Random Processes
ECE537

E. S. Sousa
Sept 3, 2024

Probability

Probability theory was motived in the classical times in order to study gambling schemes. The theory is based on a model of creating experiments with random outcomes such as flipping a coin, drawing a card from a deck, rolling a die, or many other gambling games. The concept was then generalized to the performing of any experiment, or measurement of random phenomenon, such as turning on a noise generator and observing a waveform, or sampling the value of the temperature at a particular location and point in time.

Consider the game of drawing a card at random from a shuffled 52-card deck. We define **events** as results of the draw such as "a queen occurred", "the 3 of clubs occurred", "a red card occurred", etc. Each time a draw is made we call it a trial in the experiment. Consider the experiment of drawing a card and consider a large number of repeated trials where the card is drawn at random from the full deck. Let the event "a queen" be denoted by Q. For each trial if Q occurs then we consider it a success. According to one of the classical definitions of probability (the frequency definition), the probability of the event Q is then the limit, as the number of trials approaches infinity, of the ratio of the number of successes to the total number of trials. In other words, we can define the probability of drawing a queen as the ratio of the number of trial outcomes that result in a queen to the total number of trials, as the number of trials approaches infinity. This is an old definition of probability, i.e. the definition based on frequency.

The modern definition of probability, on the other hand, is not based on this ***frequency*** interpretation, but instead is based on an abstract axiomatic model for the study of the results of the experiment. This model consists of a triplet of mathematical objects that we denote as $(\Omega, \mathcal{F}, P)$, where (**Assuming a finite set of possible outcomes**)

- Ω is a set of points where each point represents a possible ***outcome*** of the experiment. In the case where the experiment is flipping-a-coin then there are two possible outcomes denoted as H and T and we have $\Omega = \{H, T\}$. In the case of drawing a card from a 52-card deck then Ω is a set of 52 points corresponding to the 52 cards, etc.
- $\mathcal{F}$ is a collection of subsets of Ω. It is a set whose elements are subsets of Ω. Each of the elements of $\mathcal{F}$ (a subset) is called an ***event***. Note that it is not necessary that all subsets of Ω are events, i.e. elements of $\mathcal{F}$. However in cases where Ω is a finite set then typically all subsets of Ω are events. In this case the number of events is equal to the number of subsets of Ω which is 2^n, where n is the number of points in Ω. This is also called the **power set** of Ω.

So now we know that $\mathcal{F}$ is a set of elements where each element (or each point in the set) is called an event and is some subset of Ω. But there is more. The elements of $\mathcal{F}$ must satisfy a certain number of conditions as follows:

i) the set Ω is an element in $\mathcal{F}$. (Note that we are thinking of Ω as a subset of Ω).
ii) If an event $E \in \mathcal{F}$, then E^c (the complement) is also in the set $\mathcal{F}$.
iii) if $E_1 \in \mathcal{F}$ and $E_2 \in \mathcal{F}$, then we must have the union $E_1 \cup E_2$ also in the set $\mathcal{F}$, or $E_1 \cup E_2 \in \mathcal{F}$

We can also show by using De Morgan's laws that as a consequence of the conditions ii) and iii) the following proposition follows.

Proposition: If $E_1 \in \mathcal{F}$ and $E_2 \in \mathcal{F}$, then the intersection $E_1 \cap E_2 \in \mathcal{F}$.

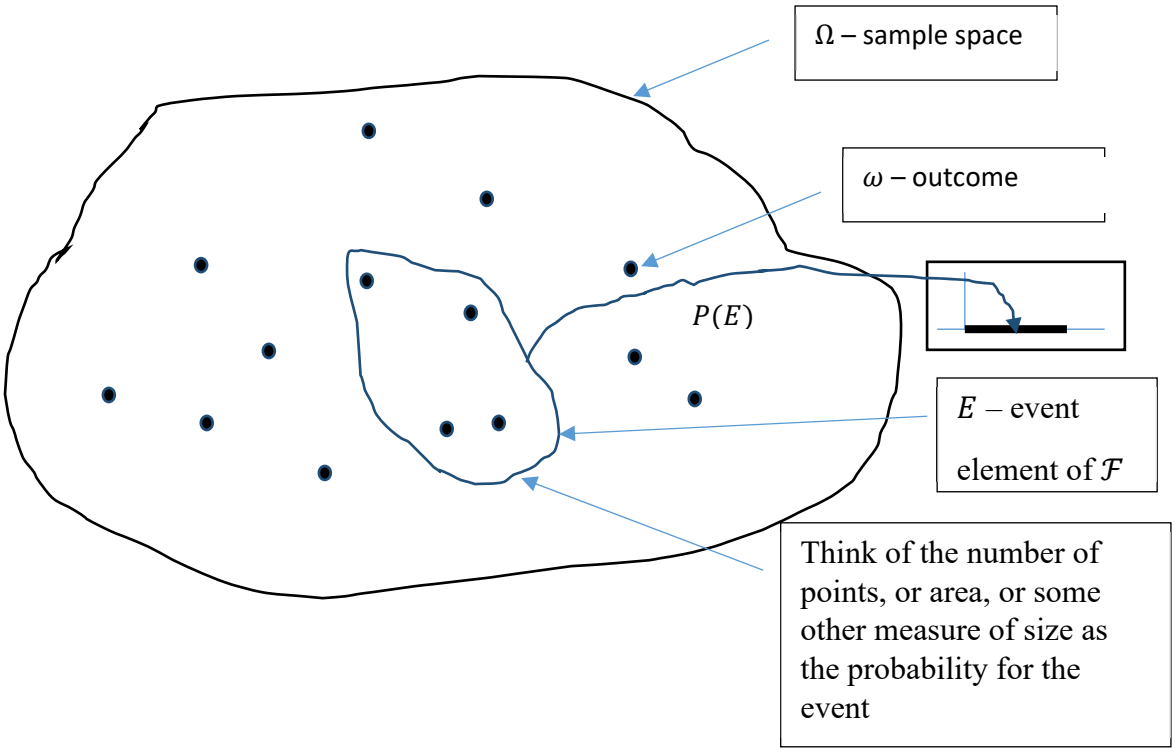
This proposition can be proved using DeMorgan's Laws (see below). This means that the requirement for the set of events $\mathcal{F}$ in a probability model is that it is closed under complement operations and either closed under union operations or closed under intersection operations. One of these two requirements (unions or intersections) implies the other.

If $\mathcal{F}$ is the set of all subsets of Ω then it is obvious that the above three conditions are satisfied. But if $\mathcal{F}$ is not the set of all subsets then we must check that the above three conditions are satisfied in order to certify that indeed $\mathcal{F}$ is a proper set of events for our probability model. $\mathcal{F}$ with the above three conditions is also called a ***field*** or an ***algebra***.

The third object in our probability model

- P is a function from the set $\mathcal{F}$ to the set $[0,1]$, (real interval) or we can write P: $\mathcal{F} \Rightarrow [0,1]$. For each element $E \in \mathcal{F}$, i.e. each event, we assign a number in the interval $[0,1]$ called the probability of E, e.g. $P(E) = 0.2$. But there are also conditions that this function must satisfy. It must satisfy the following properties:
  i) If $E = \Omega$ then $P(E) = 1$.
  ii) If $E_1 \in \mathcal{F}$, and $E_2 \in \mathcal{F}$, and $E_1 \cap E_2 = \phi$, the empty set, then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$. In this case we call the events E_1 and E_2 mutually exclusive.

We can represent the probability model $(\Omega, \mathcal{F}, P)$ by the following Figure



Example 1:

Experiment: Flip a fair coin

$$\Omega = \{H, T\}$$

$$\mathcal{F} = \{\phi, \{\mathrm{H}\}, \{\mathrm{T}\}, \{\mathrm{H}, \mathrm{T}\}\}$$

$P(\phi) = 0$, $P(\{H\}) = 1/2$, P({T})=1/2, $P(\{H, T\}) = 1$.

Note that for simplicity we usually also write $P(H)$ instead of the proper $P(\{H\})$. We refer to an event such as $\{H\}$ as a singleton event because the event consists of only one outcome, i.e. one point in Ω.

Example 2:

Draw a card at random from a deck of 52 cards.

$\Omega = \{1C, 2C, \ldots 10C, JC, QC, KC, 1D, 2D, \ldots, 10D, JD, QD, KD, \ldots.\}$ – 52 points in the set.

$\mathcal{F} =$ a set of 2^{52} events (all possible subsets of Ω). There are 2^{52} events! That is, the cardinality of the set $\mathcal{F}$ is 2^{52}.

Of all the 2^{52} possible events there are 52 that we can call singleton-events, i.e. one point in the set. If a card is chosen at random we usually assume that these 52 events have equal probabilities

- $1/52$. Using this information and the properties in the axioms we can determine the probabilities for all the 2^{52} events. The possible values that can arise are 0, 1/52, 2/52, 3/52, … 52/52 = 1. An event E is a subset of the set of 52 points and the probability of this event is the number of points in the subset divided by 52. Note that the assumption of equal probability for the singleton events is not derived mathematically, but is only an assumption that we use based on symmetry and assuming a well shuffled deck. Under this assumption we have the following:

Examples: P("3 of clubs")$=1/52$, P("a queen")$=4/52 = 1/13$, P("a spade")$=13/52=1/4$, P("a black king")$=2/52 = 1/26$, P("a red card")$=26/52=1/2$, P("a king or a red ace") $= \frac{6}{52} = 3/26$.

Example 3:

Open an English book at random, e.g. a novel, and strip all punctuation marks, treat upper and lowercase the same. Go to page 20, 17th letter. Then $\Omega = \{a, b, c, d, \ldots., z\}$.

Consider the singleton events {a}, {b}, {c}, … For simplicity refer to these as $a, b, c, ...$

These events do not have the same probability because the structure of English is such that some letters occur much more frequently than others. In studies we have found that the probability of these singleton events is approximately as follows:

$P(a) = 0.0817, P(b) = 0.01492, \ldots P(e) = .12702, ..., P(z) = 0.00074.$

Note that we can also determine P(a letter in the word $Canada$), etc. This would be computed as $P(\{a\} \cup \{c\} \cup \{n\} \cup \{d\}) = P(a) + P(c) + P(n) + P(d)$ since the events $\{a\}, \{c\}, \{n\}, \{d\}$ are mutually exclusive.

More Abstract Approach to Probability

- universal set (**possible outcomes**),
- subsets of the universal set (**events**),
- **operations** on the subsets, or events, (complement, union, intersection),
- **fields** (set operations along with the concept of **closure**),
- $\boldsymbol{\sigma}$**-fields** (set of operations with **infinite operands**),
- probability measures, or **probabilities** (mapping of events to the set [0,1]).

Set operations

Universal set: Ω

Complement $A^c = \Omega \backslash A$, i.e. $\{x \in \Omega: x \notin A\}$

Union: $A \cup B = \{x \mid x \in A$, or $x \in B\}$

Intersection: $A \cap B = \{x \mid x \in A$ and $x \in B\}$

Difference: $A \backslash B = A \cap B^c$

Empty set $\phi = \{\}$

De Morgan's Laws

$(A_1 \cap A_2)^c = A_1^c \cup A_2^c$ (DM1)

$(A_1 \cup A_2)^c = A_1^c \cap A_2^c$ (DM2)

We can also write the above as

$A_1 \cap A_2 = (A_1^c \cup A_2^c)^c$ - An intersection can be written in terms of a union and complements

$A_1 \cup A_2 = (A_1^c \cap A_2^c)^c$ - A union can be written in terms of an intersection and complements.

Field (also called an Algebra)

Consider a set of outcomes Ω.

A set of subsets of Ω, $\mathcal{F}$, is called a **field** or an **algebra** if the following two conditions hold

(i) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$
(ii) If $A_1 \in \mathcal{F}$ and $A_2 \in \mathcal{F}$, then $A_1 \cup A_2 \in \mathcal{F}$.

Properties:

1. If $\mathcal{F}$ is a field then $\phi \in \mathcal{F}$, and $\Omega \in \mathcal{F}$
2. If $A_1 \in \mathcal{F}$ and $A_2 \in \mathcal{F}$, then $A_1 \cap A_2 \in \mathcal{F}$

The above can easily be proved.

For 2. We use De Morgan's Laws, DM1.

Using DM1 and complementing both sides: $A_1 \cap A_2 = (A_1^c \cup A_2^c)^c$

Hence

(iii) $A_1 \cap A_2 \in \mathcal{F}$

For 1. Assume that $A \in \mathcal{F}$, then by (i) $A^c \in \mathcal{F}$ and by (iii) $A \cap A^c \in \mathcal{F}$. But $A \cap A^c = \phi$, hence $\phi \in \mathcal{F}$. Also by ii) $A \cup A^c \in \mathcal{F}$. But $A \cup A^c = \Omega$. Also, we can say by (i) $\phi^c = \Omega \in \mathcal{F}$.

Example (Field? or not a Field?)

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

- $\mathcal{F}_1 = \{\phi, \{2,4,6,\}, \{1,3,5\}, \Omega\}$: This is a field

- $\mathcal{F}_2 = \{\phi, \{2,4,6\}, \{1,3,5\}\}$: This is not a field. Property (i) does not hold
- $\mathcal{F}_3 = \{\phi, \{2,4,6\}, \{1,3,5\}, \{1,2,3\}, \{4,5,6\}, \Omega\}$: This is not a field. Property (ii) does not hold.
- How about $\mathcal{F}_4 = \{\phi, \{1,2,3,4\}, \{5,6\}, \Omega\}$?

Can you add some more events (i.e. subsets of Ω) to make $\mathcal{F}_3$ a field? If so, we would call it an extended field of the set $\mathcal{F}_3$.

What is the size of this extended field?

Sigma Field (Operations on an infinite number of events or subsets of Ω)

Definition: A set of subsets of Ω, $\mathcal{F}$ is a **sigma field** (or **σ-field**) if the following two conditions hold:

(i) If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.
(ii) If $A_i \in \mathcal{F}$, $i = 1,2,3,\cdots$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Note that the difference here in comparison to a field is that the set $\mathcal{F}$ is closed under countably infinite unions. By De Morgan's Laws $\mathcal{F}$ is also closed under countable intersections. This follows because De Morgan's laws also apply to countably infinite unions and intersections.

Note that the concept of sigma field is also called a **sigma algebra** in some textbooks. The text book by Leon-Garcia uses the term "**sigma field**", but the textbook by Rosenthal uses the term "sigma algebra", or σ-algebra.

Example 1:

Consider $X =$ the number facing up when rolling a die. $\Omega = \{1,2,3,4,5,6\}$. X is a random variable.

Let $A = \{X < 5\} = \{1,2,3,4\}$
Let $\mathcal{F}_1 = \{\phi, \{1,2,3,4\}, \{5,6\}, \Omega\} = \{\phi, A, A^c, \Omega\}$.
Then $\mathcal{F}_1$ is a Field. It is the smallest Field containing A. Note that $\mathcal{F}_1$ is also a σ-field. That is, some fields are also σ-fields. But there are fields that are not σ-fields, i.e. the property of closure under infinite unions does not hold.

Example of a field which is not a σ-field

Consider $\Omega = \mathcal{R}$ (set of real numbers)
Consider the case (set of certain subsets of $\mathcal{R}$) $\mathcal{F}_I =$
$\{A \mid A\ is\ the\ union\ of\ a\ finite\ set\ of\ intervals, closed\ or\ open\ intervals\ including\ semi-infinite, e.g.\ (a, \infty)\}$. Then we can show that F_I is a field, but it is not a σ-field because it is not

closed under infinite unions. For example $(1,2) \cup (3,4) \cup (5,6) \cup \ldots \cup (2n-1, 2n) \cup \cdots$ is a set that is not in $\mathcal{F}_I$ because it is not the union of a finite set of intervals.

Example 2:

Consider $\Omega = \{1,2,3,\cdots\}$, set of Natural Numbers
$\mathcal{F}_2 = \{A | \, A \subset \Omega\}$
Then $\mathcal{F}_2$ is a σ-field. In fact if we take a collection of sets $\mathcal{C}$ which is the power set of Ω then it is a σ-field. We use this σ-field often.

Example 3:

Consider $\Omega = [a, b]$, a closed interval of the set of real numbers
$\mathcal{F}_3 = \{A \mid A \subset \Omega\}$
Then $\mathcal{F}_3$ is a σ-field. Note that $\mathcal{F}_3$ is the power set of Ω.

Note that this σ-field is not commonly used in probability models. In practice in many of the models that work with, we would consider a subset of all the subsets of Ω.

Probability Measure

Now we introduce the idea of a probability measure on a σ-field $\mathcal{F}$. This is a mapping of $\mathcal{F}$ to the interval $[0,1]$. For $E \in \mathcal{F}$, $P(E) \in [0,1]$. $P(\cdot)$ must satisfy the following conditions

(i) $P(\Omega) = 1$

(ii) For any countable set of elements in $\mathcal{F}$(events) E_i, $i = 1,2,3,\cdots \in \mathcal{F}$, with $E_i \cap E_j = \phi$, for $i \neq j$. (i.e. pairwise disjoint sets) we have

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

For Example 1 above we can take the 4 elements of $\mathcal{F}$ and define $P(A) = p$, and $P(A^c) = 1 - p$, where p is any element of $[0,1]$.

In the case of Example 2 we could take the singleton sets $S_i = \{i\}, i = 1,2,3,\cdots$ and define $P(S_i) = p_i$ for any set of numbers $p_i \in [0,1]$ such that $\sum_{i=1}^{\infty} p_i = 1$. Then we use the property (ii) to define $P(E)$ for any $E \in \mathcal{F}$.

For the σ-field in Example 3 we would like to define a popular probability measure, the so-called **uniform probability measure**, where the probability for the event $(x_1, x_2) \in \mathcal{F}_3$ $(a < x_1 < x_2 < b)$ is defined as $P\big((x_1, x_2)\big) = \frac{x_2 - x_1}{b-a}$. However, we can show that it is **not possible** to define the function $P(E)$, i.e. extend the definition, probability measure, for every element of $\mathcal{F}_3$, i.e. every subset, E, of $\Omega = [\mathrm{a}, \mathrm{b}]$. This is shown (proof) in the textbook by Rosenthal. In this case in order to be able to define a probability measure on a sigma field consisting of subsets of Ω, and

containing all intervals, we must restrict the set of subsets, $\mathcal{F}$, to a smaller set than the set of all subsets, i.e. a proper subset of the set of all subsets of Ω. In other words we must define a smaller σ-field.

Subsets of Ω that are excluded in the creation of this σ-field, $\mathcal{F}$, are called non-measureable sets. Non-measureable sets are difficult to specify but they exist. Rosenthal gives an example.

σ-Fields where Ω is the real line, or an interval of the real line, e.g. $\Omega = (0, 1)$

In many cases we would like to define σ-fields that include elements that are intervals of the real line. Then we enforce the conditions for a σ-field, i.e. closure under complements, countable unions, and countable intersections, in order to obtain a field that includes these "important" intervals. Consider the following examples, as attempts to create a σ-field, starting with intervals. Note that in this space Ω, intervals are very important as events.

Consider $\Omega = (0,1)$ and attempt to create a σ-field that includes the set of elements (i.e. events) $A_{ab} = \{x \in \Omega \mid 0 < a < x < b < 1\} = (a, b)$.

Let $\mathcal{A} = \{A_{ab} \mid$ for all a and b such that $0 < a < b < 1\}$. Events are open intervals.

This is not a σ-field. It is not even a field because it is not closed under unions or complements. So if we are going to create a σ-field we need more elements.

Let $\mathcal{I} = \{A_{ab}$ together with all intersections and unions of these intervals, where $0 < a < b < 1\}$. Again, this is also not a σ-field because it is not closed under complements and also not closed under countable intersections. For example, let $A_n = \left(\frac{1}{2} - \frac{1}{n}, \frac{1}{2} + \frac{1}{n}\right)$. Then $\cap_{n=1}^{\infty} A_n = \left\{\frac{1}{2}\right\}$, i.e. a singleton set which is not in $\mathcal{I}$.

Consider Ω as above and consider the set of all open intervals $\mathcal{A}$. We wish to find the smallest σ-field that contains $\mathcal{A}$, i.e. we wish to extend the set $\mathcal{A}$ to a σ-field $\mathcal{F}$. By using countable intersections then we can show that all singleton sets must be included. For example, if $\alpha \in \Omega$, then $\{\alpha\} \in \mathcal{F}$. We can show that if $\alpha \in \Omega$ then there exists an N large enough such that $\alpha - \frac{1}{N} > 0$, and $\alpha + \frac{1}{N} < 1$. We can then show that $\cap_{n=N}^{\infty} \left(\alpha - \frac{1}{N}, \alpha + \frac{1}{N}\right) = \{\alpha\}$. We can also show that all semi-closed intervals are included. For example $(a, b] = (a, b) \cup \{b\}$. Then we can take all countable unions and intersections of open, semi-closed, or closed intervals. These must all be in $\mathcal{F}$. The smallest σ-field containing all of these sets is called a **Borel σ-Field**, or for simplicity a

Borel Field.

Axioms of Probability and Resulting Propositions

Start with a Probability Space, or Probability Triplet: $(\Omega, \mathcal{F}, P)$

- Ω is a set of outcomes $\omega \in \Omega$
- $\mathcal{F}$ is a σ-field with elements being subsets of Ω, not necessarily all such subsets.
- $P(.)$ is a probability measure defined on $\mathcal{F}$ with the following properties:
    - $P(A) \geq 0$ for any event A.
    - $P(\Omega) = 1$ (Normalization)
    - $P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$, if $A_i \cap A_j = \phi$ for any $i \neq j$.

Propositions Following From these Axioms of Probability

Proposition 1.1: $P(A^c) = 1 - P(A)$

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

Proposition 1.2: $P(A) \leq 1$

Proof: $P(A) = 1 - P(A^c) \leq 1$

Proposition 1.3: $P(\phi) = 0$.

Proof: $P(\Omega) = P(\Omega \cup \phi) = P(\Omega) + P(\phi)$, hence $P(\phi) = 0$.

Proposition 1.4: If $A_1, A_2, A_3, \cdots A_n$ are pairwise mutually exclusive, i.e. $A_i \cap A_j = \phi$, then

$$P(\cup_{n=1}^{N} A_n) = \sum_{n=1}^{N} P(A_n)$$

This follows from the countable additivity by setting $A_n = \phi$, for $n > N$. In a sense we convert an infinite sequence of events to a finite sequence of events.

Proposition 1.5 (Union bound, or sub-additivity)

If $A_1, A_2, \ldots, A_n \in \mathcal{F}$ then

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

This can be proved as follows

Let

$B_1 = A_1$

$B_2 = A_2 \backslash (A_2 \cap B_1)$

$B_3 = A_3 \backslash (A_3 \cap B_2) \backslash (A_3 \cap B_1)$

….

$B_n = A_n - (A_n \cap B_{n-1}) - \cdots - (A_n \cap B_1)$ (note that we also write $A - B$, for A/B)

Then we can show that the B_i are mutually exclusive, i.e. $B_i \cap B_j = \phi$ for $i \neq j$, and $\cup_{i=1}^{n} A_i = \cup_{i=1}^{n} B_i$. Hence, it is clear that $P(\cup_{i=1}^{n} A_i) = P(\cup_{i=1}^{n} B_i) = \sum_{i=1}^{n} P(B_i)$. But it is clear that $P(B_i) \leq P(A_i)$. Hence the result follows.

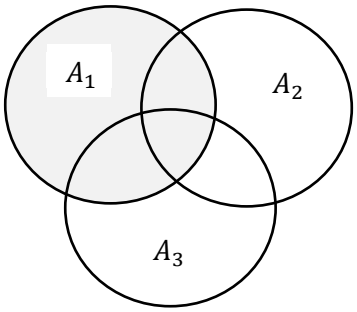
Proposition 1.6: (Inclusion-Exclusion)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Proposition 1.7: Inclusion-Exclusion for n sets

$$P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) - \sum_{1 \leq i_1 < i_2 \leq n} P\left(A_{i_1} \cap A_{i_2}\right) + \sum_{1 \leq i_1 < i_2 < i_3 \leq n} P\left(A_{i_1} \cap A_{i_2} \cap A_{i_3}\right) -$$

$$\ldots + (-1)^{n+1} P(A_1 \cap A_2 \cap \ldots \cap A_n)$$



The above can be proved by induction.

$P(\cup_{i=1}^{n+1} A_i) = P((\cup_{i=1}^{n} A_i) \cup A_{n+1}) = P(((\cup_{i=1}^{n} A_i)) + P(A_{n+1}) - P(((\cup_{i=1}^{n} A_i) \cap A_{n+1})$

$= P(((\cup_{i=1}^{n} A_i)) + P(A_{n+1}) - P((\cup_{i=1}^{n} (A_i \cap A_{n+1}))$ …..

Use induction on the first and third of the above terms and then simplify …

Proposition 1.8 (Monotonicity)

If $A \subset B$ then $P(A) \leq P(B)$

Proof: Note that $A \cup (B/A) = B$.
$P(A) + P(B/A) = P(B)$ A and B/A are mutually exclusive
Since $P(B/A) \geq 0$, then $P(A) \leq P(B)$.

Definition 1.7: Probability Space

$(\Omega, \mathcal{F}, P)$ is called a probability space, or a probability triplet.

Note that for a given Ω, whether it is finite, countably infinite, non-countably infinite, we can define many σ-fields $\mathcal{F}$, and

For a given $(\Omega, \mathcal{F})$ we can define many probability measures, or probability laws, $P(.)$.

Defining Probabilities in the Abstract Probability Model (An Aside)

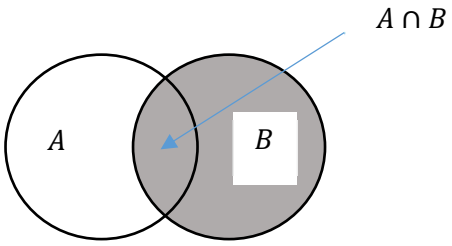
> In an abstract probability model we can define the probability function in any way that we wish as long as the axioms are satisfied. For example, if the sample space Ω is finite or countably infinite then we can define elementary events as follows: A non-empty event $\boldsymbol{E}$ is elementary if there is no other non-empty event $\boldsymbol{E'}$ such and $\boldsymbol{E'} \subset \boldsymbol{E}$. In other words there are no events that are proper subsets of $\boldsymbol{E}$. In the case that the set of events $\mathcal{F}$ is the set of all subsets of Ω then the elementary events are the singleton sets. Consider the set of elementary events $\mathcal{E}$. Then an arbitrary event is a union of some of the events in $\mathcal{E}$. Let the set of elementary events be $\mathcal{E} = \{\boldsymbol{E_1}, \boldsymbol{E_2}, \boldsymbol{E_3} \cdots\}$. Then an arbitrary probability function can be defined by specifying a set of positive numbers $p_1, p_2, p_3, \cdots$ with $\sum_k p_k = 1$, to be the probabilities of the events, i.e. $P(\boldsymbol{E_k}) = p_k$.
>
> In many practical applications we define an experiment and as a result of considerations of symmetry (e.g. rolling a die, drawing a card from a deck, spinning a wheel) we invoke the assumption that the probabilities of certain elementary events are all equal. The determination of the probability of an event E then reduces to the determination of the number of points in the set E, i.e. counting the elements in E, and taking the ratio to the number of elements in Ω. In other cases we use properties of the experiment and the concept of independence, to be discussed later to determine the probably of the different events, i.e. determining the probability function.

Conditional Probability

We consider two events A, B in a probability model. When the experiment is done we have different possibilities such as neither A, nor B occurs, or one of the two events occur, or both of the events occur. We can of course determine $P(A)$ and $P(B)$ because these are defined by the model. Now suppose we are given the information that event B has occurred. What can we say about whether or not the event A also occurred? The probability that A also occurred given that we are told that B has occurred is called the conditional probability of A given B and is written as $P(A/\mathrm{B})$. If we are told that B has occurred then we consider all the outcomes in the set B as sort of a new sample space and given this, the probability of A is actually the probability of $A \cap B$

normalized by the probability of B. Hence, we define the conditional probability of the event A given that the event B has occurred as

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



Note that we do an experiment and are not told what is the outcome, $\omega \in \Omega$. Rather, we are told that a certain event B occurred, i.e. we are told that $\omega \in B$, i.e. the outcome is in B. So in a sense we assume a restricted probability space where we replace Ω by B and we now consider a new σ-field $\mathcal{F}_B = \{E_B | E_B = E \cap B \text{ for some } E \in \mathcal{F}\}$. That is, the events in $\mathcal{F}_B$ are obtained from the events in $\mathcal{F}$ by performing an intersection with the event B. Then we define a new probability measure on B called P_B as follows: $P_B(E_B) = P(E \cap B)/P(B)$.

We start with the triplet $(\Omega, \mathcal{F}, P)$, the original probability triplet, and obtain the new triplet $(B, \mathcal{F}_B, P_B)$ which is called the conditional probability given B.

Example 1:

Suppose we roll a die and are told that the outcome is an even number. What is the probability that it is a number less or equal to 4?

$$B = \{2,4,6\}$$
$$A = \{1,2,3,4\}$$
$$P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{2,4\})}{P(\{2,4,6\})} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

Example 2:

Draw a card from a 52-card deck. What is the probability of "a queen" given that "a red card" occured?

$$P(\text{a queen}/a \text{ red card}) = \frac{P(\text{a queen \& a red card})}{P(\text{a red card})} = \frac{\frac{2}{52}}{\frac{26}{52}} = 1/13.$$

Note that $P(\text{a red card}/\text{a queen}) = 1/2$.

The Law of Total Probability

Let the set of events $\{A_1, A_2, \dots, A_n\}$ be a partition of the set Ω. This means that $\Omega = \cup_{i=1}^{n} A_i$ and $A_i \cap A_j = \phi$.

Theorem (Total probability)
Let $\{A_1, A_2, \dots, A_n\}$ be a set of events that partition Ω.
Then for any event $B \in \mathcal{F}$

$$P(B) = P(B/A_1)P(A_1) + \cdots + P(B/A_n)P(A_n)$$

This follows because we can write $B = \cup_{i=1}^{n} B \cap A_i$. And since the sets $B \cap A_i$ are mutually exclusive, then $P(B) = \sum_{i=1}^{n} P(B \cap A_i)$. But $P(B/A_i) = P(B \cap A_i)/P(A_i)$. The result follows.

Baye's Rule.

In many experiments we want to compute the conditional probability $P(B/A)$. In other cases we wish to compute $P(A/B)$. These two are related as follows:

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

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

From these we equate $P(B/A)P(A) = P(A/B)P(B)$

We can then write

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

This is known as Baye's rule. From $P(B/A)$ we compute $P(A/B)$.

Sometimes we deal with a set of events A_i, $i = 1, \dots, n$
We know that the event B occurred and wi sh to compute the conditional probability, given B, for each A_i, i.e. $P(A_i/B)$. Hence for each A_i we compute

$$P(A_i/B) = \frac{P(B/A_i)P(A_i)}{P(B)}$$

Then we can in a sense expand $P(B)$ using the law of total probability as $P(B) = \sum_{i=1}^{n} P(B/A_i)P(A_i)$. Hence the above becomes

$$P(A_i/B) = \frac{P(B/A_i)P(A_i)}{\sum_{j=1}^{n} P\big(B/A_j\big)P(A_j)}$$

Baye's rule is often expressed in the above form. The reason is that in many applications we know $P(A_j)$ and we know $P(B/A_j)$, and we wish to compute $P(A_i/B)$.

Application of Baye's Rule to Communications

In digital communications we often transmit a random symbol from an alphabet of n symbols, $\{S_1, \dots, S_n\}$. Then we observe a signal $Y \in \{S_1, \dots, S_n\}$ which occurs because of added noise in the channel. If the observed signal is equal to the transmitted signal then transmission occurred without error. On the other hand if the observed signal is different from the transmitted signal then an error occurred in the transmission. Let the transmitted signal be $S \in \{S_1, \dots, S_n\}$, and the corresponding received signal (after making a decision at the receiver) be $R \in \{S_1, \dots, S_n\}$. As a result the set of outcomes is $\Omega = \{\omega \mid \omega = (S,T)$, where $S, T \in \{S_1, \dots, S_n\}$. Note that there are n^2 possible outcomes. Define the events $A_i = \{(S_i, R) \mid R \in \{S_1, \dots, S_n\}\}$. Define $B = \big\{\big(S, R_j\big) \,\big|\, S \in \{S_1, \dots, S_n\}\}$. The typical situation is that we observe an even $B = S_j$ for some j, and we want to determine the probability that a particular symbol S_i was transmitted, i.e. $P(A_i/B)$. We can use Bayes theorem. We can write $P(A_i/B)$ as $P(\{(S_i, R) \mid R \in \{S_1, \dots, S_n\}\}/B\}$. We also usually abuse notation and write this simply as $P(S_i/B)$.

In terms of this "abused" notation Baye's theorem becomes

$$P(S_i/B) = \frac{P(B/S_i)P(S_i)}{P(B)} = \frac{P(B/S_i)P(S_i)}{\sum_{j=1}^{n} P\big(B/S_j\big)P(S_j)}$$

Note that typically $P(B/S_i)$, $P(S_i)$ are easy to compute directly and $P(S_i/B)$ is not easy to compute directly.

Independence of Events

Consider a probability triplet $(\Omega, \mathcal{F}, P)$. Suppose we have two events, $A, B \in \mathcal{F}$, and it is the case that $P(A \cap B) = P(A)P(B)$. Then we say that the two events A, and B are independent.

If two events A and B are independent then

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

Note that independence should not be confused with mutually exclusive, or disjoint events.

Example:

We construct an experiment as the flipping of two fair coins. The sample space is $\Omega = \{HH, HT, TH, TT\}$. We consider a σ-field consisting of all subsets of Ω.

There are $2^4 = 16$ possible events. Two of these events are $A = \{HH, HT\}$ and $B = \{HH, TH\}$. We assume that the singleton events $\{HH\}, \{HT\}, \{TH\}$, and $\{TT\}$ all have equal probabilities (fair coin), 1/4. Note that we may refer to the event A as "The first coin is a head" and to B as "the second coin is a head". These are independent events because $P(A \cap B) = P(HH) = 1/4$. However if $A = \{HT, TH, TT\}$, i.e. "one of the coins is a tail", and $B = \{HH, HT\}$, i.e. "the first coin is a head" then these are not independent. Because $A \cap B = \{HT\}$. Then $P(A \cap B) = 1/4$, and $P(A)P(B) = \frac{3}{4} \times \frac{2}{4} = \frac{6}{16} = 3/8$. So $P(A \cap B) \neq P(A)P(B)$.

Extension of a set of subsets to a σ-Field

Generating a σ-Field from a collection of sub-sets of Ω

In some cases we may have a collection of subsets of Ω, $\mathcal{S}$, that we wish to have as events, but $\mathcal{S}$ does not form a σ-field. This may be because any of the conditions such as closure under complements, or closure under countable unions and intersections does not hold for some set of events, i.e. elements in $\mathcal{S}$.

In this case we need to add extra events to $\mathcal{S}$ so as to make it a σ-field. In some cases we add the minimum number of elements. We will refer to the smallest σ-field containing $\mathcal{S}$ as $\sigma(\mathcal{S})$.

Finite Sample Spaces

Example

Consider $\Omega = \{1,2,3,4,5,6\}$. Let $\mathcal{S} = \{\phi, \{2,4,6\}\}$. $\mathcal{S}$ does not form a field. But it can be extended to a σ-field by adding the events $\{1,3,5\}$ and Ω. We refer to the arising σ-field as $\mathcal{F} = \sigma(\mathcal{S})$.

Note that a probability measure cannot be defined on $\mathcal{S}$ but it can be defined on $\sigma(\mathcal{S})$.

Note that for any σ-field that is finite, the number of elements, i.e. the number of different events must be a power of 2! How can we prove this? Consider a Venn diagram. When we consider all possible intersections of sets (events), the sample space will consists of a partition of elementary events. Then we consider all possible events as unions of these elementary events. For an arbitrary element (event) it is made up of a union of elementary events. Start with the partition, then we have two possibilities for each elementary event. Either it is included or it is not included. The number of such possibilities, i.e. the number of different events, is then 2^n, where n is the number of elementary events in the partition.

To make the above argument more rigorous consider a set of sub-sets of Ω, $\mathcal{S} = \{S_1, S_2, \cdots, S_k\}$ that we wish to extend to a field. Now for each element of Ω, ω_i we split $\mathcal{S}$ into two disjoint sets, $\mathcal{S}_I$ and $\mathcal{S}_O$ such that $\mathcal{S} = \mathcal{S}_I \cup \mathcal{S}_O$ and $\mathcal{S}_I \cap \mathcal{S}_O = \phi$, ω_i is in each set of $\mathcal{S}_I$ and in the complement of each set of $\mathcal{S}_O$. Then we form $E_i = E_I \cap E_O$ where E_I is the intersection of all sets in $\mathcal{S}_I$ and E_O is the intersection of all sets in $\mathcal{S}_O$.

It is clear that $\cup E_i = \Omega$. We show that for any two of these sets, E_k, E_l, either $E_k = E_l$, or $E_k \cap E_l = \phi$. In other words the two sets are either disjoint or are equal. Let us say that if $\omega_i \in S_k$ then S_k covers ω_i. Hence the sets in $\mathcal{S}_I$ are precisely the sets that cover ω_i, and no set in $\mathcal{S}_O$ covers ω_i. Hence if the collection of sets that cover ω_i is equal to the collection of sets that cover ω_j, obviously $E_i = E_j$. But if the collection of sets that cover ω_i is not equal to the collection of sets that cover ω_j, we can show in this case that $E_i \cap E_j = \phi$.

In the about we can discard some of the duplicate sets in $\mathcal{S}$ to obtain the set $\mathcal{S}_r = \{S_1, S_2, \cdots, S_r\}$ where $r \leq k$. The sets in $\mathcal{S}_r$ are elementary sets. We can them form the set of elements of $\sigma(\mathcal{S})$ as the set of all possible unions of elements in $\mathcal{S}_r$. There are 2^r such elements in $\sigma(\mathcal{S})$ since in forming a union we have two choices for each elementary set in $\mathcal{S}_r$; either include it in the union, or exclude it.

Countable Sample Spaces

The above situation also applies for any sample space that is countable.

Consider a countable sample space Ω. Index the points in Ω by i. Consider a set of subsets of Ω, $\mathcal{S}$ that we wish to extend to a σ-field. We should be able to do this since the set of all subsets of Ω does form a σ-field. For each $\omega_i \in \Omega$ partition $\mathcal{S}$ into two sets $\mathcal{S}_I$ and $\mathcal{S}_O$ as above, then form the intersections E_i as above. Then remove duplicates to obtain a partition of Ω into a set of elementary events. Finally form events from the set of elementary events as all possible unions of elementary events from the partition.

Uncountable Sample Spaces

We now consider a case where $\Omega = \mathcal{R}$, i.e. the real line. We start with a set of non-overlapping intervals, $A_i = (a_i, b_i)$ with possibly infinite number of elements. Consider $\mathcal{F}_0 = \{A_1, A_2, \ldots.\}$. We will attempt to construct $\sigma(\mathcal{F}_0) = \mathcal{F}$.

Then we try to enforce the "closure" conditions for a σ-field. Consider the point $c \in \mathcal{R}$, and consider $B_n = \left(c - \frac{1}{n}, c + \frac{1}{n}\right)$. For each n this is an open interval centered at c. Now we impose the closure condition $\cap_{n=1}^{\infty} B_n \in \mathcal{F}$. But we can show that $\cap_{n=1}^{\infty} B_n = \{c\}$, i.e. a point. Hence we have shown that the singleton set $\{c\}$ must be in any extension field.

As a result for any event $(a, b) \in \mathcal{F}_0$ then singleton sets $\{a\}$ and $\{b\}$ must also be in $\sigma(\mathcal{F}_0)$. Due to the closure with finite unions then we prove that $[a, b)$, $(a, b]$, and $[a, b]$, are all in $\sigma(\mathcal{F}_0)$. We refer to these as semi-closed, and closed intervals. We can also prove that any countable union of these intervals, of any type, is also in $\sigma(\mathcal{F}_0)$.

Now, this does not finish the construction of $\sigma(\mathcal{F}_0)$, i.e. the smallest sigma field containing $\mathcal{F}_0$. But it can be shown that such is a σ-field does not contain all the subsets of $\mathcal{R}$. There are subsets of $\mathcal{R}$ that are not included in $\sigma(\mathcal{F}_0)$, although these are difficult to construct.

Uniform Probability Measure

We can modify the above example by considering a sample space that is an interval of the real line, e.g. a closed interval $[a, b]$. Then we start with a finite set of events and extend as before, but being careful that we define sets (i.e. events) that do not extend beyond the interval $[a, b]$. Then we could define the probability measure on the set $\mathcal{F}_0$ as follows: Let (x, y) be an interval $(x, y) \subset [a, b]$. We define the probability measure on these subsets as $P\big((x, y)\big) = \frac{y-x}{b-a}$. Then we extend this measure to the $\sigma(\mathcal{F}_0)$ σ-field by using the additivity conditions for a probability measure. The resulting measure is known as the *uniform probability measure*. For this measure we can easily show for that for a countable set of non-overlapping intervals $A = \cup_{i=1}^{\infty} A_i$, then $P(A) = \sum_{i=1}^{\infty} P(A_i)$. We can also prove that for a singleton set $\{c\}$, $P(\{c\}) = 0$.

