Støddfrøðiligt grundarlag til tilgjørt vit - 2025

October 14, 2025

Contents

1 Notes

2 Linear Algebra

2.1 Matrix

A matrix is usually written as $\mathbf{A} \in \mathbb{R}^{m \times n}$ where m means the amount of rows and n is the amount of columns. An example of a matrix:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad a_{ij} \in \mathbb{R}$$

2.2 Matrix addition

Property	**Description**
Associative	$(A + B) + C = A + (B + C)$
Commutative	$A + B = B + A$
Identity	$A + O = O + A = A$, where O is the zero matrix
Inverse	$A + (-A) = O$, where $-A$ is the additive inverse

Example

2.3 Matrix multiplication

Property	**Description**
Associative	$(AB)C = A(BC)$
Distributive over Addition	$A(B + C) = AB + AC$ $(A + B)C = AC + BC$
Identity	$AI = IA = A$, where I is the identity matrix
Non-commutative	$AB \neq BA$ (in general)

Make sure to check the dimensions of the matrices, that they are compatible and that the resulting matrix has the correct dimensions. On the example below $A \cdot B = C$ notice the dimensions.

$$A \in \mathbb{R}^{n \times m}$$
$$B \in \mathbb{R}^{m \times p}$$
$$C \in \mathbb{R}^{n \times p}$$

Example

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix} \in \mathbb{R}^{2\times 3},$$
$$B = \begin{bmatrix} 0 & 2 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{3\times 2},$$
$$AB = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} 0 & 2 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 2 & 5 \end{bmatrix} \in \mathbb{R}^{2\times 2},$$
$$BA = \begin{bmatrix} 0 & 2 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 6 & 4 & 2 \\ -2 & 0 & 2 \\ 3 & 2 & 1 \end{bmatrix} \in \mathbb{R}^{3\times 3}.$$

2.4 Inverse and determinant

A matrix only has a inverse if its determinant is not 0 so always calculate it before finding the inverse. If it exists the following holds:

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$$

The determinant is easy to find for a 2×2 matrix:

$$Det(\mathbf{A}) = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11}a_{22} - a_{21}a12$$

After checking that the determinant is not zero proceed to find the inverse. If you're working with a 2×2 matrix the following setup gives the inverse, fig. **??**
For a 3×3 matrix Sarrus's rule can be used.
For matrices 3×3 and larger one can use Laplace expansion.

2.5 Transposed matrix

To find the transpose

$$\mathbf{A} = \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix}$$

then

$$\mathbf{A}^T = \begin{bmatrix} a & d \\ b & e \\ c & f \end{bmatrix}$$

2.6 Particular and general solution

2.7 Gauss elimination

When doing gaussian elimination you are allowed to do the following operations:

- Exchange of two equations (rows in the matrix representing the system of equations)
- Multiplication of an equation (row) with a constant $\lambda \in \mathbb{R}\{0\}$
- Addition of two equations (rows)

Definition 2.6 (Row-Echelon Form). A matrix is in row-echelon form if:

- All rows that contain only zeros are at the bottom of the matrix; correspondingly, all rows that contain at least one nonzero element are on top of rows that contain only zeros.
- Looking at nonzero rows only, the first nonzero number from the left (also called the pivot or the leading coefficient) is always strictly to the right of the pivot of the row above it.

Remark (Reduced Row Echelon Form). An equation system is in *reduced row-echelon form* (also: *row-reduced echelon form* or *row canonical form*) if

- It is in row-echelon form.
- Every pivot is 1.
- The pivot is the only nonzero entry in its column.

Minus-1 trick

2.8 Groups

Definition 2.7 (Group). Consider a set G and an operation $\otimes : G \times G \to G$ defined on G. Then $G := (G, \otimes)$ is called a *group* if the following hold:

1. **Closure** of G under $\otimes$: $\forall x, y \in G : x \otimes y \in G$
2. **Associativity**: $\forall x, y, z \in G : (x \otimes y) \otimes z = x \otimes (y \otimes z)$
3. **Neutral element**: $\exists e \in G \, \forall x \in G : x \otimes e = x$ and $e \otimes x = x$
4. **Inverse element**: $\forall x \in G \, \exists y \in G : x \otimes y = e$ and $y \otimes x = e$, where e is the neutral element. We often write x^{-1} to denote the inverse element of x.

Remark The inverse element is defined with respect to the operation $\otimes$ and does not necessarily mean $\frac{1}{x}$. $\diamondsuit$

2.9 Vector spaces

Definition 2.1 (Vector Space)**.** A real-valued *vector space* $V = (\mathcal{V}, +, \cdot)$ is a set $\mathcal{V}$ with two operations

$$+ : \mathcal{V} \times \mathcal{V} \to \mathcal{V} \tag{2.62}$$

$$\cdot : \mathbb{R} \times \mathcal{V} \to \mathcal{V} \tag{2.63}$$

where

1. $(\mathcal{V}, +)$ is an Abelian group.
2. Distributivity:
   (a) $\forall \lambda \in \mathbb{R},\ x, y \in \mathcal{V} : \lambda \cdot (x + y) = \lambda \cdot x + \lambda \cdot y.$
   (b) $\forall \lambda, \psi \in \mathbb{R},\ x \in \mathcal{V} : (\lambda + \psi) \cdot x = \lambda \cdot x + \psi \cdot x.$
3. Associativity (outer operation): $\forall \lambda, \psi \in \mathbb{R},\ x \in \mathcal{V} : \lambda \cdot (\psi \cdot x) = (\lambda\psi) \cdot x.$
4. Neutral element with respect to the outer operation: $\forall x \in \mathcal{V} :\ 1 \cdot x = x.$

2.10 Basis and rank

To determine a basis take the vectors that span the vector subspace and put them into matrix form:

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \\ -1 \\ -1 \\ -1 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 2 \\ -1 \\ 1 \\ 2 \\ -2 \end{bmatrix}, \quad \mathbf{x}_3 = \begin{bmatrix} 3 \\ -4 \\ 3 \\ 5 \\ -3 \end{bmatrix}, \quad \mathbf{x}_4 = \begin{bmatrix} -1 \\ 8 \\ -5 \\ -7 \\ 1 \end{bmatrix}$$

We check that the following holds:

$$\sum_{i=1}^{4} \lambda_i \mathbf{x}_i = \mathbf{0}$$

$$[\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4] = \begin{bmatrix} 1 & 2 & 3 & -1 \\ 2 & -1 & -4 & 8 \\ -1 & 1 & 3 & -5 \\ -1 & 2 & 5 & -6 \\ -1 & -2 & -3 & 1 \end{bmatrix}$$

$$\rightsquigarrow \left[\begin{array}{cccc} 1 & 2 & 3 & -3 \\ \hline 0 & 1 & 2 & -2 \\ \hline 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

The resulting matrix has three columns with pivots. Those form a basis for the subspace spanned by those four vectors.
We see that there are two dependent rows, those that are all zeros and that there is one free variable $\mathbf{x}_3$ hence the rank is 3:

$$rk(A) = 3$$

The column with no pivot, $\mathbf{x}_3$,

A matrix is full rank if the rank equals the largest possible rank for a matrix of the same dimensions.

