

Lecture 1

Basics of Linear Algebra

- Matrix Operations
- Matrix Derivative and Expectations
- Applications and Wrap-Up

Contents

- Matrix Operations
- Matrix Derivative and Expectations
- Applications and Wrap-Up

Motivation

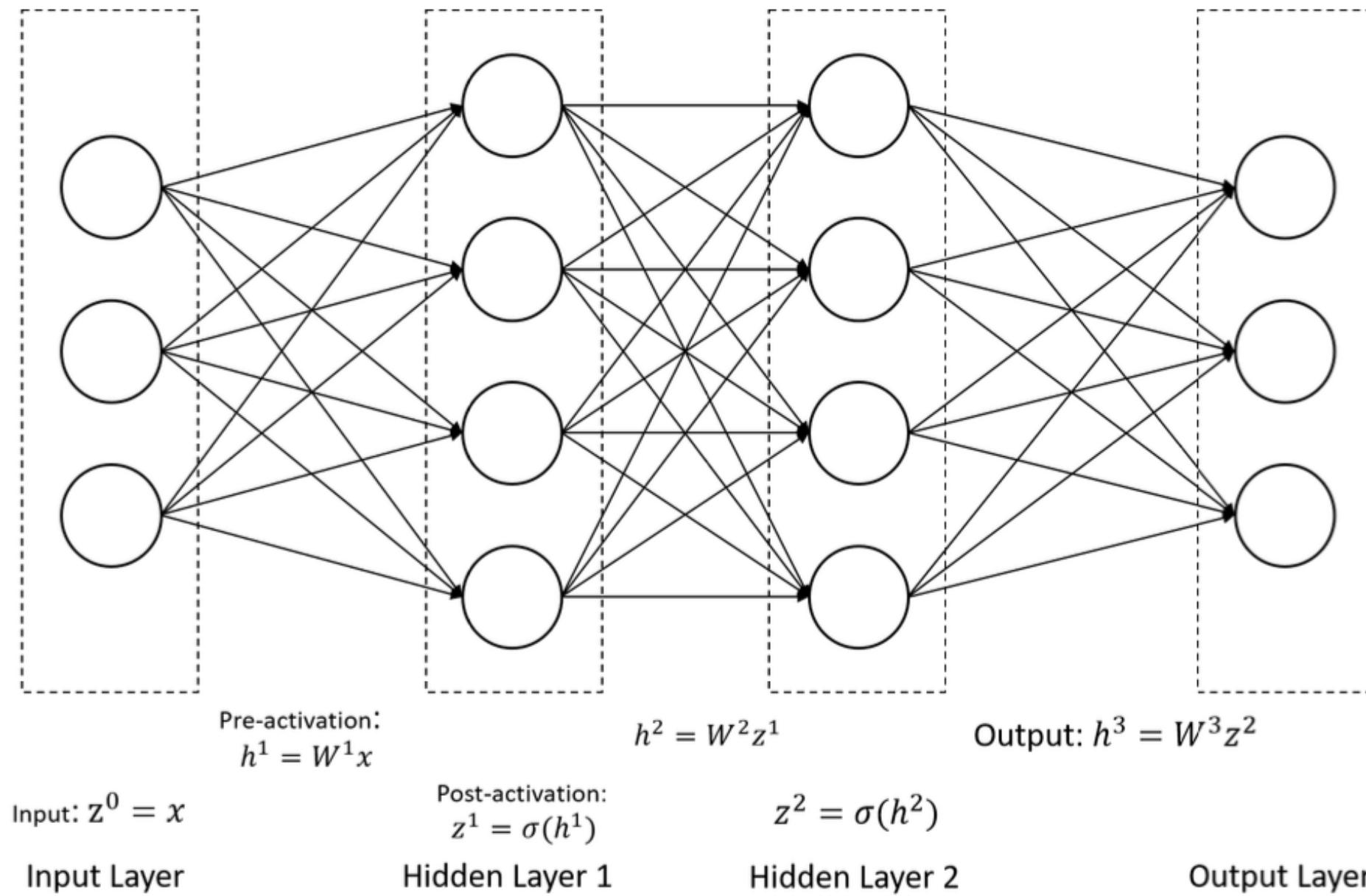


Figure: Example of a 3-layer fully-connected neural network. You should be able to understand its matrix representation.

What is a Matrix?

Let $A = (a_{ij})$ be an $m \times n$ matrix.

$A = \text{np.array}([[1, 2], [3, 4]])$

- The j th column of A is denoted by a column vector \mathbf{a}_j , i.e.,

$$\mathbf{a}_j = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{bmatrix}$$

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

- The i th row of A is denoted by a row vector $\vec{\mathbf{a}}_i$, i.e.,

$$\vec{\mathbf{a}}_i = (a_{i1}, a_{i2}, \dots, a_{in})$$

- Matrix A can be represented in terms of either its columns and rows:

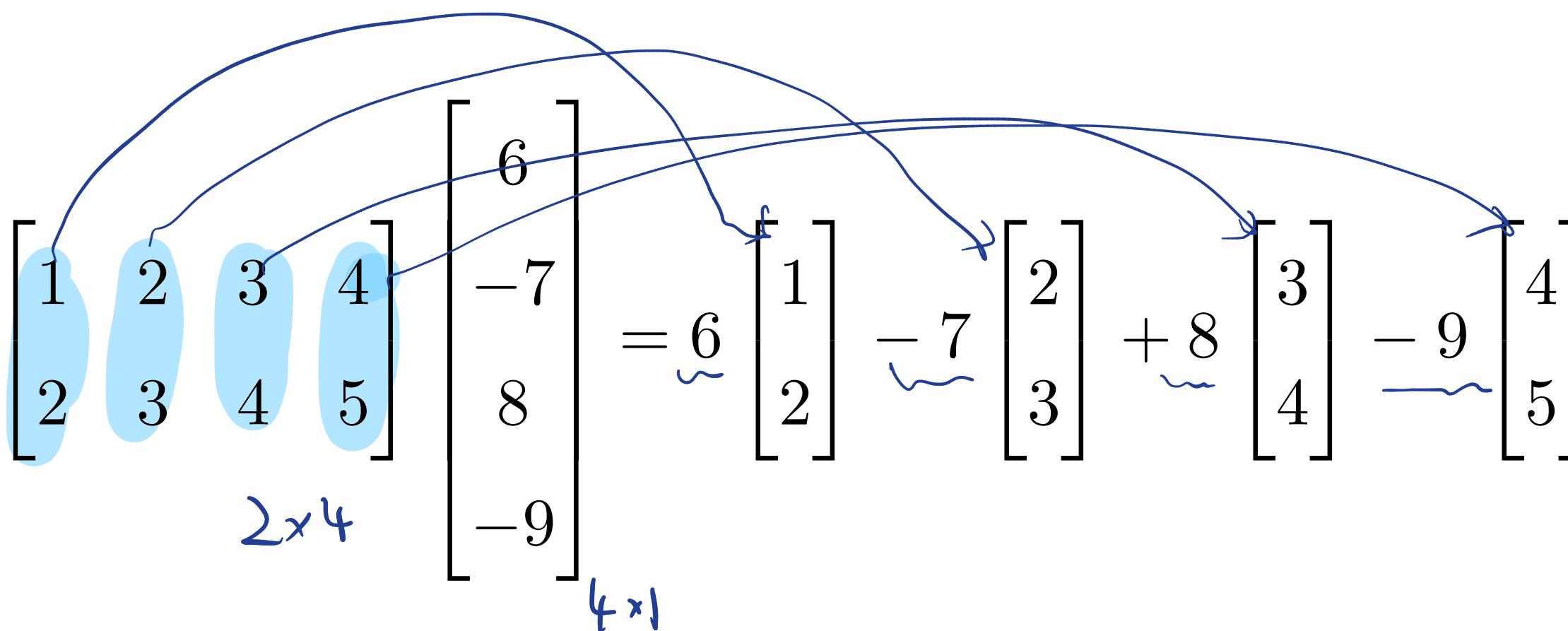
$$A = [\mathbf{a}_1, \dots, \mathbf{a}_n] = \begin{bmatrix} \vec{\mathbf{a}}_1 \\ \vec{\mathbf{a}}_2 \\ \vdots \\ \vec{\mathbf{a}}_m \end{bmatrix}$$

