MARKOV CHAINS: ROOTS, THEORY, AND APPLICATIONS

TIM MARRINAN

1. Introduction

The purpose of this paper is to develop an understanding of the theory underlying Markov chains and the applications that they have. To this end, we will review some basic, relevant probability theory. Then we will progress to the Markov chains themselves, and we will conclude with a case study analysis from two related papers.

2. Review of Probability

2.1. **Initial Definitions.** Markov chain analysis has its roots in probability theory, so we begin with a review of probability. The review will be brief and will focus mainly on the areas of probability theory that are pertinent to Markov chain analysis.

As with any discipline, it is important to be familiar with the language of probability before looking at its applications. Therefore, we will begin with a few definitions and a few more will be introduced later as necessary. In probability, the **sample space**, S, is the set of all possible outcomes for an experiment. Any subset, F, of the sample space S is known as an **event**. For example, Sheldon Ross explains in his text that if the experiment consists of a coin toss, then

$$S = \{(H)eads, (T)ails\}$$

is the sample space [4]. $F = \{H\}$ is the event that the outcome of the flip is heads and $E = \{T\}$ would be the event that the outcome of the toss is tails. Alternatively, if the experiment consists of two successive coin flips, then

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

is the sample space (where (H, T) denotes that the first coin came up heads and the second coin came up tails). $F = \{(H, H)\}$ is the event that both flips came up heads, $E = \{(H, H), (H, T), (T, H)\}$ is the event that heads shows up on at least one of the coins and so on.

The **union** of two events E and F of a sample space S, denoted $E \cup F$, is defined as the set of all outcomes that are in either E or

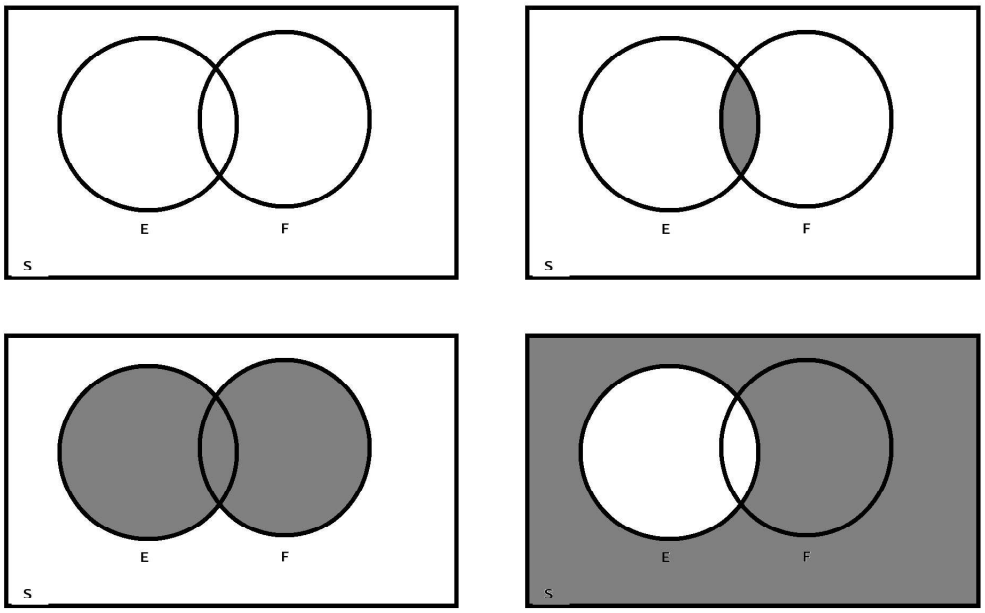

FIGURE 1. Venn Diagrams that represent (clockwise from top left): the events E and F, the intersection $E \cap F$, the complement E^c, and the union $E \cup F$

F or both. The **intersection** of E and F, denoted $E \cap F$, is defined as the outcomes that are in both E and F. The **complement** of an event E is the set of all points in the sample space that are not in E. The complement is written E^c. So if we reexamine the experiment of flipping two coins, then every outcome in either E or F or both is $F \cup E = \{(H,H),(H,T),(T,H)\}$. The only outcome simultaneously in both E and F is $F \cap E = \{(H,H)\}$. Lastly, the set of outcomes in S that are not in E is $E^c = \{(T,T)\}$. These ideas can be represented visually by a Venn diagram. In the Venn diagram we can say that the entire area in the box is the sample space. Then the interior of each of the circles represents an event. The part where the two interiors overlap is the intersection. The area that both the circles enclose together in Figure 1 is the union, and the area that is outside one of the circles is the complement to the event represented by that circle.

2.2. **Probability of Events.** In both of the coin flipping experiments, the experimenter has very little control over the outcome of the flip (although I hear that one of the chemistry professors can make a coin come up heads every time by flipping it in such a way that it doesn't actually rotate end over end, it just oscillates back and forth creating the

illusion that it is rotating [7]). Thus it is often desirable to determine the probability that a specific event or outcome will occur. Probability is essentially the fraction of times that we expect a specific event to occur. Explicitly, we write the **probability** of an event F in the sample space S as $P(F)$, and we assume that $P(F)$ is defined, along with the following three conditions:

(i) $0 \leq P(F) \leq 1$,

(ii) $P(S) = 1$,

(iii) For any sequence of events $F_1, F_2, \ldots, F_n$ that are mutually exclusive (i.e. their intersection is empty), then $P\left(\bigcup_{n=1}^{\infty} F_n\right) = \sum_{n=1}^{\infty} P(F_n)$.

The first of these conditions simply states that the event can't happen less than never or more than every time the experiment is run. The second one notes that the experiment will have some result every time. The third condition tells us that the probabilities of all the events that don't overlap add up to give us the probability that one of these events is going to occur. So with this definition of probability we can look at the first coin flip experiment and say that if the coin is well balanced and the flip is fair, then $P(\{H\}) = \frac{1}{2}$ (If the flipper is that chemistry professor then $P(\{H\}) = 1$). In the second coin flip experiment, with a fair flip the probability that each flip will yield a tail is $P(\{(T,T)\}) = \frac{1}{4}$.

From our definition of probability, we can derive the probabilities of unions and complements as

(i) $P(E \cup F) = P(E) + P(F) - P(E \cap F)$ (look at the Venn Diagram (Figure 1) for visual confirmation of this)

(ii) $P(E^c) = 1 - P(E)$.

2.3. **Conditional Probability.** The idea behind conditional probability is that we sometimes need to figure out how likely it is that an event E will happen assuming or given that another event F happens first. For example, imagine that I have eight tootsie pops, three red, three blue and two brown. Obviously you want one of my tootsie pops, but I say that in order to get one you need to pick, without looking, one that has an Indian shooting a star on it. To give you a better chance at winning the game, I tell you that one of the brown ones has the Indian, two of the red ones have him and one of the blue ones does. Naturally, you would pick one of the red tootsie pops because you have a two

out of three chance of getting to keep the sucker, and that instinctive calculation is an example of conditional probability. The probability of the sucker having an Indian on it at all is $\frac{4}{8} = \frac{1}{2}$, but the probability of the the wrapper having an Indian on it given that it is red is $\frac{2}{3}$. Now we just have to make that instinctive calculation explicit so that we can use it on more complicated examples. If E and F are events from a sample space S, we denote the **conditional probability** of E given F as $P(E|F)$ and it is calculated using the formula

$$P(E|F) = \frac{P(E \cap F)}{P(F)}. \tag{1}$$

As Ross notes, this formula makes the most sense if you think about it as follows. If the event F occurs, then in order for E to also occur the actual occurrence must be a point in the intersection, $E \cap F$ [4]. Also, since we know that F has occurred, it becomes our new sample space so we take the probability of $E \cap F$ relative to the probability of F as is shown in Equation 1. So now we have an equation that requires us to know what $P(E \cap F)$ is. The probability of the intersection of two events is usually not very obvious, but fortunately we can usually figure out the $P(E|F)$ term intuitively just like we did with the tootsie pops. If that is the case then we can rewrite Equation 1 as

$$P(E|F)P(F) = P(E \cap F). \tag{2}$$

With the tootsie pops, E was the event that the wrapper had a picture of an Indian on it, and F was the event that the wrapper was red. Therefore we can calculate the probability of the intersection of these two events as

$$\begin{aligned} P(E \cap F) &= P(E|F)P(F) \\ &= \left(\frac{2}{3}\right)\left(\frac{3}{8}\right) \\ &= \frac{1}{4}. \end{aligned}$$

which makes sense because $E \cap F$ represents the event that the wrapper is both red and has an Indian which happens on two of the eight tootsie pops ($\frac{1}{4}$ of the time). In this example it would have been easy to compute the probability of the intersection directly, but often it is necessary to use this formula.

It might occur that the conditional probability of an event E given an event F is the same as the probability of E by itself. This happens when the events E and F are independent of one another. Events are

defined as **independent** when $P(E \cap F) = P(E)P(F)$, and Equation 1 shows that this implies $P(E|F) = P(E)$ (and also $P(F|E) = P(F)$).

The definition of independence can be extended to more than two events. Ross writes that the events $E_1, E_2, \ldots, E_n$ are said to be independent if for every subset $E_{1'}, E_{2'}, \ldots, E_{r'}$, $(r' \leq n)$, of these events

$$P(E_{1'} \cap E_{2'} \cap \ldots \cap E_{r'}) = P(E_{1'})P(E_{2'}) \ldots P(E_{r'})$$

or the probability of the intersection of all the events in the subset is equal to the product of the probability of each of the events in the subset [4]. Intuitively, the events $E_1, E_2, \ldots, E_n$ are independent if the knowledge of the occurrence of any of these events has no effect on the probability of any other event. Using these newfound manipulations of probability, we can understand an important result known as Bayes' formula.

2.4. **Bayes' Formula.** Suppose that $F_1, F_2, \ldots, F_n$ are mutually exclusive events such that their union is the sample space S. In other words, exactly one of those events will occur. By writing

$$E = \bigcup_{i=1}^{n} (E \cap F_i)$$

and using the fact the the events $E \bigcap F_i, i = 1, 2, \ldots n$, are mutually exclusive, we obtain that

$$P(E) = \sum_{i=1}^{n} P(E \cap F_i) \tag{3}$$

$$= \sum_{i=1}^{n} P(E|F_i)P(F_i). \tag{4}$$

Thus, Equation 4 shows how, for given events $F_1, F_2, \ldots, F_n$ of which one and only one can occur, we can compute $P(E)$ by first "conditioning" upon which one of the F_i occurs. That is, it states that $P(E)$ is equal to a weighted average of $P(E|F_i)$, each term being weighted by the probability of the event on which it is conditioned.

Suppose now that E has occurred and we are interested in determining which one of the F_j also occurred. By Equation 4 we have that

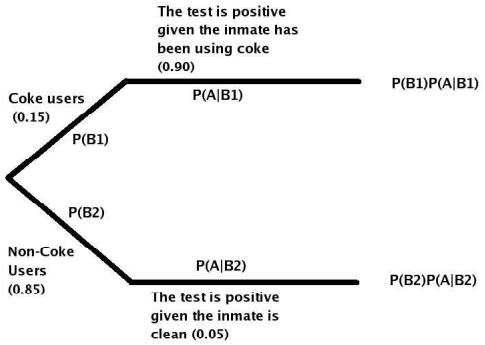

FIGURE 2. A tree diagram to help visualize the problem

$$P(F_j|E) = \frac{P(E \cap F_j)}{P(E)} \tag{5}$$

$$P(F_j|E) = \frac{P(E|F_j)P(F_j)}{\sum_{i=1}^{n} P(E|F_i)P(F_i)}. \tag{6}$$

We call Equation 6 Bayes' formula.

To help put Bayes' formula into some context, let's look at an example. Imagine that the Washington State Penitentiary is doing random drug testing among its inmates. They are testing urine for cocaine and the specific test that they are using accurately comes up positive for coke 90% of the time, but it also shows a false positive 5% of the time when no cocaine is actually present. If 15% of the inmates in the prison have been using cocaine, what is the probability that any random inmate will test positive, and what is the probability that if an inmate tested positive they were actually clean?

