Lectures on EE 3025 Chapters 4-5

John Kieffer
Department of Electrical and Computer Engineering
University of Minnesota Twin Cities
Minneapolis, MN 55455

Lecture 15

Chapters 4-5 Part 1

15.1 Two Review Examples

In Lecture 15, I presented the following two review examples over the Chapter 2-3 material.

Example 15.1. We present an example in *inventory control*. A storekeeper wishes to use probability modeling in order to figure out how much of an inventory of a certain product he should keep on his store shelves. Specifically, suppose that

- C_1 is the price in dollars he pays to the manufacturer for each item of the product.
- C_2 is the price in dollars that he sells each item of the product for.
- C_3 is the cost in dollars per product item that the storekeeper incurs as a result of having to keep an unsold item on his shelves until the beginning of the next "selling season". (Think of the product as a seasonal item, such as a swimsuit, which would typically be sold only during a certain period within the year.)
- D is the number of product items demanded during the selling season by the customers.
- I is the total number of product items stocked by the storekeeper at the beginning of selling season.
- P is the total profit in dollars realized by the storekeeper at the end of the selling season.

C_1, C_2, C_3 are assumed to be fixed numbers. The amount of inventory I is fixed and is set by the storekeeper. D is taken to be a random variable taking nonnegative integer values

$$0, 1, 2, 3, \cdots$$

according to a certain PMF $P^D(d)$. (By seeing what happened over previous selling seasons, the storekeeper could come up with a model for this PMF.) The profit P is a certain function of D and I:

$$P = \phi(D, I).$$

The goal of the storekeeper is to choose I so that his expected profit

$$E[P] = E[\phi(D, I)]$$

is maximized. It is not hard to see that we have the following functional relationship relating P to D and I:

$$P = \begin{cases} I(C_2 - C_1), & D > I \\ (DC_2 - IC_1) - C_3(I - D), & D \leq I \end{cases}$$

We can compute the expected profit as follows:

$$\begin{aligned} E[P] &= \sum_{d=I+1}^{\infty} I(C_2 - C_1)\}P^D(d) + \sum_{d=0}^{I}[(dC_2 - IC_1) - C_3(I - d)]P^D(d) \\ &= I(C_2 - C_1)P[D > I] - I(C_1 + C_3)P[D \leq I] + (C_3 + C_2)\sum_{d=0}^{I} dP^D(d) \end{aligned}$$

We can write this more compactly as:

$$E[P] = I(C_2 - C_1)P[D > I] - I(C_1 + C_3)P[D \leq I] + (C_3 + C_2)E[D|D \leq I]P[D \leq I].$$

Substitute

$$P[D > I] = 1 - P[D \leq I].$$

You then obtain

$$E[P] = I(C_2 - C_1) - I(C_3 + C_2)P[D \leq I] + (C_3 + C_2)E[D|D \leq I]P[D \leq I]. \tag{15.1}$$

To be more specific, suppose we assume that

$$\begin{aligned} C_1 &= 3 \\ C_2 &= 9 \\ C_3 &= 1 \end{aligned}$$

Also, assume that D is equiprobable over the set of values

$$\{1, 2, 3, 4, 5\}.$$

Suppose we consider the case in which I is between 1 and 5. Then

$$\begin{aligned} P[D \leq I] &= I/5 \\ E[D|D \leq I] &= 0.5(1+I) \end{aligned}$$

Substituting into (15.1), we see that

$$E[P] = 7I - I^2.$$

This gives us

$$E[P] = \begin{cases} 10, & I = 5 \\ 12, & I = 3, 4 \\ 10, & I = 2 \\ 6, & I = 1 \end{cases}$$

Notice that for $I = 3$ or $I = 4$, we get an expected profit of 12 dollars, which is the best for inventory values I between 1 and 5. The reader can check that $E[P]$ cannot become bigger than 12 if $I > 5$. Therefore, the best thing that the storekeeper can do is to make sure his inventory is either 3 or 4 at the start of the selling season.

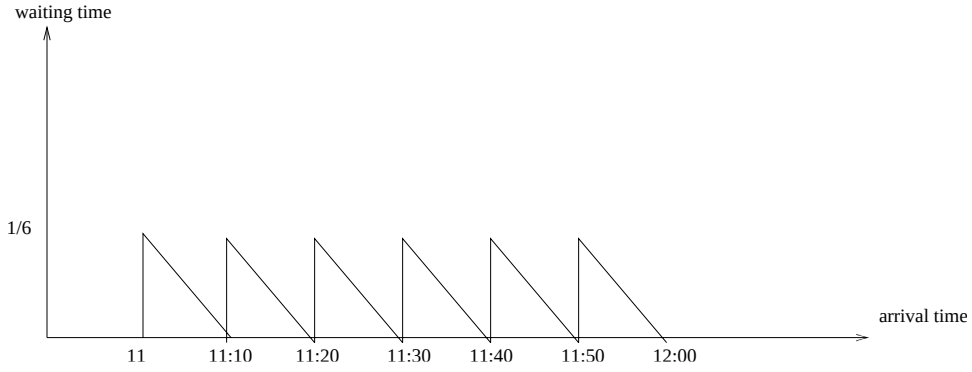
Example 15.2. In this problem, we compute how long a person would have to wait for a bus if he arrives at the bus stop at a random time within a time interval. Specifically, Bill arrives at his bus stop each day between 11am and 12 noon. His arrival time A (in hours) is a uniformly distributed RV between 11 and 12. Let us first suppose that his bus arrives at times

$$11:10, 11:20, 11:30, 11:40, 11:50, 12:00$$

His waiting time W (in hours) for the bus is then

$$W = \begin{cases} 11 + (1/6) - A, & 11 \leq A < 11 + (1/6) \\ 11 + (2/6) - A, & 11 + (1/6) \leq A < 11 + (2/6) \\ 11 + (3/6) - A, & 11 + (2/6) \leq A < 11 + (3/6) \\ 11 + (4/6) - A, & 11 + (3/6) \leq A < 11 + (4/6) \\ 11 + (5/6) - A, & 11 + (4/6) \leq A < 11 + (5/6) \\ 12 - A, & 11 + (5/6) \leq A \leq 12 \end{cases}$$

Plotting the waiting time as a function of the arrival time, you obtain the plot at the top of next page:



It should be clear from the plot that we obtain the same contribution to the overall expected waiting time from each of the 6 ten minute intervals. In other words, if we write

$$E[W] = \sum_{i=1}^{6} E[W|B_i]P[B_i], \tag{15.2}$$

where B_i is the event that Bill's arrival is in the i-th 10 minute interval, then

$$E[W|B_i] = E[W|B_1], \quad i = 1, 2, 3, 4, 5, 6.$$

Let us compute $E[W|B_1]$. Time interval B_1 goes from 11:00 to 11:10 (which is 11+(1/6) measured in hours). For A falling in interval B_1, we have

$$W = 11 + (1/6) - A.$$

Thus,

$$\begin{aligned} E[W|B_1] &= E[11 + (1/6) - A|\{A \in B_1\}] \\ &= [11 + (1/6)] - E[A|\{A \in B_1\}] \end{aligned}$$

The conditional distribution of A given $A \in B_1$ is the uniform distribution from 11 to $11 + (1/6)$. Therefore, the conditional mean is the midpoint of this interval:

$$E[A|\{A \in B_1\}] = 11 + (1/12).$$

We conclude that

$$E[W|B_1] = 1/12,$$

and therefore $E[W|B_i]$ are all equal to 1/12 hour (5 minutes). Plugging back into (15.2), we see that

$$E[W] = 1/12 \ \ (5\ minutes).$$

Exercise. Re-compute $E[W]$ assuming the bus arrives only at the times

$$11:10, 11:30, 11:50, 12:00.$$

Hint: Write

$$E[W] = \sum_{i=1}^{4} E[W|B_i]P[B_i],$$

where B_1, B_2, B_3, B_4 are respectively a 10 minute time interval, a 20 minute time interval, a 20 minute time interval, and a 10 minute time interval. You will then have

$$\begin{aligned} E[W|B_1] &= E[W|B_4] \\ E[W|B_2] &= E[W|B_3] \end{aligned}$$

which means you have to compute both $E[W|B_1]$ and $E[W|B_2]$. Your intuition should tell you what each of these are.

15.2 Joint PMF Introduction

Suppose you have two discrete RV's X and Y. Suppose you perform the underlying experiment and observe a value x for X and a value y for Y. Then you may regard the point (x, y) in the xy-plane as an observation of the random pair (X, Y). The following two events are the same:

$$\{X = x, \ \ Y = y\} = \{(X, Y) = (x, y)\}. \tag{15.3}$$

The event on the left is the event that your experiment results in value x for X and (on the same trial) value y for Y. In other words, the event on the left side of (15.3) may be thought of as

$$\{X = x\} \cap \{Y = y\},$$

the intersection of the events $\{X = x\}$ and $\{Y = y\}$. The event on the right side of (15.3) is the event that the random pair (X, Y) takes the value (x, y).

The *joint PMF* $P^{X,Y}$ of X, Y is defined for each value x of X and value y of Y by the equation

$$P^{X,Y}(x, y) \triangleq P[X = x, \ \ Y = y] = P[(X, Y) = (x, y)].$$

It satisfies the properties:

(i): $P^{X,Y}(x, y) \geq 0$, all x,y.

(ii): $\sum_{x,y} P^{X,Y}(x, y) = 1$.

(iii): For any subset E of the xy-plane, the prob that (X, Y) falls in E is computable as

$$P[(X, Y) \in E] = \sum_{(x,y)\in E} P^{X,Y}(x, y).$$

(iv): $P^X(x) = \sum_y P^{X,Y}(x, y)$, for all x.

(v): $P^Y(y) = \sum_x P^{X,Y}(x, y)$, for all y.

Notice that (iv) and (v) tell you how to compute the individual PMF's $P^X(x)$ of X and $P^Y(y)$ of Y by summing out the variable you don't want from the joint PMF. These two 1-D PMF's obtained by summing in this way from a joint PMF are called *marginal PMF's*.

Example 15.3. An urn contains three cards numbered "1", four cards numbered "2", and five cards numbered "3". Two cards are selected at random without replacement. Let X be the number on the first card selected and let Y be the number on the second card selected. The values of the random pair (X, Y) are the following nine points in the xy-plane:

$$(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (3, 3).$$

You could plot these points in the xy-plane. Instead, we will use an array of the following type similar to what we did in Chapter 1:

$$\begin{array}{c} \\ X=1 \\ X=2 \\ X=3 \end{array} \begin{array}{c} \begin{array}{ccc} Y=1 & Y=2 & Y=3 \end{array} \\ \begin{pmatrix} (X,Y)=(1,1) & (X,Y)=(1,2) & (X,Y)=(1,3) \\ (X,Y)=(2,1) & (X,Y)=(2,2) & (X,Y)=(2,3) \\ (X,Y)=(3,1) & (X,Y)=(3,2) & (X,Y)=(3,3) \end{pmatrix} \end{array}$$

We typically do this whenever X and Y take on just a finite number of values. This array is in keeping with our "discrete channel" viewpoint from Chapter 1, in which we will frequently want to consider the RV X as an input to a system and Y as the system output in response to input X. Notice how each of the nine positions in the preceding array corresponds to one of the nine possible (x, y) points which are the values of the random pair (X, Y). If in each position, we put in the appropriate joint probability, we will then get a joint PMF array which looks like

$$\begin{array}{c} \\ X=1 \\ X=2 \\ X=3 \end{array} \begin{array}{c} \begin{array}{ccc} Y=1 & Y=2 & Y=3 \end{array} \\ \begin{pmatrix} P^{X,Y}(1,1) & P^{X,Y}(1,2) & P^{X,Y}(1,3) \\ P^{X,Y}(2,1) & P^{X,Y}(2,2) & P^{X,Y}(2,3) \\ P^{X,Y}(3,1) & P^{X,Y}(3,2) & P^{X,Y}(3,3) \end{pmatrix} \end{array}$$

We can compute the joint probability that goes in each position using the following multiplication rule that follows from work we did in Chapter 1:

$$P^{X,Y}(x, y) = P^X(x)P(Y = y|X = x). \tag{15.4}$$

To see why this is true, note that the left side involves an intersection of two events

$$P^{X,Y}(x,y) = P[\{X = x\} \cap \{Y = y\}],$$

which can then be decomposed as the prob of the first event times the cond prob of the 2nd event given the 1st event. For example, we can do the following computation using (15.4):

$$P^{X,Y}(1,2) = (3/12)(4/11) = 12/132.$$

(You have 3 chances in 12 of drawing one of the 3 cards numbered "1" on the first draw, and then 4 chances in 11 of drawing one of the 4 cards numbered "2" on the second draw.) The reader may easily continue beyond this point and fill in the entire joint PMF array:

$$\begin{array}{c} \\ X=1 \\ X=2 \\ X=3 \end{array} \begin{array}{c} \begin{array}{ccc} Y=1 & Y=2 & Y=3 \end{array} \\ \begin{pmatrix} 6/132 & 12/132 & 15/132 \\ 12/132 & 12/132 & 20/132 \\ 15/132 & 20/132 & 20/132 \end{pmatrix} \end{array} \tag{15.5}$$

It is not hard to see what rules (iv) and (v) for finding the marginal PMF's become in this scenario:

- *$P^X(x)$ is obtained by computing the row sums of the joint PMF array (15.5) and $P^Y(y)$ is obtained by computing the column sums.*

The reader may check that the three row sums of array (15.5) are 33/132, 44/132, 55/132, which are respectively the same as the following PMF values for X:

$$[P^X(1) \;\; P^X(2) \;\; P^X(3)] = [3/12 \;\; 4/12 \;\; 5/12].$$

This is not a surprise, because these are easily seen to be the probs for what happens on the first draw. Notice that the joint PMF array is *symmetric*, that is, the joint prob in position (i, j) is the same as the joint prob in position (j, i). Thus, the three column sums will respectively coincide with the three row sums. That is, the Y PMF will be given by:

$$[P^Y(1) \;\; P^Y(2) \;\; P^Y(3)] = [3/12 \;\; 4/12 \;\; 5/12].$$

Is this a surprise to you? In other words, the second draw has the same individual prob dist as the first draw. What possibly confuses students initially about this example is that there are three different conditional distributions for the second draw, given each of the three possibilities for the first draw. However, the *unconditional distribution of the second draw* is not any one of these three conditional distributions; it is instead a weighted average of the 3 cond dist's. The underlying symmetry of the problem dictates that this "weighted average distribution" will be the same thing as the (unconditional) distribution of the 1st draw.

Let us now compute $P[X = Y]$, $P[X > Y]$, and $P[X < Y]$ for this example. Summing down the diagonal of the joint PMF array, we obtain

$$P[X = Y] = P^{X,Y}(1,1) + P^{X,Y}(2,2) + P^{X,Y}(3,3) = 38/132.$$

Since the array (15.5) is symmetric, $P[X > Y]$ and $P[Y > X]$ are the same. Using this fact and

$$P[X = Y] + P[X > Y] + P[Y < X] = 1,$$

we conclude that

$$P[X > Y] = P[Y > X] = (1/2)[1 - P[X = Y]] = 47/132.$$

Exercise. If you are still somewhat dubious about the fact that the probability distributions for the individual draws in sampling without replacement all coincide, you can perform a Matlab verification as follows. Write a Matlab script which will simulate one trial of the two-step random experiment of Example 15.3 (the script simulates a value of X, the result of the first draw, and then uses the simulated value of X to simulate the result Y of the second draw). By embedding your script in a for loop, you can then simulate several thousand observations of the random pair (X, Y) (from independent trials). Running this for loop will then give you a vector `x` of the simulated X observations and a vector `y` of the simulated Y observations. For each $i = 1, 2, 3$, you can then execute the Matlab commands

```
mean(x==i), mean(y==i)
```

to see if these estimates of $P(X = i)$ and $P(Y = i)$ are about the same.

Lecture 16

Chapters 4-5 Part 2

16.1 Two Joint PMF Examples

Here are two more examples on joint PMF's.

Example 16.1. Discrete RV's X, Y each take the values $0, 1, 2$. Here is the joint PMF table:

$$\begin{array}{c} \\ X=0 \\ X=1 \\ X=2 \end{array} \begin{array}{c} \begin{array}{ccc} Y=0 & Y=1 & Y=2 \end{array} \\ \left(\begin{array}{ccc} 0.1 & 0 & 0.2 \\ 0.05 & 0.2 & 0.3 \\ 0.1 & 0 & 0.05 \end{array} \right) \end{array} \tag{16.1}$$

You can check as follows that this gives a genuine PMF:

$$0.1 + 0 + 0.2 + 0.05 + 0.2 + 0.3 + 0.1 + 0 + 0.05 = 1.$$

The row sums give the marginal PMF of X:

$$\begin{aligned} P^X(0) &= 0.3 \\ P^X(1) &= 0.55 \\ P^X(2) &= 0.15 \end{aligned}$$

The column sums give the marginal PMF of Y:

$$P^Y(0) = 0.25, \quad P^Y(1) = 0.2, \quad P^Y(2) = 0.55.$$

Recall from Bayes Method the array of "forward conditional probabilities" and the array of "backward conditional probabilities". We obtain the array of forward conditional probabilities by dividing

each row of the PMF array (16.1) by the row sum of that row. This procedure yields the following array of forward conditional probabilities (after simplification):

$$\begin{array}{c} \\ X=0 \\ X=1 \\ X=2 \end{array} \begin{array}{c} \begin{array}{ccc} Y=0 & Y=1 & Y=2 \end{array} \\ \begin{pmatrix} 1/3 & 0 & 2/3 \\ 1/11 & 4/11 & 6/11 \\ 2/3 & 0 & 1/3 \end{pmatrix} \end{array} \tag{16.2}$$

For example, the entry in row 3 and column 1 of (16.2) is interpreted as the conditional probability

$$P(Y=0|X=2)=2/3.$$

Notice that the array (16.2) is a bonafide array of forward conditional probabilities because each of its rows adds up to one. You could interpret the array (16.2) as the channel matrix of a discrete channel for which X is the input to the channel and Y is the output to the channel in response to input X.

We can also obtain the array of "backward conditional probabilities" from the joint PMF (16.1) by dividing each column by the column sum for that column. This yields the following array of backward conditional probabilities (after simplification):

$$\begin{array}{c} \\ X=0 \\ X=1 \\ X=2 \end{array} \begin{array}{c} \begin{array}{ccc} Y=0 & Y=1 & Y=2 \end{array} \\ \begin{pmatrix} 2/5 & 0 & 4/11 \\ 1/5 & 1 & 6/11 \\ 2/5 & 0 & 1/11 \end{pmatrix} \end{array} \tag{16.3}$$

For example, the entry in row 2 and column 3 is interpreted as

$$P(X=1|Y=2)=6/11.$$

Notice that each column of the backward conditional probability array (16.3) sums to one.

Example 16.2. We call this the "ice cream cone" example. We perform the following two-step experiment:

Step 1: Bill eats X ice cream cones, where X has a Poisson distribution with mean 1.

Step 2: Bill flips a fair coin $X+1$ times, and then runs Y miles, where Y is the number of heads resulting from the coin flips.

Let us find the joint PMF of (X,Y). First, let us determine which (x,y) points in the xy-plane are the values of (X,Y). (There are infinitely many of them!) Notice the following:

- If $X=0$, then $Y\in\{0,1\}$.

- If $X = 1$, then $Y \in \{0, 1, 2\}$.
- If $X = 2$, then $Y \in \{0, 1, 2, 3\}$.

Continuing in this way, the reader can see that the following set S is the set of all possible values of (X, Y):

$$S = \{(i, j) : i = 0, 1, 2, \cdots; j = 0, 1, \cdots, i + 1\}.$$

For example, $(7, 5)$ is in the set S because the second coordinate is between 0 and $8 = 7 + 1$ inclusively. However, the point $(7, 10)$ is not in S.

For each point (i, j) belonging to S, we need to compute the joint PMF value $P^{X,Y}(i, j)$. We can do this by the multiplication rule from Chapter 1:

$$P^{X,Y}(i, j) = P(X = i,\ Y = j) = P^X(i)P(Y = j|X = i). \tag{16.4}$$

The probability $P^X(i)$ on the right side of (16.4) is a Poisson probability and the other probability $P(Y = j|X = i)$ is a Binomial probability. We have

$$P^X(i) = \frac{\exp(-1)}{i!} \tag{16.5}$$

$$P(Y = j|X = i) = \binom{i+1}{j}(1/2)^{i+1} \tag{16.6}$$

The formula (16.5) arises from Step 1 of the experiment and is valid because this is the probability that the Poisson RV X with mean 1 takes on the value i (see Appendix A of your textbook). The formula (16.6) arises from Step 2 of the experiment: Given $X = i$, Y conditionally has the Binomial(n, p) distribution with $n = i + 1$ and $p = 1/2$; the conditional probability of the event $\{Y = j\}$ given the event $\{X = i\}$ is then the probability that a Binomial$(n = i + 1, p = 1/2)$ RV takes on the value j, which we may see from Appendix A of the textbook to be equal to the right side of (16.6).

We have shown that the joint PMF of X, Y is given by:

$$P^{X,Y}(i, j) = \begin{cases} \binom{i+1}{j}\frac{\exp(-1)}{i!2^{i+1}}, & (i, j) \in S \\ 0, & elsewhere \end{cases}$$

It is interesting to determine the marginal PMF of Y. The values of Y are nonnegative integers

$$j = 0, 1, 2, \cdots.$$

For each such j, we compute $P^Y(j)$ by means of the following formulas:

$$P^Y(0) = \sum_{i=0}^{\infty} \frac{\exp(-1)}{i!2^{i+1}}. \tag{16.7}$$

$$P^Y(j) = \sum_{i=j-1}^{\infty} \binom{i+1}{j} \frac{\exp(-1)}{i!2^{i+1}}, \quad j = 1, 2, \cdots \tag{16.8}$$

The reader can obtain the limits on the summations in (16.7)-(16.8) by plotting the points in S and then taking a "horizontal slice" through all points (i, j) in S for which the second coordinate j is fixed. We shall examine the sums (16.7)-(16.8) further in Recitation 6. (It turns out that Y does not have a Poisson distribution, even though X has a Poisson distribution.)

