# AI-501 MATHEMATICS FOR AI

Tutorial Problems

November 8, 2024

# Week 06 - Problems

## Question 1: Linear Regression

You are given a dataset with four observations and three predictors:

| Observation | $x_1$ | $x_2$ | $x_3$ | $y$ |
|---|---|---|---|---|
| 1 | 1 | 2 | 3 | 14 |
| 2 | 2 | 4 | 6 | 28 |
| 3 | 3 | 6 | 9 | 42 |
| 4 | 4 | 8 | 12 | 56 |

**Notes:**

- There is perfect multicollinearity among the predictors: $x_2 = 2x_1$ and $x_3 = 3x_1$.

**Tasks:**

(a) Perform a standard linear regression of $y$ on $x_1$, $x_2$, and $x_3$ without regularization.

(i) Attempt to compute the regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^\top$ using the normal equations $(X^\top X)\boldsymbol{\beta} = X^\top y$.

(ii) Explain why the normal equations cannot be solved directly in this case.

(b) Compute the regression coefficients using the Moore-Penrose pseudoinverse.

(i) Calculate the pseudoinverse of $X^\top X$.

(ii) Compute $\boldsymbol{\beta} = (X^\top X)^+ X^\top y$, where $(X^\top X)^+$ denotes the pseudoinverse of $X^\top X$.

## Solution

### (a) Standard Linear Regression without Regularization

**(i) Attempt to compute the regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^\top$ using the normal equations $(X^\top X)\boldsymbol{\beta} = X^\top y$.**

First, construct the design matrix $X$ and response vector $y$:

$$X = \begin{bmatrix} 1 & 1 & 2 & 3 \\ 1 & 2 & 4 & 6 \\ 1 & 3 & 6 & 9 \\ 1 & 4 & 8 & 12 \end{bmatrix}, \quad y = \begin{bmatrix} 14 \\ 28 \\ 42 \\ 56 \end{bmatrix}$$

Compute $X^\top X$:
First, calculate the elements of $X^\top X$.
Compute sums:

$$\sum_{i=1}^{4} x_{i0} = \sum_{i=1}^{4} 1 = 4$$

$$\sum_{i=1}^{4} x_{i1} = 1 + 2 + 3 + 4 = 10$$

$$\sum_{i=1}^{4} x_{i2} = 2 + 4 + 6 + 8 = 20$$

$$\sum_{i=1}^{4} x_{i3} = 3 + 6 + 9 + 12 = 30$$

Compute cross-products:

$$\sum_{i=1}^{4} x_{i1}^2 = 1^2 + 2^2 + 3^2 + 4^2 = 30$$
$$\sum_{i=1}^{4} x_{i1}x_{i2} = (1)(2) + (2)(4) + (3)(6) + (4)(8) = 60$$
$$\sum_{i=1}^{4} x_{i1}x_{i3} = (1)(3) + (2)(6) + (3)(9) + (4)(12) = 90$$
$$\sum_{i=1}^{4} x_{i2}^2 = 2^2 + 4^2 + 6^2 + 8^2 = 120$$
$$\sum_{i=1}^{4} x_{i2}x_{i3} = (2)(3) + (4)(6) + (6)(9) + (8)(12) = 180$$
$$\sum_{i=1}^{4} x_{i3}^2 = 3^2 + 6^2 + 9^2 + 12^2 = 270$$

Now, assemble $X^\top X$:

$$X^\top X = \begin{bmatrix} 4 & 10 & 20 & 30 \\ 10 & 30 & 60 & 90 \\ 20 & 60 & 120 & 180 \\ 30 & 90 & 180 & 270 \end{bmatrix}$$

Compute $X^\top y$:

$$X^\top y = \begin{bmatrix} \sum_{i=1}^{4} y_i = 14 + 28 + 42 + 56 = 140 \\ \sum_{i=1}^{4} x_{i1}y_i = (1)(14) + (2)(28) + (3)(42) + (4)(56) = 420 \\ \sum_{i=1}^{4} x_{i2}y_i = (2)(14) + (4)(28) + (6)(42) + (8)(56) = 840 \\ \sum_{i=1}^{4} x_{i3}y_i = (3)(14) + (6)(28) + (9)(42) + (12)(56) = 1260 \end{bmatrix} = \begin{bmatrix} 140 \\ 420 \\ 840 \\ 1260 \end{bmatrix}$$

Set up the normal equations:

$$(X^\top X)\boldsymbol{\beta} = X^\top y$$

$$\begin{bmatrix} 4 & 10 & 20 & 30 \\ 10 & 30 & 60 & 90 \\ 20 & 60 & 120 & 180 \\ 30 & 90 & 180 & 270 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 140 \\ 420 \\ 840 \\ 1260 \end{bmatrix}$$

**(ii) Explain why the least-squares solution cannot be used directly in this case.**
The matrix $X^\top X$ is singular because there is perfect multicollinearity among the predictors:

$$x_2 = 2x_1, \quad x_3 = 3x_1$$

This means the columns of $X$ are linearly dependent, causing $X^\top X$ to be non-invertible. Therefore, the normal equations cannot be solved directly using standard matrix inversion methods.

## (b) Computing Regression Coefficients Using the Moore-Penrose Pseudoinverse

**(i) Calculate the pseudoinverse of $X^\top X$.**
Since $X^\top X$ is singular, we cannot compute its inverse directly. Instead, we compute its pseudoinverse using the Moore-Penrose pseudoinverse.

Alternatively, we can recognize that due to the perfect multicollinearity, the predictors can be expressed in terms of $x_1$:

$$x_2 = 2x_1, \quad x_3 = 3x_1$$

This suggests that we can reduce the problem to a regression with a single predictor, $x_1$.

**(ii) Compute $\boldsymbol{\beta} = (X^\top X)^+ X^\top y$.**

We construct a reduced design matrix $X_{\text{new}}$ containing only the intercept and $x_1$:

$$X_{\text{new}} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ 1 & x_{31} \\ 1 & x_{41} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

Compute $X_{\text{new}}^\top X_{\text{new}}$:

$$X_{\text{new}}^\top X_{\text{new}} = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}$$

Compute the inverse of $X_{\text{new}}^\top X_{\text{new}}$:
First, compute the determinant:

$$\det(X_{\text{new}}^\top X_{\text{new}}) = (4)(30) - (10)^2 = 120 - 100 = 20$$

Compute the inverse:

$$(X_{\text{new}}^\top X_{\text{new}})^{-1} = \frac{1}{20} \begin{bmatrix} 30 & -10 \\ -10 & 4 \end{bmatrix}$$

Compute $X_{\text{new}}^\top y$:

$$X_{\text{new}}^\top y = \begin{bmatrix} \sum_{i=1}^4 y_i = 140 \\ \sum_{i=1}^4 x_{i1} y_i = (1)(14) + (2)(28) + (3)(42) + (4)(56) = 420 \end{bmatrix}$$

Compute $\boldsymbol{\beta}_{\text{new}}$:

$$\boldsymbol{\beta}_{\text{new}} = (X_{\text{new}}^\top X_{\text{new}})^{-1} X_{\text{new}}^\top y = \frac{1}{20} \begin{bmatrix} 30 & -10 \\ -10 & 4 \end{bmatrix} \begin{bmatrix} 140 \\ 420 \end{bmatrix}$$

Compute the products:

$$\beta_0 = \frac{1}{20} (30 \times 140 - 10 \times 420) = \frac{1}{20}(4200 - 4200) = \frac{0}{20} = 0$$
$$\beta_1 = \frac{1}{20} (-10 \times 140 + 4 \times 420) = \frac{1}{20}(-1400 + 1680) = \frac{280}{20} = 14$$

Therefore, the estimated coefficients are:

$$\beta_0 = 0, \quad \beta_1 = 14$$

Since $x_2$ and $x_3$ are linear combinations of $x_1$, their coefficients are not uniquely determined. Any combination of $\beta_1$, $\beta_2$, and $\beta_3$ satisfying:

$$\beta_1 + 2\beta_2 + 3\beta_3 = 14$$

will fit the data perfectly.

