Lecture 6: Probability and Distributions

CSE4130: 기초머신러닝

Junsuk Choe (최준석)

Roadmap

(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

Roadmap

(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

What Do We Want?

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 $\boxed{\text{probabilistic model}}$ 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?

Probabilistic Model

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: $\boxed{\text{Sample Space } \Omega}$

2. Assigned numbers to each outcome of Ω: $\boxed{\text{Probability Law } \mathbb{P}(\cdot)}$

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

Sample Space Ω

The set of all outcomes of $\boxed{\text{my interest}}$

1. Mutually exclusive
2. Collectively exhaustive
3. At the $\boxed{\text{right granularity}}$ (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.

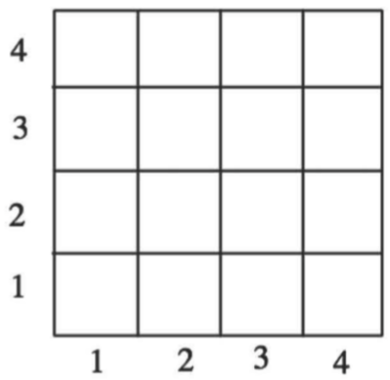
$$\implies \Omega = \{(H, R), (T, R), \\ (H, NR), (T, NR)\},$$

where R(Rain), NR(No Rain).

Examples: Sample Space Ω

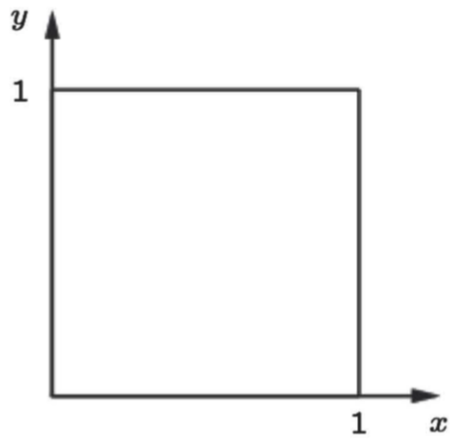
- *Discrete case*: Two rolls of a tetrahedral die

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



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

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



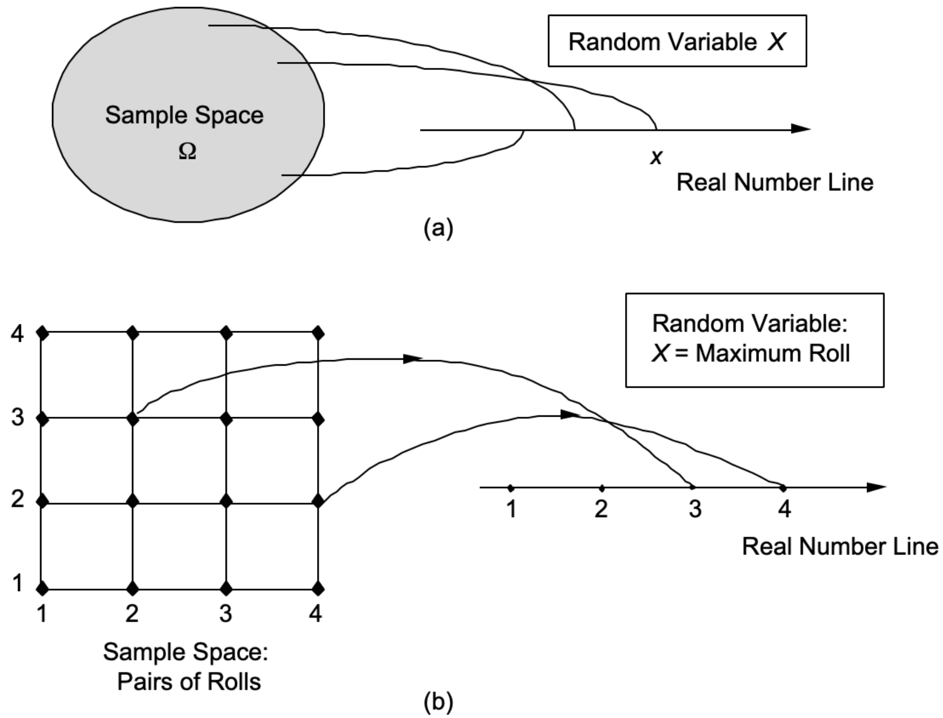
Probability Law

- Assign numbers to what? Each outcome?
- What is the probability of dropping a needle at $(0.5, 0.5)$ over the 1×1 plane?
- Assign numbers to each subset of Ω: A subset of Ω: 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 Ω. For example, in the discrete case, the power set of Ω.
- Probability Space $(\Omega, \mathcal{A}, \mathbb{P}(\cdot))$
- In general, we use the target space $\mathcal{T}$ instead.

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.



Random Variable: More Formally

- A random variable X is a $\boxed{\text{function}}$ which maps from Ω to a value x in $\mathcal{T}$.
- For any subset $S \in \mathcal{T}$, we associate $\mathbb{P}_X(S) \in [0, 1]$ (the probability) to a particular event occurring corresponding to the random variable X.
- Different random variables X, Y, ... 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,

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

Example 6.1

Conditioning: Motivating Example

- Pick a person a at random
  - event A: a's age ≤ 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.

