Lectures on EE 3025

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

Lecture 1

Chapter 1 Part 1

In Lecture 1, I gave some motivating examples concerning potential practical uses of probability. I also introduced the concepts of *random experiment*, *sample space*, and *events*.

1.1 Motivating Examples

EE 3025 helps you make decisions in the face of uncertainty caused by randomness. For an engineer, these decisions would typically be design decisions. You design some component of an engineering system based on some sort of probability model which governs how the system works.

There are hundreds (thousands?) of potential applications of probability. I give you some examples of some of these applications falling into the following general categories:

- estimation
- control
- prediction
- quality control
- reliability testing

1.1.1 Estimation

A frequent scenario in which estimation problems arise is when you are transmitting some information-bearing signal through a communication system. The following block diagram illustrates such a system:

$$x(t) \rightarrow \boxed{\begin{array}{c}\text{system}\\ h(t)\end{array}} \rightarrow y(t) = x(t) * h(t) + n(t) \rightarrow \boxed{\begin{array}{c}\text{estimation}\\ \text{filter}\\ g(t)\end{array}} \rightarrow \hat{x}(t)$$

In this diagram, $x(t)$ represents the information-bearing signal that is to be transmitted. This signal is passed through a fixed LTI system with impulse response $h(t)$. At the system output, a random noise signal $n(t)$ is added on. An estimation filter with impulse response $g(t)$ is to be designed so that its output signal $\hat{x}(t)$ is highly likely to be close to the transmitted signal $x(t)$. Later on in EE 3025, I will show you how to design some estimation filters. In order to design an estimation filter, you would need a probability model that tells you how likely it is that each different possibility for $n(t)$ will occur. As a simple example, suppose $n(t)$ is equally likely to be any one of three specific signals $n_1(t)$, $n_2(t)$, $n_3(t)$. Then you would combine these three signals $n_1(t)$, $n_2(t)$, $n_3(t)$ in some way to obtain the impulse response $g(t)$ of the estimation filter.

Discussion. In EE 3015, you maybe got the wrong impression that when you apply a deterministic input signal $x(t)$ as input to a linear system, then the output signal is also deterministic. In electronic systems (and other systems), this may not be the case: the system response to a deterministic input signal may be a *random signal*. In other words, if you apply one fixed input signal as input to a system several times, *you may get a different output signal each time*. This might be because the system generates internal random noise (typically called *ambient noise*). In the above block diagram, the system is generating internal random noise $n(t)$; this randomly generated signal $n(t)$ will appear as an additive component of the output signal—it will appear at the output *even if the input to the system is zero*!

1.1.2 Control

Suppose you again have a system generating internal noise like in our previous block diagram:

$$x(t) \rightarrow \boxed{\begin{array}{c}\text{system}\\ h(t)\end{array}} \rightarrow y(t) = x(t) * h(t) + n(t)$$

However, unlike the estimation problem considered earlier, we are going to consider a control problem. We suppose that we want the system output $y(t)$ to be some specific signal $y^*(t)$. (One possible reason for this may be that you want the signal $y^*(t)$ to be a driving signal for some other system such as a piece of machinery.) A control system design engineer would attempt to find some choice $x(t) = x^*(t)$ for the input signal so that the output signal $y(t)$ is highly likely to be close to the desired output $y^*(t)$. In order to properly generate a signal $x^*(t)$ that will do the job, the control engineer could make use of a probability model for the randomly generated internal noise signal $n(t)$. In practice, the control engineer might insert a feedback loop with a filter in it in order to help generate the desired input $x(t)$ by using a filtered form of the output—this yields what is called a *feedback stochastic control system*. Unfortunately, the design of a feedback control system can be quite complicated. Control engineers have to use a whole bunch of specialized tricks that I

guess it would not be possible to tell you about in EE 3025. But at least you are now aware of the potential uses of probability in stochastic control. There is a senior level elective course that you could take in control systems.

Remark. The word *stochastic* means the same thing as *random*. We will see this terminology again in the last 5 weeks of EE 3025. We can talk about *stochastic processes* or *random processes*. They are the same thing.

1.1.3 Prediction

There are many applications in which prediction is important. I will briefly discuss *stock market prediction*. Suppose the daily price of a share of your favorite company's stock is observed over N consecutive days:

$$x_1, x_2, \cdots, x_N$$

Since day $N+1$ has not yet occurred, the stock price x_{N+1} for that day is not yet known and therefore it must be modeled as a random quantity. One can attempt to build a prediction $\hat{x}_{N+1}$ of what x_{N+1} will be based upon the observations from the previous k days as follows:

$$\hat{x}_{N+1} = \frac{x_N + x_{N-1} + x_{N-2} + \cdots + x_{N-k+1}}{k}$$

Principles learned in EE 3025 can tell us which of the following three values of k will give us the best prediction:

- $k = 1$
- $k = 10$
- $k = 100$

In order to determine the best prediction method, you'd have to use a probability model for the different possibilities for x_{N+1}. The best prediction $\hat{x}_{N+1}$ would be the one for which $\hat{x}_{N+1}$ is most likely to be close to x_{N+1}.

Discussion. I gave three examples of possible prediction methods above. All three of these employed a simple arithmetic average of observed stock prices as the prediction for the future stock price. It could turn out that the best thing to do is to take a *weighted average*: The most recently observed stock price x_N might receive the highest weight, with the weights decreasing as you move further into the past. For example, here is such a weighted average used for prediction, based on the four most recent stock price observations:

$$\hat{x}_{N+1} = (0.4)x_N + (0.3)x_{N-1} + (0.2)x_{N-2} + (0.1)x_{N-3}.$$

We will encounter probability models later on in EE 3025 where the best prediction will be a weighted average with decreasing weights, such as the example just given.

Here are some other applications of prediction:

Weather Prediction: It is hard to predict weather more than a few days in advance. With weather satellites, one can make meteorological observations within each grid cell of the entire planet Earth partitioned into a grid of cells. With grid computing methods, one can make good weather predictions. The grid computing method you would use would be dictated by a probability model for the meteorological observations as a function of the grid cell location and time.

Fire Control: Suppose you are a fighter jet pilot. You spot an enemy aircraft taking evasive action. Your plane's computer has to make an effective prediction of where the enemy aircraft will be Δt seconds in the future, where Δt is the length of time it would take for a missile fired by you to reach this target.

Data Compression: You have a big data file. You want to store it by representing each character in the file with just a small number of bits. If you scan the file characters in raster scan order (left to right and top to bottom), then you can make a good prediction of what the next character will be based upon the previously scanned characters. Using this prediction, the next character can be represented with as few a number of bits as possible. (This data compression technique is called *arithmetic coding*. I teach it in EE 5585.)

1.1.4 Quality Control

A company manufactures N items where N is very large. The quality control engineer must extract a sample of n of these items, where n is small relative to N. The items in the sample are tested to see what fraction of them are defective. It is desired that n be chosen so that the fraction of defectives in the sample will be a good indicator of the fraction of defectives in the entire set of N items. (If n is chosen properly, then upon repeated *random* extraction of a sample of size n, it will be found that the fraction of defectives in the sample will fluctuate only a little bit about the fixed fraction of defectives in the entire set of N items.) Principles you learn in EE 3025 will help you handle quality control problems. Here's an example. Suppose you test the sample and find 1% of the items to be defective. You'd be able to answer the following two questions:

Question 1: How likely is it that the percentage of defectives in the entire set of items is close to 1%?

Question 2: If it is not very likely that the percentage of defectives in the entire set of items is close to 1%, how many more items should be sampled in order to be more sure that the percentage of defectives in the sample is highly likely to be close to the percentage of defectives in the entire set?

There are other applications in which one can see what to do by analogy with the quality control problem. One of these applications is *polling*. For example, suppose you are a pollster and you poll a few hundred potential voters, finding that 49% of them are in favor of political candidate A.

Among all potential voters in the United States, you might then be highly confident that $49 \pm p$ percent of them are in favor of candidate A. For example, if $p = 4$, you'd be saying that you're highly confident that between 45 and 53 percent of all potential voters in the United States are in favor of Candidate A. A statistical technique you learn later on in EE 3025 will enable you to determine p. This technique is called *confidence interval estimation*. (Remark: In the leadup to our recent presidential election, some pollsters only expressed confidence in their results to within an 8 point spread like I've just described. Since the two candidates were so close in preference among the general population, these pollsters produced worthless results, due to the fact that they used too small a sample size.)

1.1.5 Reliability Testing

Here's an example of reliability testing. Suppose you want to test a particular integrated circuit chip in order to see that it is doing what it was designed to do. This chip has a certain number of binary input terminals and a certain number of binary output terminals. Suppose the number of input terminals is N, where N is large. Then the number of different possible sets of inputs is 2^N, and 2^N is extremely large. It would be impossible to test all of these 2^N possibilities in order to see that the chip is performing satisfactorily in each case. Alternatively, you could do the following: you could select the binary input at each of the N input terminals randomly to be either 0 or 1; for this randomly selected set of inputs, you could then see whether the chip works OK. If the mechanism for randomly selecting the inputs is chosen appropriately, one would be able to say that the chip will work well for other sets of inputs once the chip is seen to work OK for the randomly selected inputs (without actually checking these other sets of inputs). I was asked in class to provide an example of a mechanism for randomly selecting the inputs. Here is a trivial way to do it: Flip N fair coins, and for each coin write down a 0 or 1 depending on whether that coin comes up heads or tails; each coin in this way determines a binary input for one of the N terminals. In practice, the particular random input selection mechanism I have just described is too simplistic. The way you'd do it in practice would depend upon the internal machanism of the chip.

1.2 Some Basic Concepts

1.2.1 Random Experiments

A *random experiment* is an experiment whose outcome is not known in advance of performing the experiment. We require that a random experiment be *reproducible*, that is, we should be able to perform the experiment over and over again under identical conditions (these separate performances are called *independent trials*). If the experiment is reproducible, we can gain information about how likely each possible outcome is by performing the experiment a large number of times.

Example. Here is an example of a nonreproducible experiment. You discover a new type of nuclear bomb which when exploded will destroy the entire planet Earth. The outcome of the experiment would be to measure how far out from planet Earth the radius of destruction of the bomb extends. (This example is not as farfetched as it sounds. When the first atomic bomb was about to be exploded in the New Mexico desert in the 1940's, the physicists who designed the bomb didn't know whether that bomb would destroy the entire planet Earth. This fact, which horrified me, is pointed out in an interesting book I read about the history of the Manhattan Project.)

The next section gives examples of reproducible experiments.

1.2.2 Sample Space

The *sample space* S of a random experiment is defined to be the set of all possible outcomes of the random experiment.

Example 1.1. The random experiment is "Flip a coin and see whether you get heads or tails". The sample space is

$$S = \{H, T\}.$$

Example 1.2. The random experiment is "Flip three coins and see what each of them comes up as". The sample space can be taken to be

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

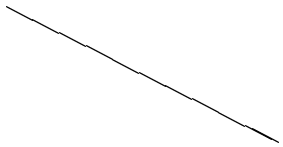
Each outcome is of the form

$$(H \text{ or } T, H \text{ or } T, H \text{ or } T).$$

We have assumed that the three coins are distinguishable (Coin 1, Coin 2, Coin 3) and that the preceding 3-tuple represents the outcome for Coins 1,2,3 in succession. There were 8 outcomes in the sample space because

$$2 * 2 * 2 = 8.$$

We can also obtain these 8 outcomes by using the following tree:



The H, T branches at the top of the tree denote the result obtained for Coin 1. The H, T branches in the middle of the tree denote the result obtained for Coin 2. The H, T branches at the bottom of the tree denote the result obtained for Coin 3. There are 8 different root-to-leaf paths you can follow in this tree. If you write down the sequence of H's and T's obtained along each such path, you obtain the 8 outcomes in S which appear as labels on the leaves at the bottom of the tree.

Example 1.3. The random experiment is "Flip a pair of dies and see what number comes up on each die". We can take the sample space to be

$$\begin{aligned} S \quad = \{ \quad & (1,1),(1,2),(1,3),(1,4),(1,5),(1,6), \\ & (2,1),(2,2),(2,3),(2,4),(2,5),(2,6), \\ & (3,1),(3,2),(3,3),(3,4),(3,5),(3,6), \\ & (4,1),(4,2),(4,3),(4,4),(4,5),(4,6), \\ & (5,1),(5,2),(5,3),(5,4),(5,5),(5,6), \\ & (6,1),(6,2),(6,3),(6,4),(6,5),(6,6) \quad \} \end{aligned}$$

The outcomes are regarded as pairs (i, j) where i, j both range between 1 and 6 independently of each other. We are assuming distinguishable dies (Die 1, Die 2). The entry i in (i, j) represents the number coming up on Die 1 and the entry j denotes the number coming up on Die 2. There are 36 outcomes because

$$6 * 6 = 36.$$

The reader can also draw a tree for obtaining these 36 outcomes. However, the tree is pretty big!

1.2.3 Events

An event can be any subset of the sample space S. It can either be described verbally or it can be specified by listing all of the outcomes in it.

Example 1.4. For the "three coin flip" experiment, we describe an event E verbally as

$$E = \{exactly\ two\ heads\ occur\}.$$

We can rewrite E as

$$E = \{HHT, HTH, THH\}.$$

There are three outcomes in E (three ways for event E to occur when you perform the experiment).

Example 1.5. For the "two die flip" experiment, we describe an event E verbally as

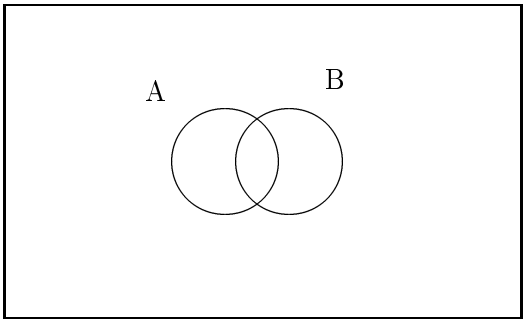
$$E = \{total\ on\ dies\ is\ 7\}.$$

We can rewrite E as

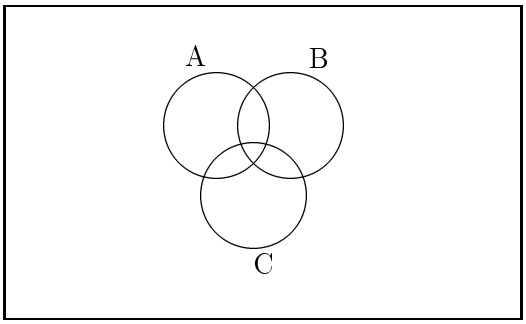
$$E = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}.$$

There are six outcomes in E (six ways for event E to occur when you perform the experiment).