Going along this way we can show that there are subsets of $[a, b]$, e.g. $U \subset [a, b]$ such that it is not possible to assign a value $P(U)$ without getting into a contradiction. For example we would break down U as $U = U_1 \cup U_2$ with $U_1 \cap U_2 = \phi$, but we would show that $P(U) \neq P(U_1) + P(U_2)$ which is a contradiction for a probability measure. These sets are known as *non-measurable* sets. The text by Rosenthal proves the existence of these sets. These non-measurable sets are of course not included in $\sigma(\mathcal{F}_0)$ but it shows that it is not possible to consider the set of all subsets of $[a, b]$ as a σ-field. Usually in applications we do not consider σ-fields that are larger than $\sigma(\mathcal{F}_0)$. We start with the set of events $\mathcal{F}_0$, (and even finite unions of these) which consists of the important events and extend it to the smallest σ-field containing $\mathcal{F}_0$.

Exponential Probability Law

Consider $\Omega = [0, \infty)$ where we start with $\mathcal{F}_0 = \{\text{all sets of the form } (x, y)\}$. We define a probability measure on these sets as $P\big((x, y)\big) = e^{-x} - e^{-y}$. Then we extend this to the smallest σ-field containning $\mathcal{F}_0$, i.e. $\sigma(\mathcal{F}_0)$, with the probability measure satisfying the additivity constrains. The resulting gives the exponential probability law. For example we can prove that for singleton sets $\{c\}$, $P(\{c\}) = 0$.

The above is the exponential probability for the parameter $\lambda = 1$. But we can define it for arbitrary values of λ as follows: $P((x, y)) = (e^{-\lambda x} - e^{-\lambda y})$.

A third example of an uncountable sample space is the example $\Omega = \mathcal{R}$, with again starting with $\mathcal{F}_0$ being all intervals $(x, y) \in \mathcal{R}$, and extending to $\sigma(\mathcal{F}_0)$. We define the probability measure on $\mathcal{F}_0$ as $P((x, y) = \int_x^y f_\sigma(x)dx$, where $f_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right)$. This is the Gaussian probability law with zero mean and standard deviation σ. Note that the extended σ-field in this case is also called a **Borel** σ-field.

Exercise:

Prove that the set of rational numbers $\mathcal{Q}$ is an event in $\mathcal{F} = \sigma(\mathcal{F}_0)$.

*Prove that the set of algebraic numbers $\mathcal{A}$ is an element in $\mathcal{F} = \sigma(\mathcal{F}_0)$. Note that an algebraic number is a number that is a root of a polynomial (with finite degree) whose coefficients are rational numbers. This is a bit more work. Just given for the more ambitious.

Random Variables

We consider the space $\mathcal{R}$ (real numbers) with the Borel σ-field, $\mathcal{F}_B$. Consider a probability space $(\Omega, \mathcal{F}, P)$ and a mapping, i.e. function $X: \Omega \to \mathcal{R}$, with the condition that for each $B \in \mathcal{F}_B$, $X^{-1}(B) \in \mathcal{F}$. Note that $X^{-1}(B) = \{\omega : X(\omega) \in B\}$. The function $X(\cdot)$ is called a random variable.

Cumulative Distribution Function

Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $X(\cdot)$ a random variable. Now, the interval $(-\infty, x]$ is of course an element of $\mathcal{F}_B$, that is the Borel σ-field (on $\mathcal{R}$) defined above. Hence according to the definition of a random variable $X^{-1}\big((-\infty, x]\big) \in \mathcal{F}$, call it A_x. Hence the probability $P(A_x) = P(X^{-1}((-\infty, x])$ is defined. We denote this probability as $F_X(x)$ and refer to it as the cumulative distribution function (CDF) for the random variable X.

Properties of CDF, $F_X(x)$

i) $0 \le F_X(x) \le 1$

ii) $F_X(x)$ is right continuous and monotone increasing.

iii) $\lim_{x \to -\infty} F_X(x) = 0$

iv) $\lim_{x \to \infty} F_x(x) = 1$

v) If $x_1 \le x_2$ then $F_X(x_1) \le F_X(x_2)$ (definition of "monotone increasing").

Note that a probability law for a random variable X, i.e. a law that specifies the definition of the probability that the random variable falls in the event $A \in \mathcal{F}_B$ is completely determine by the CDF $F_X(x)$. That is, the CDF specifies the probability for a subset of the events in $\mathcal{F}_B$, but then the additivity condition for the probability measure $P(\cdot)$ is used to determine the probability for any event in $\mathcal{F}_B$.

Probability Density Function (PDF)

Consider a random variable X with CDF $F_X(x)$. The function $F_X(x)$ may or may not be differentiable. If it is differentiable then we denote the derivative as $f_X(x)$, i.e. $f_X(x) = \frac{dF_X(x)}{dx}$ and refer to it as the Probability Density Function (PDF).

Properties of the PDF

(i) $f_X(x) \ge 0$

(ii) $F_X(x) = \int_{-\infty}^{x} f_X(u)du$

(iii) $P(a \leq X \leq b) = \int_{a}^{b} f_X(u)du$

(iv) $\int_{-\infty}^{\infty} f_X(u)du = 1$

Discrete Random Variables

Let X be a random variable such that it takes values on a discrete subset of $\mathcal{R}$, i.e. $X(\omega) \in I_X = \{x_1, x_2, \cdots\}$ for any ω in Ω. Then it can be shown that $F_X(x)$ is continuous at each point $r \notin I_X$. But it may be discontinuous (i.e. have a "jump") at any point in I_X. In order words, $F_X(x)$ is a "staircase" function with steps at the points in I_X. Note that the steps can also have "zero height", i.e. no step. In this case we define the PDF for the random variable X as $f_X(x) = \sum_k p_X(x_k)\delta(x - x_k)$, where $p_X(x_i) = P(X(\omega) = x_i)$. Note we can easily prove that $\sum_k p_X(x_k) = 1$. For discrete random variables the PDF is also called the **probability mass function (PMF).**

Hybrid Continuous and Discrete Random Variables

Note that for a discrete random variable we have a set of points $\{x_1, x_2, \cdots\}$ at which the CDF may contain a "jump", and outside these points the PDF is constant. We may also have a situation where outside a set of discrete points the CDF is continuous and differentiable (not necessarily constant), but there are possible "jumps" at the set of discrete points. In this case the PDF would be described as a continuous function with added delta functions.

Expected Value of Random Variable

We are familiar with the concept of expected value for random variables, X, as follows. If X has a density function then we define the expected value as

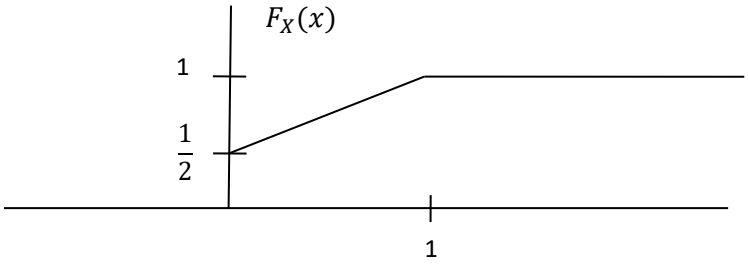
$$\mathcal{E}(X) = \int_{-\infty}^{\infty} x f_X(x)dx$$

If X is a discrete random variable on the discrete set $\{x_1, x_2, \cdots\}$ then we would write

$$\mathcal{E}(X) = \sum_{k=1}^{\infty} x_k p_X(x_k)$$

For cases where we can write a "PDF" as a hybrid of a continuous and discrete part then we can use a combination of the above, or use an expression for the PDF that includes both a "continuous" part and a part that has delta functions. For example we can specify a CDF as follows

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{1}{2} & \text{for } x = 0 \\ \frac{1+x}{2} & \text{for } 0 < x \leq 1 \\ 1 & \text{for } x > 1 \end{cases}$$



Note that F_X is not continuous, hence not differentiable, although it is piecewise differentiable. We could then write the PDF as follows:

$$f_X(x) = \frac{1}{2}\Pi\left(x - \frac{1}{2}\right) + \frac{1}{2}\delta(x)$$

Where $\Pi(x) = 1$ if $|x| \leq \frac{1}{2}$ and 0 elsewhere, is the unit rectangular function.

The above is the definition of expected value for random variables whose CDF is basically a piecewise differentiable function with "jumps" at a discrete set of points $x_1, x_2, \cdots$ This is usually enough for most cases in practice.

Formal Definition of Expected Value

In the formal definition of expected value we need to define integrals according to the **theory of integration that uses measure theory**. This is the modern theory of integration. One special case of this theory is the Lebesgue integral that we frequently hear about.

We consider **a measure space** on $\mathcal{R}$ as the triplet $(\mathcal{R}, \mathcal{F}_B, \mu)$. As above $\mathcal{R}$ is the set of real numbers and $\mathcal{F}_B$ is the smallest σ-field on $\mathcal{R}$ containing all intervals. We have also called this a Borel σ-field.

In a common case of modern integration of real functions on $\mathcal{R}$, the measure μ is defined as a function on $\mathcal{F}_B$ as $\mu: \mathcal{F}_B \to [0, \infty]$ with the property that

i) for any $A \in \mathcal{F}_B$, $\mu(A) \geq 0$, (this includes the possible case $\mu(A) = \infty$), and

ii) ii) if $A = \cup_{i=1}^{\infty} A_i$ and the A_i's are mutually exclusive (i.e. $A_i \cap A_j = \phi$ for $i \neq j$) then $\mu(A) = \sum_{i=1}^{\infty} \mu(A_i)$.

Note that this is similar to a probability measure except that we do not require the normalization condition $\mu(\mathcal{R}) = 1$. In fact $\mu(\mathcal{R})$ does not even have to be finite, i.e. we can have $\mu(\mathcal{R}) = \infty$. The best example here is the so-called Lebesgue measure $\mu = \mu_L$ on $\mathcal{R}$, where for the interval (a, b), $\mu_L\big((a, b)\big) = b - a$. For this measure $\mu_L(\mathcal{R}) = \infty$. In fact let $A = \cup_i A_i$, where the A_i are an infinite sequence of non-overlapping intervals with $\mu_L(A_i) = m_i$. Then if $\sum_{i=1}^{\infty} m_i$ is a sequence that does not converge (i.e. $\sum_{i=1}^{\infty} m_i = \infty$), we have $\mu_L(A) = \infty$. So we see that there are many sets in $\mathcal{F}_B$ (not equal to $\mathcal{R}$) with infinite measure.

So, with the above we have two triplets $(\Omega, \mathcal{F}, P)$ a probability measure space, and $(\mathcal{R}, \mathcal{F}_B, \mu)$ a measure space (non-probability). And we have the random variable $X: \Omega \to \mathcal{R}$.

An Aside on Integration Theory

Now consider the measure space $(\mathcal{R}, \mathcal{F}_B, \mu)$, and consider a real function $f: \mathcal{R} \to \mathcal{R}$, in this measure space. We say that the function f is *measurable* (in a sense, this means that it can be integrated in the modern sense of the theory of integration) if f satisfies the condition:

if $B \in \mathcal{F}_B$ then $f^{-1}(B) \in \mathcal{F}_B$.

Note that the notation $f^{-1}(B)$ refers to a subset of $\mathcal{R}$ as follows: $f^{-1}(B) = \{x \in \mathcal{R}: f(x) \in B\}$. In the case of the Riemann integral we usually define it over intervals, e.g. $\int_a^b f(x)dx$, with the possibility that $a = -\infty$ and $b = \infty$. But in the case of measure theory the integral can be defined over any set A that is an element of $\mathcal{F}_B$, i.e. not only for intervals.

First, the notation for this general integral is given as $\int_A f d\mu$. In the case that we are integrating over the whole space we could write $\int_{\mathcal{R}} f d\mu$, but for simplicity we write this simply as $\int f d\mu$.

Note also, that in the Riemann case of integration, the integral is usually defined for functions $f: \mathcal{R} \to \mathcal{R}$, whereas in the modern integration theory we can define it for functions $f: \Omega \to \mathcal{R}$ where Ω is an arbitrary set for which there is a defined σ-field, and a measure on that σ-field. In this case in order to define the integral the function f needs to satisfy the following: if $B \in \mathcal{F}_B$ then $f^{-1}(B) \in \Omega$, i.e. it needs to be measureable.

Now, we know about Riemann sums as a way to define Riemann integrals. For the integral based on measure theory we start with the simplest functions that we can integrate, i.e. the indicator functions defined as follows: Let $A \in \mathcal{F}_B$ then the indicator function for A is $i_A(x) = 1$ if $x \in A$, and 0 elsewhere. The integral of this function is then defined as $\int i_A d\mu = \mu(A)$. Then we impose linearity as follows: for any $c \in \mathcal{R}$, ci_A is a function and we define the integral as $\int ci_A d\mu = c\int i_A d\mu$. Then if $A = \cup_i A_i$ where the A_i are non-

overlapping sets (elements of $\mathcal{F}_B$) we form the function $f_A = \sum_i c_i i_{A_i}$. We refer to these functions as *simple functions*. The integral becomes $\int f_A d\mu = \sum_i c_i m_i$, where $m_i = \mu(A_i)$. For a general function f, the idea, to define the integral, then becomes that we approximate it as a sum of a large set of scaled indicator functions (simple functions), i.e. $f \approx \sum_{i=1}^{N} c_i i_{A_i}$, where the A_i are chosen so that f restricted to A_i is approximately constant with value c_i. Then $\int f d\mu \approx \sum_i c_i \mu(A_i)$. In a sense we partition the whole space into a set of "small" sets A_i where f is approximately constant over each set A_i. The sets are made small enough so as to bound the error, of the integral over each such set, to a value as small as desired. Then the integral of each set of the partition is easy to compute, i.e. the integral of a scaled indicator function. We use the additivity property to obtain the integral for the function (simple function, i.e. approximation to f, over the whole space. This needs to be made rigorous but I have indicated the conceptual idea. As a comparison, for the usual Riemann integral, we break up the domain (the x-axis) into small intervals. Over each interval the function is approximately constant and the integral for each small interval is the area of a tall and narrow rectangle, i.e. we find the areas of these small rectangles and sum them. In the case of the integral based on measure theory we break the range of the function, i.e. the y-axis into a sequence of bins, i.e. a partition of small intervals. Over each bin (small interval) we approximate the value of the function by a constant. For example if the range of the function is the interval $[Y_1, Y_2]$. Then we could partition this set as follows: $[Y_1, Y_2] = \cup_{i=1}^{n} [y_i, y_{i+1})$, where $y_1 = Y_1$ and $y_i = y_1 + \frac{Y_2 - Y_1}{n}(i-1)$. Then for each small interval $B_i = [y_i, y_{i+1})$ we let $A_i = f^{-1}(B_i)$. The integral of f is then approximately $\sum_i y_i \mu(A_i)$. This can be made rigorous to obtain the definition of the integral as a limit of these sums as $n \to \infty$. When the measure μ is chosen to be the Lebesgue measure μ_L (this is a special case in the modern theory of integration) then the resulting integral is the so-called Lebesgue integral. For piecewise continuous functions we can show that the Lebesgue integral becomes the same as the Riemann integral. But the Lebesgue integral is defined for a much larger class of functions. It is also much easier to work with in many proofs involving integration, although for problems in engineering, the Lebesgue integral is the same as the Riemann integral.

Example of a Lebesgue integrable function which is not Riemann integrable. Consider the measure space $(\Omega, \mathcal{F}, \mu)$, where $\Omega = [a, b]$ and $\mathcal{F} = \{A_r : A_r = [a, b] \cap A, \text{for } A \in \mathcal{F}_B\}$, i.e. the sets in $\mathcal{F}$ are the sets in $\mathcal{F}_B$ restricted to (intersected with) $[a, b]$. Now, consider the element in $\mathcal{F}$, $Q_r = \{\text{all rational numbers in the interval } [a, b]\}$. Now, the set of rational numbers is a countable set, hence we can see clearly that Q_r is an element of $\mathcal{F}$ because it is a countable union of singleton sets, $Q_r = \{x \in [a, b] : x \text{ is rational}\}$, and we know that singleton sets are elements in $\mathcal{F}_B$. Now assume a measure which is the Lebesgue measure μ_L restricted to the space $[a, b]$. Then we can show that $\mu_L(Q_r) = 0$. Hence, $\int i_{Q_r} d\mu_L = 0$. On the other hand the Riemann integral for the function i_{Q_r} is not defined at all, because the so-called Riemann sums (upper and lower approximations) do not converge.

Returning to the Expected Value in the Context of Integration Theory

Now that we have discussed the modern definition of integral, we return to the concept of a random variable. We consider the probability space $(\Omega, \mathcal{F}, P)$ and the measure space $(\mathcal{R}, \mathcal{F}_B, \mu)$, where the measure μ is defined as, for each $B \in \mathcal{F}_B$, $\mu(B) = P(X^{-1}(B))$. In a sense the measure μ in $\mathcal{R}$ is inherited from the measure P in Ω. With these conditions, a random variable X is a mapping $X: \Omega \to \mathcal{R}$ that is measurable. The expected value of X is then simply the integral of the function X on Ω with respect to the measure P, i.e. $\mathcal{E}(X) = \int_\Omega X dP$. Or we can consider the measure space $(\mathcal{R}, \mathcal{F}_B, \mu)$ and define the measure μ as being inherited from the measure P on Ω. The expected value then becomes $\mathcal{E}(X) = \int_\mathcal{R} X d\mu$.

To emphasize that the measure μ on $\mathcal{R}$ is in a sense inherited from the measure P on Ω, sometimes the expected value is written as $\mathcal{E}(X) = \int_\Omega X(\omega)\, dP(\omega)$, or it is written merely as $\int_\Omega X dP$, but to some extent $\int_\mathcal{R} X\, d\mu$ is preferrable because it shows that the expected value is merely an integral of a function on $\mathcal{R}$ with respect to a specific measure on $\mathcal{R}$ (in fact a probability measure meeting the normalization condition). Integrals, in the modern sense, are always with respect to a measure. The so-called Lebesgue integral is the modern integral with respect to the Lebesgue measure, i.e. a special case of a measure on $\mathcal{R}$ where the measure of an interval (a specific event in $\mathcal{F}_B$) is the length of the interval.

The above formulation was carried out in terms of integrals of functions defined on $\mathcal{R}$ because this is usually the case of interest in integration theory. We concern ourselves with functions $f: \mathcal{R} \to \mathcal{R}$. In the case of probability the random variable is defined on Ω, as the function $f: \Omega \to \mathcal{R}$. There are two points of view here.

1) Define the integral of the function on the set Ω using the probability measure P defined on Ω, i.e. $\mathcal{E}(X) = \int_\Omega X dP$, or

2) Define the integral on $\mathcal{R}$ using the inherited measure on $\mathcal{R}$, μ. In this case the random variable is actually the identity function on $\mathcal{R}$ and we are integrating it with respect to the inherited measure on $\mathcal{R}$, i.e. $\mathcal{E}(X) = \int_\mathcal{R} X_I d\mu$, where X_I is the identity function on $\mathcal{R}$, i.e. $X_I(x) = x$ for $x \in \mathcal{R}$.

The first case in the above is the usual way that we represent things in probability. However it deals with two spaces Ω and $\mathcal{R}$, whereas the second case deals with a single space $\mathcal{R}$, but with the non-Lebesgue measure on $\mathcal{R}$ inherited from Ω. It deals in a sense with real functions defined on $\mathcal{R}$, as opposed to real functions defined on Ω. I hope that this has not really confused the issue, but I am just emphasizing the connection between an expected value and an integral in the modern sense of integration.

In most engineering examples we deal with well behaved functions and the well-known Riemann integral is sufficient. Now, the above measure μ on $\mathcal{R}$ is usually not the Lebesgue measure. In fact it is the Lebesgue measure only for the special case of a random variable with uniform probability law on an interval with unit length, e.g. $[a, a+1]$ for any $a \in \mathcal{R}$. In this case the expected value would be written as $\int_a^{a+1} x dx$ (usual Riemann integral). However for a more general probability law we describe the law uniquely by the CDF and the Riemann integral becomes $\int_{\mathcal{R}} x\, f(x) dx$. The function $f(x)$ (PDF) in a sense specifies the probability measure. In the modern integral $d\mu$ plays the role of $f(x)dx$ of the Riemann case. The integral (expected value) can also be written as $\int_{\mathcal{R}} x\, dF(x)$ which is sometimes called a Riemann-Stiltjes integral.

Having defined the expected value in the above we must say that the expected value does not necessarily exist for all probability laws. For the expected value to exist, the integral $\int_{\mathcal{R}} X d\mu$ must converge. Functions X, i.e. random variables, for which this holds are called $\mathcal{L}_1$ functions. There are well known probability laws for which the expected value does not exist. One of the most well-known examples is the so-called Cauchy probability law given by the PDF $f_X(x) = \frac{1}{\pi(1+x^2)}$.

In the following, to deal with expressions that are more familiar we will just assume the Riemann case for the integral. In doing so we will assume that the random variable has a density function, i.e. the CDF is well behaved. We keep in mind that if we want it to work for the general case then we should be using the measure theory based integral, or at least the Riemann-Stiltjes version which is based on the CDF.

In the above we discussed the expected value of a random variable, X, which is the integral of X with respect to the probability measure. We can generalize this to $\mathcal{E}(g(X))$ the integral of any function of the random variable X, or $g(X)$. The requirement necessary condition is that the function $g(\cdot)$ must be measurable. But this is not a sufficient condition. It requires that such an expected value exists. In the expressions below, involving expected values we will in some cases assume that the random variable has a PDF and write $\mathcal{E}(X) = \int x f_X(x) dx$

Moments of a Random Variable

Let X be a random variable. The n^{th} moment of X is defined as $m_n = \mathcal{E}(X^n) = \int x^n f_X(x) dx$, if this integral exists. The expected value is then also called the 1^{st} moment.

Central Moments

Let X be a random variable with mean m. The n^{th} central moment of X is defined as $\mathcal{E}((X-m)^n)$. The second central moment is called the *variance*, i.e. $\mathcal{E}((X-m)^2)$. Note that the variance is

obviously always a positive quantity. The square root of the variance is called the *standard deviation* – usually denoted by σ. The variance is then denoted as σ^2. Note that for the second moment of a random variable X to exist, X must be a function in what we call the class $\mathcal{L}_2$, i.e. square integrable functions with respect to the given probability measure.

Note that since we can write the variance as $\sigma^2 = m_2 - m_1^2$ (show it), then the variance exists if and only if the second moment exists. To show this relation between the second moment and the variance we have

$\mathcal{E}((X-m)^2) = \mathcal{E}(X^2 - 2mX + m^2) = \mathcal{E}(X^2) - 2m\mathcal{E}(X) + \mathcal{E}(m^2) = m_2 - 2m^2 + m^2 = m_2 - m^2$.

Two-Random Variables

Consider a probability Space $(\Omega, \mathcal{F}, P)$. We define a pair of random variables, or vector random variable, X, on this space as a mapping $X: \omega \to \mathcal{R}^2$, i.e. a mapping from the sample space to the real plane. Each outcome $\omega \in \Omega$ maps to the point in the plane (X_1, X_2). Just as in the case of a single random variable. This mapping needs to meet a measurability condition. Note also that there are two interpretations here: i) Two real random variables, or ii) one 2-D vector random variable. Now, we need to define a σ-Field in $\mathcal{R}^2$, just as we did in $\mathcal{R}$. We refer to it as $\mathcal{F}_2$. We start by requiring that all the sub-sets of $\mathcal{R}^2$ of the form $(X_1 \le x_1, X_2 \le x_2)$ are elements in $\mathcal{F}_2$. Let one of these sub-sets be I_{x_1,x_2}. Let $\mathcal{S}_2 = \{I_{x_1,x_2}: x_1 \in \mathcal{R}, x_2 \in \mathcal{R}\}$. We then form $\mathcal{F}_2$ as the smallest σ-field containing $\mathcal{S}_2$, i.e. $\mathcal{F}_2 = \sigma(\mathcal{S}_2)$. The vector random variable X must satisfy the condition: for every $B \in \mathcal{F}_2$, $X^{-1}(B) \in \mathcal{F}$.

With this definition we can define a probability space on $\mathcal{R}^2$ as follows $(\mathcal{R}^2, \mathcal{F}_2, P_2)$, where the probability law P_2 is defined as follows: For each $B \in \mathcal{F}_2$, $P_2(B) = P(X^{-1}(B))$. In other words the space $(\mathcal{R}^2, \mathcal{F}_2, P_2)$ inherits the probability law from $(\Omega, \mathcal{F}, P)$ through the mapping X.

We may specify the probability law P_2 by specifying it on the smaller set S_2, and then extending it to $\mathcal{F}_2$ by enforcing the usual conditions for a probability law.

We introduce the notation $F_{X_1,X_2}(x_1, x_2) = P_2(I_{x_1,x_2})$. This is called the joint Cumulative Distribution Function (CDF) for the two random variables X_1, X_2, or for the vector random variable X. Note that it is a function $F_{X_1,X_2}: \mathcal{R}^2 \to [0,1]$. We can also write F_X, if we understand X to be a 2-D random vector. In most cases we also drop the subscript 2 and write P instead of P_2 if there is no confusion between the measures on Ω and $\mathcal{R}^2$. Instead of $P_2(I_{x_1,x_2})$ we usually write $P(X_1 < x_1, X_2 < x_2)$.

Properties of the Joint PDF

i) $0 \le F_X(x_1, x_2) \le 1$

ii) For a fixed x_2, $\lim_{x_1 \to -\infty} F_X(x_1, x_2) = 0$, $\lim_{x_1 \to \infty} F_X(x_1, x_2) = F_{X_2}(x_2)$

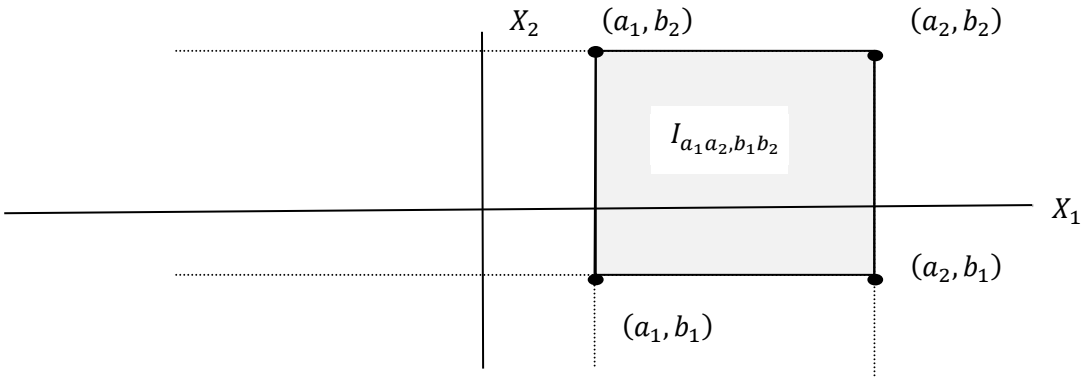
iii) For a fixed x_1, $\lim_{x_2 \to -\infty} F_X(x_1, x_2) = 0$, $\lim_{x_2 \to \infty} F_X(x_1, x_2) = F_{X_1}(x_1)$

iv) If $a_1 \leq a_2$, and $b_1 \leq b_2$, then $F_X(a_2, b_2) + F_X(a_1, b_1) - F_X(a_1, b_2) - F_X(a_2, b_1) \geq 0$

Note that in the above we defined $F_{X_2}(x_2) = P(X_1 < \infty, X_2 \leq x_2)$. But we also write this in short form as $P(X_2 \leq x_2)$. The same applies to the function $F_{X_1}(x_1)$. We are really abusing notation here, but this is standard in Engineering textbooks because it greatly simplifies notation and usually does not cause confusion.

Note that if we have a function $F_X(x_1, x_2)$ satisfying the above conditions then it determines a unique probability law for a set of two random variables. The function obviously specifies the probability for any rectangle (semi-infinite) of the form $(-\infty < X_1 \leq a, -\infty < X_2 \leq b)$, i.e. $F_X(a, b)$. Then we can determine the probability for any semi-open rectangle of the form $(a_1 < X_1 \leq a_2, b_1 < X_2 \leq b_2)$, as follows: To shorten notation we write these subsets of $\mathcal{R}^2$ as $I_{a,b}$ and $I_{a_1 a_2, b_1 b_2}$. So we have $I_{a_2,b_2} = I_{a_1 a_2, b_1 b_2} \cup I_{a_1 b_1} \cup \left(I_{a_2 b_1} \backslash I_{a_1 b_1}\right) \cup \left(I_{a_1 b_2} \backslash I_{a_1 b_1}\right)$. Note that the notation $A \backslash B$ means $A \cap B^c$. Also $P(A \backslash B) + P(A \cap B) = P(A)$. Hence $P(A \backslash B) = P(A) - P(A \cap B)$. Noting that $I_{a_2 b_1} \cap I_{a_1 b_1} = I\_(a_1 b_1)$, and following this line we get

$P\left(I_{a_1 a_2, b_1 b_2}\right) = P\left(I_{a_2,b_2}\right) - P\left(I_{a_2,b_1}\right) - P\left(I_{a_1,b_2}\right) + P\left(I_{a_1,b_1}\right)$. This can be seen clearly by drawing a diagram as in the Figure below.



Marginal Cumulative Distribution Functions

Consider a CDF for two random variables, $X = (X_1, X_2)$, $F_X(x_1, x_2) = F_{X_1, X_2}(x_1, x_2)$. We define the marginal CDFs as $F_{X_1}(x_1) = \lim_{x_2 \to \infty} F_X(x_1, x_2)$, and $F_{X_2}(x_2) = \lim_{x_1 \to \infty} F_X(x_1, x_2)$.

Discrete Random Variables

For general (continuous) random variables we have seen that the CDF describes the probability law completely. The PDF can be used but only in some cases where the CDF is at least piecewise differentiable. Hence the CDF is the most important of the two.

However for a discrete random variable X, taking values in $\{x_1, x_2, \cdots\}$, we can form the elementary events in Ω as $E_i = X^{-1}(x_i)$, and define the probabilities $p_i = P(E_i)$. For simplicity we can write this as $p_i = P_X(x_i)$. We can of course also define a CDF, but the PDF (which could be defined using delta functions, but where we may call the Probability Mass Function, PMF) specifies a random variable completely and in some sense it is the easier of the two to work with.

For two discrete random variables X_1 and X_2, or $X = (X_1, X_2)$, assume that X_1 takes values in $\{x_1, x_2, \cdots\}$ and X_2 takes values in $\{y_1, y_2, \cdots\}$. The sets in Ω $E_{i,j} = X^{-1}\left(\left(x_i, y_j\right)\right)$ are elementary events. We write the probability $P(E_{i,j})$ as $P_{X_1,X_2}(x_i, y_j)$. The marginal probabilities are then defined as

$$P_{X_1}(x_i) = \sum_j P_{X_1,X_2}(x_i, y_j)$$

and

$$P_{X_2}(y_j) = \sum_i P_{X_1,X_2}(x_i, y_j)$$

Joint Probability Density Function

The joint PDF for a pair of random variables is defined in terms of partial derivatives of the joint CDF. Of course this only applies in cases where the partial derivatives exist. To simplify notation and not use subscripts in the random variables we will write the pair of random variables as (U, V), rather than (X_1, X_2). The joint CDF for (U, V) is $F_{U,V}(u, v)$ and the joint PDF, if it exists, is defined as

$$f_{U,V}(u, v) = \frac{\partial^2 F_{U,V}(u, v)}{\partial u \partial v}$$

We have stated that for a single random variable X, the CDF defines the probability law completely, and this is similar for a pair of random variables where the joint CDF $F_{U,V}(u, v)$ defines the probability law completely. Now, with two random variables we can define conditional probabilities for one r.v. given the other. For two events A, B we have defined the conditional probability $P(A/B)$. Now consider the two r.v.'s (U, V). Suppose we are told that $U = u$, what can we say about V. What is the probability law for V? We can define events in Ω according to outcomes defined by knowledge of U. For example, if we know that $U \leq u_0$, then we can consider

the probability $P(V \leq v)$ given this information. Note that $X^{-1}(V \leq v)$ and $X^{-1}(U \leq u_0)$ are events in Ω. Hence we can write

$$P(V \leq v/U \leq u_0) = \frac{P((V \leq v) \cap (U \leq u_0))}{P(U \leq u_0)}.$$

Note that in the above we take u_0 as a fixed value and consider v as a variable. Also, for the above to be valid it must be the case that $P(U \leq u_0) > 0$, otherwise we would be dividing by 0!

We could then define a conditional CDF as follows $F_{V/U \leq u_0}(v) = P(V \leq v/U \leq u_0)$. A conditional PDF can also be defined (if the appropriate derivates of the CDF exist) as $f_{V/U \leq u_0}(v) = \frac{d}{dv} F_{V/U \leq u_0}(v)$. The key here is the observation that $U \leq u_0$. In a similar fashion we may define conditional CDF's and PDF's given any other observed event that involves only U. In particular we may consider the case $U = u_0$. Now in many cases where the joint CDF, for U and V, is continuous, the probability $P(U = u_0) = 0$. Hence we can not define a conditional probability given this event. Note that the event $U = u_0$ in Ω is defined as $\{\omega \in \Omega: U(\omega) = u_0\}$. However if we consider the small interval $\left(u_0 - \frac{\Delta}{2}, u_0 + \frac{\Delta}{2}\right)$, and if $P\left(u_0 - \frac{\Delta}{2} \leq U < u_0 + \frac{\Delta}{2}\right) > 0$, then we can define the conditional probability

$$P\left(V \leq v/u_0 - \frac{\Delta}{2} < U < u_0 + \frac{\Delta}{2}\right) = \frac{P\left((V \leq v) \cap (u_0 - \frac{\Delta}{2} < U < u_0 + \frac{\Delta}{2})\right)}{P\left(u_0 - \frac{\Delta}{2} < U < u_0 + \frac{\Delta}{2}\right)}$$

Now we can take the limit as $\Delta \to 0$, and if this limit exists it will obviously be a function of v for a fixed u_0. We can call this function the conditional CDF of V given that $U = u_0$ and write it as

$$F_{V/u}(v)$$

Note that the key here is that we fix a point u_0 which was the observed value for the r.v. U. Then we consider Y as a random variable given that $U = u_0$ and we define a CDF for Y. Various different forms of notation in different books are used to denote this conditional CDF, including $F_V(v/U = u)$, or $F_V(v/u)$, etc, where we have dropped the subscript in u_0. The subscript in u_0 was introduced to emphasize that we are fixing the value u_0, but in many cases we simply write u.

Conditional PDF

Note that the above conditional CDF is a function of v and contains in a sense u as a parameter. We may define a conditional probability density function as follows

$$f_{V/u}(v) = \frac{d}{dv} F_{V/u}(v)$$

Again we often see the notation $f_V(v/u)$ because in some calculations we first fix u as a constant but then allow it to vary and become a variable. In such cases the notation $f_V(v/u)$ indicates these two variables. The main case where this happens is in applications of a generalized law of total probability.

We can also show that the conditional PDF can be written in terms of the joint PDF for the two r.v.'s, U, V, as follows:

$$f_V(v/u) = \frac{f_{U,V}(u,v)}{f_U(u)}$$

Independence of Random Variables

We have previously defined the independence of two events A, B as the requirement that $P(A \cap B) = P(A)P(B)$. Two random variables X_1, X_2, are said to be independent if for any two events in $\mathcal{R}$, A, B, we have $P(\{X_1 \in A\} \cap \{X_2 \in B\}) = P(\{X_1 \in A\})P(\{X_2 \in B\})$. In particular, if the two events are $\{X_1 \in (-\infty, x_1]\}$ and $\{X_2 \in (-\infty, x_2]\}$, then we have the requirement $P(-\infty < X_1 \leq x_1, -\infty < X_2 \leq x_2) = P(-\infty < X_1 \leq x_1)P(-\infty < X_2 \leq x_2)$. In terms of CDFs the requirement becomes $F_{X_1,X_2}(x_1, x_2) = F_{X_1}(x_1)F_{X_2}(x_2)$.

In terms of density functions, if the CDF is sufficiently well behaved that the derivatives exist, then we have for the PDF

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$$

In the above we proceeded to show that independence of two r.v.s entails the factoring of the CDF and PDF, if it exists. On the other hand we can show that if the CDF for a pair of random variables factors, then the two r.v.s are independent.

As a result, independence of r.v.'s is equivalent to the factoring of the CDF, and PDF, if it exists.

Expected Value of a Function of Two Random Variables

Let $Z = g(X_1, X_2)$ be a function of two random variables. Z is a random variable. This can be shown as follows:

We have two functions $X: \Omega \to \mathcal{R}^2$, and $g: \mathcal{R}^2 \to \mathcal{R}$. We can also define the function, or random variable, $Z: \Omega \to \mathcal{R}$, $Z(\omega) = g(X(\omega), Y(\omega))$.
We can show that if X is a measurable function ($\Omega \to \mathcal{R}^2$), hence an r.v., and if g is a measurable function ($\mathcal{R}^2 \to \mathcal{R}$), then the function Z ($\Omega \to \mathcal{R}$) is also a measurable function, hence it is a random variable.

We can of course go through a procedure to determine the CDF and PDF of the r.v. Z in terms of the CDFs of X_1, and X_2, and g. The expected value of Z can then be computed in the usual manner. For example, if the PDF of Z exists then we can compute

$$\mathcal{E}(Z) = \int_{-\infty}^{\infty} z f_Z(z) dz$$

But we can also compute this expected value as $\mathcal{E}(Z) = \int_{\mathcal{R}^2} g(x_1, x_2) f_{X_1,X_2}(x_1, x_2) dx_1 dx_2$. Conceptually we are simply obtaining $\mathcal{E}(Z) = \int_\Omega Z dP$.

Correlation of Two Random Variables

In the above, let the two random variables be U and V. Then we may consider the special case of the function $Z = g(U, V) = UV$. We can then define the correlation of the random variables U and V which is the expected value of Z as

$$\mathcal{E}(Z) = \mathcal{E}(UV) = \int_{\mathcal{R}^2} uv f_{U,V}(u, v) du dv$$

If $\mathcal{E}(UV) = 0$ then we say that the two random variables, U, V are orthogonal.

> **Aside (Concept of Orthogonality)**
> The concept of orthogonality applies to what we call an inner product space, i.e. a space of functions where a dot product satisfying some axioms is defined.
>
> Consider a vector space over the real numbers. Axioms for inner product space
>
> Inner product: $<.,.>: V \times V \to R$
>
> $< x, x > \geq 0$ with equality iff $x = 0$
>
> $< x, y > = < y, x >$ Symmetry
> $< ax + by, z > = a < x, z > + b < y, z >$ - Linearity

An example of such a space is the set of real functions defined on an interval $[a, b]$ with the condition that, for each such function f, the integral $\int_a^b f(x)^2 dx < \infty$. Note that this space of functions can be identified with periodic functions over the real line with period equal to $b - a$. These are the functions for which we define Fourier Series. In signal processing the above integral is referred to as the energy of the signal.

In this space of functions we define an inner product as an operation on two functions as $< f, g >= \int_a^b f(x)g(x)dx$. The norm of a function is then defined as $||f|| = \left(\int_a^b f(x)^2 dx\right)^{1/2}$. The well-known Cauchy-Schwarz inequality is then $| < f, g > | \leq \left||f|\right| \cdot ||g||$. If the inner product between two functions is zero then we say that the functions are ***orthogonal***. A vector space of functions satisfying these conditions, where an inner product is defined and for each function f, $\left||f|\right| < \infty$, is also called a Hilbert space, or an $\mathcal{L}_2$ space.

In the case of vectors in three dimensional space the inner product is the usual dot product. The dot product of two vectors $\overrightarrow{v_1}$ and $\overrightarrow{v_2}$ is $|\overrightarrow{v_1}||\overrightarrow{v_2}| \cos(\theta)$. The Cauchy-Schwarz inequality basically states that $|\overrightarrow{v_1}||\overrightarrow{v_2}| \cos(\theta) \leq |\overrightarrow{v_1}||\overrightarrow{v_2}|$.

Now, in a probability space Ω with random variables $X: \Omega \to \mathcal{R}$ we can define a vector space with elements being random variables, i.e. real-valued functions on Ω. We define the inner product of two of these functions (i.e. r.v.'s) as $< X, Y >= \mathcal{E}(XY) = \int_\Omega XY dP$. The requirement for this space of functions to be a Hilbert space is that for any function in the space, X (i.e. a random variable), the second moment exists, i.e. $\int_\Omega X^2 dP < \infty$. We can also refer to the resulting space of functions as $\mathcal{L}_2(\Omega, P)$ to emphasize that the space refers to real functions on Ω and the inner product is defined in terms of an integral with respect to the probability measure P on Ω.

If $\mathcal{E}(XY) = 0$, we say that the r.v.'s X and Y are orthogonal. As a result, orthogonality between two r.v.'s is like orthogonality between any two functions in a vector space of functions, where the inner product of the two functions is (in a common case) the integral of the product of the functions. The only difference in this case of probability is that the functions are r.v.'s and the inner product is the integral of the product of two r.v.'s with respect to a probability measure.

