Machine Learning

Lecture 2 Preliminary of Algebra, Calculus, Statistics, and Probability

Hongteng Xu

中国人民大學 RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院 Gaoling School of Artificial Intelligence

Outline

Review

- **Data formulation:** Vector and matrix
- **Residual:** Vector and matrix norms
- **Basic operation:** Matrix-vector multiplication
- **Sample space:** Metric-measure space and its reconstruction

Outline

Review

- **Data formulation:** Vector and matrix
- **Residual:** Vector and matrix norms
- **Basic operation:** Matrix-vector multiplication
- **Sample space:** Metric-measure space and its reconstruction

Today

- Linear space and matrix analysis
  - More matrix operations and two important decompositions
- Derivation of (multi-dimensional) functions
  - Important concepts and derivations of important functions
- Statistics and probability theory
  - Connect data matrix with multi-dimensional random variables

Matrix Multiplication

$$\boldsymbol{B} = \boldsymbol{A}\boldsymbol{X}, \quad \boldsymbol{A} \in \mathbb{R}^{L\times M}, \ \boldsymbol{X} \in \mathbb{R}^{M\times N}, \ \boldsymbol{B} \in \mathbb{R}^{L\times N}. \tag{1}$$

- Element-wise representation:

$$b_{ij} = \sum_{k=1}^{M} a_{ik}x_{kj}, \ \forall\, i = 1, \ldots, L, \ j = 1, \ldots, N \tag{2}$$

- Block-wise representation:

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_{11} & \cdots & \boldsymbol{B}_{1n} \\ \vdots & \ddots & \vdots \\ \boldsymbol{B}_{l1} & \cdots & \boldsymbol{B}_{ln} \end{bmatrix}, \boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \cdots & \boldsymbol{A}_{1m} \\ \vdots & \ddots & \vdots \\ \boldsymbol{A}_{l1} & \cdots & \boldsymbol{A}_{lm} \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_{11} & \cdots & \boldsymbol{X}_{1n} \\ \vdots & \ddots & \vdots \\ \boldsymbol{X}_{m1} & \cdots & \boldsymbol{X}_{mn} \end{bmatrix} \tag{3}$$
$$\boldsymbol{B}_{ij} = \sum_{k=1}^{m} \boldsymbol{A}_{ik}\boldsymbol{X}_{kj}, \ \forall\, i = 1, \ldots, l, \ j = 1, \ldots, n.$$

Matrix Transposition

- You can think of it as "flipping" the rows and columns

$$\begin{bmatrix} a \\ b \end{bmatrix}^T = \begin{bmatrix} a & b \end{bmatrix} \qquad \begin{bmatrix} a & b \\ c & d \end{bmatrix}^T = \begin{bmatrix} a & c \\ b & d \end{bmatrix} \tag{4}$$

- $(\mathbf{A}^T)^T = \mathbf{A}$
- $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$
- $(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$
- $(\mathbf{A}^{-1})^T = \mathbf{A}^{-T}$
- Symmetric Matrices

$$\mathbf{A} = \mathbf{A}^T (a_{ij} = a_{ji}) \tag{5}$$

Special Cases of Matrix Multiplication: Inner and Outer Products

Given two vectors $\boldsymbol{a} \in \mathbb{R}^d$ and $\boldsymbol{b} \in \mathbb{R}^d$

- **Inner product:** $\boldsymbol{a}^T\boldsymbol{b} = \boldsymbol{a} \cdot \boldsymbol{b} = \langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_{i=1}^{d} a_i b_i$.
- **Outer product:** $\boldsymbol{a}\boldsymbol{b}^T = \boldsymbol{a} \otimes \boldsymbol{b} = [a_i b_j] \in \mathbb{R}^{d \times d}$

Special Cases of Matrix Multiplication: Inner and Outer Products

Given two vectors $\boldsymbol{a} \in \mathbb{R}^d$ and $\boldsymbol{b} \in \mathbb{R}^d$

- **Inner product:** $\boldsymbol{a}^T\boldsymbol{b} = \boldsymbol{a} \cdot \boldsymbol{b} = \langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_{i=1}^{d} a_i b_i$.
- **Outer product:** $\boldsymbol{a}\boldsymbol{b}^T = \boldsymbol{a} \otimes \boldsymbol{b} = [a_i b_j] \in \mathbb{R}^{d \times d}$

Given two matrices $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{B} \in \mathbb{R}^{m \times n}$

- **Inner product:** $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathrm{tr}(\boldsymbol{A}^T\boldsymbol{B}) = \sum_{i,j} a_{ij} b_{ij}$.
- **Outer product:** $\boldsymbol{A} \otimes \boldsymbol{B} = [a_{ij} b_{kl}] \in \mathbb{R}^{m \times n \times m \times n}$

Special Cases of Matrix Multiplication: Inner and Outer Products

Given two vectors $\boldsymbol{a} \in \mathbb{R}^d$ and $\boldsymbol{b} \in \mathbb{R}^d$

- **Inner product:** $\boldsymbol{a}^T\boldsymbol{b} = \boldsymbol{a} \cdot \boldsymbol{b} = \langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_{i=1}^d a_i b_i$.
- **Outer product:** $\boldsymbol{a}\boldsymbol{b}^T = \boldsymbol{a} \otimes \boldsymbol{b} = [a_i b_j] \in \mathbb{R}^{d \times d}$

Given two matrices $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{B} \in \mathbb{R}^{m \times n}$

- **Inner product:** $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathrm{tr}(\boldsymbol{A}^T\boldsymbol{B}) = \sum_{i,j} a_{ij} b_{ij}$.
- **Outer product:** $\boldsymbol{A} \otimes \boldsymbol{B} = [a_{ij} b_{kl}] \in \mathbb{R}^{m \times n \times m \times n}$

Kronecker product: $\boldsymbol{A} \otimes \boldsymbol{B} = [a_{ij} b_{kl}] \in \mathbb{R}^{m^2 \times n^2}$.

