Linear Algebra

February 10, 2026

Scalar: A single numerical value. It represents magnitude without direction. For example: $a = 5$ or $b = -3.14$

Vector: An ordered array of numbers arranged in a single column ($n \times 1$) or row ($1 \times n$). It is used to represent features or quantities with direction. For example:

$$X = \begin{bmatrix} 10 \\ 20 \\ 30 \end{bmatrix}$$

Matrix: A two-dimensional array of numbers arranged in rows and columns ($m \times n$, where m is the number of rows and n is the number of columns). It is used to represent multiple features, data points, or linear transformations. For example:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

Set & Subset:

- Set $S \rightarrow$ A collection of elements, $S = \{1, 2, 3\}$
- Subset $A \subseteq B \rightarrow$ All elements of A are in B, $A = \{1, 2\}$, $B = \{1, 2, 3, 4, 5\}$
- Universal Set $U \rightarrow$ The complete set under consideration, $U = \{1, 2, 3, 4, 5, 6, \dots\}$
- Empty Set $\emptyset \rightarrow$ A set without any elements.
- Union $A \cup B \rightarrow$ A set that contains all the elements from both the sets, $A = \{1, 2, 3\}$, $B = \{3, 4, 5\}$, $A \cup B = \{1, 2, 3, 4, 5\}$
- Intersection $A \cap B \rightarrow$ A set that contains only the elements that are present in both the sets, $A = \{1, 2, 3\}$, $B = \{3, 4, 5\}$, $A \cap B = \{3\}$

Vector Space: A set V of vectors together with two operations (vector addition and scalar multiplication) that satisfy the following axioms:

1. Closure under addition: $\mathbf{u} + \mathbf{v} \in V$ for all $\mathbf{u}, \mathbf{v} \in V$
2. Closure under scalar multiplication: $c\mathbf{u} \in V$ for all $\mathbf{u} \in V$ and scalar c
3. Addition is commutative: $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$
4. Addition is associative: $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$
5. Existence of zero vector: There exists $\mathbf{0} \in V$ such that $\mathbf{u} + \mathbf{0} = \mathbf{u}$
6. Existence of additive inverse: For each $\mathbf{u} \in V$, there exists $-\mathbf{u} \in V$ such that $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$
7. Distributivity: $c(\mathbf{u} + \mathbf{v}) = c\mathbf{u} + c\mathbf{v}$ and $(c + d)\mathbf{u} = c\mathbf{u} + d\mathbf{u}$

8. Scalar multiplication is associative: $(cd)\mathbf{u} = c(d\mathbf{u})$

9. Identity element: $1\mathbf{u} = \mathbf{u}$

For example: $\mathbb{R}^3$ is a vector space consisting of all 3-dimensional vectors like $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ where $x, y, z \in \mathbb{R}$. If $\mathbf{u} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}$, then $\mathbf{u} + \mathbf{v} = \begin{bmatrix} 5 \\ 7 \\ 9 \end{bmatrix} \in \mathbb{R}^3$ and $2\mathbf{u} = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} \in \mathbb{R}^3$.

Sub-space: A subset of a vector space that is itself a vector space (it must contain the zero vector, be closed under addition, and closed under scalar multiplication). For example: In $\mathbb{R}^3$, all vectors of the form $\begin{bmatrix} x \\ y \\ 0 \end{bmatrix}$ form a sub-space ($\mathbb{R}^2$ - the xy-plane), because adding any two such vectors or multiplying by a scalar keeps the third component zero.

Linear Independence: A set of vectors is linearly independent if no vector can be written as a combination of the others. In other words, $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n = \mathbf{0}$ only when all $c_i = 0$. For example: Vectors $\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ are linearly independent, but $\mathbf{v}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\mathbf{v}_2 = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$ are not (since $\mathbf{v}_2 = 2\mathbf{v}_1$).

Norm: A measure of the length or magnitude of a vector, denoted as $\|\mathbf{v}\|$. The most common is the Euclidean norm (L2 norm): $\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$. For example: The norm of $\mathbf{v} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$ is $\|\mathbf{v}\| = \sqrt{3^2 + 4^2} = \sqrt{9 + 16} = 5$. Commonly used norms:

- Lp Norm (General Norm)
- L1 Norm (Manhattan Norm)
- L2 Norm (Euclidean Norm)
- L∞ Norm (Chebyshev Distance)

Matrix Addition & Subtraction: Two matrices of the same dimensions can be added or subtracted by adding or subtracting their corresponding elements. For example: If $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$, then $A + B = \begin{bmatrix} 1+5 & 2+6 \\ 3+7 & 4+8 \end{bmatrix} = \begin{bmatrix} 6 & 8 \\ 10 & 12 \end{bmatrix}$ and $A - B = \begin{bmatrix} 1-5 & 2-6 \\ 3-7 & 4-8 \end{bmatrix} = \begin{bmatrix} -4 & -4 \\ -4 & -4 \end{bmatrix}$.

Property	Addition	Subtraction
Commutative	$A + B = B + A$	$A + B \neq B - A$
Associative	$(A + B) + C = A + (B + C)$	$A - (B - C) = (A - B) + C$
Additive Identity	$A + 0 = A$	$A - 0 = A$
Additive Inverse	$A + (-A) = 0$	$A - A = 0$

Scalar addition and subtractions are allowed as well. In that case, the scalar is added to or subtracted from all the elements. For example: $2 + A = \begin{bmatrix} 1+2 & 2+2 \\ 3+2 & 4+2 \end{bmatrix} = \begin{bmatrix} 3 & 4 \\ 5 & 6 \end{bmatrix}$.

Trace: The sum of all diagonal elements of a square matrix, denoted as $\text{tr}(A)$ or $\text{Tr}(A)$. For example: For matrix $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$, the trace is $\text{tr}(A) = 1 + 5 + 9 = 15$.

Matrix Multiplication: To multiply two matrices A and B, the number of columns in A must equal the number of rows in B. If A is $m \times n$ and B is $n \times p$, the resulting matrix C will be $m \times p$. Each element C_{ij} is computed as the dot product of the i-th row of A and the j-th column of B.

For example, multiplying a 3×2 matrix with a 2×3 matrix:

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}, \quad B = \begin{bmatrix} 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix}$$

Step-by-step calculation of $C = AB$ (which will be 3×3):

$$\begin{aligned} C_{11} &= (1)(7) + (2)(10) = 7 + 20 = 27 \\ C_{12} &= (1)(8) + (2)(11) = 8 + 22 = 30 \\ C_{13} &= (1)(9) + (2)(12) = 9 + 24 = 33 \\ C_{21} &= (3)(7) + (4)(10) = 21 + 40 = 61 \\ C_{22} &= (3)(8) + (4)(11) = 24 + 44 = 68 \\ C_{23} &= (3)(9) + (4)(12) = 27 + 48 = 75 \\ C_{31} &= (5)(7) + (6)(10) = 35 + 60 = 95 \\ C_{32} &= (5)(8) + (6)(11) = 40 + 66 = 106 \\ C_{33} &= (5)(9) + (6)(12) = 45 + 72 = 117 \end{aligned}$$

Therefore:

$$C = AB = \begin{bmatrix} 27 & 30 & 33 \\ 61 & 68 & 75 \\ 95 & 106 & 117 \end{bmatrix}$$

Property	Formula	Hold Always?
Associative	$(AB)C = A(BC)$	YES
Distributive	$A(B + C) = (AB + AC)$	YES
Identity Matrix	$AI = IA = A$	YES
Commutative	$AB = BA$	NO

Diagonal Matrix: A square matrix with non-zero elements only on the main diagonal.

$$D = \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{bmatrix}$$

Properties:

- Transpose is itself: $D^T = D$.
- Easy to compute inverse:

$$D^{-1} = \begin{bmatrix} \frac{1}{d_1} & 0 & 0 \\ 0 & \frac{1}{d_2} & 0 \\ 0 & 0 & \frac{1}{d_3} \end{bmatrix}$$

Identity Matrix: An identity matrix is a square matrix with 1s on the main diagonal and 0s everywhere else.

$$I_n = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Properties:

- Multiplication Identity: $AI = IA$
- Inverse is itself: $I^{-1} = I$

Transpose of a Matrix: If a matrix A has a shape of $m \times n$, then it's transpose A^T (or A') will have a shape of $n \times m$.

- Rows become columns.
- Columns become row.

For example:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}_{2\times 3} \quad \rightarrow \quad A^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}_{3\times 2}$$

Property	Mathematical Expression	Explanation
Double Transpose	$(A^T)^T = A$	Taking the transpose twice returns the original matrix.
Transpose of a Sum	$(A + B)^T = A^T + B^T$	Transpose of a sum is the sum of transposes.
Transpose of a Product	$(AB)^T = B^T A^T$	Transpose of a product reverses the order of multiplication.
Scalar Multiplication	$(cA)^T = cA^T$	The scalar factor c remains unchanged when transposing.
Symmetric Matrix Condition	$A = A^T$	A square matrix is symmetric if it equals to it's transpose.
Skew-Symmetric Matrix	$A^T = -A$	A matrix is skew-symmetric if it's transpose is it's negative.
Orthogonal Matrix	$A^T A = I$	A matrix is orthogonal if it's transpose is equals to it's inverse ($A^T = A^{-1}$).

Determinant: Based on determinant of a matrix -

- We can decide whether the matrix is invertible (A^{-1}).
- Whether a system of linear equations has a unique solution.
- It is also import for volume scaling in linear transformations.

For a 2×2 matrix $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$, $\det(A) = (a_{11} \times a_{22}) - (a_{12} \times a_{21})$. Example:

$$\begin{aligned} A &= \begin{bmatrix} 2 & 3 \\ 1 & 4 \end{bmatrix} \\ \det(A) &= (2 \times 4) - (3 \times 1) \\ &= 8 - 3 = 5 \end{aligned}$$

For a 3×3 matrix $A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$, $\det(A) = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$.

