Practice Set
MFML

MFML Sec9

August 26, 2025

Contents

1 Linear Algebra Fundamentals

This section covers core concepts of linear algebra that are foundational for machine learning, including matrix properties, vector spaces, and diagonalization. Questions are based on topics from the lecture slides and past exams.

Problem 1.1. Let $A = \begin{pmatrix} 4 & 2 & 12 \\ -2 & 1 & -6 \\ -12 & 6 & 36 \end{pmatrix}$.

(a) Find the rank of A.

(b) Check whether A is positive definite or not.

(c) Explain why one cannot use A to define an inner product.

Solution. (a) Rank of A

Step 1: Perform Row Reduction. We apply Gaussian elimination to transform A into its row-echelon form.

$$A = \begin{pmatrix} 4 & 2 & 12 \\ -2 & 1 & -6 \\ -12 & 6 & 36 \end{pmatrix} \xrightarrow{R_2 \leftarrow R_2 + \frac{1}{2}R_1} \begin{pmatrix} 4 & 2 & 12 \\ 0 & 2 & 0 \\ -12 & 6 & 36 \end{pmatrix}$$

$$\xrightarrow{R_3 \leftarrow R_3 + 3R_1} \begin{pmatrix} 4 & 2 & 12 \\ 0 & 2 & 0 \\ 0 & 12 & 72 \end{pmatrix}$$

$$\xrightarrow{R_3 \leftarrow R_3 - 6R_2} \begin{pmatrix} 4 & 2 & 12 \\ 0 & 2 & 0 \\ 0 & 0 & 72 \end{pmatrix}$$

Step 2: Conclusion. The resulting matrix in row-echelon form has three non-zero rows (pivot rows). Therefore, the **rank of A is 3**.

(b) Positive-Definiteness of A

Step 1: Check for Symmetry. A necessary condition for a matrix to be positive definite is that it must be symmetric (i.e., $A = A^T$).

Step 2: Inspect A. We observe that $A_{12} = 2$ but $A_{21} = -2$. Since $A_{12} \neq A_{21}$, the matrix is not symmetric.

Step 3: Conclusion. Since A is not symmetric, it **cannot be positive definite**.

(c) Why A cannot define an Inner Product

Step 1: Recall Inner Product Conditions. An inner product on $\mathbb{R}^n$ defined by a matrix M, $\langle x, y \rangle = x^T My$, requires that M be both symmetric and positive definite.

Step 2: Conclusion. As established in part (b), A is not symmetric and therefore cannot be used to define a valid inner product.

Problem 1.2. Given $b = \begin{pmatrix} -1 \\ 3 \\ 5 \end{pmatrix}$ and $A = bb^T$.

(a) Find a basis for the null space of A, $\{X \in \mathbb{R}^3 : AX = 0\}$.

(b) Find the eigenvalues of A.

Solution. (a) Basis for the Null Space

Step 1: Analyze the equation $AX = 0$. We have $AX = (bb^T)X = b(b^TX)$. Since b is a non-zero vector, $AX = 0$ if and only if the scalar product b^TX is zero. This means X must be orthogonal to b.

$$b^TX = \begin{pmatrix} -1 & 3 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = -x_1 + 3x_2 + 5x_3 = 0$$

Step 2: Express the general solution for X. From the equation, we can write $x_1 = 3x_2 + 5x_3$. The variables x_2 and x_3 are free parameters. A vector X in the null space has the form:

$$X = \begin{pmatrix} 3x_2 + 5x_3 \\ x_2 \\ x_3 \end{pmatrix} = x_2 \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} 5 \\ 0 \\ 1 \end{pmatrix}$$

Step 3: Identify the basis vectors. The two vectors that span the null space are linearly independent and form a basis.

$$\textbf{Basis} = \left\{ \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 5 \\ 0 \\ 1 \end{pmatrix} \right\}$$

(b) Eigenvalues of A

Step 1: Use the Rank and Nullity. The matrix $A = bb^T$ is a rank-one matrix. For a 3×3 matrix, the Rank-Nullity Theorem states that $\text{rank}(A) + \text{nullity}(A) = 3$. Since the rank is 1, the nullity (dimension of the null space) is 2. The nullity corresponds to the geometric multiplicity of the eigenvalue $\lambda = 0$. Thus, **two eigenvalues are 0**.

$$\lambda_2 = 0, \quad \lambda_3 = 0$$

Step 2: Find the non-zero eigenvalue. The vector b itself is an eigenvector. Let's check:

$$Ab = (bb^T)b = b(b^Tb) = (b^Tb)b$$

This is of the form $Ab = \lambda b$, where the eigenvalue is $\lambda_1 = b^Tb$.

$$\lambda_1 = b^Tb = (-1)^2 + 3^2 + 5^2 = 1 + 9 + 25 = 35$$

Step 3: Conclusion. The eigenvalues of A are $\{35, 0, 0\}$.

Problem 1.3. Determine whether the matrix $A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 1 \\ 1 & 1 & -1 \end{pmatrix}$ is diagonalizable.

Solution.

Step 1: Find the characteristic polynomial, $\det(A - \lambda I) = 0$**.**

$$\begin{aligned}\det(A - \lambda I) &= \begin{vmatrix} 1-\lambda & 0 & 0 \\ 1 & 2-\lambda & 1 \\ 1 & 1 & -1-\lambda \end{vmatrix} \\ &= (1-\lambda)\left[(2-\lambda)(-1-\lambda) - (1)(1)\right] \\ &= (1-\lambda)[-2-2\lambda+\lambda+\lambda^2-1] \\ &= (1-\lambda)(\lambda^2-\lambda-3)\end{aligned}$$

Step 2: Find the eigenvalues. The eigenvalues are the roots of this polynomial.

- From $(1-\lambda) = 0$, we get $\lambda_1 = 1$.
- From $(\lambda^2 - \lambda - 3) = 0$, using the quadratic formula:

$$\lambda = \frac{-(-1) \pm \sqrt{(-1)^2 - 4(1)(-3)}}{2(1)} = \frac{1 \pm \sqrt{1+12}}{2} = \frac{1 \pm \sqrt{13}}{2}$$

So, $\lambda_2 = \frac{1+\sqrt{13}}{2}$ and $\lambda_3 = \frac{1-\sqrt{13}}{2}$.

Step 3: Check for diagonalizability. An $n \times n$ matrix is diagonalizable if it has n distinct eigenvalues. Here, A is a 3×3 matrix, and we have found three distinct real eigenvalues: 1, $\frac{1+\sqrt{13}}{2}$, and $\frac{1-\sqrt{13}}{2}$.

Step 4: Conclusion. Since A has three distinct real eigenvalues, it is **diagonalizable**.

Problem 1.4. Show that the set $S = \{(x, y)^T : -2 \leq x \leq 2, -2 \leq y \leq 2\} \subseteq \mathbb{R}^2$ is not a subspace of $\mathbb{R}^2$.

Solution.

Step 1: Recall Subspace Conditions. For S to be a subspace of $\mathbb{R}^2$, it must satisfy three conditions:

(i) It must contain the zero vector, $\mathbf{0}$.

(ii) It must be closed under vector addition.

(iii) It must be closed under scalar multiplication.

Step 2: Test Closure under Addition. We only need to show that one of the conditions fails. Let's test closure under addition with a counterexample.

- Let vector $u = (2, 2)^T$. Since $-2 \leq 2 \leq 2$, $u \in S$.
- Let vector $v = (2, 2)^T$. Since $-2 \leq 2 \leq 2$, $v \in S$.
- Now, compute their sum: $u + v = (2+2, 2+2)^T = (4, 4)^T$.

Step 3: Conclusion. The resulting vector $(4, 4)^T$ is not in S because its components are greater than 2. Since S is not closed under vector addition, it is **not a subspace** of $\mathbb{R}^2$.

Problem 1.5. Do the vectors $v_1 = (1, -1, 1)^T$, $v_2 = (0, 2, 1)^T$, and $v_3 = (-1, 0, 1)^T$ span $\mathbb{R}^3$? Justify your answer.

Solution.

Step 1: State the Condition for Spanning $\mathbb{R}^3$. Three vectors in $\mathbb{R}^3$ span the entire space if and only if they are linearly independent. We can check for linear independence by forming a matrix with these vectors as columns and computing its determinant. If the determinant is non-zero, the vectors are linearly independent.

Step 2: Form the Matrix M.

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

Step 3: Compute the Determinant. We can use the cofactor expansion along the first row:

$$\begin{aligned} \det(M) &= 1 \cdot \begin{vmatrix} 2 & 0 \\ 1 & 1 \end{vmatrix} - 0 \cdot \begin{vmatrix} -1 & 0 \\ 1 & 1 \end{vmatrix} + (-1) \cdot \begin{vmatrix} -1 & 2 \\ 1 & 1 \end{vmatrix} \\ &= 1 \cdot (2 \cdot 1 - 0 \cdot 1) - 0 + (-1) \cdot ((-1) \cdot 1 - 2 \cdot 1) \\ &= 1 \cdot (2) - 1 \cdot (-1 - 2) \\ &= 2 - (-3) = 5 \end{aligned}$$

Step 4: Conclusion. Since $\det(M) = 5 \neq 0$, the vectors are linearly independent. Therefore, they form a basis for $\mathbb{R}^3$ and **span** $\mathbb{R}^3$.

Problem 1.6. Consider the quadratic function $f(x, y) = 2x^2 - 2xy + y^2 + 3x - 5y$. Determine if this function is convex by analyzing its Hessian matrix.

Solution.

Step 1: Recall the Convexity Condition. A twice-differentiable function is convex if and only if its Hessian matrix is positive semidefinite (PSD) everywhere. For a quadratic function, the Hessian is a constant matrix, so we just need to check if that matrix is PSD.

Step 2: Compute the Gradient. First, we find the first-order partial derivatives.

$$\begin{aligned} \frac{\partial f}{\partial x} &= 4x - 2y + 3 \\ \frac{\partial f}{\partial y} &= -2x + 2y - 5 \end{aligned}$$

Step 3: Compute the Hessian Matrix H. Next, we find the second-order partial derivatives.

$$\begin{aligned} \frac{\partial^2 f}{\partial x^2} &= 4 \\ \frac{\partial^2 f}{\partial y \partial x} &= -2 \\ \frac{\partial^2 f}{\partial x \partial y} &= -2 \\ \frac{\partial^2 f}{\partial y^2} &= 2 \end{aligned}$$

The Hessian is $H = \begin{pmatrix} 4 & -2 \\ -2 & 2 \end{pmatrix}$.

Step 4: Check if H is Positive Semidefinite. A symmetric matrix is PSD if all its eigenvalues are non-negative. Let's find the eigenvalues of H.

$$\begin{aligned}\det(H-\lambda I) = \begin{vmatrix} 4-\lambda & -2 \\ -2 & 2-\lambda \end{vmatrix} &= 0 \\ (4-\lambda)(2-\lambda) - 4 &= 0 \\ 8 - 6\lambda + \lambda^2 - 4 &= 0 \\ \lambda^2 - 6\lambda + 4 &= 0\end{aligned}$$

Using the quadratic formula:

$$\lambda = \frac{6 \pm \sqrt{36 - 4(4)}}{2} = \frac{6 \pm \sqrt{20}}{2} = 3 \pm \sqrt{5}$$

The eigenvalues are $\lambda_1 = 3 + \sqrt{5} \approx 5.236$ and $\lambda_2 = 3 - \sqrt{5} \approx 0.764$.

Step 5: Conclusion. Both eigenvalues are positive. Therefore, the Hessian matrix is positive definite (and also positive semidefinite). This proves that the function $f(x, y)$ is **convex**.

Problem 1.7. Let an operation be defined in $\mathbb{R}^2$ as $\langle u, v\rangle = u^T A v$, where $u, v \in \mathbb{R}^2$ and $A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$. Show that this operation defines a valid inner product.

Solution. To be a valid inner product, the operation must satisfy three axioms: symmetry, linearity, and positive definiteness.

Step 1: Check Symmetry: $\langle u, v\rangle = \langle v, u\rangle$. This axiom holds if and only if the matrix A is symmetric ($A = A^T$).

$$A^T = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}^T = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} = A$$

Since A is symmetric, the symmetry axiom holds.

Step 2: Check Linearity. This axiom is guaranteed because matrix multiplication is a linear operation. For any scalars c_1, c_2 and vectors u, v, w:

$$\langle c_1 u + c_2 v, w\rangle = (c_1 u + c_2 v)^T A w = c_1 u^T A w + c_2 v^T A w = c_1\langle u, w\rangle + c_2\langle v, w\rangle$$

The linearity axiom holds.

Step 3: Check Positive Definiteness: $\langle u, u\rangle \geq 0$ and $\langle u, u\rangle = 0 \iff u = \mathbf{0}$. This requires the matrix A to be positive definite. A symmetric matrix is positive definite if all its eigenvalues are strictly positive. Let's find the eigenvalues of A.

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

Using the quadratic formula:

$$\lambda = \frac{3 \pm \sqrt{9 - 4}}{2} = \frac{3 \pm \sqrt{5}}{2}$$

The eigenvalues are $\lambda_1 = \frac{3+\sqrt{5}}{2} > 0$ and $\lambda_2 = \frac{3-\sqrt{5}}{2} > 0$.

Step 4: Conclusion. Since both eigenvalues are strictly positive, the matrix A is positive definite. All three axioms are satisfied, so the operation defines a **valid inner product** on $\mathbb{R}^2$.

Problem 1.8. Find the eigenvalues and corresponding eigenvectors of the matrix $A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$.

Solution.