Some Properties of Matrix Multiplication

- Even if conformable, **AB** does not necessarily equal **BA** (i.e., matrix multiplication is not commutative)
- Matrix multiplication can be extended beyond two matrices
- matrix multiplication is associative, i.e., $\mathbf{A(BC)} = \mathbf{(AB)C}$
- Multiplication and transposition:

$$(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T \tag{6}$$

Orthogonal and Orthonormal

- If $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0,\ \|\boldsymbol{u}\|_2 \neq 0,\ \|\boldsymbol{v}\|_2 \neq 0$, $\boldsymbol{u}$ and $\boldsymbol{v}$ are **orthogonal**.
- If $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0,\ \|\boldsymbol{u}\|_2 = 1,\ \|\boldsymbol{v}\|_2 = 1$, $\boldsymbol{u}$ and $\boldsymbol{v}$ are **orthonormal**.

Orthogonal Matrix

- If square $\boldsymbol{A}$ is orthogonal, it is easy to find its inverse:

$$\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I} \quad \left(\text{i.e., } \boldsymbol{A}^{-1} = \boldsymbol{A}^T\right) \tag{7}$$

- Property:

$$\|\boldsymbol{A}\boldsymbol{v}\| = \|\boldsymbol{v}\| (\text{ does not change the magnitude of } \boldsymbol{v}) \tag{8}$$

Determinant of Matrix

- The determinant of a matrix $\boldsymbol{A}$ is denoted by $|\boldsymbol{A}|$ (or $\det(\boldsymbol{A})$ or $\det \boldsymbol{A}$).
- Determinants exist **only for square matrices**.
- 2×2

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad \det(\boldsymbol{A}) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12}$$

- 3×3

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{21} \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} + a_{31} \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix}$$

- $n \times n$

  $\det(\boldsymbol{A}) = \sum_{j=1}^{m} (-1)^{j+k} a_{jk} \det\left(\boldsymbol{A}_{jk}\right)$, for any $k : 1 \leq k \leq m$

Determinant of Matrix

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

- diagonal matrix:

$$\text{If } \boldsymbol{A} = \begin{bmatrix} a_{11} & 0 & \cdot & 0 \\ 0 & a_{22} & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & a_{nn} \end{bmatrix}, \text{then } \det(\boldsymbol{A}) = \prod_{i=1}^{n} a_{ii}$$

Inverse of Matrix

- The inverse of a matrix $\boldsymbol{A}$ is commonly denoted by $\boldsymbol{A}^{-1}$ or inv $\boldsymbol{A}$.
- The inverse of an $n \times n$ matrix $\boldsymbol{A}$ is the matrix $\boldsymbol{A}^{-1}$ such that $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I} = \boldsymbol{A}^{-1}\boldsymbol{A}$
- The matrix inverse is analogous to a scalar reciprocal
- A matrix which has an inverse is called **nonsingular**
- For some $n \times n$ matrix $\boldsymbol{A}$, an inverse matrix $\boldsymbol{A}^{-1}$ may not exist.
- A matrix which does not have an inverse is **singular**.
- An inverse of $n \times n$ matrix $\boldsymbol{A}$ exists if $|\boldsymbol{A}|$ not 0.

Inverse of Matrix

- The inverse $\boldsymbol{A}^{-1}$ of a matrix $\boldsymbol{A}$ has the property: $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$
- $\boldsymbol{A}^{-1}$ exists if only if $\det(\boldsymbol{A}) \neq 0$
- Terminology
  - **Singular matrix**: $\boldsymbol{A}^{-1}$ does not exist
  - **Ill-conditioned matrix**: $\boldsymbol{A}$ is close to being singular

Property:

$$(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$$
$$\left(\boldsymbol{A}^{T}\right)^{-1} = \left(\boldsymbol{A}^{-1}\right)^{T}$$
$$\left(\boldsymbol{A}^{-1}\right)^{-1} = \boldsymbol{A}$$

- For diagonal matrices $\mathbf{D}^{-1} = \operatorname{diag}\left\{d_1^{-1}, \ldots, d_n^{-1}\right\}$
- For orthogonal matrices $\boldsymbol{A}^{-1} = \boldsymbol{A}^{T}$

Pseudo-inverse

- The pseudo-inverse $\boldsymbol{A}^+$ of a matrix $\boldsymbol{A}$ (could be non-square, e.g., $m \times n$) is given by:

$$\boldsymbol{A}^+ = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T$$

- It can be shown that:

$$\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{I} \quad \text{(provided that } \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1} \text{ exists)}$$

Examples

- Fully-Connected Layer:

$$\boldsymbol{B} = \underbrace{\boldsymbol{A}}_{\text{Weight}}\ \underbrace{\boldsymbol{X}}_{\text{Signals}}\ . \tag{9}$$

Examples

- Fully-Connected Layer:

$$\boldsymbol{B} = \underbrace{\boldsymbol{A}}_{\text{Weight}} \underbrace{\boldsymbol{X}}_{\text{Signals}} . \tag{9}$$

- Graph Convolution:

$$\boldsymbol{B} = \underbrace{\boldsymbol{A}}_{\text{Adj./Lap.}} \underbrace{\boldsymbol{X}}_{\text{Node Attr.}} . \tag{10}$$

Examples

- Fully-Connected Layer:

$$\boldsymbol{B} = \underbrace{\boldsymbol{A}}_{\text{Weight}} \underbrace{\boldsymbol{X}}_{\text{Signals}} . \tag{9}$$

- Graph Convolution:

$$\boldsymbol{B} = \underbrace{\boldsymbol{A}}_{\text{Adj./Lap.}} \underbrace{\boldsymbol{X}}_{\text{Node Attr.}} . \tag{10}$$

- Fast Fourier Transform:

$$\boldsymbol{B} = \underbrace{\boldsymbol{A}}_{\text{Basis}} \underbrace{\boldsymbol{X}}_{\text{Signals}} . \tag{11}$$

Revisit above vector and matrix operations from a viewpoint of linear systems.

Key Tasks of Linear Systems

$$\underbrace{\boldsymbol{b}}_{\text{Output}} = \underbrace{\boldsymbol{A}}_{\text{System}} \underbrace{\boldsymbol{x}}_{\text{Input}} \tag{12}$$

- **Inverse Problem:** Both $\boldsymbol{b}$ and $\boldsymbol{A}$ are known,
  - Solve/Approximate the linear equation: $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}$ or $\min_{\boldsymbol{x}} \; d(\boldsymbol{b}, \boldsymbol{A}\boldsymbol{x})$.

