Introduction to Stochastic Processes

Thomas Laetsch

Contents

Part 1

Elements of Probability Theory

CHAPTER 1

Elementary Theory of Random Variables

1. Probability Spaces

We will often refer to a triple $(\Omega, \mathcal{B}, \mathbb{P})$ as a *probability space*. This notation means that Ω is the *sample space* (i.e., the set of all possible outcomes of some experiment); $\mathcal{B}$ is a collection of subsets $E \subseteq \Omega$, with each subset E in $\mathcal{B}$ called an *event*; $\mathbb{P}$ is the *probability*, which assigns a number to each event E in the following way:

(1) $0 \leq \mathbb{P}(E) \leq 1$ for every event E.

(2) $\mathbb{P}(\varnothing) = 0$,

(3) $\mathbb{P}(\Omega) = 1$,

(4) If $\{E_i\}_{i=1}^N$ are disjoint events for any $N \in \{1, 2, ..., \infty\}$, then

$$P\Big(\bigcup_{i=1}^{N} E_i\Big) = \sum_{i=1}^{N} P(E_i).$$

Intuitively, $\mathcal{B}$ collects together all subsets $E \subseteq \Omega$ for which it makes sense to assign a probability value $\mathbb{P}(E)$, and in this case, the subset E is called an event. Rigorously, the subsets in $\mathcal{B}$ must satisfy several technical properties; however, we won't focus on the details of $\mathcal{B}$ in these notes and rarely mention it from here on – suffice it to say that in practice, one must work in a very contrived and counterintuitive way to produce a subset of Ω which is not an event. We note that amongst analysts, $\mathcal{B}$ is called a σ-algebra (read: *sigma algebra*), though it is common for probabilists to refer to $\mathcal{B}$ as a σ-field; since reference to $\mathcal{B}$ is usually omitted in these notes, we do not choose a convention here.

In what follows throughout these notes, we discover many properties of a probability $\mathbb{P}$. To start, we present an oft used result here, with two familiar corollaries.

THEOREM 1.1. *Let E_1 and E_2 be events. If $E_1 \subseteq E_2$, then $\mathbb{P}(E_2 \setminus E_1) = \mathbb{P}(E_2) - \mathbb{P}(E_1)$.*

PROOF. This proof boils down to the set-theoretic equality $E_2 = E_1 \cup (E_2 \setminus E_1)$, where we note that the righthand side is a disjoint union of events. Therefore, by the properties of probabilities,

$$\mathbb{P}(E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2 \setminus E_1) \geq \mathbb{P}(E_1).$$

A simple rearrangement of the above equality finishes the proof. □

COROLLARY 1.2. *Let E be an event. Then*

$$\mathbb{P}(E) = 1 - \mathbb{P}(\Omega \setminus E)$$

PROOF. By Theorem 1.1, $\mathbb{P}(\Omega \setminus E) = \mathbb{P}(\Omega) - \mathbb{P}(E) = 1 - \mathbb{P}(E)$. Hence, by a simple rearrangment, $\mathbb{P}(E) = 1 - \mathbb{P}(\Omega \setminus E)$. □

COROLLARY 1.3. *Let E_1 and E_2 be events. If $E_1 \subseteq E_2$, then $\mathbb{P}(E_1) \leq \mathbb{P}(E_2)$.*

PROOF. From Theorem 1.1, we have $\mathbb{P}(E_2 \setminus E_1) = \mathbb{P}(E_2) - \mathbb{P}(E_1)$ which rearranges to $\mathbb{P}(E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2 \setminus E_1)$. From this last equality, since $\mathbb{P}(E_2 \setminus E_1) \geq 0$, we deduce that $\mathbb{P}(E_2) \geq \mathbb{P}(E_1)$, which finishes the proof. □

Because of the mix of mathematics and intuition that is ubiquitous in probability, there are often several ways to express identical expressions. Two of these which are prudent to introduce here are as follows:

- Given events E_1 and E_2, the intersection of these $E_1 \cap E_2$ is often translated into words as "the event that E_1 *and* E_2 has occurred." Moreover, when finding the probability of the intersection, it is common to use a comma between events, writing $\mathbb{P}(E_1, E_2)$ rather than $\mathbb{P}(E_1 \cap E_2)$. So, the values $\mathbb{P}(E_1 \cap E_2)$, $\mathbb{P}(E_1, E_2)$, and $\mathbb{P}(E_1 \text{ and } E_2)$ are equivalent.
- Given events E_1 and E_2, the union of these events $E_1 \cup E_2$ is often translated into words as "the event that E_1 *or* E_2 has occurred." However, there is no symbol playing the analogue of the comma in this case. Hence, we have equality of the values $\mathbb{P}(E_1 \cup E_2)$ and $\mathbb{P}(E_1 \text{ or } E_2)$.

Of course, these conventions hold by easy analogy for any number of events.

2. Independent Events

Two events E_1 and E_2 are called *independent* when $\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\,\mathbb{P}(E_2)$. More generally, any number of events $\{E_k\}_{k=1}^N$ for any $N \in \{1, 2, ..., \infty\}$ are called independent when

$$\mathbb{P}\Big(\bigcap_{j=1}^{m} E_{k_j}\Big) = \prod_{j=1}^{m} \mathbb{P}(E_{k_j}) \tag{1}$$

for all finite collections of distinct indices $k_1, ..., k_m \in \{1, 2, ..., N\}$.

EXAMPLE 2.1. To make sure the definition of independence of events is clear, suppose that E_1, E_2, and E_3 are events. The, by our definition, these random variables are independent whenever

the equalities

$$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\,\mathbb{P}(E_2),$$
$$\mathbb{P}(E_1 \cap E_3) = \mathbb{P}(E_1)\,\mathbb{P}(E_3),$$
$$\mathbb{P}(E_2 \cap E_3) = \mathbb{P}(E_2)\,\mathbb{P}(E_3), \text{ and}$$
$$\mathbb{P}(E_1 \cap E_2 \cap E_3) = \mathbb{P}(E_1)\,\mathbb{P}(E_2)\,\mathbb{P}(E_3)$$

all hold. △

EXAMPLE 2.2. Suppose that 2 fair dice are rolled. Let E_1 be the event that the first rolled die landed on 3; let E_2 be the event that the second rolled die landed on 3; and let E_3 be the event that the sum of the two dice is 7. Then $\mathbb{P}(E_1) = \mathbb{P}(E_2) = \mathbb{P}(E_3) = 1/6$; $\mathbb{P}(E_1 \cap E_2) = 1/36 = \mathbb{P}(E_1)\,\mathbb{P}(E_2)$, $\mathbb{P}(E_1 \cap E_3) = 1/36 = \mathbb{P}(E_1)\,\mathbb{P}(E_3)$, and $\mathbb{P}(E_2 \cap E_3) = 1/36 = \mathbb{P}(E_2)\,\mathbb{P}(E_3)$; however, $\mathbb{P}(E_1 \cap E_2 \cap E_3) = 0 < \mathbb{P}(E_1)\,\mathbb{P}(E_2)\,\mathbb{P}(E_3) = 1/216$. This shows that, although any two of the three events are independent, the three events together are not independent. △

3. Random Variables and Expected Value

A *random variable* X defined on a probability space $(\Omega, \mathcal{B}, \mathbb{P})$ is a function $X : \Omega \to \mathbb{R}$ such that the set $\{\omega \in \Omega \text{ s.t. } X(\omega) \leq t\}$ is an event for each fixed $t \leq \mathbb{R}$. The *state space* of X, denoted S_X, is the *image* of the function $X : \Omega \to \mathbb{R}$; that is, S_X is the set of all possible outcomes of X.

EXAMPLE 3.1. Consider an experiment where a coin is flipped three times. A reasonable sample space housing all outcomes of the experiment is then

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

where 0 represents flipping tails and 1 represents flipping heads. Let X be the random variable which counts the number of heads flipped. Then, $X : \Omega \to \mathbb{R}$ has state space $S_X = \{0,1,2,3\}$, since the possible number of heads flipped is an inter value between 0 and 3. Within these notes, we will typically repress the knowledge that X is a function and not worry about an explicit form for the sample space Ω nor the subsequent description of how X maps the elements in Ω. However, for this example, since we have an explicit representation of Ω, let us give an explicit formula for X. Indeed, by how we have chosen Ω, we can write $X(\omega_1, \omega_2, \omega_3) = \omega_1 + \omega_2 + \omega_3$ for $(\omega_1, \omega_2, \omega_3) \in \Omega$; so, for example $X(0,1,1) = 0 + 1 + 1 = 2$, so the outcome $(0,1,1)$, two heads were flipped. Moreover, we can describe many events explicitly as well; e.g., the event $\{X = 1\}$ is equal to $\{(0,0,1), (0,1,0), (1,0,0)\}$; the event $\{X \geq 2\}$ is equal to $\{(0,1,1), (1,0,1), (1,1,0), (1,1,1)\}$. △

To introduced common notation, suppose that $s, t \in \mathbb{R}$ and $A \subseteq \mathbb{R}$. If X is a random variable, events of the form $\{\omega \in \Omega \text{ s.t. } X(\omega) \leq t\}$ or $\{\omega \in \Omega \text{ s.t. } s < X(\omega) \leq t\}$ or $\{\omega \in \Omega \text{ s.t. } X(\omega) \in A\}$ etc., are used frequently and are shorthanded to $\{X \leq t\}$, $\{a < X \leq b\}$, and $\{X \in A\}$. This is by no means an exhaustive list of all such shorthands, but others are easily translated by analogy. Beyond making the notation easier, this shorthanded notation is also a more intuitive way to understand the event; i.e., $\{X \leq t\}$ is "the event that X is less than or equal to t."

The *distribution* of a random variable X, also called the *cumulative distribution function* (CDF) of X, is the function $F_X : \mathbb{R} \to \mathbb{R}$ defined by $F_X(t) = \mathbb{P}(X \leq t)$, where $\mathbb{P}(X \leq t)$ is the probability of the event $\{X \leq t\}$. Given two random variables X and Y which have the same distribution, we write $X \overset{\text{dist}}{=} Y$ and say that X and Y are *equal in distribution* or *identically distributed*.

The *expected value* of a random variable X, denoted $\mathbb{E}[X]$, is defined as

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t)\, dt - \int_{-\infty}^0 \mathbb{P}(X \leq t)\, dt \tag{2}$$

as long as the right hand side of the equality makes sense (e.g., we are not subtracting infinity from infinity). Whenever both terms $\int_0^\infty \mathbb{P}(X > t)\, dt$ and $\int_{-\infty}^0 \mathbb{P}(X \leq t)\, dt$ are finite, we will say that X is *integrable*. This expression emphasizes the dependence of the expected value on the choice of probability $\mathbb{P}$. Note that using the definition of the distribution, we can rewrite (2) as

$$\mathbb{E}[X] = \int_0^\infty \big[1 - F_X(t)\big]\, dt - \int_{-\infty}^0 F_X(t)\, dt. \tag{3}$$

Convention 3.1. Unless explicitly mentioned otherwise, we will always assume the random variables we are considering are integrable. △

Example 3.2. Suppose a coin is flipped with a probability p of landing heads. Let X be the outcome of the flip taking that value 0 if the flip is tails, and the value 1 if the coin flip is heads. Then, the cumulative distribution of X is

$$F_X(t) = \begin{cases} 0 & t < 0 \\ 1 - p & 0 \leq t < 1 \\ 1 & t \geq 1 \end{cases}$$

Therefore,

$$\mathbb{E}[X] = \int_0^\infty [1 - F_X(t)]\, dt - \int_{-\infty}^0 F_X(t)\, dt = \int_0^1 [1 - (1-p)]\, dt + \int_1^\infty [1-1]\, dt - \int_{-\infty}^0 0\, dt$$
$$= \int_0^1 p\, dt = p.$$

Thus, $\mathbb{E}[X] = p$. △

4. Jointly Distributed Random Variables

When several random variables are defined on the same probability space $(\Omega, \mathcal{B}, \mathbb{P})$, we often refer to them as *jointly distributed*. Here we briefly review some of the introductory notions of jointly distributed random variables; later, we will delve a bit deeper into these concepts in the special cases of jointly discrete or jointly continuous random variables.

4.1. Two Random Variables. Given two random variables X and Y both defined on $(\Omega, \mathcal{B}, \mathbb{P})$, the *joint distribution* of X and Y, also called the *joint cumulative distribution function* (jCDF) of X and Y, is the function $F_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ defined by $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$; as discussed above, $\mathbb{P}(X \leq x, Y \leq y)$ is the probability of the event $\{X \leq x\} \cap \{Y \leq y\}$. The random variables X and Y are called *independent* when $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B)$ for any subsets $A \subseteq \mathbb{R}$ and $B \subseteq \mathbb{R}$ such that $\{X \in A\}$ and $\{Y \in B\}$ are both events. In fact, one potentially easier way to check for independence is the following.

Proposition 4.1. *The random variables X and Y are independent if and only if $F_{X,Y}(x, y) = F_X(x)F_Y(y)$ for every x and y. Here $F_{X,Y}$ is the jCDF of X and Y, F_X is the CDF of X, and F_Y is the CDF of Y.*

Proof. One direction is easy. If X and Y are independent, then $\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x)\mathbb{P}(Y \leq y)$ for any x and y, which is equivalent to $F_{X,Y}(x, y) = F_X(x)F_Y(y)$ for every x and y. The converse statement is a bit more involved and out of the scope of these notes, but remains true nonetheless. □

4.2. Multiple Random Variables. Consider now that $\{X_k\}_{k=1}^N$ are random variables all defined on $(\Omega, \mathcal{B}, \mathbb{P})$ for $N \in \{1, 2, ..., \infty\}$. If $N < \infty$, the joint distribution of these random variables is defined as the function $F_{X_1,...,X_n} : \mathbb{R}^n \to \mathbb{R}$ defined by

$$F_{X_1,...,X_N}(x_1, ..., x_N) = \mathbb{P}(X_1 \leq x_1, ..., X_N \leq x_N).$$

In the case $N = \infty$, this distribution function is not typically defined, but there are other more sophisticated methods to explore the joint distributions of these random variables – we will not go further into this here. In any case, independence is defined by the property $\mathbb{P}(X_{k_1} \in A_1, ..., X_{k_m} \in A_m) = \mathbb{P}(X_{k_1} \in A_1) \cdots \mathbb{P}(X_{k_m} \in A_m)$ for all finite collections of distinct indices $k_1, ..., k_m \in \{1, 2, ..., n\}$. Hence the analogue of Proposition 4.1 is given as follows.

PROPOSITION 4.2. *If $N < \infty$, the random variables $X_1, X_2, ..., X_N$ are independent if and only if*

$$F_{X_{k_1},...,X_{k_m}}(x_1, ..., x_m) = F_{X_{k_1}}(x_1) \cdots F_{X_{k_m}}(x_m)$$

holds for every collection of distinct indices $k_1, ..., k_m \in \{1, ..., N\}$ and all $x_1, ..., x_m \in \mathbb{R}$.

PROOF. The proof is done by an induction proof on the number of random variables N, with the case $N = 2$ having already been done in Proposition 4.1. The details are left as an exercise. □

EXAMPLE 4.1. To make sure the statements of independence are clear, suppose that X, Y, and Z are jointly distributed random variables. By definition, these random variables are independent whenever the equalities

$$\begin{aligned}
&\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B),\\
&\mathbb{P}(X \in A, Z \in C) = \mathbb{P}(X \in A)\,\mathbb{P}(Z \in C),\\
&\mathbb{P}(Y \in B, Z \in C) = \mathbb{P}(Y \in B)\,\mathbb{P}(Z \in C), \text{ and}\\
&\mathbb{P}(X \in A, Y \in B, Z \in C) = \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B)\,\mathbb{P}(Z \in C)
\end{aligned}$$

always hold for all allowable subsets A, B, and C of $\mathbb{R}$. Hence, the statement of Proposition 4.2 claims that this independence is equivalent to the equalities

$$\begin{aligned}
&F_{X,Y}(x, y) = F_X(x)\,F_Y(y),\\
&F_{X,Z}(x, z) = F_X(x)\,F_Z(z),\\
&F_{Y,Z}(y, z) = F_Y(y)\,F_Z(z), \text{ and}\\
&F_{X,Y,Z}(x, y, z) = F_X(x)\,F_Y(y)\,F_Z(z)
\end{aligned}$$

always holding for all values $x, y, z \in \mathbb{R}$. △

4.3. Identically Distributed and iid. For $N \in \{1, 2, ..., \infty\}$, a collection of random variables $\{X_k\}_{k=1}^N$ which may each be defined on different probability spaces, we say that these are *identically distributed* when each of their distributions are equal. That is, when $F_{X_i} = F_{X_j}$ for all $i, j \in \{1, ..., N\}$. If it happens that each of these random variables are defined on the same probability space, they are independent, and are identically distributed, we then refer to them by the initialism *iid*.

EXAMPLE 4.2. We run an experiment where we continually flip a coin, presuming that the outcomes of the flips are independent of each other. Letting X_i be the outcome of the ith flip

(0 for tails, 1 for heads, say), the random variables $X_1, X_2, X_3, ...$ are iid. Indeed, we assumed independence, and if p is the probability of the coin landing heads, then

$$F_{X_i}(x) = \begin{cases} 0 & x < 0 \\ 1-p & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

for every i, from which we deduce that they are identically distributed. △

5. Discrete Random Variables

A random variable X is called *discrete* when its state space S_X is a discrete set (this is always the case if the state space is finite). In the discrete case, we define the function $p_X : S_X \to \mathbb{R}$ by $p_X(x) = \mathbb{P}(X = x)$. This function p_X is called the *mass function*, or *probability mass function* (PMF), of X. One nice property here is that the expectation of a discrete random variable has a familiar and easy form.

LEMMA 5.1. *Let X be a discrete random variable with state space S_X and mass function p_X. Then for any function $g : \mathbb{R} \to \mathbb{R}$ such that the sum $\sum_{x \in S_X} g(x)p_X(x)$ is defined, the equality*

$$\mathbb{E}[g(X)] = \sum_{x \in S_X} g(x)\, p_X(x)$$

holds.

PROOF. Note that since $p_X(x) = \mathbb{P}(X = x)$, the above equality is equivalent to

$$\mathbb{E}[g(X)] = \sum_{x \in S_X} g(x)\, \mathbb{P}(X = x)$$

Let $Y = g(X)$. We first will deduce that

$$\sum_{y \in S_Y} y\, \mathbb{P}(Y = y) = \sum_{x \in S_X} g(x)\, \mathbb{P}(X = x).$$

To this end, for every $y \in S_Y$, let $T_y = \{x \in S_X \text{ s.t. } g(x) = y\}$. From the definition of T_y, we have the two equalities

$$\mathbb{P}(Y = y) = \sum_{x \in T_y} \mathbb{P}(X = x) \quad \text{and} \quad \bigcup_{y \in S_Y} T_y = S_X.$$

Therefore,

$$\sum_{y\in S_Y} y\,\mathbb{P}(Y=y) = \sum_{y\in S_Y}\sum_{x\in T_y} y\mathbb{P}(X=x) = \sum_{y\in S_Y}\sum_{x\in T_y} g(x)\,\mathbb{P}(X=x)$$
$$= \sum_{x\in \bigcup_{y\in S_Y} T_y} g(x)\,\mathbb{P}(X=x) = \sum_{x\in S_X} g(x)\,\mathbb{P}(X=x),$$

which proves the claim. Next, we will deduce that $\mathbb{E}[Y] = \sum y\,\mathbb{P}(Y=y)$. Our argument is valid, though some of the underpinning details and theory about interchanging limits of integrals and sums are not flushed out. We have,

$$\sum_y y\,\mathbb{P}(Y=y) = \sum_y \int_0^y dt\,\mathbb{P}(Y=y) = \sum_{y>0}\int_0^y dt\,\mathbb{P}(Y=y) - \sum_{y\le 0}\int_y^0 dt\,\mathbb{P}(Y=y)$$
$$= \int_0^\infty \sum_{y>t} \mathbb{P}(Y=y)\,dt - \int_{-\infty}^0 \sum_{y\le t} \mathbb{P}(Y=y)\,dt = \int_0^\infty \mathbb{P}(Y>t)\,dt - \int_{-\infty}^0 \mathbb{P}(Y\le t)\,dt$$
$$= \mathbb{E}[Y].$$

Putting these pieces together,

$$\mathbb{E}[g(X)] = \sum_{y\in S_Y} y\,\mathbb{P}(Y=y) = \sum_{x\in S_X} g(x)\,\mathbb{P}(X=x),$$

concluding the proof. □

5.1. Jointly Discrete. Two discrete random variables X and Y defined on the same sample space Ω will be called *jointly discrete*. The *joint mass function*, or the *joint probability mass function* (jPMF), $p_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ is defined as $p_{X,Y}(x,y) = \mathbb{P}(X=x, Y=y)$. From the joint mass function, we can find the mass functions p_X and p_Y of X and Y, often called the *marginal mass functions*, by

$$p_X(x) = \sum_{y\in S_Y} p_{X,Y}(x,y) \qquad \text{and} \qquad p_Y(y) = \sum_{x\in S_X} p_{X,Y}(x,y).$$

The expected value of jointly discrete random variables has a nice form.

LEMMA 5.2. *Let X and Y be jointly discrete random variables with respective state spaces S_X and S_Y, and with joint mass function $p_{X,Y}$. For any bounded function $g : \mathbb{R}^2 \to \mathbb{R}$,*

$$\mathbb{E}[g(X,Y)] = \sum_{x\in S_X}\sum_{y\in S_Y} g(x,y)\,\mathbb{P}(X=x, Y=y) = \sum_{x\in S_X}\sum_{y\in S_Y} g(x,y)\,p_{X,Y}(x,y).$$

PROOF. The details here are similar to the proof of Lemma 5.1 and hence omitted. The idea is that if we define $Z = g(X,Y)$, then Z is a discrete random variable, so by Lemma 5.1 we have

$\mathbb{E}[Z] = \sum_{z \in S_Z} z\,\mathbb{P}(Z = z)$. From here, we perform a few manipulations to show that

$$\sum_{z \in S_Z} z\,\mathbb{P}(Z = z) = \sum_{x \in S_X} \sum_{y \in S_Y} g(x, y)\,\mathbb{P}(X = x, Y = y).$$

□

The mass function reduces nicely in the case X and Y are independent.

PROPOSITION 5.3. *Let X and Y be discrete random variables (defined on the same sample space Ω). Then X and Y are independent if and only if $p_{X,Y}(x, y) = p_X(x)\,p_Y(y)$ for every x and y.*

PROOF. By the definition of independence, the equality $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\,\mathbb{P}(Y = y)$ holds. Conversely, if $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\,\mathbb{P}(Y = y)$ holds for every x and y, it is easy to show that $\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x)\,\mathbb{P}(Y \leq y)$ holds for every $x, y \in \mathbb{R}$, proving that X and Y are independent. □

Once again, we don't need to restrict our study to two jointly discrete random variables. Given several discrete random variables $X_1, X_2, ..., X_N$ with $N < \infty$, we can define the joint mass function as

$$p_{X_1,...,X_N}(x_1, ..., x_N) = \mathbb{P}(X_1 = x_1, ..., X_N = x_N).$$

From this, the marginal mass functions can again be recovered by summing over all other variables; for example,

$$p_{X_1}(x_1) = \sum_{x_2 \in S_{X_2}} \cdots \sum_{x_N \in S_{X_N}} p_{X_1,X_2,...,X_N}(x_1, x_2, ..., x_N).$$

The straightforward generalizations of Lemma 5.2 and Proposition 5.3 holds in the case of several discrete random variables.

Once again, we don't need to restrict our study to two jointly discrete random variables. Given several random variables $X_1, X_2, ..., X_N$ with $N < \infty$, we say that they are *jointly discrete* when they are jointly distributed and each is a discrete random variable. The direct and obvious analogues for defining the joint distribution $F_{X_1,...,X_N}$ and mass function $p_{X_1,...,X_N}$ are used. Moreover, the following results still hold over:

- The (marginal) mass function p_{X_i} of each of the random variables X_i can be found by summing the joint mass function $p_{X_1,...,X_N}$ over all other variables x_j with $j \neq i$.
- For any function $g : \mathbb{R}^N \to \mathbb{R}$,

$$\mathbb{E}[g(X_1, ..., X_N)] = \sum_{x_1 \in S_{X_1}} \cdots \sum_{x_N \in S_{x_N}} g(x_1, ..., x_N) p_{X_1,...,X_N}(x_1, ..., x_N)$$

so long as the sum on the right hand side is defined.

- The random variables $X_1, ..., X_N$ are independent if and only if

$$p_{X_1,...,X_N}(x_1, ..., x_N) = p_{X_1}(x_1) \cdots p_{X_N}(x_N)$$

for all points $x_1, ..., x_N$.

6. Continuous Random Variables

A random variable X is called *continuous* when its distribution F_X is *absolutely continuous*. Absolute continuity of F_X means that there exists a function $f_X : \mathbb{R} \to \mathbb{R}$ such that F_X is differentiable at almost every x with

$$F_X' = f_X \quad \text{and} \quad F_X(t) - F_X(t_0) = \int_{t_0}^{t} f_X(x)\, dx.$$

The function f_X is called the *density*, or *probability density function* (PDF), of X. Notice here that since F_X is non-decreasing, it must hold that its derivative f_X must be non-negative. From here we can review a few easily derived and familiar results.

LEMMA 6.1. *Suppose that X is a continuous random variable with distribution F_X and with density f_X. Then the following hold.*

(1) *$F_X(t) = \int_{-\infty}^{t} f_X(x)\, dx$ and, in particular, $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$.*
(2) *$\mathbb{P}(X = t) = 0$ for any $t \in \mathbb{R}$. This further implies that for any $a \leq b$, the values $\mathbb{P}(a < X < b)$, $\mathbb{P}(a \leq X < b)$, $\mathbb{P}(a \leq X \leq b)$, and $\mathbb{P}(a < X \leq b)$ are all equal and, in fact, are equal to $\int_a^b f_X(x)\, dx$.*

PROOF. (1) We have

$$\int_{-\infty}^{t} f_X(x)\, dx = \lim_{t_0 \to -\infty} \int_{t_0}^{t} f_X(x)\, dx = \lim_{t_0 \to -\infty} F_X(t) - F_X(t_0) = F_X(t) - 0 = F_X(t).$$

Further,

$$\int_{-\infty}^{\infty} f_X(x)\, dx = \lim_{t \to \infty} \int_{-\infty}^{t} f_X(x)\, dx = \lim_{t \to \infty} F_X(t) = 1$$

(2) Let $t \in \mathbb{R}$. Then

$$\mathbb{P}(X = t) = \lim_{\epsilon \to 0} \mathbb{P}(t - \epsilon < X \leq t + \epsilon) = \lim_{\epsilon \to 0} [F_X(t + \epsilon) - F_X(t - \epsilon)] = F_X(t) - F_X(t) = 0$$

Here, the equality $\lim_{\epsilon \to 0}[F_X(t + \epsilon) - F_X(t - \epsilon)] = F_X(t) - F_X(t)$ only holds because we know that F_X is continuous at every $t \in \mathbb{R}$ (since absolute continuity implies continuity) – this argument would not have worked without this continuity, which is why it does not work in the discrete

setting. Now, since $\mathbb{P}(X = t) = 0$, then we must have equality between $\mathbb{P}(a < X < b)$ and $\mathbb{P}(a \leq X < b)$ since $\mathbb{P}(a \leq X < b) = \mathbb{P}(X = a) + \mathbb{P}(a < X < b)$. This argument similarly shows that $\mathbb{P}(a < X < b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b)$. Since $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$ and $F_X(b) - F_X(a) = \int_a^b f_X(x)\,dx$, the claim follows. □

Next, we state the analogue of Lemma 5.1 in the continuous setting.

LEMMA 6.2. *Suppose that X is a continuous random variable with density f_X. Then for any function $g : \mathbb{R} \to \mathbb{R}$ such that the integral $\int_{-\infty}^{\infty} g(x) f_X(x)\,dx$ is defined, the equality*

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx$$

holds.

PROOF. The general proof of this would require a bit too much background, so we do not include a full proof in these notes. However, we can prove a very specific case here, that $\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx$. To this end,

$$\begin{aligned}
\int_{-\infty}^{\infty} x f_X(x)\,dx &= \int_{-\infty}^{\infty} \Big[\int_0^x dt\Big] f_X(x)\,dx = \int_{-\infty}^{\infty}\int_0^x f_X(x)\,dt\,dx \\
&= \int_0^{\infty}\int_0^x f_X(x)\,dt\,dx - \int_{-\infty}^{0}\int_x^0 f_X(x)\,dt\,dx \\
&= \int_0^{\infty}\int_t^{\infty} f_X(x)\,dx\,dt - \int_{-\infty}^{0}\int_{-\infty}^{t} f_X(x)\,dx\,dt \\
&= \int_0^{\infty} \mathbb{P}(X > t)\,dt - \int_0^{\infty} \mathbb{P}(X \leq t)\,dt = \mathbb{E}[X]
\end{aligned}$$

concluding the proof. □

REMARK 6.1. The definition of a continuous random variable is a bit more involved than for a discrete random variable. However, one can easily build an analogy between the two by interchanging the probability mass function $p(x)$ in the discrete case with the differential of the density $f(x)\,dx$ in the continuous case, as well as exchanging the sum $\sum$ in the discrete case with the integral $\int$ in the continuous case. Of course, this is just for intuition, not a rigorous statement. △

6.1. Jointly Continuous. Take two jointly distributed random variables X and Y with joint distribution $F_{X,Y}$. We say that these random variables are *jointly continuous* when there exists a function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ such that the following hold: $\frac{\partial}{\partial x} F_X(x, y)$ exists for almost all points x and is absolutely continuous in the variable y, $\frac{\partial}{\partial y} F_{X,Y}(x, y)$ exists for almost all points y and is absolutely continuous in the variable x, and $\frac{\partial^2}{\partial x\,\partial y} F_{X,Y}(x, y) = \frac{\partial^2}{\partial y\,\partial x} F_{X,Y}(x, y) = f_{X,Y}(x, y)$ for

almost all points x, y. The function $f_{X,Y}$ is called the *joint density*, or *joint probability density function* (jPDF), of X and Y. As before, the joint density $f_{X,Y}$ must be non-negative. This might not seem as obvious in this case, but notice that if we fix x, then as a function of y, the distribution $F_{X,Y}(x, y)$ is increasing, and similarly if we fixed y and considered only letting x vary. It is also worth noting that if $F_{X,Y}$ and $f_{X,Y}$ satisfy the assumptions listed, then the fundamental theorem of calculus quickly implies

$$F_{X,Y}(s,t) = \int_{-\infty}^{t}\int_{-\infty}^{s} f_{X,Y}(x,y)\,dx\,dy.$$

LEMMA 6.3. *Let X and Y be jointly continuous random variables with joint density $f_{X,Y}$. Then X and Y are each continuous random variables and their respective densities can be found by*

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

PROOF. We have

$$\begin{aligned} F_X(s) = \mathbb{P}(X \le s) &= \lim_{t\to\infty} \mathbb{P}(X \le s, Y \le t) = \lim_{t\to\infty} \int_{-\infty}^{s}\int_{-\infty}^{t} f_{X,Y}(x,y)\,dy\,dx \\ &= \int_{-\infty}^{s}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\,dx = \int_{-\infty}^{s}\left[\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\right]dx \end{aligned}$$

This shows that $F_X(s) = \int_{-\infty}^{s} f_X(x)\,dx$ where $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$, proving that f_X is the density of X. The symmetric argument works to conclude that f_Y is the density of Y. □

REMARK 6.2. By analogy of the discrete case, the densities f_X and f_Y recovered using Lemma 6.3 are often called the *marginal* densities of X and Y respectively. However, the extra adjective marginal is superfluous since f_X and f_Y are, in fact, just the densities of X and Y. △

LEMMA 6.4. *Let X and Y be jointly continuous random variables with joint density $f_{X,Y}$. For any function $g : \mathbb{R}^2 \to \mathbb{R}$ such that the integral $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dx\,dy$ is defined, the equality*

$$\mathbb{E}[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dx\,dy$$

holds.

PROOF. As before, the proof of this fact is out of the scope of these notes. It remains true nonetheless. □

LEMMA 6.5. *Let D be some domain in $\mathbb{R}^2$. Then*

$$\mathbb{P}((X,Y) \in D) = \iint_D f_{X,Y}(x,y)\,dx\,dy.$$

In particular, if $a \leq b$ and $c \leq d$, then

$$\mathbb{P}(a < X \leq b, c < Y \leq d) = \int_c^d \int_a^b f_{X,Y}(x,y)\,dx\,dy.$$

PROOF. The quickest proof is to start with Lemma 6.4 and make the following argument (which may seem a bit mysterious at this point, but hopefully will become clearer in the future when we become familiar with indicator functions). Define the indicator function on D, denoted $1_D : \mathbb{R}^2 \to \mathbb{R}$, by

$$1_D(x,y) = \begin{cases} 1 & (x,y) \in D \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\mathbb{P}((X,Y) \in D) = \mathbb{E}[1_D(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 1_D(x,y) f_{X,Y}(x,y)\,dx\,dy = \iint_D f_{X,Y}(x,y)\,dx\,dy.$$

This concludes the proof. □

Furthering the analogies to jointly discrete random variables, the joint density of jointly continuous random variables reduces nicely in the case of independence.

PROPOSITION 6.6. *Let X and Y be jointly continuous random variables. X and Y are independent if and only if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for every x and y.*

PROOF. This argument can be made as follows. From Proposition 4.1, X and Y are independent if and only if $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ for every x and y. If $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ then

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_X(x)F_Y(y) = F_X'(x)F_Y'(y) = f_X(x)f_Y(y).$$

showing that if X and Y are independent then $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. On the other hand, if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for every x and y, then

$$\begin{aligned} F_{X,Y}(x,y) &= \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(s,t)\,dsdt = \int_{-\infty}^{y}\int_{-\infty}^{x} f_X(s)f_Y(t)\,dsdt \\ &= \int_{-\infty}^{x} f_X(s)\,ds \int_{-\infty}^{y} f_Y(t)\,dt = F_X(x)F_Y(y). \end{aligned}$$

showing that if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ then X and Y are independent. □

As usual, we don't need to restrict our study to two jointly continuous random variables. Given several random variables $X_1, X_2, ..., X_N$ with $N < \infty$, we say that they are *jointly continuous* when there exists a joint density function $f_{X_1,...,X_N} : \mathbb{R}^n \to \mathbb{R}$ such that the obvious analogues of

the two-variable case hold over for the joint distribution $F_{X_1,...,X_N}$ and joint density $f_{X_1,...,X_N}$. In particular, the following results still hold over:

- Each of the random variables X_i for $1 \leq i \leq N$ is continuous and the (marginal) density f_{X_i} can be found by integrating the joint density over all other variables x_j with $j \neq i$.
- For any function $g : \mathbb{R}^N \to \mathbb{R}$,

  $$\mathbb{E}[g(X_1, ..., X_N)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, ..., x_N) f_{X_1,...,X_N}(x_1, ..., x_N)\, dx_1 \cdots dx_N$$

  so long as the integral on the right hand side is defined.
- For any domain $D \subseteq \mathbb{R}^N$, it holds that

  $$\mathbb{P}((X_1, ..., X_N) \in D) = \int \cdots \int_D f_{X_1,...,X_N}(x_1, ..., x_N)\, dx_1 \cdots dx_N$$

- The random variables $X_1, ..., X_N$ are independent if and only if

  $$f_{X_1,...,X_N}(x_1, ..., x_N) = f_{X_1}(x_1) \cdots f_{X_N}(x_N)$$

  for all points $x_1, ..., x_N$.

7. Variance, Covariance, and Correlation

Let X and Y be jointly distributed random variables with means $\mathbb{E}[X] = \mu_X$ and $\mathbb{E}[Y] = \mu_Y$. The *covariance* of X and Y, denoted $\text{Cov}(X, Y)$, is defined as

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y], \tag{4}$$

so long as the above expressions are defined. The covariance of a random variable with itself is called the *variance*. Explicitly, we find the variance of X, denoted $\text{Var}(X)$, by

$$\text{Var}(X) = \text{Cov}(X, X) = \mathbb{E}[(X - \mu_X)(X - \mu_X)] = \mathbb{E}[X^2] - \mathbb{E}[X]^2, \tag{5}$$

whenever the above expressions are defined. The *correlation* of X and Y, denoted $\text{corollary}(X, Y)$, is defined as

$$\text{corollary}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}. \tag{6}$$

defined whenever $\text{Cov}(X, Y)$ is defined and the denominator is non-zero. One can show that the correlation is always a value in the range -1 to 1. When one refers to *positively correlated* random variables, it means that the correlation between the random variables is positive, and similarly,

they are called *negatively correlated* with the correlation is negative. If the correlation between random variables is 0, then they are said to be *uncorrelated*.

The square root of the variance of random variables appears frequently in probability theory and is given the name *standard deviation*. So, the value $\sqrt{\mathrm{Var}(X)}$ is the standard deviation of X, and will often be denoted by the symbol σ_X.

8. Important Examples

8.1. Bernoulli Random Variables and Indicator Functions. A discrete random variable X which takes on only the two values 0 or 1 (i.e., $S_X = \{0, 1\}$) is called a *Bernoulli* random variable. If X is a Bernoulli random variable, we denote it symbolically by $X \overset{\text{dist}}{=} \mathrm{Ber}(p)$, where $p = \mathbb{P}(X = 1)$.

PROPOSITION 8.1. *Suppose that* $X \overset{\text{dist}}{=} \mathrm{Ber}(p)$. *Then* $\mathbb{E}[X] = p$ *and* $\mathrm{Var}(X) = p(1-p)$.

PROOF. We have

$$\mathbb{E}[X] = 0 \cdot \mathbb{P}(X = 0) + 1 \cdot \mathbb{P}(X = 1) = 0 + 1 \cdot p = p.$$

Further, since $X^2 = X$ in this case (since $0^2 = 0$ and $1^2 = 1$), $\mathbb{E}[X^2] = \mathbb{E}[X] = p$. Therefore, $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1-p)$. □

If W is any set and $A \subseteq W$ then the *indicator function* on A is the function denoted $1_A : W \to \mathbb{R}$ and defined by

$$1_A(x) = \begin{cases} 1 & x \in A \\ 0 & \text{otherwise} \end{cases}$$

One of the useful properties of indicator functions is the following.

LEMMA 8.2. *Let* $D \subseteq \mathbb{R}^n$ *be some domain and* $1_D : \mathbb{R}^n \to \mathbb{R}$ *be the indicator function on* D. *Then for any integrable function* $f : \mathbb{R}^n \to \mathbb{R}$,

$$\int \cdots \int_D f(x_1, ..., x_n)\, dx_1 \cdots dx_n = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} 1_D(x_1, ..., x_n) f(x_1, ..., x_n)\, dx_1 \cdots dx_n$$

PROOF. Since 1_D takes the value 1 whenever the argument is in D and takes the value 0 otherwise, multiplying 1_D and f results in the function f whenever the argument is in D and in the value 0 when the argument is outside of D. From this observation, the result follows immediately. □

As is clear from the definition, an indicator function only outputs two values, 0 or 1. Hence, if it so happens that E is an event of some sample space Ω, then the indicator function $1_E : \Omega \to \mathbb{R}$

is, in fact, a Bernoulli random variable. Moreover, the following two lemmas give extremely useful manipulations when wanting to write a probability as an expected value.

LEMMA 8.3. *Let E be an event and 1_E the indicator function on E. Then $\mathbb{E}[1_E] = \mathbb{P}(E)$.*

PROOF. As mentioned above, 1_E is a Bernoulli random variable. Therefore $\mathbb{E}[1_E]$ is equal to $\mathbb{P}(1_E = 1)$. However, the event $\{1_E = 1\}$ is exactly the event E (in prose, $1_E = 1$ is true if and only if E is true). Therefore $\mathbb{E}[1_E] = \mathbb{P}(E)$. □

LEMMA 8.4. *Let X be a random variable and $A \subseteq \mathbb{R}$. Then $\mathbb{P}(X \in A) = \mathbb{E}[1_A(X)]$.*

PROOF. We have already proved that $\mathbb{P}(X \in A) = \mathbb{E}[1_E]$ where E is the event $\{X \in A\}$. So, we only need to convince ourselves that $1_E = 1_A(X)$. Note that 1_A is the indicator function on $A \subseteq \mathbb{R}$ and is thusly the function $1_A : \mathbb{R} \to \mathbb{R}$ such that $1_A(x) = 1$ if and only if $x \in A$ and 0 otherwise. Therefore $1_A(X)$ gives the value 1 if and only if $X \in A$ and 0 otherwise, which is exactly the same as the indicator function 1_E when $E = \{X \in A\}$. □

8.2. Uniform Random Variables. We start with the intuitionally simplest example. Roughly defined, a random variable X is called *uniformly distributed* (or simply *uniform*) on its state space S_X when all outcomes are equally likely. We break this into a few cases for clarity:

Discrete Case with Finite State Space. A discrete random variable X with a finite state space S_X is called uniform when $\mathbb{P}(X = k) = 1/\#S_X$ for each $k \in S_X$ and 0 elsewhere. Here, $\#S_X$ indicates the number of elements in the set S_X. Note that this is equivalent to defining the probability mass function for X since $p_X(k) = \mathbb{P}(X = k)$. For example, suppose that X is uniformly distributed on the set $\{1, 2, 3\}$. This tells us that $\mathbb{P}(X = 1) = \mathbb{P}(X = 2) = \mathbb{P}(X = 3) = \frac{1}{3}$.

Continuous Case on an Interval. A continuous random variable X with state space some interval $S_X = I$ in $\mathbb{R}$ is called *uniform* on I, denoted $X \overset{\text{dist}}{=} \text{Unif}(I)$, when the density of X is defined as $f_X(t) = 1/(\text{length of } I)$ whenever $t \in I$, and 0 elsewhere. For example, if $I = (2, 5]$ and $X \overset{\text{dist}}{=} \text{Unif}(2, 5]$, then the density of a uniform random variable on I is

$$f(t) = \begin{cases} \frac{1}{5-2} & 2 < t \leq 5 \\ 0 & \text{elsewhere} \end{cases} = \begin{cases} \frac{1}{3} & 2 < t \leq 5 \\ 0 & \text{elsewhere} \end{cases}$$

PROPOSITION 8.5. *Let $X \overset{\text{dist}}{=} \text{Unif}(a, b)$ be a uniform random variable on the interval from a to b. Then, $\mathbb{E}[X] = \frac{1}{2}(a + b)$ and $\text{Var}(X) = \frac{1}{12}(b - a)^2$.*

PROOF. We have

$$\mathbb{E}[X] = \frac{1}{b-a}\int_a^b t\,dt = \frac{1}{2(b-a)}(b^2-a^2) = \frac{1}{2}(b+a).$$

We next consider $\mathbb{E}[X^2]$,

$$\mathbb{E}[X^2] = \frac{1}{b-a}\int_a^b t^2\,dt = \frac{1}{3(b-a)}(b^3-a^3) = \frac{1}{3}(b^2+ab+a^2).$$

Hence, we calculate the variance,

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1}{3}(b^2+ab+a^2) - \frac{1}{4}(b+a)^2 = \frac{1}{12}(b-a)^2$$

□

Continuous Case on a Region. Suppose that $D \subseteq \mathbb{R}^N$ is some bounded domain and $X_1, ..., X_N$ are jointly continuous random variables. We call these random variables uniform on D when the joint density is defined by

$$f_{X_1,...,X_N}(x_1,...,x_N) = \begin{cases} \frac{1}{\mathrm{Vol}(D)} & (x_1,...,x_N) \in D \\ 0 & \text{otherwise} \end{cases}$$

where $\mathrm{Vol}(D)$ is the volume of D, which can be calculated by the integral $\int\cdots\int_D dx_1\cdots dx_N$.

8.3. Binomial Random Variables. A discrete random variable X is called *Binomial* with parameters $n \in \mathbb{N}$ and $p \in (0,1)$, denoted $X \overset{\text{dist}}{=} \mathrm{Bin}(n,p)$, when the state space is $S_X = \{0,1,...,n\}$ and the probability mass function is given by $p_X(k) = \binom{n}{k}p^k(1-p)^{n-k}$ for each $k \in S_X$.

PROPOSITION 8.6. *Let* $X \overset{\text{dist}}{=} \mathrm{Bin}(n,p)$. *Then* $\mathbb{E}[X] = np$ *and* $\mathrm{Var}(X) = np(1-p)$.

PROOF. Manipulating the sum representation of the expected value and using the binomial formula,

$$\mathbb{E}[X] = \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k} = np\sum_{k=0}^{n-1}\binom{n-1}{k}p^k(1-p)^{(n-1)-k}$$
$$np(p+1-p)^{n-1} = np.$$

For the variance, we first consider $\mathbb{E}[X^2]$.

$$\begin{aligned}\mathbb{E}[X^2] &= \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} = np \sum_{k=0}^{n-1} (k+1) \binom{n-1}{k} p^k (1-p)^{(n-1)-k} \\ &= np \left[\sum_{k=0}^{n-1} k \binom{n-1}{k} p^k (1-p)^{(n-1)-k} + 1 \right] = np((n-1)p+1) \\ &= n^2p^2 + np(1-p)\end{aligned}$$

where we recognized the sum in the third equality as the expected value of a $\text{Bin}(n-1,p)$ random variable. This results in the variance

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = n^2p^2 + np(1-p) - n^2p^2 = np(1-p).$$

□

8.4. Poisson Random Variables. A discrete random variable X is called *Poisson* with parameter $\lambda > 0$, denoted $X \overset{\text{dist}}{=} \text{Pois}(\lambda)$, when the state space of X is all non-negative integers $S_X = \{0, 1, 2, ...\}$ and the probability mass function is given by $p_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ for each $k \in S_X$.

PROPOSITION 8.7. *Let* $X \overset{\text{dist}}{=} \text{Pois}(\lambda)$. *Then* $\mathbb{E}[X] = \lambda$ *and* $\text{Var}(X) = \lambda$.

PROOF. We have

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} = \lambda \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = \lambda.$$

Also

$$\begin{aligned}\mathbb{E}[X^2] = \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \sum_{k=1}^{\infty} k \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} &= \lambda \left[\sum_{k=1}^{\infty} (k-1) \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} + 1 \right] \\ &= \lambda \left[\sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} + 1 \right] = \lambda[\lambda + 1] = \lambda^2 + \lambda\end{aligned}$$

Therefore,

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

□

8.5. Exponential Random Variables. A continuous random variable X is called an exponential random variable with parameter $\lambda > 0$, denoted $X \overset{\text{dist}}{=} \text{Exp}(\lambda)$, when the probability density

function is

$$f_X(t) = \begin{cases} \lambda e^{-\lambda t} & t > 0 \\ 0 & \text{otherwise} \end{cases}$$

As is clear from the density, the state space is $S_X = (0, \infty)$. Often, the parameter λ is called the *rate* of X.

PROPOSITION 8.8. *Let X be an exponential random variable with rate $\lambda > 0$; i.e., $X \overset{\text{dist}}{=} \text{Exp}(\lambda)$. Then $\mathbb{E}[X] = 1/\lambda$ and $\text{Var}(X) = 1/\lambda^2$.*

PROOF. Exercise (1). □

Exponential random variables play an invaluable role in the theory of continuous time Markov chains due to the *memoryless property* they satisfy. At a heuristic level, it is common to interpret an exponential random variable as a random alarm clock whose output is the random time (starting the count at time 0) at which the alarm rings. From this interpretation, the memoryless property tells us that if you were to notice that if you were to come upon one of these exponential alarm clocks which was already running, the distribution governing that alarm clock, even though it has been running for a potentially unknown quantity of time, is identical to the original exponential distribution – the alarm clock does not "remember" how long it has been running, so when you come upon the clock it may as well be restarting from time 0. In Chapter 2 Exercise (**??**) you will derive the memoryless property mathematically.

8.6. Normal Random Variables. A continuous random variable X is called *normal* with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, denoted $X \overset{\text{dist}}{=} \text{N}(\mu, \sigma^2)$, when the density is

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

A *standard normal* random variable is simply a normal random variable with $\mu = 0$ and $\sigma^2 = 1$.

PROPOSITION 8.9. *Let $X \overset{\text{dist}}{=} \text{N}(\mu, \sigma^2)$. Then $\mathbb{E}[X] = \mu$ and $\text{Var}(X) = \sigma^2$.*

PROOF. We start by calculating two helpful integrals. Let $\alpha \in \mathbb{R}$. Then,

$$\begin{aligned} \int_{-\infty}^{\infty} e^{-\alpha x^2}\, dt = \left[\iint_{\mathbb{R}^2} e^{-\alpha(x^2+y^2)}\, dxdy \right]^{1/2} &= \left[\iint_{\mathbb{R}^2} re^{-\alpha r^2}\, drd\theta \right]^{1/2} \\ &= \left[\frac{\pi}{\alpha} \int_0^{\infty} e^{-u}\, du \right]^{1/2} = \sqrt{\frac{\pi}{\alpha}}. \end{aligned}$$

Also,

$$\int_{-\infty}^{\infty} x^2 e^{-\alpha x^2}\,dt = -\frac{d}{d\alpha}\int_{-\infty}^{\infty} e^{-\alpha x^2}\,dx = -\frac{d}{d\alpha}\sqrt{\frac{2\pi}{\alpha}} = \frac{1}{2}\sqrt{\frac{2\pi}{\alpha^3}} = \sqrt{\frac{\pi}{2\alpha^3}}$$

Now, moving forward,

$$\begin{aligned}\mathbb{E}[X] = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} x e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}(x-\mu)e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx + \mu \\ &= \frac{\sigma^2}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-u}\,du + \mu = \mu\end{aligned}$$

where the second to last integral made the change of variable $u = (x-\mu)^2/2\sigma^2$. □

The cumulative distribution function of a standard normal random variable will be denoted by Φ. That is, if $Z \overset{\text{dist}}{=} \mathrm{N}(0,1)$, then

$$\Phi(t) = \mathbb{P}(Z \le t) = \int_{-\infty}^{t} f_Z(s)\,ds \tag{7}$$

where f_Z is the density of Z. From this, since f_Z is always positive, we realize that Φ is strictly increasing, and is therefore invertible on range $(0,1)$. Indeed, for any $p \in (0,1)$, the value $\Phi^{-1}(p)$ is well-defined; in fact, $\Phi^{-1}(p)$ gives the value t such that $\mathbb{P}(Z \le t) = p$.

THEOREM 8.10. *Let $X \overset{\text{dist}}{=} \mathrm{N}(\mu,\sigma^2)$ be a normal random variable with mean μ and variance σ^2, and let $Z \overset{\text{dist}}{=} \mathrm{N}(0,1)$ be standard normal random variable. Then*

$$X \overset{\text{dist}}{=} \sigma Z + \mu$$

In fact, this follows from equality

$$\mathbb{P}(X \le t) = \Phi\Big(\frac{t-\mu}{\sigma}\Big)$$

relating the cumulative distribution functions of X and Z.

PROOF. Let us note first that $X \overset{\text{dist}}{=} \sigma Z + \mu$ if and only if $\mathbb{P}(X \le t) = \mathbb{P}(\sigma Z + \mu \le t)$ for every $t \in \mathbb{R}$. However, $\mathbb{P}(\sigma Z + \mu \le t) = \mathbb{P}\Big(Z \le \frac{t-\mu}{\sigma}\Big) = \Phi\Big(\frac{t-\mu}{\sigma}\Big)$. So, if we can show that $\mathbb{P}(X \le t) = \Phi\Big(\frac{t-\mu}{\sigma}\Big)$, then we'll be done. To that end,

$$\mathbb{P}(X \le t) = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{t} e^{-\frac{(s-\mu)^2}{2\sigma^2}}\,ds$$

Letting $u = \frac{t-\mu}{\sigma}$, we have

$$\mathbb{P}(X \le t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{t-\mu}{\sigma}} e^{-\frac{u^2}{2}}\,du = \Phi\Big(\frac{t-\mu}{\sigma}\Big)$$

which finishes the proof. □

9. Advanced Properties of the Expected Value

9.1. Linearity and Positivity. Here we give two extremely important properties of the expected value, properties known as *linearity* and *positivity*. We describe them here in two theorems.

THEOREM 9.1 (Linearity of the Expected Value). *Let X and Y be random variables. For any scalar $\alpha \in \mathbb{R}$, it holds that*

$$\mathbb{E}[X + \alpha Y] = \mathbb{E}[X] + \alpha\,\mathbb{E}[Y].$$

PROOF. This is true in the jointly discrete case with

$$\begin{aligned}\mathbb{E}[X+\alpha Y] = \sum_{x\in S_X}\sum_{y\in S_Y}(x+\alpha y)p_{X,Y}(x,y) &= \sum_{x\in S_X}\sum_{y\in S_Y} x\,p_{X,Y}(x,y) + \alpha\sum_{x\in S_X}\sum_{y\in S_Y} y\,p_{X,Y}(x,y)\\ &= \sum_{x\in S_X} x\,p_X(x) + \alpha\sum_{y\in S_Y} y\,p_Y(y) = \mathbb{E}[X] + \alpha\mathbb{E}[Y]\end{aligned}$$

where the second to last equality followed after summing over the y variable in the first term resulting in the marginal p_X, and summing over the x variable in the second term resulting in the marginal p_Y. The analogous manipulation can be done in the jointly continuous case, using integrals instead of sums.

To fully prove the general case where X and Y may not be jointly discrete nor jointly continuous would be more involved than is appropriate for these notes. However, to give some insight into the advanced theory that is used to prove it, let us remark that the advanced mathematician has the groundwork to interpret the expected value as a more general integral which would be written $\mathbb{E}[W] = \int_\Omega W\,d\mathbb{P}$, and which we make no attempt to formalize nor define here. Suffice it to say that since an expected value is an integral and integrals have this linear property (for example, we used this in proving the jointly continuous case), it holds for expected value. □

THEOREM 9.2 (Positivity). *Let X and Y be random variables. If $\mathbb{P}(X \le Y) = 1$ then*

$$\mathbb{E}[X] \le \mathbb{E}[Y].$$

PROOF. If $\mathbb{P}(X \le Y) = 1$, then $\mathbb{P}(Y - X \ge 0) = 1$. By (2), this means that $0 \le \mathbb{E}[Y - X]$ since only the positive integral $\int_0^\infty \mathbb{P}(Y - X > t)\,dt$ is possibly non-zero; this is because the integrand in the negative integral $-\int_{-\infty}^0 \mathbb{P}(Y - X \le t)\,dt$ is 0. Using the linearity of the integral, $0 \le \mathbb{E}[Y - X] = \mathbb{E}[Y] - \mathbb{E}[X]$, which rearranges to $\mathbb{E}[X] \le \mathbb{E}[Y]$. □

10. Exercises

(1) Suppose that X is an exponential random variable with rate $\lambda > 0$; i.e., $X \overset{\text{dist}}{=} \text{Exp}(\lambda)$. Show that $\mathbb{E}[X] = 1/\lambda$ and $\text{Var}(X) = 1/\lambda^2$.

CHAPTER 2

Conditional Expectation

Throughout this chapter, we will assume that $(\Omega, \mathcal{B}, \mathbb{P})$ is a probability space.

1. Conditional Probability

DEFINITION 1.1. Let A and E be events. We define the *conditional probability* of A given E, denoted $\mathbb{P}(A \,|\, E)$, as

$$\mathbb{P}(A \,|\, E) = \frac{\mathbb{P}(A \cap E)}{\mathbb{P}(E)}$$

whenever $\mathbb{P}(E) > 0$. Otherwise, if $\mathbb{P}(E) = 0$, then we leave $\mathbb{P}(A \,|\, E)$ undefined, but use the convention that $\mathbb{P}(A \,|\, E) \times 0 = 0$, even when $\mathbb{P}(A \,|\, E)$ is undefined. $\triangle$

REMARK 1.1. We left $\mathbb{P}(A \,|\, E)$ undefined when $\mathbb{P}(E) = 0$, however defined $\mathbb{P}(A \,|\, E) \times 0 = 0$ regardless. This might seem strange, but the reason for this is that we will often see multiplication of terms of the form $\mathbb{P}(A \,|\, E)\,\mathbb{P}(E)$, which is now defined whether or not $\mathbb{P}(E) = 0$. $\triangle$

THEOREM 1.1. *Given an event E with positive probability, the function $\mathbb{P}(\,\cdot \,|\, E)$ is another probability on Ω. That is, $\mathbb{P}(\,\cdot \,|\, E)$ satisfies:*

(1) $0 \leq P(A \,|\, E) \leq 1$ *for every event A.*
(2) $\mathbb{P}(\varnothing \,|\, E) = 0$.
(3) $\mathbb{P}(\Omega \,|\, E) = 1$.
(4) *If $\{E_j\}_{j=1}^N$ are disjoint events for any $N \in \{1, 2, ..., \infty\}$, then*

$$\mathbb{P}\Big(\bigcup_{j=1}^{N} E_j \,\Big|\, E \Big) = \sum_{j=1}^{N} \mathbb{P}(E_j \,|\, E).$$

We now work through several lemmas which play an important role, both theoretically and calculationally.

LEMMA 1.2. *For any events A and E, the equality $\mathbb{P}(A \cap E) = \mathbb{P}(A \,|\, E)\mathbb{P}(E)$ holds.*

PROOF. We will prove this by considering two cases. The first case is that $\mathbb{P}(E) = 0$. This immediately implies that $\mathbb{P}(A \,|\, E)\,\mathbb{P}(E) = \mathbb{P}(A \,|\, E) \times 0 = 0$ by definition. On the other hand, since

$A \cap E \subseteq E$, it holds that $0 \leq \mathbb{P}(A \cap E) \leq \mathbb{P}(E) = 0$ and so $\mathbb{P}(A \cap E) = 0$. So, $\mathbb{P}(A \cap E) = \mathbb{P}(A \,|\, E)\mathbb{P}(E)$ in this case.