For the minimum-norm solution (Moore-Penrose pseudoinverse), the pseudoinverse gives the solution with the smallest $\|\boldsymbol{\beta}\|_2$. This results in:

$$\beta_2 = 0, \quad \beta_3 = 0$$

Thus, the estimated coefficients using the pseudoinverse are:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 14 \\ 0 \\ 0 \end{bmatrix}$$

# Question 2 (Ridge and Lasso Regression)

## Part A: First-order Optimality Conditions

Consider the following Ridge and Lasso regression loss functions:

- **Ridge regression** loss function:

$$L_{\text{ridge}}(\beta_1, \beta_2) = \sum_{i=1}^{n} \left(y_i - (\beta_1 X_{1i} + \beta_2 X_{2i})\right)^2 + \lambda(\beta_1^2 + \beta_2^2)$$

- **Lasso regression** loss function:

$$L_{\text{lasso}}(\beta_1, \beta_2) = \sum_{i=1}^{n} \left(y_i - (\beta_1 X_{1i} + \beta_2 X_{2i})\right)^2 + \lambda(|\beta_1| + |\beta_2|)$$

Derive the first-order optimality conditions for both Ridge and Lasso regression.

**Solution**

For Ridge regression, taking the derivative of the loss function with respect to $\beta_1$ and $\beta_2$ gives the first-order optimality conditions:

$$\frac{\partial L_{\text{ridge}}}{\partial \beta_1} = -2\sum_{i=1}^{n} X_{1i}\left(y_i - \beta_1 X_{1i} - \beta_2 X_{2i}\right) + 2\lambda\beta_1 = 0$$

$$\frac{\partial L_{\text{ridge}}}{\partial \beta_2} = -2\sum_{i=1}^{n} X_{2i}\left(y_i - \beta_1 X_{1i} - \beta_2 X_{2i}\right) + 2\lambda\beta_2 = 0$$

For Lasso regression, differentiating the loss function with respect to $\beta_1$ and $\beta_2$ gives:

$$\frac{\partial L_{\text{lasso}}}{\partial \beta_1} = -2\sum_{i=1}^{n} X_{1i}\left(y_i - \beta_1 X_{1i} - \beta_2 X_{2i}\right) + \lambda\text{sign}(\beta_1) = 0$$

$$\frac{\partial L_{\text{lasso}}}{\partial \beta_2} = -2\sum_{i=1}^{n} X_{2i}\left(y_i - \beta_1 X_{1i} - \beta_2 X_{2i}\right) + \lambda\text{sign}(\beta_2) = 0$$

Here, the term $\text{sign}(\beta_1)$ introduces a subgradient which makes it possible for Lasso to set coefficients to zero.

## Part B: Handling Correlated Features

Discuss how Ridge and Lasso regression handle highly correlated features.

**Solution**

For highly correlated features, Ridge regression tends to shrink both coefficients towards zero but keeps them non-zero. This helps in reducing multicollinearity but does not perform feature selection. Ridge regression distributes the penalty equally across all coefficients, meaning it retains both correlated features but shrinks their magnitudes.

In contrast, Lasso regression may set one of the correlated feature's coefficients to zero, effectively performing feature selection. Lasso introduces sparsity by adding the absolute value penalty ($\ell_1$ norm), which encourages some coefficients to become exactly zero, allowing it to discard one of the correlated features.

**Example:**

Consider two highly correlated features, $X_1$ and $X_2$. Ridge regression will result in small but non-zero coefficients for both $X_1$ and $X_2$. However, Lasso may set $\beta_1$ to zero and retain $\beta_2$, or vice versa, effectively selecting one of the two features.

## Part C: Proof of Shrinkage Behavior

Prove that Ridge regression shrinks all coefficients, while Lasso can shrink some coefficients to zero.

**Solution**

The constraint geometries imposed by the $\ell_2$ and $\ell_1$ norms explain the different behaviors of Ridge and Lasso regression.

For Ridge regression, the penalty term is $\beta_1^2 + \beta_2^2$, which forms a circular constraint (in 2D). As the constraint is quadratic, it smoothly shrinks all coefficients towards zero but never exactly to zero unless $\lambda$ is very large. This means Ridge regression shrinks all coefficients but retains all features.

For Lasso regression, the penalty term is $|\beta_1| + |\beta_2|$, which forms a diamond-shaped constraint. The sharp corners of the constraint allow some coefficients to reach exactly zero. This encourages sparsity in the model, making Lasso capable of feature selection by setting some coefficients to zero. In high dimensions, this geometry results in a sparse solution.

Therefore, while Ridge shrinks all coefficients continuously, Lasso sets some coefficients exactly to zero depending on the data and the regularization strength $\lambda$.

# Problem

You are given the following real symmetric $2 \times 2$ matrix:

$$A = \begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix}$$

**Tasks:**

(a) Compute the eigenvalues and eigenvectors of matrix $A$.

(i) Find the characteristic polynomial of $A$ and solve for its eigenvalues.

(ii) For each eigenvalue, find the corresponding eigenvectors and normalize them.

(b) Diagonalize matrix $A$ using its eigenvalues and eigenvectors.

(i) Construct a matrix $P$ whose columns are the normalized eigenvectors of $A$.

(ii) Show that $P^{-1}AP = D$, where $D$ is the diagonal matrix of eigenvalues.

(c) Use the eigenvalue decomposition to compute $A^3$.

(i) Use the fact that $A = PDP^{-1}$ and $A^k = PD^kP^{-1}$.

(ii) Compute $D^3$ and then $A^3$.

(d) Discuss the properties of matrix $A$:

(i) Is $A$ positive definite?

(ii) What can you say about the eigenvalues in terms of their signs and magnitudes?

(iii) Explain how the entries of $A$ affect its eigenvalues and eigenvectors.

# Solution

## (a) Compute the eigenvalues and eigenvectors of matrix $A$.

**(i) Find the characteristic polynomial of $A$ and solve for its eigenvalues.**

The characteristic polynomial of $A$ is obtained by calculating the determinant of $A - \lambda I$:

$$\det(A - \lambda I) = 0$$

Compute $A - \lambda I$:

$$A - \lambda I = \begin{bmatrix} 4-\lambda & -2 \\ -2 & 1-\lambda \end{bmatrix}$$

Compute the determinant:

$$\begin{aligned} \det(A - \lambda I) &= (4-\lambda)(1-\lambda) - (-2)(-2) \\ &= (4-\lambda)(1-\lambda) - 4 \\ &= (4 \times 1 - 4\lambda - \lambda \times 1 + \lambda^2) - 4 \\ &= (4 - 4\lambda - \lambda + \lambda^2) - 4 \\ &= (\lambda^2 - 5\lambda + 4) - 4 \\ &= \lambda^2 - 5\lambda \end{aligned}$$