Matrix-Vector Multiplication

For an $m \times n$ matrix A with the i th column \mathbf{a}_i , and a vector $\mathbf{u} = (u_1, u_2, \dots, u_n)^\top$, the multiplication of A and \mathbf{u} is defined as

$$A\mathbf{u} = u_1\mathbf{a}_1 + u_2\mathbf{a}_2 + \cdots + u_n\mathbf{a}_n$$

Example

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \end{bmatrix}_{2 \times 4} \begin{bmatrix} 6 \\ -7 \\ 8 \\ -9 \end{bmatrix}_{4 \times 1} = 6 \begin{bmatrix} 1 \\ 2 \end{bmatrix} - 7 \begin{bmatrix} 2 \\ 3 \end{bmatrix} + 8 \begin{bmatrix} 3 \\ 4 \end{bmatrix} - 9 \begin{bmatrix} 4 \\ 5 \end{bmatrix}$$


Inner Product

- Given a vector $\mathbf{a} = (a_1, \dots, a_n)^\top$ and a vector $\mathbf{b} = (b_1, \dots, b_n)^\top$, following the rule of matrix-vector product, we have

$$\mathbf{a}^\top \mathbf{b} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

- We call this special vector-vector multiplication the **inner product** (scalar product) of \mathbf{a} and \mathbf{b} (denoted by $\mathbf{a}^\top \mathbf{b}$ or $\langle \mathbf{a}, \mathbf{b} \rangle$)
- Properties: Commutative, bilinear
- Application: Cosine similarity, $\cos \theta = \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}$

Inner Product

- Given a vector $\mathbf{a} = (a_1, \dots, a_n)^\top$ and a vector $\mathbf{b} = (b_1, \dots, b_n)^\top$, following the rule of matrix-vector product, we have

$$\mathbf{a}^\top \mathbf{b} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

- We call this special vector-vector multiplication the **inner product** (scalar product) of \mathbf{a} and \mathbf{b} (denoted by $\mathbf{a}^\top \mathbf{b}$ or $\langle \mathbf{a}, \mathbf{b} \rangle$)
- Properties: Commutative, bilinear
- Application: Cosine similarity, $\cos \theta = \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}$

Row Perspective of Multiplication

The matrix-vector multiplication $A\mathbf{u}$ has a row formula as

$$A\mathbf{u} = \begin{bmatrix} \vec{\mathbf{a}}_1 \mathbf{u} \\ \vec{\mathbf{a}}_2 \mathbf{u} \\ \vdots \\ \vec{\mathbf{a}}_m \mathbf{u} \end{bmatrix}$$

- Consider $A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \end{bmatrix}$ and $\mathbf{u} = [6 \quad -7 \quad 8 \quad -9]^\top$.
- We calculate

$$\vec{\mathbf{a}}_1 \mathbf{u} = 6 \cdot 1 - 7 \cdot 2 + 8 \cdot 3 - 9 \cdot 4 = -20$$

$$\vec{\mathbf{a}}_2 \mathbf{u} = 6 \cdot 2 - 7 \cdot 3 + 8 \cdot 4 - 9 \cdot 5 = -22$$

- We see that $A\mathbf{u} = [-20 \quad -22]^\top$

Row Perspective of Multiplication

The matrix-vector multiplication $A\mathbf{u}$ has a row formula as

$$A\mathbf{u} = \begin{bmatrix} \vec{\mathbf{a}}_1 \mathbf{u} \\ \vec{\mathbf{a}}_2 \mathbf{u} \\ \vdots \\ \vec{\mathbf{a}}_m \mathbf{u} \end{bmatrix}$$

- Consider $A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \end{bmatrix}$ and $\mathbf{u} = [6 \quad -7 \quad 8 \quad -9]^\top$.
- We calculate

$$\vec{\mathbf{a}}_1 \mathbf{u} = 6 \cdot 1 - 7 \cdot 2 + 8 \cdot 3 - 9 \cdot 4 = -20$$

$$\vec{\mathbf{a}}_2 \mathbf{u} = 6 \cdot 2 - 7 \cdot 3 + 8 \cdot 4 - 9 \cdot 5 = -22$$

- We see that $A\mathbf{u} = [-20 \quad -22]^\top$

Row Perspective of Multiplication

The matrix-vector multiplication $A\mathbf{u}$ has a row formula as

$$A\mathbf{u} = \begin{bmatrix} \vec{\mathbf{a}}_1 \mathbf{u} \\ \vec{\mathbf{a}}_2 \mathbf{u} \\ \vdots \\ \vec{\mathbf{a}}_m \mathbf{u} \end{bmatrix}$$

- Consider $A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \end{bmatrix}$ and $\mathbf{u} = [6 \quad -7 \quad 8 \quad -9]^\top$.
- We calculate

$$\vec{\mathbf{a}}_1 \mathbf{u} = 6 \cdot 1 - 7 \cdot 2 + 8 \cdot 3 - 9 \cdot 4 = -20$$

$$\vec{\mathbf{a}}_2 \mathbf{u} = 6 \cdot 2 - 7 \cdot 3 + 8 \cdot 4 - 9 \cdot 5 = -22$$

- We see that $A\mathbf{u} = [-20 \quad -22]^\top$

Linear Systems as Matrix Equations

Write the following linear systems into compact matrix form:

$$\begin{cases} 2x_1 + x_2 + x_3 = 5 \\ 4x_1 - 6x_2 = -2 \\ -2x_1 + 7x_2 + 2x_3 = 9 \end{cases} \Rightarrow A\mathbf{x} = \mathbf{b}$$

where

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 4 & -6 & 0 \\ -2 & 7 & 2 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 5 \\ -2 \\ 9 \end{bmatrix}$$

Rank of a Matrix

- The rank of a matrix A is the number of linearly independent columns
- Equivalently, it is the number of linearly independent rows
- Example: $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$ has rank 1
- Full rank: $\text{rank}(A) = \min(m, n)$ for $A \in \mathbb{R}^{m \times n}$
- Application: Determines solvability of linear systems $A\mathbf{x} = \mathbf{b}$

Rank of a Matrix

- The rank of a matrix A is the number of linearly independent columns
- Equivalently, it is the number of linearly independent rows
- Example: $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$ has rank 1
- Full rank: $\text{rank}(A) = \min(m, n)$ for $A \in \mathbb{R}^{m \times n}$
- Application: Determines solvability of linear systems $A\mathbf{x} = \mathbf{b}$

Identity Matrix

- The identity matrix of order k , denoted by I or I_k , is a $k \times k$ square matrix whose diagonal elements are 1's and whose nondiagonal elements are 0's

$$I = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

- Properties: $AI = A$ for any compatible matrix A .