The alternate case is that $\mathbb{P}(E) > 0$. In this case

$$\mathbb{P}(A \,|\, E) = \frac{\mathbb{P}(A \cap E)}{\mathbb{P}(E)}.$$

By multiplying both sides of the equality by $\mathbb{P}(E)$, we get $\mathbb{P}(A \,|\, E)\mathbb{P}(E) = \mathbb{P}(E \cap A)$. □

LEMMA 1.3. *Let $\{E_j\}_{j=1}^N$ be disjoint events for some $N \in \{1, 2, ..., \infty\}$. For any event A, the equality*

$$\mathbb{P}\Big(A \cap \bigcup_{j=1}^{N} E_j\Big) = \sum_{j=1}^{N} \mathbb{P}(A \,|\, E_j)\mathbb{P}(E_j).$$

In particular, if it holds that $\mathbb{P}\big(\bigcup_{j=1}^{N} E_j\big) = 1$, then for any event A,

$$\mathbb{P}(A) = \sum_{j=1}^{N} \mathbb{P}(A \,|\, E_i)\, \mathbb{P}(E_i).$$

PROOF. Using basic set theory, we notice that $A \cap \bigcup_{j=1}^{N} E_j = \bigcup_{j=1}^{N} (A \cap E_j)$. Since the events $\{E_j\}_{j=1}^N$ are disjoint, so are the events $\{A \cap E_j\}_{j=1}^N$. From these observations, the properties of a probability, and Lemma 1.2,

$$\mathbb{P}\Big(A \cap \bigcup_{j=1}^{N} E_j\Big) = \mathbb{P}\Big(\bigcup_{j=1}^{N} (A \cap E_j)\Big) = \sum_{j=1}^{N} \mathbb{P}(A \cap E_j) = \sum_{j=1}^{N} \mathbb{P}(A \,|\, E_j)\mathbb{P}(E_j).$$

This establishes the first claimed equality. For the second, let $E = \bigcup_{j=1}^{N} E_j$ and assume that $\mathbb{P}(E) = 1$. We need to show that in this case, $\mathbb{P}(A \cap E) = \mathbb{P}(A)$. Notice first that $\mathbb{P}(A) = \mathbb{P}(A \cap E) + \mathbb{P}(A \setminus E)$ since $A = (A \cap E) \cup (A \setminus E)$ and the right hand side is a disjoint union. So, we need only show that $\mathbb{P}(A \setminus E) = 0$. To do so, we observe $(A \setminus E) \subseteq (\Omega \setminus E)$ and hence $0 \leq \mathbb{P}(A \setminus E) \leq \mathbb{P}(\Omega \setminus E) = \mathbb{P}(\Omega) - \mathbb{P}(E) = 1 - 1 = 0$. Therefore $\mathbb{P}(A) = \mathbb{P}(A \cap E)$. □

It is not uncommon to compound conditionings. We have already proved that if $\mathbb{P}(E) > 0$ then $\mathbb{P}(\,\cdot \,|\, E)$ is again a probability; for clarity in our discussion, let us denote $\mathbb{P}(\,\cdot \,|\, E)$ as $\mathbb{P}_E$. Now, suppose that A and F are events. Since $\mathbb{P}_E$ is a probability, it is perfectly reasonable to consider $\mathbb{P}_E(A \,|\, F)$. The natural question at this point is how to relate $\mathbb{P}_E(A \,|\, F)$ to the original probability $\mathbb{P}$. The following lemma answers this question.

LEMMA 1.4. *Suppose that $A, E,$ and F are events. Then*

$$\mathbb{P}_E(A \,|\, F) = \mathbb{P}(A \,|\, E \cap F)$$

whenever both sides of the equality are defined. Note here that we are using the notation $\mathbb{P}_E = \mathbb{P}(\,\cdot \,|\, E)$.

PROOF. Assuming both sides of the claimed equality are defined, then

$$\mathbb{P}_E(A \,|\, F) = \frac{\mathbb{P}_E(A \cap F)}{\mathbb{P}_E(F)} = \frac{\mathbb{P}(A \cap F \,|\, E)}{\mathbb{P}(F \,|\, E)} = \frac{\mathbb{P}(A \cap F \cap E)/\mathbb{P}(E)}{\mathbb{P}(F \cap E)/\mathbb{P}(E)} = \frac{\mathbb{P}(A \cap F \cap E)}{\mathbb{P}(F \cap E)} = \mathbb{P}(A \,|\, E \cap F)$$

which finishes the proof. □

In particular, combining Lemmas 1.3 and 1.4, we get

COROLLARY 1.5. *Let $\{E_j\}_{j=1}^N$ be disjoint events for some $N \in \{1, 2, ..., \infty\}$ and let E be another event. If it holds that $\mathbb{P}(\bigcup_{j=1}^N E_j \,|\, E) = 1$, then given any event A, it holds that*

$$\mathbb{P}(A \,|\, E) = \sum_{j=1}^N \mathbb{P}(A \,|\, E_j \cap E)\, \mathbb{P}(E_j \,|\, E).$$

PROOF. Applying Lemma 1.3,

$$\mathbb{P}_E(A) = \sum_{j=1}^N \mathbb{P}_E(A \,|\, E_j)\mathbb{P}_E(E_j)$$

where as above,a $\mathbb{P}_E = \mathbb{P}(\,\cdot \,|\, E)$. The result now follows since applying Lemma 1.4 for each j, $\mathbb{P}_E(A \,|\, E_j) = \mathbb{P}(A \,|\, E_j \cap E)$, and by definition, $\mathbb{P}_E(E_j) = \mathbb{P}(E_j \,|\, E)$. □

LEMMA 1.6 (Multiplication Rule). *For finitely many events $E_1, E_2, ..., E_N$,*

$$\begin{aligned} &P(E_1 \cap E_2 \cap \cdots \cap E_N) \\ &\quad = \mathbb{P}(E_1)\mathbb{P}(E_2 \,|\, E_1)\mathbb{P}(E_3 \,|\, E_2 \cap E_1) \cdots \mathbb{P}(E_N \,|\, E_1 \cap E_2 \cap \cdots \cap E_{N-1}). \end{aligned}$$

PROOF. Certainly this is true when $N = 1$, since there is nothing to prove. Also, by the definition of conditional probability, this is true when $N = 2$, since we have already proved $\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_2 \,|\, E_1)\mathbb{P}(E_1)$. For $N = 3$, we have

$$\mathbb{P}(E_1 \cap E_2 \cap E_3) = \mathbb{P}(E_3 \,|\, E_2 \cap E_1)\, \mathbb{P}(E_2 \cap E_1) = \mathbb{P}(E_3 \,|\, E_2 \cap E_1)\, \mathbb{P}(E_2 \,|\, E_1)\, \mathbb{P}(E_1).$$

Continuing this way inductively finishes the proof. □

2. Conditional Expectation on an Event

DEFINITION 2.1. Let X be a random variable and E an event. We define the *conditional expectation* of X given the *event* E as

$$\mathbb{E}[X \mid E] = \text{the expected value of } X \text{ using the probability } \mathbb{P}(\cdot \mid E)$$

whenever $\mathbb{P}(E) > 0$. Otherwise, if $\mathbb{P}(E) = 0$, then we leave $\mathbb{E}[X \mid E]$ undefined with the convention that $\mathbb{E}[X \mid E] \times 0 = 0$ even when $\mathbb{E}[X \mid E]$ is undefined. △

REMARK 2.1. For an event E with positive probability, we can use (2) to equivalently formulate the definition of the conditional expectation on E as

$$\mathbb{E}[X \mid E] = \int_0^\infty \mathbb{P}(X > t \mid E)\, dt - \int_{-\infty}^0 \mathbb{P}(X \leq t \mid E)\, dt. \tag{8}$$

This observation is extremely useful for some theoretic results which follow in the future. △

EXAMPLE 2.1. Consider rolling two fair dice. Let X be the random variable representing the outcome of the first roll and let E be the event that the sum of the two dice is 3. Let us calculate $\mathbb{E}[X \mid E]$. We have

$$\mathbb{E}[X \mid E] = \sum_{x=1}^{6} x\, \mathbb{P}(X = x \mid E)$$

Clearly, if $\mathbb{P}(X = x \mid E) = 0$ if $x \geq 3$ since the sum will always be greater than the first roll. We have

$$\mathbb{P}(X = 1 \mid E) = \frac{\mathbb{P}(X = 1,\ \text{sum is } 3)}{\mathbb{P}(\text{sum is } 3)} = \frac{1/36}{2/36} = 1/2$$

and

$$\mathbb{P}(X = 2 \mid E) = \frac{\mathbb{P}(X = 2,\ \text{sum is } 3)}{\mathbb{P}(\text{sum is } 3)} = \frac{1/36}{2/36} = 1/2.$$

Therefore,

$$\mathbb{E}[X \mid E] = \sum_{x=1}^{6} x\, \mathbb{P}(X = x \mid E) = 1 \cdot \mathbb{P}(X = 1 \mid E) + 2 \cdot \mathbb{P}(X = 1 \mid E) = 1/2(1 + 2) = 3/2.$$

△

We next present two lemmas which are often useful in practice. The first is fairly obvious; the second draws an analogy between the conditional expectation on an event and conditional probability.

LEMMA 2.1. *Suppose that E is an event with positive probability and X is a random variable such that on the event E, the random variable X is constantly equal to some fixed value x_0. That is, we suppose that $\mathbb{P}(X = x_0 \,|\, E) = 1$. Then $\mathbb{E}[X \,|\, E] = x_0$.*

PROOF. This is nearly immediate since $\mathbb{E}[X \,|\, E]$ is equal to the expected value of X with respect to the probability $\mathbb{P}(\cdot \,|\, E)$. By assumption $\mathbb{P}(X = x_0 \,|\, E) = 1$ and hence $\mathbb{E}[X \,|\, E] = x_0 \cdot \mathbb{P}(X = x_0 \,|\, E) + (\text{all other values}) \cdot \mathbb{P}(X \neq x_0 \,|\, E) = x_0 \cdot 1 + (\text{all other values}) \cdot 0 = x_0$. □

LEMMA 2.2. *If X is a random variable and E is an event with positive probability, then the equality*

$$\mathbb{E}[X \,|\, E] = \frac{\mathbb{E}[X 1_E]}{\mathbb{P}(E)}$$

holds.

PROOF. Since we are using the probability $\mathbb{P}(\,\cdot \,|\, E)$, we can write the expectation as,

$$\begin{aligned}
\mathbb{E}[X \,|\, E] &= \int_0^\infty \mathbb{P}(X > t \,|\, E)dt - \int_{-\infty}^0 \mathbb{P}(X \leq t \,|\, E)dt \\
&= \frac{1}{\mathbb{P}(E)} \left(\int_0^\infty \mathbb{P}(\{X > t\} \cap E)dt - \int_{-\infty}^0 \mathbb{P}(\{X \leq t\} \cap E)dt \right) \\
&= \frac{1}{\mathbb{P}(E)} \left(\int_0^\infty \mathbb{P}(X 1_E > t)dt - \int_{-\infty}^0 \mathbb{P}(X 1_E \leq t)dt \right) \\
&= \frac{1}{\mathbb{P}(E)} \mathbb{E}[X 1_E]
\end{aligned}$$

which is what needed to be shown. □

EXAMPLE 2.2. Let X be an exponential random variable representing the minutes after noon that Bucky arrives at a bus stop; assume the mean is 10 minutes. In this example, we will calculated the expected arrival time of Bucky, given that he arrives at least 15 minutes after noon. That is, we will calculate $\mathbb{E}[X \,|\, X \geq 15]$. To start, let's note that the parameter λ of the exponential random variable is $\lambda = 1/\mathbb{E}[X] = 1/10$. Using Lemma 2.2,

$$\mathbb{E}[X \,|\, X \geq 15] = \frac{\mathbb{E}[X 1_{X \geq 15}]}{\mathbb{P}(X \geq 15)}$$

For the denominator,

$$\mathbb{P}(X \geq 15) = \int_{15}^\infty \lambda e^{-\lambda t}\, dt = e^{-15\lambda} = e^{-15/10}$$

For the numerator, we first observe that $X1_{X\geq 15} = f(X)$ where

$$f(x) = \begin{cases} x & x \geq 15 \\ 0 & \text{otherwise} \end{cases}$$

Therefore

$$\mathbb{E}[X1_{X\geq 15}] = \mathbb{E}[f(X)] = \int_0^\infty f(x)\lambda e^{-\lambda x}\,dx = \lambda \int_{15}^\infty x e^{-\lambda x}\,dx = \frac{(15\lambda+1)e^{-15\lambda}}{\lambda} = 25e^{-15/10}.$$

Putting these together,

$$\mathbb{E}[X \mid X \geq 15] = \frac{25e^{-15/10}}{e^{-15/10}} = 25.$$

△

3. Conditional Expectation – Some Perspective

In what follows for the remainder of this chapter, we move to understand conditional expectations of the form $\mathbb{E}[Y \mid X_1, ..., X_N]$ where $Y, X_1, ..., X_N$ are (jointly distributed) random variables. We will spend this short section discussing one useful and intuitive perspective before ushering in the technical details.

A predictive interpretation goes as follows. Suppose that $X_1, ..., X_N$ are random variables which we have some grasp on (say, the outputs are easily observable quantities) and Y is a random variable which we want to try to predict using the knowledge of $X_1, ..., X_N$. Mathematically, this means that our goal is to find a function $g : \mathbb{R}^N \to \mathbb{R}$ such that $g(X_1, ..., X_N) \approx Y$, where the approximate equality means that, in some suitable sense, $g(X_1, ..., X_N)$ is the function of $X_1, ..., X_N$ which is nearest to Y. Once found, this function $g(X_1, ..., X_N)$ is the conditional expectation; i.e., $\mathbb{E}[Y \mid X_1, ..., X_N] = g(X_1, ..., X_N)$. From this perspective, evaluating g at the scalar values $x_1, ..., x_N \in \mathbb{R}$ we would interpret the output value $g(x_1, ..., x_N) = y$ as: given that $X_1 = x_1, X_2 = x_2, ...$, and $X_N = x_N$, the value we expect for Y is y.

Such a predictive scenario is familiar in life. For example, you could desire to predict the price Y of a car when given certain features of the car: $X_1 =$ age of car, $X_2 =$ mpg, $X_3 =$ miles driven, etc. Note that from a probabilistic perspective, this example has the set of all cars as the sample space with some probability associated to drawing a particular car from that set.

One important point to make here is that, although $g : \mathbb{R}^N \to \mathbb{R}$ is a deterministic (non-random) function, when evaluating g at the random variables $X_1, ..., X_N$, the result $g(X_1, ..., X_N) = \mathbb{E}[Y \mid X_1, ..., X_N]$ is a random variable. From a purely mathematical standpoint, this is because evaluating g at $X_1, ..., X_N$ is, in fact, a composition $g \circ (X_1, ..., X_N)$, and a function composed

with random variables results in a random variable. However, this is also understandable from our predictive perspective since $g(X_1, ..., X_N)$ is supposed to model the random variable Y, hence it is sensibly a random variable itself. While this point is often hidden in calculations and notation, we give a quick example here, further badgering its importance.

EXAMPLE 3.1. Suppose that we are considering only one input random variable X and have $g(X) = \mathbb{E}[Y \mid X]$ where $g : \mathbb{R} \to \mathbb{R}$ is the function $g(x) = x^2 - 1$. Evaluating g at the scalar value $x \in \mathbb{R}$ results in the scalar $x^2 - 1$; however, evaluating g at the random variable X results in the random variable $X^2 - 1 = \mathbb{E}[Y \mid X]$. Effectively, it might seem that we have simply made the notation change $x \to X$, but let us consider a couple of the many mathematical consequences of this change. First, recall that $X : \Omega \to \mathbb{R}$ is a function. When we write $g(X)$, we are in fact composing the function g with the function X; i.e., $g(X) = g \circ X$. Therefore, at the foundational level, $g(X) = g \circ X$ is a function taking in a value $\omega \in \Omega$ and outputting the value $g \circ X(\omega) = g(X(\omega)) = X(\omega)^2 - 1$. Second, consider taking the expected value $\mathbb{E}[g(X)]$ versus $\mathbb{E}[g(x)]$ for some $x \in \mathbb{R}$. In the first case, since X is a random variable, the result is $\mathbb{E}[g(X)] = \mathbb{E}[X^2 - 1] = \mathbb{E}[X^2] - 1$, and $\mathbb{E}[X^2] - 1 \neq g(X)$ except in the special case that X is constant. In the second case, for each $x \in \mathbb{R}$, $\mathbb{E}[g(x)] = \mathbb{E}[x^2 - 1] = x^2 - 1 = g(x)$, since for each fixed x, the value $x^2 - 1$ is deterministic (non-random). △

We end this section by noting a couple properties of conditional expectation that we can intuit from the predictive perspective, even though we will discuss them in rigorous detail below. The first property is that if $Y = f(X_1, ..., X_N)$ (that is, Y is a function of $X_1, ..., X_N$ from the start) then certainly $\mathbb{E}[Y \mid X_1, ..., X_N] = f(X_1, ..., X_N) = Y$, since certainly the nearest function of $X_1, ..., X_N$ to Y is $f(X_1, ..., X_N)$. Written more suggestively, this property states $\mathbb{E}[f(X_1, ..., X_N) \mid X_1, ..., X_N] = f(X_1, ..., X_N)$. The second property we can surmise is that if Y is independent of $X_1, ..., X_N$ then $\mathbb{E}[Y \mid X_1, ..., X_N]$ is a constant value. While this might not be as obvious as the previously mentioned property, it still is sensible after a bit of consideration. Indeed, suppose Y is independent of $X_1, ..., X_N$ and $g(X_1, ..., X_N) = \mathbb{E}[Y \mid X_1, ..., X_N]$. Then, as mentioned above, $g(x_1, ..., x_N) = y$ means that given $X_1 = x_1, ..., X_N = x_N$, the value we expect for Y is y. However, since Y is independent of $X_1, ..., X_N$, then changing the values $x_1, ..., x_N$ should not affect our guess for Y (otherwise, we would be asserting that the output value of Y has some dependence on the values of the X_is, conflicting with our assumption that Y is independent of the X_is). Hence, it is sensible that the function g should remain at constant value y_0; i.e., $g(x_1, ..., x_N) = y_0$ for all values $x_1, ..., x_N$. Therefore, $\mathbb{E}[Y \mid X_1, ..., X_N] = g(X_1, ..., X_N) = y_0$. What we discover below is that this constant value is y_0 is $\mathbb{E}[Y]$, which is the single fixed value nearest to Y in our sense

of *nearest*. You might argue that in this independent case, $g(X_1, ..., X_N)$ equalling the constant value y_0 contradicts our previous discussion that $g(X_1, ..., X_N)$ is a random variable; however, $g(X_1, ..., X_N)$ is a random variable for which it so happens that $\mathbb{P}(g(X_1, ..., X_N) = y_0) = 1$.

4. Discrete Case

Let Y be a random variable. We start learning how to find the function $g : \mathbb{R}^N \to \mathbb{R}$ such that $g(X_1, ..., X_N) = \mathbb{E}[Y \mid X_1, ..., X_N]$ with the simplest case where $X_1, ..., X_N$ are all discrete. In this case, let S_i be the state space for X_i for $i = 1, ..., N$. We hence define $g(x_1, ..., x_N) = \mathbb{E}[Y \mid X_1 = x_1, ..., X_N = x_N]$ for each $x_1 \in S_1, ..., x_N \in S_N$. Note that $\{X_1 = x_1, ..., X_N = x_N\}$ is an event, and we have already defined how to find the conditional expectation on an event, so we already have the technology to interpret $\mathbb{E}[Y \mid X_1 = x_1, ..., X_N = x_N]$.

Example 4.1. Suppose that X_1 and X_2 are both Bernoulli random variables. Suppose further that Y is a random variable such that given the event $\{X_1 = 0, X_2 = 0\}$, it happens that Y is distributed as $\text{Bin}(3, 1/2)$; on the event $\{X_1 = 0, X_2 = 1\}$, it happens that Y is distributed as $\text{Pois}(1)$; on the event $\{X_1 = 1, X_2 = 0\}$, it happens that Y is distributed as $\text{N}(2, 4)$; and on the event $\{X_1 = 1, X_2 = 1\}$, it happens that Y is constantly 0. By definition, for $x_1, x_2 \in \{0, 1\}$,

$$g(x_1, x_2) = \mathbb{E}[Y \mid X_1 = x_1, X_2 = x_2] = \begin{cases} \mathbb{E}[Y \mid X_1 = 0, X_2 = 0] & \text{if } x_1 = 0, x_2 = 0 \\ \mathbb{E}[Y \mid X_1 = 0, X_2 = 1] & \text{if } x_1 = 0, x_2 = 1 \\ \mathbb{E}[Y \mid X_1 = 1, X_2 = 0] & \text{if } x_1 = 1, x_2 = 0 \\ \mathbb{E}[Y \mid X_1 = 1, X_2 = 1] & \text{if } x_1 = 1, x_2 = 1 \end{cases}$$

$$= \begin{cases} 3/2 = \mathbb{E}[\text{Bin}(3, 1/2)] & \text{if } x_1 = 0, x_2 = 0 \\ 1 = \mathbb{E}[\text{Pois}(1)] & \text{if } x_1 = 0, x_2 = 1 \\ 2 = \mathbb{E}[\text{N}(2, 4)] & \text{if } x_1 = 1, x_2 = 0 \\ 0 = \mathbb{E}[0] & \text{if } x_1 = 1, x_2 = 1 \end{cases}$$

Therefore, the conditional expectation of Y with respect to X_1 and X_2 is

$$\mathbb{E}[Y \mid X_1, X_2] = g(X_1, X_2) = \begin{cases} 3/2 & \text{if } X_1 = 0, X_2 = 0 \\ 1 & \text{if } X_1 = 0, X_2 = 1 \\ 2 & \text{if } X_1 = 1, X_2 = 0 \\ 0 & \text{if } X_1 = 1, X_2 = 1 \end{cases}$$

△

REMARK 4.1. In the previous example, we made a supposition of the form: *on (or given) the event E, the random variable Y is distributed as D.* (For example, this occurred when we wrote: on the event $\{X_1 = 0, X_2 = 0\}$, it happens that Y is distributed as Bin$(3, /12)$). What this means is that the distribution of Y with respect to the probability $\mathbb{P}(\cdot \,|\, E)$ is identical to the distribution of the random variable D (with respect to whatever probability is defined on its sample space). Therefore, by the definition of the conditional expectation on an event, it then follows that $\mathbb{E}[Y \,|\, E] = \mathbb{E}[D]$. Because this is a frequently occurring construct, the aforementioned phrase "*on the event E, the random variable Y is distributed as D*" is often symbolically shorthanded by simply writing $Y \,|\, E \overset{\text{dist}}{=} D$. △

REMARK 4.2. In the previous example our goal was to find a function $g : \mathbb{R}^2 \to \mathbb{R}$ such that $g(X_1, X_2) = \mathbb{E}[Y \,|\, X_1, X_2]$. Our solution found $g(x_1, x_2)$ for $x_1, x_2 \in \{0, 1\}$ but ignored all other values $x_1, x_2 \in \mathbb{R} \setminus \{0, 1\}$. However, this was not an egregious oversight. Since after all is done, we are concerned with $g(X_1, X_2)$ with X_1 and X_2 only taking values in $\{0, 1\}$, we don't worry about how the function g behaves outside of these values – evaluating g at X_1 and X_2 essentially ignores the behaivior of g at any points outside of the state spaces of X_1 and X_2. Because of this, there is no uniqueness assured for the function g – we could define it multiple ways for $x_1, x_2 \notin \mathbb{R} \setminus \{0, 1\}$; however, this ambiguity in the definition of the function g is illusory in practice since there is a uniqueness for $g(X_1, X_2)$, after evaluating g at X_1 and X_2. △

PROPOSITION 4.1. *Let $X_1, ..., X_n$ be a discrete random variables with joint mass function $p(x_1, ..., x_N) = \mathbb{P}(X_1 = x_1, ..., X_N = x_N)$. For any event E and random variable Y, the equalities*

$$\mathbb{P}(E) = \sum_{x_N \in S_N} \cdots \sum_{x_1 \in S_1} \mathbb{P}(E \,|\, X_1 = x_1, ..., X_N = x_N)\, p(x_1, ..., x_N) \tag{9}$$

and

$$\mathbb{E}[Y] = \sum_{x_N \in S_N} \cdots \sum_{x_1 \in S_1} \mathbb{E}[Y \,|\, X_1 = x_1, ..., X_N = x_N]\, p(x_1, ..., x_N). \tag{10}$$

hold, where S_i is the state space of X_i for $i = 1, ..., N$.

PROOF. Equation (9) holds immediately from Lemma 1.3 above since all events of the form $\{X_1 = x_1, ..., X_N = x_N\}$ for $x_i \in S_i$ partition the sample space Ω. We will prove (10) while conditioning on a single random variable X to make the notation less egregous, but the argument

for several variables works analogously. By what we have just proved,

$$\begin{aligned}
\mathbb{E}[Y] &= \int_0^\infty \mathbb{P}(Y > t)\,dt - \int_{-\infty}^0 \mathbb{P}(Y \le t)\,dt \\
&= \int_0^\infty \Big[\sum_{x \in S_X} \mathbb{P}(Y > t \,|\, X = x)\, p_X(x) \Big]\, dt - \int_{-\infty}^0 \Big[\sum_{x \in S_X} \mathbb{P}(Y \le t \,|\, X = x)\, p_X(x) \Big]\, dt \\
&= \sum_{x \in S_X} \left[\int_0^\infty \mathbb{P}(Y > t \,|\, X = x)\,dt - \int_{-\infty}^0 \mathbb{P}(Y \le t \,|\, X = x)\,dt \right] p_X(x) \\
&= \sum_{x \in S_X} \mathbb{E}[Y \,|\, X = x]\, p_X(x).
\end{aligned}$$

This concludes the proof. □

4.1. Jointly Discrete Case. Assume that X and Y are jointly discrete with joint mass function $p(x, y) = \mathbb{P}(X = x, Y = y)$. Let us calculate $\mathbb{E}[Y \,|\, X]$ in this case. For $x \in S_X$, we have

$$\mathbb{E}[Y \,|\, X = x] = \text{ expected value of } Y \text{ with respect to } \mathbb{P}(\cdot \,|\, X = x) = \sum_{y \in S_Y} y\, \mathbb{P}(Y = y \,|\, X = x).$$

From this equation, we define the conditional mass function $p(y \,|\, x)$ as

$$p(y \,|\, x) = \mathbb{P}(Y = y \,|\, X = x) \tag{11}$$

from which we arrive at the equation $\mathbb{E}[Y \,|\, X = x] = \sum_{y \in S_Y} y\, p(y \,|\, x)$. Let us summarize these observations in the following lemma.

THEOREM 4.2. *Suppose that X and Y are jointly discrete random variables with conditional mass function $p(y \,|\, x) = \mathbb{P}(Y = y \,|\, X = x)$ for each $y \in S_Y$ and $x \in S_X$. If we define $g(x) = \mathbb{E}[Y \,|\, X = x]$ for each $x \in S_X$, then $g(x) = \sum_{y \in S_Y} y\, p(y \,|\, x)$ and $\mathbb{E}[Y \,|\, X] = g(X)$.*

PROOF. This lemma follows straight from the preceding calculations and the definition of $\mathbb{E}[Y \,|\, X]$ in the discrete setting. □

Before extending these concepts to more than two random variables, we make a final observation.

LEMMA 4.3. *Let X and Y be jointly discrete random variables with joint mass function $p(x, y) = \mathbb{P}(X = x, Y = y)$ and conditional mass function $p(y \,|\, x) = \mathbb{P}(Y = y \,|\, X = x)$. Then the equality*

$$p(y \,|\, x) = \frac{p(x, y)}{p_X(x)} = \frac{p(x, y)}{\sum_{y \in S_Y} p(x, y)}$$

holds. As usual, the function p_X is the (marginal) mass function of X. In particular,

$$\mathbb{E}[Y \,|\, X = x] = \frac{\sum\limits_{y \in S_Y} y\, p(x,y)}{p_X(x)}$$

for each $x \in S_X$.

Proof. By definition, $p(y \,|\, x) = \mathbb{P}(Y = y \,|\, X = x)$, where we are assuming $y \in S_Y$ and $x \in S_X$. Now,

$$\mathbb{P}(Y = y \,|\, X = x) = \frac{\mathbb{P}(Y = y, X = x)}{\mathbb{P}(X = x)} = \frac{p(x,y)}{p_X(x)}.$$

Since we have already proved that $p_X(x) = \sum\limits_{y \in S_Y} p(x,y)$, the equality follows. In particular,

$$\mathbb{E}[Y \,|\, X = x] = \sum_{y \in S_Y} y\, p(y \,|\, x) = \sum_{y \in S_Y} \frac{y\, p(x,y)}{p_X(x)} = \frac{\sum\limits_{y \in S_Y} y\, p(x,y)}{p_X(x)}$$

finishing the proof. □

Example 4.2. Suppose that X and Y are jointly discrete random variables with joint mass function

$$p(x,y) = \begin{cases} \frac{3}{13y} & x, y \in \{1,2,3\} \text{ with } 1 \le y \le x \\ 0 & \text{otherwise} \end{cases}$$

In this example we will find $\mathbb{E}[Y \,|\, X]$. To start,

$$p_X(x) = \sum_{y \in S_Y} p(x,y) = \frac{3}{13} \sum_{y=1}^{x} \frac{1}{y} = \begin{cases} \frac{3}{13} = \frac{3}{13}\Big(1\Big) & x = 1 \\ \frac{9}{26} = \frac{3}{13}\Big(1 + \frac{1}{2}\Big) & x = 2 \\ \frac{33}{78} = \frac{3}{13}\Big(1 + \frac{1}{2} + \frac{1}{3}\Big) & x = 3 \end{cases}$$

Next,

$$\sum_{y \in S_Y} y\, p(x,y) = \begin{cases} \frac{3}{13} = \frac{3}{13}\Big(1\Big) & x = 1 \\ \frac{6}{13} = \frac{3}{13}\Big(1 + 2 \times \frac{1}{2}\Big) & x = 2 \\ \frac{9}{13} = \frac{3}{13}\Big(1 + 2 \times \frac{1}{2} + 3 \times \frac{1}{3}\Big) & x = 3 \end{cases}$$

Therefore

$$\mathbb{E}[Y \,|\, X = x] = \frac{\sum\limits_{y \in S_Y} y\, p(x,y)}{p_X(x)} = \begin{cases} 1 = \frac{3/13}{3/13} & x = 1 \\ \frac{4}{3} = \frac{6/13}{9/26} & x = 2 \\ \frac{18}{11} = \frac{9/13}{33/78} & x = 3 \end{cases}$$

from which we finally deduce

$$\mathbb{E}[Y \,|\, X] = \begin{cases} 1 & X = 1 \\ \frac{4}{3} & X = 2 \\ \frac{18}{11} & X = 3 \end{cases}$$

△

EXAMPLE 4.3. Suppose that N is a discrete random variable with the state space of all non-negative integers $\{1, 2, 3, ...\}$. Define the random variable X by $X = \frac{1}{2}N(N+1)$. Note that X is a function of N here, so $X = f(N)$ where $f(n) = n(n+1)/2$. From the prediction perspective given above, we should then expect that $\mathbb{E}[X \,|\, N] = X$. We quickly show that this is true. For any $n \in \{1, 2, 3, ...\}$, we have $\mathbb{E}[X \,|\, N = n] = \mathbb{E}\Big[\frac{1}{2}N(N+1) \,\Big|\, N = n\Big] = \mathbb{E}\Big[\frac{1}{2}n(n+1) \,\Big|\, N = n\Big] = \frac{1}{2}n(n+1)$, since for a fixed n, the value $\frac{1}{2}n(n+1)$ is constant (see Lemma 2.1). Therefore, letting $g(n) = \mathbb{E}[X \,|\, N = n]$ we find that $g(n) = \frac{1}{2}n(n+1)$. Hence $\mathbb{E}[X \,|\, N] = g(N) = \frac{1}{2}N(N+1) = X$. △

To finish this section, we consider the case of more than two jointly discrete random variables. Suppose that $X_1, ..., X_N, Y$ are jointly discrete random variables. Following the analogous arguments as above, we find that

$$\mathbb{E}[Y \,|\, X_1 = x_1, ..., X_n = x_n] = \sum_{y \in S_Y} y\, \mathbb{P}(Y = y \,|\, X_1 = x_1, ..., X_N = x_N) = \sum_{y \in S_Y} y\, p(y \,|\, x_1, ..., x_N)$$

where, analogous to before, we define the conditional mass function

$$p(y \,|\, x_1, ..., x_N) = \mathbb{P}(Y = y \,|\, X_1 = x_1, ..., X_N = x_N).$$

In particular, defining $g : \mathbb{R}^N \to \mathbb{R}$ by

$$g(x_1, ..., x_N) = \mathbb{E}[Y \,|\, X_1 = x_1, ..., X_N = x_N] = \sum_{y \in S_Y} y\, p(y \,|\, x_1, ..., x_N)$$

we find that $\mathbb{E}[Y \,|\, X_1, ..., X_N] = g(X_1, ..., X_N)$. That is, as expected, Theorem 4.2 holds over in the case of multiple jointly discrete random variables. It should also be clear that Lemma 4.3 holds over in this case as well, resulting in the equalities

$$p(y \,|\, x_1, ..., x_N) = \frac{p(x_1, ..., x_N, y)}{\sum\limits_{y \in S_Y} p(x_1, ..., x_N, y)}$$

whenever the denominator is non-zero and where $p(x_1, ..., x_N, y)$ is the joint mass function of $X_1, ..., X_N, Y$; and hence

$$\mathbb{E}[Y \mid X_1 = x_1, ..., X_N = x_N] = \frac{\sum_{y \in S_Y} y\, p(x_1, ..., x_N, y)}{\sum_{y \in S_Y} p(x_1, ..., x_N, y)}$$

Although we did not explicitly write it, do note that $\sum_{y \in S_Y} p(x_1, ..., x_N, y)$ results in the (marginal) mass function $p_{X_1,...,X_N}(x_1, ..., x_N)$ for $X_1, ..., X_N$.

5. Jointly Continuous Case

Let X and Y be jointly continuous random variables with joint density $f(x, y)$. Motivated by Lemma 4.3, we define the conditional density $f(y \mid x)$ as

$$f(y \mid x) = \frac{f(x, y)}{f_X(x)} = \frac{f(x, y)}{\int_{-\infty}^{\infty} f(x, y)\, dy} \tag{12}$$

for every x such that $f_X(x) > 0$; otherwise, if $f_X(x) = 0$, we define $f(y \mid x) = 0$. From here, we *define* the conditional expectation of Y with respect to X to be $\mathbb{E}[Y \mid X] = g(X)$ where

$$g(x) = \int_{-\infty}^{\infty} y\, f(y \mid x)\, dy \tag{13}$$

Compare this definition int the jointly continuous case to Theorem 4.2 – you will see a direct analogue.

NOTATION 5.1. In the case that X and Y are jointly continuous, $\mathbb{P}(X = x) = 0$ since X is a continuous random variable. Hence, the expression $\mathbb{E}[Y \mid X = x]$ is not defined. However, because of the clear analogies with the discrete case, it is common to write $\mathbb{E}[Y \mid X = x]$ regardless. How to interpret this is that $\mathbb{E}[Y \mid X = x]$ is the function $g(x)$ in (13). △

As is expected, we can easily analogize our way into the setting with multiple random variables $X_1, ..., X_N, Y$ which are jointly continuous. In this case, letting $f(x_1, ..., x_N, y)$ be the joint mass function and defining the conditional mass function

$$f(y \mid x_1, ..., x_N) = \frac{f(x_1, ..., x_N, y)}{\int_{-\infty}^{\infty} f(x_1, ..., x_N, y)\, dy}$$

whenever the denominator is non-zero; otherwise, when the denominator is zero, define $f(y \mid x_1, ..., x_N) = 0$; note that $\int_{-\infty}^{\infty} f(x_1, ..., x_N, y)\, dy$ is the (marginal) joint density of $X_1, ..., X_N$. Hence we define the conditional expectation of Y with respect to $X_1, ..., X_N$ as $\mathbb{E}[Y \mid X_1, ..., X_N] = g(X_1, ..., X_N)$

where

$$g(x_1, ..., x_N) = \int_{-\infty}^{\infty} y\, f(y \mid x_1, ..., x_N)\, dy. \tag{14}$$

As is the case for two random variables discussed in Notation 5.1, we will abuse notation and often write $\mathbb{E}[Y \mid X_1 = x_1, ..., X_N = x_N]$ as the function $g(x_1, ..., x_N)$ in (14).

We now present the analogue to (10) above.

PROPOSITION 5.1. *Let $X_1, ..., X_N, Y$ be jointly continuous and let $f(x_1, ..., x_N)$ be the marginal density of $X_1, ..., X_N$. The equality*

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbb{E}[Y \mid X_1 = x_1, ..., X_N = x_N] f(x_1, ..., x_N)\, dx_1 \cdots dx_N. \tag{15}$$

holds.

PROOF. Let $\rho(x_1, ..., x_N, y)$ be the joint density of $X_1, ..., X_N, Y$. We start by observing

$$\mathbb{E}[Y \mid X_1 = x_1, ..., X_N = x_n] f_X(x_1, ..., x_N) = \int_{-\infty}^{\infty} y\, \rho(x_1, ..., x_N, y)\, dy$$

since either $f_X(x_1, ..., x_N) \neq 0$, in which case

$$\begin{aligned}\mathbb{E}[Y \mid X_1 = x_1, ..., X_N = x_n] f_X(x_1, ..., x_N) &= \frac{\int_{-\infty}^{\infty} y\, \rho(x_1, ..., x_N, y)\, dy}{f(x_1, ..., x_N)} f(x_1, ..., x_N) \\ &= \int_{-\infty}^{\infty} y\, \rho(x_1, ..., x_N, y)\, dy.\end{aligned}$$

or, if $f(x_1, ..., x_N) = 0$, then both sides of the equality are 0. Therefore,

$$\begin{aligned}&\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbb{E}[Y \mid X_1 = x_1, ..., X_N = x_N] f(x_1, ..., x_N)\, dx_1 \cdots dx_N \\ &\quad = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} y\, \rho(x_1, ..., x_N, y)\, dy \right] dx_1 \cdots dx_N \\ &\quad = \int_{-\infty}^{\infty} y \left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \rho(x_1, ..., x_N, y)\, dx_1 \cdots dx_N \right] dy \\ &\quad = \int_{-\infty}^{\infty} y\, f_Y(y)\, dy = \mathbb{E}[Y].\end{aligned}$$

The penultimate inequality followed from the fact that after integrating over $x_1, ..., x_N$, we were left with the marginal density for Y. □

6. Nearness of Random Variables and the General Case

When first introducing the concept of a conditional expectation, the emergent theme was that $g(X_1, ..., X_N) = \mathbb{E}[Y \mid X_1, ..., X_N]$ can be understood as meaning that $g(X_1, ..., X_N)$ is the function

of $X_1, ..., X_N$ which is closest to Y by some notion of distance. Here we describe what that notion of distance is between random variables and from this, one way to interpret conditional expectations from the vantage point of linear algebra. We will relegate the heavy linear algebra details to the appendices for those interested. The culmination of these details are Theorems 6.1 and 6.6 below; hence, the proofs of these Theorems appear in the appendices, but are omitted here.

REMARK 6.1. The assumption throughout this section is that all random variables we consider have a finite variance, otherwise this perspective needs to be slightly modified to completely address what follows. That said, most if not all random variables we consider in these notes will have a finite variance, so we do not need to concern ourselves beyond this assumption. △

DEFINITION 6.1. Given random variables X and Y, we will define the (L^2-)*distance* between them, denoted $\|X - Y\|$, as

$$\|X - Y\| = \sqrt{\mathbb{E}\Big[\big|X - Y\big|^2\Big]}$$

△

Upon some cursory examination, you will notice that this has a similar form to the standard Euclidean distance used in $\mathbb{R}^n$ – the square-root of the sum of the squares. In fact, this observation is correct; this distance between random variables is the generalization of the Euclidean distance.

THEOREM 6.1. *Given random variables $X_1, ..., X_N, Y$ there exists a unique random variable of the form $g(X_1, ..., X_N)$ such that*

$$\big\|g(X_1, ..., X_N) - Y\big\| \leq \big\|h(X_1, ..., X_N) - Y\big\| \tag{16}$$

for all other random variables of the form $h(X_1, ..., X_N)$. Poetically, this says there exists a unique function of $X_1, ..., X_N$ which is closest to Y. Moreover, this random variable $g(X_1, ..., X_N)$ in (16) is the unique random variable satisfying the orthogonality property

$$\mathbb{E}\big[g(X_1, ..., X_N)\, h(X_1, ..., X_N)\big] = \mathbb{E}\big[Y\, h(X_1, ..., X_N)\big] \tag{17}$$

for all other random variables of the form $h(X_1, ..., X_N)$.

REMARK 6.2. To be clear, by considering a random variable of the form $g(X_1, ..., X_N)$, what this notation implicitly means is that there is a function $g : \mathbb{R}^N \to \mathbb{R}$ which is evaluated at (composed with) $X_1, ..., X_N$. Note that it is possible for such a function to be constant, for example it is possible that $g(x_1, ..., x_N) = \mathbb{E}[Y]$ since the expected value $\mathbb{E}[Y]$ is simply some scalar value; of course, in this case $g(X_1, ..., X_N) = \mathbb{E}[Y]$. △

DEFINITION 6.2. The random variable $g(X_1, ..., X_N)$ defined in (16) or (17) is the *conditional expectation* of Y with respect to $X_1, ..., X_N$ and hence we denote $g(X_1, ..., X_N)$ by $\mathbb{E}[Y \mid X_1, ..., X_N]$. Keeping consistent with previous notation, it is not uncommon to denote the function g by $g(x_1, ..., x_N) = \mathbb{E}[Y \mid X_1 = x_1, ..., X_N = x_N]$. △

COROLLARY 6.2 (Law of Total Expectation). *For random variables $X_1, ..., X_N, Y$, the equality*

$$\mathbb{E}\Big[\mathbb{E}[Y \mid X_1, ..., X_N]\Big] = \mathbb{E}[Y]$$

holds.

PROOF. Let $g(X_1, ..., X_N) = \mathbb{E}[Y \mid X_1, ..., X_N]$. Define the constant random variable $h(X_1, ..., X_N) = 1$, from which by (17),

$$\begin{aligned}\mathbb{E}[Y] = \mathbb{E}[Y \cdot 1] = \mathbb{E}[Y\, h(X_1, ..., X_N)] = \mathbb{E}[g(X_1, ..., X_N)\, h(X_1, ..., X_N)] \\ = \mathbb{E}[g(X_1, ..., X_N) \cdot 1] = \mathbb{E}[g(X_1, ..., X_N)],\end{aligned}$$

concluding the proof. □

COROLLARY 6.3. *For random variables $X_1, ..., X_N, Y$, the equality $Y = \mathbb{E}[Y \mid X_1, ..., X_N]$ holds if and only if Y is a function of $X_1, ..., X_N$; that is, if and only if $Y = g(X_1, ..., X_N)$ for some function $g : \mathbb{R}^N \to \mathbb{R}$.*

PROOF. Since $\mathbb{E}[Y \mid X_1, ..., X_N]$ is a function of $X_1, ..., X_N$, then certainly if $Y = \mathbb{E}[Y \mid X_1, ..., X_N]$ then Y is a function of $X_1, ..., X_N$. To prove the other direction, assume that $Y = g(X_1, ..., X_N)$. Then,

$$\|g(X_1, ..., X_N) - Y\| = \|Y - Y\| = \|0\| = 0 \leq \|h(X_1, ..., X_N) - Y\|$$

for any other random variable of the form $h(X_1, ...X_N)$. By (16), this implies that $g(X_1, ..., X_N) = \mathbb{E}[Y \mid X_1, ..., X_N]$ which is equivalent to $Y = \mathbb{E}[Y \mid X_1, ..., X_N]$. □

COROLLARY 6.4. *Suppose that $X_1, ..., X_N$, and Y are random variables such that Y is independent from X_i for each $i = 1, ..., N$. Then $\mathbb{E}[Y \mid X_1, ..., X_N] = \mathbb{E}[Y]$.*

PROOF. Consider any random variable of the form $h(X_1, ..., X_N)$. Since Y is independent of the X_is, then Y is independent of $h(X_1, ..., X_N)$. Therefore

$$\mathbb{E}[Y\, h(X_1, ..., X_N)] = \mathbb{E}[Y]\, \mathbb{E}[h(X_1, ..., X_N)] = \mathbb{E}\Big[\mathbb{E}[Y]\, h(X_1, ..., X_N)\Big]$$

Define $g : \mathbb{R}^N \to \mathbb{R}$ as the constant function $g(x_1, ..., x_N) = \mathbb{E}[Y]$. Our above equalities then imply that

$$\mathbb{E}[Y\, h(X_1, ..., X_N)] = \mathbb{E}\Big[\mathbb{E}[Y]\, h(X_1, ..., X_N)\Big] = \mathbb{E}[g(X_1, ..., X_N)\, h(X_1, ..., X_N)]$$

which, by (17), implies that $g(X_1, ..., X_N) = \mathbb{E}[Y \,|\, X_1, ..., X_N]$. Since $g(X_1, ..., X_N) = \mathbb{E}[Y]$, we have finished the proof. □

COROLLARY 6.5. *Let $X_1, ..., X_N, Y$ be random variables. For any random variable of the form $h(X_1, ..., X_N)$, the equality*

$$\mathbb{E}[h(X_1, ..., X_N)\, Y \,|\, X_1, ..., X_N] = h(X_1, ..., X_N)\, \mathbb{E}[Y \,|\, X_1, ..., X_N]$$

holds.

PROOF. For simplicity of notation, we consider only one random variable X being conditioned on and let $g(X) = \mathbb{E}[Y \,|\, X]$; the general case holds by the analogous argument. We want to show that $\mathbb{E}[h(X)\, Y \,|\, X] = h(X)\, g(X)$. To do so, by (17), we can argue that $\mathbb{E}[h(X)\, Y\, k(X)] = \mathbb{E}[h(X)\, g(X)\, k(X)]$ for any random variable of the form $k(X)$. However, noting that the product $h(X)k(X)$ is itself a random variable which is a function of X, then the equality $\mathbb{E}[Y\, h(X)k(X)] = \mathbb{E}[g(X)\, h(X)k(X)]$ holds by (17), which, after very minor rearrangement, is what we needed to show. □

We conclude this section with what is the apparent analogue of Section 9.1 above; that conditional expectations satisfy linearity and positivity. As mentioned at the outset of this section, the proof of Theorem 6.6 is below in the appendices, since the techniques required for its proof come directly from the linear algebraic techniques discussed therein.

THEOREM 6.6. *Let $X_1, ..., X_N$ be random variables. Considered as a function inputting and outputting random variables, the conditional expectation on $X_1, ..., X_N$ is linear and satisfies positivity. That is, for any random variables Y_1 and Y_2 and scalar $\alpha \in \mathbb{R}$, the equality*

$$\mathbb{E}[Y_1 + \alpha Y_2 \,|\, X_1, ..., X_N] = \mathbb{E}[Y_1 \,|\, X_1, ..., X_N] + \alpha\, \mathbb{E}[Y_2 \,|\, X_1, ..., X_N]$$

holds; if $Y_1 \leq Y_2$, then the inequality $\mathbb{E}[Y_1 \,|\, X_1, ..., X_N] \leq \mathbb{E}[Y_2 \,|\, X_1, ..., X_N]$ holds.

7. Exercises

(1) Let X be an exponential random variable with rate $\lambda > 0$. In this problem you will show that X satisfies the *memoryless property*. Let $s \geq 0$ and $t > 0$. Show that

$$\mathbb{P}(X > t + s \,|\, X > s) = e^{-\lambda t}.$$

(2) In this problem you will show several manifestations of *Bayes' Formula*. Let O and E be events, both with positive probability. While working on this problem, to gain some intuition towards why the formula is often used, think of O as an outcome or observation and E as evidence which may contribute to the chances of O occurring.

(a) Show that the equality

$$\mathbb{P}(O \,|\, E) = \frac{\mathbb{P}(E \,|\, O)\,\mathbb{P}(O)}{\mathbb{P}(E)}$$

holds.

(b) Suppose that $\{O_j\}_{j=1}^N$ with $N \in \{1, 2, ..., \infty\}$ are disjoint events such that $\mathbb{P}(\bigcup_{j=1}^{N} O_j) = 1$. Show that the equality

$$\mathbb{P}(O \,|\, E) = \frac{\mathbb{P}(E \,|\, O)\,\mathbb{P}(O)}{\sum_{j=1}^{N} \mathbb{P}(E \,|\, O_j)\,\mathbb{P}(O_j)}$$

holds.

(c) As a special case to the previous part, justify that the equality

$$\mathbb{P}(O \,|\, E) = \frac{\mathbb{P}(E \,|\, O)\,\mathbb{P}(O)}{\mathbb{P}(E \,|\, O)\,\mathbb{P}(O) + \mathbb{P}(E \,|\, O^c)\,\mathbb{P}(O^c)}$$

holds, where O^c is the complement of O in Ω; that is, $O^c = \Omega \setminus O$.

(3) Suppose that X is a discrete random variable. If Y is another random variable which is independent from X, by explicitly calculating $\mathbb{E}[Y \,|\, X = x]$, show that $\mathbb{E}[Y \,|\, X] = \mathbb{E}[Y]$. *Hint*: Remember that $\mathbb{E}[Y \,|\, X = x] = \int_0^\infty \mathbb{P}(Y > t \,|\, X = x)\, dt - \int_{-\infty}^0 \mathbb{P}(Y \leq t \,|\, X = x)\, dt$.

(4) Suppose that X is a discrete random variable and $Y = f(X)$ for some function $f : \mathbb{R} \to \mathbb{R}$. By explicitly calculating $\mathbb{E}[Y \,|\, X = x]$, show that $\mathbb{E}[Y \,|\, X] = Y$.

(5) Suppose that X and Y are jointly continuous random variables. If X and Y are independent, by explicitly calculating $\mathbb{E}[Y \,|\, X = x]$, show that $\mathbb{E}[Y \,|\, X] = \mathbb{E}[Y]$. *Hint*: It might be useful to first show that when X and Y are independent, then the equality $f(y \,|\, x) = f_Y(y)$ holds where $f(y \,|\, x)$ is the conditional density and f_Y is the (marginal) density of Y.

(6) Suppose that X and Y are jointly continuous random variables and $Y = f(X)$ for some function $f : \mathbb{R} \to \mathbb{R}$. By explicitly calculating $\mathbb{E}[Y \,|\, X = x]$, show that $\mathbb{E}[Y \,|\, X] = Y$.

(7) Suppose that X and Y are jointly continuous random variables with joint density

$$f(x, y) = \begin{cases} ye^{-xy} & 0 < x < \infty, 1 < y < 2 \\ 0 & \text{otherwise} \end{cases}$$

(a) Given that $X > 1$, what is the expected value of Y? That is, calculate $\mathbb{E}[Y \mid X > 1]$. *Hint*: Use Lemma 2.2. Notice that $Y 1_{X>1} = f(Y)g(X)$ where $f(y) = y$ and $g(x)$ is the indicator function on $(1, \infty)$; that is, $g(x) = \begin{cases} 1 & x > 1 \\ 0 & \text{otherwise} \end{cases}$.

(b) Given that $Y > 1.5$, what is the expected value of X?

(c) Given that $X > Y$, what is the expected value of X? For this part, you are only required to set up the requisite integrals, but not required to evaluate them. *Hint*: Notice that $X 1_{X>Y} = g(X, Y)$ where

$$g(x, y) = \begin{cases} x & x > y \\ 0 & \text{otherwise} \end{cases}$$

(8) Suppose that Y is a normal random variable with mean $\mu = 3$ and variance $\sigma^2 = 1$; i.e., $Y \overset{\text{dist}}{=} \mathrm{N}(3, 1)$. Also suppose that X is a binomial random variable with $n = 2$ and $p = 1/4$; i.e., $X \overset{\text{dist}}{=} \mathrm{Bin}(2, 1/4)$. Suppose X and Y are independent random variables. Find both the CDF and expected value of Y^X. *Hint*: Consider using Proposition 4.1 by conditioning on the events $\{X = j\}$ for $j = 0, 1, 2$.

(9) Let X_1, X_2, Y be independent random variables. Suppose further that X_1 and X_2 are identically distributed random variables with common PMF

$$p(x) = \begin{cases} 2/5 & x = 1 \\ 3/5 & x = 2 \\ 0 & \text{otherwise} \end{cases}$$

and that Y is binomial random variable with $n = 2$ and $p = 1/3$; i.e., $Y \overset{\text{dist}}{=} \mathrm{Bin}(2, 1/3)$. Find the following conditional expectation

$$\mathbb{E}\left[\frac{X_2}{X_1 + Y} \,\middle|\, X_1, X_2 \right]$$

(10) Suppose that X and Y are jointly continuous random variables with joint density

$$f(x, y) = \begin{cases} y e^{-xy} & 0 < x < \infty, 1 < y < 2 \\ 0 & \text{otherwise} \end{cases}$$

Find $\mathbb{E}[Y \mid X]$ and $\mathbb{E}[X \mid Y]$.

(11) Let X and Y be random variables. The *conditional variance* of Y given X, denoted $\mathrm{Var}(Y \mid X)$, is defined as

$$\mathrm{Var}(Y \mid X) = \mathbb{E}[Y^2 \mid X] - \mathbb{E}[Y \mid X]^2.$$

Show that $\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}(\mathbb{E}[Y \mid X])$. (This equality you are showing is known as the Law of Total Variance). *Hint*: From the Law of Total Expectation (Corollary 6.2), you get $\mathrm{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \mathbb{E}\Big[\mathbb{E}[Y^2 \mid X]\Big] - \mathbb{E}\Big[\mathbb{E}[Y \mid X]\Big]^2$.

(12) Let $\{X_i\}_{i=1}^{\infty}$ be iid random variables and suppose that N is another random variable, independent of the X_is, and has as its state space the positive integers $\{1, 2, ...\}$. Show that the equality

$$\mathbb{E}\Big[\sum_{i=1}^{N} X_i\Big] = \mathbb{E}[N]\,\mathbb{E}[X_i]$$

holds. Note that it is unimportant which i is chosen for $\mathbb{E}[X_i]$ since the random variables $\{X_i\}_{i=1}^{\infty}$ were assumed identically distributed. (This equality you are showing is called Wald's Lemma). *Hint*: Show that $\mathbb{E}\Big[\sum_{i=1}^{N} X_i \,\Big|\, N\Big] = N\,\mathbb{E}[X_i]$ and then use the Law of Total Expectation (Corollary 6.2).

(13) Given random variables X_1, X_2, Y with $\mathbb{E}[Y \mid X_1, X_2] = 5X_1 + X_1X_2$ and $\mathbb{E}[Y^2 \mid X_1, X_2] = 25X_1^2X_2^2 + 15$, find

$$\mathbb{E}[(X_1Y + X_2)^2 \mid X_1, X_2].$$

Hint: You should find Theorem 6.6, Corollary 6.3, and Corollary 6.5 particularly useful here.

Part 2

Discrete Time Markov Chains

CHAPTER 3

Introduction to Discrete Time Markov Chains

1. Discrete Time Chains

For this chapter, we let $T = \{t_0, t_1, t_2, ...\}$ be a discrete subset of $\mathbb{R}$, indexed by the non-negative integers, and ordered such that $t_0 < t_1 < t_2 < \cdots$. We will call T the set of *time variables* and each value $t \in T$ we will call a *time*. For each time $t \in T$, there is a prescribed discrete random variable X_t, where the collection of all these variables will be denoted $(X_t)_{t\in T}$, often referred to as a *process*. The set $S = \bigcup_{t\in T} S_t$, where S_t is the state space of X_t, will be called the *state space* of the process $(X_t)_{t\in T}$. It is assumed that each element in S is isolated from the others. To further refine how we refer to the process $(X_t)_{t\in T}$, the assumptions made on T and S allow us to call $(X_t)_{t\in T}$ a *discrete time chain*, where "discrete time" refers to the the fact that the set of time variables T is a discrete, and "chain" refers to the fact that each point in S is an isolated point.

Example 1.1. Consider two teams A and B playing a point-earning game where points are earned when a team scores a goal. Let t_0 be the time of the start of the game; for $n \geq 1$, let t_n be the time of the nth goal scored (by either team); and let $T = \{t_0, t_1, t_2, ...\}$. For each $t \in T$, let X_t be the difference of team A's score to team B's score at time t; i.e., $X_t = A_t - B_t$ where A_t is team A's score at time t and B_t is team B's score at time t. Assuming that each time a team scores, they earn a single point and that there is no limit to the number of points a particular team can make, the state space S of the process $(X_t)_{t\in T}$ can reasonably be assumed as the set of integers $S = \mathbb{Z}$. Alternatively, we could consider the processes $(A_t)_{t\in T}$ and $(B_t)_{t\in T}$ of the scores of team A and B, respectively, in which case their state spaces are all non-negative integers $\{0, 1, 2, ...\}$, assuming that a team can only score points, not loose them. Each of these processes $(X_t)_{t\in T}$, $(A_t)_{t\in T}$, and $(B_t)_{t\in T}$ are an example of a discrete time chain. △

Convention 1.1. It is very common to assume that the set of time variables T of a discrete time chain $(X_t)_{t\in T}$ is the set of non-negative integers $T = \{0, 1, 2, ...\}$. There is nothing lost in this assumption since it can be interpreted as simply exchanging (more cumbersome) notation X_{t_n} for X_n. When it is the case that $T = \{0, 1, 2, ...\}$, we will usually write $(X_t)_{t=0}^{\infty}$ rather than $(X_t)_{t\in T}$. Along the same lines, when we consider T a set of consecutive integers starting at some value k,

$T = \{k, k+1, k+2, ...\}$, we will write $(X_t)_{t=k}^{\infty}$. In any case, since we lose no generality moving between notations, we can use whichever notation best matches our intuition of the scenario we are modeling with a discrete time chain. △

2. The Markov Property for Discrete Time Chains

Let $(X_t)_{t\in T}$ be a discrete time chain with state space S and times $T = \{t_0, t_1, t_2, ...\}$.

