ECE 5510: Random Processes
Lecture Notes
Fall 2009

Dr. Neal Patwari
University of Utah
Department of Electrical and Computer Engineering

Contents

Lecture 1

Today: (1) Syllabus, (2) Course Overview, (3) Application Assignment Intro

1 Course Overview

Randomness is all around us; in many engineering and scientific applications.

- communications,
- controls,
- manufacturing,
- economics,
- imaging,
- biology,
- the Internet,
- systems engineering.

In all of these applications, we have what we call random variables. These are things which vary across time in an unpredictable manner.

Sometimes, these things we truly could never determine beforehand. For example, thermal noise in a receiver is truly unpredictable.

In other cases, perhaps we could have determined if we had taken the effort. For example, whether or not a machine is going to fail today could have been determined by a maintenance checkup at the start of the day. But in this case, if we do not perform this checkup, we can consider whether or not the machine fails today as a random variable, simply because it appears random to us.

The study of probability is all about taking random variables and quantifying what can be known about them. Probability is a set of tools which take random variables and output deterministic numbers which answer particular questions. So while the underlying variable or process may be random, we as engineers are able to 'measure' them.

For example:

- The expected value is a tool which tells us, if we observed lots of realizations of the random variable, what its average value would be.
- Probability of the random variable being in an interval or set of values quantifies how often we should expect the random value to fall in that interval or set.
- The variance is a tool which tells us how much we should expect it to vary.
- The correlation or covariance (between two random variables) tells us how closely two random variables follow each other.

The study of random processes is simply the study of random variables sequenced by continuous or discrete time (or space), which represent the temporal (or spatial) variation of a random variable.

This class is all about quantifying *what can be known* about random variables and random processes.

We will discuss in class this outline of the topics covered in this course.

1. Review of probability for individual random variables
   (a) Probabilities on sets, Bayes' Law, independence
   (b) Distributions: pdfs, CDFs, conditional pdfs
   (c) Continuous vs. discrete-valued random variables
   (d) Expectation and moments
   (e) Transformation of (functions of) random variables
2. Joint probability for multiple random variables and sequences of random variables
   (a) Random vectors
   (b) Joint distributions
   (c) Expectation and moments (covariance and correlation)
   (d) Transformation of (functions of) random vectors
3. Random process models
   (a) Bernoulli processes
   (b) Poisson processes
   (c) Gaussian processes
   (d) Markov chains
4. Spectral analysis of random processes
   (a) Filtering of random processes
   (b) Power spectral density
   (c) Continuous vs. discrete time

Lecture 2

Today: (1) Events as Sets (2) Axioms and Properties of Probability (3) Independence of Sets

2 Events as Sets

All probability is defined on sets. In probability, we call these sets *events*. A set is a collection of elements. In probability, we call these *outcomes*.

Def'n: *Event*
A collection of outcomes. Order doesn't matter, and there are no duplicates.

2.0.1 Set Terminology vs. Probability Terminology

Set Theory	Probability Theory	Probability Symbol
universe	sample space (certain event)	S
element	outcome (sample point)	s
set	event	E
disjoint sets	disjoint events	$E_1 \cap E_2 = \emptyset$
null set	null event	$\emptyset$

2.1 Introduction

There are different ways to define an event (set):

- List them: $A = \{0, 5, 10, 15, \ldots\}$; $B = \{Tails, Heads\}$
- As an interval: $[0, 1]$, $[0, 1)$, $(0, 1)$, $(a, b]$. Note overlap with coordinates!
- An existing event set name: $\mathbb{N}$, $\mathbb{R}^2$, $\mathbb{R}^n$
- By rule: $C = \{x \in \mathbb{R} : x \geq 0\}$, $D = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < R^2\}$. Note Y&G uses '|' instead of the colon ':', which I find confusing.

2.1.1 Important Events

Here's an important event: $\emptyset = \{\}$, the *null event* or the *empty set*.

Here's the opposite: S is used to represent the set of everything possible in a given context, the *sample space*.

- $S = B$ above for the flip of a coin.
- $S = \{1, 2, 3, 4, 5, 6\}$ for the roll of a (6-sided) die.
- $S = \{Adenine, Cytosine, Guanine, Thymine\}$ for the nucleotide found at a particular place in a strand of DNA.
- $S = C$, *i.e.*, non-negative real numbers, for your driving speed (maybe when the cop pulls you over).

2.2 Finite, Countable, and Uncountable Event Sets

We denote the size of, *i.e.*, the number of items in, a set A as $|A|$. If $|A|$ is less than infinity then set A is said to be *finite*. But there are two kinds of infinite sets:

Countably Infinite: The set can be listed. That is, each element could be assigned a unique positive integer. Eg. $\{1, 2, 3, \ldots\}$, or $\{\frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, \ldots\}$, set A above. Even the set $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$ is countably infinite. Easy way: They can be seen as discrete points on the real line.

Uncountably Infinite: There's no way to list the elements. They fill the real line. Eg., $\mathbb{R}$, or any interval of the real line, $[a, b]$ for $b > a$.

Finite and countably infinite sample spaces are called *discrete* or *countable*, respectively; while uncountably infinite sample spaces are *continuous*. This is the difference we'll see for the rest of the semester between discrete and continuous random variables.

2.3 Operating on Events

We can operate on one or more events:

- Complement: $A^c = \{x \in S : x \notin A\}$. We must know the sample space S!
- Union: $A \cup B = \{x : x \in A \,\text{or}\, x \in B\}$. Merges two events together.
- Intersection: $A \cap B = \{x : x \in A \,\text{and}\, x \in B\}$. Limits to outcomes in both events.

Also: $A - B = A \cap B^c$, *i.e.*, the outcomes in A that are not in B.
Note: Venn diagrams are *great* for intuition: however, you *cannot* use them to prove anything!
Some more properties of events and their operators:

$$\begin{aligned}
A \cup B &= B \cup A \\
(A^c)^c &= A \\
A \cup S &= S \\
A \cap S &= A \\
A \cap A^c &= \emptyset \\
A \cup (B \cup C) &= (A \cup B) \cup C \\
A \cap (B \cup C) &= (A \cap B) \cup (A \cap C) \\
A \cup (B \cap C) &= (A \cup B) \cap (A \cup C) \\
(A \cap B) \cup (C \cap D) &= (A \cup C) \cap (A \cup D) \cap (B \cup C) \cap (B \cup D) \\
(A \cup B) \cap (C \cup D) &= (A \cap C) \cup (A \cap D) \cup (B \cap C) \cup (B \cap D)
\end{aligned} \tag{1}$$

These last four lines say that you can essentially "multiply out" or do "FOIL" by imagining one of the union/intersection to be multiplication and the other to be addition. Don't tell people that's how you're doing it, just do it.

Example: Prove $A \cup B = (A \cap B) \cup (A \cap B^c) \cup (A^c \cap B)$. By the way, this is a relation required later for a proof of a formula for the probability of a union of two sets.

Solution: Working on the RHS,

$$\begin{aligned}
A \cup B &= [(A \cap B) \cup (A \cap B^c)] \cup (A^c \cap B) \\
&= [A \cap (B \cup B^c)] \cup (A^c \cap B) \\
&= [A \cap S] \cup (A^c \cap B) \\
&= A \cup (A^c \cap B) \\
&= (A \cup A^c) \cap (A \cup B) \\
&= S \cap (A \cup B) \\
&= (A \cup B)
\end{aligned} \tag{2}$$

We use notation to save writing:

$$\begin{aligned}\bigcup_{i=1}^{n} A_i &= A_1 \cup A_2 \cup \cdots \cup A_n \\ \bigcap_{i=1}^{n} A_i &= A_1 \cap A_2 \cap \cdots \cap A_n\end{aligned}$$

DO NOT use addition to represent the union, and DO NOT use multiplication to represent the intersection. An example of why this is confusing:

$$\{1\} + \{1\} = \{1\}.$$

But remember when you see it in Y&G, that $P[AB] = P[A \cap B]$.

This leads to one of the most common written mistakes – exchanging unions and plusses when calculating probabilities. Don't write $P[A] + P[B]$ when you really mean $P[A \cup B]$. Don't add sets and numbers: for example, if A and B are sets, don't write $P[A] + B$.

2.4 Disjoint Sets

The words from today we'll use most often in this course are *disjoint* and *mutually exclusive*:

- Two events A_1 and A_2 are disjoint if $A_1 \cap A_2 = \emptyset$.
- A collection of events $A_1, \ldots, A_n$ are mutually exclusive if for all pairs $i, j \in \{1, \ldots n\}$ (for which $i \neq j$), $A_i \cap A_j = \emptyset$. That is, every pair is disjoint.

Some disjoint events: A and A^c; A and $\emptyset$. Proof?

3 Axioms and Properties of Probability

You're familiar with functions, like $f(x) = x^2$, which assign a number output to each number input. Probability assigns a number output to each event input. Here's how it does that.

- 0. Define an experiment. Eg., measure the nucleotide at a particular spot on a DNA molecule.
- 1. List each possible outcome of that experiment. This list is the *sample space* S. Eg., for DNA, the nucleotide must be $S = \{a, c, g, t\}$.
- 2. An *event* E is defined as any subset of S. It is anything we might be interested in knowing the probability of. An event $E \in \mathcal{F}$.
  Def'n: *Field of events*
  The field of events, $\mathcal{F}$, is a list (set) of all events for which we could possibly calculate the probability.

  Eg, for the above S, the field of events is

$$\begin{aligned}\mathcal{F} = \{&\emptyset, \{a\}, \{c\}, \{g\}, \{t\}, \{a, c\}, \{a, g\}, \{a, t\}, \{c, g\}, \{c, t\}, \{g, t\}, \\ &\{a, c, g\}, \{a, c, t\}, \{a, g, t\}, \{c, g, t\}, S\}\end{aligned}$$

- 3. Define the probability $\mathcal{P}$ of each event.

3.1 How to Assign Probabilities to Events

As long as we follow three intuitive rules (*axioms*) our assignment can be called a *probability model*.

Axiom 1: For any event A, $P[A] \geq 0$.

Axiom 2: $P[S] = 1$.

Axiom 3: For any countable collection $A_1, A_2, \ldots$ of mutually exclusive events,

$$P\left[\bigcup_{i=1}^{\infty} A_i\right] = P[A_1] + P[A_2] + \cdots .$$

This last one is really more complicated than it needs to be at our level. Many books just state it as:
Axiom 3 in other books: $P[E \cup F] = P[E] + P[F]$ for disjoint events E and F.
This could be extended to the Y&G Axiom 3, but you need more details for this proof [Billingsley 1986].

Example: DNA Measurement

Consider the DNA experiment above. We measure from a strand of DNA its first nucleotide. Let's assume that each nucleotide is equally likely. Using axiom 3,

$$P[\{a, c, g, t\}] = P[\{a\}] + P[\{c\}] + P[\{g\}] + P[\{t\}]$$

But since $P[\{a, c, g, t\}] = P[S]$, by Axiom 2, the LHS is equal to 1. Also, we have assumed that each nucleotide is equally likely, so

$$1 = 4P[\{a\}]$$

So $P[\{a\}] = 1/4$.

Def'n: *Discrete Uniform Probability Law*
In general, for event A in a discrete sample space S composed of equally likely outcomes,

$$P[A] = \frac{|A|}{|S|}$$

3.2 Other Properties of Probability Models

1. $P[A^c] = 1 - P[A]$. Proof:
   First, note that $A \cup A^c = S$ from above. Thus

   $$P[A \cup A^c] = P[S]$$

   Since $A \cap A^c = \emptyset$ from above, these two events are disjoint.

   $$P[A] + P[A^c] = P[S]$$

   Finally from Axiom 2,

   $$P[A] + P[A^c] = 1$$

And we have proven what was given.

Note that this implies that $P[S^c] = 1 - P[S]$, and from axiom 2, $P[\emptyset] = 1 - 1 = 0$.

2. For any events E and F (not necessarily disjoint),

$$P[E \cup F] = P[E] + P[F] - P[E \cap F]$$

Essentially, by adding $P[E] + P[F]$ we double-count the area of overlap. The $-P[E \cap F]$ term corrects for this. Proof: Do on your own using these four steps:

(a) Show $P[A] = P[A \cap B] + P[A \cap B^c]$.

(b) Same thing but exchange A and B.

(c) Show $P[A \cup B] = P[A \cap B] + P[A \cap B^c] + P[A^c \cap B]$.

(d) Combine and cancel.

3. If $A \subset B$, then $P[A] \leq P[B]$. Proof:
Let $B = (A \cap B) \cup (A^c \cap B)$. These two events are disjoint since $A \cap B \cap A^c \cap B = \emptyset \cap B = \emptyset$. Thus:

$$P[B] = P[A \cap B] + P[A^c \cap B] = P[A] + P[A^c \cap B] \geq P[A]$$

Note $P[A \cap B] = P[A]$ since $A \subset B$, and the inequality in the final step is due to the Axiom 1.

4. $P[A \cup B] \leq P[A] + P[B]$. Proof: directly from 2 and Axiom 1. This is called the *Union Bound*. **On your own:** Find intersection bounds for $P[A \cap B]$ from 2.

3.3 Independence

Def'n: *Independence of a Pair of Sets*
Sets A and B are independent if $P[A \cap B] = P[A]\,P[B]$.

Example: Set independence
Consider $S = \{1, 2, 3, 4\}$ with equal probability, and events $A = \{1, 2\}$, $B = \{1, 3\}$, $C = \{4\}$.

1. Are A and B independent? $P[A \cap B] = P[\{1\}] = 1/4 = (1/2)(1/2)$. Yep.

2. Are B and C independent? $P[B \cap C] = P[\emptyset] = 0 \neq (1/2)(1/4)$. Nope.

Lecture 3

Today: (1) Conditional Probability, (2) Trees, (3) Total Probability

4 Conditional Probability

Example: Three Card Monte

(Credited to Prof. Andrew Yagle, U. of Michigan.)

There are three two-sided cards: red/red, red/yellow, yellow/yellow. The cards are mixed up and shuffled, one is selected at random, and you look at one side of that card. You see red. **What is the probability that the other side is red?**

Three possible lines of reasoning on this:

1. Bottom card is red only if you chose the red/red card: $P = 1/3$.
2. You didn't pick the yellow/yellow card, so either the red/red card or the red/yellow card: $P = 1/2$.
3. There are five sides which we can't see, two red and three yellow: $P = 2/5$.

Which is correct?

Def'n: *Conditional Probability,* $P[A|B]$
P [event A occurs, GIVEN THAT event B occurred]

For events A and B, when $P[B] > 0$,

$$P[A|B] \triangleq \frac{P[A \cap B]}{P[B]} = \frac{P[A \cap B]}{P[A \cap B] + P[A^c \cap B]}$$

Notes:

1. Given that B occurs, now we know that either $A \cap B$ occurs, or $A^c \cap B$ occurs.
2. We're defining a new probability model, knowing more about the world. Instead of $P[\cdot]$, we call this model $P[\cdot|B]$. All of our Axioms STILL APPLY!
3. NOT TO BE SAID OUT LOUD because its not mathematically true in any sense. But you can remember which probability to put on the bottom, by thinking of the | as / – you know what to put in the denominator when you do division.

Note: Conditional probability always has the form, $\frac{x}{x+y}$. If $P[A|B] = \frac{x}{x+y}$ then $P[A^c|B] = \frac{y}{x+y}$. Note the two terms add to one.

Example: Three Card Monte
BR = Bottom red; TR = Top red; BY = Bottom yellow; TY = Top yellow.

$$\begin{aligned} P[BR\,|\,TR] &= \frac{P[BR \text{ and } TR]}{P[TR]} \\ &= \frac{P[BR \text{ and } TR]}{P[BR \text{ and } TR] + P[BY \text{ and } TR]} \\ &= \frac{2/6}{2/6 + 1/6} = \frac{1/3}{1/2} = 2/3. \end{aligned}$$

4.1 Conditional Probability *is* Probability

For an event B with positive probability, $P[B] > 0$, the conditional probability defined above is a valid probability law.

Proof: It must obey the three axioms. Let $A \in \mathcal{F}$,

1. Since $P[A \cap B] \geq 0$ by axiom 1, and $P[B] > 0$, then
$$P[A|B] = \frac{P[A \cap B]}{P[B]} \geq 0.$$

2. For the sample space S,
$$P[S|B] = \frac{P[S \cap B]}{P[B]} = \frac{P[B]}{P[B]} = 1.$$

3. If events $A_1, A_2, \ldots$ are disjoint,
$$\begin{aligned} P\left[\bigcup_{i=1}^{\infty} A_i|B\right] &= \frac{P[(\bigcup_{i=1}^{\infty} A_i) \cap B]}{P[B]} = \frac{P[\bigcup_{i=1}^{\infty} (A_i \cap B)]}{P[B]} \\ &= \frac{\sum_{i=1}^{\infty} P[A_i \cap B]}{P[B]} = \sum_{i=1}^{\infty} P[A_i|B]. \end{aligned}$$

4.2 Conditional Probability and Independence

We know that for any two sets A and B, that $P[A \cap B] = P[A|B]\,P[B]$. Recall that independent sets have the property $P[A \cap B] = P[A]\,P[B]$. So, independent sets also have the property that

$$\begin{aligned} P[A|B]\,P[B] &= P[A]\,P[B] \\ P[A|B] &= P[A] \end{aligned}$$

Thus the following are equivalent (t.f.a.e.):

1. $P[A \cap B] = P[A]\,P[B]$,
2. $P[A|B] = P[A]$, and
3. $P[B|A] = P[B]$,

If one is true, all of them are true. If one is false, all are false.

4.3 Bayes' Rule

We know that $P[A|B] = \frac{P[A \cap B]}{P[B]}$ so it is also clear that $P[A|B]\,P[B] = P[A \cap B]$. But equivalently, $P[A \cap B] = P[B|A]\,P[A]$ So

Def'n: *Bayes' Rule*

$$P[A|B] = \frac{P[B|A]\,P[A]}{P[B]}$$

Not a whole lot different from definition of Conditional Probability. But, it does say explicitly how to get from $P[B|A]$ to $P[A|B]$.

Example: Neural Impulse Actuation
A embedded sensor is used to monitor a neuron in a human brain. We monitor the sensor reading for 100 ms and see if there was a spike within the period. If the person thinks of flexing his knee, we will see a spike with probability 0.9. If the person is not thinking of flexing his knee (due to background noise), we will see a spike with probability 0.01. For the average person, the probability of thinking of flexing a knee is 0.001 within a given period.

1. What is the probability that we will measure a spike? Answer: Let S be the event that a spike is measured, and its complement as NS. Let the event that the person is thinking about flexing be event T, and its complement is event NT.

$$\begin{aligned} P[S] &= P[S|T]P[T] + P[S|NT]P[NT] \\ &= 0.9 * 0.001 + 0.01 * 0.999 = 0.0009 + 0.00999 = 0.01089 \end{aligned}$$

2. What is the probability that the person wants to flex his knee, given that a spike was measured? Answer:

$$\begin{aligned} P[T|S] &= \frac{P[T \cap S]}{P[S]} = \frac{P[S|T]P[T]}{P[S]} \\ &= \frac{0.9 * 0.001}{0.01089} = 0.0826 \end{aligned}$$

3. What is the probability that the person wants to flex his knee, given that no spike was measured? Answer:

$$\begin{aligned} P[T|NS] &= \frac{P[T \cap NS]}{P[NS]} = \frac{P[NS|T]P[T]}{P[S]} \\ &= \frac{0.1 * 0.001}{1 - 0.01089} \approx 1.01 \cdot 10^{-4} \end{aligned}$$

It wouldn't be a good idea to create a system that sends a signal to flex his knee, given that a spike was measured. Some other system design should be considered.

Def'n: *a priori*
prior to observation

For example, prior to observation, we know $P[T] = 0.001$, and $P[NT] = 1 - 0.001 = 0.999$.

Def'n: *a posteriori*
after observation

For example, after observation, we know $P[T|S] = 0.0826$, and $P[NT|NS] = 1 - 0.0001 = 0.9999$.

4.4 Trees

A graphical method for organizing prior, posterior information. See Figure 1.

Example: Two fair coins, one biased coin
(From Rong-Rong Chen) There are three coins in a box. One is a two-headed coin, another is a

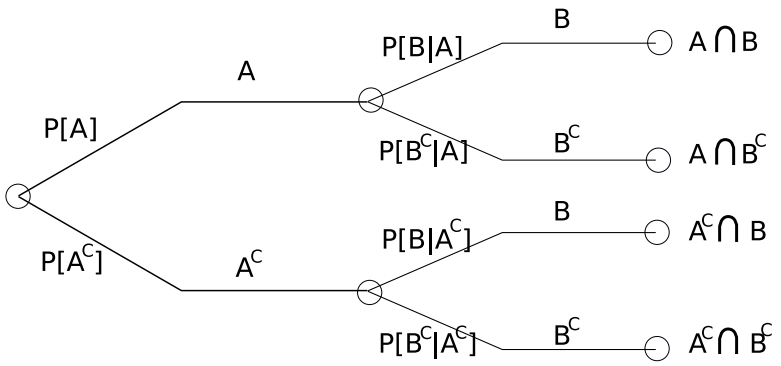


Figure 1: Tree.

fair coin, and the third is a biased coin that comes up heads 75 percent of the time. When one of the three coins is selected at random and flipped, it shows heads. What is the probability that it was the two-headed coin?

Answer:

$$\begin{aligned} P[C_2|H] &= \frac{P[H|C_2]\,P[C_2]}{P[H|C_2]\,P[C_2] + P[H|C_f]\,P[C_f] + P[H|C_b]\,P[C_b]} \\ &= \frac{1/3}{1/3 + 1/2(1/3) + 3/4(1/3)} = 4/9 \end{aligned}$$

5 Partitions and Total Probability

Def'n: *Partition*
A countable collection of mutually exclusive events $C_1, C_2, \ldots$ is a partition if $\bigcup_{i=1}^{\infty} C_i = S$.

Examples:

1. For any set C, the collection C, C^c.
2. The collection of all simple events for countable sample spaces. Eg., {1}, {2}, {3}, {4}, {5}, {6}.

As we have seen, we use a partition $C_1, C_2, \ldots$ to separate $P[A]$ into smaller parts:

$$P[A] = \sum_{i=1}^{\infty} P[A \cap C_1]$$

From the definition of the conditional probability, we have that $P[A \cap C_i] = P[A|C_i]\,P[C_i]$, so we have the **Law of Total Probability**:

$$P[A] = \sum_{i=1}^{\infty} P[A|C_i]\,P[C_i]$$

Note: You must use a partition!!!

Lecture 4

Today: (1) Combinations and Permutations, (2) Discrete random variables

6 Combinations

Example: What is the probability that two people in this room will have the same birthday?
Assume that each day of the year is equally likely, and that each person's birthday is independent.

1. How many ways are there for n people to have their birthday? Answer: Each one is chosen independently, assume 365 days per year. So: 365^n.
2. How many ways are there to have all n people have **unique** birthdays? The first one can happen in 365 ways, the second has 364 left, and so on: ${}_{365}P_n = 365!/(365-n)!$.
3. Discrete uniform probability law:

$$P\left[\exists \text{no duplicate birthdays}\right] = \frac{365!/(365-n)!}{365^n}$$

4. See Fig. 2.

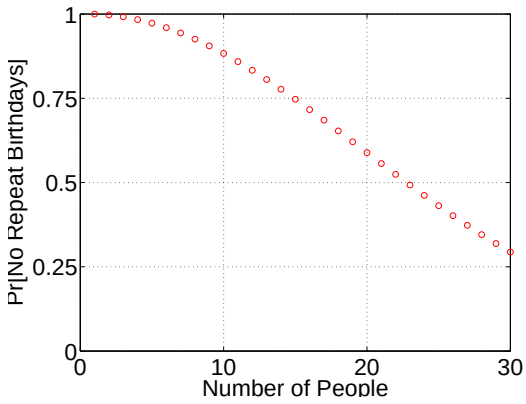


Figure 2: Tree.

Example: Poker (five card stud)
Evaluate the probabilities of being dealt poker hands. The standard deck has 52 cards, 13 cards of each suit; and there are four suits (hearts, diamonds, clubs, and spades). The thirteen cards are, in order, A, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K. The ace (A) can also be higher than the king (K).

1. How many different hands are there? A: 52 choose 5, or 2,598,960.

2. $P\left[StraightFlush\right]$? (a straight flush consists of five cards of the same suit, in a row: (4,5,6,7,8), all hearts, for example.) A: Starting at the (A, 2, 3, 4, 5) hand through the (10, J, Q, K, A) hand, there are 10 straight flushes in each suit. So $P\left[straightflush\right] = \frac{40}{2{,}598{,}960} \approx 1.5 \times 10^{-5}$

3. $P\left[Flush\right]$? (A flush is any 5 cards of the same suit, not including any straight flushes.) There are 13 of each suit, so 13 choose 5, and four suits, so $\binom{13}{5}4 - 40 = 5148 - 40 = 5108$ ways to have a flush. $P\left[Flush\right] = \frac{5{,}108}{2{,}598{,}960} \approx 0.0020$.

4. $P\left[Straight\right]$? (A straight is any five cards of any suits in a row, not including the straight flushes) Again, there are 10 possible sequences. For each card in each sequence, its suit can be chosen in 4 ways. So, there are $10 * 4^5 - 40 = 10240 - 40 = 10200$ ways to have a straight, so $P\left[Straight\right] = \frac{10{,}200}{2{,}598{,}960} \approx 0.0039$.

5. $P\left[OnePair\right]$? Suppose we have 2 K's. The K's can be chosen in $\binom{4}{2}$ ways = 6 ways. The face value of the pair can be chosen in 13 ways. Now we need to choose 3 more cards, none of which match each other. So we choose 3 cards out of the 12 remaining values, which is $\binom{12}{3=220}$ ways. But also each card can be 1 of the 4 suits, and so there are 4*4*4 ways to choose the suits. Total number of ways is 6 x 13 x 220 x 4 x 4 x 4 = 1,098,240. So $P\left[OnePair\right] = \frac{1{,}098{,}240}{2{,}598{,}960} \approx 0.42$.

Note these are the probabilities at the deal, not after any exchange or draw of additional cards.

7 Discrete Random Variables

Def'n: *Random Variable*
A random variable is a mapping from sample space S to the real numbers. In symbols, X is a random variable if $X : S \to \mathbb{R}$.

In other words, if the sample space is not already numeric, then a real number is assigned to each outcome. For example,

- For a coin-flipping experiment with $S = \{Heads, Tails\}$ we might assign $X(Heads) = 1$ and $X(Tails) = 0$.
- For an experiment involving a student's grade, $S = \{F, D, C, B, A\}$, we might assign $X(F) = 0$, $X(D) = 1$, $X(C) = 2$, $X(B) = 3$, $X(A) = 4$.

For outcomes that already have numbers associated with them, *i.e.*, temperature, voltage, or the number of a die roll, we can just use those numbers as the random variable.

We write S_X to indicate the set of values which X may take: $S_X = \{x : X(s) = x, \forall s \in S\}$. This is the 'range of X'.

Def'n: *Discrete Random Variable*
X is a discrete random variable if S_X is a countable set.

7.1 Probability Mass Function

Def'n: *probability mass function (pmf)*
The probability mass function (pmf) of the discrete random variable X is $P_X(x) = P[X = x]$.

Note the use of capital X to represent the random variable name, and lowercase x to represent a particular value that it may take (a dummy variable). Eg., we may have two different random variables R and X, and we might use $P_R(u)$ and $P_X(u)$ for both of them.

Example: Die Roll
Let Y be the sum of the roll of two dice. What is $P_Y(y)$?

Die 1 \ Die 2	**1**	**2**	**3**	**4**	**5**	**6**
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

Noting the numbers of rolls which sum to each number, 2 through 12, we have:

$$P_Y(y) = \begin{cases} 1/36 & \text{if } y = 2, 12 \\ 2/36 & \text{if } y = 3, 11 \\ 3/36 & \text{if } y = 4, 10 \\ 4/36 & \text{if } y = 5, 9 \\ 5/36 & \text{if } y = 6, 8 \\ 6/36 & \text{if } y = 7 \\ 0 & \text{o.w.} \end{cases}$$

Check: What is $\sum_y P_Y(y)$? $\frac{2(1+2+3+4+5)+6}{36} = \frac{2(15)+6}{36} = 1$.

- Always put, 'zero otherwise'.
- Always check your answer to make sure the pmf sums to 1.

Def'n: *Bernoulli Random Variable*
A r.v. X is Bernoulli (with parameter p) if its pmf has the form,

$$P_X(x) = \begin{cases} 1-p & \text{if } x = 0 \\ p & \text{if } x = 1 \\ 0 & \text{o.w.} \end{cases}$$

This is the most basic, and most common pmf! Experiments are often binary. Eg., success/failure, in range / out of range, disease / no disease, packet or bit error / no error, etc.

Def'n: *Geometric Random Variable*
A r.v. X is Geometric (with parameter p) if its pmf has the form,

$$P_X(x) = \begin{cases} p(1-p)^{x-1} & \text{if } x = 1, 2, \ldots \\ 0 & \text{o.w.} \end{cases}$$

This pmf derives from the following Bernoulli random process: repeatedly measure independent bernoulli random variables until you measure a success (and then stop). Then, X is the number of measurements (including the final success). Graphic: the tree drawn in Y&G for Example 2.11.

