ML Foundations

WEEK 0

Stats 1

① $P(A \text{ or } B) = 1/2$ for 5 question $\rightarrow (1/2)^5 = \boxed{0.03125}$

② $P(A) = 0.3 \quad P(B) = 0.4$

$P(A \cup B) = 0.6$

$P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.1$

$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.1}{0.4} = \boxed{0.25}$

③

category	prob.$^{<50}$	Freq.	tot. prob.
1	0.6	0.3	0.18
2	0.5	0.2	0.1
3	0.2	0.1	0.02
4	0.8	0.4	0.32

sum = 0.62

$P(4|<50) = \frac{0.32}{0.62} = 0.51613$

④ $M_1 \text{ — } 0.3 \text{ — } 0.02 \rightarrow 0.2449$

$M_2 \text{ — } 0.45 \text{ — } 0.03 \rightarrow 0.5510$

$M_3 \text{ — } 0.25 \text{ — } 0.02 \rightarrow 0.2041$

total probability = $0.006 + 0.0135 + 0.005 = 0.0245$

⑥ $\binom{2}{6} = \frac{6!}{2! \cdot 4!} = \frac{6 \times 5}{2} = 15$

⑧ $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$

$\text{Var}(X) = 2.917 + 0.25$

$= 3.167$

$E(X^2) = 15.167 \quad (E(X))^2 = 12.25$

$E(Y^2) = 0.5 \quad (E(Y))^2 = 0.25$

⑨ 60, 40, 40, 60

$1/3 \rightarrow 60$

$2/3 \rightarrow 40$

$E(X^2) = 2266.67, \quad E(X)^2 = 2177.78$

$\text{Var}(X) = 88.89, \quad SD(X) = 9.428$

M	T	W	Th	F	Sa	Su
–		–		–		3
	–		–		–	2
		–		–		2

⑦ $\frac{3}{4} \int_0^{\infty} (0.25)^x \, dx \Rightarrow \lim_{a \to \infty} \int_0^a (0.25)^x \, dx \Rightarrow \lim_{a \to \infty} \left[\frac{(0.25)^x}{\ln(0.25)} \right]_0^a \Rightarrow \left(\lim_{a \to \infty} \frac{(0.25)^a}{\ln(0.25)} - \frac{1}{\ln(0.25)} \right) \cdot \frac{3}{4}$

$= 0 - \frac{1}{\ln(0.25)} \Rightarrow \frac{-3}{4(\ln(0.25))}$

Stats 2

① $\text{Var}(X) = E(X^2) - (E(X))^2$

$E(X^2) = \int_0^1 4x^5 \, dx \Rightarrow \frac{4}{6} [x^6]_0^1 = \frac{2}{3}$

$E(X) = \int_0^1 4x^4 \, dx \Rightarrow \frac{4}{5} [x^5]_0^1 = 4/5$

$\text{Var}(X) = \frac{2}{3} - \frac{16}{25} = \frac{2}{75}$

② $k \int_{y=0}^{\infty} \int_{x=0}^{\infty} e^{-x} \cdot e^{-y} \, dx \, dy = 1 \Rightarrow k \int_{y=0}^{\infty} e^{-y} \underbrace{\left[-e^{-x}\right]_0^{\infty}}_{(-e^{\infty})-(-e^{0}) = 0-(-1)=1} dy = 1 \Rightarrow k \int_0^{\infty} e^{-y} \, dy = 1 \Rightarrow k = 1$

$\int_{y=0}^{5} \int_{x=5}^{\infty} e^{-x} \, e^{-y} \, dx \, dy \Rightarrow \int_0^5 e^{-y} \left[-e^{-x}\right]_5^{\infty} dy = e^{-5} \int_0^5 e^{-y} \, dy \Rightarrow e^{-5}(-e^{-5}+1)$

③ $X \sim Bin(1000, 1/6) \quad E[X/1000] = 1/6 \quad Var(X/1000) = \left(\frac{1}{1000}\right)^2 \cdot Var(X) = \frac{5}{36000}$

chebyshev's $= P(|X-\mu| \geq k) \leq Var(X)/k^2 \quad k = 0.2$

$P(|X/1000 - 1/6| \geq 0.2) \leq \frac{5}{36000} \cdot 25 = \boxed{\frac{5}{1440}}$

④ $L(x_1, x_2, \ldots, x_n) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right) \qquad \hat{\mu} = 0$

$\log(L) = -n \log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \Rightarrow \frac{\partial \log(L)}{\partial \sigma} = \frac{-n}{\sigma} + \frac{1}{\sigma^3} \Sigma (x_i - \hat{\mu})^2$

$0 = -n + \frac{1}{\sigma^2} \Sigma (x_i - \hat{\mu})^2 \Rightarrow \hat{\sigma}^2 = \frac{\Sigma (x_i - \hat{\mu})^2}{n} = \boxed{2/3}$

⑤ Likelihood $\sim N(\bar{y}, 25/10) \qquad$ Prior $\sim N(50, 25)$

Post. mean $= \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \left(\frac{\mu_0}{\sigma_0^2} + \frac{\Sigma x_i}{\sigma^2}\right) \qquad \mu_0 = 50 \quad \sigma_0^2 = 25 \quad \sigma = 25$

$= \frac{1}{\frac{1}{25} + \frac{10}{25}} \left(\frac{50}{25} + \frac{650}{25}\right) = \frac{25}{11} \cdot \frac{700}{25} = 63.636$

Maths 2

① $\lim_{x \to 0^+} f(x) = 0 \qquad \lim_{x \to 0^-} f(x) = 0 \qquad f(0) = 0$ ✓

$\lim_{h \to 0^-} \frac{f(0+h) - f(0)}{h} = \frac{f(h)}{h} = 1 \quad ; \quad \lim_{h \to 0^+} \frac{f(0+h) - f(0)}{h} = \lim_{h \to 0^+} \frac{h^2}{h} = 0$ ×

$\lim_{x \to 1^+} f(x) = \lim_{x \to 1^-} x^3 = 1 \qquad \lim_{x \to 1^-} f(x) = \lim_{x \to 1^-} x^2 = 1 \qquad f(1) = 1$ ✓

$\lim_{h \to 0^-} \frac{f(1+h) - f(1)}{h} = \frac{(1+h)^2 - 1}{h} = \frac{1 + 2h + h^2 - 1}{h} = \frac{h(2+h)}{h} = \lim_{h \to 0} 2 + h = 2$

$\lim_{h \to 0^+} \frac{f(1+h) - f(1)}{h} = \frac{(1+h)^3 - 1}{h} = \frac{1 + 3h + 3h^2 + h^3 - 1}{h} = \frac{h(3 + 3h + h^2)}{h} = \lim_{h \to 0} 3 + 3h + h^2 = 3$ ×

② $f'(x) = \begin{cases} -2x + 2 & 0 \leq x \leq 50 \\ 3x^2 & -50 \leq x < 0 \end{cases}$

$0 = -2x + 2 \Rightarrow x = 1 \qquad f(1) = 4$ glob max $\qquad f(50) = -2397$ loc min

$0 = 3x^2 \Rightarrow x = 0 \qquad f(-50) \approx -(50)^3$ glob min

③ $f'(x) = 3x^2 - 6$

$0 = 3x^2 - 6 \Rightarrow 2 = x^2 \Rightarrow x = \sqrt{2}, -\sqrt{2}$

$f''(x) = 6x \quad ; \quad f''(\sqrt{2}) = 6\sqrt{2}$ (min) $\quad ; \quad f''(-\sqrt{2}) = -6\sqrt{2}$ (max)

④ $A^2 = \begin{bmatrix} 2 & 2 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 2 & 2 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 4 & 8 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} 2 & 2 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 8 & 24 \\ 0 & 8 \end{bmatrix}$

$AB = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \begin{bmatrix} b_1 & b_2 \\ b_3 & b_4 \end{bmatrix} = \begin{bmatrix} a_1 b_1 + a_2 b_3 & a_1 b_2 + a_2 b_4 \\ a_3 b_1 + a_4 b_3 & a_3 b_2 + a_4 b_4 \end{bmatrix}$

$A = \begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix} \cdot B = \begin{bmatrix} g & h & i \\ j & k & l \end{bmatrix}$

$\begin{matrix} (ab+gj) & (ab+hk) & (ab+il) \\ (cd+gj) & (cd+hk) & (cd+il) \\ (ef+gj) & (ef+hk) & (ef+il) \end{matrix}$

⑤ $A = \left[\begin{array}{cccc|c} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \end{array}\right]$

⑦ $A = \left[\begin{array}{ccc|c} -1 & 1 & 2 & 1 \\ 2 & 1 & -2 & -1 \\ 0 & 3 & c & d \end{array}\right] \xrightarrow[R_2+2R_1]{-R_1} \left[\begin{array}{cccc} 1 & -1 & -2 & -1 \\ 0 & 3 & 2 & 1 \\ 0 & 3 & c & d \end{array}\right] \xrightarrow{R_2/3} \left[\begin{array}{cccc} 1 & -1 & -2 & -1 \\ 0 & 1 & 2/3 & 1/3 \\ 0 & 3 & c & d \end{array}\right] \xrightarrow[R_3-3R_2]{R_1+R_2} \left[\begin{array}{cccc} 1 & 0 & -4/3 & -2/3 \\ 0 & 1 & 2/3 & 1/3 \\ 0 & 0 & c-2 & d-1 \end{array}\right]$

⑧ $\begin{bmatrix} c & 1 & -1 \\ -1 & 0 & -3 \\ -2 & -1 & c \end{bmatrix}$

$\det(A) = 3c + (-c+6) - (1)$

$0 = 2c + 5 \Rightarrow c = 5/2 = 2.5$

⑨ $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad A^T = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$

$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

WEEK 1

→ Why and when ML?
- ↳ Programming / Human labour fails ?
  - ① scale / speed / cost of human labour
  - ② inability to express rules using language
  - ③ don't know exact rules
- ↳ ML can succeed when
  - ① Have lots of example data
  - ② Have some structural idea on the rules

→ Machine learning applications
- ↳ e-mail inbox spam classification
- ↳ recommender systems (e-commerce, social media, streaming services, etc.)
- ↳ smart assistants
- ↳ robot AIs
- ↳ games (chess, go)
- ↳ social media marketing

→ Types of Models in ML
- ↳ Predictive Model
  - ↳ Regression ; e.g. → model the price of a house based on its area
    - ↳ real-valued output
  - ↳ Classification ; e.g. → whether a house is closer than 2kms to metro based on price
    - ↳ discrete valued output
- ↳ Probabilistic Model
  - ↳ goal is not to predict future
  - ↳ to evaluate how likely an event is

→ Learning Algorithms
- ↳ learning algorithms : Data → Models
  - ↳ choose from a collection of models, with same structure but different parameters

Notation

- $\mathbb{R}$: real numbers, $\mathbb{R}_+$: Positive reals, $\mathbb{R}^d$: d-dimensional vector of reals.
- $\mathbf{x}$: vector. x_j: j^{th} co-ordinate. $\|\mathbf{x}\|$: Length of vector $\mathbf{x}$.
- $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$: Collection of n vectors.
- x_j^i : j^{th} co-ordinate of i^{th} vector.
- $(x_1)^2$: Square of the first co-ordinate of the vector $\mathbf{x}$
- $\mathbf{1}(2 \text{ is even}) = 1, \mathbf{1}(2 \text{ is odd}) = 0.$

↳ true is 1 ↳ false is 0

→ Supervised Learning

↳ curve-fitting at its core

↳ Given $\{(x^1, y^1), (x^2, y^2), \ldots, (x^n, y^n)\}$, find a model f such that $f(x^i)$ is 'close' to y^i

↳ Regression

→ eg. Predict house price from room, area, distance

→ takes training data: $\{(x^1, y^1), (x^2, y^2), \ldots, (x^n, y^n)\}$

→ algorithm outputs a model $f : \mathbb{R}^d \to \mathbb{R}$

→ Loss $= \frac{1}{n} \sum_{i=1}^{n} (f(x^i) - y^i)^2$

→ $f(x) = w^T x + b = \sum_{j=1}^{d} w_j x_j + b$

↳ Classification

→ e.g. Predict if rooms >3 from area and price

→ training data

→ $x^i \in \mathbb{R}^d$, $y^i \in \{+1, -1\}$

→ Algorithm outputs a model $f : \mathbb{R}^d \to \{+1, -1\}$

→ Loss $= \frac{1}{n} \sum_{i=1}^{n} 1(f(x^i) \neq y^i)$ → fraction of misclassified instances

→ $f(x) = \text{sign}(w^T x + b)$ ⟶ a.k.a. linear separator

↳ uses training data to get model f

↳ " test data not in the training data for model evaluation

↳ " validation data for model selection

→ Unsupervised Learning

↳ much more vague than supervised learning

↳ usually used to 'understand data'

↳ build models that compress, explain and group data

↳ Dimensionality Reduction

→ compression and simplification

→ Data: $\{x^1, x^2, \ldots, x^n\}$

$x^i \in \mathbb{R}$

Encoder $f : \mathbb{R}^d \to \mathbb{R}^{d'}$

Decoder $g : \mathbb{R}^{d'} \to \mathbb{R}^d$

Goal : $g(f(x^i)) \approx x^i$

Loss $= \frac{1}{n} \sum_{i=1}^{n} \| g(f(x^i)) - x^i \|^2$

↳ Density Estimation

→ probabilistic model

→ Data: $\{x^1, x^2, \ldots, x^n\}$

$x^i \in \mathbb{R}$

Probability mapping $P : \mathbb{R}^d \to \mathbb{R}_+$ that 'sums' to one

Goal : $P(x)$ is large if $x \in$ Data, and low otherwise

Loss $= \frac{1}{n} \sum_{i=1}^{n} -\log(P(x))$ ⟶ a.k.a. negative log likelihood

WEEK 2

Sets

- $\mathbb{R}$ → real numbers
  $\mathbb{R}_+$ → +ve real numbers
  $\mathbb{Z}$ → integers
  $\mathbb{Z}_+$ → +ve integers

- $[a,b] = \{x \in \mathbb{R} : a \le x \le b\}$
  $(a,b) = \{x \in \mathbb{R} : a < x < b\}$

- $\mathbb{R}^d$ → set of d-dimensional vectors $= \mathbb{R} \times \mathbb{R} \times \dots \times \mathbb{R}$ ← d times
- $[a,b]^d \to \{x \in \mathbb{R}^d : x_i \in [a,b],\ i \in \{1,2,\dots,d\}\}$

Metric Space

- A set with a distance function associated with it
- Default metric space = $\mathbb{R}^d$: $D(x,y) = \|x-y\| = \sqrt{(x_1-y_1)^2 + \dots + (x_d-y_d)^2}$
- An open ball with radius ϵ with center x
  - $B(x,\epsilon) = \{y \in \mathbb{R}^d : D(x,y) < \epsilon\}$
  - closed ball → $\bar{B}(x,\epsilon) = \{y \in \mathbb{R}^d : D(x,y) \le \epsilon\}$
    (x: center, ϵ: radius)

Logic Statements / Modifiers

- $\forall$: for all
- $\exists$: there exists
- $\Rightarrow$: implies
- $\Leftrightarrow$: equivalent

Sequence

- ordered collection of elements
- example: $x_n = \left(1 + \frac{4}{2^n},\ 3 - \frac{4}{2^n}\right)$ → each element is in $\mathbb{R}^2$ and as n increases, it will approach $(1,3)$
- $x_1, x_2, \dots$ where $x_i \in \mathbb{R}^d$

$$\lim_{i \to \infty} x_i = x^* \Leftrightarrow \forall \epsilon > 0, \exists N \text{ s.t. } x_n \in B(x^*, \epsilon)\ \forall n \ge N$$

- interpretation: after a point N, all the terms of sequence stay within a ball. whatever ϵ, there exists some integer N, such that after N, x_n is going to stay within a ball of radius ϵ centered x^*.