Conditional Probability

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

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

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

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

Roadmap

(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

Discrete Random Variables

- The values that a random variable X takes is discrete (i.e., finite or countably infinite).
- Then, $p_X(x) := \mathbb{P}(X = x) := \mathbb{P}\Big(\{\omega \in \Omega \mid X(w) = x\}\Big)$, 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.

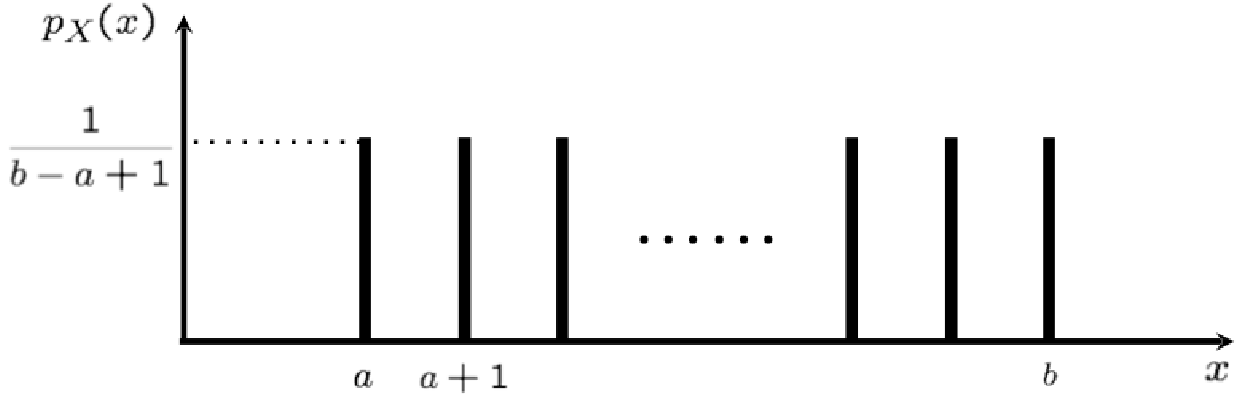
- Models a trial that results in binary results, e.g., success/failure, head/tail
- Very useful for an $\boxed{\text{indicator rv}}$ of an event A. Define a rv $\mathbf{1}_A$ as:

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

[1] 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.
- $p_X(i) = \frac{1}{b-a+1}$, $i \in \Omega$.

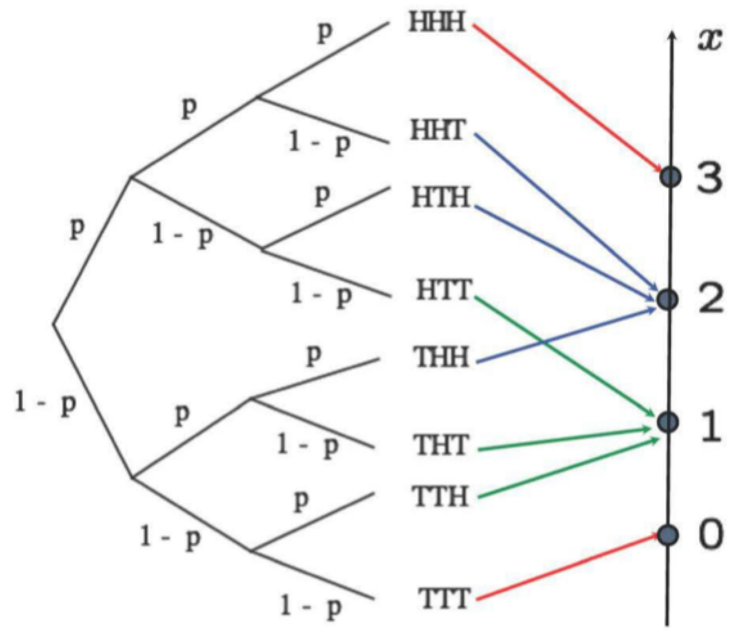


- Models complete ignorance (I don't know anything about X)

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.

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



Poisson X with parameter λ

- $\mathit{Binomial}(n, p)$: 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 $np = \lambda$

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

- Is this a legitimate PMF?

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

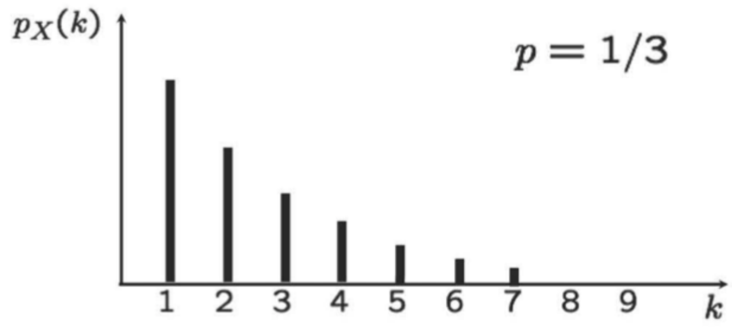
- Prove this:

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

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

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

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

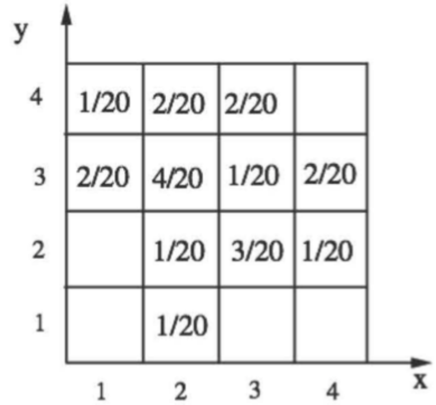
- $\sum_x \sum_y p_{X,Y}(x,y) = 1$

- $\boxed{\text{Marginal PMF.}}$

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

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

Example.