Example:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

$$\begin{aligned} \det(A) &= 1 \times \begin{vmatrix} 5 & 6 \\ 8 & 9 \end{vmatrix} - 2 \times \begin{vmatrix} 4 & 6 \\ 7 & 9 \end{vmatrix} + 3 \times \begin{vmatrix} 4 & 5 \\ 7 & 8 \end{vmatrix} \\ &= 1 \times (5 \times 9 - 6 \times 8) - 2 \times (4 \times 9 - 6 \times 7) + 3 \times (4 \times 8 - 5 \times 7) \\ &= 1 \times (45 - 48) - 2 \times (36 - 42) + 3 \times (32 - 35) \\ &= -3 + 12 - 9 = 0 \end{aligned}$$

$\det(A) = 0$	$\det(A) \neq 0$
Singular Matrix	Non-singular Matrix
Not invertible	Invertible
Row/Column vectors are linearly dependent (not orthonormal)	Linearly independent

Inverse Matrix (2×2): Inverse of a 2×2 matrix $A = \begin{bmatrix} a11 & a12 \\ a21 & a22 \end{bmatrix}$ is $A^{-1} = \frac{1}{\det(A)} \begin{bmatrix} a22 & -a12 \\ -a21 & a11 \end{bmatrix}$.

Example:

$$A = \begin{bmatrix} 2 & 3 \\ 1 & 4 \end{bmatrix}$$

$$\begin{aligned} \det(A) &= (2 \times 4) - (3 \times 1) \\ &= 5 \end{aligned}$$

$$\begin{aligned} A^{-1} &= \frac{1}{5} \begin{bmatrix} 4 & -3 \\ -1 & 2 \end{bmatrix} \\ &= \begin{bmatrix} \frac{4}{5} & -\frac{3}{5} \\ -\frac{1}{5} & \frac{2}{5} \end{bmatrix} \end{aligned}$$

$$\begin{aligned} AA^{-1} &= \begin{bmatrix} \frac{8}{5} - \frac{3}{5} & -\frac{6}{5} + \frac{6}{5} \\ \frac{4}{5} - \frac{4}{5} & -\frac{3}{5} + \frac{8}{5} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= I \end{aligned}$$

Sparse Matrix and Dense Matrix:

- **Sparse Matrix:** Matrix that has mostly zero elements and a few non-zero elements. Characteristics -
  - Most elements are zero (typically, more that 50%).
  - Memory efficient.
  - Specialized operations (compressed storage formats, optimized computation).
  - Commonly found in Natural Language Processing (NLP), Recommendation Systems, Computer Vision, Anomaly Detection, Genomics & Bioinformatics, Graph Based Data (Social Networks, Knowledge Graphs), High Dimensional Feature Spaces, etc.

– **Dense Matrix:** Matrix that has mostly non-zero elements. Characteristics -

- Few or no zero elements.
- Standard storage format (2D array/matrix).
- Computationally expensive for large sizes.
- Used in general numerical computation.

Rank of a Matrix: The number of linearly independent row or column. It represents the dimensions of row space or column space. Example:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix}$$

Step-1: Identify the column vectors.

$$C_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad C_2 = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}, \quad C_3 = \begin{bmatrix} 3 \\ 6 \\ 9 \end{bmatrix}$$

We check if these columns are linearly independent by expressing one column as a combination of others.

- Observe that $C_2 = 2C_1$ and $C_3 = 3C_1$.
- Since all columns are multiple of C_1, there is only one linearly independent column.

Thus, the rank of A is 1.

Properties of Rank:

- Row rank = Column rank.
- If $\det(A) \neq 0$, rank is n (full rank).
- If $\det(A) = 0$, rank is $< n$ (rank-deficient). For this type of matrices, it is easy to get the rank by transforming the matrix into **row-echelon** form. Then, just count the number of non-zero rows.
- Only a zero matrix has a rank of 0.

Variance, Covariance, Covariance Matrix, & Correlation:

– **Variance:** Measures how much a single variable varies from its mean. It quantifies the spread of data points.

$$\text{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Example: For dataset $X = [2, 4, 6]$, mean $\bar{x} = 4$.

$$\begin{aligned} \text{Var}(X) &= \frac{1}{3}[(2-4)^2 + (4-4)^2 + (6-4)^2] \\ &= \frac{1}{3}[4 + 0 + 4] = \frac{8}{3} \approx 2.67 \end{aligned}$$

– **Covariance:** A statistical measurement that quantifies the relationship between two random variables. It indicates whether they increase together (positive) or inversely (negative).

$$\text{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Example: For datasets $X = [1, 2, 3]$ and $Y = [2, 4, 6]$, with $\bar{x} = 2$ and $\bar{y} = 4$.

$$\begin{aligned}\text{Cov}(X, Y) &= \frac{1}{3}[(1-2)(2-4) + (2-2)(4-4) + (3-2)(6-4)] \\ &= \frac{1}{3}[(-1)(-2) + 0 + (1)(2)] \\ &= \frac{1}{3}[2 + 0 + 2] = \frac{4}{3} \approx 1.33\end{aligned}$$

- $\text{Cov}(X, Y) > 0$, a positive covariance. X and Y tend to increase or decrease together (bidirectional). Example: Height and Weight.
- $\text{Cov}(X, Y) < 0$, a negative covariance. When X increases, Y tends to decrease, and vice versa (bidirectional). Example: Speed of a car and time to reach a destination.
- $\text{Cov}(X, Y) = 0$, no linear relationship. Changes in X don't linearly relate to changes in Y. Example: Shoe size and IQ.

– **Covariance Matrix:** A square matrix showing covariances between multiple variables. For variables $X_1, X_2, \ldots, X_n$, the covariance matrix is:

$$\Sigma = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \cdots & \text{Var}(X_n) \end{bmatrix}$$

Example: For two variables $X = [1, 2, 3]$ and $Y = [2, 4, 6]$ with $\text{Var}(X) = \frac{2}{3}$, $\text{Var}(Y) = \frac{8}{3}$, and $\text{Cov}(X, Y) = \frac{4}{3}$:

$$\Sigma = \begin{bmatrix} \frac{2}{3} & \frac{4}{3} \\ \frac{4}{3} & \frac{8}{3} \end{bmatrix}$$

- $\text{Var}(X_1 \pm X_2) = \text{Var}(X_1) + \text{Var}(X_2) \pm \text{Cov}(X_1, X_2)$.
- $\text{Cov}(X_1 + X_2, X_3 - X_4) = \text{Cov}(X_1, X_3) + \text{Cov}(X_1, -X_4) + \text{Cov}(X_2, X_3) + \text{Cov}(X_2, -X_4)$ (here $\pm$ doesn't apply, means $-X_4 = X_4$).

– **Correlation:** A normalized measure of the linear relationship between two variables, ranging from -1 to 1. It is dimensionless and scale-independent.

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

where $\sigma_X = \sqrt{\text{Var}(X)}$ and $\sigma_Y = \sqrt{\text{Var}(Y)}$ are standard deviations.

Example: Using previous data $X = [1, 2, 3]$ and $Y = [2, 4, 6]$ with $\text{Var}(X) = \frac{2}{3}$, $\text{Var}(Y) = \frac{8}{3}$, and $\text{Cov}(X, Y) = \frac{4}{3}$:

$$\begin{aligned}\sigma_X &= \sqrt{\frac{2}{3}} \approx 0.816 \\ \sigma_Y &= \sqrt{\frac{8}{3}} \approx 1.633 \\ \text{Corr}(X, Y) &= \frac{\frac{4}{3}}{\sqrt{\frac{2}{3}} \cdot \sqrt{\frac{8}{3}}} \\ &= \frac{\frac{4}{3}}{\sqrt{\frac{16}{9}}} = \frac{\frac{4}{3}}{\frac{4}{3}} = 1\end{aligned}$$

A correlation (ρ) of 1 means perfect positive linear relationship (as one increases, the other increases proportionally).

Eigenvalue: Eigenvalues are special numbers associated with square matrix that indicate how the matrix scales or transforms vectors. For a matrix A, an eigenvalue λ is scalar that satisfies the equation:

$$Av = \lambda v \quad \left| \begin{array}{l} A = \text{square matrix} \\ v = \text{eigenvector (non-zero)} \\ \lambda = \text{eigenvalue (scalar)} \end{array} \right.$$

Since λv represents scalar multiplication of the vector v, λ must be a scalar. Eigenvalues are found by solving the characteristic equation:

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

Example:

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

$$\begin{aligned} A - \lambda I &= \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 2-\lambda & -1 \\ -1 & 2-\lambda \end{bmatrix} \end{aligned}$$

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

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

The sign of eigenvalues determines the definiteness of a symmetric matrix A, which relates to the quadratic form $x^T Ax$ for all non-zero vectors x:

- **Positive Definite:** All eigenvalues $\lambda > 0 \Rightarrow x^T Ax > 0$ for all $x \neq 0$.
- **Negative Definite:** All eigenvalues $\lambda < 0 \Rightarrow x^T Ax < 0$ for all $x \neq 0$.
- **Positive Semi-Definite:** All eigenvalues $\lambda \geq 0 \Rightarrow x^T Ax \geq 0$ for all $x \neq 0$.
- **Negative Semi-Definite:** All eigenvalues $\lambda \leq 0 \Rightarrow x^T Ax \leq 0$ for all $x \neq 0$.
- **Indefinite:** Mixed positive and negative eigenvalues $\Rightarrow x^T Ax$ can be positive, negative, or zero. Occurs in **saddle-point** problems and optimization.

Convexity:

- **Jensen's Inequality:**

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y), \quad \forall x, y \text{ and } \lambda \in [0, 1]$$

Example:

$$f(x) = x^2$$

$$\begin{aligned} &\textit{Assume},\\ &x = 1, y = 3, \text{ and } \lambda = 0.5 \end{aligned}$$

$$\begin{aligned} LHS &= f(\lambda x + (1-\lambda)y) \\ &= f(0.5(1) + (1-0.5)3) \\ &= f(0.5 + 1.5) \\ &= f(2) \\ &= 2^2 = 4 \end{aligned}$$

$$\begin{aligned} RHS &= \lambda f(x) + (1-\lambda) f(y) \\ &= 0.5(1^2) + (1-0.5)(3^2) \\ &= 0.5 + 4.5 = 5 \end{aligned}$$

$LHS < RHS$, proving that $f(x)$ is convex.