Vector Space

- collection of vectors that must satisfy certain properties
  - most crucial property: if V is a vector space, $u \in V$, $v \in V$, $\alpha, \beta \in \mathbb{R}$ then $\alpha u + \beta v \in V$
- $x, y \in V$; $x \cdot y = x^T y = \sum_{i=1}^d x_i y_i$ → (dot product)
- $\|x\|^2 = x \cdot x = x^T x = \sum_{i=1}^d (x_i)^2$ → (norm)
- x & y are perpendicular (orthogonal) if $x \cdot y = x^T y = \sum_{i=1}^d x_i y_i = 0$

→ Functions and Graphs

↳ graph of a function in n-dimensional space is $n+1$ dimensions

example of 1-dimensional function $f : \mathbb{R} \to \mathbb{R}$ $f(x) = x^2$ $G_f = \{(x, x^2) : x \in \mathbb{R}\} \subseteq \mathbb{R}^2$

↳ contour plots of 2-dimensional function

e.g. $f : \mathbb{R}^2 \to \mathbb{R}$ $f(x) = x_1 + x_2$ values it can take = $\{-1, 0, 1\}$ (maybe more)

$f(x) = -1$, plot all possible x_1, x_2 s.t. $x_1 + x_2 = -1$

$f(x) = 0$, " " " " $x_1 + x_2 = 0$

→ Continuity

↳ $f : \mathbb{R} \to \mathbb{R}$ is continuous at $x^* \in \mathbb{R}$ if $\forall$ sequences converging to x^*, $f(x_i)$ converges to $f(x^*)$

$$\lim_{i \to \infty} x_i = x^* \Rightarrow \lim_{i \to \infty} f(x_i) = f(x^*) \qquad \lim_{x \to x^*} f(x) = f(x^*)$$

→ Differentiability

↳ $f : \mathbb{R} \to \mathbb{R}$ is differentiable at $x^* \in \mathbb{R}$ if $\lim_{x \to x^*} \frac{f(x) - f(x^*)}{x - x^*}$ exists.

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

→ Derivatives and Linear Approximation

↳ let $f : \mathbb{R} \to \mathbb{R}$

$$f'(x^*) = \lim_{x \to x^*} \frac{f(x) - f(x^*)}{x - x^*}$$

$$f'(x^*) \approx \frac{f(x) - f(x^*)}{x - x^*} \quad \text{——— around } x = x^*$$

$$f(x) \approx \underbrace{f(x^*) + f'(x^*)(x - x^*)}_{L_{x^*}[f](x)} \quad \text{——— around } x = x^*$$

→ Higher Order Approximation

↳ $f(x) \approx f(x^*) + f'(x^*)(x - x^*) + \frac{1}{2} f''(x^*)(x - x^*)^2$ ——— quadratic approximation

ex. give lin. approx. of $f(x) = e^{\sqrt{1+x}}$ around $x = 1$

$$f(x) = f(x^*) + f'(x^*)(x - x^*)$$

$$f(x) = f(1) + f'(1)(x - 1)$$

$$= \boxed{e^{\sqrt{2}} + \frac{e^{\sqrt{2}}}{2\sqrt{2}}(x - 1)}$$

$$f'(x) = e^{\sqrt{1+x}} \cdot \frac{1}{2\sqrt{1+x}}$$

$$f'(1) = \frac{e^{\sqrt{2}}}{2\sqrt{2}}$$

Geometry of Lines

- A line in $\mathbb{R}^d \subseteq \mathbb{R}^d$
- A line through the point $u \in \mathbb{R}^d$ along the vector $v \in \mathbb{R}^d$
  - $= \{x \in \mathbb{R}^d : x = u + \alpha v \text{ for } \alpha \in \mathbb{R}\}$

A line through the points $u, u' \in \mathbb{R} = \{x \in \mathbb{R}^d : x = u + \alpha(u' - u) \text{ for } \alpha \in \mathbb{R}\}$

Geometry of (Hyper)planes

- A $(d-1)$-dimensional hyperplane $\in \mathbb{R}^d$
- A hyperplane normal to the vector $w \in \mathbb{R}^d$ with $b \in \mathbb{R}$
  - $= \{x \in \mathbb{R}^d : w^T x = b\}$
  - $= \{x \in \mathbb{R}^d : \sum_{i=1}^{d} w_i x_i = b\}$

Gradient and Linear Approximations

- $v, x \in \mathbb{R}^d$ and x is close to v

$$f(x) \approx f(v) + \nabla f(v)^T (x - v)$$
$$= f(v) + \sum_{i=1}^{d} \frac{\partial f}{\partial x_i}(v) \cdot (x_i - v_i)$$

- $f : \mathbb{R}^2 \to \mathbb{R}$ $\quad f(y_1, y_2) - f(v_1, v_2) \approx \frac{\partial f}{\partial x_1}(v)(y_1 - v_1)$
- e.g. → approx of $f(x_1, x_2) = x_1^2 + x_2^2$ around $(3, 1)$?

$$f(x) = f(3,1) + [6, 2] \begin{bmatrix} x_1 - 3 \\ x_2 - 1 \end{bmatrix}$$
$$= 10 + 6x_1 - 18 + 2x_2 - 2$$
$$= 6x_1 + 2x_2 - 10$$

$\nabla f(x) = (2x_1, 2x_2)$
$\nabla f(v) = (6, 2)$
$f(v) = 10$

Directional Derivative

- $D_u[f](v)$ → directional derivative of f at the point v, along u

$$D_u[f](v) = \nabla f(v)^T (u)$$

Cauchy-Schwarz Inequality

- two d-dimensional vectors a & b
  - $-\|a\| \cdot \|b\| \leq a^T b \leq \|a\| \cdot \|b\|$
    - $a = \alpha b$, $\alpha < 0$
    - $a = \alpha b$, $\alpha > 0$ → to maximize $a^T b$

Direction of Steepest Ascent

- Find a direction u, that maximises the rate of change of f as you move from v along u.

Maximise $D_u[f](v)$

$f(x, y) = x e^{xy}$
$\nabla f(x, y) = (e^{xy} + xy e^{xy}, x^2 e^{xy})$
$f(1, 0) = 1$
$\nabla f(1, 0) = (1, 1)$

$$L(x, y) = f(1,0) + \frac{\partial f}{\partial x}(1,0)(x-1) + \frac{\partial f}{\partial y}(1,0)(y-0)$$
$$= 1 + 1(x-1) + y = x + y$$

$f(x) = \sqrt{x+4}$ $\quad f(5) = 3$
$f'(x) = \frac{1}{2\sqrt{x+4}}$ $\quad f'(5) = 1/6$

$$L_a[f](x) = f(a) + f'(a)(x - a)$$
$$= 3 + \frac{1}{6}(x - 5)$$
$$= \frac{19}{6}$$

WEEK 3

→ Four Fundamental Subspaces

↳ Column Space

$$A = \begin{bmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & & | \end{bmatrix}$$

$C(A) = \text{span}(u_1, u_2, \ldots, u_n) = \{\text{linear combinations of the vectors } u_1, u_2, \ldots, u_n\}$

Solving $Ax = b$

↳ For what b does $Ax = b$ have a solution?

↳ $\forall\, b \in C(A)$

e.g.

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 3 \\ 3 & 1 & 4 \\ 4 & 1 & 5 \end{bmatrix}$$

→ some vectors in $\mathbb{R}^4$ are not in $C(A)$ because $Ax = b$ is "4 equations in 3 unknowns"

→ col 3 = col 1 + col 2. So, $C(A) = \text{span}(\text{col 1}, \text{col 2})$

$C(A)$ is 2-dim. subspace of $\mathbb{R}^4$.

→ Pivot columns in RREF

↳ Null Space

$$N(A) = \{x \mid Ax = 0\}$$

e.g.

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 3 \\ 3 & 1 & 4 \\ 4 & 1 & 5 \end{bmatrix}$$

→ column 1 + column 2 = column 3. So the vector $(1, 1, -1)$ is in $N(A)$.

Note: If A is invertible, then $N(A)$ has "zero" only, and $C(A)$ is the whole space. $Ax = b$ has a unique solution.

Else, $N(A)$ has $x_n \neq 0$ and $Ax = b$ solutions are of the form

$x = x_p + x_n$, s.t. $Ax_p = b$, $Ax_n = 0$

Note: Gaussian elimination to find $N(A)$. Solve $Ax = 0$. Take the pivot variables and set them to basis entry

Rank = # of pivot columns — Rank = $\dim(C(A))$

Nullity = # of free variables — Nullity = $\dim(N(A))$

↳ Row Space

↳ column space of A^T ⇔ span of rows of A

↳ col rank = $\dim(C(A))$ — row rank = $\dim(R(A))$

col rank = row rank

↳ Left Null Space

↳ $N(A^T) = \{y \mid A^T y = 0\} = \{y \mid y^T A = 0\}$

↳ linear combination of rows leading to zero vector

↳ A is $m \times n$ matrix

$\dim(C(A)) + \dim(N(A)) =$ # of col. in $A = n$

$\dim(C(A^T)) + \dim(N(A^T)) =$ " " rows in $A = m$

$\dim(C(A^T)) = r \Rightarrow r + \dim(N(A^T)) = m \Rightarrow \dim(N(A^T)) = m - r$

↳ e.g.

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix}_{2\times 2}$$

$C(A)$ = line through $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$

$$A^T = \begin{bmatrix} 1 & 3 \\ 2 & 6 \end{bmatrix}$$

$N(A)$ = line through $\begin{bmatrix} -2 \\ 1 \end{bmatrix}$

$C(A^T)$ = line through $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$N(A^T)$ = line through $\begin{bmatrix} -3 \\ 1 \end{bmatrix}$

→ Orthogonality

↳ length of a vector $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$, $\|x\|^2 = x_1^2 + x_2^2 + \cdots + x_n^2$

↳ Orthogonal vector

↳ $x \perp y$ if $x^T y = 0$ (inner product)

↳ if $\{v_1, v_2, \ldots, v_n\}$ are mutually orthogonal, then they are linearly independent. [To check, solve for $Ax = 0$.

↳ Orthonormal vector

↳ $\{u, v\}$ are orthonormal if $u^T v = 0$ and $\|u\| = \|v\| = 1$

↳ Orthogonal subspaces

↳ U, V are orthogonal subspaces if $x^T y = 0 \ \forall x \in U, v \in V$

↳ e.g.

$$U = \text{span}\left\{\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}\right\} \quad \text{and} \quad V = \text{span}\left\{\begin{pmatrix} 0 \\ 0 \\ 2 \\ 1 \end{pmatrix}\right\}$$

↳ Orthogonality w.r.t four fundamental subspaces of a matrix A

① $R(A) \perp N(A) \Leftrightarrow C(A^T) \perp N(A)$

→ $x \in N(A) \Rightarrow Ax = 0$

$$\begin{bmatrix} - \text{ row 1} - \\ - \text{ row 2} - \\ \vdots \\ - \text{ row m} - \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = 0$$

implies row 1 $\perp x$, row 2 $\perp x$, ...

② $C(A) \perp N(A^T)$

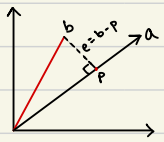
→ because $C(A^T) \perp N(A)$

→ Projections

↳ Why?

Inconsistent system of linear equations, i.e. for a system $Ax = b$, $b \notin C(A)$. In such situations, it makes sense to project b onto C(A)

↳ Projection onto a line



$p = \hat{x}a$; $e = b - p = b - \hat{x}a$; $e \perp a$

$(b - \hat{x}a) \perp a$

$a(b - \hat{x}a) = 0 \Rightarrow a^Tb - a^T\hat{x}a = 0$

$\hat{x} = \frac{a^Tb}{a^Ta}$

projection of b onto $a = \left(\frac{a^Tb}{a^Ta}\right)a$

Cauchy-Schwarz inequality

$$\|e\|^2 = \|b-p\|^2 \geq 0 \Rightarrow \left\|b - \frac{a^Tb}{a^Ta}\right\|^2 = b^Tb - \frac{2(a^Tb)^2}{a^Ta} + \left(\frac{a^Tb}{a^Ta}\right)^2 a^Ta$$

$$= \frac{(b^Tb)(a^Ta) - (a^Tb)^2}{(a^Ta)} \geq 0 \Rightarrow (b^Tb)(a^Ta) \geq (a^Tb)^2$$

$$= \boxed{|a^Tb| \leq \|a\| \cdot \|b\|}$$

↳ Projection matrix

$$p = \left(\frac{a^Tb}{a^Ta}\right)a = \left(\frac{aa^T}{a^Ta}\right)b \qquad \text{Let } P = \frac{aa^T}{a^Ta}$$

Then, projection of b onto a is $P \cdot b$

↳ just left multiply by the projection matrix.

e.g.

$$a = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad P = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

① P is symmetric

② $P^2 = P$ i.e., $P^2b = Pb$

③ C(P) = line through a ; N(P) = plane orthogonal to a

Rank (P) = 1

→ Least Squares

↳ $2x = b_1$, $3x = b_2$, $4x = b_3$ } system is solvable if b is on line through $\begin{bmatrix}2\\3\\4\end{bmatrix}$

- if b is such that the system is inconsistent
  ↳ pick a subset of equations and solve. Problem: large errors in some inputs and no errors in other.
  ↳ solution: minimize average error
  ↳ $E^2 = (2x-b_1)^2 + (3x-b_2)^2 + (4x-b_3)^2$

↳ minimize E^2

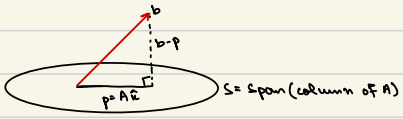
$$\frac{dE^2}{dx} = 0 \Leftrightarrow 2\left[2(2x-b_1) + 3(3x-b_2) + 4(4x-b_3)\right] = 0$$

leads to $\hat{x} = \frac{2b_1 + 3b_2 + 4b_3}{(2)^2 + (3)^2 + (4)^2} = \frac{a^T b}{a^T a}$, with $a = \begin{bmatrix}2\\3\\4\end{bmatrix}$

→ taking a derivative and finding the minima is the same as taking a projection onto the column space

↳ Projection onto a subspace

→ project b onto $C(A)$



orthogonal vector $e = b - p = b - A\hat{x}$

→ $e \perp$ ∀ vectors in $C(A)$ and $C(A) \perp N(A^T)$, therefore

$e \in N(A^T) \Rightarrow A^T e = 0 \Rightarrow A^T(b - A\hat{x}) = 0$

→ $\boxed{A^T A\hat{x} = A^T b}$ → equation to solve for proj. of b onto $C(A)$.

