2} \binom{3}{1}}{\binom{6}{3}} = \frac{3 \cdot 3}{20} = \frac{9}{20} \quad \therefore E[X_i X_j] = \frac{9}{20} \rightarrow \frac{9}{20} \times \frac{1}{3} = \frac{3}{20}$

\uparrow total ways of choosing 3 kids from 6 children

we want combinations

$$\text{cov}(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j] = \frac{3}{20} - \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{3}{20} - \frac{1}{4} = \frac{3}{20} - \frac{5}{20} = -\frac{2}{20} = -\frac{1}{10}$$

(iii) $\text{Var}(X)$

NOTE : From stats 100B

Variance of sum = sum of variances

$$\text{var}(X) = \text{var}(X_1 + X_2 + X_3)$$

$$\begin{aligned} &= \sum_{i=1}^3 \text{var}(X_i) + 2 \sum_{1 \leq i < j \leq 3} \text{cov}(X_i, X_j) \\ &= \left(\frac{1}{4} + \frac{1}{4} + \frac{1}{4}\right) + 2 \cdot \left(-\frac{1}{10} - \frac{1}{10} - \frac{1}{10}\right) \end{aligned}$$

$$= \frac{3}{4} + 2\left(-\frac{3}{10}\right) = \frac{3}{4} - \frac{6}{10} = \frac{15}{20} - \frac{12}{20} = \frac{3}{20}$$

Final Answers

i $\text{Cov}(X_i, X_i) = \frac{1}{4}$

ii $\text{cov}(X_i, X_j) = -\frac{2}{20} = -\frac{1}{10}$

iii $\text{var}(X) = \frac{3}{20}$

Problem 3 Multivariate Derivatives

- (a) (1 points) Let $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$, and $\mathbf{A} \in \mathbb{R}^{n \times m}$. What is $\nabla_{\mathbf{x}} \mathbf{x}^T \mathbf{A} \mathbf{y}$?
- (b) (1 points) Let $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$, and $\mathbf{A} \in \mathbb{R}^{n \times m}$. What is $\nabla_{\mathbf{y}} \mathbf{x}^T \mathbf{A} \mathbf{y}$?
- (c) (1 points) Let $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$, and $\mathbf{A} \in \mathbb{R}^{n \times m}$. What is $\nabla_{\mathbf{A}} \mathbf{x}^T \mathbf{A} \mathbf{y}$?
- (d) (1 points) Let $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times n}$, and let $f = \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x}$. What is $\nabla_{\mathbf{x}} f$?
- (e) (1 points) Let $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times n}$ and $f = \text{tr}(\mathbf{AB})$. What is $\nabla_{\mathbf{A}} f$?
- (f) (2 points) Let $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times n}$ and $f = \text{tr}(\mathbf{BA} + \mathbf{A}^T \mathbf{B} + \mathbf{A}^2 \mathbf{B})$. What is $\nabla_{\mathbf{A}} f$?
- (g) (3 points) Let $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times n}$ and $f = \|\mathbf{A} + \lambda \mathbf{B}\|_F^2$. What is $\nabla_{\mathbf{A}} f$?

(a) $\nabla_{\mathbf{x}} \mathbf{x}^T \mathbf{A} \mathbf{y}$

$$(1 \times n)(n \times m)(m \times 1) = 1 \times 1 \text{ scalar} \therefore \nabla_{\mathbf{x}} \mathbf{x}^T \mathbf{A} \mathbf{y} = \mathbf{A} \mathbf{y}$$

(b) $\nabla_{\mathbf{y}} \mathbf{x}^T \mathbf{A} \mathbf{y} = \nabla_{\mathbf{y}} \text{tr}(\mathbf{x}^T \mathbf{A} \mathbf{y}) = \nabla_{\mathbf{y}} \text{tr}(\mathbf{y} \mathbf{x}^T \mathbf{A})$

we can introduce the trace b/c trace of scalar is a scalar

$$\nabla_{\mathbf{y}} \text{tr}(\mathbf{y} \mathbf{x}^T \mathbf{A}) = (\mathbf{x}^T \mathbf{A})^T = \mathbf{A}^T \mathbf{x}$$