- **Hessain Matrix:** A square-symmetric matrix of second order partial derivatives of scalar-valued function. It describes the local curvature of a function and is widely used in optimization, machine learning and numerical analysis.

$$H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial x \partial z} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} & \frac{\partial^2 f}{\partial y \partial z} \\ \frac{\partial^2 f}{\partial z \partial x} & \frac{\partial^2 f}{\partial z \partial y} & \frac{\partial^2 f}{\partial z^2} \end{bmatrix} \quad H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial {x_1}^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial {x_2}^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \cdots & \cdots & \ddots & \cdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial {x_n}^2} \end{bmatrix}$$

Hessain Type	Eigenvalues λ	Local Min/Max	Convexity
Positive Definite	All $\lambda > 0$	Local Minimum	Strictly Convex
Negative Definite	All $\lambda < 0$	Local Maximum	Strictly Concave
Positive Semi-Definite	All $\lambda \geq 0$	Possibly Local Minimum (Need Higher-Order Test)	Convex (Possibly flat in some directions)
Negative Semi-Definite	All $\lambda \leq 0$	Possibly Local Maximum (Need Higher-Order Test)	Concave (Possibly flat in some directions)
Indefinite	Mixed ($\pm$)	Saddle-Point (Neither Max nor Min)	Strictly Convex

Example:

$$f(x, y) = x^2 + y^2$$

$$\begin{aligned} \frac{\partial^2 f}{\partial x^2} &= \frac{\partial}{\partial x}(\frac{\partial f}{\partial x}) = \frac{\partial}{\partial x}(2x) = 2 \\ \frac{\partial^2 f}{\partial y^2} &= \frac{\partial}{\partial y}(\frac{\partial f}{\partial y}) = \frac{\partial}{\partial y}(2y) = 2 \\ \frac{\partial^2 f}{\partial x \partial y} &= \frac{\partial}{\partial y}(\frac{\partial f}{\partial x}) = \frac{\partial}{\partial y}(2x) = 0 \end{aligned}$$

$$
\begin{aligned}
H(f) &= \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \\
\det(H - \lambda I) &= 0 \\
\Rightarrow \begin{vmatrix} 2-\lambda & 0 \\ 0 & 2-\lambda \end{vmatrix} &= 0 \\
\Rightarrow (2-\lambda)(2-\lambda) &= 0 \\
\Rightarrow \lambda &= 2
\end{aligned}
$$

Since $f(x, y) > 0$, it's strictly convex.

System of Linear Equations:

- **Linear Equation:** Algebraic equation of a line where all the variables have a maximum exponent of 1. Example: $2x + 3y = 6$.
- **System of Linear Equations:** A set of two or more linear equations involving the same variables. Example:

$$
\begin{cases} 2x + 3y = 6 \\ x - y = 4 \end{cases}
$$

  - **Homogeneous System:** A system of linear equations where all the constant terms are 0. Example:

$$
\text{General Form} \begin{cases} a_1x + b_1y + c1_z = 0 \\ a_2x + b_2y + c2_z = 0 \\ a_3x + b_3y + c3_z = 0 \\ \vdots \end{cases} \qquad \begin{matrix} \text{Three Variable} \\ \text{Example} \end{matrix} \begin{cases} x + y + z = 0 \\ 2x - y + z = 0 \\ x + 2y - 3z = 0 \end{cases}
$$

Types of Solution:

- **Unique Solution:** The equations intersect at a single point (consistent and independent). Lines have different slopes.
- **No Solution:** The equations represent parallel lines that never intersect (inconsistent). Lines have the same slope but different y-intercepts.
- **Infinitely Many Solutions:** The equations represent the same line (consistent and dependent). All coefficients are proportional.

Commonly used methods to solve these systems:

- **Substitution Method:** Solve one equation for one variable and substitute into the other equation. Example:

$$
\begin{aligned}
&\text{Given:} \\
x + 2y &= 3 \quad \dots(1) \\
2x + 3y &= 6 \quad \dots(2)
\end{aligned}
$$

From equation (1), solve for x:

$$
x = 3 - 2y \quad \dots(3)
$$

Substitute (3) into equation (2):

$$
\begin{aligned}
2(3 - 2y) + 3y &= 6 \\
6 - 4y + 3y &= 6 \\
-y &= 0 \\
y &= 0
\end{aligned}
$$

Substitute $y = 0$ into equation (3):

$$x = 3 - 2(0)$$
$$x = 3$$

$$\therefore \text{Solution: } (x, y) = (3, 0)$$

- **Cramer's Rule:** A method for solving systems of linear equations using determinants. For a system $Ax = b$, each variable is computed as the ratio of two determinants: $x_i = \frac{D_i}{D}$, where $D = \det(A)$ and D_i is the determinant of the matrix formed by replacing the i-th column of A with the constant vector b. Example:

Given:

$$\begin{aligned} x + y + z &= 6 \quad \dots(1) \\ 2x + 3y + z &= 14 \quad \dots(2) \\ x + 2y + 3z &= 14 \quad \dots(3) \end{aligned}$$

Coefficient Matrix: $A = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 1 \\ 1 & 2 & 3 \end{bmatrix}$ Constant Vector: $b = \begin{bmatrix} 6 \\ 14 \\ 14 \end{bmatrix}$

$$\begin{aligned} D = \det(A) &= 7 - 5 + 1 \\ &= 3 \end{aligned}$$

$$\begin{aligned} D_x &= \begin{vmatrix} 6 & 1 & 1 \\ 14 & 3 & 1 \\ 14 & 2 & 3 \end{vmatrix} \\ &= 42 - 28 - 14 \\ &= 0 \end{aligned}$$

$$\begin{aligned} D_y &= \begin{vmatrix} 1 & 6 & 1 \\ 2 & 14 & 1 \\ 1 & 14 & 3 \end{vmatrix} \\ &= 28 - 30 + 14 \\ &= 12 \end{aligned}$$

$$\begin{aligned} D_z &= \begin{vmatrix} 1 & 1 & 6 \\ 2 & 3 & 14 \\ 1 & 2 & 14 \end{vmatrix} \\ &= 14 - 14 + 6 \\ &= 6 \end{aligned}$$

$$\begin{aligned} x &= \frac{D_x}{D} = \frac{0}{3} = 0 \\ y &= \frac{D_y}{D} = \frac{12}{3} = 4 \\ z &= \frac{D_z}{D} = \frac{6}{3} = 2 \end{aligned}$$

$$\therefore \text{Solution: } (x, y, z) = (0, 4, 2)$$

When $\det(A) = 0$,

- **Case 1: Infinitely Many Solutions**
  - All D_i ($D_x, D_y, D_z, \dots$) are 0.
  - This means the equations are dependent (e.g., one is a combination of others), and the system has infinitely many solutions (like a plane of solutions).
- **Case 2: No Solution**
  - At least one of D_i is 0.
  - This means the equations contradict each other (e.g., parallel planes that never intersect), and the system is inconsistent.

- **Gaussian Elimination:** A systematic method to transform a system of linear equations into row echelon form (upper triangular matrix) by eliminating coefficients below the diagonal using elementary row operations. The solution is then obtained through back-substitution. Example:

Given:

$$\begin{aligned} x + y + z &= 6 \quad \dots (1) \\ 2x + 3y + z &= 14 \quad \dots (2) \\ x + 2y + 3z &= 14 \quad \dots (3) \end{aligned}$$

Augmented Matrix:

$$A|b = \left[\begin{array}{ccc|c} 1 & 1 & 1 & 6 \\ 2 & 3 & 1 & 14 \\ 1 & 2 & 3 & 14 \end{array}\right]$$

Step 1: Eliminate first column below diagonal

$$\begin{aligned} R_2' &= R_2 - 2R_1 \\ R_3' &= R_3 - R_1 \end{aligned}$$

$$A|b = \left[\begin{array}{ccc|c} 1 & 1 & 1 & 6 \\ 0 & 1 & -1 & 2 \\ 0 & 1 & 2 & 8 \end{array}\right]$$

Step 2: Eliminate second column below diagonal

$$R_3'' = R_3' - R_2'$$

$$A|b = \left[\begin{array}{ccc|c} 1 & 1 & 1 & 6 \\ 0 & 1 & -1 & 2 \\ 0 & 0 & 3 & 6 \end{array}\right]$$

Back-substitution:

$$\begin{aligned} &\text{From row 3: } 3z = 6 \Rightarrow z = 2 \\ &\text{From row 2: } y - z = 2 \Rightarrow y = 2 + z = 4 \\ &\text{From row 1: } x + y + z = 6 \Rightarrow x = 6 - y - z = 0 \end{aligned}$$

$$\therefore \text{Solution: } (x, y, z) = (0, 4, 2)$$

Elementary Row Operations (Constraints):

- **Row Replacement:** Replace any row with a linear combination of rows. Example: $R_2' = R_2 + 3R_1 - 2R_3$.
- **Row Swap:** Interchange any two rows. Example: $R_1 \leftrightarrow R_3$.
- **Row Scaling:** Multiply a row by any **non-zero** scalar. Example: $R_2' = 5R_2$.
- **Not Allowed:** Multiplying a row by zero (destroys information and makes the system unsolvable).

These operations preserve the solution set of the system (equivalent systems).

