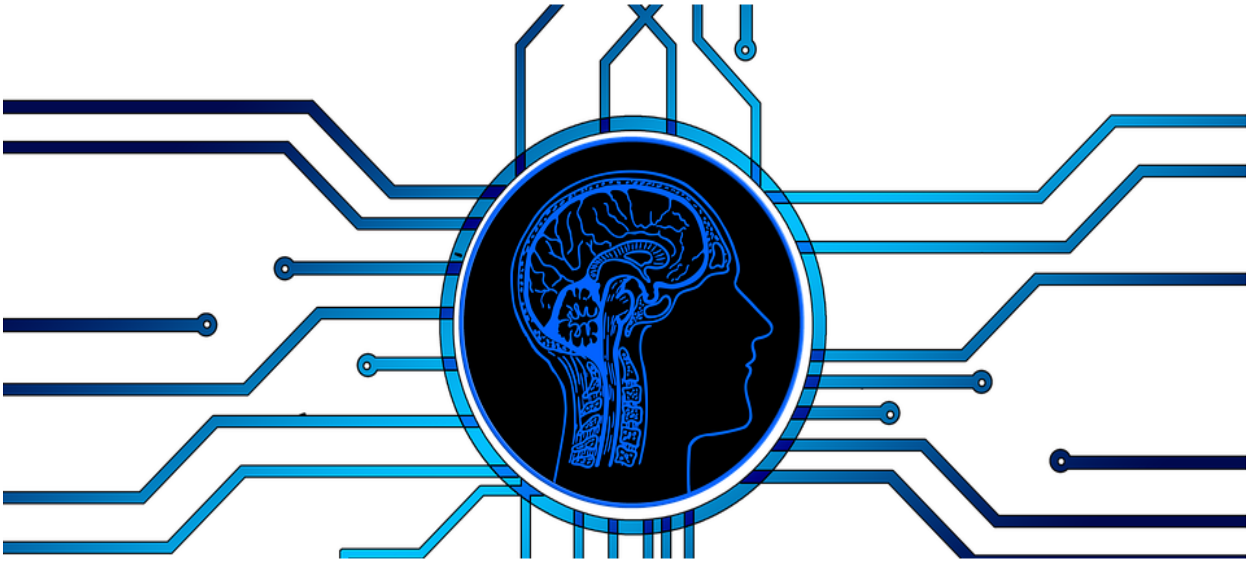
Advanced Applied Mathematics
- Machine Learning -

Ji Yong-Hyeon



Department of Information Security, Cryptology, and Mathematics
College of Science and Technology
Kookmin University

December 6, 2023

Contents

Chapter 1

Linear Algebra

1.1 Matrices

- A system of linear equations

$$\begin{cases} x_1, \ldots, x_n : \text{unknowns} \\ \#\text{ of unknowns} = n \\ \#\text{ of equations} = m \end{cases}$$

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \qquad\qquad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{cases}$$

$$\Longleftrightarrow \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} \qquad A\mathbf{x} = \mathbf{b}$$

$$\Longleftrightarrow x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} + \cdots + x_n \begin{bmatrix} a_{1m} \\ a_{2m} \\ \vdots \\ a_{mn} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \qquad x_1\mathbf{C}_1 + \cdots + x_n\mathbf{C}_n = \mathbf{b}$$

- Matrix operation

  (i) scalar multiplication: kA

  (ii) addition: $A + B$

  (iii) multiplication: AB

- Properties
  - Associative: $(A + B) + C = A + (B + C)$, $A(BC) = (AB)C$
  - Distributive: $(AB)C = A(BC)$

  - (in general) not commutative: $AB \neq BA$

- Transpose of A: A^T

$$(a_{ij})_{m\times n} \longrightarrow (a^t_{ij})_{n\times m} = (a_{ji})_{n\times m}$$

- Square Matrices

1.2 Solving Systems of Linear Equations

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

Remark 1.1. $A\mathbf{x} = \mathbf{b} \iff [A \mid \mathbf{b}]$.

Example 1.1.

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 2 \\ 2 & 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 5 \end{bmatrix} \iff \left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 1 & -1 & 2 & 2 \\ 2 & 0 & 3 & 5 \end{array}\right]$$

$$\xLeftrightarrow[R_3\leftarrow R_3-2R_1]{R_2\leftarrow R_2-R_1} \left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 0 & -2 & 1 & -1 \\ 0 & -2 & 1 & -1 \end{array}\right]$$

$$\xLeftrightarrow[R_2\leftarrow -\frac{1}{2}R_2]{R_3\leftarrow R_3-R_2} \left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 0 & 1 & -1/2 & 1/2 \\ 0 & 0 & 0 & 0 \end{array}\right] \quad \text{Row-Echelon Form (REF)}$$

$$\xLeftrightarrow{R_1\leftarrow R_1-R_2} \left[\begin{array}{ccc|c} 1 & 0 & 3/2 & 5/2 \\ 0 & 1 & -1/2 & 1/2 \\ 0 & 0 & 0 & 0 \end{array}\right] \quad \text{Reduced Row-Echelon Form (RREF)}$$

$$\iff \begin{cases} x_1 = -\frac{3}{2}x_3 + \frac{5}{2} \\ x_2 = \frac{1}{2}x_3 + \frac{1}{2} \end{cases}.$$

Let $x_3 = \lambda$ then

$$\mathbf{x} = \begin{bmatrix} -\frac{3}{2}\lambda + \frac{5}{2} \\ \frac{1}{2}\lambda + \frac{1}{2} \\ \lambda \end{bmatrix} = \begin{bmatrix} \frac{5}{2} \\ \frac{1}{2} \\ 0 \end{bmatrix} + \lambda \begin{bmatrix} -\frac{3}{2} \\ \frac{1}{2} \\ 1 \end{bmatrix}.$$

1.3 Vector Space

1.4 Linear Independence

1.5 Basis and Rank

1.6 Linear Mappings

Linear Mapping

Definition 1.1. Let V, W are vector spaces. A mapping

$$\begin{array}{rrcl} \Phi & : \quad V & \longrightarrow & W \\ & \lambda\mathbf{x} + \psi\mathbf{y} & \longmapsto & \Phi(\lambda\mathbf{x} + \psi\mathbf{y}) = \lambda\Phi(\mathbf{x}) + \psi\Phi(\mathbf{y}) \end{array}$$

is called a **linear mapping** (or **vector space homomorphism** / **linear transformation**).

Coordinate

Definition 1.2. Let V be a vector space with $\dim V = n$, and let $\mathscr{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ be a ordered basis of V. Then

$$\forall \mathbf{x} \in V : \exists \text{representation} : \quad \mathbf{x} = \sum_{i=1}^{n} \alpha_i \mathbf{b}_i = \begin{bmatrix} a_1 & \cdots a_n \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_n \end{bmatrix}.$$

Then $\begin{bmatrix} \alpha_1 & \cdots & \alpha_n \end{bmatrix}^T \in \mathbb{R}^n$ is a coordinate vector of $\mathbf{x}$ w.r.t. $\mathscr{B}$.

Example 1.2. Let $\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Then $\mathbf{x} = \begin{bmatrix} 2 \\ 3 \end{bmatrix} = 2\begin{bmatrix} 1 \\ 0 \end{bmatrix} + 3\begin{bmatrix} 0 \\ 1 \end{bmatrix} = 2\mathbf{e}_1 + 3\mathbf{e}_2$.

1.6.1 Matrix Representation of Linear Mappings

Transformation Matrix

Definition 1.3. Consider vector spaces V, W with corresponding (ordered basis) $\mathscr{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_n)$ and $\mathscr{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_m\}$. Let $\Phi : V \longrightarrow W$ be a linear mapping such that $\Phi(\mathbf{b}_j) = \sum_{i=1}^{m} \alpha_{ij}\mathbf{c}_i$. Let $A_\Phi = [\alpha_{ij}]_{m\times n}$. Note that

$$\begin{aligned} \Phi(\mathbf{x}) = \Phi(x_1\mathbf{b}_1 + \cdots + x_n\mathbf{b}_n) &= \sum_{i=1}^{n} x_i\Phi(\mathbf{b}_i) = \sum_{j=1}^{n} x_j \left(\sum_{i=1}^{m} \alpha_{ij}\mathbf{c}_i \right) \\ &= \begin{bmatrix} \sum_{j=1}^{n} \alpha_{ij}x_j \\ \vdots \\ \sum_{j=1}^{n} \alpha_{ij}x_j \end{bmatrix}_{\mathscr{C}} \\ &= \begin{bmatrix} \alpha_{11} & \cdots & \alpha_{1n} \\ \vdots & \ddots & \vdots \\ \alpha_{m1} & \cdots & \alpha_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}_{\mathscr{B}}. \end{aligned}$$

1.6.2 Basis Change

Basis Change

Theorem 1.1. *For a linear mapping $\Phi : V \longrightarrow W$, ordered bases*

$$\mathscr{B} = (\boldsymbol{b}_1, \dots, \boldsymbol{b}_n), \quad \tilde{\mathscr{B}} = (\tilde{\boldsymbol{b}}_1, \cdots \tilde{\boldsymbol{b}}_n)$$

of V and

$$\mathscr{C} = (\boldsymbol{c}_1, \dots, \boldsymbol{c}_m), \quad \tilde{\mathscr{C}} = (\tilde{\boldsymbol{c}}_1, \cdots \tilde{\boldsymbol{c}}_m)$$

of W, and a transformation matrix $\boldsymbol{A}_\Phi = \left[a_{ij}\right]_{m\times n}$ w.r.t. $\mathscr{B}$ and $\mathscr{C}$, the corresponding transformation matrix $\tilde{\boldsymbol{A}}_\Phi = \left[\tilde{a}_{ij}\right]_{m\times n}$ w.r.t. the bases $\tilde{\mathscr{B}}$ and $\tilde{\mathscr{C}}$ is given

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

$$\begin{tikzcd}
V \arrow[r, "\Phi"] & W \\
\mathscr{B} \arrow[r, "\boldsymbol{A}_\Phi"] & \mathscr{C} \\
\tilde{\mathscr{B}} \arrow[u, "\boldsymbol{S}"] \arrow[r, "\tilde{\boldsymbol{A}}_\Phi"] & \tilde{\mathscr{C}} \arrow[u, "\boldsymbol{T}"']
\end{tikzcd}
\qquad
\begin{tikzcd}
V \arrow[r, "\Phi"] & W \\
\mathscr{B} \arrow[r, "\boldsymbol{A}_\Phi"] & \mathscr{C} \arrow[d, "\boldsymbol{T}^{-1}"] \\
\tilde{\mathscr{B}} \arrow[u, "\boldsymbol{S}"] \arrow[r, "\tilde{\boldsymbol{A}}_\Phi"] & \tilde{\mathscr{C}}
\end{tikzcd}$$

Proof. Let

$$\mathbf{S} := \left[s_{ij}\right]_{n\times n} = \left[\tilde{\mathbf{b}}_1\ \tilde{\mathbf{b}}_2\ \cdots\ \tilde{\mathbf{b}}_n\right]_{\mathscr{B}}, \quad \text{and} \quad \mathbf{T} := \left[t_{lk}\right]_{m\times m} = \left[\tilde{\mathbf{c}}_1\ \tilde{\mathbf{c}}_2\ \cdots\ \tilde{\mathbf{c}}_m\right]_{\mathscr{C}}.$$

That is,

$$\tilde{\mathbf{b}}_j = \begin{bmatrix} s_{1j} \\ \vdots \\ s_{nj} \end{bmatrix}_{\mathscr{B}} = \sum_{i=1}^{n} s_{ij}\mathbf{b}_j \quad \text{and} \quad \tilde{\mathbf{c}}_k = \begin{bmatrix} t_{1k} \\ \vdots \\ t_{mk} \end{bmatrix}_{\mathscr{C}} = \sum_{l=1}^{m} t_{lk}\mathbf{c}_l$$

for $j = 1, \dots, n$ and $k = 1, \dots, m$, respectively. We must show that

$$\mathbf{T}\tilde{\mathbf{A}}_\Phi = \mathbf{A}_\Phi\mathbf{S} \in M_{m\times n}(\mathbb{R}).$$

(i) $(\mathbf{T}\tilde{\mathbf{A}}_\Phi)$ For $j = 1, 2, \dots, n$,

$$\Phi(\tilde{\mathbf{b}}_j) = \sum_{k=1}^{m} \tilde{a}_{kj}\tilde{\mathbf{c}}_k = \sum_{k=1}^{m} \left[\tilde{a}_{kj}\left(\sum_{l=1}^{m} t_{lk}\mathbf{c}_l\right)\right] = \sum_{l=1}^{m} \left[\left(\sum_{k=1}^{m} t_{lk}\tilde{a}_{kj}\right)\mathbf{c}_l\right].$$

(ii) $(\mathbf{A}_\Phi\mathbf{S})$ For $j = 1, 2, \dots, n$,

$$\Phi(\tilde{\mathbf{b}}_j) = \Phi\left(\sum_{i=1}^{n} s_{ij}\mathbf{b}_j\right) = \sum_{i=1}^{n} \left[s_{ij}\Phi(\mathbf{b}_i)\right] = \sum_{i=1}^{n} \left[s_{ij} \sum_{i=1}^{m} a_{li}\mathbf{c}_l\right] = \sum_{l=1}^{m}\left(\sum_{i=1}^{n} a_{li}s_{ij}\right)\mathbf{c}_l.$$

Hence

$$\sum_{k=1}^{m} t_{lk}\tilde{a}_{kj} = \sum_{i=1}^{n} a_{li}s_{ij} \implies \mathbf{T}\tilde{\mathbf{A}_\Phi} = \mathbf{A}_\Phi\mathbf{S} \implies \tilde{\mathbf{A}}_\Phi = \mathbf{T}^{-1}\mathbf{A}_\Phi\mathbf{S}.$$

□

Example 1.3. Let

$$y_1\mathbf{e}_1 + y_2\mathbf{e}_2 = \Phi(x_1\mathbf{e}_1 + x_2\mathbf{e}_2) = (x_1 + 5x_2)\mathbf{e}_1 + 6x_2\mathbf{e}_2.$$

Then

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathbf{A}_\Phi \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \text{where} \quad \mathbf{A}_\Phi = \begin{bmatrix} 1 & 5 \\ 0 & 6 \end{bmatrix}.$$

We define

$$\tilde{\mathscr{B}} = \begin{bmatrix} \tilde{\mathbf{b}}_1 \, \tilde{\mathbf{b}}_2 \end{bmatrix} := \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}.$$

$$\tilde{\mathbf{A}}_\Phi = \mathbf{T}^{-1}\mathbf{A}_\Phi\mathbf{S} = \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 0 & 5 \\ 0 & 6 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 6 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\Phi\left(\tilde{x}_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \tilde{x}_2 \begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) = 6\tilde{x}_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \tilde{x}_2 \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Similarity

Definition 1.4. Let $\mathbf{A}, \tilde{\mathbf{A}} \in M_{n\times n}(\mathbb{R})$. $\mathbf{A}, \tilde{\mathbf{A}}$ are **similar** if

$$\exists \mathbf{S} \in M_{n\times n}(\mathbb{R}) : \tilde{\mathbf{A}} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}.$$

1.6.3 Image and Kernel

Image and Kernel

Definition 1.5. Let $\Phi : V \to W$ be a linear mapping.

