

## Lecture 6: Probability and Distributions

Yi, Yung (이윤)

Mathematics for Machine Learning

<https://yung-web.github.io/home/courses/mathml.html>

KAIST EE

April 7, 2021

- (1) Construction of a Probability Space
- (2) Discrete and Continuous Probabilities
- (3) Sum Rule, Product Rule, and Bayes' Theorem
- (4) Summary Statistics and Independence
- (5) Gaussian Distribution
- (6) Conjugacy and the Exponential Family
- (7) Change of Variables/Inverse Transform

- (1) Construction of a Probability Space
- (2) Discrete and Continuous Probabilities
- (3) Sum Rule, Product Rule, and Bayes' Theorem
- (4) Summary Statistics and Independence
- (5) Gaussian Distribution
- (6) Conjugacy and the Exponential Family
- (7) Change of Variables/Inverse Transform

**Modeling:** Approximate reality with a simple (mathematical) model

- Experiment
  - Flip two coins
- Observation: a random outcome
  - for example,  $(H, H)$
- All outcomes
  - $\{(H, H), (H, T), (T, H), (T, T)\}$

- 
- **Our goal:** Build up a **probabilistic model** for an experiment with random outcomes
  - **Probabilistic model?**
    - Assign a number to each outcome or a set of outcomes
    - Mathematical description of an uncertain situation
  - Which model is good or bad?

**Goal:** Build up a probabilistic model. Hmm... How?

The first thing: What are the *elements* of a probabilistic model?

## Elements of Probabilistic Model

1. All outcomes of my interest: Sample Space  $\Omega$
2. Assigned numbers to each outcome of  $\Omega$ : Probability Law  $\mathbb{P}(\cdot)$

**Question:** What are the conditions of  $\Omega$  and  $\mathbb{P}(\cdot)$  under which their induced probability model becomes "legitimate"?

# Sample Space $\Omega$

The set of all outcomes of my interest

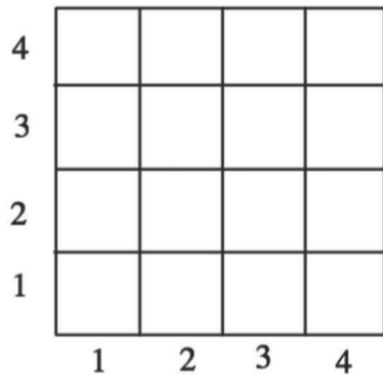
1. Mutually exclusive
2. Collectively exhaustive
3. At the right granularity (not too concrete, not too abstract)

1. Toss a coin. What about this?  
 $\Omega = \{H, T, HT\}$
2. Toss a coin. What about this?  $\Omega = \{H\}$
3. (a) Just figuring out prob. of H or T.  
 $\implies \Omega = \{H, T\}$   
  
(b) The impact of the weather (rain or no rain) on the coin's behavior.  
  
 $\implies \Omega = \{(H, R), (T, R), (H, NR), (T, NR)\}$ ,  
where R(Rain), NR(No Rain).

## Examples: Sample Space $\Omega$

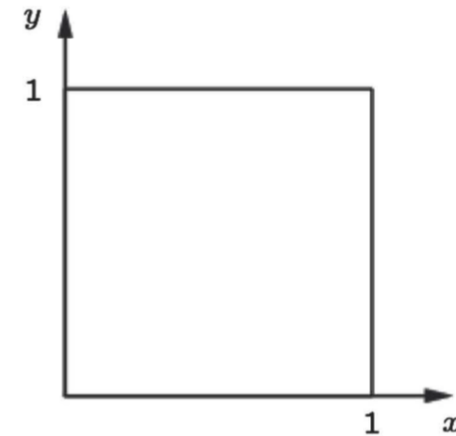
- *Discrete case:* Two rolls of a tetrahedral die

-  $\Omega = \{(1, 1), (1, 2), \dots, (4, 4)\}$



- *Continuous case:* Dropping a needle in a plain

-  $\Omega = \{(x, y) \in \mathbb{R}^2 \mid 0 \leq x, y \leq 1\}$

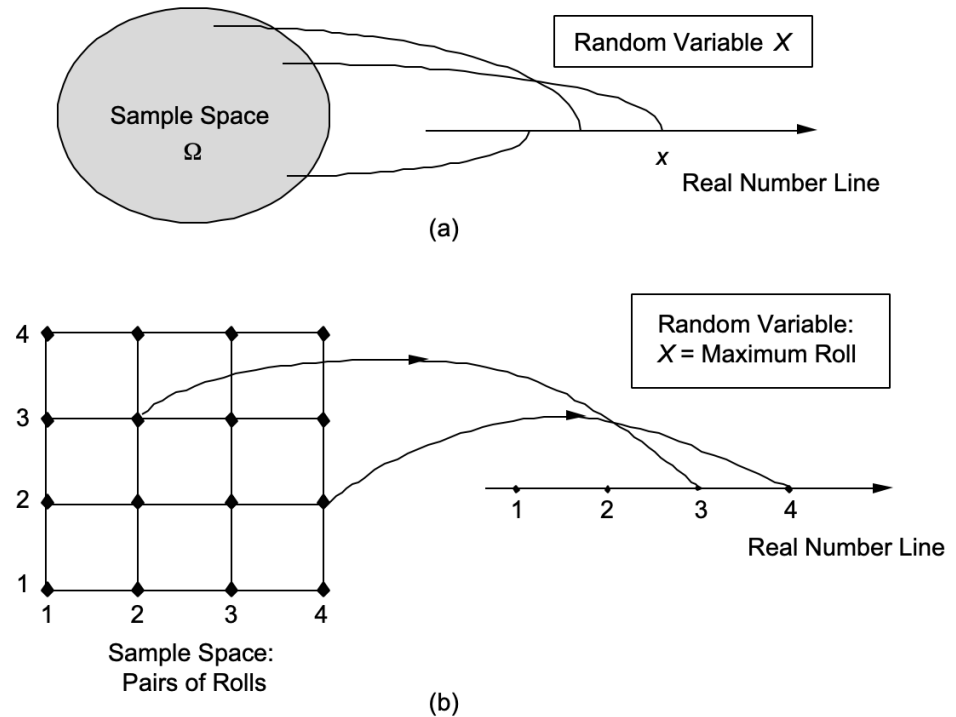


- Assign numbers to what? Each outcome?
- What is the probability of dropping a needle at  $(0.5, 0.5)$  over the  $1 \times 1$  plane?
- Assign numbers to each **subset** of  $\Omega$ : A subset of  $\Omega$ : **an event**
- $\mathbb{P}(A)$ : Probability of an event  $A$ .
  - This is where probability meets set theory.
  - Roll a dice. What is the probability of odd numbers?  
 $\mathbb{P}(\{1, 3, 5\})$ , where  $\{1, 3, 5\} \subset \Omega$  is an event.
- **Event space  $\mathcal{A}$** : The collection of subsets of  $\Omega$ . For example, in the discrete case, the power set of  $\Omega$ .
- **Probability Space  $(\Omega, \mathcal{A}, \mathbb{P}(\cdot))$**



# Random Variable: Idea

- In reality, many outcomes are **numerical**, e.g., stock price.
- Even if not, very convenient if we map numerical values to random outcomes, e.g., '0' for male and '1' for female.



- Mathematically, a random variable  $X$  is a function which maps from  $\Omega$  to  $\mathbb{R}$ .
- **Notation.** Random variable  $X$ , numerical value  $x$ .
- Different random variables  $X$ ,  $Y$ , etc can be defined on the same sample space.
- For a fixed value  $x$ , we can associate an **event** that a random variable  $X$  has the value  $x$ , i.e.,  $\{\omega \in \Omega \mid X(\omega) = x\}$
- Generally,

$$\mathbb{P}_X(S) = \mathbb{P}(X \in S) = \mathbb{P}(X^{-1}(S)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in S\})$$

- Pick a person  $a$  at random
  - event  $A$ :  $a$ 's age  $\leq 20$
  - event  $B$ :  $a$  is married
- (Q1) What is the probability of  $A$ ?
- (Q2) What is the probability of  $A$ , given that  $B$  is true?
- Clearly the above two should be different.
- **Question.** How should I change my belief, given some additional information?
- Need to build up a new theory, which we call **conditional probability**.

- $\mathbb{P}(A \mid B)$ :  $\mathbb{P}(\cdot \mid B)$  should be a new **probability law**.

