Newton School of Technology

Deep Learning Foundations

7

Newton School of Technology

Join the lecture online on your dashboard.

Newton School of Technology

Let’s start with a minute of silence.

Newton School of Technology

आचार्यात् पादं आधत्ते पादं शिष्यः स्वमेधया ।
पादं सब्रह्मचारिभ्यः पादं कालक्रमेण च ॥

Meaning:

A student acquires knowledge in four equal parts:

- One-fourth from the teacher
- One-fourth through self-reflection and independent thinking
- One-fourth through discussions with peers and fellow learners
- One-fourth over time through personal experience

Concept Recap

Newton School of Technology

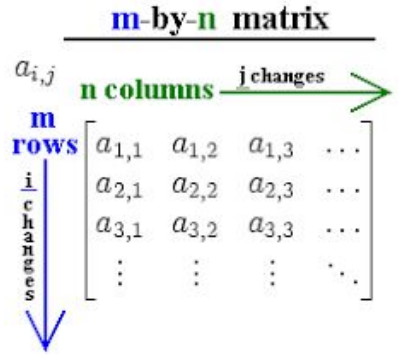
Foundation of Deep Learning:

- Linear Algebra Basics
- Scalars
- Vectors
- Tensors
- Matrices and Operations
- Norms

Linear Algebra for Deep Learning – Lecture 2

Newton School of Technology

Matrix Structure • Decompositions • Eigenvalues • SVD • Applications in ML



Will Cover:

- How matrices encode structure
- Why decompositions matter
- How PCA/SVD compress and denoise data
- How linear algebra reveals geometry behind deep learning

Rank

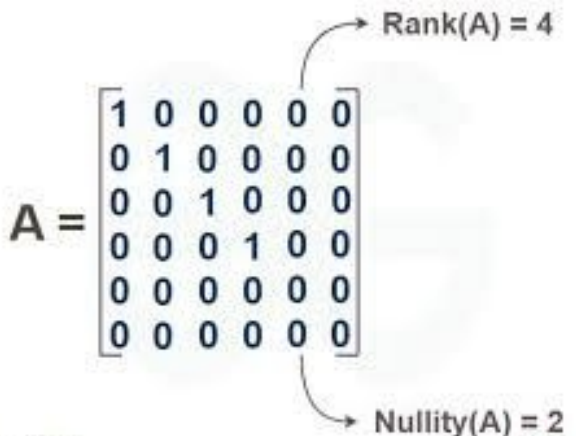
Newton School of Technology

Definition:

The rank of a matrix is the number of linearly independent rows or columns.

Rank = “how much unique information” the matrix contains.

- High rank → more information, more dimensions
- Low rank → redundancy, compression possible



ML Interpretation

- Rank tells you how many features matter.
- Many real ML datasets are low-rank due to correlations.
- SVD/PCA use this fact for dimensionality reduction.

Rank

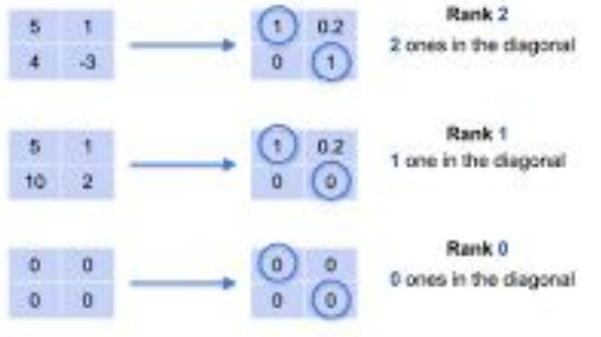
Newton School of Technology

Example

Matrix A = [1 2 3]
[2 4 6]

Row 2 = 2 × Row 1 → rank = 1

Only one direction of information.



Practice Question

If a matrix is 100×100 but rank is only 5, what does that mean?

Column Space, Row Space & Null Space

Newton School of Technology

Column Space (C(A)):

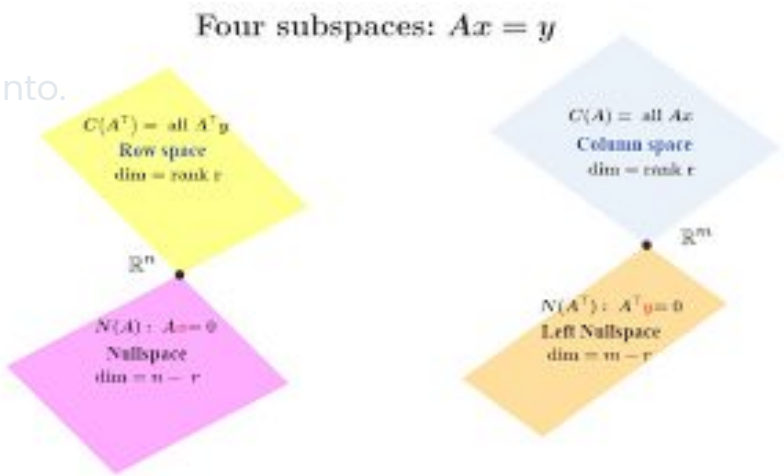
- All possible outputs of Ax.
- Represents the directions a matrix can transform vectors into.

Row Space:

- Describes constraints on x in Ax = b.

Null Space (N(A)):

- All vectors x such that Ax = 0.
- Represents directions the matrix kills (maps to zero).