(1) The **kernel (null) space** is defined by

$$\ker(\Phi) := \Phi^{-1}(\mathbf{0}_W) = \left\{\mathbf{v} \in V : \Phi(\mathbf{v}) = \mathbf{0}_W\right\}.$$

(2) The **image (range)** is defined by

$$\operatorname{Im}(\Phi) := \Phi[V] = \left\{\mathbf{w} \in W : (\exists \mathbf{v} \in V)\, \Phi(\mathbf{v}) = \mathbf{w}\right\}.$$

Remark 1.2.

(1) $\mathbf{0}_V \in \ker(\Phi) \implies \ker\Phi \neq \emptyset$.

(2) $\ker(\Phi) \subseteq V$ is a subspace of V.

(3) $\operatorname{Im}(\Phi) \subseteq W$ is a subspace of W.

(4) $\Phi : V \rightarrowtail W \iff \ker(\Phi) = \{\mathbf{0}_V\}$.

Remark 1.3 (Null Space and Column Space)**.** Let $\mathbf{A} \in M_{m\times n}(\mathbb{R})$ and

$$\begin{array}{cccc} \Phi : & \mathbb{R}^n & \longrightarrow & \mathbb{R}^m \\ & \mathbf{x} & \longmapsto & \mathbf{A}\mathbf{x} \end{array}$$

(1) The **column space** is the image of Φ, the span of the columns of $\mathbf{A}$,

$$\begin{aligned} \operatorname{Im}(\Phi) = \left\{\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathbb{R}^n\right\} &= \left\{[\mathbf{a}_1, \dots, \mathbf{a}_n]\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} : x_i \in \mathbb{R}\right\} \\ &= \left\{\sum_{i=1}^{n} x_i \mathbf{a}_i : x_i \in \mathbb{R}\right\} \\ &= \operatorname{span}\langle \mathbf{a}_1, \dots, \mathbf{a}_n\rangle \subseteq \mathbb{R}^m. \end{aligned}$$

(2) $\operatorname{rank}(\mathbf{A}) = \dim(\operatorname{Im}(\Phi))$.

(3) The **null space** $\ker(\Phi)$ is $\{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{0}\}$.

Example 1.4 (Image and Kernel of Linear Mapping)**.** The mapping

$$\Phi : \mathbb{R}^4 \to \mathbb{R}^2 : \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \mapsto \begin{bmatrix} 1 & 2 & -1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} x_1 + 2x_2 - x_3 \\ x_1 + x_4 \end{bmatrix}$$
$$= x_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + x_2 \begin{bmatrix} 2 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} -1 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

is linear. Then

(1) $\operatorname{Im}(\Phi) = \operatorname{span}\left\langle \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\rangle = \mathbb{R}^2$

(2) Since

$$\begin{bmatrix} 1 & 2 & -1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \rightsquigarrow \cdots \rightsquigarrow \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & -\frac{1}{2} & -\frac{1}{2} \end{bmatrix} \xrightarrow{\text{Minus-1 Trick}} \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix},$$

we have

$$\ker(\Phi) = \operatorname{span}\left\langle \begin{bmatrix} 1 \\ -1/2 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ -1/2 \\ 0 \\ -1 \end{bmatrix} \right\rangle.$$

Rank-Nullity Theorem (Fundamental Theorem of Linear Mapping)

Theorem 1.2. *Let $\Phi : V \to W$ be a linear mapping for vector spaces V, W. Then*

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

1.7 Affine Spaces

$$\Phi(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$$

$\operatorname{Im}\Phi$ is not a subspace if $\mathbf{b} \neq 0$.

Chapter 2

Analytic Geometry

2.1 Norm

Norm

Definition 2.1. A **norm** on a vector space V is a function

$$\begin{array}{cccccc} \|\cdot\| & : & V & \longrightarrow & \mathbb{R} \\ & & \mathbf{x} & \longmapsto & \|\mathbf{x}\| \end{array}$$

such that for all $\lambda \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y} \in V$ the following hold:

(i) (Absolutely homogeneous) $\|\lambda x\| = |\lambda| \|\mathbf{x}\|$

(ii) (Triangle inequality) $\left\|\mathbf{x} + \mathbf{y}\right\| \leq \|\mathbf{x}\| + \left\|\mathbf{y}\right\|$

(iii) (Positive definite) $\begin{cases} \|\mathbf{x}\| > 0 & : \mathbf{x} \neq \mathbf{0} \\ \|\mathbf{x}\| = 0 & : \mathbf{x} = \mathbf{0} \end{cases}$

Example 2.1 (Manhattan Norm)**.** The Manhattan norm on $\mathbb{R}^n$ is defined for $\mathbf{x} \in \mathbb{R}^n$ as

$$\|\mathbf{x}\|_1 := \sum_{i=1}^{n} |x_i| .$$

The Manhattan norm is also called ℓ_1 norm.

Example 2.2 (Euclidean Norm)**.** The Manhattan norm on $\mathbb{R}^n$ is defined for $\mathbf{x} \in \mathbb{R}^n$ as

$$\|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\mathbf{x}^T\mathbf{x}}.$$

The Euclidean norm is also called ℓ_2 norm.

2.2 Inner Products

2.2.1 General Inner Product

Dot Product (Scalar Product)

Definition 2.2. The **dot product (scalar product)** in $\mathbb{R}^n$ is given by

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T \mathbf{y} = \sum_{i=1}^{n} x_i y_i .$$

Bilinaer Mapping

Definition 2.3. Let V be a vector space and $\Omega : V \times V \to \mathbb{R}$ is a **bilienar mapping** if for all $\alpha, \beta \in \mathbb{R}$,

(i) $\Omega(\alpha \mathbf{x}_1 + \beta \mathbf{x}_2, \mathbf{y}) = \alpha \Omega(\mathbf{x}, \mathbf{y}) + \beta \Omega(\mathbf{x}_2, y)$.

(ii) $\Omega(\mathbf{y}, \alpha \mathbf{y}_1 + \beta \mathbf{y}_2) = \alpha \Omega(\mathbf{x}, \mathbf{y}_1) + \beta \Omega(\mathbf{x}, y_2)$.

Remark 2.1.

(1) Ω is called **symmetric** if $\forall \mathbf{x}, \mathbf{y} \in V : \Omega(\mathbf{x}, \mathbf{y}) = \Omega(\mathbf{y}, \mathbf{x})$.

(2) Ω is called **positive definite** if $\begin{cases} \Omega(\mathbf{x}, \mathbf{x}) > 0 & : \mathbf{x} \in V \setminus \{\mathbf{0}\} \\ \Omega(\mathbf{x}, \mathbf{x}) = 0 & : \mathbf{x} = \mathbf{0}. \end{cases}$.

Inner Product

Definition 2.4. A positive definite, symmetric bilinear mapping $\Omega : V \times V \to \mathbb{R}$ is called an **inner product** on vector space V.

Example 2.3 (Inner Product That Is Not Dot Product)**.** Consider $V = \mathbb{R}^2$. We define

$$\langle \mathbf{x}, \mathbf{y} \rangle := x_1 y_1 - (x_1 y_2 + x_2 y_1) + 2 x_2 y_2 = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

Then

(i) (positive definite)

$$\langle \mathbf{x}, \mathbf{x} \rangle = x_1^2 - 2x_1 x_2 + x_2^2 + x_2^2 = (x_1 - x_2)^2 + x_2^2 \geq 0.$$

Moreover, $\langle \mathbf{x}, \mathbf{x} \rangle = 0 \iff \mathbf{x} = 0$.

(ii) (symmetric) It holds.

(iii) (bilinear) It holds.

2.2.2 Symmetric, Positive Definite Matrices

Symmetric, Positive Defintie Matrix

Definition 2.5. Let V be a vector space with $\dim V = n$. A symmetric matrix $\mathbf{A} \in M_{n\times n}(\mathbb{R})$ is called **symmetric, positive definite** if $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x} \in V$ and

$$\begin{cases} \mathbf{x}^T\mathbf{A}\mathbf{x} > 0 & : \mathbf{x} \in V \setminus \{\mathbf{0}\} \\ \mathbf{x}^T\mathbf{A}\mathbf{x} = 0 & : \mathbf{x} = \mathbf{0}. \end{cases}$$

Remark 2.2. $\mathbf{A}$ is positive **semi-definite** if $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ only.

Theorem 2.1. *Let V be a vector space with $\dim V = n$ and $\mathscr{B}$ an ordered basis of V. A bilinear mapping $\langle\cdot,\cdot\rangle : V \times V \to \mathbb{R}$ is an inner product if and only if*

$$\exists \textit{symmetric, positive definite matrix } \boldsymbol{A} \in M_{n\times n}(\mathbb{R}) : \langle \boldsymbol{x}, \boldsymbol{y}\rangle = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{y}.$$

Remark 2.3. Let $\mathbf{A}$ be a symmetric, positive definite matrix.

(1) $\ker \mathbf{A} = \{0\}$ because

$$\mathbf{x} \neq \mathbf{0} \implies \mathbf{x}^T\mathbf{A}\mathbf{x} > 0 \implies \mathbf{A}\mathbf{x} \neq \mathbf{0}.$$

(2) The diagonal element a_{ii} of $\mathbf{A}$ are positive because

$$a_{ii} = \mathbf{e}_i^T\mathbf{A}\mathbf{e}_i = \langle \mathbf{e}_i, \mathbf{e}_i\rangle > 0.$$

2.2.3 Lengths and Distances

Remark 2.4 (Cauchy-Schwarz Inequality)**.**

$$\left|\langle \mathbf{x}, \mathbf{y}\rangle\right| \leq ||\mathbf{x}||||\mathbf{y}||.$$

Distance and Metric

Definition 2.6. Consider an inner product space $(V, \langle\cdot,\cdot\rangle)$. Let $\mathbf{x}, \mathbf{y} \in V$. Then

$$d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}|| = \sqrt{\langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y}\rangle}.$$

is called **distance** between $\mathbf{x}$ and $\mathbf{y}$. The mapping

$$\begin{array}{rcccl} d & : & V \times V & \longrightarrow & \mathbb{R} \\ & & (\mathbf{x}, \mathbf{y}) & \longmapsto & d(\mathbf{x}, \mathbf{y}) \end{array}$$

is called a **metric**

2.2.4 Angles and Orthogonality

Angle

Definition 2.7. Assume that $\mathbf{x}, \mathbf{y} \in V \setminus \{\mathbf{0}\}$. Then

$$-1 \leq \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{||\mathbf{x}||||\mathbf{y}||} \leq 1.$$

And

$$\exists ! \theta \in [0, \pi] : \cos\theta = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{||\mathbf{x}||||\mathbf{y}||}.$$

The number θ is the **angle**.

Example 2.4. Consider $\mathbf{x} = (1, 1)$ and $\mathbf{y} = (-1, 1)$ on $\mathbb{R}^2$.

(1) Dot Product:

$$\mathbf{x} \cdot \mathbf{y} = (1, 1) \cdot (-1, 1) = -1 + 1 = 0.$$

(2) Inner Product:

$$\langle \mathbf{x}, \mathbf{y} \rangle := \mathbf{x}^T \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{y} \implies \cos\theta = -\frac{1}{3}.$$

Orthogonal Matrix

Definition 2.8. A square matrix $\mathbf{A} \in M_{n\times n}(\mathbb{R})$ is an **orthogonal matrix** if and only if

$$\mathbf{A}\mathbf{A}^T = I_n = \mathbf{A}^T\mathbf{A},$$

that is, $\mathbf{A}^{-1} = \mathbf{A}^T$.

Remark 2.5.

(1) $\mathbf{A}^T\mathbf{A} = [A_i^T A_j]_{n\times n} = \left[\langle \mathbf{A}_i, \mathbf{A}_j \rangle\right]_{n\times n}$, where $\langle \mathbf{A}_i, \mathbf{A}_j \rangle = \begin{cases} 1 & : i = j \\ 0 & : i \neq j. \end{cases}$

(i) Column vectors of $\mathbf{A}$ are orthogonal each other.

(ii) $\langle \mathbf{A}_i, \mathbf{A}_i \rangle = 1 \implies ||\mathbf{A}_i|| = 1$.

(2) Let $\mathbf{A}$ is orthogonal. Then a linear mapping

$$\begin{array}{rcccc} \Phi & : & \mathbb{R}^n & \longrightarrow & \mathbb{R}^m \\ & & \mathbf{x} & \longmapsto & \mathbf{A}\mathbf{x} \end{array}$$

has **length preserving** property, i.e., $||\mathbf{x}|| = ||\mathbf{A}\mathbf{x}||$ because

$$||\mathbf{A}\mathbf{x}||^2 = \langle \mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{x} \rangle = (\mathbf{A}\mathbf{x})^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\mathbf{x} = \langle \mathbf{x}, \mathbf{x} \rangle = ||\mathbf{x}||^2.$$

Φ has also **angle preserving** property because

$$\cos\theta = \frac{(\mathbf{A}\mathbf{x}^T)(\mathbf{A}\mathbf{y})}{||\mathbf{A}\mathbf{x}||||\mathbf{A}\mathbf{y}||} = \frac{\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{y}}{\sqrt{\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x}\mathbf{y}^T\mathbf{A}^T\mathbf{A}\mathbf{y}}} = \frac{\mathbf{x}^T\mathbf{y}}{||\mathbf{x}||||\mathbf{y}||}.$$

2.3 Orthonormal Basis

Orthnormal Bais

Definition 2.9. Consider an n-dimensional vector space V and a basis $\{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ of V. The basis is called an **orthonormal basis (ONB)** if

$$\langle \mathbf{b}_i, \mathbf{b}_j \rangle = \begin{cases} 0 & : i \neq j \\ 1 & : i = j \end{cases}, \quad \text{i.e.,} \quad \langle \mathbf{b}_i, \mathbf{b}_j \rangle = \delta_{ij}$$

for all $i, j = 1, \dots, n$.

Orthogonal Complement

Definition 2.10. Consider a d-dimensional vector space V and an m-dimensional subspace $U \subseteq V$. The **orthogonal complement** is

$$U^\perp := \left\{ \mathbf{v} \in V : (\forall \mathbf{u} \in U) \; \langle \mathbf{v}, \mathbf{u} \rangle = 0 \right\}$$

is a $(d - m)$-dimensional subspace of V.

Remark 2.6.

(1) $U \cap U^\perp = \{\mathbf{0}\}$.

(2) Any vector $\mathbf{x} \in V$ can be uniquely decomposed into

$$\mathbf{x} = \sum_{i=1}^{m} \lambda_m \mathbf{b}_m + \sum_{j=1}^{d-m} \psi_j \mathbf{b}_j^\perp, \quad \lambda_i, \psi_j \in \mathbb{R},$$

where $(\mathbf{b}_1, \dots, \mathbf{b}_m)$ is a basis of U and $(\mathbf{b}_1^\perp, \dots, \mathbf{b}_{d-m}^\perp)$ is a basis of $U^\perp$.