Step 1: Find the Characteristic Polynomial.

$$\begin{aligned} \det(A - \lambda I) &= \begin{vmatrix} 2-\lambda & -1 & 0 \\ -1 & 2-\lambda & 0 \\ 0 & 0 & 3-\lambda \end{vmatrix} = 0 \\ &= (3-\lambda) \begin{vmatrix} 2-\lambda & -1 \\ -1 & 2-\lambda \end{vmatrix} \\ &= (3-\lambda)[(2-\lambda)^2 - 1] \\ &= (3-\lambda)[\lambda^2 - 4\lambda + 4 - 1] \\ &= (3-\lambda)(\lambda^2 - 4\lambda + 3) = 0 \end{aligned}$$

Step 2: Solve for Eigenvalues. We can factor the quadratic part:

$$(3-\lambda)(\lambda-3)(\lambda-1) = 0$$

The eigenvalues are $\lambda_1 = 1$, and $\lambda_2 = 3$ (with an algebraic multiplicity of 2).

Step 3: Find Eigenvector for $\lambda_1 = 1$. Solve $(A - I)v = 0$:

$$\begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

This gives $v_1 - v_2 = 0 \implies v_1 = v_2$, and $2v_3 = 0 \implies v_3 = 0$. An eigenvector is $v_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$.

Step 4: Find Eigenvectors for $\lambda_2 = 3$. Solve $(A - 3I)v = 0$:

$$\begin{pmatrix} -1 & -1 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

This gives $-v_1 - v_2 = 0 \implies v_1 = -v_2$. The variable v_3 is free. We can find two linearly independent eigenvectors.

- Let $v_2 = 1, v_3 = 0 \implies v_1 = -1$. Eigenvector is $v_2 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$.
- Let $v_2 = 0, v_3 = 1 \implies v_1 = 0$. Eigenvector is $v_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$.

Step 5: Conclusion. The eigenvalues are $\{1, 3, 3\}$. The corresponding eigenvectors are $\{(1, 1, 0)^T\}$, and $\{(-1, 1, 0)^T, (0, 0, 1)^T\}$.

Problem 1.9. Solve the following system of linear equations using Gaussian elimination:

$$\begin{aligned} x + y + z &= 6 \\ 2x - y + z &= 3 \\ x + 2y - z &= 4 \end{aligned}$$

Solution.

Step 1: Form the Augmented Matrix.

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 6 \\ 2 & -1 & 1 & 3 \\ 1 & 2 & -1 & 4 \end{array}\right]$$

Step 2: Perform Forward Elimination.

- $R_2 \leftarrow R_2 - 2R_1$:
$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 6 \\ 0 & -3 & -1 & -9 \\ 1 & 2 & -1 & 4 \end{array}\right]$$
- $R_3 \leftarrow R_3 - R_1$:
$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 6 \\ 0 & -3 & -1 & -9 \\ 0 & 1 & -2 & -2 \end{array}\right]$$
- Swap R_2 and R_3 for a nicer pivot:
$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 6 \\ 0 & 1 & -2 & -2 \\ 0 & -3 & -1 & -9 \end{array}\right]$$
- $R_3 \leftarrow R_3 + 3R_2$:
$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 6 \\ 0 & 1 & -2 & -2 \\ 0 & 0 & -7 & -15 \end{array}\right]$$

Step 3: Perform Back Substitution. The system is now:

$$\begin{aligned} x + y + z &= 6 \\ y - 2z &= -2 \\ -7z &= -15 \end{aligned}$$

- From the third equation: $z = 15/7$.
- Substitute into the second equation: $y - 2(15/7) = -2 \implies y = -2 + 30/7 = 16/7$.
- Substitute into the first equation: $x + 16/7 + 15/7 = 6 \implies x + 31/7 = 42/7 \implies x = 11/7$.

Step 4: Conclusion. The unique solution is $x = 11/7$, $y = 16/7$, $z = 15/7$.

Problem 1.10. Let $A = \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}$ and $B = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$. Compute $(A+B)^2$ and $A^2 + 2AB + B^2$ and determine if they are equal.

Solution.

Step 1: Compute $A + B$.

$$A + B = \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 1 & 4 \end{pmatrix}$$

Step 2: Compute $(A + B)^2$.

$$(A+B)^2 = \begin{pmatrix} 2 & 2 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} 2 & 2 \\ 1 & 4 \end{pmatrix} = \begin{pmatrix} 2(2)+2(1) & 2(2)+2(4) \\ 1(2)+4(1) & 1(2)+4(4) \end{pmatrix} = \begin{pmatrix} 6 & 12 \\ 6 & 18 \end{pmatrix}$$

Step 3: Compute A^2, B^2, and AB.

$$A^2 = \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 8 \\ 0 & 9 \end{pmatrix}$$

$$B^2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}$$

$$AB = \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1(1)+2(1) & 1(0)+2(1) \\ 0(1)+3(1) & 0(0)+3(1) \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ 3 & 3 \end{pmatrix}$$

Step 4: Compute $A^2 + 2AB + B^2$.

$$A^2+2AB+B^2 = \begin{pmatrix} 1 & 8 \\ 0 & 9 \end{pmatrix} + 2\begin{pmatrix} 3 & 2 \\ 3 & 3 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 8 \\ 0 & 9 \end{pmatrix} + \begin{pmatrix} 6 & 4 \\ 6 & 6 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 8 & 12 \\ 8 & 16 \end{pmatrix}$$

Step 5: Conclusion. $(A+B)^2 = \begin{pmatrix} 6 & 12 \\ 6 & 18 \end{pmatrix}$ is **not equal** to $A^2 + 2AB + B^2 = \begin{pmatrix} 8 & 12 \\ 8 & 16 \end{pmatrix}$. This is because matrix multiplication is not commutative ($AB \neq BA$), so the standard binomial expansion does not hold.

2 Principal Component Analysis (PCA)

PCA is a fundamental technique for dimensionality reduction. The problems in this section focus on calculating the covariance matrix and finding the principal components, which correspond to the eigenvectors of the covariance matrix.

Problem 2.1. The observed data are: (x_k, y_k) for $k = 1, 2, 3, 4$ given by $(2, 8), (5, 13), (7, 19), (10, 25)$.

(a) Compute the sample covariance matrix Σ.

(b) Describe the first principal component. (You may use the fact that the normalized eigenvector for the largest eigenvalue is approximately $v \approx (0.414, 0.909)^T$).

Solution. (a) Covariance Matrix

Step 1: Compute the means.

$$\bar{x} = \frac{2+5+7+10}{4} = \frac{24}{4} = 6$$
$$\bar{y} = \frac{8+13+19+25}{4} = \frac{65}{4} = 16.25$$

Step 2: Compute the centered data vectors. Let $X_c = X - \bar{x}$ and $Y_c = Y - \bar{y}$.

$$X_c = (2-6, 5-6, 7-6, 10-6)^T = (-4, -1, 1, 4)^T$$
$$Y_c = (8-16.25, 13-16.25, 19-16.25, 25-16.25)^T = (-8.25, -3.25, 2.75, 8.75)^T$$

Step 3: Calculate sample variances and covariance $(1/(N-1))$.

$$S_{xx} = \frac{1}{3}\sum X_{c,i}^2 = \frac{(-4)^2 + (-1)^2 + 1^2 + 4^2}{3} = \frac{16+1+1+16}{3} = \frac{34}{3} \approx 11.33$$
$$S_{yy} = \frac{1}{3}\sum Y_{c,i}^2 = \frac{(-8.25)^2 + (-3.25)^2 + 2.75^2 + 8.75^2}{3} = \frac{162.75}{3} = 54.25$$
$$S_{xy} = \frac{1}{3}\sum X_{c,i}Y_{c,i} = \frac{(-4)(-8.25) + (-1)(-3.25) + (1)(2.75) + (4)(8.75)}{3} = \frac{74}{3} \approx 24.67$$

Step 4: Assemble the covariance matrix.

$$\Sigma = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{xy} & S_{yy} \end{pmatrix} = \begin{pmatrix} 34/3 & 74/3 \\ 74/3 & 162.75/3 \end{pmatrix} \approx \begin{pmatrix} 11.33 & 24.67 \\ 24.67 & 54.25 \end{pmatrix}$$

(b) First Principal Component

Step 1: Definition. The first principal component (PC1) is the direction of maximum variance, defined by the eigenvector corresponding to the largest eigenvalue of the covariance matrix Σ.

Step 2: Identify the PC1 direction. The problem gives this eigenvector as $v_1 \approx (0.414, 0.909)^T$. This vector defines the principal axis.

Step 3: Define the projection (score). The score of a centered data point $(x_k - \bar{x}, y_k - \bar{y})$ on PC1 is its projection onto this vector:

$$\text{PC1}_{\text{score}}(k) = v_1^T \begin{pmatrix} x_k - \bar{x} \\ y_k - \bar{y} \end{pmatrix} = 0.414(x_k - 6) + 0.909(y_k - 16.25)$$

This formula provides the 1D representation of the original 2D data.

Problem 2.2. You are given a centered 3D dataset with the following sample covariance matrix: $\Sigma = \begin{pmatrix} 13 & 8 & 8 \\ 8 & 13 & 8 \\ 8 & 8 & 13 \end{pmatrix}$.

(a) Find all eigenvalues of Σ.

(b) Find the normalized eigenvector corresponding to the largest eigenvalue. This is the first principal component (PC1).

(c) Find the percentage of total variance explained by the first principal component.

Solution. (a) Finding Eigenvalues

Step 1: Set up the characteristic equation $\det(\Sigma - \lambda I) = 0$**.**

$$\begin{vmatrix} 13-\lambda & 8 & 8 \\ 8 & 13-\lambda & 8 \\ 8 & 8 & 13-\lambda \end{vmatrix} = 0$$

Step 2: Simplify the determinant. Perform the row operation $R_1 \leftarrow R_1 - R_2$:

$$\begin{vmatrix} 5-\lambda & -(5-\lambda) & 0 \\ 8 & 13-\lambda & 8 \\ 8 & 8 & 13-\lambda \end{vmatrix} = (5-\lambda)\begin{vmatrix} 1 & -1 & 0 \\ 8 & 13-\lambda & 8 \\ 8 & 8 & 13-\lambda \end{vmatrix} = 0$$

This shows that $\lambda_1 = 5$ is one eigenvalue.

Step 3: Solve the remaining determinant.

$$\begin{aligned} 1((13-\lambda)^2 - 64) - (-1)(8(13-\lambda) - 64) &= 0 \\ (\lambda^2 - 26\lambda + 169 - 64) + (104 - 8\lambda - 64) &= 0 \\ \lambda^2 - 34\lambda + 145 &= 0 \\ (\lambda - 5)(\lambda - 29) &= 0 \end{aligned}$$

This gives the other two roots: $\lambda_2 = 5$ and $\lambda_3 = 29$.

Step 4: Conclusion. The eigenvalues are $\{29, 5, 5\}$.

(b) Finding PC1

Step 1: Identify the largest eigenvalue. The largest eigenvalue is $\lambda_{max} = 29$. PC1 is its corresponding eigenvector.

Step 2: Solve $(\Sigma - 29I)v = 0$**.**

$$\begin{pmatrix} -16 & 8 & 8 \\ 8 & -16 & 8 \\ 8 & 8 & -16 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

The first row gives $-16v_1 + 8v_2 + 8v_3 = 0 \implies -2v_1 + v_2 + v_3 = 0$. The second row gives $8v_1 - 16v_2 + 8v_3 = 0 \implies v_1 - 2v_2 + v_3 = 0$. Subtracting the second simplified equation from the first gives: $(-2v_1 - v_1) + (v_2 - (-2v_2)) + (v_3 - v_3) = 0 \implies -3v_1 + 3v_2 = 0 \implies v_1 = v_2$. Substituting $v_1 = v_2$ into $-2v_1 + v_2 + v_3 = 0$ gives $-2v_1 + v_1 + v_3 = 0 \implies v_3 = v_1$. The eigenvector is of the form $k(1,1,1)^T$. Let's choose $v = (1,1,1)^T$.

Step 3: Normalize the eigenvector. Length $\|v\| = \sqrt{1^2 + 1^2 + 1^2} = \sqrt{3}$.

$$v_{PC1} = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

(c) Percentage of Variance Explained

Step 1: Calculate the total variance. The total variance is the sum of the eigenvalues.

$$\text{Total Variance} = 29 + 5 + 5 = 39$$

Step 2: Calculate the percentage explained by PC1.

$$\text{Percentage} = \frac{\lambda_{max}}{\text{Total Variance}} \times 100\% = \frac{29}{39} \times 100\% \approx 74.36\%$$

Problem 2.3. A 2D dataset has a covariance matrix $\Sigma = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}$. Find the principal components (eigenvectors) and their corresponding variances (eigenvalues).

Solution.

Step 1: Find the Eigenvalues (Variances).

$$\begin{aligned} \det(\Sigma - \lambda I) = \begin{vmatrix} 5-\lambda & 2 \\ 2 & 2-\lambda \end{vmatrix} &= 0 \\ (5-\lambda)(2-\lambda) - 4 &= 0 \\ \lambda^2 - 7\lambda + 10 - 4 &= 0 \\ \lambda^2 - 7\lambda + 6 &= 0 \\ (\lambda - 6)(\lambda - 1) &= 0 \end{aligned}$$

The eigenvalues (variances) are $\lambda_1 = 6$ and $\lambda_2 = 1$.

Step 2: Find the First Principal Component (Eigenvector for $\lambda_1 = 6$).

$$(\Sigma - 6I)v = \begin{pmatrix} -1 & 2 \\ 2 & -4 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

This gives $-v_1 + 2v_2 = 0 \implies v_1 = 2v_2$. Let $v_2 = 1$, then $v_1 = 2$. The unnormalized eigenvector is $(2, 1)^T$. Normalizing gives $v_{PC1} = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 \\ 1 \end{pmatrix}$.

