Maths for ML

Nipun Batra and teaching staff

IIT Gandhinagar

August 22, 2025

Maths for ML

1. Given a vector of $\boldsymbol{\epsilon}$, we can calculate $\sum \epsilon_i^2$ using $\boldsymbol{\epsilon}^T \boldsymbol{\epsilon}$

Maths for ML

1. Given a vector of $\boldsymbol{\epsilon}$, we can calculate $\sum \epsilon_i^2$ using $\boldsymbol{\epsilon}^T \boldsymbol{\epsilon}$

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_N \end{bmatrix}_{N \times 1}$$

Maths for ML

1. Given a vector of $\boldsymbol{\epsilon}$, we can calculate $\sum \epsilon_i^2$ using $\boldsymbol{\epsilon}^T \boldsymbol{\epsilon}$

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_N \end{bmatrix}_{N \times 1}$$

$$\boldsymbol{\epsilon}^T = \begin{bmatrix} \epsilon_1, \epsilon_2, \dots, \epsilon_N \end{bmatrix}_{1 \times N}$$

Maths for ML

1. Given a vector of $\boldsymbol{\epsilon}$, we can calculate $\sum \epsilon_i^2$ using $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_N \end{bmatrix}_{N\times 1}$$

$$\boldsymbol{\epsilon}^T = \begin{bmatrix} \epsilon_1, \epsilon_2, \dots, \epsilon_N \end{bmatrix}_{1\times N}$$

$$\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = \sum \epsilon_i^2$$

Maths for ML

2.

$$(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$$

Maths for ML

2. $$(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$$

3. For a scalar s
$$s = s^T$$

Maths for ML

4. Derivative of a scalar s wrt a vector $\boldsymbol{\theta}$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_N \end{bmatrix}$$

Maths for ML

4. Derivative of a scalar s wrt a vector $\boldsymbol{\theta}$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_N \end{bmatrix}$$

$$\frac{\partial s}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial s}{\partial \theta_1} \\ \frac{\partial s}{\partial \theta_2} \\ \vdots \\ \frac{\partial s}{\partial \theta_N} \end{bmatrix}$$

Linear Functions: Row Vector Times Column Vector

Definition: Setup

Configuration:

- $\boldsymbol{A}^T$ is a row vector ($1 \times n$ matrix)

Linear Functions: Row Vector Times Column Vector

Definition: Setup

Configuration:

- $\boldsymbol{A}^T$ is a row vector ($1 \times n$ matrix)
- $\boldsymbol{\theta}$ is a column vector ($n \times 1$ matrix)

Linear Functions: Row Vector Times Column Vector

Definition: Setup

Configuration:

- $\boldsymbol{A}^T$ is a row vector ($1 \times n$ matrix)
- $\boldsymbol{\theta}$ is a column vector ($n \times 1$ matrix)
- $\boldsymbol{A}^T\boldsymbol{\theta}$ produces a scalar

Linear Functions: Row Vector Times Column Vector

Definition: Setup

Configuration:

- $\boldsymbol{A}^T$ is a row vector ($1 \times n$ matrix)
- $\boldsymbol{\theta}$ is a column vector ($n \times 1$ matrix)
- $\boldsymbol{A}^T\boldsymbol{\theta}$ produces a scalar

Example: Concrete Example

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}_{2\times 1}, \quad \boldsymbol{A}^T = \begin{bmatrix} A_1 & A_2 \end{bmatrix}_{1\times 2}$$

Linear Functions: Row Vector Times Column Vector

Key Points: Matrix Multiplication Result

$$\boldsymbol{A}^T\boldsymbol{\theta} = A_1\theta_1 + A_2\theta_2$$

Linear Functions: Row Vector Times Column Vector

Key Points: Matrix Multiplication Result

$$\boldsymbol{A}^T\boldsymbol{\theta} = A_1\theta_1 + A_2\theta_2$$

This is a scalar! (Linear combination of parameters)

Linear Functions: Row Vector Times Column Vector

Key Points: Matrix Multiplication Result

$$\boldsymbol{A}^T\boldsymbol{\theta} = A_1\theta_1 + A_2\theta_2$$

This is a scalar! (Linear combination of parameters)

Important: ML Relevance

This form appears everywhere in ML:

Linear Functions: Row Vector Times Column Vector

Key Points: Matrix Multiplication Result

$$\boldsymbol{A}^T\boldsymbol{\theta} = A_1\theta_1 + A_2\theta_2$$

This is a scalar! (Linear combination of parameters)

Important: ML Relevance

This form appears everywhere in ML:

- Linear regression: $\boldsymbol{w}^T\boldsymbol{x}$
- Neural networks: $\boldsymbol{w}^T\boldsymbol{h} + b$
- Loss functions: $\boldsymbol{c}^T\boldsymbol{\theta}$