Def'n: *Zipf r.v.*
A r.v. R with parameters s and $|S_R| = N$ is Zipf if it has the p.m.f.

$$P_R(r) = \frac{1/r^s}{H_{N,s}}$$

where $H_{N,s} = \sum_{n=1}^{N} \frac{1}{r^s}$ is a normalization constant (known as the Nth generalized Harmonic number).

This is also called a 'power law' p.m.f.

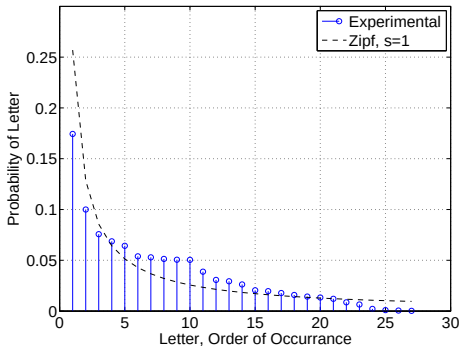


Figure 3: The experimental pmf of letters (and spaces) in Shakespeare's "Romeo and Juliet", compared to the Zipf pmf with $s = 1$.

Example: Zipf's Law
Given a finite sample space S of English words, rank each outcome by its probability. (Pick a random word in a random book.) Define r.v. $R(s)$ to be the rank of word $s \in S$ such that $R(s_1) = 1$ for the outcome s_1 which has maximizes $P[\{s_1\}]$, $R(s_2) = 2$ for the outcome $s_2 \neq s_1$ which has maximizes $P[\{s_2 \neq s_1\}]$, etc.

Often, in real life, the p.m.f. of R has the Zipf p.m.f. The word 'the' is said to appear 7% of the time, and 'of' appears 3.5% of the time. Half of words have probability less than 10^{-6}.

Also:

- web sites (ordered by page views)
- web sites (ordered by hypertext links)
- blogs (ordered by hypertext links)
- city of residence (ordered by size of city)
- given names (ordered by popularity)

7.2 Cumulative Distribution Function (CDF)

Def'n: *Cumulative Distribution Function (CDF)*
The CDF, $F_X(x)$ is defined as $F_X(x) = P[\{X : X \leq x\}]$.

Properties of the CDF:

1. $F_X(x) = \sum_{\{u \in S_X : u \leq x\}} P_X(x)$.
2. $\lim_{x \to +\infty} F_X(x) = 1$, and $\lim_{x \to -\infty} F_X(x) = 0$.
3. For all $b \geq a$, $F_X(b) \geq F_X(a)$ (*the CDF is non-decreasing*) Specifically, $F_X(b) - F_X(a) = P[\{a < X \leq b\}]$.

Example: Sum of Two Dice
Plot the CDF $F_Y(y)$ of the sum of two dice. A: Figure 4.

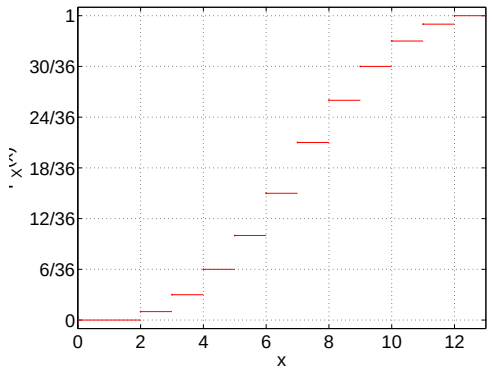


Figure 4: CDF for the sum of two dice.

Example: Geometric r.v. (Time to Success)
What is the CDF of the Geometric r.v.? If u is a positive integer,

$$F_X(u) = \sum_{x=1}^{u} p(1-p)^{x-1} = p\sum_{x=0}^{u-1}(1-p)^x = p\frac{1-(1-p)^u}{1-(1-p)} = 1-(1-p)^u$$

In general,

$$F_X(u) = \begin{cases} 0 & \text{if } x < 1 \\ 1-(1-p)^{\lfloor u \rfloor} & \text{if } x \geq 1 \end{cases}$$

which has limits: $F_X(-\infty) = 0$, $\lim_{u \to +\infty} F_X(u) = 1-(1-p)^u = 1$.

Lecture 5

Today: (1) Expectation, (2) Families of discrete r.v.s

7.3 Recap of Critical Material

- A random variable (r.v.) is a mapping, $X : S \to \mathbb{R}$, where the range is $S_X \subset \mathbb{R}$.
- For a discrete r.v. the range S_X is countable.
- pmf: $P_X(x) = P\left[\{s \in S : X(s) = x\}\right] = P\left[X(s) = x\right] = P\left[X = x\right]$.
- CDF, $F_X(x)$ is defined as $F_X(x) = P\left[\{X : X \leq x\}\right]$.

7.4 Expectation

Section 2.5 of Y&G.

Def'n: *Expected Value*
The expected value of discrete r.v. X is

$$E_X\left[X\right] = \sum_{x \in S_X} x P_X(x)$$

$E_X\left[X\right]$ is also referred to as μ_X. It is a parameter which describes the 'center of mass' of the probability *mass* function.

Note: Y&G uses $E[X]$ to denote expected value. This is somewhat ambiguous, as we will see later. The first (subscript) X refers to the pmf we'll be using, the second X refers to what to put before the pmf in the summation.

Example: Bernoulli r.v. Expectation
What is the $E_X\left[X\right]$, a Bernoulli r.v.?

$$E_X\left[X\right] = \sum_{x \in S_X} x P_X(x) = \sum_{x=0,1} x P_X(x) = (0) \cdot P_X(0) + (1) \cdot P_X(1) = p$$

Example: Geometric r.v. Expectation
What is the $E_T\left[T\right]$, a Geometric r.v.?

$$E_T\left[T\right] = \sum_{t \in S_T} t P_T(t) = \sum_{t=1}^{\infty} tp(1-p)^{t-1} = p \sum_{t=1}^{\infty} t(1-p)^{t-1}$$

You'd need this series formula: $\sum_{i=1}^{\infty} iq^i = \frac{q}{(1-q)^2}$ to get:

$$E_T\left[T\right] = \frac{p}{1-p} \frac{1-p}{(1-(1-p))^2} = \frac{p}{p^2} = 1/p$$

Note: The expected value is a constant! More specifically, once you take the expected value w.r.t. X, you will no longer have a function of X.

Def'n: *Expected Value of a Function*
The expected value of a function $g(X)$ of a discrete r.v. X is

$$E_X\left[g(X)\right] = \sum_{x \in S_X} g(x) P_X(x)$$

The expected value is a linear operator. Consider $g(X) = aX + b$ for $a, b \in \mathbb{R}$. Then

$$\begin{aligned} E_X[g(X)] &= E_X[aX+b] = \sum_{x\in S_X}(ax+b)P_X(x) = \sum_{x\in S_X}[axP_X(x)+bP_X(x)] \\ &= \sum_{x\in S_X} axP_X(x) + \sum_{x\in S_X} bP_X(x) \\ &= a\sum_{x\in S_X} xP_X(x) + b\sum_{x\in S_X} P_X(x) = aE_X[X]+b \end{aligned} \tag{3}$$

Note: The expected value of a constant is a constant: $E_X[b] = b$

Note: The ability to separate an expected value of a sum into a sum of expected values would also have held if $g(X)$ was a more complicated function.
For example, let $g(X) = X^2 + \log X + X$:

$$E_X[g(X)] = E_X\left[X^2 + \log X + X\right] = E_X\left[X^2\right] + E_X[\log X] + E_X[X]$$

You can see how this would have come from the same procedure as in (3).

7.5 Moments

We've introduced taking the expected value of a function of a r.v., $g(X)$. Let's consider some common functions $g(X)$.

1. Let $g(X) = X$. We've already done this! The mean $\mu_X = E_X[X]$.
2. Let $g(X) = X^2$. The value $E_X\left[X^2\right]$ is called the *second moment*.
3. Let $g(X) = X^n$. The value $E_X[X^n]$ is called the nth moment.
4. Let $g(X) = (X-\mu_X)^2$. This is the *second central moment*. This is also called the variance. What are the units of the variance?
5. Let $g(X) = (X-\mu_X)^n$. This is the nth central moment.

Some notes:

1. "Moment" is used by analogy to the moment of inertia of a mass. Moment of inertia describes how difficult it is to get a mass rotating about its center of mass, and is given by:

$$I \triangleq \int\int\int_V r^2 \rho dx\, dy\, dz$$

   where ρ is the mass density, and r is the distance from the center.

2. Standard deviation $\sigma = \sqrt{\mathrm{Var}_X[X]}$.
3. Variance in terms of 1st and 2nd moments.

$$E_X[(X-\mu_X)^2] = E_X[X^2 - 2X\mu_X + \mu_X^2] = E_X[X^2] - 2E_X[X]\mu_X + \mu_X^2 = E_X[X^2] - (E_X[X])^2.$$

4. Variance of a linear combination of X:

$$\mathrm{Var}_X\left[aX+b\right] = E_X\left[(aX+b-E_X\left[aX+b\right])^2\right] = E_X\left[(aX-aE_X\left[X\right])^2\right] = a^2\mathrm{Var}_X\left[X\right]$$

   THE MULTIPLYING CONSTANT IS SQUARED, THE ADDITIONAL CONSTANT DROPS. VARIANCE IS NOT A LINEAR OPERATOR!!!

5. Note $E_X\left[g(X)\right] \neq g\left(E_X\left[X\right]\right)$!

A summary:

Expression	X **is a discrete r.v.**
$E_X[X]$	$= \sum_{x\in S_X} xP_X(x)$
$E_X[g(X)]$	$= \sum_{x\in S_X} g(x)P_X(x)$
$E_X[aX+b]$	$= aE_X[X]+b$
$E_X[X^2]$	$= \sum_{x\in S_X} x^2P_X(x)$
$\mathrm{Var}_X[X] = E_X[(X-\mu_X)^2]$, $\mu_X = E_X[X]$	$= \sum_{x\in S_X}(x-\mu_X)^2P_X(x)$

7.6 More Discrete r.v.s

We already introduced the Bernoulli pmf, the Geometric pmf, and the Zipf pmf.

Example: Bernoulli Moments
What is the 2nd moment and variance of of X, a Bernoulli r.v.?

Solution: To be completed in class.

Example: Mean of the Zipf distribution
What is the mean of a Zipf random variable R? Recall that

$$P_R(r) = \frac{1/r^s}{H_{N,s}}$$

where $H_{N,s} = \sum_{n=1}^{N} \frac{1}{r^s}$.

Solution: To be completed in class.

Def'n: *Binomial r.v.*
A r.v. K is Binomial with success probability p and number of trials n if it has the pmf,

$$P_K(k) = \begin{cases} \binom{n}{k}p^k(1-p)^{n-k}, & k = 1\ldots n \\ 0, & o.w. \end{cases}$$

A binomial r.v. stems from n independent Bernoulli r.v.s. Specifically, it is the number of successes in n trials (where each trial has success with probability p). Mathematically, $K = \sum_{i=1}^{n} X_i$ where X_i for $i = 1\ldots n$ are bernoulli r.v.s.

- Since success (1) has probability p, and failure has probability $(1-p)$, what is the probability of this particular event: First, we have k successes and then $n-k$ failures? A: p^kp^{n-k}.
- Order of the n Bernoulli trials may vary; how many ways are there to arrange k successes into n slots? A: $\binom{n}{k}$.

Without proof, we state that $E_K[K] = np$. We'd need a good table of sums in order to prove this directly from the above formula. Intuitively, it makes sense that we're adding n bernoulli random variables, and each has $E_{X_i}[X_i] = p$, so the expected value of the sum is np.

These r.v.s derive from the Bernoulli:

Name	Parameters	Description
Binomial	p, n	The sum (number of successes) of n indep. binomial r.v.s.
Geometric	p	The number of trials up to & including the first success.
Pascal	p, k	The number of trials up to & including the k^{th} success.

The Pascal pmf is Definition 2.8 on page 58 of your book, please convince yourself that it makes sense to you.

Lecture 6

Today: (1) Cts r.v.s, (2) Expectation of Cts r.v.s, (3) Method of Moments

8 Continuous Random Variables

Def'n: *Continuous r.v.*
A r.v. is continuous-valued if its range S_X is uncountably infinite (*i.e.*, not countable).

E.g., the 'Wheel of Fortune', for which $X \in [0, 1)$. **pmfs are meaningless**. Why? Because $P[X = x] = 0$. Why is that?

Lemma: Let $x \in [0, 1)$. (Eg., $x = 0.5$). Then $P[\{x\}] = 0$.
Proof: Proof by contradiction. Suppose $P[\{x\}] = \epsilon > 0$. Let $N = \lceil \frac{1}{\epsilon} \rceil + 1$. (Eg., $\epsilon = 0.001 \rightarrow N = 1001$). Then

$$P\left[\bigcup_{n=0}^{N-1} \left\{\frac{n}{N}\right\}\right] = \sum_{n=0}^{N-1} P\left[\left\{\frac{n}{N}\right\}\right] = \sum_{n=0}^{N-1} \epsilon = N\epsilon > 1.$$

Contradiction! Thus $P[\{x\}] = 0, \forall x \in S$.

However, CDFs *are* still meaningful.

8.1 Example CDFs for Continuous r.v.s

Example: CDF for the wheel of fortune
What is the CDF $F_X(x) = P[[0, x]]$?

By 'uniform' we mean that the probability is proportional to the size of the interval.

$$F_X(x) = P[[0, x]] = a(x - 0)$$

for some constant a. Since we know that $lim_{x \rightarrow +\infty} F_X(x) = 1$, we know that for $x = 1$, $F_X(x) =$

$a(1-0)=1$. Thus $a=1$ and

$$F_X(x) = P\left[[0,x]\right] = \begin{cases} 0, & x<0 \\ x, & 0 \le x < 1 \\ 1, & x \ge 1 \end{cases}$$

In general a *uniform random variable* X with $S_X=[a,b)$ has

$$F_X(x) = P\left[[a,x]\right] = \begin{cases} 0, & x<a \\ \frac{x-a}{b-a}, & a \le x < b \\ 1, & x \ge b \end{cases}$$

8.2 Probability Density Function (pdf)

Def'n: *Probability density function (pdf)*
The pdf of a continuous r.v. X, $f_X(x)$, can be written as the derivative of its CDF:

$$f_X(x) = \frac{\partial F_X(x)}{\partial x}$$

Properties:

1. $f_X(a)$ is the *density*. Not a probability!.
2. $\epsilon \cdot f_X(a)$ is approximately the probability that X falls in an ϵ-wide window around a. Its a good approximation if $\epsilon \approx 0$.
3. $F_X(x) = \int_{-\infty}^{x} f_X(u)du$. (Fundamental theorem of calculus)
4. $\forall x,\ f_X(x) \ge 0$. (from non-decreasing property of F_X)
5. $\int_{-\infty}^{\infty} f_X(u)du = 1$. (from limit property of F_X)
6. $P\left[a < X \le b\right] = \int_a^b f_X(x)dx$.

Draw a picture of a pmf and pdf, emphasizing *mass* vs. *density*.

8.3 Expected Value (Continuous)

Def'n: *Expected Value (Continuous)*
The expected value of continuous r.v. X is

$$E_X\left[X\right] = \int_{x \in S_X} x f_X(x)dx$$

It can still be seen as the 'center of mass'.

8.4 Examples

Def'n: *Pareto pdf*
A r.v. X is Pareto if it has CDF,

$$F_X(x) = \begin{cases} 1-(x_{min}/x)^k, & x > x_{min} \\ 0, & o.w. \end{cases}.$$

Example: Pareto pdf and expected value

1. Find the pdf.

$$\begin{aligned} f_X(x) &= \frac{\partial}{\partial x}F_X(x) = \frac{\partial}{\partial x}\left[1-\left(\frac{x_{min}}{x}\right)^k\right] \\ &= -x_{min}^k \frac{\partial}{\partial x}x^{-k} = kx_{min}^k x^{-k-1} = k\frac{x_{min}^k}{x^{k+1}} \end{aligned}$$

   So for an arbitrary x,

$$f_X(x) = \begin{cases} k\frac{x_{min}^k}{x^{k+1}}, & x \geq x_{min} \\ 0, & o.w. \end{cases}$$

2. Find the expected value.

$$\begin{aligned} E_X[X] &= \int_{x_{min}}^{\infty} xk\frac{x_{min}^k}{x^{k+1}}dx = kx_{min}^k \int_{x_{min}}^{\infty}\frac{1}{x^k}dx \\ &= kx_{min}^k \frac{-1}{k-1}\left[\frac{1}{x^{k-1}}\right|_{x_{min}}^{\infty} = x_{min}^k \frac{k}{k-1}\frac{1}{x_{min}^{k-1}} = \frac{k}{k-1}x_{min} \end{aligned}$$

Def'n: *Exponential r.v.*
A r.v. T is exponential with parameter λ if it has CDF,

$$F_T(t) = \begin{cases} 0, & t < 0 \\ 1-e^{-\lambda t}, & t \geq 0 \end{cases}$$

Many times, time delays are modeled as Exponential.

Example: Exponential pdf and $E_T[T]$

1. What is the pdf of T, if it is exponential with parameter λ? For $t \geq 0$,

$$f_T(t) = \frac{\partial}{\partial t}F_T(t) = \frac{\partial}{\partial t}\left[1-e^{-\lambda t}\right] = \lambda e^{-\lambda t}$$

   So for an arbitrary t,

$$f_X(x) = \begin{cases} \lambda e^{-\lambda t}, & x \geq 0 \\ 0, & o.w. \end{cases}$$

2. What is the expected value of T?

$$\begin{aligned} E_T[T] &= \int_{t=0}^{\infty} t\lambda e^{-\lambda t} dt = \lambda \int_{t=0}^{\infty} t e^{-\lambda t} dt \\ &= \lambda \frac{1}{\lambda^2} = \frac{1}{\lambda} \end{aligned}$$

8.5 Expected Values

Def'n: *Expected Value (Continuous)*
The expected value of continuous r.v. X is

$$E_X[X] = \int_{x \in S_X} x f_X(x) dx$$

It is still the 'center of mass'.
Def'n: *Gaussian r.v.*
A r.v. X is Gaussian with mean μ and variance σ^2 if it has pdf,

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

- This pdf is also known as the Normal distribution (except in engineering).
- Zero-mean unit-variance Gaussian is also known as the 'standard Normal'.

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

- Matlab generates standard Normal r.v.s with the 'randn' command.

Incidentally, the CDF of a standard Normal r.v. X is,

$$\Phi(x) \triangleq F_X(x) = \int_{-\infty}^{x} f_X(u) du = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du$$

Theorem: If X is a Gaussian r.v. with mean μ and standard deviation σ, then the CDF of X is

$$\Phi\left(\frac{x-\mu}{\sigma}\right)$$

Proof: Left to do on your own.

Example: Prove that the expected value of a Gaussian r.v. is μ

$$E_X[X] = \int_{x=-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} dx$$

Let $u = \frac{(x-\mu)^2}{2\sigma^2}$. Then $du = \frac{x-\mu}{\sigma^2}dx$. That is, $\sigma^2 du = (x-\mu)dx$

$$\begin{aligned}
E_X[X] &= \int_{x=-\infty}^{\infty}(x-\mu+\mu)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/(2\sigma^2)}dx \\
E_X[X] &= \mu+\int_{x=-\infty}^{\infty}(x-\mu)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/(2\sigma^2)}dx \\
E_X[X] &= \mu+\int_{x=-\infty}^{\infty}\frac{\sigma}{\sqrt{2\pi}}e^{-u}du \\
E_X[X] &= \mu+\frac{\sigma}{\sqrt{2\pi}}\left[-e^{-u}\right|_{x=-\infty}^{\infty} \\
E_X[X] &= \mu+\frac{\sigma}{\sqrt{2\pi}}\left[-e^{-\frac{x-\mu}{\sigma^2}}\right|_{x=-\infty}^{\infty}=\mu.
\end{aligned}$$

9 Method of Moments

Example: Ceiling of an Exponential is a Geometric
If Y is an Exponential(λ) random variable, show that $M = \lceil Y \rceil$ is a Geometric(p) random variable, with $p = 1 - e^{-\lambda}$.

Answer: Find the pmf of M. Note that $f_X(x) = \lambda e^{-\lambda x}$ when $x \geq 0$, and zero otherwise. First note that

$$P_M(0) = P[\{X = 0\}] = 0$$

Then that for $m = 1, 2, \ldots$,

$$\begin{aligned}
P_M(m) &= P[\{m-1 < X \leq m\}] = \int_{m-1}^{m}\lambda e^{-\lambda x}dx \\
&= \left[-e^{-\lambda x}\right|_{m-1}^{m} = -e^{-\lambda m}+e^{-\lambda(m-1)} = e^{-\lambda(m-1)}(1-e^{-\lambda})
\end{aligned}$$

Let $p = 1 - e^{-\lambda}$, then $P_M(m) = (1-p)^{m-1}p$, for $m = 1, 2, \ldots$, and zero otherwise. Since it has the Geometric pmf with parameter $p = 1 - e^{-\lambda}$, it is Geometric.

'Derived Distributions' or 'Transformation of r.v.s'.
Given: a distribution of X, either pdf, CDF, or pmf. And, a function $Y = g(X)$.

Find: the distribution of Y.

9.1 Discrete r.v.s Method of Moments

For discrete r.v. X,

$$P_Y(y) = P[\{Y = y\}] = P[\{g(X) = y\}] = P\left[\{X \in g^{-1}(y)\}\right] = \sum_{x \in g^{-1}(y)} P_X(x).$$

Just find, for each value y in the range of Y, the values of x for which $y = g(x)$. Sum $P_X(x)$ for those values of x. This gives you $P_Y(y)$.

Note: Write these steps out EACH TIME.

Example: Overtime
Overtime of a made-up game proceeds as follows.

- Overtime consists of independent rounds $1, 2, \ldots$ and stops when a team wins in its round.
- In odd rounds, Team 1 (the first team to go) has a chance to win and will win with probability p.
- In even rounds, Team 2 has the chance to win, and will win with probability p.
- Define T to be the number of the winning team.

What is the pmf $P_T(t)$?

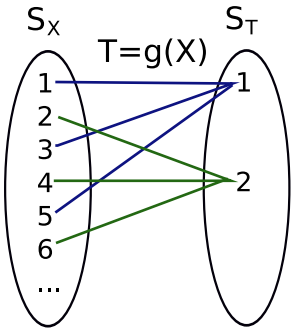


Figure 5: Overtime function $T = g(X)$. Team 1 wins if overtime ends in an odd-numbered round, and Team 2 wins if overtime ends in an even-numbered round.

Solution: Define X to be the ending round. X is Geometric, because we know that the number of trials to the first success is a Geometric r.v. Then, the event $T = 1$ happens when X is odd, and $T = 2$ happens when X is even. So we could write $g(X) = \begin{cases} 1, & X = odd \\ 2, & X = even \end{cases}$

$$\begin{aligned} P_T(1) &= P[T=1] = P[\{g(X)=1\}] = P\left[\{X \in g^{-1}(1)\}\right] = \sum_{x\in\{1,3,\ldots\}} P_X(x) \\ &= \sum_{x\in\{1,3,\ldots\}} p(1-p)^{x-1} = p \sum_{x'\in\{0,1,\ldots\}} [(1-p)^2]^{x'} = \frac{p}{1-(1-p)^2} \\ &= \frac{p}{1-(1-2p+p^2)} = \frac{p}{p(2-p)} = \frac{1}{2-p} \end{aligned} \tag{4}$$

Similarly,

$$\begin{aligned} P_T(2) &= P[T=2] = P[\{g(X)=2\}] = P\left[\{X \in g^{-1}(2)\}\right] = \sum_{x\in\{2,4,\ldots\}} P_X(x) \\ &= \sum_{x\in\{2,4,\ldots\}} p(1-p)^{x-1} = p(1-p) \sum_{x'\in\{0,1,\ldots\}} [(1-p)^2]^{x'} = \frac{p(1-p)}{1-(1-p)^2} \\ &= \frac{p(1-p)}{1-(1-2p+p^2)} = \frac{p(1-p)}{p(2-p)} = \frac{1-p}{2-p} \end{aligned} \tag{5}$$

Do they add to one? $\frac{1-p}{2-p} + \frac{1}{2-p} = \frac{2-p}{2-p} = 1$. Yep.

Lecture 7

Today: (1) Method of Moments, cts r.v.s (Y&G 3.7), (2) Jacobian Method, cts. r.v.s (Kay handout)

9.2 Method of Moments, continued

Last time: For discrete r.v. X, and a function $Y = g(X)$,

$$P_Y(y) = P\left[\{Y = y\}\right] = P\left[\{g(X) = y\}\right] = P\left[\{X \in g^{-1}(y)\}\right]$$

Def'n: *Many-to-One*
A function $Y = g(X)$ is many-to-one if, for some value y, there is more than one value of x such that $y = g(x)$, or equivalently, $\{g^{-1}(y)\}$ has more than one element.

Def'n: *One-to-One*
A function $Y = g(X)$ is one-to-one if, for every value y, there is exactly one value x such that $y = g(x)$.

Bottom line: know that the set $\{g^{-1}(y)\}$ can have either one, or many, elements.

Example: One-to-One Transform
Let X be a discrete uniform r.v. on $S_X = \{1, 2, 3\}$, and $Y = g(X) = 2X$. What is $P_Y(y)$? Answer: We can see that $S_Y = \{2, 4, 6\}$. Then

$$P_Y(y) = P\left[\{Y = y\}\right] = P\left[\{2X = y\}\right] = P\left[\{X = y/2\}\right] = P_X(y/2) = \begin{cases} 1/3, & y = 2, 4, 6 \\ 0, & o.w. \end{cases}$$

9.3 Continuous r.v.s Method of Moments

For continuous r.v. X and a transformation $g(X)$, find $f_Y(y)$ by:

1. Find the CDF by the method of moments.

$$F_Y(y) = P\left[\{Y \leq y\}\right] = P\left[\{g(X) \leq y\}\right] = P\left[\{X \in g^{-1}(\{Y : Y \leq y\})\}\right]$$

   For example (starting from f_X) integrate the pdf of X over the set of X which 'causes' in $Y = g(X) \leq y$.

2. Then find the pdf by taking the derivative of the CDF.

Example: One-to-One Transform (continuous)
Let X be a discrete uniform r.v. on $S_X = [0, 1)$, and $Y = g(X) = 2X$. (Is this function one-to-one or many-to-one?) What is $P_Y(y)$? Answer: We can see that $f_X(x) = 1$ when $0 \leq X < 1$ and zero otherwise, and that $S_Y = [0, 2)$. Then

1. Find the CDF by the method of moments. For $0 \le y < 2$,

$$F_Y(y) \;=\; P\left[\{Y \le y\}\right] = P\left[\{2X \le y\}\right] = P\left[\{X \le y/2)\}\right] = \int_{x=0}^{y/2} 1dx = y/2.$$

2. Then find the pdf by taking the derivative of the CDF.

$$f_Y(y) = \tfrac{\partial}{\partial y} F_Y(y) = \tfrac{\partial}{\partial y} y/2 = 1/2,$$

for $0 \le y < 2$ and zero otherwise.

Example: Uniform and $1/X$
Let X be Uniform$(0, 2)$, and $Y = g(X) = 1/X$. (Is this function one-to-one or many-to-one?) Compute $f_Y(y)$.

1. Note $f_X(x) = 1/2$ for $0 < X < 2$ and 0 otherwise.

2. Find the CDF by the method of moments.

$$\begin{aligned} F_Y(y) \;&=\; P\left[\{Y \le y\}\right] = P\left[\{\frac{1}{X} \le y\}\right] = P\left[X \ge \frac{1}{y}\right] \\ &=\; \int_{1/y}^{2} (1/2)dx = (1/2)\; x|_{1/y}^{2} = 1 - \frac{1}{2y} \end{aligned}$$

For $0 < 1/y < 2$, or $y > 1/2$. So

$$F_Y(y) = \begin{cases} 0, & y < 1/2 \\ 1 - \frac{1}{2y}, & y \ge 1/2 \end{cases}$$

Check: F_Y is continuous at $1/2$, starts at 0, and ends at 1.

3. Then find the pdf by taking the derivative of the CDF.

$$f_Y(y) = \tfrac{\partial}{\partial y} F_Y(y) = \tfrac{\partial}{\partial y}\left(1 - \frac{1}{2y}\right) = \begin{cases} \frac{1}{2y^2}, & y > 1/2 \\ 0, & o.w. \end{cases}$$

Example: Absolute Value
For an arbitrary pdf $f_X(x)$, and $Y = g(X) = |X|$, find $f_Y(y)$ in terms of $f_X(x)$. (Is this function one-to-one or many-to-one?)