2.11 Linear mappings

A **linear mapping** (or linear transformation) $\Phi : V \to W$ between vector spaces V and W over the same field preserves vector addition and scalar multiplication. For all $\mathbf{x}, \mathbf{y} \in V$ and $\lambda, \psi \in \mathbb{R}$:

$$\Phi(\mathbf{x} + \mathbf{y}) = \Phi(\mathbf{x}) + \Phi(\mathbf{y}), \quad \Phi(\lambda\mathbf{x}) = \lambda\Phi(\mathbf{x}).$$

This can be summarized as:

$$\Phi(\lambda\mathbf{x} + \psi\mathbf{y}) = \lambda\Phi(\mathbf{x}) + \psi\Phi(\mathbf{y}).$$

2.12 Special Cases of Linear Mappings

- **Isomorphism**: A linear mapping $\Phi : V \to W$ that is both injective and surjective (bijective). Two vector spaces are isomorphic if and only if $\dim(V) = \dim(W)$.
- **Endomorphism**: A linear mapping $\Phi : V \to V$ from a vector space to itself.
- **Automorphism**: An endomorphism that is also bijective.
- **Identity Mapping**: $\mathrm{id}_V : V \to V, \mathbf{x} \mapsto \mathbf{x}$.

2.13 Injective, Surjective, and Bijective Mappings

- **Injective (One-to-One)**: $\Phi(\mathbf{x}) = \Phi(\mathbf{y}) \implies \mathbf{x} = \mathbf{y}$.
- **Surjective (Onto)**: $\Phi(V) = W$.
- **Bijective**: Both injective and surjective. A bijective mapping Φ has an inverse Φ^{-1}.

2.14 Matrix Representation of Linear Mappings

For finite-dimensional vector spaces V and W with ordered bases $B = (\mathbf{b}_1, \ldots, \mathbf{b}_n)$ and $C = (\mathbf{c}_1, \ldots, \mathbf{c}_m)$, the **transformation matrix** $\mathbf{A}_\Phi$ of Φ is defined by:

$$\Phi(\mathbf{b}_j) = \sum_{i=1}^{m} \alpha_{ij} \mathbf{c}_i, \quad \text{where } \mathbf{A}_\Phi(i,j) = \alpha_{ij}.$$

If $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are the coordinate vectors of $\mathbf{x} \in V$ and $\Phi(\mathbf{x}) \in W$ with respect to B and C, then:

$$\hat{\mathbf{y}} = \mathbf{A}_\Phi \hat{\mathbf{x}}.$$

2.15 Example: Transformation Matrix

For a linear mapping $\Phi : V \to W$ with bases $B = (\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$ and $C = (\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4)$, if:

$$\Phi(\mathbf{b}_1) = \mathbf{c}_1 - \mathbf{c}_2 + 3\mathbf{c}_3 - \mathbf{c}_4, \quad \Phi(\mathbf{b}_2) = 2\mathbf{c}_1 + \mathbf{c}_2 + 7\mathbf{c}_3 + 2\mathbf{c}_4, \quad \Phi(\mathbf{b}_3) = 3\mathbf{c}_2 + \mathbf{c}_3 + 4\mathbf{c}_4,$$

the transformation matrix $\mathbf{A}_\Phi$ is:

$$\mathbf{A}_\Phi = \begin{bmatrix} 1 & 2 & 0 \\ -1 & 1 & 3 \\ 3 & 7 & 1 \\ -1 & 2 & 4 \end{bmatrix}.$$

2.16 Basis Change

If the bases in V and W are changed to $\tilde{B}$ and $\tilde{C}$, the new transformation matrix $\tilde{\mathbf{A}}_\Phi$ is related to the original $\mathbf{A}_\Phi$ by:

$$\tilde{\mathbf{A}}_\Phi = \mathbf{T}^{-1} \mathbf{A}_\Phi \mathbf{S},$$

where $\mathbf{S}$ maps coordinates from $\tilde{B}$ to B, and $\mathbf{T}$ maps coordinates from $\tilde{C}$ to C.

2.17 Example: Basis Change

For a linear mapping $\Phi : \mathbb{R}^3 \to \mathbb{R}^4$ with standard bases B and C, and new bases $\tilde{B}$ and $\tilde{C}$, the transformation matrices $\mathbf{S}$ and $\mathbf{T}$ are constructed by expressing the new basis vectors in terms of the old basis vectors. The new transformation matrix is:

$$\tilde{\mathbf{A}}_\Phi = \mathbf{T}^{-1} \mathbf{A}_\Phi \mathbf{S}.$$

2.18 Image and Kernel

For a linear mapping $\Phi : V \to W$:

- **Kernel (Null Space)**: $\ker(\Phi) = \{\mathbf{v} \in V \mid \Phi(\mathbf{v}) = \mathbf{0}_W\}$. The kernel is a subspace of V.
- **Image (Range)**: $\text{Im}(\Phi) = \{\Phi(\mathbf{v}) \mid \mathbf{v} \in V\}$. The image is a subspace of W.

2.19 Rank-Nullity Theorem

The **Rank-Nullity Theorem** states:

$$\dim(\ker(\Phi)) + \dim(\text{Im}(\Phi)) = \dim(V).$$

This theorem is fundamental in linear algebra and has several important consequences:

- If $\dim(\text{Im}(\Phi)) < \dim(V)$, then $\ker(\Phi)$ is non-trivial.
- If $\dim(V) = \dim(W)$, then Φ is injective if and only if it is surjective if and only if it is bijective.

2.20 Example: Image and Kernel

For a linear mapping $\Phi : \mathbb{R}^4 \to \mathbb{R}^2$ defined by:

$$\Phi(\mathbf{x}) = \begin{bmatrix} 1 & 2 & -1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \mathbf{x},$$

the image is the span of the columns of the transformation matrix:

$$\text{Im}(\Phi) = \text{span}\left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}.$$

The kernel is found by solving $\mathbf{Ax} = \mathbf{0}$:

$$\ker(\Phi) = \text{span}\left\{ \begin{bmatrix} 0 \\ \frac{1}{2} \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ \frac{1}{2} \\ 0 \\ 1 \end{bmatrix} \right\}.$$

2.21 Key Theorems and Results

- Two finite-dimensional vector spaces V and W are isomorphic if and only if $\dim(V) = \dim(W)$.
- The composition of linear mappings is linear. If $\Phi : V \to W$ and $\Psi : W \to X$ are linear, then $\Psi \circ \Phi : V \to X$ is also linear.
- If $\Phi : V \to W$ is an isomorphism, then $\Phi^{-1} : W \to V$ is also an isomorphism.