Gradient of Linear Function: Key Result

Key Points: Computing the Gradient

Goal: Find $\frac{\partial \boldsymbol{A}^T\boldsymbol{\theta}}{\partial \boldsymbol{\theta}}$ where $\boldsymbol{A}^T\boldsymbol{\theta} = A_1\theta_1 + A_2\theta_2$

Gradient of Linear Function: Key Result

Key Points: Computing the Gradient

Goal: Find $\frac{\partial \boldsymbol{A}^T\boldsymbol{\theta}}{\partial \boldsymbol{\theta}}$ where $\boldsymbol{A}^T\boldsymbol{\theta} = A_1\theta_1 + A_2\theta_2$

Example: Step-by-Step Calculation

$$\frac{\partial \boldsymbol{A}^T\boldsymbol{\theta}}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial}{\partial \theta_1}(A_1\theta_1 + A_2\theta_2) \\ \frac{\partial}{\partial \theta_2}(A_1\theta_1 + A_2\theta_2) \end{bmatrix}$$

Gradient of Linear Function: Key Result

Key Points: Computing the Gradient

Goal: Find $\frac{\partial \boldsymbol{A}^T\boldsymbol{\theta}}{\partial \boldsymbol{\theta}}$ where $\boldsymbol{A}^T\boldsymbol{\theta} = A_1\theta_1 + A_2\theta_2$

Example: Step-by-Step Calculation

$$\frac{\partial \boldsymbol{A}^T\boldsymbol{\theta}}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial}{\partial \theta_1}(A_1\theta_1 + A_2\theta_2) \\ \frac{\partial}{\partial \theta_2}(A_1\theta_1 + A_2\theta_2) \end{bmatrix}$$

$$= \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}_{2\times 1} = \boldsymbol{A}$$

Gradient of Linear Function: Key Result

Important: Fundamental Rule

$$\boxed{\frac{\partial \boldsymbol{A}^T\boldsymbol{\theta}}{\partial \boldsymbol{\theta}} = \boldsymbol{A}}$$

This is one of the most important rules in ML optimization!

Quadratic Forms and Their Derivatives

Quadratic Forms: Introduction

Definition: Quadratic Form Derivative Rule

Key Result: For matrix $\boldsymbol{Z}$ of form $\boldsymbol{X}^T\boldsymbol{X}$:

Quadratic Forms: Introduction

Definition: Quadratic Form Derivative Rule

Key Result: For matrix $\boldsymbol{Z}$ of form $\boldsymbol{X}^T\boldsymbol{X}$:

$$\frac{\partial}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}^T \boldsymbol{Z} \boldsymbol{\theta}) = 2\boldsymbol{Z}^T \boldsymbol{\theta}$$

Quadratic Forms: Introduction

Definition: Quadratic Form Derivative Rule

Key Result: For matrix $\boldsymbol{Z}$ of form $\boldsymbol{X}^T\boldsymbol{X}$:

$$\frac{\partial}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}^T \boldsymbol{Z} \boldsymbol{\theta}) = 2\boldsymbol{Z}^T\boldsymbol{\theta}$$

Example: Understanding $\boldsymbol{X}^T\boldsymbol{X}$ Matrices

Quadratic Forms: Introduction

Definition: Quadratic Form Derivative Rule

Key Result: For matrix $\boldsymbol{Z}$ of form $\boldsymbol{X}^T\boldsymbol{X}$:

$$\frac{\partial}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}^T \boldsymbol{Z} \boldsymbol{\theta}) = 2\boldsymbol{Z}^T\boldsymbol{\theta}$$

Example: Understanding $\boldsymbol{X}^T\boldsymbol{X}$ Matrices

Starting with:

$$\boldsymbol{X} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad \boldsymbol{X}^T = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

Quadratic Forms: Introduction

Key Points: Computing $Z = X^T X$

Quadratic Forms: Introduction

Key Points: Computing $Z = X^T X$

$$Z = X^T X = \begin{bmatrix} a^2 + c^2 & ab + cd \\ ab + cd & b^2 + d^2 \end{bmatrix}_{2\times 2}$$

Quadratic Forms: Introduction

Key Points: Computing $\boldsymbol{Z} = \boldsymbol{X}^T\boldsymbol{X}$

$$\boldsymbol{Z} = \boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} a^2 + c^2 & ab + cd \\ ab + cd & b^2 + d^2 \end{bmatrix}_{2\times 2}$$

Important: Symmetric Property

Quadratic Forms: Introduction

Key Points: Computing $Z = X^T X$

$$\boldsymbol{Z} = \boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} a^2 + c^2 & ab + cd \\ ab + cd & b^2 + d^2 \end{bmatrix}_{2\times 2}$$

Important: Symmetric Property

Key Observation: $Z_{ij} = Z_{ji} \Rightarrow \boldsymbol{Z}^T = \boldsymbol{Z}$ (symmetric matrix)

Maths for ML

Let

$$\boldsymbol{Z} = \boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} e & f \\ f & g \end{bmatrix}_{2\times 2}$$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}_{2\times 1}$$

