Outline of this course:

1) Linear Algebra

operations, determents, solution of linear system, block metrics, eigen value, eigen vector, SVD, other decomposition. vector space. orthogonality principle.

2) probability theory

definition of probability space, random variables, expectations. conditional prob., conditional expectations, Decision Theory, estimation.

3) Optimization

Calculus of variation

4) Machine Learning algorithm

Schedule. Mon. 12:00 - 13:20 13:40 - 15:00

3	5	7	9	11
HW	HW	MT	HW	HW

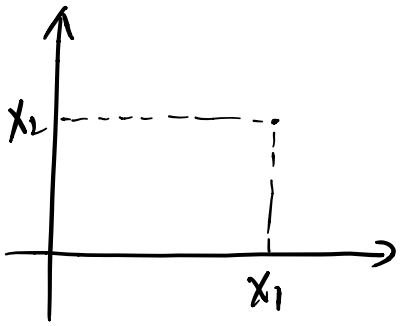
↑ no course

HW — 1 week

MT — 2 days

www.cs.rutgers.edu/~gm463/courses/CS596

LINEAR ALGEBRA.



$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ $\begin{pmatrix} x_2 \\ x_1 \end{pmatrix}$

order is important.

$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ Vector $\mathbb{R}^k$

sequence of real number — vector space of infinite dimensions

✡ $X = Y$

$$X + Y = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{bmatrix} \qquad a * X = \begin{bmatrix} ax_1 \\ ax_2 \\ \vdots \\ ax_n \end{bmatrix}$$

$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ $\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$

$ax = b$

1) $a \neq 0$ $\quad x = \frac{b}{a} = a^{-1}b$

2) $a = 0$, $b = 0$ $\quad$ any real #

3) $a = 0$, $b \neq 0$ $\quad$ no solutions

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1K}x_K = b_1$$
$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2K}x_K = b_2$$
$$\vdots$$
$$a_{K1}x_1 + a_{K2}x_2 + \cdots + a_{KK}x_K = b_K$$

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & & & \\ a_{K1} & a_{K2} & \cdots & a_{KK} \end{bmatrix} \leftarrow \text{matrics} \qquad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix}$$

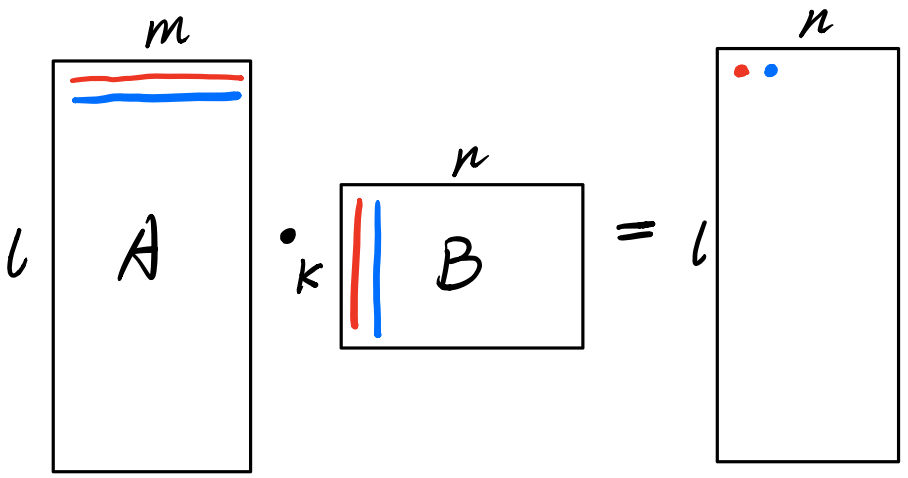
$$A \cdot K = b$$ Inner product.

$$A_{l \times m} \qquad B_{k \times n}$$

☆ $k = l$, $m = n$ $\Rightarrow$ $A + B = C$

$$A_{ij} + B_{ij} = C_{ij}$$

☆ $m = k$ $A \cdot B$



$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & \vdots & & \\ a_{K1} & a_{K2} & \cdots & a_{KK} \end{bmatrix} \qquad A^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{K1} \\ a_{12} & a_{22} & \cdots & a_{K2} \\ \vdots & \vdots & & \vdots \\ a_{1K} & a_{2K} & & a_{KK} \end{bmatrix}$$

$$(A_1 \cdot A_2)^T = A_2^T \cdot A_1^T$$

$$(A_1 \cdot A_2 \cdot \cdots A_k)^T = A_k^T \cdots A_2^T \cdot A_1^T$$

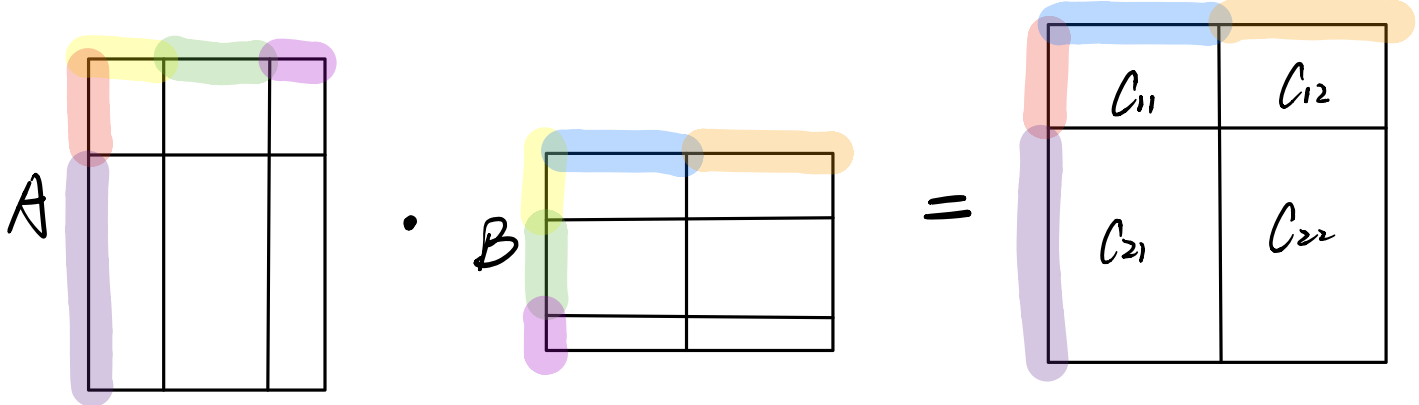
$$A = \begin{array}{|c|c|c|} \hline A_{11} & A_{12} & A_{13} \\ \hline A_{21} & A_{22} & A_{23} \\ \hline A_{31} & A_{32} & A_{33} \\ \hline A_{41} & A_{42} & A_{43} \\ \hline \end{array} \qquad B = \begin{array}{|c|c|} \hline B_{11} & B_{12} \\ \hline B_{21} & B_{22} \\ \hline B_{31} & B_{32} \\ \hline \end{array}$$

A and B have ~~the~~ same dimensions.

BUT. sub-matrics in A and B are different.

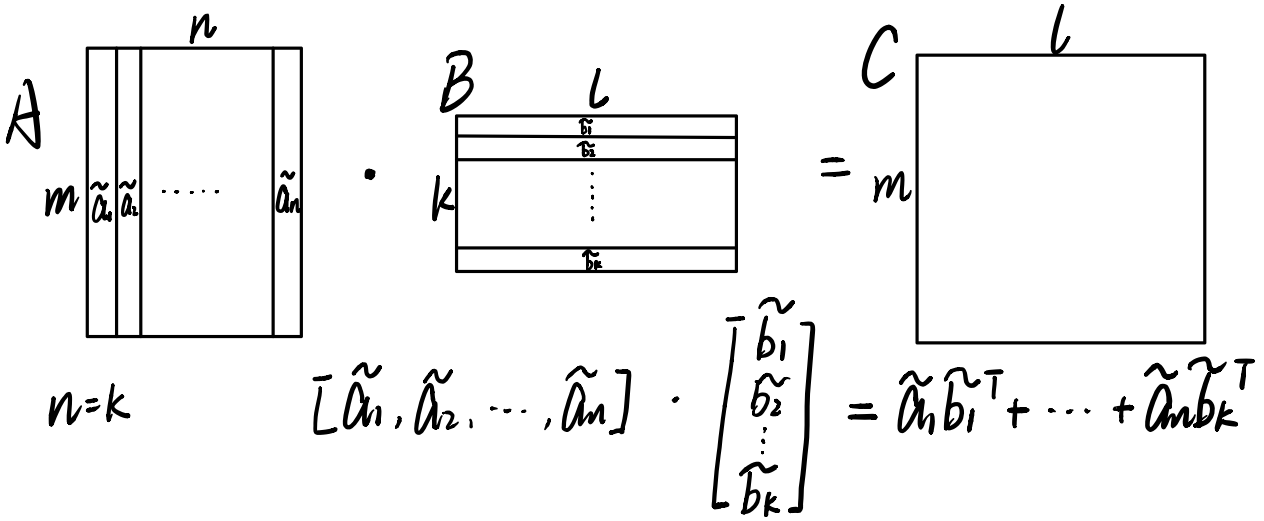
To make any operations between A and B, both A and B need to be sub-divided into same sub-matrics.

$$\begin{array}{|c|c|} \hline A_{11} & A_{12} \\ \hline A_{21} & A_{22} \\ \hline \end{array} + \begin{array}{|c|c|} \hline B_{11} & B_{12} \\ \hline B_{21} & B_{22} \\ \hline \end{array} = \begin{array}{|c|c|} \hline A_{11}+B_{11} & A_{12}+B_{12} \\ \hline A_{21}+B_{21} & A_{22}+B_{22} \\ \hline \end{array}$$



outer product

$$\begin{pmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_K^T \end{pmatrix} \cdot (b_1\ b_2\ \cdots\ b_n) = \begin{bmatrix} a_1^T b_1 & a_1^T b_2 & \cdots & a_1^T b_n \\ a_2^T b_1 & a_2^T b_2 & \cdots & a_2^T b_n \\ \vdots & \vdots & & \\ a_K^T b_1 & a_K^T b_2 & \cdots & a_K^T b_n \end{bmatrix}$$



$n=k$

$$[\tilde{a}_1, \tilde{a}_2, \cdots, \tilde{a}_n] \cdot \begin{bmatrix} \tilde{b}_1 \\ \tilde{b}_2 \\ \vdots \\ \tilde{b}_K \end{bmatrix} = \tilde{a}_1 \tilde{b}_1^T + \cdots + \tilde{a}_n \tilde{b}_K^T$$

$$A \cdot X = b$$

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1K} \\ \vdots & & \\ a_{K1} & \cdots & a_{KK} \end{bmatrix} \quad X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{bmatrix} \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix}$$

$$(A^T A)X = A^{-1} b$$

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

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ \vdots & & & \\ a_{i1} & a_{i2} & a_{ij} & a_{iK} \\ \vdots & & & \\ a_{K1} & a_{K2} & \cdots & a_{KK} \end{bmatrix}$$

$A - \mathrm{col}(a_{ij}) - \mathrm{row}(a_{ij}) = M_{ij}$

By this way, we can reduce the metric from x to x-1.

determine (A)

$|A|$

$$\det(A) = a_{i1} A_{ij} + \cdots + a_{ik} A_{ik}$$

$$A_{ij} = (-1)^{i+j} \det(M_{ij})$$

co-efficient.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{bmatrix} \Rightarrow A^T = \begin{bmatrix} A_{11} & A_{21} & \cdots & A_{k1} \\ A_{12} & A_{22} & \cdots & A_{k2} \\ \vdots & \vdots & & \vdots \\ A_{1k} & A_{2k} & \cdots & A_{kk} \end{bmatrix} \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}$$

$$A^{-1} = \frac{A^T}{\det(A)} \quad \text{for } x = A^{-1}b$$

$$x = \frac{A^T}{\det(A)} \cdot b$$

photo 3 cypher

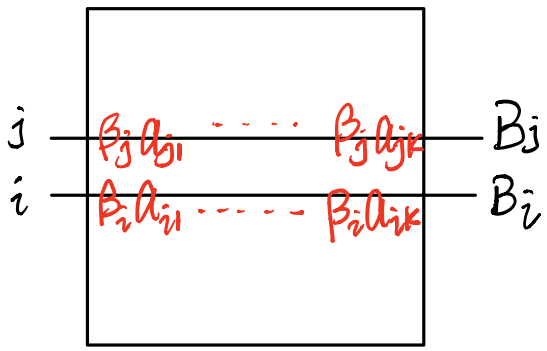
Gauss elimination.

photo 4

$Ax = b$ $\hat{x}$

$A\hat{x}$ b

having a solution does not mean use it.



$row(a_j) \cdot \beta_j$ add to $row(a_i) \cdot \beta_i$

$= \beta_i \det(A)$

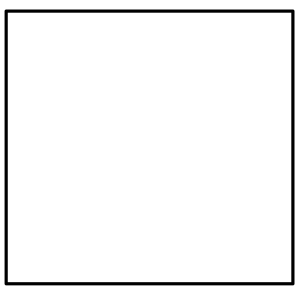
$$\det(A \cdot B) = \det(A) \cdot \det(B)$$

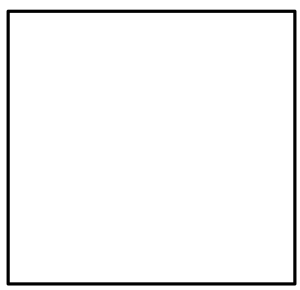
$$\det\begin{pmatrix} A_{11} & 0 & 0 & 0 \\ * & A_{22} & 0 & 0 \\ * & * & A_{33} & 0 \\ * & * & * & A_{44} \end{pmatrix} = \det(A_{11}) \cdots \det(A_{44})$$

Shur's Formula

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

$$\det(A) = \det(A_{11}) \cdot \det(A_{22} - A_{21} A_{11}^{-1} A_{12})$$
$$= \det(A_{22}) \cdot \det(A_{11} - A_{12} A_{22}^{-1} A_{21})$$

$A =$ 

$A^T =$ 

$A \cdot A^T = I$

↑

rotation matrics

orthonormal matrics

?