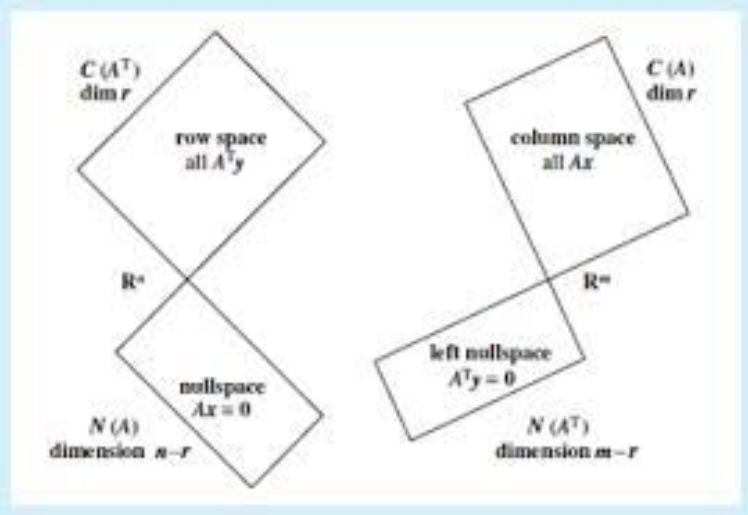


Column Space, Row Space & Null Space

Newton School of Technology

Think of the matrix as a teacher:

- Some concepts it emphasizes → column space
- Some constraints it imposes → row space
- Some mistakes it ignores → null space



Why It Matters in ML

- Null space reveals redundant features
- Column space determines representational capacity
- Row space shows restrictions on outputs
- Used in solving linear systems for optimization

Practice Question

What does a large null space imply about the matrix?

Linear Independence & Basis

Newton School of Technology

Linear Independence

A set of vectors is linearly independent if none of them can be written as a combination of the others.

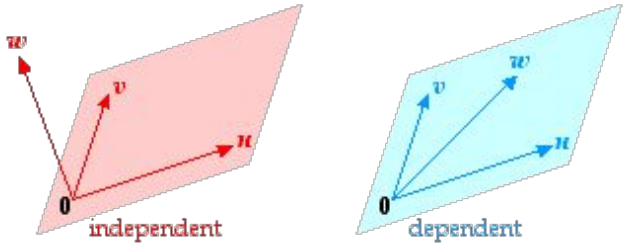
Why It Matters

ML data often has:

- Redundant features (dependent vectors)
- Only a few true underlying dimensions (independent ones)

Basis

A minimal set of independent vectors that spans a space.



ML Interpretation

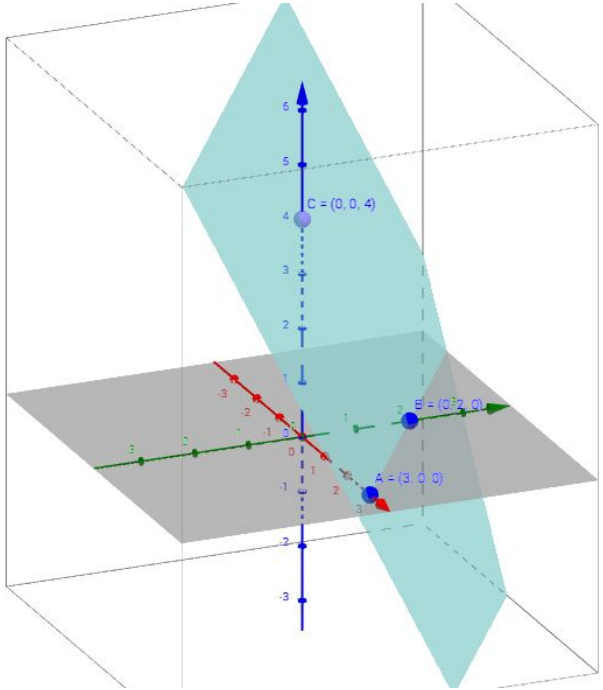
Basis vectors = “core directions” of your data.

8

Span

The Span of a set of vectors is the set of all possible linear combinations you can create with them.

- Two non-parallel vectors in 3D span a 2D plane.
- Three independent vectors in 3D span the entire space.
- If a vector y is in the span of {v_1, ..., $v_□$}, then y can be reached using those vectors.



Linear Dependence

Newton School of Technology

Linearly Dependent

A set of vectors is dependent if at least one vector can be written as a linear combination of the others. It adds no "new" information or directions to the span.

Linearly Independent

No vector in the set can be formed by combining the others. Every vector provides a unique, necessary direction to define the space.

$$c_1 v_1 + c_2 v_2 + \cdots + c_n v_n = 0$$

Dependent if there exists a non-trivial solution (not all coefficients are zero).

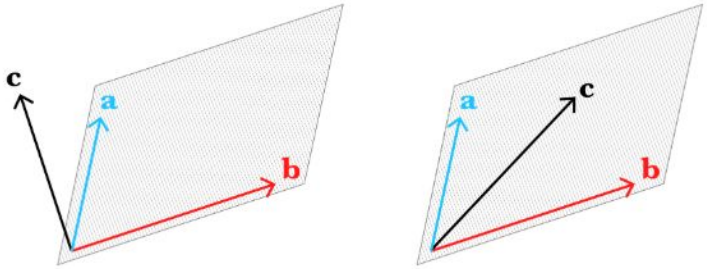
Visualizing Redundancy

Newton School of Technology

In Machine Learning, linear dependence is synonymous with redundant features.