Inverse Matrix (3×3): The inverse of a 3×3 matrix can be found using **elementary row operations** on the augmented matrix $[A|I]$. Transform the left side to I and the right side becomes A^{-1}. Example:

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

$$\begin{aligned} \text{Goal: } AA^{-1} &= I \\ \text{Augmented Matrix} &= [A|I] \\ &= \left[\begin{array}{ccc|ccc} 1 & 2 & -1 & 1 & 0 & 0 \\ -2 & 0 & 1 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 & 0 & 1 \end{array}\right] \end{aligned}$$

$$\begin{aligned} R_2' &= R_2 + 2R_1 \\ R_3' &= R_3 - R_1 \end{aligned}$$

$$A|I = \left[\begin{array}{ccc|ccc} 1 & 2 & -1 & 1 & 0 & 0 \\ 0 & 4 & -1 & 2 & 1 & 0 \\ 0 & -3 & 1 & -1 & 0 & 1 \end{array}\right]$$

$$R_3'' = 4R_3' + 3R_2'$$

$$A|I = \left[\begin{array}{ccc|ccc} 1 & 2 & -1 & 1 & 0 & 0 \\ 0 & 4 & -1 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 3 & 4 \end{array}\right]$$

$$R_2''' = R_2' + R_3''$$

$$A|I = \left[\begin{array}{ccc|ccc} 1 & 2 & 0 & 3 & 3 & 4 \\ 0 & 4 & 0 & 4 & 4 & 4 \\ 0 & 0 & 1 & 2 & 3 & 4 \end{array}\right]$$

$$R_2^{(4)} = \frac{1}{4} R_2'''$$

$$A|I = \left[\begin{array}{ccc|ccc} 1 & 2 & 0 & 3 & 3 & 4 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 2 & 3 & 4 \end{array}\right]$$

$$R_1' = R_1 - 2R_2^{(4)}$$

$$A|I = \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 1 & 2 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 2 & 3 & 4 \end{array}\right]$$

$$\therefore A^{-1} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 1 \\ 2 & 3 & 4 \end{bmatrix}$$

Verification:

$$\begin{aligned} AA^{-1} &= \begin{bmatrix} 1 & 2 & -1 \\ -2 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 1 \\ 2 & 3 & 4 \end{bmatrix} \\ &= \begin{bmatrix} 1+2-2 & 1+2-3 & 2+2-4 \\ -2+0+2 & -2+0+3 & -4+0+4 \\ 1-1+0 & 1-1+0 & 2-1+0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I \end{aligned}$$

LU Decomposition: A method of factorizing a square matrix A into the product of a lower triangular matrix L and an upper triangular matrix U, such that $A = LU$. This decomposition is particularly useful for:

- Efficiently solving multiple systems of linear equations with the same coefficient matrix
- Computing determinants: $\det(A) = \det(L) \times \det(U)$
- Finding matrix inverses
- Numerical stability in computations

The decomposition process uses Gaussian Elimination without row swaps (or with partial pivoting if needed). The matrix L contains the multipliers used during elimination, and U is the resulting upper triangular matrix.
Example: Find the LU decomposition of matrix A and use it to solve $Ax = b$.

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 4 & 3 & 3 \\ 8 & 7 & 9 \end{bmatrix}, \quad b = \begin{bmatrix} 4 \\ 10 \\ 24 \end{bmatrix}$$

Step 1: Perform Gaussian Elimination to get U and record multipliers for L

$$\text{Original matrix: } A = \begin{bmatrix} 2 & 1 & 1 \\ 4 & 3 & 3 \\ 8 & 7 & 9 \end{bmatrix}$$

Eliminate first column:

$$m_{21} = \frac{4}{2} = 2, \quad m_{31} = \frac{8}{2} = 4$$
$$R_2' = R_2 - 2R_1, \quad R_3' = R_3 - 4R_1$$

$$\begin{bmatrix} 2 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 3 & 5 \end{bmatrix}$$

Eliminate second column:

$$m_{32} = \frac{3}{1} = 3$$
$$R_3'' = R_3' - 3R_2'$$

$$U = \begin{bmatrix} 2 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix}$$

Step 2: Construct L using the multipliers

$$L = \begin{bmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & m_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 4 & 3 & 1 \end{bmatrix}$$

$$\text{Verification: } LU = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 4 & 3 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 1 & 1 \\ 4 & 3 & 3 \\ 8 & 7 & 9 \end{bmatrix} = A$$

Step 3: Solve $Ax = b$ using $A = LU$

Since $Ax = b$ and $A = LU$, we have $LUx = b$
Let $Ux = y,$ then $Ly = b$

First, solve $Ly = b$ (forward substitution):

$$\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 4 & 3 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 4 \\ 10 \\ 24 \end{bmatrix}$$

$$y_1 = 4$$
$$2y_1 + y_2 = 10 \Rightarrow y_2 = 10 - 8 = 2$$
$$4y_1 + 3y_2 + y_3 = 24 \Rightarrow y_3 = 24 - 16 - 6 = 2$$

$$\therefore y = \begin{bmatrix} 4 \\ 2 \\ 2 \end{bmatrix}$$

Then, solve $Ux = y$ (back-substitution):

$$\begin{bmatrix} 2 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \\ 2 \end{bmatrix}$$

$$2x_3 = 2 \Rightarrow x_3 = 1$$
$$x_2 + x_3 = 2 \Rightarrow x_2 = 2 - 1 = 1$$
$$2x_1 + x_2 + x_3 = 4 \Rightarrow 2x_1 = 4 - 1 - 1 = 2 \Rightarrow x_1 = 1$$

$$\therefore \text{Solution: } (x_1, x_2, x_3) = (1, 1, 1)$$

Edge Cases and Constraints in LU Decomposition:

- **Standard LU Form:** The basic LU decomposition assumes row operations of the form $R_i' = R_i - m_{ij}R_j$, where $m_{ij} = \frac{a_{ij}}{a_{jj}}$ is the multiplier. This requires no row scaling before elimination.
- **When Row Scaling is Needed:** If you need to perform $R_2' = 2R_2 - 3R_1$, this indicates you're scaling row 2 by a factor of 2 before elimination. To convert this to standard LU form, we factor out the scaling into a diagonal matrix D, giving $A = LDU$.

  Example: Solve $Ax = b$ where $A = \begin{bmatrix} 2 & 4 \\ 3 & 7 \end{bmatrix}$ and $b = \begin{bmatrix} 6 \\ 13 \end{bmatrix}$, using $R_2' = 2R_2 - 3R_1$:

$$R_2' = 2\begin{bmatrix} 3 & 7 \end{bmatrix} - 3\begin{bmatrix} 2 & 4 \end{bmatrix} = \begin{bmatrix} 0 & 2 \end{bmatrix}$$

$$\text{Matrix after operation: } \begin{bmatrix} 2 & 4 \\ 0 & 2 \end{bmatrix}$$

Since $R_2' = 2R_2 - 3R_1$, factor out 2:

$$R_2' = 2(R_2 - \frac{3}{2}R_1)$$

This means row 2 was scaled by 2.

To extract this scaling:

$$L = \begin{bmatrix} 1 & 0 \\ \frac{3}{2} & 1 \end{bmatrix} \text{ (multiplier is } \frac{3}{2})$$

$$D = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \text{ (scaling factor from } 2R_2)$$

$$U = \begin{bmatrix} 2 & 4 \\ 0 & 1 \end{bmatrix} \text{ (normalized row 2: } \frac{R_2'}{2})$$

$$\text{Verification: } A = LDU$$
$$= \begin{bmatrix} 1 & 0 \\ \frac{3}{2} & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 2 & 4 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 \\ \frac{3}{2} & 1 \end{bmatrix} \begin{bmatrix} 2 & 4 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 3 & 7 \end{bmatrix} = A$$

To solve $Ax = b$ using $LDUx = b$:

Step 1: Solve $Ly = b$ (forward substitution)

$$\begin{bmatrix} 1 & 0 \\ \frac{3}{2} & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 13 \end{bmatrix}$$

$$y_1 = 6$$
$$\frac{3}{2}y_1 + y_2 = 13 \Rightarrow y_2 = 13 - 9 = 4$$

Step 2: Solve $DUx = y$ (let $z = Ux$, solve $Dz = y$)

$$\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 4 \end{bmatrix}$$

$$z_1 = 6$$
$$2z_2 = 4 \Rightarrow z_2 = 2$$

Step 3: Solve $Ux = z$ (back-substitution)

$$\begin{bmatrix} 2 & 4 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 2 \end{bmatrix}$$

$$x_2 = 2$$
$$2x_1 + 4x_2 = 6 \Rightarrow x_1 = \frac{6 - 8}{2} = -1$$

$$\therefore \text{Solution: } (x_1, x_2) = (-1, 2)$$

- **Zero Pivot Problem:** If a diagonal element (pivot) is zero, standard LU decomposition fails. Solutions:
  - **Row Swapping (Partial Pivoting):** Leads to PLU decomposition where $PA = LU$ and P is a permutation matrix. Example: Decompose $A = \begin{bmatrix} 0 & 1 & 2 \\ 2 & 3 & 4 \\ 4 & 5 & 6 \end{bmatrix}$ (note: $a_{11} = 0$)

Cannot proceed with standard LU since pivot $a_{11} = 0$

Swap $R_1 \leftrightarrow R_2$ to get largest pivot:

$$PA = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 2 \\ 2 & 3 & 4 \\ 4 & 5 & 6 \end{bmatrix} = \begin{bmatrix} 2 & 3 & 4 \\ 0 & 1 & 2 \\ 4 & 5 & 6 \end{bmatrix}$$

Now perform LU on PA :

$$m_{31} = \frac{4}{2} = 2, \quad R_3' = R_3 - 2R_1$$

$$\begin{bmatrix} 2 & 3 & 4 \\ 0 & 1 & 2 \\ 0 & -1 & -2 \end{bmatrix}$$

$$m_{32} = \frac{-1}{1} = -1, \quad R_3'' = R_3' - (-1)R_2 = R_3' + R_2$$

$$U = \begin{bmatrix} 2 & 3 & 4 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{bmatrix}$$

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

$$\therefore PA = LU \text{ (PLU decomposition)}$$

- **Singular Matrix:** If $\det(A) = 0$, the matrix has no unique LU decomposition (or the decomposition exists but U will have a zero on the diagonal, making the system unsolvable or having infinitely many solutions).

  Example: $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$ (note: row 2 is 2 times row 1)

  $$\det(A) = 1 \times 4 - 2 \times 2 = 0$$

  Attempt LU:

  $$m_{21} = \frac{2}{1} = 2, \quad R_2' = R_2 - 2R_1$$

  $$U = \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}, \quad L = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}$$

  The zero in U_{22} indicates the matrix is singular.

  If solving $Ax = b$:

  - If b is consistent with the dependency, infinitely many solutions
  - If b is inconsistent, no solution

- **Non-Square Matrices:** LU decomposition can be extended to rectangular matrices ($m \times n$), but L will be $m \times m$ and U will be $m \times n$ (or vice versa depending on the variant).

Example: Decompose $A = \begin{bmatrix} 2 & 1 & 3 \\ 4 & 3 & 5 \end{bmatrix}$ (2×3 matrix)

$$m_{21} = \frac{4}{2} = 2, \quad R_2' = R_2 - 2R_1$$

$$U = \begin{bmatrix} 2 & 1 & 3 \\ 0 & 1 & -1 \end{bmatrix}_{2\times 3}, \quad L = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}_{2\times 2}$$

$$\text{Verification: } LU = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 3 \\ 0 & 1 & -1 \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 1 & 3 \\ 4 & 3 & 5 \end{bmatrix} = A$$

- **Best Practice:** For numerical stability, always use **partial pivoting** (swap rows to get the largest pivot), resulting in PLU decomposition. This minimizes rounding errors in computer implementations.

