REVIEW

MATRIX CALCULUS

FUNCTIONS	EXAMPLE	VALUE	FIRST DER	SECOND DER
$f: \mathbb{R} \to \mathbb{R}$	x^2	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$
$f: \mathbb{R}^d \to \mathbb{R}$	LOSS	$\mathbb{R}$	$\mathbb{R}^d$ (GRADIENT)	$\mathbb{R}^{d \times d}$ $(\mathbb{S}^d)$ (HESSIAN)
$f: \mathbb{R}^d \to \mathbb{R}^p$	PROJECTION/ NEURALNET - LAYER	$\mathbb{R}^p$	$\mathbb{R}^{d \times p}$ (JACOBIAN)	$\mathbb{R}^{d \times p \times p}$ (HIGHER DIM TENSOR)

GRADIENT

$$\nabla_x f(\bar{x}) = \begin{bmatrix} \frac{\partial f(\bar{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_d} \end{bmatrix}$$

$$\bar{x} = (x_1, \dots, x_d)$$

DIRECTION TO INCREASE THE VALUE OF $f(x)$ THE MOST

$$\nabla_A f(A) = \begin{bmatrix} \frac{\partial f}{\partial a_{11}}(A) & & \\ & \ddots & \\ & & \frac{\partial}{\partial a_{mn}} f(A) \end{bmatrix}$$

GENERALIZATION

HESSIAN

$\nabla_x^2 f(x) \qquad f : \mathbb{R}^d \to \mathbb{R}$

$$\nabla_x^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_d\, x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_d \partial x_d} \end{bmatrix}$$

Ex

$$\nabla_x b^T x = \begin{bmatrix} \vdots \\ \frac{\partial b^T x}{\partial x_i} \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \frac{\partial}{\partial x_i}(b_1 x_1 + b_2 x_2 + \cdots + b_d x_d) \\ \end{bmatrix} = \begin{bmatrix} \vdots \\ b_i \\ \vdots \end{bmatrix} = b$$

PRODUCT RULE (MULTIVARIABLE)

$$\nabla_x \, x^T A x = \nabla_x \, x^T A x + \nabla_x \, x^T A x = A x + A^T x = 2 A x \quad (\text{IF } A = A^T)$$

$$\nabla_A \log \overbrace{|A|}^{\det A} = A^{-1}$$

PROBABILITY THEORY

ELEMENTS

→ SAMPLE SPACE Ω ex $\{HH, HT, TH, TT\}$

→ EVENT $A \subseteq \Omega$ (SAMPLE SET)

→ EVENT SPACE SET OF ALL SUBSET

$$F = \{A_1, \dots, A_n\}$$

→ PROB MEASURE $P: F \to \mathbb{R}$ $(0, 1)$

$$P(A) \geq 0 \quad \forall A \in F$$

$$P(\Omega) = 1$$

IF $A_1, A_2, \ldots$ DISJOINT SET OF EVENTS ($A_i \cap A_j = \emptyset$ WHEN $i \neq j$)
THEN

$$P\left(\bigcup A_i\right) = \sum P(A_i)$$

CONDITIONAL

LET B ANY EVENT SUCH THAT $P(B) \neq 0$

$$P(A|B) := \frac{P(A \cap B)}{P(B)}$$ → INTERSECTION

- $A \perp B$ IF AND ONLY IF $P(A \cap B) = P(A)P(B)$ ($\perp$ INDIPENDENT)
- $A \perp B$ " $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$