- If feature A is just double feature B, it doesn't help the model learn anything new—it just takes up memory and can destabilize calculations.

"Independence is information efficiency."



Span and Matrix Rank

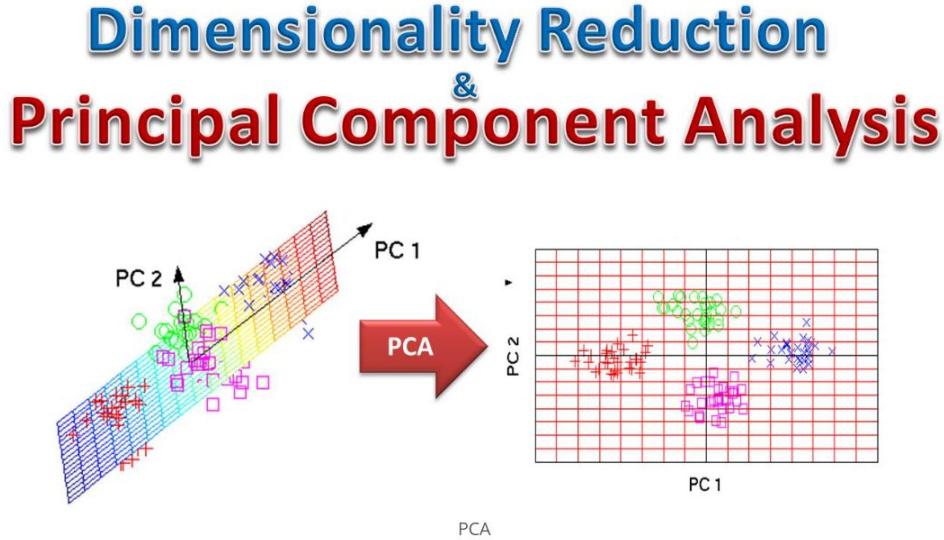
Newton School of Technology

Concept	Definition	ML Interpretation
Column Space	The span of the columns of a matrix.	All possible outputs of a model layer.
Rank	The number of linearly independent columns.	The true "intrinsic" dimensionality of data.
Singularity	When columns are dependent (Determinant = 0).	Information loss or redundant layers.

Example:
Principal Component Analysis

PCA uses these concepts to simplify data:

- Identify Span: Find the subspace where data has the most variance.
- Remove Dependence: Discard features that are highly correlated (dependent).
- Efficiency: Lower the dimensionality without losing major information.



PCA

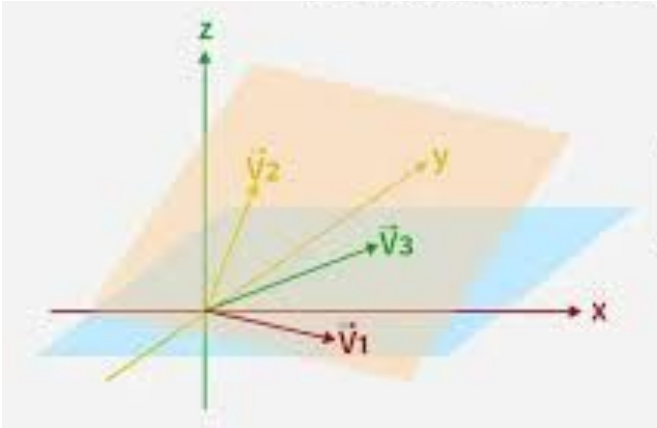
Practice Question

Newton School of Technology

Which of these sets is linearly independent?

A. {(1,2), (2,4)}

B. {(1,2), (2,3)}



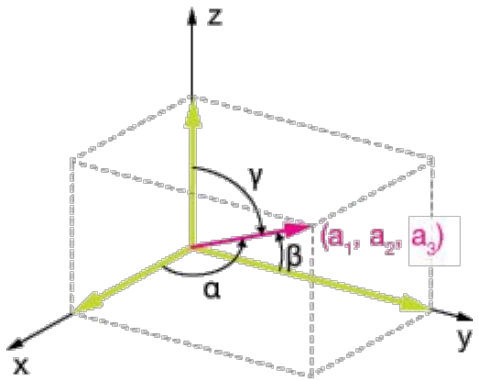
Orthogonality

Two vectors are orthogonal if:

$$x \cdot y = 0$$

Geometric Meaning:

They form a 90° angle → no projection of one onto the other.



Why ML Cares:

- Orthogonal embeddings reduce feature overlap
- Decorrelated features improve model stability
- Orthogonal weight matrices improve gradient flow
- PCA produces orthogonal principal components

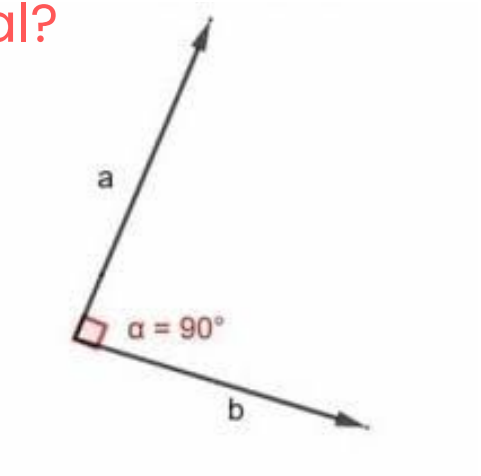
Example

(1, 2) and (2, −1) → dot product = 0 → orthogonal.