2.4 Orthogonal Projections

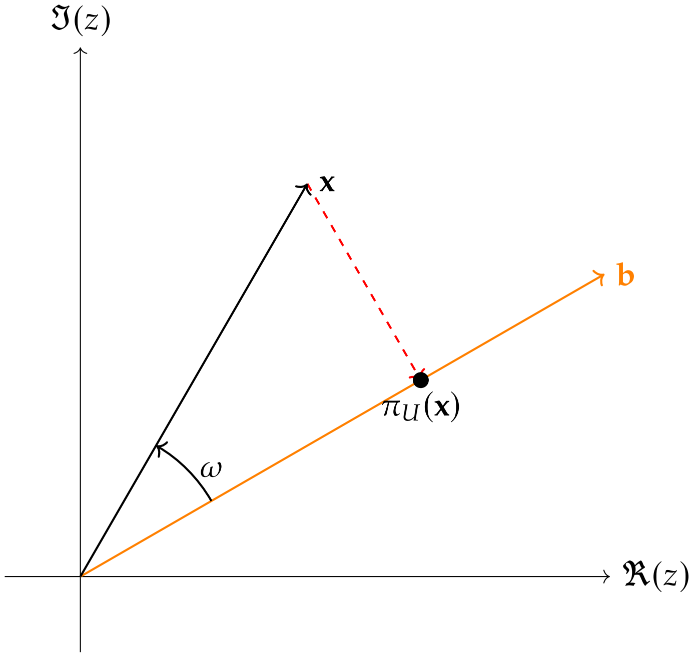
> “To minimize the compression loss, we have to find the most informative dimensions in the data”

Projection

Definition 2.11. Let V be a vector space and $U \subseteq V$ a subspace of V. A linear mapping $\pi : V \longrightarrow U$ is called a **projection** if

$$\pi^2 = \pi \circ \pi = \pi.$$

2.4.1 Projection onto One-Dimensional Subspaces (Lines)



We determine the coordinate λ, the projection $\pi_U(\mathbf{x}) \in U$, and the projection matrix $\mathbf{P}_\pi$ that maps any $\mathbf{x} \in \mathbb{R}^n$ onto U:

(Step 1) Finding the coordinate λ. $\pi_U \in U \Rightarrow \pi_U(\mathbf{x}) = \lambda\mathbf{b}$. Note that

$$\begin{aligned} 0 &= \langle \mathbf{x} - \pi_U(\mathbf{x}), \mathbf{b} \rangle \\ &= \langle \mathbf{x} - \lambda\mathbf{b}, \mathbf{b} \rangle \quad \because \pi_U(\mathbf{x}) = \lambda\mathbf{b} \\ &= \langle \mathbf{x}, \mathbf{b} \rangle - \lambda\langle \mathbf{b}, \mathbf{b} \rangle \quad \text{by bilinearity of the inner product.} \end{aligned}$$

Thus

$$\lambda = \frac{\langle \mathbf{x}, \mathbf{b} \rangle}{\langle \mathbf{b}, \mathbf{b} \rangle} = \frac{\langle \mathbf{b}, \mathbf{x} \rangle}{||\mathbf{b}||^2} = \frac{\mathbf{b}^T\mathbf{x}}{\mathbf{b}^T\mathbf{b}}.$$

If $||\mathbf{b}|| = 1$, then the coordinate λ of the projection is given by $\mathbf{b}^T\mathbf{x}$.

(Step 2) Finding the projection point $\pi_U(\mathbf{x}) \in U$ and the projection matrix $\mathbf{P}_\pi$. Note that

$$\langle \mathbf{b}, \mathbf{x} \rangle \mathbf{b} = \left(\mathbf{b}^T\mathbf{x}\right)\mathbf{b} = \left(\sum_j b_j x_j\right)\left(\sum_i b_i \mathbf{e}_i\right) = \sum_i \left(\sum_j b_i b_j x_j\right)\mathbf{e}_i = \sum_{ij} (\mathbf{b}\mathbf{b}^T)_{ij} x_i \mathbf{e}_i = \mathbf{b}\mathbf{b}^T\mathbf{x}$$

$$\pi_U(\mathbf{x}) = \lambda\mathbf{b} = \underbrace{\left(\frac{\langle \mathbf{x}, \mathbf{b} \rangle}{||\mathbf{b}||^2}\right)}_{\in\mathbb{R}} \mathbf{b} = \mathbf{P}_\pi\mathbf{x}, \quad \text{where} \quad \mathbf{P}_\pi = \left(\frac{\mathbf{b}\mathbf{b}^T}{||\mathbf{b}||^2}\right).$$

Example 2.5 (Projection onto a Line)**.** Find the projection matrix $\mathbf{P}_\pi$ onto the line through the origin spanned by $\mathbf{b} = \begin{bmatrix} 1 & 2 & 2 \end{bmatrix}^T$, where $\mathbf{b}$ is a direction and a basis of the one-dimensional subspace (line through origin).

Sol. Note that

$$\mathbf{b}\mathbf{b}^T = \begin{bmatrix}1\\2\\2\end{bmatrix}\begin{bmatrix}1 & 2 & 2\end{bmatrix} = \begin{bmatrix}1 & 2 & 2\\2 & 4 & 4\\2 & 4 & 4\end{bmatrix},$$

$$||\mathbf{b}||^2 = \mathbf{b}^T\mathbf{b} = \begin{bmatrix}1 & 2 & 2\end{bmatrix}\begin{bmatrix}1\\2\\2\end{bmatrix} = 1 + 2^2 + 2^2 = 9.$$

Thus

$$\mathbf{P}_\pi = \frac{\mathbf{b}\mathbf{b}^T}{\mathbf{b}^T\mathbf{b}} = \frac{1}{9}\begin{bmatrix}1 & 2 & 2\\2 & 4 & 4\\2 & 4 & 4\end{bmatrix}.$$

For $\mathbf{x} = \begin{bmatrix}1 & 1 & 1\end{bmatrix}^T \in \mathbb{R}^3$, the projection is

$$\pi_U(\mathbf{x}) = \mathbf{P}_\pi\mathbf{x} = \frac{1}{9}\begin{bmatrix}1 & 2 & 2\\2 & 4 & 4\\2 & 4 & 4\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix} = \frac{1}{9}\begin{bmatrix}5\\10\\10\end{bmatrix} \in \text{span}\left\langle\begin{bmatrix}1\\2\\2\end{bmatrix}\right\rangle.$$

□

2.4.2 Projection onto General Subspaces

Assume that

$$U = \text{span}\langle\mathbf{b}_1, \ldots, \mathbf{b}_m\rangle \subseteq V = R^n.$$

Then $\pi_U(\mathbf{x}) = \sum_{i=1}^m \lambda_i\mathbf{b}_i$.

We find the projection $\pi_U(\mathbf{x})$ and the projection matrix $\mathbf{P}_\pi$:

(Step 1) Find the coordinates $\lambda_1, \ldots, \lambda_m$ of projection w.r.t. the basis of U, such that the linear combination

$$\pi_U(\mathbf{x}) = \sum_{i=1}^m \lambda_i\mathbf{b}_i = \mathbf{B}\boldsymbol{\lambda} \quad \text{with}$$
$$\mathbf{B} = \begin{bmatrix}\mathbf{b}_1, \ldots, \mathbf{b}_m\end{bmatrix} \in M_{n\times m}(\mathbb{R}), \quad \boldsymbol{\lambda} = \begin{bmatrix}\lambda_1, \ldots, \lambda_m^T\end{bmatrix} \in \mathbb{R}^m$$

is closest to $\mathbf{x} \in \mathbb{R}^n$. We obtain m simulationeous conditions

$$\langle\mathbf{b}_1, \mathbf{x} - \pi_U(\mathbf{x})\rangle = \mathbf{b}_1^T(\mathbf{x} - \pi_U(\mathbf{x})) = 0$$
$$\vdots$$
$$\langle\mathbf{b}_m, \mathbf{x} - \pi_U(\mathbf{x})\rangle = \mathbf{b}_m^T(\mathbf{x} - \pi_U(\mathbf{x})) = 0$$

which, with $\pi_U(\mathbf{x}) = \mathbf{B}\boldsymbol{\lambda}$, can be written as

$$\mathbf{b}_1^T(\mathbf{x} - \mathbf{B}\boldsymbol{\lambda}) = 0$$
$$\vdots$$
$$\mathbf{b}_m^T(\mathbf{x} - \mathbf{B}\boldsymbol{\lambda}) = 0$$

such that we obtain a homogeneous linear equation system

$$\begin{aligned} \begin{bmatrix} \mathbf{b}_1^T \\ \vdots \\ \mathbf{b}_m^T \end{bmatrix} \begin{bmatrix} \mathbf{x} - \mathbf{B}\lambda \end{bmatrix} = \mathbf{0} &\iff \mathbf{B}^T(\mathbf{x} - \mathbf{B}\lambda) = 0 \\ &\iff \mathbf{B}^T\mathbf{B}\lambda = \mathbf{B}^T\mathbf{x}. \end{aligned}$$

Thus the coordinate (coefficient) is

$$\lambda = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{x}.$$

(Step 2) Find the projection $\pi_U(\mathbf{x}) \in U$.

$$\pi_U(\mathbf{x}) = \mathbf{B}\lambda = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{x}.$$

(Step 3) Find the projection $\mathbf{P}_\pi$.

$$\mathbf{P}_\pi = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T.$$

Example 2.6 (Projection onto a Two-dimensional Subspace)**.** For a subspace

$$U = \operatorname{span}\left\langle \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} \right\rangle \subseteq \mathbb{R}^3 \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix} \in \mathbb{R}^3,$$

find the coordinates λ of $\mathbf{x}$ in terms of the subspace U, the projection point $\pi_U(\mathbf{x})$ and the projection matrix $\mathbf{P}_\pi$.

Sol.

$$\mathbf{x} = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix} \implies \mathbf{P}_\pi\mathbf{x} = 5\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + (-3)\begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}.$$

□

2.4.3 Gram-Shmidt Orthogonalization

The *Gram-Schmidt orthogonalization* method iteratively constructs an orthogonal basis $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ from any basis $(\mathbf{b}_1, \ldots, \mathbf{b}_n)$ of V as follows:

$$\begin{aligned} \mathbf{u}_1 &:= \mathbf{b}_1 \\ \mathbf{u}_2 &:= \mathbf{b}_2 - \pi_{\operatorname{span}\langle \mathbf{u}_1 \rangle}(\mathbf{b}_2) \\ &\vdots \\ \mathbf{u}_k &:= \mathbf{b}_k - \pi_{\operatorname{span}\langle \mathbf{u}_1, \ldots, \mathbf{u}_{k-1} \rangle}(\mathbf{b}_k), \quad k = 2, \ldots, n. \end{aligned}$$

If we normalize $\mathbf{u}_k$ at each step, that is

$$\hat{\mathbf{u}}_k := \frac{\mathbf{u}_k}{||\mathbf{u}_k||},$$

we obtain an orthonormal basis.

Chapter 3

Matrix Decompositions

3.1 Determinant and Trace

Determinant

Definition 3.1. The **determinant** of a square matrix $\mathbf{A} \in M_{n\times n}(\mathbb{R})$ is a function

$$\begin{array}{rrcl} \det & : \quad M_{n\times n} & \longrightarrow & \mathbb{R} \\ & A & \longmapsto & \det(A) \end{array}.$$

Remark 3.1.

(1) $(n = 2)$

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \implies \det \mathbf{A} = ad - bc.$$

(2) $(n = 3)$

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

Theorem 3.1. *Let $A \in M_{n\times n}(\mathbb{R})$. $\exists A^{-1} \iff \det(A) \neq 0$.*

Upper and Lower Triangluar Matrix

Definition 3.2.

(1) **U** is an upper triangular matrix if $u_{ij} = 0$ for $i > j$.

(2) **L** is an lower triangular matrix if $l_{ij} = 0$ for $i < j$.

Remark 3.2. Note that $\det \mathbf{U} = \sum_{i=1}^{n} u_{ii}$ and $\det \mathbf{L} = \sum_{i=1}^{n} l_{ii}$.

Proposition 3.2.

(1) $\det(\boldsymbol{AB}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$

(2) $\det(\boldsymbol{A}) = \det(\boldsymbol{A}^T)$

(3) $\det(\boldsymbol{A}^{-1}) = \left[\det(\boldsymbol{A})\right]^{-1}$

(4) $\boldsymbol{B} = \boldsymbol{S}^{-1}\boldsymbol{AS} \implies \det(\boldsymbol{A}) = \det(\boldsymbol{B})$

(5) $\det(\lambda\boldsymbol{A}) = \lambda^n \det(\boldsymbol{A})$ *for* $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$

Theorem 3.3. *Let* $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$*. Then*

$$\det(\boldsymbol{A}) \neq 0 \iff \operatorname{rank}(\boldsymbol{A}) = n.$$

In other words, $\boldsymbol{A}$ *is invertible if and only if it is full rank.*

Trace

Definition 3.3. The **trace** of a square matrix $\mathbf{A} \in M_{n\times n}(\mathbb{R})$ is defined as

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

i.e., the trace is the sum of the diagonal elements of **A**.

Proposition 3.4.

(1) $\operatorname{tr}(\boldsymbol{A}+\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{B})$ *for* $\boldsymbol{A}, \boldsymbol{B} \in M_{n\times n}(\mathbb{R})$

(2) $\operatorname{tr}(\alpha\boldsymbol{A}) = \alpha\operatorname{tr}(\boldsymbol{A})$ *for* $\alpha \in \mathbb{R}, \boldsymbol{A} \in M_{n\times n}(\mathbb{R})$

(3) $\operatorname{tr}(\boldsymbol{I}_n) = n$

(4) $\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA})$ *for* $\boldsymbol{A} \in M_{n\times k}(\mathbb{R}), \boldsymbol{B} \in M_{k\times n}(\mathbb{R})$

Proof. (4) Let $\mathbf{A} = [a_{ij}]_{n\times k}$ and $\mathbf{B} = [b_{ij}]_{k\times n}$, and let

$$\mathbf{AB} := \mathbf{C} = [c_{ij}]_{n\times n} \quad \text{with} \quad c_{ij} = \sum_{l=1}^{n} a_{il}b_{lj},$$
$$\mathbf{BA} := \mathbf{D} = [d_{ij}]_{k\times k} \quad \text{with} \quad d_{ij} = \sum_{l=1}^{k} b_{il}a_{lj}.$$

Then

$$\operatorname{tr}(\mathbf{AB}) = \sum_{l=1}^{m} c_{ll}$$

□

Charateristic Polynomial

Definition 3.4. Let $\lambda \in \mathbb{R}$ and $\mathbf{A} \in M_{n \times n}(\mathbb{R})$. Then

$$p_{\mathbf{A}}(\lambda) := \det(\mathbf{A} - \lambda \mathbf{I}_n) = \sum_{i=0}^{n} c_i \lambda^n \quad \text{with} \quad c_i = \begin{cases} \det(\mathbf{A}) & : i = 0 \\ (-1)^n \operatorname{tr}(\mathbf{A}) & : i \in (0, n) \\ (-1)^n & : i = n \end{cases}$$

is the **characteristic polynomial** of **A**.