Covariance of Two Random Variables

The covariance between two random variables X, Y, is defined as

$$\text{cov}(X, Y) = \mathcal{E}\big((X - m_X)(Y - m_Y)\big),$$

where $m_X = \mathcal{E}(X)$, and $m_Y = \mathcal{E}(Y)$. This can also be simplified as $\text{cov}(X,Y) = \mathcal{E}(XY) - m_X m_Y$. Work this out! If the covariance of two random variables X, and Y, is zero, then we say that the two r.v.'s are uncorrelated.

$\text{cov}(X,Y) = 0 \quad \Leftrightarrow \quad X$, and Y are uncorrelated
$\mathcal{E}(XY) = 0 \quad \Leftrightarrow \quad X$, and Y are orthogonal

If the means of X and Y are zero then *uncorrelated* and *orthogonal* are the same thing.

Variance of a Sum of Random Variables

Let X and Y be random variables. Form the sum $Z = X + Y$.

Then the variance of Z is $\text{var}(Z) = \mathcal{E}((Z - m_Z)^2) = \mathcal{E}((X + Y - (m_X + m_Y))^2)$

$$= \mathcal{E}((X+Y)^2 - 2(m_X + m_Y)(X+Y) + (m_X + m_Y)^2)$$
$$= \mathcal{E}((X+Y)^2) - (m_X + m_Y)^2$$
$$= \mathcal{E}(X^2) - m_X^2 + \mathcal{E}(Y^2) - m_Y^2 + 2\mathcal{E}(XY) - 2m_X m_Y$$
$$= \text{var}(X) + \text{var}(Y) + 2\text{cov}(X,Y)$$

If $X$and Y are uncorrelated, then $\text{var}(Z) = \text{var}(X) + \text{var}(Y)$, i.e. the variance of the sum is equal to the sum of the variances.

Independence and Uncorrelatedness of Random Variables

If two random variables X and Y are independent then they are also uncorrelated. This follows from the fact that $\mathcal{E}(XY) = \mathcal{E}(X)\mathcal{E}(Y)$ if X and Y are independent. To prove this for the general case we need to carefully discuss integrals in $\mathcal{R}^2$ using the probability measure defined by the joint CDF. We will not do it here. In the specialized case that the CDF is well behaved and the joint PDF for X and Y is defined, the PDF factors as $f_{X,Y}(x,y) = f_X(x) f_Y(y)$. The expected value then becomes

$$\mathcal{E}(XY) = \int_{\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X,Y}(x,y) = \int_{\infty}^{\infty}\int_{-\infty}^{\infty} xy f_X(x) f_Y(y) dxdy$$
$$= \int_{-\infty}^{\infty} x f_X(x) dx \int_{-\infty}^{\infty} y f_Y(y) dy = \mathcal{E}(X)\mathcal{E}(Y)$$

Hence $\text{cov}(X,Y) = \mathcal{E}(XY) - \mathcal{E}(X)\mathcal{E}(Y) = 0$, i.e.
Independent => Uncorrelated. However, the converse is not true.

Uncorrelatedness does not Imply Independence

To conclude, independence of two random variables implies uncorrelatedness. However, the converse is not necessarily true. The common example is as follows: Assume $\Omega = [0, \pi]$ with the

uniform probability measure P. Define two random variables on X, Y, on Ω, as follows: $X = \cos(\omega)$, $Y = \sin(\omega)$. X and Y are uncorrelated as we can determine from $\int_0^{\pi} \cos(\omega)\sin(\omega) \cdot \frac{1}{\pi} d\omega = 0$. On the other hand the two r.v.'s are not independent. For example pick ϵ small. $P(Y \leq \epsilon / 0 \leq X < \epsilon) \neq P(Y \leq \epsilon)$. In fact for sufficiently small ϵ the result on the left equals 0, (X small means that ω is close to $\frac{\pi}{2}$ and Y is close to 1) whereas on the right it is not equal to 0.

Correlation Coefficient

Let X and Y be two random variables. The correlation coefficient, $\rho_{X,Y}$, is defined as

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where $\sigma_X = \sqrt{\text{var}(X)}$, $\sigma_Y = \sqrt{\text{var}(Y)}$.

Note that as a result of the Cauchy-Schwarz inequality for two random variables X, Y,

$$|\mathcal{E}(XY)|^2 \leq \mathcal{E}(X^2)\mathcal{E}(Y^2)$$

If we define two new random variables by subtracting the means, $X' = X - m_X$, $Y' = Y - m_Y$. With these transformation $\mathcal{E}(X^2) < \infty$ if and only if $\mathcal{E}(X'^2) < \infty$. The same holds for Y. Hence the Cauchy-Schwarz inequality applies equally well to second moments of X, Y, as to second moments of X', Y', i.e. correlations and co-variances of X and Y. Hence we have

$$|\text{cov}(X,Y)|^2 = |\mathcal{E}(X'Y')|^2 \leq \mathcal{E}\left(X'^2\right)\mathcal{E}\left(Y'^2\right) = \sigma_X^2 \sigma_Y^2$$

$$\frac{|\text{cov}(X,Y)|}{\sigma_X \sigma_Y} \leq 1$$

or

$$-1 \leq \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \leq 1$$

$$-1 \leq \rho_{X,Y} \leq 1.$$

Random Vectors

Random vectors are a generalization of a random variable X (a mapping to $\mathcal{R}$) to n dimensions. The random vector is a mapping from Ω to $\mathcal{R}^n$, i.e. $\boldsymbol{X}: \Omega \to \mathcal{R}^n$. First we introduce a σ-field in $\mathcal{R}^n$ as the smallest σ-field containing all n dimensional regions $R(x_1, x_2, \cdots, x_n) = (-\infty < X_1 \leq x_1, -\infty < X_2 \leq x_2, \cdots, -\infty < X_n \leq x_n)$ call it $\mathcal{F}_n$. Then the random variable $\boldsymbol{X}$ is any such

function that is measurable, i.e. for any event $B \in \mathcal{F}_n$, $\boldsymbol{X}^{-1}(B) \in \mathcal{F}$. The previous case that we considered, "joint random variables" corresponds to the case $n = 2$.

As for the case $n = 2$, we let $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$ and define the corresponding CDF as a function $F_{\boldsymbol{X}}: \mathcal{R}^n \to [0,1]$.

$$F_{\boldsymbol{X}}(x_1, x_2, \ldots, x_n) = P(X_1 \in (-\infty, x_1], X_2 \in (-\infty, x_2], \cdots, X_n \in (-\infty, x_n])$$

Marginal CDF's can be defined for all orders less than n, and for any subset of the n variables, by taking the limit for the other arguments as they approach infinity. For example, we can define a marginal CDF of order two (for variables X_1 and X_2) as follows:

$$F_{\boldsymbol{X}}(x_1, x_2) = \lim_{x_3 \to \infty, x_4 \to \infty, \cdots, x_n \to \infty} F_{\boldsymbol{X}}(x_1, x_2, \cdots, x_n)$$

Probability Mass Function

Joint Probability Mass Function (PMF)

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = P_{X_1, \cdots, X_n}(x_1, \cdots, x_n) = P(X_1 = x_1, \cdots, X_n = x_n)$$

Marginal PMFs can be defined for any subset of the variables by summing over the complement of the set of variables, e.g.

$$P_{X_1, X_3}(x_1, x_3) = \sum_{x_2} \sum_{x_4, \cdots, x_n} P_{X_1, \cdots, X_n}(x_1, \cdots, x_n)$$

Conditional PMF

$$P_{X_n}(x_n / x_1, \cdots, x_{n-1}) = \frac{P_{X_1, \cdots, X_n}(x_1, \cdots, x_n)}{P_{X_1, \cdots, X_{n-1}}(x_1, \cdots, x_{n-1})}$$

Probability Density Function

Assuming that the joint CDF is well behaved and various partial derivatives exist we can define the joint PDF as follows:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{X_1, \cdots, X_n}(x_1, \cdots, x_n) = \frac{\partial^n F_{X_1, \cdots, X_n}(x_1, \cdots, x_n)}{\partial x_1 \cdots \partial x_n}$$

Similarly, marginal densities can be defined by "integrating out" some of the variables, for example

$$f_{X_1X_3}(x_1, x_3) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1 \cdots X_n}(x_1, x_2, x_3, \cdots, x_n) d_{x_2} d_{x_4} \cdots d_{x_n}$$

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1, \cdots, X_n}(x_1, \cdots x_n) dx_2 \cdots dx_n$$

Conditional PDF's can also be defined as a generalization of the 2-variable case

$$f_{X_n}(x_n / x_1, \cdots, x_{n-1}) = \frac{f_{X_1, \cdots X_n}(x_1, \cdots, x_n)}{f_{X_1, \cdots, X_{n-1}}(x_1, \cdots x_{n-1})}$$

Independence

The set of random variables $X_1, \cdots, X_n$ is independent if an only if the joint CDF factors, as follows

$$F_{\boldsymbol{X}}(x_1, x_2, \cdots, x_n) = F_{X_1}(x_1) F_{X_2}(x_2) \cdots, F_{X_n}(x_n)$$

The PDF factors in a similar manner

$$f_{\boldsymbol{X}}(x_1, x_2, \cdots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n)$$

The Expected Value of a Vector of Random Variables* (This is more advanced, it is indented)

> Let $(\Omega, \mathcal{F}, P)$ be a probability space.
> Recall that we define the vector of random variables $\boldsymbol{X} = (X_1, \cdots, X_n)$ as a mapping $\boldsymbol{X}: \Omega \rightarrow \mathcal{R}^n$. The expected value of $\boldsymbol{X}$ is defined as the integral $\int_\Omega \boldsymbol{X} dP$. Let us see how this integral is defined. Note that in the case of a single r.v. we partitioned the real line $\mathcal{R}$ into a set of small intervals over which the function X is approximately constant. Let the set of intervals be B_k. We found the set of events $X^{-1}(B_k) = E_k$. Then we approximated the function $X \approx \sum_k c_k I_k$, where $c_k \in B_k$ (e.g. pick the middle point), and $I_k = i(E_k)$, i.e. the indicator function. The expected value was then $\mathcal{E}(X) \approx \sum_k c_k P_k$, where $P_k = P(E_k)$. We take the limit as the intervals B_k are made very small and this becomes $\int_\Omega X dP$. This of course is not a mathematically rigorous definition but it gives the idea.
>
> In the n-dimensional case we partition $\mathcal{R}^n$ into a union of very small "hyper-cubes" in n-dimensional space. We will just call them n-dimensional cubes, or nD cubes. These nD cubes will be indexed by k, and referred to as C_k. Over each nD cube (i.e. C_k), the vector X is approximately constant with value equal to the vector U_k. Consider the set of events in $\mathcal{F}$ $E_k = X^{-1}(C_k)$. The vector random variable X can then be approximated as $X \approx$

$\sum_k U_k I_k$, where $I_k = i(E_k)$, is the indicator function of E_k. The expected value is then $\mathcal{E}(X) \approx \sum_k U_k P(E_k)$. We then take the limit as the nD cubes become small and obtain the formal definition of the expected value of the vector random variable X as $\mathcal{E}(X) = \int_\Omega X dP$. Again, this is not a rigorous definition but it gives the idea behind the formal definition of the expected value of a vector random variable.

By analyzing the above we see that an integral of a vector-valued function with respect to a measure is really a sum of small vector quantities. With some reflection we can expect that the integral (expected value) should follow the relation $\mathcal{E}(\boldsymbol{X}) = \int_\Omega \boldsymbol{X} dP = (\int_\Omega X_1 dP, \cdots, \int_\Omega X_n dP) = (\mathcal{E}(X_1), \cdots, \mathcal{E}(X_n))$. In a sense the integral of a vector function over a measure space is equal to the vector of integrals of the component functions over the same measure space. From now on we will simply write for a vector random variable.

$$\mathcal{E}(\boldsymbol{X}) = \big(\mathcal{E}(X_1), \mathcal{E}(X_2), \cdots, \mathcal{E}(X_n)\big)$$

Linear Operations on Random Variables.

If X is a random variable we can define $Z = cX$ as another random variable, and we can show that, since the expected value is really an integral of the r.v. (a real-valued function on Ω) and that since the integral is a linear operation then $\mathcal{E}(Z) = c\mathcal{E}(X)$.

In the same manner we can show that if $\boldsymbol{X} = (X_1, \cdots, X_n)^T$ is a random vector then the expected value of a linear combination of the r.v.,s X_k is equal to a linear combination of the expected values. If we think of the r.v. $\boldsymbol{X}$ as a column vector, and $\boldsymbol{c}$ as a column vector of coefficients, then $\mathcal{E}(\boldsymbol{c}^T \boldsymbol{X}) = \sum_k c_k \mathcal{E}(X_k)$, where the superscript T denotes the transpose. In other words the expected value of a linear combination of random variables is the same linear combination of the expected value of the random variables. This is all a consequence of the expected value of a random variable being simply an integral over a measure space. Integrals are linear operators, hence expected values are also linear operators.

The above "dot product" of the vector $\boldsymbol{c}$ with the random vector $\boldsymbol{X}$ can be generalized to matrix operations. Let $\boldsymbol{A}$ be an $m \times n$ matrix. Then we can transform the random vector $\boldsymbol{X}$ to the random vector $\boldsymbol{Z}$ as follows: $\boldsymbol{Z} = \boldsymbol{AX}$. We can then compute

$$\mathcal{E}(\boldsymbol{Z}) = \mathcal{E}(\boldsymbol{AX}) = \boldsymbol{A}\mathcal{E}(\boldsymbol{X})$$

This also applies to multiplication of the random vector $\boldsymbol{X}$ by the constant matrix $\boldsymbol{B}$, on the right, or

$$\mathcal{E}(\boldsymbol{XB}) = \mathcal{E}(\boldsymbol{X})\boldsymbol{B}$$

Note that the above point of view for expected values is elegant and works for all probability laws for a random vector. However, we know that the probability law for the random vector $\boldsymbol{X}$ can be stated in terms of a joint CDF for the n random variables. In cases where the joint CDF is well behaved, i.e. the appropriate partial derivatives exist, we can also determine a joint PDF. The computation of the expected value of $\boldsymbol{X}$ would then involve an n-dimensional integral. For example in the case of $\mathcal{R}^2$ we would compute $\mathcal{E}(X_i) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_i f_X(x_1, x_2)dx_1dx_2$, for $i = 1,2$.

Correlation Matrix for Random Vector *X*

Let $\boldsymbol{X}$ be a random vector which we represent as a column vector, $\boldsymbol{X} = (X_1, \cdots, X_n)^T$. Note that we are defining $\boldsymbol{X}$ as a column vector.
We can compute the second moment of any of the components and also the correlation between any two components X_i, X_j, of the random vector $\boldsymbol{X}$. All this information is contained in the correlation matrix $\boldsymbol{R_X}$ which we define as follows:

$$\boldsymbol{R_X} = \text{cor}(\boldsymbol{X}) = \left(\mathcal{E}\left(X_i X_j\right)\right) = \mathcal{E}[\boldsymbol{X}\boldsymbol{X}^T]$$

Note that $\boldsymbol{R_X}$ is an $n \times n$ matrix. Also note that this matrix is symmetric, i.e. if $\boldsymbol{R_X} = (r_{ij})$ then $r_{ij} = r_{ji}$ for all i, j. Note also that r_{ii} is the second moment of X_i. If the random variables X_i $(i = 1, \cdots, n)$ are uncorrelated, i.e. any two are uncorrelated, then the off-diagonal elements of the correlation matrix can be obtained as $r_{ij} = \mathcal{E}(X_i)\mathcal{E}(X_j)$, $i \neq j$.

Covariance Matrix for *X*

Let $\boldsymbol{X}$ be a random vector which we represent as a column vector as above. We can compute the variance of any of the components and also the covariance between any two components X_i, X_j, of the random vector $\boldsymbol{X}$. All this information is contained in the covariance matrix $\boldsymbol{C_X}$ which we define as follows:

$$\boldsymbol{C_X} = \text{cov}(\boldsymbol{X}) = \mathcal{E}[(\boldsymbol{X} - \boldsymbol{m_X})(\boldsymbol{X} - \boldsymbol{m_X})^T]$$

where $\boldsymbol{m_X} = \mathcal{E}(\boldsymbol{X})$. Note that $\boldsymbol{C_X}$ is an $n \times n$ matrix. Also note that this matrix is symmetric, i.e. if $\boldsymbol{C_X} = (v_{ij})$ then $v_{ij} = v_{ji}$ for all i, j. Note also that v_{ii} is the variance of X_i. If the random variables X_i $(i = 1, \cdots, n)$ are uncorrelated, i.e. any two are uncorrelated, then the covariance matrix is a diagonal matrix.

The Correlation Matrix is Non-Negative Definite

Consider the random vector $\boldsymbol{X} = (X_1, \cdots, X_n)^T$, and the constant vector $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)^T$, $(\alpha \neq 0)$, where the α_i are real numbers. Now, take the linear combination $Z = \alpha_1 X_1 + \cdots + \alpha_n X_n$, i.e.

$Z = \boldsymbol{\alpha}^T \boldsymbol{X}$. Then Z^2 is a positive random variable and we have $\mathcal{E}(Z^2) \geq 0$. But we can write Z^2, after expansion, as $\sum_{i=1}^n \sum_{j=1}^n \alpha_i X_i \alpha_j X_j$. Taking the expected value we obtain $\mathcal{E}(Z^2) = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \mathcal{E}(X_i X_j) = \sum_{i=1}^n \sum_{j=1}^n r_{ij} \alpha_i \alpha_j$, where $\boldsymbol{R_X} = (r_{ij})$ is the correlation matrix for $\boldsymbol{X}$. Note that this can also be written in matrix notation as $\sum_{i=1}^n \sum_{j=1}^n r_{ij} \alpha_i \alpha_j = \boldsymbol{\alpha}^T \boldsymbol{R_X} \boldsymbol{\alpha}$. The conclusion is that for any real constant vector α we have

$$\boldsymbol{\alpha}^T \boldsymbol{R_X} \boldsymbol{\alpha} \geq 0$$

A matrix satisfying the above property is called a non-negative definite matrix. As a result we can say that any correlation matrix is non-negative definite.

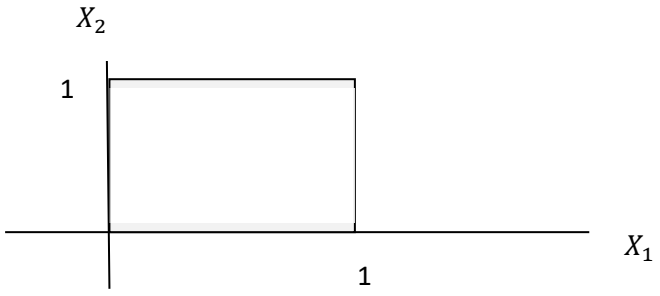
Covariance Matrix

Note that if we go through the above with a random vector $\boldsymbol{X}$ with zero mean then we obtain the covariance matrix $\boldsymbol{C_X} = (C_{ij}) = (\mathcal{E}(X_i X_j))$. In matrix form, assuming that $\boldsymbol{X}$ is represented as a column vector we can write $\boldsymbol{C_X} = \mathcal{E}(\boldsymbol{X}\boldsymbol{X}^T)$. Note from the above that the covariance matrix is obviously also symmetric and non-negative definite, i.e. for any real vector $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}^T \boldsymbol{C_X} \boldsymbol{\alpha} \geq 0$. Note that for a random vector with zero mean the covariance matrix and the correlation matrix are equal. The covariance matrix for the random vector $\boldsymbol{X}$ is equal to the correlation matrix for the random variable $\boldsymbol{X'}$, where $\boldsymbol{X'} = \boldsymbol{X} - \boldsymbol{m_X}$, and $\boldsymbol{m_X} = \mathcal{E}(\boldsymbol{X})$.

Gaining Intuition about the Joint CDF, PDF

Independence – Example 1

Consider a probability space Ω with $\boldsymbol{X} = (X_1, X_2)$ where the probability law is the uniform distribution on $[0,1] \times [0,1]$, i.e. $f_X(x_1, x_2) = 1$ on $[0,1] \times [0,1]$ and zero elsewhere. Are X_1 and X_2 independent?



We can see that if we are given the value of X_2, it does not seem to give us any information on X_1, i.e. where X_1 is likely to lie uniformly in the interval $[0,1]$.

We can check this formally as follows:
The CDF is

$$F_{\boldsymbol{X}}(x_1, x_2) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f_{\boldsymbol{X}}(x_1, x_2)dx_1dx_2$$

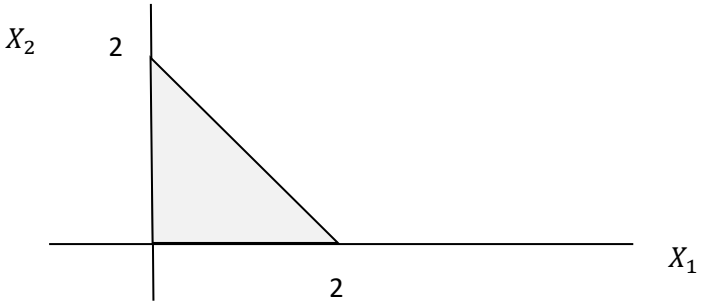
This is equal to 0 for X in quadrants 2, 3, and 4. In the shaded square we have $F_{\boldsymbol{X}}(x_1, x_2) = x_1x_2$. In $(1, \infty) \times [0,1]$ we have $F_{\boldsymbol{X}}(x_1, x_2) = x_2$, and in $[0,1] \times (0, \infty)$ we have $F_{\boldsymbol{X}}(x_1, x_2) = x_1$.

We can check that $F_{\boldsymbol{X}}(x_1, x_2) = F_{X_1}(x_1)F_{X_2}(x_2)$, where, for $i = 1,2$

$$F_{X_i}(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } 0 \le x \le 1 \\ 1 & \text{for } x > 1 \end{cases}$$

Example 1b

Now, let us consider the probability law given by the PDF $f_{\boldsymbol{X}}(x_1, x_2) = 1/2$ for (x_1, x_2) in the shaded region below, and 0 elsewhere.



It seems that now the two random variables are not independent because if we are told that, for example one of the random variables is large, i.e. close to 2, then the other one is likely to be small, i.e. close to 0. For example, if $X_2 > 1$, then the probability $P(X_1 < 1/X_2 > 1) \neq P(X_1 < 1)$. In fact, $P(X_1 < 1) = 3/4$. But $P(X_1 < 1/X_2 > 1) = 1$. Hence the events $\{X_1 < 1\}$ and $\{X_2 > 1\}$ are not independent. The random variables X_1 and X_2 are therefore not independent.

Note: $P(X_1 < 1/X_2 > 1) = \frac{P(X_1<1 \,\&\, X_2>1)}{P(X_2>1)} = \frac{\frac{1}{4}}{\frac{1}{4}} = 1$

Independence – Example 2

Consider two discrete random variables X and Y, with joint Probability Mass Function (PMF) given in the following Table

	Y					
X		1	2	3	4	
	1	.025	.025	.0125	.0375	.1
	2	.05	.05	.25	.075	.2
	3	.075	.075	.0375	.1125	.3
	4	.1	.1	.05	.15	.4
		.25	.25	.125	.375	1

The marginal PMF's are shown in the shaded row and column. We can see that the PMF's are the products of the marginal PMF's. Hence the random variables X and Y are independent.

On the other hand, consider the following case

	V					
U		1	2	3	4	
	1	.252	0	0	0	.252
	2	.062	.188	0	0	.25
	3	.061	.062	.124	0	.247
	4	.063	.062	.063	.063	.251
		.438	.312	.187	.063	

In this case the random variables $U$and V are not independent. In fact if $U = 1$ then we know that $V = 1$, i.e. $P(V = 1) \neq P(V = 1/U = 1)$. $P(V = 1) = .438$, but

$$P(V = 1/U = 1) = \frac{P(U=1,V=1)}{P(U=1)}.\frac{252}{.252} = 1.$$

Example – Jointly Gaussian Random Variables

We will generate random variables using the computer. Later we will discuss this topic in more detail. For now we will just use some built-in functions in MATLAB.

We have defined probability according to a model involving a sample space, events, and a probability measure. A random variable is a function mapping sample space points to the real line. We can simulate the experiment using a computer algorithm. For example we can consider a 32 bit computer. Taken some integer α and repeatedly take powers $\alpha, \alpha^2, \alpha^3, \cdots$, where the evaluation is $\mathrm{mod}(2^{32})$. By using a frequency definition of probability and normalizing the numbers to a maximum of 1, we conclude that the outcomes when mapped to the interval $[0,1]$ constitute a random variable with uniform distribution on the interval [0,1]. All computer languages have function that return a value that models such a random variable. Let the random variable be X. We can then apply a function $g(\cdot)$ to the random variable X to obtain $Y = g(X)$ with Y having an arbitrary distribution. In other words, it is possible to find a function $g(\cdot)$ so that the CDF for Y is an arbitrary distribution. This is a procedure that can be used to generate random variables using the computer. According to our probability model we perform an experiment and obtain an outcome ω. Then we map ω to a real number x. With the computer "performing the experiment"

means that we call a function. This function then returns a number x which is an instance of a random variable X.

We can generate a random vector (X, Y) where X and Y are independent, by repeating the experiment assuming that different calls to the above function that generates a random number produce independent random variables with uniform distribution. Whether they experiment (calling the function) produces independent random variables or not can be verified by statistics.

Below we generate pairs of random variables with marginal Gaussian distributions and a given correlation.

First we generate two independent random variables with mean zero, variance 1, and Gaussian Distribution. The density function for a Gaussian distributed random variable X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$$

Where $m = \mathcal{E}(X)$, and $\sigma^2 = \text{var}(X)$.

We use the Matlab function $\text{normalrnd}(m, v, N, n)$ to generate an array of independent Gaussian distributed random numbers with mean m, variance v, N rows, n columns. In our case we will generate N samples, where each sample is a vector of $n = 2$ components.

We refer to the random variables (i.e. random vectors) as (U, V). Hence we generate N 2-D random vectors with zero mean and $\text{var}(U) = \text{var}(V) = 1$.

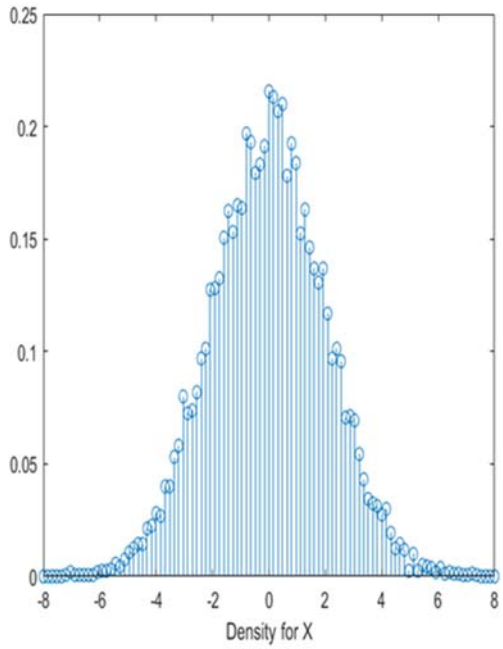
For each 2D random vector we perform a transformation to a new 2D random vector (X, Y), where $X = \sigma_X U$, and $Y = \sigma_Y\left(\rho U + \sqrt{1-\rho^2}V\right)$. Note that $\text{var}(X) = \sigma_X^2$, $\text{var}(Y) = \sigma_Y^2$, and the correlation coefficient can be determined as $\frac{\mathcal{E}(XY)}{\sigma_x \sigma_y} = \rho$.

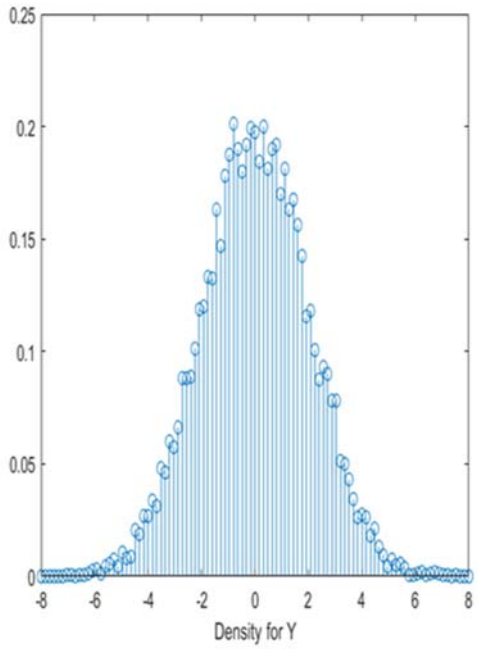
Below we give the code to generate these in Matlab.

```
clear;
nbins=100;      %   Number of bins for Histogram
N=100000;       %   Number of samples
sigmax = 1;     %   Standard deviation for random variable X
sigmay =2;      %   Standard deviation for random variable Y
range = 4*max(sigmax,sigmay)    %  Range for plotting
rho = .5;       %   Correlation coefficient for X and Y
x=(-nbins/2:nbins/2)*2*range/nbins;  % Set of points in x-axis for Histogram
dx=2*range/nbins;   % Size of one bin
dy=dx;
% Generate Gaussian distributed random vectors with mean 0, variance 1
% For each row we have one vector (U,V)
```

```
R=normrnd(0,1,N,2);
X = sigmax*R(:,1);  % scale the X r.v.
Y = sigmay*(rho*R(:,1)+sqrt(1-rho^2)*R(:,2));  % Generate Y that is correlated
with X
Hx=hist(X,x)/(N*dx);
Hy = hist(Y,x)/(N*dy);
stem(x,Hx);
xlabel("Density for X")
sum(Hx)*dx
stem(x,Hy)
xlabel("Density for Y");
sum(Hy)*dy
mx = sum(X)/N       % Estimate of the mean for X
my = sum(Y)/N       % Estimate of the mean for Y
varX = X'*X/N       % Estimate of variance for X
varY = Y'*Y/N       % Estimate of variance for Y
corXY = X'*Y/(N*sqrt(varX*varY))    % Estimate of correlation coefficient
plot(X,Y);
box off;
xlim([-range range]);
ylim([-range range]);
xlabel("X")
ylabel("Y")
```

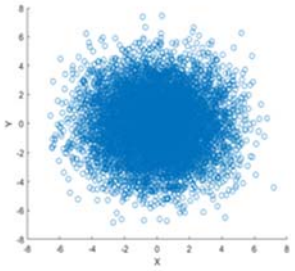
The following is the histogram for X. The histogram approaches the density as the number of samples approaches infinity and the size of the bins approaches 0.



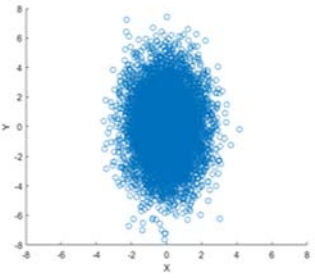


The following are scatter plots of the random samples (X_i, Y_i). For each case we give the parameters σ_X, σ_Y, and ρ.

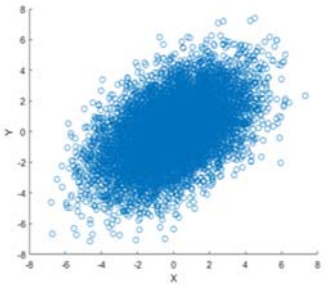
$\sigma_X = 1,\ \sigma_Y = 1, \rho = 0$



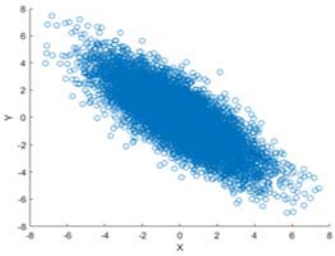
$\sigma_X = 1,\ \sigma_Y = 2, \rho = 0$



$\sigma_X = 2,\ \sigma_Y = 2, \rho = .5$



$\sigma_X = 2,\ \sigma_Y = 2, \rho = -.8$



Complex Random Variables

We have talked about the case of two random variables, or one vector random variable. This is the same thing, but different points of view. Another way of saying this is that the random variable instead of being a real number, i.e. real valued, it is vector valued. In the case of a 2D vector we had $\boldsymbol{X}: \Omega \to \mathcal{R}^2$. We are now going to introduce **a third point of view** that is also useful. The random variable X is a mapping from the sample space Ω to the complex numbers C. In this case we will call it Z instead of X. So we have $Z: \Omega \to C$, where C represents the set of complex numbers. We can specify a probability law for the random variable Z by treating it as a pair of real numbers X and Y. In order to study the probability law for Z we need to introduce a σ-field in C, which is the smallest σ-field containing all subsets of C that are semi-infinite rectangles. We will refer to this σ-field as $\mathcal{F}_C$. Note that this is similar to the case of $\mathcal{R}^2$. A semi-infinite rectangle is the set $\{Z : \mathrm{Re}(Z) \le a, \mathrm{Im}(Z) \le b\}$. So this set is defined by the complex number $a + jb$, where $j = \sqrt{-1}$. We can therefore define the CDF for Z as $F_Z(x, y) = P(X \le x, Y \le y)$, where $Z = X + jY$. Note that this is really all the same as the case that we have considered previously with respect to a 2D vector random variable in $\mathcal{R}^2$. Now, when we dealt with random variables in $\mathcal{R}^2$ we could evaluate all sorts of expected values such as $\mathcal{E}(\boldsymbol{X})$, $\mathcal{E}(X_1)$, $\mathcal{E}(X_1X_2)$, etc. This means that we write $\boldsymbol{X} = (X_1, X_2)$ then we can evaluate the expected value of any function of $X_1, X_2, \boldsymbol{X}$, etc. These functions can be vector valued or real valued. For example we can find the expected value of $X_1^2 + X_2^2$, i.e. real-valued. Or we could determine the expected value of $\boldsymbol{AX}$, where $\boldsymbol{A}$ is a constant matrix, i.e. $\boldsymbol{AX}$ *is a* vector valued random variable.

Now, the case of consideration of complex valued random variables is convenient when we have to deal with complex numbers and it is desirable to perform complex operations on these such as multiplication. This is especially important in the computation of certain correlations. For example suppose we have two complex valued random variables Z_1, and Z_2. We can define a correlation as follows: $\mathrm{cor}(Z_1, Z_2) = \mathcal{E}(Z_1 Z_2^*) = \mathcal{E}\big((X_1 + jY_1)(X_2 + jY_2)^*\big) = \mathcal{E}(X_1X_2 + Y_1Y_2) + j\mathcal{E}(X_2Y_1 - X_1Y_2)$. This correlation is important in physical problems modeling signals with complex numbers, such as the analysis of complex signals that we are familiar with in Electrical Engineering.

Covariance Matrix for a Complex Random Vector

Consider a complex random vector $\boldsymbol{X}$ represented as a column vector.
The covariance matrix of $\boldsymbol{X}$ (defined as a column vector) is defined as

$$\mathrm{cov}(\boldsymbol{X}) = \mathcal{E}((\boldsymbol{X} - \boldsymbol{m_X})(\boldsymbol{X} - \boldsymbol{m_X})^H)$$

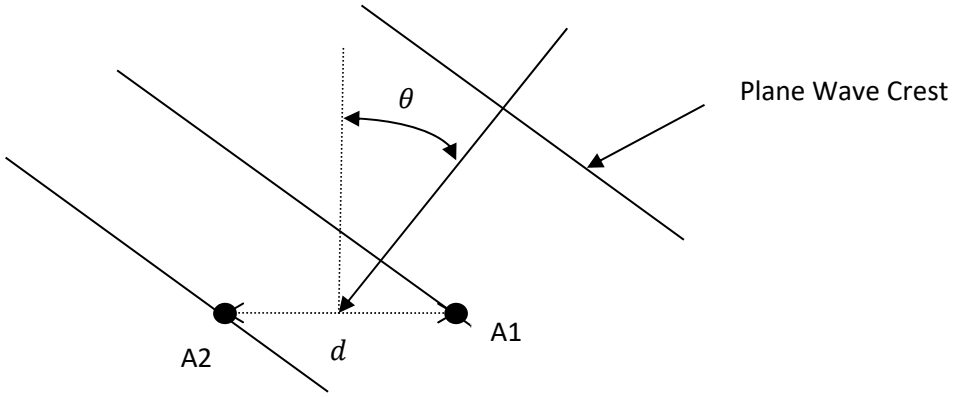
where $\boldsymbol{m_X} = \mathcal{E}(\boldsymbol{X})$ and the superscript H denotes the complex conjugate of the transpose, also called the Hermitian transpose, i.e. for a complex vector $\boldsymbol{Z}$, $\boldsymbol{Z}^H = (\boldsymbol{Z}^T)^*$

We can see that the diagonal elements are real numbers because they are equal to the product of a complex number with its complex conjugate. On the other hand the off-diagonal elements are not

necessarily real. The **covariance matrix is a Hermitian matrix**, i.e. it is equal to the complex conjugate of its transpose.

Example

Consider a radio system with two antennas separated by a distance d. A plane wave arrives from a direction θ as shown in the following Figure



There are two antenna elements A1 and A2 pointing out of the page. A plane wave arrives from a random direction θ with respect to the normal to the line connecting the two antenna elements. Assume that the plane wave signal at the antenna A1 is represented by a phasor (a complex number) $X_1 = Ae^{j\phi}$. The phasor for the signal at $A2$ is then $X_2 = Ae^{j\phi}e^{-j\frac{2\pi d}{\lambda}\sin\theta}$. Now, we model the pair (ϕ, θ) as outcomes in a probability space $\Omega = [0,2\pi] \times [0,2\pi]$ with uniform probability law. X_1 and X_2 are then two complex random variables. We can verify that the expected values $\mathcal{E}(X_1) = \mathcal{E}(X_2) = 0$. The covariance matrix for the complex vector random variable $\boldsymbol{X}$ is then

$$\text{cov}(\boldsymbol{X}) = \mathcal{E}(\boldsymbol{X}\boldsymbol{X}^H)$$

This yields the following

$$\text{cov}(\boldsymbol{X}) = \begin{pmatrix} \mathcal{E}(|X_1|^2 & \mathcal{E}(X_1X_2^*) \\ \mathcal{E}(X_1^*X_2) & \mathcal{E}(|X_2|^2) \end{pmatrix} = \begin{pmatrix} A^2 & A^2\mathcal{E}\left(e^{j\frac{2\pi d}{\lambda}\sin\theta}\right) \\ A^2\mathcal{E}\left(e^{-j\frac{2\pi d}{\lambda}\sin\theta}\right) & A^2 \end{pmatrix}$$

The expected value for the off-diagonal elements can be calculated as $\int_0^{2\pi}\frac{1}{2\pi}e^{-j\frac{2\pi d}{\lambda}\sin\theta}d\theta = J_0\left(\frac{2\pi d}{\lambda}\right)$, where $J_0(\cdot)$ is a Bessel Function. Hence the above simplifies to

$$\text{cov}(X) = A^2 \begin{pmatrix} 1 & J_0\left(\frac{2\pi d}{\lambda}\right) \\ J_0\left(\frac{2\pi d}{\lambda}\right) & 1 \end{pmatrix}$$

In wireless systems in some cases the goal is to choose an antenna element spacing, d, so as to minimize the off-diagonal elements of the above matrix.

Note that in this case we did not have to concern ourselves with probability densities for these complex random variables, which are difficult to determine. We were only interested in the covariance matrix. This is typical in many applications in signal processing. We work mostly with covariances.

Cross-Covariance

The cross-covariance matrix for two complex random vectors $\boldsymbol{X}$, $\boldsymbol{Y}$, is defined as

$$\text{cov}(\boldsymbol{X}, \boldsymbol{Y}) = \mathcal{E}\left((\boldsymbol{X} - \boldsymbol{m_X})(\boldsymbol{Y} - \boldsymbol{m_Y})^H\right)$$

Theorem

Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two constant matrices and $\boldsymbol{X}$, and $\boldsymbol{Y}$, random vectors, then

$$\text{cov}(\boldsymbol{AX}, \boldsymbol{BY}) = \boldsymbol{A}\text{cov}(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{B}^H$$

This follows from the linearity property of the expected value operator. For simplicity assume that $\mathcal{E}(\boldsymbol{X}) = \mathcal{E}(\boldsymbol{Y}) = 0$. This is very common in applications. Then we have

$$\text{cov}(\boldsymbol{AX}, \boldsymbol{BY}) = \mathcal{E}(\boldsymbol{AX}(\boldsymbol{BY})^H) = \mathcal{E}(\boldsymbol{AXY}^H\boldsymbol{B}^H) = \boldsymbol{A}\mathcal{E}(\boldsymbol{XY^H})\boldsymbol{B^H} = \boldsymbol{A}\text{cov}(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{B}^H$$

Characteristic Function

The characteristic function for a random variable X is defined as

$$\Phi_X(\omega) = \mathcal{E}\left(e^{j\omega X}\right)$$

Note that in the case that the PDF for X exists, this is obtained as

$$\Phi_X(\omega) = \int_{-\infty}^{\infty} f_X(x) e^{j\omega x} dx$$

which can be viewed as the Fourier Transform (FT) of the PDF. Note that usually we defined a FT using a negative sign in the exponent, but it can equally be defined with a positive sign. Hence we

may say that the characteristic function for a random variable is the FT of the probability density function. Note that if the density function does not exist we could use a more general definition of Fourier Transforms based on measure theory and the same comment would apply. In such a case we would write

$$\Phi_X(\omega) = \int_{\Omega} e^{j\omega x} dP(x)$$

In this case we would say that we are taking the FT of a measure, i.e. **measures have Fourier Transforms.**

The characteristic function is very useful in the study of sums of independent random variables. If X and Y and two independent random variables and we define $Z = X + Y$. Then the characteristic function of Z is

$$\Phi_Z(\omega) = \mathcal{E}\left(e^{j\omega Z}\right) = \mathcal{E}\left(e^{j\omega(X+Y)}\right) = \mathcal{E}\left(e^{j\omega X} e^{j\omega Y}\right).$$

Assuming that the joint PDF exists, then

$$\mathcal{E}\left(e^{j\omega X} e^{j\omega Y}\right) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{j\omega x} e^{j\omega y} f_{XY}(x, y) dx dy$$

.