We will call the event that an inmate uses cocaine B_1, the event that the inmate doesn't use cocaine B_2, and the event that the test is positive A. These events and their probabilities can be visualized using the tree diagram in Figure 2. Then from Equation 4 the probability that the test comes up positive is

$$\begin{aligned} P(A) &= P(B_1)P(A|B_1) + P(B_2)P(A|B_2) \\ &= (0.15)(0.90) + (0.85)(0.05) \\ &= 0.1775, \end{aligned}$$

and the probability that the inmate was actually clean given that they tested positive is

$$\begin{aligned} P(B_2|A) &= \frac{P(B_2)P(A|B_2)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2)} \\ &= \frac{(0.85)(0.05)}{(0.15)(0.90) + (0.85)(0.05)} \\ &= 0.2394. \end{aligned}$$

2.5. **Random Variables.** Bayes' formula is very useful for working with available data to find the probability of specific outcomes, but sometimes it is more valuable in the grand scheme of things to analyze functions of outcomes instead of the specific outcomes. Sheldon Ross explains the idea with this brief example, "in tossing dice we are often interested in the sum of the two dice and are not really concerned about that actual outcome [4]". In other words, we may be interested in knowing that the sum is seven and not be concerned over whether the actual outcome was $(1, 6)$ or $(2, 5)$, etc. These quantities of interest, or more formally, these real-valued functions defined on the sample space, are known as **random variables**.

Since all of the possible values of these random variables are outcomes of the experiment, it is possible to assign probabilities to the random variables based on the outcomes that they include. Let's say that we have two egg cartons in the refrigerator that each hold four eggs and we want to whip up a tasty scramble. Assume that if there are fewer than three eggs, it isn't worth our time to make the scramble. Since we are playing NBA Hangtime and on a 21 game win streak, we don't want to check the frig unless we have a 60% chance of finding four or more eggs. What is the probability that there are at least four eggs in the refrigerator?

To solve our pickle, we can create a random variable. We need the further assumption that the number of eggs in the cartons is completely random and there is an equally likely chance of any number being there. This is a safe assumption because in this example we live in a house with seven other college-age guys who could have eaten out for the past week or come in late last night and made a seven-egg omelet with equal likelihood. Random variables are often represented with a capital letter, so if we let X represent a random variable that is defined as the total number of eggs in the refrigerator then we can find the probability that there are four or more eggs left, $P(X > 3)$.

If we let $P\{(1, 1)\}$ mean that there is one egg in the first carton and one in the second, we can spell out the probabilities so that

$$
\begin{aligned}
(7) \qquad P(X=0) &= P\{(0,0)\} = \frac{1}{25} \\
P(X=1) &= P\{(1,0),(0,1)\} = \frac{2}{25} \\
P(X=2) &= P\{(2,0),(1,1),(0,2)\} = \frac{3}{25} \\
P(X=3) &= P\{(3,0),(2,1),(1,2),(0,3)\} = \frac{4}{25} \\
P(X=4) &= P\{(4,0),(1,3),(2,2),(3,1),(4,0)\} = \frac{5}{25} \\
P(X=5) &= P\{(4,1),(3,2),(2,3),(4,1)\} = \frac{4}{25} \\
P(X=6) &= P\{(4,2),(3,3),(2,4)\} = \frac{3}{25} \\
P(X=7) &= P\{(4,3),(3,4)\} = \frac{2}{25} \\
P(X=8) &= P\{(4,4)\} = \frac{1}{25}.
\end{aligned}
$$

Since we want to know if there are at least four eggs,

$$
\begin{aligned}
P(X>3) &= P(X=4 \cup X=5 \cup X=6 \cup X=7 \cup X=8) \\
&= P(X=4)+P(X=5)+P(X=6)+P(X=7)+P(X=8) \\
&= \frac{15}{25},
\end{aligned}
$$

and since this is equivalent to 60%, the probability of having eggs is large enough for us to get up and try to make some food.

Ordinarily, when taking the probability of a random variable we look at equations in the form of $P(X=a)$. In the last example, our equation was of the form $P(X>a)$ so we intuitively used what is known as a **cumulative distribution function** to find the answer. The cumulative distribution function, F, of a random variable X is defined for any real number b $(-\infty < b < \infty)$, by $F(b) = P\{X \leq b\}$. Some properties of the cumulative distribution function F are

$$
\begin{aligned}
(i) &\quad F(b) \text{ is a non-decreasing function of } b \\
(ii) &\quad \lim_{b\to\infty} F(b) = 1 \\
(iii) &\quad \lim_{b\to-\infty} F(b) = 0.
\end{aligned}
$$

This means F is a function whose output is the probability of the random variable being less than or equal to a value, b. We indirectly

used a cumulative distribution function to find the likelihood of their being eggs in the refrigerator. First we found the probability that there was less than or equal to three eggs. Then we interpreted part (ii) of the definition of a probability to say that the probability of there not being less than or equal to three eggs was $1 - P(X \leq 3)$. Since $1 - P(X \leq 3)$ is the same as $P(X > 3)$, we can use a cumulative distribution function just as we did.

It is worth noting that in the egg example, we had a finite number of outcomes for the random variable. This is not always the case. When the random variable has a finite (or countably infinite) number of possible values, it is called a **discrete** random variable, and when the possible values are uncountable it is known as a **continuous** random variable. The cumulative distribution function can be used to represent both types of random variables and all probability questions about a random variable X can be answered in terms of the cumulative distribution function, F. For example,

$$P(a < X \leq b) = F(b) - F(a) \qquad \text{for all } a < b.$$

The cumulative distribution function is not the only way of gaining information about a random variable. For a discrete random variable X we define the **probability mass function**, $p(a)$, of X by

$$p(a) = P(X = a).$$

This function tells us the probability of our random variable taking on a specific value. Equation 7 essentially gives the probability mass function for the random variable in that example. Since the random variable is discrete, there are only a countable number of values for a such that $p(a)$ is positive. If we let x_i represent these values for $i = 1, 2, \ldots$ then we can say

$$\sum_{i=1}^{\infty} p(x_i) = 1,$$

and we can write the equation for the cumulative distribution function in terms of the probability mass function so

$$F(a) = \sum_{x_i \leq a} p(x_i).$$

To help visualize the probability mass function and the cumulative distribution function let's look at a brief example. Say that Y is a discrete random variable that is defined to be the number of pieces of candy you pick up in one try of the crane game that they often have

at grocery stores and pizza places. Let us also say that Y has the probability mass function

$$p(0) = 1/2, \quad p(1) = 1/4, \quad p(2) = 1/6, \quad p(3) = 1/12,$$

and nobody ever wins more than three pieces of candy. Then we graph $p(y)$ as can be seen in Figure 3.

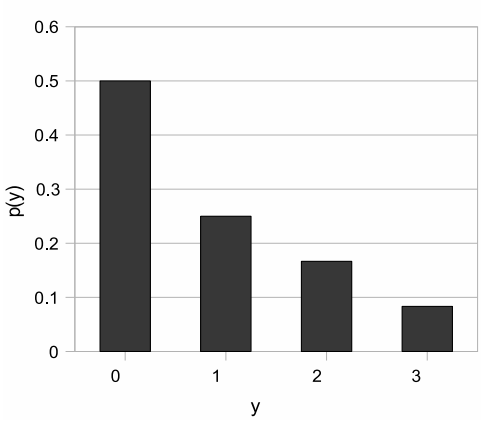

FIGURE 3. Graph of $p(y)$

Then from this probability mass function we can assemble the cumulative distribution function

$$F(y) = \begin{cases} 1/2 & y < 1 \\ 3/4 & 1 \leq y < 2 \\ 11/12 & 2 \leq y < 3 \\ 1 & 3 \leq y, \end{cases}$$

and represent it graphically as in Figure 4.

The probability mass function and the cumulative distribution as we used them work only for discrete random variables. However, they both have analogous functions for continuous random variables. We have not seen any continuous random variables in the examples so far but think of them this way: If we were to make a random variable X that was defined to be the height of every person in the world it would be convenient to group people into categories such as 5 feet 6 inches, 5 feet 7 inches, 5 feet 8 inches, etc. because we usually only measure people to the nearest inch. In this case, Y would be a discrete random variable, because there would be a countable number of possible heights. However, if we could measure height exactly, then we would find that nobody is the same height and that there are an infinite number of different heights. In this scenario, Y would be a continuous random variable.

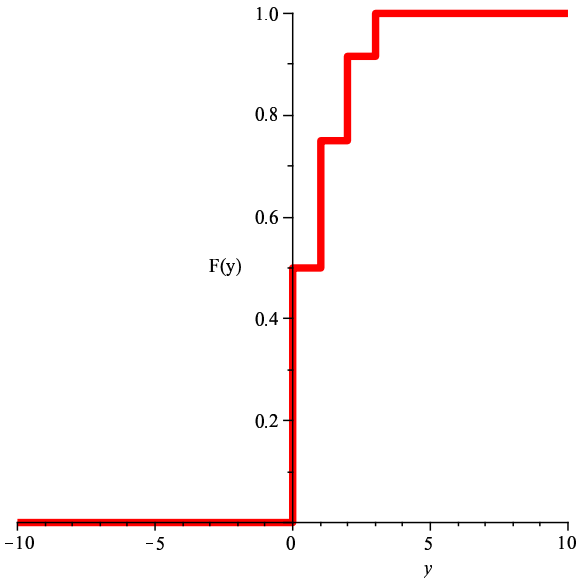

FIGURE 4. Graph of $F(y)$

Formally, we call a random variable, X, a continuous random variable if there exists a non-negative function, $f(x)$, defined for all real numbers x in the set $(-\infty, \infty)$, having the property that for any set B of real numbers $P\{X \epsilon B\} = \int_B f(x)dx$. The function $f(x)$ is called the **probability density function** of the random variable X, and is the continuous random variable's version of the probability mass function. As such, all probability statements about X can be answered in terms of $f(x)$. If B is the interval $[a, b]$ then

$$P(a \leq X \leq b) = \int_a^b f(x)dx,$$

and if we let $a = b$ then

$$\begin{aligned} P(X = a) &= \int_a^a f(x)dx \\ &= F(a) - F(a) \\ &= 0. \end{aligned}$$

So we see that in the continuous case, the probability of achieving any particular value is zero (because there are uncountably many) [4].

In the continuous case, the cumulative distribution function behaves the same as it does in the discrete case. From the preceding equations you may guess how it differs from the function in the discrete case. Since the probability density function is a continuous function, $F(a)$

becomes an integral from $-\infty$ to a instead of a summation from 0 to a and it is represented graphically as a smooth curve instead of the step-wise function that is shown in Figure 4.

In this section, there has been much talk about random variables without many examples of actual random variables in action. For the purpose of this paper, it suffices to say that there are some common random variables of both the discrete and continuous types that appear frequently in mathematics and daily life. If they come up they will be introduced, but for those who are curious please refer to Sheldon Ross' excellent text on probability models [4].

2.6. **Expectation of a Random Variable.** Remember back to the example where we were trying to win candy in the crane game. Using the probability mass function, we were able to determine the probability of each of the possible outcomes; the probability of getting zero, one, two or three pieces of candy. What if we were instead concerned with how many pieces of candy we could actually expect to get on one quarter? The most natural way to figure this out would be to take a weighted average; to multiply each outcome by its individual probability and add up those values. Conveniently, this is precisely how we compute expected value in probability. The **expected value** of a discrete random variable, X, that has the probability mass function, $p(x)$, is

$$E[X] = \sum_{x} xp(x). \tag{8}$$

The expected value is also sometimes referred to as the mean, in which case it is represented by μ. If we spend our one quarter on the crane game, then we can expect to get

$$\begin{aligned} E[X] &= \sum_{x=0}^{3} xp(x) \\ &= (0)\left(\frac{1}{2}\right) + (1)\left(\frac{1}{4}\right) + (2)\left(\frac{1}{6}\right) + (3)\left(\frac{1}{12}\right) \\ &= 0 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} \\ &= \frac{13}{12} \text{ pieces of candy.} \end{aligned}$$

To adapt this function for the continuous case we need to replace the probability mass function, $p(x)$, of the discrete random variable with the probability density function, $f(x)$, of the continuous random

variable. Then we need to change the summation to integration to ensure continuity. The new equation is

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx \tag{9}$$

for a continuous random variable.