We can pictorially represent events using *Venn Diagrams*. The following Venn Diagram represents two events A, B as the interiors of the two circles; the entire rectangular box denotes the sample space S.



Here is what happens to the Venn Diagram when we put in a third event C:



Lecture 2

Chapter 1 Part 2

In Lecture 2, I talked about the *calculus of events*, and started material on *probability models*.

2.1 Calculus of Events

You can combine events to get other events using the three operations of union $\cup$, intersection $\cap$, and complementation ^{c}:

$$\begin{aligned} \cup_i E_i &= \{\omega \in S : \omega \in E_i \; for \; at \; least \; one \; i\} \\ \cap_i E_i &= \{\omega \in S : \omega \in E_i \; for \; all \; i\} \\ E^c &= \{\omega \in S : \omega \notin E\} \end{aligned}$$

If you look on pages 4-5 of your textbook, you will find examples of illustrations of these operations using Venn Diagrams.

We say that a given event E *occurs* on a given performance (trial) of the random experiment if the observed outcome ω belongs to E. With this in mind, we can say the following:

- The union event $\cup_i E_i$ occurs if and only if event E_i occurs for at least one i.
- The intersection event $\cap_i E_i$ occurs if and only if event E_i occurs for all i.
- The complementary event E^c occurs if and only if the event E does not occur.

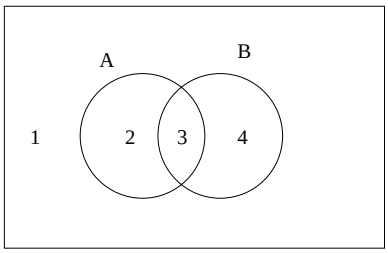
"Exclusive or" Event. You can think of the event $A \cup B$ as meaning "A or B or both". That is, the "or" is the "inclusive or" which includes the possibility that both A, B occur. Sometimes we want the "exclusive or" event, meaning that exactly one of the events A, B occurs. This is commonly written as

$$A \Delta B.$$

In terms of our three operations, we can rewrite this as

$$A\Delta B = (A \cap B^c) \cup (B \cap A^c).$$

In the Venn Diagram below, the exclusive or event $A\Delta B$ consists of Region 2 together with Region 4.



Difference Event. Another event that pops up a lot is the event $A - B$, which consists of the outcomes in A with the outcomes in B taken away:

$$A - B = \{\omega \in S : \omega \in A, \omega \notin B\}.$$

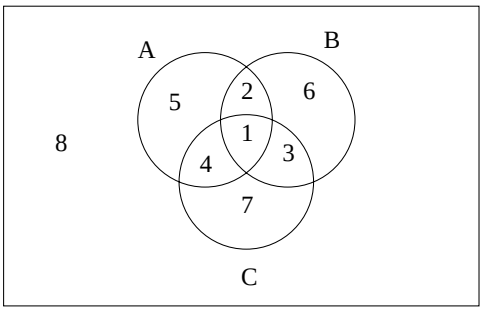
This is the same thing as

$$A - B = A \cap B^c,$$

which is Region 2 in the above Venn Diagram.

Useful Fact: Given k events, there are exactly $2^{(2^k)}$ events that arise from these events by means of the three operations $\cup$, $\cap$, ^{c}.

Example 2.1. Given events A, B, C as in the Venn Diagram below. By the preceding "Useful Fact," there should be $2^{(2^3)} = 256$ events we can build up from these. I explain how to do this.



First, note that the sample space, represented by the rectangular box in the Venn Diagram, is partitioned into the eight events labeled 1 through 8. These events are represented in terms of A, B, C as follows:

$$\begin{aligned} 1 &= A \cap B \cap C \\ 2 &= A \cap B \cap C^c \\ 3 &= B \cap C \cap A^c \\ 4 &= A \cap C \cap B^c \\ 5 &= A \cap B^c \cap C^c \\ 6 &= B \cap A^c \cap C^c \\ 7 &= C \cap A^c \cap B^c \\ 8 &= A^c \cap B^c \cap C^c \end{aligned}$$

These eight events give rise to $2^8 = 256$ events by choosing every possible subset of the eight events (there are $2^8 = 256$ subsets of a set of size 8) and taking the unions of the events in each subset. Some of these 256 events may not be very interesting, but some of them are. Here are a couple of interesting ones:

$$\begin{aligned} 2 \cup 3 \cup 4 &= \{exactly\ two\ of\ A, B, C\ occur\} \\ 5 \cup 6 \cup 7 &= \{exactly\ one\ of\ A, B, C\ occur\} \end{aligned}$$

2.1.1 Laws About Events

Sometimes an event can be computed using the three operations $\cup$,$\cap$, ^{c} in two different ways. In such a case, we have a law expressing equality between two event formation methods. There are a lot of these laws. You may have gone over sum of them in calculus. Here are a few:

$$\begin{aligned} A \cap (\cup_i E_i) &= \cup_i (A \cap E_i) \\ A \cup (\cap_i E_i) &= \cap_i (A \cup E_i) \\ (\cup_i E_i)^c &= \cap_i E_i^c \\ (\cap_i E_i)^c &= \cup_i E_i^c \\ (E^c)^c &= E \end{aligned}$$

The first two laws are *distributive laws*. The third and fourth laws are *DeMorgan's Laws*. The third law is perhaps the most important of these five laws. It says that $\cup_i E_i$ does not occur if and only if all of the events E_i do not occur. This is pretty clear if you realize that $\cup_i E_i$ is the event that at least one of the E_i's occur:

$$\{at\ least\ one\}^c = \{none\}.$$

If you have just two or three events, it's pretty easy to demonstrate the truth of a law by referring to a Venn Diagram. The following example illustrates this technique.

Example 2.2. Suppose we want to prove that

$$A \cap (B \cup C) = (A \cap C) \cup (A \cap B). \tag{2.1}$$

Refer to the preceding Venn Diagram. Event $B \cup C$ consists of regions $1, 2, 3, 4, 6, 7$. Event A consists of regions $1, 2, 4, 5$. Intersecting these, we get regions $1, 2, 4$, which is the left side of equation (2.1). By similar reasoning, we leave it to the reader to show that the right side of (2.1) also consists of regions $1, 2, 4$.

2.2 Definition of Probability Model

- We say that a sequence of events $\{E_i\}$ is *mutually exclusive* if

  $$E_i \cap E_j = \phi, \quad i \neq j,$$

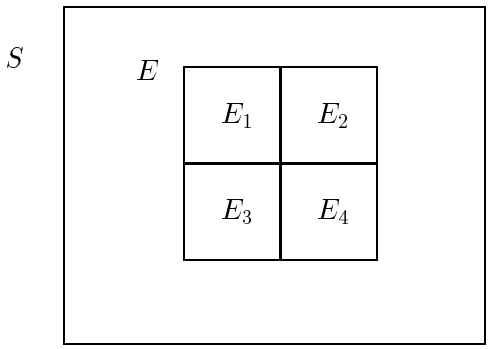
  where ϕ is the empty set. In other words, no two of the events have any outcomes in common. In particular, if we have two events A, B, we say that A, B are mutually exclusive if and only if

  $$A \cap B = \phi.$$

- We say that event E is the *disjoint union* of the sequence of events $\{E_i\}$ if

  $$E = \cup_i E_i$$

  and the events $\{E_i\}$ are mutually exclusive. For example, in the Venn Diagram below, event E is the disjoint union of events E_1, E_2, E_3, E_4.



We obtain a *probability model* for a given random experiment by assigning to each event E a number $P(E)$ (which we call the probability of the event E) so that the following axioms are satsisfied:

Axiom 1: $0 \leq P(E) \leq 1$ for every E.

Axiom 2: $P(S) = 1$.

Axiom 3: Anytime an event E is a disjoint union of events $\{E_i\}$, we must have

$$P(E) = \sum_i P(E_i).$$

Axiom 3 says probabilities act like areas: "the probability (area) of the whole is the sum of the probabilities (areas) of the parts." Intuitively, the probability of an event should reflect our feeling about how likely that event will occur upon repeated trials. For example, if $P(E) = 1/2$ is the assigned probability, this is probably because we expect that E will occur on roughly one half of a large number of trials. We discuss this intuitive notion of probability further later in this section.

2.3 Types of Probability Models

Discrete Probability Models

In a discrete probability model, the sample space consists of a finite or infinite sequence of outcomes:

$$S = \{\omega_1, \omega_2, \omega_3, \cdots\}.$$

The discrete probability model is completely and uniquely specified once you assign a nonnegative probability $P(\omega_i)$ to each outcome ω_i so that the probabilities of the outcomes add up to 1:

$$\sum_i P(\omega_i) = 1.$$

In fact, the probability $P(E)$ of any event E is then uniquely computable via the formula

$$P(E) = \sum_{\omega_i \in E} P(\omega_i),$$

by Axiom 3.

Example 2.3. Consider the discrete probability model in which the outcomes are the positive integers

$$S = \{1, 2, 3, \cdots\},$$

and in which the probabilities of the outcomes are given by the formula

$$P(i) = 2^{-i}, \quad i = 1, 2, 3, \cdots.$$

For this to be a legitimate probability model, you just have to show that

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots = 1.$$

(Can you do that?) Let us compute the probability of getting an odd number:

$$P(\{1,3,5,7,\cdots\}) = \frac{1}{2} + \frac{1}{2^3} + \frac{1}{2^5} + \frac{1}{2^7} + \cdots. \tag{2.2}$$

Recall how to sum a geometric series:

$$a + ar + ar^2 + ar^3 + \cdots = \frac{a}{1-r},$$

if the modulus of the ratio r is less than one. Applying this summation formula to (2.2),

$$P(\{1,3,5,7,\cdots\}) = \frac{(1/2)}{1-(1/4)} = 2/3.$$

Exercise. In Example 2.3, use Matlab to compute the probability of getting a prime number (to four decimal places). That is, compute

$$\frac{1}{2} + \frac{1}{2^3} + \frac{1}{2^5} + \frac{1}{2^7} + \frac{1}{2^{11}} + \frac{1}{2^{13}} + \frac{1}{2^{17}} + \cdots.$$

Remark. The reader may be wondering what random experiments would yield the probability model Example 2.3. Here is one such random experiment:

"Flip a fair coin until you get heads for the first time; count the number of flips required."

Can you satisfy yourself that this experiment does indeed give us our probability model?

Equiprobable Probability Models

In an equiprobable probability model, there are finitely many outcomes in the sample space, and they are "equally likely," that is, they are all assigned the same probability. This requirement leads to a unique probability model. To see this, suppose there are k outcomes in the sample space S. The k probabilities of these outcomes must add up to one and they are equal. This forces each of these probabilities to be equal to $1/k$. We have proved that for an equiprobable probability model, the probability of each outcome is equal to the reciprocal of the number of outcomes in S. It is then easy to prove that the probability of any event E can be computed via the formula

$$P(E) = \frac{number\ of\ outcomes\ in\ E}{number\ of\ outcomes\ in\ S}. \tag{2.3}$$

(The outcomes in E all have the same probability and so in order to sum up these probabilities, you can simply multiply the probability of any one of these outcomes by the number of outcomes in E.)

Example 2.4. Go back to the "three coin flip" experiment of Lecture 1. Suppose the coins are all fair. This is the tipoff that the resulting probability model will be an equiprobable one. We determined earlier that the sample space is

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

There are 8 of these outcomes so they each have probability equal to 1/8. We can now easily compute the probability of any event connected with this experiment simply by dividing the number of outcomes in the event by 8. We can therefore say that

$$\begin{aligned} P(\textit{three heads occur}) &= P(HHH) = \frac{1}{8}, \\ P(\textit{exactly two heads occur}) &= P(HHT, HTH, THH) = \frac{3}{8}, \\ P(\textit{exactly one head occurs}) &= P(TTH, THT, HTT) = \frac{3}{8}, \\ P(\textit{no heads occur}) &= P(TTT) = \frac{1}{8}. \end{aligned}$$

Exercise. Flip two fair coins. Prove that the probability of getting a total of seven is 1/6.

Warning. If you do not have an equiprobable probability model, do not use formula (2.3) to compute probabilities. It won't be valid!

Independent Discrete Probability Models

In an independent discrete probability model, each outcome can be written in the form of a k-tuple

$$(x_1, x_2, \cdots, x_k)$$

for some positive integer k, where the separate entries x_i are chosen independently of each other (they arise from k independent trials of different experiments or the same experiment). The probability of each outcome is computed as a product

$$P(x_1, x_2, \cdots, x_k) = \prod_{i=1}^{k} P_i(x_i),$$

where for each $i = 1, 2, \cdots, k$, we have a probability model P_i governing the selection of the i-th coordinate of $(x_1, x_2, \cdots, x_k)$. (If the k independent trials are all of the same experiment, then all of the P_i's are the same.)