THE $P(A)$ DOESN'T CHANGE WHETHER $P(B)$ OCCURS OR NOT

RANDOM VARS

$\omega_0 = HHHTHTTHTT$

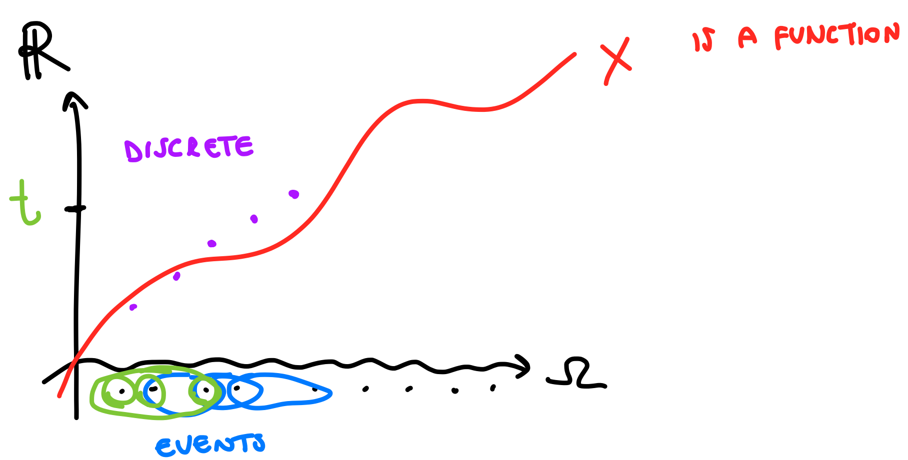
A RV IS $X: \Omega \to \mathbb{R}$ MAPPING FROM OUTCOME TO $\mathbb{R}$

OF HEADS $X(\omega_0) = 5$

OF TOSSES UNTIL TAILS: $X(\omega_0)=4$

$\text{VAL}(X) := X(\Omega)$

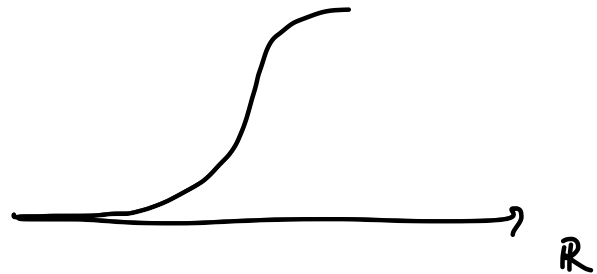
Ex



CUMULATIVE DISTRIBUTION FUNCTION (CDF)