↳ even if $Ax = b$ is inconsistent, $A^T A\hat{x} = A^T b$ has a solution.

→ if columns of A are lin. independent, then $A^T A$ is invertible.
Solving $A^T A\hat{x} = A^T b$ when $(A^T A)$ is invertible $\Rightarrow \hat{x} = (A^T A)^{-1} A^T b$
Projection $P = A\hat{x} = A(A^T A)^{-1} A^T b$

→ if $b \in C(A)$, then $b = Ax$
↳ projection $= A(A^T A)^{-1} A^T A x = Ax = b$

→ if $b \in N(A^T)$, then
$p = A(A^T A)^{-1} A^T b = 0$ because $A^T b = 0$

→ if A is square and invertible $\Leftrightarrow C(A) = \mathbb{R}^n$
$p = A(A^T A)A^T b = AA^{-1}(A^T)^{-1} A^T b = b$

→ if Rank $(A) = 1$, then
$$\hat{x} = \frac{a^T b}{a^T a}$$

↳ Projection matrix

$$P = A(A^TA)^{-1}A^T$$

→ $P^T = P$; $(A(A^TA)^{-1}A^T)^T = A^T(A^TA)^{-1}A = P$

→ $P^2 = P$; $P^2 = A(A^TA)^{-1}A^T\,A(A^TA)^{-1}A^T = A(A^TA)^{-1}A = P$

→ projection matrix is symmetric and satisfies $P^2 = P$

↳ if a matrix A is symmetric and satisfies $A^2 = A$, then A is a projection matrix.

→ Pb = projection of b onto the column space of P

→ Examples of least squares

↳ One-dimensional example:

dataset: $(x_1, b_1) \dots (x_m, b_m)$

$b_i = \theta x_i + \theta'$ ← offset

① system of equations:

$$\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{bmatrix} \begin{bmatrix} \theta' \\ \theta'' \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \qquad A\theta = b \text{, where } \theta = \begin{bmatrix} \theta' \\ \theta'' \end{bmatrix}$$

② least squares approach: minimize $E^2 = \|b - A\theta\|^2 = (b_1 - \theta' x_1 - \theta'')^2 + \dots + (b_m - \theta' x_m - \theta'')^2$

$$(\hat{\theta}', \hat{\theta}'') = \arg\min_{\tilde{\theta}} \|b - A\theta\|^2$$

e.g. $A\theta = b$ (constant offset)

$$A = \begin{bmatrix} -1 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} \theta' \\ \theta'' \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix}$$

$A^TA\hat{\theta} = A^Tb$, where $\hat{\theta} = \begin{bmatrix} \hat{\theta}' \\ \hat{\theta}'' \end{bmatrix}$

$$A^TA = \underset{2\times3}{\begin{bmatrix} -1 & 1 & 2 \\ 1 & 1 & 1 \end{bmatrix}} \underset{3\times2}{\begin{bmatrix} -1 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix}} = \begin{bmatrix} 6 & 2 \\ 2 & 3 \end{bmatrix}; \quad \underset{2\times3}{\begin{bmatrix} -1 & 1 & 2 \\ 1 & 1 & 1 \end{bmatrix}} \underset{3\times1}{\begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix}} = \begin{bmatrix} 6 \\ 5 \end{bmatrix}$$

$$\begin{bmatrix} 6 & 2 \\ 2 & 3 \end{bmatrix}\hat{\theta} = \begin{bmatrix} 6 \\ 5 \end{bmatrix} \qquad \left[\begin{array}{cc|c} 6 & 2 & 6 \\ 2 & 3 & 5 \end{array}\right] \rightsquigarrow \left[\begin{array}{cc|c} 1 & 1/3 & 1 \\ 2 & 3 & 5 \end{array}\right] \rightsquigarrow \left[\begin{array}{cc|c} 1 & 1/3 & 1 \\ 0 & 7/3 & 3 \end{array}\right]$$

$\hat{\theta}' = 4/7 \quad \hat{\theta}'' = 9/7$

WEEK 4

→ Linear Regression

↳ Given data $\{(x_1,y_1),\dots,(x_n,y_n)\}$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, $i = 1,\dots,n$

$$L(\theta) = \frac{1}{2}\sum_{i=1}^{n}(x_i^T\theta - y_i)^2 \rightarrow \text{Loss Function}$$

Minimise L: Define $A = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}$ ← feature matrix , $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$

$$A\theta - y_i = \begin{bmatrix} x_1^T\theta - y_1 \\ \vdots \\ x_n^T\theta - y_n \end{bmatrix} \qquad (A\theta - Y)^T(A\theta - Y) = \sum_{i=1}^{n}(x_i^T\theta - y_i)^2$$

Minimising L: $\nabla_\theta L(\theta) = 0 \iff \nabla_\theta((A\theta - Y)^T(A\theta - Y)) = 0$

$$= A^T(A\theta - Y) = 0$$

$$= (A^TA)\theta = A^TY \rightarrow \text{Least squares solution}$$

$$\theta = (A^TA)^{-1}A^TY \quad \text{only iff } A \text{ is Full rank}$$

↳ Maximum Likelihood and least squares regression end up solving the same equation, under a linear model.

→ Polynomial Regression

↳ given one-dimensional data: $\{(x_1,y_1),\dots,(x_n,y_n)\}$ $x_i, y_i \in \mathbb{R}\ \forall i$

↳ polynomial of degree m

Transformed features:

$$\hat{y}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots + \theta_m x^m$$

$$= \sum_{j=0}^{m}\theta_j \phi_j(x) \text{, where } \phi_j(x) = x^j$$

→ for given x, transformed feature vector $\phi(x) = (1, x, x^2, \dots, x^m)$

$$\hat{y}(x) = \theta^T\phi(x)$$

$$A = \begin{bmatrix} \phi(x_1)^T \\ \phi(x_2)^T \\ \vdots \\ \phi(x_n)^T \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$(A^TA)\theta = A^TY$ → using the transformed features

→ Regularised version of linear regression is a.k.a. **Ridge regression**

↳ instead of solving $\min_\theta \frac{1}{2}\sum_{i=1}^{n}(x_i^T\theta - y_i)^2$,

we solve → $\min_\theta \bar{L}(\theta) = \frac{1}{2}\sum_{i=1}^{n}(x_i^T\theta - y_i)^2 + \underbrace{\lambda\|\theta\|^2}_{\text{regularisation term}}$

→ Eigenvalues and Eigenvectors

↳ **Motivation**: ODEs

$$\frac{dv}{dt} = 4v - 5w \quad , v = 8 \text{ at } t = 0$$

$$\frac{dw}{dt} = 2v - 3w \quad , w = 5 \text{ at } t = 0$$

$$u(t) = \begin{bmatrix} v(t) \\ w(t) \end{bmatrix} \; ; \; u(0) = \begin{bmatrix} 8 \\ 5 \end{bmatrix}$$

$$\frac{du}{dt} = Au \text{ , where } A = \begin{bmatrix} 4 & -5 \\ 2 & -3 \end{bmatrix} \text{ and } u = u(0) \text{ at } t = 0$$

→ if want solutions of this form:

$$v(t) = e^{\lambda t} y \quad , \quad w(t) = e^{\lambda t} z$$

or $u(t) = e^{\lambda t} x$, where $x = \begin{bmatrix} y \\ z \end{bmatrix}$

Substitute:

$$\lambda e^{\lambda t} y = 4e^{\lambda t} y - 5e^{\lambda t} z$$
$$\lambda e^{\lambda t} z = 2e^{\lambda t} y - 3e^{\lambda t} z$$

$$\boxed{\begin{aligned} 4y - 5z &= \lambda y \\ 2y - 3z &= \lambda z \end{aligned}} \iff \boxed{Ax = \lambda x}$$

eigenvalue equation; eigenvalue; eigenvector

→ can solve $\frac{du}{dt} = Au$ using solutions of the form $u(t) = e^{\lambda t} x$

if $Ax = \lambda x$ can be solved.

↳ For a matrix A, λ is eigenvalue and $x \neq 0$ is eigenvector if $Ax = \lambda x$

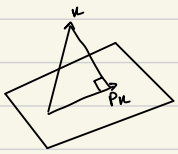
if x is eigenvector then Ax either stretches or shrinks x, but in the same direction.

→ if $\lambda = 0$, then $x \in N(A)$

↳ **examples**:

① Projection matrix P

↳ project onto a plane



(i) $Px = x$ for any x in the plane

↳ $\lambda = 1$ is an eigenvalue and any x in the plane is an eigenvector

(ii) if an x is $\perp$ to the plane, $Px = 0$

So, $\lambda = 0$ is an eigenvalue and any $x \perp$ to the plane is an eigenvector.

② Permutation matrix

↳ identity matrix with the rows shuffled.

e.g. $B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

$Bx = x$ for $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $Bx = -x$ for $x = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

↳ Finding eigenvalues:

↳ $Ax = \lambda x \Leftrightarrow (A-\lambda I)x = 0 \Leftrightarrow (A-\lambda I)$ is singular

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

characteristic polynomial of matrix A

→ If A is $n \times n$, then characteristic polynomial is of degree n

↳ $(a_{11}-\lambda) \dots (a_{nn}-\lambda) = 0$

"n" roots of characteristic polynomial = eigenvalues

→ $Ax = \lambda x$

$\Rightarrow Ax - \lambda x = 0$

$\Rightarrow Ax - \lambda I x = 0$

$\Rightarrow (A-\lambda I)x = 0$

$\Rightarrow A - \lambda I = 0$

$\det(A-\lambda I) = 0$

For eigenvectors: Given λ, want $(A-I\lambda)x = 0$

i.e., $x \in \underbrace{N(A-I\lambda)}$

find using Gaussian elimination

↳ example

① $A = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$ $\lambda I = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$ $A - \lambda I = \begin{bmatrix} 3-\lambda & 1 \\ 1 & 3-\lambda \end{bmatrix}$

$\det(A-\lambda I) = (3-\lambda)^2 - 1 = 9 - 6\lambda + \lambda^2 - 1 = \lambda^2 - 6\lambda + 8$

$\lambda^2 - 6\lambda + 8 = 0 \rightarrow \lambda^2 - 4\lambda - 2\lambda + 8 = 0 \rightarrow \lambda(\lambda-4) - 2(\lambda-4) = 0$

$(\lambda-2)(\lambda-4) = 0$

eigenvalues: $\lambda = 2$, $\lambda = 4$

$\lambda_1 + \lambda_2 = \text{trace}(A)$

$\lambda_1 \cdot \lambda_2 = \det(A)$

eigenvectors: $A - 4I = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$; $(A-4I)x = 0$ for $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$A - 2I = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$; $(A-2I)x = 0$ for $x = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

② variation of ①

$B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ $B + 3I = A$

$Ax = (B+3I)x = \lambda x + 3x = (\lambda+3)x$

eigenvectors of B = eigenvectors of A

↳ Remarks

① Suppose $Ax = \lambda_1 x$ and $Bx = \lambda_2 x$

Then, $(A+B)x \neq (\lambda_1 + \lambda_2)x$ because x can be diff. for A and B.

② Symmetric matrix has "real" eigenvalues.

for matrix $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, eigenvalues: $\lambda = i$, $\lambda_2 = -i$

③ A matrix with linearly dependent eigenvectors:

$A = \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix}$ $\lambda = \lambda_2 = 3$ $x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

→ Similarity and Diagonalisation

↳ Matrix A is diagonalisable if $\exists$ an invertible matrix S such that $S^{-1}AS = \Lambda$ ⟶ diagonal matrix

↳ **Matrix is diagonalisable if it has enough independent eigenvectors**

Suppose A is $n \times n$ matrix with n lin. independent eigenvectors $= \{\kappa_1, \kappa_2, \dots, \kappa_n\}$

Let $S = \begin{bmatrix} | & | & & | \\ \kappa_1 & \kappa_2 & \cdots & \kappa_n \\ | & | & & | \end{bmatrix}$ S is invertible because rank$(S) = n$

$$AS = A\begin{bmatrix} | & & | \\ \kappa_1 & \cdots & \kappa_n \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ A\kappa_1 & \cdots & A\kappa_n \\ | & & | \end{bmatrix}$$

$$= \begin{bmatrix} | & | & & | \\ \lambda_1\kappa_1 & \lambda_2\kappa_2 & \cdots & \lambda_n\kappa_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ \kappa_1 & \cdots & \kappa_n \\ | & & | \end{bmatrix}\begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} = S\Lambda, \quad \Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$$

So, $AS = S\Lambda \Leftrightarrow S^{-1}AS = \Lambda$

↳ eigenvectors corresponding to different eigenvalues are linearly independent.