Calculus:

- **Differential Calculus:** The branch of calculus that deals with the study of rates of change and slopes of curves. It focuses on finding derivatives, which measure how a function changes as its input changes. The derivative of a function $f(x)$ at a point x is defined as:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

This represents the instantaneous rate of change of f with respect to x.

Common Differentiation Rules:

1. **Constant Rule:** If $f(x) = c$ (where c is a constant), then $f'(x) = 0$.
   Example: If $f(x) = 7$, then $f'(x) = 0$.
2. **Power Rule:** If $f(x) = x^n$, then $f'(x) = nx^{n-1}$.
   Example:
$$f(x) = x^5$$
$$f'(x) = 5x^{5-1} = 5x^4$$
3. **Constant Multiple Rule:** If $f(x) = c \cdot g(x)$, then $f'(x) = c \cdot g'(x)$.
   Example:
$$f(x) = 3x^2$$
$$f'(x) = 3 \cdot 2x = 6x$$
4. **Sum/Difference Rule:** If $f(x) = g(x) \pm h(x)$, then $f'(x) = g'(x) \pm h'(x)$.
   Example:
$$f(x) = x^3 + 2x^2 - 5x + 7$$
$$f'(x) = 3x^2 + 4x - 5$$
5. **Product Rule:** If $f(x) = g(x) \cdot h(x)$, then $f'(x) = g'(x)h(x) + g(x)h'(x)$.

Example:

$$\begin{aligned} f(x) &= (x^2+1)(x^3-2) \\ g(x) &= x^2+1, \quad g'(x) = 2x \\ h(x) &= x^3-2, \quad h'(x) = 3x^2 \\ f'(x) &= (2x)(x^3-2) + (x^2+1)(3x^2) \\ &= 2x^4 - 4x + 3x^4 + 3x^2 \\ &= 5x^4 + 3x^2 - 4x \end{aligned}$$

6. **Quotient Rule:** If $f(x) = \dfrac{g(x)}{h(x)}$, then $f'(x) = \dfrac{g'(x)h(x) - g(x)h'(x)}{[h(x)]^2}$.
   Example:

$$\begin{aligned} f(x) &= \frac{x^2+1}{x-3} \\ g(x) &= x^2+1, \quad g'(x) = 2x \\ h(x) &= x-3, \quad h'(x) = 1 \\ f'(x) &= \frac{(2x)(x-3) - (x^2+1)(1)}{(x-3)^2} \\ &= \frac{2x^2 - 6x - x^2 - 1}{(x-3)^2} \\ &= \frac{x^2 - 6x - 1}{(x-3)^2} \end{aligned}$$

7. **Chain Rule:** If $f(x) = g(h(x))$, then $f'(x) = g'(h(x)) \cdot h'(x)$.
   Example:

$$\begin{aligned} f(x) &= (x^2+3x)^5 \\ \text{Let } u &= x^2+3x, \text{ then } f(x) = u^5 \\ \frac{df}{du} &= 5u^4, \quad \frac{du}{dx} = 2x+3 \\ f'(x) &= \frac{df}{du} \cdot \frac{du}{dx} \\ &= 5(x^2+3x)^4 \cdot (2x+3) \end{aligned}$$

8. **Exponential Rule:** If $f(x) = e^{g(x)}$, then $f'(x) = e^{g(x)} \cdot g'(x)$.
   Example:

$$\begin{aligned} f(x) &= e^{3x^2+2x} \\ f'(x) &= e^{3x^2+2x} \cdot (6x+2) \\ &= (6x+2)e^{3x^2+2x} \end{aligned}$$

9. **Logarithmic Rule:** If $f(x) = \ln(g(x))$, then $f'(x) = \dfrac{g'(x)}{g(x)}$.
   Example:

$$\begin{aligned} f(x) &= \ln\left(x^2+1\right) \\ f'(x) &= \frac{2x}{x^2+1} \end{aligned}$$

10. **Trigonometric Functions:**
    - $\dfrac{d}{dx}\sin(x) = \cos(x)$

- $\frac{d}{dx}\cos(x) = -\sin(x)$
- $\frac{d}{dx}\tan(x) = \sec^2 x$

Example:

$$f(x) = \sin\left(2x^2\right)$$
$$f'(x) = \cos\left(2x^2\right) \cdot 4x = 4x\cos\left(2x^2\right)$$

11. **Implicit Differentiation:** A technique used to find $\frac{dy}{dx}$ when y cannot be easily expressed as an explicit function of x. Instead of solving for y first, we differentiate both sides of the equation with respect to x, treating y as an implicit function of x, and then solve for $\frac{dy}{dx}$.
Key Steps:
(a) Differentiate both sides of the equation with respect to x
(b) Apply the chain rule to terms involving y: $\frac{d}{dx}[f(y)] = f'(y) \cdot \frac{dy}{dx}$
(c) Collect all terms with $\frac{dy}{dx}$ on one side
(d) Factor out $\frac{dy}{dx}$ and solve

Example 1: Find $\frac{dy}{dx}$ for the circle $x^2 + y^2 = 25$.

$$x^2 + y^2 = 25$$

Differentiate both sides with respect to x:

$$\frac{d}{dx}(x^2 + y^2) = \frac{d}{dx}(25)$$
$$\frac{d}{dx}(x^2) + \frac{d}{dx}(y^2) = 0$$
$$2x + 2y\frac{dy}{dx} = 0 \quad \text{(chain rule on } y^2\text{)}$$

Solve for $\frac{dy}{dx}$:

$$2y\frac{dy}{dx} = -2x$$
$$\frac{dy}{dx} = -\frac{2x}{2y} = -\frac{x}{y}$$

Example 2: Find $\frac{dy}{dx}$ for $x^3 + y^3 = 6xy$.

$$x^3 + y^3 = 6xy$$

Differentiate both sides:

$$\frac{d}{dx}(x^3 + y^3) = \frac{d}{dx}(6xy)$$
$$3x^2 + 3y^2\frac{dy}{dx} = 6\left(x\frac{dy}{dx} + y\right) \quad \text{(product rule on RHS)}$$
$$3x^2 + 3y^2\frac{dy}{dx} = 6x\frac{dy}{dx} + 6y$$

Collect $\frac{dy}{dx}$ terms:

$$3y^2\frac{dy}{dx} - 6x\frac{dy}{dx} = 6y - 3x^2$$
$$(3y^2 - 6x)\frac{dy}{dx} = 6y - 3x^2$$
$$\frac{dy}{dx} = \frac{6y - 3x^2}{3y^2 - 6x} = \frac{3(2y - x^2)}{3(y^2 - 2x)}$$
$$\frac{dy}{dx} = \frac{2y - x^2}{y^2 - 2x}$$

Example 3: Find the slope of the tangent line to $\sin(xy) + x = y^2$ at point $(0, 0)$.

$$\sin(xy) + x = y^2$$

Differentiate:

$$\frac{d}{dx}[\sin(xy)] + \frac{d}{dx}(x) = \frac{d}{dx}(y^2)$$
$$\cos(xy) \cdot \frac{d}{dx}(xy) + 1 = 2y\frac{dy}{dx}$$
$$\cos(xy) \cdot \left(x\frac{dy}{dx} + y\right) + 1 = 2y\frac{dy}{dx}$$
$$x\cos(xy)\frac{dy}{dx} + y\cos(xy) + 1 = 2y\frac{dy}{dx}$$

Solve for $\frac{dy}{dx}$:

$$x\cos(xy)\frac{dy}{dx} - 2y\frac{dy}{dx} = -y\cos(xy) - 1$$
$$[x\cos(xy) - 2y]\frac{dy}{dx} = -y\cos(xy) - 1$$
$$\frac{dy}{dx} = \frac{-y\cos(xy) - 1}{x\cos(xy) - 2y}$$

At point $(0, 0)$:

$$\left.\frac{dy}{dx}\right|_{(0,0)} = \frac{-0 \cdot \cos(0) - 1}{0 \cdot \cos(0) - 2(0)}$$
$$= \frac{-1}{0} \quad \text{(undefined - vertical tangent)}$$

12. **Mixed Partial Derivatives:** For a function of multiple variables $f(x, y)$, mixed partial derivatives involve taking partial derivatives with respect to different variables in succession. For most functions (those with continuous second derivatives), the order of differentiation doesn't matter, known as **Clairaut's Theorem** or **Schwarz's Theorem**:

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}$$

Notation:

- $f_{xy} = \frac{\partial^2 f}{\partial y \partial x}$ (differentiate first with respect to x, then y)
- $f_{yx} = \frac{\partial^2 f}{\partial x \partial y}$ (differentiate first with respect to y, then x)

Example 1: Find all second-order partial derivatives of $f(x, y) = x^3y^2 + 2xy$.