$$F_X(x) = P(X<x)$$

$P[\{\omega : X(\omega)<t\}]$ MEASURE PROB. ON ω



DISCRETE RV: VAL(X) COUNTABLE

$$P(X=k) := P(\{\omega \mid X(\omega)=k\})$$

PROBABILITY MASS FUNCTION

$$p_X : Val(X) \to [0,1]$$

$$p_X(x) := P(X=x)$$

$$\sum_{x \in Val(X)} p_X(x) = 1$$

CONTINUOUS RV: VAL(X) UNCOUNTAPLE

$$P(a \le X \le b) := P(\{\omega \mid a \le X(\omega) \le b\})$$

PROBABILITY DENSITY FUNCTION

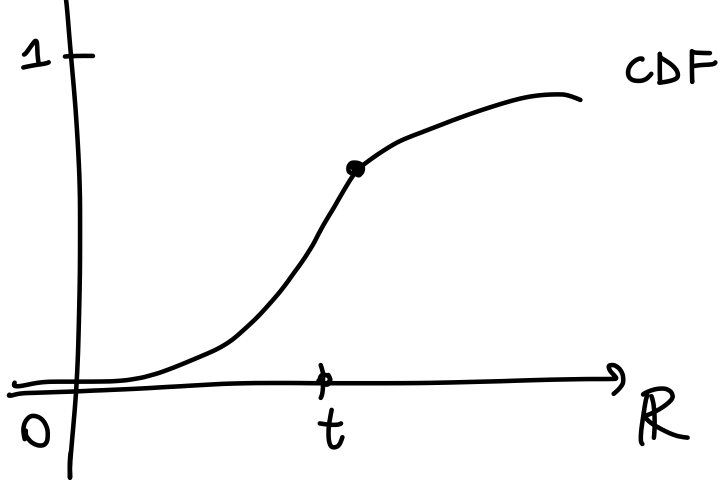
$$f_X : \mathbb{R} \to \mathbb{R}$$

$$f_X(x) := \frac{d}{dx} F_X(x)$$

$$f_X(x) \neq P(X=x)$$

$$\int_{-\infty}^{+\infty} \underbrace{f_X(x)\,dx}_{P(x \le X \le x+dx)} = 1$$

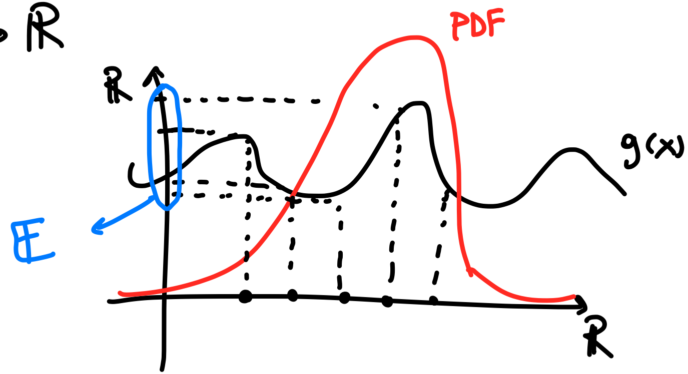
PDF



PMF

EXPECTED VALUE AND VARIANCE

$g : \mathbb{R} \to \mathbb{R}$



$\mathbb{E}[g(X)]$

LET X BE A DISCRETE RV WITH PMF p_X

$$\mathbb{E}[g(X)] := \sum_{x \in \mathrm{Val}(X)} g(x)\, p_X(x)$$

IF CONTINUOUS (PDF)

$$\mathbb{E}[g(x)] := \int_{-\infty}^{+\infty} g(x) f_X(x)\, dx$$

SAMPLES

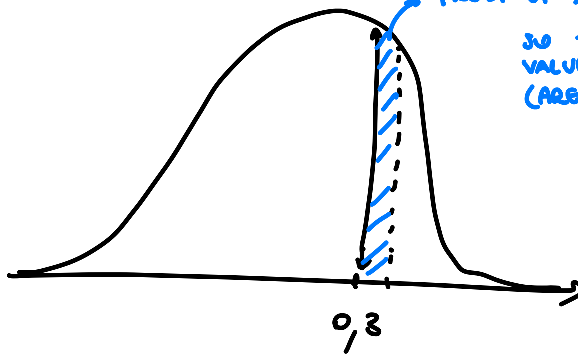
$$\lim_{N\to+\infty} \underbrace{\frac{1}{N}\sum_{i=1}^{N} g(x^{(i)})}_{\text{MONTE CARLO ESTIMATE}} \to \int_{-\infty}^{\infty} g(x)\,p(x)\,dx$$

RN

DENSITY

SOME FUNCTION OF X

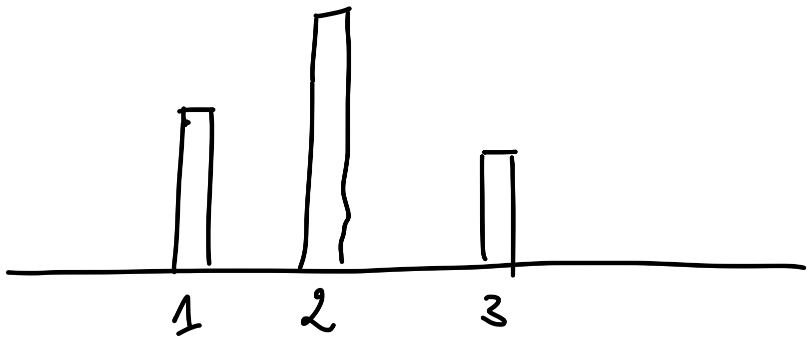
LAW OF LARGE NUMBERS
(TRUE EXPECTATIONS)



PROB. OF RN

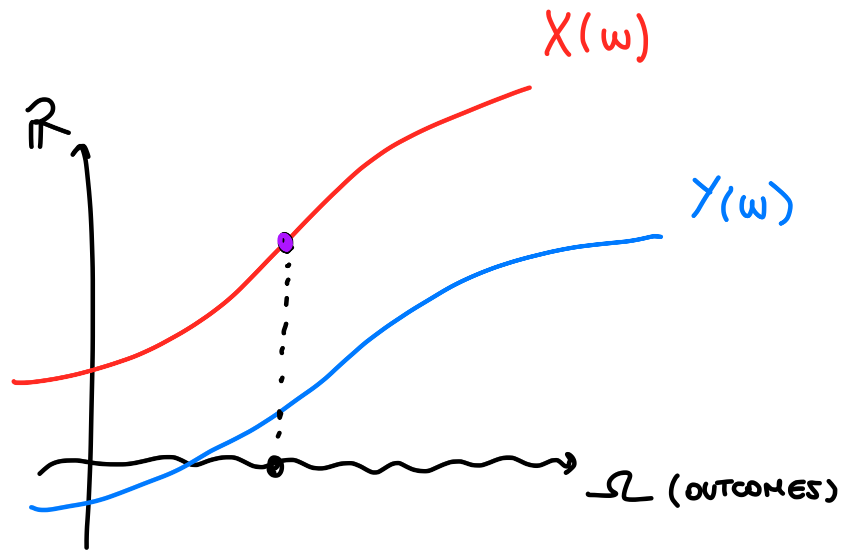
SO THE PROB. OF A SINGLE VALUE RN IS 0
(AREA IS NULL)