Identity Matrix

- The identity matrix of order k , denoted by I or I_k , is a $k \times k$ square matrix whose diagonal elements are 1's and whose nondiagonal elements are 0's

$$I = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

- Properties: $AI = A$ for any compatible matrix A .

Inverse of a Matrix

- Let A be a $k \times k$ matrix. The inverse of A , denoted by A^{-1} , is another $k \times k$ matrix such that

$$AA^{-1} = A^{-1}A = I$$

- If the inverse exists, it is unique
- Existence: A^{-1} exists if and only if $\det(A) \neq 0$ (or equivalently, $\text{rank}(A) = k$)
- For 2×2 matrix:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Inverse of a Matrix

- Let A be a $k \times k$ matrix. The inverse of A , denoted by A^{-1} , is another $k \times k$ matrix such that

$$AA^{-1} = A^{-1}A = I \quad BA^2 = AB = I$$

Assume B, C $CA = AC = I$

$$BAC = (BA)C = C$$
$$\Rightarrow B(AC) = B$$

- If the inverse exists, it is unique
- Existence: A^{-1} exists if and only if $\det(A) \neq 0$ (or equivalently, $\text{rank}(A) = k$)
- For 2×2 matrix:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Inverse of a Matrix

- Let A be a $k \times k$ matrix. The inverse of A , denoted by A^{-1} , is another $k \times k$ matrix such that

$$AA^{-1} = A^{-1}A = I$$

- If the inverse exists, it is unique
- Existence: A^{-1} exists if and only if $\det(A) \neq 0$ (or equivalently, $\text{rank}(A) = k$)
- For 2×2 matrix:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Inverse of a Matrix

- Let A be a $k \times k$ matrix. The inverse of A , denoted by A^{-1} , is another $k \times k$ matrix such that

$$AA^{-1} = A^{-1}A = I$$

- If the inverse exists, it is unique
- Existence: A^{-1} exists if and only if $\det(A) \neq 0$ (or equivalently, $\text{rank}(A) = k$)
- For 2×2 matrix:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$\det(A) = ad - bc \quad A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Transpose of a Matrix

- Let A be an $n \times k$ matrix. The transpose of A , denoted by A^\top , is a $k \times n$ matrix whose columns are the rows of A

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nk} \end{bmatrix} \Rightarrow A^\top = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & & \vdots \\ a_{1k} & a_{2k} & \cdots & a_{nk} \end{bmatrix}$$

- Properties: $(A^\top)^\top = A$, $(AB)^\top = B^\top A^\top$

Transpose of a Matrix

- Let A be an $n \times k$ matrix. The transpose of A , denoted by A^\top , is a $k \times n$ matrix whose columns are the rows of A

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nk} \end{bmatrix} \Rightarrow A^\top = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & & \vdots \\ a_{1k} & a_{2k} & \cdots & a_{nk} \end{bmatrix}$$

- Properties: $(A^\top)^\top = A$, $(AB)^\top = B^\top A^\top$

$A: n \times k$ $B: k \times m$

$$\begin{aligned} ((AB)^\top)_{ij} &= (AB)_{j,i} = \sum_e a_{je} b_{ei} \\ &= \sum_e b_{ei} a_{je} = (B^\top A^\top)_{j,i} \end{aligned}$$

Symmetric Matrices

- Let A be a $k \times k$ matrix. A is said to be symmetric if

$$A = A^\top$$

- Examples: Covariance matrices, Hessian matrices
- Properties: Real eigenvalues, orthogonal eigenvectors
- Spectral theorem: $A = Q\Lambda Q^\top$ where Q is orthogonal and Λ is diagonal

Symmetric Matrices

- Let A be a $k \times k$ matrix. A is said to be symmetric if

$$A = A^\top$$

$$a = n \times 1$$

$$A = aa^\top$$

$$A^\top = (aa^\top)^\top$$

$$= (a^\top)^\top a^\top$$

$$= aa^\top = A$$

- Examples: Covariance matrices, Hessian matrices

- Properties: Real eigenvalues, orthogonal eigenvectors

- Spectral theorem: $A = Q\Lambda Q^\top$ where Q is orthogonal and Λ is diagonal

Symmetric Matrices

- Let A be a $k \times k$ matrix. A is said to be symmetric if

$$A = A^\top$$

- Examples: Covariance matrices, Hessian matrices
- Properties: Real eigenvalues, orthogonal eigenvectors
- Spectral theorem: $A = Q\Lambda Q^\top$ where Q is orthogonal and Λ is diagonal

Symmetric Matrices

- Let A be a $k \times k$ matrix. A is said to be symmetric if

$$A = A^\top$$

- Examples: Covariance matrices, Hessian matrices
- Properties: Real eigenvalues, orthogonal eigenvectors
- Spectral theorem: $A = Q\Lambda Q^\top$ where Q is orthogonal and Λ is diagonal

Idempotent Matrices

- Let A be a $k \times k$ matrix. A is called idempotent if

$$A = AA$$

- If A is also symmetric, then A is called symmetric idempotent
- If A is symmetric idempotent, then $I - A$ is also symmetric idempotent
- Example: Projection matrices $P = X(X^\top X)^{-1}X^\top$

Idempotent Matrices

- Let A be a $k \times k$ matrix. A is called idempotent if

$$A = AA$$

- If A is also symmetric, then A is called symmetric idempotent
- If A is symmetric idempotent, then $I - A$ is also symmetric idempotent
- Example: Projection matrices $P = X(X^\top X)^{-1}X^\top$

Idempotent Matrices

$$(\mathbf{I} - \mathbf{A})^\top = \mathbf{I}^\top - \mathbf{A}^\top$$

- Let A be a $k \times k$ matrix. A is called idempotent if

$$= \mathbf{I} - \mathbf{A}$$

$$\mathbf{A} = \mathbf{AA}$$

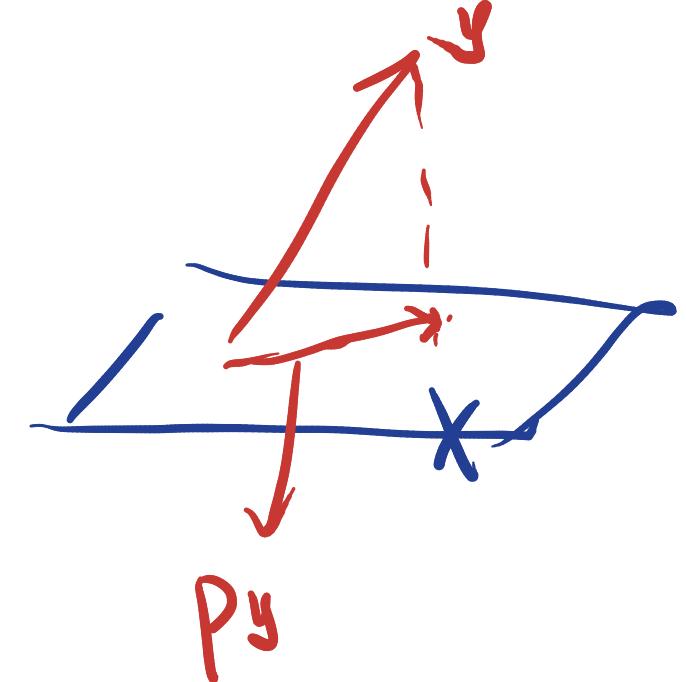
- If A is also symmetric, then A is called symmetric idempotent
- If A is symmetric idempotent, then $\mathbf{I} - \mathbf{A}$ is also symmetric

idempotent

$$\begin{aligned} (\mathbf{I} - \mathbf{A})(\mathbf{I} - \mathbf{A}) &= \mathbf{I}(\mathbf{I} - \mathbf{A}) - \mathbf{A}(\mathbf{I} - \mathbf{A}) = \mathbf{I} - \mathbf{A} + (-\mathbf{A} + \mathbf{AA}) \\ &= \mathbf{I} - \mathbf{A} + (-\mathbf{A} + \mathbf{A}) \\ &= \mathbf{I} - \mathbf{A}. \end{aligned}$$

