E1 222 Stochastic Models and Applications

P.S. Sastry
sastry@iisc.ac.in

Reference Material

- V.K. Rohatgi and A.K.Md.E. Saleh, An Introduction to probability and Statistics, Wiley, 2nd edition, 2018
- S.Ross, 'Introduction to Probability Models', Elsevier, 12th edition, 2019.
- P G Hoel, S Port and C Stone, Introduction to Probability Theory, 1971.
- P G Hoel, S Port and C Stone, Introduction to Stochastic Processes, 1971.

Course Prerequisites / Background needed

- Calculus
  - continuity, differentiability, derivatives, functions of several variables, partial derivatives, integration, multiple integrals or integration over $\Re^n$, convergence of sequences and series, Taylor series
- Matrix theory
  - vector spaces, linear independence, linear transformations, matrices, rank, determinant, eigen values and eigen vectors
- In addition, knowledge of basic probability is assumed. I assume all students are familiar with the following:

  Random experiment, sample space, events, conditional probability, independent events, simple combinatorial probability computations

  But we would review the basic probability in the first two classes.

Course grading

- Mid-Term Tests and Assignments: 70%
  Final Exam: 30%
- Three mid-term tests for 20 marks each. We will have 2-3 assignments for 10 marks. (Tentative)
- Please remember this is essentially a Maths course

Probability Theory

- Probability Theory – branch of mathematics that deals with modeling and analysis of random phenomena.
- Random Phenomena – "individually not predictable but have a lot of regularity at a population level"
- E.g., Recommender systems are useful for Amazon or Netflix because at a population level customer behaviour can be predicted well.
- Example random phenomena: Tossing a coin, rolling a dice etc – familiar to you all
- It is also useful in many engineering systems, e.g., for taking care of noise.
- Probability theory is also needed for Statistics that deals with making inferences from data.

- In many engineering problems one needs to deal with random inputs where probability models are useful
  - Dealing with dynamical systems subjected to noise (e.g., Kalman filter)
  - Policies for decision making under uncertainty
  - Pattern Recognition, prediction from data

    $\vdots$
- We may use probability models for analysing algorithms. (e.g., average case complexity of algorithms)
- We may deliberately introduce randomness in an algorithm
  (e.g., ALOHA protocol, Primality testing)

  $\vdots$

This is only a 'sample' of possible application scenarios!

Review of basic probability

We assume all of you are familiar with the terms:

random experiment, outcomes of random experiment, sample space, events etc.

We use the following Notation:

- Sample space – Ω
  Elements of Ω are the outcomes of the random experiment
  We write $\Omega = \{\omega_1, \omega_2, \cdots\}$ when it is countable
- An event is, by definition, a subset of Ω
- Set of all possible events – $\mathcal{F} \subset 2^\Omega$ (power set of Ω)
  Each event is a subset of Ω

Probability axioms

Probability (or probability measure) is a function that assigns a number to each event and satisfies some properties.

Formally, $P : \mathcal{F} \to \Re$ satisfying

A1 Non-negativity: $P(A) \geq 0, \forall A \in \mathcal{F}$

A2 Normalization: $P(\Omega) = 1$

A3 σ-additivity: If $A_1, A_2, \cdots \in \mathcal{F}$ satisfy $A_i \cap A_j = \phi, \forall i \neq j$ then

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Events satisfying $A_i \cap A_j = \phi, \forall i \neq j$ are said to be **mutually exclusive**

Probability axioms

$P : \mathcal{F} \to \Re, \ \ \mathcal{F} \subset 2^{\Omega}$ (Events are subsets of Ω)

A1 $P(A) \geq 0, \ \forall A \in \mathcal{F}$

A2 $P(\Omega) = 1$

A3 If $A_i \cap A_j = \phi, \forall i \neq j$ then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

- For these axioms to make sense, we are assuming

  $$\text{(i). } \Omega \in \mathcal{F} \ \text{ and } \ \text{(ii). } A_1, A_2, \cdots \in \mathcal{F} \Rightarrow (\cup_i A_i) \in \mathcal{F}$$

  When $\mathcal{F} = 2^{\Omega}$ this is true.

Simple consequences of the axioms

- Notation: A^c is complement of A.
  $C = A + B$ implies A, B are mutually exclusive and C is their union.
- Let $A \subset B$ be events. Then $B = A + (B - A)$.
  Now we can show $P(A) \leq P(B)$:

  $$P(B) = P\left(A + (B - A)\right) = P(A) + P(B - A) \geq P(A)$$

  This also shows $P(B - A) = P(B) - P(A)$ when $A \subset B$.

- There are many such properties (I assume familiar to you) that can be derived from the axioms.
- Here are a few important ones. (Proof is left to you as an exercise!)

Case of finite Ω – Example

- Let $\Omega = \{\omega_1, \cdots, \omega_n\}$, $\mathcal{F} = 2^{\Omega}$, and P is specified through 'equally likely' assumption.
- That is, $P(\{\omega_i\}) = \frac{1}{n}$. (Note the notation)
- Suppose $A = \{\omega_1, \omega_2, \omega_3\}$. Then

$$P(A) = P(\{\omega_1\} \cup \{\omega_2\} \cup \{\omega_3\}) = \sum_{i=1}^{3} P(\{\omega_i\}) = \frac{3}{n} = \frac{|A|}{|\Omega|}$$

- We can easily see this to be true for any event, A.
- This is the usual familiar formula: number of favourable outcomes by total number of outcomes.
- Thus, 'equally likely' is one way of specifying the probability function (in case of finite Ω).
- An obvious point worth remembering: specifying P for singleton events fixes it for all other events.

Review of basic probability

We use the following Notation:

- Sample space – Ω
  Elements of Ω are the outcomes of the random experiment
  We write $\Omega = \{\omega_1, \omega_2, \cdots\}$ when it is countable
- An event is, by definition, a subset of Ω
- Set of all possible events – $\mathcal{F} \subset 2^\Omega$ (power set of Ω)
  Each event is a subset of Ω
  For now, we take $\mathcal{F} = 2^\Omega$ (power set of Ω)

Probability axioms

Probability (or probability measure) is a function that assigns a number to each event and satisfies some properties.
$P : \mathcal{F} \to \Re, \ \ \mathcal{F} \subset 2^\Omega$

A1 $P(A) \geq 0, \forall A \in \mathcal{F}$

A2 $P(\Omega) = 1$

A3 If $A_i \cap A_j = \phi, \forall i \neq j$ then $P(\cup_{i=1}^\infty A_i) = \sum_{i=1}^\infty P(A_i)$

Some consequences of the axioms

- $0 \leq P(A) \leq 1$
- $P(A^c) = 1 - P(A)$
- If $A \subset B$ then, $P(A) \leq P(B)$
- If $A \subset B$ then, $P(B - A) = P(B) - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Case of finite Ω – Example

- Let $\Omega = \{\omega_1, \cdots, \omega_n\}$, $\mathcal{F} = 2^\Omega$, and P is specified through 'equally likely' assumption.
- That is, $P(\{\omega_i\}) = \frac{1}{n}$. (Note the notation)
- Suppose $A = \{\omega_1, \omega_2, \omega_3\}$. Then

$$P(A) = P(\{\omega_1\} \cup \{\omega_2\} \cup \{\omega_3\}) = \sum_{i=1}^{3} P(\{\omega_i\}) = \frac{3}{n} = \frac{|A|}{|\Omega|}$$

- We can easily see this to be true for any event, A.
- This is the usual familiar formula: number of favourable outcomes by total number of outcomes.
- Thus, 'equally likely' is one way of specifying the probability function (in case of finite Ω).
- An obvious point worth remembering: specifying P for singleton events fixes it for all other events.

Case of Countably infinite Ω

- Let $\Omega = \{\omega_1, \omega_2, \cdots\}$.
- Once again, any $A \subset \Omega$ can be written as mutually exclusive union of singleton sets.
- Let $q_i, i = 1, 2, \cdots$ be numbers such that $q_i \geq 0$ and $\sum_i q_i = 1$.
- We can now set $P(\{\omega_i\}) = q_i, i = 1, 2, \cdots$. (Assumptions on q_i needed to satisfy $P(A) \geq 0$ and $P(\Omega) = 1$).
- This fixes P for all events: $P(A) = \sum_{\omega \in A} P(\{\omega\})$
- This is how we normally define a probability measure on countably infinite Ω.
- This can be done for finite Ω too.

Example: countably infinite Ω

- Consider a random experiment of tossing a biased coin repeatedly till we get a head. We take the outcome of the experiment to be the number of tails we had before the first head.
- Here we have $\Omega = \{0, 1, 2, \cdots\}$.
- A (reasonable) probability assignment is:

$$P(\{k\}) = (1-p)^k p, k = 0, 1, \cdots$$

  where p is the probability of head and $0 < p < 1$. (We assume you understand the idea of 'independent' tosses here).
- In the notation of previous slide, $q_i = (1-p)^i p$
- Easy to see we have $q_i \geq 0$ and $\sum_{i=0}^{\infty} q_i = 1$.

Case of uncountably infinite Ω

- We would mostly be considering only the cases where Ω is a subset of $\Re^d$ for some d.
- Note that now an event need not be a countable union of singleton sets.
- For now we would only consider a simple intuitive extension of the 'equally likely' idea.
- Suppose Ω is a finite interval of $\Re$. Then we will take $P(A) = \frac{m(A)}{m(\Omega)}$ where $m(A)$ is length of the set A.
- We can use this in higher dimensions also by taking $m(\cdot)$ to be an appropriate 'measure' of a set.
- For example, in $\Re^2$, $m(A)$ denotes area of A, in $\Re^3$ it would be volume and so on.

  (There are many issues that need more attention here).

Example: Uncountably infinite Ω

Problem: A rod of unit length is broken at two random points. What is the probability that the three pieces so formed would make a triangle.

- Let us take left end of the rod as origin and let x, y denote the two successive points where the rod is broken.
- Then the random experiment is picking two numbers x, y with $0 < x < y < 1$.
- We can take $\Omega = \{(x, y) \ : \ 0 < x < y < 1\} \subset \Re^2$.
- For the pieces to make a triangle, sum of lengths of any two should be more than the third.

- The lengths are: $x, (y-x), (1-y)$. So we need

$$x + (y-x) > (1-y) \;\Rightarrow\; y > 0.5$$

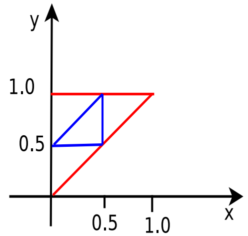
$$x + (1-y) > (y-x) \;\Rightarrow\; y < x + 0.5;$$

$$(y-x) + 1 - y > x \;\Rightarrow\; x < 0.5$$

- So the event of interest is:
  $A = \{(x,y) \;:\; y > 0.5;\; x < 0.5;\; y < x + 0.5\}$

- We have

$$\begin{aligned} \Omega &= \{(x,y) \;:\; 0 < x < y < 1\} \\ A &= \{(x,y) \;:\; y > 0.5;\; x < 0.5;\; y < x + 0.5\} \end{aligned}$$



- We can visualize it as follows
- The required probability is area of A divided by area of Ω which gives the answer as 0.25

- Everything we do in probability theory is always in reference to an underlying probability space: $(\Omega, \mathcal{F}, P)$ where
  - Ω is the sample space
  - $\mathcal{F} \subset 2^{\Omega}$ set of events; each event is a subset of Ω
  - $P : \mathcal{F} \to [0,1]$ is a probability (measure) that assigns a number between 0 and 1 to every event (satisfying the three axioms).

Conditional Probability

- Let B be an event with $P(B) > 0$. We define conditional probability, conditioned on B, of any event, A, as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(AB)}{P(B)}$$

- The above is a notation. "$A \mid$ B" does not represent any set operation! (This is an abuse of notation!)
- Given a B, conditional probability is a new probability assignment to any event.
- That is, given B with $P(B) > 0$, we define a new probability $P_B : \mathcal{F} \to [0,1]$ by

$$P_B(A) = \frac{P(AB)}{P(B)}$$

- Conditional probability is a probability. What does this mean?
- The new function we defined, $P_B : \mathcal{F} \to [0, 1]$, $P_B(A) = \frac{P(AB)}{P(B)}$, satisfies the three axioms of probability.
- $P_B(A) \geq 0$ and $P_B(\Omega) = 1$.
- If A_1, A_2 are mutually exclusive then A_1B and A_2B are also mutually exclusive and hence

$$
\begin{aligned}
P_B(A_1 + A_2) &= \frac{P((A_1 + A_2)B)}{P(B)} = \frac{P(A_1B + A_2B)}{P(B)} \\
&= \frac{P(A_1B) + P(A_2B)}{P(B)} = P_B(A_1) + P_B(A_2)
\end{aligned}
$$

- Once we understand condional probability is a new probability assignment, we go back to the 'standard notation'

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$

- Note $P(B|B) = 1$ and $P(A|B) > 0$ only if $P(AB) > 0$.
- Now the 'new' probability of each event is determined by what it has in common with B.
- If we know the event B has occurred, then based on this knowledge we can readjust probabilities of all events and that is given by the conditional probability.
- Intuitively it is as if the sample space is now reduced to B because we are given the information that B has occurred.
- This is a useful intuition as long as we understand it properly.
- It is not as if we talk about conditional probability only for subsets of B. Conditional probability is also with respect to the original probability space. Every element of $\mathcal{F}$ has conditional probability defined.

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$

- Suppose $P(A \mid B) > P(A)$
  Does it mean "B **causes** A"?

$$
\begin{aligned}
P(A \mid B) > P(A) &\Rightarrow P(AB) > P(A)P(B) \\
&\Rightarrow \frac{P(AB)}{P(A)} > P(B) \\
&\Rightarrow P(B \mid A) > P(B)
\end{aligned}
$$

- Hence, conditional probabilities cannot actually capture causal influences.
- There are probabilistic methods to capture causation (but far beyond the scope of this course!)

- In a conditional probability, the conditioning event can be any event (with positive probability)
- In particular, it could be intersection of events.
- We think of that as conditioning on multiple events.

$$P(A \mid B, C) = P(A \mid BC) = \frac{P(ABC)}{P(BC)}$$

- The conditional probability is defined by

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$

- This gives us a useful identity

$$P(AB) = P(A \mid B)P(B)$$

- We can iterate this for multiple events

$$P(ABC) = P(A \mid BC)P(BC) = P(A \mid BC)P(B \mid C)P(C)$$

- Let $B_1, \cdots, B_m$ be events such that $\cup_{i=1}^{m} B_i = \Omega$ and $B_i B_j = \phi, \forall i \neq j$.
- Such a collection of events is said to be a partition of Ω. (They are also sometimes said to be mutually exclusive and collectively exhaustive).
- Given this partition, any other event can be represented as a mutually exclusive union as

$$A = AB_1 + \cdots + AB_m$$

To explain the notation again

$$A = A \cap \Omega = A \cap (B_1 \cup \cdots \cup B_m) = (A \cap B_1) \cup \cdots \cup (A \cap B_m)$$

Hence, $A = AB_1 + \cdots + AB_m$

Total Probability rule

- Let $B_1, \cdots, B_m$ be a partition of Ω.
- Then, for any event A, we have

$$\begin{aligned} P(A) &= P(AB_1 + \cdots + AB_m) \\ &= P(AB_1) + \cdots + P(AB_m) \\ &= P(A \mid B_1)P(B_1) + \cdots + P(A \mid B_m)P(B_m) \end{aligned}$$

- The formula (where B_i form a partition)

$$P(A) = \sum_i P(A \mid B_i)P(B_i)$$

is known as **total probability rule** or total probability law or total probability formula.
- This is a very useful in many situations. ("arguing by cases")

Example: Polya's Urn

An urn contains r red balls and b black balls. We draw a ball at random, note its color, and put back that ball along with c balls of the same color. We keep repeating this process. Let R_n (B_n) denote the event of drawing a red (black) ball at the n^{th} draw. We want to calculate the probabilities of all these events.

- It is easy to see that $P(R_1) = \frac{r}{r+b}$ and $P(B_1) = \frac{b}{r+b}$.
- For R_2 we have, using total probability rule,

$$\begin{aligned} P(R_2) &= P(R_2 \mid R_1)P(R_1) + P(R_2 \mid B_1)P(B_1) \\ &= \frac{r+c}{r+c+b}\,\frac{r}{r+b} + \frac{r}{r+b+c}\,\frac{b}{r+b} \\ &= \frac{r(r+c+b)}{(r+c+b)(r+b)} = \frac{r}{r+b} = P(R_1) \end{aligned}$$

- Similarly we can show that $P(B_2) = P(B_1)$.
- One can show by mathematical induction that $P(R_n) = P(R_1)$ and $P(B_n) = P(B_1)$ forall n. (Left as an exercise for you!)
- This does not depend on the value of c!

Bayes Rule

- Another important formula based on conditional probability is Bayes Rule:

$$P(A \mid B) = \frac{P(AB)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}$$

- This allows one to calculate $P(A \mid B)$ if we know $P(B \mid A)$.
- Useful in many applications because one conditional probability may be more easier to obtain (or estimate) than the other.
- Often one uses total probability rule to calculate the denominator in the RHS above:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)}$$

Example: Bayes Rule

Let D and D^c denote someone being diagnosed as having a disease or not having it. Let T_+ and T_- denote the events of a test for it being positive or negative. (Note that $T_+^c = T_-$). We want to calculate $P(D|T_+)$.

- We have, by Bayes rule,

$$P(D|T_+) = \frac{P(T_+|D)P(D)}{P(T_+|D)P(D) + P(T_+|D^c)P(D^c)}$$

- The probabilities $P(T_+|D)$ and $P(T_+|D^c)$ can be obtained through, for example, laboratory experiments.
- $P(T_+|D)$ is called the true positive rate and $P(T_+|D^c)$ is called false positive rate.
- We also need $P(D)$, the probability of a random person having the disease.

- Let us take some specific numbers
- Let: $P(D) = 0.5$, $P(T_+|D) = 0.99$, $P(T_+|D^c) = 0.05$.

$$P(D|T_+) = \frac{0.99 * 0.5}{0.99 * 0.5 + 0.05 * 0.5} = 0.95$$

  That is pretty good.
- But taking $P(D) = 0.5$ is not realistic. Let us take $P(D) = 0.1$.

$$P(D|T_+) = \frac{0.99 * 0.1}{0.99 * 0.1 + 0.05 * 0.9} = 0.69$$

- Now suppose we can improve the test so that $P(T_+|D^c) = 0.01$

$$P(D|T_+) = \frac{0.99 * 0.1}{0.99 * 0.1 + 0.01 * 0.9} = 0.92$$

- These different cases are important in understanding the role of false positives rate.

- $P(D)$ is the probability that a random person has the disease. We call it the prior probability.
- $P(D|T_+)$ is the probability of the random person having disease once we do a test and it came positive. We call it the posterior probability.
- Bayes rule essentially transforms the prior probability to posterior probability.

- In many applications of Bayes rule the same generic situation exists
- Based on a measurement we want to predict (what may be called) the state of nature.
- For another example, take a simple communication system.
  - D can represent the event that the transmitter sent bit 1.
  - T_+ can represent an event about the measurement we made at the receiver.
  - We want the probability that bit 1 is sent based on the measurement.
  - The knowledge we need is $P(T_+|D)$, $P(T_+|D^c)$ which can be determined through experiment or modelling of channel.

$$P(D|T_+) = \frac{P(T_+|D)P(D)}{P(T_+|D)P(D) + P(T_+|D^c)P(D^c)}$$

- Not all applications of Bayes rule involve a ‘binary’ situation
- Suppose D_1, D_2, D_3 are the (exclusive) possibilities and T is an event about a measurement.

$$\begin{aligned} P(D_1|T) &= \frac{P(T|D_1)P(D_1)}{P(T)} \\ &= \frac{P(T|D_1)P(D_1)}{P(T|D_1)P(D_1) + P(T|D_2)P(D_2) + P(T|D_3)P(D_3)} \\ &= \frac{P(T|D_1)P(D_1)}{\sum_i P(T|D_i)P(D_i)} \end{aligned}$$

$$P(D|T_+) = \frac{P(T_+|D)P(D)}{P(T_+|D)P(D) + P(T_+|D^c)P(D^c)}$$

- In the binary situation we can think of Bayes rule in a slightly modified form too.

$$\frac{P(D|T_+)}{P(D^c|T_+)} = \frac{P(T_+|D)}{P(T_+|D^c)} \frac{P(D)}{P(D^c)}$$

- This is called the odds-likelihood form of Bayes rule (The ratio of $P(A)$ to $P(A^c)$ is called odds for A)

Independent Events

- Two events A, B are said to be independent if

$$P(AB) = P(A)P(B)$$

- Note that this is a definition. Two events are independent if and only if they satisfy the above.
- Suppose $P(A), P(B) > 0$. Then, if they are independent

$$P(A|B) = \frac{P(AB)}{P(B)} = P(A); \text{ similarly } P(B|A) = P(B)$$

- This gives an intuitive feel for independence.
- Independence is an important (often confusing!) concept.

Example: Independence

A class has 20 female and 30 male course (MTech) students and 6 female and 9 male research (PhD) students. Are gender and degree independent?

- Let F, M, C, R denote events of female, male, course, research students
- From the given numbers, we can easily calculate the following:

$$P(F) = \frac{26}{65} = \frac{2}{5};\ P(C) = \frac{50}{65} = \frac{10}{13};\ P(FC) = \frac{20}{65} = \frac{4}{13}$$

- Hence we can verify

$$P(F)P(C) = \frac{2}{5}\frac{10}{13} = \frac{4}{13} = P(FC)$$

and conclude that F and C are independent. Similarly we can show for others.

- In this example, if we keep all other numbers same but change the number of male research students to, say, 12 then the independence no longer holds. $(\frac{26}{68}\frac{50}{68} \neq \frac{20}{68})$
- One needs to be careful about independence!
- We always have an underlying probability space $(\Omega, \mathcal{F}, P)$
- Once that is given, the probabilities of all events are fixed.
- Hence whether or not two events are independent is a matter of 'calculation'

- If A and B are independent then so are A and B^c.
- Using $A = AB + AB^c$, we have

$$P(AB^c) = P(A) - P(AB) = P(A)(1 - P(B)) = P(A)P(B^c)$$

- This also shows that A^c and B are independent and so are A^c and B^c.
- For example, in the previous problem, once we saw that F and C are independent, we can conclude M and C are also independent (because in this example we are taking $F^c = M$).

- Consider the random experiment of tossing two fair coins (or tossing a coin twice).
- $\Omega = \{HH, HT, TH, TT\}$.
  Suppose we employ 'equally likely idea'.
- That is, $P(\{HH\}) = \frac{1}{4}$, $P(\{HT\}) = \frac{1}{4}$ and so on
- Let A = 'H on 1st toss' $= \{HH, HT\}$ $(P(A) = \frac{1}{2})$
  Let B = 'T on second toss' $= \{HT, TT\}$ $(P(B) = \frac{1}{2})$
- We have $P(AB) = P(\{HT\}) = 0.25$
- Since $P(A)P(B) = \frac{1}{2}\frac{1}{2} = \frac{1}{4} = P(AB)$,
  A, B are independent.
- Hence, in multiple tosses, assuming all outcomes are equally likely implies outcome of one toss is independent of another.

- In multiple tosses, assuming all outcomes are equally likely is alright if the coin is fair.
- Suppose we toss a biased coin two times.
- Then the four outcomes are, obviously, not 'equally likely'
- How should we then assign these probabilities?
- If we assume tosses are independent then we can assign probabilities easily.

- Consider toss of a biased coin:
  $\Omega^1 = \{H, T\}$, $P(\{H\}) = p$ and $P(\{T\}) = 1 - p$.
- If we toss this twice then $\Omega^2 = \{HH, HT, TH, TT\}$ and we assign
  $P(\{HH\}) = p^2$, $P(\{HT\}) = p(1-p)$,
  $P(\{TH\}) = (1-p)p$, $P(\{TT\}) = (1-p)^2$.
- $P(\{HH, HT\}) = p^2 + p(1-p) = p$
- This assignment ensures that $P(\{HH\})$ equals product of probability of H on 1st toss and H on second toss.
- Essentially, Ω^2 is a cartesian product of Ω^1 with itself and essentially we used products of the corresponding probabilities.
- For any independent repetitions of a random experiment we follow this.
  (We will look at it more formally when we consider multiple random variables).

- In many situations calculating probabilities of intersection of events is difficult.
- One often **assumes** A and B are independent to calculate $P(AB)$.
- As we saw, if A and B are independent, then $P(A|B) = P(A)$
- This is often used, at an intuitive level, to justify assumption of independence.

Independence of multiple events

- Events $A_1, A_2, \cdots, A_n$ are said to be (totally) independent if for any k, $1 \le k \le n$, and any indices $i_1, \cdots, i_k$, we have
$$P(A_{i_1} \cdots A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k})$$
- For example, A, B, C are independent if
$$P(AB) = P(A)P(B);\ P(AC) = P(A)P(C);$$
$$P(BC) = P(B)P(C);\ P(ABC) = P(A)P(B)P(C)$$

Pair-wise independence

- Events $A_1, A_2, \cdots, A_n$ are said to be pair-wise independent if
$$P(A_i A_j) = P(A_i)P(A_j),\ \forall i \ne j$$
- Events may be pair-wise independent but not (totally) independent.
- Example: Four balls in a box inscribed with '1', '2', '3' and '123'. Let E_i be the event that number 'i' appears on a radomly drawn ball, $i = 1, 2, 3$.
- Easy to see: $P(E_i) = 0.5,\ i = 1, 2, 3.$
- $P(E_i E_j) = 0.25\ (i \ne j) \Rightarrow$ pairwise independent
- But, $P(E_1 E_2 E_3) = 0.25 \ne (0.5)^3$

Recap

- Everything we do in probability theory is always in reference to an underlying probability space: $(\Omega, \mathcal{F}, P)$ where
  - Ω is the sample space
  - $\mathcal{F} \subset 2^{\Omega}$ set of events; each event is a subset of Ω
  - $P : \mathcal{F} \to [0, 1]$ is a probability (measure) that satisfies the three axioms:
    - A1 $P(A) \ge 0, \forall A \in \mathcal{F}$
    - A2 $P(\Omega) = 1$
    - A3 If $A_i \cap A_j = \phi, \forall i \ne j$ then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Recap

- When $\Omega = \{\omega_1, \omega_2, \cdots\}$ (is countable), then probability is generally assigned by
$$P(\{\omega_i\}) = q_i,\ i = 1, 2, \cdots,\ \text{with } q_i \ge 0,\ \sum_i q_i = 1$$
- When Ω is finite with n elements, a special case is $q_i = \frac{1}{n}, \forall i$. (All outcomes equally likely)

Recap

- Conditional probability of A given (or conditioned on) B is

$$P(A|B) = \frac{P(AB)}{P(B)}$$

- This gives us the identity: $P(AB) = P(A|B)P(B)$
- This holds for multiple event, e.g., $P(ABC) = P(A|BC)P(B|C)P(C)$
- Given a partition, $\Omega = B_1 + B_2 + \cdots + B_m$, for any event, A,

$$P(A) = \sum_{i=1}^{m} P(A|B_i)P(B_i) \quad \text{(Total Probability rule)}$$

Recap

- Bayes Rule

$$P(D|T) = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^c)P(D^c)}$$

- Bayes rule can be viewed as transforming a prior probability into a posterior probability.

Recap: Independence

- Two events A, B are said to be independent if

$$P(AB) = P(A)P(B)$$

- Suppose $P(A), P(B) > 0$. Then, if they are independent

$$P(A|B) = \frac{P(AB)}{P(B)} = P(A); \;\; \text{similarly } P(B|A) = P(B)$$

- If A, B are independent then so are $A \& B^c$, $A^c \& B$ and $A^c \& B^c$.

Independence of multiple events

- Events $A_1, A_2, \cdots, A_n$ are said to be (totally) independent if for any k, $1 \le k \le n$, and any indices $i_1, \cdots, i_k$, we have

$$P(A_{i_1} \cdots A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k})$$

- For example, A, B, C are independent if

$$P(AB) = P(A)P(B); \;\; P(AC) = P(A)P(C);$$

$$P(BC) = P(B)P(C); \;\; P(ABC) = P(A)P(B)P(C)$$

Pair-wise independence

- Events $A_1, A_2, \cdots, A_n$ are said to be pair-wise independent if

$$P(A_i A_j) = P(A_i)P(A_j), \ \forall i \neq j$$

- Events may be pair-wise independent but not (totally) independent.
- Example: Four balls in a box inscribed with '1', '2', '3' and '123'. Let E_i be the event that number 'i' appears on a radomly drawn ball, $i = 1, 2, 3$.
- Easy to see: $P(E_i) = 0.5, \ i = 1, 2, 3$.
- $P(E_i E_j) = 0.25 \ (i \neq j) \Rightarrow$ pairwise independent
- But, $P(E_1 E_2 E_3) = 0.25 \neq (0.5)^3$

Conditional Independence

- Events A, B are said to be (conditionally) independent given C if

$$P(AB|C) = P(A|C)P(B|C)$$

- If the above holds

$$P(A|BC) = \frac{P(ABC)}{P(BC)} = \frac{P(AB|C)P(C)}{P(BC)}$$

$$= \frac{P(A|C) \ P(B|C)P(C)}{P(BC)} = P(A|C)$$

- Events may be conditionally independent but not independent. (e.g., 'independent' multiple tests for confirming a disease)
- It is also possible that A, B are independent but are not conditionally independent given some other event C.

Use of conditional independence in Bayes rule

- We can write Bayes rule with multiple conditioning events.

$$P(A|BC) = \frac{P(BC|A)P(A)}{P(BC|A)P(A) + P(BC|A^c)P(A^c)}$$

- The above gets simplified if we assume $P(BC|A) = P(B|A)P(C|A)$, $P(BC|A^c) = P(B|A^c)P(C|A^c)$
- Consider the old example, where now we repeat the test for the disease.
- Take: $A = D$, $B = T_+^1$, $C = T_+^2$.
- Assuming conditional independence we can calculate the new posterior probability using the same information we had about true positive and false positive rate.

- Let us consider the example with $P(T_+|D) = 0.99$, $P(T_+|D^c) = 0.05$. $P(D) = 0.1$.
- Recall that we got $P(D|T_+) = 0.69$.
- Let us suppose the same test is repeated.

$$P(D \mid T_+^1 T_+^2) = \frac{P(T_+^1 T_+^2 \mid D)P(D)}{P(T_+^1 T_+^2 \mid D)P(D) + P(T_+^1 T_+^2 \mid D^c)P(D^c)}$$

$$= \frac{P(T_+^1 \mid D)P(T_+^2 \mid D)P(D)}{P(T_+^1 \mid D)P(T_+^2 \mid D)P(D) + P(T_+^1 \mid D^c)P(T_+^2 \mid D^c)P(D^c)}$$

$$= \frac{0.99 * 0.99 * 0.1}{0.99 * 0.99 * 0.1 + 0.05 * 0.05 * 0.9} = 0.97$$

Sequential Continuity of Probability

- For function, $f : \Re \to \Re$, it is continuous at x if and only if $x_n \to x$ implies $f(x_n) \to f(x)$.
- Thus, for continuous functions,

$$f\left(\lim_{n\to\infty} x_n\right) = \lim_{n\to\infty} f(x_n)$$

- We want to ask whether the probability, which is a function whose domain is $\mathcal{F}$, is also continuous like this.
- That is, we want to ask the question

$$P\left(\lim_{n\to\infty} A_n\right) \stackrel{?}{=} \lim_{n\to\infty} P(A_n)$$

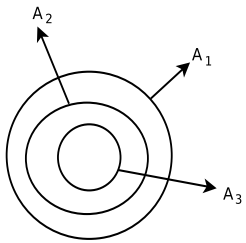
- For this, we need to first define limit of a sequence of sets.

- For now we define limits of only monotone sequences. (We will look at the general case later in the course)
- A sequence, $A_1, A_2, \cdots$, is said to be monotone decreasing if

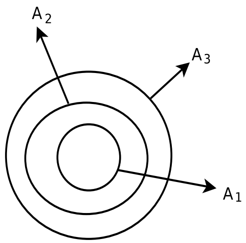
$$A_{n+1} \subset A_n, \ \forall n \quad (\text{denoted as } A_n \downarrow)$$

- A sequence, $A_1, A_2, \cdots$, is said to be monotone increasing if

$$A_n \subset A_{n+1}, \ \forall n \quad (\text{denoted as } A_n \uparrow)$$



$A_n \downarrow$



$A_n \uparrow$

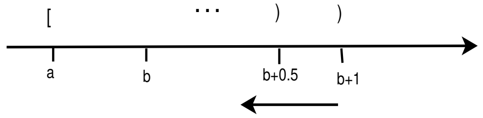
- Let $A_n \downarrow$. Then we define its limit as

$$\lim_{n\to\infty} A_n = \cap_{k=1}^{\infty} A_k$$

- This is reasonable because, when $A_n \downarrow$, we have $A_n \subset A_{n-1} \subset A_{n-2} \cdots$ and hence, $A_n = \cap_{k=1}^{n} A_k$.
- Similarly, when $A_n \uparrow$, we define the limit as

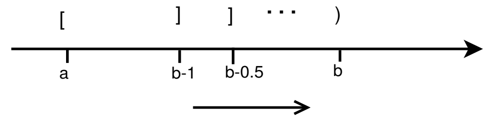
$$\lim_{n\to\infty} A_n = \cup_{k=1}^{\infty} A_k$$

- Let us look at simple examples of monotone sequences of subsets of $\Re$.
- Consider a sequence of intervals:
  $A_n = [a,\ b + \frac{1}{n}), n = 1, 2, \cdots$ with $a, b \in \Re,\ a < b$.



- We have $A_n \downarrow$ and $\lim A_n = \cap_i A_i = [a,\ b]$
- Why? – because
  - $b \in A_n, \forall n \Rightarrow b \in \cap_i A_i$, and
  - $\forall \epsilon > 0,\ b + \epsilon \notin A_n$ after some $n \Rightarrow b + \epsilon \notin \cap_i A_i$.
    For example, $b + 0.01 \notin A_{101} = [a,\ b + \frac{1}{101})$.

- We have shown that $\cap_n [a,\ b + \frac{1}{n}) = [a,\ b]$
- Similarly we can get $\cap_n (a - \frac{1}{n},\ b] = [a,\ b]$
- Now consider $A_n = [a,\ b - \frac{1}{n}]$.



- Now, $A_n \uparrow$ and $\lim A_n = \cup_n A_n = [a,\ b)$.
- Why? – because
  - $\forall \epsilon > 0, \exists n$ s.t. $b - \epsilon \in A_n \Rightarrow b - \epsilon \in \cup_n A_n$;
  - but $b \notin A_n, \forall n \Rightarrow b \notin \cup_n A_n$.
- These examples also show how using countable unions or intersections we can convert one end of an interval from 'open' to 'closed' or vice versa.

- To summarize, limits of monotone sequences of events are defined as follows

$$A_n \downarrow \quad \lim_{n \to \infty} A_n = \cap_{k=1}^{\infty} A_k$$

$$A_n \uparrow \quad \lim_{n \to \infty} A_n = \cup_{k=1}^{\infty} A_k$$

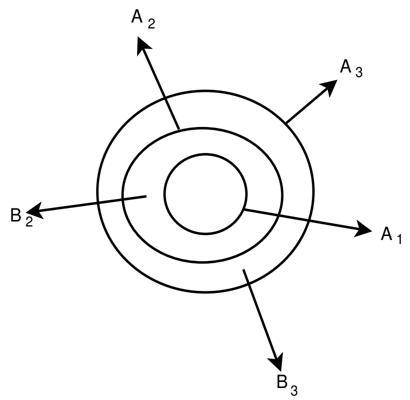
- Having defined the limits, we now ask the question

$$P\left(\lim_{n \to \infty} A_n\right) \stackrel{?}{=} \lim_{n \to \infty} P(A_n)$$

where we assume the sequence is monotone.

Theorem: Let $A_n \uparrow$. Then $P(\lim_n A_n) = \lim_n P(A_n)$
- Since $A_n \uparrow$, $A_n \subset A_{n+1}$.
- Define sets B_i, $i = 1, 2, \cdots$, by

$$B_1 = A_1,\ B_k = A_k - A_{k-1},\ k = 2, 3, \cdots$$



Theorem: Let $A_n \uparrow$. Then $P(\lim_n A_n) = \lim_n P(A_n)$

- Since $A_n \uparrow$, $A_n \subset A_{n+1}$.
- Define sets B_i, $i = 1, 2, \cdots$, by
$$B_1 = A_1, \;\; B_k = A_k - A_{k-1}, \;\; k = 2, 3, \cdots$$
- Note that B_k are mutually exclusive. Also note that
$$A_n = \cup_{k=1}^n B_k \;\; \text{and hence } P(A_n) = \sum_{k=1}^n P(B_k)$$
- We also have
$$\cup_{k=1}^n A_k = \cup_{k=1}^n B_k, \forall n \;\; \text{and hence } \cup_{k=1}^\infty A_k = \cup_{k=1}^\infty B_k$$
- Thus we get
$$\begin{aligned} P(\lim_n A_n) &= P(\cup_{k=1}^\infty A_k) = P(\cup_{k=1}^\infty B_k) \\ &= \sum_{k=1}^\infty P(B_k) = \lim_n \sum_{k=1}^n P(B_k) = \lim_n P(A_n) \end{aligned}$$

- We showed that when $A_n \uparrow$, $P(\lim_n A_n) = \lim_n P(A_n)$
- We can show this for the case $A_n \downarrow$ also.
- Note that if $A_n \downarrow$, then $A_n^c \uparrow$. Using this and the theorem we can show it. (Left as an exercise)
- This property is known as monotone sequential continuity of the probability measure.

- We can think of a simple example to use this theorem.
- We keep tossing a fair coin. (We take tosses to be independent). We want to show that never getting a head has probability zero.
- The basic idea is simple. ($(0.5)^n \to 0$)
- But to formalize this we need to specify what is our probability space and then specify what is the event (of never getting a head).
- If we toss the coin for any fixed N times then we know the sample space can be $\{0, 1\}^N$.
- But for our problem, we can not put any fixed limit on the number of tosses and hence our sample space should be for infinite tosses of a coin.

- We take Ω as set of all infinite sequences of 0's and 1's:
$$\Omega = \{(\omega_1, \omega_2, \cdots) \; : \; \omega_i \in \{0, 1\}, \; \forall i\}$$
- This would be uncountably infinite.
- We would not specify $\mathcal{F}$ fully. But assume that any subset of Ω specifiable through outcomes of finitely many coin tosses would be an event.
- Thus "no head in the first n tosses" would be an event.

- What P should we consider for this uncountable Ω? We are not sure what to take.
- So, let us ask only for some consistency. For any subset of this Ω that is specified only through outcomes of first n tosses, that event should have the same probability as in the finite probability space corresponding to n tosses.
- Consider an event here;

$$A = \{(\omega_1, \omega_2, ...) \; : \; \omega_1 = \omega_2 = 0\} \subset \Omega$$

  A is the event of tails on first two tosses.
- We are saying we must have $P(A) = (0.5)^2$.
- Now we can complete problem

- For $n = 1, 2, \cdots$, define

$$A_n = \{(\omega_1, \omega_2, ...) \; : \; \omega_i = 0, \; i = 1, \cdots, n\}$$

- A_n is the event of no head in the first n tosses and we know $P(A_n) = (0.5)^n$.
- Note that $\cap_{k=1}^{\infty} A_k$ is the event we want.
- Note that $A_n \downarrow$ because $A_{n+1} \subset A_n$.
- Hence we get

$$P(\cap_{k=1}^{\infty} A_k) = P(\lim_n A_n) = \lim_n P(A_n) = \lim_n (0.5)^n = 0$$

- The Ω we considered can be corresponded with the interval $[0, 1]$.
- Each element of Ω is an infinite sequence of 0's and 1's

$$\omega = (\omega_1, \omega_2, \cdots), \; \omega_i \in \{0, 1\} \; \forall i$$

- We can put a 'binary point' in front and thus consider ω to be a real number between 0 and 1.
- That is, we correspond ω with the real number: $\omega_1 2^{-1} + \omega_2 2^{-2} + \cdots$
- For example, the sequence $(0, 1, 0, 1, 0, 0, 0, \cdots)$ would be the number: $2^{-2} + 2^{-4} = 5/16$.
- Essentially, every number in $[0, 1]$ can be represented by a binary sequence like this and every binary sequence corresponds to a real number between 0 and 1.
- Thus, our Ω can be thought of as interval $[0, 1]$.
- So, uncountable Ω arise naturally if we want to consider infinite repetitions of a random experiment

- The P we considered would be such that probability of an interval would be its length.
- Consider the example event we considered earlier

$$A = \{(\omega_1, \omega_2, ...) \; : \; \omega_1 = \omega_2 = 0\} \subset \Omega$$

- When we view the Ω as the interval $[0, 1]$, the above is the set of all binary numbers of the form $0.00xxxxxx\cdots$.
- What is this set of numbers?
- It ranges from $0.000000\cdots$ to $0.0011111\cdots$.
- That is the interval $[0, 0.25]$.
- As we already saw, the probability of this event is $(0.5)^2$ which is the length of this interval

- We looked at this probability space only for an example where we could use monotone sequential continuity of probability.
- But this probability space is important and has lot of interesting properties.

$$\Omega = \{(\omega_1, \omega_2, \cdots) \ : \ \omega_i \in \{0,1\}, \forall i\}$$

- Here, $\frac{1}{n}\sum_{i=1}^{n}\omega_i$ the fraction of heads in the first n tosses.
- Since we are tossing a fair coin repeatedly, we should expect

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\omega_i = \frac{1}{2}$$

- We expect this to be true for 'almost all' sequences in Ω.
- That means 'almost all' numbers in $[0,1]$ when expanded as infinite binary fractions, satisfy this property.
- This is called Borel's normal number theorem and is an interesting result about real numbers.

Probability Models

- As mentioned earlier, everything in probability theory is with reference to an underlying probability space: $(\Omega, \mathcal{F}, P)$.
- Probability theory starts with $(\Omega, \mathcal{F}, P)$
- We can say that different P correspond to different models.
- Theory does not tell you how to get the P.
- The modeller has to decide what P she wants.
- The theory allows one to derive consequences or properties of the model.

- Consider the random experiment of tossing a fair coin three times.
- We can take $\Omega = \{0,1\}^3$ and can use the following P_1.

ω	$P_1(\{\omega\})$
0 0 0	1/8
0 0 1	1/8
0 1 0	1/8
0 1 1	1/8
1 0 0	1/8
1 0 1	1/8
1 1 0	1/8
1 1 1	1/8

- Now probability theory can derive many consequences:
  - The tosses are independent
  - Probability of 0 or 3 heads is 1/8 while that of 1 or 2 heads is 3/8

- Now consider a P_2 (different from P_1) on the same Ω

ω	$P_2(\{\omega\})$
0 0 0	1/4
0 0 1	1/12
0 1 0	1/12
0 1 1	1/12
1 0 0	1/12
1 0 1	1/12
1 1 0	1/12
1 1 1	1/4

- The consequences now change
  - The probability that number of heads is 0 or 1 or 2 or 3 are all same and all equal 1/4.
  - The tosses are not independent

- We can not ask which is the 'correct' probability model here.
- Such a question is meaningless as far as probability theory is concerned.
- One chooses a model based on application.
- If we think tosses are independent then we choose P_1. But if we need to model some dependence among tosses, we choose a model like P_2.

- The model P_2 accommodates some dependence among tosses.
- Outcomes of previous tosses affect the current toss.

ω	$P_2(\{\omega\})$
0 0 0	$1/4$ $(= (1/2)(2/3)(3/4))$
0 0 1	$1/12$ $(= (1/2)(2/3)(1/4))$
0 1 0	$1/12$ $(= (1/2)(1/3)(2/4))$
0 1 1	$1/12$
1 0 0	$1/12$
1 0 1	$1/12$
1 1 0	$1/12$
1 1 1	$1/4$

- It is also a useful model.

We next consider the concept of random variables. These allow one to specify and analyze different probability models.

This entire course can be considered as studying different random variables.

Random Variable

- A random variable is a real-valued function on Ω: $X : \Omega \to \Re$
- For example, $\Omega = \{H, T\}$, $X(H) = 1, \ X(T) = 0$.
- Another example: $\Omega = \{H, T\}^3$, $X(\omega)$ is numbers of H's.
- A random variable maps each outcome to a real number.
- It essentially means we can treat all outcomes as real numbers.
- We can effectively work with $\Re$ as sample space in all probability models

Recap: Monotone Sequences of Sets

- A sequence, $A_1, A_2, \cdots$, is said to be monotone decreasing if

$$A_{n+1} \subset A_n, \ \forall n \quad \text{(denoted as } A_n \downarrow)$$

- Limit of a monotone decreasing sequence is

$$A_n \downarrow: \quad \lim_{n\to\infty} A_n = \cap_{k=1}^{\infty} A_k$$

- A sequence, $A_1, A_2, \cdots$, is said to be monotone increasing if

$$A_n \subset A_{n+1}, \ \forall n \quad \text{(denoted as } A_n \uparrow)$$

- Limit of monotone increasing sequence is

$$A_n \uparrow: \quad \lim_{n\to\infty} A_n = \cup_{k=1}^{\infty} A_k$$

Recap: Monotone Sequential Continuity

- We showed that

$$P\left(\lim_{n\to\infty} A_n\right) = \lim_{n\to\infty} P(A_n)$$

when $A_n \downarrow$ or $A_n \uparrow$

Random Variable

- A random variable is a real-valued function on Ω: $X : \Omega \to \Re$
- For example, $\Omega = \{H, T\}$, $X(H) = 1, \ X(T) = 0$.
- Another example: $\Omega = \{H, T\}^3$, $X(\omega)$ is numbers of H's.
- A random variable maps each outcome to a real number.
- It essentially means we can treat all outcomes as real numbers.
- We can effectively work with $\Re$ as sample space in all probability models

- Let $(\Omega, \mathcal{F}, P)$ be our probability space and let X be a random variable defined in this probability space.
- We know X maps Ω into $\Re$.
- This random variable results in a new probability space:

$$(\Omega, \mathcal{F}, P) \ \stackrel{X}{\rightarrow} \ (\Re, \mathcal{B}, P_X)$$

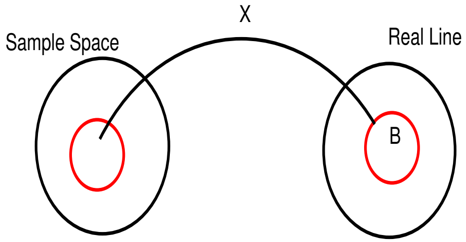
where $\Re$ is the new sample space and $\mathcal{B} \subset 2^{\Re}$ is the new set of events and P_X is a probability defined on $\mathcal{B}$.
- For now we will assume that any set of $\Re$ that we want would be in $\mathcal{B}$ and hence is an event.
- P_X is a new probability measure (which depends on P and X) that assigns probability to different subsets of $\Re$.

- Given a probability space $(\Omega, \mathcal{F}, P)$, a random variable X

$$(\Omega, \mathcal{F}, P) \xrightarrow{X} (\Re, \mathcal{B}, P_X)$$

- We define P_X:

$$P_X(B) = P\left(\{\omega \in \Omega \ : \ X(\omega) \in B\}\right), \ B \in \mathcal{B}$$



- Given a probability space $(\Omega, \mathcal{F}, P)$, a random variable X

$$(\Omega, \mathcal{F}, P) \xrightarrow{X} (\Re, \mathcal{B}, P_X)$$

- We define P_X:

$$P_X(B) = P\left(\{\omega \in \Omega \ : \ X(\omega) \in B\}\right), \ B \in \mathcal{B}$$

- We use the notation

$$[X \in B] = \{\omega \in \Omega \ : \ X(\omega) \in B\}$$

- So, now we can write

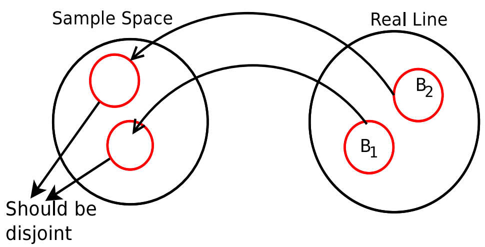
$$P_X(B) = P([X \in B]) = P[X \in B]$$

- For the definition of P_X to be proper, for each $B \in \mathcal{B}$, we must have $[X \in B] \in \mathcal{F}$.
  We will assume that. (This is trivially true if $\mathcal{F} = 2^\Omega$).
- We can easily verify P_X is a probability measure. It satisfies the axioms.

- Given a probability space $(\Omega, \mathcal{F}, P)$, a random variable X
- We define P_X:

$$P_X(B) = P[X \in B] = P\left(\{\omega \in \Omega \ : \ X(\omega) \in B\}\right)$$

- Easy to see: $P_X(B) \geq 0, \ \forall B$ and $P_X(\Re) = 1$
- If $B_1 \cap B_2 = \phi$ then $P_X(B_1 \cup B_2) = P[X \in B_1 \cup B_2] = ?$



$$P[X \in B_1 \cup B_2] = P[X \in B_1] + P[X \in B_2] = P_X(B_1) + P_X(B_2)$$

- Let us look at a couple of simple examples.
- Let $\Omega = \{H, T\}$ and $P(H) = p$.
  Let $X(H) = 1; X(T) = 0$.

$$\begin{aligned}
[X \in \{0\}] &= \{\omega \ : \ X(\omega) = 0\} = \{T\} \\
[X \in [-3.14, 0.552]\,] &= \{\omega : -3.14 \leq X(\omega) \leq 0.552\} = \{T \\
[X \in (0.62, 15.5)] &= \{\omega : 0.62 < X(\omega) < 15.5\} = \{H\} \\
[X \in [-2, 2)\,] &= \Omega
\end{aligned}$$

- Hence we get

$$P_X(\{0\}) = (1-p) = P_X([-3.14, 0.552])$$

$$P_X((0.6237, 15.5)) = p; \ \ P_X([-2, 2)) = 1$$

- Let $\Omega = \{H,T\}^3 = \{HHH, HHT, \cdots, TTT\}$.
  Let P be specified through 'equally likely' assignment.
  Let $X(\omega)$ be number of H's in ω. Thus, $X(THT) = 1$.
  (X takes one of the values: 0, 1, 2, or 3)
- We can once again write down $[X \in B]$ for different $B \subset \Re$

$$[X \in (0,1]\,] = \{HTT, THT, TTH\};$$

$$[X \in (-1.2, 2.78)\,] = \Omega - \{HHH\}$$

- Hence

$$P_X((0,1]) = \frac{3}{8};\ P_X((-1.2, 2.78)) = \frac{7}{8}$$

- A random variable defined on $(\Omega, \mathcal{F}, P)$ results in a new or induced probability space $(\Re, \mathcal{B}, P_X)$.
- The Ω may be countable or uncountable (even though we looked at only examples of finite Ω).
- Thus, we can study probability models by taking $\Re$ as sample space through the use of random variables.
- However there are some technical issues regarding what $\mathcal{B}$ we should consider.
- We briefly consider this and then move on to studying random variables.

- We want to look at the probability space $(\Re, \mathcal{B}, P_X)$.
- If we could take $\mathcal{B} = 2^{\Re}$ then everything would be simple. But that is not feasible.
- What this means is that if we want every subset of real line to be an event, we cannot construct a probability measure (to satisfy the axioms).

- Let us consider $\Omega = [0,1]$.
- This is the simplest example of uncountable Ω we considered.
- We also saw that this sample space comes up when we consider infinite tosses of a coin.
- The simplest extension of the idea of 'equally likely' is to say probability of an event (subset of Ω) is the length of the event (subset).
- But not all subsets of $[0,1]$ are intervals and length is defined only for intervals.
- We can define length of countable union of disjoint intervals to be sum of the lengths of individual intervals.
- But what about subsets that may not be countable unions of disjoint intervals ?
- Well, we say those can be assigned probability by using the axioms.

- Thus the question is the following:
- Can we construct a function $m : 2^{[0,1]} \to [0,1]$ such that
  1. $m(A) = \text{length}(A)$ if $A \subset [0,1]$ is an interval
  2. $m(\cup_i A_i) = \sum_i m(A_i)$ where $A_i \cap A_j = \phi$ whenever $i \neq j$, $(A_1, A_2, \cdots \subset [0,1])$
- The surprising answer is 'NO'
- This is a fundamental result in real analysis.
- Hence for the probability space $(\Re, \mathcal{B}, P_X)$ we cannot take $\mathcal{B} = 2^{\Re}$.
  (Recall that for countable Ω we can take $\mathcal{F} = 2^{\Omega}$).
- Now the question is what is the best $\mathcal{B}$ we can have?

σ-algebra

- An $\mathcal{F} \subset 2^{\Omega}$ is called a σ-algebra (also called σ-field) on Ω if it satisfies the following:
  1. $\Omega \in \mathcal{F}$
  2. $A \in \mathcal{F} \;\Rightarrow\; A^c \in \mathcal{F}$
  3. $A_1, A_2, \cdots \in \mathcal{F} \;\Rightarrow\; \cup_i A_i \in \mathcal{F}$
- Thus a σ-algebra is a collection of subsets of Ω that is closed under complements and countable unions (and hence countable intersections because $\cap_i A_i = (\cup_i A_i^c)^c$).
- Note that 2^{Ω} is obviously a σ-algebra
- In a Probability space $(\Omega, \mathcal{F}, P)$, if $\mathcal{F} \neq 2^{\Omega}$ then we want it to be a σ-algebra. (Why?)

- Easy to construct examples of σ-algebras
  Let $A \subset \Omega$.

$$\mathcal{F} = \{\Omega, \phi, A, A^c\} \text{ is a } \sigma\text{-algebra}$$

- For example, with $\Omega = \{1,2,3,4,5,6\}$,

$$\mathcal{F} = \{\Omega, \phi, \{1,3,5\}, \{2,4,6\}\} \text{ is a } \sigma\text{-algebra}$$

- Suppose on this Ω we want to make a σ-algebra containing $\{1,2\}$ and $\{3,4\}$.

$$\{\Omega, \phi, \{1,2\}, \{3,4\}, \{3,4,5,6\}, \{1,2,5,6\}, \{1,2,3,4\}, \{5,6\}\}$$

- This is the 'smallest' σ-algebra containing $\{1,2\}$, $\{3,4\}$

- Let $\mathcal{F}_1, \mathcal{F}_2$ be σ-algebras on Ω.
- Then, so is $\mathcal{F}_1 \cap \mathcal{F}_2$.
- It is simple to show.
  (E.g., $A \in \mathcal{F}_1 \cap \mathcal{F}_2 \;\Rightarrow\; A \in \mathcal{F}_1, A \in \mathcal{F}_2 \;\Rightarrow\; A^c \in \mathcal{F}_1, A^c \in \mathcal{F}_2 \;\Rightarrow\; A^c \in \mathcal{F}_1 \cap \mathcal{F}_2$)
- Let $G \subset 2^{\Omega}$. We denote by $\sigma(G)$ the smallest σ-algebra containing G.
- It is defined as the intersection of all σ-algebras containing G (and hence is well defined).

- Let us get back to the question we started with.
- In the probability space $(\Re, \mathcal{B}, P)$ what is the $\mathcal{B}$ we should choose.
- We can choose it to be the smallest σ-algebra containing all intervals
- That is called Borel σ-algebra, $\mathcal{B}$.
- It contains all intervals, all complements, countable unions and intersections of intervals and all sets that can be obtained through complements, countable unions and/or intersections of such sets and so on.

Borel σ-algebra

- Let $\quad G = \{(-\infty,\ x]\ :\ x \in \Re\}$
- We can define the Borel σ-algebra, $\mathcal{B}$, as the smallest σ-algebra containing G.
- We can see that $\mathcal{B}$ would contain all intervals.
  1. $(-\infty,\ x) \in \mathcal{B}$ because $(-\infty,\ x) = \cup_n(-\infty,\ x - \frac{1}{n}]$
  2. $(x,\ \infty) \in \mathcal{B}$ because $(x,\ \infty) = (-\infty,\ x]^c$
  3. $[x,\ \infty) \in \mathcal{B}$ because $[x,\ \infty) = \cap_n(x - \frac{1}{n},\ \infty)$
  4. $(x,\ y] \in \mathcal{B}$ because $(x,\ y] = (-\infty,\ y] \cap (x,\ \infty)$
  5. $[x,\ y] \in \mathcal{B}$ because $[x,\ y] = \cap_n(x - \frac{1}{n},\ y]$
  6. $[x,\ y), (x,\ y) \in \mathcal{B}$, similarly
- Thus, $\sigma(G)$ is also the smallest σ-algebra containing all intervals.

Borel σ-algebra

- We have defined $\mathcal{B}$ as

$$\mathcal{B} = \sigma\left(\{(-\infty,\ x]\ :\ x \in \Re\}\right)$$

- It is also the smallest σ-algebra containing all intervals.
- Elements of $\mathcal{B}$ are called Borel sets
- Intervals (including singleton sets), complements of intervals, countable unions and intersections of intervals, countable unions and intersections of such sets on so on are all Borel sets.
- Borel σ-algebra contains enough sets for our purposes.
- Are there any subsets of real line that are not Borel?
- YES!! Infinitely many non-Borel sets would be there!

Random Variables

- Given a probability space $(\Omega, \mathcal{F}, P)$, a random variable is a real-valued function on Ω.
- It essentially results in an induced probability space

$$(\Omega, \mathcal{F}, P) \ \stackrel{X}{\rightarrow}\ (\Re, \mathcal{B}, P_X)$$

  where $\mathcal{B}$ is the Borel σ-algebra.
- We define P_X as: for all Borel sets, $B \subset \Re$,

$$P_X(B) = P[X \in B] = P\left(\{\omega \in \Omega\ :\ X(\omega) \in B\}\right)$$

- For X to be a random variable, the folowing should also hold

$$[X \in B] = \{\omega \in \Omega\ :\ X(\omega) \in B\} \in \mathcal{F},\ \forall B \in \mathcal{B}$$

- We always assume this.

- Let X be a random variable.
- It represents a probability model with $\Re$ as the sample space.
- The probability assigned to different events (Borel subsets of $\Re$) is

$$P_X(B) = P[X \in B] = P\left(\{\omega \in \Omega \ : \ X(\omega) \in B\}\right)$$

- How does one represent this probability measure

Distribution function of a random variable

- Let X be a random variable. It distribution function is $F_X : \Re \to \Re$ defined by

$$F_X(x) = P[X \in (-\infty, x]] = P(\{\omega \in \Omega \ : \ X(\omega) \leq x\})$$

- We write the event $\{\omega \ : \ X(\omega) \leq x\}$ as $[X \leq x]$.
  We follow this notation with any such relation statement involving X
  e.g., $[X \neq 3]$ represents the event $\{\omega \in \Omega \ : \ X(\omega) \neq 3\}$.
- Thus we have

$$F_X(x) = P[X \leq x] = P(\{\omega \in \Omega \ : \ X(\omega) \leq x\}) = P_X(\ (-\infty,\ x]\ )$$

- The distribution function, F_X completely specifies the probability measure, P_X.

- The distribution function of X is given by

$$F_X(x) = P[X \leq x] = P(\{\omega \in \Omega \ : \ X(\omega) \leq x\})$$

- This is also sometimes called the cumulative distribution function.
- F_X is a real-valued function of a real variable.
- Let us look at a simple example.

- Consider tossing of a fair coin: $\Omega = \{T, H\}$, $P(\{T\}) = P(\{H\}) = 0.5$.
- Let $X(T) = 0$ and $X(H) = 1$. We want to calculate F_X
- For this we want the event $[X \leq x]$, for different x
- Let us first look at some examples:

$$\begin{aligned} [X \leq -0.5] &= \{\omega \ : \ X(\omega) \leq -0.5\} = \phi \\ [X \leq 0.25] &= \{\omega \ : \ X(\omega) \leq 0.25\} = \{T\} \\ [X \leq 1.3] &= \{\omega \ : \ X(\omega) \leq 1.3\} = \Omega \end{aligned}$$

- Thus we get

$$\begin{aligned} [X \leq x] &= \{\omega \ : \ X(\omega) \leq x\} \\ &= \begin{cases} \phi & \text{if } \ x < 0 \\ \Omega & \text{if } \ x \geq 1 \\ \{T\} & \text{if } \ 0 \leq x < 1 \end{cases} \end{aligned}$$

- We are considering: $\Omega = \{T, H\}$, $P(\{T\}) = P(\{H\}) = 0.5$.
- $X(T) = 0$ and $X(H) = 1$. We want to calculate F_X
- We showed

$$
\begin{aligned}
[X \leq x] &= \{\omega \ : \ X(\omega) \leq x\} \\
&= \begin{cases} \phi & \text{if } \ x < 0 \\ \{T\} & \text{if } \ 0 \leq x < 1 \\ \Omega & \text{if } \ x \geq 1 \end{cases}
\end{aligned}
$$

- Hence $F_X(x) = P[X \leq x]$ is given by

$$
F_X(x) = \begin{cases} 0 & \text{if } \ x < 0 \\ 0.5 & \text{if } \ 0 \leq x < 1 \\ 1 & \text{if } \ x \geq 1 \end{cases}
$$

Please note that x is a 'dummy variable'

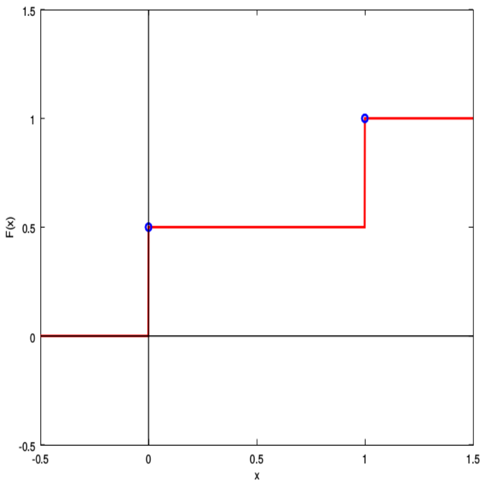
- We are considering: $\Omega = \{T, H\}$, $P(\{T\}) = P(\{H\}) = 0.5$.
- $X(T) = 0$ and $X(H) = 1$. We want to calculate F_X
- We showed

$$
\begin{aligned}
[X \leq x] &= \{\omega \ : \ X(\omega) \leq x\} \\
&= \begin{cases} \phi & \text{if } \ x < 0 \\ \{T\} & \text{if } \ 0 \leq x < 1 \\ \Omega & \text{if } \ x \geq 1 \end{cases}
\end{aligned}
$$

- Hence $F_X(y) = P[X \leq y]$ is given by

$$
F_X(y) = \begin{cases} 0 & \text{if } \ y < 0 \\ 0.5 & \text{if } \ 0 \leq y < 1 \\ 1 & \text{if } \ y \geq 1 \end{cases}
$$

- A plot of this distribution function:



- Let us look at another example.
- Let $\Omega = [0,\ 1]$ and take events to be Borel subsets of $[0,\ 1]$. (That is, $\mathcal{F} = \{B \cap [0,\ 1] \ : \ B \in \mathcal{B}\}$).
- We take P to be such that probability of an interval is its length.
- This is the 'usual' probability space whenever we take $\Omega = [0,\ 1]$.
- Let $X(\omega) = \omega$.
- We want to find the distribution function of X.

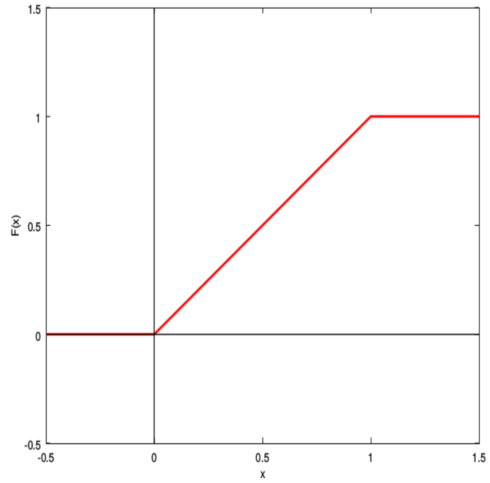
- Once again we need to find the event $[X \leq x]$ for different values of x.
- Note that the function X takes values in $[0,\ 1]$ and $X(\omega) = \omega$.

$$[X \leq x] = \{\omega \in \Omega\ :\ X(\omega) \leq x\} = \{\omega \in [0,\ 1]\ :\ \omega \leq x\}$$
$$= \begin{cases} \phi & \text{if } x < 0 \\ \Omega & \text{if } x \geq 1 \\ [0, x] & \text{if } 0 \leq x < 1 \end{cases}$$

- Hence $F_X(x) = P[X \leq x]$ is given by

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

- The plot of this distribution function:



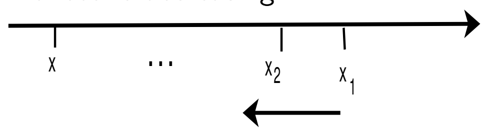
Properties of Distribution Functions

- The distribution function of random variable X is given by

$$F_X(x) = P[X \leq x] = P(\{\omega\ :\ X(\omega) \leq x\})$$

- Any distribution function should satisfy the following:
  1. $0 \leq F_X(x) \leq 1,\ \forall x$
  2. $F_X(-\infty) = 0;\ F_X(\infty) = 1$
  3. F_X is non-decreasing: $x_1 \leq x_2\ \Rightarrow\ F_X(x_1) \leq F_X(x_2)$
     $x_1 \leq x_2\ \Rightarrow\ (-\infty,\ x_1] \subset (-\infty,\ x_2]\ \Rightarrow$
     $P_X((-\infty,\ x_1]) \leq P_X((-\infty,\ x_2]\ \Rightarrow\ F_X(x_1) \leq F_X(x_2)$
  4. F_X is right continuous and has left-hand limits.

- Right continuity of F_X: $\ x_n \downarrow x\ \Rightarrow F_X(x_n) \rightarrow F_X(x)$
  - $x_n \downarrow x$ implies the sequence of events $(-\infty,\ x_n]$ is monotone decreasing.



  - Also, $\lim_n(-\infty,\ x_n] = \cap_n(-\infty,\ x_n] = (-\infty,\ x]$
  - This implies

$$\lim_n P_X((-\infty,\ x_n]) = P_X(\lim_n(-\infty,\ x_n]) = P_X((-\infty,\ x])$$

  - This in turn implies

$$\lim_{x_n \downarrow x} F_X(x_n) = F_X(x)$$

  - Using the usual notation for right limit of a function, we can write $F_X(x^+) = F_X(x), \forall x$.

- F_X is right-continuous at all x.
- Next, let us look at the lefthand limits: $\lim_{x_n \uparrow x} F_X(x_n)$
- When $x_n \uparrow x$, the sequence of events $(-\infty,\ x_n]$ is monotone increasing and

$$\lim_n (-\infty,\ x_n] = \cup_n (-\infty,\ x_n] = (-\infty,\ x)$$

- By sequential continuity of probability, we have

$$\lim_n P_X((-\infty,\ x_n]) = P_X(\lim_n (-\infty,\ x_n]\ ) = P_X(\ (-\infty,\ x_n)\ )$$

- Hence we get

$$F_X(x^-) \quad = \quad \lim_{x_n \uparrow x} F_X(x_n) = \lim_n P_X((-\infty,\ x_n]) = P_X(\ (-\infty,\ x_n)\ )$$

- Thus, at every x the left limit of F_X exists.

- F_X is right-continuous:
  $F_X(x^+) = F_X(x) = P_X(\ (-\infty,\ x]\ )$
- It has left limits: $F_X(x^-) = P_X(\ (-\infty,\ x)\ )$
- If $A \subset B$ then $P(B - A) = P(B) - P(A)$
- We have $(-\infty,\ x] - (-\infty,\ x) = \{x\}$. Hence

$$P_X(\ (-\infty,\ x]\ ) - P_X(\ (-\infty,\ x)\ ) = P_X(\{x\}) = P(\{\omega\ :\ X(\omega) = x\})$$

- Thus we get

$$F_X(x^+) - F_X(x^-) = P[X = x] = P(\{\omega\ :\ X(\omega) = x\})$$

- When F_X is discontinuous at x the height of discontinuity is the probability that X takes that value.
- And, if F_X is continuous at x then $P[X = x] = 0$

Distribution Functions

- Let X be a random variable.
- Its distribution function, $F_X : \Re \to \Re$ is given by $F_X(x) = P[X \le x]$
- The distribution function satisfies
  1. $0 \le F_X(x) \le 1,\ \forall x$
  2. $F_X(-\infty) = 0;\ F_X(\infty) = 1$
  3. F_X is non-decreasing: $x_1 \le x_2 \ \Rightarrow\ F_X(x_1) \le F_X(x_2)$
  4. F_X is right continuous and has left-hand limits.
- We also have $F_X(x^+) - F_X(x^-) = P[X = x]$
- Any real-valued function of a real variable satisfying the above four properties would be a distribution function of some random variable.

- $F_X(x) = P[X \le x] = P[X \in (-\infty,\ x]\ ]$
- Given F_X, we can, in principle, find $P[X \in B]$ for all Borel sets.
- In particular, for $a < b$,

$$\begin{aligned} P[a < X \le b] &= P[X \in (a,\ b]\ ] \\ &= P[X \in (\ (-\infty,\ b] - (-\infty,\ a]\ )\ ] \\ &= P[X \in (-\infty,\ b]\ ] - P[X \in (-\infty,\ a]\ ] \\ &= F_X(b) - F_X(a) \end{aligned}$$

- There are two classes of random variables that we would study here.
- These are called discrete and continuous random variables.
- There can be random variables that are neither discrete nor continuous.
- But these two are important classes of random variables that we deal with in this course.
- Note that the distribution function is defined for **all** random variables.

Discrete Random Variables

- A random variable X is said to be discrete if it takes only countably many distinct values.
- Countably many means finite or countably infinite.
- If $X : \Omega \to \Re$ is discrete, its (strict) range is countable
- Any random variable that is defined on finite or countable Ω would be discrete.
- Thus the family of discrete random variables includes all probability models on finite or countably infinite sample spaces.

Discrete Random Variable Example

- Consider three independent tosses of a fair coin.
- $\Omega = \{H,\ T\}^3$ and $X(\omega)$ is the number of H's in ω.
- This rv takes four distinct values, namely, $0, 1, 2, 3$.
- We denote this as $X \in \{0, 1, 2, 3\}$
- Let us find the distribution function of this rv
- Let us take some examples of $[X \leq x]$

$$[X \leq 0.72] \;=\; \{\omega \;:\; X(\omega) \leq 0.72\} = \{\omega \;:\; X(\omega) = 0\} = [X = 0]$$

$$\begin{aligned} [X \leq 1.57] &= \{\omega \;:\; X(\omega) \leq 1.57\} \\ &= \{\omega \;:\; X(\omega) = 0\} \cup \{\omega \;:\; X(\omega) = 1\} = [X = 0 \text{ or } 1] \end{aligned}$$

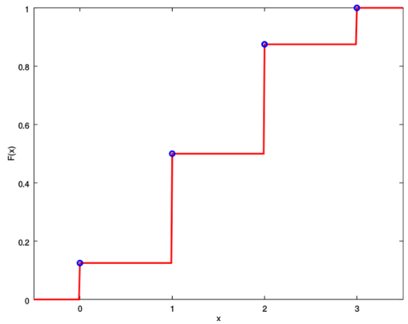
- $F_X(x) = P[X \leq x]$ (Recall $X \in \{0, 1, 2, 3\}$)
- The event $[X \leq x]$ for different x can be seen to be

$$[X \leq x] = \begin{cases} \phi & x < 0 \\ \{TTT\} & 0 \leq x < 1 \\ \{TTT, HTT, THT, TTH\} & 1 \leq x < 2 \\ \Omega - \{HHH\} & 2 \leq x < 3 \\ \Omega & x \geq 3 \end{cases}$$

- So, we get the distribution function as

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{8} & 0 \leq x < 1 \\ \frac{4}{8} & 1 \leq x < 2 \\ \frac{7}{8} & 2 \leq x < 3 \\ 1 & x \geq 3 \end{cases}$$

- The plot of this distribution function is:



- This is a stair-case function.
- It has jumps at $x = 0, 1, 2, 3$, which are the values that X takes. In between these it is constant.
- The jump at, e.g., $x = 2$ is $3/8$ which is the probability of X taking that value.

Recap: Random Variables

- Given a probability space $(\Omega, \mathcal{F}, P)$, a random variable is a real-valued function on Ω.
- It essentially results in an induced probability space

$$(\Omega, \mathcal{F}, P) \xrightarrow{X} (\Re, \mathcal{B}, P_X)$$

where $\mathcal{B}$ is the Borel σ-algebra and

$$P_X(B) = P[X \in B] = P\left(\{\omega \in \Omega \ : \ X(\omega) \in B\}\right)$$

Recap: σ-algebra

- An $\mathcal{F} \subset 2^{\Omega}$ is called a σ-algebra (also called σ-field) on Ω if it satisfies
  1. $\Omega \in \mathcal{F}$
  2. $A \in \mathcal{F} \ \Rightarrow \ A^c \in \mathcal{F}$
  3. $A_1, A_2, \cdots \in \mathcal{F} \ \Rightarrow \ \cup_i A_i \in \mathcal{F}$
- Thus a σ-algebra is a collection of subsets of Ω that is closed under complements and countable unions (and hence countable intersections).
- The Borel σ-algebra (on $\Re$), $\mathcal{B}$, is the smallest σ-algebra containing all intervals.
- We also have $\mathcal{B} = \sigma(\{(-\infty, \ x] : x \in \Re\}$

Recap: Distribution function of a random variable

- Let X be a random variable. It distribution function, $F_X : \Re \to \Re$, is defined by

$$F_X(x) = P[X \leq x] = P(\{\omega \in \Omega \ : \ X(\omega) \leq x\})$$

- The distribution function, F_X, completely specifies the probability measure, P_X.

Recap: Properties of distribution function

- The distribution function satisfies
  1. $0 \leq F_X(x) \leq 1, \ \forall x$
  2. $F_X(-\infty) = 0; \ F_X(\infty) = 1$
  3. F_X is non-decreasing: $x_1 \leq x_2 \ \Rightarrow \ F_X(x_1) \leq F_X(x_2)$
  4. F_X is right continuous and has left-hand limits.
- Any real-valued function of a real variable satisfying the above four properties would be a distribution function of some random variable.
- We also have
  $F_X(x^+) - F_X(x^-) = F_X(x) - F_X(x^-) = P[X = x]$
  $P[a < X \leq b] = F_X(b) - F_X(a)$.

- There are two classes of random variables that we would study here.
- These are called discrete and continuous random variables.
- Note that the distribution function is defined for **all** random variables.

Discrete Random Variables

- A random variable X is said to be discrete if it takes only countably many distinct values.
- Countably many means finite or countably infinite.

Discrete Random Variable Example

- Consider three independent tosses of a fair coin.
- $\Omega = \{H, \ T\}^3$ and $X(\omega)$ is the number of H's in ω.
- This rv takes four distinct values, namely, $0, 1, 2, 3$.
- We denote this as $X \in \{0, 1, 2, 3\}$
- Let us find the distribution function of this rv
- Let us take some examples of $[X \leq x]$

$$[X \leq 0.72] = \{\omega \ : \ X(\omega) \leq 0.72\} = \{\omega \ : \ X(\omega) = 0\} = [X = 0]$$

$$\begin{aligned} [X \leq 1.57] &= \{\omega \ : \ X(\omega) \leq 1.57\} \\ &= \{\omega \ : \ X(\omega) = 0\} \cup \{\omega \ : \ X(\omega) = 1\} = [X = 0 \text{ or } 1] \end{aligned}$$

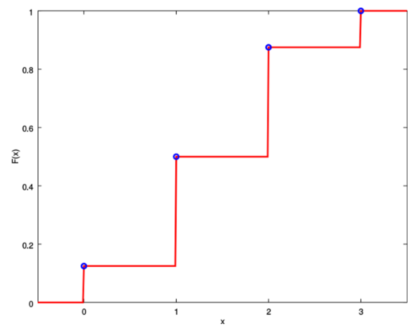
- $F_X(x) = P[X \leq x]$ (Recall $X \in \{0, 1, 2, 3\}$)
- The event $[X \leq x]$ for different x can be seen to be

$$[X \leq x] = \begin{cases} \phi & x < 0 \\ \{TTT\} & 0 \leq x < 1 \\ \{TTT, HTT, THT, TTH\} & 1 \leq x < 2 \\ \Omega - \{HHH\} & 2 \leq x < 3 \\ \Omega & x \geq 3 \end{cases}$$

- So, we get the distribution function as

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{8} & 0 \leq x < 1 \\ \frac{4}{8} & 1 \leq x < 2 \\ \frac{7}{8} & 2 \leq x < 3 \\ 1 & x \geq 3 \end{cases}$$

- The plot of this distribution function is:



- This is a stair-case function.
- It has jumps at $x = 0, 1, 2, 3$, which are the values that X takes. In between these it is constant.
- The jump at, e.g., $x = 2$ is $3/8$ which is the probability of X taking that value.

- We know that $F_X(x) - F_X(x^-) = P[X = x]$.
- For example,

$$\begin{aligned} F_X(2) - F_X(2^-) &= P[X = 2] = P(\{\omega : X(\omega) = 2\}) \\ &= P(\{THH, HTH, HHT\}) = \frac{3}{8} \end{aligned}$$

- The F_X is a stair-case function.
- It has jumps at each value assumed by X (and is constant in between)
- The height of the jump is equal to the probability of X taking that value.
- All discrete random variables would have this general form of distribution function.

- Let X be a dicrete rv and let $X \in \{a_1, a_2, \cdots, a_n\}$
- As a notation we assume: $a_1 < a_2 < \cdots < a_n$
- Let $[X = a_i] = \{\omega : X(\omega) = a_i\} = B_i$ and let $P(B_i) = q_i$.
- Since X is a function on Ω, $B_1, \cdots, B_n$ form a partition of Ω.
- Note that $q_i \geq 0$ and $\sum_{i=1}^{n} q_i = 1$.
- If $x < a_1$ then $[X \leq x] = \phi$.
- If $a_1 \leq x < a_2$ then $[X \leq x] = [X = a_1] = B_1$
- If $a_2 \leq x < a_3$ then $[X \leq x] = [X = a_1] \cup [X = a_2] = B_1 + B_2$

- Hence we can write the distribution function as

$$F_X(x) = \begin{cases} 0 & x < a_1 \\ P(B_1) & a_1 \le x < a_2 \\ P(B_1) + P(B_2) & a_2 \le x < a_3 \\ & \vdots \\ \sum_{i=1}^{k} P(B_i) & a_k \le x < a_{k+1} \\ & \vdots \\ 1 & x \ge a_n \end{cases}$$

- We can write this compactly as

$$F_X(x) = \sum_{k: a_k \le x} q_k$$

- Note that all this holds even when X takes countably infinitely many values.