2.22 Examples of Linear Mappings

- The mapping $\Phi : \mathbb{R}^2 \to \mathbb{C}, \Phi(\mathbf{x}) = x_1 + ix_2$ is a homomorphism, justifying the representation of complex numbers as tuples in $\mathbb{R}^2$.
- Linear transformations can represent geometric operations such as rotations, stretches, and reflections in $\mathbb{R}^2$.

2.23 Example: Geometric Transformations

For a rotation by 45° in $\mathbb{R}^2$, the transformation matrix is:

$$\mathbf{A}_1 = \begin{bmatrix} \cos(\pi/4) & -\sin(\pi/4) \\ \sin(\pi/4) & \cos(\pi/4) \end{bmatrix}.$$

For a stretch along the horizontal axis by a factor of 2, the transformation matrix is:

$$\mathbf{A}_2 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}.$$

2.24 Affine spaces

3 Analytic Geometry

3.1 Norms

Definition 3.1 (Norm). A **norm** on a vector space V is a function

$$\|\cdot\| : V \to \mathbb{R},$$

$$\boldsymbol{x} \mapsto \|\boldsymbol{x}\|,$$

which assigns each vector $\boldsymbol{x}$ its *length* $\|\boldsymbol{x}\| \in \mathbb{R}$, such that for all $\lambda \in \mathbb{R}$ and $\boldsymbol{x}, \boldsymbol{y} \in V$ the following hold:

- **Absolutely homogeneous:** $\|\lambda\boldsymbol{x}\| = |\lambda|\|\boldsymbol{x}\|$
- **Triangle inequality:** $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$
- **Positive definite:** $\|\boldsymbol{x}\| \geq 0$ and $\|\boldsymbol{x}\| = 0 \iff \boldsymbol{x} = \boldsymbol{0}$

3.2 Inner products

3.3 Lengths and distances

3.4 Angles and orthogonality

3.5 Orthonormal basis

3.6 Orthogonal complement

3.7 Inner product of functions

3.8 Orthogonal projections

3.9 Rotations

4 Matrix Decompositions

4.1 Determinant and trace

$$det(\mathbf{A}) = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11} \cdot a_{12} - (a_{21} \cdot a_{22})$$

if $det(\mathbf{A}) = 0$ then matrix A is not invertible and the vectors are linearly independent. In other words if $det(\mathbf{A}) \neq 0$, then there exists an A^{-1} and the vectors are linearly independent, which also follows that $\mathbf{A}$ has a full rank.

Sarrus's rule *(ONLY FOR 3X3 matrices)*:

$$det(\mathbf{A}) = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{13} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = a_{11} \cdot a_{22} \cdot a_{33} + a_{21} \cdot a_{32} \cdot a_{13} + a_{12} \cdot a_{13} \cdot a_{31}$$
$$- a_{31} \cdot a_{22} \cdot a_{13} - a_{21} \cdot a_{12} \cdot a_{33} + a_{32} \cdot a_{13} \cdot a_{11}$$

Laplace rule *Works for all matrices*
Let:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

We expand along the ***first row*** (shown in blue):

$$\det(\mathbf{A}) = a_{11} \cdot \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \cdot \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \cdot \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

Each 2×2 determinant is called a **minor**, and the sign alternates as $(+, -, +, \dots)$ across the row or column.

$$\mathbf{A} = \begin{bmatrix} + & - & + \\ - & + & - \\ + & - & + \end{bmatrix}$$

The trace of a square matrix $\mathbf{A} \in \mathbb{R}^{nxn}$ is defined as

$$tr(A) := \sum_{i=1}^{n} a_{ii}$$

i.e the trace is the sum of the diagonal elements of $\mathbf{A}$

4.2 Eigenvalues and eigenvectors

Eigenvalues and Eigenvectors *(Only for square matrices)*:

Let:

$$\mathbf{A} \cdot \vec{v} = \lambda \cdot \vec{v}$$

Where:

- $\mathbf{A}$ is a square matrix,
- $\vec{v} \neq \vec{0}$ is an eigenvector,
- λ is the corresponding eigenvalue.

—

To find the eigenvalues:

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0$$

This is called the *characteristic equation*. Solve for λ.

—

To find the eigenvectors:

$$(\mathbf{A} - \lambda \mathbf{I}) \cdot \vec{v} = \vec{0}$$

Solve this homogeneous system for $\vec{v}$.

—

Example:

Let:

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

Step 1: Find eigenvalues λ

$$\begin{aligned} \det(\mathbf{A} - \lambda \mathbf{I}) = \det \begin{bmatrix} 4-\lambda & 2 \\ 1 & 3-\lambda \end{bmatrix} &= (4-\lambda)(3-\lambda) - 2 \cdot 1 \\ &= \lambda^2 - 7\lambda + 10 = 0 \end{aligned}$$

Solve:

$$\lambda = 5, \quad \lambda = 2$$

—

Step 2: Find eigenvectors

For $\lambda = 5$:

$$(\mathbf{A} - 5\mathbf{I}) = \begin{bmatrix} -1 & 2 \\ 1 & -2 \end{bmatrix} \Rightarrow \begin{bmatrix} -1 & 2 \\ 1 & -2 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Solve:

$$-1x + 2y = 0 \Rightarrow x = 2y \Rightarrow \vec{v}_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \quad \text{(or any scalar multiple)}$$

Repeat similarly for $\lambda = 2$:

$$(\mathbf{A} - 2\mathbf{I}) = \begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix} \Rightarrow \vec{v}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

—

Result:

- Eigenvalues: $\lambda = 5, 2$
- Eigenvectors:

$$\vec{v}_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad \vec{v}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

4.3 Cholesky decomposition

Cholesky Decomposition *(Only for real, symmetric, positive definite matrices)*:

Cholesky decomposition expresses a matrix $\mathbf{A}$ as:

$$\mathbf{A} = \mathbf{L} \cdot \mathbf{L}^T$$

Where:

- $\mathbf{A}$ is a symmetric, positive definite matrix,
- $\mathbf{L}$ is a lower triangular matrix,
- $\mathbf{L}^T$ is the transpose of $\mathbf{L}$.

