Probability Basics

Lecture Notes of PRP offered by Dr. Prasad Krishnan, IIIT Hyderabad
prepared by Hari Hara Suthan and Sai Praneeth (corrections, mail to prasad.krishnan@iiit.ac.in)

Monsoon 2018

Class - 1 and 2. (31/07/18 and 3/08/18)

Note. *Reader is recommended to prove the statements in Exercise, Lemmas and Theorems.*

1 Guidelines for a good theory

The development of any Mathematical Theory involves formation of Axioms(Definitions) by observing nature using common sense. From Axioms, by using mathematically logical statements we derive Properties(Theorems). Theorems need formal proof. A good mathematical theory has to satisfy the following conditions.

1. it should be general enough
2. it should be simple enough, i.e., least number of axioms
3. it should be self consistent
4. it should have wide applicability
5. theoretical predictions should match reality

2 Probability Space $(\Omega, \mathcal{F}, P)$

2.1 Set theory prerequisites

Note: The union and intersection of an infinite sequence of sets is defined as follows

- $\bigcup\limits_{i=1}^{\infty} A_i = \{x : x \in A_i \textit{ for at least one } i\}$
- $\bigcap\limits_{i=1}^{\infty} A_i = \{x : x \in A_i \, , \forall \, i\}$
- Other basics like De Morgan's laws can be assumed.

2.2 Probability

A random experiment is an experiment for which outcome cannot be predicted with certainity. We now turn to the idea of probability space.

Definition 2.1. *The probability space $(\Omega, \mathcal{F}, P)$ is a triplet consisting of sample space Ω, Event space $\mathcal{F}$ and probabilities associated to each event in the event space*

1. *The sample space Ω consists of set of all outcomes of a random experiment*
2. *The set $\mathcal{F}$ is a set of subsets of sample space Ω called the event space satisfying the following axioms (called event space axioms)*
   - $\Omega \in \mathcal{F}$
   - *If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$*
   - *If $A_i, i = 1, 2, \cdots$, are countably infinite sequence of events (i.e., each $A_i \in \mathcal{F}$), then $\bigcup\limits_{i=1}^{\infty} A_i \in \mathcal{F}$*
3. *The function $P : \mathcal{F} \to \mathbb{R}$ (called the Probability Measure P) satisfying the following axioms*

- $P(\Omega) = 1$
- $P(A) \geq 0\,,\ \forall A \in \mathcal{F}$
- *For any* $\{A_i \in \mathcal{F} : i \in \mathcal{N}\}$ *which are pairwise disjoint(mutually exclusive) then*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Note. *Outcome and event are different. Outcome is a particular output of a random experiment. Event is a subset of sample space that lies in the event space. Outcome is an element in* Ω *and event is an element in* $\mathcal{F}$*. Probability is defined for events not for outcomes (unless an outcome itself is an event of size 1).*

Example 2.1 (Examples of sample space).

- *Suppose one is interested in knowing whether the toss produces a head or a tail, then the sample space is given by* $\Omega = \{H, T\}$*. Here, as there are only two possible outcomes, the sample space is said to be finite.*
- *Suppose one is interested in the number of tumbles before the coin hits the ground, then the sample space is the set of all natural numbers. In this case, the sample space is countably infinite and is given by* $\Omega = \mathbb{N}$*.*
- *Suppose one is interested in the speed with which the coin strikes ground, then the set of positive real numbers forms the sample space. This is an example of an uncountable sample space, which is given by* $\Omega = \mathbb{R}^+$*.*

Thus for a given random experiment, sample space is defined depending on what one is interested in observing as the outcome.

Example 2.2 (Examples of Event Space).

- *In all of the examples below, we include the events of interest in the event space, and also all the other subsets of* Ω *so that the event space axioms are satisfied.*
- *The smallest event space associated with* Ω *is the collection* $\mathcal{F} = \{\varnothing, \Omega\}$*.*
- *If* A *is any subset of* Ω *then* $\mathcal{F} = \{\varnothing, A, A^c, \Omega\}$ *is an event space.*
- *Consider the experiment of coin-toss. The sample space is* $\Omega = \{H, T, M\}$ *where* H *stands for the outcome heads,* T *for tails and* M *if outcome is neither heads nor tails. If we are interested in the event of getting a head, then* $\{H\}$ *would be our event of interest. The event space for such a case is* $\mathcal{F} = \big\{\varnothing, \{H\}, \{T, M\}, \Omega\big\}$*.*
- *If* A, B *are subsets of* Ω *then* $\mathcal{F} = \{\varnothing, A, B, A^c, B^c, \Omega\}$ *is* ***NOT*** *an event space. It satisfies the* 1^{st} *and* 2^{nd} *axioms of event space but not the* 3^{rd} *one. We know that, any finite union of subsets of* Ω *in* $\mathcal{F}$ *belongs to* $\mathcal{F}$*. Here* $A, B \in \mathcal{F}$ *but* $A \cup B \notin \mathcal{F}$*. A valid event space containing both* A, B *would be,* $\mathcal{F} = \{\varnothing,\ A,\ B,\ A^c,\ B^c,\ A \cup B,\ A^c \cup B,\ A \cup B^c,\ A^c \cup B^c,\ (A \cup B)^c,\ (A^c \cup B)^c,\ (A \cup B^c)^c,\ (A^c \cup B^c)^c,\ A \,\Delta\, B,\ (A \,\Delta\, B)^c,\ \Omega\}$ *(where* $A \Delta B = (A \cap B^c) \cup (A^c \cap B)$*).*

2.2.1 Exercise :

By only using axioms of event space prove the following

1. Show that the $\mathcal{P}(\Omega)$ (*power set* of Ω) which contains all subsets of Ω is an event space.
2. Given a sample space Ω and an event space $\mathcal{F}$ of subsets of Ω show that if $A, B \in \mathcal{F}$, the symmetric difference of A and B i.e $A \,\Delta\, B \in \mathcal{F}$.

Lemma 2.1. *If* $A_i,\ i \in \mathbb{N}$ *are subsets of a set. Then* $\bigcap_{i=1}^{\infty} A_i = \left(\bigcup_{i=1}^{\infty} A_i^c\right)^c$

Note.

- *Proof by induction can't be used when infinities are present.*
- *If* A, B *are two sets and if we want to prove* $A = B$ *then prove* $A \subseteq B$ *and* $B \subseteq A$*.*

Proof. Let $x \in \left(\bigcup_{i=1}^{\infty} A_i^c\right)^c$. So $x \notin \bigcup_{i=1}^{\infty} A_i^c$. This can be written as $x \notin A_i^c, \forall i = 1, 2, \cdots$. So $x \in A_i, \forall i = 1, 2, \cdots$. which gives $x \in \bigcap_{i=1}^{\infty} A_i$. Hence $\left(\bigcup_{i=1}^{\infty} A_i^c\right)^c \subseteq \bigcap_{i=1}^{\infty} A_i$. The other inclusion can be proved by retracing the steps backwards. □

Lemma 2.2. *If* $A_1, A_2, \cdots \in \mathcal{F}$ *then*

a) $\bigcup_{i=1}^{n} A_i \in \mathcal{F}$

b) $\bigcap\limits_{i=1}^{\infty} A_i \in \mathcal{F}$

c) $\bigcap\limits_{i=1}^{n} A_i \in \mathcal{F}$

Proof. a) $\bigcup\limits_{i=1}^{\infty} A_i \in \mathcal{F}$. (from axioms of event space). Let $A_i = \phi, \ \forall i \in \{n+1, n+2, \cdots\}$. Hence $\bigcup\limits_{i=1}^{n} A_i \in \mathcal{F}$.

b) Let $A_i \in \mathcal{F}$. So $A_i^c \in \mathcal{F}$ which gives $\bigcup\limits_{i=1}^{\infty} A_i^c \in \mathcal{F}$. Therefore $\left(\bigcup\limits_{i=1}^{\infty} A_i^c\right)^c \in \mathcal{F}$. From previous lemma we have $\bigcap\limits_{i=1}^{\infty} A_i = \left(\bigcup\limits_{i=1}^{\infty} A_i^c\right)^c$. Hence $\bigcap\limits_{i=1}^{\infty} A_i \in \mathcal{F}$.

c) Let $A_i = \Omega, \ \forall i \in \{n+1, n+2, \cdots\}$. Then $\bigcap\limits_{i=1}^{\infty} A_i = \bigcap\limits_{i=1}^{n} A_i$. Hence $\bigcap\limits_{i=1}^{n} A_i \in \mathcal{F}$. (from part b)

□

Class - 3. (07/08/18)

Lemma 2.3. *By only using axioms of event space and probability prove the following*

a) $P(\emptyset) = 0$ *(where $\emptyset$ is the null set)*

b) $P(\bigcup\limits_{i=1}^{n} A_i) = \sum\limits_{i=1}^{n} P(A_i)$ *where* $\big\{A_i : i \in \{1, 2, ..., n\}\big\}$ *are mutually exclusive*

c) $P(A \cup A^c) = P(A) + P(A^c) = 1$

d) If $A \subseteq B$, *then* $P(A) \le P(B)$

e) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ *(A, B need not be disjoint)*

Proof. a) Let $A_1 = \Omega, \ A_i = \phi, \ \forall i \in \{2, 3, \cdots\}$. We have, if $\{A_i : i \in \mathcal{N}\}$ are mutually exclusive then $P(\bigcup\limits_{i=1}^{\infty} A_i) = \sum\limits_{i=1}^{\infty} P(A_i)$. Substituting values of A_i we get $P(\Omega) = P(\Omega) + \sum\limits_{i=2}^{\infty} P(\phi)$. This can be written as $\sum\limits_{i=2}^{\infty} P(\phi) = 0$. Hence $P(\phi) = 0$. By Axiom of Probability, we must have $P(A_i) \ge 0, \forall A_i \in \mathcal{F}$. If $P(\phi) > 0$ then the infinite sum will result in infinity which contradicts the above result. Hence, we must have $P(\phi) = 0$.

b) Let $A_i = \phi, \ \forall i \in \{n+1, n+2, \cdots\}$. We have, if $\{A_i : i \in \mathcal{N}\}$ are mutually exclusive then $P(\bigcup\limits_{i=1}^{\infty} A_i) = \sum\limits_{i=1}^{\infty} P(A_i)$. Substituting values of A_i we get $P(\bigcup\limits_{i=1}^{n} A_i) = \sum\limits_{i=1}^{n} P(A_i) + \sum\limits_{i=n+1}^{\infty} P(\phi) = \sum\limits_{i=1}^{n} P(A_i)$. $\quad (\because P(\phi) = 0)$

c) A and A^c are mutually exclusive and exhaustive events. So $P(A \cup A^c) = P(A) + P(A^c) = 1$. $(\because P(\Omega) = 1)$

d) $P(B) = P(A \cup (B \setminus A)) = P(A) + P(B \setminus A)$. Hence $P(B) \ge P(A)$. $\quad (\because A \cap (B \setminus A) = \phi)$

e) $A \cup B$ can be written as union of disjoint events in three different ways. $A \cup B = A \cup (B \setminus A) = B \cup (A \setminus B) = (A \setminus B) \cup (A \cap B) \cup (B \setminus A)$. Therefore $P(A \cup B)$ can be written in three different ways as follows (use part b)

$$\begin{aligned} P(A \cup B) &= P(A) + P(B \setminus A). \quad \text{which gives } P(B \setminus A) = P(A \cup B) - P(A) \\ P(A \cup B) &= P(B) + P(A \setminus B). \quad \text{which gives } P(A \setminus B) = P(A \cup B) - P(B) \\ P(A \cup B) &= P(A \setminus B) + P(A \cap B) + P(B \setminus A) \end{aligned}$$

Substitute $P(B \setminus A)$ and $P(A \setminus B)$ in the above equation, which concludes the proof.

□

2.2.2 Examples of Probability spaces

1. **Finite sample space:** Let us consider a coin toss experiment with the probability of getting a head as p and the probability of getting a tail as $(1-p)$. Then, the sample space and the event space are $\Omega = \{H, T\}$ and $\mathcal{F} = \big\{\phi, \{H\}, \{T\}, \Omega\big\}$. The probability measure is, $P(\{H\}) = p$, $P(\{T\}) = 1 - p$. (readers are requested to check whether this is a valid probability space or not, ie they satisfy the axioms of probability space).A probability measure in which we have only two possible outcomes is called a Bernoulli measure.

2. **Countably infinite sample space:** Consider the sample space consisting of discrete arrival times. So $\Omega = \{0, 1, 2, 3, ...\}$ and $\mathcal{F} = \mathcal{P}(\Omega)$. Define the probability measure for an arbitrary event $A \in \mathcal{F}$ as

$$P(A) = \begin{cases} \dfrac{e^{-\lambda}\lambda^k}{k!}, \textit{ for some fixed } \lambda > 0 & \textit{if } |A| = 1 \\ \\ \sum\limits_{k \in A} P(\{k\}), & \textit{otherwise} \end{cases} \tag{1}$$

This probability measure is called a Poisson probability measure with parameter λ.

Lemma 2.4. *Prove that Poisson probability measure is a valid probability measure.*

Proof. Let us check whether this measure satisfies all the axioms of event space or not

(a) It is clear that $P(A) \geq 0, \forall A \in \mathcal{F}$

(b) Consider $A_i \in \mathcal{F}, \forall i = \{1, 2, \cdots\}$ are mutually exclusive events. Let $A = \bigcup\limits_{i=1}^{\infty} A_i$. Then $P(A) = \sum\limits_{k \in A} P(\{k\}) = \sum\limits_{i=1}^{\infty} \sum\limits_{k \in A_i} P(\{k\}) = \sum\limits_{i=1}^{\infty} P(A_i)$

(c) $P(\Omega) = P\left(\bigcup\limits_{k=0}^{\infty} \{k\}\right) = \sum\limits_{k=0}^{\infty} P(\{k\}) = \sum\limits_{k=0}^{\infty} \dfrac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum\limits_{k=0}^{\infty} \dfrac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1$

Hence this is a valid probability measure. □

3. **Uncountably infinite sample space:** Suppose one is interested in measurement of *temperature* or *amplitude of a noise signal* or *the sea-level at some place.* These measurements take real values and so the sample space will be $\Omega = \mathbb{R}$ (or some interval of $\mathbb{R}$ for example $\Omega = [-6, 10]$) which is an uncountable set. First note that the assignment of probabilities which we did for countably infinite case will no longer work in uncountably infinite case. Before finding the probability measure we need to fix the event space $\mathcal{F}$. Typically we are interested in the smallest event space containing all open intervals in $\mathbb{R}$. Such an event space is known as *Borel σ-algebra* or *Borel σ-field.* we let this Borel σ-field to be our event space $\mathcal{F}$. Let the probability measure function be $P : \mathcal{F} \to \mathbb{R}$, where we specify P only for open intervals (a, b) where $a \leq b$ and $a, b \in \mathbb{R} \cup \{-\infty, \infty\}$. In other words, (a, b) is a subset of $\mathbb{R} \cup \{-\infty, \infty\}$ and $(a, b) \in \mathcal{F}$. Consider an integrable function $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) \geq 0 \ \forall \ x \in \mathbb{R}$, $\int\limits_{-\infty}^{\infty} f(x)dx = 1$.

We assign probability to an open interval (a, b) of Ω as $P\big((a, b)\big) = \int\limits_{a}^{b} f(x)dx$.

Lemma 2.5. *Prove that the above function is a valid probability measure.*

Proof. (a) We have $\Omega = (-\infty, \infty)$. Therefore $P(\Omega) = \int\limits_{-\infty}^{\infty} f(x)dx = 1$.

(b) Since $f(x) \geq 0, \forall x \in \mathbb{R}$. Therefore $P\big((a, b)\big) \geq 0$.

(c) Let $A_i \in \mathbb{R}, \forall i \in \{1, 2, \cdots\}$ be mutually exclusive events. Then by definition of integrable function $P\left(\bigcup\limits_{i=1}^{\infty} A_i\right) = \int\limits_{x \in \bigcup\limits_{i=1}^{\infty} A_i} f(x)dx = \sum\limits_{i=1}^{\infty} \int\limits_{x \in A_i} f(x)dx = \sum\limits_{i=1}^{\infty} P(A_i)$.

Hence this is a valid probability measure. □

2.2.3 Exercise

1. A fair coin is tossed until head appears (assume independent tosses). What is the sample space Ω? Show that $P(\Omega) = 1$. What is the event space $\mathcal{F}$? What is the probability of getting a tail eventually? Also, find the same if the coin tossed is biased with probability of getting a head being 0.3.

2. Construct a sample space Ω and probability P to model an unfair dice in which faces 1 and 5 are equally likely, but face 6 has probability $\frac{1}{3}$. Using this model and compute the probability that toss results in face showing even number.

3. An urn contains white and black balls. When two balls are drawn without replacement, suppose the probability that both the balls are white is $\frac{1}{3}$. Find the smallest number of balls in the urn. How small can the total number of balls be if the number of black balls is even?

Class - 4. (10/08/18)

3 Conditional Probability

Definition 3.1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space "The Conditional Probability of an event A given event B" is denoted as $P_B(A)$ or $P(A|B)$ and is defined as $P(A|B) \triangleq \dfrac{P(A \cap B)}{P(B)}$ where $P(B) \neq 0$.*

Note.

- *$P(A|B)$ is just a number which is given a name "conditional probability of A given B". We have to prove it is a valid probability measure satisfying axioms. The probability measure in $P(A \cap B)$ and $P(B)$ is the probability measure defined on Ω in the probability space $(\Omega, \mathcal{F}, P)$.*
- *$P(A|B) = 1 - P(A^c|B) \neq 1 - P(A|B^c)$.*

The following lemma can be easily proved (we leave the proof to the reader).

Lemma 3.1. *Consider a probability space $(\Omega, \mathcal{F}, P : \mathcal{F} \to \mathbb{R})$ and $B \in \mathcal{F}$, then the triplet $(\Omega, \mathcal{F}, P_B)$ (where $P_B(A) = P(A|B)$ is a probability space.*

The following lemma adds another layer of understanding to the idea of conditional probability, where one can view it as a probability measure on a smaller new sample space, i.e., B itself.

Lemma 3.2. *Consider a probability space $(\Omega, \mathcal{F}, P : \mathcal{F} \to \mathbb{R})$ and $B \in \mathcal{F}$, then the triplet $(\Omega \cap B = B, \mathcal{F}_B = \{C \cap B : C \in \mathcal{F}\}, P_B : \mathcal{F}_B \to \mathbb{R})$ is a probability space.*

Proof. We verify that this triplet satisfies the axioms of probability space. To do this we use event space axioms, probability measure axioms and some of the lemmas which we have already proved.

1. Here B is the new sample space
2. Axioms of event space
   (a) We know $\phi \in \mathcal{F}$. Hence $\phi \cap B = \phi \in \mathcal{F}_B$.
   (b) Consider $A \in \mathcal{F}_B$, we have to show that $B \setminus A \in \mathcal{F}_B$. since $A \in \mathcal{F}_B$, there exist $C \in \mathcal{F}$ such that $A = C \cap B$. Since $B, C \in \mathcal{F}$ we should have $A \in \mathcal{F}$. which gives $\Omega \setminus A \in \mathcal{F}$. From the definition o f $\mathcal{F}_B$ we can write $B \cap (\Omega \setminus A) \in \mathcal{F}_B$, which gives $B \setminus A \in \mathcal{F}_B$.
   (c) Consider $A_i \in \mathcal{F}_B, \forall i = \{1, 2, \cdots\}$, so there exists $C_i \in \mathcal{F}$ such that $A_i = C_i \cap B, \forall i \in \{1, 2, \cdots\}$. We have to show that $\bigcup\limits_{i=1}^{\infty} A_i \in \mathcal{F}_B$. Consider $\bigcup\limits_{i=1}^{\infty} A_i = \bigcup\limits_{i=1}^{\infty} (C_i \cap B) = \left(\bigcup\limits_{i=1}^{\infty} C_i\right) \cap B$. Since $\bigcup\limits_{i=1}^{\infty} C_i \in \mathcal{F}$, which gives $\left(\bigcup\limits_{i=1}^{\infty} C_i\right) \cap B \in \mathcal{F}_B$. Therefore $\bigcup\limits_{i=1}^{\infty} A_i \in \mathcal{F}_B$.

   Note.
   - *With respect to the sample space Ω, $A^c = \Omega \setminus A$.*
   - *With respect to new sample space B, $A^c = B \setminus A$.*
3. Axioms of Probability measure
   (a) $P_B(B) = P(B|B) = \dfrac{P(B \cap B)}{P(B)} = \dfrac{P(B)}{P(B)} = 1$
   (b) Consider an arbitrary event $A \in \mathcal{F}_B$, we have to show that $P_B(A) \geq 0$. This follows from the definition of $P_B(A)$, since $P(B) \neq 0$ and $P(E) \geq 0, \forall E \in \mathcal{F}$.
   (c) Consider $A_i \in \mathcal{F}_B, \forall i = \{1, 2, \cdots\}$ be mutually exclusive events. We have to show that $P_B\left(\bigcup\limits_{i=1}^{\infty} A_i\right) = \sum\limits_{i=1}^{\infty} P_B(A_i)$. Consider $P_B\left(\bigcup\limits_{i=1}^{\infty} A_i\right)$ which can be written as $\dfrac{P\left(\left(\bigcup\limits_{i=1}^{\infty} A_i\right) \cap B\right)}{P(B)} = \dfrac{P\left(\bigcup\limits_{i=1}^{\infty} (A_i \cap B)\right)}{P(B)} = \dfrac{\sum\limits_{i=1}^{\infty} P(A_i \cap B)}{P(B)}$, where last equality follows from the fact that if A_i are disjoint then $A_i \cap B$ are also

disjoint and P is the probability measure in the probability space $(\Omega, \mathcal{F}, P)$. Therefore $P_B\left(\bigcup\limits_{i=1}^{\infty} A_i\right) = \sum\limits_{i=1}^{\infty} \frac{P(A_i \cap B)}{P(B)} = \sum\limits_{i=1}^{\infty} P(A_i|B) = \sum\limits_{i=1}^{\infty} P_B(A_i)$.

Therefore $(B, \mathcal{F}_B, P_B)$ is probability space.

□

3.1 Inference problems using Probability

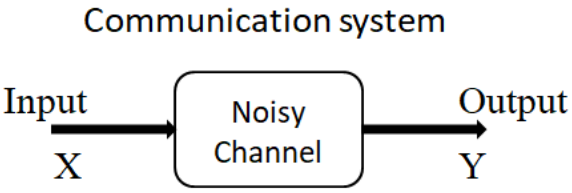

Figure 1: Basic Communication System Block Diagram

The input travels through a noisy channel and is observed as some Y at the receiver. In signal processing, noise is a general term for unwanted (and in general, unknown) changes that a signal may suffer during storage, transmission, or processing. So, Y has got some unwanted signal added to X, assuming the channel is additive noise channel. Let $Y \in A \subseteq \mathbb{R}$ i.e $A \subseteq (-\infty, \infty)$.

Typically, though we don't know the exact value of the noise that was added, we can find some statistics about it. Perhaps we know the approximate range of the noise, and so on.

We now see how conditional probabilities can be useful in such situations. For instance, consider that the inputs to the channel are one of the four possible voltage levels, i.e, $X \in \{-5, 2, 2, 5\}$ and suppose we know that the noise added to the signal X is randomly chosen between $[-1, 1]$. Then it is clear that simply observing the output Y, it is sufficient to know which X it came from. For instance, if the observed signal is 1.6, then the only input which could have caused this output is 2. Thus, we can guess the value of the input without knowing the noise, simply by knowing some statistics about the noise.