- Let X be a discrete rv with $X \in \{x_1, x_2, \cdots\}$.
- Let $q_i = P[X = x_i] \quad (= P(\{\omega \ : \ X(\omega) = x_i\})\ )$
- We have $q_i \ge 0$ and $\sum_i q_i = 1$.
- If X is discrete then there is a countable set E such that $P[X \in E] = 1$.
- The distribution function of X is specified completely by these q_i

probability mass function, f_X

- Let X be a discrete rv with $X \in \{x_1, x_2, \cdots\}$.
- The probability mass function (pmf) of X is defined by

$$f_X(x_i) = P[X = x_i]; \quad f_X(x) = 0, \quad \text{for all other } x$$

- f_X is also a real-valued function of a real variable.
- We can write the definition compactly as $f_X(x) = P[X = x]$
- The distribution function (df) and the pmf are related as

$$F_X(x) = \sum_{i: x_i \le x} f_X(x_i)$$

$$f_X(x) = F_X(x) - F_X(x^-)$$

- We can get pmf from df and df from pmf.

Properties of pmf

- The probability mass function of a discrete random variable $X \in \{x_1, x_2, \cdots\}$ satisfies
  1. $f_X(x) \ge 0, \forall x$ and $f_X(x) = 0$ if $x \ne x_i$ for some i
  2. $\sum_i f_X(x_i) = 1$
- Any function satisfying the above two would be a pmf of some discrete random variable.
- We can specify a discrete random variable by giving either F_X or f_X.
- Please remember that we have defined distribution function for any random variable. But pmf is defined only for discrete random variables

- Any discrete random variable can be specified by
  - giving the set of values of X, $\{x_1, x_2, \cdots\}$, and
  - numbers q_i such that $q_i = P[X = x_i] = f_X(x_i)$
- Note that we must have $q_i \geq 0$ and $\sum_i q_i = 1$.
- As we saw this is how we can specify a probability assignment on any countable sample space.

Computations of Probabilities for discrete rv's

- A discrete random variable is specified by giving either df or pmf. One can be obtained from the other.
- We normally specify it through the pmf.
- Given $X \in \{x_1, x_2, \cdots\}$ and f_X, we can (in principle) compute probability of any event

$$P[X \in B] = \sum_{\substack{i: \\ x_i \in B}} f_X(x_i)$$

- For example, if $X \in \{0, 1, 2, 3\}$ then

$$P[X \in [0.5,\ 1.32] \cup [2.75,\ 5.2]\ ] = f_X(1) + f_X(3)$$

- We next look at some standard discrete random variable models

Bernoulli Distribution

- Bernoulli random variable: $X \in \{0,\ 1\}$ with

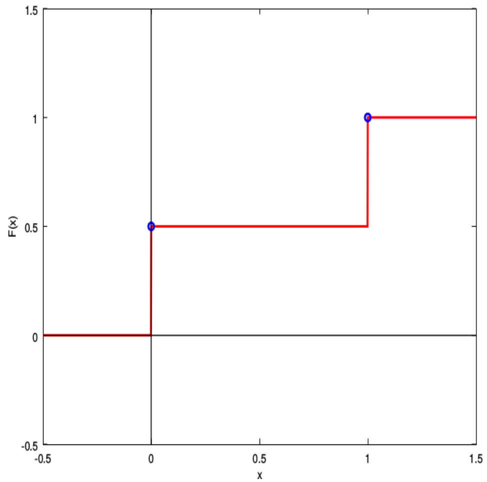
$f_X(1) = p;\ f_X(0) = 1-p;$ where $0 < p < 1$ is a parameter

- This f_X is easily seen to be a pmf
- Consider $(\Omega, \mathcal{F}, P)$ with $B \in \mathcal{F}$. (The Ω here may be uncountable).
- Consider the random variable

$$I_B(\omega) = \begin{cases} 0 & \text{if } \omega \notin B \\ 1 & \text{if } \omega \in B \end{cases}$$

- It is called indicator (random variable) of B.
- $P[I_B = 1] = P(\{\omega\ :\ I_B(\omega) = 1\}) = P(B)$
- Thus, this indicator rv has Bernoulli distribution with $p = P(B)$

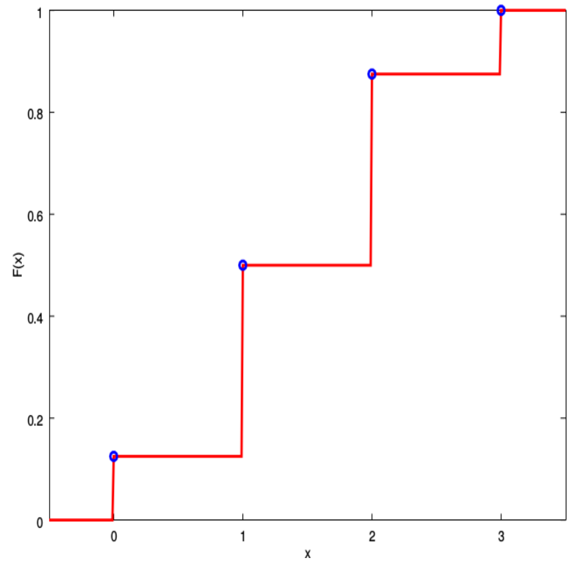
One of the df examples we saw earlier is that of Bernoulli



Binomial Distribution

- $X \in \{0,\ 1,\ \cdots,\ n\}$ with pmf
$$f_X(k) = \ {}^nC_k\ p^k\ (1-p)^{n-k},\ k = 0, 1, \cdots, n$$
where n, p are parameters (n is a $+$ve integer and $0 < p < 1$).
- This is easily seen to be a pmf
$$\sum_{k=0}^{n} {}^nC_k\ p^k\ (1-p)^{n-k} = (p + 1 - p)^n = 1$$
- Consider n independent tosses of coin whose probability of heads is p. If X is the number of heads then X has the above binomial distribution.
(Number of successes in n bernoulli trials)
- Any one outcome (a seq of length n) with k heads would have probability $p^k(1-p)^{n-k}$. There are ${}^nC_k$ outcomes with exactly k heads.

One of the df examples we considered was that of Binomial



Poisson Distribution

- $X \in \{0,\ 1,\ 2, \cdots\}$ with pmf
$$f_X(k) = \frac{\lambda^k\ e^{-\lambda}}{k!},\ k = 0, 1, 2, \cdots$$
where $\lambda > 0$ is a parameter.
- We can see this to be a pmf by
$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\ e^{-\lambda} = e^{\lambda}\ e^{-\lambda} = 1$$
- Poisson distribution is also useful in many applications

Geometric Distribution

- $X \in \{1,\ 2, \cdots\}$ with pmf
$$f_X(k) = (1-p)^{k-1}\ p,\ k = 1, 2, \cdots$$
where $0 < p < 1$ is a parameter.
- Consider tossing a coin (with prob of H being p) repeatedly till we get a head. X is the toss number on which we got the first head.
- In general waiting for 'success' in independent Bernoulli trials.

Memoryless property of geometric distribution

- Suppose X is a geometric rv. Let m, n be positive integers.
- We want to calculate $P(\,[X > m+n] \mid [X > m]\,)$
  (Remember that $[X > m]$ etc are events)
- Let us first calculate $P[X > n]$ for any positive integer n

$$\begin{aligned} P[X > n] &= \sum_{k=n+1}^{\infty} P[X = k] = \sum_{k=n+1}^{\infty} (1-p)^{k-1}\, p \\ &= p\,\frac{(1-p)^n}{1-(1-p)} = (1-p)^n \end{aligned}$$

(Does this also tell us what is df of geometric rv?)

- Now we can compute the required conditional probability

$$\begin{aligned} P[X > m+n|X > m] &= \frac{P[X > m+n, X > m]}{P[X > m]} \\ &= \frac{P[X > m+n]}{P[X > m]} \\ &= \frac{(1-p)^{m+n}}{(1-p)^m} = (1-p)^n \\ \Rightarrow P[X > m+n|X > m] &= P[X > n] \end{aligned}$$

- This is known as the memoryless property of geometric distribution
- Same as

$$P[X > m+n] = P[X > m]P[X > n]$$

- If X is a geometric random variable, it satisfies

$$P[X > m+n|X > m] = P[X > n]$$

- This is same as

$$P[X > m+n] = P[X > m]P[X > n]$$

- Does it say that $[X > m]$ is independent of $[X > n]$
- NO!
  Because $[X > m+n]$ is not equal to intersection of $[X > m]$ and $[X > n]$

Memoryless property defines geometric rv

- Suppose $X \in \{0,\ 1, \cdots\}$ is a discrete rv satisfying, for all non-negative integers, m, n

$$P[X > m+n] = P[X > m]P[X > n]$$

- We will show that X has geometric distribution
- First, note that
  $P[X > 0] = P[X > 0+0] = (P[X > 0])^2$
  $\Rightarrow P[X > 0]$ is either 1 or 0.
- Let us take $P[X > 0] = 1$ (and hence $P[X = 0] = 0$).

- We have, for any m,

$$\begin{aligned} P[X > m] &= P[X > (m-1)+1] \\ &= P[X > m-1]P[X > 1] \\ &= P[X > m-2] \left(P[X > 1]\right)^2 \end{aligned}$$

- Let $q = P[X > 1]$. Iterating on the above, we get

$$P[X > m] = P[X > 0] \left(P[X > 1]\right)^m = q^m$$

- Using this, we can get pmf of X as

$$P[X = m] = P[X > m-1] - P[X > m] = q^{m-1} - q^m = q^{m-1}(1-q)$$

- This is pmf of geometric (with $q = (1-p)$)

Continuous Random Variables

- A rv, X, is said to be continuous (or of continuous type) if its distribution function, F_X is absolutely continuous.

Absolute Continuity

- A function $g : \Re \to \Re$ is absolutely continuous on an interval, I, if given any $\epsilon > 0$ there is a $\delta > 0$ such that for any finite sequence of pair-wise disjoint subintervals, (x_k, y_k), with $x_k, y_k \in I$, $\forall k$, satisfying $\sum_k (y_k - x_k) < \delta$, we have $\sum_k |f(y_k) - f(x_k)| < \epsilon$
- A function that is absolutely continuous on a (finite) closed interval is uniformly continuous.
- If g is absolutely continuous on $[a,\ b]$ then there exists an integrable function h such that

$$g(x) = g(a) + \int_a^x h(t)\ dt, \quad \forall x \in [a,\ b]$$

- In the above, g would be differentiable almost everywhere and h would be its derivative (wherever g is differentiable).

Continuous Random Variables

- A rv, X, is said to be continuous (or of continuous type) if its distribution function, F_X is absolutely continuous.
- That is, if there exists a function $f_X : \Re \to \Re$ such that

$$F_X(x) = \int_{-\infty}^x f_X(t)\ dt, \quad \forall x$$

- f_X is called the probability density function (pdf) of X.
- Note that F_X here is continuous
- By the fundamental theorem of claculus, we have

$$\frac{dF_x(x)}{dx} = f_X(x), \ \forall x \text{ where } f_X \text{ is continuous}$$

Continuous Random Variables

- If X is a continuous rv then its distribution function, F_X, is continuous.
- Hence a discrete random variable is not a continuous rv!
- If a rv takes countably many values then it is discrete.
- However, if a rv takes uncoutably infinitely many distinct values, it does not necessarily imply it is of continuous type.
- As mentioned earlier, there would be many random variables that are neither discrete nor continuous.

Continuous Random Variables

- The df of a continuous rv is continuous.
- This implies
  $F_X(x) = F_X(x^+) = F_X(x^-)$
- Hence, if X is a continuous random variable then

$$P[X = x] = F_X(x) - F_X(x^-) = 0, \ \forall x$$

Continuous Random Variables

- A rv, X, is said to be continuous (or of continuous type) if its distribution function, F_X is absolutely continuous.
- The df of a continuous random variable can be written as

$$F_X(x) = \int_{-\infty}^{x} f_X(t) \, dt, \ \ \forall x$$

- This f_X is the probability density function (pdf) of X.

$$\frac{dF_x(x)}{dx} = f_X(x), \ \forall x \text{ where } f_X \text{ is continuous}$$

Probability Density Function

- The pdf of a continuous rv is defined to be the f_X that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t) \, dt, \ \ \forall x$$

- Since $F_X(\infty) = 1$, we must have $\int_{-\infty}^{\infty} f_X(t) \, dt = 1$
- For $x_1 \leq x_2$ we need $F_X(x_1) \leq F_X(x_2)$ and hence we need

$$\int_{-\infty}^{x_1} f_X(t) \, dt \leq \int_{-\infty}^{x_2} f_X(t) \, dt \quad \Rightarrow \quad \int_{x_1}^{x_2} f_X(t) \, dt \geq 0, \forall x_1 < x_2$$
$$\Rightarrow \quad f_X(x) \geq 0, \forall x$$

Properties of pdf

- The pdf, $f_X : \Re \to \Re$, of a continuous rv satisfies
  A1. $f_X(x) \geq 0, \ \forall x$
  A2. $\int_{-\infty}^{\infty} f_X(t)\ dt = 1$
- Any f_X that satisfies the above two would be the probability density function of a continuous rv
- Given f_X satifying the above two, define

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\ dt, \ \ \forall x$$

  This F_X satisfies
  1. $F_X(-\infty) = 0; \ F_X(\infty) = 1$
  2. F_X is non decreasing.
  3. F_X is continuous (and hence right continuous with left limits)
- This shows the the F_X is a df and hence f_X is a pdf

Continuous rv – example

- Consider a probability space with $\Omega = [0,\ 1]$ and with the 'usual' probability assignment (where probability of an interval is its length)
- Earlier we considered the rv $X(\omega) = \omega$ on this probability space.
- We found that the df for this is

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

  This is absolutely continuous and we can get the pdf as

$$f_X(x) = 1 \ \text{ if } 0 < x < 1; \ (f_X(x) = 0, \ \ \text{otherwise})$$

- On the same probability space, consider rv $Y(\omega) = 1 - \omega$.
- Let us find F_Y and f_Y.

- $Y(\omega) = 1 - \omega$.

$$\begin{aligned} [Y \leq y] &= \{\omega \ : \ Y(\omega) \leq y\} = \{\omega \in [0,\ 1] \ : \ 1 - \omega \leq y\} \\ &= \{\omega \in [0,\ 1] \ : \ \omega \geq 1 - y\} \\ &= \begin{cases} \phi & \text{if } \ y < 0 \\ \Omega & \text{if } \ y \geq 1 \\ [1-y, 1] & \text{if } \ 0 \leq y < 1 \end{cases} \end{aligned}$$

- Hence the df of Y is

$$F_Y(y) = \begin{cases} 0 & \text{if } \ y < 0 \\ y & \text{if } \ 0 \leq y < 1 \\ 1 & \text{if } \ y \geq 1 \end{cases}$$

- We have $F_X = F_Y$ and thus $f_X = f_Y$. (However, note that $X(\omega) \neq Y(\omega)$ except at $\omega = 0.5$).

- Let X be a continuous rv.
- It can be specified by giving either F_X or the pdf, f_X.
- We can, in principle, compute probability of any event as

$$P[X \in B] = \int_B f_X(t)\ dt, \ \ \forall B \in \mathcal{B}$$

- In particular, we have

$$P[X \in [a,\ b]] = P[a \leq X \leq b] = \int_a^b f_X(t)\ dt = F_X(b) - F_X(a)$$

- Since the integral over the open or closed intervals is the same, we have, for continuous rv,

$$P[a \leq X \leq b] = P[a < X \leq b] = P[a \leq X < b] \text{ etc.}$$

- Recall that for a general rv

$$F_X(b) - F_X(a) = P[a < X \leq b]$$

- If X is a continuous rv, we have

$$P[a \leq X \leq b] = \int_a^b f_X(t)\, dt$$

- Thus

$$P[x \leq X \leq x + \Delta x] = \int_x^{x+\Delta x} f_X(t)\, dt \approx f_X(x)\, \Delta x$$

- That is why f_X is called probability density function.

- For any random variable, the df is defined and it is given by

$$F_X(x) = P[X \leq x] = P[X \in (-\infty,\ x]\,]$$

- The value of $F_X(x)$ at any x is probability of some event.
- The pmf is defined only for discrete random variables as $f_X(x) = P[X = x]$
- The value of pmf is also a probability
- We use the same symbol for pdf (as for pmf), defined by

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\, dx$$

- Note that the value of pdf is not a probability.
- We can say $f_X(x)\, dx \approx P[x \leq X \leq x + dx]$

A note on notation

- The df, F_X, and the pmf or pdf, f_X, are all functions defined on $\Re$.
- Hence you should not write $F_X(X \leq 5)$.
  You should write $F_X(5)$ to denote $P[X \leq 5]$.
- For a discrete rv, X, one should not write $f_X(X = 5)$.
  It is $f_X(5)$ which gives $P[X = 5]$.
- Writing $f_X(X = 5)$ when f_X is a pdf, is particularly bad.
  Note that for a continuous rv, $P[X = 5] = 0$ and $f_X(5) \neq P[X = 5]$.

- A continuous random variable is a probability model on uncountably infinite Ω.
- For this, we take $\Re$ as our sample space.
- We can specify a continuous rv either through the df or through the pdf.
- The df, F_X, of a cont rv allows you to (consistently) assign probabilities to all Borel subsets of real line.
- We next consider a few standard continuous random variables.

Uniform distribution

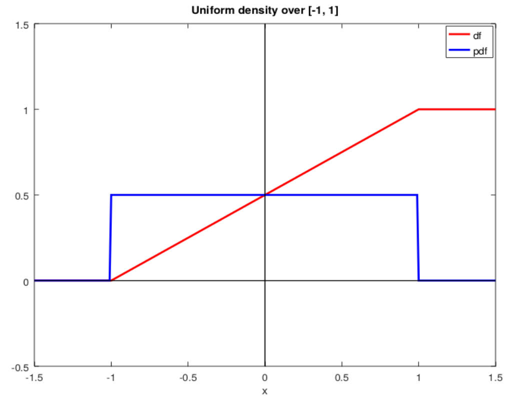
- X is uniform over $[a,\ b]$ when its pdf is

$$f_X(x) = \frac{1}{b-a}, \quad a \le x \le b$$

  ($f_X(x) = 0$ for all other values of x).
- Uniform distribution over open or closed interval is essentially the same.
- When X has this distribution, we say $X \sim U[a,\ b]$
- By integrating the above, we can see the df as

$$F_X(x) = \begin{cases} \int_{-\infty}^{x} f_X(x)\ dx = \int_{-\infty}^{x} 0\ dx = 0 & \text{if } \ x < a \\ \int_{-\infty}^{a} 0\ dx + \int_{a}^{x} \frac{1}{b-a}\ dx = \frac{x-a}{b-a} & \text{if } \ a \le x < b \\ 0 + \int_{a}^{b} \frac{1}{b-a}\ dx + 0 = 1 & \text{if } \ x \ge b \end{cases}$$

- A plot of density and distribution functions of a uniform rv is given below



- Let $X \sim U[a,\ b]$. Then $f_X(x) = \frac{1}{b-a}$, $a \le x \le b$
- Let $[c,\ d] \subset [a,\ b]$.
- Then $P[X \in [c,\ d]] = \int_c^d f_X(t)\ dt = \frac{d-c}{b-a}$
- Probability of an interval is proportional to its length.
- The earlier examples we considered are uniform over $[0,\ 1]$.

Exponential distribution

- The pdf of exponential distribution is

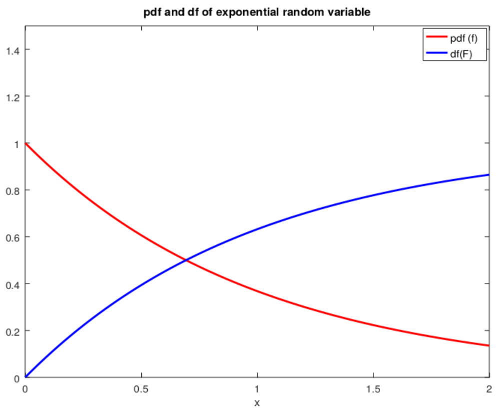
$$f_X(x) = \lambda\, e^{-\lambda x}, \quad x > 0, \ (\lambda > 0 \text{ is a parameter})$$

  (By our notation, $f_X(x) = 0$ for $x \le 0$)
- It is easy to verify $\int_0^\infty f_X(x)\ dx = 1$.
- It is easy to see that $F_x(x) = 0$ for $x \le 0$.
- For $x > 0$ we can compute F_X by integrating f_X:

$$F_X(x) = \int_0^x \lambda\, e^{-\lambda x}\ dx = \lambda \left. \frac{e^{-\lambda x}}{-\lambda} \right|_0^x = 1 - e^{-\lambda x}$$

- This also gives us: $P[X > x] = 1 - F_X(x) = e^{-\lambda x}$ for $x > 0$.

- A plot of density and distribution functions of an exponential rv is given below



exponential distribution is memoryless

- If X has exponential distribution, then, for $t, s > 0$,

$$P[X > t{+}s] = e^{-\lambda(t+s)} = e^{-\lambda t}\, e^{-\lambda s} = P[X > t]\; P[X > s]$$

- This gives us the memoryless property

$$P[X > t + s \mid X > t] = \frac{[P[X > t + s]}{P[X > t]} = P[X > s]$$

- Exponential distribution is a useful model for, e.g., life-time of components.
- If the distribution of a non-negative continuous random variable is memory less then it must be exponential.

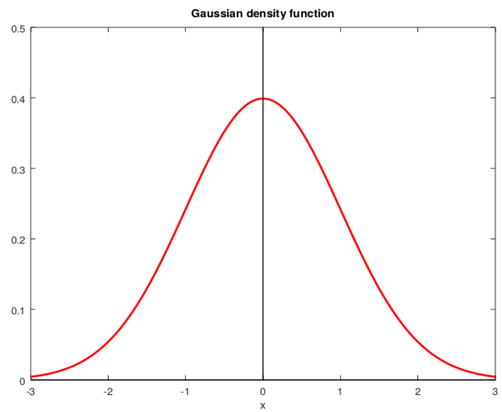
Gaussian Distribution

- The pdf of Gaussian distribution is given by

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

  where $\sigma > 0$ and $\mu \in \Re$ are parameters.
- We write $X \sim \mathcal{N}(\mu, \sigma^2)$ to denote that X has Gaussian density with parameters μ and σ.
- This is also called the Normal distribution.
- The special case where $\mu = 0$ and $\sigma^2 = 1$ is called standard Gaussian (or standard Normal) distribution.

- A plot of Gaussian density functions is given below



- $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \ -\infty < x < \infty$
- Showing that the density integrates to 1 is not trivial.
- Take $\mu = 0, \sigma = 1$. Let $I = \int_{-\infty}^{\infty} f_X(x)\, dx$. Then

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-0.5x^2}\, dx \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-0.5y^2}\, dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi}\, e^{-0.5(x^2+y^2)}\, dx\, dy \end{aligned}$$

- Now converting the above integral into polar coordinates would allow you to show $I = 1$.
  (Left as an exercise for you!)

Recap: Random Variable

- Given a probability space $(\Omega, \mathcal{F}, P)$, a random variable is a real-valued function on Ω.
- It essentially results in an induced probability space

$$(\Omega, \mathcal{F}, P) \xrightarrow{X} (\Re, \mathcal{B}, P_X)$$

where $\mathcal{B}$ is the Borel σ-algebra and

$$P_X(B) = P[X \in B] = P\left(\{\omega \in \Omega \ : \ X(\omega) \in B\}\right)$$

- For X to be a random variable

$$\{\omega \in \Omega \ : \ X(\omega) \in B\} \ \in \mathcal{F}, \ \ \forall B \in \mathcal{B}$$

Recap: Distribution Function

- Let X be a random variable. It distribution function, $F_X : \Re \to \Re$, is defined by

$$F_X(x) = P[X \le x] = P(\{\omega \in \Omega \ : \ X(\omega) \le x\})$$

- The distribution function, F_X, completely specifies the probability measure, P_X.
- The distribution function satisfies
  1. $0 \le F_X(x) \le 1, \ \forall x$
  2. $F_X(-\infty) = 0; \ F_X(\infty) = 1$
  3. F_X is non-decreasing: $x_1 \le x_2 \ \Rightarrow \ F_X(x_1) \le F_X(x_2)$
  4. F_X is right continuous and has left-hand limits.
- We also have
  $F_X(x^+) - F_X(x^-) = F_X(x) - F_X(x^-) = P[X = x]$
  $P[a < X \le b] = F_X(b) - F_X(a)$.

Recap: Discrete Random Variable

- A random variable X is said to be discrete if it takes only finitely many or countably infinitely many distinct values.
- Let $X \in \{x_1, x_2, \cdots\}$
- Its distribution function, F_X is a stair-case function with jump discontinuities at each x_i and the magnitude of the jump at x_i is equal to $P[X = x_i]$

Recap: probability mass function

- Let $X \in \{x_1, x_2, \cdots\}$.
- The probability mass function (pmf) of X is defined by

$$f_X(x_i) = P[X = x_i]; \quad f_X(x) = 0, \quad \text{for all other } x$$

- It satisfies
  1. $f_X(x) \geq 0, \forall x$ and $f_X(x) = 0$ if $x \neq x_i$ for some i
  2. $\sum_i f_X(x_i) = 1$
- We have
  $F_X(x) = \sum_{i: x_i \leq x} f_X(x_i)$
  $f_X(x) = F_X(x) - F_X(x^-)$
- We can calculate the probability of any event as

$$P[X \in B] = \sum_{\substack{i: \\ x_i \in B}} f_X(x_i)$$

Recap: continuous random variable

- X is said to be a continuous random variable if there exists a function $f_X : \Re \to \Re$ satisfying

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\, dx$$

  The f_X is called the probability density function.
- Same as saying F_X is absolutely continuous.
- Since F_X is continuous here, we have

$$P[X = x] = F_X(x) - F_X(x^-) = 0, \ \forall x$$

- A continuous rv takes uncountably many distinct values. However, not every rv that takes uncountably many values is a continuous rv

Recap: probability density function

- The pdf of a continuous rv is defined to be the f_X that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt, \quad \forall x$$

- It satisfies
  1. $f_X(x) \geq 0, \quad \forall x$
  2. $\int_{-\infty}^{\infty} f_X(t)\, dt = 1$
- We can, in principle, compute probability of any event as

$$P[X \in B] = \int_B f_X(t)\, dt, \quad \forall B \in \mathcal{B}$$

- In particular,

$$P[a \leq X \leq b] = \int_a^b f_X(t)\, dt$$

Recap: some discrete random variables

- Bernoulli: $X \in \{0, 1\}$; parameter: $p, \ 0 < p < 1$

$$f_X(1) = p; \ f_X(0) = 1 - p$$

- Binomial: $X \in \{0, 1, \cdots, n\}$; Parameters: n, p

$$f_X(x) = \ {}^nC_x p^x (1-p)^{n-x}, \ x = 0, \cdots, n$$

- Poisson: $X \in \{0, 1, \cdots\}$; Parameter: $\lambda > 0$.

$$f_X(x) = e^{-\lambda} \frac{\lambda^x}{x!}, \ x = 0, 1, \cdots$$

- Geometric: $X \in \{1, 2, \cdots\}$; Parameter: $p, \ 0 < p < 1$.

$$f_X(x) = p(1-p)^{x-1}, \ x = 1, 2, \cdots$$

Recap: Some continuous random variables

- Uniform over $[a,\ b]$: Parameters: $a, b,\ b > a$.

$$f_X(x) = \frac{1}{b-a},\ \ a \leq x \leq b.$$

- exponential: Parameter: $\lambda > 0$.

$$f_X(x) = \lambda e^{-\lambda x},\ \ x \geq 0.$$

- Gaussian (Normal): Parameters: $\sigma > 0, \mu$.

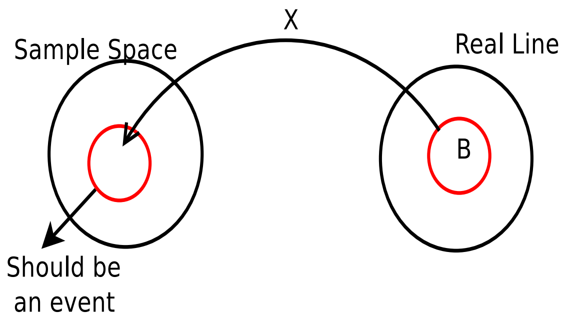
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ \ -\infty < x < \infty$$

Functions of a random variable

- We next look at random variables defined in terms of other random variables.

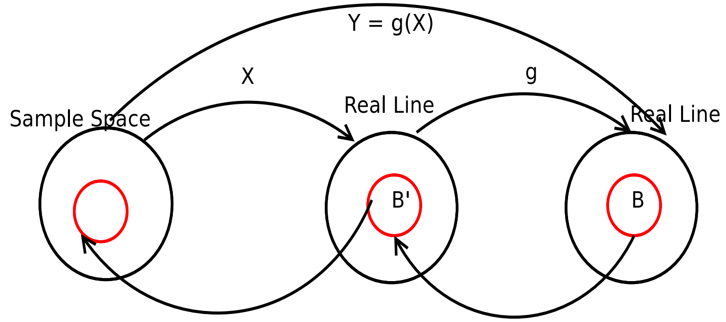
- Let X be a rv on some probability space $(\Omega, \mathcal{F}, P)$.
- Recall that $X : \Omega \to \Re$.
- Also recall that

$$[X \in B] \triangleq \{\omega\ :\ X(\omega) \in B\} \in \mathcal{F},\ \forall B \in \mathcal{B}$$



Functions of a Random Variable

- Let X be a rv on some probability space $(\Omega, \mathcal{F}, P)$. (Recall $X : \Omega \to \Re$)
- Consider a function $g : \Re \to \Re$
- Let $Y = g(X)$. Then Y also maps Ω into real line.
- If g is a 'nice' function, Y would also be a random variable
- We need: $g^{-1}(B) \triangleq \{z \in \Re\ :\ g(z) \in B\} \in \mathcal{B},\ \forall B \in \mathcal{B}$



- Let X be a rv and let $Y = g(X)$.
- The distribution function of Y is given by

$$\begin{aligned} F_Y(y) &= P[Y \le y] \\ &= P[g(X) \le y] \\ &= P[g(X) \in (-\infty, \ y] \,] \\ &= P[X \in \{z \ : \ g(z) \le y\}] \end{aligned}$$

- This probability can be obtained from distribution of X.
- Thus, in principle, we can find the distribution of Y if we know that of X

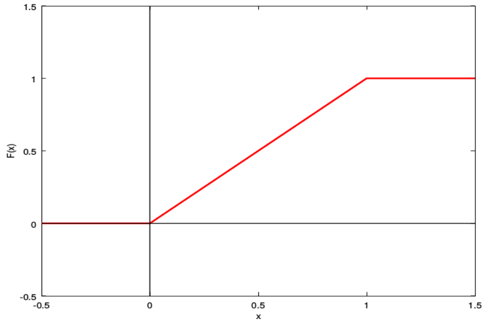
Example

- Let $Y = aX + b, \ a > 0$.
- Then we have

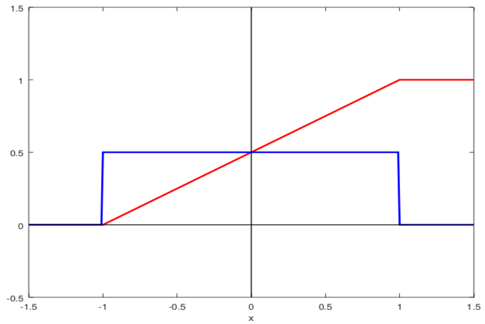
$$\begin{aligned} F_Y(y) &= P[Y \le y] \\ &= P[aX + b \le y] \\ &= P[aX \le y - b] \\ &= P\left[X \le \frac{y-b}{a}\right], \quad \text{since } a > 0 \\ &= F_X\left(\frac{y-b}{a}\right) \end{aligned}$$

- This tells us how to find df of Y when it is an affine function of X.
- If X is continuous rv, then, $f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right)$

- In many examples we would be using uniform random variables.
- Let $X \sim U[0, \ 1]$. Its pdf is $f_X(x) = 1, \ 0 \le x \le 1$.
- Integrating this we get the df: $F_X(x) = x, \ 0 \le x \le 1$



- Let $X \sim U[-1, \ 1]$. The pdf would be $f_X(x) = 0.5, \ -1 \le x \le 1$.
- Integrating this, we get the df: $F_X(x) = \frac{1+x}{2}$ for $-1 \le x \le 1$.
- These are plotted below



- Suppose $X \sim U[0,\ 1]$ and $Y = aX + b$
- The df for Y would be

$$F_Y(y) = F_X\left(\frac{y-b}{a}\right) = \begin{cases} 0 & \frac{y-b}{a} \leq 0 \\ \frac{y-b}{a} & 0 \leq \frac{y-b}{a} \leq 1 \\ 1 & \frac{y-b}{a} \geq 1 \end{cases}$$

- Thus we get the df for Y as

$$F_Y(y) = \begin{cases} 0 & y \leq b \\ \frac{y-b}{a} & b \leq y \leq a+b \\ 1 & y \geq a+b \end{cases}$$

- Hence $f_Y(y) = \frac{1}{a},\ y \in [b,\ a+b]$ and $Y \sim U[b,\ a+b]$.
- If $X \sim U[0,\ 1]$ then $Y = aX + b$, $(a > 0)$, is uniform over $[b,\ a+b]$.

- Recall that Gaussian density is $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- We denote this as $\mathcal{N}(\mu, \sigma^2)$
- Let $Y = aX + b$ where $X \sim \mathcal{N}(0,1)$. The df of Y is

$$\begin{aligned} F_Y(y) &= F_X\left(\frac{y-b}{a}\right) \\ &= \int_{-\infty}^{\frac{y-b}{a}} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}\ dx \end{aligned}$$

we make a substitution: $t = ax + b \Rightarrow x = \dfrac{t-b}{a}$, and $dx = \dfrac{1}{a}dt$

$$F_Y(y) = \int_{-\infty}^{y} \frac{1}{a\sqrt{2\pi}}\, e^{-\frac{(t-b)^2}{2a^2}}\ dt$$

- This shows that $Y \sim \mathcal{N}(b, a^2)$

- Suppose X is a discrete rv with $X \in \{x_1, x_2, \cdots\}$.
- Suppose $Y = g(X)$.
- Then Y is also discrete and $Y \in \{g(x_1), g(x_2), \cdots\}$.
- Though we use this notation, we should note:
  1. these values may not be distinct (it is possible that $g(x_i) = g(x_j)$);
  2. $g(x_1)$ may not be the smallest value of Y and so on.
- We can find the pmf of Y as

$$\begin{aligned} f_Y(y) &= p[Y = y] = P[g(X) = y] \\ &= P[X \in \{x_i\ :\ g(x_i) = y\}] \\ &= \sum_{\substack{i: \\ g(x_i)=y}} f_X(x_i) \end{aligned}$$

- Let $X \in \{1, 2, \cdots, N\}$ with $f_X(k) = \frac{1}{N},\ 1 \leq k \leq N$
- Let $Y = aX + b$, $(a > 0)$.
- Then $Y \in \{b + a,\ b + 2a, \cdots,\ b + Na\}$.
- We get the pmf of Y as

$$f_Y(b + ka) = f_X(k) = \frac{1}{N},\ 1 \leq k \leq N$$

- Suppose X is geometric: $f_X(k) = (1-p)^{k-1}p, \ k = 1, 2, \cdots$.
- Let $Y = X - 1$
- We get the pmf of Y as

$$\begin{aligned} f_Y(j) &= P[X - 1 = j] \\ &= P[X = j + 1] \\ &= (1-p)^j p, \ j = 0, 1, \cdots \end{aligned}$$

- Suppose X is geometric. $(f_X(k) = (1-p)^{k-1}p)$
- Let $Y = \max(X, 5) \quad \Rightarrow \quad Y \in \{5, 6, \cdots\}$
- We can calculate the pmf of Y as

$$\begin{aligned} f_Y(5) = P[\max(X,5) = 5] &= \sum_{k=1}^{5} f_X(k) = 1 - (1-p)^5 \\ f_Y(k) = P[\max(X,5) = k] &= P[X = k] = (1-p)^{k-1}p, \ k = 6, 7, \cdots \end{aligned}$$

- We next consider $Y = h(X)$ where

$$h(x) = \begin{cases} x & \text{if } \ x > 0 \\ 0 & \text{otherwise} \end{cases}$$

- This is written as $Y = X^+$ to indicate the function only keeps the positive part.

- Let $X \sim U[-1, \ 1]$: $F_X(x) = \frac{1+x}{2}$ for $-1 \leq x \leq 1$.
- Let $Y = X^+$. That is,

$$Y = X^+ = \begin{cases} X & \text{if } X > 0 \\ 0 & \text{otherwise} \end{cases}$$

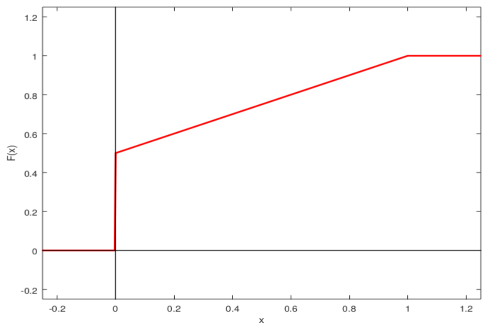
- For $y < 0$, $F_Y(y) = P[Y \leq y] = 0$ because $Y \geq 0$.
- $F_Y(0) = P[Y \leq 0] = P[X \leq 0] = 0.5$.
- For $0 < y < 1$, $F_Y(y) = P[Y \leq y] = P[X \leq y] = \frac{1+y}{2}$
- For $y \geq 1$, $F_Y(y) = 1$.
- Thus, the df of Y is

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ 0.5 & \text{if } y = 0 \\ \frac{1+y}{2} & \text{if } 0 < y < 1 \\ 1 & \text{if } y \geq 1 \end{cases}$$

- The df of Y is

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ \frac{1+y}{2} & \text{if } 0 \le y < 1 \\ 1 & \text{if } y \ge 1 \end{cases}$$

- This is plotted below



- This is neither a continuous rv nor a discrete rv.

- Let $Y = X^2$.
- For $y < 0$, $F_Y(y) = P[Y \le y] = 0$ (since $Y \ge 0$)
- For $y \ge 0$, we can get $F_Y(y)$ as

$$\begin{aligned} F_Y(y) &= P[Y \le y] = P[X^2 \le y] \\ &= P[-\sqrt{y} \le X \le \sqrt{y}] \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) + P[X = -\sqrt{y}] \end{aligned}$$

- If X is a continuous random variable, then we get

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}\left(F_X(\sqrt{y}) - F_X(-\sqrt{y})\right) \\ &= \frac{1}{2\sqrt{y}}\left[f_X(\sqrt{y}) + f_X(-\sqrt{y})\right] \end{aligned}$$

- This is the general formula for density of X^2 when X is continuous rv.

- Let $X \sim \mathcal{N}(0,1)$: $f_X(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}$
- Let $Y = X^2$. Then we know $f_Y(y) = 0$ for $y < 0$. For $y \ge 0$,

$$\begin{aligned} f_Y(y) &= \frac{1}{2\sqrt{y}}\left[f_X(\sqrt{y}) + f_X(-\sqrt{y})\right] \\ &= \frac{1}{2\sqrt{y}}\left[\frac{1}{\sqrt{2\pi}}\, e^{-\frac{y}{2}} + \frac{1}{\sqrt{2\pi}}\, e^{-\frac{y}{2}}\right] \\ &= \frac{1}{2\sqrt{y}}\,\frac{2}{\sqrt{2\pi}}\, e^{-\frac{y}{2}} \\ &= \frac{1}{\sqrt{\pi}}\left(\frac{1}{2}\right)^{0.5} y^{-0.5}\, e^{-\frac{1}{2}y} \end{aligned}$$

- This is an example of gamma density.

Gamma density

- The Gamma function is given by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\, e^{-x}\, dx$$

It can be easily verified that $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$.
- The Gamma density is given by

$$f(x) = \frac{1}{\Gamma(\alpha)}\, \lambda^\alpha x^{\alpha-1} e^{-\lambda x} = \frac{1}{\Gamma(\alpha)}\, (\lambda x)^{\alpha-1}\, \lambda e^{-\lambda x}, \quad x > 0$$

- Here $\alpha, \lambda > 0$ are parameters.
- The earlier density we saw corresponds to $\alpha = \lambda = 0.5$:

$$f_Y(y) = \frac{1}{\sqrt{\pi}}\left(\frac{1}{2}\right)^{0.5} y^{-0.5}\, e^{-\frac{1}{2}y}, \quad y > 0$$

- The gamma density with parameters $\alpha, \lambda > 0$ is given by

$$f(x) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x}, \quad x > 0$$

- If $X \sim \mathcal{N}(0, 1)$ then X^2 has gamma density with parameters $\alpha = \lambda = 0.5$.
- When α is a positive integer then the gamma density is known as the Erlang density.
- If $\alpha = 1$, gamma density becomes exponential density.

- Let $X \sim U(0,\ 1)$.
- Let $Y = \frac{-1}{\lambda} \ln(1 - X)$, where $\lambda > 0$.
- Note that $Y \geq 0$. We can find its df:

$$\begin{aligned} F_Y(y) &= P[Y \leq y] = P\left[\frac{-1}{\lambda} \ln(1 - X)\ \leq\ y\right] \\ &= P[-\ln(1 - X)\ \leq\ \lambda y] \\ &= P[\ln(1 - X)\ \geq\ -\lambda y] \\ &= P[1 - X\ \geq\ e^{-\lambda y}] \\ &= P[X\ \leq\ 1 - e^{-\lambda y}] \\ &= 1 - e^{-\lambda y},\ y \geq 0 \quad \text{(since } X \sim U(0, 1)\text{)} \end{aligned}$$

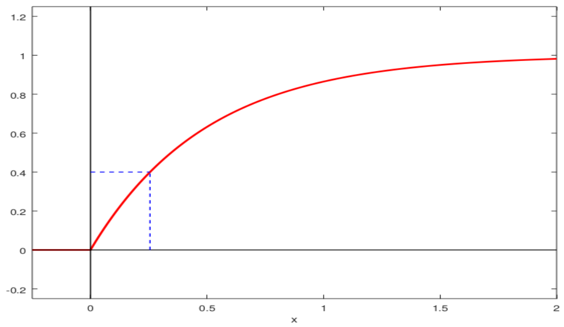
- Thus Y has exponential density
- If $X \sim U(0,\ 1)$, $\frac{-1}{\lambda} \ln(1 - X)$ has exponential density

- If $X \sim U(0,\ 1)$, $\frac{-1}{\lambda} \ln(1 - X)$ has exponential density
- This is actually a special case of a general result.
- The exponential distribution fn is $F(x) = 1 - e^{-\lambda x}$.
- This is continuous, strictly monotone and hence is invertible. The inverse function maps $[0,\ 1]$ to $\Re^+$. We derive its inverse:

$$z = 1 - e^{-\lambda x} \ \Rightarrow\ e^{-\lambda x} = 1 - z \ \Rightarrow\ x = \frac{-1}{\lambda} \ln(1 - z)$$

- Thus, the inverse of F is $F^{-1}(z) = \frac{-1}{\lambda} \ln(1 - z)$
- So, we had $Y = F^{-1}(X)$ and the df of Y was F

- We can visualize this as shown below



- Let G be a continuous invertible distribution function.
- Let $X \sim U[0,\ 1]$ and let $Y = G^{-1}(X)$.
- We can get the df of Y as

$$F_Y(y) = P[Y \leq y] = P[G^{-1}(X) \leq y] = P[X \leq G(y)] = G(y)$$

- Thus, starting with uniform rv, we can generate a rv with a desired distribution.
- Very useful in random number generation. Known as the inverse function method.
- Can be generalized to handle discrete rv also. It only involves defining an 'inverse' when F is a stair-case function. (Left as an exercise!)

- Let X be a cont rv with an invertible distribution function, say, F.
- Define $Y = F(X)$.
- Since range of F is $[0,\ 1]$, we know $0 \leq Y \leq 1$.
- For $0 \leq y \leq 1$ we can obtain $F_Y(y)$ as

$$F_Y(y) = P[Y \leq y] = P[F(X) \leq y] = P[X \leq F^{-1}(y)] = F(F^{-1}(y)) = y$$

- This means Y has uniform density.
- Has interesting applications.
  E.g., histogram equalization in image processing

- Let us sum-up the last two examples
- If $X \sim U[0,\ 1]$ and $Y = F^{-1}(X)$, then Y has df F.
- If df of X is F and $Y = F(X)$ then Y is uniform over $[0,\ 1]$.

- If $Y = g(X)$, we can compute distribution of Y, knowing the function g and the distribution of X.
- We have seen a number of examples.
- Finally, we look at a theorem that gives a formula for pdf of Y in certain special cases

- Let $g : \Re \to \Re$ be differentiable with $g'(x) > 0, \forall x$.
- Let X be a continuous rv with pdf f_X.
- Let $Y = g(X)$
- **Theorem**: With the above, Y is a continuous rv with pdf

$$f_Y(y) = f_X(g^{-1}(y)) \, \frac{d}{dy} g^{-1}(y), \;\; g(-\infty) \le y \le g(\infty)$$

- **Proof**: Since $g'(x) > 0$, g is strictly monotonically increasing and hence is invertible and g^{-1} would also be monotone and differentiable.
- So, range of Y is $[g(-\infty), \; g(\infty)]$.
- Now we have

$$F_Y(y) = P[Y \le y] = P[g(X) \le y] = P[X \le g^{-1}(y)] = F_X(g^{-1}(y))$$

- Since g^{-1} is differentiable, so is F_Y and we get the pdf as

$$f_Y(y) = \frac{d}{dy}(F_X(g^{-1}(y))) = f_X(g^{-1}(y)) \, \frac{d}{dy} g^{-1}(y)$$

- This completes the proof.

- Now, suppose $g'(x) < 0, \forall x$. Even then the theorem essentially holds.
- Now, g is strictly monotonically decreasing. So, we get

$$F_Y(y) = P[g(X) \le y] = P[X \ge g^{-1}(y)] = 1 - F_X(g^{-1}(y))$$

- Once again, by differentiating

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

  because g^{-1} is also monotone decreasing.
- The range of Y here is $[g(\infty), \; g(-\infty)]$
- We can combine both cases into one result.

- Let $g : \Re \to \Re$ be differentiable with $g'(x) > 0, \forall x$ or $g'(x) < 0, \forall x$.
- Let X be a continuous rv and let $Y = g(X)$.
- Then Y is a continuous rv with pdf

$$f_Y(y) = f_X(g^{-1}(y)) \, \left| \frac{d}{dy} g^{-1}(y) \right|, \;\; a \le y \le b$$

  where $a = \min(g(\infty), \; g(-\infty))$ and $b = \max(g(\infty), \; g(-\infty))$

- For an example, take $g(x) = ax + b$.
- This satisfies the conditions and $g^{-1}(y) = \frac{y-b}{a}$
- Hence we get

$$f_Y(y) = f_X(g^{-1}(y)) \, \left| \frac{d}{dy} g^{-1}(y) \right| = f_X\left(\frac{y-b}{a}\right) \left| \frac{1}{a} \right|$$

- This is an example we saw earlier.
- We need to find the range of Y based on range of X.

- The function $g(x) = x^2$ does not satisfy the conditions of the theorem.
- The utility of the theorem is somewhat limited.
- However, we can extend the theorem.
- Essentially, what we need is that for a any y, the equation $g(x) = y$ would have finite solutions and the derivative of g is not zero at any of these points.
- There are multiple '$g^{-1}(y)$' and we can get density of Y by summing all the terms.

- If $Y = g(x)$ and g is monotone,

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

- Let $x_o(y)$ be the solution of $g(x) = y$; then $g^{-1}(y) = x_o(y)$.
- Also, the derivative of g^{-1} is reciprocal of the derivative of g.
- Hence, we can also write the above as

$$f_Y(y) = f_X(x_o(y)) \left| g'(x_o(y)) \right|^{-1}$$

- However, the notation in the above may be confusing.

- We can now extend the theorem as follows.
- Suppose, for a given y, $g(x) = y$ has multiple solutions.
- Call them $x_1(y), \cdots, x_m(y)$. Assume the derivative of g is not zero at any of these points.
- Then we have

$$f_Y(y) = \sum_{k=1}^{m} f_X(x_k(y)) \left| g'(x_k(y)) \right|^{-1}$$

- If $g(x) = y$ has no solution (or no solution satisfying $g'(x) \neq 0$), then at that y, $f_Y(y) = 0$.

- Consider the old example $g(x) = x^2$.
- For $y > 0$, $x^2 = y$ has two solutions: $\sqrt{y}$ and $-\sqrt{y}$.
- At both these points, the absolute value of derivative of g is $2\sqrt{y}$ which is non-zero.
- Hence we get

$$f_Y(y) = (2\sqrt{y})^{-1} \left(f_X(\sqrt{y}) + f_X(-\sqrt{y}) \right)$$

- This is same as what we derived from first principles earlier.

Recap: Function of a random variable

- If X is a random variable and $g : \Re \rightarrow \Re$ is a function, then $Y = g(X)$ is a random variable.
- More formally, Y is a random variable if g is a Borel measurable function.
- We can determine distribution of Y given the function g and the distribution of X

Recap

- Let X be a rv and let $Y = g(X)$.
- The distribution function of Y is given by

$$\begin{aligned} F_Y(y) &= P[g(X) \leq y] \\ &= P[X \in \{z \ : \ g(z) \leq y\}] \end{aligned}$$

- This probability can be obtained from distribution of X.
- We have seen many specific examples of this.

Recap

- Suppose X is a discrete rv with $X \in \{x_1, x_2, \cdots\}$.
- Suppose $Y = g(X)$.
- Then Y is also discrete and $Y \in \{g(x_1), g(x_2), \cdots\}$.
- We can find the pmf of Y as

$$\begin{aligned} f_Y(y) &= p[Y = y] = P[g(X) = y] \\ &= P[X \in \{x_i \ : \ g(x_i) = y\}] \\ &= \sum_{\substack{i: \\ g(x_i)=y}} f_X(x_i) \end{aligned}$$

Recap

- Let $g : \Re \rightarrow \Re$ be differentiable with $g'(x) > 0, \forall x$ or $g'(x) < 0, \forall x$.
- Let X be a continuous rv and let $Y = g(X)$.
- Then Y is a continuous rv with pdf

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|, \ a \leq y \leq b$$

  where $a = \min(g(\infty), \ g(-\infty))$ and $b = \max(g(\infty), \ g(-\infty))$
- This theorem is useful in some cases to find the densities of functions of continuous random variables

Expectation and Moments of a random variable

- We next consider the important notion of expectation of a random variable

Expectation of a discrete rv

- Let X be a discrete rv with $X \in \{x_1, x_2, \cdots\}$
- We define its expectation by

$$E[X] = \sum_i x_i \; f_X(x_i)$$

- Expectation is essentially a weighted average.
- To make the above finite and well defined, we can stipulate the following as condition for existence of expectation

$$\sum_i |x_i| \; f_X(x_i) < \infty$$

Expectation of a Continuous rv

- If X is a continuous random variable with pdf, f_X, we define its expectation as

$$E[X] = \int_{-\infty}^{\infty} x \; f_X(x) \; dx$$

- Once again we can use the following as condition for existence of expectation

$$\int_{-\infty}^{\infty} |x| \; f_X(x) \; dx < \infty$$

- Sometimes we use the following notation to denote expectation of both kinds of rv

$$E[X] = \int_{-\infty}^{\infty} x \; dF_X(x)$$

- Though we consider only discrete or continuous rv's, expectation is defined for all random variables.

- Let us look at a couple of simple examples.
- Let $X \in \{1, 2, 3, 4, 5, 6\}$ and $f_X(k) = \frac{1}{6}, \; 1 \le k \le 6$.

$$EX = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{21}{6} = 3.5$$

- Let $X \sim U[0, \; 1]$

$$EX = \int_{-\infty}^{\infty} x \; f_X(x) \; dx = \int_0^1 x \; dx = 0.5$$

- When an rv takes only finitely many values or when the pdf is non-zero only on a bounded set, the expectation is always finite.

- The way we have defined existence of expectation, implies that expectation is always finite (when it exists).
- This may be needlessly restrictive in some situations. We redefine it as follows.
- Let X be a non-negative (discrete or continuous) random variable.
- We define its expectation by

$$EX = \sum_i x_i\, f_X(x_i) \quad \text{or} \quad EX = \int_{-\infty}^{\infty} x\, f_X(x)\, dx$$

  depending on whether it is discrete or continuous
  (In this course we will consider only discrete or continuous rv's)
- Note that the expectation may be infinite.
- But it always exists for non-negative random variables.

- Now let X be a rv that may not be non-negative.
- We define positive and negative parts of X by

$$X^+ = \begin{cases} X & \text{if } X > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$X^- = \begin{cases} -X & \text{if } X < 0 \\ 0 & \text{otherwise} \end{cases}$$

- Note that both X^+ and X^- are non-negative. Hence their expectations exist. (Also, $X(\omega) = X^+(\omega) - X^-(\omega),\ \forall \omega$).
- Now we define expectation of X by

$$EX = EX^+ - EX^-, \quad \text{if at least one of them is finite}$$

  Otherwise EX does not exist.
- Now, expectation does not exist only when $EX^+ = EX^- = \infty$

- This is the formal way of defining expectation of a random variable.
- We first note that if $\sum_i |x_i|\ f_X(x_i) < \infty$ then both EX^+ and EX^- would be finite and we can simply take the expectation as $EX = \sum_i x_i\ f_X(x_i)$.
- Also note that if X takes only finitely many values, the above always holds.
- Similar comments apply for a continuous random variable.
- This is what we do in this course because we deal with only discrete and continuous rv's.
- But to get a feel for the more formal definition, we look at a couple of examples.

- Let $X \in \{1, 2, \cdots\}$.
- Suppose $f_X(k) = \frac{C}{k^2}$.
- Since $\sum_k \frac{1}{k^2} < \infty$, we can find C so that $\sum_k f_X(k) = 1$. ($\sum_k \frac{1}{k^2} = \frac{\pi^2}{6}$ and hence $C = \frac{6}{\pi^2}$).
- Hence we get

$$\sum_k |x_k|\ f_X(x_k) = \sum_k x_k\ f_X(x_k) = \sum_k k\, \frac{C}{k^2} = \sum_k \frac{C}{k} = \infty$$

- Here the expectation is infinity.
- But by the formal definition it exists. (Note that here $X^+ = X$ and $X^- = 0$).

- Now suppose X takes values $1, -2, 3, -4, \cdots$ with probabilities $\frac{C}{1^2}$, $\frac{C}{2^2}$, $\frac{C}{3^2}$ and so on.
- Once again $\sum_k |x_k| f_X(x_k) = \infty$.
- But $\sum_k x_k f_X(x_k)$ is an alternating series.
- Here X^+ would take values $2k-1$ with probability $\frac{C}{(2k-1)^2}$, $k = 1, 2, \cdots$ (and the value 0 with remaining probability).
- Similarly, X^- would take values $2k$ with probability $\frac{C}{(2k)^2}$, $k = 1, 2, \cdots$ (and the value 0 with remaining probability).

$$EX^+ = \sum_k \frac{C}{2k-1} = \infty, \quad \text{and} \quad EX^- = \sum_k \frac{C}{2k} = \infty$$

- Hence EX does not exist.

- Consider a continuous random variable X with pdf

$$f_X(x) = \frac{1}{\pi} \frac{1}{1+x^2}, \; -\infty < x < \infty$$

- This is called (standard) Cauchy density. We can verify it integrates to 1

$$\int_{-\infty}^{\infty} \frac{1}{\pi} \frac{1}{1+x^2} \, dx = \frac{1}{\pi} \; \tan^{-1}(x)\Big|_{-\infty}^{\infty} = \frac{1}{\pi}\left(\frac{\pi}{2} - \frac{-\pi}{2}\right) = 1$$

- What would be EX?

$$EX = \int_{-\infty}^{\infty} x \, \frac{1}{\pi} \frac{1}{1+x^2} \, dx \overset{?}{=} 0 \text{ because } \int_{-a}^{a} \frac{x}{1+x^2} = 0?$$

- The question was

$$EX = \int_{-\infty}^{\infty} x \, \frac{1}{\pi} \frac{1}{1+x^2} \, dx \overset{?}{=} 0$$

- This depends on the definition of infinite integrals

$$\begin{aligned} \int_{-\infty}^{\infty} g(x) \, dx \quad &\triangleq \quad \lim_{c\to\infty, d\to\infty} \int_{-c}^{d} g(x) \, dx \\ &= \lim_{c\to\infty} \int_{-c}^{0} g(x) \, dx + \lim_{d\to\infty} \int_{0}^{d} g(x) \, dx \end{aligned}$$

$$\text{This is not same as} \quad \lim_{a\to\infty} \int_{-a}^{a} g(x) \, dx,$$

which is known as Cauchy principal value

- Here we have

$$\lim_{c\to\infty} \int_{-c}^{0} \frac{x}{1+x^2} \, dx = -\infty; \quad \lim_{d\to\infty} \int_{0}^{d} \frac{x}{1+x^2} \, dx = \infty$$

- Hence $EX = \int_{-\infty}^{\infty} x \, \frac{1}{\pi} \, \frac{1}{1+x^2} \, dx$ does not exist.
- Essentially, both halves of the integral are infinite and hence we get $\infty - \infty$ type expression which is undefined.
- However, $\lim_{a\to\infty} \int_{-a}^{a} x \, \frac{1}{\pi} \, \frac{1}{1+x^2} \, dx = 0$.

Expectation of a random variable

- Let X be a discrete rv with $X \in \{x_1, x_2, \cdots\}$. Then
$$E[X] = \sum_i x_i \; f_X(x_i)$$
- If X is a continuous random variable with pdf, f_X,
$$E[X] = \int_{-\infty}^{\infty} x \; f_X(x) \; dx$$
- Sometimes we use the following notation to denote expectation of both kinds of rv
$$E[X] = \int_{-\infty}^{\infty} x \; dF_X(x)$$
- We take the expectation to exist when the sum or integral above is absolutely convergent
- Note that expectation is defined for all random variables
- Let us calculate expectations of some of the standard distributions.

Binary random variable

- Expectation of a binary rv (e.g., Bernoulli):
$$EX = 0 \times f_X(0) + 1 \times f_X(1) = P[X = 1]$$
- Expectation of a binary random variable is same as the probability of the rv taking value 1.
- Thus, for example, $EI_A = P(A)$.

Expectation of Binomial rv

- Let $f_X(k) = \; ^nC_k p^k (1-p)^{n-k}, \; k = 0, 1, \cdots, n$.

$$EX = \sum_{k=0}^{n} k \; \frac{n!}{k!(n-k)!} \; p^k \; (1-p)^{n-k} = \sum_{k=1}^{n} k \; \frac{n!}{k!(n-k)!} \; p^k \; (1-p)^{n-k}$$
$$= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} \; p^k \; (1-p)^{n-k}$$
$$= \sum_{k=1}^{n} \frac{n(n-1)!}{(k-1)!((n-1)-(k-1))!} \; p \; p^{k-1} \; (1-p)^{(n-1)-(k-1)}$$
$$= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} \; p^{k-1} \; (1-p)^{(n-1)-(k-1)}$$
$$= np \sum_{k'=0}^{n-1} \frac{(n-1)!}{k'!((n-1)-k')!} \; p^{k'} \; (1-p)^{(n-1)-k'} = np$$

Expectation of Poisson rv

- $f_X(k) = \frac{\lambda^k}{k!} \; e^{-\lambda}, \; k = 0, 1, \cdots$
$$\begin{aligned} EX &= \sum_{k=0}^{\infty} k \; \frac{\lambda^k}{k!} \; e^{-\lambda} \\ &= \lambda \end{aligned}$$

(Left as an exercise for you!)

Expectation of Geometric rv

- $f_X(k) = (1-p)^{k-1}\, p, \quad k = 1, 2, \cdots$

$$EX = \sum_{k=1}^{\infty} k\ (1-p)^{k-1}\ p$$

- We have

$$\sum_{k=1}^{\infty} (1-p)^k = \frac{1-p}{p} = \frac{1}{p} - 1$$

- Term-wise differentiation of the above gives

$$\sum_{k=1}^{\infty} k\ (1-p)^{k-1} = \frac{1}{p^2}$$

- This gives us $EX = \frac{1}{p}$

Expectation of uniform density

- Let $X \sim U[a,b]$. $f_X(x) = \frac{1}{b-a},\ a \le x \le b$

$$\begin{aligned} EX &= \int_{-\infty}^{\infty} x\ f_X(x)\ dx \\ &= \int_a^b x\ \frac{1}{b-a}\ dx \\ &= \frac{1}{b-a}\ \left.\frac{x^2}{2}\right|_a^b \\ &= \frac{1}{b-a}\ \frac{b^2-a^2}{2} \\ &= \frac{b+a}{2} \end{aligned}$$

Expectation of exponential density

- $f_X(x) = \lambda\ e^{-\lambda x}, \quad x > 0.$

$$\begin{aligned} EX &= \int_0^{\infty} x\ \lambda\ e^{-\lambda x}\ dx \\ &= x\ \lambda\ \left.\frac{e^{-\lambda x}}{-\lambda}\right|_0^{\infty} - \int_0^{\infty} \lambda\ \frac{e^{-\lambda x}}{-\lambda}\ dx \\ &= \int_0^{\infty} e^{-\lambda x}\ dx \\ &= \left.\frac{e^{-\lambda x}}{-\lambda}\right|_0^{\infty} \\ &= \frac{1}{\lambda} \end{aligned}$$

Expectation of Gaussian density

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

$$EX = \int_{-\infty}^{\infty} x\ \frac{1}{\sigma\sqrt{2\pi}}\ e^{-\frac{(x-\mu)^2}{2\sigma^2}}\ dx$$

make a change of variable $y = \frac{x-\mu}{\sigma}$

$$\begin{aligned} &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\ (\sigma y + \mu) e^{-\frac{y^2}{2}}\ dy \\ &= \sigma \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y\ e^{-\frac{y^2}{2}}\ dy + \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\ dy \\ &= \mu \end{aligned}$$

Expectation of a function of a random variable

- Let X be a rv and let $Y = g(X)$.
- **Theorem**: $EY = \int y \, dF_Y(y) = \int g(x) \, dF_X(x)$
- That is, if X is discrete, then

$$EY = \sum_j y_j \; f_Y(y_j) = \sum_i g(x_i) f_X(x_i)$$

- If X and Y are continuous

$$EY = \int y \; f_Y(y) \; dy = \int g(x) \; f_X(x) \; dx$$

- This theorem is true for all rv's. But we will prove it in only some special cases.

- **Theorem**: Let $X \in \{x_1, x_2, \cdots x_n\}$ and let $Y = g(X)$. Then

$$EY = \sum_i g(x_i) \; f_X(x_i)$$

- **Proof**: Let $Y \in \{y_1, y_2, \cdots, y_m\}$. Each y_j would be equal to $g(x_i)$ for one or more i.
- Let $B_j = \{x_i \; : \; g(x_i) = y_j\}$. Thus,

$$f_Y(y_j) = P[Y = y_j] = P[X \in B_j] = \sum_{\substack{i: \\ x_i \in B_j}} f_X(x_i)$$

- Note that
  - B_j are disjoint
  - each x_i would be in one (and only one) of the B_j

- Now we have

$$\begin{aligned} EY &= \sum_{j=1}^{m} y_j f_Y(y_j) \\ &= \sum_{j=1}^{m} y_j \sum_{\substack{i: \\ x_i \in B_j}} f_X(x_i) \\ &= \sum_{j=1}^{m} \sum_{\substack{i: \\ x_i \in B_j}} g(x_i) f_X(x_i) \\ &= \sum_{i=1}^{n} g(x_i) f_X(x_i) \end{aligned}$$

  That completes the proof.
- The proof goes through even when X (and Y) take countably infinitely many values (because we assume the expectation sum is absolutely convergent).

- Suppose X is a continuous rv and suppose g is a differentiable function with $g'(x) > 0, \; \forall x$. Let $Y = g(X)$
- Once again we can show $EY = \int g(x) \; f_X(x) \; dx$

$$\begin{aligned} EY &= \int_{-\infty}^{\infty} y \; f_Y(y) \; dy \\ &= \int_{g(-\infty)}^{g(\infty)} y \; f_X(g^{-1}(y)) \; \frac{d}{dy} g^{-1}(y) \; dy, \end{aligned}$$

change the variable to $x = g^{-1}(y) \quad \Rightarrow \; dx = \frac{d}{dy} g^{-1}(y) \; dy$

$$= \int_{-\infty}^{\infty} g(x) \; f_X(x) \; dx$$

- We can similarly show this for the case where $g'(x) < 0, \; \forall x$

- We proved the theorem only for discrete rv's and for some restricted case of continuous rv's.
- However, this theorem is true for all random variables.
- Now, for any function, g, we can write

$$E[g(X)] = \sum_i g(x_i) f_X(x_i) \quad \text{or} \quad E[g(X)] = \int_{-\infty}^{\infty} g(x)\ f_X(x)\ d$$

Some Properties of Expectation

$$E[g(X)] = \sum_i g(x_i) f_X(x_i) \quad \text{or} \quad E[g(X)] = \int_{-\infty}^{\infty} g(x)\ f_X(x)\ dx$$

- If $X \geq 0$ then $EX \geq 0$
- $E[b] = b$ where b is a constant
- $E[ag(X)] = aE[g(X)]$ where a is a constant
- $E[aX + b] = aE[X] + b$ where a, b are constants.
- $E[ag_1(X) + bg_2(X)] = aE[g_1(X)] + bE[g_2(X)]$

- Consider the problem: $\min_c \ E[(X - c)^2]$
- We are asking what is the best constant to approximate a rv with
- We are trying to minimize (weighted) average, over all values X can take, of the square of the error
- We are interested in the best mean-square approximation of X by a constant.

$$E[(X - c)^2] = E[X^2 + c^2 - 2cX] = E[X^2] + c^2 - 2cE[X]$$

- We differentiate this and equate to zero to get the best c

$$2c^* = 2E[X] \ \Rightarrow \ c^* = E[X]$$

- We can derive this in an alternate manner too

$$\begin{aligned} E[(X - c)^2] &= E[(X - EX + EX - c)^2] \\ &= E[(X - EX)^2 + (EX - c)^2 + 2(EX - c)(X - EX)] \\ &= E[(X - EX)^2] + (EX - c)^2 + 2(EX - c)E[(X - EX)] \\ &= E[(X - EX)^2] + (EX - c)^2 + 2(EX - c)(EX - EX) \\ &= E[(X - EX)^2] + (EX - c)^2 \\ &\geq E[(X - EX)^2] \end{aligned}$$

- Thus $E[(X - c)^2] \geq E[(X - EX)^2], \ \forall c$
- So, $E[(X - c)^2]$ is minimized when $c = EX$ and the minimum value is $E[(X - EX)^2]$

Variance of a Random variable

- We define variance of X as $E[(X-EX)^2]$ and denote it as Var(X).
- By definition, Var$(X) \geq 0$.

$$\begin{aligned}\text{Var}(X) &= E[(X-EX)^2] \\ &= E\left[X^2 + (EX)^2 - 2X(EX)\right] \\ &= E[X^2] + (EX)^2 - 2(EX)E[X] \\ &= E[X^2] - (EX)^2\end{aligned}$$

- This also implies: $E[X^2] \geq (EX)^2$

Some properties of variance

- Var$(X+c) =$ Var(X) where c is a constant

$$\text{Var}(X+c) = E\left[\{(X+c) - E[X+c]\}^2\right] = E\left[(X-EX)^2\right] = \text{Var}(X)$$

- Var$(cX) = c^2$Var(X) where c is a constant

$$\text{Var}(cX) = E\left[(cX - E[cX])^2\right] = E\left[(cX - cE[X])^2\right] = c^2\text{Var}(X)$$

Variance of uniform rv

- $f_X(x) = \frac{1}{b-a}, \ a \leq x \leq b$

$$\begin{aligned}E[X^2] &= \int_a^b x^2 \, \frac{1}{b-a} \, dx \\ &= \frac{1}{b-a} \left. \frac{x^3}{3} \right|_a^b \\ &= \frac{1}{b-a} \, \frac{b^3 - a^3}{3} \\ &= \frac{b^2 + ab + a^2}{3}\end{aligned}$$

Variance of uniform rv

- We got $E[X^2] = \frac{b^2+ab+a^2}{3}$. Earlier we showed $EX = \frac{b+a}{2}$
- Now we can calculate Var(X) as

$$\begin{aligned}\text{Var}(X) &= EX^2 - (EX)^2 \\ &= \frac{b^2+ab+a^2}{3} - \frac{(b+a)^2}{4} \\ &= \frac{4(b^2+ab+a^2) - 3(b^2+2ab+a^2)}{12} \\ &= \frac{(b^2 - 2ab + a^2)}{12} \\ &= \frac{(b-a)^2}{12}\end{aligned}$$

Variance of exponential rv

- $f_X(x) = \lambda\, e^{-\lambda x},\ x > 0$

$$\begin{aligned} E[X^2] &= \int_0^\infty x^2\, \lambda\, e^{-\lambda x}\, dx \\ &= x^2\, \lambda\, \frac{e^{-\lambda x}}{-\lambda}\bigg|_0^\infty - \int_0^\infty \lambda\, \frac{e^{-\lambda x}}{-\lambda}\, 2x\, dx \\ &= \frac{2}{\lambda} \int_0^\infty x\, \lambda\, e^{-\lambda x}\, dx \\ &= \frac{2}{\lambda^2} \end{aligned}$$

- Hence the variance is now given by

$$\text{Var}(X) = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

Variance of Gaussian rv

- Let $X \sim \mathcal{N}(0,1)$. That is, $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},\ -\infty < x < \infty.$
- We know $EX = 0$. Hence Var$(X) = EX^2$.

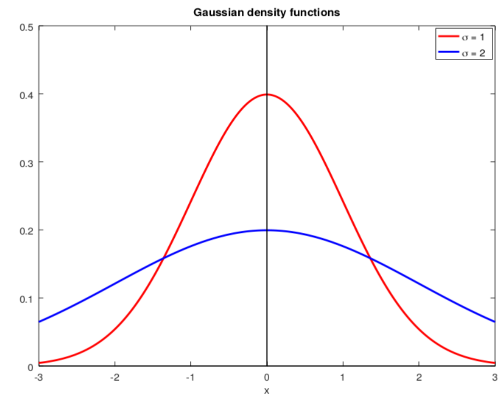
$$\begin{aligned} \text{Var}(X) &= EX^2 = \int_{-\infty}^\infty x^2\, \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx \\ &= \int_{-\infty}^\infty x \left(x\, \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \right) dx \\ &= x\, \frac{-1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \bigg|_{-\infty}^\infty + \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx \\ &= 1 \end{aligned}$$

- Let $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},\ -\infty < x < \infty.$
- Let $g(x) = \sigma x + \mu$ and hence $g^{-1}(y) = \frac{y-\mu}{\sigma}$.
- Take $\sigma > 0$ and $Y = g(X)$. By the theorem,

$$f_Y(y) = \left(\frac{d}{dy} g^{-1}(y)\right)\ f_X(g^{-1}(y)) = \frac{1}{\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

- Since $Y = \sigma X + \mu$, we get
  - $EY = \sigma EX + \mu = \mu$
  - Var$(Y) = \sigma^2$Var$(X) = \sigma^2$
- When $Y \sim \mathcal{N}(\mu, \sigma^2)$, $EY = \mu$ and Var$(Y) = \sigma^2$.

- Here is a plot of Gaussian densities with different variances



Variance of Binomial rv

- $f_X(k) = \frac{n!}{k!(n-k)!}\, p^k\, (1-p)^{n-k},\ k = 0, 1, \cdots, n$
- Here we use the identity, $EX^2 = E[X(X-1)] + EX$

$$\begin{aligned} E[X(X-1)] &= \sum_{k=0}^{n} k(k-1)\, \frac{n!}{k!(n-k)!}\, p^k\, (1-p)^{n-k} \\ &= \sum_{k=2}^{n} k(k-1)\, \frac{n!}{k!(n-k)!}\, p^k\, (1-p)^{n-k} \end{aligned}$$

$$= \sum_{k=2}^{n} \frac{n(n-1)(n-2)!}{(k-2)!((n-2)-(k-2))!}\, p^2\, p^{k-2}\, (1-p)^{(n-2)-(k-2)}$$

$$= n(n-1)p^2 \sum_{k'=0}^{n-2} \frac{(n-2)!}{k'!((n-2)-k')!}\, p^{k'}\, (1-p)^{(n-2)-k'}$$

$$= n(n-1)p^2$$

- When X is binomial rv, we showed, $E[X(X-1)] = n(n-1)p^2$
- Hence,

$$EX^2 = E[X(X-1)] + EX = n(n-1)p^2 + np = n^2p^2 + np(1-p)$$

- Now we can calculate the variance

$$\text{Var(X)} = EX^2 - (EX)^2 = n^2p^2 + np(1-p) - (np)^2 = np(1-p)$$

Variance of a geometric random variable

- $X \in \{1, 2, \cdots\}$ and $f_X(k) = (1-p)^{k-1}p,\ k = 1, 2, \cdots$
- Here also, it is easier to calculate $E[X(X-1)]$

$$E[X(X-1)] = \sum_{k=1}^{\infty} k(k-1)(1-p)^{k-1}p = p(1-p)\sum_{k=1}^{\infty} k(k-1)(1-p)^{k-2}$$

- We know

$$\sum_{k=1}^{\infty}(1-p)^k = \frac{1-p}{p} \;\Rightarrow\; \sum_{k=1}^{\infty} k(k-1)(1-p)^{k-2} = \frac{d^2}{dp^2}\left(\frac{1-p}{p}\right)$$

Now you can compute $E[X(X-1)]$ and hence $E[X^2]$ and hence Var(X) and show it to be equal to $\frac{1-p}{p^2}$.
(Left as an exercise)

moments of a random variable

- We define the k^{th} order moment of a rv, X, by

$$m_k = E[X^k] = \int x^k\, dF_X(x)$$

- $m_1 = EX$ and $m_2 = EX^2$ and so on
- We define the k^{th} central moment of X by

$$s_k = E[(X - EX)^k] = \int (x - EX)^k\, dF_X(x)$$

- $s_1 = 0$ and $s_2 = \text{Var}(X)$.
- Not all moments may exist for a given random variable. (For example, m_1 does not exist for Cauchy rv)

- **Theorem**: If $E\left[|X|^k\right] < \infty$ then $E\left[|X|^s\right] < \infty$ for $0 < s < k$.
- For example, if third order moment exists then so do first and second order moments
- **Proof**: We prove it when X is continuous rv. Proof for discrete case is similar.

$$
\begin{aligned}
E\left[|X|^s\right] &= \int_{-\infty}^{\infty} |x|^s \, f_X(x) \, dx \\
&= \int_{|x|<1} |x|^s \, f_X(x) \, dx + \int_{|x|\geq 1} |x|^s \, f_X(x) \, dx \\
&\leq \int_{|x|<1} f_X(x) \, dx + \int_{|x|\geq 1} |x|^s \, f_X(x) \\
&\leq P[|X|^s < 1] + \int_{|x|\geq 1} |x|^k \, f_X(x) \\
&\quad \text{since for } |x| \geq 1,\ |x|^s < |x|^k \text{ when } s < k \\
&< \infty \ \text{ because } E\left[|X|^k\right] = \int_{\infty}^{\infty} |x|^k f_X(x) \, dx < \infty
\end{aligned}
$$

Recap: Expectation

- Let X be a discrete rv with $X \in \{x_1, x_2, \cdots\}$. Then

$$E[X] = \sum_i x_i \, f_X(x_i)$$

- If X is a continuous random variable with pdf, f_X,

$$E[X] = \int_{-\infty}^{\infty} x \, f_X(x) \, dx$$

- Sometimes we use the following notation to denote expectation of both kinds of rv

$$E[X] = \int_{-\infty}^{\infty} x \, dF_X(x)$$

- We take the expectation to exist when the sum or integral above is absolutely convergent
- Note that expectation is defined for all random variables

Recap: Expectation of a function of a random variable

- Let X be a rv and let $Y = g(X)$. Then,
- $EY = \int y \, dF_Y(y) = \int g(x) \, dF_X(x)$
- That is, if X is discrete, then

$$EY = \sum_j y_j \, f_Y(y_j) = \sum_i g(x_i) f_X(x_i)$$

- If X and Y are continuous

$$EY = \int y \, f_Y(y) \, dy = \int g(x) \, f_X(x) \, dx$$

- This is true for all rv's.

Recap: Properties of Expectation

$$E[g(X)] = \sum_i g(x_i) f_X(x_i) \quad \text{or} \quad E[g(X)] = \int_{-\infty}^{\infty} g(x) \, f_X(x) \, dx$$

- If $X \geq 0$ then $EX \geq 0$
- $E[b] = b$ where b is a constant
- $E[ag(X)] = aE[g(X)]$ where a is a constant
- $E[aX + b] = aE[X] + b$ where a, b are constants.
- $E[ag_1(X) + bg_2(X)] = aE[g_1(X)] + bE[g_2(X)]$
- $E\left[(X - c)^2\right] \geq E\left[(X - EX)^2\right], \ \forall c$

Recap: Variance of random variable

$$\text{Var}(X) = E\left[(X - EX)^2\right] = E[X^2] - (EX)^2$$

- Properties of Variance:
  - $\text{Var}(X) \geq 0$
  - $\text{Var}(X + c) = \text{Var}(X)$
  - $\text{Var}(cX) = c^2\text{Var}(X)$

Recap: Moments of a random variable

- The k^{th} (order) moment of X is

$$m_k = E[X^k] = \int x^k \, dF_X(x)$$

- The k^{th} central moment of X is

$$s_k = E[(X - EX)^k] = \int (x - EX)^k \, dF_X(x)$$

- If moment of order k is finite then so is moment of order s for $s < k$.

Moment generating function

- The moment generating function (mgf) of rv X, $M_X : \Re \to \Re$, is defined by

$$M_X(t) = Ee^{tX} = \sum_i e^{tx_i} f_X(x_i) \quad \text{or} \quad \int e^{tx} f_X(x) \, dx, \quad t \in \Re$$

- We say the mgf exists if $E[e^{tX}] < \infty$ for t in some interval around zero
- The mgf may not exist for some random variables.