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

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

$$\mathbb{P}(X = Y) = 1/20 + 1/20 = 2/20$$

Conditional PMF

- 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 \ x	1	2	3	4
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		

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

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

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

Continuous RV and Probability Density Function (PDF)

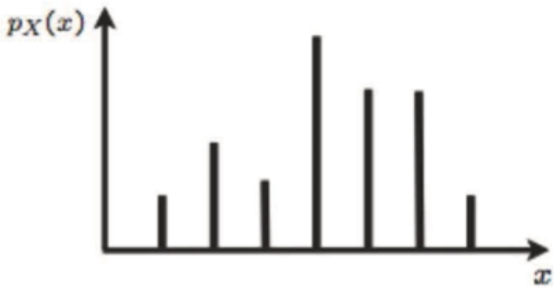
- Many cases when random variable have “continuous values”, e.g., velocity of a car

Continuous Random Variable

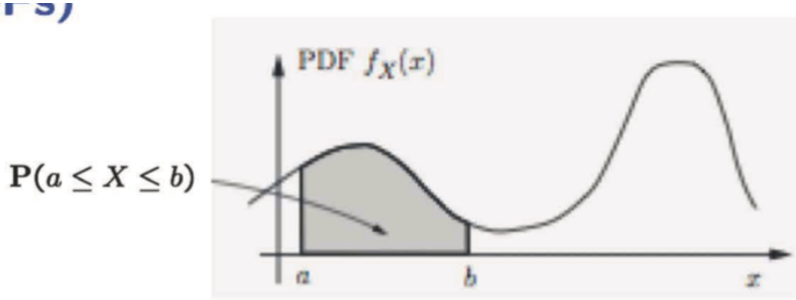
A rv X is continuous if $\exists$ a function f_X, called $\boxed{\text{probability density function (PDF)}}$, s.t.

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

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

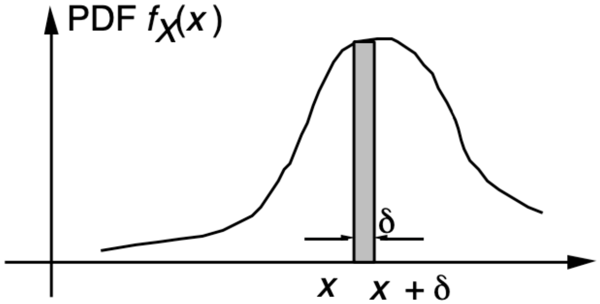


- $\mathbb{P}(a \le X \le b) = \sum_{x:a\le x\le b} p_X(x)$
- $p_X(x) \ge 0, \sum_x p_X(x) = 1$



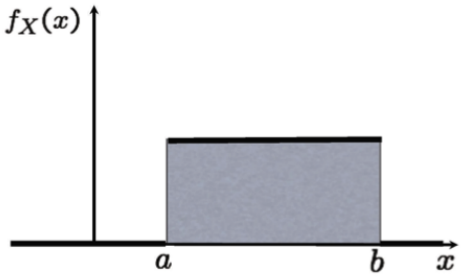
- $\mathbb{P}(a \le X \le b) = \int_a^b f_X(x)dx$
- $f_X(x) \ge 0, \int_{-\infty}^{\infty} f_X(x)dx = 1$