- Example: Projection matrices $P = X(X^\top X)^{-1}X^\top$

Idempotent Matrices



- Let A be a $k \times k$ matrix. A is called idempotent if

$$A = AA$$

$$PPy = Py, \forall y$$

↑

$$PP = P.$$

- If A is also symmetric, then A is called symmetric idempotent
- If A is symmetric idempotent, then $I - A$ is also symmetric idempotent
- Example: Projection matrices $P = X(X^\top X)^{-1}X^\top$

$$P^\top = P$$

$$\begin{aligned} PP &= \left(X(X^\top X)^{-1}X^\top \right) \left(X(X^\top X)^{-1}X^\top \right) \\ &= X(X^\top X)^{-1} \cancel{\left(X^\top X \right)} \cancel{\left(X^\top X \right)^{-1}} X^\top = X(X^\top X)^{-1}X^\top = P \end{aligned}$$

• $Ax = b$. What if this system have no solution?

$$\min_x \|Ax - b\|_2^2 = \sum_{i=1}^m (a_i^T x - b_i)^2 \rightarrow F(x)$$

$$\frac{\partial F(x)}{\partial x} = 2 A^T (Ax - b) = 0$$

$$\Rightarrow A^T A x = A^T b \quad (\text{normal equation})$$

$$x^* = (A^T A)^{-1} A^T b \quad (\text{Assume } A^T A \text{ inv.})$$

$$Ax^* \approx b$$

$$C(A) = \text{span}\{a_1, \dots, a_n\}$$

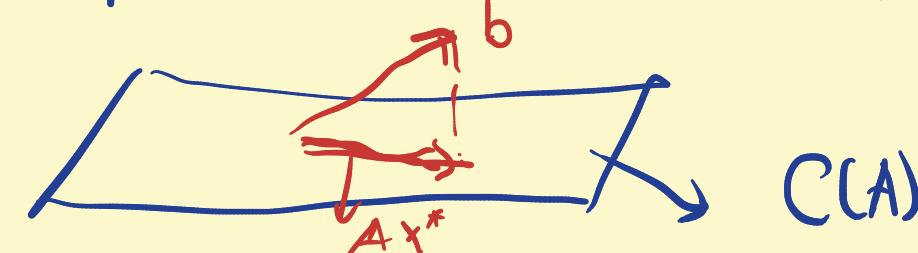
$$Ax^* = A(A^T A)^{-1} A^T b$$

↓
projection matrix

$$A = [a_1, \dots, a_n]$$

$$\textcircled{1} \quad Ax^* = b : b \in \text{column space of } A \Rightarrow \|Ax^* - b\|_2^2 = 0$$

\textcircled{2} otherwise,



$$\arg \min_{z \in C(A)} \|z - b\|_2^2$$

$$P = A(A^T A)^{-1} A^T$$

① $b \in C(A) \Leftrightarrow \exists x \text{ s.t. } Ax = b$

$$Pb = A(A^T A)^{-1} A^T b$$

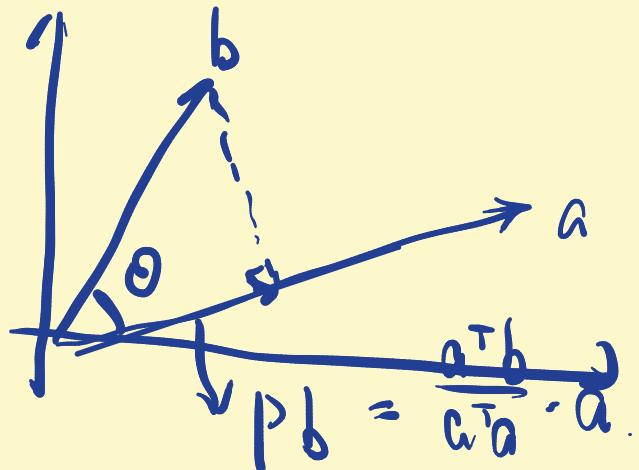
$$= A(A^T A)^{-1} A^T A x$$

$$= Ax = b$$

② $A = [a] \in \mathbb{R}^{m \times 1}$

$$P = A(A^T A)^{-1} A^T = a(a^T a)^{-1} a^T = \frac{aa^T}{a^T a}$$

$$Pb = \frac{aa^T b}{a^T a} = \frac{\langle a, b \rangle}{\|a\|_2^2} \cdot a = \frac{\|a\| \|b\| \cos \theta}{\|a\|_2^2} \cdot a = \frac{a}{\|a\|} \cdot \frac{\|b\| \cos \theta}{\|a\|_2}$$



Orthonormal Matrices

- Let A be a $k \times k$ matrix. If A is an orthonormal matrix, then

$$A^\top A = I$$

- As a consequence, if A is an orthonormal matrix, then

$$A^{-1} = A^\top$$

- Properties: Preserves norms and angles ($\|Ax\| = \|x\|$)
- Examples: Rotation matrices, permutation matrices

Orthonormal Matrices

- Let A be a $k \times k$ matrix. If A is an orthonormal matrix, then

$$A^\top A = I$$

- As a consequence, if A is an orthonormal matrix, then

$$A^{-1} = A^\top$$

- Properties: Preserves norms and angles ($\|Ax\| = \|x\|$)
- Examples: Rotation matrices, permutation matrices

Orthonormal Matrices

- Let A be a $k \times k$ matrix. If A is an orthonormal matrix, then

$$A^\top A = I$$

- As a consequence, if A is an orthonormal matrix, then

$$A^{-1} = A^\top$$

$$\begin{aligned}\|Ax\|_2^2 &= \langle Ax, Ax \rangle \\ &= x^\top A^\top A x\end{aligned}$$

- Properties: Preserves norms and angles ($\|Ax\| = \|x\|$)

$$= x^\top x$$

- Examples: Rotation matrices, permutation matrices

$$= \|x\|_2^2$$

Orthonormal Matrices

- Let A be a $k \times k$ matrix. If A is an orthonormal matrix, then

$$A^\top A = I$$

- As a consequence, if A is an orthonormal matrix, then

$$A^{-1} = A^\top$$

- Properties: Preserves norms and angles ($\|Ax\| = \|x\|$)
- Examples: Rotation matrices, permutation matrices

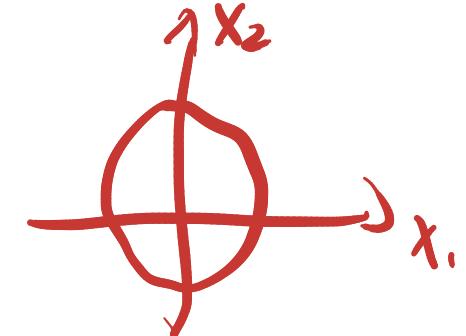
Quadratic Forms

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad y = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$y^T A y = 1 = x_1^2 + x_2^2$$