Now consider an even more interesting situation where the noise is some random number from $[-2, 2]$. And suppose the output was $Y = 3.7$. Then we see that the input could have been both 5 as well as 2, since $2 + 1.7 = 3.7$, and also $5 - 1.3 = 3.7$. So the receiver has to pick for its input-estimate either 2 or 5. If we had a probability distribution for the noise, then we can ask the question,

- Given that the output was 3.7, what is the most probable input, i.e., for what value of x is the probability

  $$p(X = x|Y = 3.7)$$

  maximized? In other words, what is the value of x for which the probability $p(Noise = 3.7 - x)$ is maximized?

Thus we are attempting a probabilistic inference about an unknown quantity (here, the channel input), using some observations (here, the channel output), and some known statistics about the system (here, the noise probabilities). Conditional probabilities are thus naturally applicable in such situations.

4 Total Probability Theorem

Consider a probability space $(\Omega, \mathcal{F}, P)$. Let $A_i \in \mathcal{F}, \forall i = \{1, 2, \cdots, n\}$ be a partition of Ω. The total probability theorem states that $P(B) = \sum\limits_{i=1}^{n} P(B|A_i)P(A_i)$

Note. *$A_i \in \mathcal{F}, \forall i = \{1, 2, \cdots, n\}$ is a partition of Ω iff*

1. *$\bigcup\limits_{i=1}^{n} A_i = \Omega$ (mutually exhaustive events)*
2. *$A_i \cap A_j = \emptyset, i \neq j$ (mutually exclusive events).*

Proof. Given that $A_i \in \mathcal{F}, \forall i = \{1, 2, \cdots, n\}$ are mutually exclusive and exhaustive events then $(B \cap A_i), \forall i = \{1, 2, \cdots, n\}$ are mutually exclusive and exhausts the event B. Therefore $P(B) = P\left(\bigcup\limits_{i=1}^{\infty} (B \cap A_i)\right) = \sum\limits_{i=1}^{n} P(B \cap A_i) = \sum\limits_{i=1}^{n} P(B|A_i)P(A_i)$. □

Remark. *we can more generally start from a countably infinite set of events $A_i, i \in \mathbb{N}$, which partitions Ω. The theorem still holds.*

Class - 5. (14/08/18)

5 Bayes Theorem

Consider a probability space $(\Omega, \mathcal{F}, P)$. Let $A_i \in \mathcal{F}$, such that $P(A_i) \neq 0, \forall i = \{1, 2, \cdots, n\}$ be (or more generally a countably infinite set of events) a partition of Ω. The Bayes theorem states that $P(A_i|B) = \dfrac{P(B|A_i)P(A_i)}{\left(\sum\limits_{i=1}^{n} P(B|A_i)P(A_i)\right)}$.

Proof. $P(A_i|B) = \dfrac{P(A_i \cap B)}{P(B)} = \dfrac{P(B|A_i)P(A_i)}{P(B)} = \dfrac{P(B|A_i)P(A_i)}{\left(\sum\limits_{i=1}^{n} P(B|A_i)P(A_i)\right)}$. (from the total probability theorem).

□

6 Independent Events

Definition 6.1. *Consider a probability space $(\Omega, \mathcal{F}, P)$. Two events $A, B \in \mathcal{F}$ are said to be independent events if $P(A \cap B) = P(A)P(B)$ otherwise they are said to be dependent events.*

Definition 6.2. *Consider a probability space $(\Omega, \mathcal{F}, P)$. The events $A_1, A_2, A_3, \cdots \in \mathcal{F}$ are said to be independent events if for any subset $\{A_{i_1}, A_{i_2}, \cdots, A_{i_m}\} \subseteq \{A_1, A_2, \cdots\}$, $P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_m}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_m})$.*

Note.

- *Intuitively two events are independent if occurrence of one event doesn't effect the occurrence of the other.*
- *If two events A, B are independent and $P(A) \neq 0, P(B) \neq 0$ then $P(A|B) = P(A)$ and $P(B|A) = P(B)$.*
- *Independent Events and mutually exclusive events are different !*
- *Two events A, B are said to be mutually exclusive(disjoint) events if $A \cap B = \phi$.*

Example 6.1.

1. ***Example for independent and mutually exclusive events:*** *In any random experiment with probability space $(\Omega, \mathcal{F}, P)$ consider two events $A, B \in \mathcal{F}$ where $B = \phi$. Here $P(A \cap B) = P(\phi) = 0$, $P(A)P(B) = 0$. Therefore $P(A \cap B) = P(A)P(B)$ and $A \cap B = \phi$.*
2. ***Example for independent but not mutually exclusive events:*** *Consider a random experiment of tossing a coin coin and rolling a die. $\Omega = \{H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6\}$. Let $\mathcal{F} = \mathcal{P}(\Omega)$. Consider the uniform probability distribution ie $P(\{a\}) = \frac{1}{12}$, where $a \in \Omega$. Let A be the event of getting a head on the coin and B be the event of getting an even number on the die. Therefore $A = \{H1, H2, H3, H4, H5, H6\}$, $B = \{H2, H4, H6, T2, T4, T6\}$. $P(A \cap B) = P(\{H2, H4, H6\}) = \frac{1}{4}$, $P(A) = \frac{1}{2}$ and $P(B) = \frac{1}{2}$. Therefore $P(A \cap B) = P(A)P(B)$ and $A \cap B \neq \phi$.*
3. ***Example for not independent and mutually exclusive events:*** *Consider a random experiment of tossing a coin. $\Omega = \{H, T\}$. Lt $\mathcal{F} = \mathcal{P}(\Omega)$. Consider the uniform probability distribution ie $P(\{H\}) = \frac{1}{2}$ and $P(\{T\}) = \frac{1}{2}$. Let $A = \{H\}$ and $B = \{T\}$. $P(A \cap B) = 0$, $P(A)P(B) = \frac{1}{4}, A \cap B = \phi$. Therefore $P(A \cap B) \neq P(A)P(B)$ and $A \cap B = \phi$.*
4. ***Example for not independent and not mutually exclusive events:*** *Consider a random experiment of rolling a die. $\Omega = \{1, 2, 3, 4, 5, 6\}$. Let $\mathcal{F} = \mathcal{P}(\Omega)$. Consider the uniform probability distribution ie $P(\{a\}) = \frac{1}{6}$, where $a \in \Omega$. Let A be the event of getting a number which is divisible by 3, B be the event of getting a number greater than 2. Therefore $A = \{3, 6\}$, $B = \{3, 4, 5, 6\}$. $P(A \cap B) = P(\{3, 6\}) = \frac{1}{3}$, $P(A) = \frac{1}{3}$ and $P(B) = \frac{2}{3}$ and $P(A)P(B) = \frac{2}{9}$. Therefore $P(A \cap B) \neq P(A)P(B)$ and $A \cap B \neq \phi$.*

6.1 Conditionally Independent Events

Definition 6.3. *Two events A, B are said to be conditionally independent given a third event C if $P\left((A \cap B)|C\right) = P(A|C)P(B|C)$.*

Note.
- *Conditionally Independent events $\not\Rightarrow$ Independent events.*
- *Independent events $\not\Rightarrow$ Conditionally Independent events.*

Example 6.2 (Independent events $\not\Rightarrow$ Conditionally Independent events)**.** *Consider a random experiment of tossing two coins. $\Omega = \{HH, HT, TH, TT\}$. Let $\mathcal{F} = \mathcal{P}(\Omega)$. Consider the uniform probability measure.*

- *Let A denotes the event of getting a head on first coin. Therefore $A = \{HH, HT\}$.*
- *Let B denotes the event of getting a head on second coin. Therefore $B = \{HH, TH\}$.*
- *Let C denotes the event of getting a head on both coins. Therefore $C = \{HH\}$.*
- *Let D denotes the event of getting exactly one head in both coins. Therefore $D = \{HT, TH\}$.*

Let us check independence and conditional independence of events A, B.

- *$P(A \cap B) = \dfrac{1}{4}$ and $P(A)P(B) = \left(\dfrac{1}{2}\right)\left(\dfrac{1}{2}\right) = \dfrac{1}{4}$. Therefore A, B are independent events.*
- *$P\big((A \cap B)|C\big) = \dfrac{P\big((A \cap B) \cap C\big)}{P(C)} = \dfrac{1/4}{1/4} = 1$. Consider $P(A|C) = \dfrac{P(A \cap C)}{P(C)} = 1$. Consider $P(B|C) = \dfrac{P(B \cap C)}{P(C)} = 1$. So $P(A|C)P(B|C) = 1$. Therefore A, B are conditionally independent events given the event C.*
- *$P\big((A \cap B)|D\big) = \dfrac{P\big((A \cap B) \cap D\big)}{P(D)} = \dfrac{0}{1/4} = 0$. Consider $P(A|D) = \dfrac{P(A \cap D)}{P(D)} = \dfrac{1/4}{2/4} = \dfrac{1}{2}$. Consider $P(B|D) = \dfrac{P(B \cap D)}{P(D)} = \dfrac{1/4}{2/4} = \dfrac{1}{2}$. So $P(A|D)P(B|D) = \dfrac{1}{4}$. Therefore A, B are not conditionally independent events given the event D.*

Example 6.3 (Conditionally Independent events $\not\Rightarrow$ Independent events)**.** *(Exercise)*

7 Theorem (Continuity of Probability)

Consider the probability space $(\Omega, \mathcal{F}, P)$. Consider $B_i \in \mathcal{F}, \forall i \in \mathbb{N}$.

a) Let $B_1 \subset B_2 \subset B_3 \cdots$ be a countably infinite sequence of events such that each event contains the prior event. Then $P\left(\bigcup\limits_{i=1}^{\infty} B_i\right) = \lim\limits_{i\to\infty} P(B_i)$.

b) Let $B_1 \supset B_2 \supset B_3 \cdots$ be a countably infinite sequence of events such that each event is contained in the prior event. Then $P\left(\bigcap\limits_{i=1}^{\infty} B_i\right) = \lim\limits_{i\to\infty} P(B_i)$.

Note.
- *$\bigcup\limits_{i=1}^{\infty} B_i = \lim\limits_{n\to\infty} \bigcup\limits_{i=1}^{n} B_i$, similarly $\bigcap\limits_{i=1}^{\infty} B_i = \lim\limits_{n\to\infty} \bigcap\limits_{i=1}^{n} B_i$.*
- *If $B_1 \subseteq B_2$ then $P(B_1) \leq P(B_2)$.*
- *$\dot{\bigcup}$ represents the disjoint union.*
- *In probability theory if we encounter probability and limits then try to apply Continuity of Probability theorem. (if it is applicable).*

Proof. a) Define a new set of events $D_1 = B_1, D_2 = B_2 \setminus B_1, D_3 = B_3 \setminus B_2, \cdots$. Then $D_1, D_2, \cdots$ are disjoint events.

$$P\left(\bigcup_{i=1}^{\infty} B_i\right) = P\left(\dot{\bigcup}_{i=1}^{\infty} D_i\right) = \sum_{i=1}^{\infty} P(D_i). \tag{2}$$

$$\lim_{n\to\infty} P(B_n) = \lim_{n\to\infty} P\left(\dot{\bigcup}_{i=1}^{n} D_i\right) = \lim_{n\to\infty} \sum_{i=1}^{n} P(D_i) = \sum_{i=1}^{\infty} P(D_i) \tag{3}$$

From Equations (2) and (3) we have $P\left(\bigcup\limits_{i=1}^{\infty} B_i\right) = \lim\limits_{n\to\infty} P(B_n) = \lim\limits_{i\to\infty} P(B_i)$. (changing the running index from n to i).

b) Consider $A_i = B_i^c$ then the given set events can be written as $A_1 \subset A_2 \subset A_3 \subset \cdots$. Now apply the previous result, $P\left(\bigcup\limits_{i=1}^{\infty} A_i\right) = \lim\limits_{i\to\infty} P(A_i)$. By substituting $A_i = B_i^c$ we get, $P\left(\bigcup\limits_{i=1}^{\infty} B_i^c\right) = \lim\limits_{i\to\infty} P(B_i^c)$. By applying De-morgans law we get $P\left(\left(\bigcap\limits_{i=1}^{\infty} B_i\right)^c\right) = \lim\limits_{i\to\infty} P(B_i)^c$, which can be written as $\left(1 - P\left(\bigcap\limits_{i=1}^{\infty} B_i\right)\right) = \lim\limits_{i\to\infty} (1 - P(B_i))$. Hence $P\left(\bigcap\limits_{i=1}^{\infty} B_i\right) = \lim\limits_{i\to\infty} P(B_i)$. (Here we have used the facts that $P(E) + P(E^c) = 1$ and $\lim(constant) = constant$).

□

Note. *The next theorem is not discussed in the class.*

Theorem 7.1. *Let $A_i, i = 1, 2, \cdots, n$ be set of events corresponding to the probability space $(\Omega, \mathcal{F}, P)$. Then*

a) $P\left(\bigcap\limits_{i=1}^{n} A_i\right) = \prod\limits_{i=1}^{n} P(A_i|A_{i-1}, \cdots, A_1)$

b) $P\left(\bigcap\limits_{i=1}^{\infty} A_i\right) = \prod\limits_{i=1}^{\infty} P(A_i|A_{i-1}, \cdots, A_1)$

Proof. a) **Base Case:** Let $i = 1$. Then $P(A_1) = P(A_1|A_1) = P(A_1)$.

Induction Step: Consider the given statement is true for $n = k$. Therefore

$$P\left(\bigcap_{i=1}^{k} A_i\right) = \prod_{i=1}^{k} P(A_i|A_{i-1}, \cdots, A_1)$$

$$\begin{aligned} \textit{Consider } P\left(\bigcap_{i=1}^{k+1} A_i\right) &= P\left(\left(\bigcap_{i=1}^{k} A_i\right) \bigcap A_{k+1}\right) \\ &= P\left(\bigcap_{i=1}^{k} A_i\right) P\left(A_{k+1}|\bigcap_{i=1}^{k} A_i\right) \\ &= \prod_{i=1}^{k} P\left(A_i|A_{i-1}, \cdots, A_1\right) P\left(A_{k+1}|\bigcap_{i=1}^{k} A_i\right) \\ &= \prod_{i=1}^{k+1} P\left(A_i|A_{i-1}, \cdots, A_1\right) \end{aligned}$$

Therefore the statement is true for $n = k + 1$. Hence the proof.

b) (Proof by mathematical induction is applicable only to countably finite number of events).

We know that $A_1 \supset (A_2 \cap A_1) \supset (A_3 \cap A_2 \cap A_1) \supset \cdots \cap (A_i \cap A_{i-1} \cdots \cap A_1)$. Let $B_i = \bigcap\limits_{j=1}^{i} A_j$. Then $\bigcap\limits_{j=1}^{i} B_j = \bigcap\limits_{j=1}^{i} A_j$. Since $B_1 \supset B_2 \supset B_3 \supset \cdots$ from the continuity of probability theorem we have

$$\begin{aligned} P\left(\bigcap_{i=1}^{\infty} B_i\right) &= \lim_{i\to\infty} P(B_i) \\ P\left(\bigcap_{i=1}^{\infty} A_i\right) &= \lim_{i\to\infty} P\left(\bigcap_{j=1}^{i} A_j\right) \\ &= \lim_{i\to\infty} \prod_{j=1}^{i} P(A_j|A_{j-1}, \cdots, A_1) \textit{ (from first part)} \\ &= \prod_{j=1}^{\infty} P(A_j|A_{j-1}, \cdots, A_1) \\ &= \prod_{i=1}^{\infty} P(A_i|A_{i-1}, \cdots, A_1) \textit{ (changing the running index from } j \textit{ to } i \textit{)} \end{aligned}$$

Hence the proof.

□

Random Variables

Lecture Notes of PRP offered by Dr. Prasad Krishnan, IIIT Hyderabad
prepared by Hari Hara Suthan and Sai Praneeth.Ch (corrections, mail to prasad.krishnan@iiit.ac.in)

Monsoon 2018

Class 6. (17/08/18)

Note. *Reader is recommended to prove the statements in Exercise, Lemmas and Theorems.*

1 Random Variables

Consider a probability space corresponding to some random experiment. In many occasions, we may be interested only in events of a particular kind, but not in all events. For instance, consider the sample space of a random experiment in which we switch on a digital source (one which generates 0s and 1s) and obtain 10000 samples. Suppose we are not interested in the precise 10000 samples, but only in the number of 1s we obtain. Thus, considering a probability space over the power set of the sample space containing all the 2^{10000} possible outcomes is overkill, given that we only need to track subsets of outcomes indicating the 10001 different possible number of 1s in the outcome.

For such situations, the notion of a *random variable* is quite useful, as we shall see. Furthermore, we shall also see that we will be able model a variety of random quantities of interest in different disciplines using the idea of random variables, and study them via the well-developed tools of mathematics applicable to the real-number system. We therefore begin our formal study of real-valued random variables with this understanding.

Definition 1.1. *Consider a probability space $(\Omega, \mathcal{F}, P)$. A function $X : \Omega \to \mathbb{R}$ is said to be a (real-valued) random variable if $X^{-1}\big((-\infty, x]\big) \in \mathcal{F}, \forall x \in \mathbb{R}$, where for any $A \subset \mathbb{R}$, $X^{-1}(A)$ is defined as follows*

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\},$$

i.e., $X^{-1}(A)$ is the set of all pre-images of all the individual elements of A.

Equivalently, a random variable X (w.r.t Ω) is also defined as any *measurable* function from $\Omega \to \mathbb{R}$, where the word *measurable* means (roughly) that the preimages of an event-space of $\mathbb{R}$ should be an event space in Ω (this is just a rough definition, unless we want to go deeper into mathematics to formalize this). To put it simply, it means that for X to be a random variable, we want $X^{-1}(A) \in \mathcal{F}$, for all $A \subset \mathbb{R}$ of interest. If this happen then X is called a measurable function. The formal definition above is more accurate.

Following definition 1.1, we develop the following 'short-hand' for referring to certain events of interest. For any $x \in \mathbb{R}$, let us define

$$(X \leq x) \triangleq \{\omega \in \Omega : X(\omega) \leq x\},$$
$$(X = x) \triangleq \{\omega \in \Omega : X(\omega) = x\},$$

and so on. In other words, the notation $P(X \leq x)$ actually means $P\{\omega \in \Omega : X(\omega) \leq x\}$, which is well-defined according to the probability measure P defined on the event space $\mathcal{F}$.

Remark. *Non-real-valued random variable X can also be defined similarly. For instance, a random variable whose co-domain is $\mathbb{C}$ is a complex-valued random variable, and so on. However, that the function should necessarily be measurable. We mostly deal with real valued random variables in this course.*

Example 1.1 (Example for random variable and non random variable)**.** *Let $\Omega = \{a, b, c\}$ be the sample space and $\mathcal{F} = \{\phi, \{a\}, \{b, c\}, \Omega\}$ be the event space. Define a function $X : \Omega \to \mathbb{R}$ such that*

$$X(\omega) = \begin{cases} 0 & \omega = a \\ 1 & \omega = b \text{ or } \omega = c \end{cases}$$

It is easy to see that

$$X^{-1}((-\infty, x]) = \begin{cases} \phi & x < 0 \\ \{a\} & 0 \leq x < 1 \\ \Omega & 1 \leq x < \infty \end{cases}$$

Here $X^{-1}\big((-\infty, x]\big) \in \mathcal{F}, \forall x \in \mathbb{R}$. Hence X is a random variable with respect to the given event space $\mathcal{F}$.

Let's define another function $Y : \Omega \to \mathbb{R}$ such that

$$Y(\omega) = \begin{cases} 0 & \omega = b \\ 1 & \omega = a \ or \ \omega = c \end{cases}$$

It is easy to see that

$$Y^{-1}((-\infty, y]) = \begin{cases} \phi & y < 0 \\ \{b\} & 0 \leq y < 1 \\ \Omega & 1 \leq y < \infty \end{cases}$$

Here $Y^{-1}((-\infty, 0.5]) = \{b\} \notin \mathcal{F}$. Hence Y is not a random variable with respect to the given event space $\mathcal{F}$.

Note. *Here both X, Y are random variables with respect to the event space $\mathcal{F}_1 =$ power set (Ω).*

Remark. *Any real valued function will be a random variable with respect to the largest event space $\big($power set$(\Omega)\big)$. (Why? - Try to answer, dear reader!)*

The following theorem classifies the reason why it is sufficient to consider only intervals of the form $(-\infty, x], \forall x \in \mathbb{R}$ in the definition of random variable.

Theorem 1.1. *Consider a probability space $(\Omega, \mathcal{F}, P)$. Let X be the random variable defined on Ω. If $(X \leq x) \in \mathcal{F}, \forall x \in \mathbb{R}$, then $\forall a, b \in \mathbb{R}$ such that $a < b$*

1. *$(X < a) \in \mathcal{F}$.*
2. *$(X = a) = (X \leq a) \setminus (X < a) \in \mathcal{F}$.*
3. *$(X \in (a, b]) = (X \leq b) \setminus (X \leq a) \in \mathcal{F}$.*
4. *$(X \in [a, b]) = (X \leq b) \setminus (X < a) \in \mathcal{F}$.*
5. *$(X \in [a, b)) = (X < b) \setminus (X < a) \in \mathcal{F}$.*
6. *$(X \in (a, b)) = (X < b) \setminus (X \leq a) \in \mathcal{F}$.*

Note. *From axioms of event space we have that, if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$. If $A, B \in \mathcal{F}$ then $A \cup B, A \cap B \in \mathcal{F}$. We can write $A \setminus B$ as $A \cap B^c$.*

Proof. Consider two events $(X \leq a)$ and $(X \leq b)$ in the event space $\mathcal{F}$.

1. Let $a_i = a - \dfrac{1}{i}, \forall i \in \{1, 2, \cdots\}$, then $(X \leq a_i) \in \mathcal{F}, \forall i \in \{1, 2, \cdots\}$. From axioms of event space we have $\bigcup\limits_{i=1}^{\infty} (X \leq a_i) \in \mathcal{F}$. It is clear that $\bigcup\limits_{i=1}^{\infty} (X \leq a_i) = (X < a)$. Therefore $(X < a) \in \mathcal{F}$. Simillarly $(X < b) \in \mathcal{F}$.
2. Since $(X \leq a), (X < a) \in \mathcal{F}$, we can write $(X = a) \in \mathcal{F}$.
3. Since $(X \leq a), (X \leq b) \in \mathcal{F}$, we can write $(X \in (a, b]) \in \mathcal{F}$.
4. Since $(X < a), (X \leq b) \in \mathcal{F}$, we can write $(X \in [a, b]) \in \mathcal{F}$.
5. Since $(X < a), (X \ < b) \in \mathcal{F}$, we can write $(X \in [a, b)) \in \mathcal{F}$.
6. Since $(X \leq a), (X < b) \in \mathcal{F}$, we can write $(X \in (a, b)) \in \mathcal{F}$.

□

Lemma 1.2. *Consider a probability space $(\Omega, \mathcal{F}, P)$. Let X be the random variable defined on Ω. Let $(X \leq x) \in \mathcal{F}, \forall x \in \mathbb{R}$. Prove that it is sufficient to know $P(X \leq x), \forall x \in \mathbb{R}$ to calculate probability of any interval in $\mathbb{R}$.*

Proof. From the axioms of probability measure we have $P(A \dot\cup B) = P(A) + P(B)$, where $A, B \in \mathcal{F}$ and $\dot\cup$ denotes the disjoint union. Here we have $P(X \leq x), \forall x \in \mathbb{R}$.