- **Definition.**

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad \text{for } \mathbb{P}(B) > 0.$$

- Note that this is a **definition**, not a **theorem**.
- All other properties of the law  $\mathbb{P}(\cdot)$  is applied to the conditional law  $\mathbb{P}(\cdot \mid B)$ .
- For example, for two disjoint events  $A$  and  $C$ ,

$$\mathbb{P}(A \cup C \mid B) = \mathbb{P}(A \mid B) + \mathbb{P}(C \mid B)$$

- (1) Construction of a Probability Space
- (2) Discrete and Continuous Probabilities
- (3) Sum Rule, Product Rule, and Bayes' Theorem
- (4) Summary Statistics and Independence
- (5) Gaussian Distribution
- (6) Conjugacy and the Exponential Family
- (7) Change of Variables/Inverse Transform

- The values that a random variable  $X$  takes is discrete (i.e., finite or countably infinite).
- Then,  $p_X(x) := \mathbb{P}(X = x) := \mathbb{P}(\{\omega \in \Omega \mid X(\omega) = x\})$ , which we call **probability mass function** (PMF).
- Examples: Bernoulli, Uniform, Binomial, Poisson, Geometric

## Bernoulli $X$ with parameter $p \in [0, 1]$

- Only **binary** values

$$X = \begin{cases} 0, & \text{w.p.}^1 \quad 1 - p, \\ 1, & \text{w.p.} \quad p \end{cases}$$

In other words,  $p_X(0) = 1 - p$  and  $p_X(1) = p$  from our PMF notation.

- Models a trial that results in binary results, e.g., success/failure, head/tail
- Very useful for an **indicator rv** of an event  $A$ . Define a rv  $1_A$  as:

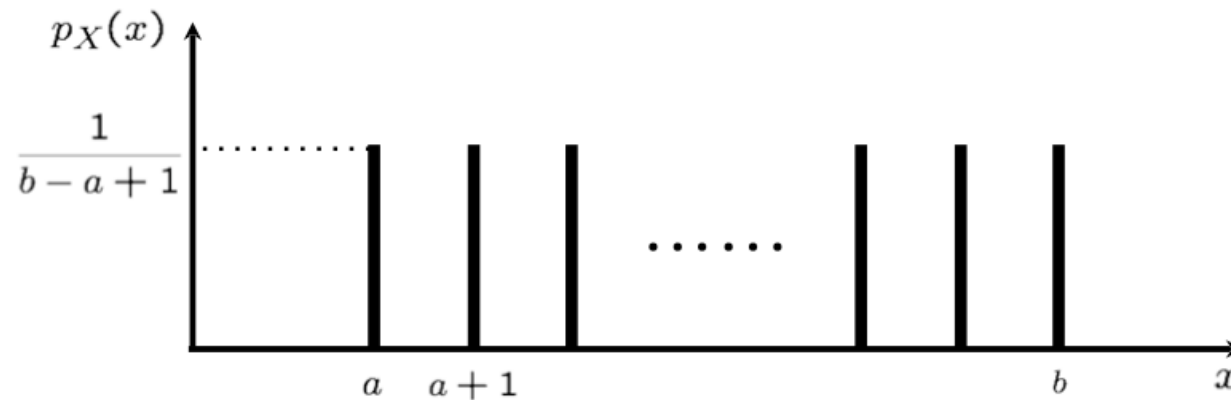
$$1_A = \begin{cases} 1, & \text{if } A \text{ occurs,} \\ 0, & \text{otherwise} \end{cases}$$

---

<sup>1</sup>with probability

## Uniform $X$ with parameter $a, b$

- integers  $a, b$ , where  $a \leq b$
- Choose a number of  $\Omega = \{a, a + 1, \dots, b\}$  uniformly at random.
- $p_X(i) = \frac{1}{b-a+1}$ ,  $i \in \Omega$ .



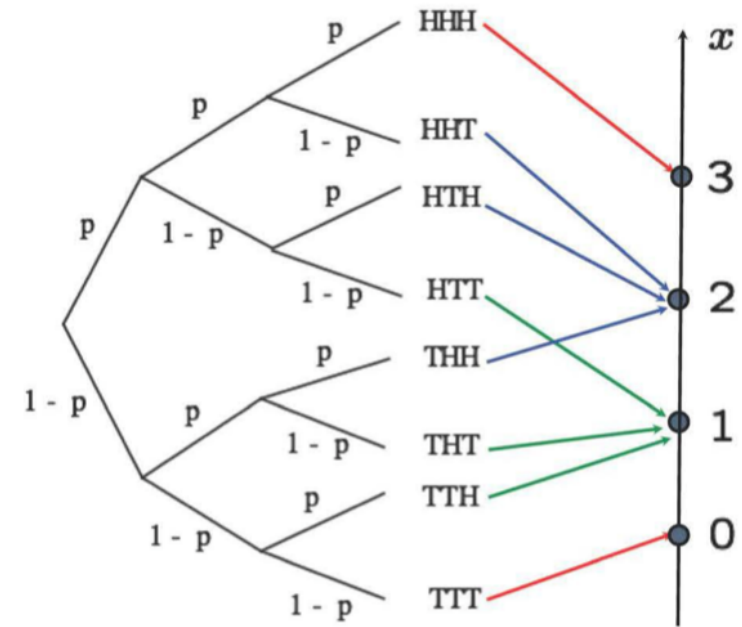
- Models complete ignorance (I don't know anything about  $X$ )



## Binomial $X$ with parameter $n, p$

- Models the number of successes in a given number of independent trials
- $n$  independent trials, where one trial has the success probability  $p$ .

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$



- *Binomial*( $n, p$ ): Models the number of successes in a given number of independent trials with success probability  $p$ .
- Very large  $n$  and very small  $p$ , such that  $np = \lambda$

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, \dots$$

- Is this a legitimate PMF?

$$\sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \left( 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} \dots \right) = e^{-\lambda} e^{\lambda} = 1$$

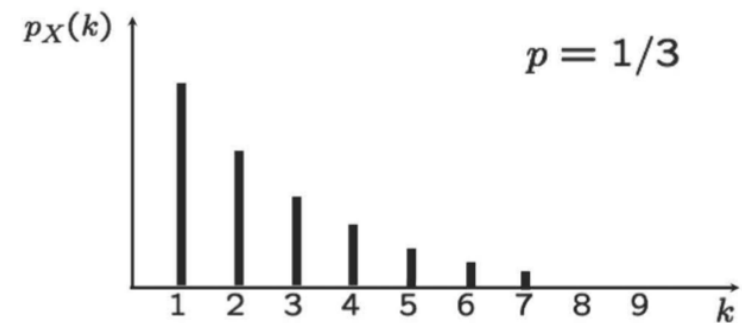
- Prove this:

$$\lim_{n \rightarrow \infty} p_X(k) = \binom{n}{k} (1/n)^k (1 - 1/n)^{n-k} = e^{-\lambda} \frac{\lambda^k}{k!}$$

## Geometric $X$ with parameter $p$

- Experiment: infinitely many independent Bernoulli trials, where each trial has success probability  $p$
- Random variable: number of trials until the **first success**.
- Models waiting times until something happens.

$$p_X(k) = (1 - p)^{k-1}p$$



- **Joint PMF.** For two random variables  $X, Y$ , consider two events  $\{X = x\}$  and  $\{Y = y\}$ , and

$$p_{X,Y}(x,y) := \mathbb{P}(\{X = x\} \cap \{Y = y\})$$

- $\sum_x \sum_y p_{X,Y}(x,y) = 1$

- **Marginal PMF.**

$$p_X(x) = \sum_y p_{X,Y}(x,y),$$

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

Example.

					y
4	1/20	2/20	2/20		
3	2/20	4/20	1/20	2/20	
2		1/20	3/20	1/20	
1		1/20			
	1	2	3	4	x

$$p_{X,Y}(1,3) = 2/20$$

$$p_X(4) = 2/20 + 1/20 = 3/20$$

$$\mathbb{P}(X = Y) = 1/20 + 4/20 + 3/20 = 8/20$$

- Conditional PMF

$$p_{X|Y}(x|y) := \mathbb{P}(X = x|Y = y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

for  $y$  such that  $p_Y(y) > 0$ .