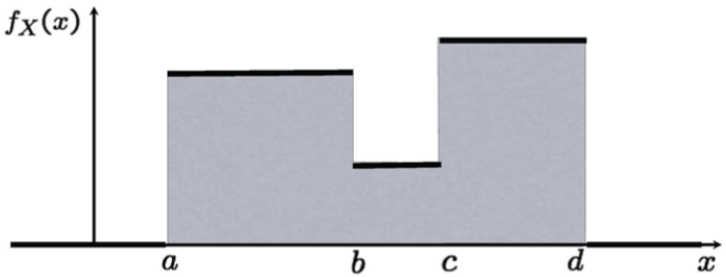
PDF and Examples



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

Examples



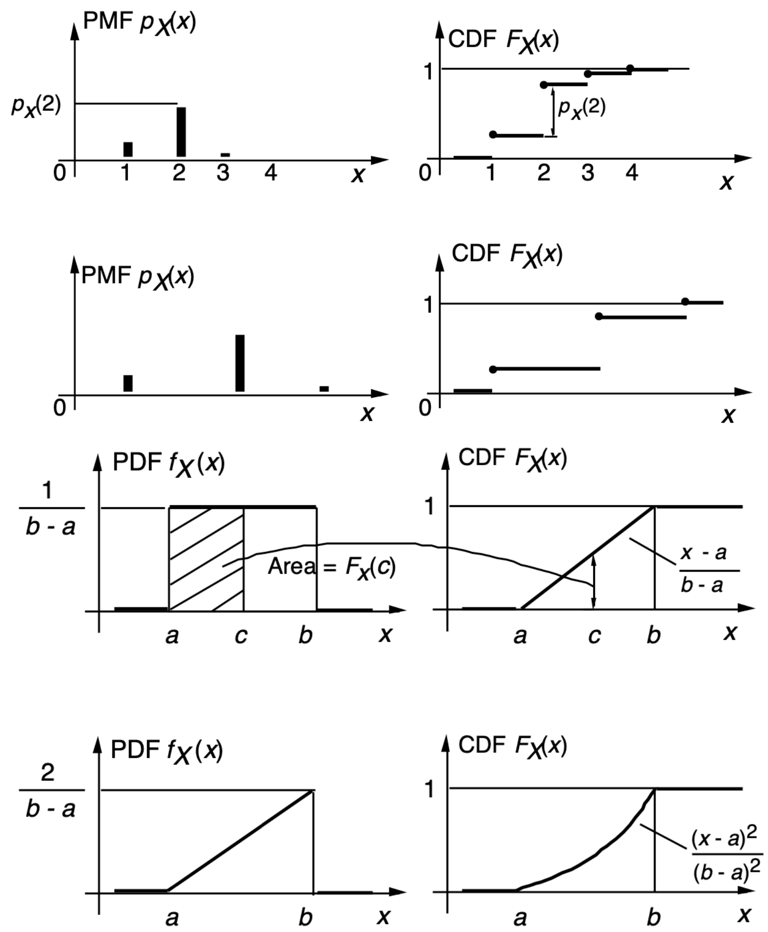


Cumulative Distribution Function (CDF)

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

$$F_X(x) = \mathbb{P}(X \le x) = \begin{cases} \sum_{k \le x} p_X(k), & \text{discrete} \\ \int_{-\infty}^{x} f_X(t)dt, & \text{continuous} \end{cases}$$

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



CDF Properties

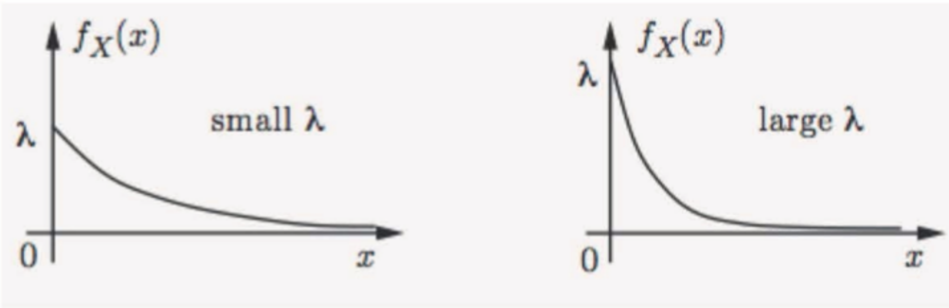
- Non-decreasing
- $F_X(x)$ tends to 1, as $x \to \infty$
- $F_X(x)$ tends to 0, as $x \to -\infty$

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

- A rv X is called exponential with λ, 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?



Continuous: Joint PDF and CDF (1)

Jointly Continuous

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

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

1. The joint PDF is used to calculate probabilities

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

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

Continuous: Joint PDF and CDF (2)

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

Continuous: Conditional PDF given a RV

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

- 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 $X = (X_1, X_2, \ldots, X_D)$,

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

  - 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: $\mathbb{P}(X|Y)$?

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

$$
\begin{aligned}
f_{X,Y}(x,y) &= f_X(x)f_{Y|X}(y|x) \\
&= f_Y(y)f_{X|Y}(x|y) \\
{\color{red}f_{X|Y}(x|y)} &= \frac{f_X(x)f_{Y|X}(y|x)}{f_Y(y)} \\
f_Y(y) &= \int f_X(x')f_{Y|X}(y|x')dx'
\end{aligned}
$$

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

Bayes Rule for Mixed Case

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

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

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

Roadmap

(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

Independence

- Occurrence of A provides no new information about B. Thus, knowledge about A does no 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, $\boxed{A \perp\!\!\!\perp B}$

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

- Q1. A and B disjoint $\Longrightarrow A \perp\!\!\!\perp B$?
  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.

Conditional Independence

- 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, $\boxed{A \perp\!\!\!\perp B|C}$

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

- 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 $A \perp\!\!\!\perp B|C$, $A \perp\!\!\!\perp B$?

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

- 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.

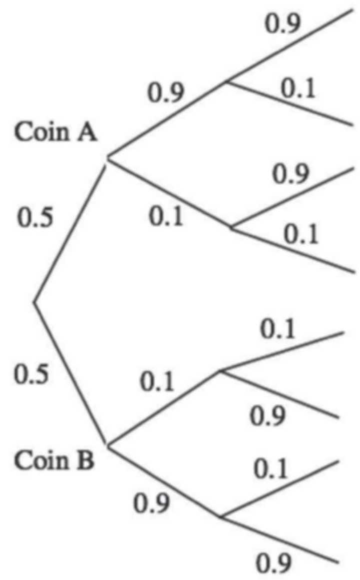
$A \perp\!\!\!\perp B|C \to A \perp\!\!\!\perp B$?