16.2 Joint PDF Introduction

Suppose X and Y are both continuous RV's. Jointly, suppose the random pair (X, Y) ranges continuously through some region of the xy-plane. The joint PDF $f_{X,Y}(x, y)$ is then the function of two variables satisfying the properties:

(i): $f_{X,Y}(x, y) \geq 0$ for all x, y.

(ii): $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy = 1$.

(iii): For every subregion E of the xy-plane,

$$P[(X, Y) \in E] = \iint_E f_{X,Y}(x, y) dx dy. \tag{16.9}$$

(iv): The marginal PDF $f_X(x)$ can be computed by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy, \quad -\infty < x < \infty.$$

(v): The marginal PDF $f_Y(y)$ can be computed by

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx, \quad -\infty < y < \infty.$$

(Note: It turns out that if (X, Y) is jointly continuously distributed, then the individual RV's X, Y are each continuously distributed over the real line. This follows from the formulas in (iv) and (v) for computing the marginal densities.)

Example 16.3. Suppose you have some region R of the xy-plane of positive finite area. Suppose your random experiment is to "select a point (X, Y) at random from R". Then, in the absence of any further information, you would assume that the joint density $f_{X,Y}(x, y)$ is constant over R. This is what the joint density would have to be in this case:

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{area(R)}, & (x, y) \in R \\ 0, & (x, y) \notin R \end{cases}$$

The area of R can be computed by the double integral

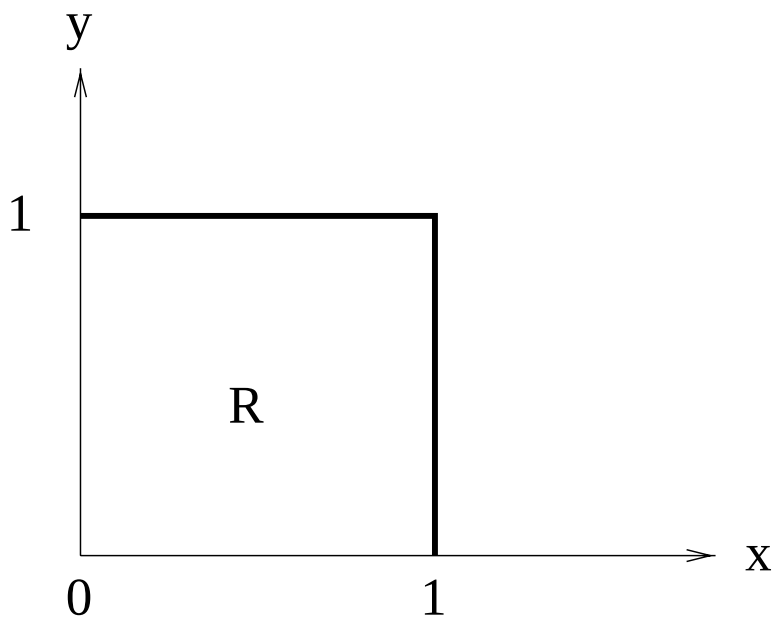
$$area(R) = \iint_R dxdy.$$

The reader can easily show that the constant $1/area(R)$ is the unique constant value over R that will make the property(ii) of joint density be true, namely, the property that the double integral of the joint PDF be equal to one. Using formula (16.9), one can easily show in this case that

$$P[(X,Y) \in E] = \frac{area(E)}{area(R)},$$

for any subregion E of R. The random pair (X,Y) of this example is said to be *uniformly distributed over* R. We shall occasionally use uniformly distributed pairs of RV's as examples to illustrate various concepts encountered in Chapter 4.

Example 16.4. Let R be the unit square region sketched below.



A pair of random variables X, Y has the following joint density $f_{X,Y}(x,y)$:

$$f_{X,Y}(x,y) = \begin{cases} C(x+y), & (x,y) \in R \\ 0, & elsewhere \end{cases}$$

We will do the following:

(a) Find C.

(b) Compute $P[0 \leq X \leq 1/4,\ 3/4 \leq Y \leq 1]$

(c) Compute $P[X > 2Y]$

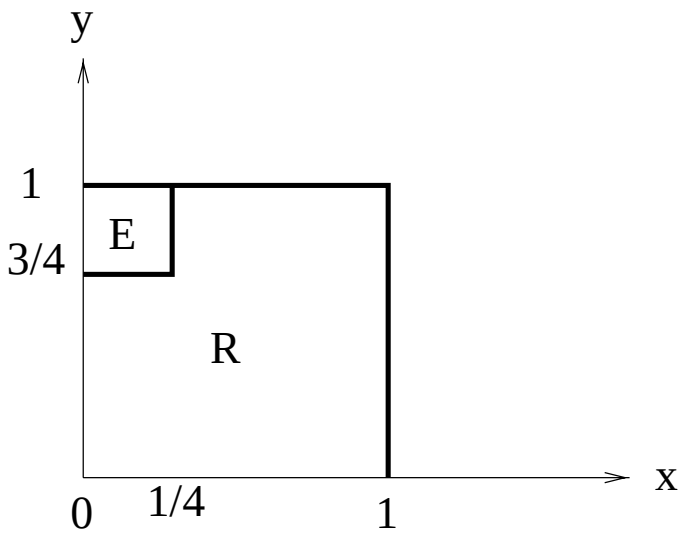
(d) Compute $P[X^2 + Y^2 \leq 1]$

(e) Find $f_X(x)$ and $f_Y(y)$.

Solution to (a). The PDF must integrate to 1 over the entire xy-plane. Therefore,

$$C = \frac{1}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy} = \frac{1}{\int_0^1\int_0^1 (x+y)dxdy} = 1.$$

Solution to (b). Let E be the square subregion of R sketched below.

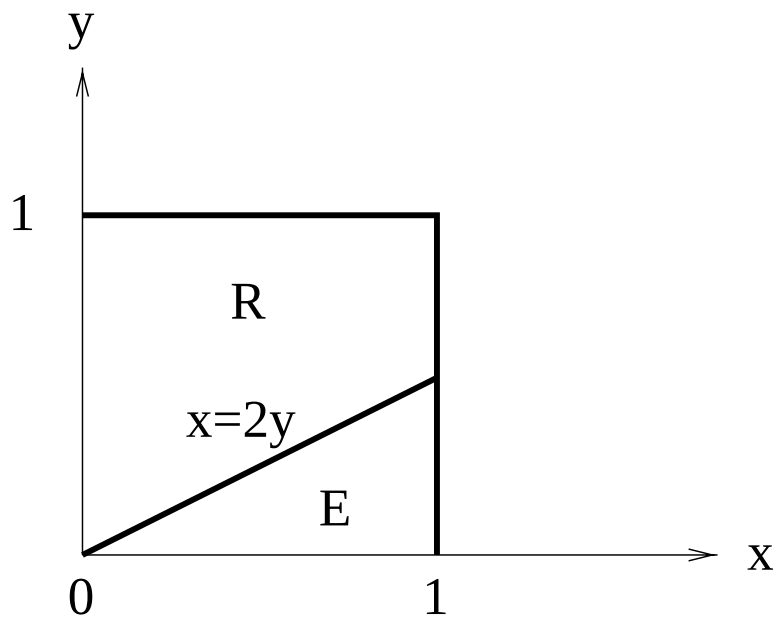


Then, we have

$$\begin{aligned} P[0 \leq X \leq 1/4,\ 3/4 \leq Y \leq 1] &= P[(X,Y) \in E] \\ &= \iint_E f_{X,Y}(x,y)dydx \\ &= \int_0^{1/4}\int_{3/4}^1 (x+y)dydx \end{aligned}$$

$$= \int_0^{1/4} \left[xy + y^2/2\right]_{y=3/4}^{y=1} dx$$
$$= \int_0^{1/4} (x/4 + 7/32)dx = 1/16.$$

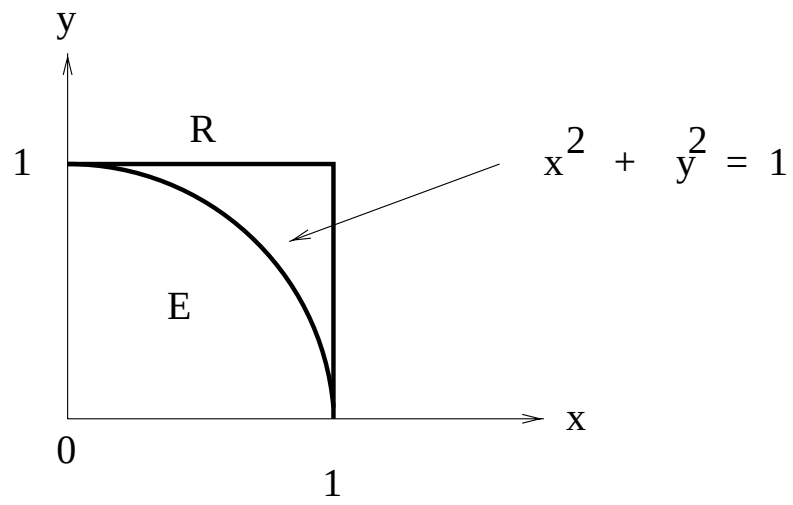
Solution to (c). Let E be the following triangular subregion of R:



We have

$$P[X > 2Y] = \iint_E f_{X,Y}(x,y)dydx$$
$$= \int_0^1 \int_0^{x/2} (x+y)dydx$$
$$= \int_0^1 \left[xy + y^2/2\right]_{y=0}^{y=x/2} dx$$
$$= \int_0^1 (5x^2/8)dx = 5/24.$$

Solution to (d). Let E be the circular sector sketched at the top of the next page:



Then

$$\begin{aligned} P[X^2+Y^2 \le 1] &= \iint_E f_{X,Y}(x,y)dydx \\ &= \int_0^1 \int_0^{\sqrt{1-x^2}} (x+y)dydx \\ &= \int_0^1 \left[xy + y^2/2\right]_{y=0}^{y=\sqrt{1-x^2}} dx \\ &= \int_0^1 [x\sqrt{1-x^2} + (1-x^2)/2]dx \\ &= \left[(-1/3)(1-x^2)^{3/2} + (1/2)(x - x^3/3)\right]_0^1 = 2/3 \end{aligned}$$

Solution to (e). Let x be fixed in the range $0 \le x \le 1$. Then:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_0^1 (x+y)dy = x + (1/2).$$

The density $f_X(x)$ must be zero for all other x, because for such x the vertical slice through x does not touch the region R. We conclude

$$f_X(x) = \begin{cases} x + (1/2), & 0 \le x \le 1 \\ 0, & elsewhere \end{cases}$$

By symmetry,

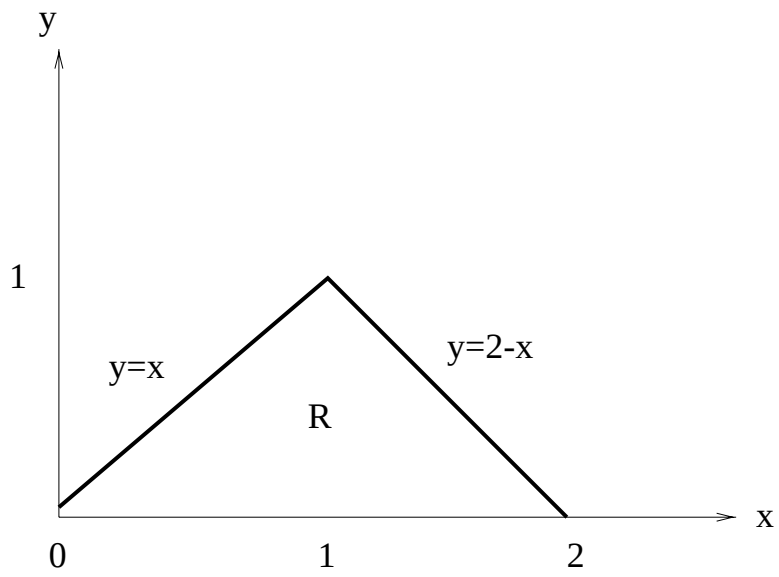
$$f_Y(y) = \begin{cases} y + (1/2), & 0 \le y \le 1 \\ 0, & elsewhere \end{cases}$$

Lecture 17

Chapters 4-5 Part 3

17.1 Another Joint PDF Example

Example 17.1. Let R be the triangular region sketched below:



Let (X, Y) be uniformly distributed over R. The area of R is 1. Therefore, the joint density is given by

$$f_{X,Y}(x,y) = \begin{cases} 1, & (x,y) \in R \\ 0, & elsewhere \end{cases}$$

Are the random variables X and Y individually uniformly distributed? Let us see whether or not this is true. To find $f_X(x)$, we fix an x in the interval $0 \leq x \leq 2$ and integrate the joint density with respect to y. The range over which y is integrated depends upon whether $0 \leq x \leq 1$ or $1 < x \leq 2$. In the first case, y ranges from $y = 0$ to $y = x$ (visualize a vertical slice cutting through R, extending upward from an x satisfying $0 \leq x \leq 1$). In the second case, y ranges from $y = 0$ to $y = 2 - x$ (visualize a vertical slice cutting through R, extending upward from an x satisfying $1 < x \leq 2$). This gives us

$$\begin{aligned} f_X(x) &= \int_0^x 1\,dy = x, \quad 0 \leq x \leq 1 \\ &= \int_0^{2-x} 1\,dy = 2 - x, \quad 1 < x \leq 2 \end{aligned}$$

The complete description of $f_X(x)$ is therefore:

$$f_X(x) = \begin{cases} x, & 0 \leq x \leq 1 \\ 2 - x, & 1 < x \leq 2 \\ 0, & elsewhere \end{cases}$$

To find the marginal density $f_Y(y)$, fix a y satisfying $0 \leq y \leq 1$ on the y-axis and make a horizonal slice through y and the region R. The slice extends from $x = y$ to $x = 2 - y$. This gives us

$$f_Y(y) = \int_y^{2-y} 1\,dy = 2 - 2y, \quad 0 \leq y \leq 1.$$

The complete description of $f_Y(y)$ is therefore:

$$f_Y(y) = \begin{cases} 2 - 2y, & 0 \leq y \leq 1 \\ 0, & elsewhere \end{cases}$$

We conclude from the preceding example that if (X, Y) is uniformly distributed over a subregion of the plane, then it may happen that the marginal densities are *not* uniform.

Exercise. Let (X, Y) be uniformly distributed over the rectangular rection whose four vertices are the points (a, c), (a, d), (b, c), (b, d). (Assume that $a < b$ and $c < d$.) Prove that X must be Uniform(a, b) and Y must be Uniform(c, d).

As a result of the preceding exercise, you see one special case in which a joint uniform distribution yields marginal uniform distributions, namely, when the joint distribution is over a rectangle whose two dimensions are parallel to the x and y axes.

17.2 Application: MAP Detector for Discrete Channel Model

Suppose we have a discrete communication channel model: this means the input is discrete and the output is discrete. At the receiving end of the channel, one can try to design a detector as indicated in the following block diagram:

$$X \to \boxed{\text{channel}} \to Y \to \boxed{\text{detector}} \to \hat{X}$$

The detector output $\hat{X}$ is a function of the channel output Y, and is intended to estimate X. In general, there may be many possible detectors that one could use. For example, suppose that X takes the values $0, 1, 2$ and that Y also takes the values $0, 1, 2$. To define the detector to be used, one would have to fill in the question marks below:

$$\begin{aligned} Y = 0 &\Rightarrow \hat{X} = ? \\ Y = 1 &\Rightarrow \hat{X} = ? \\ Y = 2 &\Rightarrow \hat{X} = ? \end{aligned}$$

Each question mark can be filled in with one of three possible estimates for X (namely $0, 1$, or 2). Thus, in this case, there are $3 * 3 * 3 = 27$ possible detectors that could be used. We want to find the detector for which the error probability $P[X \neq \hat{X}]$ is minimized. This is the detector that does the best job, and is the detector we will want to use in our system. This best detector is called the MAP detector. (We will explain later what the initials "MAP" stand for.) One could try to find the MAP detector by a brute force approach, that is, one could compute $P[X \neq \hat{X}]$ for every possible detector until the detector is found which minimizes $P[X \neq \hat{X}]$. This approach will be inefficient when X and Y take a large number of values. (If X, Y each take 10 values, you'd have to check $(10)^{10}$ different detectors!) Fortunately, we can narrow our search. It is known that the MAP detector, if $Y = y$, generates the estimate $\hat{X} = x$ for which the joint PMF $P^{X,Y}(x, y)$ is maximized over all possible inputs x to the channel (as y in $P^{X,Y}(x, y)$ is held fixed). In other words, the MAP detector chooses the most likely channel input for the given channel output. Rather than give a general proof that the MAP detector defined in this way will minimize $P[X \neq \hat{X}]$, we will now work out an example of MAP detector design; during the workout of this example, the reader may become convinced that choosing the MAP detector the way we do is the proper thing to do.

Example 17.2. Let the channel matrix for a discrete channel be

$$\begin{array}{c} \\ X = 0 \\ X = 1 \\ X = 2 \end{array} \begin{array}{c} \begin{array}{ccc} Y = 0 & Y = 1 & Y = 2 \end{array} \\ \begin{pmatrix} 3/42 & 36/42 & 3/42 \\ 9/21 & 6/21 & 6/21 \\ 18/37 & 18/37 & 1/37 \end{pmatrix} \end{array} \tag{17.1}$$

We assume that the channel input probabilities are

$$[P^X(0) \;\; P^X(1) \;\; P^X(2)] = [.42 \;\; .21 \;\; .37].$$

Let us design the MAP detector for the given channel and input to the channel. First, we compute the joint PMF matrix by multiplying each row of the channel matrix (17.1) by the corresponding input probability. This gives us the following joint PMF matrix:

$$\begin{array}{c} \\ X=0 \\ X=1 \\ X=2 \end{array} \begin{array}{c} \begin{array}{ccc} Y=0 & Y=1 & Y=2 \end{array} \\ \begin{pmatrix} 0.03 & 0.36 & 0.03 \\ 0.09 & 0.06 & 0.06 \\ 0.18 & 0.18 & 0.01 \end{pmatrix} \end{array} \tag{17.2}$$

We can now easily design the MAP detector from the array (17.2), as follows:

- For each possible MAP detector input $y \in \{0, 1, 2\}$, look down the column labelled $Y = y$ of the joint PMF array to find the largest entry in that column. (If there are two or more entries which are the largest, choose any of them.)
- Take the MAP detector output for input $Y = y$ to be the input x value corresponding to the largest entry that was chosen in the $Y = y$ column.

In the following, I have placed a box around the largest entry in each column of the joint PMF array:

$$\begin{array}{c} \\ X=0 \\ X=1 \\ X=2 \end{array} \begin{array}{c} \begin{array}{ccc} Y=0 & Y=1 & Y=2 \end{array} \\ \begin{pmatrix} 0.03 & \boxed{0.36} & 0.03 \\ 0.09 & 0.06 & \boxed{0.06} \\ \boxed{0.18} & 0.18 & 0.01 \end{pmatrix} \end{array}$$

Looking at the x-value corresponding to the location of each boxed in value, we see that the MAP detector can be described as follows:

$$\begin{aligned} Y = 0 &\Rightarrow \hat{X} = 2 \\ Y = 1 &\Rightarrow \hat{X} = 0 \\ Y = 2 &\Rightarrow \hat{X} = 1 \end{aligned}$$

Now let us compute the error probability for the MAP detector. We have

$$\begin{aligned} P[X \neq \hat{X}] &= 1 - P[X = \hat{X}] \\ &= 1 - (P[X = 0, \hat{X} = 0] + P[X = 1, \hat{X} = 1] + P[X = 2, \hat{X} = 2]) \\ &= 1 - (P[X = 0, Y = 1] + P[X = 1, Y = 2] + P[X = 2, Y = 0]) \\ &= 1 - (.36 + .06 + .18) = .4 \end{aligned}$$

From this calculation, it should be clear that if we had chosen our detector in any other way, then the error probability would have been bigger than .4. For, notice that the three probabilities

$P[X = 2, Y = 0]$, $P[X = 0, Y = 1]$, and $P[X = 1, Y = 2]$ are the largest probabilities in the first, second, and third columns of the $P^{X,Y}$ matrix, respectively. If the detector had been chosen differently, at least one of these probabilities would have been replaced in the calculation above by some smaller probability in the same column, thereby increasing $P[X \neq \hat{X}]$.

A Useful Formula

From the calculation of the error probability $P[X \neq \hat{X}]$ that we did above, the reader can see that, in general, the error probability for the MAP detector is computed in the following way:

$$\text{minimum } P[X \neq \hat{X}] = 1 - (\text{sum of largest prob. in each col. of } P^{X,Y} \text{ array})$$

Discussion. We explain where the terminology "MAP" comes from. We can factor the joint PMF as

$$P^{X,Y}(x, y) = P[X = x,\ Y = y] = P^Y(y)P[X = x|Y = y]. \tag{17.3}$$

For each fixed y that can be the input to the MAP detector, the MAP detector chooses as its output that x value for which $P^{X,Y}(x, y)$ is a maximum. By the preceding equation, this will be the same as the x which maximizes the "backward probability" $P[X = x|Y = y]$. (This is because the factor $P^Y(y)$ on the right side of (17.3) is treated as a constant when y is held fixed.) In Latin, the backward probability $P[X = x|Y = y]$ is referred to as "aposteriori probability". Thus, the MAP detector operates on the principle of "maximum backward probability" or "maximum aposteriori probability", which is MAP for short.