- $\sum_x p_{X|Y}(x|y) = 1$

- Multiplication rule.

$$\begin{aligned} p_{X,Y}(x,y) &= p_Y(y)p_{X|Y}(x|y) \\ &= p_X(x)p_{Y|X}(y|x) \end{aligned}$$

- $p_{X,Y,Z}(x,y,z) = p_X(x)p_{Y|X}(y|x)p_{Z|X,Y}(z|x,y)$

y				
4	1/20	2/20	2/20	
3	2/20	4/20	1/20	2/20
2		1/20	3/20	1/20
1		1/20		
	1	2	3	4
	x			

$$p_{X|Y}(2|2) = \frac{1}{1+3+1}$$

$$p_{X|Y}(3|2) = \frac{3}{1+3+1}$$

$$\mathbb{E}[X|Y = 3] = 1(2/9) + 2(4/9) + 3(1/9) + 4(2/9)$$

# Continuous RV and Probability Density Function (PDF)

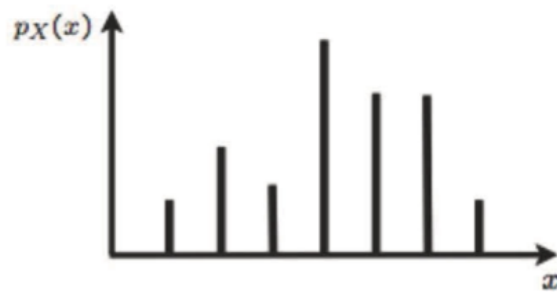
- Many cases when random variable have “continuous values”, e.g., velocity of a car

## Continuous Random Variable

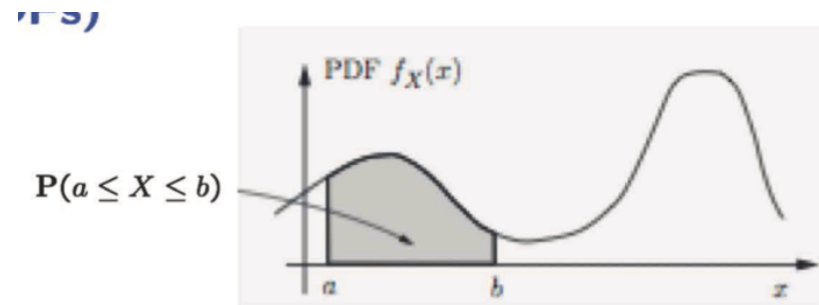
A rv  $X$  is **continuous** if  $\exists$  a function  $f_X$ , called **probability density function (PDF)**, s.t.

$$\mathbb{P}(X \in B) = \int_B f_X(x) dx$$

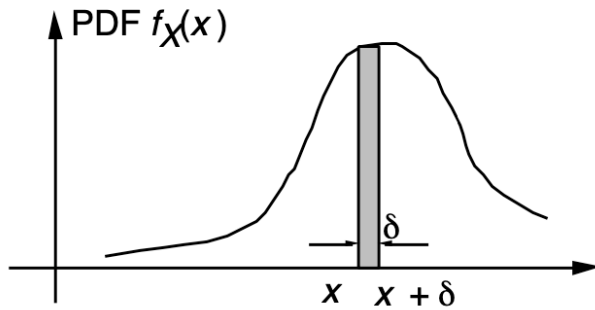
- All of the concepts and methods (expectation, PMFs, and conditioning) for discrete rvs have continuous counterparts



- $\mathbb{P}(a \leq X \leq b) = \sum_{x: a \leq x \leq b} p_X(x)$
- $p_X(x) \geq 0, \sum_x p_X(x) = 1$

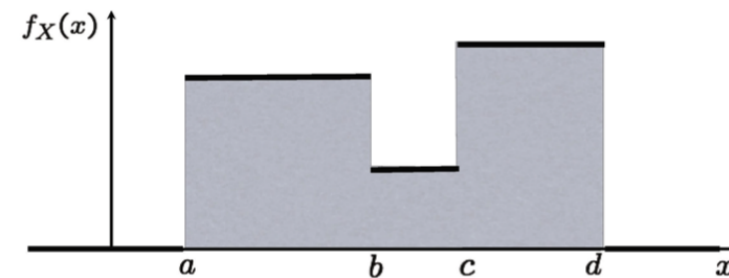
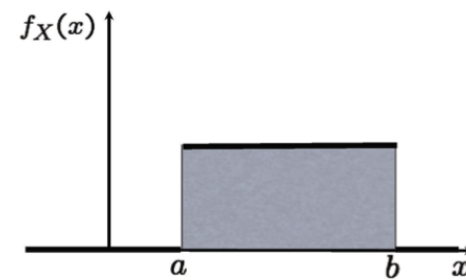


- $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x) dx$
- $f_X(x) \geq 0, \int_{-\infty}^{\infty} f_X(x) dx = 1$



- $\mathbb{P}(a \leq X \leq a + \delta) \approx f_X(a) \cdot \delta$
- $\mathbb{P}(X = a) = 0$

## Examples



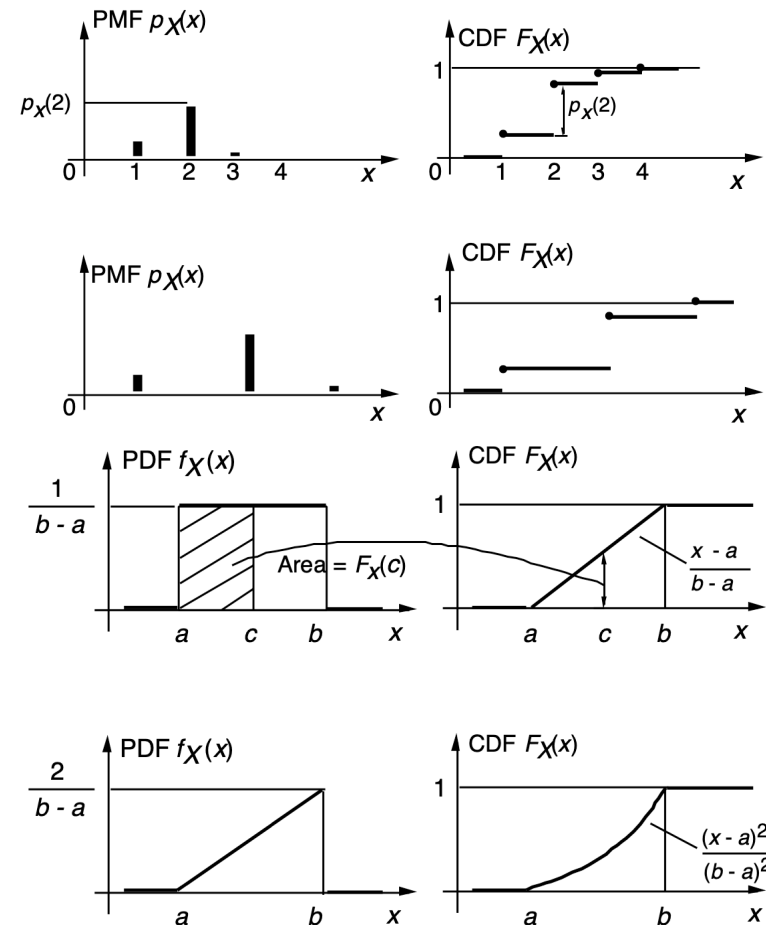
# Cumulative Distribution Function (CDF)

- Discrete: PMF, Continuous: PDF
- Can we describe all rvs with a single mathematical concept?

$$F_X(x) = \mathbb{P}(X \leq x) =$$

$$\begin{cases} \sum_{k \leq x} p_X(k), & \text{discrete} \\ \int_{-\infty}^x f_X(t) dt, & \text{continuous} \end{cases}$$

- always well defined, because we can always compute the probability for the event  $\{X \leq x\}$
- CCDF (Complementary CDF):  $\mathbb{P}(X > x)$





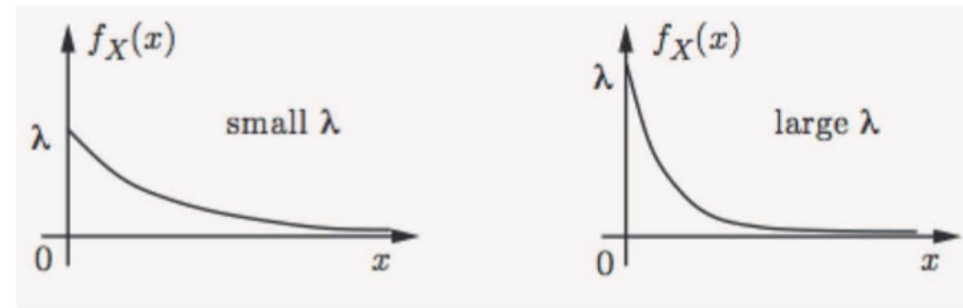
- Non-decreasing
- $F_X(x)$  tends to 1, as  $x \rightarrow \infty$
- $F_X(x)$  tends to 0, as  $x \rightarrow -\infty$

## Exponential RV with parameter $\lambda > 0$ : $\exp(\lambda)$

- A rv  $X$  is called **exponential with  $\lambda$** , if

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad \text{or} \quad F_X(x) = 1 - e^{-\lambda x}$$

- Models a waiting time
- CCDF  $\mathbb{P}(X \geq x) = e^{-\lambda x}$  (waiting time decays exponentially)
- $\mathbb{E}[X] = 1/\lambda$ ,  $\mathbb{E}[X^2] = 2/\lambda^2$ ,  $\text{var}[X] = 1/\lambda^2$
- (Q) What is the discrete rv which models a waiting time?