Maths for ML

Let

$$\boldsymbol{Z} = \boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} e & f \\ f & g \end{bmatrix}_{2\times 2}$$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}_{2\times 1}$$

$$\boldsymbol{\theta}^T\boldsymbol{Z}\boldsymbol{\theta} = \begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix}_{1\times 2} \begin{bmatrix} e & f \\ f & g \end{bmatrix}_{2\times 2} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}_{2\times 1}$$

Maths for ML

Let

$$\boldsymbol{Z} = \boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} e & f \\ f & g \end{bmatrix}_{2\times 2}$$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}_{2\times 1}$$

$$\boldsymbol{\theta}^T\boldsymbol{Z}\boldsymbol{\theta} = \begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix}_{1\times 2} \begin{bmatrix} e & f \\ f & g \end{bmatrix}_{2\times 2} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}_{2\times 1}$$

$$\boldsymbol{\theta}^T\boldsymbol{Z}\boldsymbol{\theta} = \begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix}_{1\times 2} \begin{bmatrix} e\theta_1 + f\theta_2 \\ f\theta_1 + g\theta_2 \end{bmatrix}_{2\times 1}$$

Maths for ML

Let

$$\boldsymbol{Z} = \boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} e & f \\ f & g \end{bmatrix}_{2\times 2}$$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}_{2\times 1}$$

$$\boldsymbol{\theta}^T\boldsymbol{Z}\boldsymbol{\theta} = \begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix}_{1\times 2} \begin{bmatrix} e & f \\ f & g \end{bmatrix}_{2\times 2} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}_{2\times 1}$$

$$\boldsymbol{\theta}^T\boldsymbol{Z}\boldsymbol{\theta} = \begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix}_{1\times 2} \begin{bmatrix} e\theta_1 + f\theta_2 \\ f\theta_1 + g\theta_2 \end{bmatrix}_{2\times 1}$$

$$\boldsymbol{\theta}^T\boldsymbol{Z}\boldsymbol{\theta} = e\theta_1^2 + 2f\theta_1\theta_2 + g\theta_2^2$$

Maths for ML

Let

$$\boldsymbol{Z} = \boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} e & f \\ f & g \end{bmatrix}_{2\times 2}$$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}_{2\times 1}$$

$$\boldsymbol{\theta}^T\boldsymbol{Z}\boldsymbol{\theta} = \begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix}_{1\times 2} \begin{bmatrix} e & f \\ f & g \end{bmatrix}_{2\times 2} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}_{2\times 1}$$

$$\boldsymbol{\theta}^T\boldsymbol{Z}\boldsymbol{\theta} = \begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix}_{1\times 2} \begin{bmatrix} e\theta_1 + f\theta_2 \\ f\theta_1 + g\theta_2 \end{bmatrix}_{2\times 1}$$

$$\boldsymbol{\theta}^T\boldsymbol{Z}\boldsymbol{\theta} = e\theta_1^2 + 2f\theta_1\theta_2 + g\theta_2^2$$

The term $\boldsymbol{\theta}^T\boldsymbol{Z}\boldsymbol{\theta}$ is a scalar.

Maths for ML

$$\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}^T \boldsymbol{Z} \boldsymbol{\theta} = \frac{\partial}{\partial \theta} (e\theta_1^2 + 2f\theta_1\theta_2 + g\theta_2^2)$$

Maths for ML

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}^T \boldsymbol{Z} \boldsymbol{\theta} &= \frac{\partial}{\partial \theta}(e\theta_1^2 + 2f\theta_1\theta_2 + g\theta_2^2) \\
&= \begin{bmatrix} \frac{\partial}{\partial \theta_1}(e\theta_1^2 + 2f\theta_1\theta_2 + g\theta_2^2) \\ \frac{\partial}{\partial \theta_2}(e\theta_1^2 + 2f\theta_1\theta_2 + g\theta_2^2) \end{bmatrix}
\end{aligned}
$$

Maths for ML

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}^T \boldsymbol{Z} \boldsymbol{\theta} &= \frac{\partial}{\partial \theta}(e\theta_1^2 + 2f\theta_1\theta_2 + g\theta_2^2) \\
&= \begin{bmatrix} \frac{\partial}{\partial \theta_1}(e\theta_1^2 + 2f\theta_1\theta_2 + g\theta_2^2) \\ \frac{\partial}{\partial \theta_2}(e\theta_1^2 + 2f\theta_1\theta_2 + g\theta_2^2) \end{bmatrix} \\
&= \begin{bmatrix} 2e\theta_1 + 2f\theta_2 \\ 2f\theta_1 + 2g\theta_2 \end{bmatrix} = 2 \begin{bmatrix} e & f \\ f & g \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_1 \end{bmatrix}
\end{aligned}
$$