- The mgf of X is: $M_X(t) = E[e^{tX}]$.
- If $M_X(t)$ exists (for $t \in [-a, \, a]$ for some $a > 0$) then all its derivatives also exist.
- Then we can get the moments of X by successive differentiation of $M_X(t)$.

$$\left.\frac{dM_X(t)}{dt}\right|_{t=0} = \left.\frac{d}{dt}E\left[e^{tX}\right]\right|_{t=0} = E[Xe^{tX}]\big|_{t=0} = EX$$

- In general

$$\left.\frac{d^k M_X(t)}{dt^k}\right|_{t=0} = E[X^k]$$

- We can easily see this by expanding e^{tX} in Taylor series:

$$M_X(t) = Ee^{tX} = E\left[1 + \frac{tX}{1!} + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \frac{t^4X^4}{4!} + \cdots\right.$$
$$= 1 + \frac{t}{1!}EX + \frac{t^2}{2!}EX^2 + \frac{t^3}{3!}EX^3 + \frac{t^4}{4!}EX^4 + \cdots$$

- Now we can do term-wise differentiation. For example

$$\frac{d^3M_X(t)}{dt^3} = 0+0+0+\frac{3*2*1*t^0}{3!}EX^3+\frac{4*3*2*t}{4!}EX^4+\cdots$$

- Hence we get

$$\left.\frac{d^3M_X(t)}{dt^3}\right|_{t=0} = E[X^3]$$

Example – Moment generating function for Poisson

- $f_X(k) = \frac{\lambda^k}{k!}e^{-\lambda}, \; k = 0, 1, \cdots$

$$M_X(t) = E[e^{tX}] = \sum_{k=0}^{\infty} e^{tk}\,\frac{\lambda^k}{k!}\,e^{-\lambda}$$
$$= e^{-\lambda}\sum_{k=0}^{\infty}\frac{1}{k!}\left(\lambda e^t\right)^k$$
$$= e^{-\lambda}\,e^{\lambda e^t} = e^{\lambda(e^t-1)}$$

- Now, by differentiating it we can find EX

$$EX = \left.\frac{dM_X(t)}{dt}\right|_{t=0} = \left.e^{\lambda(e^t-1)}\;\lambda e^t\right|_{t=0} = \lambda$$

(Exercise: Differentiate it twice to find EX^2 and hence show that variance is λ).

mgf of exponential rv

- $f_X(x) = \lambda e^{-\lambda x}, \; x > 0$

$$M_X(t) = E[e^{tX}] = \int_0^\infty e^{tx}\,\lambda e^{-\lambda x}\,dx$$
$$= \int_0^\infty \lambda e^{-x(\lambda-t)}\,dx$$

This is finite if $t < \lambda$

$$= \left.\frac{\lambda e^{-x(\lambda-t)}}{-(\lambda-t)}\right|_0^\infty$$
$$= \frac{\lambda}{\lambda-t}, \quad t<\lambda$$

- We can use this to compute EX

$$EX = \left.\frac{dM_X(t)}{dt}\right|_{t=0} = \left.\frac{d}{dt}\left(\frac{\lambda}{\lambda-t}\right)\right|_{t=0} = \left.\frac{\lambda}{(\lambda-t)^2}\right|_{t=0} = \frac{1}{\lambda}$$

- For mgf to exist we need $E[e^{tX}] < \infty$ for $t \in [-a, \; a]$ for some $a > 0$.
- If $M_X(t)$ exists then all moments of X are finite.
- However, all moments may be finite but the mgf may not exist.
- When mgf exists, it uniquely determines the df
- We are not saying moments uniquely determine the distribution; we are saying mgf uniquely determines the distribution

Characteristic Function

- The characteristic function of X is defined by

$$\phi_X(t) = E[e^{itX}] = \int e^{itx} \, dF_X(x) \quad (i = \sqrt{-1})$$

- If X is continuous rv,

$$\phi_X(t) = E[e^{itX}] = \int_{-\infty}^{\infty} e^{itx} \, f_X(x) \, dx$$

- Characteristic function always exists because $|e^{itx}| = 1, \forall t, x$
- For example,

$$\left| \int_{-\infty}^{\infty} e^{itx} \, f_X(x) \, dx \right| \le \int_{-\infty}^{\infty} |e^{itx}| \, |f_X(x)| \, dx = \int_{-\infty}^{\infty} f_X(x) \, dx = 1$$

- We would consider ϕ_X later in the course

Generating function

- Let $X \in \{0, 1, 2, \cdots\}$
- The (probability) generating function of X is defined by

$$P_X(s) = \sum_{k=0}^{\infty} f_X(k) s^k, \quad s \in \Re$$

- This infinite sum converges (absolutely) for $|s| \le 1$.
- We have

$$P_X(s) = f_X(0) + f_X(1)s + f_X(2)s^2 + f_X(3)s^3 + \cdots$$

- The pmf can be obtained from the generating function

- $P_X(s) = f_X(0) + f_X(1)s + f_X(2)s^2 + f_X(3)s^3 + \cdots$
- Let $P'_X(s) \triangleq \frac{dP_X(s)}{ds}$ and so on
- We get

$$P'_X(s) = 0 + f_X(1) + f_X(2) \, 2s + f_X(3) \, 3s^2 + \cdots$$

$$P''_X(s) = 0 + 0 + f_X(2) \, 2 * 1 + f_X(3) \, 3 * 2s^1 + \cdots$$

Hence, we get

$$f_X(0) = P_X(0); \; f_X(1) = \frac{P'_X(0)}{1!}; \; f_X(2) = \frac{P''_X(0)}{2!}$$

- The moments (when they exist) can be obtained from the generating function: $P_X(s) = \sum_{k=0}^{\infty} f_X(k)s^k$

$$P'_X(s) = \sum_{k=0}^{\infty} k \; f_X(k) \; s^{k-1} \;\Rightarrow\; P'_X(1) = EX$$

$$P''_X(s) = \sum_{k=0}^{\infty} k(k-1) \; f_X(k) \; s^{k-2} \;\Rightarrow\; P''_X(1) = E[X(X-1)]$$

- For (positive integer valued) discrete random variables, it is more convenient to deal with generating functions than mgf.

Example – Generating function for binomial rv

- $f_X(k) = \frac{n!}{k!(n-k)!}\, p^k\, (1-p)^{n-k},\ k = 0, 1, \cdots, n$

$$\begin{aligned} P_X(s) &= \sum_{k=0}^{n} \frac{n!}{k!(n-k)!}\, p^k\, (1-p)^{n-k}\, s^k \\ &= \sum_{k=0}^{n} \frac{n!}{k!(n-k)!}\, (sp)^k\, (1-p)^{n-k} \\ &= (sp + (1-p))^n = (1 + p(s-1))^n \end{aligned}$$

- From the above, we get $P'_X(s) = n(sp + (1-p))^{n-1} p$
- Thus,

$$EX = P'_X(1) = np; \quad f_X(1) = P'_X(0) = n(1-p)^{n-1}p$$

- Let $p \in (0,\ 1)$. The number $x \in \Re$ that satisfies

$$P[X \le x] \ge p \quad \text{and} \quad P[X \ge x] \ge 1 - p$$

is called the quantile of order p or the $100p^{th}$ percentile of rv X.
- Suppose x is a quantile of order p. Then we have
  - $p \le P[X \le x] = F_X(x)$
  - $1 - p \le 1 - P[X < x] = 1 - (P[X \le x] - P[X = x])$
    $\Rightarrow\ 1 - p \le 1 - F_X(x) + P[X = x]$
    $\Rightarrow F_X(x) \le p + P[X = x]$
- Thus, x satisfies (if it is quantile of order p)

$$p \le F_X(x) \le p + P[X = x]$$

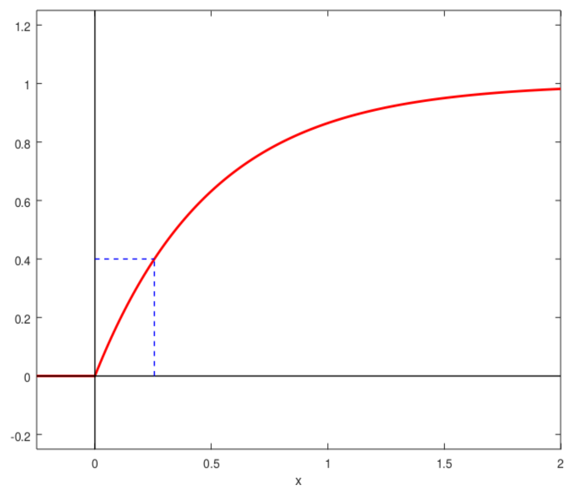
- Note that for a given p there can be multiple values for x to satisfy the above.

- If x is a quantile of order p then

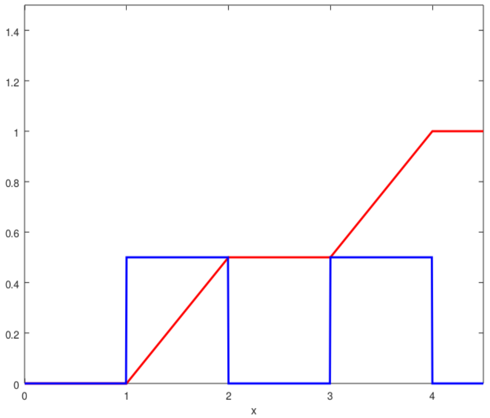
$$p \le F_X(x) \le p + P[X = x]$$

- If X is continuous rv, we need to satisfy $p = F_X(x)$.
- In general, for a given p, there may be multiple x that satisfy the above.
- Let us see some examples.

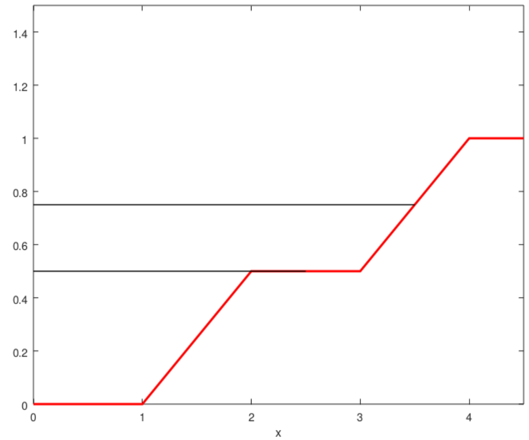
- Let X be continuous rv.
- If the df is strictly monotone then $F_X(x) = p$ would have a unique solution.



- For continuous rv, X, F_X need not be strictly monotone.
- Consider a pdf: $f_X(x) = 0.5, \; x \in [1, \; 2] \cup [3, \; 4]$
- The pdf and the corresponding df are:

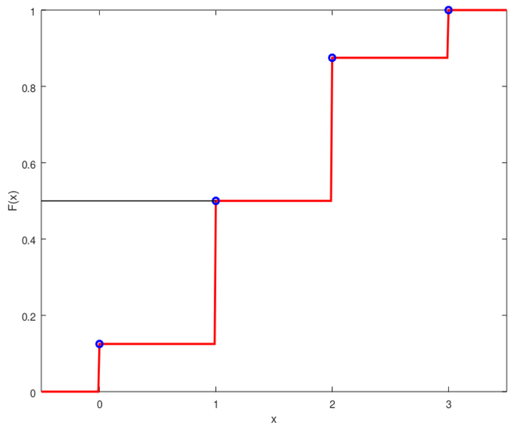


- For this df, for $p = 0.5$, the quantile of order p is not unique because there many x with $F_X(x) = 0.5$
  But for $p = 0.75$ it is unique.



- Let $X \in \{x_1, x_2, \cdots\}$
- Given a p we want to calculate quantile of order p
- Suppose there is a x_i such that $F_X(x_i) = p$.
- Then, for $x_i \leq x < x_{i+1}$, $F_X(x) = p$
- For $x_i \leq x \leq x_{i+1}$, we have $p \leq F_X(x) \leq p + P[X = x]$
- So, quantile of order p is not unique and all such x qualify.

- This situation is illustrated below

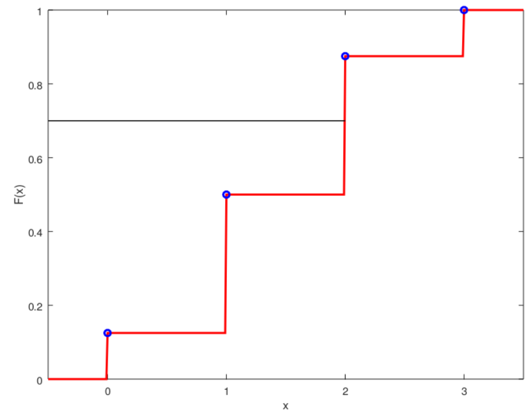


- Now suppose p is such that $F_X(x_{i-1}) < p < F_X(x_i)$.
- Let $F_X(x_{i-1}) = p - \delta_1$ and $F_X(x_i) = p + \delta_2$. (Note that $\delta_1, \delta_2 > 0$)
- Then $P[X = x_i] = F_X(x_i) - F_X(x_{i-1}) = \delta_2 + \delta_1$
- Hence we have

$$p < p + \delta_2 = F_X(x_i) < p + \delta_2 + \delta_1 = p + P[X = x_i]$$

- Hence, x_i is quantile of order p.
- For any $x < x_i$ we would have $F_X(x) \leq F_X(x_{i-1}) < p$.
- For any x, with $x_i < x < x_{i+1}$ we have $p + P[X = x] = p < F_X(x) = p + \delta_2$.
- Similarly, for $x \geq x_{i+1}$ we have $F_X(x) > p + P[X = x]$.
- Thus quantile of order p is unique here.

- This situation is illustrated below



Median of a distribution

- For $p = 0.5$ quantile of order p is called the median.
- For a continuous rv, median, x satisfies: $F_X(x) = 0.5$.
- For a discrete rv, it satisfies: $0.5 \leq F_X(x) \leq 0.5 + P[X = x]$.
- As we saw, median need not be unique.
- Recall that the (standard) Cauchy density is given by

$$f_X(x) = \frac{1}{\pi} \frac{1}{1 + x^2}, \ -\infty < x < \infty$$

- One can show that $\int_{-\infty}^{0} f_X(x)\, dx = 0.5$ and hence the median is at the origin.

- If we want to find c to minimize $E\left[(X - c)^2\right]$ then the solution is $c = EX$.
- We saw this earlier.
- Suppose we want to find c to minimize $E\left[|(X - c)|\right]$
- Then we would get c to be the median. (Exercise: Show this for discrete and continuous rv)

Markov Inequality

- Let $g : \Re \to \Re$ be a non-negative function. Then

$$P[g(X) > c] \leq \frac{E[g(X)]}{c}, \quad (c > 0)$$

- **Proof**: We prove it for continuous rv. Proof is similar for discrete rv

$$\begin{aligned} E[g(X)] &= \int_{-\infty}^{\infty} g(x)\ f_X(x)\ dx \\ &= \int_{g(x) \leq c} g(x)\ f_X(x)\ dx + \int_{g(x) > c} g(x)\ f_X(x)\ dx \\ &\geq \int_{g(x) > c} g(x)\ f_X(x)\ dx \quad \text{because } g(x) \geq 0 \\ &\geq c \int_{g(x) > c} f_X(x)\ dx = c\ P[g(X) > c] \end{aligned}$$

Thus, $P[g(X) > c] \leq \frac{E[g(X)]}{c}$

Markov Inequality

$$P[g(X) > c] \leq \frac{E[g(X)]}{c}, \quad (c > 0)$$

- In all such results an underlying assumption is that the expectation is finite.
- Let $g(x) = |x|^k$ where k is a positive integer. We have $g(x) \geq 0,\ \forall x$. Let $c > 0$.
- We know that $|x| > c \Rightarrow |x|^k > c^k$ and vice versa.
- Now we get,

$$P[|X| > c] = P[|X|^k > c^k] \leq \frac{E\left[|X|^k\right]}{c^k}$$

- Markov inequality is often used in this form.

Chebyshev Inequality

- Markov Inequality:

$$P[|X| > c] \leq \frac{E\left[|X|^k\right]}{c^k}$$

- Take $|X|$ as $|X - EX|$ and take $k = 2$

$$P[|X - EX| > c] \leq \frac{E\left[|X - EX|^2\right]}{c^2} = \frac{\text{Var}(X)}{c^2}$$

- This is known as the Chebyshev inequality.

- The Chebyshev inequality is

$$P[|X - EX| > c] \leq \frac{\text{Var}(X)}{c^2}$$

- Let $EX = \mu$ and let $\text{Var}(X) = \sigma^2$. Take $c = k\sigma$
- We call, σ, square root of variance, as standard deviation.
- Now, Chebyshev inequality gives us

$$P[|X - \mu| > k\sigma] \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

- This is true for all random variables and the RHS above does not depend on the distribution of X.

- **Markov inequality**: For a non-negative function, g,

$$P[g(X) > c] \leq \frac{E[g(X)]}{c}$$

- A specific instance of this is

$$P[|X| > c] \leq \frac{E\left[|X|^k\right]}{c^k}$$

- **Chebyshev inequality**

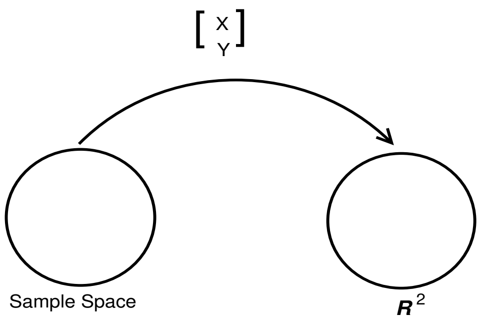
$$P[|X - EX| > c] \leq \frac{\text{Var}(X)}{c^2}$$

- With $EX = \mu$ and Var$(X) = \sigma^2$, we get

$$P[|X - \mu| > k\sigma] \leq \frac{1}{k^2}$$

A pair of random variables

- Let X, Y be random variables on the same probability space $(\Omega, \mathcal{F}, P)$
- Each of X, Y maps Ω to $\Re$.
- We can think of the pair of radom variables as a vector-valued function that maps Ω to $\Re^2$.

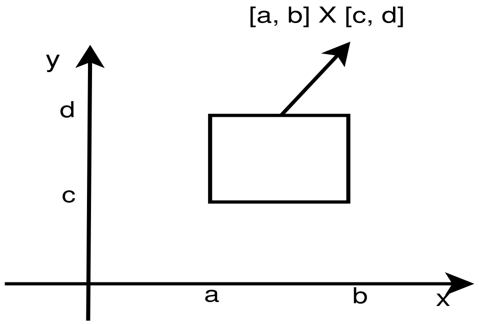


- Just as in the case of a single rv, we can think of the induced probability space for the case of a pair of rv's too.
- That is, by defining the pair of random variables, we essentially create a new probability space with sample space being $\Re^2$.
- The events now would be the Borel subsets of $\Re^2$.
- Recall that $\Re^2$ is cartesian product of $\Re$ with itself.
- So, we can create Borel subsets of $\Re^2$ by cartesian product of Borel subsets of $\Re$.

$$\mathcal{B}^2 = \sigma\left(\{B_1 \times B_2 \ : \ B_1, B_2 \in \mathcal{B}\}\right)$$

where $\mathcal{B}$ is the Borel σ-algebra we considered earlier, and $\mathcal{B}^2$ is the set of Borel sets of $\Re^2$.

- Recall that $\mathcal{B}$ is the smallest σ-algebra containing all intervals.
- Let $I_1, I_2 \subset \Re$ be intervals. Then $I_1 \times I_2 \subset \Re^2$ is known as a cylindrical set.



- $\mathcal{B}^2$ is the smallest σ-algebra containing all cylindrical sets.
- We saw that $\mathcal{B}$ is also the smallest σ-algebra containing all intervals of the form $(-\infty, x]$.
- Similarly $\mathcal{B}^2$ is the smallest σ-algebra containing cylindrical sets of the form $(-\infty, x] \times (-\infty, y]$.

- Let X, Y be random variables on the probability space $(\Omega, \mathcal{F}, P)$
- This gives rise to a new probability space $(\Re^2, \mathcal{B}^2, P_{XY})$ with P_{XY} given by

$$\begin{aligned} P_{XY}(B) &= P[(X,Y) \in B], \ \forall B \in \mathcal{B}^2 \\ &= P\left(\{\omega \ : \ (X(\omega).Y(\omega)) \in B\}\right) \end{aligned}$$

- Recall that for a single rv, the resulting probability space is $(\Re, \mathcal{B}, P_X)$ with

$$P_X(B) \ = \ P[X \in B] = P(\{\omega \ : \ X(\omega) \in B\})$$

- In the case of a single rv, we define a distribution function, F_X which essentailly assigns probability to all intervals of the form $(-\infty, \ x]$.
- This F_X uniquely determines $P_X(B)$ for all Borel sets, B.
- In a similar manner we define a joint distribution function F_{XY} for a pair of random varibles.
- $F_{XY}(x, y)$ would be $P_{XY}((-\infty, x] \times (-\infty, y])$.
- F_{XY} fixes the probability of all cylindrical sets of the form $(-\infty, x] \times (-\infty, y]$ and hence uniquely determines the probability of all Borel sets of $\Re^2$.

Joint distribution of a pair of random variables

- Let X, Y be random variables on the same probability space $(\Omega, \mathcal{F}, P)$
- The joint distribution function of X, Y is $F_{XY} : \Re^2 \to \Re$, defined by

$$\begin{aligned} F_{XY}(x,y) &= P[X \le x, Y \le y] \quad (= P_{XY}((-\infty, x] \times (-\infty, y])) \\ &= P\left(\{\omega \ : \ X(\omega) \le x\} \cap \{\omega \ : \ Y(\omega) \le y\}\right) \end{aligned}$$

- The joint distribution function is the probability of the intersection of the events $[X \le x]$ and $[Y \le y]$.

Properties of Joint Distribution Function

- Joint distribution function:

$$F_{XY}(x,y) = P[X \le x, Y \le y]$$

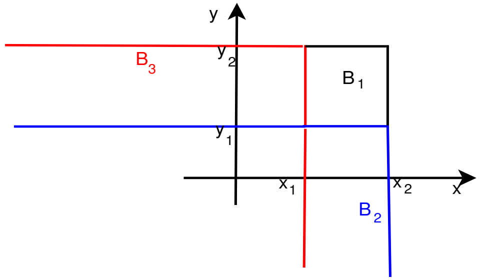
  - $F_{XY}(-\infty, y) = F_{XY}(x, -\infty) = 0, \forall x, y;$
    $F_{XY}(\infty, \infty) = 1$
    (These are actually limits: $\lim_{x \to -\infty} F_{XY}(x,y) = 0, \forall y$)
  - F_{XY} is non-decresing in each of its arguments
  - F_{XY} is right continuous and has left-hand limits in each of its arguments
- These are straight-forward extensions of single rv case
- But there is another crucial property satisfied by F_{XY}.

- Recall that, for the case of a single rv, the probability of X being in any interval is given by the difference of F_X values at the end points of the interval.
- Let $x_1 < x_2$. Then

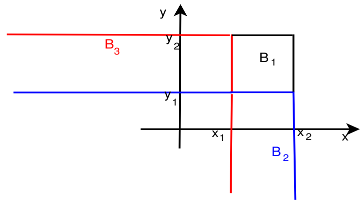
$$P[x_1 < X \le x_2] = F_X(x_2) - F_X(x_1)$$

- The LHS above is a probability.
  The RHS is non-negative because F_X is non-decreasing.
- We will now derive a similar expression in the case of two random variables.
- Here, the probability we want is that of the pair of rv's being in a cylindrical set.

- Let $x_1 < x_2$ and $y_1 < y_2$. We want $P[x_1 < X \le x_2, y_1 < Y \le y_2]$.
- Consider the Borel set $B = (-\infty, x_2] \times (-\infty, y_2]$.



$$\begin{aligned}
B &\triangleq (-\infty, x_2] \times (-\infty, y_2] = B_1 + (B_2 \cup B_3) \\
B_1 &= (x_1,\ x_2] \times (y_1,\ y_2] \\
B_2 &= (-\infty, x_2] \times (-\infty, y_1] \\
B_3 &= (-\infty, x_1] \times (-\infty, y_2] \\
B_2 \cap B_3 &= (-\infty, x_1] \times (-\infty, y_1]
\end{aligned}$$



$$\begin{aligned}
P[(X,Y) \in B] &= P[X \le x_2, Y \le y_2] = F_{XY}(x_2, y_2) \\
&= P[(X,Y) \in B_1 + (B_2 \cup B_3)] \\
&= P[(X,Y) \in B_1] + P[(X,Y) \in (B_2 \cup B_3)]
\end{aligned}$$

$$\begin{aligned}
P[(X,Y) \in B_2] &= P[X \le x_2, Y \le y_1] = F_{XY}(x_2, y_1) \\
P[(X,Y) \in B_3] &= P[X \le x_1, Y \le y_2] = F_{XY}(x_1, y_2) \\
P[(X,Y) \in B_2 \cap B_3] &= P[X \le x_1, Y \le y_1] = F_{XY}(x_1, y_1)
\end{aligned}$$

$$\begin{aligned}
P[(X,Y) \in B_1] &= F_{XY}(x_2, y_2) - P[(X,Y) \in (B_2 \cup B_3)] \\
&= F_{XY}(x_2, y_2) - F_{XY}(x_2, y_1) - F_{XY}(x_1, y_2) + F_{XY}(x_1, y_1)
\end{aligned}$$

- What we showed is the following.
- For $x_1 < x_2$ and $y_1 < y_2$

$$P[x_1 < X \leq x_2, y_1 < Y \leq y_2] = F_{XY}(x_2, y_2) - F_{XY}(x_2, y_1) - F_{XY}(x_1, y_2) + F_{XY}(x_1, y_1)$$

- This means F_{XY} should satisfy

$$F_{XY}(x_2, y_2) - F_{XY}(x_2, y_1) - F_{XY}(x_1, y_2) + F_{XY}(x_1, y_1) \geq 0$$

  for all $x_1 < x_2$ and $y_1 < y_2$
- This is an additional condition that a function has to satisfy to be the joint distribution function of a pair of random variables

Properties of Joint Distribution Function

- Joint distribution function: $F_{XY} : \Re^2 \to \Re$

$$F_{XY}(x, y) = P[X \leq x, Y \leq y]$$

- It satisfies
  1. $F_{XY}(-\infty, y) = F_{XY}(x, -\infty) = 0, \forall x, y;$ $F_{XY}(\infty, \infty) = 1$
  2. F_{XY} is non-decreasing in each of its arguments
  3. F_{XY} is right continuous and has left-hand limits in each of its arguments
  4. For all $x_1 < x_2$ and $y_1 < y_2$

$$F_{XY}(x_2, y_2) - F_{XY}(x_2, y_1) - F_{XY}(x_1, y_2) + F_{XY}(x_1, y_1) \geq 0$$

- Any $F : \Re^2 \to \Re$ satisfying the above would be a joint distribution function.

Recap: Random Variables

- Given a probability space $(\Omega, \mathcal{F}, P)$, a random variable is a real-valued function on Ω.
- It essentially results in an induced probability space

$$(\Omega, \mathcal{F}, P) \xrightarrow{X} (\Re, \mathcal{B}, P_X)$$

  where $\mathcal{B}$ is the Borel σ-algebra and

$$P_X(B) = P[X \in B] = P\left(\{\omega \in \Omega \ : \ X(\omega) \in B\}\right)$$

Recap: Distribution function of a random variable

- Let X be a random variable. It distribution function, $F_X : \Re \to \Re$, is defined by

$$F_X(x) = P[X \leq x] = P(\{\omega \in \Omega \ : \ X(\omega) \leq x\})$$

- The distribution function, F_X, completely specifies the probability measure, P_X.

Recap: Properties of distribution function

- The distribution function satisfies
  1. $0 \leq F_X(x) \leq 1, \ \forall x$
  2. $F_X(-\infty) = 0; \ F_X(\infty) = 1$
  3. F_X is non-decreasing: $x_1 \leq x_2 \ \Rightarrow \ F_X(x_1) \leq F_X(x_2)$
  4. F_X is right continuous and has left-hand limits.
- Any real-valued function of a real variable satisfying the above four properties would be a distribution function of some random variable.
- We also have
  $F_X(x^+) - F_X(x^-) = F_X(x) - F_X(x^-) = P[X = x]$
  $P[a < X \leq b] = F_X(b) - F_X(a)$.

Recap: Discrete Random Variable

- A random variable X is said to be discrete if it takes only finitely many or countably infinitely many distinct values.
- Let $X \in \{x_1, x_2, \cdots\}$
- Its distribution function, F_X is a stair-case function with jump discontinuities at each x_i and the magnitude of the jump at x_i is equal to $P[X = x_i]$

Recap: probability mass function

- Let $X \in \{x_1, x_2, \cdots\}$.
- The probability mass function (pmf) of X is defined by

$$f_X(x_i) = P[X = x_i]; \quad f_X(x) = 0, \quad \text{for all other } x$$

- It satisfies
  1. $f_X(x) \geq 0, \forall x$ and $f_X(x) = 0$ if $x \neq x_i$ for some i
  2. $\sum_i f_X(x_i) = 1$
- We have
  $F_X(x) = \sum_{i: x_i \leq x} f_X(x_i)$
  $f_X(x) = F_X(x) - F_X(x^-)$
- We can calculate the probability of any event as

$$P[X \in B] = \sum_{\substack{i: \\ x_i \in B}} f_X(x_i)$$

Recap: continuous random variable

- X is said to be a continuous random variable if there exists a function $f_X : \Re \to \Re$ satisfying

$$F_X(x) = \int_{-\infty}^{x} f_X(x) \, dx$$

  The f_X is called the probability density function.
- Same as saying F_X is absolutely continuous.
- Since F_X is continuous here, we have

$$P[X = x] = F_X(x) - F_X(x^-) = 0, \ \forall x$$

- A continuous rv takes uncountably many distinct values. However, not every rv that takes uncountably many values is a continuous rv

Recap: probability density function

- The pdf of a continuous rv is defined to be the f_X that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt, \quad \forall x$$

- It satisfies
  1. $f_X(x) \geq 0, \quad \forall x$
  2. $\int_{-\infty}^{\infty} f_X(t)\, dt = 1$
- We can, in principle, compute probability of any event as

$$P[X \in B] = \int_B f_X(t)\, dt, \quad \forall B \in \mathcal{B}$$

- In particular,

$$P[a \leq X \leq b] = \int_a^b f_X(t)\, dt$$

Recap: Function of a random variable

- If X is a random variable and $g : \Re \to \Re$ is a function, then $Y = g(X)$ is a random variable.
- More formally, Y is a random variable if g is a Borel measurable function.
- We can determine distribution of Y given the function g and the distribution of X

Recap

- Let X be a rv and let $Y = g(X)$.
- The distribution function of Y is given by

$$\begin{aligned} F_Y(y) &= P[g(X) \leq y] \\ &= P[X \in \{z \,:\, g(z) \leq y\}] \end{aligned}$$

- This probability can be obtained from distribution of X.

Recap

- Suppose X is a discrete rv with $X \in \{x_1, x_2, \cdots\}$.
- Suppose $Y = g(X)$.
- Then Y is also discrete and $Y \in \{g(x_1), g(x_2), \cdots\}$.
- We can find the pmf of Y as

$$\begin{aligned} f_Y(y) &= p[Y = y] = P[g(X) = y] \\ &= P[X \in \{x_i \,:\, g(x_i) = y\}] \\ &= \sum_{\substack{i: \\ g(x_i) = y}} f_X(x_i) \end{aligned}$$

Recap

- Let $g : \Re \to \Re$ be differentiable with $g'(x) > 0, \forall x$ or $g'(x) < 0, \forall x$.
- Let X be a continuous rv and let $Y = g(X)$.
- Then Y is a continuous rv with pdf

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|, \; a \le y \le b$$

  where $a = \min(g(\infty), \; g(-\infty))$ and $b = \max(g(\infty), \; g(-\infty))$
- This theorem is useful in some cases to find the densities of functions of continuous random variables

Recap: Expectation

- Let X be a discrete rv with $X \in \{x_1, x_2, \cdots\}$. Then

$$E[X] = \sum_i x_i \; f_X(x_i)$$

- If X is a continuous random variable with pdf, f_X,

$$E[X] = \int_{-\infty}^{\infty} x \; f_X(x) \; dx$$

- Sometimes we use the following notation to denote expectation of both kinds of rv

$$E[X] = \int_{-\infty}^{\infty} x \; dF_X(x)$$

- We take the expectation to exist when the sum or integral above is absolutely convergent
- Note that expectation is defined for all random variables

Recap: Expectation of a function of a random variable

- Let X be a rv and let $Y = g(X)$. Then,
- $EY = \int y \; dF_Y(y) = \int g(x) \; dF_X(x)$
- That is, if X is discrete, then

$$EY = \sum_j y_j \; f_Y(y_j) = \sum_i g(x_i) f_X(x_i)$$

- If X and Y are continuous

$$EY = \int y \; f_Y(y) \; dy = \int g(x) \; f_X(x) \; dx$$

- This is true for all rv's.

Recap: Properties of Expectation

$$E[g(X)] = \sum_i g(x_i) f_X(x_i) \quad \text{or} \quad E[g(X)] = \int_{-\infty}^{\infty} g(x) \; f_X(x) \; dx$$

- If $X \ge 0$ then $EX \ge 0$
- $E[b] = b$ where b is a constant
- $E[ag(X)] = aE[g(X)]$ where a is a constant
- $E[aX + b] = aE[X] + b$ where a, b are constants.
- $E[ag_1(X) + bg_2(X)] = aE[g_1(X)] + bE[g_2(X)]$
- $E\left[(X - c)^2\right] \ge E\left[(X - EX)^2\right], \; \forall c$

Recap: Variance of random variable

- $$\text{Var}(X) = E\left[(X - EX)^2\right] = E[X^2] - (EX)^2$$
- Properties of Variance:
  - $\text{Var}(X) \geq 0$
  - $\text{Var}(X + c) = \text{Var}(X)$
  - $\text{Var}(cX) = c^2\text{Var}(X)$

Recap: Moments of a random variable

- The k^{th} (order) moment of X is
$$m_k = E[X^k] = \int x^k \, dF_X(x)$$
- The k^{th} central moment of X is
$$s_k = E[(X - EX)^k] = \int (x - EX)^k \, dF_X(x)$$
- If moment of order k is finite then so is moment of order s for $s < k$.

Recap: Moment Generating function

- The moment generating function – $M_X : \Re \to \Re$
$$M_X(t) = Ee^{tX} = \sum_i e^{tx_i} f_X(x_i) \;\text{ or }\; \int e^{tx} f_X(x) \, dx, \;\; t \in \Re$$
- We say the mgf exists if $E[e^{tX}] < \infty$ for t in some interval around zero
- If $M_X(t)$ exists (for $t \in [-a, \; a]$ for some $a > 0$) then all its derivatives also exist and
$$\left.\frac{d^k M_X(t)}{dt^k}\right|_{t=0} = E[X^k]$$

Generating function

- For $X \in \{0, 1, 2, \cdots\}$ the (probability) generating function of X is defined by
$$P_X(s) = \sum_{k=0}^{\infty} f_X(k) s^k, \;\; s \in \Re$$
- We get the pmf from it as
$$f_X(0) = P_X(0); \; f_X(1) = \frac{P'_X(0)}{1!}; \; f_X(2) = \frac{P''_X(0)}{2!}$$
- We can also get the moments:
$$P'_X(1) = EX, \quad P''_X(1) = E[X(X-1)]$$

quantiles of a distribution

- Let $p \in (0,\ 1)$. The number $x \in \Re$ that satisfies

$$P[X \leq x] \geq p \quad \text{and} \quad p[X \geq x] \geq 1-p$$

is called the quantile of order p or the $100p^{th}$ percentile of rv X.
- If x is quantile of order p, it satisfies

$$p \leq F_X(x) \leq p + P[X = x]$$

- For a given p there can be multiple values for x to satisfy the above.
- For $p = 0.5$, it is called the median.

Recap: some moment inequalities

- **Markov inequality**: For a non-negative function, g,

$$P[g(X) > c] \leq \frac{E[g(X)]}{c}$$

- A specific instance of this is

$$P[|X| > c] \leq \frac{E\left[|X|^k\right]}{c^k}$$

- **Chebyshev inequality**

$$P[|X - EX| > c] \leq \frac{\text{Var}(X)}{c^2}$$

- With $EX = \mu$ and $\text{Var}(X) = \sigma^2$, we get

$$P[|X - \mu| > k\sigma] \leq \frac{1}{k^2}$$

Recap: A pair of random variables

- Let X, Y be random variables on the probability space $(\Omega, \mathcal{F}, P)$
- We can think of X, Y together as a vector-valued function mapping Ω to $\Re^2$.
- This gives rise to a new probability space $(\Re^2, \mathcal{B}^2, P_{XY})$ with P_{XY} given by

$$\begin{aligned} P_{XY}(B) &= P[(X,Y) \in B],\ \forall B \in \mathcal{B}^2 \\ &= P\left(\{\omega\ :\ (X(\omega).Y(\omega)) \in B\}\right) \end{aligned}$$

Recap: Joint distribution function

- Let X, Y be random variables on the same probability space $(\Omega, \mathcal{F}, P)$
- The joint distribution function of X, Y is $F_{XY} : \Re^2 \to \Re$, defined by

$$\begin{aligned} F_{XY}(x,y) &= P[X \le x, Y \le y] \quad (= P_{XY}((-\infty, x] \times (-\infty, y])) \\ &= P\left(\{\omega \,:\, X(\omega) \le x\} \cap \{\omega \,:\, Y(\omega) \le y\}\right) \end{aligned}$$

- The joint distribution function is the probability of the intersection of the events $[X \le x]$ and $[Y \le y]$.

Recap: Properties of Joint Distribution Function

- Joint distribution function: $F_{XY} : \Re^2 \to \Re$

$$F_{XY}(x,y) = P[X \le x, Y \le y]$$

- It satisfies
  1. $F_{XY}(-\infty, y) = F_{XY}(x, -\infty) = 0, \forall x, y$; $F_{XY}(\infty, \infty) = 1$
  2. F_{XY} is non-decreasing in each of its arguments
  3. F_{XY} is right continuous and has left-hand limits in each of its arguments
  4. For all $x_1 < x_2$ and $y_1 < y_2$

$$F_{XY}(x_2,y_2) - F_{XY}(x_2,y_1) - F_{XY}(x_1,y_2) + F_{XY}(x_1,y_1) \ge 0$$

- Any $F : \Re^2 \to \Re$ satisfying the above would be a joint distribution function.

- Let X, Y be two discrete random variables (defined on the same probability space).
- Let $X \in \{x_1, \cdots x_n\}$ and $Y \in \{y_1, \cdots, y_m\}$.
- We define the joint probability mass function of X and Y as

$$f_{XY}(x_i, y_j) = P[X = x_i, Y = y_j]$$

  ($f_{XY}(x,y)$ is zero for all other values of x, y)
- The f_{XY} would satisfy
  - $f_{XY}(x,y) \ge 0, \ \forall x, y$ and $\sum_i \sum_j f_{XY}(x_i, y_j) = 1$
- This is a straight-forward extension of the pmf of a single discrete rv.

Example

- Let $\Omega = (0, \ 1)$ with the 'usual' probability.
- So, each ω is a real number between 0 and 1
- Let $X(\omega)$ be the digit in the first decimal place in ω and let $Y(\omega)$ be the digit in the second decimal place.
- If $\omega = 0.2576$ then $X(\omega) = 2$ and $Y(\omega) = 5$
- Easy to see that $X, Y \in \{0, 1, \cdots, 9\}$.
- We want to calculate the joint pmf of X and Y

Example

- What is the event $[X = 4]$?

$$[X = 4] = \{\omega \ : \ X(\omega) = 4\} = [0.4,\ 0.5)$$

- What is the event $[Y = 3]$?

$$[Y = 3] = [0.03, 0.04) \cup [0.13, 0.14) \cup \cdots \cup [0.93, 0.94)$$

- What is the event $[X = 4, Y = 3]$?
  It is the intersection of the above

$$[X = 4, Y = 3] = [0.43, 0.44)$$

- Hence the joint pmf of X and Y is

$$f_{XY}(x, y) = P[X = x, Y = y] = 0.01,\ x, y \in \{0, 1, \cdots, 9\}$$

Example

- Consider the random experiment of rolling two dice.
  $\Omega = \{(\omega_1, \omega_2) \ : \ \omega_1, \omega_2 \in \{1, 2, \cdots, 6\}\}$
- Let X be the maximum of the two numbers and let Y be the sum of the two numbers.
- Easy to see $X \in \{1, 2, \cdots, 6\}$ and $Y \in \{2, 3, \cdots, 12\}$
- What is the event $[X = m, Y = n]$? (We assume m, n are in the correct range)

$$[X = m, Y = n] = \{(\omega_1, \omega_2) \in \Omega \ : \ \max(\omega_1, \omega_2) = m,\ \omega_1 + \omega_2 = n\}$$

- For this to be a non-empty set, we must have $m < n \leq 2m$
- Then $[X = m, Y = n] = \{(m,\ n - m),\ (n - m,\ m)\}$
- Is this always true? No! What if $n = 2m$?
  $[X = 3, Y = 6] = \{(3,\ 3)\}$,
  $[X = 4, Y = 6] = \{(4,\ 2),\ (2,\ 4)\}$
- So, $P[X = m, Y = n]$ is either $2/36$ or $1/36$ (assuming m, n satisfy other requirements)

Example

- We can now write the joint pmf.
- Assume $1 \leq m \leq 6$ and $2 \leq n \leq 12$. Then

$$f_{XY}(m, n) = \begin{cases} \frac{2}{36} & \text{if } \ m < n < 2m \\ \frac{1}{36} & \text{if } \ n = 2m \end{cases}$$

  ($f_{XY}(m, n)$ is zero in all other cases)
- Does this satisfy requirements of joint pmf?

$$\begin{aligned} \sum_{m,n} f_{XY}(m, n) &= \sum_{m=1}^{6} \sum_{n=m+1}^{2m-1} \frac{2}{36} + \sum_{m=1}^{6} \frac{1}{36} \\ &= \frac{2}{36} \sum_{m=1}^{6} (m - 1) + \frac{1}{36} 6 \\ &= \frac{2}{36}(21 - 6) + \frac{6}{36} = 1 \end{aligned}$$

Joint Probability mass function

- Let $X \in \{x_1, x_2, \cdots\}$ and $Y \in \{y_1, y_2, \cdots\}$ be discrete random variables.
- The joint pmf: $f_{XY}(x, y) = P[X = x, Y = y]$.
- The joint pmf satisfies:
  - $f_{XY}(x, y) \geq 0, \forall x, y$ and
  - $\sum_i \sum_j f_{XY}(x_i, y_j) = 1$
- Given the joint pmf, we can get the joint df as

$$F_{XY}(x, y) = \sum_{\substack{i: \\ x_i \leq x}} \sum_{\substack{j: \\ y_j \leq y}} f_{XY}(x_i, y_j)$$

- Given sets $\{x_1, x_2, \cdots\}$ and $\{y_1, y_2, \cdots\}$.
- Suppose $f_{XY} : \Re^2 \to [0,\ 1]$ be such that
  - $f_{XY}(x, y) = 0$ unless $x = x_i$ for some i and $y = y_j$ for some j, and
  - $\sum_i \sum_j f_{XY}(x_i, y_j) = 1$
- Then f_{XY} is a joint pmf.
- This is because, if we define

$$F_{XY}(x, y) = \sum_{\substack{i:\\ x_i \le x}} \sum_{\substack{j:\\ y_j \le y}} f_{XY}(x_i, y_j)$$

  then F_{XY} satisfies all properties of a df.
- We normally specify a pair of discrete random variables by giving the joint pmf

- Given the joint pmf, we can (in principle) compute the probability of any event involving the two discrete random variables.

$$P[(X, Y) \in B] = \sum_{\substack{i,j:\\ (x_i, y_j) \in B}} f_{XY}(x_i, y_j)$$

- Now, events can be specified in terms of relations between the two rv's too

$$[X < Y + 2] = \{\omega\ :\ X(\omega) < Y(\omega) + 2\}$$

- Thus,

$$P[X < Y + 2] = \sum_{\substack{i,j:\\ x_i < y_j + 2}} f_{XY}(x_i, y_j)$$

- Take the example: 2 dice, X is max and Y is sum
- $f_{XY}(m, n) = 0$ unless $m = 1, \cdots, 6$ and $n = 2, \cdots, 12$. For this range

$$f_{XY}(m, n) = \begin{cases} \frac{2}{36} & \text{if } \ m < n < 2m \\ \frac{1}{36} & \text{if } \ n = 2m \end{cases}$$

- Suppose we want $P[Y = X + 2]$.

$$\begin{aligned} P[Y = X + 2] &= \sum_{\substack{m,n:\\ n = m+2}} f_{XY}(m, n) = \sum_{m=1}^{6} f_{XY}(m, m+2) \\ &= \sum_{m=2}^{6} f_{XY}(m, m+2) \quad \text{since we need } m + 2 \le 2m \\ &= \frac{1}{36} + 4\,\frac{2}{36} = \frac{9}{36} \end{aligned}$$

Joint density function

- Let X, Y be two continuous rv's with df F_{XY}.
- If there exists a function f_{XY} that satisfies

$$F_{XY}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x', y')\ dy'\ dx', \ \ \forall x, y$$

  then we say that X, Y have a joint probability density function which is f_{XY}
- Please note the difference in the definition of joint pmf and joint pdf.
- When X, Y are discrete we defined a joint pmf
- We are not saying that if X, Y are continuous rv's then a joint density exists.
- We use joint density to mean joint pdf

properties of joint density

- The joint density (or joint pdf) of X, Y is f_{XY} that satisfies

$$F_{XY}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x',y')\ dy'\ dx', \ \ \forall x,y$$

- Since F_{XY} is non-decreasing in each argument, we must have $f_{XY}(x,y) \geq 0$.
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x',y')\ dy'\ dx' = 1$ is needed to ensure $F_{XY}(\infty,\infty) = 1$.

properties of joint density

- The joint density f_{XY} satisfies the following
  1. $f_{XY}(x,y) \geq 0, \ \ \forall x,y$
  2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x',y')\ dy'\ dx' = 1$
- These are very similar to the properties of the density of a single rv

Example: Joint Density

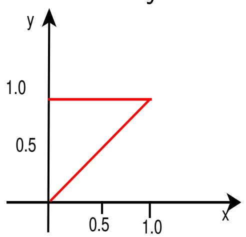
- Consider the function

$$f(x,y) = 2, \ 0 < x < y < 1 \ \ (f(x,y) = 0, \ \text{otherwise})$$

- Let us show this is a density

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\ dx\ dy = \int_{0}^{1} \int_{0}^{y} 2\ dx\ dy = \int_{0}^{1} 2\ x|_{0}^{y}\ dy = \int_{0}^{1} 2y\ dy = 1$$

- We can say this density is uniform over the region



properties of joint density

- The joint density f_{XY} satisfies the following
  1. $f_{XY}(x,y) \geq 0, \ \ \forall x,y$
  2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x',y')\ dy'\ dx' = 1$
- Any function $f_{XY} : \Re^2 \to \Re$ satisfying the above two is a joint density function.
- Given f_{XY} satisfying the above, define

$$F_{XY}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x',y')\ dy'\ dx', \ \ \forall x,y$$

- Then we can show F_{XY} is a joint distribution.

- $f_{XY}(x,y) \geq 0$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x',y')\ dy'\ dx' = 1$
- Define

$$F_{XY}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x',y')\ dy'\ dx', \quad \forall x, y$$

- Then, $F_{XY}(-\infty, y) = F_{XY}(x, -\infty) = 0, \forall x, y$ and $F_{XY}(\infty, \infty) = 1$
- Since $f_{XY}(x,y) \geq 0$, F_{XY} is non-decreasing in each argument.
- Since it is given as an integral, the above also shows that F_{XY} is continuous in each argument.
- The only property left is the special property of F_{XY} we mentioned earlier.

$$\Delta \triangleq F_{XY}(x_2,y_2) - F_{XY}(x_1,y_2) - F_{XY}(x_2,y_1) + F_{XY}(x_1,y_1).$$

- We need to show $\Delta \geq 0$ if $x_1 < x_2$ and $y_1 < y_2$.
- We have

$$\begin{aligned}
\Delta &= \int_{-\infty}^{x_2} \int_{-\infty}^{y_2} f_{XY}\ dy\ dx - \int_{-\infty}^{x_1} \int_{-\infty}^{y_2} f_{XY}\ dy\ dx \\
&\quad - \int_{-\infty}^{x_2} \int_{-\infty}^{y_1} f_{XY}\ dy\ dx + \int_{-\infty}^{x_1} \int_{-\infty}^{y_1} f_{XY}\ dy\ dx \\
&= \int_{-\infty}^{x_2} \left(\int_{-\infty}^{y_2} f_{XY}\ dy - \int_{-\infty}^{y_1} f_{XY}\ dy \right) dx \\
&\quad - \int_{-\infty}^{x_1} \left(\int_{-\infty}^{y_2} f_{XY}\ dy - \int_{-\infty}^{y_1} f_{XY}\ dy \right) dx
\end{aligned}$$

- Thus we have

$$\begin{aligned}
\Delta &= \int_{-\infty}^{x_2} \left(\int_{-\infty}^{y_2} f_{XY}\ dy - \int_{-\infty}^{y_1} f_{XY}\ dy \right) dx \\
&\quad - \int_{-\infty}^{x_1} \left(\int_{-\infty}^{y_2} f_{XY}\ dy - \int_{-\infty}^{y_1} f_{XY}\ dy \right) dx \\
&= \int_{-\infty}^{x_2} \int_{y_1}^{y_2} f_{XY}\ dy\ dx - \int_{-\infty}^{x_1} \int_{y_1}^{y_2} f_{XY}\ dy\ dx \\
&= \int_{x_1}^{x_2} \int_{y_1}^{y_2} f_{XY}\ dy\ dx \geq 0
\end{aligned}$$

- This actually shows

$$P[x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2] = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f_{XY}\ dy\ dx$$

- What we showed is the following
- Any function $f_{XY} : \Re^2 \to \Re$ that satisfies
  - $f_{XY}(x,y) \geq 0, \ \forall x, y$
  - $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x,y)\ dx\ dy = 1$

  is a joint density function.
- This is because now $F_{XY}(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{XY}(x,y)\ dx\ dy$ would satisfy all conditions for a df.
- Convenient to specify joint density (when it exists)
- We also showed

$$P[x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2] = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f_{XY}\ dy\ dx$$

- In general

$$P[(X,Y) \in B] = \int_B f_{XY}(x,y)\ dx\ dy, \quad \forall B \in \mathcal{B}^2$$

- Let us consider the example

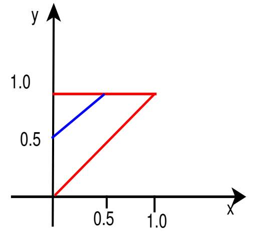
$$f(x,y) = 2, \;\; 0 < x < y < 1$$

- Suppose wee want probability of $[Y > X + 0.5]$

$$\begin{aligned} P[Y > X + 0.5] &= P\left[(X,Y) \in \{(x,y) : y > x + 0.5\}\right] \\ &= \int_{\{(x,y)\,:\, y > x+0.5\}} f_{XY}(x,y)\; dx\; dy \\ &= \int_{0.5}^{1} \int_{0}^{y-0.5} 2\; dx\; dy \\ &= \int_{0.5}^{1} 2(y-0.5) dy \end{aligned}$$

$$= 2\; \frac{y^2}{2}\bigg|_{0.5}^{1} - y|_{0.5}^{1} = 1 - 0.25 - 1 + 0.5 = 0.25$$

- We can look at it geometrically



- The probability of the event we want is the area of the small triangle divided by that of the big triangle.

Marginal Distributions

- Let X, Y be random variables with joint distribution function F_{XY}.
- We know $F_{XY}(x,y) = P[X \le x, Y \le y]$.
- Hence

$$F_{XY}(x,\infty) = P[X \le x, Y \le \infty] = P[X \le x] = F_X(x)$$

- We define the marginal distribution functions of X, Y by

$$F_X(x) = F_{XY}(x,\infty); \quad F_Y(y) = F_{XY}(\infty, y)$$

- These are simply distribution functions of X and Y obtained from the joint distribution function.

Marginal mass functions

- Let $X \in \{x_1, x_2, \cdots\}$ and $Y \in \{y_1, y_2, \cdots\}$
- Let f_{XY} be their joint mass function.
- Then

$$P[X = x_i] = \sum_j P[X = x_i, Y = y_j] = \sum_j f_{XY}(x_i, y_j)$$

(This is because $[Y = y_j],\; j = 1, \cdots,$ form a partition and $P(A) = \sum_i P(AB_i)$ when B_i is a partition)
- We define the marginal mass functions of X and Y as

$$f_X(x_i) = \sum_j f_{XY}(x_i, y_j); \quad f_Y(y_j) = \sum_i f_{XY}(x_i, y_j)$$

- These are mass functions of X and Y obtained from the joint mass function

marginal density functions

- Let X, Y be continuous rv with joint density f_{XY}.
- Then we know $F_{XY}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x', y')\ dy'\ dx'$
- Hence, we have

$$\begin{aligned} F_X(x) = F_{XY}(x, \infty) &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{XY}(x', y')\ dy'\ dx' \\ &= \int_{-\infty}^{x} \left(\int_{-\infty}^{\infty} f_{XY}(x', y')\ dy' \right)\ dx' \end{aligned}$$

- Since X is a continuous rv, this means

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\ dy$$

We call this the marginal density of X.

- Similarly, marginal density of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\ dx$$

- These are pdf's of X and Y obtained from the joint

Example

- Rolling two dice, X is max, Y is sum
- We had, for $1 \leq m \leq 6$ and $2 \leq n \leq 12$,

$$f_{XY}(m, n) = \begin{cases} \frac{2}{36} & \text{if } m < n < 2m \\ \frac{1}{36} & \text{if } n = 2m \end{cases}$$

- We know, $f_X(m) = \sum_n f_{XY}(m, n),\ m = 1, \cdots, 6.$
- Given m, for what values of n, $f_{XY}(m, n) > 0$?
  We can only have $n = m + 1, \cdots, 2m$.
- Hence we get

$$f_X(m) = \sum_{n=m+1}^{2m} f_{XY}(m, n) = \sum_{n=m+1}^{2m-1} \frac{2}{36} + \frac{1}{36} = \frac{2}{36}(m-1) + \frac{1}{36} = \frac{2m-1}{36}$$

Example

- Consider the joint density

$$f_{XY}(x, y) = 2,\ 0 < x < y < 1$$

- The marginal density of X is: for $0 < x < 1$,

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\ dy = \int_{x}^{1} 2\ dy = 2(1 - x)$$

Thus, $f_X(x) = 2(1 - x),\ 0 < x < 1$

- We can easily verify this is a density

$$\int_{-\infty}^{\infty} f_X(x)\ dx = \int_{0}^{1} 2(1 - x)\ dx = (2x - x^2)\Big|_0^1 = 1$$

We have: $f_{XY}(x, y) = 2,\ \ 0 < x < y < 1$

- We can similarly find density of Y.
- For $0 < y < 1$,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\ dx = \int_{0}^{y} 2\ dx = 2y$$

- Thus, $f_Y(y) = 2y,\ 0 < y < 1$ and

$$\int_{0}^{1} 2y\ dy = 2\ \frac{y^2}{2}\Big|_0^1 = 1$$

- If we are given the joint df or joint pmf/joint density of X, Y, then the individual df or pmf/pdf are uniquely determined.
- However, given individual pdf of X and Y, we cannot determine the joint density. (same is true of pmf or df)
- There can be many different joint density functions all having the same marginals

Conditional distributions

- Let X, Y be rv's on the same probability space
- We define the conditional distribution of X given Y by

$$F_{X|Y}(x|y) = P[X \leq x | Y = y]$$

  (For now ignore the case of $P[Y = y] = 0$).
- Note that $F_{X|Y} : \Re^2 \to \Re$
- $F_{X|Y}(x|y)$ is a notation. We could write $F_{X|Y}(x, y)$.

- Conditional distribution of X given Y is

$$F_{X|Y}(x|y) = P[X \leq x | Y = y]$$

  It is the conditional probability of $[X \leq x]$ given (or conditioned on) $[Y = y]$.
- Consider example: rolling 2 dice, X is max, Y is sum

$$P[X \leq 4 | Y = 3] = 1; \quad P[X \leq 4 | Y = 9] = 0$$

- This is what conditional distribution captures.
- For every value of y, $F_{X|Y}(x|y)$ is a distribution function in the variable x.
- It defines a new distribution for X based on knowing the value of Y.

- Let: $X \in \{x_1, x_2, \cdots\}$ and $Y \in \{y_1, y_2, \cdots\}$. Then

$$F_{X|Y}(x|y_j) = P[X \leq x | Y = y_j] = \frac{P[X \leq x, Y = y_j]}{P[Y = y_j]}$$

  (We define $F_{X|Y}(x|y)$ only when $y = y_j$ for some j).
- For each y_j, $F_{X|Y}(x|y_j)$ is a df of a discrete rv in x.
- Since X is a discrete rv, we can write the above as

$$\begin{aligned} F_{X|Y}(x|y_j) &= \frac{P[X \leq x, Y = y_j]}{P[Y = y_j]} = \frac{\sum_{i: x_i \leq x} P[X = x_i, Y = y_j]}{P[Y = y_j]} \\ &= \sum_{i: x_i \leq x} \left(\frac{f_{XY}(x_i, y_j)}{f_Y(y_j)} \right) \end{aligned}$$

Conditional mass function

- We got

$$F_{X|Y}(x|y_j) = \sum_{i:x_i \le x} \left(\frac{f_{XY}(x_i, y_j)}{f_Y(y_j)} \right)$$

- Since X is a discrete rv, what is inside the summation above is the pmf corresponding to the df, $F_{X|Y}$.
- We define the conditional mass function of X given Y as

$$f_{X|Y}(x_i|y_j) = \frac{f_{XY}(x_i, y_j)}{f_Y(y_j)} = P[X = x_i|Y = y_j]$$

- The conditional mass function is

$$f_{X|Y}(x_i|y_j) = P[X = x_i|Y = y_j] = \frac{f_{XY}(x_i, y_j)}{f_Y(y_j)}$$

- This gives us the useful identity

$$f_{XY}(x_i, y_j) = f_{X|Y}(x_i|y_j) f_Y(y_j)$$

($P[X = x_i, Y = y_j] = P[X = x_i|Y = y_j]P[Y = y_j]$)
- This gives us the total proability rule for rv's

$$f_X(x_i) = \sum_j f_{XY}(x_i, y_j) = \sum_j f_{X|Y}(x_i|y_j) f_Y(y_j)$$

- This is same as

$$P[X = x_i] = \sum_j P[X = x_i|Y = y_j]P[Y = y_j]$$

($P(A) = \sum_j P(A|B_j)P(B_j)$ when $B_1, \cdots$ form a partition)

- We have

$$f_{XY}(x_i, y_j) = f_{X|Y}(x_i|y_j) f_Y(y_j) = f_{Y|X}(y_j|x_i) f_X(x_i)$$

- This gives us Bayes rule for discrete rv's

$$\begin{aligned} f_{X|Y}(x_i|y_j) &= \frac{f_{Y|X}(y_j|x_i) f_X(x_i)}{f_Y(y_j)} \\ &= \frac{f_{Y|X}(y_j|x_i) f_X(x_i)}{\sum_i f_{XY}(x_i, y_j)} \\ &= \frac{f_{Y|X}(y_j|x_i) f_X(x_i)}{\sum_i f_{Y|X}(y_j|x_i) f_X(x_i)} \end{aligned}$$

Recap: Joint Distribution Function

- Given X, Y rv on same probability space, joint distribution function: $F_{XY} : \Re^2 \to \Re$

$$F_{XY}(x, y) = P[X \le x, Y \le y]$$

- It satisfies
  1. $F_{XY}(-\infty, y) = F_{XY}(x, -\infty) = 0, \forall x, y;$ $F_{XY}(\infty, \infty) = 1$
  2. F_{XY} is non-decreasing in each of its arguments
  3. F_{XY} is right continuous and has left-hand limits in each of its arguments
  4. For all $x_1 < x_2$ and $y_1 < y_2$

$$F_{XY}(x_2, y_2) - F_{XY}(x_2, y_1) - F_{XY}(x_1, y_2) + F_{XY}(x_1, y_1) \ge 0$$

- Any $F : \Re^2 \to \Re$ satisfying the above would be a joint distribution function.

Recap: Joint Probability mass function

- $X \in \{x_1, x_2, \cdots\}$, $Y \in \{y_1, y_2, \cdots\}$
- The joint pmf: $f_{XY}(x, y) = P[X = x, Y = y]$.
- The joint pmf satisfies:
  - A1 $f_{XY}(x, y) \geq 0, \forall x, y$ and non-zero only for x_i, y_j pairs
  - A2 $\sum_i \sum_j f_{XY}(x_i, y_j) = 1$
- Given the joint pmf, we can get the joint df as

$$F_{XY}(x, y) = \sum_{\substack{i: \\ x_i \leq x}} \sum_{\substack{j: \\ y_j \leq y}} f_{XY}(x_i, y_j)$$

- Any $f_{XY} : \Re^2 \to [0,\ 1]$ satisfying $A1$ and $A2$ above is a joint pmf. (The F_{XY} satisfies all properties of df).
- Given the joint pmf, we can (in principle) compute the probability of any event involving the two discrete random variables.

$$P[(X, Y) \in B] = \sum_{\substack{i,j: \\ (x_i, y_j) \in B}} f_{XY}(x_i, y_j)$$

Recap joint density

- Two cont rv X, Y have a joint density f_{XY} if

$$F_{XY}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x', y')\ dy'\ dx', \quad \forall x, y$$

- The joint density f_{XY} satisfies the following
  1. $f_{XY}(x, y) \geq 0, \quad \forall x, y$
  2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x', y')\ dy'\ dx' = 1$
- Any function $f_{XY} : \Re^2 \to \Re$ satisfying the above two is a joint density function. (Then the above F_{XY} can be shown to be a joint df).
- We also have

$$P[x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2] = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f_{XY}\ dy\ dx$$

and, in general,

$$P[(X, Y) \in B] = \int_B f_{XY}(x, y)\ dx\ dy, \quad \forall B \in \mathcal{B}^2$$

Recap Marginals

- Marginal distribution functions of X, Y are

$$F_X(x) = F_{XY}(x, \infty); \quad F_Y(y) = F_{XY}(\infty, y)$$

- X, Y discrete with joint pmf f_{XY}. The marginal pmfs are

$$f_X(x) = \sum_y f_{XY}(x, y); \quad f_Y(y) = \sum_x f_{XY}(x, y)$$

- If X, Y have joint pdf f_{XY} then the marginal pdf are

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

Recap Conditional distribution

- Let: $X \in \{x_1, x_2, \cdots\}$ and $Y \in \{y_1, y_2, \cdots\}$. Then

$$F_{X|Y}(x|y_j) = P[X \leq x | Y = y_j] = \frac{P[X \leq x, Y = y_j]}{P[Y = y_j]}$$

(We define $F_{X|Y}(x|y)$ only when $y = y_j$ for some j).
- For each y_j, $F_{X|Y}(x|y_j)$ is a df of a discrete rv in x.
- The pmf corresponding to this df is called conditional pmf

$$f_{X|Y}(x_i|y_j) = P[X = x_i | Y = y_j] = \frac{f_{XY}(x_i, y_j)}{f_Y(y_j)}$$

Recap Bayes rule for discrete rv's

- The conditional mass function is

$$f_{X|Y}(x_i|y_j) = P[X = x_i|Y = y_j] = \frac{f_{XY}(x_i, y_j)}{f_Y(y_j)}$$

- This gives us the useful identity

$$f_{XY}(x_i, y_j) = f_{X|Y}(x_i|y_j) f_Y(y_j)$$

- This gives us the total probability rule for rv's

$$f_X(x_i) = \sum_j f_{XY}(x_i, y_j) = \sum_j f_{X|Y}(x_i|y_j) f_Y(y_j)$$

- Also gives us Bayes rule for discrete rv

$$f_{X|Y}(x_i|y_j) = \frac{f_{Y|X}(y_j|x_i) f_X(x_i)}{\sum_i f_{Y|X}(y_j|x_i) f_X(x_i)}$$

Example: Conditional pmf

- Consider the random experiment of tossing a coin n times.
- Let X denote the number of heads and let Y denote the toss number on which the first head comes.
- For $1 \leq k \leq n$

$$\begin{aligned} f_{Y|X}(k|1) &= P[Y = k|X = 1] = \frac{P[Y = k, X = 1]}{P[X = 1]} \\ &= \frac{p(1-p)^{n-1}}{{}^nC_1 p(1-p)^{n-1}} \\ &= \frac{1}{n} \end{aligned}$$

- Given there is only one head, it is equally likely to occur on any toss.

- Let X, Y be continuous rv's with joint density, f_{XY}.
- We once again want to define conditional df

$$F_{X|Y}(x|y) = P[X \leq x|Y = y]$$

- But the conditioning event, $[Y = y]$ has zero probability.
- Hence we define conditional df as follows

$$F_{X|Y}(x|y) = \lim_{\delta \to 0} P[X \leq x|Y \in [y, \ y + \delta]]$$

- This is well defined if the limit exists.
- The limit exists for all y where $f_Y(y) > 0$ (and for all x)

- The conditional df is given by (assuming $f_Y(y) > 0$)

$$\begin{aligned} F_{X|Y}(x|y) &= \lim_{\delta \to 0} P[X \leq x|Y \in [y, \ y + \delta]] \\ &= \lim_{\delta \to 0} \frac{P[X \leq x, Y \in [y, y + \delta]]}{P[Y \in [y, y + \delta]]} \\ &= \lim_{\delta \to 0} \frac{\int_{-\infty}^{x} \int_{y}^{y+\delta} f_{XY}(x', y') \ dy' \ dx'}{\int_{y}^{y+\delta} f_Y(y') \ dy'} \\ &= \lim_{\delta \to 0} \frac{\int_{-\infty}^{x} f_{XY}(x', y) \ \delta \ dx'}{f_Y(y) \ \delta} \\ &= \int_{-\infty}^{x} \frac{f_{XY}(x', y)}{f_Y(y)} \ dx' \end{aligned}$$

- We define conditional density of X given Y as

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

- Let X, Y have joint density f_{XY}.
- The conditional df of X given Y is

$$F_{X|Y}(x|y) = \lim_{\delta \to 0} P[X \le x | Y \in [y,\ y+\delta]]$$

- This exists if $f_Y(y) > 0$ and then it has a density:

$$F_{X|Y}(x|y) = \int_{-\infty}^{x} f_{X|Y}(x'|y)\ dx'$$

- This conditional density is given by

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

- We (once again) have the useful identity

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

Example

$$f_{XY}(x,y) = 2,\ 0 < x < y < 1$$

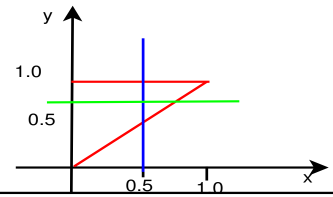
- We saw that the marginal densities are

$$f_X(x) = 2(1-x),\ 0 < x < 1; \quad f_Y(y) = 2y,\ 0 < y < 1$$

- Hence the conditional densities are given by

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{1}{y},\ 0 < x < y < 1 \\ f_{Y|X}(y|x) &= \frac{f_{XY}(x,y)}{f_X(x)} = \frac{1}{1-x},\ 0 < x < y < 1 \end{aligned}$$

- We can see this intuitively like this



- The identity $f_{XY}(x,y) = f_{X|Y}(x|y) f_Y(y)$ can be used to specify the joint density of two continuous rv's
- We can specify the marginal density of one and the conditional density of the other given the first.
- This may actually be the model of how the the rv's are generated.

Example

- Let X be uniform over $(0,\ 1)$ and let Y be uniform over 0 to X. Find the density of Y.
- What we are given is

$$f_X(x) = 1,\ 0 < x < 1; \quad f_{Y|X}(y|x) = \frac{1}{x}, 0 < y < x < 1$$

- Hence the joint density is: $f_{XY}(x,y) = \frac{1}{x},\ 0 < y < x < 1.$
- Hence the density of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\ dx = \int_y^1 \frac{1}{x}\ dx = -\ln(y),\ 0 < y < 1$$

- We can verify it to be a density

$$-\int_0^1 \ln(y)\ dy = -y\ln(y)|_0^1 + \int_0^1 y\ \frac{1}{y}\ dy = 1$$

- We have the identity
$$f_{XY}(x,y) = f_{X|Y}(x|y)\ f_Y(y)$$
- By integrating both sides
$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\ dy = \int_{-\infty}^{\infty} f_{X|Y}(x|y)\ f_Y(y)\ dy$$
- This is a continuous analogue of total probability rule.
- But note that, since X is continuous rv, $f_X(x)$ is **NOT** $P[X=x]$
- In case of discrete rv, the mass function value $f_X(x)$ is equal to $P[X=x]$ and we had
$$f_X(x) = \sum_y f_{X|Y}(x|y) f_Y(y)$$
- It is as if one can simply replace pmf by pdf and summation by integration!!
- While often that gives the right result, one needs to be very careful

- We have the identity
$$f_{XY}(x,y) = f_{X|Y}(x|y)\ f_Y(y) = f_{Y|X}(y|x) f_X(x)$$
- This gives rise to Bayes rule for continuous rv
$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)} \\ &= \frac{f_{Y|X}(y|x) f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x)\ dx} \end{aligned}$$
- This is essentially identical to Bayes rule for discrete rv's. We have essentially put the pdf wherever there was pmf

- To recap, we started by defining conditional distribution function.
$$F_{X|Y}(x|y) = P[X \le x | Y = y]$$
- When X, Y are discrete, we define this only for $y = y_j$. That is, we define it only for all values that Y can take.
- When X, Y have joint density, we defined it by
$$F_{X|Y}(x|y) = \lim_{\delta \to 0} P[X \le x | Y \in [y, y+\delta]]$$
This limit exists and $F_{X|Y}$ is well defined if $f_Y(y) > 0$. That is, essentially again for all values that Y can take.
- In the discrete case, we define $f_{X|Y}$ as the pmf corresponding to $F_{X|Y}$. This conditional pmf can also be defined as a conditional probability
- In the continuous case $f_{X|Y}$ is the density corresponding to $F_{X|Y}$.
- In both cases we have: $f_{XY}(x,y) = f_{X|Y}(x|y) f_Y(y)$
- This gives total probability rule and Bayes rule for random variables

- Now, let X be a continuous rv and let Y be discrete rv.
- We can define $F_{X|Y}$ as
$$F_{X|Y}(x|y) = P[X \le x | Y = y]$$
This is well defined for all values that y takes. (We consider only those y)
- Since X is continuous rv, this df would have a density
$$F_{X|Y}(x|y) = \int_{-\infty}^{x} f_{X|Y}(x'|y)\ dx'$$
- Hence we can write
$$\begin{aligned} P[X \le x, Y = y] &= F_{X|Y}(x|y) P[Y = y] \\ &= \int_{-\infty}^{x} f_{X|Y}(x'|y)\ f_Y(y)\ dx' \end{aligned}$$

- We now get

$$F_X(x) = P[X \le x] = \sum_y P[X \le x, Y = y]$$
$$= \sum_y \int_{-\infty}^{x} f_{X|Y}(x'|y)\ f_Y(y)\ dx'$$
$$= \int_{-\infty}^{x} \sum_y f_{X|Y}(x'|y)\ f_Y(y)\ dx'$$

- This gives us

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

- This is another version of total probability rule.
- Earlier we derived this when X, Y are discrete.
- The formula is true even when X is continuous
  Only difference is we need to take f_X as the density of X.

- When X, Y are discrete we have

$$f_X(x) = \sum_y f_{X|Y}(x|y) f_Y(y)\ (P[X = x] = \sum_y P[X = x|Y = y]P[Y = y$$

- When X is continuous and Y is discrete, we defined $f_{X|Y}(x|y)$ to be the density corresponding to $F_{X|Y}(x|y) = P[X \le x|Y = y]$
- Then we once again get

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

  Now, f_X is density (and not a mass function).
- Suppose $Y \in \{1, 2, 3\}$ and $f_Y(i) = \lambda_i$; let $f_{X|Y}(x|i) = f_i(x)$

$$f_X(x) = \lambda_1 f_1(x) + \lambda_2 f_2(x) + \lambda_3 f_3(x)$$

  Called a mixture density model

- Continuing with X continuous rv and Y discrete. We have

$$F_{X|Y}(x|y) = P[X \le x|Y = y] = \int_{-\infty}^{x} f_{X|Y}(x'|y)\ dx'$$

- We also have

$$P[X \le x, Y = y] = \int_{-\infty}^{x} f_{X|Y}(x'|y)\ f_Y(y)\ dx'$$

- Hence we can define a 'joint density'

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

- This is a kind of mixed density and mass function.
- We will not be using such 'joint densities' here

- Continuing with X continuous rv and Y discrete
- Can we define $f_{Y|X}(y|x)$?
- Since Y is discrete, this (conditional) mass function is

$$f_{Y|X}(y|x) = P[Y = y|X = x]$$

  But the conditioning event has zero prob
  We now know how to handle it

$$f_{Y|X}(y|x) = \lim_{\delta \to 0} P[Y = y|X \in [x, x + \delta]]$$

- For simplifying this we note the following:

$$P[X \le x, Y = y] = \int_{-\infty}^{x} f_{X|Y}(x'|y)\ f_Y(y)\ dx'$$

$$\Rightarrow\ P[X \in [x, x+\delta], Y = y] = \int_{x}^{x+\delta} f_{X|Y}(x'|y)\ f_Y(y)\ dx'$$

- We have

$$
\begin{aligned}
f_{Y|X}(y|x) &= \lim_{\delta \to 0} P[Y = y | X \in [x, x+\delta]] \\
&= \lim_{\delta \to 0} \frac{P[Y = y, X \in [x, x+\delta]]}{P[X \in [x, x+\delta]]} \\
&= \lim_{\delta \to 0} \frac{\int_x^{x+\delta} f_{X|Y}(x'|y)\ f_Y(y)\ dx'}{\int_x^{x+\delta} f_X(x')\ dx'} \\
&= \lim_{\delta \to 0} \frac{f_{X|Y}(x|y)\delta\ f_Y(y)}{f_X(x)\ \delta} \\
&= \frac{f_{X|Y}(x|y)\ f_Y(y)}{f_X(x)}
\end{aligned}
$$

- This gives us further versions of total probability rule and Bayes rule.

- First let us look at the total probability rule possibilities
- When X is continuous rv and Y is discrete rv, we derived

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

  Note that f_Y is mass fn, f_X is density and so on.
- Since $f_{X|Y}$ is a density (corresponding to $F_{X|Y}$),

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

- Hence we get

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

- Earlier we derived the same formula when X, Y have a joint density.

- Let us review all the total probability formulas

$$\mathbf{1.}\ f_X(x) = \sum_y f_{X|Y}(x|y) f_Y(y)$$

- We first derived this when X, Y are discrete.
- But now we proved this holds when Y is discrete
  If X is continuous the $f_X, f_{X|Y}$ are densities; If X is also discrete they are mass functions

$$\mathbf{2.}\ f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x)\ dx$$

- We first proved it when X, Y have a joint density
  We now know it holds also when X is cont and Y is discrete. In that case f_Y is a mass function

- When X is continuous rv and Y is discrete rv, we derived

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

- This once again gives rise to Bayes rule:

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

- Earlier we showed this hold when X, Y are both discrete or both continuous.
- Thus Bayes rule holds in all four possible scenarios
- Only difference is we need to interpret f_X or $f_{X|Y}$ as mass functions when X is discrete and as densities when X is a continuous rv
- In general, one refers to these always as densities since the actual meaning would be clear from context.

Example

- Consider a communication system. The transmitter puts out 0 or 5 volts for the bits of 0 and 1, and, volage measured by the receiver is the sent voltage plus noise added by the channel.
- We assume noise has Gaussian density with mean zero and variance σ^2.
- We may want the probability that the sent bit is 1 when measured voltage at the receiver is x to decide what is sent.
- Let X be the measured voltage and let Y be sent bit.
- We want to calculate $f_{Y|X}(1|x)$.
- We want to use the Bayes rule to calculate this

- We need $f_{X|Y}$. What does our model say?
- $f_{X|Y}(x|1)$ is Gaussian with mean 5 and variance σ^2 and $f_{X|Y}(x|0)$ is Gaussian with mean zero and variance σ^2

$$P[Y=1|X=x] = f_{Y|X}(1|x) = \frac{f_{X|Y}(x|1)\ f_Y(1)}{f_X(x)}$$

- We need $f_Y(1), f_Y(0)$. Let us take them to be same.
- In practice we only want to know whether $f_{Y|X}(1|x) > f_{Y|X}(0|x)$
- Then we do not need to calculate $f_X(x)$. We only need ratio of $f_{Y|X}(1|x)$ and $f_{Y|X}(0|x)$.

- The ratio of the two probabilities is

$$\begin{aligned} \frac{f_{Y|X}(1|x)}{f_{Y|X}(0|x)} &= \frac{f_{X|Y}(x|1)\ f_Y(1)}{f_{X|Y}(x|0)\ f_Y(0)} \\ &= \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-5)^2}}{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-0)^2}} \\ &= e^{-0.5\sigma^{-2}(x^2-10x+25-x^2)} \end{aligned}$$

- We are only interested in whether the above is greater than 1 or not.
- The ratio is greater than 1 if $10x > 25$ or $x > 2.5$
- So, if $X > 2.5$ we will conclude bit 1 is sent. Intuitively obvious!

- We did not calculate $f_X(x)$ in the above.
- We can calculate it if we want.
- Using total probability rule

$$\begin{aligned} f_X(x) &= \sum_y f_{X|Y}(x|y) f_Y(y) \\ &= f_{X|Y}(x|1) f_Y(1) + f_{X|Y}(x|0) f_Y(0) \\ &= \frac{1}{2}\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-5)^2}{2\sigma^2}} + \frac{1}{2}\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \end{aligned}$$

- It is a mixture density

- As we saw, given the joint distribution we can calculate all the marginals.
- However, there can be many joint distributions with the same marginals.
- Let F_1, F_2 be one dimensional df's of continuous rv's with f_1, f_2 being the corresponding densities.
  Define a function $f : \Re^2 \to \Re$ by

$$f(x,y) = f_1(x)f_2(y)\left[1 + \alpha(2F_1(x) - 1)(2F_2(y) - 1)\right]$$

  where $\alpha \in (-1, 1)$.
- First note that $f(x,y) \geq 0, \; \forall \alpha \in (-1,1)$.
  For different α we get different functions.
- We first show that $f(x,y)$ is a joint density.
- For this, we note the following

$$\int_{-\infty}^{\infty} f_1(x)\; F_1(x)\; dx = \left.\frac{(F_1(x))^2}{2}\right|_{-\infty}^{\infty} = \frac{1}{2}$$

$$f(x,y) = f_1(x)f_2(y)\left[1 + \alpha(2F_1(x) - 1)(2F_2(y) - 1)\right]$$

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\; dx\; dy &= \int_{-\infty}^{\infty} f_1(x)\; dx \int_{-\infty}^{\infty} f_2(y)\; dy \\ &\quad +\alpha \int_{-\infty}^{\infty} (2f_1(x)F_1(x) - f_1(x))\; dx \int_{-\infty}^{\infty} (2f_2(y)F_2(y) - f_2(y)) \\ &= 1\end{aligned}$$

because $2\int_{-\infty}^{\infty} f_1(x)\; F_1(x)\; dx = 1$. This also shows

$$\int_{-\infty}^{\infty} f(x,y)dx = f_2(y); \quad \int_{-\infty}^{\infty} f(x,y)dy = f_1(x)$$

- Thus infinitely many joint distributions can all have the same marginals.
- So, in general, the marginals cannot determine the joint distribution.
- An important special case where this is possible is that of independent random variables

Independent Random Variables

- Two random variable X, Y are said to be independent if for all Borel sets B_1, B_2, the events $[X \in B_1]$ and $[Y \in B_2]$ are independent.
- If X, Y are independent then

$$P[X \in B_1, Y \in B_2] = P[X \in B_1]\; P[Y \in B_2], \quad \forall B_1, B_2 \in \mathcal{B}$$

- In particular

$$F_{XY}(x,y) = P[X \leq x, Y \leq y] = P[X \leq x]P[Y \leq y] = F_X(x)\; F_Y(y)$$

- **Theorem**: X, Y are independent if and only if $F_{XY}(x,y) = F_X(x)F_Y(y)$.

- Suppose X, Y are independent discrete rv's

$$f_{XY}(x,y) = P[X=x, Y=y] = P[X=x]P[Y=y] = f_X(x)f_Y(y)$$

The joint mass function is a product of marginals.