Exercise. Consider the discrete channel with input alphabet $\{0, 1, 2\}$, output alphabet $\{0, 1, 2\}$, and channel matrix

$$\begin{bmatrix} 1-p & p/2 & p/2 \\ p/2 & 1-p & p/2 \\ p/2 & p/2 & 1-p \end{bmatrix}$$

Let the input X to the channel be equiprobable. Find a range of p values for which the MAP detector simply takes $\hat{X} = Y$. (In other words, you make a guess that the input is the observed output.)

Remark. In Example 17.2, the best error probability is 0.40. This may seem to the reader to not be very good performance. (For example, if we run thousands of inputs through the channel, the MAP detector is only going to correctly guess about 60% of them.) To get better performance, one can consider *batch processing*. For example, suppose we try to design a detector which processes two outputs at a time (in response to two inputs at a time), as indicated in the following block diagram:

$$X_1, X_2 \to \boxed{\text{channel}} \to Y_1, Y_2 \to \boxed{\text{detector}} \to \hat{X}_1, \hat{X}_2$$

One could see whether, for the channel model in Example 17.2, there is such a "batch processing detector" which achieves average error probability

$$(P[X_1 \neq \hat{X}_1] + P[X_2 \neq \hat{X}_2])/2 < 0.40.$$

One can always design a communication system for which such an improvement is possible, provided

- The size of the batch is big enough (that is, the batch may have to be of size bigger than two);
- You encode the batch of inputs appropriately before sending them through the channel; and
- You are not trying to transmit data at a rate which would exceed the capacity of the channel.

In a more advanced course (EE 5581, Information Theory) you would learn about the communication principles encapsulated in the three preceding "bullets". In particular, you would be able to formulate the notion of the capacity of a channel in a precise mathematical way. You would also learn about how to properly encode the channel inputs (using linear block error-correcting codes).

We have discussed the MAP detector for the discrete communication channel model. Later, we will be able to develop the MAP detector for the semicontinuous communication channel model; this is the channel model in which the input is discrete and the output is continuous. The semicontinuous channel is very common in practice. For example, the additive noise channel model is a semicontinuous channel model when the noise you add to the input is continuous (such as Gaussian noise).

Lecture 18

Chapters 4-5 Part 4

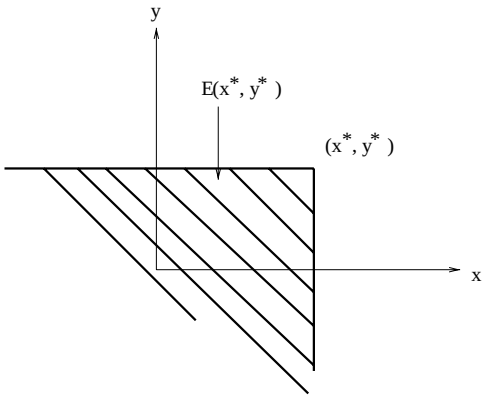
In Lecture 18, I talk about the joint CDF $F_{X,Y}(x,y)$ and I explain about the concept of independent RV's X, Y.

18.1 Joint CDF $F_{X,Y}(x,y)$

Let X, Y be an arbitrary pair of random variables. The joint cumulative distribution function (joint CDF) of these variables is the function $F_{X,Y}(x,y)$ defined by

$$F_{X,Y}(x,y) \triangleq P[X \leq x,\ Y \leq y], \quad all\ (x,y)\ in\ xy-plane.$$

In order to lend more insight to this definition, for each point (x^*, y^*) in the plane, let $E(x^*, y^*)$ be the subregion of the xy-plane sketched as follows:



Notice that $E(x^*, y^*)$ is the infinite quadrant of the xy-plane which extends to the left and below the point (x^*, y^*). Mathematically, we can describe this set as

$$E(x^*, y^*) = \{(x, y) : x \leq x^*, \;\; y \leq y^*\}.$$

We have the following geometric interpretation of the joint CDF value $F_{X,Y}(x^*, y^*)$:

$$F_{X,Y}(x^*, y^*) = P[(X, Y) \in E(x^*, y^*)].$$

The concept of joint CDF $F_{X,Y}(x, y)$ makes sense no matter what type of RV X is and no matter what type of RV Y is. Thus, X could be discrete, continuous or mixed, and Y could be discrete, continuous, or mixed, which gives 9 possibilities for the pair (X, Y). For this reason, the joint CDF is sometimes useful when we want to put forth some new concept for the pair (X, Y): we can define the concept using the joint CDF and have the concept be valid in general without having to consider a number of special cases.

If X, Y are discrete, we can use the joint PMF to compute the joint CDF as the following double sum:

$$F_{X,Y}(x, y) = \sum_{u \leq x} \sum_{v \leq y} P^{X,Y}(u, v).$$

On the other hand, if X, Y are jointly continuous, we can use the joint PDF to compute the joint CDF as the following double integral:

$$F_{X,Y}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u, v) du dv. \tag{18.1}$$

The concept of joint CDF has proved to be more useful for the case of jointly continuous RV's X, Y than for the case of discrete RV's. Consequently, in the section which follows, we discuss special things we can do concerning the joint CDF for jointly continuous RV's.

18.1.1 Joint CDF of Jointly Continuous RV's

Throughout this section, we take X, Y to be jointly continuous RV's. It is clear from (18.1) that we can obtain the joint PDF $f_{X,Y}(x, y)$ from the joint CDF $F_{X,Y}(x, y)$ as follows:

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y). \tag{18.2}$$

Or, one can do the partial derivatives in (18.2) in the reverse order.

Example 18.1. Suppose (X, Y) is jointly distributed in the square

$$R = \{(x, y) : 0 \leq x \leq 1; \; 0 \leq y \leq 1\}.$$

Also, suppose the joint CDF is specified in R as

$$F_{X,Y}(x,y) = (1/2)x^2y + (1/2)xy^2, \quad (x,y) \in R. \tag{18.3}$$

We find $f_{X,Y}(x,y)$ using (18.2). Taking the partial derivative of $F_{X,Y}(x,y)$ with respect to y, we obtain

$$\frac{\partial F_{X,Y}(x,y)}{\partial y} = (1/2)x^2 + xy.$$

Taking the partial with respect to x then gives:

$$\frac{\partial\{(1/2)x^2 + xy\}}{\partial x} = x + y.$$

We conclude that

$$f_{X,Y}(x,y) = \begin{cases} x+y, & (x,y) \in R \\ 0, & elsewhere \end{cases}$$

Example 18.2. We work out the joint CDF for the joint density

$$f_{X,Y}(x,y) = \begin{cases} e^{-(x+y)}, & x \geq 0,\ y \geq 0 \\ 0, & elsewhere \end{cases}$$

If (x,y) is in the 1st quadrant, then

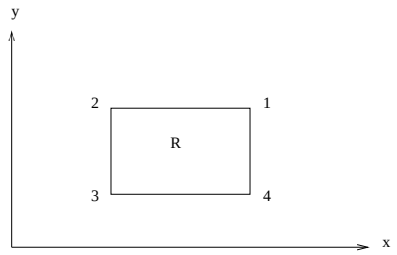
$$F_{X,Y}(x,y) = \int_0^x \int_0^y e^{-(x+y)} dy dx = (1-e^{-x})(1-e^{-y}).$$

Since the joint CDF vanishes outside of the first quadrant, we may write

$$F_{X,Y}(x,y) = (1-e^{-x})(1-e^{-y})u(x)u(y).$$

Useful Formula

Let R be the rectangular region in the following sketch:

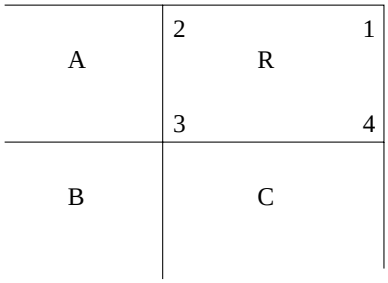


Then,

$$P[(X,Y) \in R] = F_{X,Y}(1) - F_{X,Y}(2) + F_{X,Y}(3) - F_{X,Y}(4). \tag{18.4}$$

In other words, the probability (X,Y) falls in R is computable from the joint CDF by adding together the joint CDF values at the upper left and lower right corner points of R, and then subtracting off from this sum the joint CDF values at the remaining two corner points of R.

Proof of Useful Formula (18.4): We use the following figure:



We have

$$\begin{aligned} F(1) &= P(R) + P(A) + P(B) + P(C) \\ F(2) &= P(A) + P(B) \\ F(3) &= P(B) \\ F(4) &= P(B) + P(C) \end{aligned}$$

(We have dropped the subscripts X,Y from $F_{X,Y}$ and abbreviated $P[(X,Y) \in A]$ as $P(A)$, etc.) Plugging these expressions for $F(1)$ through $F(4)$ in the right side of (18.4), we obtain cancellation of all terms except the left side of (18.4).

Example 18.3. In the previous example, suppose that $b > a > 0$, and that $d > c > 0$. Let us compute

$$P[a \leq X \leq b, \;\; c \leq Y \leq d].$$

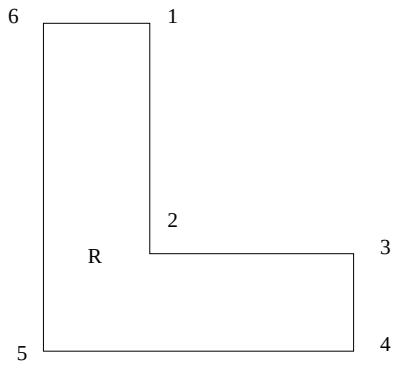
By our useful formula, we obtain

$$F(b,d) - F(a,d) + F(a,c) - F(b,c),$$

which becomes

$$(1-e^{-b})(1-e^{-d}) - (1-e^{-a})(1-e^{-d}) + (1-e^{-a})(1-e^{-c}) - (1-e^{-b})(1-e^{-c}) = (e^{-a}-e^{-b})(e^{-c}-e^{-d}).$$

Exercise. Let R be the L-shaped region in the xy-plane depicted below, and let F be the joint CDF of a pair of jointly continuous RV's X, Y.



Show that

$$P[(X,Y) \in R] = F(1) - F(2) + F(3) - F(4) + F(5) - F(6).$$

Hint: Partition R into two rectangles, and then use formula (18.4) on each of the rectangles.

18.2 Independent X, Y

We define RV's X, Y to be (statistically) independent if the following holds:

$$F_{X,Y}(x,y) = F_X(x)F_Y(y), \quad for\ all\ x, y. \tag{18.5}$$

In other words, independence means the joint CDF factors into the product of the two marginal CDF's. If X, Y are not independent, then we say that they are (statistically) dependent.

The defining statement (18.5) for independence is equivalent to saying that

$$P[X \in A, Y \in B] = P[X \in A]P[Y \in B],$$

for any two subsets A, B of the real line. In particular, we can say for independent RV's X, Y that

$$P[a \le X \le b, \ \ c \le Y \le d] = P[a \le X \le b]P[c \le Y \le d].$$

The merit to defining independence via equation (18.5) is that the definition makes sense regardless of the type of RV's X, Y are. In the special cases of discrete X, Y or of jointly continuous X, Y, then we can give simpler formulations of the independence concept, which avoid dealing with the joint CDF. We do this in the two sections which follow.

18.2.1 Independent Discrete RV's

To verify independence for discrete X, Y, you just have to show that the joint PMF factors as

$$P^{X,Y}(x, y) = P^X(x)P^Y(y), \tag{18.6}$$

for all values x of X and all values y of Y.

Example 18.4. Let the joint PMF table of discrete RV's X, Y be as follows:

$$\begin{array}{c} \\ X=0 \\ X=2 \end{array} \begin{array}{c} \begin{array}{ccc} Y=1 & Y=2 & Y=3 \end{array} \\ \begin{pmatrix} 0.15 & 0.06 & 0.09 \\ 0.35 & 0.14 & 0.21 \end{pmatrix} \end{array}$$

I explain three methods for showing that X, Y are independent.

Method 1: *Calculate the row sums and the column sums. In each position in the joint PMF matrix, check that the entry is equal to the product of the row sum for that row and the column sum for that column.* In our case here, I just enter in the row sums and column sums as headers to the rows and columns:

$$\begin{array}{c} \\ 0.3 \\ 0.7 \end{array} \begin{array}{c} \begin{array}{ccc} 0.5 & 0.2 & 0.3 \end{array} \\ \begin{pmatrix} 0.15 & 0.06 & 0.09 \\ 0.35 & 0.14 & 0.21 \end{pmatrix} \end{array}$$

In each of the 6 positions, I do a product check, for example, two of these checks would be:

$$\begin{aligned} 0.15 &= (0.30)(0.5) \\ 0.14 &= (0.7)(0.2) \end{aligned}$$

The reader can do the other four checks.

Method 2: *Divide each row by the row sum, and see if all the resulting "normalized rows" are equal to the same probability vector.* In our case here, dividing each row by its row sum gives the same prob vector, namely,

$$[0.5, 0.2, 0.3].$$

Method 3: *Divide each column by the column sum, and see if all the resulting "normalized columns" are equal to the same probability vector.* In our case here, dividing each column by its column sum gives the same prob vector

$$\begin{bmatrix} 0.3 \\ 0.7 \end{bmatrix}$$

Some brief remarks are in order as to why the three separate methods discussed in Example 18.4 work. Method 1 works because it verifies (18.6). (The row sums are the $P^X(x)$ values and the column sums are the $P^Y(y)$ values.) Method 2 works because you are verifying that

$$P(Y = y|X = x) = P(Y = y), \tag{18.7}$$

for all x, y. (Dividing each row by the row sum gives the conditional probabilities on the left side of (18.7), and (18.7) is equivalent to (18.6) if one writes down the ratio $P^{X,Y}(x, y)/P^X(x)$ for the left side of (18.7).) Method 3 works by reasoning similar to what was just given for Method 2. (Dividing each column by the column sum gives the backward conditional probs $P(X = x|Y = y)$.)

18.2.2 Independent Jointly Continuous RV's

To verify independence for jointly continuous X, Y, you just have to show that the joint PDF factors as

$$f_{X,Y}(x, y) = f_X(x)f_Y(y), \ \ all\ (x, y)\ in\ xy - plane.$$

We present two nice rules which allow us to tell by inspection of the joint density whether X and Y are statistically independent or statistically dependent. You should apply these two rules in the order indicated.

Cartesian Product Rule: We call a subset S of the xy-plane a *Cartesian product set* if it takes the form

$$S = A \times B = \{(x, y) : x \in A,\ y \in B\},$$

for some subsets A, B of the real line.[1] Suppose the joint density $f_{X,Y}(x, y)$ of random variables X, Y takes positive values over a subregion of the xy-plane which *is not* a Cartesian product set. Then, the random variables X, Y are statistically dependent.

Factorization Rule: Suppose the joint density $f_{X,Y}(x, y)$ of random variables X, Y takes positive values over a subregion of the xy-plane which *is* a Cartesian product set $S = A \times B$. Then X, Y are independent if and only if there is some factorization of the joint density

$$f_{X,Y}(x, y) = \Phi(x)\Psi(y) \tag{18.8}$$

which is valid over S, where $\Phi(x)$ is a function of x alone that is defined for $x \in A$, and $\Psi(y)$ is a function of y alone that is defined for $y \in B$. The function $\Phi(x)$ turns out to be a constant multiple of the density $f_X(x)$, and the function $\Psi(y)$ turns out to be a constant multiple of the density $f_Y(y)$.

[1] For example, any rectangular region whose two dimensions are parallel to the coordinate axes is a Cartesian product set.

Example 18.5. Suppose the joint density is of the form

$$f_{X,Y}(x,y) = Ce^{-(16x^2+9y^2)/288},$$

for some positive constant C. The density factors as

$$f_{X,Y}(x,y) = \left[\sqrt{C}e^{-x^2/18}\right]\left[\sqrt{C}e^{-y^2/32}\right]$$

over the entire xy-plane, and the entire xy-plane is a Cartesian product set. By the Factorization Rule, X and Y are independent, and the X, Y marginal densities take the form

$$\begin{aligned} f_X(x) &= C_1 e^{-x^2/18} \\ f_Y(y) &= C_2 e^{-y^2/32} \end{aligned}$$

where C_1 and C_2 are constants. It is clear from the form of these two marginal densities that X is Gaussian with mean $\mu_X = 0$ and variance $\sigma_X^2 = 9$, whereas Y is Gaussian with mean $\mu_Y = 0$ and variance $\sigma_Y^2 = 16$.

Example 18.6. Assume that

$$f_{X,Y}(x,y) = \begin{cases} e^{-(x+y)}, & x \geq 0,\ y \geq 0 \\ 0, & elsewhere \end{cases}$$

The region of positivity of this density is the first quadrant of the xy-plane, which is a Cartesian product set. We have an obvious factorization

$$e^{-(x+y)} = e^{-x}e^{-y}$$

over the first quadrant. Therefore X, Y are independent and the two marginal densities are

$$\begin{aligned} f_X(x) &= e^{-x}u(x) \\ f_Y(y) &= e^{-y}u(y) \end{aligned}$$

Example 18.7. Let the joint density be equal to $x + y$ over the unit square

$$\{(x,y) : 0 \leq x \leq 1,\ 0 \leq y \leq 1\}, \tag{18.9}$$

and zero elsewhere. The unit square is a Cartesian product set. However, the random variables X and Y are statistically dependent because $x + y$ does not factor into a function of x alone times a function of y alone over the unit square.

Example 18.8. Let $f_{X,Y}(x,y) = Cxy$ over the unit square (18.9) and zero elsewhere. The unit square is the Cartesian product set

$$\{x : 0 \leq x \leq 1\} \times \{y : 0 \leq y \leq 1\}.$$

There is clearly a factorization over the unit square:

$$f_{X,Y}(x,y) = (\sqrt{C}x)(\sqrt{C}y). \tag{18.10}$$

The Factorization Rule tells us that X, Y are independent. The two marginal densities $f_X(x)$ and $f_Y(y)$ are respectively constant multiples of x and y over the sets $\{x : 0 \leq x \leq 1\}$ and $\{y : 0 \leq y \leq 1\}$. With a little more work, we see that these two densities are:

$$f_X(x) \triangleq \begin{cases} 2x, & 0 \leq x \leq 1 \\ 0, & elsewhere \end{cases}$$

$$f_Y(y) \triangleq \begin{cases} 2y, & 0 \leq y \leq 1 \\ 0, & elsewhere \end{cases}$$

Therefore, C is 4.

Example 18.9. Let the joint density be equal to Cxy over the triangular region $\{(x,y) : 0 \leq x \leq y \leq 1\}$, and equal to zero elsewhere. This triangular region is not a Cartesian product set. By the Cartesion Product Rule, X, Y are statistically dependent. This example is tricky: Some students might try to conclude that X and Y are independent based on the Factorization Rule, but the factorization (18.10) does not hold over a Cartesian product set, only over a region which is not a Cartesian product set.

Example 18.10. Suppose the region where $f_{X,Y}(x,y)$ takes positive values is the region inside the circle $x^2 + y^2 = 1$. This region is not a Cartesian product set. Therefore, X, Y are dependent.

Exercise. Let the joint density be

$$f_{X,Y}(x,y) = \begin{cases} 4/9, & 1 \leq x \leq 2, \ 1 \leq y \leq 2 \\ 2/9, & 3 \leq x \leq 4, \ 1 \leq y \leq 2 \\ 2/9, & 1 \leq x \leq 2, \ 3 \leq y \leq 4 \\ 1/9, & 3 \leq x \leq 4, \ 3 \leq y \leq 4 \\ 0, & elsewhere \end{cases}$$

First, argue that the region of positivity of $f_{X,Y}(x,y)$ is a Cartesian product set. Then, apply the Factorization Rule to show that X, Y are independent.

Lecture 19

Chapters 4-5 Part 5

19.1 Computing $E[\phi(X,Y)]$

Let $\phi(X,Y)$ be a function of the random pair (X,Y). Then you compute $E[\phi(X,Y)]$ as follows:

- If X and Y are discrete, you compute the double sum

$$E[\phi(X,Y)] = \sum_{x,y} \phi(x,y) P^{X,Y}(x,y).$$

- If X, Y are jointly continuous, you compute the double integral

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

19.1.1 Special Case: $E[X+Y]$

Possibly the most common function $\phi(X,Y)$ needed in probability is the sum function

$$\phi(X+Y) = X+Y.$$

In this case, you can write

$$E[X+Y] = E[X] + E[Y]. \tag{19.1}$$

In other words, you can take the expected value term by term. To compute the two separate terms of the right hand side of (19.1), you'd only need to know the marginal dist's of X and Y. Here I prove (19.1) for the case in which X, Y are jointly continuous:

$$\begin{aligned} E[X+Y] &= \int\int (x+y) f_{X,Y}(x,y) dxdy \\ &= \int [\int x f_{X,Y}(x,y) dy] dx + \int [\int y f_{X,Y}(x,y) dx] dy \end{aligned}$$

$$\begin{aligned} &= \int x[\int f_{X,Y}(x,y)dy]dx + \int y[\int f_{X,Y}(x,y)dx]dy \\ &= \int x f_X(x)dx + \int y f_Y(y)dy \\ &= E[X] + E[Y] \end{aligned}$$

You can easily modify this argument for the case when X, Y are discrete. In addition, equation (19.1) is true no matter what type of RV X is and no matter what type of RV Y is.