Both discrete and continuous random variables were broadly described as real-valued functions defined on the sample space. This definition seems to imply that we should be able to create new random variables that are functions of the old random variables (as long as they are still real-valued and defined on the sample space). This turns out to be true, and as a result we can find the expected value of functions of random variables.

Suppose $g(X)$ is some function of the random variable X, then $g(X)$ is itself a random variable and its expected value is

$$\begin{aligned} E[g(X)] &= \sum_x g(x)p(x) \text{ if } X \text{ is discrete} \\ &\text{or} \\ E[g(X)] &= \int_{-\infty}^{\infty} g(x)f(x)dx \text{ if } X \text{ is continuous.} \end{aligned}$$

2.7. **Jointly Distributed Random Variables.** Sometimes we can gain more interesting information from our data if we look at probability statements concerning two or more random variables instead of single ones. This will become more apparent when we begin our discussion of Markov chains. To deal with such probabilities, Ross defines, for any two random variables X and Y, the **joint cumulative probability distribution function** of X and Y by

$$F(a,b) = P(X \leq a, Y \leq b), \quad -\infty < a,\ b < \infty \text{ [4].} \tag{10}$$

Jointly distributed random variables behave much like the single ones do, with the difference being that operations with two random variables need a double integral or a double summation. For instance, a probability for jointly distributed random variables X and Y can be represented as

$$\begin{aligned} P(a \leq X \leq b, c \leq Y \leq d) &= \sum_{X=a}^{b} \sum_{Y=c}^{d} p(x,y) \\ &\text{where } p(x,y) \text{ is the joint probability mass function, or} \\ P(X \in A, Y \in B) &= \int_B \int_A f(x,y)dxdy \\ &\text{where } f(x,y) \text{ is the joint probability density function.} \end{aligned}$$

The analogous formulas for expectation in the case of jointly distributed random variables are

$$\begin{aligned} E[g(X,Y)] &= \sum_y \sum_x g(x,y)p(x,y), \\ \text{and} & \\ E[g(X,Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dx. \end{aligned}$$

If it turns out that you actually wanted the distribution of one variable, but are stuck with two, you can find distribution of one, (X in this case), as follows:

$$\begin{aligned} F_X(a) &= P(X \leq a) \\ &= P(X \leq a, Y < \infty) \\ &= F(a, \infty). \end{aligned}$$

Similarly the cumulative distribution of Y is given by $F_Y(b) = F(\infty, b)$.

To get a better idea of what it means for random variables to be jointly distributed, let us look at an example that Ross presents on pages 50 and 51 of his text. Say that at a party, N men throw their hats into the center of a room. The hats are mixed up and each man randomly selects one. Let X denotes the number of men that select their own hats. Then we can use jointly distributed random variables to find the expected number of men who select their own hats. $E[X]$ can be best computed by noting that

$$X = X_1 + X_2 + \cdots + X_N,$$

where

$$X_i = \begin{cases} 1, & \text{if the } i\text{th man selects his own hat} \\ 0, & \text{otherwise.} \end{cases}$$

Now, because the ith man is equally likely to select any of the N hats, it follows that

$$P(X_i = 1) = P(i\text{th man selects his own hat}) = \frac{1}{N},$$

and so

$$E[X_i] = (1)P(X_i = 1) + (0)P(X_i = 0) = \frac{1}{N}$$

from the definition of expected value. Hence, we obtain that

$$\begin{aligned} E[X] &= E[X_1 + X_2 + \cdots + X_N] \\ &= E[X_1] + E[X_2] + \cdots + E[X_N] \\ &= \left(\frac{1}{N}\right) N \\ &= 1. \end{aligned}$$

So no matter how many men are at the party, Ross shows us that on the average exactly one of the men will select his own hat [4]. Many texts on probability transition from random variables to a discussion of a group of limit theorems that are very important to probability theory. These theorems give us approximations for the expected value and distribution of our random variables as the number of random variables that we have goes to infinity. For our purposes, this discussion will bog us down so readers are referred to Ross [4] or Samuel Karlin's text "A First Course in Stochastic Processes" [3] for further information.

2.8. **Conditional Probability and Conditional Expectation.** It was previously defined that for any two events E and F, the conditional probability of E given F was $P(E|F) = \frac{P(E \cap F)}{P(F)}$ as long as $P(F) > 0$. Now conditional probability will be used with the discrete random variables X and Y to define the **conditional probability mass function** of X given that $Y = y$, by

$$\begin{aligned} p_{X|Y}(x|y) &= P(X = x|Y = y) \\ &= \frac{P(X = x, Y = y)}{P(Y = y)} \\ &= \frac{p(x, y)}{p_Y(y)}, \end{aligned}$$

for all values of y such that $P\{Y = y\} > 0$. Following from this definition of the conditional probability mass function, it is easy to make the jump to the conditional probability distribution function and the conditional expectation. The **conditional probability distribution function** of X given that $Y = y$ is defined, for all y such that $P\{Y = y\} > 0$, by

$$\begin{aligned} F_{X|Y}(x|y) &= P(X \leq x|Y = y) \\ &= \sum_{a \leq x} p_{X|Y}(a|y). \end{aligned}$$

Finally, the **conditional expectation** of X given that $Y = y$ is defined by

$$\begin{aligned} E[X|Y = y] &= \sum_x xP(X = x|Y = x) \\ &= \sum_x xp_{X|Y}(x|y). \end{aligned}$$

These definitions say in essence exactly the same thing as before with the exception that everything is now conditional on the event that $Y = y$. If X is independent of Y, then these equations are the same as the unconditional ones. If X is independent of Y, then

$$\begin{aligned} p_{X|Y}(x|y) &= P(X = x|Y = y) \\ &= P(X = x). \end{aligned}$$

It is important to note that conditional expectations possess all of the properties of ordinary expectations. After looking at the conditional probabilities in the discrete case, the transition to conditional probabilities for the continuous case is very smooth. The **conditional probability density function** of the continuous random variable X given that the continuous random variable $Y = y$ is written as

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

when the random variables have the joint probability density function $f(x, y)$ and the conditional probability density function is defined for all values of y such that $f_Y(y) > 0$. Also the conditional expectation in the continuous case, defined for all values of y such that $f_Y(y) > 0$, is

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

Using these definitions Ross goes on to prove a result that allows the expectation of an unconditional random variable, X, to be found using the conditional expectation of X given that the random variable $Y = y$. In the text Ross proves for the discrete case that

$$E[X] = E\big[E[X|Y]\big] \tag{11}$$

as follows:

$$\begin{aligned} E\big[E[X|Y]\big] &= \sum_y E[X|Y=y]P(Y=y) \\ \sum_y E[X|Y=y]P(Y=y) &= \sum_y \sum_x xP(X=x|Y=y)P(Y=y) \\ &= \sum_y \sum_x x\frac{P(X=x,Y=y)}{P(Y=y)}P(Y=y) \\ &= \sum_y \sum_x xP(X=x,Y=y) \\ &= \sum_x x\sum_y P(X=x,Y=y) \\ &= \sum_x xP(X=x) \\ &= E[X] \quad [4]. \end{aligned}$$

This result is to be referred to as the **conditional expectation identity**, and we can begin to see its use in the following example which also comes from Ross' text [4].

Say that a crotchety old prospector is holed up in his personal mine, when suddenly his canary dies from the noxious fumes. The prospector realizes that he too will die if he doesn't get out of the mine in time. The fumes are starting to affect his brain, and the years of toiling beneath the earth left him with a slight case of dementia so the prospector is unsure which way will lead him to safety. The room he is in has three doors. Although he does not know it, the first door leads to a tunnel that will take him to safety after two hours of travel. The second door leads to the gold pit that will bring him back to his original location after three hours. The third door leads to the furnace room and back to his original location after five hours. If we assume that in his confused state the prospector is equally likely to choose any one of the doors each time he is in that room, we can use the conditional expectation identity to find the expected length of time until the miner reaches safety. For a picture of the prospector's predicament, see Figure 5.

To set up the problem, let X denote the time until the miner reaches safety, and let Y denote the door he initially chooses. This gives us:

$$\begin{aligned} E[X] &= E[X|Y=1]P(Y=1) + E[X|Y=2]P(Y=2) \\ &\quad +E[X|Y=3]P(Y=3) \\ &= \frac{1}{3}(E[X|Y=1] + E[X|Y=2] + E[X|Y=3]). \end{aligned} \tag{12}$$

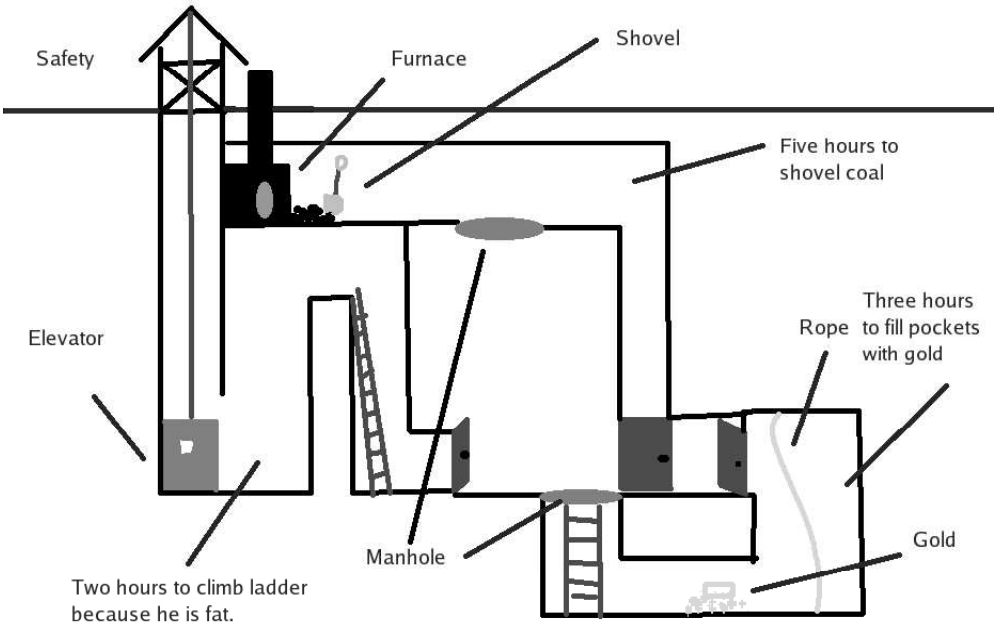

FIGURE 5. Diagram of the miner's escape options

However, we can also write

$$\begin{aligned} E[X|Y=1] &= 2, \\ E[X|Y=2] &= 3+E[X], \\ E[X|Y=3] &= 5+E[X], \end{aligned}$$

because if the prospector chooses the second door, he spends three hours in that tunnel and then returns to the main room. But once he returns to the mine the problem is as before, and hence his expected time until safety is just $E[X]$. This makes $E[X|Y=2] = 3+E[X]$, and the same argument works for $E[X|Y=3]$. If we substitute these equalities into Equation 12, we get

$$\begin{aligned} E[X] &= \frac{1}{3}(2+3+E[X]+5+E[X]) \\ (3)(E[X]) &= 10+(2)(E[X]) \\ E[X] &= 10 \text{ hours until he reaches safety.} \end{aligned}$$

2.9. **Computing Probabilities by Conditioning.** Not only can we obtain expectations by first conditioning on an appropriate random variable, but we can also use this approach to compute probabilities. To see this, let E denote an arbitrary event and define the indicator

random variable, X, by

$$X = \begin{cases} 1, & \text{if } E \text{ occurs} \\ 0, & \text{if } E \text{ does not occur} \end{cases}$$

It follows from the definition of X that

$$\begin{aligned} E[X] &= P(E), \\ E[X|Y=y] &= P(E|Y=y), \text{ for any random variable } Y. \end{aligned}$$

Therefore, we obtain that

$$\begin{aligned} P(E) &= \sum_y P(E|Y=y)P(Y=y), \text{ if } Y \text{ is discrete} \\ &= \int_{-\infty}^{\infty} P(E|Y=y)f_Y(y)dy, \text{ if } Y \text{ is continuous.} \end{aligned} \tag{13}$$

2.10. **Applications of Conditional Probability.** This method for computing probability using conditioning is quite useful in a number of real world applications. One such application is to a list model. Imagine that we have a stack of reference books that are available for borrowing. At each unit of time a book is randomly selected and then is returned to the top of the stack. The probability that any individual book is requested may not be known, but if we are going to start at the top of the stack and look down until we find the requested book, it might be useful to know the expected position of the book in the stack. We can compute this by conditioning on which book is selected.