1. $P(X < a) = P\left(\dot\bigcup_{i=1}^{\infty}\left(X \leq a - \dfrac{1}{i}\right)\right) = \lim\limits_{i \to \infty} P\left(X \leq a - \dfrac{1}{i}\right)$. (from Continuity of probability theorem).
2. $P(X \leq a) = P((X < a)\dot\cup(X = a)) = P(X < a) + P(X = a)$. Hence $P(X = a) = P(X \leq a) - P(X < a)$.

3. $P(X \leq b) = P\left((X \leq a)\dot{\cup}(X \in (a,b])\right) = P(X \leq a) + P(X \in (a,b])$. Hence $P(X \in (a,b]) = P(X \leq b) - P(X \leq a)$.

4. $P(X \leq b) = P\left((X < a)\dot{\cup}(X \in [a,b])\right) = P(X < a) + P(X \in [a,b])$. Hence $P(X \in [a,b]) = P(X \leq b) - P(X < a)$.

5. $P(X < b) = P\left((X < a)\dot{\cup}(X \in [a,b))\right) = P(X < a) + P(X \in [a,b))$. Hence $P(X \in [a,b)) = P(X < b) - P(X < a)$.

6. $P(X < b) = P\left((X \leq a)\dot{\cup}(X \in (a,b))\right) = P(X \leq a) + P(X \in (a,b))$. Hence $P(X \in (a,b)) = P(X < b) - P(X \leq a)$.

Therefore it is sufficient to know $P(X \leq x), \forall x \in \mathbb{R}$ to calculate probability of any interval in $\mathbb{R}$. □

Corollary 1.2.1. *With respect to the Borel $\sigma-$ algebra of $\mathbb{R}$ to ensure pre-images of all intervals of $\mathbb{R}$ as events in the probability space $(\Omega, \mathcal{F}, P)$ and to calculate there respective probabilities it is sufficient*

- *to ensure $(X \leq x) \in \mathcal{F}, \forall x \in \mathbb{R}$*
- *to know $P(X \leq x), \forall x \in \mathbb{R}$.*

Class 7. (24/08/18)

2 Cumulative Distribution Function(CDF)(F_X)

From the above corollary it is sufficient to know $P(X \leq x)$ to calculate all relevant properties. Hence we give a special name to this function as Cumulative Distribution Function as it finds cumulative(collective) probability of of all the events upto a particular point x. Even though the random experiment, $\Omega, \mathcal{F}, P$ are unknown we can completely characterize X by just knowing its CDF

Definition 2.1. *Consider a probability space $(\Omega, \mathcal{F}, P)$. Let X be a real valued random variable. A function $F_X : \mathbb{R} \to \mathbb{R}$ is said to be a Cumulative Distribution Function(CDF) of the random variable X if $F_X(x) = P(X \leq x)$.*

2.1 Properties of the CDF of a RV

Theorem 2.1 (Properties of CDF)**.** *The CDF $F_X(x) = P(X \leq x)$ of the random variable X satisfies the following properties*

1. *F_X is a non-decresing funciton of $x \in \mathbb{R}$.*
2. *$\lim\limits_{x \to -\infty} F_X(x) = 0$.*
3. *$\lim\limits_{x \to \infty} F_X(x) = 1$.*
4. *F_X is a right continuous function.*

Proof. 1. Let $x_1, x_2 \in \mathbb{R}$ such that $x_1 < x_2$. Consider $F_X(x_2) = P(X \leq x_2) = P((X \leq x_1)\dot{\cup}(X \in (x_1, x_2])) = P((X \leq x_1) + P(X \in (x_1, x_2]) = F_X(x_1) + P(X \in (x_1, x_2])$, Since $P(X \in (x_1, x_2]) \geq 0$ we can write $F_X(x_2) \geq F_X(x_1)$. Therefore If $x_1 < x_2$ then $F_X(x_1) \leq F_X(x_2)$. Hence F_X is a non-decreasing function of $x \in \mathbb{R}$.

2. Define $B_i = (X \leq i), \forall i \in \mathbb{Z}$. Then we can write $B_{-1} \supset B_{-2} \supset \cdots$. It is clear that $\left(\bigcap\limits_{i=1}^{\infty} B_{-i}\right) = \phi$. By continuity of probability theorem we have $\lim\limits_{n\to\infty} P(B_{-n}) = P\left(\bigcap\limits_{i=1}^{\infty} B_{-i}\right) = P(\phi) = 0$. Therefore, we can write $\lim\limits_{n\to\infty} P(B_{-n}) = \lim\limits_{n\to\infty} P(X \leq -n) = \lim\limits_{n\to-\infty} P(X \leq n) = \lim\limits_{x\to-\infty} P(X \leq x)$. Now, since $P(X \leq x) = F_X(x)$. therefore we have proved $\lim\limits_{x\to-\infty} F_X(x) = 0$.

3. Define $B_i = (X \leq i), \forall i \in \mathbb{Z}$. Then we can write $B_1 \subset B_2 \subset \cdots$. It is clear that $\left(\bigcup\limits_{i=1}^{\infty} B_i\right) = \Omega$. By continuity of probability theorem we have $\lim\limits_{n\to\infty} P(B_n) = P\left(\bigcup\limits_{i=1}^{\infty} B_i\right) = P(\Omega) = 1$. We can write $\lim\limits_{n\to\infty} P(B_n) = \lim\limits_{n\to\infty} P(X \leq n) = \lim\limits_{x\to\infty} P(X \leq x)$. Using the fact that $P(X \leq x) = F_X(x)$, therefore we have proved $\lim\limits_{x\to\infty} F_X(x) = 1$.(Here $n \to \infty$ is a countable sequence and $x \to \infty$ is a uncountable sequence).

4. Let $B_i = \left(X \leq x + \frac{1}{i}\right), i \in \mathbb{N}$. Then $B_1 \supset B_2 \supset \cdots$. It is clear that $\bigcap_{i=1}^{\infty} B_i = (X \leq x)$. By using continuity of probability we can write $F_X(x) = P(X \leq x) = P\left(\bigcap_{i=1}^{\infty} B_i\right) = \lim_{n\to\infty} P(B_n) = \lim_{n\to\infty} P\left(X \leq x + \frac{1}{n}\right) = \lim_{\epsilon\to 0} P(X \leq x + \epsilon) = \lim_{\epsilon\to 0} F_X(x+\epsilon)$. Therefore $F_X(x) = \lim_{\epsilon\to 0} F_X(x+\epsilon)$. Hence F_X is a right continuous function.

□

Why should we specifically claim only right-continuity? We now show that left continuity does not automatically hold, using the following lemma and corollary.

Lemma 2.2. *For any $x \in \mathbb{R}$,*

$$P(X = x) = F_X(x) - P(X < x) = F_X(x) - \lim_{\epsilon\to 0} F_X(x-\epsilon)$$

Proof. The first inequality in the statement follows because $(X = x)\dot{\cup}(X < x) = (X \leq x)$. Thus we prove the second equality only.

Let $B_i = \left(X \leq x - \frac{1}{i}\right), i \in \mathbb{N}$. Then $B_1 \subset B_2 \subset \cdots$. It is clear that $\bigcup_{i=1}^{\infty} B_i \neq (X \leq x)$, since the event $(X = x)$ is not included in any of the B_is. On the other hand, $\bigcup_{i=1}^{\infty} B_i = (X < x)$, as for any small $\epsilon > 0$, the event $(X \leq x - \epsilon)$ is included in the event $(X \leq x - \frac{1}{i})$ for any $i \geq \frac{1}{\epsilon}$.

Thus, by using the theorem of continuity of probability, we can write

$$P(X < x) = P\left(\bigcup_{i=1}^{\infty} B_i\right) = \lim_{n\to\infty} P(B_n) = \lim_{n\to\infty} P\left(X \leq x - \frac{1}{n}\right) = \lim_{\epsilon\to 0} P(X \leq x - \epsilon) = \lim_{\epsilon\to 0} F_X(x-\epsilon).$$

Therefore $P(X < x) = \lim_{\epsilon\to 0} F_X(x-\epsilon)$. Hence F_X need not be a left continuous function unless $P(X = 0) = 0$.

□

By the above lemma, we thus have the following corollary.

Corollary 2.2.1. *The CDF $F_X(x)$ is continuous (i.e., left-continuous) if and only if $P(X = x) = 0, \forall x \in \mathbb{R}$.*

Proof. $F_X(x)$ is already right-continuous. For left continuity, we need that

$$F_X(x) = \lim_{\epsilon\to 0} F_X(x-\epsilon),$$

where $P(X < x) = \lim_{\epsilon\to 0} F_X(x-\epsilon)$. The proof follows from Lemma 2.2. □

Indeed, in the next section, we will study random variables which are *continuous* and others which are discrete (having non-continuous CDFs).

2.2 Why we can study RVs through their CDFs alone?

The following theorem shows that *any* function which satisfies the CDF properties, is infact a CDF of *some* random variable.

Theorem 2.3. *Let $F : \mathbb{R} \to \mathbb{R}$ be any function satisfying*

1. *For $a \leq b$, $a, b \in \mathbb{R}, F_X(a) \leq F_X(b)$.*
2. *$\lim_{x\to -\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$.*
3. *$\lim_{\epsilon\to 0} F_X(x+\epsilon) = 0$. (right continuous).*

then there exists some random variable X such that $F_X = F$.

Proof. proof is beyond the scope of this course. □

Remark. *Theorem 2.3 is a special case of the so-called Kolmogorov's extension theorem. It is important to mention this here, because of the following reason. Since we can capture the essential probabilistic structure in the random variable by its cdf, and hence discussing about functions which satisfy the properties as in Theorem 2.3, we are indirectly discussing random variables themselves.*

Class 8. (28/08/18)

3 Types of random variables

As with all things under the sun, we like to classify random variables as well. This presentation of the classification is based on the structure of the CDF. However, many books do a simpler classification based on the values taken by the random variable (we mention this in parantheses).

3.1 Continuous Random Variables

Definition 3.1. *A random variable X is said to be a* ***continuous type random variable*** *if $F_X(x)$ is a continuous function of x (roughly speaking, X can take values in an uncountable set).*

We mostly consider continuous random variables whose F_X is also differentiable, in which case $\frac{dF_X(x)}{dx}$ is well defined. Therefore we give it a special name called the **probability density function(pdf)**, and a special notation $f_X(x)$. Thus we have the pdf of a continuous random variable X as

$$f_X(x) \triangleq \frac{dF_X(x)}{dx}.$$

We can show without much difficulty that $f_X(x)$ satisfies the following properties.

- $f_X(x) \geq 0, \forall x$.
- $\int\limits_{-\infty}^{\infty} f_X(x)dx = 1$ (this property means that the integral can be computed and the value is equal to 1).

And indeed, as we would expect, similar to Theorem 2.3, if we have a function $f(x)$ satisfying the above two properties, then there exists some random variable X such that $f(x)$ is the pdf of the random variable X. Therefore, more often than not, we rely upon the pdf $f_X(x)$ to describe the continuous random variable X.

By definition, we know how to obtain $f_X(x)$ if we know the CDF $F_X(x)$. Now, suppose we the pdf $f_X(x)$. Then the following expressions naturally follow by definition of $f_X(x)$.

We can further obtain the following properties for a continuous random variable X.

1. $P(X = x) = 0, \forall x \in \mathbb{R}$. (From Lemma 2.2.1 we have, $F_X(x)$ is continuous iff $P(X = x) = 0, \forall x \in \mathbb{R}$).
2. For any $a \in \mathbb{R}, F_X(a) = \int\limits_{-\infty}^{a} f_X(x)dx$.
3. $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a) = \int\limits_{-\infty}^{b} f_X(x)dx - \int\limits_{-\infty}^{a} f_X(x)dx = \int\limits_{a}^{b} f_X(x)dx$.
4. $P(a < X \leq b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X < b) = \int\limits_{a}^{b} f_X(x)dx$. (by the first property).
5. $P(x < X \leq x+\Delta x) \approx f_X(x)\Delta x$, where Δx is infinitesimally small (This is simply by definition of the integral (in the Reimann sense)).

Remark. *With respect to point (1) above, note that $P(event) = 0$, doesn't necessarily mean that the event never happens. For instance, consider a random experiment of selecting a point in the interval $[0.1]$ of real numbers and thus $\Omega = [0, 1]$. The performance of the experiment ensures some outcome in $[0, 1]$. Consider the uniform probability measure on $[0, 1]$. Hence $P(any\ outcome) = 0$, however we do get some outcomes.*

Note (Very Important). *The pdf at $x \in \mathbb{R}$, $f_X(x)$, is* ***not*** *equal to $P(X = x)$. The pdf $f_X(x)$ is rather like the rate of change of probability of the random variable X with x. That is why if we need the probability of X being in some interval, we have to integrate the pdf $f_X(x)$.*

3.2 Discrete Random Variables

We now come to the notion of discrete random variables.

Definition 3.2. *A random variable X is said to be a* ***discrete type random variable*** *if $F_X(x)$ is a step function (or equivalently, a staircase function) (roughly speaking, X can take a finite or a countably infinite number of values).*

Since $F_X(x)$ is in the shape of a staircase function, it takes countably many steps (or discontinues), and is flat elsewhere. This is the reason we say that X takes only a countable set of values. Let us call these values as $\{x_i : i \in S\}$, where S is some countable set.

Since X takes values only in $\{x_i : i \in S\}$, we have $P(X = x) = 0, \forall x \notin \{x_i : i \in S\}$. What about $P(X = x_i)$? Clearly, we know from Lemma 2.2 that $P(X = x_i) = F_X(x_i) - P(X < x_i)$. Thus, it is only for some $x_i \in \{x_i : i \in S\}$ that we can have discontinuities in $F_X(x)$, i.e., $P(X = x_i) \geq 0$.

These values of $P(X = x_i)$ are thus special for us, and at least some of them are non-zero, and hence we have to define a *probability mass function* to capture these.

Definition 3.3 (Probability mass function (PMF)). *Consider a discrete type random variable X taking values in $\{x_i : i \in S\}$ for some countable set S. The* ***probability mass function*** *of X is denoted by $P_X : \mathbb{R} \to \mathbb{R}$ and is defined as $P_X(x_i) \triangleq P(X = x_i)$ and $P_X(x) = 0, \forall x \notin \{x_i : i \in S\}$.*

Clearly, the PMF satisfies the following properties, and indeed any function which satisfies the below properties can serve as the PMF of some discrete random variable X which takes values from a countable set of values $\{x_i : i \in S\}$ (where S is a countable set)

- $P_X(x_i) \geq 0, \forall x_i$.
- $\sum\limits_{i \in S} P_X(x_i) = 1$.

The following properties of the PMF of a discrete random variable X and its CDF can be easily proved from the definition.

1. For any $a \in \mathbb{R}$,
$$F_X(a) = P(X \leq a) = \sum_{x_i \leq a} P(X = x_i) = \sum_{x_i \leq a} P_X(x_i).$$
2. $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a)$.
3. $P(a \leq X \leq b) = F_X(b) - F_X(a) + P(X = a)$.
4. $P(a < X < b) = F_X(b) - F_X(a) - P(X = b)$.
5. $P(a \leq X < b) = F_X(b) - F_X(a) + P(X = a) - P(X = b)$.

3.3 Mixed Random Variable

Definition 3.4. *A random variable X is said to be* ***mixed type random variable*** *if $F_X(x)$ has both countable number of discontinuities and is also continuous in some intervals.*

We will not pursue Mixed Random Variables much in this course, but essentially the techniques applied to both discrete and continuous random variables apply in this case.

Class 9. (31/08/18) In many situations, we may have some function of a random variable, and then we are interested in studying the probabilities of values taken by such a function. Hence we give the following definition.

4 Measurable function of random variable

Lemma 4.1 (A (measurable) function of random variable is a random variable). *Let $X : \Omega \to \mathbb{R}$ be a random variable. Let $g : \mathbb{R} \to \mathbb{R}$ be a measurable function (i.e., a function such that preimages of event spaces in $\mathbb{R}$ gives an event space of $\mathbb{R}$). Consider the composition $g(X)$ denoted by Y, which is a function from $\Omega \to \mathbb{R}$ defined as $Y(\omega) = g(X(\omega))$. Then Y is a random variable.*

Remark. *The above lemma says that the following set*

$$(Y \leq y) \triangleq \{\omega \in \Omega : Y(\omega) \leq y\}$$

is an event in the original probability space for every $y \in \mathbb{R}$.

Remark. *Note that the expression $Y = g(X)$ signifies that Y is a function from Ω to $\mathbb{R}$ such that $Y(\omega) = g(X(\omega))$ for each $\omega \in \mathbb{R}$.*

We will not prove this lemma, since we are going to study only measurable functions g here. However the proof is reasonably straightforward.

Note. *Note that we can break-down Y as $Y : \Omega \xrightarrow{X} \mathbb{R} \xrightarrow{g} \mathbb{R}$. Hence $\Omega \xrightarrow{Y = g \circ X} \mathbb{R}$.*

4.1 The CDF of $Y = g(X)$

To capture the behaviour of Y, we look at the CDF of Y. Clearly, since $Y = g(X)$, it is clear that the CDF of Y depends on that of X. Hence, for any $y \in \mathbb{R}$, we have the following expressions.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P\left(\{\omega \in \Omega : Y(\omega) \leq y\}\right) \\ &= P\left(\{\omega \in \Omega : g(X(\omega)) \leq y\}\right) \\ &= P(g(X) \leq y) \\ &= P(X \in A), \end{aligned}$$

where A denotes the set of **all** $x \in \mathbb{R}$ such that $g(x) \leq y$.

Remark. *Note that the capital alphabets refer to Random Variables, while the small letters are indicated real numbers which the random variable can take. There is nothing special about the notation y, it is just a number and not related to the random variable Y, except as a matter of convenience.*

Example 4.1. *Let $Y = g(X) = X^2$. Then*

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P\left(\{\omega \in \Omega : Y(\omega) \leq y\}\right) \\ &= P\left(\{\omega \in \Omega : (X(\omega))^2 \leq y\}\right) \\ &= P\left(\{\omega \in \Omega : -\sqrt{y} \leq X(\omega) \leq \sqrt{y}\}\right) \\ &= P(X \in [-\sqrt{y}, \sqrt{y}]) \\ &= P(X \leq \sqrt{y}) - P(X < -\sqrt{y}) \\ &= F_X(\sqrt{y}) - \lim_{\epsilon \to 0} F_X(-\sqrt{y} - \epsilon). \end{aligned}$$

Note that the above method is a general method for understanding the distribution of the random variable Y. However, in the special cases of discrete and continuous random variables, we may **also** be interested in focusing on the PMF or pdf directly, rather than computing the CDF first. These techniques are also especially relevant in the situation when we cannot or do not have closed form expressions for the CDF of X. This is what we do now. We consider three cases.

Discrete-to-Discrete: If X is discrete type random variable then $Y = g(X)$ is also a discrete type random variable irrespective of g.

Since X is a discrete type random variable, it takes values in some countable set, say $\{x_i : i \in S\}$ (for some countable set S). Thus $Y = g(X)$ must also be a discrete random variable. Hence, we can focus on the PMF of Y.

$$P_Y(y) = P(Y = y) = P(\{x_i : g(x_i) = y\}) = \sum_{x_i : g(x_i) = y} P(X = x_i) = \sum_{x_i : g(x_i) = y} P_X(x_i).$$

Continuous-to-Discrete: If X is a continuous type random variable and $g : \mathbb{R} \to \mathbb{R}$ is a discrete measurable function then $Y = g(X)$ is a discrete type random variable.

Since X is a continuous type random variable we consider its pdf $f_X(x)$. Since g is a discrete measurable function so $Y = g(X)$ is a discrete type random variable. The pmf of Y is given by

$$\begin{aligned} P_Y(y) = P(Y = y) = P(g(x) = y) = P(\{X = x : g(x) = y\}) \\ = P(\{\omega \in \Omega : X(\omega) = x \text{ and } g(x) = y\}) = \int_{x \in all\ intervals\ : g(x) = y} f_X(x) dx. \end{aligned}$$

Now, typically there are only a countable number of non-zero length intervals to be considered in the above final integral expression.

Remark. *What are the practical examples of functions g transform a continuous random variable to a discrete one? These are typically the 'quantizing' functions. In other words, whenever x falls in some interval, $g(x)$ is one real number, and so on. This can split the real number line (corresponding to X) into a countable number of intervals, such that in each interval the function g takes one value.*

Class 10. (04/09/18)

Continuous-to-Continuous:If X is a continuous type random variable and $g : \mathbb{R} \to \mathbb{R}$ is a non-discretizing (or non-quantizing, say a continuous function) measurable function then $Y = g(X)$ is a continuous type random variable.

Remark. *What do we mean by g being* ***non-discrete?*** *We mean that there are countably many x_i such that $g(x_i) = y$ for any $y \in \mathbb{R}$.*

Since X is a continuous type random variable we consider its pdf $f_X(x)$. Since g is a non-discretizing function so $Y = g(X)$ is a continuous type random variable. The pdf of Y is given by $f_Y(y)$. In that case, we have, for small $\Delta y > 0$,

$$P(y < Y \leq \Delta y) = \int\limits_{y}^{y+\Delta y} f_Y(y)dy \approx f_Y(y)\Delta y. \tag{1}$$

Since g is non-discrete, we will have the following expressions involving the countable disjoint union and hence the sum

$$\begin{aligned} P(y \leq Y \leq y + \Delta y) &= P\left(\dot{\bigcup}_{x_i:g(x_i)=y}\big(X \in [x_i, x_i + \Delta x_i]\big)\right) & (2)\\ &= \sum_{x_i:g(x_i)=y} P\big(X \in [x_i, x_i + \Delta x_i]\big) \approx \sum_{x_i:g(x_i)=y} f_X(x_i)|\Delta x_i| & (3)\end{aligned}$$

Note. *In the interval $y \leq Y \leq y + \Delta y$ the pre-images of y under g are x_i and the pre-images of $y + \Delta y$ under g are $x_i + \Delta x_i$. Note that even if $\Delta y > 0$, by the nature of the function g, we could have $\Delta x_i > 0$ or otherwise. Further the lengths of the intervals around x_i can also vary based on the slope of the function g around that point x_i Also, here Δy and each Δx_i is sufficiently small such that all the events $\big(X \in [x_i, x_i + \Delta x_i]\big)$ are disjoint.*

From the Equations (1) and (3) we have (for small Δy)

$$\begin{aligned} f_Y(y)\Delta y &= \lim_{\Delta x_i \to 0} \sum_{x_i:g(x_i)=y} f_X(x_i)|\Delta x_i| \\ f_Y(y) &= \lim_{\Delta x_i \to 0} \sum_{x_i:g(x_i)=y} f_X(x_i)\left|\frac{\Delta x_i}{\Delta y}\right| = \lim_{\Delta x_i \to 0} \sum_{x_i:g(x_i)=y} f_X(x_i)\left|\frac{\Delta y}{\Delta x_i}\right|^{-1} \\ \textit{We have } \lim_{\Delta x_i \to 0}\left|\frac{\Delta y}{\Delta x_i}\right| &= \left|\frac{dg(x)}{dx}\right|_{x=x_i} = |g'(x_i)| \end{aligned}$$

Therefore the pdf of the random variable $Y = g(X)$ is given as $f_Y(y) = \sum\limits_{x_i:g(x_i)=y} f_X(x_i)|g'(x_i)|^{-1}$ (where $^{-1}$ indicates the reciprocal).