- Suppose $f_{XY}(x,y) = f_X(x)f_Y(y)$. Then

$$\begin{aligned} F_{XY}(x,y) &= \sum_{x_i \le x, y_j \le y} f_{XY}(x_i, y_j) = \sum_{x_i \le x, y_j \le y} f_X(x_i) f_Y(y_j) \\ &= \sum_{x_i \le x} f_X(x_i) \sum_{y_j \le y} f_Y(y_j) = F_X(x)F_Y(y) \end{aligned}$$

- So, X, Y are independent if and only if $f_{XY}(x,y) = f_X(x)f_Y(y)$

- Let X, Y be independent continuous rv

$$\begin{aligned} F_{XY}(x,y) &= F_X(x)F_Y(y) = \int_{-\infty}^{x} f_X(x')\, dx' \int_{-\infty}^{y} f_Y(y')\, dy' \\ &= \int_{-\infty}^{y} \int_{-\infty}^{x} \left(f_X(x') f_Y(y') \right)\, dx'\, dy' \end{aligned}$$

- This implies joint density is product of marginals.
- Now, suppose $f_{XY}(x,y) = f_X(x)f_Y(y)$

$$\begin{aligned} F_{XY}(x,y) &= \int_{-\infty}^{y} \int_{-\infty}^{x} f_{XY}(x', y')\, dx'\, dy' \\ &= \int_{-\infty}^{y} \int_{-\infty}^{x} f_X(x') f_Y(y')\, dx'\, dy' \\ &= \int_{-\infty}^{x} f_X(x')\, dx' \int_{-\infty}^{y} f_Y(y')\, dy' = F_X(x)F_Y(y) \end{aligned}$$

- So, X, Y are independent if and only if $f_{XY}(x,y) = f_X(x)f_Y(y)$

- Let X, Y be independent.
- Then $P[X \in B_1 | Y \in B_2] = P[X \in B_1]$.
- Hence, we get $F_{X|Y}(x|y) = F_X(x)$.
- This also implies $f_{X|Y}(x|y) = f_X(x)$.
- This is true for all the four possibilities of X, Y being continuous/discrete.

Recap: Joint Distribution Function

- Given X, Y rv's on same probability space, joint distribution function: $F_{XY} : \Re^2 \to \Re$

$$F_{XY}(x,y) = P[X \le x, Y \le y]$$

- It satisfies
  1. $F_{XY}(-\infty, y) = F_{XY}(x, -\infty) = 0, \forall x, y;$ $F_{XY}(\infty, \infty) = 1$
  2. F_{XY} is non-decreasing in each of its arguments
  3. F_{XY} is right continuous and has left-hand limits in each of its arguments
  4. For all $x_1 < x_2$ and $y_1 < y_2$

     $$F_{XY}(x_2,y_2) - F_{XY}(x_2,y_1) - F_{XY}(x_1,y_2) + F_{XY}(x_1,y_1) \ge 0$$

- Any $F : \Re^2 \to \Re$ satisfying the above would be a joint distribution function.

Recap: Joint Probability mass function

- $X \in \{x_1, x_2, \cdots\}$, $Y \in \{y_1, y_2, \cdots\}$
- The joint pmf: $f_{XY}(x, y) = P[X = x, Y = y]$.
- The joint pmf satisfies:
  - A1 $f_{XY}(x, y) \geq 0, \forall x, y$ and non-zero only for x_i, y_j pairs
  - A2 $\sum_i \sum_j f_{XY}(x_i, y_j) = 1$
- Given the joint pmf, we can get the joint df as

$$F_{XY}(x, y) = \sum_{\substack{i: \\ x_i \leq x}} \sum_{\substack{j: \\ y_j \leq y}} f_{XY}(x_i, y_j)$$

- Any $f_{XY} : \Re^2 \to [0,\ 1]$ satisfying $A1$ and $A2$ above is a joint pmf. (The F_{XY} satisfies all properties of df).
- Given the joint pmf, we can (in principle) compute the probability of any event involving the two discrete random variables.

$$P[(X, Y) \in B] = \sum_{\substack{i,j: \\ (x_i, y_j) \in B}} f_{XY}(x_i, y_j)$$

Recap joint density

- Two cont rv X, Y have a joint density f_{XY} if

$$F_{XY}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x', y')\ dy'\ dx', \quad \forall x, y$$

- The joint density f_{XY} satisfies the following
  1. $f_{XY}(x, y) \geq 0, \quad \forall x, y$
  2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x', y')\ dy'\ dx' = 1$
- Any function $f_{XY} : \Re^2 \to \Re$ satisfying the above two is a joint density function. (Then the above F_{XY} can be shown to be a joint df).
- We also have

$$P[x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2] = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f_{XY}\ dy\ dx$$

and, in general,

$$P[(X, Y) \in B] = \int_B f_{XY}(x, y)\ dx\ dy, \quad \forall B \in \mathcal{B}^2$$

Recap Marginals

- Marginal distribution functions of X, Y are

$$F_X(x) = F_{XY}(x, \infty); \quad F_Y(y) = F_{XY}(\infty, y)$$

- X, Y discrete with joint pmf f_{XY}. The marginal pmfs are

$$f_X(x) = \sum_y f_{XY}(x, y); \quad f_Y(y) = \sum_x f_{XY}(x, y)$$

- If X, Y have joint pdf f_{XY} then the marginal pdf are

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

Recap Conditional distributions

- Let X, Y be continuous or discrete random variables

$$F_{X|Y}(x|y) = \lim_{\delta \to 0} P[X \leq x | Y \in [y, y + \delta]]$$

$(= P[X \leq x | Y = y]$ when Y is discrete)
- This is well defined for all values that Y can assume.
- For each y, $F_{X|Y}(x|y)$ is a df in x.
- If X, Y have a joint density or if X is continuous and Y is discrete, $F_{X|Y}$ would be absolutely continuous and would have a density.

Recap Contional density (or mass) fn

- Let X be a discrete random variable. Then

$$f_{X|Y}(x|y) = \lim_{\delta \to 0} P[X = x | Y \in [y,\ y+\delta]\ ]$$

$(= P[X = x|Y = y]$ if Y is discrete)

- This will be the mass function corresponding to the df $F_{X|Y}$.
- Let X be a continuous rv. Then we define conditional density $f_{X|Y}$ by

$$F_{X|Y}(x|y) = \int_{-\infty}^{x} f_{X|Y}(x'|y)\ dx'$$

This exists if X, Y have a joint density or when Y is discrete.

Recap

- When X, Y are both discrete or they have a joint density

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

- When X, Y are discrete or continuous (all four possibilities)

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

Here $f_{X|Y}, f_X$ are densities when X is continuous and mass functions when X is discrete. Similarly for $f_{Y|X}, f_Y$

- The above relation gives rise to the total probability rules and Bayes rule for rv's

Recap

- If Y is discrete

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

- If X is continuous, the $f_X, f_{X|Y}$ are densities; If X is also discrete, they are mass functions
- If Y is continuous

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

- If X is continuous, the $f_X, f_{X|Y}$ are densities; If X is also discrete, they are mass functions (Where needed we assume the conditional density exists)

Recap Bayes rule

- When X, Y are continuous or discrete (all four possibilities)

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

- This gives rise to Bayes rule:

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

- We need to interpret f_X or $f_{X|Y}$ as mass functions when X is discrete and as densities when X is a continuous and so on

Recap Independent Random variables

- X and Y are said to be independent if events $[X \in B_1]$, $[Y \in B_2]$ are independent for all $B_1, B_2 \in \mathcal{B}$.
- X and Y are independent if and only if
  1. $F_{XY}(x, y) = F_X(x)\ F_Y(y)$
  2. $f_{XY}(x, y) = f_X(x)\ f_Y(y)$
- This also implies $F_{X|Y}(x|y) = F_X(x)$ and $f_{X|Y}(x|y) = f_X(x)$

More than two rv

- Everything we have done so far is easily extended to multiple random variables.
- Let X, Y, Z be rv on the same probability space.
- We define joint distribution function by

$$F_{XYZ}(x, y, z) = P[X \leq x, Y \leq y, Z \leq z]$$

- If all three are discrete then the joint mass function is

$$f_{XYZ}(x, y, z) = P[X = x, Y = y, Z = z]$$

- If they are continuous , they have a joint density if

$$F_{XYZ}(x, y, z) = \int_{-\infty}^{z} \int_{-\infty}^{y} \int_{-\infty}^{x} f_{XYZ}(x', y', z')\ dx'\ dy'\ dz'$$

- Easy to see that joint mass function satisfies
  1. $f_{XYZ}(x, y, z) \geq 0$ and is non-zero only for countably many tuples.
  2. $\sum_{x,y,z}\ f_{XYZ}(x, y, z) = 1$
- Similarly the joint density satisfies
  1. $f_{XYZ}(x, y, z) \geq 0$
  2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty}\ f_{XYZ}(x, y, z)\ dx\ dy\ dz = 1$
- These are straight-forward generalizations
- The properties of joint distribution function such as it being non-decreasing in each argument etc are easily seen to hold here too.
- Generalizing the special property of the df (relating to probability of cylindrical sets) is a little more complicated. (An exercise for you!)

- Now we get many different marginals:

$$F_{XY}(x, y) = F_{XYZ}(x, y, \infty); \quad F_Z(z) = F_{XYZ}(\infty, \infty, z) \quad \text{and so on}$$

- Similarly we get

$$\begin{aligned} f_{YZ}(y, z) &= \int_{-\infty}^{\infty} f_{XYZ}(x, y, z)\ dx; \\ f_X(x) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XYZ}(x, y, z)\ dy\ dz \end{aligned}$$

- Any marginal is a joint density of a subset of these rv's and we obtain it by integrating the (full) joint density with respect to the remaining variables.
- We obtain the marginal mass functions for a subset of the rv's also similarly where we sum over the remaining variables.

- We have to be a little careful in dealing with these when some random variables are discrete and others are continuous.
- Suppose X is continuous and Y, Z are discrete. We do not have any joint density or mass function as such.
- However, the joint df is always well defined.
- Suppose we want marginal joint distribution of X, Y. We know how to get F_{XY} by marginalization.
- Then we can get f_X (a density), f_Y (a mass fn), $f_{X|Y}$ (conditinal density) and $f_{Y|X}$ (conditional mass fn)
- With these we can generally calculate most quantities of interest.

- Like in case of marginals, there are different types of conditional distributions now.
- We can always define conditional distribution functions like

$$
\begin{aligned}
F_{XY|Z}(x, y|z) &= P[X \leq x, Y \leq y | Z = z] \\
F_{X|YZ}(x|y, z) &= P[X \leq x | Y = y, Z = z]
\end{aligned}
$$

- In all such cases, if the conditioning random variables are continuous, we define the above as a limit.
- For example when Z is continuous

$$F_{XY|Z}(x, y|z) = \lim_{\delta \to 0} P[X \leq x, Y \leq y | Z \in [z, z + \delta]]$$

- If X, Y, Z are all discrete then, all conditional mass functions are defined by appropriate conditional probabilities. For example,

$$f_{X|YZ}(x|y, z) = P[X = x | Y = y, Z = z]$$

- Thus the following are obvious

$$
\begin{aligned}
f_{XY|Z}(x, y|z) &= \frac{f_{XYZ}(x, y, z)}{f_Z(z)} \\
f_{X|YZ}(x|y, z) &= \frac{f_{XYZ}(x, y, z)}{f_{YZ}(y, z)} \\
f_{XYZ}(x, y, z) &= f_{Z|YX}(z|y, x) f_{Y|X}(y|x) f_X(x)
\end{aligned}
$$

- For example, the first one above follows from

$$P[X = x, Y = y | Z = z] = \frac{P[X = x, Y = y, Z = z]}{P[Z = z]}$$

- When X, Y, Z have joint density, all such relations hold for the appropriate (conditional) densities. For example,

$$
\begin{aligned}
F_{Z|XY}(z|x, y) &= \lim_{\delta \to 0} \frac{P[Z \leq z, X \in [x, x + \delta], Y \in [y, y + \delta]]}{P[X \in [x, x + \delta, Y \in [y, y + \delta]]} \\
&= \lim_{\delta \to 0} \frac{\int_{-\infty}^{z} \int_{x}^{x+\delta} \int_{y}^{y+\delta} f_{XYZ}(x', y', z') \, dy' \, dx' \, dz'}{\int_{x}^{x+\delta} \int_{y}^{y+\delta} f_{XY}(x', y') \, dy' \, dx'} \\
&= \int_{-\infty}^{z} \frac{f_{XYZ}(x, y, z')}{f_{XY}(x, y)} \, dz'
\end{aligned}
$$

- Thus we get

$$f_{XYZ}(x, y, z) = f_{Z|XY}(z|x, y) f_{XY}(x, y) = f_{Z|XY}(z|x, y) f_{Y|X}(y|x) f_X(x)$$

- We can similarly talk about the joint distribution of any finite number of rv's
- Let $X_1, X_2, \cdots, X_n$ be rv's on the same probability space.
- We denote it as a vector $\mathbf{X}$ or $\underline{X}$. We can think of it as a mapping, $\mathbf{X} : \Omega \to \Re^n$.
- We can write the joint distribution as

$$F_{\mathbf{X}}(\mathbf{x}) = P[\mathbf{X} \leq \mathbf{x}] = P[X_i \leq x_i, \; i = 1, \cdots, n]$$

- We represent by $f_{\mathbf{X}}(\mathbf{x})$ the joint density or mass function. Sometimes we also write it as $f_{X_1 \cdots X_n}(x_1, \cdots, x_n)$
- We use similar notation for marginal and conditional distributions

Independence of multiple random variables

- Random variables $X_1, X_2, \cdots, X_n$ are said to be independent if the the events $[X_i \in B_i], \; i = 1, \cdots, n$ are independent.
  (Recall definition of independence of a set of events)
- Independence implies that the marginals would determine the joint distribution.

Example

- Let a joint density be given by

$$f_{XYZ}(x, y, z) = K, \quad 0 < z < y < x < 1$$

First let us determine K.

$$\begin{aligned}
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XYZ}(x, y, z) \; dz \; dy \; dx &= \int_0^1 \int_0^x \int_0^y K \; dz \; dy \; dx \\
&= K \int_0^1 \int_0^x y \; dy \; dx \\
&= K \int_0^1 \frac{x^2}{2} \; dx \\
&= K \frac{1}{6} \quad \Rightarrow K = 6
\end{aligned}$$

Example

- Let a joint density be given by

$$f_{XYZ}(x, y, z) = K, \quad 0 < z < y < x < 1$$

First let us determine K.

$$\begin{aligned}
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XYZ}(x, y, z) \; dz \; dy \; dx &= \int_{x=0}^1 \int_{y=0}^x \int_{z=0}^y K \; dz \; dy \; dx \\
&= K \int_{x=0}^1 \int_{y=0}^x y \; dy \; dx \\
&= K \int_0^1 \frac{x^2}{2} \; dx \\
&= K \frac{1}{6} \quad \Rightarrow K = 6
\end{aligned}$$

$$f_{XYZ}(x,y,z) = K, \quad 0 < z < y < x < 1$$

- Suppose we want to find the (marginal) joint distribution of X and Z.

$$\begin{aligned} f_{XZ}(x,z) &= \int_{-\infty}^{\infty} f_{XYZ}(x,y,z)\ dy \\ &= \int_{z}^{x} K\ dy, \quad 0 < z < x < 1 \\ &= 6(x-z), \quad 0 < z < x < 1 \end{aligned}$$

- We got the joint density as

$$f_{XZ}(x,z) = 6(x-z), \quad 0 < z < x < 1$$

- We can verify this is a joint density

$$\begin{aligned} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XZ}(x,z)\ dz\ dx &= \int_0^1 \int_0^x 6(x-z)\ dz\ dx \\ &= \int_0^1 \left(6x\ z|_0^x - 6\ \left.\frac{z^2}{2}\right|_0^x \right)\ dx \\ &= \int_0^1 \left(6x^2 - 6\ \frac{x^2}{2} \right)\ dx \\ &= 3\ \left.\frac{x^3}{3}\right|_0^1 = 1 \end{aligned}$$

- The joint density of X, Y, Z is

$$f_{XYZ}(x,y,z) = 6, \quad 0 < z < y < x < 1$$

- The joint density of X, Z is

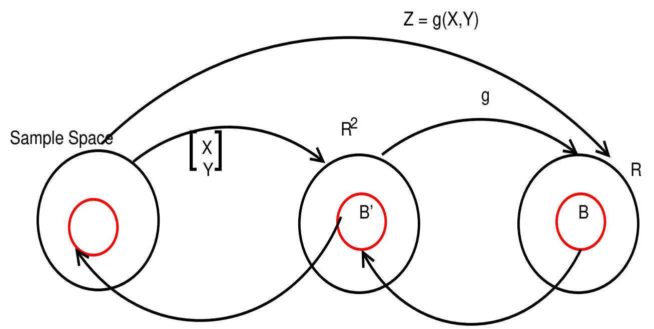
$$f_{XZ}(x,z) = 6(x-z), \quad 0 < z < x < 1$$

- Hence,

$$f_{Y|XZ}(y|x,z) = \frac{f_{XYZ}(x,y,z)}{f_{XZ}(x,z)} = \frac{1}{x-z}, \quad 0 < z < y < x < 1$$

Functions of multiple random variables

- Let X, Y be random variables on the same probability space.
- Let $g : \Re^2 \to \Re$.
- Let $Z = g(X,Y)$. Then Z is a rv
- This is analogous to functions of a single rv



- let $Z = g(X, Y)$
- We can determine distribution of Z from the joint distribution of X, Y

$$F_Z(z) = P[Z \le z] = P[g(X, Y) \le z]$$

- For example, if X, Y are discrete, then

$$f_Z(z) = P[Z = z] = P[g(X, Y) = z] = \sum_{\substack{x_i, y_j: \\ g(x_i, y_j) = z}} f_{XY}(x_i, y_j)$$

- Let X, Y be discrete rv's. Let $Z = \min(X, Y)$.

$$\begin{aligned} f_Z(z) &= P[\min(X, Y) = z] \\ &= P[X = z, Y > z] + P[Y = z, X > z] + P[X = Y = z] \\ &= \sum_{y > z} P[X = z, Y = y] + \sum_{x > z} P[X = x, Y = z] \\ &\qquad + P[X = z, Y = z] \\ &= \sum_{y > z} f_{XY}(z, y) + \sum_{x > z} f_{XY}(x, z) + f_{XY}(z, z) \end{aligned}$$

- Now suppose X, Y are independent and both of them have geometric distribution with the same parameter, p.
- Such random variables are called **independent and identically distributed** or **iid** random variables.

- Now we can get pmf of Z as (note $Z \in \{1, 2, \cdots\}$)

$$\begin{aligned} f_Z(z) &= P[X = z, Y > z] + P[Y = z, X > z] + P[X = Y = z] \\ &= P[X = z]P[Y > z] + P[Y = z]P[X > z] + P[X = z]P[Y = z] \\ &= p(1-p)^{z-1}(1-p)^z * 2 + \left(p(1-p)^{z-1}\right)^2 \\ &= 2p(1-p)^{z-1}(1-p)^z + \left(p(1-p)^{z-1}\right)^2 \\ &= 2p(1-p)^{2z-1} + p^2(1-p)^{2z-2} \\ &= p(1-p)^{2z-2}(2(1-p) + p) \\ &= (2-p)p(1-p)^{2z-2} \end{aligned}$$

- We can show this is a pmf

$$\begin{aligned} \sum_{z=1}^{\infty} f_Z(z) &= \sum_{z=1}^{\infty} (2-p)p(1-p)^{2z-2} \\ &= (2-p)p \sum_{z=1}^{\infty} (1-p)^{2z-2} \\ &= (2-p)p \, \frac{1}{1-(1-p)^2} \\ &= (2-p)p \, \frac{1}{2p - p^2} = 1 \end{aligned}$$

- Let us consider the max and min functions, in general.
- Let $Z = \max(X, Y)$. Then we have

$$\begin{aligned} F_Z(z) &= P[Z \leq z] = P[\max(X, Y) \leq z] \\ &= P[X \leq z, Y \leq z] \\ &= F_{XY}(z, z) \\ &= F_X(z) F_Y(z), \quad \text{if } X, Y \text{ are independent} \\ &= (F_X(z))^2, \quad \text{if they are iid} \end{aligned}$$

- This is true of all random variables.
- Suppose X, Y are iid continuous rv. Then density of Z is

$$f_Z(z) = 2F_X(z) f_X(z)$$

- Suppose X, Y are iid uniform over $(0,\ 1)$
- Then we get df and pdf of $Z = \max(X, Y)$ as

$$F_Z(z) = z^2, 0 < z < 1; \quad \text{and} \quad f_Z(z) = 2z, 0 < z < 1$$

$F_Z(z) = 0$ for $z \leq 0$ and $F_Z(z) = 1$ for $z \geq 1$ and $f_Z(z) = 0$ outside $(0, 1)$

- This is easily generalized to n radom variables.
- Let $Z = \max(X_1, \cdots, X_n)$

$$\begin{aligned} F_Z(z) &= P[Z \leq z] = P[\max(X_1, X_2, \cdots, X_n) \leq z] \\ &= P[X_1 \leq z, X_2 \leq z, \cdots, X_n \leq z] \\ &= F_{X_1 \cdots X_n}(z, \cdots, z) \\ &= F_{X_1}(z) \cdots F_{X_n}(z), \quad \text{if they are independent} \\ &= (F_X(z))^n, \quad \text{if they are iid} \\ & \qquad \text{where we take } F_X \text{ as the common df} \end{aligned}$$

- For example if all X_i are uniform over $(0, 1)$ and ind, then $F_Z(z) = z^n,\ 0 < z < 1$

- Consider $Z = \min(X, Y)$ and X, Y independent

$$F_Z(z) = P[Z \leq z] = P[\min(X, Y) \leq z]$$

- It is difficult to write this in terms of joint df of X, Y.
- So, we consider the following

$$\begin{aligned} P[Z > z] &= P[\min(X, Y) > z] \\ &= P[X > z, Y > z] \\ &= P[X > z] P[Y > z], \quad \text{using independence} \\ &= (1 - F_X(z))(1 - F_Y(z)) \\ &= (1 - F_X(z))^2, \quad \text{if they are iid} \end{aligned}$$

$$\text{Hence,} \quad F_Z(z) = 1 - (1 - F_X(z))(1 - F_Y(z))$$

- We can once again find density of Z if X, Y are continuous

- Suppose X, Y are iid uniform $(0,\ 1)$.
- $Z = \min(X, Y)$

$$F_Z(z) = 1 - (1 - F_X(z))^2 = 1 - (1 - z)^2, 0 < z < 1$$

- Notice that $P[X > z] = (1 - z)$.
- We get the density of Z as

$$f_Z(z) = 2(1 - z), \ \ 0 < z < 1$$

- $\min$ fn is also easily generalized to n random variables
- Let $Z = \min(X_1, X_2, \cdots, X_n)$

$$\begin{aligned} P[Z > z] &= P[\min(X_1, X_2, \cdots, X_n) > z] \\ &= P[X_1 > z, \cdots, X_n > z] \\ &= P[X_1 > z] \cdots P[X_n > z], \quad \text{using independence} \\ &= (1 - F_{X_1}(z)) \cdots (1 - F_{X_n}(z)) \\ &= (1 - F_X(z))^n, \quad \text{if they are iid} \end{aligned}$$

- Hence, when X_i are iid, the df of Z is

$$F_Z(z) = 1 - (1 - F_X(z))^n$$

where F_X is the common df

- Let X, Y be independent
- Let $Z = \max(X, Y)$ and $W = \min(X, Y)$.
- We want joint distribution function of Z and W.

$$F_{ZW}(z, w) = P[Z \le z, W \le w]$$

- This is difficult to find. But we can easily find

$$P[\max(X, Y) \le z, \min(X, Y) > w]$$

- Remaining details are left as an exercise for you!!

- Let $X, Y \in \{0, 1, \cdots\}$
- Let $Z = X + Y$. Then we have

$$\begin{aligned} f_Z(z) &= P[X + Y = z] = \sum_{\substack{x,y: \\ x+y=z}} P[X = x, Y = y] \\ &= \sum_{k=0}^{z} P[X = k, Y = z - k] \\ &= \sum_{k=0}^{z} f_{XY}(k, z - k) \end{aligned}$$

- Now suppose X, Y are independent. Then

$$f_Z(z) = \sum_{k=0}^{z} f_X(k) f_Y(z - k)$$

- Now suppose X, Y are independent Poisson with parameters λ_1, λ_2. And, $Z = X + Y$.

$$\begin{aligned} f_Z(z) &= \sum_{k=0}^{z} f_X(k) f_Y(z-k) \\ &= \sum_{k=0}^{z} \frac{\lambda_1^k}{k!} e^{-\lambda_1} \, \frac{\lambda_2^{z-k}}{(z-k)!} e^{-\lambda_2} \\ &= e^{-(\lambda_1+\lambda_2)} \, \frac{1}{z!} \sum_{k=0}^{z} \frac{z!}{k!(z-k)!} \, \lambda_1^k \, \lambda_2^{z-k} \\ &= e^{-(\lambda_1+\lambda_2)} \, \frac{1}{z!} (\lambda_1 + \lambda_2)^z \end{aligned}$$

- Z is Poisson with parameter $\lambda_1 + \lambda_2$

- Let X, Y have a joint density f_{XY}. Let $Z = X + Y$

$$\begin{aligned} F_Z(z) &= P[Z \leq z] = P[X + Y \leq z] \\ &= \int\int_{\{(x,y):x+y\leq z\}} f_{XY}(x,y) \; dy \; dx \\ &= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{z-x} f_{XY}(x,y) \; dy \; dx \\ & \quad \text{change of variable: } t = x + y \\ & \quad \quad dt = dy; \quad \text{when } (y = z - x), \; t = z \\ &= \int_{x=-\infty}^{\infty} \int_{t=-\infty}^{z} f_{XY}(x, t-x) \; dt \; dx \\ &= \int_{-\infty}^{z} \left(\int_{-\infty}^{\infty} f_{XY}(x, t-x) \; dx \right) \; dt \end{aligned}$$

- This gives us

$$f_Z(z) = \int_{-\infty}^{\infty} f_{XY}(x, z-x) \; dx$$

- X, Y have joint density f_{XY}. $Z = X + Y$. Then

$$f_Z(z) = \int_{-\infty}^{\infty} f_{XY}(x, z-x) \; dx$$

- Now suppose X and Y are independent. Then

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) \; f_Y(z-x) \; dx$$

  Density of sum of independent random variables is the convolution of their densities.

$$f_{X+Y} = f_X * f_Y \quad \text{(Convolution)}$$

- Suppose X, Y are iid exponential rv's.

$$f_X(x) = \lambda \, e^{-\lambda x}, \quad x > 0$$

- Let $Z = X + Y$. Then, density of Z is

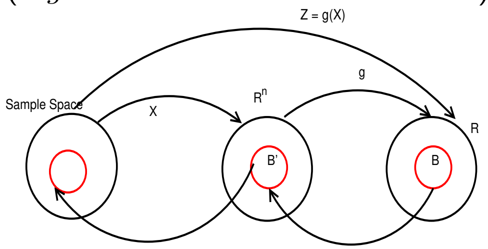
$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_X(x) \; f_Y(z-x) \; dx \\ &= \int_0^z \lambda \, e^{-\lambda x} \; \lambda \, e^{-\lambda(z-x)} \; dx \\ &= \lambda^2 \, e^{-\lambda z} \int_0^z dx = \lambda^2 \, z \, e^{-\lambda z} \end{aligned}$$

- Thus, sum of independent exponential random variables has gamma distribution:

$$f_Z(z) = \lambda z \; \lambda e^{-\lambda z}, \quad z > 0$$

Recap

- Given $X_1, \cdots, X_n$, random variables on the same probability space, $Z = g(X_1, \cdots, X_n)$ is a rv (if $g : \Re^n \rightarrow \Re$ is borel measurable).



- We can determine distribution of Z from the joint distribution of all X_i

$$F_Z(z) = P[Z \leq z] = P[g(X_1, \cdots, X_n) \leq z]$$

Recap

- $X_1, \cdots, X_n$ are said to be independent if events $[X_1 \in B_1], \cdots, [X_n \in B_n]$ are independent.
- If $X_1, \cdots, X_n$ are indepedent and all of them have the same distribution function then they are said to be iid – independent and identically distributed

Recap

- Let $X_1, \cdots, X_n$ be independent and $Z = \max(X_1, \cdots, X_n)$

$$\begin{aligned} F_Z(z) &= \prod_{i=1}^{n} F_{X_i}(z) \\ &= (F(z))^n, \quad \text{if they are iid} \end{aligned}$$

Recap

- Let $X_1, \cdots, X_n$ be independent and $Z = \min(X_1, \cdots, X_n)$

$$\begin{aligned} F_Z(z) &= 1 - \prod_{i=1}^{n} (1 - F_{X_i}(z)) \\ &= 1 - (1 - F(z))^n, \quad \text{if they are iid} \end{aligned}$$

Recap

- Let X, Y be random variables with joint density f_{XY}
- $Z = X + Y$

$$f_Z(z) = \int_{-\infty}^{\infty} f_{XY}(t, z - t)\ dt$$

- If X, Y are independent

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(t)\ f_Y(z - t)\ dt$$

  Density of sum of independent random variables is the convolution of their densities.
- Sum of independent exponential random variables has gamma density.

Recall problem from last class

- Let X, Y be independent
- Let $Z = \max(X, Y)$ and $W = \min(X, Y)$.
- We want joint distribution function of Z and W.

$$F_{ZW}(z, w) = P[Z \leq z, W \leq w]$$

- This is difficult to find. But we can easily find

$$P[\max(X, Y) \leq z, \min(X, Y) > w]$$

- Remaining details are left as an exercise for you!!

- X, Y iid with df F and density f
  $Z = \max(X, Y)$ and $W = \min(X, Y)$.
- We want joint distribution function of Z and W.
- We can use the following

$$P[Z \leq z] = P[Z \leq z, W \leq w] + P[Z \leq z, W > w]$$

$$\begin{aligned} P[Z \leq z, W > w] &= P[w < X, Y \leq z] = (F(z) - F(w))^2 \\ P[Z \leq z] &= P[X \leq z, Y \leq z] = (F(z))^2 \end{aligned}$$

- So, we get F_{ZW} as

$$\begin{aligned} F_{ZW}(z, w) &= P[Z \leq z, W \leq w] \\ &= P[Z \leq z] - P[Z \leq z, W > w] \\ &= (F(z))^2 - (F(z) - F(w))^2 \end{aligned}$$

- Is this correct for all values of z, w?

- We have $P[w < X, Y \leq z] = (F(z) - F(w))^2$ only when $w \leq z$.
- Otherwise it is zero.
- Hence we get F_{ZW} as

$$F_{ZW}(z, w) = \begin{cases} (F(z))^2 & \text{if} \quad w > z \\ (F(z))^2 - (F(z) - F(w))^2 & \text{if} \quad w \leq z \end{cases}$$

- We can get joint density of Z, W as

$$\begin{aligned} f_{ZW}(z, w) &= \frac{\partial^2}{\partial z\ \partial w} F_{ZW}(z, w) \\ &= 2f(z)f(w), \quad w \leq z \end{aligned}$$

Order Statistics

- Let $X_1, \cdots, X_n$ be iid with density f.
- Let $X_{(k)}$ denote the k^{th} smallest of these.
- That is, $X_{(k)} = g_k(X_1, \cdots, X_n)$ where $g_k : \Re^n \to \Re$ and the value of $g_k(x_1, \cdots, x_n)$ is the k^{th} smallest of the numbers $x_1, \cdots, x_n$.
- $X_{(1)} = \min(X_1, \cdots, X_n), \quad X_{(n)} = \max(X_1, \cdots, X_n)$
- The joint distribution of $X_{(1)}, \cdots X_{(n)}$ is called the order statistics.
- We calculated the order statistics for the case $n = 2$.
- It can be shown that

$$f_{X_{(1)} \cdots X_{(n)}}(x_1, \cdots x_n) = n! \prod_{i=1}^{n} f(x_i), \quad x_1 < x_2 < \cdots < x_n$$

- Let $X_1, \cdots, X_n$ be iid with df F and density f.
- $P[X_i \leq y] = F(y)$ for any i and y.
- Since they are independent, we have, e.g.,

$$P[X_1 \leq y, X_2 > y, X_3 \leq y] = (F(y))^2(1 - F(y))$$

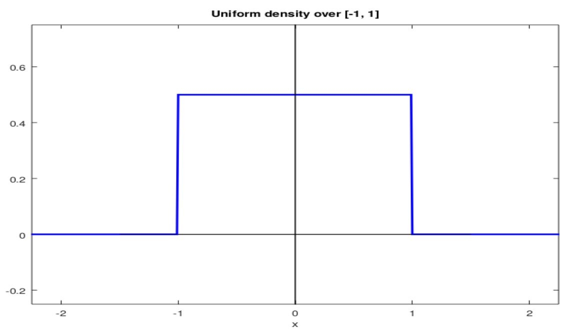
- Hence, probability that exactly k of these n random variables are less than or equal to y is ${}^nC_k(F(y))^k(1 - F(y))^{n-k}$
- Now the event $[X_{(k)} \leq y]$ is same as the event "at least k of these are less than or equal to y"
- Hence we get

$$F_{X_{(k)}}(y) \quad = \quad \sum_{j=k}^{n} {}^nC_j(F(y))^j(1 - F(y))^{n-j}$$

We can get the density by differentiating this.

Distribution of sums of independent rv

- Suppose X, Y are iid uniform over $(-1,\ 1)$.
- let $Z = X + Y$. We want f_Z.
- The density of X, Y is



- f_Z is convolution of this density with itself.

- $f_X(x) = 0.5, \quad -1 < x < 1.$ f_Y is also same
- Note that Z takes values in $[-2,\ 2]$

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(t)\ f_Y(z - t)\ dt$$

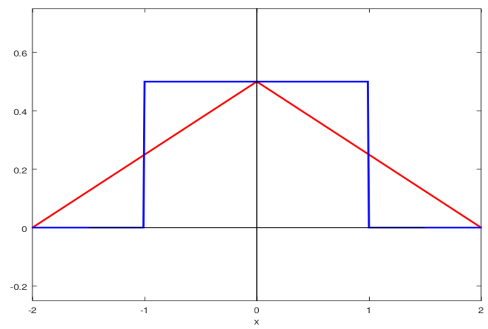
- For the integrand to be non-zero we need
  - $-1 < t < 1 \quad \Rightarrow \quad t < 1, \quad t > -1$
  - $-1 < z - t < 1 \quad \Rightarrow \quad t < z + 1, \quad t > z - 1$
  - Hence we need: $t < \min(1, z + 1), \quad t > \max(-1, z - 1)$
  - Hence, for $z < 0$, we need $-1 < t < z + 1$ and, for $z \geq 0$ we need $z - 1 < t < 1$
- Thus we get

$$f_Z(z) = \begin{cases} \int_{-1}^{z+1} \frac{1}{4} dt = \frac{z+2}{4} & \text{if} \quad -2 \leq z < 0 \\ \int_{z-1}^{1} \frac{1}{4} dt = \frac{2-z}{4} & \text{if} \quad 2 \geq z \geq 0 \end{cases}$$

- Thus, the density of sum of two ind rv's that are uniform over $(-1,\ 1)$ is

$$f_Z(z) = \begin{cases} \frac{z+2}{4} & \text{if } \ -2 < z < 0 \\ \frac{2-z}{4} & \text{if } \ 0 < z < 2 \end{cases}$$

- This is a triangle with vertices $(-2, 0), (0, 0.5), (2, 0)$



Independence of functions of random variable

- Suppose X and Y are independent.
- Then $g(X)$ and $h(Y)$ are independent
- This is because $[g(X) \in B_1] = [X \in \tilde{B}_1]$ for some Borel set, $\tilde{B}_1$ and similarly $[h(Y) \in B_2] = [Y \in \tilde{B}_2]$
- Hence, $[g(X) \in B_1]$ and $[h(Y) \in B_2]$ are independent.

Independence of functions of random variable

- This is easily generalized to functions of multiple random variables.
- If $\mathbf{X}, \mathbf{Y}$ are vector random variables (or random vectors), independence implies $[\mathbf{X} \in B_1]$ is independent of $[\mathbf{Y} \in B_2]$ for all borel sets B_1, B_2 (in appropriate spaces).
- Then $g(\mathbf{X})$ would be independent of $h(\mathbf{Y})$.
- That is, suppose $X_1, \cdots, X_m, Y_1, \cdots, Y_n$ are independent.
- Then, $g(X_1, \cdots, X_m)$ is independent of $h(Y_1, \cdots, Y_n)$.

- Let X_1, X_2, X_3 be independent continuous rv
- $Z = X_1 + X_2 + X_3$.
- Can we find density of Z?
- Let $W = X_1 + X_2$.
- Then $Z = W + X_3$ and W and X_3 are independent.
- Exercise for you: Find density of $X_1 + X_2 + X_3$ where X_1, X_2, X_3 are iid uniform over $(0,\ 1)$.

Sum of independent gamma rv

- Gamma density with parameters $\alpha > 0$ and $\lambda > 0$ is given by
$$f(x) = \frac{1}{\Gamma(\alpha)}\ \lambda^{\alpha}\ x^{\alpha-1}\ e^{-\lambda x}, \quad x > 0$$
We will call this $Gamma(\alpha, \lambda)$.
- The α is called the shape parameter and λ is called the rate parameter.
- For $\alpha = 1$ this is the exponential density.
- Let $X \sim Gamma(\alpha_1, \lambda)$, $Y \sim Gamma(\alpha_2, \lambda)$. Suppose X, Y are independent.
- Let $Z = X + Y$. Then $Z \sim Gamma(\alpha_1 + \alpha_2, \lambda)$.

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)\ f_Y(z-x)\ dx$$
$$= \int_0^z \frac{1}{\Gamma(\alpha_1)}\ \lambda^{\alpha_1}\ x^{\alpha_1-1}\ e^{-\lambda x} \frac{1}{\Gamma(\alpha_2)}\ \lambda^{\alpha_2}\ (z-x)^{\alpha_2-1}\ e^{-\lambda(z-x)}\ dx$$
$$= \frac{\lambda^{\alpha_1+\alpha_2}\ e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^z z^{\alpha_1-1} \left(\frac{x}{z}\right)^{\alpha_1-1} z^{\alpha_2-1} \left(1 - \frac{x}{z}\right)^{\alpha_2-1}\ dx$$

change the variable: $t = \frac{x}{z}$ $(\Rightarrow \ \ z^{-1}dx = dt)$

$$= \frac{\lambda^{\alpha_1+\alpha_2}\ e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\ z^{\alpha+\alpha_2-1} \int_0^1 t^{\alpha_1-1}\ (1-t)^{\alpha_2-1}\ dt$$
$$= \frac{1}{\Gamma(\alpha_1+\alpha_2)}\ \lambda^{\alpha_1+\alpha_2}\ z^{\alpha_1+\alpha_2-1}\ e^{-\lambda z}$$

Because
$$\int_0^1 t^{\alpha_1-1}\ (1-t)^{\alpha_2-1}\ dt = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}$$

- If X, Y are independent gamma random variables then $X + Y$ also has gamma distribution.
- If $X \sim Gamma(\alpha_1, \lambda)$, and $Y \sim Gamma(\alpha_2, \lambda)$, then $X + Y \sim Gamma(\alpha_1 + \alpha_2, \lambda)$.
- Exercise for you: Show that sum of independent Gaussian random variables has gaussian density.
- The algebra is a little involved.
- First take the two gaussians to be zero-mean.
- There is a calculation trick that is often useful with Gaussian density

A Calculation Trick

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2K}\left[x^2 - 2bx + c\right]\right)\ dx$$
$$= \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2K}\left[(x-b)^2 + c - b^2\right]\right)\ dx$$
$$= \int_{-\infty}^{\infty} \exp\left(-\frac{(x-b)^2}{2K}\right)\ \exp\left(-\frac{(c-b^2)}{2K}\right)\ dx$$
$$= \exp\left(-\frac{(c-b^2)}{2K}\right)\sqrt{2\pi K}$$

because
$$\frac{1}{\sqrt{2\pi K}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-b)^2}{2K}\right)\ dx = 1$$

- We next look at a general theorem that is quite useful in dealing with functions of multiple random variables.
- This result is only for continuous random variables.

- Let $X_1, \cdots, X_n$ be continuous random variables with joint density $f_{X_1 \cdots X_n}$. We define $Y_1, \cdots Y_n$ by

$$Y_1 = g_1(X_1, \cdots, X_n) \quad \cdots \quad Y_n = g_n(X_1, \cdots, X_n)$$

We think of g_i as components of $g : \Re^n \to \Re^n$.

- We assume g is continuous with continuous first partials and is invertible.
- Let h be the inverse of g. That is

$$X_1 = h_1(Y_1, \cdots, Y_n) \quad \cdots \quad X_n = h_n(Y_1, \cdots, Y_n)$$

- Each of g_i, h_i are $\Re^n \to \Re$ functions and we can write them as

$$y_i = g_i(x_1, \cdots, x_n); \quad \cdots \quad x_i = h_i(y_1, \cdots, y_n)$$

We denote the partial derivatives of these functions by $\frac{\partial x_i}{\partial y_j}$ etc.

- The jacobian of the inverse transformation is

$$J = \frac{\partial(x_1, \cdots, x_n)}{\partial(y_1, \cdots, y_n)} = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}$$

- We assume that J is non-zero in the range of the transformation
- **Theorem**: Under the above conditions, we have

$$f_{Y_1 \cdots Y_n}(y_1, \cdots, y_n) = |J| f_{X_1 \cdots X_n}(h_1(y_1, \cdots, y_n), \cdots, h_n(y_1, \cdots, y_n))$$

Or, more compactly, $f_{\mathbf{Y}}(\mathbf{y}) = |J| f_{\mathbf{X}}(h(\mathbf{y}))$

Proof of Theorem

- Let $B = (-\infty, \; y_1] \times \cdots \times (-\infty, \; y_n] \subset \Re^n$. Then

$$\begin{aligned} F_{\mathbf{Y}}(\mathbf{y}) &= F_{Y_1 \cdots Y_n}(y_1, \cdots y_n) = P[Y_i \le y_i, \; i = 1, \cdots, n] \\ &= \int_B f_{Y_1 \cdots Y_n}(y_1', \cdots, y_n') \, dy_1' \, \cdots \, dy_n' \end{aligned}$$

- Define

$$\begin{aligned} g^{-1}(B) &= \{(x_1, \cdots, x_n) \in \Re^n \; : \; g(x_1, \cdots, x_n) \in B\} \\ &= \{(x_1, \cdots, x_n) \in \Re^n \; : \; g_i(x_1, \cdots, x_n) \le y_i, \; i = 1 \cdots n\} \end{aligned}$$

- Then we have

$$\begin{aligned} F_{Y_1 \cdots Y_n}(y_1, \cdots y_n) &= P[g_i(X_1, \cdots, X_n) \le y_i, \; i = 1, \cdots n] \\ &= \int_{g^{-1}(B)} f_{X_1 \cdots X_n}(x_1', \cdots, x_n') \, dx_1' \, \cdots \, dx_n' \end{aligned}$$

Proof of Theorem

- $B = (-\infty,\ y_1] \times \cdots \times (-\infty,\ y_n]$.
- $g^{-1}(B) = \{(x_1, \cdots, x_n) \in \Re^n\ :\ g(x_1, \cdots, x_n) \in B\}$

$$\begin{aligned} F_{\mathbf{Y}}(y_1, \cdots, y_n) &= P[g_i(X_1, \cdots, X_n) \le y_i,\ i = 1, \cdots, n] \\ &= \int_{g^{-1}(B)} f_{X_1 \cdots X_n}(x'_1, \cdots, x'_n)\ dx'_1\ \cdots\ dx'_n \end{aligned}$$

change variables: $y'_i = g_i(x'_1, \cdots, x'_n), i = 1, \cdots n$

$$(x'_1, \cdots x'_n) \in g^{-1}(B) \ \Rightarrow\ (y'_1, \cdots, y'_n) \in B$$

$$x'_i = h_i(y'_1, \cdots, y'_n), \quad dx'_1 \cdots dx'_n = |J| dy'_1 \cdots dy'_n$$

$$F_{\mathbf{Y}}(y_1, \cdots, y_n) = \int_B f_{X_1 \cdots X_n}(h_1(\mathbf{y}'), \cdots, h_n(\mathbf{y}'))\ |J| dy'_1 \cdots dy'_n$$

$$\Rightarrow\ f_{Y_1 \cdots Y_n}(y_1, \cdots, y_n) = f_{X_1 \cdots X_n}(h_1(\mathbf{y}), \cdots, h_n(\mathbf{y}))\ |J|$$

- $X_1, \cdots X_n$ are continuous rv with joint density

$$Y_1 = g_1(X_1, \cdots, X_n) \quad \cdots \quad Y_n = g_n(X_1, \cdots, X_n)$$

- The transformation is continuous with continuous first partials and is invertible and

$$X_1 = h_1(Y_1, \cdots, Y_n) \quad \cdots \quad X_n = h_n(Y_1, \cdots, Y_n)$$

- We assume the Jacobian of the inverse transform, J, is non-zero
- Then the density of $\mathbf{Y}$ is

$$f_{Y_1 \cdots Y_n}(y_1, \cdots, y_n) = |J| f_{X_1 \cdots X_n}\left(h_1(y_1, \cdots, y_n), \cdots, h_n(y_1, \cdots, y_n)\right)$$

- Called multidimensional change of variable formula

- Let X, Y have joint density f_{XY}. Let $Z = X + Y$.
- We want f_Z. For the theorem we need two functions.
- To use the theorem, we need an invertible transformation of $\Re^2$ onto $\Re^2$ of which one component is $x + y$.
- Take $Z = X + Y$ and $W = X - Y$. This is an invertible.
- $X = (Z + W)/2$ and $Y = (Z - W)/2$. The Jacobian is

$$J = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}$$

- Hence we get

$$f_{ZW}(z, w) = \frac{1}{2} f_{XY}\left(\frac{z+w}{2}, \frac{z-w}{2}\right)$$

- Now we get density of Z as

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{2} f_{XY}\left(\frac{z+w}{2}, \frac{z-w}{2}\right)\ dw$$

- let $Z = X + Y$ and $W = X - Y$. Then

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{2} f_{XY}\left(\frac{z+w}{2}, \frac{z-w}{2}\right)\ dw$$

change the variable: $t = \dfrac{z+w}{2} \quad \Rightarrow dt = \dfrac{1}{2}\ dw$

$$\Rightarrow \quad w = 2t - z \quad \Rightarrow z - w = 2z - 2t$$

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_{XY}(t, z - t)\ dt \\ &= \int_{-\infty}^{\infty} f_{XY}(z - t, t)\ dt, \quad \text{by using } t = \frac{z-w}{2} \text{ above} \end{aligned}$$

- We get same result as earlier. If, X, Y are independent

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(t)\ f_Y(z - t)\ dt$$

- let $Z = X + Y$ and $W = X - Y$. We got

$$f_{ZW}(z, w) = \frac{1}{2} f_{XY}\left(\frac{z+w}{2}, \frac{z-w}{2}\right)$$

- Now we can calculate f_W also.

$$f_W(w) = \int_{-\infty}^{\infty} \frac{1}{2} f_{XY}\left(\frac{z+w}{2}, \frac{z-w}{2}\right) dz$$

change the variable: $t = \frac{z+w}{2} \quad \Rightarrow dt = \frac{1}{2}\, dz$

$$\Rightarrow \quad z = 2t - w \quad \Rightarrow z - w = 2t - 2w$$

$$f_W(w) = \int_{-\infty}^{\infty} f_{XY}(t, t-w)\, dt$$

$$= \int_{-\infty}^{\infty} f_{XY}(t+w, t) dt, \quad \text{using } t = \frac{z-w}{2} \text{ above}$$

Example

- Let X, Y be iid $U(0,\ 1)$. Let $Z = X - Y$.

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(t)\ f_Y(t-z)\ dt$$

- For the integrand to be non-zero (note $Z \in (-1, 1)$)
  - $0 < t < 1 \ \Rightarrow \ t > 0,\ t < 1$
  - $0 < t - z < 1 \ \Rightarrow \ t > z,\ t < 1 + z$
  - $\Rightarrow \quad \max(0,\ z) < t < \min(1,\ 1+z)$
- Thus, we get density as

$$f_Z(z) = \begin{cases} \int_0^{1+z} 1\ dt = 1 + z, & \text{if } -1 < z < 0 \\ \int_z^1 1\ dt = 1 - z, & 0 < z < 1 \end{cases}$$

- This we have when $X, Y \sim U(0, 1)$ iid

$$f_{X-Y}(z) = 1 - |z|, \quad -1 < z < 1$$

- We showed that

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_{XY}(t, z-t)\ dt = \int_{-\infty}^{\infty} f_{XY}(z-t, t)\ dt$$

$$f_{X-Y}(w) = \int_{-\infty}^{\infty} f_{XY}(t, t-w)\ dt = \int_{-\infty}^{\infty} f_{XY}(t+w, t) dt$$

- Suppose X, Y are discrete. Then we have

$$f_{X+Y}(z) = P[X+Y=z] = \sum_k P[X=k, Y=z-k]$$

$$= \sum_k f_{XY}(k, z-k)$$

$$f_{X-Y}(w) = P[X-Y=w] = \sum_k P[X=k, Y=k-w]$$

$$= \sum_k f_{XY}(k, k-w)$$

Distribution of product of random variables

- We want density of $Z = XY$.
- We need one more function to make an invertible transformation
- A possible choice: $Z = XY \quad W = Y$
- This is invertible: $X = Z/W \quad Y = W$

$$J = \begin{vmatrix} \frac{1}{w} & \frac{-z}{w^2} \\ 0 & 1 \end{vmatrix} = \frac{1}{w}$$

- Hence we get

$$f_{ZW}(z, w) = \left|\frac{1}{w}\right|\ f_{XY}\left(\frac{z}{w}, w\right)$$

- Thus we get the density of product as

$$f_Z(z) = \int_{-\infty}^{\infty} \left|\frac{1}{w}\right|\ f_{XY}\left(\frac{z}{w}, w\right)\ dw$$

example

- Let X, Y be iid $U(0,\ 1)$. Let $Z = XY$.

$$f_Z(z) = \int_{-\infty}^{\infty} \left|\frac{1}{w}\right| \ f_X\left(\frac{z}{w}\right) \ f_Y(w)\ dw$$

- We need: $0 < w < 1$ and $0 < \frac{z}{w} < 1$. Hence

$$f_Z(z) = \int_z^1 \left|\frac{1}{w}\right| \ dw = \int_z^1 \frac{1}{w} \ dw = -\ln(z), \ \ 0 < z < 1$$

- X, Y have joint density and $Z = XY$. Then

$$f_Z(z) = \int_{-\infty}^{\infty} \left|\frac{1}{w}\right| \ f_{XY}\left(\frac{z}{w}.w\right) \ dw$$

Suppose X, Y are discrete and $Z = XY$

$$f_Z(0) \ = \ P[X = 0 \text{ or } Y = 0] = \sum_x f_{XY}(x, 0) + \sum_y f_{XY}(0, y$$

$$f_Z(k) \ = \ \sum_{y \neq 0} P\left[X = \frac{k}{y}, Y = y\right] = \sum_{y \neq 0} f_{XY}\left(\frac{k}{y}, y\right), \ k \neq 0$$

- We cannot always interchange density and mass functions!!

- We wanted density of $Z = XY$.
- We used: $Z = XY$ and $W = Y$.
- We could have used: $Z = XY$ and $W = X$.
- This is invertible: $X = W$ and $Y = Z/W$.

$$J = \begin{vmatrix} 0 & 1 \\ \frac{1}{w} & \frac{-z}{w^2} \end{vmatrix} = -\frac{1}{w}$$

- This gives

$$f_{ZW}(z, w) \ = \ \left|\frac{1}{w}\right| \ f_{XY}\left(w, \frac{z}{w}\right)$$

$$f_Z(z) \ = \ \int_{-\infty}^{\infty} \left|\frac{1}{w}\right| \ f_{XY}\left(w, \frac{z}{w}\right) \ dw$$

- The f_Z should be same in both cases.

Distributions of quotients

- X, Y have joint density and $Z = X/Y$.
- We can take: $Z = X/Y \quad W = Y$
- This is invertible: $X = ZW \quad Y = W$

$$J = \begin{vmatrix} w & z \\ 0 & 1 \end{vmatrix} = w$$

- Hence we get

$$f_{ZW}(z, w) = |w| \ f_{XY}(zw, w)$$

- Thus we get the density of quotient as

$$f_Z(z) = \int_{-\infty}^{\infty} |w| \ f_{XY}(zw, w)\ dw$$

example

- Let X, Y be iid $U(0,\ 1)$. Let $Z = X/Y$.
  Note $Z \in (0, \infty)$

$$f_Z(z) = \int_{-\infty}^{\infty} |w|\ f_X(zw)\ f_Y(w)\ dw$$

- We need $0 < w < 1$ and $0 < zw < 1 \ \Rightarrow w < 1/z$.
- So, when $z \leq 1$, w goes from 0 to 1; when $z > 1$, w goes from 0 to $1/z$.
- Hence we get density as

$$f_Z(z) = \begin{cases} \int_0^1 w\ dw = \frac{1}{2}, & \text{if } 0 < z \leq 1 \\ \int_0^{1/z} w\ dw = \frac{1}{2z^2}, & 1 < z < \infty \end{cases}$$

- X, Y have joint density and $Z = X/Y$

$$f_Z(z) = \int_{-\infty}^{\infty} |w|\ f_{XY}(zw, w)\ dw$$

- SupposeX, Y are discrete and $Z = X/Y$

$$\begin{aligned} f_Z(z) &= P[Z = z] = P[X/Y = z] \\ &= \sum_y P[X = yz, Y = y] \\ &= \sum_y f_{XY}(yz, y) \end{aligned}$$

- We chose: $Z = X/Y$ and $W = Y$.
- We could have taken: $Z = X/Y$ and $W = X$
- The inverse is: $X = W$ and $Y = W/Z$

$$J = \begin{vmatrix} 0 & 1 \\ -\frac{w}{z^2} & \frac{1}{z} \end{vmatrix} = -\frac{w}{z^2}$$

- Thus we get the density of quotient as

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} \left|\frac{w}{z^2}\right|\ f_{XY}\left(w, \frac{w}{z}\right)\ dw \\ &\quad \text{put } \ t = \frac{w}{z} \ \ \Rightarrow dt = \frac{dw}{z}, \quad w = tz \\ &= \int_{-\infty}^{\infty} |t| f_{XY}(tz, t)\ dt \end{aligned}$$

- We can show that the density of quotient is same in both these approches.

Exchangeable Random Variables

- $X_1, X_2, \cdots, X_n$ are said to be exchangeable if their joint distribution is same as that of any permutation of them.
- let $(i_1, \cdots, i_n)$ be a permutation of $(1, 2, \cdots, n)$. Then joint df of $(X_{i_1}, \cdots, X_{i_n})$ should be same as that $(X_1, \cdots, X_n)$
- Take $n = 3$. Suppose $F_{X_1X_2X_3}(a, b, c) = g(a, b, c)$. If they are exchangeable, then

$$\begin{aligned} F_{X_2X_3X_1}(a, b, c) &= P[X_2 \leq a, X_3 \leq b, X_1 \leq c] \\ &= P[X_1 \leq c, X_2 \leq a, X_3 \leq b] \\ &= g(c, a, b) = g(a, b, c) \end{aligned}$$

- The df or density should be “symmetric” in its variables if the random variables are exchangeable.

- Consider the density of three random variables

$$f(x,y,z) = \frac{2}{3}(x+y+z), \quad 0 < x,y,z < 1$$

- They are exchangeable (because $f(x,y,z) = f(y,x,z)$)
- If random variables are exchangeable then they are identically distributed.
$F_{XYZ}(a,\infty,\infty) = F_{XYZ}(\infty,\infty,a) \quad \Rightarrow \quad F_X(a) = F_Z(a)$
- The above example shows that exchangeable random variables need not be independent. The joint density is not factorizable.

$$\int_0^1 \int_0^1 \frac{2}{3}(x+y+z)\, dy\, dz = \frac{2(x+1)}{3}$$

- So, the joint density is not the product of marginals

Expectation of functions of multiple rv

- **Theorem**: Let $Z = g(X_1, \cdots X_n) = g(\mathbf{X})$. Then

$$E[Z] = \int_{\Re^n} g(\mathbf{x})\, dF_{\mathbf{X}}(\mathbf{x})$$

- That is, if they have a joint density, then

$$E[Z] = \int_{\Re^n} g(\mathbf{x})\, f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}$$

- Similarly, if all X_i are discrete

$$E[Z] = \sum_{\mathbf{x}} g(\mathbf{x})\, f_{\mathbf{X}}(\mathbf{x})$$

- Let $Z = X + Y$. Let X, Y have joint density f_{XY}

$$\begin{aligned} E[X+Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y)\, f_{XY}(x,y)\, dx\, dy \\ &= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{XY}(x,y)\, dy\, dx \\ &\quad + \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{XY}(x,y)\, dx\, dy \\ &= \int_{-\infty}^{\infty} x\, f_X(x)\, dx + \int_{-\infty}^{\infty} y\, f_Y(y)\, dy \\ &= E[X] + E[Y] \end{aligned}$$

- Expectation is a linear operator.
- This is true for all random variables.

Recap

- $X_1, \cdots X_n$ are continuous rv with joint density

$$Y_1 = g_1(X_1, \cdots, X_n) \quad \cdots \quad Y_n = g_n(X_1, \cdots, X_n)$$

- The transformation is continuous with continuous first partials and is invertible and

$$X_1 = h_1(Y_1, \cdots, Y_n) \quad \cdots \quad X_n = h_n(Y_1, \cdots, Y_n)$$

- We assume the Jacobian of the inverse transform, J, is non-zero
- Then the density of $\mathbf{Y}$ is

$$f_{Y_1 \cdots Y_n}(y_1, \cdots, y_n) = |J| f_{X_1 \cdots X_n}\left(h_1(y_1, \cdots, y_n), \cdots, h_n(y_1, \cdots, y_n)\right)$$

- Called multidimensional change of variable formula

Recap

- One can use the theorem to find densities of sum, difference, product and quotient of random variables.

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_{XY}(t, z-t)\, dt = \int_{-\infty}^{\infty} f_{XY}(z-t, t)\, dt$$

$$f_{X-Y}(z) = \int_{-\infty}^{\infty} f_{XY}(t, t-z)\, dt = \int_{-\infty}^{\infty} f_{XY}(t+z, t) dt$$

$$f_{X*Y}(z) = \int_{-\infty}^{\infty} \left|\frac{1}{t}\right| f_{XY}\left(\frac{z}{t}, t\right)\, dt = \int_{-\infty}^{\infty} \left|\frac{1}{t}\right| f_{XY}\left(t, \frac{z}{t}\right)\, dt$$

$$f_{(X/Y)}(z) = \int_{-\infty}^{\infty} |t|\; f_{XY}(zt, t)\, dt = \int_{-\infty}^{\infty} \left|\frac{t}{z^2}\right| f_{XY}\left(t, \frac{t}{z}\right)\, dt$$

Recap

- $X_1, X_2, \cdots, X_n$ are said to be exchangeable if their joint distribution is same as that of any permutation of them.
- If the random variables are exchangeable then the joint distribution function remains the same on permutation of arguments.
- Exchangeable random variables are identically distributed but they may not be independent.

Recap

- Let $Z = g(X_1, \cdots X_n) = g(\mathbf{X})$. Then

$$E[Z] = \int_{\Re^n} g(\mathbf{x})\, dF_{\mathbf{X}}(\mathbf{x})$$

- For example, if they have a joint density, then

$$E[Z] = \int_{\Re^n} g(\mathbf{x})\, f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}$$

- This gives us: $E[X+Y] = E[X] + E[Y]$
- In general, $E\left[g_1(\mathbf{X}) + g_2(\mathbf{X})\right] = E\left[g_1(\mathbf{X}\right] + E\left[g_2(\mathbf{X})\right]$

- We saw $E[X+Y] = E[X] + E[Y]$.
- Let us calculate $\text{Var}(X+Y)$.

$$\begin{aligned}
\text{Var}(X+Y) &= E\left[((X+Y) - E[X+Y])^2\right] \\
&= E\left[((X-EX) + (Y-EY))^2\right] \\
&= E\left[(X-EX)^2\right] + E\left[(Y-EY)^2\right] \\
&\qquad +2E\left[(X-EX)(Y-EY)\right] \\
&= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)
\end{aligned}$$

where we define **covariance** between X, Y as

$$\text{Cov}(X,Y) = E\left[(X-EX)(Y-EY)\right]$$

- We define **covariance** between X and Y by

$$\begin{aligned} \text{Cov}(X,Y) &= E\left[(X-EX)(Y-EY)\right] \\ &= E\left[XY - X(EY) - Y(EX) + EX\ EY\right] \\ &= E[XY] - EX\ EY \end{aligned}$$

- Note that $\text{Cov}(X,Y)$ can be positive or negative
- X and Y are said to be uncorrelated if $\text{Cov}(X,Y) = 0$
- If X and Y are uncorrelated then

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$$

- Note that $E[X+Y] = E[X] + E[Y]$ for all random variables.

Example

- Consider the joint density

$$f_{XY}(x,y) = 2, \quad 0 < x < y < 1$$

- We want to calculate $\text{Cov}(X,Y)$

$$EX = \int_0^1 \int_x^1 x\ 2\ dy\ dx = 2\int_0^1 x\ (1-x)\ dx = \frac{1}{3}$$

$$EY = \int_0^1 \int_0^y y\ 2\ dx\ dy = 2\int_0^1 y^2\ dy = \frac{2}{3}$$

$$E[XY] = \int_0^1 \int_0^y xy\ 2\ dx\ dy = 2\int_0^1 y\ \frac{y^2}{2}\ dy = \frac{1}{4}$$

- Hence, $\text{Cov}(X,Y) = E[XY] - EX\ EY = \frac{1}{4} - \frac{2}{9} = \frac{1}{36}$

Independent random variables are uncorrelated

- Suppose X, Y are independent. Then

$$\begin{aligned} E[XY] &= \int\int x\ y\ f_{XY}(x,y)\ dx\ dy \\ &= \int\int x\ y\ f_X(x)\ f_Y(y)\ dx\ dy \\ &= \int x f_X(x)\ dx \int y f_Y(y)\ dy = EX\ EY \end{aligned}$$

- Then, $\text{Cov}(X,Y) = E[XY] - EX\ EY = 0.$
- X, Y independent $\Rightarrow$ X, Y uncorrelated

Uncorrelated random variables may not be independent

- Suppose $X \sim \mathcal{N}(0,1)$ Then, $EX = EX^3 = 0$
- Let $Y = X^2$ Then,

$$E[XY] = EX^3 = 0 = EX\ EY$$

- Thus X, Y are uncorrelated.
- Are they independent? No
  e.g.,

$$P[X > 2\ |Y < 1] = 0 \neq P[X > 2]$$

- X, Y are uncorrealted does not imply they are independent.

- We define the **correlation coefficient** of X, Y by

$$\rho_{XY} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\ \text{Var}(Y)}}$$

- If X, Y are uncorrelated then $\rho_{XY} = 0$.
- We will show that $|\rho_{XY}| \leq 1$
- Hence $-1 \leq \rho_{XY} \leq 1, \ \forall X, Y$

- We have $E\left[(\alpha X + \beta Y)^2\right] \geq 0, \ \forall \alpha, \beta \in \Re$

$$\begin{aligned}
\alpha^2 E[X^2] + \beta^2 E[Y^2] + 2\alpha\beta E[XY] &\geq 0, \quad \forall \alpha, \beta \in \Re \\
\text{Take} \quad \alpha = -\frac{E[XY]}{E[X^2]} & \\
\frac{(E[XY])^2}{E[X^2]} + \beta^2 E[Y^2] - 2\beta \frac{(E[XY])^2}{E[X^2]} &\geq 0, \quad \forall \beta \in \Re \\
\Rightarrow \ 4\left(\frac{(E[XY])^2}{E[X^2]}\right)^2 - 4E[Y^2]\frac{(E[XY])^2}{E[X^2]} &\leq 0 \\
\Rightarrow (E[XY])^2 \leq E[X^2]E[Y^2] &
\end{aligned}$$

- We showed that

$$(E[XY])^2 \leq E[X^2]E[Y^2]$$

- Take $X - EX$ in place of X and $Y - EY$ in place of Y in the above algebra.
- This gives us

$$(E[(X - EX)(Y - EY)])^2 \leq E[(X - EX)^2]E[(Y - EY)^2]$$

$$\Rightarrow \quad (\text{Cov}(X,Y))^2 \leq \text{Var}(X)\text{Var}(Y)$$

- Hence we get

$$\rho_{XY}^2 = \left(\frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}\right)^2 \leq 1$$

- The equality holds here only if $E\left[(\alpha X + \beta Y)^2\right] = 0$

Thus, $\quad |\rho_{XY}| = 1$ only if $\alpha X + \beta Y = 0$

- Correlation coefficient of X, Y is ± 1 only when Y is a linear function of X

Linear Least Squares Estimation

- Suppose we want to approximate Y as an affine function of X.
- We want a, b to minimize $E\left[(Y - (aX + b))^2\right]$
- For a fixed a, what is the b that minimizes $E\left[((Y - aX) - b)^2\right]$?
- We know the best b here is: $b = E[Y - aX] = EY - aEX$.
- So, we want to find the best a to minimize $J(a) = E\left[(Y - aX - (EY - aEX))^2\right]$

- We want to find a to minimize

$$\begin{aligned} J(a) &= E\left[(Y - aX - (EY - aEX))^2\right] \\ &= E\left[((Y - EY) - a(X - EX))^2\right] \\ &= E\left[(Y - EY)^2 + a^2(X - EX)^2 - 2a(Y - EY)(X - EX)\right] \\ &= \text{Var}(Y) + a^2\text{Var}(X) - 2a\text{Cov}(X, Y) \end{aligned}$$

- So, the optimal a satisfies

$$2a\text{Var}(X) - 2\text{Cov}(X, Y) = 0 \quad \Rightarrow \quad a = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

- The final mean square error, say, J^* is

$$\begin{aligned} J^* &= \text{Var}(Y) + a^2\text{Var}(X) - 2a\text{Cov}(X, Y) \\ &= \text{Var}(Y) + \left(\frac{\text{Cov}(X, Y)}{\text{Var}(X)}\right)^2 \text{Var}(X) - 2\frac{\text{Cov}(X, Y)}{\text{Var}(X)}\text{Cov}(X, Y) \\ &= \text{Var}(Y) - \frac{(\text{Cov}(X, Y))^2}{\text{Var}(X)} \\ &= \text{Var}(Y)\left(1 - \frac{(\text{Cov}(X, Y))^2}{\text{Var}(Y)\,\text{Var}(X)}\right) \\ &= \text{Var}(Y)\left(1 - \rho_{XY}^2\right) \end{aligned}$$

- The best mean-square approximation of Y as a 'linear' function of X is

$$Y = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}X + \left(EY - \frac{\text{Cov}(X, Y)}{\text{Var}(X)}EX\right)$$

- Called the line of regression of Y on X.
- If $\text{cov}(X, Y) = 0$ then this reduces to approximating Y by a constant, EY.
- The final mean square error is

$$\text{Var}(Y)\left(1 - \rho_{XY}^2\right)$$

- If $\rho_{XY} = \pm 1$ then the error is zero
- If $\rho_{XY} = 0$ the final error is $\text{Var}(Y)$

- The covariance of X, Y is

$$\text{Cov}(X, Y) = E[(X - EX)\,(Y - EY)] = E[XY] - EX\,EY$$

Note that $\text{Cov}(X, X) = \text{Var}(X)$

- X, Y are called uncorrelated if $\text{Cov}(X, Y) = 0$.
- X, Y independent $\Rightarrow$ X, Y uncorrelated.
- Uncorrelated random variables need not necessarily be independent
- Covariance plays an important role in linear least squares estimation.
- Informally, covariance captures the 'linear dependence' between the two random variables.

Covariance Matrix

- Let $X_1, \cdots, X_n$ be random variables (on the same probability space)
- We represent them as a vector $\mathbf{X}$.
- As a notation, all vectors are column vectors: $\mathbf{X} = (X_1, \cdots, X_n)^T$
- We denote $E[\mathbf{X}] = (EX_1, \cdots, EX_n)^T$
- The $n \times n$ matrix whose $(i,j)^{th}$ element is Cov(X_i, X_j) is called the covariance matrix (or variance-covariance matrix) of $\mathbf{X}$. Denoted as $\Sigma_{\mathbf{X}}$ or Σ_X

$$\Sigma_{\mathbf{X}} = \begin{bmatrix} \text{Cov}(X_1,X_1) & \text{Cov}(X_1,X_2) & \cdots & \text{Cov}(X_1,X_n) \\ \text{Cov}(X_2,X_1) & \text{Cov}(X_2,X_2) & \cdots & \text{Cov}(X_2,X_n) \\ \vdots & \vdots & \vdots & \vdots \\ \text{Cov}(X_n,X_1) & \text{Cov}(X_n,X_2) & \cdots & \text{Cov}(X_n,X_n) \end{bmatrix}$$

Covariance matrix

- If $\mathbf{a} = (a_1, \cdots, a_n)^T$ then $\mathbf{a}\,\mathbf{a}^T$ is a $n \times n$ matrix whose $(i,j)^{th}$ element is $a_i a_j$.
- Hence we get

$$\Sigma_{\mathbf{X}} = E\left[(\mathbf{X} - E\mathbf{X})\,(\mathbf{X} - E\mathbf{X})^T\right]$$

- This is because $\left((\mathbf{X} - E\mathbf{X})\,(\mathbf{X} - E\mathbf{X})^T\right)_{ij} = (X_i - EX_i)(X_j - EX_j)$ and $(\Sigma_{\mathbf{X}})_{ij} = E[(X_i - EX_i)(X_j - EX_j)]$

- Recall the following about vectors and matrices
- let $\mathbf{a}, \mathbf{b} \in \Re^n$ be column vectors. Then

$$\left(\mathbf{a}^T\mathbf{b}\right)^2 = \left(\mathbf{a}^T\mathbf{b}\right)^T\left(\mathbf{a}^T\mathbf{b}\right) = \mathbf{b}^T\mathbf{a}\,\mathbf{a}^T\mathbf{b} = \mathbf{b}^T\left(\mathbf{a}\,\mathbf{a}^T\right)\mathbf{b}$$

- Let A be an $n \times n$ matrix with elements a_{ij}. Then

$$\mathbf{b}^T A \mathbf{b} = \sum_{i,j=1}^{n} b_i b_j a_{ij}$$

where $\mathbf{b} = (b_1, \cdots, b_n)^T$
- A is said to be positive semidefinite if $\mathbf{b}^T A \mathbf{b} \geq 0, \ \forall \mathbf{b}$

- Σ_X is a real symmetric matrix
- It is positive semidefinite.
- Let $\mathbf{a} \in \Re^n$ and let $Y = \mathbf{a}^T\mathbf{X}$.
- Then, $EY = \mathbf{a}^T E\mathbf{X}$. We get variance of Y as

$$\begin{aligned} \text{Var}(Y) &= E[(Y - EY)^2] = E\left[\left(\mathbf{a}^T\mathbf{X} - \mathbf{a}^T E\mathbf{X}\right)^2\right] \\ &= E\left[\left(\mathbf{a}^T(\mathbf{X} - E\mathbf{X})\right)^2\right] \\ &= E\left[\mathbf{a}^T(\mathbf{X} - E\mathbf{X})\,(\mathbf{X} - E\mathbf{X})^T\mathbf{a}\right] \\ &= \mathbf{a}^T E\left[(\mathbf{X} - E\mathbf{X})\,(\mathbf{X} - E\mathbf{X})^T\right]\mathbf{a} \\ &= \mathbf{a}^T \Sigma_X \mathbf{a} \end{aligned}$$

- This gives $\mathbf{a}^T \Sigma_X \mathbf{a} \geq 0, \ \forall \mathbf{a}$
- This shows Σ_X is positive semidefinite

- $Y = \mathbf{a}^T\mathbf{X} = \sum_i a_i X_i$ – linear combination of X_i's.
- We know how to find its mean and variance

$$EY = \mathbf{a}^T E\mathbf{X} = \sum_i a_i EX_i;$$
$$\text{Var}(Y) = \mathbf{a}^T \Sigma_X \mathbf{a} = \sum_{i,j} a_i a_j \text{Cov}(X_i, X_j)$$

- Specifically, by taking all components of $\mathbf{a}$ to be 1, we get

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i,j=1}^n \text{Cov}(X_i, X_j) = \sum_{i=1}^n \text{Var}(X_i) + \sum_{i=1}^n \sum_{j \neq i} \text{Cov}(X_i, X_j)$$

- If X_i are independent, variance of sum is sum of variances.

- Covariance matrix Σ_X positive semidefinite because

$$\mathbf{a}^T \Sigma_X \mathbf{a} = \text{Var}(\mathbf{a}^T\mathbf{X}) \geq 0$$

- Σ_X would be positive definite if $\mathbf{a}^T \Sigma_X \mathbf{a} > 0, \ \forall \mathbf{a} \neq 0$
- It would fail to be positive definite if $\text{Var}(\mathbf{a}^T\mathbf{X}) = 0$ for some nonzero $\mathbf{a}$.
- $\text{Var}(Z) = E[(Z - EZ)^2] = 0$ implies $Z = EZ$, a constant.
- Hence, Σ_X fails to be positive definite only if there is a non-zero linear combination of X_i's that is a constant.

- Covariance matrix is a real symmetric positive semidefinite matrix
- It have real and non-negative eigen values.
- It would have n linearly independent eigen vectors.
- These also have some interesting roles.
- We consider one simple example.

- Let $Y = \mathbf{a}^T\mathbf{X}$ and assume $||\mathbf{a}|| = 1$
- Y is projection of $\mathbf{X}$ along the direction $\mathbf{a}$.
- Suppose we want to find a direction along which variance is maximized
- We want to maximize $\mathbf{a}^T \Sigma_X \mathbf{a}$ subject to $\mathbf{a}^T\mathbf{a} = 1$
- The lagrangian is $\mathbf{a}^T \Sigma_X \mathbf{a} + \eta(1 - \mathbf{a}^T\mathbf{a})$
- Equating the gradient to zero, we get

$$\Sigma_X \mathbf{a} = \eta \mathbf{a}$$

- So, $\mathbf{a}$ should be an eigen vector (with eigen value η).
- Then the variance would be $\mathbf{a}^T \Sigma_X \mathbf{a} = \eta \mathbf{a}^T\mathbf{a} = \eta$
- Hence the direction is the eigen vector corresponding to the highest eigen value.

Joint moments

- Given two random variables, X, Y
- The joint moment of order (i, j) is defined by

$$m_{ij} = E[X^i Y^j]$$

$m_{10} = EX$, $m_{01} = EY$, $m_{11} = E[XY]$ and so on
- Similarly joint central moments of order (i, j) are defined by

$$s_{ij} = E\left[(X - EX)^i (Y - EY)^j\right]$$

$s_{10} = s_{01} = 0$, $s_{11} = \text{Cov}(X, Y)$, $s_{20} = \text{Var}(X)$ and so on
- We can similarly define joint moments of multiple random variables

- We can define moment generating function of X, Y by

$$M_{XY}(s, t) = E\left[e^{sX+tY}\right], \quad s, t \in \Re$$

- This is easily generalized to n random variables

$$M_{\mathbf{X}}(\mathbf{s}) = E\left[e^{\mathbf{s}^T \mathbf{X}}\right], \quad \mathbf{s} \in \Re^n$$

- Once again, we can get all the moments by differentiating the moment generating function

$$\left.\frac{\partial}{\partial s_i} M_{\mathbf{X}}(\mathbf{s})\right|_{\mathbf{s}=0} = EX_i$$

- More generally

$$\left.\frac{\partial^{m+n}}{\partial s_i^n \, \partial s_j^m} M_{\mathbf{X}}(\mathbf{s})\right|_{\mathbf{s}=0} = EX_i^n X_j^m$$

Conditional Expectation

- Suppose X, Y have a joint density f_{XY}
- Consider the conditional density $f_{X|Y}(x|y)$. This is a density in x for every value of y.
- Since it is a density, we can use it in an expectation integral: $\int g(x) \; f_{X|Y}(x|y) \; dx$
- This is like expectation of $g(X)$ since $f_{X|Y}(x|y)$ is a density in x.
- However, its value would be a function of y.
- That is, this is a kind of expectation that is a function of Y (and hence is a random variable)
- It is called conditional expectation.
- We will now define it formally

- Let X, Y be discrete random variables (on the same probability space).
- The conditinal expectation of $h(X)$ conditioned on Y is a function of Y, and its value for any y is defined by

$$\begin{aligned} E[h(X)|Y = y] &= \sum_x h(x) \; f_{X|Y}(x|y) \\ &= \sum_x h(x) \; P[X = x|Y = y] \end{aligned}$$

- What this means is that we define $E[h(X)|Y] = g(Y)$ where

$$g(y) = \sum_x h(x) \; f_{X|Y}(x|y)$$

- Thus, $E[h(X)|Y]$ is a random variable

- Let X, Y have joint density f_{XY}.
- The conditinal expectation of $h(X)$ conditioned on Y is a function of Y, and its value for any y is defined by

$$E[h(X)|Y = y] = \int_{-\infty}^{\infty} h(x)\ f_{X|Y}(x|y)\ dx$$

- Once again, what this means is that $E[h(X)|Y] = g(Y)$ where

$$g(y) = \int_{-\infty}^{\infty} h(x)\ f_{X|Y}(x|y)\ dx$$

A simple example

- Consider the joint density

$$f_{XY}(x, y) = 2, \quad 0 < x < y < 1$$

- We calculated the conditional densities earlier

$$f_{X|Y}(x|y) = \frac{1}{y}, \quad f_{Y|X}(y|x) = \frac{1}{1-x}, \quad 0 < x < y < 1$$

- Now we can calculate the conditional expectation

$$\begin{aligned} E[X|Y = y] &= \int_{-\infty}^{\infty} x\ f_{X|Y}(x|y)\ dx \\ &= \int_{0}^{y} x\ \frac{1}{y}\ dx = \frac{1}{y}\ \frac{x^2}{2}\Big|_0^y = \frac{y}{2} \end{aligned}$$

- This gives: $E[X|Y] = \frac{Y}{2}$
- We can show $E[Y|X] = \frac{1+X}{2}$

- The conditional expectation is defined by

$$\begin{aligned} E[h(X)|Y = y] &= \sum_x h(x)\ f_{X|Y}(x|y), \quad X, Y \text{ are discrete} \\ E[h(X)|Y = y] &= \int_{-\infty}^{\infty} h(x); f_{X|Y}(x|y)\ dx, \quad X, Y \text{ have joint densit} \end{aligned}$$

- We can actually define $E[h(X, Y)|Y]$ also as above. That is,

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

- It has all the properties of expectation:
  1. $E[a|Y] = a$ where a is a constant
  2. $E[ah_1(X) + bh_2(X)|Y] = aE[h_1(X)|Y] + bE[h_2(X)|Y]$
  3. $h_1(X) \geq h_2(X) \ \Rightarrow \ E[h_1(X)|Y] \geq E[h_2(X)|Y]$

- Conditional expectation also has some extra properties which are very important
  - $E\left[E[h(X)|Y]\right] = E[h(X)]$
  - $E[h_1(X)h_2(Y)|Y] = h_2(Y)E[h_1(X)|Y]$
  - $E[h(X, Y)|Y = y] = E[h(X, y)|Y = y]$
- We will justify each of these.
- The last property above follows directly from the definition.