DEFINITION 2.1. The process $(X_t)_{t\in T}$ is said to satisfy the *Markov property* when either (and hence both) of the equivalent properties hold

$$\mathbb{E}[X_{s_{n+1}} \,|\, X_{s_0}, X_{s_1}, ..., X_{s_n}] = \mathbb{E}[X_{s_{n+1}} \,|\, X_{s_n}] \tag{18}$$

or

$$\mathbb{P}(X_{s_{n+1}} = j \,|\, X_{s_0} = j_0, X_{s_1} = j_1, ..., X_{s_n} = j_n) = \mathbb{P}(X_{s_{n+1}} = j \,|\, X_{s_n} = j_n) \tag{19}$$

for all collections of times $s_0, s_1, ..., s_n, s_{n+1} \in T$ with $s_0 < s_1 < \cdots < s_n$, and states $j_0, ..., j_n, j \in S$, as long as both sides of the equality in (19) are defined. Note that we are using the variable s_k rather than t_k because we are not necessarily assuming that $t_0 = s_0$ and $t_1 = s_1$, etc.; it could be that $s_0 = t_5$, $s_1 = t_{27}$, and so on, for example. If the discrete time chain $(X_t)_{t\in T}$ satisfies the Markov property, then we say that it is a *discrete time Markov chain.* △

REMARK 2.1. To find $\mathbb{E}[X_{s_{n+1}} \,|\, X_{s_0}, X_{s_1}, ..., X_{s_n}]$, you consider $\mathbb{E}[X_{s_{n+1}} \,|\, X_{s_0} = j_0, X_{s_1} = j_1, ..., X_{s_n} = j_n]$ for all states $j_0, j_1, ..., j_n \in S$; in turn, this expected value is defined by the probability $\mathbb{P}(X_{s_{n+1}} = j \,|\, X_{s_0} = j_0, X_{s_1} = j_1, ..., X_{s_n} = j_n)$ for all states $j \in S$. Similarly, to find $\mathbb{E}[X_{t_{n+1}} \,|\, X_{t_n}]$, you consider $\mathbb{E}[X_{s_{n+1}} \,|\, X_{s_n} = j_n]$ for all states $j_n \in S$; in turn, this expected value is defined by the probability $\mathbb{P}(X_{s_{n+1}} = j \,|\, X_{s_n} = j_n)$ for all states $j \in S$. Considering these observations, it quickly follows that (18) and (19) are, indeed, equivalent. △

REMARK 2.2. To understand (18) in our interpretation of conditional expectations, the left hand side $\mathbb{E}[X_{s_{n+1}} \,|\, X_{s_0}, X_{s_1}, ..., X_{s_n}]$ is a random variable of the form $g(X_{s_0}, ..., X_{s_{n+1}})$, whereas the right hand side $\mathbb{E}[X_{s_{n=1}} \,|\, X_{s_n}]$ is a random variable of the form $h(X_{s_n})$. So, the equality between these two when $(X_t)_{t\in T}$ satisfies the Markov property means that these random variables are equal $g(X_{s_0}, ..., X_{s_n}) = h(X_{s_n})$; simply put, $g(X_{s_0}, ..., X_{s_n})$ only truly depends on the variable X_{s_n}. △

It is useful here to discuss the colloquial mantra commonly used to described the Markov property. It goes as follows: the process $(X_t)_{t\in T}$ satisfies the Markov property when *given the present, the future and the past are independent.* To interpret this in the symbology of (18) and

(19), the times in consideration $s_0 < s_1 < \cdots < s_n < s_{n+1}$ are inutitively categorized into three bins: the past times $s_0, ..., s_{n-1}$; the present time s_n; and the future time s_{n+1}. Hence, if the process satisfies the Markov property, given that you know the past and present $X_{s_0}, X_{s_1}, ..., X_{s_n}$ of the process, any probabilistic questions you have about the future of the process $X_{s_{n+1}}$ only depends on the present information X_{s_n}, and forgets about the past variables, $X_{s_0}, ..., X_{s_{n-1}}$.

EXAMPLE 2.1. Suppose that $\{Y_i\}_{i=1}^{\infty}$ are independent Bernoulli random variables and define $X_n = \sum_{i=1}^{n} Y_i$ for $n = 1, 2, 3, ...$ The process $(X_n)_{n=1}^{\infty}$ is a discrete time Markov chain. Indeed, it is clear that $(X_n)_{n=1}^{\infty}$ is a discrete time chain. To see that it satisfies the Markov property, consider

$$\mathbb{P}(X_{n+1} = j \,|\, X_1 = j_1, X_2 = j_2, \cdots, X_n = j_n)$$

working under the assumption that the event we are conditioning on has positive probability. Since $X_n = j_n = \sum_{i=1}^{n} Y_i$ and $X_{n+1} = \sum_{i=1}^{n+1} Y_i = Y_{n+1} + \sum_{i=1}^{n} Y_i = Y_{n+1} + X_n$ we have that

$$\begin{aligned}\mathbb{P}(X_{n+1} = j \,|\, X_1 = j_1, X_2 = j_2, \cdots, X_n = j_n) &= \mathbb{P}(Y_{n+1} + X_n = j \,|\, X_1 = j_1, X_2 = j_2, \cdots, X_n = j_n) \\ &= \mathbb{P}(Y_{n+1} = j - j_n \,|\, X_1 = j_1, X_2 = j_2, \cdots, X_n = j_n) = \mathbb{P}(Y_{n+1} = j - j_n)\end{aligned}$$

where the last equality followed by the assumption that Y_{n+1} is independent of $Y_1, ..., Y_n$ and hence of $X_1, ..., X_n$. Similarly,

$$\begin{aligned}\mathbb{P}(X_{n+1} = j \,|\, X_n = j_n) = \mathbb{P}(Y_{n+1} + X_n = j \,|\, X_n = j_n) = \mathbb{P}(Y_{n+1} = j - j_n \,|\, X_n = j_n) \\ = \mathbb{P}(Y_{n+1} = j - j_n).\end{aligned}$$

This shows that

$$\mathbb{P}(X_{n+1} = j \,|\, X_1 = j_1, X_2 = j_2, \cdots, X_n = j_n) = \mathbb{P}(X_{n+1} = j \,|\, X_n = j_n)$$

from which we conclude that $(X_n)_{n=1}^{\infty}$ satisfies the Markov property. △

You may rightfully object to the method we used to show that the Markov property held for $(X_t)_{t\in T}$ in the previous example. The objection raised is that instead of arguing the case for any collection of ordered times $s_0, s_1, ..., s_n, s_{n+1} \in T$ as in the definition of the Markov property (see (18) and (19)), we only argued for the consecutive times $t_0, t_1, ..., t_n, t_{n+1}$. However, as we will see in the next theorem, using consecutive times is still a valid way to show the Markov property holds.

THEOREM 2.1. *Let $(X_t)_{t\in T}$ be a discrete time chain where, as previously noted, $T = \{t_0, t_1, t_2, ...\}$. The process $(X_t)_{t\in T}$ satisfies the Markov process if and only if either (and hence both) of the two*

equivalent conditions are satisfied

$$\mathbb{E}[X_{t_{n+1}} \,|\, X_{t_0}, X_{t_1}, ..., X_{t_n}] = \mathbb{E}[X_{t_{n+1}} \,|\, X_{t_n}] \tag{20}$$

or

$$\mathbb{P}(X_{t_{n+1}} = j \,|\, X_{t_0} = j_0, X_{t_1} = j_1, ..., X_{t_n} = j_n) = \mathbb{P}(X_{t_{n+1}} = j \,|\, X_{t_n} = j_n) \tag{21}$$

for all $n \in \mathbb{N}$ and states $j_0, j_1, ..., j_{n+1} \in S$ as long as both side of (21) is defined.

PROOF. We will argue that (19) and (21) are equivalent. Certainly if (19) holds, then (21) holds since we could have chosen the times s_i in (19) to be the t_i. So, the important direction is showing that if (21) holds, then so does (19). To prove the general case of this direction is a bit notationally cumbersome, so we will instead argue by an example that emphasizes the technique one would use to prove the general case. Suppose that (21) holds. Let us show that with this assumption the equality $\mathbb{P}(X_{t_4} = j \,|\, X_{t_0} = j_0, X_{t_2} = j_2) = \mathbb{P}(X_{t_4} = j \,|\, X_{t_2} = j_2)$ follows (this is the special case with $s_0 = t_0, s_1 = t_2$, and $s_2 = t_4$). By using Corollary 1.5 and assuming (21) holds, we find

$$\begin{aligned}
&\mathbb{P}(X_{t_4} = j \,|\, X_{t_0} = j_0, X_{t_2} = j_2) \\
&= \sum_{j_1, j_3 \in S} \mathbb{P}(X_{t_4} = j \,|\, X_{t_0} = j_0, X_{t_1} = j_1, X_{t_2} = j_2, X_{t_3} = j_3)\, \mathbb{P}(X_{t_1} = j_1, X_{t_3} = j_3 \,|\, X_{t_0} = j_0, X_{t_2} = j_2) \\
&= \sum_{j_1, j_3 \in S} \mathbb{P}(X_{t_4} = j \,|\, X_{t_3} = j_3)\, \mathbb{P}(X_{t_3} = j_3 \,|\, X_{t_0} = j_0, X_{t_1} = j_1, X_{t_2} = j_2)\, \mathbb{P}(X_{t_1} = j_1 \,|\, X_{t_0} = j_0, X_{t_2} = j_2) \\
&= \sum_{j_1, j_3 \in S} \mathbb{P}(X_{t_4} = j \,|\, X_{t_3} = j_3)\, \mathbb{P}(X_{t_3} = j_3 \,|\, X_{t_2} = j_2)\, \mathbb{P}(X_{t_1} = j_1 \,|\, X_{t_0} = j_0, X_{t_2} = j_2) \\
&= \sum_{j_3 \in S} \mathbb{P}(X_{t_4} = j \,|\, X_{t_3} = j_3)\, \mathbb{P}(X_{t_3} = j_3 \,|\, X_{t_2} = j_2) \underbrace{\sum_{j_1 \in S} \mathbb{P}(X_{t_1} = j_1 \,|\, X_{t_0} = j_0, X_{t_2} = j_2)}_{1} \\
&= \sum_{j_3 \in S} \mathbb{P}(X_{t_4} = j \,|\, X_{t_3} = j_3)\, \mathbb{P}(X_{t_3} = j_3 \,|\, X_{t_2} = j_2)
\end{aligned}$$

To finish our argument, we need to show that

$$\sum_{j_3 \in S} \mathbb{P}(X_{t_4} = j \,|\, X_{t_3} = j_3)\, \mathbb{P}(X_{t_3} = j_3 \,|\, X_{t_2} = j_2) = \mathbb{P}(X_{t_4} = j \,|\, X_{t_2} = j_2)$$

which will follow by Corollary 1.5 once we show that

$$\mathbb{P}(X_{t_4} = j \,|\, X_{t_3} = j_3) = \mathbb{P}(X_{t_4} = j \,|\, X_{t_2} = j_2, X_{t_3} = j_3).$$

Therefore, to show this last equality, arguing similarly as above,

$$\begin{aligned}
&\mathbb{P}(X_{t_4} = j \,|\, X_{t_2} = j_2, X_{t_3} = j_3) \\
&= \sum_{j_0, j_1 \in S} \mathbb{P}(X_{t_4} = j \,|\, X_{t_0} = j_0, X_{t_1} = j_1, X_{t_2} = j_2, X_{t_3} = j_3)\, \mathbb{P}(X_{t_0} = j_0, X_{t_1} = j_1 \,|\, X_{t_2} = j_2, X_{t_3} = j_3) \\
&= \sum_{j_0, j_1} \mathbb{P}(X_{t_4} = j_4 \,|\, X_{t_3} = j_3)\, \mathbb{P}(X_{t_0} = j_0, X_{t_1} = j_1 \,|\, X_{t_2} = j_2, X_{t_3} = j_3) \\
&= \mathbb{P}(X_{t_4} = j_4 \,|\, X_{t_3} = j_3) \sum_{j_0, j_1} \mathbb{P}(X_{t_0} = j_0, X_{t_1} = j_1 \,|\, X_{t_2} = j_2, X_{t_3} = j_3) \\
&= \mathbb{P}(X_{t_4} = j_4 \,|\, X_{t_3} = j_3).
\end{aligned}$$

Finally, putting all these pieces together,

$$\begin{aligned}
\mathbb{P}(X_{t_4} = j \,|\, X_{t_0} = j_0, X_{t_2} = j_2) &= \sum_{j_3 \in S} \mathbb{P}(X_{t_4} = j \,|\, X_{t_3} = j_3)\, \mathbb{P}(X_{t_3} = j_3 \,|\, X_{t_2} = j_2) \\
&= \sum_{j_3 \in S} \mathbb{P}(X_{t_4} = j \,|\, X_{t_3} = j_3, X_{t_2} = j_2)\, \mathbb{P}(X_{t_3} = j_3 \,|\, X_{t_2} = j_2) \\
&= \mathbb{P}(X_{t_4} = j_4 \,|\, X_{t_2} = j_2).
\end{aligned}$$

This shows the argument for this special case. However, the general argument works similarly by summing over the missing variables. □

COROLLARY 2.2. *Suppose that $(X_t)_{t \in T}$ is a discrete time Markov chain. Let $r, s, t \in T$ with $r < s < t$. Then for any states $i, j \in S$,*

$$\mathbb{P}(X_t = j \,|\, X_r = i) = \sum_{k \in S} \mathbb{P}(X_t = j \,|\, X_s = k)\, \mathbb{P}(X_s = k \,|\, X_r = i).$$

PROOF. As in the proof of Theorem 2.1, this follows from Corollary 1.5 since by the Markov property, $\mathbb{P}(X_t = j \,|\, X_s = k) = \mathbb{P}(X_t = j \,|\, X_r = i, X_s = j)$. □

3. The Transition Matrix for Stationary Discrete Time Markov Chains

We will be studying transition probabilities related to discrete time Markov chains in this section.

NOTATION 3.1. For convenience, given a time $t \in T$, we let t_+ denote the consecutive time in T. For example, if $T = \{t_0, t_1, t_2, ...\}$ and if $t = t_n$, then $t_+ = t_{n+1}$. △

DEFINITION 3.1. Suppose that $(X_t)_{t\in T}$ is a discrete time Markov chain with state space S. For states $i, j \in S$, define the *one step transition probabilities* as

$$p_{ij}(t) = \mathbb{P}(X_{t_+} = j \,|\, X_t = i)$$

for each $t \in T$ such that the right hand conditional probability is defined. △

The one step transition probability $p_{ij}(t)$ represents the probability of the process moving from state i to state j in one time step starting at time t. While this one step transition probabilities very well may depend on the time t at which you are finding the probability (clearly indicated in the notation), we will focus heavily on those processes for which $p_{ij}(t)$ is independent of t and hence write p_{ij} with no concern of what time t the step is taking place.

DEFINITION 3.2. A discrete time Markov chain $(X_t)_{t\in T}$ will be called *stationary* or *time homogenous* when the one step transition probabilities $p_{ij}(t)$ have no dependence on t; i.e., $p_{ij}(t)$ is constant for all times t for which it is defined. For this reason, we can ignore t in the notation and will simply write p_{ij} as the probability of the process moving from state i to state j in one time step. △

DEFINITION 3.3. Given a stationary discrete time Markov chain $(X_t)_{t=0}^{\infty}$, the *transition matrix* related to this process is defined as the matrix $\mathbf{P}$ indexed by the state space S where the (i, j)th element of $\mathbf{P}$ is the (i, j)th one-step transition probability p_{ij}. Note that convention dictates that the (i, j)th entry of $\mathbf{P}$ means the element located at the intersection of the ith row and jth column. △

EXAMPLE 3.1. Suppose that we have 3 marbles, labelled A, B, C, distributed between two urns, urn I and urn II. At each step, you randomly select a letter A, B, or C, at which point the corresponding marble is transplanted from the urn it is currently in into the other urn. Let X_n count the number of marbles in urn I at step n, where X_0 is how many of the marbles were in urn I before we started playing this game. Then $(X_n)_{n=0}^{\infty}$ is a stationary Markov chain with state space $S = \{0, 1, 2, 3\}$, and the one-step transition probabilities are given by,

$$\mathbb{P}(X_{n+1} = j \,|\, X_n = i) = \begin{cases} \frac{i}{3} & j = i - 1 \\ \frac{3-i}{3} & j = i + 1 \\ 0 & \text{otherwise} \end{cases}$$

To see why this is true, if i marbles are in urn I, then $3 - i$ marbles are in urn II. Since you are equally likely to choose any one marble, you then have a probability of $i/3$ for choosing a marble in urn I, in which case on the next step you have moved that marble to urn II and have one fewer

marbles in urn I; similarly, you have a $(3-i)/3$ probability of choosing a marble in urn II, in which case on the next step you have moved that marble to urn I and will have one greater marble in urn I. From these one-step transition probabilities, we easily construct the transition matrix,

$$\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} & p_{02} & p_{03} \\ p_{10} & p_{11} & p_{12} & p_{13} \\ p_{20} & p_{21} & p_{22} & p_{23} \\ p_{30} & p_{31} & p_{32} & p_{33} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 2/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

△

REMARK 3.1. The transition matrix is always $N \times N$ where N is the number of elements in S (possibly countably infinite). The row of the transition matrix represents the state the process is moving from, while the column represents the state the process is moving to. Therefore, it is important to remember that once you've chosen an order for the state space, we must keep it consistent in order to make sense of our future calculations.

To solidify the difference in transition matrices based on ordering of the state space, let's reuse the Markov chain from Example 3.1. We have 4! possible choices for ordering the four elements in $S = \{0, 1, 2, 3\}$, and therefore have 4! options for the transition matrix $\mathbf{P}$. Here are two of these choices:

S	0	1	2	3
0	0	1	0	0
1	1/3	0	2/3	0
2	0	2/3	0	1/3
3	0	0	1	0

S	3	0	1	2
3	0	0	0	1
0	0	0	1	0
1	0	1/3	0	2/3
2	1/3	0	2/3	0

The first and third quadrant of these tables are the ordering of the state space S, and the fourth quadrant consists of the entries in the corresponding transition matrix P. Notice that the first table agrees with the transition matrix in the example with $S = \{0, 1, 2, 3\}$, whereas the second table is adjusted for the ordering $S = \{3, 0, 1, 2\}$. △

NOTATION 3.2. For these notes, unless explicitly mentioned otherwise, we choose the order by increasing size of the states. If we run into an example where the state space is not clearly linearly ordered, we will make it clear which order we are using for the state space S. △

EXAMPLE 3.2. A coin is continually flipped. You keep a running tally of the number of heads flipped, letting X_n be the number of heads you've counted on the nth flip. Let us analyze the process $(X_n)_{n=1}^{\infty}$. The state space S of this process is all non-negative integers, $S = \{0, 1, 2, 3, ...\}$, since the number of heads you count at each flip can range over these states as n increases. Let p be the probability that the coin flips heads and $q = 1 - p$ be the probability of flipping tails; we will reasonably assume that p and q do not change in time. If at the nth flip we have counted k heads, then it is clear that on the $(n+1)$st flip we will have counted $k+1$ heads with probability p, the probability that the $(n+1)$st flip is heads; whereas our count will stay at k with probability q, the probability that the $(n+1)$st flip is tails. In symbols, we have that $\mathbb{P}(X_{n+1} = k+1 \,|\, X_n = k) = p$ and $\mathbb{P}(X_{n+1} = k \,|\, X_n = k) = q$. Further, note that if we are given $X_n = k$, then these one step probabilities for X_{n+1} don't depend on the how the process arrived at $X_n = k$; this means that this process satisfies the Markov property. What we can now conclude is that $(X_n)_{n=1}^{\infty}$ is a discrete time Markov chain: discrete time since the time variables are discrete ($n = 1, 2, 3, ...$); a chain because the state space $S = \{0, 1, 2, ...\}$ is comprised of isolated points; and, as we previously noted, it is a Markov process. In fact, we already observed that $(X_n)_{n=1}^{\infty}$ was a discrete time Markov chain in Example 2.1 above had we let Y_i be the outcome of the ith flip (1 if heads, 0 if tails). Moreover, this discrete time Markov chain is stationary since the one step probabilities don't depend on the time variable n. The transition matrix $\mathbf{P}$ for this process is $N \times N$ where $N = \infty$ in this case since we are indexing it by $S = \{0, 1, 2, 3, ...\}$. We get

$$\mathbf{P} = \begin{pmatrix} q & p & 0 & 0 & \cdots \\ 0 & q & p & 0 & \cdots \\ 0 & 0 & q & p & \ddots \\ 0 & 0 & 0 & q & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}$$

which is easily derived from the discussion above. Again, emphasizing that the order of the state space is important, note that we chose the transition matrix consistent with our ordering convention

S	0	1	2	3	$\cdots$
0	q	p	0	0	$\cdots$
1	0	q	p	0	$\cdots$
2	0	0	q	p	$\ddots$
3	0	0	0	q	$\ddots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\ddots$	$\ddots$

△

It turns out that the study of stationary discrete time Markov chains is inextricable from the study of their corresponding transition matrices. Because of this, the study of matrices having the same form as transition matrices is paramount, so we give such matrices a label.

DEFINITION 3.4. A square matrix $\mathbf{P}$ is called a *stochastic matrix* when every entry is non-negative and the sum of the entries along each row equals 1. △

LEMMA 3.1. *The transition matrix* $\mathbf{P}$ *of a stationary discrete time Markov chain is a stochastic matrix.*

PROOF. Let $(X_t)_{t\in T}$ be a stationary discrete time Markov chain with transition matrix $\mathbf{P}$ with entries $p_{ij} = \mathbb{P}(X_{t+} = j \,|\, X_t = i)$. Since the entries of $\mathbf{P}$ are probability values, they must be non-negative. Moreover, for $i \in S$, the sum along the ith row of $\mathbf{P}$ is

$$\sum_{j\in S} p_{ij} = \sum_{j\in S} \mathbb{P}(X_{t+} = j \,|\, X_t = i) = 1$$

showing that $\mathbf{P}$ is a stochastic matrix. □

4. Multistep Transition Probabilities

We previously defined the one-step transition probabilities as $p_{ij} = \mathbb{P}(X_{t+} = j \,|\, X_t = i)$. Here we will define multi-step transition probabilities, and using the transition matrix, how to calculate them.

DEFINITION 4.1. Recall that $T = \{t_0, t_1, t_2, ...\}$. Let $(X_t)_{t\in T}$ be a stationary discrete time Markov chain with state space S. For each $m \in \mathbb{N}$, we define the *m-step transition probability* from state i to j by

$$p_{ij}^{(m)} = \mathbb{P}(X_{t_{n+m}} = j \,|\, X_{t_n} = i) \tag{22}$$

where nth time step t_n is arbitrary, as long as the right hand probability is defined. Verbosely, $p_{ij}^{(m)}$ is the probability that the process moves from state i to state j in m time steps. △

THEOREM 4.1. *Let $(X_t)_{t\in T}$ be a stationary discrete time Markov chain with transition matrix $\mathbf{P}$. For any $m \in \mathbb{N}$, the equality*

$$p_{ij}^{(m)} = [\mathbf{P}^m]_{ij}$$

holds for the m-step transition probabilities. That is, $p_{ij}^{(m)}$ is the (i,j)th entry in the mth power of the transition matrix $\mathbf{P}^m$.

PROOF. For the case $m = 1$, there is nothing to show, since it follows by definition. Let's work the case $m = 2$. Here,

$$\begin{aligned} p_{ij}^{(2)} &= \mathbb{P}(X_{t_{n+2}} = j \,|\, X_{t_n} = i) = \sum_{k\in S} \mathbb{P}(X_{t_{n+2}} = j \,|\, X_{t_n} = i, X_{t_{n+1}} = k)\, \mathbb{P}(X_{t_{n+1}} = k \,|\, X_{t_n} = i) \\ &= \sum_{k\in S} \underbrace{\mathbb{P}(X_{t_{n+2}} = j \,|\, X_{t_{n+1}} = k)}_{p_{kj}} \underbrace{\mathbb{P}(X_{t_{n+1}} = k \,|\, X_{t_n} = i)}_{p_{ik}} = \sum_{k\in S} p_{ik}p_{kj} = [\mathbf{P}\,\mathbf{P}]_{ij} = [\mathbf{P}^2]_{ij} \end{aligned}$$

where the last equality follows from matrix multiplication and the definition of $\mathbf{P}$. Next, let's work $m = 3$. Here,

$$\begin{aligned} p_{ij}^{(3)} &= \mathbb{P}(X_{t_{n+3}} = j \,|\, X_{t_n} = i) = \sum_{k\in S} \mathbb{P}(X_{t_{n+3}} = j \,|\, X_{t_n} = i, X_{t_{n+1}} = k)\, \mathbb{P}(X_{t_{n+1}} = k \,|\, X_{t_n} = i) \\ &= \sum_{k\in S} \underbrace{\mathbb{P}(X_{t_{n+3}} = j \,|\, X_{t_{n+1}} = k)}_{p_{kj}^{(2)}} \underbrace{\mathbb{P}(X_{t_{n+1}} = k \,|\, X_{t_n} = i)}_{p_{ik}} = \sum_{k\in S} p_{ik}^{(2)}p_{kj} = [\mathbf{P}^2\,\mathbf{P}]_{ij} = [\mathbf{P}^3]_{ij}. \end{aligned}$$

From here, the result for general m follows similarly by building from the previous case and using induction. □

COROLLARY 4.2. *If $(X_t)_{t\in T}$ is a stationary discrete time Markov chain, then the stationary property of this process guarantees that the m-step transition probabilities $p_{ij}^{(m)}$ do not depend on the time step t_n in (22), as long as the probability is defined.*

PROOF. In Theorem 4.1, we were able to show that $p_{ij}^{(m)} = [\mathbf{P}^m]_{ij}$ without concern for the time step t_n. Hence the value of $p_{ij}^{(m)}$ did not depend on this time step as long as we were able to define the conditional probability used in the proof of the theorem. □

EXAMPLE 4.1. Let $(X_t)_{t\in T}$ be a stationary discrete time Markov chain with state space $S = \{0,1,2\}$ and transition matrix

$$\mathbf{P} = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/3 & 1/4 & 5/12 \\ 0 & 1/3 & 2/3 \end{pmatrix}$$

We will calculate $\mathbb{P}(X_{t_4} = 0, X_{t_2} = 1 \,|\, X_{t_1} = 2)$. We have

$$\begin{aligned}\mathbb{P}(X_{t_4} = 0, X_{t_2} = 1 \,|\, X_{t_1} = 2) &= \mathbb{P}(X_{t_4} = 0 \,|\, X_{t_1} = 2, X_{t_2} = 1)\, \mathbb{P}(X_{t_2} = 1 \,|\, X_{t_1} = 2) \\ &= \mathbb{P}(X_{t_4} = 0 \,|\, X_{t_2} = 1)\, \mathbb{P}(X_{t_2} = 1 \,|\, X_{t_1} = 2) = p_{10}^{(2)}\, p_{21}.\end{aligned}$$

We already are given that $p_{21} = [\mathbf{P}]_{21} = 1/3$. We now have a couple methods at our disposal to find $p_{10}^{(2)}$. One method is using Theorem 4.1, whence we find

$$\mathbf{P}^2 = \begin{pmatrix} 5/12 & 3/8 & 5/24 \\ 1/4 & 53/144 & 55/144 \\ 1/9 & 11/36 & 7/12 \end{pmatrix} \implies p_{10}^{(2)} = [\mathbf{P}^2]_{10} = 1/4.$$

Alternatively, we could attack this directly to find

$$\begin{aligned}p_{10}^{(2)} = \mathbb{P}(X_{t_4} = 0 \,|\, X_{t_2} = 1) &= \sum_{k=0}^{2} \mathbb{P}(X_{t_4} = 0 \,|\, X_{t_2} = 1, X_{t_3} = k)\, \mathbb{P}(X_{t_3} = k \,|\, X_{t_2} = 1) \\ &= \sum_{k=0}^{2} \mathbb{P}(X_{t_4} = 0 \,|\, X_{t_3} = k)\, \mathbb{P}(X_{t_3} = k \,|\, X_{t_2} = 1) = \sum_{k=0}^{3} p_{1k}\, p_{k0} \\ &= (1/3)(1/2) + (1/4)(1/3) + (5/12)(0) = 1/4.\end{aligned}$$

Using whichever method, we find $\mathbb{P}(X_{t_4} = 0, X_{t_2} = 1 \,|\, X_{t_1} = 2) = p_{10}^{(2)}\, p_{21} = (1/4)(1/3) = 1/12$. $\triangle$

COROLLARY 4.3 (Chapman-Kolmogorov Equation). *Let $(X_t)_{t\in T}$ be a stationary discrete time Markov chain with state space S. Then for any states $i, j \in S$ and positive integers m, n, the equality*

$$p_{ij}^{(m+n)} = \sum_{k\in S} p_{ik}^{(m)} p_{kj}^{(n)}$$

of the multistep transition probabilities holds.

PROOF. The quickest proof is simply noting that this is matrix multiplication:

$$p_{ij}^{(m+n)} = [\mathbf{P}^{m+n}]_{ij} = [\mathbf{P}^m \mathbf{P}^n]_{ij} = \sum_{k\in S} [\mathbf{P}^m]_{ik} [\mathbf{P}^n]_{kj} = \sum_{k\in S} p_{ik}^{(m)} p_{kj}^{(n)}$$

where, as usual, $\mathbf{P}$ is the transition matrix of the process. However, let us give an intuitive argument as well. The value $p_{ij}^{(m+n)}$ represents the probability the process moves from state i to state j in $m+n$ time steps. For each $k \in S$, the value $p_{ik}^{(m)} p_{kj}^{(n)}$ represents the probability the process moves from state i to state k in m time steps, and then from state k to state j in n time steps. Thus, by summing $p_{ik}^{(m)} p_{kj}^{(n)}$ over all possible states k, we are taking into account all possible paths the process made during the first m and last n time steps, arriving at $p_{ij}^{(m+n)}$. □

5. The Probability Space of a Stationary Discrete Time Markov Chain

Throughout the material we have presented regarding processes $(X_t)_{t\in T}$, we have implicitly understood that there exists a sample space Ω on which X_t is defined for every $t \in T$. Of course, if there were not one sample space Ω on which every X_t is defined, then there would not be a probability $\mathbb{P}$ defined such that expressions such those in (18) and (19) would make sense – remember that a probability $\mathbb{P}$ needs a sample space Ω to be defined on. Looking back, a process $(X_t)_{t\in T}$ can be determined as a discrete time chain without regards to the probability $\mathbb{P}$; however, the determination of whether or not $(X_t)_{t\in T}$ satisfies the Markov property, or whether or not it is stationary depends heavily on the chosen probability $\mathbb{P}$. It is possible that there are several probabilities defined on Ω, some of which $(X_t)_{t\in T}$ satisfies stationarity or the Markov property with respect to, some of which it does not. When Ω is the sample space on which X_t is defined for each $t \in T$, we will often say that Ω is the sample space of the process, or that $(X_t)_{t\in T}$ is *a process on* Ω.

THEOREM 5.1. *For any stochastic matrix* $\mathbf{P}$ *indexed by the set* S*, there exists a sample space* Ω*; there exists a probability* $\mathbb{P}$ *defined on* Ω*; and there exists a stochastic process* $(X_t)_{t\in T}$ *whose sample space is* Ω*, whose state space is* S*, and such that the following are satisfied:* $\mathbb{P}(X_{t_0} = j) > 0$ *for every* $j \in S$*, and with respect to* $\mathbb{P}$*,* $(X_t)_{t\in T}$ *is a stationary discrete time Markov chain with transition matrix* $\mathbf{P}$*.*

PROOF. The proof of this theorem (and the more general version using methods developed largely by the famous mathematician Kolmogorov) is beyond the scope of our background within these notes. □

There are several lessons to glean from Theorem 5.1. One is that given a stochastic matrix $\mathbf{P}$, we can always assume that there is some stationary discrete time Markov chain in the background for which $\mathbf{P}$ is the transition matrix. Another which will be useful in the following section is that given a stationary discrete time Markov chain $(X_t)_{t\in T}$ whose transition matrix is $\mathbf{P}$, we can assume

that the underlying probability being used $\mathbb{P}$ is such that $\mathbb{P}(X_{t_0} = j) > 0$ for every j in the state space S of the process – in a certain sense, we can define most other corresponding probabilities, those for which $(X_t)_{t\in T}$ is a stationary discrete time Markov chain with the same transition matrix $\mathbf{P}$, in terms of $\mathbb{P}$.

CONVENTION 5.1. Henceforth, if not explicitly mentioned otherwise, when given a stationary discrete time Markov chain $(X_t)_{t\in T}$ (explicitly or implicitly via a transition matrix $\mathbf{P}$), we assume that $(X_t)_{t\in T}$ was developed as in Theorem 5.1 with the probability $\mathbb{P}$ described therein. △

6. Initial Distributions

Throughout this section we let $(X_t)_{t\in T}$ be a stationary discrete time Markov chain with state space S. Further, we let $\mathbb{P}$ be the probability described in Theorem 5.1, implicitly assumed from Convention 5.1, and let $\mathbf{P}$ be the transition matrix of $(X_t)_{t\in T}$ with respect to $\mathbb{P}$.

DEFINITION 6.1. A function $\nu : S \to \mathbb{R}$ such that $\nu(j) \geq 0$ and $\sum_{j\in S} \nu(j) = 1$ will be called a *probability mass function* on S, or a *probability distribution* on S. △

There is now an ambiguity of definitions for a probability mass function, since it could refer to the probability mass function of a discrete random variable, or to the probability mass function just defined without reference to a corresponding random variable. However, there is a connection between these two, and the context which we are considering will keep confusion minimal.

NOTATION 6.1. The fundamental probability mass functions we will consider are those which put the entire *mass* onto one state $j \in S$. Explicitly, adopting *delta notation*: for each $j \in S$, we let $\delta_j : S \to \mathbb{R}$ be the probability mass function defined such that $\delta_j(j) = 1$ and $\delta_j(i) = 0$ for $i \neq j$. We will call δ_j a *delta mass function* on the state j. △

The reason that the delta mass functions are fundamental is because all other mass functions can be written as a linear combination of them.

LEMMA 6.1. *Let $\nu : S \to \mathbb{R}$ be a probability mass function. Then*

$$\nu = \sum_{j\in S} \nu(j)\, \delta_j$$

That is, for any $i \in S$, $\nu(i) = \sum_{j\in S} \nu(j)\, \delta_j(i)$.

PROOF. This is immediate since $\delta_j(i) = 1$ if and only if $i = j$, otherwise it is equal to 0, so

$$\nu(i) = \nu(i)\,\delta_i(i) = \underbrace{\sum_{j \neq i} \nu(j)\,\delta_j(i)}_{0} + \nu(i)\,\delta_i(i) = \sum_{j \in S} \nu(j)\,\delta_j(i).$$

□

DEFINITION 6.2. If $\nu : S \to \mathbb{R}$ is a probability mass function, we say that ν is the *initial distribution* of the process when ν is the probability mass function for the initial random variable X_{t_0} of the process. △

We now state a more general version of Theorem 5.1 with an initial distribution in mind.

THEOREM 6.2. *For any stochastic matrix* $\mathbf{P}$ *indexed by the set* S *and probability mass function* $\nu : S \to \mathbb{R}$*, there exists a sample space* Ω*; there exists a probability* $\mathbb{P}_\nu$ *defined on* Ω*; and there exits a stochastic process* $(X_t)_{t \in T}$ *whose sample space is* Ω*, whose state space is* S*, and such that the following are satisfied: With respect to* $\mathbb{P}_\nu$*,* $(X_t)_{t \in T}$ *is a stationary discrete time Markov chain with initial distribution* ν*, and the transition matrix of* $(X_t)_{t \in T}$ *is* $\mathbf{P}$*.*

PROOF. As in the case of Theorem 5.1, we omit the proof of this theorem. □

NOTATION 6.2. If we are considering a delta mass function $\delta_j : S \to \mathbb{R}$, we will write $\mathbb{P}_j$ instead of writing $\mathbb{P}_{\delta_j}$ for the probability described in Theorem 6.2. Notice that $\mathbb{P}_j$ is the probability to consider when the corresponding stationary discrete time Markov chain $(X_t)_{t \in T}$ in Theorem 6.2 is considered to start in state j; i.e., $\mathbb{P}_j(X_{t_0} = i)$ equals 1 if and only if $i = j$, otherwise equal to 0. △

LEMMA 6.3. *Let* $(X_t)_{t \in T}$ *be a stationary discrete time Markov chain with transition matrix* $\mathbf{P}$ *and let* $\mathbb{P}$ *be the probability from Theorem 5.1. We can define* $\mathbb{P}_j$ *introduced in Notation 6.2 by*

$$\mathbb{P}_j(E) = \mathbb{P}(E \,|\, X_{t_0} = j). \tag{23}$$

PROOF. The claim here is that by defining $\mathbb{P}_j$ as in (23), $(X_t)_{t \in T}$ is a stationary discrete time Markov chain with transition matrix $\mathbf{P}$ and initial distribution δ_j. For this, let $j_0, ..., j_n, j, k \in S$ be states, then we want to show

$$\mathbb{P}_j(X_{t_{n+1}} = k \,|\, X_{t_0} = j_0, ..., X_{t_n} = j_n) = \mathbb{P}_j(X_{t_{n+1}} = k \,|\, X_{t_n} = j_n)$$

whenever $\mathbb{P}_j(X_{t_0} = j_0, ..., X_{t_n} = j_n) > 0$. Of course, since $\mathbb{P}_j(X_{t_0} = j_0) = \mathbb{P}(X_{t_0} = j_0 \,|\, X_{t_0} = j) = 0$ unless $j_0 = j$, we must have $j_0 = j$ (which also shows that the initial distribution with respect to

$\mathbb{P}_j$ is δ_j). Then,

$$\mathbb{P}_j(X_{t_{n+1}} = k \,|\, X_{t_0} = j_0, ..., X_{t_n} = j_n) = \mathbb{P}(X_{t_{n+1}} = k \,|\, X_{t_0} = j_0, ..., X_{t_n} = j_n, X_{t_0} = j)$$
$$\mathbb{P}(X_{t_{n+1}} = k \,|\, X_{t_0} = k, ..., X_{t_n} = j_n) = \mathbb{P}(X_{t_{n+1}} = k \,|\, X_{t_n} = j_n).$$

From this calculation, it easily follows that $\mathbb{P}_j$ inherits the desired properties from $\mathbb{P}$ (making $(X_t)_{t\in T}$ stationary and Markovian) and hence the result follows. □

NOTATION 6.3. Suppose that $\nu : S \to \mathbb{R}$ is a probability mass function. We then use the symbol $\vec{\nu}$ to represent the row vector indexed by S whose jth entry is $\nu(j)$. For example, suppose $S = \{0, 1, 2\}$ and $\nu : S \to \mathbb{R}$ is defined by $\nu(0) = 1/2$, $\nu(1) = 1/3$, and $\nu(2) = 1/6$. Then $\vec{\nu} = \begin{pmatrix} 1/2 & 1/3 & 1/6 \end{pmatrix}$. Notice that the ordering of the entries in $\vec{\nu}$ is consistent with our ordering of the elements in S. △

PROPOSITION 6.4. *Let $(X_t)_{t\in T}$ be a stationary discrete time Markov chain with transition matrix $\mathbf{P}$ and initial distribution ν with respect to $\mathbb{P}_\nu$. Then for any $m \in \mathbb{N}$ and any state j,*

$$\mathbb{P}_\nu(X_{t_m} = j) = \sum_{k\in S} \nu(k)\, p_{kj}^{(m)} = [\vec{\nu}\,\mathbf{P}^m]_j$$

where $[\vec{\nu}\,\mathbf{P}^m]_j$ means the jth element of the vector-matrix multiplication of $\vec{\nu}$ and $\mathbf{P}^m$.

PROOF. The second equality above follows by the first simply by the definition of matrix multiplication. To show the first equality, we have

$$\mathbb{P}_\nu(X_{t_m} = j) = \sum_{k\in S} \underbrace{\mathbb{P}_\nu(X_{t_m} = j \,|\, X_{t_0} = k)}_{p_{kj}^{(m)}} \underbrace{\mathbb{P}_\nu(X_{t_0} = k)}_{\nu(k)}$$

which finishes the proof. □

EXAMPLE 6.1. Suppose that $(X_t)_{t=0}^{\infty}$ is a stationary discrete time Markov chain with transition matrix

$$\mathbf{P} = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/3 & 1/4 & 5/12 \\ 0 & 1/3 & 2/3 \end{pmatrix}$$

indexed by the state space $S = \{0, 1, 2\}$ and initial distribution $\nu(0) = 1/2, \nu(1) = 1/3$, and $\nu(2) = 1/6$. Let us calculate $\mathbb{P}_\nu(X_4 = 1, X_3 = 2, X_2 = 2)$. We have

$$\mathbb{P}_\nu(X_4 = 1, X_3 = 2, X_2 = 2) = \mathbb{P}_\nu(X_4 = 1 \,|\, X_3 = 2, X_2 = 2)\, \mathbb{P}_\nu(X_3 = 2 \,|\, X_2 = 2)\, \mathbb{P}_\nu(X_2 = 2)$$
$$= \mathbb{P}_\nu(X_4 = 1 \,|\, X_3 = 2)\, \mathbb{P}_\nu(X_3 = 2 \,|\, X_2 = 2)\, \mathbb{P}_\nu(X_2 = 2) = p_{21}\, p_{22}\, [\vec{\nu}\,\mathbf{P}^2]_2$$

We already have that $p_{21} = 1/3$ and $p_{22} = 2/3$. We found $\mathbf{P}^2$ in Example 4.1, so

$$\vec{\nu}\mathbf{P}^2 = \begin{pmatrix} 1/2 & 1/3 & 1/6 \end{pmatrix} \begin{pmatrix} 5/12 & 3/8 & 5/24 \\ 1/4 & 53/144 & 55/144 \\ 1/9 & 11/36 & 7/12 \end{pmatrix} = \begin{pmatrix} * & * & 71/216 \end{pmatrix}$$

where we used $*$ above since the only entry we need to know the value of was the one indexed by 2. Hence $[\vec{\nu}\mathbf{P}^2]_2 = 71/216$. Finally, we find,

$$\mathbb{P}_\nu(X_4 = 1, X_3 = 2, X_2 = 2) = (1/3)(2/3)(71/216) = 71/972.$$

△

7. Clarifications

In the last couple sections, we have introduced notions which, at times, has been fairly abstract. Here we try to describe how these abstract ideas will manifest for our purposes.

To start, we have made clear that a discrete time chain $(X_t)_{t\in T}$ can not be considered a Markov process nor stationary without making these considerations with respect to some probability $\mathbb{P}$. We have also said that if we know that $(X_t)_{t\in T}$ is a stationary discrete time Markov chain, which implicitly means that there is some extant probability in the background which we are using to make the Markov and stationary classifications, then we may as well assume that the background probability $\mathbb{P}$ is given by the probability described in Theorem 5.1. It may be suspect that we frivolously change the background probability for convenience; if we were concerned with events in the sample space that did not only involve the evolution of the process $(X_t)_{t\in T}$, this assumption would not be legitimate. However, because we will only be considering the evolution of the system, the calculations and results we are interested in will reduce to analysis of the transition matrix and, possibly, the initial distribution of the process. Therefore, if the original probability is such that $(X_t)_{t\in T}$ is a stationary discrete time Markov chain with transition matrix $\mathbf{P}$, we can apply Theorem 5.1 to create the probability $\mathbb{P}$ from the transition matrix $\mathbf{P}$, and hence any questions regarding only analysis of the transition matrix will be answered identically for the original probability and the probability $\mathbb{P}$ (or any probability such that the process is a stationary discrete time Markov chain with the same transition matrix). Moreover, if there is an initial distribution ν that is important for our analysis, then using $\mathbb{P}_\nu$ described in Theorem 6.2 will ensure that the process $(X_t)_{t\in T}$ is still a stationary discrete time Markov chain with transition matrix $\mathbf{P}$ and the initial distribution ν. Therefore, in calculations where the transition matrix and initial distribution are all that need to be analysed, we can assume that $\mathbb{P}_\nu$ is the background probability without concern.

A second point to make here is that if $(X_t)_{t\in T}$ is a stationary discrete time Markov chain with respect to some probability $\mathbb{P}$, it is certainly possible (and even likely) that $\mathbb{P}(X_{t+} = j \,|\, X_t = i)$ is not defined for many times t because $\mathbb{P}(X_t = i) = 0$. The times t where it is defined will likely depend on the initial distribution of the process. Nonetheless, we only worry about the value $\mathbb{P}(X_{t+} = j \,|\, X_t = i)$ where it is defined, and, for stationarity, that when these values are defined, they do not dependent on the time t. You may even encounter scenarios where, depending on the initial distribution, there are some states $j \in S$ such that $\mathbb{P}(X_t = j) = 0$ for all times t. In this case, you can simply ignore that those states ever existed in the first place, since the process will never reach them; however, if you are considering many initial distributions, instead of worrying about which states won't be included for each initial distribution, it is common practice to using the probability $\mathbb{P}$ from Theorem 5.1 to create the transition matrix $\mathbf{P}$, and then use $\mathbf{P}$ as the transition matrix underlying the process.

EXAMPLE 7.1. Consider the stationary discrete time Markov chain $(X_t)_{t\in T}$ with state space $S = \{0, 1, 2\}$ and transition matrix

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \end{pmatrix}$$

Notice that if the process starts initially in either states 1 or 2, it will never be able to reach state 0; i.e., if $X_{t_0} \in \{1, 2\}$, then $\mathbb{P}(X_t = 0) = 0$ for every t. So, if we only ever care about initial distributions which start in state 1 or state 2, then we can reduce the problem to considering only the state space $\{1, 2\}$ and the transition matrix $\begin{pmatrix} 1/2 & 1/2 \\ 0 & 1 \end{pmatrix}$. At that, if we know that $X_{t_0} = 2$, then it stays in 2 evermore, never reaching states 0 or 1. So again, if our only concern has the initial distribution $X_{t_0} = 2$, then we can reduce to the state space $\{2\}$ and the transition matrix $\begin{pmatrix} 1 \end{pmatrix}$. However, it is common to want to consider many possible initial distributions, some which may have the process start in any of the possible states, so it is still common to use $\mathbf{P}$ as the transition matrix for the process without reduction. △

8. Another Perspective of the Markov Property

THEOREM 8.1. *Let $(X_t)_{t\in T}$ be a discrete time chain on the state space S, and let $\mathbb{P}$ be a probability. The following are equivalent.*

(1) *The process satisfies the Markov property with respect to the probability $\mathbb{P}$.*

(2) *For every state $j \in S$ and every time $t_m \in T$ such that $\mathbb{P}(X_{t_m} = j) > 0$, with respect to the probability $\mathbb{P}(\,\cdot \,|\, X_{t_m} = j)$, the process $(X_{t_k})_{k=m}^{\infty}$ is independent of the past states $X_{t_0}, X_{t_1}, ..., X_{t_{m-1}}$.*

Poetically, this says that the process is Markov if and only if given its present, its future and past are independent.

PROOF. (2) $\implies$ (1). Without losing generality, assume that $t_0 = 0$. Given any times $t_0 \leq t_1 < \cdots < t_n < t$ and states $j_1, ..., j_n, j \in S$, we want to show

$$\mathbb{P}(X_t = j \,|\, X_{t_1} = j_1, ..., X_{t_n} = j_n) = \mathbb{P}(X_t = j \,|\, X_{t_n} = j_n).$$

We have

$$\begin{aligned}
&\mathbb{P}(X_t = j \,|\, X_{t_1} = j_1, ..., X_{t_n} = j_n) \\
&= \frac{\mathbb{P}(X_t = j, X_{t_1} = j_1, ..., X_{t_{n-1}} = j_{n-1} \,|\, X_{t_n} = j_n)\mathbb{P}(X_{t_n} = j_n)}{\mathbb{P}(X_{t_0} = j_0, X_{t_1} = j_1, ..., X_{t_n} = j_n)} \\
&= \frac{\mathbb{P}(X_t = j \,|\, X_{t_n} = j_n)\mathbb{P}(X_{t_1} = j_1, ..., X_{t_{n-1}} = j_{n-1} \,|\, X_{t_n} = j_n)\mathbb{P}(X_{t_n} = j_n)}{\mathbb{P}(X_{t_1} = j_1, ..., X_{t_n} = j_n)} \\
&= \frac{\mathbb{P}(X_t = j \,|\, X_{t-1} = j)\,\mathbb{P}(X_{t_1} = j_1, ..., X_{t_n} = j_n)}{\mathbb{P}(X_{t_1} = j_1, ..., X_{t_n} = j_n)} \\
&= \mathbb{P}(X_t = j \,|\, X_{t_n} = j_{t_n}).
\end{aligned}$$

The first equality came by manipulating the definition of conditional probability, the second equality used assumption (2) of independence with respect to $\mathbb{P}(\,\cdot \,|\, X_{t-1} = j_{t-1})$, and the third equality by another manipulation of the definition of conditional probability. What we have shown is that assuming (2), we can conclude the process has the Markov property.

(1) $\implies$ (2). This follows similarly. As before, we have

$$\begin{aligned}
&\mathbb{P}(X_t = j \,|\, X_{t_n} = j_n) \\
&= \frac{\mathbb{P}(X_t = j \,|\, X_{t-1} = j)\,\mathbb{P}(X_{t_1} = j_1, ..., X_{t_n} = j_n)}{\mathbb{P}(X_{t_1} = j_1, ..., X_{t_n} = j_n)} \\
&= \frac{\mathbb{P}(X_t = j \,|\, X_{t_n} = j_n)\mathbb{P}(X_{t_1} = j_1, ..., X_{t_{n-1}} = j_{n-1} \,|\, X_{t_n} = j_n)\mathbb{P}(X_{t_n} = j_n)}{\mathbb{P}(X_{t_1} = j_1, ..., X_{t_n} = j_n)}
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}(X_t = j \,|\, X_{t_1} = j_1, ..., X_{t_n} = j_n) \\
&= \frac{\mathbb{P}(X_t = j, X_{t_1} = j_1, ..., X_{t_{n-1}} = j_{n-1} \,|\, X_{t_n} = j_n)\mathbb{P}(X_{t_n} = j_n)}{\mathbb{P}(X_{t_1} = j_1, ..., X_{t_n} = j_n)}
\end{aligned}$$

Since we are assuming the Markov property, it must be that all the expressions above are equal, which implies that

$$\begin{aligned}&\mathbb{P}(X_t = j, X_{t_1} = j_1, ..., X_{t_{n-1}} = j_{n-1} \,|\, X_{t_n} = j_n)\\ &\quad = \mathbb{P}(X_t = j \,|\, X_{t_n} = j_n)\, \mathbb{P}(X_{t_1} = j_1, ..., X_{t_{n-1}} = j_{n-1} \,|\, X_{t_n} = j_n)\end{aligned}$$

This is enough for us to deduce that $\{X_{t_n}, X_t\}$ is independent of $\{X_{t_1}, X_{t_2}, ..., X_{t_n}\}$ using the probability $\mathbb{P}(\,\cdot \,|\, X_{t_n} = j_n)$. Following the analogous argument, we can deduce this for any finite collection of future times $\{X_{t_n}, X_{t_n+1}, ...\}$ and past times $\{X_{t_0}, X_{t_0+1}, ..., X_{t_n}\}$. Since t_n was chosen arbitrarily, the result follows. □

9. Exercises

(1) Consider the two team goal-scoring game described in Example 1.1. Describe how the processes $(A_t)_{t\in T}$, $(B_t)_{t\in T}$, and $(X_t)_{t\in T}$ might not satisfy the Markov property if the *momentum* of a team is a factor in their performance. Here *momentum* means that a team that has a streak of goals will be more likely to score the subsequent goal (due to psychological reasons such as confidence); so, for example, if momentum exists and team A scored the previous three goals in a row, their chance of scoring the next goal of the game is greater than if team B scored some (or all) of the previous three goals.

(2) Let $\{Y_i\}_{i=1}^{\infty}$ be a collection of independent Bernoulli random variables. Define $X_n = \sum_{i=1}^{n} Y_i$ for $n = 1, 2, ...$ In Example 2.1, we showed that the process $(X_n)_{n=1}^{\infty}$ satisfies the Markov property and hence concluded that it is a discrete time Markov chain. Explain why $(X_n)_{n=1}^{\infty}$ will be stationary only in the case that the Y_is are identically distributed, otherwise, if the Y_is are not identically distributed, then $(X_n)_{n=1}^{\infty}$ is not stationary.

(3) Suppose that $\{Y_i\}_{i=1}^{\infty}$ are iid random variables such that $\mathbb{P}(Y_i = 1) = p$ and $\mathbb{P}(Y_i = -1) = 1 - p$. Define the process $(X_n)_{n=0}^{\infty}$ by the following recursive relationship $X_0 = 0$ and

$$X_n = \begin{cases} -2 & \text{if } X_{n-1} + Y_n < -2 \\ X_{n-1} + Y_n & \text{if } -2 \le X_{n-1} + Y_n \le 2 \\ 2 & \text{if } X_{n-1} + Y_n > 2 \end{cases}$$

for $n \ge 1$. Show that $(X_n)_{n=0}^{\infty}$ is a stationary discrete time Markov chain, find its state space S, and calculate its transition matrix $\mathbf{P}$ (making sure the entries in $\mathbf{P}$ are ordered consistently with the ordering you gave for S).

(4) Use this exercise to convince yourself that using different probabilities, the same discrete time chain may produce different stationary discrete time Markov chains with different transition matrices (we only consider two probabilities here in this problem; there are many other probabilities that can be chosen for which the process is not stationary or does not satisfy the Markov property). Consider two states 0 or 1 which a process $(X_t)_{t=0}^{\infty}$ moves between. At each time step, $t = 0, 1, 2, 3, ...$, a coin is flipped; at time $t = 0$, if the coin is heads then the process starts in state 1, or if the coin is tails the process starts in state 0; for $t \geq 1$, if the coin is heads, then the process stays in whichever state it is currently in, or if the coin is tails, then the process switches states.

   (a) Suppose that we assign the underlying probability $\mathbb{P}$ being one which would consider the coin to be fair and each flip independent. Give a brief justification why $(X_t)_{t=0}^{\infty}$ is a stationary discrete time Markov chain with respect to this probability $\mathbb{P}$, and find the transition matrix $\mathbf{P}$ for the process.

   (b) Suppose that we assign the underlying probability $\mathbb{P}$ begin one which would consider the coin to be biased such that the probability of flipping a heads is $1/3$; also suppose that with respect to this probability the coin flips are independent. Give a brief justification why $(X_t)_{t=0}^{\infty}$ is a stationary discrete time Markov chain with respect to this probability $\mathbb{P}$, and find the transition matrix $\mathbf{P}$ for the process.

(5) Suppose that $\mathbf{A}$ and $\mathbf{B}$ are two $N \times N$ stochastic matrices. Show that the product $\mathbf{AB}$ is also a stochastic matrix. (If you are comfortable with mathematical induction, try using this to prove that for any stochastic matrix $\mathbf{P}$ and for any $m \in \mathbb{N}$, $\mathbf{P}^m$ is again a stochastic matrix). *Hint*: Let a_{ij} and b_{ij} be the entries of $\mathbf{A}$ and $\mathbf{B}$, respectively. Then $[\mathbf{AB}]_{ij} = \sum_{k=1}^{N} a_{ik} b_{kj}$, so the sum along the ith row of $\mathbf{AB}$ is

$$\sum_{m=1}^{N} [\mathbf{AB}]_{im} = \sum_{m=1}^{N} \sum_{k=1}^{N} a_{ik} b_{km}.$$

Change the order of summation.

(6) Consider continually rolling a three-sided die, with sides labelled: $1, 2, 3$. Let X_0 be the outcome of the first roll. Then we recursively define the process $(X_n)_{n=0}^{\infty}$ as follows: For each subsequent step $n + 1$, we define X_{n+1} by the rule

$$X_{n+1} = \begin{cases} \max\{(n+1)\text{st die roll value}, X_n\} & X_n \leq 2 \\ 0 & X_n = 3 \end{cases}$$

Assume that each roll of the die is independent of other rolls and that the die is fair (i.e., that each side of the die comes up with 1/3 probability).

(a) Briefly justify that this is a stationary discrete time Markov chain and find the transition matrix $\mathbf{P}$.

(b) For each $i \in S$, find $\mathbb{P}_i(X_3 = 0)$. *Note*: $\mathbb{P}_i(X_3 = 0) = \mathbb{P}(X_3 = 0 \,|\, X_0 = i)$.

(c) For each $j \in S$, find $\mathbb{P}_0(X_3 = j)$.