Step 3: Find the Second Principal Component (Eigenvector for $\lambda_2 = 1$).

$$(\Sigma - 1I)v = \begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

This gives $2v_1 + v_2 = 0 \implies v_2 = -2v_1$. Let $v_1 = 1$, then $v_2 = -2$. The unnormalized eigenvector is $(1, -2)^T$. Normalizing gives $v_{PC2} = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ -2 \end{pmatrix}$.

Problem 2.4. Explain the relationship between the singular value decomposition (SVD) of a centered data matrix $X \in \mathbb{R}^{D \times N}$ and the eigendecomposition of its covariance matrix $S = \frac{1}{N} X X^T$.

Solution.

Step 1: Define the SVD of X. Let the SVD of the centered data matrix X be $X = U\Sigma V^T$, where:

- $U \in \mathbb{R}^{D\times D}$ is an orthogonal matrix whose columns are the left singular vectors.
- $\Sigma \in \mathbb{R}^{D\times N}$ is a rectangular diagonal matrix of singular values σ_i.
- $V \in \mathbb{R}^{N\times N}$ is an orthogonal matrix whose columns are the right singular vectors.

Step 2: Substitute the SVD into the Covariance Matrix Formula.

$$S = \frac{1}{N}XX^T = \frac{1}{N}(U\Sigma V^T)(U\Sigma V^T)^T = \frac{1}{N}(U\Sigma V^T)(V\Sigma^T U^T)$$

Step 3: Simplify the Expression. Since V is orthogonal, $V^TV = I$.

$$S = \frac{1}{N}U\Sigma(V^TV)\Sigma^T U^T = \frac{1}{N}U(\Sigma\Sigma^T)U^T$$

Step 4: Interpret the Result. The expression $S = U(\frac{1}{N}\Sigma\Sigma^T)U^T$ is the eigendecomposition of S.

- The columns of U (the left singular vectors of X) are the eigenvectors of the covariance matrix S. These are the **principal components**.
- The matrix $\frac{1}{N}\Sigma\Sigma^T$ is a diagonal matrix. Its diagonal entries are the eigenvalues of S. The eigenvalue λ_i is related to the singular value σ_i by $\lambda_i = \frac{\sigma_i^2}{N}$.

Problem 2.5. You are performing PCA on a dataset with $D = 10,000$ dimensions (features) and $N = 500$ data points. Why is it computationally inefficient to compute the covariance matrix $S = \frac{1}{N}XX^T$? What is the more efficient alternative?

Solution.

Step 1: Analyze the Inefficiency. The centered data matrix X has dimensions $D \times N$, which is 10000×500.

- The covariance matrix $S = \frac{1}{N}XX^T$ has dimensions $(D \times N) \times (N \times D) = D \times D$.
- In this case, S would be a 10000×10000 matrix.
- Computing the eigendecomposition of such a large matrix is computationally very expensive, typically scaling cubically with the dimension D. This is infeasible.

Step 2: Propose the Efficient Alternative. For the case where $N \ll D$, we can work with a much smaller matrix. Instead of XX^T, we consider the $N \times N$ matrix X^TX.

Step 3: Relate the Eigendecompositions. Let c_m be an eigenvector of $\frac{1}{N}X^TX$ with eigenvalue λ_m.

$$\frac{1}{N}(X^TX)c_m = \lambda_m c_m$$

Left-multiplying by X gives:

$$\frac{1}{N}X(X^TX)c_m = \lambda_m Xc_m \implies \frac{1}{N}(XX^T)(Xc_m) = \lambda_m(Xc_m)$$

Step 4: Conclusion. This shows that if c_m is an eigenvector of the smaller $N \times N$ matrix, then $b_m = Xc_m$ is an eigenvector of the large $D \times D$ covariance matrix S with the same eigenvalue λ_m. We can efficiently find the eigenvectors of the 500×500 matrix $X^T X$ and then recover the required eigenvectors (principal components) of S via a simple matrix-vector multiplication.

Problem 2.6. PCA was performed on a dataset of university students with three standardized features: 'Study Hours', 'Exam Score', and 'Party Hours'. The first principal component (PC1) was found to be $v_1 = (0.6, 0.7, -0.4)^T$.

(a) What are the factor loadings for PC1?

(b) Provide a meaningful interpretation of what this principal component represents.

Solution. (a) Factor Loadings

Step 1: Definition. The factor loadings are the individual components of the principal component vector. They represent the correlation between the original variables and the principal component.

Step 2: Identify Loadings. The loadings for PC1 are:

- Loading on 'Study Hours': **0.6**
- Loading on 'Exam Score': **0.7**
- Loading on 'Party Hours': **-0.4**

(b) Interpretation of PC1

Step 1: Analyze the Signs and Magnitudes.

- 'Study Hours' and 'Exam Score' have strong positive loadings (0.6 and 0.7). This indicates that students with high scores on PC1 tend to have high values for both study hours and exam scores.
- 'Party Hours' has a moderate negative loading (-0.4). This indicates that students with high scores on PC1 tend to have low values for party hours.

Step 2: Synthesize the Meaning. The first principal component captures the primary axis of variation in the data. In this case, it represents a "studiousness" or "academic diligence" dimension. A high score on PC1 corresponds to a diligent student who studies a lot, gets high scores, and parties less. A low (or large negative) score on PC1 would correspond to a less academically focused student who parties more, studies less, and has lower exam scores.

Problem 2.7. PCA is performed on a 5-dimensional dataset. The eigenvalues of the covariance matrix are found to be: $\lambda_1 = 8.0, \lambda_2 = 3.5, \lambda_3 = 1.0, \lambda_4 = 0.4, \lambda_5 = 0.1$.

(a) What percentage of the total variance is explained by the first two principal components?

(b) How many principal components should be retained to explain at least 95% of the total variance?

Solution. (a) Variance explained by first two PCs

Step 1: Calculate Total Variance. The total variance is the sum of all eigenvalues.

$$\text{Total Variance} = 8.0 + 3.5 + 1.0 + 0.4 + 0.1 = 13.0$$

Step 2: Calculate Variance Explained by PC1 and PC2. The variance explained by the first two components is the sum of their corresponding eigenvalues.

$$\text{Variance}_{\text{PC1+PC2}} = \lambda_1 + \lambda_2 = 8.0 + 3.5 = 11.5$$

Step 3: Calculate the Percentage.

$$\text{Percentage} = \frac{\text{Variance}_{\text{PC1+PC2}}}{\text{Total Variance}} \times 100\% = \frac{11.5}{13.0} \times 100\% \approx 88.46\%$$

(b) Number of components for 95% variance

Step 1: Calculate Cumulative Variance Explained. We calculate the cumulative sum of eigenvalues and the corresponding percentage of total variance.

- PC1: $8.0/13.0 = 61.54\%$
- PC1+PC2: $(8.0 + 3.5)/13.0 = 11.5/13.0 = 88.46\%$
- PC1+PC2+PC3: $(11.5 + 1.0)/13.0 = 12.5/13.0 = 96.15\%$

Step 2: Conclusion. To explain at least 95% of the total variance, we need to retain the **first three principal components**, as they cumulatively explain 96.15% of the variance.

Problem 2.8. A centered 2D data point is $x = (4, -2)^T$. After PCA, it is projected onto the first principal component $v_1 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^T$.

(a) Find the 1D representation (score) of the point.

(b) Find the reconstructed 2D point $\tilde{x}$.

(c) Calculate the squared reconstruction error $\|x - \tilde{x}\|^2$.

Solution. (a) Find the Score

Step 1: Project x onto v_1. The score z is the dot product of the data point with the principal component vector.

$$z = v_1^T x = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 4 \\ -2 \end{pmatrix} = \frac{4}{\sqrt{2}} - \frac{2}{\sqrt{2}} = \frac{2}{\sqrt{2}} = \sqrt{2}$$

(b) Reconstruct the 2D point

Step 1: Use the reconstruction formula $\tilde{x} = z \cdot v_1$.

$$\tilde{x} = \sqrt{2} \cdot \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

(c) Calculate Squared Reconstruction Error

Step 1: Calculate the difference vector $x - \tilde{x}$.

$$x - \tilde{x} = \begin{pmatrix} 4 \\ -2 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ -3 \end{pmatrix}$$

Step 2: Calculate the squared norm of the difference vector.

$$\|x - \tilde{x}\|^2 = 3^2 + (-3)^2 = 9 + 9 = 18$$

Problem 2.9. Given the 2D dataset: $X_1 = (1, 2)$, $X_2 = (3, 3)$, $X_3 = (5, 7)$. Perform a complete PCA.

(a) Center the data.

(b) Compute the sample covariance matrix.

(c) Find the first principal component (the normalized eigenvector for the largest eigenvalue).

(d) Calculate the PC1 score for each of the centered data points.

Solution. (a) Center the Data

Step 1: Compute the means.

$$\bar{x} = \frac{1+3+5}{3} = \frac{9}{3} = 3$$
$$\bar{y} = \frac{2+3+7}{3} = \frac{12}{3} = 4$$

Step 2: Subtract the means.

$$X_1' = (1-3, 2-4) = (-2, -2)$$
$$X_2' = (3-3, 3-4) = (0, -1)$$
$$X_3' = (5-3, 7-4) = (2, 3)$$

(b) Compute Sample Covariance Matrix

Step 1: Calculate variances and covariance $(1/(N-1))$.

$$S_{xx} = \frac{1}{2}[(-2)^2 + 0^2 + 2^2] = \frac{4+0+4}{2} = 4$$
$$S_{yy} = \frac{1}{2}[(-2)^2 + (-1)^2 + 3^2] = \frac{4+1+9}{2} = 7$$
$$S_{xy} = \frac{1}{2}[(-2)(-2) + (0)(-1) + (2)(3)] = \frac{4+0+6}{2} = 5$$

Step 2: Assemble the matrix. $\Sigma = \begin{pmatrix} 4 & 5 \\ 5 & 7 \end{pmatrix}$.

(c) Find the First Principal Component

Step 1: Find the eigenvalues.

$$\det(\Sigma - \lambda I) = \begin{vmatrix} 4-\lambda & 5 \\ 5 & 7-\lambda \end{vmatrix} = (4-\lambda)(7-\lambda) - 25 = 0$$
$$\lambda^2 - 11\lambda + 28 - 25 = 0 \implies \lambda^2 - 11\lambda + 3 = 0$$
$$\lambda = \frac{11 \pm \sqrt{121 - 12}}{2} = \frac{11 \pm \sqrt{109}}{2}$$

The largest eigenvalue is $\lambda_1 = \frac{11+\sqrt{109}}{2} \approx 10.72$.

Step 2: Find the eigenvector for λ_1. Solve $(\Sigma - \lambda_1 I)v = 0$.

$$(4 - 10.72)v_1 + 5v_2 = 0 \implies -6.72v_1 + 5v_2 = 0 \implies v_2 \approx 1.344v_1$$

Let $v_1 = 1$, then $v_2 = 1.344$. Unnormalized vector is $(1, 1.344)^T$.

Step 3: Normalize. $\|v\| = \sqrt{1^2 + 1.344^2} \approx 1.676$.

$$v_{PC1} = \frac{1}{1.676}\begin{pmatrix} 1 \\ 1.344 \end{pmatrix} \approx \begin{pmatrix} 0.597 \\ 0.802 \end{pmatrix}$$

(d) Calculate PC1 Scores

Step 1: Project centered data onto PC1. Score $= (v_{PC1})^T X'$.

$$\text{Score}_1 = (0.597)(-2) + (0.802)(-2) = -1.194 - 1.604 = \textbf{-2.798}$$
$$\text{Score}_2 = (0.597)(0) + (0.802)(-1) = \textbf{-0.802}$$
$$\text{Score}_3 = (0.597)(2) + (0.802)(3) = 1.194 + 2.406 = \textbf{3.600}$$

Problem 2.10. A dataset was reduced to one dimension using PCA. The first principal component is $v_1 = (0.6, 0.8)^T$. The PC1 scores for three centered points are $z_1 = -5$, $z_2 = 0$, $z_3 = 5$. Reconstruct the approximate 2D coordinates of these centered data points.

Solution.

Step 1: Recall the Reconstruction Formula. The reconstructed data point in the original high-dimensional space is given by $\tilde{x} = Bz$, where B is the matrix of principal components and z is the low-dimensional representation (the score). Here, $B = v_1$.

$$\tilde{x}' = v_1 \cdot z$$

Step 2: Reconstruct Point 1 ($z_1 = -5$).

$$\tilde{x}'_1 = \begin{pmatrix} 0.6 \\ 0.8 \end{pmatrix}(-5) = \begin{pmatrix} -3.0 \\ -4.0 \end{pmatrix}$$

Step 3: Reconstruct Point 2 ($z_2 = 0$).

$$\tilde{x}'_2 = \begin{pmatrix} 0.6 \\ 0.8 \end{pmatrix}(0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Step 4: Reconstruct Point 3 ($z_3 = 5$).

$$\tilde{x}'_3 = \begin{pmatrix} 0.6 \\ 0.8 \end{pmatrix}(5) = \begin{pmatrix} 3.0 \\ 4.0 \end{pmatrix}$$

Step 5: Conclusion. The reconstructed 2D coordinates for the centered data points are approximately $(-3, -4)$, $(0, 0)$, and $(3, 4)$.

3 Optimization (Unconstrained)

This section covers iterative methods for finding the minimum of a function, which is central to training machine learning models. We focus on Gradient Descent and its variants.

Problem 3.1. Consider the loss function $L(X) = \frac{1}{2}X^TAX + b^TX$, where $X = \begin{pmatrix} x \\ y \end{pmatrix}$, $A = \begin{pmatrix} 2 & 3 \\ 4 & 1 \end{pmatrix}$, $b = \begin{pmatrix} -3 \\ -2 \end{pmatrix}$. Starting from $X_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ with step size $\gamma = 0.05$, compute the first two iterations (X_1, X_2) for:

(a) Plain Gradient Descent (GD).

(b) Gradient Descent with momentum ($\beta = 0.5$ and initial momentum $m_0 = 0$).