Maths for ML

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}^T \boldsymbol{Z} \boldsymbol{\theta} &= \frac{\partial}{\partial \theta}(e\theta_1^2 + 2f\theta_1\theta_2 + g\theta_2^2) \\
&= \begin{bmatrix} \frac{\partial}{\partial \theta_1}(e\theta_1^2 + 2f\theta_1\theta_2 + g\theta_2^2) \\ \frac{\partial}{\partial \theta_2}(e\theta_1^2 + 2f\theta_1\theta_2 + g\theta_2^2) \end{bmatrix} \\
&= \begin{bmatrix} 2e\theta_1 + 2f\theta_2 \\ 2f\theta_1 + 2g\theta_2 \end{bmatrix} = 2 \begin{bmatrix} e & f \\ f & g \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_1 \end{bmatrix} \\
&= 2\boldsymbol{Z}\boldsymbol{\theta} = 2\boldsymbol{Z}^T\boldsymbol{\theta}
\end{aligned}
$$

Matrix Rank and Invertibility

Matrix Rank: Fundamental Concept

Definition: What is Matrix Rank?

Matrix Rank: Fundamental Concept

Definition: What is Matrix Rank?

Rank = Maximum number of linearly independent rows (or columns)

Matrix Rank: Fundamental Concept

Definition: What is Matrix Rank?

Rank $=$ Maximum number of linearly independent rows (or columns)

Key Points: Two Equivalent Perspectives

For an $r \times c$ matrix:

Matrix Rank: Fundamental Concept

Definition: What is Matrix Rank?

Rank = Maximum number of linearly independent rows (or columns)

Key Points: Two Equivalent Perspectives

For an $r \times c$ matrix:

- **Row perspective:** r row vectors, each with c elements

Matrix Rank: Fundamental Concept

Definition: What is Matrix Rank?

Rank = Maximum number of linearly independent rows (or columns)

Key Points: Two Equivalent Perspectives

For an $r \times c$ matrix:

- **Row perspective:** r row vectors, each with c elements
- **Column perspective:** c column vectors, each with r elements

Matrix Rank: Fundamental Concept

Example: Maximum Rank Rules

- If $r < c$: Maximum rank $= r$ (more columns than rows)

Matrix Rank: Fundamental Concept

Example: Maximum Rank Rules

- If $r < c$: Maximum rank $= r$ (more columns than rows)
- If $r > c$: Maximum rank $= c$ (more rows than columns)

Matrix Rank: Fundamental Concept

Example: Maximum Rank Rules

- If $r < c$: Maximum rank $= r$ (more columns than rows)
- If $r > c$: Maximum rank $= c$ (more rows than columns)
- If $r = c$: Maximum rank $= r = c$ (square matrix)

Maths for ML: Matrix Rank

- Given a matrix $\boldsymbol{A}$:

$$\begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 7 & 8 \end{bmatrix}$$

Maths for ML: Matrix Rank

- Given a matrix $\boldsymbol{A}$:

$$\begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 7 & 8 \end{bmatrix}$$

- What is the rank?

Maths for ML: Matrix Rank

- Given a matrix $\boldsymbol{A}$:

$$\begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 7 & 8 \end{bmatrix}$$

- What is the rank?
- $r = c = 3$. Thus, rank is $<= 3$

Maths for ML: Matrix Rank

- Given a matrix $\boldsymbol{A}$:

$$\begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 7 & 8 \end{bmatrix}$$

- What is the rank?
- $r = c = 3$. Thus, rank is $<= 3$
- $\text{Row}(3) = 3 \times \text{Row}(1) + 2 \times \text{Row}(2)$.

Maths for ML: Matrix Rank

- Given a matrix $\boldsymbol{A}$:

$$\begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 7 & 8 \end{bmatrix}$$

- What is the rank?
- $r = c = 3$. Thus, rank is $<= 3$
- $\mathrm{Row}(3) = 3 \times \mathrm{Row}(1) + 2 \times \mathrm{Row}(2)$.
- Thus, $\mathrm{Row}(3)$ is linearly dependent on $\mathrm{Row}(1)$ and $\mathrm{Row}(2)$.

Maths for ML: Matrix Rank

- Given a matrix $\boldsymbol{A}$:

$$\begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 7 & 8 \end{bmatrix}$$

- What is the rank?
- $r = c = 3$. Thus, rank is $<= 3$
- $\mathrm{Row}(3) = 3 \times \mathrm{Row}(1) + 2 \times \mathrm{Row}(2)$.
- Thus, $\mathrm{Row}(3)$ is linearly dependent on $\mathrm{Row}(1)$ and $\mathrm{Row}(2)$.
- rank($\boldsymbol{A}$)$=2$

Maths for ML: Matrix Rank

What is the rank of

$$\boldsymbol{X} = \begin{bmatrix} 1 & 2 & 4 & 4 \\ 3 & 4 & 8 & 0 \end{bmatrix}$$

Maths for ML: Matrix Rank

What is the rank of

$$\boldsymbol{X} = \begin{bmatrix} 1 & 2 & 4 & 4 \\ 3 & 4 & 8 & 0 \end{bmatrix}$$

Since $\boldsymbol{X}$ has fewer rows than columns, its maximum rank is equal to the maximum number of linearly independent rows. And because neither row is linearly dependent on the other row, the matrix has 2 linearly independent rows; so its rank is 2.