3.2 Eigenvalues and Eigenvectors

3.2.1 Eigenvalues and Eigenvectors

Eigenvalue and Eigenvetor

Definition 3.5. Let $\mathbf{A} \in M_{n\times n}(\mathbb{R})$. Then $\lambda \in \mathbb{R}$ is an **eigenvalue** of $\mathbf{A}$ and $\mathbf{v} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ is the corresponding eigenvector of $\mathbf{A}$ if

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}.$$

Theorem 3.5. *TFAE(The following are equivalent):*

(1) λ *is an eigenvalue of* $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$.

(2) $\exists \boldsymbol{v} \in \mathbb{R}^n \setminus \{\mathbf{0}\} : \boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$.

(3) $\operatorname{rank}(\boldsymbol{A} - \lambda\boldsymbol{I}_n) < n$.

(4) $\det(\boldsymbol{A} - \lambda\boldsymbol{I}_n) = 0$.

Theorem 3.6. $\lambda \in \mathbb{R}$ *is an eigenvalue of* $\boldsymbol{A} \iff \lambda$ *is a root of the characteristic polynomial* $p_{\boldsymbol{A}}(\lambda)$ *of* $\boldsymbol{A}$.

Example 3.1 (Computing Eigenvalue, Eigenvectors, and Eigenspaces)**.** Find the eigenvalues and eigenvectors of the 2×2 matrix

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

Sol. **(Step 1) Characteristic Polynomial and Eigenvalues.**

$$p_{\mathbf{A}}(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}_2) = \begin{vmatrix} 4-\lambda & 2 \\ 13-\lambda & \end{vmatrix} = (4-\lambda)(3-\lambda) - 2 = \lambda^2 - 7\lambda + 10$$
$$= (\lambda - 2)(\lambda - 5).$$

Thus, we obtain roots $\lambda_1 = 2$ and $\lambda_2 = 5$.

(Step 2) Eigenvalues and Eigenspaces. We solve $\begin{bmatrix} 4-\lambda & 2 \\ 1 & 3-\lambda \end{bmatrix} \mathbf{x} = \mathbf{0}$.

(i) $(\lambda_1 = 2)$

$$\begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \implies \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \implies C(\lambda_1) = \operatorname{span}\left\langle \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\rangle.$$

(ii) $(\lambda_1 = 5)$

$$\begin{bmatrix} -1 & 2 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \implies \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \implies C(\lambda_2) = \operatorname{span}\left\langle \begin{bmatrix} 2 \\ 1 \end{bmatrix} \right\rangle.$$

□

Defective

Definition 3.6. A square matrix $\mathbf{A} \in M_{n\times n}(\mathbb{R})$ is **defective** if it possesses fewer than n linearly independent eigenvectors.

Remark 3.3.

(1) $\mathbf{A}$ has n distinct eigenvalue $\Rightarrow$ $\mathbf{A}$ is not defective.

(2) For a defective matrix $\mathbf{A} \in M_{n\times n}(\mathbb{R})$, the sum of the dimension of eigenspaces $< n$.

(3) A defective matrix has at lest one eigenvalue λ_i with an algebraic multiplicity $m > 1$ and a geometric multiplicity of less than m. Note that

$$\text{“Algebraic Multiplicity”} \geq \text{“Geometric Multiplicity”}$$

(4) $\mathbf{A}$ is defective iff $\sum_i \dim C(\lambda_i) \neq n$.

Theorem 3.7.

(1) $\mathbf{A}, \mathbf{A}^T$ have the same eigenvalues.

(2) Similar matrices have the same eigenvalues.

(3) Symmetric, positive definite matrices always have positive real eigenvalues.

Proof. (1) Since $(\mathbf{A} - \lambda I)^T = \mathbf{A}^T - \lambda I$ and $\det(\mathbf{A}) = \det(\mathbf{A}^T)$,

$$\det(\mathbf{A}^T - \lambda\mathbf{I}) = \det((\mathbf{A} - \lambda\mathbf{I})^T) = \det(\mathbf{A} - \lambda\mathbf{I}).$$

(2) Let $\hat{\mathbf{A}} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}$. Since

$$\hat{\mathbf{A}} - \lambda\mathbf{I} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S} - \mathbf{S}^{-1}\lambda\mathbf{I}\mathbf{S} = \mathbf{S}^{-1}[\mathbf{A} - \lambda\mathbf{I}]\mathbf{S},$$

we have

$$\det(\hat{\mathbf{A}} - \lambda\mathbf{I}) = \det(\mathbf{S}^{-1}[\mathbf{A} - \lambda\mathbf{I}]\mathbf{S}) = \det(\mathbf{S}^{-1})\det(\mathbf{A} - \lambda\mathbf{I})\det(\mathbf{S}) = \det(\mathbf{A} - \lambda\mathbf{I}).$$

(3) Let $\mathbf{A}$ is symmetric, positive definite matrix. Let $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. Then

$$\mathbf{x}^T A\mathbf{x} = \mathbf{x}^T\lambda\mathbf{x} = \lambda||\mathbf{x}|| \geq 0.$$

Since $\mathbf{x} \neq 0 \implies ||\mathbf{x}|| > 0 \wedge \mathbf{x}^T\mathbf{A}\mathbf{x} > 0$, we have $\lambda > 0$.

□

Example 3.2 (Defective Matrix)**.** Let

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

Then

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 2-\lambda & 1 & 0 \\ 0 & 2-\lambda & 0 \\ 0 & 0 & 3-\lambda \end{vmatrix} = (3-\lambda)(2-\lambda)^2 = 0 \implies \begin{cases} \lambda_1 = 3 \\ \lambda_2 = 2. \end{cases}$$

And so

$$(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{x}_1 = 0 \iff \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{x}_1 = \mathbf{0} \qquad \implies \mathbf{x}_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

$$(\mathbf{A} - \lambda_2\mathbf{I})\mathbf{x}_2 = 0 \iff \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{x}_1 = \mathbf{0} \qquad \implies \mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

Example 3.3. Let

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \in M_{2\times2}(\mathbb{R}).$$

Then

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} -\lambda & 1 \\ -1 & -\lambda \end{vmatrix} = \lambda^2 + 1 = 0.$$

And so

$$(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{x}_1 = \mathbf{0} \iff \begin{bmatrix} -i & 1 \\ -1 & -i \end{bmatrix} \mathbf{x}_1 = \mathbf{0} \qquad \implies \mathbf{x}_1 = \begin{bmatrix} 1 \\ i \end{bmatrix},$$

$$(\mathbf{A} - \lambda_2\mathbf{I})\mathbf{x}_2 = \mathbf{0} \iff \begin{bmatrix} i & 1 \\ -1 & i \end{bmatrix} \mathbf{x}_1 = \mathbf{0} \qquad \implies \mathbf{x}_1 = \begin{bmatrix} 1 \\ -i \end{bmatrix}.$$

Example 3.4. Let

$$\mathbf{A} = \begin{bmatrix} 2 & 3-3i \\ 3+3i & 5 \end{bmatrix} \in M_{2\times2}(\mathbb{C}).$$

Then

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 2-\lambda & 3-3i \\ 3+3i & 5-\lambda \end{vmatrix} = \lambda^2 - 7\lambda - 8 = (\lambda+1)(\lambda-8) = 0 \implies \begin{cases} \lambda_1 = 8 \\ \lambda_2 = -1. \end{cases}$$

And so

$$(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{x}_1 = \mathbf{0} \iff \begin{bmatrix} -6 & 3-3i \\ 3+3i & -3 \end{bmatrix} \mathbf{x}_1 = \mathbf{0} \qquad \implies \mathbf{x}_1 = \begin{bmatrix} 1 \\ 1+i \end{bmatrix},$$

$$(\mathbf{A} - \lambda_2\mathbf{I})\mathbf{x}_2 = \mathbf{0} \iff \begin{bmatrix} 3 & 3-3i \\ 3+3i & 6 \end{bmatrix} \mathbf{x}_1 = \mathbf{0} \qquad \implies \mathbf{x}_1 = \begin{bmatrix} 1-i \\ -i \end{bmatrix}.$$

3.2.2 Complex Matrices

Consider complex vector

$$\mathbf{w} = (w_1, \ldots, w_n) \in \mathbb{C}^n \quad \text{with} \quad w_j = x_j + iy_j,$$

where $x_j, y_j \in \mathbb{R}$. Then

(1) Norm:

$$||\mathbf{w}||^2 = \sum_{j=1}^{n} \left|w_j\right|^2 \quad \text{with} \quad \left|w_j\right| = \sqrt{x_j^2 + y_j^2}.$$

(2) Inner Product: For $\mathbf{w}, \mathbf{z} \in \mathbb{C}^n$,

$$\langle \mathbf{w}, \mathbf{z} \rangle := \overline{\mathbf{w}^T}\mathbf{z} = \sum_{j=1}^{n} \overline{w}_j z_j.$$

Note that $\langle \mathbf{z}, \mathbf{z} \rangle = \sum_{j=1}^{n} \overline{z}_j z_j = \sum_{j=1}^{n} \left|z_j\right|^2 = ||\mathbf{z}||^2$.

Hermition

Definition 3.7. Let $\mathbf{A} \in M_{n\times n}(\mathbb{C})$. Then

$$\mathbf{A}^H : \overline{\mathbf{A}}^T.$$

is called **Hermition** of $\mathbf{A}$.

Example 3.5.

$$\mathbf{A} = \begin{bmatrix} 1 & 1+i \\ 1-i & i \end{bmatrix} \implies \overline{\mathbf{A}} = \begin{bmatrix} 1 & 1-i \\ 1+i & -i \end{bmatrix} \implies \mathbf{A}^H = \overline{\mathbf{A}}^T = \begin{bmatrix} 1 & 1+i \\ 1-i & -i \end{bmatrix}.$$

Hermitian Matrix

Definition 3.8. $\mathbf{A}$ is a **Hermitian matrix** if $\mathbf{A} = \mathbf{A}^H$.

Remark 3.4.

(1) A real symmetric matrix $\mathbf{A}$ is a Hermitian matrix.

(2) A Hermitian matrix has real eigenvalues.

H1

Theorem 3.8. $A = A^H \implies (\forall x \in \mathbb{C}^n)\; x^H A x \in \mathbb{R}$.

Proof. Suppose that $\mathbf{A} = \mathbf{A}^H$. Let $\mathbf{y} := \mathbf{x}^H\mathbf{A}\mathbf{x}$. We must show that

$$\mathbf{y} = \mathbf{y}^H, \quad \text{i.e.,} \quad \mathbf{y} = \overline{\mathbf{y}} \; (\Rightarrow \mathbf{y} \in \mathbb{R}).$$

$$\mathbf{y}^H = \left(\mathbf{x}^H\mathbf{A}\mathbf{x}\right)^H = \mathbf{x}^H\mathbf{A}^H(\mathbf{x}^H)^H = \mathbf{x}^H\mathbf{A}\mathbf{x} = \mathbf{y}.$$

□

H2

Theorem 3.9. *If A is Hermitian, then every eigenvalue is real.*

Proof. Let $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ with $\mathbf{v} \neq \mathbf{0}$. By Theorem H1,

$$\mathbf{v}^H\mathbf{A}\mathbf{v} = \mathbf{v}^H(\lambda\mathbf{v}) = \lambda\mathbf{v}^H\mathbf{v} = \lambda||\mathbf{v}||^2 \implies \lambda = \frac{\mathbf{v}^H\mathbf{A}\mathbf{v}}{||\mathbf{v}||^2} \in \mathbb{R}.$$

□

H3

Theorem 3.10. *If $A \in M_{n\times n}(\mathbb{C})$ is Hermitian, then two eigenvectors corresponding to different eigenvalues are orthogonal.*

Proof. For a Hermitian matrix $\mathbf{A}$, let

$$\mathbf{A}\mathbf{v}_1 = \lambda_1\mathbf{v}_1, \quad \mathbf{A}\mathbf{v}_2 = \lambda_2\mathbf{v}_2$$

with $\lambda_1 \neq \lambda_2$. Then

$$\begin{aligned}\mathbf{v}_1\mathbf{A}\mathbf{v}_2 &= \mathbf{v}_1^H\lambda_2\mathbf{v}_2 = \lambda_2\mathbf{v}_1^H\mathbf{v}_2,\\ \mathbf{v}_1\mathbf{A}\mathbf{v}_2 &= \mathbf{v}_1^H\mathbf{A}^H\mathbf{v}_2 = (\mathbf{A}\mathbf{v}_1)^H\mathbf{v}_2 = (\lambda_1\mathbf{v}_1)^H\mathbf{v}_2 = \lambda_1\mathbf{v}_1^H\mathbf{v}_2.\end{aligned}$$

Thus,

$$\begin{aligned}&\lambda_1\mathbf{v}_1^H\mathbf{v}_2 = \lambda_2\mathbf{v}_1^H\mathbf{v}_2\\ \iff &\lambda_1\langle\mathbf{v}_1,\mathbf{v}_2\rangle - \lambda_2\langle\mathbf{v}_1,\mathbf{v}_2\rangle = 0\\ \iff &(\lambda_1-\lambda_2)\langle\mathbf{v}_1,\mathbf{v}_2\rangle = 0\\ \iff &\langle\mathbf{v}_1,\mathbf{v}_2\rangle = 0 \quad \because \lambda_1 \neq \lambda_2\\ \iff &\mathbf{v}_1 \perp \mathbf{v}_2.\end{aligned}$$

□

Spectral Theorem

Spectral Theorem

Theorem 3.11. *Let $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$ is symmetric. Then*

$$\exists \textit{orthonormal basis of the corresponding vector space } V \textit{ consisting of}$$

eigenvalues of $\boldsymbol{A}$*, and each eigenvalue is real.*

Proof. By Theorem H1, every eigenvalue is real. We remain to show that eigenvalues generate orthonormal basis.

(i) All eigenvalues are distinct, say, $\lambda_1 \neq \lambda_2 \neq \cdots \neq \lambda_n$. By Theorem H3,

$$\mathbf{v}_i \neq \mathbf{v}_j \quad \text{if} \quad i \neq j.$$

Then $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ is orthogonal basis of $\mathbb{R}^n$.

(ii) $\lambda_1, \lambda_2, \dots, \lambda_k$ are distinct k eigenvalues with $k < n$. Consider

$$\begin{aligned}
C(\lambda_1) &:= \text{span}\langle \mathbf{v}_{1,1}, \mathbf{v}_{1,2}, \dots, \mathbf{v}_{1,n_1} \rangle, \\
C(\lambda_2) &:= \text{span}\langle \mathbf{v}_{2,1}, \mathbf{v}_{2,2}, \dots, \mathbf{v}_{1,n_2} \rangle, \\
&\vdots \\
C(\lambda_k) &:= \text{span}\langle \mathbf{v}_{k,1}, \mathbf{v}_{k,2}, \dots, \mathbf{v}_{1,n_k} \rangle.
\end{aligned}$$

By Gram-Schmidt orthogonalization process, we have orthogonal basis of $C(\lambda_i)$ as follows:

$$\left\{\mathbf{w}_{1,1}, \cdots, \mathbf{w}_{1,n}, \cdots, \mathbf{w}_{k,1}, \cdots, \mathbf{w}_{k,n_k}\right\}.$$

Note that

$$\sum_{i=1}^{k} \dim C(\lambda_i) = n_1 + \cdots + n_k = n$$

if $\mathbf{A}$ is Hermitian.