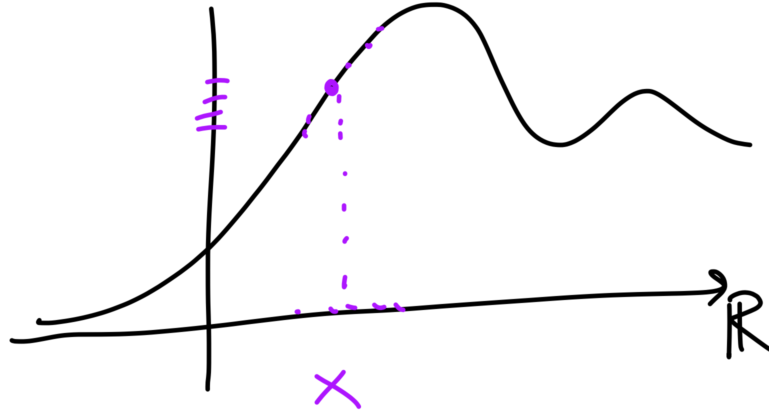
$P(x = 0.3) \neq$



OUTCOMES

EVENTS $\leftarrow$ PROB

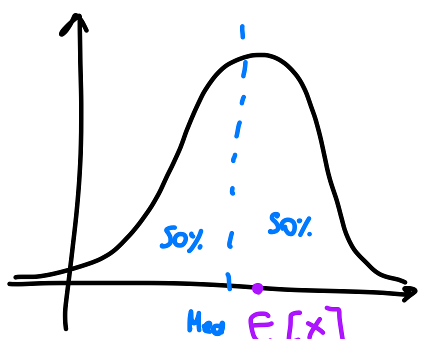




DEFINED ON TOP OF $X(\omega), Y(\omega)$

VARIANCE

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



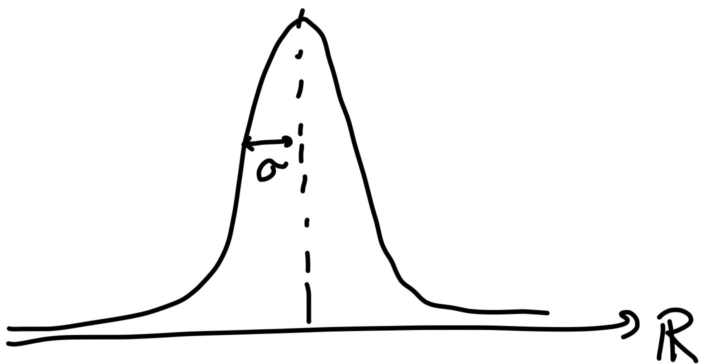
IF SIMMETRICAL (THE DISTRIBUTION) THEN $\mathrm{Med} = E[X]$

DISTRIBUTION EX

DISTRIBUTION	PDF OR PMF	MEAN	VARIANCE
BERNOULLI (p)	$\begin{cases} p & \text{if } x=1 \\ 1-p & \text{if } x=0 \end{cases}$	p	$p(1-p)$
BINOMIAL (n,p)	$\binom{n}{k} p^k (1-p)^{n-k}$ for $k=0,1,\dots,n$	np	$np(1-p)$
GEOMETRIC (p)	$p(1-p)^{k-1}$ for $k=1,\dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
POISSON (λ)	$\frac{e^{-\lambda}\lambda^k}{k!}$ for $k=0,1,\dots$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
GAUSSIAN (μ,σ^2)	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for all $x \in (-\infty, \infty)$	μ	σ^2
EXPONENTIAL (λ)	$\lambda e^{-\lambda x}$ FOR ALL $x \geq 0, \lambda \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