↳ Remarks:

① $S^{-1}AS = \Lambda$ or $A = S\Lambda S^{-1}$

S is not unique

② $A = S\Lambda S^{-1}$

Suppose col. of S is y. Col. of $S\Lambda = \lambda_1 y$. Col. of $AS = Ay$

$Ay = \lambda_1 y \Rightarrow \lambda_1$ is an eigenvalue and y is an eigenvector

③ Powers of A

Suppose λ is an eigenvalue and κ is an eigenvector

$A^2\kappa = A(A\kappa) = \lambda A\kappa = \lambda^2\kappa$

Then, $S^{-1}A^2S = \Lambda^2$

④ Not all matrices are diagonalisable

↳ Fibonacci sequence

$F_{k+2} = F_{k+1} + F_k$ Q. What is F_{100}?

system of equations: $F_{k+2} = F_{k+1} + F_k$

$F_{k+1} = F_{k+1}$

$$u_k = \begin{bmatrix} F_{k+1} \\ F_k \end{bmatrix}, \quad u_{k+1} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} u_k \Leftrightarrow \begin{bmatrix} F_{k+2} \\ F_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} F_{k+1} \\ F_k \end{bmatrix}$$

Let $A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ $\quad u_0 \xrightarrow{A} u_1 \xrightarrow{A} u_2 \xrightarrow{A} \dots$

$u_k = A^k u_0$ is a solution to $u_{k+1} = Au_k$

If A is diagonalisable, then $u_0 = c_1\kappa_1 + c_2\kappa_2 + \dots + c_n\kappa_n$

$u_k = A^k u_0 = c_1\lambda_1^k\kappa_1 + c_2\lambda_2^k\kappa_2 + \dots + c_n\lambda_n^k\kappa_n$

$A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ $\quad \lambda_1 = \frac{1+\sqrt{5}}{2}, \quad \lambda_2 = \frac{1-\sqrt{5}}{2}$

with $x_1 = \begin{bmatrix} \lambda_1 \\ 1 \end{bmatrix}$ and $x_2 = \begin{bmatrix} \lambda_2 \\ 1 \end{bmatrix}$

$$(A - \lambda_1 I)\begin{bmatrix} \lambda_1 \\ 1 \end{bmatrix} = \begin{bmatrix} \lambda_1^2 - \lambda_1 - 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$(A - \lambda_2 I)\begin{bmatrix} \lambda_2 \\ 1 \end{bmatrix} = \begin{bmatrix} \lambda_2^2 - \lambda_2 - 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

So, x_1 and x_2 are the eigenvectors

→ u_0 as a linear combination of x_1, x_2:

$$u_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} = c_1 \begin{bmatrix} (1+\sqrt{5})/2 \\ 1 \end{bmatrix} + c_2 \begin{bmatrix} (1-\sqrt{5})/2 \\ 1 \end{bmatrix}$$

$$c_1 = \frac{1}{\sqrt{5}}, \quad c_2 = -\frac{1}{\sqrt{5}}$$

$$u_k = c_1 \lambda_1^k x_1 + c_2 \lambda_2^k x_2$$

$$\begin{bmatrix} F_{k+1} \\ F_k \end{bmatrix} = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2}\right)^k \begin{bmatrix} (1+\sqrt{5})/2 \\ 1 \end{bmatrix} + \left(\frac{-1}{\sqrt{5}}\right)\left(\frac{1-\sqrt{5}}{2}\right)^k \begin{bmatrix} (1-\sqrt{5})/2 \\ 1 \end{bmatrix}$$

$$F_k = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2}\right)^k - \underbrace{\frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2}\right)^k}_{\text{becomes negligible as } k \text{ increases}}$$

$$F_k \approx \frac{1}{5} \left(\frac{1+\sqrt{5}}{2}\right)^k \qquad F_{100} \approx \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2}\right)^{100} \approx 3.54 \text{ e}20$$

↳ Orthogonally diagnolisable

→ If A is a real symmetric matrix, then:

① Eigenvalues of A are real.

② Eigenvectors corresponding to diff. eigenvalues are linearly independent.

③ A is orthogonally diagnolisable.

→ Matrix A is orthogonally diagnolisable if

$\exists\, Q$ satisfying $Q^T Q = I$ s.t. $A = Q \Lambda Q^T$

$$A = \begin{bmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} \begin{bmatrix} - & x_1^T & - \\ & \vdots & \\ - & x_n^T & - \end{bmatrix}$$

→ Remarks:

① A is diagnolisable + orthogonal matrix for diagnolisation ⇒ orthogonally diagnolisable

diagnolisable: $A = Q \Lambda Q^{-1}$; orthogonal matrix for diagnolisation: $Q^T Q = I$; orthogonally diagnolisable: $A = Q \Lambda Q^T$

② Real symmetric matrix is orthogonally diagnolisable

→ e.g. $A = \begin{bmatrix} 1 & -2 \\ -2 & -2 \end{bmatrix}$ $\det\left(\begin{bmatrix} 1-\lambda & -2 \\ -2 & -2-\lambda \end{bmatrix}\right) = 0$; $\lambda_1 = 2$, $\lambda_2 = -3$

$(A - 2I)x = 0$ $\begin{bmatrix} -1 & -2 \\ -2 & -4 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $x_1 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$ $x_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$(A + 3I)x = 0$ $\begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$q_1 = \frac{x_1}{\|x_1\|} = \frac{1}{\sqrt{5}}\begin{bmatrix} -2 \\ 1 \end{bmatrix}$, $q_2 = \frac{x_2}{\|x_2\|} = \frac{1}{\sqrt{5}}\begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$Q = \frac{1}{\sqrt{5}}\begin{bmatrix} -2 & 1 \\ 1 & 2 \end{bmatrix}$

WEEK 5

→ Complex Matrices

→ $\mathbb{C}^n$: complex counterpart of $\mathbb{R}^n$

$(x_1, x_2, \dots, x_n) \in \mathbb{C}^n$, then $x_i \rightarrow$ complex numbers

→ **complex conjugate** of $(a+ib)$ is $(a-ib)$

→ linear combinations : $c_1x_1 + c_2x_2 + \dots + c_nx_n = 0$ ($\in \mathbb{C}^n$)

(c_i: complex numbers)

→ **Inner product and length**

In $\mathbb{R}^n$, $\underbrace{\|x\|^2}_{\text{length}} = \underbrace{x^Tx}_{\text{inner product}}$

In $\mathbb{C}^n$, $x \cdot y = \bar{x}^T y = \bar{x}_1y_1 + \bar{x}_2y_2 + \dots + \bar{x}_ny_n$ (conjugate)

$\bar{x}^T \cdot y \neq \bar{y}^T \cdot x$

length of a complex vector : For $x \in \mathbb{C}^n$, define $\|x\|^2 = \bar{x}^T \cdot x$

→ ① $x \cdot y = \overline{y \cdot x}$

② $x \cdot (cy) = c(x \cdot y)$

③ $(cx) \cdot y = \bar{c}(x \cdot y)$

→ **Conjugate transpose**

A^* = conjugate transpose of A

$A^* = \bar{A}^T = \overline{A^T}$

→ Remarks

→ Real matrix A, $A^* = A^T$

→ $(AB)^* = B^* A^*$

→ $x \cdot y = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = x^* y$

→ **Hermitian Matrix**

↳ A is hermitian matrix if

$A^* = A$

↳ complex equivalent of symmetric matrix in $\mathbb{R}^n$

→ Hermitian Matrices

→ Matrix A is Hermitian if $A^* = A$

e.g. $A = \begin{bmatrix} 2 & 3-3i \\ 3+3i & 5 \end{bmatrix}$ $\quad A^* = \bar{A}^T = \begin{bmatrix} 2 & 3+3i \\ 3-3i & 5 \end{bmatrix}^T = \begin{bmatrix} 2 & 3-3i \\ 3+3i & 5 \end{bmatrix}$

→ Diagnol entries in Hermitian matrices have to be real

→ **Properties**:

① If A is Hermitian, then all eigenvalues are real.

e.g. $A = \begin{bmatrix} 2 & 3-3i \\ 3+3i & 5 \end{bmatrix}$ $\quad (A - \lambda I) = \begin{bmatrix} 2-\lambda & 3-3i \\ 3+3i & 5-\lambda \end{bmatrix}$

$$\begin{aligned} \det(A - \lambda I) &= (2-\lambda)(5-\lambda) - (3+3i)(3-3i) \\ &= (2-\lambda)(5-\lambda) - (9 - 9i^2) = (2-\lambda)(5-\lambda) - 18 \\ &= 10 - 2\lambda - 5\lambda + \lambda^2 - 18 = \boxed{(\lambda - 8)(\lambda + 1)} \end{aligned}$$

Proof: $Ax = \lambda x$; $(Ax)^* = (\lambda x)^* = \bar{\lambda} x^*$

$\Leftrightarrow x^* A^* = \bar{\lambda} x^* \Leftrightarrow x^* A^* x = \bar{\lambda} x^* x \Leftrightarrow x^* A x = \bar{\lambda} x^* x$

$\Leftrightarrow x^* \lambda x = \bar{\lambda} x^* x \Leftrightarrow \lambda x^* x = \bar{\lambda} x^* x \Leftrightarrow \lambda = \bar{\lambda}$

② If A is Hermitian, then eigenvectors corresponding to diff. eigenvalues are orthogonal i.e., If $Ax = \lambda_1 x$ and $Ay = \lambda_2 y$ where $\lambda_1 \neq \lambda_2$ then $x \cdot y = \bar{x}^T y = 0$

Proof: $Ax = \lambda_1 x$ and $Ay = \lambda_2 y$, $\lambda_1 \neq \lambda_2$

show: $x \cdot y = 0$

$\rightarrow x \cdot Ay = x \cdot \lambda_2 y = \lambda_2 (x \cdot y)$

$\rightarrow x \cdot Ay = x^* Ay = x^* A^* y = (Ax)^* y = (\lambda_1 x)^* y = \bar{\lambda}_1 (x \cdot y) = \lambda_1 (x \cdot y)$

$\rightarrow x \cdot Ay = \lambda_2 (x \cdot y) = \lambda_1 (x \cdot y)$

Since $\lambda_1 \neq \lambda_2$, we have $x \cdot y = x^* y = 0$

→ Remarks

① equivalent of Hermitian matrices in $\mathbb{R}^n$ → symmetric matrix

↳ All real symmetric matrices are Hermitian.

② If no eigenvalues are repeated, then A is diagonalisable.

→ Unitary Matrices

→ A matrix is unitary if it is a square matrix and has orthonormal columns.

→ In $\mathbb{R}^n$, $Q^T Q = I$, then Q is orthogonal matrix

↳ $\begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_n \\ | & | & & | \end{bmatrix}$ then $v_i^T v_j \ \forall i \neq j$ & $\|v_i\| = 1 \quad \forall i = 1, 2, \ldots, n$

→ In $\mathbb{C}^n$, $U^* U = I$, then U is unitary and $U^{-1} = U^*$

→ e.g. $U = \begin{bmatrix} 1/\sqrt{2} & i/\sqrt{2} \\ i/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$, $U = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$

→ Properties:

① length unchanged

↳ $\|Ux\| = \|x\|$

Proof:

$Ux \cdot Uy = (Ux)^* Uy = x^* U^* Uy = x^* y = x \cdot y$

② Eigenvalues of unitary matrix have absolute value 1

Proof:

$Ux = \lambda x \quad ; \|Ux\| = \|x\| \Rightarrow \|\lambda x\| = \|x\| \Rightarrow |\lambda| = 1$

③ Eigenvectors corresponding to diff. eigenvalues are orthogonal.

Proof:

$Ux = \lambda_1 x$, $Uy = \lambda_2 y$, $\lambda_1 \neq \lambda_2$

$x \cdot y = Ux \cdot Uy = (\lambda_1 x) \cdot (\lambda_2 y) = \bar{\lambda}_1 \lambda_2 (x \cdot y)$

$\Rightarrow (\bar{\lambda}_1 \lambda_2 - 1)(x \cdot y) = 0$

$\Rightarrow$ for $(x \cdot y) = 0$, λ_1 needs to be equal to λ_2.

→ Diagonalization of Hermitian Matrices

→ $A^* = \bar{A}^T$

Matrix A is Hermitian if $A^* = A$

Matrix U is Unitary if $U^*U = I$ and U is square matrix.

→ A is unitary diagonalizable $\exists$ matrix U s.t.

$A = U \Lambda U^*$, where Λ is a diagonal matrix

→ Any nxn matrix is similar to an upper triangular matrix

↳ $A = UTU^*$ (U: unitary, T: upper triangular)

→ **Shur's Theorem**

↳ $A = UTU^*$

↳ Proof for n=3:

Let $p(\lambda)$ be characteristic polynomial of A.

Let λ_1 be a root of $p(\lambda)$. Let z_1 be corresponding eigenvector.

Extend $\{z_1\}$ to a basis, and make it orthonormal (Gramm-Schmidt)

Let $\{z_1, u, v\}$ be the orthonormal basis.

↳ i.e., $\|z_1\| = \|u\| = \|v\| = 1$

$z_1^* u = z_1^* v = u^* v = 0$

Let $U_1 = \begin{bmatrix} | & | & | \\ z_1 & u & v \\ | & | & | \end{bmatrix}$ $\quad AU_1 = A\begin{bmatrix} | & | & | \\ z_1 & u & v \\ | & | & | \end{bmatrix} = \begin{bmatrix} | & | & | \\ \lambda_1 z_1 & Au & Av \\ | & | & | \end{bmatrix}$

$Az_1 = \lambda_1 z_1$

$U_1^* A U_1 = \begin{bmatrix} — \bar{z}^T — \\ — \bar{u}^T — \\ — \bar{v}^T — \end{bmatrix}\begin{bmatrix} | & | & | \\ \lambda_1 z_1 & Au & Av \\ | & | & | \end{bmatrix} = \begin{bmatrix} \lambda_1 & * & * \\ 0 & & \\ 0 & & B \end{bmatrix}$ (* : some non-zero entries; B: (2x2) matrix)

$\bar{z}_1^T \lambda_1 z_1 = \lambda_1 \bar{z}_1^T z_1 = \lambda_1$; $\bar{u}^T z_1 = 0$ and $\bar{v}^T z_1 = 0$

$U_1^* A U_1 = \begin{bmatrix} \lambda_1 & * & * \\ 0 & & \\ 0 & & B \end{bmatrix}$

→ Repeat the procedure with B to get eigenvalue λ_2 of B and a unitary matrix P s.t. $P^* B P = \begin{bmatrix} \lambda_2 & * \\ 0 & \lambda_3 \end{bmatrix}$

Let $U_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & & \\ 0 & & P \end{bmatrix}$. Then, $U^* = \begin{bmatrix} 1 & 0 & 0 \\ 0 & & \\ 0 & & P^* \end{bmatrix}$

$U_2^* U_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ since $P^*P = I$

$U_2^*(U_1^* A U_1) U_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & & \\ 0 & & P^* \end{bmatrix}\begin{bmatrix} \lambda_1 & * & * \\ 0 & & \\ 0 & & B \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & & \\ 0 & & P \end{bmatrix} = \begin{bmatrix} \lambda_1 & * & * \\ 0 & \lambda_2 & * \\ 0 & 0 & \lambda_3 \end{bmatrix}$

$= \begin{bmatrix} \lambda_1 & * & * \\ 0 & & \\ 0 & & P^* B P \end{bmatrix}$

→ e.g. $A = \begin{bmatrix} 5 & 8 & 16 \\ 5 & 0 & 9 \\ -3 & -5 & -10 \end{bmatrix}$ $\quad p(\lambda) = -(\lambda-1)(\lambda+3)^2$

$\lambda_1 = 1 \; ; \; \lambda_2 = -3$

Step ① eigenvector z_1 corresponding to $\lambda_1 = \begin{bmatrix} -2 \\ -1 \\ 1 \end{bmatrix}$

Step ② $\{z_1, e_1, e_2\} = \left\{ \begin{bmatrix} -2 \\ -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}$ ← basis for $\mathbb{R}^3$