For any $f = \text{tr}(\mathbf{Cy})$ where \mathbf{C} is a matrix independent of \mathbf{y} , derivative wrt \mathbf{y} will be $\nabla_{\mathbf{y}} \text{tr}(\mathbf{Cy}) = \mathbf{C}^T$

$$\therefore \nabla_{\mathbf{y}} (\mathbf{x}^T \mathbf{A} \mathbf{y}) = (\mathbf{x}^T \mathbf{A})^T = \mathbf{A}^T \mathbf{x}$$

(c) $\nabla_{\mathbf{A}} \mathbf{x}^T \mathbf{A} \mathbf{y} = \nabla_{\mathbf{A}} \text{tr}(\mathbf{x}^T \mathbf{A} \mathbf{y}) = \nabla_{\mathbf{A}} \text{tr}(\mathbf{y} \mathbf{x}^T \mathbf{A}) = (\mathbf{y} \mathbf{x}^T)^T = \mathbf{x} \mathbf{y}^T \therefore \nabla_{\mathbf{A}} \mathbf{x}^T \mathbf{A} \mathbf{y} = (\mathbf{y} \mathbf{x}^T)^T = \mathbf{x} \mathbf{y}^T$
 we apply the trace use the $\nabla_{\mathbf{y}} \text{tr}(\mathbf{Cy}) = \mathbf{C}^T$

(d) $\nabla_{\mathbf{x}} f$ where $f = \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x}$

$$\nabla_{\mathbf{x}} f = \nabla_{\mathbf{x}} (\mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x}) = \nabla_{\mathbf{x}} (\mathbf{x}^T \mathbf{A} \mathbf{x}) + \nabla_{\mathbf{x}} (\mathbf{b}^T \mathbf{x}) = (\mathbf{A} + \mathbf{A}^T) \mathbf{x} + \mathbf{b}$$

(e) $\nabla_{\mathbf{f}}$ where $f = \text{tr}(\mathbf{AB}) \rightarrow \nabla_{\mathbf{A}} \text{tr}(\mathbf{AB}) = \mathbf{B}^T$ (use the rule $\nabla_{\mathbf{y}} \text{tr}(\mathbf{Cy}) = \mathbf{C}^T$)

(f) $\nabla_{\mathbf{A}} f$ where $f = \text{tr}(\mathbf{BA} + \mathbf{A}^T \mathbf{B} + \mathbf{A}^2 \mathbf{B})$

$$\nabla_{\mathbf{A}} f = \nabla_{\mathbf{A}} \text{tr}(\mathbf{BA} + \mathbf{A}^T \mathbf{B} + \mathbf{A}^2 \mathbf{B}) = \overset{\textcircled{1}}{\nabla_{\mathbf{A}} \text{tr}(\mathbf{BA})} + \overset{\textcircled{2}}{\nabla_{\mathbf{A}} \text{tr}(\mathbf{A}^T \mathbf{B})} + \overset{\textcircled{3}}{\nabla_{\mathbf{A}} \text{tr}(\mathbf{A}^2 \mathbf{B})}$$

$$\textcircled{1} \quad \nabla_{\mathbf{A}} \text{tr}(\mathbf{A}^T \mathbf{B}) = \mathbf{B}^T \quad \textcircled{2} \quad \nabla_{\mathbf{A}} \text{tr}(\mathbf{A}^T \mathbf{B})^T = \text{tr}(\mathbf{B}^T \mathbf{A}) = (\mathbf{B}^T)^T = \mathbf{B} \quad \textcircled{3} \quad \nabla_{\mathbf{A}} \text{tr}(\mathbf{A}^2 \mathbf{B}) = \mathbf{AB}^T + \mathbf{BA}^T$$