—

Example:

Let:

$$\mathbf{A} = \begin{bmatrix} 4 & 12 & -16 \\ 12 & 37 & -43 \\ -16 & -43 & 98 \end{bmatrix}$$

We want to find:

$$\mathbf{A} = \mathbf{L} \cdot \mathbf{L}^T, \quad \text{where } \mathbf{L} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix}$$

—

Step 1: Compute elements of $\mathbf{L}$

$$l_{11} = \sqrt{4} = 2$$
$$l_{21} = \frac{12}{l_{11}} = \frac{12}{2} = 6$$
$$l_{31} = \frac{-16}{l_{11}} = \frac{-16}{2} = -8$$

$$l_{22} = \sqrt{37 - l_{21}^2} = \sqrt{37 - 36} = \sqrt{1} = 1$$
$$l_{32} = \frac{-43 - (l_{31} \cdot l_{21})}{l_{22}} = \frac{-43 - (-8 \cdot 6)}{1} = \frac{-43 + 48}{1} = 5$$

$$l_{33} = \sqrt{98 - (l_{31}^2 + l_{32}^2)} = \sqrt{98 - (64 + 25)} = \sqrt{9} = 3$$

—

Final result:

$$\mathbf{L} = \begin{bmatrix} 2 & 0 & 0 \\ 6 & 1 & 0 \\ -8 & 5 & 3 \end{bmatrix}, \quad \mathbf{A} = \mathbf{L} \cdot \mathbf{L}^T$$

4.4 Eigendecomposition and diagonalization

Eigendecomposition / Diagonalization *(For square matrices with enough linearly independent eigenvectors)*:

If a matrix $\mathbf{A}$ has n linearly independent eigenvectors, then it is **diagonalizable**, and we can write:

$$\mathbf{A} = \mathbf{P} \cdot \mathbf{D} \cdot \mathbf{P}^{-1}$$

Where:

- $\mathbf{P}$ is the matrix of eigenvectors (each column is an eigenvector),
- $\mathbf{D}$ is a diagonal matrix with eigenvalues on the diagonal,
- $\mathbf{A}$ must have n linearly independent eigenvectors (i.e., $\mathbf{P}$ is invertible).

—

Example:
Let:

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

Step 1: Find eigenvalues
Solve:

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0$$

$$\det \begin{bmatrix} 4-\lambda & 1 \\ 2 & 3-\lambda \end{bmatrix} = (4-\lambda)(3-\lambda) - 2 \cdot 1$$
$$= \lambda^2 - 7\lambda + 10 = 0 \Rightarrow \lambda_1 = 5, \quad \lambda_2 = 2$$

—

Step 2: Find eigenvectors
For $\lambda = 5$:

$$(\mathbf{A} - 5\mathbf{I}) = \begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix} \Rightarrow \vec{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

For $\lambda = 2$:

$$(\mathbf{A} - 2\mathbf{I}) = \begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix} \Rightarrow \vec{v}_2 = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

—

Step 3: Form $\mathbf{P}$ and $\mathbf{D}$

$$\mathbf{P} = \begin{bmatrix} 1 & -1 \\ 1 & 2 \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} 5 & 0 \\ 0 & 2 \end{bmatrix}$$

Then:

$$\mathbf{A} = \mathbf{P} \cdot \mathbf{D} \cdot \mathbf{P}^{-1}$$

—

Check (optional):
Compute $\mathbf{P}^{-1}$, then verify:

$$\mathbf{P} \cdot \mathbf{D} \cdot \mathbf{P}^{-1} = \mathbf{A}$$

—

Result:

- Eigenvalues: $\lambda = 5, 2$
- Eigenvectors:
$$\vec{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \vec{v}_2 = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$
- $\mathbf{A} = \mathbf{P} \cdot \mathbf{D} \cdot \mathbf{P}^{-1}$

4.5 Singular value decomposition

Singular Value Decomposition (SVD) *(Works for any real $m \times n$ matrix)*:
Every real matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ can be decomposed as:

$$\mathbf{A} = \mathbf{U} \cdot \mathbf{\Sigma} \cdot \mathbf{V}^T$$

Where:

- $\mathbf{U} \in \mathbb{R}^{m \times m}$ is an orthogonal matrix ($\mathbf{U}^T\mathbf{U} = \mathbf{I}$),
- $\mathbf{V} \in \mathbb{R}^{n \times n}$ is an orthogonal matrix,
- $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$ is a diagonal matrix with non-negative **singular values** on the diagonal,
- The singular values are square roots of eigenvalues of $\mathbf{A}^T\mathbf{A}$ (or $\mathbf{A}\mathbf{A}^T$).

—

Example:
Let:

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \quad \text{(a symmetric } 2 \times 2 \text{ matrix)}$$

—

Step 1: Compute $\mathbf{A}^T\mathbf{A}$

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$$

—

Step 2: Find eigenvalues of $\mathbf{A}^T\mathbf{A}$

$$\det\left(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}\right) = \det \begin{bmatrix} 10-\lambda & 6 \\ 6 & 10-\lambda \end{bmatrix} = (10-\lambda)^2 - 36 = 0$$

$$\Rightarrow \lambda_1 = 16, \quad \lambda_2 = 4$$

—

Step 3: Compute singular values

Remember! $\sigma_1 \geq \sigma_2 \geq \sigma_n \geq 0$

$$\sigma_1 = \sqrt{16} = 4, \quad \sigma_2 = \sqrt{4} = 2$$

So:

$$\mathbf{\Sigma} = \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}$$

—

Step 4: Find right singular vectors ($\mathbf{V}$)
Eigenvectors of $\mathbf{A}^T\mathbf{A}$:
- For $\lambda = 16$: $\vec{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ - For $\lambda = 4$: $\vec{v}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

Normalize:

$$\mathbf{V} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

—

Step 5: Find left singular vectors ($\mathbf{U}$)

Use:

$$\mathbf{u}_i = \frac{1}{\sigma_i}\mathbf{A}\vec{v}_i$$

Compute:

$$\mathbf{u}_1 = \frac{1}{4} \cdot \mathbf{A} \cdot \begin{bmatrix}1\\1\end{bmatrix} = \frac{1}{4}\begin{bmatrix}3+1\\1+3\end{bmatrix} = \frac{1}{4}\begin{bmatrix}4\\4\end{bmatrix} = \begin{bmatrix}\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\end{bmatrix}$$

$$\mathbf{u}_2 = \frac{1}{2} \cdot \mathbf{A} \cdot \begin{bmatrix}-1\\1\end{bmatrix} = \frac{1}{2}\begin{bmatrix}-3+1\\-1+3\end{bmatrix} = \frac{1}{2}\begin{bmatrix}-2\\2\end{bmatrix} = \begin{bmatrix}\frac{-1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\end{bmatrix}$$

So:

$$\mathbf{U} = \begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}}\\\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix}$$

—

Final Result:

$$\mathbf{A} = \mathbf{U} \cdot \mathbf{\Sigma} \cdot \mathbf{V}^T$$

Where:

$$\mathbf{U} = \begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}}\\\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix}4 & 0\\0 & 2\end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}}\\\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix}$$

4.6 Matrix approximation (Summary Mistral)

4.7 Introduction

The Singular Value Decomposition (SVD) allows a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ to be factorized as:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top,$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices, and $\boldsymbol{\Sigma}$ is a diagonal matrix containing singular values σ_i.