1. Find the CDF by the method of moments. For $y \ge 0$,

$$\begin{aligned} F_Y(y) \;&=\; P\left[\{Y \le y\}\right] = P\left[\{|X| \le y\}\right] = P\left[-y \le X \le y\right] \\ &=\; \int_{-y}^{y} f_X(x)dx = F_X(y) - F_X(-y) \end{aligned}$$

So

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ F_X(y) - F_X(-y), & y \ge 0 \end{cases}$$

Check: $F_Y(y)$ is continuous at $y = 0$, and starts at 0 and ends at 1.

2. Then find the pdf by taking the derivative of the CDF.

$$f_Y(y) = \tfrac{\partial}{\partial y}\left[F_X(y) - F_X(-y)\right] = f_X(y) + f_X(-y),$$

for $y > 0$ and 0 otherwise.

10 Jacobian Method

Going back to the example where $Y = g(X) = 2X$. From the Matlab-generated example shown in the Figure below, you can see that the pdf becomes less 'dense'.

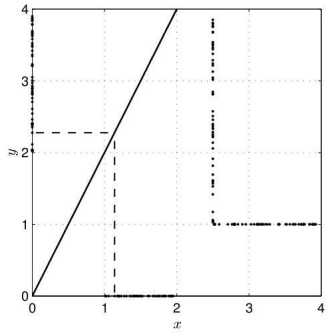


Figure: Uniform(1, 2) r.v.s X are plotted as dots on the x-axis. They are mapped with $Y = 2X$ and plotted as points on the Y-axis. Compare the "density" of points on each axis [S. M. Kay, "Intuitive Probability ...", 2006].

What would have happened if the slope was higher than 2? What if the slope was lower than 2? *Answer: the density is inversely proportional to slope.* What if the slope was curvy (non-linear)? (The density of Y would be higher when the slope was lower, and the density would be lower when the slope was higher.)

Thus, without proof:

Def'n: *Jacobian method*
For a cts r.v. X and a one-to-one and differentiable function $Y = g(X)$,

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

This slope correction factor is really called the Jacobian.

Example: Uniform and $1/X$
Let X be Uniform$(0, 2)$, and $Y = g(X) = 1/X$. (Is this function one-to-one or many-to-one?) Compute $f_Y(y)$.

$$x = 1/y = g^{-1}(y)$$

Taking the derivative,

$$\frac{\partial g^{-1}(y)}{\partial y} = -\frac{1}{y^2}$$

Now, plugging into the formula,

$$f_Y(y) = f_X(1/y) \left| \frac{\partial g^{-1}(y)}{\partial y} \right| = \begin{cases} \frac{1}{2} \left| -\frac{1}{y^2} \right|, & y > 1/2 \\ 0, & o.w. \end{cases}$$

So finally, the form of $f_Y(y)$ is

$$f_Y(y) = \begin{cases} \frac{1}{2y^2}, & y > 1/2 \\ 0, & o.w. \end{cases}$$

Example: Absolute Value
For an arbitrary pdf $f_X(x)$, and $Y = g(X) = |X|$, find $f_Y(y)$ in terms of $f_X(x)$.

We cannot find this pdf using the Jacobian method – the function is not differentiable at $x = 0$.

Def'n: *Jacobian method for Many-to-One*
For a cts r.v. X and a many-to-one and differentiable function $Y = g(X)$ with multiple inverse functions $x_1 = g_1^{-1}(y)$, $x_2 = g_2^{-1}(y), \ldots,$

$$f_Y(y) = f_X(g_1^{-1}(y)) \left| \frac{\partial g_1^{-1}(y)}{\partial y} \right| + f_X(g_2^{-1}(y)) \left| \frac{\partial g_2^{-1}(y)}{\partial y} \right| + \ldots$$

Example: Square of a Gaussian r.v.
Consider $X \sim \mathcal{N}(0,1)$ and $g(X) = X^2$.

There are two possible inverse functions,

$$g_1^{-1}(y) = -\sqrt{y}, \quad g_2^{-1}(y) = \sqrt{y}$$

Taking their derivatives,

$$\frac{\partial g_1^{-1}(y)}{\partial y} = -\frac{1}{2\sqrt{y}}, \quad \frac{\partial g_2^{-1}(y)}{\partial y} = \frac{1}{2\sqrt{y}}$$

So plugging into the formula,

$$\begin{aligned} f_Y(y) &= f_X(-\sqrt{y}) \left| -\frac{1}{2\sqrt{y}} \right| + f_X(\sqrt{y}) \left| \frac{1}{2\sqrt{y}} \right| = \frac{1}{\sqrt{2\pi}} e^{-y/2} \frac{1}{2\sqrt{y}} + \frac{1}{\sqrt{2\pi}} e^{-y/2} \frac{1}{2\sqrt{y}} \\ f_Y(y) &= \begin{cases} \frac{1}{\sqrt{2\pi y}} e^{-y/2}, & y \geq 0 \\ 0, & o.w. \end{cases} \end{aligned}$$

Lecture 8

Today: (1) Expectation for cts. r.v.s (Y&G 3.3), (2) Conditional Distributions (Y&G 2.9, 3.8)

11 Expectation for Continuous r.v.s

Expression	X **is a discrete r.v.**	X **is a continuous r.v.**
$E_X[X]$	$= \sum_{x \in S_X} x P_X(x)$	$= \int_{S_X} x f_X(x)$
$E_X[g(X)]$	$= \sum_{x \in S_X} g(x) P_X(x)$	$= \int_{S_X} g(x) f_X(x)$
$E_X[aX + b]$	$= aE_X[X] + b$	$= aE_X[X] + b$
$E_X[X^2]$ 2^{nd} moment	$= \sum_{x \in S_X} x^2 P_X(x)$	$= \int_{S_X} x^2 f_X(x)$
$\mathrm{Var}_X[X] = E_X[(X - \mu_X)^2]$, $\mu_X = E_X[X]$	$= \sum_{x \in S_X} (x - \mu_X)^2 P_X(x)$	$= \int_{S_X} (x - \mu_X)^2 f_X(x)$

Example: Variance of Uniform r.v.
Let X be a continous uniform r.v. on (a, b), with $a, b > 0$.

1. What is $E_X[X]$? It is

$$\int_a^b \frac{x}{b-a} dx = \frac{1}{2(b-a)} x^2 \Big|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}.$$

2. What is $E_X\left[\frac{1}{X}\right]$?
$$\int_a^b \frac{1}{b-a}\frac{1}{x}dx = \frac{1}{b-a}(\ln b - \ln a) = \frac{1}{b-a}\ln\frac{b}{a}.$$

3. What is $E_X\left[X^2\right]$? It is
$$\int_a^b \frac{x^2}{b-a}dx = \frac{1}{3(b-a)}\,x^3\Big|_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{b^2+ab+a^2}{3}.$$

4. What is $\mathrm{Var}_X[X]$? It is
$$\mathrm{Var}_X[X] = E_X[X^2] - (E_X[X])^2 = \frac{b^2+ab+a^2}{3} - \frac{b^2+2ab+a^2}{4} = \frac{b^2-2ab+a^2}{12} = \frac{(b-a)^2}{12}$$

12 Conditional Distributions

Def'n: *Conditional pdf*
For a continuous r.v. X with pdf $f_X(x)$, and an event $B \subset S_X$ with $P[B] > 0$, the conditional pdf $f_{X|B}(x)$ is defined as
$$f_{X|B}(x) = \begin{cases} \frac{f_X(x)}{P[B]}, & x \in B \\ 0, & o.w. \end{cases}$$

Note: Remember the "zero otherwise"!

Def'n: *Conditional pmf*
For discrete r.v. X with pdf $P_X(x)$, and an event $B \subset S_X$ with $P[B] > 0$, the conditional pdf $P_{X|B}(x|B)$ is the probability mass function of X given that event B occurred, and is
$$P_{X|B}(x) = \begin{cases} \frac{P_X(x)}{P[B]}, & x \in B \\ 0, & o.w. \end{cases}$$

12.1 Conditional Expectation and Probability

Expression	X **is a discrete r.v.**	X **is a continuous r.v.**
	Given a partition $B_1, \ldots, B_m$ of the event space S_X	
Law of Total Prob.	$P_X(x) = \sum_{i=1}^m P_{X\|B_i}(x)P[B_i]$	$f_X(x) = \sum_{i=1}^m f_{X\|B_i}(x)P[B_i]$
$E_{X\|B}[g(X)]$	$= \sum_{x\in S_X} g(x)P_{X\|B}(x)$	$= \int_{S_{X\|B}} g(x)f_{X\|B}(x)dx$
$E_{X\|B}[X]$	$= \sum_{x\in S_X} xP_{X\|B}(x)$	$= \int_{S_{X\|B}} xf_{X\|B}(x)dx$
$\mathrm{Var}_{X\|B}[X\|B]$	$= \sum_{x\in S_X}(x-\mu_{X\|B})^2P_{X\|B}(x)$	$= \int_{S_{X\|B}}(x-\mu_{X\|B})^2f_{X\|B}(x)dx$

Essentially, $X|B$ is a new r.v. Thus we treat it the same as any other r.v. – it has a (conditional) pdf, expected values and variance of its own.

The law of total probability is just a re-writing of the original law of total probability to use $P_{X|B_i}(x)$ instead of $P[\{X = x\}|B_i]$.

Example: Lifetime of Hard Drive
Company Flake produces hard drives, and one out of every six are defective (D) and fail more quickly than the good hard drives (G). The defective drives have pdf of T, time to failure

$$f_{T|D}(t) = \begin{cases} 0.5e^{-0.5t}, & t \geq 0 \\ 0, & o.w. \end{cases}$$

while the good drives have pdf of T, time to failure

$$f_{T|D}(t) = \begin{cases} 0.1e^{-0.1t}, & t \geq 0 \\ 0, & o.w. \end{cases}$$

The drives are indistinguishable when purchased. What is the pdf of T? Solution: For $t \geq 0$,

$$\begin{aligned} f_T(t) &= f_{T|D}(t)P[D] + f_{T|G}(t)P[G] \\ &= 0.5e^{-0.5t}(1/6) + 0.1e^{-0.1t}(5/6) \\ &= \frac{1}{12}(e^{-0.5t} + e^{-0.1t}) \end{aligned}$$

Recall: A r.v. T is *exponential* with parameter λ if

$$\begin{aligned} F_T(t) &= \begin{cases} 0, & t < 0 \\ 1 - e^{-\lambda t}, & t \geq 0 \end{cases} \\ f_T(t) &= \begin{cases} \lambda e^{-\lambda t}, & t \geq 0 \\ 0, & o.w. \end{cases} \end{aligned}$$

Example: Conditional Exponential Given Delay
Assume that you are wait for the 1:00 bus starting at 1:00. The actual arrival time in minutes past the hour is X, an exponential r.v. with parameter λ.

1. Let D be the event that you have waited for 5 minutes without seeing the bus. What is the conditional pdf of X (the additional time you will wait) given D?
2. What is the conditional pdf of $Y = X - 5$ given D?

Solution:

$$\begin{aligned} P[D] &= P[X > 5] = 1 - F_X(5) = 1 - (1 - e^{-\lambda 5}) = e^{-5\lambda} \\ f_{X|D}(x) &= \begin{cases} \frac{f_X(x)}{P[D]}, & x \in D \\ 0, & o.w. \end{cases} = \begin{cases} \lambda e^{-\lambda(x-5)}, & X > 5 \\ 0, & o.w. \end{cases} \end{aligned}$$

For Y, Let's use the Jacobian method to find the pdf of $Y|D$. The inverse equation is $g^{-1}(Y) = X = Y + 5$. Thus

$$\frac{\partial}{\partial y}(y + 5) = 1$$

So

$$f_{Y|D}(y|D) = f_{X|D}(g^{-1}(y)|D)|\frac{\partial}{\partial y}g^{-1}(y)| = \begin{cases} \lambda e^{-\lambda y}, & y > 0 \\ 0, & o.w. \end{cases}$$

The same as the original (un-conditional) distribution of X!

Note: Exponential forgetfulness
An exponential r.v. is said to have no 'memory'. The pdf of X is the same as the conditional pdf of $X - r$ given $X > r$. Try this in Matlab!

Example: Y&G 3.8.4
W is Gaussian with mean 0 and variance $\sigma^2 = 16$. Given the event $C = \{W > 0\}$,

1. What is $f_{W|C}(w)$? First, since W is zero-mean and f_W is symmetric, $P[C] = 1/2$. Then,

$$f_{W|C}(w) = \begin{cases} 2\frac{1}{\sqrt{32\pi}}e^{-w^2/(32)}, & w > 0 \\ 0, & o.w. \end{cases}$$

2. What is $E_{W|C}[W|C]$?

$$E_{W|C}[W|C] = \int_{w=0}^{\infty} w\frac{2}{\sqrt{32\pi}}e^{-w^2/32}dw$$

   Making the substitution $v = w^2/32$, $dv = 2w/32$,

$$E_{W|C}[W|C] = \frac{32}{\sqrt{32\pi}}\int_0^{\infty} e^{-v} = \sqrt{\frac{32}{\pi}}$$

3. What is $\mathrm{Var}_{W|C}[W|C]$?

$$E_{W|C}\left[W^2|C\right] = \int_0^{\infty} 2w^2\frac{1}{\sqrt{32\pi}}e^{-w^2/32}$$

   Knowing that $\frac{2w^2}{\sqrt{32\pi}}e^{-w^2/32}$ is an even expression, the integral for $w > 0$ is the same as the integral for $w < 0$ – half the total.

$$E_{W|C}\left[W^2|C\right] = \int_{-\infty}^{\infty} w^2\frac{1}{\sqrt{32\pi}}e^{-w^2/32} = \mathrm{Var}_W[W] = 16.$$

   The variance is then

$$\mathrm{Var}_{W|C}[W|C] = E_{W|C}\left[W^2|C\right] - (E_{W|C}[W|C])^2 = 16 - \frac{32}{\pi}$$

Lecture 9

Today: (1) Joint distributions: Intro

13 Joint distributions: Intro (Multiple Random Variables)

Often engineering problems can't be described with just one random variable. And random variables are often related to each other. For example:

1. ICs are made up of resistances, capacitances, inductances, and transistor characteristics, all of which are random, dependent on the outcome of the manufacturing process. A voltage reading at some point in the IC may depend on many of these parameters.
2. A chemical reaction may depend on the concentration of multiple reactants, which may change randomly over time.
3. Control systems for vehicles may measure from many different sensors to determine what to do to control the vehicle.

13.1 Event Space and Multiple Random Variables

Def'n: *Multiple Random Variables*
A set of n random variables which result from the same experiment or measurement are a mapping from the sample space S to $\mathbb{R}^n$

Example: Two dice are rolled.
An outcome $s \in S$ is the result of the experiment of rolling two dice. Let $X_1(s)$ be the number on die 1, and $X_2(s)$ be the number on die 2. The coordinate (X_1, X_2) is a function of s and lies in a two-dimensional space $\mathbb{R}^2$ or more specifically, $\{1,2,3,4,5,6\}^2$.

Example: Digital Communications Receiver.
We receive a packet from another radio. Let $X_i(s)$ be the ith recorded sample in that packet, for $i = 0, \ldots, n$, assuming we need n total samples to recover the data in the packet.

We still can use S_{X_1} as the event space for X_1, to describe the range or possible values of X_1.

13.2 Joint CDFs

Def'n: *Joint Cumulative Distribution Function*
The Joint CDF of the random variables X_1 and X_2 is

$$F_{X_1,X_2}(x_1, x_2) = P\left[\{X_1 \le x_1\} \cap \{X_2 \le x_2\}\right] = P\left[X_1 \le x_1, X_2 \le x_2\right]$$

Note that the book uses X and Y. You can name these two r.v.s anything you want! I like my notation because it more easily shows how you can add more r.v.s into the picture. Eg.,

- $F_{X_1,X_2,X_3}(x_1, x_2, x_3) = P\left[X_1 \le x_1, X_2 \le x_2, X_3 \le x_3\right]$
- $F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P\left[X_1 \le x_1, \ldots, X_n \le x_n\right]$

Let's consider $F_{X_1,X_2}(x_1, x_2)$. It calculates the probability of the event that (X_1, X_2) fall in the area shown in Fig. 6(a). This area is lower and left of (x_1, x_2).

Example: Probability in a rectangular area
Let's consider the event $A = \{a < X_1 \le b\} \cap \{c < X_2 \le d\}$, which is shown in Fig. 6(b). How can we calculate this from the CDF?

$$P\left[A\right] = F_{X_1,X_2}(b, d) - F_{X_1,X_2}(a, d) - F_{X_1,X_2}(b, c) + F_{X_1,X_2}(a, c)$$

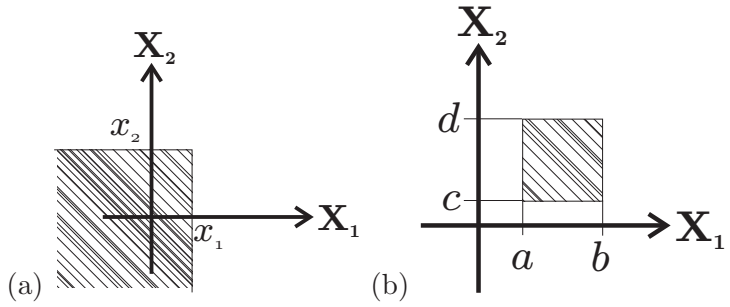


Figure 6: (a) A 2-D joint CDF gives the probability of (X_1, X_2) in the area shown. (b) The smaller area shown can also be calculated from the joint CDF.

Properties:

1. **Limits**: Taking the limit as $x_i \to \infty$ removes x_i from the Joint CDF. Taking the limit as $x_i \to -\infty$ for any i will result in zero. To see this, consider the joint CDF as $P\left[\{X_1 \le x_1\} \cap \{X_2 \le x_2\}\right]$ and consider what the probability is when one set becomes S or $\emptyset$.

$$\begin{aligned} lim_{x_1 \to \infty} F_{X_1,X_2}(x_1, x_2) &= F_{X_2}(x_2) && (6) \\ lim_{x_1 \to -\infty} F_{X_1,X_2}(x_1, x_2) &= 0 \\ lim_{x_2 \to \infty} F_{X_1,X_2}(x_1, x_2) &= F_{X_1}(x_1) && (7) \\ lim_{x_2 \to -\infty} F_{X_1,X_2}(x_1, x_2) &= 0 \\ lim_{x_1 \text{ and } x_2 \to \infty} F_{X_1,X_2}(x_1, x_2) &= 1 \end{aligned}$$

The numbered lines are called 'marginal' CDFs. We used to call them just CDFs when we only had one r.v. to worry about. Now, so we don't get confused, we call them marginal CDFs.

13.2.1 Discrete / Continuous combinations

Since we have multiple random variables, some may be continuous and some may be discrete. When $n = 2$, we could have:

- X_1 continuous and X_2 continuous.
- X_1 discrete and X_2 discrete.
- X_1 continuous and X_2 discrete.

The first two you can imagine are relevant. But the third is very often relevant in ECE. Consider the digital communication system shown in Figure 7. The binary symbol X_1 results in a continuous r.v. X_4 at the receiver.

Because of the combinations, we prefer to use the CDF whenever possible. FYI, there is an area called 'Measure Theory' which is used by probability theorists to provide unifying notation so that discrete and continuous r.v.s can be treated equally, and at the same time.

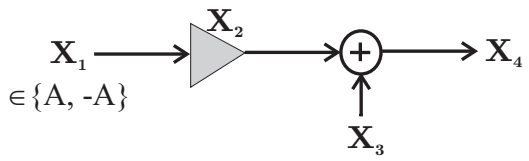


Figure 7: Here, a binary signal X_1 is transmitted at signal amplitude A or $-A$ (discrete r.v.). It is multiplied by the loss of the fading channel X_2 (continuous r.v.). Then thermal noise X_3 (continuous r.v.) adds to the remaining signal. The final received signal has amplitude X_4 (continuous r.v.).

13.3 Joint pmfs and pdfs

Def'n: *Joint pmf*
For discrete random variables X_1 and X_2, we define their joint probability mass function as

$$P_{X_1,X_2}(x_1,x_2) = P\left[\{X_1 = x_1\} \cap \{X_2 = x_2\}\right]$$

Def'n: *Joint pdf*
For continuous random variables X_1 and X_2, we define their joint probability density function as

$$f_{X_1,X_2}(x_1,x_2) = \frac{\partial^2}{\partial x_1 \partial x_2} F_{X_1,X_2}(x_1,x_2)$$

Properties of the pmf and pdf still hold:

- sum / integrate to 1.
- non-negativity of pdf / pmf.

These are still good checks!

13.4 Marginal pmfs and pdfs

When X_1 and X_2 are discrete, the **'marginal pmf'** of X_1 is the 'pmf' we were talking about in Chapter 2. It is the probability mass function for one of the random variables, averaging out the other random variable(s). Mathematically,

$$\begin{aligned} P_{X_1}(x_1) &= \sum_{x_2 \in S_{X_2}} P_{X_1,X_2}(x_1,x_2) \\ P_{X_2}(x_2) &= \sum_{x_1 \in S_{X_1}} P_{X_1,X_2}(x_1,x_2) \end{aligned}$$

By summing over the other r.v., we effectively eliminate it. This is the same as the property from the joint CDF.

When X_1 and X_2 are continuous, the **'marginal pdf'** of continuous random variable X_1 is

$$\begin{aligned} f_{X_1}(x_1) &= \int_{x_2 \in S_{X_2}} f_{X_1,X_2}(x_1, x_2) \\ f_{X_2}(x_2) &= \int_{x_1 \in S_{X_1}} f_{X_1,X_2}(x_1, x_2) \end{aligned}$$

13.5 Independence of pmfs and pdfs

We studied independence of sets. We had that sets A and B are independent if and only if $P[A \cap B] = P[A]\,P[B]$. Now, we have random variables, and they can be independent also.

Random variables X_1 and X_2 are independent if and only if, for all x_1 and x_2,

$$\begin{aligned} P_{X_1,X_2}(x_1, x_2) &= P_{X_1}(x_1) P_{X_2}(x_2) \\ f_{X_1,X_2}(x_1, x_2) &= f_{X_2}(x_2) f_{X_1}(x_1) \end{aligned}$$

We are still looking at a probability of an intersection of events on the left, and a product of probabilities of events on the right (for the discrete case). For the cts. case, it is similar, but with probability densities.

Example: Binary r.v.s
(X_1, X_2) are in $\{1, 2\}^2$. We measure that:

- $P_{X_1,X_2}(1,1) = 0.5$, $P_{X_1,X_2}(1,2) = 0.1$.
- $P_{X_1,X_2}(2,1) = 0.1$, $P_{X_1,X_2}(2,2) = 0.3$.

What are the following?

- $P_{X_1}(1)$? A: $= P_{X_1,X_2}(1,1) + P_{X_1,X_2}(1,2) = 0.5 + 0.1 = 0.6$.
- $P_{X_1}(2)$? A: $= P_{X_1,X_2}(2,1) + P_{X_1,X_2}(2,2) = 0.1 + 0.3 = 0.4$.
- $P_{X_2}(1)$? A: $= P_{X_1,X_2}(1,1) + P_{X_1,X_2}(2,1) = 0.5 + 0.1 = 0.6$.
- $P_{X_2}(2)$? A: $= P_{X_1,X_2}(1,2) + P_{X_1,X_2}(2,2) = 0.1 + 0.3 = 0.4$.
- Are X_1 and X_2 independent? No, look at

$$P_{X_1}(1) P_{X_2}(1) = 0.36 \neq 0.5 = P_{X_1,X_2}(1,1).$$

Example: Uniform on a triangle
Let X_1, X_2 have joint pdf

$$f_{X_1,X_2}(x_1, x_2) = \begin{cases} A, & 0 \leq X_1 < 1 \text{ and } 0 \leq X_2 < 1 - X_1 \\ 0, & o.w. \end{cases}$$

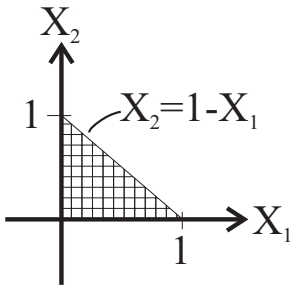


Figure 8: Area of (X_1, X_2) over which the pdf $f_{X_1,X_2}(x_1, x_2)$ is equal to A.

1. What is A?

$$\begin{aligned} 1 &= \int_{x_1=0}^{1} \int_{x_2=0}^{1-x_1} A dx_1 dx_2 \\ &= A \int_{x_1=0}^{1} (1 - x_1) dx_1 \\ &= A \left(x_1 - x_1^2/2\right)\Big|_{x_1=0}^{1} dx_1 \\ &= A/2 \end{aligned}$$

   Thus $A = 2$.

2. What is $f_{X_1}(x_1)$? A: $= \int_{x_2=0}^{1-x_1} 2dx_2 = 2(1 - x_1)$, for $0 \le x_1 < 1$, and zero otherwise.

3. What is $f_{X_2}(x_2)$? A: $= \int_{x_1=0}^{1-x_2} 2dx_1 = 2(1 - x_2)$, for $0 \le x_2 < 1$, and zero otherwise. Note that when you take the integral over x_1, you consider x_2 to be a constant. Given this value of x_2, what are the limits on x_1? It still must be above zero, but it is zero past $1 - x_2$.

4. Are X_1 and X_2 independent? No, since within the triangle,

$$f_{X_1,X_2}(x_1, x_2) = 2 \neq 4(1 - x_1)(1 - x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$$

   Note that $f_{X_1}(x_1) f_{X_2}(x_2)$ is non-zero for all $0 \le X_1 < 1$ and $0 \le X_2 < 1$, which is a square area, while $f_{X_1,X_2}(x_1, x_2)$ is only non-zero for the triangular area. This next tip generalizes this observation:

QUICK TIP: Any time the support (non-zero probability portion) of one random variable depends on another random variable, the two r.v.s are NOT independent. That is, if the non-zero probability area of two r.v.s is non-square, they must be dependent.

Lecture 10

Today: (1) Conditional Joint pmfs and pdfs

13.6 Review of Joint Distributions

This is Sections 4.1-4.5.

For two random variables X_1 and X_2,

- Joint CDF: $F_{X_1,X_2}(x_1,x_2) = P\left[\{X_1 \le x_1\} \cap \{X_2 \le x_2\}\right]$ It is the probability that both events happen simultaneously.
- Joint pmf: $P_{X_1,X_2}(x_1,x_2) = P\left[\{X_1 = x_1\} \cap \{X_2 = x_2\}\right]$ It is the probability that both events happen simultaneously.
- Joint pdf: $f_{X_1,X_2}(x_1,x_2) = \frac{\partial^2}{\partial x_1 \partial x_2} F_{X_1,X_2}(x_1,x_2)$

The pmf and pdf still integrate/sum to one, and are non-negative.

Now, to find a probability, you must double sum or double integrate. For example, for event $B \in S$,

- Discrete case: $P[B] = \sum\sum_{(X_1,X_2)\in B} P_{X_1,X_2}(x_1,x_2)$
- Continuous Case: $P[B] = \int\int_{(X_1,X_2)\in B} f_{X_1,X_2}(x_1,x_2)$

We also talked about marginal distributions:

- Marginal pmf: $P_{X_2}(x_2) = \sum_{X_1 \in S_{X_1}} P_{X_1,X_2}(x_1,x_2)$
- Marginal pdf: $f_{X_2}(x_2) = \int_{X_1 \in S_{X_1}} f_{X_1,X_2}(x_1,x_2)$

Finally we talked about independence of random variables. Two random variables X_1 and X_2 are independent iff for all x_1 and x_2,

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

Example: Collisions in Packet Radio
A packet radio protocol specifies that radios will start transmitting their packet randomly, and independently, within a transmission period of duration w ms. Let T_1 and T_2 be the times that radios 1 and 2 start transmitting their packet. Both packets are of length s ms. What is the probability that they collide, *i.e.*, that at any time, both radios are transmitting simultaneously?

Solution:

- Packets finish transmitting at $T_1 + s$ and $T_2 + s$.
- If $T_1 < T_2$, Then the packets collide if $T_1 + s > T_2$.
- If $T_2 < T_1$, Then the packets collide if $T_2 + s > T_1$.
- $P[C] = 1 - P[C^c]$

- $$\begin{aligned} P[C^c] &= 2\int_{t_1=s}^{w}\int_{t_2=0}^{t_1-s}\frac{1}{w^2}dt_2dt_1 \\ &= 2\int_{t_1=s}^{w}\frac{1}{w^2}\int_{t_2=0}^{t_1-s}dt_2dt_1 \\ &= \frac{2}{w^2}\int_{t_1=s}^{w}(t_1-s)dt_1 \\ &= \frac{2}{w^2}\left[\frac{1}{2}t_1^2-st_1\right|_{t_1=s}^{w} \\ &= \frac{2}{w^2}\left[(\frac{1}{2}w^2-sw)-(\frac{1}{2}s^2-s^2)\right] \\ &= \frac{1}{w^2}\left[w^2-2sw+s^2\right] \\ &= \frac{(w-s)^2}{w^2}=[1-(s/w)]^2 \end{aligned}$$

  This is a real system design question!

  - For a given s, if you make w longer, you'll have a higher probability of successful transmission of both packets.
  - But as you make w longer, and we need to send many packets to send a whole file, then you're increasing the time until completion, and thus decreasing the bit rate.
  - Wired and wireless networks.
  - This model can account for non-independent and non-uniform T_1 and T_2. Real world events introduce non-independence.