But due to independence of the r.v.'s X, Y, the PDF factors as

$$f_{XY}(x, y) = f_X(x) f_Y(y)$$

.

Substituting this in the double integral, the double integral becomes a product of two integrals and we obtain

$$\mathcal{E}\left(e^{j\omega X} e^{j\omega Y}\right) = \mathcal{E}(e^{j\omega X}) \mathcal{E}(e^{j\omega Y})$$

Hence

$$\Phi_Z(\omega) = \Phi_X(\omega) \Phi_Y(\omega)$$

From the theory of Fourier Transforms we know that there is an Inverse Fourier Transform, IFT, which can be used to obtain the density function from the characteristic function. In this case it becomes

$$f_Z(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi_Z(\omega) e^{-j\omega z} d\omega$$

The FT has many properties that are often tabulated. An important property is that the FT of the convolution of two functions is equal to the product of the Fourier Transforms of the two functions. Consider the two functions to be $f_X(x)$ and $f_Y(x)$.

The convolution is defined as $(f_X * f_Y)(x) = \int_{-\infty}^{\infty} f_X(t) f_Y(x-t) dt$.

According to a property of the FT, the FT of $(f_X * f_Y)(x)$ is $\Phi_X(\omega)\Phi_Y(\omega)$. As a result we can say that if the random variables X and Y are independent then the PDF for the sum random variable is given by the convolution of the two PDFs.

The procedure to find the PDF of $Z = X + Y$, where X, Y, are independent is

1. Determine $\Phi_X(\omega)$, and $\Phi_Y(\omega)$, i.e. evaluate the FTs of the densities.
2. Form the product $\Phi_Z(\omega) = \Phi_X(\omega)\Phi_Y(\omega)$
3. Find the Inverse Fourier Transform (IFT) of $\Phi_Z(\omega)$.

Characteristic Function of a Vector Random Variable

We can also define the characteristic function for the vector random variable $\boldsymbol{X} = (X_1, \cdots, X_n)^T$ as

$$\Phi_{\boldsymbol{X}}(\boldsymbol{\omega}) = \mathcal{E}\left(e^{j\boldsymbol{\omega}^T\boldsymbol{X}}\right) = \mathcal{E}\left(e^{j(\omega_1 X_1 + \cdots + \omega_n X_n)}\right)$$

Where $\boldsymbol{\omega}$ is now an n-component column vector.

If the joint density function for X exists then this becomes an n-dimensional Fourier Transform as follows

$$\Phi_{\boldsymbol{X}}(\boldsymbol{\omega}) = \int_{\mathcal{R}^n} e^{j(\omega_1 x_1 + \cdots + \omega_n x_n)} f_{\boldsymbol{X}}(x_1, \cdots, x_n) dx_1 \cdots dx_n$$

The density can be obtained from the characteristic function through the n-dimensional IFT as

$$f_{\boldsymbol{X}}(x_1, \cdots, x_n) = \frac{1}{(2\pi)^n} \int_{\mathcal{R}^n} \Phi_{\boldsymbol{X}}(\boldsymbol{\omega}) e^{-j(\omega_1 x_1 + \cdots + \omega_n x_n)} \, d\omega_1 \cdots d\omega_n$$

Note that this can also be used to study sums of independent vector random variables, but its use is not as common as in the case of ordinary random variables, where it is typical in studying the PDF of the sum of an arbitrary number of independent random variables.

Transformation of a Random Variable

Let X be a random variable and $g(\cdot)$ be a function $g: \mathcal{R} \to \mathcal{R}$ that is measurable. Then we may form $Y = g(X)$. Y is then also a random variable. A common problem is to find the CDF or the PDF of Y from the CDF or PDF of X. Usually we first determine the CDF and then differentiate to obtain the PDF.

The CDF for Y is obtained as follows:

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$$

This can be computed as follows:

$$P(g(X) \leq y) = \int_{g(x) \leq y} f_X(x)dx$$

This means that we integrate the PDF for X over a set $\{x: g(x) \leq y\}$.

There is one case where this is relatively easy, i.e. if the function $g(\cdot)$ is monotone increasing. In this case the inverse of $g(\cdot)$ exists and the condition $g(X) \leq y$ is equivalent to the condition $X \leq g^{-1}(y)$.

Hence

$$P(g(X) \leq y) = P\big(X \leq g^{-1}(y)\big) = F_X(g^{-1}(y))$$

So we have

$$F_Y(y) = F_X(g^{-1}(y))$$

The PDF for Y is then obtained in the normal manner as the derivative of the CDF

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

And using the chain rule we have

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

Note that the above can easily be modified if $g(\cdot)$ is a monotone decreasing function.

For the general case where $g(\cdot)$ is neither monotone increasing or monotone decreasing, we have to break up the function into branches in order to compute $F_Y(y)$. The computation may be tedious but the principle is the same.

Example

For example if $g(X) = X^2$, then we need to find the probability $P(\omega \in \Omega: X^2(\omega) \leq y)$. As an example suppose that X has the uniform distribution on $[-2,2]$, i.e. $f_X(x) = \frac{1}{4}$ for $x \in [-2,2]$ and 0 elsewhere. We consider three cases for y

i) For $y < 0$, clearly $P(Y \leq y) = 0$.

ii) For $0 \leq y \leq 4$, $g(X) = X^2 \leq y$, if and only if $X \in \big[-\sqrt{y}, \sqrt{y}\big]$, hence $P(Y \leq y) = P\big(X \in \big[-\sqrt{y}, \sqrt{y}\big]\big) = \frac{1}{4}\big(2\sqrt{y}\big) = \frac{\sqrt{y}}{2}$.

iii) For $y > 4$, $P(Y \leq y) = 1$.

The CDF for Y is then

$$F_Y(y) = \begin{cases} 0 & \text{for} \quad y < 0 \\ \dfrac{\sqrt{y}}{2} & \text{for} \quad 0 \leq y \leq 4 \\ 1 & \text{for} \quad y > 4 \end{cases}$$

and the PDF for Y is

$$f_Y(y) = \begin{cases} 0 & \text{for} \quad y < 0 \\ \dfrac{1}{4\sqrt{y}} & \text{for} \quad 0 \leq y \leq 4 \\ 0 & \text{for} \quad y > 4 \end{cases}$$

Transformation of a Vector Random Variable

Transformations on vector random variables lead to new random variables for which we can determine the CDF. The simplest case is a real-valued function, i.e transform a vector valued random variable to a scalar valued random variable. For example, let $Z = g(X_1, \cdots, X_n)$. The CDF for Z is $F_Z(z) = P(Z \leq z) = P(\boldsymbol{X} \in R_z)$ where

$$R_z = (\boldsymbol{x} \in \mathcal{R}^{\mathrm{n}} : g(\boldsymbol{x}) \leq z)$$

If the density $f_{\boldsymbol{X}}(x_1, \cdots, x_n)$ exists then we can evaluate the above as follows:

$$F_Z(z) = P(\boldsymbol{X} \in R_Z) = \int_{\boldsymbol{x} \in R_z} f_{\boldsymbol{X}}(x_1, \cdots, x_n) dx_1 \cdots dx_n$$

Note that if we have to do these evaluations they may be very difficult because they involve multi-dimensional integrals over general regions that are not hyper-rectangles. Similar conceptual ideas apply if we are transforming a random vector to another random vector.

Example

Let $Z = X + Y$, where X and Y are random variables. Find the CDF and PDF for Z.

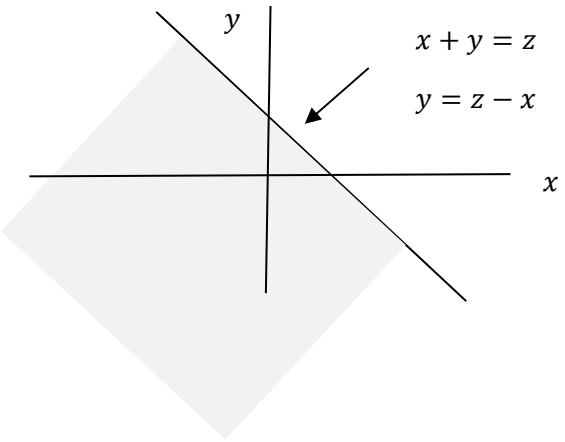
$F_Z(z) = P(Z \leq z) = P(X + Y \leq z) = P\big((X, Y) \in \text{shaded region}\big)$ (see Figure below)

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{z-v} f_{XY}(u, v) du dv$$

$$f_Z(z) = \frac{d}{dz}F_Z(z) = \frac{d}{dz}\left(\int_{-\infty}^{\infty}\int_{-\infty}^{z-v} f_{XY}(u,v)dudv\right)$$
$$= \int_{-\infty}^{\infty}\left(\frac{d}{dz}\int_{-\infty}^{z-v} f_{XY}(u,v)du\right)dv$$
$$f_Z(z) = \int_{-\infty}^{\infty} f_{XY}(z-v,v)dv$$

Now, if the random variables X and Y are independent, then $f_{XY}(z-v,v) = f_X(z-v)f_Y(v)$
Hence the above integral becomes a convolution of the PDFs for X and Y, i.e.

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-v)f_Y(v)dv$$



Invertible Functions of Random Vectors

Let $\boldsymbol{X}$ be a random vector with joint PDF $f_{\boldsymbol{X}}(\boldsymbol{x}), \quad \boldsymbol{x} = (x_1, \cdots, x_n)$.
Define $\boldsymbol{Y} = g(\boldsymbol{X})$, where $\boldsymbol{Y}$ is an n-vector, i.e.

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} g_1(X_1, \cdots, X_n) \\ \vdots \\ g_n(X_1, \cdots, X_n) \end{bmatrix}$$

Assume that $g(\cdot)$ has an inverse, i.e. there exists an $h(\cdot)$ (vector valued function of a vector) such that $h\big(g(\boldsymbol{X})\big) = \boldsymbol{X}$.

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} h_1(Y_1, \cdots, Y_n) \\ \vdots \\ h_n(Y_1, \cdots, Y_n) \end{bmatrix}$$

Now fix $\boldsymbol{Y} = \boldsymbol{y}$, evaluate $\nabla h(\boldsymbol{y}) = \begin{bmatrix} \frac{\partial h_1}{\partial y_1} & \cdots & \frac{\partial h_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_n}{\partial y_1} & \cdots & \frac{\partial h_n}{\partial y_n} \end{bmatrix}$

The Jacobian is defined as $J_h(\boldsymbol{y}) = \det\big(\nabla h(\boldsymbol{y})\big) = \begin{vmatrix} \frac{\partial h_1}{\partial y_1} & \cdots & \frac{\partial h_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_n}{\partial y_1} & \cdots & \frac{\partial h_n}{\partial y_n} \end{vmatrix}$

Then we have

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}\big(h(\boldsymbol{y})\big)|J_h(\boldsymbol{y})|$$

We can also use the Jacobian with respect to the function $g(\cdot)$

$$J_g(\boldsymbol{x}) = \det\big(\nabla g(\boldsymbol{x})\big) = \begin{vmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_n} \end{vmatrix}$$

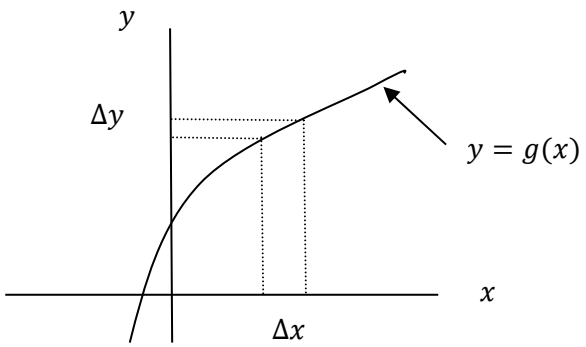
Then

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{f_{\boldsymbol{X}}\big(h(\boldsymbol{y})\big)}{|J_g(\boldsymbol{x})|} = \frac{f_{\boldsymbol{X}}\big(h(\boldsymbol{y})\big)}{|J_g(\boldsymbol{h}(\boldsymbol{y}))|}$$

Note, when we transform a single random variable with a function for which the inverse exists, we have $Y = g(X)$, or $X = h(Y)$. Sometimes we derive the density function directly using the following heuristic (for monotone increasing)

$$f_Y(y)dy = f_X(x)dx$$

Then formally dividing by dy we obtain $f_Y(y) = f_X(x)\frac{dx}{dy} = f_X(x)\left(\frac{dy}{dx}\right)^{-1} = f_X\big(h(y)\big)\left(\frac{dy}{dx}\right)^{-1}$.
To see this more clearly consider the Figure below.



$$P(x < X < x + \Delta x) = P(y < Y < y + \Delta y)$$

$$f_X(x)\Delta x = f_Y(y)\Delta y$$

$$f_X(x) = f_Y(y)\frac{\Delta y}{\Delta x}$$

Take the limit as $\Delta x \to 0$ $$f_X(x) = f_Y(y)\frac{dy}{dx}$$

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

Now $\frac{dy}{dx}$ plays the role of $J(\boldsymbol{X})$ and $\left(\frac{dy}{dx}\right)^{-1} = \frac{dx}{dy}$ plays the role of $J(\boldsymbol{Y})$, where $\lim_{x\to 0}\frac{\Delta x}{\Delta y} = \frac{dx}{dy}$.

For a monotone decreasing function we would write the above as $f_Y(y)dy = -f_X(x)dx$. To handle both cases we would then replace $\frac{dy}{dx}$ by $\left|\frac{dy}{dx}\right|$.

To summarize the above, the Jacobian generalizes the derivative to n-dimensional pace. In one dimensional space an interval of length Δx is mapped into an interval of length Δy by the function $g(\cdot)$. Hence the amount of probability in the $X-$ domain, i.e. $f_X(x)\Delta x$ is mapped to $f_Y(y)\Delta y$ in the Y- domain. $f_Y(y)\frac{\Delta y}{\Delta x} = f_X(x)$. For a function $\mathcal{R}^n \to \mathcal{R}^n$, the Jacobian $J_g(x)$ plays the role of $\lim_{\Delta x\to 0}\frac{\Delta y}{\Delta x} = \frac{dy}{dx} = \frac{dg}{dx}$.

Example

Consider the above scenario in two dimensions. To avoid subscripts we will replace $\boldsymbol{X} = (X_1, X_2)$ with (X, Y), and $\boldsymbol{Y} = (Y_1, Y_2)$ with (U, V), i.e. $(X, Y) \to (U, V)$. The function $g(\cdot)$ is defined as follows
$U = X + Y,\ V = \frac{X}{X+Y}$. Note that the inverse is obtaining by solving these two equations for X and Y. $X = UV, Y = U(1-V)$.

Assume that X and Y are independent, and $f_X(x) = \frac{\mu e^{-\mu x}(\mu x)^{m-1}}{(m-1)!}\Phi(x)$, and $f_Y(y) = \frac{\mu e^{-\mu y}(\mu y)^{k-1}}{(k-1)!}\Phi(y)$, where $\Phi(\cdot)$ is the step function. These are Erlang distributions with parameters, μ, m and k, where $\mu > 0$, and $m \geq 1, k \geq 1$. Then we need to consider $U \in [0, \infty)$ and $V \in [0,1]$.

The Jacobian is $J(x,y) = \begin{vmatrix} \frac{\partial U}{\partial X} & \frac{\partial U}{\partial Y} \\ \frac{\partial V}{\partial X} & \frac{\partial V}{\partial Y} \end{vmatrix} = \begin{vmatrix} \frac{1}{y} & \frac{1}{-x} \\ \frac{}{(x+y)^2} & \frac{}{(x+y)^2} \end{vmatrix} = -\frac{x}{(x+y)^2} - \frac{y}{(x+y)^2} = -\frac{1}{x+y} = -\frac{1}{u}$

The joint PDF for U, V, is then $f_{UV}(u,v) = \frac{1}{|J(x,y)|}\left(\frac{\mu e^{-\mu x}(\mu x)^{m-1}}{(m-1)!} \cdot \frac{\mu e^{-\mu y}(\mu y)^{k-1}}{(k-1)!}\right)_{\substack{x=uv \\ y=u(1-v)}}$

$$= u\frac{e^{-\mu uv}(\mu uv)^{m-1}}{(m-1)!} \cdot \frac{\mu e^{-\mu u(1-v)}(\mu u(1-v))^{k-1}}{(k-1)!}$$

$$= \frac{u(\mu uv)^{m-1}}{(m-1)!}\frac{\mu e^{-\mu u}\big(\mu u(1-v)\big)^{k-1}}{(k-1)!}$$

$$= \frac{\mu e^{-\mu u}(\mu u)^{m+k-1}}{(m-1)!} \cdot \frac{v^{m-1}(1-v)^{k-1}}{(k-1)!}$$

$$= \frac{\mu e^{-\mu u}(\mu u)^{m+k-1}}{(m+k-1)!} \cdot \frac{(m+k-1)!}{(m-1)!\,(k-1)!}v^{m-1}(1-v)^{k-1}$$

$$= f_U(u)f_V(v)$$

Where $f_U(u) = \frac{\mu e^{-\mu u}(\mu u)^{m+k-1}}{(m+k-1)!}$ for $u \geq 0$, 0 elsewhere

And $f_V(v) = \frac{(m+k-1)!}{(m-1)!(k-1)!}v^{m-1}(1-v)^{k-1}$, for $0 \leq v \leq 1$, 0 elsewhere

Hence U is Erlang and V is Beta.

Linear Transformation of a Vector Random Variable

Assume $\boldsymbol{X} = (X_1, \cdots, X_n)$ is a vector random variable. Let $\boldsymbol{Y} = \boldsymbol{AX}$, where $\boldsymbol{A}$ is an $n \times n$ matrix. Note that $Y_j = a_{j1}X_1 + \cdots + a_{jn}X_n$

Since $\frac{dY_j}{dX_i} = a_{ji}$, the Jacobian is $J(x) = \begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{vmatrix} = \det(A)$

Assuming that $\det(A) \neq 0$, i.e. A is invertible, we have

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{|\det(\boldsymbol{A})|}\Bigg|_{\boldsymbol{x}=\boldsymbol{A}^{-1}\boldsymbol{y}} = \frac{f_{\boldsymbol{X}}(\boldsymbol{A}^{-1}(\boldsymbol{y}))}{|\det(\boldsymbol{A})|}$$

Note that in the above expression $\boldsymbol{x}, \boldsymbol{y}$ are n-vectors.

Covariance Matrix of a Linear Transformation

Assume $\boldsymbol{X}$, $\boldsymbol{Y}$ are random vectors (column vectors) with $\boldsymbol{Y} = \boldsymbol{AX}$, where $\boldsymbol{A}$ is a constant matrix. Also for simplicity we assume that $\mathcal{E}(\boldsymbol{X}) = 0$. Then the covariance of $\boldsymbol{Y}$ can be determined as follows:

$$\text{cov}(\boldsymbol{Y}) = \mathcal{E}(\boldsymbol{YY}^T) = \mathcal{E}(\boldsymbol{AX}(\boldsymbol{AX})^T) = \boldsymbol{A}\mathcal{E}(\boldsymbol{XX}^T)\boldsymbol{A}^T = \boldsymbol{AC_X A}^T$$

where $\boldsymbol{C_X} = \text{cov}(\boldsymbol{X})$.

Karhunen-Loeve (KL) Expansion

Let $\boldsymbol{X}$ be an n-dimensional random vector with zero mean and covariance matrix $\boldsymbol{C_X}$. Assume that $\det(\boldsymbol{C_X}) \neq 0$, i.e. $\boldsymbol{C_X}$ is non-singular. A general covariance matrix, as we have seen before, is positive semi-definite, i.e. for any constant n-vector $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}^T\boldsymbol{C_X}\boldsymbol{\alpha} \geq 0$. However if $\boldsymbol{C_X}$ is non-singular then it is positive definite, i.e. for any non-zero such $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}^T\boldsymbol{C_X}\boldsymbol{\alpha} > 0$.
Since $\boldsymbol{C_X}$ is symmetric and positive definite we can form the diagonalization of the matrix as

$$\boldsymbol{C_X} = \boldsymbol{P\Lambda P}^T$$

where $\boldsymbol{P}$ is a matrix whose columns are the eigenvectors of $\boldsymbol{C_X}$, and $\boldsymbol{\Lambda}$ is a diagonal matrix whose diagonal elements are the eigenvalues. If the eigenvalues are distinct the eigenvectors are orthogonal. If the eigenvalues are not distinct then we can still choose a set of orthogonal eigenvectors. We can also, assume that the eigenvectors are normalized. Hence $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix.

Define $\boldsymbol{Y} = \boldsymbol{P}^T\boldsymbol{X}$, then $\text{cov}(\boldsymbol{Y}) = \mathcal{E}(\boldsymbol{YY}^T) = \mathcal{E}(\boldsymbol{P}^T\boldsymbol{X}(\boldsymbol{P}^T\boldsymbol{X})^T) = \mathcal{E}(\boldsymbol{P}^T\boldsymbol{XX}^T\boldsymbol{P}) = \boldsymbol{P}^T\boldsymbol{C}_X\boldsymbol{P} = \boldsymbol{P}^T(\boldsymbol{P\Lambda P}^T)\boldsymbol{P} = \boldsymbol{\Lambda}$

As a result we conclude that for a random vector with non-singular covariance matrix we can find a matrix A, such that the random vector $Y = AX$ is a random vector with uncorrelated components.

Joint Gaussian Random Variables (Gaussian Random Vectors)

A single random variable X is said to be Gaussian distributed if the PDF is given as follows:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-m)^2\right)$$

Where $m = \mathcal{E}(X)$, and σ is the standard deviation, i.e. square root of the variance.

Gaussian Random Vectors
A random vector $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^{\boldsymbol{T}}$ is said to be Gaussian distributed if **for every non-zero vector** $\alpha = (\alpha_1, \cdots, \alpha_n)$**, the linear combination** $Y = \sum_{i=1}^{n}\alpha_i X_i$ **is Gaussian** distributed. Another

way to say that a random vector is Gaussian distributed is to say that the components are jointly Gaussian distributed.

There are other equivalent definitions of a jointly Gaussian distributed set of n random variables. A common one that applies if the covariance matrix, $\boldsymbol{C_X}$ which we denote here as $\boldsymbol{K}$, is non-singular, is that the random vector $\boldsymbol{X}$ is Gaussian if the joint PDF of its components X_i is given as follows:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{K}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T\boldsymbol{K}^{-1}(\boldsymbol{x}-\boldsymbol{m})\right) \qquad (*)$$

where $\boldsymbol{m} = \mathcal{E}(\boldsymbol{X})$ is the mean and $\boldsymbol{K}$ is the covariance matrix. $|\boldsymbol{K}|$ is the determinant of $\boldsymbol{K}$. Note that $\boldsymbol{x} = (x_1, \cdots, x_n)$ is a vector in this case. Note that as we have discussed before $\boldsymbol{K} = \left(\mathcal{E}\big((X_i - m_i)(X_j - m_j)\big)\right)$, and this matrix is symmetric. We represent this probability law by $\mathcal{N}(\boldsymbol{m}, \boldsymbol{K})$.

Theorem: Let $\boldsymbol{X}$ be a random vector with non-singular covariance matrix. Then the random variable $Y = \boldsymbol{\alpha^T X}$, where $\boldsymbol{\alpha}$ is a non-zero constant vector, is Gaussian distributed if and only if the joint PDF for $\boldsymbol{X}$ is given by the above (*).

This theorem states that the definition of a Gaussian vector is either that the joint PDF is given by (*) or all non-zero linear combinations yield random variables that are Gaussian.

Uncorrelated Gaussian Vectors

Now, assume that the covariance matrix, $\boldsymbol{K}$, for a Gaussian random vector is a diagonal matrix with components in the diagonal given by $\mathcal{E}((X_i - m_i)^2) = \sigma_i^2$, where $m_i = \mathcal{E}(X_i)$. Then we can show that the expression in the exponent, in the above (*), becomes $-\frac{1}{2}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}(x_i - m_i)^2$. The determinant of $\boldsymbol{K}$ is then $|\boldsymbol{K}| = \Pi_{i=1}^{n}\sigma_i^2$. The PDF then factors as $f_X(x) = \Pi_{i=1}^{n} f_{X_i}(x_i)$. This means that the random variables X_i are independent. The joint PDF is equal to the product of the marginal PDF's. As we have shown previously the converse is also true for any joint probability law, including the Gaussian, i.e., if the components X_i are independent r.v.'s then the covariance matrix is a diagonal matrix.

The Characteristic Function of a Gaussian Random Variable

As we have introduced before the characteristic function of a random variable is defined as $\Phi_X(\omega) = \mathcal{E}\left(e^{j\omega X}\right)$. For a Gaussian random variable we can evaluate this as

$$\Phi_X(\omega) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) e^{j\omega x} dx$$

This integral can be evaluated using standard techniques to obtain

$$\Phi_X(\omega) = \exp\left(j\omega m - \frac{\sigma^2\omega^2}{2}\right)$$

In a similar manner we previously defined the characteristic function for a random vector $\boldsymbol{X}$ as $\Phi_{\boldsymbol{X}}(\boldsymbol{\omega}) = \mathcal{E}(e^{j\boldsymbol{\omega}^T\boldsymbol{X}})$, where $\boldsymbol{\omega}$ is column a vector.

For a Gaussian random vector the characteristic function can be evaluated to yield

$$\Phi_{\boldsymbol{X}}(\boldsymbol{\omega}) = \exp\left(j\boldsymbol{\omega}^T\boldsymbol{m} - \frac{\boldsymbol{\omega}^T\boldsymbol{K}\boldsymbol{\omega}}{2}\right)$$

where $\boldsymbol{K}$ is the covariance matrix, and $\boldsymbol{m}$ is the mean. We can show that if the components of the random vector $\boldsymbol{X}$ are uncorrelated, i.e. $\boldsymbol{K}$ is a diagonal matrix, then the above exponent can be written as a sum of terms each corresponding to the exponent in the characteristics function of a Gaussian random variable.

$$j\boldsymbol{\omega}^T\boldsymbol{m} - \frac{\boldsymbol{\omega}^T\boldsymbol{K}\boldsymbol{\omega}}{2} = j\sum_{i=1}^{n}\omega_i m_i - \frac{1}{2}\sum_{i=1}^{n}\sigma_i^2\omega_i^2 = \sum_{i=1}^{n}\left(j\omega_i m_i - \frac{1}{2}\sigma_i^2\omega_i^2\right)$$

The characteristic function then factors as a product $\Phi_{\boldsymbol{X}}(\boldsymbol{\omega}) = \Pi_{i=1}^{n}\Phi_{X_i}(\omega_i)$. Taking the n-dimensional Fourier Transform we would find the joint PDF factors as $f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{X_1}(x_1)\cdots f_{X_n}(x_n)$. Hence the random variables X_i, i.e. the components of $\boldsymbol{X}$, are independent.

In conclusion if a random vector is Gaussian and the covariance matrix is a diagonal matrix then the components of the random vector are independent random variables.

Linear Combination of Independent Gaussian Random Variables

Theorem:
Let the set of r.v.s X_i $(i = 1, \cdots, n)$ be Gaussian distributed, i.e. $\mathcal{N}(m_i, \sigma_i^2)$, and **independent**. Then for any constant vector, $\boldsymbol{\alpha} \neq 0$, the linear combination $Y = \sum_{i=1}^{n}\alpha_i X_i$ is a Gaussian distributed random variable.

This follows easily from the consideration of the characteristic function. $\Phi_Y(\omega) = \mathcal{E}\left(e^{j\omega Y}\right) = \mathcal{E}\left(e^{j\omega\sum_{i=1}^{n}\alpha_i X_i}\right) = \Pi_{i=1}^{n}\mathcal{E}\left(e^{j\omega\alpha_i X_i}\right)$, where the last equality follows from the independence assumption for the X_i. Now, it can easily be shown that if a random variable is Gaussian, e.g. X_i,

then a scaled version, $\alpha_i X_i$ is also Gaussian. The mean also scales by the same factor. Hence, if X_i is $\mathcal{N}(m_i, \sigma_i^2)$, then $\alpha_i X_i$ is $\mathcal{N}(\alpha_i m_i, \alpha_i^2 \sigma_i^2)$. Hence, $\mathcal{E}\left(e^{j\omega\alpha_i X_i}\right) = e^{j\omega\alpha_i m_i - \frac{\alpha_i^2\sigma_i^2\omega^2}{2}}$. Now, taking the product of these we obtain

$$\mathcal{E}\left(e^{j\omega Y}\right) = e^{j(\omega\sum_{i=1}^n \alpha_i m_i - \frac{1}{2}\omega^2\sum_{i=1}^n \alpha_i^2\sigma_i^2)}$$

and this is the characteristic function of a Gaussian distributed random variable with mean equal to $\sum_{i=1}^n \alpha_i m_i$ and variance equal to $\sum_{i=1}^n \alpha_i^2\sigma_i^2$.

A similar statement applies to the transformation $\boldsymbol{Y} = \boldsymbol{AX} + \boldsymbol{b}$, where $\boldsymbol{A}$ is a constant matrix and $\boldsymbol{b}$ is a constant vector. If $\boldsymbol{X}$ is a Gaussian distributed random vector then so is $\boldsymbol{Y}$. If $\boldsymbol{A}$ is an $r \times n$ matrix and $\boldsymbol{b}$ is an r-vector, then $\boldsymbol{Y}$ is an r dimensional Gaussian random vector. To show this we consider a linear combination $\sum_{i=1}^n \alpha_i Y_i = \boldsymbol{\alpha}^T\boldsymbol{Y}$.

Now $\boldsymbol{\alpha}^T\boldsymbol{Y} = \boldsymbol{\alpha}^T(\boldsymbol{AX} + \boldsymbol{b}) = (\boldsymbol{\alpha}^T\boldsymbol{A})\boldsymbol{X} + \boldsymbol{\alpha}^T\boldsymbol{b}$

From the last expression we see a linear combination of the X_i plus a constant which is Gaussian by assumption since the constant merely changes the mean.

The converse is not true unless $\boldsymbol{A}$ is a non-singular square matrix.

For example, assume that $\boldsymbol{X} = (X_1, X_2)$ where X_1 is Gaussian but X_2 is not Gaussian. If we define $\boldsymbol{Y} = \boldsymbol{AX}$, where

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 2 & 0 \end{bmatrix}$$

Then $\boldsymbol{Y} = \begin{bmatrix} X_1 \\ 2X_1 \end{bmatrix}$ is Gaussian, but $\boldsymbol{X}$ is obviously not Gaussian. Why is $\boldsymbol{Y}$ a Gaussian vector? According to our definition $\boldsymbol{Y}$ is Gaussian if every linear combination $\alpha_1 Y_1 + \alpha_2 Y_2$ is Gaussian. But $\alpha_1 Y_1 + \alpha_2 Y_2 = \alpha_1 X_1 + \alpha_2(2X_1) = (\alpha_1 + 2\alpha_2)X_1$ which is Gaussian because it is a scaled version of a Gaussian r.v.

Infinitely Decomposable Random Variables

The characteristic function of a Gaussian random variable X with mean m and variance σ^2 can be written as $\Phi_X(\omega) = e^{\left(\frac{j\omega m}{N} - \frac{\sigma^2\omega^2}{2N}\right)N} = \Pi_{i=1}^N e^{\left(\frac{j\omega m}{N} - \frac{\sigma^2\omega^2}{2N}\right)}$ for an arbitrary integer N. This means that a Gaussian random variable can be written as a sum of N independent Gaussian variables $X = X_1 + \cdots + X_N$, where the X_i are $\mathcal{N}\left(\frac{m}{N}, \frac{\sigma^2}{N}\right)$. This property is called infinitely decomposable, or also infinitely divisible. Distributions that are infinitely decomposable can be the limit distributions for a sum of a large number of independent random variables, e.g. the *Central Limit Theorem*.

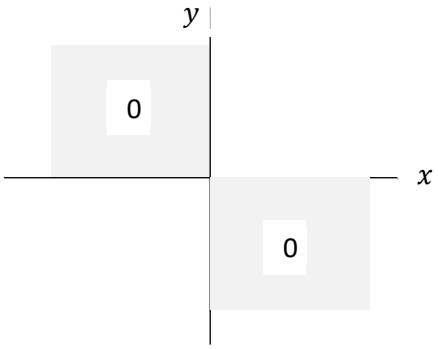
Marginally Gaussian does not Imply Jointly Gaussian

Consider the random variables X, Y with the following PDF

$$f_{XY}(x,y) = \begin{cases} \frac{1}{\pi\sigma_X\sigma_Y}\exp\left(-\frac{1}{2}\left(\frac{x^2}{\sigma_X^2}+\frac{y^2}{\sigma_Y^2}\right)\right) & \text{for } xy > 0 \\ 0 & \text{for } xy \le 0 \end{cases}$$

i.e. $f_{XY}(x,y) = 0$ on the 2nd and 4th quadrants.

Then the marginal PDFs for X and Y are both Gaussian, but not the joint PDF.
Note that on the 1st and 3rd quadrants the PDF is basically that of a 0 mean 2-dimensional Gaussian random vector, scaled by 2. For the marginal PDF for X we "collapse" the probability on the X axis, where for the marginal PDF for Y we "collapse" the probability on the Y axis.



Inequalities in Calculation of Probabilities

In many applications it is difficult to calculate a probability and wish to obtain an approximation by means of an upper bound. There are a few of these well known inequalities, or bounds on the probability.

Markov Inequality

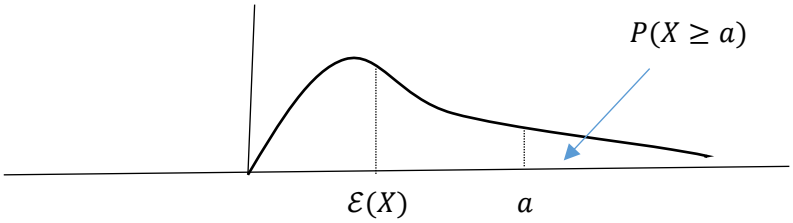
If X is an non-negative random variable, i.e. $P(X < 0) = 0$, with finite expected value, then for any $a > 0$

$$P(X \ge a) \le \frac{\mathcal{E}(X)}{a}$$

Proof: Assuming the density exists, but it applies in all cases,

$$aP(X \ge a) = \int_a^\infty a f_X(x)dx \le \int_a^\infty x f_X(x)dx \le \int_0^\infty x f_X(x)dx = \mathcal{E}(X)$$

Note that we get an essential property from this bound, as $a \to \infty$, the bound approaches 0. Note that if the bound predicted that the probability would be, for example, less than ½, it would be true but not a useful bound.



Example

Let X be a Poisson random variable with parameter $\lambda = 1/2$. Find an upper bound for $P(X \geq 3)$. Note that for a Poisson r.v. $\mathcal{E}(X) = \lambda = 1/2$. We can also calculate $P(X \geq 3) = \sum_{k=3}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = 1 - \big(P(X = 0) + P(X = 1) + P(X = 2)\big) = 0.0144$. The bound predicts

$$P(X \geq 3) \leq \left(\frac{1}{2}\right)/3 = \frac{1}{6} = 0.167$$

The bound is somewhat loose.

There are several other bounds that are related to this one but are applicable in more restrictive case. Assume that X is a positive random variable, $r \geq 1$, and $\mathcal{E}(X^r)$ exists, then we have

$$P(X \geq a) \leq \frac{\mathcal{E}(X^r)}{a^r}$$

The proof is similar to the above

$$a^r P(X \geq a) = \int_a^{\infty} a^r f_X(x)dx \leq \int_a^{\infty} x^r f_X(x)dx \leq \int_0^{\infty} x^r f_X(x)dx = \mathcal{E}(X^r)$$

Example

Let X be a Poisson distributed r.v. as above. Using the inequality with $r = 2$ we obtain

$$P(X \geq 3) \leq \frac{\mathcal{E}(X^2)}{3^2} = \frac{3/4}{3^2} = 0.0833$$

This is tighter that the Markov bound.

Chebychev Inequality

A related bound to the above is the following. This applies to random variables for which the variance exists, which is the same as the requirement that the 2nd moment exists (as above),

Let X be a random variable with mean m_X and variance σ_X^2, then $P(|X - m_X| \geq a) \leq \frac{\sigma_X^2}{a^2}$.

Proof: Let $Y = (X - m_X)^2$, then

$$P(|X - m_X| \geq a) = P(|X - m_X|^2 \geq a^2) = P(Y \geq a^2) \leq \frac{\mathcal{E}(Y)}{a^2} = \frac{\sigma_X^2}{a^2}$$

Where the last inequality is due to the Markov inequality.

Again using the example of the Poisson distribution above ($\lambda = 1/2$),

$$P(X \geq 3) = P\left(X - \frac{1}{2} \geq 3 - \frac{1}{2}\right) = P\left(\left|X - \frac{1}{2}\right| \geq 2.5\right) \leq \frac{\sigma_X^2}{2.5^2} = \frac{1/2}{6.25} = 0.08$$

We can see that the Chebychev bound is tighter than the Markov bound but it requires that the variance exists, which is the same as requiring that the 2nd moment exists. There are tighter bounds if higher moments exist.

Chernoff Bound

This is another bound that is similar to the above but is more restrictive in the requirements for the random variable X.

Let X be an arbitrary random variable with the requirement that for any $s > 0$, $\mathcal{E}(e^{sX})$ exists. Then for all $s > 0$, $P(X > a) \leq e^{-sa}\mathcal{E}(e^{sX})$.

Note that this is a stronger requirement on X than requiring that all the moments exist. In fact if we let $M_X(s) = \mathcal{E}(e^{sX})$ then $\mathcal{E}(e^{sX}) = \mathcal{E}\left(1 + sX + \frac{(sX)^2}{2!} + \frac{(sX)^3}{3!} + \cdots\right) = \sum_{k=0}^{\infty} \frac{\mathcal{E}(sX)^k}{k!}$
Hence $M_X'(s)|_{s=0} = \mathcal{E}(X)$, and in general for $k \geq 1$, $M^{(k)}(s)|_{s=0} = \mathcal{E}(X^k)$, where the superscript (k) denotes the k^{th} derivative. As a result $M_X(s)$ is also called the moment generating function.

Proof

Pick $s > 0$

$e^{sa}P(X > a) = \int_a^{\infty} e^{sa} f_X(x)dx \leq \int_a^{\infty} e^{sx} f_X(x)dx \leq \int_{-\infty}^{\infty} e^{sx} f_X(x)dx = \mathcal{E}(e^{sX})$.

The result follows.

Note that in some applications we write this bound in terms of a general parameter s and then optimize it with respect to s, i.e. we find the minimum over all $s > 0$, i.e. we evaluate

$$P(X > a) \leq \min_{s>0} e^{-sa}\mathcal{E}(e^{sX}) = \min_{s>0} e^{-sa}M_X(s)$$

Example

We consider the same example as above with the Poisson distribution with $\lambda = 1/2$. Let $a = 3$. Then $M_X(s) = \sum_{k=0}^{\infty} e^{sk}\frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=0}^{\infty}\frac{(\lambda e^s)^k}{k!}e^{-\lambda} = e^{\lambda(e^s-1)}$.
Hence $s^{-sa}M_X(s) = e^{-sa}e^{\lambda(e^s-1)} = e^{\lambda(e^s-1)-as}$. Now to minimize over s we take the derivative and set to 0.

$$e^{\lambda(e^s-1)-as}(\lambda e^s - a) = 0$$
$$\lambda e^s - a = 0$$
$$s = \ln\left(\frac{a}{\lambda}\right) = \ln\left(\frac{3}{1/2}\right) = \ln 6$$

Hence $P(X \geq 3) < \exp\left(\frac{1}{2}\left(e^{\ln 6} - 1\right) - 3(\ln 6)\right) = \exp\left(\frac{1}{2}(6-1) - 3\ln 6\right) = \exp(2.5 - 3\ln 6) = 0.0564$.

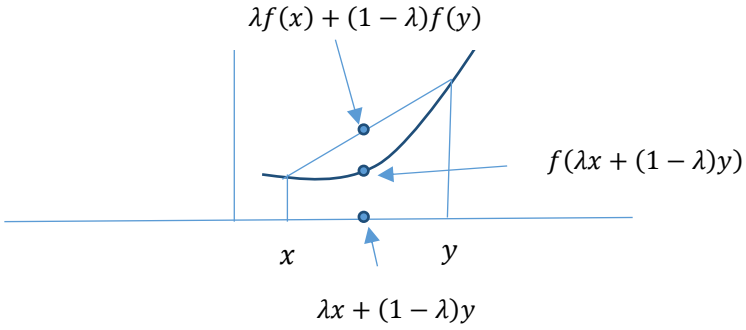
Note that this is tighter than the Chebychev bound which was 0.08.