SPACE OVER THE DEFINED DISTRIBUTION : $\mathbb{R}, \mathbb{R}^2, \dots$

PARAMETER , SHAPE OF THE DISTRIBUTION



WE CAN HAVE 2 RN

<u>CDF</u> $F_{XY}(x,y) = P(X \leq x, Y \leq y)$

<u>PMF</u> $p_{XY}(x,y) = p(X=x, Y=y)$

<u>MARGINAL PMF</u> $p_X(x) = \sum_y p_{XY}(x,y)$

<u>BIVARIATE PDF</u> $f_{XY}(x,y) = \dfrac{\partial^2 F_{XY}(x,y)}{\partial x \partial y}$

<u>MARGINAL PDF</u> $f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy$

FOR INSTANCE, THE JOINT PDF, PMF

	DISC	CONT
$p(x,y)$		
$p(x) =$	$\sum_y p(x,y)$	$= \int p(x,y)\,dy$
$p(y) =$	$\sum_x p(x,y)$	$= \int p(x,y)\,dx$

BAYES' THEOREM

GIVEN THE CONDITIONAL PROBABILITY OF AN EVENT $P(x|y)$

WANT TO FIND THE "REVERSE" CONDITIONAL PROBABILITY, $P(y|x)$

$$P(y|x) = \frac{P(x|y)\,P(y)}{P(x)} \qquad \left(P(x|y) = \frac{P(x \cap y)}{P(y)} \right)$$

WHERE $p(x) = \sum_{y' \in \text{value } y} P(x|y')\,P(y')$

X AND Y ARE CONTINUOUS

$$f(y|x) = \frac{f(x|y)\,f(y)}{f(x)}$$

WHERE $f(x) = \int_{y' \in \text{value } y} f(x|y') f(y') dy'$

$$P(x, y) = P(x) \, P(y|x)$$

JOINT MARGINED CONDITIONAL

CHAIN RULE

$$= P(y) \, P(x|y)$$

BAYES

$$P(y|x) = \frac{P(y) \, P(x|y)}{P(x)} = \frac{P(y) P(x|y)}{\sum_{y'} P(y') P(x|y')}$$

JOINT $\rightarrow \sum p(x, y')$

INDIPENDENCE (OF 2 RN)

X, Y ARE INDIP.

$$p_{XY}(x,y) = p_X(x)\, p_Y(y)$$

$$p_{Y|X}(x,y) = p_Y(y)$$

FOR CONT. RN

$$p_{XY}(x,y) \to f_{XY}(x,y)$$

EX

TOSS A COIN AND SPIN 1-7 NUMBERS. PROB OF GETTING ODD NUMBER AND A TAIL?

$$p_{XY}(x,y) = p_X(x)\, p_Y(y) = \frac{1}{2} \cdot \frac{4}{7} = \frac{2}{7}$$

Ex

X, Y : 2 CONT. RN

g, $\mathbb{R}^2 \to \mathbb{R}$: A FUNC. OF Y AND X

$$\mathbb{E}(g(x,y)) = \int_{x \in VAL(X)} \int_{y \in VAL(Y)} g(x,y) f_{XY}(x,y)\, dx\, dy$$