Pop Quiz #1

Answer this!

What is the rank of a 3×3 matrix $\boldsymbol{A}$ formed by the outer product of two non-zero vectors, $\boldsymbol{u}$ (3×1) and $\boldsymbol{v}^T$ (1×3)?

$$\boldsymbol{A} = \boldsymbol{u}\boldsymbol{v}^T = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix}$$

A) 0

B) 1

C) 2

D) 3

Pop Quiz #1

Answer this!

What is the rank of a 3×3 matrix $\boldsymbol{A}$ formed by the outer product of two non-zero vectors, $\boldsymbol{u}$ (3×1) and $\boldsymbol{v}^T$ (1×3)?

$$\boldsymbol{A} = \boldsymbol{u}\boldsymbol{v}^T = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix}$$

A) 0

B) 1

C) 2

D) 3

Answer: B) 1

Maths for ML: Rank of an Outer Product

Key Points: Matrix Formation

First, let's construct the matrix $\boldsymbol{A} = \boldsymbol{u}\boldsymbol{v}^T$:

$$\boldsymbol{A} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix}$$

Maths for ML: Rank of an Outer Product

Key Points: Matrix Formation

First, let's construct the matrix $\boldsymbol{A} = \boldsymbol{u}\boldsymbol{v}^T$:

$$\boldsymbol{A} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} = \begin{bmatrix} u_1 v_1 & u_1 v_2 & u_1 v_3 \\ u_2 v_1 & u_2 v_2 & u_2 v_3 \\ u_3 v_1 & u_3 v_2 & u_3 v_3 \end{bmatrix}$$

Maths for ML: Rank of an Outer Product

- **Look at the columns:** Each column is just a scalar multiple of the original vector $\boldsymbol{u}$.

$$\text{Column } 1 = v_1 \boldsymbol{u}, \quad \text{Column } 2 = v_2 \boldsymbol{u}, \quad \text{Column } 3 = v_3 \boldsymbol{u}$$

Maths for ML: Rank of an Outer Product

- **Look at the columns:** Each column is just a scalar multiple of the original vector $\boldsymbol{u}$.

$$\text{Column } 1 = v_1 \boldsymbol{u}, \quad \text{Column } 2 = v_2 \boldsymbol{u}, \quad \text{Column } 3 = v_3 \boldsymbol{u}$$

- **Look at the rows:** Similarly, each row is a scalar multiple of the original vector $\boldsymbol{v}^T$.

$$\text{Row } 1 = u_1 \boldsymbol{v}^T, \quad \text{Row } 2 = u_2 \boldsymbol{v}^T, \quad \text{Row } 3 = u_3 \boldsymbol{v}^T$$

Maths for ML: Rank of an Outer Product

- **Look at the columns:** Each column is just a scalar multiple of the original vector $\boldsymbol{u}$.

$$\text{Column 1} = v_1 \boldsymbol{u}, \quad \text{Column 2} = v_2 \boldsymbol{u}, \quad \text{Column 3} = v_3 \boldsymbol{u}$$

- **Look at the rows:** Similarly, each row is a scalar multiple of the original vector $\boldsymbol{v}^T$.

$$\text{Row 1} = u_1 \boldsymbol{v}^T, \quad \text{Row 2} = u_2 \boldsymbol{v}^T, \quad \text{Row 3} = u_3 \boldsymbol{v}^T$$

Important: Conclusion

Since all rows and columns are linearly dependent on a single vector, the maximum number of linearly independent rows (or columns) is one. Therefore, the rank of the matrix is **1**.

Maths for ML: Matrix Inverse

Suppose $\boldsymbol{A}$ is an $n \times n$ matrix. The inverse of $\boldsymbol{A}$ is another $n \times n$ matrix, denoted $\boldsymbol{A}^{-1}$, that satisfies the following conditions.

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_n$$

where $\boldsymbol{I}_n$ is the identity matrix.

Maths for ML: Matrix Inverse

There are two ways to determine whether the inverse of a square matrix exists.

- If the rank of an $n \times n$ matrix is less than n, the matrix does not have an inverse.

Maths for ML: Matrix Inverse

There are two ways to determine whether the inverse of a square matrix exists.

- If the rank of an $n \times n$ matrix is less than n, the matrix does not have an inverse.
- When the determinant for a square matrix is equal to zero, the inverse for that matrix does not exist.

Maths for ML: Matrix Inverse

There are two ways to determine whether the inverse of a square matrix exists.