Symmetric $\mathbf{G} 2\mathbf{AB}$

$$\nabla_{\mathbf{A}} f = \mathbf{B}^T + \mathbf{B} + 2\mathbf{AB}$$

(g) $\nabla_{\mathbf{A}} f$ where $f = \|\mathbf{A} + \lambda \mathbf{B}\|_F^2$

$$\nabla_{\mathbf{A}} f = \nabla_{\mathbf{A}} \|\mathbf{A} + \lambda \mathbf{B}\|_F^2 = \nabla_{\mathbf{A}} \text{tr}((\mathbf{A} + \lambda \mathbf{B})^T (\mathbf{A} + \lambda \mathbf{B})) = \text{tr}(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{A}^T \mathbf{B} + \lambda \mathbf{B}^T \mathbf{A} + \lambda^2 \mathbf{B}^T \mathbf{B})$$

$$\nabla_{\mathbf{A}} f = 2\mathbf{A} + \lambda \mathbf{B} + \lambda \mathbf{B} + 0 \rightarrow \nabla_{\mathbf{A}} f = 2\mathbf{A} + 2\lambda \mathbf{B}$$

Problem #4

4. (10 points) Deriving least-squares with matrix derivatives.

In least-squares, we seek to estimate some multivariate output \mathbf{y} via the model

$$\hat{\mathbf{y}} = \mathbf{W}\mathbf{x}$$

In the training set we're given paired data examples $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$ from $i = 1, \dots, n$. Least-squares is the following quadratic optimization problem:

$$\min_{\mathbf{W}} \frac{1}{2} \sum_{i=1}^n \|\mathbf{y}^{(i)} - \mathbf{W}\mathbf{x}^{(i)}\|^2$$

Derive the optimal \mathbf{W} .

Where \mathbf{W} is a matrix, and for each example in the training set, both $\mathbf{x}^{(i)}$ and $\mathbf{y}^{(i)} \forall i = 1, \dots, n$ are vectors.

Hint: you may find the following derivatives useful:

$$\begin{aligned} \frac{\partial \text{tr}(\mathbf{WA})}{\partial \mathbf{W}} &= \mathbf{A}^T \\ \frac{\partial \text{tr}(\mathbf{WAW}^T)}{\partial \mathbf{W}} &= \mathbf{WA}^T + \mathbf{WA} \end{aligned}$$

$$\min_{\mathbf{W}} \frac{1}{2} \sum_{i=1}^n \|\mathbf{y}^{(i)} - \mathbf{W}\mathbf{x}^{(i)}\|^2 = \frac{1}{2} \sum_{i=1}^n (\mathbf{y}^{(i)} - \mathbf{W}\mathbf{x}^{(i)})^T (\mathbf{y}^{(i)} - \mathbf{W}\mathbf{x}^{(i)})$$

$$\begin{aligned} &= \frac{1}{2} \sum_{i=1}^n \mathbf{y}^{(i)\top} \mathbf{y}^{(i)} - \mathbf{x}^{(i)\top} \mathbf{W}^T \mathbf{y}^{(i)} - \mathbf{y}^{(i)\top} \mathbf{W} \mathbf{x}^{(i)} + (\mathbf{W} \mathbf{x}^{(i)})^T (\mathbf{W} \mathbf{x}^{(i)}) \\ &= \frac{1}{2} \sum_{i=1}^n (-2 \mathbf{y}^{(i)\top} \mathbf{W} \mathbf{x}^{(i)} + \mathbf{x}^{(i)\top} \mathbf{W}^T \mathbf{W} \mathbf{x}^{(i)}) \end{aligned}$$

Terms dependent on w
[$\mathbf{y}^{(i)\top} \mathbf{y}^{(i)}$ isn't dependent
on w at all]

Written in matrix form

$$\min_{\mathbf{W}} \frac{1}{2} \sum_{i=1}^n \|\mathbf{y}^{(i)} - \mathbf{W}\mathbf{x}^{(i)}\|^2 = \min_{\mathbf{W}} \frac{1}{2} \|\mathbf{Y} - \mathbf{WX}\|_F^2 \quad \begin{aligned} \mathbf{Y} &= \text{matrix of target vectors } \mathbf{y}^{(i)} \\ \mathbf{X} &= \text{matrix of target vectors } \mathbf{x}^{(i)} \end{aligned}$$