(Note: Per convention, assume the gradient is $\nabla L(X) = AX + b$).

Solution. Gradient Formula: Per the problem's convention, we use $\nabla L(X) = AX + b$.
(a) Plain Gradient Descent: $X_{k+1} = X_k - \gamma \nabla L(X_k)$

Iteration 1: Calculate X_1.

- **Gradient at X_0:** $\nabla L(X_0) = AX_0 + b = \begin{pmatrix} 2 & 3 \\ 4 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} -3 \\ -2 \end{pmatrix} = \begin{pmatrix} 5 \\ 5 \end{pmatrix} + \begin{pmatrix} -3 \\ -2 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$.
- **Update step:** $X_1 = X_0 - \gamma \nabla L(X_0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - 0.05 \begin{pmatrix} 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 - 0.1 \\ 1 - 0.15 \end{pmatrix} = \begin{pmatrix} 0.90 \\ 0.85 \end{pmatrix}$.

Iteration 2: Calculate X_2.

- **Gradient at X_1:** $\nabla L(X_1) = AX_1 + b = \begin{pmatrix} 2 & 3 \\ 4 & 1 \end{pmatrix}\begin{pmatrix} 0.90 \\ 0.85 \end{pmatrix} + \begin{pmatrix} -3 \\ -2 \end{pmatrix} = \begin{pmatrix} 1.8 + 2.55 \\ 3.6 + 0.85 \end{pmatrix} + \begin{pmatrix} -3 \\ -2 \end{pmatrix} = \begin{pmatrix} 4.35 \\ 4.45 \end{pmatrix} + \begin{pmatrix} -3 \\ -2 \end{pmatrix} = \begin{pmatrix} 1.35 \\ 2.45 \end{pmatrix}$.
- **Update step:** $X_2 = X_1 - \gamma \nabla L(X_1) = \begin{pmatrix} 0.90 \\ 0.85 \end{pmatrix} - 0.05 \begin{pmatrix} 1.35 \\ 2.45 \end{pmatrix} = \begin{pmatrix} 0.90 - 0.0675 \\ 0.85 - 0.1225 \end{pmatrix} = \begin{pmatrix} 0.8325 \\ 0.7275 \end{pmatrix}$.

Final Answer: $X_1 = (0.90, 0.85)^T$ and $X_2 = (0.8325, 0.7275)^T$.
(b) Gradient Descent with Momentum Update Rules: $m_{k+1} = \beta m_k + \nabla L(X_k)$ and $X_{k+1} = X_k - \gamma m_{k+1}$.

Iteration 1: Calculate X_1.

- **Gradient at X_0:** $\nabla L(X_0) = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$ (from part a).
- **Momentum update:** $m_1 = \beta m_0 + \nabla L(X_0) = 0.5 \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$.
- **Position update:** $X_1 = X_0 - \gamma m_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - 0.05 \begin{pmatrix} 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 0.90 \\ 0.85 \end{pmatrix}$.

Iteration 2: Calculate X_2.

- **Gradient at X_1:** $\nabla L(X_1) = \begin{pmatrix} 1.35 \\ 2.45 \end{pmatrix}$ (from part a).
- **Momentum update:** $m_2 = \beta m_1 + \nabla L(X_1) = 0.5 \begin{pmatrix} 2 \\ 3 \end{pmatrix} + \begin{pmatrix} 1.35 \\ 2.45 \end{pmatrix} = \begin{pmatrix} 1 \\ 1.5 \end{pmatrix} + \begin{pmatrix} 1.35 \\ 2.45 \end{pmatrix} = \begin{pmatrix} 2.35 \\ 3.95 \end{pmatrix}$.
- **Position update:** $X_2 = X_1 - \gamma m_2 = \begin{pmatrix} 0.90 \\ 0.85 \end{pmatrix} - 0.05 \begin{pmatrix} 2.35 \\ 3.95 \end{pmatrix} = \begin{pmatrix} 0.90 - 0.1175 \\ 0.85 - 0.1975 \end{pmatrix} = \begin{pmatrix} 0.7825 \\ 0.6525 \end{pmatrix}$.

Final Answer: $X_1 = (0.90, 0.85)^T$ and $X_2 = (0.7825, 0.6525)^T$.

Problem 3.2. Consider the function $J(w_1, w_2) = w_1^2 + 5w_2^2$. Starting from $w^{(0)} = (5, 1)^T$, perform one iteration of the AdaGrad algorithm to find $w^{(1)}$. Use a global learning rate $\alpha = 1.0$ and a small constant $\epsilon = 10^{-8}$ for numerical stability.

Solution.

Step 1: Define AdaGrad Update Rules. The AdaGrad update rules for each parameter w_i are:

- Accumulate squared gradients: $A_{i,t} \leftarrow A_{i,t-1} + (\frac{\partial J}{\partial w_i})^2$
- Update parameter: $w_{i,t} \leftarrow w_{i,t-1} - \frac{\alpha}{\sqrt{A_{i,t}+\epsilon}} \frac{\partial J}{\partial w_i}$

We initialize the accumulator $A_0 = (0, 0)^T$.

Step 2: Compute the Gradient at the Starting Point. The gradient of $J(w_1, w_2)$ is $\nabla J = (2w_1, 10w_2)^T$. At $w^{(0)} = (5, 1)^T$, the gradient is $\nabla J(w^{(0)}) = \begin{pmatrix} 2(5) \\ 10(1) \end{pmatrix} = \begin{pmatrix} 10 \\ 10 \end{pmatrix}$.

Step 3: Update the Gradient Accumulator A.

$$A_{1,1} = A_{1,0} + (10)^2 = 0 + 100 = 100$$
$$A_{2,1} = A_{2,0} + (10)^2 = 0 + 100 = 100$$

Step 4: Update the Parameters to find $w^{(1)}$.

$$w_1^{(1)} = w_1^{(0)} - \frac{\alpha}{\sqrt{A_{1,1} + \epsilon}} \frac{\partial J}{\partial w_1} \approx 5 - \frac{1.0}{\sqrt{100}}(10) = 5 - 1 = 4.0$$
$$w_2^{(1)} = w_2^{(0)} - \frac{\alpha}{\sqrt{A_{2,1} + \epsilon}} \frac{\partial J}{\partial w_2} \approx 1 - \frac{1.0}{\sqrt{100}}(10) = 1 - 1 = 0.0$$

Step 5: Conclusion. After one iteration of AdaGrad, the new parameter vector is $w^{(1)} = (4.0, 0.0)^T$.

Problem 3.3. Consider the function $J(w_1, w_2) = (w_1 - 10)^2 + w_2^2$. Starting from $w^{(0)} = (0, 2)^T$, perform one iteration of the Adam algorithm to find $w^{(1)}$. Use $\alpha = 0.5$, $\beta_1 = 0.9$ (for momentum), $\beta_2 = 0.999$ (for squared gradient), and $\epsilon = 10^{-8}$.

Solution.

Step 1: Define Adam Update Rules. For iteration $t = 1$, we initialize the first moment $m_0 = \mathbf{0}$ and second moment $v_0 = \mathbf{0}$. The steps are:

- $g_1 = \nabla J(w_0)$
- $m_1 = \beta_1 m_0 + (1 - \beta_1) g_1 = (1 - \beta_1) g_1$
- $v_1 = \beta_2 v_0 + (1 - \beta_2) g_1^2 = (1 - \beta_2) g_1^2$ (element-wise)
- $\hat{m}_1 = \frac{m_1}{1-\beta_1^1}$, $\hat{v}_1 = \frac{v_1}{1-\beta_2^1}$
- $w_1 = w_0 - \alpha \frac{\hat{m}_1}{\sqrt{\hat{v}_1} + \epsilon}$

Step 2: Compute Gradient at $w^{(0)}$. The gradient of $J(w_1, w_2)$ is $\nabla J = (2(w_1 - 10), 2w_2)^T$. At $w^{(0)} = (0, 2)^T$, the gradient is $g_1 = \begin{pmatrix} 2(0-10) \\ 2(2) \end{pmatrix} = \begin{pmatrix} -20 \\ 4 \end{pmatrix}$.

Step 3: Update First and Second Moments (m_1, v_1).

$$m_1 = (1 - 0.9) \begin{pmatrix} -20 \\ 4 \end{pmatrix} = 0.1 \begin{pmatrix} -20 \\ 4 \end{pmatrix} = \begin{pmatrix} -2 \\ 0.4 \end{pmatrix}$$
$$v_1 = (1 - 0.999) \begin{pmatrix} (-20)^2 \\ 4^2 \end{pmatrix} = 0.001 \begin{pmatrix} 400 \\ 16 \end{pmatrix} = \begin{pmatrix} 0.4 \\ 0.016 \end{pmatrix}$$

Step 4: Correct for Bias.

$$\hat{m}_1 = \frac{m_1}{1 - 0.9} = \frac{1}{0.1} \begin{pmatrix} -2 \\ 0.4 \end{pmatrix} = \begin{pmatrix} -20 \\ 4 \end{pmatrix}$$
$$\hat{v}_1 = \frac{v_1}{1 - 0.999} = \frac{1}{0.001} \begin{pmatrix} 0.4 \\ 0.016 \end{pmatrix} = \begin{pmatrix} 400 \\ 16 \end{pmatrix}$$

Step 5: Update Parameters to find $w^{(1)}$.

$$w_1^{(1)} = 0 - \alpha \frac{\hat{m}_{1,1}}{\sqrt{\hat{v}_{1,1}} + \epsilon} \approx 0 - 0.5 \frac{-20}{\sqrt{400}} = 0 - 0.5 \frac{-20}{20} = 0.5$$
$$w_2^{(1)} = 2 - \alpha \frac{\hat{m}_{1,2}}{\sqrt{\hat{v}_{1,2}} + \epsilon} \approx 2 - 0.5 \frac{4}{\sqrt{16}} = 2 - 0.5 \frac{4}{4} = 1.5$$

Step 6: Conclusion. After one iteration of Adam, the new parameter vector is $w^{(1)} = (0.5, 1.5)^T$.

Problem 3.4. Consider the loss function for a dataset of two points: $L(w) = (w \cdot x_1 - y_1)^2 + (w \cdot x_2 - y_2)^2$, where $x_1 = 1, y_1 = 2$ and $x_2 = 3, y_2 = 4$. The parameter is a scalar w. Starting from $w_0 = 1$ with learning rate $\alpha = 0.1$:

(a) Compute one update step using Batch Gradient Descent.

(b) Compute one update step using Stochastic Gradient Descent (SGD), assuming the second data point (x_2, y_2) is chosen.

Solution. (a) Batch Gradient Descent

Step 1: Find the full gradient. The loss is $L(w) = (w - 2)^2 + (3w - 4)^2$. The gradient is $\frac{dL}{dw} = 2(w - 2) + 2(3w - 4)(3) = 2w - 4 + 18w - 24 = 20w - 28$.

Step 2: Evaluate the gradient at $w_0 = 1$.

$$\nabla L(w_0) = 20(1) - 28 = -8$$

Step 3: Perform the update.

$$w_1 = w_0 - \alpha \nabla L(w_0) = 1 - 0.1(-8) = 1 + 0.8 = 1.8$$

(b) Stochastic Gradient Descent (SGD)

Step 1: Find the loss for the chosen point. For point 2, the loss is $L_2(w) = (3w - 4)^2$.

Step 2: Find the gradient for that point.

$$\frac{dL_2}{dw} = 2(3w - 4)(3) = 18w - 24$$

Step 3: Evaluate the gradient at $w_0 = 1$.

$$\nabla L_2(w_0) = 18(1) - 24 = -6$$

Step 4: Perform the update.

$$w_1 = w_0 - \alpha \nabla L_2(w_0) = 1 - 0.1(-6) = 1 + 0.6 = 1.6$$

Step 5: Conclusion. The batch update gives $w_1 = 1.8$, while the SGD update using the second point gives $w_1 = 1.6$.

Problem 3.5. You are minimizing the function $f(x) = (x-3)^2$ using gradient descent, starting at $x_0 = 1$. The learning rate is $\alpha = 0.1$. Calculate the first three iterates, x_1, x_2, x_3.

Solution.

Step 1: Define the Update Rule and Gradient. The update rule is $x_{k+1} = x_k - \alpha f'(x_k)$. The gradient (derivative) is $f'(x) = 2(x - 3)$.

Step 2: Calculate x_1.

- $f'(x_0) = f'(1) = 2(1 - 3) = -4$.
- $x_1 = x_0 - \alpha f'(x_0) = 1 - 0.1(-4) = 1 + 0.4 = 1.4$.

Step 3: Calculate x_2.

- $f'(x_1) = f'(1.4) = 2(1.4 - 3) = 2(-1.6) = -3.2$.
- $x_2 = x_1 - \alpha f'(x_1) = 1.4 - 0.1(-3.2) = 1.4 + 0.32 = 1.72$.

Step 4: Calculate x_3.

- $f'(x_2) = f'(1.72) = 2(1.72 - 3) = 2(-1.28) = -2.56$.
- $x_3 = x_2 - \alpha f'(x_2) = 1.72 - 0.1(-2.56) = 1.72 + 0.256 = 1.976$.

Step 5: Conclusion. The first three iterates are $x_1 = 1.4$, $x_2 = 1.72$, and $x_3 = 1.976$.

Problem 3.6. Consider the function $J(w_1, w_2) = w_1^2 + 25w_2^2$. Starting from $w^{(0)} = (5, 1)^T$, perform one iteration of the RMSProp algorithm. Use a learning rate $\alpha = 0.1$, a decay rate $\rho = 0.9$, and $\epsilon = 10^{-8}$.

Solution.