Set the characteristic equation:

$$\lambda^2 - 5\lambda = 0$$

Factor out $\lambda$:

$$\lambda(\lambda - 5) = 0$$

Thus, the eigenvalues are:

$$\lambda_1 = 0, \quad \lambda_2 = 5$$

**(ii) For each eigenvalue, find the corresponding eigenvectors and normalize them.**
**For $\lambda = 0$:**
Solve $(A - 0I)\mathbf{v} = \mathbf{0}$:

$$\begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Set up the system of equations:

$$\begin{cases} 4v_1 - 2v_2 = 0 \\ -2v_1 + v_2 = 0 \end{cases}$$

From the second equation:

$$v_2 = 2v_1$$

Substitute back into the first equation to verify consistency:

$$4v_1 - 2(2v_1) = 0 \implies 4v_1 - 4v_1 = 0$$

Thus, the eigenvector corresponding to $\lambda = 0$ is:

$$\mathbf{v}_1 = v_1 \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Normalize $\mathbf{v}_1$:

$$\|\mathbf{v}_1\| = \sqrt{v_1^2 + (2v_1)^2} = v_1\sqrt{1+4} = v_1\sqrt{5}$$

Choose $v_1 = \dfrac{1}{\sqrt{5}}$, so the normalized eigenvector is:

$$\mathbf{v}_1 = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

**For $\lambda = 5$:**
Solve $(A - 5I)\mathbf{v} = \mathbf{0}$:

$$\begin{bmatrix} -1 & -2 \\ -2 & -4 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Set up the system of equations:

$$\begin{cases} -1v_1 - 2v_2 = 0 \\ -2v_1 - 4v_2 = 0 \end{cases}$$

From the first equation:

$$v_1 = -2v_2$$

Substitute back into the second equation to verify consistency:

$$-2(-2v_2) - 4v_2 = 0 \implies 4v_2 - 4v_2 = 0$$

Thus, the eigenvector corresponding to $\lambda = 5$ is:

$$\mathbf{v}_2 = v_2 \begin{bmatrix} -2 \\ 1 \end{bmatrix}$$

Normalize $\mathbf{v}_2$:

$$\|\mathbf{v}_2\| = \sqrt{(-2v_2)^2 + v_2^2} = v_2\sqrt{4+1} = v_2\sqrt{5}$$

Choose $v_2 = \frac{1}{\sqrt{5}}$, so the normalized eigenvector is:

$$\mathbf{v}_2 = \frac{1}{\sqrt{5}} \begin{bmatrix} -2 \\ 1 \end{bmatrix}$$

## (b) Diagonalize matrix $A$ using its eigenvalues and eigenvectors.

**(i) Construct a matrix $P$ whose columns are the normalized eigenvectors of $A$.**

$$P = \left[ \mathbf{v}_1 \,\middle|\, \mathbf{v}_2 \right] = \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix}$$

**(ii) Show that $P^{-1}AP = D$, where $D$ is the diagonal matrix of eigenvalues.**

Since $A$ is symmetric, the eigenvectors are orthogonal, and $P$ is an orthogonal matrix. Therefore, $P^{-1} = P^\top$.

Compute $D$:

$$D = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 5 \end{bmatrix}$$

Verify that $P^\top AP = D$:

$$P^\top AP = D$$

Compute $P^\top$:

$$P^\top = \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{-2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix}$$

Compute $P^\top A$:

$$P^\top A = \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{-2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix} = \begin{bmatrix} \frac{(4)(1)+(-2)(2)}{\sqrt{5}} & \frac{(-2)(1)+(1)(2)}{\sqrt{5}} \\ \frac{(4)(-2)+(-2)(1)}{\sqrt{5}} & \frac{(-2)(-2)+(1)(1)}{\sqrt{5}} \end{bmatrix}$$

Simplify:

$$P^\top A = \begin{bmatrix} \frac{4-4}{\sqrt{5}} & \frac{-2+2}{\sqrt{5}} \\ \frac{-8-2}{\sqrt{5}} & \frac{4+1}{\sqrt{5}} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ \frac{-10}{\sqrt{5}} & \frac{5}{\sqrt{5}} \end{bmatrix}$$

Compute $P^\top AP$:

$$P^\top AP = \begin{bmatrix} 0 & 0 \\ \frac{-10}{\sqrt{5}} & \frac{5}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ \frac{-10}{\sqrt{5}} \times \frac{1}{\sqrt{5}} + \frac{5}{\sqrt{5}} \times \frac{2}{\sqrt{5}} & \frac{-10}{\sqrt{5}} \times \frac{-2}{\sqrt{5}} + \frac{5}{\sqrt{5}} \times \frac{1}{\sqrt{5}} \end{bmatrix}$$

Simplify:

$$P^\top AP = \begin{bmatrix} 0 & 0 \\ \left(\frac{-10}{5} + \frac{10}{5}\right) & \left(\frac{20}{5} + \frac{5}{5}\right) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 5 \end{bmatrix} = D$$

Therefore:

$$P^{-1}AP = D$$

## (c) Use the eigenvalue decomposition to compute $A^3$.

**(i) Use the fact that $A = PDP^{-1}$ and $A^k = PD^kP^{-1}$.**

We have:

$$A^3 = PD^3P^{-1}$$

**(ii) Compute $D^3$ and then $A^3$.**

Compute $D^3$:

$$D^3 = \begin{bmatrix} 0^3 & 0 \\ 0 & 5^3 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 125 \end{bmatrix}$$

Compute $A^3$:

$$A^3 = PD^3P^\top$$

Compute $PD^3$:

$$PD^3 = \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 125 \end{bmatrix} = \begin{bmatrix} 0 & \frac{-250}{\sqrt{5}} \\ 0 & \frac{125}{\sqrt{5}} \end{bmatrix}$$

Compute $A^3$:

$$A^3 = \begin{bmatrix} 0 & \frac{-250}{\sqrt{5}} \\ 0 & \frac{125}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{-2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix} = \begin{bmatrix} \left(0 \times \frac{1}{\sqrt{5}} + \frac{-250}{\sqrt{5}} \times \frac{-2}{\sqrt{5}}\right) & \left(0 \times \frac{2}{\sqrt{5}} + \frac{-250}{\sqrt{5}} \times \frac{1}{\sqrt{5}}\right) \\ \left(0 \times \frac{1}{\sqrt{5}} + \frac{125}{\sqrt{5}} \times \frac{-2}{\sqrt{5}}\right) & \left(0 \times \frac{2}{\sqrt{5}} + \frac{125}{\sqrt{5}} \times \frac{1}{\sqrt{5}}\right) \end{bmatrix}$$

Simplify:

$$A^3 = \begin{bmatrix} \frac{500}{5} & \frac{-250}{5} \\ \frac{-250}{5} & \frac{125}{5} \end{bmatrix} = \begin{bmatrix} 100 & -50 \\ -50 & 25 \end{bmatrix}$$

## (d) Discuss the properties of matrix $A$.

**(i) Is $A$ positive definite?**

A matrix is positive definite if all its eigenvalues are positive. Since $A$ has eigenvalues $\lambda_1 = 0$ and $\lambda_2 = 5$, it is **not positive definite** because one of its eigenvalues is zero (not strictly positive).

**(ii) What can you say about the eigenvalues in terms of their signs and magnitudes?**

- **Signs:** One eigenvalue is zero, and the other is positive. - **Magnitudes:** $\lambda_1 = 0$, $\lambda_2 = 5$. - The zero eigenvalue indicates that $A$ is singular (non-invertible).

**(iii) Explain how the entries of $A$ affect its eigenvalues and eigenvectors.**

- The negative off-diagonal elements introduce coupling between the variables. - The combination of diagonal and off-diagonal elements determines the spread of the eigenvalues. - The structure of $A$ leads to one eigenvalue being zero, reflecting redundancy or linear dependence in the system. - The eigenvectors indicate the directions of variance, influenced by the relative magnitudes of the entries in $A$.

# Week 07 - Problems

## 1. Basic Concepts of SVD

**Question:** What is the Singular Value Decomposition (SVD) of a matrix? Given a matrix $A \in \mathbb{R}^{m \times n}$, write down the SVD of $A$.

**Solution:** The SVD of a matrix $A$ is a factorization $A = U\Sigma V^T$, where $U \in \mathbb{R}^{m \times m}$, $\Sigma \in \mathbb{R}^{m \times n}$ (a diagonal matrix with non-negative real numbers on the diagonal), and $V \in \mathbb{R}^{n \times n}$. The columns of $U$ are called the left singular vectors, and the columns of $V$ are the right singular vectors. The diagonal entries of $\Sigma$ are the singular values.

## 2. Geometric Interpretation of SVD

**Question:** Explain the geometric interpretation of SVD in terms of linear transformations. What do the matrices $U$, $\Sigma$, and $V^T$ represent in terms of rotating, scaling, and projecting vectors?