- Let y be a $k \times 1$ vector, and let A be a $k \times k$ matrix. The function

$$y^T A y = \sum_{i=1}^k \sum_{j=1}^k a_{ij} y_i y_j$$



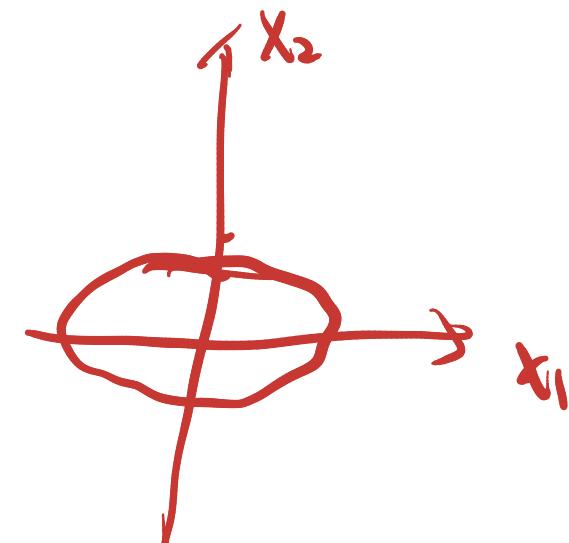
is called a quadratic form

$$A = \begin{pmatrix} 3 & 0 \\ 0 & 5 \end{pmatrix} \quad y = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

- Geometric interpretation: Ellipsoids in k -dimensional space

$$y^T A y = 3x_1^2 + 5x_2^2$$

- Example: Energy in physical systems, Mahalanobis distance



Quadratic Forms

- Let y be a $k \times 1$ vector, and let A be a $k \times k$ matrix. The function

$$y^\top A y = \sum_{i=1}^k \sum_{j=1}^k a_{ij} y_i y_j$$

is called a quadratic form

- Geometric interpretation: Ellipsoids in k -dimensional space
- Example: Energy in physical systems, Mahalanobis distance

$$\|x - y\|_2^2 = (x - y)^\top (x - y)$$

$$\|x - y\|_A^2 = \sqrt{\frac{1}{2} (x - y)^\top A (x - y)}$$

Positive Definite and Positive Semidefinite Matrices

Let A be a $k \times k$ matrix.

- A is said to be *positive definite* if
 - (a) $A = A^\top$ (A is symmetric)
 - (b) $\mathbf{y}^\top A \mathbf{y} > 0 \quad \forall \mathbf{y} \in \mathbb{R}^k, \mathbf{y} \neq 0$
- A is said to be *positive semidefinite* if:
 - (a) $A = A^\top$ (A is symmetric)
 - (c) $\mathbf{y}^\top A \mathbf{y} \geq 0 \quad \forall \mathbf{y} \in \mathbb{R}^k$
- Tests: Eigenvalues > 0 (positive definite), eigenvalues ≥ 0 (positive semidefinite)
- Application: Convex optimization, kernel methods

Positive Definite and Positive Semidefinite Matrices

Let A be a $k \times k$ matrix.

- A is said to be *positive definite* if
 - (a) $A = A^\top$ (A is symmetric)
 - (b) $\mathbf{y}^\top A \mathbf{y} > 0 \quad \forall \mathbf{y} \in \mathbb{R}^k, \mathbf{y} \neq 0$
- A is said to be *positive semidefinite* if:
 - (a) $A = A^\top$ (A is symmetric)
 - (c) $\mathbf{y}^\top A \mathbf{y} \geq 0 \quad \forall \mathbf{y} \in \mathbb{R}^k$
- Tests: Eigenvalues > 0 (positive definite), eigenvalues ≥ 0 (positive semidefinite)
- Application: Convex optimization, kernel methods

Positive Definite and Positive Semidefinite Matrices

Let A be a $k \times k$ matrix.

- A is said to be *positive definite* if

- (a) $A = A^\top$ (A is symmetric)
- (b) $\mathbf{y}^\top A \mathbf{y} > 0 \quad \forall \mathbf{y} \in \mathbb{R}^k, \mathbf{y} \neq 0$

- A is said to be *positive semidefinite* if:

- (a) $A = A^\top$ (A is symmetric)
- (c) $\mathbf{y}^\top A \mathbf{y} \geq 0 \quad \forall \mathbf{y} \in \mathbb{R}^k$

- Tests: Eigenvalues > 0 (positive definite), eigenvalues ≥ 0 (positive semidefinite)
- Application: Convex optimization, kernel methods

$$\begin{array}{c} A \\ (\lambda, x) \Rightarrow \begin{array}{l} \downarrow \text{eigenvalues} \\ \downarrow \text{eigenvector} \end{array} \\ A x = \lambda x \\ (A - \lambda I) x = 0 \\ \downarrow \\ \det(A - \lambda I) = 0 \end{array}$$

Positive Definite and Positive Semidefinite Matrices

Let A be a $k \times k$ matrix.

- A is said to be *positive definite* if

$$(a) A = A^\top \text{ (A is symmetric)}$$

$$(b) \mathbf{y}^\top A \mathbf{y} > 0 \quad \forall \mathbf{y} \in \mathbb{R}^k, \mathbf{y} \neq 0$$

- A is said to be *positive semidefinite* if:

$$(a) A = A^\top \text{ (A is symmetric)}$$

$$(c) \mathbf{y}^\top A \mathbf{y} \geq 0 \quad \forall \mathbf{y} \in \mathbb{R}^k$$

- Tests: Eigenvalues > 0 (positive definite), eigenvalues ≥ 0 (positive semidefinite)
- Application: Convex optimization, kernel methods

$$\textcircled{1} \quad A = B^\top B$$

$$\mathbf{y}^\top A \mathbf{y} = \mathbf{y}^\top B^\top B \mathbf{y} = \|B\mathbf{y}\|_2^2 \geq 0$$

$$\textcircled{2} \quad A = c_1 b_1 b_1^\top + c_2 b_2 b_2^\top + \dots + c_m b_m b_m^\top$$

$$\mathbf{y}^\top A \mathbf{y} = \sum_{i=1}^m c_i \mathbf{y}^\top b_i b_i^\top \mathbf{y}$$

$$= \sum_{i=1}^m c_i (b_i^\top \mathbf{y})^2 \geq 0 \quad c_1, \dots, c_m \geq 0$$

Trace of a Matrix

Let A be a $k \times k$ matrix. The *trace* of A , denoted by $\text{trace}(A)$ or $\text{tr}(A)$, is the sum of the diagonal elements of A ; thus,

$$\text{trace}(A) = \sum_{i=1}^k a_{ii}$$

Properties:

1. If A is an $m \times n$ matrix and B is an $n \times m$ matrix, then

$$\begin{aligned} & \sum_{i=1}^k (AB)_{i,i} \\ &= \sum_{i=1}^k \sum_{e=1}^n a_{i,e} B_{e,i} \end{aligned}$$