Step 1: Define RMSProp Update Rules. The RMSProp update rules are:

- Accumulate squared gradients: $A_{i,t} \leftarrow \rho A_{i,t-1} + (1-\rho)(\frac{\partial J}{\partial w_i})^2$
- Update parameter: $w_{i,t} \leftarrow w_{i,t-1} - \frac{\alpha}{\sqrt{A_{i,t}+\epsilon}} \frac{\partial J}{\partial w_i}$

We initialize the accumulator $A_0 = (0,0)^T$.

Step 2: Compute the Gradient at the Starting Point. The gradient of $J(w_1, w_2) = w_1^2 + 25w_2^2$ is $\nabla J = (2w_1, 50w_2)^T$. At $w^{(0)} = (5,1)^T$, the gradient is $\nabla J(w^{(0)}) = \begin{pmatrix} 2(5) \\ 50(1) \end{pmatrix} = \begin{pmatrix} 10 \\ 50 \end{pmatrix}$.

Step 3: Update the Gradient Accumulator A.

$$A_{1,1} = \rho A_{1,0} + (1-\rho)(10)^2 = 0.9(0) + 0.1(100) = 10$$
$$A_{2,1} = \rho A_{2,0} + (1-\rho)(50)^2 = 0.9(0) + 0.1(2500) = 250$$

Step 4: Update the Parameters to find $w^{(1)}$.

$$w_1^{(1)} = w_1^{(0)} - \frac{\alpha}{\sqrt{A_{1,1}+\epsilon}} \frac{\partial J}{\partial w_1} \approx 5 - \frac{0.1}{\sqrt{10}}(10) = 5 - \frac{1}{\sqrt{10}} \approx 5 - 0.316 = 4.684$$
$$w_2^{(1)} = w_2^{(0)} - \frac{\alpha}{\sqrt{A_{2,1}+\epsilon}} \frac{\partial J}{\partial w_2} \approx 1 - \frac{0.1}{\sqrt{250}}(50) = 1 - \frac{5}{\sqrt{250}} \approx 1 - 0.316 = 0.684$$

Step 5: Conclusion. After one iteration of RMSProp, the new parameter vector is approximately $w^{(1)} = (4.684, 0.684)^T$.

Problem 3.7. You want to minimize $f(x) = 2x^2 + 4x + 5$. At the current point $x_t = 1$, the descent direction is the negative gradient, $d_t = -f'(x_t)$. Find the optimal step size α_t using an exact line search.

Solution.

Step 1: Define the Line Search Problem. The goal of exact line search is to find the α_t that minimizes the function along the search direction:

$$\alpha_t = \arg\min_{\alpha > 0} f(x_t + \alpha d_t)$$

Step 2: Calculate the Descent Direction at $x_t = 1$. The gradient is $f'(x) = 4x + 4$. At $x_t = 1$, the gradient is $f'(1) = 4(1)+4 = 8$. The descent direction is $d_t = -f'(1) = -8$.

Step 3: Create the Function of α. We define a new function $H(\alpha) = f(x_t + \alpha d_t) = f(1 - 8\alpha)$.

$$\begin{aligned} H(\alpha) &= 2(1-8\alpha)^2 + 4(1-8\alpha) + 5 \\ &= 2(1 - 16\alpha + 64\alpha^2) + 4 - 32\alpha + 5 \\ &= 2 - 32\alpha + 128\alpha^2 + 9 - 32\alpha \\ &= 128\alpha^2 - 64\alpha + 11 \end{aligned}$$

Step 4: Minimize $H(\alpha)$. To find the minimum, we take the derivative with respect to α and set it to zero.

$$H'(\alpha) = 256\alpha - 64 = 0$$

$$256\alpha = 64 \implies \alpha = \frac{64}{256} = \frac{1}{4}$$

Step 5: Conclusion. The optimal step size for this iteration is $\alpha_t = 0.25$.

Problem 3.8. You are minimizing $f(x) = x^2$ starting at $x_0 = 4$ using gradient descent with exponential learning rate decay: $\alpha_t = \alpha_0 e^{-kt}$. Use $\alpha_0 = 0.5$ and $k = 0.1$. Compute the first two iterates, x_1 and x_2.

Solution.

Step 1: Define Update Rules. The update is $x_{t+1} = x_t - \alpha_t f'(x_t)$. The gradient is $f'(x) = 2x$. The learning rate for step t (to find x_{t+1}) is $\alpha_t = 0.5e^{-0.1t}$.

Step 2: Calculate x_1 (using $t = 0$).

- Learning rate $\alpha_0 = 0.5e^{-0.1(0)} = 0.5$.
- Gradient at $x_0 = 4$: $f'(4) = 2(4) = 8$.
- Update: $x_1 = x_0 - \alpha_0 f'(x_0) = 4 - 0.5(8) = 4 - 4 = 0$.

Step 3: Calculate x_2 (using $t = 1$).

- Learning rate $\alpha_1 = 0.5e^{-0.1(1)} \approx 0.5 \times 0.9048 = 0.4524$.
- Gradient at $x_1 = 0$: $f'(0) = 2(0) = 0$.
- Update: $x_2 = x_1 - \alpha_1 f'(x_1) = 0 - 0.4524(0) = 0$.

Step 4: Conclusion. The first iterate is $x_1 = 0$. Since the first step lands exactly on the minimum, the gradient becomes zero and all subsequent iterates, including x_2, will also be 0.

Problem 3.9. For the function $f(x, y) = 3x^2 - 4xy + 2y^2$, compute the gradient vector ∇f and the Hessian matrix H. Evaluate both at the point $(1, 2)$.

Solution.

Step 1: Compute the Gradient Vector. The gradient is the vector of first-order partial derivatives.

$$\frac{\partial f}{\partial x} = 6x - 4y$$
$$\frac{\partial f}{\partial y} = -4x + 4y$$

So, the gradient vector is $\nabla f(x, y) = \begin{pmatrix} 6x - 4y \\ -4x + 4y \end{pmatrix}$.

Step 2: Compute the Hessian Matrix. The Hessian is the matrix of second-order partial derivatives.

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}(6x - 4y) = 6$$
$$\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}(6x - 4y) = -4$$
$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial x}(-4x + 4y) = -4$$
$$\frac{\partial^2 f}{\partial y^2} = \frac{\partial}{\partial y}(-4x + 4y) = 4$$

So, the Hessian matrix is $H = \begin{pmatrix} 6 & -4 \\ -4 & 4 \end{pmatrix}$.

Step 3: Evaluate at the point $(1, 2)$.

- Gradient: $\nabla f(1, 2) = \begin{pmatrix} 6(1) - 4(2) \\ -4(1) + 4(2) \end{pmatrix} = \begin{pmatrix} 6 - 8 \\ -4 + 8 \end{pmatrix} = \begin{pmatrix} -2 \\ 4 \end{pmatrix}$.
- Hessian: The Hessian is constant for a quadratic function, so $H(1, 2) = \begin{pmatrix} 6 & -4 \\ -4 & 4 \end{pmatrix}$.

Step 4: Conclusion. At $(1, 2)$, the gradient is $(-2, 4)^T$ and the Hessian is $\begin{pmatrix} 6 & -4 \\ -4 & 4 \end{pmatrix}$.

4 Optimization (Constrained)

This section addresses optimization problems where the solution must satisfy certain equality or inequality constraints. We will use the method of Lagrange Multipliers and the Karush-Kuhn-Tucker (KKT) conditions.

Problem 4.1. A signal tower is at (0,0) and the cost to place a receiver at (x, y) is $f(x, y) = 2(x^2 + y^2)$. The receiver must be placed in the region $R = \{(x, y) : 1 \leq x + y \leq 2\}$.

(a) Formulate the primal optimization problem.

(b) Formulate the Lagrangian dual problem.

Solution. (a) Primal Problem Formulation

Step 1: Write the objective function. The objective is to minimize the cost:

$$\text{Minimize } f(x, y) = 2(x^2 + y^2)$$

Step 2: Write constraints in standard form $g_i(x) \leq 0$. The region $1 \leq x + y \leq 2$ gives two constraints:

- $1 \leq x + y \implies 1 - x - y \leq 0$. Let $g_1(x, y) = 1 - x - y$.
- $x + y \leq 2 \implies x + y - 2 \leq 0$. Let $g_2(x, y) = x + y - 2$.

Step 3: State the full primal problem.

$$\begin{aligned} \textbf{Minimize} \quad & 2(x^2 + y^2) \\ \textbf{subject to} \quad & 1 - x - y \leq 0 \\ & x + y - 2 \leq 0 \end{aligned}$$

(b) Lagrangian Dual Problem Formulation

Step 1: Write the Lagrangian function. Introduce non-negative Lagrange multipliers $\lambda_1, \lambda_2 \geq 0$.

$$\begin{aligned} \mathcal{L}(x, y, \lambda_1, \lambda_2) &= f(x, y) + \lambda_1 g_1(x, y) + \lambda_2 g_2(x, y) \\ &= 2(x^2 + y^2) + \lambda_1(1 - x - y) + \lambda_2(x + y - 2) \end{aligned}$$

Step 2: Define the dual function $D(\lambda_1, \lambda_2)$. The dual function is the infimum (minimum) of the Lagrangian with respect to the primal variables. To find it, we set the partial derivatives with respect to x and y to zero.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial x} = 4x - \lambda_1 + \lambda_2 = 0 &\implies x = \frac{\lambda_1 - \lambda_2}{4} \\ \frac{\partial \mathcal{L}}{\partial y} = 4y - \lambda_1 + \lambda_2 = 0 &\implies y = \frac{\lambda_1 - \lambda_2}{4} \end{aligned}$$

Substitute these back into $\mathcal{L}$ to get $D(\lambda_1, \lambda_2)$.

$$\begin{aligned} D(\lambda_1, \lambda_2) &= 2\left(\frac{\lambda_1 - \lambda_2}{4}\right)^2 + 2\left(\frac{\lambda_1 - \lambda_2}{4}\right)^2 + (\lambda_2 - \lambda_1)\left(\frac{\lambda_1 - \lambda_2}{4} + \frac{\lambda_1 - \lambda_2}{4}\right) + \lambda_1 - 2\lambda_2 \\ &= 4\frac{(\lambda_1 - \lambda_2)^2}{16} - (\lambda_1 - \lambda_2)\frac{2(\lambda_1 - \lambda_2)}{4} + \lambda_1 - 2\lambda_2 \\ &= \frac{(\lambda_1 - \lambda_2)^2}{4} - \frac{(\lambda_1 - \lambda_2)^2}{2} + \lambda_1 - 2\lambda_2 \\ &= -\frac{(\lambda_1 - \lambda_2)^2}{4} + \lambda_1 - 2\lambda_2 \end{aligned}$$

Step 3: Formulate the dual problem. The dual problem is to maximize the dual function subject to the constraints on the multipliers.

$$\begin{aligned} \textbf{Maximize} \quad & D(\lambda_1, \lambda_2) = -\frac{1}{4}(\lambda_1 - \lambda_2)^2 + \lambda_1 - 2\lambda_2 \\ \textbf{subject to} \quad & \lambda_1 \geq 0, \quad \lambda_2 \geq 0 \end{aligned}$$

Problem 4.2. Consider the optimization problem: Minimize $f(x, y) = (x - 4)^2 + (y - 4)^2$ subject to $x + y \leq 4$ and $x \geq 1$.

(a) Is this a convex optimization problem? Does Slater's condition hold?

(b) Write down the KKT conditions for this problem.

(c) Solve the KKT conditions to find the optimal point (x^*, y^*).

Solution. (a) Convexity and Slater's Condition

Step 1: Check convexity. The objective function $f(x, y)$ is a paraboloid, which is convex. The constraints, written in standard form as $g_1(x, y) = x + y - 4 \leq 0$ and $g_2(x, y) = 1 - x \leq 0$, are both linear and therefore convex. The problem is a convex optimization problem.

Step 2: Check Slater's condition. We need a point (x, y) where all inequality constraints are strictly satisfied: $x + y < 4$ and $x > 1$. The point $(x, y) = (2, 1)$ works, since $2 + 1 = 3 < 4$ and $2 > 1$. Therefore, Slater's condition holds and strong duality is guaranteed.

(b) KKT Conditions

Step 1: Form the Lagrangian. $\mathcal{L}(x, y, \lambda_1, \lambda_2) = (x-4)^2+(y-4)^2+\lambda_1(x+y-4)+\lambda_2(1-x)$.

Step 2: List the KKT conditions.

(a) **Stationarity:** $\nabla_x \mathcal{L} = 2(x - 4) + \lambda_1 - \lambda_2 = 0$ and $\nabla_y \mathcal{L} = 2(y - 4) + \lambda_1 = 0$.

(b) **Primal Feasibility:** $x + y - 4 \leq 0$ and $1 - x \leq 0$.

(c) **Dual Feasibility:** $\lambda_1 \geq 0, \lambda_2 \geq 0$.

(d) **Complementary Slackness:** $\lambda_1(x + y - 4) = 0$ and $\lambda_2(1 - x) = 0$.

(c) Solving the KKT conditions

Case 1: Both constraints inactive ($\lambda_1 = 0, \lambda_2 = 0$). Stationarity gives $2(x - 4) = 0 \implies x = 4$ and $2(y - 4) = 0 \implies y = 4$. Check feasibility: $x + y - 4 = 4 + 4 - 4 = 4 \not\leq 0$. Impossible.

Case 2: g_1 active, g_2 inactive ($\lambda_1 > 0, \lambda_2 = 0$). $x + y - 4 = 0 \implies y = 4 - x$. Also $x > 1$. From stationarity: $2(y - 4) + \lambda_1 = 0 \implies 2((4 - x) - 4) + \lambda_1 = 0 \implies -2x + \lambda_1 = 0 \implies \lambda_1 = 2x$. $2(x-4)+\lambda_1 = 0 \implies 2(x-4)+2x = 0 \implies 4x-8 = 0 \implies x = 2$. This gives $y = 4 - 2 = 2$ and $\lambda_1 = 4$. This point $(x, y) = (2, 2)$ with $\lambda_1 = 4, \lambda_2 = 0$ satisfies all KKT conditions. This is our solution.