**Solution:** The SVD decomposes the matrix into three parts: $V^T$ rotates the input data, $\Sigma$ scales the data along the principal axes, and $U$ rotates the scaled data to the output space. The matrix $V^T$ gives the directions in the input space, $\Sigma$ represents the magnitude of stretching along these directions, and $U$ aligns these transformed vectors to the output space.

## 3. SVD and Rank of a Matrix

**Question:** Explain how SVD can be used to determine the rank of a matrix. How do the singular values in the diagonal matrix $\Sigma$ relate to the rank of matrix $A$?

**Solution:** The rank of matrix $A$ is the number of non-zero singular values in $\Sigma$. These singular values represent the extent to which each independent direction contributes to the transformation described by $A$. If some singular values are zero, they indicate that the matrix compresses the input space along those dimensions, reducing the rank.

## 4. Low-Rank Approximation via SVD

**Question:** Given the SVD of a matrix $A \in \mathbb{R}^{m \times n}$, how can you compute a low-rank approximation of $A$? Why is this useful in practice, especially in data compression or noise reduction? Why is this approximation optimal in terms of the Frobenius norm?

**Solution:**

The optimal low-rank approximation theorem states that for a given matrix $A \in \mathbb{R}^{m \times n}$ and a desired rank $k$, the low-rank approximation $A_k$ that minimizes the approximation error is given by:

$$A_k = \arg \min_{\text{rank}(B) \leq k} \|A - B\|_F$$

where $\| \cdot \|_F$ denotes the Frobenius norm.

Given the SVD of the matrix $A = U\Sigma V^T$, the best rank-$k$ approximation is constructed by keeping only the largest $k$ singular values from $\Sigma$, along with their corresponding singular vectors from $U$ and $V$. The low-rank approximation is given by:

$$A_k = U_k \Sigma_k V_k^T$$

where $U_k \in \mathbb{R}^{m \times k}$, $\Sigma_k \in \mathbb{R}^{k \times k}$, and $V_k \in \mathbb{R}^{n \times k}$. The reason this approximation is optimal in the Frobenius norm is that the singular values $\sigma_1, \sigma_2, \ldots, \sigma_k$ in $\Sigma_k$ capture the largest contributions to the matrix's variance. Hence, minimizing the sum of the squared errors in the approximation yields:

$$\|A - A_k\|_F^2 = \sum_{i=k+1}^{\min(m,n)} \sigma_i^2$$

This result shows that the error is the sum of the squares of the discarded singular values, making $A_k$ the best approximation in terms of minimizing the Frobenius norm.

This approximation is useful in applications like data compression, where it reduces the storage or transmission cost, and in noise reduction, where it can help filter out small singular values that may correspond to noise rather than signal.

## 5. SVD, Pseudoinverse and Solution of Linear Systems

**Question:** How can the SVD be used to compute the Moore-Penrose pseudoinverse of a matrix $A \in \mathbb{R}^{m\times n}$? Write the formula for the pseudoinverse using SVD.

**Solution:** Given the SVD of $A = U\Sigma V^T$, the Moore-Penrose pseudoinverse $A^+$ is given by:

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

where $\Sigma^+$ is obtained by taking the reciprocal of each non-zero singular value in $\Sigma$ and transposing the matrix.

Using the SVD $A = U\Sigma V^T$, the least-squares solution is given by:

$$\mathbf{x} = A^+\mathbf{b} = V\Sigma^+ U^T\mathbf{b}$$

The solution is unique if all singular values of $A$ are non-zero, meaning that the matrix $A$ has full column rank.

## 6. Condition Number and Stability

**Question:** Explain how the singular values of a matrix relate to its condition number. Why is the condition number important for numerical stability?

**Solution:** The condition number $\kappa(A)$ of a matrix $A$ is defined as the ratio of the largest singular value to the smallest non-zero singular value:

$$\kappa(A) = \frac{\sigma_{\max}}{\sigma_{\min}}$$

A high condition number indicates that the matrix is ill-conditioned, meaning that small changes in the input can cause large changes in the output, leading to numerical instability. I will explain this during the tutorial.

## 7. Compute the SVD of a 2x2 Matrix

**Question:** Given the matrix

$$A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$$

compute the Singular Value Decomposition (SVD) of $A$. Provide the matrices $U$, $\Sigma$, and $V^T$.

**Solution:** 1. Compute $A^TA$:

$$A^TA = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} 10 & 6 \\ 6 & 10 \end{pmatrix}$$

2. Compute the eigenvalues of $A^TA$. The characteristic equation is:

$$\det(A^TA - \lambda I) = 0 \implies \det\left(\begin{pmatrix} 10 & 6 \\ 6 & 10 \end{pmatrix} - \lambda\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) = 0$$

$$\det\begin{pmatrix} 10-\lambda & 6 \\ 6 & 10-\lambda \end{pmatrix} = (10-\lambda)^2 - 36 = 0 \implies \lambda^2 - 20\lambda + 64 = 0$$

The eigenvalues are $\lambda_1 = 16$, $\lambda_2 = 4$.

3. The singular values of $A$ are $\sigma_1 = \sqrt{16} = 4$, $\sigma_2 = \sqrt{4} = 2$.

4. Compute the eigenvectors of $A^TA$ corresponding to the eigenvalues. For $\lambda_1 = 16$, solve:

$$(A^TA - 16I)v_1 = 0 \implies \begin{pmatrix} -6 & 6 \\ 6 & -6 \end{pmatrix}\begin{pmatrix} v_1^1 \\ v_1^2 \end{pmatrix} = 0$$

This gives $v_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

For $\lambda_2 = 4$, solve:

$$(A^TA - 4I)v_2 = 0 \implies \begin{pmatrix} 6 & 6 \\ 6 & 6 \end{pmatrix}\begin{pmatrix} v_2^1 \\ v_2^2 \end{pmatrix} = 0$$

This gives $v_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix}$.

5. Thus, $V = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$.

6. The matrix $U$ is computed as $U = AV\Sigma^{-1}$. Perform the matrix multiplication to get $U$:

$$U = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

So, the SVD is:

$$A = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}^T$$

## 8. PCA using SVD

**Question:** How can Singular Value Decomposition (SVD) be used to perform Principal Component Analysis (PCA)? Outline the steps to compute PCA using SVD for a dataset.

**Solution:** PCA can be performed using SVD by first centering the data (subtracting the mean from each data point). Then, apply SVD to the centered data matrix $X = U\Sigma V^T$. The columns of $V$ represent the principal components (directions of maximum variance), and the singular values in $\Sigma$ correspond to the magnitude of the variance along each principal component. The first $k$ columns of $V$ can be used to reduce the dimensionality of the data by projecting the data onto the top $k$ principal components.

## 9. Variance Explained by Principal Components

**Question:** Given the SVD of a dataset $X$, how do you compute the proportion of the total variance explained by the first $k$ principal components? Provide an expression for this variance.

**Solution:** The proportion of variance explained by the first $k$ principal components is given by the sum of the squares of the first $k$ singular values divided by the sum of the squares of all the singular values:

$$\text{Variance explained by top } k \text{ components} = \frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{r} \sigma_i^2}$$

where $\sigma_i$ are the singular values, and $r$ is the rank of the matrix.

## 10. Connection Between Eigenvalues and SVD in PCA

**Question:** Show the relationship between the eigenvalue decomposition of the covariance matrix and the SVD of the data matrix when performing PCA. How do the eigenvalues and eigenvectors of the covariance matrix relate to the singular values and singular vectors from SVD?

**Solution:** The covariance matrix $\frac{1}{n}X^TX$ has the same eigenvectors as the right singular vectors from the SVD of $X$, and the eigenvalues are the squares of the singular values divided by $n$. Specifically, if $X = U\Sigma V^T$, then $X^TX = V\Sigma^T\Sigma V^T$, where $V$ contains the eigenvectors and the diagonal elements of $\Sigma^T\Sigma$ (i.e., the singular values squared) are the eigenvalues.