Jensen's Inequality

Jensen's inequality is yet another similar bound which generalizes the previous bounds.

Let X be a random variable with finite mean. Let $f: \mathcal{R} \to \mathcal{R}$ be a convex function, i.e. f satisfies the following:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$



For $x, y, \lambda \in \mathcal{R}$, and $0 < \lambda < 1$. Then $f\left(\mathcal{E}(X)\right) \leq \mathcal{E}(f(X))$.

Idea of proof:

Consider a random variable X on the X-axis. Consider the fixed point $\mathcal{E}(X)$ on the X-axis. Draw a tangent to the curve $f(x)$ at this fixed point. Then let the random variable X vary close to the fixed point. Then $f(X) \geq f\big(\mathcal{E}(X)\big) + f'\big(\mathcal{E}(X)\big)\big(X - \mathcal{E}(X)\big)$. This is like the first term of a Taylor expansion about the point $E(X)$ which is a constant. Take the expected value of both sides and obtain

$$\mathcal{E}\big(f(X)\big) \geq \mathcal{E}(f\big(\mathcal{E}(X)\big) + \mathcal{E}\left(f'\big(\mathcal{E}(X)\big)\big(X - \mathcal{E}(X)\big)\right)$$

Note that in the above $\mathcal{E}(X)$ is a constant. Hence $f(\mathcal{E}(X))$, and $f'(\mathcal{E}(X))$ are constants. Hence the first term on the right side is $f(\mathcal{E}(X))$ and the second term on the right side is $f'\big(\mathcal{E}(X)\big)\big(\mathcal{E}(X) - \mathcal{E}(X)\big) = 0$. Thus we have $\mathcal{E}\big(f(X)\big) \geq f\big(\mathcal{E}(X)\big)$.

Alternate proof:

Let X be a random variable with expected value $m = \mathcal{E}(X)$. Let the function $f(\cdot)$ be as above.

Consider the random variables $Y = g(X)$ and $Z = h(X)$. Assume that for all x $h(x) \geq g(x)$.
We say that $h(\cdot)$ majorizes $g(\cdot)$. Then $\mathcal{E}\big(g(X)\big) \leq \mathcal{E}(h(X))$

Note $\mathcal{E}\big(h(X) - g(X)\big) = \int_{-\infty}^{\infty}\big(h(x) - g(x)\big)f_X(x)dx \geq 0$, i.e. $\mathcal{E}\big(h(X)\big) \geq \mathcal{E}(g(X))$.

Consider the two functions, $f(x)$ and $g(x) = f(m) + f'(m)(x - m)$. Note that $f(\cdot)$ majorizes $g(\cdot)$. Hence we must have $\mathcal{E}\big(f(X)\big) \geq \mathcal{E}(g(X))$. But $\mathcal{E}\big(g(X)\big) = \mathcal{E}\big(f(m) + f'(m)(X - m)\big) = f(m) = f(\mathcal{E}(X))$.
Hence $\mathcal{E}\big(f(X)\big) \geq f(\mathcal{E}(X))$.

Convergence of Sequences of Random Variables

If $Z_1, Z_2, \cdots$ is a sequence of random variables and Z is a random variable, all on a probability space $(\Omega, \mathcal{F}, P)$, what does it mean to say that the sequence $Z_1, Z_2, \cdots$ converges to Z as $n \to \infty$? There are different notions of convergence including what we refer to as strong and weak convergence and convergence in distribution.

Convergence of Sequences of Real Numbers

Let us consider a sequence of real numbers, $r_1, r_2, \cdots$
This sequence converges to a number r, if for every $\epsilon > 0$, there exists and integer N such that $|r - r_n| < \epsilon$ for all $n \geq N$. For us to verify convergence using this definition we need to know the limit, r in this case.

In some cases we have a sequence that we would like to test for convergence but we don't know the limit.
Suppose we don't know the limit but still would like to verify convergence. We can use the so-called *Cauchy Criteria* for convergence of a sequence.

Cauchy Criteria for Convergence of a Sequence: The sequence $\{r_n\}$ converges to some value if give any $\epsilon > 0$ there exists an integer N, such that for all integers $m, n \geq N$, $|r_n - r_m| \leq \epsilon$.

Sequences of Real Functions

Since random variables are functions, in order to consider convergence of sequences of random variables, first we consider the notion of convergence of a sequence of real functions on $\mathcal{R}$. Let $f_1, f_2, \cdots$ be a sequence of functions, $f_n: \mathcal{R} \to \mathcal{R}$. We consider the following different types of convergence of the sequence of functions $\{f_n\}$ to a function f.

Pointwise convergence:

A sequence of functions f_n converges pointwise to the function f if for each $x \in \mathcal{R}$ $\lim_{n\to\infty} f_n(x) = f(x)$.

This means that for each $x \in \mathcal{R}$. Pick $\epsilon > 0$, then $\exists$ an integer N, such that $\forall\, n > N$, $|f(x) - f_n(x)| < \epsilon$.

Here the rate of convergence depends on the point x. For example consider functions on the interval $(0,1)$. Let $f_n(x) = x^n$. Then the sequence f_n converges pointwise to $f = 0$. But the convergence gets slower and slower as x approaches 1. For a given ϵ, if we consider an x closer 1, then the N in the convergence test becomes larger.

We can select different values for x for which the converge rate is arbitrarily slow! In other words we need a greater value of N in the convergence test.

To show this consider $x = 1 - \epsilon_1$. To test for convergence pick $\epsilon > 0$. Then consider $x^N = (1 - \epsilon_1)^N$. Now set $(1 - \epsilon_1)^N < \epsilon$. Solve for N and obtain $N > \frac{\log(\epsilon)}{\log(1-\epsilon_1)}$. We can see that the required N becomes larger as ϵ_1 approaches 0, i.e. x approaches 1.

In the modern analysis of convergence of sequences of functions, e.g. in Fourier Series, we usually do not consider what happens to the limit function in some set of discrete points and generally

disregard what happens to the convergence in sets of measure zero. In other words we do not require convergence at all the points in the domain of the functions.

In these cases a sequence of functions is considered to converge to some specific function if it converges at all points in the domain except for a set of points with measure equal to 0. Sometimes we refer to this type of convergence as ***convergence almost everywhere.***

Uniform convergence:

Another type of convergence is **uniform convergence**
A sequence of functions f_n converges uniformly to the function f if the following holds: Pick $\epsilon > 0$, $\exists$ an integer N such that $\forall \;\; n > N, \; |f(x) - f_n(x)| < \epsilon, \quad \forall\, x \in \mathcal{R}$
The idea here is that the convergence rate, i.e. the integer N, can be selected so that it holds uniformly for all x in the domain of the function.
Note that **uniform convergence** implies **pointwise convergence** but not conversely.

Convergence in norm: (e.g. $\mathcal{L}_2$ norm)

We should think of functions as vectors with infinitely many components, generally non-countable. Then the norm of a function f is synonymous to the length of a vector, and we write $||f||$. We can then define the distance between two functions, f, and g, as the norm of the difference $||f - g||$. The convergence of a sequence of functions according to the norm is then defined as follows:

A sequence of functions f_n converges to the function f is the following holds

For every $\epsilon > 0, i.e.$ pick $\epsilon > 0$, $\exists$ an integer N, such that $\forall \;\; n > N, \;\; \big||f - f_n|\big| < \epsilon$.
Note for example, for the $\mathcal{L}_2$ norm, $\big||f|\big| = \left(\int_{-\infty}^{\infty} |f(x)|^2 dx\right)^{1/2}$.
For this norm $\mathcal{L}_2$ we also refer to the convergence of the sequence of functions as mean-square convergence.

Sequences of Random Variables

Almost Sure Convergence

Consider a sequence of random variables X_n, $n = 1,2, \cdots$, and the random variable X.
Assume that there exists a set N that is an event in Ω with $P(N) = 0$, such that for each $\omega \in \Omega \backslash N$ the sequence $X_n(\omega)$ converges to $X(\omega)$. This means that as a sequence of functions X_n converges to X on all points of Ω, except for the points on a set of *probability* zero (i.e. set of measure 0). This is the same as **pointwise converge** of real functions except for a set of measure zero. In probability theory we refer to this type of convergence as *almost sure convergence (a.s.).*

Alternative names for this type of convergence are convergence *with probability 1 (w.p. 1)*, or *almost everywhere (a.e.)* convergence.

We can indicate this type of convergence as

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1$$

In more precise terms, the above means

$$P(\{\omega: \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1$$

In other words we consider the set, $C = \{\omega: \lim_{n\to\infty} X_n(\omega) = X(\omega)\}$, and if $P(C) = 1$, then we say that X_n converges to X almost surely. Now C is not necessarily the whole sample space Ω

In probability this is also referred to as **strong convergence** in contrast to weak convergence to be discussed below.

Example

Consider a probability space with $\Omega = [0,1]$ with the uniform probability measure. E.g. for an interval I, $P(I)$ is the length of the interval.

Let $X_n(\omega) = \omega^n$.

Then the sequence of random variables X_n converges to 0 for $\omega \in [0,1)$. For $\omega = 1$, it converges to 1. Hence the sequence of random variables X_n converges to the random variable $X = 0$ on the set $\Omega\backslash\{1\}$. Since $\{1\}$ is a set with probability 0 then the sequence X_n converges to X almost surely.

Another Example

Consider $\Omega = [0,1]$ with uniform measure and fix an integer N

$$X_n(\omega) = \sin^n(2\pi N\omega)$$

Let $2\pi N\omega = \frac{k\pi}{2}$, k odd, i.e. $\omega = \frac{k}{4N}$.

Hence $X_n(\omega) = 1$ for $\omega = \frac{k}{4N}$ for $k = 1,3,5,\cdots 4N-1$.
We can see that $X_n \to 0$ for $\omega \in \Omega\backslash\mathrm{E}$
where $E = \{\omega: \omega = \frac{k}{4N}, k = 1,3,5,\cdots 4N-1\}$.
Since $P(E) = 0$. We can say that $X_n \to 0$ a.s. Note that N can be made arbitrarily large.

The above examples are relatively simple and not very useful in practice. But they are simple and easy to construct on a simple sample space.

To consider more realistic and useful examples we need to discuss unions and intersections with an infinite sequence of events A_n, $n = 1,2,\cdots$. As we test for convergence we will consider at each integer N the set of points for which the random variables for, $n > N$, are within a certain value ϵ of the limit. This set of points is an event. As a result we are going to consider infinite sequences of events and their "limits".

The easiest sequences to consider are what we call **"increasing sequences" and "decreasing sequences"** of events.

Increasing sequences:
A sequence of sets $\{A_n\}$ is increasing if $A_1 \subset A_2 \subset \cdots$. where $A_i \subset A_j$ means that A_i is a subset of A_j. Note that this includes the case $A_i = A_j$, i.e. it does not need to be a proper subset.

Decreasing sequences:
In a similar manner a sequence of sets $\{A_n\}$ is decreasing if $A_1 \supset A_2 \supset \cdots$.

Now, can we define a limit to the above sequences? For the increasing sequence the limit A is the set consisting of all points in any of the sets, i.e. $U = \{\omega: \omega \in A_k \text{ for some integer } k\}$. Another way to write this is as follows $U = \cup_{i=1}^{\infty} A_i$.

Similarly for the decreasing sequence we define the limit set (event) as the set $I = \{\omega: \omega \in A_k$ for all $k\}$. We can write $I = \cap_{i=1}^{\infty} A_i$ is the set of elements ω, such that $\omega \in A_i$ for all i.

Note that $U \supset A_k$ for any k, whereas $I \subset A_k$ for any k.

For example, consider and arbitrary sequence of events A_n. Then define a sequence of events $B_m = \cup_{n=m}^{\infty} A_n$. In other words, B_1 is the union of all the A_n's. B_2 is the union of all the A_n starting with A_2, B_m is the union of all the A_n's starting with A_m. Note that when we consider a union of sets, if we add sets to the union, the union set may get larger. If we take away sets from the union, the union set may get smaller. It is the opposite with intersections. Adding sets will result in a possibly smaller set. Removing sets will possibly result in a larger intersection set. In this case the above sequence B_m can possibly "decrease" as m increases because as m increases B_m is the union of a smaller collection of sets. Considering all m, we can write this as $B_1 \supset B_2 \supset \cdots \supset B_m \supset \cdots$. So this is a decreasing sequence of sets. Note that this means that, using the axioms of probability, $P(B_1) \geq P(B_2) \geq \cdots \geq P(m) \geq \cdots$.

Now consider intersections $B_m = \cap_{n=m}^{\infty} A_n$. In this case $B_1 \subset B_2 \subset \cdots \subset B_m \subset \cdots$. The B_m's get larger because we are intersecting less sets as m increases. So this is an increasing sequence of sets.

Now, consider an increasing sequence of sets, e.g. A_n. In a sense these sets get larger and larger. How do we refer to the limit? One way is to define the limit as $\lim_{n\to\infty} A_n = \cup_{n=1}^{\infty} A_n$. For a decreasing sequence A_n we can write $\lim_{n\to\infty} A_n = \cap_{n=1}^{\infty} A_n$. Note that for an increasing sequence the limit can be "as large" as Ω, although it need not be Ω, whereas for a decreasing sequence the limit can be "as small" as the empty set, ϕ.

What happens if we have an arbitrary sequence of sets A_n that is neither increasing, nor decreasing? Then we can create a decreasing sequence as follows: $S_n = \cup_{k=n}^{\infty} A_k$. Note that S_n is a decreasing sequence, because as n increases there are less sets in the union (S_n is the union of all sets in the sequence starting at the n^{th}). On the other hand $I_n = \cap_{k=n}^{\infty} A_k$ is an increasing sequence because as n increases there are less sets in the intersection.

Now consider the above sequence S_n. This is a decreasing sequence and we can talk about its limit, i.e. $\cap_{n=1}^{\infty} S_n$. Actually we should note that $\cap_{n=1}^{\infty} S_n = \cap_{n=N}^{\infty} S_n$ for any N, i.e. it does not matter where we start with the intersections, so we may as well start with $n = 1$.

So now we put the above two together and we can talk about the limit, $\cap_{n=1}^{\infty} S_n = \cap_{n=1}^{\infty} (\cup_{k=n}^{\infty} A_k) = \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k$.

We refer to this limit as the **lim sup**, i.e. $\limsup A_n = \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k$.

In a similar manner we can create a sequence of increasing sets $I_n = \cap_{k=n}^{\infty} A_k$. The limit of these increasing sequence is then $\cup_{n=1}^{\infty} I_n$. Note that it does not matter where we start the union because the I_n increase as n increases.

We refer to the limit as **lim inf**, $\liminf A_n = \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k$.

Conclusion: We can talk about the limit of a sequence of sets A_n only if it is an increasing or decreasing sequence. However we can talk about the lim sup, or the lim inf of an arbitrary sequence of sets A_n. In some cases we may have lim sup A_n = lim inf A_n hence we may say that $\lim A_n$ exists.

Example 1: Consider the set $\Omega = [0,1]$. Let $A_n = \left[\frac{1}{n}, 1 - \frac{1}{n}\right]$. Note that A_n is an increasing sequence of sets (events) because for $m > n$ $A_m \supset A_n$, i.e. $\left[\frac{1}{m}, 1 - \frac{1}{m}\right] \supset \left[\frac{1}{n}, 1 - \frac{1}{n}\right]$.

We can determine that $\cup_{n=1}^{\infty} A_n = (0,1)$. This is true because for any $x \in (0,1)$ we can find an n such that $x \in A_n$, and for any $x \in \cup_{n=1}^{\infty} A_n$ we can find an n such that $x \in A_n$, hence $x \in \left[\frac{1}{n}, 1-\frac{1}{n}\right]$ for some n, hence $x \in (0,1)$.

Example 2: Modify the above so that $A_n = \left[\frac{1}{n}, 1-\frac{1}{n}\right] \cup \{0,1\}$. Now we can see that the increasing sequence has a limit that is equal to Ω, $\cup_{n=1}^{\infty} A_n = [0,1] = \Omega$.

Example 3: Consider the decreasing sequence $A_n = \left[\frac{1}{2}-\frac{1}{n}, \frac{1}{2}+\frac{1}{n}\right]$. We can see that the limit is $\cap_{n=1}^{\infty} A_n = \left\{\frac{1}{2}\right\}$. There is no other point in $[0,1]$ other than $\frac{1}{2}$ that is in the intersection.

Example 4: Consider the decreasing sequence $A_n = (1-\frac{1}{n}, 1)$. We can show that $\cap_{n=1}^{\infty} = \phi$, the empty set.

Example 5: Consider the sequence $A_n = \left[\frac{1}{4}-\frac{(-1)^n}{8}, \frac{3}{4}+\frac{(-1)^n}{8}\right]$.
$\limsup A_n = \left[\frac{1}{8}, \frac{7}{8}\right]$ $\liminf A_n = \left[\frac{3}{8}, \frac{5}{8}\right]$.

Example 6: Now consider the sequence $A_n = \left(\frac{1}{4}-\frac{(-1)^n}{2^{n+2}}, \frac{3}{4}+\frac{(-1)^n}{2^{n+1}}\right)$. We have $A_1 = \left(\frac{3}{8}, \frac{1}{2}\right), A_2 = \left(\frac{1}{8}, 1\right), A_3 = \left(\frac{1}{4}+\frac{1}{2^5}, \frac{3}{4}-\frac{1}{2^4}\right)$, ...

We can see that this sequence is neither increasing nor decreasing. $\limsup A_n = \left[\frac{1}{4}, \frac{3}{4}\right]$, and $\liminf A_n = \left[\frac{1}{4}, \frac{3}{4}\right]$

Infinitely Often Occurrence

Now consider the following set $S = \cap_{m=1}^{\infty} (\cup_{n=m}^{\infty} A_n)$. How do we characterize the points in S? A point ω is in S if it is in infinitely many of the A_n's. If ω is in infinitely many of the A_n's then the set $\cup_{n=m}^{\infty} A_n$ must contain ω for any m, i.e. $B_m = \cup_{n=m}^{\infty} A_n$ must contain ω for each m. If this is not true then ω would occur in only a finite number of the A_n's. In set theory we refer to the set S as $S = \limsup A_n = \cap_{m=1}^{\infty} \cup_{n=m}^{\infty} A_n$. We could also write this as $\{A_n \text{ i.o.}\} = \limsup A_n$, where i.o. denotes infinitely often. In other words, lets say we have an infinite sequence of events (sets), A_n. Then $\{A_n$ i.o.$\}$ is equal to the set of points ω that are in infinitely many of the A_n's. Another way to say this more precisely is as follows: For each $\omega \in \limsup A_n$, pick any arbitrarily large integer m, then there exists an integer $m' > m$, such that $\omega \in A_{m'}$.

Almost Always Occurrence

Along the same line as above we define the set $I = \cup_{m=1}^{\infty} \cap_{n=m}^{\infty} A_n$. Now I is a union of increasing sets (as m increases, the sets $\cap_{n=m}^{\infty} A_n$, get larger). How do we characterize the points in I? A point ω is in I if it is in all the A_n's except for a finite number of the A_n's. We say that ω is in almost all (a.a.) of the A_n's. In set theory we refer to the set I as $I = \liminf A_n = \cup_{m=1}^{\infty} \cap_{n=m}^{\infty} A_n$. Another way to write this is $\{A_n \text{ a.a.}\} = \liminf A_n = \cup_{m=1}^{\infty} \cap_{n=m}^{\infty} A_n$.

By using DeMorgan's laws we can show the following:

$$\{A_n \text{ i.o.}\}^c = (\cap_{m=1}^{\infty} (\cup_{n=m}^{\infty} A_n))^c = \cup_{m=1}^{\infty} (\cup_{n=m}^{\infty} A_n)^c = \cup_{m=1}^{\infty} \cap_{n=m}^{\infty} A_n^c = \{A_n^c \text{ a.a.}\}$$

Or

$$\{A_n \text{ i.o.}\} = \{A_n^c \text{ a.a.}\}^c$$

Condition for Almost Sure Convergence of a Sequence of Random Variables

Lemma (5.2.1 in Rosenthal)

Let $Z_1, Z_2, \cdots$ be a sequence of random variables. Let Z be a random variable. Pick an arbitrary $\epsilon > 0$, and define the events $A_n(\epsilon) = \{\omega: |Z_n - Z| > \epsilon\}$. Suppose that for each $\epsilon > 0$, we have $P(A_n(\epsilon) \text{ i.o.}) = 0$. Then $P(Z_n \to Z) = 1$, i.e. Z_n converges to Z almost surely.
Note that $\{Z_n \to Z\}$ means $\{\omega: Z_n(\omega) \to Z(\omega)\}$.

Note that using DeMorgan's laws as above we could define $B_n(\epsilon) = \{\omega: |Z_n - Z| \leq \epsilon\}$, i.e. $B_n(\omega) = A_n(\omega)^c$, or $A_n(\epsilon) = B_n(\epsilon)^c$. Then we can write, using the above, $P(A_n(\epsilon) \text{ i.o.}) = P((A_n(\epsilon)^c \text{ a.a.})^c) = 1 - P(B_n(\epsilon) \text{ a.a.})$. Hence $P(A_n(\epsilon) \text{ i.o.}) = 0$ if and only if $P(B_n(\epsilon) \text{ a.a.}) = 1$. So we can say, suppose that for $\epsilon > 0$ we have $P(B_n(\epsilon) \text{ a. a.}) = 1$. Then $P(Z_n \to Z) = 1$, i.e. Z_n converges almost surely.

Example of Almost Sure Convergence (More Realistic than the Previous Example)

Consider an experiment where $\Omega = \{H, T\} \times \{H, T\} \times \cdots$. Ω is the product set of an infinite number of sets $\{H, T\}$. For example, the experiment may consist of flipping a non-fair coin many times (infinitely many times with $P(\text{heads}) = \frac{1}{n}$ at the n^{th} flip, with the outcomes of the different flips being independent.
Each $\omega \in \Omega$ is an infinite sequence $b_1, b_2, \cdots$, where the $b_i \in \{H, T\}$. Let Z_n be a random variable, for each n, with $Z_n(\omega) = 1$ if $b_n = H$, and 0 otherwise. The probability law for Z_n is $P(Z_n = 1) = 1/n$, i.e. $P(Z_n = 0) = 1 - 1/n$. The Z_n are independent random variables.

Let $Z = 0$ be a random variable.

Now, let us check to see if Z_n converges to Z almost surely. We can see that $P(Z_n = 0) = 1 - 1/n$, which approaches 1. But is the convergence of Z_n to $Z = 0$ almost sure convergence?

We define $A_n(\epsilon) = \{|Z_n| < \epsilon\}$.

Now for almost sure convergence we need $P(A_n(\epsilon)\text{ a.a.}) = 1$.
But

$$P(A_n(\epsilon)\text{ a.a.}) = P\left(\cup_{m=1}^{\infty} \cap_{n=m}^{\infty} A_n(\epsilon)\right)$$

Now the sets $B_m = \cap_{n=m}^{\infty} A_n(\epsilon)$ form a monotone increasing sequence of sets. According to the property of the probability measure (countable additivity) $P(\cup_{m=1}^{\infty} B_m) = \lim_{m\to\infty} P(B_m)$.
This is also referred to as continuity. Let E_n be a sequence of events such that $E_1 \subset E_2 \subset \cdots$.
Then for $E = \cup_{i=1}^{\infty} E_i$, $P(E) = \lim_{i\to\infty} P(E_i)$.

Now

$$P(B_m) = P\left(\cap_{n=m}^{\infty} A_n(\epsilon)\right) = P\left(\left(\cap_{n=m}^{N} A_n(\epsilon)\right) \cap_{n=N+1}^{\infty} A_n(\epsilon)\right))$$

$$\lim_{N\to\infty} \Pi_{n=m}^{N} P\left(A_n(\epsilon)\right) = \lim_{N\to\infty} \Pi_{n=m}^{N} P(|Z_n| < \epsilon) = \lim_{N\to\infty} \Pi_{n=m}^{N} \left(1 - \frac{1}{n}\right) = \lim_{N\to\infty} \left(\frac{m-1}{m} \cdot \frac{m}{m+1} \cdots \frac{N-1}{N}\right) = 0,$$

i.e. for any m $P(B_m) = 0$. Hence $\lim_{m\to\infty} P(B_m) = 0$. This proves that $P(A_n(\epsilon)\text{ a. a.}) = 0$ and the sequence Z_n does not converge a.s. to $Z = 0$.

Borel-Cantelli Lemma

This is a very important theorem.

Let A_n be an infinite sequence of events.

(a) If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(A_n \text{ i.o.}) = 0$.

(b) If $\sum_{i=1}^{\infty} P(A_n) = \infty$ and $A_1, A_2, \cdots$ is an independent set of events then $P(A_n \text{ i.o.}) = 1$.

This means that if $A_1, A_2, \cdots$ are independent, then $P(\cap_{m=1}^{\infty} \cup_{n=m}^{\infty} A_n)$ is always equal to 0 or 1.

Corollary

Let $Z_1, Z_2, \cdots$, and Z be random variables. Suppose that for each $\epsilon > 0$, we have $\sum_n P(|Z_n - Z| > \epsilon) < \infty$. Then $P(Z_n \to Z) = 1$, i.e. Z_n converges to Z almost surely.

Example – Borel-Cantelli Lemma

We consider the above scenario of an infinite sequence of coin tosses (not fair coins) where $\Omega = [H,T] \times [H,T] \times \cdots$, and the Z_n are defined in the same manner, i.e. independent random variables with $P(Z_n = 1) = \frac{1}{n}$, $P(Z_n = 0) = 1 - \frac{1}{n}$. Using the Borel-Cantelli Lemma with $A_n = \{|Z_n| > \epsilon\}$, $P(|Z_n| > \epsilon) = P(Z_n = 1) = \frac{1}{n}$. Hence, $\sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$. Then $P(A_n \text{ i.o.}) = 1$, i.e. on a set with probability equal to 1, the sequence $Z_n(\omega)$ does not converge almost surely. This is the case because we cannot find an integer $N(\epsilon)$ such that for all $n > N(\epsilon)$, $|Z_n| < \epsilon$.

Convergence in Probability

Consider a sequence of random variables Z_n, and a random variable Z. Then the sequence Z_n converges in probability to Z if for any $\epsilon > 0$, we have $P(|Z_n - Z| \geq \epsilon) \to 0$ as $n \to \infty$.
We may denote convergence in probability as $Z_n \xrightarrow{p} Z$.

Note that the above means $\forall\ \epsilon > 0, \forall\ \delta\ \exists$ an integer N, such that for $n > N$, $P(|Z_n - Z| \geq \epsilon) \leq \delta$.

Example

Consider the above example where Z_n are a sequence of independent random variables with $P(Z_n = 1) = \frac{1}{n}$, $P(Z_n = 0) = 1 - \frac{1}{n}$, and $Z = 0$. Then $P(|Z_n - Z| \geq \epsilon) = P(Z_n = 1) = \frac{1}{n}$. But $\frac{1}{n} \to 0$ as $n \to \infty$. Hence the sequence Z_n converges in Probability to 0. Recall from the above that this sequence does not converge to $Z = 0$ almost surely.

Almost Sure Convergence => Convergence in Probability

Let $Z_1, Z_2, \cdots$ be a sequence of random variables, and Z a random variable. If $Z_n \to Z$ almost surely then $Z_n \to Z$ in probability.

Proof: One can show that a necessary and sufficient condition for $Z_n \to Z$ a.s. is that $P(\sup_{k \geq n} |Z_k - Z| \geq \epsilon) \to 0$ as $n \to \infty$.
Now, we note that $\left\{ \sup_{k \geq n} |Z_k - Z| \geq \epsilon \right\} \supset \{|Z_n - Z| \geq \epsilon\}$. Hence $P\left(\sup_{k \geq n} |Z_k - Z| \geq \epsilon \right) \geq P(|Z_n - Z| \geq \epsilon)$ for all n. So, if $\lim_{n \to \infty} P(\sup_{k \geq n} |Z_k - Z| \geq \epsilon) = 0$, then $\lim_{n \to \infty} P(|Z_n - Z| \geq \epsilon) = 0$.

Convergence in Distribution

A sequence of random variables, $Z_1, Z_2, \cdots$ is said to *converge in distribution* to the random variable Z if $\lim_{n\to\infty} F_{Z_n}(z) = F_Z(z)$ for all z in $\mathcal{R}$ at which $F_Z(z)$ is continuous.

Convergence in distribution is sometimes denoted as $Z_n \xrightarrow{d} Z$.

Example – Convergence in Distribution

Let $S_n = \sum_{i=1}^{n} X_i$, where the X_i are independent Bernoulli random variables with parameter p, i.e. $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$.

Let $Z_n = \frac{S_n - np}{\sqrt{np(1-p)}}$

Then according to the Laplace-De Moivre theorem we can show that Z_n converges to the standard Gaussian distribution with mean equal to zero and standard deviation equal to 1.

Example: Use the convergence property to approximate a Binomial distribution by a Gaussian.

Toss a die 12,000 times. What is the probability that the number of 6's lies in the interval $(1800, 2100)$?

$$P(k\ 6\text{'s}) = \binom{12000}{k}\left(\frac{1}{6}\right)^k\left(1-\frac{1}{6}\right)^{12000-k}$$
$$\approx \Phi\left(\frac{2100-2000}{\sqrt{12000(1/6)(5/6)}}\right) - \Phi\left(\frac{1800-2000}{\sqrt{12000(1/6)(5/6)}}\right) \approx \Phi\left(\sqrt{6}\right) - \Phi\left(-2\sqrt{6}\right)$$
$$\approx 0.992$$

Where $\Phi(\cdot)$ is the CDF for the Normal Distributed random variable.

Convergence in Distribution is Weaker than Convergence in Probability.

Example

Let Y_n be a sequence of independent random variables with a distribution equal to the Normal distribution. Let $Y = Y_1$. Then the distributions for all the Y_n are equal. Hence the Y_n converge to Y in distribution.

Let $Z_n = Y_n - Y$, $n \geq 2$. Z_n is Gaussian distributed with $\text{var}(Z_n) = \text{var}(Y_n) + \text{var}(Y) = 2$.
Now pick $\epsilon > 0$, $P(|Z_n| \leq \epsilon) = \alpha$, for some $\alpha < 1$, and α is independent of n.

Hence $\lim_{n\to\infty} P(|Z_n| \leq \epsilon) = \alpha < 1)$. Hence we cannot have the condition that that Y_n converges to Y in probability because it would require that this limit equals 1. In fact this counter-example argument works for any distribution with non-zero variance.

Note that convergence in distribution is a very different type of convergence of a sequence of random variables. For all the other types of convergence we are talking about the convergence of a sequence of functions on a space Ω. But as with ordinary real valued functions there are different types of convergence. In the case of distributions of random variables, and convergence in distribution, it is not a convergence of a sequence of random variables, because a random variable is a function and we are not talking about functions. It is a case of convergence of a property of a random variable, the CDF. So, in a sense convergence of random variables in distribution is really not convergence of random variables. In a sense "convergence in distribution" is a misnomer, but it is widely used.

Mean Square Convergence

If we think of a sequence of random variables Z_n as really a sequence of functions on the space Ω, then there are many ways to define convergence of these functions to the function Z (i.e. random variable Z). In order to consider limits of sequences of functions we need to discuss whether or not a function is "close" to another function, and to do this we need the concept of a metric. A space of functions on which there is a concept of metric, or a distance defined between any two functions, is known as a ***metric space***. There is also the concept of the norm of a function which in a sense defines a "size" for a function, i.e. it is like length of a vector. For the function f we use the notation $||f||$ for the norm. Such a norm needs to obey a set of axioms. A space of functions for which a norm is defined is known as a ***normed space***.

For example if we have a space of functions for which the integral $\left(\int_R |f(x)|^p dx\right)^{\frac{1}{p}}$ is finite for any function $f(\cdot)$ then we refer to the space of functions as an $\mathcal{L}_p$ normed space. We may denote the norm of f as $\left|||f|\right|_p$. The case $p = 2$, is the most common case and is used in many applications. The resulting space of functions $(p = 2)$ is known as a Hilbert space. In the case of random variables the function space is the space of real factions defined on the probability space Ω, with the appropriate conditions of measurability. The corresponding norms would be $\left(\int_\Omega |X|^p dP\right)^{\frac{1}{p}}$. For $p = 2$, the norm is the square root of the second moment for the random variable.

Now, for any normed space we can define a metric with a distance between any two functions f, g being as follows: $d(f,g) = \left||f - g|\right|$. We can then speak about convergence of a sequence of functions according to this metric. In the case of random variables and in the case that we use the above metric with $p = 2$, we then speak about *mean-square convergence* of the sequence of random varables.

Consider the sequence of random variables Z_n and the random variable Z. The distance squared (i.e. the square of the metric) between Z_n and Z according to the $\mathcal{L}_2$ metric is then $\int_\Omega |Z_n - Z|^2 dP$. Note that this is really the mean of the square of the random variable $Z_n - Z$. If we refer to this random variable as an "error", then the metric is the "mean-square error". A sequence of random variables Z_n then converges to the random variable Z in the mean square sense, if it converges according to the metric defined using the $\mathcal{L}_2$ norm. In other words

$$\lim_{n\to\infty} \int_\Omega |Z_n - Z|^2 dP = 0$$

Note that this integral can also be performed over the real line as

$$\int_{-\infty}^{\infty} x^2 f_n(x) dx$$

where $f_n(x)$ is the PDF for the random variable $E_n = Z_n - Z$. Note that in the above discussion we have referred to many norms and in a sense each such norm leads to a different type of convergence. But the one that is the most important is the one as in $\mathcal{L}_2$, known in the case of a probability space as *mean-square convergence*. We can also denote this type of convergence as $Z_n \xrightarrow{m.s.} Z$.

Example 1: Mean Square Convergence

We consider a sequence of independent random variables, Z_n, defined on the probability $\Omega = \{(\omega_1, \omega_2, \cdots): \omega_i \in \{0,1\}\}$, with $Z_n(\omega) = \omega_n$ and $P(Z_n = 1) = \frac{1}{n}$, $P(Z_n = 0) = 1 - \frac{1}{n}$.
Let $Z = 0$ be a random variable defined on Ω.
Consider $\mathcal{E}((Z_n - Z)^2) = \mathcal{E}(Z_n^2) - \mathcal{E}(2Z_n Z) + \mathcal{E}(Z^2) = \frac{1}{n}$.
Clearly $\lim_{n\to\infty} \mathcal{E}((Z_n - Z)^2) = \lim_{n\to\infty} \frac{1}{n} = 0$. Hence $Z_n \xrightarrow{ms} Z$.
Note that we saw in the above that Z_n does not converge to $Z = 0$ a.s.

Example 2

Consider the case in Example 1 but now with $Z_n(\omega) = n\omega_n$. $P(Z_n = n) = \frac{1}{n}$, $P(Z_n = 0) = 1 - \frac{1}{n}$, and $Z = 0$.
$\mathcal{E}((Z_n - Z)^2) = \mathcal{E}(Z_n^2) = n^2\left(\frac{1}{n}\right) + 0\left(1 - \frac{1}{n}\right) = n.$

Note that $\lim_{n\to\infty} \mathcal{E}((Z_n - Z)^2) \neq 0$, hence Z_n does not converge to Z as $n \to \infty$ in the mean square sense.

Note that in this case Z_n converges to Z in probability, because for every $\epsilon > 0$, $\lim_{n\to\infty} P(|Z_n - Z| < \epsilon) = \lim_{n\to\infty} 1 - \frac{1}{n} = 1$.

Note that in this example all that we have proved is that Z_n does not converge to $Z = 0$ in the mean square sense. We have not proved that Z_n does not converge to Z for any random variable Z. We only tested the case $Z = 0$.

MS Convergence Implies Convergence in Probability

Lemma: If a sequence of random variables Z_n converges to Z in the mean square sense then it also converges to Z in probability.

Proof:

Convergence in Probability means that for $\epsilon > 0$, $\lim_{n\to\infty} P(|Z_n - Z| \leq \epsilon) = 1$. This is the same as $\lim_{n\to\infty} P(|Z_n - Z| > \epsilon) = 0$.

Now, $P(|Z_n - Z| > \epsilon) = P(|Z_n - Z|^2 > \epsilon^2) \leq \frac{\mathcal{E}((Z_n - Z)^2)}{\epsilon^2}$.

Where the last inequality follows from the Markov inequality.

If the sequence Z converges to Z in the mean square sense then the right hand side approaches 0 as $n \to \infty$. Hence the sequence Z_n also converges Z in probability.

Almost Sure Convergence dos not Necessarily Imply Mean Square Convergence

Example

Consider a sequence of random variables as in the above but instead of $Z_n = n\omega_n$, we define $Z_n = n^2\omega_n$. We also let $P(Z_n = n^2) = \frac{1}{n^2}$, $P(Z_n = 0) = 1 - \frac{1}{n^2}$. Now this sequence converges to $Z = 0$ almost surely.

Proof:
Let $A_n(\epsilon) = (Z_n > \epsilon)$. Then $P\big(A_n(\epsilon)\big) = P(Z_n = n^2) = \frac{1}{n^2}$. Now, consider $\sum_{n=1}^{\infty} P\big(A_n(\epsilon)\big)$. Since the series $\sum_n 1/n^2$ converges, we have $\sum_{n=1}^{\infty} P\big(A_n(\epsilon)\big) < \infty$. Hence by the Borel-Cantelli Lemma, $P(A_n(\epsilon)\text{ i.o.}) = 0$. Hence, except for a set of probability 0 we cannot have the event $(Z_n > \epsilon)$ infinitely often. Hence $Z_n \xrightarrow{a.s.} 0$.

What about mean square convergence, i.e. $Z_n \xrightarrow{m.s.} Z$?

Consider the requirement for mean square convergence

$$\mathcal{E}((Z_n - Z)^2) = \mathcal{E}(Z_n^2) = (n^2)^2 \cdot \frac{1}{n^2} = n^2$$

This limit is not equal to 0!
Hence Z_n does not converge to $Z = 0$ in the mean square sense.

Mean Square Convergence: Cauchy Criterion

In the above we considered questions of convergence given a sequence of random variables Z_n and a limiting random variable Z. For each case we determined whether or not there was convergence. In some cases we may want to determine whether or not a sequence Z_n converges to some random variable Z which we are not given. How do we determine if the sequence converges without knowing what is the limit? There is a criterion, called the **Cauchy criterion**, that we can use to determine whether or not such convergence occurs without considering the limiting random variable.

Cauchy Criteria:

The sequence of random variables Z_n converges in the mean square sense to some random variable Z, if $\lim_{m,m\to\infty} \mathcal{E}((Z_m - Z_n)^2) = 0$.

More precisely this means that given $\epsilon > 0$, there exists an integer N such that for all $m > N$, and $n > N$, $\mathcal{E}((X_n - X_m)^2) < \epsilon$.

Note that we do not refer to any limiting random variable Z, rather the claim is that the random variables, for very large n, m are all close to each other. If we consider them as points in a metric space then they all cluster around some point. Let us define a sequence of random

To gain more insight into the different types of convergence we consider a sample space $\Omega = [0,1]$ with uniform probability measure. Consider a sequence of intervals I_n as follows: $I_1 = [0,1]$, $I_2 = \left[0, \frac{1}{2}\right]$, $I_2 = \left[\frac{1}{2}, \frac{1}{2} + \frac{1}{3} \bmod 1\right]$, ... Now the n^{th} interval has width $\frac{1}{n}$ and is equal to $\left[a_n, a_n + \frac{1}{n}\right]$ but if $a_n + \frac{1}{n} > 1$, then we replace it with $[a_n, 1] \cup \left[0, a_n + \frac{1}{n} - 1\right]$. This becomes easy to see if we map the interval $[0,1]$ into a circle with circumference equal to 1. Then we lay out a sequence of intervals where the n^{th} interval has length $\frac{1}{n}$ and it starts where the previous interval ended. Now consider a sequence of random variables Z_n where $Z_n(\omega) = 1$ if $\omega \in I_n$ and zero otherwise. In other words Z_n is the indicator function for the intervals I_n. The random variable X_n can be viewed as a rectangle of width $\frac{1}{n}$ and height 1. As we think of the sequence of random variables we visualize a sequence of rectangles that are being shifted to the right (cyclic shifts).

Summary of Convergence Relations

We have seen the following relations between the different types of convergence: almost sure, in probability, in distribution, and mean square.

Assume that we have a sequence of random variables X_n and a limiting random variable X. We have shown the following

We have shown that almost sure convergence implies convergence in probability

$$\left(X_n \xrightarrow{a.s.} X\right) \Rightarrow \left(X_n \xrightarrow{p} X\right)$$

Also, convergence in mean square implies convergence in probability

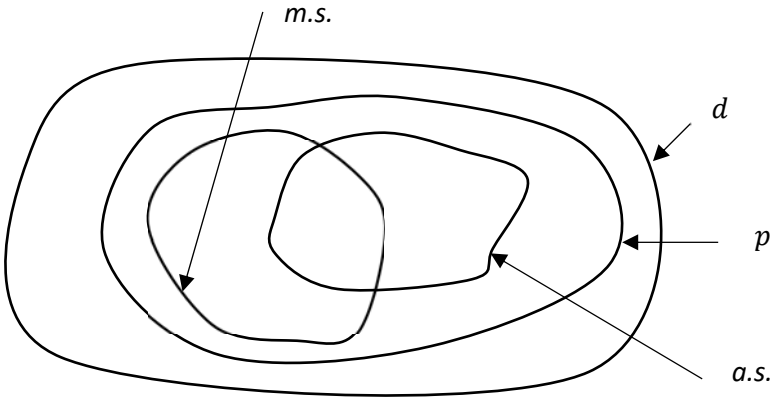
$$\left(X_n \xrightarrow{m.s.} X\right) \Rightarrow (X_n \xrightarrow{p} X)$$

And we have shown that convergence in probability implies convergence in distribution

$$\left(X_n \xrightarrow{p} X\right) \Rightarrow (X_n \xrightarrow{d} X)$$

But we also showed that there are cases of almost sure convergence which do not constitute mean square convergence.