4.8 Rank-1 Matrices

A rank-1 matrix $\boldsymbol{A}_i$ is constructed as the outer product of the i-th left and right singular vectors:

$$\boldsymbol{A}_i = \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top.$$

The matrix $\boldsymbol{A}$ can be expressed as a sum of rank-1 matrices:

$$\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top,$$

where r is the rank of $\boldsymbol{A}$.

4.9 Rank-k Approximation

A rank-k approximation of $\boldsymbol{A}$ is given by:

$$\widehat{\boldsymbol{A}}(k) = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top.$$

This approximation is optimal in the spectral norm sense, as stated by the **Eckart-Young Theorem**:

$$\widehat{\boldsymbol{A}}(k) = \operatorname{argmin}_{\mathrm{rk}(\boldsymbol{B})=k} \|\boldsymbol{A} - \boldsymbol{B}\|_2,$$

with the error $\|\boldsymbol{A} - \widehat{\boldsymbol{A}}(k)\|_2 = \sigma_{k+1}$.

4.10 Applications

The rank-k approximation is used for:

- Dimensionality reduction
- Data compression
- Noise filtering
- Principal Component Analysis (PCA)

4.11 Example: Image Compression

For an image represented by $\boldsymbol{A} \in \mathbb{R}^{1432\times1910}$, a rank-5 approximation requires only 0.6% of the original storage, demonstrating the efficiency of SVD for compression.

4.12 Spectral Norm

The spectral norm of $\boldsymbol{A}$ is defined as:

$$\|\boldsymbol{A}\|_2 = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = \sigma_1.$$

4.13 Key Takeaways

- SVD enables principled matrix approximation by projecting $\boldsymbol{A}$ onto a lower-dimensional space.
- The Eckart-Young Theorem guarantees the optimality of the rank-k approximation.
- Applications include image processing, clustering, and regularization.

4.14 Matrix approximation (Summary ChatGPT)

Setup. Given a matrix $A \in \mathbb{R}^{m\times n}$ with singular value decomposition (SVD) $A = U\Sigma V^\top$, let $\{u_i\}_{i=1}^r$ and $\{v_i\}_{i=1}^r$ denote the left- and right-singular vectors corresponding to nonzero singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, where $r = \operatorname{rank}(A)$. The SVD expresses A as a sum of rank-1 outer products $\{u_i v_i^\top\}$ scaled by σ_i.

Rank-1 building blocks. Define the rank-1 matrices

$$A_i := u_i v_i^\top \in \mathbb{R}^{m\times n}.$$

Then the full matrix can be written as the finite sum

$$A = \sum_{i=1}^{r} \sigma_i A_i = \sum_{i=1}^{r} \sigma_i \, u_i v_i^\top.$$

This decomposition isolates interpretable "directions" in the data, each contributing σ_i units of energy/strength.

Low-rank (truncated) approximation. For a target rank $k < r$, the truncated SVD (rank-k approximation) is

$$A^{(k)} := \sum_{i=1}^{k} \sigma_i \, u_i v_i^\top, \qquad \operatorname{rank}\Big(A^{(k)}\Big) = k,$$

obtained by keeping only the top k singular values/vectors. This provides a principled lossy compression of A.

Spectral norm. The spectral norm of a matrix A is

$$\|A\|_2 := \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2},$$

and satisfies

$$\|A\|_2 = \sigma_1.$$

Thus the action of A on vectors is dominated by its largest singular value.

Eckart–Young theorem. Among all rank-k matrices B, the truncated SVD $A^{(k)}$ uniquely minimizes the spectral-norm error:

$$A^{(k)} \;=\; \arg\min_{\operatorname{rank}(B)=k} \|A - B\|_2, \qquad \|A - A^{(k)}\|_2 = \sigma_{k+1}.$$

Hence the best achievable spectral-norm error at rank k is exactly the next singular value.

Interpretation and uses. Truncated SVD projects A onto the manifold of rank-$\leq k$ matrices with provably minimal spectral-norm distortion. It underpins dimensionality reduction, denoising/regularization, topic modeling, and data compression. In practice, storing k singular values and the corresponding singular vectors can dramatically reduce storage compared to the full $m \times n$ matrix, while preserving the dominant structure of the data.

Toy example (ratings). In a small movie–ratings matrix, the rank-1 terms can align with coherent latent "themes" (e.g., sci-fi vs. art-house). A rank-2 approximation $A^{(2)}$ may already reconstruct the table well, indicating that only two latent factors are needed to explain the observed preferences.

5 Vector Calculus

5.1 Differentiation of Univariate Functions

Consider a scalar function $f : \mathbb{R} \to \mathbb{R}$, where $x \in \mathbb{R}$.

—

Derivative:
The derivative of f at x is defined as

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

if the limit exists.

—

Basic differentiation rules:

- **Constant rule:** $\frac{d}{dx} c = 0$
- **Power rule:** $\frac{d}{dx} x^n = nx^{n-1}$
- **Sum rule:** $\frac{d}{dx}[f(x) + g(x)] = f'(x) + g'(x)$
- **Product rule:** $\frac{d}{dx}[f(x)g(x)] = f'(x)g(x) + f(x)g'(x)$
- **Quotient rule:** $\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \frac{f'(x)g(x) - f(x)g'(x)}{[g(x)]^2}$
- **Chain rule:** $\frac{d}{dx} f(g(x)) = f'(g(x)) \cdot g'(x)$

—

Higher-order derivatives:
Derivatives can be extended to higher orders:

$$f^{(n)}(x) = \frac{d^n}{dx^n} f(x)$$

—

Taylor series expansion:
If f is infinitely differentiable at $x = a$, it can be approximated near a by its Taylor series:

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots$$

—

Remainder term:
The error of truncating the series at order N is given by the remainder $R_N(x)$, which often can be bounded or expressed via Lagrange's form:

$$R_N(x) = \frac{f^{(N+1)}(\xi)}{(N+1)!}(x-a)^{N+1} \quad \text{for some } \xi \text{ between } a \text{ and } x$$

—

Example:
For $f(x) = e^x$ expanded at $a = 0$:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots$$

—

Interpretation:
- $f'(x)$ measures the instantaneous rate of change of f at x. - Taylor series expresses f as an infinite polynomial centered at a. - Linearization is the first-order truncation of the Taylor series.