- 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

$$\begin{aligned}
\mathbb{P}(H_1) &= \mathbb{P}(B)\mathbb{P}(H_1|B) + \mathbb{P}(B^c)\mathbb{P}(H_1|B^c) \\
&= \frac{1}{2}0.9 + \frac{1}{2}0.1 = \frac{1}{2} \\
\mathbb{P}(H_2) &= \mathbb{P}(H_1) \quad \text{(because of symmetry)} \\
\mathbb{P}(H_1 \cap H_2) &= \mathbb{P}(B)\mathbb{P}(H_1 \cap H_2|B) + \mathbb{P}(B^c)\mathbb{P}(H_1 \cap H_2|B^c) \\
&= \frac{1}{2}(0.9 \times 0.9) + \frac{1}{2}(0.1 \times 0.1) \neq \frac{1}{4}
\end{aligned}$$



Independence for Random Variables

- 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), $X \perp\!\!\!\perp Y|Z(conditionalindependence)$

Expectation/Variance

- 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

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

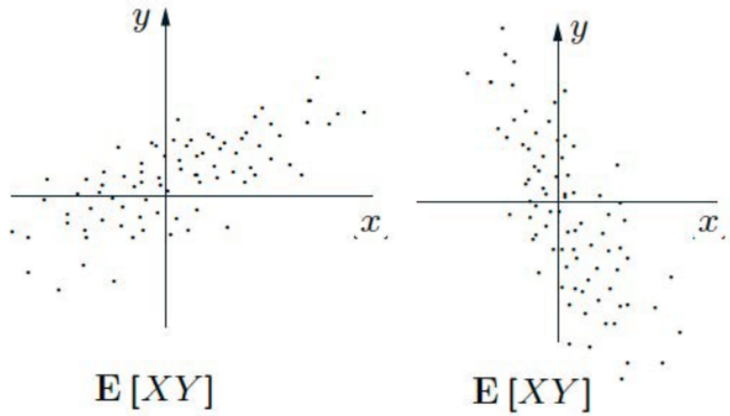
Properties

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

Covariance

- 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 $\mathbb{E}[XY]$?
- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] = 0$ when $X \perp\!\!\!\perp Y$
- Increase: more data points when $xy > 0$ (both positive or negative)



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

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

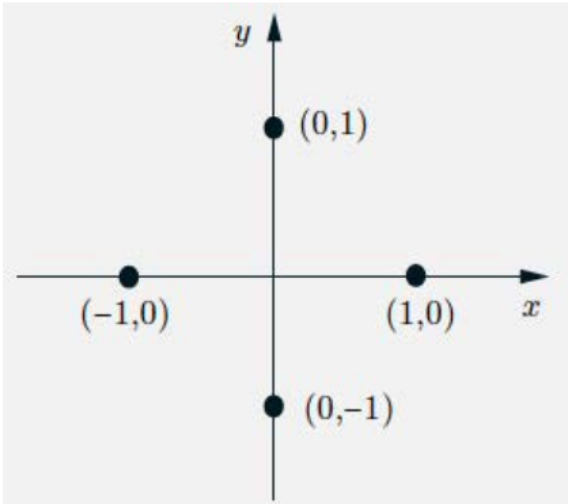
Covariance

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

- 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 $Y = 0$.



Properties

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

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

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

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

Correlation Coefficient: Bounded Dimensionless Metric

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

Correlation Coefficient

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

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

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

Expectation, Covariance, Variance

- $\mathbb{E}(\boldsymbol{X}) := \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_n) \end{pmatrix}$
- Covariance of $\boldsymbol{X} \in \mathbb{R}^n$ and $\boldsymbol{Y} \in \mathbb{R}^m$

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

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

$$\boldsymbol{\Sigma}_{\boldsymbol{X}} := \mathrm{var}[\boldsymbol{X}] = \begin{pmatrix} \mathrm{cov}(X_1, X_1) & \mathrm{cov}(X_1, X_2) & \cdots \mathrm{cov}(X_1, X_n) \\ \vdots & \vdots & \vdots \\ \mathrm{cov}(X_n, X_1) & \mathrm{cov}(X_n, X_2) & \cdots \mathrm{cov}(X_n, X_n) \end{pmatrix}$$

  - We call $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ covariance matrix of $\boldsymbol{X}$.

Data Matrix and Data Covariance Matrix

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

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

- (data) covariance matrix L10(1)

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

Covariance Matrix and Data Covariance Matrix

- Question. Relation between covariance matrix and data covariance matrix?
- Covaiance matrix for a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_D)^\mathsf{T}$,

$$\boldsymbol{\Sigma}_{\boldsymbol{Y}} = \begin{pmatrix} \mathrm{cov}(Y_1, Y_1) & \mathrm{cov}(Y_1, Y_2) & \cdots \mathrm{cov}(Y_1, Y_D) \\ \vdots & \vdots & \vdots \\ \mathrm{cov}(Y_D, Y_1) & \mathrm{cov}(Y_n, Y_2) & \cdots \mathrm{cov}(Y_D, Y_D) \end{pmatrix}$$

- Data convariance matrix $\boldsymbol{S} \in \mathbb{R}^{D \times D}$
  - Each Y_i has N samples $\begin{pmatrix} x_{i,1} & \cdots & x_{i,N} \end{pmatrix}$