□

Spectral Decomposition

Theorem 3.12. *Let A be a real symmetric. Then*

$$A = PDP^T,$$

where $\mathbf{D} = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$ *is diagonal and* $\mathbf{P}$ *orthogonal matrix.*

Proof. Let $\lambda_1, \ldots, \lambda_n$ are solutions, counting multiplicity, of $\det(\mathbf{A} - \lambda \mathbf{I}_n) = 0$, and let $\mathbf{v}_1, \cdots, \mathbf{v}_n$ are eigenvectors corresponding to $\lambda_1, \ldots, \lambda_n$, respectively. Since $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is orthogonal basis of $\mathbb{R}^n$,

$$\mathbf{P} := \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{bmatrix}$$

be a orthogonal matrix, and so $\mathbf{P} = \mathbf{P}^T$. Then

$$\begin{aligned} \mathbf{AP} = \mathbf{A} \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{bmatrix} = \begin{bmatrix} \mathbf{A}\mathbf{v}_1 & \cdots & \mathbf{A}\mathbf{v}_n \end{bmatrix} &= \begin{bmatrix} \lambda_1 \mathbf{v}_1 & \cdots & \lambda_n \mathbf{v}_n \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} \\ &= \mathbf{PD}. \end{aligned}$$

Hence

$$\mathbf{AP} = \mathbf{PD} \implies \mathbf{A} = \mathbf{PDP}^{-1} = \mathbf{PDP}^T.$$

□

Remark 3.5. Let $\mathbf{A}$ be a real symmetric matrix. Then

$$\begin{aligned} \mathbf{A} = \mathbf{PDP}^T &= \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_n^T \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{v}_1 \lambda_1 & \ldots & \lambda_n \mathbf{v}_n \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_n^T \end{bmatrix} \\ &= \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^T. \end{aligned}$$

- We call $\lambda_i [\mathbf{v}_i \mathbf{v}_i^T]$ the principal component as an approximation of $\mathbf{A}$.

Cholesky Decomposition

Theorem 3.13. *Let A be a symmetric, positive definite matrix. Then*

$$A = LL^T,$$

where $\mathbf{L}$ is a lower triangular matrix with positive diagonal elements.

Proof. Let $\mathbf{A}v_i = \lambda_i v_i$ with $v_i \neq 0$ for $i = 1, \dots, n$. By spectral decomposition, we have

$$\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^T = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}.$$

Note that

$$\begin{aligned} \mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^T &= \mathbf{P}\sqrt{\mathbf{D}}\sqrt{\mathbf{D}}\mathbf{P}^T \\ &= \mathbf{P}\sqrt{\mathbf{D}}\sqrt{\mathbf{D}}^T\mathbf{P}^T \\ &= (\mathbf{P}\sqrt{\mathbf{D}})(\mathbf{P}\sqrt{\mathbf{D}})^T \\ &= \mathbf{L}\mathbf{L}^T. \end{aligned}$$

□

3.3 Eigendecomposition and Diagonalization

Diagonalizable

Definition 3.9. A matrix $\mathbf{A} \in M_{n \times n}(\mathbb{R})$ is **diagonalizable** if

$$\exists \mathbf{P} \in M_{n \times n}(\mathbb{R}) : \mathbf{D} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P},$$

i.e., if it is similar to a diagonal matrix.

Eigendecomposition

Theorem 3.14. *A square matrix $A \in M_{n \times n}(\mathbb{R})$ can be factorized into*

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

where $P \in M_{n \times n}(\mathbb{R})$ and D is a diagonal matrix whose diagonal entries are the eigenvalues of A, if and only if the eigenvectors of A form a basis of $\mathbb{R}^n$.

Proof. Let $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\mathbf{A}$ and $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are corresponding eigenvectors of $\mathbf{A}$. Let $\mathbf{P} = \begin{bmatrix} \mathbf{v}_1 & \ldots & \mathbf{v}_n \end{bmatrix}$ and $\mathbf{D} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$. Then

$$\begin{aligned} \mathbf{AP} &= \begin{bmatrix} \mathbf{Av}_1 & \cdots & \mathbf{Av}_n \end{bmatrix}, \\ \mathbf{PD} &= \begin{bmatrix} \mathbf{v}_1 & \ldots & \mathbf{v}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} = \begin{bmatrix} \lambda_1 \mathbf{v}_1 & \cdots & \lambda_n \mathbf{v}_n \end{bmatrix}. \end{aligned}$$

Since $\mathbf{Av}_i = \lambda_i \mathbf{v}_i$ for all $i = 1, \ldots, n$, we have

$$\mathbf{AP} = \mathbf{PD} \implies \mathbf{A} = \mathbf{APP}^{-1}.$$

□

3.4 Singular Value Decomposition

SVD Theorem

Theorem 3.15. *Let $A \in M_{m\times n}(\mathbb{R})$ be a rectangular matrix of rank $r \in \big[0, \min(m,n)\big]$. The SVD of $\boldsymbol{A}$ is a decomposition of the form*

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$$

with

(i) *an orthogonal matrix $\boldsymbol{U} \in M_{m\times m}$ with column vectors $\boldsymbol{u}_i$ for $i = 1, \dots, m$,*

(ii) *and an orthogonal matrix $\boldsymbol{V} \in M_{n\times n}$ with column vectors $\boldsymbol{v}_j$ for $j = 1, \dots, n$.*

(iii) *Moreover, $\boldsymbol{\Sigma} \in M_{m\times n}(\mathbb{R})$ with $\Sigma_{ii} = \begin{cases} \sigma_i \geq 0 & : i = j, \\ 0 & : i \neq j. \end{cases}$*

Remark 3.6.

$$\begin{tikzcd} \mathbb{R}^n \arrow[r, "\mathbf{V}^T", "\text{basis change}"'] & \mathbb{R}^n \arrow[r, "\mathbf{\Sigma}", "\text{scaling(embedding/projection)}"'] & \mathbb{R}^m \arrow[r, "\mathbf{U}", "\text{basis change}"'] & \mathbb{R}^m \end{tikzcd}$$

Remark 3.7.

(1) Since $\mathbf{U}$ is orthogonal, $\mathbf{U}\mathbf{U}^T = \mathbf{I}_m$

(2) Since $\mathbf{V}$ is orthogonal, $\mathbf{V}\mathbf{V}^T = \mathbf{I}_n$

(3)

$$\mathbf{\Sigma} = \begin{cases} \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \vdots & & \vdots \\ 0 & 0 & \sigma_m & 0 & \cdots & 0 \end{bmatrix} & : m < n \\ \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_n \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix} & : m > n \end{cases}$$

3.4.1 Construction of SVD

Let $\mathbf{A} \in M_{m\times n}(\mathbb{R})$.

(Step 1) **Find a symmetric, positive semi-definite matrix.** Let $\mathbf{S} := \mathbf{A}^T\mathbf{A} \in M_{n\times n}(\mathbb{R})$. Then

(i) $\mathbf{S}$ is symmetric: $\mathbf{S}^T = (\mathbf{A}^T\mathbf{A})^T = \mathbf{A}^T(\mathbf{A}^T)^T = \mathbf{A}^T\mathbf{A} = \mathbf{S}$.

(ii) $\mathbf{S}$ is positive semi-definite: for $\mathbf{v} \in \mathbb{R}$,

$$\mathbf{v}^T\mathbf{S}\mathbf{v} = \mathbf{v}^T\mathbf{A}^T\mathbf{A}\mathbf{v} = (\mathbf{A}\mathbf{v})^T(\mathbf{A}\mathbf{v}) = ||\mathbf{A}\mathbf{v}||^2 \geq 0.$$

(Step 2) **Spectral Decomposition.**

$$\mathbf{S} = \mathbf{A}^T\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^T = \mathbf{P}\begin{bmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_n\end{bmatrix}\mathbf{P}^T \quad \text{with} \quad \mathbf{P}\mathbf{P}^T = \mathbf{I}_n.$$

(Step 3) **Assume the SVD of A$\in M_{m\times n}(\mathbb{R})$ exists,** i.e., $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. Then

$$\begin{aligned}\mathbf{S} = \mathbf{A}^T\mathbf{A} &= (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)^T(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)\\ &= \mathbf{V}\mathbf{\Sigma}^T(\mathbf{U}^T\mathbf{U})\mathbf{\Sigma}\mathbf{V}^T\\ &= \mathbf{V}\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{V} \quad \text{by orthogonality of } \mathbf{U}\\ &= \mathbf{V}\begin{bmatrix}\sigma_1^2 & & 0\\ & \sigma_2^2 & \\ 0 & & \ddots\end{bmatrix}\mathbf{V}^T\end{aligned}$$

Thus

$$\mathbf{P} = \mathbf{V} \quad \text{and} \quad \lambda_i = \sigma_i^2.$$

(Step 4) Find $\mathbf{U}$ s.t.

$$\begin{aligned}\mathbf{S} = \mathbf{A}\mathbf{A}^T &= (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)^T\\ &= \mathbf{U}\mathbf{\Sigma}(\mathbf{V}^T\mathbf{V})\mathbf{\Sigma}^T\mathbf{U}^T\\ &= \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^T\mathbf{U}^T \quad \text{by orthogonality of } \mathbf{V}\\ &= \mathbf{U}\begin{bmatrix}\sigma_1^2 & & 0\\ & \sigma_2^2 & \\ 0 & & \ddots\end{bmatrix}\mathbf{U}^T.\end{aligned}$$

Note that $\mathbf{A}$ and $\mathbf{A}^T$ have the same eigenvalues. Let

$$\mathbf{V} := \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_n\end{bmatrix},$$

where $\mathbf{v}_i$ is eigenvector of $\mathbf{A}^T\mathbf{A}$ for $i = 1, \ldots, n$. Then

$$i \neq j \implies \langle \mathbf{A}\mathbf{v}_i, \mathbf{A}\mathbf{v}_j\rangle = \mathbf{v}_i^T\mathbf{A}^T\mathbf{A}\mathbf{v}_j = \mathbf{v}_i^T\lambda_j\mathbf{v}_j = \lambda_j\mathbf{v}_i^T\mathbf{v}_j = 0,$$

and so $\{\mathbf{A}\mathbf{v}_1, \ldots, \mathbf{A}\mathbf{v}_r\}$ forms a orthogonal basis of $\operatorname{Im}(\mathbf{A}) \in \mathbb{R}^m$. Since

$$||\mathbf{A}\mathbf{v}_i||^2 = \langle \mathbf{A}\mathbf{v}_i, \mathbf{A}\mathbf{v}_i \rangle = \lambda_i \mathbf{v}_i \mathbf{v}_i^T \mathbf{v}_i = \lambda_i ||\mathbf{v}_i||^2 = \lambda_i,$$

we have

$$\mathbf{u}_i := \frac{\mathbf{A}\mathbf{v}_i}{||\mathbf{A}\mathbf{u}_i||} = \frac{1}{\sqrt{\lambda_i}} \mathbf{A}\mathbf{v}_i$$

for $i = 1, \ldots, r$. Therefore

$$\mathbf{A}\mathbf{V} = \mathbf{A} \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_r \end{bmatrix} = \begin{bmatrix} \sigma_1 \mathbf{u}_1 & \cdots & \sigma_r \mathbf{u}_r \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_2 \end{bmatrix} \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{bmatrix} = \mathbf{U}\boldsymbol{\Sigma}.$$

Hence

$$\mathbf{A}\mathbf{V} = \mathbf{U}\boldsymbol{\Sigma} \implies \mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T.$$

Example 3.6 (Computing the SVD)**.** Find the singular value decomposition of

$$\mathbf{A} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \in M_{2\times 3}(\mathbb{R}).$$

Sol. The SVD requires us to compute the right-singular vectors v_j , the singular values σ_k, and the left-singular vectors u_i.

(Step 1) **Right-singular vectors as the eigenbasis of $\mathbf{A}^T\mathbf{A}$.**

(i) Create real symmetric matrix.

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} -1 & 0 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}.$$

(ii) Spectral Decomposition.

$$\begin{aligned} \det\left(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}_3\right) &= \begin{vmatrix} 1-\lambda & -1 & 0 \\ -1 & 2-\lambda & -1 \\ 0 & -1 & 1-\lambda \end{vmatrix} \\ &= (1-\lambda)\left[(2-\lambda)(1-\lambda) - 1\right] - (-1)\left[\lambda - 1\right] \\ &= (1-\lambda)(2 - 3\lambda + \lambda^2 - 1 - 1) \\ &= (1-\lambda)(-3\lambda + \lambda^2) \\ &= \lambda(1-\lambda)(\lambda - 3) = 0. \end{aligned}$$

Let $\lambda_1 = 3, \lambda_2 = 1$ and $\lambda_3 = 0$.

(a) $(\lambda_1 = 3)$

$$\begin{bmatrix} -2 & -1 & 0 \\ -1 & -1 & -1 \\ 0 & -1 & -2 \end{bmatrix} \mathbf{v}_1 = \mathbf{0} \implies \mathbf{v}_1 = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} \implies \hat{\mathbf{v}}_1 = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}.$$

(b) $(\lambda_2 = 1)$

$$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 1 & -1 \\ 0 & -1 & 0 \end{bmatrix} \mathbf{v}_2 = \mathbf{0} \implies \mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \implies \hat{\mathbf{v}}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}.$$

(c) $(\lambda_3 = 0)$

$$\begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix} \mathbf{v}_3 = \mathbf{0} \implies \mathbf{v}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \implies \hat{\mathbf{v}}_3 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Thus,

$$\mathbf{A}\mathbf{A}^T = \mathbf{P}\mathbf{D}\mathbf{P}^T = \begin{bmatrix} 1/\sqrt{6} & 1/\sqrt{2} & 1/\sqrt{3} \\ -2/\sqrt{6} & 0 & 1/\sqrt{3} \\ 1/\sqrt{6} & -1/\sqrt{2} & 1/\sqrt{3} \end{bmatrix} \begin{bmatrix} 3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{6} & -2/\sqrt{6} & 1/\sqrt{6} \\ 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \end{bmatrix}.$$

Here, let $\mathbf{V} := \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{bmatrix} = \mathbf{P}$.