(7) Let $(X_n)_{n=0}^{\infty}$ be a discrete time chain with state space S, and that $\nu : S \to \mathbb{R}$ is a probability mass function. Suppose that $\mathbb{P}_\nu$ is a probability (on the sample space of the process) such that $(X_t)_{t\in T}$ is a stationary discrete time Markov chain with transition matrix $\mathbf{P}$. For a time $n_1 \geq 1$, explain why with respect to the probability $\mathbb{P}_\nu$, the process $(X_n)_{n=n_1}^{\infty}$ is a stationary discrete time Markov chain with transition matrix $\mathbf{P}$ and initial distribution $\nu_1 : S \to \mathbb{R}$, where $\nu_1(j) = \mathbb{P}_\nu(X_{n_1} = j)$ for each $j \in S$. In other words, the process $(X_n)_{n=n_1}^{\infty}$ under the probability $\mathbb{P}_\nu$ can be considered the same stationary discrete time Markov chain as the process $(X_n)_{n=0}^{\infty}$ under the probability $\mathbb{P}_{\nu_1}$.

(8) Let $(X_t)_{t\in T}$ be a stationary discrete time Markov chain with state space S and one-step transition probabilities p_{ij}. For each state $j \in S$, define

$$\tau_j = \min\{m \geq 0 : X_{t_m} = j\},$$

with the convention that $\min \varnothing = \infty$; notice that τ_j represents the number of steps it takes the process to first hit state j. Using Exercise (7), deduce the following equations

$$\mathbb{P}_i(\tau_j = t_m) = \sum_{k\in S} p_{ik}\, \mathbb{P}_k(\tau_j = t_{m-1})$$

and

$$\mathbb{E}[\tau_j \,|\, X_{t_0} = i] = 1 + \sum_{k\in S} p_{ik}\, \mathbb{E}[\tau_j \,|\, X_{t_0} = k]$$

We are assuming the time variables are $T = \{t_0, t_1, t_2, ...\}$ and m is a positive integer.

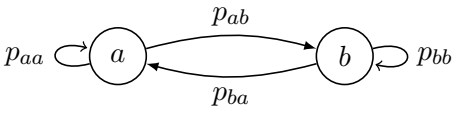
CHAPTER 4

Jump Diagrams and Communication Classes

1. Jump Diagrams

Let $\mathbf{P}$ be a stochastic matrix indexed by the states S and let p_{ij} be the (i,j)th element of $\mathbf{P}$ for $i,j \in S$. A useful way to visualize the information stored within the stochastic matrix $\mathbf{P}$ is through a *jump diagram*. The jump diagram is a directed graph where each state in S is a vertex of the graph and a weighted and directed edge emanates from state i towards state j whenever p_{ij} is non-zero; the weight of this edge is the value of p_{ij}.

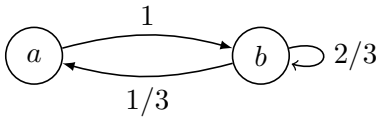
EXAMPLE 1.1. Let $\mathbf{P}$ be a 2×2 stochastic matrix indexed by the states $S = \{a, b\}$. The corresponding jump diagram will have the form



So, if for example,

$$\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1/3 & 2/3 \end{pmatrix}$$

then the corresponding jump diagram is

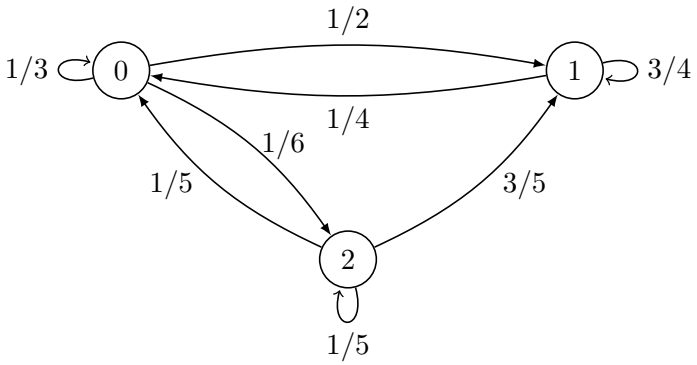


where any edges of weight 0 are omitted. △

EXAMPLE 1.2. Consider the stochastic matrix

$$\mathbf{P} = \begin{pmatrix} 1/3 & 1/2 & 1/6 \\ 1/4 & 3/4 & 0 \\ 1/5 & 3/5 & 1/5 \end{pmatrix}$$

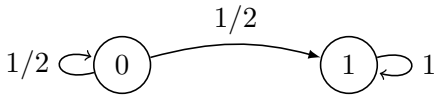
We will assume that the matrix is labelled by the states $S = \{0, 1, 2\}$ (of course, without prior knowledge of what the actual states are, you can use any three symbols you choose to label the states). Then this stochastic matrix corresponds to the jump diagram



△

REMARK 1.1. It is important to notice that given a jump diagram, we can recreate the stochastic matrix it corresponds to. In fact, we realize that neither jump diagrams nor stochastic matrices are a more fundamental object than the other since given either one, you can create the other. △

EXAMPLE 1.3. Consider the jump diagram



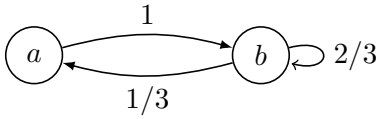
then the corresponding stochastic matrix is

$$\mathbf{P} = \begin{pmatrix} 1/2 & 1/2 \\ 0 & 1 \end{pmatrix}$$

indexed by the state space $S = \{0, 1\}$. △

DEFINITION 1.1. Let n be a positive integer. An *allowable path* of length n from state i to state j is an ordered tuple of states $(k_0, k_1, ..., k_n)$ where $k_0 = i$, $k_n = j$, and there is an edge (an arrow in the jump diagram) with positive weight from k_p to k_{p+1} for each $p = 0, 1, ..., n$. In the case that there is an allowable path from i to j of length n, we say that the system can move from i to j in n *steps* (or *jumps*) and call each edge between consecutive states in the path a *step* (or *jump*) of the path. △

EXAMPLE 1.4. It is notationally and analytically easier to consider a path as a tuple of states $(k_0, k_1, ..., k_n)$, however it is often nice to interpret such paths as moves within the jump diagram following the directed edges (the arrows). For example, in the jump diagram



the tuple (a, b, b, a, b, b) is an allowable path from a to b of length 5. But it is easy to see this as the five moves: $a \to b$ then $b \to b$ then $b \to a$ then $a \to b$ and finally $b \to b$. Alternatively, the tuple (a, b, a, a, b) is not an allowable path since the third step would be $a \to a$, for which there is no arrow of positive weight from a to a. △

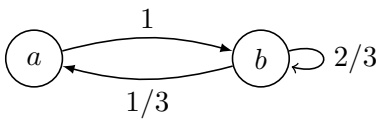
PROPOSITION 1.1. *Let n be some positive integer. With respect to the stochastic matrix $\mathbf{P}$, there is an allowable path from state i to state j of length n if and only if $[\mathbf{P}^n]_{ij} > 0$. Moreover, $[\mathbf{P}^n]_{ij}$ can be found by summing over all terms $p_{ik_1}p_{k_1k_2}\cdots p_{k_{n-1}j}$ where $(i, k_1, ..., k_{n-1}, j)$ is an allowable path from i to j of length n.*

PROOF. Suppose that $(k_0, k_1, ..., k_n)$ is some tuple of length n path (we're not assuming that it is or is not allowable at this point) from $i = k_0$ to $j = k_n$. This path is allowable if and only if $p_{k_pk_{p+1}} > 0$ for every $p = 0, 1, ..., n-1$, which in turn happens if and only if the product $p_{i\,k_1}p_{k_1k_2}\cdots p_{k_{n-1}j}$ is positive. Hence we deduce that there is an allowable path from i to j of length n if and only if there is at least one such product $p_{ik_1}p_{k_1k_2}\cdots p_{k_{n-1}\,j}$ which is positive as we vary $k_1, ..., k_{n-1}$ through all possible states. Realizing that by matrix multiplication

$$[\mathbf{P}^n]_{ij} = \sum_{k_1,...,k_{n-1}\in S} p_{i\,k_1}p_{k_1k_2}\cdots p_{k_{n-1}\,j}$$

we learn that $[\mathbf{P}^n]_{ij} > 0$ if and only if one of these products $p_{i\,k_1}p_{k_1k_2}\cdots p_{k_{n-1}\,j}$ is positive, which happens if and only if there is an allowable path from i to j of length n. Hence the first part of the proposition follows. From here, we see that the sum for $[\mathbf{P}^n]_{ij}$ is only contributed to by those terms which are positive, and thus we can sum such terms only over the allowable paths. □

EXAMPLE 1.5. Consider again the jump diagram



corresponding to the transition matrix

$$\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1/3 & 2/3 \end{pmatrix}$$

The only allowable paths from a to b of length 4 are: (a,b,b,b,b), (a,b,b,a,b), and (a,b,a,b,b). Considering Proposition 1.1, we can thus surmise that

$$[\mathbf{P}^4]_{ab} = p_{ab}p_{bb}p_{bb}p_{bb} + p_{ab}p_{bb}p_{ba}p_{ab} + p_{ab}p_{ba}p_{ab}p_{bb} = (1)(2/3)^3 + (1)^2(2/3)(1/3) + (1)^2(1/3)(2/3)$$
$$= 20/27$$

Indeed, we have

$$\mathbf{P}^4 = \begin{pmatrix} 7/27 & 20/27 \\ 20/81 & 61/81 \end{pmatrix}$$

where we see the (a,b)th entry agrees with our calculation. △

2. Communication Classes

We start here by introducing a convention that is a natural extension to m-step transition probabilities and hence to allowable paths. We have so far defined $p_{ij}^{(m)}$ as the m-step transition probability from state i to j for some stationary discrete time Markov chain. We learned in Theorem 4.1 that $p_{ij}^{(m)} = [\mathbf{P}^m]_{ij}$ for the transition matrix $\mathbf{P}$, and hence these transition probabilities can be defined in terms of a stochastic matrix, simple as the entries of that matrix raised to the appropriate power. We then learned in Proposition 1.1 that for a stochastic matrix, $[\mathbf{P}^m]_{ij}$ has an interpretation for the allowability of paths along a jump diagram. Keeping these interpretations in mind, we naturally define the 0th step transition probability $p_{ij}^{(0)}$ as $[\mathbf{P}^0]_{ij}$. From linear algebra, we recall that $\mathbf{P}^0$ is defined as the identity matrix $\mathbf{I}$, hence using Kronecker delta notation, we have $p_{ij}^{(0)} = \delta_{ij}$, where δ_{ij} is 1 when $i = j$, and 0 otherwise. In terms of Markov chains, we can interpret this as the 0th step transition probability: $p_{ij}^{(0)}$ is the probability that process moves from state i to state j in 0 steps – from this standpoint, the $p_{ij}^{(0)} = \delta_{ij}$ makes perfect sense. In terms of allowable paths, we interpret this as all 0 length paths are allowable: there is an allowable path from any state to itself of length 0.

Definition 2.1. Let $\mathbf{P}$ be a stochastic matrix indexed by the states S. We will say that j is *accessible* to state i with respect to $\mathbf{P}$ when there is an allowable path from i to j; note that i is accessible to i for every $i \in S$ by our length 0 path convention.. To denote that j is accessible to i, we write $i \longrightarrow j$. If j is accessible to i and i is accessible to j, we say that i and j *communicate*, denoted $i \longleftrightarrow j$. For a state i, the *communication class* of i induced by $\mathbf{P}$ is a subset C of the state space S containing all states j which communicate with i. That is, the communication class of i is $C = \{j \in S \text{ s.t. } i \longleftrightarrow j\}$. △

LEMMA 2.1. *Accessibility is a transitive relation. That is, given a stochastic matrix* $\mathbf{P}$ *indexed by* S*, if* $i, j, k \in S$ *with* $i \longrightarrow j$ *and* $j \longrightarrow k$*, then* $i \longrightarrow k$*.*

PROOF. If $i, j, k \in S$ with $i \longrightarrow j$ and $j \longrightarrow k$, then there exist $m, n \in \mathbb{N}$ such that $[\mathbf{P}^m]_{ij} > 0$ and $[\mathbf{P}^n]_{jk} > 0$. Hence

$$0 < [\mathbf{P}^m]_{ij} \, [\mathbf{P}^n]_{jk} \leq \sum_{l \in S} [\mathbf{P}^m]_{il} \, [\mathbf{P}^n]_{lk} = [\mathbf{P}^{m+n}]_{ik},$$

showing that there is an allowable path from i to k of length $m + n$, hence $i \longrightarrow k$. Notice that had we been considering this result in terms of multi-step transition probabilities of some stationary discrete time Markov chain, the last equality is an example of an application of the Chapman-Kolmogorov Equation, Corollary 4.3. □

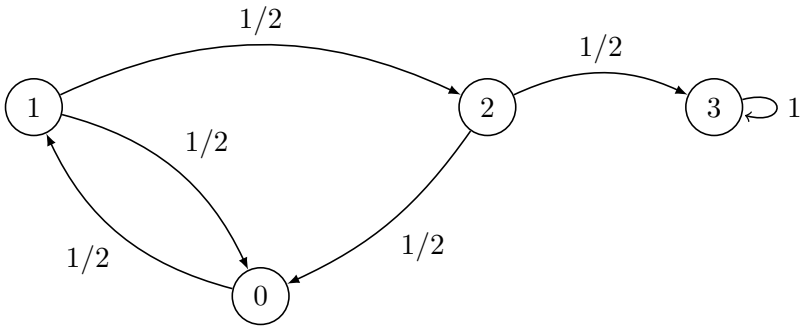
THEOREM 2.2. *Let* $\mathbf{P}$ *be a stochastic matrix indexed by the state space* S*. The collection of distinct communication classes induced by* $\mathbf{P}$ *are a partition of* S*.*

PROOF. A basic result in set theory is that implies that if we can show that the communication relation $i \longleftrightarrow j$ is an equivalence relation on S, then the set of distinct equivalence classes partition S. Realizing that the communication classes are the equivalence classes, we will hence be done. What this proof will boil down to then is to show: (reflexivity) $i \longleftrightarrow i$ for every $i \in S$; (symmetry) if $i \longleftrightarrow j$ then $j \longleftrightarrow i$ for every $i, j \in S$; and (transitivity) if $i \longleftrightarrow j$ and $j \longleftrightarrow k$ then $i \longleftrightarrow k$ for every $i, j, k \in S$. First, reflexivity follows by the length 0 path convention and transitivity follows easily by Lemma 2.1. What remains is symmetry, but this is immediate since if $i \longleftrightarrow j$ then, by definition, $i \longrightarrow j$ and $j \longrightarrow i$, which means $j \longrightarrow i$ and $i \longrightarrow j$, and hence $j \longleftrightarrow i$. □

DEFINITION 2.2. Let $\mathbf{P}$ be a stochastic matrix indexed by the state space S. If there is only one distinct communication class C induced by $\mathbf{P}$, then we call $\mathbf{P}$ *irreducible*. Necessarily, if $\mathbf{P}$ is irreducible, then $C = S$. △

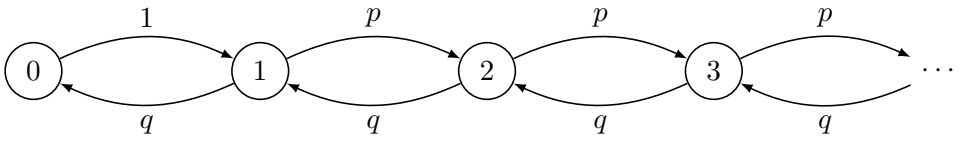
DEFINITION 2.3. Let $\mathbf{P}$ be a stochastic matrix indexed by the state space S and let C be some communication class induced by $\mathbf{P}$. We call C an *open* communication class when there exists some $i \in C$ and some $j \in S \setminus C$ such that $i \longrightarrow j$. If C is not an *open* communication class, it is called a *closed* communication class. Visually, C is open whenever there is an allowable path along the jump diagram leading out of C; otherwise, it is closed. △

EXAMPLE 2.1. Suppose the following is a jump diagram for the stochastic matrix $\mathbf{P}$ indexed by the state space $S = \{0, 1, 2, 3\}$.



In this example there are two distinct communication classes $O = \{0, 1, 2\}$ and $C = \{3\}$. In this case, O is open since there is a path from any element in O to an element outside of O (note that $i \longrightarrow 3$ for every $i \in O$); however, C is closed since there is no allowable path from 3 to any other element. △

EXAMPLE 2.2. Suppose that $p, q \in [0, 1]$ with $p + q = 1$, and $\mathbf{P}$ is a stochastic matrix with jump diagram



If $p > 0$ and $q > 0$, then $\mathbf{P}$ is irreducible since there is only one distinct communication class $C = \{0, 1, 2, ...\} = S$, because any two states communicate. If $p = 0$ and $q = 1$, then there are infinitely many distinct communication classes $C = \{0, 1\}$, $O_1 = \{2\}$, $O_2 = \{3\}$, ...; in this case, C is closed, but every other communication class is open. If $p = 1$ and $q = 0$, there are again infinitely many distinct communication classes $O_0 = \{0\}$, $O_1 = \{1\}$, $O_2 = \{2\}$, ...; in this case, every communication class is open. △

3. Reduction of a Stochastic Matrix

We will rely heavily on the notation in Section 2 of Appendix A. Importantly, if $\mathbf{P}$ is indexed by S and $A \subseteq S$, then by $\mathbf{P}_A$, we denote the submatrix of $\mathbf{P}$ created by using only those entries in $\mathbf{P}$ indexed by A.

LEMMA 3.1. *Suppose that $\mathbf{P}$ is a stochastic matrix indexed by S and $C \subseteq \mathbf{P}$ is a communication class induced by $\mathbf{P}$. Then C is closed if and only if either $C = S$ or $[\mathbf{P}^m]_{ij} = 0$ for every $i \in C$ and $j \in S \setminus C$ and for every positive integer m.*

PROOF. The case $C = S$ (i.e., $\mathbf{P}$ is irreducible) is Exercise 1. Otherwise, note that $[\mathbf{P}^m]_{ij} > 0$ for some positive integer m if and only if there is an allowable path from i to j. Since C is closed

if and only if there are no allowable paths from any $i \in C$ to any $j \in S \setminus C$, the result trivially follows. □

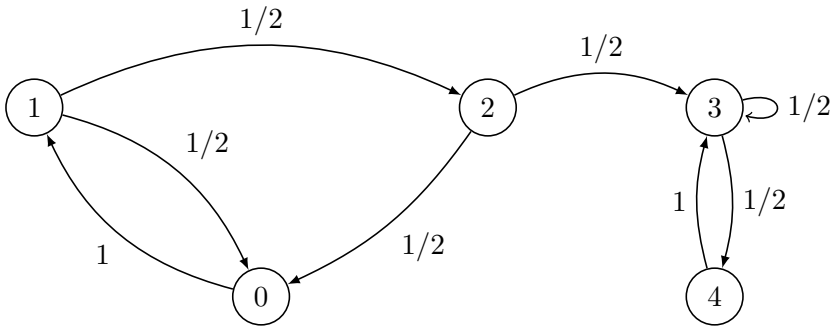
COROLLARY 3.2. *Suppose that* $\mathbf{P}$ *is a stochastic matrix indexed by* S *and* $C \subseteq \mathbf{P}$ *is a communication class induced by* $\mathbf{P}$*. Then* C *is a closed communication class if and only if* $\mathbf{P}_C$ *is a stochastic matrix.*

PROOF. Since every entry in $\mathbf{P}_C$ is an entry in $\mathbf{P}$, it follows that all entries in $\mathbf{P}_C$ are non-negative. So, we only need to show that the entries along the every row of $\mathbf{P}_C$ sum to one. If $C = S$, then there is nothing to show. Let $i \in C$. By Lemma 3.1, for every $j \notin S$, $[\mathbf{P}]_{ij} = 0$. Hence

$$\begin{aligned}\text{Sum along } i\text{th row of } \mathbf{P} = 1 = \sum_{j\in S}[\mathbf{P}]_{ij} &= \sum_{j\in S\setminus C}\overbrace{[\mathbf{P}]_{ij}}^{0} + \sum_{j\in C}[\mathbf{P}]_{ij} \\ &= \sum_{j\in C}[\mathbf{P}]_{ij} = \text{Sum along } i\text{th row of } \mathbf{P}_C\end{aligned}$$

This shows that $\mathbf{P}_C$ is a stochastic matrix. □

EXAMPLE 3.1. Suppose the following is a jump diagram for the stochastic matrix $\mathbf{P}$ indexed by the state space $S = \{0, 1, 2, 3, 4\}$.



In this example there are two distinct communication classes $O = \{0, 1, 2\}$ and $C = \{3, 4\}$. The corresponding transition matrix is

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Considering the sub-matrices indexed by the communication classes, we have

$$\mathbf{P}_O = \begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1/2 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{P}_C = \begin{pmatrix} 1/2 & 1/2 \\ 1 & 0 \end{pmatrix}$$

where we see that the the sub-matrix for the open communication class $\mathbf{P}_{C_1}$ is not stochastic (the last row does not sum to 1), whereas for the closed communication class $\mathbf{P}_{C_2}$ is stochastic. So, applying Corollary 3.2, we can conclude that C_1 is open and C_2 is closed. Of course, we also know that C_1 is open since there are allowable paths from C_1 to C_2; whereas C_2 is closed since there is no allowable paths from C_2 leading out of C_2. △

Let us finish this section by noting another way to interpret Lemma 3.1 in terms of the matrix $\mathbf{P}$ overall. Suppose that there exist exactly N distinct closed communication classes $C_1, C_2, ..., C_N$ and M distinct open communication classes $O_1, O_2, ..., O_M$ induced by $\mathbf{P}$. Then, as a direct consequence of the above results, we can arrange the order of S in such a way that $\mathbf{P}$ has the block matrix form

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_{C_1} & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{P}_{C_2} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & \cdots & \mathbf{P}_{C_N} & 0 & 0 & \cdots & 0 \\ * & * & * & * & \mathbf{P}_{O_1} & * & \cdots & * \\ * & * & * & * & * & \mathbf{P}_{O_2} & \cdots & * \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ * & * & * & * & * & \cdots & * & \mathbf{P}_{O_M} \end{pmatrix}$$

where $*$ symbolizes a potentially non-zero matrix of appropriate dimension, and the 0s represent 0 matrices of the appropriate dimensions.

Example 3.2. As an illustration, notice that in the previous example, Example 3.1, had we reordered S as $S = \{3, 4, 0, 1, 2\}$, then $\mathbf{P}$ would have the form

$$\mathbf{P} = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \mathbf{P}_C & 0 \\ * & \mathbf{P}_O \end{pmatrix}$$

where here the 0 in the right most block matrix represents the 2×3 zero matrix, and $*$ represents a 3×2 matrix with zeros everywhere except for the value $1/2$ in the bottom left entry. △

4. The Period of States

DEFINITION 4.1. Suppose that $\mathbf{P}$ is a stochastic matrix indexed by S. Given a state $i \in S$, we say that the *period* of i induced by $\mathbf{P}$ is

$$\begin{aligned} d(i) &= \gcd\{\text{positive integers } m \text{ s.t. } [\mathbf{P}^m]_{ii} > 0\} \\ &= \gcd\{\text{allowable path lengths from } i \text{ to } i \text{ with length at least } 1\} \end{aligned}$$

The convention here is that $\gcd\{\varnothing\} = 0$; i.e., if there is no allowable path from i to i of length at least 1, then $d(i) = 0$. If $d(i) = 1$, then we say i is *aperiodic*. △

THEOREM 4.1. *Let $\mathbf{P}$ be a stochastic matrix indexed by S. If states $i, j \in S$ communicate, then $d(i) = d(j)$. That is, all states within the same communication class have the same period.*

PROOF. For convenience, we write $d_i = d(i)$ and $d_j = d(j)$. We will show that $d_i = d_j$ by showing that $d_i \leq d_j$ and that $d_j \leq d_i$. Since by assumption $i \longleftrightarrow j$, there are positive integers m_1 and m_2 such that $[\mathbf{P}^{m_1}]_{ij} > 0$ and $[\mathbf{P}^{m_2}]_{ji} > 0$. Hence $[\mathbf{P}^{m_1+m_2}]_{ii} > 0$ since the Chapmann-Kolmogorov equations yield

$$[\mathbf{P}^{m_1+m_2}]_{ii} = \sum_{k \in S} [\mathbf{P}^{m_1}]_{ik} [\mathbf{P}^{m_2}]_{ki} \geq [\mathbf{P}^{m_1}]_{ij} [\mathbf{P}^{m_2}]_{ji} > 0.$$

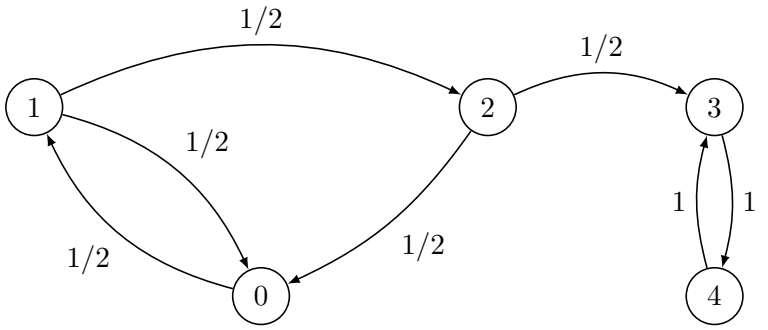
By the definition of the period, d_i divides $m_1 + m_2$. Further, given another positive integer m_3 such that $[\mathbf{P}^{m_3}]_{jj} > 0$, then $[\mathbf{P}^{m_1+m_2+m_3}]_{ii} > 0$ since, similar to before,

$$[\mathbf{P}^{m_1+m_2+m_3}]_{ii} = \sum_{k,l \in S} [\mathbf{P}^{m_1}]_{ik} [\mathbf{P}^{m_3}]_{kl} [\mathbf{P}^{m_2}]_{li} \geq [\mathbf{P}^{m_1}]_{ij} [\mathbf{P}^{m_3}]_{jj} [\mathbf{P}^{m_2}]_{ji} > 0.$$

Therefore d_i also divides $m_1 + m_2 + m_3$. Since d_i divides $m_1 + m_2$ and d_i divides $m_1 + m_2 + m_3$, then d_i divides m_3. This further implies that d_i divides any positive integer m such that $[\mathbf{P}^m]_{jj} > 0$ because m_3 was chosen arbitrarily. By definition, d_j is the greatest of all such divisors, so therefore $d_i \leq d_j$. By the symmetry of this argument (just redo the argument interchanging i and j), we also see that $d_j \leq d_i$, proving that $d_i = d_j$. □

DEFINITION 4.2. If C is a communication class, the *period* of C is defined as $d(i)$ for any (hence all) $i \in C$. If $d(i) = 1$ for the states $i \in C$, then we say that C is *aperiodic*. △

EXAMPLE 4.1. Suppose the following is a jump diagram for the stochastic matrix $\mathbf{P}$ indexed by the state space $S = \{0, 1, 2, 3, 4\}$.

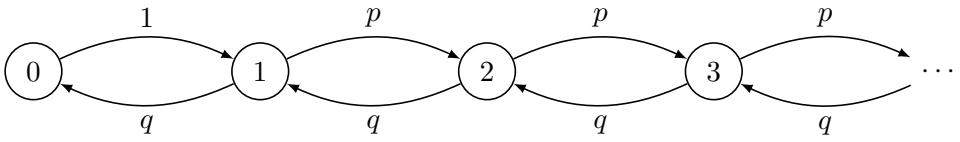


With the two distinct communication classes $O = \{0, 1, 2\}$ and $C = \{3, 4\}$, we only need to find the period of any element in O and any element in C to find the period of all states.

To start, let us find $d(0)$. We see that there are allowable paths from 0 to 0 of length 3 and length 4, and thus $d(0) = \gcd\{3, 4, \cdots\}$ where $\cdots$ represents other allowable path lengths. But, since the greatest common divisor of 3 and 4 is 1, regardless of other allowable path lengths, $d(0) = \gcd\{3, 4, \cdots\} = 1$. Hence, $1 = d(0) = d(1) = d(2)$ and O is aperiodic.

Next, let us find $d(3)$. Considering all allowable path lengths from 3 to 3, we have find lengths 2, 4, 6, ...; generally, a allowable path length will be of the form $2n$ for any positive integer n. Thus, $d(3) = \gcd\{2n \text{ s.t. } n = 1, 2, 3, ...\} = 2$. Hence $2 = d(3) = d(4)$, from which we also see that C is not aperiodic. △

EXAMPLE 4.2. Suppose that $p, q \in [0, 1]$ with $p + q = 1$, and $\mathbf{P}$ is a stochastic matrix with jump diagram



If $p > 0$ and $q > 0$, then $\mathbf{P}$ is irreducible since there is only one distinct communication class $C = \{0, 1, 2, ...\} = S$. We can quickly realize that $d(0) = \gcd\{2n \text{ s.t. } n = 1, 2, 3, ...\} = 2$, so every state has period 2.

If $p = 0$ and $q = 1$, then there are infinitely many distinct communication classes $C = \{0, 1\}$, $O_1 = \{2\}$, $O_2 = \{3\}$, ... Similar to before, we will find that $d(0) = \gcd\{2n \text{ s.t. } n = 1, 2, 3, ...\} = 2$ so $2 = d(0) = d(1)$. However, for any state $i \notin O$, we find that there is no allowable path from i to i of length at least 1, so $d(i) = \gcd\{\varnothing\} = 0$. Thus $d(i) = 0$ for ever state $i \neq 0, 1$.

If $p = 1$ and $q = 0$, there are again infinitely many distinct communication classes $O_0 = \{0\}$, $O_1 = \{1\}$, $O_2 = \{2\}$, ... However, we see that for any state, there is no allowable path from a state to itself of length at least 1, and therefore every state has period 0. △

5. A Probability Perspective

Suppose that $(X_t)_{t\in T}$ is a stationary discrete time Markov chain with state space S. Associated to this process is a transition matrix $\mathbf{P}$, which as a stochastic matrix and hence subject to the analysis we have thusfar developed in this chapter. Here we will introduce some of these previously developed concepts from the probabilist perspective in terms of the process $(X_t)_{t\in T}$, and introduce two important probabilistic classifications of states: transience and recurrence.

To start, we note that since the transition matrix $\mathbf{P}$ corresponds to a jump diagram, then so does the process $(X_t)_{t\in T}$. So, it is common to interpret the jump diagram as the diagram imparting possible jumps of the process itself, reusing much of the same vocabulary in terms of the process. In particular, we will often say things as "the jump diagram for the process," to mean the jump diagram corresponding to the transition matrix of the process.

Further, we have been rather (needlessly) careful about using the notation $[\mathbf{P}^m]_{ij}$ when looking at the ijth entry of matrix $\mathbf{P}^m$. However, since we have shown that $[\mathbf{P}^m]_{ij}$ is the m-step transition probability $p_{ij}^{(m)}$ from i to j of the process $(X_t)_{t\in T}$ for which $\mathbf{P}$ is a transition matrix, we can interpret all previous results involving terms of the form $[\mathbf{P}^m]_{ij}$ as the corresponding multi-step transition probabilities of the process. In particular, we saw many manipulations that had a similar form as

$$[\mathbf{P}^m]_{ij}[\mathbf{P}^n]_{jk} \leq \sum_{l\in S}[\mathbf{P}^m]_{il}\,[\mathbf{P}^n]_{lk} = [\mathbf{P}^{m+n}]_{ik}$$

which in terms of the multi-step transition probabilities is

$$p_{ij}^{(m)}p_{jk}^{(n)} \leq \sum_{l\in S} p_{il}^{(m)}p_{lk}^{(n)} = p_{ik}^{(m+n)}.$$

As mentioned before, the last equality is a direct application of the Chapman-Kolmogorov equation. For intuition in terms of the process and multi-step transition probabilities, this manipulation says that the the probability of the process moving from state i to state j in m steps and then from state j to state k in the subsequent n steps is less than or equal to the probability that the process moves from state i to state k in $m+n$ steps (without regards to where it landed after the first m steps).

5.1. Hitting Times.

Definition 5.1. For a state $j \in S$, define the *first hitting time* of j by the process, denoted τ_j, as the random variable

$$\tau_j = \min\{m \geq 0 \ \text{ s.t. } \ X_{t_m} = j\}$$

where the values of m considered are non-negative integers and, by convention, define $\tau_j = \infty$ whenever there is no m such that $X_{t_m} = j$ (i.e., we are defining $\min \varnothing = \infty$). △

The value τ_j can be interpreted as the length of the path the process $(X_t)_{t\in T}$ follows until it first hits states j. Of course, if the process starts in state j, then the length of the path will be 0. The randomness of τ_j comes from the randomness of the process itself, since the paths that the process follows are not deterministic. To further this point, the state space of τ_j is $\{0, 1, 2, ..., \infty\}$, where for $m < \infty$, the event $\{\tau_j = m\}$ equivalent to the event that t_m is the first time X_{t_m} equals j; i.e.,

$$\{\tau_j = m\} = \bigcap_{k=0}^{m-1} \{X_{t_k} \neq j\} \cap \{X_{t_m} = j\}$$

and where, in the case $m = \infty$, the event $\{\tau_j = \infty\}$ is the event that the process never hits state j.

LEMMA 5.1. *There is an allowable path from state i to state j in the jump diagram for $(X_t)_{t\in T}$ if and only if $\mathbb{P}_i(\tau_j < \infty) > 0$. That is, $i \longrightarrow j$ if and only if there is a positive probability that the process starts from state i and hits state j in finitely many steps.*

PROOF. Suppose first that $i \longrightarrow j$, meaning that there is some finite integer $m \geq 0$ such that $p_{ij}^{(m)} > 0$. By Exercise 2, $\mathbb{P}_i(X_{t_m} = j) < \mathbb{P}_i(\tau_j < \infty)$, showing that $i \longrightarrow j$ implies that $\mathbb{P}_i(\tau_j < \infty) > 0$. On the other hand, suppose that $\mathbb{P}_i(\tau_j < \infty) > 0$. Since

$$\mathbb{P}_i(\tau_j < \infty) = \sum_{m=0}^{\infty} \mathbb{P}_i(\tau_j = m)$$

it must be that one of the terms $\mathbb{P}_i(\tau_j = m)$ is non-zero. By Exercise 2, $\mathbb{P}_i(\tau_j = m) < \mathbb{P}_i(X_{t_m} = j)$, showing that $p_{ij}^{(m)} > 0$, and hence implying that $i \longrightarrow j$. □

LEMMA 5.2. *Suppose that the state $j \in S$ belongs to an open communication class C with respect to the process $(X_t)_{t\in T}$ (i.e., with respect to its transition matrix $\mathbf{P}$). Then for any state $i \notin C$ such that $j \longrightarrow i$, it holds that $\mathbb{P}_i(\tau_j = \infty) = 1$. Intuitively, this says that once the process steps outside of an open communication class, it can never return.*

PROOF. First, notice that for $i \notin C$ with $j \longrightarrow i$, it holds that $p_{ij}^{(m)} = 0$ for every non-negative integer m. If this were not the case, then by definition $i \longrightarrow j$, from which we could conclude that $i \longleftrightarrow j$ (since we already assumed that $j \longrightarrow i$), contradicting that i and j are not in the same communication class. Since $\mathbb{P}_i(\tau_j = m) \leq p_{ij}^{(m)}$ and $p_{ij}^{(m)} = 0$ for every m, we have that

$$\mathbb{P}_i(\tau_j < \infty) \leq \sum_{m=0}^{\infty} \overbrace{p_{ij}^{(m)}}^{0} = 0.$$

This equivalently shows that $\mathbb{P}_i(\tau_j = \infty) = 1$. □

LEMMA 5.3. *Suppose that the state $j \in S$ belongs to an closed communication class C with respect to the process $(X_t)_{t\in T}$. Then for any state $i \notin C$, it holds that $\mathbb{P}_j(\tau_i = \infty) = 1$. Intuitively, this says that once the process enters a closed communication class, it can never leave.*

PROOF. By Lemma 3.1, $p_{ij}^{(m)} = 0$ for every non-negative integer m. Hence the result follows as in the proof of Lemma 5.2. □

Closely related to, but subtly different than the first hitting time is the first return time.

5.2. Return Times.

DEFINITION 5.2. For a state $j \in S$, the *first return time* to j by the process, denoted ρ_j, is the random variable defined by

$$\rho_j = \min\{m \geq 1 \text{ s.t. } X_{t_m} = j\}$$

where the values of m considered are the positive integers, and the standard convention applies, where if there is no such m, then $\rho_j = \infty$. △

The only difference is that instead of considering $m \geq 0$ in the minimum we used defining τ_j, we are considering $m \geq 1$ for ρ_j. In words, the first hitting time considers the evolution of the system from the outset, whereas the first return time only considers the evolution of the process after the first move. In particular, if the process does not start in state j, then τ_j and ρ_j will be equal since it will need to move at least once to hit j. However, it is most common to consider return times when the process starts in state j and we are curious about when the process next returns to state j (hence the name return time) – a curiosity that could not be addressed with hitting times since $\mathbb{P}_j(\tau_j = 0) = 1$.

LEMMA 5.4. *Suppose that with respect to the process $(X_t)_{t\in T}$, the state j belongs to an open communication class. Then $\mathbb{P}_j(\rho_j = \infty) > 0$. In particular, $\mathbb{E}_j[\rho_j] = \infty$.*

PROOF. If j belongs to an open communication class C, then there is some state $i \notin C$ such that $j \longrightarrow i$. In fact, we easily see that taking any allowable path of shortest length from j to i, say of length m, the path does not hit j after the first step. By this choice of m, notice that $p_{ji}^{(m)} > 0$ is the probability that the process follows a path from j to i of length m, and hence never returns to j along the way. By Lemma 5.2, once the process leaves the open communication class of j, it never returns (and thusly will never hit j afterward), we deduce that the probability that the

process starts at j and never returns is bounded below by the probability that the process follows a path of length m from j to i; i.e., $0 < p_{ji}^{(m)} \leq \mathbb{P}_j(\rho_i = \infty)$. □

EXAMPLE 5.1. Consider a stationary discrete time Markov chain $(X_t)_{t\in T}$ on the states $S = \{0, 1\}$ with transition matrix

$$\mathbf{P} = \begin{pmatrix} 1/3 & 2/3 \\ 2/3 & 1/3 \end{pmatrix}$$

We will calculate $\mathbb{E}_i[\tau_j]$ and $\mathbb{E}_i[\rho_j]$ for $i, j \in S$. △

5.3. Transience and Recurrence. For this section, we will assume that $(X_t)_{t\in T}$ is a stationary discrete time Markov chain on the states S with m-step transition probabilities $p_{ij}^{(m)}$ for $i, j \in S$.

DEFINITION 5.3. Given a state $j \in S$, we say that j is *transient* with respect to the process when $\mathbb{P}_j(\rho_j = \infty) > 0$. Intuitively, j is transient when there is a positive probability that starting from the state j, the process will never return to state j. If j is not transient, we call j *recurrent*; note that recurrence can be equivalently defined by $\mathbb{P}_j(\rho_j = \infty) = 0$. Intuitively, the state j is recurrent when starting from state j, the process will return to state j with probability 1. △

THEOREM 5.5. *For a state $j \in S$, the following are equivalent.*

(1) *j is transient.*
(2) *$\mathbb{P}_j(X_t = j$ for infinitely many times $t) = 0$.*
(3) *$\mathbb{P}_j(X_t = j$ for infinitely many times $t) < 1$.*
(4) $\sum_{m=0}^{\infty} p_{ii}^{(m)} < \infty$.

THEOREM 5.6. *For a state $j \in S$, the following are equivalent.*

(1) *j is recurrent.*
(2) *$\mathbb{P}_j(X_t = j$ for infinitely many times $t) = 1$.*
(3) *$\mathbb{P}_j(X_t = j$ for infinitely many times $t) > 0$.*
(4) $\sum_{m=0}^{\infty} p_{ii}^{(m)} = \infty$.

COROLLARY 5.7. *Suppose that $i, j \in S$ are states such that $i \longleftrightarrow j$. Then i and j are either both transient, or are both recurrent. That is, all elements within the same communication class are either all transient, or all recurrent.*

PROOF. We will show that if $i \longleftrightarrow j$, then i is transient if and only if j is transient; the result for recurrent then follows by the contrapositive. If $i \longleftrightarrow j$, then there are integers m_1

and m_2 such that $p_{ij}^{(m_1)} > 0$ and $p_{ji}^{(m_2)} > 0$. Suppose that i is transient. Then $\sum_{m=0}^{\infty} p_{ii}^{(m)} < \infty$, which certainly means that $\sum_{m=m_1+m_2}^{\infty} p_{ii}^{(m)} < \infty$. Arguing as we have several times in the past, the Chapman-Kolmogorov equations imply that $p_{ij}^{(m_1)} p_{jj}^{(k)} p_{ji}^{(m_2)} \leq p_{ii}^{(m_1+m_2+k)}$. We therefore have that

$$\sum_{k=0}^{\infty} p_{ij}^{(m_1)} p_{jj}^{(k)} p_{ji}^{(m_2)} \leq \sum_{m=m_1+m_2} p_{ii}^{(m)} < \infty$$

which, since $p_{ij}^{(m_1)} > 0$ and $p_{ji}^{(m_2)} > 0$, implies that $\sum_{k=0}^{\infty} p_{jj}^{(k)} < \infty$. Thus, if $i \longleftrightarrow j$ and i is recurrent, then so is j. By symmetry (swapping all is and js in the previous argument), we also deduce that if $i \longleftrightarrow j$ and j is recurrent, then so is i. □

DEFINITION 5.4. Since for a communication class C, every element is either transient, or every element is recurrent, we will call the communication class transient when all its elements are transient, or will call it recurrent when all its elements are recurrent. △

THEOREM 5.8. *Let $C \subset S$ be a communication class defined by the stationary discrete time Markov chain $(X_t)_{t \in T}$. The following hold.*

(1) *If C is open, then every element in C is transient. Equivalently, if C is recurrent, then C is closed.*
(2) *If C is closed and contains only finitely many elements, then C is recurrent.*

PROOF. (1) Suppose that C is an open communication class and $j \in C$. Then, as we have argued before, there is a positive probability that with respect to $\mathbb{P}_j$, the process only hits j at time t_0 and never again. Briefly, this argument was made by observing that there is an allowable path from j to some state outside of C which never hits j at any subsequent step, and once the process leaves the open communication class it does not come back. What this means is that $\mathbb{P}_j(X_t = j \text{ for only the time } t_0) > 0$, and hence $\mathbb{P}_j(X_t = j \text{ for infinitely many times } t) < 1$, proving that j is transient.

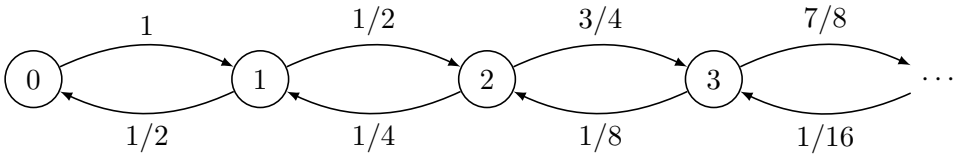
(2) If C is closed with only finitely many states, then by the Pigeon-hole principle, there must be some state $j \in C$ where $\mathbb{P}_j(X_t = j \text{ for infinitely many times } t) = 1$. Indeed, with respect to $\mathbb{P}_j$, the process starts in C, and since C is closed, all subsequent jumps of the process occur within C; since there are only finitely many states in C and the process makes infinitely many jumps, it must hit at least one of the states in C infinitely many times. This implies that j is recurrent, further implying that C is recurrent. □

We have now proved that every open communication class is transient; however, have only proved that closed communication classes with *finitely* many elements are recurrent. We can not extend this result to closed communication classes with arbitrarily many elements, since it can happen that, even though the communication class is closed, it is still transient; on the other hand, it can also happen that it is recurrent. To emphasize this, the next example is of a closed communication class in which all elements are transient.

EXAMPLE 5.2 (Closed and Transient). Let $S = \{0, 1, 2, ...\}$, the collection of all non-negative integers. From any state $k \in S$, the one step transition probabilities for a stationary discrete time Markov chain $(X_n)_{n=0}^{\infty}$ are given by

$$p_{kj} = \begin{cases} 1 & k = 0, j = 1 \\ 1 - 2^{-k} & 1 < j = k + 1 \\ 2^{-k} & j = k - 1 \\ 0 & \text{otherwise} \end{cases}$$

The jump diagram for this chain looks like,



This process is clearly irreducible, hence the single communication class (all of S) is closed. We now show that each state is transient. In fact, if $k \in S$, then

$$\begin{aligned} \mathbb{P}_k(\rho_k = \infty) &\geq \mathbb{P}(\text{starting from } k \text{ we only move to the right at each step}) \\ &= \prod_{n=k}^{\infty} (1 - 2^{-n}) > 0. \end{aligned}$$

The fact that the infinite product $\prod_{n=k}^{\infty}(1 - 2^{-n})$ is positive is from a standard result in analysis which says that for any sequence $\{a_n \in (0, 1) : n \geq k\}$, the product $\prod_{n=k}^{\infty}(1 - a_n) > 0$ if and only if $\sum_{n=k}^{\infty} a_n < \infty$. In this case, $a_n = 2^{-n}$, and the sum is just the familiar geometric series, which is finite. △

6. Exercises

(1) Suppose that $\mathbf{P}$ is a stochastic matrix. Explain why if $\mathbf{P}$ is irreducible, then the one communication class (induced by $\mathbf{P}$) is necessarily closed.

(2) Suppose that $(X_t)_{t\in T}$ is a stationary discrete time Markov chain with state space S, that $i, j \in S$, and τ_j is the first hitting time of j by the process.
   (a) Explain why for any time $t \in T$, $\mathbb{P}_i(X_t = j) < \mathbb{P}_i(\tau_j < \infty)$.
   (b) Explain why for any non-negative integer m, $\mathbb{P}_i(\tau_j = m) < \mathbb{P}_i(X_{t_m} = j)$.

CHAPTER 5

Stopping and Restarting a Stationary Discrete Time Markov Chain

Henceforth we will restrict our time variables to the non-negative integers, which as we have mentioned before, is not actually a theoretical restriction. As previously mentioned, the notation $(X_m)_{m=0}^{\infty}$ and $(X_t)_{t=0}^{\infty}$ are truly interchangeable since m and t are playing the roll of a dummy variable. That said, we give preference to the indexing variables m, k, and n, since we have used these variables in the past to indicate the number of jumps of the process, as opposed to using the variable t, which we have used mostly for indicating the time at which a jump of the process occurred; that is, we had used notation of the form t_m where m represented the number of jumps and t_m represented the time of jump m. The reason we choose to use this indexing-by-number-of-jumps convention here is that the correct way to "generalize" what follows to time variables which are not indexed by the integers is to consider the number of jumps of the process rather the physical time at which the jump took place.

1. Stopping Times

DEFINITION 1.1. Let $(X_m)_{m=0}^{\infty}$ be a stationary discrete time Markov chain with respect to the probability $\mathbb{P}$. A random variable τ taking values in $\{0, 1, 2, ..., \infty\}$ is called a *stopping time* with respect to this process when the event $\{\tau = m\}$ depends only on the first t steps of the process $X_0, X_1, X_2, ..., X_t$ (in particular, it does not depend on future states $X_{m+1}, X_{m+2}, ...$). If further $\mathbb{P}(\tau < \infty) = 1$, we call τ a *finite stopping time*. △

REMARK 1.1. A good intuition to equip ourselves with for a stopping time τ goes as follows: τ is a stopping time when it represents a strategy which can be used to decide when to stop the process without having to look into the future to make the decision. For example, if $(X_m)_{m=0}^{\infty}$ represents the evolution of our wealth during some gambling game (i.e., X_m is our wealth after round t of the game), then we could use the strategy that we will stop playing when our wealth first hits \$100. Certainly we do not need to peer into the future to use this strategy, and hence the first time τ at which our wealth hits \$100 is indeed a stopping time. △

EXAMPLE 1.1 (Non-Example). For the stationary discrete time Markov chain $(X_m)_{m=0}^{\infty}$ on the states S, consider the random variable κ_j defined by

$$\kappa_j = \max\{m \geq 0 \text{ s.t. } X_m \in A\}$$

where $A \subset S$ is some collection of states. In words, κ_j is the last time the process visits state j. Certainly κ_j takes a value in $\{0, 1, 2, ..., \infty\}$; however, κ_j is not a stopping time. Intuitively this is true since to use κ_j as a strategy to stop the process, we would need to know the future evolution of the process. More rigorously, consider the event $\{\kappa_j = t\}$ for any non-negative integer t. Then, we can reinterpret this event as

$$\{\kappa_j = m\} = \{X_m = j, X_{m+1} \neq j, X_{m+2} \neq j, X_{m+3} \neq j, ...\}$$

which depends on the future states $X_{m+1}, X_{m+2}, ...$, precluding the possibility that κ_j is a stopping time. △

EXAMPLE 1.2. Any constant, non-negative integer m is a stopping time with respect to any stationary discrete time Markov chain $(X_t)_{t=0}^{\infty}$. Intuitively this is clear since using the strategy that we will stop the process after m steps does not require us to peer into the future of the process to decide whether or not to stop. (E.g., choosing to stop gambling after playing 5 rounds does not require you to look into the future). A more rigorous argument is given in the exercises. △

1.1. Hitting and Return Times as Stopping Times. Within these notes, the most encountered stopping times will be hitting and return times. Before observing that hitting and return times are, in fact, stopping times, we give a slight generalization the notion of a hitting time.

DEFINITION 1.2. Let $(X_m)_{m=0}^{\infty}$ be a stationary discrete time Markov chain with state space S. For a non-empty collection of states $A \subset S$, the *first hitting time* of A is the random variable τ_A defined as

$$\tau_A = \min\{m \geq 0 \text{ s.t. } X_t \in A\}.$$

where values of m are taken over the non-negative integers. As before, if there is no such $m \geq 0$ such that $X_m \in A$, we define $\tau_A = \infty$. △

The intuition of this general notion of hitting time is consistent with our previous intuition, where the value τ_A represents the number of steps taken by the process until it first lands on any of the states in A. In the special case that $A = \{j\}$ for some state j, then τ_A is simply our previously introduced notion of a stopping time τ_j. Of course, we could similarly generalize the notion of a

return time to include return times to a collection of states A; however, this will not be helpful for us in the future, so we neglect doing so.

PROPOSITION 1.1. *Let $(X_m)_{m=0}^{\infty}$ be a stationary discrete time Markov chain on the states S. For any state $j \in S$ and any non-empty collection of states $A \subset S$, the first hitting time τ_A of A, and the first return time ρ_j to j are stopping times.*

PROOF. Exercise 1. □

2. The Strong Markov Property

After introducing stopping times, we have the opportunity to explore the Markov property of stationary discrete time Markov chains more deeply. For this section, we assume that, with respect to some probability $\mathbb{P}$, the process $(X_m)_{m=0}^{\infty}$ is a stationary discrete time Markov chain with transition matrix $\mathbf{P}$ and state space S.

THEOREM 2.1 (Strong Markov Property). *Let τ be a stopping time for the process $(X_m)_{m=0}^{\infty}$ and j be a state such that $\mathbb{P}(X_\tau = j) > 0$. With respect to the probability $\mathbb{P}(\,\cdot\,|\,\tau < \infty, X_\tau = j)$, the process $(X_{\tau+m})_{m=0}^{\infty}$ is again a stationary discrete time Markov chain with transition matrix $\mathbf{P}$ and initial distribution δ_j, and is independent of the past states $X_0, X_1, ..., X_{\tau-1}$. In particular, the process $(X_{\tau+m})_{m=0}^{\infty}$ with respect to $\mathbb{P}(\cdot\,|\,\tau < \infty, X_\tau = j)$ behaves as the original process $(X_m)_{m=0}^{\infty}$ with respect to $\mathbb{P}_j$.*

PROOF. Suppose that E is an event that depends only on the past states $X_0, X_1, ..., X_\tau$; and let F be any event depending only on the future states $(X_{\tau+m})_{m=0}^{\infty}$. Then, for any non-negative integer m, with respect to the probability $\mathbb{P}(\cdot\,|\,\tau = m, X_\tau = j)$ certainly E and F are independent and $(X_{\tau+m})_{m=0}^{\infty}$ again a stationary discrete time Markov chain with transition matrix $\mathbf{P}$ and initial distribution δ_j, since when $\tau = m$, we are back in the setting of Theorem 8.1 and Exercise (7) from Chapter 3. Now, consider the following

$$\begin{aligned}
&\mathbb{P}(\{X_\tau = j_0, X_{\tau+1} = j_1, ..., X_{\tau+n} = j_n\} \cap E\,|\,\tau = m, X_\tau = j) \\
&\quad = \mathbb{P}(X_\tau = j_0, X_{\tau+1} = j_1, ..., X_{\tau+n} = j_n\,|\,\tau = m, X_\tau = j)\,\mathbb{P}(E\,|\,\tau = m, X_\tau = j) \\
&\quad = \mathbb{P}(X_m = j_0, X_{m+1} = j_1, ..., X_{m+n} = j_n\,|\,X_m = j)\,\mathbb{P}(E\,|\,\tau = m, X_\tau = j) \\
&\quad = \mathbb{P}_j(X_0 = j_0, X_1 = j_1, ..., X_n = j_n)\,\mathbb{P}(E\,|\,\tau = m, X_\tau = j).
\end{aligned}$$

Therefore, since we are conditioning on $\{\tau < \infty\}$, the events $\{\tau = m\}$ for each $m = 0, 1, 2, ...$ partition our sample space with probability 1, and hence we get that

$$\begin{aligned}
&\mathbb{P}(\{X_\tau = j_0, X_{\tau+1} = j_1, ..., X_{\tau+n} = j_n\} \cap E \,|\, \tau < \infty, X_\tau = j) = \\
&= \mathbb{P}_j(X_0 = j_0, X_1 = j_1, ..., X_n = j_n) \sum_{m=0}^{\infty} \mathbb{P}(E \,|\, \tau = m, X_\tau = j)\mathbb{P}(\tau = m \,|\, T < \infty, X_\tau = j) \\
&= \mathbb{P}_j(X_0 = j_0, X_1 = j_1, ..., X_n = j_n)\, \mathbb{P}(E \,|\, \tau < \infty, X_\tau = j).
\end{aligned}$$

By setting $E = \Omega$, this proves that, with respect to $\mathbb{P}(\,\cdot \,|\, \tau < \infty, X_\tau = j)$, the process $(X_{\tau+t})_{t=0}^{\infty}$ is again a stationary discrete time Markov chain with transition matrix $\mathbf{P}$ and initial distribution δ_j; note that with $E = \Omega$ we have shown $\mathbb{P}_j(X_0 = j_0, X_1 = j_1, ..., X_n = j_n) = \mathbb{P}(X_\tau = j_0, X_{\tau+1} = j_1, ..., X_{\tau+n} = j_n \,|\, \tau < \infty, X_\tau = j)$. This also implies the independence of past events E and the future states $(X_{\tau+t})_{t=0}^{\infty}$ with respect to the probability $\mathbb{P}(\,\cdot \,|\, \tau < \infty, X_\tau = j)$. □

COROLLARY 2.2. *For any constant non-negative integer k, the process $(X_m)_{m=k}^{\infty} = (X_{k+m})_{m=0}^{\infty}$ is a stationary discrete time Markov chain with identical transition matrix $\mathbf{P}$ and initial distribution defined by $\nu(j) = \mathbb{P}(X_m = j)$. In particular, this means that $(X_m)_{m=k}^{\infty}$ with respect to $\mathbb{P}$ behaves as the original process $(X_m)_{m=0}^{\infty}$ with respect to $\mathbb{P}_\nu$. Moreover, given any state j such that $\mathbb{P}(X_m = j) > 0$, with respect to the probability $\mathbb{P}(\,\cdot \,|\, X_m = j)$, the process $(X_m)_{m=k}^{\infty}$ is independent of the past steps $X_0, ..., X_{k-1}$ and behaves like the original process $(X_m)_{m=0}^{\infty}$ with respect to $\mathbb{P}_j$.*

PROOF. This follows from the fact that a constant time k is a finite stopping time, and hence we can apply the Strong Markov Property in this case. □

3. First Step Analysis: Theory

Here we continue to assume that $(X_m)_{m=0}^{\infty}$ is a stationary discrete time Markov chain with respect to $\mathbf{P}$, with transition matrix $\mathbf{P}$, and with state space S. *First step analysis* is a paradigm which attempts to make tractable calculations involving the behavior of the process by considering the first step made by the process. In this section we will cover several results related to this paradigm with a few examples; in the following section, we give several more examples.

NOTATION 3.1. Consistent with the notation for probabilities, we will write $\mathbb{E}_j$ as the expected value with respect to the probability $\mathbb{P}_j$ for $j \in S$. That is, $\mathbb{E}_j$ is the expected value given that the process starts in state j. △

THEOREM 3.1. *Let $A \subset S$, $i \in S$, and $m \geq 1$. Then,*

$$\mathbb{P}_i(\tau_A = m) = \begin{cases} 0 & i \in A \\ \sum_{j \in S} p_{ij}\, \mathbb{P}_j(\tau_A = m-1) & i \notin A. \end{cases}$$

As a consequence,

$$\mathbb{E}_i[\tau_A] = \begin{cases} 0 & i \in A \\ 1 + \sum_{j \in S} p_{ij}\, \mathbb{E}_j[\tau_A] & i \notin A \end{cases}$$

PROOF. It is clear by definition that $\mathbb{P}_i(\tau_A = m) = 0$ if $i \in A$ and $m \geq 1$, since then $\tau_A = 0 \neq m$. If $i \notin A$, then we write

$$\mathbb{P}_i(\tau_A = m) = \sum_{j \in S} \mathbb{P}_i(\tau_A = m \,|\, X_1 = j) \underbrace{\mathbb{P}_i(X_1 = j)}_{p_{ij}} = \sum_{j \in S} p_{ij}\, \mathbb{P}_i(\tau_A = m \,|\, X_1 = j)$$

We now show that $\mathbb{P}_i(\tau_A = m \,|\, X_1 = j) = \mathbb{P}_j(\tau_A = m-1)$.

$$\begin{aligned} \mathbb{P}_i(\tau_A = m \,|\, X_1 = j) &= \mathbb{P}_i(X_m \in A, X_k \notin A \text{ for } 1 \leq k < m \,|\, X_1 = j) \\ &= \mathbb{P}_j(X_{m-1} \in A, X_k \notin A \text{ for } 1 \leq k < m-1) \\ &= \mathbb{P}_j(\tau_A = m-1) \end{aligned}$$

where the second equality comes from Corollary 2.2. Putting these pieces together gives us

$$\mathbb{P}_i(\tau_A = m) = \sum_{j \in S} p_{ij}\, \mathbb{P}_j(\tau_A = m-1).$$

which concludes the first claim.