Another application given in Ross' text is to Ploya's Urn Model [4]. Before we examine that general model, suppose that a coin may be chosen at random from a huge bin of coins representing a uniform spread over all possible values of p, the coin's probability of coming up heads. The chosen coin is then flipped n times. That experiment is a specific example of the following model.

Let's suppose that n independent trials, each of which is a success with probability of p, are performed. Let us compute the conditional probability that the $(r+1)$st trial will result in a success given a total of k success (and $r-k$ failures) in the first r trials. However, let us now suppose that whereas the trials all have the same success probability p, its value is not predetermined but is chosen according to a uniform distribution on $(0,1)$. In this case, by conditioning on the actual value

of p, we have that

$$P\{(r=1)\text{st trial is a success}|k \text{ successes in first } r \text{ trials}\}$$
$$= \frac{P\{(r=1)\text{st trial is a success}, k \text{ successes in first } r \text{ trials}\}}{P\{k \text{ successes in first } r \text{ trials}\}}$$
$$= \frac{\int_0^1 P\{(r+1)\text{st trial is a success}, k \text{ successes in first } r|p\}dp}{1/(r+1)}$$

(To reach the next step, we used the standard representation of a binomial distribution. For an explanation of this distribution, please refer to Ross [4].)

$$= (r+1)\int_0^1 \binom{r}{k} p^{k=1}(1-p)^{r-k}dp$$
$$= (r+1)\binom{r}{k}\frac{(k+1)!(r-k)!}{(r+2)!}$$
$$= \frac{k+1}{r+2}.$$

This says that if the first r trials result in k successes, then the next trial will be a success with probability $(k+1)/(r+2)$.

Using this equation we can describe the model as follows: There is an urn which initially contains one white and one black ball. At each stage a ball is randomly drawn and is then replaced along with another ball of the same color. Thus, for instance, if of the first r balls drawn, k were white, then the urn at the time of the $(r+1)$st draw would consist of $k+1$ white and $r-k+1$ black, and thus the next ball would be white with probability $(k+1)/(r+2)$. If we identify the drawing of a white ball with a successful trial, then we see that this yields an alternate description of the original model. This description is what Ross and others are referring to when they talk about "Ploya's Urn Model" [4].

Using a generalization of the preceding application to situations in which each trial has more than two possible outcomes, we can derive the Bose-Einstein Distribution. In it we suppose that n independent trials, each resulting in one of m possible outcomes $1, \ldots, m$ with respective probabilities $p_1, \ldots, p_m$, are performed. For example, using another urn model, consider that this time the urn starts with one of each of m types of balls. Balls are then randomly drawn and are replaced along with another of the same type. Hence, if in the first n drawings there have been a total of x_j type j balls drawn, then the urn immediately

before the $(n+1)$st draw will contain $x_j + 1$ type j balls out of a total of $m+n$, and so the probability of a type j on the $(n+1)$st draw will be given by $\frac{x_j+1}{n+m}$ according to the Bose-Einstein Distribution which will not be derived here. According to Ross, the Bose-Einstein Distribution basically says that for one of these situations, each possible outcome is equally likely [4].

3. Markov Chains

3.1. **Initial Definitions.** From the idea of jointly distributed random variables, we get the following definition. A **stochastic process** $\{X(t), t\epsilon T\}$ is a collection of random variables. That is, for each $t\epsilon T$, $X(t)$ is a random variable. The index t is often interpreted as time and, as a result, we refer to $X(t)$ as the state of the process at time t. The set T is called the index set of the process. If T is countable then the stochastic process is said to be a discrete-time process and if T is an interval of the real line, then the stochastic process is said to be a continuous-time process. The term stochastic process is used in the definition of a Markov chain, but the ideas of states and their associated times will become clearer when we see some examples of Markov chains.

According to Wolfram's MathWorld, a **Markov Chain** is a stochastic process $\{X_n, n = 0, 1, 2, \ldots\}$ having the property that given the present state, the future is conditionally independent of the past [6]. This stochastic process takes on a finite or countable number of possible values or states. Unless otherwise noted, this set of possible values will be written as the set of nonnegative integers $\{1, 2, 3, \ldots\}$. If $X_n = i$, then the process is said to be in state i at time n. If we suppose that whenever the process is in state i, there is a fixed probability P_{ij} that it will be in state j next, then we can write the defining property of Markov chains as

$$P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0\} = P_{ij} \tag{14}$$

for all states $i_0, i_1, \ldots, i_{n-1}, i, j$ and all $n \geq 0$. Since probabilities are nonnegative and the process must make a transition into some state, we have that

$$\begin{aligned} P_{ij} &\geq 0, \text{ for } i, j \geq 0 \text{ and} \\ \sum_{j=0}^{\infty} P_{ij} &= 1, \text{ for } i = 0, 1, \ldots \end{aligned}$$

The basic idea behind Markov chains can be explained with a simplistic model for weather prediction. Suppose that every day can be classified as either rainy or sunny, then those would be the two states of our stochastic process. In order for this process to be a Markov chain with those states, we would have to be able to say that the probability that it would rain or be sunny tomorrow depended only on whether it was rainy or sunny today. In that case, our model would be a Markov chain and we would assemble the different probabilities into a transition matrix.

For any Markov chain, we will let $\mathbf{P}$ represent the matrix of one-step transition probabilities P_{ij}, so that

$$\mathbf{P} = \begin{Vmatrix} P_{00} & P_{01} & P_{02} & \dots \\ P_{10} & P_{11} & P_{12} & \dots \\ \vdots & \vdots & \vdots & \\ P_{i0} & P_{i1} & P_{i2} & \dots \\ \vdots & \vdots & \vdots & \end{Vmatrix}.$$

To help understand how the matrix $\mathbf{P}$ works, let's look at an example that is adapted from Ross' text [4]. Suppose that the Whitman men's soccer team is playing in a tournament. In that tournament, the chance of them winning their next game depends only upon whether or not they win their current game. Suppose also that if they win their current game, then they will win their next game with probability 0.4; and if they do not win their current game then they will win the next one with probability 0.2.

If we say that the process is in state 0 when they are winning and state 1 when they are losing, then the preceding is a two-state Markov chain whose transition probabilities are given by

$$\mathbf{P} = \begin{Vmatrix} 0.4 & 0.6 \\ 0.2 & 0.8 \end{Vmatrix}.$$

If we examine $\mathbf{P}$, we can see that the rows of the matrix represent the state that the team is currently in (winning or losing their current game), and the columns represent the state that the are going to (winning or losing their next game). So if we want to know the probability that the team loses their next game given that they lose their current game, we see that entry $P_{11} = 0.8$, which is the probability of that event. It is also important to note that the sum of the probabilities along any row of $\mathbf{P}$ is equal 1. This makes sense because it means that from any given state, the team either stays in that same state or moves

to the other one. Probabilities in this example don't come from actual data.

We have already defined the one-step transition probabilities P_{ij}. We now define the **n-step transition probabilities** P_{ij}^n to be the probability that a process proceeds from state i to state j over n additional steps. Symbolically this is written as

$$P_{ij}^n = P\{X_{n+k} = j | X_k = i\}, \text{ for } n \geq 0, i, j \geq 0.$$

P_{ij}^1 is equivalent to P_{ij} and using that equivalency, the Chapman-Kolmogorov equations provide a method for computing the n-step transition probabilities. These equations are

$$P_{ij}^{n+m} = \sum_{k=0}^{\infty} P_{ik}^n P_{kj}^m \text{ for all } n, m \geq 0, \text{ and for all } i, j. \tag{15}$$

These equations are most easily understood by noting that $P_{ik}^n P_{kj}^m$ represents the probability that starting in i, the process will go to state j in $n + m$ transitions through a path which takes it into state k at the nth transition. Then when you sum over all intermediate states k, you end up with the probability that the process will be in state j after $n + m$ one-step transitions [4].

If we tie this idea of n-step transitional probability into the previous example, we could use it to find the probability that the soccer team will win their game after next (their 2nd game from now) given that they win their current game. The probability that we want to find is P_{00}^2. So

$$\mathbf{P}^2 = \mathbf{P} * \mathbf{P} = \begin{Vmatrix} 0.4 & 0.6 \\ 0.2 & 0.8 \end{Vmatrix} * \begin{Vmatrix} 0.4 & 0.6 \\ 0.2 & 0.8 \end{Vmatrix} = \begin{Vmatrix} 0.36 & 0.64 \\ 0.24 & 0.76 \end{Vmatrix}$$

which tells us that they will win their second game from now with probability $P_{00}^2 = 0.36$.

Examples of Markov chains are not always obvious at first glance. Thus, one important concept is how to turn things that don't look like Markov chains into states that are suitable for treatment as a Markov chain. The states of most Markov chains can be defined intuitively. Sometimes, however, the probability of moving from one of these "states" to another depends on more than the current state. This means we haven't successfully created a Markov chains and we need to create states that encompass more than one of these intuitive states in order for the Markov property to apply. For example, if we continue with the soccer team example from earlier, but now we say that the probability of the team winning their next game depends on the outcome of their current game and the one before it. This might be

a more realistic model for winning probabilities because it would take into account hot streaks. But, if we define the states incorrectly then it won't be a Markov Chain. In order for it to still follow the Markov property, we need to define the states as

state 0 if they win both their current game and won the one before it,
state 1 if they win their current game but did not win the one before it,
state 2 if they did not win their current game but they won the one before it, and
state 3 if they lost both their current game and the one before it.

Then we can assign probabilities based on the outcome of the last two games. Let's say that they win with a probability of 0.6 if they won their last two games, with a probability of 0.4 if they win their current game but did not win their last game, with a probability of 0.3 if they lose their current game, but won the one previous to that, and with a probability of 0.2 if they lost their last two games. Using the above states gives us a transition probability matrix of

$$\mathbf{P} = \begin{Vmatrix} 0.6 & 0 & 0.4 & 0 \\ 0.4 & 0 & 0.6 & 0 \\ 0 & 0.3 & 0 & 0.7 \\ 0 & 0.2 & 0 & 0.8 \end{Vmatrix},$$

where the zero entries tell us that it is impossible to make that one step transition. For example, the team can not move from state 0 to state 1, because of how the states were defined so that transition has a probability of 0. This is very similar to an example that Ross discusses in his text [4].

3.2. **Definitions About States.** We are going to need a larger vocabulary to delve further into the interesting parts of Markov chain theory, so we will look at a few definitions that all come from Ross [4]. In a Markov chain or other stochastic process, state j is said to be **accessible** from state i if $P_{ij}^n > 0$ for some $n \geq 0$. This implies that state j is accessible from state i if and only if starting in i, it is possible that the process will ever enter state j. If two states are accessible to each other they are said to **communicate**, which we write as $i \leftrightarrow j$. Additionally, the relation of communication satisfies the following three

properties:

(i) State i communicates with state i, for all $i \geq 0$.

(ii) If state i communicates with state j, then state j communicates with state i.

(iii) If state i communicates with state j, and state j communicates with state k, then state i communicates with state k.

According to these properties, all of the states in the last soccer team example communicate. Even though you can't get from state 0 to state 3 in one step, you could potentially go from state 0 to state 2, and from state 2 to state 3. Similar arguments apply for the communication of the other states. If two states in a Markov chain communicate, they are said to be in the same **class**, and if the chain has only one class it is said to be **irreducible**. (Note that classes must be either identical or disjoint. They cannot overlap.) Since all the states in the last example communicate, they are all in the same class, and therefore that Markov chain is irreducible.

The term irreducible refers to the fact that if there is more than one class in a Markov chain, we can essentially think of the classes as individual states. As we defined earlier, one state can be accessible from another one without the two actually communicating, so it is possible that one class is accessible from another. As long as the two do not communicate they are still separate classes. Thus, since you cannot travel freely between them, you can get stuck in one of the classes and your movement can be restricted to the states in that class. For this reason the entire class is sometimes called an absorbing state, where once entered there is a probability of one that you will stay in that state.