Remark. *Suppose at some point y and for some $x_i : g(x_i) = y$, the value of $g'(x_i) = 0$. Then the above final expression has an issue. This can be overcome simply by looking at the expressions prior, where we had $\frac{\Delta x_i}{\Delta y}$. As g is smooth, this would indicate that the inverse function of g (if such exists) is non-differentiable at y. Hence we can simply ignore the contribution due to x_i. This is a crude explanation, typically such situations will not arise.*

Note. *Discrete-to-Continuous case is not possible. (Check !)*
Class 11. (11/09/18)

5 Mean(Expectation) and Variance of a random variable

Definition 5.1 (Mean of discrete random variable)**.** *Consider a discrete type random variable X taking values in $\{x_i : i \in S\}$ for some countable set S with PMF $P_X(x_i)$. The Mean or Expectation of X is denoted by $\mathbb{E}(X)$ and is defined as $\mathbb{E}(X) = \sum\limits_{x_i \in S} x_i P_X(x_i)$.*

$\mathbb{E}(X)$ is usually denoted by the symbol μ.

Definition 5.2 (Mean of continuous random variable)**.** *Consider a continuous type random variable X with pdf $f_X(x)$. The Mean or Expectation of X is denoted by $\mathbb{E}(X)$ and is defined as $\mathbb{E}(X) = \int\limits_{-\infty}^{\infty} x f_X(x)dx$.*

Remark. *If $\sum\limits_{x_i \in S} |x_i| P_X(x_i) < \infty$ (equivalently, for continuous, $\int\limits_{-\infty}^{\infty} |x| f_X(x) dx < \infty$) then the mean is finite. There exists distributions for which the mean is not finite. For instance, check the so-called Cauchy distribution.*

Definition 5.3 (Variance of a random variable)**.** *The variance of a random variable is denoted by $Var(X)$ and is defined as $Var(X) = \mathbb{E}((X-\mu)^2)$.*

$Var(X)$ is usually denoted as σ^2. The below equalities is true (its proof is shown below, in which the last-but-one equality is based on the so-called 'Linearity of Expectation' theorem, which will be shown later (in Section 10).

$$Var(X) = \mathbb{E}((X-\mu)^2) = \mathbb{E}(X^2 + \mu^2 - 2\mu X) = \mathbb{E}(X^2) - \mu^2 = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$$

Note. *$\mathbb{E}((X-\mu)^n)$ is called as " n^{th} order moment of X about the mean" or " n^{th} order central moment of X" (for $n \in \{1, 2, \cdots\}$). So $Var(X)$ is the second order central moment. $\mathbb{E}(X^n)$ is called as n^{th} order moment of X about the origin (for $n \in \{1, 2, \cdots\}$).*

Standard deviation of X is defined as $\sqrt{Var(X)}$. Therefore $\sigma = \sqrt{Var(X)}$ (note that $\sqrt{}$ is non-negative by default).

5.1 Mean of function of random variable

Lemma 5.1. *Let X be a discrete type random variable with PMF $P_X(x)$. Let $g : \mathbb{R} \to \mathbb{R}$ be a measurable function and $Y = g(X)$ then $\mathbb{E}(Y) = \sum\limits_{x_i} g(x_i) P_X(x_i)$.*

Proof. Since X is a discrete type random variable, consider X takes values in the countable set $\mathcal{X}$. We know that Y is a discrete type random variable with PMF $P_Y(y_i) = \sum\limits_{x_{i,j} \in \mathcal{X} : g(x_{i,j}) = y} P_X(x_{i,j})$. Thus, we have

$$\begin{aligned}\mathbb{E}(Y) &= \sum_{y_i} y_i P_Y(y_i) = \sum_{y_i} y_i \sum_{x_{i,j} : g(x_{i,j}) = y_i} P_X(x_{i,j}) = \sum_{y_i} \sum_{x_{i,j} : g(x_{i,j}) = y_i} y_i P_X(x_{i,j}) \\ &= \sum_{y_i} \sum_{x_{i,j} : g(x_{i,j}) = y_i} g(x_{i,j}) P_X(x_{i,j}) \\ &= \sum_{x_j} g(x_j) P_X(x_j).\end{aligned}$$

The last equality above follows because as we run through all possible preimages $x_{i,j} \in \mathcal{X}$ of all possible y_is, we are also running through all possible values of $x_{i,j}$s. □

Lemma 5.2. *Let X be a continuous type random variable with pdf $f_X(x)$. Let $g : \mathbb{R} \to \mathbb{R}$ be a discrete measurable function and $Y = g(X)$. Then $\mathbb{E}(Y) = \int\limits_{-\infty}^{\infty} g(x) f_X(x) dx$.*

Proof. We know that Y is a discrete type random variable with PMF $P_Y(y) = \int\limits_{x \in all\ intervals\ : g(x) = y} f_X(x) dx$.

$$\begin{aligned}\mathbb{E}(Y) &= \sum_{y_i} y_i P_Y(y_i) = \sum_{y_i} y_i \int\limits_{all\ intervals\ : g(x) = y_i} f_X(x) dx \\ &= \sum_{y_i} \int\limits_{all\ intervals\ : g(x) = y_i} y_i f_X(x) dx \\ &= \sum_{y_i} \int\limits_{all\ intervals\ : g(x) = y_i} g(x) f_X(x) dx \\ &= \int\limits_{-\infty}^{\infty} g(x) f_X(x) dx,\end{aligned}$$

where the reason for the above last equality is the same as given in the previous lemma. □

Lemma 5.3. *Let X be a continuous type random variable with pdf $P_X(x)$. Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous measurable function and $Y = g(X)$ then $\mathbb{E}(Y) = \int\limits_{-\infty}^{\infty} g(x) f_X(x)$.*

Proof. The technique is similar to Lemma 5.1. The difference is that we have to 'discretize' Y as well. In other words, $E(Y) = \int y f_Y(y)dy = \lim_{\Delta y \to 0} \sum y_i f_Y(y_i)\Delta y$, by approximating the integral using a Reimann sum. Now for the last expression, proceed similar to proof of Lemma 5.1 using the fact that $f_Y(y_i)\Delta y = \sum_{x_{i,j}:g(x_{i,j})=y_i} f_X(x_{i,j})|\Delta x|$ (as $\Delta x \to 0$). □

Class 12. (14/09/18)

6 Standard/Common random variables and their distributions

In many real world problems some random variables are often relevant to model several situations. So we give some formal names to them and study their properties like CDF,pdf, PMF, mean, variance, moment generating function. Note that in general, the CDF $F_X(x)$ is sufficient to give the description of the random variable. In the case of continuous random variables, the pdf suffices, and in the case of discrete random variables, the PMF is sufficient.

Note. *for sketches of CDF, pdf, PMF of standard random variables refer to any standard text book (or draw them yourself or using a software like Octave to see how they look, how they change with the parameters, etc.).*

Remark. *The reader should verify the properties of the pdf, PMF, and the values of mean and expectation of each of the distributions mentioned here. There are other classic standard distributions,but these mentioned here are some of the most used ones.*

6.1 Standard continuous distributions

6.1.1 Uniform random variable (or Uniform Distribution on one dimension)

A continuous type random variable is said to be a *uniform random variable* if its pdf is

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & x \in [a,b] \\ 0 & otherwise \end{cases}$$

for some interval $[a,b]$.

If X is a uniformly distributed random variable in the interval $[a,b]$, then we represent it as $X \sim U(a,b)$.

$$\begin{aligned} \mathbb{E}(X) &= \int_{-\infty}^{\infty} x f_X(x)dx = \int_a^b \frac{x}{b-a}dx = \frac{1}{b-a}\int_a^b xdx = \frac{1}{2(b-a)}.\big[x^2\big]_a^b = \frac{b+a}{2} \\ \mathbb{E}(X^2) &= \int_{-\infty}^{\infty} x^2 f_X(x)dx = \int_a^b \frac{x^2}{b-a}dx = \frac{1}{b-a}\int_a^b x^2dx = \frac{1}{3(b-a)}.\big[x^3\big]_a^b = \frac{b^2+ba+a^2}{3} \\ Var(X) &= \mathbb{E}(X^2) - \big(\mathbb{E}(X)\big)^2 = \frac{b^2+ba+a^2}{3} - \Big(\frac{b+a}{2}\Big)^2 = \frac{(b-a)^2}{12} \end{aligned}$$

6.1.2 Exponential Random variable (Exponential Distribution)

A continuous type random variable is said to be a *exponential random variable* if its pdf is

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & 0 \le x < \infty \\ 0 & otherwise \end{cases}$$

for some $\lambda > 0$.

If X is a exponentially distributed random variable with parameter λ, then we represent it as $X \sim exp(\lambda)$. The following involves only basic integral evaluations.

$$\begin{aligned} \mathbb{E}(X) &= \int_{-\infty}^{\infty} x f_X(x)dx = \frac{1}{\lambda} \\ \mathbb{E}(X^2) &= \int_{-\infty}^{\infty} x^2 f_X(x)dx = \frac{2}{\lambda^2} \\ Var(X) &= \mathbb{E}(X^2) - \big(\mathbb{E}(X)\big)^2 = \frac{1}{\lambda^2} \end{aligned}$$

6.1.3 Normal Distribution or Gaussian Distribution

A continuous type random variable is said to be a *Gaussian random variable* if its pdf is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \; e^{\dfrac{-(x-\mu)^2}{2\sigma^2}} \quad , \quad \forall x \in \mathbb{R}$$

for some $\mu \geq 0, \quad \sigma^2 > 0$.

If X is a Gaussian distributed random variable with parameters μ, σ^2, then we represent it as $X \sim \mathcal{N}(\mu, \sigma^2)$.

$$\mathbb{E}(X) = \mu$$
$$Var(X) = \sigma^2$$

Note. *The special case of the Gaussian distribution with $\mu = 0, \sigma^2 = 1$ is called the Standard Normal Distribution. A random variable $X \sim \mathcal{N}(0,1)$ is called a Standard Normal Variable.*

Class 13. (18/09/18)

6.2 Standard discrete distributions

6.2.1 Constant random variable (F_X has single jump)

The constant random variable takes only one value, say $a \in \mathbb{R}$. The PMF is given by

$$P_X(x) = \begin{cases} 1 & x = a \\ 0 & otherwise \end{cases}$$

Thus the CDF, $F_X(x) = \begin{cases} 0 & x < a. \\ 1 & x \geq a. \end{cases}$

$$\begin{aligned} \mathbb{E}(X) &= \sum_{x_i} x_i P_X(x_i) = aP(a) + \sum_{x_i \neq a} x_i P_X(x_i) = a(1) + 0 = a \\ \mathbb{E}(X^2) &= \sum_{x_i} x_i^2 P_X(x_i) = a^2 P(a) + \sum_{x_i \neq a} x_i^2 P_X(x_i) = a^2 \\ Var(X) &= \mathbb{E}(X^2) - \big(\mathbb{E}(X)\big)^2 = a^2 - (a)^2 = 0 \end{aligned}$$

This can be easily understood by intuition. The random variable takes a single value i.e it does not *deviate from its mean*, and so the variance will be zero.

Exercise : Let X be a non-negative random variable. If $Var(X) = 0$, show that $P(X = c) = 1$ for some $c \geq 0$.

6.2.2 Bernoulli random variable (or Bernoulli distribution) (F_X has two jumps)

A Bernoulli random variable is one which can take two values (typically these two values are 0 and 1). For $p \in [0,1]$ the pmf of a Bernoulli random variable are given below

$$P_X(x) = \begin{cases} 1-p & x = 0 \\ p & x = 1 \\ 0 & otherwise \end{cases}$$

Thus the CDF is $F_X(x) = \begin{cases} 0 & x < 0. \\ 1-p & 0 \leq x < 1. \\ 1 & x \geq 1. \end{cases}$

Note that this distribution is parameterized by the value p. If X is a Bernoulli random variable, we say $X \sim Bernoulli(p)$.

$$\begin{aligned} \mathbb{E}(X) &= \sum_{x_i} x_i P_X(x_i) = 0.P(0) + 1.P(1) = 0.(1-p) + 1.p = p \\ \mathbb{E}(X^2) &= \sum_{x_i} x_i^2 P_X(x_i) = 0.P(0) + 1.P(1) = 0.(1-p) + 1.p = p \\ Var(X) &= \mathbb{E}(X^2) - \big(\mathbb{E}(X)\big)^2 = p - p^2 = p(1-p) \end{aligned}$$

6.2.3 Binomial random variable (F_X has $(n+1)$ jumps)

Definition 6.1. *A discrete type random variable is said to be Binomial random variable if its pmf is*

$$P_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^x & x \in \{0,1,2,\cdots,n\} \\ 0 & otherwise \end{cases}$$

for some $p \in [0,1]$.

Here n, p are the parameters of the Binomial random variable. If X is binomial random variable with parameters n, p then we represent it as $X \sim B(n,p)$.

$$\mathbb{E}(X) = np$$
$$Var(X) = np(1-p)$$

Note. *Bernoulli random variable is a special case of Binomial random variable with $n = 1$. Therefore if X is a Bernoulli random variable with parameter p, then it can be repreented as $X \sim B(1,p)$. More interestingly, at some point later, we will see that the Binomial random variable is equal to the sum of independent Bernoulli random variables.*

6.2.4 Poisson random variable (F_X has countably infinite jumps)

Definition 6.2. *A discrete type random variable is said to be Poisson random variable if its pmf is*

$$P_X(x) = \begin{cases} \dfrac{e^{-\lambda}\lambda^x}{x!} & x \in \{0,1,2,\cdots\} \\ 0 & otherwise \end{cases}$$

for some $\lambda > 0$.

Here λ is the parameter of the Poisson random variable. If X is Poisson random variable with parameter λ then we represent it as $X \sim Poisson(\lambda)$.

$$\begin{aligned}
\mathbb{E}(X) &= \sum_k kP_X(k) = \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = \lambda \sum_{(k-1)=0}^{\infty} \frac{e^{-\lambda}\lambda^{(k-1)}}{(k-1)!} = \lambda e^{-\lambda}e^{\lambda} = \lambda \\
\mathbb{E}(X^2) &= \sum_k k^2 P_X(k) = \sum_k (k^2 - k + k)P_X(k) = \sum_{k=0}^{\infty} k(k-1)\frac{e^{-\lambda}\lambda^k}{k!} + \mathbb{E}(X) \\
&= e^{-\lambda}\left(0 + 0 + \left(2.1.\frac{\lambda^2}{2!}\right) + \left(3.2.\frac{\lambda^3}{3!}\right) + \cdots\right) + \lambda \\
&= e^{-\lambda}\lambda^2(1 + \lambda + \lambda^2 + \cdots) + \lambda = e^{-\lambda}\lambda^2 e^{\lambda} + \lambda = \lambda^2 + \lambda \\
Var(X) &= \mathbb{E}(X^2) - \big(\mathbb{E}(X)\big)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.
\end{aligned}$$

Class 14. (25/09/18)

7 Jointly distributed RVs (or) Random Vector

In previous lectures we have defined a single random variable (one dimensional random variable) and discussed its properties. There are situations where we need to handle multiple random variables (defined on the same probability space) simultaneously. This leads to the notion of Random vector (or) Jointly distributed random variables. An $n-$ dimensional random vector is a vector with n random variables as $X = (X_1, X_2, \cdots, X_n)$. By using this notion we can find Joint distribution of n random variables, Marginal distribution of fewer than n random variables, Conditional distribution of some random variables given some other random variables which are useful in real life problems.

Definition 7.1. *Consider the probability space $(\Omega, \mathcal{F}, P)$. Let $X_i : \Omega \to \mathbb{R}, \forall i \in \{1,2,\cdots,n\}$ be RVs, then $X_1, X_2, X_3, ...X_n$ are said to be jointly distributed with **joint CDF** denoted as $F_{X_1,X_2,...X_n}(x_1,x_2,...x_n)$ and is defined as*

$$F_{X_1,X_2,...X_n}(x_1,x_2,...x_n) \triangleq P\big((X_1 \le x_1),(X_2 \le x_2),....(X_n \le x_n)\big)$$

Note. *In the right hand side of the above equation comma $(,)$ represents intersection $(\cap)$.*

Therefore

$$\begin{aligned}F_{X_1,X_2,...X_n}(x_1,x_2,...x_n) &= P\big((X_1 \le x_1) \cap (X_2 \le x_2) \cap \cdots (X_n \le x_n)\big)\\ &= P\left(\{\omega \in \Omega : X_1(\omega) \le x_1\} \cap \{\omega \in \Omega : X_2(\omega) \le x_2\} .... \cap \{\omega \in \Omega : X_n(\omega) \le x_n\}\right)\\ &= P\left(\{\omega \in \Omega : X_1(\omega) \le x_1, X_2(\omega) \le x_2, \cdots, X_n(\omega) \le x_n\}\right)\end{aligned}$$

Note. *In the rest of the section we will consider $n = 2$, which can be generalized to an arbitrary n.*

7.1 Properties of Joint CDF

Let X and Y be jointly distributed RVs. Then the joint CDF will be $F_{X,Y}(x, y) = P(X \le x, Y \le y)$.

1. $F_{X,Y}(x, y)$ is non-decreasing in both X and Y.

   - If $x_1 < x_2$, then $F_{X,Y}(x_1, y) \le F_{X,Y}(x_2, y)\ \forall\ y \in \mathbb{R}$.

     Proof.

     $$\begin{aligned}\textit{Consider } (X \le x_2, Y \le y) &= (X \le x_1, Y \le y)\dot{\bigcup}(x_1 < X \le x_2, Y \le y)\\ P(X \le x_2, Y \le y) &= P(X \le x_1, Y \le y) + P(x_1 < X \le x_2, Y \le y)\\ F_{X,Y}(x_2, y) &= F_{X,Y}(x_1, y) + P(x_1 < X \le x_2, Y \le y)\\ \textit{Therefore } F_{X,Y}(x_2, y) &\le F_{X,Y}(x_1, y)\end{aligned}$$

     □

   - If $y_1 < y_2$, then $F_{X,Y}(x, y_1) \le F_{X,Y}(x, y_2)\ \forall\ x \in \mathbb{R}$. (Proof is similar to the above proof).
   - If $x_1 < x_2$, and $y_1 < y_2$, where $x_1, x_2, y_1, y_2 \in \mathbb{R}$. Then

     $$\begin{aligned}F_{X,Y}(x_1, y_1) \le F_{X,Y}(x_2, y_1) \le F_{X,Y}(x_2, y_2)\\ F_{X,Y}(x_1, y_1) \le F_{X,Y}(x_1, y_2) \le F_{X,Y}(x_2, y_2)\end{aligned}$$

     (Proof is similar to the above proof).

2. - $\lim\limits_{\substack{x\to\infty\\ y\to\infty}} F_{X,Y}(x, y) = 1$

     Proof. Define $B_i = (X \le i, Y \le i), \forall i \in \mathbb{Z}$. Then we can write $B_1 \subset B_2 \subset \cdots$. It is clear that $\left(\bigcup\limits_{i=1}^{\infty} B_i\right) = \Omega$. By continuity of probability theorem we have $\lim\limits_{n\to\infty} P(B_n) = P\left(\bigcup\limits_{i=1}^{\infty} B_i\right) = P(\Omega) = 1$. We can write $\lim\limits_{n\to\infty} P(B_n) = \lim\limits_{n\to\infty} P(X \le n, Y \le n) = \lim\limits_{\substack{x\to\infty\\ y\to\infty}} P(X \le x, Y \le y)$. Using the fact that $P(X \le x, Y \le y) = F_{X,Y}(x, y)$, therefore we have proved $\lim\limits_{\substack{x\to\infty\\ y\to\infty}} F_{X,Y}(x, y) = 1$.

     (Not part of the proof) Note: The last inequality, $\lim\limits_{n\to\infty} P(X \le n, Y \le n) = \lim\limits_{\substack{x\to\infty\\ y\to\infty}} P(X \le x, Y \le y)$, does require a proof, but here we are assuming it to be true. To prove this equality, we need to understand the definition of the limit operator. Those interested are encouraged to figure this out by themselves how to apply the definition of the limit, or ask the instructor. □

   - $\lim\limits_{\substack{x\to-\infty\\ y\to-\infty}} F_{X,Y}(x, y) = 0$. (Proof is similar to the above proof).
   - $\lim\limits_{x\to-\infty} F_{X,Y}(x, y) = 0$.

     Proof. Define $B_i = (X \le i, Y \le y), \forall i \in \mathbb{Z}^+$ (positive integers) and for some $y \in \mathbb{Z}^+$. Then we can write $B_{-1} \supset B_{-2} \supset \cdots$. It is clear that $\left(\bigcap\limits_{i=1}^{\infty} B_{-i}\right) = \phi$. By continuity of probability theorem we have $\lim\limits_{n\to\infty} P(B_{-n}) = P\left(\bigcap\limits_{i=1}^{\infty} B_{-i}\right) = P(\phi) = 0$. We can write $\lim\limits_{n\to\infty} P(B_{-n}) = \lim\limits_{n\to\infty} P(X \le -n, Y \le y) = \lim\limits_{x\to\infty} P(X \le -x, Y \le y)$. Using the fact that $P(X \le x, Y \le y) = F_{X,Y}(x, y)$, therefore we have proved $\lim\limits_{x\to\infty} F_{X,Y}(-x, y) = 0$. Therefore $\lim\limits_{x\to-\infty} F_{X,Y}(x, y) = 0$. □

- $\lim_{y \to -\infty} F_{X,Y}(x, y) = 0$. (Proof is similar to the above proof).

3. $F_{X,Y}(x, y)$ is Right continuous in x, y. (Proofs are similar to the one dimensional case)
   - $\lim_{\epsilon_1 \to 0+} F_{X,Y}(x + \epsilon_1, y) = F_{X,Y}(x, y)$
   - $\lim_{\epsilon_2 \to 0+} F_{X,Y}(x, y + \epsilon_2) = F_{X,Y}(x, y)$
   - $\lim_{\substack{\epsilon_1 \to 0+ \\ \epsilon_2 \to 0+}} F_{X,Y}(x + \epsilon_1, y + \epsilon_2) = F_{X,Y}(x, y)$ (infer from above two).

4. Using the joint CDFs, one can recover the CDF of the individual random variables. These individual CDFs are then called the *marginal* CDF (although all that this word means is that we started from the joint distribution and obtained the distribution of a single random variable). This is expressed as follows.
   - $F_X(x) = \lim_{y \to \infty} F_{X,Y}(x, y)$
   - $F_Y(y) = \lim_{x \to \infty} F_{X,Y}(x, y)$.

   (Proof hint: apply continuity of probability)

Class 15. (28/09/18)

7.2 Types of Jointly distributed RVs

- If X and Y are discrete RVs then X and Y are said to be Jointly discrete random variables.
- If X and Y are continuous RVs then X and Y are said to be Jointly continuous random variables.
- If X and Y are neither both discrete RVs nor both continuous RVs, then X and Y are said to be Jointly mixed random variables.