For the second claim, if $i \in A$, then $\tau_A = 0$ and hence $\mathbb{E}_i[\tau_A] = 0$. Otherwise, if $i \notin A$,

$$\mathbb{E}_i[\tau_A] = \sum_{j \in S} \mathbb{E}_i[\tau_A \,|\, X_1 = j] \overbrace{\mathbb{P}_i(X_1 = j)}^{p_{ij}} = \sum_{j \in S} p_{ij}\, \mathbb{E}_i[\tau_A \,|\, X_1 = j].$$

We now show that $\mathbb{E}_i[\tau_A \,|\, X_1 = j] = 1 + \mathbb{E}_j[\tau_A]$.

$$\begin{aligned}
\mathbb{E}_i[\tau_A \,|\, X_1 = j] &= \sum_{m=1}^{\infty} m\, \mathbb{P}_i(\tau_A = m \,|\, X_1 = j) \\
&= \sum_{m=1}^{\infty} m\, \mathbb{P}_j(\tau_A = m-1) \\
&= \sum_{m=1}^{\infty} (m-1+1)\, \mathbb{P}_j(\tau_A = m-1) \\
&= \underbrace{\sum_{m=1}^{\infty} (m-1)\, \mathbb{P}_j(\tau_A = m-1)}_{\mathbb{E}_j[\tau_A]} + \underbrace{\sum_{m=1}^{\infty} \mathbb{P}_j(\tau_A = m-1)}_{1} \\
&= 1 + E_j[\tau_A]
\end{aligned}$$

Therefore,

$$\mathbb{E}_i[\tau_A] = \sum_{j \in S} p_{ij}(1 + \mathbb{E}_j[\tau_A]) = \sum_{j \in S} p_{ij} + \sum_{j \in S} p_{ij}\, \mathbb{E}_j[\tau_A] = 1 + \sum_{j \in S} p_{ij}\, \mathbb{E}_j[\tau_A].$$

which concludes the proof. □

The following corollary relies heavily on the notation used above, and introduced in Appendix A.2.

COROLLARY 3.2. *Let $A \subseteq S$ and τ_A the first hitting time of A by our process $(X_t)_{t=t_0}^{\infty}$. Define the column vector $\mathbf{E}$ indexed by S with ith entry $\mathbb{E}_i[\tau_A]$. Then,*

$$(\mathbf{I} - \mathbf{P}_{A^c})\mathbf{E}_{A^c} = \mathbf{1}$$

where we are indexing with $A^c = S \setminus A$, $\mathbf{I}$ is the identity matrix indexed by A^c, and $\mathbf{1}$ is the column vector indexed by A^c in which every entry is 1. In particular, if $\mathbf{I} - \mathbf{P}_{A^c}$ is invertible, then

$$\mathbf{E}_{A^c} = (\mathbf{I} - \mathbf{P}_{A^c})^{-1}\mathbf{1}.$$

If further ν is any initial distribution, then

$$\mathbb{E}_\nu[\tau_A] = \vec{\nu}_{A^c}(\mathbf{I} - \mathbf{P}_{A^c})^{-1}\mathbf{1} = \vec{\nu}_{A^c}\, \mathbf{E}_{A^c}$$

PROOF. Assume that $\#(S) = N$, $\#(A) = m$, and that $A^c = \{j_1, ..., j_{N-m}\}$. From Theorem 3.1 we have for each $j \in A^c$,

$$\begin{aligned}
&\mathbb{E}_j[\tau_A] = 1 + \sum_{k\in S} p_{jk}\, \mathbb{E}_k[\tau_A] \\
&\implies \mathbb{E}_j[\tau_A] - \sum_{j\in S} p_{j,k}\mathbb{E}_k[\tau_A] = 1 \\
&\implies (1-p_{j,j})\mathbb{E}_j[\tau_A] - \sum_{k\neq j} p_{j,k}\mathbb{E}_k[\tau_A] = 1.
\end{aligned}$$

Considering each of the $N-m$ values $\mathbb{E}_j[\tau_A]$ for $j \in A^c$ as unknowns of this linear equation, we have a system of $N-m$ equations with $N-m$ unknowns:

$$\begin{array}{rrcrl}
(1-p_{j_1,j_1})\mathbb{E}_{j_1}[\tau_A] & -p_{j_1,j_2}\mathbb{E}_{j_2}[\tau_A] & -\cdots & -p_{j_1,j_{N-m}}\mathbb{E}_{j_{N-m}}[\tau_A] & = 1 \\
-p_{j_2,j_1}\mathbb{E}_{j_1}[\tau_A] & +(1-p_{j_2,j_2})\mathbb{E}_{j_2}[\tau_A] & -\cdots & -p_{j_2,j_{N-m}}\mathbb{E}_{j_{N-m}}[\tau_A] & = 1 \\
\vdots & \vdots & & \vdots & = \vdots \\
-p_{j_{N-m},j_1}\mathbb{E}_{j_1}[\tau_A] & -p_{j_1,j_2}\mathbb{E}_{j_2}[\tau_A] & -\cdots & +(1-p_{j_{N-m},j_{N-m}})\mathbb{E}_{j_{N-m}}[\tau_A] & = 1
\end{array}$$

which is equivalent to the matrix equation,

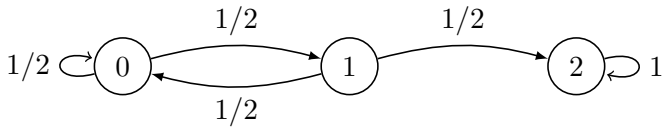
$$(\mathbf{I} - \mathbf{P}_{A^c})\mathbf{E}_{A^c} = \mathbf{1}$$

For the second claim, if $\vec{\nu}$ is the starting vector with $\nu_i = \mathbb{P}_\nu(X_{t_0} = i)$, then

$$\begin{aligned}
\mathbb{E}[\tau_A] &= \sum_{j\in S} \mathbb{P}(X_{t_0} = j)\mathbb{E}[\tau_A \,|\, X_{t_0} = j] = \sum_{j\in S} \nu_j\, \mathbb{E}_j[\tau_A] \\
&= \sum_{j\in A} \nu_j\, \mathbb{E}_j[\tau_A] + \sum_{j\in A^c} \nu_j\, \mathbb{E}_j[\tau_A] = \sum_{j\in A^c} \nu_j\, \mathbb{E}_j[\tau_A] = \vec{\nu}_{A^c}\, \mathbf{E}_{A^c}.
\end{aligned}$$

Here, for the second to last equality, we used the fact that if $j \in A$, then $\mathbb{E}_j[\tau_A] = 0$. The last equality follows by taking the vector dot product. □

EXAMPLE 3.1. Suppose we continually flip a fair coin until the first time a run of two heads in a row appears. We will analyze this game using a stationary discrete time Markov chain $(X_m)_{m=0}^{\infty}$ where X_m represents the number of heads in a row at the mth flip. Sensibly, the state space for this chain is $S = \{0, 1, 2\}$ with jump diagram



and transition matrix

$$\mathbf{P} = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 0 & 1 \end{pmatrix}$$

We will find the average number of flips it will take to achieve the run of 2 heads. For this, we let τ_2 be the first hitting time of state 2. In this case, we are considering $A = \{2\}$ and hence

$$\mathbf{P}_{A^c}: \quad \begin{array}{c|ccc} S & 0 & 1 & 2 \\ \hline 0 & 1/2 & 1/2 & 0 \\ 1 & 1/2 & 0 & 1/2 \\ 2 & 0 & 0 & 1 \end{array} \implies \mathbf{P}_{A^c} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 0 \end{pmatrix}$$

Therefore,

$$\mathbf{I} - \mathbf{P}_{A^c} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 0 \end{pmatrix} = \begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1 \end{pmatrix}.$$

Similarly

$$\mathbf{E}_{A^c}: \quad \begin{array}{c|c} S & \mathbf{E} \\ \hline 0 & \mathbb{E}_0[\tau_2] \\ 1 & \mathbb{E}_1[\tau_2] \\ 2 & \mathbb{E}_2[\tau_2] \end{array} \implies \mathbf{E}_{A^c}: \begin{pmatrix} \mathbb{E}_0[\tau_2] \\ \mathbb{E}_1[\tau_2] \end{pmatrix}$$

Therefore, we want to solve the matrix equation

$$\begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1 \end{pmatrix} \begin{pmatrix} \mathbb{E}_0[\tau_2] \\ \mathbb{E}_1[\tau_2] \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

We have

$$\begin{pmatrix} \mathbb{E}_0[\tau_2] \\ \mathbb{E}_1[\tau_2] \end{pmatrix} = \begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 4 \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 6 \\ 4 \end{pmatrix}$$

This shows that starting from state 0, the average number of steps it takes to arrive at state 2 is 6; whereas, starting from state 1, the average number of steps it takes to arrive at state 2 is 4. △

Continuing to work with theory, we next consider a fairly general result which falls under the umbrella of first step analysis. Along with the previous results of Theorem 3.1 and Corollary 3.2, notice that these results let us *linearize* the problem at hand; i.e., we create a system of linear equations to solve.

PROPOSITION 3.3. *Suppose that E is an event such that for every $i, j \in S$ with $p_{ij} > 0$ one of three possibilities occurs: $\mathbb{P}_i(E \mid X_1 = j) = 0$, $\mathbb{P}_i(E \mid X_1 = j) = 1$, or $\mathbb{P}_i(E \mid X_1 = j) = \mathbb{P}_j(E)$. Then for any state $i \in S$, we have that*

$$\mathbb{P}_i(E) = \sum_{j \in S_1(i)} p_{ij} + \sum_{j \in S_2(i)} p_{ij} \mathbb{P}_j(E)$$

where $S_1(i)$ are all the states $j \in S$ such that $\mathbb{P}_i(E \mid X_1 = j) = 1$ and $S_2(i)$ are all states $j \in S$ where $\mathbb{P}_i(E \mid X_1 = j) = \mathbb{P}_j(E)$. In particular, this implies we can convert calculating the values of $\mathbb{P}_i(E)$ into a matrix equation of the form

$$(\mathbf{I} - \mathbf{B})\mathbf{q} = \mathbf{b}$$

where $\mathbf{q}$ is the column vector indexed by S such that for each $i \in S$, the ith entry of $\mathbf{q}$ is $\mathbb{P}_i(E)$; $\mathbf{B}$ is the matrix indexed by S whose ijth entry is

$$[\mathbf{B}]_{ij} = \begin{cases} 0 & j \in S_1(i) \\ p_{ij} & j \in S_2(i) \end{cases}$$

and $\mathbf{b}$ is the column vector indexed by S whose ith entry is $\sum_{j \in S_1(i)} p_{ij}$.

PROOF. Let us note that if E is an event satisfying the assumptions of this proposition, then for any state $i \in S$, we have that the sets $S_1(i) = \{j \in S \text{ s.t. } \mathbb{P}_i(E \mid X_1 = j) = 1\}$, $S_2(i) = \{j \in S \text{ s.t. } \mathbb{P}_i(E \mid X_1 = j) = \mathbb{P}_j(E)\}$, and $Z(i) = \{j \in S \text{ s.t. } \mathbb{P}_i(E \mid X_1 = j) = 0\}$ partition the state space S. Therefore,

$$\begin{aligned} \mathbb{P}_i(E) = \sum_{j \in S} p_{ij}\, \mathbb{P}_i(E \mid X_1 = j) &= \sum_{j \in S_1(i)} p_{ij}\, \mathbb{P}_i(E \mid X_1 = j) + \sum_{j \in S_2(i)} p_{ij}\, \mathbb{P}_i(E \mid X_1 = j) \\ &\quad + \sum_{j \in Z(i)} p_{ij}\, \mathbb{P}_i(E \mid X_1 = j) \\ &= \sum_{j \in S_1(i)} p_{ij} \cdot 1 + \sum_{j \in S_2(i)} p_{ij}\, \mathbb{P}_j(E) + \sum_{j \in Z(i)} p_{ij} \cdot 0 \\ &= \sum_{j \in S_1(i)} p_{ij} + \sum_{j \in S_2(i)} p_{ij}\, \mathbb{P}_j(E) \end{aligned}$$

□

EXAMPLE 3.2. Consider the transition matrix of some Markov chain $(X_m)_{m=0}^{\infty}$ indexed by the state space $S = \{0, 1, 2\}$.

$$\mathbf{P} = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \end{pmatrix}$$

In this example we will find:

(1) $\mathbb{P}_i$(the chain eventually ends in state 2) for $i = 0, 1, 2$, and
(2) $\mathbb{P}_i$(when the chain eventually jumps to 2, it came from 1) for $i = 0, 1, 2$.

Solutions. (1) Start by noting

$$\mathbb{P}_i(\text{the chain eventually ends in state 2}) = \mathbb{P}_i(X_m = 2 \text{ for some } m).$$

Considering the first step, we have

$$\mathbb{P}_i(X_m = 2 \text{ for some } m) = \sum_{j \in S} p_{ij}\, \mathbb{P}_i(X_m = 2 \text{ for some } m \,|\, X_1 = j)$$

Let's make the observations

$$\mathbb{P}_2(X_m = 2 \text{ for some } m) = 1, \quad \text{and}$$

$$\mathbb{P}_i(X_m = 2 \text{ for some } m \,|\, X_1 = j) = \mathbb{P}_j(X_m = 2 \text{ for some } m) \quad \text{for } i = 0, 1 \text{ and } j \in S$$

This last equality happens since, if on the first step the chain moves to state j, then watching the evolution of the process from the first step and beyond is equivalent to watching the evolution of the process had it initially started in state j (this is an application of Corollary 2.2); in particular, the probability that the chain eventually hits state 2 starting from state i and given that it moves to state j in the first step is the same as the probability that the chain eventually hits state 2 starting from state j. For convenience, let $E = \{X_m = 2 \text{ for some } m\}$. We have

$$\begin{aligned} \mathbb{P}_0(E) &= p_{00}\, \mathbb{P}_0(E \,|\, X_1 = 0) + p_{01}\, \mathbb{P}_0(E \,|\, X_1 = 1) + p_{02}\, \mathbb{P}_0(E \,|\, X_1 = 2) \\ &= \frac{1}{2}\, \mathbb{P}_0(E) + \frac{1}{4}\, \mathbb{P}_1(E) + \frac{1}{4}\, \mathbb{P}_2(E) = \frac{1}{2}\, \mathbb{P}_0(E) + \frac{1}{4}\, \mathbb{P}_1(E) + \frac{1}{4} \end{aligned}$$

and

$$\begin{aligned} \mathbb{P}_1(E) &= p_{10}\, \mathbb{P}_1(E \,|\, X_1 = 0) + p_{11}\, \mathbb{P}_1(E \,|\, X_1 = 1) + p_{12}\, \mathbb{P}_1(E \,|\, X_1 = 2) \\ &= 0 + \frac{1}{2}\, \mathbb{P}_1(E) + \frac{1}{2}\, \mathbb{P}_2(E) = \frac{1}{2}\, \mathbb{P}_1(E) + \frac{1}{2}. \end{aligned}$$

We are left with the easily solvable system of equations,

$$\begin{aligned}\mathbb{P}_0(E) &= \frac{1}{2}\,\mathbb{P}_0(E) + \frac{1}{4}\,\mathbb{P}_1(E) + \frac{1}{4}\\ \mathbb{P}_1(E) &= \frac{1}{2}\,\mathbb{P}_1(E) + \frac{1}{2}\\ \mathbb{P}_2(E) &= 1\end{aligned}$$

Solving this easy system of equations leads to $\mathbb{P}_0(E) = 1$, $\mathbb{P}_1(E) = 1$, and $\mathbb{P}_2(E) = 1$. This shows that $\mathbb{P}_i(X_m = 2 \text{ for some } m) = 1$ regardless of which starting state i initiates our chain.

Note: We could have rearranged our system of equations as

$$\begin{aligned}\Big(1 - \frac{1}{2}\Big)\mathbb{P}_0(E) - \frac{1}{4}\,\mathbb{P}_1(E) + 0 \cdot \mathbb{P}_2(E) &= \frac{1}{4}\\ 0 \cdot \mathbb{P}_0(E) + \Big(1 - \frac{1}{2}\Big)\mathbb{P}_1(E) + 0 \cdot \mathbb{P}_2(E) &= \frac{1}{2}\\ 0 \cdot \mathbb{P}_0(E) + 0 \cdot \mathbb{P}_1(E) + \Big(1 - 0\Big)\mathbb{P}_2(E) &= 1\end{aligned}$$

resulting in the matrix equation

$$\left[\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix} - \begin{pmatrix}1/2 & 1/4 & 0\\ 0 & 1/2 & 0\\ 0 & 0 & 0\end{pmatrix}\right]\begin{pmatrix}\mathbb{P}_0(E)\\ \mathbb{P}_1(E)\\ \mathbb{P}_2(E)\end{pmatrix} = \begin{pmatrix}1/4\\ 1/2\\ 1\end{pmatrix}$$

of the form $(\mathbf{I} - \mathbf{B})\mathbf{q} = \mathbf{b}$, suggested by Proposition 3.3.

(2) For this, we notice that we're looking for

$$\begin{aligned}&\mathbb{P}_i(\text{when the chain eventually jumps to 2, it came from 1})\\ &\qquad = \mathbb{P}_i(X_{m+1} = 2 \text{ and } X_m = 1 \text{ for some } m).\end{aligned}$$

We again consider the first step

$$\begin{aligned}&\mathbb{P}_i(X_{m+1} = 2 \text{ and } X_m = 1 \text{ for some } m)\\ &= \sum_{j=0}^{2} p_{ij}\,\mathbb{P}_i(X_{m+1} = 2 \text{ and } X_m = 1 \text{ for some } m \,|\, X_1 = j)\end{aligned}$$

Let's make a few observations

$$\mathbb{P}_0(X_{m+1} = 2 \text{ and } X_m = 1 \text{ for some } m \,|\, X_1 = 2) = 0,$$

$$\mathbb{P}_1(X_{m+1} = 2 \text{ and } X_m = 1 \text{ for some } m \,|\, X_1 = 2) = 1,$$

$$\mathbb{P}_2(X_{m+1} = 2 \text{ and } X_m = 1 \text{ for some } m) = 0, \text{ and}$$

$$\mathbb{P}_i(X_{m+1} = 2 \text{ and } X_m = 1 \text{ for some } m \,|\, X_1 = j) = \mathbb{P}_j(X_{m+1} = 2 \text{ and } X_m = 1 \text{ for some } m)$$

where the last equality holds for $i = 0, 1$ and $j = 0, 1$ for similar reasons as discussed in part (1). For convenience, let $E = \{X_{m+1} = 2 \text{ and } X_m = 1 \text{ for some } m\}$. Then,

$$\begin{aligned}\mathbb{P}_0(E) &= p_{00}\,\mathbb{P}_0(E \,|\, X_1 = 0) + p_{01}\,\mathbb{P}_0(E \,|\, X_1 = 1) + p_{11}\,\mathbb{P}_0(E \,|\, X_1 = 2) \\ &= \frac{1}{2}\,\mathbb{P}_0(E) + \frac{1}{4}\,\mathbb{P}_1(E) + \frac{1}{4}\cdot 0 = \frac{1}{2}\,\mathbb{P}_0(E) + \frac{1}{4}\,\mathbb{P}_1(E)\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}_1(E) &= p_{10}\,\mathbb{P}_1(E \,|\, X_1 = 0) + p_{11}\,\mathbb{P}_1(E \,|\, X_1 = 1) + p_{12}\,\mathbb{P}(E \,|\, X_1 = 2) \\ &= 0 + \frac{1}{2}\,\mathbb{P}_1(E) + \frac{1}{2}\cdot 1 = \frac{1}{2}\mathbb{P}_1(E) + \frac{1}{2}.\end{aligned}$$

This leaves us with the easily solvable system of equations

$$\begin{aligned}\mathbb{P}_0(E) &= \frac{1}{2}\,\mathbb{P}_0(E) + \frac{1}{4}\,\mathbb{P}_1(E) \\ \mathbb{P}_1(E) &= \frac{1}{2}\mathbb{P}_1(E) + \frac{1}{2} \\ \mathbb{P}_2(E) &= 0\end{aligned}$$

resulting in the solutions $\mathbb{P}_2(E) = 0, \mathbb{P}_1(E) = 1$, and $\mathbb{P}_0(E) = \frac{1}{2}$. △

To conclude this section, we give a more general statement of first step analysis which, when appropriately interpreted, houses all previously introduced theories (i.e., our previous first step analysis results can be realized as corollaries to the following theorem). However, for our purposes, most (if not all) examples will be of the guise introduced in the first step linearization previously discussed in this section.

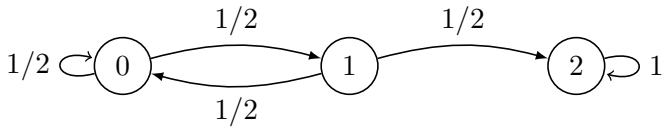
THEOREM 3.4 (First Step Analysis - General). *The equality*

$$\mathbb{E}_i[f(X_0, X_1, X_2, ...) \,|\, X_1 = j] = \mathbb{E}_j[f(i, X_0, X_1, ...)]$$

holds for any reasonably nice, real-valued function f.

4. First Step Analysis: More Examples

EXAMPLE 4.1. Here we revisit Example 3.1, but approach it in a way that feels more like first step analysis using Theorem 3.1. Recall that we are investigating a stationary discrete time Markov chain $(X_m)_{m=0}^{\infty}$ with jump diagram



and transition matrix

$$\mathbf{P} = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 0 & 1 \end{pmatrix}.$$

We will find $\mathbb{E}_i[\tau_2]$ for $i = 0, 1$, where τ_2 is the first hitting time of the state 2. Using Theorem 3.1, we have

$$\begin{aligned} \mathbb{E}_0[\tau_2] &= p_{00}(1 + \mathbb{E}_0[\tau_2]) + p_{01}(1 + \mathbb{E}_1[\tau_2]) + p_{02}(1 + \mathbb{E}_2[\tau_2]), \quad \text{and} \\ \mathbb{E}_1[\tau_2] &= p_{10}(1 + \mathbb{E}_0[\tau_2]) + p_{11}(1 + \mathbb{E}_1[\tau_2]) + p_{12}(1 + \mathbb{E}_2[\tau_2]) \end{aligned}$$

Since $\mathbb{E}_2[\tau_2] = 0$ (since on the event that $X_0 = 2$, the first hitting time of 2 is 0). With this and using the values for the one-step transition probabilities, we have

$$\begin{aligned} \mathbb{E}_0[\tau_2] &= \frac{1}{2}(1 + \mathbb{E}_0[\tau_2]) + \frac{1}{2}(1 + \mathbb{E}_1[\tau_2]) + 0\,(1 + 0) \\ &= \frac{1}{2}\mathbb{E}_0[\tau_2] + \frac{1}{2}\mathbb{E}_1[\tau_2] + 1 \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}_1[\tau_2] &= \frac{1}{2}(1 + \mathbb{E}_0[\tau_2]) + 0 \cdot (1 + \mathbb{E}_1[\tau_2]) + \frac{1}{2}(1 + 0) \\ &= \frac{1}{2}\mathbb{E}_0[\tau_2] + 1. \end{aligned}$$

This gives us the system of equations

$$\begin{aligned} \frac{1}{2}\mathbb{E}_0[\tau_2] - \frac{1}{2}\mathbb{E}_1[\tau_2] &= 1 \\ -\frac{1}{2}\mathbb{E}_0[\tau_2] + \mathbb{E}_1[\tau_2] &= 1 \end{aligned}$$

which solving out for $\mathbb{E}_0[\tau_2]$ and $\mathbb{E}_1[\tau_2]$ results in $\mathbb{E}_0[\tau_2] = 6$ and $\mathbb{E}_1[\tau_2] = 4$. Note that we could have converted this system of equations to the matrix equation

$$\begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1 \end{pmatrix} \begin{pmatrix} \mathbb{E}_0[\tau_2] \\ \mathbb{E}_1[\tau_2] \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

which is the same matrix equation we solved in Example 3.1. △

EXAMPLE 4.2. Consider a stationary discrete time Markov chain $(X_t)_{t=t_0}^{\infty}$ with transition matrix

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ .1 & .3 & .5 & .1 \\ .2 & .1 & .6 & .1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

indexed by the state space $S = \{-1, 0, 2, 4\}$. The *absorbing* states are $A = \{-1, 4\}$. In this example we do the following.

(1) Find $\mathbb{E}_i[\tau_A]$ for each $i \in S$, where τ_A is the first hitting time of A by the process.
(2) Find the probability that when the process is absorbed, it gets absorbed into state 4.
(3) Find the probability that the step before the process is absorbed, it was in state 2.

Solutions. (1) For $i \in A$, $\mathbb{E}_i[\tau_A] = 0$, so we need only work to find $\mathbb{E}_0[\tau_A]$ and $\mathbb{E}_2[\tau_A]$. Using the first step analysis of Theorem 3.1 we have

$$\begin{aligned} \mathbb{E}_0[\tau_A] &= p_{0,-1}\,(1 + \mathbb{E}_{-1}[\tau_A]) + p_{00}\,(1 + \mathbb{E}_0[\tau_A]) + p_{02}\,(1 + \mathbb{E}_2[\tau_A]) + p_{04}\,(1 + \mathbb{E}_4[\tau_A]) \\ &= (.1)(1+0) + (.3)(1 + \mathbb{E}_0[\tau_A])) + (.5)(1 + \mathbb{E}_2[\tau_A]) + (.1)(1+0) \\ &= 1 + .3\,\mathbb{E}_0[\tau_A] + .5\,\mathbb{E}_2[\tau_A] \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}_2[\tau_A] &= p_{2,-1}\,(1 + \mathbb{E}_{-1}[\tau_A]) + p_{20}\,(1 + \mathbb{E}_0[\tau_A]) + p_{22}\,(1 + \mathbb{E}_2[\tau_A]) + p_{24}\,(1 + \mathbb{E}_4[\tau_A]) \\ &= (.2)(1+0) + (.1)(1 + \mathbb{E}_0[\tau_A])) + (.6)(1 + \mathbb{E}_2[\tau_A]) + (.1)(1+0) \\ &= 1 + .1\,\mathbb{E}_0[\tau_A] + .6\,\mathbb{E}_2[\tau_A] \end{aligned}$$

This results in the system of equations

$$\begin{aligned} .7\,\mathbb{E}_0[\tau_A] - .5\,\mathbb{E}_2[\tau_A] &= 1 \\ -.1\,\mathbb{E}_0[\tau_A] + .4\,\mathbb{E}_2[\tau_A] &= 1 \end{aligned}$$

which has solutions $\mathbb{E}_0[\tau_A] = 90/23$ and $\mathbb{E}_2[\tau_A] = 80/23$. Alternatively, we could have used Corollary 3.2 and solved the matrix equation

$$\left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} .3 & .5 \\ .1 & .6 \end{pmatrix}\right] \begin{pmatrix} \mathbb{E}_0[\tau_A] \\ \mathbb{E}_2[\tau_A] \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

which would have resulted in the same solutions

$$\begin{pmatrix} \mathbb{E}_0[\tau_A] \\ \mathbb{E}_2[\tau_A] \end{pmatrix} = \begin{pmatrix} .7 & -.5 \\ -.1 & .4 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{.23} \begin{pmatrix} .4 & .5 \\ .1 & .7 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 90/23 \\ 80/23 \end{pmatrix}$$

In either case, we have

$$\mathbb{E}_{-1}[\tau_A] = 0,\ \mathbb{E}_0[\tau_A] = 90/23,\ \mathbb{E}_2[\tau_A] = 80/23,\ \mathbb{E}_4[\tau_A] = 0.$$

(2) We are looking for $\mathbb{P}_i(E)$ for each $i \in S$ where E is the event that the chain gets absorbed into state 4. Since 4 is an absorbing state, then $E = \{X_m = 4 \text{ for some } m\}$. Clearly $\mathbb{P}_4(E) = 1$ and $\mathbb{P}_{-1}(E) = 0$. Let's further notice that for $i = 0, 2$, we have $\mathbb{P}_i(E \,|\, X_1 = j) = \mathbb{P}_j(E)$; this is similar to the discussion in Example 3.2: if on the first step the chain moves to state j, then watching the evolution of the process from the first step and beyond is equivalent to watching the evolution of the process had it initially started in state j; in particular, the probability of chain eventually hits state 4 starting from state i given that it moves to state j in the first step, is the same as the probability that the chain eventually hits state 4 starting from state j. Therefore, using the first step analysis suggested by Proposition 3.3, we have

$$\begin{aligned} \mathbb{P}_0(E) &= p_{0,-1}\,\mathbb{P}_{-1}(E) + p_{00}\,\mathbb{P}_0(E) + p_{02}\,\mathbb{P}_2(E) + p_{04}\,\mathbb{P}_4(E) \\ &= .1 \cdot 0 + .3\,\mathbb{P}_0(E) + .5\,\mathbb{P}_2(E) + .1 \cdot 1 = .3\,\mathbb{P}_0(E) + .5\,\mathbb{P}_2(E) + .1 \end{aligned}$$

and

$$\begin{aligned} \mathbb{P}_2(E) &= p_{2,-1}\,\mathbb{P}_{-1}(E) + p_{20}\,\mathbb{P}_0(E) + p_{22}\,\mathbb{P}_2(E) + p_{24}\,\mathbb{P}_4(E) \\ &= .2 \cdot 0 + .1\,\mathbb{P}_0(E) + .6\,\mathbb{P}_2(E) + .1 \cdot 1 = .1\,\mathbb{P}_0(E) + .6\,\mathbb{P}_2(E) + .1 \end{aligned}$$

This leaves us with the system of equations

$$\begin{aligned} .7\,\mathbb{P}_0(E) - .5\,\mathbb{P}_4(E) &= .1 \\ -.1\,\mathbb{P}_0(E) + .4\,\mathbb{P}_4(E) &= .1 \end{aligned}$$

which admits solutions $\mathbb{P}_0(E) = 9/23$ and $\mathbb{P}_2(E) = 8/23$. We have thus found the solutions

$$\mathbb{P}_{-1}(E) = 0, \mathbb{P}_0(E) = 9/23, \mathbb{P}_2(E) = 8/23, \text{ and } \mathbb{P}_4(E) = 1.$$

(3) For each $i \in S$, we are looking for $\mathbb{P}_i(E)$ where E is the event that the step prior to the chain begin absorbed, the chain was in state 2. We realize that this event can be written as

$$E = \{X_{m+1} \in A \text{ and } X_m = 2 \text{ for some } m\}$$

Working similarly to the previous part, let's make a few observations:

$$\begin{aligned}
&\mathbb{P}_{-1}(E) = \mathbb{P}_4(E) = 0, \\
&\mathbb{P}_0(E \,|\, X_{t_0+1} = -1) = \mathbb{P}_0(E \,|\, X_{t_0+1} = 4) = 0, \\
&\mathbb{P}_2(E \,|\, X_{t_0+1} = 4) = \mathbb{P}_2(E \,|\, X_{t_0+1} = -1) = 1, \text{ and} \\
&\mathbb{P}_i(E \,|\, X_{t_0+1} = j) = \mathbb{P}_j(E) \text{ for all other } i, j \in S.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{P}_0(E) = p_{0,-1}\,\mathbb{P}_0(E \,|\, X_{t_0+1} = -1) + p_{00}\,\mathbb{P}_0(E \,|\, X_{t_0+1} = 0) \\
+ p_{02}\,\mathbb{P}_0(E \,|\, X_{t_0+1} = 2) + p_{04}\,\mathbb{P}_0(E \,|\, X_{t_0+1} = 4) \\
= .3\,\mathbb{P}_0(E) + .5\,\mathbb{P}_2(E)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{P}_2(E) = p_{2,-1}\,\mathbb{P}_2(E \,|\, X_{t_0+1} = -1) + p_{20}\,\mathbb{P}_2(E \,|\, X_{t_0+1} = 0) \\
+ p_{22}\,\mathbb{P}_2(E \,|\, X_{t_0+1} = 2) + p_{24}\,\mathbb{P}_2(E \,|\, X_{t_0+1} = 4) \\
= .3 + .1\,\mathbb{P}_0(E) + .6\,\mathbb{P}_2(E).
\end{aligned}$$

This results in the system of equations

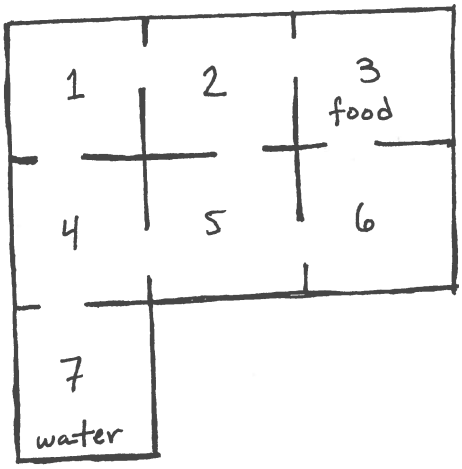
$$\begin{aligned}
.7\,\mathbb{P}_0(E) - .5\,\mathbb{P}_2(E) = 0 \\
-.1\,\mathbb{P}_0(E) + .4\,\mathbb{P}_2(E) = .3
\end{aligned}$$

with solutions $\mathbb{P}_0(E) = 15/23$ and $\mathbb{P}_2(E) = 21/23$. We have thus found the solutions

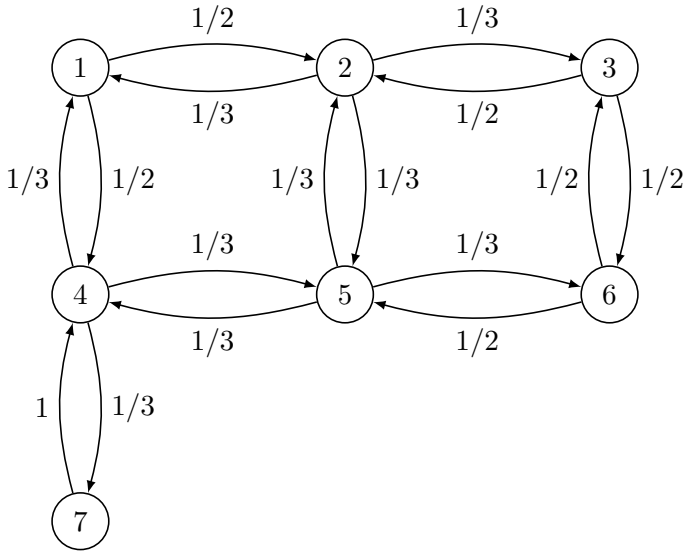
$$\mathbb{P}_{-1}(E) = 0, \mathbb{P}_0(E) = 15/23, \mathbb{P}_2(E) = 21/23, \text{ and } \mathbb{P}_4(E) = 0.$$

△

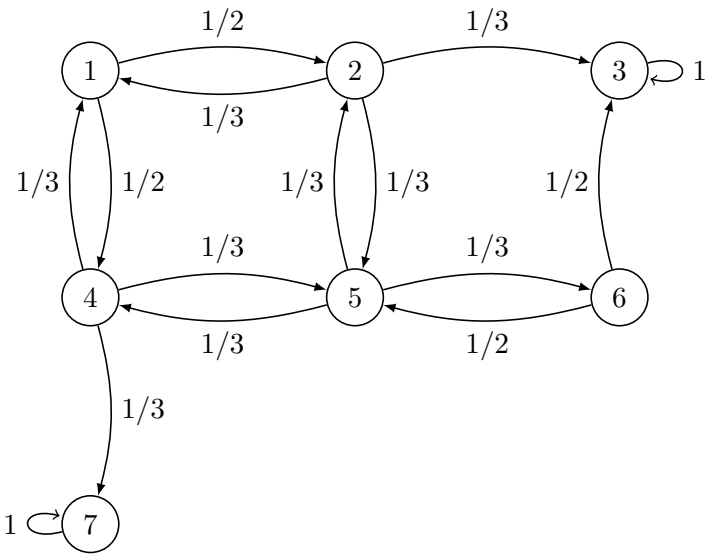
EXAMPLE 4.3. Consider a mouse in the following maze.



Suppose that the motion of the mouse can be described by a stationary discrete time Markov chain $(X_m)_{m=0}^{\infty}$ where X_m represents the room occupied by the mouse on its mth move. Assuming that the mouse is equally likely to pass through any door in whichever room it currently occupies, the jump diagram looks like,



Our goal in this example is to find the probability that, starting from room 1, what the probability is that the mouse gets the food before getting water. In Example 4.2, we were able to find the probability that the process hits one absorption state before another. At first, this question appears more difficult since neither states 3 nor 7 are absorption states. However, let's note that for this question, we do not care what moves the mouse makes after hitting either room 3 or 7; we only care the moves prior to hitting these states. Because of this, it is reasonable to consider the alternate process $(\tilde{X}_m)_{m=0}^{\infty}$ with jump diagram



which makes 3 and 7 into absorption states (leaving other jumps unchanged), since this will only affect the moves of the mouse after hitting either 3 or 7. With this alternate process, we can follow the method of the previous example and find $\mathbb{P}_1(\tilde{X}_m = 3 \text{ for some } m)$. As suggested by Proposition 3.3 and as we did in Examples 3.2 and 4.2, we will set up a system of equations for $\mathbb{P}_i(E)$ with $i \in S$ and $E = \{\tilde{X}_m = 3 \text{ for some } m\}$, and from this solve for $\mathbb{P}_1(E)$. Clearly, $\mathbb{P}_3(E) = 1$ and $\mathbb{P}_7(E) = 0$. From this, we find

$$\begin{aligned}
\mathbb{P}_1(E) &= \frac{1}{2}\,\mathbb{P}_2(E) + \frac{1}{2}\,\mathbb{P}_4(E), \\
\mathbb{P}_2(E) &= \frac{1}{3}\,\mathbb{P}_1(E) + \frac{1}{3}\,\mathbb{P}_5(E) + \frac{1}{3}, \\
\mathbb{P}_4(E) &= \frac{1}{3}\,\mathbb{P}_1(E) + \frac{1}{3}\,\mathbb{P}_5(E), \\
\mathbb{P}_5(E) &= \frac{1}{3}\,\mathbb{P}_4(E) + \frac{1}{3}\,\mathbb{P}_2(E) + \frac{1}{3}\,\mathbb{P}_6(E), \text{ and} \\
\mathbb{P}_6(E) &= \frac{1}{2}\,\mathbb{P}_5(E) + \frac{1}{2}.
\end{aligned}$$

This is a system of five linear equations with five unknowns, which is easily solved as the matrix equation

$$\begin{pmatrix}
1 & -1/2 & -1/2 & 0 & 0 \\
-1/3 & 1 & 0 & -1/3 & 0 \\
-1/3 & 0 & 1 & -1/3 & 0 \\
0 & -1/3 & -1/3 & 1 & -1/3 \\
0 & 0 & 0 & -1/2 & 1
\end{pmatrix}
\begin{pmatrix}
\mathbb{P}_1(E) \\ \mathbb{P}_2(E) \\ \mathbb{P}_4(E) \\ \mathbb{P}_5(E) \\ \mathbb{P}_6(E)
\end{pmatrix}
=
\begin{pmatrix}
0 \\ 1/3 \\ 0 \\ 0 \\ 1/2
\end{pmatrix}$$

with solution

$$\begin{pmatrix} \mathbb{P}_1(E) \\ \mathbb{P}_2(E) \\ \mathbb{P}_4(E) \\ \mathbb{P}_5(E) \\ \mathbb{P}_6(E) \end{pmatrix} = \begin{pmatrix} 1 & -1/2 & -1/2 & 0 & 0 \\ -1/3 & 1 & 0 & -1/3 & 0 \\ -1/3 & 0 & 1 & -1/3 & 0 \\ 0 & -1/3 & -1/3 & 1 & -1/3 \\ 0 & 0 & 0 & -1/2 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 1/3 \\ 0 \\ 0 \\ 1/2 \end{pmatrix} = \begin{pmatrix} 7/12 \\ 3/4 \\ 5/12 \\ 2/3 \\ 5/6 \end{pmatrix}$$

In particular, $\mathbb{P}_1(E) = 7/12$. △

EXAMPLE 4.4. In the previous examples, we have concerned ourselves with hitting times and probabilities where we considered the states we are hitting to be absorption states (even in Example 4.3 where there were no absorption states, we considered an alternate process making certain states into absorption states). In this example, we run through similar calculations without considering any states as absorption states and see that we arrive at the correct solutions without any added work (basically, we show that altering the process to include absorption states isn't necessary in many applications). Consider the transition matrix

$$\mathbf{P} = \begin{pmatrix} 0 & 1/4 & 1/4 & 1/2 \\ 1/2 & 0 & 0 & 1/2 \\ 1/3 & 1/3 & 0 & 1/3 \\ 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

of some stationary discrete time Markov chain $(X_m)_{m=0}^{\infty}$ with state space $S = \{0, 1, 2, 3\}$. In this example we will find

(1) For each $i \in S$, the expected number of steps the process takes before hitting either state 2 or state 3 assuming that the process starts from state i.
(2) For each $i \in S$, the probability that the process hits state 3 before hitting state 2 assuming that the process starts from state i.
(3) For each $i \in S$, the probability that the process was in state 2 the step before first hitting either state 1 or state 3.

Solutions. (1) Let $A = \{2, 3\}$ and τ_A be the first hitting time of A by the process. We are looking for $\mathbb{E}_i[\tau_A]$ for each $i \in S$. Let's agree that $\mathbb{E}_2[\tau_A] = \mathbb{E}_3[\tau_A] = 0$, so the only work we really

need to do is finding $\mathbb{E}_0[\tau_A]$ and $\mathbb{E}_1[\tau_A]$. We have

$$\begin{aligned}\mathbb{E}_0[\tau_A] &= \frac{1}{4}\,\mathbb{E}_0[\tau_A \,|\, X_1 = 1] + \frac{1}{4}\,\mathbb{E}_0[\tau_A \,|\, X_1 = 2] + \frac{1}{2}\,\mathbb{E}_0[\tau_A \,|\, X_1 = 3] \\ &= \frac{1}{4}\,(1 + \mathbb{E}_1[\tau_A]) + \frac{1}{4}\,(1 + \mathbb{E}_2[\tau_A]) + \frac{1}{2}\,(1 + \mathbb{E}_3[\tau_A]) \\ &= \frac{1}{4}\,\mathbb{E}_1[\tau_A] + 1.\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_1[\tau_A] &= \frac{1}{2}\,\mathbb{E}_1[\tau_A \,|\, X_1 = 0] + \frac{1}{2}\,\mathbb{E}_1[\tau_A \,|\, X_1 = 3] \\ &= \frac{1}{2}\,(1 + \mathbb{E}_0[\tau_A]) + \frac{1}{2}\,(1 + \mathbb{E}_3[\tau_A]) \\ &= \frac{1}{2}\,\mathbb{E}_0[\tau_A] + 1.\end{aligned}$$

This leaves us with the system of equations

$$\begin{aligned}\mathbb{E}_0[\tau_A] &= \frac{1}{4}\,\mathbb{E}_1[\tau_A] + 1 \\ \mathbb{E}_1[\tau_A] &= \frac{1}{2}\,\mathbb{E}_0[\tau_A] + 1\end{aligned}$$

which has solutions $\mathbb{E}_0[\tau_A] = 10/7$ and $\mathbb{E}_1[\tau_A] = 12/7$.

(2) Let E be the event that the process hits 3 before hitting 2. Notice that $\mathbb{P}_2(E) = 0$, $\mathbb{P}_3(E) = 1$, and $\mathbb{P}_i(E \,|\, X_1 = j) = \mathbb{P}_j(E)$ for any states $i = 0, 1$ and $j \in S$. So, we have left to find $\mathbb{P}_0(E)$ and $\mathbb{P}_1(E)$. We have

$$\begin{aligned}\mathbb{P}_0(E) &= \frac{1}{4}\,\mathbb{P}_0(E \,|\, X_1 = 1) + \frac{1}{4}\,\mathbb{P}_0(E \,|\, X_1 = 2) + \frac{1}{2}\,\mathbb{P}_0(E \,|\, X_1 = 3) \\ &= \frac{1}{4}\,\mathbb{P}_1(E) + \frac{1}{4}\,\mathbb{P}_2(E) + \frac{1}{2}\,\mathbb{P}_3(E) \\ &= \frac{1}{4}\,\mathbb{P}_1(E) + \frac{1}{2}\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}_1(E) &= \frac{1}{2}\,\mathbb{P}_1(E \,|\, X_1 = 0) + \frac{1}{2}\,\mathbb{P}_1(E \,|\, X_1 = 3) \\ &= \frac{1}{2}\,\mathbb{P}_0(E) + \frac{1}{2}\,\mathbb{P}_3(E) \\ &= \frac{1}{2}\,\mathbb{P}_0(E) + \frac{1}{2}\end{aligned}$$

This leaves us with the system of equations

$$\mathbb{P}_0(E) = \frac{1}{4}\,\mathbb{P}_1(E) + \frac{1}{2}$$
$$\mathbb{P}_1(E) = \frac{1}{2}\,\mathbb{P}_0(E) + \frac{1}{2}$$

which has solutions $\mathbb{P}_0(E) = 5/7$ and $\mathbb{P}_1(E) = 6/7$.

(3) Let E be the event that the process was in state 2 the step prior to first hitting states 1 or 3. Notice first that $\mathbb{P}_1(E) = \mathbb{P}_3(E) = 0$, so we only need to worry about calculating $\mathbb{P}_0(E)$ and $\mathbb{P}_2(E)$. Let us make the following first step observations:

$$\mathbb{P}_0(E \,|\, X_1 = 0) = \mathbb{P}_2(E \,|\, X_1 = 0) = \mathbb{P}_0(E),$$
$$\mathbb{P}_0(E \,|\, X_1 = 2) = \mathbb{P}_2(E),$$
$$\mathbb{P}_0(E \,|\, X_1 = 1) = \mathbb{P}_0(E \,|\, X_1 = 3) = 0,$$
$$\mathbb{P}_2(E \,|\, X_1 = 1) = \mathbb{P}_2(E \,|\, X_1 = 3) = 1.$$

Therefore, we have

$$\begin{aligned}\mathbb{P}_0(E) &= \frac{1}{4}\,\mathbb{P}_0(E \,|\, X_1 = 1) + \frac{1}{4}\,\mathbb{P}_0(E \,|\, X_1 = 2) + \frac{1}{2}\,\mathbb{P}_0(E \,|\, X_1 = 3)\\ &= \frac{1}{4}\,\mathbb{P}_2(E)\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}_2(E) &= \frac{1}{3}\,\mathbb{P}_2(E \,|\, X_1 = 0) + \frac{1}{3}\,\mathbb{P}_2(E \,|\, X_1 = 1) + \frac{1}{3}\,\mathbb{P}_2(E \,|\, X_1 = 3)\\ &= \frac{1}{3}\,\mathbb{P}_0(E) + \frac{2}{3}.\end{aligned}$$

This leaves us with the system of equations

$$\mathbb{P}_0 = \frac{1}{4}\,\mathbb{P}_2(E)$$
$$\mathbb{P}_2(E) = \frac{1}{3}\,\mathbb{P}_0(E) + \frac{2}{3}$$

which has solutions $\mathbb{P}_0(E) = 8/44$ and $\mathbb{P}_2(E) = 8/11$. △

5. Exercises

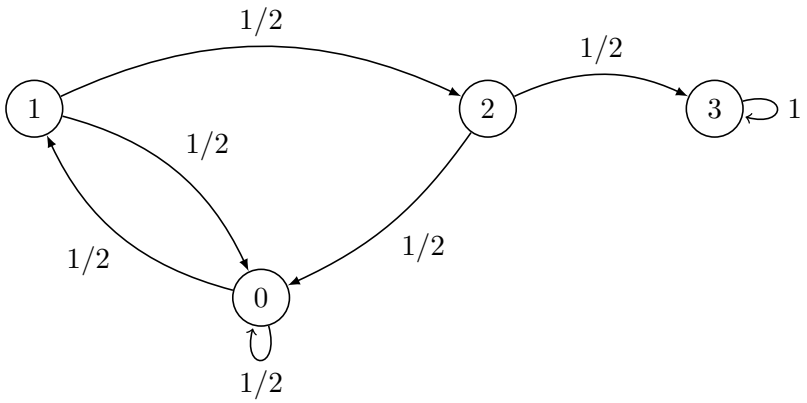
(1) Let $(X_m)_{m=0}^{\infty}$ be a stationary discrete time Markov chain on the states S. Also let $j \in S$ be any states and $A \subset S$ be any non-empty collection of states., the first hitting time τ_A of A, and the first return time ρ_j to j are stopping times.

(a) For any non-empty collection of states $A \subset S$, show that the first hitting time τ_A of A is a stopping time (with respect to the process).

(b) For any state $j \in S$, show that the first return time ρ_j of j is a stopping time (with respect to the process).

(2) (Motivated by Example 4.12 in the 11th edition of Ross' *Introduction to Probability Models*) You play a gambling game where you either win or lose during each round, and the result of winning or losing during each round is independent of the outcomes of the other rounds. You decide to play by the strategy where once you win three times in a row, you will quit playing; until you win three times in a row, you will keep playing. Let N be the number of rounds you play until you quit. In this exercise you will find the probability that you play at most 6 rounds; i.e., you will find $\mathbb{P}(N \leq 6)$. Assume that you are equally likely to win or lose during each round.

(a) Let $(X_n)_{n=0}^{\infty}$ be a stationary discrete time Markov chain where X_n tells you the current number of wins you have in a row at the nth round. For example, if on the $(n-2)$nd round you lost, but you won during the $(n-1)$st round and won during the nth round, then $X_n = 2$; whereas, regardless of the previous wins, if on the nth round you are still playing and lose, then $X_n = 0$ since your current run of wins has restarted. Explain why the jump diagram

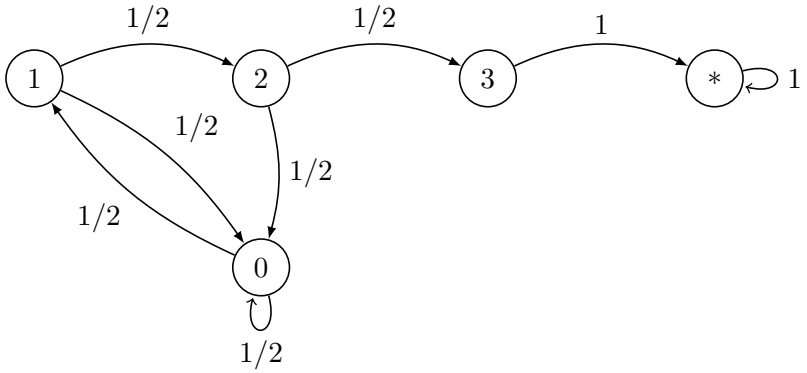


works well for our set up (in particular, why are we choosing to have state 3 "trap" the process once it is there?).

(b) Let $\mathbf{P}$ be the transition matrix corresponding to the jump diagram of the previous part. Using the sensible initial condition that $X_0 = 0$, explain why $\mathbb{P}(N \leq 6) = [\mathbf{P}^6]_{0,3}$.

(c) Find $\mathbb{P}(N \leq 6)$.

(d) Suppose instead you wanted to find $\mathbb{P}(N = 6)$. For this, consider a process $(X_n)_{n=0}^{\infty}$ defined analogously as part (a), but with jump diagram



Note that the state space here is $S = \{0, 1, 2, 3, *\}$ where the value of $*$ is unimportant (as long as it is not $0, 1, 2$ or 3). Again, assuming $X_0 = 0$, if $\mathbf{P}$ is the transition matrix for this process, explain why $\mathbb{P}(N = 6) = [\mathbf{P}^6]_{0,3}$. (Note: In reality, you could have used the same process as part (b) here if you wanted to, since $\mathbb{P}(N = 6) = \mathbb{P}(N \leq 6) - \mathbb{P}(N \leq 5)$; however, this part emphasizes a different method).

(e) Find $\mathbb{P}(N = 6)$.

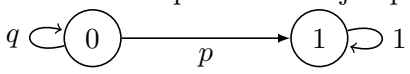
(3) Suppose that $p \in (0, 1)$ is the probability of a certain coin landing heads when flipped; hence $q = 1 - p$ is the probability of that coin landing tails. Let $(X_m)_{m=0}^{\infty}$ be a stationary discrete time Markov chain where X_m is the tallied number of heads which have occurred through the mth flip.

(a) Draw the jump diagram and give the transition matrix for this process.

(b) Next, show that the probability that a heads is eventually flipped is 1 in two ways:

(i) Finding $\mathbb{P}_0(X_m \geq 1 \text{ for some } m)$ using a first step analysis argument (similar to the arguments in Example 3.2).

(ii) By considering an alternate Markov process with jump diagram



and finding the average first hitting time of the state 1 starting from 0. Why does this method also give you the correct answer?

(c) Continuing to consider the process $(X_m)_{m=0}^{\infty}$ whose transition matrix $\mathbf{P}$ you found in (a), show that for any $m \geq 1$ and any $k = 1, ..., m$, it holds that $[\mathbf{P}^m]_{0,k} = \binom{m}{k} p^k q^{n-k}$. *Hint*: You know that with respect to $\mathbb{P}_0$, you have $X_m \stackrel{d}{=} \text{Bin}(m, p)$.

(4) Let $(X_m)_{m=0}^{\infty}$ be a stationary discrete time Markov chain with state space $S = \{1, 2, 3, 4\}$ and transition matrix

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 1/8 & 1/4 & 1/8 \\ 1/4 & 1/2 & 1/8 & 1/8 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Let $A = \{1, 4\}$ be the *absorption* states. *Hint.* Before tackling the parts below, you may want to look at Examples 3.1, 4.1, 3.2, and 4.2 for inspiration.

(a) For each $i \in S$, given that the process starts in state i, find the average time until *absorption*. That is, find $\mathbb{E}_i[\tau_A]$ for each state i.

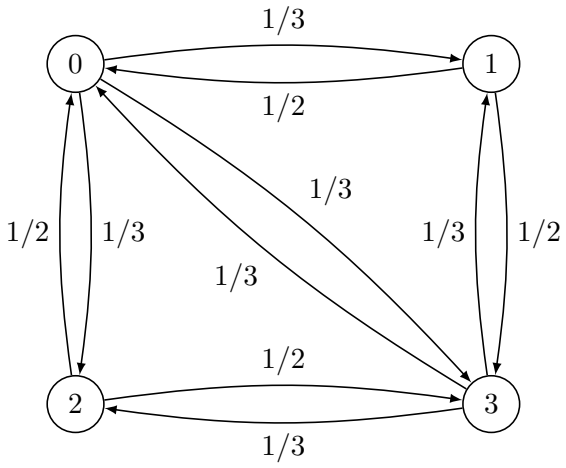
(b) For each $i \in S$, find the probability that, if the process starts in state i, it will be *absorbed*. That is, for each $i \in S$, find $\mathbb{P}_i(E)$ where E is the event $E = \{X_m \in A \text{ for some } m\}$.

(c) For each $i \in S$, find the probability that, if the process starts in state i, it will land in state 4 before landing in state 1.

(d) For each $i \in S$, find the probability that, if the process starts in state i, the last state the it was in before being absorbed was 3.

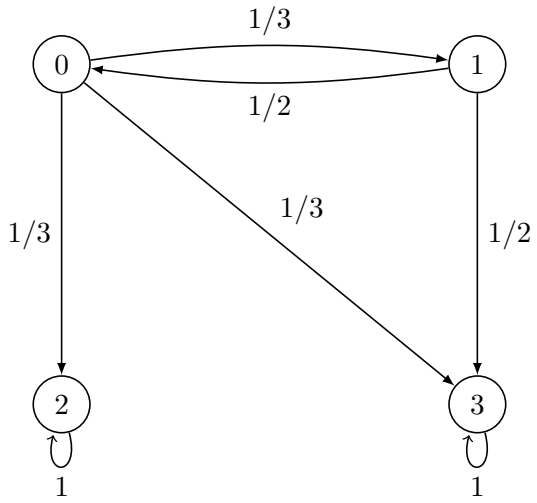
(e) If the initial distribution ν that corresponds to the starting vector for this process is $\vec{\nu} = (0, 1/4, 3/4, 0)$, find $\mathbb{E}_\nu[X_3]$. How would this answer change if we assumed our state space was $S = \{-1, 2, 5, 7\}$? How would this answer change if $\mathbf{P}$ given above was actually indexed by the ordering $S = \{2, 3, 1, 4\}$?

(5) Let $(X_m)_{m=0}^{\infty}$ be a stationary discrete time Markov chain with jump diagram



(a) Starting from state 0, find the expected time it takes the process to hit state 2 or 3, and find the probability that the process will hit state 3 before hitting state 2.