## Jointly Continuous

Two continuous rvs are **jointly continuous** if a non-negative function  $f_{X,Y}(x, y)$  (called joint PDF) satisfies: for **every** subset  $B$  of the two dimensional plane,

$$\mathbb{P}((X, Y) \in B) = \iint_{(x,y) \in B} f_{X,Y}(x, y) dx dy$$

1. The joint PDF is used to calculate probabilities

$$\mathbb{P}((X, Y) \in B) = \iint_{(x,y) \in B} f_{X,Y}(x, y) dx dy$$

Our particular interest:  $B = \{(x, y) \mid a \leq x \leq b, c \leq y \leq d\}$

2. The marginal PDFs of  $X$  and  $Y$  are from the joint PDF as:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx$$

3. The joint CDF is defined by  $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$ , and determines the joint PDF as:

$$f_{X,Y}(x, y) = \frac{\partial^2 F_{x,y}}{\partial x \partial y}(x, y)$$

4. A function  $g(X, Y)$  of  $X$  and  $Y$  defines a new random variable, and

$$\mathbb{E}[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy$$

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

- Similarly, for  $f_Y(y) > 0$ ,

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

- Remember: For a fixed event  $A$ ,  $\mathbb{P}(\cdot|A)$  is a legitimate probability law.
- Similarly, For a fixed  $y$ ,  $f_{X|Y}(x|y)$  is a legitimate PDF, since

$$\int_{-\infty}^{\infty} f_{X|Y}(x|y) dx = \frac{\int_{-\infty}^{\infty} f_{X,Y}(x,y) dx}{f_Y(y)} = 1$$

- Sum Rule

$$p_X(x) = \begin{cases} \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y) & \text{if discrete} \\ \int_{y \in \mathcal{Y}} f_{X,Y}(x, y) dy & \text{if continuous} \end{cases}$$

- Generally, for  $X = (X_1, X_2, \dots, X_D)$ ,

$$p_{X_i}(x_i) = \int p_X(x_1, \dots, x_i, \dots, x_D) d\mathbf{x}_{-i}$$

- Computationally challenging, because of high-dimensional sums or integrals

- Product Rule

$$p_{X,Y}(x, y) = p_X(x) \cdot p_{Y|X}(y|x)$$

**joint** dist. = **marginal** of the first  $\times$  **conditional** dist. of the second given the first

- Same as  $p_Y(y) \cdot p_{X|Y}(x|y)$

# Bayes Rule

- $X$ : state/cause/original value  $\rightarrow$   $Y$ : result/resulting action/noisy measurement
- Model:  $\mathbb{P}(X)$  (prior) and  $\mathbb{P}(Y|X)$  (cause  $\rightarrow$  result)
- Inference:  $\mathbb{P}(X|Y)$ ?

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

$$p_{X|Y}(x|y) = \frac{p_X(x)p_{Y|X}(y|x)}{p_Y(y)}$$

$$p_Y(y) = \sum_{x'} p_X(x')p_{Y|X}(y|x')$$

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

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

$$f_Y(y) = \int f_X(x')f_{Y|X}(y|x')dx'$$

$$\underbrace{p_{X|Y}(x|y)}_{\text{posterior}} = \frac{\overbrace{p_{Y|X}(y|x)}^{\text{likelihood}} \overbrace{p_X(x)}^{\text{prior}}}{\underbrace{p_Y(y)}_{\text{evidence}}}$$

$K$ : discrete,  $Y$ : continuous

- Inference of  $K$  given  $Y$

$$p_{K|Y}(k|y) = \frac{p_K(k)f_{Y|K}(y|k)}{f_Y(y)}$$

$$f_Y(y) = \sum_{k'} p_K(k')f_{Y|K}(y|k')$$

- Inference of  $Y$  given  $K$

$$f_{Y|K}(y|k) = \frac{f_Y(y)p_{K|Y}(k|y)}{p_K(k)}$$

$$p_K(k) = \int f_Y(y')p_{K|Y}(k|y')dy'$$



- (1) Construction of a Probability Space
- (2) Discrete and Continuous Probabilities
- (3) Sum Rule, Product Rule, and Bayes' Theorem
- (4) **Summary Statistics and Independence**
- (5) Gaussian Distribution
- (6) Conjugacy and the Exponential Family
- (7) Change of Variables/Inverse Transform

- Occurrence of  $A$  provides no new information about  $B$ . Thus, knowledge about  $A$  does not change my belief about  $B$ .

$$\mathbb{P}(B|A) = \mathbb{P}(B)$$

- Using  $\mathbb{P}(B|A) = \mathbb{P}(B \cap A)/\mathbb{P}(A)$ ,

Independence of  $A$  and  $B$ ,  $A \perp\!\!\!\perp B$

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$$

- **Q1.**  $A$  and  $B$  disjoint  $\implies A \perp\!\!\!\perp B$ ?  
No. Actually, really dependent, because if you know that  $A$  occurred, then, we know that  $B$  did not occur.
- **Q2.** If  $A \perp\!\!\!\perp B$ , then  $A \perp\!\!\!\perp B^c$ ? Yes.

- Remember: for a probability law  $\mathbb{P}(\cdot)$ , given, say  $B$ ,  $\mathbb{P}(\cdot|B)$  is a new probability law.
- Thus, we can talk about independence under  $\mathbb{P}(\cdot|B)$ .
- Given that  $C$  occurs, occurrence of  $A$  provides no new information about  $B$ .

$$\mathbb{P}(B|A \cap C) = \mathbb{P}(B|C)$$

Conditional Independence of  $A$  and  $B$  given  $C$ ,  $A \perp\!\!\!\perp B|C$

$$\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C) \times \mathbb{P}(B|C)$$

- **Q1.** If  $A \perp\!\!\!\perp B$ , then  $A \perp\!\!\!\perp B|C$ ? Suppose that  $A$  and  $B$  are independent. If you heard that  $C$  occurred,  $A$  and  $B$  are still independent?
- **Q2.** If  $A \perp\!\!\!\perp B|C$ ,  $A \perp\!\!\!\perp B$ ?

$$A \perp\!\!\!\perp B \rightarrow A \perp\!\!\!\perp B|C?$$

- Two independent coin tosses
  - $H_1$ : 1st toss is a head
  - $H_2$ : 2nd toss is a head
  - $D$ : two tosses have different results.
- $\mathbb{P}(H_1|D) = 1/2, \mathbb{P}(H_2|D) = 1/2$
- $\mathbb{P}(H_1 \cap H_2|D) = 0,$
- No.