Ex

$g(x,y) = 3x$

$f_{x,y} = 4xy \qquad 0 < x < 1,\ 0 < y < 1$

$$\mathbb{E}(g(x,y)) = \int_0^1 \int_0^1 12x^2y\, dx\, dy$$

COVARIANCE

$$\mathrm{VAR}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

$$\mathrm{COV}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

$$\mathrm{COV}(X,X) = \mathrm{VAR}[X]$$

MULTIVARIATE GAUSSIAN

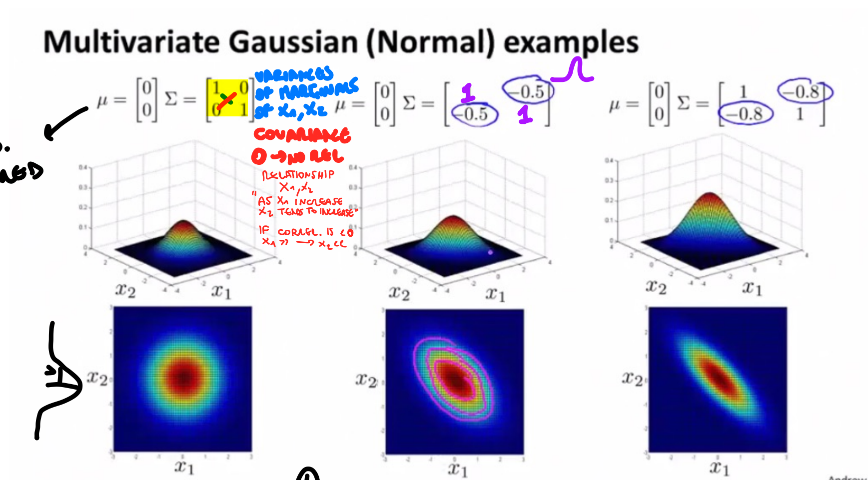
$X \in \mathbb{R}^n$ OBSERVATION

COVARIANCE MAT → - INVERTIBLE
- POSITIVE SEMIDEFINITE

MEAN: μ

$$p(x;\underbrace{\mu,\Sigma}_{\text{PARAMS}}) = \frac{1}{\underbrace{(2\pi)^{n/2}}_{\text{NORMALIZING CONSTANT}}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}\underbrace{(x-\mu)^T\Sigma^{-1}(x-\mu)}_{\text{QUADRATIC FORM}}\right]$$

$$\iint p(x;\mu,\Sigma)\,dx_1\,dx_2$$



VARIANCES OF MARGINALS OF x_1, x_2

COVARIANCE 0 → NO REL

RELATIONSHIP x_1, x_2

"AS x_1 INCREASE x_2 TENDS TO INCREASE"

IF CORREL. IS < 0 $x_1 \nearrow \rightarrow x_2 \searrow$

THE D. IS CENTERED IN $\vec{0}$

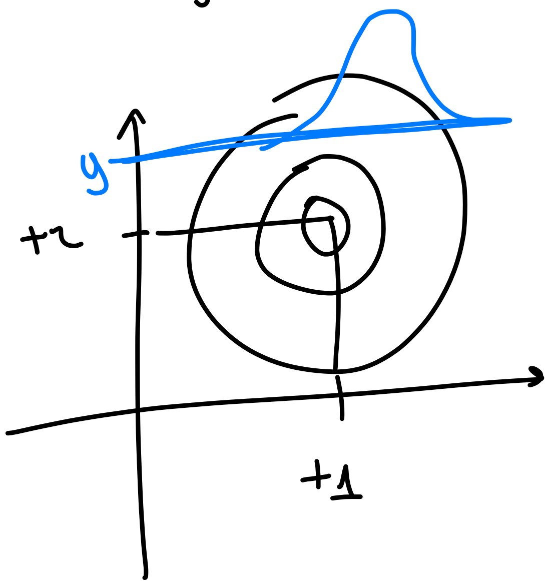
ST. DEV OF 1

CONDITIONAL EXPECTATION

$\mathbb{E}[X]$ — CONSTANT

$\mathbb{E}[X|Y]$ — RV

$\mathbb{E}[X|Y=y]$ — FUNCTION OF y



$\mathbb{E}[X|Y=y]$

LAW OF TOTAL $\mathbb{E}$

$$\mathbb{E}[X] = \mathbb{E}\left[\mathbb{E}[X|Y]\right]$$ TRUE FOR ANY X AND ANY Y

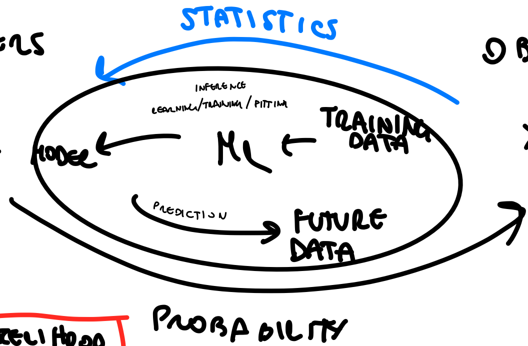
$$P(a|b) = \frac{P(b|a)\,P(a)}{P(b)}$$

$$P(a|b,c) = \frac{P(b|a,c)\,P(a|c)}{P(b|c)}$$

STATISTICS

PARAMETERS

μ, Σ



OBSERVATION (DATA)

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

- METHOD OF MOMENTS
- MAXIMUM LIKELIHOOD ESTIMATION

TRAINING DATA

(x, y)

MAX LIKELIHOOD ESTIMATION (MLE)

EX GAUSSIAN DATA

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

$X = x^{(i)}, \dots, x^{(n)}, \quad x^{(i)} := i^{th}$ example $\in \mathbb{R}^d$, SAMPLE INDIPENDENTLY AND IDENTICALLY DISTRIBUTED

I.I.D.