Example 2.5. Suppose our random experiment is to flip 3 separate coins and then to record whether each coin comes up H or T. Intuitively, since the coins act independently of each other, our

probability model should be an independent probability model. The sample space is the following set of 3-tuples:

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

So far, this is just like the sample space for flipping 3 fair coins. However, to make things more interesting, we suppose that not all of the three coins are fair coins. Suppose Coin 1 is fair. The outcome for Coin 1 is then governed by the equiprobable model

$$P_1(H) = 1/2, \quad P_1(T) = 1/2.$$

We suppose that Coin 2 and Coin 3 are unfair (biased) coins governed by the respective probability models

$$P_2(H) = 1/3, \quad P_2(T) = 2/3.$$

$$P_3(H) = 2/3, \quad P_3(T) = 1/3.$$

Then the probability model for S is

$$\begin{aligned}
P(HHH) &= P_1(H)P_2(H)P_3(H) = (1/2)(1/3)(2/3) = 1/9, \\
P(HHT) &= P_1(H)P_2(H)P_3(T) = (1/2)(1/3)(1/3) = 1/18, \\
P(HTH) &= P_1(H)P_2(T)P_3(H) = (1/2)(2/3)(2/3) = 2/9, \\
P(HTT) &= P_1(H)P_2(T)P_3(T) = (1/2)(2/3)(1/3) = 1/9, \\
P(THH) &= P_1(T)P_2(H)P_3(H) = (1/2)(1/3)(2/3) = 1/9, \\
P(THT) &= P_1(T)P_2(H)P_3(T) = (1/2)(1/3)(1/3) = 1/18, \\
P(TTH) &= P_1(T)P_2(T)P_3(H) = (1/2)(2/3)(2/3) = 2/9, \\
P(TTT) &= P_1(T)P_2(T)P_3(T) = (1/2)(2/3)(1/3) = 1/9.
\end{aligned}$$

Exercise. For the preceding model, verify that the eight probabilities add up to one. Then compute the following:

$$\begin{aligned}
P(three\ heads\ occur) &= ? \\
P(exactly\ two\ heads\ occur) &= ? \\
P(exactly\ one\ head\ occurs) &= ? \\
P(no\ heads\ occur) &= ?
\end{aligned}$$

Empirical Probability Models

Suppose we have a finite sample space

$$S = \{\omega_1, \omega_2, \cdots, \omega_k\}.$$

Maybe you're unsure what probability model to take on this sample space. One can always come up with an *empirical model* based upon performing the underlying random experiment a certain number of times. Let us explain how an empirical model is obtained. Suppose we perform n independent trials of the random experiment. Then you can define a probability model as follows:

$$P(\omega_i) = \frac{number\ of\ trials\ in\ which\ \omega_i\ occurs}{n}, \quad i = 1, 2, \cdots, k.$$

That is, each probability is simply the frequency with which the given outcome occurs in the n trials. The problem with an empirical model is that it can change every time you perform the independent trials anew. If you are using a probability model as a design tool (such as in applications we discussed in Lecture 1), you naturally want to have a fixed probability model. It may be that when you look at a whole bunch of different empirical models for your random experiment, you will be able to decide upon one particular model such that all of the empirical models are approximately equal to the particular model. Then that particular probability model would be a good model to select as your fixed probability model that will be used by you now and forever (in order to make predictions, form estimates, or whatever it is that your application requires).

Example 2.6. Execute the Matlab command

```
floor(2*rand(1,10))
```

You will see printed out on your Matlab screen 10 numbers, each equal to 0 or 1. You will find that the 10 numbers you obtain will vary as you execute the preceding command over and over again. If you use 0 to represent "heads" and 1 to represent "tails", you can regard the 10 numbers you obtain as the result of 10 independent trials of the experiment

"Flip a fair coin and see if heads or tails comes up."

Suppose you get the following by executing the Matlab command:

$$0, 1, 1, 0, 1, 0, 0, 0, 1, 0.$$

We can think of this as representing the following results from flipping a fair coin 10 times:

$$H, T, T, H, T, H, H, H, T, H.$$

The resulting empirical probability model is

$$P(H) = 6/10, \quad P(T) = 4/10.$$

For a fair coin, the fixed probability model you would probably want to use is

$$P(H) = 1/2, \quad P(T) = 1/2.$$

The empirical model we got varies slightly from this. If we had taken a larger number of trials, the variation between the empirical model and the $(1/2, 1/2)$ model would have been less. Try the Matlab command

```
floor(2*rand(1,10000))
```

instead (which simulates the result of 10000 fair coin flips) and see if your resulting empirical model is closer to the $(1/2, 1/2)$ model.

In Lecture 3, I will give some more types of probability models.

Lecture 3

Chapter 1 Part 3

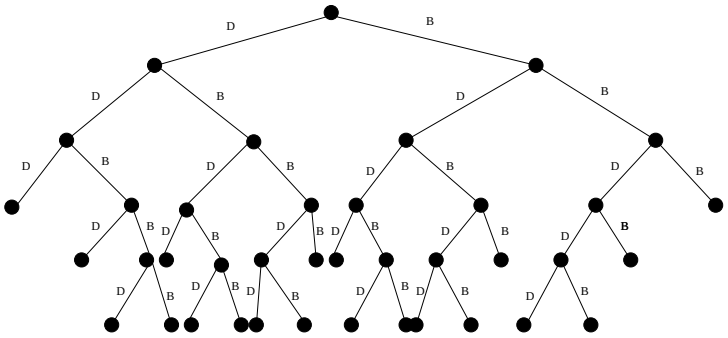
In this lecture, I covered some more types of probability models. I also gave some indications on how to do probability calculations using Venn Diagrams.

3.1 Types of Probability Models (Continued)

Tree Models

Trees can be used to represent multiple step experiments. The branches from the root node at the top of the tree represent possibilities for the first step. Branching at the next level of the tree represents the second step, etc. The branches emanating from a fixed node are given probability labels adding up to one. You obtain the different outcomes of the experiment by following each root-to-leaf path; the product of probabilities along such a path yields the probability of that particular outcome.

Example 3.1. The Dodgers and Braves play a best 3 out of 5 playoff series. (The first team to win 3 games wins the series.) We assume that the two teams are equally matched. The tree below will be used to obtain the probability model for this random experiment.



Following all paths in the tree from the root to the leaf vertices, there are 20 outcomes in the sample space S:

$$S = \{DDD, DDBD, DDBBD, DDBBB, DBDD, DBDBD,$$
$$DBDBB, DBBDD, DBBDB, DBBB, BDDD, BDDBD, BDDBB,$$
$$BDBDD, BDBDB, BDBB, BBDDD, BBDDB, BBDB, BBB\}$$

Each branch of the tree should be assigned a probability label of 1/2. Multiplying along each root-to-leaf path, we then get the following probability model for this experiment:

$P[DDD] = 1/8$	$P[DDBD] = 1/16$	$P[DDBBD] = 1/32$	$P[DDBBB] = 1/32$
$P[DBDD] = 1/16$	$P[DBDBD] = 1/32$	$P[DBDBB] = 1/32$	$P[DBBDD] = 1/32$
$P[DBBDB] = 1/32$	$P[DBBB] = 1/16$	$P[BDDD] = 1/16$	$P[BDDBD] = 1/32$
$P[BDDBB] = 1/16$	$P[BDBDD] = 1/16$	$P[BDBDB] = 1/32$	$P[BDBB] = 1/16$
$P[BBDDD] = 1/32$	$P[BBDDB] = 1/32$	$P[BBDB] = 1/16$	$P[BBB] = 1/8$

We can use this model to compute the probability of any event associated with this random experiment. For example, suppose E is the event that the series lasts exactly four games. Since

$$E = \{DDBD, DBDD, DBBB, BDDD, BDBB, BBDB\},$$

we can compute

$$\begin{aligned} P(E) &= P(DDBD) + P(DBDD) + P(DBBB) + P(BDDD) + P(BDBB) + P(BBDB) \\ &= 6(1/16) = 3/8 \end{aligned}$$

That is, if a large number of "best 3 of 5" series are played involving equally matched teams, we'd expect that about 37.5% of these series would take exactly 4 games to play.

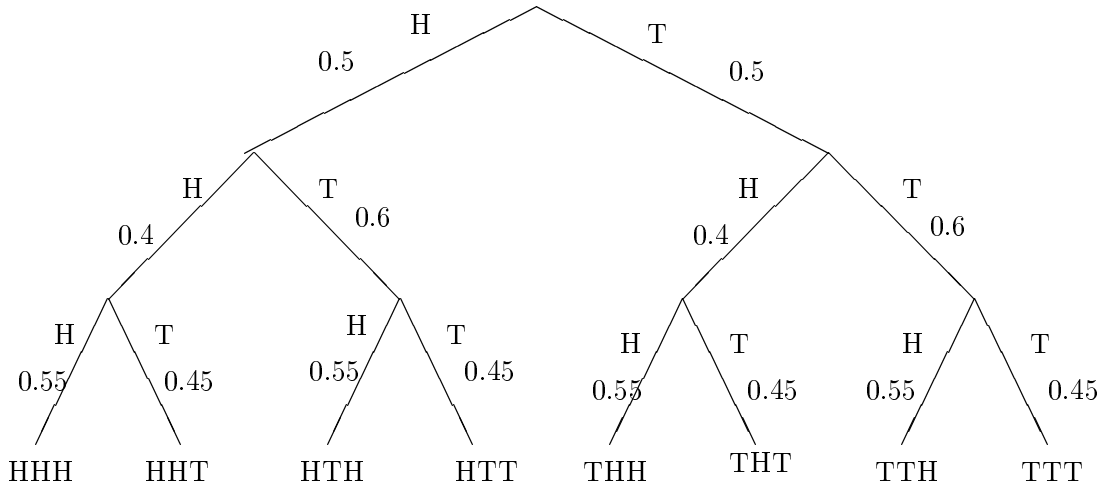
Example 3.2. In Example 3.1, let us suppose that the two teams are not equally matched. Assume that

$$P(D) = 0.6, \quad P(B) = 0.4$$

are the respective probabilities with which the Dodgers and Braves will be the winner of any game, respectively. In the tree at the bottom of the previous page, you would then assign a probability label 0.6 to any branch labelled D, and a probability label 0.4 to any branch labelled B. This would give us a different probability model than the one in Example 3.1. Using this new model, we would re-compute the probability of the series going exactly 4 games as follows:

$$\begin{aligned} P(E) &= P(DDBD) + P(DBDD) + P(DBBB) + P(BDDD) + P(BDBB) + P(BBDB) \\ &= P(D)P(D)P(B)P(D) + P(D)P(B)P(D)P(D) + \cdots + P(B)P(B)P(D)P(B) \\ &= 3(0.6)^3(0.4) + 3(0.4)^3(0.6) = 0.3744. \end{aligned}$$

Example 3.3. We have three coins. Coin 1 is a fair coin. Coin 2 has $P(H) = 0.4$. Coin 3 has $P(H) = 0.55$. Consider the following 3-step experiment. On Step 1, Coin 1 is tossed and the outcome (H or T) is recorded. On Step 2, Coin 2 is tossed and the outcome (H or T) is recorded. On Step 3, Coin 3 is tossed and the outcome (H or T) is recorded. The tree for this experiment is the following.



The sample space is

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

Multiplying along the 8 root-to-leaf paths in the tree, you get the following probability model for this experiment:

$$P(HHH) = (0.5)(0.4)(0.55) = 0.11 \tag{3.1}$$
$$P(HHT) = (0.5)(0.4)(0.45) = 0.09 \tag{3.2}$$
$$P(HTH) = (0.5)(0.6)(0.55) = 0.165 \tag{3.3}$$
$$P(HTT) = (0.5)(0.6)(0.45) = 0.135 \tag{3.4}$$
$$P(THH) = (0.5)(0.4)(0.55) = 0.11 \tag{3.5}$$
$$P(THT) = (0.5)(0.4)(0.45) = 0.09 \tag{3.6}$$
$$P(TTH) = 7(0.5)(0.6)(0.55) = 0.165 \tag{3.7}$$
$$P(TTT) = (0.5)(0.6)(0.45) = 0.135 \tag{3.8}$$

This probability model can be used to compute the probability of any event connected with this experiment. To illustrate, let us compute the probability that there are exactly two heads. Letting

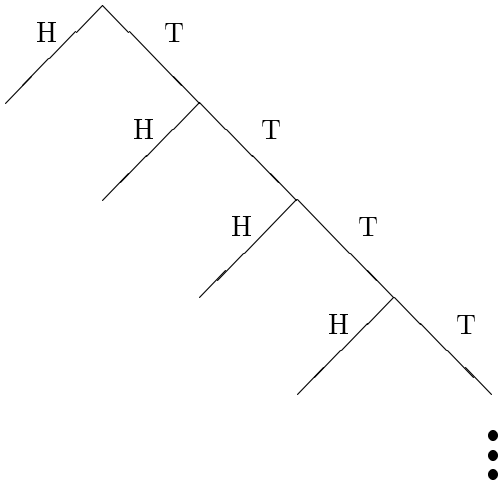
$$E = \{HHT, HTH, THH\},$$

we have

$$P(E) = P(HHT) + P(HTH) + P(THH) = 0.09 + 0.165 + 0.11 = 0.365.$$

Another way to derive the probability model of this example is to consider the model as an independent model, as in Example 2.5. Conversely, we could have attacked Example 2.5 using a tree model as we did in the present example. This just goes to show you that there may be more than one way to derive the probability model for a given random experiment.

Example 3.4. The tree used to derive a probability model may be infinite. Here is an example where this occurs. Our experiment is to keep flipping a fair coin until we first obtain a head, at which point the experiment stops.



The three dots at the bottom of the above tree for this experiment indicate that the tree keeps going on forever. The probability label on every branch is 1/2. The sample space is infinite:

$$S = \{H, TH, TTH, TTTH, TTTTH, \cdots\}.$$

We obtain the probabilities of these outcomes by multiplying along the root-to-leaf path in the tree corresponding to each outcome:

$$\begin{aligned} P(H) &= 1/2 \\ P(TH) &= (1/2)^2 \\ P(TTH) &= (1/2)^3 \\ P(TTTH) &= (1/2)^4 \\ P(TTTTH) &= (1/2)^5, \quad etc. \end{aligned}$$

Derived Models

By applying a function to the sample space, you can go from the original probability model for an experiment to a new model, which we call a *derived model*. Let the sample space for the experiment be S, which we assume to be discrete. Let X be any function mapping S into the real line. (The function X is called a *discrete random variable*.) Instead of the original experiment, in which the outcome is an element ω of the set S, we can consider a new experiment in which the outcome instead is $X(\omega)$, a point on the real line. Suppose the values of X are the real numbers

$$x_1, x_2, \cdots, x_k.$$

Then the new sample space is

$$S_{new} = \{x_1, x_2, \cdots, x_k\}.$$

The derived probability model is denoted by the notation P^X; it is a probability model defined on S_{new}. The derived model P^X is obtained from the original probability model P on S via the following formula:

$$P^X(x_i) = P[\{\omega \in S : X(\omega) = x_i\}] = \sum_{\omega \in S,\ X(\omega)=x_i} P(\omega). \tag{3.9}$$

In other words, you add up the probabilities of all outcomes ω of the original experiment which are mapped by the function X into the real value x_i.