- If the rank of an $n \times n$ matrix is less than n, the matrix does not have an inverse.
- When the determinant for a square matrix is equal to zero, the inverse for that matrix does not exist.

A square matrix that has an inverse is said to be nonsingular or invertible; a square matrix that does not have an inverse is said to be singular.

Maths for ML: Matrix Inverse

There are two ways to determine whether the inverse of a square matrix exists.

- If the rank of an $n \times n$ matrix is less than n, the matrix does not have an inverse.
- When the determinant for a square matrix is equal to zero, the inverse for that matrix does not exist.

A square matrix that has an inverse is said to be nonsingular or invertible; a square matrix that does not have an inverse is said to be singular.
Not every square matrix has an inverse; but if a matrix does have an inverse, it is unique.

Generalizing Derivatives: Gradients and Jacobians

Derivatives of $\mathbb{R}^n \to \mathbb{R}$: The Gradient

Definition: Recap: Derivative of a Scalar Function

For a function $f : \mathbb{R}^n \to \mathbb{R}$ that takes a vector $\boldsymbol{\theta} \in \mathbb{R}^n$ and returns a scalar, its derivative is the **gradient**.

Derivatives of $\mathbb{R}^n \to \mathbb{R}$: The Gradient

Definition: Recap: Derivative of a Scalar Function

For a function $f : \mathbb{R}^n \to \mathbb{R}$ that takes a vector $\boldsymbol{\theta} \in \mathbb{R}^n$ and returns a scalar, its derivative is the **gradient**.

$$\nabla f(\boldsymbol{\theta}) = \frac{\partial f}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial f}{\partial \theta_1} \\ \frac{\partial f}{\partial \theta_2} \\ \vdots \\ \frac{\partial f}{\partial \theta_n} \end{bmatrix}_{n \times 1}$$

Derivatives of $\mathbb{R}^n \to \mathbb{R}$: The Gradient

Definition: Recap: Derivative of a Scalar Function

For a function $f : \mathbb{R}^n \to \mathbb{R}$ that takes a vector $\boldsymbol{\theta} \in \mathbb{R}^n$ and returns a scalar, its derivative is the **gradient**.

$$\nabla f(\boldsymbol{\theta}) = \frac{\partial f}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial f}{\partial \theta_1} \\ \frac{\partial f}{\partial \theta_2} \\ \vdots \\ \frac{\partial f}{\partial \theta_n} \end{bmatrix}_{n \times 1}$$

Note: By convention in ML, the gradient is a column vector.

Derivatives of $\mathbb{R}^n \to \mathbb{R}$: The Gradient

Definition: Recap: Derivative of a Scalar Function

For a function $f : \mathbb{R}^n \to \mathbb{R}$ that takes a vector $\boldsymbol{\theta} \in \mathbb{R}^n$ and returns a scalar, its derivative is the **gradient**.

$$\nabla f(\boldsymbol{\theta}) = \frac{\partial f}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial f}{\partial \theta_1} \\ \frac{\partial f}{\partial \theta_2} \\ \vdots \\ \frac{\partial f}{\partial \theta_n} \end{bmatrix}_{n \times 1}$$

Note: By convention in ML, the gradient is a column vector.

Important: Geometric Intuition

The gradient vector $\nabla f(\boldsymbol{\theta})$ points in the direction of the **steepest ascent** of the function f at point $\boldsymbol{\theta}$. The magnitude $||\nabla f(\boldsymbol{\theta})||$ gives the rate of that increase.

From $\mathbb{R}^n \to \mathbb{R}$ to $\mathbb{R}^n \to \mathbb{R}^m$

Important: Handling Vector Outputs

What if our function takes a vector and also *outputs* a vector?

From $\mathbb{R}^n \to \mathbb{R}$ to $\mathbb{R}^n \to \mathbb{R}^m$

Important: Handling Vector Outputs

What if our function takes a vector and also *outputs* a vector?

Let $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$. We can think of $\boldsymbol{f}$ as a stack of m scalar-valued functions:

$$\boldsymbol{f}(\boldsymbol{\theta}) = \begin{bmatrix} f_1(\boldsymbol{\theta}) \\ f_2(\boldsymbol{\theta}) \\ \vdots \\ f_m(\boldsymbol{\theta}) \end{bmatrix}_{m\times 1}$$

.

From $\mathbb{R}^n \to \mathbb{R}$ to $\mathbb{R}^n \to \mathbb{R}^m$

Important: Handling Vector Outputs

What if our function takes a vector and also *outputs* a vector?

Let $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$. We can think of $\boldsymbol{f}$ as a stack of m scalar-valued functions:

$$\boldsymbol{f}(\boldsymbol{\theta}) = \begin{bmatrix} f_1(\boldsymbol{\theta}) \\ f_2(\boldsymbol{\theta}) \\ \vdots \\ f_m(\boldsymbol{\theta}) \end{bmatrix}_{m \times 1}$$

Question: How do we differentiate $\boldsymbol{f}$ with respect to $\boldsymbol{\theta}$?