$$\begin{aligned} \min_{\mathbf{W}} \frac{1}{2} \|\mathbf{Y} - \mathbf{WX}\|_F^2 &= \text{tr}((\mathbf{Y} - \mathbf{WX})^T (\mathbf{Y} - \mathbf{WX})) = \text{tr}(\mathbf{Y}^T \mathbf{Y} - \mathbf{Y}^T \mathbf{WX} - \mathbf{X}^T \mathbf{W}^T \mathbf{Y} + (\mathbf{WX})^T (\mathbf{WX})) \\ &= \text{tr}(\mathbf{Y}^T \mathbf{Y}) - 2 \text{tr}(\mathbf{Y}^T \mathbf{WX}) + \text{tr}((\mathbf{WX})^T (\mathbf{WX})) \end{aligned}$$

$$\frac{\partial}{\partial \mathbf{W}} [\text{tr}(\mathbf{Y}^T \mathbf{Y}) - 2 \text{tr}(\mathbf{Y}^T \mathbf{WX}) + \text{tr}((\mathbf{WX})^T (\mathbf{WX}))] = 0 - \mathbf{Y} \mathbf{X}^T + 2 \mathbf{W} \mathbf{X} \mathbf{X}^T [=] 0$$

$$-\mathbf{Y} \mathbf{W}^T + \mathbf{W} \mathbf{X} \mathbf{X}^T = 0$$

$$\mathbf{W} \mathbf{X} \mathbf{X}^T = \mathbf{Y} \mathbf{W}^T \rightarrow \mathbf{W} = \mathbf{Y} \mathbf{X}^T (\mathbf{X} \mathbf{X}^T)^{-1}$$

Problem #5

5. (10 points) Regularized least squares

In lecture, we worked through the following least squares problem

$$\arg \min_{\theta} \frac{1}{2} \sum_{i=1}^N (y^{(i)} - \theta^T \hat{\mathbf{x}}^{(i)})^2$$

However, the least squares has a tendency to overfit the training data. One common technique used to address the overfitting problem is regularization. In this problem, we work through one of the regularization techniques namely ridge regularization which is also known as the regularized least squares problem. In the regularized least squares we solve the following optimization problem

$$\arg \min_{\theta} \frac{1}{2} \sum_{i=1}^N (y^{(i)} - \theta^T \hat{\mathbf{x}}^{(i)})^2 + \frac{\lambda}{2} \|\theta\|_2^2$$

where λ is a tunable regularization parameter. From the above cost function it can be observed that we are seeking least squares solution with a smaller 2-norm. Derive the solution to the regularized least squares problem, i.e Find θ^* .

$$\begin{aligned}
 & \arg \min_{\theta} \frac{1}{2} \sum_{i=1}^N (y^{(i)} - \theta^T \mathbf{x}^{(i)})^2 + \frac{\lambda}{2} \|\theta\|_2^2 \\
 \hookrightarrow \text{vectorization} \quad & \mathcal{L}(\theta) = \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\theta\|_2^2 + \frac{\lambda}{2} \|\theta\|_2^2 \\
 & = \frac{1}{2} (\mathbf{Y} - \mathbf{X}\theta)^T (\mathbf{Y} - \mathbf{X}\theta) + \frac{\lambda}{2} \theta^T \theta \\
 & = \frac{1}{2} (\mathbf{Y}^T \mathbf{Y} - 2\mathbf{Y}^T \mathbf{X}\theta + \theta^T \mathbf{X}^T \mathbf{X}\theta) + \frac{\lambda}{2} \theta^T \theta \\
 \nabla_{\theta} \mathcal{L}(\theta) = \frac{\partial \mathcal{L}(\theta)}{\partial \theta} & = \frac{\partial}{\partial \theta} \left[\frac{1}{2} (\mathbf{Y}^T \mathbf{Y}) - \mathbf{Y}^T \mathbf{X}\theta + \frac{1}{2} \theta^T \mathbf{X}^T \mathbf{X}\theta + \frac{\lambda}{2} \theta^T \theta \right] \\
 & = -\mathbf{X}^T \mathbf{Y} + \mathbf{X}^T \mathbf{X}\theta + \lambda \theta [=] 0 \\
 -\mathbf{X}^T \mathbf{Y} + \mathbf{X}^T \mathbf{X}\theta + \lambda \theta & = 0 \\
 (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})\theta & = \mathbf{X}^T \mathbf{Y} \\
 \theta^* = \hat{\theta} & = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y}
 \end{aligned}$$