Gram - Schmidt procedure:

$$w_1 = \frac{z_1}{\|z_1\|} = \frac{1}{\sqrt{6}} \begin{bmatrix} -2 \\ -1 \\ 1 \end{bmatrix}$$

$$v_2 = x_2 - \langle x_2, w_1 \rangle w_1 \; ; \; w_2 = \frac{v_2}{\|v_2\|}$$

$$w_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} - \left(\frac{-1}{3}\right) \begin{bmatrix} -2 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1/3 \\ -1/3 \\ 1/3 \end{bmatrix}$$

$$w_3 = \begin{bmatrix} 0 \\ 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

Step ③ $U_1 = \begin{bmatrix} -2/\sqrt{6} & 1/\sqrt{3} & 0 \\ -1/\sqrt{6} & -1/\sqrt{3} & 1/\sqrt{2} \\ 1/\sqrt{6} & 1/\sqrt{3} & 1/\sqrt{2} \end{bmatrix}$ $\quad U_1^* A U_1 = U_1^{\dagger} A U_1 = \begin{bmatrix} 1 & -8\sqrt{2} & -12\sqrt{3} \\ 0 & -3 & 0 \\ 0 & \sqrt{6} & -3 \end{bmatrix}$ → B

Step ④ Find an eigenvalue of B

$\lambda_2 = -3$, $e_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ eigenvector corresponding to λ_2

extend $\{e_2\} = \{e_2, e_1\} = \left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}$

unitary matrix $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

$$P^* B P = P^{\dagger} B P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} -3 & 0 \\ \sqrt{6} & -3 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} -3 & \sqrt{6} \\ 0 & -3 \end{bmatrix}$$

$$U_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & & \\ 0 & & P \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

$$U_2^* U_1^* A U_1 U_2 = \begin{bmatrix} 1 & -8\sqrt{2} & -12\sqrt{3} \\ 0 & -3 & \sqrt{6} \\ 0 & 0 & -3 \end{bmatrix}$$

→ **Spectral Theorem**

→ Hermitian matrix A is unitarily diagonalisable

i.e., $\exists$ unitary U s.t. $U^*AU = D$ → diagonal matrix with real numbers

→ Proof:

Schur's theorem: $U^*AU = T$

$T^* = U^*A^*U = U^*AU = T$ because $A^* = A$

$T \rightarrow$ upper Δ matrix and $T^* \rightarrow$ lower triangular matrix

$T^* = T \Rightarrow T$ is a diagonal matrix

↳ $T_{ii} = T^*_{ii} = \overline{T}_{ii} \Rightarrow T_{ii}$ is a real number

→ example.

$$A = \begin{bmatrix} 2 & 1-i \\ 1+i & 3 \end{bmatrix} \qquad A^* = A$$

$p(\lambda) = (\lambda - 1)(\lambda - 4)$ eigenvalues: $\lambda_1 = 1$, $\lambda_2 = 4$

eigenvectors: $z_1 = \begin{bmatrix} -1+i \\ 1 \end{bmatrix}$; $z_2 = \begin{bmatrix} 1-i \\ 2 \end{bmatrix}$

$$\frac{z_1}{\|z_1\|} = u_1 = \frac{1}{\sqrt{3}}\begin{bmatrix} -1+i \\ 1 \end{bmatrix} \quad ; \quad u_2 = \frac{1}{\sqrt{6}}\begin{bmatrix} 1-i \\ 2 \end{bmatrix}$$

$$U = \begin{bmatrix} \frac{-1+i}{\sqrt{3}} & \frac{1-i}{\sqrt{6}} \\ 1/\sqrt{3} & 2/\sqrt{6} \end{bmatrix} \quad ; \quad U^*AU = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$$

→ Remark. Hermitian $\Rightarrow$ unitarily diagonalisable but that does not mean only unitarily diagonalisable matrices are Hermitian.

WEEK 6

→ Singular Value Decomposition

→ Spectral theorem

↳ If A is a real symmetric matrix, then:

① All eigenvalues of A are real

② A is orthogonally diagonalisable

→ Every matrix cannot be diagonalised but any "real" $m \times n$ matrix A can be decomposed in SVD form i.e.,

A can be written as $A = Q_1 \Sigma Q_2^T$, where Q_1 and Q_2 are orthogonal

(A: $m \times n$, Q_1: $m \times m$, Q_2^T: $n \times n$)

$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}, \text{ where } D = \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{bmatrix}, \ \sigma_i > 0$$

→ Proof:

If A is $m \times n$, A^T is $n \times m$, A^TA is symmetric and real

⇒ there exists a basis of orthonormal eigenvectors $\{u_1, \dots, u_n\}$ corresponding to eigenvalues $\{\lambda_1, \dots, \lambda_n\}$

$$A^TA u_i = \lambda_i u_i \ , \ \forall \ i = 1, \dots, n$$

(u_i: orthonormal eigenvector)

$$(A^TA u_i) \cdot u_i = (\lambda_i u_i) \cdot u_i = \lambda_i \quad \text{since } \|u_i\|^2 = 1$$

Order the eigenvalues of A^TA: $\lambda_1, \lambda_2, \dots, \lambda_r, \lambda_{r+1}, \dots, \lambda_n$

$$\lambda_1 > 0, \dots, \lambda_r > 0 \ ; \ \lambda_{r+1} = \dots = \lambda_n = 0$$

σ_i's are singular values and $\forall \ i = 1, \dots r$, $\sigma_i = \sqrt{\lambda_i}$

Let $y_i = \frac{1}{\sigma_i} A u_i$, for $i = 1, \dots, r$, $y_i \in \mathbb{R}^m$

$$\|y_i\| = \frac{1}{\sigma_i} \|Au_i\| = \frac{\sqrt{\lambda_i}}{\sigma_i} = 1$$

(because $\|Au_i\|^2 = \lambda_i$)

$$y_i \cdot y_j = \frac{1}{\sigma_i \sigma_j} (Au_i) \cdot (Au_j) = \frac{1}{\sigma_i \sigma_j} u_i^T A^T A u_j = \frac{1}{\sigma_i \sigma_j} u_i^T \lambda_j u_j$$

$$= \frac{\lambda_j}{\sigma_i \sigma_j} u_i^T u_j = 0 \quad \text{since } \{u_1, \dots, u_n\} \text{ are orthonormal}$$

→ So we have a set $\{y_1, \dots, y_r\}$ of orthonormal vectors.

→ Since $r \leq m$, set $\{y_1, \dots, y_r\}$ is not a basis. So extend the set $\{y_1, \dots, y_r\}$ to form orthonormal basis of $\mathbb{R}^m$ giving the set $\{y_1, \dots, y_m\}$

$$Q_1 = \begin{bmatrix} | & & | \\ y_1 & \dots & y_m \\ | & & | \end{bmatrix} \quad \text{and} \quad Q_2 = \begin{bmatrix} | & & | \\ u_1 & \dots & u_n \\ | & & | \end{bmatrix}$$

$$\Sigma = Q_1^T A Q_2$$

$$= \begin{bmatrix} - & y_1^T & - \\ & \vdots & \\ - & y_m^T & - \end{bmatrix} \begin{bmatrix} | & & | \\ Ax_1 & \cdots & Ax_n \\ | & & | \end{bmatrix}$$

$\Sigma_{ij} = y_i^T (Ax_j)$, where $\sigma_i = \sqrt{\lambda_i}$ and x_i = a basis vector, λ_i = eigenvalue

For $j \le r$:

$$y_j = \frac{1}{\sigma_j} Ax_j$$

$$y_i^T (Ax_j) = \begin{cases} \sigma_j & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

For $j > r$:

$$\|Ax_j\|^2 = \lambda_j = 0$$

$$\Rightarrow y_i^T Ax_j = 0$$

$$\Sigma = \left[\begin{array}{ccc|c} \sigma_1 & & 0 & \\ & \ddots & & 0 \\ 0 & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right]$$

→ Remarks:

① $A = Q_1 \Sigma Q_2^T$

$AA^T = Q_1 \Sigma \Sigma^T Q_1^T$ } → eigen decomposition of AA^T

→ Q_1 contains eigenvectors of AA^T

② $A^T A = Q_2 \Sigma^T \Sigma Q_2^T$

→ Q_2 contains eigenvectors of AA^T

→ Procedure of SVD:

① Find eigenvalues and eigenvectors of $A^T A$. Using eigenvectors form an orthonormal basis.

② Use $\sigma_1 y_1 = Ax_1$, $\sigma_2 y_2 = Ax_2$, where $\{x_1, x_2\}$ is the basis from step ① , and $\sigma_i = \sqrt{\lambda_i}$. Find y_1 and y_2 .

③ $A = Q_1 \Sigma Q_2^T$ s.t.

$$Q_1 = \begin{bmatrix} | & & | \\ y_1 & \cdots & y_m \\ | & & | \end{bmatrix}, \quad Q_2 = \begin{bmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_n} \end{bmatrix}$$

→ example:

$$A = \begin{bmatrix} \sqrt{2} & 1 \\ 0 & \sqrt{2} \end{bmatrix}$$ Find SVD of A.

Is A diagonalisable? No, because $\lambda_1 = \lambda_2 = \sqrt{2}$

$$A - \sqrt{2}I = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \qquad \text{eigenvector} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

① $A^TA = \begin{bmatrix} 2 & \sqrt{2} \\ \sqrt{2} & 3 \end{bmatrix}$ $\lambda_1 = 4, \lambda_2 = 1$ $\sigma_1 = 2, \lambda_2 = 1$ $\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$

eigenvectors of A^TA: $A^TA - 4I = \begin{bmatrix} -2 & \sqrt{2} \\ \sqrt{2} & -1 \end{bmatrix}$ $z_1 = \begin{bmatrix} 1 \\ \sqrt{2} \end{bmatrix}$

$$A^TA - I = \begin{bmatrix} 1 & \sqrt{2} \\ \sqrt{2} & 2 \end{bmatrix} \qquad z_2 = \begin{bmatrix} \sqrt{2} \\ -1 \end{bmatrix}$$

normalise eigenvectors: $x_1 = \frac{1}{\sqrt{3}}\begin{bmatrix} 1 \\ \sqrt{2} \end{bmatrix}$ and $x_2 = \frac{1}{\sqrt{3}}\begin{bmatrix} \sqrt{2} \\ -1 \end{bmatrix}$

$$Q_2 = \frac{1}{\sqrt{3}}\begin{bmatrix} 1 & \sqrt{2} \\ \sqrt{2} & -1 \end{bmatrix}$$

② Use $\sigma_1 y_1 = Ax_1$ and $\sigma_2 y_2 = Ax_2$

to obtain $y_1 = \frac{1}{\sqrt{3}}\begin{bmatrix} \sqrt{2} \\ 1 \end{bmatrix}$ $y_2 = \frac{1}{\sqrt{3}}\begin{bmatrix} 1 \\ -\sqrt{2} \end{bmatrix}$

$$Q_1 = \frac{1}{\sqrt{3}}\begin{bmatrix} \sqrt{2} & 1 \\ 1 & -\sqrt{2} \end{bmatrix}$$

→ Positive Definiteness

→ A function that is that has first derivate 0 at (0,0) and is strictly positive at other points is called positive definite.

→ example:

$$f(x,y) = 2x^2 + 4xy + y^2$$

$$\frac{\partial f}{\partial x} = 4x + 4y \qquad \frac{\partial f}{\partial y} = 4x + 2y$$

(0,0) is a stationary point.

$\frac{\partial^2 f}{\partial x^2} = 4$ $\frac{\partial^2 f}{\partial y^2} = 2$ $\frac{\partial^2 f}{\partial x \partial y} = 4$ $\Rightarrow f$ has a minima at (0,0)

→ conditions for a function to be positive definite.

let $f(x,y)$ be of the form $ax^2 + 2bxy + cy^2$

① If $f > 0$, then $a > 0$

② If $f > 0$, then $c > 0$

③ If $f > 0$, then $ac > b^2$

→ Remarks:

① If $ac = b^2$, then $f(x,y) = ax^2 + 2bxy + cy^2$ is { positive semi-definite if $a > 0$ / negative semi-definite if $a < 0$

② We have saddle point at $(0,0)$ if $ac < b^2$

→ $ax^2 + 2bxy + cy^2 = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$

let $v = \begin{bmatrix} x \\ y \end{bmatrix}$ and $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$

Then, $ax^2 + 2bxy + cy^2 = v^T A v$

In general,

$$\begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

At $v = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$, $f(v) = 0$ i.e., $\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$ is a stationary point of f

→ Positive Definite Matrices

→ $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ is positive definite if $a > 0$ and $ac - b^2 > 0$

If $a > 0$ and $ac - b^2 > 0$, then both eigenvalues of A are > 0.

$ac - b^2 = \det(A) = \lambda_1 \lambda_2 > 0$, $\text{Trace}(A) = \lambda_1 + \lambda_2 = a + c > 0$

→ A real symmetric $n \times n$ matrix is positive definite if:

$v^T A v > 0 \quad \forall\, v \in \mathbb{R}^n, v \neq 0$

WEEK 7

→ Principal Component Analysis

→ Feature Selection : start with as many features as you can and then find the good subset of features.

→ Idea behind PCA → project data onto a lower dimensional subspace s.t.
① Reconstruction error is minimised.
② Variance of projected data is maximised.

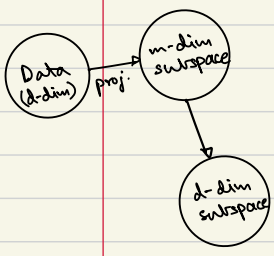
→ Problem:

Given : Dataset $D : \{x_1, \dots, x_n\}$, $x_i \in \mathbb{R}^d$

Goal : Project D onto a m-dimensional subspace.
↳ m : input parameter

PCA algorithm

① let $B = \{u_1, \dots, u_m\}$ orthonormal basis for m-dimensional subspace
↳ fix a subspace, find the best projection
↳ how to choose subspace optimally? Later...