$\det(A^T) = \det(A)$

$\det(A^T \cdot A) = \det(I) = 1$

$\det(A^T) = \det(A) = 1$

symmatic

$$\begin{bmatrix} a_1^T a_1 & a_1^T a_2 & \cdots & a_1^T a_K \\ a_2^T a_1 & a_2^T a_2 & \cdots & a_2^T a_K \\ \vdots & & & \\ a_K^T a_1 & & & \end{bmatrix}$$

$$b^T a = a^T b$$

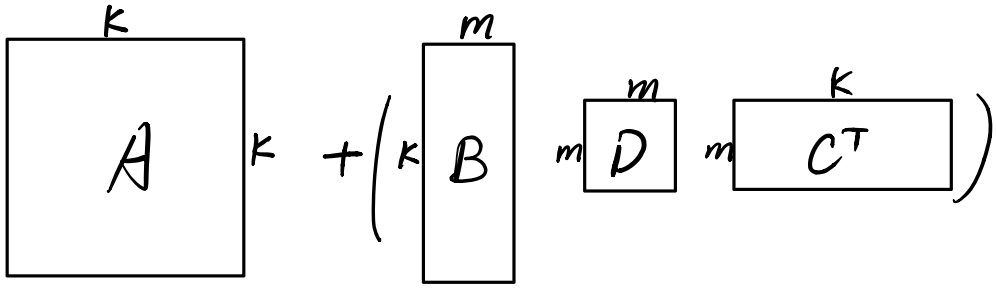
Degree of freedom. $K^2 - \frac{k(k+1)}{2}$ (only for rotation matics).

$$Ax = b$$
$$x = A^{-1} b$$

Matrix Inversim Lemma

$$A^{-1} - A^{-1}B(D^{-1} + C^T A^{-1} B)^{-1} C^T A^{-1}$$

$2k^2m + 2km^2 + 2m^3$



$$A = \begin{bmatrix} A_{11} & 0 & 0 \\ 0 & A_{22} & 0 \\ 0 & 0 & A_{33} \end{bmatrix} \qquad A^{-1} = \begin{bmatrix} A_{11}^{-1} & 0 & 0 \\ 0 & A_{22}^{-1} & 0 \\ 0 & 0 & A_{33}^{-1} \end{bmatrix}$$

Shur's Inversim Formula.

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} A_{11}^{-1} & \\ & \end{bmatrix} + \begin{bmatrix} -\bar{E} \\ I \end{bmatrix} \Delta^{-1} \begin{bmatrix} \bar{F} & I \end{bmatrix}$$

$$\Delta = A_{22} - A_{21} A_{11}^{-1} A_{12} \qquad \bar{E} = A_{11}^{-1} A_{12} \qquad \bar{F} = A_{21} A_{11}^{-1}$$

$k^2 m$ vs k^3

$$AX = 0$$

$$\begin{bmatrix} a_1 & a_2 & \cdots & a_k \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = 0$$

$$x_1 g_1 + x_2 g_2 + \cdots + x_k g_k = 0$$

r_C = longest subset of collum that are linearly independent.

r_C = Column rank r_r = row rank

$r_C = r_r$ = matrix rank

$$\overset{k}{X} \to \overset{m}{Y}$$

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$X_1 \to Y_1 \qquad X_1 + X_2 \to Y_1 + Y_2$

$X_2 \to Y_2 \qquad \alpha X_1 \to \alpha Y_1$

$Y = AX$ $(m \times k)$ linear mapping can represent by matrix

$m = k$ vector define direction in space

$X = \alpha_1 X_1 + \cdots + \alpha_k X_k$ decomposing x into k basis.

$Y = \beta_1 X_1 + \cdots + \beta_k X_k$

$$\beta_1 X_1 + \cdots + \beta_k X_k = A(\alpha_1 X_1 + \cdots + \alpha_k X_k)$$
$$= \alpha_1 \lambda_1 X_1 + \cdots + \alpha_k \lambda_k X_k$$

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & 0 \\ & & \ddots & \\ 0 & & & \lambda_k \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix}$$

$AX = \lambda X$

$BX = 0 \qquad \det(B) = 0$

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

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

$$\det\begin{pmatrix} \lambda-a_{11} & -a_{12} & \cdots & -a_{1k} \\ -a_{21} & \lambda-a_{22} & \cdots & -a_{2k} \\ & & \ddots & \\ 0 & & & \lambda-a_{kk} \end{pmatrix} = C_k\lambda^k + C_{k-1}\lambda^{k-1} + \cdots + C_1\lambda^1 + C_0$$

$(\lambda-\lambda_1)(\lambda-\lambda_2)\cdots(\lambda-\lambda_k)$

$C_0 = (-1)^k \lambda_1\lambda_2\cdots\lambda_k$

$\det(A) = \lambda_1\lambda_2\cdots\lambda_k$

CHARACTERISTIC POLYNOMIAL

$\det(\alpha B) = \alpha^k \det(B)$

$$\frac{1}{\lambda^k}\det(\lambda I - A) = C_k + \frac{C_{k-1}}{\lambda} + \cdots + \frac{C_0}{\lambda^k}$$

$$\det\left(I - \frac{A}{\lambda}\right) = C_k + \frac{C_{k-1}}{\lambda} + \cdots + \frac{C_0}{\lambda^k} \qquad C_k = \det(I) = 1$$

$C_0 = \det(-A) = (-1)^k \det(A)$

$(\lambda-\lambda_1)\cdots(\lambda-\lambda_k) = \lambda^k - (\lambda_1+\cdots+\lambda_k)\lambda^{k-1} + \cdots$

$C_{k-1} = -(\lambda_1+\cdots+\lambda_k)$

$B = TAT^{-1}$ similarity transformation

A & B are similar.

$$\begin{aligned} \det(\lambda I - B) &= \det(\lambda I - TAT^{-1}) \\ &= \det(\lambda TT^{-1} - TAT^{-1}) \\ &= \det(T(\lambda I - A)T^{-1}) \\ &= \det(T)\det(\lambda I - A)\det(T^{-1}) \\ &= \det(\lambda I - A) \end{aligned}$$

$\det(T)\det(T^{-1}) = 1$

$T \cdot T^{-1} = I$

$\lambda_i \lambda_i = A X_i = A T^{-1} T X_i$

$\lambda_i X_i = B Y_i$

$T A T^{-1} (T X_i) = \lambda_i (T X_i)$

$B \qquad\qquad Y_i$

$y^T A \stackrel{?}{=} \lambda y^T$

$A^T y = \lambda y$

$\det(\lambda I - A^T) = \det((\lambda I - A)^T)$

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

$\lambda_1, \lambda_2, \cdots, \lambda_k$

$Y_1 \; X_1 \; Y_2 \; X_2 \qquad Y_k \; X_k$

A $k \times k$ $\quad \lambda_1, \lambda_2, \cdots, \lambda_r$ eigenvalues

$r \leq k$

$X_1 \quad X_2 \qquad X_r$

linearly independent

$X_1, X_2, \cdots, X_m$

$X_{m+1}, X_{m+2}, \cdots, X_r$

$X_{m+1} = \alpha_1 X_1 + \cdots + \alpha_m X_m$

$A X_{m+1} = \alpha_1 A X_1 + \cdots + \alpha_m A X_m$

$\lambda_{m+1} X_{m+1} = \alpha_1 \lambda_1 X_1 + \cdots + \alpha_m \lambda_m X_m$

$0 = \alpha_1 (\lambda_{m+1} - \lambda_1) X_1 + \alpha_2 (\lambda_{m+1} - \lambda_2) X_2 + \cdots + \alpha_m (\lambda_{m+1} - \lambda_m) X_m$

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} k_1 \\ k_2 \end{pmatrix} = \begin{pmatrix} k_1 \\ k_2 \end{pmatrix} \qquad \begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 1 \\ 0 & 1+\epsilon \end{pmatrix}\begin{pmatrix} k_1 \\ k_2 \end{pmatrix} = \begin{pmatrix} k_1 \\ k_2 \end{pmatrix} \quad \begin{cases} k_1 + k_2 = k_1 \\ k_2 = k_2 \end{cases} \checkmark \quad \begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{pmatrix} 1+\epsilon \\ 0 \end{pmatrix}$$

if $\epsilon = 0$. only one eigen vacter.

$x_1, x_2, \cdots, x_k$ linearly independene.

$\lambda_1, \lambda_2, \cdots, \lambda_k$

$$A\bar{[}x_1 \; x_2 \cdots x_k] = \bar{[}Ax_1 \; Ax_2 \cdots Ax_k]$$

$$\bar{[}\lambda_1 x_1 \; \lambda_2 x_2 \cdots \lambda_k x_k]$$

$$\underbrace{\bar{[}x_1 \; x_2 \cdots x_k]}_{T} \underbrace{\begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & 0 \\ 0 & & \ddots \\ & & & \lambda_k \end{bmatrix}}_{\Lambda}$$

$AT = T \cdot \Lambda \Rightarrow A = T\Lambda T^{-1}$ diagonalizelism of A

$T^{-1}A = \Lambda T^{-1}$

$$\begin{bmatrix} x_1^T \\ & y_2^T \\ & & \ddots \end{bmatrix} \qquad A = \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & 0 \\ 0 & & \ddots \\ & & & \lambda_k \end{bmatrix} \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_k^T \end{bmatrix}$$

left & right eigenvecter. related.

$$Y_1^T A = \lambda_1 Y_1^T$$

$Y_i^T x_j = 0 \quad i \neq j$

$Y_i^T x_i = 1$

symmatric matrix.

$A^T = A$. $AX_i = \lambda_i X_i$

$(X_i^*)^T A X_i = \lambda_i (X_i^*)^T X_i$

$(X_i^*)^T X_i = \|X_i\|^2$ ← real number.

$(X_i)^T A X_i^* = \lambda_i^* \|X_i\|^2$

$(X_i^*)^T A^T X_i = \lambda_i^* \|X_i\|^2$

$\lambda_i \|X_i\|^2 = \lambda_i^* \|X_i\|^2$

X_i is eigenvector ⇓ $X_i \neq 0$

$\lambda_i = \lambda_i^*$ ☆ λ_i is real number.

$AX_i = \lambda_i X_i$

$X_i^T A = (A^T X_i)^T = (\lambda_i X_i)^T = \lambda_i X_i^T$

$A = P \cdot \Lambda \cdot P^T$ $P = [X_1\ X_2 \cdots X_k]$

A is symmatric

$\min_{X \neq 0} \frac{X^T A X}{X^T X}$ $\frac{X^T P \cdot \Lambda P^T X}{X^T P P^T X}$ $y = P^T \cdot X$

$= \frac{Y^T \cdot \Lambda Y}{Y^T Y}$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} \qquad \Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots \\ & & & \lambda_k \end{bmatrix}$$

$$\begin{bmatrix} y_1 & y_2 & \cdots & y_k \end{bmatrix} \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_k \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix}$$

$$\begin{pmatrix} \lambda_1 y_1 & \lambda_2 y_2 & \cdots & \lambda_k y_k \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{pmatrix} = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_k y_k^2$$

$$\lambda_{min} = \min \{\lambda_1, \lambda_2, \cdots \lambda_k\}$$

$$\frac{\lambda_1 y_1^2 + \cdots + \lambda_k y_k^2}{y_1^2 + \cdots + y_k^2} \geq \lambda_{min}$$

$$\lambda_1 y_1^2 \geq \lambda_{min} y_1^2$$

$$\frac{\lambda_{min}(y_1^2 + \cdots + x_k^2)}{(y_1^2 + \cdots + y_k^2)}$$

$A = A^T$

Positive definite.

$x \neq 0 \quad x^T A x > 0$

nonnegative definite

$x \neq 0 \quad x^T A x \geq 0$

λ_{min} of A

$$\frac{x^T A x}{x^T x} \geq \lambda_{min} \geq 0 \quad (\text{or } > 0)$$

$$x^T A x \geq \lambda_{min} \cdot x^T x$$

$x^T Q x + bx$

if $A \neq A^T$ A is square matrix

$X^T A X \gtrless 0$?