First-order partial derivatives:

$$f_x = \frac{\partial f}{\partial x} = 3x^2y^2 + 2y$$
$$f_y = \frac{\partial f}{\partial y} = 2x^3y + 2x$$

Second-order partial derivatives:

$$f_{xx} = \frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}(3x^2y^2 + 2y) = 6xy^2$$
$$f_{yy} = \frac{\partial^2 f}{\partial y^2} = \frac{\partial}{\partial y}(2x^3y + 2x) = 2x^3$$

Mixed partial derivatives:

$$f_{xy} = \frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}(3x^2y^2 + 2y) = 6x^2y + 2$$
$$f_{yx} = \frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial x}(2x^3y + 2x) = 6x^2y + 2$$

Note: $f_{xy} = f_{yx}$ (Clairaut's Theorem verified)

Example 2: Find mixed partials for $f(x, y, z) = x^2yz + xe^{yz}$.

$$f_x = 2xyz + e^{yz}$$
$$f_y = x^2z + xze^{yz}$$
$$f_z = x^2y + xye^{yz}$$

$$f_{xy} = \frac{\partial}{\partial y}(2xyz + e^{yz}) = 2xz + ze^{yz}$$
$$f_{yx} = \frac{\partial}{\partial x}(x^2z + xze^{yz}) = 2xz + ze^{yz}$$

$$f_{xz} = \frac{\partial}{\partial z}(2xyz + e^{yz}) = 2xy + ye^{yz}$$
$$f_{zx} = \frac{\partial}{\partial x}(x^2y + xye^{yz}) = 2xy + ye^{yz}$$

$$f_{yz} = \frac{\partial}{\partial z}(x^2z + xze^{yz}) = x^2 + xze^{yz} + xye^{yz}$$
$$f_{zy} = \frac{\partial}{\partial y}(x^2y + xye^{yz}) = x^2 + xze^{yz} + xye^{yz}$$

Critical Points and Extrema: Critical points are points where the derivative is zero or undefined. These points are candidates for local maxima, local minima, or saddle points. **Definition:** A critical point of $f(x)$ occurs at $x = c$ if:

- $f'(c) = 0$ (horizontal tangent), or
- $f'(c)$ is undefined (vertical tangent or cusp)

Types of Extrema:

- **Local Maximum:** $f(c) \geq f(x)$ for all x near c
- **Local Minimum:** $f(c) \leq f(x)$ for all x near c
- **Global Maximum:** $f(c) \geq f(x)$ for all x in the domain
- **Global Minimum:** $f(c) \leq f(x)$ for all x in the domain

First Derivative Test:

- If $f'(x)$ changes from positive to negative at $x = c$, then f has a local maximum at c
- If $f'(x)$ changes from negative to positive at $x = c$, then f has a local minimum at c
- If $f'(x)$ does not change sign at $x = c$, then f has neither a max nor min at c

Second Derivative Test:

- If $f'(c) = 0$ and $f''(c) > 0$, then f has a local minimum at c
- If $f'(c) = 0$ and $f''(c) < 0$, then f has a local maximum at c
- If $f'(c) = 0$ and $f''(c) = 0$, the test is inconclusive (use first derivative test)

Example 1: Find and classify all critical points of $f(x) = x^3 - 6x^2 + 9x + 1$.

$$f(x) = x^3 - 6x^2 + 9x + 1$$
$$f'(x) = 3x^2 - 12x + 9$$

$$\text{Set } f'(x) = 0\text{:}$$
$$3x^2 - 12x + 9 = 0$$
$$3(x^2 - 4x + 3) = 0$$
$$3(x-1)(x-3) = 0$$
$$x = 1, 3 \quad \text{(critical points)}$$

$$f''(x) = 6x - 12$$

$$\text{At } x = 1:$$
$$f''(1) = 6(1) - 12 = -6 < 0 \quad \Rightarrow \text{ local maximum}$$
$$f(1) = 1 - 6 + 9 + 1 = 5$$

$$\text{At } x = 3:$$
$$f''(3) = 6(3) - 12 = 6 > 0 \quad \Rightarrow \text{ local minimum}$$
$$f(3) = 27 - 54 + 27 + 1 = 1$$

$$\therefore \text{Local max at } (1, 5), \text{ local min at } (3, 1)$$

Example 2: Find extrema of $f(x) = \frac{x^2}{x-1}$ on the interval $[2, 4]$.

$$f(x) = \frac{x^2}{x-1}$$

Using quotient rule:

$$\begin{aligned} f'(x) &= \frac{2x(x-1) - x^2(1)}{(x-1)^2} \\ &= \frac{2x^2 - 2x - x^2}{(x-1)^2} \\ &= \frac{x^2 - 2x}{(x-1)^2} \\ &= \frac{x(x-2)}{(x-1)^2} \end{aligned}$$

$$\begin{aligned} \text{Set } f'(x) &= 0\text{:} \\ x(x-2) &= 0 \\ x &= 0, 2 \end{aligned}$$

In $[2, 4]$, only $x = 2$ is a critical point

Evaluate at critical point and endpoints:

$$\begin{aligned} f(2) &= \frac{4}{1} = 4 \quad \text{(critical point)} \\ f(4) &= \frac{16}{3} \approx 5.33 \quad \text{(right endpoint)} \end{aligned}$$

$$\therefore \text{Absolute min at } (2, 4), \text{ absolute max at } (4, \frac{16}{3})$$

Concavity and Inflection Points: Concavity describes the direction a curve bends. The second derivative determines concavity.

Concavity:

- **Concave Up:** If $f''(x) > 0$ on an interval, the graph curves upward (like $\cup$)
- **Concave Down:** If $f''(x) < 0$ on an interval, the graph curves downward (like $\cap$)

Inflection Point: A point where the concavity changes (from up to down or vice versa). Occurs where $f''(x) = 0$ or $f''(x)$ is undefined, AND the concavity changes.

Steps to Find Inflection Points:

1. Find $f''(x)$
2. Solve $f''(x) = 0$ and identify where $f''(x)$ is undefined
3. Test intervals around these points to verify concavity change
4. If concavity changes, it's an inflection point

Example 1: Find intervals of concavity and inflection points for $f(x) = x^3 - 6x^2 + 9x + 1$.

$$\begin{aligned} f(x) &= x^3 - 6x^2 + 9x + 1 \\ f'(x) &= 3x^2 - 12x + 9 \\ f''(x) &= 6x - 12 \end{aligned}$$

Set $f''(x) = 0$:

$$\begin{aligned} 6x - 12 &= 0 \\ x &= 2 \quad \text{(potential inflection point)} \end{aligned}$$

Test intervals:

$$\begin{aligned} \text{For } x < 2: &\quad f''(1) = 6 - 12 = -6 < 0 \quad \text{(concave down)} \\ \text{For } x > 2: &\quad f''(3) = 18 - 12 = 6 > 0 \quad \text{(concave up)} \end{aligned}$$

At $x = 2$:

$$f(2) = 8 - 24 + 18 + 1 = 3$$

$\therefore$ Inflection point at $(2, 3)$
Concave down on $(-\infty, 2)$
Concave up on $(2, \infty)$

Example 2: Analyze $f(x) = x^4 - 4x^3$ for concavity and inflection points.

$$\begin{aligned} f(x) &= x^4 - 4x^3 \\ f'(x) &= 4x^3 - 12x^2 \\ f''(x) &= 12x^2 - 24x \\ &= 12x(x - 2) \end{aligned}$$

Set $f''(x) = 0$:

$$\begin{aligned} 12x(x - 2) &= 0 \\ x &= 0, 2 \end{aligned}$$

Test intervals:

$$\begin{aligned} x < 0: &\quad f''(-1) = 12 + 24 = 36 > 0 \quad \text{(up)} \\ 0 < x < 2: &\quad f''(1) = 12 - 24 = -12 < 0 \quad \text{(down)} \\ x > 2: &\quad f''(3) = 108 - 72 = 36 > 0 \quad \text{(up)} \end{aligned}$$

$$\begin{aligned} f(0) &= 0 \\ f(2) &= 16 - 32 = -16 \end{aligned}$$

$\therefore$ Inflection points at $(0, 0)$ and $(2, -16)$
Concave up: $(-\infty, 0) \cup (2, \infty)$
Concave down: $(0, 2)$

Example 3: Complete analysis of $f(x) = xe^{-x}$.

$$f(x) = xe^{-x}$$

First derivative (product rule):

$$\begin{aligned} f'(x) &= (1)e^{-x} + x(-e^{-x}) \\ &= e^{-x}(1-x) \end{aligned}$$

Critical points: $f'(x) = 0$

$$\begin{aligned} e^{-x}(1-x) &= 0 \\ x &= 1 \quad (e^{-x} \neq 0) \end{aligned}$$

Second derivative:

$$\begin{aligned} f''(x) &= -e^{-x}(1-x) + e^{-x}(-1) \\ &= e^{-x}[-(1-x)-1] \\ &= e^{-x}(x-2) \end{aligned}$$

At $x = 1$:

$$\begin{aligned} f''(1) &= e^{-1}(-1) < 0 \quad \Rightarrow \text{ local max} \\ f(1) &= e^{-1} \approx 0.368 \end{aligned}$$

Inflection: $f''(x) = 0$

$$\begin{aligned} e^{-x}(x-2) &= 0 \\ x &= 2 \\ f(2) &= 2e^{-2} \approx 0.271 \end{aligned}$$

Concavity test:

$$\begin{aligned} x < 2: &\quad f''(1) < 0 \quad \text{(concave down)} \\ x > 2: &\quad f''(3) = e^{-3}(1) > 0 \quad \text{(concave up)} \end{aligned}$$

$\therefore$ Local max at $(1, e^{-1})$
Inflection point at $(2, 2e^{-2})$
Concave down on $(-\infty, 2)$
Concave up on $(2, \infty)$

Regression Analysis: A supervised learning technique that models the relationship between one or more independent variables (X) and a dependent variable (Y). The goal is to find a function that best predicts continuous values based on input data by minimizing the prediction error.

Simple Linear Regression: Models the relationship using a straight line: $y = mx + c$, where:

- y = dependent variable (output/prediction)
- x = independent variable (input/feature)
- m = slope (rate of change of y with respect to x)
- c = y-intercept (value of y when $x = 0$)

Example: Predict house prices based on area.

House Area (sq. ft)	Price (taka)
1000	200000
1500	250000
2000	300000
2200	?

Methods to Calculate Slope and Intercept:

1. **Ordinary Least Squares (OLS):** Minimizes the sum of squared residuals (vertical distances between actual and predicted values). This is the most common analytical solution.

   Formulas:

$$m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$c = \bar{y} - m\bar{x}$$

   where $\bar{x} = \frac{1}{n}\sum x_i$ and $\bar{y} = \frac{1}{n}\sum y_i$ are the means.

   Calculation for the example:

$$n = 3, \quad \bar{x} = \frac{1000 + 1500 + 2000}{3} = 1500$$
$$\bar{y} = \frac{200000 + 250000 + 300000}{3} = 250000$$

$$\begin{aligned} m &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \\ &= \frac{(1000 - 1500)(200000 - 250000) + (1500 - 1500)(250000 - 250000)}{(1000 - 1500)^2 + (1500 - 1500)^2 + (2000 - 1500)^2} \\ &\quad + \frac{(2000 - 1500)(300000 - 250000)}{(1000 - 1500)^2 + (1500 - 1500)^2 + (2000 - 1500)^2} \\ &= \frac{(-500)(-50000) + 0 + (500)(50000)}{250000 + 0 + 250000} \\ &= \frac{25000000 + 25000000}{500000} = \frac{50000000}{500000} = 100 \end{aligned}$$

$$\begin{aligned} c &= \bar{y} - m\bar{x} \\ &= 250000 - 100(1500) = 250000 - 150000 = 100000 \end{aligned}$$

$$\therefore y = 100x + 100000$$

$$\begin{aligned} \text{For } x = 2200: \; y &= 100(2200) + 100000 \\ &= 220000 + 100000 = 320000 \end{aligned}$$

2. **Normal Equation (Matrix Form):** For multiple linear regression with n samples and p features, we represent the problem as $y = X\beta + \epsilon$, where X is the design matrix, β is the parameter vector, and ϵ is the error term.