Practice Question

Newton School of Technology

If x = [3, 4] and y = [4, −3], are they orthogonal?



Projection Of a Vector

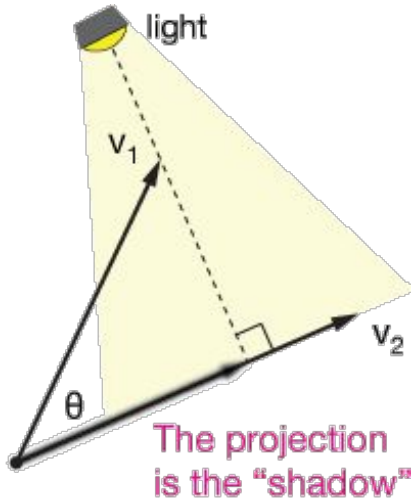
Newton School of Technology

Projection of vector b onto a:

$$\text{proj}_a(b) = (a \cdot b / a \cdot a)\, a$$

Projection = shadow of one vector onto another direction.

Analogy:
Shining a flashlight on an object → shadow cast onto a surface = projection.



When Does Projection Matter?

- When you want to isolate the useful direction
- When removing irrelevant components
- When reducing dimensions

Practice Question

Newton School of Technology

If vector b lies entirely in the null space of a matrix A, what is its projection onto the column space?

Eigenvalues & Eigenvectors

Newton School of Technology

For matrix A:

$$A v = \lambda v$$

Where:

- v = eigenvector
- λ = eigenvalue

Eigenvectors are directions that A stretches/compresses.

- Eigenvalues are how much it stretches/compresses.

Geometric Meaning

- Matrix transformation → one direction stays the same, only its length changes.

ML Interpretation:

Eigenvectors show:

- Main directions of variation in data (PCA)
- Stable directions in optimization
- Important components in embeddings

9

Practice Question

Newton School of Technology

If A v = 5v, what does eigenvalue 5 mean?

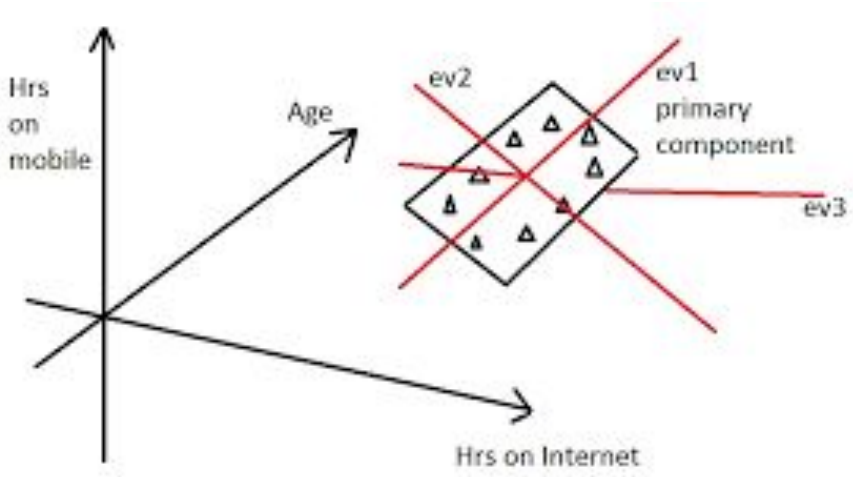
Eigenvalue Intuition

Newton School of Technology

Simplified Story

Imagine a matrix A is a “machine” that transforms

- For most vectors → direction changes
- For special vectors → direction stays the same
- Only length changes



These special vectors are eigenvectors

The amount of stretching = eigenvalue

Analogy:

Think of A as a filter:

- It changes most images
- But some images remain in the same “style” after filtering → these are eigenvectors

Deep Learning Uses

- PCA → find directions with maximum variance
- Spectral clustering → use eigenvectors of Laplacian
- Stability in RNNs → eigenvalues determine exploding/vanishing behavior

1
0
1

Practice Question

Newton School of Technology

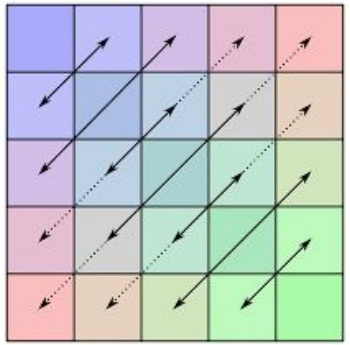
Why do eigenvectors matter in PCA?

Symmetric Matrices

Newton School of Technology

A matrix A is symmetric if:

$$A^T = A$$



Why Symmetric Matrices Matter

- All eigenvalues are real
- Eigenvectors are orthogonal
- Decomposition is stable & interpretable

ML Applications

- Covariance matrix ($\Sigma = X^T X$) → symmetric
- PCA uses covariance matrix
- Many optimization problems use symmetric matrices
- Graph Laplacians in GNNs are symmetric

Symmetric matrices are the nicest matrices to work with:

- Clear geometric interpretation
- Diagonalizable

Practice Question

Newton School of Technology

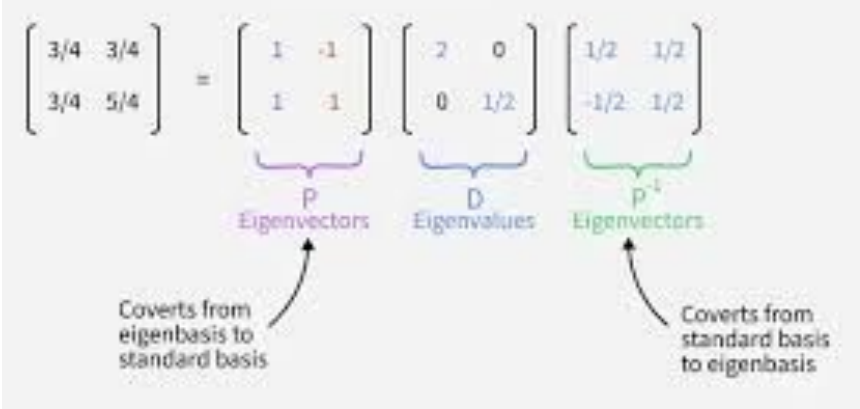
Why does PCA always use a symmetric matrix?

Diagonalization

Newton School of Technology

A square matrix A ∈ R^{n×n} is diagonalizable if:

$$A = PDP^{-1}$$



where:

$$P = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}$$

(eigenvectors)

$$D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

Diagonalization

Newton School of Technology

Mathematical Significance

- Converts a complex matrix into a simple scaling matrix
- Eases computation of:
  - A^K
  - e^A
  - Matrix exponentials
  - Stability analysis in dynamical systems
- Allows easier interpretation of linear transformations

Condition:

A matrix is diagonalizable if it has linearly n independent eigenvectors.

Intuition Behind Diagonalization

Newton School of Technology

Diagonalization finds a new coordinate system where a matrix acts as pure scaling rather than mixing components.

This is expressed through the eigenvector relation:

$$Av_i = \lambda_i v_i$$

This means:

- Each eigenvector v_i points in a direction that the matrix does not rotate or distort
- The transformation only stretches or shrinks that direction by λ_i
- The matrix becomes easier to understand in a basis formed by its eigenvectors

Geometric Interpretation

- A matrix normally mixes coordinates (shearing, rotating, stretching)
- Diagonalization “untangles” the transformation into independent dimensions
- In the eigenvector basis:
  - No mixing

Singular Value Decomposition (SVD)

Newton School of Technology

Formal Definition

For any matrix A:

$$A = U\Sigma V^\top$$

Where:

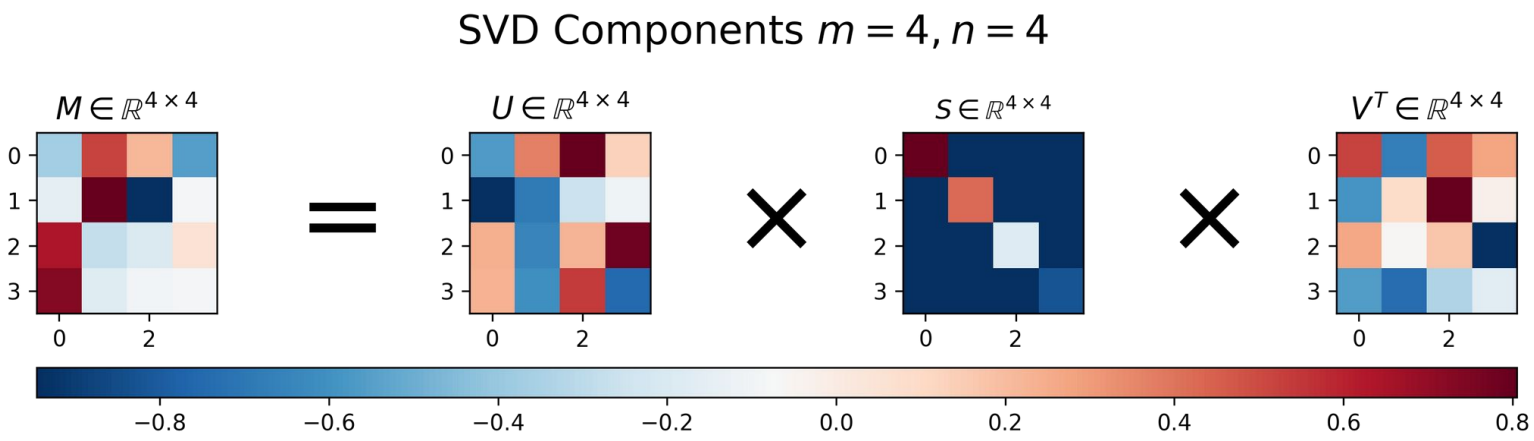
- U: left singular vectors (orthonormal)
- Σ: diagonal matrix of singular values

$$\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots)$$

- V: right singular vectors (orthonormal)

Key Properties

- Works for every matrix (not just square)
- Singular values represent “energy” or importance of directions


SVD Components $m = 4, n = 4$
$M \in \mathbb{R}^{4 \times 4}$
$U \in \mathbb{R}^{4 \times 4}$
$s \in \mathbb{R}^{4 \times 4}$
$V^T \in \mathbb{R}^{4 \times 4}$
0
1
2
3
0
2
=
×
×
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8