Example 3.5. Let the original experiment be the "three coin flip" experiment of Example 2.4. The original sample space is

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

The probability of each outcome is 1/8. Let X be the function on S which counts the number of heads:

$$X(HHH) = 3, \quad X(HHT) = 2, \quad X(HTH) = 2, \quad X(HTT) = 1$$
$$X(THH) = 2, \quad X(THT) = 1, \quad X(TTH) = 1, \quad X(TTT) = 0$$

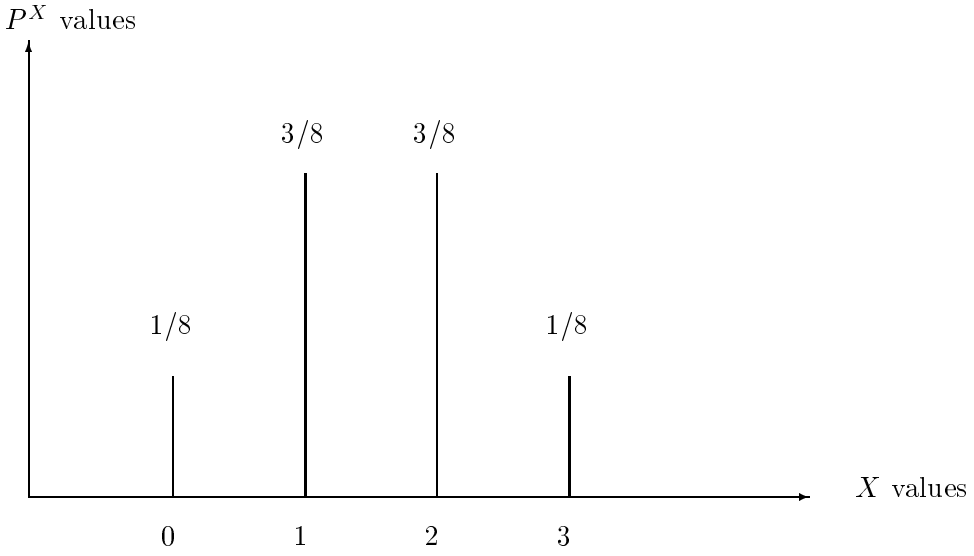
The possible values of X are therefore the real numbers $0, 1, 2, 3$. Our new sample space is

$$S_{new} = \{0, 1, 2, 3\}.$$

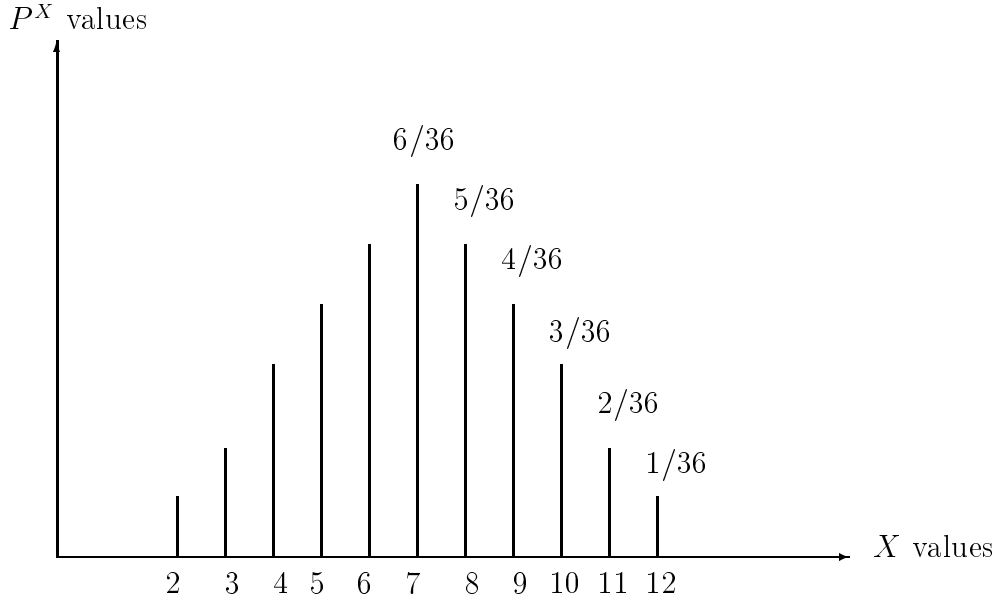
If you look at the last half of Example 2.4, you will see that what we are computing there is the probability model P^X on S_{new}, according to formula (3.9):

$$P^X(0) = 1/8, \quad P^X(1) = 3/8, \quad P^X(2) = 3/8, \quad P^X(3) = 1/8.$$

In Chapter 2, where we deal with discrete random variables, it can be convenient to represent a derived model P^X via a plot of the P^X values (vertical axis) versus the X values (horizontal axis). Here is the P^X plot you obtain for this example:



Example 3.6. Let the original experiment be the two fair die toss. The sample space S was given in Example 1.3, and consists of 36 outcomes each having probability 1/36. Let X be the total of the numbers on the two dies. The plot of the derived model is the following, and the reader should verify this result:



The largest P^X probability takes place at the central value 7 with the probabilities descending symmetrically on the two sides of the central value as you move away from the central value. Typically, when you take X to be the total in tossing finitely many fair dies, or the number of

heads in flipping finitely many fair coins, you get a P^X plot which descends symetrically on two sides of either one central value (like in this example) or two central values (like in the previous example).

Continuous Probability Models

In a continuous probability model, the outcomes of the experiment are n-tuples of real numbers

$$(x_1, x_2, \cdots, x_n)$$

for some fixed positive integer n. To compute a probability $P(E)$, where E is a subregion of the n-dimensional space consisting of all n-tuples, you would perform an n-fold integral over E of the form

$$P(E) = \iint \cdots \int_E f(x_1, x_2, \cdots, x_n) dx_1 dx_2 \cdots dx_n,$$

where $f(x_1, x_2, \cdots, x_n)$ is a nonnegative function of n variables called a *probability density function.* You will see various continuous probability models later in the course when we cover Chapters 3,4,5. For the time being, here is possibly the simplest example of a continuous probability model.

Example 3.7. We take the dimension n of the continuous probability model to be $n = 1$ (one-dimensional model). Then the outcome of the random experiment is a real number. In particular, let us consider the following random experiment: Pick a point at random from the interval of real numbers $[0, 1]$; this is the outcome of the experiment. The sample space is

$$S = [0, 1].$$

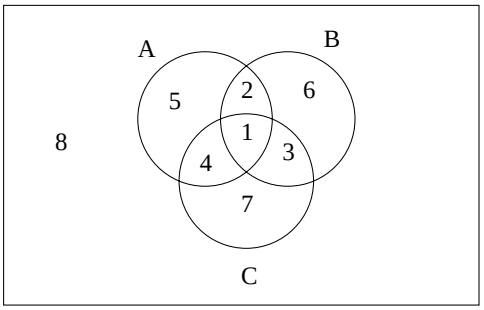
The events can be subintervals of $[0, 1]$ (or unions of them). The probability density function $f(x)$ for this experiment turns out to be the unit rectangular pulse from $x = 0$ to $x = 1$:

$$f(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & elsewhere \end{cases}$$

Let $[c, d]$ be a subinterval of the interval $[0, 1]$. We can then compute the probability $P([c, d])$ that the selected outcome falls in the interval $[c, d]$ via integration as follows:

$$P([c, d]) = \int_c^d f(x) dx = \int_c^d dx = d - c.$$

In other words, the probability assigned by this model to any subinterval of the interval $[0, 1]$ is simply the length of the subinterval. We remark that the Matlab command "`rand(1,1)`" simulates the outcome of this particular random experiment.



3.2 Probability Calculations Via Venn Diagram Reasoning

Let us consider again the general Venn Diagram for three events A, B, C given above. As we have remarked before, there are 256 events in this diagram determined by A, B, C via the operations of union, intersection, and complementation. If you are given the following 7 probabilities

$$P(A), P(B), P(C), P(A \cap B), P(A \cap C), P(B \cap C), P(A \cap B \cap C), \tag{3.10}$$

then you can compute the probability of any of these 256 events. First, by simple Venn Diagram reasoning (thinking of probabilities as areas), you can compute the probabilities

$$P(1), P(2), P(3), P(4), P(5), P(6), P(7), P(8) \tag{3.11}$$

from the probabilities (3.10). Any of the 256 events determined by A, B, C is a disjoint union of some subset of the events $1, 2, 3, 4, 5, 6, 7, 8$. Therefore, we are able to compute the probability of any of these 256 events once we have determined the probabilities (3.11). Using "area reasoning", we can compute (3.11) from (3.10) by starting in the middle of the Venn Diagram and then working our way outwards:

$$\begin{aligned}
P(1) &= P(A \cap B \cap C) \\
P(2) &= P(A \cap B) - P(1) \\
P(3) &= P(B \cap C) - P(1) \\
P(4) &= P(A \cap C) - P(1) \\
P(5) &= P(A) - (P(1) + P(2) + P(4)) \\
P(6) &= P(B) - (P(1) + P(2) + P(3)) \\
P(7) &= P(C) - (P(1) + P(3) + P(4)) \\
P(8) &= 1 - (P(1) + P(2) + P(3) + P(4) + P(5) + P(6) + P(7))
\end{aligned}$$

Example 3.8. Suppose the probabilities (3.10) as given as:

$$P(A) = P(B) = P(C) = 0.5$$

$$P(A \cap B) = P(A \cap C) = P(B \cap C) = 0.3$$
$$P(A \cap B \cap C) = 0.18$$

Then using the equations prior to this example, you compute

$$\begin{aligned} P(1) &= 0.18 \\ P(2) &= 0.12 \\ P(3) &= 0.12 \\ P(4) &= 0.12 \\ P(5) &= 0.08 \\ P(6) &= 0.08 \\ P(7) &= 0.08 \\ P(8) &= 0.22 \end{aligned}$$

We can now compute the probabilities of any of the 256 events determined by A, B, C. For example,

```
P[at least one of A,B,C occur]=P(1)+P(2)+P(3)+P(4)+P(5)+P(6)+P(7)=0.78
P[none of A,B,C occur] = P(8)= 0.22
P[exactly one of A,B,C occur] = P(5)+P(6)+P(7)= 0.24
P[exactly two of A,B,C occur] = P(2)+P(3)+P(4) = 0.36
```

Sometimes you are not directly given the probabilities (3.10), but instead you are given equations relating these probabilities. You can then re-express these equations in terms of the variables $P(1), P(2), \cdots, P(8)$, and solve the equations simultaneously for these variables. The following example is of this type.

Example 3.9. In Problem 3 of Homework Set 1, you are given the following relationships among the probabilities (3.10):

- $P(A) = 0.25,\ P(B) = 0.2,\ P(C) = 0.25$
- $P(A \cap B) = 0.1,\ P(A \cap B \cap C) = 0.05,\ P(A \cap C) = 2P(B \cap C)$
- The probability that at least two of the events A, B, C occur is 0.3.

If you think about it for a few moments, you will see that under these assumptions, the following system of 8 equations in 8 unknowns must hold:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 2 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} P(1) \\ P(2) \\ P(4) \\ P(3) \\ P(6) \\ P(5) \\ P(7) \\ P(8) \end{bmatrix} = \begin{bmatrix} 1 \\ 0.25 \\ 0.2 \\ 0.25 \\ 0.1 \\ 0.05 \\ 0 \\ 0.3 \end{bmatrix}$$

This system has a unique solution for $P(1), P(2), \cdots, P(8)$ which is left as an exercise for you to find.

Lecture 4

Chapter 1 Part 4

In Lecture 4, I talked about *Laws of Probability*, *Independent Events*, and *Application to Relay Circuits*.

4.1 Laws of Probability

There are quite a number of probability laws that can be proved from the axioms. Here are some of the most common laws.

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

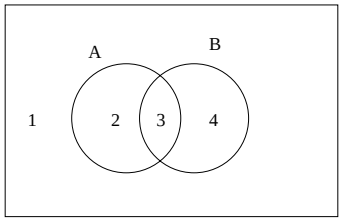
(ii): $P(A - B) = P(A) - P(A \cap B)$

(iii): $P(E^c) = 1 - P(E)$

(iv): $E \subset F \Rightarrow P(E) \leq P(F)$

If you are trying to prove a probability law involving a small number of events, Venn Diagrams can help you to come up with a proof. To illustrate, let us prove Law(i) this way.

Proof of Law(i): We use the Venn Diagram



The following facts are obvious from the Venn Diagram:

$$\begin{aligned} P(A) &= P(2) + P(3) \\ P(B) &= P(3) + P(4) \\ P(A \cap B) &= P(3) \\ P(A \cup B) &= P(2) + P(3) + P(4) \end{aligned}$$

From the first three of these equations, we obtain

$$P(A) + P(B) - P(A \cap B) = P(2) + P(3) + P(4),$$

which is $P(A \cup B)$. This completes the proof of Law(i).

Proof of Law(iii): S is the disjoint union of E and E^c. Therefore,

$$1 = P(S) = P(E) + P(E^c).$$

Solving for $P(E^c)$ in terms of $P(E)$, we obtain Law(iii).

Proof of Law(iv): Event E is assumed to be inside event F. Therefore, event F is the disjoint union of event E and event $F - E$. (If you need to, draw a Venn Diagram to help you see this.) This allows us to write down the equation

$$P(F) = P(E) + P(F - E).$$

Since $P(F - E) \geq 0$, it clearly follows that $P(E) \leq P(F)$.

Exercise. Prove Law(ii).