   Solution:

$$\beta = (X^T X)^{-1} X^T y$$

For simple linear regression with intercept:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \beta = \begin{bmatrix} c \\ m \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Example calculation:

$$X = \begin{bmatrix} 1 & 1000 \\ 1 & 1500 \\ 1 & 2000 \end{bmatrix}, \quad y = \begin{bmatrix} 200000 \\ 250000 \\ 300000 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 \\ 1000 & 1500 & 2000 \end{bmatrix} \begin{bmatrix} 1 & 1000 \\ 1 & 1500 \\ 1 & 2000 \end{bmatrix} = \begin{bmatrix} 3 & 4500 \\ 4500 & 7250000 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 1 & 1 & 1 \\ 1000 & 1500 & 2000 \end{bmatrix} \begin{bmatrix} 200000 \\ 250000 \\ 300000 \end{bmatrix} = \begin{bmatrix} 750000 \\ 1175000000 \end{bmatrix}$$

$$\begin{aligned} (X^T X)^{-1} &= \frac{1}{3(7250000) - 4500^2} \begin{bmatrix} 7250000 & -4500 \\ -4500 & 3 \end{bmatrix} \\ &= \frac{1}{1500000} \begin{bmatrix} 7250000 & -4500 \\ -4500 & 3 \end{bmatrix} \\ &= \begin{bmatrix} \frac{29}{6} & \frac{-3}{1000} \\ \frac{-3}{1000} & \frac{1}{500000} \end{bmatrix} \end{aligned}$$

$$\begin{aligned} (X^T X)^{-1} X^T y &= \begin{bmatrix} \frac{29}{6} & \frac{-3}{1000} \\ \frac{-3}{1000} & \frac{1}{500000} \end{bmatrix} \begin{bmatrix} 750000 \\ 1175000000 \end{bmatrix} \\ &= \begin{bmatrix} 3625000 - 3525000 \\ -2250 + 2350 \end{bmatrix} \\ &= \begin{bmatrix} 100000 \\ 100 \end{bmatrix} = \beta \end{aligned}$$

3. **Gradient Descent:** An iterative optimization algorithm that minimizes the cost function (mean squared error) by updating parameters in the direction of steepest descent.

   What is a Gradient?

   Gradient is a vector of partial derivatives with respect to all the input parameters of a function, where each partial derivative represents the rate of change of the function with respect to that variable considering all the other parameters being stationary.

   Example:

   Say,

   $$f(x, y) = 2x + 3y$$

   $$\text{Gradient, } \nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}$$

$$\frac{\partial f}{\partial x} = 2$$
$$\frac{\partial f}{\partial y} = 3$$

The gradient above can be interpreted as:

- $\frac{\partial f}{\partial x}$: If x increases by 1 unit the output of the function increases by 2 (holding y constant). For example, $f(1,2) = 2 + 6 = 8$ and $f(2,2) = 4 + 6 = 10$. Here, x increased by 1 unit (1 to 2) and the output increased by 2 ($10 - 8 = 2$).
- $\frac{\partial f}{\partial y}$: If y increases by 1 unit the output of the function increases by 3 (holding x constant).
- $\begin{bmatrix} 2 \\ 3 \end{bmatrix}$ is the direction of the steepest ascent of this function.
- The maximum rate of change of this function is $||\nabla f|| = \sqrt{2^2 + 3^2} = \sqrt{13}$.

Gradient Descent intuition:

(a) Start with random parameter values.

(b) Calculate the gradient (the maximum ascent direction).

(c) Move a small step in the opposite direction (maximum descent).

(d) Repeat until we reach the minimum.

The learning rate α controls the step size - too large and we might overshoot, too small and convergence is slow.

Gradient Descent Variants: Based on the sample size to compute gradient, gradient descent can be of three types.

- **Batch Gradient Descent:** It is the vanilla gradient descent. Here every epoch the whole training dataset is used to compute gradient and update the weights. Batch gradient descent is computationally very expensive.
- **Mini-Batch Gradient Descent:** Instead of using the whole dataset at once, it is divided into multiple parts of same size (usually 16, 32, 64, 128, ...) and weights are updated. Each epoch has multiple steps which depends on the batch size.
- **Stochastic Gradient Descent (SGD):** Unlike previous methods, here only one sample is used to update weights at a time. So, each epoch has n steps where n is the total number of samples.

In case of **Mini-Batch Gradient Descent** and **SGD**, the training dataset is randomly shuffled at the beginning of each epoch.

Cost Function:

$$J(m, c) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - (mx_i + c))^2$$

Note: The $\frac{1}{2}$ factor is included for mathematical convenience. When we take the derivative of the squared term, we get a factor of 2 from the chain rule, which cancels with the $\frac{1}{2}$. This simplifies the gradient formulas. The standard MSE uses $\frac{1}{n}$, but for gradient descent, $\frac{1}{2n}$ is conventional because:

- It doesn't change the location of the minimum (just scales the function)
- The derivative of $(y - \hat{y})^2$ gives $2(y - \hat{y})$, and $2 \times \frac{1}{2} = 1$, eliminating the coefficient

- Results in cleaner gradient expressions

Update Rules:

$$m := m - \alpha \frac{\partial J}{\partial m} = m - \alpha \frac{1}{n} \sum_{i=1}^{n} (mx_i + c - y_i)x_i$$

$$c := c - \alpha \frac{\partial J}{\partial c} = c - \alpha \frac{1}{n} \sum_{i=1}^{n} (mx_i + c - y_i)$$

where α is the learning rate (e.g., 0.01).

Derivation of gradients:

$$\begin{aligned} \frac{\partial J}{\partial m} &= \frac{\partial}{\partial m} \left[\frac{1}{2n} \sum_{i=1}^{n} (y_i - mx_i - c)^2 \right] \\ &= \frac{1}{2n} \sum_{i=1}^{n} 2(y_i - mx_i - c)(-x_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} (mx_i + c - y_i)x_i \end{aligned}$$

The factor of 2 from differentiation cancels with the $\frac{1}{2}$ in the cost function.

Algorithm:

(a) Initialize $m = 0, c = 0$

(b) Repeat until convergence:

- Compute predictions: $\hat{y}_i = mx_i + c$
- Compute gradients using all training samples
- Update m and c simultaneously

(c) Stop when $|J^{(t)} - J^{(t-1)}| < \epsilon$ (e.g., $\epsilon = 10^{-6}$)

Advantages: Works well with large datasets, can handle online learning.

Disadvantages: Requires choosing learning rate, may converge slowly.

4. **Regularized Methods:** Add penalty terms to prevent overfitting in complex models.

Ridge Regression (L2):

$$J(m, c) = \sum_{i=1}^{n} (y_i - (mx_i + c))^2 + \lambda m^2$$

Solution: $\beta = (X^T X + \lambda I)^{-1} X^T y$

Lasso Regression (L1):

$$J(m, c) = \sum_{i=1}^{n} (y_i - (mx_i + c))^2 + \lambda |m|$$

where $\lambda > 0$ is the regularization parameter. Larger λ forces slope toward zero, preventing overfitting.

Comparison of Methods:

- **OLS/Normal Equation:** Best for small to medium datasets, provides exact solution, requires $(X^T X)^{-1}$ to exist

- **Gradient Descent:** Best for large datasets or when matrix inversion is expensive, flexible and scalable
- **Regularized Methods:** Best when dealing with multicollinearity or to prevent overfitting in high-dimensional data

Loss Functions and Cost Functions: These quantify how well a model's predictions match the actual values. Understanding them is crucial for training and evaluating regression models.

Terminology:

- **Loss Function:** Measures the error for a *single* data point: $L(y_i, \hat{y}_i)$
- **Cost Function:** Aggregates the loss over the *entire* dataset: $J = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{y}_i)$

Common Loss/Cost Functions for Regression:

1. **Mean Squared Error (MSE):** Most commonly used for linear regression. Penalizes larger errors more heavily due to squaring.

   Loss: $L(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$

   Cost:
   $$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

   Properties:

   - Differentiable everywhere (smooth optimization)
   - Sensitive to outliers (large errors dominate)
   - Units are squared (e.g., price^2)
   - Convex function (guarantees global minimum)

   Example: For predictions $\hat{y} = [100, 150, 200]$ and actual $y = [110, 145, 205]$:
   $$\begin{aligned}\text{MSE} &= \frac{1}{3}[(110-100)^2 + (145-150)^2 + (205-200)^2] \\ &= \frac{1}{3}[100 + 25 + 25] = \frac{150}{3} = 50\end{aligned}$$

2. **Root Mean Squared Error (RMSE):** Square root of MSE, brings units back to original scale.

   Formula:
   $$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

   Properties:

   - Same units as target variable (easier to interpret)
   - Still sensitive to outliers
   - Commonly reported metric

   Example (using previous data): RMSE $= \sqrt{50} \approx 7.07$

3. **Mean Absolute Error (MAE):** Uses absolute differences instead of squaring.

   Loss: $L(y_i, \hat{y}_i) = |y_i - \hat{y}_i|$

   Cost:
   $$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

   Properties:

- More robust to outliers than MSE
- Same units as target variable
- Not differentiable at zero (optimization challenges)
- All errors weighted equally

Example: $\text{MAE} = \frac{1}{3}[|10| + |-5| + |5|] = \frac{20}{3} \approx 6.67$

4. **Huber Loss:** Combines MSE and MAE - quadratic for small errors, linear for large errors.

   Loss:

$$L_\delta(y_i, \hat{y}_i) = \begin{cases} \frac{1}{2}(y_i - \hat{y}_i)^2 & \text{if } |y_i - \hat{y}_i| \leq \delta \\ \delta|y_i - \hat{y}_i| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases}$$

   where δ is a threshold parameter (typically 1.0).