A state i is said to be **recurrent** if and only if there is a probability of 1 that a process, starting in state i, will at some point return to state i. And finally, in a related definition, a state i is said to be **transient** if the probability that the process, starting in state i, will at some point return to state i is less than 1. In the terms that we have already defined, for a state i to be transient there exists another state j that is accessible from i, but the two states do not communicate and therefore are not in the same class.

3.3. **Limiting Probabilities.** From the first example about the Whitman men's soccer team, the 1-step transition matrix was

$$\mathbf{P} = \begin{Vmatrix} 0.4 & 0.6 \\ 0.2 & 0.8 \end{Vmatrix}.$$

When talking about n-step transition probabilities, we showed how this matrix could be used to find the 2-step transition probabilities. We then generalized this technique for transitions of n steps using the Chapman-Kolmogorov equations. If we look at the transition matrices for a higher number of transitions, the probabilities appear as if they are reaching a limit. For example, the 4-step transition matrix $\mathbf{P}^{(4)}$ from that same example would be

$$\mathbf{P}^{(4)} = \begin{Vmatrix} 0.2512 & 0.7488 \\ 0.2496 & 0.7504 \end{Vmatrix},$$

and the 8-step transition matrix looks like

$$\mathbf{P}^{(8)} = \begin{Vmatrix} 0.2500 & 0.7500 \\ 0.2499 & 0.7501 \end{Vmatrix}.$$

From these matrices it seems that the entries approach a limit as the number of steps approaches infinity. As it turns out, this is exactly the case. However, getting to the theorem that states this formally requires a few more definitions.

State i is said to have **period** d if $P_{ii}^n = 0$ whenever n is not divisible by d, and d is the smallest positive integer with this property. For instance, starting in i it may be possible for the process to enter state i only when $n = 2, 4, 6, 8, \ldots$, in which case state i has period 2. A state with period 1 is said to be **aperiodic**.

If state i is recurrent, then it is said to be **positive recurrent** if, starting in i, the expected time until the process returns to state i is finite. (It could be shown that in a finite-state Markov chain all recurrent states are positive recurrent.) Also, positive recurrent, aperiodic states are called **ergodic**.

Using these definitions, Ross arrives at the following theorem [4].

Theorem 1. *For an irreducible, ergodic Markov chain,* $\lim_{n\to\infty} P_{ij}^n$ *exists and is independent of* i*. Furthermore, letting*

$$\pi_j = \lim_{n\to\infty} P_{ij}^n, \qquad j \geq 0$$

then π_j *is the unique nonnegative solution of*

$$\pi_j \quad = \quad \sum_{i=0}^{\infty} \pi_i P_{ij}^n, \qquad \textit{for } j \geq 0$$

and additionally,

$$\sum_{j=0}^{\infty} \pi_j = 1.$$

In words, this theorem says that the limit of each transition probability exists, and that it is independent of the initial state i. It also tells us that the limiting probability of the process being in state j at time n is equal to the fraction of the total time that the process will be in state j.

A complete proof of Theorem 1 is beyond the scope of this paper, but a partial proof will help the reader believe an integral part of the theorem. To that effect, Professor Robert Fontenot completed a proof showing that if the conditional probability $P_{ij}^{(n)} \to \pi_j$, as $n \to \infty$ then π_j can be thought of as the unconditional probability of being in state j after a long time as follows:

Proof 1.

$$\begin{aligned} P(X_{n+1} = j) &= \sum_i P(X_0 = i) P_{ij}^{(n)} \\ \lim_{n\to\infty} P(X_{n+1} = j) &= \sum_i P(X_0 = i)\pi_j \\ &= \pi_j(\sum_i P(X_0 = i)) \\ &= \pi_j(1) \\ &= \pi_j \qquad [2] \end{aligned}$$

To understand how this theorem can change how we see probabilities, let's look at an example. Assume that you like to ride your bike really fast because it makes you feel cool. Unfortunately, when you ride your bike really fast you have a hard time avoiding the pokey seed pods that are notorious for popping bicycle tires. Therefore, if you are riding your bike really fast one day, you will continue to ride fast the next day with a probability of 0.6. Otherwise you have a flat tire and you will not be riding at all. If you get a flat tire it makes you stop and think about whether feeling cool is really worth all the time it takes to walk to the store and get a new tube or patch the old one, so sometimes after you fix your tire you start to ride slowly for a time. Let's say that if you pop your tire one day, you get lazy and don't fix it with a probability of 0.5, you decide to ride slow for the next day with a probability of 0.2, and you fix the tire and forget your inhibitions about riding fast with a probability of 0.3. Lastly, if you have decided to ride slow one

day, the probability that you think riding slow is silly and start riding fast the next day is 0.7, the probability that you get a flat despite your best efforts to be careful is 0.2 and the probability that you continue to ride slow after a day of biking at such an indolent pace is 0.1. These probabilities lend themselves to a very neat representation as a three state Markov chain that has the following 1-step transition probability matrix:

$$\mathbf{P} = \begin{bmatrix} 0.6 & 0.4 & 0 \\ 0.3 & 0.5 & 0.2 \\ 0.7 & 0.2 & 0.1 \end{bmatrix}.$$

From this matrix we can see that being in state 0 represents riding fast, being in state 1 represents having a flat and being in state 2 represents puttering along. Then, from Theorem 1, we can write the limiting probabilities of each state as the following system of equations

$$\begin{aligned} \pi_0 &= (0.6)\pi_0 + (0.3)\pi_1 + (0.7)\pi_2, \\ \pi_1 &= (0.4)\pi_0 + (0.5)\pi_1 + (0.2)\pi_2, \\ \pi_2 &= (0.0)\pi_0 + (0.2)\pi_1 + (0.1)\pi_2, \\ 1 &= \pi_0 + \pi_1 + \pi_2. \end{aligned}$$

The equations for the limiting probabilities are the sums down the columns of the matrix where each entry in the column is multiplied by the limiting probability of that state to which it transitions. Now we can solve this system of equations to reach the actual limiting probabilities. In this case they are,

$$\begin{aligned} \pi_0 &= \frac{51}{95} = 0.5368 \\ \pi_1 &= \frac{36}{95} = 0.3789 \\ \pi_2 &= \frac{8}{95} = 0.0842. \end{aligned}$$

We can check our solutions by substituting them back into the system of equations, but it is easy to see that they do indeed sum to 1. What these solutions tell us is that, in the long run, we spend more than 50% of our time riding really fast, almost 40% not riding at all because of flat tires, and only around 8% of our time riding like a retiree on their way to a shuffleboard match. The beauty of limiting probabilities is that they allow us to make these generalizations because the longer we let the chain run, the closer the system comes to following them. Thus, we can call these limiting probabilities constants, (they are sometimes referred to as *stationary probabilities*), if we are willing to let our system

run indefinitely. This in turn makes it easier for us to interpret other things about the systems in fields like queuing theory, which is the study of efficiency and waiting in lines.

3.4. **A useful result.** Here is another example of how limiting probabilities can be used in conjunction with Markov chains. Suppose a wealthy alumnus was going to donate money to the soccer program based on how well the soccer team performed this season. Suppose the alum was offering \$200 for every game that the team won and \$50 for every game that the team played and didn't win during a ten game season. How much money could Mike Washington, the soccer coach, expect to receive from the benefactor? The following result and proof from Ross show that if, in a process governed by a Markov chain, a reward $r(j)$ is earned whenever the chain is in state j, then the average award per unit time is the sum of all the awards multiplied by the amount of time that the chain spends in each state, i.e. $\sum_j r(j)\pi_j$ [4].

Theorem 2. *Let $\{X_n, n \geq 1\}$ be an irreducible Markov chain with stationary probabilities π_j, $j \geq 0$. Let r be a bounded function on the state space, which means that the the states of the chain are the range of r. Then, with probability* 1,

$$\lim_{N\to\infty} \frac{\sum_{n=1}^{N} r(X_n)}{N} = \sum_{j=0}^{\infty} r(j)\pi_j. \tag{16}$$

Note that r is not necessarily a continuous function.

Proof 2. *If we let $a_j(N)$ be the amount of time that the Markov chain spends in state j during time periods $1, \ldots, N$, then*

$$\sum_{n=1}^{N} r(x_n) = \sum_{j=0}^{\infty} a_j(N) r(j).$$

Since $a_j(N)$ is the amount of time that chain is in state j, $a_j(N)/N$ is the fraction of time that the chain is in state j. This is fraction is the same as π_j, so the result follows from Equation 16 upon dividing by N and letting $N \to \infty$.

Using the probabilities from the simplest soccer example, we get the following system of equations:

$$\begin{aligned} \pi_0 &= (0.4)\pi_0 + (0.2)\pi_1 \\ \pi_1 &= (0.6)\pi_0 + (0.8)\pi_1 \\ \pi_0 + \pi_1 &= 1, \end{aligned}$$

and stationary probabilities

$$\pi_0 = \frac{0.2}{1 + 0.2 - 0.4} = 0.25, \qquad \pi_1 = \frac{0.6}{1 + 0.2 - 0.4} = 0.75,$$

where π_0 is the probability of winning, and π_1 is the probability of losing. So to find the amount of money the soccer team can expect to get, Mike would solve the equation

$$(200\pi_0 + 50\pi_1)10 = (200(0.25) + 50(0.75))10$$

to come up with \$875.

3.5. **Unconditional Markov Probabilities.** Another important twist on Markov chains is that they can be used to find unconditional probabilities. Up until this point, all the probabilities that we discussed were conditional. For instance, P_{ij}^n is the probability that the state at time n is j given that the initial state at time 0 is i. If the unconditional probability is desired, it is necessary to specify the probability distribution of the initial state. We can think of the initial probability distribution as the probabilities that a Markov chain has before it is actually in any state. For example, if the soccer team from previous examples has not yet played a game, perhaps we could say that the probability of them winning their first game is based on the strength of their incoming freshmen or the number of veteran players they graduated last year. Then our estimates of these probabilities are what we would use as the initial probability distribution. Let us denote this by

$$\begin{aligned} \alpha_i &= P(X_0 = i) \text{ for } i \geq 0, \\ &\text{where } \sum_{i=0}^{\infty} \alpha_i = 1. \end{aligned}$$

So each α_i is the probability of starting in state i. Then Ross tells us that all unconditional probabilities may be computed by conditioning on the initial state so that,

$$\begin{aligned} P(X_n = j) &= \sum_{i=0}^{\infty} P(X_n = j | X_0 = i) P(X_0 = i) \\ &= \sum_{i=0}^{\infty} P_{ij}^n \alpha_i \qquad [4]. \end{aligned}$$

In other words, the probability of being in state j at time n is the sum over all the different initial states, of the probability of going from state i to state j in n steps multiplied by the probability of starting in state i.

For example, to find the probability that the soccer team will win their fourth game from now (unconditionally), and again using the simplest transition matrix that we have used for the soccer team examples, we need to make up a distribution for the initial states. Don't forget that the one-step transition matrix and the four-step transition matrix are

$$\mathbf{P} = \begin{Vmatrix} 0.4 & 0.6 \\ 0.2 & 0.8 \end{Vmatrix},$$

and

$$\mathbf{P}^{(4)} = \begin{Vmatrix} 0.2512 & 0.7488 \\ 0.2496 & 0.7504 \end{Vmatrix}.$$

Now, let's say that $\alpha_0 = 0.3$ and $\alpha_1 = 0.7$. Then the probability that they will win their fourth game from now is

$$\begin{aligned} P\{X_4 = 0\} &= (0.3)P_{00}^4 + (0.7)P_{10}^4 \\ &= (0.3)(0.2512) + (0.7)(0.2496) \\ &= 0.25008. \end{aligned}$$

3.6. **Markov Chains in Genetics: The Hardy-Weinberg Law.** Many of the examples in this paper have been rather simplistic and contrived for the sake of understanding concepts, but if we turn to the field of genetics we can look at a more natural example from Ross' text that is still simple enough to further our understanding of the Markov chains [4]. Specifically, let us think about a large population of individuals with respect to a particular pair of genes. If each gene is classified as being of type A or of type a and we assume that Aa is the same as aA, then there are three possible gene combinations for each individual in the population, AA, aa, and Aa. Assume that the proportions of individuals who have these gene pairs are, respectively, p_0, q_0, and r_0 such that $p_0 + q_0 + r_0 = 1$. If the population being examined is comprised of college-aged humans, (or really any sexually mature animals), they are going to mate at some point. When this happens, each individual contributes one of his or her genes, chosen at random to their offspring. If the mating occurs randomly with each individual equally likely to mate with any other individual, (pretend everyone is naked, blindfolded, and bumbling around trying to mate with anything they bump into), then we can determine the proportions of individuals in the next generation whose genes are AA, aa, and Aa. To do this let us focus on an individual of the next generation and determine the probabilities for the gene pair of that individual.