Geometric Interpretation of SVD

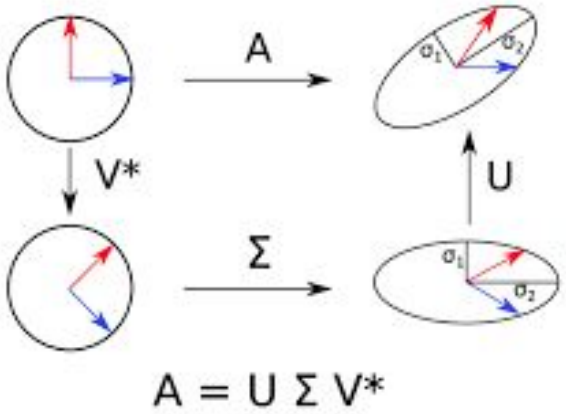
Newton School of Technology

Three Transformations

SVD breaks the transformation into:

$$A = U\Sigma V^\top$$

- $V^\top$: rotates the input
- Σ: scales (stretches/shrinks)
- U: rotates the output



Geometric Meaning

- Every matrix can be expressed as rotation → scaling → rotation
- Singular values give axis-aligned scaling
- Orthogonal matrices preserve lengths and angles
- Shows intrinsic geometric structure of data transformation

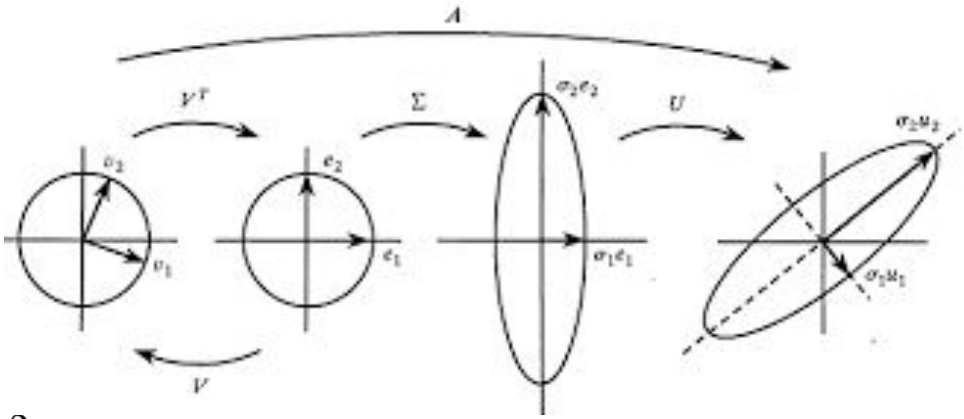
Geometric Interpretation of SVD

Newton School of Technology

PCA Connection

Apply SVD to centered data matrix X:

$$X = U\Sigma V^\top$$



- Columns of V: principal directions
- Singular values give importance of components

Explained Variance:

The variance captured by component i is:

$$\text{Variance}_i = \sigma_i^2$$

Why PCA Works

- Projects data onto directions of maximum variance
- New axes are orthogonal → no redundancy
- Dimensionality reduction retains major structure
- Reduces noise

1

1

1

Low-Rank Approximation

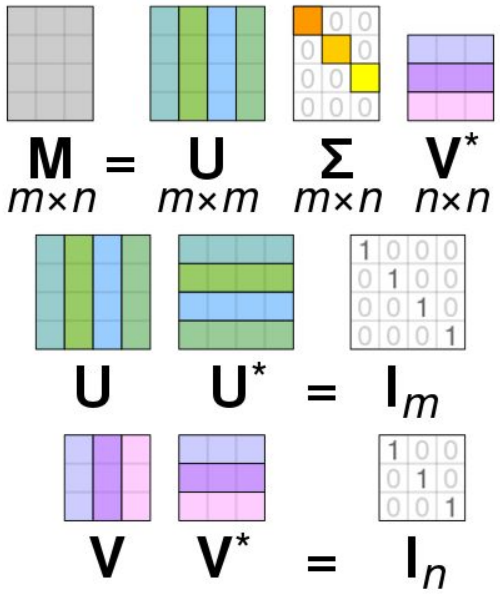
Newton School of Technology

Core Idea

Any matrix AAA can be approximated by keeping only the top-kkk singular values from its SVD:

$$A \approx A_k = U_k \Sigma_k V_k^\top$$

This gives the best rank-kkk approximation in terms of minimizing reconstruction error.



Why Low-Rank Works

- Real-world data often has hidden low-dimensional structure
- Noise appears in small singular values
- Keeping major singular values captures essential patterns

Eckart–Young Theorem

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

The best rank-k approximation is obtained by truncating SVD:

ML Relevance

- Compressing large embedding matrices

PCA: Step-by-Step Procedure

Newton School of Technology

Step 1 — Center the Data

Given data matrix X:

$$Xcentered = X - \mu$$

Each feature has mean zero.

Step 2 — Compute Covariance

$$C = \frac{1}{n} X_{\text{centered}}^{\top} X_{\text{centered}}$$

This is a symmetric matrix.