② Extend B to an orthonormal basis for $\mathbb{R}^d$.
Extended basis $B' = \{u_1, \dots, u_m, u_{m+1}, \dots, u_d\}$

③ Any vector $x \in \mathbb{R}^d$ can be written using B' as:

$$x = \alpha_1 u_1 + \dots + \alpha_d u_d \text{ , where } \alpha_j = x^T u_j \quad \forall j = 1, \dots, d$$

Expressing each datapoint in $D = \{x_1, \dots, x_n\}$ using B':

$$x_i = \sum_{j=1}^{d} (x_i^T u_j) u_j \leftarrow \text{original datapoint}$$

Approximate x_i by $\tilde{x}_i$ as :

$$\tilde{x}_i = \sum_{j=1}^{m} z_{ij} u_j + \sum_{j=m+1}^{d} \beta_j u_j \leftarrow \text{projected datapoint (belongs to m-dim subspace)}$$

④ Find optimal z_{ij} and β_j to minimize square error:

$$J = \frac{1}{n} \sum_{i=1}^{n} \| x_i - \tilde{x}_i \|^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} \| \sum_{j=1}^{m} (x_i^T u_j - z_{ij}) u_j + \sum_{j=m+1}^{d} (x_i^T u_j - \beta_j) u_j \|^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{j=1}^{m} (x_i^T u_j - z_{ij})^2 + \sum_{j=m+1}^{d} (x_i^T u_j - \beta_j)^2 \right]$$

To minimise J, take partial derivative w.r.t. z_{ij} and β_j

$$\frac{\partial J}{\partial z_{ij}} = 0 \Rightarrow 2 (x_i^T u_j - z_{ij}) = 0 \Rightarrow z_{ij} = x_i^T u_j$$

$$\frac{\partial J}{\partial \beta_j} = 0 \Rightarrow \frac{1}{n} \sum_{i=1}^{n} (x_i^T u_j - \beta_j) = 0 \Rightarrow \beta_j = \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right)^T}_{\bar{x}^T \rightarrow \text{mean of data}} u_j$$

→ Conclusion:

$$\tilde{x}_i = \sum_{j=1}^{m} (x_i^T u_j) u_j + \sum_{j=m+1}^{d} (\bar{x}^T u_j) u_j$$

→ $x_i - \tilde{x}_i = \sum_{j=m+1}^{d} (x_i^T u_j - \bar{x}^T u_j) u_j$ ——— $\sum_{j=1}^{m} (x_i^T u_j) u_j$ vanishes

$\|x_i - \tilde{x}_i\|^2 = \sum_{j=m+1}^{d} (x_i^T u_j - \bar{x}^T u_j)^2 = \sum_{j=m+1}^{d} ((x_i - \bar{x})^T u_j)^2$, where $\bar{x}$ is the mean of data

With optimal z_{ij}, β_{ij}, the square error becomes:

$$J^* = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=m+1}^{d} ((x_i - \bar{x})^T u_j)^2$$

$$= \frac{1}{n} \sum_{j=m+1}^{d} \sum_{i=1}^{n} ((x_i - \bar{x})^T u_j)^T ((x_i - \bar{x})^T u_j)$$

$$= \frac{1}{n} \sum_{j=m+1}^{d} \sum_{i=1}^{n} u_j^T (x_i - \bar{x})(x_i - \bar{x})^T u_j$$

$$= \sum_{j=m+1}^{d} u_j^T \left[\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \right] u_j$$

$$J^* = \sum_{j=m+1}^{d} u_j^T C u_j \text{ , where } C = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$$

→ example:

Goal: want to choose $u_{m+1}, \dots, u_d$ s.t. J^* is minimised

Constrained optimisation: $\min_u u^T C u$ s.t. $u^T u = 1$

Lagrangian $L(u, \lambda) = u^T C u + \lambda(1 - u^T u)$

primal variable, Lagrange multiplier

$\nabla_u L(u, \lambda) = 0 \Rightarrow Cu = \lambda u \Leftrightarrow u^T C u = \lambda$

$\min_u u^T C u$ is the smallest eigenvalue λ of C.

$C = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$

→ real symmetric matrix and all eigenvalues are real and there exists an orthonormal basis of eigenvectors

- let the basis of eigenvectors of C be $\{u_1, \dots, u_m, u_{m+1}, \dots u_d\}$ corresponding to eigenvalues $\{\lambda_1, \dots, \lambda_m, \lambda_{m+1}, \dots, \lambda_d\}$ where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$

- To minimise $J^* = \sum_{j=m+1}^{d} u_j^T C u_j$:

choose $\{u_{m+1}, \dots, u_d\}$ to be $(d-m)$ eigenvectors corresponding to the $(d-m)$ least eigenvalues $\{\lambda_1, \dots, \lambda_d\}$

Remaining $\{u_1 \dots u_m\}$ are chosen to be top-m eigenvectors of C.

→ **Procedure of PCA:**

① Data: $\{x_1, \dots, x_n\}$ $x_i \in \mathbb{R}^d$ $\forall i$

② Let $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $C = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$

pick the largest m eigenvalues and their eigenvectors as a proxy for "max explainability"

③ Find eigenvalues $\{\lambda_1, \dots, \lambda_d\}$ $\lambda_1 \geq \dots \geq \lambda_d$ with corresponding eigenvectors $\{u_1, \dots, u_d\}$

Essentially, the variables with the least eigenvalues in the covariance matrix are being muted.

④ Projected data: $\tilde{x}_i = \sum_{j=1}^{m}(x_i^T u_j)u_j + \sum_{j=m+1}^{d}(\bar{x}^T u_j)u_j$

→ **example:**

dataset $D = \left\{ \begin{pmatrix}-1\\-1\end{pmatrix}, \begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1\\1\end{pmatrix} \right\}$ (x_1, x_2, x_3)

project D onto "m=1"-dimensional subspace

$$\bar{x} = \frac{1}{3}\sum_{i=1}^{3} x_i = \begin{bmatrix}0\\0\end{bmatrix}$$

$$C = \frac{1}{3}\sum_{i=1}^{3}(x_i - \bar{x})(x_i - \bar{x})^T = \frac{1}{3}\sum_{i=1}^{3} x_i \cdot x_i^T$$

$$= \frac{1}{3}\left[\begin{bmatrix}-1\\-1\end{bmatrix}\begin{bmatrix}-1 & -1\end{bmatrix} + 0 + \begin{bmatrix}1\\1\end{bmatrix}\begin{bmatrix}1 & 1\end{bmatrix}\right] = \frac{2}{3}\begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}$$

eigenvalues of C: $\lambda_1 = 4/3$, $\lambda_2 = 0$

eigenvectors of C: $u_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}$, $u_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}$

Best 1-dimensional space is spanned by $u_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}$

$$\tilde{x}_i = (x_i^T u_1)u_1 + (\bar{x}^T u_2)u_2 = (x_i^T u_1)u_1$$

$$\tilde{x}_1 = \frac{1}{2}\begin{bmatrix}-1 & -1\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}-1\\-1\end{bmatrix}$$

$$\tilde{x}_2 = \frac{1}{2}\begin{bmatrix}0 & 0\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix} = 0$$

$$\tilde{x}_3 = \frac{1}{2}\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}1\\1\end{bmatrix}$$

Reconstruction error: $J^* = \frac{1}{3}\sum_{i=1}^{3}\|x_i - \tilde{x}_i\|^2 = 0$

→ PCA as maximising variance

→ Projection of a datapoint x_i onto line along u is $(x_i^T u)u$

Dataset $D = \{x_1, \dots, x_n\}$ $x_i \in \mathbb{R}^d$, mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

- x_i's projection $= (x_i^T u)u$, mean projected value $= (\bar{x}^T u)u$
- variance $= (x_i^T u - \bar{x}u)^2$
- Projected variance over all points: $\frac{1}{n}\sum_{i=1}^{n}(x_i^T u - \bar{x}^T u)^2$

$$\frac{1}{n}\sum_{i=1}^{n}((x_i-\bar{x})^T u)^2 = \frac{1}{n}\sum_{i=1}^{n} u^T(x_i-\bar{x})(x_i-\bar{x})^T u = u^T C u$$

- Goal: $\max_u \; u^T C u \quad \text{s.t.} \quad u^T u = 1$

→ maximiser of $u^T C u$ is an eigenvector of C corresponding to the largest eigenvalue of C

→ $\max_u u^T C u$ s.t $u^T u = 1 \Leftrightarrow \max_u \frac{u^T C u}{u^T u}$

Let $u = (u^{(1)}, \dots, u^{(d)})$

$$\frac{\partial u^T u}{\partial u^{(i)}} = \frac{\partial}{\partial u^{(i)}}(u^{(1)^2} + \dots + u^{(d)^2}) = 2u^{(i)}$$

$$\frac{\partial u^T C u}{\partial u^{(i)}} = \frac{\partial}{\partial u^{(i)}}\left(\sum_{i=1}^{d}\sum_{j=1}^{d} C_{ij}\, u^{(i)} u^{(j)}\right) = 2\sum_j C_{ij} u^{(j)} = 2(Cu)^{(i)}$$

(C_{ij}: (i,j)-th entry of matrix C; $u^{(i)}$: i-th co-ordinate of u)

$$\frac{\partial}{\partial u^{(i)}}\left(\frac{u^T C u}{u^T u}\right) = u^T u\, 2(Cu)^{(i)} - (u^T C u)2u^{(i)} = 0 \quad \text{——— ①}$$

① in vector form: $u^T u Cu = (u^T Cu)u \Leftrightarrow Cu = \left(\frac{u^T Cu}{u^T u}\right)u$

$$\Rightarrow Cu = \lambda u$$

So, the maximimiser of $\frac{u^T Cu}{u^T u}$ is an eigenvector of C

and $\max_u \frac{u^T Cu}{u^T u} = \lambda_1$, where λ_1 is the largest eigenvalue

→ Can be extended to cases where $m > 1$.

→ Pick the eigenvectors corresponding to the largest eigenvalues to maximise the projected variance.

→ From a "maximising variance" viewpoint, PCA picks largest-m eigenvalues of C with corresponding $\{u_1, \dots, u_m\}$ (u_i → principal directions; projected values → principal components)

→ Revisiting the example:

$$D = \left\{ \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\} \quad ; \; m = 1 \quad ; \quad u_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\text{Projections} = \left\{ \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}$$

$$x_1^T u_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = -\sqrt{2}$$

$$x_2^T u_1 = 0$$

$$x_3^T u_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \sqrt{2}$$

$$\text{Projected variance} = \frac{1}{3} \left((-\sqrt{2})^2 + (\sqrt{2})^2 \right) = \frac{4}{3}$$

← is equal to the largest eigenvalue of C.

→ PCA in higher dimensions

→ Dataset $D = \{x_1, \ldots, x_n\}$ $x_i \in \mathbb{R}^d \; \forall \, i$

Feature dimension $d \gg$ number of datapoints n

↳ easier to handle $n \times n$ matrix than $d \times d$

PCA requires finding eigenvectors of $C = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$, $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

↳ $d \times d$ matrix

Goal: Reformulate the problem as finding eigenvectors of $n \times n$ matrix

$\text{Rank}(C) \leq n$

$\Rightarrow (d - n)$ eigenvalues are zero; it is not necessary to find $(d-n)$ eigenvectors

$$\text{Let } A = \begin{bmatrix} (x_1 - \bar{x})^T \\ \vdots \\ (x_n - \bar{x})^T \end{bmatrix} \quad \text{Then, } C = \frac{1}{n} A^T A$$

Let u_i be an eigenvector of C corresponding to eigenvalue $\lambda_i > 0$

Claim: λ_i is an eigenvalue of $\frac{1}{n} A^T A$

Proof:

$$\lambda_i (A u_i) = A(\lambda_i u_i)$$

$$= A \left(\frac{1}{n} A^T A u_i \right) \rightarrow \text{since } \lambda_i \text{ is an eigenvalue of } \frac{1}{n} A^T A$$

$$\lambda_i (A u_i) = \frac{1}{n} A A^T (A u_i)$$

i.e., $\left(\frac{1}{n} A A^T \right) (A u_i) = \lambda_i (A u_i) \Rightarrow \lambda_i$ is eigenvalue of $\frac{1}{n} A A^T$

→ It is enough to find eigenvectors of $\frac{1}{n} A A^T$, where $A = \begin{bmatrix} (x_1 - \bar{x})^T \\ \vdots \\ (x_n - \bar{x})^T \end{bmatrix}$

WEEK 8

→ General form of an optimisation problem

→ $\min_{x \in \mathbb{R}^d} f(x) \quad ; \quad g_i(x) \leq 0 \;\; \forall i = 1, \dots, k \quad ; \quad h_j(x) = 0 \;\; \forall j = 1, \dots, k$

- x: variable / parameter
- $f(x)$: objective function
- $g_i(x) \leq 0$: inequality constraints
- $h_j(x) = 0$: equality constraints

→ Unconstrained Optimisation

→ example $\min_x (x-5)^2$

$f(x) = (x-5)^2 \; ; \; f'(x) = 2(x-5)$

using: $x_{t+1} = x_t + d$, where $d = -f'(x)$ the solution will not converge to the answer. The direction is correct, magnitude is not.

update: $x_{t+1} = x_t + \eta_t d$, where $d = -f'(x)$ (η_t: step size)

good step size sequence ← $\eta_t = \frac{1}{t+1}$; $\eta_t = \frac{1}{2^t}$ → bad step-size sequence

→ **Gradient Descent Algorithm**

↳ Initialise at $x_0 \in \mathbb{R}$

for $t = 1, 2, \cdots$

$x_{t+1} = x_t - \eta_t f'(x_t)$ where $\eta_t = \frac{1}{t+1}$

end

↳ Properties:

→ If $\eta_t = 1/(t+1)$, the algorithm converges

→ Gradient descent converges to a "local minimum"

→ **Taylor Series**

↳ $f(x + \eta d) = f(x) + \eta d f'(x) + \frac{\eta^2 d^2}{2} f''(x) + \cdots$

↳ Local information gives global information

↳ for small enough η:

$f(x + \eta d) \approx f(x) + \eta d f'(x)$

$\underbrace{f(x+\eta d)}_{\text{function at new point}} - \underbrace{f(x)}_{\text{function at current point}} \approx \eta d f'(x)$

To minimise $f(x)$, choose direction d s.t. $f(x+\eta d) - f(x) < 0$

$\Rightarrow$ s.t. $\eta d f'(x) < 0 \Rightarrow$ want d s.t. $d f'(x) < 0$

(η: small +ve constant)

→ Gradient Descent for multivariate functions

↳ $f(x_1, x_2) = x_1^2 + 4x_2 + 8x_2^2$

$$\nabla f\left(\begin{bmatrix}1\\3\end{bmatrix}\right) = \begin{bmatrix}2x_1\\4+16x_2\end{bmatrix}_{x_1=1,\,x_2=3} = \begin{bmatrix}2\\52\end{bmatrix}$$

↳ $\vec{x}_{t+1} = \vec{x}_t + \eta(-\nabla f(x_t))$

→ Taylor Series for multivariable functions

↳ $f(x + \eta d) = f(x) + \eta d^T \nabla f(x) + \cdots$

↳ want d s.t. $f(x+\eta d) - f(x) < 0$

⇒ want d s.t. $d^T \nabla f(x) < 0$

if $d = -\nabla f(x)$, then

$d^T \nabla f(x) = -\|\nabla f(x)\|^2$ → norm is always positive

---

WEEK 9

→ Constrained Optimisation

→ $\min_x f(x)$ s.t. $g(x) \le 0$ ——— ①

Given x^*, it is claimed to solve ① (i.e., x^* is optimal). How to check the claim?