Exercise. Prove

$$P(E \Delta F) \leq |P(E) - P(F)|.$$

Exercise. Prove the following statement about the union of any three events:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C). \quad (4.1)$$

(Hint: Refer to our earlier Venn Diagram involving events A, B, C and the eight disjoint regions $1, 2, 3, 4, 5, 6, 7, 8$. Express the left side of (4.1) and each term on the right side of (4.1) as a linear combination of the eight probabilities

$$P(1), P(2), P(3), P(4), P(5), P(6), P(7), P(8).$$

The two sides of (4.1) should then cancel each other out.)

4.2 Independent Events

Let E^{+1} denote the event E and let E^{-1} denote the event E^c.

Definition. We say that events $E_1, E_2, \cdots, E_k$ are *independent events* if

$$P(E_1^{\pm 1} \cap E_2^{\pm 1} \cap E_3^{\pm 1} \cap \cdots \cap E_k^{\pm 1}) = P(E_1^{\pm 1})P(E_2^{\pm 1})P(E_3^{\pm 1}) \cdots P(E_k^{\pm 1}), \tag{4.2}$$

where, on the left side, we make all possible choices of ± 1 in each position, and, on the right side, we make these same choices. Thus, to show that k events are independent, we have to check 2^k equations of the form (4.2).

If events $E_1, E_2, \cdots, E_k$ fail to be independent, then we say that they are *dependent events.*

Two Independent Events

The definition just given says that to show two events A, B are independent, we have to verify the four equations

$$\begin{aligned} P(A \cap B) &= P(A)P(B) && (4.3)\\ P(A^c \cap B) &= P(A^c)P(B) && (4.4)\\ P(A \cap B^c) &= P(A)P(B^c) && (4.5)\\ P(A^c \cap B^c) &= P(A^c)P(B^c) && (4.6) \end{aligned}$$

Let me show you that these just reduce to the single equation

$$P(A \cap B) = P(A)P(B).$$

There is a "brute force" way to do this (just show that equations (4.4)-(4.6) must all be true if (4.3) is true). Instead, I use a more clever approach which uses ideas that will also be of use to us later in the course. Consider the following 2×2 array:

$$\begin{array}{cc} & \begin{array}{cc} P(B) & P(B^c) \end{array} \\ \begin{array}{c} P(A) \\ P(A^c) \end{array} & \left(\begin{array}{cc} P(A \cap B) & P(A \cap B^c) \\ P(A^c \cap B) & P(A^c \cap B^c) \end{array} \right) \end{array} \tag{4.7}$$

Using a Venn Diagram, you can show the following useful facts:

- The row headings $P(A), P(A^c)$ are the row sums for the respective rows of the 2×2 array.
- The column headings $P(B), P(B^c)$ are the column sums for the respective columns of the 2×2 array.

Let us also consider the following array in which the row headings and column headings are easily verified to have this same interpretation:

$$\begin{array}{c} \\ P(A) \\ P(A^c) \end{array} \begin{array}{c} \begin{array}{cc} P(B) & P(B^c) \end{array} \\ \begin{pmatrix} P(A)P(B) & P(A)P(B^c) \\ P(A^c)P(B) & P(A^c)P(B^c) \end{pmatrix} \end{array} \tag{4.8}$$

Arrays (4.7) and (4.8) have the same row and column sums and the same upper left hand corner (the number $P(A \cap B)$, assumed to be the same as the number $P(A)P(B)$). Therefore, these two arrays are identical! (It is easy to argue that two 2×2 arrays must coincide if they have the same row and column sums and the same entry in the upper left hand corner.) Since our two arrays (4.7), (4.8) coincide, equations (4.3)-(4.6) must hold, and we are done.

Three Independent Events

To verify that events A, B, C are independent, the following eight equations would have to be verified:

$$\begin{aligned} P(A \cap B \cap C) &= P(A)P(B)P(C) \\ P(A^c \cap B \cap C) &= P(A^c)P(B)P(C) \\ P(A \cap B^c \cap C) &= P(A)P(B^c)P(C) \\ P(A \cap B \cap C^c) &= P(A)P(B)P(C^c) \\ P(A^c \cap B^c \cap C) &= P(A^c)P(B^c)P(C) \\ P(A^c \cap B \cap C^c) &= P(A^c)P(B)P(C^c) \\ P(A \cap B^c \cap C^c) &= P(A)P(B^c)P(C^c) \\ P(A^c \cap B^c \cap C^c) &= P(A^c)P(B^c)P(C^c) \end{aligned}$$

Exercise. You can pick out 5 of the preceding equations, such that if these 5 equations are true, then the remaining 3 equations are true. Which 5 equations can you pick? (There is more than one possible answer.)

Intuitiveness of Independence Concept

Suppose we have a multiple step experiment in which the outcome on any step is not contingent upon other steps. If $E_1, E_2, \cdots, E_k$ are events associated with different steps, then our intuition suggests that these events are independent.

Example 4.1. Flip three fair dies. For any i, j, k belonging to the set $\{1, 2, 3, 4, 5, 6\}$, intuition tells us that the following three events should be independent:

$$E_1 = \{die\ 1 = i\}, \quad E_2 = \{die\ 2 = j\}, \quad E_3 = \{die\ 3 = j\}.$$

Let us see whether this intuition is justified. Since

$$P(E_1)P(E_2)P(E_3) = (1/6)(1/6)(1/6) = 1/216$$

and

$$P(E_1 \cap E_2 \cap E_3) = P(i, j, k) = 1/216$$

give the same result, this is a strong suggestion that these three events are indeed independent, and this can be verified. (You'd have to verify a total of 8 equations, but since i, j, k are arbitrary, these equations will be true. The arbitrariness of i, j, k gives us 216 equations, from which the 8 equations we need are verifiable. You can fill in the details.) The fact that E_1, E_2, E_3 are independent (which is a mathematical fact) confirms our intuition that they should be independent.

Example 4.2. Go back to Example 3.3, in which three coins were tossed, one of them fair and the other two unfair. Intuition tells us that any three events of the form

$$\{Coin\ 1 = H\ or\ T\}, \{Coin\ 2 = H\ or\ T\}, \{Coin\ 3 = H\ or\ T\} \tag{4.9}$$

should be independent. In the 8 equations (3.1)-(3.8), you see how the probability of the intersection of these three events would be computed as the product of the probabilities of the individual events. (This is because we have an independent probability model, which is reflected in the fact that in the tree on page 21 the H, T probability labels across each level of the tree are the same.) These 8 equations are precisely the 8 equations you'd have to check to make sure any three events of type (4.9) are independent. Therefore, any three events of type (4.9) are indeed independent from the mathematical definition of independence, which confirms our intuition.

Example 4.3. Here we examine cases of pairs of events in which we cannot use our intuition to see whether we have independence or not. Flip two fair dies and consider the events

$$A = \{first\ die\ = 2\}, \quad B = \{total\ = 7\}.$$

Then we have

$$P(A \cap B) = P(2, 5) = 1/36.$$

On the other hand, we have

$$P(A)P(B) = (1/6)(6/36) = 1/36.$$

Since $P(A \cap B) = P(A)P(B)$, events A, B are independent but we have no intuition for saying so. To make this more evident, suppose we change the second event a little bit:

$$B_1 = \{total\ = 6\}.$$

Then

$$P(A \cap B_1) = P(2, 4) = 1/36,$$

but

$$P(A)P(B_1) = (1/6)(5/36) \neq 1/36.$$

We conclude (based upon the above mathematics) that A, B_1 are dependent events, but we have no intuition for saying so.

Probability Calculations Involving Independent Events

Events $E_1, E_2, \cdots, E_k$ determine a total $2^{(2^k)}$ events if you take all possible events formed by means of unions, intersections, and complementation. If the events $E_1, E_2, \cdots, E_k$ are independent, it is useful to know that the probability of any of these $2^{(2^k)}$ events can be uniquely determined as a combination of the probabilities

$$P(E_1), P(E_2), \cdots, P(E_k).$$

The following two examples illustrate this fact.

Example 4.4. Let A, B be independent events. Then

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

To see this, go back to Law(i) proved at the beginning of this Lecture and substitute $P(A)P(B)$ for $P(A \cup B)$.

Example 4.5. Let A, B, C be independent events. Then it is easy to show that any two of these three events are independent. (Try to prove this.) Go back to formula (4.1) developed earlier for computing the probability of the union of A, B, C. In that formula, you can make the following substitutions on the right side:

$$\begin{aligned} P(A \cap B) &= P(A)P(B) \\ P(A \cap C) &= P(A)P(C) \\ P(B \cap C) &= P(B)P(C) \\ P(A \cap B \cap C) &= P(A)P(B)P(C) \end{aligned}$$

This gives us the formula

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A)P(B) - P(A)P(C) - P(B)P(C) - P(A)P(B)P(C). \tag{4.10}$$

Here is another way to obtain the same result. Using the complementation Law(iii) several times we have

$$\begin{aligned} P(A \cup B \cup C) &= 1 - P(A^c \cap B^c \cap C^c) \\ &= 1 - P(A^c)P(B^c)P(C^c) \\ &= 1 - (1 - P(A))(1 - P(B))(1 - P(C)) \end{aligned}$$

It is a simple exercise in algebra to show that $1 - (1 - P(A))(1 - P(B))(1 - P(C))$ is the same as the right side of (4.10).

4.3 Application of Independence to Relay Circuits

Switches in Series

We start with the relay circuit

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

The circuit elements $1, 2, 3$ are switches which each have only two possible states: "on" or "off". You are attempting to have some quantity (such as information or an electrical current) flow from point A to point B. With the three switches connected in series as we have here, the only way that can happen is if all the switches are "on". We suppose that the switches operate randomly and independently, with

$$p_i = P(\{\text{switch i is "on"}\}),$$

and therefore it is automatically true that

$$1 - p_i = P(\{\text{switch i is "off"}\}).$$

Let $\{A \to B\}$ denote the event that a connection from A to B is possible. Our goal is to compute $P(\{A \to B\})$, the probability that the A to B connection can be made. Note that $\{A \to B\}$ is the intersection of three independent events:

$$\{A \to B\} = \{\text{switch 1 is "on"}\} \cap \{\text{switch 2 is "on"}\} \cap \{\text{switch 3 is "on"}\}.$$

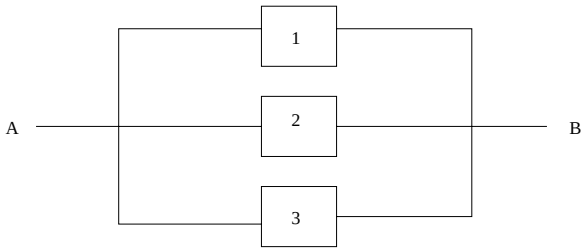
Taking the probability of both sides, we conclude that

$$P(\{A \to B\}) = p_1 p_2 p_3.$$

The argument we just made is applicable to any relay circuit consisting of k switches connected in series, where k can be any positive integer. For k switches in series, we'd have

$$P(\{A \to B\}) = p_1 p_2 p_3 \cdots p_k.$$

Switches in Parallel



For the preceding relay circuit, the connection from A to B will be operative if and only if at least one of the switches is "on". That is, we have a union event:

$$\{A \to B\} = \{\text{switch 1 is "on"}\} \cup \{\text{switch 2 is "on"}\} \cup \{\text{switch 3 is "on"}\}.$$

Complementing both sides, we have

$$\{A \to B\}^c = \{\text{switch 1 is "off"}\} \cap \{\text{switch 2 is "off"}\} \cap \{\text{switch 3 is "off"}\}.$$

Taking the probability of both sides,

$$P(\{A \to B\}^c) = 1 - P(\{A \to B\}) = (1-p_1)(1-p_2)(1-p_3),$$

and then we conclude that

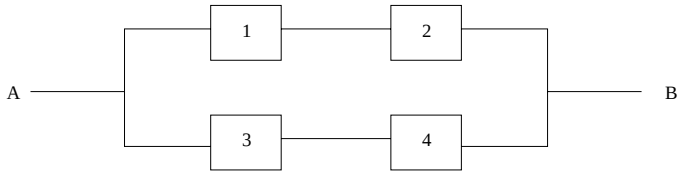
$$P(\{A \to B\}) = 1 - (1-p_1)(1-p_2)(1-p_3).$$

If we have k switches connected in parallel, then the answer is

$$P(\{A \to B\}) = 1 - \prod_{i=1}^{k}(1-p_i).$$

Combined Series/Parallel Circuits

Some relay circuits can be handled via a combination of the series and parallel approaches. Here is one:

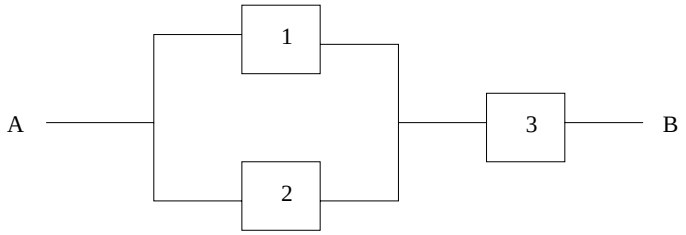


You can replace the series connection of switches 1,2 with a single switch that operates with probabiity $q_1 = p_1p_2$. Similarly, you can replace switches $3, 4$ with a single switch that operates with probability $q_2 = p_3p_4$. You then have a parallel connection of two switches that operate with probabilities q_1, q_2, respectively. We conclude that

$$P(\{A \to B\}) = 1 - (1-q_1)(1-q_2) = 1 - (1-p_1p_2)(1-p_3p_4).$$

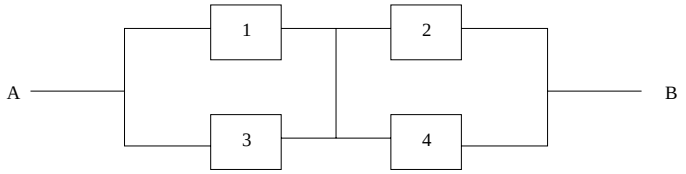
Exercise. For the relay circuit at the top of the next page, prove that

$$P(\{A \to B\}) = [1 - (1-p_1)(1-p_2)]p_3.$$



General Circuits

Some relay circuits can't be handled by our previous approaches. We cover here a method applicable to all relay circuits to determine $P(\{A \to B\})$. Let us examine the circuit:



Our general method begins by listing all paths that allow a possible connection from A to B. In this case, we have four such paths:

$$12, \quad 14, \quad 32, \quad 34.$$

A connection from A to B can be made along a given path if and only if all switches along that path are "on". This allows one to express $P(\{A \to B\})$ as the probability of a union event

$$P(\{A \to B\}) = P(E_1 \cup E_2 \cup E_3 \cup E_4),$$

where E_1, E_2, E_3, E_4 are the events

$$\begin{aligned} E_1 &= \{12 \; all \; on\} \\ E_2 &= \{14 \; all \; on\} \\ E_3 &= \{32 \; all \; on\} \\ E_4 &= \{34 \; all \; on\} \end{aligned}$$

The probability of any union event can be expressed as a linear combination of probabilities of intersection events. (We have seen this already for unions of 2 events and 3 events.) Here is what

you get for the union of 4 events:

$$P(E_1 \cup E_2 \cup E_3 \cup E_4) = \sum_{i=1}^{4} P(E_i) - \sum_{1 \le i < j \le 4} P(E_i \cap E_j) + \sum_{1 \le i < j < k \le 4} P(E_i \cap E_j \cap E_k) - P(E_1 \cap E_2 \cap E_3 \cap E_4). \tag{4.11}$$

The second summation contains 6 terms:

$$P(E_1 \cap E_2) + P(E_1 \cap E_3) + P(E_1 \cap E_4) + P(E_2 \cap E_3) + P(E_2 \cap E_4) + P(E_3 \cap E_4).$$

The third summation contains 4 terms:

$$P(E_1 \cap E_2 \cap E_3) + P(E_1 \cap E_2 \cap E_4) + P(E_1 \cap E_3 \cap E_4) + P(E_2 \cap E_3 \cap E_4)$$

Evaluating all the terms, we obtain

$$\begin{aligned}
P(E_1) &= p_1 p_2 \\
P(E_2) &= p_1 p_4 \\
P(E_3) &= p_3 p_2 \\
P(E_4) &= p_3 p_4 \\
P(E_1 \cap E_2) &= p_1 p_2 p_4 \\
P(E_1 \cap E_3) &= p_1 p_2 p_3 \\
P(E_1 \cap E_4) &= p_1 p_2 p_3 p_4 \\
P(E_2 \cap E_3) &= p_1 p_2 p_3 p_4 \\
P(E_2 \cap E_4) &= p_1 p_3 p_4 \\
P(E_3 \cap E_4) &= p_2 p_3 p_4 \\
P(E_1 \cap E_2 \cap E_3) &= p_1 p_2 p_3 p_4 \\
P(E_1 \cap E_2 \cap E_4) &= p_1 p_2 p_3 p_4 \\
P(E_1 \cap E_3 \cap E_4) &= p_1 p_2 p_3 p_4 \\
P(E_2 \cap E_3 \cap E_4) &= p_1 p_2 p_3 p_4 \\
P(E_1 \cap E_2 \cap E_3 \cap E_4) &= p_1 p_2 p_3 p_4
\end{aligned}$$

In case the reader may be confused by how we arrived at the preceding results, we compute $P(E_1 \cap E_3)$ as an example: E_1, E_3 both occur if and only if switches 1,2,3 are all on, and so $E_1 \cap E_3$ is the intersection of 3 independent events; the result $p_1 p_2 p_3$ should now be evident. Plugging back into (4.11), a lot of cancellation occurs. Putting the answer in simplest form, we got:

$$P(\{A \to B\}) = p_1 p_2 + p_1 p_4 + p_3 p_2 + p_3 p_4 - p_1 p_2 p_4 - p_1 p_2 p_3 - p_1 p_3 p_4 - p_2 p_3 p_4 + p_1 p_2 p_3 p_4.$$

We have just illustrated the use of one general method for handling relay circuits. There is a second general method, which works as follows. Suppose there are k switches in the circuit. Let

E_i^{+1} be the event that switch i is "on", and let E_i^{-1} be the event that switch i is "off". Consider the 2^k events of form

$$E_1^{\pm 1} \cap E_2^{\pm 1} \cap E_3^{\pm 1} \cap \cdots \cap E_k^{\pm 1}. \tag{4.12}$$

Using independence, the probability of each such event is easy to compute according to formula (4.2). The event $\{A \to B\}$ is a disjoint union of certain of the events of form (4.12). Find out which events these are, and then add up their probabilities—the result is $P(\{A \to B\})$. For some relay circuits, this second method works better than the method we described earlier. For an example using this second method, see Problem 6.1 in the Chapter 1 Solved Problems.

Exercise. Show that the last circuit we considered can be handled in a simpler way.

Lecture 5

Chapter 1 Part 5

In Lecture 5, we started to talk about conditional probability. I derived what the conditional probability formula $P(E|F)$ should be, and then started to look at some examples involving conditional probability. One significant application of conditional probabilities throughout the course will be to the *discrete communication channel*; I will complete these notes by explaining what this channel model is.

5.1 Conditional Probability Derivation

When we originally came up with a probability model P on our sample space S, we assumed that we did not have any advance information about where the experiment's outcome ω might lie within S. Suppose instead that we know that the outcome ω will lie in F, an event contained in S. To reflect this knowledge, we should change our probability model P to a new probability model which I will call P_F. We need to figure out what the "conditional probability model" P_F is.

For simplicity, let us assume a discrete sample space S. Then $P(\omega)$ is defined for every $\omega \in S$. We now have to define $P_F(\omega)$ for every $\omega \in S$. In order to do this, we will be guided by the following two principles:

(i): $P_F(\omega)$ should 0 for outcomes ω lying outside of F (because these outcomes cannot occur under the conditional information).

(ii): For outcomes ω in F, the $P_F(\omega)$'s should be in the same proportions as the $P(\omega)$'s. (In other words, if an outcome ω_1 in F is twice as likely under model P as some other outcome ω_2 in F, then under the conditional model P_F, outcome ω_1 will still be twice as likely as outcome ω_2.)

Because of assumption (ii), there will exist a positive constant C such that

$$P_F(\omega) = CP(\omega), \quad \omega \in F. \tag{5.1}$$

Once we figure out what the value of C is, then we will know what $P_F(\omega)$ is for every $\omega \in F$. Summing both sides of equation (5.1) over all $\omega \in F$, we obtain

$$P_F(F) = C \sum_{\omega \in F} P(\omega) = CP(F).$$

We must have $P_F(F) = 1$ (why?). Therefore, we conclude that

$$C = \frac{1}{P(F)}.$$

Here is then our formula for all the $P_F(\omega)$ values as ω ranges through S:

$$P_F(\omega) = \begin{cases} 0, & \omega \notin F \\ \frac{P(\omega)}{P(F)}, & \omega \in F \end{cases}$$

Let E be any event. We will denote the probability $P_F(E)$ by the standard notation $P(E|F)$. We call $P(E|F)$ the "conditional probability of E given F". Let us derive a formula for $P(E|F)$ based upon our preceding work. We have:

$$\begin{aligned} P_F(E) &= \sum_{\omega \in E} P_F(\omega) \\ &= \sum_{\omega \in E, \omega \notin F} P_F(\omega) + \sum_{\omega \in E, \omega \in F} P_F(\omega) \\ &= 0 + \sum_{\omega \in E, \omega \in F} \frac{P(\omega)}{P(F)} \\ &= \left(\frac{1}{P(F)}\right) \sum_{\omega \in E \cap F} P(\omega) \\ &= \frac{P(E \cap F)}{P(F)} \end{aligned}$$

We have derived the formula $P(E|F) = P(E \cap F)/P(F)$. We can reverse the roles of E and F to obtain a formula for $P(F|E)$, the conditional probability for F given E. We can also multiply both sides of $P(E|F) = P(E \cap F)/P(F)$ by $P(F)$ to obtain a formula for $P(E \cap F)$ as a product of an unconditional probability and a conditional probability. This gives us several formulas, which we list below.

Some Formulas Involving Conditional Probability

(a): $P(E|F) = \frac{P(E \cap F)}{P(F)}$

(b): $P(F|E) = \frac{P(E \cap F)}{P(E)}$

(c): $P(E \cap F) = P(E)P(F|E)$

(d): $P(E \cap F) = P(F)P(E|F)$

These four formulas are really saying the same thing. We wrote the formulas separately for emphasis. In some applications, you will compute conditional probabilities using (a) or (b). In other applications, you will already know the conditional probabilities and will be using them to compute the left side of (c),(d).

Example 5.1. Let us harken back to Example 4.3. We flip a pair of fair dice. Consider the events

$$\begin{aligned} F &= \{first\ die\ = 2\} \\ E &= \{total\ = 6\} \end{aligned}$$

Then

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{P(2,4)}{P(first\ die\ = 2)} = \frac{1/36}{1/6} = 1/6.$$

We know from earlier work that $P(E) = 5/36$, which is less than $1/6$. This is a case where given information has made an event more likely to occur than it would have been in the absence of any information. There can be other cases where given information will make an event less likely to occur.

In Example 5.1, let us change the E event slightly to

$$E = \{total\ = 7\}.$$

The reader can show that $P(E|F) = 1/6$ for this case (the above argument for our previous choice of E will work here almost word for word). However, we know from our previous work that $P(E) = 1/6$. Therefore, $P(E)$ and $P(E|F)$ are the same. In other words, we have an event that is just as likely to occur given some information as it is to occur in the absence of any information. When does this situation occur? In formula(d) above, substitute $P(E)$ for $P(E|F)$ and you will see that $P(E \cap F) = P(F)P(E)$, that is, events E, F are independent. It follows that two events are independent if and only if the conditional probability of either event given the other one is the same as the unconditional probability of that event. In this way, the concept of independent events can be considered as a byproduct of the theory of conditional probabilities. We summarize our conclusions below.

Independence and Conditional Probabilities

Let E, F be events. The following statements are equivalent:

(e): E, F are independent.

(f): $P(E|F) = P(E)$.

(g): $P(F|E) = P(F)$.

(Note: When we say that statements (e),(f),(g) above are equivalent, we mean that if any one of them is true, then the other two statements are also true.)

5.2 Sampling Without Replacement

Suppose you draw items randomly from a pool of items one by one, without putting previously selected items back into the pool before selecting the next item. This procedure is called sampling without replacement (abbreviated as "sampling w/o replacement"). Quality control would be one important application in which sampling w/o replacement would take place; see Section 1.1.4 of Lecture 1. It is natural for us to consider sampling w/o replacement here in the context of conditional probabilities, because in such a scenario it is easy to describe the likelihood of what happens on draws after the first draw in terms of conditional probabilities.

In this our first exposure to sampling w/o replacement, we illustrate some ideas using a "toy problem" involving a so-called *urn model*.

Example 5.2. Suppose we have an urn containing 2 black balls and 3 white balls. We draw two balls from the urn without replacement. We record as the outcome of this experiment only the color of each ball selected. The sample space is therefore

$$S = \{BB, BW, WB, WW\},$$

where in each outcome the first entry denotes the color of the first ball selected and the second entry denotes the color of the second ball selected. Note that each of the four outcomes is the intersection of two events. It is therefore natural for us to use formula(c) as a means to determine the probability of an outcome in S. Let $B1, B2, W1, W2$ denote the events

$$\begin{aligned} B1 &= \{Ball\ 1\ is\ black\} \\ B2 &= \{Ball\ 2\ is\ black\} \\ W1 &= \{Ball\ 1\ is\ white\} \\ W2 &= \{Ball\ 2\ is\ white\} \end{aligned}$$

Using formula(c) repeatedly, we have

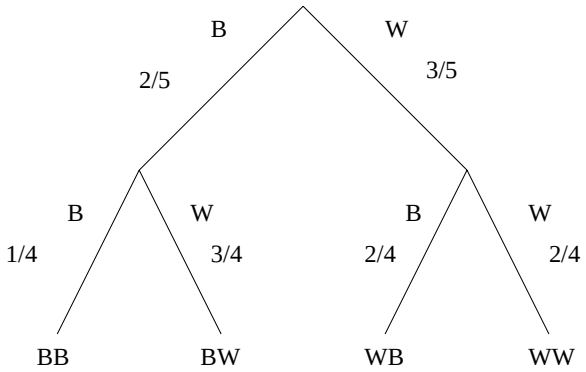
$$\begin{aligned} P(BB) &= P(B1)P(B2|B1) = (2/5)(1/4) = 0.10 \\ P(BW) &= P(B1)P(W2|B1) = (2/5)(3/4) = 0.30 \\ P(WB) &= P(W1)P(B2|W1) = (3/5)(2/4) = 0.30 \\ P(WW) &= P(W1)P(W2|W1) = (3/5)(2/4) = 0.30 \end{aligned}$$

It should be obvious how we obtained the values of $P(B1)$ and $P(W1)$. It is also not hard to see how we obtain the conditional probabilities. For example, let me explain how I obtained the values for $P(B2|B1)$ and $P(W2|B1)$. Given that the first draw results in a black ball, since this ball is set aside, the composition of the urn at the beginning of the second draw is 1 black ball and 3 white balls. You then have 1 chance in 4 of drawing a second ball which is black and therefore 3 chances in 4 of drawing a second ball which is white; that is,

$$P(B2|B1) = 1/4, \quad P(W2|B1) = 3/4.$$

Notice that these two conditional probabilities form a probability distribution (i.e., they add up to one). This is because when we condition on a fixed event (event $B1$ in this case), we obtain a conditional probability model given that event (we called this conditional model P_F back at the beginning of Lecture 5 notes; event F is $B1$ in this case).

We have just illustrated how to use the conditional probability concept to find the probability model in Example 5.2. Alternatively, we could have found this probability model by using a tree to model the experiment in Example 5.2, as we did a few lectures ago. It is interesting to look at this tree and see how it relates to the computations we just completed above. The tree is:



Look at the probability labels on the tree branches at the second level. When we earlier considered tree models, we did not have a terminology for designating these labels. Now, on the basis of our treatment of conditional probabilities up to this point, we can interpret these four second level probability labels (from left to right) as:

$$P(B2|B1), P(W2|B1), P(B2|W1), P(W2|W1).$$

In the tree above, we do not have any branches below the second level. This is because we made only two draws from the urn. If we draw more than two balls, then our tree goes deeper beyond the second level. However, we will still be able to interpret any probability label on a branch below

the first level as some sort of conditional probability. We will ultimately consider an example in which we draw more than two balls from an urn.