Key Tasks of Linear Systems

$$\underbrace{\boldsymbol{b}}_{\text{Output}} = \underbrace{\boldsymbol{A}}_{\text{System}} \underbrace{\boldsymbol{x}}_{\text{Input}} \tag{12}$$

- **Inverse Problem:** Both $\boldsymbol{b}$ and $\boldsymbol{A}$ are known,
  - Solve/Approximate the linear equation: $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}$ or $\min_{\boldsymbol{x}} \ d(\boldsymbol{b}, \boldsymbol{A}\boldsymbol{x})$.
- **Modeling:** Given sets of $\boldsymbol{b}$'s and $\boldsymbol{x}$'s, denoted as $\boldsymbol{B}$ and $\boldsymbol{X}$,
  - Solve/Approximate the linear equation: $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{X}$ or $\min_{\boldsymbol{A}} \ d(\boldsymbol{B}, \boldsymbol{A}\boldsymbol{X})$.

Key Tasks of Linear Systems

$$\underbrace{\boldsymbol{b}}_{\text{Output}} = \underbrace{\boldsymbol{A}}_{\text{System}} \underbrace{\boldsymbol{x}}_{\text{Input}} \tag{12}$$

- **Inverse Problem:** Both $\boldsymbol{b}$ and $\boldsymbol{A}$ are known,
  - Solve/Approximate the linear equation: $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}$ or $\min_{\boldsymbol{x}} \; d(\boldsymbol{b}, \boldsymbol{A}\boldsymbol{x})$.
- **Modeling:** Given sets of $\boldsymbol{b}$'s and $\boldsymbol{x}$'s, denoted as $\boldsymbol{B}$ and $\boldsymbol{X}$,
  - Solve/Approximate the linear equation: $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{X}$ or $\min_{\boldsymbol{A}} \; d(\boldsymbol{B}, \boldsymbol{A}\boldsymbol{X})$.
- **Factorization:** Given $\boldsymbol{B}$,
  - Solve/Approximate the decomposition/factorization problem: $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{X}$ or $\min_{\boldsymbol{A},\boldsymbol{X}} \; d(\boldsymbol{B}, \boldsymbol{A}\boldsymbol{X})$.

Key Tasks of Linear Systems

$$\underbrace{\boldsymbol{b}}_{\text{Output}} = \underbrace{\boldsymbol{A}}_{\text{System}} \underbrace{\boldsymbol{x}}_{\text{Input}} \tag{12}$$

- **Inverse Problem:** Both $\boldsymbol{b}$ and $\boldsymbol{A}$ are known,
  - Solve/Approximate the linear equation: $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}$ or $\min_{\boldsymbol{x}} \ d(\boldsymbol{b}, \boldsymbol{A}\boldsymbol{x})$.
- **Modeling:** Given sets of $\boldsymbol{b}$'s and $\boldsymbol{x}$'s, denoted as $\boldsymbol{B}$ and $\boldsymbol{X}$,
  - Solve/Approximate the linear equation: $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{X}$ or $\min_{\boldsymbol{A}} \ d(\boldsymbol{B}, \boldsymbol{A}\boldsymbol{X})$.
- **Factorization:** Given $\boldsymbol{B}$,
  - Solve/Approximate the decomposition/factorization problem: $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{X}$ or $\min_{\boldsymbol{A},\boldsymbol{X}} \ d(\boldsymbol{B}, \boldsymbol{A}\boldsymbol{X})$.

Many ML models, algorithms, and applications fall into these paradigms.

The main technical content of this course

Orthogonal Vectors and Linear Independence

- For $\boldsymbol{A} \in \mathbb{C}^{M \times N}$, its Hermitian transport (adjoint) is denoted as $\boldsymbol{A}^H \in \mathbb{C}^{N \times M}$.
- For $\boldsymbol{A} \in \mathbb{R}^{M \times N}$, its transport is $\boldsymbol{A}^T \in \mathbb{R}^{N \times M}$.

Orthogonal Vectors and Linear Independence

- For $\boldsymbol{A} \in \mathbb{C}^{M \times N}$, its Hermitian transport (adjoint) is denoted as $\boldsymbol{A}^H \in \mathbb{C}^{N \times M}$.
- For $\boldsymbol{A} \in \mathbb{R}^{M \times N}$, its transport is $\boldsymbol{A}^T \in \mathbb{R}^{N \times M}$.
- Inner product:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^H \boldsymbol{y} = \sum\nolimits_{n=1}^{N} \bar{x}_n y_n, \; \forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^N \tag{13}$$

- **Explanation:** The projection of $\boldsymbol{y}$ along the direction of $\boldsymbol{x}$.
- $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal to each other, if $\boldsymbol{x}^H \boldsymbol{y} = 0$.
- A **set** of *nonzero* vectors is orthogonal if its vectors are pairwise orthogonal.
- Orthonormal set of vectors: orthogonal + unit norm
- **Theorem.** Vectors in an orthogonal set are **linearly-independent**.

Orthogonal Vectors and Linear Independence

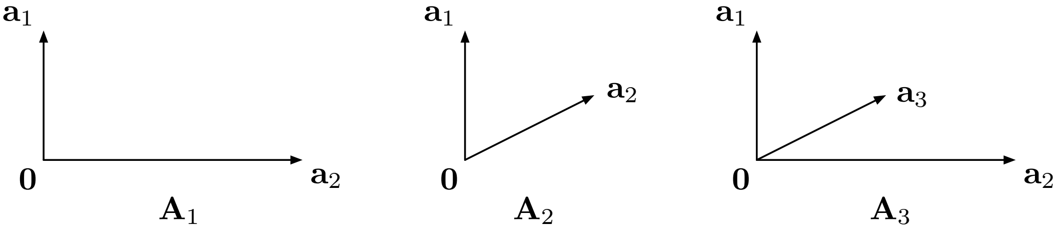
Definition (Linear Independence)

$[\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N]$ are linearly-independent $\Leftrightarrow \sum_{i=1}^{N} \boldsymbol{a}_i x_i = \mathbf{0}$ iff $\boldsymbol{x} = [x_i] = \mathbf{0}$.