## $A \perp\!\!\!\perp B|C \rightarrow A \perp\!\!\!\perp B?$

- Two coins: **Blue** and **Red**. Choose one uniformly at random, and proceed with two independent tosses.
- $\mathbb{P}(\text{head of blue}) = 0.9$  and  $\mathbb{P}(\text{head of red}) = 0.1$   
 $H_i$ :  $i$ -th toss is head, and  $B$ : blue is selected.
- $H_1 \perp\!\!\!\perp H_2|B$ ? Yes

$$\mathbb{P}(H_1 \cap H_2|B) = 0.9 \times 0.9, \quad \mathbb{P}(H_1|B)\mathbb{P}(H_2|B) = 0.9 \times 0.9$$

- $H_1 \perp\!\!\!\perp H_2$ ? No

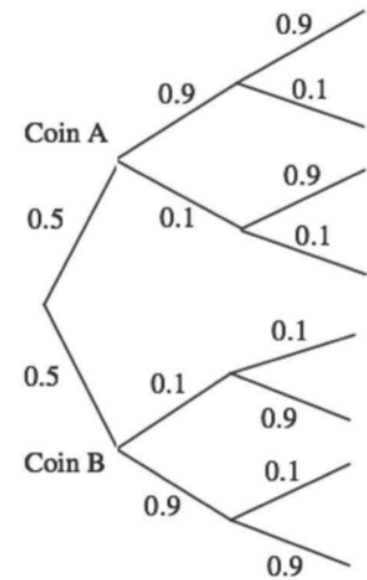
$$\mathbb{P}(H_1) = \mathbb{P}(B)\mathbb{P}(H_1|B) + \mathbb{P}(B^c)\mathbb{P}(H_1|B^c)$$

$$= \frac{1}{2}0.9 + \frac{1}{2}0.1 = \frac{1}{2}$$

$$\mathbb{P}(H_2) = \mathbb{P}(H_2) \quad (\text{because of symmetry})$$

$$\mathbb{P}(H_1 \cap H_2) = \mathbb{P}(B)\mathbb{P}(H_1 \cap H_2|B) + \mathbb{P}(B^c)\mathbb{P}(H_1 \cap H_2|B^c)$$

$$= \frac{1}{2}(0.9 \times 0.9) + \frac{1}{2}(0.1 \times 0.1) \neq \frac{1}{2}$$



- Two rvs

$$\mathbb{P}(\{X = x\} \cap \{Y = y\}) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y), \quad \text{for all } x, y$$

$$p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y)$$

$$\mathbb{P}(\{X = x\} \cap \{Y = y\} | C) = \mathbb{P}(X = x | C) \cdot \mathbb{P}(Y = y | C), \quad \text{for all } x, y$$

$$p_{X,Y|C}(x, y) = p_{X|C}(x) \cdot p_{Y|C}(y)$$

- Notation:  $X \perp\!\!\!\perp Y$  (independence),  $X \perp\!\!\!\perp Y | Z$  (conditional independence)

- Expectation

$$\mathbb{E}[X] = \sum_x x p_X(x), \quad \mathbb{E}[X] = \int_x x f_X(x) dx$$

- Variance, Standard deviation

- Measures how much the spread of PMF/PDF is

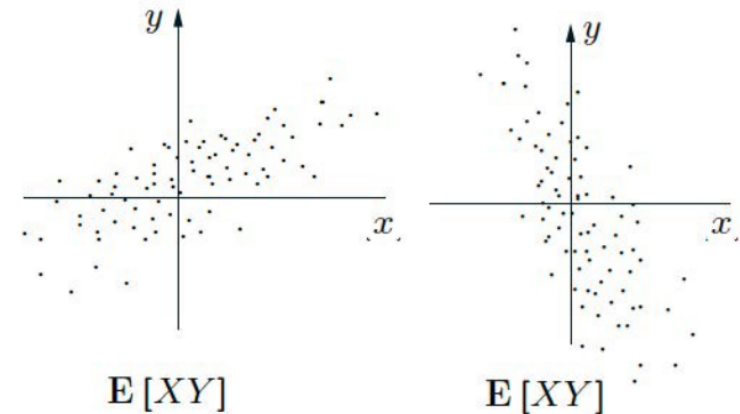
$$\text{var}[X] = \mathbb{E}[(X - \mu)^2]$$

$$\sigma_X = \sqrt{\text{var}[X]}$$

## Properties

- $\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$
- $\text{var}[aX + b] = a^2 \text{var}[X]$
- $\text{var}[X + Y] = \text{var}[X] + \text{var}[Y]$  if  $X \perp\!\!\!\perp Y$  (generally not equal)

- Goal: Given two rvs  $X$  and  $Y$ , quantify the degree of their dependence
  - Dependent: Positive (If  $X \uparrow$ ,  $Y \uparrow$ ) or Negative (If  $X \uparrow$ ,  $Y \downarrow$ )
  - Simple case:  $\mathbb{E}[X] = \mu_X = 0$  and  $\mathbb{E}[Y] = \mu_Y = 0$
- What about  $\mathbb{E}[XY]$ ? Seems good.
- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] = 0$  when  $X \perp\!\!\!\perp Y$
- More data points (thus increases) when  $xy > 0$  (both positive or negative)





## What If $\mu_X \neq 0, \mu_Y \neq 0$ ?

- Solution: Centering.  $X \rightarrow X - \mu_X$  and  $Y \rightarrow Y - \mu_Y$

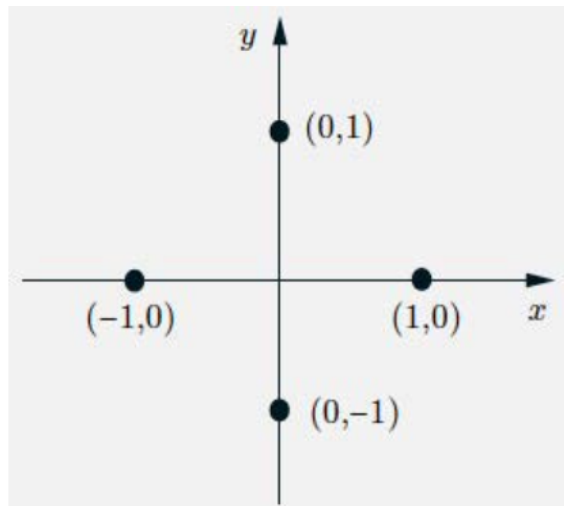
### Covariance

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])]$$

- After some algebra,  $\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- $X \perp\!\!\!\perp Y \implies \text{cov}(X, Y) = 0$
- $\text{cov}(X, Y) = 0 \implies X \perp\!\!\!\perp Y$ ? NO.
- When  $\text{cov}(X, Y) = 0$ , we say that  $X$  and  $Y$  are **uncorrelated**.

## Example: $\text{cov}(X, Y) = 0$ , but not independent

- $p_{X,Y}(1, 0) = p_{X,Y}(0, 1) = p_{X,Y}(-1, 0) = p_{X,Y}(0, -1) = 1/4$ .
- $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ , and  $\mathbb{E}[XY] = 0$ . So,  $\text{cov}(X, Y) = 0$
- Are they independent? No, because if  $X = 1$ , then we should have  $Y = 0$ .



$$\text{cov}(X, X) = 0$$

$$\text{cov}(aX + b, Y) = \mathbb{E}[(aX + b)Y] - \mathbb{E}[aX + b]\mathbb{E}[Y] = a \cdot \text{cov}(X, Y)$$

$$\text{cov}(X, Y + Z) = \mathbb{E}[X(Y + Z)] - \mathbb{E}[X]\mathbb{E}[Y + Z] = \text{cov}(X, Y) + \text{cov}(X, Z)$$

$$\text{var}[X + Y] = \mathbb{E}[(X + Y)^2] - (\mathbb{E}[X + Y])^2 = \text{var}[X] + \text{var}[Y] - 2\text{cov}(X, Y)$$

- Always bounded by some numbers, e.g.,  $[-1, 1]$
- Dimensionless metric. How? Normalization, but by what?

## Correlation Coefficient

$$\rho(X, Y) = \mathbb{E} \left[ \frac{(X - \mu_X)}{\sigma_X} \cdot \frac{Y - \mu_Y}{\sigma_Y} \right] = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}[X]\text{var}[Y]}}$$

- $-1 \leq \rho \leq 1$
- $|\rho| = 1 \implies X - \mu_X = c(Y - \mu_Y)$  (linear relation, VERY related)

Extension to Random Vectors  $\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$

- $\mathbb{E}(\mathbf{X}) := \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_n) \end{pmatrix}$

- Covariance of  $\mathbf{X} \in \mathbb{R}^n$  and  $\mathbf{Y} \in \mathbb{R}^m$

$$\text{cov}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}(\mathbf{X}\mathbf{Y}^T) - \mathbb{E}(\mathbf{X})\mathbb{E}(\mathbf{Y})^T \in \mathbb{R}^{n \times m}$$

- Variance of  $\mathbf{X}$ :  $\text{var}(\mathbf{X}) = \text{cov}(\mathbf{X}, \mathbf{X}) \in \mathbb{R}^{n \times n}$ , often denoted by  $\Sigma_{\mathbf{X}}$  (or simply  $\Sigma$ ):

$$\Sigma_{\mathbf{X}} := \text{var}[\mathbf{X}] = \begin{pmatrix} \text{cov}(X_1, X_1) & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_n) \\ \vdots & \vdots & & \vdots \\ \text{cov}(X_n, X_1) & \text{cov}(X_n, X_2) & \cdots & \text{cov}(X_n, X_n) \end{pmatrix}$$

- We call  $\Sigma_{\mathbf{X}}$  **covariance matrix** of  $\mathbf{X}$ .