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

Properties

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

- $\mathbb{E}(\boldsymbol{X} + \boldsymbol{Y}) = \mathbb{E}(\boldsymbol{X}) + \mathbb{E}(\boldsymbol{Y}) \in \mathbb{R}^n$
- $\mathbb{E}(\boldsymbol{X} - \boldsymbol{Y}) = \mathbb{E}(\boldsymbol{X}) - \mathbb{E}(\boldsymbol{Y}) \in \mathbb{R}^n$
- $\mathrm{var}(\boldsymbol{X} + \boldsymbol{Y}) = \mathrm{var}(\boldsymbol{X}) + \mathrm{var}(\boldsymbol{Y}) + \mathrm{cov}(\boldsymbol{X}, \boldsymbol{Y}) + \mathrm{cov}(\boldsymbol{Y}, \boldsymbol{X}) \in \mathbb{R}^{n \times n}$
- $\mathrm{var}(\boldsymbol{X} - \boldsymbol{Y}) = \mathrm{var}(\boldsymbol{X}) + \mathrm{var}(\boldsymbol{Y}) - \mathrm{cov}(\boldsymbol{X}, \boldsymbol{Y}) - \mathrm{cov}(\boldsymbol{Y}, \boldsymbol{X}) \in \mathbb{R}^{n \times n}$
- Assume $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$.
    - $\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{A}\mathbb{E}(\boldsymbol{X}) + \boldsymbol{b}$
    - $\mathrm{var}(\boldsymbol{Y}) = \mathrm{var}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\,\mathrm{var}(\boldsymbol{X})\boldsymbol{A}^{\mathsf{T}}$
    - $\mathrm{cov}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{A}^{\mathsf{T}}$ (Please prove)

Roadmap

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

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

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

Gaussian Random Vector

- $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^\mathsf{T}$ 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 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^\mathsf{T}$ has a joint pdf of the form:

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

  where $\boldsymbol{\Sigma}$ is symmetric and positive definite.
- We write $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, or $p_{\boldsymbol{X}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Power of Gaussian Random Vectors

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

Marginals and Conditionals of Gaussians

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

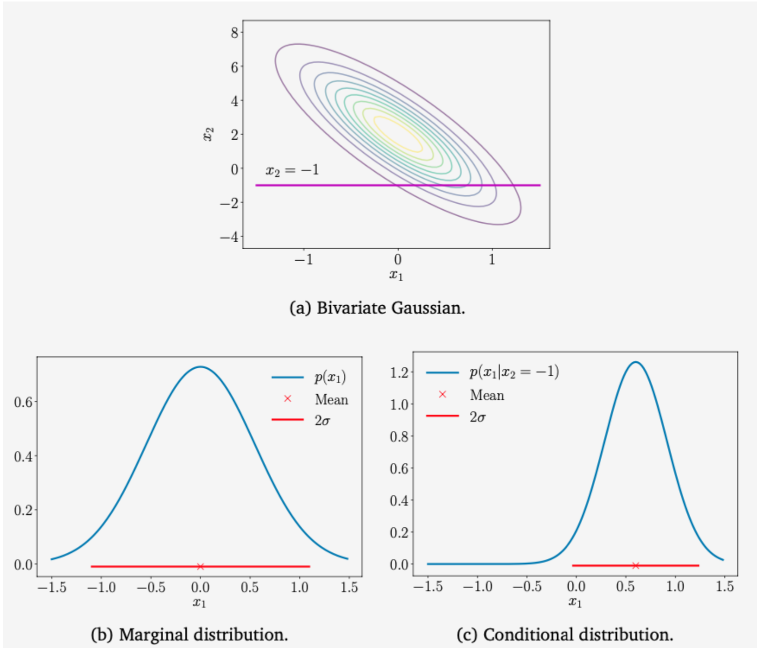
$$\boldsymbol{\Sigma_Z} = \begin{pmatrix} \boldsymbol{\Sigma_X} & \boldsymbol{\Sigma_{XY}} \\ \boldsymbol{\Sigma_{YX}} & \boldsymbol{\Sigma_Y} \end{pmatrix}, \text{ where } \boldsymbol{\Sigma_{XY}} = \mathrm{cov}(\boldsymbol{X}, \boldsymbol{Y}).$$

\- Marginal.

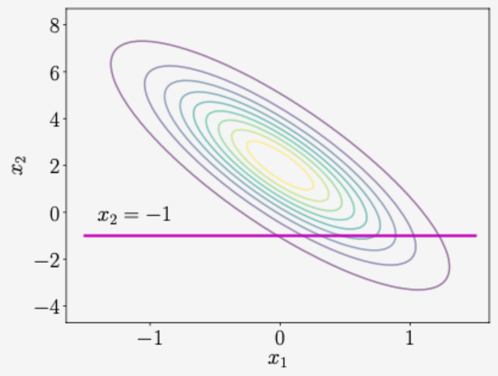
$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \int f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu_x}, \boldsymbol{\Sigma_X})$$