→ check constraint → $g(x^*) \le 0$

→ No "descent direction" should be a "feasible direction"

↑ any direction that reduces $f(x)$

↑ any direction that takes to a point that is feasible (satisfies $g(x) \le 0$)

→ If there is a direction from x^* which is a descent direction for both $f(x)$ and $g(x)$, i.e. $d^T \nabla f(x^*) < 0$ and $d^T \nabla g(x^*) < 0$, then x^* is not optimal because there is a direction that decreases $f(x)$ and also satisfies $g(x) \le 0$.

→ Necessary condition for x^* to be optimal: $\nabla f(x^*)$ and $\nabla g(x^*)$ should be antiparallel

↳ $\boxed{\nabla f(x^*) = -\lambda \nabla g(x^*)}$ where λ is a positive scalar

↑ Lagrange multiplier

→ Method of Lagrange Multiplier

→ $\min_x f(x) \quad g(x) = 0$

Method:

① $g(x^*) = 0$

② $\nabla f(x^*) = -\lambda \nabla g(x^*)$

→ example:

$f(x_1, x_2) = x_1^2 + 2x_2 + 4x_2^2$

$g(x_1, x_2) = x_1^2 + x_2^2 - 1$

$\nabla f(x_1, x_2) = (2x_1, 2 + 8x_2)$

$\nabla g(x_1, x_2) = (2x_1, 2x_2)$

Using method of lagrangian Multiplier:

$$\begin{bmatrix} 2x_1 \\ 2 + 8x_2 \end{bmatrix} = -\lambda \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} \Rightarrow \begin{matrix} 2x_1 = -\lambda 2x_1 \quad \text{——} ① \\ 2 + 8x_2 = -\lambda 2x_2 \quad \text{——} ② \end{matrix}$$

① $\Rightarrow 2x_1(1 + \lambda) = 0 \longrightarrow$ either $\lambda = -1$ or $x_1 = 0$

If $\lambda = -1$, using ②, $x_2 = -1/3$

$x_1 = \sqrt{1 - x_2^2} = +\sqrt{8}/3$ or $-\sqrt{8}/3$

$$\left\{ \begin{bmatrix} -\sqrt{8}/3 \\ -1/3 \end{bmatrix}, \begin{bmatrix} \sqrt{8}/3 \\ -1/3 \end{bmatrix} \right\}$$

If $x_1 = 0$, $x_2 = 1$ or -1

$$\left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\}$$

$f(0,1) = 6$, $f(0,-1) = 2$, $f(-\sqrt{8}/3, -1/3) = 2/3$, $f(\sqrt{8}/3, -1/3) = 2/3$

$f(0,1)$: maximizer; $f(-\sqrt{8}/3, -1/3)$ and $f(\sqrt{8}/3, -1/3)$: minimizer

→ Projected Gradient Descent

→ Gradient descent algorithm worked for unconstrained optimisation, but in the case of constrained optimisation, the algorithm might go outside the scope of constraints.

→ Solution: after every step, take a projection onto feasible set.

→ Convex Sets

→ A set $S \subseteq \mathbb{R}^d$ is a convex set if $\forall\, x_1, x_2 \in S$,

then $(\lambda x_1 + (1-\lambda) x_2) \in S \quad \forall\, \lambda \in [0,1]$

→ Properties:

① Intersection of convex sets is convex.

↳ System of linear equations is a convex set

② Convex combinations: Let $S = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^d$. $z \in \mathbb{R}^d$ is a convex combination of points in S if $\exists\, \lambda_1, \ldots, \lambda_n$ s.t. $\lambda_i \geq 0$, $\sum_{i=1}^n \lambda_i = 1$

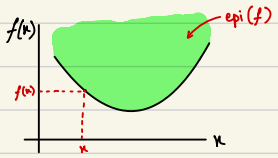
$z = \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n$.

Convex-Hull $(S) = \left\{ z : z = \sum_{i=1}^n \lambda_i x_i \text{ for some } \lambda_1, \ldots, \lambda_n \geq 0 \text{ s.t } \sum_{i=1}^n \lambda_i = 1 \right\}$

→ Convex Functions

→ $f: \mathbb{R}^d \to \mathbb{R}$ (domain: any convex set)

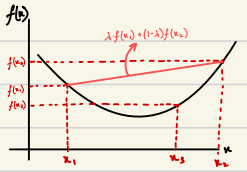
$$\underbrace{epi(f)}_{\subseteq \mathbb{R}^{d+1}} = \left\{ \begin{bmatrix} x \\ z \end{bmatrix} \in \mathbb{R}^{d+1} : z \geq f(x) \right\}$$



① **Definition:** A function $f: \mathbb{R}^d \to \mathbb{R}$ is a convex function if $epi(f)$ is a convex set

→ For any point x_3, $\lambda \cdot f(x_1) + (1-\lambda) f(x_2) \geq f(x_3)$

where $x_3 = \lambda x_1 + (1-\lambda) x_2$

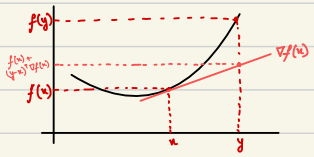


② **Definition:** A function $f: \mathbb{R}^d \to \mathbb{R}$ is a convex function iff $\forall x_1, x_2 \in \mathbb{R}^d$ and all $\lambda \in [0,1]$,

$$f(\lambda x_1 + (1-\lambda) x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

→ Assume $f: \mathbb{R}^d \to \mathbb{R}$ is differentiable

f is always greater than the linear approximation.



③ **Definition:** A function $f: \mathbb{R}^d \to \mathbb{R}$ is a convex function iff $f(y) \geq f(x) + (y-x)^T \nabla f(x)$

→ Assume $f: \mathbb{R}^d \to \mathbb{R}$ is twice differentiable

f alway has a non-negative curvature

④ **Definition:** A function $f: \mathbb{R}^d \to \mathbb{R}$ is a convex function if matrix H is a positive semi-definite matrix (→ all eigenvalues$(H) \geq 0$)

where $H \in \mathbb{R}^{d \times d}$, $H_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$.

WEEK 10

→ Properties of Convex functions

→ If f is a convex function, then all local minima of f are also global minima.

→ Necessary conditions for optimality of convex functions
($\hookrightarrow \min_x f(x)$)

$x^* \in \mathbb{R}^d$ is a global minimum of f iff $\nabla f(x^*) = 0$

Proof: if $\nabla f(x^*) \neq 0$, then f can be lowered by moving in direction $-\nabla f(x^*)$

→ If $\exists\, x^*$ s.t. $\nabla f(x^*) = 0 \Rightarrow x^*$ is a global minimum

Proof:

By definition of convexity:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \forall\, x, y$$

$$\Rightarrow f(y) \geq f(x^*) + \underbrace{\nabla f(x^*)^T (y - x^*)}_{=0} \quad \forall\, y$$

$$\Rightarrow f(y) \geq f(x^*) \quad \forall\, y \Rightarrow x^* \text{ is a global min}$$

→ Additional properties of Convex functions

① If $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ are both convex functions, then $h(x) := f(x) + g(x)$ is also a convex function.

→ Sum of convex functions is a convex function

Proof:

$$h(\lambda x + (1-\lambda) y) = f(\lambda x + (1-\lambda) y) + g(\lambda x + (1-\lambda) y) \qquad \lambda \in [0,1]$$

$$h(\lambda x + (1-\lambda) y) \leq (\lambda f(x) + (1-\lambda) f(y)) + (\lambda g(x) + (1-\lambda) g(y))$$

$$h(\lambda x + (1-\lambda) y) \leq \lambda (f(x) + g(x)) + (1-\lambda)(f(x) + g(y))$$

$$h(\lambda x + (1-\lambda) y) \leq \lambda h(x) + (1-\lambda) h(y) \longrightarrow \text{proves convexity of } h(x)$$

② Composition of functions

Let $f : \mathbb{R} \to \mathbb{R}$ be a convex non-decreasing function i.e., $f(x) \geq f(y)\ \forall\, x \geq y$

Let $g : \mathbb{R}^d \to \mathbb{R}$ be a convex function

Define $h := f \circ g \qquad h(x) = f(g(x)) \qquad h : \mathbb{R}^d \to \mathbb{R}$

→ Composition of convex with convex + non-decreasing is a convex function

③ Let $f : \mathbb{R} \to \mathbb{R}$ be convex

Let $g : \mathbb{R}^d \to \mathbb{R}$ be linear

Define $h := f \circ g$

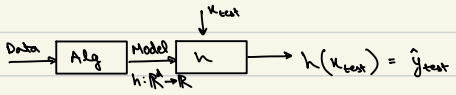
→ Composition of linear with convex is a convex function

→ Application of optimization in ML

→ Linear Regression : Dataset: $\{x_1, x_2, \dots, x_n\}$ (Training data), $\{y_1, y_2, \dots, y_n\}$ $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$ $\forall i$

goal: $h : \mathbb{R}^d \to \mathbb{R}$ $\quad \hat{y}_{test} = h(x_{test})$, $x_{test} \in \mathbb{R}^d$, $\hat{y}_{test} \in \mathbb{R}$

↳ given input during test time.



- In this case, h is a linear function
  $h(x) = w^T x$ for some $w \in \mathbb{R}^d$
- The "best" $h(x)$ from training data is characterised by
  good the w is Performance measure: Sum of Square Error
  $SSE = \sum_{i=1}^{n} (w^T x_i - y_i)^2$

Goal: $\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \frac{1}{2}(w^T x_i - y_i)^2$

$f(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2$. Is $f(w)$ convex?

$\sum_{i=1}^{n} h_i(w)$ sum of convex functions is convex

If $h_i(w)$ is convex $\forall i$, then $f(w)$ is convex

$h_i(w) = (w^T x_i - y_i)^2$

$= f(g(w))$ where $g(w) = w^T x_i - y_i$, $g : \mathbb{R}^d \to \mathbb{R}$ ← linear

$\Rightarrow$ convexity $f(z) = z^2$, $f : \mathbb{R} \to \mathbb{R}$ ← convex

Conclusion: SSE is a convex function

$f(w) = \frac{1}{2} \sum_{i=1}^{n} (w^T x_i - y_i)^2$

$= \frac{1}{2} \left\| \begin{bmatrix} (w^T x_1 - y_1) \\ \vdots \\ (w^T x_n - y_n) \end{bmatrix} \right\|_2^2$ ——— norm

$f(w) = \frac{1}{2} \|Xw - y\|_2^2$ where $X = \begin{bmatrix} — x_1 — \\ — x_2 — \\ \vdots \\ — x_n — \end{bmatrix}_{n \times d}$ $w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}_{d \times 1}$ $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1}$

$f(w) = \frac{1}{2}(Xw - y)^T (Xw - y)$

$= \frac{1}{2}(w^T X^T X w - 2 w^T X^T y + y^T y)$

$\nabla f(w) = \frac{1}{2}(2(X^T X)w - 2(X^T y)) = (X^T X)w - (X^T y)$

The global minimum at w^*: $(X^T X) w^* = (X^T y)$

$w^* = (X^T X)^{\dagger} (X^T y)$

Issue with $(X^T X)^{-1}(X^T y)$: needs an inverse computation $\sim O(d^3)$

Gradient descent can be done without inverse; iteratively

↳ $w_{t+1} = w_t - \eta_t \underbrace{\nabla f(w_t)}_{(X^T X)w - X^T y}$

↳ can be made faster by using

Stochastic Gradient Descent

↳ samples a small set of data points uniformly at random

↳ pretend sampled points form the new dataset

and compute gradient w.r.t. w

→ Constrained Optimisation

→ $\min_x f(x)$ s.t. $h(x) \leq 0$ $\qquad f: \mathbb{R}^n \to \mathbb{R}$

→ **Lagrangian function**

$L(x, \lambda) = f(x) + \lambda h(x)$

(x: vector, λ: scalar)

Fix $x \in \mathbb{R}^n$ $\quad \max_{\lambda \geq 0} L(x, \lambda) = \max_{\lambda \geq 0} f(x) + \lambda h(x)$

↳ if $h(x) \leq 0$, $L(x, \lambda)$ at max will be $f(x)$

if $h(x) > 0$, $L(x, \lambda)$ at max will be ∞

→ By maximising $L(x, \lambda)$, the solution is only possible if the condition of $h(x) \leq 0$ is met.

Therefore,

$$\boxed{\begin{aligned} &\min_x f(x) \\ &\text{s.t. } h(x) \leq 0 \end{aligned}} \equiv \min_x \left[\max_{\lambda \geq 0} L(x, \lambda) \right] \quad \text{where } L(x, \lambda) = f(x) + \lambda h(x)$$

→ $\min_x \left[\max_{\lambda \geq 0} L(x, \lambda) \right]$ ← Primal problem $\qquad x^* \to$ Primal solution

Dual Problem

$\max_{\lambda \geq 0} \left[\min_x f(x) + \lambda h(x) \right] \equiv \max_{\lambda \geq 0} g(\lambda)$, where $g(\lambda) = \min_x L(x, \lambda)$

Unconstrained Optimisation $\qquad$ Concave Function

→ **Relation b/w Primal and Dual problem**

Primal: $\min_x \max_{\lambda \geq 0} L(x, \lambda)$ $\qquad$ Dual: $\max_{\lambda \geq 0} \min_x L(x, \lambda)$

Define $J(x) = \begin{cases} f(x) & \text{if } h(x) \leq 0 \\ \infty & \text{otherwise} \end{cases}$

→ For any $\lambda \geq 0$, $L(x, \lambda) = f(x) + \lambda h(x)$

↳ $L(x, \lambda) \leq f(x)$ if $h(x) \leq 0$ and $\leq \infty$ otherwise

→ Fix $\lambda \geq 0$,

$L(x, \lambda) \leq J(x) \quad \forall x$

$\min_x L(x, \lambda) \leq \min_x J(x) = f(x^*)$ (x^*: Solution of the Primal Problem)

$$\max_{\lambda \geq 0} \min_x L(x, \lambda) \leq f(x^*)$$

→ The value at Dual optimum ≤ value at primal optimum ← Weak-Duality

→ If f, h are convex, then Strong-Duality holds

(f: objective, h: constraint)

→ $\min_x \left[\max_{\lambda \geq 0} L(x, \lambda) \right]$ (or) $\max_{\lambda \geq 0} \left[\min_x L(x, \lambda) \right]$ (easier to solve)

→ KKT conditions

→ Assume f, h are convex ⇒ Strong Duality holds
let x^*, λ^* are the primal and dual optimal solutions

$$x^* = \arg\min_x \left[\max_{\lambda \geq 0} f(x) + \lambda h(x) \right]$$

$$\lambda^* = \arg\max_{\lambda \geq 0} \left[\min_x f(x) + \lambda h(x) \right]$$