From $\mathbb{R}^n \to \mathbb{R}$ to $\mathbb{R}^n \to \mathbb{R}^m$

Important: Handling Vector Outputs

What if our function takes a vector and also *outputs* a vector?

Let $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$. We can think of $\boldsymbol{f}$ as a stack of m scalar-valued functions:

$$\boldsymbol{f}(\boldsymbol{\theta}) = \begin{bmatrix} f_1(\boldsymbol{\theta}) \\ f_2(\boldsymbol{\theta}) \\ \vdots \\ f_m(\boldsymbol{\theta}) \end{bmatrix}_{m\times 1}$$

Question: How do we differentiate $\boldsymbol{f}$ with respect to $\boldsymbol{\theta}$? We need to track how **every output** changes with respect to **every input**.

The Jacobian Matrix

Definition: The Derivative of a Vector Function

The derivative of $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ is the **Jacobian matrix $\boldsymbol{J}$**, an $m \times n$ matrix of all first-order partial derivatives.

The Jacobian Matrix

Definition: The Derivative of a Vector Function

The derivative of $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ is the **Jacobian matrix** $\boldsymbol{J}$, an $m \times n$ matrix of all first-order partial derivatives.

$$\boldsymbol{J} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial f_1}{\partial \theta_1} & \frac{\partial f_1}{\partial \theta_2} & \cdots & \frac{\partial f_1}{\partial \theta_n} \\ \frac{\partial f_2}{\partial \theta_1} & \frac{\partial f_2}{\partial \theta_2} & \cdots & \frac{\partial f_2}{\partial \theta_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial \theta_1} & \frac{\partial f_m}{\partial \theta_2} & \cdots & \frac{\partial f_m}{\partial \theta_n} \end{bmatrix}_{m \times n}$$

The Jacobian Matrix

Definition: The Derivative of a Vector Function

The derivative of $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ is the **Jacobian matrix** $\boldsymbol{J}$, an $m \times n$ matrix of all first-order partial derivatives.

$$\boldsymbol{J} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial f_1}{\partial \theta_1} & \frac{\partial f_1}{\partial \theta_2} & \cdots & \frac{\partial f_1}{\partial \theta_n} \\ \frac{\partial f_2}{\partial \theta_1} & \frac{\partial f_2}{\partial \theta_2} & \cdots & \frac{\partial f_2}{\partial \theta_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial \theta_1} & \frac{\partial f_m}{\partial \theta_2} & \cdots & \frac{\partial f_m}{\partial \theta_n} \end{bmatrix}_{m \times n}$$

Key Structure: Row i of the Jacobian is the transpose of the gradient of the i-th output function, f_i.

$$(\boldsymbol{J})_{[i,:]} = \left(\nabla f_i(\boldsymbol{\theta})\right)^T$$

Jacobian: A Concrete Example

Example: Let's Compute a Jacobian

Consider $\boldsymbol{f}\colon \mathbb{R}^2 \to \mathbb{R}^2$ with $\boldsymbol{\theta} = [\theta_1, \theta_2]^T$.

$$\boldsymbol{f}(\boldsymbol{\theta}) = \begin{bmatrix} f_1(\theta_1, \theta_2) \\ f_2(\theta_1, \theta_2) \end{bmatrix} = \begin{bmatrix} \theta_1^2 \theta_2 \\ 5\theta_1 + \sin(\theta_2) \end{bmatrix}$$

Jacobian: A Concrete Example

Example: Let's Compute a Jacobian

Consider $\boldsymbol{f}\colon \mathbb{R}^2 \to \mathbb{R}^2$ with $\boldsymbol{\theta} = [\theta_1, \theta_2]^T$.

$$\boldsymbol{f}(\boldsymbol{\theta}) = \begin{bmatrix} f_1(\theta_1, \theta_2) \\ f_2(\theta_1, \theta_2) \end{bmatrix} = \begin{bmatrix} \theta_1^2 \theta_2 \\ 5\theta_1 + \sin(\theta_2) \end{bmatrix}$$

The Jacobian $\boldsymbol{J}$ will be a 2×2 matrix.

$$\boldsymbol{J} = \begin{bmatrix} \frac{\partial f_1}{\partial \theta_1} & \frac{\partial f_1}{\partial \theta_2} \\ \frac{\partial f_2}{\partial \theta_1} & \frac{\partial f_2}{\partial \theta_2} \end{bmatrix}$$

Jacobian: A Concrete Example

Example: Let's Compute a Jacobian

Consider $\boldsymbol{f}\colon \mathbb{R}^2 \to \mathbb{R}^2$ with $\boldsymbol{\theta} = [\theta_1, \theta_2]^T$.