—

Final Result (Boxed):

$$\boxed{f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n}$$

5.2 Partial Differentiation and Gradients

Partial differentiation is used when dealing with functions of multiple variables, such as $f(x, y)$ or $f(x, y, z)$.

We differentiate with respect to one variable at a time, treating the others as constants.

—

Notation:

$$\frac{\partial f}{\partial x}, \quad \frac{\partial f}{\partial y}, \quad \frac{\partial f}{\partial z}, \quad \text{etc.}$$

—

Example:
Let:

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

Partial derivative w.r.t. x:

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

Partial derivative w.r.t. y:

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

—

Gradient Vector:
The **gradient** of a scalar function $f(x, y, \dots)$ is a vector of its partial derivatives:

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \vdots \\ \frac{\partial f}{\partial z} \end{bmatrix}$$

So for the above function:

$$\nabla f(x, y) = \begin{bmatrix} 2xy + 3y^2 \\ x^2 + 6xy \end{bmatrix}$$

—

Interpretation:

- The gradient points in the direction of **steepest ascent**.
- Its magnitude gives the rate of increase in that direction.
- When $\nabla f = \vec{0}$, you may be at a maximum, minimum, or saddle point (check with second derivatives).

—

Final Result (Boxed):

$$\boxed{\nabla f(x, y) = \begin{bmatrix} 2xy + 3y^2 \\ x^2 + 6xy \end{bmatrix}}$$

5.3 Gradients of Vector-Valued Functions

Vector-valued functions are functions where the output is a vector rather than a scalar. That is:

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix} \quad \text{where } \mathbf{x} \in \mathbb{R}^n, \quad \mathbf{f}(\mathbf{x}) \in \mathbb{R}^m$$

—

Gradient generalization: The Jacobian matrix

The gradient of a vector-valued function is expressed using the **Jacobian matrix**, defined as:

$$\mathbf{J_f}(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

- Each row: gradient of a scalar function $f_i(\mathbf{x})$ - Each column: partials w.r.t. a variable x_j

—

Example:
Let:

$$\mathbf{f}(x, y) = \begin{bmatrix} f_1(x, y) \\ f_2(x, y) \end{bmatrix} = \begin{bmatrix} x^2 + y \\ xy \end{bmatrix}$$

Then the Jacobian is:

$$\mathbf{J_f}(x, y) = \begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \end{bmatrix} = \begin{bmatrix} 2x & 1 \\ y & x \end{bmatrix}$$

—

Interpretation:

- The Jacobian maps small changes in inputs ($d\mathbf{x}$) to changes in outputs ($d\mathbf{f}$).
- In machine learning, it's used for backpropagation in neural networks.
- When $m = 1$, the Jacobian reduces to the gradient row vector.

—

Final Result (Boxed):

$$\boxed{\mathbf{J_f}(x, y) = \begin{bmatrix} 2x & 1 \\ y & x \end{bmatrix}}$$

5.4 Gradients of Matrices

When dealing with functions whose inputs and/or outputs are matrices, the concept of gradient extends naturally but requires careful notation.

—

Function of a matrix:
Consider a scalar-valued function $f : \mathbb{R}^{m \times n} \to \mathbb{R}$, where the input is a matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$.

The **gradient of f with respect to $\mathbf{X}$** is the matrix of partial derivatives:

$$\nabla_{\mathbf{X}} f = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{12}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \frac{\partial f}{\partial x_{21}} & \frac{\partial f}{\partial x_{22}} & \cdots & \frac{\partial f}{\partial x_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \frac{\partial f}{\partial x_{m2}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

—

Interpretation:

- Each element $\frac{\partial f}{\partial x_{ij}}$ measures how f changes when the element x_{ij} changes, keeping all other elements fixed.

- The gradient points in the direction of steepest ascent of f in the space of matrices.

—

Example:
Let

$$f(\mathbf{X}) = \text{trace}(\mathbf{X}^T\mathbf{A}\mathbf{X}) = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{m} x_{ki}a_{ij}x_{kj}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is a constant matrix and $\mathbf{X} \in \mathbb{R}^{m\times n}$.

—

Gradient of f w.r.t. $\mathbf{X}$:

$$\nabla_{\mathbf{X}} f = \mathbf{X}(\mathbf{A} + \mathbf{A}^T)$$

If $\mathbf{A}$ is symmetric, this simplifies to

$$\nabla_{\mathbf{X}} f = 2\mathbf{X}\mathbf{A}$$

—

Final Result (Boxed):

$$\boxed{\nabla_{\mathbf{X}}\text{trace}(\mathbf{X}^T\mathbf{A}\mathbf{X}) = \mathbf{X}(\mathbf{A} + \mathbf{A}^T)}$$

—

Note: For matrix functions producing matrix outputs (e.g., $\mathbf{F}(\mathbf{X}) \in \mathbb{R}^{p\times q}$), the gradient generalizes to a *tensor* or a collection of Jacobians, which is more advanced.

5.5 Useful Identities for Computing Gradients

In matrix calculus and multivariate analysis, several standard identities simplify the computation of gradients.

Below are some commonly used results for scalar-valued functions $f(\mathbf{x})$ or $f(\mathbf{X})$:

—

Vector and matrix variables:

Let $\mathbf{x} \in \mathbb{R}^n$ be a vector, $\mathbf{X} \in \mathbb{R}^{m\times n}$ a matrix, and $\mathbf{A}, \mathbf{B}$ constant matrices of appropriate dimensions.

—

Identities:

1. $\nabla_{\mathbf{x}}(\mathbf{a}^T\mathbf{x}) = \mathbf{a}$ where $\mathbf{a} \in \mathbb{R}^n$
2. $\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$
3. $\nabla_{\mathbf{X}}\text{trace}(\mathbf{A}^T\mathbf{X}) = \mathbf{A}$
4. $\nabla_{\mathbf{X}}\text{trace}(\mathbf{X}^T\mathbf{A}\mathbf{X}) = \mathbf{A}\mathbf{X} + \mathbf{A}^T\mathbf{X}$
5. $\nabla_{\mathbf{X}}\|\mathbf{X}\|_F^2 = 2\mathbf{X}$
6. $\nabla_{\mathbf{x}}\|\mathbf{x}\|_2 = \dfrac{\mathbf{x}}{\|\mathbf{x}\|_2}$ for $\mathbf{x} \neq \mathbf{0}$
7. $\nabla_{\mathbf{x}}(\mathbf{b}^T\mathbf{x} + c) = \mathbf{b}$ where $\mathbf{b} \in \mathbb{R}^n, c \in \mathbb{R}$

—

Remarks:

- In identity 2, if $\mathbf{A}$ is symmetric, then $\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = 2\mathbf{A}\mathbf{x}$.
- The Frobenius norm is defined as $\|\mathbf{X}\|_F = \sqrt{\text{trace}(\mathbf{X}^T\mathbf{X})}$.
- These identities are widely used in optimization, statistics, and machine learning.