Exercise. We have a standard 52 card deck of playing cards. You are dealt two cards at random without replacement. Compute the probability that you obtain exactly one K card (K=king). I will get you started on a solution. Take the sample space as

$$S = \{(K, K), (K, NK), (NK, K), (NK, NK)\},$$

where "NK" denotes a "nonking" card, i.e., a card which is not a king. You need to compute

$$P(K, NK) + P(NK, K).$$

I will compute $P(K, NK)$ for you. You can compute $P(NK, K)$.

$$P(K, NK) = P(K1)P(NK2|K1) = (4/52)(48/51).$$

To see why this is true, note that initially you have 4 king cards, and 48 cards which are not kings. The first card then has 4 chances in 52 of being a king, that is,

$$P(K1) = 4/52.$$

After being dealt a king card as the first card, the deck now contains 3 king cards and 48 cards which are not kings. Therefore, you have 48 chances in 51 of drawing a nonking card as your second card:

$$P(NK2|K1) = 48/51.$$

5.3 Discrete Communication Channel Model

In data communications, conditional probabilities are used to describe the various likelihoods with which a communication channel will generate possible outputs given each fixed possible input. Let us conceptualize such a channel via the following block diagram:

$$input \rightarrow \boxed{\text{channel}} \rightarrow output$$

There is a finite set of possible inputs and a finite set of possible outputs. For the sake of illustration, let us suppose that there are two possble inputs, which come from the set

$$\{a_1, a_2\}, \tag{5.2}$$

and that there are three possible outputs, which come from the set

$$\{b_1, b_2, b_3\}. \tag{5.3}$$

Suppose we perform the random experiment in which we select an input at random from the set (5.2) and transmit it through the channel, which results in a randomly generated output from the set (5.3). The sample space S of this random experiment consists of $2 * 3 = 6$ input-output pairs of the form (a_i, b_j). There is a convenient convention for representing these pairs pictorially:

$$\begin{array}{c} \\ a_1 \\ a_2 \end{array} \begin{array}{c} \begin{array}{ccc} b_1 & b_2 & b_3 \end{array} \\ \begin{pmatrix} (a_1,b_1) & (a_1,b_2) & (a_1,b_3) \\ (a_2,b_1) & (a_2,b_2) & (a_2,b_3) \end{pmatrix} \end{array}$$

Notice that the 6 outcomes in the sample space appear within the parentheses as a 2×3 array. The row headings are the possible inputs and the column headings are the possible outputs. If the input is a_i and the output is b_j, then the input-output pair (a_i, b_j) appears at the intersection of the row with heading a_i and the column with heading b_j.

To specify the probability model for our random experiment, we need to be able to compute six probabilities of the form $P(a_i, b_j)$. It is also convenient to put these probabilities in a 2×3 array as follows:

$$\begin{array}{c} \\ a_1 \\ a_2 \end{array} \begin{array}{c} \begin{array}{ccc} b_1 & b_2 & b_3 \end{array} \\ \begin{pmatrix} P(a_1,b_1) & P(a_1,b_2) & P(a_1,b_3) \\ P(a_2,b_1) & P(a_2,b_2) & P(a_2,b_3) \end{pmatrix} \end{array} \tag{5.4}$$

Let $P(b_j|a_i)$ denote the conditional probability that the channel output is b_j given that the channel input is a_i. Let $P(a_i)$ denote the probability with which input a_i is chosen. From our earlier work with conditional probabilities in the Lecture 5 notes, we have the following six equations via which each $P(a_i, b_j)$ can be computed:

$$P(a_i, b_j) = P(a_i)P(b_j|a_i), \quad i = 1, 2; j = 1, 2, 3.$$

It will be convenient for us to view these computations from the matrix point of view. To do this, let us put the 6 conditional proabilities $P(b_j|a_i)$ in the following array:

$$\begin{array}{c} \\ a_1 \\ a_2 \end{array} \begin{array}{c} \begin{array}{ccc} b_1 & b_2 & b_3 \end{array} \\ \begin{pmatrix} P(b_1|a_1) & P(b_2|a_1) & P(b_3|a_1) \\ P(b_1|a_2) & P(b_2|a_2) & P(b_3|a_2) \end{pmatrix} \end{array} \tag{5.5}$$

In array (5.5), row 1 represents the conditional probability model for the different channel outputs that can arise when the channel input is a_1. Row 2 represents the conditional probability model for the different channel outputs that can arise when the channel input is a_2. Since each row of (5.5) is a conditional probability model, *each row of (5.5) sums up to one.* In order to specify a discrete channel model, you would be given a matrix like (5.5) in which each row sums up to one. This matrix is called *the channel matrix.* Given the channel matrix (5.5) and the probabilities $P(a_1), P(a_2)$ of the inputs, then it is easy to describe how you compute the array (5.4) which gives

the probability model on our sample S of input-output pairs: You multiply each row of the channel matrix (5.5) by the probability $P(a_i)$ for the a_i that is the row header for that row. Equivalently, we can do the following matrix product:

$$\begin{pmatrix} P(a_1) & 0 \\ 0 & P(a_2) \end{pmatrix} \begin{pmatrix} P(b_1|a_1) & P(b_2|a_1) & P(b_3|a_1) \\ P(b_1|a_2) & P(b_2|a_2) & P(b_3|a_2) \end{pmatrix} = \begin{pmatrix} P(a_1,b_1) & P(a_1,b_2) & P(a_1,b_3) \\ P(a_2,b_1) & P(a_2,b_2) & P(a_2,b_3) \end{pmatrix}$$

Conclusion: *Suppose we are given the channel matrix which specifies a given discrete channel model. Suppose we are also given the probabilities with which the channel inputs are to be selected. Perform the following random experiment: Select a channel input at random, transmit it through the channel, and record the resulting input-output pair. The probability model for this experiment can then be computed as the matrix product*

$$D * C,$$

where D is the diagonal matrix which has the input probabilities on the diagonal, and C is the channel matrix.

Example 5.3. Suppose the channel matrix is given as

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 0.90 & 0.10 \\ 0.05 & 0.95 \end{pmatrix} \end{array}$$

What does this mean? From the row and column headings, we see that the possible inputs are $0, 1$ and the possible outputs are also $0, 1$. From the first row of the channel matrix, we see that when the channel input is 0, the channel output will also be 0 with probability 0.90. That is, when the channel input is 0, the channel will operate correctly 90% of the time. (If we were to transmit thousands of 0's through this channel, then we would see that about 90% of the resulting outputs are 0 and about 10% of them are incorrect (i.e., equal to 1). From the second row of the channel matrix, we see that when the channel input is 1, the channel output will also be 1 with probability 0.95. That is, when the channel input is 1, the channel will operate correctly 95% of the time. Suppose we select a channel input equally likely to be 0 or 1:

$$P(0) = 1/2, \quad P(1) = 1/2.$$

Having selected the channel input randomly, we then transmit this input through the channel, and then record the pair consisting of this input with the observed output as the outcome of our random experiment. The probability model for this random experiment is then the matrix product:

$$\begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix} \begin{pmatrix} 0.90 & 0.10 \\ 0.05 & 0.95 \end{pmatrix} = \begin{pmatrix} 0.45 & 0.05 \\ 0.025 & 0.475 \end{pmatrix}$$

At this point, it helps to put the row and column headings back in, which gives us

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 0.45 & 0.05 \\ 0.025 & 0.475 \end{pmatrix} \end{array}$$

The sample space of our experiment consists of the four possible input-output pairs:

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

From our preceding array with the row and column headings, we can now read off the probabilities for the probability model for our experiment as follows:

$$\begin{aligned} P(0,0) &= 0.45 \\ P(0,1) &= 0.05 \\ P(1,0) &= 0.025 \\ P(1,1) &= 0.475 \end{aligned}$$

As we go through the course, we will do more and more things with the discrete channel model. (This is just the beginning!)

Lecture 6

Chapter 1 Part 6

In this set of Lecture notes, I finish Chapter 1. The principal topic is *Bayes Method*. As a byproduct of developing Bayes Method, I obtain some Laws of Conditional Probability, which I will state separately with some examples at the end of these notes.

6.1 Bayes Method Explained

Let events $\{A_i\}$ partition the sample space S. (This means that S is a disjoint union of these events.) Let events $\{B_j\}$ also partition S. Suppose you are given the $P(A_i)$'s and the $P(B_j|A_i)$'s (these conditional probabilities are called *forward conditional probabilities*). The goal of Bayes Method is to do the following:

- Compute the $P(B_j)$'s.
- Compute the $P(A_i|B_j)$'s, which are called the *backward conditional probabilities*.

As a by-product of Bayes Method, we will also compute the probabilities $P(A_i \cap B_j)$, called *joint probabilities*.

In order to explain how Bayes Method works, it will be convenient for us to deal with three different matrices, described as follows.

Matrix of Forward Conditional Probabilities: This is the matrix in which the matrix element in Row i and Column j is $P(B_j|A_i)$. For example, if there are two A_i's and three B_j's, the matrix of forward conditional probabilities is:

$$\begin{array}{c} \\ A_1 \\ A_2 \end{array} \begin{array}{c} \begin{array}{ccc} B_1 & B_2 & B_3 \end{array} \\ \begin{pmatrix} P(B_1|A_1) & P(B_2|A_1) & P(B_3|A_1) \\ P(B_1|A_2) & P(B_2|A_2) & P(B_3|A_2) \end{pmatrix} \end{array}$$

We use the A_i's as row headings and the B_j's as column headings for the matrix of forward cond probs. (This is a useful bookkeeping device which we use for all three types of matrices we are defining here.)

Matrix of Joint Probabilities: This is the matrix in which the matrix element in Row i and Column j is the joint probability $P(A_i \cap B_j)$. For example, if there are two A_i's and three B_j's, the matrix of joint probabilities is:

$$\begin{array}{c} \\ A_1 \\ A_2 \end{array} \begin{array}{c} \begin{array}{ccc} B_1 & B_2 & B_3 \end{array} \\ \begin{pmatrix} P(B_1 \cap A_1) & P(B_2 \cap A_1) & P(B_3 \cap A_1) \\ P(B_1 \cap A_2) & P(B_2 \cap A_2) & P(B_3 \cap A_2) \end{pmatrix} \end{array}$$

Matrix of Backward Conditional Probabilities: This is the matrix in which the matrix element in Row i and Column j is the backward conditional probability $P(A_i|B_j)$. For example, if there are two A_i's and three B_j's, the matrix of backward conditional probabilities is:

$$\begin{array}{c} \\ A_1 \\ A_2 \end{array} \begin{array}{c} \begin{array}{ccc} B_1 & B_2 & B_3 \end{array} \\ \begin{pmatrix} P(A_1|B_1) & P(A_1|B_2)) & P(A_1|B_3) \\ P(A_2|B_1) & P(A_2|B_2) & P(A_2|B_3) \end{pmatrix} \end{array}$$

Matrix Properties

(i): Each row of the matrix of forward conditional probabilities adds up to one. This is because Row i is the conditional probability model for the B_j's given A_i.

(ii): Each column of the matrix of backward conditional probabilities adds up to one. This is because Column j is the conditional probability model for the A_i's given B_j.

(iii): The sum of all of the elements of the joint probability matrix is equal to 1.

(iv): For each i, the sum of row i of the joint probability matrix is $P(A_i)$.

(v): For each j, the sum of column j of the joint probability matrix is $P(B_j)$.

Using the following Venn Diagram, it is not hard to see why the last three properties are true.

	B_1	B_2	B_3
A_1	$A_1 \cap B_1$	$A_1 \cap B_2$	$A_1 \cap B_3$
A_2	$A_2 \cap B_1$	$A_2 \cap B_2$	$A_2 \cap B_3$

Property(iii) is true because S (the entire Venn Diagram) is the disjoint union of all the $A_i \cap B_j$'s. Property(iv) is true because A_i is the disjoint union of the events $A_i \cap B_j$ in which j is allowed to vary and i is held fixed (the events in Row i of the Venn Diagram). Property(v) is true because B_j is the disjoint union of the events $A_i \cap B_j$ in which i varies and j is held fixed (the events in Column j of the Venn Diagram).

Remark. The discrete communication channel model, discussed at the end of the Lecture 5 Notes, yields a particular application of the Bayes Method machinery. The inputs to the channel yield the events A_i, the outputs to the channel yield the events B_j, and the matrix of forward conditional probabilities is the channel matrix.

Bayes Method involves three steps, which I now describe.

6.1.1 Step 1 of Bayes Method

This step consists of the computation of the joint probabilities, the $P(A_i \cap B_j)$'s. There are two ways to do this:

(i): If you want to compute the joint probabilities one by one, just plug into the right side of the following formula:

$$P(A_i \cap B_j) = P(A_i)P(B_j|A_i). \tag{6.1}$$

(ii): If you want to compute the matrix of joint probabilities all at once, then for each i, you multiply row i of the matrix of forward conditional probabilities by $P(A_i)$. Equivalently, this operation can be performed as the following matrix product:

$$\begin{pmatrix} P(A_1) & 0 \\ 0 & P(A_2) \end{pmatrix} \begin{pmatrix} P(B_1|A_1) & P(B_2|A_1) & P(B_3|A_1) \\ P(B_1|A_2) & P(B_2|A_2) & P(B_3|A_2) \end{pmatrix} = \begin{pmatrix} P(A_1 \cap B_1) & P(A_1 \cap B_2) & P(A_1 \cap B_3) \\ P(A_2 \cap B_1) & P(A_2 \cap B_2) & P(A_2 \cap B_3) \end{pmatrix}$$

In other words, you form a diagonal matrix whose diagonal entries are the $P(A_i)$'s, and then multiply this diagonal matrix (on the left) times the matrix of forward conditional probabilities; the result of this matrix product is the matrix of joint probabilities.

Formula (6.1) was covered in Lecture 5 notes. It is a special case of a Conditional Probability Law called the Multiplication Law; the general case of the Multiplication Law (which applies to intersections of possibly more than two events) will be given at the end of this set of notes.