$$\boldsymbol{f}(\boldsymbol{\theta}) = \begin{bmatrix} f_1(\theta_1, \theta_2) \\ f_2(\theta_1, \theta_2) \end{bmatrix} = \begin{bmatrix} \theta_1^2 \theta_2 \\ 5\theta_1 + \sin(\theta_2) \end{bmatrix}$$

The Jacobian $\boldsymbol{J}$ will be a 2×2 matrix.

$$\boldsymbol{J} = \begin{bmatrix} \frac{\partial f_1}{\partial \theta_1} & \frac{\partial f_1}{\partial \theta_2} \\ \frac{\partial f_2}{\partial \theta_1} & \frac{\partial f_2}{\partial \theta_2} \end{bmatrix}$$

$$\boldsymbol{J} = \begin{bmatrix} 2\theta_1\theta_2 & \theta_1^2 \\ 5 & \cos(\theta_2) \end{bmatrix}$$

Visualizing Functions: Graphs and Level Sets

Definition: The Graph of a Function

For a function $f: \mathbb{R}^n \to \mathbb{R}$, its **graph** is the set of all input-output pairs.

$$\text{Graph}(f) = \{(\boldsymbol{\theta}, f(\boldsymbol{\theta})) \mid \boldsymbol{\theta} \in \mathbb{R}^n\}$$

Visualizing Functions: Graphs and Level Sets

Definition: The Graph of a Function

For a function $f\colon \mathbb{R}^n \to \mathbb{R}$, its **graph** is the set of all input-output pairs.

$$\text{Graph}(f) = \{(\boldsymbol{\theta}, f(\boldsymbol{\theta})) \mid \boldsymbol{\theta} \in \mathbb{R}^n\}$$

This graph lives in one higher dimension, $\mathbb{R}^{n+1}$.
Example: For $f\colon \mathbb{R}^2 \to \mathbb{R}$, the graph is a surface in 3D space.

Visualizing Functions: Graphs and Level Sets

Definition: The Graph of a Function

For a function $f \colon \mathbb{R}^n \to \mathbb{R}$, its **graph** is the set of all input-output pairs.

$$\text{Graph}(f) = \{(\boldsymbol{\theta}, f(\boldsymbol{\theta})) \mid \boldsymbol{\theta} \in \mathbb{R}^n\}$$

This graph lives in one higher dimension, $\mathbb{R}^{n+1}$.
Example: For $f \colon \mathbb{R}^2 \to \mathbb{R}$, the graph is a surface in 3D space.

Definition: Level Sets (Contours)

A **level set** (or contour) is the set of all points in the *domain* where the function's output is a constant value, c.

$$\text{Level Set}_c = \{\boldsymbol{\theta} \in \mathbb{R}^n \mid f(\boldsymbol{\theta}) = c\}$$

Visualizing Functions: Graphs and Level Sets

Definition: The Graph of a Function

For a function $f\colon \mathbb{R}^n \to \mathbb{R}$, its **graph** is the set of all input-output pairs.

$$\text{Graph}(f) = \{(\boldsymbol{\theta}, f(\boldsymbol{\theta})) \mid \boldsymbol{\theta} \in \mathbb{R}^n\}$$

This graph lives in one higher dimension, $\mathbb{R}^{n+1}$.
Example: For $f\colon \mathbb{R}^2 \to \mathbb{R}$, the graph is a surface in 3D space.

Definition: Level Sets (Contours)

A **level set** (or contour) is the set of all points in the *domain* where the function's output is a constant value, c.

$$\text{Level Set}_c = \{\boldsymbol{\theta} \in \mathbb{R}^n \mid f(\boldsymbol{\theta}) = c\}$$

This set lives in the original domain space, $\mathbb{R}^n$.

Surplus - Directional Derivative

Important: Why the Gradient is the Steepest Direction

- Let's define a line through $\boldsymbol{x} \in \mathbb{R}^n$ on the function $f : \mathbb{R}^n \to \mathbb{R}$ in direction $\boldsymbol{v}$ as $\boldsymbol{c}(t) = \boldsymbol{x} + t\boldsymbol{v}$. The rate of change of f along this line is $\frac{d}{dt} f(\boldsymbol{c}(t))$.
- Using the chain rule, this derivative is $\nabla f(\boldsymbol{c}(t)) \cdot \boldsymbol{c}'(t)$. At our point $\boldsymbol{x}$ (where $t = 0$), this becomes $\nabla f(\boldsymbol{x}) \cdot \boldsymbol{v}$.
- From geometry, we know $\nabla f(\boldsymbol{x}) \cdot \boldsymbol{v} = ||\nabla f(\boldsymbol{x})|| \, ||\boldsymbol{v}|| \cos(\theta)$. Since $||\boldsymbol{v}|| = 1$, this value is maximized when $\cos(\theta) = 1$.
- This occurs when $\boldsymbol{v}$ points in the same direction as $\nabla f(\boldsymbol{x})$. Thus, the gradient points in the direction of steepest ascent.