## 9. Principal Component Analysis (PCA) using SVD

**Question:** Consider the following dataset:

$$X = \begin{pmatrix} 2 & 0 \\ 3 & 2 \\ 4 & 4 \end{pmatrix}$$

a) Center the dataset by subtracting the mean of each column.
b) Perform Singular Value Decomposition (SVD) on the centered dataset.
c) Project the data onto the first principal component.
d) Compute the proportion of the variance explained by each principal component.

**Solution:**

1. Center the data:

$$\text{Mean of column 1} = \frac{2+3+4}{3} = 3, \quad \text{Mean of column 2} = \frac{0+2+4}{3} = 2$$

The centered data matrix is:

$$X_{\text{centered}} = \begin{pmatrix} 2-3 & 0-2 \\ 3-3 & 2-2 \\ 4-3 & 4-2 \end{pmatrix} = \begin{pmatrix} -1 & -2 \\ 0 & 0 \\ 1 & 2 \end{pmatrix}$$

2. Perform SVD on the centered matrix:

$$X_{\text{centered}} = U\Sigma V^T$$

This yields:

$$U = \begin{pmatrix} -0.707 & 0 & 0.707 \\ 0 & 1 & 0 \\ 0.707 & 0 & 0.707 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 3.162 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad V = \begin{pmatrix} 0.447 & -0.894 \\ 0.894 & 0.447 \end{pmatrix}$$

3. Project the data onto the first principal component:

$$\text{First principal component} = V_1 = \begin{pmatrix} 0.447 & 0.894 \end{pmatrix}^T$$

The projected data is:

$$X_{\text{projected}} = X_{\text{centered}} V_1 = \begin{pmatrix} -1 & -2 \\ 0 & 0 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 0.447 & 0.894 \end{pmatrix} = \begin{pmatrix} -2.236 \\ 0 \\ 2.236 \end{pmatrix}$$

**Variance Explained by Principal Components**

In general the proportion of the total variance due to the $i_{th}$ principal component is given by $\frac{\sigma_i^2}{\Sigma_{k=0}^{r}\sigma_k^2}$. In this case, we only have one eigenvalue, therefore this will account for 100% of the total variance.

# Week 08 - Problems

## Calculus

**Multivariate Taylor series**

The multivariate Taylor series is a generalization of the Taylor series for functions of multiple variables, expanding a function around a point to approximate its values nearby. For any smooth function $f(\mathbf{x})$, where $\mathbf{x} = [x_1, x_2, .., x_n]^T$, the second order Taylor Series Expansion around a point $\mathbf{a} = [a_1, a_2, .., a_n]^T$ is given by:

$$f(\mathbf{a}) + \nabla f(\mathbf{a})^T(\mathbf{x} - \mathbf{a}) + \frac{1}{2}(\mathbf{x} - \mathbf{a})^T H(\mathbf{a})(\mathbf{x} - \mathbf{a}) \tag{1}$$

Where, $\nabla f(\mathbf{a})$ is the gradient of $f(\mathbf{x})$ evaluated at $\mathbf{a}$ and $H(\mathbf{a})$ represents the Hessian Matrix evaluated at $\mathbf{a}$.

## Problems

1. Given $f(x, y) = \sin(x)e^y + \cos(xy)$, find the second-order Taylor series expansion around the point $\mathbf{a} = (0, 0)$.
2. Let $f(x, y, z) = x^2 + y^2 + z^2 + 2xy + 3xz - yz$. Use the second-order Taylor expansion around the point $\mathbf{a} = (1, -1, 2)$ to approximate $f$ and determine if $\mathbf{a}$ is a local minimum, maximum, or saddle point. Use the eigenvalues of the Hessian matrix to justify your answer.
3. Consider the function $f(x, y, z) = xe^{y+z} + y\sin(z) + z\cos(x)$. Derive the second-order Taylor series expansion around $\mathbf{a} = (0, 0, 0)$. Calculate the gradient and Hessian at $\mathbf{a}$, then use them to construct the Taylor expansion.

## Solution

**Problem 1 Step 1: First Derivatives**

We compute the first derivatives of $f(x, y)$:

$$\frac{\partial f}{\partial x} = e^y \cos(x) - y\sin(xy)$$
$$\frac{\partial f}{\partial y} = \sin(x)e^y - x\sin(xy)$$

**Step 2: Second Derivatives**

Next, we calculate the second-order partial derivatives of $f(x, y)$:

$$\frac{\partial^2 f}{\partial x^2} = -e^y \sin(x) - y^2 \cos(xy)$$

Evaluating at $(0, 0)$:

$$\frac{\partial^2 f}{\partial x^2}(0, 0) = -\sin(0) - 0 = 0$$

$$\frac{\partial^2 f}{\partial y^2} = \sin(x)e^y - x^2 \cos(xy)$$

Evaluating at $(0, 0)$:

$$\frac{\partial^2 f}{\partial y^2}(0, 0) = \sin(0) - 0 = 0$$

$$\frac{\partial^2 f}{\partial x \partial y} = \cos(x)e^y - \sin(xy) - xy\cos(xy)$$

Evaluating at $(0, 0)$:

$$\frac{\partial^2 f}{\partial x \partial y}(0, 0) = \cos(0) - 0 = 1$$

**Step 3: Hessian Matrix**

The Hessian matrix $H(x, y)$ at $(0, 0)$ is:

$$H(0,0) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

**Step 4: Final Answer**

Putting everything together, we obtain:

$$f(x,y) \approx 1 + \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \frac{1}{2} \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

$$f(x,y) \approx 1 + x + xy$$

**Problem 2**

**Step 1: First Derivatives**

First, plugging **a** in the function, we get $f(1, -1, 2) = 12$.

We compute the first derivatives of $f(x, y, z)$:

$$\begin{aligned} \frac{\partial f}{\partial x} &= 2x + 2y + 3z = 6 \\ \frac{\partial f}{\partial y} &= 2y + 2x - z = -2 \\ \frac{\partial f}{\partial z} &= 2z + 3x - y = 8 \end{aligned}$$

**Step 2: Second Derivatives**

Next, we calculate each second derivative:

$$\begin{aligned} \frac{\partial^2 f}{\partial x^2} &= 2 \\ \frac{\partial^2 f}{\partial y^2} &= 2 \\ \frac{\partial^2 f}{\partial z^2} &= 2 \\ \frac{\partial^2 f}{\partial x \partial y} &= 2 \\ \frac{\partial^2 f}{\partial x \partial z} &= 3 \\ \frac{\partial^2 f}{\partial y \partial z} &= -1 \end{aligned}$$

**Step 3: Hessian Matrix**

The Hessian matrix $H(x, y, z)$ is:

$$H(x,y,z) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial x \partial z} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} & \frac{\partial^2 f}{\partial y \partial z} \\ \frac{\partial^2 f}{\partial z \partial x} & \frac{\partial^2 f}{\partial z \partial y} & \frac{\partial^2 f}{\partial z^2} \end{bmatrix} = \begin{bmatrix} 2 & 2 & 3 \\ 2 & 2 & -1 \\ 3 & -1 & 2 \end{bmatrix}$$

This Hessian matrix is constant, so we use it for any point, including $(1, -1, 2)$.

**Step 4: Eigenvalues** We want to find the eigenvalues by solving the characteristic equation:

$$\det(H - \lambda I) = 0$$

where $I$ is the identity matrix and $\lambda$ is the eigenvalue. First, we write the matrix $H - \lambda I$:

$$H - \lambda I = \begin{bmatrix} 2-\lambda & 2 & 3 \\ 2 & 2-\lambda & -1 \\ 3 & -1 & 2-\lambda \end{bmatrix}$$

Now, we calculate the determinant:

$$\det(H - \lambda I) = \det \begin{bmatrix} 2-\lambda & 2 & 3 \\ 2 & 2-\lambda & -1 \\ 3 & -1 & 2-\lambda \end{bmatrix}$$

Expanding the determinant:

$$(2-\lambda)\left[(2-\lambda)(2-\lambda) - (-1)(-1)\right] - 2\left[2(2-\lambda) - (-1)(3)\right] + 3\left[2(-1) - 3(2-\lambda)\right]$$

Simplifying the terms:

$$= (2-\lambda)\left[(2-\lambda)^2 - 1\right] - 2\left[2(2-\lambda) + 3\right] + 3\left[-2 - 6 + 3\lambda\right]$$

We then solve this cubic equation for $\lambda$ to find the eigenvalues. The eigenvalues are: $\lambda_1 = -2.11, \lambda_2 = 5.2, \lambda_3 = 2.91$, Hence **a** is not an extremum.