By strong duality,

$$f(x^*) = g(\lambda^*) \quad \Rightarrow \quad \nabla f(x^*) + \lambda^* \nabla h(x^*) = 0$$
$$= \min_x f(x) + \lambda^* h(x) \leftarrow \text{this is minimum over all possible } x$$

$$f(x^*) \leq f(x^*) + \underbrace{\lambda^* h(x^*)}_{\leq 0} \leq f(x^*)$$

$$\Rightarrow \lambda^* h(x^*) = 0$$

→ If f, h are convex ⇒ strong duality
⇒ x^*, λ^* must satisfy:

① $\nabla f(x^*) + \lambda^* \nabla h(x^*) = 0$ — stationarity condition
② $\lambda^* h(x^*) = 0$ — complementary slackness condition
③ $h(x^*) \leq 0$ — primal feasibility
④ $\lambda^* \geq 0$ — dual feasibility

→ In general, if (x^*, λ^*) satisfies the above conditions ⇒ x^* is a local optima

→ Optimisation problem with multiple constraints:

$$\min_x f(x) \quad \text{s.t.} \quad h_i(x) \leq 0 \ \forall i = 1, \dots, m \quad , \quad l_j(x) = 0 \ \forall j = 1, \dots, n$$

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{n} v_j l_j(x) \qquad \text{where } u = \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix} \quad v = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$$

(x: vector)

KKT conditions:

① $\nabla f(x^*) + \sum_{i=1}^{m} u_i^* \nabla h_i(x^*) + \sum_{j=1}^{n} v_j^* \nabla l_j(x^*) = 0$
② $u_i^* h_i(x^*) = 0 \quad \forall i$
③ $h_i(x^*) \leq 0 \ \forall i \ ; \ l_j(x^*) = 0 \ \forall j$
④ $u_i^* \geq 0 \quad \forall i$

→ Example ML algorithm: Support Vector Machine (SVM)

→ Optimisation problem

$$\min_w \ \tfrac{1}{2} \|w\|^2 \leftarrow \text{objective} \qquad \text{Dataset: } \{(x_1, y_1), \dots, (x_n, y_n)\}$$

→ Quadratic ⇒ convex . $\|w\|^2 = \sum_i w_i^2$

$$\text{s.t.} \quad w^T x_i y_i \geq 1 \quad \forall i \leftarrow \text{constraints}$$

→ Linear constraints ⇒ convex

→ Tutorial on Linear Programming

→ Example:

Primal Problem ⟷ **Dual Problem**

$$\min\ 50x_1 + 80x_2 \quad \text{s.t.} \begin{cases} 3x_1 + 0x_2 \geq 6 \\ 2x_1 + 4x_2 \geq 10 \\ 2x_1 + 5x_2 \geq 8 \\ x_1, x_2 \geq 0 \end{cases}$$

$$\max\ 6y_1 + 10y_2 + 8y_3 \quad \text{s.t.} \begin{cases} 3y_1 + 2y_2 + 2y_3 \leq 50 \\ 0y_1 + 4y_2 + 5y_3 \leq 80 \\ y_1, y_2, y_3 \geq 0 \end{cases}$$

WEEK 11

→ Continuous Random Variable

→ $X: \Omega \to \mathbb{R}$ probability function X maps the sample space Ω to $\mathbb{R}$

↳ Domain and Range are uncountable

→ PDF: $f_x(x) = \frac{P(X \in [x, x+dx])}{dx}$; $P(X = x) = 0$

→ CDF: $F_x(x) = P(X \leq x)$

→ Properties:

① $f_x(x) \geq 0$

② $\int_{-\infty}^{\infty} f_x(x)\,dx = 1$

③ $F_x(-\infty) = 0$

④ $F_x(\infty) = 1$

→ Conditional PDF

$$f_{x|A}(x) = \frac{P(X \in [x, x+dx] \mid A)}{dx} = \frac{P(x \in [x+dx])}{P(A)\,dx}$$

→ Function of random variables

example:

$f_x(x) = 1/2$ if $x \in [-1, 1]$

$Y = X/2$ $\quad f_y(y) = \frac{P(Y \in [y, y+dy])}{dy} = \frac{P(X \in [2y, 2y+2dy])}{dy}$

$= 0$ if $2y \notin [1, -1]$; $2 \cdot 1/2$ if $y \in [-1/2, 1/2]$

$X \sim$ Uniform $[-1, 1]$ $\quad Y = X/2 \in [-1/2, 1/2]$

$F_y(y) = P(Y \leq y) = P(X \leq 2y)$

$P(X \leq 2y) = \int_{-1}^{2y} 1/2\,dx = x/2 \Big|_{-1}^{2y} = y + 1/2$

$F_y(y) = y + 1/2$ $\quad f_y(y) = 1$

→ Expectation

$$X : \Omega \to \mathbb{R}$$

$$E[X] : \int_{-\infty}^{\infty} x \, f_X(x) \, dx$$

Properties of expectation:

① $E[X+Y] = E[X] + E[Y]$

② $Y = g(X)$; $E[Y] = \int_{-\infty}^{\infty} g(x) f_X(x) \, dx$

→ Variance

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

→ Conditional Expectation

$$E[X|A] = \int_{-\infty}^{\infty} x \cdot f_{X|A}(x) \, dx$$

→ Multiple Random Variables

→ Joint functions

$$f_{XY}(x,y) = \frac{P(X \in [x, x+dx], Y \in [y, y+dy])}{dx \, dy}$$

$$F_{XY}(x,y) = P(X \le x, Y \le y)$$

→ Marginal

$$f_X(x) = \int_{y=-\infty}^{\infty} f_{XY}(x,y) \, dy$$

→ Conditional

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}$$

WEEK 12

→ Bivariate and multivariate normal

→ Standard Normal Vector

$$Z_1 \sim N(0,1), \dots, Z_d \sim N(0,1) \qquad Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_d \end{bmatrix}$$

$$f_Z(z) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2} z_i^2\right) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\tfrac{1}{2}\|z\|^2\right)$$

→ Super Linear Transform of 2D-Normal

$X_1 = Z_1$; $X_2 = \rho \cdot Z_1 + \sqrt{1-\rho^2}\, Z_2$, where $\rho \in [-1, 1]$

$$X = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} Z \quad \text{, where } X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \text{ and } Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \text{ (standard normal)}$$

$$\hookrightarrow A \qquad A^{-1} = \begin{bmatrix} 1 & 0 \\ \frac{-\rho}{\sqrt{1-\rho^2}} & \frac{1}{\sqrt{1-\rho^2}} \end{bmatrix} \qquad X = AZ, \quad Z = A^{-1}X$$

$$\det(A) = \sqrt{1-\rho^2} \qquad \det(A^{-1}) = \frac{1}{\sqrt{1-\rho^2}}$$

$$E[X_1] = E[X_2] = 0$$

$$\mathrm{Cov}[X_1, X_2] = E[X_1 X_2] - E[X_1] \cdot E[X_2]$$

$$= E[X_1 \cdot X_2] = E[\rho Z_1^{2=1} + \sqrt{1-\rho^2}\, Z_1 Z_2^{=0}]$$

$$= \rho$$

$$\mathrm{Var}[X_1] = 1 \; ; \; \mathrm{Var}[X_2] = 1$$

$$\mathrm{Cov}[X] = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} = AA^T$$

$$\hookrightarrow \Sigma \qquad \det(\Sigma) = 1 - \rho^2$$

$$\Sigma^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}$$

$$f_X(x) = f_Z(A^{-1}x) \cdot |\det(A^{-1})|$$

$$= \frac{1}{\sqrt{1-\rho^2}} \cdot \frac{1}{2\pi} \exp\left(-\tfrac{1}{2}\|A^{-1}x\|^2\right)$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\tfrac{1}{2} x^T A^{-1T} A^{-1} x\right) \qquad A^{-1T}A^{-1} = \Sigma^{-1}$$

simplifies to

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(\frac{-1}{2} \cdot \frac{1}{(1-\rho^2)} [x_1, x_2] \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right)$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-1}{2} x_1^2\right) \cdot \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(\frac{-1}{2(1-\rho^2)} (x_2 - \rho x_1)^2\right)$$

$$f_X(x) = f_{X_1}(x_1) \cdot f_{X_2|X_1}(x_2 \mid x_1)$$

→ Bivariate Normal

$$X = AZ$$

$$f_X(x) = f_Z(A^{-1}x) \cdot |\det(A^{-1})|$$

$$\Sigma = AA^T = EXX^T = \mathrm{Cov}[X]$$

$$\det(A^{-1}) = \frac{1}{\det(A)} = \frac{1}{\sqrt{\det(\Sigma)}}$$

$$\boxed{f_X(x) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2} x^T \Sigma^{-1} x\right)}$$

Only positive definite matrices can be written as Σ

$$\Sigma = \begin{bmatrix} a^2 & \rho ab \\ \rho ab & b^2 \end{bmatrix} \quad \rho \in [-1, 1]$$

→ Multivariate Normal

$$X = AZ + \mu$$

$$\boxed{f_X(x) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)}$$

$$E[X] = \mu \qquad \mathrm{Cov}[X] = \Sigma$$

→ Properties of multivariate normal

$$X \sim N(\mu, \Sigma)$$

① $Y = a^T \cdot X \Rightarrow Y \sim N(a^T \cdot \mu, a^T \Sigma a)$

② $Y = B \cdot X \Rightarrow Y \sim N(B \cdot \mu, B \Sigma B^T)$

③ X_i, X_j are ind $\Leftrightarrow \mathrm{Cov}[X_i, X_j] = 0$

(⇐ only if they happen to be coordinates of multivariate random vector)

→ Estimation of parameters using ML

→ $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ → family of distributions parameterised by θ, which takes values from Θ space.

$X_1, X_2, \ldots, X_n$ drawn iid from P_θ for $\theta \in \Theta$ (→ use this data to estimate θ.)

→ $$L(\theta) = P(X_1 = x_1, \ldots, X_n = x_n \mid \theta)$$
$$= \prod_{i=1}^n f_{x_i}(x_i \mid \theta)$$
$$= \prod_{i=1}^n P_\theta(x_i) \qquad R(\theta) = -L(\theta)$$

$$\log(L(\theta)) = \sum_{i=1}^n \log(P_\theta(x_i)) \quad \leftarrow \text{maximising log likelihood} \Leftrightarrow \text{maximising likelihood}$$

→ example: $\mathcal{P} = \{\text{Bernoulli}(\theta) : \theta \in [0,1]\}$

$X_1, X_2, \ldots, X_n \in \{0, 1\}$

$$P_\theta(x) = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \end{cases} = (\theta)^x \cdot (1-\theta)^{1-x}$$

$$R(\theta) = -\sum_{i=1}^n \log(P_\theta(x_i)) = -\sum_{i=1}^n \log(\theta^{x_i} \cdot (1-\theta)^{1-x_i})$$
$$= -\sum_{i=1}^n x_i \log(\theta) + (1-x_i)\log(1-\theta) = a\log(1/\theta) + (n-a)\log(1/1-\theta) \qquad a = \sum_{i=1}^n x_i$$

$$\hat{\theta}_{ML} = \frac{a}{n}$$

→ Gaussian Mixture Models and Expectation Maximisation

→ $f_X(x) = \sum_{k=1}^{K} \pi_k N(x_k | \mu_k, \Sigma_k)$, where K = # of components

(μ_k: mean, Σ_k: variance)

π_k = what % of data is from k-th component

→ ML Estimation

Assume K is known.

Data = $\{x_1, x_2, \ldots, x_n\}$, $x_i \rightarrow d$-dimensional vector

$$P(\text{Data} | \pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K) = \prod_{n=1}^{N} \left(\sum_{k=1}^{K} \pi_k N(x_n | \mu_k, \Sigma_k) \right)$$

Goal: Find $\max_{\pi_k, \mu_k, \Sigma_k}$

$$R(\theta) = \sum_{n=1}^{N} -\log \left(\sum_{k=1}^{K} \pi_k N(x_n | \mu_k, \Sigma_k) \right)$$

↳ not easy to solve because of summation inside the log function

→ If the 'clustering' of data points were given, estimating parameters for each 'cluster' is straightforward. If the parameters of each cluster were given, assigning datapoints to clusters is also possible

→ Cluster Responsibilities

→ Start by fixing the parameters $\{\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K\}$ to some random numbers. This gives the joint density of the entire space.

→ Calculate $P(Z=k | X=x_n)$, i.e., for the n-th datapoint what is the probability that it belongs to the k-th component / cluster.

$$P(Z=k | X=x_n) = \frac{P(X=x_n | Z=k) P(Z=k)}{P(X=x_n)} = \frac{P(X=x_n | Z=k) P(Z=k)}{\sum_{j=1}^{K} P(X=x_n | Z=j) P(Z=j)}$$

$$\gamma(z_{nk}) = \frac{\pi_k N(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n | \mu_j, \Sigma_j)}$$ → Probability that n-th datapoint belongs to k-th cluster

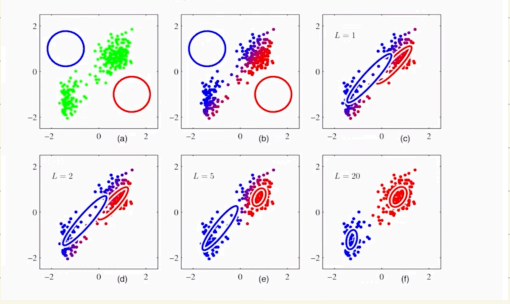
$$\mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_n}{N_k}$$, where $\gamma(z_{nk}) = \begin{cases} 1 & \text{if n-th datapoint is assigned to k-th cluster} \\ 0 & \text{otherwise} \end{cases}$

N_k → # of points assigned to k-th cluster

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (x_n - \mu_k)(x_n - \mu_k)^T$$

↳ These parameters are 'truer' than the random parameters. With this, you can iteratively come closer to an increasingly less random parameters.

Illustration



→ Markov Inequality

→ X is a positive random variable.

$$P(X \geq t) \leq \frac{E[X]}{t}$$

→ Chebyshev Inequality

→ $E[X] = \mu$; $\text{var}[X] = \sigma^2$

$$P(|X-\mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

→ Hoeffding Inequality

→ $X_1, X_2, \ldots, X_n$ are i.i.d.

$E[X_i] = \mu \qquad a \leq X_i \leq b$

$\bar{X}_n = 1/n \sum_i X_i \qquad E[\bar{X}] = n \qquad \text{Var}[\bar{X}_n] = \sigma^2/n$

$$P(|\bar{X}_n - \mu| \geq \delta) \leq \frac{\text{Var}(\bar{X}_n)}{\delta^2}$$

$$P(|\bar{X}_n - \mu| \geq \delta) \leq \frac{\sigma^2}{n\delta^2} \longrightarrow \text{Chebyshev's Inequality}$$

$$P(|\bar{X}_n - \mu| \geq \delta) \leq 2 \exp\left(\frac{-2nt^2}{(b-a)^2}\right)$$