- Expectation of a conditional expectation is the unconditional expectation
$$E[\,E[h(X)|Y]\,] = E[h(X)]$$
In the above, LHS is expectation of a function of Y.
- Let us denote $g(Y) = E[h(X)|Y]$. Then

$$\begin{aligned}
E[\,E[h(X)|Y]\,] &= E[g(Y)] \\
&= \int_{-\infty}^{\infty} g(y)\, f_Y(y)\, dy \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} h(x)\, f_{X|Y}(x|y)\, dx \right) f_Y(y)\, dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x)\, f_{XY}(x,y)\, dy\, dx \\
&= \int_{-\infty}^{\infty} h(x)\, f_X(x)\, dx \\
&= E[h(X)]
\end{aligned}$$

- Any factor that depends only on the conditioning variable behaves like a constant inside a conditional expectation

$$E[h_1(X)\, h_2(Y)|Y] = h_2(Y) E[h_1(X)|Y]$$

- Let us denote $g(Y) = E[h_1(X)\, h_2(Y)|Y]$

$$\begin{aligned}
g(y) &= E[h_1(X)\, h_2(Y)|Y = y] \\
&= \int_{-\infty}^{\infty} h_1(x) h_2(y)\, f_{X|Y}(x|y)\, dx \\
&= h_2(y) \int_{-\infty}^{\infty} h_1(x)\, f_{X|Y}(x|y)\, dx \\
&= h_2(y)\, E[h_1(X)|Y = y]
\end{aligned}$$

- A very useful property of conditional expectation is $E[\,E[X|Y]\,] = E[X]$ (Assuming all expectations exist)
- We can see this in our earlier example.

$$f_{XY}(x,y) = 2, \;\; 0 < x < y < 1$$

- We calculated: $EX = \frac{1}{3}$ and $EY = \frac{2}{3}$
- We also showed $E[X|Y] = \frac{Y}{2}$

$$E[\,E[X|Y]\,] = E\left[\frac{Y}{2}\right] = \frac{1}{3} = E[X]$$

- Similarly

$$E[\,E[Y|X]\,] = E\left[\frac{1+X}{2}\right] = \frac{2}{3} = E[Y]$$

- We have
$$E[\,E[X|Y]\,] = E[X], \quad \forall X, Y$$
- This is a useful technique to find EX.
- We can choose a Y that is useful.

Density of XY

- Let X, Y have joint density f_{XY}.
- Let $Z = XY$. We want to find density of XY directly
- Let $A_z = \{(x, y) \in \Re^2 \ : \ xy \leq z\} \subset \Re^2$.

$$\begin{aligned} F_Z(z) &= P[XY \leq z] = P[(X, Y) \in A_z] \\ &= \int\int_{A_z} f_{XY}(x, y)\ dy\ dx \end{aligned}$$

- We need to find limits for integrating over A_z
- If $x > 0$, then $xy \leq z \ \Rightarrow \ y \leq z/x$
  If $x < 0$, then $xy \leq z \ \Rightarrow \ y \geq z/x$

$$F_Z(z) = \int_{-\infty}^{0} \int_{z/x}^{\infty} f_{XY}(x, y)\ dy\ dx + \int_{0}^{\infty} \int_{-\infty}^{z/x} f_{XY}(x, y)\ dy\ dx$$

$$F_Z(z) = \int_{-\infty}^{0} \int_{z/x}^{\infty} f_{XY}(x, y)\ dy\ dx + \int_{0}^{\infty} \int_{-\infty}^{z/x} f_{XY}(x, y)\ dy\ dx$$

- Change variable from y to t using $t = xy$
  $y = t/x; \ dy = \frac{1}{x}\ dt; \ y = z/x \ \Rightarrow \ t = z$

$$\begin{aligned} F_Z(z) &= \int_{-\infty}^{0} \int_{z}^{-\infty} \frac{1}{x} f_{XY}(x, \frac{t}{x})\ dt\ dx + \int_{0}^{\infty} \int_{-\infty}^{z} \frac{1}{x} f_{XY}(x, \frac{t}{x})\ dt \\ &= \int_{-\infty}^{0} \int_{-\infty}^{z} \left|\frac{1}{x}\right| f_{XY}(x, \frac{t}{x})\ dt\ dx + \int_{0}^{\infty} \int_{-\infty}^{z} \left|\frac{1}{x}\right| f_{XY}(x, \frac{t}{x}) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{z} \left|\frac{1}{x}\right| f_{XY}\left(x, \frac{t}{x}\right)\ dt\ dx \\ &= \int_{-\infty}^{z} \int_{-\infty}^{\infty} \left|\frac{1}{x}\right| f_{XY}\left(x, \frac{t}{x}\right)\ dx\ dt \end{aligned}$$

This shows: $f_Z(z) = \int_{-\infty}^{\infty} \left|\frac{1}{x}\right| f_{XY}\left(x, \frac{z}{x}\right)\ dx$

Recap: Covariance

- The covariance of X, Y is

  $\text{Cov}(X, Y) = E[(X - EX)\ (Y - EY)] = E[XY] - EX\ EY$

  Note that $\text{Cov}(X, X) = \text{Var}(X)$
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$
- X, Y are called uncorrelated if $\text{Cov}(X, Y) = 0$.
- If X, Y are uncorrelated, $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$
- X, Y independent $\Rightarrow$ X, Y uncorrelated.
- Uncorrelated random variables need not necessarily be independent

Recap: Correlation coefficient

- The correlation coefficient of X, Y is

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\ \text{Var}(Y)}}$$

- If X, Y are uncorrelated then $\rho_{XY} = 0$.
- $-1 \leq \rho_{XY} \leq 1, \ \forall X, Y$
- $|\rho_{XY}| = 1$ iff $X = aY$

Recap: mean square estimation

- The best mean-square approximation of Y as a 'linear' function of X is

$$Y = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}\, X \;+\; \left(EY - \frac{\text{Cov}(X,Y)}{\text{Var}(X)}\, EX\right)$$

- Called the line of regression of Y on X.
- If $\text{cov}(X,Y) = 0$ then this reduces to approximating Y by a constant, EY.
- The final mean square error is

$$\text{Var}(Y)\left(1 - \rho_{XY}^2\right)$$

- If $\rho_{XY} = \pm 1$ then the error is zero
- If $\rho_{XY} = 0$ the final error is $\text{Var}(Y)$

Recap: Covariance matrix

- For a random vector, $\mathbf{X} = (X_1, \cdots, X_n)^T$, the covariance matrix is

$$\Sigma_{\mathbf{X}} = E\left[(\mathbf{X} - E\mathbf{X})\,(\mathbf{X} - E\mathbf{X})^T\right]$$

$$(\Sigma_{\mathbf{X}})_{ij} = E[(X_i - EX_i)(X_j - EX_j)]$$

- $\text{Var}(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T \Sigma_X\, \mathbf{a}$
- Σ_X is a real symmetric and positive semidefinite matrix.

Recap: Moment generating function

- For a pair of rv, the joint moment of order (i,j) is $m_{ij} = E[X^i Y^j]$
- The moment generating function of X, Y is $M_{XY}(s,t) = E\left[e^{sX+tY}\right], \;\; s,t \in \Re$
- For n rv, the joint moments are

$$m_{i_1 i_2 \cdots i_n} = E\left[X_1^{i_1} X_2^{i_2} \cdots X_n^{i_n}\right]$$

- The moment generating function of $\mathbf{X}$ is

$$M_{\mathbf{X}}(\mathbf{s}) = E\left[e^{\mathbf{s}^T\mathbf{X}}\right], \quad \mathbf{s} \in \Re^n$$

Recap: Conditional Expectation

- The conditional expectation of $h(X)$ conditioned on Y is defined by

$$E[h(X)|Y=y] \;=\; \sum_x h(x)\; f_{X|Y}(x|y), \quad X, Y \text{ are discrete}$$

$$E[h(X)|Y=y] \;=\; \int_{-\infty}^{\infty} h(x)\; f_{X|Y}(x|y)\; dx, \quad X, Y \text{ have joint density}$$

- The conditional expectation of $h(X)$ conditioned on Y is a function of Y: $E[h(X)|Y] = g(Y)$ the above specify the value of $g(y)$.
- We define $E[h(X,Y)|Y]$ also as above:

$$E[h(X,Y)|Y=y] = \int_{-\infty}^{\infty} h(x,y)\; f_{X|Y}(x|y)\; dx$$

- If X, Y are independent, $E[h(X)|Y] = E[h(X)]$

Recap: Properties of Conditional Expectation

- It has all the properties of expectation:
  - $E[a|Y] = a$ where a is a constant
  - $E[ah_1(X) + bh_2(X)|Y] = aE[h_1(X)|Y] + bE[h_2(X)|Y]$
  - $h_1(X) \geq h_2(X) \;\; \Rightarrow \;\; E[h_1(X)|Y] \geq E[h_2(X)|Y]$
- Conditional expectation also has some extra properties which are very important
  - $E\left[E[h(X)|Y]\right] = E[h(X)]$
  - $E[h_1(X)h_2(Y)|Y] = h_2(Y)E[h_1(X)|Y]$
  - $E[h(X,Y)|Y = y] = E[h(X,y)|Y = y]$

- Expectation of a conditional expectation is the unconditional expectation

$$E\left[\, E[h(X)|Y] \,\right] = E[h(X)]$$

In the above, LHS is expectation of a function of Y.

- Let us denote $g(Y) = E[h(X)|Y]$. Then

$$\begin{aligned}
E\left[\, E[h(X)|Y] \,\right] &= E[g(Y)] \\
&= \int_{-\infty}^{\infty} g(y)\; f_Y(y)\; dy \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} h(x)\; f_{X|Y}(x|y)\; dx \right)\; f_Y(y)\; dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x)\; f_{XY}(x,y)\; dy\; dx \\
&= \int_{-\infty}^{\infty} h(x)\; f_X(x)\; dx \\
&= E[h(X)]
\end{aligned}$$

- Any factor that depends only on the conditioning variable behaves like a constant inside a conditional expectation

$$E[h_1(X)\; h_2(Y)|Y] = h_2(Y)E[h_1(X)|Y]$$

- Let us denote $g(Y) = E[h_1(X)\; h_2(Y)|Y]$

$$\begin{aligned}
g(y) &= E[h_1(X)\; h_2(Y)|Y = y] \\
&= \int_{-\infty}^{\infty} h_1(x)h_2(y)\; f_{X|Y}(x|y)\; dx \\
&= h_2(y) \int_{-\infty}^{\infty} h_1(x)\; f_{X|Y}(x|y)\; dx \\
&= h_2(y)\; E[h_1(X)|Y = y]
\end{aligned}$$

Example

- Let X, Y be random variables with joint density given by

$$f_{XY}(x,y) = e^{-y}, \;\; 0 < x < y < \infty$$

- The marginal densities are:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\; dy = \int_{x}^{\infty} e^{-y}\; dy = e^{-x}, \;\; x > 0$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\; dx = \int_{0}^{y} e^{-y}\; dx = y\; e^{-y}, \;\; y > 0$$

Thus, X is exponential and Y is gamma.

- Hence we have

$$EX = 1; \quad \text{Var}(X) = 1; \quad EY = 2; \quad \text{Var}(Y) = 2$$

$$f_{XY}(x,y) = e^{-y}, \;\; 0 < x < y < \infty$$

- Let us calculate covariance of X and Y

$$E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy \; f_{XY}(x,y) \; dx \; dy$$
$$= \int_{0}^{\infty}\int_{0}^{y} xye^{-y} \; dx \; dy = \int_{0}^{\infty} \frac{1}{2} \, y^3 e^{-y} \; dy = 3$$

- Hence, Cov$(X,Y) = E[XY] - EX \; EY = 3 - 2 = 1$.
- $\rho_{XY} = \frac{1}{\sqrt{2}}$

- Recall the joint and marginal densities

$$f_{XY}(x,y) = e^{-y}, \;\; 0 < x < y < \infty$$

$$f_X(x) = e^{-x}, \; x > 0; \quad f_Y(y) = ye^{-y}, \; y > 0$$

- The conditional densities will be

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{e^{-y}}{ye^{-y}} = \frac{1}{y}, \;\; 0 < x < y < \infty$$

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{e^{-y}}{e^{-x}} = e^{-(y-x)}, \;\; 0 < x < y < \infty$$

- The conditional densities are

$$f_{X|Y}(x|y) = \frac{1}{y}; \quad f_{Y|X}(y|x) = e^{-(y-x)}, \;\; 0 < x < y < \infty$$

- We can now calculate the conditional expectation

$$E[X|Y=y] = \int x \; f_{X|Y}(x|y) \; dx = \int_0^y x \, \frac{1}{y} \; dx = \frac{y}{2}$$

Thus $E[X|Y] = \frac{Y}{2}$

$$E[Y|X=x] = \int y \; f_{Y|X}(y|x) \; dy = \int_x^{\infty} ye^{-(y-x)} \; dy$$
$$= e^x \left(-ye^{-y}\Big|_x^{\infty} + \int_x^{\infty} e^{-y} \; dy \right)$$
$$= e^x \left(xe^{-x} + e^{-x} \right) = 1 + x$$

Thus, $E[Y|X] = 1 + X$

- We got

$$E[X|Y] = \frac{Y}{2}; \quad E[Y|X] = 1 + X$$

- Using this we can verify:

$$E[\; E[X|Y] \;] = E\left[\frac{Y}{2}\right] = \frac{EY}{2} = \frac{2}{2} = 1 = EX$$

$$E[\; E[Y|X] \;] = E[1 + X] = 1 + 1 = 2 = EY$$

- A property of conditional expectation is

$$E[\ E[X|Y]\ ] = E[X]$$

- We assume that all three expectations exist.
- Very useful in calculating expectations

$$EX = \sum_y E[X|Y=y]\ f_Y(y) \quad \text{or} \quad \int E[X|Y=y]\ f_Y(y)\ dy$$

- Can be used to calculate probabilities of events too

$$P(A) = E[I_A] = E\left[\ E\left[I_A|Y\right]\ \right]$$

- Let X be geometric and we want EX.
- X is number of tosses needed to get head
- Let $Y \in \{0,\ 1\}$ be outcome of first toss. (1 for head)

$$\begin{aligned} E[X] &= E[\ E[X|Y]\ ] \\ &= E[X|Y=1]\ P[Y=1] + E[X|Y=0]\ P[Y=0] \\ &= E[X|Y=1]\ p + E[X|Y=0]\ (1-p) \\ &= 1\ p + (1+EX)(1-p) \\ \Rightarrow &\quad EX\ (1-(1-p)) = p + (1-p) = 1 \\ \Rightarrow &\quad EX = \frac{1}{p} \end{aligned}$$

- $P[X=k|Y=1] = 1$ if $k=1$ (otherwise it is zero) and hence $E[X|Y=1] = 1$

$$P[X=k|Y=0] = \begin{cases} 0 & \text{if } k=1 \\ \frac{(1-p)^{k-1}p}{(1-p)} & \text{if } k \geq 2 \end{cases}$$

Hence

$$\begin{aligned} E[X|Y=0] &= \sum_{k=2}^{\infty} k\ (1-p)^{k-2}p \\ &= \sum_{k=2}^{\infty}(k-1)\ (1-p)^{k-2}p + \sum_{k=2}^{\infty}(1-p)^{k-2}p \\ &= \sum_{k'=1}^{\infty} k'\ (1-p)^{k'-1}p + \sum_{k'=1}^{\infty}(1-p)^{k'-1}p \\ &= EX + 1 \end{aligned}$$

Another example

- Example: multiple rounds of the party game
- Let R_n denote number of rounds when you start with n people.
- We want $\bar{R}_n = E\left[R_n\right]$.
- We want to use $E\left[R_n\right] = E[\ E\left[R_n|X_n\right]\ ]$
- We need to think of a useful X_n.
- Let X_n be the number of people who got their own hat in the first round with n people.

- R_n – number of rounds when you start with n people.
- X_n – number of people who got their own hat in the first round

$$\begin{aligned} E[R_n] &= E[\, E[R_n|X_n]\,] \\ &= \sum_{i=0}^{n} E[R_n|X_n=i]\;\; P[X_n=i] \\ &= \sum_{i=0}^{n} (1+E[R_{n-i}])\;\; P[X_n=i] \\ &= \sum_{i=0}^{n} P[X_n=i] \;+\; \sum_{i=0}^{n} E[R_{n-i}]\;\; P[X_n=i] \end{aligned}$$

If we can guess value of $E[R_n]$ then we can prove it using mathematical induction

- What would be $E[X_n]$?
- Let $Y_i \in \{0,\ 1\}$ denote whether or not i^{th} person got his own hat.
- We know

$$E[Y_i] = P[Y_i = 1] = \frac{(n-1)!}{n!} = \frac{1}{n}$$

Now, $\quad X_n = \sum_{i=1}^{n} Y_i \;$ and hence $\; EX_n = \sum_{i=1}^{n} E[Y_i] = 1$

- Hence a good guess is $E[R_n] = n$.
- We verify it using mathematical induction. We know $E[R_1] = 1$

- Assume: $E[R_k] = k, \;\; 1 \le k \le n-1$

$$\begin{aligned} E[R_n] &= \sum_{i=0}^{n} P[X_n=i] \;+\; \sum_{i=0}^{n} E[R_{n-i}]\;\; P[X_n=i] \\ &= 1 \;+\; E[R_n]\;\; P[X_n=0] \;+\; \sum_{i=1}^{n} E[R_{n-i}]\;\; P[X_n=i] \\ &= 1 \;+\; E[R_n]\;\; P[X_n=0] \;+\; \sum_{i=1}^{n} (n-i)\; P[X_n=i] \end{aligned}$$

$$\begin{aligned} E[R_n]\,(1-P[X_n=0]) &= 1 + n(1-P[X_n=0]) - \sum_{i=1}^{n} i\; P[X_n=i] \\ &= 1 \;+\; n\,(1-P[X_n=0]) - E[X_n] \\ &= 1 \;+\; n\,(1-P[X_n=0]) - 1 \end{aligned}$$

$$\Rightarrow \quad E[R_n] = n$$

Analysis of Quicksort

- Given n numbers we want to sort them. Many algorithms.
- Complexity – order of the number of comparisons needed
- Quicksort: Choose a pivot. Separte numbers into two parts – less and greater than pivot, do recursively
- Separating into two parts takes $n-1$ comparisons.
- Suppose the two parts contain m and $n-m-1$. Separating both of them into two parts each takes $m+n-m-1$ comparisons
- So, final number of comparisons depends on the 'number of rounds'

quicksort details

- Given $\{x_1, \cdots, x_n\}$.
- Choose first as pivot

$$\{x_{j_1}, x_{j_2}, \cdots, x_{j_m}\}x_1\{x_{k_1}, x_{k_2}, \cdots, x_{k_{n-1-m}}\}$$

- Suppose r_n is the number of comparisons. If we get (roughly) equal parts, then

$$r_n \approx n+2r_{n/2} = n+2(n/2+2r_{n/4}) = n+n+4r_{n/4} = \cdots = n\log_2(n)$$

- If all the rest go into one part, then

$$r_n = n + r_{n-1} = n + (n-1) + r_{n-2} = \cdots = \frac{n(n+1)}{2}$$

- If you are lucky, $O(n\log(n))$ comparisons.
- If unlucky, in the worst case, $O(n^2)$ comparisons
- Question: 'on the average' how many comparisons?

Average case complexity of quicksort

- Assume pivot is equally likely to be the smallest or second smallest or m^{th} smallest.
- M_n – number of comparisons.
- Define: $X = j$ if pivot is j^{th} smallest
- Given $X = j$ we know $M_n = (n-1) + M_{j-1} + M_{n-j}$.

$$\begin{aligned} E[M_n] &= E[\ E[M_n|X]\ ] = \sum_{j=1}^{n} E[M_n|X=j]\ P[X=j] \\ &= \sum_{j=1}^{n} E[(n-1) + M_{j-1} + M_{n-j}]\ \frac{1}{n} \\ &= (n-1) + \frac{2}{n}\sum_{k=1}^{n-1} E[M_k], \quad \text{(taking } M_0 = 0) \end{aligned}$$

- This is a recurrence relation. (A little complicated to solve)

Least squares estimation

- We want to estimate Y as a function of X.
- We want an estimate with minimum mean square error.
- We want to solve (the min is over all functions g)

$$\min_g\ E\left(Y - g(X)\right)^2$$

- Earlier we considered linear functions: $g(X) = aX + b$
- The solution now turns out to be

$$g^*(X) = E[Y|X]$$

- Let us prove this.

- We want to show that for all g

$$E\left[(E[Y \mid X] - Y)^2\right] \ \le\ E\left[(g(X) - Y)^2\right]$$

- We have

$$\begin{aligned} (g(X) - Y)^2 &= \left[(g(X) - E[Y \mid X]) + (E[Y \mid X] - Y)\right]^2 \\ &= \left(g(X) - E[Y \mid X]\right)^2 \ +\ \left(E[Y \mid X] - Y\right)^2 \\ &\qquad + 2\left(g(X) - E[Y \mid X]\right)\left(E[Y \mid X] - Y\right) \end{aligned}$$

- Now we can take expectation on both sides.
- We first show that expectation of last term on RHS above is zero.

First consider the last term

$$
\begin{aligned}
& E\left[(g(X)-E[Y \mid X])(E[Y \mid X]-Y)\right] \\
= \; & E\left[\; E\left\{(g(X)-E[Y \mid X])(E[Y \mid X]-Y) \mid X\right\} \;\right] \\
& \qquad \text{because} \quad E[Z] = E[\,E[Z|X]\,] \\
= \; & E\left[\; (g(X)-E[Y \mid X]) \;\; E\left\{(E[Y \mid X]-Y) \mid X\right\} \;\right] \\
& \qquad \text{because} \quad E[h_1(X)h_2(Z)|X] = h_1(X)\, E[h_2(Z)|X] \\
= \; & E\left[\; (g(X)-E[Y \mid X]) \;\; \left(E\left\{(E[Y \mid X])|X\right\} \; - \; E\left\{Y \mid X\right\}\right) \;\right] \\
= \; & E\left[\; (g(X)-E[Y \mid X]) \;\; (E[Y \mid X] \; - \; E[Y \mid X]) \;\right] \\
= \; & 0
\end{aligned}
$$

- We earlier got

$$
\begin{aligned}
(g(X)-Y)^2 \;\; = \;\; & \big(g(X)-E[Y \mid X]\big)^2 \; + \; \big(E[Y \mid X]-Y\big)^2 \\
& + 2\big(g(X)-E[Y \mid X]\big)\big(E[Y \mid X]-Y\big)
\end{aligned}
$$

- Hence we get

$$
\begin{aligned}
E\left[\,(g(X)-Y)^2\,\right] \;\; = \;\; & E\left[\,(g(X)-E[Y \mid X])^2\,\right] \\
& + E\left[\,(E[Y \mid X]-Y)^2\,\right] \\
\geq \;\; & E\left[\,(E[Y \mid X]-Y)^2\,\right]
\end{aligned}
$$

- Since the above is true for all functions g, we get

$$
g^*(X) \; = \; E\,[Y \mid X]
$$

Sum of random number of random variables

- Let $X_1, X_2, \cdots$ be iid rv on the same probability space. Suppose $EX_i = \mu, \; \forall i$.
- Let N be a positive integer valued rv that is independent of all X_i.
- Let $S = \sum_{i=1}^{N} X_i$.
- We want to calculate ES. We can use

$$
E[S] = E[\;E[S|N]\;]
$$

- We have

$$
\begin{aligned}
E[S|N=n] \;\; = \;\; & E\left[\sum_{i=1}^{N} X_i \mid N=n\right] \\
= \;\; & E\left[\sum_{i=1}^{n} X_i \mid N=n\right] \\
& \text{since } E[h(X,Y)|Y=y] = E[h(X,y)|Y=y] \\
= \;\; & \sum_{i=1}^{n} E[X_i \mid N=n] = \sum_{i=1}^{n} E[X_i] = n\mu
\end{aligned}
$$

- Hence we get

$$
E[S|N] = N\mu \quad \Rightarrow \quad E[S] = E[N]E[X_1]
$$

- Actually, we did not use independence of X_i.

Recap: Conditional Expectation

- The conditional expectation of $h(X)$ conditioned on Y is defined by

$$E[h(X)|Y=y] = \sum_x h(x)\ f_{X|Y}(x|y), \quad X, Y \text{ are discrete}$$
$$E[h(X)|Y=y] = \int_{-\infty}^{\infty} h(x)\ f_{X|Y}(x|y)\ dx, \quad X, Y \text{ have joint density}$$

- The conditional expectation of $h(X)$ conditioned on Y is a function of Y: $E[h(X)|Y] = g(Y)$
  the above specify the value of $g(y)$.
- We define $E[h(X,Y)|Y]$ also as above:

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

- If X, Y are independent, $E[h(X)|Y] = E[h(X)]$

Recap: Properties of Conditional Expectation

- It has all the properties of expectation:
  - $E[a|Y] = a$ where a is a constant
  - $E[ah_1(X) + bh_2(X)|Y] = aE[h_1(X)|Y] + bE[h_2(X)|Y]$
  - $h_1(X) \geq h_2(X) \quad \Rightarrow \quad E[h_1(X)|Y] \geq E[h_2(X)|Y]$
- Conditional expectation also has some extra properties which are very important
  - $E\left[E[h(X)|Y]\right] = E[h(X)]$
  - $E[h_1(X)h_2(Y)|Y] = h_2(Y)E[h_1(X)|Y]$
  - $E[h(X,Y)|Y=y] = E[h(X,y)|Y=y]$

- The property of conditional expectation

$$E[\ E[X|Y]\ ] = E[X]$$

  is very useful in calculating expectations

$$EX = \sum_y E[X|Y=y]\ f_Y(y) \quad \text{or} \quad \int E[X|Y=y]\ f_Y(y)\ dy$$

  We saw many examples.
- Can be used to calculate probabilities of events too

$$P(A) = E[I_A] = E\left[\ E\left[I_A|Y\right]\ \right]$$

Sum of random number of random variables

- Let $X_1, X_2, \cdots$ be iid rv on the same probability space. Suppose $EX_i = \mu,\ \forall i$.
- Let N be a positive integer valued rv that is independent of all X_i.
- Let $S = \sum_{i=1}^{N} X_i$.
- We want to calculate ES. We can use

$$E[S] = E[\ E[S|N]\ ]$$

- We have

$$\begin{aligned}
E[S|N=n] &= E\left[\sum_{i=1}^{N} X_i \mid N=n\right] \\
&= E\left[\sum_{i=1}^{n} X_i \mid N=n\right] \\
&\qquad \text{since } E[h(X,Y)|Y=y] = E[h(X,y)|Y=y] \\
&= \sum_{i=1}^{n} E[X_i \mid N=n] = \sum_{i=1}^{n} E[X_i] = n\mu
\end{aligned}$$

- Hence we get

$$E[S|N] = N\mu \quad \Rightarrow \quad E[S] = E[N]E[X_1]$$

- Actually, we did not use independence of X_i.

Variance of random sum

- $S = \sum_{i=1}^{N} X_i$, X_i iid, ind of N. Want Var(S)

$$E[S^2] = E\left[\left(\sum_{i=1}^{N} X_i\right)^2\right] = E\left[E\left[\left(\sum_{i=1}^{N} X_i\right)^2 \mid N\right]\right]$$

- As earlier, we have

$$\begin{aligned}
E\left[\left(\sum_{i=1}^{N} X_i\right)^2 \mid N=n\right] &= E\left[\left(\sum_{i=1}^{n} X_i\right)^2 \mid N=n\right] \\
&= E\left[\left(\sum_{i=1}^{n} X_i\right)^2\right]
\end{aligned}$$

- Let $Y = \sum_{i=1}^{n} X_i, \quad X_i$ iid
- Then, Var$(Y) = n$ Var(X_1)
- Hence we have

$$E[Y^2] = \text{Var}(Y) + (EY)^2 = n\,\text{Var}(X_1) + (nEX_1)^2$$

- Using this

$$E\left[\left(\sum_{i=1}^{N} X_i\right)^2 \mid N=n\right] = E\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] = n\,\text{Var}(X_1) + (nEX_1)^2$$

- Hence

$$E\left[\left(\sum_{i=1}^{N} X_i\right)^2 \mid N\right] = N\,\text{Var}(X_1) + N^2\,(EX_1)^2$$

- $S = \sum_{i=1}^{N} X_i \quad (X_i$ iid$)$. We got

$$E[S^2] = E[\,E[S^2|N]\,] = EN\,\text{Var}(X_1) + E[N^2](EX_1)^2$$

- Now we can calculate variance of S as

$$\begin{aligned}
\text{Var}(S) &= E[S^2] - (ES)^2 \\
&= EN\,\text{Var}(X_1) + E[N^2](EX_1)^2 - (EN\;EX_1)^2 \\
&= EN\,\text{Var}(X_1) + (EX_1)^2\left(E[N^2] - (EN)^2\right) \\
&= EN\,\text{Var}(X_1) + \text{Var}(N)\,(EX_1)^2
\end{aligned}$$

Wald's formula

- Considered $S = \sum_{i=1}^{N} X_i$ with N independent of all X_i.
- With iid X_i, the formula $ES = EN\ EX_1$ is valid even under some dependence between N and X_i.
- Here is one version of Wald's formula. We assume
  1. $E[|X_i|] < \infty$, $\forall i$ and $EN < \infty$.
  2. $E\left[X_n\, I_{[N \geq n]}\right] = E[X_n] P[N \geq n]$, $\forall n$
- Let $S_N = \sum_{i=1}^{N} X_i$ and let $T_N = \sum_{i=1}^{N} E[X_i]$.
- Then, $ES_N = ET_N$.
  If $E[X_i]$ is same for all i, $ES_N = EX_1\ EN$.
- Assume X_i are iid. Suppose the event $[N \leq n-1]$ depends only on $X_1, \cdots, X_{n-1}$.
- Then the event $[N \leq n-1]$ and hence its complement $[N \geq n]$ is independent of X_n and the assumption above is satisfied.
- Such an N is an example of what is called a stopping time.

Another Example

- We toss a (biased) coin till we get k consecutive heads. Let N_k denote the number of tosses needed.
- N_1 would be geometric.
- We want $E[N_k]$. What rv should we condition on?
- Useful rv here is N_{k-1}

$$E[N_k \mid N_{k-1} = n] = (n+1)p + (1-p)(n+1+E[N_k])$$

- Thus we get the recurrence relation

$$\begin{aligned} E[N_k] &= E[\, E[N_k \mid N_{k-1}]\,] \\ &= E\left[\, (N_{k-1}+1)p + (1-p)(N_{k-1}+1+E[N_k])\,\right] \end{aligned}$$

- We have

$$E[N_k] = E\left[\, (N_{k-1}+1)p + (1-p)(N_{k-1}+1+E[N_k])\,\right]$$

- Denoting $M_k = E[N_k]$, we get

$$\begin{aligned} M_k &= pM_{k-1} + p + (1-p)M_{k-1} + (1-p) + (1-p)M_k \\ pM_k &= M_{k-1} + 1 \\ M_k &= \frac{1}{p} M_{k-1} + \frac{1}{p} \\ &= \frac{1}{p}\left(\frac{1}{p} M_{k-2} + \frac{1}{p}\right) + \frac{1}{p} = \left(\frac{1}{p}\right)^2 M_{k-2} + \left(\frac{1}{p}\right)^2 + \frac{1}{p} \\ &= \left(\frac{1}{p}\right)^{k-1} M_1 + \sum_{j=1}^{k-1} \left(\frac{1}{p}\right)^j = \sum_{j=1}^{k} \left(\frac{1}{p}\right)^j \quad (M_1 = \frac{1}{p}) \\ &= \frac{\frac{1}{p}\left(1 - \left(\frac{1}{p}\right)^k\right)}{\left(1 - \frac{1}{p}\right)} = \frac{1-p^k}{(1-p)p^k} \end{aligned}$$

- As mentioned earlier, we can use the conditional expectation to calculate probabilities of events also.

$$P(A) = E[I_A] = E\left[\, E\left[I_A|Y\right]\,\right]$$

$$E[I_A|Y=y] = P[I_A = 1|Y=y] = P(A|Y=y)$$

- Thus, we get

$$\begin{aligned} P(A) &= E[I_A] = E\left[\, E\left[I_A|Y\right]\,\right] \\ &= \sum_y P(A|Y=y)P[Y=y], \quad \text{when } Y \text{ is discrete} \\ &= \int P(A|Y=y)\ f_Y(y)\ dy, \quad \text{when } Y \text{ is continuous} \end{aligned}$$

Example

- Let X, Y be independent continuous rv
- We want to calculate $P[X \leq Y]$
- We can calculate it by integrating joint density over $A = \{(x, y) \ : \ x \leq y\}$

$$
\begin{aligned}
P[X \leq Y] &= \int\int_A f_X(x)\ f_Y(y)\ dx\ dy \\
&= \int_{-\infty}^{\infty} f_Y(y) \left(\int_{-\infty}^{y} f_X(x)\ dx \right) dy \\
&= \int_{-\infty}^{\infty} F_X(y)\ f_Y(y)\ dy
\end{aligned}
$$

- IF X, Y are *iid* then $P[X < Y] = 0.5$

- We can also use the conditional expectation method here

$$
\begin{aligned}
P[X \leq Y] &= \int_{-\infty}^{\infty} P[X \leq Y \mid Y = y]\ f_Y(y)\ dy \\
&= \int_{-\infty}^{\infty} P[X \leq y \mid Y = y]\ f_Y(y)\ dy \\
&= \int_{-\infty}^{\infty} P[X \leq y]\ f_Y(y)\ dy \\
&= \int_{-\infty}^{\infty} F_X(y)\ f_Y(y)\ dy
\end{aligned}
$$

- Consider a sequence of bernoullli trials where p, probability of success, is random.
- We first choose p uniformly over $(0, 1)$ and then perform n tosses.
- Let X be the number of heads.
- Conditioned on knowledge of p, we know distribution of X

$$P[X = k \mid p] = {}^nC_k\ p^k\ (1-p)^{n-k}$$

- Now we can calculate $P[X = k]$ using the conditioning argument.

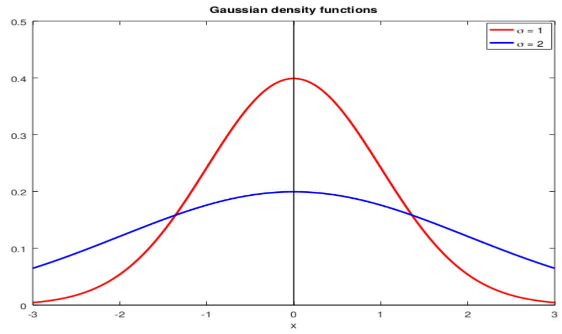
- Assuming p is chosen uniformly from $(0,\ 1)$, we get

$$
\begin{aligned}
P[X = k] &= \int [P[X = k \mid p]\ f(p)\ dp \\
&= \int_0^1\ {}^nC_k\ p^k\ (1-p)^{n-k}\ 1\ dp \\
&= {}^nC_k\ \frac{k!(n-k)!}{(n+1)!} \\
\text{because} & \quad \int_0^1 p^k\ (1-p)^{n-k}\ dp = \frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} \\
&= \frac{1}{n+1}
\end{aligned}
$$

- So, we get: $P[X = k] = \frac{1}{n+1},\ \ k = 0, 1, \cdots, n$

Gaussian or Normal distribution

- The Gaussian or normal density is given by
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$
- If X has this density, we denote it as $X \sim \mathcal{N}(\mu, \sigma^2)$.
We showed $EX = \mu$ and Var$(X) = \sigma^2$
- The density is a 'bell-shaped' curve



- Standard Normal rv — $X \sim \mathcal{N}(0, 1)$
- The distribution function of standard normal is
$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$$
- Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$
$$\begin{aligned} P[a \le X \le b] &= \int_a^b \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \\ & \quad \text{take } \ y = \frac{(x-\mu)}{\sigma} \ \Rightarrow \ dy = \frac{1}{\sigma} dx \\ &= \int_{\frac{(a-\mu)}{\sigma}}^{\frac{(b-\mu)}{\sigma}} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{y^2}{2}}\, dy \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \end{aligned}$$
- We can express probability of events involving all Normal rv using Φ.

- $X \sim \mathcal{N}(0, 1)$. Then its mgf is
$$\begin{aligned} M_X(t) &= E\left[e^{tX}\right] = \int_{-\infty}^{\infty} e^{tx}\, \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2 - 2tx)}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left((x-t)^2 - t^2\right)}\, dx \\ &= e^{\frac{1}{2}t^2}\, \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-t)^2}\, dx \\ &= e^{\frac{1}{2}t^2} \end{aligned}$$
- Now let $Y = \sigma X + \mu$. Then $Y \sim \mathcal{N}(\mu, \sigma^2)$.
The mgf of Y is
$$\begin{aligned} M_Y(t) &= E\left[e^{t(\sigma X + \mu)}\right] = e^{t\mu}\, E\left[e^{(t\sigma)X}\right] = e^{t\mu}\, M_X(t\sigma) \\ &= e^{\left(\mu t + \frac{1}{2}t^2\sigma^2\right)} \end{aligned}$$

Multi-dimensional Gaussian Distribution

- The n-dimensional Gaussian density is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{|\Sigma|^{\frac{1}{2}}\,(2\pi)^{\frac{n}{2}}}\; e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \quad \mathbf{x}\in\Re^n$$

- $\boldsymbol{\mu}\in\Re^n$ and $\Sigma\in\Re^{n\times n}$ are parameters of the density and Σ is symmetric and positive definite.
- If $X_1,\cdots,X_n$ have the above joint density, they are said to be jointly Gaussian.
- We denote this by $\mathbf{X}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)$
- We will now show that this is a joint density function.

- We begin by showing the following is a density (when M is symmetric +ve definite)

$$f_{\mathbf{Y}}(\mathbf{y}) = C\; e^{-\frac{1}{2}\mathbf{y}^TM\mathbf{y}}$$

- Let $I = \int_{\Re^n} C\; e^{-\frac{1}{2}\mathbf{y}^TM\mathbf{y}}\; d\mathbf{y}$
- Since M is real symmetric, there exits an orthogonal transform, L with $L^{-1}=L^T$, $|L|=1$ and L^TML is diagonal
- Let $L^TML = \text{diag}(m_1,\cdots,m_n)$.
- Then for any $\mathbf{z}\in\Re^n$,

$$\mathbf{z}^TL^TML\mathbf{z} = \sum_i m_i z_i^2$$

- We now get

$$\begin{aligned}
I &= \int_{\Re^n} C\; e^{-\frac{1}{2}\mathbf{y}^TM\mathbf{y}}\; d\mathbf{y}\\
&\quad \text{change variable: } \mathbf{z}=L^{-1}\mathbf{y}=L^T\mathbf{y} \;\;\Rightarrow \mathbf{y}=L\mathbf{z}\\
&= C\int_{\Re^n} e^{-\frac{1}{2}\mathbf{z}^TL^TML\mathbf{z}}\; d\mathbf{z} \quad (\text{note that } \; |L|=1)\\
&= C\int_{\Re^n} e^{-\frac{1}{2}\sum_i m_i z_i^2}\; d\mathbf{z}\\
&= C\prod_{i=1}^n \int_{\Re} e^{-\frac{1}{2}m_i z_i^2}\; dz_i = C\prod_{i=1}^n \int_{\Re} e^{-\frac{1}{2}\frac{z_i^2}{\frac{1}{m_i}}}\; dz_i\\
&= C\prod_{i=1}^n \sqrt{2\pi\,\frac{1}{m_i}}
\end{aligned}$$

- We will first relate $m_1\cdots m_n$ to the matrix M.
- By definition, $L^TML=\text{diag}(m_1,\cdots,m_n)$. Hence

$$\text{diag}\left(\frac{1}{m_1},\cdots,\frac{1}{m_n}\right) = \left(L^TML\right)^{-1} = L^{-1}M^{-1}(L^T)^{-1} = L^TM^{-1}L$$

- Since $|L|=1$, we get

$$\left|L^TM^{-1}L\right| = \left|M^{-1}\right| = \frac{1}{m_1\cdots m_n}$$

Putting all this together

$$\int_{\Re^n} C\; e^{-\frac{1}{2}\mathbf{y}^TM\mathbf{y}}\; d\mathbf{y} \;=\; C\prod_{i=1}^n\sqrt{2\pi\,\frac{1}{m_i}} = C\;(2\pi)^{\frac{n}{2}}\left|M^{-1}\right|^{\frac{1}{2}}$$

$$\Rightarrow \frac{1}{(2\pi)^{\frac{n}{2}}\left|M^{-1}\right|^{\frac{1}{2}}}\int_{\Re^n} e^{-\frac{1}{2}\mathbf{y}^TM\mathbf{y}}\; d\mathbf{y} \;=\; 1$$

- We showed the following is a density (taking $M^{-1} = \Sigma$)

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \, e^{-\frac{1}{2}\mathbf{y}^T\Sigma^{-1}\mathbf{y}}, \quad \mathbf{y} \in \Re^n$$

- Let $\mathbf{X} = \mathbf{Y} + \boldsymbol{\mu}$. Then

$$f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{Y}}(\mathbf{x} - \boldsymbol{\mu}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

- This is the multidimensional Gaussian distribution

- Consider $\mathbf{Y}$ with joint density

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \, e^{-\frac{1}{2}\mathbf{y}^T\Sigma^{-1}\mathbf{y}}, \quad \mathbf{y} \in \Re^n$$

- As earlier let $M = \Sigma^{-1}$. Let $L^T M L = \text{diag}(m_1, \cdots, m_n)$
- Define $\mathbf{Z} = (Z_1, \cdots, Z_n)^T = L^T\mathbf{Y}$. Then $\mathbf{Y} = L\mathbf{Z}$.
- Recall $|L| = 1$, $|M^{-1}| = (m_1 \cdots m_n)^{-1}$
- Then density of $\mathbf{Z}$ is

$$\begin{aligned} f_{\mathbf{Z}}(\mathbf{z}) &= \frac{1}{(2\pi)^{\frac{n}{2}} |M^{-1}|^{\frac{1}{2}}} \, e^{-\frac{1}{2}\mathbf{z}^T L^T M L \mathbf{z}} = \frac{1}{(2\pi)^{\frac{n}{2}} (\frac{1}{m_1 \cdots m_n})^{\frac{1}{2}}} e^{-\frac{1}{2}\sum_i m_i z_i^2} \\ &= \prod_{i=1}^n \sqrt{\frac{1}{2\pi}} \, \frac{1}{\sqrt{\frac{1}{m_i}}} \, e^{-\frac{1}{2}m_i z_i^2} = \prod_{i=1}^n \sqrt{\frac{1}{2\pi}} \, \frac{1}{\sqrt{\frac{1}{m_i}}} \, e^{-\frac{1}{2}\frac{z_i^2}{\frac{1}{m_i}}} \end{aligned}$$

This shows that $Z_i \sim \mathcal{N}(0, \frac{1}{m_i})$ and Z_i are independent.

- If $\mathbf{Y}$ has density $f_{\mathbf{Y}}$ and $\mathbf{Z} = L^T Y$ then $Z_i \sim \mathcal{N}(0, \frac{1}{m_i})$ and Z_i are independent. Hence,

$$\Sigma_Z = \text{diag}\left(\frac{1}{m_1}, \cdots, \frac{1}{m_n}\right) = L^T M^{-1} L$$

- Also, since $Z_i = 0$, $\Sigma_Z = E[\mathbf{Z}\mathbf{Z}^T]$.
- Since $\mathbf{Y} = L\mathbf{Z}$, $E[\mathbf{Y}] = 0$ and

$$\Sigma_Y = E[\mathbf{Y}\mathbf{Y}^T] = E[L\mathbf{Z}\mathbf{Z}^T L^T] = LE[\mathbf{Z}\mathbf{Z}^T]L^T = L(L^T M^{-1} L)L^T = M^{-1}$$

- Thus, if $\mathbf{Y}$ has density

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \, e^{-\frac{1}{2}\mathbf{y}^T\Sigma^{-1}\mathbf{y}}, \quad \mathbf{y} \in \Re^n$$

then $E\mathbf{Y} = 0$ and $\Sigma_Y = M^{-1} = \Sigma$

- Let $\mathbf{Y}$ have density

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \, e^{-\frac{1}{2}\mathbf{y}^T\Sigma^{-1}\mathbf{y}}, \quad \mathbf{y} \in \Re^n$$

- Let $\mathbf{X} = \mathbf{Y} + \boldsymbol{\mu}$. Then

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

- We have

$$E\mathbf{X} = E[\mathbf{Y} + \boldsymbol{\mu}] = \boldsymbol{\mu}$$

$$\Sigma_X = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] = E[\mathbf{Y}\mathbf{Y}^T] = \Sigma$$

Multi-dimensional Gaussian density

- $\mathbf{X} = (X_1, \cdots, X_n)^T$ are said to be jointly Gaussian if

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

- $E\mathbf{X} = \boldsymbol{\mu}$ and $\Sigma_X = \Sigma$.
- Suppose Cov$(X_i, X_j) = 0, \forall i \neq j$.
- Then $\Sigma_{ij} = 0, \forall i \neq j$. Let $\Sigma =$ diag$(\sigma_1^2, \cdots, \sigma_n^2)$.

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} \, \sigma_1 \cdots \sigma_n} e^{-\frac{1}{2}\sum_{i=1}^n \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2} = \prod_{i=1}^n \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2}$$

- This implies X_i are independent.
- If $X_1, \cdots, X_n$ are jointly Gaussian then uncorrelatedness implies independence.

- Let $\mathbf{X} = (X_1, \cdots, X_n)^T$ be jointly Gaussian:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

- Let $\mathbf{Y} = \mathbf{X} - \boldsymbol{\mu}$.
- Let $M = \Sigma^{-1}$ and L be such that $L^T M L =$ diag$(m_1, \cdots, m_n)$
- Let $\mathbf{Z} = (Z_1, \cdots, Z_n)^T = L^T Y$.
- Then we saw that $Z_i \sim \mathcal{N}(0, \frac{1}{m_i})$ and Z_i are independent.
- If $X_1, \cdots, X_n$ are jointly Gaussian then there is a 'linear' transform that transforms them into independent random variables.

Moment generating function

- Let $\mathbf{X} = (X_1, \cdots, X_n)^T$ be jointly Gaussian
- Let $\mathbf{Y} = \mathbf{X} - \boldsymbol{\mu}$ and $\mathbf{Z} = (Z_1, \cdots, Z_n)^T = L^T Y$ as earlier
- The moment generating function of $\mathbf{X}$ is given by

$$\begin{aligned}
M_{\mathbf{X}}(\mathbf{s}) &= E\left[e^{\mathbf{s}^T \mathbf{X}}\right] \\
&= E\left[e^{\mathbf{s}^T(\mathbf{Y}+\boldsymbol{\mu})}\right] = e^{\mathbf{s}^T\boldsymbol{\mu}} \, E\left[e^{\mathbf{s}^T\mathbf{Y}}\right] \\
&= e^{\mathbf{s}^T\boldsymbol{\mu}} \, E\left[e^{\mathbf{s}^T L \mathbf{Z}}\right] \\
&= e^{\mathbf{s}^T\boldsymbol{\mu}} \, E\left[e^{\mathbf{u}^T \mathbf{Z}}\right] \\
&\quad \text{where} \quad \mathbf{u} = L^T \mathbf{s} \\
&= e^{\mathbf{s}^T\boldsymbol{\mu}} \, M_{\mathbf{Z}}(\mathbf{u})
\end{aligned}$$

- Since Z_i are independent, easy to get $M_{\mathbf{Z}}$.
- We know $Z_i \sim \mathcal{N}(0, \frac{1}{m_i})$. Hence

$$M_{Z_i}(u_i) = e^{\frac{1}{2}\frac{1}{m_i}u_i^2} = e^{\frac{u_i^2}{2m_i}}$$

$$M_{\mathbf{Z}}(\mathbf{u}) = E\left[e^{\mathbf{u}^T \mathbf{Z}}\right] = \prod_{i=1}^n E\left[e^{u_i Z_i}\right] = \prod_{i=1}^n e^{\frac{u_i^2}{2m_i}} = e^{\sum_i \frac{u_i^2}{2m_i}}$$

- We derived earlier

$$M_{\mathbf{X}}(\mathbf{s}) = e^{\mathbf{s}^T\boldsymbol{\mu}} \, M_{\mathbf{Z}}(\mathbf{u}), \quad \text{where} \quad \mathbf{u} = L^T \mathbf{s}$$

- We got

$$M_{\mathbf{X}}(\mathbf{s}) = e^{\mathbf{s}^T \boldsymbol{\mu}} \; M_{\mathbf{Z}}(\mathbf{u}); \quad \mathbf{u} = L^T \mathbf{s}; \quad M_{\mathbf{Z}}(\mathbf{u}) = e^{\sum_i \frac{u_i^2}{2m_i}}$$

- Earlier we have shown $L^T M^{-1} L = \text{diag}(\frac{1}{m_1}, \cdots, \frac{1}{m_n})$ where $M^{-1} = \Sigma$. Now we get

$$\frac{1}{2} \sum_i \frac{u_i^2}{m_i} = \frac{1}{2} \mathbf{u}^T (L^T M^{-1} L) \mathbf{u} = \frac{1}{2} \mathbf{s}^T M^{-1} \mathbf{s} = \frac{1}{2} \mathbf{s}^T \Sigma \mathbf{s}$$

- Hence we get

$$M_{\mathbf{X}}(\mathbf{s}) = e^{\mathbf{s}^T \boldsymbol{\mu} + \frac{1}{2} \mathbf{s}^T \Sigma \mathbf{s}}$$

- This is the moment generating function of multi-dimensional Normal density

- Let X, Y be jointly Gaussian. For simplicity let $EX = EY = 0$.
- Let Var$(X) = \sigma_x^2$, Var$(Y) = \sigma_y^2$ and $\rho_{XY} = \rho$. $\Rightarrow$ Cov$(X, Y) = \rho \sigma_x \sigma_y$.
- Now, the covariance matrix and its inverse are given by

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho \sigma_x \sigma_y \\ \rho \sigma_x \sigma_y & \sigma_y^2 \end{bmatrix}; \quad \Sigma^{-1} = \frac{1}{\sigma_x^2 \sigma_y^2 (1 - \rho^2)} \begin{bmatrix} \sigma_y^2 & -\rho \sigma_x \sigma_y \\ -\rho \sigma_x \sigma_y & \sigma_x^2 \end{bmatrix}$$

- The joint density of X, Y is given by

$$f_{XY}(x, y) = \frac{1}{2\pi \sigma_x \sigma_y \sqrt{1 - \rho^2}} \; e^{-\frac{1}{2(1-\rho^2)} \left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2\rho x y}{\sigma_x \sigma_y} \right)}$$

- This is the bivariate Gaussian density

- Suppose X, Y are jointly Gaussian (with the density above)
- Then, all the marginals and conditionals would be Gaussian.
- $X \sim \mathcal{N}(0, \sigma_x^2)$, and $Y \sim \mathcal{N}(0, \sigma_y^2)$
- $f_{X|Y}(x|y)$ would be a Gaussian density with mean $y \rho \frac{\sigma_x}{\sigma_y}$ and variance $\sigma_x^2 (1 - \rho^2)$.
- Exercise for you – show all this starting with the joint density we have
- Note that X, Y are individually Gaussian does not mean they are jointly Gaussian (unless they are independent)

- The multi-dimensional Gaussian density has some important properties.
- If $X_1, \cdots, X_n$ are jointly Gaussian then they are independent if they are uncorrelated.
- Suppose $X_1, \cdots, X_n$ be jointly Gaussian and have zero means. Then there is an orthogonal transform $\mathbf{Y} = A\mathbf{X}$ such that $Y_1, \cdots, Y_n$ are jointly Gaussian and independent.
- $X_1, \cdots, X_n$ are jointly Gaussian if and only if $\mathbf{t}^T \mathbf{X}$ is Gaussian for for all non-zero $\mathbf{t} \in \Re^n$.
- We will prove this using moment generating functions

Recap: Multi-dimensional Gaussian density

- $\mathbf{X} = (X_1, \cdots, X_n)^T$ are said to be jointly Gaussian if

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

- $E\mathbf{X} = \boldsymbol{\mu}$ and $\Sigma_X = \Sigma$.
- The moment generating function is given by

$$M_{\mathbf{X}}(\mathbf{s}) = e^{\mathbf{s}^T \boldsymbol{\mu} + \frac{1}{2}\mathbf{s}^T \Sigma\, \mathbf{s}}$$

- When X, Y are jointly Gaussian, the joint densityis given by

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right)}$$

- The multi-dimensional Gaussian density has some important properties.
- If $X_1, \cdots, X_n$ are jointly Gaussian then they are independent if they are uncorrelated.
- Suppose $X_1, \cdots, X_n$ be jointly Gaussian and have zero means. Then there is an orthogonal transform $\mathbf{Y} = A\mathbf{X}$ such that $Y_1, \cdots, Y_n$ are jointly Gaussian and independent.
- $X_1, \cdots, X_n$ are jointly Gaussian if and only if $\mathbf{t}^T\mathbf{X}$ is Gaussian for for all non-zero $\mathbf{t} \in \Re^n$.
- We will prove this using moment generating functions

- Suppose $\mathbf{X} = (X_1, \cdots, X_n)^T$ be jointly Gaussian and let $W = \mathbf{t}^T\mathbf{X}$.
- Let μ_X and Σ_X denote the mean vector and covariance matrix of $\mathbf{X}$. Then

$$\mu_w \triangleq EW = \mathbf{t}^T \mu_X; \quad \sigma_w^2 \triangleq \mathrm{Var}(W) = \mathbf{t}^T \Sigma_X \mathbf{t}$$

- The mgf of W is given by

$$\begin{aligned} M_W(u) &= E\left[e^{uW}\right] = E\left[e^{u\,\mathbf{t}^T\mathbf{X}}\right] \\ &= M_X(u\mathbf{t}) = e^{u\mathbf{t}^T\mu_x + \frac{1}{2}u^2\mathbf{t}^T\Sigma_x\mathbf{t}} \\ &= e^{u\mu_w + \frac{1}{2}u^2\sigma_w^2} \end{aligned}$$

  showing that W is Gaussian
- Shows density of X_i is Gaussian for each i. For example, if we take $\mathbf{t} = (1, 0, 0, \cdots, 0)^T$ then W above would be X_1.

- Now suppose $W = \mathbf{t}^T\mathbf{X}$ is Gaussian for all $\mathbf{t}$.

$$M_W(u) = e^{u\mu_w + \frac{1}{2}u^2\sigma_w^2} = e^{u\,\mathbf{t}^T\mu_X + \frac{1}{2}u^2\,\mathbf{t}^T\Sigma_X\mathbf{t}}$$

- This implies

$$\begin{aligned} E\left[e^{u\,\mathbf{t}^T\mathbf{X}}\right] &= e^{u\,\mathbf{t}^T\mu_X + \frac{1}{2}u^2\,\mathbf{t}^T\Sigma_X\mathbf{t}},\ \forall u \in \Re, \forall \mathbf{t} \in \Re^n,\ \mathbf{t} \neq 0 \\ E\left[e^{\mathbf{t}^T\mathbf{X}}\right] &= e^{\mathbf{t}^T\mu_X + \frac{1}{2}\mathbf{t}^T\Sigma_X\mathbf{t}},\ \forall \mathbf{t} \end{aligned}$$

  This implies $\mathbf{X}$ is jointly Gaussian.
- This is a defining property of multidimensional Gaussian density

- Let $\mathbf{X} = (X_1, \cdots, X_n)^T$ be jointly Gaussian.
- Let A be a $k \times n$ matrix with rank k.
- Then $\mathbf{Y} = A\mathbf{X}$ is jointly Gaussian.
- We will once again show this using the moment generating function.
- Let μ_x and Σ_x denote mean vector and covariance matrix of $\mathbf{X}$. Similarly μ_y and Σ_y for $\mathbf{Y}$
- We have $\mu_y = A\mu_x$ and

$$\begin{aligned}
\Sigma_y &= E\left[(\mathbf{Y} - \mu_y)(\mathbf{Y} - \mu_y)^T\right] \\
&= E\left[(A(\mathbf{X} - \mu_x))(A(\mathbf{X} - \mu_x))^T\right] \\
&= E\left[A(\mathbf{X} - \mu_x)(\mathbf{X} - \mu_x)^T A^T\right] \\
&= A\, E\left[(\mathbf{X} - \mu_x)(\mathbf{X} - \mu_x)^T\right] A^T = A\Sigma_x A^T
\end{aligned}$$

- The mgf of $\mathbf{Y}$ is

$$\begin{aligned}
M_Y(\mathbf{s}) &= E\left[e^{\mathbf{s}^T\mathbf{Y}}\right] \quad (\mathbf{s} \in \Re^k) \\
&= E\left[e^{\mathbf{s}^T A\, \mathbf{X}}\right] \\
&= M_X(A^T\mathbf{s}) \\
&\quad \text{(Recall } \; M_X(\mathbf{t}) = e^{\mathbf{t}^T\mu_x + \frac{1}{2}\mathbf{t}^T\Sigma_x\mathbf{t}}) \\
&= e^{\mathbf{s}^T A\mu_x + \frac{1}{2}\mathbf{s}^T A\, \Sigma_x A^T\mathbf{s}} \\
&= e^{\mathbf{s}^T\mu_y + \frac{1}{2}\mathbf{s}^T\Sigma_y\mathbf{s}}
\end{aligned}$$

This shows $\mathbf{Y}$ is jointly Gaussian

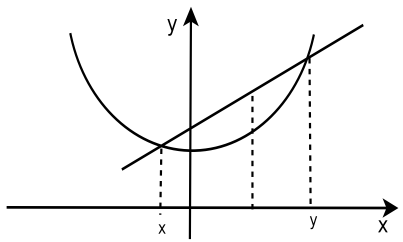
- $\mathbf{X}$ is jointly Gaussian and A is a $k \times n$ matrix with rank k.
- Then $\mathbf{Y} = A\mathbf{X}$ is jointly Gaussian.
- This shows all marginals of $\mathbf{X}$ are gaussian
- For example, if you take A to be

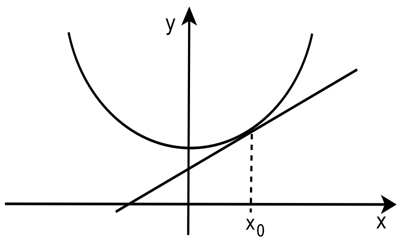
$$A = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \end{bmatrix}$$

then $\mathbf{Y} = (X_1, X_2)^T$

Jensen's Inequality

- Let $g : \Re \to \Re$ be a convex function. Then
$$g(EX) \le E[g(X)]$$
- For example, $(EX)^2 \le E\left[X^2\right]$
- Function g is convex if
$$g(\alpha x+(1-\alpha)y) \le \alpha g(x)+(1-\alpha)g(y), \quad \forall x, y, \quad \forall 0 \le \alpha \le 1$$
- If g is convex, then, given any x_0, exists $\lambda(x_0)$ such that
$$g(x) \ge g(x_0) + \lambda(x_0)(x - x_0), \ \forall x$$





Jensen's Inequality: Proof

- We have
$$g(x) \ge g(x_0) + \lambda(x_0)(x - x_0), \ \forall x$$
- Take $x_0 = EX$ and $x = X(\omega)$. Then
$$g(X(\omega)) \ge g(EX) + \lambda(EX)(X(\omega) - EX), \ \forall \omega$$
- $Y(\omega) \ge Z(\omega), \ \forall \omega \ \Rightarrow \ Y \ge Z \ \Rightarrow EY \ge EZ$
- Hence we get
$$\begin{aligned} g(X) &\ge g(EX) + \lambda(EX)(X - EX) \\ \Rightarrow E[g(X)] &\ge g(EX) + \lambda(EX)\ E[X - EX] = g(EX) \end{aligned}$$
- This completes the proof

- Consider the set of all mean-zero random variables.
- It is closed under addition and scalar (real number) multiplication.
- $\text{Cov}(X, Y) = E[XY]$ satisfies
  1. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
  2. $\text{Cov}(X, X) = \text{Var}(X) \ge 0$ and is zero only if $X = 0$
  3. $\text{Cov}(aX, Y) = a\text{Cov}(X, Y)$
  4. $\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y)$
- Thus $\text{Cov}(X, Y)$ is an inner product here.
- The Cauchy-Schwartz inequality ($|\mathbf{x}^T\mathbf{y}| \le ||\mathbf{x}||\ ||\mathbf{y}||$) gives
$$|\text{Cov}(X, Y)| \le \sqrt{\text{Cov}(X, X)\ \text{Cov}(Y, Y)} = \sqrt{\text{Var}(X)\ \text{Var}(Y)}$$
- This is same as $|\rho_{XY}| \le 1$
- A generalization of Cauchy-Schwartz inequality is Holder inequality

Holder Inequality

- For all p, q with $p, q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$
$$E[|XY|] \le (E|X|^p)^{\frac{1}{p}}\ (E|Y|^q)^{\frac{1}{q}}$$
(We assume all the expectations are finite)
- If we take $p = q = 2$
$$E[|XY|] \le \sqrt{E[X^2]\ E[Y^2]}$$
- This is same as Cauchy-Schwartz inequality. We once again get
$$\begin{aligned} \left|\text{Cov}(X, Y)\right| &= \left|E[(X - EX)(Y - EY)]\right| \\ &\le E\left[\left|(X - EX)(Y - EY)\right|\right] \\ &\le \sqrt{E[(X - EX)^2]\ E[(Y - EY)^2]} \\ &= \sqrt{\text{Var}(X)\ \text{Var}(Y)} \end{aligned}$$

Proof

- First we will show, for $p, q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$

$$|xy| \le \frac{|x|^p}{p} + \frac{|y|^q}{q}, \quad \forall x, y \in \Re$$

- For $x > 0$, $g(x) = -\log(x)$ is convex because $g''(x) = 1/x^2 \ge 0$, $\forall x$.
- Hence, for all $x_1, x_2 > 0$ and $0 \le t \le 1$,

$$\begin{aligned} -\log(tx_1 + (1-t)x_2) &\le -t\log(x_1) - (1-t)\log(x_2) \\ \Rightarrow \log(tx_1 + (1-t)x_2) &\ge \log\left(x_1^t\, x_2^{(1-t)}\right) \\ \Rightarrow tx_1 + (1-t)x_2 &\ge x_1^t\, x_2^{(1-t)} \end{aligned}$$

- We have for all $x_1, x_2 > 0$ and $0 \le t \le 1$,

$$tx_1 + (1-t)x_2 \ \ge \ x_1^t\, x_2^{(1-t)}$$

- Take $x_1 = |x|^p$, $x_2 = |y|^q$, $t = \frac{1}{p}$ (and hence $1 - t = \frac{1}{q}$)

$$\begin{aligned} (|x|^p)^{\frac{1}{p}}\ (|y|^q)^{\frac{1}{q}} &\le \frac{1}{p}\,|x|^p + \frac{1}{q}\,|y|^q \\ \Rightarrow\ |xy| &\le \frac{|x|^p}{p} + \frac{|y|^q}{q}, \quad \forall x, y \end{aligned}$$

$$|xy| \ \le \ \frac{|x|^p}{p} + \frac{|y|^q}{q}, \quad \forall x, y$$

- Take $x = X(\omega)\,(E|X|^p)^{-\frac{1}{p}}$, $y = Y(\omega)\,(E|Y|^q)^{-\frac{1}{q}}$

$$\begin{aligned} \frac{|X(\omega)Y(\omega)|}{(E|X|^p)^{\frac{1}{p}}\ (E|Y|^q)^{\frac{1}{q}}} &\le \frac{|X(\omega)|^p\ (E|X|^p)^{-1}}{p} + \frac{|Y(\omega)|^q\ (E|Y|^q)^{-1}}{q} \\ \Rightarrow \frac{|XY|}{(E|X|^p)^{\frac{1}{p}}\ (E|Y|^q)^{\frac{1}{q}}} &\le \frac{|X|^p\ (E|X|^p)^{-1}}{p} + \frac{|Y|^q\ (E|Y|^q)^{-1}}{q} \\ \Rightarrow \frac{E|XY|}{(E|X|^p)^{\frac{1}{p}}\ (E|Y|^q)^{\frac{1}{q}}} &\le \frac{1}{p} + \frac{1}{q} = 1 \\ \Rightarrow E|XY| &\le (E|X|^p)^{\frac{1}{p}}\ (E|Y|^q)^{\frac{1}{q}} \end{aligned}$$

- **Jensen's Inequality**: If g is convex and EX and $E[g(X)]$ exist

$$g(EX) \ \le \ E[g(X)]$$

- **Holder Inequality**: For $p, q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$

$$E|XY| \ \le \ (E|X|^p)^{\frac{1}{p}}\ (E|Y|^q)^{\frac{1}{q}}$$

  (assuming all expectations exist)
- For $p = q = 2$, the above is Cauchy-Schwartz inequality
- This implies $|\rho_{XY}| \le 1$
- **Minkowski's Inequality**:

$$(E|X+Y|^r)^{\frac{1}{r}} \ \le \ (E|X|^r)^{\frac{1}{r}} \ + \ (E|Y|^r)^{\frac{1}{r}}$$

Chernoff Bounds

- Recall Markov inequality. If h is positive, strictly increasing

$$P[X > a] \; = \; P[h(X) > h(a)] \; \leq \; \frac{E[h(X)]}{h(a)}$$

- Take $h(x) = e^{sx}, \; s > 0$. Then

$$P[X > a] \; \leq \; \frac{E[e^{sX}]}{e^{sa}} \; = \; \frac{M_X(s)}{e^{sa}}, \; \forall s > 0$$

- The RHS is a function of S. We can get a tight bound by using a value of s which minimizes RHS.

Hoeffding Inequality

- Often we need to deal with sums of iid random variables.
- Here is a simple version of an inequality very useful in such situations.
- Let X_i be iid and let $X_i \in [a, \; b], \; \forall i$. Let $EX_i = \mu$

$$P\left[\left|\sum_{i=1}^{n} X_i - n\mu\right| \geq \epsilon\right] \leq 2e^{-\frac{2\epsilon^2}{n(b-a)}}, \; \epsilon > 0$$

- Note we do not need knowledge of any moments of X_i to calculate the bound

- Let $X_1, X_2, \cdots$ be iid random variables
- Let $EX_i = \mu$ and let $\mathsf{Var}(X_i) = \sigma^2$
- Define $S_n = \sum_{i=1}^{n} X_i$. Then

$$ES_n = \sum_{i=1}^{n} EX_i = n\mu; \quad \text{and} \quad \mathsf{Var}(S_n) = \sum_{i=1}^{n} \mathsf{Var}(X_i) = n\sigma^2$$

- We are interested in $\frac{S_n}{n}$, average of $X_1, \cdots, X_n$.

$$E\left[\frac{S_n}{n}\right] = \frac{1}{n} ES_n = \mu, \quad \forall n$$

$$\mathsf{Var}\left(\frac{S_n}{n}\right) = \left(\frac{1}{n}\right)^2 \mathsf{Var}(S_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}, \quad \forall n$$

Weak Law of large numbers

- X_i are iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$

$$E\left[\frac{S_n}{n}\right] = \mu; \quad \text{and} \quad \text{Var}\left(\frac{S_n}{n}\right) = \frac{\sigma^2}{n}$$

- As n becomes large, variance of $\frac{S_n}{n}$ becomes close to zero
- $\frac{S_n}{n}$ 'converges' to its expectation, μ, as $n \to \infty$
- By Chebyshev Inequality

$$P\left[\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right] \leq \frac{\text{Var}(\frac{S_n}{n})}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}, \quad \forall \epsilon > 0$$

- Thus, we get

$$\lim_{n\to\infty} P\left[\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right] = 0, \quad \forall \epsilon > 0$$

- Known as weak law of large numbers

- Suppose we are tossing a (biased) coin repeatedly
- $X_i = 1$ if i^{th} toss came up head and is zero otherwise.
- $EX_i = p$ where p is the probability of heads. Variance of X_i is $p(1-p)$
- $S_n = \sum_{i=1}^n X_i$ is the number of heads in n tosses
- $\frac{S_n}{n}$ is the fraction of heads in n tosses.
- We are saying $\frac{S_n}{n}$ 'converges' to p
- The probability of head is the limiting fraction of heads when you toss the coin infinite times

$$\lim_{n\to\infty} P\left[\left|\frac{S_n}{n} - p\right| \geq \epsilon\right] = 0, \quad \forall \epsilon > 0$$

- This is true of any event.
- Consider repeatedly performing a random experiment
- X_i be the indicator of event A on i^{th} repetition
- Then $EX_i = P(A), \forall i$
- $\frac{S_n}{n}$ is the fraction of times the event A occurred.
- The fraction of times an event occurs 'converges' to its probability as you repeat the experiment infinite times

- X is a random variable and we want to find EX.
- Make multiple independent observations of X. Call them $X_1, \cdots, X_n$.
- These are called samples of X. $S_n = \sum_{i=1}^n X_i$
- $\frac{S_n}{n}$ is the sample mean – average of all samples.
- $\frac{S_n}{n}$ has the same expectation as X but has much smaller variance.
- Sample mean 'converges' to expectation ('population mean')
- This is the principle of sample surveys
- In general one can get an approximate value of expectation of X through simulations/experiments
- Known as Monte Carlo simulations

- X_i are iid, $EX_i = \mu$, Var$(X_i) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$

$$E\left[\frac{S_n}{n}\right] = \mu; \quad \text{and} \quad \text{Var}\left(\frac{S_n}{n}\right) = \frac{\sigma^2}{n}$$

- As n becomes large, variance of $\frac{S_n}{n}$ becomes close to zero
- We would like to say $\frac{S_n}{n} \to \mu$.
- We need to properly define convergence of a sequence of random variables
- One way of looking at this convergence is

$$\lim_{n\to\infty} P\left[\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right] = 0, \quad \forall \epsilon > 0$$

- There are other ways of defining convergence of random variables

- Recall convergence of real number sequences.
- A sequence of real numbers x_n is said to converge to x_0, $x_n \to x_0$, if

$$\forall \epsilon > 0, \ \exists N < \infty, \ s.t. \ |x_n - x_0| \leq \epsilon, \ \forall n \geq N$$

- To show a sequence converges using this definition, we need to know (or guess) the limit.
- Convergent sequences of real numbers satisfy the Cauchy criterion

$$\forall \epsilon > 0, \ \exists N < \infty, \ s.t. |x_n - x_m| \leq \epsilon, \ \forall n, m \geq N$$

- Now consider defining sequence of random variables X_n converging to X_0
- These are not numbers. They are, in fact functions.
- We know that $|X_n - X_0| \leq \epsilon$ is an event. We can define convergence in terms of probability of that event becoming 1.
- Or we can look at different notions of convergence of a sequence of functions to a function.

- Consider a sequence of functions g_n mapping $\Re$ to $\Re$.
- We can say $g_n \to g_0$ if $g_n(x) \to g_0(x), \ \forall x$.
- This is known as point-wise convergence
- Or we can ask for $\int |g_n(x) - g_0(x)|^2 \ dx \to 0$.
- There are multiple notions of convergence that are reasonable for a sequence of functions.
- Thus there would be multiple ways to define convergence of sequence of random variables.