Note. *In this lecture notes we consider Jointly discrete RVs, Jointly continuous RVs. In Jointly mixed RVs we only consider the case where one random variable is discrete and other random variable is continuous. We won't consider mixed-mixed, mixed-continuous, mixed-discrete cases.*

7.2.1 Joint CDF and joint pdf of jointly continuous RVs

If X and Y are jointly continuous RVs, we have the so-called joint pdf of X, Y, denoted by $f_{X,Y}(x, y)$. The relationship between joint CDF and joint pdf can be given as

$$f_{X,Y}(x, y) = \frac{\partial^2 \big(F_{X,Y}(x, y)\big)}{\partial x \partial y}. \tag{4}$$

Or in other words,

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(x, y) dy dx. \tag{5}$$

Since the individual variables are continuous, they have their own (marginal) pdfs f_X and f_Y which can be obtained as follows.

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

The above equations are easily seen to be true using Property 4 given in Section 7.1 (or by using the Total probability theorem after appropriately modelling the integrals as Reimann sums).

7.2.2 Joint CDF and joint PMF of jointly discrete RVs

Let X and Y be jointly discrete RVs, i.e $X \in \{x_i : i \in A\}, Y \in \{y_j : j \in B\}$, where A and B are countable and $x_i, y_j \in \mathbb{R}$. In this case, we have the joint PMF denoted by $P_{X,Y}(x, y)$, which denotes the probability $P(X = x_i, Y = y_i)$. The joint PMF and joint CDF of X and Y are given as

$$P_{X,Y}(x_i, y_j) = P(X = x_i, Y = y_j)$$

$$F_{X,Y}(x, y) = \sum_{x_i \leq x} \sum_{y_i \leq y} P_{X,Y}(x_i, y_j) = \sum_{\substack{x_i \leq x \\ y_i \leq y}} P_{X,Y}(x_i, y_j).$$

As in the jointly continuous case, we have the individual (marginal) PMFs obtained as

$$P(X=x)=P_X(x)=\sum_{y_j}P_{X,Y}(x,y_j),\ \text{and}\ P(Y=y)=P_Y(y)=\sum_{x_i}P_{X,Y}(x_i,y).$$

7.3 Conditional distributions of jointly distributed RVs

7.3.1 Conditional CDF of jointly distributed RVs

The joint CDF of the random variables X, Y is $F_{X,Y}(x,y)=P(X\le x,Y\le y)$. For some event B (with $P(B)>0$) from the definition of conditional probability we have $P\big((X\le x)|B\big)=\dfrac{P((X\le x)\cap B)}{P(B)}$. From this observation we can define the Conditional CDF of a random variables X given an event B as $F_{X|B}(x)=P\left((X\le x)|B\right)$.

Note. *If $B=(Y\le y)$ for some fixed $y\in\mathbb{R}$ then*

$$F_{X|(Y\le y)}(x)=P\left((X\le x)|(Y\le y)\right)=\frac{P(X\le x)\cap(Y\le y)}{P(Y\le y)}=\frac{P(X\le x,Y\le y)}{P(Y\le y)}=\frac{F_{X,Y}(x,y)}{F_Y(y)}.$$

Note that $F_{X|(Y\le y)}(x)=P(X\le x|(Y\le y)$ is a cumulative probability distribution on the random variable X given that the event $(Y\le y)$ occured. Thus, we must have that the properties of a CDF should be satisfied by $F_{X|(Y\le y)}(x)$. For instance, we must have $\lim\limits_{x\to\infty}F_{X|(Y\le y)}(x)=1$, and so on.

Class 16. (01/10/18)

7.4 Bayes theorem for jointly distributed RVs

We now describe four variants of the Bayes' theorem for the case of random variables by lifting the Bayes' theorem defined for events before. In particular we consider the cases when both X and Y are discrete or continuous, and then the case when one of them is continuous and the other one is discrete. This exercise of obtaining the form of Bayes' theorem for these situations will simply reflect the fact that Bayes' theorem hold as we expect, with the application of pdf and PMFs appropriately as per whether the random variables are discrete or continuous.

7.4.1 If X and Y are jointly discrete RVs

For any event B, $P(X=x,B)$ is a valid pmf. Since Y is also discrete, $P(Y=y)\neq 0$ for some $y\in\mathbb{R}$. So, for such a y, $P(X=x|Y=y)$ is well-defined. Hence we can apply Bayes' theorem.

$$P(X=x|Y=y)=\frac{P\big((X=x)\cap(Y=y)\big)}{P(Y=y)}=\frac{P_{X,Y}(x,y)}{P_Y(y)}.$$

Moreover we can obtain the marginal PMF of Y using $P_Y(y)=\dfrac{P_{X,Y}(x,y)}{\sum_x P(x,y)}$.

Denoting $P(X=x|Y=y)$ as $P_{X|Y}(x|y)$, *the conditional PMF of X given $Y=y$*, we can thus write from the above expressions, $P_{X,Y}(x,y)=P_{Y|X}(y|x)P_X(x)=P_{X|Y}(x|y)P_Y(y)$, where $P_{Y|X}(y|x)=P(Y=x|X=x)$. Therefore we can express the Bayes' theorem for jointly discrete RVs as

$$P_{X|Y}(x|y)=\frac{P_{Y|X}(y|x)P_X(x)}{P_Y(y)} \tag{6}$$

(of course, we need to have $P_Y(y)\neq 0$ for this to hold).

7.4.2 If X is discrete RV and Y is continuous RV

Given that Y is a continuous RV. Therefore $P(Y=y)=0,\forall y\in\mathbb{R}$. So, $P(X|Y=y)$ is not defined. So we denote by $P_{X|Y}(x|y)$ *the conditional PMF of X given that Y takes values in an infitesimally small interval around y*, i.e., $P_{X|Y}(x|y)\triangleq\lim\limits_{\Delta y\to 0}P(X=x|y<Y\le y+\Delta y)$, where $P(y<Y\le y+\Delta y)\neq 0$. We can then write the following inequalities.

$$\begin{aligned}\lim_{\Delta y\to 0}P(X=x|y<Y\le Y+\Delta y)&=\lim_{\Delta y\to 0}\frac{P\big((X=x),(y<Y\le y+\Delta y)\big)}{P(y<Y\le y+\Delta y)}\\&=\lim_{\Delta y\to 0}\frac{P\big((X=x),(y<Y\le y+\Delta y)\big)}{f_Y(y)\Delta y}\\&=\lim_{\Delta y\to 0}\frac{P\big((y<Y\le y+\Delta y)|(X=x)\big)P(X=x)}{f_Y(y)\Delta y}\end{aligned}$$

Define the *conditional pdf on Y given $X = x$* as

$$f_{Y|X}(y|x) \triangleq \lim_{\Delta y \to 0} \frac{P(y < Y \leq y + \Delta y | X = x)}{\Delta y}. \tag{7}$$

By using this definition in the above equation, we get

$$P(X = x | y < Y \leq y + \Delta y) = \lim_{\Delta y \to 0} \frac{f_{Y|X}(y|x) \Delta y P_X(x)}{f_Y(y) \Delta y}$$

$$P_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) P_X(x)}{f_Y(y)} \text{ (\textit{as } } \Delta y \text{ \textit{cancels we can remove limit})} \tag{8}$$

Note. *The reader should note the difference in the meaning allocated to the notation $P_{X|Y}(x|y)$ in the above two situations. We also see that by exchanging the terms in (8), we can get the continuous-discrete form of Bayes' theorem.*

$$f_{Y|X}(y|x) = \frac{P_{X|Y}(x|y) f_Y(y)}{P_X(x)}. \tag{9}$$

7.4.3 If X and Y are jointly continuous RVs

We finally come to the case when both X, Y are continuous RVs. Therefore $P(X = x) = 0, P(Y = y) = 0, \forall x, y \in \mathbb{R}$. Hence we define the following conditional pdf of X given Y.

$$f_{X|Y}(x|y) = \lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \frac{P(x < X \leq x + \Delta x | y < Y \leq y + \Delta y)}{\Delta x}$$

Similarly, the conditional pdf $f_{Y|X}(y|x)$ is defined (see how this definition is different from the definition in (**??**). We give now the Bayes' form

$$\begin{aligned} P(x < X \leq x + \Delta x | y < Y \leq y + \Delta y) &= \frac{(P(y < Y \leq y + \Delta y | x < X \leq x + \Delta x) P(x < X \leq x + \Delta x)}{P(y < Y \leq y + \Delta y)} \\ &= \frac{f_{Y|X}(y|x) \Delta y . f_X(x) \Delta x}{f_Y(y) \Delta y} = \frac{f_{Y|X}(y|x) f_X(x) \Delta x}{f_Y(y)} \end{aligned}$$

Thus, by definition of $f_{X|Y}(x|y)$, we have the form of Bayes' result for the case when X and Y are both continuous.

$$\begin{aligned} f_{X|Y}(x|y) &= \lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \frac{f_{Y|X}(y|x) f_X(x) \Delta x}{f_Y(y) \Delta x} \\ \therefore f_{X|Y}(x|y) &= \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)} \end{aligned}$$

Note.

- *In each expression of the Bayes' result, we get only what is expected. One can easily recollect them (as well as the multiple interpretations we give for the conditional pmf and pdf $p_{X|Y}$ and $f_{X|Y}$)simply based on whether the random variables X and Y are continuous or discrete.*
- *Bayes theorem has to be applied according to the problem requirement. For example let X_1, X_2, X_3, X_4 be jointly discrete RVs. Then*

$$\begin{aligned} P_{X_1, X_3 | X_2, X_4}(x_1, x_3 | x_2, x_4) &= \frac{P_{X_2, X_4 | X_1, X_3}(x_2, x_4 | x_1, x_3) P_{X_1, X_3}(x_1, x_3)}{P_{X_2, X_4}(x_2, x_4)} \\ &= \frac{P_{X_2 | X_1, X_3, X_4}(x_2 | x_1, x_3, x_4) P_{X_1, X_3 | X_4}(x_1, x_3 | x_4)}{P_{X_2 | X_4}(x_2 | x_4)} \end{aligned}$$

In the above expression, depending on which variables are continuous, which variables are discrete, we can have the appropriate entries on the left and right.

- *For jointly mixed RVs where at least one of the RV is of mixed type we have to use only joint CDF (neither joint pdf nor joint PMF can be used).*

Class 17. (05/10/18) Forenoon

8 Independent RVs

Two jointly distributed RVs X and Y are said to be independent RVs, if

$$F_{X,Y}(x,y) = F_X(x)F_Y(y),\ \forall x, y \in \mathbb{R}$$

Note : The above conditions means that all events $(X \leq x), \forall x \in \mathbb{R}$ and $(Y \leq y), \forall y \in \mathbb{R}$ are independent. For the case of X, Y being jointly discrete or continuous, we have a simpler condition to check, which is summarized by the following two lemmas.

Lemma 8.1. *Let X, Y be jointly discrete RVs. Prove that $P_{X,Y}(x,y) = P_X(x)P_Y(y), \forall x, y \in \mathbb{R}$ if and only if X, Y are independent RVs.*

Proof. Note that since X and Y are jointly discrete, we have

$$P(X \leq x, Y \leq y) = \sum_{x_i \leq x, y_j \leq y} P_{X,Y}(x_i, y_j).$$

Now suppose $P_{X,Y}(x_i, y_j) = P_X(x_i)P_Y(y_j)$ for all x_i, y_j. For any x, y we can then write

$$\begin{aligned} P(X \leq x, Y \leq y) &= \sum_{x_i \leq x, y_j \leq y} P_{X,Y}(x_i, y_j) \\ &= \sum_{x_i \leq x} \sum_{y_j \leq y} P_X(x_i)P_Y(y_j) = \big(\sum_{x_i \leq x} P_X(x_i)\big)\big(\sum_{y_j \leq y} P_Y(y_j)\big) = F_X(x)F_Y(y), \end{aligned}$$

which means that X and Y are independent.

Now assume we have that X and Y are independent, i.e., $F_{X,Y}(x,y) = F_X(x)F_Y(y), \forall x, y$. Then we want to prove $P_{X,Y} = P_X P_Y$.

To prove this, we first consider the following expression.

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq x) = P(X = x, Y = y) + P(X < x, Y \leq y) + P(X \leq x, Y < y) - P(X < x, Y < y). \tag{10}$$

The above expression is true, since we have that the event

$$(X \leq x), (Y \leq y) = (X \leq x) \cap (Y \leq y) = (X = x, Y = y) \dot\cup (X < x, Y \leq y) \dot\cup \big((X \leq x, Y < y) \backslash (X < x, Y < y)\big)$$

Now since X and Y are discrete, using (10), we can write

$$P_{X,Y}(x,y) = P(X = x, Y = y) = F_{X,Y}(x,y) - F_{X,Y}(x_l, y) - F_{X,Y}(x, y_l) + F_{X,Y}(x_l, y_l),$$

where x_l is the largest value that X can take which is smaller than x, and y_l is the largest value which Y can take which is smaller than y. Now using the fact that $F_{X,Y} = F_X F_Y$ and the fact that $F_X(x) - F_X(x_l) = P_X(x)$ and $F_Y(y) - F_Y(y_l) = P_Y(y)$, we can complete the above proof. We leave it to the reader to do this. □

Remark. *We remark here that if X, Y are independent, then it is true that $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ for any $A, B \subseteq \mathbb{R}$.*

Lemma 8.2. *Let X, Y be jointly continuous RVs. Prove that $f_{X,Y}(x,y) = f_X(x)f_Y(y), \forall x, y \in \mathbb{R}$ if and only if X, Y are independent RVs.*

Proof. The reader can recall the basic relationships (4) and (5) from Section 7.2.1 and complete both the if part and the only if part of the proofs.

□

Theorem 8.3. *Let X and Y be independent random variables and g, h be borel measurable functions. Then $g(X)$ and $h(Y)$ are independent random variables*

Proof. Given X, Y are random variables and g, h are borel measurable functions. We know that $g(X), h(Y)$ are also random variables. Since X, Y are independent random variables we have $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y), \forall x, y \in \mathbb{R}$. Consider

$$\begin{aligned} P\left(g(X) \leq x, h(Y) \leq y\right) &= P\left(X \in g^{-1}(-\infty, x], Y \in h^{-1}(-\infty, y]\right) \\ &= P\left(X \in g^{-1}(-\infty, x]\right) P\left(Y \in h^{-1}(-\infty, y]\right) \ \textit{(because of above remark)} \\ &= P\left(g(X) \leq x\right) P\left(h(Y) \leq y\right) \ \textit{This holds } \forall x, y \in \mathbb{R} \end{aligned}$$

Therefore $g(X), h(Y)$ are independent random variables. Hence the proof. □

Note. *If X, Y are independent random variables and h, g are borel measurable functions then $g(X, Y)$ and $h(X, Y)$ are not independent random variables in general.*

9 Functions of multiple RVs

In any random experiment if we create some RVs and working with them, there might arise situations where we need some more RVs, then we can create new RV

- by using the definition of RV as a measurable function from sample space to $\mathbb{R}$.
- as a measurable function of existing RVs.

The second technique is more useful in applications. In this case we need to find distribution of new set of RVs from distribution of old set of RVs.

Consider two jointly distributed random variables X, Y. Let $g(X, Y)$ and $h(X, Y)$ be two measurable functions. Thus $U = g(X, Y)$ and $V = h(X, Y)$ are also random variables.

We then have the following expressions for the joint CDF of U, V and also the marginal CDF of U (the reader is recommended to get the same expression for V), following the same approach as that of functions of single random variables.

$$F_{U,V}(u,v) \triangleq P(U \leq u, V \leq v) \triangleq P((X,Y) \in \{(x,y) : g(x,y) \leq u, h(x,y) \leq v\}) \quad (11)$$

$$= P\{\omega \in \Omega : g(X(\omega), Y(\omega)) \leq u, h(X(\omega), Y(\omega)) \leq v\} \quad (12)$$

And similarly,

$$F_U(u) = P((X,Y) \in \{(x,y) : g(x,y) \leq u\}).$$

The above is the most general approach, which apply to any kind of random variables, and any number of functions of any number of random variables.

We now focus on the two random variable case, and consider only a few possible subcases that are relevant to our discussions going forward.

9.1 X, Y are jointly discrete:

In this case, U, V must also be jointly discrete, irrespective of what functions g and h are. Thus, we can obtain the joint PMF of U, V from that of X, Y by the obvious calculations.

$$P_{U,V}(u,v) = \sum_{(x,y):\ g(x,y)=u,\ h(x,y)=v} P_{X,Y}(x,y).$$

This can be done for each ordered pair $(u, v) \in \mathbb{R}^2$ which the random variables (U, V) can take.

In the same way, if we are interested only in the marginal PMF of U, denoted by P_U, we can obtain for each $u \in \mathbb{R}$ that U can take,

$$P_U(u) = \sum_{(x,y):\ g(x,y)=u} P_{X,Y}(x,y).$$

Similarly one can find the marginal PMF of V.

This covers the case when X, Y are discrete.

9.2 X, Y are jointly continuous

In this case, we consider the case when U, V are also jointly continuous (i.e., g and h are continuous functions). Hence we are interested in finding the joint pdf $f_{U,V}$ given the joint pdf $f_{X,Y}$. We use a similar approach as for the case for a continuous function of a continuous random variable.

$$P(u < U \leq u + \Delta u, v < V < v + \Delta v) = \sum_{(x_i,y_i):g(x_i,y_i)=u,h(x_i,y_i)=v} P(x_i < X \leq x_i + \Delta x_i, y_i < Y < y_i + \Delta y_i)$$

We thus have from the above expression

$$f_{U,V}(u,v)\Delta_{u,v} = \sum_{(x_i,y_i):g(x_i,y_i)=u,h(x_i,y_i)=v} f_{X,Y}(x_i,y_i)|\Delta_{x_i,y_i}|,$$

where Δ_{x_i,y_i} is an infinitesimal area in the X, Y plane around (x_i, y_i) corresponding the infinitestimal area in the U, V plane around (u, v).

It turns out that we can show that (but we won't do it here)

$$\frac{|\Delta_{x_i,y_i}|}{\Delta_{u,v}} = |determinant(J(x_i,y_i))|^{-1},$$

where $J(x_i, y_i) = \begin{bmatrix} \frac{\partial g}{\partial x}(x_i, y_i) & \frac{\partial g}{\partial y}(x_i, y_i) \\ \frac{\partial h}{\partial x}(x_i, y_i) & \frac{\partial h}{\partial y}(x_i, y_i) \end{bmatrix}$ is *the Jacobian matrix* (evaluated at (x_i, y_i)).

Thus, we have

$$f_{U,V}(u, v) = \sum_{(x_i,y_i):g(x_i,y_i)=u,h(x_i,y_i)=v} f_{X,Y}(x_i, y_i)|J(x_i, y_i)|^{-1}.$$

The above is a natural generalization of the functions of single random variable. Now if there are three functions of three random variables, then we will have similar expressions with the Jacobian matrix being a 3 × 3 matrix.

However, the reader may wonder as to what will happen if there are k functions of $n > k$ random variables which are jointly continuous. In this case, the Jacobian cannot be used, since it will not be a square matrix. So what do we do in this case? How do we find the joint pdf of the k new random variables?

The simple idea is to add some $n - k$ dummy functions of the n random variables so that the Jacobian is not uniformly zero. Then after finding the joint pdf of all the new n random variables, we can always integrate this joint pdf over all the dummy $n - k$ random variables we introduced so that we get the marginal joint pdf of the k original functions of the random variables.

We illustrate this using an example consisting of a single function of two random variables X, Y which are jointly continuous with joint pdf $f_{X,Y}$.

Example 9.1. *Consider $U = g(X, Y) = X^2 + Y^2$. We need to pick a function h such that the Jacobian $J(x, y) \neq 0$ as a function of x, y. Note that we cannot pick $h(X, Y) = c$ (some constant c) as this will lead to the Jacobian being 0 uniformly for all (x, y). We cannot also pick $h(X, Y) = c(X^2 + Y^2)$ as this leads to the same situation. However we can pick $V = h(X, Y) = X$. With this, we get the Jacobian as*

$$J(x, y) = \begin{bmatrix} 2x & 2y \\ 1 & 0 \end{bmatrix}.$$

Hence we have $|det(J(x, y))|^{-1} = \frac{1}{2|y|}$. We will get

$$f_{U,V}(u, v) = \sum_{(x_i,y_i):x_i^2+y_i^2=u,x_i=v} f_{X,Y}(x_i, y_i)\frac{1}{2|y_i|} = \frac{f_{X,Y}\left(v, \sqrt{v - x_i^2}\right)}{2\sqrt{v - x_i^2}} + \frac{f_{X,Y}\left(v, -\sqrt{v - x_i^2}\right)}{2\sqrt{v - x_i^2}}.$$

Thus if we know $f_{X,Y}$ we can obtain $f_{U,V}$.

Class 18. (05/10/18) Afternoon

10 Expected value of functions of multiple random variables and Linearity of Expectation

One can show without much difficulty that

$$\mathbb{E}(U) = \mathbb{E}(g(X, Y)) = \sum_{x,y} g(x, y)p_{X,Y}(x, y)$$

for the jointly discrete case, and similarly for the jointly continuous case, $\mathbb{E}(g(X, Y)) = \int_{x,y\in\mathbb{R}} g(x, y)f_{X,Y}(x, y)dxdy$. The proof follows in a similar way as to the functions of single random variables.

Now suppose that X, Y are jointly discrete and $U = g(X, Y) = X + Y$. Then we have

$$\begin{aligned} \mathbb{E}(U) = \mathbb{E}(X + Y) &= \sum_{x,y}(x + y)P_{X,Y}(x, y) \\ &= \sum_x \sum_y xP_{X,Y}(x, y) + \sum_x \sum_y yP_{X,Y}(x, y) \\ &= \sum_x x \sum_y P_{X,Y}(x, y) + \sum_y y \sum_x P_{X,Y}(x, y) \\ &= \sum_x xP_X(x) + \sum_y yP_Y(y) = \mathbb{E}(X) + \mathbb{E}(Y). \end{aligned}$$

We can show a similar result for the jointly continuous case also, and in fact the same holds in general. Thus, by induction, we can prove that *the expectation of the sum of a finite number of random variables is the sum of the individual expectations of the individual random variables.* This property of the expectation operator is known as the *linearity property.* It is quite useful in various occasions.

The reader is encouraged to check that $Var(X + Y)$ is *not* equal to $Var(X) + Var(Y)$ in general. However, this is true if the random variables X and Y are independent.

11 Tail Bounds (or) Moment inequalities

Sometimes it is very difficult to find the distribution of random variable but some of its moments(like mean, variance) are available. In these type of situations it is very difficult to find exact probabilities of events. So the natural question is can we get some bounds on the probabilities ? The following inequalities especially bound the 'tail' probability of random variables, using only some of its data such as its mean or variance.