   Properties:

   - Less sensitive to outliers than MSE
   - Differentiable everywhere (unlike MAE)
   - Requires tuning δ parameter
   - Good balance between MSE and MAE

5. **Log-Cosh Loss:** Logarithm of the hyperbolic cosine of prediction error.

   Loss: $L(y_i, \hat{y}_i) = \log\left(\cosh\left(y_i - \hat{y}_i\right)\right)$

   Cost:

$$J = \frac{1}{n}\sum_{i=1}^{n} \log\left(\cosh\left(y_i - \hat{y}_i\right)\right)$$

   Properties:

   - Approximately equal to MSE for small errors
   - Approximately linear (like MAE) for large errors
   - Differentiable everywhere
   - Smooth and convex

6. **R^2 Score (Coefficient of Determination):** Not a loss function, but a common evaluation metric. Measures proportion of variance explained by the model.

   Formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

   where RSS = Residual Sum of Squares, TSS = Total Sum of Squares.

   Properties:

   - Range: $(-\infty, 1]$, where 1 is perfect fit
   - $R^2 = 0$ means model performs as well as mean baseline
   - Negative R^2 means model performs worse than mean
   - Higher is better (unlike loss functions)

Choosing a Loss Function:

- **MSE/RMSE:** Use when large errors are particularly undesirable and data has few outliers
- **MAE:** Use when outliers are present or all errors should be weighted equally
- **Huber/Log-Cosh:** Use for robust regression with moderate outliers
- **Consider domain:** Some fields prefer specific metrics (e.g., MAE for demand forecasting)

Correlation in Regression: Correlation measures the strength and direction of the linear relationship between two variables. The correlation coefficient, denoted by r (Pearson's correlation coefficient), quantifies this relationship and ranges from -1 to $+1$.

Interpretation of Correlation Coefficient (r) Ranges:

- $r = +1$**:** Perfect positive linear correlation. All points lie exactly on a line with positive slope. As x increases, y increases proportionally.
- $+0.7 < r < +1$**:** Strong positive correlation. Points cluster closely around an upward-sloping line. Strong tendency for y to increase as x increases.
- $+0.3 < r \leq +0.7$**:** Moderate positive correlation. Points show a general upward trend but with noticeable scatter. Moderate tendency for y to increase with x.
- $0 < r \leq +0.3$**:** Weak positive correlation. Points show slight upward trend with substantial scatter. Weak relationship between variables.
- $r = 0$**:** No linear correlation. No linear relationship between variables (though non-linear relationships may exist).
- $-0.3 < r < 0$**:** Weak negative correlation. Points show slight downward trend with substantial scatter.
- $-0.7 < r \leq -0.3$**:** Moderate negative correlation. Points show a general downward trend but with noticeable scatter.
- $-1 < r \leq -0.7$**:** Strong negative correlation. Points cluster closely around a downward-sloping line. Strong tendency for y to decrease as x increases.
- $r = -1$**:** Perfect negative linear correlation. All points lie exactly on a line with negative slope. As x increases, y decreases proportionally.

Types of Correlation:

1. **Pearson Correlation Coefficient (r):** Measures the *linear* relationship between two continuous variables. Assumes normal distribution and is sensitive to outliers.

   Formula:

$$r = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

   where $\text{Cov}(X,Y)$ is the covariance, σ_X and σ_Y are standard deviations.

   Properties:

   - Range: $[-1, +1]$
   - Measures only linear relationships
   - Sensitive to outliers
   - Assumes continuous, normally distributed data

   Example: Calculate Pearson's r for study hours (x) and exam scores (y):

Hours (x)	Score (y)
2	65
4	75
6	85
8	95

$$n = 4, \quad \bar{x} = \frac{2+4+6+8}{4} = 5, \quad \bar{y} = \frac{65+75+85+95}{4} = 80$$

$$\begin{aligned}\sum (x_i - \bar{x})(y_i - \bar{y}) &= (2-5)(65-80) + (4-5)(75-80) \\ &\quad + (6-5)(85-80) + (8-5)(95-80) \\ &= (-3)(-15) + (-1)(-5) + (1)(5) + (3)(15) \\ &= 45 + 5 + 5 + 45 = 100\end{aligned}$$

$$\sum (x_i - \bar{x})^2 = (-3)^2 + (-1)^2 + (1)^2 + (3)^2 = 9 + 1 + 1 + 9 = 20$$

$$\begin{aligned}\sum (y_i - \bar{y})^2 &= (-15)^2 + (-5)^2 + (5)^2 + (15)^2 \\ &= 225 + 25 + 25 + 225 = 500\end{aligned}$$

$$r = \frac{100}{\sqrt{20}\sqrt{500}} = \frac{100}{\sqrt{10000}} = \frac{100}{100} = 1$$

Perfect positive correlation: more study hours perfectly predict higher scores.

2. **Spearman's Rank Correlation (ρ or r_s):** Measures the *monotonic* relationship (not necessarily linear) between two variables by using ranks instead of raw values. More robust to outliers than Pearson's r.

   Formula:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

   where d_i is the difference between ranks of corresponding values.

   Properties:

   - Range: $[-1, +1]$
   - Measures monotonic relationships (including non-linear)
   - Robust to outliers
   - Works with ordinal data
   - No assumption of normality

   Example: Calculate Spearman's ρ for the data:

x	y	Rank(x)	Rank(y)	d = Rank(x) - Rank(y)
10	35	1	1	0
20	40	2	2	0
30	50	3	3	0
40	45	4	4	0

$$\sum d_i^2 = 0^2 + 0^2 + 0^2 + 0^2 = 0$$

$$\rho = 1 - \frac{6(0)}{4(16-1)} = 1 - 0 = 1$$

3. **Kendall's Tau (τ):** Another rank-based correlation measure that counts concordant and discordant pairs. Better for small sample sizes.

   Formula:
   $$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{n(n-1)}{2}}$$

   or equivalently:
   $$\tau = \frac{C - D}{C + D}$$

   where C = concordant pairs, D = discordant pairs.

   Properties:

   - Range: $[-1, +1]$
   - More conservative than Spearman (typically lower values)
   - Better for small sample sizes
   - Interpretation in terms of probability of concordance

4. **Point-Biserial Correlation (r_{pb}):** Measures the relationship between a continuous variable and a binary (dichotomous) variable.

   Formula:
   $$r_{pb} = \frac{\bar{x}_1 - \bar{x}_0}{s_x} \sqrt{\frac{n_1 n_0}{n(n-1)}}$$

   where $\bar{x}_1$ and $\bar{x}_0$ are means for groups 1 and 0, s_x is the standard deviation of the continuous variable, n_1 and n_0 are group sizes.

   Use case: Correlating exam scores (continuous) with pass/fail status (binary).

5. **Partial Correlation:** Measures the relationship between two variables while controlling for the effect of one or more other variables.

   Formula (controlling for variable z):
   $$r_{xy \cdot z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

   where r_{xy} is correlation between x and y, r_{xz} is correlation between x and z, etc.

   Use case: Finding correlation between ice cream sales and drowning incidents while controlling for temperature.

6. **Multiple Correlation (R):** Measures the strength of relationship between one dependent variable and multiple independent variables combined.

   Formula (for two predictors):
   $$R_{y \cdot x_1 x_2} = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2 r_{yx_1} r_{yx_2} r_{x_1 x_2}}{1 - r_{x_1 x_2}^2}}$$

   Properties:

   - Range: $[0, +1]$ (always positive)
   - Used in multiple regression analysis
   - R^2 is the coefficient of determination

Important Notes:

- **Correlation $\neq$ Causation:** A strong correlation does not imply that one variable causes changes in the other.
- **Outliers:** Can significantly affect Pearson's r; use Spearman's ρ for robustness.
- **Non-linear Relationships:** Pearson's r may be near zero even with strong non-linear relationships.
- **Sample Size:** Small samples may show strong correlations by chance; use significance testing.

Improvements on Gradient Descent:

- **Gradient Descent with Momentum:** In case of Mini-Batch Gradient Descent and SGD, the frequent and small weight updates creates noise as the true maximum descent direction cannot be interpreted from a small size of sample. Also, there could be such dimensions where gradient is higher than others. This causes oscillation (change of direction back-n-forth). To counter both of the cases, a momentum term is added with current gradient. It acts like a memory of past gradients causing bigger steps where the direction is consistent and cancels out/reduces oscillation.

$$v_{t+1} = \beta v_t + g_{t+1}$$
$$w_{t+1} = w_t - \alpha v_{t+1}$$

  Here,

  - v is the velocity which starts at $\begin{bmatrix}0 & 0 & \dots & 0\end{bmatrix}^T_{1\times p}$ (p is the number of parameters).
  - β is the momentum coefficient. $0 \leq \beta < 1$ (usually 0.9, 0.99 etc.)
  - g is the gradient.
  - w is the weights.
  - α is the learning rate.

  This algorithm, prioritizes newer gradients over older ones. Older gradients decays exponentially ($v_{t+1} = g_{t+1} + \beta g_t + \beta^2 g_{t-1} + \beta^3 g_{t-2} + \cdots + \beta^t g_1$).

- RMSProp:

Extras:

- **Logarithmic Properties:**
  - **Product Rule:** $\log_a bc = \log_a b \times \log_a c$.
  - **Quotient Rule:** $\log_a \frac{b}{c} = \log_a b - \log_a c$.
  - **Power Rule:** $\log_a b^c = c \log_a b$.
  - **Change of Base:** $\log_a b = \frac{\log_x b}{\log_x a}$.
  - **Logarithm of 1:** $\log_a 1 = 0$.
  - **Logarithm of Base:** $\log_a a = 1$.
- **Properties of e and Natural Logarithms $\ln x$:**
  - The constant $e \approx 2.718$ is the base of natural logarithm. $\ln x = \ln_e x$.
  - $\ln e = 1$.
  - $e^{\ln x} = x$ and $\ln e^x = x$.
  - **Derivative and Integral of e^x:** $\frac{d}{dx}e^x = e^x$ and $\int e^x dx = e^x + C$.
  - **Derivative and Integral of $\ln x$:** $\frac{d}{dx}\ln x = \frac{1}{x}$ and $\int \ln x dx = x \ln x - x + C$.