Orthogonal Vectors and Linear Independence

Definition (Linear Independence)

$[\boldsymbol{a}_1, ..., \boldsymbol{a}_N]$ are linearly-independent $\Leftrightarrow \sum_{i=1}^{N} \boldsymbol{a}_i x_i = \mathbf{0}$ iff $\boldsymbol{x} = [x_i] = \mathbf{0}$.

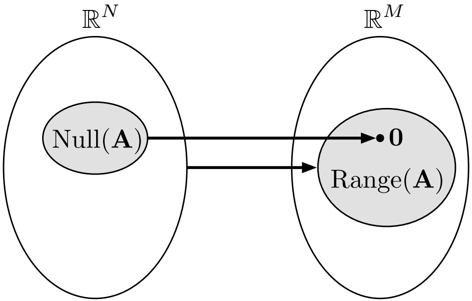


Range and Null Space

- $\text{Range}(\boldsymbol{A}) = $ The column space of $\boldsymbol{A} = \text{span}\{\boldsymbol{a}_1, ..., \boldsymbol{a}_N\} = \{\boldsymbol{A}\boldsymbol{x} | \boldsymbol{x} \in \mathbb{R}^N\}$.
- The null space of $\boldsymbol{A}$ is $\text{Null}(\boldsymbol{A}) = \{\boldsymbol{x} | \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}$.
- Column rank: dimension of column space ($\leq N$).

Range and Null Space

- $\mathrm{Range}(\boldsymbol{A})$ = The column space of $\boldsymbol{A} = \mathrm{span}\{\boldsymbol{a}_1, ..., \boldsymbol{a}_N\} = \{\boldsymbol{A}\boldsymbol{x} | \boldsymbol{x} \in \mathbb{R}^N\}$.
- The null space of $\boldsymbol{A}$ is $\mathrm{Null}(\boldsymbol{A}) = \{\boldsymbol{x} | \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}$.
- Column rank: dimension of column space ($\leq N$).



Rank of Matrix

- rank($\boldsymbol{A}$) (the rank of a m-by-n matrix $\boldsymbol{A}$) is
  = The maximal number of linearly independent columns
  = The maximal number of linearly independent rows

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \text{Rank=?} \quad \begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix} \text{Rank=?}$$

- If $\boldsymbol{A}$ is n by m, then rank($\boldsymbol{A}$)<= min(m,n)
- If n=rank($\boldsymbol{A}$), then $\boldsymbol{A}$ has full row rank
- If m=rank($\boldsymbol{A}$), then $\boldsymbol{A}$ has full column rank

Full Column Rank and Linear Independence

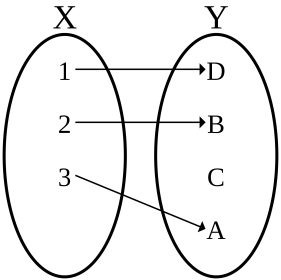
Theorem

$\boldsymbol{A}$ is of full column rank (Rank$(\boldsymbol{A}) = N$).

$\Leftrightarrow$ *$\boldsymbol{A}$ is injective.*

$\Leftrightarrow$ $[\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N]$ *is linearly-independent.*

$\Leftrightarrow$ *Null$(\boldsymbol{A}) = \{\mathbf{0}\}$.*



Linear (Vector) Space and Linear Independence

- Linear Vector Space: $\mathcal{X} \subset \mathbb{R}^N$.
- Computational closure:
  - Vector addition: $\mathcal{X} \times \mathcal{X} \mapsto \mathcal{X}$.
  - Scalar multiplication: $\mathcal{X} \times \mathbb{R} \mapsto \mathcal{X}$.

Linear (Vector) Space and Linear Independence

- Linear Vector Space: $\mathcal{X} \subset \mathbb{R}^N$.
- Computational closure:
  - Vector addition: $\mathcal{X} \times \mathcal{X} \mapsto \mathcal{X}$.
  - Scalar multiplication: $\mathcal{X} \times \mathbb{R} \mapsto \mathcal{X}$.
- The basis of a linear vector space $\mathcal{X}$: $\boldsymbol{B} = \{\boldsymbol{b}_1, ..., \boldsymbol{b}_D\}$
  - The $\boldsymbol{b}$'s are linear independent.
  - For each $\boldsymbol{x} \in \mathcal{X}$, $\exists \boldsymbol{a} \in \mathbb{R}^D$, $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{a} = \sum_{i=1}^{D} \mathbf{b}_d a_d$.

Linear (Vector) Space and Linear Independence

- Linear Vector Space: $\mathcal{X} \subset \mathbb{R}^N$.
- Computational closure:
  - Vector addition: $\mathcal{X} \times \mathcal{X} \mapsto \mathcal{X}$.
  - Scalar multiplication: $\mathcal{X} \times \mathbb{R} \mapsto \mathcal{X}$.
- The basis of a linear vector space $\mathcal{X}$: $\boldsymbol{B} = \{\boldsymbol{b}_1, ..., \boldsymbol{b}_D\}$
  - The $\boldsymbol{b}$'s are linear independent.
  - For each $\boldsymbol{x} \in \mathcal{X}$, $\exists \boldsymbol{a} \in \mathbb{R}^D$, $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{a} = \sum_{i=1}^{D} \mathbf{b}_d a_d$.
- **Question:** Can $D > N$?

Linear (Vector) Space and Linear Independence

- Linear Vector Space: $\mathcal{X} \subset \mathbb{R}^N$.
- Computational closure:
  - Vector addition: $\mathcal{X} \times \mathcal{X} \mapsto \mathcal{X}$.
  - Scalar multiplication: $\mathcal{X} \times \mathbb{R} \mapsto \mathcal{X}$.
- The basis of a linear vector space $\mathcal{X}$: $\boldsymbol{B} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_D\}$
  - The $\boldsymbol{b}$'s are linear independent.
  - For each $\boldsymbol{x} \in \mathcal{X}$, $\exists \boldsymbol{a} \in \mathbb{R}^D$, $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{a} = \sum_{i=1}^{D} \mathbf{b}_d a_d$.
- **Question:** Can $D > N$? **No.**
- $D = \dim(\mathcal{X})$, and $\mathcal{X} = \mathrm{span}(\boldsymbol{B})$.