14 Joint Conditional Probabilities

Two types of joint conditioning!

1. Conditioned on an event $B \in S$.
2. Conditioned on a random variable.

14.1 Joint Probability Conditioned on an Event

This is section 4.8.

Given event $B \in S$ which has $P[B] > 0$, the joint probability conditioned on event B is

- Discrete case:
$$P_{X_1,X_2|B}(x_1,x_2) = \begin{cases} P_{X_1,X_2}(x_1,x_2)/P[B], & (X_1,X_2)\in B \\ 0, & o.w. \end{cases}$$
- Continuous Case:
$$f_{X_1,X_2|B}(x_1,x_2) = \begin{cases} f_{X_1,X_2}(x_1,x_2)/P[B], & (X_1,X_2)\in B \\ 0, & o.w. \end{cases}$$

14.2 Joint Probability Conditioned on a Random Variable

This is section 4.9.

Given r.v.s X_1 and X_2,

- Discrete case. The conditional pmf of X_1 given $X_2 = x_2$, where $P_{X_2}(x_2) > 0$, is
$$P_{X_1|X_2}(x_1|x_2) = P_{X_1,X_2}(x_1, x_2)/P_{X_2}(x_2)$$
- Continuous Case: The conditional pdf of X_1 given $X_2 = x_2$, where $f_{X_2}(x_2) > 0$, is
$$f_{X_1|X_2}(x_1|x_2) = f_{X_1,X_2}(x_1, x_2)/f_{X_2}(x_2)$$

Note

- The joint pdf of X_1 conditioned on $X_2 = x_2$ is a pdf (a probability model) for X_1, NOT FOR X_2.
- This means that $\int_{x_1 \in S_{X_1}} f_{X_1|X_2}(x_1|x_2)dx_1 = 1$
- But $\int_{x_2 \in S_{X_2}} f_{X_1|X_2}(x_1|x_2)dx_2$ is meaningless!!!

Example: 4.17 from Y&G
Random variables X_1 and X_2 have the joint pmf shown in Example 4.17 (A triangular pattern with $P_{X_1,X_2}(1,1) = 1/4$, $P_{X_1,X_2}(2, 1 or 2) = 1/8$, $P_{X_1,X_2}(3, 1 \ldots 3) = 1/12$, $P_{X_1,X_2}(4, 1 \ldots 4) = 1/16$).

1. What is $P_{X_1}(x_1)$? A:
$$P_{X_1}(x_1) = \begin{cases} \frac{1}{4}, & X_1 = 1, 2, 3, 4 \\ 0, & o.w. \end{cases}$$
2. What is $P_{X_2|X_1}(x_2|x_1)$? A:
$$P_{X_2|X_1}(x_2|x_1) = \frac{P_{X_2,X_1}(x_2, x_1)}{P_{X_1}(x_1)} = 4P_{X_2,X_1}(x_2, x_1)$$
For each $x_1 = 1, 2, 3, 4$, there is a different pmf:
$$\begin{aligned} P_{X_2|X_1}(x_2|1) &= \begin{cases} 1, & X_2 = 1 \\ 0, & o.w. \end{cases} \\ P_{X_2|X_1}(x_2|2) &= \begin{cases} 1/2, & X_2 = 1, 2 \\ 0, & o.w. \end{cases} \\ P_{X_2|X_1}(x_2|3) &= \begin{cases} 1/3, & X_2 = 1, 2, 3 \\ 0, & o.w. \end{cases} \\ P_{X_2|X_1}(x_2|4) &= \begin{cases} 1/4, & X_2 = 1, 2, 3, 4 \\ 0, & o.w. \end{cases} \end{aligned} \tag{8}$$
Given a particular value of X_1, we have a different discrete uniform pmf as a result.

Lecture 11

Today: Joint r.v.s (1) Expectation of Joint r.v.s, (2) Covariance (both in Y&G 4.7)

15 Expectation of Joint r.v.s

We can find the expected value of a random variable or a function of a random variable similar to how we learned it earlier. However, now we have a more complex model, and the calculation may be more complicated.

Def'n: *Expected Value (Joint)*
The expected value of a function $g(X_1, X_2)$ is given by,

1. Discrete: $E_{X_1,X_2}[g(X_1,X_2)] = \sum_{X_1 \in S_{X_1}} \sum_{X_2 \in S_{X_2}} g(X_1,X_2) P_{X_1,X_2}(x_1,x_2)$
2. Continuous: $E_{X_1,X_2}[g(X_1,X_2)] = \int_{X_1 \in S_{X_1}} \int_{X_2 \in S_{X_2}} g(X_1,X_2) P_{X_1,X_2}(x_1,x_2)$

Essentially we have a function of two random variables X_1 and X_2, called $Y = g(X_1, X_2)$. Remember when we had a function of a random variable? We could either

1. Derive the pdf/pmf of Y, and then take the expected value of Y.
2. Take the expected value of $g(X_1, X_2)$ using directly the model of X_1, X_2.

Example: Expected Value of $g(X_1)$
Let's look at the discrete case:

$$\begin{aligned} E_{X_1,X_2}[g(X_1)] &= \sum_{X_1 \in S_{X_1}} \sum_{X_2 \in S_{X_2}} g(X_1) P_{X_1,X_2}(x_1,x_2) \\ &= \sum_{X_1 \in S_{X_1}} g(X_1) \sum_{X_2 \in S_{X_2}} P_{X_1,X_2}(x_1,x_2) \\ &= \sum_{X_1 \in S_{X_1}} g(X_1) P_{X_1}(x_1) \\ &= E_{X_1}[g(X_1)] \end{aligned}$$

So if you're taking the expectation of a function of only one of the r.v.s, you can either use the joint probability model (pdf or pmf) OR you can use the marginal probability model (pdf or pmf). If you already have the marginal, it's easier that way.

Also - don't carry around subscripts when not necessary.

Example: Expected Value of $aX_1 + bX_2$
Let a, b be real numbers. Let's look at the discrete case:

$$\begin{aligned} E_{X_1,X_2}[aX_1 + bX_2] &= \sum_{X_1 \in S_{X_1}} \sum_{X_2 \in S_{X_2}} (aX_1 + bX_2) P_{X_1,X_2}(x_1,x_2) \\ &= \sum_{X_1 \in S_{X_1}} \sum_{X_2 \in S_{X_2}} (aX_1 P_{X_1,X_2}(x_1,x_2) + bX_2 P_{X_1,X_2}(x_1,x_2) \\ &= a \sum_{X_1 \in S_{X_1}} \sum_{X_2 \in S_{X_2}} (X_1 P_{X_1,X_2}(x_1,x_2) + b \sum_{X_1 \in S_{X_1}} \sum_{X_2 \in S_{X_2}} X_2 P_{X_1,X_2}(x_1,x_2) \\ &= aE_{X_1,X_2}[X_1] + bE_{X_1,X_2}[X_2] \\ &= aE_{X_1}[X_1] + bE_{X_2}[X_2] \end{aligned}$$

This is because the linearity of the sum (or integral) operator. This is Theorem 4.14 in your book (when $a = b = 1$).

Def'n: *Variance of* $Y = g(X_1, X_2)$
The variance of Y is the expected value, $E_{X_1,X_2}\left[(Y - E_{X_1,X_2}[Y])^2\right]$.

Example: Variance of $Y = aX_1 + bX_2$
Define $\mu_1 = E_{X_1}[X_1]$ and $\mu_2 = E_{X_2}[X_2]$. Let's look at the discrete case. First of all, we know $E_{X_1,X_2}[Y] = aE_{X_1}[X_1] + bE_{X_2}[X_2] = a\mu_1 + b\mu_2$. Thus

$$\begin{aligned}
\mathrm{Var}_{X_1,X_2}[Y] &= E_{X_1,X_2}\left[[aX_1 + bX_2 - (a\mu_1 + b\mu_2)]^2\right] \\
&= E_{X_1,X_2}\left[[a(X_1 - \mu_1) + b(X_2 - \mu_2)]^2\right] \\
&= E_{X_1,X_2}\left[a^2(X_1 - \mu_1)^2\right] + E_{X_1,X_2}[2ab(X_1 - \mu_1)(X_2 - \mu_2)] + E_{X_1,X_2}\left[b^2(X_2 - \mu_2)^2\right] \\
&= a^2 E_{X_1}\left[(X_1 - \mu_1)^2\right] + 2abE_{X_1,X_2}[(X_1 - \mu_1)(X_2 - \mu_2)] + b^2 E_{X_2}\left[(X_2 - \mu_2)^2\right]
\end{aligned}$$

Because of theorem 4.14 in your book. Then looking at it this way:

$$\mathrm{Var}_{X_1,X_2}[X_1 + X_2] = \mathrm{Var}_{X_1}[X_1] + 2\mathrm{Cov}(X_1, X_2) + \mathrm{Var}_{X_2}[X_2]$$

we see that the variance of the sum is NOT the sum of the variances. Recall that variance is NOT a linear operator!

16 Covariance

Def'n: *The covariance of r.v.s* X_1 *and* X_2 *is given as*

$$Cov(X_1, X_2) = E_{X_1,X_2}[(X_1 - \mu_1)(X_2 - \mu_2)]$$

Repeat after me: **I WILL NOT MAKE THE EXPECTATION OF A PRODUCT OF TWO RANDOM VARIABLES EQUAL TO THE PRODUCT OF THEIR EXPECTATIONS.**

Because however the expectation of a sum is the sum of the expectations,

$$\mathrm{Cov}(X_1, X_2) = E_{X_1,X_2}[X_1X_2 - X_1\mu_2 - \mu_1X_2 - \mu_1\mu_2] = E_{X_1,X_2}[X_1X_2] - \mu_1\mu_2$$

Def'n: *Uncorrelated*
R.v.s X_1 and X_2 are called 'uncorrelated' if $\mathrm{Cov}(X_1, X_2) = 0$.

Note: If r.v.s X_1 and X_2 are independent, then they will also have $\mathrm{Cov}(X_1, X_2) = 0$. However, $\mathrm{Cov}(X_1, X_2) = 0$ does NOT imply independence!

Example: What is the Cov (X_1, X_2) of two independent r.v.s X_1 and X_2?
Let's use the continuous case:

$$\begin{aligned}
\text{Cov}(X_1, X_2) &= E_{X_1,X_2}[(X_1-\mu_1)(X_2-\mu_2)] = \int_{x_1}\int_{x_2}(x_1-\mu_1)(x_2-\mu_2)f_{X_1,X_2}(x_1,x_2)dx_1dx_2 \\
&= \int_{x_1}\int_{x_2}(x_1-\mu_1)(x_2-\mu_2)f_{X_1}(x_1)f_{X_2}(x_2)dx_1dx_2 \\
&= \int_{x_1}(x_1-\mu_1)f_{X_1}(x_1)dx_1\int_{x_2}(x_2-\mu_2)f_{X_2}(x_2)dx_2 \\
&= E_{X_1}[(X_1-\mu_1)]\,E_{X_2}[(X_2-\mu_2)] = (0)(0) = 0
\end{aligned}$$

Amendment to the original mantra: **I WILL NOT MAKE THE EXPECTATION OF A PRODUCT OF TWO RANDOM VARIABLES EQUAL TO THE PRODUCT OF THEIR EXPECTATIONS UNLESS THEY ARE UNCORRELATED OR STATISTICALLY INDEPENDENT.**

Going back to the previous expression:

$$\text{Var}_{X_1,X_2}[X_1+X_2] = \text{Var}_{X_1}[X_1] + 2\text{Cov}(X_1,X_2) + \text{Var}_{X_2}[X_2]$$

If the r.v.s are uncorrelated, *i.e.*, $\text{Cov}(X_1, X_2) = 0$, then

$$\text{Var}_{X_1,X_2}[X_1+X_2] = \text{Var}_{X_1}[X_1] + \text{Var}_{X_2}[X_2]$$

True, but often a source of confusion.

Def'n: *Correlation Coefficient*
The correlation coefficient of X_1 and X_2 is given by

$$\rho_{X_1,X_2} = \frac{\text{Cov}(X_1,X_2)}{\sqrt{\text{Var}_{X_1}[X_1]\,\text{Var}_{X_2}[X_2]}} = \frac{\text{Cov}(X_1,X_2)}{\sigma_{X_1}\sigma_{X_2}}.$$

Note: The correlation coefficient is our intuitive notion of correlation.

- It is bounded by -1 and 1.
- Close to one, we say the two r.v.s have high positive correlation.
- Close to -1, we say the two have high negative correlation.
- Close to zero, we say they have little correlation. Equal to zero, the two are uncorrelated.

2nd mantra: **I WILL NOT MAKE THE VARIANCE OF A SUM OF TWO RANDOM VARIABLES EQUAL TO THE SUM OF THE TWO VARIANCES UNLESS THEY HAVE ZERO COVARIANCE.**

16.1 'Correlation'

Def'n: *'Correlation'*
The correlation of two random variables X_1 and X_2 is denoted:

$$r_{X_1,X_2} = E_{X_1,X_2}[X_1X_2]$$

Note: This nomenclature is BAD. The popular notion of correlation is **not** represented by this definition for correlation. I will call this 'the expected value of the product of' rather than 'the correlation of'.

In our new notation,

$$\mathrm{Cov}\left(X_1, X_2\right) = r_{X_1,X_2} - \mu_1\mu_2$$

Lecture 12

Today: (1) Joint r.v. Expectation Review (2) Transformations of Joint r.v.s, Y&G 4.6 (3) Random Vectors (R.V.s), Y&G 5.2

16.2 Expectation Review

Short "quiz". Given r.v.s. X_1 and X_2, what is

1. What is $\mathrm{Var}\left[X_1 + X_2\right]$?
2. What is the definition of $\mathrm{Cov}\left(X_1, X_2\right)$?
3. What do we call two r.v.s with zero covariance?
4. What is the definition of correlation coefficient?

Note: We often define several random variables to be independent, and to have identical distributions (CDF or pdf or pmf). We abbreviate "i.i.d." for "independent and identically distributed".

17 Transformations of Joint r.v.s

Random variables are often a function of multiple other random variables. The example the book uses is a good one, of a multiple antenna receiver. How do you choose from the antenna signals?

1. Just choose the best one: This uses the $\max(X_1, X_2)$ function.
2. Add them together; $X_1 + X_2$. 'Combining'.
3. Add them in some ratio: $X_1/\sigma_1 + X_2/\sigma_2$. 'Maximal Ratio Combining'.

We may have more than one output: we'll have $Y_1 = aX_1 + bX_2$ and $Y_2 = cX_1 + dX_2$, where a, b, c, d, are constants. If we choose wisely to match to the losses in the channel, we won't loose any of the information that is contained in X_1 and X_2. Ideas that exploit this lead to space-time coding and MIMO communication systems, now seen in 802.11n.

Here, we're going to show how to come up with a model for these derived random variables.

Figure 9: A function Y of two random variables $Y = g(X_1, X_2)$, might be viewed as a 3D map of what value Y takes for any given input coordinate (X_1, X_2), like this topology map of Black Mountain, Utah. Contour lines give $Y = y$, for many values of y, which is useful to find the pre-image. One pre-image of importance is the coordinates (X_1, X_2) for which $Y \leq y$.

17.1 Method of Moments for Joint r.v.s

Let $Y = g(X_1, X_2)$. What is the pdf (and CDF) for Y? Using the method of moments, we again find the CDF and then find the pdf.

$$F_Y(y) = P\left[\{Y \leq y\}\right] = P\left[\{g(X_1, X_2) \leq y\}\right]$$

In general, it is very hard to find the 'pre-image', the area of (X_1, X_2) for which $g(X_1, X_2) \leq y$. However, for many cases, it is possible. For example, see the example $g(X_1, X_2)$, shown as a topology map in Figure 9. What is the pre-image of $g(X_1, X_2) < 7200$ feet? (Answer: Follow the 7200 contour line and cut out the mountain ridges above this line.)

Example: Max of two r.v.s

(You are doing this for a particular pdf on your HW 5; here we do it in general for any pdf). Let X_1 and X_2 be continuous r.v.s with CDF $F_{X_1,X_2}(x_1, x_2)$. (a) What is the CDF of Y, the maximum of X_1 and X_2? (b) If X_1 and X_2 are i.i.d. with $f_X(x)$ uniform on $(0, a)$, what is the pdf of Y?

$$F_Y(y) = P\left[\{Y \leq y\}\right] = P\left[\{\max(X_1, X_2) \leq y\}\right] = P\left[\{X_1 \leq y\} \cap \{X_2 \leq y\}\right]$$

The Max $< y$ bit translates to BOTH variables being less than y, as shown in Figure 10.

$$F_Y(y) = F_{X_1,X_2}(y, y)$$

Second part. If X_1 and X_2 are independent with the same CDF,

$$F_Y(y) = F_{X_1,X_2}(y, y) = F_{X_1}(y)F_{X_1}(y) = [F_{X_1}(y)]^2$$

To find the pdf,

$$f_Y(y) = \frac{\partial}{\partial y}F_Y(y) = \frac{\partial}{\partial y}[F_{X_1}(y)]^2 = 2F_{X_1}(y)f_{X_1}(y)$$

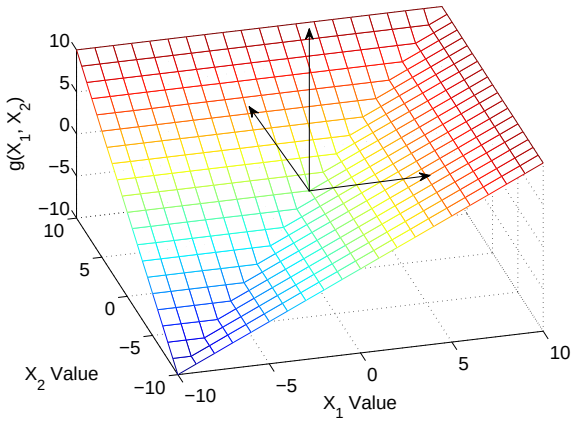


Figure 10: The function $Y = \max(X_1, X_2)$. The coordinate axes centered at (0,0,0) are shown as arrows.

For the case of uniform $(0, a)$, the CDF is $F_{X_1}(y) = y/a$, between 0 and a, and the pdf is $f_{X_1}(y) = 1/a$, between 0 and a. Thus

$$f_Y(y) = \begin{cases} y/a^2, & 0 < y < a \\ 0, & o.w. \end{cases}.$$

Note: To really show that you understand the stuff from today's lecture, you can do the following: Show that for n i.i.d continuous r.v.s, $X_1 \ldots X_n$, that the pdf of $Y = \max(X_1 \ldots X_n)$ is given by

$$f_Y(y) = n[F_{X_1}(y)]^{n-1} f_{X_1}(y)$$

This is a practical problem! Often for uniform r.v.s we don't know the limits (a, b), *e.g.*, the highest energy of a bunch of batteries, the maximum range of a transmitter, or in sales, they want to know the maximum price a person would pay for something. We can estimate that maximum by taking many independent samples, and taking the maximum Y. By some analysis, we can know the pdf of our estimate Y (the mean and variance are especially important).

Example: Sum of two r.v.s

This is actually covered in the book in Section 6.2.
Let $W = X_1 + X_2$, and X_1, X_2 have CDF $F_{X_1,X_2}(x_1, x_2)$. What is the pdf of W?

$$F_W(w) = P\left[\{W \leq w\}\right] = P\left[\{X_1 + X_2 \leq w\}\right] = P\left[\{X_2 \leq w - X_1\}\right]$$

Steps for drawing an inequality picture.

1. Make it an equality. Draw the line.
2. Change it back to an inequality. Pick points on both sides of the line, and see which side meets the inequality.

(Draw picture). Thus the CDF of W is:

$$F_W(w) = \int_{x_1=-\infty}^{\infty} \left(\int_{x_2=-\infty}^{w-x_1} f_{X_1,X_2}(x_1,x_2)dx_2 \right) dx_1$$

Now, to find the pdf,

$$\begin{aligned} f_W(w) &= \tfrac{\partial}{\partial w} F_W(w) = \tfrac{\partial}{\partial w} \int_{x_1=-\infty}^{\infty} \left(\int_{x_2=-\infty}^{w-x_1} f_{X_1,X_2}(x_1,x_2)dx_2 \right) dx_1 \\ &= \int_{x_1=-\infty}^{\infty} \left(\tfrac{\partial}{\partial w} \left(\int_{x_2=-\infty}^{w-x_1} f_{X_1,X_2}(x_1,x_2)dx_2 \right) \right) dx_1 \\ &= \int_{x_1=-\infty}^{\infty} f_{X_1,X_2}(x_1, w-x_1)dx_1 \end{aligned}$$

That last line is from the fundamental theorem of calculus, that the derivative of the integral is the function itself. You do have to be careful about using it when the limits aren't as simple. (I would provide this theorem on a HW or test if it was needed, but you should be familiar with it.)

What if X_1 and X_2 are independent?

$$f_W(w) = \int_{x_1=-\infty}^{\infty} f_{X_1}(x_1) f_{X_2}(w-x_1)dx_1$$

This is a convolution! It pops up all the time in ECE.

$$f_W(w) = \int_{x_1=-\infty}^{\infty} f_{X_1}(w-x_2) f_{X_2}(x_2)dx_2$$

We write it as $f_W(w) = f_{X_1}(x_1) * f_{X_2}(x_2)$.

Example: Sum of Exponentials
Let X_1, X_2 be i.i.d. Exponential with parameter λ. What is $f_Y(y)$ for $Y = X_1 + X_2$?

$$\begin{aligned} f_Y(y) &= \int_{x_1=-\infty}^{\infty} f_{X_1}(y-x_2) f_{X_2}(x_2)dx_2 \\ &= \int_{x_1=0}^{y} \lambda e^{-\lambda(y-x_2)} \lambda e^{-\lambda x_2} dx_2 \\ &= \lambda^2 e^{-\lambda y} \int_{x_1=0}^{y} dx_2 \\ &= \begin{cases} \lambda^2 y e^{-\lambda y}, & y > 0 \\ 0, & o.w. \end{cases} \end{aligned} \tag{9}$$

By the way, $E_Y[Y] = E_{X_1,X_2}[X_1 + X_2] = \frac{2}{\lambda} = \frac{1}{\lambda} + \frac{1}{\lambda}$, as we would expect.

18 Random Vectors

a.k.a. Multiple random variables. This is Section 5.2 in Y&G.

Def'n: *Random Vector*
A random vector (R.V.) is a list of multiple random variables $X_1, X_2, \ldots, X_n$ in a vector:

$$\mathbf{X} = [X_1, X_2, \ldots, X_n]^T$$

The transpose operator is denoted $'$ in the book, and ^{T} by me.

1. $\mathbf{X}$ is the random vector.
2. $\mathbf{x}$ is a value that the random vector takes.

Here are the Models of R.V.s:

1. The CDF of R.V. $\mathbf{X}$ is $F_{\mathbf{X}}(\mathbf{x}) = F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P[X_1 \le x_1, \ldots, X_n \le x_n]$.
2. The pmf of R.V. $\mathbf{X}$ is $P_{\mathbf{X}}(\mathbf{x}) = P_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P[X_1 = x_1, \ldots, X_n = x_n]$.
3. The pdf of R.V. $\mathbf{X}$ is $f_{\mathbf{X}}(\mathbf{x}) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F_{\mathbf{X}}(\mathbf{x})$.

We can find the marginals with multiple r.v.s just like before. To 'eliminate' a r.v. from the model, we sum (or integrate) from $-\infty$ to ∞, *i.e.*, the whole range of that r.v. Some examples:

$$\begin{aligned}
f_{X_1,X_2,X_3}(x_1,x_2,x_3) &= \int_{S_{X_4}} f_{X_1,X_2,X_3,X_4}(x_1,x_2,x_3,x_4)dx_4 \\
f_{X_2,X_3}(x_2,x_3) &= \int_{S_{X_1}}\int_{S_{X_4}} f_{X_1,X_2,X_3,X_4}(x_1,x_2,x_3,x_4)dx_4dx_1 \\
P_{X_2}(x_2) &= \sum_{S_{X_1}}\sum_{S_{X_3}}\sum_{S_{X_4}} P_{X_1,X_2,X_3,X_4}(x_1,x_2,x_3,x_4)
\end{aligned}$$

(Its hard to write these in vector notation!)

Conditional distributions: If we measure that one or more random variables takes some particular values, we can use that as 'given' information to make a new conditional model. Just divide the joint model with the marginal model for the random variables that are 'given'. Some examples:

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

18.1 Expectation of R.V.s

We can find expected values of the individual random variables or of a function of many of the random variables:

$$E_{\mathbf{X}}[X_1] = \sum_{x_1 \in S_{X_1}} \cdots \sum_{x_n \in S_{X_n}} x_1 P_{\mathbf{X}}(\mathbf{x})$$

or

$$E_{\mathbf{X}}[g(\mathbf{X})] = \sum_{x_1 \in S_{X_1}} \cdots \sum_{x_n \in S_{X_n}} g(\mathbf{x}) P_{\mathbf{X}}(\mathbf{x})$$

Notably, we may need to refer to all of the random variable expectations as $\boldsymbol{\mu}_{\mathbf{X}}$:

$$\boldsymbol{\mu}_{\mathbf{X}} = [E_{\mathbf{X}}[X_1], \ldots, E_{\mathbf{X}}[X_n]]$$

Lecture 13

Today: (1) Covariance Matrices, (2) Gaussian R.V.s

19 Covariance of a R.V.

Def'n: *Covariance Matrix*
The covariance matrix of an n-length random vector $\mathbf{X}$ is an $n \times n$ matrix $C_{\mathbf{X}}$ with (i,j)th element equal to $\mathrm{Cov}(X_i, X_j)$. In vector notation,

$$C_{\mathbf{X}} = E_{\mathbf{X}}\left[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})^T\right]$$

Example: For $\mathbf{X} = [X_1 X_2 X_3]^T$

$$C_{\mathbf{X}} = \begin{bmatrix} \mathrm{Var}_{X_1}[X_1] & \mathrm{Cov}(X_1, X_2) & \mathrm{Cov}(X_1, X_3) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}_{X_2}[X_2] & \mathrm{Cov}(X_2, X_3) \\ \mathrm{Cov}(X_3, X_1) & \mathrm{Cov}(X_3, X_2) & \mathrm{Var}_{X_3}[X_3] \end{bmatrix}$$

You can see that for two r.v.s, we'll have just the first two rows and two columns of $C_{\mathbf{X}}$ – this is what we put on the board when we first talked about covariance as a matrix. Note for $n = 1$, $C_{\mathbf{X}} = \sigma^2_{X_1}$.

20 Joint Gaussian r.v.s

We often (OFTEN) see joint Gaussian r.v.s. E.g. ECE 5520, Digital Communications, joint Gaussian r.v.s are everywhere. In addition, the joint Gaussian R.V. is extremely important in statistics, economics, other areas of engineering. We can't overemphasize its importance. In many areas of the sciences, the Gaussian r.v. is an approximation. In digital communications, control systems, and signal processing, the Gaussian r.v. can be a very accurate representation of noise.

Def'n: *Multivariate Gaussian r.v.s.*
An n-length R.V. $\mathbf{X}$ is multivariate Gaussian with mean $\mu_{\mathbf{X}}$, and covariance matrix $C_{\mathbf{X}}$ if it has the pdf,

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n det(C_{\mathbf{X}})}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu_{\mathbf{X}})^T C_{\mathbf{X}}^{-1}(\mathbf{x} - \mu_{\mathbf{X}})\right]$$

where $det()$ is the determinant of the covariance matrix, and $C_{\mathbf{X}}^{-1}$ is the inverse of the covariance matrix.

Note: The 'Inner Product' means that the transpose is in the middle. You will get one number out of an inner product. The 'Outer Product' means the transpose is on the outside of the product,

and you will get a matrix out of it.

Example: Write out in non-vector notation the pdf of a $n = 2$ bivariate Gaussian R.V. from the vector definition of a multivariate Gaussian R.V. pdf.