Step 3 — Perform SVD

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

Principal directions are columns of V.

$$Z = X_{\text{centered}} V_k$$

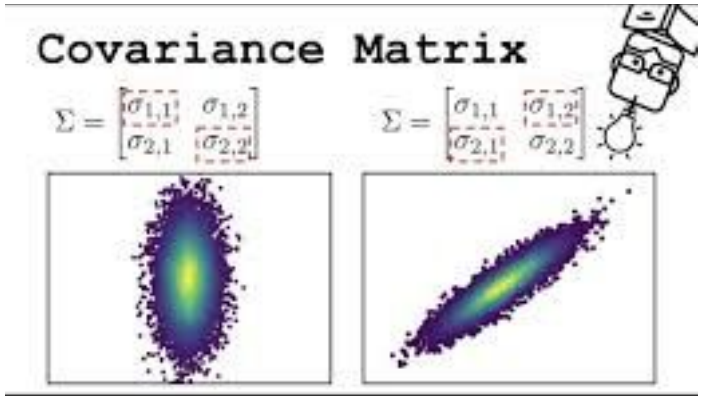
Step 4 — Dimensionality Reduction

Covariance Matrix and Principal Directions

Newton School of Technology

Covariance Matrix:

$$C = \frac{1}{n} X^\top X$$



Key properties:

- Symmetric
- Positive semi-definite
- Eigenvectors = principal directions
- Eigenvalues = variances along those directions

$$cv_i = \lambda_i v_i$$

Eigen-Decomposition of Covariance

v_i
λ_i

- : principal components
- : variance captured by component iii

Covariance Matrix and Principal Directions

Newton School of Technology

Connection to SVD

If $X = U\Sigma V^\top$, then:

$$C = V\Sigma^2 V^\top$$

Thus,

- Right singular vectors VVV = PCA directions
- Singular values give explained variance

ML Usage

- Identify dominant patterns
- Understand correlations

Matrix Compression Using SVD

Newton School of Technology

Compression Idea

Keep only top-kkk singular values:

$$A \approx U_k \Sigma_k V_k^\top$$

Storage required reduces from: $mn \quad \text{to} \quad k(m+n+1)$

where m, n are matrix dimensions.

Where Compression Helps

- Word embedding matrices (Word2Vec, GloVe)
- Vision model feature maps
- Recommender system matrices
- Large linear layers in LLMs

Why Do Models Use SVD/PCA?

Newton School of Technology

If your dataset has:

- 200 columns
- 200 image pixels
- 200 sensor readings

Most of those features are usually saying the same thing.
PCA/SVD help answer:

- “Which features actually matter?”
- “What directions carry the real information?”
- “What can we safely ignore?”

Analogy:
A movie has thousands of frames…
…but the story can be described in a few sentences.

What SVD Does to Data

Newton School of Technology

Think of your data matrix XXX as a complicated object.
SVD breaks it into 3 simple steps:

- Rotate the data → $V^\top$
- Stretch/Shrink the axes → σ_i
- Rotate back → U

Mathematically:

$$X = U\Sigma V^\top$$

SVD finds the cleanest, most meaningful directions in your data.

This is exactly how:

- image compression
- noise removal
- topic extraction

How PCA Helps You Understand Data

Newton School of Technology

PCA answers:

- "What direction has the highest variance?"
- "Where is the data mostly spread?"
- "What are the strongest patterns?"

PCA uses SVD: $X = U\Sigma V^\top$

The columns of V are the principal directions.

Example:
Plotting data in PCA space often reveals:

- hidden clusters
- natural groupings
- outliers
- structure you couldn't see before

Low-Rank Approximation: Why It Works in ML

If SVD shows these singular values:

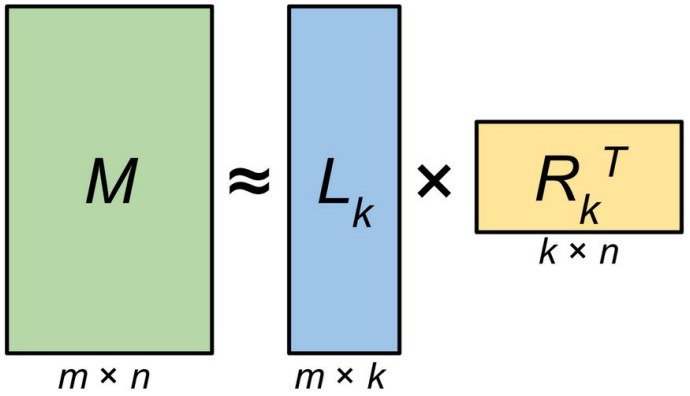
$$\sigma_1 = 100$$
$$\sigma_2 = 80$$
$$\sigma_3 = 8$$

- others: <1

Then almost all information is in the first 2–3 directions.

So we approximate:

$$A \approx U_k \Sigma_k V_k^\top$$



Why ML models love this:

- Remove noise
- Keep essential structure
- Reduce model size
- Make training faster

Analogy:

You can describe a person with:

- height
- weight
- age

You don't need 200 features (hair length, pinky size, etc.)

Real ML Examples

1. Face Recognition (Eigenfaces):
   - Faces are represented using PCA directions ("eigenfaces").
   - Only ~50 directions are enough to reconstruct a face.

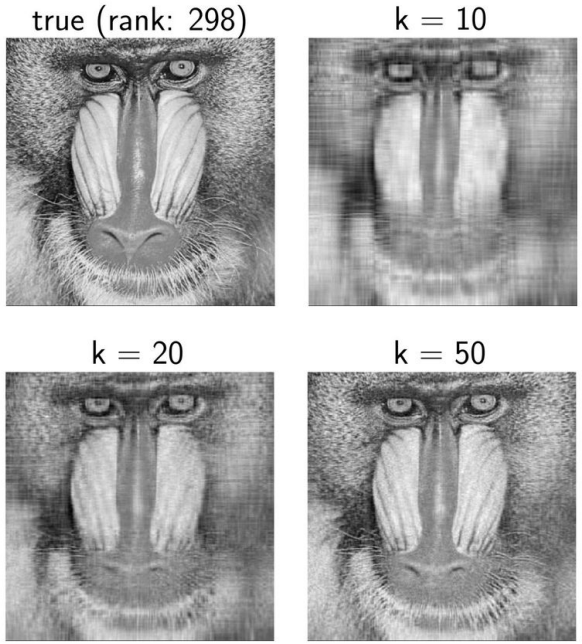
2. Image Compression (JPEG-like idea)
   - Keep only important singular values → smaller file, same image.