VAL OF FUNCTION IS DEF. — PARAMS

$$p(x;\underbrace{\mu,\Sigma}) = \frac{1}{\sqrt{2\pi}\,|\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]$$ PROB. DENSITY

$$p(x^{(1)},\dots,x^{(n)}) = \prod_{i=1}^{n} p(x^{(i)})$$ JOINT PROBABILITY

LET'S GENERALIZE

$$p(x^{(1)},\dots,x^{(n)};\mu,\Sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,|\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x^{(i)}-\mu)^T \Sigma^{-1}(x^{(i)}-\mu)\right]$$

DEFINING LIKELIHOOD FUNCTION

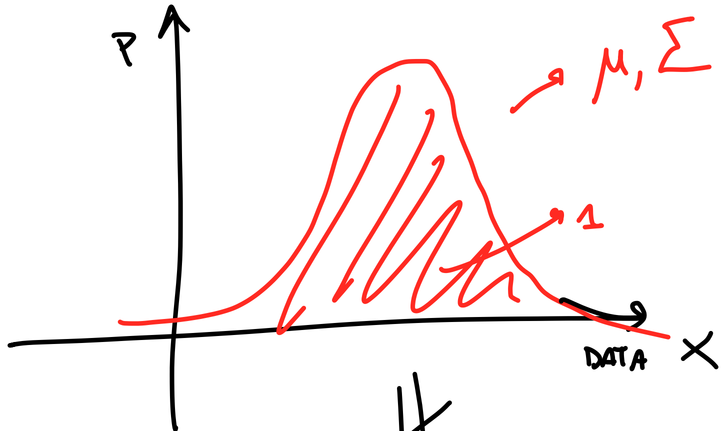
$$L(\underbrace{\mu,\Sigma}_{\text{NOW THE PARAMS ARE THE VARS}};\underbrace{x^{(1)},\dots,x^{(n)}}_{\text{PARAMETERS}}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,|\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x^{(i)}-\mu)^T \Sigma^{-1}(x^{(i)}-\mu)\right]$$

SO IT IS A FUNCTION OF μ, Σ

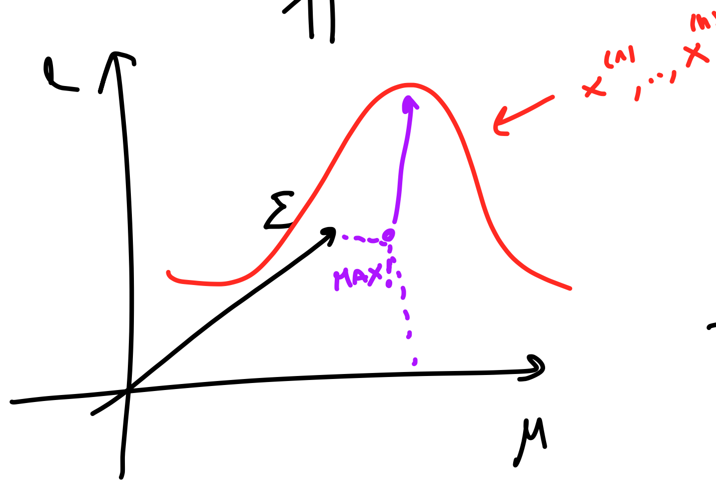
PROB. OF THE DATA GIVEN PARAMS
LIKEL. OF THE PARAMS GIVEN DATA

$$\int p(\quad)\,dx = 1$$

$$\int L(\quad)\,dx = ?$$ MAY NOT EVEN EXIST!