More generally, you can take the expected value of any linear combination of RV's term by term:

$$E\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i E[X_i].$$

In the preceding, the c_i's are constants and the X_i's are RV's.

19.1.2 Special Case: $E[g_1(X)g_2(Y)]$, for X, Y independent

Suppose X, Y are independent RV's. Then, you can easily compute $E[g_1(X)g_2(Y)]$, the expected value of any function of X times any function of Y. Here is how you do it:

$$E[g_1(X)g_2(Y)] = E[g_1(X)]E[g_2(Y)], \quad X, Y \text{ independent} \tag{19.2}$$

In other words, you can compute $E[g_1(X)]$ and $E[g_2(Y)]$ separately, and then multiply the two results together. These two separate expected values would involve only the separate marginal densities, not the joint density. It is easy to prove (19.2). Here I do it for the case in which X, Y are discrete independent RV's:

$$\begin{aligned} E[g_1(X)g_2(Y)] &= \sum_x \sum_y g_1(x)g_2(y)P^{X,Y}(x,y) \\ &= \sum_x \sum_y g_1(x)g_2(y)P^X(x)P^Y(y) \\ &= \sum_x g_1(x)P^X(x)[\sum_y g_2(y)P^Y(y)] \\ &= \sum_x g_1(x)P^X(x)E[g_2(Y)] \\ &= E[g_2(Y)]\sum_x g_1(x)P^X(x) \\ &= E[g_2(Y)]E[g_1(X)] \end{aligned}$$

Warning: Do not use formula (19.2) unless X, Y are independent!

19.2 Introduction to $r_{X,Y}$, $\sigma_{X,Y}$, $\rho_{X,Y}$

We begin to discuss three parameters of a random pair (X, Y) which are almost as famous as the three tenors. These three parameters, which will be important to us for much of the rest of the semester, are defined as follows:

- The *correlation* of X and Y is denoted $r_{X,Y}$. It is defined by

$$r_{X,Y} \stackrel{\Delta}{=} E[XY].$$

  In other words, the correlation of two RV's is just the expected value of their product.

- The *covariance* of X and Y is denoted $\sigma_{X,Y}$ or $Cov(X, Y)$. It is defined by

$$\sigma_{X,Y} = Cov(X, Y) \stackrel{\Delta}{=} E[(X - \mu_X)(Y - \mu_Y)].$$

  In other words, you center each RV X, Y about its mean by doing the operations $X - \mu_X$ and $Y - \mu_Y$. Then, you compute the correlation of $X - \mu_X$ and $Y - \mu_Y$: this is the covariance of X and Y.

- The *correlation coefficient* of X, Y is denoted $\rho_{X,Y}$. It is defined by

$$\rho_{X,Y} \stackrel{\Delta}{=} \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}.$$

  In other words, to compute the correlation coefficient, you simply divide the covariance by the product of the standard deviations of the two RV's. If in a given context, we understand what the two RV's X, Y are, it is customary to drop the subscripts from $\rho_{X,Y}$ and to refer to it simply as ρ.

In the rest of this section, we will present examples in which we compute the values of the parameters $r_{X,Y}$, $\sigma_{X,Y}$, $\rho_{X,Y}$. Before we do that, I list a few facts about these parameters.

Fact 1: The three parameters are symmetric in X, Y. That is,

$$\begin{aligned} r_{X,Y} &= r_{Y,X} \\ Cov(X, Y) &= Cov(Y, X) \\ \rho_{X,Y} &= \rho_{Y,X} \end{aligned}$$

Fact 2: Variance is a special case of covariance. That is,

$$Var(X) = Cov(X, X).$$

Fact 3: Covariance may be computed from correlation via the following formula:

$$Cov(X,Y) = E[XY] - \mu_X \mu_Y \tag{19.3}$$

Fact 4: If X, Y are independent, then the three parameters are immediately computable as:

$$\begin{aligned} r_{X,Y} &= E[XY] = E[X]E[Y] \\ Cov(X,Y) &= \sigma_{X,Y} = 0 \\ \rho_{X,Y} &= 0 \end{aligned}$$

Facts 1 and 2 are trivial consequences of the definitions of the three parameters. Fact 4 follows from formula (19.2). Fact 3 can be proved as follows:

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

In equation (19.3), notice what happens when we replace Y by X:

$$Cov(X,X) = Var(X) = E[X^2] - \mu_X^2.$$

This is the "second moment formula" for computing variance that we derived back in the Chapter 2-3 Notes. It is interesting to see that this earlier formula follows from the present formula (19.3).

Example 19.1. Let X, Y be discrete RV's each taking the values $0, 1, 2$. The following is the joint PMF array:

$$\begin{array}{c} \\ X=0 \\ X=1 \\ X=2 \end{array} \begin{array}{c} \begin{array}{ccc} Y=0 & Y=1 & Y=2 \end{array} \\ \left(\begin{array}{ccc} 0.1 & 0 & 0.2 \\ .05 & 0.2 & 0.3 \\ 0.1 & 0 & .05 \end{array} \right) \end{array}$$

We compute the values of the three parameters $r_{X,Y}$, $\sigma_{X,Y}$, $\rho_{X,Y}$. First, notice from the joint PMF array that products of the form xy are equal to zero all along the first row (where $x = 0$) and all along the first column (where $y = 0$). That leaves just 4 nonzero products in computing $E[XY]$:

$$E[XY] = (1 \cdot 1)0.2 + (1 \cdot 2)0.3 + (2 \cdot 1)0 + (2 \cdot 2).05 = 1$$

So, for our two random variables X, Y, the correlation is

$$r_{X,Y} = E[XY] = 1.$$

Let us now compute the covariance. We use formula (19.3). We need to compute μ_X, μ_Y. Taking the row sums of the joint PMF array, we see that

$$[P^X(0) \;\; P^X(1) \;\; P^X(2)] = [0.3 \;\; .55 \;\; .15]$$

from which it follows that

$$\mu_X = 1(.55) + 2(.15) = .85.$$

Taking the column sums of the joint PMF array, we obtain

$$[P^Y(0) \;\; P^Y(1) \;\; P^Y(2)] = [.25 \;\; 0.2 \;\; .55],$$

from which we obtain

$$\mu_Y = 1(0.2) + 2(.55) = 1.3.$$

We conclude that

$$\sigma_{X,Y} = r_{X,Y} - (.85)(1.3) = -0.105.$$

Finally, in order to compute the correlation coefficient $\rho_{X,Y}$ from the covariance $\sigma_{X,Y}$, we need the variance of each RV X, Y:

$$\begin{aligned} \sigma_X^2 &= E[X^2] - \mu_X^2 = 1(.55) + 4(.15) - (.85)^2 = 0.4275 \\ \sigma_Y^2 &= E[Y^2] - \mu_Y^2 = 1(0.2) + 4(.55) - (1.3)^2 = 0.7100 \end{aligned}$$

Therefore,

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} = \frac{-0.105}{\sqrt{0.4275}\sqrt{0.7100}} = -0.1906.$$

Before proceeding with further examples, now is a good time to list three important properties of $\rho_{X,Y}$:

Property 1 of $\rho_{X,Y}$: The correlation coefficient is always between -1 and 1:

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

Property 2 of $\rho_{X,Y}$: If the RV's X, Y are independent, then the correlation coefficient $\rho_{X,Y}$ is equal to 0.

Property 3 of $\rho_{X,Y}$: If the correlation coefficient $\rho_{X,Y}$ is equal to ± 1, then there is a straight line relationship between X and Y. Specifically, if $\rho_{X,Y} = +1$, there is a unique straight line $y = Ax + B$ in the xy-plane, with positive slope A, such that the random pair (X, Y) will always fall on this straight line. On the other hand, if $\rho_{X,Y} = -1$, there is a unique straight line $y = Ax + B$ in the xy-plane, with negative slope A, such that the random pair (X, Y) will always fall on this straight line.

Property 2 is just a restatement of part of Fact 4, which we have restated for emphasis. Properties 1 and 3 are a bit deeper than Property 2: we will prove Properties 1 and 3 later.

Example 19.2. Let random pair (X, Y) be chosen uniformly from the square region

$$R = \{(x, y) : 0 \leq x \leq 1; \;\; 0 \leq y \leq 1\}.$$

We compute $r_{X,Y}$, $\sigma_{X,Y}$, and $\rho_{X,Y}$. The region R is a Cartesian product set, and we have the factorization

$$f_{X,Y}(x, y) = \Phi(x)\Psi(y), \quad (x, y) \in R,$$

where

$$\Phi(x) = 1, \quad \Psi(y) = 1.$$

Therefore, X, Y are independent Uniform(0,1) RV's. By Fact 4,

$$r_{X,Y} = E[XY] = E[X]E[Y] = (1/2)(1/2) = 1/4.$$

Also from Fact 4, we can immediately say that both the covariance $\sigma_{X,Y}$ and the correlation coefficient $\rho_{X,Y}$ are equal to 0.

Example 19.3. Let random pair (X, Y) be chosen uniformly from the circular region

$$R = \{(x, y) : x^2 + y^2 \leq 1\}.$$

We compute $r_{X,Y}$, $\sigma_{X,Y}$, and $\rho_{X,Y}$. The area of R is π. Therefore, $f_{X,Y}(x, y)$ is equal to $1/\pi$ inside R, and so

$$E[XY] = \iint_R xy \left(\frac{1}{\pi}\right) dxdy.$$

I now switch the double integral from rectangular coordinates x, y to polar coordinates r, θ:

$$\begin{aligned} E[XY] &= (1/\pi) \int_0^{2\pi} \int_0^1 (r\cos\theta)(r\sin\theta) r dr d\theta \\ &= (1/\pi) \left(\int_0^1 r^3 dr \right) \left(\int_0^{2\pi} \sin\theta\cos\theta d\theta \right) \end{aligned}$$

The reader can easily show that the above integral with respect to $d\theta$ is equal to zero. We have shown that

$$r_{X,Y} = E[XY] = 0.$$

By the "center of gravity rule" (covered in the next section), the point $(E[X], E[Y])$ is the center of gravity of the region R. By the circular symmetry of the region R, the center of gravity of R is the origin $(0, 0)$ in the xy-plane. Therefore,

$$\begin{aligned} (E[X], E[Y]) &= (0, 0) \\ E[X] &= 0 \\ E[Y] &= 0 \end{aligned}$$

We conclude that

$$\begin{aligned} Cov(X,Y) &= r_{X,Y} - \mu_X \mu_Y = 0 \\ \rho_{X,Y} &= \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = 0 \end{aligned}$$

Notice that the RV's X, Y are statistically dependent because the region R is not a Cartesian product set. Yet $\rho_{X,Y} = 0$ even though the variables are dependent.

Remark. Earlier, we learned that if X, Y are independent, then $\rho_{X,Y} = 0$. Example 19.3 is important because it shows us that the converse of this statement is not true: if $\rho_{X,Y} = 0$, *we cannot conclude that the RV's X, Y are independent.* It may happen that $\rho_{X,Y} = 0$ simply due to some underlying symmetry in the joint distribution of (X, Y), without the RV's being independent.

Exercise. Here is another example you can work out for yourself in which the correlation coefficient will be equal to zero for reasons of symmetry. Let X, Y be the discrete RV's in which the random pair (X, Y) is equidistributed over the set of four points

$$S = \{(-1, 0), (1, 0), (0, 1), (0, -1)\}.$$

Show that X, Y are dependent but that $\rho_{X,Y} = 0$.

Example 19.4. Suppose X, Y are jointly continuously distributed with joint density

$$f_{X,Y}(x, y) = \left(\frac{1}{2\pi\sqrt{1-\rho^2}} \right) \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)} \right), \tag{19.4}$$

where ρ is a fixed parameter strictly between 0 and 1. The following can be shown (I do this later as part of the joint Gaussian distribution coverage):

- For each ρ satisfying $-1 < \rho < 1$, (19.4) defines a bonafide prob density function. (That is, the double integral is 1.)
- The two marginal distributions for X and Y are each Gaussian(0,1) (standard Gaussian).
- All three parameters $r_{X,Y}$, $\sigma_{X,Y}$, and $\rho_{X,Y}$ are equal to ρ.

The reader should go to page 192 of Yates-Goodman and view the plots of the surface

$$z = f_{X,Y}(x, y)$$

for various ρ; the reader can also do these plots via Matlab in the last experiment of Recitation 6. You will see that the surface plots for ρ very close to 1 or -1 are roughly concentrated above a straight line in the xy-plane. This bears out Property 3 of the correlation coefficient, which we will be proving later on. Another interesting case is $\rho = 0$: in this case, the xy term in $f_{X,Y}(x, y)$ drops

out, and then one can see by inspection of (19.4) that the joint density factors into a function of x times a function of y. It follows that we have independence of X and Y when $\rho = 0$. (Earlier, we saw that $\rho = 0$ does not necessarily imply independence, but for the density of form (19.4) this will be true.) What is nice about this example is that the marginal densities of X and Y stay fixed as you change the correlation coefficient ρ. The moral we draw from this fact is that knowing the marginal distributions of two RV's X, Y tells us nothing about how they may be correlated—the correlation coefficient could be anything between 0 and 1.

19.2.1 Center of Gravity Rule

Suppose the random pair (X, Y) is selected uniformly from a region R in the xy-plane which has finite and positive area. Then it is easy to derive the following formulas:

$$\begin{aligned} E[X] &= \frac{\iint_R x dx dy}{area(R)} \\ E[Y] &= \frac{\iint_R y dx dy}{area(R)} \end{aligned}$$

If the reader goes to any calculus book, it will be seen that the right hand sides of these equations are, respectively, the x and y coordinates of the *center of gravity* of the region R. The center of gravity is also called the *centroid*. We have proven the following "center of gravity" rule:

- *If (X, Y) is uniform over the region R, then the point $(E[X], E[Y])$ is the centroid of R.*

See Problems 6.1 and 6.2 of the Chapter 4-5 Solved Problems. These Problems are illustrations of the use of the Center of Gravity Rule.

Lecture 20

Chapters 4-5 Part 6

20.1 Positively/Negatively Correlated

X, Y are *positively correlated* if

$$0 < \rho_{X,Y} < 1. \tag{20.1}$$

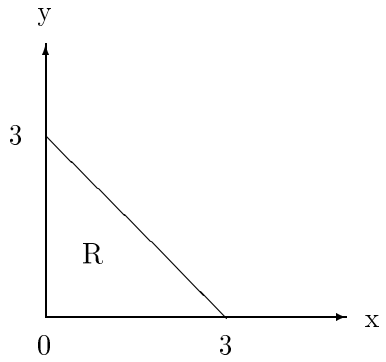
They are *negatively correlated* if

$$-1 < \rho_{X,Y} < 0. \tag{20.2}$$

As the value of X increases, suppose you expect that (on average) the value of Y will also increase; then you'd guess that X, Y are positively correlated. On the other hand, as the value of X increases, suppose you expect that (on average) the value of Y will decrease; then you'd guess that X, Y are negatively correlated.

Example 20.1. In the "ice cream example" we considered earlier, if the number of ice cream cones X that Bill eats increases, then we'd expect (on average) that he'd have to run more miles Y. So, we'd guess that X and Y are positively correlated. In a later lecture, I'll use a neat technique called the "law of iterated expectation" to actually compute $\rho_{X,Y}$ in this case and show that (20.1) holds.

Example 20.2. Suppose (X, Y) is chosen uniformly from the region R below.



As X increases, the point (X, Y) in R is forced into the lower right hand corner of R, making Y decrease. So, we'd expect that X and Y are negatively correlated. Later, we will bear out this intuition by actually computing $\rho_{X,Y}$ in this case and showing that (20.2) holds.

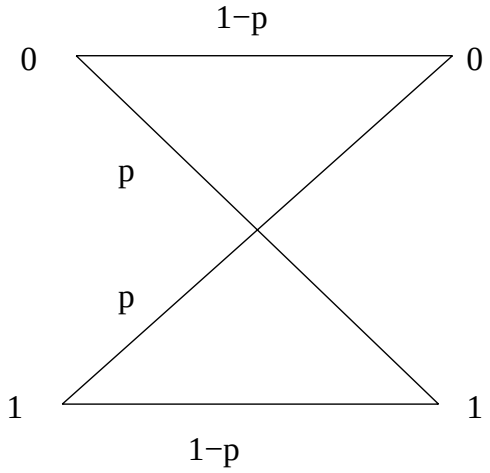
Exercise. Go to page 176 of Yates-Goodman. You will see there several other examples of X, Y where in each case you can intuit whether the variables are positively or negatively correlated.

20.2 Correlation Properties of BSC

"BSC" stands for "binary symmetric channel" model:

$$X \to \boxed{\text{BSC}} \to Y$$

The so-called line diagram of the BSC is the following:



The parameter p (the "crossover probability") is the probability that the BSC makes an error; it can be anything in the range $0 \leq p \leq 1$.

Let us take the input X to be binary equiprobable (i.e., $P(X = 0) = P(X = 1) = 1/2$). From the line diagram, the channel matrix is:

$$[p(y|x)] = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$

We want to compute $\rho = \rho_{X,Y}$ as a function of the cross-over probability. Then we shall draw some conclusions.

We know (from the first step of the Bayes Method) that the joint PMF matrix $[p_{X,Y}(x,y)]$ is obtained by multiplying the first row of the channel matrix by $P(X=0)$ and the second row by $P(X=1)$. This gives us:

$$[p_{X,Y}(x,y)] = \begin{bmatrix} (1-p)/2 & p/2 \\ p/2 & (1-p)/2 \end{bmatrix}$$

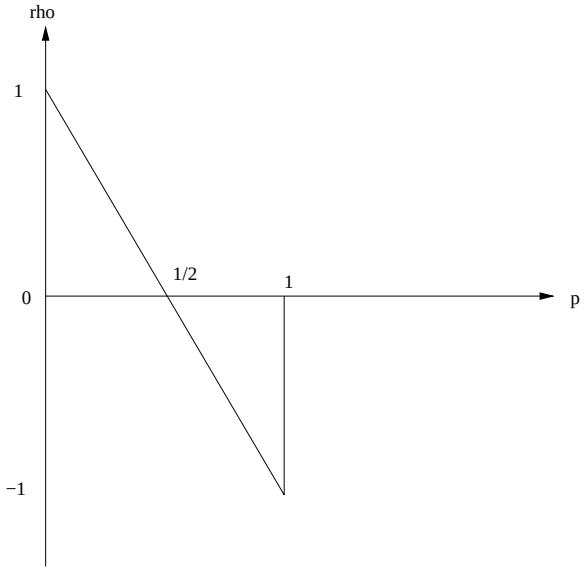
Only one xy product is nonzero, namely, when $x=1$ and $y=1$, which corresponds to the lower right hand corner of the matrix $[p_{X,Y}(x,y)]$. Therefore,

$$r_{X,Y} = (1)(1)(1-p)/2$$

We now compute covariance and correlation coefficient. The reader can easily work out that $\mu_X = \mu_Y = 1/2$ and $\sigma_X = \sigma_Y = 1/2$. Therefore,

$$\begin{aligned} \sigma_{X,Y} &= r_{X,Y} - \mu_X \mu_Y = (1-p)/2 - 1/4 \\ \rho_{X,Y} &= \sigma_{X,Y}/\sigma_X\sigma_Y = \frac{(1-p)/2 - 1/4}{1/4} = 1-2p \end{aligned}$$

The plot of ρ versus p is given by:



We will always have $-1 \leq \rho \leq 1$. Three cases of particular interest are $\rho = 0$, $\rho = 1$, $\rho = -1$. From the equation $\rho = 1 - 2p$, the reader can see that

$$\begin{aligned} \rho = 0 &\Leftrightarrow p = 1/2 \\ \rho = 1 &\Leftrightarrow p = 0 \\ \rho = -1 &\Leftrightarrow p = 1 \end{aligned}$$

- When $\rho = 0$, the channel matrix is

$$\begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$$

  from which it follows that X and Y are independent. (Any time you have a channel matrix in which the rows are all the same, then the channel input X and the channel output Y are automatically statistically independent.) This is particularly interesting because of examples showing that $\rho_{X,Y} = 0$ can occur for some dependent random variables X, Y. In the case of input and output to a BSC, this behavior can't occur — independence of input and output is equivalent to vanishing of the correlation coefficient.

- When $\rho = 1$, the channel matrix is

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

  This means that the channel perfectly transmits any input, so that $X = Y$ always holds in this case. Notice that the values of (X, Y) for this $\rho = 1$ case fall on the straight line $x = y$ of positive slope. We shall show later on that for *any* random variables X, Y, if $\rho_{X,Y} = 1$, then the values of (X, Y) must all lie on a straight line in the (x, y)-plane of positive slope (this will be the line $x = y$ if the means are zero and the standard deviations are the same).

- When $\rho = -1$, then the channel matrix is

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

  and so the channel output is the exact opposite of the channel input. In other words, $X+Y = 1$ in this case, meaning that the values of (X, Y) lie on the straight line $x + y = 1$ of negative slope. We shall show later on that for *any* random variables X, Y, if $\rho_{X,Y} = -1$, then the values of (X, Y) must all lie on a straight line in the (x, y)-plane of negative slope.

20.3 Bilinearity Properties of Covariance

The bilinearity properties of covariance, which follow, allow you to linearly expand either argument of a covariance while holding the other argument fixed.

Property 1: $Cov(X, Y + Z) = Cov(X, Y) + Cov(X, Z)$

Property 2: $Cov(X, CY) = C\,[Cov(X, Y)]$ if C is a constant.

Property 3: $Cov(X, C) = 0$ if C is constant.

Property 4: $Cov(X, C_1Y + C_2Z + C_3) = C_1[Cov(X, Y)] + C_2[Cov(X, Z)]$ if C_1, C_2, C_3 are constants.