Theorem 11.1 (Markov Inequality)**.** *Let X be a non negative random variable with $\mathbb{E}(X)$ exist. Then for any $a > 0, a \in \mathbb{R}$ we have*

$$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Proof. (applicable irrespective of whether X is continuous, discrete or mixed)

Given X is a non negative random variable. Therefore $P(X \geq 0) = 1$. Define a new random variable

$$Y = \begin{cases} 0 & if X < a \\ a & if X \geq a \end{cases}$$

$$\textit{Therefore } P(Y = 0) = P(X < a)$$
$$P(Y = a) = P(X \geq a)$$

So irrespective of the type of random variable X our new random variable Y is discrete.

$$\mathbb{E}(Y) = 0.P(Y = 0) + a.P(Y = a) = a.P(X \geq a) \tag{13}$$

from the definition of the random variable Y we have $X \geq Y$ (i.e., $X(\omega) \geq Y(\omega), \forall\omega$). Therefore $\mathbb{E}(X) \geq \mathbb{E}(Y)$ (Check !). By substituting $\mathbb{E}(Y)$ in this inequality we get $\mathbb{E}(X) \geq a.P(X \geq a)$. Therefore $P(X \geq a) \leq \dfrac{\mathbb{E}(X)}{a}$. Hence the proof. □

Proof. (alternate proof)(We will prove for continuous random variable. for discrete case proceed similarly)

Let X be a non negative continuous random variable. Then

$$\begin{aligned}
\mathbb{E}(X) &= \int_{-\infty}^{\infty} x f_X(x) dx = \int_{0}^{\infty} x f_X(x) dx \textit{ (since X is non negative)} \\
&= \int_{0}^{a} x f_X(x) dx + \int_{a}^{\infty} x f_X(x) dx \\
&\geq \int_{a}^{\infty} x f_X(x) dx \\
&\geq \int_{a}^{\infty} a f_X(x) dx = a \int_{a}^{\infty} f_X(x) dx \\
\mathbb{E}(X) &\geq a.P(X \geq a) \\
\textit{Therefore } P(X \geq a) &\leq \frac{\mathbb{E}(X)}{a}
\end{aligned}$$

Hence the proof. □

Theorem 11.2 (Chebyshev's Inequality)**.** *Let X be a random variable with $\mathbb{E}(X)$ and $\mathbb{E}\left((X-c)^2\right), \forall c \in \mathbb{R}$ exist, Then for $b > 0, b \in \mathbb{R}$ the Chebyshev's Inequality states that $P\left(|X - c| \geq b\right) \leq \dfrac{\mathbb{E}\left((X-c)^2\right)}{b^2}$. Also prove that $P\left(|X - \mathbb{E}(X)| \geq b\right) \leq \dfrac{Var(X)}{b^2}$.*

Proof. Given that X is a random variable. Then the random variable $(X - c)^2$ is a non negative random variable $\forall c \in \mathbb{R}$. From markov inequality we have $P\left((X-c)^2 \geq b^2\right) \leq \dfrac{\mathbb{E}\left((X-c)^2\right)}{b^2}, \forall b > 0$. The event $\left((X-c)^2 \geq b^2\right)$ can also be represented as $(|X - c| \geq b)$. Therefore the above inequality becomes $P\left(|X - c| \geq b\right) \leq \dfrac{\mathbb{E}\left((X-c)^2\right)}{b^2}$. Hence the proof. A special case of Chebyshev's Inequality which is useful in applications is obtained by taking $c = \mathbb{E}(X)$. then $\mathbb{E}\left((X - \mathbb{E}(X))^2\right) = Var(X)$. Hence we get $P\left(|X - \mathbb{E}(X)| \geq b\right) \leq \dfrac{Var(X)}{b^2}$. □

Theorem 11.3 (Weak law of large numbers (WLLN))**.** *Let* $X_i, \forall i = 1, 2, \cdots, n$ *are independent and identically distributed (i.i.d) random variables. Let* $S_n = \frac{\sum_{i=1}^{n} X_i}{n}$ *then* $\lim_{n \to \infty} P(|S_n - \mathbb{E}(X_1)| \geq \epsilon) = 0, \forall \epsilon \in \mathbb{R}$ *and* $\epsilon > 0$.

Proof. Given that $X_i, \forall i \in \{1, 2, \cdots, n\}$ are i.i.d RVs we have $\mathbb{E}(X_1) = \mathbb{E}(X_2) = \cdots \mathbb{E}(X_n)$ and $Var(X_1) = Var(X_2) = \cdots = Var(X_n)$. First we will find $\mathbb{E}(S_n)$ and $Var(S_n)$ and then apply Chebyshev's inequality on the RV S_n to get WLLN.

$$\begin{aligned}
\mathbb{E}(S_n) &= \mathbb{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \\
&= \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) \ (\textit{since } \mathbb{E}(aX) = a\mathbb{E}(X)) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) \ (\textit{because of Linearity of expectation}) \\
&= \frac{1}{n}n.\mathbb{E}(X_1) = \mathbb{E}(X_1) \ \textit{(since } X_i \textit{ are identically distributed)} \\
\therefore \mathbb{E}(S_n) &= \mathbb{E}(X_1) \\
Var(S_n) &= Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \\
&= \frac{1}{n^2}.Var\left(\sum_{i=1}^{n} X_i\right) \ (\textit{Since } Var(aX) = a^2.Var(X)) \\
&= \frac{1}{n^2}.\sum_{i=1}^{n} Var(X_i) \ (\textit{since } X_i \textit{ are independent }) \\
&= \frac{1}{n^2}.n.Var(X_1) \ (\textit{since } X_i \textit{ are identically distributed }) \\
\therefore Var(S_n) &= \frac{Var(X_1)}{n}
\end{aligned}$$

From Chebyshev's inequality we have $P\left(|S_n - \mathbb{E}(S_n)| \geq \epsilon\right) \leq \frac{Var(S_n)}{\epsilon^2}, \forall \epsilon \in \mathbb{R}$ and $\epsilon > 0$. By substituting $\mathbb{E}(S_n)$ and $Var(S_n)$ we get $P(|S_n - \mathbb{E}(X_1)| \geq \epsilon) \leq \frac{Var(X_1)}{n\epsilon^2}$. As $n \to \infty$ this inequality becomes $\lim_{n \to \infty} P(|S_n - \mathbb{E}(X_1)| \geq \epsilon) \leq 0$. But we have $P(\textit{any event}) \geq 0$. Therefore $\lim_{n \to \infty} P(|S_n - \mathbb{E}(X_1)| \geq \epsilon) = 0, \forall \epsilon \in \mathbb{R}$ and $\epsilon > 0$. Hence the proof. □

Class 19. (23/10/18)

12 Covariance and Cauchy-schwartz inequality

Definition 12.1 (Covariance)**.** *The Covariance of two RVs is defined as* $cov(X, Y) \triangleq \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$.

Note.

$$\begin{aligned}
cov(X, Y) &= \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \\
&= \mathbb{E}[XY - X\mathbb{E}(Y) - Y\mathbb{E}(X) + \mathbb{E}(X)\mathbb{E}(Y)] \\
&= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y) \\
&= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)
\end{aligned}$$

$cov(X, Y)$ indicates the average effect of X on Y and vice versa.

Definition 12.2 (Correlation coefficient (ρ))**.** *The Correlation coefficient (ρ) of two RVs X, Y is defined as*

$$\rho = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}}$$

Remark. *No need to remember the above term definition as it was not done in class.*

Definition 12.3 (Uncorrelated RVs)**.** *Two RVs X, Y are said to be uncorrelated if $cov(X, Y) = 0$.*

Note. *Therefore if X, Y are uncorrelated then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.*

Theorem 12.1. *Consider two RVs X, Y. Then $Var(X + Y) = Var(X) + Var(Y) - 2cov(X, Y)$.*

Proof.

$$\begin{aligned}
Var(X+Y) &= \mathbb{E}\left[[(X+Y) - \mathbb{E}(X+Y)]^2\right] \\
&= \mathbb{E}\left[[(X+Y) - \mathbb{E}(X) - \mathbb{E}(Y)]^2\right] \\
&= \mathbb{E}\left[[(X - \mathbb{E}(X)) + (Y - \mathbb{E}(Y)]^2\right] \\
&= \mathbb{E}\left[(X - \mathbb{E}(X))^2 + (Y - \mathbb{E}(Y))^2 - 2(X - \mathbb{E}(X))(Y - \mathbb{E}(Y)\right] \\
&= \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] + \mathbb{E}\left[(Y - \mathbb{E}(Y))^2\right] - 2\mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y)\right] \\
&= Var(X) + Var(Y) - 2cov(X, Y)
\end{aligned}$$

□

Corollary 12.1.1. *If X, Y are uncorrelated RVs then $Var(X + Y) = Var(X) + Var(Y)$.*

Lemma 12.2. *If X, Y are independent RVs then they are uncorrelated. The converse is not true in general.*

Proof. exercise ! (your job is to construct an example such that they are uncorrelated but not independent) □

We have in the above section discussed about when two random variables are completely uncorrelated. What about the other extreme, i.e. when they are completely correlated? What will be their relationship ? We now explore this question.

Theorem 12.3 (Cauchy-Schwartz inequality (Probability version))**.** *Consider two RVs X, Y defined in the probability space $(\Omega, \mathcal{F}, P)$ then $|\mathbb{E}(XY)| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$*

Proof. Given X, Y are two RVs, then $(X - \alpha Y)^2$ is a non-negative RV $\forall \alpha \in \mathbb{R}$. Therefore

$$\begin{aligned}
0 &\leq \mathbb{E}[(X - \alpha Y)^2] \\
0 &\leq \mathbb{E}(X^2 + \alpha^2 Y^2 - 2\alpha XY) \\
0 &\leq \mathbb{E}(X^2) + \alpha^2 \mathbb{E}(Y^2) - 2\alpha \mathbb{E}(XY)
\end{aligned}$$

This is true $\forall \alpha \in \mathbb{R}$. Consider $\alpha = \dfrac{\mathbb{E}(XY)}{\mathbb{E}(Y^2)}$. The above inequality becomes

$$\begin{aligned}
0 &\leq \mathbb{E}(X^2) + \left(\frac{\mathbb{E}(XY)}{\mathbb{E}(Y^2)}\right)^2 \mathbb{E}(Y^2) - 2\left(\frac{\mathbb{E}(XY)}{\mathbb{E}(Y^2)}\right)\mathbb{E}(XY) \\
0 &\leq \mathbb{E}(X^2) + \frac{(\mathbb{E}(XY))^2}{\mathbb{E}(Y^2)} - 2\frac{(\mathbb{E}(XY))^2}{\mathbb{E}(Y^2)} \\
0 &\leq \mathbb{E}(X^2) - \frac{(\mathbb{E}(XY))^2}{\mathbb{E}(Y^2)} \\
0 &\leq \mathbb{E}(X^2)\mathbb{E}(Y^2) - (\mathbb{E}(XY))^2 \\
(\mathbb{E}(XY))^2 &\leq \mathbb{E}(X^2)\mathbb{E}(Y^2) \\
\therefore |\mathbb{E}(XY)| &\leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}
\end{aligned}$$

Hence the proof. □

Remark. *Consider $|\mathbb{E}(XY)| = \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$. This means that $\mathbb{E}[(X - \alpha Y)^2] = 0$ for $\alpha = \dfrac{\mathbb{E}(XY)}{\mathbb{E}(Y^2)}$. Since $(X - \alpha Y)^2$ is a non negative RV and $\mathbb{E}[(X - \alpha Y)^2] = 0$, we can write $(X - \alpha Y)^2 = 0$ with probability 1. Therefore $X = \alpha Y$ with probability 1.*

Corollary 12.3.1. $|cov(X, Y)| \leq \sqrt{Var(X)Var(Y)}$

Proof.

$$\begin{aligned}|cov(X,Y)| &= |\mathbb{E}[(X-\mathbb{E}(X))\,(Y-\mathbb{E}(Y))]| \\ &\leq \sqrt{\mathbb{E}\left[(X-\mathbb{E}(X))^2\right]\mathbb{E}\left[(Y-\mathbb{E}(Y))^2\right]}\text{(by previous theorem)} \\ \therefore |cov(X,Y)| &\leq \sqrt{Var(X)Var(Y)}\end{aligned}$$

Hence the proof. □

Remark. *Following the previous remark, we can see that if $|cov(X,Y)| = \sqrt{Var(X)Var(Y)}$ then $X = \alpha Y +$ (some constant) with probability 1. This is the other extreme we were talking about, i.e., the case when the covariance is maximized. In this case, we see that X and Y have a linear relationship (with a constant shift) with each other.*

13 Moment Generating Function, Characteristic function, Chernoff bound and Central Limit Theorem

In this section, we shall define certain quantities which are very useful for us to prove a central result of probability theory, which is the *central limit theorem.* Towards that end, we first remark that we can consider complex-valued random variables in the same way as real-valued random variables. For instance, a complex valued random variable $X_r + jX_i$ essentially consists of two jointly distributed real random variables X_r and X_i, signifying the real and imaginary parts ($j = \sqrt{-1}$). In that case the mean of the complex-value random variable can be written as $\mathbb{E}(X_r + jX_i) = \mathbb{E}(X_r) + j\mathbb{E}(X_i)$. With this, we now define our quantities.

Definition 13.1. *The moment generating function (MGF) of a RV X is denoted by $m_X(t)$ where $t \in \mathbb{C}$ and is defined as $m_X(t) \triangleq \mathbb{E}(e^{tX})$. (if it exists).*

Note that e^{tX} is a complex-valued random variable in general, whose expectation is then calculated as above. However we also have the following inequalities which explain why the MGF is called as 'moment-generating'. Using Taylor series expansion, we have,

$$\begin{aligned}m_X(t) \triangleq \mathbb{E}[e^{tX}] &= \mathbb{E}\left[1 + \frac{tX}{1!} + \frac{(tX)^2}{2!} + \ldots\right] \\ &= 1 + t\frac{\mathbb{E}\left[X\right]}{1!} + t^2\frac{\mathbb{E}\left[X^2\right]}{2!} + \cdots \\ &= \sum_{i=1}^{\infty} t^i \frac{\mathbb{E}(X^i)}{i!}\end{aligned}$$

$$\text{Hence we can write } \left.\frac{d^n(m_X(t))}{dt^n}\right|_{t=0} = \mathbb{E}(X^n)$$

Therefore n^{th} moment of X is obtained from MGF as $\mathbb{E}(X^n) = \left.\frac{d^n}{dt^n}\left(m_X(t)\right)\right|_{t=0}$

Remark. *Note that the moment generating can be likened to the inverse Laplace transform for the case of continuous random variables (which have pdfs), i..e., $m_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx$. An advantage of the moment generating function is that if we can estimate the moments of a continuous random variable X, then we have an estimate for the moment generating function. In that case, it is easy for us to obtain $f_X(x)$ (the pdf) by taking a Laplace transform of the MGF. The situation is much simpler for the discrete case. If X is discrete, then we have $m_X(t) = \sum_{x_i} e^{tx_i} P_X(x_i)$ by definition. Hence, having an estimate for the moments (i.e. for the MGF) gives us the PMF directly, by looking at the coefficients f the various powers of e^t at various points x_i.*

A problem with the MGF is that it need not always exist for every distribution. For instance, for the Cauchy distribution, the MGF does not exist. However, we have a similar object which is the characteristic function of X, for which we shall show that the existence is guaranteed for all real-valued random variables.

Definition 13.2 (Characteristic function)**.** *The Characteristic function of a RV X is denoted by $\phi_X(\theta)$ and is defined as $\phi_X(\theta) \triangleq \mathbb{E}\left(e^{j\theta X}\right)$.*

Similar to MGF, using the Taylor series expansion, we can calculate n^{th} moment of X from Characteristic function as $\mathbb{E}(X^n) = \frac{1}{j^n}\frac{d^n}{d\theta^n}\left(\phi_X(\theta)\right)\Big|_{\theta=0}$

The Characteristic function of a RV always exists since

- If X is continuous RV then the integral $\int\limits_{-\infty}^{\infty} |e^{j\theta x} f_X(x)| dx$ always converges.

$\left(\int\limits_{-\infty}^{\infty} |e^{j\theta x} f_X(x)| dx = \int\limits_{-\infty}^{\infty} |e^{j\theta x}||f_X(x)| dx = \int\limits_{-\infty}^{\infty} f_X(x) dx = 1\right).$

- If X is discrete RV then the infinite sum $\sum\limits_{x=-\infty}^{\infty} |e^{j\theta x} P_X(x)|$ always converges.

$\left(\sum\limits_{x=-\infty}^{\infty} |e^{j\theta x} P_X(x)| = \sum\limits_{x=-\infty}^{\infty} |e^{j\theta x}||P_X(x)| dx = \sum\limits_{x=-\infty}^{\infty} P_X(x) dx = 1\right).$

Because we are assured of the existence of the characteristic function for any real-valued random variable, we will be using this for our proof of the central limit theorem.

Note. - *If $m_X(t)$ exist then $\phi_X(\theta) = m_X(t)|_{t=j\theta}$.*

- *If X is a continuous RV then*
  - *Characteristic function and pdf form a Fourier transform pair. Thus, the argument behind how we can obtain the distribution from the MGF carries over to the characteristic function as well.*

We end this section by showing an important property of the moment generating function for sums of independent random variables. It is easy to see that the same property continues to hold for the characteristic function as well.

Lemma 13.1. *Consider X, Y are independent RVs and $Z = X + Y$. Then $m_Z(t) = m_X(t) m_Y(t)$.*

Proof. From Theorem 8.3 we have e^{tX}, e^{tY} are independent RVs.

$$\begin{aligned} m_Z(t) &= \mathbb{E}(e^{tZ}) \\ &= \mathbb{E}(e^{t(X+Y)}) \\ &= \mathbb{E}(e^{tX} e^{tY}) \\ &= \mathbb{E}(e^{tX})\mathbb{E}(e^{tY}) \quad \text{(since } X \text{ and } Y \text{ are independent)} \\ \therefore m_Z(t) &= m_X(t) m_Y(t). \end{aligned}$$

□

13.1 MGF and Characteristic function of Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$

As both an example and for the purpose of proving the central limit theorem, we shall derive the MGF and the characteristic function of the Gaussian distribution.

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ M_X(t) &= \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} e^{tx} dx \\ \textit{Let } y &= \frac{x-\mu}{\sigma} \implies dy = \frac{dx}{\sigma} \\ M_X(t) &= \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}} e^{t(\sigma y + \mu)} dy \\ &= e^{t\mu} \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(y^2 - 2t\sigma y)}{2}} dy \\ &= e^{t\mu} \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(y^2 - 2t\sigma y + t^2\sigma^2 - t^2\sigma^2)}{2}} dy \\ &= e^{t\mu + \frac{t^2\sigma^2}{2}} \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(y-t\sigma)^2}{2}} dy \end{aligned}$$

The function in the integral is a gaussian pdf with mean $t\sigma$ and variance 1. Hence integral value is 1.

$$\therefore M_X(t) = e^{t\mu + \frac{t^2\sigma^2}{2}}$$
$$\textit{Similarly } \phi_X(\theta) = e^{j\theta\mu - \frac{\theta^2\sigma^2}{2}}$$

Class 20. (26/10/18) (Fore Noon)

13.2 Chernoff Bound

The Chernoff Bound is another bound on the tail probability of a random variable. The advantage of the Chernoff bound is that it is applicable for all random variables (unlike Markov inequality which is applicable only for non-negative RVs). Another advantage is that it has a tuneable parameter t, and one can find the best Chernoff bound by choosing an appropriate value for the parameter t which minimizes the bound.

Theorem 13.2. *Let X be a RV in probability space $(\Omega, \mathcal{F}, P)$ such that its MGF $m_X(t)$ exist. For any $t, a \in \mathbb{R}$, we must have $P(X \geq a) \leq \dfrac{m_X(t)}{e^{ta}}$.*

Proof. Given that X is a RV then e^{tX} is a non negative RV. Therefore we can apply Markov inequality on e^{tX}. Consider

$$\begin{aligned} P(X \geq a) &= P(e^{tX} \geq e^{ta}) \\ &\leq \frac{\mathbb{E}(e^{tX})}{e^{ta}} \\ \therefore P(X \geq a) &\leq \frac{m_X(t)}{e^{ta}} \end{aligned}$$

Hence the proof. □

The Chernoff Bound uses MGF of the RV to give a class of bounds for tail probability. To get best Chernoff-style upper bound minimize the function $\dfrac{m_X(t)}{e^{ta}}$ over all possible values of t.

Remark.

- *Markov inequality requires only mean value, Chebyshev inequality requires mean and variance. Chernoff bound requires MGF (which in some sense requires knowledge of the moments). So Chernoff bound requires more information than Markov and Chebyshev inequality.*
- *Even though all moments exist for certain RVs MGF won't exist (for example log-normal distribution). So Chernoff bound is not applicable for these kind of RVs.*

13.3 Central Limit Theorem (CLT)

Theorem 13.3 (CLT). *Let $X_1, X_2, \cdots$ be a sequence of independent and identically distributed (iid) RVs in the probability space $(\Omega, \mathcal{F}, P)$. Assume that $\mathbb{E}(X_i) = \mu$ and $var(X_i) = \sigma^2 (> 0), i = 1, 2, \cdots$ exist. Define a new sequence of RVs as*

$$Z_n = \frac{\sum\limits_{i=1}^{n} X_i - \mathbb{E}\left(\sum\limits_{i=1}^{n} X_i\right)}{\sqrt{var\left(\sum\limits_{i=1}^{n} X_i\right)}}, \forall n \in \mathbb{Z}^+$$

We denote the CLT statement also as $Z_n \xrightarrow{\text{in distribution}} \mathcal{N}(0,1)$.(i.e. as $n \to \infty, Z_n$ converges to standard normal RV).