Linear (Vector) Space and Linear Independence

- Linear Vector Space: $\mathcal{X} \subset \mathbb{R}^N$.
- Computational closure:
  - Vector addition: $\mathcal{X} \times \mathcal{X} \mapsto \mathcal{X}$.
  - Scalar multiplication: $\mathcal{X} \times \mathbb{R} \mapsto \mathcal{X}$.
- The basis of a linear vector space $\mathcal{X}$: $\boldsymbol{B} = \{\boldsymbol{b}_1, ..., \boldsymbol{b}_D\}$
  - The $\boldsymbol{b}$'s are linear independent.
  - For each $\boldsymbol{x} \in \mathcal{X}$, $\exists \boldsymbol{a} \in \mathbb{R}^D$, $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{a} = \sum_{i=1}^{D} \mathbf{b}_d a_d$.
- **Question:** Can $D > N$? **No.**
- $D = \dim(\mathcal{X})$, and $\mathcal{X} = \mathrm{span}(\boldsymbol{B})$.

Could you enumerate some typical linear spaces and their basis?

Components of a Vector

- Let $\mathcal{S} = \text{span}\{\boldsymbol{q}_1, ..., \boldsymbol{q}_N\}$, where $\{\boldsymbol{q}_1, ..., \boldsymbol{q}_N\}$ is an orthonormal set in $\mathbb{R}^M$.
- For a vector $\boldsymbol{v} \in \mathbb{R}^M$, its residual w.r.t. the set is

$$\boldsymbol{r} = \boldsymbol{v} - \sum_{i=1}^{N} \langle \boldsymbol{q}_i, \boldsymbol{v} \rangle \boldsymbol{q}_i. \tag{14}$$

Components of a Vector

- Let $\mathcal{S} = \text{span}\{\boldsymbol{q}_1, ..., \boldsymbol{q}_N\}$, where $\{\boldsymbol{q}_1, ..., \boldsymbol{q}_N\}$ is an orthonormal set in $\mathbb{R}^M$.
- For a vector $\boldsymbol{v} \in \mathbb{R}^M$, its residual w.r.t. the set is

$$\boldsymbol{r} = \boldsymbol{v} - \sum_{i=1}^{N} \langle \boldsymbol{q}_i, \boldsymbol{v} \rangle \boldsymbol{q}_i. \tag{14}$$

- Obviously, $\langle \boldsymbol{r}, \boldsymbol{q}_i \rangle = 0, \; \forall \boldsymbol{i} = 1, .., N$. (Derive it)

Components of a Vector

- Let $\mathcal{S} = \text{span}\{\boldsymbol{q}_1, ..., \boldsymbol{q}_N\}$, where $\{\boldsymbol{q}_1, ..., \boldsymbol{q}_N\}$ is an orthonormal set in $\mathbb{R}^M$.
- For a vector $\boldsymbol{v} \in \mathbb{R}^M$, its residual w.r.t. the set is

$$\boldsymbol{r} = \boldsymbol{v} - \sum_{i=1}^{N} \langle \boldsymbol{q}_i, \boldsymbol{v} \rangle \boldsymbol{q}_i. \tag{14}$$

- Obviously, $\langle \boldsymbol{r}, \boldsymbol{q}_i \rangle = 0, \ \forall i = 1, .., N$. (Derive it)
- The decomposition of $\boldsymbol{v}$:

$$\boldsymbol{v} = \underbrace{\sum_{i=1}^{N} \langle \boldsymbol{q}_i, \boldsymbol{v} \rangle \boldsymbol{q}_i}_{\in \mathcal{S}} + \underbrace{\boldsymbol{r}}_{\in \mathcal{S}^\perp}, \tag{15}$$

where $\mathcal{S} \oplus \mathcal{S}^\perp = \mathbb{R}^M$.

Components of a Vector

- Let $\mathcal{S} = \text{span}\{\boldsymbol{q}_1, ..., \boldsymbol{q}_N\}$, where $\{\boldsymbol{q}_1, ..., \boldsymbol{q}_N\}$ is an orthonormal set in $\mathbb{R}^M$.
- For a vector $\boldsymbol{v} \in \mathbb{R}^M$, its residual w.r.t. the set is

$$\boldsymbol{r} = \boldsymbol{v} - \sum_{i=1}^{N} \langle \boldsymbol{q}_i, \boldsymbol{v} \rangle \boldsymbol{q}_i. \tag{14}$$

- Obviously, $\langle \boldsymbol{r}, \boldsymbol{q}_i \rangle = 0, \ \forall i = 1, .., N$. (Derive it)
- The decomposition of $\boldsymbol{v}$:

$$\boldsymbol{v} = \underbrace{\sum_{i=1}^{N} \langle \boldsymbol{q}_i, \boldsymbol{v} \rangle \boldsymbol{q}_i}_{\in \mathcal{S}} + \underbrace{\boldsymbol{r}}_{\in \mathcal{S}^\perp}, \tag{15}$$

  where $\mathcal{S} \oplus \mathcal{S}^\perp = \mathbb{R}^M$.
- **Question:** $\forall \boldsymbol{x} \in \mathcal{S}$ and $\boldsymbol{y} \in \mathcal{S}^\perp$, is $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \equiv 0$?

Components of a Vector

- Let $\mathcal{S} = \text{span}\{\boldsymbol{q}_1, ..., \boldsymbol{q}_N\}$, where $\{\boldsymbol{q}_1, ..., \boldsymbol{q}_N\}$ is an orthonormal set in $\mathbb{R}^M$.
- For a vector $\boldsymbol{v} \in \mathbb{R}^M$, its residual w.r.t. the set is

$$\boldsymbol{r} = \boldsymbol{v} - \sum_{i=1}^{N} \langle \boldsymbol{q}_i, \boldsymbol{v} \rangle \boldsymbol{q}_i. \tag{14}$$

- Obviously, $\langle \boldsymbol{r}, \boldsymbol{q}_i \rangle = 0, \ \forall i = 1, .., N$. (Derive it)
- The decomposition of $\boldsymbol{v}$:

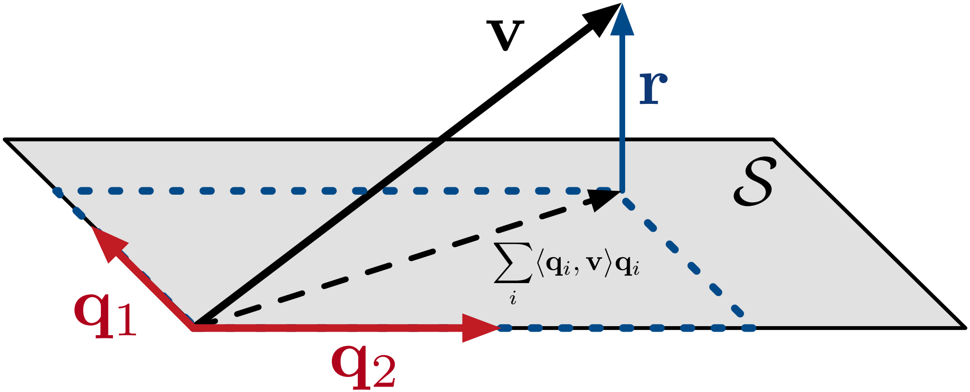
$$\boldsymbol{v} = \underbrace{\sum_{i=1}^{N} \langle \boldsymbol{q}_i, \boldsymbol{v} \rangle \boldsymbol{q}_i}_{\in \mathcal{S}} + \underbrace{\boldsymbol{r}}_{\in \mathcal{S}^\perp}, \tag{15}$$

where $\mathcal{S} \oplus \mathcal{S}^\perp = \mathbb{R}^M$.

- **Question:** $\forall \boldsymbol{x} \in \mathcal{S}$ and $\boldsymbol{y} \in \mathcal{S}^\perp$, is $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \equiv 0$? **Yes.**

Components of a Vector

- Let $\mathcal{S} = \text{span}\{\boldsymbol{q}_1, ..., \boldsymbol{q}_2\}$ and $\mathbb{R}^M = \mathbb{R}^3$.



Recall The Linear System $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$

- $\boldsymbol{A} \in \mathbb{R}^{M \times M}$, $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^{M}$

Recall The Linear System $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$

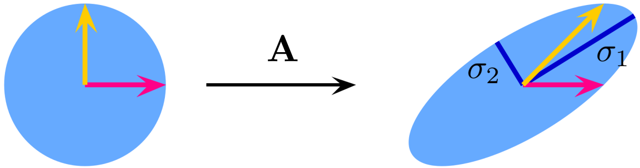
- $\boldsymbol{A} \in \mathbb{R}^{M \times M}$, $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^M$
- $\mathcal{Y} = \text{Range}(\boldsymbol{A}) = \{\boldsymbol{A}\boldsymbol{x} \mid \boldsymbol{x} \in \mathbb{R}^M\}$ The image of the domain $\mathcal{X}$ under the linear transform $\boldsymbol{A}$.

Recall The Linear System $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$

- $\boldsymbol{A} \in \mathbb{R}^{M \times M}, \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^M$
- $\mathcal{Y} = \mathrm{Range}(\boldsymbol{A}) = \{\boldsymbol{A}\boldsymbol{x} \mid \boldsymbol{x} \in \mathbb{R}^M\}$ The image of the domain $\mathcal{X}$ under the linear transform $\boldsymbol{A}$.
- When $\mathcal{X} := \{\boldsymbol{x} \in \mathbb{R}^M \mid \|\boldsymbol{x}\|_2 = 1\}$ (a unit sphere $\mathcal{S}^{M-1}$), $\mathcal{Y}$ is always a hyperellipse.

Recall The Linear System $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$

- $\boldsymbol{A} \in \mathbb{R}^{M \times M}$, $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^M$
- $\mathcal{Y} = \text{Range}(\boldsymbol{A}) = \{\boldsymbol{A}\boldsymbol{x} \mid \boldsymbol{x} \in \mathbb{R}^M\}$ The image of the domain $\mathcal{X}$ under the linear transform $\boldsymbol{A}$.
- When $\mathcal{X} := \{\boldsymbol{x} \in \mathbb{R}^M \mid \|\boldsymbol{x}\|_2 = 1\}$ (a unit sphere $\mathcal{S}^{M-1}$), $\mathcal{Y}$ is always a hyperellipse.
- A hyperellipse $\mathcal{Y} = \{\boldsymbol{A}\boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{S}^{M-1}\}$, characterized by $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M \in \mathbb{R}^M$ orthonormal directions, and $\sigma_1, \ldots, \sigma_M$ the corresponding length of the semi-axes.



Derivative of a Function

- $\lim_{h\to 0} \frac{f(a+h)-f(a)}{h}$ is called the derivative of f at a.
- We write the derivative of f with respect to x is

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

- There are many ways to write the derivative of $y = f(x)$.

  e.g. define the slope of the curve y=f(x) at the point x.

Some Important Rules of (Partial) Derivatives

- Scalar multiplication: $\partial_x[af(x)] = a\,[\partial_x f(x)]$
- Polynomials: $\partial_x\left[x^k\right] = kx^{k-1}$
- Function addition: $\partial_x[f(x) + g(x)] = [\partial_x f(x)] + [\partial_x g(x)]$
- Function multiplication: $\partial_x[f(x)g(x)] = f(x)\,[\partial_x g(x)] + [\partial_x f(x)]\,g(x)$
- Function division: $\partial_x\left[\frac{f(x)}{g(x)}\right] = \frac{[\partial_x f(x)]g(x) - f(x)[\partial_x g(x)]}{[g(x)]^2}$
- Function composition: $\partial_x[f(g(x))] = [\partial_x g(x)]\,[\partial_x f]\,(g(x))$
- Exponentiation: $\partial_x\,[e^x] = e^x$ and $\partial_x\,[a^x] = \log(a)e^x$
- Logarithms: $\partial_x[\log x] = \frac{1}{x}$

Definitions of Gradient

- Matrix-calculus Scalar-by-matrix
- Suppose that $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is a function that takes as input a matrix A of size $m \times n$ and returns a real value. Then the ***gradient*** of f (with respect to $\boldsymbol{A} \in \mathbb{R}^{m\times n}$) is the matrix of

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

- In principle, gradients are a natural extension of partial derivatives to functions of multiple variables.