$(X^T A X)^T = X^T A^T X$ $\quad$ $X^T \left(\frac{A+A^T}{2}\right) X$

$A \gtrless B$? when $A - B \geq 0$

$Ax_i = \lambda_i x_i$

$A^2 x_i = A(Ax_i) = \lambda_i A x_i = \lambda_i^2 x_i$

$A^n x_i = \lambda_i^n x_i$

$A^{-n} x_i = \lambda_i^{-n} x_i$.

$$A = \begin{bmatrix} y_1^T \\ \vdots \\ y_k^T \end{bmatrix} \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \lambda_k \end{bmatrix} [x_1 \cdots x_k]$$

$$A = \lambda_1 x_1 y_1^T + \lambda_2 x_2 y_2^T + \cdots + \lambda_k x_k y_k^T$$

$$A^n = \lambda_1^n x_1 y_1^T + \cdots + \lambda_k^n x_k y_k^T$$

$$A^n = \lambda_1^n \left(x_1 y_1^T + \left(\frac{\lambda_2}{x_1}\right)^n x_2 y_2^T + \cdots \right)$$

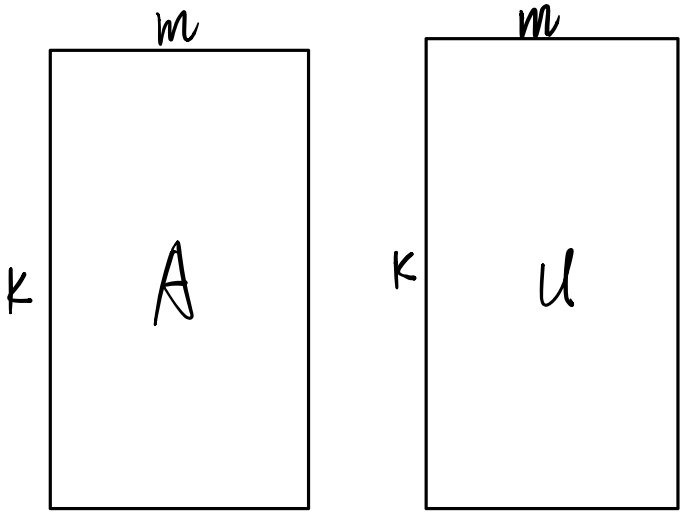
less than 1

while $n \uparrow$. $\left(\frac{\lambda_2}{\lambda_1}\right)^n \downarrow$

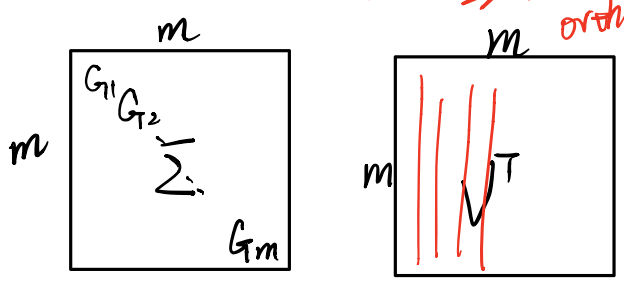
all parts dispear. same as following parts.

$$y_1 = Ax_0$$
$$y_2 = Ay_1$$
$$y_3 = Ay_2$$
$$\vdots$$
$$y_n = Ox_1$$

↑ some scaler.



col orthonormal. ⇒ row orthonormal



$$G_1 \geq G_2 \geq G_3 \geq \cdots \geq G_m \geq 0$$

$$U^T U = I$$
$$V^T V = I$$

$$A = U \Delta U^{-1}$$

↖ eigenvalue of A

A is symmetric, nonnegative definite.

$$V = U$$

$$SVD(A) = U \Sigma U^T$$

$$\Sigma = \begin{pmatrix} G_1 & & 0 \\ & G_2 & \\ 0 & & G_u \end{pmatrix} \qquad G_1 > G_2 > \cdots > G_u$$

$(u_1\ u_2\ \cdots\ u_u)$ ↑ orthonormal

$$(u_1\ u_2\ \cdots\ u_n)\cdot\begin{pmatrix}\sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n\end{pmatrix}\cdot\begin{pmatrix}u_1^T\\ u_2^T\\ \vdots\\ u_n^T\end{pmatrix}$$

$$\Downarrow$$

$$\sigma_1 u_1 u_1^T + \cdots + \sigma_n u_n u_n^T = A$$

$$A^n = \sigma_1^n u_1 u_1^T + \cdots + \sigma_n^n u_n u_n^T = \sigma_1^n\left(u_1 u_1^T + \left(\frac{\sigma_2}{\sigma_1}\right)^n u_2 u_2^T + \cdots + \left(\frac{\sigma_n}{\sigma_1}\right)^n u_n u_n^T\right)$$

$$\lim_{n\to\infty} A^n = \sigma_1^n u_1 u_1^T$$

Assume $x_0^T u_1 \neq 0$

$$z_1 = Ax_0$$

$$x_1 = \frac{z_1}{\|z_1\|}$$

$$z_2 = Ax_1$$

$$x_2 = \frac{z_2}{\|z_2\|}$$

$$\vdots$$

$x_n \to ?$

$$x_n = \frac{A^n x_0}{\|A^n x_0\|} = \frac{\sigma_1^n\left(u_1(u_1^T x_0) + \left(\frac{\sigma_2}{\sigma_1}\right)^n \cdot O(1)\right)}{\sigma_1^n\left(|u_1^T x_0| + \left(\frac{\sigma_2}{\sigma_1}\right)^n O(1)\right)}$$

$$x_n \to \frac{u_1\cdot(u_1^T x_0)}{|u_1^T x_0|} = u_1 \cdot \operatorname{sgn}(u_1^T x_0)$$

$$\downarrow$$

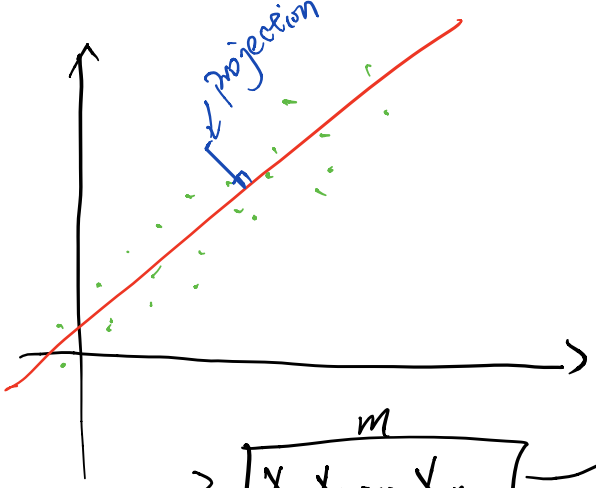
Power method.

$x_n \sim \pm u_1$

$z_{n+1} = Ax_n \approx \sigma_1(\pm u_1)$ $\quad \|z_{n+1}\| = \sigma_1\|u_1\|$

$\hookrightarrow 1$

Subspace tracking



m

$\geq$ $X_1\ X_2 \cdots X_n$

x_n — 2d-vector

$U_1 U_2$ | $G_1\ 0$ / $0\ G_2$ | V_1^T / V_2^T

$G_1 >> G_2$

$= G_1 U_1 U_1^T + G_2 U_2 U_2^T$

$U_1 U_2$ | $G_1\ 0$ / $0\ 0$ | V_1^T / V_2^T

$= G_1 U_1 U_1^T$

PCA

lose some information and keep most data. (feature)

Matrix filling

k × n : missing pairs, A, rank = r, $r < k$

k × r : $\tilde{U}$

r × n : $\tilde{V}^T$

SVD

k × r : U

r × r : $G_1, G_2, \Sigma, G_r, G_k$

r × n : V^T

see how much diff

k × r : U

r × r : $G_1, G_2, \Sigma, G_r, G_k$

r × n : V^T

PCA ?

Vector space. $\oplus$ $\odot$

V a set of vectors.

x, y — vector $\quad x \oplus y \in V$

α, β — real num.

1) $x \oplus (y \oplus z) = (x \oplus y) \oplus z$ associativity

2) $x \oplus y = y \oplus x$ commutativity

3) $x \oplus 0 = x$ Identity

4) $(-x) \oplus x = 0$

↓ not negative x, it's a notation.

if V satisfy all 9 properties then V is a vector space.

$x \in V$, $\alpha \in \mathbb{R}$

1) $\alpha \odot x \in V$

2) $\alpha \odot (\beta \odot x) = (\alpha \cdot \beta) \odot x$

3) $1 \odot x = x$

4) $\alpha \odot (x \oplus y) = (\alpha \odot x) \oplus (\alpha \odot y)$

5) $(\alpha + \beta) \odot x = (\alpha \odot x) \oplus (\beta \odot x)$

HILBERT SPACE

V

$x, y \in V$

$\langle x, y \rangle \in \mathbb{R}$

inner product is mapping from vector to $\mathbb{R}$.

$\langle x, y \rangle = \langle y, x \rangle$ symmetric

$\langle \alpha \odot x, y \rangle = \alpha \cdot \langle x, y \rangle$

$\langle x + z, y \rangle = \langle x, y \rangle + \langle z, y \rangle$

$\langle x, x \rangle \geq 0$ if $\langle x, x \rangle = 0 \Leftrightarrow x = 0$

$\mathbb{R}^K \qquad x = \begin{bmatrix} x_1 \\ \vdots \\ x_K \end{bmatrix}$

$$\langle x, y \rangle = x^T y = x_1 y_1 + \cdots + x_K y_K$$

$$\langle x, y \rangle = w_1 x_1 y_1 + \cdots + w_K x_K y_K$$

$$w_1, w_2, \cdots, w_K > 0$$

$$x^T \begin{pmatrix} w_1 & & & \\ & w_2 & & \\ & & \ddots & \\ & & & w_K \end{pmatrix} y$$

$x^T \Omega y$. Ω : symmetric & positive definite.

$$x^T \Omega y \geq 0$$

$\phi(x)$ $\qquad \int \phi(x) \psi(x) dx$

$\psi(x)$ $\qquad \int w(x) \phi(x) \psi(x) dx \quad (w(x) > 0)$

$$\iint \underbrace{\Omega(x,y)}_{\text{kernels}} \phi(x) \psi(x) dx dy$$

$X, Y \in V \rightarrow$ random variables.

$\langle x, y \rangle = \mathbb{E}[x, y]$ ← Expectation.

V . $\langle x, y \rangle$, $\|x\| = \sqrt{\langle x, x \rangle}$

def : $\cos\phi = \dfrac{\langle x, y \rangle}{\|x\| \cdot \|y\|}$

Cauchy – Schwarz inequality

$(\langle x, y \rangle)^2 \le \|x\|^2 \cdot \|y\|^2$ ← only and only if $y = r \cdot x$ (r: number)

$\langle \alpha x + y , \alpha x + y \rangle \ge 0$

$= \alpha^2 \langle x, x \rangle + 2\alpha \langle x, y \rangle + \langle y, y \rangle \ge 0$ $\qquad \alpha = -\dfrac{\langle x, y \rangle}{\langle x, x \rangle}$

$\langle y, y \rangle - \dfrac{(\langle x, y \rangle)^2}{\langle x, x \rangle} \ge 0 \rightarrow \langle y, y \rangle \cdot \langle x, x \rangle \ge (\langle x, y \rangle)^2$

$\|y\|^2 \cdot \|x\|^2 \ge (\langle x, y \rangle)^2$

$\mathbb{R}^k$ $\quad x^T y$

$(x_1 y_1 + \cdots + x_k y_k)^2 \le (x_1^2 + \cdots + x_k^2)(y_1^2 + \cdots + y_k^2)$

$\left(\int \phi(t) \psi(t) dt\right)^2 \le \left(\int \phi^2(t) dt\right)\left(\int \psi^2(t) dt\right)$

$(|E[x, y]|)^2 \le E[x^2] \cdot E[y^2]$

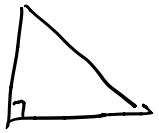
$y = rx$.

x, y $\quad \langle x, y \rangle = 0$

$\cos\phi = \dfrac{\langle x, y \rangle}{\|x\| \cdot \|y\|}$ $\qquad \|x - y\|^2 = \langle x - y, x - y \rangle$

$= \langle x, x \rangle + \langle y, y \rangle - 2\langle x, y \rangle$

$\|x - y\|^2 = \|x\|^2 + \|y\|^2$

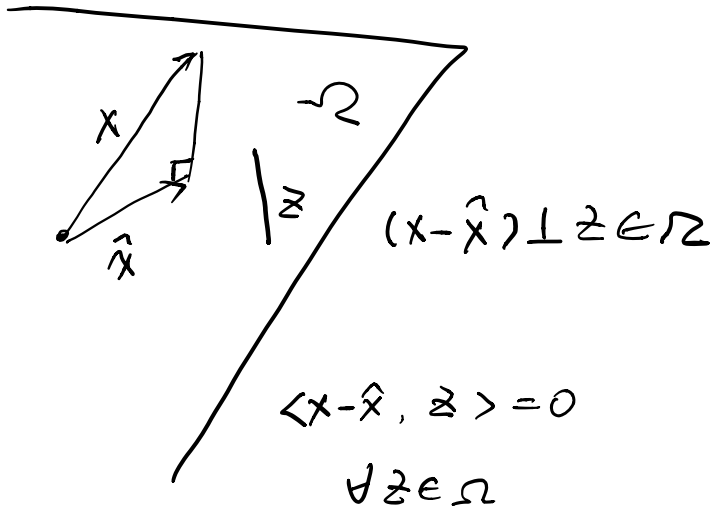


Pythagorean theorem.