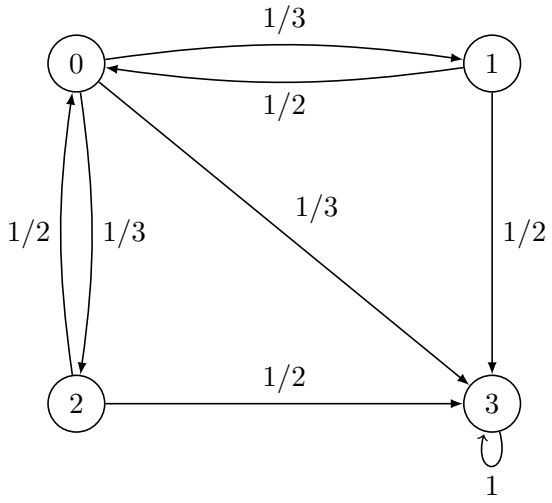
Hint: If you appreciate absorption states, you can consider answering this question for the alternate process with jump diagram



and justify why this will give you the same answer. Look at Example 4.3 for similar arguments. However, you can also work similarly to Example 4.4 without considering an alternate process.

(b) Starting from state 0, what is the probability that the process first hits state 3 coming from state 0?

Hint: If you appreciate absorption states, you can consider answering this question for the alternate process with jump diagram



and justify why this will give you the same answer. Look at Example 4.2 for inspiration. However, you can also work similarly to Example 4.4 without considering an alternate process.

(6) You continually flip a fair coin until you get three heads in a row, then you stop. Find the expected number of flips until you stop in two ways.
   (a) Using Corollary 3.2 as in Example 3.1.
   (b) Using Theorem 3.1 and first step analysis as in Example 4.1.

(7) You continually flip a fair coin until two of the three most recent flips are heads, then you stop flipping. Find the expected number of flips necessary until you stop flipping.
Hint: There is a probability of $1/4$ that the first two flips both are heads, and you're done after two flips. Otherwise, consider a Markov chain $(X_m)_{m=3}^{\infty}$ (note the starting number of flips is $n = 3$) with state space

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

Consider THH and HTH to be absorbing states and find $\mathbb{E}_i[\tau_A]$ for each $i \in S$, where $A = \{THH, HTH\}$. Then, explain why the answer you're looking for is

$$\text{Expected number of flips} = \frac{1}{4} \cdot 2 + \frac{3}{4} \cdot \Big[3 + \frac{1}{6} \sum_{i \in S} \mathbb{E}_i[\tau_A]\Big].$$

(8) Let $(X_m)_{m=0}^{\infty}$ be a stationary discrete time Markov chain on the states S. Let $A \subset S$ and τ_A be the first hitting time of A by the process.
   (a) Interpret τ_A as a function of the process $\tau_A = f(X_0, X_1, X_2, ...)$. That is, describe the function f such that $\tau_A = f(X_0, X_1, X_2, ...)$. *Hint*: You are implicitly told what f is in the definition of the hitting time τ_A.
   (b) Continue with the function f you found in the previous part. For a state $i \in S$, explain why

$$f(i, X_0, X_1, ...) = \begin{cases} 0 & i \in A \\ 1 + f(X_0, X_1, X_2, ...) & i \notin A \end{cases}$$

   (c) Use the previous part and Theorem 3.4 to re-derive the equation

$$\mathbb{E}_i[\tau_A] = \begin{cases} 0 & i \in A \\ 1 + \sum_{j \in S} p_{ij} \mathbb{E}_j[\tau_A] & i \notin A \end{cases}$$

(9) Let $(X_m)_{m=0}^{\infty}$ be a stationary discrete time Markov chain on the states S, and assume $j \in S$. Let ρ_j and τ_j be the first return time and first hitting time of state j, respectively. For a state $i \in S$, derive the equation

$$\mathbb{E}_i[\rho_j] = \begin{cases} \mathbb{E}_i[\tau_j] & i \neq j \\ 1 + \sum_{k \in S} p_{ik} \, \mathbb{E}_k[\tau_j] & i = j \end{cases}$$

CHAPTER 6

Invariant and Long Run Distributions

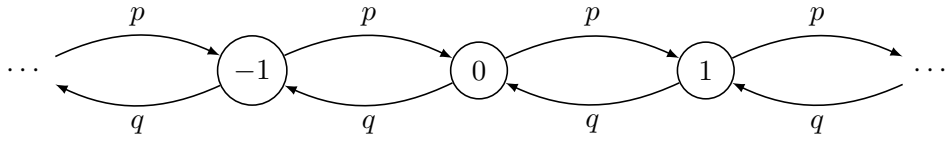
Throughout this chapter we will be studying behaviors of a stationary discrete time Markov chain $(X_m)_{m=0}^{\infty}$ on the states S with transition matrix $\mathbf{P}$. In particular, we will be concerned with long run behavior. For a large portion of the chapter, we will assume that $\mathbf{P}$ is irreducible; i.e., that all states communicate. When we are using irreducibility, we will make sure to explicitly mention it. Nearer the end of the chapter, we will examine some of what there is to say when the irreducible assumption does not hold.

1. Types of Recurrence – Positive vs Null

We have already defined what it means for a state $j \in S$ to be transient, when $\mathbb{P}_j(\rho_j = \infty) > 0$, or recurrent, when $\mathbb{P}_j(\rho_j = \infty) = 0$. From here, there is also an important division among recurrent states: positive recurrent versus null recurrent.

Definition 1.1. Suppose that $j \in S$ is recurrent, and let ρ_j be the first return time to j. If $\mathbb{E}_j[\rho_j] = \infty$, then we will say that j is *null recurrent*; otherwise, if $\mathbb{E}_j[\rho_j] < \infty$, then we say that j is *positive recurrent*. △

1.1. The $1-D$ Random Walk on the Integers. . Let $(X_n)_{n=0}^{\infty}$ be a (possibly biased) *random walk* on the integers. The state space for this is $S = \mathbb{Z}$, the integers. For some $p \in (0,1)$ and $q = 1-p$, the jump diagram for this chain is given by:



We will prove the following theorem.

Theorem 1.1. *The $1-D$ random walk is transient if $p \neq q$. Otherwise, if $p = q = 1/2$, then it is null recurrent.*

We break the proof up into two lemmas.

Lemma 1.2. *The 1-D random walk is recurrent if and only if $p = q = 1/2$. Otherwise, it is transient.*

PROOF. For this we will use Theorems 5.5 and 5.6. Since $\mathbf{P}$ is irreducible, we need only check for one of the states i. Starting from state i, it will take the process an even number of steps to return to state i; so, suppose the process starts in state i and moves $2k$ times. An allowable path from state i will return to state i in $2k$ steps if and only if half the steps were to the right and half to the left. Therefore, there are $\binom{2k}{k}$ total paths from i returning to i. There was a probability p to take each step to the right, and a probability q of taking each step to the left. Therefore, given that we started in state i, the probability we return to state i in $2k$ steps is $\binom{2k}{k}p^k q^k$. In other words, $p_{ii}^{(2k)} = \binom{2k}{k}p^k q^k$.

Using Stirling's approximation,

$$\binom{2k}{k} = \frac{(2k)!}{(k!)^2} \approx \frac{\sqrt{4\pi k}\left(\frac{2k}{e}\right)^{2k}}{2\pi k\left(\frac{k}{e}\right)^{2k}} = \frac{C}{\sqrt{k}}4^k$$

where C is some constant which does not depend on k. Hence,

$$p_{ii}^{(2k)} = \binom{2k}{k}p^k q^k \approx \frac{C}{\sqrt{k}}(4pq)^k.$$

Now, let's see how $(4pq)^k$ behaves. Recall that $q = 1 - p$, so we really want to analyze $f(p)^k$ where $f(p) = 4p(1-p)$. Using elementary calculus, you find that on the interval $0 < p < 1$, the maximum value of $f(p)$ happens when $p = 1/2$, in which case $f(1/2) = 1$. Otherwise, for $p \neq 1/2$, $0 < f(p) < 1$. Therefore,

$$\sum_{k=0}^{\infty} p_{ii}^{(2k)} \sim C\sum_{k=0}^{\infty}\frac{f(p)^k}{\sqrt{k}} = \begin{cases} \infty & \text{for } p = 1/2 \quad (\text{i.e., } f(p) = 1) \\ < \infty & \text{for } p \neq 1/2 \quad (\text{i.e., } 0 < f(p) < 1) \end{cases}$$

This shows us that each state is recurrent if and only if $p = 1/2$. Otherwise, each state is transient. □

LEMMA 1.3. *If $p = q = 1/2$ then the 1-D random walk is null-recurrent.*

PROOF. For $n \geq 0$, define the sequence $a_n = \mathbb{E}_n[\tau_0]$, where τ_0 is the first hitting time of state 0 (note that $a_0 = 0$). Using first step analysis we get the recursive relationship for $n \neq 0$: $a_n = 1 + \frac{1}{2}a_{n-1} + \frac{1}{2}a_{n+1}$. Rearranging this equation we get $\frac{1}{2}(a_{n+1} - a_n) = \frac{1}{2}(a_n - a_{n-1}) - 1$, or equivalently,

$$a_{n+1} - a_n = (a_n - a_{n-1}) - 2.$$

Writing out the first few terms,

$$\begin{aligned} a_2 - a_1 &= (a_1 - a_0) - 2 = a_1 - 2 \\ a_3 - a_2 &= (a_2 - a_1) - 2 = a_1 - 2 \cdot 2 \\ a_4 - a_3 &= (a_3 - a_2) - 2 = a_1 - 2 \cdot 3 \\ &\vdots \\ a_n - a_{n-1} &= a_1 - 2(n-1) \end{aligned}$$

For $n \geq 1$, summing these pieces together to create a telescoping series yields,

$$a_n = \sum_{k=0}^{n-1} (a_{k+1} - a_k) = a_1 + \sum_{k=1}^{n-1} (a_1 - 2k) = na_1 - n(n-1).$$

Rearranging this and using that $a_n = \mathbb{E}_n[\tau_0] \geq 0$,

$$a_1 = \frac{a_n}{n} + \frac{n(n-1)}{n} \geq \frac{n(n-1)}{n} = n - 1.$$

Since $n \geq 1$ was arbitrary, it must remain true as $n \to \infty$. The right hand side of this equation goes to ∞ as $n \to \infty$, thus it must be that $a_1 = \infty$. That is, $\mathbb{E}_1[\tau_0] = \infty$. Letting ρ_0 be the first return time to 0, we have $\mathbb{E}_0[\rho_0] = 1 + \frac{1}{2}(\underbrace{\mathbb{E}_1[\tau_0]}_{=\infty} + \underbrace{\mathbb{E}_{-1}[\tau_0]}_{\geq 0}) = \infty$. This shows that 0 is null recurrent. However, from the symmetry of the random walk about any state, it is clear that every state must be null recurrent. □

2. Invariant Distributions and the Function π

Definition 2.1. Let $\nu : S \to \mathbb{R}$ be a probability mass function on S with corresponding row vector for $\vec{\nu}$. We will call ν (or its vector form $\vec{\nu}$) an *invariant distribution* for $\mathbf{P}$ when $\vec{\nu}\mathbf{P} = \vec{\nu}$. That is, ν is an invariant distribution for $\mathbf{P}$ when $\vec{\nu}$ is a left eigenvector of $\mathbf{P}$ with eigenvalue 1. Let us also note here that is common to call ν an invariant distribution for the process $(X_m)_{m=0}^{\infty}$, meaning that it is invariant for its transition matrix $\mathbf{P}$. △

Remark 2.1. Note that the equality $\vec{\nu}\mathbf{P} = \vec{\nu}$ can be rewritten as

$$\nu_j = \sum_{i \in S} \nu_i \, p_{ij}$$

for each $j \in S$. This is simply writing out the matrix multiplication explicitly for each entry in the vector $\vec{\nu}\mathbf{P}$. △

There are two main theorems for this section, Theorems 2.1 and 2.4. The first exposes the existence and uniqueness of invariant distributions for an irreducible transition matrix, and gives a probabilistic interpretation. Before jumping into these theorems, we introduce notation which will be useful.

NOTATION 2.1. We define the function $\pi : S \to \mathbb{R}$ as $\pi(j) = 1/\mathbb{E}_j[\rho_j]$, where as usual ρ_j is the first return time to j. In the case $\mathbb{E}_j[\rho_j] = \infty$, we interpret this to mean $\pi(j) = 0$. As is often the case, we will write π_j instead of $\pi(j)$ (interpreting π in its vectorized form $\vec{\pi}$ indexed by S). △

THEOREM 2.1. *Suppose that* $\mathbf{P}$ *is irreducible and recurrent. The following hold.*

1. *Either* $\pi(j) = 0$ *for every* $j \in S$*, or* $\pi(j) > 0$ *for every* $j \in S$*. In other words, this says that every state is either null recurrent, or every state is positive recurrent.*
2. *(Invariance of* π*) Every state is positive recurrent if and only if* π *is a probability mass function, in which case* π *is an invariant distribution for* $\mathbf{P}$*.*
3. *(Uniqueness of* π*) At most one invariant distribution for* $\mathbf{P}$ *exists, and if one does exist, it is* π*.*

COROLLARY 2.2. *Suppose that* $C \subset S$ *is a communication class which is recurrent. Then all states in* C *are either positive recurrent or all states are null recurrent. Moreover, if* C *has only finitely many elements, then every state is positive recurrent.*

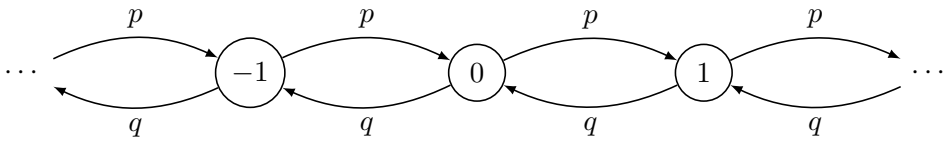
PROOF. If C is recurrent, then we have proved that it must be closed. Therefore the reduced transition matrix $\mathbf{P}_C$ is irreducible (since C is closed) and recurrent by assumption. We now apply Theorem 2.1 to $\mathbf{P}_C$ to conclude that all states in C are positive recurrent or all null recurrent. For the next part of the claim, we can argue that if C has only finitely many elements, then $\mathbb{E}_j[\rho_j] < \infty$ for at least one of the states $j \in C$ (which would then imply it to be true for all states by what we have just observed). For this argument, we could consider to the contrary that $\mathbb{E}_j[\rho_j] = \infty$ for every $j \in C$, which would result in a contradiction since there are only finitely many states for the process to jump between, so it would be impossible for each of these states to have an infinite average return time. However, for the reader's interest, we offer a deeper and alternate reference for proof: the Perron-Frobenious theorem. It turns out that an immediate corollary to Perron-Frobenious is that a finite dimensional stochastic matrix which is irreducible and recurrent must have an invariant distribution – applying this fact to $\mathbf{P}_C$ finishes the proof. □

NOTATION 2.2. Since every element recurrent communication class are either all positive recurrent or null recurrent, we often refer to the communication class itself as positive recurrent or null

recurrent. Further, if a the transition matrix is irreducible and recurrent, we will often say that it is (or its defining process is) positive recurrent or null recurrent, referring to the states which must all share these properties. △

REMARK 2.2. Amongst other implications, the Perron-Frobenius theorem tells us that if $\mathbf{P}$ is a finite dimensional, irreducible, and recurrent stochastic matrix, then the dimension of the left eigenspace of $\mathbf{P}$ corresponding to eigenvalue 1 is one-dimensional, and that this space is spanned by a (left-)eigenvector with all positive entries $\vec{\nu}$. Then, because of the eigenspace has only one dimension, every vector in this space must be of the form $c\,\vec{\nu}$ for a scalar $c \in \mathbb{R}$. Hence, choosing $c = \sum_{j \in S} \nu_j$, we find the unique invariant distribution $\pi = c\,\vec{\nu}$. △

EXAMPLE 2.1. Let us return to the one-dimensional random walk on the integers.



We have seen that if $p = q = 1/2$, then the random walk is recurrent and we gave a clever argument proving that it was null recurrent. Here is another (less clever) proof of the null recurrence. Using Theorem 2.1, if $\mathbf{P}$ is positive recurrent, then there is a unique invariant distribution π such that $\pi(i) = 1/\mathbb{E}_i[\rho_i] > 0$ for every $i \in \mathbb{Z} = S$. By the symmetry of the random walk, it is clear that $\mathbb{E}_i[\rho_i] = \mathbb{E}_j[\rho_j]$ for any states $i, j \in \mathbb{Z}$, which implies that $\pi(i) = \pi(j)$ for any integers i, j. Since we are assuming $\mathbf{P}$ is positive recurrent, π must be a probability mass function, and hence $\sum_{i \in \mathbb{Z}} \pi(i) = 1$. However, any fixed positive number summed together infinitely many times results in an infinite value, contradicting either that π is a probability mass function, or that $\pi(i) = \pi(j) > 0$ for every i, j. In light of this contradiction, Theorem 2.1 ensures that this process can not be positive recurrent and is thusly null recurrent. △

Comparing Theorem 2.1 with the fact that $\mathbb{E}_j[\rho_j] = \infty$ in the transient or null recurrent case, we arrive at the following result which largely summarizes these interplay of these results.

THEOREM 2.3. *Suppose that* $\mathbf{P}$ *is irreducible. There is no invariant distribution for* $\mathbf{P}$ *if and only if* $\mathbf{P}$ *is either transient or null recurrent. On the other hand, there is an invariant distribution for* $\mathbf{P}$ *if and only if* $\mathbf{P}$ *is positive recurrent, in which case that invariant distribution is unique and given by* π*.*

PROOF. These results follow directly from the results thus far proved in this section along with Lemma 5.4 (showing that $\pi(j) = 0$ for every $j \in S$ in the transient case). □

THEOREM 2.4 (A Law of Large Numbers). *Suppose that* $\mathbf{P}$ *is irreducible and recurrent. For non-negative integer* m *and each state* $j \in S$*, let* $N_j(k) = \sum_{m=0}^{k} 1_{\{X_m=j\}}$ *be the number of times the process has hit state* j *in* k *steps. Then*

$$\lim_{k\to\infty} \frac{N_j(k)}{k} = \pi_j \tag{24}$$

holds with probability 1*. Intuitively, this says that the for an irreducible and recurrent process, the long run frequency (or rate) at which the process hits state* j *is equal to* π_j*.*

COROLLARY 2.5. *If* $\mathbf{P}$ *is irreducible and recurrent, then for any states* $i, j \in S$

$$\lim_{k\to\infty} \frac{1}{k} \sum_{m=0}^{k} [\mathbf{P}^m]_{ij} = \lim_{k\to\infty} \frac{1}{k} \sum_{m=0}^{k} p_{ij}^{(m)} = \pi_j. \tag{25}$$

PROOF. With $N_j(k)$ be defined as in Theorem 2.4,

$$\mathbb{E}_i[N_j(k)] = \mathbb{E}_i\Big[\sum_{m=0}^{k} 1_{\{X_m=j\}}\Big] = \sum_{m=0}^{k} \mathbb{E}_i\big[1_{\{X_m=j\}}\big] = \sum_{m=0}^{k} \mathbb{P}_i(X_m = j) = \sum_{m=0}^{k} [\mathbf{P}^m]_{ij}$$

and $\mathbb{E}_i[\pi_j] = \pi_j$ (since π_j is constant), this result follows by applying $\mathbb{E}_i$ to the left- and right-hand sides of (24). □

EXAMPLE 2.2. Suppose a stationary discrete time Markov chain $(X_m)_{m=0}^{\infty}$ on the states $S = \{0, 1, 2\}$ has transition matrix

$$\mathbf{P} = \begin{pmatrix} .6 & .4 & 0 \\ .2 & .6 & .2 \\ 0 & .6 & .4 \end{pmatrix}.$$

For this process, we will do the following.

(1) Find the invariant distribution π.
(2) For each $j \in S$, we will find $\mathbb{E}_j[\rho_j]$.
(3) For each $j \in S$, we will find the long run frequency at which the process hits state j.

Solutions. (1) First, by inspection (perhaps drawing out the jump diagram), it is clear that $\mathbf{P}$ is irreducible, and since there are only finitely many states, it must be positive recurrent. Therefore, using Theorems 2.1 we are guaranteed a unique invariant distribution π. Let's write $\vec{\pi} = (\pi_1, \pi_2, \pi_3)$. We want to solve $\vec{\pi}\,\mathbf{P} = \vec{\pi}$. For this, we equate

$$\vec{\pi}\,\mathbf{P} = (.6\pi_1 + .2\pi_2, .4\pi_1 + .6\pi_2 + .6\pi_3, .3\pi_2 + .4\pi_3)$$

with $\vec{\pi} = (\pi_1, \pi_2, \pi_3)$, resulting in the equations

$$\pi_1 = .6\pi_1 + .2\pi_2$$

$$\pi_2 = .4\pi_1 + .6\pi_2 + .6\pi_3$$

$$\pi_3 = .2\pi_2 + .4\pi_3$$

which can be solved in terms of one of the three variables π_1, π_2, or π_3. Choosing π_1, we have

$$\pi_1 = \pi_1$$

$$\pi_2 = .4\pi_1$$

$$\pi_3 = .2\pi_1$$

telling us that $\vec{\pi} = (\pi_1, .4\pi_1, .2\pi_1) = \pi_1(1, .4, .2) = \pi_1(1, 2/5, 1/5)$. It remains to find π_1, which is done using the requirement that π is a probability mass function

$$1 = \pi_1 + \pi_2 + \pi_3 = \pi_1 + .4\pi_1 + .2\pi_1 = 1.6\pi_1$$

implying that $\pi_1 = 1/1.6 = 5/8$. Thus, $\vec{\pi} = (5/8)(1, 2/5, 1/5) = (5/8, 2/8, 1/8)$.

(2) Now that we have found $\vec{\pi}$, this problem is easy. Indeed, we have by Theorem 2.1 that $\pi_j = 1/\mathbb{E}_j[\rho_j]$, which rearranges to $\mathbb{E}_j[\rho_j] = 1/\pi_j$. Thus $\mathbb{E}_0[\rho_0] = 1/(5/8) = 8/5$, $\mathbb{E}_1[\rho_1] = 1/(2/8) = 8/2 = 4$, and $\mathbb{E}_2[\rho_2] = 1/(1/8) = 8$.

(3) As with the previous solution, having found $\vec{\pi}$, this problem is trivial. By Theorem 2.4, the long run frequency at which the process hits j is π_j. So, for $j = 0$, we have 5/8 (hits/steps) as the long run frequency; for $j = 1$, we have 2/8 (hits/steps) as the long run frequency; for $j = 2$, we have 1/8 (hits/steps) as the long run frequency. △

3. Long Run Results

We continue to reserve the symbol π as the function $\pi(j) = 1/\mathbb{E}_j[\rho_j]$ for each state $j \in S$. We immediately introduce one of the main results of this section, relating the "long run limit" of the stochastic matrix $\mathbf{P}$ with the function π.

THEOREM 3.1. *Suppose that* $\mathbf{P}$ *is irreducible. Then,*

(1) *If* $\mathbf{P}$ *is transient or null recurrent, then* $\lim_{m\to\infty} [\mathbf{P}^m]_{ij} = 0$ *for all states* $i, j \in S$. *Equivalently, from the perspective of transition probabilities, if* $\mathbf{P}$ *is transient or null recurrent, then the* m*-step transition probabilities* $p_{ij}^{(m)}$ *tend to* 0 *as the number of steps* m *tends to infinity.*

(2) *If* $\mathbf{P}$ *is aperiodic and positive recurrent then* $\lim_{m\to\infty}[\mathbf{P}^m]_{ij} = \pi(j)$ *for all states* $i, j \in S$. *Equivalently, from the perspective of transition probabilities, if* $\mathbf{P}$ *is aperiodic and positive recurrent, then the* m*-step transition probability* $p_{ij}^{(m)}$ *from state* i *to state* j *tends to* $\pi(j)$ *as* m *tends to infinity, independent of the state* i.

REMARK 3.1. At this point we have seen that if a state j is irreducible, then $\mathbb{E}_j[\rho_j] = \infty$, which equivalently can be understood as $\pi(j) = 0$. Because of this, Theorem 3.1 can be simplified to the statement: *If* $\mathbf{P}$ *is irreducible and either transient, null recurrent, or aperiodic and positive recurrent, then* $\lim_{m\to\infty}[\mathbf{P}^m]_{ij} = \pi(j)$. △

REMARK 3.2. Define $\mathbf{\Pi}$ as the matrix where each row is $\vec{\pi}$. That is,

$$\mathbf{\Pi} = \begin{pmatrix} \vec{\pi} & \cdots & \cdots \\ \vec{\pi} & \cdots & \cdots \\ \vdots & \ddots & \vdots \\ \vec{\pi} & \cdots & \cdots \end{pmatrix} = \begin{pmatrix} \pi(i_1) & \pi(i_2) & \cdots & \pi(i_N) \\ \pi(i_1) & \pi(i_2) & \cdots & \pi(i_N) \\ \vdots & \vdots & \ddots & \vdots \\ \pi(i_1) & \pi(i_2) & \cdots & \pi(i_N) \end{pmatrix}$$

when $S = \{i_1, ..., i_N\}$. Then, we can interpret Theorem 3.1 as the limiting result $\mathbf{P}^m \to \mathbf{\Pi}$ as $m \to \infty$ whenever $\mathbf{P}$ is irreducible and either transient, null recurrent, or aperiodic and positive recurrent. Notice, however, that $\mathbf{\Pi}$ is the zero matrix in either the transient or null recurrent cases; it is only "interesting" (non-zero) in the aperiodic and positive recurrent case. △

COROLLARY 3.2. *Suppose that* $\mathbf{P}$ *is irreducible and either transient, null recurrent, or aperiodic and positive recurrent. Let* $\nu : S \to \mathbb{R}$ *be any probability mass function on* S. *For each step* m, *let* ν_m *be the probability mass function of* X_m *with respect to* $\mathbb{P}_\nu$; *that is,* $\nu_m(j) = \mathbb{P}_\nu(X_m = j) = [\vec{\nu}\,\mathbf{P}^m]_j$ *for each state* $j \in S$. *Then* $\lim_{m\to\infty}\nu_m(j) = \pi(j)$. *That is, regardless of the initial distribution of the process, the long run distribution of the process tends to* π. *In particular, in the case that* $\mathbf{P}$ *is aperiodic and positive recurrent, the long run distribution of the process tends to the invariant distribution.*

PROOF. We have

$$\lim_{m\to\infty}\nu_m(j) = \lim_{m\to\infty}[\vec{\nu}_0\,\mathbf{P}^m]_j = \lim_{m\to\infty}\sum_{i\in S}\nu_i[\mathbf{P}^m]_{ij} = \sum_{i\in S}\nu_i\lim_{m\to\infty}[\mathbf{P}^m]_{ij} = \sum_{i\in S}\nu_i\,\pi_j = \pi_j\overbrace{\sum_{i\in S}\nu_i}^{1} = \pi_j.$$

This finishes the proof. Let's note that in the case that $\mathbf{P}$ is positive recurrent, then we have proved that π is the invariant distribution for $\mathbf{P}$, which is why ν_m tends to the invariant distribution when $\mathbf{P}$ is aperiodic and positive recurrent. □

EXAMPLE 3.1. Consider the stationary discrete time Markov chain $(X_m)_{m=0}^{\infty}$ with transition matrix

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \end{pmatrix}$$

indexed by the state space $S = \{1, 2, 3\}$. Let us note that this process is irreducible, aperiodic, and positive recurrent (irreducible since all states communicate, aperiodic since the period of any state is 1, and positive recurrent since there are finitely many states each within a closed communication class). In this example, we will do the following:

(1) Justify the existence of and find the invariant distribution π.
(2) Justify the existence of and find the long run limit $\lim_{m\to\infty} \mathbf{P}^m$.
(3) For any probability mass function $\nu : S \to \mathbb{R}$, we will find the long run distribution of the process with respect to $\mathbb{P}_\nu$ (i.e., find the limiting behavior of the mass function of X_m with respect to $\mathbb{P}_\nu$ as m tends to infinity).

Solutions. (1) As mentioned, we know that this process is irreducible and positive recurrent, which by Theorem 2.1 implies the existence of an invariant distribution π (note that we don't need the assumption of aperiodic here). We offer a derivation of π in a few ways:

Method 1. Using first step analysis.

$$\mathbb{E}_1[\rho_1] = 1 + \mathbb{E}_2[\tau_1] = 1 + 1 + \frac{1}{2}\overbrace{\mathbb{E}_1[\tau_1]}^{=0} + \frac{1}{2}\overbrace{\mathbb{E}_3[\tau_1]}^{=1} = \frac{5}{2},$$
$$\mathbb{E}_2[\rho_2] = 1 + \frac{1}{2}\overbrace{\mathbb{E}_1[\tau_2]}^{=1} + \frac{1}{2}\overbrace{\mathbb{E}_3[\tau_2]}^{=2} = \frac{5}{2},$$

and

$$\mathbb{E}_3[\rho_3] = 1 + \mathbb{E}_1[\tau_3] = 1 + 1 + \mathbb{E}_2[\tau_3] = 1 + 1 + 1 + \frac{1}{2}\mathbb{E}_1[\tau_3] + \frac{1}{2}\overbrace{\mathbb{E}_3[\tau_3]}^{=0} = 3 + \frac{1}{2}\,\mathbb{E}_1[\tau_3].$$

To finish finding $\mathbb{E}_3[\rho_3]$, note that we have the equality $1 + \mathbb{E}_1[\tau_3] = 3 + \frac{1}{2}\,\mathbb{E}_1[\tau_3]$ implying that $\mathbb{E}_1[\tau_3] = 4$. Hence $\mathbb{E}_3[\rho_3] = 5$. Therefore, we have found

$$\mathbb{E}_1[\rho_1] = 5/2,\ \mathbb{E}_2[\rho_2] = 5/2,\ \mathbb{E}_3[\rho_3] = 5$$

which gives us the invariant distribution

$$\vec{\pi} = \left(\frac{1}{\mathbb{E}_1[\rho_1]}, \frac{1}{\mathbb{E}_2[\rho_2]}, \frac{1}{\mathbb{E}_3[\rho_3]}\right) = (2/5, 2/5, 1/5).$$

Method 2. Using linear algebra. The invariant distribution must be the unique left eigenvector of $\mathbf{P}$ with eigenvalue 1 which is also a distribution (i.e., non-negative entries which sum to 1). To find the left eigenvector, we find the left null space of the matrix

$$\mathbf{P} - \mathbf{I} = \begin{pmatrix} -1 & 1 & 0 \\ 1/2 & -1 & 1/2 \\ 1 & 0 & -1 \end{pmatrix}$$

Hence we want to find solutions of

$$(x, y, z) \begin{pmatrix} -1 & 1 & 0 \\ 1/2 & -1 & 1/2 \\ 1 & 0 & -1 \end{pmatrix} = (0, 0, 0)$$

which results in solving the following system

$$-x + y/2 + z = 0$$

$$x - y = 0$$

$$y/2 - z = 0$$

which gives us $x = y$, $y = y$, and $z = y/2$ in which y is our "degree of freedom." (Note that each "degree of freedom" represents a dimension of the null space, so we know that this null space has dimension 1). Hence the vector we are searching for is $(x, y, z) = y\,(1, 1, 1/2)$. What this information relates is that the left null space of $\mathbf{P} - \mathbf{I}$ is the span of the vector $(1, 1, 1/2)$. In particular, by normalizing so that the sum of the entries is 1 (which amounts to dividing by $5/2$), we have found the invariant distribution $\vec{\pi} = (2/5, 2/5, 1/5)$.

Method 3. By inspection. Suppose that $\vec{\pi} = (x, y, z)$ is an invariant distribution. Then $\vec{\pi}\,\mathbf{P} = \vec{\pi}$. This means that since

$$(x\,y\,z) \begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \end{pmatrix} = (y/2 + z, x, y/2)$$

we must have $(x, y, z) = (y/2 + z, x, y/2)$. This leaves us with the equations $x = y/2 + z, y = x$, and $z = y/2$. This is easily solved to find $x = 2z, y = 2z$, and $z = z$. Meaning that the solution must be of the form $z(2, 2, 1)$. Normalizing (setting $z = 1/5$), we get $\pi = (2/5, 2/5, 1/5)$.

(2) We found the invariant distribution $\vec{\pi} = (2/5, 2/5, 1/5)$. Since in addition to being irreducible and positive recurrent, the process is also aperiodic, the limit $\lim_{m\to\infty} \mathbf{P}^m$ exists by Theorem

3.1 and the remarks which followed. Hence,

$$\lim_{m\to\infty} \mathbf{P}^m = \begin{pmatrix} 2/5 & 2/5 & 1/5 \\ 2/5 & 2/5 & 1/5 \\ 2/5 & 2/5 & 1/5 \end{pmatrix}$$

(3) We found the invariant distribution $\vec{\pi} = (2/5, 2/5, 1/5)$. Since in addition to being irreducible and positive recurrent, the process is also aperiodic, Corollary 3.2 tells us $\lim_{m\to\infty} \mathbb{P}_\nu(X_m = j)$ exists for each $j \in S$, is equal to $\pi(j)$, and this is true regardless of the initial distribution ν. So, letting $\nu_m(j) = \mathbb{P}_\nu(X_m = j)$, we have $\lim_{m\to\infty} \vec{\nu}_m = \vec{\pi} = (2/5, 2/5, 1/5)$. △

EXAMPLE 3.2. Let $p \in (0, 1)$ and $q = 1 - p$. Consider the stochastic matrix

$$\mathbf{P} = \begin{pmatrix} q & p & 0 & 0 & 0 & \cdots \\ q & 0 & p & 0 & 0 & \cdots \\ q & 0 & 0 & p & 0 & \cdots \\ q & 0 & 0 & 0 & p & \ddots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}$$

indexed by the non-negative integers $S = \{0, 1, 2, 3, ...\}$. In this example we will do the following.

(1) Show $\mathbf{P}$ is irreducible and aperiodic.
(2) Show that $\mathbf{P}$ is positive recurrent and find $\mathbb{E}_j[\rho_j]$ for every $j \in S$.
(3) Consider watching a (stationary discrete time Markov chain) process evolve with transition matrix $\mathbf{P}$. Approximate the probability that, after watching for some time, the process is in state 3. Does the answer depend on how the process was initially distributed over the states in S?

Solutions. (1) To convince ourselves that $\mathbf{P}$ is irreducible, we need to show that for every $i, j \in S$, it holds that $i \leftrightarrow j$. To start, let's notice that for every $j \in S$, it holds that $j \leftrightarrow 0$. This is because $j \to 0$ in one step with probability q, and there is a path from 0 to j with positive probability (following the path $0 \to 1 \to 2 \to \cdots \to j$). Now, if $i, j \in S$, then $i \leftrightarrow 0 \leftrightarrow j$ implying that $i \leftrightarrow j$, which shows that $\mathbf{P}$ is irreducible.

To find the period of each state, we need only find the period of any one state, since all states communicate and hence share the same period. For this, we will find the period of 0. We can get from $0 \to 0$ in 1 step (with probability q), and therefore the period of 0 must be 1. Hence every state has period 1 and we deduce that $\mathbf{P}$ is aperiodic.

(2) Since $\mathbf{P}$ is irreducible, we know that $\mathbf{P}$ is positive recurrent if and only if an invariant distribution π exists. Let us try to find $\vec{\pi} = (\pi_0, \pi_1, \pi_2, \pi_3, ...)$ by inspection.

$$(\pi_0, \pi_1, \pi_2, \pi_3, ...) \begin{pmatrix} q & p & 0 & 0 & 0 & \cdots \\ q & 0 & p & 0 & 0 & \cdots \\ q & 0 & 0 & p & 0 & \cdots \\ q & 0 & 0 & 0 & p & \ddots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix} = (q, p\,\pi_0, p\,\pi_1, p\,\pi_2, p\,\pi_3, ...)$$

This shows that if $\vec{\pi}$ is an invariant distribution, then $\pi_0 = q$, $\pi_1 = p\,\pi_0 = pq$, $\pi_2 = p\,\pi_1 = p^2 q$, etc. In fact, we see that in general, $\pi_j = p^j q$. Thus, by our calculations, $\vec{\pi} = (q, pq, p^2 q, p^3 q, ...)$ is a left eigenvector of $\mathbf{P}$ with eigenvalue 1. To check that π is truly the invariant distribution, we need to check that the entries sum to 1:

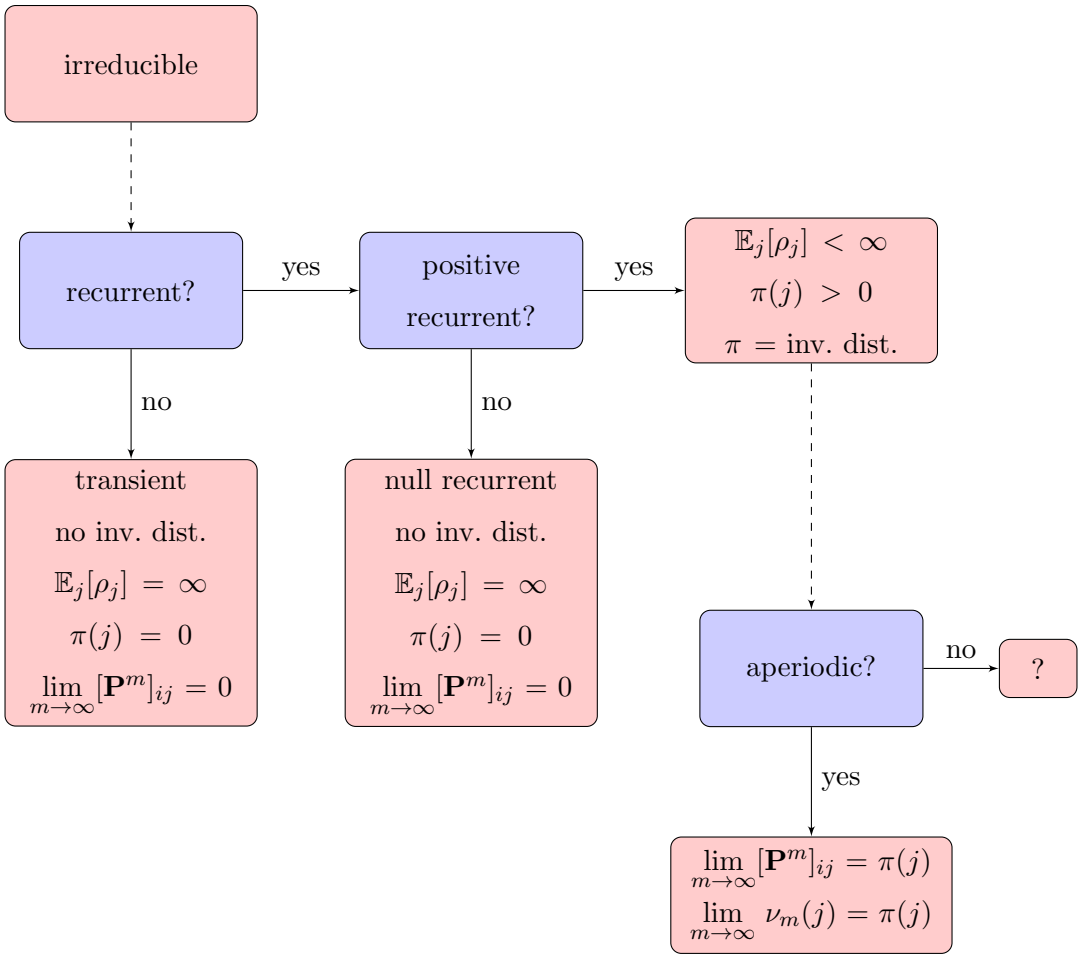
$$\sum_{j=0}^{\infty} p^j q = q \sum_{j=0}^{\infty} p^j = q\,\frac{1}{1-p} = 1$$

where the last equality follows from the fact that $q = 1 - p$. Therefore we have found an invariant distribution π for $\mathbf{P}$, implying that $\mathbf{P}$ is positive recurrent (and that π is the unique invariant distribution). In particular, we find that $\mathbb{E}_j[\rho_j] = \frac{1}{\pi(j)} = \frac{1}{qp^j} < \infty$.

(3) Since we have shown that $\mathbf{P}$ is irreducible, aperiodic, and positive recurrent, Corollary 3.2 implies that $\lim_{m\to\infty} \mathbb{P}_\nu(X_m = j) = \pi(j)$ for any initial distribution $\nu : S \to \mathbb{R}$. In particular, for large values of m, $\mathbb{P}_\nu(X_m = 3) \approx \pi(3) = p^3 q$. △

4. Summary Flow Chart

Before moving forward, we dedicate this section to a single flow chart summarising many of our results from this chapter. When reading this flow diagram, any claims such as $\mathbb{E}_j[\rho_j] = \infty$, $\lim_{m\to\infty} [\mathbf{P}^m]_{ij} = 0$, $\pi(j) > 0$, etc., are to be understood as being true for all states $i, j \in S$; also, as frequently used above, we use the symbol $\nu_m(j) = \mathbb{P}_\nu(X_m = j)$ (by Corollary 3.2, the initial distribution ν will be unimportant in the flow chart).



5. Considerations in the Non-irreducible Case

Suppose that $\mathbf{P}$ is not necessarily irreducible. Then we can't immediately use most of our results concerning invariant and long run distributions. However, we can still glean useful results in these cases by simply "reducing" the matrix $\mathbf{P}$ into the irreducible submatrices $\mathbf{P}_C$ for each communication class C.

LEMMA 5.1. *Let $C \subset S$ be a closed communication class (induced by $\mathbf{P}$). Then all previous results in this chapter concerning the case when $\mathbf{P}$ irreducible on the states S can be used for $\mathbf{P}_C$ on the states C.*

PROOF. We have shown that if C is a closed communication class, then $\mathbf{P}_C$ is itself a stochastic matrix. Further, since C is a communication class, $\mathbf{P}_C$ is irreducible. Thus $\mathbf{P}_C$ is an irreducible stochastic matrix indexed by the states C, and hence everything we have done thus far with irreducible stochastic matrices will hold for $\mathbf{P}_C$ restricted to the states C. □

THEOREM 5.2. *For every communication class C which is positive recurrent, there is an invariant distribution π_C for $\mathbf{P}$ defined by $\pi_C(j) = 1/\mathbb{E}_j[\rho_j]$ for every $j \in C$ and $\pi_C(j) = 0$ for $j \in S \setminus C$. Moreover, if C_1 and C_2 are distinct closed communication classes which are both positive recurrent, then the vector representations of the invariant distributions $\vec{\pi}_{C_1}$ and $\vec{\pi}_{C_2}$ are orthogonal (i.e., the dot-product of these two vectors is zero).*

PROOF. If C is closed and positive recurrent, then we have seen that the invariant distribution of $\mathbf{P}_C$ has the form $j \mapsto 1/\mathbb{E}_j[\rho_j]$ for each $j \in C$. Reminding ourselves of the definition of an invariant distribution, this means that $\sum_{i \in C} (1/\mathbb{E}_i[\rho_i])\, [\mathbf{P}_C]_{ij} = 1/\mathbb{E}_j[\rho_j]$ whenever $j \in C$; in fact, if $i, j \in C$, then $[\mathbf{P}_C]_{ij} = \mathbf{P}_{ij}$, so we could have instead written $\sum_{i \in C} (1/\mathbb{E}_i[\rho_i])\, \mathbf{P}_{ij} = 1/\mathbb{E}_j[\rho_j]$ when $j \in C$. Hence, with π_C defined as above,

$$[\vec{\pi}_C \mathbf{P}]_j = \sum_{i \in S} \pi_C(i)\, \mathbf{P}_{ij} = \sum_{i \notin C} 0 \cdot \mathbf{P}_{ij} + \sum_{i \in C} (1/\mathbb{E}_i[\rho_i])\, \mathbf{P}_{ij} = \begin{cases} 0 & j \notin C \\ 1/\mathbb{E}_j[\rho_j] & j \in C \end{cases}$$

Note that the 0 for $j \notin C$ comes from the fact that $\mathbf{P}_{ij} = 0$ if $i \in C$ and $j \notin C$ because C is closed. Since the last equality or our calculations is the definition of $\pi_C(j)$, we have shown that $[\vec{\pi}_C \mathbf{P}]_j = \pi_C(j)$, proving that π_C is an invariant distribution for $\mathbf{P}$. Orthogonality of these invariant distributions π_{C_1} and π_{C_2} for distinct communication classes C_1 and C_2 follows easily by definition and the fact that C_1 and C_2 are disjoint. Indeed, $\pi_{C_1}(j) = 0$ for every $j \in C_2$ since $C_2 \subset S \setminus C_1$, and symmetrically, $\pi_{C_2}(j) = 0$ for every $j \in C_1$; from this, the orthogonality follows quickly. □

DEFINITION 5.1. The invariant distrubiton π_C described in Theorem 5.2 is called the *canonical invariant distribution* for $\mathbf{P}$ corresponding to the communication class C. △

COROLLARY 5.3. *If $\mathbf{P}$ has at least two invariant distributions whose vector representations are linearly independent, then it has infinitely many distinct invariant distributions. In particular, if there are at least two distinct closed communcation classes which are positive recurrent, there are infinitely many invariant distributions of $\mathbf{P}$.*

PROOF. Suppose that there are at least two two distinct closed communcation classes which are positive recurrent, and label two of them C_1 and C_2. Let π_{C_1} and π_{C_2} be the corresponding canonical invariant distributions for $\mathbf{P}$. Then π_{C_1} and π_{C_2} are two invariant distributions whose vector representations $\vec{\pi}_{C_1}$ and $\vec{\pi}_{C_2}$ are orthogonal and hence linearly independent. We will prove this Corollary using π_{C_1} and π_{C_2}, but the argument only uses the fact that they are invariant distributions of $\mathbf{P}$ which are linearly independent, and thus holds for the more general case. Now,

for any number $p \in (0,1)$, the linearity of matrix multiplication implies

$$[p\,\vec{\pi}_{C_1} + (1-p)\,\vec{\pi}_{C_2}]\mathbf{P} = p[\vec{\pi}_{C_1}\mathbf{P}] + (1-p)[\vec{\pi}_{C_2}\mathbf{P}] = p\,\vec{\pi}_{C_1} + (1-p)\,\vec{\pi}_{C_2}$$

This shows that the vector $p\,\vec{\pi}_{C_1} + (1-p)\,\vec{\pi}_{C_2}$ is a left eigenvector of $\mathbf{P}$ with eigenvalue 1. The argument that this vector is, in fact, an invariant distribution thus reduces to showing that each entry is non-negative and the sum over the entries equals 1; we leave this as an easy exercise. We therefore conclude that for any $p \in (0,1)$, we have that $p\,\pi_{C_1} + (1-p)\,\pi_{C_2}$ is an invariant distribution. Since the vectors $\vec{\pi}_{C_1}$ and $\vec{\pi}_{C_2}$ are linearly independent, each distinct $p \in (0,1)$ corresponds to a distinct invariant distribution for $\mathbf{P}$. □

REMARK 5.1. To distil some of these results in an interesting linear algebraic way, what Theorem 5.2 implies is that the dimension of the eigenspace of $\mathbf{P}$ corresponding to eigenvalue 1 is bounded below by the number of closed positive recurrent communication classes of $\mathbf{P}$. From this, we see that △

To this point we have only considered closed communication classes. We end this section considering a result for open communication classes.

THEOREM 5.4. *Suppose that O is an open communication class for the stochastic matrix $\mathbf{P}$. Then the reduced matrix $\mathbf{P}_O$ does not have an invariant distribution. Moreover, $\lim_{m\to\infty}[\mathbf{P}_O^m]_{ij} = 0$ for every $i, j \in O$.*

PROOF. Using techniques from linear algebra, one can show that the eigenvalues of $\mathbf{P}_O$ all have magnitude strictly less than 1. This implies that $\mathbf{P}_O$ can not have an invariant distribution (since the eigenvalue of such an invariant distribution is 1) and further (by another linear algebraic argument) that the $\lim_{m\to\infty}[\mathbf{P}_O^m]_{ij} = 0$. However, let us argue that this must be true with intuition. Suppose that an invariant distribution π_O exists for $\mathbf{P}_O$. Then $\pi_O\,\mathbf{P}_O = \pi_O$ and as before, we can canonically extend π_O to an invariant distribution π for $\mathbf{P}$ by simply setting all components indexed by states outside of O with 0. That is $\pi(j) = \pi_O(j)$ when $j \in O$, otherwise $\pi(j) = 0$. Now, since $\vec{\pi}\,\mathbf{P} = \vec{\pi}$, this means that if the process starts with initial distribution π, then as the process evolves its distribution constantly stays π. However, this implies that (since $\pi(j) = 0$ for every $j \notin O$), that the process will never leave O, which contradicts the fact that regardless of where in O the process starts, it will leave O with positive probability (this is equivalent to the openness of O). So we realize there can be no such invariant distribution for $\mathbf{P}_O$. Continuing along this same argument, starting from any state in O, there is a positive probability that the process will leave O in finitely many steps, which indicates that with some positive probability, the process moves

out of the open communication class and thusly never comes back. Speaking roughly, this means that as the process evolves the probability the process stays in O is "escaping" outside of O, which eventually leads to the fact that in the long run, the probability of starting in O and ending in O should be 0; i.e., $\lim_{m\to\infty}[\mathbf{P}_O^m]_{ij} = 0$ for every $i, j \in O$. □

6. Exercises

(1) For each of the following irreducible stochastic matrices $\mathbf{P}$ (or related jump diagram), either justify why there is no invariant distribution, or find the invariant distribution if one exists. Further for every $i, j \in S$, find $\lim_{m\to\infty}[\mathbf{P}^m]_{ij}$ if the limit exists.

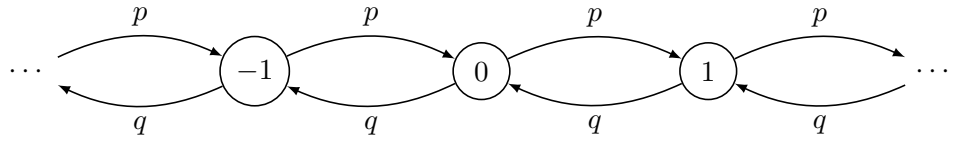
(a) The transition matrix

$$\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

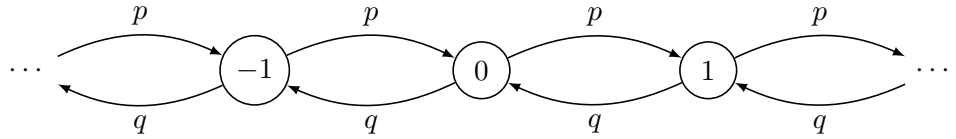
(b) The transition matrix

$$\mathbf{P} = \begin{pmatrix} 1/2 & 1/2 \\ 1 & 0 \end{pmatrix}.$$

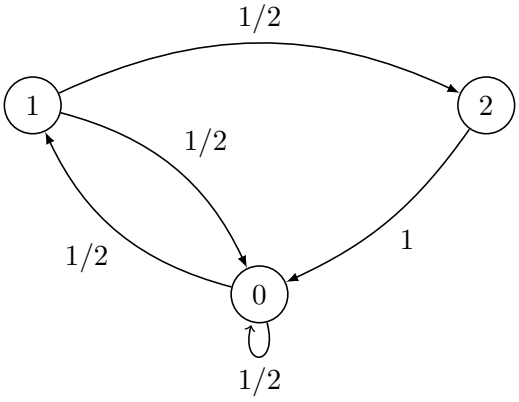
(c) The 1-D walk for $p \in (0,1)$, $q = 1-p$, and with $p \neq q$,



(d) The 1-D walk for $p = q = 1/2$,



(e) The process with jump diagram:



(2) Suppose that within a certain society, the class ranking of individuals is modeled by a stationary discrete time Markov chain with state space $S = \{0, 1, 2\}$ (0 being the lowest ranking and 2 being the highest). The class motility is monitored each year and the following transition matrix is given

$$\mathbf{P} = \begin{pmatrix} 8/10 & 2/10 & 0 \\ 3/10 & 5/10 & 2/10 \\ 0 & 4/10 & 6/10 \end{pmatrix}$$

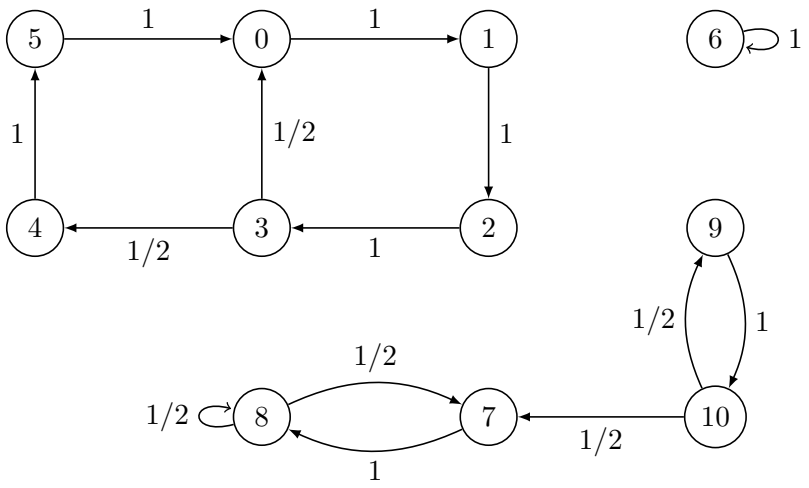
How to read this is that each year an average individual with class ranking 0 will stay at rank 0 with probability 8/10, or move up to rank 1 with probability 2/10; an average individual with class ranking 1 will move back to class rank 0 with probability 3/10, stay at rank 1 with probability 5/10, and move up to class rank 2 with probability 2/10; finally, an average individual with class rank 3 will move back to class rank 2 with probability 4/10, and stay at class rank 3 with probability 6/10. Assuming this model is correct, after letting the process evolve for many decades, what is the approximate probability that a randomly selected average individual has class rank 0? Class rank 1? Class rank 2?

(3) Within a certain society of individuals, there is a certain gene pair where each gene in the pair can have one of two types: A or a. This leaves $\{AA, Aa, aa\}$ as the possible gene pairs an individual has. Assuming that each offspring in the society gets one of the two genes in the pair from each of its two parents, we can reasonably model the gene pair motility through descendancy as a stationary discrete time Markov chain $(X_n)_{n=0}^{\infty}$ with states $S = \{AA, Aa, aa\}$ (if you are concered that S is not a subset of $\mathbb{R}$, you can, for example, set $S = \{0, 1, 2\}$ and have state j represent j copies of gene a), where X_n represents the gene pair of the nth descendant (we will ignore issues such as death of an individual before having another descendant). Assuming the that transition matrix for this process is

$$\mathbf{P} = \begin{pmatrix} .6 & .4 & 0 \\ .2 & .6 & .2 \\ 0 & .6 & .4 \end{pmatrix}$$

For each gene pair $g \in S$, what is the long run probability that the (far future) descendants have gene pair g?

(4) Consider the following jump diagram



which is for the stochastic matrix $\mathbf{P}$.

(a) Justify why each closed communication class is positive recurrent.

(b) For each closed communication class C, find the invariant distribution for $\mathbf{P}_C$.

(c) From the previous part, for each closed communication class C, find the canonical invariant distribution for $\mathbf{P}$ generated by C. Quickly verify that these distributions are orthogonal.

(d) For any communication class C which is either closed and aperiodic or open, find $\lim_{m\to\infty} \mathbf{P}_C^m$.

Part 3

Continuous Time Markov Chains

CHAPTER 7

Introduction to Continuous Time Markov Chains

We now turn our attention to the setting where the time variables are from a "continuum" of values; namely, when T is an interval subset of $\mathbb{R}$. However, before we jump in too deep, we start by presenting some theory and results which will let us generalize a transition matrix to the continuous time setting.

1. Markov Semigroups and Their Generators

In discrete time, much of the work involved studying Markov chains boiled down to studying the transition matrices. We still reserve the symbol S for some discrete set, the elements of which we continue to call states.

DEFINITION 1.1. For every $t \geq 0$, suppose that $\mathbf{P}(t)$ is a square matrix indexed by S. We say that $(\mathbf{P}(t))_{t\geq 0}$ is a (matrix-valued) *Markov semigroup* when the following three properties hold.

(1) $\lim_{t\to 0+} \mathbf{P}(t) = \mathbf{I} = \mathbf{P}(0)$, where $\mathbf{I}$ is the identity matrix (indexed by S).
(2) For all $s, t \geq 0$, it holds that $\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s)$.
(3) For every $t \geq 0$, $\mathbf{P}(t)$ is a stochastic matrix. That is, $\sum_{j\in S} [\mathbf{P}(t)]_{ij} = 1$ for each $i \in S$ and $[\mathbf{P}(t)]_{ij} \geq 0$ for every $i, j \in S$.