Definitions of Gradient

- $f : \mathcal{X} \subset \mathbb{R}^n \mapsto \mathbb{R}$
- Size of gradient is always the same as the size of variable

$$\nabla_x f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \frac{\partial f(\boldsymbol{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^n \text{ if } \boldsymbol{x} \in \mathbb{R}^n \tag{16}$$

Definitions of Gradient

- $f : \mathcal{X} \subset \mathbb{R}^n \mapsto \mathbb{R}^m$
- Size of gradient is always the same as the size of variable

$$\nabla_x f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix}$$

Definitions of Gradient

- $f : \mathcal{X} \subset \mathbb{R}^n \mapsto \mathbb{R}^m$
- Size of gradient is always the same as the size of variable

$$\nabla_x f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix} = \boldsymbol{J}_f^T(\boldsymbol{x}) \tag{17}$$

Examples

$$\frac{\partial \boldsymbol{x}^T \boldsymbol{a}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{a}^T \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}$$

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{X} \boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{a}\boldsymbol{b}^T$$

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{X}^T \boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{b}\boldsymbol{a}^T$$

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{X} \boldsymbol{a}}{\partial \boldsymbol{X}} = \frac{\partial \boldsymbol{a}^T \boldsymbol{X}^T \boldsymbol{a}}{\partial \boldsymbol{X}} = \boldsymbol{a}\boldsymbol{a}^T$$

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

Hessian Matrix

- Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a function that takes a vector in $\mathbb{R}^n$ and returns a real number. Then the Hessian matrix with respect to $\boldsymbol{x}$, written $\nabla^2_{\boldsymbol{x}} f(\boldsymbol{x})$ or simply as $\boldsymbol{H}$ is the $n \times n$ matrix of partial derivatives,

$$\nabla^2_{\boldsymbol{x}} f(\boldsymbol{x}) \in \mathbb{R}^{n \times n} = \begin{bmatrix} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1^2} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n^2} \end{bmatrix} \tag{18}$$

References

- More knowledge about matrix analysis will be introduced later
  - SVD
  - Eigen decomposition
- Matrix Cookbook

  https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Statistical Machine Learning: Connecting Statistics and Algebra

A typical ML scenario:

- $\boldsymbol{X} \in \mathbb{R}^{D \times N}$ are a set of samples.
- Each sample $\boldsymbol{x} \sim \mu_{\mathcal{X}}$ is a **random variable**.
- How to estimate $P_{\mathcal{X}}$ via a model $\hat{p}_\theta$ based on the data $\boldsymbol{X}$?

Mean, (Co)variance, and Their Unbiased Estimation

Suppose that we observed a set of i.i.d. samples $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$, each $\boldsymbol{x}_i \sim P$

- Mean

$$\mu = \mathbb{E}_P[X] \tag{19}$$

- Unbiased estimation of mean: the average of samples

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \tag{20}$$

Mean, (Co)variance, and Their Unbiased Estimation

Suppose that we observed a set of i.i.d. samples $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$, each $\boldsymbol{x}_i \sim P$

- Mean

$$\mu = \mathbb{E}_P[X] \tag{19}$$

- Unbiased estimation of mean: the average of samples

$$\hat{\mu} = \frac{1}{n} \sum\nolimits_{i=1}^{n} \boldsymbol{x}_i \tag{20}$$

- Variance

$$\sigma^2 = \mathbb{V}_P[X] = \mathbb{E}_P[(X-\mu)^2] \underbrace{= \mathbb{E}_P[X^2] - \mathbb{E}_P^2[X]}_{\text{Derive it}} \tag{21}$$

- Unbiased estimation of variance:

$$\hat{\sigma}^2 = \frac{1}{n} \sum\nolimits_{i=1}^{n} (\boldsymbol{x}_i - \mu)^2 \quad \text{or} \quad \hat{\sigma}^2 = \frac{1}{n-1} \sum\nolimits_{i=1}^{n} (\boldsymbol{x}_i - \hat{\mu})^2 \tag{22}$$

Properties of Mean and Variance

- $\mathbb{E}[aX] = a\mathbb{E}[X]$
- $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$
- $\mathbb{V}[X + a] = \mathbb{V}[X]$
- $\mathbb{V}[aX + bY] = a^2\mathbb{V}[X] + b^2\mathbb{V}[Y] + 2ab\mathrm{Cov}(X, Y)$.

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \tag{23}$$

- Linear combinations

$$\mathbb{V}\left[\sum_{i=1}^{K} X_i\right] = \sum_{i,j=1}^{K} \mathrm{Cov}(X_i, X_j). \tag{24}$$

Method of Moments

- Suppose that we have a model $\theta \in \mathbb{R}^D$, which works to define a data distribution $P(X; \theta)$.
- The parameters can be determined by solving the equations corresponding to the top-D moments of the distribution

$$\mu_d = \mathbb{E}_P[X^d] = f_d(\theta), \quad d = 1, ..., D. \tag{25}$$

- In practice, given a set of samples $\boldsymbol{X} = [x_i]$, we can estimate $\{\mu_d\}$ as

$$\hat{\mu}_d = \frac{1}{n} \sum_{i=1}^{n} x_i^d, \quad d = 1, ..., D. \tag{26}$$

Ideally, solving the D equations in (25) provides us with a good estimation of θ.