(Step 2) **Singular-value matrix.** Let

$$\sigma_1 := \sqrt{\lambda_1} = \sqrt{3}, \quad \sigma_2 := \sqrt{\lambda_2} = 1, \quad \sigma_3 := \sqrt{\lambda_3} = \sqrt{0} = 0.$$

Then

$$\mathbf{\Sigma} := \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \end{bmatrix} = \begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

(Step 3) **Left-singular vectors as the normalized image of the right- singular vectors.**

$$\mathbf{u}_1 := \frac{1}{\sigma_1}\mathbf{A}\mathbf{v}_1 = \frac{1}{\sqrt{3}}\begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}\frac{1}{\sqrt{6}}\begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} = \frac{1}{3\sqrt{2}}\begin{bmatrix} -3 \\ 3 \end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ 1 \end{bmatrix},$$

$$\mathbf{u}_2 := \frac{1}{\sigma_2}\mathbf{A}\mathbf{v}_2 = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ -1 \end{bmatrix}.$$

Thus,

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 & -1 \\ 1 & -1 \end{bmatrix}.$$

By Step 1-3, we have

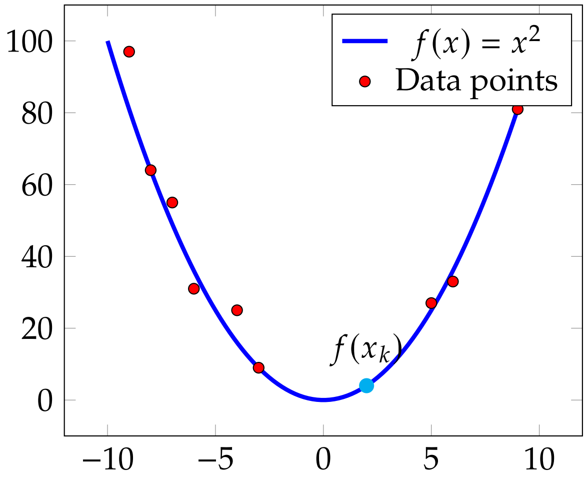
$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 & -1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} 1/\sqrt{6} & -2/\sqrt{6} & 1/\sqrt{6} \\ 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \end{bmatrix}.$$

□

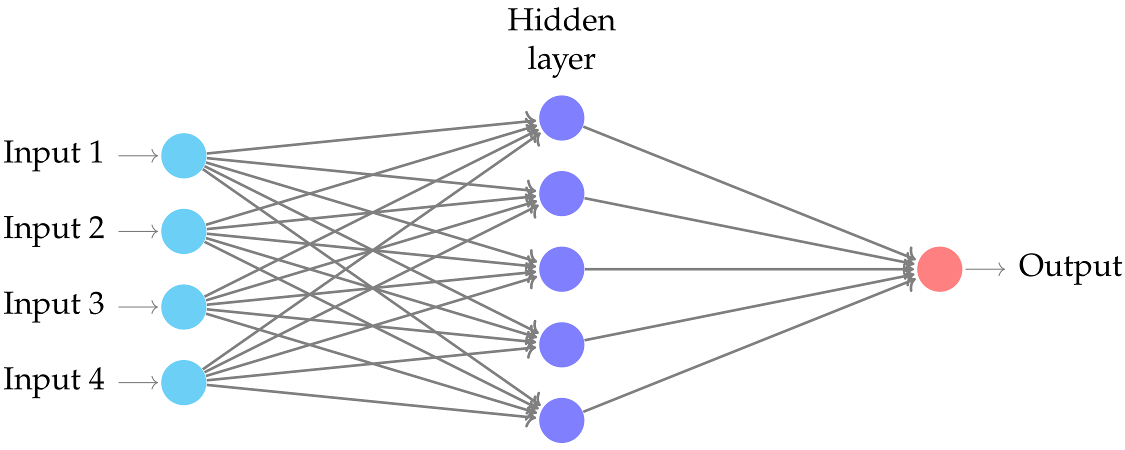
Chapter 4

Vector Calculus (Multi-Variate Calculus)

Note. Evaluate $f(x_k)$ using the date set $\left\{(\mathbf{x}_i, f(\mathbf{x}_i))\right\}_{i=1}^{N}$.



Note (Neural Network)**.**



4.1 Differentiation of Univariate Functions

Derivative

Definition 4.1. For $h > 0$ the **derivative** of f at x is defined as

$$\frac{\mathrm{d}f}{\mathrm{d}x} := \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}.$$

4.1.1 Taylor Series

Taylor Polynomial

Definition 4.2. The **Taylor polynomial** of degree n of $f : \mathbb{R} \to \mathbb{R}$ at x_0 is defined as

$$T_n(x) := \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k,$$

where $f^{(k)}(x_0)$ is the k-th derivative of f at x_0 and $\frac{f^{(k)}(x_0)}{k!}$ are the coefficients of the polynomial.

Taylor Series

Definition 4.3. For a smooth function $f \in C^\infty$, $f : \mathbb{R} \to \mathbb{R}$, the the **Taylor series** of degree n of $f : \mathbb{R} \to \mathbb{R}$ at x_0 is defined as

$$T_\infty(x) := \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k.$$

For $x_0 = 0$, we obtain the **Maclaurin series** as a special case of the Taylor series.

Example 4.1.

(1) $e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots = \sum_{k=0}^{\infty} \frac{x^k}{k!}..$

(2) $\cos x = 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 + \cdots = \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k)!} x^{2k}.$

(3) $\sin x = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 + \cdots = \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k+1)!} x^{2k+1}.$

4.1.2 Differentiation Rules

Chain Rule

Theorem 4.1. *Let I, J be intervals in $\mathbb{R}$, let $g : J \longrightarrow \mathbb{R}$ and $f : I \longrightarrow \mathbb{R}$ be functions such that $f[I] \subseteq J$, and let $a \in I$. Then $\exists f'(a) \exists g'(f(a)) \implies \exists (g \circ f)'(a)$ and*

$$(g \circ f)'(a) = g'(f(a))f'(a).$$

4.2 Partial Differentiation and Gradients

Partial Derivative

Definition 4.4. For a function

$$\begin{array}{rrcl} f : & \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ & \mathbf{x} = (x_1, \cdots, x_n) & \longmapsto & y = f(\mathbf{x}) \end{array} .$$

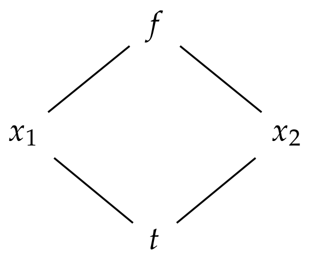
of n variables $x_1, \ldots, x_n$ we define the **partial derivatives** as

$$\begin{aligned} \frac{\partial f}{\partial x_1} &= \lim_{h \to 0} \frac{f(x_1 + h, x_2, \ldots, x_n) - f(\mathbf{x})}{h} \\ &\vdots \\ \frac{\partial f}{\partial x_n} &= \lim_{h \to 0} \frac{f(x_1, \ldots, x_{n-1}, x_n + h) - f(\mathbf{x})}{h} \end{aligned}$$

and collect them in the row vector

$$\nabla_{\mathbf{x}} f = \text{grad } f = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} & \frac{\partial f(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} \in M_{1 \times n}(\mathbb{R}).$$

Example 4.2 (Chain Rule). $g : \mathbb{R} \xrightarrow{\mathbf{x}} \mathbb{R}^2 \xrightarrow{f} \mathbb{R} : t \mapsto \mathbf{x}(t) = \big(x_1(t), x_2(t)\big) \mapsto f(x_1(t), x_2(t))$



$$\begin{aligned} \frac{\mathrm{d}g}{\mathrm{d}t} = \frac{\mathrm{d}f(x_1(t), x_2(t))}{\mathrm{d}t} = &= \frac{\partial f}{\partial x_1}\frac{\mathrm{d}x_1}{\mathrm{d}t} + \frac{\partial f}{\partial x_2}\frac{\mathrm{d}x_2}{\mathrm{d}t} \\ &= \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\mathrm{d}x_1}{\mathrm{d}t} \\ \frac{\mathrm{d}x_2}{\mathrm{d}t} \end{bmatrix} \\ &= \nabla f(x_1, x_2) \cdot \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} \end{aligned}$$

Example 4.3.

4.3 Gradients of Vector-Valued Functions

Vector-valued Function (Vector Field)

Definition 4.5.

$$\begin{aligned} \mathbf{f} \ : \quad & \mathbb{R}^n \longrightarrow \mathbb{R}^m \\ \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}_{n\times 1} & \longmapsto \mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} f_1(x_1, \cdots, x_n) \\ \vdots \\ f_m(x_1, \cdots, x_n) \end{bmatrix}_{m\times 1} \end{aligned}$$

Remark 4.1. The partial derivative of a vector-valued function $\mathbf{f} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ with respect to $x_i \in \mathbb{R}$, $i = 1, \ldots, n$, is given as the vector

$$\frac{\partial \mathbf{f}}{\partial x_i} = \lim_{h\to 0} \frac{\mathbf{f}(x_1, \cdots, x_i + h, \cdots, x_n) - \mathbf{f}(\mathbf{x})}{h} = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix} = \begin{bmatrix} \lim_{h\to 0} \frac{f_1(x_1,\cdots,x_i+h,\cdots,x_n) - f_1(\mathbf{x})}{h} \\ \vdots \\ \lim_{h\to 0} \frac{f_m(x_1,\cdots,x_i+h,\cdots,x_n) - f_m(\mathbf{x})}{h} \end{bmatrix} \in \mathbb{R}^m$$

Jacobian

Definition 4.6. The collection of all first-order partial derivatives of a vector-valued function $\mathbf{f} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is called the **Jacobian**.

$$\mathbf{J} = \nabla_{\mathbf{x}} \mathbf{f} = \left[\frac{\partial f_i}{\partial x_j} \right]_{m\times n} = \begin{bmatrix} \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix} \in M_{m\times n}(\mathbb{R}).$$

In other words,

$$\begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \mathbf{J} = \begin{bmatrix} \nabla_{\mathbf{x}} f_1 \\ \vdots \\ \nabla_{\mathbf{x}} f_m \end{bmatrix}.$$

Remark 4.2.

- The Jocobian approximates a nonlinear transformation locally with a linear transformation.
- The determinant of the Jacobian of **f** can be used to compute the magnifier between two area.

Example 4.4 (Gradient of a Least-Squares Loss in a Linear Model). Consider the linear model

$$\mathbf{y} = \mathbf{\Phi}\boldsymbol{\theta},$$

where

(i) $\boldsymbol{\theta} \in \mathbb{R}^D$ is a parameter vector,

(ii) $\mathbf{\Phi} \in M_{N\times D}(\mathbb{R})$ are input features and

(iii) $\mathbf{y} \in \mathbb{R}^N$ are corresponding observations.

Define the functions

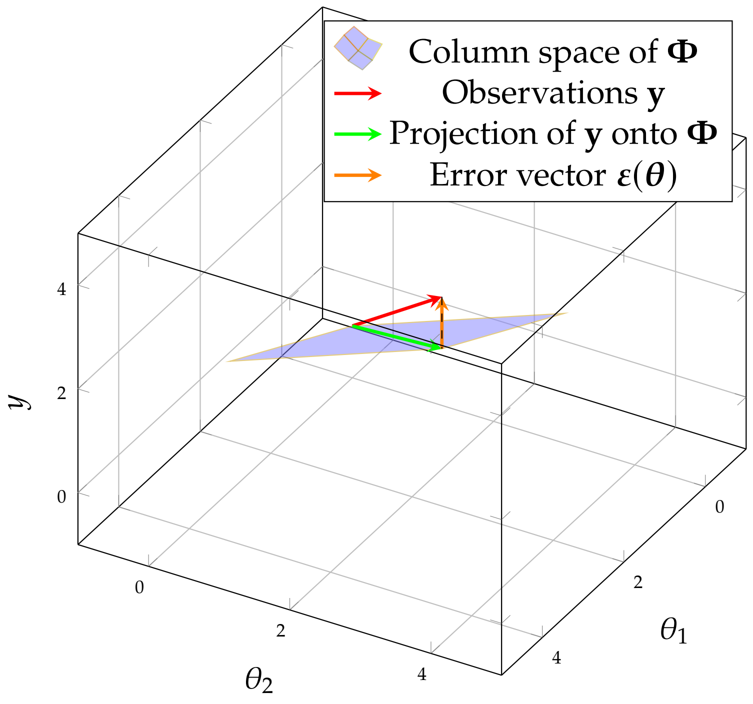
$$\begin{aligned} L(\boldsymbol{\varepsilon}) &= \mathbb{R}^N \to \mathbb{R} := \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = ||\boldsymbol{\varepsilon}||^2, \\ \boldsymbol{\varepsilon}(\boldsymbol{\theta}) &= \mathbb{R}^D \to \mathbb{R}^N := \mathbf{y} - \mathbf{\Phi}\boldsymbol{\theta}. \end{aligned}$$

L is called a *least-squares loss* function. Consider $L \circ \boldsymbol{\varepsilon} : \mathbb{R}^D \to \mathbb{R}$. Then

$$\begin{aligned} \tfrac{\partial L}{\partial \boldsymbol{\theta}} = \tfrac{\partial L}{\partial \varepsilon}\tfrac{\partial \varepsilon}{\partial \boldsymbol{\theta}} \iff \nabla_{\boldsymbol{\theta}} L &= \nabla_{\boldsymbol{\varepsilon}} L \nabla_{\boldsymbol{\theta}} \boldsymbol{\varepsilon} = 2\boldsymbol{\varepsilon}^T(-\mathbf{\Phi}) \quad (2\boldsymbol{\varepsilon}^T \in M_{1\times N}(\mathbb{R}),\ -\mathbf{\Phi} \in M_{N\times D}(\mathbb{R})) \\ &= -2(\mathbf{y}^T - \boldsymbol{\theta}^T\mathbf{\Phi}^T)\mathbf{\Phi} \in M_{1\times D}(\mathbb{R}). \end{aligned}$$

Note that

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} L = 0 \iff -2(\mathbf{y}^T - \boldsymbol{\theta}^T\mathbf{\Phi}^T)\mathbf{\Phi} = 0 &\iff \mathbf{y}^T\mathbf{\Phi} = \boldsymbol{\theta}^T\mathbf{\Phi}^T\mathbf{\Phi} \\ &\iff \mathbf{\Phi}^T\mathbf{y} = \mathbf{\Phi}^T\mathbf{\Phi}\boldsymbol{\theta} \\ &\iff \boldsymbol{\theta} = \left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\mathbf{y}. \end{aligned}$$



4.4 Useful Identities for Computing Gradients

Proposition 4.2. *Let $\boldsymbol{x}, \boldsymbol{a} \in \mathbb{R}^n$ and $\boldsymbol{B} \in M_n(\mathbb{R})$.*

(1) $\frac{\partial}{\partial x}\left(\boldsymbol{x}^T\boldsymbol{a}\right) = \boldsymbol{a}^T$

(2) $\frac{\partial}{\partial x}\left(\boldsymbol{a}^T\boldsymbol{x}\right) = \boldsymbol{a}^T$

(3) $\frac{\partial}{\partial X}\left(\boldsymbol{a}^T\boldsymbol{X}\boldsymbol{b}\right) = \boldsymbol{a}\boldsymbol{b}^T$

(4) $\frac{\partial}{\partial x}\left(\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}\right) = \boldsymbol{x}^T(\boldsymbol{B} + \boldsymbol{B}^T)$

(5) $\frac{\partial}{\partial s}\left[(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^T\boldsymbol{W}(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})\right] = -2(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^T\boldsymbol{W}\boldsymbol{A}$ *for symmetric* $\boldsymbol{W}$.