Case 3: g_1 inactive, g_2 active ($\lambda_1 = 0, \lambda_2 > 0$). $1 - x = 0 \implies x = 1$. Also $x + y < 4 \implies y < 3$. From stationarity: $2(y - 4) + \lambda_1 = 0 \implies 2(y - 4) = 0 \implies y = 4$. This contradicts $y < 3$. Impossible.

Case 4: Both constraints active ($\lambda_1 > 0, \lambda_2 > 0$). $x = 1$ and $y = 3$. From stationarity: $2(y-4) + \lambda_1 = 0 \implies 2(3-4) + \lambda_1 = 0 \implies \lambda_1 = 2$. $2(x-4) + \lambda_1 - \lambda_2 = 0 \implies 2(1-4) + 2 - \lambda_2 = 0 \implies -4 - \lambda_2 = 0 \implies \lambda_2 = -4$. This contradicts dual feasibility ($\lambda_2 \geq 0$). Impossible.

Conclusion. The unique solution is $(x^*, y^*) = (2, 2)$.

Problem 4.3. Formulate the Lagrangian dual of the following optimization problem: Minimize $f(x, y) = x^2 + 2y^2$ subject to $x + y \leq 3$ and $x - 2y = 0$.

Solution.

Step 1: Identify Objective and Constraints in Standard Form.

- Objective: $f(x, y) = x^2 + 2y^2$.
- Inequality Constraint: $g(x, y) = x + y - 3 \leq 0$.
- Equality Constraint: $h(x, y) = x - 2y = 0$.

Step 2: Formulate the Lagrangian. Introduce a non-negative multiplier $\lambda \geq 0$ for the inequality and an unrestricted multiplier ν for the equality.

$$\begin{aligned}\mathcal{L}(x, y, \lambda, \nu) &= f(x, y) + \lambda g(x, y) + \nu h(x, y) \\ &= x^2 + 2y^2 + \lambda(x + y - 3) + \nu(x - 2y)\end{aligned}$$

Step 3: Find the Dual Function $D(\lambda, \nu)$. The dual function is the infimum of $\mathcal{L}$ over the primal variables x, y. We find this by setting the partial derivatives to zero.

$$\frac{\partial \mathcal{L}}{\partial x} = 2x + \lambda + \nu = 0 \implies x = -\frac{\lambda + \nu}{2}$$
$$\frac{\partial \mathcal{L}}{\partial y} = 4y + \lambda - 2\nu = 0 \implies y = -\frac{\lambda - 2\nu}{4}$$

Step 4: Substitute x and y back into $\mathcal{L}$. This will give $D(\lambda, \nu)$.

$$\begin{aligned}D(\lambda, \nu) &= \left(-\frac{\lambda + \nu}{2}\right)^2 + 2\left(-\frac{\lambda - 2\nu}{4}\right)^2 + \lambda\left(-\frac{\lambda + \nu}{2} - \frac{\lambda - 2\nu}{4} - 3\right) + \nu\left(-\frac{\lambda + \nu}{2} - 2\left(-\frac{\lambda -\right.\right. \\ &= \frac{(\lambda + \nu)^2}{4} + \frac{2(\lambda - 2\nu)^2}{16} - \frac{\lambda}{4}(2\lambda + 2\nu + \lambda - 2\nu) - 3\lambda + \frac{\nu}{2}(-\lambda - \nu + \lambda - 2\nu) \\ &= \frac{(\lambda + \nu)^2}{4} + \frac{(\lambda - 2\nu)^2}{8} - \frac{3\lambda^2}{4} - 3\lambda - \frac{3\nu^2}{2} \\ &= \frac{2(\lambda^2 + 2\lambda\nu + \nu^2) + (\lambda^2 - 4\lambda\nu + 4\nu^2)}{8} - \frac{6\lambda^2}{8} - 3\lambda - \frac{12\nu^2}{8} \\ &= \frac{2\lambda^2 + 4\lambda\nu + 2\nu^2 + \lambda^2 - 4\lambda\nu + 4\nu^2 - 6\lambda^2 - 12\nu^2}{8} - 3\lambda \\ &= \frac{-3\lambda^2 - 6\nu^2}{8} - 3\lambda\end{aligned}$$

Step 5: Formulate the Dual Problem.

$$\begin{aligned}\textbf{Maximize} \quad & D(\lambda, \nu) = -\frac{3}{8}\lambda^2 - \frac{3}{4}\nu^2 - 3\lambda \\ \textbf{subject to} \quad & \lambda \geq 0\end{aligned}$$

Problem 4.4. Use the method of Lagrange multipliers to find the maximum value of $f(x, y) = xy$ subject to the constraint $x + y = 10$.

Solution.

Step 1: Set up the Lagrangian. The constraint is $g(x, y) = x + y - 10 = 0$.

$$\mathcal{L}(x, y, \lambda) = f(x, y) - \lambda g(x, y) = xy - \lambda(x + y - 10)$$

Step 2: Find the Gradient and Set to Zero.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial x} &= y - \lambda = 0 \implies y = \lambda \\
\frac{\partial \mathcal{L}}{\partial y} &= x - \lambda = 0 \implies x = \lambda \\
\frac{\partial \mathcal{L}}{\partial \lambda} &= -(x + y - 10) = 0 \implies x + y = 10
\end{aligned}$$

Step 3: Solve the System. From the first two equations, we get $x = y$. Substituting this into the third equation:

$$x + x = 10 \implies 2x = 10 \implies x = 5$$

Therefore, $y = 5$ as well.

Step 4: Conclusion. The point that maximizes the function is $(5, 5)$. The maximum value is $f(5, 5) = 5 \times 5 = 25$.

Problem 4.5. Find the point on the plane $2x+y-z = 5$ that is closest to the origin $(0, 0, 0)$ by formulating this as a constrained optimization problem and solving it with Lagrange multipliers.

Solution.

Step 1: Formulate the Optimization Problem. Minimizing the distance from the origin to a point (x, y, z) is equivalent to minimizing the squared distance. This avoids square roots and simplifies differentiation.

- **Objective Function:** Minimize $f(x, y, z) = x^2 + y^2 + z^2$.
- **Constraint:** The point must be on the plane, so $g(x, y, z) = 2x + y - z - 5 = 0$.

Step 2: Set up the Lagrangian.

$$\mathcal{L}(x, y, z, \lambda) = (x^2 + y^2 + z^2) - \lambda(2x + y - z - 5)$$

Step 3: Find the Gradients and Set to Zero.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial x} &= 2x - 2\lambda = 0 \implies x = \lambda \\
\frac{\partial \mathcal{L}}{\partial y} &= 2y - \lambda = 0 \implies y = \lambda/2 \\
\frac{\partial \mathcal{L}}{\partial z} &= 2z + \lambda = 0 \implies z = -\lambda/2 \\
\frac{\partial \mathcal{L}}{\partial \lambda} &= -(2x + y - z - 5) = 0 \implies 2x + y - z = 5
\end{aligned}$$

Step 4: Solve the System of Equations. Substitute the expressions for x, y, z in terms of λ into the constraint equation:

$$2(\lambda) + (\lambda/2) - (-\lambda/2) = 5$$

$$2\lambda + \lambda/2 + \lambda/2 = 5$$

$$3\lambda = 5 \implies \lambda = 5/3$$

Step 5: Find the Optimal Point (x^*, y^*, z^*).

$$x^* = \lambda = 5/3$$
$$y^* = \lambda/2 = 5/6$$
$$z^* = -\lambda/2 = -5/6$$

Step 6: Conclusion. The point on the plane closest to the origin is $(5/3, 5/6, -5/6)$.

Problem 4.6. Find the extreme values of the function $f(x, y, z) = x - 2y + 2z$ on the sphere $x^2 + y^2 + z^2 = 1$.

Solution.

Step 1: Identify Objective and Constraint.

- Objective: $f(x, y, z) = x - 2y + 2z$.
- Constraint: $g(x, y, z) = x^2 + y^2 + z^2 - 1 = 0$.

Step 2: Set up the Lagrange Multiplier Equation $\nabla f = \lambda \nabla g$.

- $\nabla f = (1, -2, 2)^T$.
- $\nabla g = (2x, 2y, 2z)^T$.

The system of equations is:

$$1 = \lambda(2x) \quad (1)$$
$$-2 = \lambda(2y) \quad (2)$$
$$2 = \lambda(2z) \quad (3)$$
$$x^2 + y^2 + z^2 = 1 \quad (4)$$

Step 3: Solve the System. From equations (1), (2), and (3), we can express x, y, z in terms of λ (note that $\lambda \neq 0$, otherwise the equations would be inconsistent):

$$x = \frac{1}{2\lambda}, \quad y = \frac{-2}{2\lambda} = \frac{-1}{\lambda}, \quad z = \frac{2}{2\lambda} = \frac{1}{\lambda}$$

Now substitute these into the constraint equation (4):

$$\left(\frac{1}{2\lambda}\right)^2 + \left(\frac{-1}{\lambda}\right)^2 + \left(\frac{1}{\lambda}\right)^2 = 1$$

$$\frac{1}{4\lambda^2} + \frac{1}{\lambda^2} + \frac{1}{\lambda^2} = 1$$

$$\frac{1+4+4}{4\lambda^2} = 1 \implies \frac{9}{4\lambda^2} = 1 \implies 4\lambda^2 = 9 \implies \lambda = \pm\frac{3}{2}$$

Step 4: Find the Candidate Points for Each Value of λ.

- **Case 1:** $\lambda = 3/2$**.**

$$x = \frac{1}{2(3/2)} = \frac{1}{3}, \quad y = \frac{-1}{3/2} = -\frac{2}{3}, \quad z = \frac{1}{3/2} = \frac{2}{3}$$

  Candidate point $P_1 = (1/3, -2/3, 2/3)$.

- **Case 2:** $\lambda = -3/2$**.**

$$x = \frac{1}{2(-3/2)} = -\frac{1}{3}, \quad y = \frac{-1}{-3/2} = \frac{2}{3}, \quad z = \frac{1}{-3/2} = -\frac{2}{3}$$

  Candidate point $P_2 = (-1/3, 2/3, -2/3)$.

Step 5: Evaluate the Objective Function at the Candidate Points.

- At P_1: $f(1/3, -2/3, 2/3) = \frac{1}{3} - 2(-\frac{2}{3}) + 2(\frac{2}{3}) = \frac{1+4+4}{3} = \frac{9}{3} = 3$.
- At P_2: $f(-1/3, 2/3, -2/3) = -\frac{1}{3} - 2(\frac{2}{3}) + 2(-\frac{2}{3}) = \frac{-1-4-4}{3} = \frac{-9}{3} = -3$.

Step 6: Conclusion. The **maximum** value is 3, and the **minimum** value is -3.

Problem 4.7. Write the KKT conditions for the following optimization problem: Minimize $f(x) = \frac{1}{2}x^2 - 4x$ subject to the constraint $x \geq 2$.

Solution.

Step 1: Standardize the Constraint. The constraint $x \geq 2$ must be in the form $g(x) \leq 0$.

$$2 - x \leq 0$$

So, $g(x) = 2 - x$.

Step 2: Form the Lagrangian. Introduce a non-negative Lagrange multiplier $\lambda \geq 0$.

$$\mathcal{L}(x, \lambda) = (\frac{1}{2}x^2 - 4x) + \lambda(2 - x)$$

Step 3: Write the KKT conditions.

(a) **Stationarity:** $\frac{d\mathcal{L}}{dx} = x - 4 - \lambda = 0$.

(b) **Primal Feasibility:** $2 - x \leq 0$ (or $x \geq 2$).

(c) **Dual Feasibility:** $\lambda \geq 0$.

(d) **Complementary Slackness:** $\lambda(2 - x) = 0$.

Step 4: (Optional) Solve the system. From complementary slackness, either $\lambda = 0$ or $x = 2$.

- If $\lambda = 0$: Stationarity gives $x - 4 = 0 \implies x = 4$. This satisfies primal feasibility ($4 \geq 2$). This is a valid solution.
- If $x = 2$: Stationarity gives $2 - 4 - \lambda = 0 \implies \lambda = -2$. This violates dual feasibility ($\lambda \geq 0$).

The optimal solution is $x = 4$.

Problem 4.8. Find the minimum of $f(x, y) = x^2 + y^2$ subject to $x - y = 3$ and $x \leq 2$.

Solution.

Step 1: Standardize Constraints and Form Lagrangian.

- Equality: $h(x, y) = x - y - 3 = 0$.
- Inequality: $g(x, y) = x - 2 \leq 0$.
- Lagrangian: $\mathcal{L}(x, y, \lambda, \nu) = x^2 + y^2 + \lambda(x - 2) + \nu(x - y - 3)$, with $\lambda \geq 0$.

Step 2: Write the KKT Conditions.

(a) **Stationarity:**

$$\frac{\partial \mathcal{L}}{\partial x} = 2x + \lambda + \nu = 0$$
$$\frac{\partial \mathcal{L}}{\partial y} = 2y - \nu = 0 \implies \nu = 2y$$

(b) **Primal Feasibility:** $x - 2 \leq 0$ and $x - y - 3 = 0$.

(c) **Dual Feasibility:** $\lambda \geq 0$.

(d) **Complementary Slackness:** $\lambda(x - 2) = 0$.

Step 3: Solve by Cases based on Complementary Slackness.