$\neq$



$$P(X)$$

$$L(\mu, \Sigma)$$

THEN $\ddot{\mu}$ AND $\ddot{\Sigma}$ USED FOR PREDICTIONS

$$L(\theta; X) = \prod_{i=1}^{n} L(\theta; x^{(i)})$$

ESTIMATED (OUTPUT OF ESTIM. PROCESS)

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \prod_{i=1}^{n} L(\theta; x^{(i)})$$

MONOTONIC INC. FUNC. (SAME θ)

$$= \arg\max_{\theta} \log \prod_{i=1}^{n} L(\theta; x^{(i)})$$

STD RECIPE :) $\Rightarrow$

$$= \arg\max_{\theta} \sum_{i=1}^{n} \ell(\theta; x^{(i)}) \qquad \ell(\theta) = \log L(\theta)$$

NOW LET'S ASSUME M-GAUSSIAN ℓ

$$\hat{\mu}, \hat{\Sigma} = \arg\max_{\mu, \Sigma} \sum_{i=1}^{n} \log \left[\frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right] \right]$$

$$= \arg\max_{\mu, \Sigma} \left[\sum_{i=1}^{n} \left\{ K - \frac{1}{2} \log|\Sigma| - \frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right] \right]$$

LET'S MAX TO BOTH VARS

μ

$$\nabla_\mu \sum_{i=1}^{n} K - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x^{(i)}-\mu)^T \Sigma^{-1}(x^{(i)}-\mu)$$

$$\nabla_x x^T A x = 2Ax$$

Σ IS SIMM $\Sigma = \Sigma^T$

$$\sum_{i=1}^{n} -\frac{1}{2}\left[\underbrace{x^{(i)T}\Sigma^{-1}x^{(i)}}_{0} - x^{(i)T}\Sigma^{-1}\mu - \mu^T\Sigma^{-1}x^{(i)} + \mu^T\Sigma^{-1}\mu \right]$$

$$= \sum_{i=1}^{n} -\frac{1}{2}\left[-x^{(i)}\Sigma^{-1} - \Sigma^{-1}x^{(i)} + 2\Sigma^{-1}\mu \right]$$

$$= \sum_{i}^{n} \left[x^{(i)}\Sigma^{-1} - \Sigma^{-1}\mu \right] = 0$$

$$\underbrace{n}_{\text{SUMMATION}}\Sigma^{-1}\mu = \sum_{i=1}^{n}\Sigma^{-1}x^{(i)}$$

$$\cancel{\Sigma^{-1}}\mu = \cancel{\Sigma^{-1}}\left[\frac{1}{n}\sum_{i=1}^{n} x^{(i)}\right]$$

$$\Rightarrow \boxed{\mu = \frac{1}{2}\sum_{i=1}^{n} x^{(i)}}$$

Σ

$$\frac{d}{d\Sigma}\Sigma^{-1} ?$$

$$S = \Sigma^{-1} \qquad \nabla_S \mathcal{L} = 0 \iff \nabla_{\Sigma^{-1}}\mathcal{L} = 0$$

SOLVE FOR S, INVERT AND FIND Σ

$$\frac{d}{dS} \to S^{-1}$$

$$\nabla_S \sum_{i=1}^{n} \frac{1}{2}\log|S| - \frac{1}{2}(x^{(i)}-\mu)^T S (x^{(i)}-\mu)$$

$$= -\frac{1}{2}\left[nS^{-1} - \sum_{i=1}^{n}\left[(x^{(i)}-\mu)(x^{(i)}-\mu)^T\right]\right] = 0 \quad \ldots$$

$$\nabla_A x^T A x = ? \quad [xx^T]$$

$$\nabla_A A = \begin{bmatrix}\frac{d}{da_{11}}A & & \\ & \ddots & \\ & & \frac{d}{da_{nm}}A\end{bmatrix} \qquad x^T A = (x_1 a_{11} + x_2 a_{21} + \cdots + x_n a_{n1}, \ldots, x_1 a_{1m} + \cdots + x_n a_{nm})^T$$

$$x^T A x = \left(x_1^T x_1 a_{11} + \cdots + x_n^T x_n a_{n1}, \ldots, x_1^T x_n a_{1m} + \cdots + x_n^T x_n a_{nm} \right)^T$$

$$\nabla_A x^T A x = \nabla_A \left(x_1^T x_1 a_{11} + \cdots + x_n^T x_n a_{n1}, \ldots, x_1^T x_n a_{1m} + \cdots + x_n^T x_n a_{nm} \right)^T$$

$$\nabla_A x^T A x = x^T x$$

$$\ldots \; S^{-1} = \frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - \mu)(x^{(i)} - \mu)^T = \Sigma$$

$$\approx \mathbb{E}\left((x - E[x]\right)^2$$

DONE!