Convergence in Probability

- A sequence of random variables, X_n, is said to **converge in probability** to a random variable X_0 is

$$\lim_{n\to\infty} P\left[|X_n - X_0| > \epsilon\right] = 0, \ \forall \epsilon > 0$$

This is denoted as $X_n \xrightarrow{P} X_0$

- We would mostly be considering convergence to a constant.
- By the definition of limit, the above means

$$\forall \delta > 0, \ \exists N < \infty, \ s.t. \ P\left[|X_n - X_0| > \epsilon\right] < \delta, \ \forall n > N$$

- We only need marginal distributions of individual X_n to decide whether a sequence converges to a constant in probability

Example: Partial sums of iid random variables

- X_i are iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$
- Then we saw

$$P\left[\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right] \leq \frac{\sigma^2}{n\epsilon^2}, \quad \forall \epsilon > 0$$

- Hence we have $\frac{S_n}{n} \xrightarrow{P} \mu$
- Weak law of large numbers says that sample mean converges in probability to the expectation

Example

- Let $\Omega = [0,\ 1]$ with the usual probability measure and let $X_n = I_{[0,\ 1/n]}$.
- $P[X_n = 1] = \frac{1}{n} = 1 - P[X_n = 0]$
- The probability of X_n taking value 1 is decreasing with n
- A good guess is that it converges to zero

$$P[|X_n - 0| > \epsilon] = P[X_n = 1] = \frac{1}{n}$$

  which goes to zero as $n \to \infty$.
- Hence, $X_n \xrightarrow{P} 0$

Example

- Let $X_1, X_2, \cdots$ be a sequence of iid random variable which are uniform over $(0,\ 1)$.
- Let $M_n = \max(X_1, X_2, \cdots, X_n)$
- Does M_n converge in probability?
- A reasonable guess for the limit is 1

$$P\left[|M_n - 1| \geq \epsilon\right] = P\left[M_n \leq 1 - \epsilon\right] = (1-\epsilon)^n$$

- This implies $M_n \xrightarrow{P} 1$
- Suppose $Z_n = \min(X_1, X_2, \cdots, X_n)$.
  Then $Z_n \xrightarrow{P} 0$

Some properties of convergence in probability

- $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{P} Y \;\Rightarrow\; P[X = Y] = 1$
- $X_n \xrightarrow{P} X \;\Rightarrow\; P[|X_n - X_m| > \epsilon] \to 0$ as $n, m \to \infty$
- Suppose $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$ Then the following hold
  1. $aX_n \xrightarrow{P} aX$
  2. $X_n + Y_n \xrightarrow{P} X + Y$
  3. $X_nY_n \xrightarrow{P} XY$
- We omit the proofs

Recap: Multi-dimensional Gaussian density

- $\mathbf{X} = (X_1, \cdots, X_n)^T$ are said to be jointly Gaussian if

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} \, |\Sigma|^{\frac{1}{2}}} \; e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

- $E\mathbf{X} = \boldsymbol{\mu}$ and $\Sigma_X = \Sigma$.
- The moment generating function is given by

$$M_{\mathbf{X}}(\mathbf{s}) = e^{\mathbf{s}^T \boldsymbol{\mu} \, + \, \frac{1}{2}\mathbf{s}^T \Sigma \, \mathbf{s}}$$

- When X, Y are jointly Gaussian, the joint density is given by

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \; e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right)}$$

Recap

- If $X_1, \cdots, X_n$ are jointly Gaussian then they are independent if they are uncorrelated.
- When $X_1, \cdots, X_n$ be jointly Gaussian (with zero means), there is an orthogonal transform $\mathbf{Y} = A\mathbf{X}$ such that $Y_1, \cdots, Y_n$ are jointly Gaussian and independent.
- $X_1, \cdots, X_n$ are jointly Gaussian if and only if $\mathbf{t}^T\mathbf{X}$ is Gaussian for for all non-zero $\mathbf{t} \in \Re^n$.
- If $X_1, \cdots, X_n$ are jointly Gaussian and A is a $k \times n$ matrix of rank k, then, $\mathbf{Y} = A\mathbf{X}$ is jointly gaussian

Recap: Moment inequalities

- **Jensen's Inequality**: If g is convex and EX and $E[g(X)]$ exist

$$g(EX) \;\leq\; E[g(X)]$$

- **Holder Inequality**: For $p, q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$

$$E|XY| \;\leq\; (E|X|^p)^{\frac{1}{p}} \; (E|Y|^q)^{\frac{1}{q}}$$

(assuming all expectations exist)
- For $p = q = 2$, the above is Cauchy-Schwartz inequality
- This implies $|\rho_{XY}| \leq 1$
- **Minkowski's Inequality**:

$$(E|X+Y|^r)^{\frac{1}{r}} \;\leq\; (E|X|^r)^{\frac{1}{r}} \;+\; (E|Y|^r)^{\frac{1}{r}}$$

Recap

- **Chernoff Bounds**

$$P[X > a] \;\leq\; \frac{E[e^{sX}]}{e^{sa}} \;=\; \frac{M_X(s)}{e^{sa}}, \;\forall s > 0$$

- **Hoeffding Inequality** X_i iid, $X_i \in [a,\, b]$, $\forall i$ and $EX_i = \mu$

$$P\left[\left|\sum_{i=1}^{n} X_i - n\mu\right| \geq \epsilon\right] \leq 2e^{-\frac{2\epsilon^2}{n(b-a)}}, \;\epsilon > 0$$

Recap: Weak Law of large numbers

- X_i are iid, $EX_i = \mu$, Var$(X_i) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$

$$E\left[\frac{S_n}{n}\right] = \mu; \quad \text{and} \quad \text{Var}\left(\frac{S_n}{n}\right) = \frac{\sigma^2}{n}$$

- By Chebyshev Inequality

$$P\left[\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right] \leq \frac{\text{Var}(\frac{S_n}{n})}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}, \quad \forall \epsilon > 0$$

$$\Rightarrow \quad \lim_{n\to\infty} P\left[\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right] = 0, \quad \forall \epsilon > 0$$

Recap: Convergence in Probability

- A sequence of random variables, X_n, is said to **converge in probability** to a random variable X_0 is

$$\lim_{n\to\infty} P\left[|X_n - X_0| > \epsilon\right] = 0, \ \forall \epsilon > 0$$

This is denoted as $X_n \stackrel{P}{\to} X_0$

- By the definition of limit, the above means

$$\forall \delta > 0, \ \exists N < \infty, \ s.t. \ P\left[|X_n - X_0| > \epsilon\right] < \delta, \ \forall n > N$$

- We only need marginal distributions of individual X_n to decide whether a sequence converges to a constant in probability

- We mentioned point-wise convergence of a sequence of functions

$$g_n \ \to \ g_0 \quad \text{if} \quad g_n(x) \ \to \ g_0(x), \ \ \forall x$$

- Since random variables are also functions we can define convergence like this.
- We can demand $X_n(\omega) \ \to \ X_0(\omega), \ \forall \omega$
- Such pointise convergence is too restrictive.
- But we can demand that it should be satisfied for almost all ω

- A sequence of random variables, X_n, is said to converge **almost surely** or **with probability one** to X if

$$P\left(\{\omega \ : \ X_n(\omega) \to X(\omega)\}\right) = 1$$

or equivalently

$$P\left(\{\omega \ : \ X_n(\omega) \nrightarrow X(\omega)\}\right) = 0$$

- Denoted as $X_n \stackrel{a.s.}{\to} X$ or $X_n \stackrel{w.p.1}{\to} X$ or $X_n \to X_0 \ \ (w.p.1)$
- We are saying that for 'almost all' ω, $X_n(\omega)$ converges to $X(\omega)$
- We will first try and write the event $\{\omega \ : \ X_n(\omega) \nrightarrow X(\omega)\}$ in a proper form

- Recall convergence of real number sequences.
- A sequence of real numbers x_n is said to converge to x_0, $x_n \rightarrow x_0$, if

$$\forall \epsilon > 0, \ \exists N < \infty, \ s.t. \ |x_n - x_0| < \epsilon, \ \forall n \geq N$$

This is equivalent to

$$\forall \epsilon > 0, \ \exists N < \infty, \ \forall k \geq 0 \ |x_{N+k} - x_0| < \epsilon$$

- So, $x_n \nrightarrow x_0$ means

$$\exists \epsilon \ \ \forall N \ \ \exists k \ \ |x_{N+k} - x_0| \geq \epsilon$$

- Note that given any ω, $X_n(\omega)$ is real number sequence.
- Hence $X_n(\omega) \rightarrow X(\omega)$ is same as

$$\forall \epsilon > 0 \ \ \exists N < \infty \ \ \forall k \geq 0 \ \ |X_{N+k}(\omega) - X(\omega)| < \epsilon$$

This is equivalent to

$$\forall r > 0, r \text{ integer} \ \ \exists N < \infty \ \ \forall k \geq 0 \ |X_{N+k}(\omega) - X(\omega)| < \frac{1}{r}$$

- Hence, $X_n(\omega) \nrightarrow X(\omega)$ is same as

$$\exists r \ \ \forall N \ \ \exists k \, |X_{N+k}(\omega) - X(\omega)| \geq \frac{1}{r}$$

- Hence we can write this event as

$$\cup_{r=1}^{\infty} \cap_{N=1}^{\infty} \cup_{k=0}^{\infty} \left\{ \omega \ : \ |X_{N+k}(\omega) - X(\omega)| \geq \frac{1}{r} \right\}$$

- The event $\{\omega \ : \ X_n(\omega) \nrightarrow X(\omega)\}$ can be expressed as

$$\cup_{r=1}^{\infty} \cap_{N=1}^{\infty} \cup_{k=0}^{\infty} \left[|X_{N+k} - X| \geq \frac{1}{r} \right]$$

- Hence X_n converges almost surely to X iff

$$P\left(\cup_{r=1}^{\infty} \cap_{N=1}^{\infty} \cup_{k=0}^{\infty} \left[|X_{N+k} - X| \geq \frac{1}{r} \right] \right) = 0$$

- This is same as

$$P\left(\cap_{N=1}^{\infty} \cup_{k=0}^{\infty} \left[|X_{N+k} - X| \geq \frac{1}{r} \right] \right) = 0, \ \ \forall r > 0, \text{ integer}$$

- Same as

$$P\left(\cap_{N=1}^{\infty} \cup_{k=0}^{\infty} \ [|X_{N+k} - X| \geq \epsilon] \right) = 0, \ \ \forall \epsilon > 0$$

- A sequence X_n is said to converge **almost surely** or **with probability one** to X if

$$P\left(\{\omega \ : \ X_n(\omega) \rightarrow X(\omega)\}\right) = 1$$

- We can also write it as

$$P[X_n \rightarrow X] = 1$$

- We showed that this is equivalent to

$$P\left(\cap_{N=1}^{\infty} \cup_{k=0}^{\infty} \ [|X_{N+k} - X| \geq \epsilon] \right) = 0, \ \ \forall \epsilon > 0$$

- Same as

$$P\left(\cap_{N=1}^{\infty} \cup_{k=N}^{\infty} \ [|X_k - X| \geq \epsilon] \right) = 0, \ \ \forall \epsilon > 0$$

- let $A_k = [|X_k - X| \geq \epsilon]$
- Let $B_N = \cup_{k=N}^{\infty} A_k$.
- Then, $B_{N+1} \subset B_N$ and hence $B_N \downarrow$.
- Hence, $\lim B_N = \cap_{N=1}^{\infty} B_N$.
- We saw that $X_n \stackrel{a.s.}{\to} X$ is same as

$$P\left(\cap_{N=1}^{\infty} \cup_{k=N}^{\infty} \; [|X_k - X| \geq \epsilon]\right) = 0, \quad \forall \epsilon > 0$$

$$\Leftrightarrow \quad P\left(\lim_{N \to \infty} \cup_{k=N}^{\infty} \; [|X_k - X| \geq \epsilon]\right) = 0, \quad \forall \epsilon > 0$$

$$\Leftrightarrow \lim_{N \to \infty} \; P\left(\cup_{k=N}^{\infty} \; [|X_k - X| \geq \epsilon]\right) = 0, \quad \forall \epsilon > 0$$

- X_n converges to X almost surely iff

$$\lim_{n \to \infty} \; P\left(\cup_{k=n}^{\infty} \; [|X_k - X| \geq \epsilon]\right) = 0, \quad \forall \epsilon > 0$$

- To show convergence with probability one using this one needs to know the joint distribution of $X_n, X_{n+1}, \cdots$
- Contrast this with $X_n \stackrel{P}{\to} X$ which is

$$\lim_{n \to \infty} P\left[|X_n - X| > \epsilon\right] = 0, \; \forall \epsilon > 0$$

- This also shows that

$$X_n \stackrel{a.s.}{\to} X \quad \Rightarrow \quad X_n \stackrel{P}{\to} X$$

- Almost sure convergence is a stronger mode of convergence

simple example: almost sure convergence

- Let $\Omega = [0, \; 1]$ with the usual probability measure and let $X_n = I_{[0, \; 1/n]}$.

$$X_n(\omega) = \begin{cases} 1 & \omega \leq \frac{1}{n} \\ 0 & \text{otherwise} \end{cases}$$

- Since $X_n \stackrel{P}{\to} 0$, zero is the only candidate for limit
- $X_n(\omega) = 1$ only when $n \leq 1/\omega$.
- Given any ω, for all $n > 1/\omega$, $\quad X_n(\omega) = 0$
- Hence, $\{\omega \; : \; X_n(\omega) \to 0\} = (0, 1]$

$$P[X_n \to X_0] = P(\{\omega \; : \; X_n(\omega) \to 0\}) = P((0, 1]) = 1$$

- Hence $X_n \stackrel{a.s}{\to} 0$

- X_n converges to X almost surely iff

$$P\left(\cap_{N=1}^{\infty} \cup_{k=N}^{\infty} \; [|X_k - X| \geq \epsilon]\right) = 0, \quad \forall \epsilon > 0$$

$$\Leftrightarrow \quad \lim_{n \to \infty} \; P\left(\cup_{k=n}^{\infty} \; [|X_k - X| \geq \epsilon]\right) = 0, \quad \forall \epsilon > 0$$

- We normally do not specify X_n as functions over Ω
- We are only given the distributions
- How do we establish convergence almost surely

- Let $A_1, A_2, \cdots$ be a sequence of events.
- How do we define limit of this sequence ?
- Define sequences

$$B_n = \cup_{k=n}^{\infty} A_k \qquad C_n = \cap_{k=n}^{\infty} A_k$$

- These are monotone: $B_n \downarrow$, $C_n \uparrow$. They have limits.
- Define

$$\lim\ \sup A_n \triangleq \lim\ B_n = \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k$$
$$\lim\ \inf A_n \triangleq \lim\ C_n = \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k$$

- If $\lim\ \sup A_n = \lim\ \inf A_n$ then we define that as $\lim A_n$. Otherwise we say the sequence does not have a limit
- Note that $\lim\ \sup A_n$ and $\lim\ \inf A_n$ are events
- Note that $X_n \stackrel{a.s.}{\to} X$ iff

$$P\left(\cap_{N=1}^{\infty} \cup_{k=N}^{\infty} \left[|X_k - X| \ge \epsilon\right]\right) = 0, \quad \forall \epsilon > 0$$

$$\Leftrightarrow \quad P\left(\lim\ \sup\ \left[|X_n - X| \ge \epsilon\right]\right) = 0, \quad \forall \epsilon > 0$$

- We can show $\lim\ \inf A_n \subset \lim\ \sup A_n$

$$\begin{aligned}
\omega \in \lim\ \inf A_n \quad &\Rightarrow \quad \omega \in \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k \\
&\Rightarrow \quad \exists m,\ \omega \in A_k,\ \forall k \ge m \\
&\Rightarrow \quad \omega \in \cup_{j=n}^{\infty} A_j,\ \forall n \\
&\Rightarrow \quad \omega \in \cap_{n=1}^{\infty} \cup_{j=n}^{\infty} A_j \\
&\Rightarrow \quad \omega \in \lim\ \sup A_n
\end{aligned}$$

- We can characterize $\lim\ \inf A_n$ as follows

$$\begin{aligned}
\omega \in \lim\ \inf A_n \quad &\Rightarrow \quad \omega \in \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k \\
&\Rightarrow \quad \exists m,\ \omega \in A_k,\ \forall k \ge m \\
&\Rightarrow \quad \omega \text{ belongs to all but finitely many of } A_n
\end{aligned}$$

Thus, $\lim\ \inf A_n$ consists of all points that are there in all but finitely many A_n.

- We can characterize $\lim\ \sup A_n$ as follows

$$\begin{aligned}
\omega \in \lim\ \sup A_n \quad &\Rightarrow \quad \omega \in \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k \\
&\Rightarrow \quad \omega \in \cup_{k=n}^{\infty} A_k,\ \forall n \\
&\Rightarrow \quad \omega \text{ belongs to infinitely many of } A_n
\end{aligned}$$

Thus $\lim\ \sup A_n$ consists of points that are in infinitely many A_n
One refers to $\lim\ \sup A_n$ also as 'A_n infinitely often' or 'A_n $i.o.$'
- What is the difference between
Points that belong to all but finitely many A_n and
Points that belong to infinitely many A_n
- There can be ω that are there in infinitely many of A_n and are also not there in infinitely many of the A_n

Example

- Consider the following sequence of sets: $A, B, A, B, \cdots$
- Recall

$$\lim\ \sup A_n = \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k \quad \lim\ \inf A_n = \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k$$

$$\cup_{k=n}^{\infty} A_k = A \cup B,\ \forall n \quad \Rightarrow \quad \lim\ \sup A_n = A \cup B$$

$$\cap_{k=n}^{\infty} A_k = A \cap B,\ \forall n \quad \Rightarrow \quad \lim\ \inf A_n = A \cap B$$

example

- Consider the sets $A_n = [0,\ 1 + \frac{(-1)^n}{n})$
  The sequence is

$$[0,\ 0),\ \left[0,\ 1 + \frac{1}{2}\right),\ \left[0,\ 1 - \frac{1}{3}\right),\ \left[0,\ 1 + \frac{1}{4}\right) \cdots$$

- Guess: $\lim\sup A_n = [0,\ 1]$ and $\lim\inf A_n = [0,\ 1)$
- First note that $[0,\ 1 + \frac{1}{n+1}) \subset \cup_{k=n}^{\infty} A_k \subset [0,\ 1 + \frac{1}{n})$.
  Hence

$$x \in [0,\ 1]\ \Rightarrow x \in \cup_{k=n}^{\infty} A_k,\ \forall n\ \Rightarrow x \in \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k\ \Rightarrow x \in \lim\ \sup A_n$$

- Given any $\epsilon > 0$, $1 + \epsilon \notin [0,\ 1 + \frac{1}{n})$ if $\epsilon > \frac{1}{n}$ or $n > \frac{1}{\epsilon}$.
- Hence, given any $\epsilon > 0$, $\exists n$ such that $1 + \epsilon \notin \cup_{k=n}^{\infty} A_k$.
- This proves $\lim\sup A_n = [0,\ 1]$

- Now let us consider: $\lim\ \inf A_n = \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k$.
- Recall $A_n = [0,\ 1 + \frac{(-1)^n}{n})$
- First note that $[0,\ 1 - \frac{1}{n}) \subset \cap_{k=n}^{\infty} A_k \subset [0,\ 1 - \frac{1}{n+1})$
- Given any $\epsilon > 0$, $1 - \epsilon < 1 - \frac{1}{n}$ if $n > \frac{1}{\epsilon}$
- Hence, given any $\epsilon > 0$, $\exists n$ such that $1 - \epsilon \in \cap_{k=n}^{\infty} A_k$
- Hence $1 - \epsilon \in \lim\ \inf A_n$
- It is easy to see $1 \notin \cap_{k=n}^{\infty} A_k$ for ay n.
- Hence $1 \notin \lim\ \inf A_n$
- This proves $\lim\ \inf A_n = [0,\ 1)$
- Since $\lim\ \sup A_n \neq \lim\ \inf A_n$, this sequence does not have a limit

- $X_n \stackrel{a.s.}{\rightarrow} X$ iff

$$P\left(\cap_{N=1}^{\infty} \cup_{k=N}^{\infty}\ [|X_k - X| \geq \epsilon]\right) = 0, \quad \forall \epsilon > 0$$

- Let $A_n^{\epsilon} = [|X_n - X| \geq \epsilon]$
- Then $X_n \stackrel{a.s.}{\rightarrow} X$ iff

$$P\left(\lim\ \sup\ A_n^{\epsilon}\right) = 0,\ \forall \epsilon > 0$$

- The question now is: can we get probability of $\lim\ \sup A_n$
- We look at an important result that allows us to do this

Borel-Cantelli Lemma

- **Borel-Cantelli lemma**: Given sequence of events, $A_1, A_2, \cdots$
  1. If $\sum_{i=1}^{\infty} P(A_i) < \infty$, then, $P(\lim\ \sup A_n) = 0$
  2. If $\sum_{i=1}^{\infty} P(A_i) = \infty$ and A_i are independent, $P(\lim\ \sup A_n) = 1$

  Proof:
- We will first show: $P(\cup_{i=n}^{\infty} A_i) \leq \sum_{i=n}^{\infty} P(A_i), \ \forall n$
- We have the result: $P(\cup_{i=n}^{N} A_i) \leq \sum_{i=n}^{N} P(A_i), n \leq N$
- For any n, let $B_N = \cup_{i=n}^{N} A_i$. Then $B_N \subset B_{N+1}$.
- $\lim_{N\to\infty} B_N = \cup_{k=n}^{\infty} A_k$

$$P(\cup_{i=n}^{\infty} A_i) = P(\lim_{N\to\infty} \cup_{i=n}^{N} A_i) = \lim_{N\to\infty} P(\cup_{i=n}^{N} A_i) \leq \lim_{N\to\infty} \sum_{i=n}^{N} P(A_i) = \sum_{i=n}^{\infty} P(A_i)$$

- If $\sum_{k=1}^{\infty} P(A_k) < \infty$, then, $\lim_{n\to\infty} \sum_{k=n}^{\infty} P(A_k) = 0$

$$\begin{aligned} 0 \leq P(\lim\ \sup A_n) &= P\left(\cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k\right) \\ &= P\left(\lim_{n\to\infty} \cup_{k=n}^{\infty} A_k\right) \\ &= \lim_{n\to\infty} P\left(\cup_{k=n}^{\infty} A_k\right) \\ &\leq \lim_{n\to\infty} \sum_{k=n}^{\infty} P(A_k) \\ &= 0, \quad \text{if} \quad \sum_{k=1}^{\infty} P(A_k) < \infty \end{aligned}$$

- This completes proof of first part of Borel-Cantelli lemma

- Let $\sum_{k=1}^{\infty} P(A_k) = C < \infty$
- It means given any $\epsilon > 0$, $\exists n$

$$\left| \sum_{k=1}^{n} P(A_k) - C \right| < \epsilon \quad \Rightarrow \quad \left| \sum_{k=n}^{\infty} P(A_k) \right| < \epsilon$$

- This implies

$$\lim_{n\to\infty} \sum_{k=n}^{\infty} P(A_k) = 0$$

- For the second part of the lemma:

$$\begin{aligned} P(\lim\ \sup A_n) &= P\left(\cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k\right) \\ &= P\left(\lim_{n\to\infty} \cup_{k=n}^{\infty} A_k\right) \\ &= \lim_{n\to\infty} P\left(\cup_{k=n}^{\infty} A_k\right) \\ &= \lim_{n\to\infty} \left(1 - P\left(\cap_{k=n}^{\infty} A_k^c\right)\ \right) \\ &= \lim_{n\to\infty} \left(1 - \prod_{k=n}^{\infty} (1 - P(A_k))\right) \\ &\quad \text{because } A_k \text{ are independent} \\ &= 1 \ - \ \lim_{n\to\infty} \prod_{k=n}^{\infty} (1 - P(A_k)) \end{aligned}$$

- We can compute that limit as follows

$$\lim_{n\to\infty}\prod_{k=n}^{\infty}(1-P(A_k)) \le \lim_{n\to\infty}\prod_{k=n}^{\infty}e^{-P(A_k)}, \quad \text{since } 1-x\le e^{-x}$$
$$= \lim_{n\to\infty} e^{-\sum_{k=n}^{\infty}P(A_k)}$$
$$= 0$$

because
$\sum_{k=1}^{\infty}P(A_k)=\infty \quad \Rightarrow \quad \lim_{n\to\infty}\sum_{k=n}^{\infty}P(A_k)=\infty$

- This finally gives us

$$P(\lim\ \sup A_n) = 1 - \lim_{n\to\infty}\prod_{k=n}^{\infty}(1-P(A_k)) = 1$$

- Given a sequence X_n we want to know whether it converges to X
- Let $A_k^\epsilon = [|X_k - X| \ge \epsilon]$
- $X_n \xrightarrow{P} X$ if

$$\lim_{k\to\infty}P[|X_k - X|\ge\epsilon]=0 \quad \text{same as} \quad \lim_{k\to\infty}P(A_k)=0, \ \forall\epsilon>0$$

- By Borel-Cantelli lemma

$$\sum_{k=1}^{\infty}P(A_k)<\infty \quad \Rightarrow \quad P(\lim\ \sup A_k)=0 \quad \Rightarrow \quad X_k \xrightarrow{a.s.} X$$

- Consider a sequence X_n with

$$P[X_n=0]=1-a_n; \quad P[X_n=c_n]=a_n$$

- We want to investigate convergence to 0.
- $A_n^\epsilon = [|X_n - 0| > \epsilon] = [X_n = c_n], \ \forall\epsilon>0$
- Hence $P(A_n^\epsilon)=a_n, \ \forall\epsilon>0$.
- If $a_n\to 0$ then $X_n \xrightarrow{P} 0$. (e.g., $a_n=\frac{1}{n},\frac{1}{n^2}$)
- If $\sum a_n<\infty$, $X_n \xrightarrow{a.s.} 0$ (e.g., $a_n=\frac{1}{n^2}$)

- Consider a sequence X_n with

$$P[X_n=0]=1-\frac{1}{n}; \quad P[X_n=1]=\frac{1}{n}$$

- We can easily conclude $X_n \xrightarrow{P} 0$.
- But since, $\sum_n \frac{1}{n}=\infty$, Borel-Cantelli lemma is not really useful here
- We saw one example where such X_n converge almost surely.
- But, if X_n are independent, then by Borel-Cantelli lemma, they do not converge
- Convergence (to a constant) in probability depends only on distribution of individual X_n.
- Convergence almost surely depends on the joint distribution

Strong Law of Large Numbers

- Let X_n be iid, $EX_n = \mu$, Var$(X_n) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$
- We saw weak law of large numbers:

$$\frac{S_n}{n} \xrightarrow{P} \mu$$

- Strong law of large numbers says:

$$\frac{S_n}{n} \xrightarrow{a.s.} \mu$$

- Let $A_n^\epsilon = \left[\left|\frac{S_n}{n} - \mu\right| > \epsilon\right]$
- As we saw, by Chebyshev inequality

$$P\left[\left|\frac{S_n}{n} - \mu\right| > \epsilon\right] \leq \frac{\sigma^2}{n\epsilon^2}$$

- This shows $P(A_n^\epsilon) \rightarrow 0$ and thus we get weak law
- To prove strong law using Borel-Cantelli lemma, we need $\sum P(A_n^\epsilon) < \infty$
- Since $\sum_n \frac{\sigma^2}{n\epsilon^2} = \infty$, the Chebyshev bound is not useful
- We need a bound: $P[|\frac{S_n}{n} - \mu|] \leq c_n$ such that $\sum_n c_n < \infty$.

Recap: Convergence in Probability

- A sequence of random variables, X_n, is said to **converge in probability** to a random variable X_0 is

$$\lim_{n\rightarrow\infty} P\left[|X_n - X_0| > \epsilon\right] = 0, \ \forall\epsilon > 0$$

This is denoted as $X_n \xrightarrow{P} X_0$
- By the definition of limit, the above means

$$\forall\delta > 0, \ \exists N < \infty, \ s.t. \ P\left[|X_n - X_0| > \epsilon\right] < \delta, \ \forall n > N$$

- We only need marginal distributions of individual X_n to decide whether a sequence converges to a constant in probability

Recap: Weak Law of large numbers

- X_i are iid, $EX_i = \mu$, Var$(X_i) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$

$$E\left[\frac{S_n}{n}\right] = \mu; \quad \text{and} \quad \text{Var}\left(\frac{S_n}{n}\right) = \frac{\sigma^2}{n}$$

Weak law of large numbers states

$$\frac{S_n}{n} \xrightarrow{P} \mu$$

Recap: almost sure convergence

- A sequence of random variables, X_n, is said to converge **almost surely** or **with probability one** to X if

$$P\left(\{\omega \ : \ X_n(\omega) \to X(\omega)\}\right) = 1$$

or equivalently

$$P\left(\{\omega \ : \ X_n(\omega) \nrightarrow X(\omega)\}\right) = 0$$

- Denoted as $X_n \stackrel{a.s.}{\to} X$ or $X_n \stackrel{w.p.1}{\to} X$ or $X_n \to X_0 \ \ (w.p.1)$
- We can also write it as

$$P[X_n \to X] = 1$$

Recap

- The sequence X_n converges to X almost surely iff

$$P\left(\cap_{N=1}^{\infty} \cup_{k=0}^{\infty} \ [|X_{N+k} - X| \geq \epsilon]\right) = 0, \ \ \forall \epsilon > 0$$

Same as

$$P\left(\cap_{N=1}^{\infty} \cup_{k=N}^{\infty} \ [|X_k - X| \geq \epsilon]\right) = 0, \ \ \forall \epsilon > 0$$

- Equivalently

$$\lim_{n\to\infty} \ P\left(\cup_{k=n}^{\infty} \ [|X_k - X| \geq \epsilon]\right) = 0, \ \ \forall \epsilon > 0$$

- $X_n \stackrel{P}{\to} X$ iff

$$\lim_{n\to\infty} P\left[|X_n - X| > \epsilon\right] = 0, \ \forall \epsilon > 0$$

$$X_n \stackrel{a.s.}{\to} X \quad \Rightarrow \quad X_n \stackrel{P}{\to} X$$

- Almost sure convergence is a stronger mode of convergence

Recap: lim sup and lim inf

- Let $A_1, A_2, \cdots$ be a sequence of events.
- We define

$$\lim\ \sup A_n \ \triangleq \ \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k$$
$$\lim\ \inf A_n \ \triangleq \ \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k$$

- If $\lim\ \sup A_n = \lim\ \inf A_n$ then that is $\lim A_n$. Otherwise the sequence does not have a limit
- $\lim\ \sup A_n$ and $\lim\ \inf A_n$ are events
- $\lim\ \inf A_n \ \subset \ \lim\ \sup A_n$

Recap

- $X_n \stackrel{a.s.}{\to} X$ iff

$$P\left(\cap_{N=1}^{\infty} \cup_{k=N}^{\infty} \ [|X_k - X| \geq \epsilon]\right) = 0, \ \ \forall \epsilon > 0$$

- Let $A_k^\epsilon = [|X_k - X| \geq \epsilon]$.
- Hence, $X_n \stackrel{a.s.}{\to} X$ iff

$$P\left(\lim\ \sup \ A_n^\epsilon\right) = 0, \ \ \forall \epsilon > 0$$

Recall: Borel-Cantelli Lemma

- **Borel-Cantelli lemma**: Given sequence of events, $A_1, A_2, \cdots$
  1. If $\sum_{i=1}^{\infty} P(A_i) < \infty$, then, $P(\lim\ \sup A_n) = 0$
  2. If $\sum_{i=1}^{\infty} P(A_i) = \infty$ and A_i are independent, $P(\lim\ \sup A_n) = 1$

- Given a sequence X_n we want to know whether it converges to X
- Let $A_k^\epsilon = [|X_k - X| \geq \epsilon]$
- $X_n \stackrel{a.s.}{\rightarrow} X$ if

$$P\left(\lim\ \sup\ A_n^\epsilon\right) = 0, \ \ \forall \epsilon > 0$$

- By Borel-Cantelli lemma

$$\sum_{k=1}^{\infty} P(A_k) < \infty \ \ \Rightarrow \ \ P(\lim\ \sup A_k) = 0 \ \ \Rightarrow \ \ X_k \stackrel{a.s.}{\rightarrow} X$$

If A_k are ind

$$\sum_{k=1}^{\infty} P(A_k) = \infty \ \ \Rightarrow \ \ P(\lim\ \sup A_k) = 1 \ \ \Rightarrow \ \ X_k \stackrel{a.s.}{\nrightarrow} X$$

Strong Law of Large Numbers

- Let X_n be iid, $EX_n = \mu$, Var$(X_n) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$
- We saw weak law of large numbers:

$$\frac{S_n}{n} \stackrel{P}{\rightarrow} \mu$$

- Strong law of large numbers says:

$$\frac{S_n}{n} \stackrel{a.s.}{\rightarrow} \mu$$

- Let $A_n^\epsilon = \left[\left|\frac{S_n}{n} - \mu\right| > \epsilon\right]$
- As we saw, by Chebyshev inequality

$$P\left[\left|\frac{S_n}{n} - \mu\right| > \epsilon\right] \leq \frac{\sigma^2}{n\epsilon^2}$$

- This shows $P(A_n^\epsilon) \rightarrow 0$ and thus we get weak law
- To prove strong law using Borel-Cantelli lemma, we need $\sum P(A_n^\epsilon) < \infty$
- Since $\sum_n \frac{\sigma^2}{n\epsilon^2} = \infty$, the Chebyshev bound is not useful
- We need a bound: $P[|\frac{S_n}{n} - \mu|] \leq c_n$ such that $\sum_n c_n < \infty$.

- Let us assume X_i have finite fourth moment

$$\left(\sum_{i=1}^{n}(X_i - \mu)\right)^4 = \sum_{i=1}^{n}(X_i - \mu)^4 + \sum_{i}\sum_{j>i} \frac{4!}{2!2!} (X_i - \mu)^2 (X_j - \mu)^2 + T$$

Where T represent a number of terms such that every term in it contains a factor like $(X_i - \mu)$
Note that $E[(X_i - \mu)(X_j - \mu)^3] = 0$ etc. because X_i are independent.

- Hence we get

$$E\left[\left(\sum_{i=1}^{n}(X_i - \mu)\right)^4\right] = nE[(X_i - \mu)^4] + 3n(n-1)\sigma^4 \leq C'n^2$$

- Now we can get, using Markov inequality

$$\begin{aligned} P\left[\left|\frac{S_n}{n} - \mu\right| > \epsilon\right] &= P\left[|S_n - n\mu| > n\epsilon\right] \\ &= P\left[\left|\sum_{i=1}^{n}(X_i - \mu)\right| > n\epsilon\right] \\ &\leq \frac{E\left(\sum_{i=1}^{n}(X_i - \mu)\right)^4}{(n\epsilon)^4} \\ &\leq \frac{C'n^2}{n^4\epsilon^4} = \frac{C}{n^2} \end{aligned}$$

- Since $\sum_n \frac{C}{n^2} < \infty$, we get $\frac{S_n}{n} \xrightarrow{a.s.} \mu$

- Strong law of large numbers says

$$\frac{S_n}{n} \xrightarrow{a.s.} \mu \quad \text{where } S_n = \sum_{i=1}^{n} X_i, \;\; X_i \text{ iid}, \;\; EX_i = \mu$$

- We proved it assuming finite fourth moment of X_i.
- This is only for illustration
- Strong law holds without any such assumptions on moments
- Strong law of large numbers says that sample mean converges to the expectation with probability one.

Convergence in r^{th} mean

- We say that a sequence X_n converges in r^{th} mean to X if $E[|X_n|^r] < \infty$, $\forall n$, $E[|X|^r] < \infty$ and

$$E\left[|X_n - X|^r\right] \to 0 \quad \text{as } \; n \to \infty$$

- Denoted as $X_n \xrightarrow{r} X$
- Consider our old example of binary random variables

$$P[X_n = 1] = \frac{1}{n} \qquad P[X_n = 0] = 1 - \frac{1}{n}$$

- All moments of X_n are finite and we have

$$E\left[|X_n - 0|^2\right] = \frac{1}{n} \to 0$$

- Hence $X_n \xrightarrow{2} 0$.
- In this example X_n converges in r^{th} mean for all r

- Suppose $X_n \xrightarrow{r} X$. Then, by Markov inequality

$$P\left[|X_n - X| > \epsilon\right] \leq \frac{E\left[|X_n - X|^r\right]}{\epsilon^r} \rightarrow 0$$

- Hence

$$X_n \xrightarrow{r} X \quad \Rightarrow \quad X_n \xrightarrow{P} X$$

- In general, neither of convergence almost surely and in r^{th} mean imply the other.
- We can generate counter examples for this easily.
- However, if all X_n take values in a bounded interval, then almost sure convergence implies r^{th} mean convergence

- Consider sequence X_n where X_n are independent with

$$P[X_n = 0] = 1 - a_n; \quad P[X_n = c_n] = a_n$$

- Assume $a_n \rightarrow 0$ so that $X_n \xrightarrow{P} 0$
- By Borel-Cantelli lemma

$$X_n \xrightarrow{a.s.} 0 \quad \Leftrightarrow \quad \sum_n a_n < \infty$$

- For convergence in r^{th} mean we need

$$E\left[|X_n - 0|^r\right] = (c_n)^r \ a_n \quad \rightarrow \ 0$$

- Take $a_n = \frac{1}{n}$ and $c_n = 1$. Then $X_n \xrightarrow{r} 0$ but the sequence does not converge almost surely.
- Take $a_n = \frac{1}{n^2}$ and $c_n = e^n$. Then $X_n \xrightarrow{a.s.} 0$ but the sequence does not converge in r^{th} mean for any r.

- Let $X_n \xrightarrow{r} X$. Then
  1. $E[|X_n|^r] \rightarrow E[|X|^r]$
  2. $X_n \xrightarrow{s} X, \ \forall s < r$
- The proofs are straight-forward but we omit the proofs

Convergence in distribution

- Let F_n be the df of X_n, $n = 1, 2, \cdots$. Let X be a rv with df F.
- Sequence X_n is said to converge to X **in distribution** if

$$F_n(x) \rightarrow F(x), \ \ \forall x \ \text{ where } F \text{ is continuous}$$

- We denote this as

$$X_n \xrightarrow{d} X, \quad \text{or} \quad X_n \xrightarrow{L} X, \quad \text{or } F_n \xrightarrow{w} F$$

- This is also known as **convergence in law** or weak convergence
- Note that here we are essentially talking about convergence of distribution functions.
- Convergence in probability implies convergence in distribution
- The converse is not true. (e.g., sequence of iid random variables)

Examples

- $X_1, X_2, \cdots$ be iid; uniform over $(0,\ 1)$
- $N_n = \min(X_1, \cdots, X_n)$, $Y_n = nN_n$. Does Y_n converge in distribution?

$$P[N_n > a] = (P[X_i > a])^n = (1-a)^n,\ \ 0 < a < 1$$

$$P[Y_n > y] = P[N_n > y/n] = \left(1 - \frac{y}{n}\right)^n, \quad \text{if } \ n > y$$

- Hence for any y

$$\lim_{n\to\infty} P[Y_n > y] = \lim_{n\to\infty} \left(1 - \frac{y}{n}\right)^n = e^{-y}$$

- The sequence converges in distribution to an exponential rv

Examples

- Let $\{X_n\}$ be *iid* with density $f(x) = e^{-x+\theta}, x > \theta > 0$.
- Let $N_n = \min(X_1, \cdots X_n)$. Does N_n converge in probability?
- Guess for limit: θ

$$P[|N_n - \theta| > \epsilon] = P[N_n > \theta + \epsilon] = (P[X_i > \theta + \epsilon])^n$$

$$P[X_i > \theta + \epsilon] = \int_{\theta+\epsilon}^{\infty} e^{-x+\theta}\, dx = e^{-\epsilon}$$

$$P[N_n > \theta + \epsilon] = \left(e^{-\epsilon}\right)^n \to 0, \text{ as } n \to \infty, \ \forall \epsilon > 0$$

- Hence $N_n \xrightarrow{P} \theta$
- Does it converge almost surely?

Examples

- $EX_n = m_n$ and $\text{Var}(X_n) = \sigma_n^2$, $n = 1, 2, \cdots$
- Want a sufficient condition for $X_n - m_n$ to converge in probability
- Note that $E[X_n - m_n] = 0$, and $\text{Var}(X_n - m_n) = \sigma_n^2$, $\forall n$

$$P[|X_n - m_n| > \epsilon] \le \frac{\sigma_n^2}{\epsilon^2}$$

- Hence, a sufficient condition is $\sigma_n^2 \to 0$.
- What is a sufficient condition for convergece almost surely?

- We have seen different modes of convergence
- $X_n \xrightarrow{d} X$ iff

$$F_n(x) \to F(x), \ \ \forall x \ \text{ where } F \text{ is continuous}$$

- $X_n \xrightarrow{P} X$ iff

$$\lim_{n\to\infty} P\left[|X_n - X| > \epsilon\right] = 0, \ \forall \epsilon > 0$$

- $X_n \xrightarrow{r} X$ iff

$$E\left[|X_n - X|^r\right] \to 0 \quad \text{ as } \ n \to \infty$$

- $X_n \xrightarrow{a.s} X$ iff

$$P[X_n \to X] = 1 \ \text{ or } \ P[\limsup |X_n - X| > \epsilon] = 0$$

- We have the following relations among different modes of convergence

$$X_n \xrightarrow{r} X \;\Rightarrow\; X_n \xrightarrow{P} X \;\Rightarrow\; X_n \xrightarrow{d} X$$

$$X_n \xrightarrow{a.s.} X \;\Rightarrow\; X_n \xrightarrow{P} X \;\Rightarrow\; X_n \xrightarrow{d} X$$

- All the implications are one-way and we have seen counter examples
- In general, almost sure convergence does not imply convergence in r^{th} mean and vice versa

- Strong and weak laws of large numbers are very useful examples of convergence of sequences of random variables.
- Given X_i are iid, $EX_i = \mu$, Var$(X_i) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$
  - Weak law of large numbers: $\frac{S_n}{n} \xrightarrow{P} \mu$
  - strong law of large numbers: $\frac{S_n}{n} \xrightarrow{a.s.} \mu$
- Another useful result is the Central Limit Theorem (CLT)
- CLT is about (normalized) sums of of independent random variables converging to the Gaussian distribution

Central Limit Theorem

- Given X_i are iid, $EX_i = \mu$, Var$(X_i) = \sigma^2$, $n = 1, 2, \cdots$

$$S_n = \sum_{i=1}^n X_i \;\Rightarrow\; ES_n = n\mu, \; \text{Var}(S_n) = n\sigma^2$$

- Given any rv Y, let $Z = \frac{Y - EY}{\sqrt{\text{Var}(Y)}}$
- Then, $EZ = 0$ and Var$(Z) = 1$.
- Define $\tilde{S}_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$ $E\tilde{S}_n = 0$, Var$(\tilde{S}_n) = 1$, $\forall n$
- Central Limit Theorem states: $\tilde{S}_n \xrightarrow{d} \mathcal{N}(0,1)$

$$\lim_{n\to\infty} P[\tilde{S}_n \leq a] = \Phi(a) \triangleq \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}\, dx$$

- Take X_i iid, $EX_i = 0$, Var$(X_i) = 1$, $n = 1, 2, \cdots$
- $S_n = \sum_{i=1}^n X_i$
- Strong law of large numbers implies

$$\frac{S_n}{n} \xrightarrow{a.s.} 0$$

- Central Limit Theorem implies

$$\frac{S_n}{\sqrt{n}} \xrightarrow{a.s.} \mathcal{N}(0,1)$$

Central Limit Theorem

- Given X_i are iid, $EX_i = \mu$, Var$(X_i) = \sigma^2$, $n = 1, 2, \cdots$

$$S_n = \sum_{i=1}^{n} X_i \quad \tilde{S}_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

- Central Limit Theorem states: $\tilde{S}_n \xrightarrow{d} \mathcal{N}(0,1)$
- We use characteristic functions for proving CLT

Characteristic Function

- Given rv X, its characteristic function, ϕ_X, is defined by

$$\phi_X(u) = E\left[e^{iuX}\right] = \int e^{iux}\, dF_X(x) \quad (i = \sqrt{-1})$$

- Since $|e^{iux}| \leq 1$, ϕ_X exists for all random variables

Properties of characteristic function

$$\phi_X(u) = E\left[e^{iuX}\right] = \int e^{iux}\, dF_X(x) \quad (i = \sqrt{-1})$$

- ϕ is continuous; $|\phi(u)| \leq \phi(0) = 1$; $\phi(-u) = \phi^*(u)$
- If $Y = aX + b$, $\phi_Y(u) = e^{iub}\phi_X(ua)$
- If $E|X|^r < \infty$, ϕ would be differentiable r times and

$$\phi^{(r)}(u) = E[(iX)^r e^{iuX}]$$

- Let $\mu_r = E[X^r]$ and let $\nu_r = E[|X|^r]$
- If ν_r is finite, then

$$\phi_X(u) = \sum_{s=0}^{r-1} \mu_s \frac{(iu)^s}{s!} + \rho(u)\, \mu_r \frac{(iu)^r}{r!}$$

where $|\rho(u)| \leq 1$ and $\rho(u) \to 1$ as $u \to 0$
- If all moments exist, then

$$\phi_X(u) = \sum_{s=0}^{\infty} \mu_s \frac{(iu)^s}{s!}$$

- We denote by ϕ_F characteristic function of df F
- Let F_n be a sequence of distribution functions
- **Continuity theorem**
  - If $F_n \to F$ then $\phi_{F_n} \to \phi_F$
  - If $\phi_{F_n} \to \psi$ and ψ is continuous at zero, then ψ would be characteristic function of some df, say, F, and $F_n \to F$

Characteristic function example

- Let X be binomial rv

$$\begin{aligned}\phi_X(u) = E\left[e^{iuX}\right] &= \sum_{k=0}^{n} {}^nC_k\, p^k\, (1-p)^{n-k}\, e^{iuk} \\ &= \sum_{k=0}^{n} {}^nC_k\, (pe^{iu})^k\, (1-p)^{n-k} \\ &= \left(pe^{iu} + (1-p)\right)^n\end{aligned}$$

- Let $X \sim \mathcal{N}(0,1)$

$$\begin{aligned}\phi_X(u) = E\left[e^{iuX}\right] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{iux}\, e^{-\frac{x^2}{2}}\, dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}((x-iu)^2 - i^2u^2)}\, dx \\ &= e^{-\frac{u^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}((x-iu)^2}\, dx \\ &= e^{-\frac{u^2}{2}}\end{aligned}$$

Recap: Modes of convergence

- $X_n \xrightarrow{d} X$ iff

$$F_n(x) \to F(x), \;\; \forall x \;\; \text{where } F \text{ is continuous}$$

- $X_n \xrightarrow{P} X$ iff

$$\lim_{n\to\infty} P\left[|X_n - X| > \epsilon\right] = 0, \; \forall \epsilon > 0$$

- $X_n \xrightarrow{r} X$ iff

$$E\left[|X_n - X|^r\right] \to 0 \quad \text{as} \;\; n \to \infty$$

- $X_n \xrightarrow{a.s} X$ iff

$$P[X_n \to X] = 1 \;\; \text{or} \;\; P[\limsup |X_n - X| > \epsilon] = 0$$

Recap

- We have the following relations among different modes of convergence

$$X_n \xrightarrow{r} X \;\Rightarrow\; X_n \xrightarrow{P} X \;\Rightarrow\; X_n \xrightarrow{d} X$$

$$X_n \xrightarrow{a.s.} X \;\Rightarrow\; X_n \xrightarrow{P} X \;\Rightarrow\; X_n \xrightarrow{d} X$$

- All the implications are one-way and we have seen counter examples
- In general, almost sure convergence does not imply convergence in r^{th} mean and vice versa

Recap

- Given X_i are iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$, $S_n = \sum_{i=1}^{n} X_i$
- Weak law of large numbers: $\frac{S_n}{n} \xrightarrow{P} \mu$
- strong law of large numbers: $\frac{S_n}{n} \xrightarrow{a.s.} \mu$
- Central Limit Theorem: $\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$

Recap

- Take X_i iid, $EX_i = 0$, $\text{Var}(X_i) = 1$, $n = 1, 2, \cdots$
- $S_n = \sum_{i=1}^{n} X_i$
- Strong law of large numbers implies

$$\frac{S_n}{n} \xrightarrow{a.s.} 0$$

- Central Limit Theorem implies

$$\frac{S_n}{\sqrt{n}} \xrightarrow{a.s.} \mathcal{N}(0,1)$$

Recap: Characteristic Function

- Given rv X, its characteristic function, ϕ_X, is defined by

$$\phi_X(u) = E\left[e^{iuX}\right] = \int e^{iux}\, dF_X(x) \quad (i = \sqrt{-1})$$

- Since $|e^{iux}| \leq 1$, ϕ_X exists for all random variables
  - ϕ is continuous; $|\phi(u)| \leq \phi(0) = 1$; $\phi(-u) = \phi^*(u)$
  - If $Y = aX + b$, $\phi_Y(u) = e^{iub}\phi_X(ua)$
  - If $E|X|^r < \infty$, ϕ would be differentiable r times and

$$\phi^{(r)}(u) = E[(iX)^r e^{iuX}]$$

Recap

- Let $\mu_r = E[X^r]$ and let $\nu_r = E[|X|^r]$
- If ν_r is finite, then

$$\phi_X(u) = \sum_{s=0}^{r-1} \mu_s \frac{(iu)^s}{s!} + \rho(u)\ \mu_r \frac{(iu)^r}{r!}$$

  where $|\rho(u)| \leq 1$ and $\rho(u) \to 1$ as $u \to 0$
- If all moments exist, then

$$\phi_X(u) = \sum_{s=0}^{\infty} \mu_s \frac{(iu)^s}{s!}$$

Recap

- We denote by ϕ_F characteristic function of df F
- Let F_n be a sequence of distribution functions
- **Continuity theorem**
  - If $F_n \to F$ then $\phi_{F_n} \to \phi_F$
  - If $\phi_{F_n} \to \psi$ and ψ is continuous at zero, then ψ would be characteristic function of some df, say, F, and $F_n \to F$

- Given X_i iid, $EX_i = \mu$, Var$(X_i) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$
- Let $\tilde{S}_n = \frac{S_n - ES_n}{\sqrt{\text{var}(S_n)}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$
- **(Lindberg-Levy) Central Limit Theorem**

$$\lim_{n\to\infty} P\left[\tilde{S}_n \leq x\right] = \lim_{n\to\infty} P\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\ dt, \quad \forall x$$

Proof:

- Without loss of generality let us assume $\mu = 0$.
- We use characteristic function of $\tilde{S}_n$ for the proof.
- Let ϕ be the characteristic function of X_i. Then

$$\phi_{S_n}(t) = (\phi(t))^n \quad \text{and} \quad \phi_{\tilde{S}_n}(t) = \left(\phi\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n$$

- Recall that we can expand ϕ in a Taylor series

$$\phi(u) = \sum_{s=0}^{r-1} \mu_s \frac{(iu)^s}{s!} + \rho(u)\ \mu_r \frac{(iu)^r}{r!}, \quad \rho(u) \to 1, \text{ as } u \to 0$$

- Here we assume: $EX_i = 0$ and $EX_i^2 = \sigma^2$

$$\phi(t) = 1\ +\ 0\ -\ \frac{1}{2}\,\rho(t)\,\sigma^2\,t^2$$

$$\begin{aligned}
\phi\left(\frac{t}{\sigma\sqrt{n}}\right) &= 1 - \frac{1}{2}\,\rho\left(\frac{t}{\sigma\sqrt{n}}\right)\,\sigma^2\,\frac{t^2}{\sigma^2 n} \\
&= 1 - \frac{1}{2}\frac{t^2}{n} + \frac{1}{2}\frac{t^2}{n}\left(1 - \rho\left(\frac{t}{\sigma\sqrt{n}}\right)\right) \\
&= 1 - \frac{1}{2}\frac{t^2}{n} + o\left(\frac{1}{n}\right)
\end{aligned}$$

- Hence we get

$$\begin{aligned}\lim_{n\to\infty} \phi_{\tilde{S}_n}(t) &= \lim_{n\to\infty} \left(\phi\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n \\ &= \lim_{n\to\infty} \left(1 - \frac{1}{2}\frac{t^2}{n} + o\left(\frac{1}{n}\right)\right)^n \\ &= e^{-\frac{t^2}{2}}\end{aligned}$$

which is the characteristic function of standard normal
- By Continuity theorem, distribution function of $\tilde{S}_n$ converges to that of standard Normal rv

$$\lim_{n\to\infty} P\left[\tilde{S}_n \le x\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt, \quad \forall x$$

- What CLT says is that sums of iid random variables, when appropriately normalized, would always approach the Gaussian distribution.
- It allows one to approximate distribution of sums of independent rv's
- Let X_i be iid and $S_n = \sum_{i=1}^n X_i$

$$P[S_n \le x] = P\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \le \frac{x - n\mu}{\sigma\sqrt{n}}\right] \approx \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right)$$

- Thus, S_n is well approximated by a normal rv with mean $n\mu$ and variance $n\sigma^2$, if n is large

Example

- Twenty numbers are rounded off to the nearest integer and added. What is the probability that the sum obtained differs from true sum by more than 3.
- A reasonable assumption is round-off errors are independent and uniform over $[-0.5,\ 0.5]$
- Take $Z = \sum_{i=1}^{20} X_i$, $X_i \sim U[-0.5,\ 0.5]$, X_i iid.
- Then Z represents the error in the sum.

- $Z = \sum_{i=1}^{20} X_i$, $X_i \sim U[-0.5,\ 0.5]$, X_i iid
- $EX_i = 0$ and $\text{Var}(X_i) = \frac{1}{12}$.
- Hence, $EZ = 0$ and $\text{Var}(Z) = \frac{20}{12} = \frac{5}{3}$

$$\begin{aligned}P[|Z| \le 3] &= P[-3 \le Z \le 3] \\ &= P\left[\frac{-3}{\sqrt{\frac{5}{3}}} \le \frac{Z - EZ}{\sqrt{\text{Var}(Z)}} \le \frac{3}{\sqrt{\frac{5}{3}}}\right] \\ &\approx \Phi\left(\frac{3}{\sqrt{\frac{5}{3}}}\right) - \Phi\left(\frac{-3}{\sqrt{\frac{5}{3}}}\right) \\ &\approx \Phi(2.3) - \Phi(-2.3) \\ &= 0.9893 - 0.0107 \approx 0.98\end{aligned}$$

- Hence probability that the sum differs from true sum by more than 3 is 0.02

- We can approximate binomial rv with Gaussian for large n
- Binomial random variable with parameters n, p is a sum of n independent Bernoulli variables: $S_n = \sum_{i=1}^n X_i; \;\; X_i \in \{0,\ 1\},\ P[X_i = 1] = p,\ X_i$ ind
- Hence we can approximate distribution of S_n by

$$\begin{aligned} P[S_n \le x] &= P\left[\frac{S_n - np}{\sqrt{np(1-p)}} \le \frac{x - np}{\sqrt{np(1-p)}}\right] \\ &\approx \Phi\left(\frac{x-np}{\sqrt{np(1-p)}}\right) \end{aligned}$$

- For large n, binomial rv is like a Gaussian rv with mean np and variance $np(1-p)$
- The approximation is quite good in practice

- S_n be binomial with parameters n, p

$$P[S_n \le x] \approx \Phi\left(\frac{x-np}{\sqrt{np(1-p)}}\right)$$

- For example, with $p = 0.95$

$$P[S_{110} \le 100] \approx \Phi\left(\frac{100 - 110 * 0.95}{\sqrt{110 * 0.05 * 0.95}}\right) \approx \Phi(-1.97) = 0.025$$

- Since S_n is integer-valued, the LHS above is same for all x between two consecutive integers; but RHS changes
- To get a good approximation, to calculate $P[S_n \le m]$ one uses $P[S_n \le m + 0.5]$ in the above approximation formula

- CLT allows one to get rate of convergence of law of large numbers
- Let X_i iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$
- By Law of large numbers, $\frac{S_n}{n} \to \mu$.
- Now, by CLT

$$\begin{aligned} P\left[\left|\frac{S_n}{n} - \mu\right| > \epsilon\right] &= P\left[|S_n - n\mu| > n\epsilon\right] \\ &= P\left[\left|\frac{S_n - n\mu}{\sigma\sqrt{n}}\right| > \frac{n\epsilon}{\sigma\sqrt{n}}\right] \\ &\approx 1 - \left(\Phi\left(\frac{n\epsilon}{\sigma\sqrt{n}}\right) - \Phi\left(-\frac{n\epsilon}{\sigma\sqrt{n}}\right)\right) \\ &= 2\left(1 - \Phi\left(\frac{n\epsilon}{\sigma\sqrt{n}}\right)\right) \end{aligned}$$

(because $\Phi(-x) = (1 - \Phi(x))$)

Example: Opinion Polls

- let p denote the fraction of population that prefers product A to product B
- We want to estimate p
- We conduct a sample survey by asking n people
- We want to make a statement such as

  $p = 0.34 \pm 0.07$ *with a confidence of* 95%
- Here, the 0.34 would be the sample mean. The other two numbers can be fixed using CLT

- $X_i \in \{0,\ 1\}$ iid, $EX_i = p$, $S_n = \sum_{i=1}^{n} X_i$
- Now, by CLT, we have

$$\begin{aligned} P\left[\left|\frac{S_n}{n} - p\right| > \epsilon\right] &= P\left[|S_n - np| > n\epsilon\right] \\ &= 2\left(1 - \Phi\left(\frac{n\epsilon}{\sqrt{np(1-p)}}\right)\right) \end{aligned}$$

- Suppose we want to satisfy

$$P\left[\left|\frac{S_n}{n} - p\right| > \epsilon\right] = \delta$$

- We can calculate any one of ϵ, δ or n given the other two using the earlier equation.
- But we need value of p for it!

- Fortunately, $\sqrt{p(1-p)}$ does not change too much with p
- It attains its maximum value of 0.5 at $p = 0.5$
- It is 0.458 at $p = 0.3$ and is 0.4 at $p = 0.2$
- One normally fixes this variance as 0.5 or 0.45 to calculate the sample size, n.
- There are other ways of handling it

- We have

$$P\left[\left|\frac{S_n}{n} - p\right| > \epsilon\right] = 2\left(1 - \Phi\left(\frac{\epsilon\sqrt{n}}{\sqrt{p(1-p)}}\right)\right)$$

- Suppose $n = 900$ and $\epsilon = 0.025$.
  Let us approximate $\sqrt{p(1-p)} = 0.45$. Then

$$2\left(1 - \Phi\left(\frac{0.025 * 30}{0.45}\right)\right) = 2(1 - \Phi(1.66)) \approx 0.1$$

- If we took $\sqrt{p(1-p)} = 0.5$ we get the value as 0.14
- If we use Chebyshev inequality with variance as 0.5 we get the bound as 0.8
- If we change ϵ to 0.05, then at variance equal to 0.5 the probability becomes about 0.02 while the Chebyshev bound would be about 0.2

Confidence intervals

- Let X_i iid, $EX_i = \mu$, Var$(X_i) = \sigma^2$, $S_n = \sum_{i=1}^{n} X_i$.
- Using CLT, we get

$$P\left[\left|\frac{S_n}{n} - \mu\right| > c\right] = 2\left(1 - \Phi\left(\frac{c\sqrt{n}}{\sigma}\right)\right)$$

- If the RHS above is δ, then we can say that $\frac{S_n}{n} \in [\mu - c,\ \mu + c]$ with probability $(1 - \delta)$
- This interval is called the $100(1-\delta)\%$ confidence interval.

$$P\left[\left|\frac{S_n}{n} - \mu\right| > c\right] = 2\left(1 - \Phi\left(\frac{c\sqrt{n}}{\sigma}\right)\right)$$

- Suppose $c = \frac{1.96\sigma}{\sqrt{n}}$
- Then

$$P\left[\left|\frac{S_n}{n} - \mu\right| > \frac{1.96\sigma}{\sqrt{n}}\right] = 2\left(1 - \Phi\left(1.96\right)\right) = 0.05$$

- Denoting $\bar{X} = \frac{S_n}{n}$, the 95% confidence interval is $\left[\bar{X} - \frac{1.96\sigma}{\sqrt{n}},\ \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right]$
- One generally uses an estimate for σ obtained from X_i
- In analyzing any experimental data the confidence intervals or the variance term is important

central limit theorem

- CLT essentially states that sum of many independent random variables behaves like a Gaussian random variable
- It is very useful in many statistics applications.
- We stated CLT for iid random variables.
- While independence is important, all rv need not have the same distribution.
- Essentially, the variances should not die out.

- We have been considering sequences: $X_n,\ n = 1, 2, \cdots$
- We have so far considered only the asymptotic properties or limits of such sequences.
- Any such sequence is an example of what is called a random process or stochastic process
- Given n rv, they are completely characterized by their joint distribution.
- How doe we specify or characterize an infinite collection of random variables?
- We need the joint distribution of every finite subcollection of them.

Markov Chains

- Let $X_n,\ n = 0, 1, \cdots$ be a sequence of discrete random variables taking values in S.
  Note that S would be countable
- We say it is a Markov chain if

$$P[X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1} \cdots X_0 = x_0] = P[X_{n+1} = x_{n+1} | X_n = x_n], \forall$$

- We can write it as

$$P[X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} \cdots X_0] = P[X_{n+1} = x_{n+1} | X_n = x_n], \forall x_i$$

- Conditioned on X_n, X_{n+1} is independent of $X_{n-1}, X_{n-2}, \cdots$
- We think of X_n as state at n
- For a Markov chain, given the current state, the future evolution is independent of the history of how you reached the current state

Example

- Let X_i be iid discrete rv taking integer values.
- Let $Y_0 = 0$ and $Y_n = \sum_{i=1}^{n} X_i$
- $Y_n, \; n = 0, 1, \cdots$ is a Markov chain with state space as integers
- Note that $Y_{n+1} = Y_n + X_{n+1}$ and X_{n+1} is independent of $Y_0, \cdots, Y_n$.

$$P[Y_{n+1} = y | Y_n = x, Y_{n-1}, \cdots] = P[X_{n+1} = y - x]$$

- Thus, Y_{n+1} is conditionally independent of $Y_{n-1}, \cdots$ conditioned on Y_n

- In this example, we can think of X_n as the number of people or things arriving at a facility in the n^{th} time interval.
- Then Y_n would be total arrivals till end of n^{th} time interval.
- Number of packets coming into a network switch, number people joining the queue in a bank, number of infections till date are all Markov chains.
- This is a useful model for many dynamic systems or processes

- The Markov property is: given current state, the future evolution is independent of the history of how we came to current state.
- It essentially means the current state contains all needed information about history
- We are considering the case where states as well as time are discrete.
- It can be more general and we discuss some of them

Transition Probabilities

- Let $\{X_n, \; n = 0, 1, \cdots\}$ be a Markov Chain with (countable) state space S

$$Pr[X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} \cdots X_0] = Pr[X_{n+1} = x_{n+1} | X_n = x_n], \forall x$$

(Notice change of notation)

- Define function $P : S \times S \to [0, \; 1]$ by

$$P(x, y) = Pr[X_{n+1} = y | X_n = x]$$

- P is called the state transition probability function. It satisfies
  - $P(x, y) \geq 0, \; \forall x, y \in S$
  - $\sum_{y \in S} P(x, y) = 1, \forall x \in S$
- If S is finite then P can be represented as a matrix

- The state transition probability function is given by

$$P(x,y) = Pr[X_{n+1} = y | X_n = x]$$

- In general, this can depend on n though our notation does not show it
- If the value of that probability does not depend on n then the chain is called homogeneous
- For a homogeneous chain we have

$$Pr[X_{n+1} = y | X_n = x] = Pr[X_1 = y | X_0 = x], \ \forall n$$

- In this course we will consider only homogeneous chains

Initial State Probabilities

- Let $\{X_n\}$ be a Markov Chain with state space S
- Define function $\pi_0 : S \to [0, \ 1]$ by

$$\pi_0(x) = Pr[X_0 = x]$$

- It is the pmf of the rv X_0
- Hence it satisfies
  - $\pi_0(x) \geq 0, \ \forall x \in S$
  - $\sum_{x \in S} \pi_0(x) = 1$
- From now on, without loss of generality, we take $S = \{0, 1, \cdots\}$

- Let X_n be a (homogeneous) Markov chain
- Then we have

$$\begin{aligned} Pr[X_0 = x_0, X_1 = x_1] &= Pr[X_1 = x_1 | X_0 = x_0] \ Pr[X_0 = x_0], \forall x_0, x_1 \\ &= P(x_0, x_1)\pi_0(x_0) = \pi_0(x_0)P(x_0, x_1) \end{aligned}$$

- Now we can extend this as

$$\begin{aligned} Pr[X_0 = x_0, X_1 = x_1, X_2 = x_2] &= Pr[X_2 = x_2 | X_1 = x_1, X_0 = x_0] \cdot \\ &\quad Pr[X_0 = x_0, X_1 = x_1] \\ &= Pr[X_2 = x_2 | X_1 = x_1] \cdot \\ &\quad Pr[X_0 = x_0, X_1 = x_1] \\ &= P(x_1, x_2) \ P(x_0, x_1)\pi_0(x_0) \\ &= \pi_0(x_0)P(x_0, x_1)P(x_1, x_2) \end{aligned}$$

- This calculation is easily generalized to any number of time steps

$$\begin{aligned} Pr[X_0 = x_0, \cdots X_n = x_n] &= Pr[X_n = x_n | X_{n-1} = x_{n-1}, \cdots X_0 = x_0] \cdot \\ &\quad Pr[X_{n-1} = x_{n-1}, \cdots X_0 = x_0] \\ &= Pr[X_n = x_n | X_{n-1} = x_{n-1}] \cdot \\ &\quad Pr[X_{n-1} = x_{n-1}, \cdots X_0 = x_0] \\ &= P(x_{n-1}, x_n) Pr[X_{n-1} = x_{n-1}, \cdots X_0 = x_0] \\ &= P(x_{n-1}, x_n) Pr[X_{n-1} = x_{n-1} | X_{n-2} = x_{n-2}] \cdot \\ &\quad Pr[X_{n-2} = x_{n-2}, \cdots X_0 = x_0] \\ &\ \vdots \\ &= \pi_0(x_0)P(x_0, x_1) \cdots P(x_{n-1}, x_n) \end{aligned}$$

- We showed

$$Pr[X_0 = x_0, \cdots X_n = x_n] = \pi_0(x_0)P(x_0, x_1) \cdots P(x_{n-1}, x_n)$$

- This shows that the transition probabilities, P, and initial state probabilities, π_0, completely specify the chain.
- They give us the joint distribution of any finite subcollection of the rv's
- Suppose you want joint distribution of $X_{i_1}, \cdots X_{i_k}$
- Let $m = \max(i_1, \cdots, i_k)$
- We know how to get joint distribution of $X_0, \cdots, X_m$.
- The joint distribution of $X_{i_1}, \cdots X_{i_k}$ is now calculated as a marginal distribution from the joint distribution of $X_0, \cdots, X_m$

Recap: Central Limit Theorem

- Given X_i iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$
- Let $\tilde{S}_n = \frac{S_n - ES_n}{\sqrt{\text{Var}(S_n)}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$
- **(Lindberg-Levy) Central Limit Theorem**

$$\lim_{n\to\infty} P\left[\tilde{S}_n \leq x\right] = \lim_{n\to\infty} P\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt, \;\; \forall x$$

- It allows us to approximate distributions of sums of independent random variables

$$P[S_n \leq x] \approx \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right)$$

- For example, binomial rv is well approximated by normal for large n
- CLT is also important to get information on rate of convergence of law of large numbers.

Recap: Markov Chain

- Let $X_n, \; n = 0, 1, \cdots$ be a sequence of discrete random variables taking values in S.
- We say it is a Markov chain if

$$Pr[X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1} \cdots X_0 = x_0] = Pr[X_{n+1} = x_{n+1} | X_n = x_n],$$

- We can write it as

$$f_{X_{n+1}|X_n,\cdots X_0}(x_{n+1}|x_n, \cdots, x_0) = f_{X_{n+1}|X_n}(x_{n+1}|x_n), \;\forall x_i$$

- For a Markov chain, given the current state, the future evolution is independent of the history of how you reached the current state

Recap: Transition Probabilities

- Let $\{X_n, \; n = 0, 1, \cdots\}$ be a Markov Chain with (countable) state space S
- Transition probability function is $P : S \times S \to [0, \; 1]$

$$P(x, y) = Pr[X_{n+1} = y | X_n = x]$$

  The chain is said to be homogeneous when this is not a function of time.
- It satisfies
  - $P(x, y) \geq 0, \; \forall x, y \in S$
  - $\sum_{y \in S} P(x, y) = 1, \forall x \in S$
- If S is finite then P can be represented as a matrix

Recap: Initial State Probabilities

- Let $\{X_n\}$ be a Markov Chain with state space S
- Initial state probabilities $\pi_0 : S \to [0,\ 1]$

$$\pi_0(x) = Pr[X_0 = x]$$

It satisfies
  - $\pi_0(x) \geq 0, \ \forall x \in S$
  - $\sum_{x \in S} \pi_0(x) = 1$

- The P and π_0 determine all joint distributions

$$\begin{aligned}
Pr[X_0 = x_0, \cdots X_n = x_n] &= Pr[X_n = x_n | X_{n-1} = x_{n-1}, \cdots X_0 = x_0] \cdot \\
&\qquad Pr[X_{n-1} = x_{n-1}, \cdots X_0 = x_0] \\
&= Pr[X_n = x_n | X_{n-1} = x_{n-1}] \cdot \\
&\qquad Pr[X_{n-1} = x_{n-1}, \cdots X_0 = x_0] \\
&= P(x_{n-1}, x_n) Pr[X_{n-1} = x_{n-1}, \cdots X_0 = x_0] \\
&= P(x_{n-1}, x_n) Pr[X_{n-1} = x_{n-1} | X_{n-2} = x_{n-2}] \cdot \\
&\qquad Pr[X_{n-2} = x_{n-2}, \cdots X_0 = x_0] \\
&\ \vdots \\
&= \pi_0(x_0) P(x_0, x_1) \cdots P(x_{n-1}, x_n)
\end{aligned}$$

- We showed

$$Pr[X_0 = x_0, \cdots X_n = x_n] = \pi_0(x_0) P(x_0, x_1) \cdots P(x_{n-1}, x_n)$$

- This shows P, and π_0, determine joint distribution of $X_0, \cdots, X_m$ for any m
- Suppose you want joint distribution of $X_{i_1}, \cdots X_{i_k}$
- Let $m = \max(i_1, \cdots, i_k)$
- We know how to get joint distribution of $X_0, \cdots, X_m$.
- The joint distribution of $X_{i_1}, \cdots X_{i_k}$ is now calculated as a marginal distribution from the joint distribution of $X_0, \cdots, X_m$
- This shows that the transition probabilities, P, and initial state probabilities, π_0, completely specify the chain.

Example: 2-state chain

- Let $S = \{0, 1\}$.
- We can write the transition probabilities as a matrix

$$P = \begin{bmatrix} P(0,0) & P(0,1) \\ P(1,0) & P(1,1) \end{bmatrix} = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

- Now we can calculate the joint distribution, e.g., of X_1, X_2 as

$$\begin{aligned}
Pr[X_1 = 0, X_2 = 1] &= \sum_{x=0}^{1} Pr[X_0 = x, X_1 = 0, X_2 = 1] \\
&= \sum_{x=0}^{1} \pi_0(x) P(x, 0) P(0, 1) \\
&= \pi_0(0)(1-p)p + \pi_0(1)qp
\end{aligned}$$

- We can similarly calculate probabilities of any events involving these random variables

$$
\begin{aligned}
Pr[X_2 \neq X_0] &= Pr[X_2 = 0, X_0 = 1] + Pr[X_2 = 1, X_0 = 0] \\
&= \sum_{x=0}^{1} (\pi_0(1)P(1,x)P(x,0) + \pi_0(0)P(0,x)P(x,1))
\end{aligned}
$$

- We have the formula

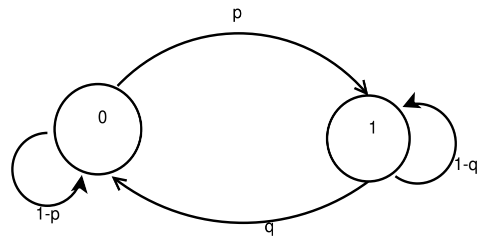
$$Pr[X_0 = x_0, \cdots X_n = x_n] = \pi_0(x_0)P(x_0, x_1)\cdots P(x_{n-1}, x_n)$$

- This can easily be seen through a graphical notation.

- Consider the 2-state chain with $S = \{0, 1\}$ and

$$P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

- We can represent the chain through a graph as shown below



- The nodes represent states. The edges show possible transitions and the probabilities

$$Pr[X_0 = 0, X_1 = 1, X_2 = 1, X_3 = 0] = \pi_0(0)p(1-q)q$$

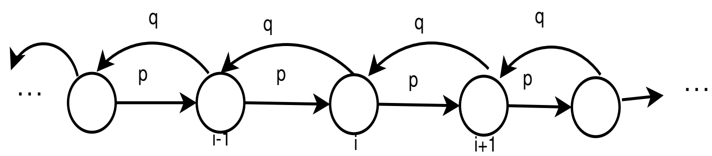
An example

- A man has 4 umbrellas. carries them from home to office and back when needed. Probability of rain in the morning and evening is same, namely, p.
- What should be the state?
- $S = \{0, 1, \cdots, 5\}$. The transition probabilities are

$$
P = \begin{array}{c|ccccc}
 & 0 & 1 & 2 & 3 & 4 \\
\hline
0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 1-p & p \\
2 & 0 & 0 & 1-p & p & 0 \\
3 & 0 & 1-p & p & 0 & 0 \\
4 & 1-p & p & 0 & 0 & 0
\end{array}
$$

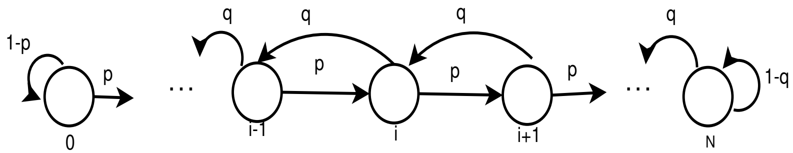
Birth-Death chain

- The following Markov chain is known as a birth-death chain



- In general, birth-death chains may have self-loops on states
- Random walk: $X_i \in \{-1, +1\}$, iid, $S_n = \sum_{i=1}^{n} X_i$
- We can have 'reflecting boundary' at 0
- Queuing chains can also be birth-death chains

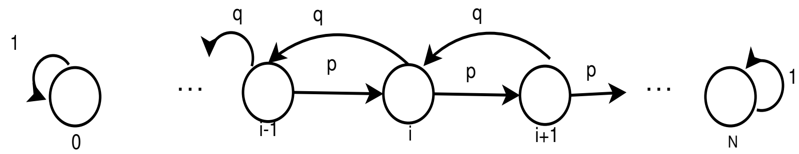
- We can have birth-death chains with finite state space also



- This chain keeps visiting all the states again and again

Gambler's Ruin chain

- The following chain is called Gambler's ruin chain



- Here, the chain is ultimately absorbed either in 0 or in N
- Here state can be the current funds that the gambler has

- The transition probabilities we defined earlier are also called one step transition probabilities

$$P(x,y) = Pr[X_{n+1} = y | X_n = x] = Pr[X_1 = y | X_0 = x]$$

- We can define transition probabilities for multiple steps, that is, $Pr[X_n = y | X_0 = x]$
- We first look at one consequence of markov property
- The Markov property implies that it is the most recent past that matters. For example

$$Pr[X_{n+m} = y | X_n = x, X_0] = Pr[X_{n+m} = y | X_n = x]$$

- We consider a simple case

$$\begin{aligned} Pr[X_3 = y | X_1 = x, X_0 = z] &= \frac{Pr[X_3 = y, X_1 = x, X_0 = z]}{Pr[X_1 = x, X_0 = z]} \\ &= \frac{\sum_w \pi_0(z) P(z,x) P(x,w) P(w,y)}{\pi_0(z) P(z,x)} \\ &= \sum_w P(x,w) P(w,y) \end{aligned}$$

- We also have

$$\begin{aligned} Pr[X_3 = y | X_1 = x] &= Pr[X_2 = y | X_0 = x] \\ &= \frac{\sum_w \pi_0(x) P(x,w) P(w,y)}{\pi_0(x)} \\ &= \sum_w P(x,w) P(w,y) \end{aligned}$$

- Thus we get

$$Pr[X_3 = y | X_1 = x, X_0 = z] = Pr[X_3 = y | X_1 = x]$$

- Using similar algebra, we can show that

$$\begin{aligned} Pr[X_{m+n} = y | X_m = x, X_0 = z] &= Pr[X_{m+n} = y | X_m = x] \\ &= Pr[X_n = y | X_0 = x] \end{aligned}$$

- Or, in general,

$$f_{X_{m+n}|X_m,\cdots,X_0}(y|x,\cdots) = f_{X_{m+n}|X_m}(y|x)$$

- Using the same algebra, we can show

$$\begin{aligned} \Pr[X_{m+n} = y | X_m = x,\ X_{m-k} \in A_k,\ k = 1,\cdots,m] &= \\ Pr[X_{m+n} = y | X_m = x] \end{aligned}$$

$$\begin{aligned} Pr[X_{m+n+r} \in B_r,\ r = 0,\cdots,s \mid X_m = x,\ X_{m-k} \in A_k,\ k = 1,\cdots,m] \\ = Pr[X_{m+n+r} \in B_r,\ r = 0,\cdots,s \mid X_m = x] \end{aligned}$$

- Now we get

$$\begin{aligned} Pr[X_{m+n} = y | X_0 = x] &= \sum_z Pr[X_{m+n} = y, X_m = z | X_0 = x] \\ &= \sum_z Pr[X_{m+n} = y | X_m = z, X_0 = x]\ Pr[X_m = z | X_0 = x] \\ &= \sum_z Pr[X_{m+n} = y | X_m = z]\ Pr[X_m = z | X_0 = x] \\ &= \sum_z Pr[X_n = y | X_0 = z]\ Pr[X_m = z | X_0 = x] \end{aligned}$$

Chapman-Kolmogorov Equations

- Define: $P^n(x,y) = Pr[X_n = y | x_0 = x]$
- These are called n-step transition probabilities.
- From what we showed, n-step transition probabilities satisfy

$$P^{m+n}(x,y) = \sum_z P^m(x,z) P^n(z,y)$$

- These are known as Chapman-Kolmogorov equations
- This relationship is intuitively clear

- Specifically, using Chapman-Kolmogorov equations,

$$P^2(x,y) = \sum_z P(x,z) P(z,y)$$

- For a finite chain, P is a matrix
- Thus $P^2(x,y)$ is the $(x,y)^{th}$ element of the matrix, $P \times P$
- That is why we use P^n for n-step transition probabilities

- Define: $\pi_n(x) = Pr[X_n = x]$.
- Then we get

$$\begin{aligned}\pi_n(y) &= \sum_x Pr[X_n = y | X_0 = x] \ Pr[X_0 = x] \\ &= \sum_x \pi_0(x) P^n(x, y)\end{aligned}$$

- In particular

$$\begin{aligned}\pi_{n+1}(y) &= \sum_x Pr[X_{n+1} = y | X_n = x] \ Pr[X_n = x] \\ &= \sum_x \pi_n(x) P(x, y)\end{aligned}$$

Hitting times

- Let y be a state.
- We define hitting time for y as the random variable

$$T_y = \min\{n > 0 \ : \ X_n = y\}$$

- T_y is the first time that the chain is in state y (after $t = 0$ when the chain is initiated).
- It is easy to see that $Pr[T_y = 1 | X_0 = x] = P(x, y)$.
- We often need conditional probability conditioned on the initial state.
- Notation: $P_z(A) = Pr[A | X_0 = z]$
- We write the above as $P_x(T_y = 1) = P(x, y)$

$$T_y = \min\{n > 0 \ : \ X_n = y\}$$

- We can now get

$$\begin{aligned}P_x(T_y = 2) &= \sum_{z \neq y} P(x, z) \ P(z, y) = \sum_{z \neq y} P(x, z) \ P_z(T_y = 1) \\ P_x(T_y = m) &= Pr[T_y = m | X_0 = x] \\ &= \sum_{z \neq y} Pr[T_y = m | X_1 = z, X_0 = x] \ Pr[X_1 = z | X_0 = x] \\ &= \sum_{z \neq y} P(x, z) Pr[T_y = m | X_1 = z] \\ &= \sum_{z \neq y} P(x, z) P_z(T_y = m - 1)\end{aligned}$$

- Similarly we can get:

$$P^n(x, y) = \sum_{m=1}^{n} P_x(T_y = m) P^{n-m}(y, y)$$

- We can derive this as

$$\begin{aligned}P^n(x, y) &= Pr[X_n = y | X_0 = x] \\ &= \sum_{m=1}^{n} Pr[T_y = m, X_n = y | X_0 = x] \\ &= \sum_{m=1}^{n} Pr[X_n = y | T_y = m, X_0 = x] \ Pr[T_y = m | X_0 = x] \\ &= \sum_{m=1}^{n} Pr[X_n = y | X_m = y] \ Pr[T_y = m | X_0 = x] \\ &= \sum_{m=1}^{n} P^{n-m}(y, y) \ P_x(T_y = m)\end{aligned}$$

transient and recurrent states

- Define $\rho_{xy} = P_x(T_y < \infty)$.
- It is the probability that starting in x you will visit y
- Note that

$$\rho_{xy} = \lim_{n\to\infty} P_x(T_y < n) = \sum_{n=1}^{\infty} P_x(T_y = n)$$

Definition: A state y is called transient if $\rho_{yy} < 1$; it is called recurrent if $\rho_{yy} = 1$.

- Intuitively, all transient states would be visited only finitely many times while recurrent states are visited infinitely often.
- For any state y define

$$I_y(X_n) = \begin{cases} 1 & \text{if } X_n = y \\ 0 & \text{otherwise} \end{cases}$$

- Now, the total number of visits to y is given by

$$N_y = \sum_{n=1}^{\infty} I_y(X_n)$$

- We can get distribution of N_y as

$$\begin{aligned} P_x(N_y \ge 1) &= P_x(T_y < \infty) = \rho_{xy} \\ P_x(N_y \ge 2) &= \sum_m P_x(T_y = m) P_y(T_y < \infty) \\ &= \rho_{yy} \sum_m P_x(T_y = m) = \rho_{yy}\, \rho_{xy} \\ P_x(N_y \ge m) &= \rho_{yy}^{m-1}\, \rho_{xy} \\ P_x(N_y = m) &= P_x(N_y \ge m) - P_x(N_y \ge m+1) \\ &= \rho_{yy}^{m-1}\, \rho_{xy} - \rho_{yy}^{m}\, \rho_{xy} = \rho_{xy}\, \rho_{yy}^{m-1}(1 - \rho_{yy}) \\ P_x(N_y = 0) &= 1 - P_x(N_y \ge 1) = 1 - \rho_{xy} \end{aligned}$$

- Notation: $E_x[Z] = E[Z | X_0 = x]$
- Define

$$\begin{aligned} G(x,y) &\triangleq E_x[N_y] \\ &= E_x\left[\sum_{n=1}^{\infty} I_y(X_n)\right] \\ &= \sum_{n=1}^{\infty} E_x\left[I_y(X_n)\right] \\ &= \sum_{n=1}^{\infty} P^n(x,y) \end{aligned}$$

- $G(x,y)$ is the expected number of visits to y for a chain that is started in x.

Theorem:

(i). Let y be transient. Then

$$P_x(N_y < \infty) = 1, \ \forall x \quad \text{and} \quad G(x,y) = \frac{\rho_{xy}}{1 - \rho_{yy}} < \infty, \ \forall x$$

(ii) Let y be recurrent. Then

$$P_y[N_y = \infty] = 1, \quad \text{and} \quad G(y,y) = E_y[N_y] = \infty$$

$$P_x[N_y = \infty] = \rho_{xy}, \quad \text{and} \quad G(x,y) = \begin{cases} 0 & \text{if } \rho_{xy} = 0 \\ \infty & \text{if } \rho_{xy} > 0 \end{cases}$$

Recap: Markov Chain

- Let $X_n,\ n = 0, 1, \cdots$ be a sequence of discrete random variables taking values in S.
- We say it is a Markov chain if

$$Pr[X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1} \cdots X_0 = x_0] = Pr[X_{n+1} = x_{n+1} | X_n = x_n],$$

- We can write it as

$$f_{X_{n+1}|X_n,\cdots X_0}(x_{n+1}|x_n, \cdots, x_0) = f_{X_{n+1}|X_n}(x_{n+1}|x_n),\ \forall x_i$$

- For a Markov chain, given the current state, the future evolution is independent of the history of how you reached the current state

Recap: Transition Probabilities

- Transition probability function is $P : S \times S \to [0,\ 1]$

$$P(x, y) = Pr[X_{n+1} = y | X_n = x]$$

  The chain is said to be homogeneous when this is not a function of time.
- For a homogeneous chain

$$Pr[X_{n+1} = y | X_n = x] = Pr[X_1 = y | X_0 = x],\ \forall n$$

- P satisfies
  - $P(x, y) \geq 0,\ \forall x, y \in S$
  - $\sum_{y \in S} P(x, y) = 1, \forall x \in S$
- If S is finite then P can be represented as a matrix

Recap: Initial State Probabilities

- Initial state probabilities $\pi_0 : S \to [0,\ 1]$

$$\pi_0(x) = Pr[X_0 = x]$$

  It satisfies
  - $\pi_0(x) \geq 0,\ \forall x \in S$
  - $\sum_{x \in S} \pi_0(x) = 1$
- The P and π_0 together determine all joint distributions

Recap

- The Markov property implies

$$\begin{aligned} Pr[X_{m+n} = y | X_m = x, X_0 = z] &= Pr[X_{m+n} = y | X_m = x] \\ &= Pr[X_n = y | X_0 = x] \end{aligned}$$

- Or, in general,

$$f_{X_{m+n}|X_m,\cdots,X_0}(y|x, \cdots) = f_{X_{m+n}|X_m}(y|x)$$

- Further, we can show

$$\mathrm{Pr}[X_{m+n} = y | X_m = x,\ X_{m-k} \in A_k,\ k = 1, \cdots, m] = Pr[X_{m+n} = y | X_m = x]$$

$$Pr[X_{m+n+r} \in B_r,\ r = 0, \cdots, s \mid X_m = x,\ X_{m-k} \in A_k,\ k = 1, \cdots, m] = Pr[X_{m+n+r} \in B_r,\ r = 0, \cdots, s \mid X_m = x]$$

Recap: Chapman-Kolmogorov Equations

- The n-step transition probabilities are defined by

$$P^n(x,y) = Pr[X_n = y | X_0 = x]$$

- These n-step transition probabilities satisfy

$$P^{m+n}(x,y) = \sum_z P^m(x,z) P^n(z,y)$$

- These are known as Chapman-Kolmogorov equations
- For a finite chain, the n-step transition probability matrix is n-fold product of the transition probability matrix
- We also have

$$\pi_n(x) \triangleq Pr[X_n = x] = \sum_x \pi_0(x) P^n(x,y)$$

Recap: Hitting times

- We define hitting time for y as the random variable

$$T_y = \min\{n > 0 \ : \ X_n = y\}$$

- Using this defintion, we can derive

$$P_x(T_y = m) = \sum_{z \neq y} P(x,z) P_z(T_y = m-1)$$

(Notation: $P_z(A) = Pr[A | X_0 = z]$)

$$P^n(x,y) = \sum_{m=1}^{n} P_x(T_y = m) P^{n-m}(y,y)$$

Recap: transient and recurrent states

- Define $\rho_{xy} = P_x(T_y < \infty)$.
- It is the probability that starting in x you will visit y
- Note that

$$\rho_{xy} = \lim_{n \to \infty} P_x(T_y < n) = \sum_{n=1}^{\infty} P_x(T_y = n)$$

Definition: A state y is called transient if $\rho_{yy} < 1$; it is called recurrent if $\rho_{yy} = 1$.