6.1.2 Step 2 of Bayes Method

This step consists of the computation of the $P(B_j)$'s. There are two ways to do this:

(i): If you want to compute the $P(B_j)$'s one by one, just plug into the right side of the following formula:

$$P(B_j) = \sum_i P(A_i)P(B_j|A_i). \tag{6.2}$$

(ii): If you want to compute the $P(B_j)$'s all at once, multiply the matrix of forward cond probs on the left by the row vector whose entries are the $P(A_i)$'s:

$$(P(A_1)\ \ P(A_2)) \begin{pmatrix} P(B_1|A_1) & P(B_2|A_1) & P(B_3|A_1) \\ P(B_1|A_2) & P(B_2|A_2) & P(B_3|A_2) \end{pmatrix} = (P(B_1)\ \ P(B_2)\ \ P(B_3))\,.$$

Equation (6.2) is the "Law of Total Probability". It is easy to prove. We defer its proof and a discussion of other applications of this Law to the end of the Lecture 6 Notes.

6.1.3 Step 3 of Bayes Method

This step consists of the computation of the backward conditional probabilities, the $P(A_i|B_j)$'s. There are two ways to do this:

(i): If you want to compute the $P(A_i|B_j)$'s one by one, just plug into the right side of the formula

$$P(A_i|B_j) = \frac{P(A_i)P(B_j|A_i)}{\sum_{i'} P(A_{i'})P(B_j|A_{i'})}. \tag{6.3}$$

(ii): If you want to compute the backward conditional probabilities all at once, then for each j, divide Column j of the joint probability matrix by the column sum for that column (which is $P(B_j)$). These column operations yield the matrix of backward conditional probabilities. You can also accomplish this by multiplying the joint probability matrix on the right by a diagonal matrix as follows:

$$\begin{pmatrix} P(B_1 \cap A_1) & P(B_2 \cap A_1) & P(B_3 \cap A_1) \\ P(B_1 \cap A_2) & P(B_2 \cap A_2) & P(B_3 \cap A_2) \end{pmatrix} \begin{pmatrix} 1/P(B_1) & 0 & 0 \\ 0 & 1/P(B_2) & 0 \\ 0 & 0 & 1/P(B_3) \end{pmatrix} =$$

$$\begin{pmatrix} P(A_1|B_1) & P(A_1|B_2) & P(A_1|B_3) \\ P(A_2|B_1) & P(A_2|B_2) & P(A_2|B_3) \end{pmatrix} \tag{6.4}$$

Remark. Equation (6.3) is called *Bayes Law*. It is easy to prove. First, write

$$P(A_i|B_j) = \frac{P(A_i \cap B_j)}{P(B_j)} = \frac{P(A_i)P(B_j|A_i)}{P(B_j)}.$$

Then, substitute for $P(B_j)$ the right side of (6.2).

6.2 Bayes Method Examples

Example 6.1. We consider the discrete channel model with binary input and output and channel matrix

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} \end{array}$$

This is a famous channel model called the *binary symmetric channel* (BSC). The parameter p is called the *crossover probability*. It is the probability that the channel makes an error (in which either a transmitted 0 is received as a 1, or vice-versa). Suppose a binary channel input is selected at random according to the probability model

$$P(input\ = 0) = 0.6, \quad P(input\ = 1) = 0.4.$$

Let us answer the following two questions via Bayes Methodology.

(a) Compute p if it is measured that

$$P(output\ = 0) = 0.59.$$

(b) Compute p if it is measured that

$$P(input\ = 0 | output\ = 0) = 0.97.$$

Solution to (a). The output probability distribution is

$$(0.6,\ 0.4) \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} = (0.6(1-p) + (0.4)p,\ (0.6)p + 0.4(1-p)).$$

Setting

$$0.6(1-p) + (0.4)p = 0.59,$$

we see that $p = 0.05$.

Solution to (b). The joint probability matrix is

$$\begin{pmatrix} 0.6(1-p) & (0.6)p \\ (0.4)p & 0.4(1-p) \end{pmatrix}.$$

If we divide the left column by the column sum of that column, the top left corner of the new column will be

$$\frac{0.6(1-p)}{0.6(1-p) + (0.4)p},$$

which we set equal to the backward conditional probability 0.97. Solving this equation, one obtains $p = 0.0443$.

Example 6.2. At a certain university, the Statistics Department has tried three different texts in Stat 101. These texts are by Professors Mean, Median, and Mode, respectively. Of the 1000 students who took Stat 101, 500 of them used Professor Mean's book, 300 of them used Professor Median's book, and 200 of them used Professor Mode's book. A survey showed that 200 students were satisfied with Mean's book, 150 were satisfied with Median's book, and 160 were satisfied with Mode's book. Given that a randomly selected student was satisfied with his/her textbook, let us determine which of the three textbooks that the student was most likely to have used.

Solution. Consider the following 5 events:

E_1 is the event that the student took Stat 101 using Mean's text
E_2 is the event that the student took Stat 101 using Median's text
E_3 is the event that the student took Stat 101 using Mode's text
S is the event that the student was satisfied with his/her text
NS is the event that the student was not satisfied with his/her text

Here is the forward cond prob matrix:

	S	NS
E_1	0.40	0.60
E_2	0.50	0.50
E_3	0.80	0.20

Multiplying the three rows, respectively, by

$$P(E_1) = 0.50, \quad P(E_2) = 0.30, \quad P(E_3) = 0.20,$$

we obtain the following joint probability matrix:

	S	NS
E_1	0.20	0.30
E_2	0.15	0.15
E_3	0.16	0.04

Finally, dividing each column by the column sum, we obtain the following matrix of backward conditional probabilities:

	S	NS
E_1	20/51	30/49
E_2	15/51	15/49
E_3	16/51	4/49

In the first column, the biggest entry is 20/51, that is,

$$P(E_1|S) = 20/51.$$

We conclude that the student most likely used Mean's text.

Example 6.3. In testing a new drug, 1000 sick people were tested. 600 of the sick people were given the drug and the remaining 400 people were given a placebo. In each case, some of them were cured and some not cured, according to the following table:

$$\begin{array}{l} \\ \textit{given drug} \\ \textit{given placebo} \end{array} \begin{array}{c} \begin{array}{cc} \textit{cured} & \textit{not cured} \end{array} \\ \begin{pmatrix} 450 & 150 \\ 200 & 200 \end{pmatrix} \end{array}$$

Suppose one of the 1000 people is chosen at random, and we want to see what is the probability that this person was in the placebo group given that they are not cured. We obtain the matrix of

forward cond probs by dividing each row of the preceding table by the row sum:

$$\begin{array}{c} \\ \textit{given drug} \\ \textit{given placebo} \end{array} \begin{array}{c} \begin{array}{cc} \textit{cured} & \textit{not cured} \end{array} \\ \begin{pmatrix} 450/600 & 150/600 \\ 200/400 & 200/400 \end{pmatrix} \end{array}$$

To obtain the joint probability matrix, we multiply the first row of the preceding matrix by 0.60 and the second row by 0.40. (This is because the selected person belongs to the placebo group with a probability of 400/1000.) This gives us the following matrix of joint probabilities:

$$\begin{array}{c} \\ \textit{given drug} \\ \textit{given placebo} \end{array} \begin{array}{c} \begin{array}{cc} \textit{cured} & \textit{not cured} \end{array} \\ \begin{pmatrix} 0.45 & 0.15 \\ 0.20 & 0.20 \end{pmatrix} \end{array}$$

Summing the two columns,

$$P(\textit{cured}) = 0.65, \quad P(\textit{not cured}) = 0.35.$$

If we want the matrix of backward cond probs, we can normalize each column of the preceding table by dividing the columns by 0.65, 0.35, respectively.

$$\begin{array}{c} \\ \textit{given drug} \\ \textit{given placebo} \end{array} \begin{array}{c} \begin{array}{cc} \textit{cured} & \textit{not cured} \end{array} \\ \begin{pmatrix} 45/65 & 15/35 \\ 20/65 & 20/35 \end{pmatrix} \end{array}$$

Thus, for example, we can say that

$$P(\textit{given placebo}|\textit{not cured}) = 20/35.$$

6.3 Two Conditional Probability Laws

Multiplication Law

The Multiplication Law says that for any finite number of events $E_1, E_2, \cdots, E_k$, the probability of the intersection can be broken down as a product as follows:

$$P(E_1 \cap E_2 \cap E_3 \cap \cdots \cap E_k) = P(E_1)P(E_2|E_1)P(E_3|E_1 \cap E_2) \cdots P(E_k|E_1 \cap E_2 \cap \cdots \cap E_{k-1}).$$

In other words, after the first factor $P(E_1)$ of the first event E_1 on the right side, the remaining factors are the conditional probabilities of each remaining event given all the previous events in the list $E_1, E_2, \cdots, E_k$. If you have just two events, the multiplication law says

$$P(E_1 \cap E_2) = P(E_1)P(E_2|E_1),$$

which is something we proved in Lecture 5 Notes. For three events, we have

$$P(E_1 \cap E_2 \cap E_3) = P(E_1)P(E_2|E_1)P(E_3|E_1 \cap E_2).$$

The reader can easily write down what the Multiplication Law would mean for four events. It is not hard to see that the Multiplication Law for k events follows from the Multiplication Law for $k-1$ events. For example, we could prove the Multiplication Law for three events as follows:

$$\begin{aligned} P(E_1 \cap E_2 \cap E_3) &= P(E_1 \cap E_2)P(E_3|E_1 \cap E_2) \\ &= P(E_1)P(E_2|E_1)P(E_3|E_1 \cap E_2) \end{aligned}$$

Example 6.4. Suppose an experiment consists of multiple steps. You arrive after the experiment starts and can observe the results of the last few steps. Your goal is to determine what was the most likely result of the earlier steps which you could not observe. Solving a problem like this is a nice application of the multiplication theorem. To illustrate, suppose we have an urn containing 6 red, 6 white, and 6 blue balls. Four balls are selected at random from the urn, one after the other, w/o replacement. It is observed that the second, third, and fourth balls selected are white, white, red, in that order. Given this information, was the first ball most likely to have been red, white, or blue? We let the notation Ri ($i = 1, 2, 3, 4$) denote the event of getting a red ball on draw i. Similarly, we let Wi and Bi denote the events of getting a white ball and a blue ball on draw i, respectively. We want to determine which of the following 3 numbers is the greatest:

$$P(R1|W2 \cap W3 \cap R4), \quad P(W1|W2 \cap W3 \cap R4), \quad P(B1|W2 \cap W3 \cap R4) \tag{6.5}$$

If we multiply these three numbers by $P(W2 \cap W3 \cap R4)$, we obtain the 3 numbers

$$P(R1 \cap W2 \cap W3 \cap R4), \quad P(W1 \cap W2 \cap W3 \cap R4), \quad P(B1 \cap W2 \cap W3 \cap R4). \tag{6.6}$$

Therefore, if we can determine which of the three numbers (6.6) is the greatest, the corresponding number in (6.5) will be the greatest. By the multiplication theorem, we have

$$\begin{aligned} P(R1 \cap W2 \cap W3 \cap R4) &= P(R1)P(W2|R1)P(W3|R1 \cap W2)P(R4|R1 \cap W2 \cap W3) \\ &= (6/18)(6/17)(5/16)(5/15) \\ P(W1 \cap W2 \cap W3 \cap R4) &= P(W1)P(W2|W1)P(W3|W1 \cap W2)P(R4|W1 \cap W2 \cap W3) \\ &= (6/18)(5/17)(4/16)(6/15) \\ P(B1 \cap W2 \cap W3 \cap R4) &= P(B1)P(W2|B1)P(W3|B1 \cap W2)P(R4|B1 \cap W2 \cap W3) \\ &= (6/18)(6/17)(5/16)(6/15) \end{aligned}$$

The third of these numbers is the biggest. Therefore, the first ball selected most likely was blue.

Law of Total Probability

Let $\{A_i\}$ be a partition of the sample space, and let B be any event. The Law of Total Probability says that

$$P(B) = \sum_i P(A_i)P(B|A_i). \tag{6.7}$$

In other words, this law allows you to express the "total probability" $P(B)$ as a weighted average of the conditional probabilities with which B occurs when conditioned on each A_i. The Law on Total Probability is easy to prove. First, we remark that

$$P(B) = \sum_i P(A_i \cap B).$$

This is because B is a disjoint union of the $B \cap A_i$'s. In this preceding sum, just substitute

$$P(A_i \cap B) = P(A_i)P(B|A_i)$$

and you have the result (6.7).

Example 6.5. A prisoner is given 100 black balls and 100 white balls and two identical urns. He is told to distribute the balls in the urns any way he wants. The urns are then taken away and returned to the prisoner at a later time so that the prisoner does not know which urn is which. He is then asked to choose an urn at random and then to select a ball at random from the chosen urn. If the ball is white, the prisoner is allowed to go free. The "prisoner's dilemma" (a classic probability problem) is then to determine how the prisoner should distribute the 200 balls in the two urns in order to maximize his chance of going free. Suppose he chooses x of the 100 black balls and y of the 100 white balls to put in Urn 1 (x, y unknown). The remaining $200 - x - y$ balls go in Urn 2. The event that he selects Urn 1 ($U1$) has prob 1/2 and the event that he selects Urn 2 ($U2$) has prob 1/2. Let W be the event that he selects a white ball. We have, by the Law of Total Probability,

$$\begin{aligned} P(W) &= P(U1)P(W|U1) + P(U2)P(W|U2) \\ &= (1/2)\left(\frac{y}{x+y}\right) + (1/2)\left(\frac{100-y}{200-x-y}\right) \end{aligned}$$

You want this expression to be as big as possible. What should x, y be to achieve this? If you get stuck, consult the Chapter 1 Solved Problems.

Remarks. There are other problems where the Law of Total Probability gives a nice solution. Another such problem is the "Monte Hall Problem." See Recitation 2 Instructions for a description of this problem. We will see a generalization of the Law of Total Problability later on in the course called the "Law of Total Expectation". The Law of Total Expectation will ultimately be an important tool for us when we consider optimal filter design problems.