Orthogonality principle

$\forall$. $\langle x, y \rangle$

Ω $x, y \in \Omega$. $\alpha x + \beta y \in \Omega$



$$\|x-y\|^2 = \langle x-y, x-y \rangle = \langle x-\hat{x}+\hat{x}-y,\ x-\hat{x}+\hat{x}-y \rangle$$

$$\langle x-\hat{x}, x-\hat{x} \rangle + \langle \hat{x}-y, \hat{x}-y \rangle + 2\langle x-\hat{x}, \hat{x}-y \rangle$$

$$\|x-\hat{x}\|^2 + \|\hat{x}-y\|^2 \geq \|x-\hat{x}\|^2$$

$$\Omega = \{\alpha_1 z_1 + \cdots + \alpha_k z_k\} \quad z_1, z_2, \cdots, z_k$$

$$\hat{x} = \alpha_1 z_1 + \cdots + \alpha_k z_k$$

$$\langle x - \alpha_1 z_1 - \cdots - \alpha_k z_k,\ z_1 \rangle = 0$$

z_2
$\vdots$
z_k

$$\begin{bmatrix} \langle x, z_1 \rangle \\ \langle x, z_2 \rangle \\ \vdots \\ \langle x, z_k \rangle \end{bmatrix} = \begin{bmatrix} \langle z_1, z_1 \rangle & \langle z_1, z_2 \rangle & \cdots \\ \langle z_2, z_1 \rangle & \langle z_2, z_2 \rangle & \cdots \\ \vdots & & \\ & & \langle z_k, z_k \rangle \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix}$$

solve this equation ↑

Application: Kalman filter

Total probability

$A_1, A_2, \cdots, A_k \quad A_i \cap A_j = \emptyset \ (i \neq j)$

$A_1 \cup A_2 \cup \cdots \cup A_k = (H)$

event B. $\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(B|A_i)\,\mathbb{P}(A_i)$ $\quad B = \{x < m < x + dx\}$

density probability

random events $H_1, H_2 \cdots H_k$

H_1	$f_1(x)$	$\mathbb{P}(H_1)$
H_2	$f_2(x)$	$\mathbb{P}(H_2)$
$\vdots$		
H_k	$f_k(x)$	$\mathbb{P}(H_k)$

$\Sigma = 1$

$$f(x)dx = \sum_{i=1}^{k} \mathbb{P}(x < m \le x + dx \mid H_i)\,\mathbb{P}(H_i)$$

$$= \sum_{i=1}^{k} f_i(x)\,\mathbb{P}(H_i)$$

$$f(x) = f_1(x|A_1)\,\mathbb{P}(H_1) + f_2(x|A_2)\,\mathbb{P}(H_2) + \cdots + f_k(x|A_k)\,\mathbb{P}(H_k)$$

Posteenor Probability

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)\,\mathbb{P}(A_j)}{\sum_{i=1}^{k} \mathbb{P}(B|A_i)\,\mathbb{P}(A_i)}$$

$$\mathbb{P}(H_j \mid m = x) = \frac{f_j(x)\,\mathbb{P}(H_j)}{\sum_{i=1}^{k} f_i(x)\,\mathbb{P}(H_i)}$$

(0) non of these guys 10%

(1) George 50%

(2) Jim 20%

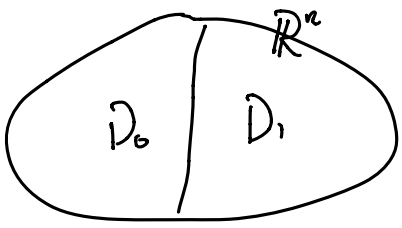
(3) John 20%

$H_0 : f_0(x), \mathbb{P}(H_0)$

$H_1 : f_1(x), \mathbb{P}(H_1)$

find a decision strategy.

$X = [0.1, 0.33, \cdots]$ D_0. $X \to D_0$ or D_1



Randomized

$\delta_0(x)$, $\delta_1(x)$

$\delta_0(x)$ ↙ H_0 ; $\delta_1(x)$ ↘ H_1

$$\delta_0(x) = \begin{cases} 1 & x \in H_0 \\ 0 & x \in H_1 \end{cases} \qquad \delta_1(x) = \begin{cases} 1 & x \in H_1 \\ 0 & x \in H_0 \end{cases}$$

Bayesian Approach

$H_0 = f_0(x)$, $\mathbb{P}(H_0)$ $\{D_0, H_0\}$ C_{00}

$H_1 = f_1(x)$, $\mathbb{P}(H_1)$ $\{D_0, H_1\}$ C_{01}

$\{D_1, H_0\}$ C_{10}

$\{D_1, H_1\}$ C_{11}

$$e(\delta_0, \delta_1) = C_{00}\mathbb{P}(D_0, H_0) + C_{01}\mathbb{P}(D_0, H_1) + C_{10}\mathbb{P}(D_1, H_0) + C_{11}\mathbb{P}(D_1, H_1)$$

for a specific data

optimal ≠ best optimal is <u>average</u> minimal cost.

$$\mathbb{P}(D_i \& H_j) = \mathbb{P}(D_i \mid H_j) \cdot \mathbb{P}(H_j) \qquad \mathbb{P}(D_i \mid H_j) = \int \delta_i(x) f_j(x) dx$$

$$e(\delta_0, \delta_1) = \sum_{i=0}^{1} \sum_{j=0}^{1} C_{ij} \mathbb{P}(H_j) \int \delta_i(x) f_j(x) dx$$

$$= \int \Big(\delta_0(x) \underbrace{\{ C_{00} \mathbb{P}(H_0) f_0(x) + C_{01} \mathbb{P}(H_1) f_1(x) \}}_{C_0(x)} + \delta_1(x) \underbrace{\{ C_{10} \mathbb{P}(H_0) f_0(x) + C_{11} \mathbb{P}(H_1) f_1(x) \}}_{C_1(x)} \Big) dx$$

$$= \int (\delta_0(x) C_0(x) + \delta_1(x) C_1(x)) dx$$

$C_1 > C_0$, $\delta_1 = 0, \delta_0 = 1$

$C_1 < C_0$, $\delta_1 = 1, \delta_0 = 0$

$C_1 = C_0$, throw a coin

$$\delta_0 C_0 \geq \delta_0 \min(C_0, C_1)$$
$$+ \qquad +$$
$$\delta_1 C_1 \geq \delta_1 \min(C_0, C_1)$$
$$\delta_0 C_0 + \delta_1 C_1 \geq \min(C_0, C_1)$$

$$\frac{f_1(x)}{f_0(x)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{(C_{10} - C_{00}) \mathbb{P}(H_0)}{(C_{01} - C_{11}) \mathbb{P}(H_1)}$$

← Likelihood Ratio Test

minimal average cost.

$H_0 : f_0(x), \mathbb{P}(H_0)$

$H_1 : f_1(x), \mathbb{P}(H_1)$

$$e(\delta_0, \delta_1) = \sum_{i=0}^{1} \sum_{j=0}^{1} C_{ij} \mathbb{P}(D_i \& H_j)$$

$$\frac{f_1(x)}{f_0(x)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{(C_0 - C_{00}) \mathbb{P}(H_0)}{(C_{01} - C_{11}) \mathbb{P}(H_1)}$$

$C_{00} = C_{11} = 0$

$C_{01} = C_{10} = 1$

$$\mathbb{P}(D_0 \& H_1) + \mathbb{P}(D_1 \& H_0)$$

$$\frac{f_1(x)}{f_0(x)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)}$$

$$\mathbb{P}(H_1) f_1(x) \underset{H_0}{\overset{H_1}{\gtrless}} \mathbb{P}(H_0) f_0(x)$$

$$\frac{\mathbb{P}(H_1) f_1(x)}{C(x)} \underset{<}{\geq} \frac{\mathbb{P}(H_0) f_0(x)}{C(x)}$$

$$\frac{\mathbb{P}(H_1) f_1(x)}{\mathbb{P}(H_1) f_1(x) + \mathbb{P}(H_0) f_0(x)} \underset{<}{\geq} \frac{\mathbb{P}(H_0) f_0(x)}{\mathbb{P}(H_1) f_1(x) + \mathbb{P}(H_0) f_0(x)}$$

$$\mathbb{P}(H_1|x) \underset{<}{\geq} \mathbb{P}(H_0|x)$$

$$\text{sign}\left(\mathbb{P}(H_1|x) - \mathbb{P}(H_0|x)\right) \underset{H_0}{\overset{H_1}{\gtrless}} 0$$

$+1 \Rightarrow H_1$

$-1 \Rightarrow H_0$

* present & future,

$H_0 : f_0(x), \mathbb{P}(H_0)$ $\quad x_1^0, x_2^0 \cdots x_{n_0}^0$

$H_1 : f_1(x), \mathbb{P}(H_1)$ $\quad x_1^1, x_2^1 \cdots x_{n_1}^1$

classification.

$u(x, \theta)$

$$\frac{1}{n_0 + n_1}\left(\sum_{i=1}^{n_0}\left(1 + u(x_i^0, \theta)\right)^2 + \sum_{i=1}^{n_1}\left(1 - u(x_i^1, \theta)\right)^2\right)$$

$$\frac{n_0}{n_0+n_1}\frac{1}{n_0}\sum_{i=1}^{n_0}\left(1 + u(x_i^0, \theta)\right)^2$$

$n_0, n_1 \to \infty$

$$\mathbb{P}(H_0) + \mathbb{E}_0\left[(1+u(x,\theta))^2\right] \xrightarrow[\text{neural network}]{\text{large}} \mathbb{P}(H_0) + \mathbb{E}_0\left[(1+u(x))^2\right]$$

$$\min_u \mathbb{P}(H_0)\,\mathbb{E}_0[(1+u(x))^2] + \mathbb{P}(H_1)\,\mathbb{E}[(1-u(x))^2]$$

$$u(x) = \frac{\mathbb{P}(H_1|x) - \mathbb{P}(H_0|x)}{\mathbb{P}(H_1|x) + \mathbb{P}(H_0|x)} \rightarrow \text{equal to Bayesian.}$$

Hinge - cost

Cross - Entropy

* only two classification.

multiple hyposis

$H_0 = f_0(x), \mathbb{P}(H_0)$

$H_1 = f_1(x), \mathbb{P}(H_1)$

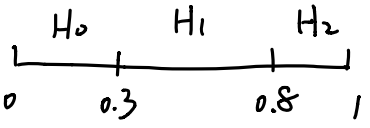
$\vdots$

$H_{k-1} = f_{k-1}(x), \mathbb{P}(H_{k-1})$

$\delta_0(x), \delta_1(x), \cdots, \delta_{k-1}(x)$

$\delta_0(x), \delta_1(x), \delta_2(x)$

0.3 0.5 0.2



$\{D_i \& H_j\}$ C_{ij}

$$e(\delta_0, \delta_1 \cdots \delta_{k-1}) = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} C_{ij}\,\mathbb{P}(D_i|H_j)\,\mathbb{P}(H_j)$$

$$\int \delta_i(x) f_j(x)\,dx$$

$$\int \sum_{i=0}^{k-1} \delta_i(x) \frac{\left(\sum_{j=0}^{k-1} C_{ij}\,\mathbb{P}(H_j) f_j(x)\right)}{C_i(x)}\,dx$$

$$= \int \left(\sum_{i=0}^{k-1} \delta_i(x)\, C_i(x)\right) dx$$

$$\delta_0 \cdot C_0 + \delta_1 \cdot C_1 + \cdots + \delta_{K-1} \cdot C_{K-1} \geq \min\{C_0, C_1, \cdots, C_{K-1}\}$$

$$\delta_0 C_0 \geq \min\{C_0, C_1, \cdots C_{K-1}\}\,\delta_0$$

$$\delta_1 C_1 \geq \delta_1 (\quad)$$

$$\vdots$$

$$\delta_{K-1} C_{K-1} \geq \delta_{K-1} (\quad)$$

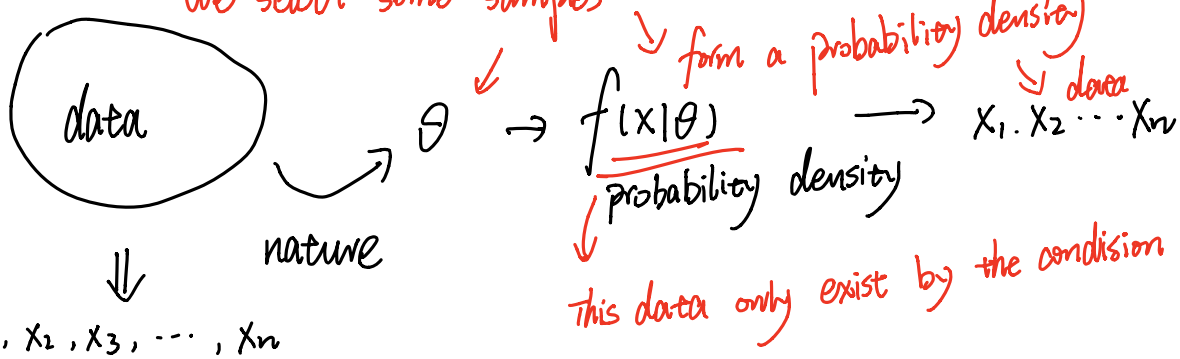
$C_0 = \min \rightarrow \delta_0 = 1,\ \delta_1 = \delta_2 = \cdots = \delta_{K-1} = 0$

$C_1 = \min \rightarrow \delta_1 = 1,\ \delta_0 = \delta_2 = \cdots = \delta_{K-1} = 0$

PARAMETER ESTIMATION

$X_1, X_2, \cdots, X_n$ $\quad \theta$

car (x, y)



we don't know the data.

It comes from nature.

we want to estimate it.

Estimate is a function of data.

$X \longrightarrow \vartheta(X)$

data estimation

$X_1, X_2, \cdots, X_n$

$\hat{\mu} = \frac{X_1 + X_2 + \cdots + X_n}{n}$ good one

$\hat{\mu}' = \frac{X_1 + X_2^2 + X_3^3 + \cdots + X_n^n}{n}$ bad one

} both estimation.

1. $f(x|\theta)$ we can describe the behavor of θ.

$\pi(\theta)$ probability density. We treat θ as random.

prior density

hard to know

$f(x|\theta) \cdot \pi(\theta) = f(x, \theta)$

$f(\theta|x) = \frac{f(x,\theta)}{f(x)} = \frac{f(x,\theta)}{\int f(x,\theta)\, d\theta}$

post density

Bayesian Estimation Theory

X, θ, $\hat{\theta}$

data from θ — true — estimate

$C(\theta, \hat{\theta})$ cost function

$\| $

$C(\theta, \hat{\theta}(x))$

random, random

Average cost $= \iint C(\theta, \hat{\theta}(x)) f(x,\theta)\, dx\, d\theta = e(\hat{\theta})$

joint pdf

function of a function of x.