Ross tells us that, "randomly choosing a parent and then randomly choosing one of its genes is equivalent to just randomly choosing a gene

from the total gene population", so we can condition on the gene pair of the parent to find the probability that a randomly chosen gene will be of type A [4]. So the probability of picking gene A is

$$\begin{aligned} P(A) &= P(A|AA)p_0 + P(A|aa)q_0 + P(A|Aa)r_0 \\ &= p_0 + \frac{1}{2}r_0, \end{aligned}$$

and picking a gene of type a has a probability of

$$\begin{aligned} P(a) &= P(a|AA)p_0 + P(a|aa)q_0 + P(a|Aa)r_0 \\ &= q_0 + \frac{1}{2}r_0. \end{aligned}$$

So if the mating in the population is random, it follows that the probability of getting the gene pair AA in the offspring is

$$\begin{aligned} p &= P(A)P(A) \\ &= (p_0 + r_0/2)^2, \end{aligned}$$

the probability of getting a baby that has the genes aa is

$$\begin{aligned} q &= P(a)P(a) \\ &= (q_0 + r_0/2)^2, \end{aligned}$$

and the probability of getting Aa is

$$\begin{aligned} r &= 2P(A)P(a) \\ &= 2(p_o + r_0/2)(q_0 + r_0/2). \end{aligned}$$

This last probability makes sense because it does not matter whether the child gets the genes Aa or aA. We know that p, q, and r are the probabilities of each individual getting those specific gene pairs. Then, since each member of the next generation gets their genes independent of each other member, these probabilities are also the fractions of the next generation that have each gene pair type.

As of yet this model is not a Markov chain, but if we look at the next generation of offspring after this one, we can begin to see the chain materialize. Since we are assuming random mating, the fractions of genes that are of type A or type a are the same in this third generation that they were in the second generation, namely $p + r/2$ and $q + r/2$. Ross shows this fact algebraically as follows:

$$\begin{aligned} p + r/2 &= (p_0 + r_0/2)^2 + (p_0 + r_0/2)(q_0 + r_0/2) \\ &= (p_0 + r_0/2)(p_0 + r_0/2 + q_0 + r_0/2) \\ &= p_0 + r_0/2 \text{ since } p_0 + q_0 + r_0 = 1 \\ &= P(A), \end{aligned}$$

and he goes on to note, "thus, the fractions of the gene pool that are A and a are the same as in the initial generation" [4]. Following from this conclusion we can say that under random mating all the successive generations after the first one will have the percentages of the population having gene pairs AA, aa, and Aa fixed at the values p, q, and r. This result is what is known as the Hardy-Weinberg law.

Now we can generalize the genetics probability example using a Markov chain. Suppose that the gene pair population has stabilized, as per the Hardy-Weinberg law, at percentages p, q, and r. If we simplify nature a bit, we can assume that each individual has exactly one offspring. For any given individual we will let the random variable X_n denote the genetic pair present in the individual's n^{th} generation decedent (i.e. if you were the individual being monitored, X_2 would be the genetic make-up of your grandchild). Then, if we condition on the state of a randomly chosen mate, we can create a transition probability matrix for the genetic make-up of any individual's descendants as follows:

$$\mathbf{P} = \begin{array}{c} \\ AA \\ aa \\ Aa \end{array} \begin{array}{c} \begin{array}{ccc} AA & aa & Aa \end{array} \\ \left\| \begin{array}{ccc} p+\frac{r}{2} & 0 & q+\frac{r}{2} \\ 0 & q+\frac{r}{2} & p+\frac{r}{2} \\ \frac{p}{2}+\frac{r}{4} & \frac{q}{2}+\frac{r}{4} & \frac{p}{2}+\frac{q}{2}+\frac{r}{2} \end{array} \right\| \end{array},$$

where the limiting probabilities for this Markov chain are conveniently also p, q, and r. These limiting probabilities are equal to the fractions of the individual's descendants that are in each of these genetic states.

4. The Problem of n Liars and Markov Chains

4.1. **The Problem.** All the theory that we have discussed about Markov chains has been presented with the intent of being able to understand problems in the field of Markov chain analysis. The first significant problem that we will examine comes from a paper written by William Feller in 1951 entitled, "The Problem of n Liars and Markov Chains [1]." In his paper, Feller generalizes and expounds upon the following problem, first treated by A.S. Eddington: "If A, B, C, D each speak the truth once in three times (independently), and A *affirms that* B *denies that* C *declares that* D *is a liar, what is the probability that* D *was telling the truth* [1]?"

Eddington treated this question as a simple probability problem, but Feller realized that it lends itself to representation by "the simplest Markov chain, and that natural variations of the same problem correspond to more general chains [1]." As Feller began his study of the n Liars by analyzing Eddington's treatment, so will we. From the

initial setup of the problem, there are only eight different statements that A can make or deny. Both of the aforementioned mathematicians concluded that the probability we are searching for is $\frac{13}{41}$, so we will work with this as our goal. To make things a little less confusing, let's reword the initial question. Instead of saying "A affirms that B denies that C declares that D is a liar, what is the probability that D was telling the truth," it is logically equivalent to write "A says that B says that C says that D is truthful." Additionally, let us call this assertion statement Q, so we do not have to repeatedly write it out. According to the formulation of the problem, the states are ordered so that D makes his statement at time 0, C makes his statement at time 1, and so on. Also, assume that each person knows only the statement of the last speaker, but not the past history of the system. So as part of our assumption, C knows whether or not D is lying. This will be fundamental in turning the problem into a Markov chain.

With this breakdown, the question really becomes one of conditional probability. What is the probability that D is actually telling the truth given that statement Q has been made, or $P(D = T|Q)$.

4.2. **The Eight Cases.** Since we have turned the problem into one of conditional probability, it is easier to see the eight cases that Feller spoke of. The first four cases are the four different ways that statement Q could be made while D is actually telling the truth. If that condition is satisfied, then the four cases are as follows:

$$\begin{aligned}
(i) \quad & P((A = T) \cap (B = T) \cap (C = T) \cap (D = T) \\
(ii) \quad & P((A = T) \cap (B = L) \cap (C = L) \cap (D = T) \\
(iii) \quad & P((A = L) \cap (B = T) \cap (C = L) \cap (D = T) \\
(iv) \quad & P((A = L) \cap (B = L) \cap (C = T) \cap (D = T),
\end{aligned}$$

where $A = T$ means that A is telling the truth from what their knowledge of the situation. We know that each person tells the truth once in three times, we can fill the times when a person is truthful with a probability of $\frac{1}{3}$ and the times when they are lying with $\frac{2}{3}$. Since these are the only four cases when statement Q was made and D is telling

the truth, we can sum the four probabilities to get

$$\begin{aligned} P(D = T \cap Q) &= \left(\frac{1}{3}\right)\left(\frac{1}{3}\right)\left(\frac{1}{3}\right)\left(\frac{1}{3}\right) + \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\left(\frac{2}{3}\right)\left(\frac{1}{3}\right) \\ &\quad + \left(\frac{2}{3}\right)\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\left(\frac{1}{3}\right) + \left(\frac{2}{3}\right)\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)\left(\frac{1}{3}\right) \\ &= \left(\frac{1}{81}\right) + \left(\frac{4}{81}\right) + \left(\frac{4}{81}\right) + \left(\frac{4}{81}\right) \\ &= \left(\frac{13}{81}\right). \end{aligned}$$

The next four cases are the ones when D is in fact lying, and they are written as

$$\begin{aligned} (v) &\quad P((A = T) \cap (B = T) \cap (C = L) \cap (D = L) \\ (vi) &\quad P((A = T) \cap (B = L) \cap (C = T) \cap (D = L) \\ (vii) &\quad P((A = L) \cap (B = T) \cap (C = T) \cap (D = L) \\ (viii) &\quad P((A = L) \cap (B = L) \cap (C = L) \cap (D = L). \end{aligned}$$

Which means that the probability of D being a liar while statement Q is made is

$$\begin{aligned} P(D = L \cap Q) &= \left(\frac{1}{3}\right)\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\left(\frac{2}{3}\right) + \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)\left(\frac{2}{3}\right) \\ &\quad + \left(\frac{2}{3}\right)\left(\frac{1}{3}\right)\left(\frac{1}{3}\right)\left(\frac{2}{3}\right) + \left(\frac{2}{3}\right)\left(\frac{2}{3}\right)\left(\frac{2}{3}\right)\left(\frac{2}{3}\right) \\ &= \left(\frac{4}{81}\right) + \left(\frac{4}{81}\right) + \left(\frac{4}{81}\right) + \left(\frac{16}{81}\right) \\ &= \left(\frac{28}{81}\right). \end{aligned}$$

Remembering back to our earliest definitions in probability, we find that

$$\begin{aligned} P(Q) &= P(D = T \cap Q) \cup P(D = L \cap Q) \\ &= \frac{13}{81} + \frac{28}{81} \\ &= \frac{41}{81}, \end{aligned}$$

and since another early definition tells us that we can write $P(D = T|Q)$ as $\frac{P(Q \cap D = T)}{P(Q)}$, we are left with $\frac{13}{81}/\frac{41}{81} = \frac{13}{81} * \frac{81}{41} = \frac{13}{41}$ as we expected. We now have a definite answer to the probability that D is telling the truth

in the case of four liars. When we make the Markov chain, we can use this probability to test our formulation.

4.3. **Modeling the Liars as a Markov Chain.** In this original four person model of the liars problem, each person issues a statement which may be true or false and may contradict each other. If for a moment, we believe everything that each person says, each statement would imply either that D is telling the truth of lying. So when A makes a statement about what B and the others have said, he implies (rightly or wrongly) that D is either lying or telling the truth. According to Feller, "in a continued process, an even number of denials cancel, and the implication of statement like 'A_1 asserts that A_2 denies that $A_3 \cdots$' depends on the evenness of the total number of denials [1]."

Feller then sets up the chance process with two states. At any time the observed state is 1 if the last statement made implies that D is honest. Alternately the observed state is 2 if the last statement implies that D is a liar. So that we can think of the problem in terms of steps for our transition matrix. Feller says that the statements are issued at times $0, 1, 2, \ldots$, and that the initial state at time 0 is 1 if D tells the truth and 2 if D lies. As we noted before, only the first two people know the initial state, because for the Markov property to apply each person can have knowledge only of what the last statement was. Since there are only two states, at time n the observed state changes or remains the same according as the n^{th} speaker tells the truth or lies. With this setup, Feller gives us the following Markov chain,

We have a process with two possible states, 1 *and* 2. *Initially* (or at time 0) *the probabilities of the two states are* α *and* β, *respectively* ($\alpha + \beta = 1$). *Whatever the development up to time* n, *there is probability* p *that at time* n *the observed state does not undergo a change and probability* $q = 1 - p$ *that it does. We seek the conditional probabilities* x_n *and* y_n *that the process actually started from state* 1, *given that at time* n *the observed state is* 1 *or* 2, *respectively* [1].

In this setup, the initial probability distribution represented by α and β is the probability that D is lying or telling the truth. Since each person in the problem tells the truth one in three times, the probability of starting in state 1, α, is $\frac{1}{3}$. Similarly, the probability that the observed state does not change is the same as the probability of the next person in line telling the truth, so $p = \frac{1}{3}$ as well. This makes $\beta = q = \frac{2}{3}$, because the initial probability distribution and the transition probabilities both need to sum to one. So the transition probabilities are the

probabilities that the person speaking lies about what they heard or not. They really say nothing about whether or not D is telling the truth even though it may initially seem like they should. D tells the truth with a probability of $\frac{1}{3}$ regardless of what anyone says about him.