Remark. *Another form of CLT states that $\sum\limits_{i=1}^{n} X_i$ converges to Normal(Gaussian) distribution as $n \to \infty$ with mean $= \lim\limits_{n\to\infty} \mathbb{E}\left[\sum\limits_{i=1}^{n} X_i\right]$ and variance $= \lim\limits_{n\to\infty} var\left[\sum\limits_{i=1}^{n} X_i\right]$ (provided these limits exist). There are many versions of CLT. This is the simplest form of CLT.*

Proof. We prove the theorem by showing convergence of the MGF of Z_n to the MGF of the standard normal random variable (which we have already derived). Given that $X_i, \forall i \in \mathbb{Z}^+$ are iid RVs with mean μ and variance σ^2. Therefore

$$\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = n\mu$$

$$\sqrt{var\left(\sum_{i=1}^{n} X_i\right)} = \sqrt{n\sigma^2} = \sigma\sqrt{n}$$

$$\therefore Z_n = \frac{\left(\sum\limits_{i=1}^{n} X_i\right) - n\mu}{\sigma\sqrt{n}}$$

We consider that MGF of X_i exist, $\forall i \in \mathbb{Z}^+$. (In the following remark we will relax this condition also). Consider MGF of Z_n.

$$\begin{aligned}
m_{Z_n}(t) &= \mathbb{E}(e^{Z_n t}) \\
&= \mathbb{E}\left(e^{\frac{\left(\left(\sum\limits_{i=1}^{n} X_i\right) - n\mu\right)t}{\sigma\sqrt{n}}} \right) \\
&= \mathbb{E}\left[e^{\left(\frac{-n\mu t}{\sigma\sqrt{n}}\right)} e^{\left[\frac{1}{\sigma\sqrt{n}}\left(\sum\limits_{i=1}^{n} X_i\right)t\right]} \right] \\
&= e^{\left(\frac{-n\mu t}{\sigma\sqrt{n}}\right)} \mathbb{E}\left[\prod_{i=1}^{n} e^{\left(\frac{X_i t}{\sigma\sqrt{n}}\right)} \right] \\
&= e^{\left(\frac{-n\mu t}{\sigma\sqrt{n}}\right)} \prod_{i=1}^{n} \mathbb{E}\left[e^{\left(\frac{X_i t}{\sigma\sqrt{n}}\right)} \right] \quad (\textit{ If } X, Y \textit{ are independent then } \mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y))
\end{aligned}$$

$$\left(\textit{Since } X_i, \forall i \in \mathbb{Z}^+ \textit{ are independent, functions of } X_i \textit{ that is } e^{\left(\frac{X_i t}{\sigma\sqrt{n}}\right)} \textit{ are also independent}\right).$$

$$\begin{aligned}
&= e^{\left(\frac{-n\mu t}{\sigma\sqrt{n}}\right)} \left[\mathbb{E}\left(e^{\left(\frac{X_1 t}{\sigma\sqrt{n}}\right)} \right) \right]^n \quad (\textit{ since } X_i, \forall i \in \mathbb{Z}^+ \textit{ are identically distributed}). \\
\therefore m_{Z_n}(t) &= e^{\left(\frac{-n\mu t}{\sigma\sqrt{n}}\right)} \left[m_{X_1}\left(\frac{t}{\sigma\sqrt{n}}\right) \right]^n \\
\ln m_{Z_n}(t) &= \frac{-n\mu t}{\sigma\sqrt{n}} + n \ln\left[m_{X_1}\left(\frac{t}{\sigma\sqrt{n}}\right) \right] (\textit{ By taking natural log on both sides}) \\
&\left(\textit{Apply } m_X(t) = 1 + \frac{\mathbb{E}(X)t}{1!} + \frac{\mathbb{E}(X^2)t}{2!} + \cdots = 1 + \mu t + \frac{(\sigma^2 + \mu^2)t^2}{2!} + \cdots \right) \\
\ln m_{Z_n}(t) &= \frac{-n\mu t}{\sigma\sqrt{n}} + n \ln\left[1 + \frac{\mu t}{\sigma\sqrt{n}} + \frac{(\sigma^2 + \mu^2)t^2}{2\sigma^2 n} + \cdots \right] \\
&\left(\textit{Apply } \ln(1+a) = a - \frac{a^2}{2} + \frac{a^3}{3} - \cdots \right) \left(\textit{here } a = \left[\frac{\mu t}{\sigma\sqrt{n}} + \frac{(\sigma^2 + \mu^2)t^2}{2\sigma^2 n} + \cdots \right] \right)
\end{aligned}$$

$$\therefore \ln m_{Z_n}(t) = \frac{-n\mu t}{\sigma\sqrt{n}} + n\left\{\left[\frac{\mu t}{\sigma\sqrt{n}} + \frac{(\sigma^2+\mu^2)t^2}{2\sigma^2 n} + \cdots\right] - \frac{1}{2}\left[\frac{\mu t}{\sigma\sqrt{n}} + \frac{(\sigma^2+\mu^2)t^2}{2\sigma^2 n} + \cdots\right]^2 + \cdots\right\}$$

$(By\ writing\ only\ terms\ of\ t\ and\ t^2\ explicitly)$

$$\ln m_{Z_n}(t) = \frac{-n\mu t}{\sigma\sqrt{n}} + n\left\{\left[\frac{\mu t}{\sigma\sqrt{n}}\right] + \left[\frac{(\sigma^2+\mu^2)t^2}{2\sigma^2 n} - \frac{\mu^2 t^2}{2\sigma^2 n}\right] + \cdots\right\}$$

$$\ln m_{Z_n}(t) = \frac{-n\mu t}{\sigma\sqrt{n}} + n\left\{\left[\frac{\mu t}{\sigma\sqrt{n}}\right] + \left[\frac{\sigma^2 t^2}{2\sigma^2 n}\right] + \cdots\right\} (\ other\ terms\ have\ positive\ powers\ of\ n\ in\ denominator)$$

$$\ln m_{Z_n}(t) = \frac{-n\mu t}{\sigma\sqrt{n}} + \left[\frac{n\mu t}{\sigma\sqrt{n}}\right] + \left\{\left[\frac{\sigma^2 t^2}{2\sigma^2}\right] + \cdots\right\} (\ other\ terms\ have\ positive\ powers\ of\ n\ in\ denominator)$$

$$\ln m_{Z_n}(t) = \left\{\left[\frac{t^2}{2}\right] + \cdots\right\} (\ other\ terms\ have\ positive\ powers\ of\ n\ in\ denominator)$$

$$\lim_{n\to\infty} \ln m_{Z_n}(t) = \frac{t^2}{2}\ (As\ n \to \infty\ other\ terms\ will\ become\ 0)$$

$$\therefore \lim_{n\to\infty} m_{Z_n}(t) = e^{\frac{t^2}{2}}\ \ (This\ is\ mgf\ of\ standard\ gaussian\ RV)$$

$$\therefore Z_n \xrightarrow{in\ distribution} \mathcal{N}(0,1).$$

□

Remark. *The above proof can be done similarly by considering characteristic functions instead of MGFs by taking $t = j\theta$. In this case also we will get the same result. We know that characteristic function exist for all RVs. So this proof is more general that it is applicable to RVs for which MGF doesn't exist.*

Note. *We have considered iid RVs for CLT. But in the proof we have used only one property of independent RVs which is $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. This property also holds for uncorrelated RVs. So our statement of CLT is also applicable to sequence of identically distributed and uncorrelated RVs.*

Use of the CLT

- A number of engineering applications utilize the Gaussian distribution in assuming the distribution of some variables, and this assumption is validated by the CLT. For instance, In practical communication system noise is due to addition of so many negligible noise sources which are modeled as Rvs (which are very difficult to characterize individually). Because of CLT the combined noise effect is Gaussian.
- The pdf of a sum of random variables is the convolution of their individual marginal pdfs. In the CLT, we see that the sum of i.i.d random variables (except for scaling and shifting factors) converges to a Gaussian RV. This means that if we take a function and keep convolving it with itself. This will converge to Gaussian.

14 Convergence of Random Variable sequences

In this Section, we want to discuss some aspects of convergence of a sequence of random variables. We have already looked at two specific cases of this. One in the WLLN, where we showed that a sequence of rvs converging to a constant, and the other in the CLT. To make things precise, first we discuss the convergence aspects of real number sequences.

14.1 Convergence of a sequence of real numbers

Consider a sequence of real numbers $x_1, x_2, \cdots, x_n, \cdots$. where $x_i \in \mathbb{R}, \forall i \in \mathbb{Z}^+$(positive integers).

Definition 14.1 (Limit operator)**.** *The sequence x_n is said to converge to a if the following condition is true*

- *for any $\epsilon > 0, \exists n_0 \in \mathbb{Z}^+$ such that $\forall n > n_0, |x_n - a| < \epsilon$.*

We then say that $\lim_{n\to\infty} x_n = a$

Noe that the above definition allows us to check convergence to a given number a. However the next lemma shows that finding a is not necessary if we have to only check convergence of the sequence. We give this lemma without proof.

Definition 14.2 (Cauchy sequence)**.** *A sequence $x_n \in \mathbb{R}, n \in \mathbb{Z}^+$ is said to be a* ***Cauchy sequence*** *if $\forall \epsilon > 0, \exists N \in \mathbb{Z}^+$ such that if $m, n \geq N$ then $|x_n - x_m| < \epsilon$.*

Lemma 14.1 (Cauchy convergence criterion)**.** *A sequence $x_n \in \mathbb{R}, n \in \mathbb{Z}^+$ is convergent iff x_n is a Cauchy sequence.*

Convergence of functions with Co-domain $\mathbb{R}$

We now discuss the convergence of a sequence of functions.

Definition 14.3. *The sequence $f_n : A_n \to \mathbb{R}, n \in \mathbb{Z}^+$ of functions converges to a function $f : A \to \mathbb{R}$ if the sequence of real numbers $f_1(x), f_2(x), \cdots$ converges to $f(x), \forall x \in A$.*

Class 21. (26/10/18) (After Noon) We are now ready to discuss the convergence aspects of sequences of random variables.

14.2 Convergence of sequences of RVs

Consider a sequence of RVs $X_1, X_2, \cdots, X_n, \cdots$ which are defined in the same probability space $(\Omega, \mathcal{F}, P)$. Each RV may have different distribution. For some RVs moments may exist (or) may not exist (or) only few moments may exist. We present a number of ways in which a sequence of RVs converges to one RV, which are called as modes of convergence.

14.3 1. Absolute convergence ($X_n \xrightarrow{\textit{absolute sense}} X$)

A sequence of RVs $X_n, n \in \mathbb{Z}^+$ converges to a random variable X in absolute sense if the functions $X_n, n \in \mathbb{Z}^+$ converges to the function X.

14.4 2. Almost sure convergence ($X_n \xrightarrow{\textit{almost surely}} X$)

A sequence of RVs $X_n, n \in \mathbb{Z}^+$ converges to a random variable X almost surely if $P\left(\left\{\omega \in \Omega : \lim\limits_{n\to\infty} X_n(\omega) \neq X(\omega)\right\}\right) = 0$. We can also write this as $P\left(\left\{\omega \in \Omega : \lim\limits_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1$.

Example 14.1. *Let $X_n, n \in \mathbb{Z}^+$ be a sequence of RVs defined in the probability space $(\Omega, \mathcal{F}, P)$ as*

$$X_n(\omega) = \begin{cases} 1 + \frac{1}{n} & \forall \omega \in \Omega \backslash \{\omega_0\} \\ 0 & \omega_0 \end{cases}$$

where $P(\omega_0) = 0$. Therefore

$$\lim_{n\to\infty} X_n(\omega) = \begin{cases} 1 & \forall \omega \in \Omega \backslash \{\omega_0\} \\ 0 & \omega_0 \end{cases}$$

Consider a new RV X which is defined as

$$X(\omega) = \begin{cases} 1 & \forall \omega \in \Omega \\ 0 & otherwise \end{cases}$$

Hence $P\left(\left\{\omega \in \Omega : \lim\limits_{n\to\infty} X_n(\omega) \neq X(\omega)\right\}\right) = P(\omega_0) = 0$. Therefore $X_n \xrightarrow{almost\ surely} X$. Note that X_n does not converge in the absolute sense to X in this case also, as the functions X_n does not converge to X at ω_0.

14.5 Convergence in r^{th} moment :

A sequence of RVs $X_n, n \in \mathbb{Z}^+$ converges to a random variable X in r^{th} moment if $\lim\limits_{n\to\infty} \mathbb{E}\left(|X_n - X|^r\right) = 0$. (provided r^{th} moment of all the RVs exist).

Note. *If $r = 2$ we get convergence in mean square sense. $X_n \xrightarrow{mean\ square} X$ if $\lim\limits_{n\to\infty} \mathbb{E}\left(|X_n - X|^2\right) = 0$.*

Example 14.2. *Let $X_n, n \in \mathbb{Z}^+$ be a sequence of i.i.d RVs in the probability space $(\Omega, \mathcal{F}, P)$. Let $S_n = \dfrac{\sum\limits_{i=1}^{n} X_i}{n}, n \in \mathbb{Z}^+$ be a new sequence of RVs, then from* ***Theorem 11.3*** *we have $\mathbb{E}(S_n) = \mathbb{E}(X_1)$ and $Var(S_n) = \mathbb{E}[((S_n - \mathbb{E}(S_n))^2] = \dfrac{Var(X_1)}{n}$. Consider $\lim\limits_{n\to\infty} \mathbb{E}[((S_n - \mathbb{E}(S_n))^2] = \lim\limits_{n\to\infty} \dfrac{Var(X_1)}{n} = 0$. Therefore $S_n \xrightarrow{mean\ square} \mathbb{E}(X_1)$.*

14.6 Convergence in probability ($X_n \xrightarrow{\textit{in probability}} X$)

A sequence of RVs $X_n, n \in \mathbb{Z}^+$ converges to a random variable X in probability if $\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$, for any $\epsilon > 0$.

Example 14.3. *Weak Law of Large Numbers. Refer proof of* ***Theorem 11.3***

Example 14.4. *Let $X_n, n \in \mathbb{Z}^+$ be a sequence of RVs defined in the probability space $(\Omega, \mathcal{F}, P)$ with PMF*

$$P_{X_n}(x) = \begin{cases} 1 - \dfrac{1}{n} & x = 0 \\ \dfrac{1}{n} & x = n \\ 0 & otherwise \end{cases}$$

Therefore for any $\epsilon > 0$ we have $(\forall n \in \mathbb{Z}^+)$

$$P(|X_n| > \epsilon) = \frac{1}{n}.$$

It is clear that $\lim_{n\to\infty} P(|X_n| > \epsilon) = 0$ which is same as $\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$, where X is a random variable taking value 0 with probability 1. Therefore $X_n \xrightarrow{in\ probability} X$.

14.7 Convergence in distribution($X_n \xrightarrow{\textit{in distribution}} X$)

A sequence of RVs $X_n, n \in \mathbb{Z}^+$ converges to a random variable X in distribution if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \forall x \in \mathbb{R}$, where F_{X_n} are CDFs of RVs $X_n, \forall n \in \mathbb{Z}^+$ and F_X is the CDF of RV X.

Note. *There are situations where a sequence of RVs whose CDFs converges to a function which doesn't satisfy the properties of CDF, then we say that this sequence of RVs doesn't converge in distribution.*

Example 14.5. *Let $X_n, n \in \mathbb{Z}^+$ be a sequence of RVs defined in the probability space $(\Omega, \mathcal{F}, P)$ as*

$$X_n(\omega) = \begin{cases} \dfrac{1}{n} & \forall \omega \in \Omega \\ 0 & otherwie \end{cases}$$

Therefore CDFs of these RVs are $(\forall n \in \mathbb{Z}^+)$

$$F_{X_n}(x) = \begin{cases} 0 & x < \dfrac{1}{n} \\ 1 & x \geq \dfrac{1}{n} \end{cases}$$

$$\lim_{n\to\infty} F_{X_n}(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

Consider a function

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

$F_X(x)$ satisfies all the properties of CDF. So it is the valid CDF of some RV X. Therefore $X_n \xrightarrow{in\ distribution} X$

Remark. $\left(X_n \xrightarrow{r^{th}\ moment} X\right) \Rightarrow \left(X_n \xrightarrow{(r-1)^{th}\ moment} X\right) \Rightarrow \cdots \Rightarrow \left(X_n \xrightarrow{2^{nd}\ moment} X\right) \Rightarrow \left(X_n \xrightarrow{1^{st}\ moment} X\right)$

Remark. $\left(X_n \xrightarrow{absolute\ sense} X\right) \Rightarrow \left(X_n \xrightarrow{almost\ surely} X\right) \Rightarrow \left(X_n \xrightarrow{in\ probability} X\right) \Rightarrow \left(X_n \xrightarrow{in\ distribution} X\right)$

Remark. $\left(X_n \xrightarrow{absolute\ sense} X\right) \Rightarrow \left(X_n \xrightarrow{r^{th}\ moment} X\right) \Rightarrow \left(X_n \xrightarrow{in\ probability} X\right) \Rightarrow \left(X_n \xrightarrow{in\ distribution} X\right)$

Note. *The converses to the above remarks are not true in general.*

Remark. *Let $X_n, n \in \mathbb{Z}^+$ be a sequence of i.i.d RVs in the probability space $(\Omega, \mathcal{F}, P)$. Let $S_n = \frac{\sum\limits_{i=1}^{n} X_i}{n}, n \in \mathbb{Z}^+$ be a new sequence of RVs. Then*

1. *$S_n \xrightarrow{almost\ surely} \mathbb{E}(X_1)$. This is known as Strong Law of Large Numbers (SLLN : we have not proved this in class as it requires more complicated mathematical machinery).*

2. *$S_n \xrightarrow{in\ probability} \mathbb{E}(X_1)$. This is known as Weak Law of Large Numbers (WLLN).*

Class 22. (30/10/18)

15 Jointly Gaussian Random Vector

We are now interested in delving deeper into Gaussians. In this section, we want to develop an understanding for a finite number of random variables which have a 'joint Gaussian' distribution. What this means and why this is important, will be revealed in the forthcoming sections.

We first prove the following simple result, which states that the sum of random variables which are Gaussian RVs is also a Gaussian RV.

Lemma 15.1. *If $X_i, i = 1, 2, \dots, l$ are independent gaussian RVs, then the RV $\sum\limits_{i=1}^{l} \alpha_i X_i + \beta$ $(\alpha_i, \beta \in \mathbb{R})$ is also gaussian distributed with mean $= \sum\limits_{i=1}^{l} \alpha_i \mathbb{E}(X_i) + \beta$ and variance $= \sum\limits_{i=1}^{l} \alpha_i^2 Var(X_i)$.*

Proof. Let $\sigma_i^2 = Var(X_i)$ and $\mu_i = \mathbb{E}(X_i)$. We use MGFs to prove it. Recall that the MGF of a Gaussian RV with mean μ and variance σ^2 is given by $e^{t\mu + \frac{\sigma^2 t^2}{2}}$. We now show that the MGF of the random variable $U = \sum\limits_{i=1}^{l} \alpha_i \mathbb{E}(X_i) + \beta$ also has the same form. Hence we can claim the distribution of U is also Gaussian with the requisite parameters.

$$\begin{aligned}
\mathbb{E}(e^{tU}) &= \mathbb{E}\left(e^{t\left(\sum\limits_{i=1}^{l} \alpha_i X_i + \beta\right)}\right) \\
&= e^{t\beta}\mathbb{E}\left(e^{\left(\sum\limits_{i=1}^{l} t\alpha_i X_i\right)}\right) \\
&= e^{t\beta}\mathbb{E}\left(\prod_{i=1}^{l} e^{(t\alpha_i X_i)}\right) \\
&= e^{t\beta}\prod_{i=1}^{l}\mathbb{E}\left(e^{(\alpha_i t)X_i}\right) (\textit{ Since } X_i, \forall i \in \{1, 2, \cdots\} \textit{ are independent}) \\
&= e^{t\beta}\prod_{i=1}^{l} M_{X_i}(\alpha_i t) \\
&= e^{t\beta}\prod_{i=1}^{l} e^{\left(\alpha_i t\mu_i + \frac{\alpha_i^2 t^2 \sigma_i^2}{2}\right)} \\
&= e^{t\beta} e^{\left(\sum\limits_{i=1}^{l}\left(\alpha_i t\mu_i + \frac{\alpha_i^2 t^2 \sigma_i^2}{2}\right)\right)} \\
&= e^{\left(\sum\limits_{i=1}^{l}(\alpha_i t\mu_i + t\beta) + \sum\limits_{i=1}^{l}\left(\frac{\alpha_i^2 t^2 \sigma_i^2}{2}\right)\right)} \\
&= e^{\left\{t\left(\sum\limits_{i=1}^{l}\alpha_i\mu_i + \beta\right) + \frac{t^2}{2}\left(\sum\limits_{i=1}^{l}(\alpha_i^2\sigma_i^2)\right)\right\}}.
\end{aligned}$$

This is MGF of a gaussian distribution with mean $= \sum\limits_{i=1}^{l} \alpha_i \mu_i + \beta$ and variance $= \sum\limits_{i=1}^{l} (\alpha_i^2 \sigma_i^2)$.Hence the proof. □

We start now with a few preliminary definitions. A *random vector* $\underline{X} = (X_1, ..., x_n)^T$ is nothing but a finite collection of jointly distributed random variables X_is. For the random vector $\underline{X}$ we can define the mean vector, denoted by $\mu_{\underline{X}}$ or $\mathbb{E}(\underline{X})$, as the vector $(\mathbb{E}(X_1), ..., \mathbb{E}(X_n))^T$.

We also need the idea of the covariance matrix of a random vector $\underline{X}$. The *covariance matrix*, $K_{\underline{X}}$ of $\underline{X}$ is an $n \times n$ matrix in which the $(i,j)^{th}$ entry is the covariance of the random variables X_i and X_j. By definition, the covariance matrix is a real matrix, furthermore it is also a *symmetric* matrix, i.e., $K_{\underline{X}}^T = K_{\underline{X}}$. Also observe that the diagonal entries of the covariance matrix are nothing but the variances of the random variables X_i. Finally the reader can easily check that the covariance matrix $K_{\underline{X}}$ can be expressed as $\mathbb{E}((\underline{X} - \mu_{\underline{X}})_{n\times 1}(\underline{X} - \mu_{\underline{X}})^T_{1\times n}))$.

As an example of a random vector, consider a random vector $\underline{W} = (W_1, ..., W_l)^T$ consisting of i.i.d standard normal random variables $W_i : i = 1, .., l$. Clearly the pdf of $\underline{W}$ can be written as

$$f_{\underline{W}}(\underline{w}) = f_{\underline{W}}(w_1, ..., w_l) = \prod_{i=1}^{l} f_{W_i}(w_i) = \prod_{i=1}^{l} \frac{1}{\sqrt{2\pi}} e^{-\frac{w_i^2}{2}} = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{||w||^2}{2}} = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{\underline{w}^T \underline{w}}{2}}$$

where $||w||^2 = \sum_{i=1}^{l} w_i^2$. What about the covariance matrix $K_{\underline{W}}$? Since the random variables W_i are iid standard normal, we have that the covariance matrix is just an $l \times l$ identity matrix.

We now proceed to define Joint Gaussian distributions.

Definition 15.1 (Jointly Gaussian Random Vector). *A vector $\underline{Z} = (Z_1, Z_2, \ldots, Z_n)^T$ of RVs which are jointly distributed is called a Gaussian Random Vector (or a jointly Gaussian random vector, or equivalently having a joint Gaussian distribution) if the following relationship is true.*

$$\underline{Z} = A\underline{W} + \underline{b},$$

for some $\underline{W}_{l\times 1}$ being a vector of iid standard normal RVs, and $\underline{b}_{n\times 1}$ is a constant real vector, and A is an $n \times l$ real matrix.

Note that for such a random vector, the mean vector is $\mathbb{E}(\underline{Z}) = A\mathbb{E}(\underline{W}) + \underline{b}$ (by linearity of expectation, and since $\underline{W}$ has zero mean). The covariance matrix $K_{\underline{Z}}$ can also be easily obtained as follows.

$$\begin{aligned}
K_{\underline{Z}} &= \mathbb{E}\left[(\underline{Z} - \mathbb{E}(\underline{Z}))(\underline{Z} - \mathbb{E}(\underline{Z}))^T\right] \\
&= \mathbb{E}\left[\{A\underline{W} - A\mathbb{E}(\underline{W})\}\{A\underline{W} - A\mathbb{E}(\underline{W})\}^T\right] \\
&= \mathbb{E}\left[A\{\underline{W} - \mathbb{E}(\underline{W})\}\{\underline{W} - \mathbb{E}(\underline{W})\}^T A^T\right] \\
&= A\mathbb{E}\left[\{\underline{W} - \mathbb{E}(\underline{W})\}\{\underline{W} - \mathbb{E}(\underline{W})\}^T\right] A^T \\
&= AI_lA^T \\
&\quad (\textit{Since covariance matrix of iid standard normal random vector is identity matrix}) \\
\therefore K_{\underline{Z}} &= AA^T.
\end{aligned}$$

Remark. *$det(k_{\underline{Z}}) = det(AA^T) = det(A)det(A^T) = [det(A)]^2$. Therefore $det(A) = \sqrt{det(K_{\underline{Z}})}$*

We are interested in finding the pdf of $\underline{Z}$. We focus on the special case when A is an $n \times n$ non-singular matrix and $\underline{b}$ is an n-length vector. The other cases (when A is non-square and/or singular) are left for the reader to dwell upon and arrive at the distribution.

In this case, we can rely upon our prior knowledge of computing the joint pdf of n functions of n RVs. We thus have that

$$f_{\underline{Z}}(\underline{z}) = \sum_{\underline{w}:A\underline{w}+\underline{b}=\underline{z}} f_{\underline{W}}(\underline{w}) \left|det(J)\Big|_{\underline{w}}\right|^{-1},$$

where $det(J)\Big|_{\underline{w}}$ is the determinant of the Jacobian matrix evaluated at the appropriate $\underline{w}$ within the same. It is easy to see that in this case we have only one $\underline{w}$ such that $A\underline{w} + \underline{b} = \underline{z}$, which is $\underline{w} = A^{-1}(\underline{z} - \underline{b})$. Further, the Jacobian J in this case can be easily calculated to be the matrix A itself, which when evaluated at $\underline{w}$ remains A.