$\iint C(\theta, \hat{\theta}(x)) f(\theta|x) \cdot f(x)\, dx\, d\theta$

$\int \left(\int C(\theta, \hat{\theta}(x)) f(\theta|x)\, d\theta \right) f(x)\, dx$

$\int \left(\int C(\theta, u) f(\theta|x)\, d\theta \right) f(x)\, dx = \int G(x, u) f(x)\, dx$

number $\rightarrow \min_{x,u} G(x,u)$ $\Delta x \to 0$ $\Delta u \to 0$

function $\rightarrow \min_{u} G(x,u) : \phi(x)$ $G(x,u) \geq \phi(x)$ $G(x, u_0(x)) = \phi(x)$

$$G_T(x,u) = (u - \log(1+x^2))^2 + x^4$$

$$\min_{u,x} G_T(x,u) = 0 \qquad u = x = 0$$

$$\min_{u} G_T(x,u) = x^4 \qquad U_0(x) = \log(1+x^2)$$

$$e(\hat{\theta}) = \int \underbrace{\left(\int C(\theta, \hat{\theta}(x)) f(\theta|x)\, d\theta \right)}_{G_T(x,\hat{\theta}(x)) \geq \phi(x)} f(x)\, dx$$

$$e(\hat{\theta}) = \int G_T(x, \hat{\theta}(x)) f(x)\, dx \geq \int \phi(x) f(x)\, dx$$

a known number

find $\hat{\theta}$ satisfy above.

$$\hat{\theta}_0(x) = \arg\min_u G_T(x,u)$$

$$G_T(x,u) = \int C(\theta, u) f(\theta|x)\, d\theta.$$

$$\nabla_u G_T(x,u) = 0$$

$$\int \nabla_u C(\theta,u) f(\theta|x)\, d\theta = 0 \qquad C(\theta,\hat{\theta}) = \|\theta - \hat{\theta}\|^2$$

MMSE

minimal min square estimator

$$(\theta-u)^T(\theta-u) = \theta^T\theta - 2u^T\theta + u^Tu$$

$$\nabla_u \phi(u_1, u_2 \cdots u_k) = \begin{pmatrix} \frac{\partial \phi}{\partial u_1} \\ \frac{\partial \phi}{\partial u_2} \\ \vdots \\ \frac{\partial \phi}{\partial u_k} \end{pmatrix}$$

vector

$$-2\theta + 2u$$

$$= -2(u-\theta)$$

vector

$$(-2)\int (u-\theta) f(\theta|x)\, d\theta = 0$$

vector vector

$$\int u f(\theta|x)\, d\theta = \int \theta f(\theta|x)\, d\theta$$

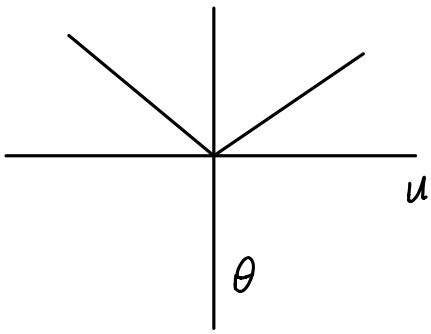
$$u = E[\theta|x]$$

$$\hat{\theta}_{MMSE}(x) = E[\theta|x]$$

θ scalar

$C(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$ $\quad C(\theta, u) = |u - \theta|$

$$\frac{d|u-\theta|}{du} = \text{sgn}(u-\theta)$$



$$\int \text{sgn}(u-\theta) f(\theta|x)\, d\theta = \int_{-\infty}^{u} 1 \cdot f(\theta|x)\, d\theta + \int_{u}^{\infty} -1 \cdot f(\theta|x)\, d\theta$$

cdf

$$F(u|x) - (1 - F(u|x)) = 0$$

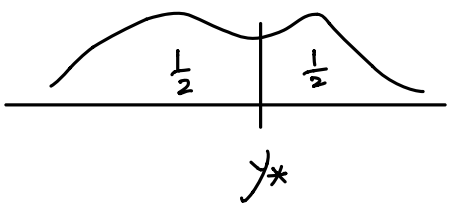
$$F(u|x) = \frac{1}{2}$$

y, $f(y)$, $F(y)$

median $F(y_*) = \frac{1}{2}$

$$C(\theta, \hat{\theta}) = |\theta_1 - \hat{\theta}_1| + |\theta_2 - \hat{\theta}_2| + \dots$$

median of median of



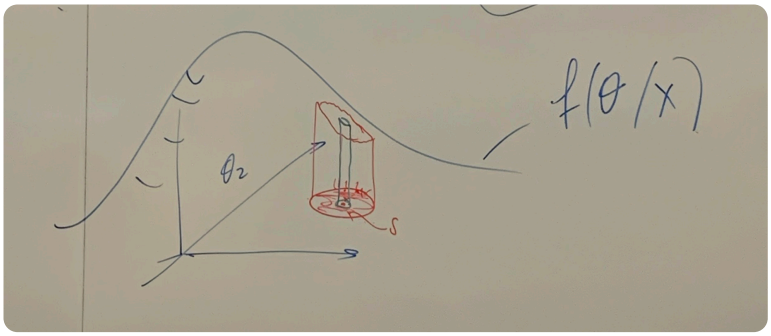
MAP Estimator

$$C(\theta, \hat{\theta}) = \begin{cases} 0 & \|\theta - \hat{\theta}\| = \delta \\ 1 & \text{otherwise} \end{cases}$$

$$G(x, u) = \int C(\theta, u) f(\theta|x)\, d\theta$$

$$= \int_{\|\theta-u\| \le \delta} 0 \cdot f(\theta|x)\, d\theta + \int_{\|\theta-u\| > \delta} 1 \cdot f(\theta|x)\, d\theta$$

$$+ \int_{\|\theta-u\| \le \delta} f(\theta|x)\, d\theta - \int_{\|\theta-u\| \le \delta} f(\theta|x)\, d\theta$$



$$= 1 - \int_{\|\theta - u\| \le \delta} f(\theta|x)\, d\theta$$

δ is very small.

$$\approx B_\delta \cdot f(u|x)$$

$$\hat{\theta}_{MAP}(x) = \arg\max_\theta f(\theta|x)$$

$$\arg\max_\theta \frac{f(x,\theta)}{f(x)} = \arg\max_\theta f(x,\theta)$$

$$\arg\max_\theta f(x|\theta)\,\pi(\theta) \qquad \theta \in \Omega$$

$$\pi(\theta) = \begin{cases} \frac{1}{Vol(\Omega)} & \theta \in \Omega \\ 0 & \text{otherwise.} \end{cases}$$

$$\arg\max_\theta f(x|\theta)\,\frac{1}{Vol(\Omega)}$$

number

$$\hat{\theta}_{MLE}(x)$$

$\hat{\theta}(x)$ unbiased estimator

average : $\int \hat{\theta}(x) f(x|\theta)\, dx = \theta$

$$\int (\hat{\theta}(x) - \theta)(\hat{\theta}(x) - \theta)^T f(x|\theta)\, dx = \Sigma_{\hat{\theta}} \ge (I_F(\theta))^{-1}$$

Fisher information matrix

Grammer – Rao Lower Bound

CRLB

$$\int \left(\frac{\nabla_\theta f(x|\theta)}{f(x|\theta)}\right)(\quad)^T f(x|\theta)\, dx) = I_F(\theta)$$

$X_1, X_2, \cdots, X_n$ independent

$X_i \sim \mathcal{N}(\mu, 1)$ Gaussian.

$\pi(\mu) \sim \mathcal{N}(\mu_0, \sigma^2)$

$$f(x_1, x_2, \cdots x_n / \mu) = \prod_{i=1}^{n} \frac{e^{-\frac{1}{2}(x_i-\mu)^2}}{\sqrt{2\pi}} = \left(\frac{1}{\sqrt{2\pi}}\right)^n \cdot e^{-2\sum_{i=1}^{n}(x_i-\mu)^2}$$

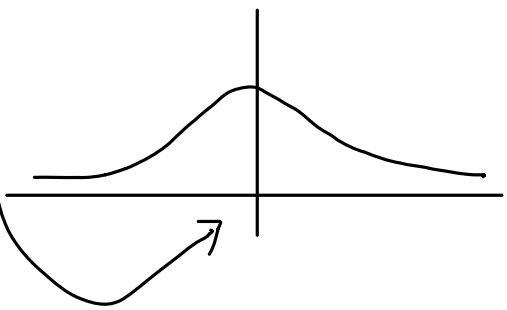
$$\min_{\mu} \sum_{i=1}^{n}(x_i-\mu)^2 \rightarrow \sum_{i=1}^{n}(-2)(x_i-\mu) = 0$$

$$\frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

$$f(\mu / x_1, x_2 \cdots x_k) = \mathcal{N}\left(\frac{\sum_{i=1}^{n} x_i + \frac{\mu_0}{\sigma^2}}{n+\frac{1}{\sigma^2}}, \frac{1}{n+\frac{1}{\sigma^2}}\right)$$

$$\mu_{MMSE} = E[\mu / x] = \frac{\sum_{i=1}^{n} x_i + \frac{\mu_0}{\sigma^2}}{n+\frac{1}{\sigma^2}}$$

$\mu_{median} =$



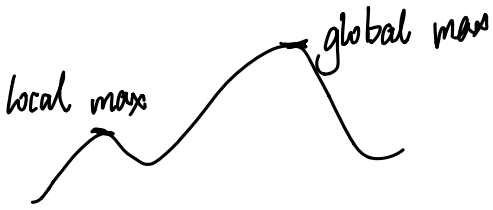
Optimal problem

1) function optimize with constraint
2) linear programming
3) optimize functional (function of function)
   Calculus of variations

$g(x_1, x_2 \cdots x_n) = g(x) \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$

$\max_x g(x)$ exist a x_* that $g(x_*) \geq g(x)$ ↗ global maximum.

Supreme $\sup_x g(x)$ x_* may not exist but there's a max limit.



there's no alg to find global max.
It's always a local max.

Hessian matrix = $H_{ij} = \frac{\partial^2 g}{\partial x_i \partial x_j}$ matrix function of X.

$H(x_i) \rightarrow$ positive definite (min)
negative definite (max).
neither (subtle points)

positive def: $x^T Q x > 0$ (Q: pos def matrix)
$x \neq 0$

$X = X_* + \delta$

$g(x) = g(X_* + \delta) = g(X_*) + (\nabla_x g(X_*))^T \delta + \frac{1}{2}\delta^T H(X_*)\delta$ ($\nabla_x g(X_*) = 0$)

$g(X_* + \delta) - g(X_*) = \frac{1}{2}\delta^T H(X_*)\delta$ compare 0

$\max_x g(x_1, x_2, \cdots, x_N) = g(x)$

↓ x has to satisfy

$w(x) = 0$

$\max_{x_1, x_2} g(x_1, x_2)$

$\max_{x_1} \{\max_{x_2} g(x_1, x_2)\} \rightarrow \tilde{g}(x_1)$ → It's like going on the ridge while climbing a mountain.

$\max_{x_2} \{\max_{x_1} g(x_1, x_2)\} \rightarrow \tilde{g}(x_2)$

$\max_{x,y} g(x, y)$ $x = 5$

$x + y = 1$ $y = 1 - x$ $\max_x g(x, 1-x)$ These three are equal.

Lagrauge multiplier method

$\max_x g(x)$ — $w_1(x) = 0$, $w_2(x) = 0$

$\tilde{g}(x) = g(x) + \lambda_1 w_1(x) + \lambda_2 w_2(x)$

Lagrauge factors.

$\max_x \tilde{g}(x) \rightarrow x_*(\lambda_1, \lambda_2)$

$w_1(x_*(\lambda_1, \lambda_2)) = 0$

$w_2(x_*(\lambda_1, \lambda_2)) = 0$

$\Rightarrow \lambda_1^*, \lambda_2^* \rightarrow x_*(\lambda_1^*, \lambda_2^*)$

possible solution.

$\mathcal{X} = \{x : w_1(x) = 0, w_2(x) = 0\}$

$\max_{x \in \mathcal{X}} g(x)$

$x \in \mathcal{X}$

$g(x) = g(x) + \lambda_1^* w_1(x) + \lambda_2^* w_2(x) = \tilde{g}(x)$

($w_1(x) = 0$, $w_2(x) = 0$)

$\tilde{g}(x) \leq \tilde{g}(x_*(\lambda_1^*, \lambda_2^*))$

$= g(x_*(\lambda_1^*, \lambda_2^*)) + \lambda_1^* w_1(x_*(\lambda_1^*, \lambda_2^*)) + \lambda_2^* w_2(x_*(\lambda_1^*, \lambda_2^*))$

Example.

Q positive definite

$x \neq 0$

$g(x) = x^T Q x > 0$

$\min_x x^T Q x = 0$

$x = 0$

$A^T x = b$

$a_1^T x = b_1$

$a_2^T x = b_2$

$\vdots$

$a_k^T x = b_k$

k = dimension of x.

find $\min_x x^T Q x$.

$x^T Q x + \lambda_1 (a_1^T x - b_1) + \lambda_2 (a_2^T x - b_2) + \cdots + \lambda_k (a_k^T x - b_k)$

take gradient : $2Qx + \lambda_1 a_1 + \lambda_2 a_2 + \cdots + \lambda_k a_k = 0$

$[a_1 \ a_2 \cdots a_k] \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_k \end{bmatrix}$

A λ

$$2Qx + A\lambda = 0 \rightarrow \quad x_* = -\tfrac{1}{2}Q^{-1}A\lambda$$

$$A^T x = -\tfrac{1}{2}A^T Q^{-1} A\lambda = b \quad \Rightarrow \quad \lambda^* = -2(A^T Q^{-1} A)^{-1} b$$

$$x_* = -\tfrac{1}{2}Q^{-1}A \cdot -2(A^T Q^{-1} A)^{-1} b$$

$$= Q^{-1}A(A^T Q^{-1} A)^{-1} b$$

LINEAR PROGRAMMING.