Feller suggests that we use x_n to be the probability that D was actually telling the truth, given that at time n the person who is speaking claims that D told the truth. Since this is the case in our four person case (which means $n = 3$), x_3 is the probability that we found to be $\frac{13}{41}$ previously. So our goal will be to duplicate this result by calculating a general formula for x_n, and solving for our specific parameters.

4.4. **The Transition Matrix.** At each step of the game, there are two possible transitions the chain can make. If it is in state 1 the two possibilities are that the person can tell the truth and the chain can stay in state 1 or the person can lie so that the chain changes to state 2. If the chain starts in state 2, then the two possibilities are $2 \to 1$ or $2 \to 2$, and so the corresponding transition probabilities are,

$$\begin{aligned} p &= p_{11} = p_{22} \\ q &= p_{12} = p_{21}. \end{aligned} \tag{17}$$

We will let $p_{jk}^{(n)}$ represent the probability that the system is in state k after n steps if it started in state j. Since the one-step and two-step transition probabilities can be written as

$$p_{jk}^{(1)} = p_{jk},$$

and

$$p_{jk}^{(2)} = p_{j1}p_{1k} + p_{j2}p_{2k}. \tag{18}$$

Feller suggests that, generally, we can calculate the n-step transition probability using a recursion formula. He suggests that

$$p_{jk}^{(n)} = p_{j1}^{(n)}p_{1k} + p_{j2}^{(n)}p_{2k}, \tag{19}$$

because this is just the formula for matrix multiplication using the transition matrix implied by Equation 17 [1]. As we typically do with Markov chains, we will denote this matrix $\mathbf{P}$. This makes each $p_{jk}^{(n)}$ an element of $\mathbf{P}^n$.

We already know that the initial probabilities of states 1 and 2 are are α and β, so we can write $a_k^{(n)}$, the probability of observing the state k at time n, as

$$a_k^{(n)} = \alpha p_{1k}^{(n)} + \beta p_{2k}^{(n)}. \tag{20}$$

This probability is the sum of the probability of starting in either state multiplied by the probability of moving from that state to state k in n steps.

We now have everything we need to get formulas for the conditional probabilities x_n and y_n. Using the notation that we have defined here and the definition of conditional probability we can find that they are

$$x_n = \frac{\alpha p_{11}^{(n)}}{a_1^{(n)}} \quad \text{and} \quad y_n = \frac{\alpha p_{12}^{(n)}}{a_2^{(n)}}, \tag{21}$$

just as Feller did, but as he points out, for explicit formulas we must calculate $p_{jk}^{(n)}$ [1]. We can do this with the old linear algebra trick of diagonalization. If we diagonalize the transition matrix $\mathbf{P}$, we can find a relatively simple formula for $\mathbf{P}^n$. Then each $p_{jk}^{(n)}$ is just an entry in that matrix. To diagonalize $\mathbf{P}$ we need to find its eigenvalues and eigenvectors. Then we can create three matrices, R, D, and R^{-1}, such that the columns of R are the eigenvectors of $\mathbf{P}$, R^{-1} is the inverse of R, D is a diagonal matrix with the eigenvalues of $\mathbf{P}$ as the entries on its main diagonal, and the product of the three matrices is $\mathbf{P}$. If we can do that then $\mathbf{P}^n = RD^nR^{-1}$, and we have our formula.

By taking the determinant of $\mathbf{P}$ and finding its eigenvectors we get that

$$R = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix},$$

$$D = \begin{bmatrix} p-q & 0 \\ 0 & p+q \end{bmatrix},$$

and

$$R^{-1} = \begin{bmatrix} -1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}.$$

When we put them into the appropriate form we get

$$\begin{aligned} \mathbf{P}^n &= \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} (p-q)^n & 0 \\ 0 & (p+q)^n \end{bmatrix} \begin{bmatrix} -1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} \\ &= \left(\frac{1}{2}\right) \begin{bmatrix} (p-q)^n + (p+q)^n & -(p-q)^n + (p+q)^n \\ -(p-q)^n + (p+q)^n & (p-q)^n + (p+q)^n \end{bmatrix}. \end{aligned} \tag{22}$$

Feller equivalently writes this matrix as

$$\mathbf{P}^n = \left(\frac{1}{2}\right) \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + \left(\frac{1}{2}\right) (p-q)^n \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \tag{23}$$

That seems to make the next substitutions less explicit and more confusing so for our purposes we will use the matrix in Equation 22 [1].

If we now substitute from Equations 22 and 20 into 21, we come to the formula

$$x_n = \frac{\alpha(\frac{1}{2})((p-q)^n + (p+q)^n)}{\alpha(\frac{1}{2})((p-q)^n + (p+q)^n) + \beta(\frac{1}{2})(-(p-q)^n + (p+q)^n)},$$

which looks rather messy, but if we remember that $q = 1 - p$ and $\alpha + \beta = 1$, we can cancel terms to rewrite the equation as

$$x_n = \frac{\alpha(1 + (p-q)^n)}{1 + (\alpha - \beta)(p-q)^n}. \tag{24}$$

By similar methods we get the formula for y_n, which is,

$$y_n = \frac{\alpha(1 - (p-q)^n)}{1 - (\alpha - \beta)(p-q)^n}. \tag{25}$$

We now have two general formulas that tell us the conditional probability that D is telling the truth given that the last person affirms or denies that D is truthful no matter how many people are in the chain. This is great news because calculating the probability explicitly for large numbers of people, as we did for four people, would be incredibly time consuming. We can verify the accuracy of our formulas by evaluating the equations at $n = 3$ and $\alpha = p = \frac{1}{3}$. We see that $x_n = \frac{13}{41}$, just as it did when we crunched the numbers earlier and for the same parameters, $y_n = \frac{7}{20}$. Upon further inspection Feller realized that if we let $n \to \infty$ the value of the conditional probabilities tend towards the initial probability distribution ($x_n \to \alpha$ and $y_n \to \beta$) [1]. With the values of p, q, α, and β given in the problem, this limiting probability is demonstrated as follows:

$$\begin{aligned}
x_n &= \frac{(\frac{1}{3})(1 - (-\frac{1}{3})^n)}{1 + (-\frac{1}{3})(-\frac{1}{3})^n} \\
x_1 &= \frac{1}{2} = 0.5 \\
x_2 &= \frac{3}{10} = 0.3 \\
x_4 &= \frac{27}{82} = 0.329 \\
x_8 &= \frac{2187}{6562} = 0.333 \approx \alpha.
\end{aligned}$$

4.5. **Preferential Lying.** In his paper, Feller suggests another modification on the liars problem. Imagine that all of the people who are making statements about the initial liar, D, are friends of D or at least friendly acquaintances. They would probably like to believe that D was telling the truth unless he had lied to them sometime in their life. As such, when it arrives time for them to make a statement about the truthfulness of D, it would make sense for them to have a preference to claim the D is honest. Then Feller tells us that a transition from state $1 \to 1$ is more probable than $2 \to 2$, while $1 \to 2$ is less probable than $2 \to 1$ [1].

This is really just the same problem as before if we replace the transition probabilities in Equation 17 with the transition matrix

$$\mathbf{P} = \begin{bmatrix} p & q \\ q' & p' \end{bmatrix},$$

where $p + q = p' + q' = 1$. Then all of the previous formulas apply except that Feller's Equation 23 turns into

$$\mathbf{P}^n = \frac{1}{q+q} \begin{bmatrix} q' & q \\ q' & q \end{bmatrix} + \frac{(p'-q)^n}{q+q'} \begin{bmatrix} q & -q \\ -q' & q' \end{bmatrix},$$

through a similar process of diagonalization, and the final conditional probabilities become

$$\begin{aligned} x_n &= \frac{\alpha(q' + q(p'-q)^n)}{q' + (\alpha q - \beta q')(p'-q)^n} \\ y_n &= \frac{\alpha(q' - q(p'-q)^n)}{q' + (\beta q' - \alpha q)(p'-q)^n} \end{aligned}$$

by substitution [1].

5. Pallet Repair Policies at Labatt's Breweries

5.1. **Background.** The problem of n liars is an interesting exercise in turning an unlikely situation into a Markov chain. For a more realistic example we turn to Mr. Barry Davidson, an analyst for Labatt's Ontario Breweries. During June of 1969 Mr. Davidson was faced with a question posed be the warehouse manager concerning the repair and replacement policy for the pallets used in the shipment of beer. Should the pallets be repaired at all and if so, how often? This example comes to us courtesy of "Cases in Operations Research," by Christopher H. von Lanzenaurer [5].

Labatt's Breweries Ltd., of which Labatt's Ontario Breweries was a wholly owned subsidiary, was one of Canada's largest brewing companies with operations all across Canada. Labatt's belonged to a pallet pool with the other breweries in Ontario and Quebec and their common retail outlets. The a pallets were used for shipments to the retail outlets from the breweries and the return of empties. Due to one of Canada's socialist ideas, the use of a common bottle and pallet by breweries in Canada allowed the return of empties to any brewery.

New bottles, when shipped to the breweries from the glass manufacturers, were shipped on new pallets. The bill for the new pallets was included with the bill for the new bottles. From time to time, additional pallets could be ordered by individual breweries, if required. The number of pallets purchased by each brewery was recorded and at the end of each fiscal year, breweries that had purchased more than their share of new pallets were compensated by those breweries that had purchased less than their share. A brewery's share of new pallets purchased was determined by the number of new pallets purchased times its share of the market.

Damaged pallets were repaired by individual breweries if feasible. If a center block had been damaged, the pallet was not repaired. Often in repairing a pallet, the new nails would split a center block which would necessitate scrapping of the pallet. Approximately 10 percent of the damaged pallets were unrepairable. All pallets were identified as to time of purchase.

5.2. **Ice Cold Beer Facts.** At the end of fiscal 1969 (1969F), there were approximately $150,000$ pallets in the pool. In 1969F, $59,925$ new pallets had been purchased at an average cost of \$4.47, $32,050$ had been sold as scrap for an average price of \$0.55 and $7,771$ had been repaired by members of the pool. The average cost of repairing a pallet at Labatt's in 1969F was \$2.07. $1,721,000$ pallets of beer were moved in Ontario and Quebec in 1969F, making Labatt's share of the market approximately 33%.

Using the records for pallet repairs at Labatt's, Mr. Davidson was able to obtain the following average damage rates for the pallets,

Description of Pallets	Percentage of Pallets Damaged in Year			
	One	Two	Three	Four
New Pallets	22	45	33	—
Pallets Repaired in Year 1	—	47	48	05
Pallets Repaired in Year 2	—	—	83	17

,

which, he felt could be applied to the entire pallet pool. The foreman in charge of pallet repairs considered pallets over two years old not worth repairing and they were scrapped.

Although \$0.75 was the most ever received for damaged pallets, the warehouse manager thought that as much as \$1.50 could be obtained if the scrapped pallets were in a better condition.

With changes in the price of pallets, and the cost of repairing pallets which, due to high labor content, was rising rapidly, Mr. Davidson wanted to implement the most economical pallet replacement policy.

5.3. **Analysis.** To help Mr. Davidson with his analysis, we first need to understand his goal. Mr. Davidson is looking to minimize the cost associated with the repair and replacement of pallets for Labatt's. For our study we will assume that all the other breweries in the pool use the same policy as Labatt's and that the number of pallets in the pool is fixed at $150,000$. Then the cost in dollars is

$$\text{Cost} = \$150,000(4.47X + 2.07Y - 0.55Z), \tag{26}$$

where X is the fraction of pallets in the system that are new at the beginning of each year, Y is the fraction of pallets that are repaired each year, and Z is the fraction of pallets that are scrapped each year. This equation is simply calculating the brewery's expenditures on pallets minus the revenue generated by selling the scrap pallets to determine the money that they spend on pallets each year.