2. If the matrices are appropriately conformable, then

$$\text{trace}(AB) = \text{trace}(BA) \quad \sum_{i=1}^k (BA)_{i,i} = \sum_{i=1}^k \sum_{e=1}^n B_{i,e} A_{e,i}$$

$$\text{trace}(ABC) = \text{trace}(CAB) = \sum_e \sum_i A_{e,i} B_{i,e}$$

3. If A and B are $k \times k$ matrices and a and b are scalars, then

$$\text{trace}(aA + bB) = a\text{trace}(A) + b\text{trace}(B)$$

Trace of a Matrix

Let A be a $k \times k$ matrix. The *trace* of A , denoted by $\text{trace}(A)$ or $\text{tr}(A)$, is the sum of the diagonal elements of A ; thus,

$$\text{trace}(A) = \sum_{i=1}^k a_{ii}$$

Properties:

1. If A is an $m \times n$ matrix and B is an $n \times m$ matrix, then

$$\text{trace}(AB) = \text{trace}(BA)$$

2. If the matrices are appropriately conformable, then

$$\text{trace}(ABC) = \text{trace}(CAB) \quad \text{trace}(BCA)$$

3. If A and B are $k \times k$ matrices and a and b are scalars, then

$$\text{trace}(aA + bB) = a\text{trace}(A) + b\text{trace}(B)$$

Trace of a Matrix

Let A be a $k \times k$ matrix. The *trace* of A , denoted by $\text{trace}(A)$ or $\text{tr}(A)$, is the sum of the diagonal elements of A ; thus,

$$\text{trace}(A) = \sum_{i=1}^k a_{ii}$$

Properties:

1. If A is an $m \times n$ matrix and B is an $n \times m$ matrix, then

$$\text{trace}(AB) = \text{trace}(BA)$$

2. If the matrices are appropriately conformable, then

$$\text{trace}(ABC) = \text{trace}(CAB)$$

3. If A and B are $k \times k$ matrices and a and b are scalars, then

$$\text{trace}(aA + bB) = a\text{trace}(A) + b\text{trace}(B)$$

Rank of an Idempotent Matrix

Assume (λ, x) is eigen-pair of A .

$$Ax = \lambda x$$

$$AA = A$$

$$AAx = A(Ax) = A(\lambda x) = \lambda(Ax) = \lambda^2 x \quad \}$$

- Let A be an idempotent matrix. The rank of A is equal to its trace

$$\text{rank}(A) = \text{trace}(A)$$

$$\Rightarrow \lambda x = \lambda^2 x$$

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

- Proof sketch: Use the fact that idempotent matrices are $\Rightarrow \lambda = 0$ or $\lambda = 1$

diagonalizable with eigenvalues 0 or 1

$$\textcircled{1} \text{ Trace}(A) = \sum_{i=1}^n \lambda_i$$

$= \# \text{ of } 1_s \text{ of }$
 eigenvalues

- Application: In regression, $\text{rank}(X) = \text{trace}(H)$ where

$H = X(X^\top X)^{-1}X^\top$ is the hat matrix

$\textcircled{2} \text{ Rank}(A) = \# \text{ of } 1_s \text{ of }$
 eigenvalues

Rank of an Idempotent Matrix

- Let A be an idempotent matrix. The rank of A is equal to its trace

$$\text{rank}(A) = \text{trace}(A)$$

- Proof sketch: Use the fact that idempotent matrices are diagonalizable with eigenvalues 0 or 1
- Application: In regression, $\text{rank}(X) = \text{trace}(H)$ where $H = X(X^\top X)^{-1}X^\top$ is the hat matrix

An Important Identity for a Partitioned Matrix

Let \mathbf{X} be an $n \times p$ matrix partitioned such that

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$$

We note that

$$\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} = \mathbf{X}$$

$$\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top [\mathbf{X}_1 \ \mathbf{X}_2] = \mathbf{X}$$

$$\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top [\mathbf{X}_1 \ \mathbf{X}_2] = [\mathbf{X}_1 \ \mathbf{X}_2]$$

Consequently,

$$\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}_1 = \mathbf{X}_1 \quad \text{and} \quad \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}_2 = \mathbf{X}_2$$

Similarly,

$$\mathbf{X}_1^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{X}_1^\top \quad \text{and} \quad \mathbf{X}_2^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{X}_2^\top$$

Inverse of a Partitioned Matrix

Consider a matrix of the form

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} \mathbf{X}_1^\top \mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{X}_2 \end{bmatrix}$$

It can be shown that the inverse of this matrix is $(\mathbf{X}^\top \mathbf{X})^{-1}$ that equals

$$\begin{bmatrix} (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} + (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{X}_2 G \mathbf{X}_2^\top \mathbf{X}_1 (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} & -(\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{X}_2 G \\ -G \mathbf{X}_2^\top \mathbf{X}_1 (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} & G \end{bmatrix}$$

where

$$\mathbf{H}_1 = \mathbf{X}_1 (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \quad \text{and} \quad G = [\mathbf{X}_2^\top (\mathbf{I} - \mathbf{H}_1) \mathbf{X}_2]^{-1}$$

Application: Regression analysis with multiple groups of predictors

We will show that

$$\begin{bmatrix} (X_1^T X_1)^{-1} + (X_1^T X_1)^{-1} X_1^T X_2 G X_2^T X_1 (X_1^T X_1)^{-1} & - (X_1^T X_1)^{-1} X_1^T X_2 G \\ - G X_2^T X_1 (X_1^T X_1)^{-1} & G \end{bmatrix} \begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{bmatrix} = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} = I.$$

① We can verify

$$\begin{aligned} M_{11} &= \left[(X_1^T X_1)^{-1} + (X_1^T X_1)^{-1} X_1^T X_2 G X_2^T X_1 (X_1^T X_1)^{-1} \right] X_1^T X_1 + \left[- (X_1^T X_1)^{-1} X_1^T X_2 G \right] X_2^T X_1 \\ &= I + (X_1^T X_1)^{-1} X_1^T X_2 G X_2^T X_1 - (X_1^T X_1)^{-1} X_1^T X_2 G X_2^T X_1 = I \end{aligned}$$

$$\begin{aligned} M_{12} &= \left[(X_1^T X_1)^{-1} + (X_1^T X_1)^{-1} X_1^T X_2 G X_2^T X_1 (X_1^T X_1)^{-1} \right] X_1^T X_2 + \left[- (X_1^T X_1)^{-1} X_1^T X_2 G \right] X_2^T X_2 \\ &= (X_1^T X_1)^{-1} X_1^T X_2 + \left[(X_1^T X_1)^{-1} X_1^T X_2 \right] G \left[X_2^T X_1 (X_1^T X_1)^{-1} X_1^T X_2 - X_2^T X_2 \right] \\ &= (X_1^T X_1)^{-1} X_1^T X_2 - \left[(X_1^T X_1)^{-1} X_1^T X_2 \right] G G^{-1} \\ &= 0 \end{aligned}$$

③ Similarly $M_{21} = 0$

$$\begin{aligned} ④ M_{22} &= -G X_2^T X_1 (X_1^T X_1)^{-1} X_1^T X_2 + G X_2^T X_2 = G X_2^T [I - X_1 (X_1^T X_1)^{-1} X_1^T] X_2 \\ &= G G^{-1} = I \end{aligned}$$

Determinant

- The determinant of a square matrix A , denoted $\det(A)$ or $|A|$, is a scalar value
- Geometric interpretation: Scaling factor of the linear transformation
- For 2×2 matrix: $\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$
- Properties:
 - $\det(AB) = \det(A) \det(B)$
 - $\det(A^{-1}) = 1/\det(A)$
 - $\det(A^\top) = \det(A)$
- Application: Testing invertibility, change of variables in integration