All of the relations can be represented in the following Venn diagram



Law of Large Numbers (LLN)

The law of large numbers states that if we consider a sequence of independent random variables $X_1, X_2, \cdots$ and define the sequence of random variables $\{Z_n\}$, where $Z_n = \frac{1}{n}\sum_{k=1}^{n} X_k$. Then the sequence of random variables Z_n converges to a constant random variable depending on some conditions placed on the sequence X_n. In general, these conditions may or may not involve the

CDF's or moments for the random variables X_n. Then there is the question of the convergence criteria, e.g. is it almost sure convergence? Convergence in probability, etc. There are many theorems which have been proved over the history of the development of probability theory. Each new theorem utilizes conditions that get more and more general, i.e. less restrictive on the probabilities laws for the random variables.

In the most common application, all the random variables X_n are assumed to have the same distribution, but we may consider different conditions on the existence of the different moments, e.g. mean and variance.

In these LLN theorems some use convergence in probability, while others are based on almost sure convergence. Laws that use convergence in probability are known as the Weak Law of Large Numbers, whereas laws that use almost sure convergence are known as the Strong Law of Large Numbers.

Weak Law of Large Numbers (WLLN)

We consider an easy case of the WLLN. In this case the X_n are assumed to have the same mean m and variance v, but not necessarily the same probability law, although in many cases in practice the X_n have the same probability law.

WLLN: Let X_n be a sequence of independent random variables with $\mathcal{E}(X_n) = m$, and $\text{var}(X_n) = v < \infty$.
Then for all $\epsilon > 0$ $\lim_{n\to\infty} P\left(\left|\frac{1}{n}(X_1 + X_2 + \cdots + X_n) - m\right| \geq \epsilon\right) = 0$,
That is $\frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ converges in probability to the constant random variable m.

Proof: Set $S_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$. Then $\mathcal{E}(S_n) = m$ and $\text{var}(S_n) = \frac{v}{n}$. By the Chebychev's inequality we have

$$P\left(\left|\frac{1}{n}(X_1 + X + 2 + \cdots + X_n) - m\right| \geq \epsilon\right) \leq \frac{v}{n\epsilon^2}$$

But for a fixed ϵ the right hand side approaches zero as $n \to \infty$.

Strong Law of Large Numbers (SLLN)
We state two theorems without proof.

1) Independent Non-Identically Distributed RV

Let $X_1, X_2, \cdots$ be a sequence of independent random variables, each having the same finite mean m, and the X_i having bounded central moments of 4th order, i.e. $\mathcal{E}((X_i - m)^4 \leq \alpha < \infty$.

Then $P\left(\lim_{n\to\infty}\frac{1}{n}(X_1+X_2+\cdots+X_n)=m\right)=1$,

i.e. the sequence $Z_n=\frac{1}{n}(X_1+X_2+\cdots+X_n)$ converges to m almost surely.

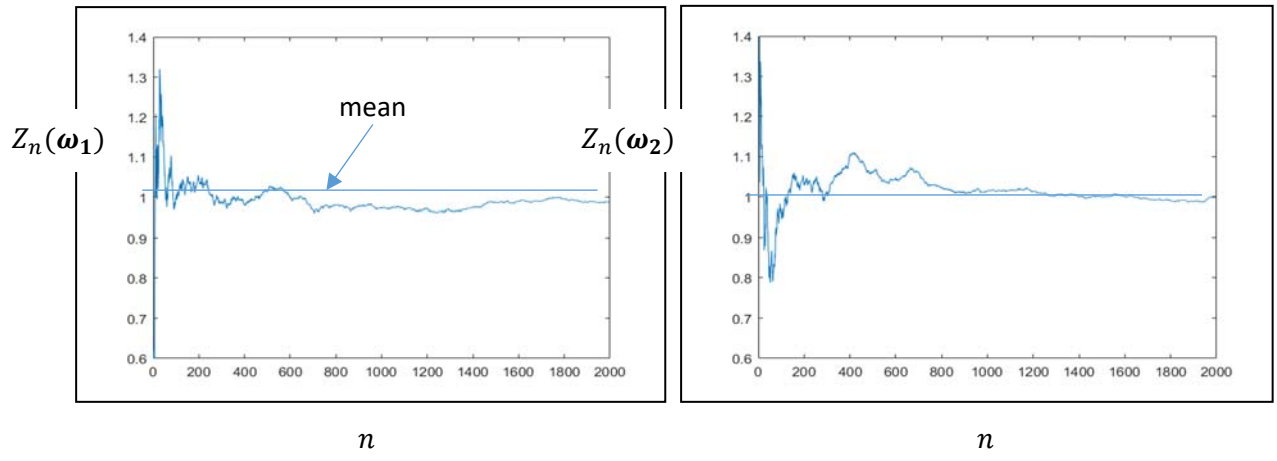
2) Identically Distributed RV
Let $X_1, X_2, \cdots$ be a sequence of independent identically distributed (i.i.d.) random variables, with finite mean m. Then

$$P\left(\lim_{n\to\infty}\frac{1}{n}(X_1+X_2+\cdots+X_n)=m\right)=1$$

In other words, the sequence of averages $\frac{1}{n}(X_1+X_2+\cdots+X_n)$ converges to m almost surely. Note that we do not assume that any of the moments other than the mean are finite.

Weak vs. Strong Law of Large Numbers

We could take several instances of a sequence of random variables, i.e. several $\omega \in \Omega$. For example consider i.i.d. Gaussian random variables with mean $= 1$, and variance $= 1$. The following are two sample outcomes, i.e. $Z_n(\boldsymbol{\omega_1})=\frac{1}{n}\sum_{i=1}^{n}X_i\,(\boldsymbol{\omega_1})$, and $Z_n(\boldsymbol{\omega_2})=\frac{1}{n}\sum_{i=1}^{n}X_i\,(\boldsymbol{\omega_2})$ generated by MATLAB for 2000 samples, i.e. $\boldsymbol{\omega_1}=(\omega_{1,1},\omega_{1,2},\cdots,\omega_{1,2000},\cdots)$, $\boldsymbol{\omega_2}=(\omega_{2,1},\omega_{2,2},\cdots,\omega_{2,2000},\cdots)$.



The Strong Law of Large Numbers guarantees that each sample outcome will result in the average converging to the mean as the number of samples taken approaches infinity.

Example

We have a biased coin with $P(\text{Heads})=p$. We would like to estimate the parameter p by an experiment where we toss the coin n times. We toss the coin n times and determine the frequency

of Heads, n_H, $f_H = \frac{n_H}{n}$. How large should n be in order to have a 0.95 probability that the relative frequency is within 0.01 of p?

Solution: If X_i denotes the i^{th} toss then $\mathcal{E}(X_i) = p$, and $\text{var}(X_i) = p(1-p) = \sigma^2$. Note that since $0 \le p \le 1$, we have $\sigma^2 \le \frac{1}{4}$. Note also that $\mathcal{E}(f_H) = p$, and $\text{var}(f_H) = \frac{\sigma^2}{n}$. Using the Chebychev inequality we have $P(|f_H - p| \ge \epsilon) \le \frac{\sigma^2}{n\epsilon^2} \le \frac{1}{4n\epsilon^2}$.
Now set $\frac{1}{4n\epsilon^2} = .05$, and using $\epsilon = .01$, we obtain $n = 50000$. Note that because we are using the Chebychev inequality this is an upper bound on n, the minimum value of n that would guarantee the above specified probability is smaller.

Central Limit Theorem

Frequently we deal with problems involving sums of large numbers of random variables

$$S_n = \sum_{i=1}^{n} X_i$$

Where the X_i are independent random variables that may or may not be identically distributed.
Question: We know that the PDF for S_n is the convolution of the PDF's for the X_i. If n is large can we find a good approximation to the probability distribution for S_n. What happens to this distribution as $n \to \infty$?

If such a distribution exists, what are the conditions for such convergence?

Assume that the X_i are identically distributed with mean m. According to the Weak Law of Large Numbers $\frac{S_n}{n} \xrightarrow{p} m$. This may also be written as $\frac{S_n - nm}{n} \xrightarrow{p} 0$. In other words the random variable $\frac{S_n - nm}{n}$ approaches the degenerate distribution, i.e. CDF is $F(x) = u(x)$ (step function), or PDF is $\delta(x)$.

Mean and Variance of a Sum of Independent Random Variables

Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables with mean m, and variance σ^2. Let

$$S_n = X_1 + X_2 + \cdots + X_n$$

Then the mean of S_n is nm, and the variance is $n\sigma^2$. Now, we define $Z_n = \frac{S_n - nm}{\sqrt{n}}$. Then we can check that $\text{var}(Z_n) = \mathcal{E}(Z_n^2) = \frac{1}{n}\text{var}(S_n) = \sigma^2$.

Theorem

The distribution of Z_n, defined above, converges to a Gaussian distribution with variance σ^2 as $n \to \infty$.

Proof:
Write $Z_n = \sum_{i=1}^{n} \frac{X_i - m}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$, where the Y_i are zero mean random variables

$$\Phi_n(\omega) = \Pi_{i=1}^{n} \mathcal{E}\left(\exp\left(\frac{j\omega Y_i}{\sqrt{n}}\right)\right) = \Phi_Y^n\left(\frac{\omega}{\sqrt{n}}\right)$$

Now, $\exp\left(\frac{j\omega Y}{\sqrt{n}}\right) = \sum_{k=0}^{\infty} \frac{\left(\frac{j\omega Y}{\sqrt{n}}\right)^k}{k!} = 1 + \frac{j\omega Y}{\sqrt{n}} + \frac{1}{2!}\left(\frac{j\omega Y}{\sqrt{n}}\right)^2 + \cdots + \frac{1}{k!}\left(\frac{j\omega Y}{\sqrt{n}}\right)^k + \cdots$

Taking the expected value we evaluate $\Phi_Y(\omega) = \mathcal{E}\left(\exp\left(\frac{j\omega Y}{\sqrt{n}}\right)\right) = \mathcal{E}\left(1 + \frac{j\omega Y}{\sqrt{n}} + \frac{1}{2!}\left(\frac{j\omega Y}{\sqrt{n}}\right)^2 + o\left(\frac{1}{n}\right)\right)$, where $o\left(\frac{1}{n}\right)$ is a quantity c_n such that $\lim_{n-\infty} \frac{c_n}{\frac{1}{n}} = 0$.

We can write this as $1 + \frac{1}{n}\left(-\frac{\omega^2\sigma^2}{2} + o\left(\frac{1}{n}\right)\right)$.

We can now write $\Phi_n(\omega) = \left(1 + \frac{z_n}{n}\right)^n$, where $z = -\frac{\omega^2\sigma^2}{2} + o\left(\frac{1}{n}\right)$

As $n \to \infty$, $z_n \to -\frac{\omega^2\sigma^2}{2}$, and $\left(1 + \frac{z_n}{n}\right)^n \to e^{-\frac{\sigma^2\omega^2}{2}}$.

This is the characteristic function for a Gaussian distributed random variable. Hence the random variable Z_n converges to a Gaussian random variable in distribution.

Motivation

In many scenarios we have many small signals adding up randomly, and the observed signal is the sum of this large number of small signals. The resulting signal is then modelled as a Gaussian random variable. In many cases we refer to this resultant signal as noise and call is Gaussian noise.

random currents in a resistor. The intensity of these currents is proportional to the temperature. The intensity is also dependent on the resistance of the resistor. We refer to the resulting noise signal as thermal noise.

Other cases include a interference in a wireless system where many interferers with small power create an interfering signal at some receiver local. The received interference signal is then modeled as a Gaussian random variable.

In general in any system that involve the processing of signals, whether is communication signals, radar system signals, biomedical signals, or geophysical exploration signals, the received signal, i.e. the signal is interest, is often modelled as being affected by what we call additive Gaussian noise.

There are also cases where some phenomenon is modelled by a discrete random variable, but when the variable takes on a large range of values, it is modelled as a continuous random variable with Gaussian distribution.

Bernoulli Distributed Random Variable

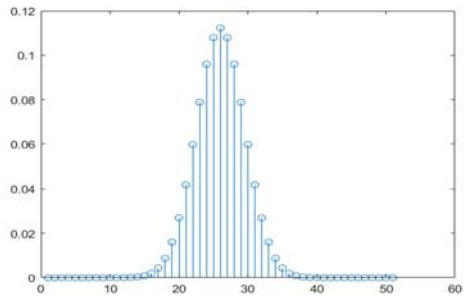
A discrete random variable X taking on only two possible values in the set $\{0,1\}$, with $P(X = 1) = p$, and $P(X = 0) = 1 - p$ is usually referred to as being Bernoulli distributed, or simply being a Bernoulli random variable. The mean of X is p, and the variance is $p - p^2 = p(1 - p)$.

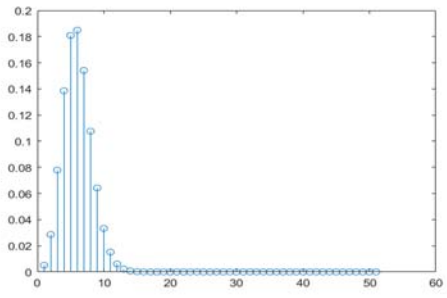
Binomial Distributed Random Variables.

Consider a sequence of n independent identically distributed Bernoulli random variables $X_1, X_2, \cdots, X_n$. Now, form the sum $Y = \sum_{i=1}^{n} X_i$. Then Y has a binomial distribution, which we denote here as $B(n, p)$. The probability mass function is

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

In the following we show the probability mass function for the case $n = 50$, and $p = 0.5$, on the left and $p = 0.1$ on the right.





The mean can easily be determined as $m_Y = \mathcal{E}(Y) = np$.
The variance is the sum of the variances of the Bernoulli r.v.'s and is obtained as $\sigma_Y^2 = np(1 - p)$.

Note let us form the zero mean and normalized random variable $Z_n = \frac{Y-nm}{\sqrt{np(1-p)}}$. Then we can show that $F_Z(z) = P(Z_n \leq z) \rightarrow \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$ as $n \rightarrow \infty$. This is known as a theorem from Laplace. In other words, the sequence of random variables Z_n converges in distribution to a Gaussian random variable with mean 0 and variance equal to 1, also known as the *Normal Distribution*. So we can conclude that for large n and a fixed p the binomial distribution, with the mean subtracted and appropriately scaled converges to a Normal Distribution.

Poisson Approximation to the Binomial

We consider a Binomial distribution with parameters n and p, and let $n \rightarrow \infty$, and $p \rightarrow 0$, in such as manner so that $np \rightarrow \lambda$, a constant. The classical example here is the modeling of an arrival time in a systems. We can start looking at the system at time $t = 0$, and observe the time that the first customer in a queuing system, or car in a road, or packet in a network, arrives. We break time into intervals of length Δt, where Δt is arbitrary small. The probability that an arrival occurs in a particular interval is specified as $p = \lambda \Delta t$. The probability that more than one arrival occurs in such an interval is approximated as 0. Hence in the time interval $[0, T]$ the number of small time intervals is approximately $n = T/\Delta t$. The probability that we have k arrivals is then

$$P(k \text{ arrivals in } [0,T]) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} (\lambda \Delta t)^k (1-\lambda \Delta t)^{n-k}$$
$$= \frac{n!}{(n-k)!\, k!} (\lambda \Delta t)^k (1-\lambda \Delta t)^{n-k}$$

$$= \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\lambda T}{n}\right)^k \left(1-\frac{\lambda T}{n}\right)^{n-k}$$
$$\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right) (\lambda T)^k \left(1-\frac{\lambda T}{n}\right)^{n-k}$$

For a fixed k, as $n \rightarrow \infty$, i.e. $\Delta t \rightarrow 0$, the first k factors all approach 1, and the last factor can be written as $\frac{\left(1-\frac{\lambda T}{n}\right)^n}{\left(1-\frac{\lambda T}{n}\right)^k}$, where we see clearly that the denominator approaches 1 since $\frac{\lambda T}{n} \rightarrow 0$, and the numerator is the well known limit that yields and exponential, in this case $e^{-\lambda T}$. As a result we have the following

$$P(k \text{ arrivals in the interval of length } T) = \frac{(\lambda T)^k e^{-\lambda T}}{k!}$$

.

From the origins in the Binomial distribution we can see that the expected value of this random variable is $np = n\lambda \Delta t = \lambda T$. However we can also evaluate the mean directly and obtain this result. If we set the time duration as $T = 1$, then the expected value of the number of arrivals is equal to λ. As a result the parameter λ is know as the arrival rate, i.e. expected number of arrivals per unit time. Of course we can also redefine the parameter λ so that λT is replaced by λ which

then becomes the expected number of arrivals in the interval of interest. The probability of the number of arrivals is then

$$P(k \text{ arrivals}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Three Limit Laws for the Binomial Distribution

Let Y_n be a random variable with binomial distribution and parameters (n, p). The have shown the following convergences in distribution

Fix $p > 0$ and let $n \to \infty$

$$\mathcal{L}\left(\frac{Y_n - \mathcal{E}(Y_n)}{n}\right) \to \mathcal{C}(0)$$

Where $\mathcal{L}(\cdot)$ stands for "law" and $\mathcal{C}(\cdot)$ denotes the constant law, i.e. degenerate distribution.

Fix $p > 0$, and let $n \to \infty$

$$\mathcal{L}\left(\frac{Y_n - \mathcal{E}(Y_n)}{\sqrt{\text{var}(Y_n)}}\right) \to \mathcal{N}(0,1)$$

Fix λ, let $p = \frac{\lambda}{n}$, and let $n \to \infty$

$$\mathcal{L}(Y_n) \to P(\lambda)$$

Where $\mathcal{C}(0)$ denotes the degenerate distribution, i.e. the CDF is $F(x) = u(x)$, $\mathcal{N}(0,1)$ is the normal distribution, and $\mathcal{P}(\lambda)$ is the Poisson distribution with parameter λ.

Characteristic Functions for Some Limiting Distributions

Degenerate Law (Point Distribution): CDF: $F(x) = u(x-a)$, PDF: $f(x) = \delta(x-a)$.

$$\Phi_d(\omega) = \exp(j\omega a)$$

Gaussian Law:

$$\Phi_n(\omega) = \exp\left(j\omega a - \frac{\sigma^2\omega^2}{2}\right)$$

Poisson Distribution:

$$\Phi_{\mathrm{X}}(\omega) = \mathcal{E}\left(e^{j\omega X}\right) = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} e^{j\omega k} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\left(e^{j\omega}\lambda\right)^k}{k!} = e^{-\lambda} \exp\left(\lambda e^{j\omega}\right)$$
$$= \exp\left(\lambda\left(e^{j\omega} - 1\right)\right)$$

The characteristic functions of all three have an exponential form with the property that each distribution can be represented as the distribution of the sum of two independent random variables of the same type as we will see shortly. This means that we can in fact say that each distribution can be represented as the distribution of a sum of n independent random variables of the same type of distribution.

Degenerate Law:

Let X_1 and X_2 be random variables with the degenerate law with parameters a_1 and a_2 respectively. Then the sum $X = X_1 + X_2$ has a degenerate distribution with parameter $a = a_1 + a_2$. Note that

$$e^{j\omega a_1} e^{j\omega a_2} = e^{j\omega(a_1 + a_2)}$$

Hence $\Phi_X(\omega) = \Phi_{X_1}(\omega)\Phi_{X_2}(\omega)$.

In terms of distributions we can also write that the distribution of the sum is the convolution of the individual distributions, or $\mathcal{C}(a_1) * \mathcal{C}(a_2) = \mathcal{C}(a_1 + a_2)$.

Sum of Two Gaussian Random Variables

The sum of two independent Gaussian random variables is a Gaussian r.v. The probability density for the sum is the convolution of the densities for the components

$$\Phi_X(\omega) = \exp\left(j\omega m_1 - \frac{\sigma_1^2\omega^2}{2}\right) \cdot \exp\left(j\omega m_2 - \frac{\sigma_2^2\omega^2}{2}\right) = \exp\left(j\omega(m_1 + m_2) - \frac{(\sigma_1^2 + \sigma_2^2)\omega^2}{2}\right)$$

$$\mathcal{N}(a_1, \sigma_1^2) * \mathcal{N}(a_2, \sigma_2^2) = \mathcal{N}(a_1 + a_2, \sigma_1^2 + \sigma_2^2)$$

Sum of Two Poisson Random Variables

First we determine the characteristic function for a Poisson r.v.

$$\Phi_{\mathrm{X}}(\omega) = \exp\left(\lambda_1\left(e^{j\omega} - 1\right)\right) \cdot \exp\left(\lambda_2\left(e^{j\omega} - 1\right)\right) = \exp\left((\lambda_1 + \lambda_2)\left(e^{j\omega} - 1\right)\right)$$

As a result of the above properties these three laws can be limit distribution laws.

Random Processes

So far, we have talked about probability spaces $(\Omega, \mathcal{F}, P)$, and random variables as real-valued functions on this space, i.e. $X: \Omega \to \mathcal{R}$. We expressed this function as $X(\omega)$. We then defined several random variables on the same probability space such as $X(\omega), Y(\omega)$, and $Z(\omega)$, and considered the joint CDFs and PDFs. We extended these considerations to sequences of random variables $X_n(\omega)$ – in fact we considered infinite sequences of random variables. This is a set of random variables indexed by the integers. Note that when we have several random variables defined, for each outcome (of an experiment, ω) we get a set of real numbers such as $X(\omega), Y(\omega)$, and $Z(\omega)$, or in the case of an infinite sequence of random variables we get an infinite sequence of real numbers, i.e. $X_n(\omega), n = 1, \cdots$. We may refer to such a sequence as *a realization.* A random process is a generalization of these ideas where now we consider an indexing set as the set of real numbers instead of a discrete set or the set of integers, although we will consider the case of the integers as the indexing set, as a special case of a random process. In practice, i.e. in applications, the most common case is that the indexing set corresponds to time, although it can also correspond to space. In the case of indexing by the integers we refer to the random process as a *discrete time* random process.

In the above discussion a realization resulted from a specific outcome in the experiment, i.e ω. In the study of random processes we often use the letter ζ to refer to an outcome in the probability space, instead of ω. We will follow this practice here. Hence instead of the elements of the probability space Ω being referred to as ω, we will refer to then as ζ.

Now a realization as discussed above is a collection, or family of random variables indexed by a set. In the case of the indexing set being the real numbers we represent the indexing parameter as t, standing for time. Hence a realization is represented as $X(t)$. So we now think of a realization as a function of the continuous variable t, i.e. the real number t. We also refer to a ***realization*** as a ***sample path***, or ***sample function***

So we have three main cases

1. A finite set of random variables, X, Y, Z, etc. Or $X_1, X_2, \cdots, X_n$
2. Infinite sequence of random variables X_n, $n = 1,2, \cdots$
3. Random process, $X(t)$.

In the above, case 2 is also called a discrete random process.

We may also want to show the dependence on the outcome, ω, or ζ. In this case we would represent the above as

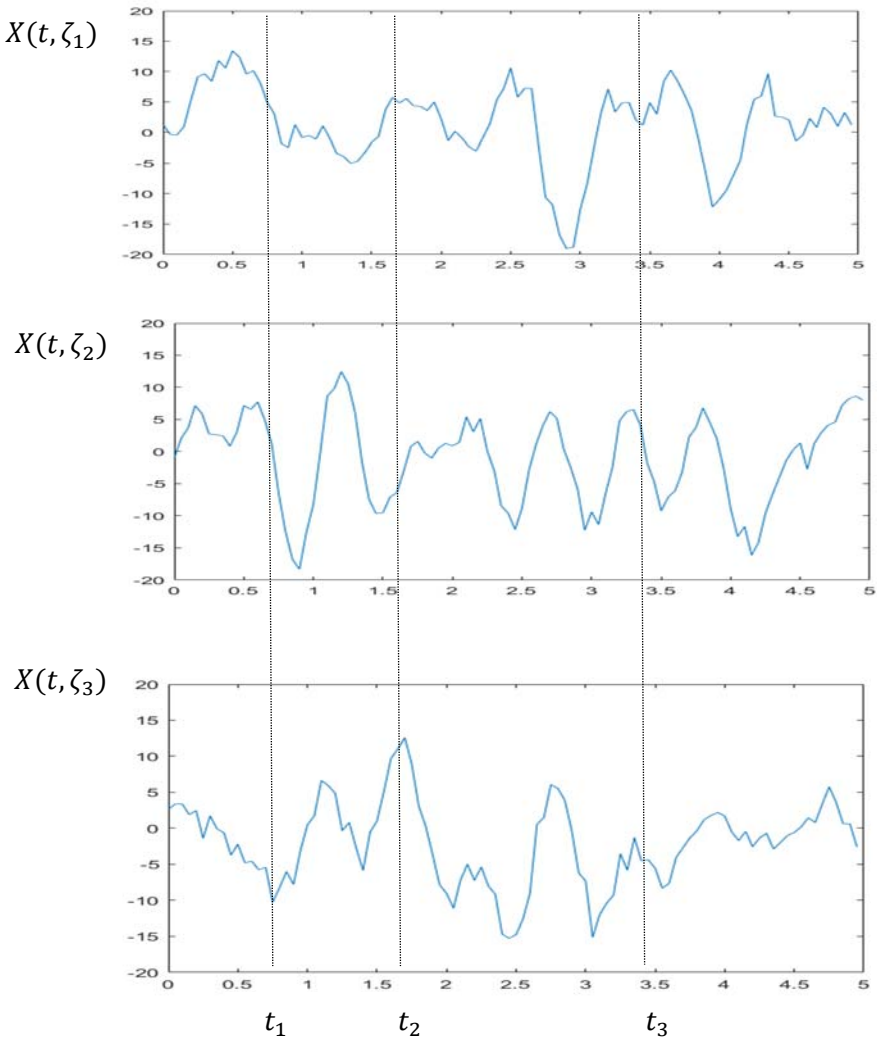
1. $X(\omega), Y(\omega), Z(\omega)$, etc
2. $X_n(\omega)$
3. $X(t, \zeta)$. We could also use the notation $X_t(\omega)$, but it is not as common.

Note that the case 2 can also be represented as $X(n, \zeta)$.

Examples of random processes (i.e. experiments)

1. Turn on a noise generator: A realization is a function of time denoting the noise signal.
2. Observe the temperature during the day.
3. Observe wind speed.
4. Observe the value of a stock in the stock market as a function of time.
5. Observe the value of the stock at the market closing time. This is now a discrete time process.

Realizations of a continuous time process



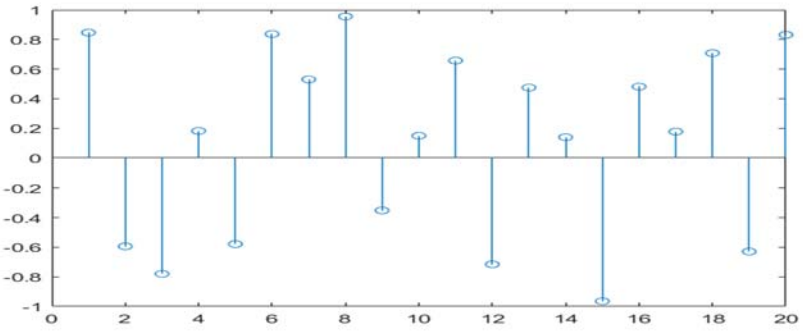
Note that for a fixed t, e.g. t_1, $X(t_1,\zeta)$ is a random variable

For a fixed ζ, e.g. ζ_1, $X(t,\zeta_1)$ is a realization, or sample function.

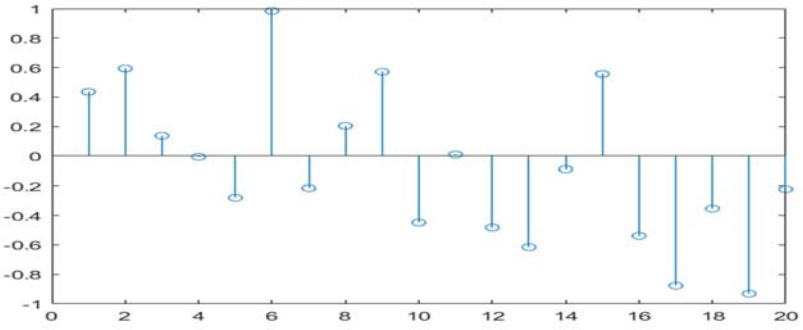
Discrete Time Process

$X(n,\zeta)$, or $X_n(\zeta)$

$X_n(\zeta_1)$



$X_n(\zeta_2)$



Again, for a fixed n, e.g. n_1, $X_{n_1}(\zeta)$ is a random variable.

For a fixed ζ, e.g. ζ_1, $X_n(\zeta_1)$ is a realization.

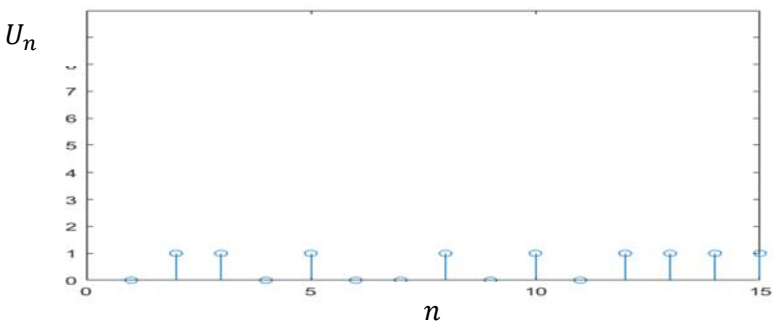
Note that the indexing set for the discrete time random process can be any subset of the integers.

Example: Discrete Time Process

Consider an experiment where we toss a coin repeatedly. A sample outcome is $\zeta = (\nu_1, \nu_2, \cdots)$, where $\nu_i \in \{H, T\}$. We create a sequence of random variables

$$U_n = \begin{cases} 1 & \text{if } \nu_n = H \\ 0 & \text{if } \nu_n = T \end{cases}$$

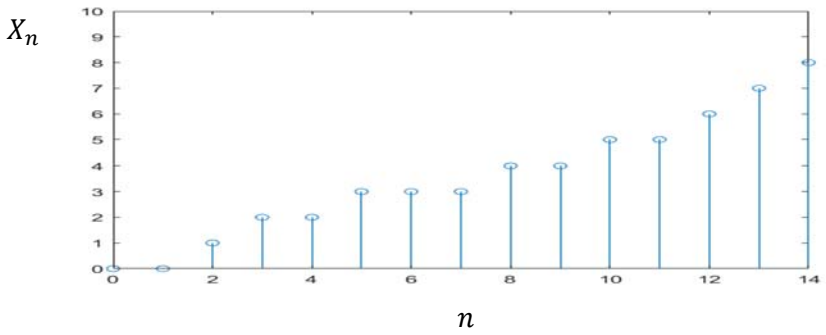
Note that U_n is a discrete time process. The following is a realization (sample function). For simplicity we omit the ζ, and simply write X_n instead of $X_n(\zeta)$.



We can define a second process X_n in terms of U_n as follows:

Set $X_0 = 0$

$X_n = \sum_{i=0}^{n} U_i$, for $n = 1,2,\cdots$



Example

Let ζ be selected at random, with uniform distribution, from the interval $\Omega = [0,1]$. Let $b_1, b_2, \cdots$ be the binary expansion of ζ, i.e. $\zeta = \sum_{i=1}^{\infty} b_i 2^{-i}$.
Definite a discrete random process

$X_n(\zeta) = b_n$, $n = 1,2,\cdots$

E.g. $X_n(0.625) = 1\ 0\ 1\ 0\ 0\ 0\ 0\cdots$

Example of a Continuous Time Random Process

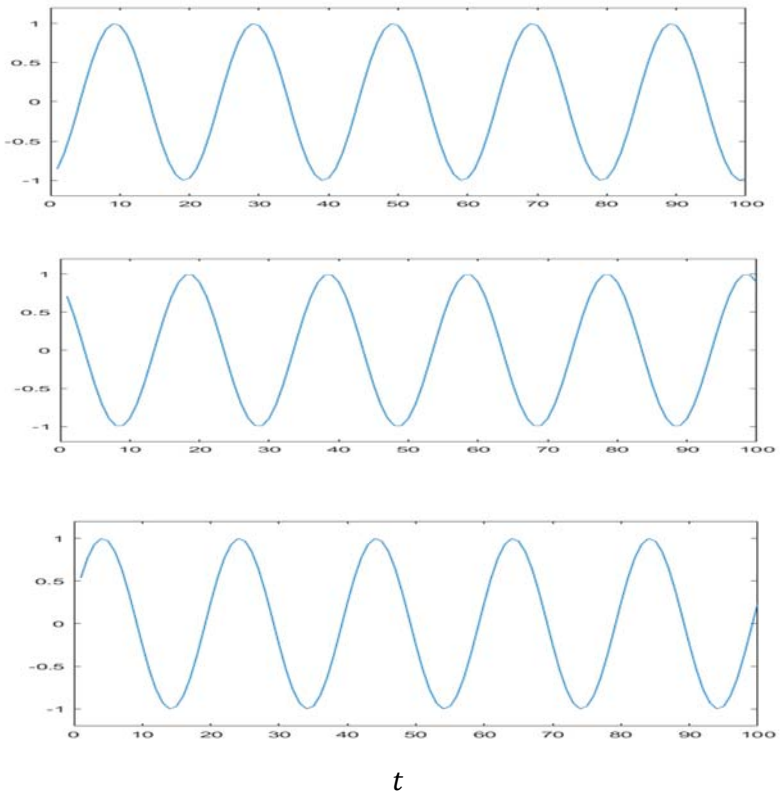
Let $X(t,\theta) = \cos(2\pi f_0 t + \theta)$. Note that we are using θ instead of ζ to index the sample functions. f_0 is the frequency of the carrier in a communication system and θ is the phase angle.

Each realization is a sinusoidal signal with a constant frequency f_0 and a random phase, θ.

$\cos(2\pi f_0 t + \theta_1)$

$\cos(2\pi f_0 t + \theta_2)$

$\cos(2\pi f_0 t + \theta_3)$



t

Probability Law for a Random Process

In the case of a random variable the probability law is specified fully be the CDF.
In the case of a finite set of random variables, the probability law is specified fully by the joint CDF.

In the case of a random process we have a family of random variables that typically is infinite. In this case the probability law is specified fully if it is specified for every possible finite subset of the index set.

Consider a random process $X(t)$, and let $t_1, t_2, \cdots, t_n)$ be an arbitrary fixed set of points in time. Then $X(t_1), X(t_2), \cdots, X(t_n)$ is a set of n random variables. The joint CDF is then $F_{X_{t_1}, X_{t_2}, \cdots, X_{t_n}}(x_1, x_2, \cdots, x_n)$. We must specify the CDFs for any arbitrary n, and an arbitrary fixed set of points in time $t_1, t_2, \cdots, t_n$.

In many applications we do not know these joint CDFs and instead resort to working with various moments of these distributions, such as mean, and variance, and correlations between any two random variables, i.e. for any pair t_1 and t_2.

Mean and Variance

Consider a random process $X(t)$. For a fixed point in time t, $X(t)$ is a random variable for which we may define a mean $m_X(t)$.

$$m_X(t) = \mathcal{E}(X(t))$$

Similarly at each fixed point in time t we may define the variance

$$\sigma_X^2(t) = \mathcal{E}\left\{\left(X(t) - m_X(t)\right)^2\right\}$$

Example

Consider the process that we have displayed above

$$X(t) = \cos(2\pi f_0 t + \theta)$$

For a fixed t, $X(t)$ is a random variable that is a function of the random variable θ. If θ has the uniform distribution in $[-\pi, \pi]$, then we can determine

$$m_X(t) = \mathcal{E}(\cos(2\pi f_0 t + \theta)) = \int_{-\pi}^{\pi} \cos(2\pi f_0 t + \theta)\frac{1}{2\pi}d\theta = 0$$

$$\text{var}\big(X(t)\big) = \int_{-\pi}^{\pi} \cos^2(2\pi f_0 t + \theta)\frac{1}{2\pi}d\theta = \frac{1}{2}$$

Processing a random process by a linear time invariant system

A realization, or a sample function, or a random process is a signal that can be processed by any system including a linear time invariant (LTI) system. Consider a random process $X(t)$, and an LTI system with impulse response $h(t)$. If $X(t)$ is input to the system then the output is the random process $Y(t) = \int_{-\infty}^{\infty} h(t-\tau)X(\tau)d\tau$.

We can compute the mean of $Y(t)$ as follows:

$$m_Y(t) = \mathcal{E}\left(\int_{-\infty}^{\infty} h(t-\tau)X(\tau)d\tau\right) = \int_{-\infty}^{\infty} h(t-\tau)\mathcal{E}\big(X(\tau)\big)d\tau = \int_{-\infty}^{\infty} h(t-\tau)m_X(\tau)d\tau$$

Note that we are assuming that these integrals exist and we are using the linearity of the expected value operator.

Auto-correlation Function of a Random Process

A random process is a family or collection of random variables indexed by a set. For any two indices from the set we have two random variables for which we can determine the correlation in

the usual manner. For a random process, we denote the correlation corresponding to the random variables at t_1 and t_2 as $R_X(t_1, t_2)$, where

$$R_X(t_1, t_2) = \mathcal{E}\big(X(t_1)X(t_2)\big) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X_{t_1},X_{t_2}}(x, y)dxdy$$

and the last equality assumes that the joint density function exists. We also call it the auto-correlation function because it involves two points in time for the same random process.

Note that the auto-correlation function is symmetric in the variables, i.e. $R_X(t_1, t_2) = R_X(t_2, t_1)$.

Similarly for a discrete time random process we can define the auto-correlation function

$$R_X(m, n) = \mathcal{E}(X_m X_n)$$

Example

Consider a random process X_n where the random variables for different n are uncorrelated and have zero mean and equal variances σ_X^2. Find the auto-correlation function $R_X(m, n)$.

Solution:

$$R_X(m, n) = \mathcal{E}(X_m X_n) = \begin{cases} \sigma_X^2 & \text{if } m = n \\ 0 & \text{otherwise} \end{cases}$$

The process X_n is sometimes referred to as *white noise*. The variance σ_X^2 is also referred to as the power of the process.

Example

Consider a process $Y_n = X_n + \alpha X_{n-1}$, where X_n is the above white noise process. Find the autocorrelation for Y_n.

Solution:

$$R_Y(m, n) = \mathcal{E}(Y_m Y_n) = \mathcal{E}\big((X_m + \alpha X_{m-1})(X_n + \alpha X_{n-1})\big)$$
$$= \mathcal{E}(X_m X_n) + \alpha\mathcal{E}(X_m X_{n-1}) + \alpha\mathcal{E}(X_{m-1} X_n) + \alpha^2\mathcal{E}(X_{m-1} X_{n-1})$$

$$= \begin{cases} (1 + \alpha^2)\sigma_X^2 & \text{if } m = n \\ \alpha\sigma_X^2 & \text{if } n = m \pm 1 \\ 0 & \text{otherwise} \end{cases}$$

Note that the process Y_n is actually obtained as a sort of filtering operation on the process X_n. This filtering operation tends to broaden the correlation function.

Properties of the Auto-Correlation Function

1. Cauchy-Schwarz Inequality:

$$|R_X(t_1, t_2)| \leq \sqrt{\mathcal{E}(X^2(t_1))\mathcal{E}(X^2(t_2))}$$

2. The auto-correlation function for a random process has the positive semi-definite property. For any constant vector $(a_1, a_2, \cdots, a_n)$

$$\sum_{i,j} a_i a_j R_X(t_i, t_j) \geq 0$$

This can be shown as follows: To simplify notation let $X_i = X(t_i)$

$$0 \leq \mathcal{E}\left\{\left(\sum_{i=1}^{n} a_i X_i\right)^2\right\} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i\, a_j \mathcal{E}(X_i X_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j r_{ij} = \boldsymbol{a}^T \boldsymbol{R}_X \boldsymbol{a}$$

where $\boldsymbol{a} = (a_1, a_2, \cdots, a_n)$, $r_{ij} = \mathcal{E}(X_i X_j) = \mathcal{E}\left(X(t_i)X(t_j)\right) = R_X(t_i, t_j)$, and $\boldsymbol{R}_X$ is the correlation matrix for the random vector $X = (X_1, X_2, \cdots, X_n)$.

Example: Consider the process, $Y_n = X_n + \alpha X_{n-1}$, as above, where X_n is a zero mean white noise process with power σ_X^2.