When examining the problem, Professor Fontenot suggested that this cost equation could be further simplified by assuming (as we are) that the size of the pallet pool is constant. In this case, Z is some fraction of X, say $Z = \lambda X$ where $0 < \lambda \leq 1$ is the fraction of X that is sold as scrap[2]. From our numbers $59,925$ new pallets were purchased in 1969, while only $32,050$ were sold as scrap. This means that $Z = \left(\frac{32050}{59925}\right) X = (0.535)X$. It would seem that those two numbers should be equal and that λ should always equal 1, but the discrepancy can be explained. Perhaps some of the pallets were damaged to the point that they could not be sold and were thrown away. Maybe there are uses for scrap wood around the brewery so some broken pallets are kept for Labatt's own purposes or maybe employees are allowed to take unrepairable pallets home for personal use. So we no have a cost equation for the current repair and replacement policy in terms of two variables. To help Mr. Davidson's analysis we need to find the values of X and Y for the current policy. Thinking in the context of Markov chain analysis, we can try to set up some states so that the steady state probabilities of the Markov chain give us the values that we need.

If we define the states for a pallet at the beginning of the year, the current repair policy can be modeled as follows:

state 1: 0 years old
state 2: 1 year old, undamaged
state 3: 2 years old, undamaged
state 4: 1 year old, repaired in the first year of use
state 5: 2 years old, repaired in the second year of use (and possibly also the first year)
state 6: 2 years old, repaired in the first year only
state 7: 3 years old.

In this setup, being 0 years old means that a pallet is brand new. Since we are assuming that the size of the pallet pool remains constant, all scrapped pallets are replaced with new ones so when a pallet is scrapped it can also be called 0 years old. The foreman in charge of pallet repairs said that damaged pallets over two years old are not worth repairing so we don't need states for pallets more than 3 years old, and it is also pivotal to note that all of these states are conditional. For example the probability of transitioning from state 2 to state 3, $\mathbf{P}_{23}$, is actually the probability of moving from 2 to 3 given that the pallet is in state 2. It cannot go from being 1 year old and undamaged to 2 years old and undamaged if it was actually damaged in the first year of use. This may seem trivial because any transition probability requires you to be in a certain state to transition from it, but consider the following. Say that you start with 100 pallets. Using Mr. Davidson's average damage rates, 22% of those pallets are damaged and 78% remain undamaged after their first year of use. So $\mathbf{P}_{12} = \frac{(100)(0.78)}{100} = 0.78$. Then for $\mathbf{P}_{21}$, we are starting out in state 2 with only 78 pallets, because that is how many were undamaged after the first year. Additionally we can see that the percentages in Mr. Davidson's chart sum to 1 across the rows. This implies that in year 3, all of the remaining pallets that were undamaged after their first two years get damaged, as opposed to 33% of the remaining undamaged pallets being damaged. Thus, the 45% in year 2 is 45% of the original hundred, meaning that 45 pallets are damaged. Since there only 78 undamaged pallets going into year 2, $\frac{45}{78} = 57.7\%$ of the remaining pallets are damaged. Additionally, Mr. Davidson noted that 10% of the damaged pallets are unrepairable, so $\mathbf{P}_{21} = \frac{(100)(0.45)(0.1)}{78} = 0.058$. This same conditioning is carried into the other transition probabilities

giving us,

$$\mathbf{P} = \begin{bmatrix} 0.022 & 0.78 & 0 & 0.198 & 0 & 0 & 0 \\ 0.058 & 0 & 0.423 & 0 & 0.519 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.047 & 0 & 0 & 0 & 0.423 & 0.530 & 0 \\ 0.83 & 0 & 0 & 0 & 0 & 0 & 0.17 \\ 0.906 & 0 & 0 & 0 & 0 & 0 & 0.094 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

as the transition matrix.

If we remember back to steady-state probabilities we can see that solving the system of equations,

$$\begin{aligned} \pi_1 &= \pi_1(0.022) + \pi_2(0.78) + \pi_4(0.198) \\ \pi_2 &= \pi_1(0.058) + \pi_3(0.423) + \pi_5(0.519) \\ \pi_3 &= \pi_1(1) \\ \pi_4 &= \pi_1(0.047) + \pi_5(0.423) + \pi_6(0.53) \\ \pi_5 &= \pi_1(0.83) + \pi_7(0.17) \\ \pi_6 &= \pi_1(0.906) + \pi_7(0.094) \\ \pi_7 &= \pi_7(1) \\ 1 &= \pi_1 + \pi_2 + \pi_3 + \pi_4 + \pi_5 + \pi_6 + \pi_7, \end{aligned}$$

will give us the steady-state solution. This solution is

$$\begin{aligned} \pi_1 &= 0.333 \\ \pi_2 &= 0.261 \\ \pi_3 &= 0.110 \\ \pi_4 &= 0.066 \\ \pi_5 &= 0.163 \\ \pi_6 &= 0.035 \\ \pi_7 &= 0.031. \end{aligned}$$

We started solving these steady-state probabilities in search of X and Y from our cost equation. Here we can see that the fraction of pallets that are new at the beginning of each year, X, is around 33.3%. The fraction of pallets that are repaired each year, Y, is $\pi_4 + \pi_5 = 22.9\%$.

This means that our yearly pallet expenditure can be modeled by,

$$\begin{aligned} \text{Cost} &= \$150,000(4.47X + 2.07Y - 0.55(0.535)X) \\ &= \$150,000(4.47(0.333) + 2.07(0.229) - 0.55(0.535)(0.333)) \\ &= \$150,000(1.865) \\ &= \$2.797 \times 10^5. \end{aligned}$$

5.4. **Alternative Policies.** It seems like there are a lot of ways the cost associated with pallets repairs could be reduced. One option that comes to mind is keeping a closer watch on the scrappable pallets. The coefficient λ from the $Z = \lambda X$ substitution is pretty much 50% in the current numbers. If all of the unrepairable pallets were sold as scrap, then that λ would approach 1 and the cost would be significantly decrease. Having $\lambda = 1$ would bring the cost from \$2.797x10^5 to \$2.669x10^5, which is a difference of about $\$12,800$. The problem with this solution is that it doesn't address when or how often the pallets are repaired. Those are the areas where it seems like there is the most wiggle room, because if a pallet is no good, you have to scrap it. If you have beer to ship, you need to buy a pallet. But if you have a somewhat broken pallet, you have more than one option.

5.5. **Case 1: Repair Pallets Only Once.** One option is that we could change Labatt's pallet repair policy so that pallets are only repaired one time in their life. With this change, we can use the same number of states as with the current policy. The only difference will be that state 5 will be defined as "2 years old, repaired in the second year only" instead of "2 years old, repaired in the second year of use and possibly also the first year." This makes the new transition matrix

$$\mathbf{P} = \begin{bmatrix} 0.022 & 0.78 & 0 & 0.198 & 0 & 0 & 0 \\ 0.058 & 0 & 0.423 & 0 & 0.519 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.470 & 0 & 0 & 0 & 0 & 0.530 & 0 \\ 0.83 & 0 & 0 & 0 & 0 & 0 & 0.17 \\ 0.906 & 0 & 0 & 0 & 0 & 0 & 0.094 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

where you can notice that there is no longer a positive transition probability for $\mathbf{P}_{45}$. This alteration makes the steady-state probabilities

equal to

$$\begin{aligned}\pi_1 &= 0.345\\ \pi_2 &= 0.269\\ \pi_3 &= 0.114\\ \pi_4 &= 0.068\\ \pi_5 &= 0.139\\ \pi_6 &= 0.036\\ \pi_7 &= 0.027,\end{aligned}$$

and the cost function becomes

$$\begin{aligned}\text{Cost} &= \$150,000(4.47(0.345) + 2.07(0.208) - 0.55(0.535)(0.345))\\ &= \$150,000(1.87)\\ &= \$2.806 \times 10^5,\end{aligned}$$

for $\lambda = 0.535$. For $\lambda = 1$ cost would be

$$\begin{aligned}\text{Cost} &= \$150,000(4.47(0.345) + 2.07(0.208) - 0.55(0.345))\\ &= \$150,000(1.78)\\ &= \$2.674 \times 10^5.\end{aligned}$$

We can see that the cost of this policy is lower than the cost of the current policy when the number of pallets that are sold as scrap is lower and pretty much the same when the number of pallets sold as scrap is equal to the number of new pallets. A potential bonus of this "repair once" policy is that we might get more money when selling the pallets as scrap wood, because they are in better condition. We have no numbers at this time to calculate the cost difference. However, if we were able to get even ten cents more per pallet, we would reduce our cost to \$2.779x10^5 for $\lambda = 0.535$ and \$2.623x10^5 for $\lambda = 1$. This difference is on the order of tens of thousands of dollars.

5.6. **Case 2: Never Repair.** Another policy worth examining is one under which Labatt exerts no effort in repairing broken pallets. The logic behind this idea is that they would be spending no money on repairs, and since they are in the pallet pool, the cost of all the new ones they were buying would be spread out amongst everyone in the pool. Additionally, same as the last policy, we could expect to get more money for our pallets sold as scrap because they have seen less use in their life times.

To model this policy, we are going to need to redefine the states in our Markov chain. It doesn't make sense to have seven states in this

policy, because we can simplify the states as follows:

state 1: 0 years old

state 2: 1 year old, undamaged

state 3: 2 years old, undamaged,

because according to Mr. Davidson's information no new pallet makes it past year three without getting damaged. This makes the transition probability matrix equal to

$$\mathbf{P} = \begin{bmatrix} 0.22 & 0.78 & 0 \\ 0.577 & 0 & 0.423 \\ 1 & 0 & 0 \end{bmatrix}.$$

With these transition probabilities, once the system reaches equilibrium the probabilities will be

$$\begin{aligned} \pi_1 &= 0.474 \\ \pi_2 &= 0.370 \\ \pi_3 &= 0.156, \end{aligned}$$

and our cost equation tells us that we should expect to spend

$$\begin{aligned} \text{Cost} &= \$150,000(4.47(0.474) + 2.07(0) - 0.55(0.535)(0.474)) \\ &= \$150,000(1.98) \\ &= \$2.97\text{x}10^5 \text{ for } \lambda = 0.535 \\ &\quad \text{and} \\ &= \$150,000(4.47(0.474) + 2.07(0) - 0.55(0.474)) \\ &= \$150,000(1.86) \\ &= \$2.79 \times 10^5 \text{ for } \lambda = 1. \end{aligned}$$

So it would seem that the "no repair" policy actually costs the most at face value. It may still be a viable option however. If we were able to get numbers for how much more the pallets are worth as scrap or the increase in beer shipping at Labatt's now that there are extra workers because no one is repairing pallets.

5.7. **Limitations.** The analysis that we just performed regarding Labatt's policy on repair and replacement of pallets gives a good initial assessment of the options available to them, but it is not exhaustive. A number of areas of the work could have been more detailed or accurate. For example, we don't know how much more the brewery would actually get for the scrapped pallets if they were never repaired or if

they were repaired once or repaired twice. Another failing is that we assume that a pallet can only break once a year, while that does not accurately reflect the real world. We also don't fully know the brewery's priorities. If they are concerned with environmental impacts in their cost analysis, they might be hesitant for the sake of the trees to adopt a plan that involves throwing out repairable pallets. On the other hand, if they are worried about rising labor costs, then they would want to repair the pallets as little as possible.

6. Conclusion

The objective of this paper was to explore the theory of Markov chains with the hope that we would discover its value through relevant case studies. From this exploration, Markov chains seem to be more important as a stepping stone that leads to other forms of analysis, than as an analytical tool themselves. In the fields of Queuing Theory and Operations Research, for example, Markov theory is used to assume that after a long enough period of time, systems act according to their limiting probabilities. So Markov chains play a part in the assumptions of these two fields that are used by many modern industries, and for that reason this paper has been an attempt to enlighten us about the grounds that these fields are built on.

References

[1] Feller, William. *The American Mathematical Monthly* "The Problem of n Liars and Markov Chains." Vol 58, No. 9. (Nov. 1951), pp 606-608.

[2] Fontenot, Robert A. *Unpublished?* Walla Walla, WA 2007.

[3] Karlin, Samuel. *A First Course in Stochastic Processes.* Academic Press, Inc., New York, 1966.

[4] Ross, Sheldon M. *Introduction to Probability Models, Ninth Edition.* Academic Press, Inc., Orlando, FL, 2006.

[5] von Lanzenaurer, Christopher H. *Cases in Operations Research.* Holden-Day Publishing Company, New York, NY. 1975.

[6] Weisstein, Eric W. *"Markov Chain."* From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/MarkovChain.html.

[7] Woods, Brian P. *He told me* In the science building, Walla Walla, WA. 2008.