- Intuitively, all transient states would be visited only finitely many times while recurrent states are visited infinitely often.

Recap

- For any state y define

$$I_y(X_n) = \begin{cases} 1 & \text{if } X_n = y \\ 0 & \text{otherwise} \end{cases}$$

- The total number of visits to y is given by

$$N(y) = \sum_{n=1}^{\infty} I_y(X_n)$$

- We can get distribution of $N(y)$ as

$$P_x(N(y) = m) = \rho_{xy}\, \rho_{yy}^{m-1}(1 - \rho_{yy}), \ m \geq 1$$

and $P_x(N(y) = 0) = 1 - \rho_{xy}$

Recap

- Notation: $E_x[Z] = E[Z|X_0 = x]$
- Define

$$\begin{aligned} G(x,y) &\triangleq E_x[N(y)] \\ &= \sum_{n=1}^{\infty} E_x\left[I_y(X_n)\right] \\ &= \sum_{n=1}^{\infty} P^n(x,y) \end{aligned}$$

- $G(x,y)$ is the expected number of visits to y for a chain that is started in x.

Theorem:
(i). Let y be transient. Then

$$P_x(N(y) < \infty) = 1, \ \forall x \ \text{ and } \ G(x,y) = \frac{\rho_{xy}}{1-\rho_{yy}} < \infty, \ \forall x$$

(ii) Let y be recurrent. Then

$$P_y[N(y) = \infty] = 1, \ \text{ and } \ G(y,y) = E_y[N(y)] = \infty$$

$$P_x[N(y) = \infty] = \rho_{xy}, \ \text{ and } \ G(x,y) = \begin{cases} 0 & \text{if } \rho_{xy} = 0 \\ \infty & \text{if } \rho_{xy} > 0 \end{cases}$$

Proof of (i): y is transient; $\rho_{yy} < 1$

$$\begin{aligned} G(x,y) &= E_x[N(y)] = \sum_m m P_x[N(y) = m] \\ &= \sum_m m\, \rho_{xy}\, \rho_{yy}^{m-1}(1-\rho_{yy}) \\ &= \rho_{xy} \sum_{m=1}^{\infty} m\, \rho_{yy}^{m-1}(1-\rho_{yy}) \\ &= \rho_{xy} \frac{1}{1-\rho_{yy}} < \infty, \ \text{ because } \ \rho_{yy} < 1 \end{aligned}$$

$$\Rightarrow \ P_x[N(y) < \infty] = 1$$

Proof of (ii):
y recurrent $\Rightarrow$ $\rho_{yy} = 1$. Hence

$$\begin{aligned} P_y[N(y) \geq m] &= \rho_{yy}^m = 1, \ \forall m \\ \Rightarrow \ P_y[N(y) = \infty] &= \lim_{m \to \infty} P_y[N(y) \geq m] = 1 \\ \Rightarrow \ G(y,y) &= E_y[N(y)] = \infty \end{aligned}$$

$$P_x[N(y) \geq m] = \rho_{xy}\, \rho_{yy}^{m-1} = \rho_{xy}, \ \forall m$$

Hence $P_x[N(y) = \infty] = \rho_{xy}$

$$\rho_{xy} = 0 \ \Rightarrow \ P_x[N(y) \geq m] = 0, \ \forall m > 0 \ \Rightarrow \ G(x,y) = 0$$

$$\rho_{xy} > 0 \ \Rightarrow \ P_x[N(y) = \infty] > 0 \ \Rightarrow \ G(x,y) = \infty$$

- Transient states are visited only finitely many times while recurrent states are visited infinitely often
- If S is finite, it should have at least one recurrent state
- If y is transient, then, for all x

$$G(x,y) = \sum_{n=1}^{\infty} P^n(x,y) < \infty \quad \Rightarrow \quad \lim_{n\to\infty} P^n(x,y) = 0$$

- However, $\sum_y P^n(x,y) = 1, \ \forall n, \ \forall x$
- If all $y \in S$ are transient, then we get a contradiction

$$1 = \lim_{n\to\infty} \sum_{y \in S} P^n(x,y) = \sum_{y\in S} \lim_{n\to\infty} P^n(x,y) = 0$$

- A finite chain has to have at least one recurrent state
- An infinite chain can have only transient states

- We say, x leads to y if $\rho_{xy} > 0$

Theorem: If x is recurrent and x leads to y then y is recurrent and $\rho_{xy} = \rho_{yx} = 1$.

Proof:

- Take $x \neq y$, wlog. Since $\rho_{xy} > 0$, $\exists n$ s.t. $P^n(x,y) > 0$
- Take least such n. Then we have states $y_1, \cdots, y_{n-1}$, none of which is x (or y) such that

$$P(x,y_1)\ P(y_1,y_2)\cdots P(y_{n-1},y) > 0$$

- Now suppose, $\rho_{yx} < 1$. Then

$$P(x,y_1)\ P(y_1,y_2)\cdots P(y_{n-1},y)(1-\rho_{yx}) > 0$$

is the probability of starting in x but not returning to x.

- But this cannot be because x is recurrent and hence $\rho_{xx} = 1$
- Hence, if x is recurrent and x leads to y then $\rho_{yx} = 1$

- Now, $\exists n_0, n_1$ s.t. $P^{n_0}(x,y) > 0$, $P^{n_1}(y,x) > 0$.

$$\begin{aligned} P^{n_1+n+n_0}(y,y) &= P_y[X_{n_1+n+n_0} = y] \\ &\geq P_y[X_{n_1} = x, X_{n_1+n} = x, X_{n_1+n+n_0} = y] \\ &= P^{n_1}(y,x)P^n(x,x)P^{n_0}(x,y), \quad \forall n \end{aligned}$$

- We know $G(x,x) = \sum_{m=1}^{\infty} P^m(x,x) = \infty$

$$\begin{aligned} \sum_{m=1}^{\infty} P^m(y,y) &\geq \sum_{m=n_0+n_1+1}^{\infty} P^m(y,y) = \sum_{n=1}^{\infty} P^{n_1+n+n_0}(y,y) \\ &\geq \sum_{n=1}^{\infty} P^{n_1}(y,x)P^n(x,x)P^{n_0}(x,y) \\ &= \infty, \quad \text{because } x \text{ is recurrent} \\ &\Rightarrow \quad y \text{ is recurrent} \end{aligned}$$

- What we showed so far is: if x leads to y and x is recurrent, then $\rho_{yx} = 1$ and y is recurrent.
- Now, y is recurrent and y leads to x and hence $\rho_{xy} = 1$.
- This completes proof of the theorem

equivalence relation

- let R be a relation on set A. Note $R \subset A \times A$
- R is called an equivalence relation if it is
  1. reflexive, i.e., $(x, x) \in R,\ \forall x \in A$
  2. symmetric, i.e., $(x, y) \in R \ \Rightarrow \ (y, x) \in R$
  3. transitive, i.e., $(x, y), (y, z) \in R \ \Rightarrow \ (x, z) \in R$

example

- Let $A = \{\frac{m}{n} \mid m, n$ are integers$\}$
- Define relation R by

$$\left(\frac{m}{n}, \frac{p}{q}\right) \in R \text{ if } mq = np$$

- This is the usual equality of fractions
- Easy to check it is an equivalence relation.

Equivalence classes

- Let R be an equivalence relation on A.
- Then, A can be partitioned as

$$A = C_1 + C_2 + \cdots$$

  Where C_i satisfy
  - $x, y \in C_i \ \Rightarrow \ (x, y) \in R,\ \forall i$
  - $x \in C_i,\ y \in C_j,\ i \neq j \ \Rightarrow \ (x, y) \notin R$
- In our example, each equivalence class corresponds to a rational number.
- Here, C_i contains all fractions that are equal to that rational number

- The state space of any Markov chain can be partitioned into the transient and recurrent states: $S = S_T + S_R$:

$$S_T = \{y \in S \ : \ \rho_{yy} < 1\} \qquad S_R = \{y \in S \ : \ \rho_{yy} = 1\}$$

- On S_R, consider the relation: 'x leads to y' (i.e., x is related to y if $\rho_{xy} > 0$)
- This is an equivalence relation
  - $\rho_{xx} > 0,\ \forall x \in S_R$
  - $\rho_{xy} > 0 \ \Rightarrow \ \rho_{yx} > 0,\ \forall x, y \in S_R$
  - $\rho_{xy} > 0,\ \rho_{yz} > 0 \ \Rightarrow \ \rho_{xz} > 0$
- Hence we get a partition: $S_R = C_1 + C_2 + \cdots$ where C_i are equivalence classes.

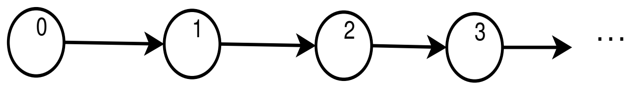
- On S_R, "x leads to y" is an equivalence relation.
- This gives rise to the partition $S_R = C_1 + C_2 + \cdots$
- Since C_i are equivalence classes, they satisfy:
  - $x, y \in C_i \quad \Rightarrow \quad x$ leads to y
  - $x \in C_i, y \in C_j, i \neq j \quad \Rightarrow \quad \rho_{xy} = 0$
- All states in any C_i lead to each other or communicate with each other
- If $i \neq j$ and $x \in C_i$ and $y \in C_j$, then, $\rho_{xy} = \rho_{yx} = 0$. x and y do not communicate with each other.

- A set of states, $C \subset S$ is said to be irreducible if x leads to y for all $x, y \in C$
- An irreducible set is also called a communicating class
- A set of states, $C \subset S$, is said to be closed if $x \in C,\ y \notin C$ implies $\rho_{xy} = 0$.
- Once the chain visits a state in a closed set, it cannot leave that set.
- We get a partition of recurrent states

$$S_R = C_1 + C_2 + \cdots$$

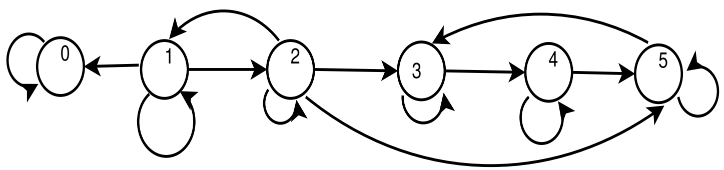
  where each C_i is a closed and irreducible set of states.
- If S is irreducible then the chain is said to be irreducible. (Note that S is trivially closed)

- In an irreducible set of states, if one state is recurrent, then, all states are recurrent.
- We saw that a finite chain has to have at least one recurrent state.
- Thus, a finite irreducible chain is recurrent.
- For example, in the umbrellas problem, the chain is irreducible and hence all states are recurrent.
- An infinite irreducible chain may be wholly transient
- Here is a trivial example of non-irreducible transient chain:



- The state space of any Markov chain can be partitioned into transient and recurrent states.
- We need not calculate ρ_{xx} to do this partition.
- By looking at the structure of transition probability matrix we can get this partition

Example



$$
\begin{array}{c|cccccc}
 & 0 & 1 & 2 & 3 & 4 & 5 \\
\hline
0 & + & - & - & - & - & - \\
1 & + & + & + & - & - & - \\
2 & - & + & + & + & - & + \\
3 & - & - & - & + & + & - \\
4 & - & - & - & - & + & + \\
5 & - & - & - & + & - & +
\end{array}
$$

- State 0 is called an absorbing state. $\{0\}$ is a closed irreducible set.
- $1, 2$ are transient states.
- We get: $S_T = \{1, 2\}$ and $S_R = \{0\} + \{3, 4, 5\}$

- If you start the chain in a recurrent state it will stay in the corresponding closed irreducible set
- If you start in one of the transient states, it would eventually get 'absorbed' in one of the closed irreducible sets of recurrent states.
- We want to know the probabilities of ending up in different sets.
- We want to know how long you stay in transient states
- We want to know what is the 'steady state'?

- let C be a closed irreducible set of recurrent states
- T_C – hitting time for C.
  $T_C = \min\{n > 0 \ : \ X_n \in C\}$
  It is the first time instant when the chain is in C
- Define $\rho_C(x) = P_x[T_C < \infty]$

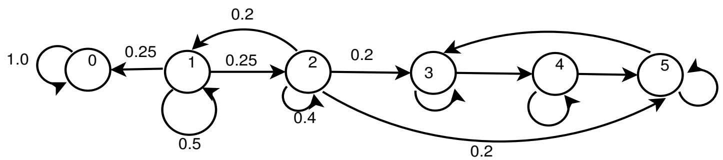
$$\text{If } x \text{ is recurrent,} \quad \rho_C(x) = \begin{cases} 1 & \text{if } x \in C \\ 0 & \text{if } x \notin C \end{cases}$$

  Because each x is in a closed irreducible set
- Suppose x is transient. Then

$$\rho_C(x) = \sum_{y \in C} P(x, y) + \sum_{y \in S_T} P(x, y) \ \rho_C(y)$$

- By solving this set of linear equations we can get $\rho_c(x), \ x \in S_T$

Example: Absorption probabilities



- $S_T = \{1, 2\}$ and $C_1 = \{0\}$, $C_2 = \{3, 4, 5\}$

$$\rho_C(x) = \sum_{y \in C} P(x, y) + \sum_{y \in S_T} P(x, y) \ \rho_C(y)$$

$$
\begin{aligned}
\rho_{C_1}(1) &= P(1, 0) + P(1, 1)\rho_{C_1}(1) + P(1, 2)\rho_{C_1}(2) \\
&= 0.25 + 0.5\rho_{C_1}(1) + 0.25\rho_{C_1}(2) \\
\rho_{C_1}(2) &= 0 + 0.2\rho_{C_1}(1) + 0.4\rho_{C_1}(2)
\end{aligned}
$$

- Solving these, we get $\rho_{C_1}(1) = 0.6, \quad \rho_{C_1}(2) = 0.2$
- What would be $\rho_{C_2}(1)$?

Expected time in transient states

- We consider a simple method to get the time spent in transient states for finite chains
- Let states $1, 2, \cdots, t$ be the transient states
- b_{ij} – the expected number of time instants spent in state j when started in i.
- Then we get

$$b_{ij} = \delta_{ij} + \sum_{k=1}^{t} P(i,k) b_{kj}$$

where $\delta_{ij} = 1$ if $i = j$ and is zero otherwise
- let B be the $t \times t$ matrix of b_{ij}, I be the $t \times t$ identity matrix and P_T be the submatrix (corresponding to the transient states) of P.
- Then the above in Matrix notation is

$$B = I + P_T B \quad \text{or} \quad B = (I - P_T)^{-1}$$

stationary distributions

- $\pi : S \to [0,\ 1]$ is a probability distribution (mass function) over S if $\pi(x) \geq 0,\ \forall x$ and $\sum_{x \in S} \pi(x) = 1$
- A probability distribution over S, π, is said to be a stationary distribution for the Markov chain with transition probabilities P if

$$\pi(y) = \sum_{x \in S} \pi(x) P(x,y), \quad \forall y \in S$$

- Suppose S is finite. Then π can be represented by a vector.
- The π is stationary if

$$\pi^T = \pi^T P \quad \text{or} \quad P^T \pi = \pi$$

- π is a stationary distribution if

$$\pi(y) = \sum_{x \in S} \pi(x) P(x,y), \quad \forall y \in S$$

- Recall $\pi_n(x) \triangleq Pr[X_n = x]$ satisfies

$$\pi_{n+1}(y) = \sum_{x \in S} Pr[X_{n+1} = y | X_n = x] Pr[X_n = x] = \sum_{x \in S} \pi_n(x) P(x,y)$$

- Hence, if $\pi_0 = \pi$ then $\pi_1 = \pi$ and hence $\pi_n = \pi,\ \forall n$
- Hence the name, stationary distribution.
- It is also called the invariant distribution or the invariant measure

- If the chain is started in stationary distribution then the distribution of X_n is not a function of time, as we saw.
- Suppose for a chain, distribution of X_n is not dependent on n. Then the chain must be in a stationary distribution.
- Suppose $\pi = \pi_0 = \pi_1 = \cdots = \pi_n = \cdots$. Then

$$\pi(y) = \pi_1(y) = \sum_{x \in S} \pi_0(x) P(x,y) = \sum_{x \in S} \pi(x) P(x,y)$$

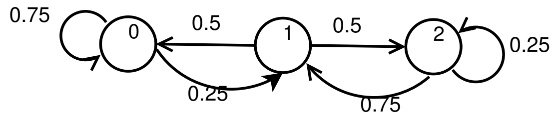
which shows π is a stationary distribution

- Suppose S is finite.
- Then π is a stationary distribution if

$$P^T \pi = \pi \quad \text{or} \quad \left(P^T - I\right)\, \pi = 0$$

- Note that each column of P^T sums to 1.
- Hence, $\left(P^T - I\right)$ would be singular
(1 is always an eigen value for a column stochastic matrix)
- A stationary distribution always exists for a finite chain.
- But it may or may not be unique.
- What about infinite chains?

Example



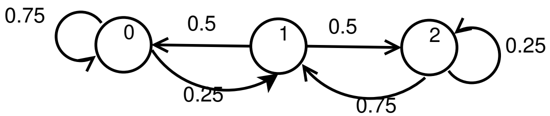
- The stationary distribution has to satisfy

$$\pi(y) = \sum_{x \in S} \pi(x) P(x, y), \quad \forall y \in S$$

- Thus we get the following linear equations

$$\begin{aligned} 0.75\pi(0) + 0.5\pi(1) &= \pi(0) \\ 0.25\pi(0) + 0.75\pi(2) &= \pi(1) \\ 0.5\pi(1) + 0.25\pi(2) &= \pi(2) \end{aligned}$$

$$\text{in addition,} \quad \pi(0) + \pi(1) + \pi(2) = 1$$

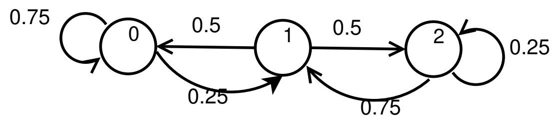


- We can also write the equations for π as

$$\left[\begin{array}{ccc} \pi(0) & \pi(1) & \pi(2) \end{array}\right] \left[\begin{array}{ccc} 0.75 & 0.25 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0.75 & 0.25 \end{array}\right] = \left[\begin{array}{ccc} \pi(0) & \pi(1) & \pi(2) \end{array}\right]$$

$$\begin{aligned} 0.75\pi(0) + 0.5\pi(1) &= \pi(0) \\ 0.25\pi(0) + 0.75\pi(2) &= \pi(1) \\ 0.5\pi(1) + 0.25\pi(2) &= \pi(2) \end{aligned}$$

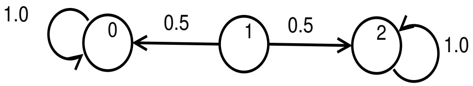
- We have to solve these along with $\pi(0) + \pi(1) + \pi(2) = 1$



$$\begin{aligned} 0.75\pi(0) + 0.5\pi(1) &= \pi(0) \;\Rightarrow\; \pi(1) = \frac{1}{2}\pi(0) \\ 0.25\pi(0) + 0.75\pi(2) &= \pi(1) \;\Rightarrow\; \pi(2) = \frac{1}{3}\pi(0) \\ 0.5\pi(1) + 0.25\pi(2) &= \pi(2) \\ \pi(0) + \pi(1) + \pi(2) &= 1 \;\Rightarrow\; \pi(0)\left(1 + \frac{1}{2} + \frac{1}{3}\right) = 1 \end{aligned}$$

- Now, $\pi(0)\left(1 + \frac{1}{2} + \frac{1}{3}\right) = 1$ gives $\pi(0) = \frac{6}{11}$
- We get a unique solution: $\left[\frac{6}{11} \;\; \frac{3}{11} \;\; \frac{2}{11}\right]$

Example2

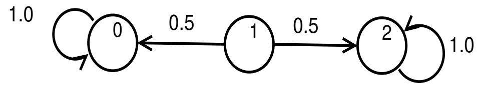


- The stationary distribution has to satisfy

$$\left[\begin{array}{ccc} \pi(0) & \pi(1) & \pi(2) \end{array}\right] \left[\begin{array}{ccc} 1.0 & 0 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0 & 1.0 \end{array}\right] = \left[\begin{array}{ccc} \pi(0) & \pi(1) & \pi(2) \end{array}\right]$$

- We also have to add the equation $\pi(0) + \pi(1) + \pi(2) = 1$
- We now do not have a unique stationary distribution

Example2



$$\pi(y) = \sum_{x \in S} \pi(x) P(x, y), \quad \forall y \in S$$

- We get the following linear equations

$$\begin{array}{rcl} \pi(0) + 0.5\pi(1) & = & \pi(0) \;\Rightarrow\; \pi(1) = 0 \\ 0.5\pi(1) + \pi(2) & = & \pi(2) \;\Rightarrow\; \pi(1) = 0 \\ \pi(0) + \pi(1) + \pi(2) & = & 1 \;\Rightarrow\; \pi(0) = 1 - \pi(2) \end{array}$$

- Now there are infinitely many solutions.
- Any distribution $[a \;\; 0 \;\; 1-a]$ with $0 \leq a \leq 1$ is a stationary distribution
- This chain is not irreducible; the previous one is irreducible

- We now explore conditions for existence and uniqueness of stationary distributions
- For finite chains stationary distribution always exists.
- For finite irreducible chains it is unique.
- But for infinite chains, it is possible that stationary distribution does not exist.
- When the stationary distribution is unique, we also want to know if the chain converges to that distribution
- The stationary distribution, when it exists, is related to expected fraction of time spent in different states.

- Let $I_y(X_n)$ be indicator of $[X_n = y]$
- Number of visits to y till n: $\quad N_n(y) = \sum_{m=1}^{n} I_y(X_m)$

$$G_n(x, y) \triangleq E_x[N_n(y)] = \sum_{m=1}^{n} E_x[I_y(X_n)] = \sum_{m=1}^{n} P^m(x, y)$$

- Expected fraction of time spent in y till n is

$$\frac{G_n(x, y)}{n} = \frac{1}{n} \sum_{m=1}^{n} P^m(x, y)$$

- We will first establish a limit for the above as $n \to \infty$

Recap: Markov Chain

- Let $X_n, \; n = 0, 1, \cdots$ be a sequence of discrete random variables taking values in S.
- We say it is a Markov chain if

$$Pr[X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1} \cdots X_0 = x_0] = Pr[X_{n+1} = x_{n+1} | X_n = x_n],$$

- We can write it as

$$f_{X_{n+1}|X_n, \cdots X_0}(x_{n+1}|x_n, \cdots, x_0) = f_{X_{n+1}|X_n}(x_{n+1}|x_n), \; \forall x_i$$

- For a Markov chain, given the current state, the future evolution is independent of the history of how you reached the current state

Recap: Transition Probabilities

- Transition probabilities: $P(x, y) = Pr[X_{n+1} = y | X_n = x]$
  Chain is homogeneous:
  $Pr[X_{n+1} = y | X_n = x] = Pr[X_1 = y | X_0 = x], \; \forall n$
- Initial probabilities $\pi_0(x) = Pr[X_0 = x]$
- Similarly, $\pi_n(x) = Pr[X_n = x]$

Recap: Chapman-Kolmogorov Equations

- n-step transition probabilities:
  $P^n(x, y) = Pr[X_n = y | X_0 = x]$
- These satisfy Chapman-Kolmogorov equations:

$$P^{m+n}(x, y) = \sum_z P^m(x, z) P^n(z, y)$$

- For a finite chain, the n-step transition probability matrix is n-fold product of the transition probability matrix

Recap: transient and recurrent states

- Hitting time for y: $T_y = \min\{n > 0 \; : \; X_n = y\}$
- $\rho_{xy} = P_x(T_y < \infty)$.
- A state y is called transient if $\rho_{yy} < 1$; it is called recurrent if $\rho_{yy} = 1$.
- $N(y)$ – total number of visits to y
- $G(x, y) = E_x[N(y)]$

Recap

Theorem:
(i). Let y be transient. Then

$$P_x(N(y) < \infty) = 1, \; \forall x \;\; \text{and} \;\; G(x,y) = \frac{\rho_{xy}}{1 - \rho_{yy}} < \infty, \; \forall x$$

(ii) Let y be recurrent. Then

$$P_y[N(y) = \infty] = 1, \;\; \text{and} \;\; G(y,y) = E_y[N(y)] = \infty$$

$$P_x[N(y) = \infty] = \rho_{xy}, \;\; \text{and} \;\; G(x,y) = \begin{cases} 0 & \text{if } \rho_{xy} = 0 \\ \infty & \text{if } \rho_{xy} > 0 \end{cases}$$

Recap

- Transient states are visited only finitely many times while recurrent states are visited infinitely often
- A finite chain should have at least one recurrent state
- We say, x leads to y if $\rho_{xy} > 0$
  Theorem: If x is recurrent and x leads to y then y is recurrent and $\rho_{xy} = \rho_{yx} = 1$.

Recap: closed and irreducible sets

- A set of states, $C \subset S$ is said to be irreducible if x leads to y for all $x, y \in C$
- An irreducible set is also called a communicating class
- A set of states, $C \subset S$, is said to be closed if $x \in C, \; y \notin C$ implies $\rho_{xy} = 0$.
- Once the chain visits a state in a closed set, it cannot leave that set.

Recap: Partition of state space

- $S = S_T + S_R$, transient and recurrent states and

$$S_R = C_1 + C_2 + \cdots$$

  where C_i are closed and irreducible
- We can calculate absorption probabilities for C_i using

$$\rho_C(x) = \sum_{y \in C} P(x,y) + \sum_{y \in S_T} P(x,y) \; \rho_C(y)$$

Recap: Stationary distribution

- π is said to be a stationary distribution for the Markov chain with transition probabilities P if

$$\pi(y) = \sum_{x \in S} \pi(x) P(x, y), \quad \forall y \in S$$

- For finite chains, $P^T \pi = \pi$
- When π is stationary distribution,
$\pi_0 = \pi \quad \Rightarrow \quad \pi_n = \pi, \ \forall n$
- If $\pi_n = \pi, \ \forall n$ then π is a stationary distribution
- For a finite chain, a stationary distribution always exists.
- The stationary distribution, when it exists, is related to expected fraction of time spent in different states.

- Let $I_y(X_n)$ be indicator of $[X_n = y]$
- Number of visits to y till n: $N_n(y) = \sum_{m=1}^{n} I_y(X_m)$

$$G_n(x, y) \triangleq E_x[N_n(y)] = \sum_{m=1}^{n} E_x[I_y(X_m)] = \sum_{m=1}^{n} P^m(x, y)$$

- Expected fraction of time spent in y till n is

$$\frac{G_n(x, y)}{n} = \frac{1}{n} \sum_{m=1}^{n} P^m(x, y)$$

- We will first establish a limit for the above as $n \to \infty$

- Suppose y is transient. Then we have

$$\lim_{n \to \infty} N_n(y) = N(y)$$

and $\quad Pr[N(y) < \infty] = 1 \qquad \lim_{n \to \infty} G_n(x, y) = G(x, y) < \infty$

$$\Rightarrow \quad \lim_{n \to \infty} \frac{N_n(y)}{n} = 0 \ (w.p.1) \quad \text{and} \quad \lim_{n \to \infty} \frac{G_n(x, y)}{n} = 0$$

- The expected fraction of time spent in a transient state is zero.
- This is intuitively obvious

- Now, let y be recurrent
- Then, $P_y[T_y < \infty] = 1$
- Define $m_y = E_y[T_y]$
- m_y is mean return time to y
- We will show that $\frac{N_n(y)}{n}$ converges to $\frac{1}{m_y}$ if the chain starts in y.
- Convergence would be with probability one.

- Consider a chain started in y
- let T_y^r be time of r^{th} visit to y, $r \geq 1$

$$T_y^r = \min\{n \geq 1 \ : \ N_n(y) = r\}$$

- Define $W_y^1 = T_y^1 = T_y$ and $W_y^r = T_y^r - T_y^{r-1}, \ \ r > 1$
- Note that $E_y[W_y^1] = E_y[T_y] = m_y$
- Also, $T_y^r = W_y^1 + \cdots + W_y^r$
- W_y^r are the "waiting times"
- By Markovian property we should expect them to be iid
- We will prove this.
- Then T_y^r/r converges to m_y by law of large umbers

- We have

$$Pr[W_y^3 = k_3 | W_y^2 = k_2, W_y^1 = k_1] \ \ =$$
$$Pr[X_{k_1+k_2+j} \neq y, \ 1 \leq j \leq k_3 - 1, \ \ X_{k_1+k_2+k_3} = y \mid B]$$

where $B = [X_{k_1+k_2} = y, \ X_{k_1} = y, \ X_j \neq y, j < k_1 + k_2, j \neq k_1]$

- Using the Markovian property, we get

$$Pr[W_y^3 = k_3 | W_y^2 = k_2, W_y^1 = k_1] \ \ =$$
$$Pr[X_{k_1+k_2+j} \neq y, \ 1 \leq j \leq k_3 - 1, \ \ X_{k_1+k_2+k_3} = y \mid X_{k_1+k_2} = y]$$
$$= Pr[X_j \neq y, \ 1 \leq j \leq k_3 - 1, \ \ X_{k_3} = y \mid X_0 = y]$$
$$= P_y[W_y^1 = k_3]$$

- In general, we get

$$Pr[W_y^r = k_r \mid W_y^{r-1} = k_{r-1}, \cdots, W_y^1 = k_1] = P_y[W_y^1 = k_r]$$

- This shows the waiting time are iid

$$\begin{aligned} P_y[W_y^2 = k_2] &= \sum_{k_1} P_y[W_y^2 = k_2 \mid W_y^1 = k_1] \ P_y[W_y^1 = k_1] \\ &= \sum_{k_1} P_y[W_y^1 = k_2] \ P_y[W_y^1 = k_1] \\ &= P_y[W_y^1 = k_2] \end{aligned}$$

$\Rightarrow$ identically distributed

$$\begin{aligned} P_y[W_y^2 = k_2, \ W_y^1 = k_1] &= P_y[W_y^2 = k_2 \mid W_y^1 = k_1] P_y[W_y^1 = k_1] \\ &= P_y[W_y^1 = k_2] \ P_y[W_y^1 = k_1] \\ &= P_y[W_y^2 = k_2] \ P_y[W_y^1 = k_1] \end{aligned}$$

$\Rightarrow$ independent

- We have shown $W_y^r, \ r = 1, 2, \cdots$ are iid
- Since $E[W_y^1] = m_y$, by strong law of large numbers,

$$\lim_{k \to \infty} \frac{T_y^k}{k} = \lim_{k \to \infty} \frac{1}{k} \sum_{r=1}^{k} W_y^r = m_y, \ \ (w.p.1)$$

- Note that this is true even if $m_y = \infty$

- For all n such that $N_n(y) \geq 1$, we have

$$T_y^{N_n(y)} \leq n < T_y^{N_n(y)+1}$$

- $N_n(y)$ is the number of visits to y till time step n
- Suppose $N_{50}(y) = 8$ – Visited y 8 times till time 50.
- So, the 8^{th} visit occurred at or before time 50.
- The 9^{th} visit has not occurred till 50.
- So, time of 9^{th} visit is beyond 50.

$$T_y^{N_n(y)} \leq n < T_y^{N_n(y)+1}$$

- Now we have

$$\frac{T_y^{N_n(y)}}{N_n(y)} \leq \frac{n}{N_n(y)} < \frac{T_y^{N_n(y)+1}}{N_n(y)}$$

- We know that
  - As $n \to \infty$, $N_n(y) \to \infty$, $w.p.1$
  - As $n \to \infty$, $\frac{T_y^n}{n} \to m_y$, $w.p.1$
- Hence we get

$$\lim_{n\to\infty} \frac{n}{N_n(y)} = m_y, \quad w.p.1$$

or

$$\lim_{n\to\infty} \frac{N_n(y)}{n} = \frac{1}{m_y}, \quad w.p.1$$

- All this is true if the chain started in y.
- That means it is true if the chain visits y once.
- So, we get

$$\lim_{n\to\infty} \frac{N_n(y)}{n} = \frac{I_{[T_y<\infty]}}{m_y}, \quad w.p.1$$

- Since $0 \leq \frac{N_n(y)}{n} \leq 1$, almost sure convergence implies convergence in mean

$$\lim_{n\to\infty} \frac{G_n(x,y)}{n} = \lim_{n\to\infty} E_x\left[\frac{N_n(y)}{n}\right] = \lim_{n\to\infty} \frac{P_x[T_y<\infty]}{m_y} = \frac{\rho_{xy}}{m_y}$$

- The fraction of time spent in each recurrent state is inversely proportional to the mean recurrence time

- Thus we have proved the following theorem
- **Theorem**:
  Let y be recurrent. Then
  1. $$\lim_{n\to\infty} \frac{N_n(y)}{n} = \frac{I_{[T_y<\infty]}}{m_y}, \quad w.p.1$$
  2. $$\lim_{n\to\infty} \frac{G_n(x,y)}{n} = \frac{\rho_{xy}}{m_y}$$

- The limiting fraction of time spent in a state is inversely proportional to m_y, the mean return time.
- Intuitively, the stationary probability of a state could be the limiting fraction of time spent in that state.
- Thus $\pi(y) = \frac{1}{m_y}$ is a good candidate for stationary distribution.
- We first note that we can have $m_y = \infty$.
  Though $P_y[T_y < \infty] = 1$, we can have $E_y[T_y] = \infty$.
- What if $m_y = \infty, \quad \forall y$?
- Does not seem reasonable for a finite chain.
- But for infinite chains??
- Let us characterize y for which $m_y = \infty$

- A recurrent state y is called **null recurrent** if $m_y = \infty$.
- y is called **positive recurrent** if $m_y < \infty$
- We earlier saw that the fraction of time spent in a transient state is zero.
- Suppose y is null recurrent. Then

$$\lim_{n\to\infty} \frac{N_n(y)}{n} = \frac{1}{m_y} = 0$$

- Thus the limiting fraction of time spent by the chain in transient and null recurrent states is zero.

- **Theorem:** Let x be positive recurrent and let x lead to y. Then y is positive recurrent.
  Proof
- Since x is recurrent and x leads to y we know $\exists n_0, n_1$ s.t. $P^{n_0}(x,y) > 0$, $P^{n_1}(y,x) > 0$ and

$$P^{n_1+m+n_0}(y,y) \geq P^{n_1}(y,x) P^m(x,x) P^{n_0}(x,y), \quad \forall m$$

Summing the above for $m = 1, 2, \cdots n$ and dividing by n

$$\frac{1}{n}\sum_{m=1}^{n} P^{n_1+m+n_0}(y,y) \geq P^{n_1}(y,x) \ \frac{1}{n}\sum_{m=1}^{n} P^m(x,x) \ \ P^{n_0}(x,y), \ \forall n$$

If we now let $n \to \infty$, the RHS goes to $P^{n_1}(y,x) \ \frac{1}{m_x} \ P^{n_0}(x,y) > 0$.

$$\frac{1}{n}\sum_{m=1}^{n} P^{n_1+m+n_0}(y,y) \geq P^{n_1}(y,x) \ \frac{1}{n}\sum_{m=1}^{n} P^m(x,x) \ \ P^{n_0}(x,y), \ \forall n$$

- We can write the LHS of above as

$$\frac{1}{n}\sum_{m=1}^{n} P^{n_1+m+n_0}(y,y) \ = \ \frac{1}{n}\sum_{m=1}^{n_1+n+n_0} P^m(y,y) - \frac{1}{n}\sum_{m=1}^{n_1+n_0} P^m(y,y)$$

$$= \frac{n_1+n+n_0}{n} \frac{1}{n_1+n+n_0} \sum_{m=1}^{n_1+n+n_0} P^m(y,y) - \frac{1}{n}\sum_{m=1}^{n_1+n_0} P^m(y,y)$$

$$\Rightarrow \ \lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} P^{n_1+m+n_0}(y,y) = \frac{1}{m_y}$$

$$\Rightarrow \ \frac{1}{m_y} \geq P^{n_1}(y,x) \ \frac{1}{m_x} \ P^{n_0}(x,y) > 0$$

which implies y is positive recurrent

- Thus, in a closed irreducible set of recurrent states, if one state is positive recurrent then all are positive recurrent.
- Hence, in the partition: $S_R = C_1 + C_2 + \cdots$, each C_i is either wholly positive recurrent or wholly null recurrent.
- We next show that a finite chain cannot have any null recurrent states.

- Let C be a finite closed set of recurrent states.
- Suppose all states in C are null recurrent. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} P^m(x,y) = 0, \quad \forall x, y \in C$$

- Since C is closed, $\sum_{y\in C} P^m(x,y) = 1, \ \forall m, \ \forall x \in C$.
- Thus we get

$$1 = \frac{1}{n} \sum_{m=1}^{n} \sum_{y\in C} P^m(x,y) = \sum_{y\in C} \frac{1}{n} \sum_{m=1}^{n} P^m(x,y), \ \forall n$$

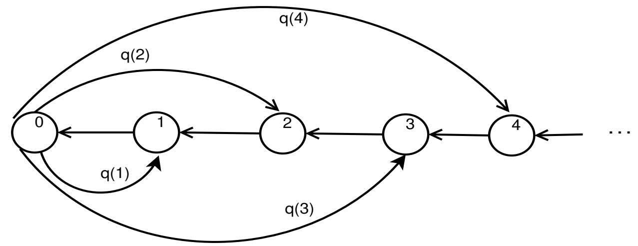
$$\Rightarrow \ 1 = \lim_{n\to\infty} \sum_{y\in C} \frac{1}{n} \sum_{m=1}^{n} P^m(x,y) = \sum_{y\in C} \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} P^m(x,y) = 0$$

where we could take the limit inside the sum because C is finite.

- If C is a finite closed set of recurrent states then all states in it cannot be null recurrent.
- Actually what we showed is that any closed finite set must have at least one positive recurrent state.
- Hence, in a finite chain, every closed irreducible set of recurrent states contains only positive recurrent states.
- Hence, a finite chain cannot have a null recurrent state.

Example of null recurrent chain

- Consider the chain with state space $\{0, 1, \cdots\}$ given by



- Here, $q(k) \geq 0, \forall k$ and $\sum_{k=1}^{\infty} q(k) = 1$. We have

$$P_0[T_0 = j{+}1] = q(j) \quad \Rightarrow m_0 = \sum_{j=2}^{\infty} j \ P_o[T_0 = j] = \sum_{j=2}^{\infty} j \ q(j{-}1)$$

(Note that $P_0[T_0 = 1] = 0$)
- So, $m_0 = \infty$ if the $q(\cdot)$ distribution has infinite expectation. For example, if $q(k) = \frac{c}{k^2}$
- Then state 0 is null recurrent. Implies chain is null recurrent

- Suppose π is a stationary distribution.
- Then $\pi(y) = 0$ if y is transient or null recurrent
- We prove this as follows

$$\begin{aligned}
\pi(y) &= \sum_x \pi(x) P^m(x,y) \ \ \forall m \\
\Rightarrow \ \pi(y) &= \frac{1}{n}\sum_{m=1}^{n}\sum_x \pi(x) P^m(x,y) = \sum_x \pi(x)\ \frac{1}{n}\sum_{m=1}^{n} P^m(x,y) \\
\Rightarrow \ \pi(y) &= \lim_{n\to\infty} \sum_x \pi(x)\ \frac{1}{n}\sum_{m=1}^{n} P^m(x,y)
\end{aligned}$$

- The proof is complete if we can take the limit inside the sum

- **Bounded Convergence Theorem**: Suppose $a(x) \geq 0, \ \forall x \in S$ and $\sum_x a(x) < \infty$. Let $b_n(x), \ x \in S$ be such that $|b_n(x)| \leq K, \ \forall x, n$ and suppose $\lim_{n\to\infty} b_n(x) = b(x), \forall x \in S$. Then

$$\lim_{n\to\infty} \sum_{x\in S} a(x)\ b_n(x) = \sum_{x\in S} a(x)\ \lim_{n\to\infty} b_n(x) = \sum_{x\in S} a(x)\ b(x)$$

- We had

$$\pi(y) = \lim_{n\to\infty} \sum_x \pi(x)\ \frac{1}{n}\sum_{m=1}^{n} P^m(x,y)$$

- We have

$$\pi(x) \geq 0; \ \ \sum_x \pi(x) = 1; \ \ 0 \leq \frac{1}{n}\sum_{m=1}^{n} P^m(x,y) \leq 1, \forall x$$

- Hence, if y is transient or null recurrent, then

$$\pi(y) = \sum_x \pi(x)\ \lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} P^m(x,y) = 0$$

- In any stationary distribution π, we would have $\pi(y) = 0$ is y is transient or null recurrent.
- Hence an irreducible transient or null recurrent chain would not have a stationary distribution.

- **Theorem** An irreducible positive recurrent chain has a unique stationary distribution given by

$$\pi(y) = \frac{1}{m_y}, \ \forall y \in S$$

- Suppose $\exists\ \pi$ such that $\pi(y) = \sum_x \pi(x) P(x,y)$. Then

$$\begin{aligned}
\pi(y) &= \sum_x \pi(x)\ P^m(x,y), \ \forall m \\
\Rightarrow \ \pi(y) &= \sum_x \pi(x)\ \frac{1}{n}\sum_{m=1}^{n} P^m(x,y), \ \forall n \\
\Rightarrow \ \pi(y) &= \lim_{n\to\infty} \sum_x \pi(x)\ \frac{1}{n}\sum_{m=1}^{n} P^m(x,y) \\
\Rightarrow \ \pi(y) &= \sum_x \pi(x)\ \lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} P^m(x,y) \\
&= \sum_x \pi(x)\ \frac{1}{m_y} = \frac{1}{m_y}
\end{aligned}$$

- To complete the proof, we need to show $\sum_y \frac{1}{m_y} = 1$.
- We also need to show $\frac{1}{m_y} = \sum_x \frac{1}{m_x} P(x, y)$
- We skip these steps in the proof.
- The theorem shows that an irreducible positive recurrent chain has a unique stationary distribution
- Corollary: An irreducible chain has a stationary distribution if and only if it is positive recurrent
- An irreducible finite chain has a unique stationary distribution

- If π^1 and π^2 are stationary distributions, then so is $\alpha\pi^1 + (1-\alpha)\pi^2$ (easily verified)
- Let C be a closed irreducible set of positive recurrent states.
  Then there is a unique stationary distribution π that satisfies $\pi(y) = 0, \ \forall y \notin C$.
- Any other stationary distribution of the chain is a convex combination of the stationary distributions concentrated on each of the closed irreducible sets of positive recurrent states.
- This answers all questions about existence and uniqueness of stationary distributions

- Consider an irreducible positive recurrent chain.
- It has a unique stationary distribution and $\frac{1}{n}\sum_{m=1}^{n} P^m(x, y)$ converges to $\pi(y)$.
- The next question is convergence of π_n

$$\lim_{n\to\infty} \pi_n(y) = \lim_{n\to\infty} \sum_x \pi_0(x) \ P^n(x, y) = ?$$

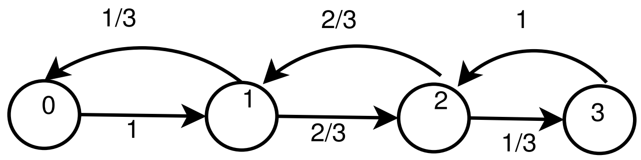
- If $P^n(x, y)$ converges to $g(y)$ then that would be the stationary distribution and π_n converges to it
- But, $\frac{1}{n}\sum_{m=1}^{n} a_m$ may have a limit though $\lim_{n\to\infty} a_n$ may not exist.
  For example, $a_n = (-1)^n$

- Consider a chain with transition probabilities

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 2/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

- One can show $\pi^T = \left[\frac{1}{8} \ \frac{3}{8} \ \frac{3}{8} \ \frac{1}{8}\right]$
- However, P^n goes to different limits based on whether n is even or odd

- The chain is the folowing



$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 2/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

- We can return to a state only after even number of time steps
- That is why P^n does not go to a limit
- Such a chain is called a periodic chain

- We define period of a state x as

$$d_x = \gcd\{n \geq 1 \ : \ P^n(x,x) > 0\}$$

- If $P(x,x) > 0$ then $d_x = 1$
- If x leads to y and y leads to x, then $d_x = d_y$
- Let $P^{n_1}(x,y) > 0$, $P^{n_2}(y,x) > 0$. Then $P^{n_1+n_2}(x,x) > 0 \ \Rightarrow \ d_x$ divides $n_1 + n_2$.
- For any n s.t. $P^n(y,y) > 0$, we get $p^{n_1+n+n_2}(x,x) > 0$
- Hence, d_x divides n for all n s.t. $P^n(y,y) > 0 \Rightarrow d_x \leq d_y$
- Similarly, $d_y \leq d_x$ and hence $d_y = d_x$
- All states in an irreducible chain have the same period.
- If the period is 1 then chain is called aperiodic

- The extra condition we need for convergence of π_n is aperiodicity
- For an aperiodic, irreducible, positive recurrent chain, there is a unique stationary distribution and π_n converges to it irrespective of what π_0 is.
- An aperiodic, irreducible, positive recurrent chain is called an ergodic chain

Recap: Stationary Distribution

- π is said to be a stationary distribution for the Markov chain with transition probabilities P if

$$\pi(y) = \sum_{x \in S} \pi(x) P(x,y), \ \ \forall y \in S$$

- When π is stationary distribution, $\pi_0 = \pi \ \ \Rightarrow \ \ \pi_n = \pi, \ \forall n$
- If $\pi_n = \pi, \ \forall n$ then π is a stationary distribution
- For a finite chain: $P^T \pi = \pi$
- A stationary distribution always exists for a finite chain

Recap

- $N_n(y)$ – number of visits to y till n
- $G_n(x, y) = E_x[N_n(y)] = \sum_{m=1}^{n} P^m(x, y)$
  – expected number of visits to y till n
- $m_y = E_y[T_y]$ – mean return time to y

$$\lim_{n\to\infty} \frac{N_n(y)}{n} = \frac{I_{[T_y<\infty]}}{m_y}, \quad w.p.1$$

$$\lim_{n\to\infty} \frac{G_n(x, y)}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} P^m(x, y) = \frac{\rho_{xy}}{m_y}$$

Recap: positive and null recurrent states

- y is positive recurrent if $m_y < \infty$
- y is null recurrent if $m_y = \infty$
- If x is positive recurrent and x leads to y, then y is positive recurrent
- In a closed irreducible set of recurrent states either all states are positive recurrent or all states are null recurrent
- A finite closed set has to have at least one positive recurrent state
- A finite chain cannot have null recurrent states

Recap: Existence of stationary distribution

- In any stationary distribution π, $\pi(y) = 0$ if y is transient or null recurrent
- An irreducible transient or null recurrent chain does not have a stationary distribution
- An irreducible positive recurrent chain has a unique stationary distribution: $\pi(y) = \frac{1}{m_y}$
- An irreducible chain has a stationary distribution iff it is positive recurrent
- For a non-irreducible chain, for each closed irreducible set of positive recurrent states, there is a unique stationary distribution concentrated on that set.
- All stationary distributions of the chain are convex combinations of these

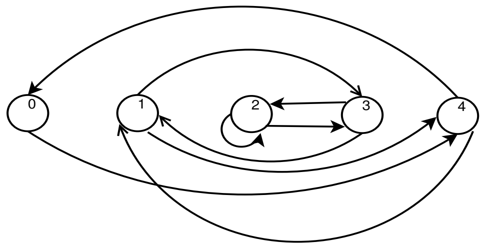
Recap: Periodic chains

- The period of a state x is $d_x = \gcd\{n \geq 1 \ : \ P^n(x, x) > 0\}$
- If x and y lead to each other, $d_x = d_y$
- In an irreducible chain, all states have the same period
- An irreducible chain is called aperiodic if the period is 1
- For an irreducible aperiodic positive recurrent chain, π_n converges to π, the unique stationary distribution, irrespective of what π_0 is.
- Also, for an irreducible, aperiodic, positive recurrent chain, $P^n(x, y)$ converges to $\frac{1}{m_y}$

Example

- Consider the umbrella problem

$$P = \begin{array}{cc} & \begin{matrix} 0 & 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1-p & p \\ 0 & 0 & 1-p & p & 0 \\ 0 & 1-p & p & 0 & 0 \\ 1-p & p & 0 & 0 & 0 \end{bmatrix} \end{array}$$



- This is an irreducible, aperiodic positive recurrent chain

- We want calculate the probability of getting caught in a rain without an umbrella.
- This would be the steady state probability of state 0 multiplied by p
- We are using the fact that this chain converges to the stationary distribution starting with any initial probabilities.

$$P = \begin{array}{cc} & \begin{matrix} 0 & 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1-p & p \\ 0 & 0 & 1-p & p & 0 \\ 0 & 1-p & p & 0 & 0 \\ 1-p & p & 0 & 0 & 0 \end{bmatrix} \end{array}$$

The stationary distribution satisfies $\pi^T P = \pi^T$

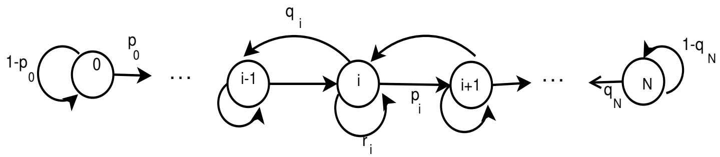
$$\begin{aligned} \pi(0) &= (1-p)\pi(4) \\ \pi(1) &= (1-p)\pi(3) + p\pi(4) \quad \Rightarrow \quad \pi(3) = \pi(1) \\ \pi(2) &= (1-p)\pi(2) + p\pi(3) \\ \pi(3) &= (1-p)\pi(1) + p\pi(2) \quad \Rightarrow \quad \pi(2) = \pi(1) \\ \pi(4) &= \pi(0) + p\pi(1) \quad \Rightarrow \quad \pi(4) = \pi(1) \end{aligned}$$

This gives $4\pi(1) + (1-p)\pi(1) = 1$ and hence

$$\pi(i) = \frac{1}{5-p} \;\; i = 1,2,3,4 \quad \text{and} \quad \pi(0) = \frac{1-p}{5-p}$$

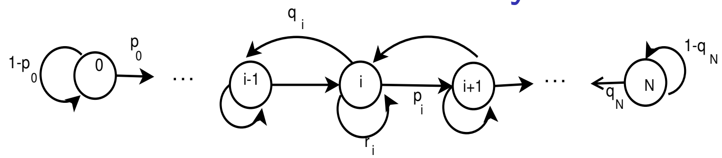
Birth-Death chains

- The following is a finite birth-death chain



- We assume $p_i, q_i > 0, \forall i$.
- Then the chain is irreducible, positive recurrent
- If we assume $r_i > 0$ at least for one i, it is also aperiodic
- We can derive a general form for its stationary probabilities

birth-death chains – stationary distribution



$$\pi(y) = \sum_x \pi(x) P(x, y)$$

$$\begin{aligned}
\pi(0) &= \pi(0)(1 - p_0) + \pi(1) q_1 \\
&\quad \Rightarrow \ \pi(1) q_1 - \pi(0) p_0 = 0 \\
\pi(1) &= \pi(0) p_0 + \pi(1)(1 - p_1 - q_1) + \pi(2) q_2 \\
&\quad \Rightarrow \pi(1) q_1 - \pi(0) p_0 = \pi(2) q_2 - \pi(1) p_1 \\
&\quad \Rightarrow \pi(2) q_2 - \pi(1) p_1 = 0 \\
\pi(2) &= \pi(1) p_1 + \pi(2)(1 - p_2 - q_2) + \pi(3) q_3 \\
&\quad \Rightarrow \pi(2) q_2 - \pi(1) p_1 = \pi(3) q_3 - \pi(2) p_2 = 0
\end{aligned}$$

- Thus we get

$$\begin{aligned}
\pi(1) q_1 - \pi(0) p_0 &= 0 \ \Rightarrow \ \pi(1) = \frac{p_0}{q_1} \pi(0) \\
\pi(2) q_2 - \pi(1) p_1 &= 0 \ \Rightarrow \ \pi(2) = \frac{p_1}{q_2} \pi(1) = \frac{p_0 p_1}{q_1 q_2} \pi(0)
\end{aligned}$$

- Iterating like this, we get

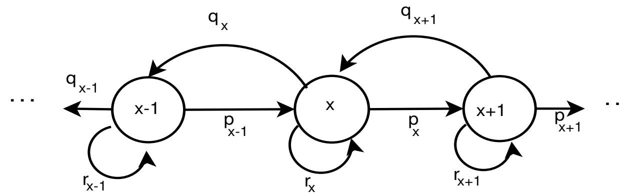
$$\pi(n) = \eta_n \, \pi(0), \quad \text{where} \quad \eta_n = \frac{p_0 p_1 \cdots p_{n-1}}{q_1 q_2 \cdots q_n}, \ n = 1, 2, \cdots, N$$

- With $\eta_0 = 1$, we get $\pi(0) \sum_{j=0}^{N} \eta_j = 1$ and hence

$$\pi(0) = \frac{1}{\sum_{j=0}^{N} \eta_j} \quad \text{and} \quad \pi(n) = \eta_n \, \pi(0), \ n = 1, \cdots, N$$

- Note that this process is applicable even for infinite chains with state space $\{0, 1, 2, \cdots\}$ (but there may not be a solution)

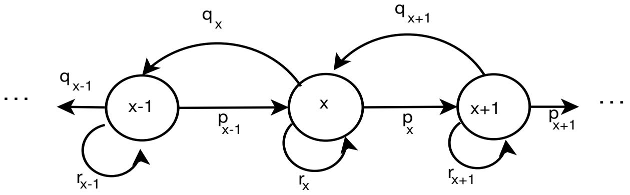
- Consider a birth-death chain



- The chain may be infinite or finite
- Let $a, b \in S$ with $a < b$. Assume $p_x, q_x > 0$, $a < x < b$.
- Define

$$U(x) = P_x[T_a < T_b], \ a < x < b, \quad U(a) = 1, \ U(b) = 0$$

- We want to derive a formula for $U(x)$
- This can be useful, e.g., in the gambler's ruin chain



$$
\begin{aligned}
U(x) &= P_x[T_a < T_b] = Pr[T_a < T_b | X_0 = x] \\
&= \sum_{y=x-1}^{x+1} Pr[T_a < T_b | X_1 = y] \; Pr[X_1 = y | X_0 = x] \\
&= U(x-1)q_x + U(x)r_x + U(x+1)p_x \\
&= U(x-1)q_x + U(x)(1 - p_x - q_x) + U(x+1)p_x
\end{aligned}
$$

$$\Rightarrow \qquad q_x[U(x) - U(x-1)] = p_x[U(x+1) - U(x)]$$

$$\Rightarrow \qquad U(x+1) - U(x) = \frac{q_x}{p_x}\,[U(x) - U(x-1)]$$

$$
\begin{aligned}
U(x+1) - U(x) &= \frac{q_x}{p_x}\,[U(x) - U(x-1)] \\
&= \frac{q_x}{p_x}\,\frac{q_{x-1}}{p_{x-1}}\,[U(x-1) - U(x-2)] \\
&= \frac{q_x q_{x-1} \cdots q_{a+1}}{p_x p_{x-1} \cdots p_{a+1}}\,[U(a+1) - U(a)]
\end{aligned}
$$

Let $\gamma_y = \dfrac{q_y q_{y-1} \cdots q_{a+1}}{p_y p_{y-1} \cdots p_{a+1}}, \;\; a < y < b, \;\; \gamma_a = 1$

Now we get

$$U(x+1) - U(x) = \frac{\gamma_x}{\gamma_a}[U(a+1) - U(a)]$$

$$U(x+1) - U(x) = \frac{\gamma_x}{\gamma_a}\,[U(a+1) - U(a)]$$

- By taking $x = b-1, \; b-2, \cdots, a+1, \; a,$

$$
\begin{aligned}
U(b) - U(b-1) &= \frac{\gamma_{b-1}}{\gamma_a}\,[U(a+1) - U(a)] \\
U(b-1) - U(b-2) &= \frac{\gamma_{b-2}}{\gamma_a}\,[U(a+1) - U(a)] \\
&\vdots \\
U(a+1) - U(a) &= \frac{\gamma_a}{\gamma_a}\,[U(a+1) - U(a)]
\end{aligned}
$$

- Adding all these we get

$$\frac{1}{\gamma_a}\,[U(a+1) - U(a)] \sum_{x=a}^{b-1} \gamma_x = U(b) - U(a) = 0 - 1$$

$$\Rightarrow \;\; U(a) - U(a+1) = \frac{\gamma_a}{\sum_{x=a}^{b-1} \gamma_x}$$

- Using these, we get

$$
\begin{aligned}
U(x) - U(x+1) &= \frac{\gamma_x}{\gamma_a}[U(a) - U(a+1)] \\
&= \frac{\gamma_x}{\gamma_a}\,\frac{\gamma_a}{\sum_{x=a}^{b-1} \gamma_x} = \frac{\gamma_x}{\sum_{x=a}^{b-1} \gamma_x}
\end{aligned}
$$

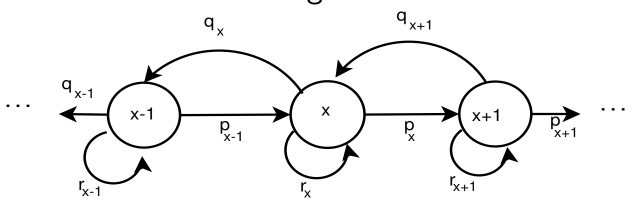
- Putting $x = b-1, \; b-2, \cdots, y$ in the above

$$
\begin{aligned}
U(b-1) - U(b) &= \frac{\gamma_{b-1}}{\sum_{x=a}^{b-1} \gamma_x} \\
U(b-2) - U(b-1) &= \frac{\gamma_{b-2}}{\sum_{x=a}^{b-1} \gamma_x} \\
&\vdots \\
U(y) - U(y+1) &= \frac{\gamma_y}{\sum_{x=a}^{b-1} \gamma_x}
\end{aligned}
$$

- Adding these we get

$$U(y) - U(b) = U(y) = \frac{\sum_{x=y}^{b-1} \gamma_x}{\sum_{x=a}^{b-1} \gamma_x}, \;\; a < y < b$$

- We are considering birth-death chains



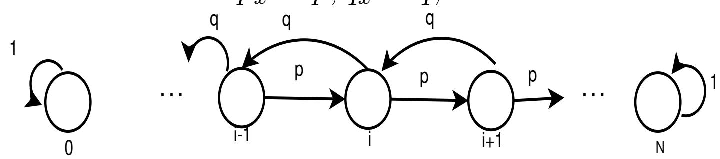
- We have derived, for $a < y < b$,

$$U(y) = P_y[T_a < T_b] = \frac{\sum_{x=y}^{b-1} \gamma_x}{\sum_{x=a}^{b-1} \gamma_x}, \quad \gamma_x = \frac{q_x q_{x-1} \cdots q_{a+1}}{p_x p_{x-1} \cdots p_{a+1}}$$

- Hence we also get

$$P_y[T_b < T_a] = \frac{\sum_{x=a}^{y-1} \gamma_x}{\sum_{x=a}^{b-1} \gamma_x}$$

- Suppose this is a Gambler's ruin chain:
$$p_x = p, q_x = q, \ \forall x$$

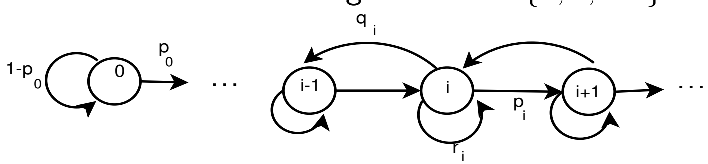


- Then, $\gamma_x = \left(\frac{q}{p}\right)^x$
- Hence, for a Gambler's ruin chain we get, e.g.,

$$P_i[T_N < T_0] = \frac{\sum_{x=0}^{i-1} \gamma_x}{\sum_{x=0}^{N-1} \gamma_x} = \frac{\left(\frac{q}{p}\right)^i - 1}{\left(\frac{q}{p}\right)^N - 1}$$

- This is the probability of gambler being successful

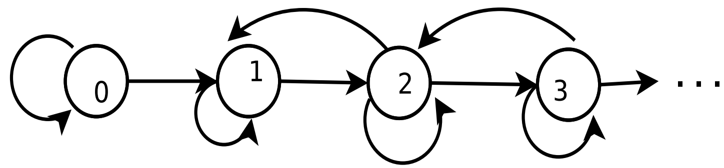
- Consider the following chain over $\{0, 1, \cdots\}$



- This is an infinite irreducible birth-death chain
- We want to know whether the chain is transient or recurrent etc.
- We can use the earlier analysis for this too.

$$\begin{aligned} P_1[T_0 < T_n] &= \frac{\sum_{x=1}^{n-1} \gamma_x}{\sum_{x=0}^{n-1} \gamma_x}, \ \forall n > 1 \\ &= \frac{\sum_{x=0}^{n-1} \gamma_x - \gamma_0}{\sum_{x=0}^{n-1} \gamma_x} = 1 - \frac{1}{\sum_{x=0}^{n-1} \gamma_x} \end{aligned}$$

- Consider this chain started in state 1.



$$[T_0 < T_n] \subset [T_0 < T_{n+1}], \quad n = 2, 3, \cdots$$

since the chain cannot hit $n+1$ without hitting n.

- Also, $1 \leq T_2 < T_3 < \cdots < T_n$ and $T_n \geq n$.
- Hence $[T_0 < \infty]$ is same as $[T_0 < T_n, \ \text{ for some } n]$

- Consider this chain started in state 1.

$$[T_0 < T_n] \subset [T_0 < T_{n+1}], \quad n = 2, 3, \cdots$$

since the chain cannot hit $n+1$ without hitting n.
- Also, $1 \leq T_2 < T_3 < \cdots < T_n$ and $T_n \geq n$.
- Hence $[T_0 < \infty]$ is same as $[T_0 < T_n, \text{ for some } n]$

$$\begin{aligned} P_1[T_0 < T_n, \text{ for some } n] &= P_1\left(\cup_{n>1} [T_0 < T_n]\right) \\ &= P_1\left(\lim_{n\to\infty} [T_0 < T_n]\right) \\ &= \lim_{n\to\infty} P_1\left([T_0 < T_n]\right) \\ &= \lim_{n\to\infty} 1 - \frac{1}{\sum_{x=0}^{n-1} \gamma_x} \\ \Rightarrow \quad P_1[T_0 < \infty] &= 1 - \frac{1}{\sum_{x=0}^{\infty} \gamma_x} \end{aligned}$$

- **Theorem**: The chain is recurrent iff $\sum_{x=0}^{\infty} \gamma_x = \infty$

Proof

- Supoose chain is recurrent. Since it is irreducible,

$$P_1[T_0 < \infty] = 1 \quad \Rightarrow \quad \sum_{x=0}^{\infty} \gamma_x = \infty$$

- Suppose $\sum_{x=0}^{\infty} \gamma_x = \infty \Rightarrow P_1[T_0 < \infty] = 1$

$$\begin{aligned} P_0[T_0 < \infty] &= P(0,0) \ + \ P(0,1)\ P_1[T_0 < \infty] \\ &= P(0,0) + P(0,1) = 1 \end{aligned}$$

- Implies state 0 is recurrent and hence the chain is recurrent because it is irreducible.
- Note that we have used the fact that the chain is infinite only to the right.

- The chain is transient if $\sum_{x=0}^{\infty} \gamma_x < \infty$
- Let $p_x = p, q_x = q \Rightarrow \quad \gamma_x = \left(\frac{q}{p}\right)^x$

$$\text{Transient if } \sum_{x=0}^{\infty} \left(\frac{q}{p}\right)^x < \infty \quad \Leftrightarrow \quad q < p$$

$$\text{Recurrent if } \sum_{x=0}^{\infty} \left(\frac{q}{p}\right)^x = \infty \quad \Leftrightarrow \quad q \geq p$$

- Intuitively clear
- This chain with $q < p$ is an example of an irreducible chain that is wholly transient

- We know the chain is recurrent if $\sum_{x=0}^{\infty} \left(\frac{q}{p}\right)^x = \infty$
- When will this chain be positive recurrent?
- We know that an irreducible chain is positive recurrent if and only if it has a stationary distribution.
- We can check if it has a stationary distribution
- The earlier equations that we derived earlier hold for this infinite case also.

- We derived earlier the equations that a stationary distribution of this chain (if it exists) has to satisfy

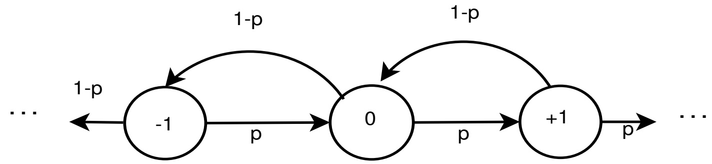
$$\pi(n) = \eta_n \, \pi(0), \quad \text{where} \quad \eta_n = \frac{p_0 p_1 \cdots p_{n-1}}{q_1 q_2 \cdots q_n}, \; n = 1, 2, \cdots,$$

- Setting $\eta_0 = 1$, we get $\pi(0) \sum_{j=0}^{\infty} \eta_j = 1$
- Hence stationary distribution exists iff $\sum_{j=0}^{\infty} \eta_j < \infty$
- Let $p_x = p, q_x = q$

$$\sum_{j=0}^{\infty} \eta_j = \sum_{j=0}^{\infty} \left(\frac{p}{q}\right)^j < \infty \quad \Leftrightarrow \quad p < q$$

- Thus in this special case, the chain is
  - transient if $p > q$; recurrent if $p \le q$
  - positive recurrent if $p < q$
  - null recurrent if $p = q$

- This analysis can handle chains which are infinite in one direction
- Consider the following random walk chain



- The state space here is $\{\cdots, -1, 0, +1, \cdots\}$
- The chain is irreducible and periodic with period 2
- $P^{2n}(0,0) = {}^{2n}C_n p^n (1-p)^n$.
- We can look at the limit of $\frac{1}{n} \sum_n P^{2n}(0,0)$
- We can show that the chain is transient if $p \neq 0.5$ and is recurrent if $p = 0.5$.

- In general, determining when an infinite chain is positive recurrent is difficult.
- The method we had works only for birth-death chains over non-negative integers.
- There is a useful general theorem.

  Foster's Theorem
  Let P be the transition probabilities of a homogeneous irreducible Markov chain with state space S. Let $h : S \to \Re$ with $h(x) \ge 0$ and
  - $\sum_{k \in S} P(i,k) h(k) < \infty \; \forall i \in F$ and
  - $\sum_{k \in S} P(i,k) h(k) \le h(i) - \epsilon \; \forall i \notin F$

  for some finite set F and some $\epsilon > 0$. Then the Markov chain is positive recurrent
- The h here is called a Lyapunov function.
- We will not prove this theorem

- Let $\{X_n, \; n \ge 0\}$ be an irreducible markov chain on a finite state space S with stationary distribution π.
- Let $r : S \to \Re$ be a bounded function.
- Suppose we want $E[r(X)]$ with respect to the stationary distribution π ($E[r(X)] = \sum_{j \in S} r(j) \pi(j)$)
- Let $N_n(j)$ be as earlier. Then

$$\frac{1}{n} \sum_{m=1}^{n} r(X_m) = \frac{1}{n} \sum_{j \in S} N_n(j) r(j)$$

$$\Rightarrow \lim_{n \to \infty} \frac{1}{n} \sum_{m=1}^{n} r(X_m) = \sum_{j \in S} r(j) \lim_{n \to \infty} \frac{N_n(j)}{n} = \sum_{j \in S} r(j) \pi(j)$$

- For this to be true for infinite S, we need some extra conditions

MCMC Sampling

- Consider a distribution over (finite) S: $\pi(x) = \frac{b(x)}{Z}$
- Since this is a distribution, $Z = \sum_{x \in S} b(x)$
- We assume, we can efficiently calculate $b(x)$ for any x but computation of Z is intractable or computationally expensive
  E.g., the Boltzmann distribution: $b(x) = e^{-E(x)/KT}$
- We want $E[g(X)]$ w.r.t. distribution π (for any g)

$$E[g(X)] = \sum_x g(x)\ \pi(x) \approx \frac{1}{n}\sum_{i=1}^{n} g(X_i), \quad X_1, \cdots X_n \sim \pi$$

- One way to generate samples is to design an ergodic markov chain with stationary distribution π
  – MCMC sampling

- Suppose $\{X_n\}$ is a an irreducible, aperiodic positive recurrent Markov chain with stationary dist $\pi(x) = \frac{b(x)}{Z}$
- Then we have

$$\lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} g(X_m) = \sum_x g(x)\pi(x)$$

- hence, if we can design a Markov chain with a given stationary distribution, we can use that to calculate the expectation.
- We can also use the chain to generate samples from distribution π

- $\{X_n\}$: Markov chain with stationary dist $\pi(x) = \frac{b(x)}{Z}$
  We can approximate the expectation as

$$\sum_x g(x)\pi(x) \approx \frac{1}{n}\sum_{i=1}^{n} g(X_{M+i})$$

  Where M is large enough to assume chain is in steady state
- When we take sample mean, $\frac{1}{n}\sum_{i=1}^{n} Z_i$, we want Z_i to be uncorrelated
- We can, for example, use

$$\sum_x g(x)\pi(x) \approx \frac{1}{n}\sum_{i=1}^{n} g(X_{M+Ki})$$

- For all these, we need to design a Markov chain with π as stationary distribution

- Let $Q = [q(i,j)]$ be the transition probability matrix of an irreducible Markov chain over S.
- Q is called the proposal distribution
- We start with arbitray X_0 and generate $X_{n+1},\ n = 0, 1, 2, \cdots$, iteratively as follows
  - If $X_n = i$, we generate Y with $Pr[Y = k] = q(i,k)$
  - Let the generated value for Y be j. Set

$$X_{n+1} = \begin{cases} j & \text{with probability } \ \alpha(i,j) \\ X_n & \text{with probability } \ 1 - \alpha(i,j) \end{cases}$$

- $\alpha(i,j)$ is called the acceptance probability
- We want to choose $\alpha(i,j)$ to make X_n an ergodic markov chain with stationary probabilities π

- The stationary distribution π satisfies (with transition probabilities P)

$$\pi(y) = \sum_x \pi(x)\ P(x, y), \ \ \forall y \in S$$

- Suppose there is a distribution $g(\cdot)$ that satisfies

$$g(y)\ P(y, x) = g(x)\ P(x, y), \ \ \forall x, y \in S$$

This is called detailed balance

- Summing both sides above over x give

$$g(y) = \sum_x g(y)\ P(y, x) = \sum_x g(x) P(x, y), \ \ \forall y$$

- Thus if $g(\cdot)$ satisfies detailed balance, then it must be the stationary distribution
- Note that it is not necessary for a stationary distribution to satisfy detailed balance

- Any stationary distribution has to satisfy

$$\pi(y) = \sum_x \pi(x)\ P(x, y), \ \ \forall y \in S$$

- If I can find a π that satisfies

$$\pi(x) P(x, y) = \pi(y) P(y, x), \ \ \forall x, y \in S, \ x \neq y$$

that would be the stationary distribution

- This is called detailed balance

- Recall our algorithm for generating $X_n, \ n = 0, 1, \cdots$
- Start with arbitrary X_0 and generate X_{n+1} from X_n
  - If $X_n = i$, we generate Y with $Pr[Y = k] = q(i, k)$
  - Let the generated value for Y be j. Set

$$X_{n+1} = \begin{cases} j & \text{with probability } \ \alpha(i, j) \\ X_n & \text{with probability } \ 1 - \alpha(i, j) \end{cases}$$

- Hence the transition probabilities for X_n are

$$\begin{aligned} P(i, j) &= q(i, j)\ \alpha(i, j), \quad i \neq j \\ P(i, i) &= q(i, i) + \sum_{j \neq i} q(i, j)\ (1 - \alpha(i, j)) \end{aligned}$$

- $\pi(i) = b(i)/Z$ is the desired stationary distribution
- So, we can try to satisfy

$$\begin{aligned} \pi(i)\ P(i, j) &= \pi(j)\ P(j, i), \ \ \forall i, j, i \neq j \\ \text{that is,} \quad b(i) q(i, j)\ \alpha(i, j) &= b(j) q(j, i)\ \alpha(j, i) \end{aligned}$$

MCMC Sampling

- Consider a distribution over (finite) S: $\pi(x) = \frac{b(x)}{Z}$
- Since this is a distribution, $Z = \sum_{x \in S} b(x)$
- We assume, we can efficiently calculate $b(x)$ for any x but computation of Z is intractable or computationally expensive
  E.g., the Boltzmann distribution: $b(x) = e^{-E(x)/KT}$
- We want $E[g(X)]$ w.r.t. distribution π (for any g)

$$E[g(X)] = \sum_x g(x)\ \pi(x) \approx \frac{1}{n} \sum_{i=1}^n g(X_i), \quad X_1, \cdots X_n \sim \pi$$

- One way to generate samples is to design an ergodic markov chain with stationary distribution π
  – MCMC sampling

- Suppose $\{X_n\}$ is a an irreducible, aperiodic positive recurrent Markov chain with stationary dist $\pi(x) = \frac{b(x)}{Z}$
- Then we have

$$\lim_{n \to \infty} \frac{1}{n} \sum_{m=1}^{n} g(X_m) = \sum_x g(x)\pi(x)$$

- hence, if we can design a Markov chain with a given stationary distribution, we can use that to calculate the expectation.
- We can also use the chain to generate samples from distribution π

- $\{X_n\}$: Markov chain with stationary dist $\pi(x) = \frac{b(x)}{Z}$
  We can approximate the expectation as

$$\sum_x g(x)\pi(x) \approx \frac{1}{n} \sum_{i=1}^{n} g(X_{M+i})$$

  Where M is large enough to assume chain is in steady state
- When we take sample mean, $\frac{1}{n}\sum_{i=1}^{n} Z_i$, we want Z_i to be uncorrelated
- We can, for example, use

$$\sum_x g(x)\pi(x) \approx \frac{1}{n} \sum_{i=1}^{n} g(X_{M+Ki})$$

- For all these, we need to design a Markov chain with π as stationary distribution

- Let $Q = [q(i,j)]$ be the transition probability matrix of an irreducible Markov chain over S.
- Q is called the proposal distribution
- We start with arbitrary X_0 and generate $X_{n+1}, \; n = 0, 1, 2, \cdots$, iteratively as follows
  - If $X_n = i$, we generate Y with $Pr[Y = k] = q(i,k)$
  - Let the generated value for Y be j. Set

$$X_{n+1} = \begin{cases} j & \text{with probability } \; \alpha(i,j) \\ X_n & \text{with probability } \; 1 - \alpha(i,j) \end{cases}$$

- $\alpha(i,j)$ is called the acceptance probability
- We want to choose $\alpha(i,j)$ to make X_n an ergodic Markov chain with stationary probabilities π

- The stationary distribution π satisfies (with transition probabilities P)

$$\pi(y) = \sum_x \pi(x) \; P(x,y), \quad \forall y \in S$$

- Suppose there is a distribution $g(\cdot)$ that satisfies

$$g(y) \; P(y,x) = g(x) \; P(x,y), \quad \forall x, y \in S$$

  This is called detailed balance
- Summing both sides above over x give

$$g(y) = \sum_x g(y) \; P(y,x) = \sum_x g(x) P(x,y), \quad \forall y$$

- Thus if $g(\cdot)$ satisfies detailed balance, then it must be the stationary distribution
- Note that it is not necessary for a stationary distribution to satisfy detailed balance

- Any stationary distribution has to satisfy

$$\pi(y) = \sum_x \pi(x)\ P(x,y), \quad \forall y \in S$$

- If I can find a π that satisfies

$$\pi(x)P(x,y) = \pi(y)P(y,x), \quad \forall x, y \in S, \ x \neq y$$

that would be the stationary distribution
- This is called detailed balance

- Recall our algorithm for generating $X_n, \ n = 0, 1, \cdots$
- Start with arbitrary X_0 and generate X_{n+1} from X_n
  - If $X_n = i$, we generate Y with $Pr[Y = k] = q(i,k)$
  - Let the generated value for Y be j. Set

$$X_{n+1} = \begin{cases} j & \text{with probability } \ \alpha(i,j) \\ X_n & \text{with probability } \ 1 - \alpha(i,j) \end{cases}$$

- Hence the transition probabilities for X_n are

$$\begin{aligned} P(i,j) &= q(i,j)\ \alpha(i,j), \quad i \neq j \\ P(i,i) &= q(i,i) + \sum_{j \neq i} q(i,j)\ (1 - \alpha(i,j)) \end{aligned}$$

- $\pi(i) = b(i)/Z$ is the desired stationary distribution
- So, we can try to satisfy

$$\begin{aligned} \pi(i)\ P(i,j) &= \pi(j)\ P(j,i), \quad \forall i, j, i \neq j \\ \text{that is,} \quad b(i)q(i,j)\ \alpha(i,j) &= b(j)q(j,i)\ \alpha(j,i) \end{aligned}$$

- We want to satisfy

$$b(i)q(i,j)\ \alpha(i,j) = b(j)q(j,i)\ \alpha(j,i)$$

- Choose

$$\alpha(i,j) = \min\left(\frac{\pi(j)q(j,i)}{\pi(i)q(i,j)}, 1\right) = \min\left(\frac{b(j)q(j,i)}{b(i)q(i,j)}, 1\right)$$

- Note that one of $\alpha(i,j), \ \alpha(j,i)$ is 1

$$\begin{aligned} \text{suppose} \quad \alpha(i,j) &= \frac{\pi(j)q(j,i)}{\pi(i)q(i,j)} < 1 \\ \Rightarrow \quad \pi(i)\ q(i,j)\ \alpha(i,j) &= \pi(j)\ q(j,i) \\ &= \pi(j)\ q(j,i)\ \alpha(j,i) \end{aligned}$$

- Note that $\pi(i)$ above can be replaced by $b(i)$

Metropolis-Hastings Algorithm

- Start with arbitrary X_0 and generate X_{n+1} from X_n
  - If $X_n = i$, we generate Y with $Pr[Y = k] = q(i,k)$
  - Let the generated value for Y be j. Set

$$X_{n+1} = \begin{cases} j & \text{with probability } \ \alpha(i,j) \\ X_n & \text{with probability } \ 1 - \alpha(i,j) \end{cases}$$

  Where $Q = [q(i,j)]$ is the transition probabilities of an irreducible chain and

$$\alpha(i,j) = \min\left(\frac{\pi(j)q(j,i)}{\pi(i)q(i,j)}, 1\right)$$

- Then $\{X_n\}$ would be an irreducible, aperiodic chain with stationary distribution π.
- Q is called the proposal chain and $\alpha(i,j)$ is called acceptance probabilities

- Consider Boltzmann distribution: $b(x) = e^{-E(x)/KT}$
- Take proposal to be uniform: from any state, we go to all other states with equal probabilities
- Then,

$$\alpha(x,y) = \min\left(\frac{b(y)}{b(x)}, 1\right) = \min\left(e^{-(E(y)-E(x))/KT}, 1\right)$$

- In state x you generate a random new state y.
  If $E(y) \leq E(x)$ you always go there;
  if $E(y) > E(x)$, accept with probability $e^{-(E(y)-E(x))/KT}$
- An interesting way to simulate Boltzmann distribution
- We could have chosen Q to be 'uniform over neighbours'

- Suppose $E : S \to \Re$ is some function.
- We want to find $x \in S$ where E is *globally* minimized.
- A gradient descent type method tries to find a locally minimizing direction and hence gives only a 'local' minimum.
- The Metropolis-Hastings algorithm gives another view point on how such optimization problems can be handled.
- We can think of E as the energy function in a Boltzmann distribution

- Let $b(x) = e^{-E(x)/T}$ where T is a parameter called 'temparature'
- $\{X_n\}$ be Markov chain with stationary dist $\pi(x) = \frac{b(x)}{Z}$
- We can find relative occupation of different states by the chain by collecting statistics during steady state
- We know

$$\frac{\pi(x_1)}{\pi(x_2)} = \frac{b(x_1)}{b(x_2)} = e^{-(E(x_1)-E(x_2))/T}$$

- We spend more time in global minimum
  We can increase the relative fraction of time spent in global minimum by decreasing T (There is a price to pay!)
- Gives rise to interesting optimization technique called simulated annealing

- In most applications of MCMC, $x \in S$ is a vector.
- One normally changes one component at a time. That is how neighbours can be defined
- A special case of proposal distribution is the conditional distribution.
- Suppose $X = (X_1, \cdots, X_N)$. To propose a value for X_i, we use $f_{X_i|X_{-i}}$
- Here the conditional distribution is calculated using the target π as the joint distribution.
- With such a proposal distribution, one can show that $\alpha(i,j)$ is always 1
- This is known as Gibbs sampling

Random process

- A random process or a stochastic process is a collection of random variables: $\{X_t, \ t \in T\}$
- Markov chain is an example. Here $T = \{0, 1, \cdots\}$
- We call T the index set.
- Normally, T is either (a subset of) set of integers or an interval on real line.
- We think of the index t as time
- Thus a random process can represent the time-evolution of the state of a system
- We assume T is infinite
- The index need not necessarily represent time. It can represent, for example, space coordinates.