Note that $\alpha\sigma_X^2 < (1+\alpha)\sigma_X^2 = \sqrt{(1+\alpha)\sigma_X^2(1+\alpha)\sigma_X^2}$

Hence $R(m, m+1) \leq \sqrt{R_X(m,m)R_X(m+1,m+1)}$ is verified.

Auto-Covariance Function

Again, as for the case of any two random variables, we define the covariance corresponding for the two random variables corresponding to the points in time t_1 and t_2 as

$$C_X(t_1, t_2) = \mathcal{E}\left(\left(X(t_1) - m_X(t_1)\right)\left(X(t_2) - m_X(t_2)\right)\right)$$

This is called the auto-covariance function for the random process because it involves two points in time for the same process.

Note that we can easily show that

$$C_X(t_1, t_2) = R_X(t_1, t_2) - m_X(t_1)m_X(t_2)$$

Correlation Coefficient

In the same way as for two random variables we define the correlation coefficient for a random process as

$$\rho_X(t_1, t_2) = \frac{C_X(t_1, t_2)}{\sqrt{C_X(t_1, t_1)C_X(t_2, t_2)}} = \frac{C_X(t_1, t_2)}{\sigma_X(t_1)\sigma_X(t_2)}$$

As for the case of two random variables we can also show

$$|\rho_X(t_1, t_2)| \leq 1$$

Example

Let $X(t) = \cos(2\pi f_0 t + \theta)$, where f_0 is the carrier frequency in a communication system and θ is a random phase with uniform distribution in $[-\pi, \pi]$. Find the auto-covariance function.

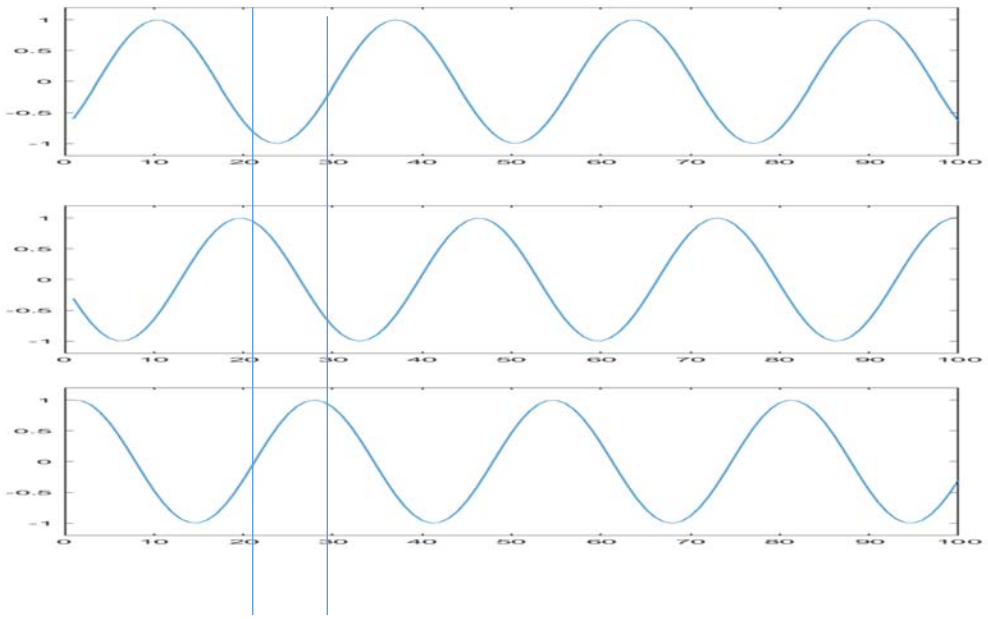
Solution:

We can show that $m_X(t) = 0$

$$C_X(t_1, t_2) = \mathcal{E}\big(X(t_1)X(t_2)\big) = \mathcal{E}(\cos(2\pi f_0 t_1 + \theta)\cos(2\pi f_0 t_2 + \theta))$$

$= \frac{1}{2\pi}\int_{-\pi}^{\pi} \cos(2\pi f_0 t_1 + \theta)\cos(2\pi f_0 t_2 + \theta)\, d\theta = \frac{1}{2}\cos\big(2\pi f_0(t_1 - t_2)\big)$.

Note that if we pick $f_0(t_1 - t_2) = k$, an integer then the covariance is maximum, whereas if $|t_1 - t_2| = \frac{1}{4f_0}$, i.e. one quarter of the period, then the covariance is zero.



Example

Let $X(t) = A\cos(\omega t + \theta)$, where A, and θ are independent random variables and θ is uniform on $[-\pi, \pi]$.

$$\mathcal{E}\big(X(t)\big) = \mathcal{E}(A\cos(\omega t + \theta)) = \mathcal{E}(A)\mathcal{E}(\cos(\omega t + \theta)) = \mathcal{E}(A) \cdot 0 = 0$$

The auto-correlation function is $\mathcal{E}\big(X(t_1)X(t_2)\big) = \frac{1}{2}\mathcal{E}(A^2)\mathcal{E}(\cos(\omega(t_1 - t_2) + \cos(\omega t_1 + \omega t_2 + 2\theta))$

Note that $\mathcal{E}(\cos(\omega t_1 + \omega t_2 + 2\theta)) = 0$, i.e. treat ω, t_1, and t_2 as constants and integrate with respect to θ in $[-\pi, \pi]$.

Hence $R_X(t_1, t_2) = \frac{1}{2}\mathcal{E}(A^2)\cos(\omega(t_1 - t_2))$.

Since $\mathcal{E}\big(X(t)\big) = 0$, the auto-covariance is equal to the auto-correlation.

Example

Consider the process $X(t)$, with $m_X(t) = 3$, and $R_X(t_1, t_2) = 9 + 4e^{-0.2|t_2 - t_1|}$.

Define $Z = X(5)$ and $W = X(8)$. Then $\mathcal{E}(Z) = m_X(5) = 3$. $\mathcal{E}(W) = m_X(8) = 3$.

$\mathcal{E}(Z^2) = R_X(5,5) = 9 + 4e^0 = 13$, and $\mathcal{E}(W^2) = R_X(8,8) = 13$.

Therefore Z and W have the same mean and second moment. The variance is $\sigma^2 = \mathcal{E}(Z^2) - \big(\mathcal{E}(Z)\big)^2 = 13 - 3^2 = 4.$

The correlation between Z and W is $\mathcal{E}(ZW) = R(5,8) = 9 + e^{-0.6} = 11.2$

$\mathcal{E}(ZW) - \mathcal{E}(Z)\mathcal{E}(W) = 11.2 - 9 = 2.2$

Example

Find the auto-correlation function of $X_n = Z_1 + Z_2 + \cdots + Z_n$, for $n = 1,2,\cdots$ if the Z_i are zero mean and uncorrelated with common variance $\text{var}(Z_i) = \sigma^2$ for all i.

Solution: $R_X(m,n) = \mathcal{E}(X_m X_n)$. Now assume that $m > n$

$$X_m = X_n + \sum_{i=n+1}^{m} Z_i$$

$$\mathcal{E}(X_m X_n) = \mathcal{E}\left(X_n\left(X_n + \sum_{i=n+1}^{m} Z_i\right)\right) = \mathcal{E}(X_n^2) + \mathcal{E}\left(X_n \sum_{i=n+1}^{m} Z_i\right)$$

$$= \mathcal{E}(X_n^2) + \sum_{i=n+1}^{m} \mathcal{E}(X_n Z_i) = \mathcal{E}(X_n^2) + 0 = n\sigma^2$$

Hence for $m > n$, $\mathcal{E}(X_m X_n) = n\sigma^2$

In general $\mathcal{E}(X_m X_n) = \sigma^2 \min(m,n)$

Two Random Processes

Let $X(t)$ and $Y(t)$ be two random processes. This means that for each ζ we have two families of random variables $X(t,\zeta)$ and $Y(t,\zeta)$. We can define moments of random variables from the two processes which we call the cross-correlation function

$$R_{XY}(t_1,t_2) = \mathcal{E}(X(t_1)Y(t_2))$$

Example

Let $X(t)$ be the input to an LTI system with impulse response $h(t)$ and $Y(t)$ be the output. Then $Y(t) = \int_{-\infty}^{\infty} h(\tau)X(t-\tau)d\tau$. Find the cross-correlation $R_{XY}(t_1,t_2)$ and the auto-correlation $R_Y(t_1,t_2)$.

Solution:

$$R_{XY}(t_1, t_2) = \mathcal{E}\big(X(t_1)Y(t_2)\big) = \mathcal{E}(X(t_1)\int_{-\infty}^{\infty} h(\tau)X(t_2 - \tau)d\tau$$
$$= \int_{-\infty}^{\infty} h(\tau)\mathcal{E}\big(X(t_1)X(t_2 - \tau)\big)d\tau$$
$$= \int_{-\infty}^{\infty} h(\tau)R_X(t_1, t_2 - \tau)d\tau$$

For the auto-correlation of $Y(t)$

$$R_Y(t_1, t_2) = \mathcal{E}\big(Y(t_1)Y(t_2)\big)$$
$$= \mathcal{E}\left(Y(t_2)\int_{-\infty}^{\infty} h(\tau)X(t_1 - \tau)d\tau\right)$$
$$= \int_{-\infty}^{\infty} h(\tau)\mathcal{E}\big(Y(t_2)X(t_1 - \tau)\big)d\tau$$
$$= \int_{-\infty}^{\infty} h(\tau)R_{XY}(t_1 - \tau, t_2)d\tau$$

Orthogonality

The processes $X(t)$ and $Y(t)$ are said to be orthogonal random processes if

$$R_{XY}(t_1, t_2) = 0$$

for all $t_1,\ t_2$.

Cross-Covariance Function

The cross-covariance function for the processes $X(t)$ and $Y(t)$ is defined as

$$C_{XY}(t_1, t_2) = \mathcal{E}\big[\big(X(t_1) - m_X(t_1)\big)\big(Y(t_2) - m_Y(t_2)\big)\big]$$

$$= R_{XY}(t_1, t_2) - m_X(t_1)m_Y(t_2)$$

Uncorrelatedness of Random Processes

The processes $X(t)$ and $Y(t)$ are said to be uncorrelated random processes if

$$C_{XY}(t_1, t_2) = 0$$

for all t_1, t_2.

Example

Let $X(t) = \cos(\omega t + \theta)$ and $Y(t) = \sin(\omega t + \theta)$, where θ is a random variable with unform distribution in $[-\pi, \pi]$. Find the cross-covariance of $X(t)$ and $Y(t)$.

Solution:
Note that $X(t)$ and $Y(t)$ are zero mean processes. The cross-covariance is

$$C_{XY}(t_1, t_2) = \mathcal{E}(\cos(\omega t_1 + \theta)\sin(\omega t_2 + \theta))$$
$$= \mathcal{E}\left(-\frac{1}{2}\sin\big(\omega(t_1 - t_2)\big) + \frac{1}{2}\sin(\omega(t_1 + t_2) + 2\theta)\right)$$
$$= -\frac{1}{2}\sin\big(\omega(t_1 - t_2)\big) = \frac{1}{2}\sin\big(\omega(t_2 - t_1)\big)$$

Note that in the above we have used the identity

$$\cos A \sin B = \frac{1}{2}(\sin(B + A) + \sin(B - A))$$

$$\mathcal{E}(\sin(\omega(t_1 + t_2) + 2\theta)) = 0.$$

Note that the last can be written as $\mathcal{E}(\sin(2\theta + c))$ where $c = \omega(t_1 + t_2)$ is a constant.

Hence $\mathcal{E}(\sin(2\theta + c)) = \frac{1}{2\pi}\int_0^{2\pi} \sin(2\theta + c)\, d\theta = 0$. (integral of a sinusoid over two periods)

Note that $X(t)$ and $Y(t)$ are not uncorrelated.

Independent Random Processes

The random processes $X(t)$ and $Y(t)$ are said to be independent if the vector random variables $\mathbf{X} = \big(X(t_1), X(t_2), \cdots, X(t_k)\big)$ and $\mathbf{Y} = \big(Y(t_1'), Y(t_2'), \cdots, Y\big(t_j'\big)\big)$ are independent for all k, j, and all choices of $t_1, t_2, \cdots, t_k$ and $t_1', t_2', \cdots, t_j'$.

This means that $F_{\mathbf{XY}}\big(x_1, x_2, \cdots, x_k, y_1, y_2, \cdots, y_j\big) = F_{\mathbf{X}}(x_1, x_2, \cdots, x_k) F_{\mathbf{Y}}(y_1, y_2, \cdots, y_j)$

Discrete Time Processes Based on Independent Identically Distributed (i.i.d.) Random Variables

Let X_n be a discrete time random process consisting of a sequence of independent, identically distributed (i.i.d.) random variables with a common CDF, $F_X(x)$. The sequence is called an i.i.d. random process.

The joint CDF for an arbitrary set of time instants $n_1, n_2, \cdots, n_k$ is given by

$$F_{X_1, X_2, \cdots, X_k}(x_1, x_2, \cdots, x_k) = P(X_1 \le x_1, X_2 \le x_2, \cdots, X_k \le x_k) = F_{X_1}(x_1) F_{X_2}(x_2) \cdots F_{X_k}(x_k)$$

Let $m = \mathcal{E}(X_i)$ be the mean for an i.i.d. process

The auto-covariance is the following:
For $n_1 \neq n_2$

$$C_X(n_1, n_2) = \mathcal{E}\left(\left(X_{n_1} - m\right)\left(X_{n_2} - m\right)\right) = \mathcal{E}\left(X_{n_1} - m\right)\mathcal{E}\left(X_{n_2} - m\right) = 0$$

For $n_1 = n_2$

$$C_X(n, n) = \mathcal{E}((X_n - m)^2) = \sigma^2$$

Hence $C_X(n_1, n_2) = \sigma^2 \delta_{n_1 n_2}$

The auto-correlation function is $R_X(n_1, n_2) = \sigma^2 \delta_{n_1 n_2} + m^2$

The Bernoulli Random Process

Let I_n be a sequence of independent Bernoulli random variables. I_n is an i.i.d. process taking values in {0,1} with probabilities $P(I_n = 1) = p$, $P(I_n = 0) = 1 - p$.

Mean $\mathcal{E}(I_n) = p$, Variance $\text{var}(I_n) = p(1 - p)$

The i.i.d. nature of the process makes it easy to compute probabilities such as the probability that the second bit in a sequence is 0 and the seventh bit is 1

$$P(I_2 = 0, I_7 = 1) = P(I_2 = 0)P(I_7 = 1) = p(1 - p)$$

Time Until the First Success in a Bernoulli Process

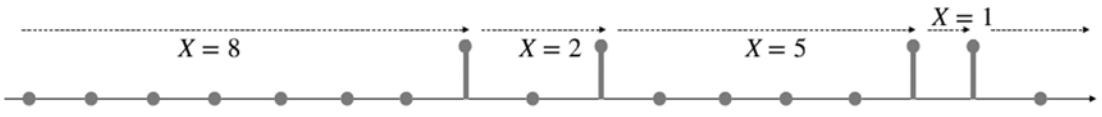
Let X be a random variable equal to the time for the first success in a Bernoulli process

Then $P(X = k) = P(\text{first } k - 1 \text{ bits equal } 0)P\left(k^{\text{th}} \text{ bit equals} = 1\right) = (1 - p)^{k-1}p$.

X has a geometric distribution

$\mathcal{E}(X) = \frac{1}{p}$, $\text{var}(X) = \mathcal{E}(k^2) - \frac{1}{p^2} = \sum_{k=1}^{\infty} k^2(1 - p)^{k-1}p - \frac{1}{p^2} = \frac{1-p}{p^2}$

Exercise: Work out the above sum.



Note X is the time to the next success

X_k is the time to the k^{th} success:
$X_1 = 8, X_2 = 10, X_3 = 15, \ldots$

Time of the k-th Success in a Bernoulli Process

Let X_k be the time of the k-th success in a Bernoulli process. For example, if a Bernoulli process realization is

0 0 1 0 1 1 0 1 0 0 1 0 0 0

Then $X_1 = 3,\ X_2 = 5,\ X_3 = 6,\ X_4 = 8,\ X_5 = 11,$ etc

Find $P(X_k = m)$

$$P(X_k = m) = P(k-1 \text{ success in first } m-1 \text{ time units}) \cdot P(\text{success at time } m)$$

$$= \binom{m-1}{k-1} p^{k-1}(1-p)^{m-k} . p = \binom{m-1}{k-1} p^k (1-p)^{m-k}$$

$$\mathcal{E}(X_k) = \sum_{m=k}^{\infty} m \binom{m-1}{k-1} p^k (1-p)^{m-k}$$

It can be shown with some work that the above sum yields $\mathcal{E}(X_k) = \frac{k}{p}$.

The variance is $\text{var}(X_k) = \sum_{m=k}^{\infty} m^2 \binom{m-1}{k-1} p^k (1-p)^{m-k} - \left(\frac{k}{p}\right)^2 = \frac{k(1-p)}{p^2}$

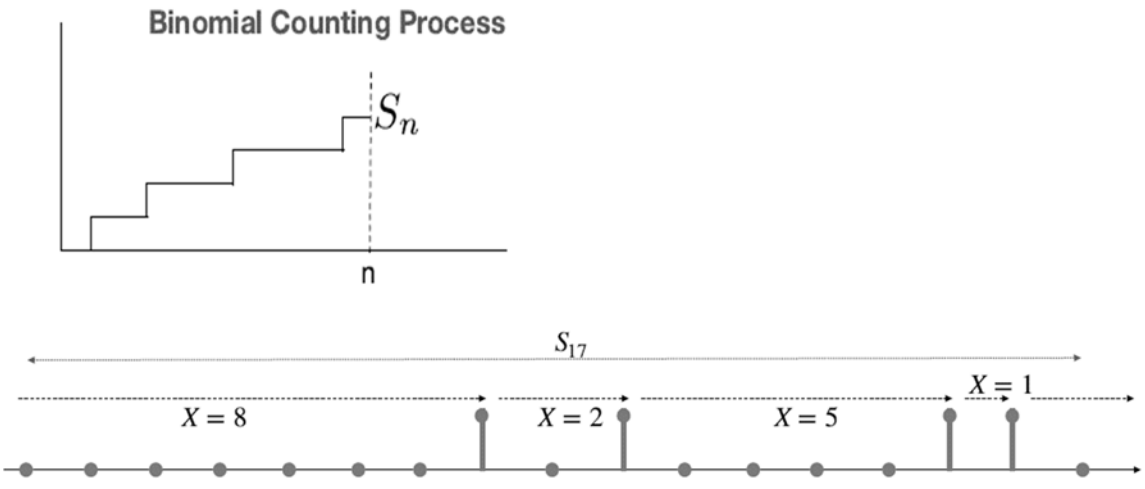
Bernoulli Process: Total Number of Successes up to Time n

Let S_n be the total number of successes in a time interval of length n. Then S_n has a binomial distribution

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

$$\mathcal{E}(S_n) = np$$

$$\text{var}(S_n) = np(1-p)$$



Processes with Independent Increments

A random process $X(t)$ is said to have *independent increments* if the increments in disjoint intervals are independent random variables, i.e., for any integer k, and any set of points in time $t_1 < t_2, \cdots, t_k$, the associated increments $X(t_2) - X(t_1), X(t_3) - X(t_2), \cdots, X(t_k) - X(t_{k-1})$ are independent random variables.

We can show that the joint PDF, or PMF of $X(t_1), X(t_2), \cdots, X(t_k)$ is given by the product of the PDF (PMF) of $X(t_1)$, and the PDFs of the above increment random variables. In other words, let $Y(t_i), i = 1, \cdots, k-1,$ be the increments, i.e.

$X(t_i) = X(t_{i-1}) + Y(t_{i-1}), i = 2, \cdots, k$

Then it can be shown that

$$f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = f_{X_1}(x_1) f_{Y_1}(x_2 - x_1) \cdots f_{Y_{k-1}}(x_k - x_{k-1})$$

See the discussion in section 9.3.3 of the textbook by Leon-Garcia

We can also define the vectors $\boldsymbol{X} = (X_1, X_2, \cdots, X_k)$ and $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_k)$
Where $Y_1 = X_1$, and $Y_i = X_i - X_{i-1}$ for $i = 2,3, \cdots k$. We can write this as $\boldsymbol{Y} = \boldsymbol{AX}$, where A is the following matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \dots & 0 \\ -1 & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & -1 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & -1 & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 & -1 & 1 \end{bmatrix}$$

Then we can consider the linear transformation of a vector random variable $Y = g(X) = AX$

We found that the PDFs can be written as

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{|\det(\boldsymbol{A})|}\Big|_{\boldsymbol{x}=\boldsymbol{A}^{-1}\boldsymbol{y}} = \frac{f_{\boldsymbol{X}}(\boldsymbol{A}^{-1}(\boldsymbol{y}))}{|\det(\boldsymbol{A})|}$$

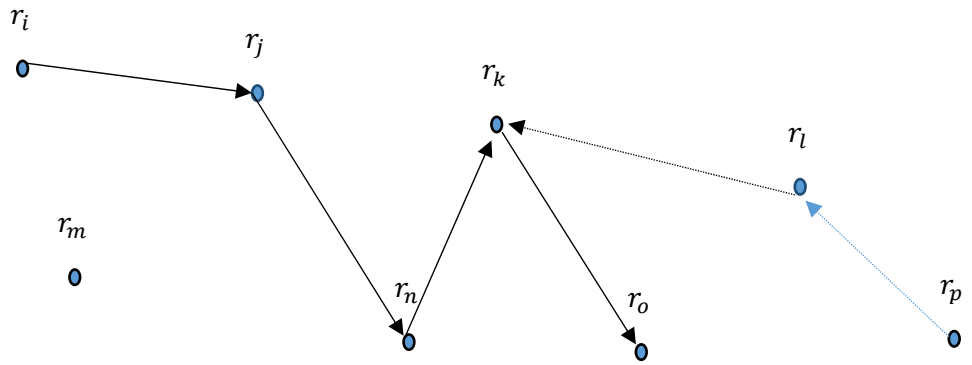
Note that the determinant of A is $|\det(A)| = 1$.

Markov Process

Loosely speaking a random process $X(t)$ is said to be a Markov process if the future of the process depends only on the present and not the past. First we consider a process which takes values in a discrete set $D = \{r_1, r_2, \cdots\}$, i.e. for any t, $X(t) \in D$.
Then the process is a Markov Process if for a sequence of points in time $t_1, t_2, \cdots, t_k, \cdots$

$$P(X(t_k) = x_k | X(t_{k-1}) = x_{k-1}, \cdots, X(t_1) = x_1) \\ = P(X(t_k) = x_k | X(t_{k-1}) = x_{k-1})$$

In the above we are considering t_{k-1} as being the present, and the times $t_{k-2}, \cdots, t_1$ as being the past. The point t_k is in the future. Hence the probabilities for the future (i.e. $P(X(t_k) = x_x)$ given that we know the present state, i.e. $X(t_{k-1})$, and the past states $X(t_{k-2}), \cdots, X(t_1)$, depend only on the present state $X(t_{k-1})$.



It is easy to show that a process with independent increments is a Markov process. However the converse is not necessarily true. A process with the Markov property does not necessarily have independent increments.

Now we assume the case of non-discrete state. In this case we work with probability density functions. The process is Markov if the following holds:

To simplify notation we will write $X(t_i)$ as simply X_i. Then we have

$$f_{X_k}(x_k|X_{k-1} = x_{k-1}, \cdots, X_1 = x_1) \\ = f_{X_k}(x_k|X_{k-1} = x_{k-1})$$

Example: Sum Processes

Consider a sequence of i.i.d. random variables $X_1, X_2, \cdots$

Now form the process $S_n = X_1 + X_2 + \cdots + X_n = S_{n-1} + X_n$, where we have set $S_0 = 0$.

Note that S_n depends on the past only through the sample S_{n-1}. Hence S_n is a Markov process.

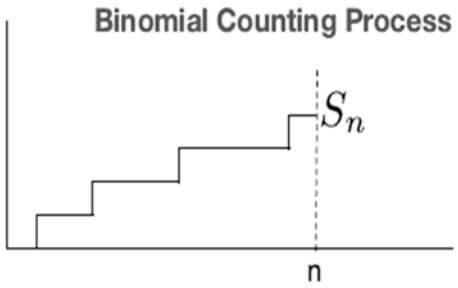
Binomial Counting Process

Let I_i be a sequence of Bernoulli i.i.d. random variables with $I_i \in \{0,1\}$

Define $S_n = I_1 + I_2 + \cdots + I_n$

Then S_n is a counting process that gives the number of successes in the first n Bernoulli trials (i.e. "success" at the i^{th} trial means $I_i = 1$).

The counting process increases with time. Given that it has a certain value at time n e.g. k, it does not matter how this value was achieved, i.e. where the k successes occurred up to time n. For any $m > n$, the distribution for S_m depends only on S_n, and not any of the $S_1, S_2, \cdots, S_{n-1}$.



S_n is a binomial random variable with parameters n and $p = P(I_i = 1)$.

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

Note that S_n is also a process of independent increments.

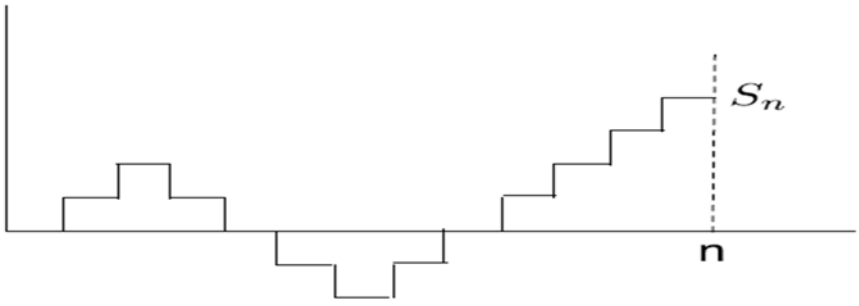
Random Walk

Let D_k be an i.i.d. process with probabilities $P(D_k = 1) = p, \quad P(D_k = -1) = 1 - p$.
Let $S_n = D_1 + D_2 + \cdots + D_n$

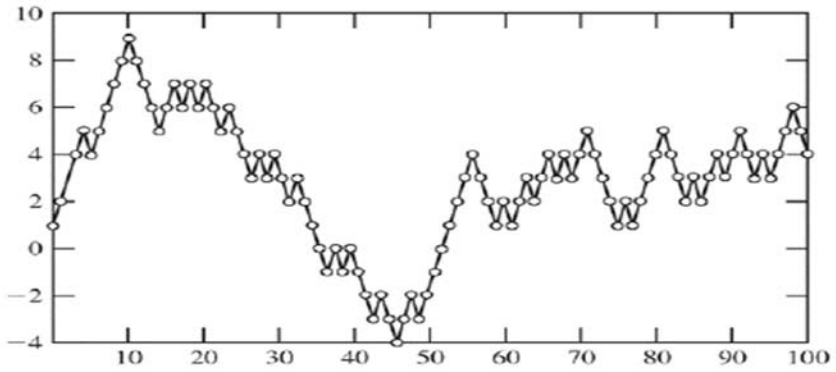
Note that the maximum of S_n is n, and the minimum is $-n$.

This process is an example of a one-dimensional random walk. We flip a fair coin with $P(\text{heads} = p)$. If "heads" arise then we take one step to the right. If "tails" arise then we take one step to the left.

Sample Path (Realization)



Probability Mass Function (Marginal) for a Random Walk



Let us determine the probabilities for the random variable S_n, i.e. the value of the process at time n.

If there are k steps with value $+1$ in the first n steps, then the other $n-k$ steps must have a value of -1. Hence $S_n = k - (n-k) = 2k - n$.

Thus,

$$P(S_n = 2k - n) = \binom{n}{k} p^k (1-p)^{n-k}$$

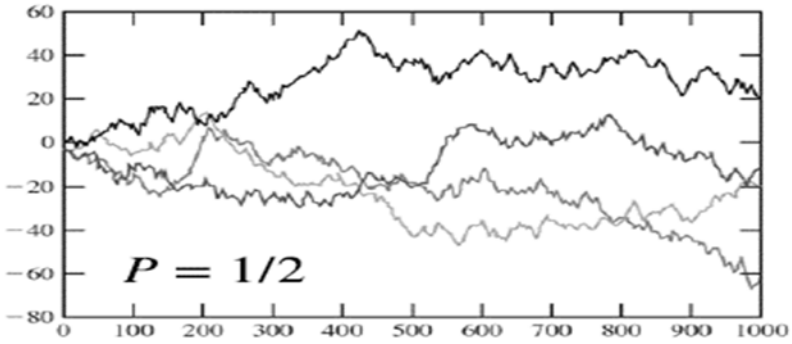
Mean of the Random Walk

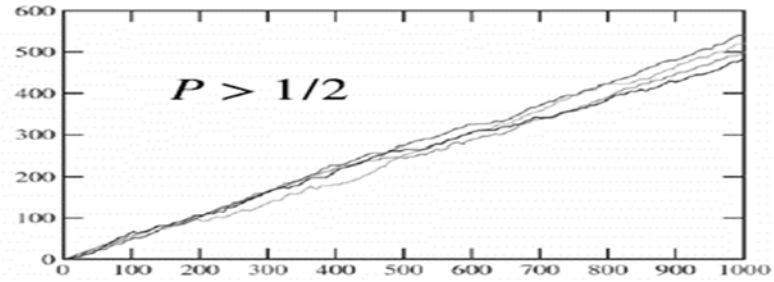
$$m_{S_n}(n) = \mathcal{E}(S_n) = \mathcal{E}(2k - n) = 2\mathcal{E}(k) - n = 2np - n = n(2p - 1)$$

For $p = \frac{1}{2}$, $\mathcal{E}(S_n) = 0$.

If $p > \frac{1}{2}$, then $\mathcal{E}(S_n) = \alpha n$, where $\alpha > 0$. Hence $\mathcal{E}(S_n) \to \infty$, as $n \to \infty$.

If $p < \frac{1}{2}$, then $\mathcal{E}(S_n) = -\alpha n$, where $\alpha > 0$. Hence $\mathcal{E}(S_n) \to -\infty$, as $n \to \infty$.





Variance of a Random Walk Process

Let S_n be a random walk process. As above we let k be the number of successes in n steps, i.e. the number of steps to the right. Then $S_n = 2k - n$. Now, fix n, the variance of S_n is then

$$\text{var}(S_n) = \text{var}(2k - n) = \mathcal{E}[(2k - n)^2] - \big(\mathcal{E}(2k - n)\big)^2$$

$$= \mathcal{E}(4k^2 - 4kn + n^2) - \big(\mathcal{E}(2k) - \mathcal{E}(n)\big)^2$$

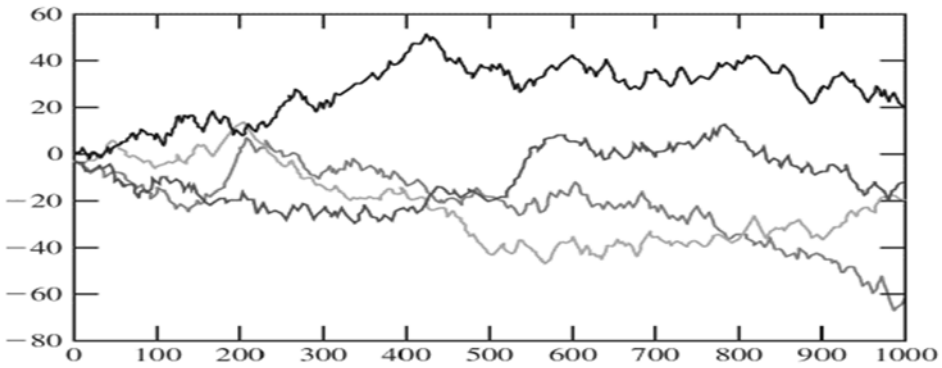
$$= 4\mathcal{E}(k^2) - 4n\mathcal{E}(k) + \mathcal{E}(n^2) - \mathcal{E}(2k)^2 + 2\mathcal{E}(2k)\mathcal{E}(n) - \mathcal{E}(n)^2 = 4\mathcal{E}(k^2) - 4\big(\mathcal{E}(k)\big)^2$$

$$= 4\text{var}(k) = 4np(1 - p)$$

For a fixed p the variance of the random walk increases with n. E.g. for $p = \frac{1}{2}$, the variance is

$$\text{var}(S_n) = n$$

The following shows different sample paths for the random walk process



Independent and Stationary Increments of Sum Processes of i.i.d. Random Variables

If S_n is the sum of n i.i.d. random variables, i.e. $S_n = X_1 + X_2 + \cdots + X_n$, where the X_i are i.i.d., then S_n has independent increments in non-overlapping intervals.

The increments in two non-overlapping intervals do not have any X_i's in common.

For $n > k$, the increment $S_n - S_k$ is the sum of $n - k$ i.i.d. random variables, hence it has the same distribution as S_{n-k}, the sum of the first $n - k$ random variables $X_1, X_2, \cdots, X_{n-k}$, that is

$$P(S_n - S_k = y) = P(S_{n-k} = y)$$

Thus, the increments in intervals of the same length have the same distribution regardless of where the interval begins. For this reason we say that the process S_n has *stationary increments*.

Joint PMF (PDF)

The independent and stationary increments property of the sum process makes it easy to compute the joint PMF/PDF for the process at any set of time points.

For simplicity suppose that the X_n are integer valued, hence S_n is also integer-valued.

We compute the joint PMF of S_n at times n_1 and n_2, where $n_1 < n_2$.

$$\begin{aligned} P(S_{n_1} = y_1, S_{n_2} = y_2) &= P(S_{n_1} = y_1, S_{n_2} - S_{n_1} = y_2 - y_1) \\ &= P(S_{n_1} = y_1)P(S_{n_2} - S_{n_1} = y_2 - y_1) \\ &= P(S_{n_1} = y_1)P(S_{n_2 - n_1} = y_2 - y_1) \end{aligned}$$

Note that the second line above follows because S_{n_1} and $S_{n_2} - S_{n_1}$ and independent.

If the X_n are continuous valued, then the above can be generalized with probability densities

$$f_{S_{n_1}S_{n_2}}(y_1, y_2) = f_{S_{n_1}}(y_1) f_{S_{n_2-n_1}}(y_2 - y_1)$$

Example – Joint PMF of Binomial Counting Process

Let S_n be the Binomial counting process. Then

$$P(S_{n_1} = y_1, S_{n_2} = y_2) = P(S_{n_1} = y_1) P(S_{n_2-n_1} = y_2 - y_1)$$

$$= \binom{n_1}{y_1} p^{y_1}(1-p)^{n_1-y_1} \cdot \binom{n_2-n_1}{y_2-y_1} p^{y_2-y_1}(1-p)^{n_2-n_1-y_2+y_1}$$

$$= \binom{n_1}{y_1} \cdot \binom{n_2-n_1}{y_2-y_1} p^{y_2}(1-p)^{n_2-y_2}$$

Find the probability that the first n_1 trials are all failures and the remaining trials are all successes.

For the first n_1 being failures, $y_1 = 0$, and for the others to be successes we have $y_2 - y_1 = y_2 = n_2 - n_1$.

Hence $P(S_{n_1} = 0, S_{n_2} = n_2 - n_1) = \binom{n_1}{0}\binom{n_2-n_1}{n_2-n_1} p^{n_2-n_1}(1-p)^{n_1} = p^{n_2-n_1}(1-p)^{n_1}$

Random Processes: Time and State

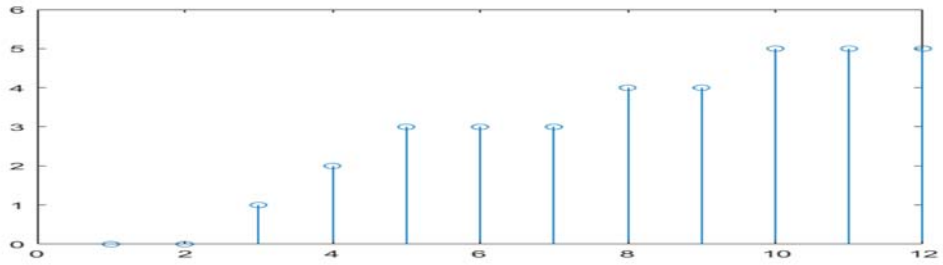
The realizations of a random process are functions of time, $y = X(t)$. Each realization can be plotted as a function in the $t - y$ plane. We refer to the value t as time and the value y as the state of the process, i.e. the variable plotted in the horizontal axis is *time*, and the variable plotted in the vertical axis is the *state*.

The time variable t and the state variable y may be considered to be discrete, i.e. take discrete values, or continuous. As a result there are 4 cases to be considered as follows:

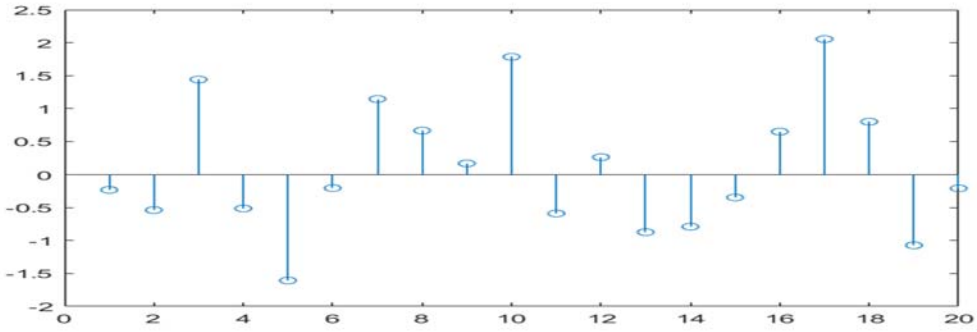
1. Discrete Time, Discrete State (DT-DS)
2. Discrete Time, Continuous State (DT-CS)
3. Continuous Time, Discrete State (CT-DS)
4. Continuous Time, Continuous State (CT-CS)

Examples of the 4 Different Types of Random Processes

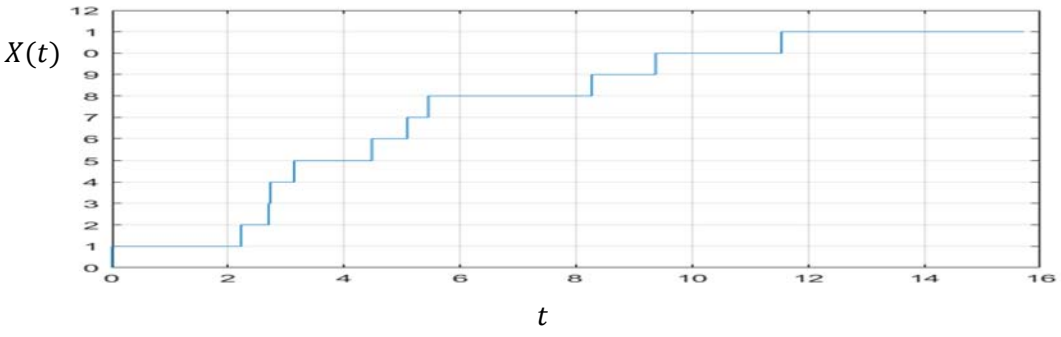
1. DT-DS: The Binomial counting process is a DT-DS process, because it is obviously a discrete time process and the values S_n are discrete. The following is a realization of the process with $p = \frac{1}{2}$.



2. DT-CS: Consider the process X_n, where the X_n are i.i.d. random variables with Gaussian distribution. This is a DT-CS process. The following is a sample function (a realization) of a discrete time Gaussian process with mean 0 and variance 1.



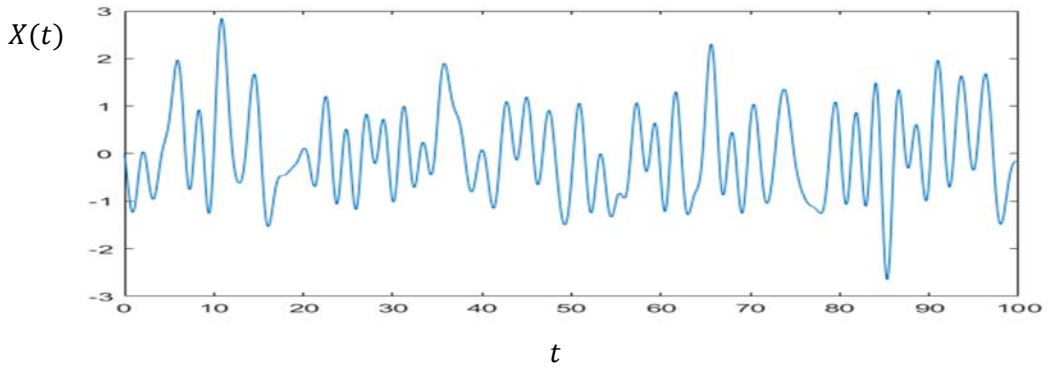
3. CT-DS: Consider the following random process $X(t)$. Each realization is determined from an infinite sequence of i.i.d. random variables, E_n, with exponential distribution with parameter λ. That is, the E_n have a PDF given by $f_E(x) = \lambda e^{-\lambda x}$. We set $X(t) = 0$ for $t \leq 0$, then $X(t) = 0$ for $t < E_1$, $X(t) = 1$ for $E_1 \leq t < E_2$, …, $X(t) = k - 1$ for $E_{k-1} \leq t < E_k$. This is known as a Poisson Process. It is a CT-DS process. The following is a realization of the Poisson process with $\lambda = 1$.