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

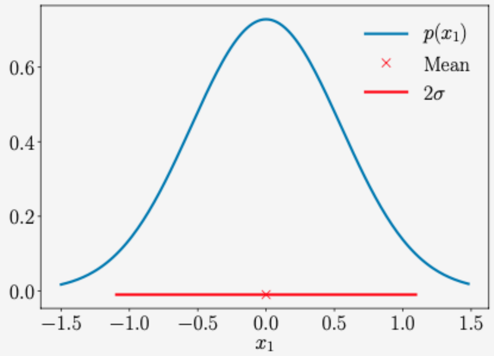
$$\boldsymbol{\mu_{X|Y}} = \boldsymbol{\mu_X} + \boldsymbol{\Sigma_{XY}} \boldsymbol{\Sigma_Y}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu_Y})$$
$$\boldsymbol{\Sigma_{X|Y}} = \boldsymbol{\Sigma_X} - \boldsymbol{\Sigma_{XY}} \boldsymbol{\Sigma_Y}^{-1} \boldsymbol{\Sigma_{YX}}$$



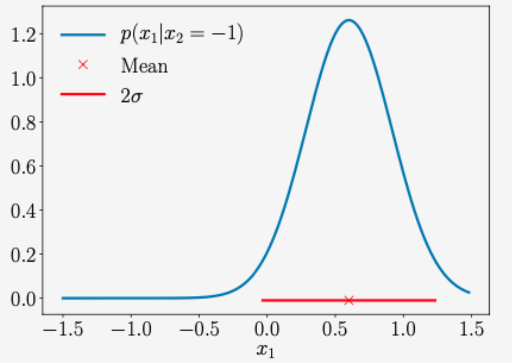
Example 6.6



(a) Bivariate Gaussian.



(b) Marginal distribution.



(c) Conditional distribution.

Product of Two Gaussian Densities

- Lemma. Up to rescaling, 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.

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

Product of Two Gaussian Densities for Random Vectors

- 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:

$$\mathcal{N}\left(\frac{\nu_1\mu_0 + \nu_0\mu_1}{\nu_0 + \nu_1}, \frac{\nu_0\nu_1}{\nu_0 + \nu_1}\right)$$

Sum of Gaussians

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

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

Mixture of Two Gaussian Densities

- $f_1(x)$ 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

- Linear transformation[2] 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 $\boxed{\text{standardized}}$:

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

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

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

[2]Strictly speaking, this is affine transformation.

Linear Transformation for Random Vectors

- $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$, where $\boldsymbol{X} \in \mathbb{R}^n$, $\boldsymbol{Y}, \boldsymbol{b} \in \mathbb{R}^m$, and $\boldsymbol{A} = \mathbb{R}^{m \times n}$

$\implies \quad \boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\Sigma\boldsymbol{A}^{\mathsf{T}})$

Roadmap

Conjugate Prior: Motivation

- Bayesian Inference

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

- 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	Dirchlet

Beta Distribution

Beta distribution

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

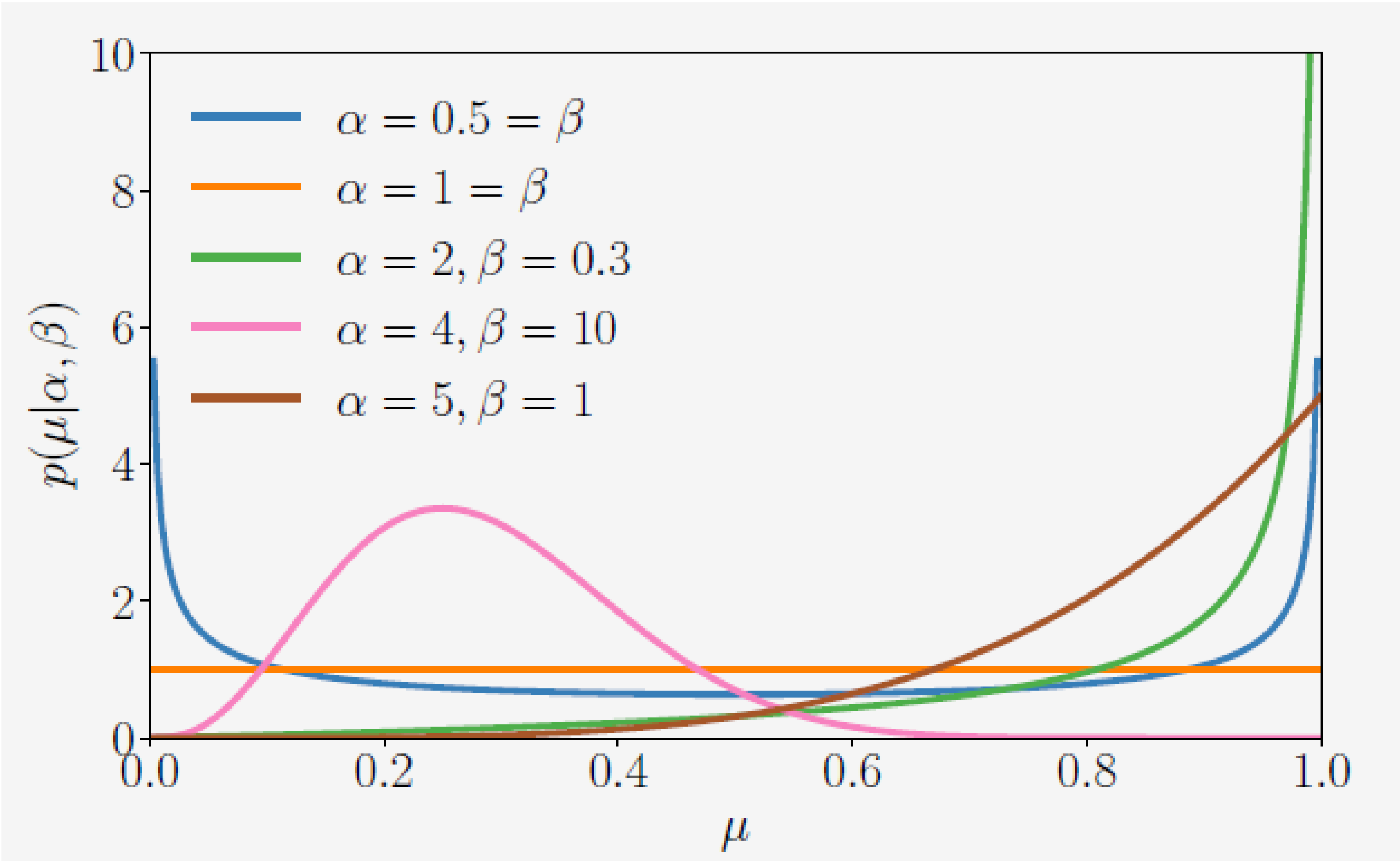
$$f_\Theta(\theta) = \begin{cases} \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, & 0 < \theta < 1, \\ 0, & \text{otherwise,} \end{cases}$$