- A random process: $\{X_t, \ t \in T\}$
- The set T can be countable e.g., $T = \{0, 1, 2, \cdots\}$
- Or, T can be continuous e.g., $T = [0, \infty)$
- These are termed **discrete-time** or **continuous-time** processes
- The random variables, X_t, may be discrete or continuous
- These are termed **discrete-state** or **continuous-state** processes
- The Markov chain we considered is a discrete-time discrete-state process

- A random process: $\{X_t, \ t \in T\}$
- We can think of this as a mapping: $X : \Omega \times T \to \Re$
- Thus, $X(\omega, \cdot)$ is a real-valued function over T.
- So, we can think of the process also as a collection of time functions.
- X can be thought of as a map that associates with each $\omega \in \Omega$ a real-valued function on T.
- These functions are called sample paths or paths of the process
- We can view the random process as a collection of random variables, or as a collection of functions
- We will denote the random variables as X_t or $X(t)$

- A finite collection of random variables is completely specified by its joint distribution
- How do we characterize a random process?
- We need to specify joint distribution of $X_{t_1}, X_{t_2}, \cdots X_{t_n}$ for all n and all $t_1, t_2, \cdots t_n \in T$..
- One can show this completely specifies the process.
- As we saw, for a Markov chain, π_0 and P together specify all such joint distributions

Distributions of a random process

- A random process: $\{X_t,\ t \in T\}$ or $X : \Omega \times T \to \Re$
- The first order distribution function of X is

$$F_X(x;t) = Pr[X_t \leq x] = F_{X_t}(x)$$

- The second order distribution function of X is

$$F_X(x_1, x_2; t_1, t_2) = Pr[X_{t_1} \leq x_1,\ X_{t_2} \leq x_2]$$

- The n^{th} order distribution function of X is

$$F_X(x_1, \cdots, x_n; t_1, \cdots t_n) = Pr[X_{t_i} \leq x_i,\ i = 1, \cdots, n]$$

- When it is a discrete-state process, all X_t would be discrete random variables
- We can specify distributions through mass functions:

$$f_X(x;t) = Pr[X_t = x] = f_{X_t}(x)$$

$$f_X(x_1, x_2; t_1, t_2) = Pr[X_{t_1} = x_1,\ X_{t_2} = x_2]$$

$$f_X(x_1, \cdots, x_n; t_1, \cdots t_n) = Pr[X_{t_i} = x_i,\ i = 1, \cdots, n]$$

- If all X_t are continuous random variables and if all distributions have density functions, then we denote joint density of $X_{t_1}, \cdots, X_{t_n}$ by $f_X(x_1, \cdots, x_n; t_1, \cdots t_n)$

- Specifying the n^{th} order distributions for all n separately is not feasible.
- Hence one needs some assumptions on the model so that these are specified implicitly.
- One example is the Markovian assumption.
- As we saw, in a Markov chain, the transition probabilities and initial state probabilities would determine all the distributions
- Another such useful assumption is what is called a process with independent increments

- A random process $\{X(t),\ t \in T\}$ is said to be a process with independent increments if

  for all $t_1 < t_2 \leq t_3 < t_4$, *the random variables* $X(t_2) - X(t_1)$ *and* $X(t_4) - X(t_3)$ *are independent*

- Note that this also implies, e.g., $X(t_1)$ is independent of $X(t_2) - X(t_1)$ for all $t_1 < t_2$.
- Now suppose this is a discrete-state process.
- Then we can write n^{th} order pmf's as

$$\begin{aligned} & Pr[X(t_1) = x_1, X(t_2) = x_2, \cdots X(t_n) = x_n] \\ & \quad = Pr[X(t_1) = x_1, X(t_2) - X(t_1) = x_2 - x_1, \cdots] \\ & \quad = Pr[X(t_1) = x_1]\ Pr[X(t_2) - X(t_1) = x_2 - x_1]\ \cdots \\ & \qquad\qquad \cdots Pr[X(t_n) - X(t_{n-1}) = x_n - x_{n-1}] \end{aligned}$$

- We only need up to second order distributions

- Let $\{X(t),\ t \in T\}$ be a discrete-state process with independent increments
- Then we specify $f_X(x;t)$ and another function

$$g(x_1, x_2; t_1, t_2) = Pr[X(t_2) - X(t_1) = x_2 - x_1]$$

- Now we can get all distributions as

$$\begin{aligned} f_X(x_1, \cdots, x_n; t_1, \cdots t_n) \\ &= Pr[X(t_i) = x_i,\ i = 1, \cdots, n] \\ &= f_X(x_1; t_1) \prod_{i=1}^{n-1} Pr[X(t_{i+1}) - X(t_i) = x_{i+1} - x_i] \\ &= f_X(x_1; t_1) \prod_{i=1}^{n-1} g(x_i, x_{i+1}; t_i, t_{i+1}) \end{aligned}$$

- Given a random process $\{X(t),\ t \in T\}$
- Its mean or mean function is defined by

$$\eta_X(t) = E[X(t)],\ \ t \in T$$

- We define the autocorrelation of the process by

$$R_X(t_1, t_2) = E[X(t_1)X(t_2)]$$

- We define the autocovariance of the process by

$$\begin{aligned} C_X(t_1, t_2) &= E\left[(X(t_1) - E[X(t_1)])(X(t_2) - E[X(t_2)])\right] \\ &= R_X(t_1, t_2) - \eta_X(t_1)\eta_X(t_2) \end{aligned}$$

Stationary Processes

- A random process $\{X(t),\ t \in T\}$ is said to be stationary if

  for all n, for all $t_1, \cdots, t_n$, for all $x_1, \cdots x_n$ and for all τ we have

$$F_X(x_1, \cdots, x_n\ ;\ t_1, \cdots, t_n) = F_X(x_1, \cdots, x_n\ ;\ t_1{+}\tau, \cdots, t_n{+}\tau)$$

- For a stationary process, the distributions are unaffected by translation of the time axis.
- This is a rather stringent condition and is often referred to as strict-sense stationarity

- A homogeneous Markov chain started in its stationary distribution is a stationary process
- As we know, if π_0 is the stationary distribution then π_n is same for all n.
- This, along with the Markov condition would imply that shift of time origin does not affect the distributions

$$\begin{aligned} Pr[X_n = x_0, X_{n+1} = x_1, \cdots X_{n+m} = x_m] \\ &= \pi_n(x_0)P(x_0, x_1) \cdots P(x_{m-1}, x_m) \\ &= \pi_0(x_0)P(x_0, x_1) \cdots P(x_{m-1}, x_m) \\ &= Pr[X_0 = x_0, X_1 = x_1, \cdots X_m = x_m] \end{aligned}$$

- Suppose $\{X(t),\ t \in T\}$ is (strict-sense) stationary
- Then the first order distribution is independent of time

$$F_X(x;t) = F_X(x;t+\tau),\ \forall x,t,\tau \quad \Rightarrow \quad \text{e.g.,} \quad F_X(x;t) = F_X(x;0)$$

- This implies $\eta_X(t) = \eta_X$, a constant
- The second order distribution has to satisfy

$$F_X(x_1,x_2;t,t+\tau) = F_X(x_1,x_2;0,\tau),\ \forall x_1,x_2,t,\tau$$

  Hence $F_X(x_1,x_2;t_1,t_2)$ can depend only on $t_1 - t_2$
- This implies

$$R_X(t,t+\tau) = E[X(t)X(t+\tau)] = R_X(\tau)$$

  Autocorrelation depends only on the time difference

- The process $\{X(t),\ t \in T\}$ is said to be wide-sense stationary if

$$\begin{aligned} F_X(x;t) &= F_X(x;t+\tau),\ \forall x,t,\tau \\ F_X(x_1,x_2;t_1,t_2) &= F_X(x_1,x_2;t_1+\tau,t_2+\tau) \end{aligned}$$

- The process is wide-sense stationary if the first and second order distributions are invariant to translation of time origin

- Let $\{X(t),\ t \in T\}$ be wide-sense stationary. Then

1. $\eta_X(t) = \eta_X$, a constant
2. $R_X(t_1,t_2)$ depends only on $t_1 - t_2$

- In many engineering applications, we call a process wide-sense stationary if the above two hold.
- In this course we take the above as the definition of wide-sense stationary process
- When the process is wide-sense stationary, we write autocorrelation as

$$R_X(\tau) = E[X(t)X(t+\tau)]$$

Ergodicity

- Suppose $X(n)$ is a discrete-time discrete-state process (like a Markov chain)
- Suppose it is wide-sense stationary.
  Then $E[X(n)]$ does not depend on n
- Ergodicity is the question of

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X(i) \stackrel{?}{=} E[X(n)] = \eta_X$$

- We proved that this is true for an irreducible, aperiodic, positive recurrent Markov chain (with a finite state space)
- The question is : do ‘time-averages’ converge to ‘ensemble-averages’
- The process is wide-sense stationary and hence all $X(n)$ have the same distribution; but they need not be independent or uncorrelated (e.g., Markov chain)

- Ergodicity is a question of whether time-averages converge to ensemble-averages?

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X(i) \overset{?}{=} E[X(n)] = \eta_X$$

Or, more generally

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} g(X(i)) \overset{?}{=} E[g(X(n))]$$

For a continuous time process we can write this as

$$\lim_{\tau\to\infty} \frac{1}{2\tau}\int_{-\tau}^{\tau} X(t)\ dt \overset{?}{=} E[X(t)] = \eta_X$$

- Essentially if there is no long-term correlation in the process this may hold.
- One sufficient condition could be that covariance between $X(t)$ and $X(t+\tau)$ decreases fast with increasing τ.

- Define

$$\eta_\tau = \frac{1}{2\tau}\int_{-\tau}^{\tau} X(t)\ dt \quad (\tau > 0)$$

- For each τ, η_τ is a rv. We write η for η_X.
- We say the process is mean-ergodic if

$$\eta_\tau \xrightarrow{P} \eta, \quad \text{as} \quad \tau \to \infty$$

- That is, if

$$\lim_{\tau\to\infty} Pr\left[|\eta_\tau - \eta| > \epsilon\right] = 0, \quad \forall \epsilon > 0$$

- Note that $E[\eta_\tau] = \eta, \ \forall \tau$.
- Hence it is enough if we show

$$\sigma_\tau^2 \triangleq E\left[(\eta_\tau - \eta)^2\right] \ \to \ 0, \quad \text{as} \quad \tau \to \infty$$

- Let $C_X(t_1, t_2)$ be the autocovariance of the process

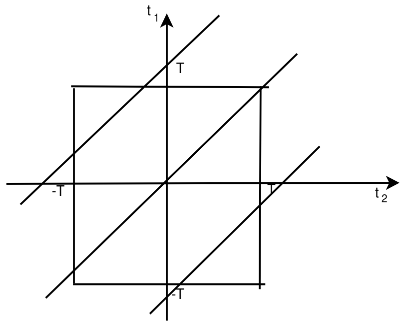
$$C_X(t_1,t_2) = E[(X(t_1)-\eta)(X(t_2)-\eta)]$$

- Assuming wide-sense stationarity, $C_X(t_1,t_2) = C_X(t_1 - t_2)$
- We can get σ_τ^2 as

$$\begin{aligned}
\sigma_\tau^2 &= E\left[(\eta_\tau - \eta)^2\right] \\
&= E\left[\frac{1}{2\tau}\int_{-\tau}^{\tau}(X(t)-\eta)\ dt\ \frac{1}{2\tau}\int_{-\tau}^{\tau}(X(t')-\eta)\ dt'\right] \\
&= \frac{1}{4\tau^2}\int_{-\tau}^{\tau}\int_{-\tau}^{\tau} E[(X(t)-\eta)(X(t')-\eta)]\ dt\ dt' \\
&= \frac{1}{4\tau^2}\int_{-\tau}^{\tau}\int_{-\tau}^{\tau} C_X(t-t')\ dt\ dt'
\end{aligned}$$

$$\text{Let} \quad I = \int_{-\tau}^{\tau}\int_{-\tau}^{\tau} C_X(t_1 - t_2)\ dt_2\ dt_1$$

- Let $z = t_1 - t_2$. We want to change the integration to be over t_2 and z



- Easy to see z goes from -2τ to 2τ
When $z \geq 0$, for a given z, t_2 goes from $-\tau$ to $\tau - z$
When $z < 0$, for a given z, t_2 goes from $-\tau - z$ to τ

- Now we get

$$
\begin{aligned}
I &= \int_{-\tau}^{\tau}\int_{-\tau}^{\tau} C_X(t_1-t_2)\ dt_2\ dt_1 \\
&= \int_{-2\tau}^{0}\int_{-\tau-z}^{\tau} C_X(z)\ dt_2\ dz \ + \ \int_{0}^{2\tau}\int_{-\tau}^{\tau-z} C_X(z)\ dt_2\ dz
\end{aligned}
$$

$$
= \int_{-2\tau}^{0} C_X(z)\ (\tau-(-\tau-z))\ dz \ + \ \int_{0}^{2\tau} C_X(z)\ (\tau-z-(-\tau))\ dz
$$

$$
\begin{aligned}
&= \int_{-2\tau}^{0} C_X(z)\ (2\tau+z)\ dz \ + \ \int_{0}^{2\tau} C_X(z)\ (2\tau-z)\ dz \\
&= \int_{-2\tau}^{2\tau} C_X(z)\ (2\tau-|z|)\ dz
\end{aligned}
$$

- Now we get σ_τ^2 as

$$
\begin{aligned}
\sigma_\tau^2 &= \frac{1}{4\tau^2}\int_{-\tau}^{\tau}\int_{-\tau}^{\tau} C_X(t-t')\ dt\ dt' \\
&= \frac{1}{4\tau^2}\int_{-2\tau}^{2\tau} C_X(z)\ (2\tau-|z|)\ dz \\
&= \frac{1}{2\tau}\int_{-2\tau}^{2\tau} C_X(z)\ \left(1-\frac{|z|}{2\tau}\right)\ dz
\end{aligned}
$$

- Hence, a sufficient condition for $\sigma_\tau^2 \to 0$ is

$$
\int_{-\infty}^{\infty} |C_X(z)|\ dz \ < \ \infty
$$

- This is a sufficient condition for the process being mean-ergodic

Poisson Process

- This is the next process we study
- This is a discrete-state continuous-time process
- The index set is the interval $[0, \infty)$ and all random variables are discrete and take non-negative integer values.

- A random process $\{N(t),\ t \geq 0\}$ is called a counting process if
  1. $N(t) \geq 0$ and is integer-valued
  2. If $s < t$ then, $N(s) \leq N(t)$

  $N(t)$ represents number of 'events' till t
- The counting process has independent increments if for all $t_1 < t_2 \leq t_3 < t_4$, $N(t_2) - N(t_1)$ is independent of $N(t_4) - N(t_3)$
- In particular, for all $s > t$, $N(s) - N(t)$ is independent of $N(t) - N(0)$
- The process is said to have stationary increments if $N(t_2) - N(t_1)$ has the same distribution as $N(t_2 + \tau) - N(t_1 + \tau)$, $\forall \tau, \forall t_2 > t_1$

- We start with two definitions of Poisson process
- **Definition 1** A counting process $\{N(t), t \geq 0\}$ is said to be a Poisson process with rate $\lambda > 0$ if
  1. $N(0) = 0$
  2. The process has stationary and independent increments
  3. $Pr[N(t) = n] = e^{-\lambda t} \frac{(\lambda t)^n}{n!}, \ n = 0, 1, \cdots$
- $N(t)$ is Poisson with parameter λt
- $E[N(t)] = \lambda t$ and hence λ is called rate
- Since the process has stationary increments and $N(0) = 0$, $(N(t+s) - N(s))$ would be Poisson with parameter λt for all $s, t > 0$.

- **Definition 2** A counting process $\{N(t), t \geq 0\}$ is said to be a Poisson process with rate $\lambda > 0$ if
  1. $N(0) = 0$
  2. The process has stationary and independent increments
  3. $Pr[N(h) = 1] = \lambda h + o(h)$ and $Pr[N(h) \geq 2] = o(h)$
- We say $g(h)$ is $o(h)$ if

$$\lim_{h \to 0} \frac{g(h)}{h} = 0$$

- This definition tells us when Poisson process may be a good model
- We will show that both definitions are equivalent

- We first show Definition 2 $\Rightarrow$ Definition 1
- For this we need to calculate distribution of $N(t)$
- Let $P_n(t) = Pr[N(t) = n]$

$$\begin{aligned}
P_0(t+h) &= Pr[N(t+h) = 0] \\
&= Pr[N(t) = 0, \ N(t+h) - N(t) = 0] \\
&= Pr[N(t) = 0] \ Pr[N(t+h) - N(t) = 0] \\
&\qquad \text{(because of independent increments)} \\
&= Pr[N(t) = 0] \ Pr[N(h) = 0] \quad \text{(stationary increments)} \\
&= P_0(t)(1 - \lambda h + o(h))
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \quad \frac{P_0(t+h) - P_0(t)}{h} &= -\lambda P_0(t) + \frac{o(h)}{h} \\
\Rightarrow \quad \frac{d}{dt} P_0(t) &= -\lambda P_0(t)
\end{aligned}$$

- Now we can solve this differential equation to get $P_0(t)$

$$\begin{aligned}
\frac{d}{dt} P_0(t) &= -\lambda P_0(t) \\
\Rightarrow \quad \frac{1}{P_0(t)} \frac{d}{dt} P_0(t) &= -\lambda \\
\Rightarrow \quad \ln(P_0(t)) &= -\lambda t + c \\
\Rightarrow \quad P_0(t) &= K e^{-\lambda t}
\end{aligned}$$

- Since $P_0(0) = Pr[N(0) = 0] = 1$, we get $K = 1$ and hence
$$P_0(t) = Pr[N(t) = 0] = e^{-\lambda t}$$
- Next we consider $P_n(t)$ for $n > 0$

$$
\begin{aligned}
P_n(t+h) &= Pr[N(t+h)=n] \\
&= Pr[N(t)=n,\ N(t+h)-N(t)=0] + \\
&\qquad Pr[N(t)=n-1,\ N(t+h)-N(t)=1] + \\
&\qquad \sum_{k=2}^{n} Pr[N(t)=n-k,\ N(t+h)-N(t)=k] \\
&= P_n(t)P_0(h) + P_{n-1}(t)P_1(h) + o(h) \\
&= P_n(t)(1-\lambda h + o(h)) + P_{n-1}(t)(\lambda h + o(h)) + o(h)
\end{aligned}
$$

$$
\Rightarrow \frac{P_n(t+h)-P_n(t)}{h} = -\lambda P_n(t) + \lambda P_{n-1}(t) + \frac{o(h)}{h}
$$

$$
\Rightarrow \frac{d}{dt}P_n(t) = -\lambda P_n(t) + \lambda P_{n-1}(t)
$$

$$
\frac{d}{dt}P_n(t) + \lambda P_n(t) = \lambda P_{n-1}(t)
$$

- We need to solve this linear ODE to obtain P_n
- The integrating factor is $e^{\lambda t}$. Let $P_n'(t) = \frac{d}{dt}P_n(t)$

$$
\begin{aligned}
e^{\lambda t}\left(P_n'(t) + \lambda P_n(t)\right) &= e^{\lambda t}\ \lambda P_{n-1}(t) \\
\Rightarrow \frac{d}{dt}\left(P_n(t)\ e^{\lambda t}\right) &= \lambda\ e^{\lambda t}\ P_{n-1}(t)
\end{aligned}
$$

- We need P_{n-1} to solve for P_n. Take $n=1$

$$
\begin{aligned}
\frac{d}{dt}\left(P_1(t)\ e^{\lambda t}\right) &= \lambda\ e^{\lambda t}\ P_0(t) = \lambda e^{\lambda t}e^{-\lambda t} = \lambda \\
\Rightarrow\ e^{\lambda t}P_1(t) = \lambda t + c\ &\Rightarrow\ P_1(t) = e^{-\lambda t}(\lambda t + c)
\end{aligned}
$$

- Since $P_1(0) = Pr[N(0)=1] = 0$, $c=0$
  Hence $P_1(t) = \lambda t\ e^{-\lambda t}$

- We showed: $P_0(t) = e^{-\lambda t}$ and $P_1(t) = \lambda t e^{-\lambda t}$
- We need to show: $P_k(t) = e^{-\lambda t}\ \frac{(\lambda t)^k}{k!}$
- Assume it is true till $k = n-1$

$$
\frac{d}{dt}\left(P_n(t)\ e^{\lambda t}\right) = \lambda\ e^{\lambda t}\ P_{n-1}(t) = \lambda e^{\lambda t}e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!} = \lambda^n \frac{t^{n-1}}{(n-1
$$

$$
\Rightarrow\ e^{\lambda t}P_n(t) = \lambda^n \frac{t^n}{n}\frac{1}{(n-1)!} + c\ \Rightarrow\ P_n(t) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}
$$

where $c=0$ because $P_n(0)=0$.

- This completes the proof that Definition 2 implies Definition 1

- **Definition 1** A counting process $\{N(t), t \geq 0\}$ is said to be a Poisson process with rate $\lambda > 0$ if
  1. $N(0) = 0$
  2. The process has stationary and independent increments
  3. $Pr[N(t)=n] = e^{-\lambda t}\ \frac{(\lambda t)^n}{n!}$, $n = 0, 1, \cdots$
- **Definition 2** A counting process $\{N(t), t \geq 0\}$ is said to be a Poisson process with rate $\lambda > 0$ if
  1. $N(0) = 0$
  2. The process has stationary and independent increments
  3. $Pr[N(h)=1] = \lambda h + o(h)$ and $Pr[N(h) \geq 2] = o(h)$

- Now we prove Definition 1 implies Definition 2
- We need to only show point(3) of Definition 2 using point (3) of Definition 1

$$\text{Let} \quad Pr[N(t) = k] = e^{-\lambda t} \frac{(\lambda t)^k}{k!}$$

$$Pr[N(h) = 1] = \lambda\, h\, e^{-\lambda h} = \lambda\, h + \lambda\, h \left(e^{-\lambda h} - 1\right) = \lambda\, h + o(h)$$

because

$$\lim_{h \to 0} \frac{\lambda\, h \left(e^{-\lambda h} - 1\right)}{h} = \lim_{h \to 0} \lambda \left(e^{-\lambda h} - 1\right) = 0$$

- We showed $Pr[N(h) = 1] = \lambda h + o(h)$

- Now we need to show $Pr[N(h) \geq 2] = o(h)$

$$\begin{aligned} Pr[N(h) \geq 2] &= 1 - Pr[N(h) = 0] - Pr[N(h) = 1] \\ &= 1 - e^{-\lambda h} - \lambda h e^{-\lambda h} \end{aligned}$$

- This goes to zero as $h \to 0$
- We can use L'Hospital rule

$$\lim_{h \to 0} \frac{1 - e^{-\lambda h} - \lambda h e^{-\lambda h}}{h} = \lim_{h \to 0} \frac{\lambda e^{-\lambda h} - \lambda e^{-\lambda h} + \lambda^2 h e^{-\lambda h}}{1} = 0$$

- This completes the proof that Definition 2 implies Definition 1

These two definitions are equivalent

- **Definition 1** A counting process $\{N(t), t \geq 0\}$ is said to be a Poisson process with rate $\lambda > 0$ if
  1. $N(0) = 0$
  2. The process has stationary and independent increments
  3. $Pr[N(t) = n] = e^{-\lambda t} \frac{(\lambda t)^n}{n!}, \; n = 0, 1, \cdots$
- **Definition 2** A counting process $\{N(t), t \geq 0\}$ is said to be a Poisson process with rate $\lambda > 0$ if
  1. $N(0) = 0$
  2. The process has stationary and independent increments
  3. $Pr[N(h) = 1] = \lambda h + o(h)$ and $Pr[N(h) \geq 2] = o(h)$

- Since the process has stationary increments, for $t_2 > t_1$,

$$\begin{aligned} Pr[N(t_2) - N(t_1) = k]] &= Pr[N(t_2 - t_1) - N(0) = k] \\ &= e^{-\lambda(t_2 - t_1)} \frac{(\lambda(t_2 - t_1))^k}{k!} \end{aligned}$$

- The first order distribution of the process is: $N(t) \sim$ Poisson(λt)
- This, along with stationary and independent increments property determines all distributions

$$\begin{aligned} &Pr[N(t_1) = n_1, N(t_2) = n_2, N(t_3) = n_3] \\ &\quad = Pr[N(t_1) = n_1]\; Pr[N(t_2) - N(t_1) = n_2 - n_1] \\ &\qquad\quad Pr[N(t_3) - N(t_2) = n_3 - n_2] \\ &= Pr[N(t_1) = n_1]\; Pr[N(t_2 - t_1) = n_2 - n_1]\; Pr[N(t_3 - t_2) = n_3 - n_2] \end{aligned}$$

where we assumed $t_1 < t_2 < t_3$

- We can easily calculate mean and autocorrelation of the process

$$\eta_N(t) = E[N(t)] = \lambda\, t \quad \Rightarrow \quad \text{not stationary}$$

With $t_2 > t_1$, we have

$$\begin{aligned} R_N(t_1, t_2) &= E[N(t_2)N(t_1)] \\ &= E[N(t_1)(N(t_2) - N(t_1) + N(t_1))] \\ &= E[N(t_1)(N(t_2) - N(t_1))] + E[N(t_1)^2] \\ &= E[N(t_1)]\ E[N(t_2) - N(t_1)] + E[N(t_1)^2] \\ &= E[N(t_1)]\ E[N(t_2 - t_1)] + E[N(t_1)^2] \\ &= \lambda t_1(\lambda(t_2 - t_1)) + (\lambda t_1 + \lambda^2 t_1^2) \\ &= \lambda t_1 + \lambda^2 t_1 t_2 \end{aligned}$$

$$\Rightarrow \quad R_N(t_1, t_2) = \lambda^2 t_1 t_2 + \lambda\ \min(t_1, t_2)$$

Inter-arrival or waiting times

- Let T_1 denote the time of first event and let T_n denote the time between n^{th} and $(n-1)st$ events.
- Let $S_n = \sum_{i=1}^{n} T_i$ – time of n^{th} event

$$Pr[T_1 > t] = Pr[N(t) = 0] = e^{-\lambda t}$$

$$\Rightarrow \quad T_1 \sim \text{exponential}(\lambda)$$

$$\begin{aligned} Pr[T_2 > t | T_1 = s] &= Pr[0 \ \text{ events in } \ (s,\ s+t] \mid T_1 = s] \\ &= Pr[0 \ \text{ events in } \ (s,\ s+t]\ ] = e^{-\lambda t} \\ \Rightarrow \ Pr[T_2 > t] &= \int Pr[T_2 > t | T_1 = s]\ f_{T_1}(s)\ ds = e^{-\lambda t} \end{aligned}$$

- T_n are iid exponential with parameter λ

- The time of n^{th} event is

$$S_n = \sum_{i=1}^{n} T_i$$

Since T_i are iid, exponential, S_n is Gamma with parameters n, λ
- Let $s < t$.

$$\begin{aligned} Pr[T_1 < s | N(t) = 1] &= \frac{Pr[T_1 < s,\ N(t) = 1]}{Pr[N(t) = 1]} \\ &= \frac{Pr[1 \text{ event in } (0,\ s),\ 0 \text{ in } [s,\ t]]}{Pr[N(t) = 1]} \\ &= \frac{\lambda s e^{-\lambda s}\ e^{-\lambda(t-s)}}{\lambda t e^{-\lambda t}} \\ &= \frac{s}{t} \end{aligned}$$

- Conditioned on $N(t) = 1$, T_1 is uniform over $[0, t]$

- This can be used, e.g., in simulating Poisson process
- We can cut time axis into small intervals of length h.
- In each interval we can decide whether or not there is an event, with prob λh.
- If there is an event, we choose its time uniformly in the interval.
- Called Bernoulli approximation of Poisson process
- We could also generate Poisson process by generating independent exponential random variables

Examples

- We look at a few simple example problems using Poisson process.

$$\begin{aligned} E[N(4) - N(2) | N(1) = 3] &= E[N(4) - N(2)] \\ &= E[N(2) - 0] = 2\lambda \end{aligned}$$

- Another example;

$$E[S_4] = E\left[\sum_{i=1}^{4} T_i\right] = \frac{4}{\lambda}$$

- The memoryless property of exponential rv can be useful

$$Pr[S_3 > t | N(1) = 2] = \begin{cases} 1 & \text{if } t < 1 \\ e^{-\lambda(t-1)} & \text{if } t \geq 1 \end{cases}$$

- We can explicitly derive this (taking $t > 1$)

$$\begin{aligned} Pr[S_3 > t | N(1) = 2] &= \frac{Pr[S_3 > t,\ N(1) = 2]}{Pr[N(1) = 2]} \\ &= \frac{Pr[2 \text{ event in } (0,\ 1],\ 0 \text{ in } (1,\ t)]}{Pr[N(1) = 2]} \\ &= \frac{Pr[2 \text{ event in } (0,\ 1]]\ Pr[0 \text{ in } (1,\ t)]}{Pr[2 \text{ event in } (0,\ 1]]} \\ &= e^{-\lambda(t-1)} \end{aligned}$$

- Here is another example

$$E[S_4 | N(1) = 2] = 1 + E[S_2] = 1 + \frac{2}{\lambda}$$

Exercise for you: calculate $Pr[S_4 > t | N(1) = 2]$ and use it to find the above expectation

Example

- Given a specific T_0 we want to guess which is the last event before T_0.
- Consider a strategy: we will wait till $T_0 - \tau$ and pick the next event as the last one before T_0.
- The probability of winning for this is

$$Pr[\text{exactly 1 event in } (T_0 - \tau,\ T_0)] = \lambda\tau e^{-\lambda\tau}$$

- We pick τ to maximize this

$$\lambda e^{-\lambda\tau} - \lambda^2 \tau e^{-\lambda\tau} = 0 \quad \Rightarrow \quad \tau = \frac{1}{\lambda}$$

- Intuitively reasonable because expected inter-arrival time is $\frac{1}{\lambda}$

- Let $\{N(t), t \geq 0\}$ be a Poisson process with rate λ
- Suppose each event can be one of two types – Typ-I or Typ-II
  - $N_1(t)$ = number of Typ-I events till t
  - $N_2(t)$ = number of Typ-II events till t
  - Note that $N(t) = N_1(t) + N_2(t),\ \forall t$
- Suppose that, independently of everything else, an event is of Typ-I with probability p and Typ-II with probability $(1 - p)$

Theorem: $\{N_1(t), t \geq 0\}$ and $\{N_2(t), t \geq 0\}$ are Poisson processes with rate λp and $\lambda(1 - p)$ respectively, and they are independent

$$
\begin{aligned}
&Pr[N_1(t) = n, N_2(t) = m] \\
&= \sum_k Pr[N_1(t) = n, N_2(t) = m \mid N(t) = k]\ Pr[N(t) = k] \\
&= Pr[N_1(t) = n, N_2(t) = m \mid N(t) = m+n]\ Pr[N(t) = m+n] \\
&= {}^{m+n}C_n\ p^n\ (1-p)^m\ e^{-\lambda t}\ \frac{(\lambda t)^{m+n}}{(m+n)!} \\
&= \frac{(m+n)!}{m!\ n!}\ p^n\ (1-p)^m\ e^{-\lambda(p+1-p)t}\ \frac{(\lambda t)^m (\lambda t)^n}{(m+n)!} \\
&= \frac{(\lambda p t)^n}{n!}\ e^{-\lambda p t}\ \frac{(\lambda(1-p)t)^m}{m!}\ e^{-\lambda(1-p)t}
\end{aligned}
$$

- This shows that $N_1(t)$ and $N_2(t)$ are independent Poisson

- The interesting issue here is that $N_1(t)$ and $N_2(t)$ are independent.
- Suppose customers arrive at a bank as a Poisson process with rate 12 per hour.
- Independently of everything, an arriving customer is male or female with equal probability.
- Q: Given that on some day 6 male customers came in the first half hour, what is the expected number of female customers in that half hour?
- The answer is 3 because the two processes are independent

- The theorem is easily generalized to multiple types for events
- Consider Poisson process with rate λ
- Suppose, independently of everything, an event is Typ-i with probability p_i, $i = 1, \cdots K$.
- Note we have $\sum_{i=1}^{K} p_i = 1$
- Let $N_i(t)$ be the number of Typ-i customers till t
- Then, these are independent Poisson processes with rates λp_i, $i = 1, \cdots, K$

- Superposition of independent Poisson processes also gives Poisson process.
- If N_1 and N_2 are independent Poisson processes with rates λ_1 and λ_2 then $N(t) = N_1(t) + N_2(t)$ is a Poisson process with rate $\lambda_1 + \lambda_2$
- We know that sum of independent Possion rv's is Poisson

- Suppose number of radioactive particles emitted is Poisson with rate λ.
- We are counting particles using a sensor
- Suppose (independent of everything) an emitted particle is detected by our sensor with probability p
- Given that we detected K particles till t what is the expected number of particles emitted?
- Let these processes be $N(t), N_1(t), N_2(t)$

$$\begin{aligned} E[N(t) | N_1(t) = K] &= E[N_1(t) + N_2(t) | N_1(t) = K] \\ &= K + E[N_2(t)] = K + \lambda(1-p)t \end{aligned}$$

where we have used independence of N_1 and N_2

- There is an interesting generalization of this.
- Events are of different types
- The type of an event can depend on the time of occurrence but it is independent of everything else.
- Suppose an event occurring at time t is Typ-i with probability $p_i(t)$.
- $p_i(t) \geq 0, \ \forall i, t$ and $\sum_{i=1}^{K} p_i(t) = 1, \ \forall t$
- $N_i(t)$ is the number of Typ-i events till t

Theorem; Then, at any t, $N_i(t), i = 1, \cdots K$ are independent Poisson random variables with

$$E[N_i(t)] = \lambda \int_0^t p_i(s) \, ds$$

Example: Tracking infections

- We use a simple model
- Individuals get infected as a Poisson process with rate λ
- Time between getting infected and showing symptoms is a random variable with known distribution function G An individual infected at s would show symptoms by t with probability $G(t-s)$
- The incubation times of different infected individuals are iid
- Define
  - $N(t)$ – total number infected till t
  - $N_1(t)$ – number showing symptoms by t
  - $N_2(t)$ – infected by t but not showing symptoms

- Define two types of events. We take t as current time and treat it as fixed
  - An event occurring at s is Typ-1 with probability $G(t-s)$
  - It is Typ-2 with probability $1 - G(t-s)$
- Then, Typ-1 individuals are those showing symptoms by t
- From our theorem,

$$E[N_1(t)] = \lambda \int_0^t G(t-s) \, ds = \lambda \int_0^t G(y) \, dy$$

$$E[N_2(t)] = \lambda \int_0^t (1 - G(t-s)) \, ds = \lambda \int_0^t (1 - G(y)) \, dy$$

- Suppose we have n_1 people showing symptoms at t
- We can approximate

$$n_1 \approx E[N_1(t)] = \lambda \int_0^t G(y)\, dy$$

- Hence we can estimate

$$\hat{\lambda} = \frac{n_1}{\int_0^t G(y)\, dy}$$

- Using this we can approximate

$$E[N_2(t)] \approx \hat{\lambda} \int_0^t (1 - G(y))\, dy$$

- The Poisson process we considered is called homogeneous because the rate is constant.
- For a non-homogeneous Poisson process the rate can be changing with time.
- But we can still use a definition similar to definition 2

$$Pr[N(t+h) - N(t) = 1] = \lambda(t)h + o(h)$$

- We still stipulate independent increments though we cannot have stationary increments now
- One can show that $N(t+s) - N(t)$ is Poisson with parameter $m(t+s) - m(t)$ where $m(\tau) = \int_0^\tau \lambda(s)\, ds$
- Suppose Y_i are iid and ind of $N(t)$. Then

$$X(t) = \sum_{i=1}^{N(t)} Y_i$$

is called a compound Poisson process

Random Walk

- Let Z_i be iid with $Pr[Z_i = +s] = Pr[Z_i = -s] = 0.5$
- Define a continuous-time process $X(t)$ by

$$\begin{aligned} X(nT) &= Z_1 + Z_2 + \cdots + Z_n \\ X(t) &= X(nT), \quad \text{for } nT \leq t < (n+1)T \end{aligned}$$

- Viewed as a discrete-time process, $X(nT)$, is a Markov chain.
- Called a (one dimensional) random walk
- It is the position after n random steps
- We defined $X(t)$ by piece-wise constant interpolation of $X(nT)$
- We could have also use piece-wise linear interpolation

- We have $EZ_i = 0$ and $E[Z_i^2] = s^2$
- Hence, $E[X(nT)] = 0$ and $E[X^2(nT)] = ns^2$
- For large n, $\frac{X(nT)}{s\sqrt{n}}$ would be Gaussian

$$Pr\left[\frac{X(nT)}{s\sqrt{n}} \leq y\right] \approx \Phi(y)$$

  where Φ is distribution function of standard Normal
- For any t, $X(t)$ is $X(nT)$ for $n = [t/T]$.
  Large n would mean large t. Hence

$$Pr[X(t) \leq ms] = Pr\left[\frac{X(t)}{s\sqrt{n}} \leq \frac{ms}{s\sqrt{n}}\right] \approx \Phi\left(\frac{m}{\sqrt{n}}\right), \quad \text{for large } t$$

- We are interested in limit of this process as $T \to 0$

- Consider $t = nT$

$$E[X^2(t)] = ns^2 = s^2 \frac{t}{T}$$

- If we let $T \to 0$ then the variance goes to infinity (the process goes to infinity) unless we let s also to go to zero.
- We actuall need s^2 to go to zero at the same rate as T.
- So, we keep $s^2 = \alpha T$ and let T go to zero.
- Define

$$W(t) = \lim_{T \to 0, s^2 = \alpha T} X(t)$$

  This is called the Wiener Process or Brownian motion.
  This result is known as Donsker's theorem
- Let us intuitively see some properties of $W(t)$

- We have seen that for $n = [t/T]$,

$$Pr[X(t) \leq ms] \approx \Phi\left(\frac{m}{\sqrt{n}}\right)$$

- Let $w = ms$ and $t = nT$. Then

$$\frac{m}{\sqrt{n}} = \frac{w/s}{\sqrt{t/T}} = \frac{w}{\sqrt{t}} \sqrt{\frac{T}{s^2}} = \frac{w}{\sqrt{\alpha t}}$$

- $W(t)$ is limit of $X(t)$ as T goes to zero
- As T goes to zero, any t is 'large n'.
- Hence we can expect

$$Pr[W(t) \leq w] = \Phi\left(\frac{w}{\sqrt{\alpha t}}\right)$$

$$\Rightarrow \quad W(t) \sim \mathcal{N}(0, \ \alpha t)$$

- We had Z_i iid and defined

$$X(nT) = Z_1 + Z_2 + \cdots + Z_n$$

- Hence we get

$$X((m+n)T) - X(nT) = Z_{n+1} + \cdots + Z_{n+m}$$

  Thus, $X(nT)$ is independent of $X((m+n)T) - X(nT)$.
- Hence the $X(nT)$ process has independent increments
- Hence, we can expect $W(t)$ to be a process with independent increments

- $X((m+n+k)T) - X((n+k)T)$ and $X((m+n)T) - X(nT)$ both are sums of m of the Z_i's
- Hence both would have the same distribution
- Thus $X(nT)$ would also have stationary increments.
- Hence we also expect $W(t)$ to have stationary increments
- Thus, $W(t)$ should be a process with stationary and independent increments and for each t, $W(t)$ is Gaussian with zero mean and variance proportional to t
- We will now formally define Brownian motion using these properties.

- Let $\{X(t),\ t \geq 0\}$ be a continuous-state continuous-time process

  This process is called a Brownian motion if
  1. $X(0) = 0$
  2. The process has stationary and independent increments
  3. For every $t > 0$, $X(t)$ is Gaussian with mean 0 and variance $\sigma^2 t$
- Let $B(t) = \frac{X(t)}{\sigma}$. Then, variance of $B(t)$ is t
- $\{B(t),\ t \geq 0\}$ is called standard Brownian Motion
- Let $Y(t) = X(t) + \mu$. Then $Y(t)$ has non-zero mean
- The mean can be a function of time
- $\{Y(t),\ t \geq 0\}$ is called Brownian motion with a drift

- Let $\{X(t),\ t \geq 0\}$ be a Brownian motion
- The process has stationary increments.
- Hence for $t_2 > t_1$, $X(t_2) - X(t_1)$ has the same distribution as $X(t_2 - t_1)$
- Thus, $X(t_2) - X(t_1)$ is Gaussian with zero mean and variance $\sigma^2(t_2 - t_1)$
- Since increments are also independent, we can show that all n^{th} order distributions are Gaussian

- We can calculate the autocorrelation function

$$\begin{aligned}
R_X(t_1, t_2) &= E[X(t_1)X(t_2)] \\
&= E[X(t_1)\ (X(t_2) - X(t_1) + X(t_1))], \quad \text{(take } t_1 < t_2\text{)} \\
&= E[X(t_1)(X(t_2) - X(t_1))] + E[X^2(t_1)] \\
&= E[X(t_1)]\ E[X(t_2) - X(t_1)] + E[X^2(t_1)] \\
&= E[X^2(t_1)] \\
&= \sigma^2\ t_1
\end{aligned}$$

- Since $E[X(t)] = 0,\ \forall t$, we have

$$\text{Cov}(X(t_1), X(t_2)) = E[X(t_1)X(t_2)] = \sigma^2\ \min(t_1, t_2)$$

- Suppose we want the joint distribution of $X(t_1), X(t_2), \cdots, X(t_n)$
- Let $t_1 < t_2 < \cdots < t_n$
- Define random variables $Y_1, \cdots, Y_n$ by

$$Y_1 = X(t_1), \;\; Y_2 = X(t_2) - X(t_1), \;\; Y_3 = X(t_3) - X(t_2), \cdots$$

- We know Y_i are independent because the process has independent increments
- This transformation is invertible
- Hence we can get joint density of $X(t_1), \cdots X(t_n)$ in terms of joint density of $Y_1, \cdots, Y_n$
- This is how we can get n^{th} order density for any continuous-state process with independent increments

$$Y_1 = X(t_1), \;\; Y_i = X(t_i) - X(t_{i-1}), \; i = 2, \cdots, n$$

- The transformation is invertible

$$\begin{aligned} X(t_1) &= Y_1 \\ X(t_2) &= Y_1 + Y_2 \\ X(t_3) &= Y_1 + Y_2 + Y_3 \\ &\vdots \\ X(t_n) &= Y_1 + Y_2 + \cdots + Y_n \end{aligned}$$

- $Y_1, \cdots Y_n$ are independent and Gaussian and hence are Jointly Gaussian
- Hence $X(t_1), \cdots, X(t_n)$ are jointly Gaussian
- Thus all n^{th} order distributions are Gaussian

- $X(t_1), X(t_2), \cdots, X(t_n)$ are jointly Gaussian.
- We can write their joint density because we know the means, variances and covariances
- We can also write the density using the transformation considered earlier
  Let $t_1 < t_2 < \cdots < t_n$

$$f_X(x_1, \cdots, x_n; t_1, \cdots, t_n) = f_{Y_1}(x_1) f_{Y_2}(x_2 - x_1) \cdots f_{Y_n}(x_n - x_{n-1})$$

- Note that $Y_i = X(t_i) - X(t_{i-1})$ is Gaussian with mean zero and variance $\sigma^2(t_i - t_{i-1})$, $i = 1, \cdots, n$
  (Take $t_0 = 0$)

- Since all joint densities are Gaussian and are easy to write, we can also calculate conditional densities

$$\begin{aligned} f_{X(s)|X(t)}(x|b) &= \frac{f_{X(s)X(t)}(x,b)}{f_{X(t)}(b)} \quad (s < t) \\ &= \frac{f_{X(s)}(x)\; f_{X(t)-X(s)}(b-x)}{f_{X(t)}(b)} \\ &\propto e^{-\frac{x^2}{2s}}\; e^{-\frac{(b-x)^2}{2(t-s)}} \quad (\text{taking } \sigma^2 = 1) \\ &\propto \exp\left(-x^2\left(\frac{1}{2s} + \frac{1}{2(t-s)}\right) + \frac{bx}{t-s}\right) \\ &\propto \exp\left(-\frac{t}{2s(t-s)}\left(x^2 - 2\frac{sb}{t}x\right)\right) \\ &\propto \exp\left(-\frac{(x - bs/t)^2}{2s(t-s)/t}\right) \end{aligned}$$

- Hence the conditional density is Gaussian with mean bs/t and variance $s(t-s)/t$

- An important result is that Brownian motion paths are continuous
- Brownian motion is the limit of random walk where both s and T tend to zero
- Intuitively the paths should be continuous.
- The paths are continuous but non-differentiable everywhere
- This is a deep result

Hitting Times

- Let T_a denote the first time Brownian motion hits a. We take $a > 0$.

$$\begin{aligned} Pr[X(t) \geq a] &= Pr[X(t) \geq a \mid T_a \leq t]\ Pr[T_a \leq t] + \\ & \qquad Pr[X(t) \geq a \mid T_a > t]\ Pr[T_a > t] \end{aligned}$$

- Since Brownian motion paths are continuous, $Pr[X(t) \geq a \mid T_a > t] = 0$
- Brownian motion is a limit of symmetric random walk. Hence if we had already hit a sometime back, then now we are as likely to be above a as below it.

$$\Rightarrow \quad Pr[X(t) \geq a \mid T_a \leq t] = \frac{1}{2}$$

Thus

$$P[X(t) \geq a] = 0.5 Pr[T_a \leq t]$$

- Hence we get

$$\begin{aligned} Pr[T_a \leq t] &= 2\ Pr[X(t) \geq a] \\ &= \frac{2}{\sqrt{2\pi t}} \int_a^\infty e^{-\frac{x^2}{2t}}\ dx \\ &= \frac{2}{\sqrt{2\pi}} \int_{a/\sqrt{t}}^\infty e^{-\frac{y^2}{2}}\ dy \end{aligned}$$

- Here we have assumed $a > 0$. For $a < 0$ the situation is similar. Hence the above is true even for $a < 0$ except that the lower limit becimes $|a|/\sqrt{t}$
- Another interesting consequence is the following

$$\begin{aligned} Pr[\max_{0 \leq s \leq t} X(s) \geq a] &= Pr[T_a \leq t], \text{ by continuity of paths} \\ &= 2 Pr[X(t) \geq a] \end{aligned}$$

Geometric Brownian Motion

- Let $\{Y(t),\ t \geq 0\}$ is a Brownian motion with drift. Define

$$X(t) = e^{Y(t)}$$

- Then, $\{X(t),\ t \geq 0\}$ is called geometric Brownian motion. It is useful in mathematial finance
- Let $X_0, X_1, \cdots$ be time series of prices of a stock.
- Let $Y_n = X_n / X_{n-1}$ and assume Y_i are iid

$$X_n = Y_n X_{n-1} = Y_n Y_{n-1} X_{n-2} = \cdots = Y_n Y_{n-1} \cdots Y_1 X_0$$

$$\Rightarrow \quad \ln(X_n) = \sum_{i=1}^n \ln(Y_i) + \ln(X_0)$$

- Since $\ln(Y_i)$ are iid, with suitable normalization, the interpolated process $\ln(X(t))$ would be Brownian motion and $X(t)$ would be geometric Brownian motion

Gaussian Processes

- A continuous-time continuous-state process $\{X(t),\ t \geq 0\}$ is said to be a Gaussian process if for all n and all $t_1, t_2, \cdots, t_n$, we have that $X(t_1), \cdots, X(t_n)$ are jointly Gaussian.
- The Brownian motion is an example of a Gaussian Process
- The Brwonian motion is a Gaussian process with

$$E[X(t)] = 0, \quad \text{Var}(X(t)) = \sigma^2 t, \quad \text{Cov}(X(s), X(t)) = \sigma^2 \min(s, t)$$

- Recall that the multivariate Gaussian density is specified by the marginal means, variances and the covariances of the random variables
- Hence, a general Gaussian process is specified by the mean function and the variance and covariance functions

- Consider the statistics of the Brownian motion process for $0 < t < 1$ under the condition that $X(1) = 0$
- Consider standard Brownian motion. $(\sigma^2 = 1)$

$$E[X(t)|X(1) = 0] = \frac{t}{1}\ 0 = 0$$

Recall that, for $s < t$, conditional density of $X(s)$ conditioned on $X(t) = b$ is gaussian with mean bs/t and variance $s(t-s)/t$

Now, for $s < t < 1$, since $E[X(s)|X(1) = 0] = 0,\ s < 1$,

$$\begin{aligned}
\text{Cov}(X(s), X(t)|X(1) = 0) &\triangleq E[X(s)X(t) \mid X(1) = 0] \\
&= E[\ E[X(s)X(t) \mid X(t), X(1) = 0] \mid X(1) = 0] \\
&= E[\ X(t)E[X(s) \mid X(t)] \mid X(1) = 0] \\
&= E[\ X(t)\frac{s}{t}X(t) \mid X(1) = 0] \\
&= \frac{s}{t}\ E[X^2(t) \mid X(1) = 0] \\
&= \frac{s}{t}\ t(1-t) \\
&= s(1-t)
\end{aligned}$$

Thus, for $0 < t < 1$, conditioned on $X(1) = 0$, this process has mean 0 and covariance function $s(1-t),\ s < t$

- Consider a process $\{Z(t),\ 0 \leq t \leq 1\}$.
- It is called Brownian Bridge process if it is a Gaussian process with mean zero and covariance function $\text{Cov}(Z(s), Z(t)) = s(1-t)$ when $s \leq t$.
- Let $X(t)$ be a standard Brownian motion process.
- Then, $Z(t) = X(t) - tX(1)$, $0 \leq t \leq 1$ is a Brownian Bridge
- Easy to see it is a Gaussian process with mean zero. For $s < t$

$$\begin{aligned}
\text{Cov}(Z(s), Z(t)) &= \text{Cov}(X(s) - sX(1), X(t) - tX(1)) \\
&= \text{Cov}(X(s), X(t)) - t\text{Cov}(X(s), X(1)) - \\
&\quad s\text{Cov}(X(1), X(t)) + st\text{Cov}(X(1), X(1)) \\
&= s - st - st + st = s(1-t)
\end{aligned}$$

White Noise

- Consider a process $\{V(t),\ t \geq 0\}$ with

$$E[V(t)] = 0; \quad \text{Var}(V(t)) = \sigma^2 \quad \text{Cov}(V(t), V(s)) = 0,\ s \neq t$$

- This is a kind of generalization of sequence of iid random variables to continuous time
- It is an example of what is called White Noise.

- Assume $V(t)$ is Gaussian. Let

$$X(t) = \int_0^t V(\tau)\, d\tau$$

- Then we get $E[X(t)] = 0$ and

$$E[X^2(t)] = \int_0^t \int_0^t E[V(t_1)V(t_2)]\, dt_1\, dt_2 = \int_0^t \sigma^2\, dt_1 = \sigma^2 t$$

$$E[X(t_1)(X(t_2) - X(t_1))] = \int_0^{t_1} \int_{t_1}^{t_2} E[V(t)V(t')]\, dt\, dt' = 0$$

- We see that $X(t)$ is a process with mean zero, variance proportional to t and having uncorrelated increments.
- One can show that it would be a Brownian motion
- The actual concept involved is rather deep

- We have considered three random processes
- Markov Chain
  – Example of Discrete-time discrete-state process
- Poisson Process
  – Example of continuous-time discrete-state process
- Brownian Motion
  – Example of continuous-time continuous-state process
- We need an example of discrete-time continuous-state process!
- Any sequence of continuous random variables would be a discrete-time continuous-state process

- Let $\{X_n,\ n = 0, 1, \cdots\}$ be a discrete-time continuous-state process.
- It is called a martingale if $E|X_n| < \infty,\ \forall n$ and

$$E[X_{n+1} \mid X_n, \cdots, X_0] = X_n,\ \forall n$$

- Suppose Z_i are iid with $Pr[Z_i = +1] = Pr[Z_i = -1] = 0.5$. Let

$$X_n = \sum_{i=1}^{n} Z_i \quad \Rightarrow \quad X_{n+1} = X_n + Z_{n+1}$$

- Since $EZ_i = 0,\ \forall i$,

$$E[X_{n+1} \mid X_n, \cdots, X_0] = E[X_n + Z_{n+1} | X_n] = X_n + E[Z_{n+1} | X_n] = X_n$$

- Hence, X_n is a martingale.
- When X_n is a martingale, we have

$$E[X_{n+1}] = E[X_n],\ \forall n$$

- $\{X_n,\ n = 0, 1, \cdots\}$ and $E|X_n| < \infty,\ \forall n$
- It is called a martingale if

$$E[X_{n+1} \mid X_n, \cdots, X_0] = X_n,\ \forall n$$

- It is called a supermartingale if

$$E[X_{n+1} \mid X_n, \cdots, X_0] \leq X_n,\ \forall n$$

- It is called a submartingale if

$$E[X_{n+1} \mid X_n, \cdots, X_0] \geq X_n,\ \forall n$$

Please note that these are 'simplified' definitions In the above, the conditioning random variables can be another sequence Y_i if $Y_1, \cdots, Y_n$ determine $X_1, \cdots, X_n$

- Martingales are useful because of the martingale convergence theorem.

  martingale convergence theorem: If X_n is a martingale with $\sup_n E|X_n| < \infty$ then X_n converges almost surely to a rv X which will have finite expectation. A positive supermartingale also converges almost surely

- Consider the 2-armed bandit problem in problem sheet 3.7
- The algorithm is

$$\begin{aligned}
p(k+1) &= p(k) + \lambda(1 - p(k)) \ \text{ if arm 1 chosen, } b(k) = 1 \\
&= p(k) - \lambda p(k) \ \text{ if arm 2 is chosen and } b(k) = 1 \\
&= p(k) \ \text{ if } b(k) = 0
\end{aligned}$$

- We get

$$\begin{aligned}
&E[p(k+1) - p(k)|p(k)] \\
&\quad = \lambda(1 - p(k))\ Pr[b(k) = 1, \text{arm } 1 \mid p(k)] \\
&\qquad -\lambda p(k)\ Pr[b(k) = 1, \text{arm } 2 \mid p(k)] \\
&\quad = \lambda(1 - p(k))\ Pr[b(k) = 1 \mid \text{arm } 1, p(k)]\ Pr[\text{arm } 1 \mid p(k)] \\
&\qquad -\lambda p(k)\ Pr[b(k) = 1 \mid \text{arm } 2, p(k)]\ Pr[\text{arm } 2 \mid p(k)]
\end{aligned}$$

- This gives us

$$\begin{aligned}
E[p(k+1) - p(k)|p(k)] &= \lambda(1 - p(k))\ d_1\ p(k) \\
&\quad -\lambda p(k)\ d_2\ (1 - p(k) \\
&= \lambda\ p(k)(1 - p(k))\ (d_1 - d_2) \\
&\geq 0, \quad \text{if} \quad d_1 > d_2
\end{aligned}$$

$$\Rightarrow \quad E[p(k{+}1) \mid p(k)] \geq p(k) \quad \Rightarrow E[p(k{+}1)] \geq E[p(k)],\ \forall k$$

- This also shows $p(k)$ is a submartingale.
- Here, $p(k)$ is bounded and $1 - p(k)$ is a supermartingale.
- So, we can conclude, the algorithm converges almost surely

- We have mentioned martingales as an example of discrete-time continuous processes
- A stochastic iterative algorithm essentially generates a discrete-time continuous-state processes.
- Martingales are very useful in analyzing convergence of many stochastic algorithms
- While we mentioned only discrete-time martingales, one can similarly have continuous-time martingales

Continuous-Time Markov Chains

- Let $\{X(t),\ t \geq 0\}$ be a continuous-time discrete-state process
- Let $X(t)$ take non-negative integer values
- It is called a continuous-time markov chain if

$$\begin{aligned} Pr[X(t+s) = j \mid X(s) = i,\ X(u) \in A_u,\ 0 \leq u < s] \\ = Pr[X(t+s) = j \mid X(s) = i] \end{aligned}$$

- Only most recent past matters
- It is called homogeneous chain if

$$Pr[X(t+s) = j \mid X(s) = i] = Pr[X(t) = j \mid X(0) = i],\ \forall s$$

- Define

$$P_{ij}(t) = Pr[X(t) = j \mid X(0) = i] = Pr[X(t+s) = j \mid X(s) = i]$$

It is the probability of going from i to j in time t

- Analogous to transition probabilities in the discrete case
- Like in the discrete case, we can show that the Markov condition implies

$$\begin{aligned} Pr[X(s) \in B_s, s \in (t, t+\tau] \mid X(t) = i, X(s'),\ 0 \leq s' < t] \\ = Pr[X(s) \in B_s, s \in (t, t+\tau] \mid X(t) = i] \end{aligned}$$

- Next we consider distribution of time spent in a state before leaving it

- By the Markov property and homogeneity we have

$$\begin{aligned} Pr\,[X(s) = i,\ s \in [t,\ t+\tau] \mid X(s') = i,\ 0 \leq s' \leq t] \\ = Pr[X(s) = i,\ s \in [t,\ t+\tau] \mid X(t) = i] \\ = Pr[X(s) = i,\ s \in [0,\ \tau] \mid X(0) = i] \end{aligned}$$

- Let $X(0) = i$ and let T_i be time spent in i before leaving it for the first time

$$\begin{aligned} Pr[X(s) = i,\ s \in [t,\ t+\tau] \mid X(s') = i,\ 0 \leq s' \leq t] \\ = Pr[T_i > t+\tau \mid T_i > t] \end{aligned}$$

$$Pr[X(s) = i,\ s \in [0,\ \tau] \mid X(0) = i] \ = \ Pr[T_i > \tau]$$

$$\Rightarrow \quad Pr[T_i > t+\tau \mid T_i > t] \ = \ Pr[T_i > \tau]$$

$\Rightarrow$ T_i is memoryless and hence exponential

- Once you transit into a state, the time spent in it is exponentially distributed.
- So, the chain can be viewed as follows
- Once you transit to a state, it spends time, say, $T_i \sim$ exponential(ν_i) in it.
- Then, when it leaves i, it transits to state j with probability, say, z_{ij}
- We would have $z_{ij} \geq 0$, $\sum_j z_{ij} = 1$. Also, $z_{ii} = 0$
- Note that $P_{ij}(t)$ is different from these z_{ij}

Example: Birth-Death process

- This is generalization of birth-death chains we saw earlier to continuous time
- From i the process can only go to $i+1$ or $i-1$
- A birth event takes it to $i+1$ and a death event takes it to $i-1$
- An example would be: $X(t)$ is number of people in a queuing system.
- A birth event would be a new person joining the queue.
- A death event would be a person leaving after finishing service

- Suppose, in state n, time till next arrival or birth event is exponential(λ_n).
- Let time till next departure or death event be exponential(μ_n)
  We assume that these two are independent
- Now, these λ_n and μ_n completely determine ν_n and z_{ij} and hence completely specify the chain
- $z_{i,i+1}$ is the probability that when the system changes state it goes to $i+1$
- Hence it is the probability that a birth event occurs before a death event.
- Let $W_1 \sim$ exponential(λ_i) and $W_2 \sim$ exponential(μ_i) be independent. Then

$$z_{i,i+1} = Pr[W_1 < W_2] = \frac{\lambda_i}{\lambda_i + \mu_i}; \quad \Rightarrow \quad z_{i,i-1} = \frac{\mu_i}{\lambda_i + \mu_i}$$

- The time spent in state i, T_i, is exponential(ν_i)
- The chain would be in state i till either a birth event or a death event occurs
- Hence, $T_i = \min(W_1, W_2)$
- Hence, $T_i \sim$ exponential($\lambda_i + \mu_i$).
- Thus, $\nu_i = \lambda_i + \mu_i$
- We have taken state space to be non-negative integers.
- Hence, $\mu_0 = 0$ and $\nu_0 = \lambda_0$ and $z_{01} = 1$

- Suppose $\lambda_n = \lambda, \ \forall n$ and $\mu_n = 0, \ \forall n$
- It is called pure birth process
- The process spend time $T_i \sim$ exponential(λ) in state i and then moves to state $i+1$
- This is the Poisson process

- Consider a queuing system
- Suppose people joining the queue is a Poisson process with rate λ
- Suppose the time to service each customer is independent and exponential with parmeter μ.
- We assume that the arrival and service processes are independent.
- Then this is a birth death process with

$$\lambda_n = \lambda, \ n \geq 0 \quad \text{and} \quad \mu_n = \mu, \ n \geq 1$$

- This is known as an $M/M/1$ queue
- A variation: $M/M/K$ queue

$$\lambda_n = \lambda, \ n \geq 0 \quad \text{and} \quad \mu_n = \begin{cases} n\mu & 1 \leq n \leq K \\ K\mu & n > K \end{cases}$$

- Consider an example of some calculations with continuous Markov chains
- Consider a Birth-Death process. Let Y_i be the time that a chain currently in i takes to reach state $i+1$ for the first time.
- We want to calculate $E[Y_i]$. (Note that $E[Y_0] = 1/\lambda_0$)
- The chain may directly go to $i+1$ or it may go to $i-1$ and then to i and then to $i+1$ or ...
- Define

$$I_i = \begin{cases} 1 & \text{if first transition out of } i \text{ is to } i+1 \\ 0 & \text{if first transition out of } i \text{ is to } i-1 \end{cases}$$

- We can find $E[Y_i]$ by conditioning on I_i.

- Time spent in i is exponential with rate $\lambda_i + \mu_i$.
- Hence, expected time till transition out of i is $1/(\lambda_i + \mu_i)$
- If this transition is to $i+1$ then that is the expected time to reach $i+1$

$$E\left[Y_i \mid I_i = 1\right] = \frac{1}{\lambda_i + \mu_i}$$

- Suppose this transition is to $i-1$.
- Then the expected time to reach $i+1$ is this time plus expected time to reach i from $i-1$ plus expected time to reach $i+1$ from i

$$E\left[Y_i \mid I_i = 0\right] = \frac{1}{\lambda_i + \mu_i} + E\left[Y_{i-1}\right] + E[Y_i]$$

- We also have

$$Pr[I_i = 1] = z_{i,i+1} = \frac{\lambda_i}{\lambda_i + \mu_i}; \quad Pr[I_i = 0] = \frac{\mu_i}{\lambda_i + \mu_i}$$

- Now we can calculate $E[Y_i]$ as

$$\begin{aligned} E[Y_i] &= Pr[I_i = 1]\; E\left[Y_i \mid I_i = 1\right] + Pr[I_i = 0]\; E\left[Y_i \mid I_i = 0\right] \\ &= \frac{\lambda_i}{\lambda_i + \mu_i}\, \frac{1}{\lambda_i + \mu_i} + \frac{\mu_i}{\lambda_i + \mu_i} \left(\frac{1}{\lambda_i + \mu_i} + E\left[Y_{i-1}\right] + E[Y_i] \right) \\ &= \frac{1}{\lambda_i + \mu_i} \;+\; \frac{\mu_i}{\lambda_i + \mu_i} \left(E\left[Y_{i-1}\right] + E[Y_i] \right) \end{aligned}$$

$$\begin{aligned} E[Y_i] \left(1 - \frac{\mu_i}{\lambda_i + \mu_i} \right) &= \frac{1}{\lambda_i + \mu_i} \;+ \frac{\mu_i}{\lambda_i + \mu_i} \left(E[Y_{i-1}] \right) \\ E[Y_i] &= \frac{1}{\lambda_i} + \frac{\mu_i}{\lambda_i} E[Y_{i-1}] \end{aligned}$$

- Thus we get

$$E\left[Y_i\right] = \frac{1}{\lambda_i} + \frac{\mu_i}{\lambda_i} E[Y_{i-1}], \; i \geq 1$$

- Since $E[Y_0] = 1/\lambda_0$, we have a formula for $E[Y_i]$
- For example,

$$E[Y_1] = \frac{1}{\lambda_1} + \frac{\mu_1}{\lambda_1\, \lambda_0}; \quad E[Y_2] = \frac{1}{\lambda_2} + \frac{\mu_2}{\lambda_2} \left(\frac{1}{\lambda_1} + \frac{\mu_1}{\lambda_1\, \lambda_0} \right)$$

- Expected time to go from i to j, $i < j$ can now be computed as

$$E[Y_i] + E[Y_{i+1}] + \cdots + E[Y_{j-1}]$$

- Note that these are only for birth-death processes

- Consider the transition probabilities, $P_{ij}(t)$
- These satisfy Chapmann-Kolmogorov equation

$$\begin{aligned} & P_{ij}(t+s) = Pr[X(t+s) = j \mid X(0) = i] \\ &= \sum_k Pr[X(t+s) = j \mid X(s) = k, X(0) = i] Pr[X(s) = k \mid X(0) = i] \\ &\quad = \sum_k Pr[X(t+s) = j \mid X(s) = k] Pr[X(s) = k \mid X(0) = i] \\ &\quad = \sum_k Pr[X(t) = j \mid X(0) = k] Pr[X(s) = k \mid X(0) = i] \\ &\quad = \sum_k P_{kj}(t)\; P_{ik}(s) \end{aligned}$$

- For finite chain, P is a matrix and $P(t+s) = P(t)\; P(s)$

- Chapmann-Kolmogorov equation gives

$$P_{ij}(t+s) = \sum_k P_{ik}(s)\; P_{kj}(t)$$

- Hence we get

$$\begin{aligned} P_{ij}(t+h) - P_{ij}(t) &= \sum_k P_{ik}(h)\; P_{kj}(t) - P_{ij}(t) \\ &= \sum_{k \neq i} P_{ik}(h)\; P_{kj}(t) - (1 - P_{ii}(h)) P_{ij}(t) \end{aligned}$$

- Define

$$q_{ik} = \lim_{h \to 0} \frac{P_{ik}(h)}{h}, \; i \neq k, \quad \text{and} \quad q_{ii} = \lim_{h \to 0} \frac{1 - P_{ii}(h)}{h}$$

- Then, assuming limit and sum can be interchanged,

$$\lim_{h \to 0} \frac{P_{ij}(t+h) - P_{ij}(t)}{h} = \sum_{k \neq i} q_{ik}\; P_{kj}(t) - q_{ii}\; P_{ij}(t)$$

- By definition, $1 - P_{ii}(h)$ is the probability that the chain that started in i is not in i at h.
- This is equivalent to there being a transition in the time h and transitions out of i occur at the rate of ν_i. Also, two or more transitions in h is $o(h)$
- Hence
$$1 - P_{ii}(h) = \nu_i h + o(h)$$
- Thus $q_{ii} = \nu_i$. It is rate of transition out of i
- We also have
$$\nu_i = q_{ii} = \lim_{h \to 0} \frac{1 - P_{ii}(h)}{h} = \lim_{h \to 0} \frac{\sum_{j \neq i} P_{ij}(h)}{h} = \sum_{j \neq i} q_{ij}$$

- By definition, $P_{ij}(h) = q_{ij} h + o(h), \ i \neq j$
- Hence q_{ij} is the rate at which transitions out of i into j are occurring.
- Transitions out of i occur with rate ν_i and z_{ij} fraction of these are into j
- Hence, $q_{ij} = \nu_i \, z_{ij}, \ i \neq j$
- Thus, we got
$$\nu_i = \sum_{j \neq i} q_{ij}, \quad z_{ij} = \frac{q_{ij}}{\sum_{j \neq i} q_{ij}}, \quad q_{ii} = \sum_{j \neq i} q_{ij}$$
- The $\{q_{ij}\}$ are called the infinitesimal generator of the process.
- A continuous time Markov Chain is specified by these q_{ij}

- Consider a Birth-Death process.
- We got earlier
$$\nu_i = \lambda_i + \mu_i, \quad z_{i,i+1} = \frac{\lambda_i}{\lambda_i + \mu_i}, \quad z_{i,i-1} = \frac{\mu_i}{\lambda_i + \mu_i}$$
- Now we can calculate q_{ij}
$$q_{i,i+1} = (\lambda_i + \mu_i) \, \frac{\lambda_i}{\lambda_i + \mu_i} = \lambda_i, \quad q_{i,i-1} = (\lambda_i + \mu_i) \, \frac{\mu_i}{\lambda_i + \mu_i} = \mu_i$$
- This is intuitively obvious
- We specify a birth-death chain by birth rate (rate of transition from i to $i+1$), λ_i and death rate (rate of transition from i to $i-1$), μ_i.

- The Chapmann-Kolmogorov equations give us
$$P_{ij}(t+h) - P_{ij}(t) = \sum_{k \neq i} P_{ik}(h) \, P_{kj}(t) - (1 - P_{ii}(h)) P_{ij}(t)$$
- Using this we derived
$$P'_{ij}(t) = \sum_{k \neq i} q_{ik} \, P_{kj}(t) - q_{ii} \, P_{ij}(t)$$
Called Kolmogorov Backward equation
- We can solve these ODEs to get $P_{ij}(t)$
- For a birth-death chain the equation becomes
$$P'_{ij}(t) = \lambda_i P_{i+1,j}(t) + \mu_i P_{i-1,j}(t) - (\lambda_i + \mu_i) P_{ij}(t)$$

Poisson process as a special case

- Consider the case: $\lambda_i = \lambda$ and $\mu_i = 0.\ \forall i$.
- This would be a Poisson process with rate λ.
- Taking $i = 0$, the differential equation becomes

$$P'_{0j}(t) = \lambda P_{1j}(t) - \lambda P_{0j}(t)$$

- $P_{0j}(t)$ is the probability of j events in an interval of length t which is same as what we had called $P_j(t)$.
- Similarly, $P_{1j}(t)$ is same as what we called $P_{j-1}(t)$ there
- Now one can see that the above ODE is what we got for Poisson process.

- Consider a two-state Birth-Death chain.
- We would have $\mu_0 = \lambda_1 = 0$. Let $\lambda_0 = \lambda$ and $\mu_1 = \mu$
- The two states can be a machine working or failed.
- λ is rate of failure. Time till next failure is exponential(λ)
- μ is rate of repair. Time for repair is exponential(μ)
- We may want to calculate $P_{00}(T)$, the probability that the machine would be working at a time T units later given it is in working condition now
- We can calculate it by solving the ODE's

$$P'_{ij}(t) = \lambda_i P_{i+1,j}(t) + \mu_i P_{i-1,j}(t) - (\lambda_i + \mu_i) P_{ij}(t)$$

- For the two state chain, these equations are

$$\begin{aligned}
P'_{00}(t) &= \lambda_0 P_{10}(t) - \lambda_0 P_{00}(t) \\
P'_{01}(t) &= \lambda_0 P_{11}(t) - \lambda_0 P_{01}(t) \\
P'_{10}(t) &= \mu_1 P_{00}(t) - \mu_1 P_{10}(t) \\
P'_{11}(t) &= \mu_1 P_{01}(t) - \mu_1 P_{11}(t)
\end{aligned}$$

- As is easy to see, we get a system of equations like this for any finite chain.
- Solving these we can show

$$P_{00}(t) = \frac{\lambda}{\lambda + \mu} e^{-(\lambda+\mu)t} + \frac{\mu}{\lambda + \mu}$$

- Consider a finite chain
- Then the transition probabilities can be represented as a matrix
- The Chapmann-Kolmogorov equation gives

$$P(t+s) = P(t)\ P(s)$$

- Differentiating the above with respect to t

$$P'(t+s) = P'(t) P(s)$$

- Putting $t = 0$ in the above we get

$$P'(s) = P'(0)\ P(s) = \bar{Q}\ P(s), \quad \text{where} \quad \bar{Q} = P'(0)$$

- The solution for this is

$$P(t) = e^{t\bar{Q}}, \quad \text{because} \quad P(0) = I$$

- This is the expression for calculating $P_{ij}(t)$ for any t and i, j

- Let us examine the matrix $\bar{Q} = [\bar{q}_{ij}]$

$$\bar{Q} = P'(0) = \lim_{h \downarrow 0} \frac{P(h) - P(0)}{h} = \lim_{h \downarrow 0} \frac{P(h) - I}{h}$$

- This gives us

$$\begin{aligned} \text{for } \ k \neq j, \ \ \bar{q}_{kj} &= \lim_{h \downarrow 0} \frac{P_{kj}(h) - 0}{h} = q_{kj} \\ \bar{q}_{jj} &= \lim_{h \downarrow 0} \frac{P_{jj}(h) - 1}{h} = -q_{jj} = -\nu_j \end{aligned}$$

- Thus this $\bar{Q}$ matrix has q_{ik} as off-diagonal entries and $-q_{jj}$ as diagonal entries
- So, each row here sums to zero
- We normally write it as Q and call it the infinitesimal generator of the process

- The Kolmogorov backward equation is

$$P'_{ij}(t) = \sum_{k \neq i} q_{ik} \ P_{kj}(t) - q_{ii} \ P_{ij}(t)$$

- The above can be written in a matrix form

$$P'(t) = QP(t)$$

- The off-diagonal entries of Q are q_{ik} and diagonal entries are $-q_{ii}$
- From the above equation, $P'(0) = Q$
- So, what we did is to write the backward equation in matrix form

- For the backward equation, we started with

$$P_{ij}(t+h) = \sum_{k} P_{ik}(h) \ P_{kj}(t)$$

- The Chapmann-Kolmogorov equation also gives us

$$P_{ij}(t+h) = \sum_{k} P_{ik}(t) \ P_{kj}(h)$$

- Similar algebra as earlier gives us

$$P'_{ij}(t) = \sum_{k \neq j} P_{ik}(t) \ q_{kj} - q_{jj} \ P_{ij}(t)$$

  (under some assumptions about interchanging limit and summation)
- This is known as Kolmogorov forward equation
- For finite chains, both forward and backward equations are same
- For infinite chains there are some differences

- We can define transient and recurrent states as in the discrete case.
- However, we need to be careful about defining hitting times or first passage times
- We define

$$T_i = \min\{t > 0 \ : \ X(t) \neq i\} \qquad f_i = \min\{t \ : \ t > T_i, \ X(t) = i\}$$

- For a chain started in i we take f_i as first return time to i
- A state i is said to be
  - Transient if $Pr[f_i < \infty \mid X(0) = i] < 1$
  - Recurrent if $Pr[f_i < \infty \mid X(0) = i] = 1$

- Most of the other definitions are also similar to the case of discrete chains
- The chain is said to be irreducible if for all i, j there is a positive probability of going from i to j in some finite time: $P_{ij}(t) > 0$ for some t
- A recurrent state is positive recurrent if mean time to return is finite: $E[f_i \mid X(0) = i] < \infty$
  Otherwise it is null recurrent
- An irreducible positive recurrent chain would have a unique stationary distribution
- There is no concept of periodicity in the continuous time case
- An irreducible positive recurrent chain would be called an ergodic chain

- Define

$$\pi_j(t) = Pr[X(t) = j] = \sum_i \pi_i(0) P_{ij}(t)$$

  This also analogous to the discrete case
- The above equation is true for general infinite chains.
- In the finite case, we can get a more compact expression
- For a finite chain, taking π as a row vector,

$$\pi(t) = \pi(0)\ P(t) = \pi(0)\ e^{Qt}$$

- We say π is a stationary distribution if

$$\pi(0) = \pi \quad \Rightarrow \quad \pi(t) = \pi,\ \forall t$$

- Hence, if we start the chain in the stationary distribution, $\pi'(t) = 0$
- We get from the earlier equation

$$\pi_j(t) = \sum_i \pi_i(0) P_{ij}(t) \quad \text{and hence} \quad \pi'_j(t) = \sum_i \pi_i(0) P'_{ij}(t)$$

- Using the forward equation for $P'_{ij}(t)$

$$\begin{aligned} \sum_i \pi_i(0) \left(\sum_{k \neq j} q_{kj} P_{ik}(t) - q_{jj} P_{ij}(t) \right) &= 0 \\ \Rightarrow \quad \sum_{k \neq j} q_{kj} \pi_k - \pi_j \sum_{k \neq j} q_{jk} &= 0 \end{aligned}$$

  when π is a stationary distribution and $\pi(0) = \pi$

- What we showed is that any stationary distribution π has to satisfy

$$\sum_{k \neq j} q_{kj} \pi_k = \pi_j \sum_{k \neq j} q_{jk}$$

- We can interpret this (as we did in discrete case)
- q_{kj} is the rate of transition from k to j and π_k is the fraction present in k.
- Hence $\sum_{k \neq j} q_{kj} \pi_k$ is the net flow into j
- $\pi_j \sum_{k \neq j} q_{jk}$ is the net flow out of j
- At steady state the flows have to be balanced
- The above equation is known as a balance equation