Actually, Properties 1-3 are special cases of Property 4; we have stated Properties 1-3 separately for emphasis. We have fixed the first argument in these properties, and linearly expanded the second argument. You can also do the reverse (because covariance is symmetric).

The properties are easy to prove. For example, Property 1 can be proved by taking the expected value of both sides of the identity

$$(X - \mu_X)(Y + Z - \mu_Y - \mu_Z) = (X - \mu_X)(Y - \mu_Y) + (X - \mu_X)(Z - \mu_Z)$$

The properties allow one to expand the variance of a sum:

$$\begin{aligned} \mathrm{V}ar[X + Y] &= \mathrm{C}ov(X + Y, X + Y) \\ &= \mathrm{C}ov(X, X) + \mathrm{C}ov(X, Y) + \mathrm{C}ov(Y, X) + \mathrm{C}ov(Y, Y) \end{aligned}$$

This gives us the important formula

$$\boxed{\mathrm{V}ar[X + Y] = \sigma_X^2 + 2\sigma_{X,Y} + \sigma_Y^2}$$

The preceding formula extends to sums of more than two terms:

$$\boxed{\mathrm{V}ar[X_1 + X_2 + X_3 + \ldots + X_n] = \left(\sum_{i=1}^n \mathrm{V}ar[X_i]\right) + 2\left(\sum_{i<j} \mathrm{C}ov(X_i, X_j)\right)}$$

From the preceding formulae, we see that the variance of a sum is, in general, *not equal* to the sum of the variances of the separate terms. However, if the summands are uncorrelated (in particular, if the summands are independent), the variance of a sum will be equal to the sum of the variances of the separate terms.

Example 20.3. Let

$$\begin{aligned} \rho_{X,Y} &= -1/2 \\ \sigma_X &= 2 \\ \sigma_Y &= 3 \end{aligned}$$

Then

$$\begin{aligned} Var(3X - Y + 5) &= Var(3X - Y) \\ &= Cov(3X - Y, 3X - Y) \\ &= 9 * Cov(X, X) - 6 * Cov(X, Y) + Cov(Y, Y) \\ &= 9\sigma_X^2 - 6\sigma_{X,Y} + \sigma_Y^2 \\ &= 9\sigma_X^2 - 6\sigma_X\sigma_Y\rho_{X,Y} + \sigma_Y^2 \\ &= 9 * 4 - 6 * 2 * 3 * (-1/2) + 9 = 63 \end{aligned}$$

Also,

$$\begin{aligned} Cov(3X - Y + 4, 5X + Y - 7) &= Cov(3X - Y, 5X + Y) \\ &= 15 * Cov(X, X) - 2 * Cov(X, Y) - Cov(Y, Y) \\ &= 15\sigma_X^2 - 2\sigma_X\sigma_Y\rho_{X,Y} - \sigma_Y^2 \\ &= 15 * 4 - 2 * 2 * 3 * (-1/2) - 9 = 57 \end{aligned}$$

Remark. Later, we will develop an easier way to compute covariances of linear combinations of RV's using matrices.

Lecture 21

Chapters 4-5 Part 7

21.1 Application: Correlation Receiver

Earlier, in Section 17.2, we discussed how to design the MAP detector (MAP receiver) for obtaining an estimate of channel input X based upon channel output Y. In this section, we discuss how to design a different type of receiver called the *correlation receiver*.

Let RV X be the input to a channel and RV Y be the resulting output. The correlation receiver can then be applied to Y in order to construct an estimate $\hat{X}$ of X as shown in the following block diagram:

$$X \rightarrow \boxed{\text{channel}} \rightarrow Y \rightarrow \boxed{\begin{array}{c}\text{correlation}\\ \text{receiver}\end{array}} \rightarrow \hat{X} = CY$$

C is a constant that is computed in order to minimize the "mean-square estimation error"

$$E[(X - \hat{X})^2].$$

(This minimization criterion is what distinguishes the correlation receiver from the MAP receiver: the MAP receiver minimizes the estimation error probability $P[X \neq \hat{X}]$ instead of minimizing $E[(X - \hat{X})^2]$. The estimation error criterion that one uses determines the type of receiver you get when you minimize with respect to this criterion.)

Let us find the minimizing choice of C in the correlation receiver output $\hat{X} = CY$. Expanding $E[(X - \hat{X})^2]$,

$$\begin{aligned} E[(X - \hat{X})^2] &= E[(X - CY)^2] \\ &= E[X^2 - 2CXY + C^2Y^2] \\ &= E[X^2] - 2CE[XY] + C^2E[Y^2] \end{aligned}$$

The first derivative of the estimation error with respect to C is:

$$-2E[XY] + 2CE[Y^2].$$

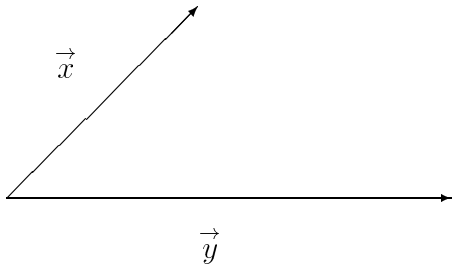
Setting this equal to 0 and solving for C, we see that

$$C = \frac{E[XY]}{E[Y^2]} = \frac{r_{X,Y}}{E[Y^2]}. \tag{21.1}$$

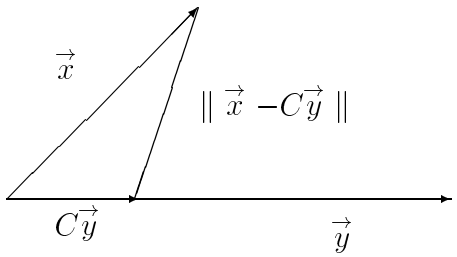
Because of the presence of the correlation $r_{X,Y}$ in formula (21.1), the reader can now see why our receiver is called the "correlation receiver".

21.1.1 Geometric Interpretation

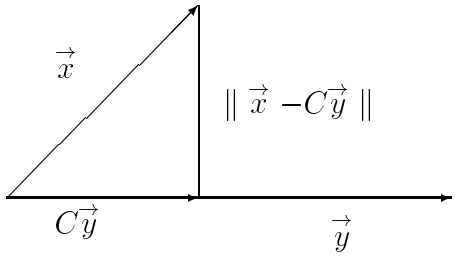
Let $\vec{x}$ and $\vec{y}$ be geometric vectors (such as you encountered in your freshman-sophomore physics and calculus courses). Draw them in a diagram as two sides of a triangle:



Any vector $C\vec{y}$, where C is a scalar, would point in the same direction as $\vec{y}$ if $C > 0$ and would point in the opposite direction as $\vec{y}$ if $C < 0$. The length of the vector $\vec{x} - C\vec{y}$ is denoted $\| \vec{x} - C\vec{y} \|$; it can be interpreted as the length of the line in our diagram which connects the end of vector $\vec{x}$ to the end of vector $C\vec{y}$:



It is geometrically clear that as C varies, the length $\| \vec{x} - C\vec{y} \|$ becomes a minimum when the end of vector $C\vec{y}$ lies at the base of a perpendicular dropped from the end of vector $\vec{x}$:



This unique position of the vector $C\vec{y}$ is called the *projection of vector* $\vec{x}$ *on vector* $\vec{y}$. You learned in physics/calculus that this projection is expressible as

$$\left(\vec{x} \cdot \vec{u}\right) \vec{u},$$

where $\vec{u}$ is the unit vector in the same direction as $\vec{y}$:

$$\vec{u} = \frac{\vec{y}}{\| \vec{y} \|}.$$

Summarizing, we now have the following useful fact:

Useful Fact 1: The multiple $C\vec{y}$ of vector $\vec{y}$ which makes $\| \vec{x} - C\vec{y} \|^2$ a minimum is

$$\begin{aligned} C\vec{y} &= \left(\vec{x} \cdot \frac{\vec{y}}{\| \vec{y} \|} \right) \frac{\vec{y}}{\| \vec{y} \|} \\ &= \left(\frac{\vec{x} \cdot \vec{y}}{\| \vec{y} \|^2} \right) y \end{aligned}$$

Let us now make the following correspondences between the "world of random variables" X, Y and the "world of geometric vectors" $\vec{x}, \vec{y}$:

$$\begin{aligned} X &\leftrightarrow \vec{x} \\ Y &\leftrightarrow \vec{y} \\ E[XY] &\leftrightarrow \vec{x} \cdot \vec{y} \\ \sqrt{E[Y^2]} &\leftrightarrow \| \vec{y} \| \end{aligned}$$

In other words, we want to think of the correlation $E[XY]$ of RV's X, Y in the same way we think of the dot product $\vec{x} \cdot \vec{y}$ of geometric vectors $\vec{x}, \vec{y}$. In particular, this means that the second moment $E[Y^2]$, which is the correlation of Y with itself, will be thought of in the same way as

$$\vec{y} \cdot \vec{y} = \| \vec{y} \|^2,$$

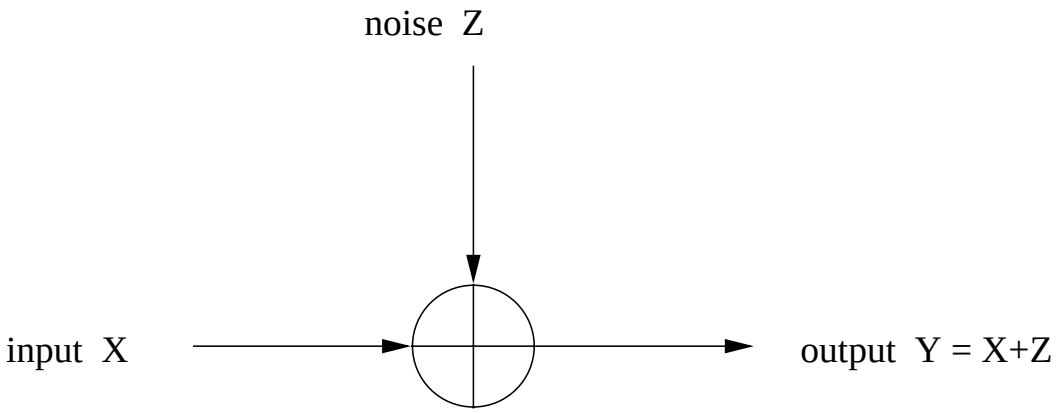
the square of the length of geometric vector $\vec{y}$. Using the given correspondences, we can now re-interpret our earlier Useful Fact 1 as the following useful fact:

Useful Fact 2: The multiple CY of random variable Y which makes $E[(X-CY)^2]$ a minimum is

$$CY = \left(\frac{E[XY]}{E[Y^2]}\right) Y \tag{21.2}$$

Equation (21.2) gives us the output of the correlation receiver, which we have interpreted as the *projection of X on Y*.

Example 21.1. The channel we will be using is an "additive noise channel":



The channel input random variable X is assumed to have mean 0 and variance σ_X^2. The channel noise random variable Z is independent of the input X, and is assumed to have mean 0 and variance σ_Z^2. Let us determine the form of the correlation receiver output

$$\hat{X} = CY.$$

Using formula (21.1),

$$\begin{aligned} C &= \frac{E[XY]}{E[Y^2]} \\ &= \frac{E[X(X+Z)]}{E[(X+Z)^2]} \\ &= \frac{E[X^2]+E[X]E[Z]}{E[X^2]+2E[X]E[Z]+E[Z^2]} \\ &= \frac{\sigma_X^2}{\sigma_X^2+\sigma_Z^2} \end{aligned}$$

That is, the correlation receiver output may be expressed as

$$\hat{X} = \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_Z^2} \right) Y.$$

21.2 More About ρ

In Section 19.2, I gave some properties of the correlation coefficient ρ without proof (Properties 1 and 3). In this section, I will prove these properties. My proofs will use the following new property of ρ:

Property 4 of ρ: The correlation coefficient ρ remains unchanged (except possibly for a change in sign) under scaling and translation of its two arguments. More precisely, let X, Y be two RV's and suppose we scale and translate them to obtain two new RV's U, V as follows:

$$\begin{aligned} U &= AX + B \\ V &= CY + D, \end{aligned}$$

where A, B, C, D are real constants with $A \neq 0$ and $C \neq 0$. Then,

$$\rho_{U,V} = \begin{cases} \rho_{X,Y}, & AC > 0 \\ -\rho_{X,Y}, & AC < 0 \end{cases}$$

Proof of Property 4. Use the bilinearity properties given in Section 20.3 to establish the following facts:

$$\begin{aligned} Cov(AX + B, CY + D) &= Cov(AX, CY) = (AC)Cov(X, Y) \\ Var(AX + B) &= A^2 Var(X) \\ Var(CY + D) &= C^2 Var(Y) \end{aligned}$$

It follows that

$$\rho_{U,V} = \left(\frac{AC}{|AC|} \right) \rho_{X,Y},$$

from which Property 4 is apparent.

We now state and prove the following theorem, which establishes Property 1 of Section 19.2 and an improved version of Property 3 of Section 19.2.

Theorem: The correlation coefficient always satisfies the inequality

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

Furthermore, if $\rho_{X,Y} = 1$, then (X, Y) can be considered as always falling on the straight line

$$\frac{x - \mu_X}{\sigma_X} = \frac{y - \mu_Y}{\sigma_Y}$$

in the xy-plane, whereas if $\rho_{X,Y} = -1$, then (X, Y) can be considered as always falling on the straight line

$$\frac{x - \mu_X}{\sigma_X} = -\frac{y - \mu_Y}{\sigma_Y}.$$

Proof of Theorem. Suppose we can prove that the following identity is true:

$$E\left[\left\{\left(\frac{X - \mu_X}{\sigma_X}\right) - \rho\left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right\}^2\right] = 1 - \rho^2,$$

where $\rho = \rho_{X,Y}$. Then, the Theorem is obviously true. To prove this identity, we make the change of variable

$$\begin{aligned} U &= \frac{X - \mu_X}{\sigma_X} \\ V &= \frac{Y - \mu_Y}{\sigma_Y} \end{aligned}$$

By Property 4, $\rho_{U,V} = \rho_{X,Y} = \rho$, and our identity reduces to the following much more simple form:

$$E[(U - \rho V)^2] = 1 - \rho^2.$$

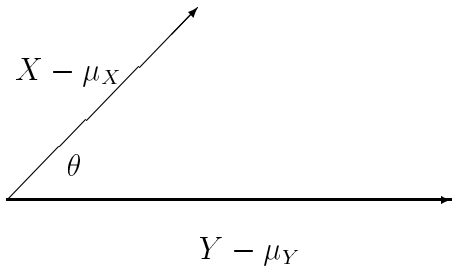
It is almost a trivial matter to prove this fact (using the fact that U, V each have mean 0 and variance 1):

$$\begin{aligned} E[(U - \rho V)^2] &= E[U^2] - 2\rho E[UV] + \rho^2 E[V^2] \\ &= 1 - 2\rho^2 + \rho^2 = 1 - \rho^2 \end{aligned}$$

21.2.1 Geometric Interpretation of ρ

In Section 21.1.1, we introduced geometric notions which helped us to visualize what the correlation receiver is doing. In this section, we will use similar notions to give a geometric interpretation of the correlation coefficient ρ.

Given RV's X, Y, let us think of $X - \mu_X$ and $Y - \mu_Y$ as if they are geometric vectors in the following diagram:



The dot product of two geometric vectors is the product of the lengths of the vectors times the cosine of the angle between them:

$$(X - \mu_X) \cdot (Y - \mu_Y) = \|X - \mu_X\| \|Y - \mu_Y\| \cos\theta. \tag{21.3}$$

As we pointed out in Section 21.1.1, we want to think of dot product as expected value of the product of the RV's, and we want to think of length as the square root of the second moment of the RV. Under these correspondences, we interpret equation (21.3) as:

$$E[(X - \mu_X)(Y - \mu_Y)] = \sqrt{E[(X - \mu_X)^2]}\sqrt{E[(Y - \mu_Y)^2]}\cos\theta,$$

which is the same thing as saying that

$$\sigma_{X,Y} = \sigma_X \sigma_Y \cos\theta.$$

By definition of $\rho_{X,Y}$, this tells us that

$$\rho_{X,Y} = \cos\theta. \tag{21.4}$$

That is, we can interpret the correlation coefficient $\rho_{X,Y}$ as the cosine of the angle between $X - \mu_X$ and $Y - \mu_Y$. Since $\cos\theta$ will always be between -1 and 1, this interpretation of $\rho_{X,Y}$ helps us to appreciate why the inequality

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

should be true.

There are applications in which the interpretation (21.4) of $\rho_{X,Y}$ can be useful. One of these applications (to be discussed more in a later lecture) is *decorrelation*. In decorrelation, you take linear combinations of X and Y in order to obtain RV's U, V which are uncorrelated in the sense that $\rho_{U,V} = 0$. Geometrically, this means we seek those linear combinations that will make the angle between $U - \mu_U$ and $V - \mu_V$ equal to 90 degrees. In other words, before decorrelation, the angle between $X - \mu_X$ and $Y - \mu_Y$ was not 90 degrees; decorrelation changes this angle to 90 degrees.

21.3 Conditioning X or Y on an (X, Y) event

Let X, Y be RV's. Let B be a subset of the xy-plane such that

$$P((X, Y) \in B) > 0.$$

Then it is no problem to condition X (or Y) on event B, because B is an event of positive probability.

For example, suppose we want to compute the conditional probability

$$P(a \leq X \leq b|B),$$

which is an abbreviation for

$$P(\{a \leq X \leq b\}|\{(X, Y) \in B\}),$$

the conditional probability of event $\{a \leq X \leq b\}$ given event $\{(X, Y) \in B\}$. Chapter 1 tells us that we may compute this probability as the following ratio:

$$P(a \leq X \leq b|B) = \frac{P(\{a \leq X \leq b\} \cap \{(X, Y) \in B\})}{P((X, Y) \in B)}.$$

Both the probability in the numerator and the probability in the denominator can easily be computed if we know how X, Y are jointly distributed, which gives us the following formulas:

- If X, Y are each discrete, then

$$P(a \leq X \leq b|B) = \frac{\sum_{\{(x,y): a \leq x \leq b, \ (x,y) \in B\}} P^{X,Y}(x, y)}{\sum_{(x,y) \in B} P^{X,Y}(x, y)}. \tag{21.5}$$

- If X, Y are jointly continuous, then

$$P((a \leq X \leq b|B) = \frac{\iint_{\{(x,y): a \leq x \leq b, \ (x,y) \in B\}} f_{X,Y}(x, y)dxdy}{\iint_B f_{X,Y}(x, y)dxdy}. \tag{21.6}$$

It is equally easy to compute

$$E(X|B),$$

the conditional expected value of X given that (X, Y) falls in B. This is done as follows:

- If X, Y are each discrete, then

$$E(X|B) = \frac{\sum\limits_{(x,y)\in B} xP^{X,Y}(x,y)}{\sum\limits_{(x,y)\in B} P^{X,Y}(x,y)}. \tag{21.7}$$

- If X, Y are jointly continuous, then

$$E(X|B) = \frac{\iint_B xf_{X,Y}(x,y)dxdy}{\iint_B f_{X,Y}(x,y)dxdy}. \tag{21.8}$$

The reader can also easily see how to compute expressions like

$$P(a \leq Y \leq b|B),$$

and

$$E(Y|B).$$

You simply reverse the roles of X and Y in the four formulas (21.5)-(21.8).

Example 21.2. Let X, Y be independent RV's, each exponentially distributed with $E(X) = 1$ and $E(Y) = 2$. Then:

$$\begin{aligned}
P(X \leq 1|X+Y \leq 2) &= \frac{\int_0^1 \int_0^{2-x} (0.5)\exp(-x-y/2)dydx}{\int_0^2 \int_0^{2-x} (0.5)\exp(-x-y/2)dydx} = 0.8575 \\
E(X|X+Y \leq 2) &= \frac{\int_0^2 \int_0^{2-x} (0.5)x\exp(-x-y/2)dydx}{\int_0^2 \int_0^{2-x} (0.5)\exp(-x-y/2)dydx} = 0.5134 \\
P(Y \leq 1|X+Y \leq 2) &= \frac{\int_0^1 \int_0^{2-y} (0.5)\exp(-x-y/2)dxdy}{\int_0^2 \int_0^{2-y} (0.5)\exp(-x-y/2)dxdy} = 0.7650 \\
E(Y|X+Y \leq 2) &= \frac{\int_0^2 \int_0^{2-y} (0.5)y\exp(-x-y/2)dxdy}{\int_0^2 \int_0^{2-y} (0.5)\exp(-x-y/2)dxdy} = 0.6452
\end{aligned}$$

Remark. Notice that in all four formulas (21.5)-(21.8), the denominator on the right hand side is simply $P((X,Y) \in B)$. The fact that this probability is positive makes these calculations straightforward. If this probability is equal to zero, then another approach would have to be found (because then all four formulas (21.5)-(21.8) would yield an indeterminate form 0/0). For example, if Y is continuous, the event $\{Y = y\}$ is an event of probability zero for any fixed real number y, yet we will need to make sense of expressions like

$$P(a \leq X \leq b|Y = y)$$

and

$$E(X|Y = y).$$

This will be done during our next lecture.

21.4 Conditioning one RV on another: Discrete Case

Let X, Y be discrete RV's. In this section, we explain how you find the conditional PMF of X given any value of Y and how you find the conditional PMF of Y given any value of X.