- **Case 1: Constraint is inactive ($\lambda = 0$).** The stationarity equations become: $2x + \nu = 0$ and $\nu = 2y$. This implies $2x + 2y = 0 \implies x = -y$. Substitute this into the equality constraint: $x - (-x) - 3 = 0 \implies 2x = 3 \implies x = 3/2$. Then $y = -3/2$. Now, check the primal feasibility for the inequality: $x - 2 \leq 0 \implies 3/2 - 2 \leq 0 \implies -1/2 \leq 0$. This is true. So, $(x, y) = (3/2, -3/2)$ with $\lambda = 0$ and $\nu = 2y = -3$ is a valid KKT point.
- **Case 2: Constraint is active ($\lambda > 0$).** This means $x - 2 = 0 \implies x = 2$. From the equality constraint: $2 - y - 3 = 0 \implies y = -1$. Now find the multipliers. From stationarity: $\nu = 2y = 2(-1) = -2$. And $2x + \lambda + \nu = 0 \implies 2(2) + \lambda + (-2) = 0 \implies 4 + \lambda - 2 = 0 \implies \lambda = -2$. This contradicts dual feasibility ($\lambda \geq 0$), so this case is impossible.

Step 4: Conclusion. The only valid solution from the KKT conditions is from Case 1. The optimal point is $(x^*, y^*) = (3/2, -3/2)$.

Problem 4.9. Solve using KKT conditions: Minimize $f(x, y) = x^2 + y^2$ subject to $x + y \geq 2$.

Solution.

Step 1: Standardize and Form Lagrangian. The constraint is $g(x, y) = 2 - x - y \leq 0$.

$$\mathcal{L}(x, y, \lambda) = x^2 + y^2 + \lambda(2 - x - y)$$

Step 2: Write KKT Conditions.

(a) **Stationarity:** $2x - \lambda = 0 \implies x = \lambda/2$; $2y - \lambda = 0 \implies y = \lambda/2$.

(b) **Primal Feasibility:** $2 - x - y \leq 0$.

(c) **Dual Feasibility:** $\lambda \geq 0$.

(d) **Complementary Slackness:** $\lambda(2 - x - y) = 0$.

Step 3: Solve the System.

- **Case 1:** $\lambda = 0$**.** Stationarity implies $x = 0, y = 0$. Check primal feasibility: $2 - 0 - 0 = 2 \not\leq 0$. This case is impossible.
- **Case 2:** $\lambda > 0$**.** Complementary slackness implies the constraint must be active: $2 - x - y = 0$. Substitute the stationarity results ($x = y = \lambda/2$) into the active constraint:

$$2 - (\lambda/2) - (\lambda/2) = 0 \implies 2 - \lambda = 0 \implies \lambda = 2$$

This gives $x = 2/2 = 1$ and $y = 2/2 = 1$.

Step 4: Verify Solution. The point $(x, y) = (1, 1)$ with $\lambda = 2$ satisfies all conditions: Stationarity ($2(1) - 2 = 0$), Primal Feasibility ($2 - 1 - 1 = 0 \leq 0$), Dual Feasibility ($2 \geq 0$), and Complementary Slackness ($2(0) = 0$).

Step 5: Conclusion. The optimal point is $(x^*, y^*) = (1, 1)$.

Problem 4.10. For the problem "Minimize $x^2 + y^2$ subject to $x + y \geq 2$", we found the primal solution is $(x^*, y^*) = (1, 1)$ and the dual solution is $\lambda^* = 2$. Verify that strong duality holds by calculating the primal and dual objective values.

Solution.

Step 1: Calculate the Primal Objective Value. The primal objective is $f(x, y) = x^2 + y^2$. At the optimal point $(1, 1)$:

$$p^* = f(1, 1) = 1^2 + 1^2 = 2$$

Step 2: Formulate the Dual Function. The Lagrangian is $\mathcal{L}(x, y, \lambda) = x^2 + y^2 + \lambda(2 - x - y)$. The dual function $D(\lambda) = \inf_{x,y} \mathcal{L}$. The stationarity conditions are $x = \lambda/2$ and $y = \lambda/2$. Substituting these into the Lagrangian:

$$\begin{aligned} D(\lambda) &= (\lambda/2)^2 + (\lambda/2)^2 + \lambda(2 - \lambda/2 - \lambda/2) \\ &= \frac{\lambda^2}{4} + \frac{\lambda^2}{4} + \lambda(2 - \lambda) \\ &= \frac{\lambda^2}{2} + 2\lambda - \lambda^2 = 2\lambda - \frac{\lambda^2}{2} \end{aligned}$$

Step 3: Calculate the Dual Objective Value. The dual problem is to maximize $D(\lambda)$ subject to $\lambda \geq 0$. The optimal dual variable is $\lambda^* = 2$.

$$d^* = D(\lambda^*) = D(2) = 2(2) - \frac{2^2}{2} = 4 - 2 = 2$$

Step 4: Conclusion. Since the primal optimal value $p^* = 2$ is equal to the dual optimal value $d^* = 2$, **strong duality holds**.

5 Support Vector Machines (SVM)

SVMs are powerful classification models that work by finding the hyperplane that maximizes the margin between classes. This section includes problems on formulating the SVM optimization problem for different scenarios.

Problem 5.1. You are given a set of 2D data points.

- Class +1: $(3, 4), (5, 2)$
- Class -1: $(1, 1), (3, 0)$

Formulate the complete primal optimization problem for a soft-margin linear SVM. Use a regularization parameter $C = 10$.

Solution.

Step 1: State the General Primal Problem. For a soft-margin SVM, we want to find $w = (w_1, w_2)^T$ and b that solve:

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{subject to} \quad & y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \text{for } i = 1, \ldots, N \\ & \xi_i \geq 0, \quad \text{for } i = 1, \ldots, N \end{aligned}$$

Step 2: Define the Specific Objective Function. We have $N = 4$ points, $C = 10$, and $\|w\|^2 = w_1^2 + w_2^2$.

$$\text{Minimize} \quad \frac{1}{2}(w_1^2 + w_2^2) + 10(\xi_1 + \xi_2 + \xi_3 + \xi_4)$$

Step 3: Define the Specific Constraints.

- For $x_1 = (3, 4), y_1 = +1$: $1(w_1(3)+w_2(4)+b) \geq 1-\xi_1 \implies 3w_1+4w_2+b \geq 1-\xi_1$.
- For $x_2 = (5, 2), y_2 = +1$: $1(w_1(5)+w_2(2)+b) \geq 1-\xi_2 \implies 5w_1+2w_2+b \geq 1-\xi_2$.
- For $x_3 = (1, 1), y_3 = -1$: $-1(w_1(1)+w_2(1)+b) \geq 1-\xi_3 \implies -w_1-w_2-b \geq 1-\xi_3$.
- For $x_4 = (3, 0), y_4 = -1$: $-1(w_1(3) + w_2(0) + b) \geq 1 - \xi_4 \implies -3w_1 - b \geq 1 - \xi_4$.

Step 4: Conclusion. The complete primal optimization problem is:
Minimize over w_1, w_2, b, ξ_i:

$$\frac{1}{2}(w_1^2 + w_2^2) + 10(\xi_1 + \xi_2 + \xi_3 + \xi_4)$$

Subject to:

$$\begin{aligned} 3w_1 + 4w_2 + b &\geq 1 - \xi_1 \\ 5w_1 + 2w_2 + b &\geq 1 - \xi_2 \\ -w_1 - w_2 - b &\geq 1 - \xi_3 \\ -3w_1 - b &\geq 1 - \xi_4 \\ \xi_1, \xi_2, \xi_3, \xi_4 &\geq 0 \end{aligned}$$

Problem 5.2. Consider a 1D dataset that is not linearly separable:

- Class +1: $x = -1, x = 1$
- Class -1: $x = 0$

Use the feature mapping $\phi(x) = (x, x^2)$ to transform the data into a 2D feature space.

(a) List the transformed data points in the new feature space.

(b) Show that the data is now linearly separable and find a simple separating hyperplane by inspection.

(c) Formulate the hard-margin SVM primal problem for the transformed data.

Solution. (a) Transforming the Data

Step 1: Apply the feature map $\phi(x) = (x, x^2)$. Let the new coordinates be (z_1, z_2).

- For $x_1 = -1$ (+1): $\phi(-1) = (-1, 1)$.
- For $x_2 = 1$ (+1): $\phi(1) = (1, 1)$.
- For $x_3 = 0$ (-1): $\phi(0) = (0, 0)$.

Step 2: Conclusion. Transformed dataset: Class +1 at $\{(-1, 1), (1, 1)\}$, Class -1 at $\{(0, 0)\}$.

(b) Linear Separability

Step 1: Inspect the Transformed Points. The points for Class +1 are at a "height" of $z_2 = 1$, while the point for Class -1 is at the origin.

Step 2: Find a Separating Hyperplane. A horizontal line, such as $z_2 = 0.5$, clearly separates the classes. This corresponds to the hyperplane equation $0 \cdot z_1 + 1 \cdot z_2 - 0.5 = 0$. So, a valid separator is $w = (0, 1)^T$ and $b = -0.5$.

(c) Hard-Margin SVM Formulation

Step 1: State the Objective. Minimize $\frac{1}{2}\|w\|^2 = \frac{1}{2}(w_1^2 + w_2^2)$.

Step 2: State the Constraints. Use the hard-margin constraint $y_i(w^T z_i + b) \geq 1$.

- For $z_1 = (-1, 1), y_1 = +1$: $1(w_1(-1) + w_2(1) + b) \geq 1 \implies -w_1 + w_2 + b \geq 1$.
- For $z_2 = (1, 1), y_2 = +1$: $1(w_1(1) + w_2(1) + b) \geq 1 \implies w_1 + w_2 + b \geq 1$.
- For $z_3 = (0, 0), y_3 = -1$: $-1(w_1(0) + w_2(0) + b) \geq 1 \implies -b \geq 1 \implies b \leq -1$.

Step 3: Conclusion. The complete hard-margin SVM problem is:

$$\begin{aligned} \textbf{Minimize} \quad & \frac{1}{2}(w_1^2 + w_2^2) \\ \textbf{subject to} \quad & -w_1 + w_2 + b \geq 1 \\ & w_1 + w_2 + b \geq 1 \\ & b \leq -1 \end{aligned}$$

Problem 5.3. Explain two primary reasons why the dual formulation of the SVM optimization problem is often preferred over the primal formulation in practice.

Solution. The dual formulation is often preferred for two key reasons:

Reason 1: Enabling the Kernel Trick. The dual objective function depends only on the dot products of the input data points $(x_i \cdot x_j)$. This structure allows for the "kernel trick," where we can replace the dot product with a kernel function, $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$. This implicitly maps the data to a higher-dimensional feature space and allows the SVM to learn non-linear decision boundaries without ever explicitly computing the high-dimensional feature vectors $\phi(x)$. This is not possible in the primal formulation, which depends directly on the weight vector w.

Reason 2: Computational Efficiency in High Dimensions. In the primal problem, the number of optimization variables is determined by the dimensionality of the feature space, D (i.e., the components of w). In the dual problem, the number of optimization variables is the number of data points, N (i.e., the Lagrange multipliers α_i). In many machine learning problems, the number of features is much larger than the number of training samples ($D \gg N$). In such cases, solving the dual problem with N variables is computationally much more efficient than solving the primal with D variables.

Problem 5.4. An SVM classifier uses the decision function $f(x) = w^T x + b$. For a given weight vector $w = (1, 1)^T$ and bias $b = -1.5$, calculate the Hinge Loss for the following data points:

(a) $x_1 = (2, 2)^T$, with true label $y_1 = +1$.

(b) $x_2 = (1, -1)^T$, with true label $y_2 = -1$.

(c) $x_3 = (0, 1)^T$, with true label $y_3 = +1$.

Solution.

Step 1: Recall the Hinge Loss Formula. The Hinge Loss is defined as $L = \max(0, 1 - y_i(w^T x_i + b))$.

Step 2: Calculate Loss for Point (a).

- Compute the score: $w^T x_1 + b = (1)(2) + (1)(2) - 1.5 = 2.5$.
- Compute the margin product: $y_1(\text{score}) = (+1)(2.5) = 2.5$.
- Compute the loss: $L_1 = \max(0, 1 - 2.5) = \max(0, -1.5) = 0$.
- **Interpretation:** The point is correctly classified with a margin greater than 1, so the loss is zero.

Step 3: Calculate Loss for Point (b).

- Compute the score: $w^T x_2 + b = (1)(1) + (1)(-1) - 1.5 = -1.5$.
- Compute the margin product: $y_2(\text{score}) = (-1)(-1.5) = 1.5$.
- Compute the loss: $L_2 = \max(0, 1 - 1.5) = \max(0, -0.5) = 0$.
- **Interpretation:** The point is correctly classified with a margin greater than 1, so the loss is zero.

Step 4: Calculate Loss for Point (c).

- Compute the score: $w^T x_3 + b = (1)(0) + (1)(1) - 1.5 = -0.5$.
- Compute the margin product: $y_3(\text{score}) = (+1)(-0.5) = -0.5$.
- Compute the loss: $L_3 = \max(0, 1 - (-0.5)) = \max(0, 1.5) = 1.5$.

- **Interpretation:** The point is misclassified, resulting in a positive hinge loss.

Problem 5.5. Explain the role of the regularization parameter C in a soft-margin SVM. What is the effect of using a very large C versus a very small C?

Solution.

Step 1: Role of C. The parameter C in the soft-margin SVM objective function, $\min \frac{1}{2}\|w\|^2 + C\sum \xi_i$, controls the trade-off between maximizing the margin and minimizing the classification error on the training data. It is a penalty term for misclassification.

Step 2: Effect of a Very Large C.

- A large value of C places a high penalty on margin violations (i.e., on non-zero slack variables ξ_i).
- The optimization will prioritize classifying all training points correctly, even if it means choosing a hyperplane with a smaller margin.
- This can lead to a model with a **narrow margin** and a complex decision boundary that may **overfit** the training data. The behavior approaches that of a hard-margin SVM.

Step 3: Effect of a Very Small C.