4. CT-CS: Consider the process $X(t) = A\cos(2\pi f_0 t + \theta)$, where A and θ are random variables. This is a CT-CS process. Another example of a CT-CS process is the following. Consider a sequence of i.i.d. Gaussian distributed random variables, X_n. Then we can form the process

$$X(t) = \sum_{k=-\infty}^{\infty} X_k \frac{\sin(t-k)}{t-k}$$

In other words, the sequence of random variables X_k is used as samples to reconstruction a continuous time signal $X(t)$. This is known as a Gaussian Process. It can be used to model a noise signal. The following is a sample function of a Gaussian process using the above construction.



Derivation of the Poisson Counting Process

We now derive the Poisson Process as a limit of the Binomial Counting Process.

We consider a scenario where events occur at random instants of time at an average rate of λ events per second.

$N(t)$ = number of event occurrences in the time interval $[0, t]$.

$N(t)$ is a non-decreasing, integer-valued, continuous-time random process.

Divide time into small intervals of duration $\delta_n = t/n$.

Assume the following two conditions:

- The probability of more than one event occurrence in an interval of duration δ_n is negligible compared to the probability of 0 or 1 events occurring

$$P(\text{2 or more events in } \delta_n \text{ seconds}) = 0$$

- The occurrence of an event in one of the intervals (of duration δ_n) is independent of the occurrence of an event in any other interval (of duration δ_n).

The first assumption implies that the occurrence of an event in an interval is a Bernoulli trial. The second assumption implies that the Bernoulli trials are independent.

As a result, with the two assumptions, the counting process $N(t)$ can be modelled as a Binomial counting process.

The expected value of the number of events in the interval $[0,t]$ = np.

The variance of the number of events in the interval $[0,t]$ is $np(1-p)$. The "distance" between any two events is geometrically distributed.

Now Take the Limit as $n \to \infty$

Consider $p = \lambda\delta_n = \lambda t/n$. As n increases p decreases.

Let S_n be the number of events that occur in the interval $[0,t]$. As a result of the independence assumption the characteristic function for S_n factors as a product of n characteristic functions of Bernoulli random variables

$$\Phi_{S_n}(\omega) = \mathcal{E}\left(e^{j\omega S_n}\right) = \Pi_{k=1}^{n}\mathcal{E}\left(e^{j\omega X_n}\right) = \Pi_{k=1}^{n}\left(pe^{j\omega} + (1-p)\right) = \left(1 + p\left(e^{j\omega} - 1\right)\right)^n$$

$$= \left(1 + \frac{\lambda t}{n}\left(e^{j\omega} - 1\right)\right)^n$$

Now as $n \to \infty$ $\Phi_{S_n}(\omega) \to \exp\left(\lambda t\left(e^{j\omega} - 1\right)\right)$

This is the characteristic function of a Poisson random variable the parameter λt. In other words the process $N(t)$, at t, has a Poisson distribution with parameter λt, i.e.

$$P(N(t) = k) = \frac{(\lambda t)^k}{k!}e^{-\lambda t}$$

For $k = 0,1,2,\cdots$

$$\mathcal{E}\left(N(t)\right) = \lambda t$$

$$\text{var}\left(N(t)\right) = \lambda t$$

Independent and Stationary Increments of Poisson Process

The Poisson Process has independent and stationary increments. This is a consequence of the Poisson process being a limit of the Binomial counting process.

As a consequence of the independent and stationary increments we can write the joint PDF for the process $N(t)$ at any number of points. For example, if $t_1 < t_2$, then

$$P(N(t_1) = i, N(t_2) = j)$$

$$= P(N(t_1) = i \;\&\; N(t_2) - N(t_1) = j - i)$$

$$= P(N(t_1) = i) \cdot P(N(t_2) - N(t_1) = j - i)$$

$$= P(N(t_1) = i) \cdot P(N(t_2 - t_1) = j - i)$$

$$\frac{(\lambda t_1)^i e^{-\lambda t_1}}{i!} \cdot \frac{\big(\lambda(t_2 - t_1)\big)^{j-i} e^{-\lambda(t_2 - t_1)}}{(j-i)!}$$

Example

Patients arrive at a dentist office at an average rate of 2 customers per hour. We model the arrival process as a Poisson process. What is the probability that 2 patients arrive in the first hour and 3 patients arrive in the next 2 hours.

We denote the number of arrivals in the first interval as X_1, and the number of arrivals in the second interval as X_2. Using the independence of arrivals in the two intervals

$$P(X_1 = 2, X_2 = 3) = P(X_1 = 2)P(X_2 = 3) = \frac{(\lambda t_1)^2 e^{-\lambda t_1}}{2!} \cdot \frac{(\lambda t_2)^3 e^{-\lambda t_2}}{3!}$$

Note that $\lambda t_1 = 2 \times 1 = 2$, and $\lambda t_2 = 2 \times 2 = 4$. Substitute in the above to obtain 0.05.

Auto-Covariance of the Poisson Process

$$C_N(t_1, t_2) = \mathcal{E}\big((N(t_1) - \lambda t_1)(N(t_2) - \lambda t_2)\big)$$

$$= \mathcal{E}\big[(N(t_1) - \lambda t_1)\big(N(t_2) - N(t_1) - \lambda(t_2 - t_1) + (N(t_1) - \lambda t_1)\big)\big]$$

$$= \mathcal{E}\left((N(t_1) - \lambda t_1)\big(N(t_2) - N(t_1) - \lambda(t_2 - t_1)\big)\right) + \mathcal{E}((N(t_1) - \lambda t_1)^2)$$

$$= \mathcal{E}((N(t_1) - \lambda t_1)^2) = \mathrm{var}\big(N(t_1)\big) = \lambda t_1$$

Note that the first term of the second last expression is zero due to the independent increment assumption..

We can also write the above succinctly as $C_N(t_1, t_2) = \lambda \min(t_1, t_2)$.

Inter-Arrival Times for the Poisson Process

To determine the distribution for the inter-arrival times we again start with a Binomial counting process using very small intervals of length δ_n.

Let M be the random variable equal to the time of the first occurrence, in units of intervals of length δ_n, i.e. first arrival.

$$P(M > m) = (1 - p)^m$$

Where p is the probability of an occurrence in an interval.

$$P(M \leq m) = 1 - (1 - p)^m.$$

This is the CDF for a geometrically distributed random variable. The expected value is $1/p$ as we have shown before.

Let $\delta_n = 1/n$, i.e. we create n sub-intervals per unit time. Then if the expected number of occurrences per unit time is λ, the expected number of occurrences per sub-interval is $p = \frac{\lambda}{n}$.

$$P(T \leq t) = P(M\delta_n \leq m\delta_n) = P(M \leq m) = 1 - \left(1 - \frac{\lambda}{n}\right)^m$$

$$t = m\delta_n = \frac{m}{n}$$

$$= 1 - \left(1 - \frac{\lambda}{n}\right)^{nt} = 1 - \left(1 - \frac{\lambda t}{nt}\right)^{nt}$$

and as $n \to \infty$, $P(T \leq t) \to 1 - e^{-\lambda t}$. This is the CDF for an exponential random variable with parameter λ, i.e. $F_T(t) = 1 - e^{-\lambda t}$. The PDF is $f_T(t) = \lambda e^{-\lambda t}u(t)$, where $u(t)$ is the step function. The random variable T is also known as the inter-arrival time for the Poisson process.

Memoryless Property

The Poisson process also has what we call the memoryless property for the inter-arrival time distribution. A random variable T, is call memoryless if the following holds

$$P(T > t + s | T > s) = P(T > t)$$

In other words suppose that after a given occurrence we have waited s for the next occurrence. Then the probability that we have to wait another t seconds (i.e. the total wait is $s + t$) is equal to the probability that we wait for t seconds after an occurrence. Another way to say this is that if we have waited a long time for the next occurrence it does not mean that the next occurrent is likely to be soon.

Suppose that the inter-arrival time is a constant equal to T. Then if we have waited for a time that is close to T then the remaining wait is small. Hence such a process is not memoryless.

The left side of the above equation can also be written as (i.e. using the definition of conditional probability)

$$\frac{P(T > t + s \,\&\, T > s)}{P(T > s)} = \frac{P(T > t + s)}{P(T > s)}$$

since the event $(T > t + s) \subset (T > s)$, i.e. $(T > t + s) \cap (T > s) = (T > t + s)$.

Hence

$$\frac{P(T > t + s)}{P(T > s)} = P(T > t)$$

Thus $P(T > t + s) = P(T > s)P(T > t)$.

Let $g(x) = P(T > x)$. Then $g(t+s) = g(s)g(t)$. The only real function that satisfies this relation is the exponential function, i.e. $g(x) = e^{cx}$, where c is a constant.

Example

The number of taxis arriving at the taxi stand at an airport in any time interval of t seconds is a Poisson random variable with expected value equal to $2t$. A passenger has been waiting for 3 minutes for a taxi, What is the probability that a taxi will arrive in the next minute?

Solution: For a Poisson process the inter-arrival times are exponentially distributed with parameter 2. Hence

$$P(T \leq 4|T > 3) = 1 - P(T > 4|T > 3) = 1 - P(T > 1) = P(T \leq 1) = 1 - e^{-2\times 1} = 1 - e^{-2}$$

Note that for the second equality we invoked the memoryless property.

Note also that the assumption that the process can be modelled as a Poisson process is one that should be validated in practice. It is only a mathematical model applied to the real world.

Memoryless Exponential Becomes Geometric Distribution for Discrete Time

As we have stated the exponential distribution is the only memoryless distribution for a continuous random variable, i.e.

$$P(T > t+s|T > s) = \frac{P(T > t+s, T > s)}{P(T > s)} = \frac{P(T > t+s)}{P(T > s)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t}$$

Now assume that t is a discrete time random variable, i.e. $t = n\Delta t$, for $n = 0,1,2,\cdots$

Let $s = m\Delta t$. Then the above would become, setting $T = M\Delta t$

$$P(M\Delta t > n\Delta t + m\Delta t|M\Delta t > m\Delta t) = \frac{e^{-\lambda(n\Delta t+m\Delta t)}}{e^{-\lambda m\Delta t}} = e^{-\lambda n\Delta t} = \left(e^{-\lambda\Delta t}\right)^n = (1-p)^n$$

where we have set $e^{-\lambda\Delta t} = 1 - p$, i.e. we have defined $p = 1 - e^{-\lambda\Delta t}$.

Since Δt is a constant, if we consider the random variable to be the integer M then we have

$$P(M > n + m|M > m) = (1-p)^n$$

For $m = 0$, we have $P(M > n|M > 0) = P(M > n) = (1-p)^n$.

$P(M \leq n) = 1 - (1-p)^n$.

Note that this is the CDF for the geometric distribution. In other words, the exponential distribution, when restricted to a discrete set of points in time, becomes the geometric distribution.

Example

The chance of winning the jackpot in weekly lottery is 10^{-6}. I have been playing the lottery every week for the past 10 years and have never won. What is the probability that I will win the lottery over the next year.

$P(\text{win in the next 53 weeks}|\text{did not win in the past 10 years}) =$
$P(\text{win in the next 53 weeks}) = 1 - P(\text{no win in the next 53 weeks}) = 1 - (1 - 10^{-6})^{53}$.

Poisson Process: Time of Second Arrival

We start at the time origin, 0, and let T_1 be the time of the first arrival, and T_2 be the inter-arrival time between the 1st and 2nd arrivals. Then T_1 and T_2 are independent random variables with exponential distribution with parameter λ, i.e. the PDF for both is $f_T(t) = \lambda e^{-\lambda t}u(t)$. Now, let $E_2 = T_1 + T_2$. The PDF for E_2 is the convolution of the PDFs for T_1 and T_2, i.e. $(f_T * f_T)(t)$.

We evaluate

$$f_{E_2}(t) = (f_T * f_T)(t) = \int_{-\infty}^{\infty} f_T(\tau) f_T(t-\tau) d\tau = \int_{-\infty}^{\infty} \lambda^2 u(\tau) e^{-\lambda\tau} e^{-\lambda(t-\tau)} u(t-\tau) d\tau$$

$$\lambda^2 \int_0^t e^{-\lambda\tau} e^{-\lambda(t-\tau)} d\tau = \lambda^2 e^{-\lambda t} \int_0^t d\tau$$

For $t \geq 0$, and 0 for $t < 0$.

$$= \lambda^2 t e^{-\lambda t} u(t)$$

This is the PDF for the so-called Erlang-2 distribution.

Poisson Process: Time of the n^{th} arrival

The above generalizes to the time of the n^{th} arrival. Let T_k be the inter-arrival time between the $(k-1)^{\text{th}}$ and k^{th} arrivals. Then $E_n = T_1 + T_2 + \cdots + T_n$ is the time of the n^{th} arrival. Since E_n is the sum of n independent random variables, the PDF for E_n is the convolution of the PDFs for the T_k's. The characteristic function for T is $\Phi_T(\omega) = \mathcal{E}\left(e^{j\omega T}\right) = \int_{-\infty}^{\infty} e^{j\omega t} \lambda e^{-\lambda t} u(t) dt$

$$= \int_0^{\infty} \lambda\, e^{-(\lambda - j\omega)t} dt = \frac{\lambda}{\lambda - j\omega}$$

The characteristic function for E_n is then $\Phi_{E_n}(\omega) = \left(\frac{\lambda}{\lambda - j\omega}\right)^n$. The PDF for E_n is then obtained as an inverse Fourier Transform and can be shown to be

$$f_{E_n}(t) = \frac{(\lambda t)^{n-1}}{(n-1)!} \lambda e^{-\lambda t}$$

For $t \geq 0$.

Note that in the above we have used the step function $u(t) = 1$, for $t \geq 0$, and 0 elsewhere. We have also stated that the PDF for the exponential random variable is $f_T(t) = \lambda e^{-\lambda t}$. This assumes that we implicitly acknowledge that the random variable takes only positive values. However strictly speaking we should always consider the PDF as ranging over the whole real line, and write instead $f_T(t) = \lambda e^{-\lambda t} u(t)$. This is important when we have expressions that involve the integral of $f_T(t)$, especially with convolutions.

The above PDF is known as the Erlang-n PDF.

Poisson Process: Conditional Arrival Time (Given one Arrival)

Suppose we are given that for a Poisson process, the event $N(t) = 1$, i.e. in the interval $[0, t]$ there is one arrival, occurred. What is the probability density function for the arrival time, T, i.e. $f_T(t_1|N(t) = 1)$? Obviously whatever it is, we know that the PDF for this conditional arrival time is 0 for $t_1 > t$.

Consider $t_1 < t$.

$$F_T(t_1|N(t) = 1) = P(T \leq t_1|N(t) = 1) = \frac{P(N(t)=1, N(t_1)=1)}{P(N(t)=1)} = \frac{P(N(t_1)=1, \text{zero arrivals in } [t_1,t])}{P(N(t)=1)} = \frac{\lambda t_1 e^{-\lambda t_1} e^{-\lambda(t-t_1)}}{\lambda t e^{-\lambda t}} = \frac{t_1}{t}$$

For $t_1 > t$ $P(T \leq t_1|N(t) = 1) = 1$

Hence the PDF is $f_T(t_1|N(t) = 1) = \frac{1}{t}$, that is, the arrival time is a uniform random variable in the interval $[0, t]$. In other words, given that we have one arrival in the interval $[0, t]$, the time of arrival, T, is a random variable with a uniform distribution in the interval $[0, t]$.

Poisson Process: Conditional Arrival Times (Given n Arrivals)

Now, suppose we are given that n arrivals occurred in the interval $[0, t]$. Denote the arrival instants by the random variables $T_1, T_2, \cdots, T_n$, where T_1 is the time for the first arrival, T_2 is the time for the second arrival, etc. What is the PDF for the vector $(T_1, T_2, \cdots, T_n)$, i.e. $f_{T_1,T_2,\cdots,T_n}(t_1, t_2, \cdots, t_n \mid N(t) = n)$?

First we compute the CDF

Consider the region R_n in $\mathcal{R}^n$, defined as follows:

$$R_n = \{(t_1, t_2, \cdots, t_n) \in \mathcal{R}^n : 0 \leq t_1 < t_2 < \cdots < t_n \leq t\}$$

We compute the CDF at points in R_n

$$F_{T_1,\cdots,T_n}(T_1 \leq t_1, T_2 \leq t_2, \cdots, T_n \leq t_n|N(t) = n)$$

$$= \frac{P(T_1 \leq t_1, T_2 \leq t_2, \cdots, T_n \leq t_n, N(t) = n)}{P(N(t) = n)}$$

$$= \frac{[(\lambda t_1)e^{-\lambda t_1}][\lambda(t_2 - t_1)e^{-\lambda(t_2-t_1)}] \cdots [\lambda(t_n - t_{n-1})e^{-\lambda(t_n-t_{n-1})}]e^{-\lambda(t-t_n)}}{\frac{(\lambda t)^n e^{-\lambda t}}{n!}}$$

$$= \frac{t_1(t_2 - t_1)\cdots(t_n - t_{n-1})\lambda^n e^{-\lambda t}}{\frac{(\lambda t)^n e^{-\lambda t}}{n!}}$$

$$= \frac{n!\, t_1(t_2 - t_1)\cdots(t_n - t_{n-1})}{t^n}$$

The PDF can be obtained by taking the partial derivatives with respect to $t_1, t_2, \cdots, t_n$, in the region R_n to obtain

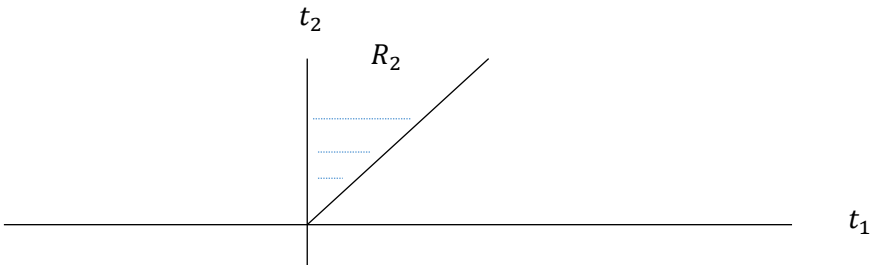
$$f_{T_1,\cdots,T_n}(t_1, \cdots, t_n | N(t) = n) = \frac{n!}{t^n}$$

Outside R_n the CDF is a constant and the PDF is equal to 0.

In other words

$$f_{T_1,T_2,\cdots,T_n}(t_1, t_2, \cdots, t_n \mid N(t) = n) = \begin{cases} \frac{n!}{t^n} & \text{for } (t_1, t_2, \cdots, t_n) \in R_n \\ 0 & \text{elsewhere} \end{cases}$$

Note that for $N = 2$, the region R_2 is shown below



Example

Suppose that two customers arrive at a shop during a two minute period. Find the probability that the two customers arrived in the first minute?

Solution 1: Assume that the customers arrive according to a Poisson process with parameter λ arrivals per minute.

$$P(N(1) = 2 | N(2) = 2) = \frac{P(N(1) = 2, N(2) = 2)}{P(N(2) = 2)}$$

$$= \frac{P(N(1) = 2, N(2-1) = 0)}{P(N(2) = 2)} = \frac{\left(\frac{\lambda^2}{2!}\right) e^{-\lambda} \cdot e^{-\lambda}}{\frac{(2\lambda)^2}{2!} e^{-2\lambda}} = \frac{1}{4}$$

Solution 2:

Use the joint distribution for (T_1, T_2).

Over the region R_2, the conditional PDF is

$$f_{T_1,T_2}(t_1, t_2|N(t) = 2) = \frac{2!}{t^2} = \frac{2!}{2^2} = \frac{1}{2}$$

Hence

$$P(N(1) = 2|N(2) = 2) = \int_{R_2} \frac{1}{2} dt_2 dt_1$$

$$= \int_{t_1=0}^{1} \int_{t_2=t_1}^{1} \frac{1}{2} dt_2 dt_1 = \frac{1}{2} \int_{t_1=0}^{1} (1 - t_1) dt_1 = \frac{1}{2} \left(1 - \frac{t_1^2}{2} |_0^1\right) = \frac{1}{4}$$

Order Statistics

The set of arrival times $T_1, \cdots, T_n$ in the interval $[0, t]$, for a Poisson process, given that $N(t) = n$, can also be generated in a different manner. Consider a set of n independent random variables $U_1, \cdots, U_n$ with uniform PDF $f_U(u) = \frac{1}{t}$ on $[0, t]$ and 0 elsewhere. The joint PDF is given by

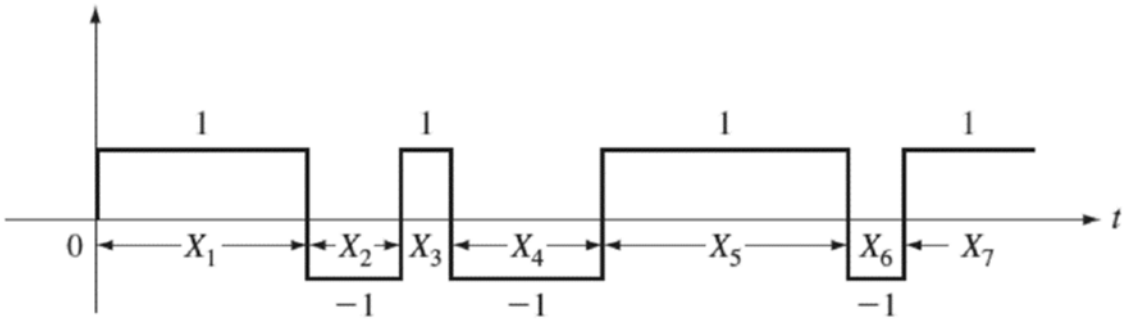
$$f_{\boldsymbol{U}}(u_1, \cdots, u_n) = \frac{1}{t^n}$$

If we now introduce the so-called order statistics, i.e. we define a new random vector $\boldsymbol{V} = (V_1, \cdots, V_n)$, where $\boldsymbol{V}$ is obtained by ordering in U_i in increasing order. Then since for each $(U_1, \cdots, U_n)$ there are $n!$ permutations, $f_{\boldsymbol{V}}(x_1, \cdots, x_n) = n!\, f_{\boldsymbol{U}}(x_1, \cdots, x_n) = n!/t^n$. This is the same joint PDF as for the arrival times $\boldsymbol{T} = (T_1, \cdots, T_n)$ for a Poisson process (given $N(t) = n$) as derived above. In other words for some calculations we may treat the random variables $T_1, \cdots, T_n$ as if they were obtained as n independent random variables from a uniform density in $[0, t]$.

The Random Telegraph Signal

The random telegraph signal is a random process $X(t)$ that takes values ± 1. The process is derived from a Poisson process as follows:

$X(0) = 1$ with probability $1/2$, $X(0) = -1$ with probability $1/2$. Then $X(t)$ changes polarity with each occurring event in a Poisson process with rate λ. The following is a sample function with $X(0) = 1$.



What is the PMF for $X(t)$?

We use the law of total probability as follows:

$$P(X(t)=1)=P(X(t)=1|X(0)=1)\cdot P(X(0)=1)+P(X(t)=1|X(0)=-1)\cdot P(X(0)=-1)$$

Now, here is the key

$P(X(t)=1|X(0)=1)=P(\text{there is an even number of polarity changes in } [0,t])$

$$=\sum_{k=\text{even}}^{\infty}\frac{(\lambda t)^k e^{-\lambda t}}{k!}=e^{-\lambda t}\sum_{k\ \text{even}}^{\infty}\frac{(\lambda t)^k}{k!}$$

Note that

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

Hence $P(X(t)=1|X(0)=1)=e^{-\lambda t}\left(\frac{1}{2}\left(e^{\lambda t}+e^{-\lambda t}\right)\right)=\frac{1}{2}\left(1+e^{-2\lambda t}\right)$

In the same way

$$P(X(t)=1|X(0)=-1)=P(\text{there is an odd number of polarity changes in } [0,t])$$

$$=\frac{1}{2}\left(1-e^{-2\lambda t}\right)$$

Hence $P(X(t)=1)=\frac{1}{2}\left(\frac{1}{2}\left(1+e^{-2\lambda t}\right)\right)+\frac{1}{2}\left(\frac{1}{2}\left(1-e^{-2\lambda t}\right)\right)=\frac{1}{2}$

Similarly we can show that $P(X(t)=-1)=\frac{1}{2}$.

Note that as a result of symmetry none of the above is surprising.

Telegraph Signal Mean and Variance

The mean of the telegraph signal is clearly $\mathcal{E}\big(X(t)\big) = \frac{1}{2} \times 1 + \frac{1}{2} \times (-1) = 0$

The variance is equal to the second moment

$$\mathcal{E}\big(X^2(t)\big) = \mathcal{E}((\pm 1)^2) = \mathcal{E}(1) = 1$$

Telegraph Signal Auto-Covariance

$$C_X(t_1, t_2) = \mathcal{E}(X(t_1)X(t_2)$$

$$= 1 \cdot P\big(X(t_1) = X(t_2)\big) - 1 \cdot P\big(X(t_1) \neq X(t_2)\big)$$

$$= P(\text{even \# transitions in } [t_1, t_2] - P(\text{odd \# transitions in } [t_1, t_2])$$

$$= \frac{1}{2}\big(1 + e^{-\lambda(t_2 - t_1)}\big) - \frac{1}{2}\big(1 - e^{-\lambda(t_2 - t_1)}\big) = e^{-\lambda(t_2 - t_1)}$$

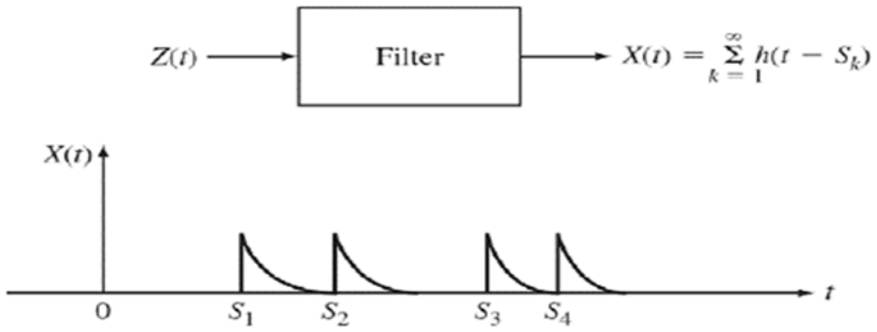
In the above we have assumed $t_1 < t_2$. For arbitrary t_1, t_2 we can write the result as follows:

$$C_X(t_1, t_2) = e^{-\lambda|t_2 - t_1|}$$

Note that $C_X(t_1, t_2) \to 1$ as $|t_2 - t_1| \to 0$, and $C_X(t_1, t_2) \to 0$ as $|t_2 - t_1| \to \infty$. This is a common property of the auto-covariance function for many random processes but not for all.

Shot Noise Process

Shot noise is another process that is generated from a Poisson process. Consider a Poisson process with the points in time being given as $S_n = E_1 + E_2 + \cdots + E_n$, $n = 1,2, \ldots$, where the E_n are exponential random variables representing the inter-arrival times. We may represent the Poisson process as $N(t) = \sum_{k=1}^{\infty} u(t - S_k)$ $(t \geq 0)$, where $u(t)$ is the step function. We may represent $N(t)$ as the output of a filter with impulse response $u(t)$ when the input is $Z(t) = \sum_{k=1}^{\infty} \delta(t - S_k)$. For a general filter with impulse response $h(t)$ the output process is $X(t) = \sum_{k=1}^{\infty} h(t - S_k)$. We refer to this process as shot noise.



Mean of Shot Noise Process

We now compute the mean of the process $X(t)$, i.e. $\mathcal{E}\big(X(t)\big)$. In computing this mean we may first condition the probability distribution for $X(t)$ on the event $N(t) = k$, and then uncondition on the random variable $N(t)$. This amounts to using the law of total probability. Assume that $h(t) = 0$ for $t < 0$, i.e. the filter is causal, then

$$\mathcal{E}(X(t)|N(t) = k) = \mathcal{E}\left(\sum_{i=1}^{\infty} h(t - S_i)\right) = \mathcal{E}\left(\sum_{i=1}^{k} h(t - S_i)\right) = \sum_{i=1}^{k} \mathcal{E}\big(h(t - S_i)\big)$$

Now, as we discussed before we may model the S_i, $i = 1, \cdots, k$, as independent random variables with uniform PDF on $[0, t]$. In this case $\mathcal{E}\big(h(t - S_i)\big) = \int_0^t \frac{1}{t} h(t - s)ds = \frac{1}{t}\int_0^t h(s)ds$

Hence $\mathcal{E}(X(t)|N(t) = k) = \frac{k}{t}\int_0^t h(s)ds$

Now we treat k as a random variable and uncondition to obtain

$$\mathcal{E}\big(X(t)\big) = \mathcal{E}\big(\mathcal{E}(X(t)|N(t) = k)\big) = \mathcal{E}\left(\frac{k}{t}\int_0^t h(s)ds\right) = \frac{\mathcal{E}(k)}{t}\int_0^t h(s)ds = \frac{\lambda t}{t}\int_0^t h(s)ds$$
$$= \lambda \int_0^t h(s)ds$$

Gaussian Random Process

A random process $X(t)$ is said to be a Gaussian random process if for any integer k and any set of points in time $t_1, \cdots, t_k$, the random vector $\boldsymbol{X} = (X_1, \cdots, X_k) = \big(X(t_1), \cdots, X(t_k)\big)$ is a Gaussian random vector as we have defined previously. Let $\boldsymbol{m} = \mathcal{E}(\boldsymbol{X})$ be the mean, and $\boldsymbol{C}$ the co-variance matrix. Then the joint PDF is

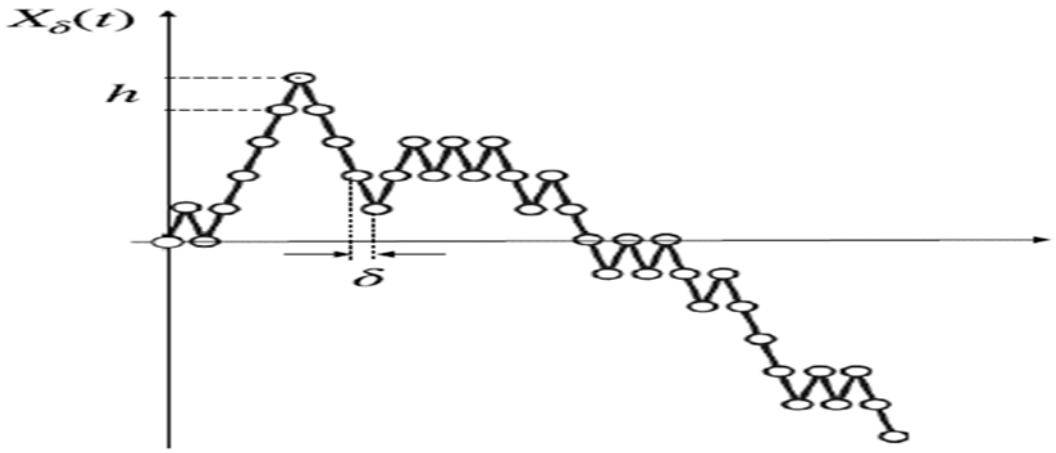
$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T \boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{m})}}{(2\pi)^{\frac{k}{2}}|C|^{\frac{1}{2}}}$$

Where $|\boldsymbol{C}|$ denotes the determinant of the matrix $\boldsymbol{C}$. Note that this holds if the covariance matrix is non-singular. Also note that all the elements of the covariance matrix can be obtained from the auto-covariance function $C_X(t_1, t_2)$. If the covariance matrix is singular then the above PDF needs to be modified. As a result the probability law for a Gaussian random process can be specified completely by the mean function, $m_X(t)$, and the auto-covariance function.

As a result of the Gaussian property, and as we have seen previously, since linear operations on a Gaussian vector produce a Gaussian vector, linear operations on a Gaussian random process result in a Gaussian random process. These operations include scaling by a constant, addition of two Gaussian processes, an integral, or a derivative, result in a Gaussian random process, although we need to define what we mean by the integral or derivative of a Gaussian random process.

The Construction of Brownian Motion

We have previously discussed a random process consisting of Bernoulli trials at discrete points in time $k\Delta t$. We set $X(0) = 0$, and at each point in time we perform a Bernoulli trial, and if success occurs then we increment $X(k\Delta t)$ by $+1$, otherwise we increment by -1. We referred to this process as a random walk. We will now generalize as follows: Let the increment in time be $\Delta t = \delta$, and let the step change be $= \pm h$. We set $X(0) = 0$, and then $X(k\delta) = X((k-1)\delta) \pm h$, according to the trial at the k^{th} step. Note that this is a discrete time random process. To make it a continuous time process we perform linear interpolation between the points $k\delta$, i.e. we connect the points with a straight line. We refer to this process as $X_\delta(t)$. We will let δ and h approach 0. In doing this we obtain a random process where the sample functions are continuous functions of time and at the same time exhibit random behaviour, in fact independent increments, over very short time intervals. A sample function of the process $X_\delta(t)$ is shown in the following Figure.



Consider the times $t = n\delta$. At these times $X_\delta(t) = h(D_1 + D_2 + \cdots + D_n) = hS_n$, where the $D_k \in \{-1,1\}$ indicate the k^{th} jump.

At these points the mean is $\mathcal{E}\big(X(t)\big) = h\mathcal{E}(S_n) = kn\mathcal{E}(D_k) = hn\big(p(1) + (1-p)(-1)\big) = hn(2p-1)$, where we have considered the general case of $P(D_k = 1) = p$, and $P(D_k = -1) = 1 - p$.

The variance of D_k is $\mathcal{E}(D_k^2) - \big(\mathcal{E}(D_k)\big)^2 = 1 - (2p-1)^2 = 4p - 4p^2 = 4p(1-p)$.

Hence $\text{var}\big(X(t)\big) = h^2 n \cdot \text{var}(D_k) = 4nh^2p(1-p)$. For the common case of $p = \frac{1}{2}$, we have $\mathcal{E}\big(X(t)\big) = 0$, and $\text{var}\big(X(t)\big) = nh^2$.

Now, consider $p = 1/2$, and let $\delta \to 0$ and $h \to 0$ in such a manner that $h = \sqrt{\alpha\delta}$.

Note that α is an intensity parameter. Note also that the slope of the line segment between any two adjacent discrete point is $\frac{h}{\delta} = \sqrt{\frac{\alpha}{\delta}}$, i.e. the slope goes to infinity as $\delta \to 0$. We could choose

an h equal to any function of δ. But this choice results in the variance of $X(t)$ to grow proportional to t.

For example we could let $\delta = \frac{1}{n}$, where $n \to \infty$, and $h = \sqrt{\alpha\delta}$. Then $\mathcal{E}\big(X(t)\big) = \lim_{\delta\to 0} \mathcal{E}\big(X_\delta(t)\big) = 0$, and $\text{var}\big(X(t)\big) = \lim_{\delta\to 0} \text{var}\big(X_\delta(t)\big) = \lim_{\delta\to 0}\big(\sqrt{\alpha\delta}\big)^2\left(\frac{t}{\delta}\right) = \alpha t$.

In the limit these results apply to all t.

Gaussian Process as Limit of the Above Brownian Motion

As $\delta \to 0$ in the above, $X(t)$ approaches the sum of an infinite number of independent random variables since $n = \frac{t}{\delta} \to \infty$.

By the Central Limit Theorem the PDF of $X(t)$ approaches the Gaussian PDF with mean 0 and variance $\sigma^2 = \alpha t$.

$$F_{X(t)}(x) = \frac{1}{\sqrt{2\pi\alpha t}} e^{-\frac{x^2}{2\alpha t}}$$

$X(t)$ inherits the property of independent and stationary increments from the random walk process from which it is derived. We can show that for any set of k points $t_1, t_2, \cdots, t_k$, the random vector $(X(t_1), X(t_2), \cdots, X(t_k))$ approaches a Gaussian random vector as $\delta \to 0$. One way to do this is to use the result for processes with independent increments that we have listed previously

$$f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = f_{X_1}(x_1) f_{Y_1}(x_2 - x_1) \cdots f_{Y_{k-1}}(x_k - x_{k-1})$$

Where the X_i are the random variables at the points $X(t_i)$ and $Y_i = X_i - X_{i-1}$, are the increments.

Each of the factors in the above, right side, approaches a Gaussian, because the increments are Gaussian. Then we simplify the expression to obtain a result for the PDF on the left side.

$$\frac{1}{\sqrt{2\pi\alpha t_1}} \exp\left(-\frac{x_1^2}{2\alpha t_1}\right) \frac{1}{\sqrt{2\pi\alpha(t_2 - t_1)}} \exp\left(-\frac{(x_2 - x_1)^2}{2\alpha(t_2 - t_1)}\right) \cdots \frac{1}{\sqrt{2\pi\alpha(t_n - t_{n-1})}} \exp\left(-\frac{(x_n - x_{n-1})^2}{2\alpha(t_n - t_{n-1})}\right)$$

$$= \frac{\exp\left(-\left(\frac{x_1^2}{2\alpha t_1} + \frac{(x_2 - x_1)^2}{2\alpha(t_2 - t_1)} + \cdots + \frac{(x_n - x_{n-1})^2}{2\alpha(t_n - t_{n-1})}\right)\right)}{(2\pi\alpha)^{\frac{n}{2}}\sqrt{t_1(t_2 - t_1)\cdots(t_n - t_{n-1})}}$$

Let $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$, and $\boldsymbol{y} = (x_1, x_2 - x_1, \cdots, x_n - x_{n-1})$. Then the argument of the exponential function can be written as $\frac{-1}{2\alpha}\boldsymbol{y}^T\Lambda^{-1}\boldsymbol{y}$, where Λ is a diagonal matrix with the i^{th} element being equal to $t_i - t_{i-1}$, $i = 1,2, \cdots, n$, and where we set $t_0 = 0$. Now, we can write $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{A}$ is the following matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \dots & 0 \\ -1 & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & -1 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & -1 & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 & -1 & 1 \end{bmatrix}$$

The exponent then becomes $\frac{-1}{2\alpha}\boldsymbol{y}^T\Lambda^{-1}\boldsymbol{y} = \frac{-1}{2\alpha}(\boldsymbol{Ax})^T\Lambda^{-1}\boldsymbol{Ax} = \frac{-1}{2\alpha}\boldsymbol{x}^T\boldsymbol{A}^T\Lambda^{-1}\boldsymbol{Ax} = -\frac{\mathbf{1}}{\mathbf{2}\boldsymbol{\alpha}}\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{C}^{-\mathbf{1}}\boldsymbol{x}$, where $\boldsymbol{C} = \boldsymbol{A}^{-\mathbf{1}}\boldsymbol{\Lambda}(\boldsymbol{A}^{\boldsymbol{T}})^{-\mathbf{1}}$.

Note that $\det(C) = \det(A^{-1})\det(\Lambda)\det(A^{T^{-1}}) = \det(\Lambda)$, because $\det(A) = \det(A^{-1}) = \det A^T = 1$.
With all the above we obtain

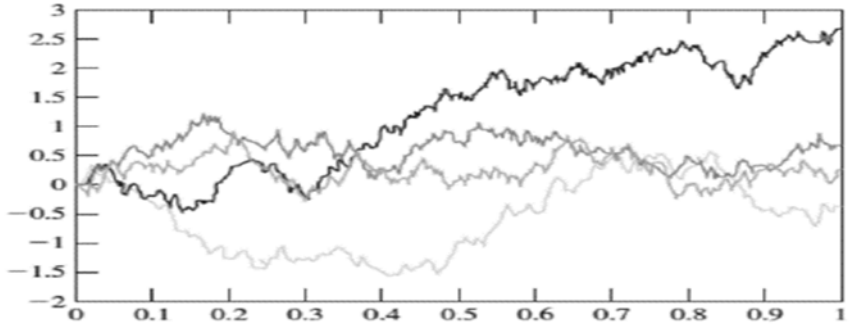
$$f_{X_1,\cdots,X_n}(x_1,\cdots,x_n) = \frac{\exp\left(-\frac{1}{2\alpha}\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{C}^{-\mathbf{1}}\boldsymbol{x}\right)}{(2\pi\alpha)^{\frac{n}{2}}\sqrt{\det C}}$$

Hence the random vector $\boldsymbol{x}$ is Gaussian.

The Wiener Random Process

The continuous time process, $X(t)$, above, is also called the Wiener random process. It has the property that $X(0) = 0$, zero mean for all time, and the variance increases with time linearly.

The Wiener process is used to model Brownian motion, the motion of particles suspended in a fluid that move under the rapid and random impact of neighbouring particles. Sample functions of the Wiener process are shown in the Figure below.



Using the independent increments property, as we did for the Poisson process, we can show that the auto-covariance for the Wiener process is

$$C_X(t_1,t_2) = \alpha\min(t_1,t_2)$$

The Wiener process and the Poisson process have the same auto-covariance function, but the sample functions are very different.

Summary of the Properties of the Wiener Process

- $X(0) = 0$
- $\mathcal{E}\big(X(t)\big) = 0$
- $\text{var}\big(X(t)\big) = \alpha t$
- $C_X(t_1, t_2) = \alpha \min(t_1, t_2)$
- Gaussian process
- Independent and stationary increments
- Continuous sample functions
- Not differentiable (because at each step in the limiting process there are “corners” in the sample path at the points $k\delta$)
- Integral of White Noise (more on this later)