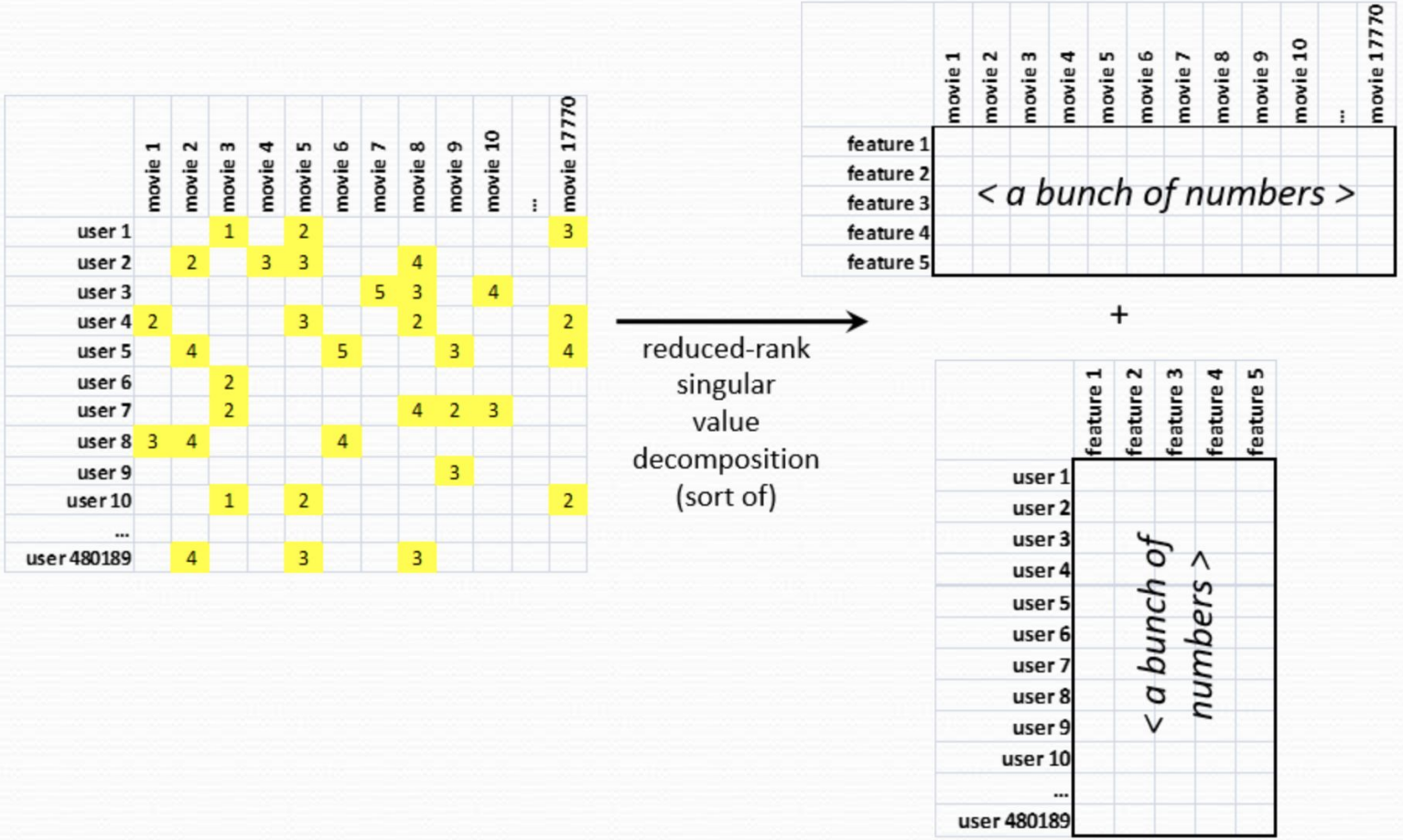
3. NLP Embedding Compression
   - Large embedding matrices can be approximated using low-rank SVD.

4. Recommender Systems
   - User–item matrix → SVD reveals latent interests.

5. Noise Removal
   - Drop small singular values to remove noise from signals/images.

6. Data Visualization
   - PCA → convert messy high-dimensional data into 2D/3D shapes.




movie 1
movie 2
movie 3
movie 4
movie 5
movie 6
movie 7
movie 8
movie 9
movie 10
...
movie 17770
user 1
user 2
user 3
user 4
user 5
user 6
user 7
user 8
user 9
user 10
...
user 480189
reduced-rank singular value decomposition (sort of)
feature 1
feature 2
feature 3
feature 4
feature 5
< a bunch of numbers >
+
< a bunch of numbers >

Summary

Newton School of Technology

- Eigenvalues and eigenvectors reveal how a matrix transforms space.
- Diagonalization simplifies a matrix into independent scaling directions.
- SVD breaks any matrix into rotations + scaling and is fundamental in ML.
- PCA uses SVD to find the most informative directions in data.
- Low-rank approximation keeps only the essential structure, reducing noise and size.
- These tools power applications like compression, visualization, and feature reduction.