1. Find $det(C_{\mathbf{X}})$. Since

$$C_{\mathbf{X}} = \begin{bmatrix} \text{Var}[X_1] & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_2, X_1) & \text{Var}[X_2] \end{bmatrix} = \begin{bmatrix} \sigma_{X_1}^2 & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \sigma_{X_2}^2 \end{bmatrix}$$

   Thus

$$det(C_{\mathbf{X}}) = \sigma_{X_1}^2\sigma_{X_2}^2 - \rho^2\sigma_{X_1}^2\sigma_{X_2}^2 = \sigma_{X_1}^2\sigma_{X_2}^2(1-\rho^2)$$

2. Find $C_{\mathbf{X}}^{-1}$.

$$C_{\mathbf{X}}^{-1} = \frac{1}{\sigma_{X_1}^2\sigma_{X_2}^2(1-\rho^2)} \begin{bmatrix} \sigma_{X_2}^2 & -\rho\sigma_{X_1}\sigma_{X_2} \\ -\rho\sigma_{X_1}\sigma_{X_2} & \sigma_{X_1}^2 \end{bmatrix}$$

3. Plug these into the pdf:

$$\begin{aligned}
f_{\mathbf{X}}(\mathbf{x}) &= \frac{1}{\sqrt{(2\pi)^n det(C_{\mathbf{X}})}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{X}})^T C_{\mathbf{X}}^{-1}(\mathbf{x}-\mu_{\mathbf{X}})\right] \\
&= \eta \exp\left[-\frac{1}{2\sigma_{X_1}^2\sigma_{X_2}^2(1-\rho^2)}[x_1-\mu_1, x_2-\mu_2] \begin{bmatrix} \sigma_{X_2}^2 & -\rho\sigma_{X_1}\sigma_{X_2} \\ -\rho\sigma_{X_1}\sigma_{X_2} & \sigma_{X_1}^2 \end{bmatrix} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}\right] \\
&= \eta \exp\left[-\frac{1}{2\sigma_{X_1}^2\sigma_{X_2}^2(1-\rho^2)}[x_1-\mu_1, x_2-\mu_2] \begin{bmatrix} \sigma_{X_2}^2(x_1-\mu_1) - \rho\sigma_{X_1}\sigma_{X_2}(x_2-\mu_2) \\ -\rho\sigma_{X_1}\sigma_{X_2}(x_1-\mu_1) + \sigma_{X_1}^2(x_2-\mu_2) \end{bmatrix}\right] \\
&= \eta \exp\left[-\frac{\sigma_{X_2}^2(x_1-\mu_1)^2 - 2\rho\sigma_{X_1}\sigma_{X_2}(x_2-\mu_2)(x_1-\mu_1) + \sigma_{X_1}^2(x_2-\mu_2)^2}{2\sigma_{X_1}^2\sigma_{X_2}^2(1-\rho^2)}\right] \\
&= \eta \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{X_1-\mu_1}{\sigma_{X_1}}\right)^2 - 2\rho\left(\frac{X_1-\mu_1}{\sigma_{X_1}}\right)\left(\frac{X_2-\mu_2}{\sigma_{X_2}}\right) + \left(\frac{X_2-\mu_2}{\sigma_{X_2}}\right)^2\right]\right\}
\end{aligned}$$

   where $\eta = \frac{1}{\sqrt{(2\pi)^2\sigma_{X_1}^2\sigma_{X_2}^2(1-\rho^2)}} = \frac{1}{2\pi\sigma_{X_1}\sigma_{X_2}\sqrt{(1-\rho^2)}}$.

This is the form given to us in Y&G 4.11. See p. 192 for plots of the Gaussian pdf when means are zero, and variances are 1, and ρ varies.

In Y&G page 192-193, it points out to us that we can rewrite the final bivariate Gaussian form as follows:

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}} \frac{1}{\sqrt{2\pi\tilde{\sigma}_2^2}} e^{-\frac{(x_2-\tilde{\mu}_2(x_1))^2}{2\tilde{\sigma}_2^2}}$$

where

$$\tilde{\mu}_2(x_1) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1), \quad \text{and } \tilde{\sigma}_2^2 = \sigma_2^2(1-\rho^2)$$

This form helps see for the case of $n = 2$: (1) Any **marginal pdf** of a multivariate Gaussian R.V. is also (multivariate) Gaussian. (2) Any **conditional pdf** (conditioned on knowing one or more values) of a multivariate Gaussian R.V. is also Gaussian.

First, to see that the marginal is Gaussian,

$$\begin{aligned} f_{X_1}(x_1) &= \int_{x_2=-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)dx_2 \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}} \int_{x_2=-\infty}^{\infty} \frac{1}{\sqrt{2\pi\tilde{\sigma}_2^2}} e^{-\frac{(x_2-\tilde{\mu}_2(x_1))^2}{2\tilde{\sigma}_2^2}} dx_2 \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}} \end{aligned}$$

Next, the conditional pdf is

$$f_{X_2|X_1}(x_2|x_1) = f_{X_1,X_2}(x_1,x_2)/f_{X_1}(x_1) = \frac{1}{\sqrt{2\pi\tilde{\sigma}_2^2}} e^{-\frac{(x_2-\tilde{\mu}_2(x_1))^2}{2\tilde{\sigma}_2^2}}$$

20.1 Linear Combinations of Gaussian R.V.s

We also can prove that: *Any linear combination of multivariate Gaussian R.V.s is also multivariate Gaussian.* A linear combination is any sum or weighted sum. **Once you know that a R.V. (or a r.v.) is Gaussian, all you need to do is find its mean vector and covariance matrix – no need to use the method-of-moments or Jacobian method to find the pdf.**

We will talk in lecture 13 about how to find the mean and covariance matrix of a linear combination expressed in vector notation: $\mathbf{Y} = A\mathbf{X}$, where $\mathbf{X}$ is a n-length and $\mathbf{Y}$ is and m-length random vectors, and A is an $m \times n$ matrix of constants. We will present a formula for $\mu_{\mathbf{Y}}$ given $\mu_{\mathbf{X}}$, and a formula for $C_{\mathbf{Y}}$ given $C_{\mathbf{X}}$ that works for any distribution. If we know $\mathbf{X}$ is multivariate Gaussian, then $\mathbf{Y}$ is also Gaussian. Since we know exactly the mean vector and covariance matrix are for $\mathbf{Y}$, we know all there is to know about its joint distribution!

Lecture 14

Today: (1) Linear Combinations of R.V.s (2) Decorrelation Transform

21 Linear Combinations of R.V.s

Consider two random vectors:

$$\begin{aligned} \mathbf{X} &= [X_1,\ldots,X_m]^T \\ \mathbf{Y} &= [Y_1,\ldots,Y_n]^T \end{aligned}$$

Let each r.v. Y_i be a *linear combination* of the random variables in vector $\mathbf{X}$. Specifically, create an $n \times m$ matrix A of known real-valued constants:

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,m} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n,1} & A_{n,2} & \cdots & A_{n,m} \end{bmatrix}$$

Then the vector $\mathbf{Y}$ is given as the product of A and $\mathbf{X}$:

$$\mathbf{Y} = A\mathbf{X} \tag{10}$$

We can represent many types of systems as linear combinations. Just for some specific motivation, some examples:

- Multiple antenna transceivers, such as 802.11n. The channel gain between each pair of antennas is represented as a matrix A. Then what is received is a linear combination of what is sent. Note that A in this case would be a complex matrix.
- Secret key generation. In application assignment 4, you will come up with linear combinations in order to eliminate correlation between RSS samples.
- Finance. A mutual fund or index is a linear combination of many different stocks or equities. $A_{i,j}$ is the quantity of stock j contained in mutual fund i.
- Finite impulse response (FIR) filters, for example, for audio or image processing. Each value in matrix A would be a filter tap. Matrix A would have special structure: each row has identical values but delayed one column (shifted one element to the right).

Let's study what happens to the mean and covariance when we take a linear transformation.

Mean of a Linear Combination The expected value is a linear operator. Thus the constant matrix A can be brought outside of the expected value.

$$\mu_{\mathbf{Y}} = E_{\mathbf{Y}}[\mathbf{Y}] = E_{\mathbf{X}}[A\mathbf{X}] = AE_{\mathbf{X}}[\mathbf{X}] = A\mu_{\mathbf{X}}$$

The result? Just apply the transform A to the vector of means of each component.

Covariance of a Linear Combination Use the definition of covariance matrix to come up with the covariance of $\mathbf{Y}$.

$$\begin{aligned} C_{\mathbf{Y}} &= E_{\mathbf{Y}}\left[(\mathbf{Y}-\mu_{\mathbf{Y}})(\mathbf{Y}-\mu_{\mathbf{Y}})^T\right] \\ &= E_{\mathbf{X}}\left[(A\mathbf{X}-A\mu_{\mathbf{X}})(A\mathbf{x}-A\mu_{\mathbf{X}})^T\right] \end{aligned}$$

Now, we can factor out A from each term inside the expected value. But note that $(CD)^T = D^T C^T$ (This is a linear algebra relationship you should know).

$$\begin{aligned} C_{\mathbf{Y}} &= E_{\mathbf{X}}\left[A(\mathbf{X}-\mu_{\mathbf{X}})(A(\mathbf{x}-\mu_{\mathbf{X}}))^T\right] \\ &= E_{\mathbf{X}}\left[A(\mathbf{X}-\mu_{\mathbf{X}})(\mathbf{x}-\mu_{\mathbf{X}})^T A^T\right] \end{aligned}$$

Because the expected value is a linear operator, we again can bring the A and the A^T outside of the expected value.

$$\begin{aligned} C_{\mathbf{Y}} &= AE_{\mathbf{X}}\left[(\mathbf{X}-\mu_{\mathbf{X}})(\mathbf{x}-\mu_{\mathbf{X}})^T\right] A^T \\ &= AC_{\mathbf{X}}A^T \end{aligned} \tag{11}$$

This is the final, simple result: if $\mathbf{Y} = A\mathbf{X}$, then $C_{\mathbf{Y}} = AC_{\mathbf{X}}A^T$. In other words, if we know the mean and covariance of a R.V. $\mathbf{X}$, we can come up with any linear transform of the components

of R.V. $\mathbf{X}$, and immediately (with a couple of matrix multiplies) the mean and covariance of that new R.V.!

Example: Three Sums of i.i.d. r.v.s
Let X_1, X_2, and X_3 be independent random variables, each with mean μ and variance σ^2. Let $Y_1 = X_1$, $Y_2 = X_1 + X_2$, $Y_3 = X_1 + X_2 + X_3$. Also, let vector $\mathbf{Y} = [Y_1, Y_2, Y_3]^T$

1. Find $\mu_Y = E_{\mathbf{Y}}[\mathbf{Y}]$, the mean vector of $\mathbf{Y}$.
2. Find $C_{\mathbf{Y}}$, the covariance matrix of $\mathbf{Y}$.
3. Find the correlation coefficient ρ_{Y_2,Y_3}.

Solution: First, write down the particular $\mathbf{Y} = A\mathbf{X}$ transform here:

$$\begin{aligned} \mathbf{X} &= [X_1, X_2, X_3]^T \\ \mathbf{Y} &= [Y_1, Y_2, Y_3]^T \\ A &= \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \end{aligned}$$

Next, what are the mean and covariance of $\mathbf{X}$?

$$\begin{aligned} \mu_{\mathbf{X}} &= [\mu, \mu, \mu]^T \\ C_{\mathbf{X}} &= \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix} \end{aligned}$$

Finally, we can get to the questions:

1. $\mu_Y = A\mu_{\mathbf{X}} = [\mu, 2\mu, 3\mu]^T$.
2. $C_{\mathbf{Y}} = AC_{\mathbf{X}}A^T$:

$$C_{\mathbf{Y}} = \begin{bmatrix} \sigma^2 & 0 & 0 \\ \sigma^2 & \sigma^2 & 0 \\ \sigma^2 & \sigma^2 & \sigma^2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \sigma^2 & \sigma^2 & \sigma^2 \\ \sigma^2 & 2\sigma^2 & 2\sigma^2 \\ \sigma^2 & 2\sigma^2 & 3\sigma^2 \end{bmatrix}$$

   Note that $C_{\mathbf{Y}}$ is symmetric! Check your calculator results this way.

3. $$\rho_{Y_2,Y_3} = \frac{\mathrm{Cov}(Y_2, Y_3)}{\sqrt{\mathrm{Var}[Y_2]\,\mathrm{Var}[Y_3]}} = \frac{2\sigma^2}{\sqrt{3\sigma^2 2\sigma^2}} = \frac{2}{\sqrt{6}}$$

   Note this is between -1 and +1; another check to make sure your solution is okay.

22 Decorrelation Transformation of R.V.s

We can, if we wanted, make up an arbitrary linear combination $\mathbf{Y} = A\mathbf{X}$ of a given R.V. $\mathbf{X}$ in order to get a desired covariance matrix for $\mathbf{Y}$. One in particular which is often desired is a diagonal covariance matrix, which indicates that all pairs of components (i, j) with $i \neq j$ have zero covariance.

$$C_{\mathbf{Y}} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

In other words, all pairs of components are uncorrelated. In short, we often say the random variable $\mathbf{Y}$ is an uncorrelated random vector.

Why should we want this? Let's go back to those examples.

- Multiple antenna transceivers. The channels between each pair of antennas cause one linear transform H. We might want to come up with a linear combination of antenna elements which give us back the uncorrelated signals sent on the transmit antennas.
- Secret key generation. Again, to eliminate correlation between RSS samples over time to improve the secret key.
- Finance. Come up with mutual funds which are uncorrelated; thus achieving better *diversification*.
- Finite impulse response (FIR) filters. Come up with what is called a "whitening filter", which takes correlated noise and spreads it across the frequency spectrum.

Let's review the goal, which is to find a matrix A for a linear transformation $\mathbf{Y} = A\mathbf{X}$ which causes $C_{\mathbf{Y}}$ to be a diagonal matrix.

22.1 Singular Value Decomposition (SVD)

The solution is to use the singular value decomposition. You needed to learn this for your linear algebra class, and probably thought you'd never use it again. It says that any matrix C can be written as

$$C = U\Lambda V^T$$

where U and V are unitary matrices, and Λ is a diagonal matrix. (A unitary matrix is one that has the property that $U^T U = I$, the identity matrix. It is an orthogonal transformation, if you've heard of that.)

Covariance matrices like $C_{\mathbf{X}}$ have two properties which make things simpler for us: it is symmetric and positive semi-definite. This simplifies the result; it means that $V = U$ in the above equation, so

$$C_{\mathbf{X}} = U\Lambda U^T$$

Also, for positive semi-definite matrices, all of the diagonal elements of Λ are non-negative. The columns of U are called the eigenvectors and the diagonal elements of Λ are called the eigenvalues.

22.2 Application of SVD to Decorrelate a R.V.

Using this linear algebra result, we can come up with the desired transform. The answer is as follows:

1. Take the SVD of the known covariance matrix $C_{\mathbf{X}}$ to find U and Λ.
2. Let $A = U^T$, that is, define vector $\mathbf{Y}$ as $\mathbf{Y} = U^T\mathbf{X}$.

What happens then? Well, we know that

$$C_{\mathbf{Y}} = AC_{\mathbf{X}}A^T = U^T C_{\mathbf{X}} U$$

But since we can re-write $C_{\mathbf{X}}$ as $U\Lambda U^T$,

$$C_{\mathbf{Y}} = U^T U \Lambda U^T U$$

Since U is a unitary matrix,

$$C_{\mathbf{Y}} = I\Lambda I = \Lambda.$$

We now have a transformed R.V. $\mathbf{Y}$ with a diagonal covariance matrix, in other words, an uncorrelated random vector!

22.3 Mutual Fund Example

It is supposed to be good to "diversify" one's savings. That is, own securities which rise and fall "independently" from one another. We can read this as wanting uncorrelated securities. But individual stocks are often correlated; when banking companies fail (for example) auto companies can't sell cars, so they also go down. A financial company might offer groups of mutual funds which enable a person to "diversify" their savings as follows. Let each mutual fund be a linear combination of stocks. And, create a family of mutual funds, each which gains or loses in a way uncorrelated with the others in the family. Then, a person could diversify by owning these mutual funds.

Problem Statement Consider the five stocks shown in Figure 11. Let X_j be the percent gain (or loss if it is negative) for stock j on a given day. Assume that mutual fund i can invest in them, either by buying or short-selling the stocks. (By short-selling, we mean that you effectively buy a negative quantity of that stock – if it goes down in value, you make money). Mutual fund i can own any linear combination of the five stocks, that is,

$$Y_i = A_{i,GM}X_{GM} + A_{i,MSFT}X_{MSFT} + A_{i,GOOG}X_{GOOG} + A_{i,LLL}X_{LLL} + A_{i,HD}X_{HD}$$

If you wanted high volatility, you'd pick the linear combination so that Var $[Y_i]$ was high. If you wanted low volatility, you'd pick the linear combination so that Var $[Y_i]$ was low.

Let

$$\begin{aligned}\mathbf{X} &= [X_{GM}, X_{MSFT}, X_{GOOG}, X_{LLL}, X_{HD}]^T \\ \mathbf{Y} &= [Y_1, Y_2, Y_3, Y_4, Y_5]^T\end{aligned}$$

Problem: (a) Find the matrix A which results in a decorrelated vector $\mathbf{Y}$. (b) Identify the two "mutual funds" with the highest and lowest volatility.

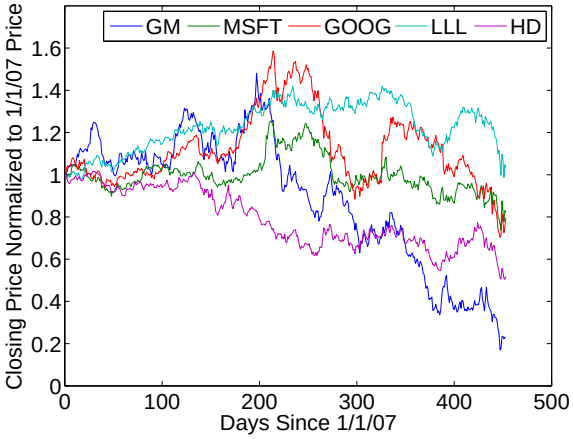


Figure 11: Normalized closing price of General Motors (GM), Microsoft (MSFT), Google (GOOG), L-3 Communications Holdings (LLL), and Home Depot (HD). Prices are normalized to the price on January 1, 2007.

Method and Solution Since we are not given the mean vector or covariance matrix of $\mathbf{X}$, we need to estimate them. We do this as follows. Denoting the realization of the random vector on day i as $\mathbf{X}_i$,

$$\begin{aligned}\hat{\mu}_{\mathbf{X}} &= \frac{1}{K}\sum_{i=1}^{K}\mathbf{X}_i \\ \hat{C}_{\mathbf{X}} &= \frac{1}{K-1}\sum_{i=1}^{K}(\mathbf{X}_i-\hat{\mu}_{\mathbf{X}})(\mathbf{X}_i-\hat{\mu}_{\mathbf{X}})^T\end{aligned}$$

We find (using Matlab) that

$$\hat{C}_{\mathbf{X}} = \begin{bmatrix} 0.00190 & 0.00045 & 0.00047 & 0.00031 & 0.00059 \\ 0.00045 & 0.00044 & 0.00031 & 0.00019 & 0.00023 \\ 0.00047 & 0.00031 & 0.00061 & 0.00017 & 0.00023 \\ 0.00031 & 0.00019 & 0.00017 & 0.00025 & 0.00018 \\ 0.00059 & 0.00023 & 0.00023 & 0.00018 & 0.00050 \end{bmatrix}$$

Next, we compute the SVD. In Matlab, this is computed using `[U, Lambda, V] = svd(C_X)`. We find that:

$$U = V = \begin{bmatrix} -0.831 & 0.500 & -0.237 & -0.054 & 0.009 \\ -0.276 & -0.457 & 0.142 & -0.680 & -0.482 \\ -0.303 & -0.695 & -0.493 & 0.414 & 0.104 \\ -0.184 & -0.207 & 0.312 & -0.323 & 0.849 \\ -0.327 & -0.124 & 0.764 & 0.509 & -0.188 \end{bmatrix}$$

$$\Lambda = \begin{bmatrix} 0.00253 & 0 & 0 & 0 & 0 \\ 0 & 0.00058 & 0 & 0 & 0 \\ 0 & 0 & 0.00028 & 0 & 0 \\ 0 & 0 & 0 & 0.00020 & 0 \\ 0 & 0 & 0 & 0 & 0.00013 \end{bmatrix}$$

Now, $U^T\mathbf{X}$ gives us the decorrelated vector $\mathbf{Y}$. The first "mutual fund" thus corresponds to the first column of U. It short sells a little bit of each stock, but mostly the first (GM). It's daily percentage change has the highest variance of any of the "mutual funds". The fifth (last) mutual fund buys LLL and a little bit of GOOG, and short sells a lot of MSFT and a little of HD. It variance is much smaller than the variance of the first mutual fund: about 1/20 of the variance. In terms of standard deviation, mutual fund 1 has $\sqrt{0.00253} = 0.050$ and mutual fund 2 has $\sqrt{0.00013} = 0.011$. The daily percentage changes are much smaller with the fifth compared to the first mutual fund.

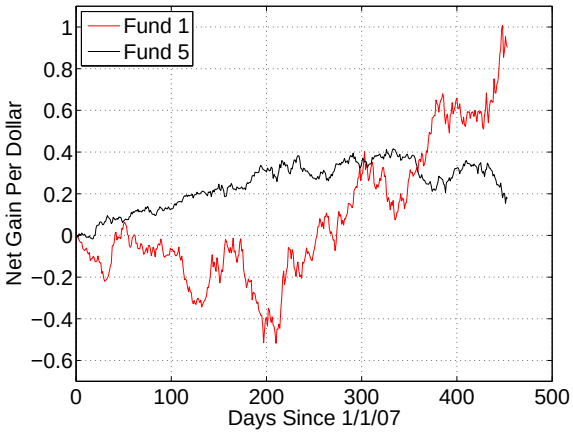


Figure 12: Comparison of two "mutual funds" of stocks, composed of two different linear combinations of General Motors (GM), Microsoft (MSFT), Google (GOOG), L-3 Communications Holdings (LLL), and Home Depot (HD) stocks. Vertical axis shows net gain or loss per dollar invested on 1/1/07.

These two mutual funds have net gains, over the entire (almost) 2 years, shown in Figure 12. The figure is showing net gain per dollar invested, so for example if you put \$100 in fund 1, you'd have about \$200 at the end of the period. The main results are

- The daily volatility is higher in fund 1 than in fund 5.
- Although fund 1 results in a higher net gain, we did not design the funds for highest gain (we assume that you can't predict future gains and losses from past gains and losses).
- The daily gains and losses seem uncorrelated between the two funds.

22.4 Linear Transforms of R.V.s Continued

Example: Average and Difference of Two i.i.d. R.V.s

We represent the arrival of two people for a meeting as r.v.s X_1 and X_2. Let $\mathbf{X} = [X_1, X_2]^T$. Assume that the two people arrive independently, with the same variance σ^2 and mean μ. Consider the average arrival time $Y_1 = (X_1 + X_2)/2$, and the difference between the arrival times, $Y_2 = X_1 - X_2$. The latter is a wait time that one person must wait before the second person arrives. Show that the average time and the difference between the times are uncorrelated.

You can take these steps to solve this problem:

1. Let $\mathbf{Y} = [Y_1, Y_2]^T$. What is the transform matrix A in the relation $\mathbf{Y} = A\mathbf{X}$?
2. What is the mean matrix $\mu_{\mathbf{Y}} = E_{\mathbf{Y}}[\mathbf{Y}]$? (Note this isn't really needed to answer the question, but is good practice anyway.)
3. What is the covariance matrix of $\mathbf{Y}$?
4. How does the covariance matrix show that the two are uncorrelated?

Lecture 15

Today: (1) Random Process Intro (2) Binomial R.P. (3) Poisson R.P.

23 Random Processes

This starts into Chapter 10, 'Stochastic' Processes. As Y&G says, "The word stochastic means random." So I prefer 'Random Processes'. We've covered Random Vectors, which have many random variables. So what's new?

- Before we had a few random variables, X_1, X_2, X_3. Now we have possibly infinitely many: $X_1, X_2, \ldots$.
- In addition, we may not be taking samples - we may have a continuously changing random variable, indexed by time t. We'll denote this as $X(t)$.

Def'n: *Random Process*
A random process $X(t)$ consists of an experiment with a probability measure $P[\cdot]$, a sample space S, and a function that assigns a time (or space) function $x(t, s)$ to each outcome s in the sample space.

Recall that we used S to denote the event space, and every $s \in S$ is a possible 'way' that the outcome could occur.

23.1 Continuous and Discrete-Time

Types of Random Processes: A random process (R.P.) can be either

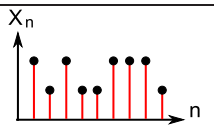
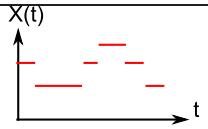
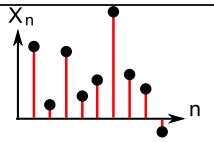
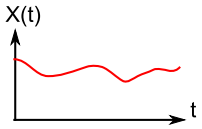
1. **Discrete-time**: Samples are taken at particular time instants, for example, $t_n = nT$ where n is an integer and T is the sampling period. In this case, rather than referring to $X(t_n)$, we abbreviate it as X_n. (This matches exactly our previous notation.) In this case, we also call it a *random sequence*.
2. **Continuous-time**: Uncountably-infinite values exist, for example, for $t \in (0, \infty)$.

Types of Random Processes: A random process (R.P.) can be still be

1. **Discrete-valued**: The sample space S_X is countable. That is, each value in the R.P. is a discrete r.v. (For example, our R.P. can only take integer values, or we allow a finite number of decimal places.)

2. **Continuous-valued**: The sample space S_X is uncountably infinite.

Draw a example plot here of each of the following:

	Discrete-Time	Continuous-Time
Discrete-Valued		
Continuous-Valued		

23.2 Examples

For each of these examples, say whether this is continuous/discrete time, and continuous/discrete valued (there may be multiple 'right' answers):

- **Stock Values**: Today, we invest \$1000 in one stock. Let $X(t)$ be the random process of the value of our investment. Based on the particular stock s that we picked, at time t we could measure $x(t, s)$.

- **RSS Measurements**: The measured received signal strength is a function of time and the node doing the measurement. Define $X(t, i)$ as the RSS measured at time t at node $i \in \{a, b\}$.

- **Temperature over time**: I put a wireless temperature sensor outside the MEB and record the temperature on my laptop, $W(t)$. Based on the 'weather' s we might record $W(t, s)$.

- **Temperature over space I**: As I drive, I record the temperature that my car reports to me $W(t)$. But this is also $W(\mathbf{z})$, since my position is a function of time.

- **Temperature over space II**: I get from the weather forecasters the temperature all across Utah, $W(\mathbf{z})$.

- **Imaging**: I take a photo $Q(\mathbf{z})$.

- **Counting Traffic**: We monitor a router on the Internet and count the number of packets which have passed through since we started, a R.P. we might call $N(t)$. This is affected by the traffic offered by people, and where those people are, and what they're doing on the Internet, what we might call $s \in S$.

- **Radio Signal**: We measure the RF signal at a receiver tuned to an AM radio station, $A(t)$.

23.3 Random variables from random processes

Section 10.3 in Y&G.

- **Discrete-time**: We sample our signal $X(t)$ at times t_n, resulting in samples $\{X(t_n)\}$. We might have negative and positive n, if we've been sampling for a long time: $n = \ldots, -1, 0, 1, \ldots$, or just one sided: $n = 0, 1, \ldots$.
- **Continuous-time**: We have a continuous-time signal, $X(t)$, but we still might be interested in how the signal compares at two different times t and s, for example, so we could ask questions about $X(t)$ and $X(s)$, *e.g.*, are they independent, correlated, or what is their joint distribution.

23.4 i.i.d. Random Sequences

An i.i.d. random sequence is just a random sequence in which each X_i is from the same marginal distribution, i.e.,

$$f_{X_0}(x_0) = f_{X_1}(x_1) = \cdots = f_{X_n}(x_n) = f_X(x)$$

Example: i.i.d. Gaussian Sequence
The resistance, X_i of the ith resistor coming off the assembly line is measured for each resistor i. We model X_i as i.i.d. Gaussian with mean R and variance σ^2. Find the joint pdf of R.V.

$$\mathbf{X} = [X_1, X_2, \ldots, X_n]$$

Answer: Since $\{X_i\}$ are i.i.d., their joint pdf is the product of their marginal pdfs:

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \prod_{i=1}^{n} f_{X_i}(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-R)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-R)^2} \end{aligned}$$

Example: i.i.d. Bernoulli Sequence
Let X_n be a sequence of i.i.d Bernoulli random variables, each one is equal to 1 with probability p and equal to zero with probability $1-p$. This is so important it is defined as the 'Bernoulli process'. Examples: Buy a lottery ticket every day and let X_n be your success on day n.

These are all models. Don't take these independence assumptions for granted! There are often dependencies between random variables. But simple models allow for engineering analysis...

23.5 Counting Random Processes

What if we take the Bernoulli process, starting at X_0, and run it through a summer? This is a counting process. Graphic:

iid Bernoulli r.v.s $X_i \longrightarrow \sum_{i=0}^{n} \longrightarrow K_n$

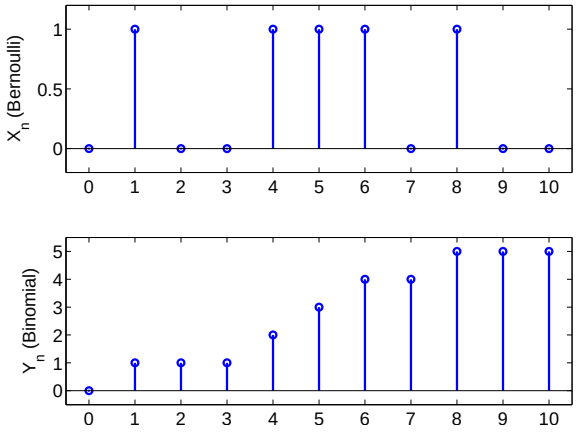


Figure 13: An example of a Bernoulli process, $X_0, \ldots, X_{10}$ and the resulting Binomial process, $Y_0, \ldots, Y_{10}$.