—

Final Summary (Boxed):

$$\boxed{\begin{aligned}\nabla_{\mathbf{x}}(\mathbf{a}^T\mathbf{x}) &= \mathbf{a} \\ \nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) &= (\mathbf{A} + \mathbf{A}^T)\mathbf{x} \\ \nabla_{\mathbf{X}}\text{trace}(\mathbf{A}^T\mathbf{X}) &= \mathbf{A} \\ \nabla_{\mathbf{X}}\text{trace}(\mathbf{X}^T\mathbf{A}\mathbf{X}) &= \mathbf{A}\mathbf{X} + \mathbf{A}^T\mathbf{X} \\ \nabla_{\mathbf{X}}\|\mathbf{X}\|_F^2 &= 2\mathbf{X}\end{aligned}}$$

5.6 Backpropagation and Automatic Differentiation

Backpropagation is an efficient algorithm for computing gradients of functions defined by computational graphs, commonly used in training neural networks.

Automatic Differentiation (AD) systematically applies the chain rule to compute derivatives of complex functions programmatically.

—

Key idea:

Given a function $y = f(\mathbf{x})$ composed of intermediate variables $\{v_i\}$, the chain rule allows gradients to be propagated backward from output to inputs:

$$\frac{\partial y}{\partial x_j} = \sum_i \frac{\partial y}{\partial v_i} \cdot \frac{\partial v_i}{\partial x_j}$$

Backpropagation exploits this by storing intermediate derivatives and reusing them efficiently.

—

Computational graph example:
Consider

$$z = f(x, y) = (x + y) \cdot (xy)$$

We can decompose it into operations:

$$\begin{cases} a = x + y \\ b = xy \\ z = a \cdot b \end{cases}$$

—

Forward pass: compute a, b, z.
Backward pass: compute gradients using chain rule:

$$\frac{\partial z}{\partial a} = b, \quad \frac{\partial z}{\partial b} = a$$

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial a}\frac{\partial a}{\partial x} + \frac{\partial z}{\partial b}\frac{\partial b}{\partial x} = b \cdot 1 + a \cdot y = b + ay$$

$$\frac{\partial z}{\partial y} = \frac{\partial z}{\partial a}\frac{\partial a}{\partial y} + \frac{\partial z}{\partial b}\frac{\partial b}{\partial y} = b \cdot 1 + a \cdot x = b + ax$$

—

Automatic Differentiation:
- AD breaks computations into elementary operations. - Each operation's derivative is known. - AD combines these derivatives using the chain rule to compute gradients accurately and efficiently. - It can be implemented in two modes: *forward mode* and *reverse mode* (backpropagation is reverse mode AD).

—

Summary:

Backpropagation efficiently computes $\nabla_{\mathbf{x}} f(\mathbf{x})$ via chain rule.
Automatic Differentiation automates this process for complex functions.

5.7 Higher Order Derivatives

Higher order derivatives extend the concept of differentiation to derivatives of derivatives.

—

Second-order derivatives:
For a scalar-valued function $f(x)$, the second derivative is:

$$\frac{d^2 f}{dx^2} = \frac{d}{dx}\left(\frac{df}{dx}\right)$$

For multivariate functions $f(\mathbf{x})$ where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, second-order derivatives are partial derivatives of partial derivatives:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right)$$

—

Hessian matrix:

The collection of all second-order partial derivatives forms the Hessian matrix $\mathbf{H} \in \mathbb{R}^{n \times n}$:

$$\mathbf{H}(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

If f is twice continuously differentiable, Clairaut's theorem states:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

making the Hessian symmetric.

—

Example:

Consider

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

First-order partial derivatives:

$$\frac{\partial f}{\partial x} = 6xy, \quad \frac{\partial f}{\partial y} = 3x^2 + 6y^2$$

Second-order partial derivatives:

$$\frac{\partial^2 f}{\partial x^2} = 6y, \quad \frac{\partial^2 f}{\partial y^2} = 12y, \quad \frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x} = 6x$$

Hessian matrix:

$$\mathbf{H}(f) = \begin{bmatrix} 6y & 6x \\ 6x & 12y \end{bmatrix}$$

—

Interpretation:

- The Hessian provides information about the curvature of f.

- It is used in optimization to classify critical points (minima, maxima, saddle points).

—

Final Result (Boxed):

$$\boxed{\mathbf{H}(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}}$$

5.8 Linearization and Multivariate Taylor Series

Linearization approximates a multivariate function near a point by its first-order Taylor expansion.

—

Linear approximation:

Given a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, the linearization of f around a point $\mathbf{a} \in \mathbb{R}^n$ is:

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \nabla f(\mathbf{a})^T (\mathbf{x} - \mathbf{a})$$

where $\nabla f(\mathbf{a})$ is the gradient of f at $\mathbf{a}$.

—

Multivariate Taylor series:

If f is k-times differentiable, the Taylor series expansion of f around $\mathbf{a}$ up to order k is:

$$f(\mathbf{x}) = \sum_{|\alpha| \leq k} \frac{D^\alpha f(\mathbf{a})}{\alpha!} (\mathbf{x} - \mathbf{a})^\alpha + R_k(\mathbf{x})$$

where:

- $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ is a multi-index with $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n$,
- $D^\alpha f(\mathbf{a}) = \left. \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_n^{\alpha_n}} \right|_{\mathbf{x}=\mathbf{a}}$,
- $\alpha! = \alpha_1! \alpha_2! \cdots \alpha_n!$,
- $(\mathbf{x} - \mathbf{a})^\alpha = (x_1 - a_1)^{\alpha_1} (x_2 - a_2)^{\alpha_2} \cdots (x_n - a_n)^{\alpha_n}$,
- $R_k(\mathbf{x})$ is the remainder term.

—

Second-order Taylor expansion:

Up to second order, the expansion is:

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \nabla f(\mathbf{a})^T (\mathbf{x} - \mathbf{a}) + \frac{1}{2} (\mathbf{x} - \mathbf{a})^T \mathbf{H}(f)(\mathbf{a}) (\mathbf{x} - \mathbf{a})$$

where $\mathbf{H}(f)(\mathbf{a})$ is the Hessian matrix of f at $\mathbf{a}$.