△

REMARK 1.1. Property (1) is referred to as *continuity* of $(\mathbf{P}(t))_{t\geq 0}$. Property (2) is referred to as the *semigroup property* of $(\mathbf{P}(t))_{t\geq 0}$; in fact, we will see that property (2) is equivalent to the Chapmann-Kolmogorov equations for a continuous time Markov chain. △

LEMMA 1.1. *If $(\mathbf{P}(t))_{t\geq 0}$ is a Markov semigroup, then for any $s, t \geq 0$, we have $\mathbf{P}(s)\mathbf{P}(t) = \mathbf{P}(t)\mathbf{P}(s)$.*

PROOF. Using the semigroup property, we have

$$\mathbf{P}(s)\mathbf{P}(t) = \mathbf{P}(s+t) = \mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s),$$

which is what we wanted to show. □

THEOREM 1.2. *Suppose that* $(\mathbf{P}(t))_{t\geq 0}$ *is a Markov semigroup. Then the derivative*

$$\dot{\mathbf{P}}(t) = \lim_{h\to 0} \frac{1}{h}\Big[\mathbf{P}(t+h) - \mathbf{P}(t)\Big]$$

exists for every $t \geq 0$ *(interpreted as the righthand derivative at* $t = 0$*). Moreover, defining* $\mathbf{L}$ *as* $\mathbf{L} = \dot{\mathbf{P}}(0)$*, then* $\dot{\mathbf{P}}(t) = \mathbf{L}\mathbf{P}(t) = \mathbf{P}(t)\mathbf{L}$ *for every* $t \geq 0$*.*

IDEA OF THE PROOF. Assume first that $\dot{\mathbf{P}}(0)$ exists and $\mathbf{L}$ is defined by $\dot{\mathbf{P}}(0) = \mathbf{L}$. We will show that this implies the result. Indeed, for any $t > 0$ we have from the semigroup property,

$$\frac{1}{h}\Big[\mathbf{P}(t+h) - \mathbf{P}(t)\Big] = \frac{1}{h}\Big[\mathbf{P}(t)\mathbf{P}(h) - \mathbf{P}(t)\Big] = \Big[\frac{1}{h}\Big(\mathbf{P}(h) - \mathbf{I}\Big)\Big]\mathbf{P}(t) = \Big[\frac{1}{h}\Big(\mathbf{P}(h) - \mathbf{P}(0)\Big)\Big]\mathbf{P}(t)$$

From the last equality, we see

$$\dot{\mathbf{P}}(t) = \lim_{h\to 0} \frac{1}{h}\Big[\mathbf{P}(t+h) - \mathbf{P}(t)\Big] = \lim_{h\to 0}\Big[\frac{1}{h}\Big(\mathbf{P}(h) - \mathbf{P}(0)\Big)\Big]\mathbf{P}(t) = \dot{\mathbf{P}}(0)\mathbf{P}(t) = \mathbf{L}\mathbf{P}(t).$$

Had we factored $\mathbf{P}(t)$ to the left (which we could have equally done by Lemma 1.1), we would have also found $\dot{\mathbf{P}}(t) = \mathbf{P}(t)\mathbf{L}$. Therefore, it is sufficient to prove that $\mathbf{P}(t)$ is differentiable at 0. We relegate this (slightly tedious) proof to the appendices. □

DEFINITION 1.2. Let $\mathbf{L}$ be a matrix indexed by S. We will call $\mathbf{L}$ a *Markov generator* when the following three properties hold.

(1) Each non-diagonal element is non-negative. That is, $[\mathbf{L}]_{ij} \geq 0$ for all $i \neq j$.
(2) The diagonal elements are non-positive. That is, $[\mathbf{L}]_{ii} \leq 0$ for all $i \in S$.
(3) The sum of each row of $\mathbf{L}$ is 0. That is, $\sum_{j\in S} [\mathbf{L}]_{ij} = 0$ for each $i \in S$.

△

THEOREM 1.3. *Suppose that for every* $t \geq 0$*,* $\mathbf{P}(t)$ *is a matrix indexed by* S*. Then* $(\mathbf{P}(t))_{t\geq 0}$ *is a Markov semigroup if and only if there is a Markov generator* $\mathbf{L}$ *such that* $\mathbf{P}(t) = e^{t\mathbf{L}} = \sum_{k=0}^{\infty} \frac{t^k}{k!}\mathbf{L}^k$*. (Here,* $\mathbf{L}^k$ *is the* k*th power of the matrix* $\mathbf{L}$*, and as usual, we define* $\mathbf{L}^0 = \mathbf{I}$*). Moreover, if* $\mathbf{P}(t) = e^{t\mathbf{L}}$*, then* $\dot{\mathbf{P}}(t) = \mathbf{L}\mathbf{P}(t) = \mathbf{P}(t)\mathbf{L}$ *for every* $t \geq 0$*.*

IDEA OF THE PROOF. A fact from linear algebra and differential equations is that as long as the series defining $e^{t\mathbf{L}}$ converges absolutely (which it always will if S is finite or we can show that $\mathbf{L}$ is a Markov generator), then $e^{t\mathbf{L}}$ is differentiable in t and $\frac{d}{dt}e^{t\mathbf{L}} = \mathbf{L}e^{t\mathbf{L}} = e^{t\mathbf{L}}\mathbf{L}$.

From Theorem 1.2, we know that if $(\mathbf{P}(t))_{t\geq 0}$ is a Markov semigroup, then there exists some matrix $\mathbf{L}$ such that $\dot{\mathbf{P}}(t) = \mathbf{L}\mathbf{P}(t)$ with boundary condition $\mathbf{P}(0) = \mathbf{I}$. By the uniqueness of solutions to such differential equations, since $e^{t\mathbf{L}}$ also satisfies this differential equation with the

same boundary conditions, then it must be that $\mathbf{P}(t) = e^{t\mathbf{L}}$. We must show that $\mathbf{L}$ is a Markov generator. To do so, let $\mathbf{1}$ be the column vector of all 1s indexed by S. Since $\mathbf{P}(t)$ is a stochastic matrix, $\mathbf{P}(t)\mathbf{1} = \mathbf{1}$ for every t. In particular, $\mathbf{L1} = \frac{d}{dt}\big|_{t=0}\mathbf{P}(t)\mathbf{1} = \frac{d}{dt}\mathbf{1} = \mathbf{0}$, where $\mathbf{0}$ is the column vector of all 0s indexed by S. Since the ith entry in $\mathbf{L1}$ is the sum of the ith row of $\mathbf{L}$, we see that the sum of each row of $\mathbf{L}$ must be 0. Finally, to convince ourselves that the diagonal elements of $\mathbf{L}$ are non-positive and the off-diagonal elements are non-negative, let us consider $\mathbf{P}(t)$ for t very near 0. We know that at $t = 0$, we have $\mathbf{P}(0) = \mathbf{I}$, where the diagonals all start as 1 and the off-diagonals all start at 0. Since $\mathbf{P}(t)$ is continuous in t and is a stochastic matrix, then we realize that all diagonal elements must be non-increasing (decreasing away from 1) as the off-diagonals are non-decreasing (increasing away from 0) for times very near 0. Hence the time derivative at $t = 0$ of these elements must leave a non-positive diagonal value and non-negative off diagonal values; but, $\dot{\mathbf{P}}(0) = \mathbf{L}$.

For the converse, noting that $e^{(t+s)\mathbf{L}} = e^{t\mathbf{L}}e^{s\mathbf{L}}$ and $\lim_{t\to 0+} e^{t\mathbf{L}} = e^{0\cdot\mathbf{L}} = \mathbf{I}$, we see that $e^{t\mathbf{L}}$ satisfies all the properties of a Markov semigroup except that it potentially does not form a stochastic matrix for each t; this is where we must use the assumption of $\mathbf{L}$ being a Markov generator. Working backwards from before, since $\mathbf{L}$ is a Markov generator, we find that $\mathbf{L1} = \mathbf{0}$ implying that $\frac{d}{dt}e^{t\mathbf{L}}\mathbf{1} = \mathbf{0}$ for every t; hence $e^{t}\mathbf{1}$ is constant in time. Since $e^{0\cdot\mathbf{L}}\mathbf{1} = \mathbf{I1} = \mathbf{1}$ and $e^{t\mathbf{L}}\mathbf{1}$ is constant in t, we must have that $e^{t\mathbf{L}}\mathbf{1} = \mathbf{1}$ for every t, further implying that each row of $e^{t\mathbf{L}}$ sums to 1. Since the off-diagonal elements of $\mathbf{L}$ are non-negative (and $\mathbf{L}$ represents the infinitesimal change in $e^{t\mathbf{L}}$), it must be that the off-diagonal entries in $e^{t\mathbf{L}}$ are non-decreasing in time. In particular, since at time $t = 0$, the off diagonal entries of $e^{t\mathbf{L}}$ are 0, then the off-diagonal entries of $e^{t\mathbf{L}}$ are non-negative for all time t. Using the same methods as the proof of Proposition 1.4 below, we can prove that the diagonal elements of $e^{t\mathbf{L}}$ are bounded above 0 for all time t. Thus, if $\mathbf{L}$ is a Markov generator, then $e^{t\mathbf{L}}$ is a stochastic matrix for all times t. Hence, $e^{t\mathbf{L}}$ defines a Markov semigroup. □

Before moving on, we finish this section with the following proposition and its corollary, which give us a useful "growth" behavior for Markov semi-groups.

PROPOSITION 1.4. *Let $(\mathbf{P}(t))_{t\geq 0}$ be a Markov semi-group with generator $\mathbf{L}$ indexed by S. Let $i \in S$ and $\lambda_i = -[\mathbf{L}]_{ii}$ (noting that λ_i is non-negative since the diagonal elements of $\mathbf{L}$ are non-positive). Then, for every $t > 0$, it holds that $[\mathbf{P}(t)]_{ii} \geq e^{-t\lambda_i} > 0$.*

PROOF. We have $\frac{d}{dt}[\mathbf{P}(t)]_{ii} = [\mathbf{LP}(t)]_{ii} = \sum_{j\in S}[\mathbf{L}]_{ij}[\mathbf{P}(t)]_{ji}$. Now, expanding the sum,

$$\frac{d}{dt}[\mathbf{P}(t)]_{ii} = -\lambda_i[\mathbf{P}(t)]_{ii} + \sum_{j\neq i}[\mathbf{L}]_{ij}[\mathbf{P}(t)]_{ji} = -\lambda_i[\mathbf{P}(t)]_{ii} + h(t)$$

where as defined $h(t) = \sum_{j \neq i} [\mathbf{L}]_{ij} [\mathbf{P}(t)]_{ji}$, which implies $h(t) \geq 0$. From ordinary differential equations, the solution to the equation $\dot{y}(t) = -\lambda_i \, y(t) + h(t)$ is

$$y(t) = e^{-\lambda_i t} \Big[y(0) + \int_0^t e^{\lambda_i s} h(s) \, ds \Big].$$

Hence, replacing $y(t)$ with $[\mathbf{P}(t)]_{ii}$ (and noting that in this case, $y(0) = 1$) we have

$$[\mathbf{P}(t)]_{ii} = e^{-\lambda_i t} \Big[1 + \underbrace{\int_0^t e^{\lambda_i s} h(s) ds}_{\geq 0} \Big] \geq e^{-\lambda_i t}.$$

□

COROLLARY 1.5. *Let $(\mathbf{P}(t))_{t \geq 0}$ be the Markov semi-group of a continuous time Markov chain indexed by the state space S. Let $i, j \in S$ with $i \neq j$. If there exists some $t \geq 0$ such that $[\mathbf{P}(t)]_{ij} > 0$, then it happens that $[\mathbf{P}(t)]_{ij} > 0$ for all $t > 0$.*

PROOF. Suppose that $t > 0$ and $[P(t)]_{i,j} > 0$. Let $s \geq 0$. Using the semi-group property,

$$[\mathbf{P}(t+s)]_{ij} = [\mathbf{P}(s)\mathbf{P}(t)]_{ij} = \sum_{k \in S} [\mathbf{P}(s)]_{ik} [\mathbf{P}(t)]_{ki} \geq [\mathbf{P}(s)]_{ii} [\mathbf{P}(t)]_{ij}.$$

By the previous proposition, $[\mathbf{P}(s)]_{ii} > 0$ and by assumption $[\mathbf{P}(t)]_{i,j} > 0$, which shows that $[\mathbf{P}(t+s)]_{i,j} > 0$. So now, we know that if for any $t > 0$, $[\mathbf{P}(t)]_{ij} > 0$, then for any future time this must also be true. On the other hand, if $[\mathbf{P}(t)]_{ij} = 0$ for some $t > 0$, then by what we just showed, it must be true that $[\mathbf{P}(s)]_{ij} = 0$ for every $0 \leq s \leq t$. In particular, $[\mathbf{P}(t)]_{ij}$ must then satisfy $\frac{d}{dt}\big|_{t=0} [\mathbf{P}(t)]_{ij} = 0$. Therefore, $[\mathbf{P}(t)]_{ij}$ the unique solution to the linear differential equation $\frac{d}{dt}[\mathbf{P}(t)]_{ij} = [\mathbf{L}\mathbf{P}(t)]_{ij}$ with initial conditions $[\mathbf{P}(0)]_{ij} = 0$ and $\frac{d}{dt}\big|_{t=0} [\mathbf{P}(t)]_{ij} = 0$. You can now check that $[\mathbf{P}(t)]_{ij} = 0$ is that solution (since it must be unique!). □

2. The Markov Property and Stationarity for Continuous Time Chains

Let our time variables T be an interval in $\mathbb{R}$ and continue to let our state space S be discrete. As in the discrete time case, a process $(X_t)_{t \in T}$ is a collection of random variables indexed by the time variables; unlike the discrete time case, T is an interval of real numbers, hence we say such a process is a *continuous time* process (as t moves through a "continuum" of values). When the state space of each of the random variables X_t in a process $(X_t)_{t \in T}$ is a discrete set S, we will continue to use the word *chain*, and call S the state space for the process itself. Most often, the time variables will be an infinite half-line of the form $T = [a, \infty)$, and in this case, instead of writing $(X_t)_{t \in T}$, we will write $(X_t)_{t \geq a}$.

As you will soon realize, we will almost exclusively consider processes on the time interval $[0, \infty)$, hence following the definition of the Markov property below, we will abruptly start looking at processes of the form $(X_t)_{t\geq 0}$. Realize that, much like in the discrete time setting, we can simply re-index our process to force the time variables to be $[0, \infty)$ without losing any structure (e.g., when considering $(X_t)_{t\geq a}$, use $(X_{t+a})_{t\geq 0}$ instead), and thusly we don't lose generality by this convention.

DEFINITION 2.1. Suppose that $(X_t)_{t\in T}$ is a continuous time chain on the state space S. We will say that this process satisfies the *Markov property* when for all times $t_0 < t_1 < \cdots < t_n < t$,

$$\mathbb{P}(X_t = j \,|\, X_{t_0} = j_0, X_{t_1} = j_1, ..., X_{t_n} = j_n) = \mathbb{P}(X_t = j \,|\, X_{t_n} = j_n).$$

When it happens that the continuous time chain satisfies the Markov property, we will call the process a *continous time Markov chain.* △

REMARK 2.1. This definition of the Markov property looks identical to the discrete time setting, but because the time variables here are being drawn from an interval, we can't interpret "consecutive" time variables in the same manner as we had in the discrete time setting. △

REMARK 2.2. As in the discrete time setting, note that there is a background probability $\mathbb{P}$ which the Markov property is defined with respect to. Much like we have done in the discrete time setting, we will frequently change the probability (typically by conditioning on events), so it's important to know when altering the probability that the Markov property still holds for the process. △

DEFINITION 2.2. Suppose that with respect to the probability $\mathbb{P}$, the process $(X_t)_{t\geq 0}$ is a continuous time Markov chain with state space S. We call $(X_t)_{t\geq 0}$ *time homogenous* or *stationary* when $\mathbb{P}(X_{t+s} = j \,|\, X_s = i) = \mathbb{P}(X_t = j \,|\, X_0 = i)$ for any $s, t \geq 0$ and $i, j \in S$. Equivalently, for any times $0 \leq s < t$, the process $(X_t)_{t\geq 0}$ is time homogenous when $\mathbb{P}(X_t = j \,|\, X_s = i) = \mathbb{P}(X_{t-s} = j \,|\, X_0 = i)$ depends only on the increment of time $t - s$. △

REMARK 2.3. Recall that for a discrete time Markov chain, we use the word stationary when the probability of the chain being in state j after $n + k$ steps given that we are in state i after n steps depends only on the number of interim time steps k, but does not depend on the number of steps n it took the chain to reach state i. Definition 2.2 is truly the continuous time analog, since we use stationary to mean that the probability of being in state j at time $t + s$ knowing that the chain was in state i at time s only depends on the interim time t, but does not depend on the time s it takes the chain to reach state i. △

DEFINITION 2.3. We will say that a process $(X_t)_{t\geq 0}$ is *right continuous* when

$$\mathbb{P}\Big(\lim_{t\to t_0+} X_t = X_{t_0}\Big) = 1 \tag{RC}$$

holds for every $t_0 \geq 0$. To shorthand these assumption, when the process is assumed right continuous, we may say it satisfies RC. △

Since we typically use processes $(X_t)_{t\geq 0}$ to model the evolution of physical systems, many of which have this right continuity and lefthand limits behavior, assuming these properties is not too restrictive. In the following example we give a process which does not satisfy this property, but does not model any meaningful physical system.

EXAMPLE 2.1. One can create a stationary continuous time Markov chain $(X_t)_{t\geq 0}$ such that for every $s \neq t$, the random variables X_s and X_t are independent. In this case there is no reason why this process will be right continuous (since the the process can jump around between the states infinitely many times within any time interval, as independence here means that the state X_t has no regard to the the behavior of the process except at time t). △

LEMMA 2.1. *Let $f : S \to \mathbb{R}$ be a bounded function and suppose that $(X_t)_{t\geq 0}$ is a stationary continuous time Markov chain satisfying RC. Then $\lim_{t\to t_0+} \mathbb{E}[f(X_t)] = \mathbb{E}[f(X_{t_0})]$.*

PROOF. This is an immediate consequence of the dominated convergence theorem since $\lim_{t\to t_0+} f(X_t) = f(X_{t_0})$ and $\lim_{t\to t_0-} f(X_t)$ exists with probability 1. □

DEFINITION 2.4. Let $(X_t)_{t\geq 0}$ be a stationary continuous time Markov chain with state space S. Define the *transition semigroup* $(\mathbf{P}(t))_{t\geq 0}$ of $(X_t)_{t\geq 0}$ by

$$[\mathbf{P}(t)]_{ij} = \mathbb{P}(X_t = j \,|\, X_0 = i)$$

for every $i, j \in S$. That is, $[\mathbf{P}(t)]_{ij}$ gives the probability that, starting from from state i, the chain will be in state j at time t. △

REMARK 2.4. Note that with the assumption that the process is stationary, for any times s and t,

$$[\mathbf{P}(t)]_{ij} = \mathbb{P}(X_{s+t} = j \,|\, X_s = i).$$

△

Many of the definitions from the discrete time case require little or no adjustments to import into the real of continuous time. In particular, we have the following immediate analogues.

THEOREM 2.2 (Markov Semigroups are Transition Semigroups). *Suppose S is a finite subset of $\mathbb{R}$ and let $\mathbf{L}$ be a Markov generator indexed by S. Then there exists a sample space Ω, a stochastic process $(X_t)_{t\geq 0}$ defined on Ω with state space S, and a probability $\mathbb{P}$ defined on Ω such that for every $i \in S$, it holds that $\mathbb{P}(X_0 = i) > 0$, and with respect to $\mathbb{P}_i = \mathbb{P}(\,\cdot \,|\, X_0 = i)$, the process $(X_t)_{t\geq 0}$ is a stationary continuous time Markov chain with transition semigroup $(e^{t\mathbf{L}})_{t\geq 0}$.*

COROLLARY 2.3. *Suppose S is a finite subset of $\mathbb{R}$ and let $\mathbf{L}$ be a Markov generator indexed by S. Let $\mathbb{P}$ be the probability described in Theorem 2.2 such that with respect to $\mathbb{P}_i$, the process $(X_t)_{t\geq 0}$ is a stationary continuous time Markov chain with state space S and initial distribution δ_i. Then for any probability mass function $\nu : S \to [0,1]$, with respect to the probability $\mathbb{P}_\nu$ defined by*

$$\mathbb{P}_\nu = \sum_{i\in S} \nu(i)\, \mathbb{P}_i$$

the process $(X_t)_{t\geq 0}$ is a stationary continuous time Markov chain with transition semigroup $(e^{t\mathbf{L}})_{t\geq 0}$ and initial distribution ν.

DEFINITION 2.5. In light of Theorem 2.2, if $\mathbf{L}$ is the Markov generator such that the transition semigroup of $(X_t)_{t\geq 0}$ is $(e^{t\mathbf{L}})_{t\geq 0}$, then we say that $\mathbf{L}$ is the generator of $(X_t)_{t\geq 0}$. △

NOTATION 2.1. Similar to before, corresponding to each Markov semigroup $(\mathbf{P}(t))_{t\geq 0}$ indexed by the set S we will always assume that probability $\mathbb{P}$ as in Theorem 2.2. We will henceforth use the notation $\mathbb{P}_\nu$ to denote the probability in Corollary 2.3 under which $(X_t)_{t\geq 0}$ is a stationary continuous time Markov chain with initial distribution ν; by $\mathbb{E}_\nu$ we will denote the expected value with respect to $\mathbb{P}_\nu$. Further, when $\nu = \delta_j$ for some $j \in S$, then we will continue to write $\mathbb{P}_j$ and $\mathbb{E}_j$ rather than $\mathbb{P}_{\delta_j}$ or $\mathbb{E}_{\delta_j}$. Finally, when we want to view ν as a (row) vector indexed by S, we will write $\vec{\nu}$, and potentially refer to $\vec{\nu}$ as the *starting vector* of the process. △

This following proposition is the continuous time analogue of Theorem **??**.

THEOREM 2.4. *Let $(X_t)_{t\geq 0}$ be a stationary continuous time Markov chain with state space S with transition semigroup $(\mathbf{P}(t))_{t\geq 0}$. If ν is the initial distribution of this process, then*

$$\mathbb{P}_\nu(X_t = i) = [\vec{\nu}\,\mathbf{P}(t)]_i$$

From this, we have another way to classify the transition matrix $\mathbf{P}(t)$ by the way it acts on vectors. This will be useful when generalizing the notions to the case where S is no longer discrete.

THEOREM 2.5. *Suppose that $(X_t)_{t\geq 0}$ is a stationary continuous time Markov chain with state space S and transition semigroup $(\mathbf{P}(t))_{t\geq 0}$. Let $f : S \to \mathbb{R}$ be a bounded function. By $\vec{f}$, denote f as a (column) vector indexed by S; that is, the ith entry of $\vec{f}$ is $f(i)$. For any initial distribution ν, we have*

$$\mathbb{E}_\nu[f(X_t)] = \vec{\nu}\,\mathbf{P}(t)\vec{f}.$$

In particular, for any $j \in S$,

$$\mathbb{E}_j[f(X_t)] = [\mathbf{P}(t)\vec{f}\,]_j.$$

Equivalently, letting $\vec{\mathbb{E}}[f(X_t)]$ be the vector indexed by S such that the jth entry is $\mathbb{E}_j[f(X_t)]$, we have

$$\vec{\mathbb{E}}[f(X_t)] = \mathbf{P}(t)\vec{f}.$$

PROOF. By the definition of expected value and Theorem 2.4,

$$\mathbb{E}_\nu[f(X_t)] = \sum_{k\in S} f(k)\,\mathbb{P}_\nu(X_t = k) = \sum_{k\in S} [\vec{\nu}\,\mathbf{P}(t)]_k f(k) = \vec{\nu}\,\mathbf{P}(t)\vec{f}.$$

If it happens that $\nu = \delta_j$, then $\vec{\nu}\,\mathbf{P}(t)\vec{f}$ gives the jth entry of $\mathbf{P}(t)\vec{f}$ and hence $\mathbb{E}_j[f(X_t)] = [\mathbf{P}(t)\vec{f}\,]_j$. □

We are now at the point to state the theorem which has undoubtedly been foreshadowed by the definitions introduced so far in this chapter.

THEOREM 2.6 (Transition Semigroups are Markov Semigroups). *Let $(X_t)_{t\geq 0}$ be a stationary continuous time Markov chain satisfying RC with state space S and transition semigroup $(\mathbf{P}(t))_{t\geq 0}$. Then, $(\mathbf{P}(t))_{t\geq 0}$ is a Markov semigroup.*

PROOF. Using Theorem 2.5 and Lemma 2.1, for any initial distribution ν and bounded function $f : S \to \mathbb{R}$,

$$\vec{\nu}\,\mathbf{P}(t)\vec{f} = \mathbb{E}_\nu[f(X_t)] \xrightarrow{t\to 0+} \mathbb{E}_\nu[f(X_0)] = \vec{\nu}\,\mathbf{P}(0)\vec{f}$$

Since $[\mathbf{P}(0)]_{ij} = \mathbb{P}_i(X_0 = j) = \delta_{ij}$, we learn that

$$\vec{\nu}\,\mathbf{P}(t)\vec{f} \xrightarrow{t\to 0+} \vec{\nu}\,\mathbf{I}\vec{f}$$

which implies $\lim_{t\to 0+} \mathbf{P}(t) = \mathbf{I} = \mathbf{P}(0)$. Indeed, by taking $\nu = \delta_i$ and $f = \delta_j$, $\vec{\nu}\,\mathbf{P}(t)\vec{f} = [\mathbf{P}(t)]_{ij}$ and $\vec{\nu}\,\mathbf{I}\vec{f} = [\mathbf{I}]_{ij}$, implying that $[\mathbf{P}(t)]_{ij} \to [\mathbf{I}]_{ij}$ as $t \to 0+$.

From here, we show the semigroup property.

$$
\begin{aligned}
&[\mathbf{P}(s+t)]_{i,j} \\
&= \mathbb{P}(X_{t+s} = j | X_0 = i) = \sum_{k \in S} \mathbb{P}(X_{t+s} = j | X_s = k, X_0 = i) \mathbb{P}(X_s = k | X_0 = i) \\
&= \sum_{k \in S} \mathbb{P}(X_{t+s} = j | X_s = k) \mathbb{P}(X_s = k | X_0 = i) \text{ by Markov property} \\
&= \sum_{k \in S} \mathbb{P}(X_t = j | X_0 = k) \mathbb{P}(X_s = k | X_0 = i) \text{ by stationarity} \\
&= \sum_{k \in S} [\mathbf{P}(t)]_{k,j} [\mathbf{P}(s)]_{i,k} = [\mathbf{P}(s)\mathbf{P}(t)]_{i,j}
\end{aligned}
$$

Finally, the fact that $\mathbf{P}(t)$ is a stochastic matrix for every t follows trivially from the fact that the ith row of $\mathbf{P}(t)$ represents the probability mass function of X_t with respect to the probability $\mathbb{P}_i$. □

3. Restarting a Continuous Time Markov Chain

For this section we will continue to assume that $(X_t)_{t \geq 0}$ is a continuous time Markov process with finite state space S and generator $\mathbf{L}$. We also assume that $(X_t)_{t \geq 0}$ is right continuous; i.e., satisfies RC.

THEOREM 3.1. *Let $\mathbb{P}_{\nu_0}$ be the probability such that with respect to $\mathbb{P}_{\nu_0}$, the process $(X_t)_{t \geq 0}$ is a stationary continuous time Markov chain with generator $\mathbf{L}$ and initial distribution ν_0. For any fixed time time $t_1 \geq 0$, let $\nu_1 : S \to [0,1]$ be the mass function $\nu(j) = \mathbb{P}_{\nu_0}(X_{t_1} = j)$. Then with respect to $\mathbb{P}_{\nu_0}$, the process $(X_{t_1+t})_{t \geq 0}$ is again a stationary continuous time Markov chain with generator $\mathbf{L}$ and initial distribution ν_1 (here, the initial distribution of $(X_{t_1+t})_{t \geq 0}$ refers to time $t = 0$ and hence the mass function of X_{t_1}). Further, suppose that for some state j it holds that $\mathbb{P}_{\nu_0}(X_{t_1} = j) > 0$. Then with respect to the probability $\mathbb{P}_{\nu_0}(\cdot \,|\, X_{t_1} = j)$, the process $(X_{t_1+t})_{t \geq 0}$ is a stationary continuous time Markov chain with generator $\mathbf{L}$ and initial distribution δ_j, and is independent of the past $(X_t)_{0 \leq t \leq t_1}$.*

DEFINITION 3.1. A random variable τ taking values in $[0, \infty]$ is called a *stopping time* for $(X_t)_{t \geq 0}$ whenever the event $\{\tau \leq s\}$ depends only on the process X_t for values of $t \leq s$. Just as before, the poetic interpretation is that τ is a stopping time whenever the value of τ depends only on past and present values of the process, but does not depend on future values. △

As in the discrete time case, a stopping time in the continuous time can still be understood as a realistic stopping strategy. The following intuitive example is to help "visualize" the distinguished behavior of a stopping time.

EXAMPLE 3.1. Suppose that you're a store owner and one day you decide to count the number of customers entering the store. So, you astutely let N_t be the number of customers who have entered your store by time t. Now, you have to decide when to close your store. One of your two children, Cora, makes the following suggestion, "we should close the store after the 5th customer arrives." Your other delinquent child, Tom, suggests an alternative, "we should close the store after the last customer arrives." At first glance, one might believe that Tom's suggestion is better, since if you wait until the last customer arrives to close, then you potentially make more money than closing after the 5th arrival. However, how will you ever decide which customer is the last? You would certainly have to see into the future to decide whether or not another customer is going to enter. So, Tom's suggestion – while pleasant to hope for – will give a closing time which is not a realistic strategy (unless you can see into the future). Cora's suggestion on the other hand is a perfectly realistic strategy. To decide whether or not the arrival of a customer is the 5th arrival, you need only the information up to the point of the 5th arrival; you won't need to be a fortune teller at all. So the closing time that Cora suggests is in fact a stopping time with respect to $(N_t)_{t\geq 0}$.

To state the above scenario in slightly more mathematical detail, one might assume that $(N_t)_{t\geq 0}$ is a Markov process with state space $\mathbb{N}_0 = \{0, 1, 2, ...\}$. Cora's suggested closing time is given by $\tau_5 = \inf\{t \geq 0 : N_t = 5\}$ whereas Tom's closing time is given by $\tau_L = \inf\{t : N_t = N_{t+s} \text{ for all } s \geq 0\}$. Notice that the event $\{\tau_5 \leq t\}$ depends only on the process up until time t, you needn't check the customers behaviors after time t. However, the event $\{\tau_L \leq t\}$ needs you to compare the number of arrivals at time t, represented by N_t, with the number of arrivals after time t, represented by N_{t+s}. Therefore τ_5 is a stopping time, whereas τ_L is not. △

DEFINITION 3.2. We will say that $(X_t)_{t\geq 0}$ satisfies the *strong Markov property* whenever the following holds. *Given any stopping time T and state $i \in S$, with respect to the probability $\mathbb{P}(\cdot \,|\, \tau < \infty, X_\tau = i)$, the process $(X_{\tau+t})_{t\geq 0}$ is again a stationary continuous time Markov chain with generator $\mathbf{L}$ and initial distribution δ_i, which is independent of the past states $(X_t)_{t\leq\tau}$.* △

Intuitively, what the strong Markov property allows us to do is treat a stopping time τ as if it were a fixed time, instead of a random variable with regards to the process $(X_t)_{t\geq 0}$. The first property simply says that if we know that we are in state i at time τ, then watching the evolution of the chain after time τ would be the same as watching the evolution of $(X_t)_{t\geq 0}$ if it had started in

state i to begin with. The second property says that if you know the state of the chain at stopping time τ (again called the "present", even though the time is random rather than fixed), then the future path traveled by the process is independent of the past path the process took to arrive at it's present position. These are properties that you have hopefully come to associate with Markov processes, so hopefully aren't too mysterious. Unlike in discrete time, the strong Markov property is more constraining than the standard Markov property. However, the examples where the strong Markov property fails will not bother us much.

THEOREM 3.2. *Any right continuous stationary continuous time Markov chain $(X_t)_{t\geq 0}$ with only finitely many states satisfies the strong Markov property.*

The interested reader can find a version of the above theorem in [**?**, Theorem 9.1.12].

4. The Embedded Discrete Time Markov Chain: The Jump-Hold Description

This section sets out to prove a very useful (intuitively and calculationally) description of how a stationary continuous time Markov chain behaves. We suppose that $(X_t)_{t\geq 0}$ is a stationary continuous time Markov chain with generator $\mathbf{L}$ indexed by the finite state space S.

The idea is as follows.

The Jump-Hold Idea: Let $(X_t)_{t\geq 0}$ be a stationary continuous time Markov chain with state space S and generator $\mathbf{L}$. There exists a stochastic matrix $\mathbf{R}$ corresponding to $\mathbf{L}$ such that the "jumping behavior" of the continuous time Markov chain is given by that of a discrete time Markov chain with transition matrix $\mathbf{R}$. That is, if we ignore how long the chain is in a particular state and only pay attention to its behavior when jumping between states, it looks like a discrete time Markov chain with transition matrix $\mathbf{R}$. Conversely, if we know what state the continuous time Markov chain is currently in, then the amount of time the chain will "hold" before making its next jump is given by an exponential random variable with parameter depending on its current state. Moreover, if we know what state the chain is currently in, then the time the chain holds until its next jump, and the location of its next jump are independent. Simply, a continuous time Markov chain acts as a discrete time Markov chain which, once it arrives in state, stays in that state for a certain exponentially distributed period of time before jumping again to a new state, which will be chosen independently from the time it waited to jump.

DEFINITION 4.1. Let $\mathbf{L}$ be a Markov generator indexed by S. For each $i \in S$, let $\lambda_i = -[\mathbf{L}]_{ii}$ be the magnitude of the ith diagonal element of $\mathbf{L}$. The *embedded stochastic matrix* $\mathbf{R}$ is the matrix

indexed by S defined by:

$$i \neq j: \ [\mathbf{R}]_{ij} = \begin{cases} \dfrac{[\mathbf{L}]_{ij}}{\lambda_i} & \lambda_i \neq 0 \\ 0 & \lambda_i = 0 \end{cases}$$

$$i = j: \ [\mathbf{R}]_{ii} = \begin{cases} 0 & \lambda_i \neq 0 \\ 1 & \lambda_i = 0 \end{cases}$$

△

EXAMPLE 4.1. Let $\mathbf{L}$ be given by

$$\mathbf{L} = \begin{pmatrix} -3 & 1 & 2 \\ 1 & -2 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

indexed by $S = \{1, 2, 3\}$. Then $\lambda_1 = 3$, $\lambda_2 = 2$, and $\lambda_3 = 0$. From here, we find $\mathbf{R}$ to be,

$$\mathbf{R} = \begin{pmatrix} 0 & 1/3 & 2/3 \\ 1/2 & 0 & 1/2 \\ 0 & 0 & 1 \end{pmatrix}$$

You will notice something in this case which is true in general: $\mathbf{R}$ is a stochastic matrix. △

LEMMA 4.1. *For any Markov generator* $\mathbf{L}$*, the corresponding discrete time transition matrix* $\mathbf{R}$ *is a stochastic matrix.*

PROOF. We need to show that the elements in $\mathbf{R}$ are all non-negative, and that each row of $\mathbf{R}$ have elements which sum to 1. To start, let S be the indexing set for $\mathbf{L}$. If $i \in S$ is such that $\lambda_i = 0$, then every element in that row is 0 except the diagonal element which is 1. So, we really only need to check with $\lambda_i \neq 0$. In this case, we have $[\mathbf{R}]_{ij} = [\mathbf{L}]_{ij}/\lambda_i$ when $i \neq j$, and $[\mathbf{R}]_{ii} = 0$. Since for $i \neq j$, $[\mathbf{L}]_{i,j} \geq 0$ and $\lambda_i > 0$, we immediately have that every element is non-negative. Moreover,

$$\sum_{j \in S} [\mathbf{R}]_{ij} = 0 + \sum_{i \neq j} [\mathbf{R}]_{ij} = \sum_{j \neq i} \frac{[\mathbf{L}]_{ij}}{\lambda_i} = \frac{\sum_{j \neq i} [\mathbf{L}]_{ij}}{\lambda_i} = 1$$

where the last equality holds since the sum of the elements along the i^{th} row of $\mathbf{L}$ equal zero, hence,

$$0 = \sum_{j \in S} [\mathbf{L}]_{ij} = -\lambda_i + \sum_{i \neq j} [\mathbf{L}]_{ij}$$

implying that $\sum\limits_{i \neq j} [\mathbf{L}]_{ij} = \lambda_i$. □

At this point we are ready to make rigorous the Jump-Hold idea with several theorems. Before this we need some preliminary definitions.

DEFINITION 4.2. $(X_t)_{t\geq 0}$ be a stationary continuous time Markov chain with state space S. The sequence of stopping times stopping times $J_0, J_1, J_2, ...$ defined recursively by

$$J_{k+1} = \inf\{t > J_k \text{ s.t. } X_t \neq X_{J_k}\}$$

with J_0 defined to be 0 are called the *jump times* of the process. The random variables $S_0, S_1, S_2, ...$ defined by $S_k = J_{k+1} - J_k$ are called the *sojourn times* or *interim times*. △

REMARK 4.1. Note that for each k, the jump time J_k gives the time at which the process makes it kth jump. Related, the sojourn time $S_k = J_{k+1} - J_k$ gives the amount of time the process occupies the state it lands in on the kth jump before making its next jump. Further, recognize that since $S_k = J_{k+1} - J_k$, it also holds that $J_{k+1} = S_0 + S_1 + \cdots + S_k$. △

For the following "jump-hold theorems," suppose $(X_t)_{t\geq 0}$ be a stationary continuous time Markov chain satisfying RC with finite state space S. Let $\mathbf{L}$ be the generator of this process, $\mathbf{R}$ the corresponding embedded stochastic matrix, $J_0, J_1, J_2, ...$ be the jump times, and $S_0, S_1, S_2, ...$ the sojourn times. Further, continuing with our standard notation, we let $\lambda_i = -[\mathbf{L}]_{ii}$ for each $i \in S$.

THEOREM 4.2 (Deconstructing a Stationary Continuous Time Markov Chain). *The following hold.*

(1) *The discrete time process $(X_{J_n})_{n=0}^{\infty}$ is a stationary discrete time Markov chain with transition matrix $\mathbf{R}$.*
(2) *For any $i \in S$, conditioned on the event $X_{J_k} = i$, the sojourn time S_k is an exponential random variable with parameter λ_k. Further, for any states $i_0, i_1, ..., i_k \in S$, conditioned on the event $\{X_0 = i_0, X_1 = i_1, ..., X_k = i_k\}$, the variables $S_0, S_1, ..., S_k$ are independent*
(3) *Given the current location of the process, the length of time until the next jump and the state the process jumps to next are independent. That is, for any state $i \in S$ and any jump time J_k, given the event $\{X_{J_k} = i\}$, the variables S_k and $X_{J_{k+1}}$ are independent. In particular,*

$$\begin{aligned}&\mathbb{P}(S_k > t, X_{J_{k+1}} = j \,|\, X_{J_k} = i)\\ &\quad = \mathbb{P}(S_k > t \,|\, X_{J_k} = i)\mathbb{P}(X_{J_{k+1}} = j \,|\, X_{J_k} = i) = e^{-\lambda_i t}\mathbf{R}_{ij}\end{aligned}$$

for any states $i, j \in S$.

THEOREM 4.3 (Constructing a Stationary Continuous Time Markov Chain). *Suppose that $(Y_n)_{n=0}^{\infty}$ is a stationary discrete time Markov chain with transition matrix $\mathbf{R}$. Define the random variables $S_0, S_1, S_2, ...$ in such a way that for any states $i_0, i_1, ..., i_k \in S$, conditioned on the event $\{Y_0 = i_0, Y_1 = i_1, ..., Y_k = i_k\}$, it holds that $S_0, S_1, ..., S_k$ are independent and each S_m is an exponential random variable with parameter λ_{i_m} for $m = 1, 2, ..., k$. Define $J_0 = 0$ and $J_{k+1} = S_0 + S_1 + \cdots + S_k$. Further, define the counting process $(N_t)_{t\geq 0}$ by,*

$$N_t = \max\{k \geq 0 \quad s.t. \quad S_0 + S_1 + \cdots + S_{k-1} \leq t\} = \max\{k \geq 0 \quad s.t. \quad J_k \leq t\}$$

Then the continuous time process $(X_t)_{t\geq 0}$ defined by $X_t = Y_{N_t}$ is a continuous time Markov process with infinitesimal generator $\mathbf{L}$. Moreover, $S_0, S_1, S_2, ...$ are the sojourn times and $J_0, J_1, J_2, ...$ are the jump times of the process $(X_t)_{t\geq 0}$; further, N_t represents the number of jumps made by the process $(X_t)_{t\geq 0}$ by time t.

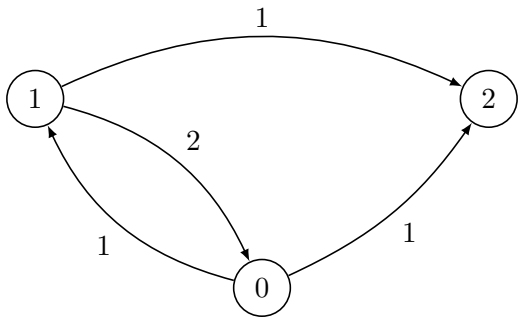
5. Rate Diagrams

Suppose that $\mathbf{L}$ is a Markov generator indexed by the finite set S. Similar to a jump diagram corresponding to a stochastic matrix, a Markov generator has an associated *rate diagram*. As with the jump diagrams, the rate diagram organizes the states S as the vertices of a graph and uses arrows as edges in the graph, connecting together appropriate states. In the rate diagram, two edges i and j in S are connected by an arrow from i to j whenever $[\mathbf{L}]_{ij}$ is positive; in this case, the weight we give to the arrow from i to j is simply $[\mathbf{L}]_{ij}$. Note that since the diagonal elements of $\mathbf{L}$ are always non-positive, there is never an arrow drawn from any edge to itself.

EXAMPLE 5.1. Suppose we are given the Markov generator

$$\mathbf{L} = \begin{pmatrix} -2 & 1 & 1 \\ 2 & -3 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

indexed by $S = \{0, 1, 2\}$. Then the corresponding rate diagram is

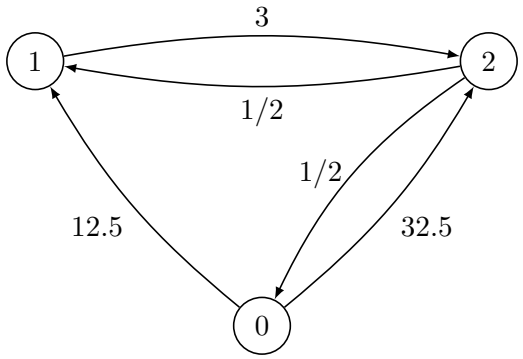


△

EXAMPLE 5.2. Suppose we are given the Markov generator

$$\mathbf{L} = \begin{pmatrix} -45 & 12.5 & 32.5 \\ 0 & -3 & 3 \\ 1/2 & 1/2 & -1 \end{pmatrix}$$

indexed by $S = \{0, 1, 2\}$. Then the corresponding rate diagram is

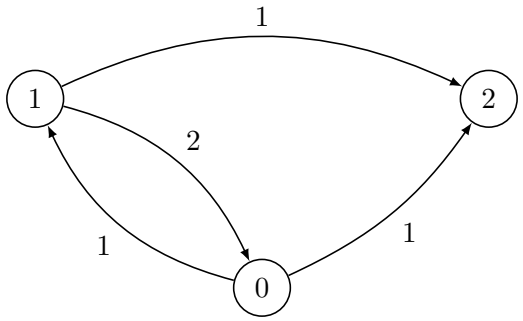


△

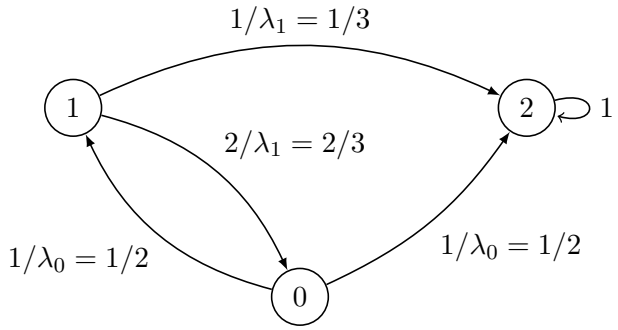
It is important to note here that there is a one-to-one correspondence between a rate diagram and a Markov generator $\mathbf{L}$. That is, given a Markov generator $\mathbf{L}$, you can create the corresponding rate diagram; given a rate diagram, you can construct the corresponding generator $\mathbf{L}$. Indeed, once given a rate diagram, you know all off-diagonal entries of the corresponding generator $\mathbf{L}$; since the sum along each row of $\mathbf{L}$ must be 0, you can then deduce the diagonal elements as well.

A very useful aspect of rate diagrams is that we can easily infer the jump-hold behavior of corresponding process $(X_t)_{t\geq 0}$ whose generator corresponds to the jump diagram. Indeed, if $i \in S$ and λ_i is the sum of all arrow weights emanating from i, then the corresponding jump diagram of $\mathbf{R}$ will have matching arrows emanating from i, simply scaled by $1/\lambda_i$; this is unless $\lambda_i = 0$, in which case there will be one arrow from i to itself with weight 1. Moreover, the exponential "clock" at state i, the "clock" which tells the process how long to wait in state i before jumping again, will have parameter λ_i.

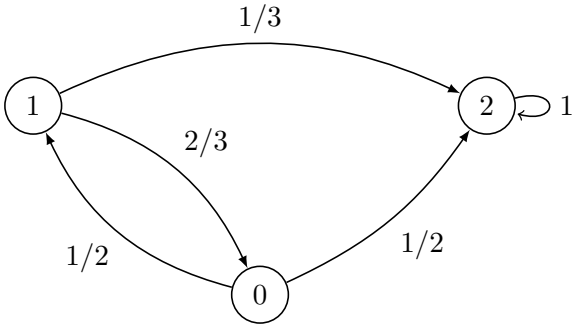
EXAMPLE 5.3. Suppose we are given the rate diagram



Then $\lambda_0 = 1 + 1 = 2$, $\lambda_1 = 1 + 2 = 3$, and $\lambda_2 = 0$. Therefore, the jump diagram of the embedded stochastic matrix $\mathbf{R}$ is



Written neatly,



Moreover, we also have that at 0, there is an exponential $\lambda_0 = 2$ clock; at state 1 there is an exponential $\lambda_1 = 3$ clock; at state 2 there is a $\lambda_2 = 0$ clock. Note that we define an Exp(0) random variable to be constantly ∞ (i.e., the clock never rings). △

6. Invariant Distributions and Long Run Results

Throughout this section, let $(X_t)_{t\geq 0}$ be a right continuous stationary continuous time Markov chain with finite state space S and transition semigroup $(\mathbf{P}(t))_{t\geq 0}$ with generator $\mathbf{L}$. As usual, we let $\lambda_i = -[\mathbf{L}]_{ii}$ for each $i \in S$.

Definition 6.1. A probability mass function $\pi : S \to [0, 1]$ is called an *invariant distribution* for $(\mathbf{P}(t))_{t\geq 0}$ when $\vec{\pi}\,\mathbf{P}(t) = \vec{\pi}$ for every $t \geq 0$. △

PROPOSITION 6.1. *A probability mass function $\pi : S \to [0,1]$ is an invariant distribution for $(\mathbf{P}(t))_{t\geq 0}$ if and only if $\vec{\pi}\,\mathbf{L} = \vec{\mathbf{0}}$, where $\vec{\mathbf{0}}$ is the vector of all zeros. In other words, π is an invariant distribution for $(\mathbf{P}(t))_{t\geq 0}$ if and only if $\vec{\pi}$ is a left eigenvector of $\mathbf{L}$ with eigenvalue 0.*

PROOF. Suppose π is invariant for $(\mathbf{P}(t))_{t\geq 0}$. Then, since $\mathbf{P}(t) = e^{t\mathbf{L}}$, it must be that $\vec{\pi}\, e^{t\mathbf{L}} = \vec{\pi}$. Taking the derivative of the lefthand side of the last equality gives

$$\frac{d}{dt}\Big|_0 \left(\vec{\pi} e^{t\mathbf{L}}\right) = \vec{\pi}\mathbf{L},$$

while taking the derivative of the righthand side (i.e., constant vector π) results in $\vec{\mathbf{0}}$. Therefore, if π is an invariant distribution, we get

$$\vec{\pi}\,\mathbf{L} = \vec{\mathbf{0}}.$$

For the converse, suppose that $\vec{\pi}\,\mathbf{L} = \vec{\mathbf{0}}$. Then,

$$\vec{\mathbf{0}} = \vec{\pi}\,\mathbf{L}\, e^{t\mathbf{L}} = \frac{d}{dt}\left(\vec{\pi}\, e^{t\mathbf{L}}\right).$$

Since the derivative is $\vec{\mathbf{0}}$, this proves that $\vec{\pi}\, e^{t\mathbf{L}}$ is a constant vector. Since for $t = 0$, $\vec{\pi}\, e^{0\,\mathbf{L}} = \vec{\pi}$, it holds that $\vec{\pi}\, e^{t\mathbf{L}} = \vec{\pi}$ for every t. □

DEFINITION 6.2. The Markov semigroup $(\mathbf{P}(t))_{t\geq 0}$ (or the corresponding Markov chain) is called *irreducible* whenever the embedded stochastic matrix $\mathbf{R}$ is irreducible. Intuitively, this says that $(\mathbf{P}(t))_{t\geq 0}$ is irreducible whenever the underlying jump diagram of the chain is irreducible. △

LEMMA 6.2. *The following are equivalent.*

1. *Markov semigroup $(\mathbf{P}(t))_{t\geq 0}$ is irreducible.*
2. *There exists some $t > 0$ such that $[\mathbf{P}(t)]_{ij} > 0$ for every $i, j \in S$.*
3. *For all times $t > 0$, it holds that $[\mathbf{P}(t)]_{ij} > 0$ for every $i, j \in S$.*

PROOF. The underlying jump diagram is irreducible if and only if there is a positive probability of getting from any state i to any other state j, which happens if and only if there is some $t > 0$ such that $[\mathbf{P}(t)]_{ij} > 0$. However, we have proved that if $[\mathbf{P}(t)]_{ij} > 0$ for any $t > 0$, then $[\mathbf{P}(t)]_{ij} > 0$ for every t. Since this is true for every $i, j \in S$, the result now follows. □

DEFINITION 6.3. For each state $i \in S$, let ρ_i be the *first return time* to i, defined by

$$\rho_i = \inf\{t \geq J_1 \ \text{ s.t. } \ X_t = i\}$$

where, as before, J_1 is the first jump time of the process. △

THEOREM 6.3. *Suppose that $(\mathbf{P}(t))_{t\geq 0}$ is irreducible. Then, with the assumption that the state space S is finite, we are guaranteed that $\mathbb{E}_i[\rho_i] < \infty$ for every $i \in S$.*

PROOF. If there is only one state, then $\mathbf{L} = (0)$ and there is nothing to prove. So, let's assume that there is more than one state. Let λ be the smallest of the parameters $\lambda_i = -[\mathbf{L}]_{ii}$ dictating the behavior of the exponential clocks at each state i. Since we are assuming irreducibility, $\lambda > 0$. The alternate process whose jump behavior is still given by $\mathbf{R}$, but the "hold behavior" is controlled by exponential clocks with parameter λ at every state jumps like the process we are considering, but on average holds in each state as long or longer than the original process (recall that the smaller the parameter λ of an exponential random variable, the larger the expected value $1/\lambda$). In particular, for this alternate process $\mathbb{E}_i[\rho_i]$ will be as large or larger than for the original process. However, since the expected time of holding in each state is $1/\lambda$ for this alternate process, it is easy to deduce that $\mathbb{E}_i[\rho_i]$ for the alternate process is equal to the expected return time for the embedded discrete time process (which is finite as we proved in the theory of discrete time Markov chains) multiplied by the expected hold time $1/\lambda$ in each state. Therefore $\mathbb{E}_i[\rho_i]$ is finite for the alternate process, and hence is finite for the original process. □

NOTATION 6.1. Similar to before, we will reserve the symbol π as the function $\pi : S \to [0, 1]$ where $\pi(i)$ is defined as

$$\pi(i) = \frac{1}{\lambda_i \, \mathbb{E}_i[\rho_i]}$$

for each $i \in S$. △

THEOREM 6.4. *Suppose that $(\mathbf{P}(t))_{t\geq 0}$ is irreducible and S is finite. Then*

1. *$\pi(i) > 0$ for every $i \in S$ and $\sum_{i\in S} \pi(i) = 1$.*
2. *There exists exactly one invariant distribution for $(\mathbf{P}(t))_{t\geq 0}$, and that invariant distribution is π.*
3. *$\lim_{t\to\infty} [\mathbf{P}(t)]_{ij} = \pi(j)$ for all states $i, j \in S$. In other words, the long time limit of $\mathbf{P}(t)$ results in the matrix $\mathbf{\Pi}$ in which every row is the vector $\vec{\pi}$.*
4. *Regardless of initial distribution, the long run fraction of time the process is in state j is $\pi(j)$. In symbols, given any initial distribution ν, we have*

$$\mathbb{P}_\nu\left(\lim_{T\to\infty} \frac{1}{T}\int_0^T 1_{\{X_t=j\}}\, dt = \pi(j)\right) = 1.$$

7. Reducing a Continuous Time Markov Chain

In the discrete time setting, we learned in Section **??** that we can focus our results for irreducible transition matrices $\mathbf{P}$ to the reduced transition matrices $\mathbf{P}_C$ for each closed communication class C. Here we make the analogous observations, mostly unproved since the arguments follow those in Section **??** and hence are redundant here. For the remainder of this section, we assume that $(\mathbf{P}(t))_{t\geq 0}$ is a Markov semigroup indexed by the finite state space S with generator $\mathbf{L}$.

DEFINITION 7.1. We define a subset $C \subset S$ as a (open/closed/recurrent/transient) *communication class* whenever C is a (open/closed/recurrent/transient) communication class of the embedding stochastic jump matrix $\mathbf{R}$. △

REMARK 7.1. Note that since we are currently assuming that S is finite, a communication class $C \subset S$ is positive recurrent if and only if it is recurrent, which happens if and only if it is closed. Therefore, we did not bother worrying about the extended adjectives "positive" or "null" prefixing the label "recurrent." △

NOTATION 7.1. Let $C \subset S$. We let $\mathbf{P}_C(t)$ be the reduced matrix of $\mathbf{P}(t)$ indexed only by the elements in C. △

From the jump-hold behavior of a continuous time Markov chain, we recover the following theorem.

THEOREM 7.1. *Suppose that $(X_t)_{t\geq 0}$ is the stationary continuous time Markov chain with transition matrix $(\mathbf{P}(t))_{t\geq 0}$. Suppose further that C is a closed communication class and that $\nu : S \to [0,1]$ is a probability mass function such that $\nu(j) = 0$ for any $j \notin C$. If the initial distribution of $(X_t)_{t\geq 0}$ is ν, then the process evolves as an irreducible stationary continuous time Markov chain with state space C, transition semigroup $(\mathbf{P}_C(t))_{t\geq 0}$, and initial distribution given by the restriction of ν to C.*

An immediate corollary to this is the following.

COROLLARY 7.2. *The subset $C \subset S$ is a closed communication class if and only if $(\mathbf{P}_C(t))_{t\geq 0}$ is a Markov semigroup.*

We further recover the following results, whose proofs are nearly identical to those given in Section **??**.

LEMMA 7.3. *If C is a closed communication class, then $(\mathbf{P}_C(t))_{t\geq 0}$ is irreducible.*

PROOF. See Lemma **??** □

LEMMA 7.4. *Suppose that there exist exactly N distinct closed communication classes $C_1, C_2, ..., C_N$ and M open communication classes $O_1, O_2, ..., 0_M$ for $(\mathbf{P}(t))_{t\geq 0}$. Then by reordering the states S, we can arrange $(\mathbf{P}(t))_{t\geq 0}$ such to have the block-matrix form*

$$\mathbf{P}(t) = \begin{pmatrix} \mathbf{P}_{C_1}(t) & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{P}_{C_2}(t) & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & \cdots & \mathbf{P}_{C_N}(t) & 0 & 0 & \cdots & 0 \\ * & * & * & * & \mathbf{P}_{O_1}(t) & * & \cdots & * \\ * & * & * & * & * & \mathbf{P}_{O_2}(t) & \cdots & * \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ * & * & * & * & * & \cdots & * & \mathbf{P}_{O_M}(t) \end{pmatrix}$$

where $*$ symbolizes a potentially non-zero entry.

PROOF. See Lemma **??**. □

REMARK 7.2. As before, there is no theoretical benefit of reordering $\mathbf{P}(t)$ as we did in Lemma 7.4. However, it does give a visual intuition and insight in recognizing how the communication classes and the transition semigroup are inner-organized. △

THEOREM 7.5. *Suppose that C is a closed communication class. Then the results of Theorem 6.4 hold for $(\mathbf{P}_C(t))_{t\geq 0}$.*

PROOF. By the assumption that C is closed, then $\mathbf{P}_C(t))_{t\geq 0}$ is an irreducible Markov semigroup. Therefore, we can simply apply the results of Theorem 6.4. □

LEMMA 7.6. *Suppose that C is a closed communication class. Then there exists an invariant distribution $\pi_C : C \to [0,1]$ for $(\mathbf{P}_C(t))_{t\geq 0}$. Moreover, by extending π_C to a probability mass function on $\pi : S \to [0,1]$ by defining $\pi(j) = \pi_C(j)$ for $j \in C$ and $\pi(i) = 0$ for all $i \notin C$, then π is an invariant distribution for $(\mathbf{P}(t))_{t\geq 0}$.*

PROOF. See Lemma **??**. □

DEFINITION 7.2. Suppose that C is a closed communication class. Let π_C be the invariant distribution for $(\mathbf{P}_C(t))_{t\geq 0}$ and define π as the extension of π_C to $\mathbf{P}$ as in Lemma 7.6. We will say that π is a *canonical invariant distribution* for $(\mathbf{P}(t))_{t\geq 0}$ generated by C. △

THEOREM 7.7. *Suppose that* $(\mathbf{P}(t))_{t\geq 0}$ *has* N *distinct closed communication classes. Then the* N *canonical invariant distributions* $\pi_1, \pi_2, ..., \pi_N$ *for* $(\mathbf{P}(t))_{t\geq 0}$ *generated by these communication classes are mutually orthogonal. Moreover, if* $\theta_1, ..., \theta_N$ *are non-negative numbers which sum to* 1*, then* $\theta_1\,\pi_1 + \cdots + \theta_N\,\pi_N$ *is also an invariant distribution.*

8. When the State Space is Discrete and Infinite

Most of the previous results we've thus far presented in the setting of a finite state space S still hold over when S is discrete and has infinitely many states as long as we make the following assumption.