$x_1, x_2, \cdots x_n$

$$\min_{x_1 \cdots x_m} \{g_1 x_1 + g_2 x_2 + \cdots + g_m x_m\}$$

$$G = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_m \end{bmatrix} \qquad X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

Constraints:

linear

$$a_{11}x_1 + \cdots + a_{1m}x_m = b_1$$

$$\vdots$$

$$a_{m1}x_1 + \cdots + a_{mm}x_m = b_m$$

$$A_e X = b_e$$

equality

non linear

$$\alpha_{11}x_1 + \cdots + a_{1n}x_n \leq \beta_1$$

$$\vdots$$

$$\alpha_{k1}x_1 + \cdots + \alpha_{kn}x_n \leq \beta_k$$

$$\min_x G^T X$$

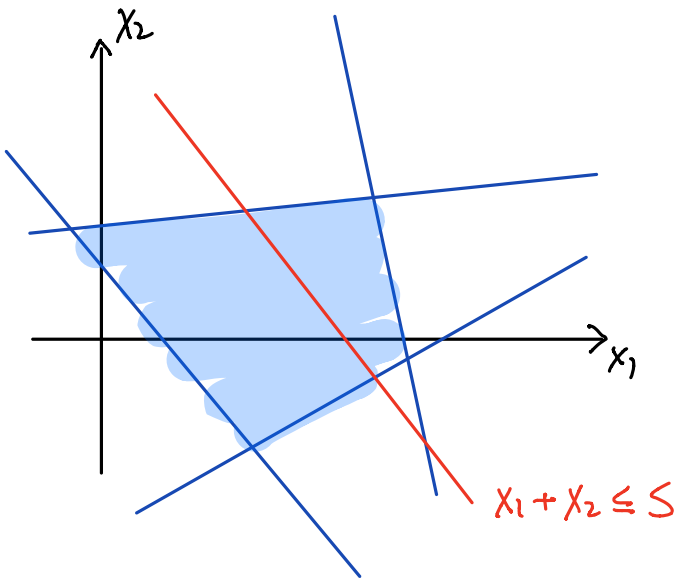
$$Ax \leq b$$

$$\min_{x_1, x_2} \{x_1 + x_2\}$$

$$x_1 + 2x_2 \leq 2$$

$$3x_1 + x_2 \leq 0$$

$$\vdots$$



$$k_1 = |C_{11}x_1 + \cdots + C_{1n}x_n|$$
$$k_2 = |C_{21}x_1 + \cdots + C_{2n}x_n| \qquad Aex = be$$
$$\vdots \qquad Ax \le b$$
$$k_s = |C_{s1}x_1 + \cdots + C_{sn}x_n|$$

$$\min_{x_1 \cdots x_n} \{ \max \{k_1, k_2 \cdots k_s\}\}$$

$$-\delta \le C_{11}x_1 + \cdots + C_{1n}x_n \le \delta$$
$$\vdots$$
$$-\delta \le C_{s1}x_1 + \cdots + C_{sn}x_n \le \delta$$
$$g_1x_1 + g_2x_2 + \cdots + g_nx_n + g_\delta \cdot \delta$$

CALCULUS OF VARIATION

$F(x, y)$

$g(x)$ $\qquad \min_{g(x)} \int_{x_1}^{x_2} F(x, g(x))\,dx \qquad g_0(x) + \epsilon\delta(x)$

$$\Omega(\epsilon) = \int_{x_1}^{x_2} F(x, g_0(x) + \epsilon\delta(x))\,dx \ge \Omega(0)$$
$$\Omega'(0) = 0$$

$$\Omega'(\epsilon) = \int_{x_1}^{x_2} \frac{d}{d\epsilon} F(x, g_0(x) + \epsilon\delta(x))\,dx$$

$$= \int_{x_1}^{x_2} \frac{\partial}{\partial y} F(x, g_0(x))\,\delta(x)\,dx = 0$$

any function.

$$\int_{x_1}^{x_2} \phi(x) \cdot \delta(x)\,dx = 0 \Rightarrow \phi(x) = 0$$

$$\frac{\partial}{\partial y} F(x, g_0(x)) = 0$$

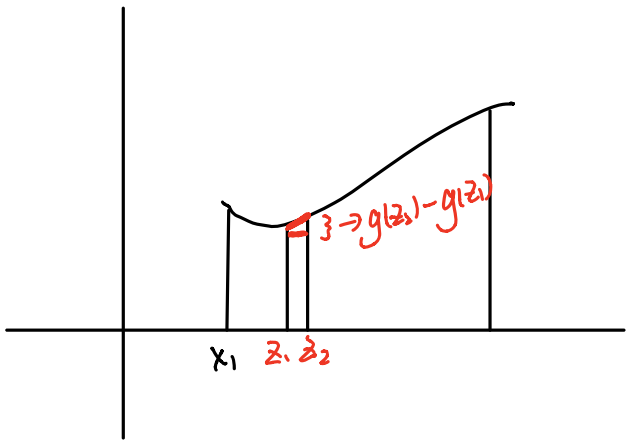
$$\frac{d}{dy} F(x, y) = F_y(x, y) = 0$$

$F(x, y, z)$ $\qquad g(x) = g_0(x) + \epsilon \delta(x)$

$$\min_{g(x)} \int_{x_1}^{x_2} F(x, g(x), g'(x))\, dx.$$

$$g(x_1) = \alpha_1$$
$$g(x_2) = \alpha_2$$

$$\frac{d}{dy} F(x, g_0(x), g_0'(x)) - \frac{d}{dx}\left(\frac{\partial}{\partial \epsilon} F(x, g_0(x), g_0'(x)\right) = 0.$$



$$\sqrt{(g(z_2) - g(z_1))^2 + (z_2 - z_1)^2}$$

$$\sum \sqrt{1 + \left(\frac{g(z_2) - g(z_1)}{z_2 - z_1}\right)^2}\,(z_2 - z_1)$$

$$l(g) = \int_{x_1}^{x_2} \sqrt{(g'(x))^2 + 1}\, dx$$

$$F(x, y, z) = \sqrt{1 + z^2}$$

$$\frac{d}{dx}\left(\frac{g'(x)}{\sqrt{1 + (g'(x))^2}}\right) = 0$$

$$\frac{g'(x)}{\sqrt{1 + (g'(x))^2}} = C \qquad g(x) = \sqrt{\frac{C^2}{1 - C^2}}\, x + b.$$

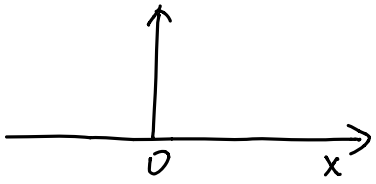
SVM : approximate non-linear function

Kernels

1) Approx function
2) classification
   image compare

Kernels approximate dirac delta function.

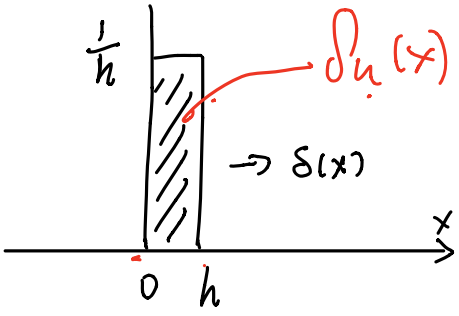
Dirac delta function $\delta(x)$



$$\int_{-\varepsilon}^{\varepsilon} \delta(x)dx = 1$$

$$\int \phi(x)\,\delta(x)dx = \phi(0)$$ → a way of definition

$$\int \phi(x)\,\delta'(x)dx = -\phi'(0)$$



$$\int_{-\infty}^{\infty} g(x)dx = 1. \quad g(x,h) = g\left(\frac{x}{h}\right)\frac{1}{h}$$

$$\int_{-\infty}^{\infty} \phi(x)\,\delta_h(x)dx = \frac{1}{h}\int_0^h \phi(x)dx \xrightarrow{h\to 0} \frac{1}{h}\int_0^h \phi(0)dx = \phi(0)\frac{1}{h}\int_0^h dx$$

($\delta_h(x)$)

$$g(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \quad \text{(gaussian approximation)}$$

$$\int \phi(x)\,\delta(x)\,dx = \phi(0)$$

$$\delta(x) = \delta(x_1)\cdot\delta(x_2)\cdots\delta(x_k) \quad \text{multi-dimension of } \delta(x)$$

$$\delta(x_1) = \frac{e^{-\frac{1}{2}\frac{x^2}{h^2}}}{\sqrt{2\pi h^2}}$$

$$\delta(x,h) = \frac{e^{-\frac{1}{2}\frac{x_1^2}{h_1^2}}}{\sqrt{2\pi h^2}} \cdot \frac{e^{-\frac{1}{2}\frac{x_2^2}{h_2^2}}}{\sqrt{2\pi h^2}} \cdots = \frac{e^{-\frac{1}{2h^2}\|x\|^2}}{\sqrt{(2\pi)^k (h^2)^k}}$$

$$\int \phi(x)\,\delta(\underbrace{x - x_0}_{z})\,dx = \phi(x_0)$$

$$\int \phi(z+x_0)\,\delta(z)\,dz = \phi(x_0) \rightarrow x = z + x_0$$

$$\phi(x_0) = \int \phi(x)\,\delta(x - x_0)\,dx \qquad K(x,h) \text{ approximation of } \delta(x)$$

$$\approx \int \phi(x)\,K(x - x_0, h)\,dx.$$

$$\int_{-\infty}^{+\infty} f(x)\cdot\delta(x_0 - x)\,dx = f(x_0)$$

$$\Downarrow$$

$$f(x_0) = \mathbb{E}_x[\delta(y - x)] \approx \mathbb{E}_{x_0}[K(x_0 - x, h)]$$

$\mathcal{X}$: $x_1, x_2, x_3 \cdots x_k$

$$\mathbb{E}[x] \approx \frac{x_1 + x_2 + \cdots + x_k}{k} \qquad \mathbb{E}[x^2] \approx \frac{x_1^2 + x_2^2 + \cdots + x_k^2}{k}$$

$$\bar{E}[g(x)] \approx \frac{g(x_1)+g(x_2)+\cdots+g(x_k)}{k}$$

$$\bar{E}_x[k(x,h)] \approx \frac{k(x_1,h)+k(x_2,h)+\cdots+k(x_k,h)}{k}$$

$$\bar{E}_x[k(Y-x,h)] \approx \frac{k(y-x_1,h)+k(y-x_2,h)+\cdots+k(y-x_k,h)}{k}$$

$$f(x) \approx \frac{k(x-x_1,h)+k(x-x_2,h)+\cdots+k(x-x_k,h)}{k}$$

kernel function

approximation of $\delta(x)$

classification

class 1 & 2 $\{ x_1^1, x_2^1, \cdots, x_k^1 \quad x_1^2, x_2^2, \cdots, x_k^2$

$$\frac{\hat{f}_1(x)}{\hat{f}_2(x)} \gtrless 1 \quad \text{(likelyhood ratio)}$$

greater than 1 $\rightarrow X^1$

smaller than 1 $\rightarrow X^2$

OR. maybe make estimation of the entire ratio,

MERCER KERNELS

$k(x,y)$ — scalar function

1) $k(x,y) = k(y,x)$

2) positive definite : $\iint \phi(x)\phi(y)\,k(x,y)\,dx\,dy > 0$. $\phi(x) \neq 0$

$\phi(x)$ scalar function

$\iint \phi(x)\phi(y)\,k(x,y)\,dx\,dy = 0$ if and only if $\phi(x) = 0$

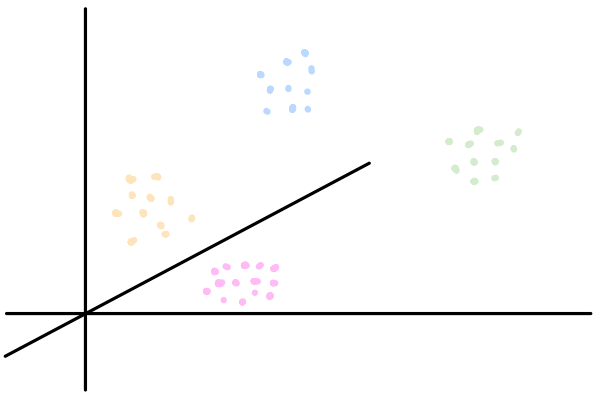
different kernels.

$k(x,y) = (x^T y + c)^2$

☆ $k(x,y) = e^{-\frac{1}{2h}\|x-y\|^2}$

$k(x,y) = e^{-\frac{1}{2h}\|x-y\|}$

$k(x,y) = e^{-\frac{1}{h}\|x-y\|_{L_1}}$ $\quad \|z\|_{L_1} = |z_1|+|z_2|+\cdots+|z_k|$



$x_0 \in \mathbb{R}^n \longrightarrow k(x, x_0)$

$\phi(x) \in V$

$\phi(x) = \alpha_1 k(x, x_1) + \cdots + \alpha_m k(x, x_m)$

↑ vector space.