- A small value of C places a low penalty on margin violations.
- The optimization will prioritize finding a hyperplane with a large margin, even if it misclassifies some training points.
- This leads to a model with a **wide margin** and a simpler decision boundary that is more tolerant of outliers. It can lead to **underfitting** if C is too small.

Problem 5.6. A linear SVM for a 2D dataset has been trained, yielding the optimal parameters $w = (1, 2)^T$ and $b = -5$. The original dataset contained the following points. Which of them are support vectors?

- Point A: $x_A = (1, 1)^T$, label $y_A = -1$.
- Point B: $x_B = (3, 1)^T$, label $y_B = +1$.
- Point C: $x_C = (4, 2)^T$, label $y_C = +1$.

Solution.

Step 1: Recall the Support Vector Condition. A point x_i is a support vector if it lies on or within the margin, meaning it satisfies the condition $y_i(w^T x_i + b) \leq 1$.

Step 2: Check Point A: $x_A = (1, 1)^T, y_A = -1$.

- Calculate the score: $w^T x_A + b = (1)(1) + (2)(1) - 5 = -2$.
- Calculate the margin product: $y_A(\text{score}) = (-1)(-2) = 2$.
- Since $2 > 1$, Point A is not a support vector. (It is correctly classified and outside the margin).

Step 3: Check Point B: $x_B = (3, 1)^T, y_B = +1$.

- Calculate the score: $w^T x_B + b = (1)(3) + (2)(1) - 5 = 0$.

- Calculate the margin product: $y_B(\text{score}) = (+1)(0) = 0$.
- Since $0 \leq 1$, this point violates the margin. It is a **support vector**.

Step 4: Check Point C: $x_C = (4, 2)^T, y_C = +1$**.**

- Calculate the score: $w^T x_C + b = (1)(4) + (2)(2) - 5 = 3$.
- Calculate the margin product: $y_C(\text{score}) = (+1)(3) = 3$.
- Since $3 > 1$, Point C is not a support vector. (It is correctly classified and outside the margin).

Step 5: Conclusion. Based on the definition, Point B is the only support vector among the three.

Problem 5.7. For the SVM solution from the previous problem ($w = (1, 2)^T, b = -5$), what is the width of the margin between the two classes?

Solution.

Step 1: Recall the Margin Width Formula. The margin is defined by the hyperplanes $w^T x + b = 1$ and $w^T x + b = -1$. The perpendicular distance between them, which is the total margin width, is given by $\frac{2}{\|w\|}$.

Step 2: Calculate the Norm of w**.**

$$\|w\| = \sqrt{w_1^2 + w_2^2} = \sqrt{1^2 + 2^2} = \sqrt{1 + 4} = \sqrt{5}$$

Step 3: Calculate the Margin Width.

$$\text{Width} = \frac{2}{\|w\|} = \frac{2}{\sqrt{5}}$$

Step 4: Conclusion. The width of the margin is $\frac{2}{\sqrt{5}} \approx 0.894$.

Problem 5.8. Using the trained SVM model with $w = (1, 2)^T$ and $b = -5$, classify the new data point $x_{new} = (4, 0)^T$. Also, calculate its signed distance to the decision hyperplane.

Solution.

Step 1: Recall the Classification Rule. A new point x is classified based on the sign of the decision function $f(x) = w^T x + b$. If $f(x) > 0$, it belongs to Class +1. If $f(x) < 0$, it belongs to Class -1.

Step 2: Calculate the Decision Function Value.

$$f(x_{new}) = w^T x_{new} + b = (1)(4) + (2)(0) - 5 = 4 - 5 = -1$$

Step 3: Classify the Point. Since $f(x_{new}) = -1 < 0$, the point $(4, 0)^T$ is classified as belonging to **Class -1**.

Step 4: Calculate the Signed Distance. The signed distance of a point x to the hyperplane $w^T x + b = 0$ is given by the formula $\frac{w^T x + b}{\|w\|}$.

- We already know $w^T x_{new} + b = -1$.

- We calculate the norm $\|w\| = \sqrt{1^2 + 2^2} = \sqrt{5}$.
- The distance is $\frac{-1}{\sqrt{5}}$.

Step 5: Conclusion. The point is classified as Class -1. Its signed distance from the decision boundary is $-\frac{1}{\sqrt{5}}$, meaning it is on the negative side of the hyperplane at a perpendicular distance of $\frac{1}{\sqrt{5}}$.

Problem 5.9. An SVM was trained on a dataset, and the dual solution identified three support vectors with non-zero Lagrange multipliers (α_i):

- $x_1 = (2, 2)^T, y_1 = +1, \alpha_1 = 0.5$
- $x_2 = (1, 0)^T, y_2 = -1, \alpha_2 = 0.3$
- $x_3 = (3, 0)^T, y_3 = -1, \alpha_3 = 0.2$

Calculate the optimal weight vector w.

Solution.

Step 1: Recall the Formula for w. The optimal weight vector w can be computed as a linear combination of the support vectors, weighted by their corresponding labels and dual variables:

$$w = \sum_{i \in SV} \alpha_i y_i x_i$$

Step 2: Calculate the Contribution of Each Support Vector.

$$\alpha_1 y_1 x_1 = 0.5(+1) \begin{pmatrix} 2 \\ 2 \end{pmatrix} = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix}$$
$$\alpha_2 y_2 x_2 = 0.3(-1) \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -0.3 \\ 0 \end{pmatrix}$$
$$\alpha_3 y_3 x_3 = 0.2(-1) \begin{pmatrix} 3 \\ 0 \end{pmatrix} = \begin{pmatrix} -0.6 \\ 0 \end{pmatrix}$$

Step 3: Sum the Contributions to find w.

$$w = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix} + \begin{pmatrix} -0.3 \\ 0 \end{pmatrix} + \begin{pmatrix} -0.6 \\ 0 \end{pmatrix} = \begin{pmatrix} 1.0 - 0.3 - 0.6 \\ 1.0 + 0 + 0 \end{pmatrix} = \begin{pmatrix} 0.1 \\ 1.0 \end{pmatrix}$$

Step 4: Conclusion. The optimal weight vector is $w = (0.1, 1.0)^T$.

Problem 5.10. Using the optimal weight vector $w = (0.1, 1.0)^T$ from the previous problem, find the bias term b. Assume the first support vector, $x_1 = (2, 2)^T$, lies exactly on the margin (i.e., its slack variable is zero).

Solution.

Step 1: Recall the Margin Condition for a Support Vector. For a support vector x_k with $\xi_k = 0$, the following equality holds:

$$y_k(w^T x_k + b) = 1$$

Step 2: Use the Given Support Vector. We are given $x_1 = (2, 2)^T$ and $y_1 = +1$.

$$(+1)\left(\begin{pmatrix}0.1\\1.0\end{pmatrix}^T\begin{pmatrix}2\\2\end{pmatrix} + b\right) = 1$$

Step 3: Solve for b.

$$\begin{aligned}(0.1)(2) + (1.0)(2) + b &= 1\\ 0.2 + 2.0 + b &= 1\\ 2.2 + b &= 1\\ b &= 1 - 2.2 = -1.2\end{aligned}$$

Step 4: Conclusion. The optimal bias term is $b = -1.2$. The final decision boundary is $0.1x_1 + 1.0x_2 - 1.2 = 0$.

6 Minimization and Maximization using Matrices

This section focuses on using calculus concepts, specifically the gradient and the Hessian matrix, to find and classify critical points of multivariate functions. This is the mathematical basis for many optimization algorithms.

Problem 6.1. Find and classify the critical points of the function $f(x, y) = \sin(x) + \sin(y)$ for $x, y \in [0, 2\pi]$.

Solution.

Step 1: Find the Critical Points by Setting the Gradient to Zero. First, we compute the gradient of $f(x, y)$:

$$\nabla f(x, y) = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix} = \begin{pmatrix} \cos(x) \\ \cos(y) \end{pmatrix}$$

Set the gradient to the zero vector:

$$\cos(x) = 0 \implies x = \frac{\pi}{2}, \frac{3\pi}{2} \quad \text{for } x \in [0, 2\pi]$$
$$\cos(y) = 0 \implies y = \frac{\pi}{2}, \frac{3\pi}{2} \quad \text{for } y \in [0, 2\pi]$$

This gives four critical points in the specified domain: $(\frac{\pi}{2}, \frac{\pi}{2})$, $(\frac{\pi}{2}, \frac{3\pi}{2})$, $(\frac{3\pi}{2}, \frac{\pi}{2})$, and $(\frac{3\pi}{2}, \frac{3\pi}{2})$.

Step 2: Compute the Hessian Matrix. We compute the second-order partial derivatives to form the Hessian:

$$H(x, y) = \begin{pmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{pmatrix} = \begin{pmatrix} -\sin(x) & 0 \\ 0 & -\sin(y) \end{pmatrix}$$

Step 3: Classify Each Critical Point using the Second Derivative Test. The determinant of the Hessian is $D = (-\sin(x))(-\sin(y)) = \sin(x)\sin(y)$.

- **Point $(\frac{\pi}{2}, \frac{\pi}{2})$:** $f_{xx} = -\sin(\frac{\pi}{2}) = -1$. $D = \sin(\frac{\pi}{2})\sin(\frac{\pi}{2}) = (1)(1) = 1$. Since $D > 0$ and $f_{xx} < 0$, this is a **local maximum**.
- **Point $(\frac{\pi}{2}, \frac{3\pi}{2})$:** $D = \sin(\frac{\pi}{2})\sin(\frac{3\pi}{2}) = (1)(-1) = -1$. Since $D < 0$, this is a **saddle point**.
- **Point $(\frac{3\pi}{2}, \frac{\pi}{2})$:** $D = \sin(\frac{3\pi}{2})\sin(\frac{\pi}{2}) = (-1)(1) = -1$. Since $D < 0$, this is a **saddle point**.
- **Point $(\frac{3\pi}{2}, \frac{3\pi}{2})$:** $f_{xx} = -\sin(\frac{3\pi}{2}) = -(-1) = 1$. $D = \sin(\frac{3\pi}{2})\sin(\frac{3\pi}{2}) = (-1)(-1) = 1$. Since $D > 0$ and $f_{xx} > 0$, this is a **local minimum**.

Problem 6.2. Find and classify the critical point of the function $f(x, y) = 2x^2 + \ln(y) - 2xy$. Note the domain requires $y > 0$.

Solution.

Step 1: Compute the Gradient.

$$\frac{\partial f}{\partial x} = 4x - 2y$$
$$\frac{\partial f}{\partial y} = \frac{1}{y} - 2x$$

Step 2: Find the Critical Point by Setting the Gradient to Zero.

$$4x - 2y = 0 \implies y = 2x \quad (1)$$
$$\frac{1}{y} - 2x = 0 \quad (2)$$

Substitute (1) into (2):

$$\frac{1}{2x} - 2x = 0 \implies \frac{1}{2x} = 2x \implies 1 = 4x^2 \implies x^2 = \frac{1}{4}$$

Since the domain of $\ln(y)$ requires $y > 0$, and $y = 2x$, we must have $x > 0$. Thus, we take the positive root: $x = 1/2$. Then $y = 2(1/2) = 1$. The only critical point is $(1/2, 1)$.

Step 3: Compute the Hessian Matrix.

$$f_{xx} = 4, \quad f_{yy} = -\frac{1}{y^2}, \quad f_{xy} = -2$$

The Hessian is $H(x, y) = \begin{pmatrix} 4 & -2 \\ -2 & -1/y^2 \end{pmatrix}$.

Step 4: Classify the Critical Point $(1/2, 1)$**.** Evaluate the Hessian at the point:

$$H(1/2, 1) = \begin{pmatrix} 4 & -2 \\ -2 & -1 \end{pmatrix}$$

Calculate the determinant: $D = (4)(-1) - (-2)^2 = -4 - 4 = -8$.

Step 5: Conclusion. Since $D = -8 < 0$, the critical point $(1/2, 1)$ is a **saddle point**.

Problem 6.3. Find and classify the critical points of the function $f(x, y) = (x^2 + y^2)e^{-x}$.

Solution.

Step 1: Compute the Gradient using the Product Rule.

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

Step 2: Find the Critical Points by Setting the Gradient to Zero. Since e^{-x} is never zero, we only need to solve:

$$2x - x^2 - y^2 = 0 \quad (1)$$
$$2y = 0 \implies y = 0 \quad (2)$$

Substitute $y = 0$ into equation (1):

$$2x - x^2 = 0 \implies x(2 - x) = 0$$

This gives $x = 0$ or $x = 2$. The critical points are **(0,0)** and **(2,0)**.

Step 3: Compute the Hessian Matrix.

$$f_{xx} = \frac{\partial}{\partial x}(e^{-x}(2x - x^2 - y^2)) = -e^{-x}(2x - x^2 - y^2) + e^{-x}(2 - 2x) = e^{-x}(x^2 - 4x + 2 - y^2)$$
$$f_{yy} = \frac{\partial}{\partial y}(2ye^{-x}) = 2e^{-x}$$
$$f_{xy} = \frac{\partial}{\partial y}(e^{-x}(2x - x^2 - y^2)) = e^{-x}(-2y) = -2ye^{-x}$$

Step 4: Classify Each Critical Point.

- **Point (0,0):** $H(0,0) = e^0 \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$. $D = (2)(2) - 0 = 4$. Since $D > 0$ and $f_{xx} > 0$, this is a **local minimum**.
- **Point (2,0):** $H(2,0) = e^{-2} \begin{pmatrix} 2^2 - 4(2) + 2 - 0^2 & 0 \\ 0 & 2 \end{pmatrix} = e^{-2} \begin{pmatrix} -2 & 0 \\ 0 & 2 \end{pmatrix}$. $D = (e^{-2})(-2) \cdot (e^{-2})(2) = -4e^{-4}$. Since $D < 0$, this is a **saddle point**.