**Step 5: final answer**

Putting everything together:

$$f(x,y,z) \approx 12 + \begin{bmatrix} 6 & -2 & 8 \end{bmatrix} \begin{bmatrix} x-1 \\ y+1 \\ z-2 \end{bmatrix} + \begin{bmatrix} x-1 & y+1 & z-2 \end{bmatrix} \begin{bmatrix} 2 & 2 & 3 \\ 2 & 2 & -1 \\ 3 & -1 & 2 \end{bmatrix} \begin{bmatrix} x-1 \\ y+1 \\ z-2 \end{bmatrix}$$

**Problem 3 Step 1: First Derivatives**

We compute the first derivatives of $f(x,y,z)$:

$$\frac{\partial f}{\partial x} = e^{y+z} - z\sin(x)$$
$$\frac{\partial f}{\partial y} = xe^{y+z} + \cos(z)$$
$$\frac{\partial f}{\partial z} = xe^{y+z} + y\cos(z) - \sin(x)$$

**Step 2: Second Derivatives**

Next, we calculate each second derivative:

$$\frac{\partial^2 f}{\partial x^2} = -z\cos(x)$$
$$\frac{\partial^2 f}{\partial y^2} = xe^{y+z}$$
$$\frac{\partial^2 f}{\partial z^2} = xe^{y+z} - y\sin(z)$$
$$\frac{\partial^2 f}{\partial x \partial y} = e^{y+z}$$
$$\frac{\partial^2 f}{\partial x \partial z} = e^{y+z} - \sin(x)$$
$$\frac{\partial^2 f}{\partial y \partial z} = xe^{y+z} - y\sin(z)$$

**Step 3: Hessian Matrix**

The Hessian matrix $H(x,y,z)$ is:

$$H(x,y,z) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial x \partial z} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} & \frac{\partial^2 f}{\partial y \partial z} \\ \frac{\partial^2 f}{\partial z \partial x} & \frac{\partial^2 f}{\partial z \partial y} & \frac{\partial^2 f}{\partial z^2} \end{bmatrix}$$

Substituting the second derivatives, we have:

$$H(x,y,z) = \begin{bmatrix} -z\cos(x) & e^{y+z} & e^{y+z} - \sin(x) \\ e^{y+z} & xe^{y+z} & xe^{y+z} - y\sin(z) \\ e^{y+z} - \sin(x) & xe^{y+z} - y\sin(z) & xe^{y+z} - y\sin(z) \end{bmatrix}$$

**Step 4: Evaluate at a** $= (0, 0, 0)$
Substitute $x = 0$, $y = 0$, and $z = 0$ into the Hessian matrix:

$$H(0,0,0) = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

**Step 5: Final Answer**

$$f(x,y,z) \approx x + y - z + \frac{1}{2} \begin{bmatrix} x & y & z \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

## Overview of Integration as Anti-Derivative

Integration is the inverse operation of differentiation. If $F(x)$ is an anti-derivative of $f(x)$, then the integral of $f(x)$ over an interval $[a, b]$ is represented as:

$$\int f(x)\,dx = F(x) + C$$

where $C$ is the constant of integration.

### Fundamental Theorem of Calculus

- **Part 1:** If $F(x)$ is an anti-derivative of $f(x)$, then:

$$\frac{d}{dx} \int_a^x f(t)\,dt = f(x)$$

- **Part 2:** The definite integral of $f(x)$ from $a$ to $b$ is given by:

$$\int_a^b f(x)\,dx = F(b) - F(a)$$

  where $F$ is any anti-derivative of $f$.

### Use Cases of Integration in Machine Learning and AI

There are hundreds of use-cases but we will just list two here.

- **Probabilistic Models**: Integration is essential for probability distributions, especially for computing marginal probabilities, normalization constants, and posterior distributions in Bayesian models.

$$p(y) = \int p(y|x)p(x)\,dx$$

- **Expectation and Variance Calculation**: Calculating the expected value of a continuous random variable $X$ with probability density function $p(x)$ requires integration:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xp(x)\,dx$$

  Variance also involves integration:

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 p(x)\,dx$$

**Evaluation of Integration as Quadrature**

**Quadrature by Summation**

Quadrature refers to methods used for the numerical approximation of definite integrals. By discretizing the integral, we approximate it as a finite sum over equally spaced points. Let's consider the integral of a function $f(x)$ over an interval $[a, b]$:

$$\int_a^b f(x)\,dx$$

We can approximate this integral by dividing the interval $[a, b]$ into $n$ subintervals of equal width $\Delta x = \frac{b-a}{n}$. Then, the integral becomes a summation:

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{n} f(x_i)\,\Delta x$$

where $x_i = a + (i-1)\Delta x$ represents the points at which $f(x)$ is evaluated.

**Problems on Basic Conceptual Integration**

**Problem:** Find the anti-derivative of the function $f(x) = 3x^2 + 2x + 1$ and evaluate the definite integral over $[1, 3]$.
**Solution:** The anti-derivative $F(x)$ of $f(x) = 3x^2 + 2x + 1$ is:

$$F(x) = x^3 + x^2 + x + C$$

Evaluating $\int_1^3 (3x^2 + 2x + 1)\,dx$:

$$F(3) - F(1) = (3^3 + 3^2 + 3) - (1^3 + 1^2 + 1) = 39 - 3 = 36$$

**Problem:** Let $X$ be a continuous random variable with a probability density function $f(x) = 4x^3$ for $x \in [0, 1]$. Verify that $f(x)$ is a valid probability density function and compute $\mathbb{E}[X]$, the expected value of $X$.
**Solution:**

- First, verify $f(x)$ is a valid probability density function by ensuring it integrates to 1:

$$\int_0^1 4x^3\,dx = x^4\big|_0^1 = 1$$

- To compute $\mathbb{E}[X]$:

$$\mathbb{E}[X] = \int_0^1 x \cdot 4x^3\,dx = \int_0^1 4x^4\,dx = \left.\frac{4x^5}{5}\right|_0^1 = \frac{4}{5}$$

# Convex Sets and Functions

In the lecture, we studied the convex functions. Here, we briefly define the convex set.

## Convex Set – Overview

A subset $C$ of a vector space $\mathbb{R}^n$ is called a **convex set** if, for any two points $x, y \in C$, the line segment joining $x$ and $y$ is entirely contained within $C$. Formally, this means that for any $x, y \in C$ and for any $\theta \in [0, 1]$,

$$\theta x + (1 - \theta)y \in C$$

In other words, if we take any two points in the set and form a convex combination of them, the result will still lie within the set. The expression $\theta x + (1-\theta)y$ is known as a **convex combination** of $x$ and $y$, where $\theta$ is a weight that determines the relative position between $x$ and $y$ along the line segment joining them. Below are some examples of convex sets.

### 1. Euclidean Space $\mathbb{R}^n$

The entire $\mathbb{R}^n$ space is a convex set. For any two points $x, y \in \mathbb{R}^n$, any convex combination $\theta x + (1-\theta)y$ for $\theta \in [0, 1]$ lies within $\mathbb{R}^n$.

### 2. Line Segments

A line segment connecting two points $x$ and $y$ in $\mathbb{R}^n$ is convex. By definition, all points of the form $\theta x + (1 - \theta)y$ for $\theta \in [0, 1]$ are contained within the line segment between $x$ and $y$, making it a convex set.