What is the process K_n? What is its pmf?

$$P_{K_n}(k_n) = P\left[K_n = k_n\right] = P\left[\sum_{i=0}^{n} X_i = k_n\right] = P\left[\text{there were } k_n \text{ successes out of } n \text{ trials}\right]$$

This is a Binomial r.v., like what we've seen previously.

$$P_{K_n}(k_n) = \binom{n}{k_n} p^{k_n} (1-p)^{n-k_n}$$

Overall, this is called a 'Binomial process'. It is a counting process.

Def'n: *Discrete-time counting process*
A (discrete time) random process X_n is a counting process if it has these three properties:

- X_n is defined as zero for $n < 0$.
- X_n is integer-valued for all n (discrete r.v.s).
- X_n is non-decreasing with n.

Def'n: *Continuous-time counting process*
A (continuous time) random process $X(t)$ is a counting process if it has these three properties:

- $X(t)$ is defined as zero for $t < 0$.
- $X(t)$ is integer-valued for all t (discrete r.v.s).
- $X(t)$ is non-decreasing with time t.

23.6 Derivation of Poisson pmf

Now, time is continuous. I want to know, how many arrivals have happened in time T. Eg, how many packets have been transmitted in my network. Here's how but I could translate it to a Bernoulli process question:

- Divide time into intervals duration T/n (eg, $T/n = 1$ minute).
- Define the arrival of one packet in an interval as a 'success' in that interval.
- Assume that success in each interval is i.i.d.
- Sum the Bernoulli R.P. to get a Binomial process.

Problem: Unless the time interval T/n is really small, you might actually have **more than one packet arrive** in each interval. Since Binomial can only account for 1 or 0, this doesn't represent your total number of packets exactly.
Short story: As the time interval goes to zero, the probability of more than one arrival in that interval becomes negligible, so it becomes an accurate representation of the experiment. And, as the time interval goes to zero, the limit of the Binomial pmf $P_{K_n}(k_n)$ approaches the Poisson pmf, for the continuous time r.v. $K(t)$:

$$P_{K(t)}(k) = \frac{(\lambda T)^k e^{-\lambda R}}{k!}$$

23.6.1 Let time interval go to zero

Long Story: Let's define $K(T)$ to be the number of packets which have arrived by time T for the real, continuous time process. Let's define Y_n as our Binomial approximation, in which T is divided into n identical time bins. Let's show what happens as we divide T into more and more time bins, *i.e.*, as $n \to \infty$. In this case, we should get more and more exact.

Each time bin has width T/n. If the average arrival rate is λ, then the probability of a success in a really small time bin is $p = \lambda T/n$. Thus,

$$P_{Y_n}(k) = \binom{n}{k}(p)^k(1-p)^{n-k} = \binom{n}{k}(\lambda T/n)^k(1-\lambda T/n)^{n-k}$$

Let's write this out:

$$\begin{aligned} P_{Y_n}(k) &= \frac{n(n-1)\cdots(n-k+1)}{k!}\left(\frac{\lambda T}{n}\right)^k\left(1-\frac{\lambda T}{n}\right)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{n^k}\frac{(\lambda T)^k}{k!}\left(1-\frac{\lambda T}{n}\right)^{n-k} \end{aligned}$$

Now, lets take the limit as $n \to \infty$ of each term.

$$\lim_{n\to\infty}\frac{n(n-1)\cdots(n-k+1)}{n^k} = \lim_{n\to\infty}\frac{n}{n}\frac{n-1}{n}\cdots\frac{n-k+1}{n} = 1(1)\cdots(1)$$

For the right-most term,

$$\lim_{n\to\infty}\left(1-\frac{\lambda T}{n}\right)^{n-k} = \frac{\left(1-\frac{\lambda T}{n}\right)^n}{\left(1-\frac{\lambda T}{n}\right)^k}$$

But the limit of the denominator is 1, while $\lim_{n\to\infty}(1-\frac{\lambda T}{n})^n$ is equal to $e^{-\lambda T}$ (See a table of limits). Thus

$$P_{K(T)}(k) = \lim_{n\to\infty} P_{Y_n}(k) = \frac{(\lambda T)^k}{k!}e^{-\lambda T}$$

Which shows that $P_{K(T)}(k)$ is Poisson.

Lecture 16

Today: Poisson Processes: (1) Indep. Increments, (2) Exponential Interarrivals

24 Poisson Process

The left hand side of this table covers discrete-time Bernoulli and Binomial R.P.s, which we have covered. We also mentioned the Geometric pmf in the first part of this course. Now, we are covering the right-hand column, which answer the same questions but for continuous-time R.P.s.

	Discrete Time	Continuous-Time
What is this counting process called?	"Bernoulli"	"Poisson"
How long until my first arrival/success?	Geometric p.m.f.	Exponential p.d.f.
After a set amount of time, how many arrivals/successes have I had?	Binomial p.m.f.	Poisson p.m.f.

24.1 Last Time

This is the marginal pmf of Y_n during a Binomial counting process:

$$P_{Y_n}(k_n) = \binom{n}{k_n} p^{k_n}(1-p)^{n-k_n}$$

24.2 Independent Increments Property

In the Binomial process, Y_n, we derived the pmf by assuming that we had independent Bernoulli trials at each trial i. In the Poisson process,

- If we consider any two non-overlapping intervals, they are independent. For example, consider the number of arrivals in the intervals $(0, T_1)$ and (T_2, T_3), where $0 \le T_1 \le T_2 \le T_3$. Then the numbers of arrivals in the two intervals is independent.

Example: What is the joint pmf of K_1, K_2, the number of arrivals in the two above intervals?
Let $\Delta_1 = T_1$ and $\Delta_2 = T_3 - T_2$. Then

$$P_{K(\Delta_1)}(k_1) = \frac{(\lambda\Delta_1)^{k_1}}{k_1!}e^{-\lambda\Delta_1} \quad P_{K(\Delta_2)}(k_2) = \frac{(\lambda\Delta_2)^{k_2}}{k_2!}e^{-\lambda\Delta_2}$$

Since they are independent, the joint pmf is just the product of the two.

24.3 Exponential Inter-arrivals Property

Theorem: For a Poisson process with rate λ, the time until the first arrival, T_1, is an exponential r.v. with parameter λ.
Proof: First, consider the probability that $T_1 > t_1$:

$$P[T_1 > t_1] = P[K(t_1) = 0] = P_{K(t_1)}(0) = \frac{(\lambda t_1)^0}{0!} e^{-\lambda t_1} = e^{-\lambda t_1}$$

So, the CDF of T_1 is 1 minus this probability,

$$P[T_1 \leq t_1] = 1 - e^{-\lambda t_1},$$

for $t_1 > 0$. Then the pdf is the derivative of the CDF,

$$f_{T_1}(t_1) = \frac{\partial}{\partial t_1} P[T_1 \leq t_1] = \begin{cases} \lambda e^{-\lambda t_1}, & t_1 > 0 \\ 0, & o.w. \end{cases}$$

In general, start the clock at any particular time – it doesn't matter, since each non-overlapping interval is independent.

Example: What is the conditional probability that $T_1 > t + \delta$ given that $T_1 > t$?
First compute $P[T_1 > t]$:

$$\begin{aligned} P[T_1 > t] &= \int_{t_1=t}^{\infty} \lambda e^{-\lambda t_1} dt_1 \\ &= \left[-e^{-\lambda t_1} \right|_{t_1=t}^{\infty} \\ &= e^{-\lambda t} - 0 = e^{-\lambda t} \end{aligned}$$

Now compute the conditional probability:

$$P[T_1 > t + \delta | T_1 > t] = \frac{P[\{T_1 > t + \delta\} \cap \{T_1 > t\}]}{P[T_1 > t]}$$

We just computed the denominator. What is the set in the numerator? You can see that $\{T_1 > t\}$ is redundant. If $T_1 > t + \delta$, then it must be also $T_1 > t$, so

$$P[T_1 > t + \delta | T_1 > t] = \frac{P[T_1 > t + \delta]}{e^{-\lambda t}}$$

What is the numerator? We've already derived it for t.

$$P[T_1 > t + \delta | T_1 > t] = \frac{e^{-\lambda(t+\delta)}}{e^{-\lambda t}} = \frac{e^{-\lambda t} e^{-\lambda \delta}}{e^{-\lambda t}} = e^{-\lambda \delta}$$

What is the conditional CDF?

$$P[T_1 < t + \delta | T_1 > t] = 1 - e^{-\lambda \delta}$$

What is the conditional pdf?

$$f_{T_1|T_1>t}(t_1) = \begin{cases} \lambda e^{-\lambda t_1}, & t_1 > t \\ 0, & o.w. \end{cases}$$

This is **not a function of** t, the time we already know that the first arrival did not arrive. This is the same analysis you did for the HW problem on waiting for a bus that arrives with an exponential dist'n.

24.4 Inter-arrivals

Apply this to T_2. If you are given that $T_1 = t_1$, what is the pdf of T_2? It is exponential, not a function of t_1.

24.5 Examples

Example: What is the pdf of the time of the 2nd arrival?
Since T_1 is the time of the 1st arrival, and T_2 is the time between the first and 2nd arrival, the 2nd arrival arrives at $T_1 + T_2$. Note that T_1 and T_2 are independent, and are both exponential with parameter λ. Thus the pdf of $T_1 + T_2$ is just the convolution of the exponential pdf with itself. We have done this problem before in Lecture 12.

$$f_{T_1+T_2}(y) = \begin{cases} \lambda^2 y e^{-\lambda y}, & y > 0 \\ 0, & o.w. \end{cases}$$

Example: ALOHA Packet Radio

Wikipedia: the ALOHA network was created at the University of Hawaii in 1970 under the leadership of Norman Abramson and Franklin Kuo, with DARPA funding. The ALOHA network used packet radio to allow people in different locations to access the main computer systems.

Assumptions of ALOHA net:

1. There are many computers, each with a packet radio, silent unless it has a packet that needs to be sent. Packets have duration T.

2. During a packet transmission, if any other computer sends a packet that overlaps in time, it is called a collision.

3. Packets are offered at (total) rate λ, and are a Poisson process.

Questions:

1. Given that computer 1 transmits a packet at time t, what is the probability that it is received, $P_{R|T}$?

2. Define the success rate as $R = \lambda P_{R|T}$. What λ should we require to maximize R?

Solution: This is the probability that no collision occurs. Another packet collides if it was sent any time between $t - T$ and $t + T$:

$$P_{R|T} = P\left[\text{ Success } | \text{ Transmit at } t\right] = P\left[T_1 > 2T\right] = e^{-\lambda 2T}$$

Second part: $R = \lambda e^{-\lambda 2T}$. To maximize R w.r.t. λ, differentiate and set to zero:

$$0 \quad = \quad \frac{\partial}{\partial \lambda} R = \frac{\partial}{\partial \lambda} \lambda e^{-\lambda 2T} = e^{-\lambda 2T} - 2\lambda T e^{-\lambda 2T} = e^{-\lambda 2T}(1 - 2\lambda T)$$

Thus $\lambda = 1/(2T)$ provides the best rate R. Furthermore, the rate R at this value of λ is

$$R = \frac{1}{2T} e^{-1} \approx 0.184 \frac{1}{T}$$

If packets were stacked one right next to the other, the maximum packet rate could be $\frac{1}{T}$. The highest success rate of this system (ALOHA) is 18.4% of that! This is called the multiple-access channel (MAC) problem – we can't get good throughput when we just transmit a packet whenever we want.

Lecture 17

Today: Random Processes: (1) Autocorrelation & Autocovariance, (2) Wide Sense Stationarity, (3) PSD

As a brief overview of the rest of the semester.

- We're going to talk about autocovariance (and autocorrelation, a similar topic), that is, the covariance of a random signal with itself at a later time. The autocovariance tells us something about our ability to predict future values (k in advance) of Y_k. The higher $C_Y[k]$ is, the more the two values separated by k can be predicted.

- Autocorrelation is critical to the next topic: What does the random signal look like in the frequency domain? More specifically, what will it look like in a spectrum analyzer (set to average). This is important when there are specific limits on the bandwidth of the signal (imposed, for example, by the FCC) and you must design the process in order to meet those limits.

- We can analyze what happens to the spectrum of a random process when we pass it through a filter. Filters are everywhere in ECE, so this is an important tool.

- We'll also discuss new random processes, including Gaussian random processes, and Markov chains. Markov chains are particularly useful in the analysis of many discrete-time engineered systems, for example, computer programs, networking protocols, and networks like the Internet.

25 Expectation of Random Processes

25.1 Expected Value and Correlation

Def'n: *Expected Value of a Random Process*
The expected value of continuous time random process $X(t)$ is the deterministic function

$$\mu_X(t) = E_{X(t)}\left[X(t)\right]$$

for the discrete-time random process X_n,

$$\mu_X[n] = E_{X_n}\left[X_n\right]$$

Example: What is the expected value of a Poisson process?
Let Poisson process $X(t)$ have arrival rate λ. We know that

$$\begin{aligned}
\mu_X(t) &= E_{X(t)}\left[X(t)\right] = \sum_{x=0}^{\infty} x\frac{(\lambda t)^x}{x!}e^{-\lambda t} \\
&= e^{-\lambda t}\sum_{x=0}^{\infty} x\frac{(\lambda t)^x}{x!} \\
&= e^{-\lambda t}\sum_{x=1}^{\infty} \frac{(\lambda t)^x}{(x-1)!} \\
&= (\lambda t)e^{-\lambda t}\sum_{x=1}^{\infty} \frac{(\lambda t)^{x-1}}{(x-1)!} \\
&= (\lambda t)e^{-\lambda t}\sum_{y=0}^{\infty} \frac{(\lambda t)^{y}}{(y)!} \\
&= (\lambda t)e^{-\lambda t}e^{\lambda t} = \lambda t \qquad (12)
\end{aligned}$$

This is how we intuitively started deriving the Poisson process - we said that it is the process in which on average we have λ arrivals per unit time. Thus we'd certainly expect to see λt arrivals after a time duration t.

Example: What is the expected value of X_n, the number of successes in a Bernoulli process after n trials?
We know that X_n is Binomial, with mean np. This is the mean function, if we consider it to be a function of n: $\mu_X[n] = E_X\left[X_n\right] = np$.

25.2 Autocovariance and Autocorrelation

These next two definitions are the most critical concepts of the rest of the semester. Generally, for a time-varying signal, we often want to know two things:

- How to predict its future value. (It is not deterministic.) We will use the 'autocovariance' to determine this.

- How much 'power' is in the signal (and how much power within particular frequency bands). We will use the 'autocorrelation' to determine this.

Def'n: *Autocovariance*
The autocovariance function of the continuous time random process $X(t)$ is

$$C_X(t,\tau) = \text{Cov}\left(X(t), X(t+\tau)\right)$$

For a discrete-time random process, X_n, it is

$$C_X[m,k] = \text{Cov}\left(X_m, X_{m+k}\right)$$

Note that $C_X(t,0) = \text{Var}_X\left[X(t)\right]$, and $C_X[m,0] = \text{Var}_X\left[X_m\right]$.

Def'n: *Autocorrelation*
The autocorrelation function of the continuous time random process $X(t)$ is

$$R_X(t,\tau) = E_X\left[X(t)X(t+\tau)\right]$$

For a discrete-time random process, X_n, it is

$$R_X[m,k] = E_X\left[X_m X_{m+k}\right]$$

These two definitions are related by:

$$\begin{aligned} C_X(t,\tau) &= R_X(t,\tau) - \mu_X(t)\mu_X(t+\tau) \\ C_X[m,k] &= R_X[m,k] - \mu_X[m]\mu_X[m+k] \end{aligned}$$

Example: Poisson R.P. Autocorrelation and Autocovariance
What is $R_X(t,\tau)$ and $C_X(t,\tau)$ for a Poisson R.P. $X(t)$?

$$R_X(t,\tau) = E_X\left[X(t)X(t+\tau)\right]$$

Look at this very carefully! It is a product of two samples of the Poisson R.P. Are $X(t)$ and $X(t+\tau)$ independent?

- NO! They are not – $X(t)$ counts the arrivals between time 0 and time t. $X(t+\tau)$ counts the arrivals between time 0 and time $t+\tau$. The intervals $(0,t)$ and $(0,t+\tau)$ do overlap! So they are not the "non-overlapping intervals" required to apply the independent increments property.

See Figure 14. But, we can re-work the product above to include two terms which allow us to apply the independent increments property. WATCH THIS:

$$R_X(t,\tau) = E_X\left[X(t)[X(t+\tau) - X(t) + X(t)]\right]$$

Note that because of the independent increments property, $[X(t+\tau) - X(t)]$ is independent of $X(t)$.

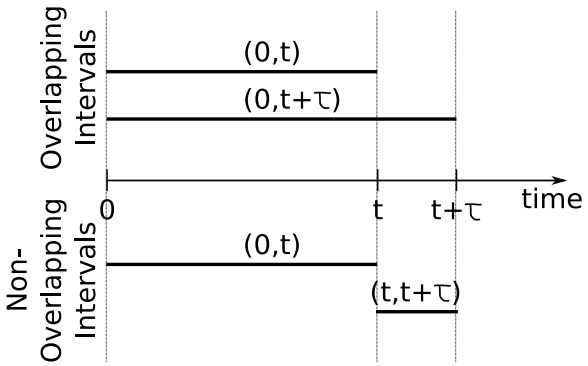


Figure 14: Comparison of non-overlapping and overlapping intervals. Converting problems which deal with overlapping intervals of time to a form which is in terms of non-overlapping intervals is a key method for analysis.

$$R_X(t,\tau) = E_X\left[X(t)[X(t+\tau)-X(t)]\right] + E_X\left[X(t)X(t)\right] \tag{13}$$

See what we did? The first expected value is now a product of two r.v.s which correspond to non-overlapping intervals. LEARN THIS TRICK! This one trick, converting products with non-independent increments to products of independent increments, will help you solve a lot of of the autocovariance problems you'll see.

$$\begin{aligned} &= E_X\left[X(t)\right] E_X\left[X(t+\tau)-X(t)\right] + E_X\left[X^2(t)\right] \\ &= \mu_X(t)\mu_X(\tau) + E_X\left[X^2(t)\right] \\ &= \mu_X(t)\mu_X(\tau) + \mathrm{Var}_X\left[X(t)\right] + \left[\mu_X(t)\right]^2 \end{aligned} \tag{14}$$

We will not derive it right now, but $\mathrm{Var}_X\left[X(t)\right] = \lambda t$. So

$$R_X(t,\tau) = \lambda t \lambda\tau + \lambda t + \lambda^2 t^2 = \lambda t\left[\lambda(t+\tau)+1\right]$$

Then

$$\begin{aligned} C_X(t,\tau) &= R_X(t,\tau) - \mu_X(t)\mu_X(t+\tau) \\ &= \lambda t\left[\lambda(t+\tau)+1\right] - \mu_X(t)\mu_X(t+\tau) \\ &= \lambda t\left[\lambda(t+\tau)+1\right] - \lambda t\lambda(t+\tau) \\ &= \lambda t \end{aligned}$$

The autocovariance at time t is the same as the variance of $X(t)$! We will revisit this later, but all random processes which have the independent increments property exhibit this trait.

Example: Example 10.22 from Y&G
We select a phase Θ to be uniform on $[0, 2\pi)$. Define:

$$X(t) = A\cos(2\pi f_c t + \Theta)$$

What are the mean and covariance functions of $X(t)$?

Solution: Note that we need two facts for this derivation. First, for any integer k, real valued α,

$$E_\Theta\left[cos(\alpha + k\Theta)\right] = 0$$

Also, we'll need the identity $\cos A \cos B = \frac{1}{2}[\cos(A-B) + \cos(A+B)]$.

Because of the first fact,

$$\mu_X(t) = 0$$

From the 2nd fact,

$$\begin{aligned} C_X(t,\tau) &= R_X(t,\tau) = E_X\left[A\cos(2\pi f_c t + \Theta)A\cos(2\pi f_c(t+\tau) + \Theta)\right] \\ &= \frac{A^2}{2}E_X\left[\cos(2\pi f_c\tau) + \cos(2\pi f_c(2t+\tau) + 2\Theta)\right] \\ &= \frac{A^2}{2}E_X\left[\cos(2\pi f_c\tau) + \cos(2\pi f_c(2t+\tau) + 2\Theta)\right] \\ &= \frac{A^2}{2}E_X\left[\cos(2\pi f_c\tau)\right] \\ &= \frac{A^2}{2}\cos(2\pi f_c\tau) \end{aligned}$$

25.3 Wide Sense Stationary

Section 10.10. We're going to skip 10.9 because it is a poorly-written section, and it ends up being confusing. Besides, it does not have many practical applications.

Def'n: *Wide Sense Stationary (WSS)*
A R.P. $X(t)$ is wide-sense stationary if its mean function and covariance function (or correlation function) is **not** a function of t, *i.e.*,

$$E_X\left[X(t)\right] = \mu_X, \quad \text{and} \quad R_X(t,\tau) = R_X(s,\tau) \doteq R_X(\tau)$$

and

$$E_X\left[X_n\right] = \mu_X, \quad \text{and} \quad R_X[m,k] = R_X[n,k] \doteq R_X[k]$$

Intuitive Meaning: The mean does not change over time. The covariance and correlation of a signal with itself at a time delay does not change over time.

Note that R_X and μ_X not a function of t also means that C_X is not a function of t.

Example: WSS of past two examples
Are the Poisson process and/or Random phase sine wave process WSS? Solution:

- Poisson process: no.
- Random phase sine wave process: yes.

Processes that we'll see on a circuit, for example, we want to know that they have certain properties. WSS is one of them.

25.3.1 Properties of a WSS Signal

- $R_X(0) \geq 0$.
- $R_X(0)$ ($R_X[0]$) is the average power of $X(t)$ (X_n). Think of $X(t)$ as being a voltage or current signal, going through a 1 Ω resistor. $R_X(0)$ is the power dissipated in the resistor.
- $R_X(0) \geq |R_X(\tau)|$.
- $R_X(\tau) = R_X(-\tau)$.

For example, what is the power of the random phase sine wave process? Answer: $\frac{A^2}{2}$, where A was the amplitude of the sinusoid.

26 Power Spectral Density of a WSS Signal

Now we make specific our talk of the spectral characteristics of a random signal. What would happen if we looked at our WSS random signal on a spectrum analyzer? (That is, set it to average)

Def'n: *Fourier Transform*
Functions $g(t)$ and $G(f)$ are a Fourier transform pair if

$$\begin{aligned} G(f) &= \int_{t=-\infty}^{\infty} g(t)e^{-j2\pi ft}dt \\ g(t) &= \int_{f=-\infty}^{\infty} G(f)e^{j2\pi ft}df \end{aligned}$$

PLEASE SEE TABLE 11.1 ON PAGE 413.

We don't directly look at the FT on the spectrum analyzer; we look at the *power* in the FT. Power for a complex value G is $S = |G|^2 = G^* \cdot G$.

Theorem: If $X(t)$ is a WSS random process, then $R_X(\tau)$ and $S_X(f)$ are a Fourier transform pair, where $S_X(f)$ is the power spectral density (PSD) function of $X(t)$:

$$\begin{aligned} S_X(f) &= \int_{\tau=-\infty}^{\infty} R_X(\tau)e^{-j2\pi f\tau}d\tau \\ R_X(\tau) &= \int_{f=-\infty}^{\infty} S_X(f)e^{j2\pi f\tau}df \end{aligned}$$

Proof: Omitted: see Y&G p. 414-415. This theorem is called the Wiener-Khintchine theorem, please use this name whenever talking to non-ECE friends to convince them that this is hard.

Short story: You can't just take the Fourier transform of $X(t)$ to see the averaged spectrum analyzer output when $X(t)$ is a random process. But, with the work we've done so far this semester, you can come up with an autocorrelation function $R_X(\tau)$, which is a non-random function. Then, the FT of the autocorrelation function shows what the average power vs. frequency will be on the spectrum analyzer.

Lecture 18

Today: (1) Exam Return, (2) Review of Lecture 17

27 Review of Lecture 17

Def'n: *Expected Value of a Random Process*
The expected value of continuous time random process $X(t)$ is the deterministic function

$$\mu_X(t) = E_{X(t)}\left[X(t)\right]$$

for the discrete-time random process X_n,

$$\mu_X[n] = E_{X_n}\left[X_n\right]$$

Def'n: *Autocovariance*
The autocovariance function of the continuous time random process $X(t)$ is

$$C_X(t,\tau) = \mathrm{Cov}\left(X(t), X(t+\tau)\right)$$

For a discrete-time random process, X_n, it is

$$C_X[m,k] = \mathrm{Cov}\left(X_m, X_{m+k}\right)$$

Def'n: *Autocorrelation*
The autocorrelation function of the continuous time random process $X(t)$ is

$$R_X(t,\tau) = E_X\left[X(t)X(t+\tau)\right]$$

For a discrete-time random process, X_n, it is

$$R_X[m,k] = E_X\left[X_m X_{m+k}\right]$$

Def'n: *Wide Sense Stationary (WSS)*
A R.P. $X(t)$ is wide-sense stationary if its mean function and covariance function (or correlation function) is **not** a function of t, *i.e.*,

$$E_X\left[X(t)\right] = \mu_X, \quad \text{and} \quad R_X(t,\tau) = R_X(s,\tau) \doteq R_X(\tau)$$

and

$$E_X\left[X_n\right] = \mu_X, \quad \text{and} \quad R_X[m,k] = R_X[n,k] \doteq R_X[k]$$

Meaning: The mean does not change over time. The covariance and correlation of a signal with itself at a time delay does not change over time.

We did two examples:

1. Random phase sinusoid process.
2. Poisson process.

Example: Redundant Bernoulli Trials
Let $X_1, X_2, \ldots$ be a sequence of Bernoulli trials with success probability p. Then let $Y_2, Y_3, \ldots$ be a process which describes if the number of successes in the past two trials, *i.e.*,

$$Y_k = X_k + X_{k-1}$$

for $k = 2, 3, \ldots$. (a) What is the mean of Y_k? (b) What is the autocovariance and autocorrelation of Y_k? (c) Is it a WSS R.P.?

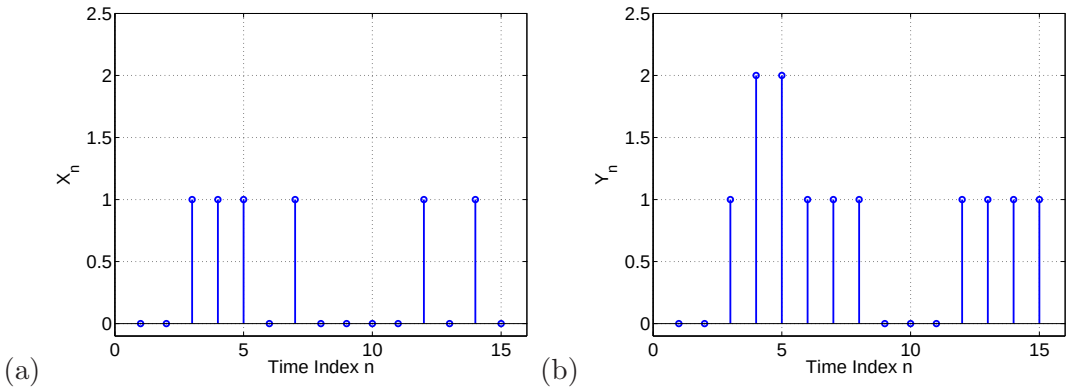


Figure 15: Realization of the (a) Bernoulli and (b) filtered Bernoulli process example covered in previous lecture.

Solution:

$$\mu_Y[n] = E_{Y_n}[Y_n] = E_{Y_n}[X_n + X_{n-1}] = 2p$$

Then, for the autocorrelation,

$$\begin{aligned} R_Y[m,k] &= E_Y[Y_m Y_{m+k}] = E_Y[(X_m + X_{m-1})(X_{m+k} + X_{m+k-1})] \\ &= E_X[X_m X_{m+k} + X_m X_{m+k-1} + X_{m-1} X_{m+k} + X_{m-1} X_{m+k-1}] \\ &= E_X[X_m X_{m+k} + X_m X_{m+k-1} + X_{m-1} X_{m+k} + X_{m-1} X_{m+k-1}] \end{aligned}$$

Let's take the case when $k = 0$:

$$\begin{aligned} R_Y[m,0] &= E_X\left[X_m^2 + X_m X_{m-1} + X_{m-1} X_m + X_{m-1}^2\right] \\ &= E_X\left[X_m^2\right] + 2E_X[X_m X_{m-1}] + E_X\left[X_{m-1}^2\right] \\ &= 2[p(1-p) + p^2] + 2p^2 = 2p + 2p^2 \end{aligned} \tag{15}$$

Let's take the case when $k = 1$:

$$\begin{aligned} R_Y[m,1] &= E_X[X_m X_{m+1} + X_m X_m + X_{m-1} X_{m+1} + X_{m-1} X_m] \\ &= p^2 + p + p^2 + p^2 = p + 3p^2 \end{aligned}$$

This would be the same for $k = -1$:

$$\begin{aligned} R_Y[m,1] &= E_X[X_m X_{m-1} + X_m X_{m-2} + X_{m-1} X_{m-1} + X_{m-1} X_{m-2}] \\ &= p^2 + p + p^2 + p^2 = p + 3p^2 \end{aligned}$$

But for $k = 2$,

$$\begin{aligned} R_Y[m,1] &= E_X\left[X_m X_{m+2} + X_m X_{m+1} + X_{m-1} X_{m+2} + X_{m-1} X_{m+1}\right] \\ &= p^2 + p^2 + p^2 + p^2 = 4p^2 \end{aligned}$$

For any $|k| \geq 2$, the autocorrelation will be the same. So:

$$R_Y[m,k] = \begin{cases} 2p + 2p^2 & k = 0 \\ p + 3p^2 & k = -1, 1 \\ 4p^2 & o.w. \end{cases}$$

Note that R_Y is even and that $R_Y[m,k]$ is not a function of m. Is R_Y WSS? Yes.