$\phi(x) = \alpha_1 k(x, x_1) + \cdots + \alpha_m k(x, x_m)$

$\psi(x) = \beta_1 k(x, x_1') + \cdots + \beta_{m'} k(x, x_{m'}')$

$$\langle \phi, \psi \rangle \triangleq \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \cdot \beta_j k(x_i, x_j')$$

$$\langle \phi, \phi \rangle = \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \cdot \alpha_j k(x_i, x_j) \stackrel{?}{\geq} 0$$

$\phi(x) \in V$

$\phi(x) = \alpha_1 k(x, x_1) + \cdots + \alpha_m k(x, x_m)$

$k(x, x_0)$

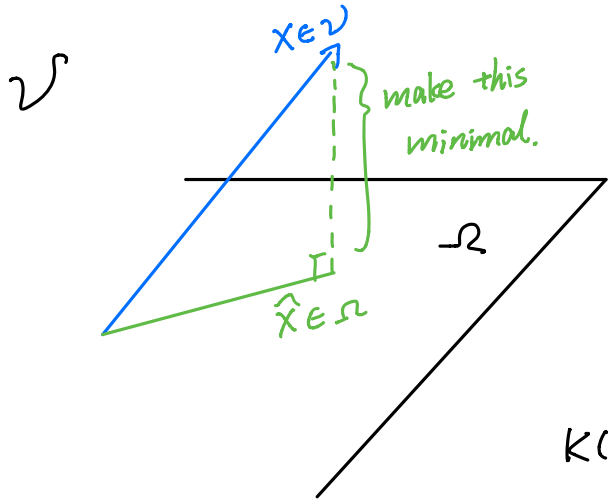
$\langle k(x, x_0), \phi(x) \rangle = \alpha_1 \langle k(x, x_1), k(x, x_0) \rangle +$

$\cdots + \alpha_m \langle k(x, x_m), k(x, x_0) \rangle$

$\alpha_1 k(x_0, x_1) + \cdots + \alpha_m k(x_0, x_m) = \phi(x_0)$

$$\begin{bmatrix} k(x_1, y_1) & k(x_1, y_2) & \cdots \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}$$

$k(x, y)$ is positive definite.



$\Omega = \{\alpha_1 z_1 + \cdots + \alpha_n z_n\}$

$z_1, z_2 \cdots z_n \in \mathcal{V}$

$x - \hat{x} \perp z_1, z_2 \cdots z_n$

$K(x, x_1)$, $K(x, x_2)$ $\cdots$ $K(x, x_n)$
z_1 z_2 z_n

$\Omega = \{\alpha_1 K(x, x_1) + \cdots + \alpha_n K(x, x_n)\}$

$\phi(x) \in \mathcal{V}$ $\quad \phi(x) = \beta_1 K(x, x_1') + \cdots + \beta_m K(x, x_m')$

$\phi(x) - \hat{\phi}(x) \perp K(x, x_1), K(x, x_2) \cdots K(x, x_n)$

$\langle \phi(x) - \hat{\phi}(x), K(x, x_i) \rangle = 0$

Representer Theorem.

$\phi(x_i) = \langle \phi(x), K(x, x_i) \rangle = \langle \hat{\phi}(x), K(x, x_i) \rangle = \hat{\phi}(x_i)$

$\phi(x_i) = \hat{\phi}(x_i)$

$(y_1, x_1), (y_2, x_2) \cdots (y_k, x_k)$ — data

$\phi(x)$ $\qquad$ $l(y, \phi(x)) = (y - \phi(x))^2$

$\min_{\phi(x) \in \mathcal{V}} \sum_{i=1}^{k} l(y_i, \hat{\phi}(x_i))$ $\qquad \phi(x) \in \mathcal{V}$

every kernel function define a $\mathcal{V}$ (vector space)

$\phi(x) = \alpha_1 K(x, x_1') + \alpha_2 K(x, x_2') + \cdots + \alpha_m K(x, x_m')$

$\phi(x) = \alpha_1 K(x, x_1) + \alpha_2 K(x, x_2) + \cdots + \alpha_n K(x, x_n)$

$\min_{\alpha_1 \cdots \alpha_n} \sum_{i=1}^{k} (y_i - \alpha_1 K(x_i, x_1) - \alpha_2 K(x_i, x_2) - \cdots - \alpha_n K(x_i, x_n))^2$

Known numbers.

EASY TO OPTIMISE.

$\phi(x) = \alpha_1^* K(x, x_1) + \cdots + \alpha_n^* (x, x_n)$ $\qquad$ problem comes to linear.

Learning Algorithm.

$$G(\theta) = \begin{bmatrix} G_1(\theta_1, \cdots \theta_k) \\ G_2(\theta_1, \cdots \theta_k) \\ \vdots \\ G_k(\theta_1, \cdots, \theta_k) \end{bmatrix}$$

$$G(\theta) = 0$$

Assume $G(\theta)$ is continue in θ.

$$v_t = v_{t-1} - \mu G(v_{t-1}) \quad (\mu: \text{scaler})$$

$$v_t \to v_\infty$$

$$\lim_{t\to 0} v_t = \lim_{t\to\infty} v_{t-1} - \mu \lim_{t\to\infty} G(v_{t-1})$$

$$v_\infty \quad v_\infty \quad G(v_\infty)$$

$$v_\infty = v_\infty - \mu G(v_\infty) \to G(v_\infty) = 0$$

$$g(\theta) = g(\theta_1, \theta_2, \cdots \theta_k)$$

$$\min_\theta g(\theta)$$

$$\nabla_\theta g(\theta) = 0$$

$$\nabla_\theta g(\theta) = \begin{bmatrix} \frac{dg}{d\theta_1}(\theta_1, \theta_2, \cdots \theta_k) \\ \vdots \\ \frac{dg}{d\theta_k}(\theta_1, \theta_2 \cdots \theta_k) \end{bmatrix}$$

Gradient Descent

$$v_t = v_{t-1} - \mu \nabla_\theta g(v_{t-1})$$

(μ: step size)

$$H(z, \theta) = \begin{bmatrix} H_1(z_1, z_2 \cdots z_n \theta_1 \theta_2 \cdots \theta_k) \\ \vdots \\ H_n(z_1, z_2 \cdots z_n \theta_1 \theta_2 \cdots \theta_k) \end{bmatrix}$$

Assume.

z is random

pdf of z : $f(z)$

$$G(\theta) = E_z[H(z, \theta)] = 0$$

$$G(\theta) = \int H(z, \theta) f(z) dz$$

we don't know

we have to know data. $z_1, z_2 \cdots$

$$v_t = v_{t-1} - \mu \underbrace{E_z[H(z, v_{t-1})]}_{G(v_{t-1})} \to \theta_t = \theta_{t-1} - \mu H(z_t, \theta_{t-1})$$

$G(\theta) = 0$

deterministic

data point

random (z_t is random)

$$\theta_t = \bar{\theta}_t + v_t$$

$$\bar{\theta}_t = \bar{\theta}_{t-1} - \mu \overline{H(z_t, \theta_{t-1})} \rightarrow \bar{\theta}_t = \bar{\theta}_{t-1} - \mu \overline{H(z_t, \bar{\theta}_{t-1})}$$

not the same, but only cause an error of $O(\mu^2)$

$$\theta_t \rightarrow \theta_\infty = \theta_\infty + v_\infty$$

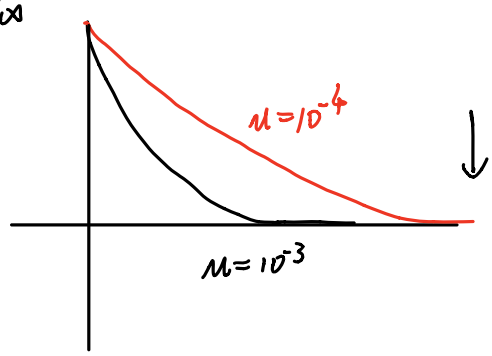
θ_∞: solution to $G(\theta) = 0$

v_∞: small bias ← we want to make this small.

$$\mathbb{E}[v_\infty v_\infty^T] = \mu Q + O(\mu^2)$$

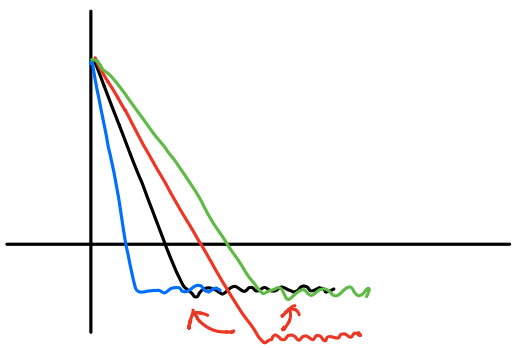
$$\theta_t = \theta_{t-1} - \mu H(z_t, \theta_{t-1})$$

v_∞ $\qquad$ $\|\theta_t - \theta_\infty\|^2$



↓ converge

← we can not evaluate Alg. by this.



$\Delta_1 = \mu_1 C_1 + O(\mu^2)$

$\Delta_2 \qquad \mu_1' \cdot C_1 = \Delta_2 \quad \mu_1' = \frac{\Delta_2}{C_1} = \frac{\Delta_2}{\Delta_1}\mu_1$

blue better than BLACK.

Green worse than BLACK.

$$\theta_t = \theta_{t-1} - \mu H(z_t, \theta_{t-1}) \qquad \underbrace{\mathbb{E}_z[H(z, \theta)]}_{G(\theta)} = 0$$

we are trying to solve this.

$h(z,\theta)$ $\quad g(\theta) = \mathbb{E}_z[h(z,\theta)]$ $\quad$ x pdf $\quad$ / data : $z_1, z_2 \cdots$

$$\theta_t = \theta_{t-1} - \mu \nabla_\theta h(z_t, \theta_{t-1})$$

SG (stochastic gradient) $\quad$ learning rate (step size)

$\nabla_\theta g(\theta) = 0$

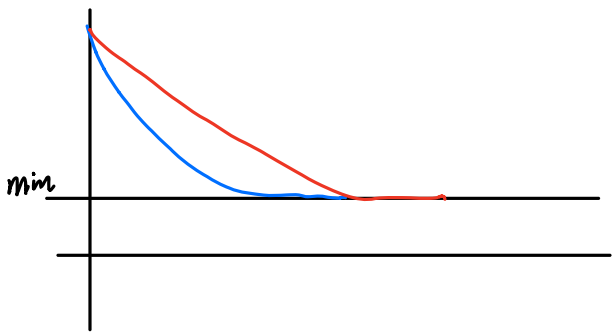
$G(\theta) = \mathbb{E}_z[\underbrace{\nabla_\theta h(z,\theta)}_{H(z,\theta)}]$

$G(\theta) = \int H(z,\theta) f(z)\, dz$

$\theta_t^i = \bar{\theta}_t^i + v_t^i \quad g(\theta_t^i)$

$\hookrightarrow v_\infty$

$$g(\theta_t^i) = g(v_\infty) + (\nabla_\theta g(v_\infty))^T v_t + \frac{1}{2} v_t^T (\nabla_\theta g(\theta_\infty)) v_t$$

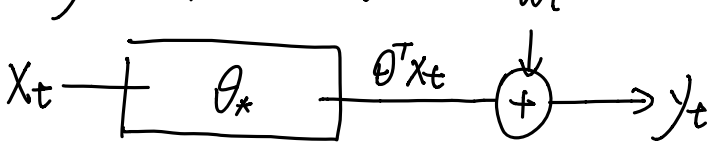


Avg (blue − min) vs Avg (red − min)

we need make sure μ is the same time step.

LINEAR REGRASSION.

$y_t = \theta_*^T x_t + w_t$



$\{w_t\}$ is independent from $\{x_t\}$

need to estimate θ_*.

$$g(\theta) = \mathbb{E}_z[(y - \theta^T x)^2] \qquad z = \begin{bmatrix} y \\ x \end{bmatrix}$$

$$= \mathbb{E}[(\theta_* - \theta)^T x + w)^2] \quad (\theta_* - \theta)^T \overline{x x^T} (\theta_* - \theta) + 2(\theta_* - \theta)^T \overline{xw} + \underbrace{\overline{w^2}}_{\sigma_w^2}$$

$$g(\theta) = (\theta_* - \theta)^T \underline{R_x} (\theta_* - \theta) + \underline{\sigma_w^2}$$ don't know

data: (y_1, x_1), (y_2, x_2) ...

$$\theta_t = \theta_{t-1} - \mu (y_t - \theta_{t-1}^T x_t) x_t$$

LMS

$$g(\theta) = \mathbb{E}_z[\underbrace{(y - \theta^T x)^2}_{\downarrow}]$$

$h(z, \theta)$ $z = \begin{bmatrix} y \\ x \end{bmatrix}$

$$\mathbb{E}[(y_t - \theta^T x_t) x_t] = 0$$

$$\theta_t^2 = \theta_{t-1}^2 - \mu \, sign(y_t - (\theta_{t-1}^2)^T x_t) x_t$$

$$\mathbb{E}[sign(y_t - \theta_*^T x_t) \cdot x_t] = 0$$

$$\mathbb{E}[sign(w_t) x_t] = \mathbb{E}[sign(w_t)] \cdot \mathbb{E}[x_t]$$

$$\theta_t^3 = \theta_{t-1}^3 - \mu (y_t - (\theta_{t-1}^3)^T x_t) \, sign(x_t)$$

$$g(\theta) = \mathbb{E}_z[h(z, \theta)] \qquad h(z, \theta), p(z, \theta)$$

$$g(\theta) = \underbrace{(\mathbb{E}_z[h(z, \theta)])^2}_{\alpha} + \underbrace{(\mathbb{E}_z[p(z, \theta)])^2}_{\beta}$$ ← Tensorflow can not solve

$$\nabla_\theta g(\theta) = 2\alpha \, \mathbb{E}_z[\nabla_\theta h(z, \theta)] + 2\beta \, \mathbb{E}_z[\nabla_\theta p(z, \theta)] = 0$$

$$\mathbb{E}_z[\alpha \nabla_\theta h(z, \theta) + \beta \nabla_\theta p(z, \theta)] = 0$$

$$\alpha - \mathbb{E}_z[h(z, \theta)] = 0$$

$$\mathbb{E}_z[\alpha - h(z, \theta)] = 0$$

$$\mathbb{E}_z[\beta - p(z, \theta)] = 0$$

$$\mathbb{E}_z[H(z, \theta)] = 0$$

$$\theta_t = \theta_{t-1} - \mu H(z_t, \theta_{t-1})$$

self-defined Alg. to solve this.

$$\begin{bmatrix} \theta \\ \alpha \\ \beta \end{bmatrix}_t = \begin{bmatrix} \theta \\ \alpha \\ \beta \end{bmatrix}_{t-1} - \mu \begin{bmatrix} \alpha_{t-1} \nabla_\theta h(z_t, \theta_{t-1}) + \beta_{t-1} \nabla \rho(z_t, \theta_{t-1}) \\ \alpha_{t-1} - h(z_t, \theta_{t-1}) \\ \beta_{t-1} - \rho(z_t, \theta_{t-1}) \end{bmatrix}$$

$$v_t = v_{t-1} - \mu [\bar{\mathbb{E}}_z \bar{\nabla} H(z, v_{t-1})]$$

A: $$\theta_t = \theta_{t-1} - \mu \frac{[H(z_t, \theta_{t-1}) + H(z_{t-1}, \theta_{t-1}) + \cdots + H(z_{t-m}, \theta_{t-1})]}{m+1}$$

B: $$\theta_t = \theta_{t-1} - \mu H(z_t, \theta_{t-1})$$

converge almost the same.