The expression $P^{X|Y}(x|y)$ denotes the conditional PMF of X given $Y = y$. In the expression $P^{X|Y}(x|y)$, it is to be understood that y is fixed and x varies through the values of X. This conditional PMF is defined by:

$$P^{X|Y}(x|y) \triangleq P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P^{X,Y}(x,y)}{P^{Y}(y)}.$$

Keep in mind that even though we are calling $P^{X|Y}(x|y)$ a *conditional* PMF, it is a bonafide PMF in its own right, that is,

$$\sum_x P^{X|Y}(x|y) = 1.$$

We can reverse the roles of X and Y in all of the preceding. Thus, the expression $P^{Y|X}(y|x)$ denotes the conditional PMF of Y given $X = x$. In the expression $P^{Y|X}(y|x)$, it is to be understood that x is fixed and y varies through the values of Y. This conditional PMF is defined by:

$$P^{Y|X}(y|x) \triangleq P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{P^{X,Y}(x,y)}{P^{X}(x)}.$$

Keep in mind that even though we are calling $P^{Y|X}(y|x)$ a *conditional* PMF, it is a bonafide PMF in its own right, that is,

$$\sum_y P^{Y|X}(y|x) = 1.$$

One particularly attractive subcase is when the discrete RV's X and Y take just finitely many values. Then we know that we can put the joint PMF values $P^{X,Y}(x,y)$ in an array, and from this array we can obtain any possible conditional PMF as follows:

- For a fixed value y of Y, you find the conditional PMF of X given $Y = y$ by finding the column of the joint PMF array with the heading "$Y = y$", and then dividing through this column by the column sum.
- For a fixed value x of X, you find the conditional PMF of Y given $X = x$ by finding the row of the joint PMF array with the heading "$X = x$", and then dividing through this row by the row sum.

Example 21.3. (This is Problem 7.1 of the Chapter 4-5 Solved Problems.) Let discrete X, Y have the following joint PMF array:

$$\begin{array}{c} \\ X=1 \\ X=2 \\ X=3 \\ X=4 \end{array} \begin{array}{c} \begin{array}{cccc} Y=1 & Y=2 & Y=3 & Y=4 \end{array} \\ \left(\begin{array}{cccc} .10 & .05 & .05 & .05 \\ .05 & .10 & .05 & .05 \\ .05 & .05 & .10 & .05 \\ .05 & .05 & .05 & .10 \end{array} \right) \end{array}$$

(a) Compute $P(2 \leq X \leq 3 | Y = 2)$ and $E(X|Y = 2)$.

Solution. Divide the $Y = 2$ column of the joint PMF array by the column sum 0.25. The conditional PMF of X given $Y = 2$ is then

$$P^{X|Y}(x|2) = \begin{cases} 1/5, & x = 1 \\ 2/5, & x = 2 \\ 1/5, & x = 3 \\ 1/5, & x = 4 \end{cases}$$

Therefore:

$$P(2 \leq X \leq 3 | Y = 2) = P^{X|Y}(2|2) + P^{X|Y}(3|2) = 3/5$$

$$\begin{aligned} E(X|Y=2) &= P^{X|Y}(1|2) * 1 + P^{X|Y}(2|2) * 2 + P^{X|Y}(3|2) * 3 + P^{X|Y}(4|2) * 4 \\ &= (1/5) * 1 + (2/5) * 2 + (1/5) * 3 + (1/5) * 4 = 12/5 \end{aligned}$$

(b) Compute $P(Y \leq 2 | X = 4)$ and $E(Y|X = 4)$.

Solution. Divide the $X = 4$ row of the joint PMF array by the row sum 0.25. The conditional PMF of Y given $X = 4$ is then

$$P^{Y|X}(y|4) = \begin{cases} 1/5, & y = 1 \\ 1/5, & y = 2 \\ 1/5, & y = 3 \\ 2/5, & y = 4 \end{cases}$$

$$P(Y \leq 2|X = 4) = P^{Y|X}(1|4) + P^{Y|X}(2|4) = 2/5.$$

$$\begin{aligned} E(Y|X = 4) &= P^{Y|X}(1|4) * 1 + P^{Y|X}(2|4) * 2 + P^{Y|X}(3|4) * 3 + P^{Y|X}(4|4) * 4 \\ &= (1/5) * 1 + (1/5) * 2 + (1/5) * 3 + (2/5) * 4 = 14/5 \end{aligned}$$

Lecture 22

Chapters 4-5 Part 8

22.1 Conditioning one RV on another: Continuous Case

Let X, Y be jointly continuous RV's. For each value of Y, there is a conditional density (conditional PDF) for X given that value of Y. Similarly, for each value of X, there is a conditional density (conditional PDF) for Y given that value of X. In this lecture, I explain how to find these conditional densities and how to do conditional probability and conditional expected value computations using these conditional densities. I will also cover the law of iterated expectation.

22.1.1 Conditional Density $f_{X|Y}(x|y)$

Let y be any value of RV Y. The *conditional density of X given $Y = y$* is denoted $f_{X|Y}(x|y)$ and is defined by

$$f_{X|Y}(x|y) \triangleq \frac{f_{X,Y}(x,y)}{f_Y(y)}. \tag{22.1}$$

In the expression $f_{X|Y}(x|y)$, we are regarding y as being fixed and we are regarding x as a variable which ranges through the values of RV X. Even though we are calling $f_{X|Y}(x|y)$ a *conditional density*, it is a bonafide density function in its own right, that is,

$$\int_{-\infty}^{\infty} f_{X|Y}(x|y)dx = 1.$$

The conditional density $f_{X|Y}(x|y)$ is used to compute conditional probabilities and conditional expected values in the following way:

$$\begin{aligned} P(a \leq X \leq b|Y = y) &= \int_a^b f_{X|Y}(x|y)dx \\ E(X|Y = y) &= \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx \end{aligned}$$

It should be mentioned that you can also do the previous calculations directly, *without finding the conditional PDF* $f_{X|Y}(x|y)$:

$$\begin{aligned} P(a \leq X \leq b|Y = y) &= \frac{\int_a^b f_{X,Y}(x,y)dx}{\int_{-\infty}^{\infty} f_{X,Y}(x,y)dx} \\ E(X|Y = y) &= \frac{\int_{-\infty}^{\infty} x f_{X,Y}(x,y)dx}{\int_{-\infty}^{\infty} f_{X,Y}(x,y)dx} \end{aligned}$$

Remark. In a later lecture, I will prove why formula (22.1) is correct. (The proof uses L'Hospital's Rule from calculus.)

22.1.2 Conditional Density $f_{Y|X}(y|x)$

Let x be any value of RV X. The *conditional density of* Y *given* $X = x$ is denoted $f_{Y|X}(y|x)$ and is defined by

$$f_{Y|X}(y|x) \triangleq \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

In the expression $f_{Y|X}(y|x)$, we are regarding x as being fixed and we are regarding y as a variable which ranges through the values of RV Y. Even though we are calling $f_{Y|X}(y|x)$ a *conditional density*, it is a bonafide density function in its own right, that is,

$$\int_{-\infty}^{\infty} f_{Y|X}(y|x)dy = 1.$$

The conditional density $f_{Y|X}(y|x)$ is used to compute conditional probabilities and conditional expected values in the following way:

$$\begin{aligned} P(a \leq Y \leq b|X = x) &= \int_a^b f_{Y|X}(y|x)dy \\ E(Y|X = x) &= \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy \end{aligned}$$

It should be mentioned that you can also do the previous calculations directly, *without finding the conditional PDF* $f_{Y|X}(y|x)$:

$$P(a \leq Y \leq b|X = x) = \frac{\int_a^b f_{X,Y}(x,y)dy}{\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy} \tag{22.2}$$

$$E(X|Y=y) = \frac{\int_{-\infty}^{\infty} y f_{X,Y}(x,y)dy}{\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy} \tag{22.3}$$

22.1.3 Worked Examples

Example 22.1. Let (X,Y) have joint density

$$f_{X,Y}(x,y) = C\exp\left(-0.25[x^2 - 2xy + 2y^2]\right),$$

where C is the unique positive real constant that makes this a joint density. (This is a special case of the *joint Gaussian density.*) Let us first find the conditional PDF of X given $Y=1$, namely, we want to find $f_{X|Y}(x|1)$. Instead of finding $f_{X|Y}(x|1)$ by plugging into formula (22.1), I show you here another approach that is sometimes easier. We can think of our conditional density as having the form

$$f_{X|Y}(x|1) = C' f_{X,Y}(x,1), \quad -\infty < x < \infty,$$

where the constant C' is chosen so that $f_{X|Y}(x|1)$ integrates to 1. Note that

$$x^2 - 2xy + 2y^2 = (x-y)^2 + y^2,$$

so that we may manipulate $f_{X|Y}(x|1)$ into the form

$$f_{X|Y}(x|1) = C''\exp\left(-\frac{(x-1)^2}{4}\right).$$

This is clearly the form of a Gaussian density function. We immediately conclude that the conditional distribution of X given $Y=1$ is the Gaussian distribution with mean 1 and variance 2. This gives us the complete description of this conditional density as

$$f_{X|Y}(x|1) = \left(\frac{1}{\sqrt{2\pi}\sqrt{2}}\right)\exp\left(-\frac{(x-1)^2}{4}\right).$$

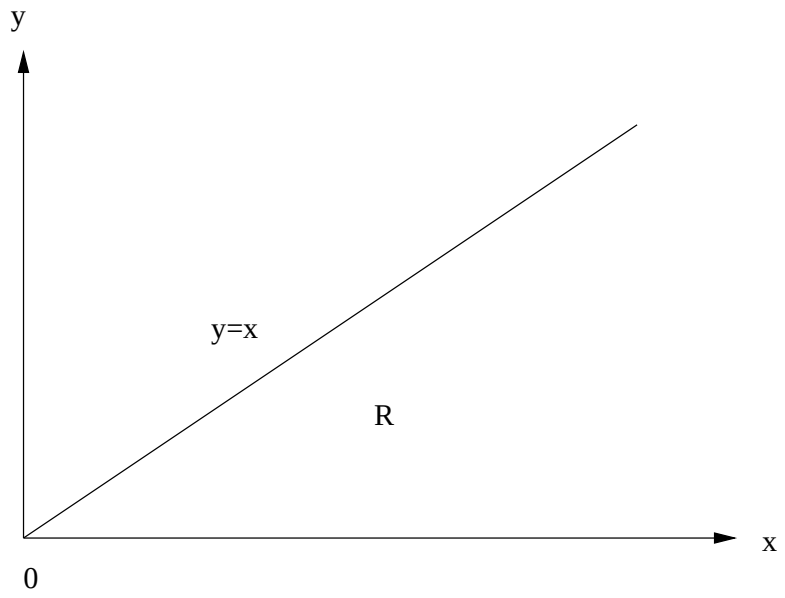
It is also immediate that

$$\begin{aligned} E(X|Y=1) &= 1 \\ Var(X|Y=1) &= 2 \end{aligned}$$

Exercise. In the preceding example, use the same technique to determine the precise expression for $f_{Y|X}(y|1)$ without doing any integration. Also, give the values of $E(Y|X=1)$ and $Var(Y|X=1)$ without doing any computation.

Moral. If (X, Y) is joint Gaussian, then any of its conditional PDF's are one-dimensional Gaussian densities. (This is valid because the technique of Example 22.1 will apply to any joint Gaussian PDF.)

Example 22.2. Let R be the infinite triangular region below.



Let $f(x, y)$ be the joint PDF of random variables X, Y as follows:

$$f(x,y) = \begin{cases} Ce^{-(x+y)}, & (x,y) \in R \\ 0, & elsewhere \end{cases}$$

(The value of the positive constant C is not needed in this problem.) Let us find the conditional PDF of Y given $X = 2$; that is, we are going to find $f_{Y|X}(y|2)$. Locate the point $x = 2$ on the x-axis in the above plot, and then move up from there along a vertical slice through region R that goes from $y = 0$ to $y = 2$. Given $X = 2$, this tells us that Y can only vary from 0 to 2. Plugging $x = 2$ into the joint density, we see that the conditional PDF $f_{Y|X}(y|2)$ takes the form

$$f_{Y|X}(y|2) = C' \exp(-y), \;\; 0 \le y \le 2 \; (zero \; elsewhere).$$

Let us now go one step further and compute $E(Y|X = 2)$. First, we need to evaluate the constant C':

$$C' = \frac{1}{\int_0^2 \exp(-y)dy} = 1.1565.$$

We then have:

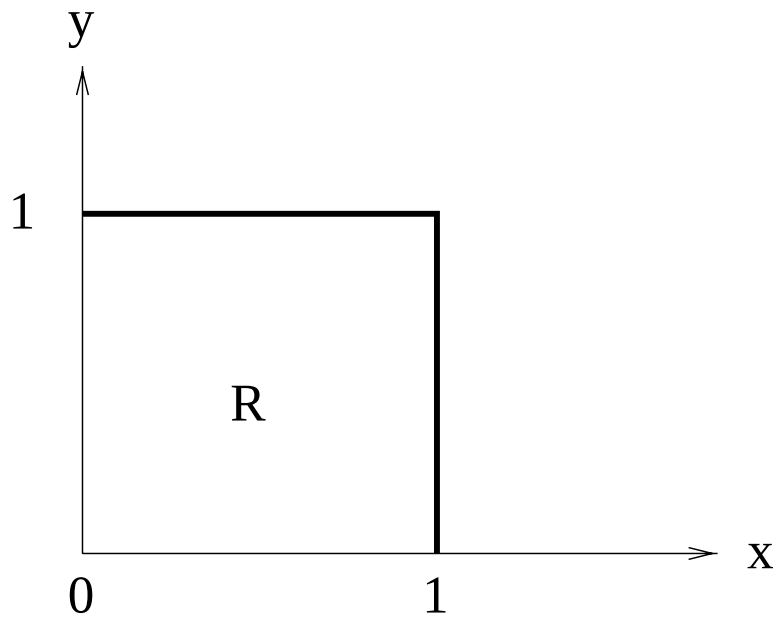
$$E(Y|X=2) = \int_0^2 y f_{Y|X}(y|2)dy = \int_0^2 (1.1565)y\exp(-y)dy = 0.6870.$$

Exercise. In the preceding example, prove that

$$E(Y|X=x) = \frac{1 - xe^{-x} - e^{-x}}{1 - e^{-x}}, \quad x \geq 0.$$

If you get stuck, go to Problem 7.3 of the Chapter 4-5 Solved Problems.

Example 22.3. Let R be the region below.



Let random variables X, Y have the joint density

$$f(x,y) = \begin{cases} x+y, & (x,y) \in R \\ 0, & elsewhere \end{cases}$$

As a change of pace, let us compute $P(0 \leq Y \leq 1/4|X = 1/2)$ and $E(Y|X = 1/2)$ *without* finding $f_{Y|X}(y|1/2)$. Using formulas (22.2)-(22.3), we have:

$$P(0 \leq Y \leq 1/4|X = 1/2) \quad = \quad \frac{\int_0^{1/4}(0.5+y)dy}{\int_0^1 (0.5+y)dy} = 0.1562$$

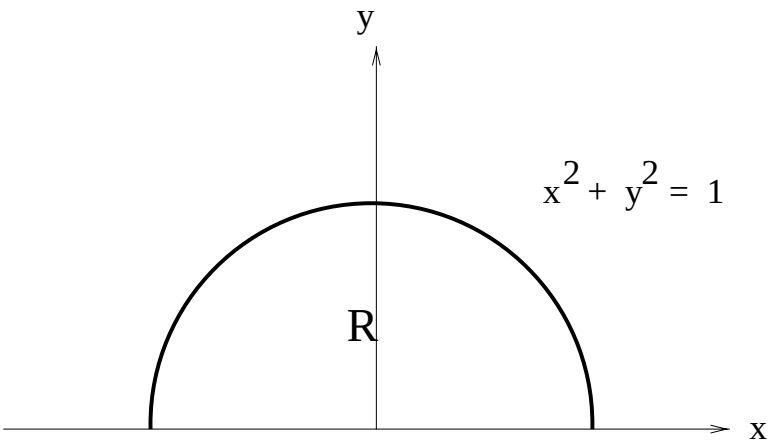
$$E(Y|X=1/2) = \frac{\int_0^1 y(0.5+y)dy}{\int_0^1 (0.5+y)dy} = 0.5833$$

Exercise. In the preceding example, prove that

$$E(X|Y=y) = \frac{1/3+y/2}{y+1/2}, \ \ 0 \le y \le 1.$$

If you get stuck, consult Problem 7.2 of the Chapter 4-5 Solved Problems.

Exammple 22.4. Let the random point (X,Y) be chosen uniformly from the semicircular region R as follows:



Notice that X varies from -1 to 1. Let us fix $X=x$ for an arbitrary x satisfying $-1 \le x \le 1$. We will now find the conditional distribution of Y given $X=x$. To see what this would be, locate the point x on the x-axis in the above figure and then go up along a vertical slice through R. This slice goes in the y direction from $y=0$ to $y=\sqrt{1-x^2}$. Along this slice, the joint density is constant. Therefore, we come to the important conclusion that given $X=x$, Y is conditionally uniformly distributed from $y=0$ to $y=\sqrt{1-x^2}$. We can immediately conclude from this that

$$E(Y|X=x) = (1/2)\sqrt{1-x^2}, \ \ -1 \le x \le 1,$$

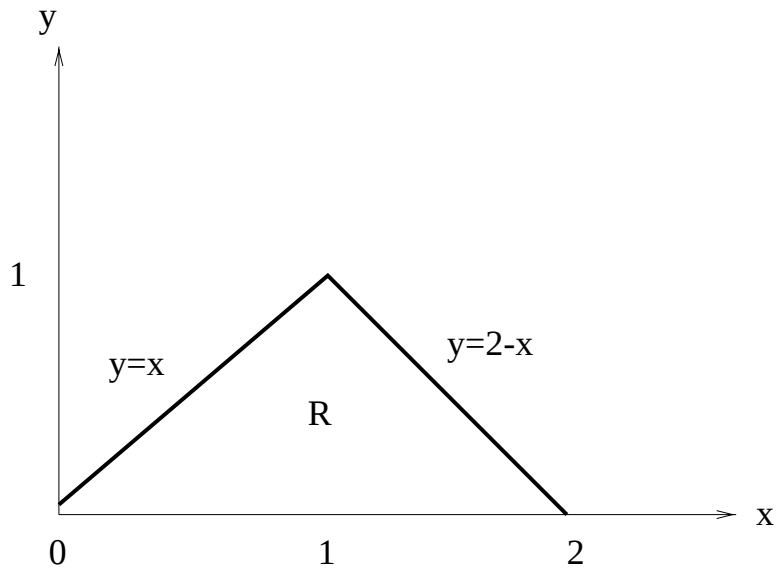
since the mean of a uniform distribution is the midpoint of the interval over which the distribution extends. Similarly, for each fixed y satisfying $0 \le y \le 1$, we'd be able to argue that the conditional

distribution of X given $Y = y$ is a uniform distribution extending from $x = -\sqrt{1-y^2}$ to $x = \sqrt{1-y^2}$. We'd be able to easily conclude from this that

$$\begin{aligned} E(X|Y=y) &= 0 \\ Var(X|Y=y) &= \frac{1-y^2}{3} \end{aligned}$$

Moral. We conclude from Example 22.4 that if (X, Y) is jointly uniformly distributed in region R, then every single conditional PDF is a one-dimensional uniform density. The interval over which each of these conditional uniform distributions extends is determined by where the appropriate slice (horizontal or vertical) through R begins and ends.

Exercise. Let R be the triangular region:



Using the above "Moral," draw the following conclusion by inspection:

$$E[Y|X=x] = \begin{cases} x/2, & 0 \le x \le 1 \\ (2-x)/2, & 1 < x \le 2 \end{cases}$$

You should also be able to draw the conclusion that

$$E[X|Y=y] = 1, \;\; 0 \le y \le 1.$$

22.2 Law of Iterated Expectation

Suppose X, Y are two RV's. The notation $E(Y|X)$ will denote the random variable which takes the value $E(Y|X = x)$ when X takes the value x. Thus, the random variable $E(Y|X)$ is a function of the random variable X.

The *law of iterated expectation* is the following formula:

$$E[E(Y|X)] = E(Y).$$

In other words, to calculate $E(Y)$, you can first calculate $E(Y|X)$ (the first expected value) and then you can compute the expected value of the random variable $E(Y|X)$ (the second expected value). Because two expected value operations are involved, you see how this law got its name.

The law of iterated expectation is particularly suited to two step experiments in which X is observed as the result of the first step, and then Y is observed conditioned on the X value as the second step of the experiment.

Here is the easy proof of the law of iterated expectation. First, you can write

$$E[E(Y|X)] = \int_{-\infty}^{\infty} E[Y|X = x] f_X(x) dx. \tag{22.4}$$

You can then substitute

$$E[Y|X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy.$$

The right side of (22.4) then becomes

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_X(x) f_{Y|X}(y|x) dy dx. \tag{22.5}$$

Then, using the fact that

$$f_{X,Y}(x, y) = f_X(x) f_{Y|X}(y|x),$$

it is a simple matter to show that the right side of (22.5) is equal to $E(Y)$.