What is the autocovariance?

$$C_Y[m,k] = R_Y[m,k] - \mu_Y[m]\mu_Y[m+k] = R_Y[m,k] - 4p^2$$

Which in this case is:

$$C_Y[m,k] = \begin{cases} 2p - 2p^2 & k = 0 \\ p - p^2 & k = -1, 1 \\ 0 & o.w. \end{cases}$$

The autocovariance tells us something about our ability to predict future values (k in advance) of Y_k. The higher $C_Y[k]$ is, the more the two values separated by k can be predicted. Here, $k = 0$ is very high (complete predictability) while $k = 1$ is half, or halfway predictable. Finally, for $k > 1$ there is no (zero) predictability.

Lecture 19

Today: (1) Discussion of AA 5, (2) Several Example Random Processes (Y&G 10.12)

28 Random Telegraph Wave

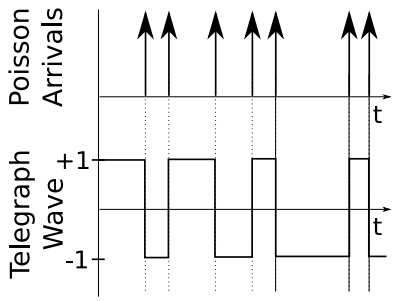


Figure 16: The telegraph wave process is generated by switching between +1 and -1 at every arrival of a Poisson process.

This was originally used to model the signal sent over telegraph lines. Today it is still useful in digital communications, and digital control systems. We model each flip as an arrival in a Poisson process. It is a model for a binary time-varying signal:

$$X(t) = X(0)(-1)^{N(t)}$$

Where $X(0)$ is -1 with prob. 1/2, and 1 with prob. 1/2, and $N(t)$ is a Poisson counting process with rate λ, (the number of arrivals in a Poisson process at time t). $X(0)$ is independent of $N(t)$ for any time t. See Figure 16.

1. What is $E_{X(t)}[X(t)]$?

$$\begin{aligned}\mu_X(t) &= E_{X(t)}\left[X(0)(-1)^{N(t)}\right] = E_X[X(0)]\,E_N\left[(-1)^{N(t)}\right] \\ &= 0 \cdot E_N\left[(-1)^{N(t)}\right] = 0 \end{aligned} \tag{16}$$

2. What is $R_X(t,\delta)$? (Assume $\tau \geq 0$.)

$$\begin{aligned}R_X(t,\tau) &= E_X\left[X(0)(-1)^{N(t)}X(0)(-1)^{N(t+\tau)}\right] \\ &= E_X\left[(X(0))^2(-1)^{N(t)+N(t+\tau)}\right] \\ &= E_N\left[(-1)^{N(t)+N(t+\tau)}\right] \\ &= E_N\left[(-1)^{N(t)+N(t)+(N(t+\tau)-N(t))}\right] \\ &= E_N\left[(-1)^{2N(t)}(-1)^{(N(t+\tau)-N(t))}\right] \\ &= E_N\left[(-1)^{2N(t)}\right]E_N\left[(-1)^{(N(t+\tau)-N(t))}\right] \\ &= E_N\left[(-1)^{(N(t+\tau)-N(t))}\right]\end{aligned}$$

Remember the trick you see inbetween lines 3 and 4? $N(t)$ and $N(t+\tau)$ represent the number of arrivals in **overlapping intervals**. Thus $(-1)^{N(t)}$ and $(-1)^{N(t+\tau)}$ are NOT independent. But $N(t)$ and $N(t+\tau)-N(t)$ DO represent the number of arrivals in non-overlapping intervals, so we can proceed to simplify the expected value of the product (in line 6) to the product of the expected values (in line 7). This difference is just the number of arrivals in a period τ, call it $K = N(t+\tau) - N(t)$, and it must have a Poisson pmf with parameter λ and time τ. Thus the expression is $E_K\left[(-1)^K\right]$ is given by

$$\begin{aligned}R_X(t,\tau) &= \sum_{k=0}^{\infty}(-1)^k\frac{(\lambda\tau)^k}{k!}e^{-\lambda\tau} \\ &= e^{-\lambda\tau}\sum_{k=0}^{\infty}\frac{(-\lambda\tau)^k}{k!} = e^{-\lambda\tau}e^{-\lambda\tau} = e^{-2\lambda\tau}\end{aligned} \tag{17}$$

If $\tau < 0$, we would have had $e^{2\lambda\tau}$. Thus

$$R_X(t,\tau) = e^{-2\lambda|\tau|} = R_X(\tau)$$

It is WSS. Note $R_X(0) \geq 0$, that it is also symmetric and decreasing as it goes away from 0. What is the power in this R.P.? (Answer: Avg power $= R_X(0) = 1$.) Does that make sense?

You could also have considered $R_X(t,\tau)$ to be a question of, whether or not $X(t)$ and $X(t+\tau)$ have the same sign – if so, their product will be one, if not, their product will be zero.

29 Gaussian Processes

We talked about i.i.d. random sequences prior to Exam 2. Now, we're going to talk specifically about Gaussian i.i.d. random sequences. Let X_n be an i.i.d. Gaussian sequence with mean μ and variance σ^2. See Figure 17 for an example.

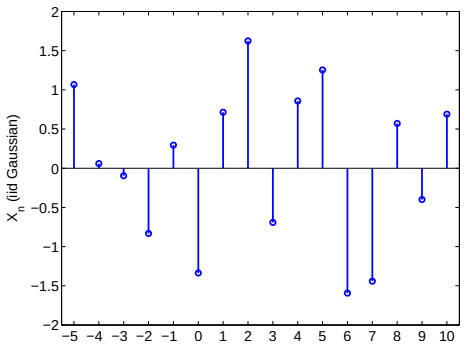


Figure 17: An example of an i.i.d. Gaussian random sequence X_n, with $\mu = 0$ and variance $\sigma^2 = 1$.

Example: Gaussian i.i.d. Random Sequence
What are the autocovariance and autocorrelation functions?

Solution: Since $R_X[m,k] = E_X[X_m X_{m+k}]$, we have to separate into cases $k = 0$ or $k \neq 0$. In the former case, $R_X[m,k] = E_X\left[X_m^2\right]$. In the latter case, $R_X[m,k] = E_X[X_m]\,E_X[X_{m+k}]$. So

$$R_X[m,k] = E_X[X_m X_{m+k}] = \begin{cases} \sigma^2 + \mu^2, & k = 0 \\ \mu^2, & o.w. \end{cases}$$

Then,

$$C_X[m,k] = R_X[m,k] - \mu_X(m)\mu_X(k) = \begin{cases} \sigma^2, & k = 0 \\ 0, & o.w. \end{cases}$$

29.1 Discrete Brownian Motion

Now, let's consider the motion. It is often important to model motion. Eg., the diffusion of gasses. Eg., the motion of an ant. Eg., the mobility of a cell phone user. Also, the motion of a stock price or any commodity.

Let $X_0, X_1, \ldots$ be an i.i.d. Gaussian process with $\mu = 0$ and variance $\sigma^2 = T\alpha$ where T is the sampling period and α is a scale factor. We model an object's motion in 1-D as

$$Y_n = \sum_{i=0}^{n} X_i$$

Say that $Y_n = Y(t_n) = Y(nT)$. Then $Y_n = Y_{n-1} + X_n$, so X_n represents the motion that occurred between time $(n-1)T$ and nT. This says that the motion in one time period is independent of the motion in another time period. This may be a bad model for humans, and even ants, but it turns out to be a very good model for gas molecules.

You could do two dimensional motion by having Y_n represent the motion along one axis, and Z_n represent motion along another axis.

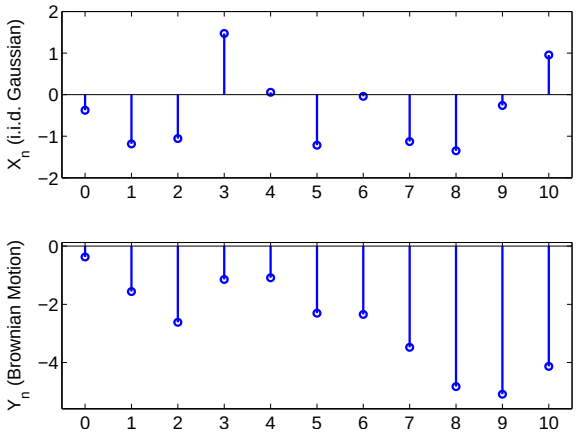


Figure 18: Realization of (a) an i.i.d. Gaussian random sequence and (b) the resulting Brownian motion random sequence.

What does the discrete Brownian motion R.P. remind you of? It is a **discrete counting process**, just like the Binomial R.P. or the Poisson R.P.! On Homework 9, you will compute the mean and autocovariance functions of the discrete Brownian motion R.P. My advice: see the derivation of the autocovariance function $C_X(t,\tau)$ for the Poisson R.P. and alter it.

29.2 Continuous Brownian Motion

Now, lets decrease the sampling interval T. Now, we have more samples, but within each interval, the motion has smaller and smaller variance.

Def'n: *Continuous Brownian Motion Process*
A cts. Brownian motion random process $W(t)$ has the properties

1. $W(0) = 0$,
2. for any $\tau > 0$, $W(t+\tau) - W(t)$ is Gaussian with zero mean and variance $\tau\alpha$,
3. and the *independent increments* property: the change in any interval is independent of the change in any other non-overlapping interval.

The discrete Brownian motion R.P. is just a sampled version of this $W(t)$.

Example: Mean and autocovariance of cts. Brownian Motion
What is $\mu_W(t)$ and $C_W(t,\tau)$?

$$\begin{aligned}
\mu_W(t) &= E_W\left[W(t)\right] = E_W\left[W(t) - W(0)\right] = 0 \\
C_X(t,\tau) &= E_W\left[W(t)W(t+\tau)\right] = E_W\left[W(t)[W(t) + W(t+\tau) - W(t)]\right] \\
&= E_W\left[W^2(t)\right] + E_W\left[W(t)[W(t+\tau) - W(t)]\right] \\
&= \alpha t + 0 \cdot 0 = \alpha t
\end{aligned}$$

Assuming $\tau \geq 0$. If $\tau < 0$,

$$\begin{aligned} C_X(t,\tau) &= E_W\left[W(t)W(t+\tau)\right] = E_W\left[[W(t)-W(t+\tau)+W(t+\tau)]W(t+\tau)\right] \\ &= E_W\left[W^2(t+\tau)\right] + E_W\left[[W(t)-W(t+\tau)]W(t+\tau)\right] \\ &= \alpha(t+\tau) \end{aligned}$$

Overall, you could write

$$C_X(t,\tau) = \alpha \min(t, t+\tau)$$

Is this the same as $R_X(t,\tau)$? Yes, since the mean function is zero.

29.3 Continuous White Gaussian process

Now, let's discuss the continuous version of the i.i.d. Gaussian random sequence. If we made the sample time smaller and smaller, and the samples were still i.i.d Gaussian, we'd eventually see a autocovariance function like an impulse function. So, we talk about a process $W(t)$ as a White Gaussian r.p. when it has these properties:

1. $W(t)$ is WSS.
2. $W(t)$ has autocorrelation function $R_W(\tau) = \eta_0\delta(\tau)$, where $\delta(\tau)$ is the impulse function centered at zero, *i.e.*,
$$\delta(\tau) = \lim_{\epsilon \to 0} \begin{cases} 1/\epsilon, & -\epsilon/2 \leq \tau \leq \epsilon/2 \\ 0, & o.w. \end{cases}$$
and η_0 is a constant.

These properties have consequences:

1. This means that $W(t)$ is zero-mean and uncorrelated with (and independent of) $W(t+\tau)$ for any $\tau \neq 0$, no matter how small. How do you draw a sample of $W(t)$? Ans: blow chalk dust on the board. Or, see Figure 19.
2. The average power of the White Gaussian r.p. $R_W(0)$ is infinite. It is not a r.p. that exists in nature! But it is a good approximation that we make quite a bit of use of in communications and controls. Eg, what do you see/hear when you turn the TV/radio on to a station that doesn't exist? This noise isn't completely white, but it can be modeled as a white noise process going through a narrowband filter.

Example: Lossy, Noisy AM Radio Channel
Consider a noise-free AM radio signal $A(t)$: We might model $A(t)$ as a zero-mean WSS random process with autocovariance $C_A(\tau) = Pe^{-10|\tau|}$, where P is the transmit power in Watts. At a receiver, we receive an attenuated, noisy version of this signal, $S(t)$,

$$S(t) = \alpha A(t) + W(t)$$

where α is the attenuation, and $W(t)$ is a noise process, modeled as a white Gaussian random process with $R_W(\tau) = \eta_0\delta(\tau)$. Also, we know that $W(t_1)$ is independent of $A(t_2)$ for all times t_1 and t_2.

1. What is the mean function $\mu_S(t)$?

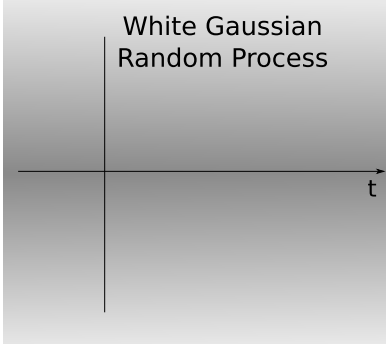


Figure 19: A realization of a white Gaussian process. Get it?

2. What is the autocovariance function $R_S(\tau)$?

SUMMARY: You can calculate the cross-correlation for a wide variety of r.p.s. This tells you about the power, and the joint pdf of two samples of the r.p. taken at different points in time. Many processes, like position of something in motion, are correlated over time, and autocorrelation and autocovariance are critical to be able to model and analyze them.

Lecture 20

Today: (1) Power Spectral Density; (2) Filtering of R.P.s

30 Power Spectral Density of a WSS Signal

Now we make specific our talk of the spectral characteristics of a random signal. **What would happen if we looked at our WSS random signal on a spectrum analyzer (set to average)?** We first need to remind ourselves what the "frequency domain" is.

Def'n: *Fourier Transform*
Functions $g(t)$ and $G(f)$ are a Fourier transform pair if

$$\begin{aligned} G(f) &= \int_{t=-\infty}^{\infty} g(t)e^{-j2\pi ft}dt \\ g(t) &= \int_{f=-\infty}^{\infty} G(f)e^{j2\pi ft}df \end{aligned}$$

PLEASE SEE TABLE 11.1 ON PAGE 413.

Def'n: *Power Spectral Density (PSD)*
The PSD of WSS R.P. X(t) is defined as:

$$S_X(f) = \lim_{T\to\infty} \frac{1}{2T} E\left[\left|\int_{-T}^{T} X(t)e^{-j2\pi ft}dt\right|^2\right]$$

Note three things. Firstly, the a spectrum analyzer records $X(t)$ for a finite duration of time $2T$ (in

this notation) before displaying $X(f)$, the FT of $X(t)$. We define PSD in the limit as T goes large. Secondly, this expected value is "sitting in" for the time average. This is valid if the process has the "ergodic" property, which we do not cover specifically in this course, although it is described on page 378 of Y&G. Thirdly, we don't directly look at the FT on the spectrum analyzer; we look at the *power* in the FT. Power for a (possibly complex) voltage signal G is $S = |G|^2 = G^* \cdot G$.

Theorem: If $X(t)$ is a WSS random process, then $R_X(\tau)$ and $S_X(f)$ are a Fourier transform pair, where $S_X(f)$ is the power spectral density (PSD) function of $X(t)$:

$$\begin{aligned} S_X(f) &= \int_{\tau=-\infty}^{\infty} R_X(\tau) e^{-j2\pi f\tau} d\tau \\ R_X(\tau) &= \int_{f=-\infty}^{\infty} S_X(f) e^{j2\pi f\tau} df \end{aligned}$$

Proof: Omitted: see Y&G p. 414-415. This theorem is called the Wiener-Khintchine theorem, please use the formal name whenever talking to non-ECE friends to convince them that this is hard.

Short story: You can't just take the Fourier transform of $X(t)$ to see what would be on the spectrum analyzer when $X(t)$ is a random process. But, with the work we've done so far this semester, you can come up with an autocorrelation function $R_X(\tau)$, which is a non-random function. Then, the FT of the autocorrelation function shows what the average power vs. frequency will be on the spectrum analyzer.

Four Properties of the PSD:

1. **Units**: $S_X(f)$ has units of power per unit frequency, *i.e.*, Watts/Hertz.
2. **Non-negativity**: $S_X(f) \geq 0$ for all f.
3. **Average total power**: $\int_{f=-\infty}^{\infty} S_X(f)df = E_X\left[X^2(t)\right] = R_X(0)$ is the average power of the r.p. $X(t)$.
4. **Even function**: $S_X(-f) = S_X(f)$.

Intuition for (1) and (3) above: PSD shows us how much energy exists in a band. We can integrate over a certain band to get the total power within the band.

Example: Power Spectral Density of white Gaussian r.p.

1. What is $S_W(f)$ for white Gaussian noise process $W(t)$?
2. What is the power (in Watts) contained in the band $[10, 20]$ MHz if $\eta_0 = 10^{-12}$ W/Hz?

(a) Answer: Recall $R_W(\tau) = \eta_0\delta(\tau)$. Thus

$$S_X(f) = \int_{\tau=-\infty}^{\infty} \eta_0\delta(\tau) e^{-j2\pi f\tau} d\tau$$

Note that

$$\begin{aligned}\int_{\tau=-\infty}^{\infty} \delta(\tau-\tau_0)g(\tau)d\tau &= g(\tau_0)\\ \int_{\tau=-\infty}^{\infty} \delta(\tau)g(\tau)d\tau &= g(0)\end{aligned}$$

So

$$S_X(f) = \eta_0 e^{-j2\pi f 0} = \eta_0$$

It is constant across frequency! Remember what a spectrum analyzer looks like without any signal connected, with averaging turned on - just a low constant. This is the practical effect of thermal noise, which is well-modeled as a white Gaussian random process (with a high cutoff frequency).

(b) Answer: Integrate $S_X(f) = \eta_0$ under the range $[10, 20]$ MHz, to get $10 \times 10^6 \eta_0 = 10^{-5}$ W or 1 μW.

Example: PSD of the Random Telegraph Wave process
Recall that in the last lecture, for $Y(t)$ a random telegraph wave, we showed that

$$R_X(\tau) = e^{-2\lambda|\tau|}$$

Now, find its PSD.

$$S_X(f) = \mathfrak{F}\left\{e^{-2\lambda|\tau|}\right\}$$

From Table 11.1, $\mathfrak{F}\left\{ae^{-a|\tau|}\right\} = \frac{2a^2}{a^2+(2\pi f)^2}$, so

$$\begin{aligned}S_X(f) &= \frac{1}{2\lambda}\mathfrak{F}\left\{2\lambda e^{-2\lambda|\tau|}\right\}\\ &= \frac{1}{2\lambda}\frac{2(2\lambda)^2}{4\lambda^2+(2\pi f)^2}\\ &= \frac{4\lambda}{4\lambda^2+(2\pi f)^2}\end{aligned}$$

Example: PSD of the random phase sinusoid
Recall that we had a random phase Θ uniform on $[0, 2\pi)$ and,

$$X(t) = A\cos(2\pi f_c t + \Theta)$$

We found that $\mu_X = 0$ and

$$C_X(t,\tau) = \frac{A^2}{2}\cos(2\pi f_c \tau)$$

So to find the PSD,

$$S_X(f) = \mathfrak{F}\left\{\frac{A^2}{2}\cos(2\pi f_c\tau)\right\} = \frac{A^2}{2}\mathfrak{F}\left\{\cos(2\pi f_c\tau)\right\}$$

We find in table the exact form for the Fourier transform that we're looking for:

$$S_X(f) = \frac{A^2}{4}\left[\delta(f-f_c)+\delta(f+f_c)\right]$$

(Draw a plot).

31 Linear Time-Invariant Filtering of WSS Signals

In our examples, we've been generally talking about filtered signals. For example:

1. Brownian Motion is an integral (low pass filter) of White Gaussian noise
2. Some of the HW 8 problems.
3. Problems like this: $X_1, X_2, \ldots$ an i.i.d random sequence. Let $Y_2, Y_3, \ldots$ be $Y_k = X_k + X_{k-1}$ (also a low-pass filter)

We're going to limit ourselves to the study of linear time-invariant (LTI) filters. What is a linear filter? It is a linear combination of the inputs. This includes sums, multiply by constants, and integrals. What is a time-invariant filter? Its definition doesn't change over the course of the experiment. It has a single definition.

Well, LTI filters are more generally represented by their impulse response, $h(t)$ or $h[n]$. The output is a convolution of the input and the filter's impulse response.

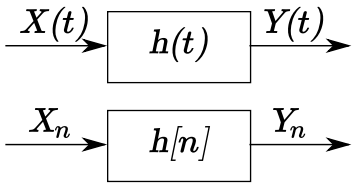


Figure 20: Filtering of a continous R.P. $X(t)$ by a filter with impulse response $h(t)$ to generate output R.P. $Y(t)$; or filtering of a discrete R.P. X_n by a filter with impulse response $h[n]$ to generate output R.P. Y_n.

Here, we're going to be more general. Let $X(t)$ be a WSS random process with autocorrelation $R_X(\tau)$. Let $X(t)$ be the input to a filter $h(t)$. Let $Y(t)$ be the output of the filter. (See Figure 20.) What is the mean function and autocorrelation of $Y(t)$?

$$\begin{aligned} Y(t) &= (h \star X)(t) = \int_{\tau=-\infty}^{\infty} h(\tau) X(t-\tau) d\tau \\ \mu_Y(t) &= E_Y\left[Y(t)\right] = E_Y\left[\int_{\tau=-\infty}^{\infty} h(\tau) X(t-\tau) d\tau\right] \end{aligned}$$

Why can you exchange the order of an integral and an expected value? An expected value *is* an integral.

$$\mu_Y(t) = \int_{\tau=-\infty}^{\infty} h(\tau) E_Y\left[X(t-\tau)\right] d\tau = \mu_X \int_{\tau=-\infty}^{\infty} h(\tau) d\tau$$

For the autocorrelation function,

$$\begin{aligned} R_Y(\tau) &= E_Y\left[Y(t)Y(t+\tau)\right] = E_Y\left[\int_{\alpha=-\infty}^{\infty} h(\alpha) X(t-\alpha) \int_{\beta=-\infty}^{\infty} h(\beta) X(t+\tau-\beta) d\beta d\alpha\right] \\ &= \int_{\alpha=-\infty}^{\infty} h(\alpha) \int_{\beta=-\infty}^{\infty} h(\beta) E_Y\left[X(t-\alpha)X(t+\tau-\beta)\right] d\beta d\alpha \\ &= \int_{\alpha=-\infty}^{\infty} h(\alpha) \int_{\beta=-\infty}^{\infty} h(\beta) R_X(\tau+\alpha-\beta) d\beta d\alpha \end{aligned} \tag{18}$$

This is a convolution & correlation of R_X with $h(t)$:

$$R_Y(\tau) = h(\tau) \star R_X(\tau) \star h(-\tau)$$

31.1 In the Frequency Domain

What if we took the Fourier transform of both sides of the above equation?

$$\mathfrak{F}\{R_Y(\tau)\} = \mathfrak{F}\{h(\tau) \star R_X(\tau) \star h(-\tau)\}$$

Well the Fourier transform of the autocorrelation function is the PSD; and the Fourier transform of a convolution is the product of the Fourier transforms.

$$S_Y(f) = \mathfrak{F}\{h(\tau)\} S_X(f) \mathfrak{F}\{h(-\tau)\}$$

So

$$S_Y(f) = H(f) S_X(f) H^*(f)$$

Or equivalently

$$S_Y(f) = |H(f)|^2 S_X(f)$$

Example: White Gaussian Noise through a moving average filter
Example 11.2 in your book. A white Gaussian noise process $W(t)$ with autocorrelation function $R_X(\tau) = \eta_0 \delta(\tau)$ is passed through a moving average filter,

$$h(t) = \begin{cases} 1/T, & 0 \le t \le T \\ 0, & o.w. \end{cases}$$

What are the mean and autocorrelation functions?

Recall that $X(t)$ is a zero mean process. So, $\mu_Y(t) = \mu_X \int_{\tau=-\infty}^{\infty} h(\tau) d\tau = 0$.

1. What is $H(f)$?

$$\begin{aligned} H(f) &= \mathfrak{F}\{h(t)\} = \mathfrak{F}\left\{\text{rect}\left(\frac{t}{T} - \frac{1}{2}\right)\right\} \\ &= \mathfrak{F}\left\{\text{rect}\left(\frac{t}{T}\right)\right\} e^{-j2\pi f(T/2)} \\ &= T\text{sinc}(fT) e^{-j\pi fT} \end{aligned} \tag{19}$$

2. What is $S_X(f)$? It is $\mathfrak{F}\{\eta_0 \delta(\tau)\} = \eta_0$.

3. What is $S_Y(f)$?

$$\begin{aligned} S_Y(f) &= |H(f)|^2 S_X(f) = \left| T\text{sinc}(fT) e^{-j\pi fT} \right|^2 \\ &= T^2 \text{sinc}^2(fT) \eta_0 \end{aligned} \tag{20}$$

4. What is $R_Y(f)$?

$$R_Y(\tau) = \mathfrak{F}^{-1}\left\{T^2 \text{sinc}^2(fT) \eta_0\right\} = T\eta_0 \wedge (\tau/T) \tag{21}$$

where

$$\wedge(\tau/T) = \begin{cases} 1 - |\tau|/T, & -T < \tau < T \\ 0, & o.w. \end{cases}$$

Note: you should add this to your table:

$$\mathfrak{F}\{\wedge(\tau/T)\} = T\text{sinc}^2(fT)$$

Does this agree with what we would have found directly? Yes, please verify this using the convolution and correlation equation in (18).

$$\begin{aligned} R_Y(\tau) &= \int_{\alpha=0}^{T} \frac{1}{T} \int_{\beta=0}^{T} \frac{1}{T} \eta_0 \delta(\tau - \beta + \alpha) d\beta d\alpha \\ &= \frac{\eta_0}{T^2} \int_{\alpha=0}^{T} [u(\tau+\alpha) - u(T - (\tau+\alpha))] d\alpha \\ &= \frac{\eta_0}{T^2} \int_{\alpha=0}^{T} [u(\alpha+\tau) - u((T-\tau) - \alpha))] d\alpha \end{aligned}$$

where $u(t)$ is the unit step function. Drawing a graph, we can see that for $\tau > 0$,

$$R_Y(\tau) = \begin{cases} \frac{\eta_0}{T^2}(T - |\tau|), & \tau \leq T \\ 0, & o.w. \end{cases}$$

Summary of today's lecture: We can find the power spectral density just by taking the Fourier transform of the autocorrelation function. This shows us what we'd see on a spectrum analyzer, if we averaged over time. Finally, we looked at running a R.P. through a general, LTI filter, and saw that we can analyse the output in the frequency domain.

Lecture 21

Today: (1) LTI Filtering, continued; (2) Discrete-Time Filtering

32 LTI Filtering of WSS Signals

Continued from Lecture 20.

32.1 Addition of r.p.s

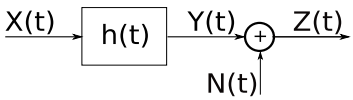


Figure 21: Continuous-time filtering with the addition of noise.

If $Z(t) = Y(t) + N(t)$ for two WSS r.p.s (which are uncorrelated with each other), then we also have that $S_Z(f) = S_Y(f) + S_N(f)$. A typical example is when noise is added into a system at a receiver, onto a signal $Y(t)$ which is already a r.p.

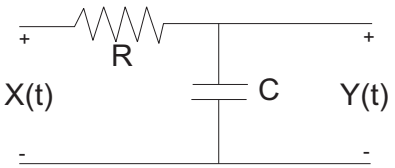


Figure 22: An RC filter, with input $X(t)$ and output $Y(t)$.