Proof. (1) Let

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{a} = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \sum_{i=1}^n a_i x_i.$$

Then

$$\nabla_{\mathbf{x}} f = \begin{bmatrix} \frac{\partial}{\partial x_1} f & \cdots & \frac{\partial}{\partial x_n} f \end{bmatrix} = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} = \mathbf{a}^T.$$

(2) Let

$$\nabla_{\mathbf{x}}\left(\mathbf{a}^T\mathbf{x}\right) \overset{\mathbf{a}^T\mathbf{x}\in\mathbb{R}}{=} \nabla_{\mathbf{x}}(\mathbf{a}^T\mathbf{x})^T = \nabla_{\mathbf{x}}\left(\mathbf{x}^T\mathbf{a}\right) = \mathbf{a}^T.$$

(3)

(4) Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is defined by

$$\begin{aligned} f(\mathbf{x}) = \mathbf{x}^T\mathbf{B}\mathbf{x} &= \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}\begin{bmatrix} B_{11} & \cdots & B_{1n} \\ \vdots & \ddots & \vdots \\ B_{n1} & \cdots & B_{nn} \end{bmatrix}\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \\ &= \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}\begin{bmatrix} \sum_{s=1}^n B_{1s}x_s \\ \vdots \\ \sum_{s=1}^n B_{ns}x_s \end{bmatrix} \\ &= \sum_{r=1}^n x_r\left(\sum_{s=1}^n B_{rs}x_s\right) \\ &= \sum_{r,s=1}^n x_r B_{rs} x_s. \end{aligned}$$

Recall that Kronecker $\delta_{ij} = \begin{cases} 1 & : i = j, \\ 0 & : i \neq j. \end{cases}$ and $\frac{\partial x_i}{\partial x_j} = \delta_{ij}$. Then

$$\begin{aligned}
\frac{\partial f}{\partial x_i} &= \frac{\partial}{\partial x_i}\left(\sum_{r,s=1}^{n} x_r B_{rs} x_s\right) \\
&= \sum_{r,s=1}^{n} \frac{\partial}{\partial x_i}(x_r B_{rs} x_s) \\
&= \sum_{r,s=1}^{n} \left(\frac{\partial x_r}{\partial x_i}(B_{rs} x_s) + x_r \frac{\partial (B_{rs} x_s)}{\partial x_i}\right) \qquad \text{Product Rule for Differentiation} \\
&= \sum_{r,s} (\delta_{ri} B_{rs} x_s + x_r B_{rs} \delta_{si}) \\
&= \sum_{s}\sum_{r} \delta_{ri} B_{rs} x_s + \sum_{r}\sum_{s} \delta_{si} x_r B_{rs} \\
&= \sum_{s} \cancelto{1}{\delta_{ii}} B_{is} x_s + \sum_{r} \cancelto{1}{\delta_{ii}} x_r B_{ri} \\
&= [\mathbf{B}\mathbf{x}]_i + \left[\mathbf{x}^T \mathbf{B}\right]_i \\
&= \left[\mathbf{x}^T \mathbf{B}^T\right]_i + \left[\mathbf{x}^T \mathbf{B}\right]_i \qquad \because \mathbf{B}\mathbf{x} \in \mathbb{R} \Rightarrow (\mathbf{B}\mathbf{x})^T = \mathbf{B}\mathbf{x} \\
&= \left[\mathbf{x}^T (\mathbf{B}^T + \mathbf{B})\right]_i .
\end{aligned}$$

Thus

$$\nabla_{\mathbf{x}} f = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_i} & \cdots & \frac{\partial f}{\partial x_D} \end{bmatrix} = \mathbf{x}^T(\mathbf{B}^T + \mathbf{B}).$$

(5) Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is defined by

$$\begin{aligned}
f(\mathbf{s}) = (\mathbf{x} - \mathbf{A}\mathbf{s})^T \mathbf{W} (\mathbf{x} - \mathbf{A}\mathbf{s}) &= \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} B_{11} & \cdots & B_{1n} \\ \vdots & \ddots & \vdots \\ B_{n1} & \cdots & B_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \\
&= \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} \sum_{s=1}^{n} B_{1s} x_s \\ \vdots \\ \sum_{s=1}^{n} B_{ns} x_s \end{bmatrix} \\
&= \sum_{i,j=1}^{n} [\mathbf{x} - \mathbf{A}\mathbf{s}]_i W_{ij} [\mathbf{x} - \mathbf{A}\mathbf{s}]_j \\
&= \sum_{i,j} \left(x_i \sum_{r} A_{ir} s_r\right) W_{ij} \left(x_j - \sum_{t} A_{jt} s_t\right)
\end{aligned}$$

□

Chapter 5

Probability and Distributions

This chapter covers the study of probability and statistics as tools to understand and model uncertainty and observations.

5.1 Probability vs Statistics

In this section, we explore the differences between probability and statistics and their applications in Machine Learning.

5.1.1 Probability

Definition 5.1 (Probability)**.** Probability is the study of uncertainty. It provides a mathematical framework to model and analyze the likelihood of various outcomes.

A **random variable** is a fundamental concept in probability, representing the uncertain outcomes quantitatively.

5.1.2 Statistics

Definition 5.2 (Statistics)**.** Statistics is the discipline that concerns the collection, analysis, interpretation, and presentation of data. In the context of Machine Learning, it involves inferring the processes that generate the data.

5.2 Machine Learning and Data

Machine Learning is closely related to statistics as it often involves creating functions that can predict or categorize data based on observed inputs.

5.3 Key Concepts in Probability

This section outlines the key concepts and definitions used in the study of probability.

- Random Variable X
- Probability Distribution $\mathcal{D}$

5.3.1 Probability Distributions

The probability distribution of a random variable X is a description of the probabilities associated with each of its possible values.

Example 5.1. Consider a random variable X representing the roll of a die, with X taking values from 1 to 6, each with a probability of $\frac{1}{6}$.

Exercise 5.1. Show that the probabilities in a distribution sum up to 1.

5.3.2 Sample Space and Events

Definition 5.3 (Sample Space)**.** The sample space of an experiment or random trial is the set of all possible outcomes.

Definition 5.4 (Event)**.** An event is a set of outcomes of an experiment to which a probability is assigned.

5.3.3 Joint and Marginal Distributions

The joint distribution of a pair of random variables (X, Y) gives the probability that each variable simultaneously falls within any specified range or discrete set of values.

$X \backslash Y$	0	1	Pr[X]
0	1/4	1/2	3/4
1	1/8	1/8	1/4
Pr[Y]	3/8	5/8	

Table 5.1: Joint distribution of X and Y.

5.3.4 Independence and Conditional Probability

Definition 5.5 (Independence)**.** Two events are independent if the occurrence of one does not affect the probability of the occurrence of the other.

Definition 5.6 (Conditional Probability)**.** The probability of an event given that another event has occurred is called the conditional probability.

5.4 Bayes' Theorem

Bayes' Theorem is a fundamental theorem in probability that describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

Theorem 5.1 (Bayes' Theorem)**.** *For any two events A and B, if $P(B) \neq 0$, then*

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

Proof. Starting from the definition of conditional probability:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Similarly, we have:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}.$$

Thus, by rearranging the terms, we get:

$$P(A \mid B)P(B) = P(B \mid A)P(A).$$

Dividing both sides by $P(B)$, we obtain the statement of Bayes' Theorem. □

5.5 Conditional Probability and the Binomial Distribution

Conditional probability is a measure of the probability of an event occurring given that another event has already occurred. The notation $p(y \mid x)$ represents the probability of event y occurring given that event x has occurred. This can be formally defined as follows:

$$p(y \mid x) = \begin{cases} \text{(1) Probability of } y \text{ given } x \\ \text{(2) Likelihood of } x \text{ given } y \end{cases}$$

5.5.1 Binomial Distribution

The binomial distribution is a discrete probability distribution that models the number of successes in a sequence of independent experiments.

Definition 5.7 (Binomial Distribution)**.** A random variable X follows a binomial distribution $\mathcal{B}(n, p)$, denoted by $X \sim \mathcal{B}(n, p)$, if the probability mass function of X is given by:

$$\Pr[X = k] = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{for} \quad k = 0, 1, \cdots, n,$$

where n is the number of trials, p is the probability of success on a single trial, and k is the number of successes.

Example 5.2. Consider a dice with an unknown fixed number of sides marked with a dollar sign ($). Let X denote the number of $ signs observed in n trials, such that $X \in \{1, 2, \cdots, n\}$. If p is the probability of observing a $ sign on a single trial, the distribution of X can be represented as follows:

X	0	1	k	n
$\Pr[X]$	$\binom{n}{k}p^k(1-p)^{n-k}$			

Question: If the actual number of $ signs on the dice, denoted by Y, is unknown, and we observe two $ signs out of 10 trials, what would be our best guess for Y? The probability $\Pr[X = 2 \mid Y = y]$ represents the likelihood of observing exactly 2 $ signs given a specific number y of $ signs on the dice.

The estimation problem can be approached from two perspectives:

Hard	Easy
$\max_y \Pr[Y = y \mid X = 2]$	$\Pr[X = 2 \mid Y = y]$

The "hard" approach involves maximizing the probability of Y given the observation $X = 2$, while the "easy" approach involves directly computing the probability of observing $X = 2$ given a particular value of Y.

5.6 Properties of Random Variables

A random variable X is a variable whose value is subject to variations due to chance. We denote by $X \sim p(x) = \Pr[X = x]$ the probability mass function (pmf) of the random variable X.

5.6.1 Expected Value and Variance

The expected value and variance are two fundamental concepts in the theory of random variables.

Definition 5.8 (Expected Value)**.** The expected value of a function $g(x)$ of a random variable X is given by

$$\mathbb{E}[g(x)] = \sum_x p(x)g(x),$$

where $p(x)$ is the probability mass function of X.

Definition 5.9 (Mean and Variance)**.** The mean or expected value of a random variable X is defined as

$$\mathbb{E}[X] = \sum_x p(x)x,$$

and the variance is defined as

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \left(\mathbb{E}[X]\right)^2.$$

5.6.2 Covariance and Correlation

Covariance and correlation are measures of how much two random variables change together.

Definition 5.10 (Covariance)**.** The covariance of two random variables X and Y is defined as

$$\mathrm{Cov}[X, Y] = \mathbb{E}_{X,Y}\left[(X - \mathbb{E}_X[X])(Y - \mathbb{E}_Y[Y])\right] = \mathbb{E}_{X,Y}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y].$$

For the special case of X with itself, it simplifies to

$$\mathrm{Cov}[X, X] = \mathrm{Var}[X].$$

Definition 5.11 (Correlation)**.** The correlation coefficient between X and Y is given by

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}[X]}\sqrt{\mathrm{Var}[Y]}} \in [-1, 1].$$

The covariance and correlation can have special values under certain conditions:

$$\mathrm{Cov}(X,Y) = \begin{cases} 1 & : X = Y, \\ -1 & : X = -Y, \\ 0 & : \text{if } X, Y \text{ are independent.} \end{cases}$$

Example 5.3. Consider a discrete distribution of random variables X and Y with the following joint probability distribution:

$Y\backslash X$	-1	0	1	$\Pr[Y]$
0	0	$\frac{1}{3}$	0	$\frac{1}{3}$
1	$\frac{1}{3}$	0	$\frac{1}{3}$	$\frac{2}{3}$
$\Pr[X]$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	

Using the definitions above, we can compute the expected values of X and Y as follows:

$$\mathbb{E}[X] = \sum_x p(x)x = (-1)\cdot\frac{1}{3} + 0\cdot\frac{1}{3} + 1\cdot\frac{1}{3} = 0,$$
$$\mathbb{E}[Y] = \sum_y p(y)y = 0\cdot\frac{1}{3} + 1\cdot\frac{2}{3} = \frac{2}{3}.$$

The covariance of X and Y is computed to be

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \sum_{x,y} p(x,y)xy - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0\cdot\frac{2}{3} = 0.$$

This implies that X and Y are uncorrelated since their covariance is zero.

5.7 Multidimensional Random Variables

In the multidimensional case, we consider random vectors and their associated expected values, covariance matrices, and variance matrices.

5.7.1 Expected Value of a Random Vector

Let $\mathbf{X}$ be a random vector in $\mathbb{R}^D$ represented as:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_D \end{bmatrix}.$$

The expected value of $\mathbf{X}$ is a vector in $\mathbb{R}^D$ whose elements are the expected values of the individual random variables that make up $\mathbf{X}$:

$$\mathbb{E}[\mathbf{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_D] \end{bmatrix}.$$

5.7.2 Covariance Matrix

For random vectors $\mathbf{X} \in \mathbb{R}^D$ and $\mathbf{Y} \in \mathbb{R}^E$, the covariance matrix is defined as:

$$\begin{aligned}\operatorname{Cov}(\mathbf{X}, \mathbf{Y}) &= \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^\top\right] \\ &= \mathbb{E}[\mathbf{X}\mathbf{Y}^\top] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{Y}]^\top.\end{aligned}$$

This matrix contains the covariances between each pair of elements in the two random vectors.

5.7.3 Variance Matrix

The variance matrix for $\mathbf{X}$, also known as the covariance matrix of $\mathbf{X}$ with itself, is given by:

$$\begin{aligned}\operatorname{Var}[\mathbf{X}] &= \operatorname{Cov}[\mathbf{X}, \mathbf{X}] \\ &= \mathbb{E}[\mathbf{X}\mathbf{X}^\top] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]^\top \\ &= \begin{bmatrix} \operatorname{Cov}[X_1, X_1] & \cdots & \operatorname{Cov}[X_1, X_D] \\ \vdots & \ddots & \vdots \\ \operatorname{Cov}[X_D, X_1] & \cdots & \operatorname{Cov}[X_D, X_D] \end{bmatrix}.\end{aligned}$$

The covariance between any two elements X_i and X_j of $\mathbf{X}$ is symmetrical, such that $\operatorname{Cov}[X_i, X_j] = \operatorname{Cov}[X_j, X_i]$, and it is defined as:

$$\operatorname{Cov}[X_i, X_j] = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j].$$

The variance matrix $\operatorname{Var}[\mathbf{X}]$ is symmetric and positive semidefinite, meaning that for any vector $\mathbf{x} \in \mathbb{R}^D$, it holds that:

$$\mathbf{x}^\top \operatorname{Var}[\mathbf{X}]\mathbf{x} \geq 0.$$

5.8 Probability Distributions and Independence

5.8.1 Independent and Identically Distributed Random Variables

Random variables $X_1, \ldots, X_n$ are said to be independent and identically distributed (i.i.d.) if they satisfy the following conditions:

(1) **Mutual Independence:** Each pair of variables is independent, which means that for all i, j with $i \neq j$, the joint probability $p(x_i, x_j)$ can be expressed as the product of their individual probabilities: $p(x_i, x_j) = p(x_i)p(x_j)$.