- $N$ : number of samples,  $D$ : number of measurements (or original features)
- iid dataset  $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$  whose mean is 0 (well-centered), where each  $\mathbf{x}_i \in \mathbb{R}^D$ , and its corresponding data matrix

$$\mathbf{X} = (\mathbf{x}_1 \cdots \mathbf{x}_N) = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,N} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{D,1} & x_{D,2} & \cdots & x_{D,N} \end{pmatrix} \in \mathbb{R}^{D \times N}$$

- (data) covariance matrix

L10(1)

$$\mathbf{S} = \frac{1}{N} \mathbf{X} \mathbf{X}^T = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^T \in \mathbb{R}^{D \times D}$$

- **Question.** Relation between covariance matrix and data covariance matrix?

- Covariance matrix for a random vector  $\mathbf{Y} = (Y_1, \dots, Y_D)^T$ ,

$$\Sigma_{\mathbf{Y}} = \begin{pmatrix} \text{cov}(Y_1, Y_1) & \text{cov}(Y_1, Y_2) & \cdots & \text{cov}(Y_1, Y_D) \\ \vdots & \vdots & & \vdots \\ \text{cov}(Y_D, Y_1) & \text{cov}(Y_D, Y_2) & \cdots & \text{cov}(Y_D, Y_D) \end{pmatrix}$$

- Data covariance matrix  $\mathbf{S} \in \mathbb{R}^{D \times D}$ 
  - Each  $Y_i$  has  $N$  samples  $(x_{i,1} \cdots x_{i,N})$

$$\begin{aligned} \mathbf{S}_{ij} = \text{cov}(Y_i, Y_j) &= \frac{1}{N} \sum_{k=1}^N x_{i,k} \cdot x_{j,k} \\ &= \text{average covariance (over samples) btwn features } i \text{ and } j \end{aligned}$$



For two random vectors  $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^n$ ,

- $\mathbb{E}(\mathbf{X} + \mathbf{Y}) = \mathbb{E}(\mathbf{X}) + \mathbb{E}(\mathbf{Y}) \in \mathbb{R}^n$
- $\text{var}(\mathbf{X} + \mathbf{Y}) = \text{var}(\mathbf{X}) + \text{var}(\mathbf{Y}) \in \mathbb{R}^{n \times n}$
- Assume  $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$ .
  - $\mathbb{E}(\mathbf{Y}) = \mathbf{A}\mathbb{E}(\mathbf{X}) + \mathbf{b}$
  - $\text{var}(\mathbf{Y}) = \text{var}(\mathbf{A}\mathbf{X}) = \mathbf{A} \text{var}(\mathbf{X}) \mathbf{A}^T$
  - $\text{cov}(\mathbf{X}, \mathbf{Y}) = \Sigma_{\mathbf{X}} \mathbf{A}^T$  (Please prove)

- (1) Construction of a Probability Space
- (2) Discrete and Continuous Probabilities
- (3) Sum Rule, Product Rule, and Bayes' Theorem
- (4) Summary Statistics and Independence
- (5) **Gaussian Distribution**
- (6) Conjugacy and the Exponential Family
- (7) Change of Variables/Inverse Transform

# Normal (also called Gaussian) Random Variable

- Why important?
  - Central limit theorem (중심극한정리)
    - One of the most remarkable findings in the probability theory
  - Convenient analytical properties
  - Modeling aggregate noise with many small, independent noise terms

- Standard Normal  $\mathcal{N}(0, 1)$

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

- $\mathbb{E}[X] = 0$
- $\text{var}[X] = 1$

- General Normal  $\mathcal{N}(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

- $\mathbb{E}[X] = \mu$
- $\text{var}[X] = \sigma^2$

- $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$  with the mean vector  $\boldsymbol{\mu} = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_n) \end{pmatrix}$  and the covariance matrix  $\boldsymbol{\Sigma}$ .

- A Gaussian random vector  $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$  has a joint pdf of the form:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \exp \left( -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right),$$

where  $\boldsymbol{\Sigma}$  is symmetric and positive definite.

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

- Marginals of Gaussians are Gaussians
- Conditionals of Gaussians are Gaussians
- Products of Gaussian Densities are Gaussians.
- A sum of two Gaussians is Gaussian if they are independent
- Any linear/affine transformation of a Gaussian is Gaussian.

# Marginals and Conditionals of Gaussians

- $\mathbf{X}$  and  $\mathbf{Y}$  are Gaussians with mean vectors  $\mu_{\mathbf{X}}$  and  $\mu_{\mathbf{Y}}$ , respectively.
- Gaussian random vector  $\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$  with  $\mu = \begin{pmatrix} \mu_{\mathbf{X}} \\ \mu_{\mathbf{Y}} \end{pmatrix}$  and the covariance matrix

$$\Sigma_{\mathbf{Z}} = \begin{pmatrix} \Sigma_{\mathbf{X}} & \Sigma_{\mathbf{XY}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{Y}} \end{pmatrix}, \text{ where } \Sigma_{\mathbf{XY}} = \text{cov}(\mathbf{X}, \mathbf{Y}).$$

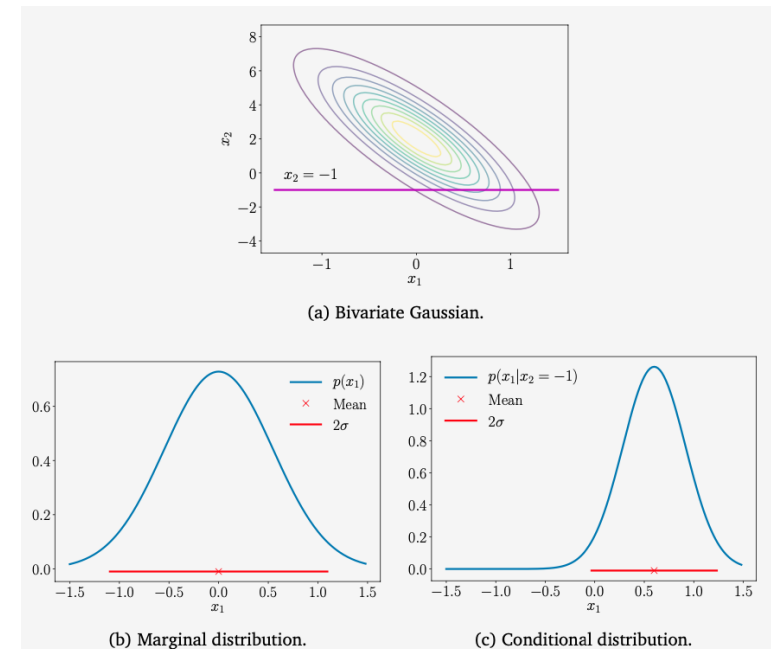
- Marginal.

$$f_{\mathbf{X}}(\mathbf{x}) = \int f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) d\mathbf{y} \sim \mathcal{N}(\mu_{\mathbf{X}}, \Sigma_{\mathbf{X}})$$

- Conditional.  $\mathbf{X} | \mathbf{Y} \sim \mathcal{N}(\mu_{\mathbf{X}|\mathbf{Y}}, \Sigma_{\mathbf{X}|\mathbf{Y}}),$

$$\mu_{\mathbf{X}|\mathbf{Y}} = \mu_{\mathbf{X}} + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}^{-1}(\mathbf{Y} - \mu_{\mathbf{Y}})$$

$$\Sigma_{\mathbf{X}|\mathbf{Y}} = \Sigma_{\mathbf{X}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{YX}}$$



- **Lemma.** Up to recaling, the pdf of the form  $\exp(-\frac{1}{2}ax^2 - 2bx + c)$  is  $\mathcal{N}(\frac{b}{a}, \frac{1}{a})$ .
- Using the above Lemma, the product of two Gaussians  $\mathcal{N}(\mu_0, \nu_0)$  and  $\mathcal{N}(\mu_1, \nu_1)$  is Gaussian up to rescaling.