Method of Moments

- Suppose that we have a model $\theta \in \mathbb{R}^D$, which works to define a data distribution $P(X; \theta)$.
- The parameters can be determined by solving the equations corresponding to the top-D moments of the distribution

$$\mu_d = \mathbb{E}_P[X^d] = f_d(\theta), \quad d = 1, ..., D. \tag{25}$$

- In practice, given a set of samples $\boldsymbol{X} = [x_i]$, we can estimate $\{\mu_d\}$ as

$$\hat{\mu}_d = \frac{1}{n}\sum_{i=1}^{n} x_i^d, \quad d = 1, ..., D. \tag{26}$$

  Ideally, solving the D equations in (25) provides us with a good estimation of θ.
- **Could you enumerate its drawbacks?**

Why Does ML Require Lots of Data: Law of Large Numbers

Weak law (converge in probability)

- For $\{X_i\}_{i=1}^n$,

$$\lim_{n\to\infty} \bar{X}_n = \lim_{n\to\infty} \frac{\sum_{i=1}^n X_i}{n} \xrightarrow{P} \underbrace{\mathbb{E}[X]}_{\mu}. \tag{27}$$

Why Does ML Require Lots of Data: Law of Large Numbers

Weak law (converge in probability)

- For $\{X_i\}_{i=1}^n$,

$$\lim_{n\to\infty} \bar{X}_n = \lim_{n\to\infty} \frac{\sum_{i=1}^n X_i}{n} \xrightarrow{P} \underbrace{\mathbb{E}[X]}_{\mu}. \tag{27}$$

- Equivalent representation

$$\lim_{n\to\infty} \mathrm{P}(|\bar{X}_n - \mu| < \epsilon) = 1, \quad \forall \epsilon > 0. \tag{28}$$

Why Does ML Require Lots of Data: Law of Large Numbers

Strong law (Kolmogorov's law, converge almost surely)

- For $\{X_i\}_{i=1}^n$,

$$\lim_{n\to\infty} \bar{X}_n \xrightarrow{a.s.} \mu \tag{29}$$

or equivalently,

$$\mathrm{P}\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1. \tag{30}$$

Why Does ML Require Lots of Data: Law of Large Numbers

Strong law (Kolmogorov's law, converge almost surely)

- For $\{X_i\}_{i=1}^n$,

$$\lim_{n\to\infty} \bar{X}_n \xrightarrow{a.s.} \mu \tag{29}$$

or equivalently,

$$\mathrm{P}\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1. \tag{30}$$

Variance reduction: Suppose that $\mathbb{V}[X_i] = \sigma^2$ for $i = 1, 2, \ldots$.

$$\mathbb{V}[\bar{X}_n] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^n X_i\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \tag{31}$$

Why Does ML Like Gaussian Distribution: Central Limit Theorem

Lindeberg-Lévy CLT.

- Suppose $\{X_i\}_{i=1}^{n}$ is a sequence of i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}[X_i] = \sigma^2 < \infty$, then

$$\lim_{n \to \infty} \sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \tag{32}$$

where $\xrightarrow{d}$ means converge in distribution.

ML: Frequentist Statistic Viewpoint

Given a set of samples $\boldsymbol{X} = [\boldsymbol{x}_i]$, we assume that the samples are sampled from a distribution $P(\boldsymbol{x}|\theta)$ (a model with parameter θ).

- **Principle:** Assume a **deterministic** model, learn the model via maximum likelihood estimation (**MLE**):

$$
\begin{aligned}
\hat{\theta}_{MLE} &= \arg\max_{\theta} \underbrace{\log P(\boldsymbol{X}|\theta)}_{\text{Loglike}(\theta;\boldsymbol{X})} \\
&= \arg\max_{\theta} \sum_{i=1}^{n} \log P(\boldsymbol{x}_i|\theta) \quad \text{(i.i.d. Assumption)}
\end{aligned}
\tag{33}
$$

ML: Frequentist Statistic Viewpoint

Given a set of samples $\boldsymbol{X} = [\boldsymbol{x}_i]$, we assume that the samples are sampled from a distribution $P(\boldsymbol{x}|\theta)$ (a model with parameter θ).

- **Principle:** Assume a **deterministic** model, learn the model via maximum likelihood estimation (**MLE**):

$$\begin{aligned} \hat{\theta}_{MLE} &= \arg\max_{\theta} \underbrace{\log P(\boldsymbol{X}|\theta)}_{\text{Loglike}(\theta;\boldsymbol{X})} \\ &= \arg\max_{\theta} \sum_{i=1}^{n} \log P(\boldsymbol{x}_i|\theta) \quad \text{(i.i.d. Assumption)} \end{aligned} \tag{33}$$

- Deterministic model $\rightarrow$ (pointwise) MLE $\rightarrow$ optimization.

ML: Bayesian Statistic Viewpoint

- **Principle:** Assume a **probabilistic** model — the model θ yields a prior distribution.
- Bayes' theorem

$$\underbrace{P(\theta|\boldsymbol{X})}_{\text{Posterior}(\theta)} = \frac{P(\boldsymbol{X}|\theta)P(\theta)}{P(X)} \propto \underbrace{P(\boldsymbol{X}|\theta)}_{\text{Likelihood}(\theta)} \underbrace{P(\theta)}_{\text{Prior}(\theta)} . \tag{34}$$

- Maximum A Posterior (**MAP**) estimation

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|\boldsymbol{X}) = \arg\max_{\theta} P(\boldsymbol{X}|\theta)P(\theta) \tag{35}$$

ML: Bayesian Statistic Viewpoint

- **Principle:** Assume a **probabilistic** model — the model θ yields a prior distribution.
- Bayes' theorem

$$\underbrace{P(\theta|\boldsymbol{X})}_{\text{Posterior}(\theta)} = \frac{P(\boldsymbol{X}|\theta)P(\theta)}{P(X)} \propto \underbrace{P(\boldsymbol{X}|\theta)}_{\text{Likelihood}(\theta)} \underbrace{P(\theta)}_{\text{Prior}(\theta)} . \tag{34}$$

- Maximum A Posterior (**MAP**) estimation

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|\boldsymbol{X}) = \arg\max_{\theta} P(\boldsymbol{X}|\theta)P(\theta) \tag{35}$$

- Probabilistic model $\rightarrow$ (distribution) MAP $\rightarrow$ optimization (Variational Inference) or sampling (MCMC).
- The influence of prior decays with the increase of the number of samples.

Frequentist v.s. Bayesian

Frequentist

- **Pros:** more efficient in general, avoid the design of prior.
- **Cons:** non-robust to sparse data, cannot quantify the uncertainty of the estimation.

Bayesian

- **Pros:** prior makes it (relatively) robust to sparse data, quantify the uncertainty of the estimation (obtain the distribution of θ)
- **Cons:** require sophisticated design of prior, time-consuming in general.

In Summary

- We review more matrix operations and their properties
- A viewpoint of linear algebra is provided and connected with machine learning
- Some statistical concepts are provided and connected with matrix operations and machine learning

Next...

- Linear regression model
- Learning, evaluation, and some theoretical results.