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 $Uniform[0, 1]$

Beta Distribution



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, $p(x \mid \theta) = \binom{N}{x}\theta^{x}(1-\theta)^{N-x}$ (likelihood)
- Posterior $\propto$ (likelihood) $\times$ (prior)

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

Sufficient Statistics

- A statistic of a random variable $\boldsymbol{X}$ is a deterministic function of $\boldsymbol{X}$.
- Example. For $\boldsymbol{X} = \begin{pmatrix} X_1 & X_2 & \dots & X_n \end{pmatrix}^\mathsf{T}$, the sample mean $T(\boldsymbol{X}) = \frac{1}{N}(X_1 + \cdots + X_n)$ 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 $T = T(\boldsymbol{X})$ is said to be sufficient for $\boldsymbol{X}$ with its pdf or pmf $p_{\boldsymbol{X}}(\boldsymbol{x};\theta)$,[3] if the conditional distribution of $\boldsymbol{X}$ given $T(\boldsymbol{X}) = t$ is independent of θ for all t.

[3]The parameter can be a vector, but we do not use $\boldsymbol{\theta}$ for simplicity.

Poisson Example

- X_1, X_2: independent Poisson variables with common parameter λ.
- Claim. $T(\boldsymbol{X}) = X_1 + X_2$ is a sufficient statistic for inference of λ.
- Joint distribution

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

  $$\mathbb{P}(t) = \frac{(n\lambda)^t}{t!}e^{-n\lambda}$$
- Conditional dist. of X_1 given $X_1 + X_2 = t$

  $$\mathbb{P}(x_1|X_1 + X_2 = t) = \frac{t!}{n^t x_1!(t - x1)!}$$
- Independent of $\lambda \implies T$ is a sufficient statistic.

Fisher-Neyman Factorization Theorem

Factorization Theorem

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

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

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

  $\mathbb{P}_{\boldsymbol{X}}(x_1, \ldots, x_n) = \lambda^{\sum x_i} e^{-n\lambda} / \prod (x_i!)$
- $T(\boldsymbol{X}) = \sum X_i$ is a sufficient statistic.

Exponential Family: Motivation

- 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 (μ, σ^2) from data

L3. Consider a family of distributions which satisfy "nice" properties
  - Example. Exponential family

Exponential Family: Definition

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

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

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

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

Example 6.13 Gaussian as Exponential Family

Example 6.14 Bernoulli as Exponential Family

Roadmap

Knowing Distributions of Functions of RVs

- 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 univarate random variables for simplicity.

CDF-based Technique

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

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

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

$$
\begin{aligned}
F_Y(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(X^2 \leq y) = \mathbb{P}(X \leq \sqrt{y}) = F_X(\sqrt{y}) \\
&= \int_0^{\sqrt{y}} 3t^2 \mathrm{d}t = y^{\frac{3}{2}}, \quad 0 \leq y \leq 1
\end{aligned}
$$

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

How to Get Random Samples of a Given Distribution? (1)

- Assume that $f_X(x) = e^{-x}$ 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 $\boxed{U := F(X)}$, 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.

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

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

How to Get Random Samples of a Given Distribution? (2)

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 $x = F^{-1}(u)$.

Change-of-Variables Technique: Univariate

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

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

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}\int_a^{U^{-1}(y)} f_X(x)\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}y}\int_a^{U^{-1}(y)} f_X(U^{-1}(y))U^{-1'}(y)\mathrm{d}y$$
$$= f_X(U^{-1}(y)) \cdot \frac{\mathrm{d}}{\mathrm{d}y}U^{-1}(y)$$

- Including the case when U is strcitly decreasing,

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

Change-of-Variables Technique: Multivariate

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

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

Example 6.17

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

consider $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$, where $\boldsymbol{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. p

Questions?

References

[1] This lecture slide is mainly based upon `https://yung-web.github.io/home/courses/mathml.html` (made by Prof. Yung Yi, KAIST EE)

[2] Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong. Mathematics for machine learning. Cambridge University Press, 2020.