Here is a generalization of the law of iterated expectation that is easy to prove based upon the method just used:

$$E[\phi(X)\psi(Y)] = E[\phi(X)E[\psi(Y)|X]]. \tag{22.6}$$

In equation (22.6), $\phi(X)$ can be any function of RV X and $\psi(Y)$ can be any function of RV Y. For example, if you take $\phi(X) = X$ and $\psi(Y) = Y$, you obtain the following useful formula for computing correlation:

$$E[XE(Y|X)] = E[XY].$$

Example 22.5. Let's go back to our ice cream example. Bill eats X ice cream cones, where X is Poisson with mean 1. Given that $X = x$, Bill runs Y miles, where Y is the number of heads in

tossing a fair coin $x+1$ times. Notice that $E[Y|X=x]$ is the mean of a Binomial(n,p) distribution with $n=x+1$ and $p=1/2$. By Appendix A, this is

$$E[Y|X=x] = np = (x+1)/2.$$

We conclude that

$$E(Y|X) = (X+1)/2.$$

Therefore,

$$E(Y) = E[E(Y|X)] = E[(X+1)/2] = (E[X]+1)/2 = 1.$$

The law of iterated expectation has allowed us to see that the expected number of ice cream cones that Bill eats is 1. If we tried to compute $E(Y)$ directly from the PMF of Y, we would have a difficult time because it is not easy to find the PMF of Y. (We tried to find some PMF values for Y during one of our earlier recitations.) Let us go further and compute the correlation $E(XY)$. We obtain

$$E(XY) = E[XE(Y|X)] = E[X(X+1)/2] = (1/2)(E[X^2]+E[X]) = 1.5.$$

(In this last calculation, I used the fact that the mean and variance of X are both 1.) We can also compute the second moment of Y:

$$E[Y^2] = E[E(Y^2|X)] = E[Var(Y|X) + E(Y|X)^2].$$

From Appendix A,

$$Var(Y|X=x) = np(1-p) = (x+1)/4,$$

and so

$$Var(Y|X) = (X+1)/4.$$

This gives us

$$E[Y^2] = E[(X+1)/4 + (X+1)^2/4] = 7/4.$$

The variance of Y is therefore

$$Var(Y) = 7/4 - 1^2 = 3/4.$$

Remark. The reader will find several more worked examples on law of iterated expectation in Section 8 of the Chapter 4-5 Solved Problems.

Lecture 23

Chapters 4-5 Part 9

23.1 Odds and Ends

Here I take the time to discuss some theoretical issues that were deferred from earlier lectures.

23.1.1 Justification of Conditional PDF Formula

I show you why the formula

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

is valid for a pair of jointly continuous RV's (X, Y). This means I need to show you why the formula

$$P[X \in A|Y = y] = \int_A \left(\frac{f_{X,Y}(x,y)}{f_Y(y)} \right) dx, \tag{23.1}$$

is true for every event $\{X \in A\}$. The left hand side cannot be evaluated directly because the event $\{Y = y\}$ has probability zero. Instead, I will evaluate it as

$$P[X \in A|Y = y] = \lim_{\Delta y \to 0} P[X \in A|y \le Y \le y + \Delta y]. \tag{23.2}$$

By Chapter 1,

$$P[X \in A|y \le Y \le y + \Delta y] = \frac{P[X \in A, \ \ y \le Y \le y + \Delta y]}{P[y \le Y \le y + \Delta y]}.$$

The numerator is

$$P[X \in A, \ \ y \le Y \le y + \Delta y] = \int_y^{y+\Delta y} \int_A f_{X,Y}(x,y) dx dy, \tag{23.3}$$

and the denominator is

$$P[y \leq Y \leq y + \Delta y] = \int_y^{y+\Delta y} f_Y(y)dy. \tag{23.4}$$

Notice that both the numerator and the denominator approach 0 as $\Delta y \to 0$. Therefore, the limit in (23.2) is an indeterminate of the form 0/0. In this case, calculus tells us that we can use L'Hospital's Rule to compute the limit. We must divide the derivative of the numerator (with respect to Δy) by the derivative of the denominator, and then let $\Delta y \to 0$ to obtain the limit in (23.2). By the fundamental theorem of calculus, the derivative of the numerator (23.3) is

$$\int_A f_{X,Y}(x, y + \Delta y)dx,$$

and the derivative of the denominator (23.4) is

$$f_Y(y + \Delta y).$$

The limit of the quotient of the derivatives is therefore

$$\frac{\int_A f_{X,Y}(x, y)dx}{f_Y(y)},$$

which gives us formula (23.1), completing our proof.

23.1.2 Factoring Joint PDF/PMF; Independence

Suppose we have a two-step experiment in which we observe the value of RV X on the first step, and then we observe the value of RV Y given the value of X on the second step. In such a scenario, we would probably not be given the joint distribution of (X, Y) "up front". Instead, we would be given the distribution of X followed by the conditional distribution of Y given each possible observed value of X; we could then combine these two distributions by multiplication to obtain the joint distribution; this gives a factorization of the joint distribution into two parts. If X, Y are discrete, this factorization takes the form:

$$P^{X,Y}(x, y) = P^X(x)P^{Y|X}(y|x). \tag{23.5}$$

If X, Y are jointly continuous, this factorization takes the form:

$$f_{X,Y}(x, y) = f_X(x)f_{Y|X}(y|x). \tag{23.6}$$

It is instructive to see what these formulas tell us when X, Y are independent. If X, Y are independent, then in the case of discrete X, Y, we obtain the factorization of the joint PMF into the product of the marginal PMF's:

$$P^{X,Y}(x, y) = P^X(x)P^Y(y). \tag{23.7}$$

Comparing (23.7) to (23.5), we see that

$$P^{Y|X}(y|x) = P^Y(y)$$

if and only if the discrete RV's X, Y are independent. Or, we can reverse the roles of X and Y and conclude that

$$P^{X|Y}(x|y) = P^X(x)$$

if and only if the discrete RV's X, Y are independent. If (X, Y) are jointly continuous, one can make similar conclusions: From the factorization

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

and equation (23.6), one concludes that

$$f_{Y|X}(y|x) = f_Y(y)$$

if and only if X, Y are independent, or, equivalently,

$$f_{X|Y}(x|y) = f_X(x)$$

if and only if X, Y are independent. We can summarize these conclusions as follows:

Conclusion: X, Y are independent if and only if every conditional distribution is equal to the marginal (unconditional) distribution that you obtain by dropping the condition.

Example 23.1. Previously, we discussed how to find out whether X, Y are dependent or independent without using conditional distributions. Now, using the above Conclusion, we can sometimes decide very quickly that two RV's are dependent or independent by appealing to conditional distributions. Suppose, for example, that (X, Y) is jointly continuously distributed over the entire semicircular region

$$R = \{(x, y) : x^2 + y^2 \leq 1;\ x \geq 0\}.$$

X ranges from 0 to 1. Clearly, when $X = 0$, the conditional distribution of Y will range from -1 to 1. However, when $X = 1$, then Y will always be 0. I have picked out two conditional distributions that are different. That is enough to conclude that X, Y must be dependent RV's. (Actually, more than just two conditional distributions are different: as x varies from 0 to 1, all of the conditional PDF's $f_{Y|X}(y|x)$ are different, because the corresponding vertical slices through R have different starting and ending y values. Thus, no two of the conditional distributions of Y given X are the same!)

23.2 Distribution of Sum of Independent RV's

Let $X_1, X_2, \cdots, X_n$ be independent RV's. (Although I have not yet defined independence of RV's for more than two RV's, what I mean by independence is that

$$P[X_1 \in A_1, X_2 \in A_2, \cdots, X_n \in A_n] = \prod_{i=1}^{n} P[X_i \in A_i]$$

for all choices of events $\{X_i \in A_i\}$.) Let

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

be the sum of all these independent RV's. I will prove later that the PDF of X is obtained by convolution of the separate PDF's of the X_i's:

$$f_X = f_{X_1} * f_{X_2} * \cdots * f_{X_n}. \tag{23.8}$$

Recall that the Laplace transform of a convolution is the product of the separate Laplace transforms. Therefore, we can say from equation (23.8) that

$$\mathcal{L}[f_X] = \prod_{i=1}^{n} \mathcal{L}[f_{X_i}], \tag{23.9}$$

where $\mathcal{L}$ denotes the Laplace transform operator. If you replace Laplace variable s by $-s$, then you obtain the moment generating function. Therefore, we can also say that

$$M_X(s) = \prod_{i=1}^{n} M_{X_i}(s). \tag{23.10}$$

Example 23.2. Let X be the number of heads on the toss of 3 fair coins. From earlier in the course, we already know that X is Binomial(n, p) with $n = 3$ and $p = 1/2$. Here, we show another way to obtain this result using convolution. We can write

$$X = X_1 + X_2 + X_3,$$

where X_i is equal to 1 if the i-th coin comes up heads and is equal to 0 otherwise. Each X_i has PDF

$$(1/2)\delta(x) + (1/2)\delta(x - 1).$$

The Laplace transform is

$$0.5 + 0.5e^{-s}.$$

By (23.9), we see that the PDF of X is the inverse Laplace transform of

$$(0.5 + 0.5e^{-s})^3 = (1/8) + (3/8)e^{-s} + (3/8)e^{-2s} + (1/8)e^{-3s}.$$

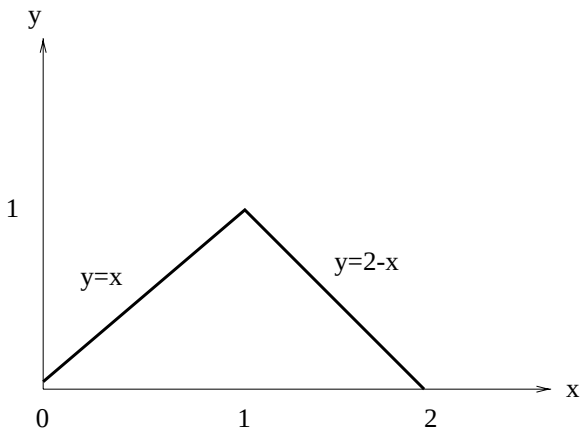
Inverting, we have

$$f_X(x) = (1/2)\delta(x) + (3/8)\delta(x-1) + (3/8)\delta(x-2) + (1/8)\delta(x-3).$$

Example 23.3. Let X_1, X_2 be independent Uniform$(0,1)$ RV's. Let

$$X = X_1 + X_2.$$

The PDF's of X_1 and X_2 are the same, namely, a rectangular pulse from $x = 0$ to $x = 1$ of amplitude 1. If we convolute this rectangular pulse with itself, we obtain a symmetric triangular pulse that starts at $x = 0 + 0 = 0$ and ends at $x = 1 + 1 = 2$. This triangular pulse is the density of X, and since it must have area one under it, the plot of $f_X(x)$ must be as follows:



Exercise. Re-work Example 23.3 where you again assume that X_1 is Uniform$(0,1)$, but instead you assume that X_2 is Uniform$(0,2)$. Determine $f_X(x)$. (Hint: $f_X(x)$ is a symmetric trapezoidal pulse starting at $x = 0$ and ending at $x = 1 + 2 = 3$.)

Example 23.4. Let $X = X_1 + X_2 + X_3$, where the X_i's are independent and each X_i has the exponential distribution with mean 1, that is

$$f_{X_i}(x) = \exp(-x)u(x).$$

The Laplace transform of the preceding is $1/(s+1)$, and therefore

$$f_X(x) = \mathcal{L}^{-1}\left[\frac{1}{(s+1)^3}\right] = (x^2/2)\exp(-x)u(x).$$

Exercise. Re-work Example 23.4, where you again assume that X_1 is exponential with mean 1, but instead you assume that X_2 is exponential with mean 2 and X_3 is exponential with mean 3. Hint: $f_X(x)$ is the inverse transform of

$$\left(\frac{1}{s+1}\right)\left(\frac{1/2}{s+1/2}\right)\left(\frac{1/3}{s+1/3}\right).$$

Example 23.5. Let $X_1, X_2, \cdots, X_n$ be independent Gaussian RV's. Let us prove that

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

is also a Gaussian RV. Letting μ_i be the mean of X_i and letting σ_i^2 be the variance of X_i, we see from Chapter 6 of your textbook that the moment generating function of X_i is

$$M_{X_i}(s) = \exp(\mu_i s + 0.5\sigma_i^2 s^2).$$

Using equation (23.10), we see that

$$M_X(s) = \prod_{i=1}^{n} \exp(\mu_i s + 0.5\sigma_i^2 s^2) = \exp\left((\sum_{i=1}^{n}\mu_i)s + 0.5(\sum_{i=1}^{n}\sigma_i^2)s^2\right).$$

From the preceding equation, we see that $M_X(s)$ has the form of a Gaussian MGF. We conclude that X must be Gaussian with mean

$$\mu_1 + \mu_2 + \cdots + \mu_n$$

and variance

$$\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2.$$

Remark. You can find more worked examples of this type in Section 6.4 of your textbook and in Section 9 of the Chapter 4-5 Solved Problems.

Lecture 24

Chapters 4-5 Part 10

24.1 Distribution of Max/Min of Independent RV's

Let $X_1, X_2, \cdots, X_n$ be independent RV's.

Distribution of Max: Let

$$X = \max(X_1, X_2, \cdots, X_n).$$

Then the PDF $f_X(x)$ of X is given by the following formula:

$$f_X(x) = \frac{d}{dx}\left\{\prod_{i=1}^{n} F_{X_i}(x)\right\}. \tag{24.1}$$

Distribution of Min: Let

$$X = \min(X_1, X_2, \cdots, X_n).$$

Then the PDF $f_X(x)$ of X is given by the following formula:

$$f_X(x) = -\frac{d}{dx}\left\{\prod_{i=1}^{n}(1 - F_{X_i}(x))\right\}. \tag{24.2}$$

Proof of (24.1). Let X be the maximum of the X_i's. Note that

$$\{X \leq x\} = \cap_{i=1}^{n}\{X_i \leq x\}. \tag{24.3}$$

(If the biggest of a bunch of numbers is $\leq x$ then every single number is $\leq x$ and vice-versa.) The events on the right side of (24.3) are independent. The product of an intersection of independent events is the product of the probabilities of the separate events. Therefore,

$$P[X \leq x] = \prod_{i=1}^{n} P[X_i \leq x].$$

The probabilities on each side of the preceding equation are all CDF's and so

$$F_X(x) = \prod_{i=1}^{n} F_{X_i}(x).$$

Differentiating both sides with respect to x, you get (24.1).

Proof of (24.2). Let X be the minimum of the X_i's. Note that

$$\{X > x\} = \cap_{i=1}^{n} \{X_i > x\}. \tag{24.4}$$

(If the smallest of a bunch of numbers is $> x$ then every single number is $> x$ and vice-versa.) The events on the right side of (24.4) are independent. The product of an intersection of independent events is the product of the probabilities of the separate events. Therefore,

$$P[X > x] = \prod_{i=1}^{n} P[X_i > x].$$

The probabilities on each side of the preceding equation are all $1-$ CDF's and so

$$1 - F_X(x) = \prod_{i=1}^{n} (1 - F_{X_i}(x)).$$

Differentiating both sides with respect to x, and then multiplying by -1, you get (24.2).

Example 24.1. Consider the system

$$A \to \boxed{1} \to \boxed{2} \to \boxed{3} \to B,$$

with subsystems 1, 2, 3 connected in series. The object of this system is for something to flow from point A to point B. For $i = 1, 2, 3$, let the random lifetime T_i of subsystem i be exponentially distributed. Let us find the PDF of T_{AB}, the lifetime of the connection from A to B. We have

$$T_{AB} = \min(T_1, T_2, T_3).$$

Plug into equation (24.2) to get the PDF f_{T_AB}(t):

$$f_{T_{AB}}(t) = -\frac{d}{dt} \left\{ \prod_{i=1}^{3} (1 - F_{T_i}(t)) \right\}.$$

It is easy to compute that

$$1 - F_{T_i}(t) = \exp(-a_i t), \quad t \geq 0$$

where a_i is the reciprocal of the expected lifetime of T_i. Therefore,

$$f_{T_{AB}}(t) = -\frac{d}{dt}\exp(-(a_1+a_2+a_3)t),$$

which simplifies to

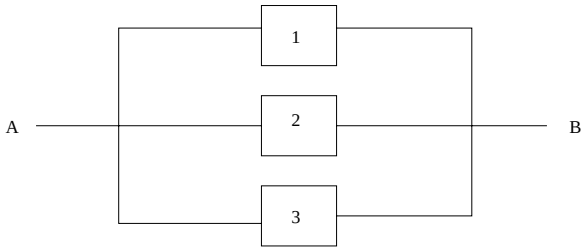
$$f_{T_{AB}}(t) = (a_1+a_2+a_3)\exp(-(a_1+a_2+a_3)t)u(t).$$

That is, T_{AB} is exponentially distributed with parameter $a_1+a_2+a_3$. We obtain the following nice formula expressing the expected lifetime of the overall system in terms of the expected lifetimes of its subsystems:

$$E[T_{AB}] = \left(\frac{1}{E[T_1]}+\frac{1}{E[T_2]}+\frac{1}{E[T_3]}\right)^{-1}.$$

What do you think this formula would become for a system consisting of n subsystems connected in series, where n can be any positive integer ≥ 2?

Example 24.2. Consider the following system with subsystems 1, 2, 3 connected in parallel:



The object of this system is for something to flow from point A to point B. For $i = 1, 2, 3$, let the random lifetime T_i of subsystem i be exponentially distributed. Let us find the PDF of T_{AB}, the lifetime of the connection from A to B. We have

$$T_{AB} = \max(T_1, T_2, T_3).$$

Plug into equation (24.1) to get the PDF f_{T_AB}(t):

$$f_{T_{AB}}(t) = \frac{d}{dt}\prod_{i=1}^{3}(1-\exp(-a_i t)).$$

Use the product rule of differentiation to finish. You obtain

$$\begin{aligned} f_{T_{AB}}(t) \quad = \quad & [a_1\exp(-a_1t)(1-\exp(-a_2t))(1-\exp(-a_3t)) + \\ & a_2\exp(-a_2t)(1-\exp(-a_1t))(1-\exp(-a_3t)) + \\ & a_3\exp(-a_3t)(1-\exp(-a_1t))(1-\exp(-a_2t))]\,u(t) \end{aligned}$$

24.2 Application to Mean-Square Receiver Design

As we get toward the end of our Chapter 4-5 material, we will now and again consider various applications. In the present section, we consider the application to *mean-square receiver design.* The following block diagram gives us the scenario we are operating under in this application:

$$X \rightarrow \boxed{\text{channel}} \rightarrow Y \rightarrow \boxed{\begin{array}{c}\text{MS}\\ \text{receiver}\end{array}} \rightarrow \hat{X} = \phi(Y)$$

The channel, random input X to channel, and random output Y from channel are fixed. Our job is to design a receiver which converts the channel output into an estimate $\hat{X}$ of X, which can in general be any function $\phi(Y)$ of Y. The receiver is referred to as "mean-square receiver" (MS receiver for short) because in order to see how good a job the receiver is doing, we measure the so-called mean-square estimation error, defined by the formula

$$\text{mean-square estimation error} \triangleq E[(X - \hat{X})^2]$$

The closer the mean-square estimation error is to zero, the better the job that the mean-square receiver is doing. Here are five common types of MS receivers:

Default Receiver: The *default receiver* simply declares that the estimate is

$$\hat{X} = 0.$$

When you use a default receiver, you are ignoring the value of Y coming into the receiver and you are ignoring any information about the probability distribution of X. The mean-square estimation error of the default receiver is

$$E[(X - \hat{X})^2] = E[(X - 0)^2] = E[X^2],$$

the second moment of the channel input RV X.

Blind Receiver: The *blind receiver* declares that the estimate is

$$\hat{X} = \mu_X.$$

"Blind" refers to the fact that you are ignoring the value of Y coming into the receiver. However, the blind receiver does use information about the probability distribution of X, namely, the mean of X. The mean-square estimation error of the blind receiver is

$$E[(X - \hat{X})^2] = E[(X - \mu_X)^2] = \sigma_X^2,$$

the variance of X. We know from the Chapter 2-3 material that $Var(X) \leq E[X^2]$. Therefore, the blind receiver is doing a better job of estimating X than the default receiver.

Correlation Receiver: The *correlation receiver* was covered in Section 21.1. It is defined by

$$\hat{X} \triangleq \left(\frac{E[XY]}{E[Y^2]}\right) Y. \tag{24.5}$$

As discussed in Section 21.1, we can think of the correlation receiver geometrically as the projection of X on Y. We can also think of the correlation receiver as the unique constant multiple of Y which is closest to X in the mean-square sense (that is, of all $\hat{X}$'s which are constant multiples of Y, $E[(X-\hat{X})^2]$ is minimized for the correlation receiver output $\hat{X}$ given by (24.5)).

Straight Line Receiver: The purpose of the *straight line receiver* is to produce an estimate of X of the form

$$\hat{X} = AX + B, \tag{24.6}$$

where A, B are constants chosen to give the smallest mean-square estimation error $E[(X-\hat{X})^2]$ among all estimates of "straight-line" form (24.6). The straight line receiver operates in two steps. In the first step, you project $X - \mu_X$ on $Y - \mu_Y$. What this does is give you the best mean-square estimate of $X - \mu_X$ which is a constant multiple of $Y - \mu_Y$. The result of this first step is precisely what you get from the right side of formula (24.5) when you substitute $X - \mu_X$ for X and $Y - \mu_Y$ for Y, namely

$$\left(\frac{E[(X-\mu_X)(Y-\mu_Y)]}{E[(Y-\mu_Y)^2]}\right)(Y-\mu_Y) = \frac{\sigma_{X,Y}}{\sigma_Y^2}(Y-\mu_Y) = \rho(\sigma_X/\sigma_Y)(Y-\mu_Y)$$

is what you get from Step 1, where ρ is the correlation coefficient $\rho_{X,Y}$. Notice that the result of our first step, since it can be regarded as an estimate of $X - \mu_X$, can be adjusted to obtain an estimate of X if we add μ_X to it; this is Step 2. In other words, the second step of forming the straight line receiver estimate $\hat{X}$ is to add μ_X to the result of Step 1. This gives us the following formula for the straight line receiver estimate, which we can take as a definition:

$$\hat{X} \triangleq \mu_X + \rho(\sigma_X/\sigma_Y)(Y-\mu_Y).$$

Minimum MS Receiver: The *minimum mean-square receiver* (minimum MS receiver for short) is the receiver defined by

$$\hat{X} \triangleq E(X|Y),$$

where $E(X|Y)$ is the conditional expectation random variable discussed in our earlier section on the law of iterated expectation. That is, if the minimum MS receiver input is $Y = y$, then the value of the estimate $\hat{X}$ is very intuitive because it is

$$E(X|Y = y),$$

the conditional expected value of X given the condition that $Y = y$. We call this receiver the "minimum" MS receiver because we will prove in the following that it gives the smallest MS estimation error of all possible receivers.