### 3. Convex Polyhedrons

A **convex polyhedron** is the intersection of a finite number of half-spaces in $\mathbb{R}^n$, such as a triangle or a quadrilateral in $\mathbb{R}^2$, or a polytope in higher dimensions. For example, in $\mathbb{R}^2$, any triangle or quadrilateral is a convex set since the line segment joining any two points within the shape will also lie within the shape.

### 4. Convex Hull

The **convex hull** of a set of points $\{x_1, x_2, \ldots, x_k\}$ in $\mathbb{R}^n$ is the smallest convex set containing all the points. Formally, it is defined as:

$$\text{conv}(x_1, x_2, \ldots, x_k) = \left\{ \sum_{i=1}^{k} \theta_i x_i \,\middle|\, \theta_i \geq 0, \sum_{i=1}^{k} \theta_i = 1 \right\}$$

The convex hull is itself a convex set because any convex combination of points within the set remains in the set.

### 5. Balls and Spheres

A **ball** (or disk in 2D) is the set of all points within a certain distance (radius) from a central point. Mathematically, a ball of radius $r$ centered at $c$ in $\mathbb{R}^n$ is defined as:

$$B(c, r) = \{x \in \mathbb{R}^n \mid \|x - c\| \leq r\}$$

This is a convex set because, for any two points within the ball, the line segment joining them lies entirely within the ball.

### 6. Half-Spaces

A **half-space** in $\mathbb{R}^n$ is defined by a linear inequality. For a vector $a \in \mathbb{R}^n$ and a scalar $b$, the set

$$H = \{x \in \mathbb{R}^n \mid a^T x \leq b\}$$

is a convex set. For any two points $x, y \in H$, any convex combination $\theta x + (1 - \theta)y$ will also satisfy $a^T(\theta x + (1 - \theta)y) \leq b$, meaning the combination remains in $H$.

## Convex Functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is called **convex** if its domain is a convex set and, for all $x, y \in \text{dom}(f)$ and $\theta \in [0, 1]$, it satisfies the inequality (Jensen's Inequality):

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y)$$

This property implies that the line segment joining any two points on the graph of $f$ lies above the graph itself. In simpler terms, the function does not "dip below" any line segment connecting two points on its curve. Please also review the first-order and second-order conditions for convexity.

### Examples of Convex Functions

- **Linear Functions:** Any linear function of the form $f(x) = a^T x + b$, where $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$, is convex (and also concave). This is because it satisfies the convexity inequality as an equality.
- **Quadratic Functions with Positive Semi-Definite Hessian:** A function of the form $f(x) = x^T Ax + b^T x + c$, where $A$ is a positive semi-definite matrix, is convex. The positive semi-definite nature of $A$ ensures that the function curves upward, satisfying the convexity conditions.
- **Exponential Function:** The function $f(x) = e^x$ is convex on $\mathbb{R}$. It satisfies the convexity inequality because the slope of the exponential function increases as $x$ increases.
- **Norm Functions:** The Euclidean norm $f(x) = \|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$ is convex. Norms are convex functions as they satisfy the triangle inequality.
- **Logarithmic Functions:** The negative logarithm function $f(x) = -\log(x)$ is convex on $(0, \infty)$. Since the second derivative of $-\log(x)$ is positive, it satisfies the convexity condition.

### Problems

**Problem 1: Verifying Convexity of a Quadratic Function**

**Problem Statement:** Let $f(x) = x^2 + 4x + 5$. Determine if $f(x)$ is convex over $\mathbb{R}$.

**Solution:** To verify convexity, we can check the second derivative of $f(x)$.

$$f'(x) = 2x + 4$$

$$f''(x) = 2$$

Since $f''(x) = 2 > 0$ for all $x \in \mathbb{R}$, the function $f(x) = x^2 + 4x + 5$ is convex.

**Problem 2: Proving Convexity of the Exponential Function**

**Problem Statement:** Show that the function $f(x) = e^x$ is convex over $\mathbb{R}$.

**Solution:** To determine convexity, we calculate the second derivative of $f(x) = e^x$.

$$f'(x) = e^x$$

$$f''(x) = e^x$$

Since $f''(x) = e^x > 0$ for all $x \in \mathbb{R}$, the function $f(x) = e^x$ is convex.

**Problem 3: Verifying Convexity Using the First-Order Condition**

**Problem Statement:** Let $f(x) = \log(x)$ for $x > 0$. Verify if $f(x)$ is concave over $(0, \infty)$ using the first-order condition.

**Solution:** To verify concavity, we need to check if $f(y) \leq f(x) + f'(x)(y - x)$ for any $x, y > 0$.

1. First, calculate the derivative of $f(x)$:

$$f'(x) = \frac{1}{x}$$

2. Now, substitute into the first-order condition:

$$f(y) \leq f(x) + f'(x)(y - x)$$

Expanding this, we have:

$$\log(y) \leq \log(x) + \frac{1}{x}(y - x)$$

This inequality holds for $x, y > 0$, verifying that $\log(x)$ is concave over $(0, \infty)$.

**Problem 4: Proving the Convexity (Advanced)**

Show that the following functions $f : \mathbb{R}^n \to \mathbb{R}$ are convex.

(a) $f(x) = \|Ax - b\|$, where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $\|\cdot\|$ is a norm on $\mathbb{R}^m$

(b) $f(x) = -\left(\det(A_0 + x_1 A_1 + \cdots + x_n A_n)\right)^{1/m}$ on the domain $\{x \mid A_0 + x_1 A_1 + \cdots + x_n A_n \succ 0\}$, where $A_i \in S^m$ (the space of $m \times m$ symmetric matrices).

**Solutions:**

**(a)**
We know that a function $g(y) = \|y\|$ is convex for any norm $\|\cdot\|$. Since $g(y) = \|y\|$ is convex, we can show that $f(x) = \|Ax - b\|$ is also convex as follows.
Observe that $f(x) = g(Ax - b)$, where $g(y) = \|y\|$. Since $g(y)$ is convex and $Ax - b$ is an affine function of $x$, $f(x) = g(Ax - b)$ is also convex. Thus, $f(x) = \|Ax - b\|$ is convex as it is a composition of a convex function $g(y) = \|y\|$ with an affine transformation $Ax - b$.
**(b)**
To show that $f(x)$ is convex, we use the fact that the determinant function raised to the power $\frac{1}{m}$ is concave on the set of positive definite matrices. Specifically, the function $g(Y) = (\det(Y))^{1/m}$ is concave for $Y \succ 0$.
The expression $A_0 + x_1 A_1 + \cdots + x_n A_n$ is an affine transformation of $x$ which maps into the space of symmetric matrices. Since the composition of a concave function with a concave transformation is convex when the function is negated, $f(x) = -(\det(A_0 + x_1 A_1 + \cdots + x_n A_n))^{1/m}$ is convex on the domain $\{x \mid A_0 + x_1 A_1 + \cdots + x_n A_n \succ 0\}$.
Therefore, $f(x)$ is convex in $x$ on the specified domain.

# Week 09 - Problems

## 1 LP and QP

Consider a network with $n$ nodes and directed links between each pair of nodes. Let $x_{ij}$ denote the flow from node $i$ to node $j$. The cost of the flow along the link from node $i$ to node $j$ is given by $c_{ij}x_{ij}$, where $c_{ij}$ are given constants. The total cost across the network is:

$$C = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}x_{ij}.$$

Each link flow $x_{ij}$ is subject to a lower bound $l_{ij}$ (typically nonnegative) and an upper bound $u_{ij}$.

The external supply at node $i$ is represented by $b_i$, where $b_i > 0$ means an external flow enters the network at node $i$, and $b_i < 0$ implies that an amount $|b_i|$ flows out of the network from node $i$. We assume that $\mathbf{1}^T\mathbf{b} = 0$, meaning the total external supply equals the total external demand.