—

Example:
Let

$$f(x, y) = e^{xy}$$

Linearize f around $\mathbf{a} = (0, 0)$.

$$f(0, 0) = e^0 = 1$$

Gradient at $(0, 0)$:

$$\nabla f(0, 0) = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}_{(0,0)} = \begin{bmatrix} ye^{xy} \\ xe^{xy} \end{bmatrix}_{(0,0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

So the linearization is:

$$f(x, y) \approx 1 + [0 \quad 0] \begin{bmatrix} x - 0 \\ y - 0 \end{bmatrix} = 1$$

—

Interpretation: The linearization approximates f near $\mathbf{a}$ by a plane tangent to the function graph.

—

Final Result (Boxed):

$$\boxed{f(\mathbf{x}) \approx f(\mathbf{a}) + \nabla f(\mathbf{a})^T (\mathbf{x} - \mathbf{a})}$$

6 Probability Distribution

6.1 Construction of Probability Space

A probability space is a triplet $(\Omega, \mathcal{F}, P)$:

- Ω: sample space – set of all possible outcomes
- $\mathcal{F}$: sigma-algebra – collection of events (subsets of Ω)
- P: probability measure – assigns a value $P : \mathcal{F} \to [0, 1]$

Target space: Let $\mathcal{T} \subseteq \mathbb{R}^n$ be the space in which random variables take values. For a random variable $X : \Omega \to \mathcal{T}$, we define:

$$P_X(B) = P(X^{-1}(B)) \quad \text{for } B \subseteq \mathcal{T}$$

Example: Tossing a fair coin:

- $\Omega = \{\text{Heads}, \text{Tails}\}$
- $\mathcal{F} = \mathcal{P}(\Omega)$
- $P(\text{Heads}) = P(\text{Tails}) = 0.5$

—

6.2 Discrete and Continuous Probabilities

Discrete random variable:

- Has a countable number of outcomes
- Probability mass function (PMF): $P(X = x_i) = p_i$
- Must satisfy: $\sum_i p_i = 1$

Continuous random variable:

- Takes values in $\mathcal{T} \subseteq \mathbb{R}^n$
- Described by a probability density function (PDF) $p(x)$
- Must satisfy: $\int_{\mathcal{T}} p(x)\, dx = 1$

Example (Discrete): $X \sim \text{Bernoulli}(p) \Rightarrow P(X = 1) = p,\ P(X = 0) = 1 - p$

Example (Continuous): $X \sim \text{Uniform}(0, 1) \Rightarrow p(x) = 1$ for $x \in (0, 1)$

—

6.3 Sum Rule, Product Rule, Bayes' Theorem

Sum Rule:

$$P(A) = \sum_B P(A, B) = \sum_B P(A|B)P(B)$$

Product Rule:

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

Bayes' Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{if } P(B) > 0$$

Example: Suppose:

- $P(\text{Rain}) = 0.2$
- $P(\text{Umbrella}|\text{Rain}) = 0.9$
- $P(\text{Umbrella}|\text{No Rain}) = 0.1$

Then:

$$P(\text{Umbrella}) = 0.9 \cdot 0.2 + 0.1 \cdot 0.8 = 0.26$$

$$P(\text{Rain}|\text{Umbrella}) = \frac{0.9 \cdot 0.2}{0.26} \approx 0.692$$

—

6.4 Summary Statistics and Independence

Mean (Expectation):

$$\mathbb{E}[X] = \begin{cases} \sum_x x \cdot p(x) & \text{(discrete)} \\ \int_{\mathcal{T}} x \cdot p(x)\, dx & \text{(continuous)} \end{cases}$$

Variance:

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Covariance:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Independence: Two variables X and Y are independent if:

$$P(X, Y) = P(X) \cdot P(Y) \quad \text{or} \quad p(x, y) = p(x)p(y)$$

Example: Let $X, Y \sim \text{Uniform}(0, 1)$, independent. Then:

$$\mathbb{E}[X] = \mathbb{E}[Y] = 0.5, \quad \text{Var}(X) = \text{Var}(Y) = \frac{1}{12}$$

$$\text{Cov}(X, Y) = 0$$

—

6.5 Gaussian Distribution

Univariate Gaussian:

$$X \sim \mathcal{N}(\mu, \sigma^2), \quad p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Multivariate Gaussian:

$$\vec{X} \sim \mathcal{N}(\vec{\mu}, \mathbf{\Sigma}), \quad p(\vec{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^\top \mathbf{\Sigma}^{-1}(\vec{x}-\vec{\mu})\right)$$

Example: Let $X \sim \mathcal{N}(0, 1)$. Then:

$$\mathbb{E}[X] = 0, \quad \text{Var}(X) = 1$$

—

6.6 Conjugacy and the Exponential Family

Exponential Family Form:

$$p(x|\theta) = h(x) \exp\left(\eta(\theta)^\top T(x) - A(\theta)\right)$$

Where:

- $T(x)$: sufficient statistic
- $\eta(\theta)$: natural parameter
- $A(\theta)$: log-partition function
- $h(x)$: base measure

Conjugate prior: A prior $p(\theta)$ is conjugate to the likelihood $p(x|\theta)$ if the posterior is in the same family.

Example:

- Likelihood: $x \sim \text{Bernoulli}(\theta)$
- Prior: $\theta \sim \text{Beta}(\alpha, \beta)$
- Posterior: $\theta|x \sim \text{Beta}(\alpha + x, \beta + 1 - x)$

—

6.7 Change of Variables / Inverse Transform

Let $Y = g(X)$, where g is differentiable and invertible.

1D Change of Variables:

$$p_Y(y) = p_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|$$

Multivariate Case:

$$p_Y(\vec{y}) = p_X(\vec{x}) \cdot \left| \det \left(\frac{\partial \vec{x}}{\partial \vec{y}} \right) \right| \quad \text{(Jacobian determinant)}$$

Inverse Transform Sampling:

- Sample $u \sim \text{Uniform}(0, 1)$
- Set $x = F^{-1}(u)$
- Then $x \sim p(x)$

Example 1: Change of Variables (1D)
Let $X \sim \text{Uniform}(0, 1)$, and define $Y = -\log(X)$. Then:

$$x = e^{-y}, \quad \frac{dx}{dy} = -e^{-y} \quad \Rightarrow \quad p_Y(y) = e^{-y}, \quad y \geq 0$$

Example 2: Inverse Transform Sampling (Exponential)
Given CDF $F(x) = 1 - e^{-\lambda x}$, we invert:

$$x = F^{-1}(u) = -\frac{1}{\lambda} \ln(1 - u)$$

So to sample:

- Sample $u \sim \text{Uniform}(0, 1)$
- Set $x = -\frac{1}{\lambda} \ln(1 - u)$