Hence we have

$$
\begin{aligned}
f_{\underline{Z}}(\underline{z}) &= f_{\underline{W}}(A^{-1}(\underline{z}-\underline{b}))\frac{1}{det(A)} \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}det(A)}e^{-\left(\frac{||A^{-1}(\underline{z}-\underline{b})||^2}{2}\right)} \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}det(A)}e^{-\left(\dfrac{(\underline{z}-\underline{b})^T(AA^T)^{-1}(\underline{z}-\underline{b})}{2}\right)} \\
\therefore f_{\underline{Z}}(\underline{z}) &= \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{det(K_{\underline{Z}})}}e^{-\left(\frac{(\underline{z}-\underline{b})^T K_{\underline{Z}}^{-1}(\underline{z}-\underline{b})}{2}\right)}
\end{aligned}
$$

is the pdf of the jointly Gaussian random vector $\underline{Z}$.

Remark. *Note from the above expression that if the compoents of $\underline{Z}$ are uncorrelated, then $K_{\underline{Z}}$ is a diagonal matrix. In that case it is not difficult to see that the joint pdf $f_{\underline{Z}}(\underline{z})$ can be written as the product of the individual pdfs $f_{Z_i}(z_i),\ i = 1, .., n$. This means that jointly Gaussian random variables which are uncorrelated are also independent. This is a surprising but extremely useful property of the jointly Gaussian random variables, which we will also use henceforth.*

Class 23. (02/11/18)

15.1 Finding the matrix A given $K_{\underline{Z}}$

Assume that we have an input-output equation of the form $\underline{Y} = \underline{X} + \underline{Z}$, where all these are vectors of length n, $\underline{X}$ is the input and $\underline{Y}$ is the output, and $\underline{Z}$ is a noise vector with covariance matrix $\underline{Z}$. If the covariance matrix $K_{\underline{Z}}$ is a diagonal matrix, then this means that the components of $\underline{Z}$ are uncorrelated, which means that they are also independent. In this case, we can decode for each component X_i separately, using the i/o equation $Y_i = X_i + Z_i$. There is no loss of performance according to this decoding technique, because the other $Z_j : j \neq i$ does not contain any information about Z_i.

If $K_{\underline{Z}}$ is not a diagonal matrix, then it means that the components of $\underline{Z}$ are correlated with each other. This means that separate decoding for the X_is will result in a loss of performance, and thus the detection process to estimate $\underline{X}$ has to be done in one-shot, i.e., the decoder has to search through all possible transmissions for $\underline{X}$ and find the one which optimizes some cost-function, for instance minimizing the probability of error.

To overcome this latter situation of having a correlated noise vector, we can propose the following method. Suppose the decoder knows a matrix A such that $AA^T = K_{\underline{Z}}$. Then we can uncorrelate the noise in the equation by premultiplying the observed vector $\underline{Y}$ by A^{-1}. What happens then?

$$A^{-1}\underline{Y} = A^{-1}\underline{X} + A^{-1}\underline{Z}.$$

Now the covariance matrix of $A^{-1}Z$ happens to then be $A^{-1}K_{\underline{Z}}(A^{-1})^T = A^{-1}AA^T(A^{-1})^T = I$. Thus we see that the covariance matrix of the modified noise vector becomes an identity matrix, effectively render it uncorrelated. This means now we can decode for the vector $A^{-1}\underline{X}$ component-wise. To recover $\underline{X}$ we then only have to do premultiply the obtained vector by A.

Thus it is important to know, given the covariance matrix $K_{\underline{Z}}$, how to find out a matrix A such that $AA^T = K_{\underline{Z}}$.

Now suppose that we can find out a decomposition of the matrix $K_{\underline{Z}}$ as $K_{\underline{Z}} = UDU^T$, where $U_{n\times n}$ is an *orthonormal matrix* (an orthonormal matrix U is one such that $UU^T = I$), and D is a diagonal matrix. We can then assume that $A = UD^{\frac{1}{2}}U^T$, where $D^{\frac{1}{2}}$ is a *square-root matrix* of D, obtained by simply taking the square roots of the diagonal elements of D. It is easy to check that in that case we indeed have $AA^T = K_{\underline{Z}}$.

How do we now obtain the matrices U and D? Here linear algebra comes to our help. It turns out that we can take the matrix D to be the diagonal matrix in which the diagonal entries are the eigen values of $K_{\underline{Z}}$ (taken along with their algebraic repetition). The *normalized* eigen vectors (magnitude of each vector is made to be 1 by dividing the eigen vector by its magnitude) of $K_{\underline{Z}}$ are to be taken as the column vectors of U. With such assumptions we can indeed show that the equation $K_{\underline{Z}} = UDU^T$ holds. We skip the proof of this fact here as it involves a fair bit of linear algebra.

Random Processes

Course Notes of PRP offered by Dr. Prasad Krishnan, IIIT Hyderabad
prepared by Hari Hara Suthan C (corrections, mail to prasad.krishnan@iiit.ac.in)

Monsoon 2018

Class 24. (09/11/18) (Fore Noon)

1 Introduction

Definition 1.1 (Random Process (RP)). *Consider a probability space $(\Omega, \mathcal{F}, P)$. Let $\{X(t) : t \in \mathbb{R}\}$ be an indexed collection of Random variables. $\{X(t) : t \in \mathbb{R}\}$ is called a **Random Process** if for every $n \in \mathbb{Z}^+$, for any $t_1, t_2, \cdots t_n \in \mathbb{R}$ the Random variables $X(t_1), X(t_2), \cdots, X(t_n)$ have a joint distribution.*

Note. *Strictly speaking the Random Process has to be represented as $\{X(\omega, t) : \omega \in \Omega, t \in \mathbb{R}\}$. If we fix $t = t_1$ then we get a Random variable $\{X(\omega, t_1) : \omega \in \Omega\}$. If we fix $\omega = \omega_1$ then we get a deterministic(non-random) function of the variable t that is $\{X(\omega_1, t) : t \in \mathbb{R}\}$ which is called a single realization of the Random Process $X(\omega, t)$. If we sample a signal we get a real number. If we sample a Random Process we get a Random Variable. Collection of signals(wave forms) is called as ensemble. For all our practical purposes we treat t as time index.*

Definition 1.2 (Mean of a RP). *The mean of a random process $X(t)$ is denoted by $\mu_X(t)$ and is defined as $\mu_X(t) = \mathbb{E}[X(t)]$. (first order moment of the Random Process $X(t)$).*

The expectation operation kills the randomness in the random variable or random process. The expectation of a random process is the expectation of random variables present at all value of t. Therefore expectation of a random process is a deterministic function of time t.

Definition 1.3 (Auto correlation function of a RP). *The Auto correlation function of a random process $X(t)$ is denoted by $R_{XX}(t_1, t_2)$ and is defined as $R_{XX}(t_1, t_2) = \mathbb{E}[X(t_1)X(t_2)]$.*

Note. *$R_{XX}(t, t) = \mathbb{E}(X^2(t))$ (second order moment of the random process $X(t)$).*
As the expectation operation 'kills' the randomness $R_{XX}(t_1, t_2)$ is a deterministic function of t_1, t_2.

Definition 1.4 (Auto covariance function of a RP). *The Auto covariance function of a random process $X(t)$ is denoted by $K_{XX}(t_1, t_2)$ and is defined as $K_{XX}(t_1, t_2) = \mathbb{E}[\{X(t_1) - \mu_X(t_1)\}\{X(t_2) - \mu_X(t_2)\}]$.*

$$\begin{aligned}
K_{XX}(t_1, t_2) &= \mathbb{E}[\{X(t_1) - \mu_X(t_1)\}\{X(t_2) - \mu_X(t_2)\}] \\
&= \mathbb{E}[X(t_1)X(t_2) - X(t_1)\mu_X(t_2) - X(t_2)\mu_X(t_1) + \mu_X(t_1)\mu_X(t_2)] \\
&= \mathbb{E}[X(t_1)X(t_2)] - \mathbb{E}[X(t_1)]\mu_X(t_2) - \mathbb{E}[X(t_2)]\mu_X(t_1) + \mu_X(t_1)\mu_X(t_2) \\
&= \mathbb{E}[X(t_1)X(t_2)] - \mu_X(t_1)\mu_X(t_2) - \mu_X(t_2)\mu_X(t_1) + \mu_X(t_1)\mu_X(t_2) \\
\therefore K_{XX}(t_1, t_2) &= R_{XX}(t_1, t_2) - \mu_X(t_1)\mu_X(t_2)
\end{aligned}$$

Definition 1.5 (Cross correlation function of two Random Processes). *The Cross correlation function of two random process $X(t), Y(t)$ is denoted by $R_{XY}(t_1, t_2)$ and is defined as $R_{XY}(t_1, t_2) = \mathbb{E}[X(t_1)Y(t_2)]$.*

Definition 1.6 (Cross covariance function of two Random Processes). *The Cross covariance function of two random process $X(t), Y(t)$ is denoted by $K_{XY}(t_1, t_2)$ and is defined as $K_{XY}(t_1, t_2) = \mathbb{E}[\{X(t_1) - \mu_X(t_1)\}\{Y(t_2) - \mu_Y(t_2)\}]$.*

Class 25. (09/11/18) (After Noon)

Definition 1.7 (Stationary RP (or) Strict Sense Stationary RP). *A RP is said to be Strict Sense Stationary(SSS) RP if*

$$F_{X(t_1+\tau), X(t_2+\tau), \cdots, X(t_n+\tau)}(x_1, x_2, \cdots, x_n) = F_{X(t_1), X(t_2), \cdots X(t_n)}(x_1, x_2, \cdots, x_n)$$

holds $\forall t_i \in \mathbb{R}, i = 1, 2, \cdots, n$ and $n \in \mathbb{Z}^+, \forall \tau \in \mathbb{R}$.

Definition 1.8 (Wide Sense Stationary RP (or) Weak Sense Stationary RP)**.** *A RP is said to be Wide Sense Stationary(WSS) if*

1. $\mu_X(t)$ *is a constant function (independent of time).*
2. $R_{XX}(t_1, t_2) = R_{XX}(t_1+\tau, t_2+\tau)$, $\forall t_1, t_2, \tau \in \mathbb{R}$ *(i.e., the autocorrelation is the function of difference between the two sampling instants).*

Class 26. (13/11/18) Class 27. (16/11/18)

Note. *Refer to Gaussian random processes notes.*

Gaussian Noise Process (Additive White Gaussian Noise (AWGN))

Prasad Krishnan- Probability and Random Processes 2017 - IIIT Hyderabad

Note 1 *$\sum$ means $\sum_{k=-\infty}^{\infty}$ in general, unless some other limits are otherwise specified. Do not worry if the limits used in class were different, it does not matter as the results still go through.*

Note 2 *All matrices have appropriate dimensions (i.e. proper dimensions to allow multiplication of matrices).*

Definition 1 (Jointly Gaussian Random Vector GRV) *A random vector $\boldsymbol{Z} = (Z_1, ..., Z_n)$ is jointly Gaussian if*

$$\boldsymbol{Z} = A\boldsymbol{W} + \boldsymbol{b}$$

for some matrix A, some vector $\boldsymbol{b}$ and some $\boldsymbol{W}$ containing iid standard normal RVs.

Lemma 1 *If $\boldsymbol{Z}$ is a GRV, then so is $B\boldsymbol{Z}$ for any real matrix B.*

Definition 2 (Jointly Gaussian Random Process GRP) *A random process $Z(t)$ is a GRP if the random vector $(Z(t_1), ..., Z(t_n))$ is a GRV for all distinct $t_1, .., t_n$, for all $n \in \mathbb{Z}_+$.*

Other definitions to recall : Stationary RP, Wide-sense Stationary RP.

Theorem 1 *[Proved in class]If $Z(t)$ is a GRP and it is also WSS, then $Z(t)$ is also Stationary.*

Note 3 *Stationarity roughly means that the properties of the RP are constant through time (this is not a definition but a rough way of understanding).*

1 Examples and a unproved result

Example 1 *Let $Z(t) = \sum_{k=1}^{m} Z_k \phi_k(t)$ for some finite energy functions $\phi_k(t), k = 1, .., m$ and Z_k are independent Gaussian RVs. Then $Z(t)$ is a Gaussian Process.(Easy to check : Use definitions above and Lemma 1)*

Example 2 *Let $Z(t) = \sum Z_k\phi_k(t)$ (infinite sum), for some finite energy functions $\phi_k(t)$, and $Z_k s$ are independent zero-mean Gaussian RVs such that $\sum Var(Z_k) < \infty$. Then $Z(t)$ is a Gaussian RP. (Check using MGFs and Lemma 1).*

Definition 3 *A set of (possibly complex) functions $\phi_k(t), k \in \mathbb{Z}$ are orthonormal if $\int_{t=-\infty}^{t=\infty} \phi_i(t)\phi_j^*(t)dt = 1$ if $i = j$ and 0 if $i \neq j$, where $*$ indicates the conjugate. Clearly, if $\phi_k s$ are real, then $\phi_k^*(t) = \phi_k(t)$*

Theorem 2 *[Not proved in class] Let $\phi_k(t) : k \in \mathbb{Z}$ be a set of orthonormal functions. Then any Gaussian process $Z(t)$ can be written as*

$$Z(t) = \sum Z_k\phi_k(t).$$

where $Z_k s$ are some independent Gaussians.

Because of this above theorem, we can assume that all Gaussian RPs are essentially of the form of Example 2.

Theorem 3 *Let $h(t)$ be the finite energy impulse response of a filter. For $Z(t) = \sum Z_k\phi_k(t)$ (as in Example 2) is a Gaussian random process, the random process defined as*

$$V(t) = \int_{\tau=-\infty}^{\infty} Z(\tau)h(t-\tau)d\tau$$

is also a Gaussian RP.

The proof of this theorem is by expression $V(t)$ in the form of Example 2. Hence samples of this process are Gaussian.

2 Motivation and analysis regarding the filtered RP of a GRP and its samples

Henceforth $\phi_k(t)$s will always be orthonormal waveforms.

We have a communication channel in which we want to transmit a random process $U(t) = \sum U_k\phi_k(t)$. The information is completely encoded in the U_k RVs (which can take one of finite set of values, for instance $U_k \in \{+a, -a\}$ with uniform distribution) and the $\phi_k(t)$ waveforms are of the form $p(t-kT)$ (for some constant T), i.e. some pulse $p(t)$ shifted by kT (typically in practice). Note that $\phi_k(t)$s are known both to the transmitter and the receiver. If the channel is noiseless, then the decoding of U_k is simple. It is easy to see that

$$U_k = \int_{t=-\infty}^{\infty} U(t)\phi_k(t)dt.$$

Now suppose that the channel is noisy with additive noise process with zero-mean $Z(t)$, which is modelled as a Gaussian process (because of CLT). Because of Theorem 2, we can write $Z(t) = \sum Z_k\phi_k(t)$, where Z_ks are independent Gaussians with zero-mean.

Thus, now we have the receiver output (with noise) $Y(t) = U(t) + Z(t)$. We thus have

$$Y_k = U_k + Z_k = \int_{t=-\infty}^{\infty} U(t)\phi_k(t), \quad k \in \mathbb{Z}$$

These are our output samples. From these we have to recover the U_ks. To do this, we need to obtain the properties of Z_ks, their mean (which we know is zero), variance, independence properties, etc. The reason for knowing the independence properties amongst Z_ks is that if the Z_ks are not independent, then we would have to do a joint-detection for U_ks using several or all Y_ks together. This is much harder than using only one received sample Y_k to decode the corresponding U_k. Also, we want to know what is the variance of Z_k, because this will tell us the noise which we have to 'overcome' in the channel. The following sections discuss these aspects by making several further engineering assumptions which are valid about the noise process $Z(t)$. (Note that we have already assumed $Z(t)$ is a Gaussian process by CLT).

3 Stationary Gaussian Noise

We assume that $Z(t)$ is WSS (and hence also stationary by Theorem 1). This is a valid engineering assumption, in the sense that we can expect the properties of the noise process remain roughly same over time.

Hence we have for $\tau \in \mathbb{R}$,

$$R_Z(\tau) \triangleq \mathbb{E}(Z(t)Z(t+\tau)) = \mathbb{E}(Z(t')(t'+\tau))$$

for any t, t'. We can then define the Power Spectral Density of the random process $Z(t)$ as follows.

Definition 4 (Power Spectral Density of a WSS RP $Z(t)$)

$$S_Z(f) \triangleq \textit{Fourier transform of } R_Z(\tau) = \int_{\tau=-\infty}^{\infty} R_Z(\tau) e^{-j2\pi f\tau} d\tau.$$

Note that $R_Z(\tau)$ is a real function (as $Z(t)$ takes only real-values for fixed t) and also symmetric (by definition). Hence, $S_Z(f)$ must also be symmetric and real.

4 AWGN (Additive White Gaussian Noise)

We want to understand whether $Z_{k_1} = \int_{t=-\infty}^{\infty} Z(t)\phi_{k_1}(t)dt$ and $Z_{k_2} = \int_{t=-\infty}^{\infty} Z(t)\phi_{k_2}(t)dt$ are independent or not. Note that since Z_{k_1} and Z_{k_2} are both zero-mean Gaussians, thus it is sufficient to check their correlation (since covariance = correlation for zero mean RVs). Thus we have the correlation as follows.

$$\begin{aligned}
\mathbb{E}(Z_{k_1}Z_{k_2}) &= \mathbb{E}\left(\int_{\tau=-\infty}^{\infty} Z(\tau)\phi_{k_1}(\tau)d\tau \int_{t=-\infty}^{\infty} Z(t)\phi_{k_2}(t)dt\right) &(1)\\
&= \int_{\tau=-\infty}^{\infty}\int_{t=-\infty}^{\infty} \mathbb{E}(Z(\tau)Z(t))\phi_{k_1}(\tau)\phi_{k_2}(t)dtd\tau &(2)\\
&= \int_{\tau=-\infty}^{\infty}\int_{t=-\infty}^{\infty} R_Z(t-\tau)\phi_{k_1}(\tau)\phi_{k_2}(t)dtd\tau &(3)\\
&= \int_{t=-\infty}^{\infty} \phi_{k_2}(t)\left(\int_{\tau=-\infty}^{\infty} \phi_{k_1}(\tau)R_Z(t-\tau)d\tau\right)dt &(4)
\end{aligned}$$

Let $\int_{\tau=-\infty}^{\infty}\phi_{k_1}(\tau)R_Z(t-\tau)d\tau = r(t)$. Clearly, $r(t)$ is the convolution of $R_Z(t)$ and $\phi_{k_1}(t)$. Since both $R_Z(t)$ and $\phi_{k_1}(t)$ are both real, $r(t)$ is also a real function. Let $\hat{r}(f), \hat{\phi}_{k_1}(f)$ denote the Fourier transforms of $r(t)$ and $\phi_{k_1}(t)$. Then

$$\begin{aligned}
r(t) &= \int_{f=-\infty}^{\infty} \hat{r}(f)e^{j2\pi ft}df\\
&= \int_{f=-\infty}^{\infty} S_Z(f)\hat{\phi}_{k_1}(f)e^{j2\pi ft}df\\
&= r(t)^* = \left(\int_{f=-\infty}^{\infty} S_Z(f)\hat{\phi}_{k_1}(f)e^{j2\pi ft}df\right)^*\\
&= \int_{f=-\infty}^{\infty} S_Z(f)\hat{\phi}_{k_1}^*(f)e^{-j2\pi ft}df \qquad \text{(since } S_Z(f) \text{ is also real)}
\end{aligned}$$

Using the above in (4), we have

$$\begin{aligned}
\mathbb{E}(Z_{k_1}Z_{k_2}) &= \int_{t=-\infty}^{\infty} \phi_{k_2}(t)\left(\int_{f=-\infty}^{\infty} S_Z(f)\hat{\phi}_{k_1}^*(f)e^{-j2\pi ft}df\right)dt &(5)\\
&= \int_{f=-\infty}^{\infty} S_Z(f)\hat{\phi}_{k_1}^*(f)\left(\int_{t=-\infty}^{\infty} \phi_{k_2}(t)e^{-j2\pi ft}dt\right)df &(6)\\
\mathbb{E}(Z_{k_1}Z_{k_2}) &= \int_{f=-\infty}^{\infty} S_Z(f)\hat{\phi}_{k_1}^*(f)\hat{\phi}_{k_2}(f)df. &(7)
\end{aligned}$$

Now, suppose $S_Z(f)$ is constant (equal to $\frac{N_0}{2}$ for some real N_0) in the 'band of interest' (i.e., the frequency intervals where ϕ_{k_1}, ϕ_{k_2} are non-zero). This assumption will be used in Communications literature. This is a valid engineering assumption as long as the signal (sample functions of $U(t)$) varies 'much more slowly' compared to the sample functions of the noise $Z(t)$. This is typically ensured because the noise sample functions vary much more faster than the signal sample functions by taking long enough pulse-widths ($p(t)$ should have much smaller bandwidth compared to a noise waveform). Idealistically speaking, we may assume $S_Z(f) = \frac{N_0}{2}, \forall f$ (This is called as '**White Noise**' Process, since the PSD is constant for all frequencies and hence the noise is 'white')

This corresponds to the autocorrelation $R_Z(\tau) = \delta(\tau)$. Since the autocorrelation is measuring how much samples of $Z(t)$ at time-gap of τ are correlated to each other, '$R_Z(\tau) = \delta(\tau)$' means that the process $Z(t)$ is ideally random in the sense that samples taken at non-zero gap τ (however small) are completely uncorrelated.

Thus, by (7)

$$\begin{aligned}
\mathbb{E}(Z_{k_1} Z_{k_2}) &= \frac{N_0}{2} \int_{f=-\infty}^{\infty} \hat{\phi}^*_{k_1}(f) \hat{\phi}_{k_2}(f) df &\quad (8)\\
&= \frac{N_0}{2} \int_{f=-\infty}^{\infty} \phi_{k_2}(t) \phi_{k_1}(t) dt &\quad (9)\\
&= \frac{N_0}{2} \delta_{k_1,k_2} &\quad (10)
\end{aligned}$$

where $\delta_{k_1,k_2} = 1$ if $k_1 = k_2$ and 0 otherwise. Note that (9) follows by a method similar to (5) and (6), and (10) follows because of orthonormality of $\phi_k(t)$s.

Note that if $k_1 = k_2$, then $E(Z^2_{k_1}) = \frac{N_0}{2}$ is the variance of the noise sample Z_{k_1}. Also, since $\mathbb{E}(Z_{k_1} Z_{k_2}) = 0$ if $k_1 \neq k_2$, this means the random variables Z_{k_1} and Z_{k_2} are uncorrelated (and hence independent) if $k_1 \neq k_2$.

Because of this reason we have the following channel model for the Channel with Additive White Gaussian Noise (This is called the **AWGN Channel**).

$$Y_k = U_k + Z_k, \quad Z_k \sim \mathcal{N}(0, N_0/2).$$

Communication theory for AWGN Channels proceeds from this point. For detecting user samples U_k, the sample Y_k alone is sufficient (as long as U_ks are independent) as Z_ks are mutually uncorrelated.