**Proof.**

$$\begin{aligned} & \exp\left(-(x - \mu_0)^2/2\nu_0\right) \times \exp\left(-(x - \mu_1)^2/2\nu_1\right) \\ &= \exp\left[-\frac{1}{2}\left(\left(\frac{1}{\nu_0} + \frac{1}{\nu_1}\right)x^2 - 2\left(\frac{\mu_0}{\nu_0} + \frac{\mu_1}{\nu_1}\right)x + c\right)\right] \\ &\Rightarrow \mathcal{N}\left(\overbrace{\frac{1}{\nu_0^{-1} + \nu_1^{-1}}}^{=\nu}, \left(\frac{\mu_0}{\nu_0} + \frac{\mu_1}{\nu_1}\right)\right) = \mathcal{N}\left(\frac{\nu_1\mu_0 + \nu_0\mu_1}{\nu_0 + \nu_1}, \frac{\nu_0\nu_1}{\nu_0 + \nu_1}\right) \end{aligned}$$

- Similar results for the matrix version.
- The product of the densities of two Gaussian vectors  $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$  and  $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$  is Gaussian up to rescaling.
- The resulting Gaussian is given by:

$$\mathcal{N}\left(\boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}_0(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1}\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1}\boldsymbol{\Sigma}_0\right)$$

Compare the above to this:

$$\mathcal{N}\left(\frac{\nu_1\boldsymbol{\mu}_0 + \nu_0\boldsymbol{\mu}_1}{\nu_0 + \nu_1}, \frac{\nu_0\nu_1}{\nu_0 + \nu_1}\right)$$



# Formula: Conditional and Marginal Gaussians

If we have a marginal Gaussian distribution for  $\mathbf{x}$  and a conditional Gaussian distribution for  $\mathbf{y}$  given  $\mathbf{x}$  in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \quad (\text{B.42})$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \quad (\text{B.43})$$

then the marginal distribution of  $\mathbf{y}$ , and the conditional distribution of  $\mathbf{x}$  given  $\mathbf{y}$ , are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) \quad (\text{B.44})$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \quad (\text{B.45})$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}. \quad (\text{B.46})$$

If we have a joint Gaussian distribution  $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$  with  $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$  and we define the following partitions

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \quad (\text{B.47})$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix} \quad (\text{B.48})$$

then the conditional distribution  $p(\mathbf{x}_a|\mathbf{x}_b)$  is given by

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}) \quad (\text{B.49})$$

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \quad (\text{B.50})$$

and the marginal distribution  $p(\mathbf{x}_a)$  is given by

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}). \quad (\text{B.51})$$

---

<sup>1</sup>Source: Pattern Recognition and Machine Learning, Springer by Christopher M. Bishop

- $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{X}})$  and  $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{Y}}, \boldsymbol{\Sigma}_{\mathbf{Y}})$

$$\Rightarrow a\mathbf{X} + b\mathbf{Y} \sim \mathcal{N}(a\boldsymbol{\mu}_{\mathbf{X}} + b\boldsymbol{\mu}_{\mathbf{Y}}, a^2\boldsymbol{\Sigma}_{\mathbf{X}} + b^2\boldsymbol{\Sigma}_{\mathbf{Y}})$$

- $f_1(x)$  is the density of  $\mathcal{N}(\mu_1, \sigma_1^2)$  and  $f_2(x)$  is the density of  $\mathcal{N}(\mu_2, \sigma_2^2)$
- **Question.** What are the mean and the variance of the random variable  $Z$  which has the following density  $f(x)$ ?

$$f(x) = \alpha f_1(x) + (1 - \alpha) f_2(x)$$

Answer:

$$\mathbb{E}(Z) = \alpha \mu_1 + (1 - \alpha) \mu_2$$

$$\text{var}(Z) = \left( \alpha \sigma_1^2 + (1 - \alpha) \sigma_2^2 \right) + \left( [\alpha \mu_1^2 + (1 - \alpha) \mu_2^2] - [\alpha \mu_1 + (1 - \alpha) \mu_2]^2 \right)$$

- Linear transformation<sup>2</sup> preserves normality

## Linear transformation of Normal

If  $X \sim \mathcal{N}(\mu, \sigma^2)$ , then for  $a \neq 0$  and  $b$ ,  $Y = aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$ .

- Thus, every normal rv can be **standardized**:

If  $X \sim \mathcal{N}(\mu, \sigma^2)$ , then  $Y = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$

- Thus, we can make the **table** which records the following CDF values:

$$\Phi(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(Y < y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y e^{-t^2/2} dt$$

---

<sup>2</sup>Strictly speaking, this is affine transformation.

- $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$ , where  $\mathbf{X} \in \mathbb{R}^n$ ,  $\mathbf{Y}, \mathbf{b} \in \mathbb{R}^m$ , and  $\mathbf{A} \in \mathbb{R}^{m \times n}$

$$\Rightarrow \mathbf{Y} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$$

- (1) Construction of a Probability Space
- (2) Discrete and Continuous Probabilities
- (3) Sum Rule, Product Rule, and Bayes' Theorem
- (4) Summary Statistics and Independence
- (5) Gaussian Distribution
- (6) Conjugacy and the Exponential Family
- (7) Change of Variables/Inverse Transform

- Bayesian Inference

$$\underbrace{p(\theta | D)}_{\text{posterior}} = \frac{\overbrace{p(D | \theta)}^{\text{likelihood}} \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(D)}_{\text{evidence}}}$$

- The forms of likelihood and prior come from a model.
- **Question.** Given a form of likelihood, how can I choose a prior such that the resulting posterior has the same form as the prior?
  - Such prior is called **conjugate prior** (to the given likelihood)
  - **Pros:** Algebraic calculation of posterior and even analytical description is often possible.
  - **Cons:** A restricted form of prior, which may lead to distorted understanding about data interpretation.

# Conjugate Priors: Definition and Examples

- **Definition.** A prior is **conjugate** for the likelihood function if the posterior is of the same form/type as the prior.
- Representative conjugate priors

Likelihood	Prior	Posterior
Poisson	Gamma	Gamma
Bernoulli	Beta	Beta
Binomial	Beta	Beta
Normal	Normal/inverse Gamma	Normal/inverse Gamma
Normal	Normal/inverse Wishart	Normal/inverse Wishart
Exponential	Gamma	Gamma
Multinomial	Dirichlet	Dirichlet



## Beta distribution

A continuous rv  $\Theta$  follows a beta distribution with integer parameters  $\alpha, \beta > 0$ , if

$$f_{\Theta}(\theta) = \begin{cases} \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}, & 0 < \theta < 1, \\ 0, & \text{otherwise,} \end{cases}$$

where  $B(\alpha, \beta)$ , called Beta function, is a normalizing constant, given by

$$B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1} (1 - \theta)^{\beta-1} d\theta = \frac{(\alpha - 1)! (\beta - 1)!}{(\alpha + \beta - 1)!}$$

- Beta distribution models a continuous random variable over a finite interval  $[0, 1]$ .
- A special case of  $Beta(1, 1)$  is *Uniform* $[0, 1]$

## Example: Beta-Binomial Conjugacy

- Assume that the parameter  $\Theta \sim \text{Beta}(\alpha, \beta)$  (prior):  $p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$
- $\theta \sim \Theta$  and  $X \sim \text{Bin}(N, \theta)$ . Thus,  $p(x | \theta) = \binom{N}{x} \theta^x (1-\theta)^{N-x}$  (likelihood)
- Posterior  $\propto$  (likelihood)  $\times$  (prior)

$$\begin{aligned} p(\theta | x = h) &\propto \binom{N}{h} \theta^h (1-\theta)^{N-h} \times \theta^{\alpha-1} (1-\theta)^{\beta-1} \\ &= \theta^{h+\alpha-1} (1-\theta)^{(N-h)+\beta-1} \\ &\sim \text{Beta}(h + \alpha, N - h + \beta) \end{aligned}$$

- A **statistic** of a random variable  $\mathbf{X}$  is a deterministic function of  $\mathbf{X}$ .
- **Example.** For  $\mathbf{X} = (X_1 \ X_2 \ \dots \ X_n)^T$ , the sample mean  $T(\mathbf{X}) = \frac{1}{N}(X_1 + \dots + X_n)$  is a statistic.
- **Question.** Does a statistic contain all the information for the inference from data? (e.g., the parameter estimation of a distribution based on data)
- **Sufficient statistics:** carry all the information for the inference
- **Definition.** A statistic  $T = T(\mathbf{X})$  is said to be **sufficient** for  $\mathbf{X}$  with its pdf or pmf  $p_{\mathbf{X}}(\mathbf{x}; \theta)$ ,<sup>3</sup> if the conditional distribution of  $\mathbf{X}$  given  $T(\mathbf{X}) = t$  is **independent** of  $\theta$  for all  $t$ .