Determinant

- The determinant of a square matrix A , denoted $\det(A)$ or $|A|$, is a scalar value
- Geometric interpretation: Scaling factor of the linear transformation
- For 2×2 matrix: $\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$
- Properties:
 - $\det(AB) = \det(A) \det(B)$
 - $\det(A^{-1}) = 1/\det(A)$
 - $\det(A^\top) = \det(A)$
- Application: Testing invertibility, change of variables in integration

Determinant

- The determinant of a square matrix A , denoted $\det(A)$ or $|A|$, is a scalar value
- Geometric interpretation: Scaling factor of the linear transformation
- For 2×2 matrix: $\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$
- Properties:
 - $\det(AB) = \det(A) \det(B)$
 - $\det(A^{-1}) = 1/\det(A)$ $\det(A A^{-1}) = \det(I) = 1 = \det(A) \det(A^{-1})$
 - $\det(A^\top) = \det(A)$
- Application: Testing invertibility, change of variables in integration

Determinant

- The determinant of a square matrix A , denoted $\det(A)$ or $|A|$, is a scalar value
- Geometric interpretation: Scaling factor of the linear transformation
- For 2×2 matrix: $\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$
- Properties:
 - $\det(AB) = \det(A) \det(B)$
 - $\det(A^{-1}) = 1/\det(A)$
 - $\det(A^\top) = \det(A)$
- Application: Testing invertibility, change of variables in integration

Contents

- Matrix Operations
- Matrix Derivative and Expectations
- Applications and Wrap-Up

Matrix Derivatives

(Matrix Cookbook)

Let A be a $k \times k$ matrix of constants, a be a $k \times 1$ vector of constants, and y be a $k \times 1$ vector of variables.

1. If $z = a^\top y$, then

$$\frac{\partial z}{\partial y} = \frac{\partial(a^\top y)}{\partial y} = a$$

2. If $z = y^\top y$, then

$$\frac{\partial z}{\partial y} = \frac{\partial(y^\top y)}{\partial y} = 2y$$

3. If $z = a^\top A y$, then

$$\frac{\partial z}{\partial y} = \frac{\partial(a^\top A y)}{\partial y} = A^\top a$$

4. If $z = y^\top A y$ and A is symmetric, then

$$\frac{\partial z}{\partial y} = \frac{\partial(y^\top A y)}{\partial y} = 2A y$$

$$\begin{aligned}\frac{\partial z}{\partial y} &= \left(\frac{\partial z}{\partial y_i} \right)_i \\ &= \left(\frac{\partial}{\partial y_i} \sum_j c_{ij} y_j \right)_i \\ &= \left(\frac{\partial}{\partial y_i} a_{ii} y_i \right)_i = (a_i)_{ii} = a_i\end{aligned}$$

Matrix Derivatives

Let A be a $k \times k$ matrix of constants, a be a $k \times 1$ vector of constants, and y be a $k \times 1$ vector of variables.

1. If $z = a^\top y$, then

$$\frac{\partial z}{\partial y} = \frac{\partial(a^\top y)}{\partial y} = a$$

2. If $z = y^\top y$, then

$$\frac{\partial z}{\partial y} = \frac{\partial(y^\top y)}{\partial y} = 2y$$

3. If $z = a^\top A y$, then

$$\frac{\partial z}{\partial y} = \frac{\partial(a^\top A y)}{\partial y} = A^\top a$$

4. If $z = y^\top A y$ and A is symmetric, then

$$\frac{\partial z}{\partial y} = \frac{\partial(y^\top A y)}{\partial y} = 2Ay$$

Matrix Derivatives

Let A be a $k \times k$ matrix of constants, a be a $k \times 1$ vector of constants, and y be a $k \times 1$ vector of variables.

1. If $z = a^\top y$, then

$$\frac{\partial z}{\partial y} = \frac{\partial(a^\top y)}{\partial y} = a$$

2. If $z = y^\top y$, then

$$\frac{\partial z}{\partial y} = \frac{\partial(y^\top y)}{\partial y} = 2y$$

3. If $z = a^\top A y$, then

$$\frac{\partial z}{\partial y} = \frac{\partial(a^\top A y)}{\partial y} = A^\top a$$

4. If $z = y^\top A y$ and A is symmetric, then

$$\frac{\partial z}{\partial y} = \frac{\partial(y^\top A y)}{\partial y} = 2A y$$

$$z = b^\top y \quad b = A^\top a$$
$$\Rightarrow \frac{\partial z}{\partial y} = b = A^\top a$$

Matrix Derivatives

Let A be a $k \times k$ matrix of constants, a be a $k \times 1$ vector of constants, and y be a $k \times 1$ vector of variables.

1. If $z = a^\top y$, then

$$\frac{\partial z}{\partial y} = \frac{\partial(a^\top y)}{\partial y} = a$$

2. If $z = y^\top y$, then

$$\frac{\partial z}{\partial y} = \frac{\partial(y^\top y)}{\partial y} = 2y$$

3. If $z = a^\top A y$, then

$$\frac{\partial z}{\partial y} = \frac{\partial(a^\top A y)}{\partial y} = A^\top a$$

4. If $z = y^\top A y$ and A is symmetric, then

$$\frac{\partial z}{\partial y} = \frac{\partial(y^\top A y)}{\partial y} = 2A y$$

Let $Z = \mathbf{y}^T \mathbf{A} \mathbf{y}$

$$\frac{\partial Z}{\partial y_e} = \frac{\partial}{\partial y_e} \sum_{(i,j)} a_{i,j} y_i y_j$$

$$= \frac{\partial}{\partial y_e} \left[\sum_{i=j=e} a_{e,e} y_e^2 + \sum_{\substack{i=j \\ i \neq e}} a_{i,j} y_i y_j + \sum_{\substack{j=e \\ i \neq e}} a_{i,j} y_i y_j \right]$$

$$= 2 a_{e,e} y_e + \sum_{j \neq e} a_{e,j} y_j + \sum_{i \neq e} a_{i,e} y_i$$

$$= \sum_j a_{e,j} y_j + \sum_i a_{i,e} y_i$$

$$\Rightarrow \frac{\partial Z}{\partial y} = \left(\sum_j a_{e,j} y_j \right)_e + \left(\sum_i a_{i,e} y_i \right)_e$$

$$= A_y + A^T y \quad \overline{\Downarrow} \quad \geq A_y$$

if assume A symmetric

More Derivative Rules

- Application: Gradient descent optimization

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla f(\mathbf{w}_t)$$

where $\nabla f(\mathbf{w})$ is the gradient of the objective function

- Example: For linear regression with loss $L(\mathbf{w}) = \|\mathbf{y} - X\mathbf{w}\|^2$, the gradient is

$$\nabla L(\mathbf{w}) = -2X^\top(\mathbf{y} - X\mathbf{w})$$

- Chain rule for matrix derivatives: If $z = f(\mathbf{y})$ and $\mathbf{y} = g(\mathbf{x})$, then

$$\frac{\partial z}{\partial \mathbf{x}} = \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \right)^\top \frac{\partial z}{\partial \mathbf{y}}$$

More Derivative Rules

- Application: Gradient descent optimization

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla f(\mathbf{w}_t)$$

where $\nabla f(\mathbf{w})$ is the gradient of the objective function

- Example: For linear regression with loss $L(\mathbf{w}) = \|\mathbf{y} - X\mathbf{w}\|^2$, the gradient is

$$\nabla L(\mathbf{w}) = -2X^\top(\mathbf{y} - X\mathbf{w})$$

- Chain rule for matrix derivatives: If $z = f(\mathbf{y})$ and $\mathbf{y} = g(\mathbf{x})$, then

$$\frac{\partial z}{\partial \mathbf{x}} = \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \right)^\top \frac{\partial z}{\partial \mathbf{y}}$$

Expectations of Random Vectors

Let A be a $k \times k$ matrix of constants, a be a $k \times 1$ vector of constants, and y be a $k \times 1$ random vector with mean μ and nonsingular variance–covariance matrix V .