(2) **Identical Distribution:** All variables share the same probability distribution.

5.8.2 Conditional Independence

Definition 5.12 (Conditionally Independent)**.** Two random variables X and Y are conditionally independent given a third variable Z if:

$$p(x, y \mid z) = p(x \mid z)p(y \mid z).$$

This means that knowing the value of Z renders X and Y independent of each other.

5.9 Gaussian Distribution

Note (Gaussian Distribution)**.** A random variable X with mean μ and variance σ^2 has the Gaussian distribution given by:

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

Multivariate Gaussian Distribution

Definition 5.13. Let $\mathbf{X} \in \mathbb{R}^D$, and let $\boldsymbol{\mu} \in \mathbb{R}^D$ and $\boldsymbol{\Sigma} \in M_D(\mathbb{R})$ be the mean vector and covariance matrix, respectively. The multivariate Gaussian distribution of $\mathbf{X}$ is then defined as:

$$p_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{X}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right].$$

We write $p(\mathbf{x}) = \mathcal{N}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x})$ or $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Remark 5.1. Note that $|\sigma| = \sqrt{\sigma^2}$ for the scalar case, and $\sqrt{\boldsymbol{\Sigma}} = |\boldsymbol{\Sigma}|^{1/2}$ denotes the matrix square root of the determinant of $\boldsymbol{\Sigma}$.

Remark 5.2 (Marginals and Conditionals of Gaussians are Gaussians)**.** Let X and Y be two multivariate random variables, that may have. We write the Gaussian distribution in terms of the concatenated states $\begin{bmatrix}\mathbf{x}^T & \mathbf{y}^T\end{bmatrix}$,

$$p(\mathbf{x},\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy}\end{bmatrix}\right).$$

where

$$\begin{cases} \boldsymbol{\Sigma}_{xx} = \operatorname{Cov}[\mathbf{x},\mathbf{x}] & : \text{the marginal covariance matrix of } \mathbf{x}, \\ \boldsymbol{\Sigma}_{yy} = \operatorname{Cov}[\mathbf{y},\mathbf{y}] & : \text{the marginal covariance matrix of } \mathbf{y}, \\ \boldsymbol{\Sigma}_{xy} = \operatorname{Cov}[\mathbf{x},\mathbf{y}] & : \text{the cross-covariance matrix between } \mathbf{x} \text{ and } \mathbf{y}. \end{cases}$$

Remark 5.3. The conditional distribution $p(\mathbf{x} \mid \mathbf{y})$ is also Gaussian and given by

$$p(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}) \quad \text{with} \quad \begin{cases} \boldsymbol{\mu}_{x|y} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y}-\boldsymbol{\mu}_y) \\ \boldsymbol{\Sigma}_{x|y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}. \end{cases}$$

Example 5.4. Consider the bivariate Gaussian distribution

$$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix}0 \\ 2\end{bmatrix}, \begin{bmatrix}0.3 & -1 \\ -1 & 5\end{bmatrix}\right).$$

Then

$$\mu_{x_1|x_2=1} = 0 + (-1)\cdot\frac{1}{5}\cdot(-1-2) = 0.6$$

$$\sigma^2_{x_1|x_2=1} = 0.3 - (-1)\cdot\frac{1}{5}\cdot(-1) = 0.1.$$

Therefore, the conditional Gaussian is given by $p(x_1 \mid x_2 = -1) = \mathcal{N}(0.6, 0.1)$.

Chapter 6

Continuous Optimization

Minimum

If $y = f(x)$ has the minimum $y*$ at $x = x*$ (i.e., $y* = f(x*)$),

$$\begin{cases} y^* := \min_x f(x) \\ x^* := \arg\min_x f(x) \end{cases}$$

6.1 Optimization Using Gradient Descent

We consider the problem of solving for the minimum of a real-valued function

$$\min_{\mathbf{x}} f(\mathbf{x}),$$

where $f : \mathbb{R}^D \to \mathbb{R}$ is an objective function that captures the machine learning problem at hand.

Definition 6.1.

Theorem 6.1.

Chapter 7

Linear Regression

This section introduces linear regression, a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal is to find a linear function that best fits a set of data points, minimizing the difference between the observed values and those predicted by the model.

Key concepts covered include:

- **Regression Analysis:** The process of fitting a curve to the data points.
- **Noise and Variability:** Accounting for random variation and measurement errors in the data.
- **Model Selection:** Choosing the right complexity for the model to avoid overfitting or underfitting.
- **Optimization:** Techniques for finding the parameters that minimize the loss function.
- **Uncertainty Modeling:** Assessing the confidence in the model's predictions.

Regression is crucial in many fields, such as finance, engineering, and medicine, due to its ability to predict and explain complex phenomena.

7.1 Problem Formulation

7.2 Parameter Estimation

Consider the linear regression setting (**??**) and assume we are given a training set $\mathcal{D} = \{(x_1, y_1), \dots, (x_N, y_N)\}$ consisting of N inputs $x_n \in \mathbb{R}^D$ and corresponding observations/targets $y_n \in \mathbb{R}$, $n = 1, \dots, N$. The corresponding graphical model is given in Figure 9.3. Note that y_i and y_j are conditionally independent given their respective inputs x_i, x_j, so that the likelihood factorizes according to

$$p(\mathcal{Y}|\mathcal{X}, \theta) = p(y_1, \dots, y_N|x_1, \dots, x_N, \theta) = \prod_{n=1}^{N} p(y_n|x_n, \theta) = \prod_{n=1}^{N} \mathcal{N}(y_n|x_n^\top \theta, \sigma^2), \quad (9.5a)$$

where we defined $\mathcal{X} = \{x_1, \ldots, x_N\}$ and $\mathcal{Y} = \{y_1, \ldots, y_N\}$ as the sets of training inputs and corresponding targets respectively. The likelihood and the factors $p(y_n|x_n, \theta)$ are Gaussian due to the noise distribution; see (9.3).

In the following, we will discuss how to find optimal parameters θ^* in $\mathbb{R}^D$ for the linear regression model (9.4). Once the parameters θ^* are found, we can predict function values by using this parameter estimate in (9.4) so that at an arbitrary test input x_*, the distribution of the corresponding target y_* is

$$p(y_*|x_*, \theta^*) = \mathcal{N}(y_*|x_*^\top \theta^*, \sigma^2). \tag{9.6}$$

In the following, we will have a look at parameter estimation by maximizing the likelihood, a topic that we already covered to some degree in Section 8.3.

7.3 Maximum Likelihood Estimation

A widely used approach to finding the desired parameters θ_{ML} is *maximum likelihood estimation*, where we find parameters θ_{ML} that maximize the likelihood $p(\mathcal{Y}|\mathcal{X}, \theta)$. Intuitively, maximizing the likelihood means maximizing the predictive distribution of the training data given the model parameters. We obtain the maximum likelihood parameters as

$$\theta_{ML} \in \arg\max_\theta p(\mathcal{Y}|\mathcal{X}, \theta). \tag{7.1}$$

Remark. The likelihood $p(y|x, \theta)$ is not a probability distribution in θ: it is simply a function of the parameters θ but does not integrate to 1 (i.e., it is unnormalized), and may not even be integrable with respect to θ. However, the likelihood in (1) is a normalized probability distribution in y.

7.3.1 Log-Transformation of the Likelihood

To find the desired parameters θ_{ML} that maximize the likelihood, we typically perform gradient ascent (or gradient descent on the negative likelihood). In the case of linear regression we consider here, however, a closed-form solution exists, which makes iterative gradient descent unnecessary. In practice, instead of maximizing the likelihood directly, we apply the log-transformation to the likelihood function and minimize the negative log-likelihood.

$$\mathcal{L}(\theta) := -\log p(\mathcal{Y}|\mathcal{X}, \theta) = -\log \prod_{n=1}^{N} \mathcal{N}(y_n|x_n^\top \theta, \sigma^2), \tag{7.2}$$

where we exploited that the likelihood factorizes over the number of data points due to our independence assumption on the training set.

Remark (Log-Transformation). Since the likelihood is a product of N Gaussian distributions, the log-transformation is especially useful as it does not suffer from numerical underflow, and the differentiation rules turn our problem simpler. The log-transform will turn the product into a sum of log-probabilities such that the corresponding gradient is a sum of individual gradients, instead of a repeated application of the product rule.

7.3.2 Computing the Negative Log-Likelihood

The negative log-likelihood function is also called *error function*. The squared error is often used as a measure of distance. Recall from Section 3.1 that $\|x\|_2 = x^\top x$ if we choose the dot product as the inner product. Ignoring the possibility of duplicate data points, $\mathcal{L}(\theta)$ is given by

$$\mathcal{L}(\theta) := \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - x_n^\top\theta)^2. \tag{7.3}$$

The maximum likelihood estimator θ_{ML} solves $\nabla_\theta\mathcal{L}(\theta) = 0$. Setting the gradient to 0 is a necessary and sufficient condition, and we obtain a global minimum since the Hessian $\nabla^2_\theta(\theta) = \frac{1}{\sigma^2}X^\top X$ is positive definite.

7.3.3 Maximum Likelihood Estimation

A widely used approach to finding the desired parameters $\boldsymbol{\theta}_{ML}$ is *maximum likelihood estimation*, where we find parameters $\boldsymbol{\theta}_{ML}$ that maximize the likelihood $p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})$. Intuitively, maximizing the likelihood means maximizing the predictive distribution of the training data given the model parameters. We obtain the maximum likelihood parameters as

$$\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}). \tag{7.4}$$

Remark. The likelihood $p(y|x, \boldsymbol{\theta})$ is not a probability distribution in $\boldsymbol{\theta}$: it is simply a function of the parameters $\boldsymbol{\theta}$ but does not integrate to 1 (i.e., it is unnormalized), and may not even be integrable with respect to $\boldsymbol{\theta}$. However, the likelihood in (9.7) is a normalized probability distribution in y.

To find the desired parameters $\boldsymbol{\theta}_{ML}$ that maximize the likelihood, we typically perform gradient ascent (or gradient descent on the negative likelihood). In the case of linear regression we consider here, however, a closed-form solution exists, which makes iterative gradient descent unnecessary. In practice, instead of maximizing the likelihood directly, we apply the log-transformation to the likelihood function and minimize the negative log-likelihood.

Remark (Log-Transformation). Since the likelihood (9.5b) is a product of N Gaussian distributions, the log-transformation is especially useful as it (a) does not suffer from numerical underflow, and (b) the differentiation rules will turn our simpler. More specifically, numerical underflow will be a problem when we multiply N probabilities, where N is the number of data points, since we cannot represent very small numbers, such as 10^{-256}. Furthermore, the log-transform will turn the product into a sum of log-probabilities such that the corresponding gradient is a sum of individual gradients, instead of a repeated application of the product rule (5.46) to compute the gradient of a product of N terms.

To find the optimal parameters $\boldsymbol{\theta}_{ML}$ of our linear regression problem, we minimize the negative log-likelihood

$$-\log p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) = -\log\prod_{n=1}^{N} p(y_n|x_n, \boldsymbol{\theta}), \tag{7.5}$$

where we exploited that the likelihood (9.5b) factorizes over the number of data points due to our independence assumption on the training set. In the linear regression model

(9.4), the likelihood is Gaussian (due to the Gaussian additive noise term), such that we arrive at

$$\log p(y_n|x_n, \boldsymbol{\theta}) = -\frac{1}{2\sigma^2}(y_n - x_n^T\boldsymbol{\theta})^2 + \text{const}, \tag{7.6}$$

where the constant includes all terms independent of $\boldsymbol{\theta}$. Using (9.9) in the negative log-likelihood (9.8), we obtain (ignoring the constant terms)

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - x_n^T\boldsymbol{\theta})^2, \tag{7.7}$$

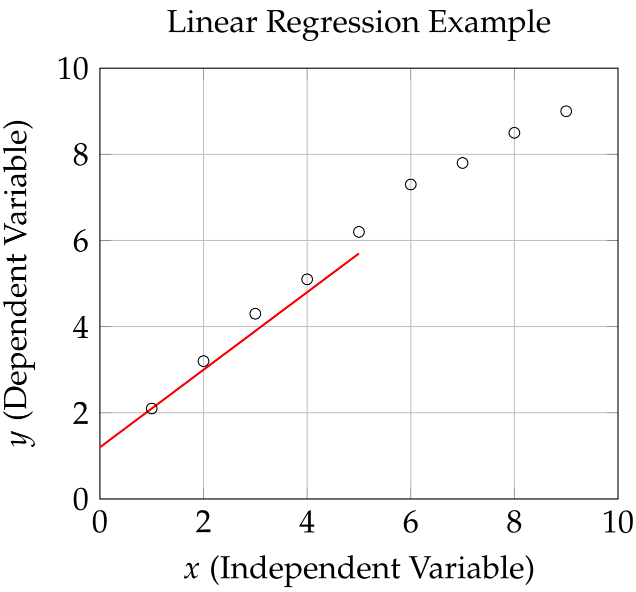
where we define the design matrix $\mathbf{X} = [x_1, \dots, x_N]^T \in \mathbb{R}^{N\times D}$ as the collection of training inputs and $\mathbf{y} = [y_1, \dots, y_N]^T \in \mathbb{R}^N$ as a vector that collects all training targets. Note that the n-th row in the design matrix $\mathbf{X}$ corresponds to the training input x_n. In (9.10b), we used the fact that the sum of squared errors between the observations y_n, and the corresponding model prediction $x_n^T\boldsymbol{\theta}$ equals the squared distance between $\mathbf{y}$ and $\mathbf{X}\boldsymbol{\theta}$. With (9.10b), we now have a concrete form of the negative log-likelihood function we need to optimize. We immediately see that (9.10b) is quadratic in $\boldsymbol{\theta}$. This means that we can find a unique global solution for $\boldsymbol{\theta}_{ML}$ for minimizing the negative log-likelihood $\mathcal{L}$. We can find the global optimum by computing the gradient of $\mathcal{L}$, setting it to 0 and solving for $\boldsymbol{\theta}$.

7.4 Introduction

Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables. The formula for a simple linear regression (with one independent variable) is:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{7.8}$$

where y is the dependent variable, x is the independent variable, β_0 is the intercept, β_1 is the slope, and ϵ is the error term.



7.5 Methodology

The parameters β_0 and β_1 are estimated using the least squares approach, which minimizes the sum of squared residuals.

7.6 Example

Consider a dataset where we want to predict a person's weight based on their height. Here, weight would be our dependent variable (y), and height would be our independent variable (x).

7.7 Results

After fitting the linear regression model, we can use the estimated parameters to make predictions. For instance, if the estimated parameters are $\beta_0 = 50$ and $\beta_1 = 0.75$, then for a person who is 170 cm tall, their predicted weight would be:

$$\text{Weight} = 50 + 0.75 \times 170 \tag{7.9}$$

7.8 Conclusion

Linear regression is a fundamental tool in statistical analysis and helps in understanding the linear relationship between variables.

Bibliography

[1] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for Machine Learning*. 1st ed. Cambridge, U.K.: Cambridge University Press, 2020.