---

<sup>3</sup>The parameter can be a vector, but we do not use  $\theta$  for simplicity.

- $X_1, X_2$ : independent Poisson variables with common parameter  $\lambda$  which is the expectation.

- **Claim.**  $T(\mathbf{X}) = X_1 + X_2$  is a sufficient statistic for inference of  $\lambda$ .

- Joint distribution

$$\mathbb{P}(x_1, x_2) = \frac{\lambda^{x_1+x_2}}{x_1!x_2!} e^{-2\lambda}$$

- Conditional dist. of  $X_1$  given  $X_1 + X_2 = t$

$$\mathbb{P}(x_1 | X_1 + X_2 = t) = \frac{1}{x_1!(t-x_1)!} \left( \frac{1}{\sum_{y=0}^t \frac{1}{y!(t-y)!}} \right)^{-1}$$

- Independent of  $\lambda \implies T$  is a sufficient statistic.

## Factorization Theorem

A necessary and sufficient condition for a statistic  $T$  to be sufficient for  $X$  with its pdf or pmf  $p_{\mathbf{X}}(\mathbf{x}; \theta)$  is that there exist non-negative functions  $g_{\theta}$  and  $h$  such that

$$p_{\mathbf{X}}(\mathbf{x}; \theta) = g_{\theta}(T(\mathbf{x}))h(\mathbf{x}).$$

- **Example.** Continuing the Poisson example, suppose that  $X_1, \dots, X_n$  are iid according to a Poisson distribution with parameter  $\lambda$ . Then, with  $\mathbf{X} = (X_1, \dots, X_n)$ ,

$$\mathbb{P}_{\mathbf{X}}(x_1, \dots, x_n) = \lambda^{\sum x_i} e^{-n\lambda} / \prod (x_i!)$$

- $T(\mathbf{X}) = \sum X_i$  is a sufficient statistic.

- Three levels of abstraction when we use a distribution to model a random phenomenon
- L1.** Fix a particular named distribution with fixed parameters
  - **Example.** Use a Gaussian with zero mean and unit variance,  $\mathcal{N}(0, 1)$
- L2.** Use a parametric distribution and infer the parameters from data
  - **Example.** Use a Gaussian with unknown mean and variance,  $\mathcal{N}(\mu, \sigma^2)$ , and infer  $(\mu, \sigma^2)$  from data
- L3.** Consider a family of distributions which satisfy “nice” properties
  - **Example.** Exponential family

An **exponential family** is a family of probability distributions, parameterized by  $\theta \in \mathbb{R}^D$ , has the form

$$p_{\mathbf{X}}(\mathbf{x}; \theta) = h(\mathbf{x}) \exp \left( \langle \theta, T(\mathbf{x}) \rangle - A(\theta) \right),$$

where  $\mathbf{X} \in \mathbb{R}^n$  and  $T(\mathbf{x}) : \mathbb{R}^n \mapsto \mathbb{R}^D$  is a vector of sufficient statistics.

- Nothing but a particular form of  $g_{\theta}(\cdot)$  in the F-N factorization theorem
- $\langle \theta, T(\mathbf{x}) \rangle$  is an inner product, e.g., the standard dot product.
- Essentially, it is of the form:  $p_{\mathbf{X}}(\mathbf{x}; \theta) \propto \exp(\theta^T T(\theta))$
- $A(\theta)$ : normalization constant, called **log-partition function**.
- Why Useful?
  - Parametric form of conjugate priors (see pp. 190 in the text), offering sufficient statistics, etc.

- Gaussian as exponential family, a random variable  $X \sim \mathcal{N}(\mu, \sigma^2)$ .
  - Let  $T(\mathbf{x}) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$  and  $\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix}$

$$p(\mathbf{x} \mid \boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\theta}^\top T(\mathbf{x})\right) = \exp\left(\frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2}\right) = \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$$



- (1) Construction of a Probability Space
- (2) Discrete and Continuous Probabilities
- (3) Sum Rule, Product Rule, and Bayes' Theorem
- (4) Summary Statistics and Independence
- (5) Gaussian Distribution
- (6) Conjugacy and the Exponential Family
- (7) Change of Variables/Inverse Transform

- If  $X \sim \mathcal{N}(0, 1)$ , what is the distribution of  $Y = X^2$ ?
- If  $X_1, X_2 \sim \mathcal{N}(0, 1)$ , what is the distribution of  $Y = \frac{1}{2}(X_1 + X_2)$ ?
- Two techniques
  - CDF-based technique
  - Change-of-Variable technique
- In this lecture note, we focus on the case of univariate random variables for simplicity.

**S1.** Find the CDF:  $F_Y(y) = \mathbb{P}(Y \leq y)$

**S2.** Differentiate the CDF to get the pdf  $f_Y(y)$ :  $f_Y(y) = \frac{d}{dy} F_Y(y)$

- **Example.**  $f_X(x) = -3x^2$ ,  $0 \leq x \leq 1$ . What is the pdf of  $Y = X^2$ ?

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X^2 \leq y) = \mathbb{P}(X \leq \sqrt{y}) = F_X(\sqrt{y})$$

$$= \int_0^{\sqrt{y}} 3t^2 dt = y^{\frac{3}{2}}, \quad 0 \leq y \leq 1$$

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{3}{2} \sqrt{y}, \quad 0 \leq y \leq 1$$

- Assume that  $X \sim \exp(1)$ , i.e.,  $f_X(x) = e^{-x}$  and  $F_X(x) = 1 - e^{-x}$ . How to make a programming code that gives random samples following the distribution  $X$ ?
- **Theorem. Probability Integral Theorem.** Let  $X$  be a continuous rv with a **strictly monotonic** CDF  $F(\cdot)$ . Then, if we define a new rv  $U$  as  $U := F(X)$ , then  $U$  follows the uniform distribution over  $[0,1]$ .
- **Proof.** Will show that  $F_U(u) = u$ , which is the CDF of a standard uniform rv.

$$F_U(u) = \mathbb{P}(U \leq u) = \mathbb{P}(F(X) \leq u) \stackrel{(*)}{=} \mathbb{P}(X \leq F^{-1}(u)) = F(F^{-1}(u)) = u,$$

where  $(*)$  is due to the strict monotonicity of  $F(\cdot)$ .

Pseudo Code of getting a random sample with the distribution  $F(\cdot)$ .

**Step 1.** Get a random sample  $u$  over  $[0, 1]$  (most of software packages include this capability of generating a random number generation)

**Step 2.** Get a value  $x = F^{-1}(u)$ .

- Chain rule of calculus:  $\int f(g(x))g'(x)dx = \int f(u)du$ , where  $u = g(x)$ .
- Consider a rv  $X \in [a, b]$  and an invertible, strictly increasing function  $U$ .

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(U(X) \leq y) = \mathbb{P}(X \leq U^{-1}(y)) = \int_a^{U^{-1}(y)} f_X(x)dx$$

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} \int_a^{U^{-1}(y)} f_X(x)dx = \frac{d}{dy} \int_a^{U^{-1}(y)} f_X(U^{-1}(y))U^{-1'}(y)dy \\ &= f_X(U^{-1}(y)) \cdot \frac{d}{dy} U^{-1}(y) \end{aligned}$$

- Including the case when  $U$  is strictly decreasing,

$$f_Y(y) = f_X(U^{-1}(y)) \cdot \left| \frac{d}{dy} U^{-1}(y) \right|$$

- **Theorem.** Let  $f_{\mathbf{X}}(\mathbf{x})$  is the pdf of multivariate continuous random vector  $\mathbf{X}$ . If  $\mathbf{Y} = U(\mathbf{X})$  is differentiable and invertible, the pdf of  $\mathbf{Y}$  is given as:

$$f(\mathbf{y}) = f_{\mathbf{X}}(U^{-1}(\mathbf{y})) \cdot \left| \det \left( \frac{d}{d\mathbf{y}} U^{-1}(\mathbf{y}) \right) \right|$$

- **Example.** For a bivariate rv  $\mathbf{X}$  with its pdf  $f\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = \frac{1}{2\pi} \exp\left(-\frac{1}{2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right)$ ,

consider  $\mathbf{Y} = \mathbf{A}\mathbf{X}$ , where  $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ . Then, we have the following pdf of  $\mathbf{Y}$ :

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{2\pi} \exp\left(-\frac{1}{2} \mathbf{y}^T (\mathbf{A}^{-1})^T \mathbf{A}^{-1} \mathbf{y}\right) |ad - bc|^{-1}$$

Questions?



1)