ASSUMPTION 8.1. Suppose that S is discrete and infinite. We assume that a stationary continuous time Markov chain $(X_t)_{t\geq 0}$ with state space S is right continuous (satisfies RC), satisfies the strong Markov property, and that there is a Markov generator $\mathbf{L}$ such that the transition semigroup of $(X_t)_{t\geq 0}$ is generated by $\mathbf{L}$. △

With this assumption in tow, the only previous results we need to alter from their immediate analogue in the present infinite S setting are those which assume that a closed communication class is positive recurrent. Certainly a closed communication class with only finitely many elements is positive recurrent, but not all closed communication classes have only finitely many elements here, so we need to concern ourselves with the possibility of positive and null recurrence as well.

DEFINITION 8.1. Let $i \in S$ be recurrent (i.e., belongs to a recurrent communication class) and ρ_i be the first return time of $(X_t)_{t\geq 0}$ to i (i.e., $\rho_i = \inf\{t > J_1 \text{ s.t. } X_t = i\}$). If $\mathbb{E}_i[\rho_i] < \infty$ then we say that i is *positive recurrent*; otherwise, if $\mathbb{E}_i[\rho_i] = \infty$, we say that i is *null recurrent*. △

You will notice in Theorem 6.4, one of the results is that $\pi(i) = \frac{1}{\lambda_i\,\mathbb{E}_i[\rho_i]} > 0$ for every i whenever $(\mathbf{P}(t))_{t\geq 0}$ is irreducible. This is using the fact that $\mathbb{E}_i[\rho_i] < \infty$ since all states will be positive recurrent as there are only finitely many states (belonging to the closed communication class S). Now that we have the very real possibility that $\mathbb{E}_i[\rho_i] = \infty$ (either if i is transient or null recurrent), we need to make some adjustments to the statement of Theorem 6.4.

REMARK 8.1. We will continue to define $\pi(i) = \frac{1}{\lambda_i\mathbb{E}_i[\rho_i]}$ here, defined as 0 when $\mathbb{E}_i[\rho_i] = \infty$. △

THEOREM 8.1. *Suppose that* $(\mathbf{P}(t))_{t\geq 0}$ *is irreducible. Then*

1. *If* $\pi(i) > 0$ *for any* $i \in S$*, then* $\pi(i) > 0$ *for every* $i \in S$*, and in this case,* $\sum_{i\in S}\pi(i) = 1$.
2. *There exists exactly one invariant distribution for* $(\mathbf{P}(t))_{t\geq 0}$ *if and only if it is positive recurrent. In this case, the unique invariant distribution is* π.

(3) $\lim_{t\to\infty}[\mathbf{P}(t)]_{ij} = \pi(j)$ *for all states* $i, j \in S$. *In other words, the long time componentwise limit of* $\mathbf{P}(t)$ *results in the matrix* $\mathbf{\Pi}$ *in which every row is the vector* $\vec{\pi}$.

(4) *Regardless of initial distribution, the long run fraction of time the process is in state* j *is* $\pi(j)$. *In symbols, given any initial distribution* ν, *we have*

$$\mathbb{P}_\nu\left(\lim_{T\to\infty}\frac{1}{T}\int_0^T 1_{\{X_t=j\}}\,dt = \pi(j)\right) = 1.$$

REMARK 8.2. Since we are now in the setting that it might hold that $\pi(i) = 0$, there is a possibility that the limiting in Theorem 8.1 (3) is to 0. Similarly, in Theorem 8.1 (4), it is possible that the long run fraction of time spent in a particular state is 0. △

From this, we immediately can re-deduce some previous results regrading the reduction of a non-irreducible process.

THEOREM 8.2. *The results of Theorem 7.5, Lemma 7.6, and Theorem 7.7 still hold true when the assumption of a communication class being closed closed communication class is replaced by the assumption that a communication class is closed and positive recurrent.*

9. Exercises

(1) Exponentiate the matrix

$$\mathbf{L} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$$

to find the Markov semigroup $(\mathbf{P}(t))_{t\geq 0}$ with generator $\mathbf{L}$.

(2) Consider the following time-dependent matrices $(\mathbf{P}(t))_{t\geq 0}$

$$\mathbf{P}(t) = \begin{pmatrix} e^{-2t} & \frac{1}{2}(1-e^{-2t}) & \frac{1}{2}(1-e^{-2t}) \\ \frac{t^2}{4(t^2+1)} & \frac{1}{t^2+1} & \frac{3t^2}{4(t^2+1)} \\ \frac{2-2e^{-t/2}}{3} & \frac{1-e^{-t/2}}{3} & e^{-t/2} \end{pmatrix}$$

Define $\mathbf{L} = \frac{d}{dt}\big|_{t=0}\mathbf{P}(t)$, and explain why $\mathbf{L}$ is not an Markov generator and therefore $(\mathbf{P}(t))_{t\geq 0}$ is not a Markov semigroup.

(3) Find the generator $\mathbf{L}$ of the Markov semigroup $(\mathbf{P}(t))_{t\geq 0}$, defined by

$$\mathbf{P}(t) = \begin{pmatrix} \frac{2+e^{-3t}}{3} & \frac{1-e^{-3t}}{3} \\ \frac{2-2e^{-3t}}{3} & \frac{1+2e^{-3t}}{3} \end{pmatrix}.$$

(4) A frog is jumping around on four lilly pads. Assuming that the frog initially started jumping on the lilly pads at time 0, let X_t be the current lilly pad the frog is on at time

t; number the lilly pads so that the state space of this process $(X_t)_{t=0}^{\infty}$ is $S = \{1, 2, 3, 4\}$. Assume that $(X_t)_{t\geq 0}$ is a stationary continuous time Markov chain.

(a) Justify why it is a reasonable assumption that $(X_t)_{t\geq 0}$ is right continuous (i.e., satisfies RC).

(b) Suppose that the generator of this process is

$$\mathbf{L} = \begin{pmatrix} -2 & 1 & 0 & 1 \\ 1 & -3 & 1 & 1 \\ 2 & 0 & -3 & 1 \\ 1 & 1 & 0 & -2 \end{pmatrix}$$

Find the embedded stochastic matrix $\mathbf{R}$.

(c) Draw the rate diagram and the embedded jump diagram.

(d) You look and notice that the frog is currently on lilly pad 3. You look again five minutes later and the frog is again on lilly pad 3. What is the probability that the frog did not leave lilly pad 3 during that 5 minute period? (Assume that the rates from $\mathbf{L}$ are in minutes).

(e) Getting bored, you decide that you won't watch the frog for much longer. You decide to stop watching with one of two stopping strategies: either leave at time α, the fifth time from now that the frog lands on lilly pad 2; or leave at time β, the second to last time the frog jumps to lilly pad 1 within the next twenty minutes. Which time α or β is a stopping time? Why is it so, and why is the other not?

(5) Let $(\mathbf{P}(t))_{t\geq 0}$ be a Markov semigroup with generator $\mathbf{L}$ and embedded stochastic matrix $\mathbf{R}$. Recall that the jump-hold description tells us that the corresponding Markov chain $(X_t)_{t\geq 0}$ has "jump behavior" governed by the stochastic matrix $\mathbf{R}$. Justify that for states $i, j \in S$, we have that $i \to j$ with respect to $\mathbf{R}$ (i.e., with respect to the embedded jump behavior) if and only if there exists some $t > 0$ such that $[\mathbf{P}(t)]_{ij} > 0$.

(6) Consider the state space $S = \{1, 2, 3\}$ and the Markov generator

$$\mathbf{L} = \begin{pmatrix} -2 & 1 & 1 \\ 2 & -3 & 1 \\ 1 & 0 & -1 \end{pmatrix}$$

(a) Confirm that the corresponding semigroup $(\mathbf{P}(t))_{t\geq 0}$ is irreducible.

(b) Find an invariant distribution for $(\mathbf{P}(t))_{t\geq 0}$.

(c) Find $\lim_{t\to\infty} \mathbf{P}(t)$.

(d) Find the long run fraction of time the process is in state 2.

(e) For each state $i \in S$, find $\mathbb{E}_i[\rho_i]$.

(7) Let $\mathbf{L}$ be the following Markov generator

$$\mathbf{L} = \begin{pmatrix} -1 & 1 \\ 2 & -2 \end{pmatrix}$$

indexed by $S = \{0, 2\}$. Note that we can diagonalize $\mathbf{L}$ as

$$\mathbf{L} = \begin{pmatrix} 1 & 1 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & -3 \end{pmatrix} \begin{pmatrix} 2/3 & 1/3 \\ 1/3 & -1/3 \end{pmatrix}$$

(a) Confirm that the corresponding Markov semigroup is

$$\mathbf{P}(t) = \begin{pmatrix} \frac{2+e^{-3t}}{3} & \frac{1-e^{-3t}}{3} \\ \frac{2-2e^{-3t}}{3} & \frac{1+2e^{-3t}}{3} \end{pmatrix}$$

(b) Draw the rate and embedded jump diagrams corresponding to $\mathbf{L}$. Also, find the parameters λ_i for each exponential clock at each state $i \in S$.

(c) Suppose that $(X_t)_{t\geq 0}$ is the corresponding continuous time Markov chain. Find $\mathbb{P}_\nu(X_{2.5} > X_3)$ where $\vec{\nu} = (1/2, 1/2)$.

(d) As a function of time t, find $\mathbb{E}_\nu[X_t]$ and $\mathrm{Var}_\nu(X_t)$, where $\vec{\nu} = (1/2, 1/2)$.

(e) Let $J_0, J_1, J_2, ...$ be the jump times of the process. For each $i, j \in S$ and $k = 1, 2, ...$, find $\mathbb{P}_i(X_{J_k} = j)$. *Hint*: You should be able to find a pattern considering the cases k even and k odd.

(f) Given that $X_{J_2} = 2$, find the probability that the second sojourn time S_2 is at most 1; also find the probability that the third sojourn time S_3 is smaller than the second. That is, find $\mathbb{P}(S_2 \leq 1 \,|\, X_{J_2} = 2)$ and $\mathbb{P}(S_3 < S_2 \,|\, X_{J_2} = 2)$.

(g) Find the invariant distribution π in two ways: i) Finding $\lim_{t\to\infty} \mathbf{P}(t)$ and deducing π from this limit, and ii) solving $\vec{\pi}\,\mathbf{L} = \vec{\mathbf{0}}$.

(h) For each state $i \in S$, find the long run fraction of time the process spends in state i.

(i) For each state $i \in S$, find $\mathbb{E}_i[\rho_i]$.

(8) Suppose that $(X_t)_{t\geq 0}$ is a stationary discrete time Markov chain with generator $\mathbf{L}$ and state space S. Let $\mathbf{R}$ be the embedded stochastic matrix, and for every $i \in S$, let $\lambda_i = -[\mathbf{L}]_{ii}$. Further, let $S_0, S_1, S_2, ...$ be the sojourn times.

(a) Let $i \in S$. Show that

$$\mathbb{P}_i(S_n \leq t) = \sum_{j \in S} (1 - e^{-\lambda_j t})[\mathbf{R}^n]_{ij}.$$

Hint: $\mathbb{P}_i(S_n \leq t) = \sum_{j \in S} \mathbb{P}_i(S_n \leq t \,|\, X_{J_n} = j)\, \mathbb{P}_i(X_{J_n} = j)$.

(b) If the generator

$$\mathbf{L} = \begin{pmatrix} -2 & 1 & 1 \\ 2 & -3 & 1 \\ 1 & 0 & -1 \end{pmatrix}$$

is indexed by $S = \{1, 2, 3\}$, find the cumulative distribution function of the sojourn time S_3 with respect to $\mathbb{P}_i$ for each $i \in S$.

(9) Let $(X_t)_{t \geq 0}$ be a right continuous stationary discrete time Markov chain with state space S and transition semigroup $(\mathbf{P}(t))_{t \geq 0}$. Suppose that $\tau \geq 1$ is a stopping time with respect to this process. If $(X_t)_{t \geq 0}$ satisfies the strong Markov property, show that

$$\mathbb{P}(X_{\tau+2} = j, X_{\tau-1} = k \,|\, X_\tau = l) = [\mathbf{P}(2)]_{lj}\, [\mathbf{P}(1)]_{kl}\, \frac{\mathbb{P}(X_{\tau-1} = k)}{\mathbb{P}(X_\tau = j)}$$

and make sure to note where you're using the strong Markov property.

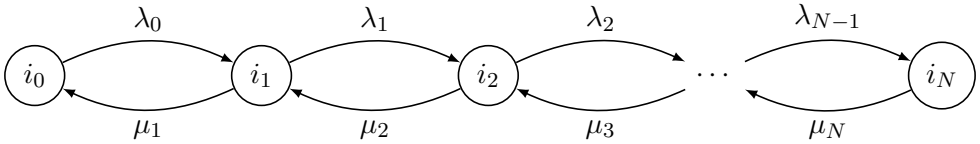
CHAPTER 8

Birth, Death, and Renewal Processes

1. Birth-Death Processes

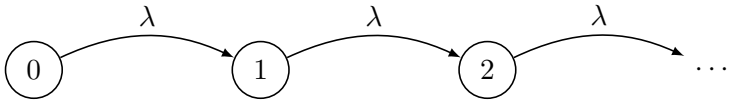
Birth-death processes are a specific type of continuous time Markov chain which are extremely useful in practice, and because of their nice form allow for quite a bit of theory development. For this section we will assume that the state space $S = \{i_0, i_1, ..., i_N\}$ with the possibility that $N = \infty$, when we have a discrete, but infinite state space. As in the previous chapter $(X_t)_{t\geq 0}$ will denote our stationary continuous time Markov chain, $(\mathbf{P}(t))_{t\geq 0}$ its transition semigroup, and $\mathbf{L}$ its infinitesimal generator.

DEFINITION 1.1. A *birth-death process* is a stationary continuous time Markov chain in which all jumps happen between consecutive neighbors. That is, the process has a rate diagram of the form:



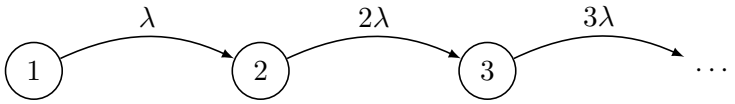
where the $\lambda_k \geq 0$ and $\mu_{k+1} \geq 0$ for all k. In the case that $\mu_k = 0$ for every k, then the process is known as a *pure birth process*, where as if $\lambda_k = 0$ for every k, then the process is known as a *pure death process*. △

EXAMPLE 1.1 (Poisson Process). A pure birth process when S is the collection of all non-negative integers and constant birth rate $\lambda > 0$



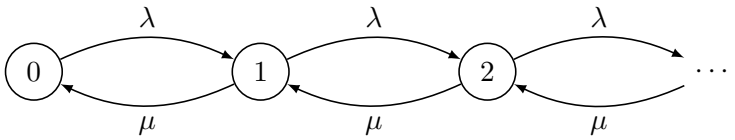
is a *Poisson process* with parameter λ. △

EXAMPLE 1.2 (Yule Process). A pure birth process when S is all positive integers such that $\lambda_k = k\lambda$ for some fixed $\lambda > 0$



is called a *Yule process*. △

EXAMPLE 1.3 ($M/M/1$ Queue). A birth-death process when S is all non-negative integers with fixed birth rate $\lambda > 0$ and death rate $\mu > 0$



is called an $M/M/1$ *queue*. The story of an $M/M/1$ queue goes as follows. Requests coming to a single server at an average rate λ, and the server will process each request in the order they are received with an average rate μ. The $M/M/1$ process $(X_t)_{t\geq 0}$ then represents the number of requests which are currently in the "system", meaning the number of requests which are with the operator or waiting to be processed. △

LEMMA 1.1. *A birth-death process is irreducible if and only if $\lambda_k > 0$ and $\mu_k > 0$ for every $k \in S$.*

PROOF. It is obvious from the rate diagram for a birth-death process that every state communicates with every other state if and only if none of the birth rates λ_k or death rates μ_k are zero. □

THEOREM 1.2. *Let $(\mathbf{P}(t))_{t\geq 0}$ be the transition semigroup for an irreducible birth-death process $(X_t)_{t\geq 0}$. Define $\theta_0 = 1$ and for $1 \leq n \leq N$,*

$$\theta_n = \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n}.$$

Further, define $\Theta = \sum_{n=0}^{N} \theta_n$.

1. *Each state is positive recurrent if and only if $\Theta < \infty$.*
2. *If $\Theta < \infty$, then the unique invariant distribution π of $(\mathbf{P}(t))_{t\geq 0}$ can be found as $\pi(j) = \dfrac{\theta_j}{\Theta}$ for every $j \in S$.*

PROOF. Without losing generality, let's assume that $S = \{0, 1, ..., N\}$ with $N = \infty$ possible. We will first show that $\vec{\pi}\,\mathbf{L} = \vec{\mathbf{0}}$ if and only if $\vec{\pi}$ is a scalar multiple of the vector $(1, \theta_1, \theta_2, ..., \theta_N)$.

To do so, the system of equations we will need to solve is given by,

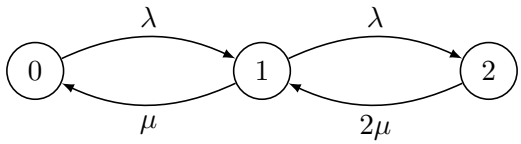
$$\begin{aligned}
0 &= -\lambda_0\pi(0) + \mu_1\pi(1) \\
0 &= \lambda_0\pi(0) - (\lambda_1 + \mu_1)\pi(1) + \mu_2\pi(2) \\
&\vdots \\
0 &= \lambda_{N-2}\pi(N-2) - (\lambda_{N-1} + \mu_{N-1})\pi(N-1) + \mu_N\pi(N) \\
0 &= \lambda_{N-1}\pi(N-1) - \mu_N\pi(N)
\end{aligned}$$

Starting from the top and solving down, the first equation tells us that $\pi(1)/\pi(0) = \lambda_0/\mu_1 = \theta_1$. Using this, dividing the second equation through by π_0, and solving for $\pi(2)/\pi(0)$ we have $\pi(2)/\pi(0) = \lambda_0\lambda_1/(\mu_1\mu_2) = \theta_2$. Continuing in this fashion, we find that $\pi(n)/\pi(0) = \theta_n$. In particular, $\vec{\pi}\,\mathbf{L} = \vec{\mathbf{0}}$ if and only if $\pi(n) = \pi(0)\,\theta_n$. Therefore, π has the form

$$\pi = (\pi_0, \pi_0\theta_1, \pi_0\theta_2, ..., \pi_0\theta_N) = \pi_0\,(1, \theta_1, \theta_2, ..., \theta_N).$$

We now are in a position to finish the proof of the theorem. With $\vec{\pi}\,\mathbf{L} = \vec{\mathbf{0}}$, Proposition 6.1 tells us that π will be an invariant distribution for $(\mathbf{P}(t))_{t\geq 0}$ if and only if we can normalize the entries of π so that they sum to 1. From what we have shown, this means we need to normalize the vector $(\theta_0, \theta_1, \theta_2, ..., \theta_N)$. Since $\sum_{n=0}^{N} \theta_n = \Theta$, we realize that $1/\Theta$ must be the normalizing factor. Hence, we see that π is an invariant distribution if and only if $\Theta < \infty$. From the results of Theorem 8.1, everything now follows. □

Example 1.4. A facility has two machines, each with fail independently at a time given by an exponential distribution with failure rate μ. A single repair person can repair a failed machine in a time exponentially distributed with repair rate λ. We also assume that time time of repair is independent of the failure times. If X_t is the number of operational machines at the facility at time $t \geq 0$, we aim to find the fraction of time are both machines not operational during the (long) lifetime of the facility. To answer this, let us first model X_t as a birth death process. The reasonable rate diagram for this model is given by



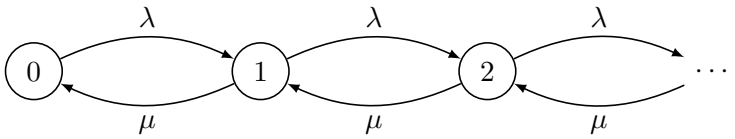
The reason that $\mu_2 = 2\mu$ is because if $T_1 \overset{d}{=} \text{Exp}(\mu)$ and $T_2 \overset{d}{=} \text{Exp}(\mu)$ are the failure times of the two machines, then with both machines operational (i.e., the process is in state 2), the first failure

time will happen at $\min(T_1, T_2) \stackrel{d}{=} \text{Exp}(\mu + \mu) = \text{Exp}(2\mu)$. Let us now use Theorem 6.4 to find the fraction of time the machines will be non-operational. The long run fraction of time there are no operational machines will be given by $\pi(0)$, where π is the invariant distribution for this process (which is guaranteed to exist since it is irreducible and has only finitely many states). Using Proposition 1.2, we will have $\pi(0) = \theta_0/\Theta = 1/\Theta$ where $\Theta = \theta_0 + \theta_1 + \theta_2$ and $\theta_0 = 1$, $\theta_1 = \lambda/\mu$ and $\theta_2 = \lambda^2/(2\mu^2)$. Hence,

$$\pi_0 = \frac{1}{1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{2\mu^2}} = \frac{2\mu^2}{2\mu^2 + 2\mu\lambda + \lambda^2}.$$

△

EXAMPLE 1.5. Consider the $M/M/1$ queue



for some $\lambda > 0$ and $\mu > 0$. We find that $\theta_n = \lambda^n/\mu^n = (\lambda/\mu)^n$. Then

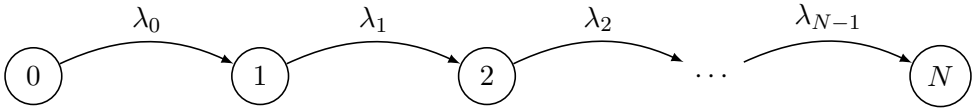
$$\Theta = \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n = \begin{cases} \frac{\mu}{\mu - \lambda} & \lambda < \mu \\ \infty & \lambda \geq 0 \end{cases}$$

where we used the geometric series formula to deduce the last equality. In particular, there is an invariant distribution for this process if and only if $\lambda < \mu$. Moreover, relating to the story corresponding to an $M/M/1$ queue (see Example 1.3), if $\lambda < \mu$, then the long run fraction of time there are no requests in the system is $\pi(0) = \frac{1}{\Theta} = \frac{\mu - \lambda}{\mu}$, and the long run fraction of time there are $n \geq 1$ requests in the system is $\pi(n) = \frac{\theta_n}{\Theta} = \frac{\lambda^n(\mu - \lambda)}{\mu^{n+1}}$. △

2. The Poisson Process

We will focus on a very special pure birth process, called the Poisson process, in this section. However, we start with a quick observation about any pure birth process.

PROPOSITION 2.1. *Consider a pure birth process $(X_t)_{t \geq 0}$ with rate diagram*



where, as usual, $N = \infty$ is perfectly allowed. Assume that $\lambda_i > 0$ for every $1 \leq i \leq N - 1$. Let $S_0, S_1, ...$ be the sojourn times of this process. With respect to the probability $\mathbb{P}_i$, the sojourn times $S_0, S_1, S_2, ..., S_{N-i}$ are independent exponential random variables with parameters

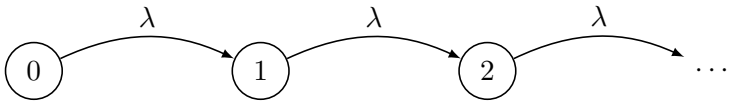
$\lambda_i, \lambda_{i+1}, \lambda_{i+2}, ..., \lambda_N$, respectively. If $N = \infty$ we interpret this as the claim that the sojourn times $S_0, S_1, S_2, ...$ are independent exponential random variables with parameters $\lambda_i, \lambda_{i+1}, \lambda_{i+2}, ...$ If $N < \infty$, then λ_N is defined as 0, and hence S_{N-i} is constantly ∞ (i.e., state N is an absorption state, and hence the process "holds" there eternally).

PROOF. Since $(X_t)_{t\geq 0}$ is a pure birth process, then

$$\mathbb{P}_i = \mathbb{P}(\,\cdot \,|\, X_{J_0} = i, X_{J_1} = i+1, ..., X_{J_{N-i}} = N).$$

From Theorem 4.2 (2), the result follows. □

DEFINITION 2.1 (Poisson process, first definition). The pure birth process $(N_t)_{t\geq 0}$ defined by the rate diagram



is called a *Poisson process* with parameter λ (it is assumed $\lambda > 0$). △

We now give an alternate definition of the Poisson process that is more common (and justifies the name Poisson begin involved with this process).

DEFINITION 2.2 (Poisson process, second definition). Let $(N_t)_{t\geq 0}$ be a right continuous stochastic process whose state space S is the non-negative integers. Suppose further that the following hold.

(1) For every $0 \leq s < t$, the random variable $N_t - N_s$ is a Poisson random variable with parameter $\lambda(t-s)$.
(2) (Independent Increments) The evolutions of the process during non-overlapping increments of time are independent. That is, for any sequence of times $0 \leq t_0 < t_1 < \cdots < t_n$, the random variables $N_{t_1} - N_{t_0}, N_{t_2} - N_{t_1}, ..., N_{t_n} - N_{t_{n-1}}$ are independent with respect to $\mathbb{P}_i$ for any state $i \in S$.

Then we call $(N_t)_{t\geq 0}$ a *Poisson process* with parameter λ. If we also assume

(3) $N_0 = 0$ (i.e., when we are referencing this process with respect to $\mathbb{P}_0$)

we call $(N_t)_{t\geq 0}$ a *standard Poisson process* with parameter λ. △

REMARK 2.1. Suppose that $(N_t)_{t\geq 0}$ is a Poisson process. If $N_0 \neq 0$, then the *centered* process $(N_t - N_0)_{t\geq 0}$ is a standard Poisson process, so it is common to always assume that the Poisson process is standard. △

THEOREM 2.2. *Definitions 2.1 and 2.2 define the same process.*

PROOF. We prove below in Proposition **??** that the standard process $(N_t)_{t\geq 0}$ in Definition 2.2 is a stationary continuous time Markov chain; the proof of this heavily uses the independent increments. By Definition 2.2 (1), we easily find that the transition semigroup of $(N_t)_{t\geq 0}$ is

$$\mathbf{P}(t) = \begin{pmatrix} e^{-\lambda t} & \lambda t e^{-\lambda t} & \frac{\lambda^2 t^2}{2} e^{-\lambda t} & \frac{\lambda^3 t^3}{3!} e^{-\lambda t} & \frac{\lambda^4 t^4}{4!} e^{-\lambda t} & \cdots \\ 0 & e^{-\lambda t} & \lambda t e^{-\lambda t} & \frac{\lambda^2 t^2}{2} e^{-\lambda t} & \frac{\lambda^3 t^3}{3!} e^{-\lambda t} & \cdots \\ 0 & 0 & e^{-\lambda t} & \lambda t e^{-\lambda t} & \frac{\lambda^2 t^2}{2!} e^{-\lambda t} & \cdots \\ 0 & 0 & 0 & e^{-\lambda t} & \lambda t e^{-\lambda t} & \cdots \\ 0 & 0 & 0 & 0 & e^{-\lambda t} & \ddots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}$$

By taking the time derivative of this transition semigroup and setting time to 0, we find that the generator of this semigroup is

$$\mathbf{L} = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & 0 & \cdots \\ 0 & -\lambda & \lambda & 0 & 0 & \cdots \\ 0 & 0 & -\lambda & \lambda & 0 & \cdots \\ 0 & 0 & 0 & -\lambda & \lambda & \cdots \\ 0 & 0 & 0 & 0 & -\lambda & \ddots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}$$

You will recognize $\mathbf{L}$ as the generator of the pure birth process in Definition 2.1. Therefore Definition 2.1 and Definition 2.2 both define stationary continuous time Markov chains with identical generators, and hence must be defining the same process. □

We now have two alternate but equivalent perspectives of the Poisson process $(N_t)_{t\geq 0}$ which will help us deduce several facts. For one, the proof of Theorem 2.2 gives us the generator $\mathbf{L}$ and the the corresponding transition semigroup $\mathbf{P}(t) = e^{t\mathbf{L}}$ of the Poisson process. Further, from Proposition 2.1 and Definition 2.1, we immediately get the following Proposition.

PROPOSITION 2.3. *Let $(N_t)_{t\geq 0}$ be a Poisson process with parameter $\lambda > 0$. For any state i, with respect to $\mathbb{P}_i$, the sojourn times $S_0, S_1, S_2, \ldots$ are independent identically distributed exponential random variables with parameter λ.*

Moreover, from Definition 2.2, we can easily conclude the following.

PROPOSITION 2.4. *Suppose that $(N_t)_{t\geq 0}$ is a Poisson process. Let $0 \leq s < t$. Then for any state $i \in S$,*

(1) $\mathbb{E}_i[N_t - N_s] = \lambda(t-s)$. *In particular, if $i = 0$, then $\mathbb{E}_0[N_t] = \lambda t$.*
(2) $\mathrm{Var}_i(N_t - N_s) = \lambda(t-s)$. *In particular, if $i = 0$, then $\mathrm{Var}_0(N_t) = \lambda t$.*
(3) $\mathrm{Cov}_i(N_t, N_s) = \mathrm{Var}_i(N_s)$. *In particular, if $i = 0$, then $\mathrm{Cov}_0(N_t, N_s) = \lambda s$.*

PROOF. The important point here is that $N_t - N_s$ is distributed as a Poisson random variable with parameter $\lambda(t-s)$. If $i = 0$, then with $N_0 = 0$ and so $N_t - N_0 = N_t$ is distributed as a Poisson random variable with parameter λt. For (3), we also need to employ the fact that $N_t - N_s$ and N_s are independent random variables. With this observation, $\mathrm{Cov}_i(N_t, N_s) = \mathrm{Cov}_i(N_t - N_s + N_s, N_s) = \mathrm{Cov}_i(N_t - N_s, N_s) + \mathrm{Cov}_i(N_s, N_s) = 0 + \mathrm{Var}_i(N_s)$. □

Continuing in the vein, we have the following conditional result.

PROPOSITION 2.5. *Let $(N_t)_{t\geq 0}$ be a Poisson process. For any initial state $i \in S$ and bounded or non-negative function $f : [0, \infty) \to \mathbb{R}$, if $s < t$ then*

$$\mathbb{E}_i[f(N_t) \,|\, N_s] = \sum_{m=0}^{\infty} f(m + N_s) \frac{\lambda^m (t-s)^m}{m!} e^{-\lambda(t-s)}.$$

In particular, this implies

$$\mathbb{E}_i[N_t \,|\, N_s] = \lambda(t-s) + N_s.$$

PROOF. Let $0 \leq s < t$ and recall that the independent increment property implies that $N_t - N_s$ is independent from N_s. Now, for any state $k \geq i$,

$$\begin{aligned}\mathbb{E}_i[f(N_t) \,|\, N_s = k] = \mathbb{E}_i[f(N_t - N_s + N_s) \,|\, N_s = k] = \mathbb{E}_i[f(N_t - N_s + k) \,|\, N_s = k] \\ = \mathbb{E}_i[f(N_t - N_s + k)].\end{aligned}$$

Since $N_t - N_s$ is distributed as a Poisson random variable with parameter $\lambda(t-s)$,

$$\mathbb{E}_i[f(N_t - N_s + k)] = \sum_{m=0}^{\infty} f(m + k) \frac{\lambda^m (t-s)^m}{m!} e^{-\lambda(t-s)}.$$

This shows that

$$\mathbb{E}_i[f(N_t) \,|\, N_s = k] = \sum_{m=0}^{\infty} f(m + k) \frac{\lambda^m (t-s)^m}{m!} e^{-\lambda(t-s)}$$

from which the first claim follows by replacing k with N_s (which is what is needed to find $\mathbb{E}[f(N_t) \mid N_s]$) on the righthand side of the last equality. For the second claim, setting $f(x) = x$,

$$\begin{aligned}\mathbb{E}_i[N_t \mid N_s] &= \sum_{m=0}^{\infty}(m + N_s)\frac{\lambda^m(t-s)^m}{m!}e^{-\lambda(t-s)} \\ &= \sum_{m=0}^{\infty} m\,\frac{\lambda^m(t-s)^m}{m!}e^{-\lambda(t-s)} + N_s\sum_{m=0}^{\infty}\frac{\lambda^m(t-s)^m}{m!}e^{-\lambda(t-s)} \\ &= \lambda(t-s) + N_s.\end{aligned}$$

which proves the "in particular" claim. However, we could have proved this more simply with the following (equivalent, but seemingly nicer) manipulation

$$\mathbb{E}[N_t \mid N_s] = \mathbb{E}[N_t - N_s \mid N_s] + \mathbb{E}[N_s \mid N_s] = \mathbb{E}[N_t - N_s] + N_s = \lambda(t-s) + N_s$$

which uses several properties of conditional expectation we are now familiar with. □

2.1. Sums of Independent Poisson Processes.

2.2. Compound Poisson Process. We mention here one generalization of the Poisson process, called compound Poisson process. To motivate the definition, let us make a quick observation. If $(N_t)_{t\geq 0}$ is a Poisson process, then $N_t = \sum_{i=1}^{N_t} 1$, where we use the standard summation convention that $\sum_{i=l}^{m}$ is 0 whoever $m < l$. This equality between N_t and the sum of 1 with itself N_t times is clear since we will always recover the value N_t with this sum (note that this observation holds true for any process taking values in the non-negative integers). We now generalize this by adding some randomness to the summand.

Definition 2.3. Let $(N_t)_{t\geq 0}$ be a Poisson process and $\{R_i\}_{i=1}^{\infty}$ a sequence of i.i.d. random variables which are also independent of the process. A *compound Poisson process* is the continuous time stochastic process $(X_t)_{t\geq 0}$ defined by

$$X_t = \sum_{i=1}^{N_t} R_i.$$

△

Example 2.1. Suppose that customer arrivals to a certain online store according to a standard Poisson process $(N_t)_{t\geq 0}$ with rate $\lambda = 4$ customers per hour. Each customer can either leaving without purchase, buy item A for \$5, buy item B for \$10, or buy both items A and B for the discounted price of \$13. We assume that the purchase choice of each customer is independent of the

other customers. Let $\{R_i\}_{i=1}^{\infty}$ be a sequence of i.i.d. random variables taking values in $\{0, 5, 10, 13\}$, where R_i represents the amount the ith customer pays to the store. Then, the compound Poisson process $(X_t)_{t\geq 0}$ defined by $X_t = \sum_{i=1}^{N_t} R_i$ represents the amount of money made by the store at time t. In Exercise 4 (8) we explore this example in more detail. △

THEOREM 2.6. *Let $(X_t)_{t\geq 0}$ be a compound Poisson process $X_t = \sum_{i=1}^{N_t} R_i$, as in Definition 2.3. Assume that $\lambda > 0$ is the parameter of the Poisson process (N_t). Then the following hold.*

(1) $\mathbb{E}_0[X_t] = \mathbb{E}_0[N_t]\,\mathbb{E}_0[R_i] = \lambda t\,\mathbb{E}_0[R_1]$.
(2) $\mathrm{Var}_0(X_t) = \mathbb{E}[N_t]\,\mathrm{Var}(R_1) + \mathbb{E}[R_1]^2\,\mathrm{Var}(N_t) = \lambda t\,\mathbb{E}_0[R_1^2]$.

PROOF. This is an application of Lemma **??**. Indeed, (1) is immediate from the referenced Lemma, and for (2),

$$\mathrm{Var}_0(X_t) = \mathbb{E}[N_t]\,\mathrm{Var}(R_1) + \mathbb{E}[R_1]^2\,\mathrm{Var}(N_t) = \lambda t(\mathrm{Var}(R_1) + \mathbb{E}[R_1]^2) = \lambda t\mathbb{E}[R_1^2].$$

This concludes the proof. □

3. Renewal Processes

This section moves towards a useful class of generalized Poisson processes, called *renewal processes*, which takes a step away from Markov processes. In the arena of renewal processes, we consider a pure birth type process on the non-negative integers, where the sojourn times are no longer required to be exponentially distributed. By the jump-hold description, if the sojourn times are not exponentially distributed, then the process will not be a Markov chain. However, we still have techniques to help understand such processes, and our goal will be to understand some of these techniques.

We already have a understood vocabulary to address many aspects of continuous time Markov chains, much of which we can carry over to renewal processes by analogy. However, it is apparent that the vocabulary in regards to renewal processes takes its own originality.

DEFINITION 3.1. Let $\{S_k\}_{k=1}^{\infty}$ be i.i.d. random variables which take only positive values. Let $J_0 = 0$ and for $n \in \mathbb{N}$, let $J_n = S_0 + S_1 + \cdots + S_{n-1}$. For $t \geq 0$, define $N_t = \max\{k : S_0 + S_1 + \cdots + S_{k-1} \leq t\}$. Then for each k we call S_k an *inner-renewal time*, we call J_n the nth *renewal time* corresponding to $\{S_k\}_{k=0}^{\infty}$, and we call $(N_t)_{t\geq 0}$ the *renewal process* corresponding to $\{S_k\}_{k=0}^{\infty}$. △

EXAMPLE 3.1. Let $\lambda > 0$ and suppose that the inner-renewal times have exponential distributions $S_k \overset{\text{dist}}{=} \mathrm{Exp}(\lambda)$. Then the corresponding renewal process $(N_t)_{t\geq 0}$ is the now-familiar Poisson process. △

DEFINITION 3.2. For a renewal process $(N_t)_{t\geq 0}$, the associated *renewal function*, $M(t)$, is given by $M(t) = \mathbb{E}[N_t]$. △

THEOREM 3.1 (Elementary Renewal Theorem). *Let $\{S_k\}_{k=0}^{\infty}$ be inner-renewal times. Given the corresponding renewal process $(N_t)_{t\geq 0}$ and renewal function $M(t)$ the following to equalities hold.*

$$\lim_{t\to\infty} \frac{N_t}{t} = \frac{1}{\mathbb{E}[S_0]}$$

with probability 1*, and*

$$\lim_{t\to\infty} \frac{M(t)}{t} = \frac{1}{\mathbb{E}[S_0]}.$$

Here, as usual, $1/\infty$ is to be understood as 0*.*

REMARK 3.1. In the Elementary Renewal Theorem (ERT), we say that

$$\lim_{t\to\infty} \frac{N_t}{t} = \frac{1}{\mathbb{E}[S_0]}$$

with probability 1. This is because for every $t \geq 0$, N_t is a random variable, where as $1/\mathbb{E}[S_0]$ is just a fixed constant. Therefore, the limit on the lefthand side of the equality is the convergence of random variables, which should result in a random variable X. The ERT then states that $\mathbb{P}(X = 1/\mathbb{E}[S_0]) = 1$. On the other hand, the limit

$$\lim_{t\to\infty} \frac{M(t)}{t} = \frac{1}{\mathbb{E}[S_0]}$$

is truly a converge of real numbers, since $M(t)/t$ is just some value in $\mathbb{R}$ for each t. △

EXAMPLE 3.2. The timing belt in a car engine should be replaced occasionally throughout the life of a car. A sensible replacement strategy for a car owners to follow is for the timing belt to be replaced after running for a certain fixed interval of time $T > 0$, unless the timing belt breaks before this time T, in which case the belt should be replaced when broken. Assume that each timing belt will break while being used according to an exponential random variable with rate $\lambda > 0$. Letting S_k represent the time until replacement of the kth timing belt, we have the sensible kth replacement modeled by $S_k = \min\{X_k, T\}$, where $\{X_k\}_{k=0}^{\infty}$ are iid exponential random variables with parameter λ. Now we can model the number of replacements by the renewal process $(N_t)_{t\geq 0}$ with inner-renewal times $\{S_k\}_{k=0}^{\infty}$. In this model, N_t represents the number of timing belt replacements done by time t.

As many are unfortunately aware of, if the timing belt breaks before replacement, the cost of replacement is significantly more than being replaced when not broken (this is due to damage to the engine caused by the break). Assume that it costs $a > 0$ dollars to replace the belt when it is

unbroken, and $b > a$ dollars to replace the belt when it is broken. Then the cost due to timing belt replacement at time t is

$$C_t = a \sum_{k=0}^{N_t-1} 1_{\{T<X_k\}} + b \sum_{k=0}^{N_t-1} 1_{\{X_k \le T\}}$$

where, as usual, the sum $\sum_{k=0}^{-1}$ is defined as 0. To make sense of this cost function, the sum $\sum_{k=0}^{N_t-1} 1_{\{T<X_k\}}$ counts the number of replacements that have happened by time t in which the belt was unbroken (since $T < X_k$, where X_k is the exponential breaking time of the belt); the sum $\sum_{k=0}^{N_t-1} 1_{\{X_k \le T\}}$ counts the number of replacements that have happened by time t in which the belt was broken (since $X_k \le T$). Therefore, with this belt replacement strategy, the expected cost at time t is,

$$\begin{aligned}\mathbb{E}[C_t] &= a\,\mathbb{E}[N_t - 1]\mathbb{P}(T < X_k) + b\,\mathbb{E}[N_t - 1]\mathbb{P}(X_k \le T) \\ &= a\,\mathbb{E}[N_t - 1]e^{-\lambda T} + b\,\mathbb{E}[N_t - 1](1 - e^{-\lambda T}) = \mathbb{E}[N_t - 1]\big[ae^{-\lambda T} + b(1 - e^{-\lambda T})\big]\end{aligned}$$

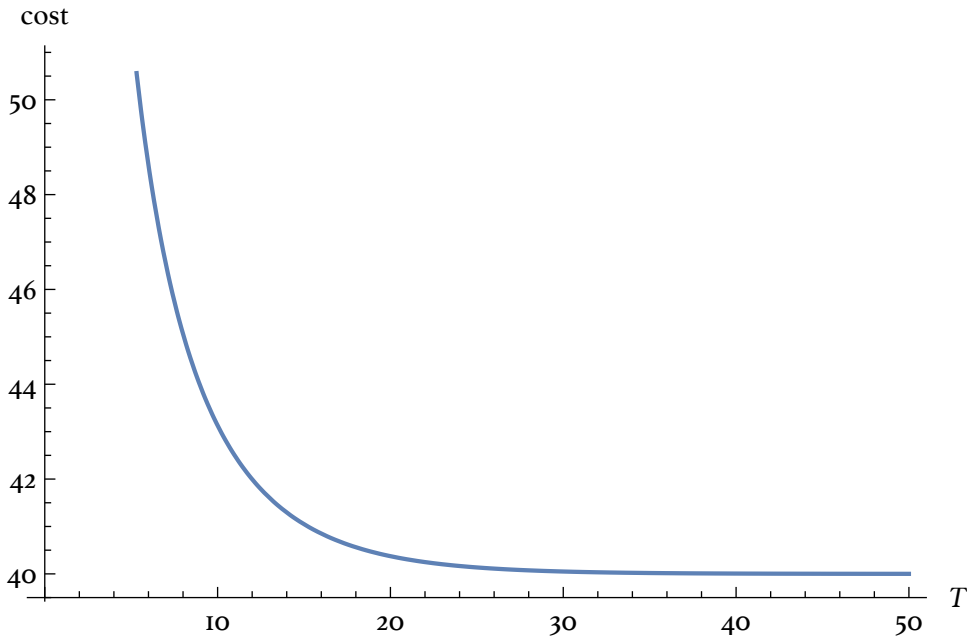
According to the Elementary Renewal Theorem 3.1, the long run cost per time is then

$$\lim_{t\to\infty} \frac{\mathbb{E}[C_t]}{t} = \lim_{t\to\infty} \frac{\mathbb{E}[N_t - 1]}{t}\big[ae^{-\lambda T} + b(1 - e^{-\lambda T})\big] = \frac{ae^{-\lambda T} + b(1 - e^{-\lambda T})}{\mathbb{E}[S_1]}.$$

You will show in Exercise 4 (9) that $\mathbb{E}[S_1] = (1 - e^{-\lambda T})/\lambda$. Therefore, we have

$$c(T) = \lambda\Big[\frac{a\,e^{-\lambda T}}{1 - e^{-\lambda T}} + b\Big]$$

The graph of $c(T)$ as a function of T (with $\lambda = 1/5, a = 100$, and $b = 200$) is



△

3.1. Lifetimes. If the current time is t, then N_t renewals have already occurred. The previous renewal took place at time W_{N_t} and the next renewal will take place at time W_{N_t+1}. It is of interest to ask how much longer from now until the next renewal takes place, how long has it been until now since the last renewal occurred, and what is the total lifetime of the current renewal.

DEFINITION 3.3. Let $(N_t)_{t\geq 0}$ be a renewal process with inner-renewal times $\{S_k\}_{k=0}^{\infty}$ and renewal times $\{J_n\}_{n=0}^{\infty}$.

(1) The *excess lifetime* or *residual lifetime* is the time-dependent random variable

$$\gamma_t = J_{N_t+1} - t$$

(2) The *current lifetime* or *age* is the time-dependent random variable

$$\delta_t = t - J_{N_t}$$

(3) The current *total lifetime* is the time-dependent random variable

$$\beta_t = \gamma_t + \delta_t = J_{N_t+1} - J_{N_t} = S_{N_t}$$

△

PROPOSITION 3.2. *We have equality between the following events.*

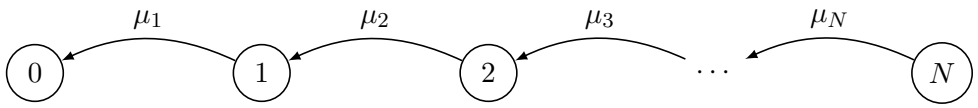
(1) *For* $x > 0$, $\{\gamma_t > x\} = \{N_{t+x} - N_t = 0\}$.

(2) *For* $0 < x < t$, $\{\delta_t \geq x\} = \{N_t - N_{t-x} = 0\}$.

(3) *For* $0 < x < t$ *and* $y > 0$, $\{\delta_t \geq x, \gamma_t > y\} = \{N_{t+y} - N_{t-x} = 0\}$.

4. Exercises

(1) Consider the following rate diagram for a pure death process $(X_t)_{t\geq 0}$:

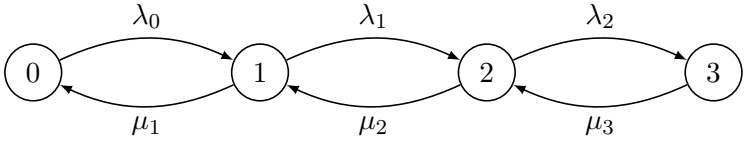


(a) For $t \geq 0$, what is $\mathbb{P}_2(X_t = 0)$?

(b) For $t \geq 0$, what is $\mathbb{P}_3(X_t = 0)$?

(c) Do you see a general formula for $\mathbb{P}_N(X_t = 0)$?

(2) Consider a birth-death process $(X_t)_{t\geq 0}$ with rate diagram given by



(a) Draw a possible sample path of $(X_t)_{t\geq 0}$.

(b) Find the generator $\mathbf{L}$ for this process.

(c) For each $i, j \in S$, explicitly write out $[\dot{\mathbf{P}}(t)]_{ij} = [\mathbf{L}\mathbf{P}(t)]_{ij}$ and $[\dot{\mathbf{P}}(t)]_{ij} = [\mathbf{P}(t)\mathbf{L}]_{ij}$. That is, for each fixed $i, j \in S$ you should write out the differential equation in terms of the λ's, μ's, and terms of the form $[\mathbf{P}(t)]_{km}$.

(d) Assuming that none of the λs nor μs are 0, what is invariant distribution for this process?

(3) (From Taylor and Karlin, chapter IV, number 4.7) A system consists of 3 machines and 2 repairmen. At most 2 machines can operate at any time. The amount of time that an operating machine works before breaking down is exponentially distributed with mean 5 hours. The amount of time that it takes a single repairman to fix a machine is exponentially distributed with mean 4 hours. Only one repairman can work on a failed machine at any given time. Let X_t be the number of machines in operating condition at time t.

(a) Calculate the long run probability distribution of the process (X_t).

(b) If an operating machine produces 100 units of output per hour, what is the long run output per hour from the factory?

(4) A server with N cores can process up to N requests simultaneously where each core independently process a request with a mean time of $1/\mu$ for some $\mu > 0$. Requests enter the server at rate λ for some $\lambda > 0$, unless the server is full (with all N cores currently

processing other requests) in which case the request is rejected. Let X_t be the number of cores which are occupied at time t and assume that $(X_t)_{t\geq 0}$ is a birth-death process. In terms of λ and μ, approximately what percentage of the time is the server busy in the long run?

(5) Let $(N_t)_{t\geq 0}$ be a standard Poisson process with parameter $\lambda > 0$. Let $0 \leq s < t$.
   (a) For each non-negative integers k and n, find $\mathbb{P}(N_s = k, N_t = n)$. Your answer will depend on λ, s, and t. Once you have done this, realize that you have found the joint mass function of N_s and N_t. *Hint*: $\mathbb{P}(N_s = k, N_t = n) = \mathbb{P}(N_s = k, N_t - N_s = n - k)$.
   (b) For each non-negative integers k and n with $k \leq n$, find

$$\mathbb{P}(N_s = k \,|\, N_t = n).$$

   Your answer will depend on λ, s, and t.

(6) Let $(N_t)_{t\geq 0}$ be a standard Poisson process. For $0 \leq s < t$, find $\mathbb{E}_0[N_s \,|\, N_t]$. *Hint*: First find $\mathbb{E}_0[N_s \,|\, N_t = n]$ by justifying the equality

$$\mathbb{E}_0[N_s \,|\, N_t = n] = \sum_{k=0}^{n} k\, \mathbb{P}_0(N_s = k \,|\, N_t = n)$$

and use Exercise (5) (b).

(7) Let $(N_t)_{t\geq 0}$ be a standard Poisson process with parameter $\lambda > 0$. For every positive integer n, show that N_n can be written as the sum of n independent Poisson random variables with parameter λ.

(8) In Example 2.1, a compound Poisson process $(X_t)_{t\geq 0}$ is used to described the money earned by a certain shop as a function of time t. Let $p : \{0, 5, 10, 13\} \to [0, 1]$ be the probability mass function of R_i (it doesn't matter which i since the R_i are assumed identically distributed).
   (a) Find $\mathbb{E}[X_t]$ and $\mathrm{Var}(X_t)$ in terms of the probability mass function p.
   (b) If $p(0) = 1/2,\ p(5) = 1/4,\ p(10) = 1/16$, and $p(13) = 3/16$, how much income can the online store expect each day?

(9) Here we expand on Example 3.2.
   (a) In the derivation of the average long run cost per unit time, we used the Elementary Renewal Theorem to conclude that

$$\lim_{t\to\infty} \frac{\mathbb{E}[C_t]}{t} = \frac{ae^{-\lambda T} + b(1 - e^{-\lambda T})}{\mathbb{E}[S_1]}$$

which, when we look at what limit was taking place, suggests that

$$\lim_{t\to\infty} \frac{\mathbb{E}[N_t - 1]}{t} = \frac{1}{\mathbb{E}[S_1]}.$$

Using the Elementary Renewal Theorem, show that this is true. In fact, show that for any constant value α

$$\lim_{t\to\infty} \frac{\mathbb{E}[N_t + \alpha]}{t} = \frac{1}{\mathbb{E}[S_1]}.$$

(where $\alpha = -1$ was just a special case for this example).

(b) Recall that $S_1 = \min\{X_1, T\}$, where X_1 is an exponential random variable with parameter $\lambda > 0$. Calculate $\mathbb{E}[S_1]$ (which will explicitly depend on T), and hence show that our formula

$$c(T) = \lambda\Big[\frac{ae^{-\lambda T}}{1 - e^{-\lambda T}} + b\Big]$$

is valid.

(c) In the graph of $c(T)$ given (with $\lambda = 1/5$, $a = 100$, and $b = 200$) it is clear that the minimal value of $c(T)$ occurs when $T = \infty$. Show that this is true using calculus; find the derivative of $c(T)$ as a function of T, and see that $c(T)$ is always negative (meaning the the cost per time is always decreasing) for $T > 0$.

(d) Suppose instead of the timing belt having exponential breaking times, the breaking times were modelled by a uniform random variable on the interval $(5, 7)$, where we'll assume the units of time are years. Then the sojourn times are $S_i = \min\{U_i, T\}$ where $\{U_i\}_{i=1}^{\infty}$ are the iid uniform breaking times of the belts. Rederive the cost per time function $c(T)$ and minimize it with respect to T. *Note*: If $T \leq 5$, then $S_i = T$ for every i, and if $T \geq 7$ then $S_i = U_i$ for every i; so the real interest is when $5 < T < 7$.

(10) A *renewal rewards* process is the generalization of a compound Poisson process to the setting of renewal processes. The setup is as follows. Suppose that $(N_t)_{t\geq 0}$ is a renewal process with inner-renewal times $\{S_k\}_{k=0}^{\infty}$. Let $\{R_i\}_{i=1}^{\infty}$ be a collection of iid random variables (which are not necessarily independent of the renewal process). The process $(X_t)_{t\geq 0}$ defined by $X_t = \sum_{i=1}^{N_t} R_i$ is a renewal rewards process (again, as usual, if the upper limit of the sum is smaller than the lower limit, we define the sum value to be 0). The intuition for such a process is similar to that of a compound Poisson process, with exception that the arrivals $(N_t)_{t\geq 0}$ which are rewarded (with rewards R_i) occur as a renewal process rather than strictly a Poisson process.

One of the cornerstone results encountered when first being introduced to the theory of renewal rewards processes is that

$$\lim_{t\to\infty} \frac{X_t}{t} = \frac{\mathbb{E}[R_1]}{\mathbb{E}[S_1]} \tag{RR}$$

This is the version of the Elementary Renewal Theorem in the setting of renewal rewards processes. In this problem, we derive (RR).

(a) Briefly justify that

$$\lim_{t\to\infty} \frac{\sum_{i=1}^{N_t} R_k}{N_t} = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} R_i}{n}.$$

and from this, use the Strong Law of Large Numbers to conclude that

$$\lim_{t\to\infty} \frac{\sum_{i=1}^{N_t} R_k}{N_t} = \mathbb{E}[R_1].$$

Recall: The Strong Law of Large Numbers (SLLN) states that

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} Y_i}{n} = \mathbb{E}[Y_i]$$

whenever $\{Y_i\}_{i=1}^{\infty}$ are iid random variables. That is, SLLN states that the sample average converges to the true average as the sample gets larger.

(b) From here, deduce that $\lim_{t\to\infty} \frac{X_t}{t} = \frac{\mathbb{E}[R_t]}{\mathbb{E}[S_1]}$.

Hint: Notice that

$$\frac{X_t}{t} = \frac{\sum_{i=1}^{N_t} R_i}{t} = \frac{\sum_{i=1}^{N_t} R_i}{N_t} \frac{N_t}{t}$$

so you can apply part (a) to the first term and the Elementary Renewal Theorem to the second.

Part 4

Appendices

Appendices

CHAPTER A

Some Matrix Analysis

1. Matrix Multiplication

2. Matrix Subsetting

Suppose that $\mathbf{B}$ is a square matrix indexed by S. If $A \subseteq S$ is some subset, we will write $\mathbf{B}_A$ as the square sub-matrix of $\mathbf{B}$ whose entries are only those indexed by elements in A; equivalently, $\mathbf{B}_A$ is the square matrix remaining when removing all rows and columns indexed by $A^c = S \setminus A$. For example, suppose that $S = \{j, k, l, m, n\}$ and $\mathbf{B}$ is indexed by S with

$$\mathbf{B} = \begin{pmatrix} 1 & 2 & 1 & 3 & 1 \\ -1 & 5 & 1/2 & 2 & -3 \\ 9 & 8 & 7 & 6 & 2 \\ -2 & 4 & -6 & 8 & -10 \\ 4 & 3/2 & 1 & -5 & 6 \end{pmatrix}$$

Let $A = \{j, l\}$. Then by removing all rows and columns indexed by elements not in A we find,

S	j	k	l	m	n
j	1	2	1	3	1
k	-1	5	$1/2$	2	-3
l	9	8	7	6	2
m	-2	4	-6	8	-10
n	4	$3/2$	1	-5	6

$$\implies \mathbf{B}_A = \begin{pmatrix} 1 & 1 \\ 9 & 7 \end{pmatrix}$$

Alternatively, if we remove those rows and columns indexed by A, we find

S	j	k	l	m	n
j	1	2	1	3	1
k	-1	5	$1/2$	2	-3
l	9	8	7	6	2
m	-2	4	-6	8	-10
n	4	$3/2$	1	-5	6

$$\implies \mathbf{B}_{A^c} = \begin{pmatrix} 5 & 2 & -3 \\ 8 & 7 & -10 \\ 3/2 & -5 & 6 \end{pmatrix}$$

We will also apply this notation to vectors indexed by S. If $\mathbf{v} = (1, 2, 3, 4, 5)$ is a vector indexed by $S = \{j, k, l, m, n\}$ and $A = \{j, l\}$, then $\mathbf{v}_A = (1, 3)$ and $\mathbf{v}_{A^c} = (2, 4, 5)$. Indeed, with $A = \{j, l\}$,

S	j	k	l	m	n
$\mathbf{v}$	1	2	3	4	5

$\implies \mathbf{v}_A = (1, 3)$

and, with $A^c = \{k, m, n\}$,

S	j	k	l	m	n
$\mathbf{v}$	1	2	3	4	5

$\implies \mathbf{v}_{A^c} = (2, 4, 5)$