A will update every $m+1$ steps.

B will update every steps.

Even if A converge faster, it will update slower.

By using GPU, A is better.

$$G(\theta) = 0$$

$$G(\theta) = \begin{bmatrix} G_1(\theta_1 \cdots \theta_K) \\ G_2(\theta_1 \cdots \theta_K) \\ \vdots \\ G_K(\theta_1 \cdots \theta_K) \end{bmatrix} = 0$$

$$\theta_t = \theta_{t-1} - \mu G(\theta_{t-1})$$

$\{\theta_t\} \to \theta_\infty$ continue

$G(\theta) = 0$ solution for this.

we don't know distribution but have data.

$$G(\theta) = \bar{\mathbb{E}}_z [\nabla H(z, \theta)]$$

solvable algorithm. random

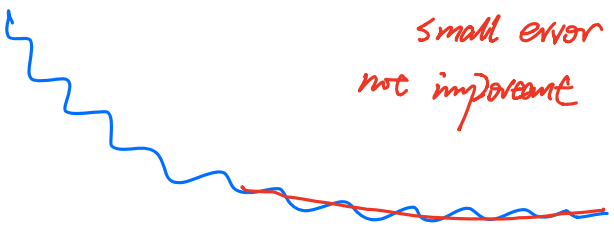
$$\theta_t = \theta_{t-1} - \mu H(z_t, \theta_{t-1})$$

$$\theta_t = \underbrace{\bar{\theta}_t}_{\text{mean}} + v_t$$

$G(\bar{\theta}_{t-1}) \to \theta_\infty$

$$\bar{\theta}_t = \bar{\theta}_{t-1} - \mu \bar{\mathbb{E}}_z [\nabla H(z_t, \bar{\theta}_{t-1})] + O(t^2)$$

small error not improvement



mean + noise

$\min g(\theta)$ $\quad \nabla_\theta g(\theta) = 0$

$\theta_t = \theta_{t-1} - \mu \nabla_\theta g(\theta_{t-1})$

μ too small converge too slow.

$\mathbb{E}[\|\mu_t\|^2] \sim \mu$

$g(\theta) = \mathbb{E}_z[h(z,\theta)]$ don't know distribution but have data.

solvable algorithm. $\theta_t = \theta_{t-1} - \mu \nabla_\theta h(z_t, \theta_{t-1})$

μ can reduce noise.

but slow down convergence.

$\theta_t = \theta_0 + \nu_t$

$\theta_t \to \theta_\infty$

$\Omega(\theta, w)$, $\Omega(\theta, 0) = \theta$

$\theta_t = \Omega(\theta_{t-1}, \mu G(\theta_{t-1}))$

$\approx \Omega(\theta_{t-1}, 0) + (J_w \Omega(\theta_{t-1}, 0))^T \mu G(\theta_{t-1})$

Jacobi matrix

$= \theta_{t-1} + \mu (J_w \Omega\ \theta_{t-1})^T G(\theta_{t-1})$

$(J_w \Omega(\theta, 0))^T G(\theta) = 0$

if $\Omega(\theta, w) = \theta + w$

$\theta_t = \theta_{t-1} - \mu G(\theta_{t-1})$

$G(\theta) = \mathbb{E}_z[H(z, \theta)]$

$\theta_t = \Omega(\theta_{t-1}, \mu H(z_t, \theta_{t-1}))$

$\theta_t = \theta_{t-1} + \mu H(z_t, \theta_{t-1})$

PCA.

$R \longrightarrow$

u_0

$u_1 = \frac{R u_0}{\|R u_0\|}$

$u_t = \frac{R u_{t-1}}{\|R u_{t-1}\|} u_t$

$R = \mathbb{E}[z z^T]$ $\quad z_1, z_2 \cdots$

$R_t = \frac{z_1 z_1^T + \cdots + z_t z_t^T}{t}$

$R_t = R_{t-1} + z_t \cdot z_t^T = R_{t-1} + \mu(z_t z_t^T - R_{t-1})$

$u_t = \frac{R_t \cdot u_{t-1}}{\|R_t u_{t-1}\|}$ complexity: k^2

$$\mu \cdot z_t z_t^T + (1-\mu) z_{t-1} z_{t-1}^T + \cdots + (1-\mu)^t z_0 z_0^T$$

evolve by time.

$$u_t = \frac{R u_{t-1}}{\| R u_{t-1} \|}$$

$$v_t = \frac{(I + \mu R) v_{t-1}}{\| (I + \mu R) v_{t-1} \|}$$

u finds eigen larget vector R.

v ---------- $I + \mu R$.

$$\lambda u = R u$$

$$(I + \mu R) u = (1 + \mu\lambda) u.$$

converge to same ~~thing~~.

$$\Omega(\theta, w) \to \Omega(v, w) = \frac{(I + w) v}{\| (I + w) v \|}$$

$$R = \frac{1}{t} \sum z z^T$$

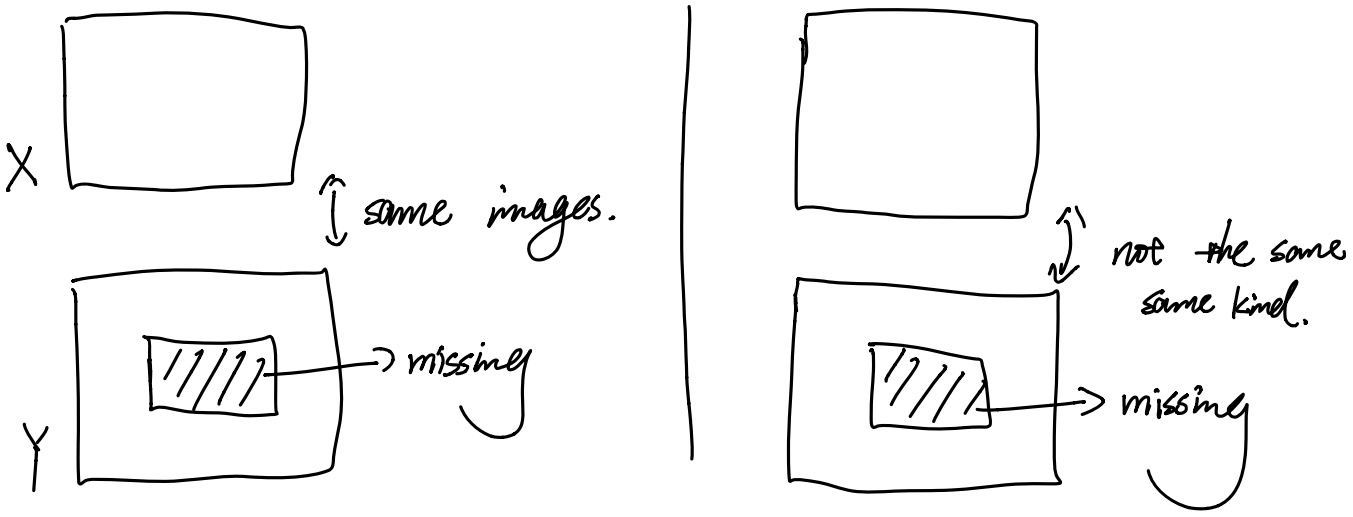
$$v_t = \frac{(I + \mu z_t z_t^T) v_{t-1}}{\| (I + \mu z_t z_t^T) v_{t-1} \|} \quad O(h)$$

find largest eigen vector

$$v_t = \frac{(I - \mu R) v_{t-1}}{\| (I - \mu R) v_{t-1} \|}$$

$$u_t = \frac{R^{-1} u_{t-1}}{\| R^{-1} u_{t-1} \|}$$

) find smallest eigen vector.



$X \xrightarrow{Y=\phi(X)} Y$

(Y, X) $\hat{Y} = \phi(X)$

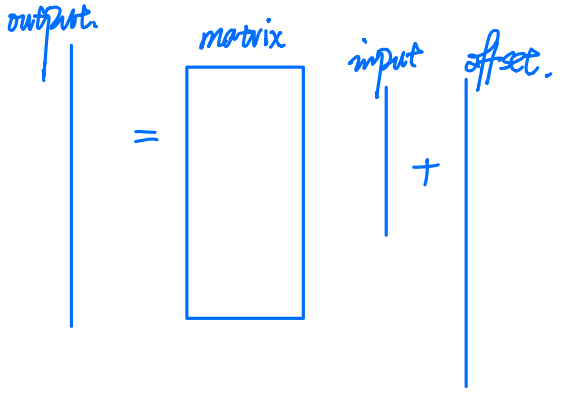
$E[C(Y, \hat{Y})]$ loss cost.

$= E_{[X]}[C(Y, \phi(X)]$

$\min_{\phi(X)} E_{[X]}[C(Y, \phi(X))]$

Don't know distribution we have data.

$\phi(X) = \phi(X, \theta)$ collection of all parameters. output of neural network.



$\min_{\theta} E_{[X]}[C(Y, \phi(X, \theta)]$

$\theta_t = \theta_{t-1} - \mu \nabla_\theta C(Y_t, \phi(X_t, \theta_{t-1}))$

along layers ↓ Back propagations.

Assume it converge to θ_*.

$\hat{Y} = \phi(X, \theta_*)$

last part — lost image

classification.

data. X_1 . X_2 ...

lable y_1 y_2 ...

$\hat{y} = \phi(x)$

$\hat{y}$ ↓ lable. ↑ new x

$\phi(x)$

$$\sum_{i=1}^{K} (1-\phi(x_i))^2 + \sum_{j=1}^{k} (1+\phi(x_j))^2$$

error for x_i error for x_j

$$E_1[(1-\phi(x^1))^2] + E_2[(1+\phi(x^2))^2]$$

$$E_1[(1-\phi(x^1,\theta))^2] + E_2[(1-\phi(x^2,\theta))^2]$$ total average cost

$$\vartheta_t = \theta_{t-1} - \mu\{\nabla\theta(1-\phi(x_t^1,\theta_{t-1}))^2 + \nabla\theta(1-\phi(x_t^2,\theta_{t-1}))^2\}.$$

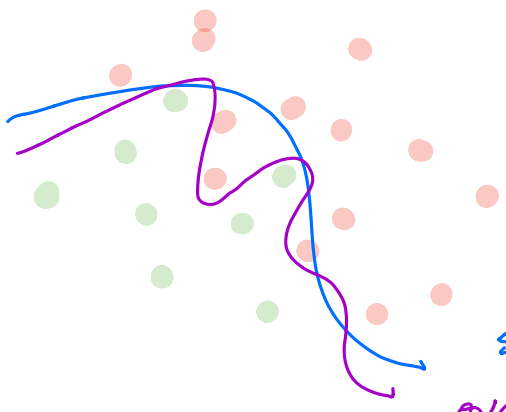
class1 (y_1^1, x_1^1), (y_2^1, x_2^1)

class2 (y_1^2, x_1^2), (y_2^2, x_2^2)

y_1^1, y_2^1: 1 ; y_1^2, y_2^2: -1

$\phi(x, \theta_*)$ ↓ lable ↑ input data

boundary → $\phi(x, \theta_*) = 0$



smooth is better.

overfitting

GANS.

$f(x)$

$h(z) \rightarrow z$

$G(z) = Y$

non-linear transformation.

$x \quad f(x)$

$z \quad h(z)$

$G(z) = Y \rightarrow g(Y)$

random ↓ also random.

scalar function $D(x)$

$\mathbb{E}_x[D(x)] - \mathbb{E}_Y[D(Y)] = 0$

D is arbitrary.

$$\max_{D(x)} \{ \mathbb{E}_x[D(x)] - \mathbb{E}_Y[D(Y)] \}$$

$$= \max_{D(x)} \{ \mathbb{E}_x[D(x)] - \mathbb{E}_z[D(G(z))] \}.$$

$$\min_{G(z)} \{ \max_{D(x)} \{ \mathbb{E}_x[D(x)] - \mathbb{E}_z[D(G(z))] \} \}$$

$$\min_G \max_D \{ \mathbb{E}_x[\phi(D(x))] + \mathbb{E}_z[\psi(D(G(z)))] \}$$

$\phi(z)$ $\psi(z)$

$$\phi(z) = \log(1-z)$$

$$\psi(z) = \log(z)$$

$G(z) \sim G(z, \theta)$

$D(x) \sim D(x, \theta)$

$$v_t = v_{t-1} + \mu \nabla_v X_t z_t \quad v_t \rightarrow v_\infty$$

$$\theta_t = \theta_{t-1} - \mu \nabla_\theta X_t z_t \quad \theta_t \rightarrow \theta_\infty$$