32.2 Partial Fraction Expansion

Lets say you come up with a PSD $S_Y(f)$ that is a product of fractions. Eg,

$$S_Y(f) = \frac{1}{(2\pi f)^2 + \alpha^2} \cdot \frac{1}{(2\pi f)^2 + \beta^2}$$

This can equivalently be written as a sum of two different fractions. You can write it as:

$$S_Y(f) = \frac{A}{(2\pi f)^2 + \alpha^2} + \frac{B}{(2\pi f)^2 + \beta^2}$$

Where you use partial fraction expansion (PFE) to find A and B. You should look in an another textbook for the formal definition of PFE. I use the "thumb method" to find A and B. The thumb method is:

1. Pull all constants in the numerators out front. The numerator should just be 1.
2. Go to the first fraction. What do you need $(2\pi f)^2$ to equal to make the denominator 0? Here, it is $-\alpha^2$.
3. Put your thumb on the first fraction, and plug in the value from (2.) for $(2\pi f)^2$ in the second fraction. The value that you get is A. In this case, $A = \frac{1}{-\alpha^2+\beta^2}$.
4. Repeat for each fraction. In the second case, $B = \frac{1}{\alpha^2-\beta^2}$

Thus

$$S_Y(f) = \frac{1}{-\alpha^2 + \beta^2}\left[\frac{1}{(2\pi f)^2 + \alpha^2} - \frac{1}{(2\pi f)^2 + \beta^2}\right]$$

32.3 Discussion of RC Filters

Example: RC Filtering of Random Processes

Let $X(t)$ be a zero-mean white Gaussian process, with $R_X(\tau) = \eta_0\delta(\tau)$, input to the filter in Figure 22.

Remember your circuits classes? This will be a review.

1. What is the frequency response of this filter, $H(f)$? Solution in two ways: (1) Use complex impedances, and treat it as a voltage divider; (2) Use the basic differential equations approach. For (1),

$$H(\omega) = \frac{1/(j\omega C)}{R + 1/(j\omega C)} = \frac{1}{j\omega RC + 1} = \frac{1/(RC)}{1/(RC) + j\omega}$$

Since $\omega = 2\pi f$,

$$H(f) = \frac{1/(RC)}{1/(RC) + j2\pi f}$$

For approach (2), we need to go back to the equation for a current through a capacitor,

$$i(t) = C\frac{dY(t)}{dt}$$

which we can then use to figure out the voltage across the resistor, and thus what $Y(t)$ must be.

$$Y(t) + R\left(C\frac{dY(t)}{dt}\right) = X(t)$$

So in the frequency domain,

$$Y(f) + RCj2\pi fY(f) = X(f)$$

Thus

$$H(f) = \frac{Y(f)}{X(f)} = \frac{1}{1 + RCj2\pi f}$$

And we get the same filter.

2. What is $S_Y(f)$

$$S_Y(f) = |H(f)|^2 S_X(f) = \frac{1/(RC)}{1/(RC) + j2\pi f}\frac{1/(RC)}{1/(RC) - j2\pi f}\eta_0 = \eta_0\frac{\frac{1}{(RC)^2}}{\frac{1}{(RC)^2} + (2\pi f)^2}$$

3. What is $R_Y(\tau)$?

$$R_Y(\tau) = \mathfrak{F}^{-1}\left\{\eta_0\frac{\frac{1}{(RC)^2}}{\frac{1}{(RC)^2} + (2\pi f)^2}\right\} = \frac{\eta_0}{2}\frac{1}{RC}e^{-\frac{1}{RC}|\tau|}$$

4. What is the average power of $Y(t)$? Answer: $\frac{\eta_0}{2}\frac{1}{RC}$

5. What is the power spectral density of $Y(t)$ at $f = 10^4/(2\pi)$, when $RC = 10^{-4}s$? Answer:

$$S_Y(10^4/(2\pi)) = \eta_0\frac{10^8}{10^8 + (10^4)^2} = \eta_0/2$$

This is the 3-dB point in the filter, that is, where the power is 1/2 its maximum. In this case, the maximum is at $f = 0$.

33 Discrete-Time R.P. Spectral Analysis

Traditionally, the discrete time case is not taught in a random process course. But real digital signals are everywhere, and so I imagine that most of the time that this analysis is necessary is with discrete-time random processes. For example, image processing, audio processing, video processing, all are mostly digital. Traditionally, communication system design has required continuous-time filters; but now, many of the radios being designed are all-digital (or mostly digital) so that filters must be designed in software.

33.1 Discrete-Time Fourier Transform

Def'n: *Discrete-Time Fourier Transform (DTFT)*
The sequence $\{\ldots, x_{-2}, x_{-1}, x_0, x_1, x_2, \ldots\}$ and the function $X(\phi)$ are a discrete-time Fourier transform pair if

$$X(\phi) = \sum_{n=-\infty}^{\infty} x_n e^{-j2\pi\phi n}, \quad x_n = \int_{-1/2}^{1/2} X(\phi) e^{+j2\pi\phi n} d\phi$$

See Table 11.2 on page 418.

Here, ϕ is a normalized frequency, with a value between $-\frac{1}{2}$ and $+\frac{1}{2}$. The actual frequency f in Hz is a function of the sampling frequency, $f_s = 1/T_s$,

$$f = f_s \phi$$

Recall that the Nyquist theorem says that if we sample at rate f_s, then we can only represent signal components in the frequency range $-f_s/2 \leq f \leq f_s/2$.

33.2 Power-Spectral Density

Def'n: *Discrete-Time Weiner-Khintchine*
If X_n is a WSS random sequence, then $R_X[k]$ and $S_X(f)$ are a discrete-time Fourier transform pair,

$$S_X(\phi) = \sum_{k=-\infty}^{\infty} R_X[k] e^{-j2\pi\phi k}, \quad R_X[k] = \int_{-1/2}^{1/2} S_X(\phi) e^{+j2\pi\phi k} d\phi$$

Theorem: (11.6) LTI Filters and PSD: When a WSS random sequence X_n is input to a linear time-invariant filter with transfer function $H(\phi)$, the power spectral density of the output Y_n is

$$S_Y(\phi) = |H(\phi)|^2 S_X(\phi)$$

Proof: In Y&G.

Note a LTI filter is completely defined by its impulse response h_n. The DTFT of h_n is $H(\phi)$.

We have two different delta functions:

- Continuous-time impulse: $\delta(t)$ (Dhirac). This is infinite at $t = 0$ in a way that makes the area under the curve equal to 1. It is zero anywhere else $t \neq 0$.
- Discrete-time impulse: δ_n (Kronecker). This has a finite value for all time (1 for $n = 0$, 0 otherwise).

Also note that: $u[n]$ is the discrete unit step function,

$$u[n] = \begin{cases} 1, & n = 0, 1, 2, \ldots \\ 0, & o.w. \end{cases}$$

Some identities of use:

- Complex exponential:
$$e^{-j2\pi\phi} = \cos(2\pi\phi) - j\sin(2\pi\phi)$$
- Cosine:
$$2\cos(2\pi\phi) = e^{-j2\pi\phi} + e^{j2\pi\phi}$$

33.3 Examples

Example: An i.i.d. random sequence X_n with zero mean and variance 10 is input to a LTI filter with impulse response $h_n = (0.75)^n u[n]$. What is the PSD and autocorrelation function of the output, Y_n?

1. What is the autocorrelation function of the input?
$$R_X[m,k] = E_X\left[X_m X_{m+k}\right] = 10\delta_k = R_X[k].$$
2. What is the PSD of the input?
$$S_X(\phi) = \text{DTFT}\left\{R_X[k]\right\} = \text{DTFT}\left\{10\delta_k\right\} = 10$$
3. What is the frequency characteristic of the filter?
$$H(\phi) = \text{DTFT}\left\{h_n\right\} = \text{DTFT}\left\{(0.75)^n u[n]\right\} = \frac{1}{1-0.75e^{-j2\pi\phi}}$$
4. What is the $|H(\phi)|^2$?
$$\begin{aligned}
|H(\phi)|^2 &= \frac{1}{1-0.75e^{-j2\pi\phi}}\left(\frac{1}{1-0.75e^{-j2\pi\phi}}\right)^* \\
&= \frac{1}{1-0.75e^{-j2\pi\phi}}\left(\frac{1}{1-0.75e^{j2\pi\phi}}\right) \\
&= \frac{1}{1-0.75e^{-j2\pi\phi}-0.75e^{j2\pi\phi}+(0.75)^2} \\
&= \frac{1}{1-0.75(e^{-j2\pi\phi}+e^{j2\pi\phi})+(0.75)^2} \\
&= \frac{1}{1+(0.75)^2-0.75(2)\cos(2\pi\phi)}
\end{aligned}$$
5. What is $S_Y(\phi)$? It is just $10|H(\phi)|^2$.
6. What is $R_Y[k]$?
$$R_Y[k] = \text{DTFT}^{-1}\left\{S_Y(\phi)\right\} = \text{DTFT}^{-1}\left\{\frac{10}{1+(0.75)^2-0.75(2)\cos(2\pi\phi)}\right\} = \frac{10}{1-(0.75)^2}(0.75)^{|k|}$$

Write this down on your table: Discrete time domain:

$$h_n = \begin{cases} 1, & n = 0, 1, \ldots, M-1 \\ 0, & o.w. \end{cases}$$

DTFT domain:

$$H(\phi) = \left(\frac{1 - e^{-j2\pi\phi M}}{1 - e^{-j2\pi\phi}} \right)$$

Example: Moving average filter
Let X_n, a i.i.d. random sequence with zero mean and variance σ^2 be input to a moving-average filter,

$$h_n = \begin{cases} 1/M, & n = 0, 1, \ldots, M-1 \\ 0, & o.w. \end{cases}$$

What is the PSD of the output of the filter?

We know that $S_X[k] = \sigma^2$.

$$H(\phi) = \frac{1}{M} \left(\frac{1 - e^{-j2\pi\phi M}}{1 - e^{-j2\pi\phi}} \right)$$

$$\begin{aligned} |H(\phi)|^2 &= \frac{1}{M^2} \left(\frac{1 - e^{-j2\pi\phi M}}{1 - e^{-j2\pi\phi}} \right) \left(\frac{1 - e^{j2\pi\phi M}}{1 - e^{j2\pi\phi}} \right) \\ &= \frac{1}{M^2} \left(\frac{1 - e^{-j2\pi\phi M} - e^{j2\pi\phi M} + 1}{1 - e^{-j2\pi\phi} - e^{j2\pi\phi} + 1} \right) \\ &= \frac{1}{M^2} \left(\frac{1 - \cos(2\pi\phi M)}{1 - \cos(2\pi\phi)} \right) \end{aligned}$$

So the output PSD is

$$S_Y(\phi) = |H(\phi)|^2 S_X[k] = \frac{\sigma^2}{M^2} \left(\frac{1 - \cos(2\pi\phi M)}{1 - \cos(2\pi\phi)} \right)$$

Lecture 22

Today: (1) Markov Property (2) Markov Chains

34 Markov Processes

We've talked about

1. iid random sequences
2. WSS random sequences

Each sample of the iid sequence has no dependence on past samples. Each sample of the WSS sequence may depend on many (possibly infinitely many) previous samples. Now we'll talk specifically about random processes for which the distribution of X_{n+1} depends at most on the most recent sample. A random property like this is said to have the "Markov property". A quick way to talk about a Markov process is to say that *given the present value, its future is independent of the past.*

It turns out, there are quite a variety of R.P.s which have the Markov property. The benefit is that you can do a lot of analysis using a program like Matlab, and come up with valuable answers for the design of systems.

34.1 Definition

Def'n: *Markov Process*
A discrete random process X_n is Markov if it has the property that

$$P\left[X_{n+1}|X_n, X_{n-1}, X_{n-2}, \ldots\right] = P\left[X_{n+1}|X_n\right]$$

A discrete random process $X(t)$ is Markov if it has the property that for $t_{n+1} > t_n > t_{n-1} > t_{n-2} > \cdots$,

$$P\left[X(t_{n+1})|X(t_n), X(t_{n-1}), X(t_{n-2}), \ldots\right] = P\left[X(t_{n+1})|X(t_n)\right]$$

Examples For each one, write $P\left[X(t_{n+1})|X(t_n), X(t_{n-1}), X(t_{n-2}), \ldots\right]$ and $P\left[X(t_{n+1})|X(t_n)\right]$:

- Brownian motion: The value of X_{n+1} is equal to X_n plus the random motion that occurs between time n and $n+1$.
- Any independent increments process.
- Gambling or investment value over time.
- Digital computers and control systems. The state is described by what is in the computer's memory; and the transitions may be non-random (described by a deterministic algorithm) or random. Randomness may arrive from input signals.

Notes:

- If you need more than one past sample to predict a future value, then the process is not Markov.
- The value X_n is also called the 'state'. The change from X_n to X_{n+1} is called the state transition.
- i.i.d. R.P.s are also Markov.

34.2 Visualization

We make diagrams to show the possible progression of a Markov process. Each state is a circle; while each transition is an arrow, labeled with the probability of that transition.

Example: Discrete Telegraph Wave R.P.
Let X_n be a Binomial R.P. with parameter p, and let $Y_n = (-1)^{X_n}$. Each time a trial is a success, the R.P. switches from 1 to -1 or vice versa. See the state transition diagram drawn in Fig. 23.

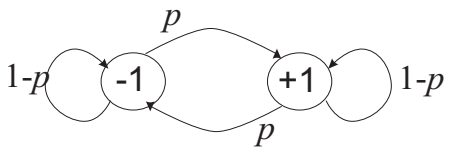


Figure 23: A state transition diagram for the Discrete Telegraph Wave.

Example: (Miller & Childers) "Collect Them All"
How many happy meal toys have you gotten? Fast food chains like to entice kids with a series of toys and tell them to "Collect them all!". Let there be four toys, and let X_n be the number out of four that you've collected after your nth visit to the chain. How many states are there? What are the transition probabilities?

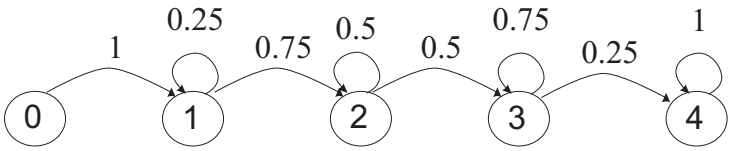


Figure 24: A state transition diagram for the "Collect Them All!" random process.

In particular we're going to be interested in cases where the transition probabilities $P[X_{n+1}|X_n]$ are not a function of n. These are called Markov chains. Also, we're going to limit ourselves to discrete-time Markov chains, and those which are discrete-valued.

34.3 Transition Probabilities: Matrix Form

This is Section 12.1.

We define $P_{i,j}$ as:

$$P_{i,j} = P[X_{n+1} = j | X_n = i]$$

They satisfy:

1. $P_{i,j} \geq 0$
2. $\sum_j P_{i,j} = 1$

Note: $\sum_i P_{i,j} \neq 1$! Don't make this mistake.

Def'n: *State Transition Probability Matrix*
The state transition probability matrix $\mathbf{P}$ of an N-state Markov chain is given by:

$$\mathbf{P}(1) = \begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,N} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,N} \\ \vdots & & \ddots & \vdots \\ P_{N,1} & P_{N,2} & \cdots & P_{N,N} \end{bmatrix}$$

Note:

- The rows sum to one; the columns may not.
- There may be N states, but they may not have values 1, 2, 3, $\ldots, N$. Thus if we don't have such values, we may create an intermediate r.v. W_n which is equal to the rank of the value of X_n, or $W_n = \text{rank} X_n$, for some arbitrary ranking system.

Example: Discrete telegraph wave
What is the the TPM of the Discrete Telegraph Wave R.P.? Use: $W_n = 1$ for $X_n = -1$, and $W_n = 2$ for $X_n = 1$:

$$\mathbf{P}(1) = \begin{bmatrix} P_{1,1} & P_{1,2} \\ P_{2,1} & P_{2,2} \end{bmatrix} = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$

Example: "Collect Them All"
What is the the TPM of the Collect them all example? Use $W_n = X_n + 1$:

$$\begin{aligned} \mathbf{P}(1) &= \begin{bmatrix} P_{1,1} & P_{1,2} & P_{1,3} & P_{1,4} & P_{1,5} \\ P_{2,1} & P_{2,2} & P_{2,3} & P_{2,4} & P_{2,5} \\ P_{3,1} & P_{3,2} & P_{3,3} & P_{3,4} & P_{3,5} \\ P_{4,1} & P_{4,2} & P_{4,3} & P_{4,4} & P_{4,5} \\ P_{5,1} & P_{5,2} & P_{5,3} & P_{5,4} & P_{5,5} \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0.25 & 0.75 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 0.75 & 0.25 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{aligned}$$

Trick: write in above (and to the right) of the matrix the actual states X_n which correspond to each column (row). This makes it easier to fill in the transition diagram. Shown below is the same $\mathbf{P}(1)$ but with red text denoting the actual states of X_n to which each column and row correspond. I do this *every time* I make up a transition probability matrix (but my notes don't show this). You should do this for your own sake; but if you were to type the matrix into Matlab, of course you wouldn't copy in the red parts.

$$\mathbf{P}(1) = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0.25 & 0.75 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 0.75 & 0.25 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

Example: Gambling $50
You start at a casino with 5 $10 chips. Each time n you bet one chip. You win with probability 0.45, and lose with probability 0.55. If you run out, you will stop betting. Also, you decide beforehand to stop if you double your money. What is the TPM for this random process? (Also draw the state transition diagram).

$$\mathbf{P}(1) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.55 & 0 & 0.45 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.55 & 0 & 0.45 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.55 & 0 & 0.45 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.55 & 0 & 0.45 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.55 & 0 & 0.45 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.55 & 0 & 0.45 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.55 & 0 & 0.45 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.55 & 0 & 0.45 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.55 & 0 & 0.45 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Example: Waiting in a finite queue
A mail server (bank) can deliver one email (customer) at each minute. But, X_n more emails (customers) arrive in minute n, where X_n is (i.i.d.) Poisson with parameter $\lambda = 1$ per minute. Emails (people) who can't be handled immediately are queued. But if the number in the queue, Y_n, is equal to 2, the queue is full, and emails will be dropped (customers won't stay and wait). Thus the number of emails in the queue (people in line) is given by

$$Y_{n+1} = \min\left(2, \max\left(0, Y_n - 1\right) + X_n\right)$$

What is the $P\left[X_n = k\right]$?

$$P\left[X_n = k\right] = \frac{(\lambda t)^k}{k!} e^{-\lambda t} = \frac{1}{ek!}$$

$P\left[X_n = 0\right] = 1/e \approx 0.37$, and $P\left[X_n = 1\right] = 1/e \approx 0.37$, and $P\left[X_n = 2\right] = 1/(2e) \approx 0.18$, and .

$$\mathbf{P}(1) = \begin{bmatrix} P_{1,1} & P_{1,2} & P_{1,3} \\ P_{2,1} & P_{2,2} & P_{2,3} \\ P_{3,1} & P_{3,2} & P_{3,3} \end{bmatrix} = \begin{bmatrix} 0.37 & 0.37 & 0.26 \\ 0.37 & 0.37 & 0.26 \\ 0 & 0.37 & 0.63 \end{bmatrix}$$

Example: Chute and Ladder
This does not infringe on the copyright held by Milton Bradley Co. on Chutes and Ladders. See Figure 25. You roll a die (a fair die) and move forward that number of squares. Then, if you land on top of a chute, you have to fall down to a lower square; if you land on bottom of a ladder, you climb up to the higher square. The object is to land on 'Winner'. You don't need to get there with an exact roll. This is a Markov Chain: your future square only depends on your present square and your roll. What are the states? They are

$$S_X = \{1, 2, 4, 5, 7\}$$

Since you'll never stay on 3 and 6, we don't need to include them as states. (We could but there would just be 0 probability of landing on them, so why bother.) This is the transition probability matrix:

$$\mathbf{P}(1) = \begin{bmatrix} 0 & 1/6 & 2/6 & 2/6 & 1/6 \\ 0 & 0 & 2/6 & 2/6 & 2/6 \\ 0 & 0 & 1/6 & 1/6 & 4/6 \\ 0 & 0 & 1/6 & 0 & 5/6 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Example: Countably Infinite Markov Chain
We can also have a countably infinite number of states. It is a discrete-valued R.P. after all; we might still have an infinite number of states. For example, if we didn't ever stop gambling at a fixed upper number. (Draw state transition diagram). There are quite a number of interesting problems in this area, but we won't get into them.

34.4 Multi-step Markov Chain Dynamics

This is Section 12.2.

34.4.1 Initialization

We might not know in exactly which state the markov chain will start. For example, for the bank queue example, we might have people lined up when the bank opens. Let's say we've measured over many days and found that at time zero, the number of people is uniformly distributed, *i.e.*,

$$P[X_0 = k] = \begin{cases} 1/3, & x = 0, 1, 2 \\ 0, & o.w. \end{cases}$$

We represent this kind of information in a vector:

$$\mathbf{p}(0) = [P[X_0 = 0], P[X_0 = 2], P[X_0 = 2]]$$

In general,

$$\mathbf{p}(n) = [P[X_n = 0], P[X_n = 2], P[X_n = 2]]$$

This vector $\mathbf{p}(n)$ is called the "state probability vector" at time n.

The only requirement is that the sum of $\mathbf{p}(n)$ is 1 for any n.

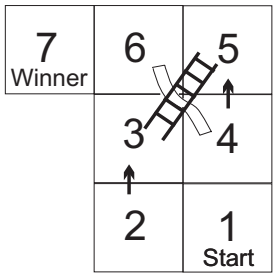


Figure 25: Playing board for the game, *Chute and Ladder*.

34.4.2 Multiple-Step Transition Matrix

Def'n: *n-step transition Matrix*
The n-step transition Matrix $\mathbf{P}(n)$ of Markov chain X_n has (i,j)th element

$$P_{i,j}(n) = P\left[X_{n+m} = j | X_m = i\right]$$

Theorem: Chapman-Kolmogorov equations
Proof: For a Markov chain, the n-step transition matrix satisfies

$$\mathbf{P}(n+m) = \mathbf{P}(n)\mathbf{P}(m)$$

This is Theorem 12.2 in Y&G.

This means, to find the two-step transition matrix, you multiply (matrix multiply) $\mathbf{P}$ and $\mathbf{P}$ together. In general, the n-step transition matrix is

$$\mathbf{P}(n) = [\mathbf{P}(1)]^n$$

Theorem: State probabilites at time n
Proof: The state probabilities at time n can be found as

$$\mathbf{p}(n) = \mathbf{p}(0)[\mathbf{P}(1)]^n$$

This is Theorem 12.4 in Y&G.

34.4.3 n-step probabilities

You start a chain in a random state. The probability that you start in each state is given by the State Probability Vector $\mathbf{p}(0)$. In each time step, your state changes, as described by the TPM $\mathbf{P}(1)$. Now, what are your probabilities in step 2?

$$\mathbf{p}(1) = \mathbf{p}(0)\mathbf{P}(1)$$

Note the transposes. This is the common way of writing this equation, which shows how you progress from one time instant to the next. What is $\mathbf{p}(2)$?

$$\mathbf{p}(2) = \mathbf{p}(1)\mathbf{P}(1) = (\mathbf{p}(0)\mathbf{P}(1))\mathbf{P}(1) = \mathbf{p}(0)\mathbf{P}^2(1)$$

Extending this,

$$\mathbf{p}(n) = \mathbf{p}(0)\mathbf{P}^n(1)$$

34.5 Limiting probabilities

Without doing lots and lots of matrix multiplication, we can more quickly find what happens in the limit. As $n \to \infty$,

$$\lim_{n\to\infty} \mathbf{p}(n) = \lim_{n\to\infty} \mathbf{p}(0)\mathbf{P}^n(1)$$

We denote the left-hand side as follows

$$\boldsymbol{\pi} = \lim_{n\to\infty} \mathbf{p}(n)$$

Depending on the Markov chain, the limit $\boldsymbol{\pi}$ may or may not exist:

1. It may not exist at all;
2. It may depend on $\mathbf{p}(0)$, the initial state probability vector.
3. It may exist regardless of $\mathbf{p}(0)$.

For example, consider the three Markov chains in Figure 12.1 in Y&G.

1. *It may not exist at all*: In 12.1(a), the chain alternates deterministically between state 0 and state 1, at each time. Since the probability of this alternation is exactly one, the n-step transition from state 0 to state 1 will be either 1 or 0 depending on whether n is even or odd. It does not converge.
2. *It may depend on* $\mathbf{p}(0)$: In 12.1(b), the chain stays where it starts, *i.e.*, $\mathbf{P}$ is the identity matrix. So $\boldsymbol{\pi} = \mathbf{p}(0)$.
3. *It may exist regardless of* $\mathbf{p}(0)$: For 12.1(a), the limit $\boldsymbol{\pi}$ exists, it is shown in Example 12.6 to be,
$$\boldsymbol{\pi} = \frac{1}{p+q}[p, q]$$

Theorem: If a finite Markov chain with TPM $\mathbf{P}$ and initial value probability $\mathbf{p}(0)$ has a limiting state vector $\boldsymbol{\pi}$, then

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$$

Proof: Proof: We showed that

$$\mathbf{p}(n+1) = \mathbf{p}(n)\mathbf{P}$$

Taking the limit of both sides,

$$\lim_{n\to\infty} \mathbf{p}(n+1) \;=\; \left(\lim_{n\to\infty} \mathbf{p}(n)\right)\mathbf{P}$$

as $n \to \infty$, the limit of both sides (since the limit exists) is $\boldsymbol{\pi}$.

What is this vector $\mathbf{v}$ called in the following expression?

$$\lambda\mathbf{v} = A\mathbf{v}$$

An eigenvector of a matrix is a vector which, when multiplied by the matrix, results in a scaled version of the original vector. Here, $\mathbf{v}$ is a eigenvector of A.

Returning to the limiting state vector $\boldsymbol{\pi}$. Since $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$, or equivalently,

$$\boldsymbol{\pi}^T = \mathbf{P}\boldsymbol{\pi}^T$$

It is clear that $\boldsymbol{\pi}^T$ is an eigenvector; in particular, the eigenvector with eigenvalue 1. (If there are more than one eigenvectors with value 1, then the limiting state may depend on the input state probability vector.) Note that if you do Matlab 'eig' command, it will return a vector with norm 1. We need a vector with sum 1.

34.6 Matlab Examples

Here, we run two examples in Matlab to see how this stuff is useful to calculate the numerical n-step and limiting probabilities. Of particular interest in these examples is how the state probability vector evolves over time.

34.6.1 Casino starting with $50

See Figure 26.

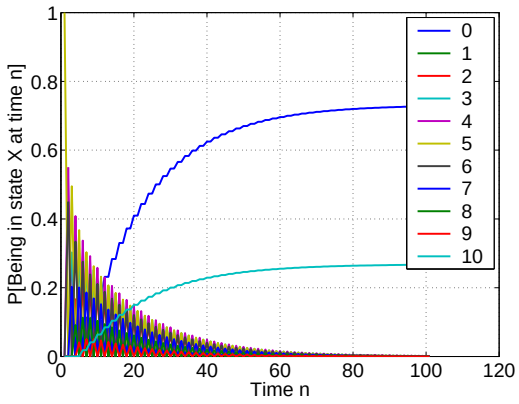


Figure 26: The probability of being in a state as a function of time, for states $0 \ldots 10$, which stand for the \$0 through \$100 values.

34.6.2 Chute and Ladder Game

See Figure 27.

34.7 Applications

- *Google's PageRank*: Each page is a state, and clicking on a hyperlink causes one to change state. If one were to click randomly (uniform across all links on the current page), and then do the same on every page one came to, what would the limiting distribution be? This would effectively take you to the more important pages more often. PageRank uses Markov chain limiting probability analysis, given a uniform probability for all links on a page. The output is one number for each page, which then ranks all pages according to its popularity.

- *Robust Computing*: The memory of a computer can be considered the state, and the program is then the means by which it transitions between different states. Robust programs must never "crash" (imagine, *e.g.*, the code running your car). Analysis of the state transition matrix can be used to verify mathematically that a program will never fail.

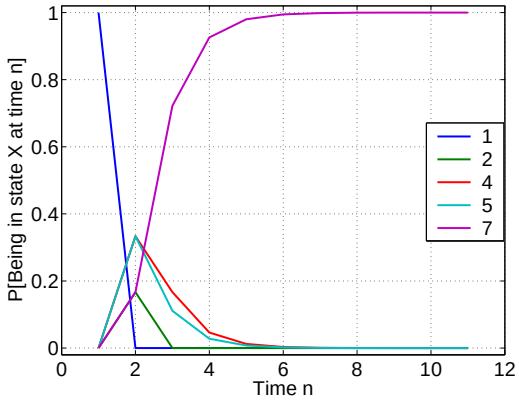


Figure 27: The probability of being on a square as a function of time, for squares 1,2,4,5, and 7.

- *Cell Biology*: Cells have a random number of 'sides', and certain numbers are more popular. But in different parts of a tissue, the proportion for each side-number is always nearly the same. It turns out this can be well-explained using a simple Markov chain. See "The Emergence of Geometric Order in Proliferating Metazoan Epithelia", Ankit Patel, Radhika Nagpal, Matthew Gibson, Norbert Perrimon, *Nature*, 442(7106):1038-41, Aug 31, 2006.