If we have a receiver, let the notation $e(receiver)$ be the mean-square estimation error that results from using this receiver.

Useful Facts

- The default receiver, blind receiver, straight line receiver, and minimum MS receiver are successively better receivers, that is, their MS estimation errors get smaller and smaller:

$$e(default\ receiver) \geq e(blind\ receiver)$$
$$e(blind\ receiver) \geq e(st\ line\ receiver)$$
$$e(st\ line\ receiver) \geq e(min\ MS\ receiver)$$

- The default receiver, correlation receiver, straight line receiver, and minimum MS receiver are successively better receivers, that is, their MS estimation errors get smaller and smaller:

$$e(default\ receiver) \geq e(corr\ receiver)$$
$$e(corr\ receiver) \geq e(st\ line\ receiver)$$
$$e(st\ line\ receiver) \geq e(min\ MS\ receiver)$$

- Sometimes the correlation receiver is better than the blind receiver, and sometimes the blind receiver is better than the correlation receiver.

Discussion. Let us see why the facts given above are true. First, let us investigate the performance of the minimum MS receiver vis-a-vis other receivers. Let $\phi(Y)$ denote the estimate of X generated by an arbitrary MS receiver. For each value y of Y, we know from Chapter 2-3 Notes that

$$E[(X-\phi(y))^2|Y=y] = E[(X-E[X|Y=y])^2|Y=y] + (\phi(y)-E[X|Y=y])^2.$$

Therefore,

$$E[(X-\phi(y))^2|Y=y] \geq E[(X-E[X|Y=y])^2|Y=y].$$

Multiplying both sides of the preceding inequality by $f_Y(y)$ are integrating from $-\infty$ to ∞, we obtain the following inequality by the law of iterated expectation:

$$E[(X-\phi(Y))^2] \geq E[(X-E(X|Y))^2]. \tag{24.7}$$

This inequality tells us that the MS estimation error of our arbitrary MS receiver (left hand side of (24.7)) is greater than or equal to the MS estimation error of the minimum MS receiver (right hand side of (24.7)). We conclude that the minimum MS receiver is indeed the best of all the MS receivers.

Secondly, let's investigate the performance of the straight line receiver vis-a-vis other MS receivers. By definition, the straight line receiver is the MS receiver yielding estimate of the straight line form

$$\hat{X} = AX + B \tag{24.8}$$

that yields the smallest MS estimation error among all MS receivers generating an estimate of the straight line form. The correlation receiver, the blind receiver, and the default receiver all yield estimates of the form (24.8). (The corr receiver yields $B = 0$, the blind receiver yields $A = 0$, and the default receiver yields $A = B = 0$.) Therefore, the straight line receiver must have MS estimation error at least as small as these other three types of receivers.

It is easy to argue that the correlation receiver has MS estimation error less than or equal to that of the default receiver. (Do you see why this is true?) Also, we have already remarked that the blind receiver is better than the default receiver in our earlier discussion of the blind receiver.

To conclude our discussion, we present a couple of examples which show us that in general the correlation receiver is not better than the blind receiver and vice-versa.

Example 24.3. Let X, Y be independent, with $\mu_X \neq 0$ and $\mu_Y = 0$. Then,

$$E[XY] = E[X]E[Y] = 0$$

and so the correlation receiver coincides with the default receiver. The blind receiver is better than the correlation receiver in this case.

Example 24.4. Let $X = Y$ and let X have positive variance. Then the correlation receiver estimate is $\hat{X} = Y = X$, which yields MS estimation error 0. The blind receiver yields MS estimation error $\sigma_X^2 > 0$. The correlation receiver is better than the blind receiver in this case.

Lecture 25

Chapters 4-5 Part 11

25.1 Application to Reliability

Before discussing how this application works, we need the following result.

Result: If X is any nonnegative continuously distributed RV, then

$$E[X] = \int_0^\infty P[X \geq x]dx. \tag{25.1}$$

Proof of Result. Let $\phi(x, y)$ be the function

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

Calculus tells us that

$$\int_0^\infty \int_0^\infty \phi(x, y) f_X(x) dy dx = \int_0^\infty \int_0^\infty \phi(x, y) f_X(x) dx dy.$$

If you evaluate these two double integrals, you will see that formula (25.1) results.

The method I am going to show you gives an easy way to determine the expected lifetime of a system built up from independently acting subsystems. Rather than give you a general description of the method, I illustrate the use of the method in a couple of examples.

Example 25.1. We first apply the method to the system in Example 24.1, in order to show you that you obtain the same answer. We have the system

$$A \to \boxed{1} \to \boxed{2} \to \boxed{3} \to B,$$

with subsystems 1, 2, 3 connected in series. The object of this system is for something to flow from point A to point B. For $i = 1, 2, 3$, let the random lifetime T_i of subsystem i be exponentially distributed. Let

$$a_i = \frac{1}{E[T_i]}, \quad i = 1, 2, 3.$$

We want to compute $E[T_{AB}]$, the expected lifetime of the A to B connection. The so-called *reliability function* $R(t)$ of the system is defined by

$$R(t) \stackrel{\Delta}{=} P[T_{A,B} \geq t], \quad t \geq 0.$$

Using formula (25.1), we can express $E[T_{AB}]$ in terms of the reliability function as

$$E[T_{AB}] = \int_0^\infty R(t)dt. \tag{25.2}$$

The reliability function $R(t)$ is easily determined by Chapter 1 techniques. To do this, define p_1, p_2, p_3 as follows:

$$p_i = P[T_i \geq t] = \exp(-a_i t).$$

We can interpret each p_i as the probability that subsystem i works at time t, and we can interpret $R(t)$ as the probability that the overall system works at time t, where we regard t as a parameter that is ≥ 0. Suppose we "freeze" the system at time t: In terms of what is happening at time t alone, we can view each subsystem $i = 1, 2, 3$ as a relay switch which is either working with prob p_i or not working with prob $1 - p_i$, and we can view the overall system as a relay circuit which is working with prob $R(t)$ and not working with prob $1 - R(t)$. Chapter 1 tells us that the probability that a relay circuit consisting of three switches in series will work is $p_1 p_2 p_3$. Thus, it is immediate that

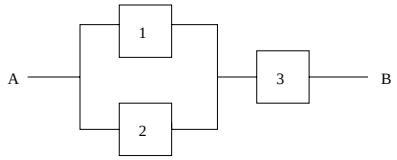
$$R(t) = p_1 p_2 p_3.$$

Using (25.2), we see that

$$E[T_{AB}] = \int_0^\infty p_1 p_2 p_3 dt = \int_0^\infty \exp(-(a_1 + a_2 + a_3)t)dt = \frac{1}{a_1 + a_2 + a_3}.$$

This is the same answer we obtained in Example 24.1.

Example 25.2. We now consider the system:



The individual components $1, 2, 3$ act independently and have exponentially distributed lifetimes, and we want to compute the mean lifetime of the overall system. "Freezing" the system in time, view each component i as a relay switch which works with probability p_i. From Chapter 1, we almost immediately conclude that

$$(1 - (1 - p_1)(1 - p_2))p_3$$

is the prob that the overall "frozen system" works. That is, the reliability function of our system is

$$R(t) = (1 - (1 - p_1)(1 - p_2))p_3.$$

The mean lifetime of the A to B connection is therefore

$$\begin{aligned} E[T_{AB}] &= \int_0^\infty R(t)dt \\ &= \int_0^\infty (1 - (1 - p_1)(1 - p_2))p_3 dt \\ &= \int_0^\infty (1 - (1 - e^{-a_1})(1 - e^{-a_2}))e^{-a_3} dt \end{aligned}$$

Here is a Matlab script that computes the mean lifetime of the A to B connection, where we assume that compoents $1, 2, 3$ have mean lifetimes $100, 200, 300$ (hours), respectively.

```
syms t
a1=1/100; a2=1/200; a3=1/300;
p1=exp(-a1*t); p2=exp(-a2*t); p3=exp(-a3*t);
R=(1-(1-p1)*(1-p2))*p3; %the "reliability function"
lifetimeAB = int(R,0,inf)
lifetimeAB =
1545/11
```

We see that the mean lifetime of the A to B connection is $1545/11 = 140.4545$ hours.

25.2 Linear Transformation of Corr/Cov Matrices

Suppose we have a linear transformation

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = A \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}, \tag{25.3}$$

where

- $X_1, X_2, \cdots, X_n$ are given RV's (the "old" RV's).
- A is the $n \times n$ linear transformation matrix.
- $Y_1, Y_2, \cdots, Y_n$ are the "new" RV's resulting from applying the linear transformation matrix A to the X_i's.

We suppose that the correlation and covariances between pairs of X_i's are known. We want to compute correlation and covariances between pairs of Y_i's. We already know one way to do this which is rather tedious: Using bilinearity properties explained in Section 20.3, we can compute each separate $Cov(Y_i, Y_j)$ as a linear combination of covariances of the X_i's. In this section, we show how matrices may be used to compute all the covariances $Cov(Y_i, Y_j)$ simultaneously via a matrix multiplication. Specifically, we will prove the following result.

Useful Result: Let

$$C_Y = [Cov(Y_i, Y_j)]$$

and

$$C_X = [Cov(X_i, X_j)]$$

be the $n \times n$ covariance matrices of the Y RV's and the X RV's, respectively. Then:

$$C_Y = AC_XA^T. \tag{25.4}$$

Furthermore, let

$$R_Y = [E(Y_iY_j)]$$

and

$$R_X = [E(X_iX_j)]$$

be the $n \times n$ correlation matrices of the Y RV's and the X RV's, respectively. Then:

$$R_Y = AR_YA^T. \tag{25.5}$$

Proof. I prove (25.5). (The proof of (25.4) is similar.) In (25.3), let's let Y be the column vector of the Y_i's and let's let X be the column vector of the X_i's. We can then rewrite (25.3) in more compact form as

$$Y = AX.$$

Notice that

$$YY^T = [Y_iY_j].$$

Let us define the expected value of a square array of RV's to be what we get when we take the expected value of each individual RV in the array. Then:

$$E[YY^T] = [E(Y_iY_j)] = R_Y.$$

Notice that

$$YY^T = (AX)(AX)^T = (AX)(X^T A^T) = A(XX^T)A^T,$$

where we used the fact that the transpose of a product of matrices is the same thing as the product of the separate transposes in the reverse order. Then:

$$R_Y = E[YY^T] = E[A(XX^T)A^T] = AE[XX^T]A^T = AR_X A^T,$$

completing the proof of (25.5). In the preceding manipulations, the operation $E[A(XX^T)A^T] = AE[XX^T]A^T$ was legitimate because the entries of A and A^T are constants and so the expectation operator E can be pulled inside these two matrices.

25.2.1 Extension to a Constant Term

We can easily extend our "Useful Result" to treat the case in which (25.3) includes an additional term on the right hand side. Accordingly, let us write our transformation equation as

$$Y = AX + B,$$

where Y, A, X are as before, but now we have added a constant column vector B on the right hand side. Let column vectors μ_X and μ_Y be the "mean vectors"

$$\mu_X = [E(X_i)]$$

and

$$\mu_Y = [E(Y_i)],$$

respectively. Then one can prove that (see Chapter 5 of your textbook):

$$\begin{aligned} \mu_Y &= A\mu_X + B \\ C_Y &= AC_X A^T \\ R_Y &= C_Y + \mu_Y \mu_Y^T \end{aligned} \tag{25.6}$$

Example 25.3. As in Example 20.3, let X, Y be RV's such that

$$\begin{aligned} \rho_{X,Y} &= -1/2 \\ \sigma_X &= 2 \\ \sigma_Y &= 3 \end{aligned}$$

Suppose we define RV's U, V as follows:

$$\begin{aligned} U &= 3X - Y + 4 \\ V &= 5X + Y - 7 \end{aligned}$$

In Example 20.3, we computed $\sigma_{U,V}$ using the bilinearity of the two arguments of the covariance function. We now rework using the matrix method of this section. Write the preceding system of equations in matrix form:

$$\begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} 3 & -1 \\ 5 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} + \begin{bmatrix} 4 \\ -7 \end{bmatrix}. \tag{25.7}$$

The covariance matrix of X and Y is

$$\begin{bmatrix} \sigma_X^2 & \sigma_{X,Y} \\ \sigma_{X,Y} & \sigma_Y^2 \end{bmatrix} = \begin{bmatrix} 4 & -3 \\ -3 & 9 \end{bmatrix}.$$

In computing the covariance matrix of U and V, we can ignore the vector $[4, -7]^T$ in (25.7). Therefore, the covariance matrix of U and V can be computed as:

$$\begin{bmatrix} \sigma_U^2 & \sigma_{U,V} \\ \sigma_{U,V} & \sigma_V^2 \end{bmatrix} = \begin{bmatrix} 3 & -1 \\ 5 & 1 \end{bmatrix} \begin{bmatrix} 4 & -3 \\ -3 & 9 \end{bmatrix} \begin{bmatrix} 3 & 5 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 63 & 57 \\ 57 & 79 \end{bmatrix}.$$

From the preceding, we see that $\sigma_{U,V} = 57$. This agrees with the answer we found in Example 20.3. Let us now go further and compute the correlation matrix of U and V. To do this, we need to know the means of X and Y. Let's take the means of X and Y to each be 1. Then the means of U and V are computed by:

$$\begin{bmatrix} \mu_U \\ \mu_V \end{bmatrix} = \begin{bmatrix} 3 & -1 \\ 5 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 4 \\ -7 \end{bmatrix} = \begin{bmatrix} 6 \\ -1 \end{bmatrix}.$$

(To obtain this, we just replaced X and Y in (25.7) with their means 1 and 1.) Using formula (25.6), we obtain the correlation matrix of U and V as follows:

$$\begin{aligned} \begin{bmatrix} E[U^2] & E[UV] \\ E[UV] & E[V^2] \end{bmatrix} &= \begin{bmatrix} \sigma_U^2 & \sigma_{U,V} \\ \sigma_{U,V} & \sigma_V^2 \end{bmatrix} + \begin{bmatrix} 6 \\ -1 \end{bmatrix} \begin{bmatrix} 6 & -1 \end{bmatrix} \\ &= \begin{bmatrix} 63 & 57 \\ 57 & 79 \end{bmatrix} + \begin{bmatrix} 36 & -6 \\ -6 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 99 & 51 \\ 51 & 80 \end{bmatrix} \end{aligned}$$

25.3 Multivariate Densities

Suppose you have RV's $X_1, X_2, \cdots, X_n$ that are jointly continuously distributed. Then joint probability calculations for these RV's would be done with their *multivariate density function*

$f(x_1, x_2, \cdots, x_n)$, which is a nonnegative function of n variables that integrates to 1 over all variables:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1, x_2, \cdots, x_n)dx_1dx_2\cdots dx_n = 1. \tag{25.8}$$

Such a joint probability calculation would be of the form

$$P[(X_1, X_2, \cdots, X_n) \in E] = \iint\cdots\int_E f(x_1, x_2, \cdots, x_n)dx_1dx_2\cdots dx_n,$$

where E is some n-dimensional region.

Example 25.4. Let Σ be any $n \times n$ symmetric matrix whose eigenvalues are all positive. Then there is a unique multivariate density function of the form

$$f(x_1, x_2, \cdots, x_n) = C\exp[-(1/2)\ \vec{x}\ \Sigma^{-1}\ \vec{x}^T], \tag{25.9}$$

where $\vec{x}$ is our shorthand for the row vector

$$\vec{x} = (x_1, x_2, \cdots, x_n),$$

and where C is the positive constant which makes the n-fold integral in (25.8) equal to 1. (You can find an expression for C in your textbook on page 229; we will only rarely have to know what the precise value of C is.) Suppose we have RV's $X_1, X_2, \cdots, X_n$ jointly distributed according to the multivariate density function (25.9). Then we say that these RV's have a *multivariate Gaussian distribution* (Section 5.7 of Chapter 5). The matrix Σ used to define our multivariate density (25.9) turns out to be the covariance matrix of the X_i's:

$$C_X = [Cov(X_i, X_j)] = \Sigma.$$

Also, the means of the X_i's turn out to be zero:

$$E[X_i] = 0, \quad i = 1, 2, \cdots, n.$$

(More generally, if $\vec{\mu}$ is an n-dimensional column vector with constant entries, a multivariate density function of the form

$$C\exp[-(1/2)(\vec{x}\ -\ \vec{\mu})\Sigma^{-1}(\vec{x}\ -\ \vec{\mu})^T]$$

would satisfy

$$\vec{\mu} = \begin{bmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_n] \end{bmatrix}.$$

In our example here, the $\vec{\mu}$ is missing from the right side of (25.9) and so the means are all zero.) We will see more about multivariate Gaussian distributions in subsequent lectures.

Remark. The *bivariate Gaussian distribution* covered in Section 4.11 of your textbook can be regarded as a special case of the multivariate Gaussian distribution. If we have a random pair (X, Y) with zero means and the bivariate Gaussian distribution, then the joint density would be of the form

$$f_{X,Y}(x,y) = C \exp\left[-(1/2) \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} \sigma_X^2 & \sigma_{X,Y} \\ \sigma_{X,Y} & \sigma_Y^2 \end{pmatrix}^{-1} \begin{pmatrix} x \\ y \end{pmatrix} \right]. \tag{25.10}$$

It is interesting to compare this expression with the expression for the bivariate Gaussian density on page 191 of your textbook. First, you can check that

$$\begin{pmatrix} \sigma_X^2 & \sigma_{X,Y} \\ \sigma_{X,Y} & \sigma_Y^2 \end{pmatrix}^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} \frac{1}{\sigma_X^2} & \frac{-\rho}{\sigma_X \sigma_Y} \\ \frac{-\rho}{\sigma_X \sigma_Y} & \frac{1}{\sigma_Y^2} \end{pmatrix}.$$

(A matrix times its inverse should be the identiy matrix.) Plugging in for the inverse of the covariance matrix in (25.10), the density (25.10) becomes

$$f_{X,Y}(x,y) = C \exp\left[\frac{-1}{2(1-\rho^2)} \left(\left\{ \frac{x}{\sigma_X} \right\}^2 - 2\rho \left\{ \frac{x}{\sigma_X} \right\} \left\{ \frac{y}{\sigma_Y} \right\} + \left\{ \frac{y}{\sigma_Y} \right\}^2 \right) \right].$$

This is precisely the bivariate Gaussian density with zero means given on page 191.

25.4 Preview

In our next few lectures, we will be getting into the *statistics* part of EE 3025. This coverage includes the *central limit theorem* (CLT), the *law of large numbers* (LLN), and *design of confidence intervals*. Let me give you a brief preview of what these topics are about.

Suppose we have a given probability distribution with mean μ and variance σ^2. (This is called our *sampling distribution*.) A sequence n RV's

$$X_1, X_2, \cdots, X_n$$

is called *a sample of size n* from our sampling distribution if the RV's are independent and if the distribution of each X_i is the sampling distribution.

Central Limit Theorem

The CLT says that, no matter what the sampling distribution is, the "normalized sum"

$$\frac{(X_1 + X_2 + \cdots + X_n) - n\mu}{\sqrt{n}\sigma}$$

has approximately a Gaussian$(0,1)$ distribution (standard Gaussian distribution) if n is large. (The approximation becomes precise in the limit as $n \to \infty$.)

Law of Large Numbers

Let $\bar{X}$ be the sample mean of our sample of size n, defined by

$$\bar{X} \triangleq \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

The LLN says that, no matter what the sampling distribution is, the event

$$\{\mu - \epsilon \leq \bar{X} \leq \mu + \epsilon\}$$

has probability close to 1 if n is large, where ϵ is any positive number that you select in advance (you can choose ϵ ahead of time to be as close to zero as you like). Moreover, this probability becomes 1 in the limit as $n \to \infty$:

$$\lim_{n\to\infty} P[\mu - \epsilon \leq \bar{X} \leq \mu + \epsilon] = 1.$$

Confidence Interval Design

Assume the mean μ of our sampling distribution is unknown. The purpose of confidence interval design is to find a sample size n for which

$$P[\bar{X} - \epsilon \leq \mu \leq \bar{X} + \epsilon] \geq p,$$

where $\epsilon > 0$ and $p < 1$ are chosen in advance. For example, you might choose ϵ to be something like 0.05 or 0.01 or 0.005. You might choose p to be something like 0.90 or 0.95. If you take $p = 0.90$, then you've achieved

$$P[\bar{X} - \epsilon \leq \mu \leq \bar{X} + \epsilon] \geq 0.90$$

and the interval

$$[\bar{X} - \epsilon, \bar{X} + \epsilon]$$

is called a *90% confidence interval for* μ. On the other hand, if you take $p = 0.95$, then you've achieved

$$P[\bar{X} - \epsilon \leq \mu \leq \bar{X} + \epsilon] \geq 0.95$$

and the interval

$$[\bar{X} - \epsilon, \bar{X} + \epsilon]$$

is called a *95% confidence interval for* μ.

The three topics CLT, LLN, and confidence interval design are linked as follows: The CLT implies that the LLN is true, and the LLN implies that any desired confidence interval will exist.

In our next lectures, you will be presented with more detailed information concerning the CLT, LLN, and confidence interval design.