1. $\mathbb{E}(a^\top y) = a^\top \mu$
2. $\mathbb{E}(Ay) = A\mu$
3. $\text{Var}(a^\top y) = a^\top Va$
4. $\text{Var}(Ay) = AVA^\top$

Note: If $V = \sigma^2 I$, then $\text{Var}(Ay) = \sigma^2 AA^\top$

5. $\mathbb{E}(y^\top Ay) = \text{trace}(AV) + \mu^\top A\mu$

Note: If $V = \sigma^2 I$, then $\mathbb{E}(y^\top Ay) = \sigma^2 \text{trace}(A) + \mu^\top A\mu$

Expectations of Random Vectors

Let A be a $k \times k$ matrix of constants, a be a $k \times 1$ vector of constants, and y be a $k \times 1$ random vector with mean μ and nonsingular variance–covariance matrix V .

$$1. \quad \mathbb{E}(a^\top y) = a^\top \mu$$

$$2. \quad \mathbb{E}(Ay) = A\mu$$

$$3. \quad \text{Var}(a^\top y) = a^\top Va$$

$$4. \quad \text{Var}(Ay) = AVA^\top$$

Note: If $V = \sigma^2 I$, then $\text{Var}(Ay) = \sigma^2 AA^\top$

$$5. \quad \mathbb{E}(y^\top Ay) = \text{trace}(AV) + \mu^\top A\mu$$

Note: If $V = \sigma^2 I$, then $\mathbb{E}(y^\top Ay) = \sigma^2 \text{trace}(A) + \mu^\top A\mu$

Expectations of Random Vectors

Let A be a $k \times k$ matrix of constants, a be a $k \times 1$ vector of constants, and y be a $k \times 1$ random vector with mean μ and nonsingular variance–covariance matrix V .

1. $\mathbb{E}(a^\top y) = a^\top \mu$
2. $\mathbb{E}(Ay) = A\mu$
3. $\text{Var}(a^\top y) = a^\top Va$
4. $\text{Var}(Ay) = AVA^\top$

Note: If $V = \sigma^2 I$, then $\text{Var}(Ay) = \sigma^2 AA^\top$

5. $\mathbb{E}(y^\top Ay) = \text{trace}(AV) + \mu^\top A\mu$

Note: If $V = \sigma^2 I$, then $\mathbb{E}(y^\top Ay) = \sigma^2 \text{trace}(A) + \mu^\top A\mu$

Expectations of Random Vectors

Let A be a $k \times k$ matrix of constants, a be a $k \times 1$ vector of constants, and y be a $k \times 1$ random vector with mean μ and nonsingular variance–covariance matrix V .

1. $\mathbb{E}(a^\top y) = a^\top \mu$
2. $\mathbb{E}(Ay) = A\mu$
3. $\text{Var}(a^\top y) = a^\top Va$
4. $\text{Var}(Ay) = AVA^\top$

Note: If $V = \sigma^2 I$, then $\text{Var}(Ay) = \sigma^2 AA^\top$

5. $\mathbb{E}(y^\top Ay) = \text{trace}(AV) + \mu^\top A\mu$

Note: If $V = \sigma^2 I$, then $\mathbb{E}(y^\top Ay) = \sigma^2 \text{trace}(A) + \mu^\top A\mu$

Expectations of Random Vectors

Let A be a $k \times k$ matrix of constants, a be a $k \times 1$ vector of constants, and y be a $k \times 1$ random vector with mean μ and nonsingular variance–covariance matrix V .

1. $\mathbb{E}(a^\top y) = a^\top \mu$
2. $\mathbb{E}(Ay) = A\mu$
3. $\text{Var}(a^\top y) = a^\top Va$
4. $\text{Var}(Ay) = AVA^\top$

Note: If $V = \sigma^2 I$, then $\text{Var}(Ay) = \sigma^2 AA^\top$

5. $\mathbb{E}(y^\top Ay) = \text{trace}(AV) + \mu^\top A\mu$

Note: If $V = \sigma^2 I$, then $\mathbb{E}(y^\top Ay) = \sigma^2 \text{trace}(A) + \mu^\top A\mu$

Applications of Matrix Expectations

- Portfolio variance: For portfolio returns \mathbf{r} with weights \mathbf{w} ,

$$\text{Var}(\mathbf{w}^\top \mathbf{r}) = \mathbf{w}^\top \Sigma \mathbf{w}$$

where Σ is the covariance matrix of returns

- Risk estimation: For quadratic loss functions
- Signal processing: For estimating power in transformed signals
- Econometrics: In GMM and other estimation methods

Applications of Matrix Expectations

- Portfolio variance: For portfolio returns \mathbf{r} with weights \mathbf{w} ,

$$\text{Var}(\mathbf{w}^\top \mathbf{r}) = \mathbf{w}^\top \Sigma \mathbf{w}$$

where Σ is the covariance matrix of returns

- Risk estimation: For quadratic loss functions
- Signal processing: For estimating power in transformed signals
- Econometrics: In GMM and other estimation methods

Applications of Matrix Expectations

- Portfolio variance: For portfolio returns \mathbf{r} with weights \mathbf{w} ,

$$\text{Var}(\mathbf{w}^\top \mathbf{r}) = \mathbf{w}^\top \Sigma \mathbf{w}$$

where Σ is the covariance matrix of returns

- Risk estimation: For quadratic loss functions
- Signal processing: For estimating power in transformed signals
- Econometrics: In GMM and other estimation methods

Applications of Matrix Expectations

- Portfolio variance: For portfolio returns \mathbf{r} with weights \mathbf{w} ,

$$\text{Var}(\mathbf{w}^\top \mathbf{r}) = \mathbf{w}^\top \Sigma \mathbf{w}$$

where Σ is the covariance matrix of returns

- Risk estimation: For quadratic loss functions
- Signal processing: For estimating power in transformed signals
- Econometrics: In GMM and other estimation methods

Contents

- Matrix Operations
- Matrix Derivative and Expectations
- Applications and Wrap-Up

Applications in AI

- Neural networks: Weight matrices and activation functions

$$\mathbf{h}^{(l)} = f(W^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)})$$

- Principal Component Analysis (PCA): Eigendecomposition of covariance matrix

$$\Sigma = Q\Lambda Q^\top$$

- Linear regression: Least squares solution

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

- Support Vector Machines: Quadratic optimization with linear constraints

Further Reading

- Strang, G. (2016). *Introduction to Linear Algebra*
- Boyd, S. & Vandenberghe, L. (2018). *Introduction to Applied Linear Algebra*
- MIT OpenCourseWare: Linear Algebra

Next lecture: Derivative of Neural Network Functions