The conservation of flow at each node implies that the total flow into node $i$ (from external supply and other nodes) minus the total flow out of node $i$ equals zero.

The problem is to minimize the total cost of flow through the network, subject to the constraints described above. Formulate this as a linear program (LP).

**Solution:**

$$\begin{array}{ll} \text{minimize} & C = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}x_{ij} \\ \text{subject to} & b_i + \sum_{j=1}^{n} x_{ij} - \sum_{j=1}^{n} x_{ji} = 0, \quad i = 1, \ldots, n, \\ & l_{ij} \le x_{ij} \le u_{ij}, \quad \forall i, j. \end{array}$$

In this formulation:

- The objective function $C = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}x_{ij}$ minimizes the total flow cost across the network, where $c_{ij}$ represents the cost per unit flow from node $i$ to node $j$, and $x_{ij}$ is the flow amount on this link.
- The flow conservation constraints $b_i + \sum_{j=1}^{n} x_{ij} - \sum_{j=1}^{n} x_{ji} = 0$ ensure that the total incoming flow (including external supply or demand $b_i$) equals the total outgoing flow at each node $i$. Here:
  - $b_i > 0$ indicates external supply at node $i$, while $b_i < 0$ represents demand.
  - $\sum_{j=1}^{n} x_{ij}$ is the total outgoing flow from node $i$, and $\sum_{j=1}^{n} x_{ji}$ is the total incoming flow.

  Since $\sum_{i=1}^{n} b_i = 0$, the total supply equals total demand across the network.
- The flow bound constraints $l_{ij} \le x_{ij} \le u_{ij}$ ensure that each flow $x_{ij}$ remains within its specified lower and upper bounds, $l_{ij}$ and $u_{ij}$, for all $i, j$. This captures the capacity limitations of each link.

Thus, this linear program finds the cost-minimizing flow configuration across the network, while ensuring that all flow conservation and capacity constraints are met. Consider the $\ell_4$-norm approximation problem:

$$\text{minimize } \|Ax - b\|_4 = \left(\sum_{i=1}^{m}(a_i^T x - b_i)^4\right)^{1/4}$$

where the matrix $A \in \mathbb{R}^{m\times n}$ (with rows $a_i^T$) and the vector $b \in \mathbb{R}^m$ are given. The goal is to find a formulation of this problem as a quadratically constrained quadratic program (QCQP).

**Solution:**

1. **Understanding the Original Problem:**
   We start with the problem:

   $$\text{minimize } \|Ax - b\|_4 = \left(\sum_{i=1}^{m}(a_i^T x - b_i)^4\right)^{1/4}$$

   where $A \in \mathbb{R}^{m\times n}$ is a matrix with rows $a_i^T$, $b \in \mathbb{R}^m$ is a vector, and $x \in \mathbb{R}^n$ is the variable. This objective represents the $\ell_4$-norm of the residual vector $Ax - b$.

2. **Objective Transformation:**
   The objective is to minimize $\|Ax - b\|_4$, which can be expanded as:

$$\|Ax - b\|_4 = \left( \sum_{i=1}^{m} (a_i^T x - b_i)^4 \right)^{1/4}.$$

   Directly minimizing this $\ell_4$-norm is challenging due to the fourth power and the root. We can simplify by introducing auxiliary variables $y_i$ and $z_i$ for each residual term $(a_i^T x - b_i)^4$.

3. **Introducing Auxiliary Variables:**
   Define auxiliary variables $y_i$ such that:

$$y_i = a_i^T x - b_i, \quad i = 1, \ldots, m.$$

   Now, instead of working with $(a_i^T x - b_i)^4$, we can work with $y_i^4$.

4. **Formulating a Quadratic Constraint:**
   To avoid the fourth power, we introduce another auxiliary variable $z_i$ that approximates $y_i^2$. We enforce $z_i \geq y_i^2$, or equivalently:

$$y_i^2 \leq z_i, \quad i = 1, \ldots, m.$$

   This allows $z_i$ to act as an upper bound on $y_i^2$, so we can approximate the original $\ell_4$-norm objective by minimizing the sum $\sum_{i=1}^m z_i^2$, related to the fourth power of the deviations.

5. **Simplifying the Objective:**
   Since $z_i^2$ bounds $y_i^4$, we approximate the original objective $\|Ax - b\|_4$ by minimizing the sum of $z_i^2$ terms:

$$\text{minimize} \sum_{i=1}^{m} z_i^2.$$

   This objective is now quadratic, aligning with QCQP form.

6. **Formulating the Final QCQP Problem:**
   The problem can now be written as a QCQP with the following objective and constraints:

$$\begin{aligned} \text{minimize} \quad & \textstyle\sum_{i=1}^m z_i^2 \\ \text{subject to} \quad & a_i^T x - b_i = y_i, \quad i = 1, \ldots, m, \\ & y_i^2 \leq z_i, \quad i = 1, \ldots, m. \end{aligned}$$

# 2 Support Vector Machine

# Classification Problems with Hyperplanes

1. **Problem 1**
   Given the hyperplane defined by the line:

$$y = x_1 + x_2$$

   Determine if the following points are correctly classified:

   - $y = 1, x = (2, 1)$
   - $y = -1, x = (-1, -2)$

   **Solution:** The hyperplane equation is $y = w^T x = x_1 + x_2$.

   - For $y = 1$ and $x = (2, 1)$:

$$w^T x = 2 + 1 = 3$$

   Since $y = 1$ and $w^T x = 3 > 0$, this point is correctly predicted.

   - For $y = -1$ and $x = (-1, -2)$:

$$w^T x = -1 + (-2) = -3$$

   Since $y = -1$ and $w^T x = -3 < 0$, this point is correctly predicted.

2. **Problem 2**
Given the hyperplane defined by the line:

$$y = 2x_1 - x_2$$

Determine if the following points are correctly classified:

- $y = 1, x = (1, 1)$
- $y = -1, x = (-2, 4)$

**Solution:** The hyperplane equation is $y = w^T x = 2x_1 - x_2$.
- For $y = 1$ and $x = (1, 1)$:

$$w^T x = 2 \cdot 1 - 1 = 1$$

Since $y = 1$ and $w^T x = 1 > 0$, this point is correctly predicted.
- For $y = -1$ and $x = (-2, 4)$:

$$w^T x = 2 \cdot (-2) - 4 = -4 - 4 = -8$$

Since $y = -1$ and $w^T x = -8 < 0$, this point is correctly predicted.

3. **Problem 3**
Given the hyperplane defined by:

$$y = -x_1 + 3x_2$$

Determine if the following points are correctly classified:

- $y = 1, x = (0, 1)$
- $y = -1, x = (1, 2)$

**Solution:** The hyperplane equation is $y = w^T x = -x_1 + 3x_2$.
- For $y = 1$ and $x = (0, 1)$:

$$w^T x = -0 + 3 \cdot 1 = 3$$

Since $y = 1$ and $w^T x = 3 > 0$, this point is correctly predicted.
- For $y = -1$ and $x = (1, 2)$:

$$w^T x = -1 + 3 \cdot 2 = -1 + 6 = 5$$

Since $y = -1$ but $w^T x = 5 > 0$, this point is incorrectly predicted.

4. **Problem 4**
Given the hyperplane defined by:

$$y = 4x_1 - x_2$$

Determine if the following points are correctly classified:

- $y = 1, x = (1, -2)$
- $y = -1, x = (-1, 5)$

**Solution:** The hyperplane equation is $y = w^T x = 4x_1 - x_2$.
- For $y = 1$ and $x = (1, -2)$:

$$w^T x = 4 \cdot 1 - (-2) = 4 + 2 = 6$$

Since $y = 1$ and $w^T x = 6 > 0$, this point is correctly predicted.
- For $y = -1$ and $x = (-1, 5)$:

$$w^T x = 4 \cdot (-1) - 5 = -4 - 5 = -9$$

Since $y = -1$ and $w^T x = -9 < 0$, this point is correctly predicted.